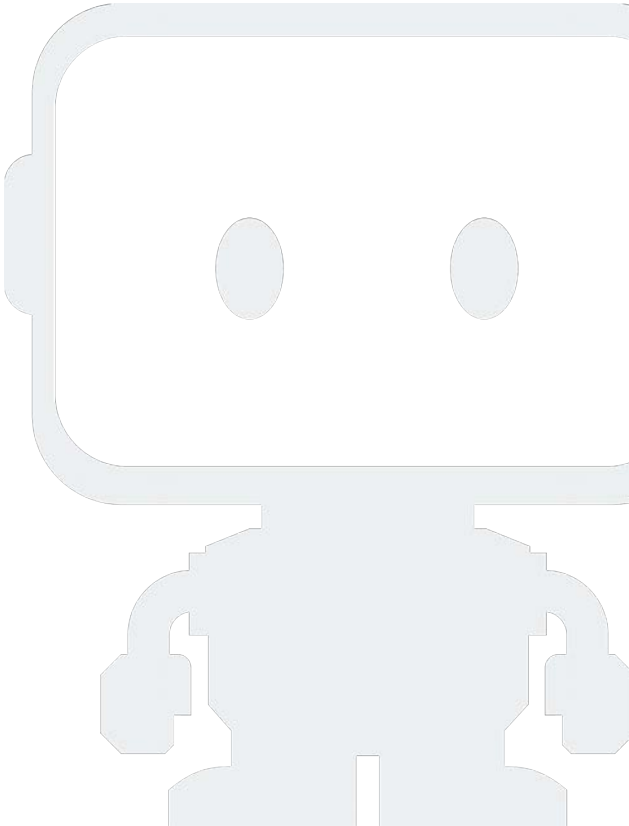
DataRobot

# Paxata

## Data Prepのドキュメント

バージョン：2021.2 2023年04月25日



# Paxata Data Prep

DataRobot Data Prepを使用すると、機械学習用に複数のソースからデータを収集、調査、準備できます。データとその準備に使用したステップを保存して共有できます。

DataRobot Data Prepについて考えるときは、次のことを考えてください。

- 準備するデータセットを保存するライブラリ
- データ準備を行うプロジェクト
- インポート、クリーンアップ、および組み合わせるデータ

このセクションでは、機械学習のためにデータをクリーニングして準備するData Prepの使い方を説明します。

| トピック | 説明... |
|---|---|
| Data Prepの開始 | Data Prepのクイックスタートおよび Data Prepのライブラリとプロジェクトの基本のツアーを完了します。 |
| データセットの操作 | Data Prepコネクターを設定します。 |
| データセットの操作 | データセットをインポートし、エクスポート、プロファイリング、更新などの他のデータセット操作を実行します。 |
| プロジェクトツールの操作 | Data Prepプロジェクトツールを使用して、データをクリーンアップして整形します。 |
| 列データの操作 | 幅広い列操作を使用して列を更新します。 |
| データソースへの接続 | 外部システムからデータをインポートしたり、外部システムにデータをエクスポートしたりできるように、データソース接続を設定します。 |
| 自動化と運用化 | ワークフローの自動化を採用して、AnswerSetの作成にかかる反復タスクの数を減らします。 |
| 高度なトピック | ClicktoPrepリンクを作成し、インタラクティブモードを使用し、Data Prepのインフラストラクチャとアプリケーションのセキュリティを理解します。 |

# Data Prepの開始

これらのセクションでは、Data Prepプロジェクトの開始を行うために、知っておくべきことすべて説明します：

| トピック | 説明... |
| --- | --- |
| Data Prepのクイックスタート | データをインポートしてプロジェクトを設定します。次に、データの準備、エクスポート、および共有。 |
| Data Prepの基本の見学 | Data Prepアプリケーションの主要なコンポーネントを見学し、回避する方法を学びます。 |
| Data Prepライブラリ | 新しいデータセットを追加し、既存のデータセットを管理し、AnswerSetを公開します。 |
| Data Prepプロジェクト | プロジェクトでデータを探索して準備します。 |

# データを準備します

DataRobot Data Prepを使用してデータを準備するには、データをインポートして開始します。ローカルデータセットをインポートするか、外部データソースに接続できます。このクイックスタートでは、ローカルデータセットのインポートについて説明します。

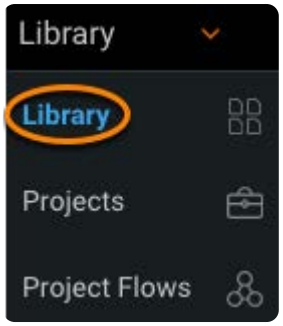
クイックスタートを完了するには、まずDataRobot Data Prepにログインします。ログインした後、次のステップを完了します：

1. 自分のライブラリへのデータの追加。
2. プロジェクトの開始
3. プロジェクトのデータの準備
4. *AnswerSet*としてデータを公開\–データのスナップショット。
5. 準備したデータのエクスポート。

## Data Prepライブラリへのデータの追加

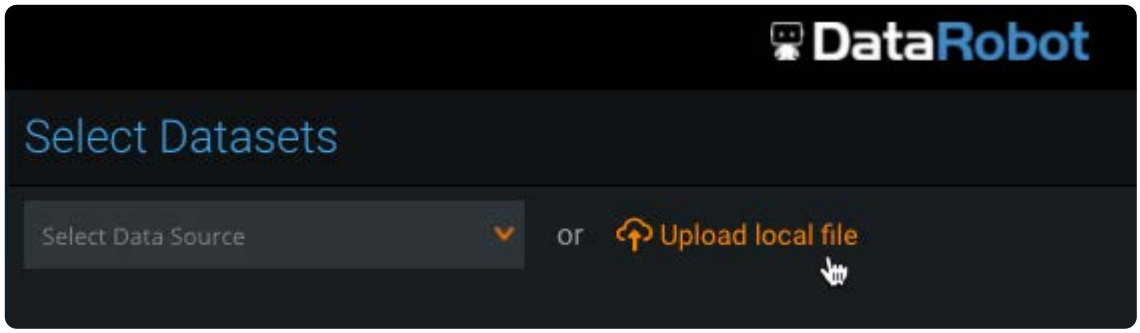
このクイックスタートでは、ローカルデータセットをライブラリにインポートします。データをプロジェクトに直接インポートしたり、外部データソースからデータをインポートしたりすることもできます。インポートに関するその他のオプションの詳細については、データセットの操作を参照してください。

1. DataRobot Data Prepで、左上にある**ライブラリ**を選択します。

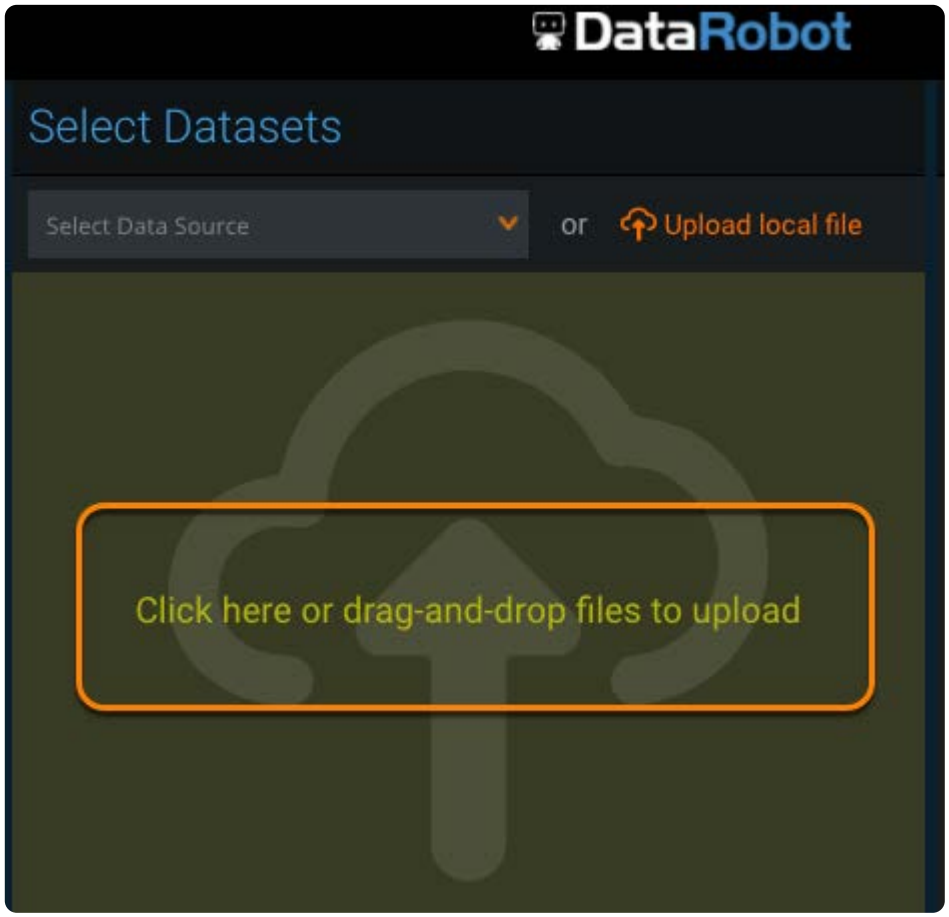



2. **ライブラリ**ページの上で、**+インポート**をクリックします。

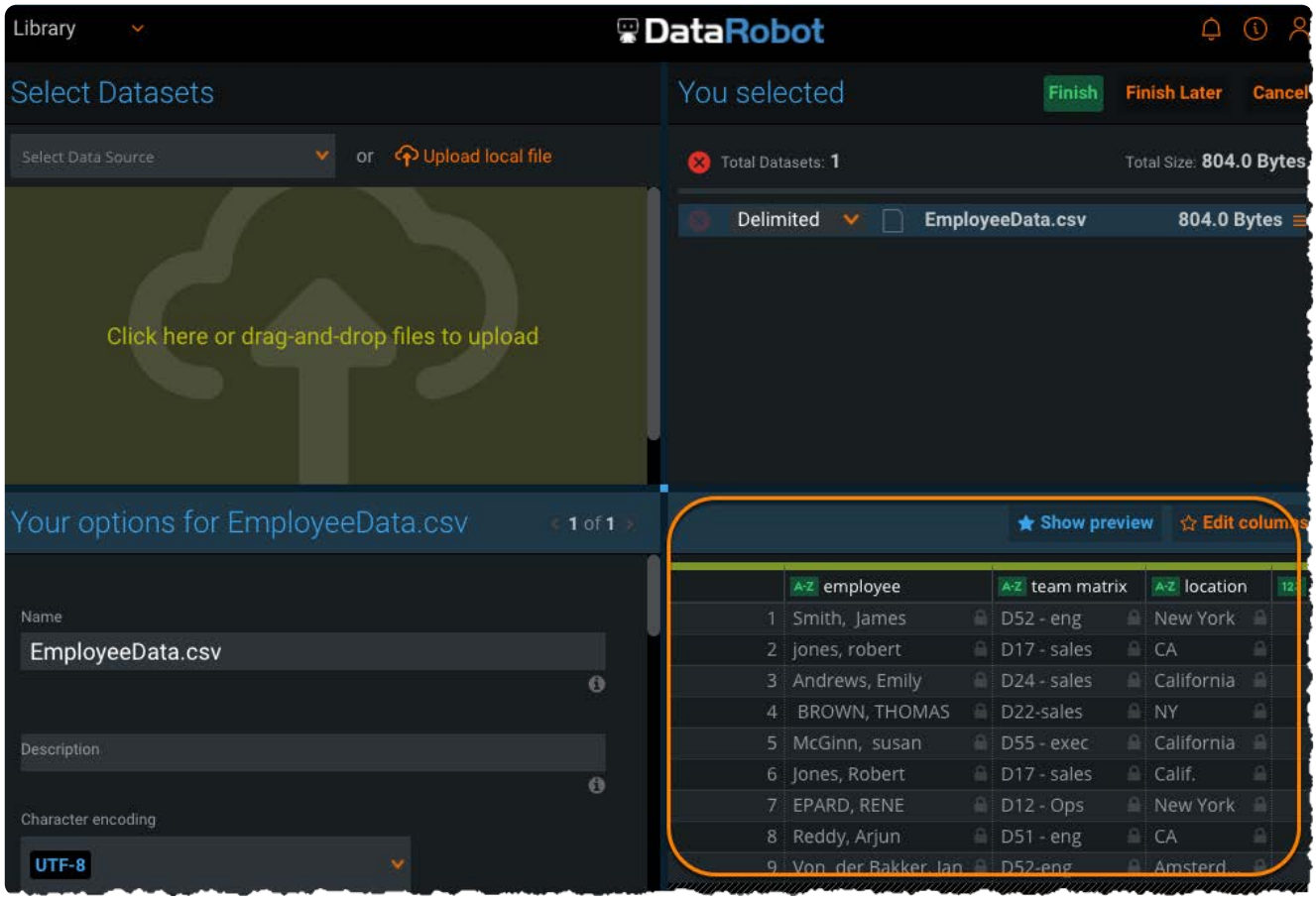

3. **ローカルファイルのアップロード**をクリックします。



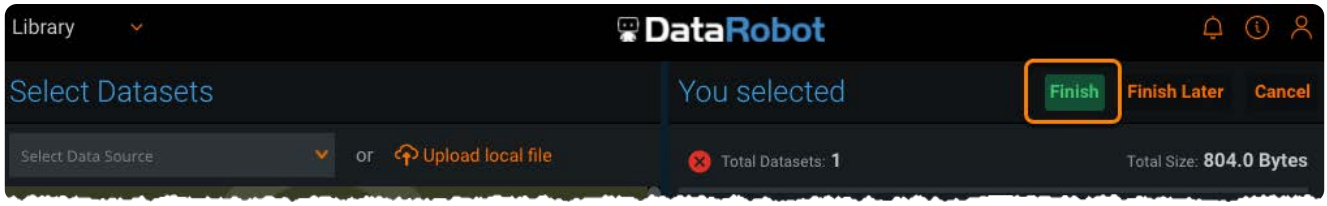

4. ファイルを参照するか、ファイルをドラッグアンドドロップ領域にドラッグします。



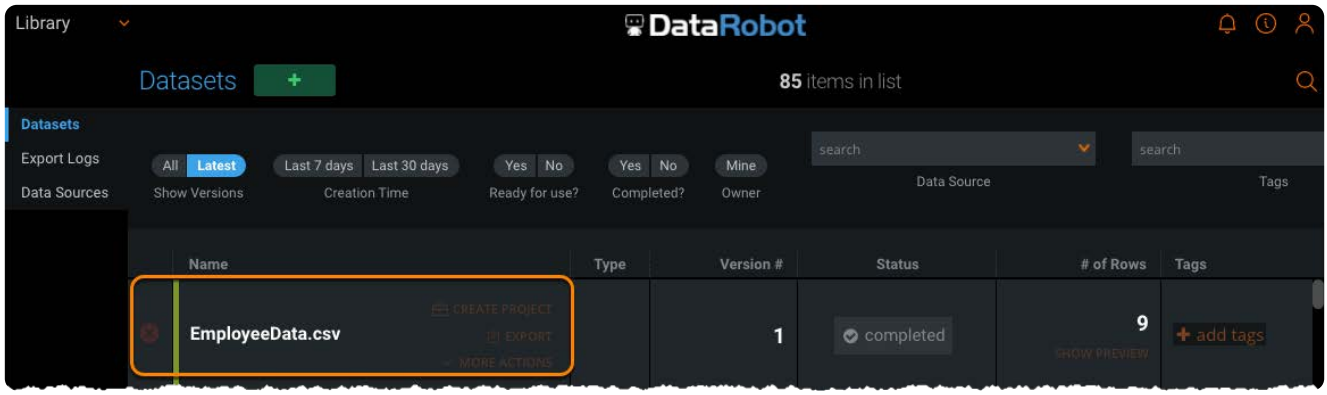

5. 右下のデータセットのプレビューを確認します。




6. データが正しいと思われる場合は、右上にある**終了**をクリックします。




Data Prepはデータセットをライブラリにインポートし、その準備を開始できます。




# Data Prepプロジェクトの開始

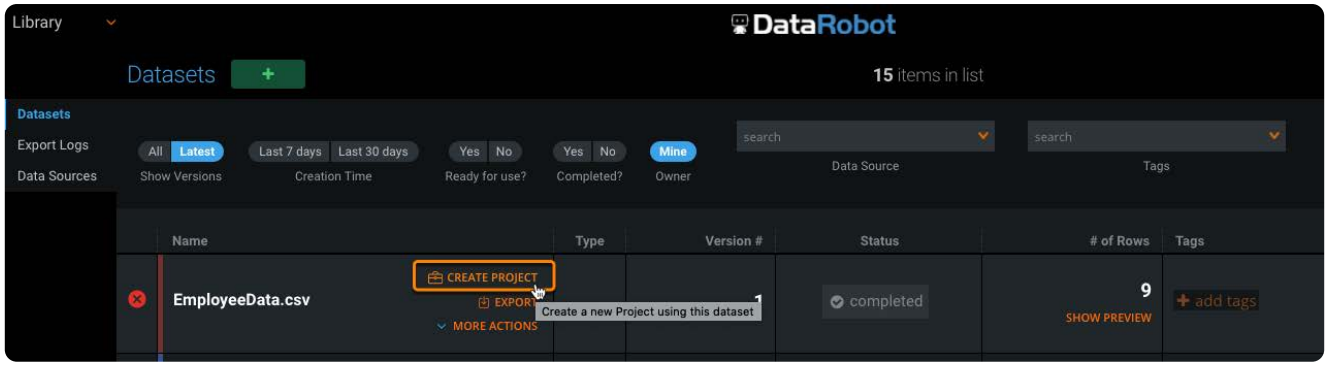
次の場所から新しいプロジェクトを開始できます。

- プロジェクトの開始点として使用するデータセットを選択する**ライブラリ**ページ。

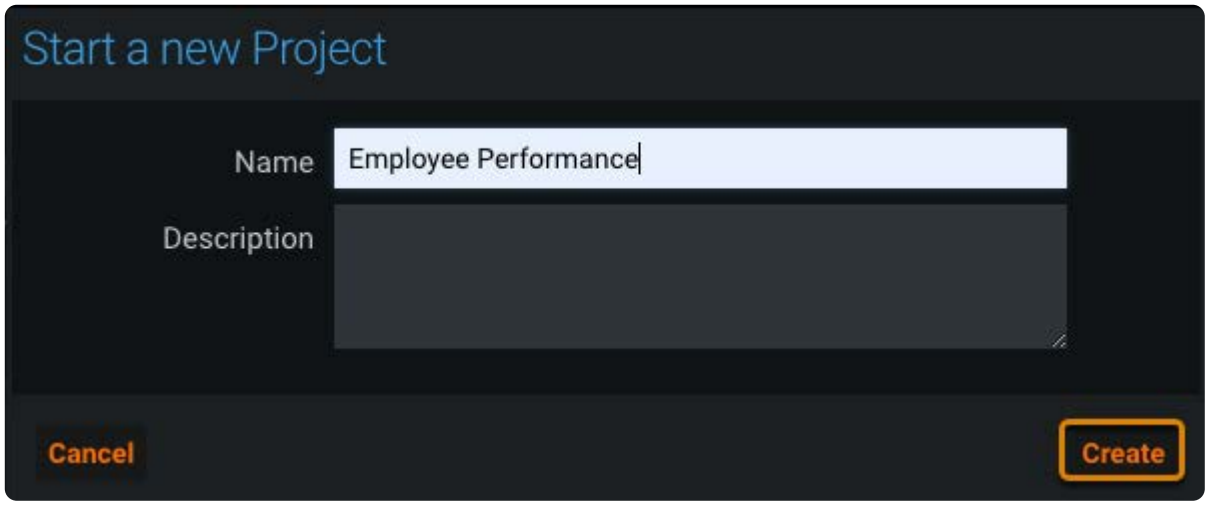
・空のプロジェクトを開始してからデータを追加する**プロジェクト**ページ。

## ライブラリからの新しいプロジェクトの開始

1. 左上にある**ライブラリ**を選択します。
2. アップロードしたデータセットを見つけて、**プロジェクトの作成**をクリックします。




3. **新しいプロジェクトの開始**ダイアログで、プロジェクトの新しい**名前**、およびオプションの**説明**を入力します。

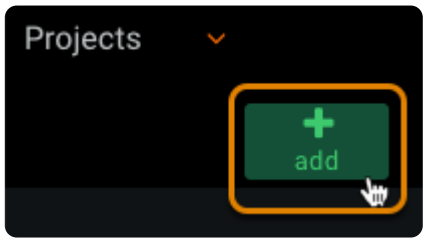



4. **作成**をクリックします。

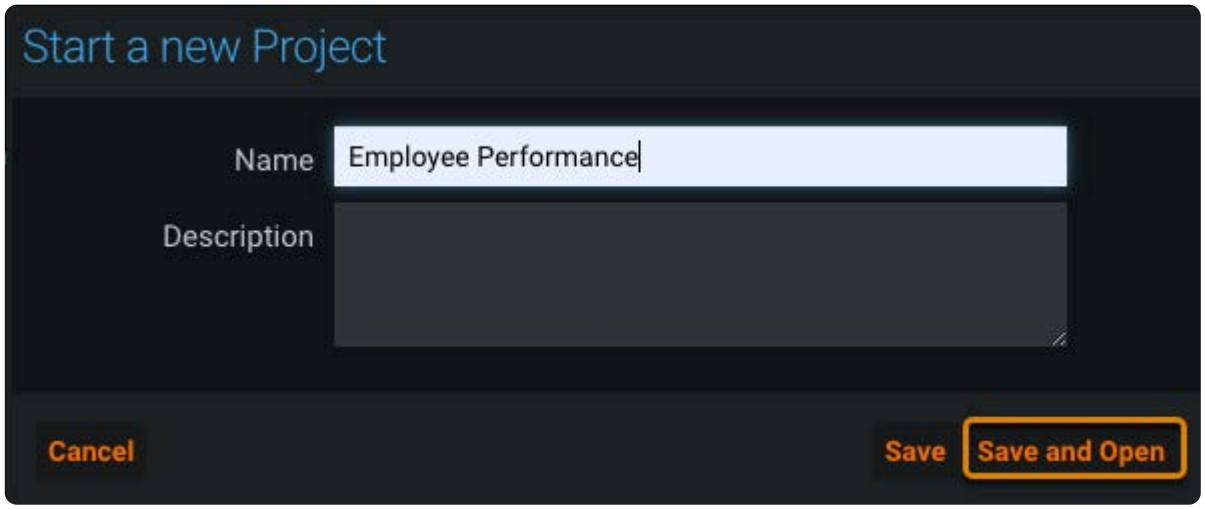
## プロジェクトページからの新しいプロジェクトの開始

ライブラリからプロジェクトを開始する代わりに、**プロジェクト**ページから開始できます。

1. 左上にある**プロジェクト**を選択します。
2. **ライブラリ**ページの上部にある**+追加**をクリックします。




3. プロジェクト**名**およびオプションの**説明**を入力します。

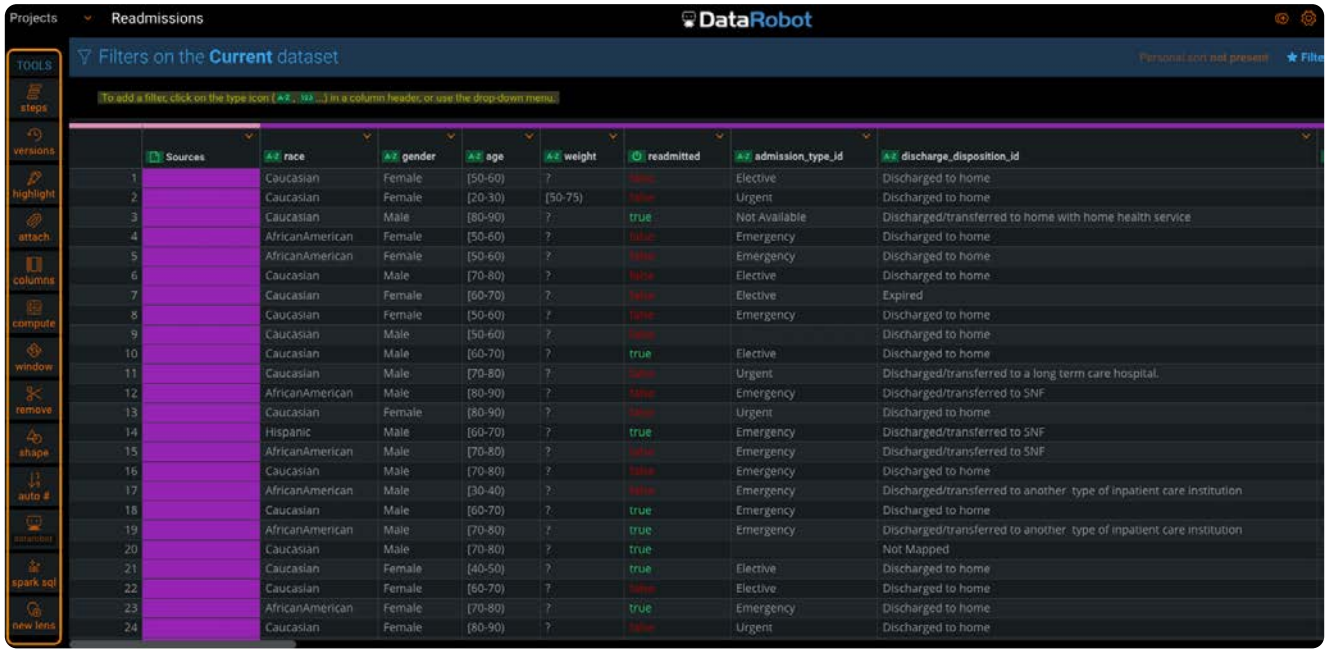



4. **保存して開く**をクリックします。

# データを準備します

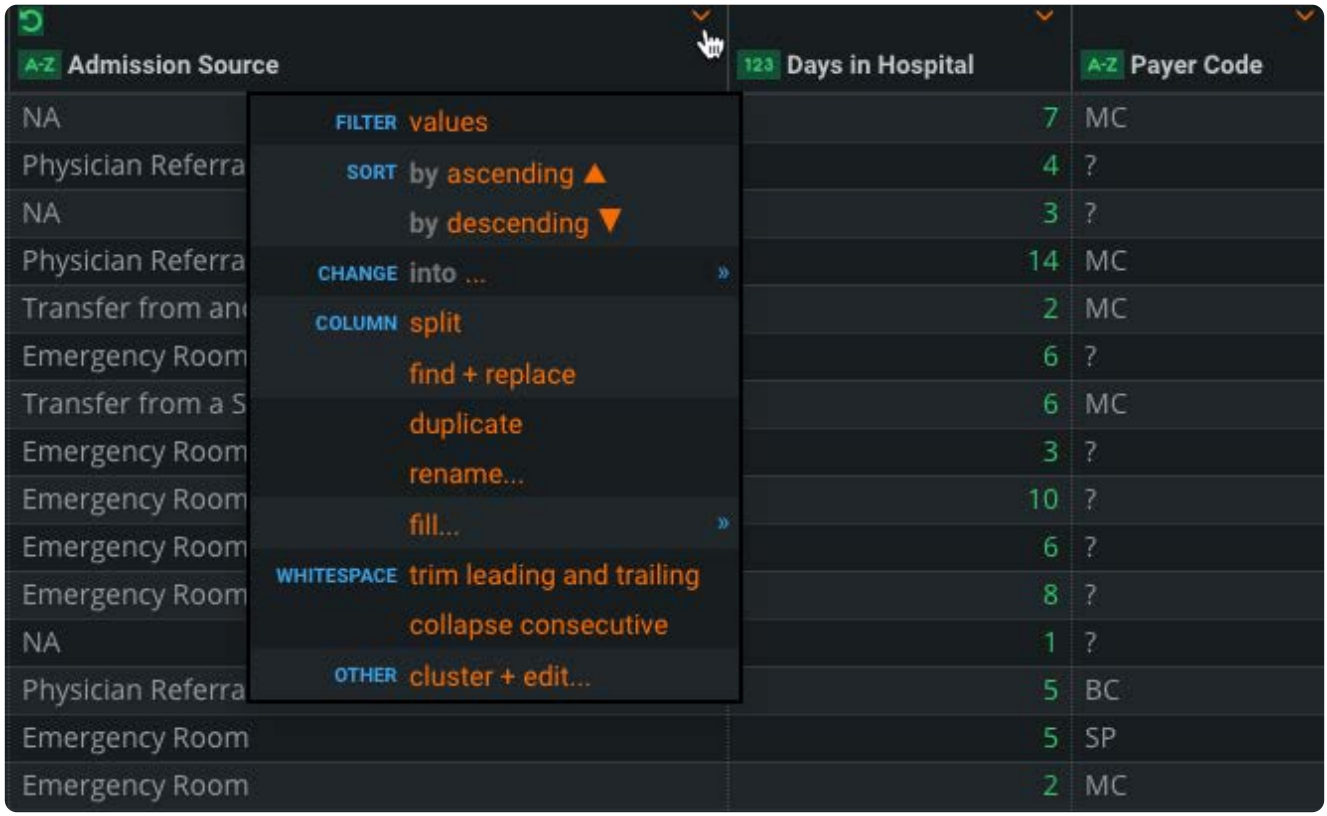
プロジェクトを開始したら、プロジェクトの準備ページでデータの準備を開始できます。

1. 左側のプロジェクト**ツール**バーを使用して、データをクリーンアップおよび変換します。



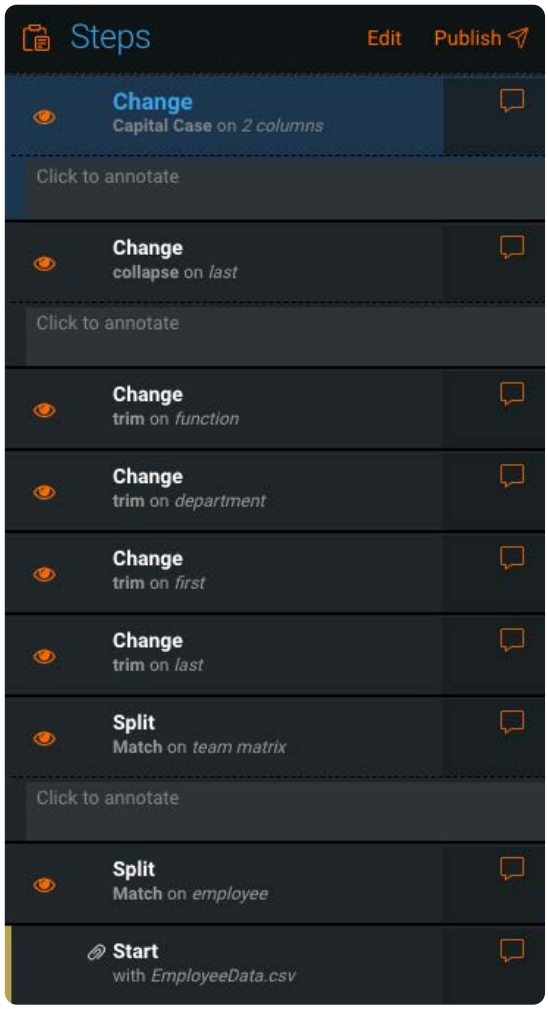

   詳細な手順については、プロジェクトツールの操作を参照してください。

2. 各列の上部にあるメニューを選択して、列操作を適用します。




詳細な手順については、列データの操作を参照してください。

3. Data Prepステップを表示、再配置、およびミュートするには、ステップツールを使用できます。




詳細な手順については、ステップの操作を参照してください。

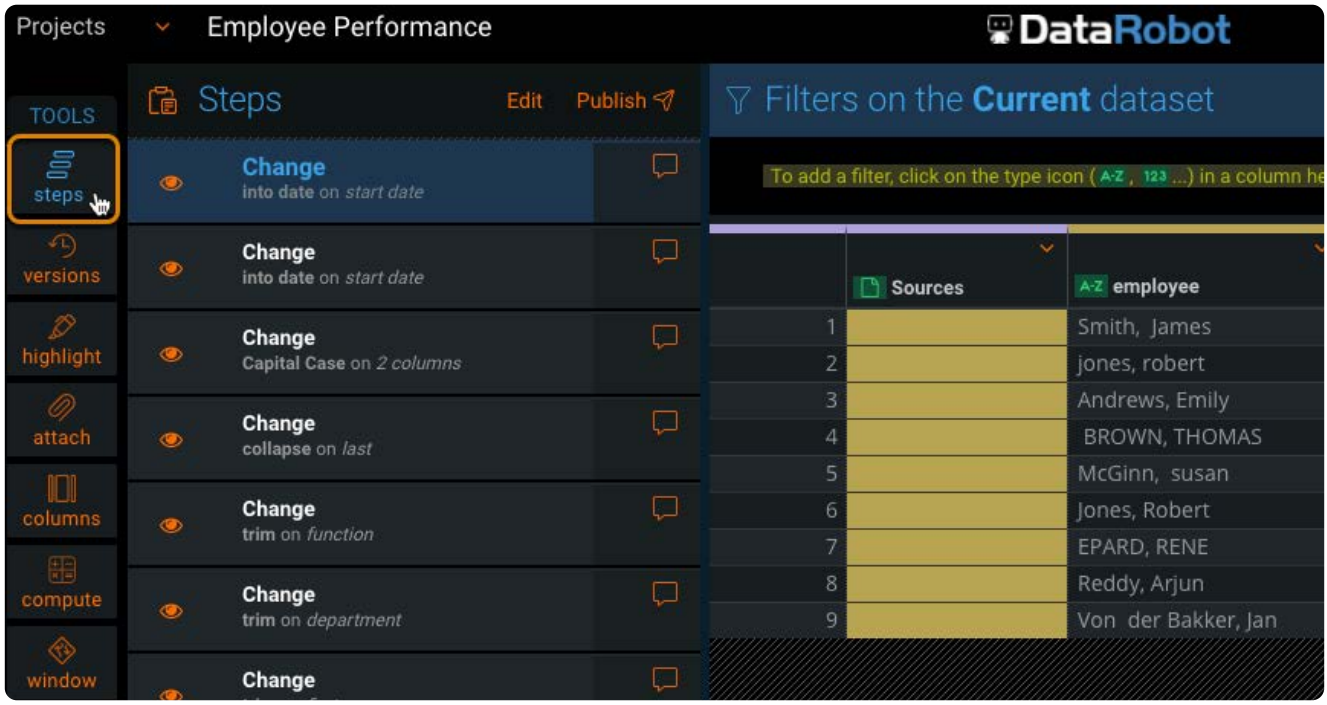
# AnswerSetの公開

準備したデータを保存して共有する準備ができたら、それを*AnswerSet*としてライブラリに公開できます。AnswerSetはデータセットに似ていますが、Data Prepの公開された結果です。いったん公開された AnswerSet は、他のプロジェクトで再利用したり、エクスポートして他のアプリケーションと共有したりできます。

プロジェクトのAnswerSetを公開するには、次のステップに従います：

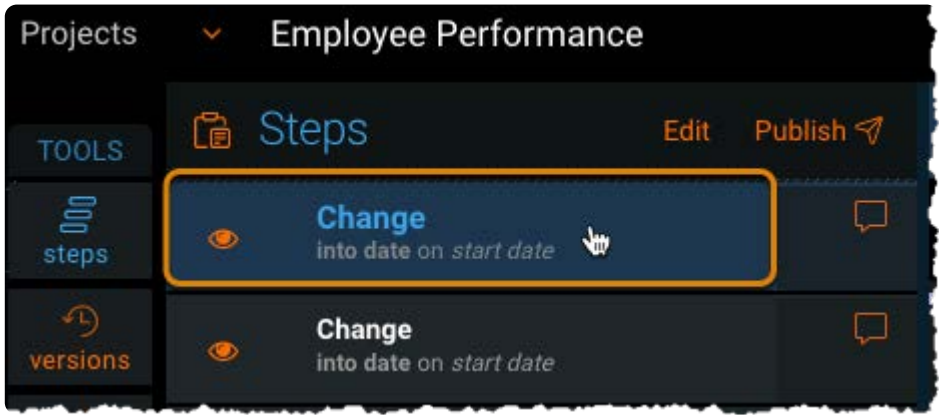
1. **ツール**バーの**ステップ**をクリックします。

   **ステップ**ペインが開きます。

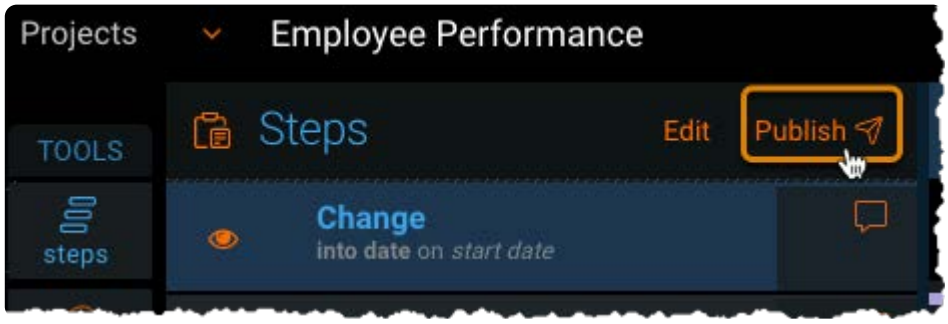



2. AnswerSetを公開するステップをクリックします。

   Data Prepのデフォルトは、プロジェクトの最後のステップ、つまり一番上のステップです。

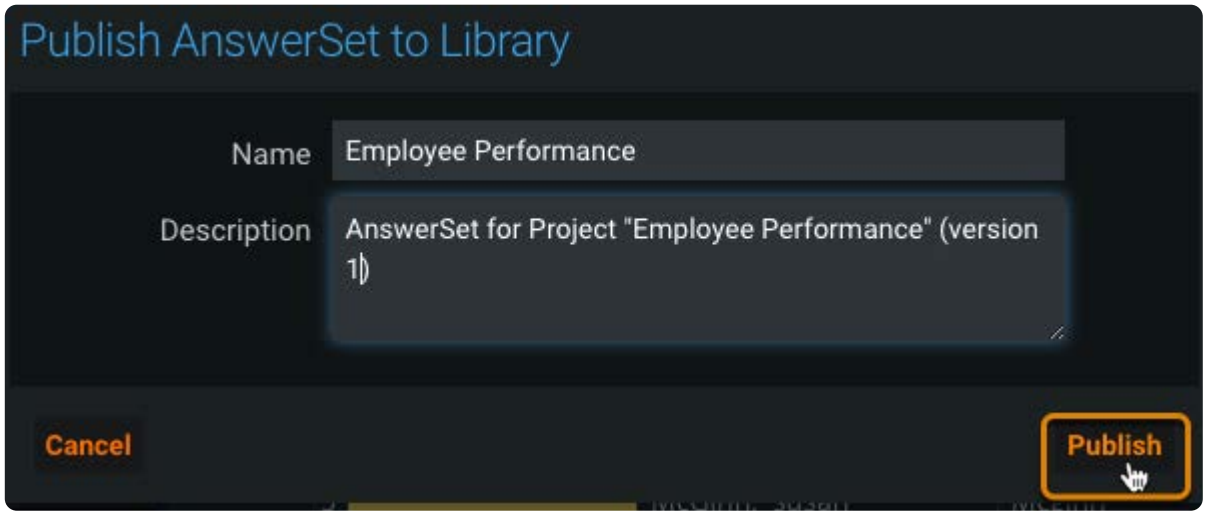



3. **ステップ**ペインの上部で、**公開する**をクリックします。




   **AnswerSetをライブラリに公開**ウィンドウが表示されます。

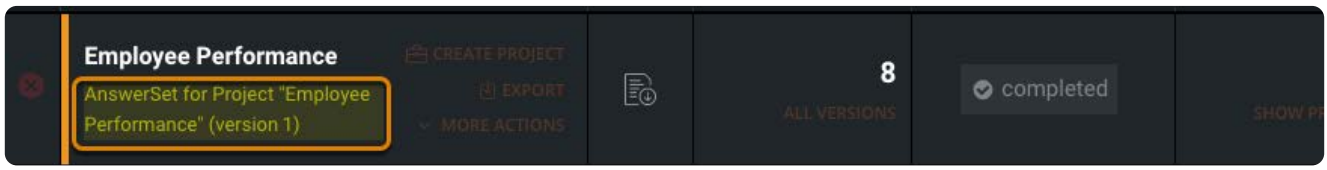
4. AnswerSetの名前を**名前**フィールドに入力し、**説明**オプションを選択して**公開する**をクリックします。




Data Prepは、AnswerSetをライブラリに公開します。「AnswerSetを公開中」というメッセージが表示されます。

5. **ライブラリに表示**をクリックし、ライブラリ内のAnswerSetを表示します。

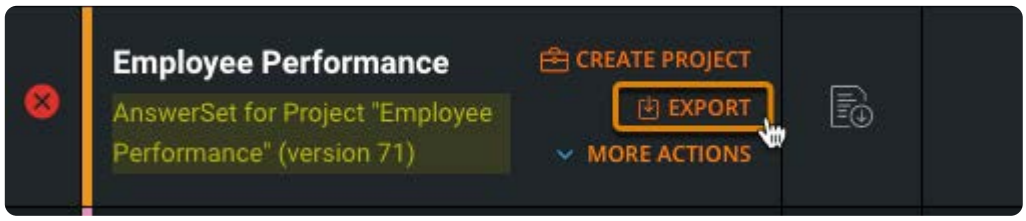

AnswerSetには、選択したステップまでのステップが含まれています。



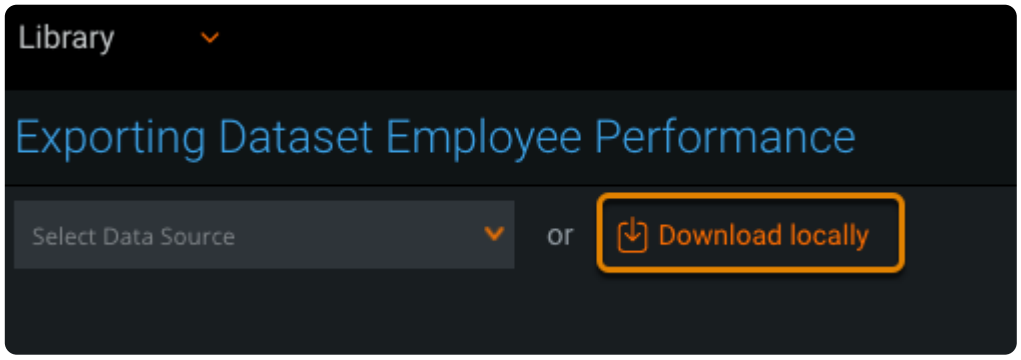

# 準備されたデータをエクスポートする

データセットとAnswerSetをローカルまたは接続されたデータソースにエクスポートできます。これらのステップは、以前に公開されたAnswerSetのローカルコピーをダウンロードする方法を示しています。

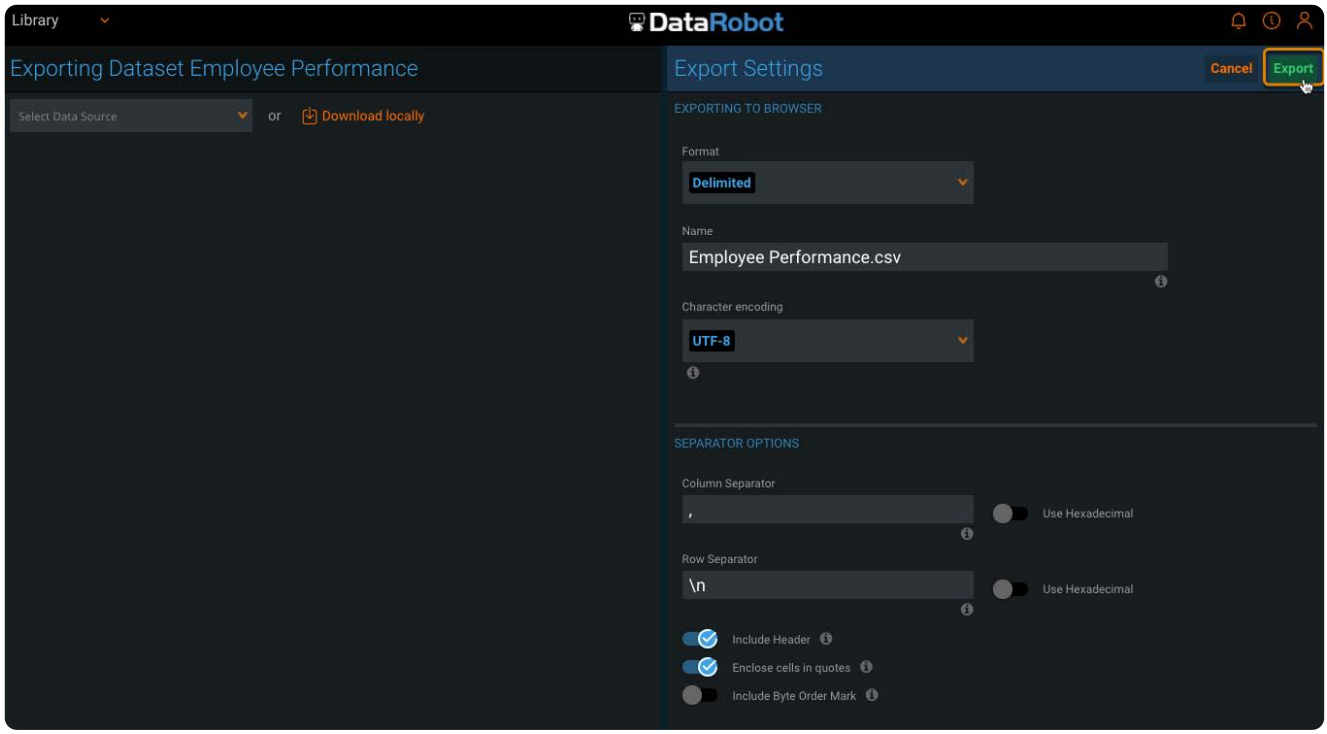
1. **ライブラリ**ページで、エクスポートするAnswerSetにカーソルを合わせて、**エクスポート**をクリックします。



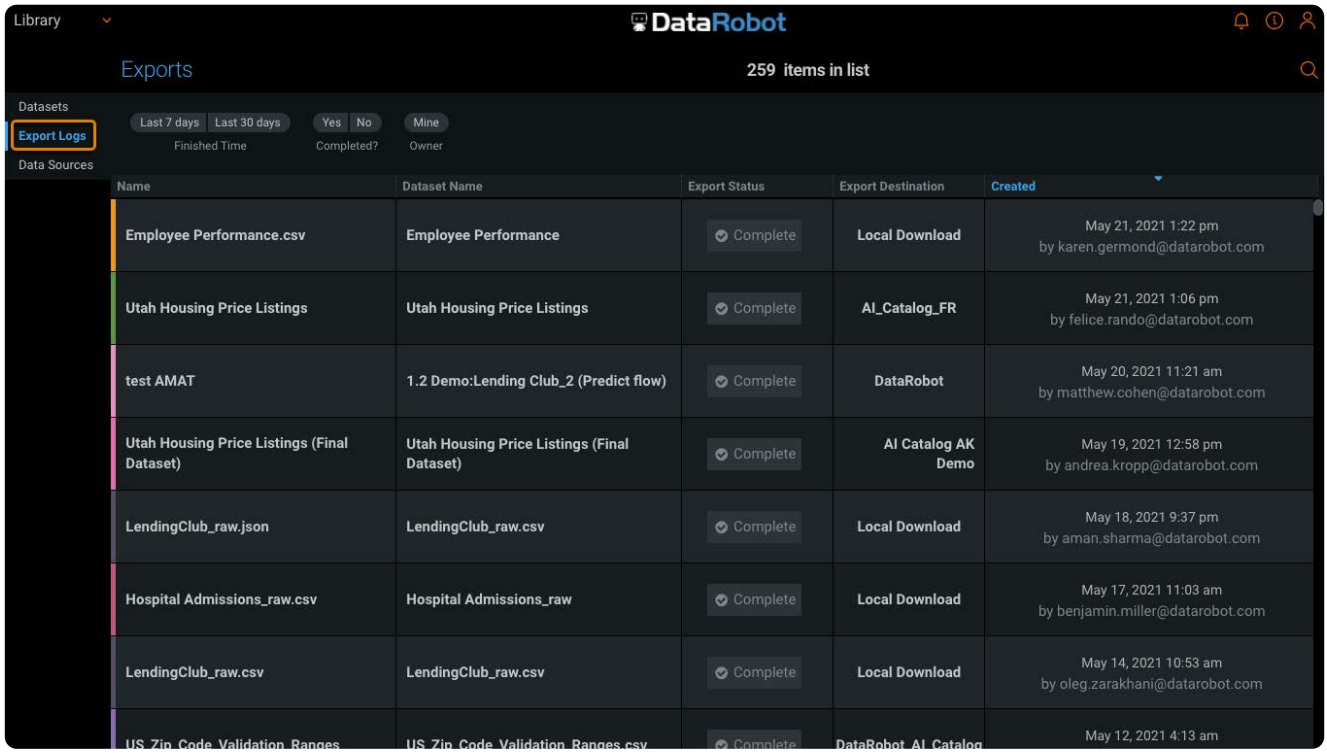

2. **エクスポート**ページで、**ローカルにダウンロード**をクリックします。




3. **エクスポート設定**ページで、**エクスポート**をクリックします。




AnswerSetは、CSVファイルとしてコンピュータにダウンロードされます。**ログのエクスポート**ページが表示されます。




詳細については、データセットのエクスポートを参照してください。

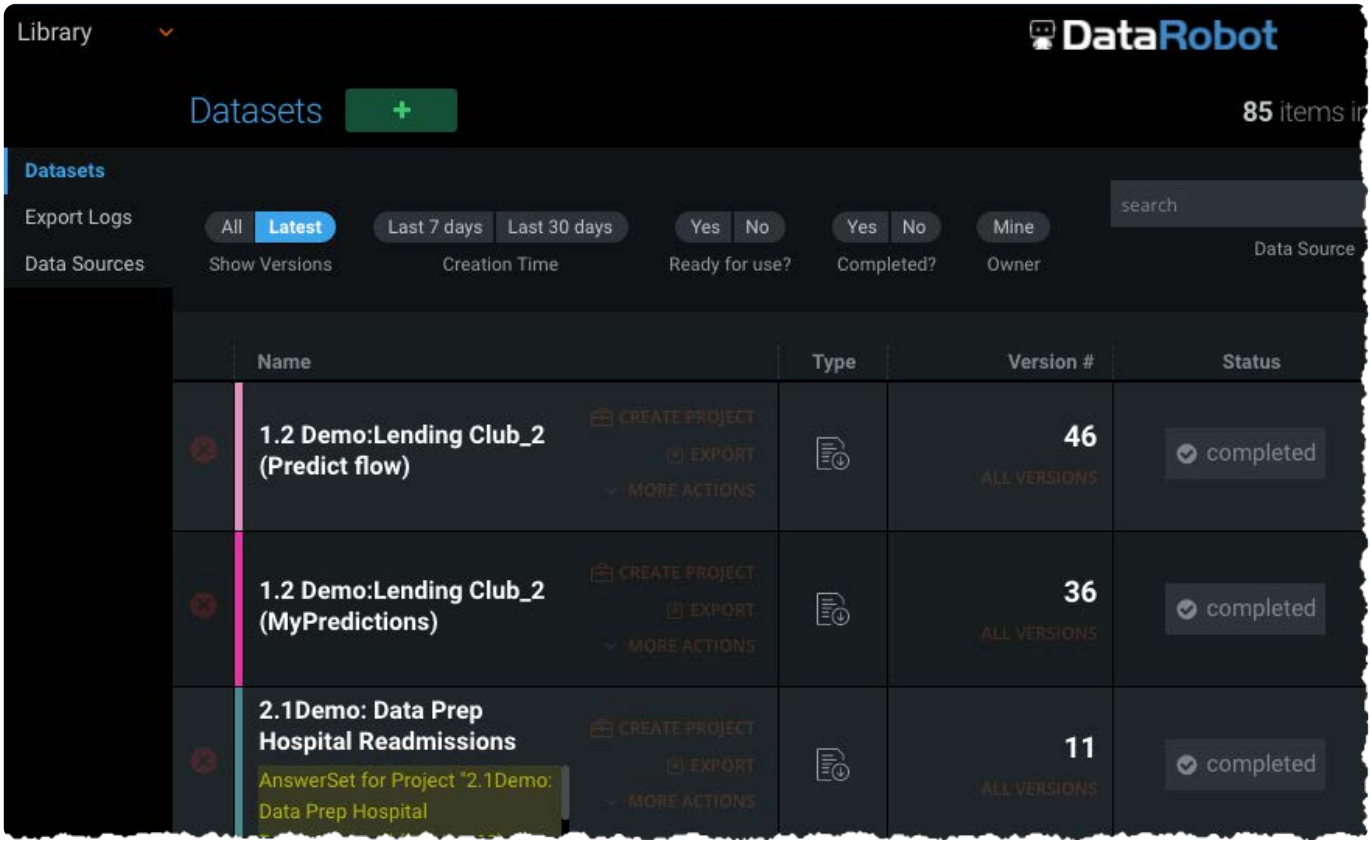
# Data Prepのベーシック機能ツアー

このトピックでは、Data Prepアプリケーションの主要コンポーネントを見ていきます。

## ライブラリ

Data Prepライブラリでは以下を実行します。

- データセットを追加および管理する。
- 準備されたデータセット、AnswerSetsを公開する。
- 自動化用のデータセットを設定する。
- 新しいバージョンを追加する。
- データセットのプロファイルを作成する。
- データセットのインポート時に発生する警告またはエラーを表示する。

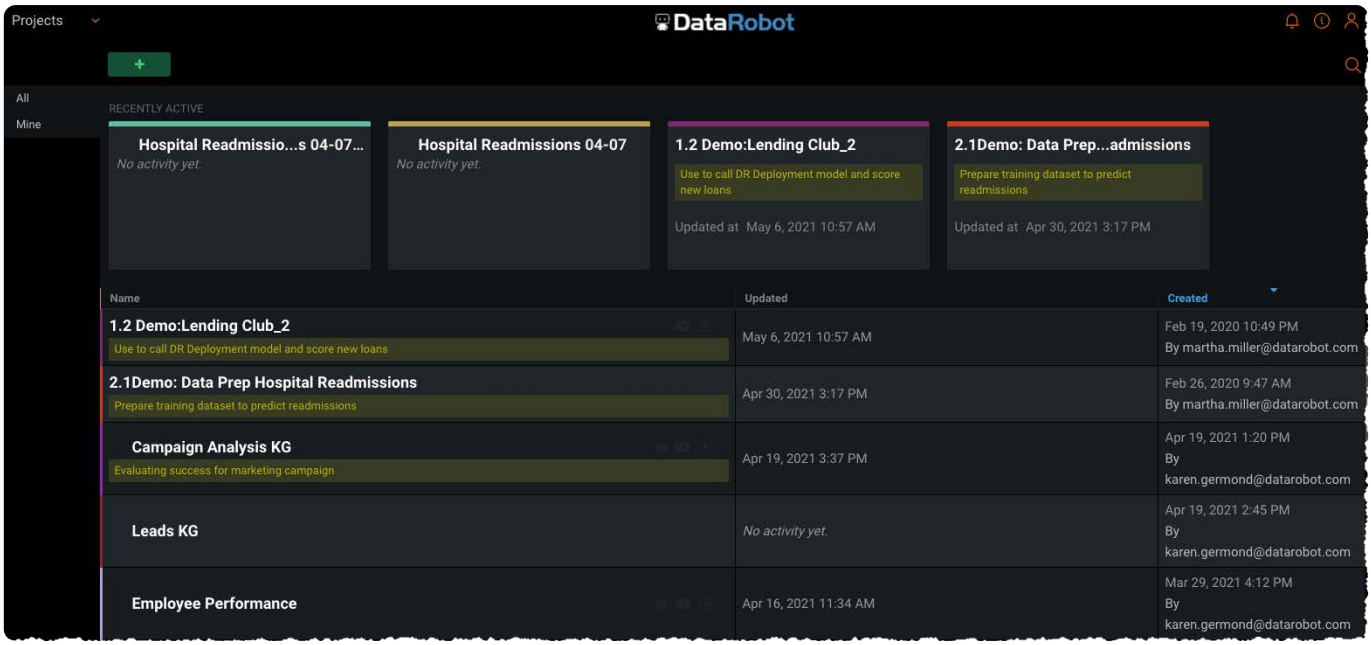



データセットをライブラリにインポートした後に、*プロジェクト*内のデータの準備を開始できます。データの準備が完了したら、*AnswerSet*（公開されたデータセット）としてライブラリに公開できます。

詳細については、Data Prepライブラリを参照してください。

# プロジェクト

**プロジェクト**ページには、ユーザーが表示する権限を持つすべてのプロジェクトが一覧表示されます。



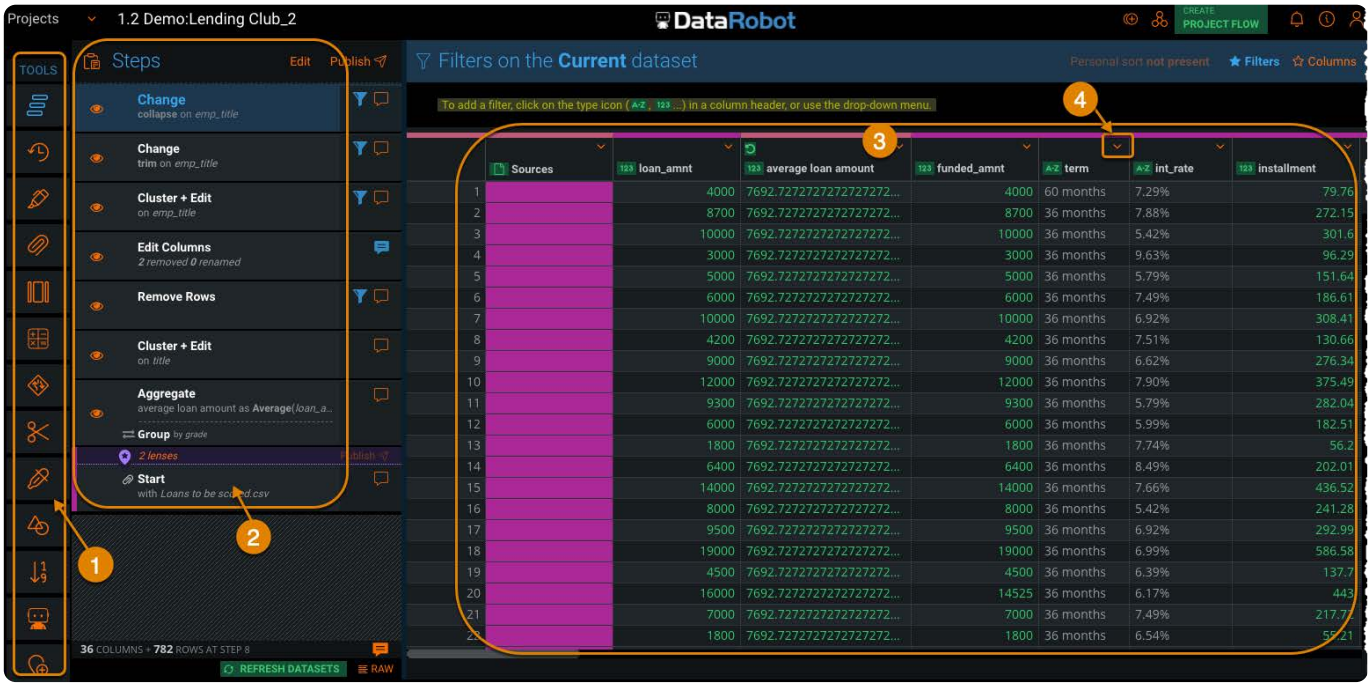

データセットをプロジェクトに追加すると、データセットを探索して、クリーンアップ、変換、または他のデータと組み合わせることができます。

ライブラリへの変更をAnswerSetとして公開し、別のプロジェクト内でエクスポートまたは使用することができます。

詳細については、Data Prepプロジェクトページを参照してください。

# プロジェクトの準備

プロジェクトを開くには、**プロジェクト**ページでプロジェクトをクリックするか、ライブラリから新しいプロジェクトを開始します。プロジェクトを開いたら、データの準備を開始できます。




| | 要素 | 説明 |
|---|---|---|
| 1 | **ツール**バー | データの準備に使用するプロジェクト**ツール**バーには、画面左側からアクセスできます。 |
| 2 | **ステップ**ツール | **ステップ**ツールには実行した各操作が保存され、ステップを再生、ミュート、および再配置できるようになります。 |
| 3 | **プレビューの表示**ペイン | データは**データプレビュー**ペインに表示されます。 |
| 4 | **列操作メニュー** | 各列の上から、列操作にアクセスして列を更新できます。 |

詳細については、Data Prepプロジェクトの準備ページを参照してください。

# データ

コンピューター上のローカルファイルまたは接続されているデータソースからデータをインポートできます。Data Prepシステム管理者は、データソースからインポートする前に、データソースを設定する必要があります。接続できるデータ ソースの例として、次のようなものがあります。

- Amazon S3 などのクラウドストレージ
- Hadoop Distributed File System（HDFS）
- MySQL などのリレーショナルデータベース
- Secure File Transfer Protocol（SFTP）

# Data Prepナビゲーション

Data Prepヘッダーは、ナビゲーション、ヘルプ、およびアカウント管理機能を提供します。

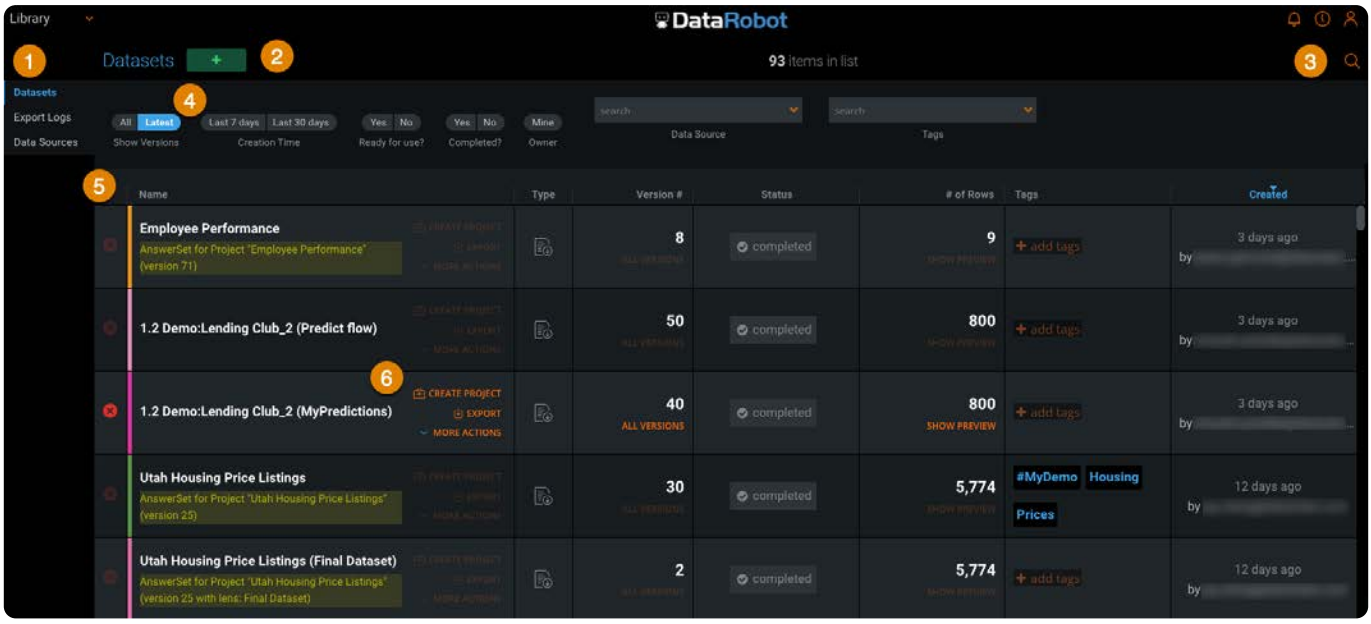

| | 要素 | 説明 |
|---|---|---|
| 1 | ナビゲーションメニュー | Data Prepページ間の移動：<br>・**ライブラリ：**インポートされた公開されたデータにアクセスします。<br>・**プロジェクト：**データの準備。<br>・**管理:** データソースに接続し、ユーザーの権限を制御します。<br>・**プロジェクトフロー：**データ準備プロセスを自動化します。<br><br>**備考：**各ユーザーが利用できるページは、ユーザーの権限に基づいています。 |
| 2 | 通知アイコン | Data Prepが警告またはエラーを生成するときに表示されます。強調表示されている場合、アイコンの上にカーソルを合わせるとメッセージが表示されます。 |
| 3 | ヘルプ | Data Prepのヘルプを取得します。 |
| 4 | ユーザーメニュー | パスワードの更新やログアウトなど、アカウント固有のオプションにアクセスします。アプリケーションのアクセスと承認を管理するために使用されるトークンを生成することもできます。Data Prepシステム管理者は、トークンを生成する必要があるときに通知します。 |
| 5 | 検索 | ページに固有の検索です。たとえば、**ライブラリ**ページではデータセットを検索でき、**プロジェクト**ページではプロジェクトを検索できます。 |

# Data Prepライブラリ

**ライブラリ**ページでは、新しいデータセットを追加したり、プロジェクトから公開するData Prep AnswerSetsを含む既存のデータセットを管理したりできます。ライブラリでは、データセットのエクスポート、自動化の設定、新しいバージョンの追加、データセットのプロファイルの作成、データセットのインポート時に発生した警告またはエラーの表示を行うこともできます。

次の表では、ライブラリレイアウトと、データセットを操作するためにライブラリで実行するアクションについて説明します。

## ライブラリのレイアウト



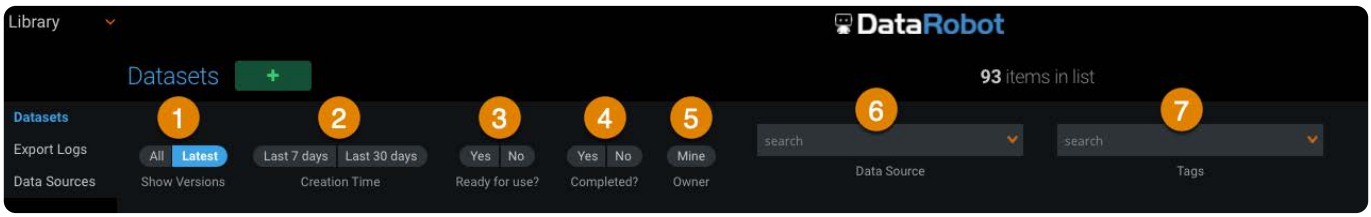

次の表は、**ライブラリ**ページのセクションについて説明しています。

| | アクション | 説明 |
|---|---|---|
| 1 | **ライブラリタブ** | 表の選択：<br>・**データセット：**データセットの管理。<br>・**ログのエクスポート：**エクスポート中に生成されたログを表示します。<br>・**データソース：**新しいデータソースを追加します。ライブラリに追加されたデータソースからのインポートおよびエクスポート |

| | アクション | 説明 |
|---|---|---|
| 2 | **新しいデータセットの追加** | ライブラリに新しいデータセットを追加するには、左上の**Datasets +**をクリックします。**データセットの選択**ページで、データソースまたはローカルデータセットを選択して、1つ以上のデータセットをインポートします。インポート中にエラーが発生した場合、ページ上のデータセットのリストの横に赤い注意アイコンが表示されます。エラーの詳細については、データセットの名前の上にカーソルを合わせ**詳細を編集する**をクリックします。 |
| 3 | **ページでデータセットを検索する** | ライブラリ内のデータセットを検索するには、右上の虫眼鏡アイコンをクリックします。表示されるフィールドに、検索するデータセットの名前の入力を開始します。入力を続けると、一致する可能性のある名前が表示されます。 |
| 4 | **ページに表示されるデータセットをフィルターする** | バージョン、作成時間、所有者などのカテゴリ別にデータセットのリストをフィルターします。 |
| 5 | **列を並べ替える** | **ライブラリ**ページには、フィルターされたデータセットが一覧表示されます。ライブラリ列は、タイプ、バージョン、ステータス、行数、タグ、作成されたデータとその作成者など、各データセットの属性を示します。 |
| 6 | **データセットアクション** | プロジェクトの作成やデータセットのエクスポートなど、データセットにアクションを実行できます。データセットを削除するには、データセットの左側にある赤いXアイコンをクリックします。 |

# ライブラリーフィルター

ページの上部にあるフィルターを使用して、ページに表示されるデータセットのリストをフィルター処理します。




次の表はデータセットのフィルター処理のオプションを示します。

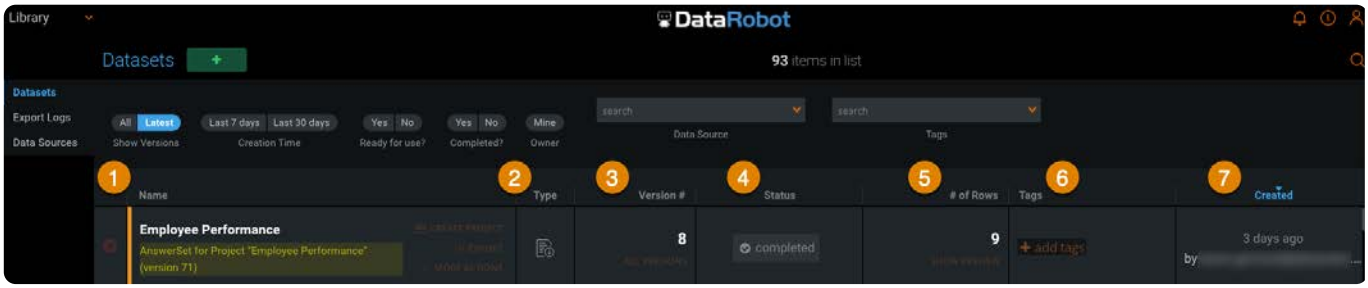
| | アクション | 説明 |
|---|---|---|
| 1 | **バージョンの表示** | すべてのデータセットとAnswerSetのすべてのバージョンを表示するか、それぞれの最新バージョンのみを表示するかを切り替えます。 |
| 2 | **作成時間** | 過去7日間または30日間を選択します。 |
| 3 | **使用する準備ができていますか？** | このフィルターは、Data Prepプロジェクトでインタラクティブモードが有効になっている場合にのみ表示されます。これを使用すると、インタラクティブ部分の読み込みが完了していてプロジェクトで使用できるデータセットがすぐにわかります。 |
| 4 | **完了しましたか？** | ライブラリへのインポートが正常に終了したすべてのデータセットを表示します。 |
| 5 | **オーナー** | Data Prepプロジェクトからインポートして作成したデータセットとAnswerSetを表示します。 |
| 6 | **データソース** | このフィールドをクリックすると、データセットのインポートに使用されるすべてのデータソースが表示されます。連続してクリックして選択することで、複数のデータ ソースを選択できます。 |
| 7 | **タグ** | タグは、データセットを整理するための説明的な言葉です。**タグ**フィールドをクリックして、ライブラリ内のデータセットに現在割り当てられているすべてのタグを表示します。特定のタグでデータセットを見つけるには、タグ名を入力して**Enter**をクリックします。複数のタグを追加した場合、検索タグの すべてを含むデータセットだけが一致したものとして返されます。データセットに新しいタグを追加するには、データセットにカーソルを合わせ、そのデータセットの**タグの追加**フィールド内をクリックします。 |

# ライブラリの列

列のヘッダーをクリックして新しいロケーションにドラッグすると、**ライブラリ**ページの新しい列の表示順序を作成できます。ライブラリ リストを特定の列で並べ替えるには、その列名をクリックします。複数の列で並べ替えるには、**Shift**を押したまま追加の列をクリックします。

**備考**

**ライブラリ**ページでアイテムを並べ替える場合、変更は一時的なものであり、**ライブラリ**ページを離れたり、ブラウザーの表示を更新したりすると保持されません。




次の表は、**ライブラリ**ページの列について説明するものです。

| | アクション | 説明 |
|---|---|---|
| 1 | **名前** | データセットがライブラリにインポートされたときのデータセットの名前を表示します。名前を変更するには、データセットの**追加アクション**にカーソルを合わせ、**詳細を編集する**をクリックします。データセット表示の**一般**ページ。データセット名を変更し、メタデータフィールドを更新できます。 |
| 2 | **タイプ** | ライブラリ内のどのデータセットがData Prepプロジェクトから作成されたAnswerSetであるかをすばやく識別できます。Data Prepプロジェクトでインタラクティブモード機能が有効になっている場合、AnswerSetは部分的なアイコンで表され、作成時にインタラクティブモードで作業していたことを示します。 |
| 3 | **バージョン#** | 各データセットまたはAnswerSetのバージョン数を表示します。複数のバージョンがある場合、データセットの**すべてのバージョン**をクリックするとすべてのバージョンが表示されます。すべてのデータセットの表示に戻るには、左上の**すべてのデータセット**をクリックします。<br><br>**すべてのバージョン**ページを調べているときにタグをフィルター処理すると、検索は**ライブラリ**ページの全部ではなく、**すべてのバージョン**ページにのみ適用されることに注意してください。<br><br>バージョン番号は、ライブラリ内のそれらのデータセットの実際の番号に必ずしも対応しているわけではありません。<br><br>バージョン番号がライブラリ内のそれらのデータセットの正確な番号と一致しない条件：<br>・インポートが完了する前にキャンセルされると、バージョン番号が自動的に生成され、その後のインポートは単にバージョン番号に追加される増分になります。<br>・データセットの特定のバージョンが削除されても、残りのデータセットのバージョン番号が減ることはありません。 |

| | アクション | 説明 |
|---|---|---|
| 4 | **ステータス** | データセットがライブラリにインポートされているときのデータセットの読み込みステータスを表します。ほとんどの場合、ステータスはすぐに「完了」に進みます。ただし、より大きなデータセットの場合、データセットが正常にインポートされ続けていることを示す暫定的な状態が表示されます。表示される中間状態は、以下の条件によっても左右されます。<br>・**データセットの行数をインポート前にあらかじめ決定できるかどうか**：ほとんどの場合、Data Prepは、インポートプロセスが開始される前にデータセット内の行数を認識しています。しかし、数を事前定義できない場合（Salesforceからインポート、およびJDBCデータソースに対するクエリなど）があります。<br>・**プロジェクトでインタラクティブモードが有効になっているかどうか**：インタラクティブ モードが有効になっている場合、ステータス アイコンは 2 つの同心円で表されます。内側の円は、データセットのインタラクティブ部分を表します。インタラクティブ部分がプロジェクトで使用可能になると、内側の円が緑のチェックマークになります。データセットの残りの部分が引き続きライブラリに読み込まれるにつれて、外側の円が緑で塗りつぶされていきます。インタラクティブな部分または残りの部分のインポート中にエラーが発生した場合、それぞれの同心円に赤い注意アイコンが表示され、データセットのどの部分がライブラリへのインポートに失敗したかを示します。ロード状態の例についてはロード状態を参照してください。障害状態の例については障害状態を参照してください。<br><br>データセットの解析オプションの選択が完了していない場合、この列に「保留」状態が表示される場合があります。この場合、**作成された**列に**クリックして終了**ボタンが表示されます。ボタンをクリックして**インポート**ページを開き、インポートを終了します。 |
| 5 | **行数** | データセット内の行数を表示します。データセットの上にカーソルを移動し、この列に表示される**プレビュー表示**リンクをクリックすることにより、データセットから行をプレビューすることができます。データセットが現在インポートプロセス中であり、行数が事前に決定されている場合、この列に表示される数は、インポートが完了するまで増加し続けます。データセットのインポートに失敗した場合、*正常にインポートされた*行の数はこの列にリスト表示されます。この場合、**プレビュー表示**はそれらの行のプレビューを表示します。 |
| 6 | **タグ** | タグは、データを整理するためにデータセットに追加できるラベルです。データセットにタグを追加するには、そのデータセットのタグ列をクリックし、タグ名を入力し、表示される**追加**リンクをクリックするか、**Enter**キーを押します。 |
| 7 | **作成完了** | データセットをインポートしたユーザーとインポートされた日時を表示します。列内に**クリックして終了**リンクが表示される場合があります。これは、解析オプションの設定が完了していないためにインポートが開始されなかったことを示します。このリンクをクリックして**インポート**ページに戻り、データセットのインポートプロセスを完了してください。 |

## 読み込み状態

| アイコン | 説明 |
| --- | --- |
|   | インタラクティブモードが有効になっておらず、行数を決定できる時に表示されるアイコン |
|  | インタラクティブモードが有効になっておらず、行数を事前に決定できない時に表示されるアイコン |
|  | インタラクティブモードが有効になっており、行数を事前に決定できない時に表示されるアイコン |
|  | インタラクティブモードが有効になっており、行数を事前に決定できない時に表示されるアイコン |
|   | 読み込みが完了すると表示されるアイコン。 |

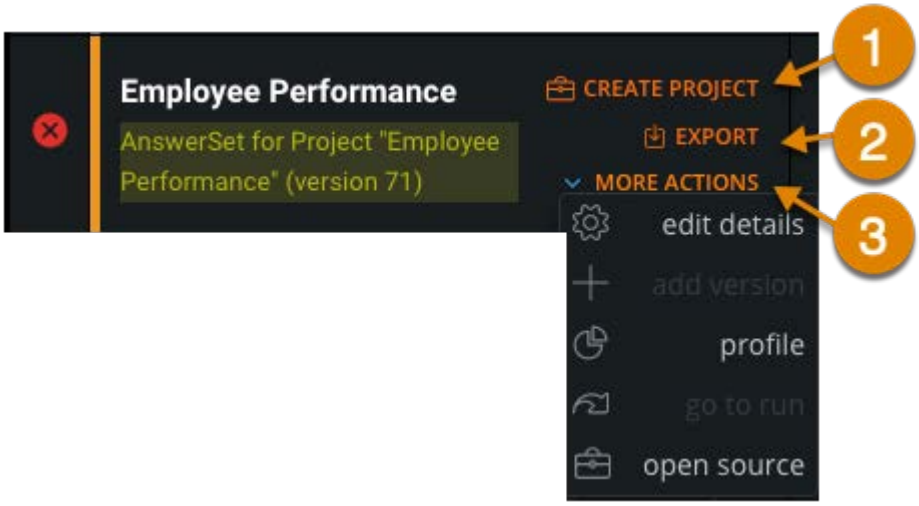
## 障害状態

| アイコン | 説明 |
| --- | --- |
|  | インタラクティブモードが有効になって**いません**：データセットのインポートに失敗しました。 |
|  | インタラクティブモード：インタラクティブ部分が正常にインポートされませんでした。 |

| アイコン | 説明 |
| --- | --- |
|  | インタラクティブモード：インタラクティブ部分は正常に完了しましたが、データセットの残りは正常にインポートできませんでした。 |

# データセットに対して実行できるアクション

データセットにカーソルを合わせると表示される3つのリンクにより、そのデータセットに対して実行できるオプションが提供されます。




次の表は、**ライブラリ**ページのデータセットに対して実行できるアクションを示します。

| | アクション | 説明 |
| --- | --- | --- |
| 1 | **プロジェクトの作成** | データセットをベースデータセットとして使用して、新しいプロジェクトを作成します。 |
| 2 | **エクスポート** | データセットをローカルでエクスポートまたはダウンロードします。 |

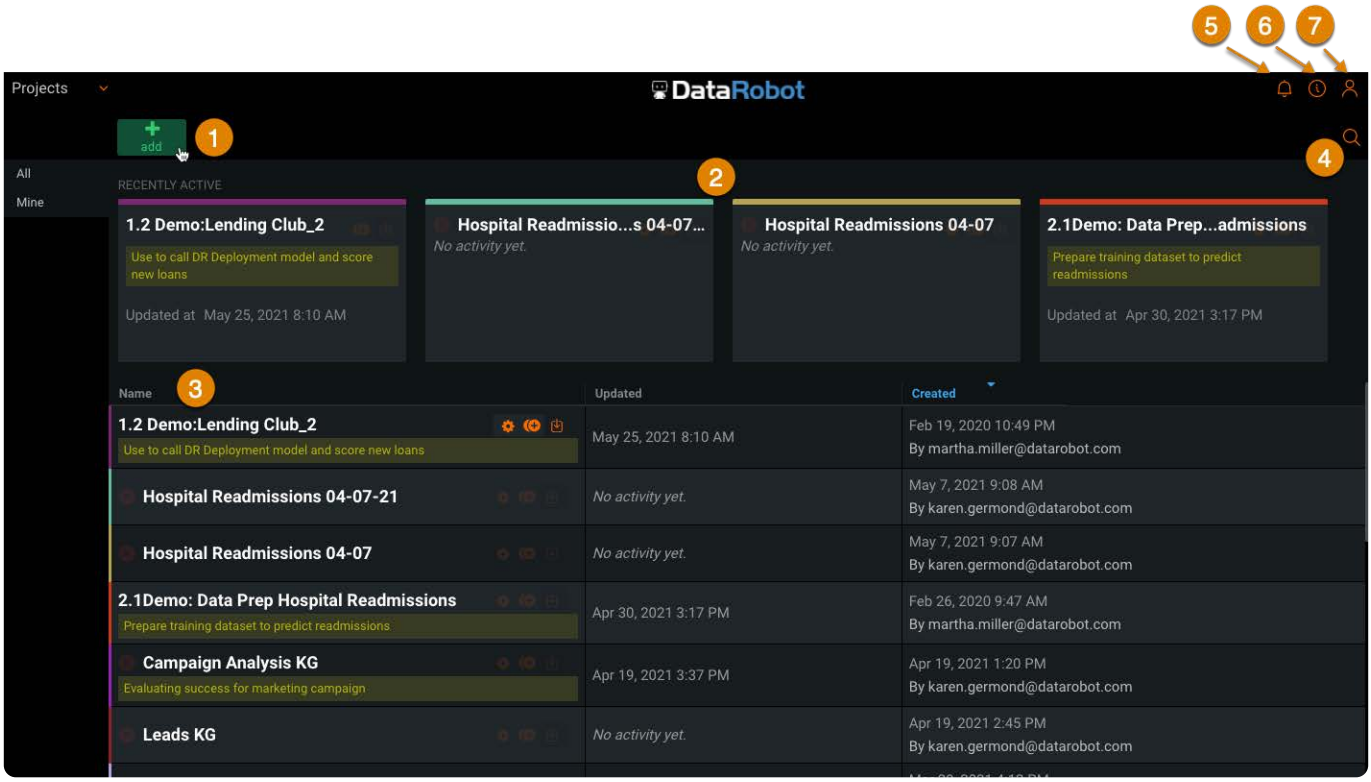
| | アクション | 説明 |
|---|---|---|
| 3 | **その他のアクション** | Data Prepアプリケーションで有効になっている機能に応じて、追加のオプションを提供します。<br>・**詳細の編集**：データセットの**一般**ページを開きます。ここで、データセットの名前とメタデータを更新できます。このページには、インポート中に発生した警告またはエラーも表示されます。注意またはエラーのあるデータセットはリストで簡単に見つけることができます。データセット名の横に注意アイコンが表示され、データセットの行の色が赤になり、ステータスアイコンが障害状態を示します。<br>・**バージョンの追加**：現在のバージョンを上書きすることなく、既存のデータセットの新しいバージョンを追加します。<br>・**自動化**：データセットを自動化します（自動化機能が有効になっている場合）。<br>・**プロファイル**：データセットのプロファイルを設定します（プロファイリング機能が有効な場合）。<br>・**オープンソース**：AnswerSet について、そのAnswerSetが作成された正確なステップでプロジェクトを開きます。 |

# Data Prepプロジェクト

Data Prepで、*プロジェクト*のデータを調査して準備します。このトピックでは、トップレベルの**プロジェクト**ページと、個々のプロジェクトの**プロジェクト準備**ページについて説明します。

## プロジェクトページ

**プロジェクト**ページは、Data Prepにログインした後に表示されるホームページです。




備考：データセットが保存されるライブラリページを開きたい場合は、左上のメニューをクリックして **ライブラリ**を選択します。

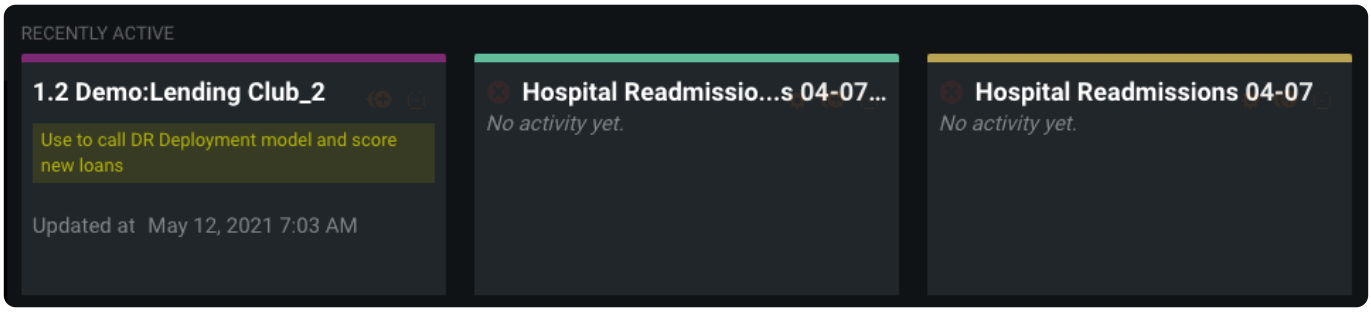
次の表は、**プロジェクト**ページのセクションについて説明しています。

| | アクション | 説明 |
|---|---|---|
| 1 | **+プロジェクトの追加** | 新しいプロジェクトを作成。一意の名前と、オプションで説明を入力します。プロジェクトを作成すると、データをインポートできます。 |

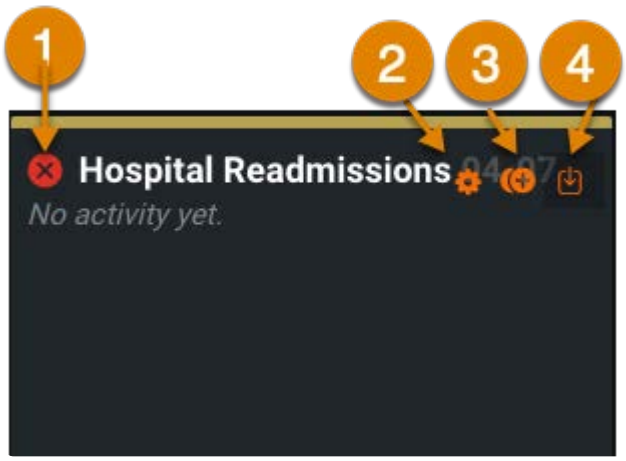
| | アクション | 説明 |
|---|---|---|
| 2 | **最近アクティブ** | 最近アクティブなプロジェクトは、ページの上部にリストされます。必要なプロジェクト名をクリックして開きます。 |
| 3 | **すべてのプロジェクトのインベントリ** | ここに、すべてのプロジェクトが一覧表示されます。デフォルトでは、最終更新日の新しい順にプロジェクトが表示されることに注意してください。**名前**列のヘッダーをクリックして、プロジェクト名に基づいてリストを並べます。**更新済み**または**作成済み**の列見出しをクリックすると、プロジェクトの更新または作成の日付に基づいて表示するようにリストの順序を変更できます。 |
| 4 | **SEARCH** | 虫眼鏡アイコンをクリックして、プロジェクトを名前で検索します。 |
| 5 | **通知アイコン** | アプリケーションがいつメッセージまたはエラーを生成したかを示します。強調表示されている場合、アイコンの上にカーソルを合わせるとメッセージが表示されます。 |
| 6 | **ヘルプ** | ・**表示／非表示：**アプリケーションのインラインヘルプのメニューを開いたり閉じたりします。<br>・**新機能:** アプリケーションの最新リリースにおけるすべての新機能の概要を一覧表示します。<br>・**使用を開始する：**新しいユーザーガイドを開きます。<br>・**ヘルプシェルフ：**すべてのヘルプドキュメントのホームページを開きます。<br>・**フィードバック：**DataRobotのカスタマーサクセスチームに連絡するメールを開きます。 |
| 7 | **ユーザーメニュー** | ・**マイアカウント:** アカウント情報を表示し、パスワードを変更するためのオプションを提供します。<br>・**トークン：**アプリケーションのアクセスと承認を管理するために使用されるトークンを生成します。DataRobotシステム管理者は、トークンを生成する必要があるときに通知します。<br>・**製品情報:** アプリケーションの現在のバージョン番号に関する詳細情報を表示します。<br>・**ログアウト:** アプリケーションからログアウトします。 |

# 最近使ったプロジェクト

最近使ったプロジェクトは、**最近アクティブ**の下にタイルとして表示されます。



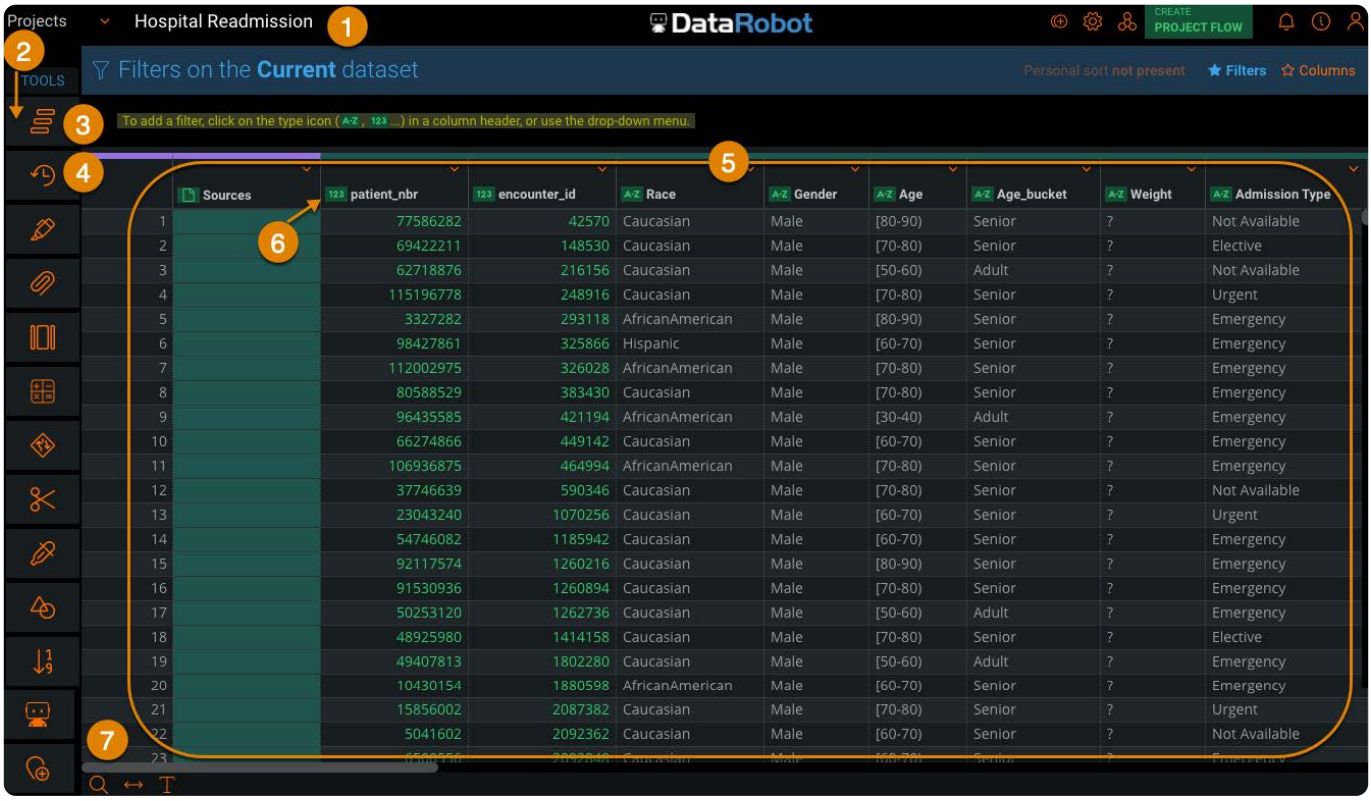

各プロジェクトタイルには、次のコントロールがあります。




| | アクション | 説明 |
|---|---|---|
| 1 | **プロジェクトを削除** | プロジェクトを削除するには、左上の赤いXアイコンをクリックします。確認ボックスで**はい**をクリックします。自分が作成したプロジェクトのみを削除できます。 |
| 2 | **名前または説明の編集** | プロジェクトの名前と説明を編集します。 |
| 3 | **このプロジェクトの最新バージョンのコピー** | プロジェクトを複製します。 |
| 4 | **プロジェクトレポートのダウンロード** | プロジェクトレポートのファイル（Word形式）をダウンロードします。レポートのファイルには、プロジェクト名とメタデータ、列データ、およびデータの準備に使用される変換手順が含まれています。 |

# プロジェクトの準備ページ

データセットまたはそのデータセットまたは**プロジェクト**ページの**最近アクティブ**のタイルをクリックするか、または**ライブラリ**ページでプロジェクトを作成して、プロジェクトを開きます。**プロジェクトの準備**ページが表示され、データセットを準備できます。




| Sources | patient_nbr | encounter_id | Race | Gender | Age | Age_bucket | Weight | Admission Type |
|---|---|---|---|---|---|---|---|---|
| 1 | 77586282 | 42570 | Caucasian | Male | [80-90) | Senior | ? | Not Available |
| 2 | 69422211 | 148530 | Caucasian | Male | [70-80) | Senior | ? | Elective |
| 3 | 62718876 | 216156 | Caucasian | Male | [50-60) | Adult | ? | Not Available |
| 4 | 115196778 | 248916 | Caucasian | Male | [70-80) | Senior | ? | Urgent |
| 5 | 3327282 | 293118 | AfricanAmerican | Male | [80-90) | Senior | ? | Emergency |
| 6 | 98427861 | 325866 | Hispanic | Male | [60-70) | Senior | ? | Emergency |
| 7 | 112002975 | 326028 | AfricanAmerican | Male | [70-80) | Senior | ? | Emergency |
| 8 | 80588529 | 383430 | Caucasian | Male | [70-80) | Senior | ? | Emergency |
| 9 | 96435585 | 421194 | AfricanAmerican | Male | [30-40) | Adult | ? | Emergency |
| 10 | 66274866 | 449142 | Caucasian | Male | [60-70) | Senior | ? | Emergency |
| 11 | 106936875 | 464994 | AfricanAmerican | Male | [70-80) | Senior | ? | Emergency |
| 12 | 37746639 | 590346 | Caucasian | Male | [70-80) | Senior | ? | Not Available |
| 13 | 23043240 | 1070256 | Caucasian | Male | [60-70) | Senior | ? | Urgent |
| 14 | 54746082 | 1185942 | Caucasian | Male | [60-70) | Senior | ? | Emergency |
| 15 | 92117574 | 1260216 | Caucasian | Male | [80-90) | Senior | ? | Emergency |
| 16 | 91530936 | 1260894 | Caucasian | Male | [70-80) | Senior | ? | Emergency |
| 17 | 50253120 | 1262736 | Caucasian | Male | [50-60) | Adult | ? | Emergency |
| 18 | 48925980 | 1414158 | Caucasian | Male | [70-80) | Senior | ? | Elective |
| 19 | 49407813 | 1802280 | Caucasian | Male | [50-60) | Adult | ? | Emergency |
| 20 | 10430154 | 1880598 | AfricanAmerican | Male | [60-70) | Senior | ? | Emergency |
| 21 | 15856002 | 2087382 | Caucasian | Male | [70-80) | Senior | ? | Urgent |
| 22 | 5041602 | 2092362 | Caucasian | Male | [60-70) | Senior | ? | Not Available |

次の表は、**プロジェクト準備**ページのセクションについて説明しています。

| | アクション | 説明 |
|---|---|---|
| 1 | **プロジェクト名** | プロジェクトに付けた名前がここに表示されます。 |
| 2 | **ツールバー** | プロジェクトツールを使用して、データのクリーンアップ、整形、組み合わせを行い、最終的に準備します。 |
| 3 | **ステップツール** | データの準備中に実行するすべてのアクションは、ステップとしてログに記録されます。ステップツールにより次のことができます。<br>・ステップを順番に表示する。<br>・ステップをミュートにする。<br>・ステップ中に起こることを編集する。<br>・データ準備ステップの順序を並べ替える。<br>・ステップを削除する。 |
| 4 | **バージョンツール** | プロジェクトを保存すると、そのたびに新しいバージョンが作成されます。バージョンツールを使うと、プロジェクトの以前のバージョンにアクセスできます。 |

| | アクション | 説明 |
|---|---|---|
| 5 | **データプレビュー** | 準備中のデータへの変更を表示します。 |
| 6 | **列メニュー** | クリックして列メニューを開きます。列の操作を使用して、データのクリーニングと標準化を行います。 |
| 7 | **グリッドツールとステータスの更新** | グリッド ツールを使用すると、データセット内の特定の列を見つけたり、列幅を指定したり、セルテキストの表示方法を調整したりすることができます。ステータスの更新は、データのプレビューグリッドまたはフィルターに影響を及ぼす変換の進行中に表示されます。更新メッセージに表示されるタスクの数は、操作が完了するにつれて動的に変化する場合があります。 |

# Data Prep用のデータソースに接続

これらのページには、Data Prepがサポートする各コネクターの設定手順が含まれています。詳細に入る前に、コネクター全般の設定について読むことができます。

設定方法については、以下のコネクターを選択してください。

| コネクター | 説明 |
| --- | --- |
| Amazon Athena | インポートソースとしてAWS Athenaに接続します。 |
| Amazon S3 | Amazon S3 オブジェクトストレージに対してデータをインポートおよびエクスポートします。 |
| Amazon Redshift | インポートおよびエクスポートソースとしてAmazon Redshiftに接続します。 |
| Amazon DynamoDB | ライブラリをインポートするには、Amazon DynamoDBに接続します。 |
| Cloudera CDH5 HDFS | インポートおよびエクスポートのためにCloudera CDH 5.16 Hadoopファイルシステム(HDFS)に接続します。 |
| Cloudera CDH6 HDFS | インポートとエクスポートのためにHadoopファイルシステム（HDFS）クラスターに接続します。 |
| Cloudera CDH5 Hive | インポートとエクスポートのためにCloudera CDH 5.16 Hiveに接続します。 |
| Cloudera CDH6 Hive | インポートとエクスポートのためにHiveデータベースに接続します。 |
| Databricks | ライブラリのインポートとエクスポートのためにDatabricksに接続します。 |
| DataRobot | ライブラリのインポートとエクスポートのためにDataRobotに接続します。 |
| Google Analytics | 利用可能なデータを参照およびインポートするには、Google Analyticsに接続します。 |
| Google BigQuery | 利用可能なデータをインポートおよびエクスポートするには、BigQueryに接続します。 |
| Google Cloud Storage | オブジェクトを参照およびインポートするには、Google Cloud Storage（GCS）に接続します。 |

| コネクター | 説明 |
| --- | --- |
| Google Cloud SQL | Data Prep JDBCコネクターを使用してCloud SQLに接続します。 |
| Google Drive | 利用可能なデータを参照およびインポートするには、Google Driveに接続します。 |
| Google Sheets | Googleスプレッドシートコネクターは使用非推奨になりました。Google Sheetsのインポートとエクスポートに対応したGoogle Drive Connectorを使用します。 |
| Hortonworks HDP2 HDFS | インポートおよびエクスポートのためにHortonworks HDP 2.6.5 Hadoopファイルシステム（HDFS）に接続します。 |
| Hortonworks HDP2 Hive | インポートおよびエクスポートのためにHortonworks HDP 2.6.5 Hiveに接続します。 |
| Hubspot | 利用可能なデータを参照およびインポートするには、HubSpotに接続します。 |
| IBM DB2 | Data Prep JDBCコネクターを使用してIBM DB2に接続します。 |
| IBM Netezza | Data Prep JDBCコネクターを使用してIBM Netezzaに接続します。 |
| JDBC | Javaデータベース接続（JDBC）ドライバーを利用してデータをインポートおよびエクスポートします。通常、このコネクターはリレーショナルデータベースに対するインポート／エクスポートに利用されますが、多くのアプリケーションでJDBCドライバーを提供しています。 |
| Jira | 利用可能なデータを参照およびインポートするには、Jiraに接続します。 |
| Marketo | インポートソースとしてMarketoに接続します。 |
| MicroStrategy | ライブラリのインポートとエクスポートのためにMicroStrategyサーバーに接続します。 |
| MongoDB | 利用可能なデータを参照およびインポートするには、MongoDBに接続します。 |
| MS Azureデータレイクストア（ADLS） | ライブラリのインポートとエクスポートのためにAzureデータレイクストレージ（ADLS）に接続します。 |
| MS Azsureデータレイクストア Gen2（ADLS Gen2） | インポートとエクスポートのためにAzureデータレイクストレージGen2に接続します。 |
| MS Azure SQL | Data Prep JDBCコネクターを使用してAzure SQLに接続します。 |

| コネクター | 説明 |
| --- | --- |
| MS Azure Synapse Analytics | ライブラリのインポートとエクスポートのためにAzure Synapse Analyticsに接続します。 |
| MS Dynamics 365 | エンティティセットのインポートのためにMicrosoft Dynamics 365リソースに接続します。 |
| MS Sharepoint | ファイルとSharePointリストのライブラリのインポートとエクスポートのためにSharePointサイトに接続します。 |
| MS SQL Server | Data Prep JDBCコネクターを使用してMS SQL Serverに接続します。 |
| MS Windows Azure Blob ストレージ (WASB) | ライブラリのインポートとエクスポートのためにAzure Blobストレージのアカウントに接続します。 |
| MySQL | Data Prep JDBCコネクターを使用してMySQLに接続します。 |
| Netsuite | 利用可能なデータを参照およびインポートするには、NetSuiteに接続します。 |
| ネットワーク共有（SMB/Samba） | インポートおよびエクスポートにサーバーメッセージブロック（SMB）プロトコルを使用してネットワーク共有に接続します。 |
| Oracle | Data Prep JDBCコネクターを使用してOracleに接続します。 |
| Oracle Marketing Cloud (Eloqua) | ライブラリのインポートのためにOracle Marketing Cloudに接続します。 |
| PostgreSQL | Data Prep JDBCコネクターを使用してPostgreSQLに接続します。 |
| PowerBI | PowerBIへの接続は、PowerBIデスクトップで設定されます。接続方法の詳細についてはこちらをご覧ください。 |
| REST API | REST APIに接続してRESTリソースをインポートします。 |
| Salesforce Lightning | インポートソースとしてSalesforce組織に接続します。 |
| Salesforce Marketing Cloud | 利用可能なデータを参照およびインポートするには、Salesforce Marketing Cloudに接続します。 |
| SAP HANA | Data Prep JDBCコネクターを使用してSAP HANAに接続します。 |

| コネクター | 説明 |
| --- | --- |
| Spark SQL | 利用可能なデータを参照、インポート、およびエクスポートするには、Spark SQLに接続します。 |
| SFTP | ライブラリのインポートおよびエクスポートのために、SSHファイル転送プロトコル（SFTP）サーバーに接続します。 |
| Snowflake | Data Prep JDBCコネクターを使用してSnowflakeに接続します。 |
| Tableau .Hyper | エクスポート先としてTableauに接続します。 |
| Tableau.tde（使用非推奨） | エクスポート先としてTableauサーバーおよびTableau Onlineに接続します。 |
| Teradata | Data Prep JDBCコネクターを使用してTeradataに接続します。 |
| Thoughtspot | Data Prep AnswerSetをThoughtSpotにエクスポートします。 |
| Vertica（HP） | Data Prep JDBCコネクターを使用してVerticaに接続します。 |
| Zendesk | 利用可能なデータを参照およびインポートするには、Zendeskに接続します。 |

# Data Prep用のData Prepコネクターのセットアップ

## Data Prepコネクターとは？

すべてのData Prepストーリーは、コネクターで始まり、コネクターで終わります。データの準備ができることは、準備に必要なデータを取得し、準備後に必要な場所にそのデータを送信できる場合にのみ価値があります。Data Prepコネクターは、Data Prepとの間でデータを送受信するためのツールです。

## Data Prepコネクターの利点

### ビジネスユーザーのための簡単なデータアクセス

異種システム上のデータへのアクセスは、コーダーにとってそれほど複雑ではありません。ほとんどのデータベース、ファイルストア、およびWebサービスには、業界標準に準拠した、十分に開発されたコードフレンドリーなインターフェイスがあります。

### コーディングをしないユーザーにとってデータのインテグレーションは難しいものです。

DataRobotはこの問題に取り組み、DataRobot Data Prepのノンコーディングユーザーに可能な限り多くのデータソースを開放しました。当社の目標は、ビジネスアナリスト（ノンコーディングユーザー）が、使用を許可されている組織内の任意のデータにアクセスできるようにすることです。

### ブラウジングとクエリーの比較

ノンコーディングユーザーを有効にするための核心的な側面の一つは、ブラウジングインターフェイスです。他のデータ準備またはETLソリューションがSQLクエリーに依存している場合、Data Prep内のすべてのデータソースを参照し、クリックするだけでデータをインポートできます。

### コントロールとガバナンス

ビジネス環境は、通常のITインフラが対応するよりもはるかに流動的ですが、それでも、特定の人は特定の情報にしかアクセスできず、その情報を特定の場所にしか送れないようにすべきです。コネクターフレームワークを使用すると、大規模で複雑な組織は、ユーザーが自分に付与された情報にのみアクセスできるようにし、速度とセルフサービスが優先される小規模な組織向けに簡単に設定できます。

# Data Prep コネクターのセットアップ

## 3層の設定

コネクターを設定する際には、上位から下位に向かって、それぞれ「コネクター」、「データソース」、「セッション」の3つの階層レベルがあります。フィールドが上位レベルで入力されている場合、下流で再度入力する必要はありません。一部のフィールドは後の段階で変更できる場合がありますが、それはコネクターによって大きく異なります。

### コネクター設定

このレベルは通常、管理者やITによって作成、管理され、以下の目的のために存在します。

- 特定のユーザーグループが特定のコネクターを利用できるようにします。
- 管理者が、ユーザーが知らない情報や、コネクター設定に依存するすべてのユーザー/データソースで同じ情報を入力できるようにします。
- また、管理者は、SSHキーなど、アクセス権を持たないユーザーから機密情報を保護することができます。

### データソース設定

このレベルは、ソースシステムデータへのアクセスがどのように管理されているかに応じて、通常、個々のユーザーまたは管理者のいずれかによって作成・管理され、次の目的で存在します。

- コネクター設定レベルでまだキャプチャされていないすべての永続的な設定を含みます。
    - 通常これには、共有データソース設定の実行時に指定されるユーザー資格情報を除くすべてが含まれます。

### セッション設定

このレベルは、ほとんどが個々のユーザーによって独占的に管理されているか、不要な場合は無視され、次の目的で存在します。

- インポート/エクスポートの実行時に情報をキャプチャします。
    - 通常、これはユーザーの資格情報に限定されます。

## コントロールの共有

- コネクターとデータソースの設定は、テナント内のグループと共有できます。
- これらの共有コントロールを使用すると、指定したグループのメンバーが設定を読み取り、更新、または削除できるかどうか、およびユーザーが設定を使用してインポートやエクスポートを実行できるかどうかを指定することもできます。

# セットアップの例

以下は、ビジネスの状況と、各チームのニーズに合わせてコネクターフレームワークをどのように設定するかに関する例です。

# 例1：

## ビジネスの状況

「SSH Key with Passphrase」によって認証されたIT管理のSFTPサーバー。キーとパスフレーズはIT部門によって保持され、複数のチームが異なるディレクトリにアクセスする必要があります。

## セットアップ

- コネクター設定
  - IT部門は1つのコネクター設定を作成し、SFTPホストとポート、SSHキーとパスフレーズを入力します。
  - 共有：なし
- データソース設定
  - チームごとに新しいデータソースを作成し、適切なルートディレクトリを指定します。
  - 共有：完全に設定された各データソースを、対応するチームと読み取り専用で共有し、必要に応じてインポートやエクスポートを許可します。
- セッション設定
  - N/A

## このアプローチの利点

- 資格情報が変更された場合でも、1か所で管理するだけで済みます。
- IT部門は資格情報を管理し、ユーザーから非公開にすることができます。
- 各チームは、データソース自体のアクセス制御を管理しなくても、必要なアクセス権を持っています。

# 例2

## ビジネスの状況

管理者が管理するSalesforce組織では、各ユーザーはSalesforceで権限を持つ情報のみにアクセスする必要があり、各ユーザーはData Prep内で自動化ジョブを実行する必要があります。

## セットアップ

- コネクター設定
  - Salesforce管理者は、1つのコネクター設定を作成し、ユーザーとパスワードを除くすべての関連情報を入力します。
  - 共有：この設定と関連する各グループと読み取り専用で共有します。
- データソース設定
  - 各ユーザーは、独自のデータソース設定を作成し、資格情報のみを入力して、セットアップを維持し、自動化ジョブで使用できるようにする必要があります。

    - 共有：なし
- セッション設定
    - N/A

## このアプローチの利点

- 管理者レベルのセットアップは管理者が行い、各ユーザーはユーザー名とパスワード、すぐに利用できる情報だけを入力する必要があります。
- 各ユーザーの認証はSalesforceで管理されます。

# Data Prep用のAmazon Athenaコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、インポートソースとしての AWS Athena に接続できます。データソースで設定する必要があるフィールドは、管理者が行ったコネクターの設定に応じて異なります。

### 一般

- **名前：**UIでユーザーに表示されるデータソースの名前。
- **説明：**UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数のAWS Athenaインスタンスに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Amazon Athenaの設定

- **Athenaリージョン：**AWSがホストする領域。
- **アクセスキー：**AWSアカウントのアクセス キー。
- **シークレットキー：**AWS アカウントのシークレットキー。

### クエリ結果ストレージの設定

- **S3バケット名：**Athenaで クエリ結果を格納するS3バケットの名称。

- **S3オブジェクトの接頭辞：**Athena が指定されたS3バケット内にクエリ結果を格納する際に使用する接頭辞。接頭辞に関する詳細については、 S3 バケットでのフォルダーの使用方法 を参照してください。
- **暗号化タイプ：**AWSサーバー側の暗号化の種類。

### クエリ結果について

アテナを使用する場合、各クエリ結果は設定されたS3バケットに保存されます。Athena はこのように動作するように設計されており、これは想定される動作です。Athenaを使用してData Prepにインポートする場合、接続が閉じた場合はクエリー結果がデフォルトでクリーンアップされます。Athena のコネクターは、当該クエリ結果を保持するインスタンスが複数存在しないように、この削除タスクを実行するように設計されています。Data Prepへのインポートからのクエリー結果をS3で引き続き利用できるようにする場合は、Athenaスタンドアロンでクエリーを実行し、結果ファイルをS3からData Prepにインポートするだけです。

## Webプロキシ設定

プロキシサーバーを介してAWSアテナに接続する場合、これらのフィールドは、プロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」、AWSAthenaへの接続をプロキシサーバー経由で行う必要がある場合は「プロキシ」。 Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Webプロキシ サーバーのホスト名またはIPアドレス。
- **プロキシサーバー：**データソースのプロキシサーバー上のポート。
- **プロキシユーザー名：**プロキシ サーバーのユーザー名。
- **プロキシ パスワード：**プロキシ サーバーのパスワード。

認証されていないプロキシ接続では、**ユーザー名**と**パスワード**を空白のままにします。

# データインポート情報

## ブラウジング経由

このコネクタではブラウジングがサポートされており、アテナクエリを使用してブラウジング可能な階層を生成します。Athenaのコスト構造については、以下の備考を参照してください。

## SQLクエリー経由

詳細については、SQLリファレンスにアクセスしてください。

# ベストプラクティス

Athena を使用するにあたっては、実行するクエリごとに課金がされます。発生する料金は、クエリが読み込むデータの量に基づきます。詳細については、 Amazon Athenaの価格設定を参照してください。

# Data Prep用のAmazon S3コネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターにより、Amazon S3オブジェクトストレージに対してデータをインポートおよびエクスポートできるようになります。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数のS3バケットに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### Amazon S3 クライアントの設定

- **バケット名:** AS3バケットは、Amazon S3に保存されているオブジェクトの集合を表します。 コネクターには、s3:ListBucket、s3:GetObject、およびs3:PutObject (エクスポートの場合のみ) のアクセス権限が必要です。 さらに 、バケットポリシーにSourceIP条件ブロックが指定されている場合は、メインコアサーバーとオートメーションコアサーバー（ある場合）のIPアドレスを含める必要があります。

  詳細については、この記事の下部にあるAWS S3バケットのアクセス許可/ポリシーの詳細を参照してください。
- **プレフィックス：**指定されたプレフィックスで始まるキーのみに結果を制限します。
- **暗号化タイプ:**サーバー側の暗号化に使用するタイプです。詳細については、AWS暗号化タイプを参照してください。

- バケットリー**ジョン：**このオプションを使用すると、ユーザーはS3バケットがホストされているリージョンを指定したり、コネクタがリージョンを自動的に決定するように選択したりできます。

## Amazon S3認証

これらのオプションで、S3 の認証方法を指定します。

- **AWS資格情報：**ユーザーのAWSアクセスキーに関連付けられたアクセスキーIDとシークレットキー。これはデフォルトの設定です。

  詳細については、AWSのセキュリティ資格情報を参照してください。

- **インスタンスプロファイル（IAMロール）：**このテナント内のすべてのユーザーを、個別に認証することなくAWSにアクセスできるようにします。

  詳細については、インスタンスプロファイル（IAMロール）を使用して Amazon EC2上のAWSリソースへのアクセスを許可するを参照してください。

> **備考**
>
> このコネクターは、EC2サーバーインスタンスから認証情報を自動的に取得します。

- **IAMクロスアカウント：**S3へのアクセスを有効にするには、設定されたS3バケットにアクセスできる別のAWSアカウントのロールを引き受けます。

  詳細については、クロスアカウントのアクセスを参照してください。

> **備考**
>
> インスタンスプロファイル（IAMロール）およびIAMクロスアカウントオプションの場合、Data PrepがアマゾンEC2ホストにインストールされている必要があります。

## Webプロキシ

プロキシサーバーを介してAmazon S3に接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」、プロキシサーバー経由でZendesk RESTエンドポイントに接続する必要がある場合は「プロキシ」。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシ ホスト:** Web プロキシ サーバーのホスト名または IP アドレスです。
- **プロキシサーバー：**データソースのプロキシサーバー上のポート。
- **プロキシ ユーザー名：**プロキシ サーバーのユーザー名です。
- **プロキシ パスワード：**プロキシ サーバーのパスワードです。 *認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

## その他の設定

**ソケットタイムアウトの秒数：**確立された接続でAmazon S3からの応答を待つ秒数です。デフォルト値は5分です。大きなファイルのエクスポートを処理するには、この値を増やします。

# データのインポートとエクスポートに関する情報

## ブラウジング経由

コネクターは、プレフィックスフィールドで定義された場所から始まるブラウズ可能なディレクトリ階層を表示します。

コネクターはワイルドカードとグロブのインポートもサポートしているため、ユーザーは複数のS3データファイルを単一のデータセットとしてData Prepにインポートできます。

## SQLクエリー経由

S3はファイルストアであるため、このデータソースではSQLクエリーはサポートされていません。AWS S3データに直接クエリーを実行したい場合は、Data PrepのAWS Athenaコネクターに関してカスタマーサクセス担当者に連絡してください。

# AWS S3バケットのアクセス許可/ポリシーの詳細

このセクションでは、S3バケットポリシーで割り当てる必要があるアクセス権限と、バケットポリシーでSourceIP条件ブロックが指定されている場合に行う必要があることを確認します。

## 必要な権限

AWS S3 コネクターでは、S3 からのデータのインポート、S3 への公開、S3 ソースからのインポートの自動化を正しく実行できるように、S3 バケット ポリシーに特定の権限が必要です。サマリーを以下に示します。

- コネクターは、ブラウジングには、バケットでの、*s3:ListBucket* の権限が必要です。
- バケットの内容をインポートするには、Data Prepに *s3:GetObject* の権限が必要です
- バケットにエクスポートするには、Data Prepに *s3:PutObject* の権限が必要です

## サンプルバケットポリシーの例

```
{
	"Version": "2012-10-17",
	"Statement": [
		{
			"Sid": "Stmt1965292834357",
			"Effect": "Allow",
			"Principal": {
				"AWS": [
					"arn:aws:iam::123456781234:user/pax01",
					"arn:aws:iam::432143214321:user/pax02",
					"arn:aws:iam::121212343434:user/pax03"
				]
			},
			"Action": "s3:ListBucket",
			"Resource": "arn:aws:s3:::paxhh-session1"
		},
		{
			"Sid": "Stmt1965293102818",
			"Effect": "Allow",
			"Principal": {
				"AWS": [
					"arn:aws:iam::123456781234:user/pax01",
					"arn:aws:iam::432143214321:user/pax02",
					"arn:aws:iam::121212343434:user/pax03"
				]
			},
			"Action": [
				"s3:DeleteObject",
				"s3:DeleteObjectVersion",
				"s3:GetObject",
				"s3:GetObjectAcl",
				"s3:GetObjectTorrent",
				"s3:GetObjectVersion",
				"s3:GetObjectVersionAcl",
				"s3:GetObjectVersionTorrent",
				"s3:PutObject",
				"s3:PutObjectAcl",
				"s3:PutObjectVersionAcl"
			],
			"Resource": "arn:aws:s3:::paxhh-session1/*"
		}
	]
}
```

## 最小のポリシー権限

S3 バケットからの読み取りに必要なMinimum（最小）のポリシー権限は次のとおりです。

```
{
	"Version": "2012-10-17",
	"Statement": [
		{ "Effect": "Allow", "Action": "s3:ListBucket", "Resource":
		"arn:aws:s3:::mybucketname"
		},
	{ "Effect": "Allow", "Action": [ "s3:ListBucket", "s3:GetObject" ],
	"Resource": "arn:aws:s3:::mybucketname/*" }
	]
}
```

S3 バケットへの書き込みに必要なMinimum（最小）のポリシー権限は次のとおりです。

```
{
     "Version": "2012-10-17",
     "Statement": [
          { "Effect": "Allow", "Action": "s3:ListBucket", "Resource":
          "arn:aws:s3:::mybucketname"
          },
     { "Effect": "Allow", "Action": [ "s3:ListBucket", "s3:GetObject",
     "s3:PutObject" ], "Resource": "arn:aws:s3:::mybucketname/*" }
     ]
}
```

S3バケットの詳細については、Amazon S3バケットの操作を参照してください。

## SourceIP 条件ブロック

バケットポリシーで**SourceIP条件ブロック**が指定されている場合、Data Prepクラウドサーバーまたは Data Prepコアサーバー（Data Prepデプロイに応じて）のIPアドレスを**SourceIP条件ブロック**に含める必要があります。さらに、自動化専用のData Prepサーバーがある場合は、オートメーションサーバーのIPアドレスも**SourceIP条件ブロック**に含める必要があります。

Data Prepのカスタマーサクセスチームに問い合わせて、Data Prepクラウドサーバーのアドレスのリストを取得してください。

条件ブロックの要素と例の詳細については、ポリシーでの条件の指定およびIdentity and Access Management（IAM）ポリシー要素のリファレンスを参照してください。

# Data Prep用のAmazon Redshiftコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、IT/DvOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、インポートおよびエクスポートソースとしてAmazon Redshiftに接続できます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

あなたは複数のRedshiftのデータウェアハウスへのデータ準備を接続することができます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### データベース URL

- **JDBCのURL：**JDBC接続文字列。URLにデータベース名を含めることができます。
- JDBC URLの例：jdbc:redshift://examplecluster.abc123xyz789_._us-west-2.redshift.amazonaws.com:5439/dev?ssl=true
- 接続文字列オプションの詳細については、このAWSドキュメントを参照してください。

## 可視性の設定

インポート中にユーザーがデータソースを参照するときに表示されるデータベース、スキーマ、およびテーブルを制御できます。データベース、スキーマ、およびテーブルの場合、次のいずれかを選択できます。

- [表示のみ]ここで指定したデータベース、スキーマ、またはテーブルだけが返されます。
- [非表示]：ここで指定したデータベース、スキーマ、テーブルが非表示になります。
- [すべて表示]：データソース内のすべてを表示するデフォルト設定です。

[表示のみ]または[非表示]オプションを選択すると、オプションを適用するデータベース、スキーマ、またはテーブルを指定するフィールドが表示されます。

> **備考**
>
> これらの設定は、ユーザーがデータソースに対してクエリーを実行する場合は適用されません。クエリー結果は、一致の完全なリストを返します。たとえば、特定のデータベースを[非表示]にした場合でも、ユーザーはそのデータベース内のテーブルからデータをプルするクエリーを実行できます。ただし、そのデータベースは、ユーザーがデータソースを参照するときに表示されません。

## インポート設定

- **インポート前のSQL：** テーブルのスキーマを決定した後、インポートの開始前に実行するSQLステートメントです。
- **インポート後のSQL：**インポートの完了後に実行するSQLステートメントです。

## エクスポート設定

- **エクスポート前のSQL：**自動作成が有効になっている場合、テーブルの作成後、エクスポートの開始前に実行するSQLステートメント。
- **エクスポート後のSQL：**エクスポートの完了後に実行するSQLステートメント。

## Redshift資格情報

ユーザー認証は、共有アカウントまたは個人アカウントを介して行うことができます。選択に応じて、次のフィールドが必須です。

- **個人アカウント：**
  - **ユーザー：**データベースへの認証に使用される個人アカウントのユーザー名です。
  - **パスワード：**データベースへの認証に使用される個人アカウントのパスワードです。
- **共有アカウント：**
  - **ユーザー：**データベースへの認証に使用される共有アカウントのユーザー名です。
  - **パスワード：**データベースへの認証に使用される共有アカウントのパスワードです。

- **ロール：**このデータベースにロールが実装されている場合、このユーザーロールを持つ認証済みユーザーは、認証後にクエリーを実行できます。

## Amazon S3 クライアントの設定

- **S3 を使用するエクスポートの指定：**このオプションでは、コネクターによって Redshift にデータをエクスポートする際、データをまず Amazon S3 にアップロードしてから Redshift にコピーするのか、あるいは Redshift に直接データを挿入するのかを指定します。
- **S3 を使用してエクスポート：** データをAmazon S3 にアップロードしてから Redshift にコピーします。これは、よりパフォーマンスの高いエクスポートを可能にするため、大規模なデータセットに推奨されるアプローチです。
- **バケット名: ー＞：** Amazon S3 に保存されたオブジェクトのコレクションを表す S3 バケットの名称です。
- **プレフィックス：**指定されたプレフィックスで始まるキーのみに結果を制限します。
- **ソケットタイムアウトの秒数：** 確立された S3 接続からの応答があるまで待機する秒数です。デフォルト値は5分で、大きなファイルのエクスポートを処理するには増やす必要がある場合があります。
- **SQLのInsertステートメントを使用してエクスポート：**コネクターはRedshiftに直接データを挿入します。このオプションを使用すると、エクスポートが遅くなります。Redshiftからのインポートのみを実行する予定の場合は、このオプションを選択すると、S3アカウントの詳細を入力する必要がなくなります。

> **備考**
>
> コネクターには、バケットに対するs3:ListBucket権限が必要です。バケットのコンテンツには、s3:ListBucket、s3:GetObject、（エクスポートの場合のみ）s3:PutObject 権限が必要です。さらに、バケットポリシーにSourceIP条件ブロックが指定されている場合は、Data Prepサーバーと自動化ジョブの実行に使用するサーバーのIPアドレスを含める必要があります。

詳細については、AmazonS3コネクタのセットアップを参照してください。

## Amazon S3 認証の設定

- **AAWS資格情報：**アクセスキーIDとシークレットキーはユーザーのAWSアクセスキーに関連付けられています。
- **インスタンスプロファイル（IAMロール）：**追加フィールドは不要です。

詳細については、AWSのセキュリティ認証を参照してください。

## Webプロキシ

プロキシサーバーを介してAmazon Redshiftに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Web プロキシ**: プロキシが不要な場合は [なし]、プロキシ サーバー経由で Amazon Redshift RESTエンドポイントに接続する必要がある場合は [プロキシ] を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Web プロキシ サーバーのホスト名または IP アドレスです。

- **プロキシポート：**データソースのプロキシサーバー上のポートです。
- **プロキシユーザー名：**プロキシ サーバーのユーザー名です。
- **プロキシパスワード：**プロキシサーバーのパスワードです。 *認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

# データのインポート／エクスポート情報

## ブラウジング経由

上で選択したデータベース、スキーマ、およびテーブルの可視性設定と、指定されたユーザー資格情報に基づき、ブラウジングエクスペリエンスは異なります。

## SQLクエリー経由

可視性のセクションで説明したように、ユーザーがクエリを介してインポートできるものへの制限は、接続に指定された資格情報で判断される承認にのみ制限されます。

# Data Prep用のAmazon DyneroDBコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Amazon DynamoDBに接続して、ライブラリのインポートを行うことができます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明：** UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数のDynamoDBアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Webプロキシ

プロキシサーバーを介してDynamoDBに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ：** プロキシが不要な場合は [なし]を選択し、プロキシサーバー経由で DynamoDBに接続する必要がある場合は「プロキシ」を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：** Webプロキシ サーバーのホスト名またはIPアドレス。
- **プロキシサーバー：** データソースのプロキシサーバー上のポート。
- **プロキシユーザー名：** プロキシ サーバーのユーザー名です。
- **プロキシ パスワード：** プロキシサーバーのパスワード。

**備考：**認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

## AWS の構成

- **AWSリージョン：**DynamoDB APIにリクエストを送信する際に使用するリージョンを選択します。
- **AWS 認証タイプ：** このオプションで AWS の認証方法を指定します。
  - **AWS 資格情報：**ユーザーのAWSアクセスキーに関連付けられているアクセスキーIDとシークレットキーの入力を各ユーザーに要求します。これはデフォルトの設定です。詳細については、AWSセキュリティ認証を参照してください。
  - **インスタンスプロファイル（IAMロール）：**このテナント内のすべてのユーザーを、個別の認証なしにAWSにアクセスできるようにします。この認証方法は、AWS VPCにデプロイされた顧客のみが利用でき、この種の認証を許可するように設定されたEC2サーバーを備えています。このアプローチの詳細については、インスタンスプロファイル（IAMロール）を使用してAmazon EC2上のAWSリソースへのアクセスを許可するを参照してください。

> **備考**
>
> このコネクターは、EC2サーバーインスタンスから認証情報を自動的に取得します。

## DynamoDBテーブル設定

**サンプル項目：** インポートするテーブルのスキーマを決定する際に使用するレコードの数（従来のリレーショナルデータベースの「行」に相当）を指定します。

**備考：** DynamoDBはリレーショナルデータベースではありませんが、Data Prepはインポート時にすべてのデータを表形式に変換します。 これを行うために、Data PrepはDynamoDBテーブル (DynamoDBはドキュメントデータベース) の最初のn個のドキュメントを調べ、それらの属性を列として扱うことができるように、どの属性が存在するかを判断します。「サンプル項目」の値は、参照するドキュメントの数を決定します。

# データインポート情報

## ブラウジング経由

指定されたAWS地域のダイナモDBのテーブルは、インポートのために用意されています。

## SQLクエリー経由

サポートされていません。

# FAQ / トラブルシューティング / 一般的な問題

特定の権限はダイナモDBのからデータをインポートするために必要とされています。以下の権限になります。

- dynamodb:ListTables

- dynamodb:Scan

# Data Prep用のCloudera CDH5 HDFSコネクター

**ユーザーペルソナ**：Data Prep管理者、データソース管理者、またはIT/DevOps

**本機能の提供について**

このコネクターは、Data Prep SaaSのお客様はご利用いただけません。

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Cloudera CDH 5.16 Hadoopファイルシステム（HDFS）に接続して、インポートとエクスポートを行うことができます。次のフィールドを使用して、接続パラメーターを定義します。

**備考**

このコネクタを構成するには、データ準備サーバーでのファイルシステムアクセスとHadoopクラスタ構成 `core-site.xml` で。この手順については、カスタマーサクセス担当者にお問い合わせください。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

データ準備を複数の HDFS クラスターに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

## シンプル構成（シンプル認証の場合のみ）

- **ユーザー名：**アプリケーションWebサーバーは、ここで指定したユーザー名でHDFSクラスターに接続します。

## 設定

- **データストアのルートディレクトリ：**クラスターの「親ディレクトリ」です。インポートおよびエクスポート操作で、コネクターはこのディレクトリに対して読み書きを行います。ルートのサブディレクトリに対するインポートとエクスポートにも対応しています。

## Kerberos認証の構成

Kerberos認証には、以下のパラメータも設定する必要があります。

- **プリンシパル：** Kerberos認証のプリンシパルです。
- **レルム：** Kerberos認証のレルムです。
- **KDC ホスト名：**Kerberos認証のキー配布センターのホスト名です。
- **Kerberos認証の構成ファイル：**Webサーバー上のKerberos認証の構成ファイルの完全修飾パスです。
- **キータブファイル：**Webサーバー上のKerberos認証のキータブ ファイルの完全修飾パスです。
- **アプリケーションユーザーの使用：** ログイン中のアプリケーションユーザーで読み取りまたは書き込みを行う場合は、このチェックボックスをオンにし、プロキシユーザーを使用する場合は、このチェックボックスをオフにします。
- **プロキシ ユーザー：**クラスターでの認証に使用するプロキシです。プロキシユーザーとして${user.name}を入力します。${user.name}は、[アプリケーションユーザーを使用]を選択した場合と同様の動作をしますが、より柔軟性に優れています。例:
- ユーザーの認証情報にドメインを追加するには、[プロキシユーザー]フィールドに `\domain_name\${user.name}` と入力します。Data Prepではユーザー名とドメインが渡されます。
- 例： `\Accounts\${user.name}` はAccountsJoeになります（Joeがユーザー名であると仮定）。
- ユーザー名にテキスト修飾子を適用するには、キー `${user.name}` に.modifierを追加します。使用できる修飾子は ToLower、ToUpper、ToLowerCase、ToUpperCase、Trim です。
- たとえば、 `${user.name.toLowerCase}` はJoeをjoeに変換します（Joeがユーザー名であると仮定）。

# データインポート情報

## ブラウジング経由

サポートされています

## SQLクエリー経由

サポートされていません

# Data Prep用のCloudera CDH6 HDFSコネクター

**ユーザーペルソナ**：Data Prep管理者、データソース管理者、またはIT/DevOps

**本機能の提供について**

このコネクターは、Data Prep SaaSのお客様はご利用いただけません。

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、HDFS クラスターに接続し、データをインポートおよびエクスポートできます。ここで設定する必要があるフィールドは、選択した認証方法（シンプルまたはKerberos）によって異なります。選択した認証の種類は、コネクター設定に基づいて作成するすべてのデータソースに適用されます。

**備考**

このコネクターを設定するには、Data Prepサーバー上のファイルシステムへのアクセスと、Hadoopクラスター設定のcore-site.xmlが必要です。この手順については、カスタマーサクセス担当者にお問い合わせください。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数の HDFS クラスターに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

## Hadoop クラスター

- **認証方法：**シンプル認証またはKerberos認証を選択します。選択した認証の種類は、コネクター設定に基づいて作成するすべてのデータソースに適用されます。シンプル認証やKerberos認証の設定については、選択に応じて以下のセクションを参照してください。
- **クラスターコアサイトXMLのパス：**Webサーバー上のcore-site.xmlの完全修飾パスです。例: /path/to/core-site.xml
- **クラスターHDFSサイトXMLのパス：** Webサーバー上のhdfs-site.xmlの完全修飾パスです。例: /path/to/hdfs-site.xml
- **Native Hadoopライブラリのパス：**Webサーバー上のネイティブHadoopライブラリの完全修飾パスです。例: /path/to/libraries

## シンプル構成（シンプル認証の場合のみ）

- **ユーザー名：**アプリケーションWebサーバーは、ここで指定したユーザー名でHDFSクラスターに接続します。

## Kerberos認証の構成

Kerberosおよびハイブリッド認証には次のパラメーターが必要です。

- **プリンシパル：** Kerberos認証のプリンシパルです。
- **レルム：** Kerberos認証のレルムです。
- **KDC ホスト名：**Kerberos認証のキー配布センターのホスト名です。
- **Kerberos認証の構成ファイル：**Webサーバー上のKerberos認証の構成ファイルの完全修飾パスです。
- **キータブファイル：**Webサーバー上のKerberos認証のキータブ ファイルの完全修飾パスです。
- **アプリケーションユーザーの使用：** ログイン中のアプリケーションユーザーで読み取り、または書き込みを行う場合は、このチェックボックスをオンにし、プロキシユーザーを使用する場合は、このチェックボックスをオフにします。
- **プロキシユーザー：**クラスターでの認証に使用されるプロキシ。${user.name}はプロキシユーザーとして入力できます。${user.name}は[アプリケーションユーザーの使用]の選択と同様に機能しますが、より柔軟性があります。例:
  - ユーザーの認証情報にドメインを追加するには、[プロキシユーザー]フィールドに\domain_name\${user.name}と入力します。Data Prepではユーザー名とドメインが渡されます。
    - 例： \Accounts${user.name}はAccountsJoeになります（Joeがユーザー名であると仮定）。
  - ユーザー名にテキスト修飾子を適用するには、キー${user.name}に.modifierを追加します。使用できる修飾子はToLower、ToUpper、ToLowerCase、ToUpperCase、Trim です。
    - たとえば、${user.name.toLowerCase}はJoeをjoeに変換します（Joeがユーザー名であると仮定）。

## 設定

- **データストアのルートディレクトリ：**クラスターの「親ディレクトリ」です。インポートおよびエクスポート操作で、データライブラリはこのディレクトリに対して読み書きを行います。ルートのサブディレクトリに対するインポートとエクスポートにも対応しています。
- **INT96をDatetimeにマッピング：**インポート時にINT96タイプのフィールドをDatetime値に変換します。

## 資格情報

- **Hiveユーザー：** シンプル認証でHiveへのアクセスに使用するユーザー名です。
- **Hiveパスワード：**シンプル認証とハイブリッド認証用にHiveへのアクセスに使用されるパスワード。

## Hiveのオプション

- **プレインポートSQL：** インポート開始前に実行する、改行で区切られたSQLステートメントです。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。
- **インポート後のSQL：**インポート処理後に実行されるSQL。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。

> **備考**
>
> インポート前およびインポート後のSQLはインポートプロセス全体で複数回実行される可能性があります。インポートが実行されるたびにこの設定に基づくSQLが実行されるため、これらの値をコネクタ／データソース設定で指定するときは注意が必要です。*

- **エクスポート前のSQL：**エクスポートプロセスの前に実行されるSQL。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。
- **ポスト エクスポートSQL：**エクスポート完了後に実行するSQLステートメントです。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。

# データインポート情報

## ブラウジング経由

- 参照：
  - 区切りデータセット：コンマ、タブ…
  - xml
  - JSON
  - エクセル：XlsおよびXLSX
  - Avro
  - Parquet
  - 固定な形式
  - ファイルを参照し、インポートするファイルを選択します
  - サポートされているデータ形式：
- ワイルドカード：
  - グロブがサポートされています

## SQLクエリー経由

SQL選択クエリの使用

## エクスポート

ブラウザ経由 でインポートの下で一覧表示されているストリームベースの形式の1つを使用してサポートされます。

# Data Prep用のCloudera CDH5 Hiveコネクター

**ユーザーペルソナ**：Data Prep管理者、データソース管理者、またはIT/DevOps

**本機能の提供について**

このコネクターは、Data Prep SaaSのお客様はご利用いただけません。

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Cloudera CDH 5.16 Hiveに接続して、インポートとエクスポートを行うことができます。次のフィールドを使用して、接続パラメーターを定義します。ここで設定する必要があるフィールドは、選択した認証方法 (シンプル、Kerberos、またはハイブリッド) によって異なります。選択した認証の種類は、コネクター設定に基づいて作成するすべてのデータソースに適用されます。

**備考**

このコネクターを設定するには、Data Prepサーバー上のファイルシステムへのアクセスと、Hadoopクラスター設定のcore-site.xmlが必要です。この手順については、カスタマーサクセス担当者にお問い合わせください。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のHiveデータベースに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

## Hadoop クラスター

- **HDFS ユーザー：** HDFS クラスター上のユーザー名。 ファイルを出力して Hive にエクスポートするために使用されます。

## Kerberos認証の構成

Kerberosおよびハイブリッド認証には次のパラメーターが必要です。

- **プリンシパル：** Kerberos認証のプリンシパルです。
- **レルム：** Kerberos認証のレルムです。
- **KDC ホスト名：**Kerberos認証のキー配布センターのホスト名です。
- **Kerberos認証の構成ファイル：**Web サーバー上のKerberos認証構成ファイルの完全修飾パスです。
- **キータブ ファイル：**Web サーバー上のKerberos認証キータブ ファイルの完全修飾パスです。
- **アプリケーション ユーザーの使用：** ログイン中のアプリケーション ユーザーで読み取りまたは書き込みを行う場合は、このチェックボックスをオンにし、プロキシ ユーザーを使用する場合は、このチェックボックスをオフにします。
- **プロキシユーザー：**クラスターでの認証に使用されるプロキシ。${user.name}はプロキシユーザーとして入力できます。${user.name}は[アプリケーションユーザーの使用]の選択と同様に機能しますが、より柔軟性があります。例:
- ユーザーの認証情報にドメインを追加するには、[プロキシユーザー]フィールドに\domain_name\${user.name}と入力します。Data Prepではユーザー名とドメインが渡されます。
  - 例： \Accounts${user.name}はAccountsJoeになります（Joeがユーザー名であると仮定）。
- ユーザー名にテキスト修飾子を適用するには、キー${user.name}に.modifierを追加します。使用できる修飾子は ToLower、ToUpper、ToLowerCase、ToUpperCase、Trim です。
  - たとえば、${user.name.toLowerCase}はJoeをjoeに変換します（Joeがユーザー名であると仮定）。

## Hive の構成

- **JDBC URL：** この URL を Hive へのアクセスに使用して、インポートおよび外部テーブルの登録を行います。Kerberos認証を使用する場合は、次の文字列をURLに追加する必要があります：";auth=kerberos;hive.server2.proxy.user=${user.name}"
  - プロキシユーザーが使用されている場合、文字列${user.name}をプロキシのユーザー名に置き換える必要があります。
- **Hive ファイルの場所：** 外部テーブルの Hive ファイルを格納する HDFS クラスター上の場所を指します。

## 資格情報

- **Hive ユーザー：** シンプル認証で Hive へのアクセスに使用するユーザー名です。

- **Hiveパスワード：**シンプル認証とハイブリッド認証用にHiveへのアクセスに使用されるパスワード。

## Hive のオプション

- **インポー ト前のSQL：** インポート開始前に実行する、改行で区切られた SQL ステートメントです。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。
- **インポート後のSQL：**インポート処理後に実行されるSQL。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。

> **備考**
>
> インポート前およびインポート後のSQLはインポートプロセス全体で複数回実行される可能性があります。インポートが実行されるたびにこの設定に基づくSQLが実行されるため、これらの値をコネクタ／データソース設定で指定するときは注意が必要です。*

- **エクスポート前のSQL：**エクスポートプロセスの前に実行されるSQL。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。
- **エクスポート後のSQL：**エクスポート完了後に実行する SQL ステートメントです。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。

# データインポート情報

## ブラウジング経由

サポートされていません

## SQLクエリー経由

SQL選択クエリの使用

# Data Prep用のCloudera CDH6 Hiveコネクター

**ユーザーペルソナ**：Data Prep管理者、データソース管理者、またはIT/DevOps

**本機能の提供について**

このコネクターは、Data Prep SaaSのお客様はご利用いただけません。

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Hiveデータベースに接続して、インポートとエクスポートを行うことができます。ここで設定する必要があるフィールドは、選択した認証方法 (シンプル、Kerberos、またはハイブリッド）によって異なります。選択した認証の種類は、コネクター設定に基づいて作成するすべてのデータソースに適用されます。

**備考**

このコネクターを設定するには、Data Prepサーバー上のファイルシステムへのアクセスと、Hadoopクラスター設定のcore-site.xmlが必要です。この手順については、カスタマーサクセス担当者にお問い合わせください。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のHiveデータベースに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

## Hadoop クラスター

- **認証方法:** シンプル認証、Kerberos認証、またはハイブリット認証を選択します。選択した認証の種類は、コネクター設定に基づいて作成するすべてのデータソースに適用されます。
- **クラスター コア サイト XML のパス:** Web サーバー上の core-site.xml の完全修飾パスです。例: /path/to/core-site.xml
- **クラスター HDFS サイト XML のパス:** Web サーバー上の hdfs-site.xml の完全修飾パスです。例: /path/to/hdfs-site.xml
- **Native Hadoop ライブラリのパス:** Web サーバー上のネイティブ Hadoop ライブラリの完全修飾パスです。例: /path/to/libraries
- **HDFS ユーザー:** HDFS クラスター上のユーザー名。Hive にエクスポートするためにファイルに書き込む際に使用されます。

## Hive の構成

- **JDBC URL:** この URL を Hive へのアクセスに使用して、インポートおよび外部テーブルの登録を行います。Kerberos認証を使用する場合は、次の文字列をURLに追加する必要があります：";auth=kerberos;hive.server2.proxy.user=${user.name}"
- プロキシユーザーが使用されている場合、文字列${user.name}をプロキシのユーザー名に置き換える必要があります
- **Hive ファイルの場所:** 外部テーブルの Hive ファイルを格納する Hadoop クラスター上の場所を指します。

## Kerberos認証の構成

Kerberosおよびハイブリッド認証には次のパラメーターが必要です。

- **プリンシパル:** Kerberos認証のプリンシパルです。
- **レルム:** Kerberos認証のレルムです。
- **KDC ホスト名:**Kerberos認証のキー配布センターのホスト名です。
- **Kerberos認証の構成ファイル:** Web サーバー上のKerberos認証の構成ファイルの完全修飾パスです。
- **キータブ ファイル:** Web サーバー上のKerberos認証のキータブ ファイルの完全修飾パスです。
- **アプリケーション ユーザーの使用:** ログイン中のアプリケーション ユーザーで読み取りまたは書き込みを行う場合は、このチェックボックスをオンにし、プロキシ ユーザーを使用する場合は、このチェックボックスをオフにします。
- **プロキシユーザー：**クラスターでの認証に使用されるプロキシ。${user.name}はプロキシユーザーとして入力できます。${user.name}は[アプリケーションユーザーの使用]の選択と同様に機能しますが、より柔軟性があります。例:
- ユーザーの認証情報にドメインを追加するには、[プロキシユーザー]フィールドに\domainname\${user.name}と入力します。Data Prepではユーザー名とドメインが渡されます。
  - 例：\Accounts${user.name}はAccountsJoeになります（Joeがユーザー名であると仮定）。
- ユーザー名にテキスト修飾子を適用するには、キー${user.name}に.modifierを追加します。使用できる修飾子は ToLower、ToUpper、ToLowerCase、ToUpperCase、Trim です。
  - たとえば、${user.name.toLowerCase}はJoeをjoeに変換します（Joeがユーザー名であると仮定）。

## 資格情報

- **Hive ユーザー:** シンプル認証で Hive へのアクセスに使用するユーザー名です。
- **Hiveパスワード：**シンプル認証とハイブリッド認証用にHiveへのアクセスに使用されるパスワードです。

## 可視性の設定

インポート中にユーザーがデータ ソースを参照する際に表示されるデータベース、スキーマ、およびテーブルを制御できます。スキーマとテーブルに対しては次の設定を選択できます。

- **[表示のみ]**ここで指定したスキーマまたはテーブルだけが返されます。
- **[非表示]**ここで指定したスキーマおよびテーブルが非表示になります。
- **[すべて表示]**：データソース内のすべてを表示するデフォルト設定です。

[表示のみ]または[非表示]オプションを選択すると、オプションを適用するスキーマまたはテーブルを指定するフィールドが表示されます。

> **備考**
>
> これらの設定は、ユーザーがデータソースに対してクエリーを実行する場合は適用されません。クエリー結果は、一致の完全なリストを返します。たとえば、特定のスキーマを "非表示" と設定しても、クエリの結果には、そのスキーマ内の全テーブルのデータが含まれて返されます。しかし、ユーザーがそのデータ ソースを参照する際には、そのスキーマはユーザーに表示されません。

## インポート設定

- **クエリの プリフェッチ サイズ:** バッチあたりの行数です。
- **行最大サイズ:** インポートおよびエクスポートする値に対して許容する Unicode 形式の最大文字数です。このサイズを超過した値は null に置き換えられます。
- **インポー ト前のSQL:** インポートプロセス前に実行する SQL ステートメントです。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。
- **インポート後のSQL：**インポート処理後に実行されるSQLです。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。

> **備考**
>
> インポート前およびインポート後のSQLはインポートプロセス全体で複数回実行される可能性があります。インポートが実行されるたびにこの設定に基づくSQLが実行されるため、これらの値をコネクタ／データソース設定で指定するときは注意が必要です。*

## エクスポート設定

- **エクスポート前のSQL：**エクスポートプロセスの前に実行されるSQL。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。
- **エクスポート後のSQL:**エクスポート完了後に実行する SQL ステートメントです。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。

# データのインポートとエクスポートに関する情報

## ブラウジング経由

テーブルを参照し、インポートするテーブルを「選択」します。

## SQLクエリー経由

SQL選択クエリの使用

# Cloudera CDH6 Impala Connector for Data Prep

**User Persona**: Data Prep Admin, Data Source Admin, or IT/DevOps

**本機能の提供について**

このコネクターは、Data Prep SaaSのお客様はご利用いただけません。

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Configuring Data Prep

This connector allows you to connect to an Impala database for imports and exports. The fields you are required to set up here depend on the authentication method you select—Simple, Kerberos, or Hybrid. The type of authentication you select will apply to all Data Sources that you create based on a connector configuration.

**Note**

Configuring this Connector requires file system access on the Data Prep Server and a core-site.xml with the Hadoop cluster configuration. Please reach out to your Customer Success representative for assistance with this step.

### General

- **Name:** Name of the data source as it will appear to users in the UI.
- **Description**: Description of the data source as it will appear to users in the UI.

**Tip**

You can connect Data Prep to multiple Impala databases. Using a descriptive name can be a big help to users in identifying the appropriate data source.

## Hadoop Cluster

- **Authentication Method:** Choose between Simple, Kerberos, or Hybrid. The type of authentication you select will apply to all Data Sources that you create based on a connector configuration.
- **Cluster Core Site XML Path:** Fully qualified path of core-site.xml on webserver. Example: /path/to/core-site.xml
- **Cluster HDFS Site XML Path:** Fully qualified path of hdfs-site.xml on webserver. Example: /path/to/hdfs-site.xml
- **Native Hadoop Library Path:** Fully qualified path of native Hadoop libraries on webserver. Example: /path/to/libraries
- **HDFS User:** The username on the HDFS cluster used to write files for export to Impala.

## Impala Configuration

- **JDBC URL:** The URL used to access Impala for import and registration of external tables. If Kerberos authentication is used, the following string must be added to the URL: ";auth=kerberos;impala.server2.proxy.user=${user.name}"
- If a proxy user is used, then the string ${user.name} must be replaced with the proxy username
- **Impala File Location:** The location on the Hadoop cluster used to store Impala files for external tables.

## Kerberos Configuration

The following parameters are required for Kerberos and Hybrid authentication.

- **Principal:** Kerberos Principal.
- **Realm:** Kerberos Realm.
- **KDC Hostname:** Kerberos Key Distribution Center Hostname.
- **Kerberos Configuration File:** Fully-qualified path of Kerberos configuration file on webserver.
- **Keytab File:** Fully-qualified path of Kerberos Keytab File on webserver.
- **Use Application User:** Check this box to read/write as the logged-in application user, or uncheck to use proxy user.
- **Proxy User:** The proxy used to authenticate with the cluster. ${user.name} can be entered as the proxy user. ${user.name} works similar to selecting Use Application User but allows for more flexibility. For example:
- To add a domain to the user's credentials, enter \domain_name\${user.name} in the Proxy User field. Data Prep will pass the username and the domain.
    - Example: \Accounts\${user.name} results in AccountsJoe (assuming Joe is the username).
- To apply a text modifier to the username, add .modifier to the key ${user.name}. The acceptable modifiers are: toLower, toUpper, toLowerCase, toUpperCase, and trim.
    - For example ${user.name.toLowerCase} converts Joe into joe (assuming Joe is the username).

## Credentials

- **Impala User:** The username used to access Impala for Simple and Hybrid authentication.
- **Impala Password:** The password used to access Impala for Simple and Hybrid authentication.

## Visibility Settings

You can control the schemas and tables that are shown to users when they browse a data source during import. For schemas and tables you can choose to:

- **"Show only"** which returns only the schemas or tables that you specify here.
- **"Hide"** which hides the schemas and tables that you specify here.
- **"Show all"** which is the default setting to display everything in the data source.

When you select the "Show only" or "Hide" options, a field is provided for specifying the schemas or tables on which you want the option enforced.

> **Note**
>
> These settings are not enforced when users query against the data source; query results still return a complete list of matches. For example, if you choose to "hide" a specific schema, users can still execute queries that pull data from tables within that schema. However, that schema will not be displayed to users when they browse the data source.

## Import Configuration

- **Query Prefetch Size:** Number of rows per batch.
- **Max Column Size:** The maximum size in Unicode characters allowed for any value for both import and export. Values larger than this will be replaced by null.
- **PRE-IMPORT SQL:** SQL to be executed before import process. This SQL may execute multiple times (for preview and import) and could be multiple SQL statements, newline-delimited.
- **POST-IMPORT SQL:** SQL to be executed after import process. This SQL may execute multiple times (for preview and import) and could be multiple SQL statements, newline-delimited.

> **Note**
>
> As the Pre- and Post-Import SQL may be executed multiple times throughout the import process, please take care when specifying these values in the Connector/Datasource Configuration as they will be executed for every import performed with this configuration.

## Export Configuration

- **PRE-EXPORT SQL:** SQL to be executed before export process. This SQL will execute once and could be multiple SQL statements, newline-delimited.
- **POST-EXPORT SQL:** SQL to be executed after export process. This SQL will execute once and could be multiple SQL statements, newline-delimited.

# Data Import and Export Information

## Via Browsing

Browse to a table and "Select" the table for import.

## Via SQL Query

Using SQL Select queries

# Data Prep用のDatabricksコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Databricksに接続して、ライブラリのインポートとエクスポートを行うことができます。次のフィールドを使用して、接続パラメーターを定義します。AzureとAWSのDatabricksに対して認定されています。

このコネクターにより、参照、クエリ、およびエクスポート操作によるインポートが可能になります。

エクスポート時にDatabricksストレージ（つまり、Databricksサービスプロバイダーに応じてADLS Gen2またはS3バケット）に直接ロードされるデータを除いて、すべてのアクションはJDBC接続を介して実行されます。

### 一般

**名前**：UIでユーザーに表示されるデータソースの名前。

**説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のデータブリックスアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Databricksサーバーの設定

- **Databricksサービスプロバイダー：**接続するDatabricksサービスのタイプに応じて、このプロパティを設定します。AzureとAWSでDatabricksをサポートしています。
  - AzureでのDatabricks
  - AWSでのDatabricks

- **Databricksサーバーの設定タイプ**：Databricksに接続するためのデータソースの設定方法に基づいて、このプロパティを設定します。
  - 基本
  - 詳細設定
- **Databricksサーバー**：Databricksサービスをホストするサーバーのホスト名です。
- **Databricksポート**：Databricksサーバーのポートです。
- **SSLを使用する**：このプロパティにHive設定ファイル（hive-site.xml）の「hive.server2.use.SSL」のプロパティで指定された値を設定します。
- **トランスポートモード**：このプロパティにHive設定ファイル（hive-site.xml）の「hive.server2.trantsports.mode」のプロパティで指定された値を設定します。
- **HTTTPのパス**：HTTPトランスポートモードでは、このプロパティにURLエンドポイントのパスコンポーネントを指定します。このプロパティには、Hive設定ファイル（hive-site.xml）のhive.server2.thrift.http.pathプロパティで指定された値を設定する必要があります。
- **タイムアウト**：操作がタイムアウトするまでの秒数です。ゼロに設定された場合、操作はタイムアウトしません。
- **JDBC Url**：高度な設定で、Databricksに接続するためのJDBC Url全体を設定します。詳細については、CData JDBCドライバーのドキュメントを参照してください。

## Databricksサーバー認証の設定

- **ユーザー**：Databricksサーバーでの認証に使用されるユーザー名です。通常、ユーザー名は「トークン」です。
- **パスワード**：Databricksでの認証に使用されるパーソナルアクセストークンです。個人用アクセストークンは、Databricksインスタンスの[ユーザー設定]ページに移動し、[アクセストークン]タブを選択することで取得できます。

## Databricksログの設定

- **冗長性**：ログファイルに含まれる詳細の量を決定する冗長性レベルです。これは、本番環境の問題をデバッグするのに非常に役立ちます。
- **ログファイル**：Paxサーバー内のドライバーログファイルのパスです。指定されたパスに含まれるすべてのディレクトリが、あらかじめ存在している必要があります

## Databricksサーバーのエクスポートストレージレイヤーの設定

### Azure

- **ADLS Gen2データストアのルートディレクトリ**：このコネクターでアクセスできる明白なルートパスです。'/'を使用して、ADLS Gen2ファイルシステムのルートフォルダ内にDatabricksデータを格納します。
- **ADLS Gen2ストレージアカウント名**：一意のAzure URLのサブドメイン名です。このストレージアカウントは、Databricksクラスターに関連付けられ、アクセスできる必要があります。ADLS Gen2ストレージ アカウント名の長さは3〜24文字にする必要があり、数字と小文字のみを使用できます。ADLS Gen2ストレージアカウント名は、Azure内で一意である必要があります。同じ名前のストレージ アカウントが 2 つ以上存在することはできません。

- **ADLS Gen2ファイルシステム名**：ストレージアカウント内のDatabricksデータを保存するADLS Gen2ファイルシステムの名前です。これは、「コンテナ」 名と呼ばれることもあります。
- **認証タイプ**：ADLS Gen2ストレージに接続する認証のタイプ（「ストレージアカウントアクセスキー」または「Active Directoryユーザー名/パスワード」のいずれか）。
- **ADLS Gen2ストレージアカウントアクセスキー**：フィールドにストレージアカウントアクセスキーを入力します。これは、「共有キー」と呼ばれることもあります。
- **ActiveDirectoryユーザー名/パスワード**：アカウントに関連付けられているAzure Directoryのユーザー名とパスワードを入力します。

> **備考**
>
> マイクロソフトアカウント内でデータを読み書きするには、Data Prepにアクセスを許可する必要があります。そうしないと、接続しようとしたときにエラーが発生します。アクセスを許可するには、コネクター設定ペインで**データソースのテスト**をクリックし、**アクセス許可**のリンクをクリックします。これにより、ログインしてアクセスを許可できるMicrosoftアカウントに移動します。その後でData Prepに戻って続行します。

## AWS

- **S3バケット名**：Amazon S3でDatabricksデータを保存するS3バケットの名前です。このS3バケットは、Databricksクラスターに関連付けられ、アクセスできる必要があります。
- **S3オブジェクトプレフィックス**：このコネクターでアクセスできる明白なルートパスです。「/」を使用して、DatabricksのデータをS3バケットのルートフォルダーに保存します。
- **認証タイプ**：S3バケットにアクセスするための認証方法です。
- **AWS資格情報**：各ユーザーが、ユーザーのAWSアクセスキーに関連付けられているアクセスキーIDとシークレットキーを入力する必要があります。これはデフォルトの設定です。
- **インスタンスプロファイル（IAMロール）**：このテナント内のすべてのユーザーが、個別の認証を必要とせずにAWSにアクセスできるようにします。
- **IAMクロスアカウント**：設定されたS3バケットにアクセスできる別のAWSアカウントのロールを推定し、S3へのアクセスを有効化します。

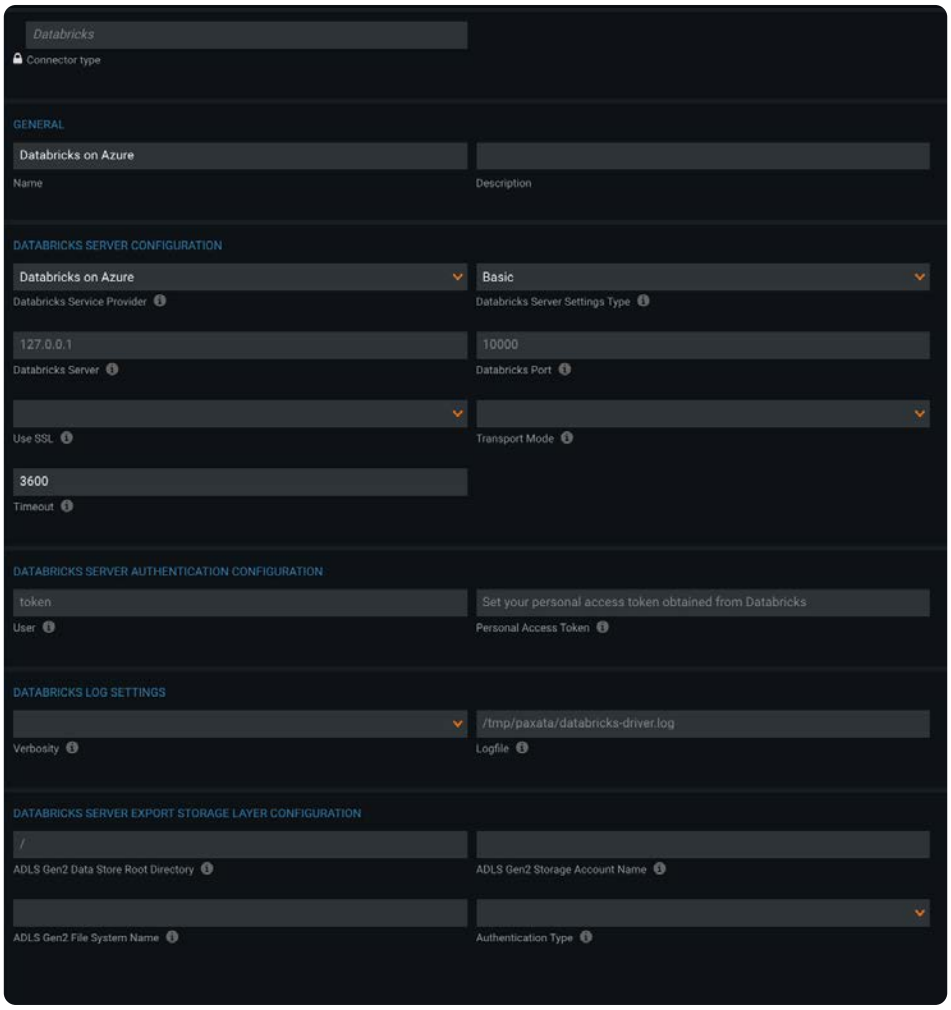
> **重要**
>
> インスタンスプロファイル（IAMロール）およびIAMクロスアカウントオプションの場合、データプレップがアマゾンEC2ホストにインストールされている必要があります。

- **暗号化タイプ**：
  - なし
  - SSE-S3
  - SSE-KMS
- **バケットリージョンロケーター**：S3 AWSバケット領域のロケーターストラテジーです。

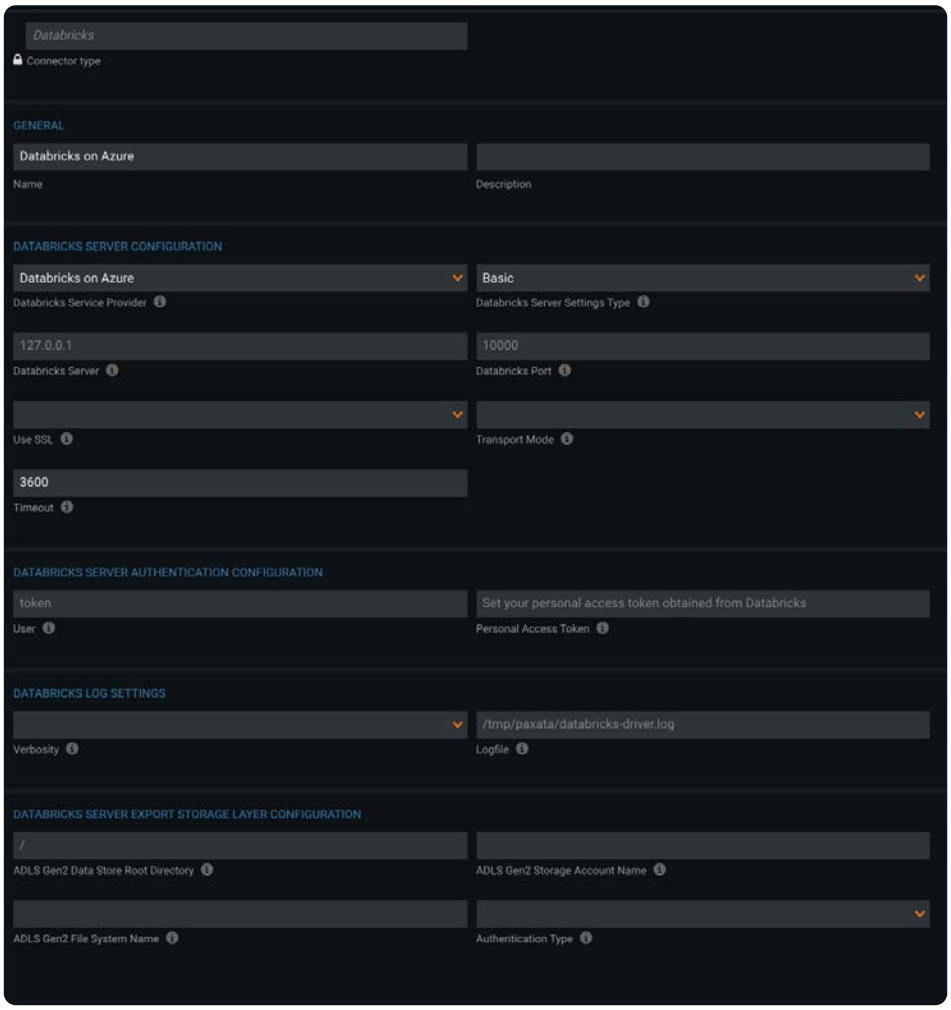
- **ソケットタイムアウトの秒数**：確立された接続でAmazon S3からの応答を待つ秒数です。デフォルト値は5分です。大きなサイズのファイルをエクスポートするには、この値を増やしてください。
- **参照**：
  - 使用可能なデータベースとテーブルのリストを表示します。
- **インポート**：
  - 参照：
    - テーブル（パーティション化および非パーティション化）を参照し、インポートする名前をクリックします。
  - クエリ：
    - 正当なSQL選択クエリの使用
- **エクスポート**：
  - データベースを参照し、テーブルをエクスポートします。

# 設定レイアウト

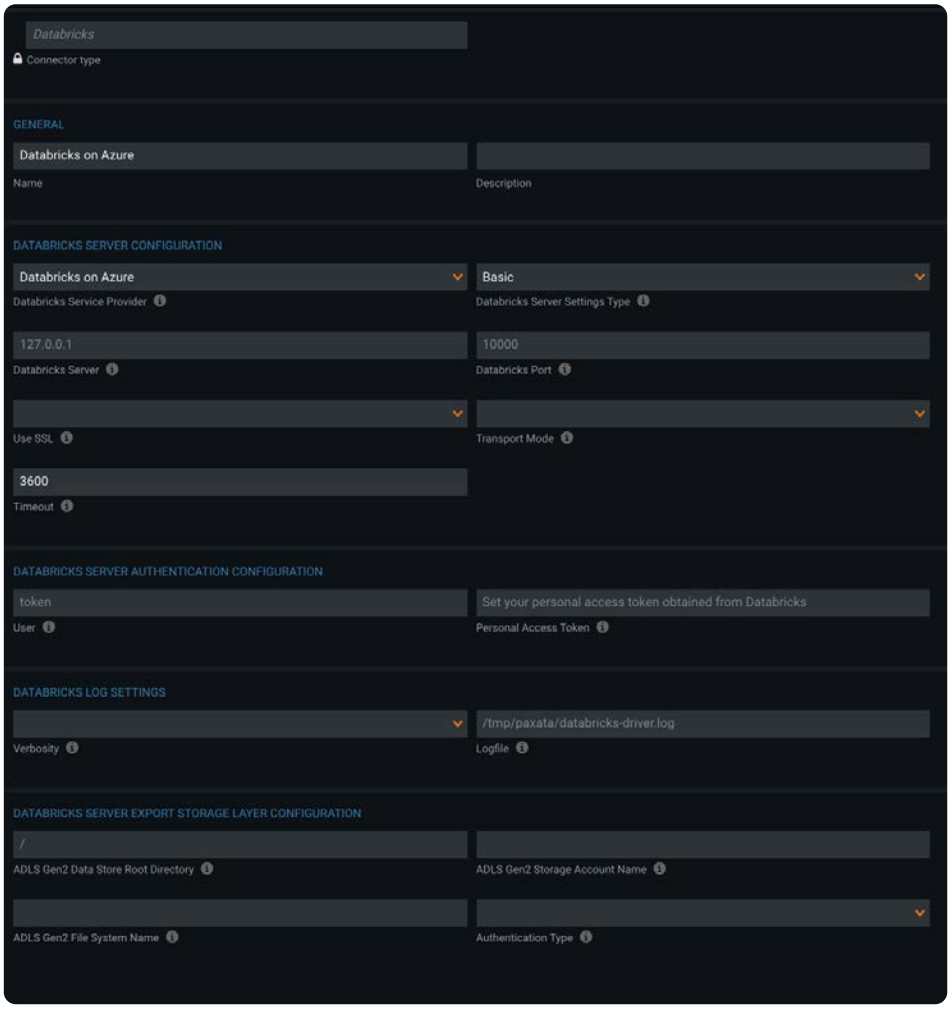
## ADLS Gen2ストレージを使用したAzure上のDatabricks



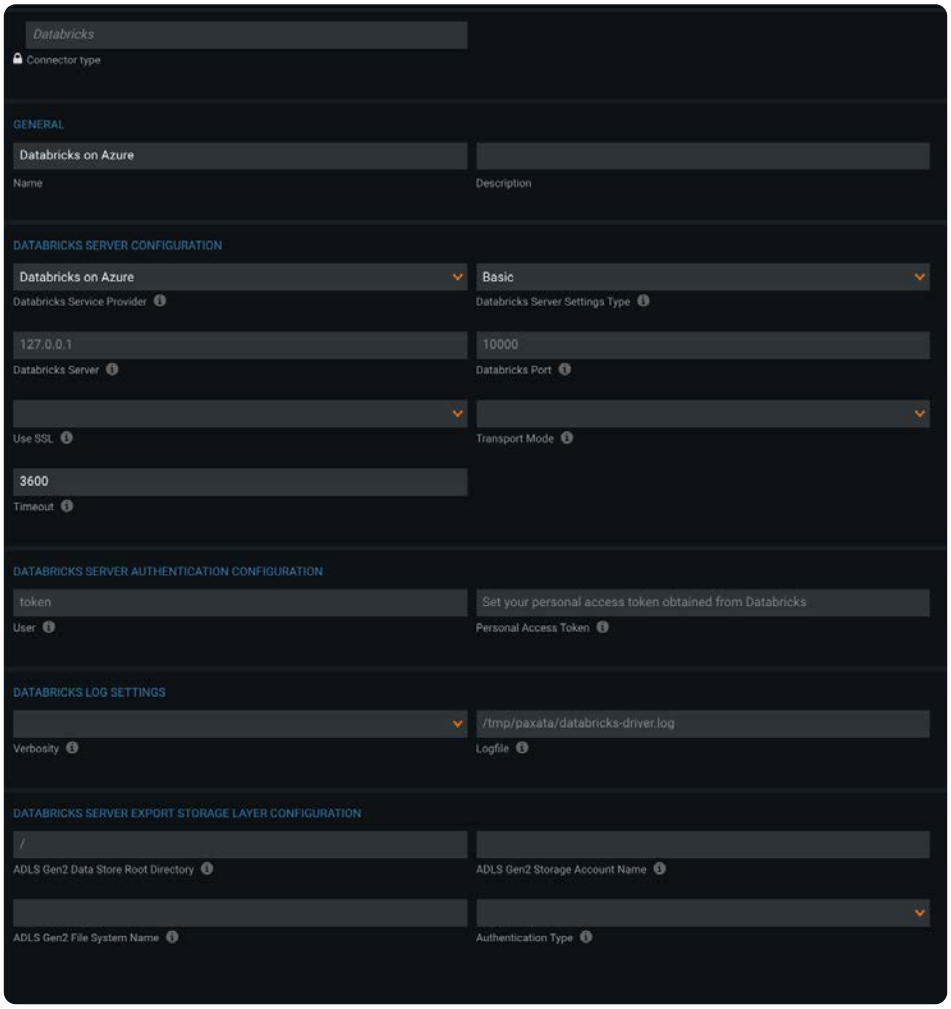
## S3バケットストレージを使用したAWS上のDatabricks



# 参照によるインポート


Databricks
Connector type
GENERAL
Databricks on Azure
Name
Description
DATABRICKS SERVER CONFIGURATION
Databricks on Azure
Databricks Service Provider
Basic
Databricks Server Settings Type
127.0.0.1
Databricks Server
10000
Databricks Port
Use SSL
Transport Mode
3600
Timeout
DATABRICKS SERVER AUTHENTICATION CONFIGURATION
token
User
Set your personal access token obtained from Databricks
Personal Access Token
DATABRICKS LOG SETTINGS
Verbosity
/tmp/paxata/databricks-driver.log
Logfile
DATABRICKS SERVER EXPORT STORAGE LAYER CONFIGURATION
/
ADLS Gen2 Data Store Root Directory
ADLS Gen2 Storage Account Name
ADLS Gen2 File System Name
Authentication Type

## 参照によるエクスポート



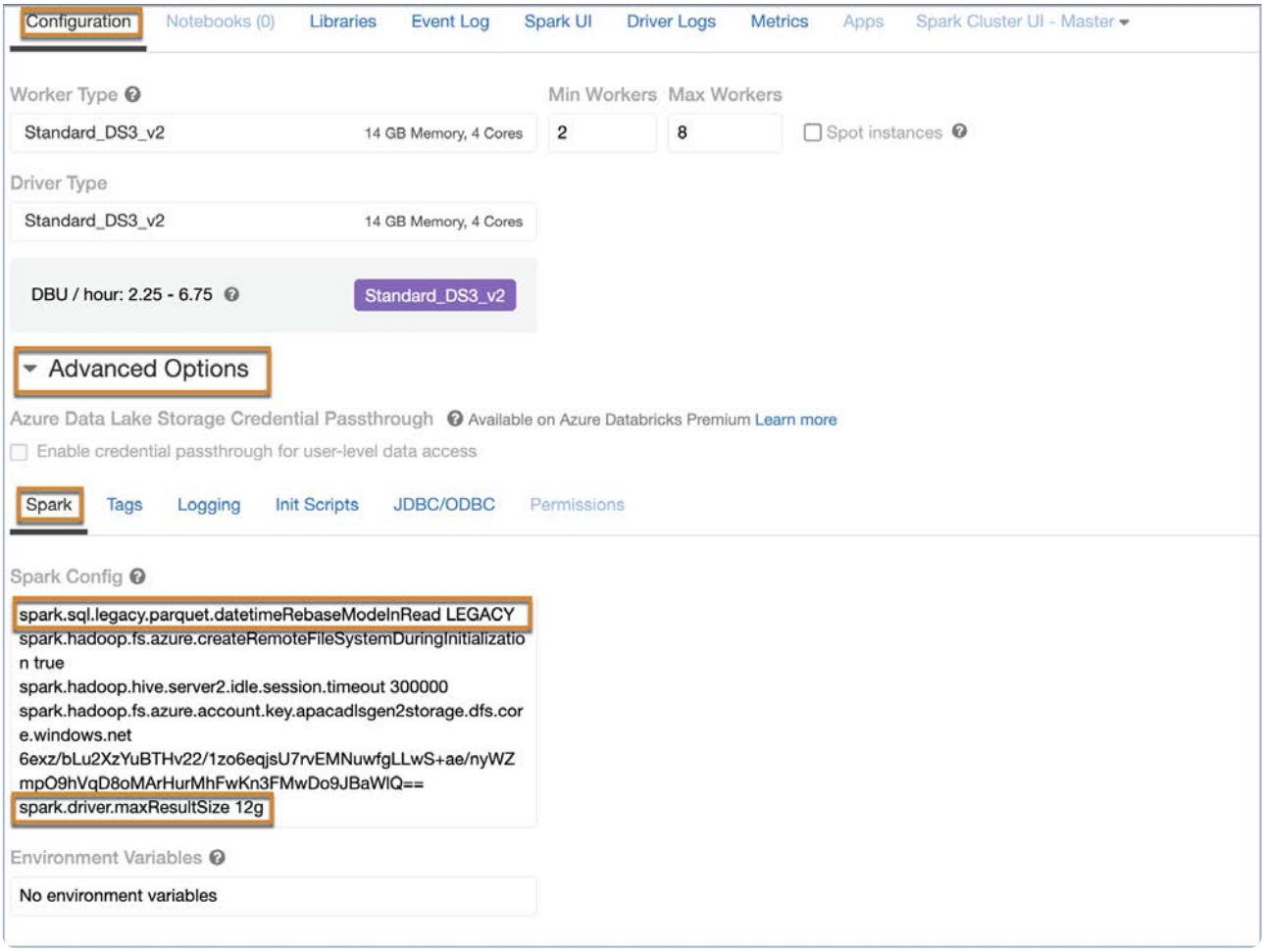

# Databricksクラスターの設定

DataRobotでDatabricksコネクターを設定することに加えて、Sparkの設定をDatabricksクラスターに追加する必要があります。

1. DataBricsクラスターの**設定**タブに移動し、**高度なオプション**を展開します。
2. **Spark**タブで、以下の構成設定を追加および保存します。

   `spark.sql.legacy.parquet.datetimeRebaseModeInRead LEGACY`

   `spark.driver.maxResultSize 12g`




# Databricksコネクターの既知の問題と制限

次の特徴量セットは、一部の運用環境では機能しない場合があります。これらの問題は、今後のリリースで修正される予定です。

- Active Directoryの資格情報を使用したAzure DatabricksインスタンスとADLS GEN2ストレージによる認証。
- クロスアカウントバケットARNを使用したAWS DatabricksインスタンスとAmazon S3サービスストレージによる認証。
- IAMロールが有効になっているAWS DatabricksインスタンスとAmazon S3サービスストレージによる認証。
- 暗号化されていないS3バケット内のSSE-KMSでデータが暗号化されているAWS Databricksインスタンスから、テーブルをインポートします。
- 暗号化されているS3バケット内のSSE-S3およびSSE-KMSでデータが暗号化されているAWS Databricksインスタンスからテーブルをインポートします。
- クロスアカウントとIAMロールで認証した後、AWS Databricksインスタンスからテーブルをインポートします。
- クロスアカウントとIAMロールで認証した後、AWS Databricksインスタンスにエクスポートします。

# Data Prep用のDataRobotコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、DataRobotに接続して、ライブラリのインポートとエクスポートを行うことができます。また、この接続によって、Data Prepから直接DataRobotプロジェクトを作成してアクセスすることもできます。

次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数のDataRobotアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### DataRobotの設定

- **サーバーのURL**：DataRobotのサーバーURL。（https://app.datarobot.comなど）

### 認証設定

- **認証タイプ**：使用する認証タイプを選択：
  - **APIキー**
    - **キー**：DataRobot APIキー

- **ユーザー資格情報**
  - **Eメール：**DataRobotで認証するためのEメールまたはユーザー名。
  - **パスワード：**DataRobotで認証するためのパスワード。
    - **注意：**多要素認証はサポートされていないため、エラーが発生します。

# データのインポート／エクスポート情報

## ブラウジング経由

- **AIカタログ**からインポートするには、**AIカタログ**オプションを選択して使用可能なすべてのデータセットを表示します。目的のデータセットを選択してプレビューを表示し、インポート設定を調整します。

> **備考**
>
> インポート中に、AIカタログからデータセットをダウンロードする権限がないことを示すエラーが表示された場合は、DataRobotで設定を調整する必要があります。DataRobotにログインして、右上の**ユーザーアイコン** → **設定** → **オプション製品** → **AIカタログのダウンロードを有効化** → **保存**をクリックします。その後でDataRobot Data Prepに戻って続行します。

- AIカタログからエクスポートするには、**AIカタログ**オプションを選択してから、**選択する**をクリックします。データセットに名前を付けて**エクスポート**をクリックします。
- データセットをDataRobotに直接エクスポートし、1つのステップでプロジェクトを作成するには、DataRobotプロジェクトの作成を参照してください。

## SQLクエリー経由

サポートされていません

# FAQ／トラブルシューティング／一般的な問題

## DataRobotで生成したモデルをエクスポートし、データが存在する場所でそのコードを実行したい場合はどうすればよいでしょうか？

Data Prepには50を超えるその他のコネクターがあり、準備されたデータを適切な場所に送信できる可能性があります。Data Prepが必要なサービス／ストレージの場所への接続をサポートしていない場合は、カスタマーサクセス担当者までお問い合わせください。

## データセットをインポートできない理由

- 問題1：Data Prepは、DataRobotの**AIカタログ**とのインテグレーションを設計しており、「スナップショット」データセットのインポートのみをサポートします。「スナップショットされていない」データセットに含まれるデータは、実際には

DataRobotに保存されず、使用時に取得されます。Data Prepの場合、DataRobotがデータソースからデータセットを完全にインポートした後、Data Prepがそのデータセットのインポートを開始することを意味します。「スナップショットされていない」データセットの場合は、データソースからData Prepに直接データをプルする方がはるかに効率的です。データセットが「スナップショット」かどうかを判断するには、**AIカタログ**に移動し、問題のデータセットを選択して、右側のパネルの[ステータス]を確認します。

- 問題2：**AIカタログ**からデータセットをダウンロードする権限がないことを示すエラーが表示された場合は、DataRobotで設定を調整する必要があります。DataRobotにログインして、右上の**ユーザーアイコン** > **設定** > **オプション製品** > **AIカタログのダウンロードを有効化** > **保存**をクリックします。その後でDataRobot Data Prepに戻って続行します。
- []問題3：「のマッピングが見つかりませんでした。のいずれかが必要です」というエラーが表示された場合は、DataRobotの設定を調整する必要があります。DataRobotにログインし、右上の**ユーザーアイコン** > **設定** > **CSVエクスポート** > **BOMを含めるのチェックを外す** > **保存**をクリックします。その後でDataRobot Data Prepに戻って続行します。

## データセットの新しいバージョンをエクスポートすると、AIカタログにそのように表示されますか？

はい、同じ名前のデータセットのバージョンが新しいデータセットとしてではなく、**AI カタログ**の_バージョン履歴_タブに表示されます。

## AIカタログへのデータエクスポートの要件

DataRobotにエクスポートされるデータセットは、次の条件を満たす必要があります。

- 少なくとも100行
- 少なくとも2列
- 有効な列名を持っている

# Data Prep用のGoogle Analyticsコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Googleアナリティクスに接続して、利用可能なデータを参照およびインポートできます。次のパラメーターを使用して、接続を設定します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のGoogle Analyticsアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Google Analytics の構成

- **OAuth 検証キー:** Google Analytics との認証に使用する検証キーです。検証キーを取得するには、[データ ソースをテスト] をクリックし、Google Analytics へのアクセスを許可するリンクをクリックします。アクセスを許可すると、アクセスコードを表示するページにリダイレクトされます。このフィールドにコードをコピーします。
- **プロファイル:** 接続先の Google Analytics プロファイルまたはビュー。プロファイルの ID または Web サイト URL を設定します。

## Webプロキシ

Google Analytics に接続するための Web プロキシ オプションを選択します。

- **Webプロキシ：**プロキシが不要な場合は「なし」、プロキシサーバー経由でGoogle アナリティクスに接続する必要がある場合は「プロキシ」。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**プロキシサーバーのホスト名またはIPアドレス。
- **プロキシ ポート:** プロキシ サーバーのポートです。
- **プロキシ ユーザー名とプロキシ パスワード:** 認証されたプロキシ接続のユーザー資格情報です。非認証プロキシ接続では、これらを空白のままにしてください。

# データインポート情報

## ブラウジング経由

- Googleアナリティクスで事前定義されたデータセットのリストを参照してデータセットを選択し、[インポート用に選択]をクリックします。
- 事前定義されたデータセット：
  - アカウント：ユーザーがアクセスできるすべてのアカウントを一覧表示します。
  - AdWords：AdWordsデータを取得します。
  - eコマース：eコマースデータを取得します。
  - イベント：イベントデータを取得します。
  - GoalCompletions：目標完了に関するデータを取得します。
  - プロファイル：ユーザーがアクセスできるすべてのプロファイルを一覧表示します。
  - セグメント：ユーザーがアクセスできるすべてのセグメントを一覧表示します。
  - SiteContent：内部サイトコンテンツデータを取得します。
  - SiteSearch：サイト内検索データを取得します。
  - SiteSpeed：サイト内の速度データを取得します。
  - トラフィック：すべてのトラフィックデータを取得します。
  - Webプロパティ：ユーザーがアクセスできるWebプロパティを一覧表示します。
  - エクスポートはサポートされていません。

## SQLクエリー経由

- SQL選択クエリーの使用
- CData JDBCドライバーのデフォルトの動作は、過去7日間のデータを取得することです。時間ウィンドウをカスタマイズするために、クエリーで直接StartDateとEndDateの値を設定できます。GoogleアナリティクスAPIでサポートされているStartDateとEndDateの入力は、「today」、「yesterday」、「NdaysAgo」（Nは数字）、および正確な日付です。

例:

```
SELECT * FROM Traffic WHERE StartDate=`2020-01-01` AND EndDate=`5daysAgo`
```

詳細については、http://cdn.cdata.com/help/DAE/jdbc/pg_table-sitecontent.htmを参照してください。

# Data Prep用のGoogle BigQueryコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、BigQueryに接続して、利用可能なデータをインポートおよびエクスポートできます。ここで設定する必要があるフィールドは、管理者がどのようにコネクターを設定したかによって異なります。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前です。
- **説明**：UIでユーザーに表示されるデータソースの説明です。

**ヒント**

Data Prepは複数のBigQueryアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### BigQueryの設定

- **OAuth検証キー**：BigQueryでの認証に使用される検証キーです。検証キーを取得するには、"Test Data Source"（データソースをテスト）リンクをクリックして、BigQueryへのアクセスを許可します。アクセスを許可すると、アクセスコードを表示するページにリダイレクトされます。このフィールドにコードをコピーします。
- **プロフィール**：接続するGCPプロジェクトのIDです。
- **テーブルを自動的に作成（オプション）**：有効にした場合、Data Prepは、エクスポートされたデータセットと名前が一致するテーブルを削除し（存在する場合）、エクスポートされたデータセットを使用してテーブルを再作成します。無効にした場合、Data Prepはテーブルがすでに作成されているという想定の下にテーブルへのエクスポートを試行します。

## エクスポート用のGoogle Cloud Storage設定

これらのフィールドは、BigQueryへのエクスポートを実行するために必要です。インポートだけを行う場合は、これらのフィールドを空白のままにすることができます。

> **備考**
>
> 両方を指定するか、どちらも空白のままにする必要があります。

- **Google Cloud Storageバケット名：**エクスポートのステージング領域として使用されるGoogle Cloud Storageバケット名です。
- **Google Cloud Storage JSON Webトークン：**Google Cloud Storageへの接続に使用されるJSON Web Token（JWT）の内容です。

## Webプロキシ

プロキシサーバーを介してBigQueryに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」を選択し、プロキシサーバー経由でBigQueryに接続する必要がある場合は「プロキシ」を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**プロキシサーバーのホスト名またはIPアドレスです。
- **プロキシ ポート:** プロキシ サーバーのポートです。
- **プロキシ ユーザー名とプロキシ パスワード:** 認証されたプロキシ接続のユーザー資格情報です。非認証プロキシ接続では、これらを空白のままにしてください。

# データインポート情報

## ブラウジング経由

- 設定で指定されたプロジェクト内のデータセットとテーブルを表示します。プロジェクトは、ブラウジングビューの最上位ディレクトリとして表示されます。
- データセット内のテーブルを参照し、インポート用テーブルを「選択」します。

## SQLクエリー経由

- SQLのSelectクエリーの使用。

## 使用法

クエリー内の各テーブル名には一重引用符で囲み、ドット区切りは一重引用符の外で行う必要があります。

**有効な構文** `SELECT * FROM `my-project`.`paxata`.`test``

**無効な構文** `SELECT * FROM `my-project.paxata.test``

# Data Prep用のGoogle Cloud Storage コネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Google Cloud Storage（GCS）に接続して、オブジェクトを参照およびインポートできます。以下のフィールドを使用してデータ ソースへの接続を作成します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のGCSアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Google Cloud Storageの設定

- **バケット名**：Google Cloud Storage バケットは、Google Cloud Storage内のオブジェクトの集合を表します。
- **オブジェクトプレフィックス**：プレフィックスは、バケット内のフォルダーまたはサブフォルダーです。バケットで使用する接頭辞を選択します。すべてのオブジェクトを表示するデフォルト値は「/」です。
- **JSON Web Token**：アカウントの認証には、Google Cloud StorageのJSON Web Tokenが必要です。Google Cloud Storageとのセキュリティで保護された接続を確立する JWT ファイルの内容を指定してください。JWT の詳細については、「サーバー間のアプリケーションで OAuth 2.0 を使う」に関するGoogleのドキュメントを参照してください。

## Webプロキシ設定

- プロキシサーバーを介してGoogle Cloud Storageに接続する場合、これらのフィールドでプロキシの詳細を定義します。
  - **Webプロキシ：**プロキシが不要な場合は「なし」、プロキシサーバー経由でGoogle Cloud Storageに接続する必要がある場合は「プロキシ」を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
  - **プロキシホスト：**Web プロキシ サーバーのホスト名または IP アドレスです。
  - **プロキシポート：**データソースのプロキシサーバー上のポートです。
  - **プロキシ ユーザー名：**プロキシ サーバーのユーザー名です。
  - **プロキシパスワード：**プロキシサーバーのパスワードです。 *認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

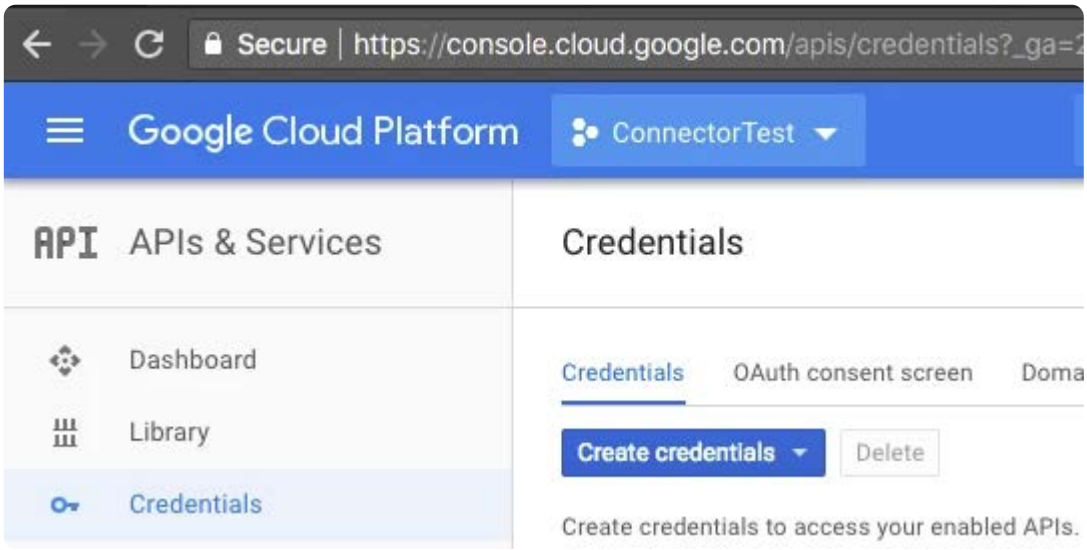
# Googleで認証する方法

Data Prep Google Cloud Storageコネクターは、サービスアカウント認証を利用します。

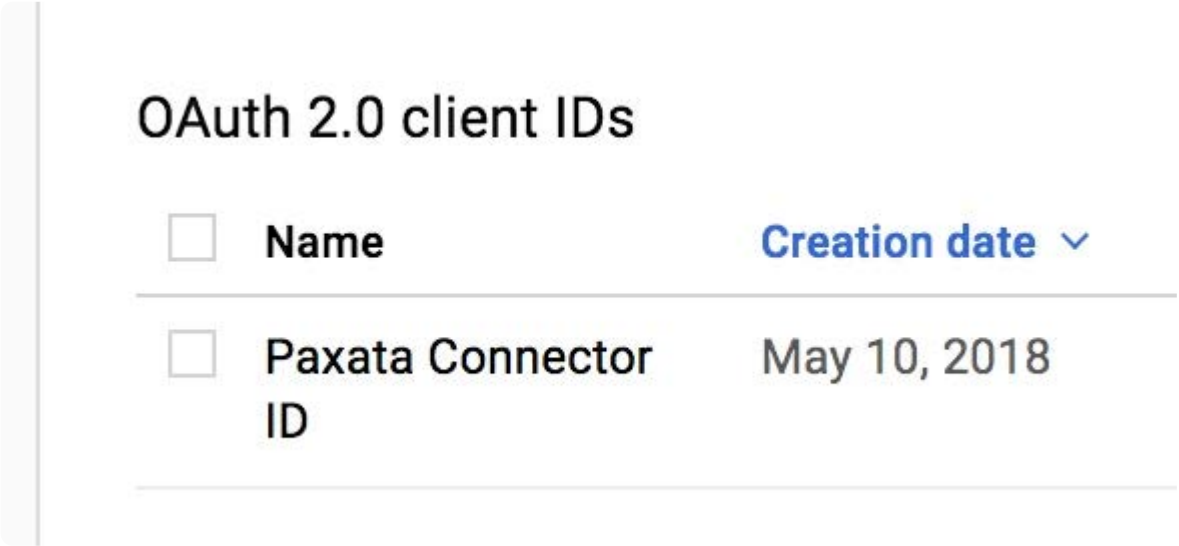
Data Prepを使用してGoogle Cloud Storageにアクセスするには、次のことを行う必要があります。

1. Cloud Storageサービス用のGoogleサービスアカウントを作成します。
   a. Google Cloud Platform Consoleで資格情報のリストを開きます：https://console.cloud.google.com/apis/credentials。
   b. **資格情報を作成する**をクリックします。
   c. **サービスアカウントキー**を選択します。
   d. **サービスアカウントキーを作成する** ウィンドウで、**サービスアカウント**の下のドロップダウンボックスをクリックしてから、**新しいサービスアカウント**をクリックします。
   e. **名前**にサービスアカウントの名前を入力します。
   f. Cloud Storageロールを選択して、サービスアカウントに必要なアクセスレベルを付与します。
   g. デフォルトの**サービスアカウントID**を使用するか、または別のアカウントを生成します。
   h. **キータイプ**：**JSON**を選択します。
   i. **作成**をクリックします。

      **サービスアカウントの作成**画面が表示され、選択した**キータイプ**の秘密鍵が自動的にダウンロードされます。ダウンロードした資格情報の場所を覚えておいてください。 10. **閉じる**をクリックします。 2. Cloud Storageサービス用の既存のサービスアカウントのJSON認証情報をダウンロードします。 1. エンドユーザーアカウントを使用してGoogle Consoleにログインします：https://console.cloud.google.com/apis/credentials。
   j. ドロップダウンリストで正しいプロジェクトが選択されていることを確認してください。

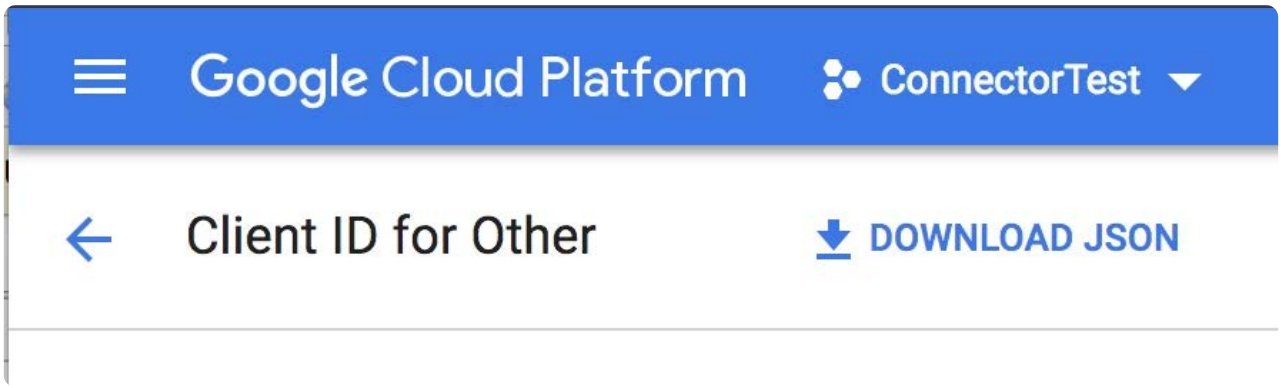



k. 「OAUTH 2.0クライアントID」セクションまでスクロールします。




l. コネクターで使用する予定の既存のID名をクリックします。

m. 表示されたページで、「JSONのダウンロード」リンクをクリックします。




n. ダウンロードした資格情報の場所を覚えておいてください。

追加のリファレンスについては、https://cloud.google.com/storage/docs/authentication#generating-a-private-keyを参照してください。

# データインポート情報

## ブラウジング経由

設定されたバケット／プレフィックス内のディレクトリとファイルを参照します。

# Data Prep用のGoogle Cloud SQLコネクター

**ユーザーペルソナ**：Data PrepユーザーまたはData Prep管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

Data Prep JDBCコネクター（MySQL、PostgreSQL、およびSQLサーバー）を使用してCloud SQLに接続できます。接続を適切に設定するための詳細については、JDBCコネクターのドキュメントとCloud SQLのドキュメントをご覧ください。

# Data Prep用のGoogle Sheetsコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Google Driveに接続して、利用可能なデータを参照およびインポートできます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前です。
- **説明**：UIでユーザーに表示されるデータソースの説明です。

**ヒント**

Data Prepは複数のGoogle Driveアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Googleドライブの設定

- **OAuth検証キー**：Google Driveでの認証に使用される検証キーです。検証キーを取得するには、「Test Data Source」（データソースをテスト）リンクをクリックして、Googleドライブへのアクセスを許可します。アクセスを許可すると、アクセスコードを表示するページにリダイレクトされます。このフィールドにコードをコピーします。

## Webプロキシ

プロキシサーバーを介してGoogleドライブに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」、プロキシサーバー経由でGoogle Driveに接続する必要がある場合は「プロキシ」を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**プロキシサーバーのホスト名またはIPアドレスです。
- **プロキシ ポート:** プロキシ サーバーのポートです。
- **プロキシ ユーザー名とプロキシ パスワード:** 認証されたプロキシ接続のユーザー資格情報です。非認証プロキシ接続では、これらを空白のままにしてください。

# Google Sheetsからのインポート

- スプレッドシートを参照してインポートすると、各Googleスプレッドシートがディレクトリアイテム（つまりフォルダー）として一覧表示され（プレフィックス「GSheet」で識別）、インポートするスプレッドシートごとに個別のデータファイルが見つかります。たとえば、5つの個別のスプレッドシートを含む1つのGoogleスプレッドシートがある場合、この参照インターフェイスを使用して、ワークブック全体を選択するのではなく、インポートする各スプレッドシートを個別に選択します。
- 次の条件に注意することが重要です。
  - インポートするGoogleスプレッドシートの名前に「%」文字を含めることはできません。
  - Google Driveのファイルのサイズ制限に注意してください。

# Data Prep用のGoogle Sheetsコネクター

Googleスプレッドシートコネクターは非推奨になりました。Googleシートのインポート・エクスポートに対応したGoogle Drive Connectorをご利用ください。

# Data Prep用のHortonworks HDP2 HDFSコネクター

**ユーザーペルソナ**：Data Prep管理者、データソース管理者、またはIT/DevOps

**本機能の提供について**

このコネクターは、Data Prep SaaSのお客様はご利用いただけません。

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、インポートおよびエクスポート用にHortonworks HDP 2.6.5 Hadoopファイルシステム（HDFS）に接続できます。次のフィールドを使用して、接続パラメーターを定義します。

**備考**

このコネクターを設定するには、Data Prepサーバー上のファイルシステムへのアクセスと、Hadoopクラスター設定のcore-site.xmlが必要です。この手順については、カスタマーサクセス担当者にお問い合わせください。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

データ準備を複数の HDFS クラスターに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

## シンプル構成（シンプル認証の場合のみ）

- **ユーザー名：**アプリケーションWebサーバーは、ここで指定したユーザー名でHDFSクラスターに接続します。

## 設定

- **データストアのルートディレクトリ：**クラスターの「親ディレクトリ」です。インポートおよびエクスポート操作で、コネクターはこのディレクトリに対して読み書きを行います。ルートのサブディレクトリに対するインポートとエクスポートにも対応しています。

## Kerberos認証の構成

Kerberos認証には、以下のパラメータも設定する必要があります。

- **プリンシパル：** Kerberos認証のプリンシパルです。
- **レルム：** Kerberos認証のレルムです。
- **KDC ホスト名：**Kerberos認証のキー配布センターのホスト名です。
- **Kerberos認証の構成ファイル：**Webサーバー上のKerberos認証の構成ファイルの完全修飾パスです。
- **キータブ ファイル:**Web サーバー上のKerberos認証のキータブ ファイルの完全修飾パスです。

[プロキシユーザー]および[アプリケーションユーザーを使用]オプションを使用すると、インパーソネーションアカウントを指定できます。HDFSでのインパーソネーションの詳細については、このドキュメントを参照してください。ここには3つのオプションがあります：特定のプロキシユーザー、修飾子を持つプロキシユーザー、または個々のアプリケーションユーザーを使用します。

- **プロキシユーザー：**ここで、すべての接続でインパーソネーションされたユーザーアカウントを指定するか、[アプリケーションユーザーの使用]ボックスをオンにして、コネクターを実行する個々のData Prepユーザーのユーザーアカウントをインパーソネーションすることができます。[アプリケーションユーザーの使用]がチェックされている場合、[プロキシユーザー]フィールドは有効になっていないことに注意してください。プロキシユーザーに${user.name}を入力することは、[アプリケーションユーザーの使用]を選択した場合と同様に機能しますが、修飾子や追加のテキストを追加できるため、より柔軟になります。例:
  - ユーザーの認証情報にドメインを追加するには、[プロキシユーザー]フィールドに\domain_name\${user.name}と入力します。Data Prepではユーザー名とドメインが渡されます。
    - 例：\Accounts${user.name}はAccounts\Joeになります（Joeがユーザー名であると仮定）。
  - ユーザー名にテキスト修飾子を適用するには、キー${user.name}に.modifierを追加します。使用できる修飾子はToLower、ToUpper、ToLowerCase、ToUpperCase、Trim です。
    - 例：${user.name.toLowerCase}はJoeをjoeに変換します（Joeがユーザー名であると仮定）。

# データインポート情報

## ブラウジング経由

サポートされています

## SQLクエリー経由

サポートされていません

# Data Prep用のHortonworks HDP2 Hiveコネクター

**ユーザーペルソナ**：Data Prep管理者、データソース管理者、またはIT/DevOps

**本機能の提供について**

このコネクターは、Data Prep SaaSのお客様はご利用いただけません。

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Hortonworks HDP 2.6.5 Hiveに接続して、インポートとエクスポートを行うことができます。次のフィールドを使用して、接続パラメーターを定義します。

**備考**

このコネクターを設定するには、Data Prepサーバー上のファイルシステムへのアクセスと、Hadoopクラスター設定のcore-site.xmlが必要です。この手順については、カスタマーサクセス担当者にお問い合わせください。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のHiveディレクトリに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

## Hadoop クラスター

- **HDFS ユーザー:** HDFS クラスター上のユーザー名。 ファイルを出力して Hive にエクスポートするために使用されます。

## Kerberos認証の構成

Kerberos認証には、以下のパラメータも設定する必要があります。

- **プリンシパル:** Kerberos認証のプリンシパルです。
- **レルム:** Kerberos認証のレルムです。
- **KDC ホスト名:**Kerberos認証のキー配布センターのホスト名です。
- **Kerberos認証の構成ファイル:** Web サーバー上のKerberos認証の構成ファイルの完全修飾パスです。
- **キータブ ファイル:**Web サーバー上のKerberos認証のキータブ ファイルの完全修飾パスです。

[プロキシユーザー]および[アプリケーションユーザーを使用]オプションを使用すると、インパーソネーションアカウントを指定できます。HDFSでのインパーソネーションの詳細については、このドキュメントを参照してください。ここには3つのオプションがあります：特定のプロキシユーザー、修飾子を持つプロキシユーザー、または個々のアプリケーションユーザーを使用します。

- **プロキシユーザー：**ここで、すべての接続でインパーソネーションされたユーザーアカウントを指定するか、[アプリケーションユーザーの使用]ボックスをオンにして、コネクターを実行する個々のData Prepユーザーのユーザーアカウントをインパーソネーションすることができます。[アプリケーションユーザーの使用]がチェックされている場合、[プロキシユーザー]フィールドは有効になっていないことに注意してください。プロキシユーザーに${user.name}を入力することは、[アプリケーションユーザーの使用]を選択した場合と同様に機能しますが、修飾子や追加のテキストを追加できるため、より柔軟になります。例:
  - ユーザーの認証情報にドメインを追加するには、[プロキシユーザー]フィールドに\domain_name${user.name}と入力します。Data Prepではユーザー名とドメインが渡されます。
    - 例：\Accounts${user.name}はAccounts\Joeになります（Joeがユーザー名であると仮定）。
  - ユーザー名にテキスト修飾子を適用するには、キー${user.name}に.modifierを追加します。使用できる修飾子はToLower、ToUpper、ToLowerCase、ToUpperCase、Trim です。
    - 例：${user.name.toLowerCase}はJoeをjoeに変換します（Joeがユーザー名であると仮定）。

## Hive の構成

Hiveコネクターを使用してデータをエクスポートすると、ファイルがHDFSに書き込まれ、Hive JDBCドライバーを介してHiveに外部テーブルが作成されます。[プロキシユーザー]フィールドは、HDFSにファイルを書き込むときにインパーソネーションのユーザーアカウントを指定しますが、Hiveでインパーソネーションを行うには、JDBC URLでもユーザーを指定する必要があります。

- **JDBC URL:** この URL を Hive へのアクセスに使用して、インポートおよび外部テーブルの登録を行います。Kerberos認証を使用する場合は、次の文字列をURLに追加する必要があります：";auth=kerberos;hive.server2.proxy.user=${user.name}。
- プロキシユーザーが使用されている場合、文字列${user.name}をプロキシのユーザー名に置き換える必要があります。

- **Hiveファイルの場所：**外部テーブルのHiveファイルの保存に使用されるHDFS内の場所。

## 資格情報

- **Hive ユーザー:** シンプル認証で Hive へのアクセスに使用するユーザー名です。
- **Hive パスワード:** シンプル認証で Hive へのアクセスに使用するパスワードです。

## Hive のオプション

- **インポー ト前のSQL:** インポート開始前に実行する、改行で区切られた SQL ステートメントです。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。
- **インポート後のSQL：**インポート処理後に実行されるSQL。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。

> **備考**
>
> インポート前およびインポート後のSQLはインポートプロセスを通して複数回実行される可能性があります。インポートが実行されるたびにこの設定に基づくSQLが実行されるため、これらの値をコネクタ／データソース設定で指定するときは注意が必要です。*

- **エクスポート前のSQL：**エクスポートプロセスの前に実行されるSQL。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。
- **エクスポート後のSQL:**エクスポート完了後に実行する SQL ステートメントです。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。

# データインポート情報

## ブラウジング経由

サポートされていません

## SQLクエリー経由

SQL選択クエリの使用

# Data Prep用のHubSpotコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、HubSpotに接続して、利用可能なデータを参照およびインポートできます。次のパラメーターを使用して、接続を設定します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。Data Prep DevOpsにこのセットの希望の旨をお知らせください。

**ヒント**

Data Prepは複数のHubspotアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Webプロキシ

プロキシサーバーを介してHubSpotに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ:** プロキシが不要な場合は [なし]を選択し、経由で HubSpot に接続する必要がある場合は [プロキシ] を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシ ホスト: Web プロキシ** サーバーのホスト名または IP アドレスです。
- **プロキシ ポート:** HubSpot プロキシ サーバーのポート番号です。
- **プロキシ ユーザー名：プロ**キシ サーバーのユーザー名です。
- **プロキシ パスワード：プロキシ** サーバーのパスワード。 *認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

## HubSpot の構成

- **OAuth 検証キー:** HubSpot との認証に使用する検証キーです。検証キーを取得するには、Test Data Source（データソースをテスト）リンクをクリックして、HubSpotへのアクセスを許可します。アクセスを許可すると、http://localhost:33333]（http://localhost:33333/） にリダイレクトされますが、ウェブページは表示されません。URL から 「code」 の URL パラメーター値をコピーします。コネクター との認証に使用する検証キーです。コピーした値を検証キーのフィールドに貼り付けます。

# データインポート情報

## ブラウジング経由

事前定義されたデータセットのリストを表示し、インポートするデータセットを「選択」します。

## SQLクエリー経由

正当なSQL選択クエリを使用します。

# Data Prep用のIBM DB2コネクター

**ユーザーペルソナ**：Data Prep管理者またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

IBM DB2に接続する機能は、Data Prep JDBCコネクターの一部です。このトピックでは、IBM DB2への接続のセットアップに固有の詳細を説明します。接続を構成するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

jdbc:db2://db2._yourdb2instancedomain.com:yourDB2instanceport/Data Prep:ConnectionRetryCount=3;LoginTimeout=10;com.ibm.db2.jcc.DB2BaseDataSource.keepAliveTimeOut=20;'

## 技術的な仕様

### ドライバー仕様

- **IBM DB2データベースドライバーの名前とバージョン：**
  - ドライバーのクラス名：com.ibm.db2.jcc.DB2Driver
  - バージョン：11.5
- **サポートされているIBM DB2データベースのバージョン：**
  - JDBC 3およびJDBC 4標準をサポート - すべてのソフトウェアエディション

### ドライバーのドキュメント

- **一般的なドライバーのドキュメント：** https://www.ibm.com/support/pages/download-initial-version-115-clients-and-drivers

# Data Prep用のIBM Netezzaコネクター

**ユーザーペルソナ**：Data Prep管理者またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

IBM Netezzaに接続する機能は、Data Prep JDBCコネクターの一部です。このトピックでは、IBM Netezzaへの接続のセットアップに固有の詳細を説明します。接続を構成するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

jdbc:netezza://YourNetezzaIPaddress:YourNetezzaPortnumber/MYDB

## 技術的な仕様

### ドライバー仕様

- **IBM Netezzaデータベースドライバーの名前とバージョン：**
  - バージョン：7.2.1.0
- **サポートされているIBM Netezzaデータベースのバージョン：**
  - 7.0.x、7.1.x、7.2.x

### ドライバーのドキュメント

- **一般的なドライバーのドキュメント：**https://www.ibm.com/support/knowledgecenter/SSULQD_7.2.1/com.ibm.nz.datacon.doc/c_datacon_introduction.html
- **バージョンの互換性** https://www.ibm.com/support/knowledgecenter/SSULQD_7.2.1/com.ibm.nz.datacon.doc/c_datacon_release_compatibility_matrix.html

# Data Prep用のJDBCコネクター

**ユーザーペルソナ**：Data Prep管理者またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターは、JDBCドライバーを利用してデータをインポートおよびエクスポートする機能を有効にします。通常、このコネクターはリレーショナルデータベースに対するインポート／エクスポートに利用されますが、多くのアプリケーションでJDBCドライバーを提供しています。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

データ準備を複数のJDBCソースに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### データベースURI

- **JDBC URI：**使用されているドライバーによって定義されたJDBC接続文字列。接続文字列オプションの詳細については、使用するドライバーのドキュメントを参照してください。JDBC接続文字列は通常、次の形式を取ります。

  jdbc://::/:;

## データベースの可視性

インポート中にユーザーがデータソースを参照するときに表示されるデータベース、スキーマ、およびテーブルを制御できます。データベース、スキーマ、およびテーブルの場合、次のいずれかを選択できます。

- [表示のみ]ここで指定したデータベース、スキーマ、またはテーブルだけが返されます。
- [非表示]：ここで指定したデータベース、スキーマ、テーブルが非表示になります。
- [すべて表示]：データソース内のすべてを表示するデフォルト設定です。

[表示のみ]または[非表示]オプションを選択すると、オプションを適用するデータベース、スキーマ、またはテーブルを指定するフィールドが表示されます。

**備考**

これらの設定は、ユーザーがデータソースに対してクエリーを実行する場合は適用されません。クエリー結果は、一致の完全なリストを返します。たとえば、特定のデータベースを[非表示]にした場合でも、ユーザーはそのデータベース内のテーブルからデータをプルするクエリーを実行できます。ただし、そのデータベースは、ユーザーがデータソースを参照するときに表示されません。

## インポート設定

- **クエリープリフェッチサイズ：**インポート中のバッチごとの行数。

**備考**

バッチサイズを大きくすると、大量のインポートのスループットは向上しますが、この値の設定が大きすぎる場合、コネクタープロセスのメモリーが不足する可能性があります。このフィールドのデフォルト値は、バッチごとに10,000行です。コネクタープロセスに付与されるメモリーの量は、Data Prepインストールのサイズによって大きく異なります。このフィールドをデフォルトより大きい値に設定する前に、Data Prep管理者に相談してください。

- **最大列サイズ：**任意の列の最大長（Unicode文字）。これより大きい値は「null」に置き換えられます。
- **プレインポートSQL：** インポート開始前に実行する、改行で区切られたSQLステートメントです。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。
- **インポート後のSQL：**インポート処理後に実行されるSQL。このSQLは（プレビューとインポートのために）複数回実行される可能性があり、改行で区切られた複数のSQLステートメントになることがあります。
- **カウントクエリーを実行：**このセレクターでは、一部のデータベーステーブルで非常に遅くなる可能性のある、インポート時のコネクターによるカウントクエリーの実行を防止できます。インポート中の行数のカウントを無効にするには、これを「False」に設定します。

**備考**

インポート前およびインポート後のSQLはインポートプロセス全体で複数回実行される可能性があります。インポートが実行されるたびにこの設定に基づくSQLが実行されるため、これらの値をコネクタ／データソース設定で指定するときは注意が必要です。

## エクスポート設定

- **エクスポートバッチサイズ：**エクスポート中のバッチごとの行数。

**備考**

バッチサイズを大きくすると、サイズの大きいインポートのスループットが向上しますが、この値の設定が大きすぎる場合、コネクタープロセスでメモリーが不足する可能性があります。このフィールドのデフォルト値は、バッチごとに10,000行です。コネクタープロセスに付与されるメモリーの量は、Data Prepインストールのサイズによって大きく異なります。このフィールドをデフォルトより大きい値に設定する前に、Data Prep管理者に相談してください。

- **最大VARCHARサイズ：** VARCHAR列の最大幅です。

**備考**

このコネクターは、最大VARCHAR幅を超え、データベースがCLOBタイプをサポートしている場合、CLOBタイプを使用して列をエクスポートしようとします。

- **テーブルを自動的に作成：** 有効 | 無効
- 有効：データ準備は、データセットをエクスポートするときに自動的に新しいテーブルを作成します。 テーブルが存在する場合、データ準備は、同じ名前で新しいテーブルを作成する前に、既存のテーブルを削除します。
- 無効：データ準備は、データセットをエクスポートするときに新しいテーブルを自動的に作成しません。 コネクタは、エクスポートされたデータセットの名前と形式に一致するテーブルが存在することを前提としています。エクスポートされたデータは既存のテーブルに追加されます。
- **エクスポート前のSQL：**エクスポートプロセスの前に実行されるSQL。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。
- **ポスト エクスポートSQL：**エクスポート完了後に実行するSQLステートメントです。このSQLは1回実行され、改行で区切られた複数のSQLステートメントになる場合があります。

## 資格情報

- **ユーザー：**データソースへのアクセスに使用するユーザー名。
- **パスワード：**データソースへのアクセスに使用するパスワード。
- **役割：**一部のアプリケーションでは、接続時に役割を指定できます。ここに役割の値を入力します（必要な場合）。このフィールドは空白のままにすることもできます。

# データのインポート／エクスポート情報

## ブラウジング経由でインポート

- 構成設定に基づいて、データベース、スキーマ、および/またはテーブルを参照します。インポートするテーブルを「選択」します。

## SQLクエリー経由でインポート

- データベースのSQL選択クエリーが必要です。
- 例：「SAMPLE_DATA」から選択*。TPCH_SF1"。お客様"

## エクスポート情報

- 構成設定に基づいてデータベースおよび/またはスキーマを参照
- データベースがカタログ、スキーマ、およびテーブルをサポートしている場合、スキーマではなくカタログの下に直接エクスポートしようとすると、エラーが発生する可能性があります。
- (オプション) テーブル名に使用される名前を編集します。

# JDBCのティア2サポート

Data PrepのSaaS以外のお客様は、JDBCコネクターで使用する独自のドライバーを提供およびインストールできます。この機能は現在、SaaSのお客様にはご利用いただけません。

Data Prepでは、特定のJDBCドライバーがData Prepでの使用に適しているかどうかを評価するのに役立つテストキットを用意しています。Data Prepカスタマーサクセスチームがお客様に代わって、このテストキットを実行します。

> **備考**
>
> このツールでの成功は、提供されたJDBCドライバーの公式サポートを保証するものではありません。このツールを使用してテストに合格したドライバーは、Data Prepアプリケーションの現在または将来のバージョンで動作する保証はありません。JDBCテストキットは、包括的なJDBCテストスイートツールではありません。

## 「ティア2」とはどういう意味ですか？

JDBCコネクターテストキットがすべてのテストに合格した場合、Data Prepは次の組み合わせの使用をサポートします。

- 特定のData Prep JDBCコネクターバージョンおよび特定のData Prepコアサーバーバージョン。これらの数値は通常、テストキットのバージョンと一致しますが、常にそうであるとは限りません。
- 特定のデータベース／アプリケーションのバージョン。
- 特定のJDBCドライバーのバージョン。
- テストされたデータ型。

- 正常にテストされた機能のみ-、現在、クエリによるインポート、ブラウズによるインポート、およびエクスポートの3つがあります。

これは他に何を意味しますか：

- Data Prepは、JDBCまたはその他のコネクター接続のデータベース／アプリケーションの認定は行いません。
- Data Prepは、データベース／アプリケーションに対して明示的または正式なテストは行いません。
- Data Prepは、ティア2データソースの潜在的な連続値について、新しいバージョンやサービスパックのテストは行いません。
- Data Prepは、Data Prepクラウドでのティア2JDBCソースの使用には対応できません。Data Prepクラウドでは、認定された（ティア1）JDBCソースのみを使用できます。

## 設定方法を教えてください。

Data Prepで使用したいJDBCドライバーがある場合は、次の手順に従い、JDBCコネクターの下でティア2データソースとして使用してください。

1. まず、カスタマーサクセス担当者に連絡して、使用したいドライバーを提供してください。テストキットを使用してドライバーをテストできます。
2. テストが成功した場合、カスタマーサクセス担当者は、Data Prepコアサーバーの正しいディレクトリにドライバーをインストールし、ドライバーのレジストリへの追加を支援します。
3. そこから、ドライバーのドキュメントと上記の詳細を参照して、新しいドライバーでJDBCコネクターを動作するように設定できます。

# Data Prep用のJDBCコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Jiraに接続して、使用可能なデータのブラウジングやインポートを行うことができます。接続を設定する際には以下のパラメーターを使用します。

### 一般

- **名前：**UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のJiraアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Jira の構成

- **Jira URL：** Jira の URLは「https://your-site-name.atlassian.net」の形式です。
- **ユーザー名：** Jira に接続するユーザーのメールアドレスです。
- **認証タイプ：**使用する認証のタイプ（パスワードまたは API トークン）です。
- **パスワード：**Jira に接続するためのパスワードです。
- **API トークン：**Jira に接続するための API トークンです。これは Cloud Jira のみで使用します。APIトークンの生成については、このJiraのドキュメントを参照してください。
- **タイムアウト：** タイムアウトエラーによって、実行中の操作が取り消されるまでの待機時間（秒単位）です。

## Webプロキシ

プロキシサーバーを介してJiraに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」を選択し、プロキシサーバー経由でJiraに接続する必要がある場合は「プロキシ」を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシ ホスト:**Web プロキシ サーバーのホスト名または IP アドレスです。
- **プロキシサーバー：**データソースのプロキシサーバー上のポート。
- **プロキシユーザー名：**プロキシサーバーのユーザー名です。
- **プロキシ パスワード：**プロキシ サーバーのパスワードです。 *認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

# データインポート情報

## ブラウジング経由

- データセットを参照し、インポートするデータセット名をクリックします。一部のデータセットには、すべてのユーザーがアクセスできない場合があることに注意してください。アカウントが事前定義されたデータセットのいずれかにアクセスできない場合、それをプレビューまたはインポートしようとするとエラーが発生します。

## SQLクエリー経由

- 正当なSQL選択クエリを使用します。

# Data Prep用のMarketoコネクター

**ユーザーペルソナ**：Data PrepユーザーまたはMarketo管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、インポートソースとしてのMarketoに接続できます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のMarketoアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### Marketoの設定

- **Marketo RESTエンドポイント**：Marketo REST APIエンドポイント。
- **OAuthクライアントID**：MarketoにおけるサービスのクライアントID。
- **OAuthクライアントシークレット**：Marketoにおけるサービスのクライアントシークレット。
- **タイムアウト**：タイムアウトエラーによって、実行中の操作が取り消されるまでの待機時間（秒単位）。デフォルトは60秒ですが、ほとんどの場合、これで十分です。

## Webプロキシ

プロキシサーバーを介してMarketo に接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ**：プロキシが不要な場合は「なし」、プロキシサーバー経由でMarketo RESTエンドポイントに接続する場合は「プロキシ」。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト**：Webプロキシサーバーのホスト名またはIPアドレス。
- **プロキシポート**：データソースのプロキシサーバー上のポート。
- **プロキシユーザー名**：プロキシサーバーのユーザー名。
- **プロキシパスワード**：プロキシサーバーのパスワード。

> **ヒント**
>
> 認証されていないプロキシ接続では、ユーザー名とパスワードを空白のままにします。

## Marketoの設定

MarketoコネクタはMarketo RESTAPIとOAuthを活用します。このステップは、Marketo管理者が行う必要があります。

接続を確立するには、Marketo管理者は次のことを行う必要があります。

- Marketoで「カスタムサービス」を作成します。
- 「カスタムサービス」のOAuthクレデンシャルを取得して、Marketoにアクセスしているクライアント（Data Prepコネクタ）を識別します。

### Marketoでの「カスタムサービス」の作成

REST APIを使用してMarketoに接続するには、カスタムサービスが必要です。以下の手順に従って、カスタムサービスを作成します。

1. Marketoアプリケーションの**管理者**領域に移動します。
2. セキュリティセクションの**ユーザーとロール**をクリックします。
3. **ロール**タブを選択し、**新しいロール**をクリックして新しいロールを作成します。
4. **ロールの名前**を入力し、ロールの権限を選択します。**アクセスAPI**権限はREST APIに固有です。
5. APIロールが作成されたので、[ユーザー]タブを選択して**新しいユーザーを招待する**をクリックします。
6. 新しいユーザー情報を入力し、APIアクセスで作成されたばかりのロールを選択します。[APIのみ]オプションを選択して、ユーザーをAPIのみのユーザーとして示すことができます。
7. 新しいユーザーが作成されたので、新しいサービスを作成する必要があります。**LaunchPoint**オプション（**管理 > インテグレーション > LaunchPoint**）をクリックします。
8. **新しいサービス**をクリックします。
9. カスタムサービスタイプを選択し、表示名と説明を入力します。

10. 作成したユーザーを選択します。

### OAuthClientId値とOAuthClientSecret値を取得します

OAuthClientIdとOAuthClientSecretを取得するには、管理者領域の[LaunchPoint]オプションに移動します。目的のサービスの[詳細の表示]リンクをクリックします。認証資格情報を含むウィンドウが表示されます。

### RESTエンドポイントのURLを取得します

RESTエンドポイントは、REST APIセクションの統合->ウェブサービスオプションのMarketo管理領域にあります。Identity Endpointは必要ないことに注意してください。

## データインポート情報

### ブラウジングおよびSQLクエリー経由

表示されるオブジェクトのリストとSQLクエリーの例については、次の表を参照してください。

**MARKETOコネクターのデータオブジェクト：**

Marketoの一部のオブジェクトは、設定に基づいて存在する場合と存在しない場合があり、500列のインポートのMaximum（最大）値に達する場合があります（これはデフォルト設定です。これが不十分な場合は、Data Prepカスタマーサクセスの連絡先に連絡してください）。以下の表には、SQLクエリーの例もいくつかあります。

| 要素 | 説明 |
| --- | --- |
| アクティビティ | Marketo組織のカスタムアクティビティ。<br>・Marketo組織に含まれる各カスタムアクティビティは、独自のオブジェクトとして返されます。<br>・各テーブル名の前には「Activity」が付けられ、その後にカスタムアクティビティの名前が続きます。 |
| ActivityBulkExports | 過去7日間に作成されたアクティビティエクスポートジョブのリストを返します。 |
| Campaigns | Marketo組織のキャンペーン。 |
| Channels | Marketo組織のチャネル。 |
| Companies | Marketo組織の企業。このオブジェクト<br><br>は、ネイティブCRM同期が有効になっていないMarketoサブスクリプションでのみ使用できます。 |

| 要素 | 説明 |
| --- | --- |
| CustomObjects | Marketo組織のカスタムオブジェクト。 |
| Emails | Marketo組織への電子メール。 |
| LeadBulkExports | 過去7日間に作成されたリードエクスポートジョブのリスト。 |
| LeadPartitions | Marketo組織のリードパーティション。 |
| Leads | Marketo組織のリードします。<br><br>組織のMarketo設定に基づいて、Leadオブジェクトには500を超える列が含まれる場合があります。デフォルトのData Prepライブラリ設定を使用して500を超える列をインポートしようとすると、切り捨てエラーが発生する場合があります。 |
| Lists | Marketo組織のリスト。 |
| NamedAccounts | Marketo組織の名前付きアカウント。<br>・このMarketoオブジェクト、「クエリーの作成」オプションを使用したSQLを使用してのみインポートできます。<br>・クエリーには「=」演算子を活用するフィルターを含める*必要があります*。<br>・例:SELECT * FROM NamedAccounts WHERE State='CA' |
| Opportunities | Marketo組織にとっての機会。<br>・このテーブルは、ネイティブCRM同期が有効になっていないMarketoサブスクリプションでのみ使用できます。<br>・クエリーには「=」演算子を活用するフィルターを含める必要があります。<br>・例:SELECT * FROM Opportunities WHERE State='CA' |
| OpportunityRoles | Marketo組織の機会のロール。<br>・このテーブルは、ネイティブCRM同期が有効になっていないMarketoサブスクリプションでのみ使用できます。<br>・クエリーには「=」演算子を活用するフィルターを含める必要があります。<br>・例:SELECT * FROM OpportunityRoles WHERE ExternalOpportunityId='Opportunity1'AND LeadId='1'AND Role='MyRole' |

| 要素 | 説明 |
| --- | --- |
| Programs | Marketo組織向けのプログラム。<br>・すべてのプログラムを参照する場合（SELECT *クエリーの実行など）、タグ列とコスト列は返されません。これらの列は、特定のプログラムIDまたは名前でフィルタリングした場合にのみ返されます。<br>・この場合、タグとコストの列は返されません：SELECT * FROM Programs<br>・この場合、タグ列とコスト列が返されます：SELECT * FROM Programs WHERE Id='1001' |
| SalesPersons | Marketo組織の営業担当者。<br>・このテーブルは、SalesPerson APIが有効になっている場合にのみ使用できます。<br>・会社を取得するときは、フィルターを指定する必要があります。有効なフィルターは、Id、ExternalSalesPersonId、またはEmailを含む検索可能な列です。<br>・例: SELECT * FROM SalesPersons WHERE ExternalSalesPersonId='sales@company.com' |
| タグ | Marketo組織のタグ。 |

# Data Prep用のMicroStrategyコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクタを使用すると、MicroStrategyサーバーに接続して、ライブラリのインポートとエクスポートを行うことができます。

コネクタの作成に使用されるパラメーターに関する情報を以下に示します。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数のマイクロストラテジーアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### MicroStrategyの設定

MicroStrategyの設定セクションで、MicroStrategyサーバーを検出して接続するために使用する情報を入力します。

- **サーバーのホスト名：**ドメイン名を含む完全修飾ホスト名、またはMicroStrategyホストのIPアドレスを使用できます。
- **サーバーポート：**MicroStrategyサーバーのポート番号。
- **SSLを使用：**MicroStrategyサーバーへの接続にSecure Sockets Layerを使用するかどうかを選択します。

## 資格情報

- **ユーザー名とパスワード：**MicroStrategyサーバーへの認証に使用されるユーザー名とパスワードです。
- **認証モード：**MicroStrategyサーバーへの接続に使用される認証モードです。

## エクスポート設定

- **エクスポートバッチサイズ：**MicroStrategyサーバーをエクスポートする際に使用するバッチサイズです。

## Webプロキシ

プロキシサーバーを介してMicroStrategyに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」を選択し、プロキシサーバー経由でMicroStrategyに接続する必要がある場合は「プロキシ」を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**プロキシサーバーのホスト名またはIPアドレス。
- **プロキシ ポート**: プロキシ サーバーのポートです。
- **プロキシ ユーザー名とプロキシ パスワード**: 認証されたプロキシ接続のユーザー資格情報です。非認証プロキシ接続では、これらを空白のままにしてください。

# インポート／エクスポート情報

- MicroStrategyはブールデータ型をサポートしません。したがって、Data Prepからエクスポートする場合、すべてのブール値は文字列（「ture」または「false」）としてエクスポートされます。
- **レポートのエクスポート：**
  - MicroStrategyレポートはインポートできますが、MicroStrategyコネクタはエクスポート時にレポートを更新する機能をサポートしていません。
- **キューブへのエクスポート：**
  - MicroStrategyコネクタは、複数のテーブルを含むキューブへのデータのエクスポートをサポートしていません。
  - MicroStrategyコネクタではREST APIが使用されるため、MicroStrategy Web UI（またはREST API以外の方法）によるエクスポート時のキューブの更新はサポートされません。
  - （Data Prep AnswerSet内の）スキーマがキューブのスキーマと一致しない場合、既存のキューブにエクスポートすることはできません。
  - AnswerSetに列が追加されると、新しい行または更新された行がキューブに追加されますが、新しい列はキューブに追加されません。
  - 依存関係のあるキューブを上書きするには、依存関係のあるキューブへの上書きが許可されるように、MicroStrategyのユーザー権限が設定されていることを確認します。

# Data Prep用のMongoDBコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、MongoDBに接続して、使用可能なデータを閲覧およびインポートできます。以下のフィールドは、接続パラメーターを定義するために使用されます。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のMongoDBデプロイに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### MongoDB の設定

- **サーバー**：MongoDBインスタンスをホストするサーバーのホスト名IPアドレスまたはIPアドレスです。レプリカセットに接続する場合は、いずれかのサーバーのホスト名またはIPアドレスを使用します。
- **ポート**：MongoDBへの接続用のポートです。初期設定のポートは27017です。
- **ユーザー**： MongoDB のユーザーです。
- **パスワード**：MongoDBユーザーのパスワードです。
- **SSL を使用**：このフィールドでSSL を有効化するかどうかを設定します。
- **タイムアウト**：操作がタイムアウトするまでの待ち時間（秒数）です。0に設定された場合、操作はタイムアウトされません。

## データベース タイプの設定

- **データベースタイプ：**スタンドアロンのMongoDBインスタンスまたはレプリカセットに接続します。
- **データベース名：** MongoDB のデータベース名です。スタンドアロンインスタンスに接続する場合は、データベース名が必要です。レプリカセットに接続する際に、このフィールドを空白のままにすると、使用可能なすべてのデータベースがインポートUIに表示されます。
- **レプリカセットに接続する場合：**
  - **レプリカセット：**セカンダリーサーバーのコンマ区切りリスト（server:port）です。これにより、サーバーとポートで設定されたサーバーに加えて、複数のサーバーを指定できます。
  - **読み取り設定：**レプリカセットからの読み取り戦略です。詳細については、読み取り設定を参照してください。

# データインポート情報

## ブラウジング経由

ブラウジングすると、MongoDB内のデータベースとコレクションのリストを表示できます。MongoDBレプリカセットに接続する場合、設定でデータベースが指定されていなければ、すべてのデータベースを参照できます。データベースが指定されている場合は、そのデータベース内のコレクションのみを参照できます。

## SQLクエリー経由

このコネクターは、CDataが提供するドライバーを使用してJDBC上に構築されているため、SQL SELECTクエリーを使用してデータをインポートできます。CDataのドライバードキュメントに記載されているように、クエリーは、MongoDBシェルで使用されるJavaScriptベースのDSLではなく、SQLを使用します。

# ベストプラクティス

- Data Prepは、MongoDBをメタデータストアとして使用します。このコネクターを使用して、Data Prep独自のMongoDBレプリカセットに接続することは、このコネクターの意図または推奨されるユースケースではありません。このように使用された場合は、コネクターおよびData Prep自体のパフォーマンスと正しい機能を保証することはできません。
- Data Prepメタデータを読み取ることを目的としている場合は、Data Prep MongoDBメタデータストア内のデータのバックアップを定期的に作成し、MongoDBの別のインスタンスでバックアップを復元してから、そのインスタンスでコネクターをポイントしてください。

# Data Prep用のMS Azure Data Lake Storage（ADLS）コネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Azure Data Lake Storage（ADLS）に接続して、データのインポートとエクスポートを行うことができます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のAzuate Data Lake Storageアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Azure Data Lake Storageの設定

- **ADL URI**：ADL サイトの URI です。
- **ルート ディレクトリ**：データのインポート/エクスポートが有効化されているディレクトリ構造の最上位を指定します。
- **アプリケーションID**：ADLサイトのアプリケーションIDです。
- **OAUTH 2.0トークンエンドポイント**：ADLサイトのOAUTH 2.0トークンエンドポイントです。
- **アプリケーションアクセスキー値**：ADLサイトのアプリケーションアクセスキー値です。詳細については、FAQ／トラブルシューティング／一般的な問題のセクションの問題1を参照してください。

# FAQ／トラブルシューティング／一般的な問題

## ADLS Gen1とADLS Gen2の両方のコネクターを同じData Prepアカウントに含めることはできますか？

可能です。2つのコネクタは共存でき、相互に干渉しません。

## 問題：テスト接続をすると失敗し、「base64」上の問題があると報告されます。

**修正方法**：2020年3月に、Azureでは**アプリケーションアクセスキー値**の形式が変更されました。新しい形式は認証用には機能しないため、Azureコマンドラインを使用して、Base64でエンコードされたバージョンの**アプリケーションアクセスキー値**を設定する必要があります。

Azure Portalの場合：

1. 新しいアプリケーションアクセスサービスアカウントを作成します。
2. 生成されたアクセスキー値をコピーします。
3. パスワードをBase64でエンコードします。
    - Macの例：echo -n '<パスワード>' | openssl base64
    - Windows：Base64エンコーダーなどのツールを使用します。
4. Base64でエンコードされたバージョンにパスワードをリセットするAzureコマンドを作成します。
    - az ad sp credential reset –name –credential-description "" –append –years 2 -p "" -o=jsonc
5. Azure Portalでコマンドプロンプトを開き、手順4のコマンドを貼り付けます。
6. このサービスアカウントにストレージの適切なACLがあることを確認してください。権限が適切でない場合、ACLエラーが表示されます。
7. Base64でエンコードした新しいパスワードを使用して**アプリケーションアクセスキー値**を設定します。

# Data Prep用のMS Azure Data Lake Storage Gen2（ADLS Gen2）コネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Azure Data Lake Storage Gen2に接続して、インポートとエクスポートを行うことができます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数のAzureデータレイクストレージGen2アカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Azure Data Lake Storage Gen2の設定

- **データストアのルートディレクトリ**：このコネクターでアクセスができる識別可能なルートパス。ファイル システムのすべてのファイルにアクセスするには、"/" を使用してください。
- **Azureストレージアカウント名**：ユニーク数のAzure URLのサブドメイン名。ストレージ アカウント名の長さは 3～24文字である必要があり、数字と小文字のみを使用できます。ストレージ アカウント名は、Azure 内でユニーク数となるものでなければなりません。同じ名前のストレージ アカウントが 2 つ以上存在することはできません。
- **ファイルシステム名**：ストレージアカウント内のファイルシステムの名前。「コンテナー」 名と呼称される場合もあります。

### Azure Data Lake Storage Gen2の認証設定

ドロップダウンから、ADLS Gen2ストレージの優先認証方法を選択し、必須フィールドに入力します。

- **ストレージアカウントのアクセスキー：**フィールドにストレージアカウントのアクセスキーを入力します。これは、「共有キー」と呼ばれることもあります。
- **Active Directoryユーザー名/パスワード：**アカウントに関連付けられているAzure Directoryのユーザー名とパスワードを入力します。

> **備考**
>
> Data PrepがMicrosoftアカウント内のデータを読み書きするためのアクセスを許可する必要があります。許可しない場合、接続しようとするとエラーが発生します。アクセスを許可するには、データソース設定パネルで[データソースのテスト]ボタンをクリックし、[アクセスの許可]のリンクをクリックします。これにより、ログインしてアクセスを許可できるMicrosoftアカウントに移動します。その後でData Prepに戻って続行します。

# データインポート情報

## ブラウジング経由

コネクターは、データストアのルートディレクトリフィールドで定義された場所から始まるブラウズ可能なディレクトリ階層を表示します。

## SQLクエリー経由

サポートされていません

# FAQ／トラブルシューティング／一般的な問題

### ADLS Gen1とADLS Gen2の両方のコネクターを同じData Prepアカウントに含めることはできますか？

可能です。2つのコネクタは共存でき、相互に干渉しません。

# Data Prep用のMS Azure SQLコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

Data PrepJDBCコネクターを使用することで、Microsoft Azure SQLデータベースに接続できます。このトピックでは、Azure SQLデータベースへの接続に向けたセットアップに関する詳細を説明しています。接続の設定に関するガイドラインについては、JDBCコネクタードキュメントも参照してください。

### JDBC URIの例

`jdbc:sqlserver://`
`serverName.database.windows.net:serverPortNumber;encrypt=true;trustServerCertificate=false;hostNameInCertificate=*.database.windows.net;loginTimeout=30;data`

### 多要素認証に対するJDBC URIの例

以下は、Active Directory多要素認証を使用した構成のJDBC URIでの例を示しています。

`jdbc:sqlserver://`
`<serverName>.database.windows.net:<serverPortNumber>;encrypt=true;trustServerCertificate=false;hostNameInCertificate=*.database.windows.net;loginTimeout=30`
詳細については、Active Directory多要素認証の設定を参照してください。

## 技術的な仕様

### ドライバー仕様

Microsoft Azure SQLデータベースドライバーの名前とバージョン：

- ドライバークラス名： `com.microsoft.sqlserver.jdbc.PxMSSQLDriver`
- バージョン： `9.2.1.jre8`

サポートされているMicrosoft Azure SQLデータベースのバージョン：

- Azure SQLデータベース
- Azure SQLマネージドインスタンス（拡張プライベートプレビュー）

# Active Directory多要素認証の設定

MS Azure SQLコネクターは、Active Directory（AD）の多要素認証（MFA）をサポートしており、2つ以上の検証要素を使用してインタラクティブに認証することができます。

MFAをインタラクティブにサポートするには、以下の接続プロパティをJDBC URLに追加します。

`authentication=ActiveDirectoryInterActive`
`clientid=<azure_registered_app's_client_id>`

## 多要素認証アプリの登録

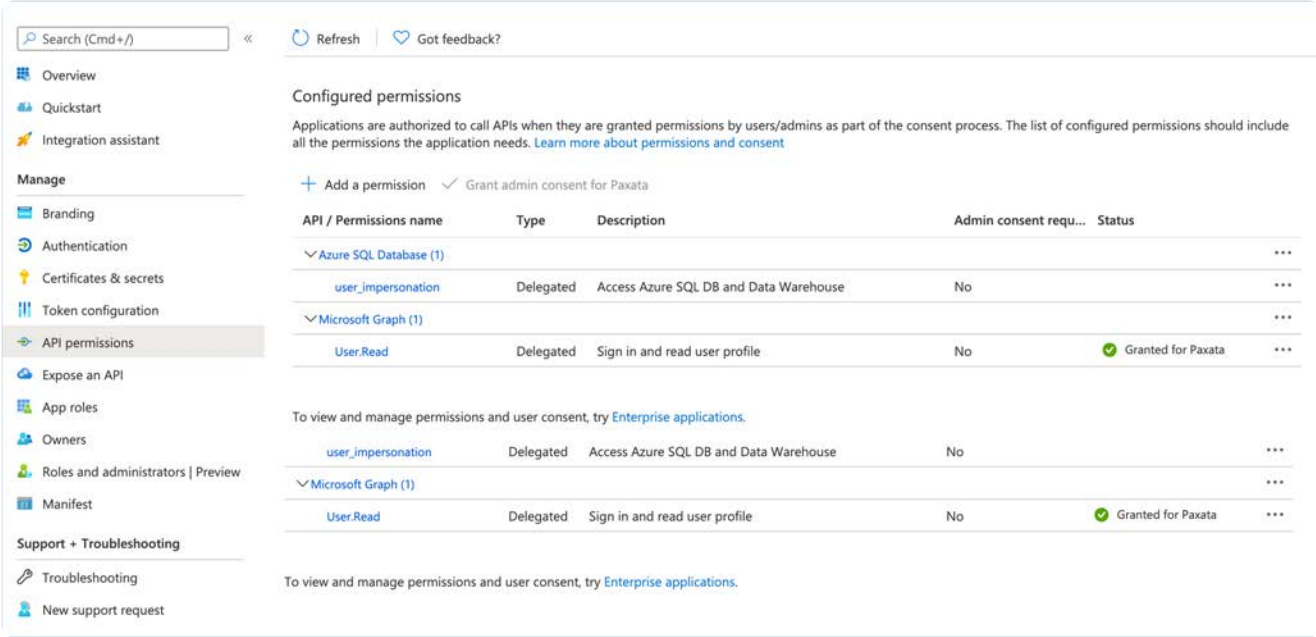
多要素認証を設定するには、Data Prep（Paxata）のMFAアプリケーションとそのユーザーに対して、Microsoft IDプラットフォームが認証および認可サービスを提供できるようにするアプリを登録する必要があります。

> **ヒント**
>
> Microsoft IDプラットフォームでアプリをセットアップするには、以下のガイドラインに従い、クイックスタート：Microsoft IDプラットフォームにアプリケーションを登録するも参照してください。

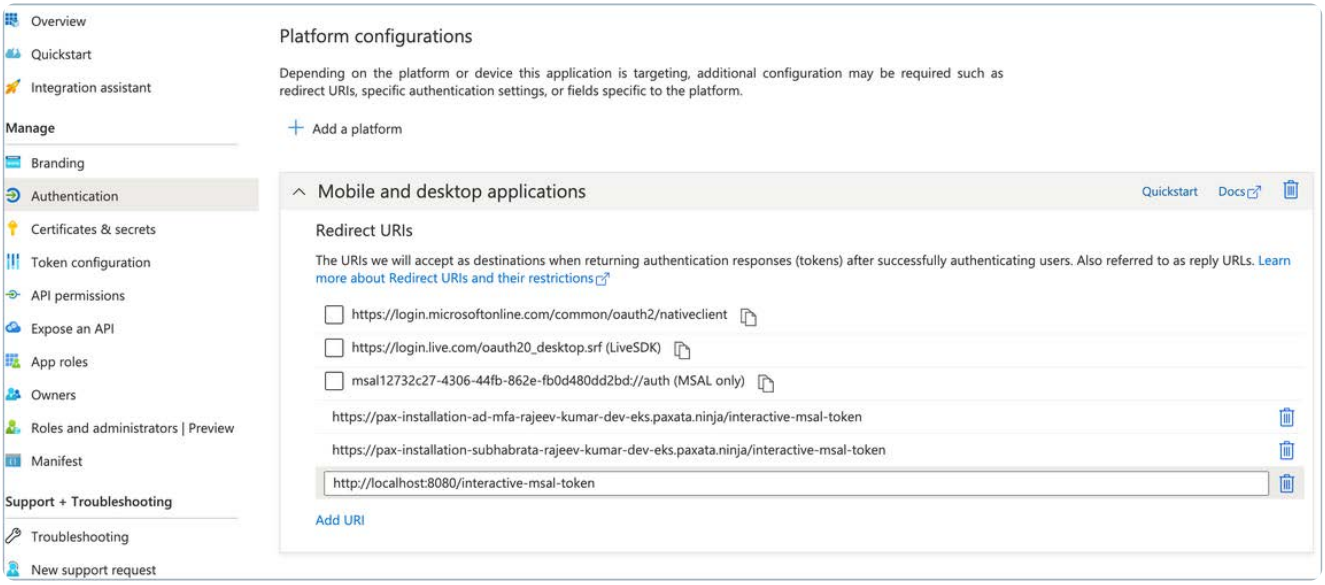
MFAアプリのアプリケーションクライアントIDがJDBC URLに必要となります。これにより、MFAプロセスの完了後、Microsoft IDプラットフォームによってData Prep（Paxata）インスタンスのエンドポイント（[リダイレクトURI形式]セクションで指定）にリダイレクトされます。

Microsoft IDプラットフォーム上でアプリを登録する際には、以下の作業を行う必要があります：

- ADユーザーがアクセス許可を付与したときに、アプリにAzure SQL Database APIを呼び出す権限があることを確認します。Microsoft IDプラットフォーム上での設定例を以下に示します。




- [**プラットフォームの構成**]ページの[**モバイルおよびデスクトップアプリケーション**]で、リダイレクトURIを設定します。Azure SQLデータベースでMFAを完了した後、Microsoft IDプラットフォームはURIを使用して、セキュリティトークンをクライアントからData Prep（Paxata）アプリケーションへとリダイレクトして送信します。




> **備考**
>
> 組織のアカウントを管理するには、インスタンスごとにアプリを登録する必要があります。組織からData Prepにマルチテナントでアクセスできる場合、すべてのテナントに対して単一のMFAアプリを使用できます。この場合、**プラットフォーム構成ページ**上の各テナントに対して、リダイレクトURIを追加する必要があります。

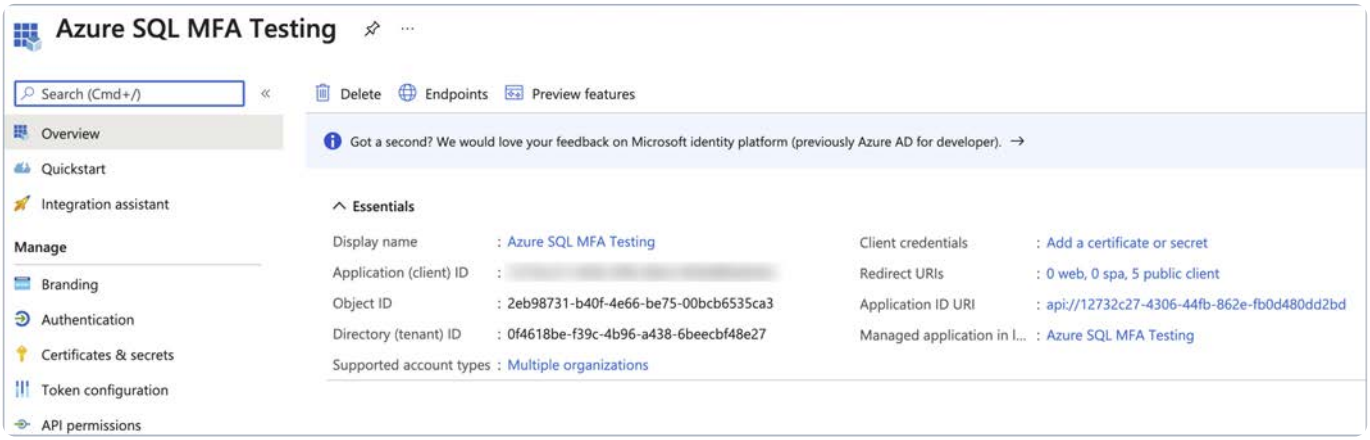
## リダイレクトURIの形式

`http(s)://<paxata_instance_host_name>/interactive-msal-token`

## リダイレクトURIの例

https://datarobot.paxata.com/interactive-msal-token

アプリを設定した後の設定され権限は、以下のようになります。

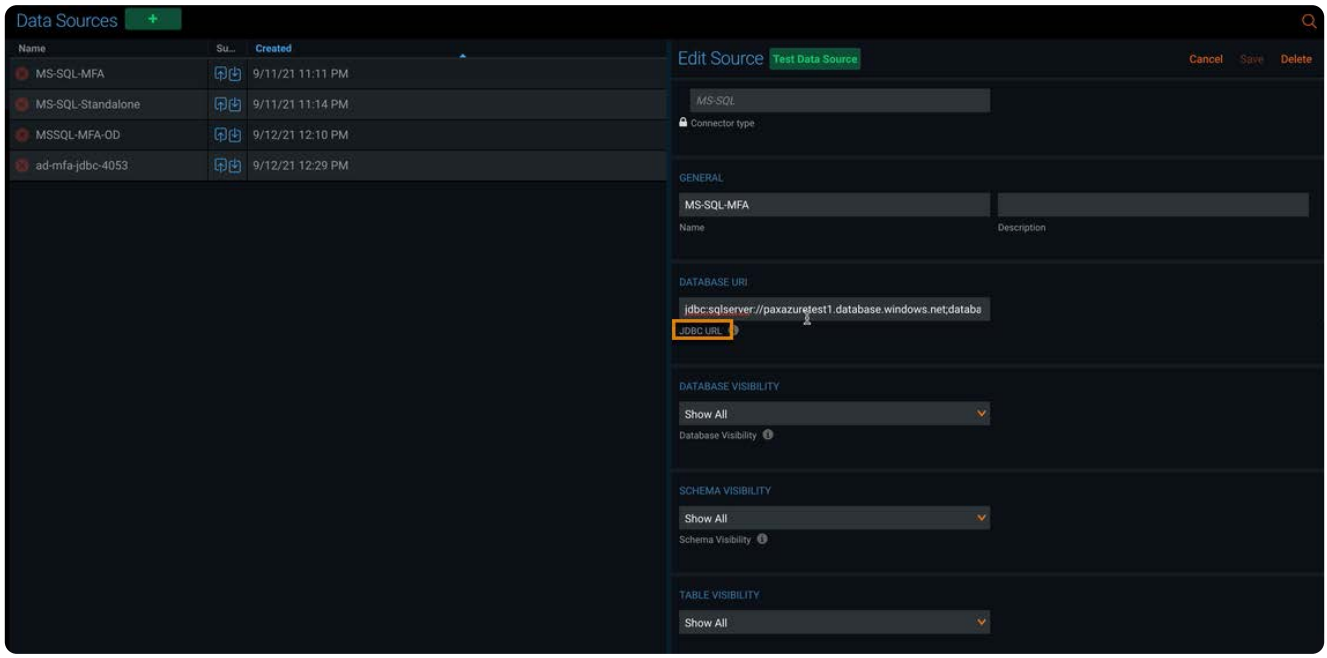



# 多要素認証プロセス

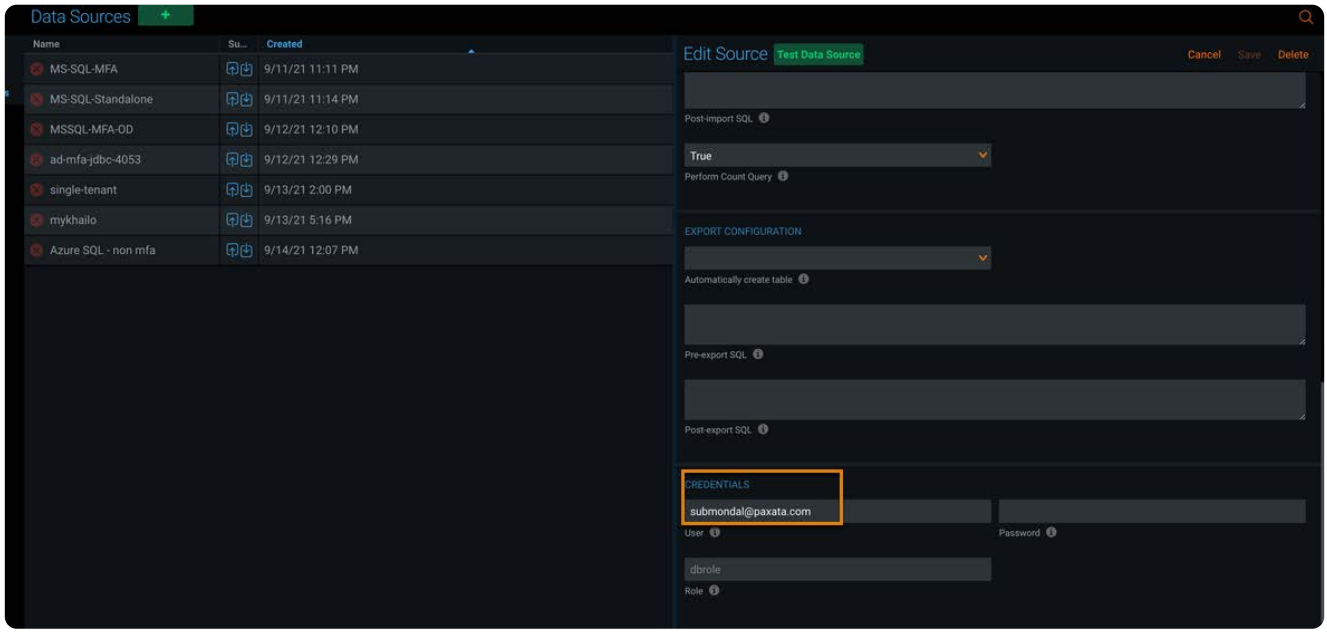
1. 次の設定を使用して、[Data Prep（Paxata）**データソース**]ページでJDBCURLを構成します。

   authentication=ActiveDirectoryInterActive

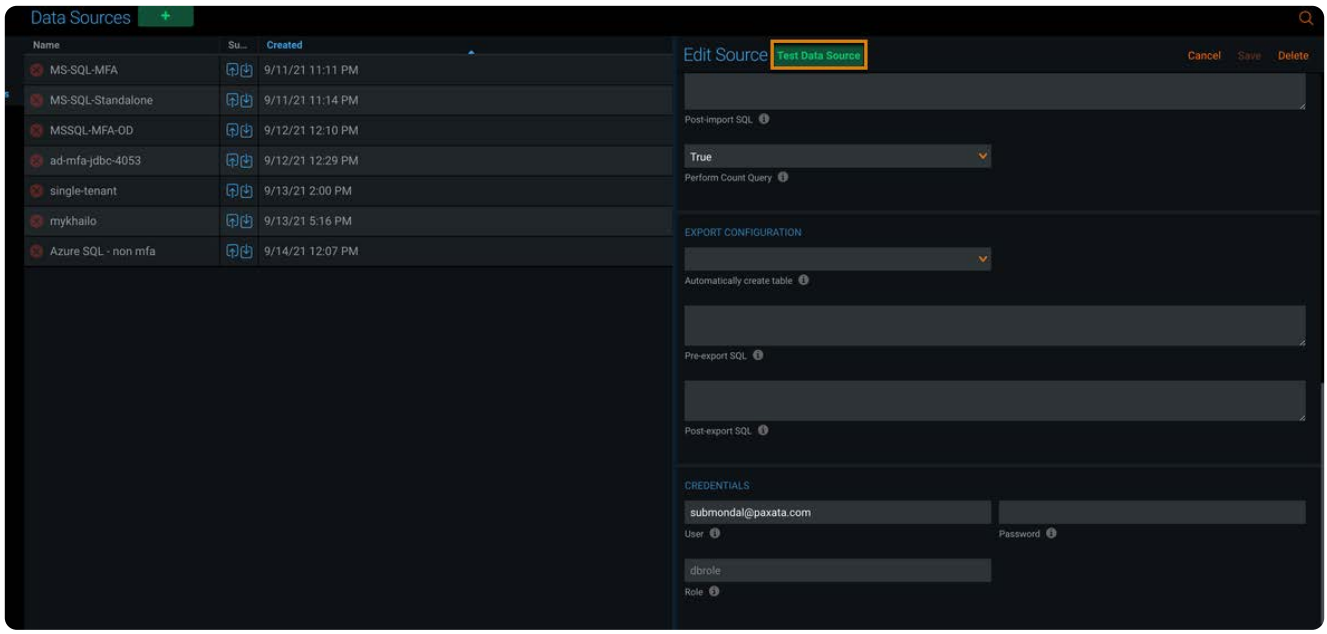
   clientid=<azure_registered_app's_client_id>




2. **資格**情報には、Active Directryユーザー名のみを入力します。パスワードは必要ありません。

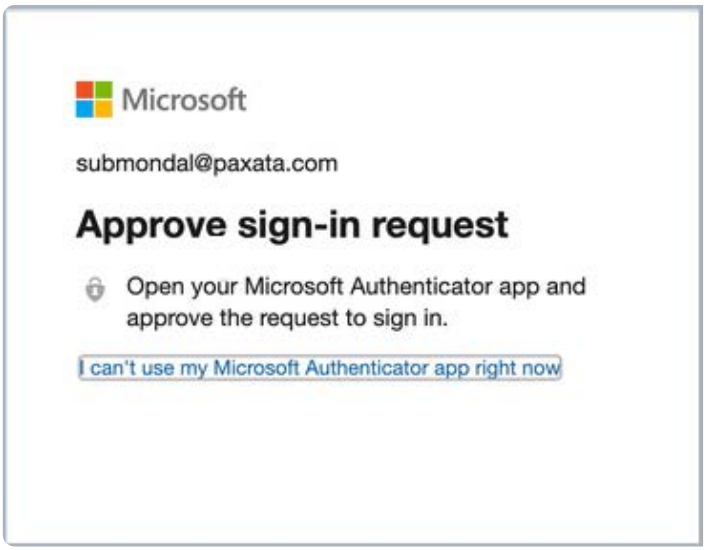



3. **データソースをテスト**をクリックします。



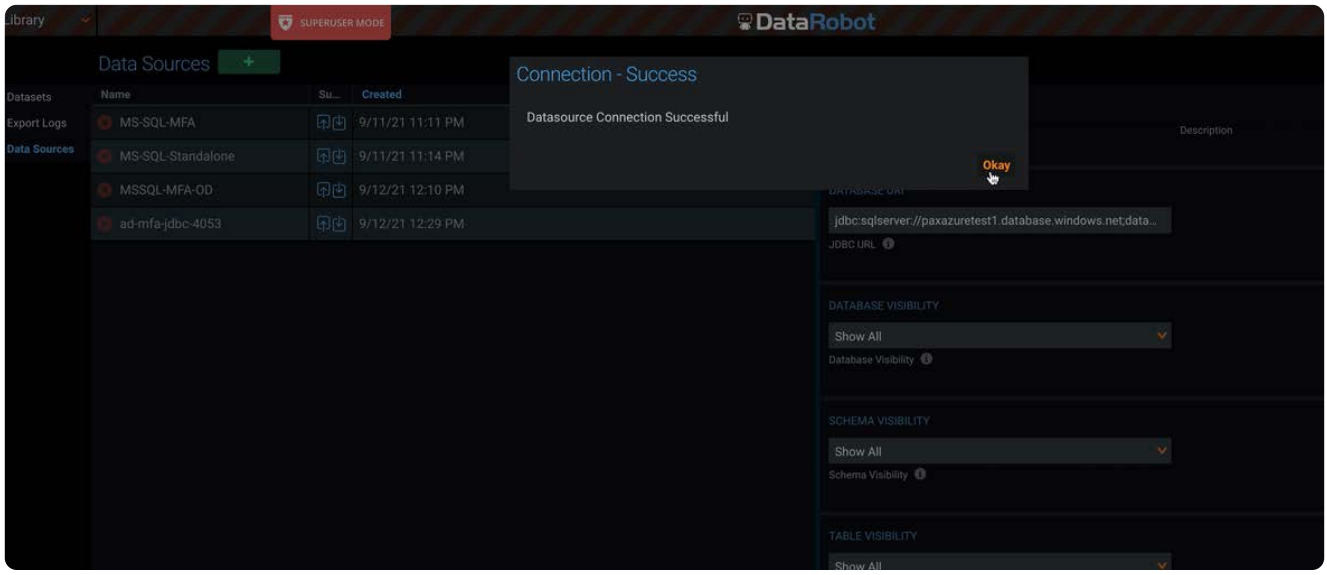

4. ブラウザーに新しいタブが開き、Active Directoryのユーザー名とパスワードが設定されたログイン画面が表示されます。**サインイン**をクリックします。


5. ユーザー名とパスワードが認証されたら、2番目の検証要素とその他に必要となる検証要素を実行します（検証要素はMFAアカウントがどのように設定されているのかによって異なります）。




6. すべての要素が検証された後、Microsoft IDプラットフォームによってData Prepアプリケーションにリダイレクトされます。


7. ウィンドウを閉じ、**データソース**設定ページに戻ります。




**備考**

MFA認証プロセスが2分以内に完了しない場合、「JDBC URLで間違ったクライアントIDが指定されている」、または「認証トークンがまだ受信されていない」、あるいは「要求されたリソースにアクセスできない」などのメッセージが表示されます。

## ドライバーのドキュメント

- 一般的なドライバーのドキュメント： https://docs.microsoft.com/en-us/sql/connect/jdbc/using-the-jdbc-driver
- ドライバーの互換性： https://docs.microsoft.com/en-us/sql/connect/jdbc/microsoft-jdbc-driver-for-sql-server-support-matrix

## 補足ドキュメント

- Azure SQLデータベースへの接続に関する一般情報： https://docs.microsoft.com/en-us/azure/sql-database/sql-database-connect-query-java

# Data Prep用のMS Azure Synapse Analytics コネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Azure Synapse Analyticsに接続して、ライブラリのインポートとエクスポートを行うことができます。エクスポートの場合、コネクターはデータをAzure Data Lakeサービスにアップロードし、SQL Data Warehouseのテーブルとしてデータを公開します。

コネクタの作成に使用されるパラメーターに関する情報を以下に示します。

### 一般

- **名前：**UIでユーザーに表示されるデータソースの名前。
- **説明：**UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは、複数のAzure Synapse Analyticsアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### データベース URL

データベース URL には、Java Database Connectivity（JDBC）の接続文字列を指定します。この文字列は、インポートおよびエクスポートするデータベースの場所をData Prepに通知します。URL にはスキーマの名前を含めることができます。

## 可視性の設定

インポート中にユーザーがデータソースを参照するときに表示されるデータベース、スキーマ、およびテーブルを制御できます。データベース、スキーマ、およびテーブルの場合、次のいずれかを選択できます。

- **[表示のみ]：** ここで指定したデータベース、スキーマ、またはテーブルだけが返されます。
- **[非表示]：** ここで指定したデータベース、スキーマ、テーブルが非表示になります。
- **[すべて表示]：**データソース内のすべてを表示するデフォルト設定です。

[表示のみ]または[非表示]オプションを選択すると、オプションを適用するデータベース、スキーマ、またはテーブルを指定するフィールドが表示されます。

> **備考**
>
> これらの設定は、ユーザーがデータソースに対してクエリーを実行する場合は適用されません。クエリー結果は、一致の完全なリストを返します。たとえば、特定のデータベースを[非表示]にすることを選択した場合、クエリー結果には、返される結果にそのデータベースが含まれます。ただし、そのデータベースは、ユーザーがデータソースを参照するときに表示されません。

## インポート設定

インポート設定では、データをData Prepにインポートする方法を指定できます。

- **クエリーフェッチサイズ：**インポート時にデータを取得する際のバッチごとの行数。
- **インポート前のSQL：** テーブルのスキーマを決定した後、インポートの開始前に実行するSQLステートメント。
- **インポート後のSQL：**インポートの完了後に実行するSQLステートメント。

## エクスポート設定

エクスポート設定では、Data Prepからデータをエクスポートする方法を指定できます。

- **エクスポート前のSQL：**自動作成が有効になっている場合、テーブルの作成後、エクスポートの開始前に実行するSQLステートメント。
- **エクスポート後のSQL：**エクスポートの完了後に実行するSQLステートメント。
- **外部データソース名：** Azure Data Lake 内のデータへのアクセスに使用する SQL Data Warehouse 内の外部データソース の名前。

> **備考**
>
> VARCHARの最大サイズは8000文字です。8000 バイトを超える値は空の文字列としてエクスポートされます。

## 資格情報

資格情報を設定することで、データ ソースへの接続時に認証する単一のユーザー アカウントを指定できます。

## Azure Data Lake の構成

Azure Data Lake設定では、Data PrepがAzure Data Lakeに接続するために必要な設定を指定します。

- **ADL URI：**ADL サイトの URI です。
- **ルートディレクトリ：**データのインポート／エクスポートを有効にするディレクトリ構造の最上位を指定します。
- **アプリケーションID：**ADLサイトのアプリケーションIDです。
- **OAUTH 2.0トークンエンドポイント：**ADLサイトのOAUTH 2.0トークンエンドポイントです。
- **アプリケーションアクセスキー値：**ADLサイトのアプリケーションアクセスキーの値です。

# データインポート情報

## ブラウジング経由

テーブルを参照し、インポートするテーブルを「選択」します。

- サポートされているデータ形式：
- 区切りデータセット：コンマ、タブ…
- xml
- JSON
- エクセル：XlsおよびXLSX
- Avro
- Parquet
- 固定フォーマット
- エクスポート
  - 使用できるエクスポート形式はJDBCの1つだけです。
  - （オプション）Snowflakeテーブル名に使用される名前の編集

## SQLクエリー経由

正当なSQL選択クエリの使用

# Data Prep用のMS Dynamics 365コネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Microsoft Dynamics 365 リソースに接続して、エンティティ セットをインポートできます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前：**UIでユーザーに表示されるデータソースの名前。
- **説明：**UIでユーザーに表示されるデータソースの説明。

**ヒント**

複数のマイクロソフトダイナミクス365リソースへのData Prepを接続することができます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### Microsoft Dynamics 365の設定

このセクションでは、Microsoft Dynamics 365 リソースを識別して接続するための情報を指定します。これらのフィールドは必須です。

- **テナントのドメイン名/ID：**これは、Microsoft Azure Active DirectoryテナントIDまたはドメイン名です。
- **リソースURL：**これは、Microsoft Dynamics365のリソースURLです。

## プロキシ設定

プロキシサーバーを介してMS Dynamics 365に接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」、プロキシサーバー経由でMS Dynamics 365 RESTエンドポイントに接続する場合は「プロキシ」。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Webプロキシサーバーのホスト名またはIPアドレス。
- **プロキシポート：**データソースのプロキシサーバー上のポート。
- **プロキシユーザー名：**プロキシサーバーのユーザー名。
- **プロキシパスワード：**プロキシサーバーのパスワード。*認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空白のままにします。

## 認証設定

このセクションでは、Microsoft Dynamics 365リソースの認証と承認に使用する情報を提供します。これらのフィールドは必須です。

- **クライアントID：**Microsoft Azure Active Directoryに登録されているアプリケーション（Webアプリ/APIまたはネイティブ）のアプリケーションID。アプリケーションの登録方法については、このドキュメントを参照してください。
- **認証タイプ：**アプリケーションがAzure AD にWebアプリ/APIアプリケーションとして登録されている場合は、[クライアント資格情報] を選択します。アプリケーションが Azure AD にネイティブ アプリケーションとして登録されている場合は、[ユーザー資格情報] を選択します。
  - **クライアントシークレット：**[認証タイプ] で [クライアント資格情報] を選択した場合、認証にこれを指定する必要があります。
  - **ユーザー名およびパスワード：**[認証タイプ] で [ユーザー資格情報] を選択した場合、認証にこれらを指定する必要があります。

# データインポート情報

## ブラウジング経由

インポート可能なCRMオブジェクトのリストを表示します。

- リストからオブジェクト名を選択して、インポートを有効にします。
- Dynamicsコネクターは、Dynamics 365 Web APIからのページ付けされた結果を自動的に処理します。

## SQLクエリー経由

サポートされていません

# Data Prep用のMS SharePointコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、SharePoint サイトに接続して、ライブラリのデータをインポートおよびエクスポートできます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前：**UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数のSharepointサイトに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### SharePointの設定

- **SharePoint サーバーの URL:** SharePoint サイトの URLです。
- 特定のSharePointサイト用のコネクターを作成するには、URLに「/sites/」を含めます。例：https://acme.sharepoint.com/sites/sales-department
- 社内のネットワーク上のSharePointの場合のみ、URLにポートを追加する必要がある場合があります。例：http://sp.your-organization.com:8080/sites/department
- **SharePointエディション：**使用されているSharePointのエディション（オンラインまたはオンプレミス）。
  - SharePointオンラインの場合、認証モードとして**シンプル認証**または**シングルサインオン**を選択する必要があります。
  - シングルサインオンの場合は、認証されているユーザーの**SSOドメイン**も指定する必要があります。

- **ユーザー名：**SharePointでの認証に使用される個人または共有アカウントのユーザー名。
- SharePointオンラインの場合、これは通常、メールアドレスの形式です。社内のネットワーク上のSharePointの場合のみ、ドメインにユーザー名を指定する必要がある場合があります。例: Accounts/JDoe
- **パスワード:** SharePoint での認証に使用される共有アカウントのパスワードです。

## Webプロキシ

プロキシサーバーを介してSharePointに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ:** プロキシが不要な場合は [なし]を選択し、経由で SharePoint に接続する必要がある場合は [プロキシ] を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Webプロキシサーバーのホスト名またはIPアドレス。
- **プロキシ ポート:** プロキシ サーバーのポートです。
- **プロキシ ユーザー名：プロ**キシ サーバーのユーザー名です。
- **プロキシ パスワード：プロキシ** サーバーのパスワードです。 *認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

# データインポート情報

## 以下を参照

- コネクターは、ファイルとリストの参照可能なディレクトリ階層を表示します。
- 階層には、サイトの「サイトコンテンツ」ページに表示されるものと同様のデータセットが含まれていると予想できます。

## クエリー

- サポートされていません。

## インポート

- ファイルインポート：サポートされています。
- SharePointリストのインポート：サポートされています。

## エクスポート

- ファイルエクスポート：サポートされています。
- リストのエクスポート：サポートされています。

# FAQ／トラブルシューティング／一般的な問題

## 問題1

- **問題：**インポート後、データのData Prep列にHTMLタグが表示されます。（"

  "、"

  "、""など）。SharePoint はフィールド内に HTML を格納するため、「拡張リッチテキスト」として設定したすべてのフィールドは、このようにインポートされます。SharePointはこのHTMLをリストビューでフォーマットされたテキストとして表示しますが、Data Prepがデータベースから元のテキストを受信する場合、HTMLタグがあり、フォーマットされたテキストではありません。

- **解決策:** Sharepoint で、インポートするリストに移動します。対象列のオプションページに移動します。ここで、許容テキストのタイプを指定できます。これを「プレーンテキスト」に設定します。

## 問題2

- **問題：**インポート後、データのData Prep列に列がありません。
- **解決方法：**SharePoint のリストでは、「Title」列は必須です。これは、構成パラメータが埋め込まれたデフォルトの列です。「Title」列を名前変更すると、この列の表示名とデータベース名とが一致しなくなります。この不一致により、Data Prepは列をインポートできません。次のいずれかのオプションを使用して、Data Prepは列をインポートできます。

  **(a)** SharePointで、列名を「タイトル」にリセットし、再度インポートします。

  **(b)** 別の列名が必要な場合は、SharePointをインポートする前に、データを[タイトル]列から希望の新しい名前の列にコピーし、[タイトル]列を非表示にします。

# Data Prep用のMS SQLサーバーコネクター

**ユーザーペルソナ**：Data Prep管理者またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

Microsoft SQL Serverに接続する機能は、Data Prep JDBCコネクターの一部です。このトピックでは、Microsoft SQLサーバーへの接続のセットアップに固有の詳細を説明します。接続を構成するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

jdbc:sqlserver://cs-mssql1-db1.yourMsSqlHost:yourMssqlPort

## 技術的な仕様

### ドライバー仕様

- Microsoft SQLデータベースドライバーの名前とバージョン：
  - ドライバークラス名：com.microsoft.sqlserver.jdbc.SQLServerDriver
  - バージョン：7.4
- サポートされているMicrosoft SQLデータベースのバージョン：
  - Microsoft SQL Server 2019
  - Microsoft SQL Server 2017
  - Microsoft SQL Server 2016
  - Microsoft SQL Server 2014
  - Microsoft SQL Server 2012

## ドライバーのドキュメント

- 一般的なドライバーのドキュメント：https://docs.microsoft.com/en-us/sql/connect/jdbc/using-the-jdbc-driver
- ドライバーの互換性：https://docs.microsoft.com/en-us/sql/connect/jdbc/microsoft-jdbc-driver-for-sql-server-support-matrix

# Data Prep用のMS Windows Azure Blob Storage (WASB) コネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Azure Blob Storageアカウントに接続して、ライブラリのインポートとエクスポートを行うことができます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前です。
- **説明**：UIでユーザーに表示されるデータソースの説明です。

**ヒント**

Data Prepは複数のAzure Blobアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Azure Blob Storageの設定

- **データストアのルートディレクトリ**：このコネクターがアクセスできるデータストアのルートパスです。コンテナ内のすべてのファイルにアクセスするには、「/」を使用します。
- **Azureストレージアカウント名**：ストレージアカウント名には、小文字と数字を含めることができます。
- **Blobサービスコンテナ名**：コンテナは、ファイルシステムのフォルダーに類似したBlobのセットを整理します。すべてのBlob はコンテナー内に存在します。
- **INT96をDatetimeにマッピングする**：インポート中にINT96フィールドの値を日時値に変換します。具体的には、これにより、Data PrepはImpalaによって書き込まれたParquetファイルを読み取ることができます。

## Azure Blob Storageの認証設定

- **認証タイプ：**2つの認証方法（共有キーと共有アクセス署名）がサポートされます。詳細については、Azureストレージサービスの認証タイプを参照してください。
- **共有キー：**アカウントのアクセスキーで認証します。
- **共有アクセス署名（SAS）：**共有アクセス署名（SAS）で認証します。

# データインポート情報

## ブラウジング経由

サポートされています

## SQLクエリー経由

サポートされていません

# Data Prep用のMySQLコネクター

**ユーザーペルソナ**：Data Prep管理者またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

MySQLに接続する機能は、Data Prep JDBCコネクターの一部です。このトピックでは、MySQLへの接続のセットアップに固有の詳細を説明します。接続を構成するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

jdbc:mysql://yourMysqlSubdomain.yourMysqlDomain.com:yourMysqlPort/yourDatabaseName

## 技術的な仕様

### ドライバー仕様

- **MySQLデータベースドライバーの名前とバージョン：**
  - ドライバーのクラス名：com.mysql.jdbc.Driver
  - バージョン：mysql-connector-java-5.1.39-bin
- **サポートされているIBM DB2データベースのバージョン：**
  - サポートされているバージョンについては、https://dev.mysql.com/doc/connector-j/5.1/en/connector-j-versions.html [dev.mysql.com]を参照してください。

### ドライバーのドキュメント

- **一般的なドライバーのドキュメント：** https://dev.mysql.com/doc/connector-j/5.1/en/ [ibm.com]

# Data Prep用のNetSuiteコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターでは、NetSuiteに接続して使用可能なデータの閲覧とインポートを行うことができます。次のパラメーターが、接続の設定に使用できます。

### 一般

- **名前：**UIでユーザーに表示されるデータソースの名前です。
- **説明：**UIでユーザーに表示されるデータソースの説明です。

**ヒント**

Data Prepは複数のNetsuiteアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### NetSuite の構成

- **NetSuite アカウント ID：** ご使用のNetSuiteアカウントIDです。これを見つけるには、NetSuiteにログインし、[設定]タブ → [インテグレーション] → [Webサービス設定]をクリックします。
- **ユーザー名：** 認証に使用される NetSuite アカウントのユーザー名です。
- **パスワード：**NetSuiteユーザーのパスワードです。
- **ロールID：**NetSuiteへの接続に使用されるロールの内部IDです。ユーザーのデフォルトロールを使用するには、これを空のままにします。ロールIDを見つけるには、アカウントIDを見つけるための上記の手順に従い、ドロップダウンリストからユーザー名を選択し、最後にユーザーアカウントに関連付けられた対応するロールを選択します。
- **タイムアウト：**操作があるまでの待機期間を示す秒数です。デフォルト値は300秒です。

## Webプロキシ設定

プロキシサーバーを介してデータソースに接続する場合、これらのフィールドでプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」を選択し、プロキシサーバー経由でMicroStrategyに接続する必要がある場合は「プロキシ」を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Webプロキシサーバーのホスト名またはIPアドレス。
- **プロキシサーバー：**データソースのプロキシサーバー上のポート。
- **プロキシ ユーザー名：**プロキシ サーバーのユーザー名です。
- **プロキシ パスワード：**プロキシ サーバーのパスワード。

**備考**

認証されていないプロキシ接続では、ユーザー名とパスワードを空白のままにします。

# データインポート情報

## ブラウジング経由

- オブジェクトを参照し、インポートするテーブルを「選択」します。

## SQLクエリー経由

- 正当なSQL選択クエリを使用します。

# Data Prep用のNetwork Share SMBコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、SMB（Server Message Block）プロトコルを使用してネットワーク共有に接続し、インポートおよびエクスポートを実行できます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

複数のSMB共有にData Prepを接続することができます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### 設定

- **共有ホスト名：**はサーバーのホスト名です。
- **共有ポート：**はサーバーのポート番号です。初期設定のポートは445です。
- **共有名：**は共有の名前です。ユーザーがインポートまたはエクスポートに使用する予定の共有の名前。共有名はパスではなく、「\」文字を含めることはできません。スペース文字を使用できます。

## 資格情報

ユーザー認証は、共有アカウントまたは個人アカウントを使用して実行できます。資格情報がデータ ソースを使用して設定されていない場合は、ユーザーは資格情報を入力するよう求められます。

- **ユーザー名：**共有での認証に使用されるユーザー名。
- **パスワード：**共有での認証に使用されるパスワード。
- **ユーザードメイン：**共有に接続するためのユーザーのドメイン。
  - SMBワークグループとも呼ばれます。
  - AD/LDAP管理アカウントでは、これはアカウントが属するADドメインです。
    - ドメインアカウント構造：
  - AD/LDAPで管理されていないアカウントの場合、ドメインは空白にすることができます。

## タイムアウト

- **読み取り/書き込みタイムアウト：**共有との間で読み取り、または書き込みを行う場合のタイムアウト時間（秒単位）。

# データインポート情報

## ブラウジング経由

設定された共有内のディレクトリとファイルを参照します。

## SQLクエリー経由

SMBはファイルストアであるため、このデータソースではSQLクエリーはサポートされていません。

# FAQ／トラブルシューティング／一般的な問題

Data Prepは、LDAPを使用するWindows共有（読み取り専用および読み取り/書き込み）およびローカルサーバーアカウントを使用するLinux共有に対してこのコネクターをテストします。SMBサービスをホストして設定するには多くの方法があり、接続の確立で問題が発生する場合があります。問題が発生した場合は、サーバーの設定ファイルのコピー（機密性の高い値は削除済み）と、サーバーのログ出力のコピーを管理者に依頼してください。

# Data Prep用のOracleコネクター

**ユーザーペルソナ**：Data PrepユーザーまたはData Prep管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

Oracleに接続する機能は、Data Prep JDBCコネクターの一部です。このトピックでは、Oracleデータベースへの接続のセットアップに固有の詳細を説明します。接続を構成するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

jdbc:oracle:thin:yourOracleHost:yourOraclePort:XE

## 技術的な仕様

### ドライバー仕様

- **Oracleデータベースのドライバーのバージョン：**19.3
- **サポートされているOraclデータベースのバージョン：** 12.2.0.1、18c、19c

### ドライバーのドキュメント

- **一般的なドライバーのドキュメント：** https：//www.oracle.com/database/technologies/appdev/jdbc.html
- **ドライバーに関するよくある質問：** https：//www.oracle.com/database/technologies/faq-jdbc.html

# FAQ／トラブルシューティング／一般的な問題

## データ型処理のトラブルシューティング

Oracleデータベースへのエクスポート中に問題が発生する可能性があるまれな項目の1つは、Data PrepとOracleデータベース間でのデータ型の処理方法の違いに関連しています。Data Prepは、Oracleにはない、混合データ型の列を処理するように設計されています。Data Prepは、列に存在する主要なデータ型に基づいて列のデータ型を決定します。これは、Data Prepが列のデータ型がブールであると識別し、データ型が混在していても列をブールとしてエクスポートする場合に問題になります。これは、OracleにエクスポートするData Prepで対処する必要があります。

# Data Prep用のOracle Marketing クラウド（Eloqua）コネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、ライブラリのインポートのためにOracle Marketing クラウドに接続できます。接続を設定するには、以下のパラメータを使用します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

複数のオラクルマーケティングクラウドアカウントへのData Prepを接続することができます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Oracle Marketing クラウドの設定

- **会社**：Oracle Marketing クラウドアカウントが登録されている会社
- **ユーザー**：Oracle Marketing クラウドアカウントのユーザー。
- **パスワード**：Oracle Marketing クラウドユーザーのパスワード。
- **タイムアウト**：操作がタイムアウトするまでの待機時間（秒単位）。デフォルト値は 300 秒です。待機時間を長くするには、この値を増やします。「0」に設定すると、操作はタイムアウトされなくなります。

## Webプロキシ設定

プロキシサーバーを介してOracle Marketing クラウドに接続する場合、これらのフィールドはプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」、プロキシサーバー経由でOracle Marketing クラウドアカウントに接続する必要がある場合は「プロキシ」。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Webプロキシサーバーのホスト名またはIPアドレス。
- **プロキシサーバー：**プロキシサーバー上のポート。
- **プロキシユーザー名：**プロキシサーバーのユーザー名。
- **プロキシ パスワード：**プロキシサーバーのパスワード。*認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空白のままにします。

# データインポート情報

## ブラウジング経由

オラクルマーケティングオブジェクトテーブルご御覧ください。

## SQLクエリー経由

- 有効なSQL選択クエリーの使用
- クエリーの制限：オラクルマーケティングクラウドでは、クエリーの複数の条件に対するサポートが制限されています。
- SINGLE WHERE CLAUSE：多くのテーブルでは、フィルター間のAND条件はサポートされていません。ANDが含まれている場合、サーバーは結果を返さず、エラーもスローしません。
- "="演算子は、WHERE clauseの日付列では使用できません。ただし、 "<"、 ">"、 "> ="、および "<="演算子を使用して、日付列をフィルタリングできます。
- 更新のみが大きい：さらに、で更新列は、「>」演算子によるフィルタリングのみをサポートします。
- これらの制限を示すオブジェクトがオラクルマーケティングオブジェクトテーブルで識別されています。

## Oracle Marketing Objects

| 名前 | 説明 | 容量制限 |
|---|---|---|
| アカウント | アカウントを参照してクエリーを実行します。 | |
| AccountGroup | アカウントグループを参照してクエリーを実行します。 | |
| Campaign | キャンペーンを参照してクエリーを実行します。 | |

| 名前 | 説明 | 容量制限 |
| --- | --- | --- |
| 連絡先 | 連絡先を参照してクエリーを実行します。 | |
| ContactEmailSubscription | 特定の連絡先のすべてのEメールグループのサブスクリプションステータスを参照してクエリーを実行します。 | |
| ContactSegment | 連絡先セグメントを参照して照会します。 | |
| ContentSection | コンテンツセクションを参照してクエリーを実行します。 | |
| カスタム | カスタムを参照してクエリーを実行します。 | |
| Eメール | Eメールを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATはGreater Thanのみ |
| EmailFooter | Eメールのフッターを参照してクエリーを実行します。 | |
| EmailGroup | Eメールグループを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATはGreater Thanのみ |
| EmailHeader | Eメールのヘッダーを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATはGreater Thanのみ |
| イベント | イベントを参照してクエリーを実行します。 | |
| ExternalActivity | 外部活動を参照してクエリーを実行します。 | |

| 名前 | 説明 | 容量制限 |
| --- | --- | --- |
| ExternalAsset | 外部資産を参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATは Greater Thanのみ |
| フォルダー | フォルダーを参照して照会します。 | |
| Form | フォームを参照して照会します。 | |
| Hyperlink | ハイパーリンクを参照してクエリーを実行します。 | |
| LandingPage | ランディングページを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATは Greater Thanのみ |
| Microsite | マイクロサイトを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATは Greater Thanのみ |
| OptionList | オプションリストを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATは Greater Thanのみ |
| AccountField | アカウントフィールドを参照してクエリーを実行します。 | |
| AccountView | アカウント表示を参照してクエリーを実行します。 | |
| アクティビティ_バウンスバック | バウンスバックアクティビティを参照してクエリーを実行します。 | |
| アクティビティ_キャンペーンメンバーシップ | キャンペーンメンバーシップアクティビティを参照してクエリーを実行します。 | |

| 名前 | 説明 | 容量制限 |
| --- | --- | --- |
| アクティビティ_メールクリックスルー | Eメールのクリックスルーアクティビティを参照しクエリーを実行します。 | |
| アクティビティ_メールオープン | Eメールオープンアクティビティを参照してクエリーを実行します。 | |
| アクティビティ_メール送信 | Eメール送信アクティビティを参照してクエリーを実行します。 | |
| アクティビティ _メール購読 | Eメール購読アクティビティを参照してクエリーを実行します。 | |
| アクティビティ_メール購読解除 | Eメール配信停止アクティビティを参照してクエリーを実行します。 | |
| アクティビティ_フォーム送信 | フォーム送信アクティビティを参照してクエリーを実行します。 | |
| アクティビティ_ページビュー | ページ表示アクティビティを参照してクエリーを実行します。 | |
| アクティビティ_ウェブ訪問 | Web訪問アクティビティを参照してクエリーを実行します。 | |
| CampaignElement | キャンペーン要素を参照してクエリーを実行します。 | |
| CampaignField | キャンペーンフィールドを参照してクエリーを実行します。 | |
| CampaignFolder | キャンペーンフォルダーを参照してクエリーを実行します。 | |
| ContactField | 連絡先フィールドを参照してクエリーを実行します。 | |
| ContactFilter | 連絡先フィルターを参照してクエリーを実行します。 | |

| 名前 | 説明 | 容量制限 |
| --- | --- | --- |
| ContactFilterFolder | 連絡先フィルターフォルダーを参照してクエリーを実行します。 | |
| ContactList | 連絡先リストを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATは<br>Greater Thanのみ |
| ContactListFolder | 連絡先リストフォルダーを参照してクエリーを実行します。 | |
| ContactScoringModelFolder | 連絡先スコアリングモデルフォルダーを参照してクエリーを実行します。 | |
| ContactSegmentData | Eloquaコンタクトセグメントデータを参照してクエリーを実行します。 | |
| ContactSegmentFolder | 連絡先セグメントフォルダーを参照してクエリーを実行します。 | |
| ContactView | 連絡先表示を参照してクエリーを実行します。 | |
| ContentSectionFolder | コンテンツセクションフォルダーを参照してクエリーを実行します。 | |
| 依存関係 | 依存関係を参照してクエリーを実行します。 | |
| DynamicContent | 動的コンテンツを参照してクエリーを実行します。 | |
| DynamicContentFolder | 動的コンテンツフォルダーを参照してクエリーを実行します。 | |
| EmailDeployment | Eメールのデプロイを参照してクエリーを実行します。 | |

| 名前 | 説明 | 容量制限 |
| --- | --- | --- |
| EmailFolder | Eメールフォルダーを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATは<br>Greater Thanのみ |
| EmailFooterFolder | Eメールフッターフォルダーを参照してクエリーを実行します。 | |
| EmailHeaderFolder | Eメールヘッダーフォルダーを参照してクエリーを実行します。 | |
| ExternalType | 外部型を参照してクエリーを実行します。 | |
| FieldMerge | フィールドマージを参照してクエリーを実行します。 | |
| FieldMergeFolder | フィールドマージフォルダーを参照してクエリーを実行します。 | |
| FormElement | フォーム要素を参照してクエリーを実行します。 | |
| FormFolder | フォームフォルダーを参照してクエリーを実行します。 | |
| FormProcessingStep | フォーム処理ステップを参照してクエリーを実行します。 | |
| HyperlinkFolder | ハイパーリンクフォルダーを参照してクエリーを実行します。 | |
| イメージ | 画像を参照してクエリーを実行します。 | |
| ImageFolder | 画像フォルダーを参照してクエリーを実行します。 | |
| ImportedFile | インポートされたファイルを参照してクエリーを実行します。 | |
| ImportedFileFolder | インポートされたファイルフォルダーを参照してクエリーを実行します。 | |

| 名前 | 説明 | 容量制限 |
| --- | --- | --- |
| LandingPageFolder | ランディングページフォルダーを参照してクエリーを実行します。 | |
| LeadScoringModel | クエリリードスコアリングモデルを参照してクエリーを実行します。 | |
| PageTag | ページタグを参照してクエリーを実行します。 | |
| PageTagGroup | ページタググループを参照してクエリーを実行します。 | |
| Program | プログラムを参照およびクエリーを実行します。 | |
| Style | スタイルを参照してクエリーを実行します。 | |
| テンプレート | テンプレートを参照してクエリーを実行します。 | |
| TemplateCategory | テンプレートカテゴリを参照してクエリーを実行します。 | |
| TrackedUrl | 追跡されたURLを参照してクエリーを実行します。 | |
| ユーザー | ユーザーを参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATはGreater Thanのみ |
| Visitor | 訪問者を参照してクエリーを実行します。 | 単一のWHERE句<br>日時は同等ではありません<br>UPDATEDATはGreater Thanのみ |
| VisitorProfileField | すべての訪問者プロファイルフィールドを参照してクエリーを実行します。 | |

# 技術的な仕様

## Bulk API

このコネクターは、可能な場合、Bulk APIの使用を自動的にネゴシエートします。

# Data Prep用のPostgreSQLコネクター

**ユーザーペルソナ**：Data PrepユーザーまたはData Prep管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

PostgreSQLに接続する機能は、Data Prep JDBCコネクターの一部です。この記事では、PostgreSQLへの接続を設定する際に特有の詳細を説明します。接続を適切に設定するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

jdbc:postgresql://yourPostgresHost:yourPostgresPort/yourDatabaseName

## 技術的な仕様

### ドライバー仕様

- **PostgreSQLデータベースドライバーの名前とバージョン：**
- ドライバークラス名：org.postgresql.Driver
- バージョン：42.2.8
- **サポートされているPostgreSQLデータベースのバージョン：**
  - 8.2以降

### ドライバーのドキュメント

- **一般的なドライバーのドキュメント：** https://jdbc.postgresql.org/documentation/head/index.html

# Data Prep用のPowerBIコネクター

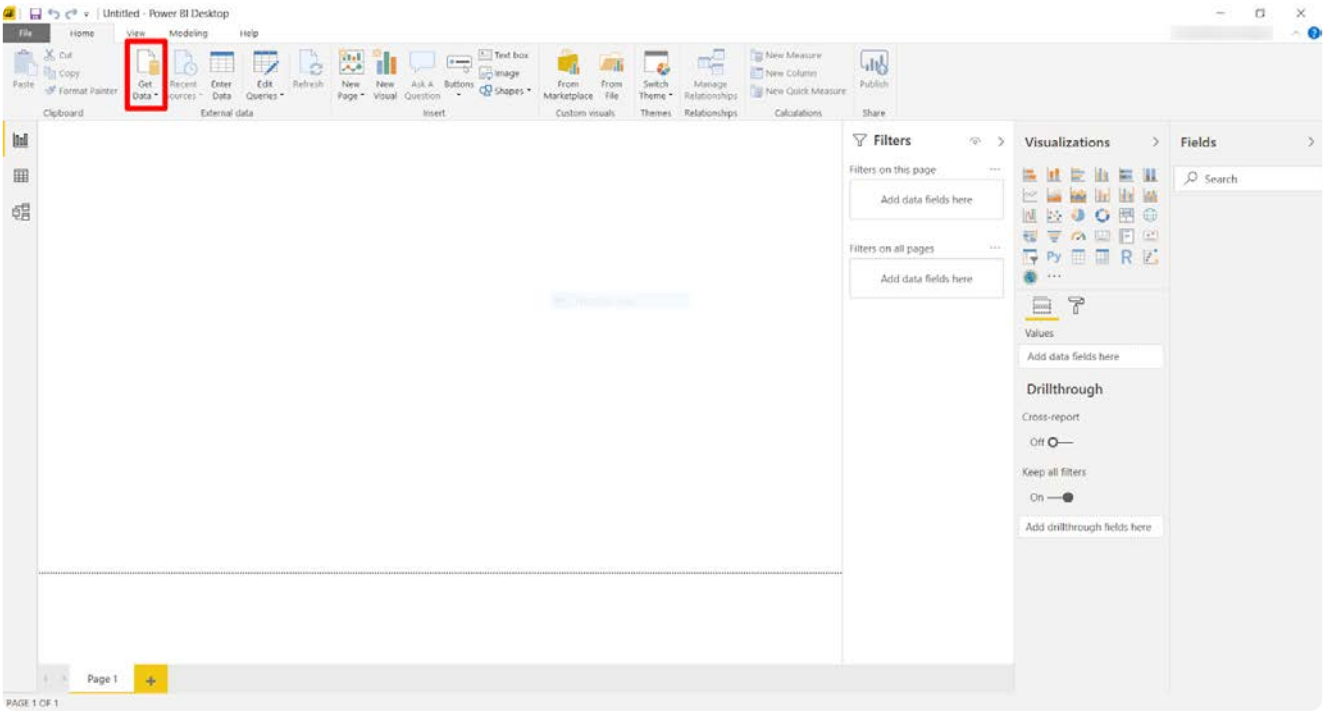
**ユーザーペルソナ**：DataData Prepユーザー

**備考**

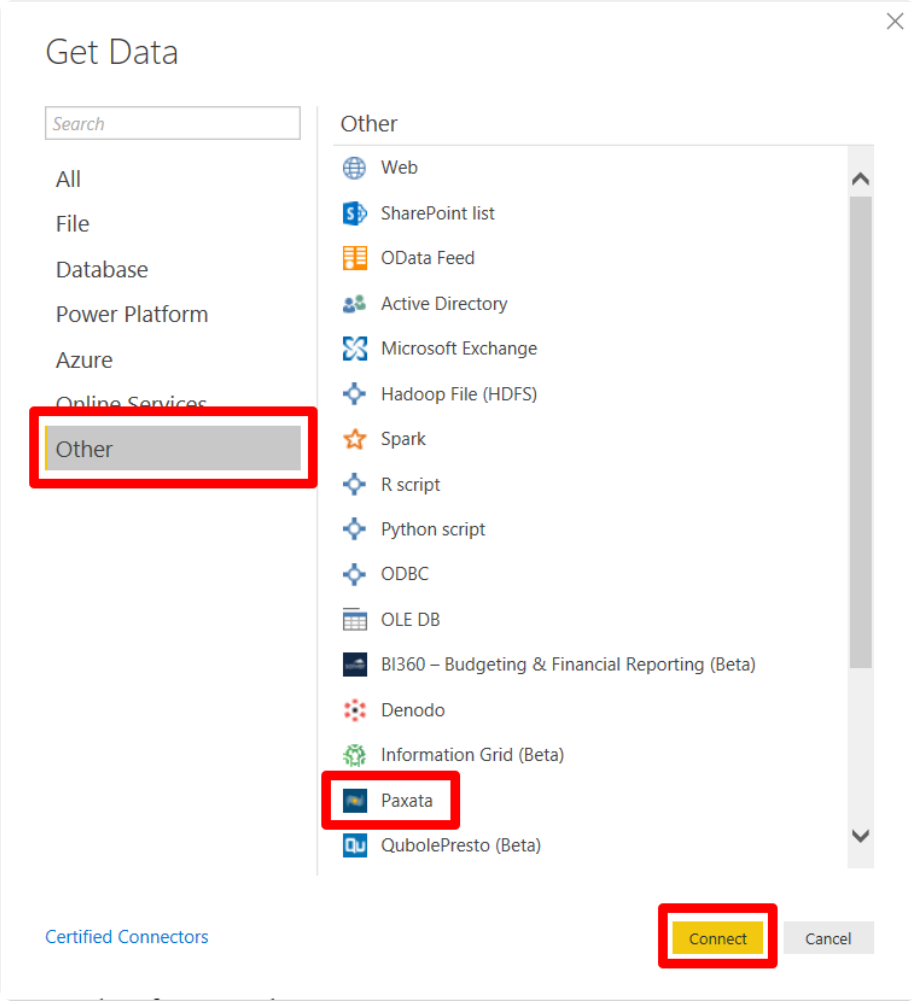
この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## 設定するには

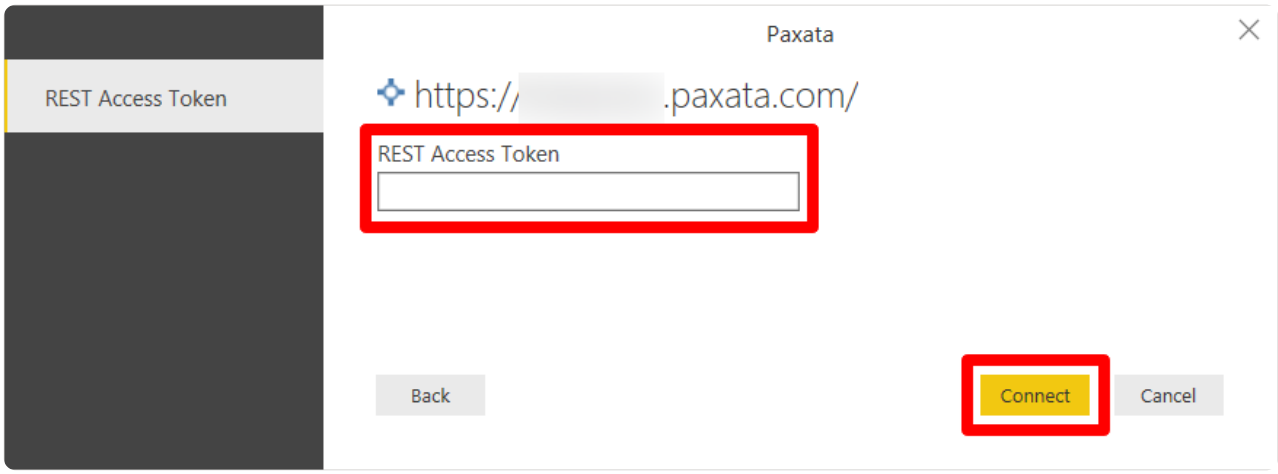
1. PowerBIデスクトップを開く
2. **データを取得する**をクリックします。



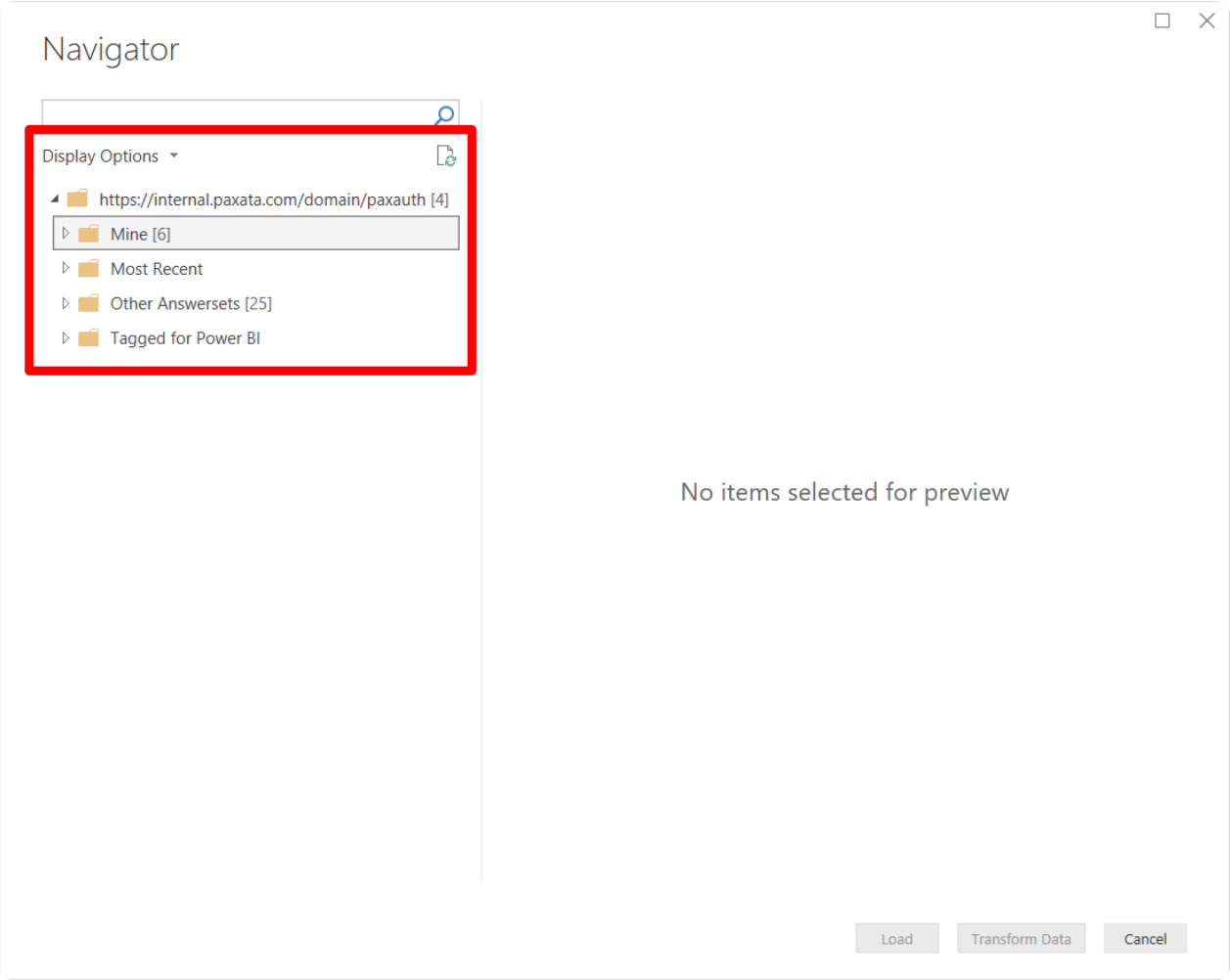

3. **その他 > Data Prep > 接続**をクリックします。




4. 「サードパーティのサービスに接続しています」という警告が表示された場合は、**継続する**をクリックします。
5. ログインするData Prep URLを入力し、**OK**をクリックします。
   a. これは、Kubernetesベースのデプロイの場合「https://yourdomain.Data Prep.com/domain/paxauth」、またはKubernetesベース以外のデプロイを使用している場合、「https://yourdomain.Data Prep.com/」となります。
   b. 先頭に「https：//」を付けることを忘れないでください
6. Data Prep REST APIのトークンを入力し、**接続**をクリックします。
   a. REST APIトークンをお持ちでない場合は、**ユーザー**メニューに移動して**トークン**をクリックします。ここで、アクセスと承認を管理するためのトークンを生成できます。
   b. 紛失した場合は、このトークンは取得できず、削除して再生成するだけであるため、保存してください。




7. 認証の処理には1分ほどかかる場合があります。完了すると、インポートするデータファイルとAnswerSetのData Prepデータライブラリを参照できる画面が表示されます。




# ベストプラクティス

- アドホックレポートの場合、このコネクターを使用することをお勧めします。
- 公開されたダッシュボードの自動レポート/更新については、デスクトップでのプロセスの自動化は不可能であるため、考慮に値するアプローチがいくつかあります。これは、いくつかの状況で成功したアプローチの1つです:

- Data Prepでデータの準備を自動化し、ADLS Gen2などの公開されたPowerBIダッシュボードから直接クエリできるストレージのロケーションにエクスポートします。
- PowerBI側で別の自動化を使用して、共有ロケーションからData Prep AnswerSetをPowerBIにインポートします。

さらに質問がある場合は、Data Prepカスタマーサクセス担当者にお問い合わせください。

# Data Prep用のREST APIコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、REST API に接続して、REST リソースをインポートできます。コネクタの作成に使用されるパラメーターに関する情報を以下に示します。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

REST APIコネクタを使用して、Data Prepを複数のソースに接続したり、同じソースの複数のインスタンスに接続することができます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Webプロキシ

プロキシサーバーを介してREST APIソースに接続する場合、これらのフィールドはプロキシの詳細を定義します。

- **Webプロキシ：** プロキシが不要な場合は [なし]、プロキシサーバー経由でMarketo RESTエンドポイントに接続する必要がある場合は [プロキシ] を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Webプロキシサーバーのホスト名またはIPアドレス。
- **プロキシポート：**データソースのプロキシサーバー上のポート。
- **プロキシユーザー名：**プロキシサーバーのユーザー名。
- **プロキシパスワード：**プロキシサーバーのパスワード。

*認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

## REST API設定

このセクションでは、REST APIリソースを見つけるために使用される情報を提供します。

これを設定する方法の例については、RESTAPI認証設定を参照してください。

- **ベースURL：**REST APIのベースURL。ベース URL にはプロトコル（http/https）、ホスト名（ポート番号はオプション）、コンテキスト パスを含める必要があります。
  - 例： *http(s)://api.domain.com(:port)/rest/v1*
- **リソース：**インポートする複数のRESTリソース。各行には、単一のRESTリソース設定を、*name:path?query*の形式で含めます。
  - *name*はインポートするリソースのユーザーに表示される名前であり、RESTリソース設定に必要です。この名前は、「Browse」のユーザーインターフェースで表示されます。アカウントの詳細。
  - *path*はリソースへのパスであり、RESTリソース設定に必要です。このパスはスラッシュ(/)で始まり、オプションで複数のセグメントをスラッシュ(/)で区切ります。例：*/resource/sub-category*
  - *query*は、リソースの取得時に使用するオプションのフィルタリング基準であり、RESTリソース設定ではオプションです。クエリー構文は、「&」で区切られたキー=値のペアである必要があります。例： *criteria=active&order=desc*または*jql=status=done*。

## REST API認証の設定

このセクションでは、REST API サービスエンドポイントへの認証に使用する情報を指定します。

- **認証タイプ：**要件に基づいて、オプションの1つを選択します。
- 認証なし：REST APIが認証を必要としない場合
- 基本認証：REST APIがユーザー名とパスワードによる認証を許可する場合
- ベアラートークン：REST APIがベアラートークンによる認証を許可する場合ベアラートークンの場合、各Webサービスはトークンへのアクセスまたはトークンの生成を異なる方法で提供する可能性があり、Webサービスのドキュメントでその検索方法を説明する必要があります。
- **ユーザー名およびパスワード：**[基本認証] に [認証タイプ] を選択した場合、これらのフィールドは認証用に提供されます。一部のWebサービスではどちらか一方のフィールドのみが必要です。そのため、ほとんどの場合、両方のフィールドが必要になりますが、設定ページでは両方を空白にすることができます。これにより、データソースへの認証中にエラーが発生する可能性がありますが、データソースの保存時にフォーム検証エラーは発生しません。
- **ベアラートークン：**[認証タイプ] に [ベアラートークン] を選択した場合、これを認証用に提供する必要があります。すべてのシステムがこれを異なる方法で処理するため、ユーザーはこのトークンを取得する方法を知っている必要があります。このトークンを取得するには、管理者の助けが必要になる場合もあります。

## REST APIのテスト接続と運用の設定

- **テスト接続と運用の方法：**Data PrepコネクターがREST APIサービスに接続できるかどうか、また、コネクタがリソースを要求する際にどのメソッドを使用するかを判断するために、リクエストで使用されるHTTP方法。`POST`自動」を選択すると、1、2、3の接続テストを行いますので、どの方法を選択してよいかわからない場合には最適な方法です。`HEAD``GET`
- `GET``HEAD``GET`選択された方法は、実際のインポートにも使用され、1または2がテストに成功した場合は3が、4が成功した場合は5がインポートに使用されることになります。`POST``POST`
- **接続タイムアウト：**REST APIに接続するためのタイムアウト（ミリ秒）。

# データインポート情報

## ブラウジング経由

リソースリストで定義されたリソース名を使用して、インポートワークフロー内のリソースをインポート可能なデータセットとして表示します。

## SQLクエリー経由

サポートされていません。

# 技術的な仕様

## ページ付け

- このコネクターは、RESTデータセットのRFC 5988ページ付けをサポートしています：https：//tools.ietf.org/html/rfc5988
- ページ付けされたREST応答の場合、ページ付けされた各応答には、結果の次のページのURLを識別するHTTPヘッダーが含まれます。
- ページ付けされたデータセットが要求されると、RESTコネクターはデータセットがページ付けされていることを自動的に識別し、データリンクをたどります。
    - データの次のページのHTTPリンクを自動的に抽出します。
    - 結果の現在のページから結果を返します。
    - 呼び出しを実行して、結果の次のページを取得します。
- Data Prep UIを使用したインポート中は、プレビューの迅速な表示を可能にし、レート制限されたAPIに対するヒットを減らすために、1ページのデータ値のみを表示します。
- インポート中に、コネクターは次のことを行います。
    - データの次のページのHTTPリンクを自動的に抽出する。
    - 結果の現在のページから結果を返す。
    - 呼び出しを実行して、結果の次のページを取得します。

## パフォーマンス

RESTコネクターのパフォーマンスは、利用するREST APIの実装に大きく依存します。

- REST APIの呼び出しごとにデータセット全体を返すことをサポートするREST APIで最高のパフォーマンスが得られます。これは、チャンク転送エンコーディングを利用するAPIの典型です。このシナリオでは、RESTコネクターは単一のAPI呼び出しを実行して完全なデータセットを取得します。
- ページ付けを利用するREST APIは、追加のREST API呼び出しが必要になるため、パフォーマンスが低下します。
  - ページ付けスタイル：RFC 5899
  - 各応答には、N個のレコードと、次のバッチを指すURLを含むHTTPヘッダーが含まれます。
- REST APIドキュメントを確認して、API呼び出しの数を減らすために設定できるMaximum（最大）ページサイズを特定します。
- 例：GitHub REST API
- APIにはレート制限がある場合があります。ページ分割された大規模なデータセットをインポートする場合、時間枠内に実行されるREST呼び出しの数に制限がかかることは珍しくありません。たとえば、GitHubでは1時間あたり5000件のリクエストが許可され、Googleドライブでは100秒あたり1000件のリクエストが許可されます。

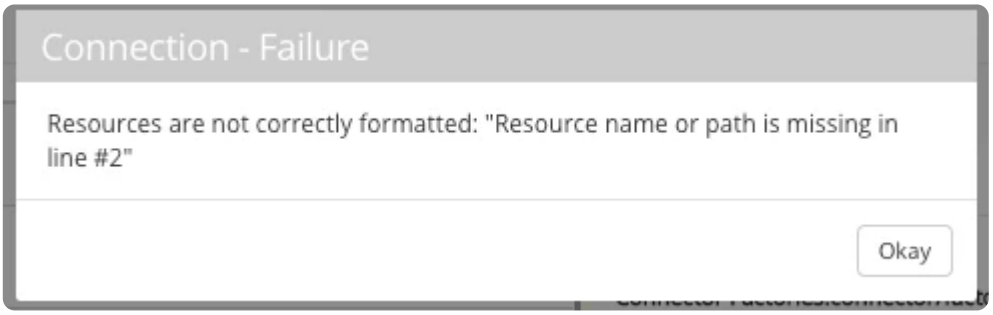
# FAQ／トラブルシューティング／一般的な問題

## OAuth認証はサポートされていますか？

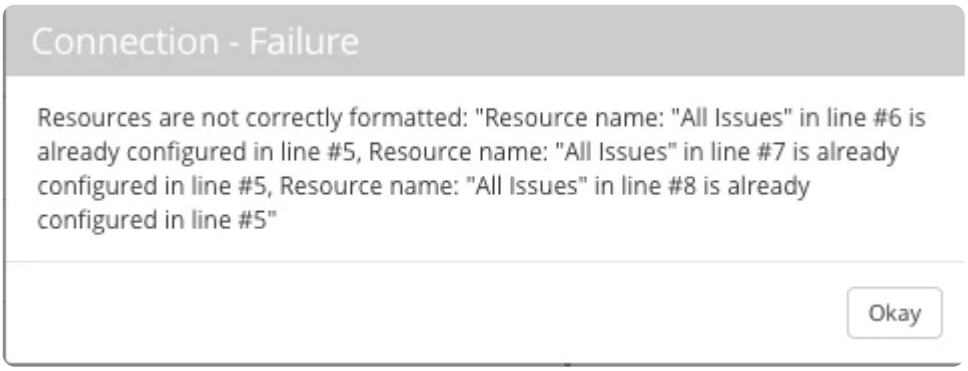
現時点ではサポートされていません。現在、ユーザー名/パスワードおよびトークン認証方法のみがサポートされています。多くのデータソースはOAuth認証のみを許可しており、これらのソースは現時点ではサポートされていません。この理由でDataSourceに接続できない場合は、Data Prep Client Successに連絡してください。

## 「テスト接続」メッセージは何を意味しますか？

- テスト接続は、リソースリストの各エントリが期待される形式と一致することを確認します。
- エントリーが期待される形式と一致しない場合、識別された形式の問題とエントリー番号を示すエラーが発生します。




- 各リソースエントリーにユニーク数のデータセット名を使用しないと、形式の検証に失敗します。




- テストによってリソースの適切な形式が確認された後、リソースリストの最初のエントリのみが、接続が正しく設定されていることを確認するために使用されます。

# 設定例

以下は、REST APIコネクターの実際の使用例です。アカウントでこれらのいずれかを自由に使用してください。ただし、企業はAPIを随時、場合によっては通知なしに変更する可能性があり、これらは完全にサポートされているデータソースではないことにご注意ください。これは、Data Prepがこれらの設定で発生する可能性のある問題のトラブルシューティングを支援できず、それらが時代遅れになる可能性があることを意味します。

## 簡単な学習例

Web上には、学習、テスト、およびプロトタイピングを目的として作成された、単純で認証されていないREST APIリソースが数多くあります。これらのソースの1つはJSONPlaceholderです。この例は単純化しすぎている可能性がありますが、REST APIコネクターを使用してRESTful Webサービスに接続する方法の構成要素を示すことを目的としています。

データセットが小さいため、レート制限は投稿されず、ページ付けもありません。

設定

- ベースURL：https://jsonplaceholder.typicode.com
- リソース：
- 投稿：/投稿
- コメント：/コメント
- アルバム：/アルバム
- 認証：認証なし
- REST APIテスト設定：自動化

[データソースのテスト]をクリックしてセットアップが機能していることを確認し、[保存]をクリックした後、このデータソースを使用してデータをData Prepにインポートできます。

# GitHubの例

GitHubは、リッチだがレート制限のあるREST APIを提供するクラウドベースのソフトウェアソースコードリポジトリです。GitHub REST APIドキュメント。（このリンクをクリックする前にGitHubにログインしてください。）

#### レート制限

- GitHub APIレート制限リリファレンス： https://developer.github.com/v3/rate_limit/（このリンクをクリックする前にGitHubにログインしてください。）
- GitHubでは、1時間あたり5000件のリクエストが許可されており、具体的な制限はサービスごとに異なります。
- レート制限は、認証されていないユーザーと認証されたユーザーで異なります。

#### ページ付け

- GitHubは、RESTデータセットのRFC 5899ページ付けをサポートしています。
- ページ分割されたREST応答の場合、ユーザーはAPI呼び出しごとに1ページのデータ(/search APIの場合は30エントリ)を受け取ります。
- ユーザーは、「ペール_ページ」APIパラメーターを使用して、ページあたりの結果数を最大100までオーバーライドできます。
- ページごとの結果カウントをMaximum（最大）許容設定に設定すると、REST API呼び出しの数が減るため、データインポートのスループットが向上します。

#### 設定

- ベースURL： https://api.github.comこのリンクをクリックする前にGitHubにログインしてください。
- リソース：
  - Mozilla Repos：「mozilla」の検索に一致するソフトウェアリポジトリのリストを取得します。
    - 注：このクエリーは、呼び出しごとにデフォルトの30レコードを使用する場合、/ サーチAPIに対するユーザーのクォータを使い果たします。
    - 期待される結果数 > 6600
    - Mozillaリポジトリ：/search/repositories？q=mozilla
- Mozillaリポジトリページ33+：結果の33ページから始まる検索セットを実行する例
  - Mozilla Repos Page 33:search/repositories?q=mozilla&page=33
- Square Repos：Square組織に属するリポジトリのリストを取得します。
  - Square Repos：/orgs/square/repos
- 組織：要求あたり100件のレコードを使用して、すべてのGitHubの組織のページ化されたリストを取得します。*警告：この場合、200万件以上のエントリをプルするために長時間実行されます。*
  - Organizations:/organizations?per_page=100

# Jiraの例

Jiraは、通常、ソフトウェア開発チームが使用するプロジェクトおよび課題追跡ソフトウェアです。Jira REST APIドキュメント。

#### レート制限、ページ付け、および設定

- レート制限は、サブスクリプションレベルによって異なります。
- Jira REST APIは、データのページごとに最大100件の結果を返すように制限されています。
- JiraはRFC 5899のページ付けをサポートしていません。
- Jira クラウドインスタンスでは、ユーザーがJIRA REST APIトークンを作成する必要がある場合があります。
- トークンを作成します。https://confluence.atlassian.com/cloud/api-tokens-938839638.html
- 認証タイプ：基本認証
- ユーザー名フィールドのユーザー名
- パスワードフィールドのAPIトークン

#### 設定

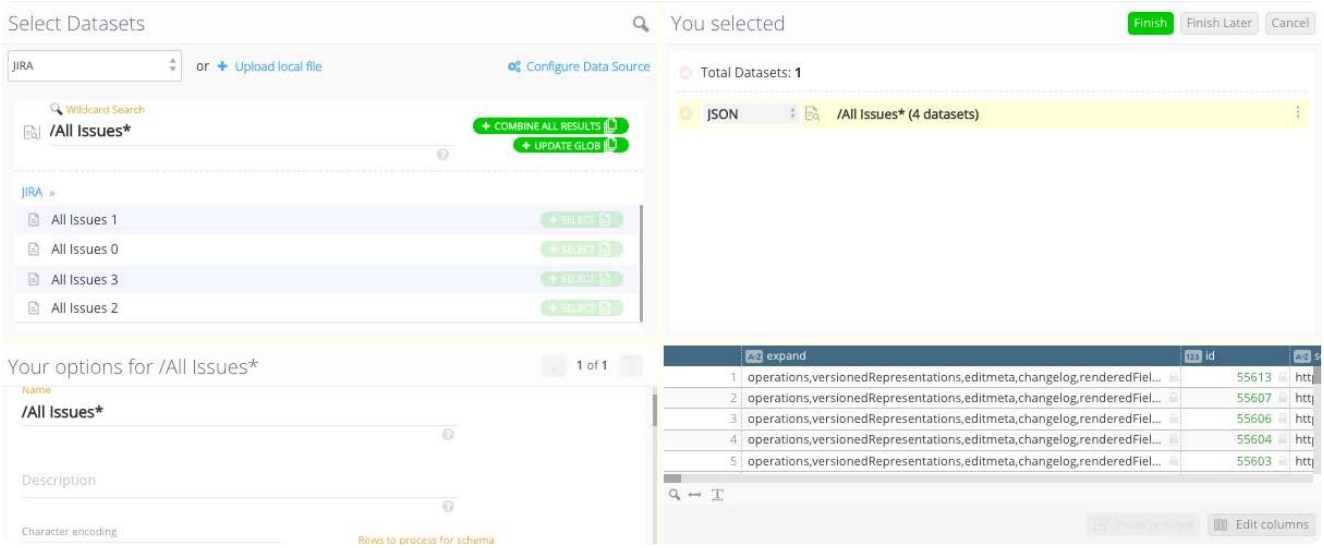
- ベースURL：
- オンプレミス： https://(hostname):(port)/rest/api/
- Jira クラウド： https://(your-domain).atlassian.net/rest/api/
- リソース
- 全プロジェクトリスト：
  - 全プロジェクト：/project
- JQLリソースの例：Jira JQLクエリーを実行して200のTo-Doタスクアイテムを取得します。
  - JQLクエリーは、設定に貼り付ける前にURLエンコードする必要がありました。
  - Connector To Do Tasks：/search?jql=Project%3D_yourProject_%20and%20statusCategory%3D%22To%20Do%22&maxResults=200
- 認証：基本（ユーザー名/パスワード）

#### JIRAのページ付け

JiraはRFC 5899のページ付けをサポートしていません。JIRAのページ付けをサポートするには：

- データのページを指定するデータソースリソースエントリを定義します。
- 「maxResults=100」を使用して、REST呼び出しごとのエントリ数を最大化します。
- 開始点の指定には「startAt=N」を使用します。Nは0から始まります。
- 例：100件の検索結果の4ページ
  - すべての問題 0：/search？jql=&startAt=0&maxResults=100 すべての問題 1：/search？jql=&startAt=100&maxResults=100 すべての問題 2：/search？jql=&startAt=200&maxResults=100 すべての問題 3：/search？jql=&startAt=300&maxResults=100

- Data Prepのワイルドカード機能を使用して、データのすべてのページを選択
- ワイルドカードパターン= "すべての問題*"




- JSONをフラット化し、サブタスクを説明するためにいくつかの行を複製するJSON解析の後、557行のデータX 298列を取得しました。

# 米国国勢調査データの例

国勢調査データWebサイトはREST APIではありませんが、REST APIコネクターを使用してHTTP経由でデータを取得できます。

- ベースURL： https://www2.census.gov/
- American Community Survey 2002のリソース例

ACS_2002_Midwest:/acs2002/2007_prod_release1/BaseTablesSubjectTables/Region/MidwestRegionBaseTables02000US2.csv

ACS_2002_US_OH_Franklin:/acs2002/2007_prod_release1/BaseTablesSubjectTables/States/Ohio/StateCounty/FranklinCountyOhio/BaseTables05000US39049.csv

ACS_2002_Base_California:/acs2002/2007_prod_release1/BaseTablesSubjectTables/States/California/CaliforniaBaseTables04000US06.csv

- 認証：なし
- 「ACS 2002 Base」データのすべてのページを選択するには、Data Prepのワイルドカード機能を使用します。 ワイルドカードパターン：
- 「ACS2002ベースA *」：アラバマ、アラスカ、アーカンソー、アリゾナのファイルを1つのデータセットとしてインポートします。
- 「ACS2002ベース*」：一致するすべての「ACS2002ベース」ファイルを1つのデータセットとしてインポートします。

# Data Prep用のSalesford Ligningコネクター

**ユーザーペルソナ**：Data PrepユーザーまたはSalesford管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Salesforce組織にインポートソースとして接続できます。次のフィールドを使用して、接続パラメーターを定義します。

**備考**

Salesforce Lightning Editionの使用は、このコネクターを使用するための前提条件 *ではありません*。コネクターは、REST API v40.0以降をサポートするすべてのSalesforce組織で機能します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のSalesforce組織（Sandbox、Dev、UATなど）に接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

## Webプロキシ

プロキシ サーバー経由で Salesforce に接続する場合、次のフィールドでプロキシの詳細を定義します。

- **Webプロキシ：**Webプロキシサーバーが必要となり、プロキシ接続を有効にする際には、以下のフィールドが必要となります。
- **プロキシ ホスト:** Web プロキシ サーバーのホスト名または IP アドレスです。
- **プロキシサーバー：**データソースのプロキシサーバー上のポートです。
- **プロキシ ユーザー名：**プロキシ サーバーのユーザー名です。
- **プロキシ パスワード：**プロキシ サーバーのパスワードです。 *認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

## ユーザー資格情報

ユーザー認証は、共有アカウントまたは個人アカウントを介して行うことができます。個人アカウントで認証を行う場合は、このデータ ソースにアクセスするための資格情報の入力を求められます。共有アカウントで認証する場合は、次のフィールドを設定する必要があります。

- **Salesforce URL:** Salesforce の URL です。デフォルトでは、https://login.salesforce.comです。
- Salesforce Sandboxに接続している場合は、URLをhttps://test.salesforce.comに設定します。
- **セッション セキュリティ:** Salesforce への接続時に使用するセッションのセキュリティ オプションです。Salesforce が信頼できるIP範囲を使用するように構成されていない場合は、*[APIセキュリティトークン]*を選択します。もしも*「信頼できるIP範囲」*が選択されている場合、APIセキュリティトークンの値は必要ありません。
- **ユーザー:**Salesforce での認証に使用される共有アカウントのユーザー名です。
- **パスワード:** Salesforce での認証に使用される共有アカウントのパスワードです。
- **API セキュリティ トークン:** API ベースのアクセスを可能にする、ユーザー固有の Salesforce API セキュリティ トークンです。このトークンは、通常、パスワードが変更されるたびに Salesforce ユーザーにメールで送信されます。これは、[セッションセキュリティ] で [信頼済み IP 範囲を使用] が選択されている場合は必要ありません。
- **OAuthアプリのフィールド**Salesforce組織でOAuthアプリとしてData Prepを設定するには、以下の「Salesforceの設定」セクションの手順に従ってください。これらの手順を完了すると、必要な値を見つけることができます。
- **使用者キー：**Data PrepのSalesforce Lightningコネクタが Salesforceに対してそれ自体を識別するために使用する値OAuth 2.0 では、この値はクライアント ID と呼ばれます。
- **コンシューマ シークレット:**Data PrepのSalesforce Lightning コネクタがコンシューマキーの所有権を確立するために使用するシークレットです。OAuth 2.0 では、この値はクライアント シークレットと呼ばれます。

# Salesforceの設定：

Salesforce Lightningコネクターは、Salesforce REST APIとOAuthを利用します。このステップは、Salesforce管理者が行う必要があります。

接続を確立するには：

- Salesforce管理者は、Salesforceで「接続済みアプリ」を作成する必要があります。
- Salesforce管理者は、Salesforceにアクセスしているクライアント(Data Prepコネクター)を識別するために、「接続済みアプリ」のOAuth認証情報を取得する必要があります。
- 個々のユーザーが認証する必要があります。

## Salesforceでの「接続済みアプリ」の作成

接続済みアプリの背景とSalesforce内のナビゲーションについては、すぐ下にリンクされているSalesforceドキュメントを使用してください。指定されたフィールドを設定する方法については、以下のガイドを使用してください。

Salesforceの手順：Salesforce Instructions: 1https://help.salesforce.com/articleView?id=connected_app_create.htm&type=51

- **OAuth設定を有効にする：**このオプションを選択します。
- **コールバックURL：**SalesforceではコールバックURLを定義する必要がありますが、Data Prepには必要ありません。
- 簡単にするために、Data Prep URLを使用してください。
- **選択されたOAuthスコープ**：
- 「**利用可能なOAuthスコープ**」で、[データへのアクセスと管理(api)]を選択し、[追加]ボタンをクリックします。
- **Webサーバーフローにシークレットを要求する：**Data PrepがOAuth使用者シークレットを安全に保存するように、このオプションを選択します。
- 必ず「保存」を押してください。

## 「接続済みアプリ」のOAuth認証情報

保存時：

- 「接続済みアプリ」のOAuth認証情報を含むページが表示されます。
- **コンシューマーキー**：これをコピーして保存します。
- **コンシューマーシークレット：**[クリックして表示]リンクをクリックします。
- **OAuth Webサーバーフローの信頼できるIP範囲**
  - 顧客がIPアドレスのリストへのアクセスをさらに制限したい場合は、Data PrepサーバーのIPアドレスがここにIP追加されます。

**データインポート情報**

## ブラウジング経由

- すべてのオブジェクトはアルファベット順にリストされています。
- カスタムオブジェクトもアルファベット順に表示され、最後に「__c」が付きます。
- 名前空間を持つオブジェクトも、名前空間の下にアルファベット順に表示されます。

- 上部にあるレポートのフォルダー

## SQLクエリー経由

Salesforceコネクターでのクエリの使用は、Salesforceオブジェクトクエリ言語のSOQLと呼ばれるSalesforce独自のクエリ言語に依存しています。これに関する情報については、https://developer.salesforce.com/docs/atlas.en-us.soql_sosl.meta/soql_sosl/sforce_api_calls_soql.htmを参照してください。

ここにいくつかのクエリの例：

1. エスケープ文字を含むクエリー

   `SELECT Id FROM Account WHERE Name LIKE 'Ter%'`

2. 関数を含むクエリ アカウントからcount(id)を選択

   `SELECT Name, MAX(Amount), MIN(Amount) FROM Opportunity GROUP BY Name`

3. エイリアスを使用したクエリー

   `SELECT a.Id, c.Id, c.name FROM Contact c, c.Account a WHERE a.name = 'MyriadPubs'`

4. Where句のnull

   `SELECT AccountId FROM Event WHERE ActivityDate != null`

5. Where句のSubQuery

   `SELECT Id, Name FROM Account WHERE Id IN ( SELECT AccountId FROM Opportunity WHERE StageName = 'Closed Lost' )`

6. GroupBy句で

   `SELECT LeadSource, COUNT(Name) FROM Lead GROUP BY LeadSource`

7. エイリアスとしてのフィールド

   `SELECT Name n FROM Opportunity`

8. TYPEOF によるクエリー `SELECT TYPEOF What WHEN Account THEN Phone, NumberOfEmployees WHEN Opportunity THEN Amount, CloseDate ELSE Name, Email END FROM Event`

9. 関係性クエリー（親へ）

   `SELECT Contact.FirstName, Contact.Account.Name FROM Contact`

10. 関係性クエリー（子へ）

    `SELECT Account.Name, (SELECT Contact.LastName FROM Account.Contacts) FROM Account`

11. カスタムオブジェクトを使用した関係性クエリー

    `SELECT Opportunity__c, Id, Opportunity__r.Name, Opportunity__r.Owner.Manager.Email, Opportunity__r.Owner.Email FROM Opportunity_Change__c`

12. Select句のサブクエリー

    `SELECT Amount, Id, Name, ( SELECT Quantity, ListPrice, PricebookEntry.UnitPrice, PricebookEntry.Name FROM OpportunityLineItems ) FROM Opportunity`

# ベストプラクティス

- Salesforceへのエクスポート：
- ほとんどの組織には、Salesforce内で情報を一括更新するためのプロセスが定義されています。このため、Data Prep Salesforceコネクターはインポートのみをサポートしています。
- データをSalesforceにエクスポートするには、データのCSVファイルをローカルにダウンロードし、組織のガイドラインに従ってSalesforceにアップロードします。
- データをSalesforceに一括ロードする方法の詳細については、https://help.salesforce.com/articleView?id=data_import_wizard.htm&type=5を参照してください。

# 技術的な仕様

- Salesforce REST API v40.0の利用

# FAQ／トラブルシューティング／一般的な問題

## Salesforceカスタムレポートの行制限

Salesforce APIは、Salesforceレポートを2000行の結果セットに制限します。

- https://help.salesforce.com/articleView?id=rd_reports_limits.htm&type=5T

## 列データ型の処理

Salesforceからデータをインポートする場合、Data Prepは、参照時とクエリ時で列のデータ型を異なる方法で処理します。

- 参照してインポート：
- 列のデータ型は、Salesforceオブジェクトのメタデータを使用して識別され、Data Prepの内部型にマッピングされます。
- SOQLクエリのインポート：
- Salesforce SOQLの結果は列の特定のデータ型を返さず、Data Prepはクエリまたはクエリ結果を解析して列のデータ型を決定しません。その結果、すべてのクエリ結果はテキストとして解釈されます。

**例**

「2013-11-13」を含む「CloseDate」列を持つSalesforce Opportunity SObject行をインポートする場合、CloseDateは次のようにインポートされます。

- 参照してインポート：DateTime
- "2013-11-13T00:00:00.000-08:00"
- SOQLクエリー：テキスト
- "2013-11-13"

# Data Prep用のSalesforce Marketing Cloudコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Salesforce Marketing Cloudに接続して、利用可能なデータを参照およびインポートできます。次のパラメーターを使用して、接続を設定します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

複数のSalesforceマーケティングクラウドアカウントへのData Prepを接続することができます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### Salesforce Marketing Cloudの設定

- **ユーザー:** Salesforce Marketing Cloud のユーザーです。
- **パスワード:** Salesforce Marketing Cloud ユーザーのパスワードです。
- **タイムアウト:** 操作がタイムアウトするまでの待機時間（秒単位）です。デフォルト値はありません。

## Webプロキシ設定

プロキシサーバーを介してSalesforce Marketing Cloudに接続する場合、これらのフィールドはプロキシの詳細を定義します。

- **Web プロキシ:** プロキシが不要な場合は [なし]、プロキシ サーバー経由で Salesforce Marketing Cloud に接続する必要がある場合は [プロキシ] を選択します。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシ ホスト:** Web プロキシサーバーのホスト名または IP アドレスです。
- **プロキシサーバー：**Saleのプロキシサーバー上のポート。
- **プロキシ ユーザー名：**プロキシ サーバーのユーザー名です。
- **プロキシ パスワード：** プロキシサーバーのパスワード。

> **備考**
>
> 認証されていないプロキシ接続では、ユーザー名とパスワードを空白のままにします。

# データのインポート／エクスポート情報

## ブラウジング経由

テーブルを参照し、インポートするテーブルを「選択」します。

## SQLクエリー経由

正当なSQL選択クエリの使用

- 日付/時刻値の場合、> と < だけがWHERE句でサポートされています。
- 日付/時刻以外の値の場合、=、!=、<>、>、>=、<、<=、INがWHERE句でサポートされています。

# Salesforceオブジェクト

| 名前 | タイプ | 説明 |
|---|---|---|
| アカウント | 表 | Marketing Cloudアカウント |
| AccountUser | 表 | アカウント内の個々のユーザーこの表は削除をサポートしていません。 |

| 名前 | タイプ | 説明 |
| --- | --- | --- |
| BusinessUnit | 表 | 大規模なEnterpriseまたはEnterprise 2.0アカウント内のユニットこの表は、クエリと更新のみをサポートします。 |
| ContentArea | 表 | ContentAreaは、再利用可能なコンテンツの定義済みセクションを表します。 |
| DataExtension | 表 | アカウント内のデータエクステンションを表します。 |
| Eメール | 表 | Marketing CloudアカウントのEメールを表します。 |
| EmailSendDefinition | 表 | メッセージ情報、送信者プロファイル、配信プロファイル、およびオーディエンス情報を含むレコード |
| FileTrigger | 表 | 将来の使用のために予約されています。この表は削除をサポートしていません。 |
| FilterDefinition | 表 | フィルターで指定されたルールに基づいてオーディエンスを定義します。この表は挿入をサポートしていません。 |
| ImportDefinition | 表 | インポートオプションの再利用可能なパターンを定義します。この表は挿入をサポートしていません。 |
| List | 表 | 購読者のマーケティングリスト |
| Portfolio | 表 | Marketing Cloudアカウントのポートフォリオ内のファイルを示します。 |
| ProgramManifestTemplate | 表 | 将来の使用のために予約されています。この表は、削除または挿入をサポートしていません。 |
| QueryDefinition | 表 | SOAP APIによってアクセスおよび実行されるSQLクエリアクティビティを表します。この表は更新または挿入をサポートしていません。 |
| ReplyMailManagementConfiguration | 表 | アカウントでの返信メール管理の詳細設定この表は削除をサポートしていません。 |

| 名前 | タイプ | 説明 |
| --- | --- | --- |
| 送信 | 表 | Eメールの送信と集計データの取得に使用されます。この表は、削除または更新をサポートしていません。 |
| SendClassification | 表 | Marketing Cloudアカウントの送信分類を表します。 |
| SenderProfile | 表 | Eメール送信定義と組み合わせて使用される送信プロファイル |
| SMSTriggeredSend | 表 | SMSトリガー送信の単一インスタンスを示します。この表は、削除または更新をサポートしていません。 |
| Subscriber | 表 | EメールまたはSMS通信を受信するように登録した人 |
| SuppressionListDefinition | 表 | さまざまなコンテキストに関連付けることができる抑制リスト |
| TriggeredSendDefinition | 表 | リストIDがすべての購読者リストIDであるTriggeredSendDefinitionを作成または更新するには、Eメール - 購読者 - すべての購読者 - 表示およびSendEmailToListの権限が必要です。 |
| 自動化 | 表示 | アカウントのAutomation Studio内に存在するオートメーションを定義します。 |
| BounceEvent | 表示 | Eメールメッセージのバウンスの特定のイベントに関連するSMTPおよびその他の情報が含まれています。 |
| ClickEvent | 表示 | メッセージに含まれるリンクのクリックに関する、日時情報、およびURL IDおよびURLが含まれます。 |
| DataExtensionField | 表示 | データ拡張内のフィールドを表します。 |
| DataExtensionTemplate | 表示 | アカウント内のデータエクステンションテンプレートを表します。 |
| DataFolder | 表示 | Marketing Cloudアカウントのフォルダーを表します。 |

| 名前 | タイプ | 説明 |
| --- | --- | --- |
| DoubleOptInMOKeyword | 表示 | DoubleOptInMOKeywordオブジェクトは、MOキーワードを定義し、モバイルユーザーがダブルオプトインワークフローを使用してSMSメッセージを購読できるようにします。 |
| FileTriggerTypeLastPull | 表示 | 将来の使用のために予約されています。 |
| ForwardedEmailEvent | 表示 | 購読者が友人への転送機能を使用して、Eメールを他の人に送信したことを示します。 |
| ForwardedEmailOptInEvent | 表示 | Forward To A Friendイベントに関連するオプトインイベントを指定します。 |
| HelpMOKeyword | 表示 | アカウントのHELP SMSキーワードに関連付けられたアクションを定義します。 |
| ImportResultsSummary | 表示 | ImportDefinitionから開始された個々のインポートに関するステータスと集計情報を含む取得専用オブジェクト |
| LinkSend | 表示 | 送信内のリンクに関する情報を提供します。 |
| ListSend | 表示 | 完了した送信のリストに関連付けられている取得専用のプロパティを指定します。 |
| ListSubscriber | 表示 | サブスクライバーのリストまたはリストのサブスクライバーを取得します。 |
| MessagingVendorKind | 表示 | SMS（ショートメッセージサービス）またはボイスメッセージベンダーのベンダーの詳細が含まれています。廃止。 |
| NotSentEvent | 表示 | Eメールメッセージの送信に失敗したときの情報が含まれています。 |
| OpenEvent | 表示 | サブスクライバーによって送信されたメッセージの開封に関する情報が含まれています。 |

| 名前 | タイプ | 説明 |
| --- | --- | --- |
| PrivateIP | 表示 | PrivateIPオブジェクトには、メッセージ送信の一部として使用されるプライベートIPアドレスに関する情報が含まれています。 |
| Publication | 表示 | 将来の使用のために予約されています。 |
| PublicationSubscriber | 表示 | パブリケーションリストのサブスクライバーについて記述します。 |
| PublicKeyManagement | 表示 | 将来の使用のために予約されています。 |
| ResultItem | 表示 | 非同期API呼び出しの結果が含まれます。 |
| ResultMessage | 表示 | 非同期呼び出しの結果を含むメッセージ |
| 役割 | 表示 | アカウントのユーザーに割り当てられるロールと権限を定義します。 |
| SendEmailMOKeyword | 表示 | MOメッセージで定義されたEメールアドレスにトリガーされたEメールメッセージを送信するアクションを定義します。 |
| SendSMSMOKeyword | 表示 | 指定されたMOキーワードを受信したときに実行するアクションを定義します。 |
| SendSummary | 表示 | 特定の送信イベントに関するサマリー情報を含む取得のみのオブジェクト |
| SentEvent | 表示 | 個々のサブスクライバーに関する情報など、送信に関連する追跡データが含まれます。 |
| SMSMTEvent | 表示 | サブスクライバーに送信される特定のSMSメッセージに関する情報が含まれます。 |

| 名前 | タイプ | 説明 |
| --- | --- | --- |
| SMSSharedKeyword | 表示 | Marketing CloudアカウントのSMSメッセージで使用するキーワードをリクエストするために使用される情報が含まれています。 |
| SMSTriggeredSendDefinition | 表示 | SMSメッセージの送信定義を定義します。 |
| SubscriberList | 表示 | 特定のサブスクライバーのリストを取得するために使用します。 |
| SubscriberSendResult | 表示 | 将来の使用のために予約されています。 |
| SuppressionListContext | 表示 | SuppressionListDefinitionを関連付けることができるコンテキストを定義します。 |
| SurveyEvent | 表示 | 調査回答が行われた日時に関する情報が含まれています。 |
| テンプレート | 表示 | Marketing CloudアカウントのEメールテンプレートを表します。 |
| TimeZone | 表示 | アプリケーションの特定のタイムゾーンを表します。 |
| TriggeredSendSummary | 表示 | 特定のトリガーされた送信の結果の概要。 |
| UnsubEvent | 表示 | サブスクライバーが実行した特定のサブスクリプション解除アクションに関する情報が含まれています。 |
| UnsubscribeFromSMSPublicationMOKeyword | 表示 | サブスクライバーがSMSパブリケーションリストからサブスクライブ解除するために使用するキーワードを定義します。 |

# Data Prep用のSAP HANAコネクター

**ユーザーペルソナ**：Data Prep管理者またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

SAP HANAに接続する機能は、Data Prep JDBCコネクターの一部です。このトピックでは、SAP HANAへの接続のセットアップに固有の詳細を説明します。接続を構成するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

`jdbc:sap://_yourSAPHanaHostOrIP_:_yourSAPHannaPort_/?currentschema=_yourRootSchema_`

## 技術的な仕様

### ドライバー仕様

- **SAP HANAデータベースドライバーのバージョン：**
  - バージョン：2.4.63
- **サポートされているSAP HANAデータベースのバージョン：**
  - SAP HANA 1.0およびSAP HANA 2.0データベース

### ドライバーのドキュメント

- **一般的なドライバーのドキュメント：** https://help.sap.com/viewer/0eec0d68141541d1b07893a39944924e/2.0.04/en-US

# Data Prep用のSpark SQLコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、SparkSQLに接続して、使用可能なデータのブラウジング、インポート、エクスポートを行うことができます。以下のフィールドは、接続パラメーターを定義するために使用されます。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のSpark SQLインスタンスに接続することができます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Spark SQLサーバー設定

- **Spark SQLサーバー**：Spark SQLデータベースをホストするサーバーのホスト名またはIPアドレスです。
- **Spark SQLポート**：Spark SQLデータベースのポートです。
- **SSLを使用する**：このプロパティにHive設定ファイル（hive-site.xml）の「hive.server2.use.SSL」のプロパティで指定された値を設定します。
- **トランスポートモード**：このプロパティにHive設定ファイル（hive-site.xml）の「hive.server2.trantsports.mode」のプロパティで指定された値を設定します。

- **HTTPパス**：このプロパティは、HTTPトランスポートモードを使用する際のURLエンドポイントのパスコンポーネントを指定するために使用されます。このプロパティには、Hive設定ファイル（hive-site.xml）のhive.server2.thrift.http.pathプロパティで指定された値を設定する必要があります。

## Spark SQLサーバー認証設定

- **ユーザー**：Spark SQLでの認証に使用されるユーザー名です。 Databricksの場合、'token'に設定します。
- **パスワード**: Spark SQLでの認証に使用されるパスワードです。Databricksの場合は、個人用アクセストークンに設定します（値は、Databricksインスタンスの[ユーザー設定]ページに移動し、[アクセストークン]タブを選択することで取得できます）。

# データインポート情報

## ブラウジング経由

- 表を表示し、インポートする表を「選択」します。

## SQLクエリー経由

- 正当なSQL Selectクエリを使用したインポートをサポートします。

# Data Prep用のSFTPコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクタを使用すると、ライブラリのインポートとエクスポートのためにSSHファイル転送プロトコル（SFTP）サーバーに接続できます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のSFTPサーバーに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### SFTP ホスト

[ソースの追加]、または [ソースの編集] フォームに [SFTP ホスト] セクションが表示されている場合は、SFTP ホストを探して接続するときに使用された情報を指定します。

- **SFTP ホスト名：**SFTP サーバーの完全修飾ホスト名（ドメイン名を含む）または IP アドレスを使用できます。
- **SFTPポート：**SFTPサーバーのソケットポートです。プロトコルは、ポート22をデフォルトとして指定します。
- **自動ホストキー検証：**SFTP サーバーからホストキーを自動的に受け入れます。
- *選択済み：*このオプションを使用すると、SFTPサーバーへの接続を自動的に信頼できるようになります。これは、SSHでStrictHostKeyChecking=noを設定するのと同じです。

- *選択解除（デフォルト設定）：*このオプションを使用すると、SFTPホスト名への自動信頼設定を無効にします。これは、より高いセキュリティ設定を表すため、デフォルトのオプションとして選択されています。
- **キープアライブ：**タイムアウトを避けるためのセッションアクティビティを有効化/無効化します。
  - *選択済み（デフォルト設定）：*SFTPコネクターとSFTPサーバー間の定期的なバックグラウンド通信を有効にして、ブラウジング、インポート、およびエクスポート中にサーバーによって接続が閉じられないようにします。
  - *選択解除：*接続の持続期間は、SFTPサーバー構成によって管理されます。アイドル状態の接続は、サーバーによって終了される場合があります。この構成では、データをインポート/エクスポートするために参照するときに、非アクティブ状態が続くのを避けるのが最善です。
- **データ圧縮：**転送中のデータ圧縮を有効にします。
- *選択済み（デフォルト設定）：*SFTPサーバーとData Prep間の転送中にデータのZLIB圧縮を有効にし、ほとんどのデータセットの転送速度を向上させます。Data Prepとサーバー間でZLIB圧縮をネゴシエートできない場合、接続は自動的に非圧縮転送に後退します。
- *選択解除：*SFTPサーバーとData Prep間の転送中のデータのZLIB圧縮を無効にします。
- **ソケットタイムアウトの秒数:** SFTP コマンド（ディレクトリのリスト、ディレクトリの作成、ログアウトなど）の実行を待機する秒数です。デフォルト値は30秒です。待機時間を長くするには、この値を増やします。
- このオプションは、SFTPサーバーディレクトリにデータファイルの非常に大きなリストが含まれている場合に使用される可能性が最も高くなります。

## 設定

- **ルートディレクトリ：**インポートおよびエクスポートのためにData Prepのブラウズインターフェイスに表示される最上位ディレクトリを定義します。ユーザーは、参照インターフェイスでこのディレクトリ内のファイルとディレクトリを表示できます。

## 認証

SFTPコネクターは、パスワード認証またはSSHキー(パスフレーズの有無にかかわらず)を使用して認証できます。オプションは次のとおりです。

- **ユーザー資格情報：**これはユーザー名とパスワードの組み合わせです。
  - *ユーザー名：*SFTPサーバーでの認証用のユーザー名です。
  - *パスワード：*指定されるユーザー名に関連付けられたパスワードです。
- **パスフレーズなしのSSHキー：**このオプションでは、SSHキーを貼り付けるだけで済みます。
  - *ユーザー名：*SFTPサーバーでの認証用のユーザー名です。
  - *SSHプライベートキー：*ユーザー名に関連付けられているSSH秘密キーの内容です。
- **パスフレーズ付きのSSHキー：**SSHキーを貼り付け、パスフレーズを入力します。
  - *ユーザー名：*SFTPサーバーでの認証用のユーザー名です。
  - *SSHプライベートキー：*ユーザー名に関連付けられているSSH秘密キーの内容です。
  - *パスフレーズ：*秘密キーの暗号化パスフレーズです。

# データのインポート／エクスポート情報

## ブラウジング経由

- コネクターは、ROOT DIRECTORYフィールドで定義されたロケーションから始まるブラウズ可能なディレクトリ階層を表示します。
- コネクターはワイルドカードとグロブのインポートもサポートしているため、ユーザーは複数のSFTPデータファイルを単一のデータセットとしてData Prepにインポートできます。

## SQLクエリー経由

- SFTPはファイルストアであるため、このデータソースではSQLクエリはサポートされていません。

# 技術的な仕様

- OpenSSHの設定されていない標準のLinux実装に対してこのコネクターをテストします。

# FAQ／トラブルシューティング／一般的な問題

SFTPはストレージの一種であると同時にプロトコルであることに注意してください。「SFTPサーバー」を使用している場合、実際に使用できるのは、SSHファイル転送プロトコルを使用してWebとインターフェイスするストレージロケーションです。これは重要な違いです。何でも（Webサービス、SFTPサービスプロバイダーなど）このプロトコルを使用してデータをWebに公開できるからです。これらのサービスは、SFTPの異なる実装を使用している場合や、従来のSFTPサーバーが実行しないことをバックグラウンドで実行している場合があります。これはすべて、SFTPサーバーがデータの接続またはインポートのいずれかで課題となるカスタム動作を行う可能性があることを意味します。

標準SFTPからのこのタイプの分散が、いくつかの問題を引き起こした例があります。顧客はSFTPコネクターを使用してベンダーの1社からデータを取得していました。ベンダーは、SFTPを介してデータを公開するサービスを使用していましたが、読み取られた後に各データファイルを削除していました。Data Prepがインポート時にデータのプレビューを提供する場合、これは、存在するデータの小さなチャンクについてデータソースにクエリを実行することによって行われます。これにより、システムはファイルを完全にインポートする前にファイルを削除しました。

# Data Prep用のSnowflakeデータウェアハウスコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

Snowflakeコネクターを使用すると、ライブラリのインポートとエクスポートにJDBCベースの接続を使用できます。次のフィールドは、接続を作成するために使用されます。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明：** UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のSnowflakeアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### データベースURI

- **JDBC URI：**JDBC接続文字列。スキーマ名はURIに含めることができます。詳細については、 https://docs.snowflake.net/manuals/user-guide/jdbc-configure.html を参照してください。

## データベース、スキーマ、およびテーブルの可視性

インポート中にユーザーがデータソースを参照するときに表示されるデータベース、スキーマ、およびテーブルを制御できます。それぞれについて、次のいずれかを選択できます。

- [表示のみ]ここで指定したデータベース、スキーマ、またはテーブルだけが返されます。
- [非表示]：ここで指定したデータベース、スキーマ、テーブルが非表示になります。
- [すべて表示]：データソース内のすべてを表示するデフォルト設定です。

「表示のみ」または「非表示」オプションを選択すると、オプションを適用するデータベース、スキーマ、または表のコンマ区切りリストを指定するためのフィールドが提供されます。

> **備考**
>
> これらの設定は、ユーザーがSQLを使用してデータソースに対してクエリを実行する場合は適用されません。クエリの結果は、一致の完全なリストを返します。たとえば、特定のデータベースを[非表示]にした場合でも、ユーザーはそのデータベース内のテーブルからデータをプルするクエリーを実行できます。ただし、そのデータベースは、ユーザーがデータソースを参照するときに表示されません。

## インポート設定

- **クエリに対するクエリのプリフェッチサイズ:** インポートでデータのプリフェッチを行うときに使用するバッチ サイズ（行数）。
- **最大列サイズ：** データをインポートまたはエクスポートするときに許容されるセルの最大サイズ（Unicode 文字数）。
- **インポート前のSQL：** テーブルのスキーマを決定した後、インポートの開始前に実行するSQLステートメント。このSQLは、データがData Prep UIでプレビューされる前にも実行されます。
- **インポート後のSQL：** インポート完了後に実行するSQLステートメント。このSQLは、データがData Prep UIでプレビューされる前にも実行されます。

## エクスポート設定

- **エクスポート方法**：Snowflakeからデータをエクスポートする方法を選択します。これらの方法は、いずれもSnowflakeがデータをインポートする方法に固有のものです。これらのオプションの詳細については、リストアップされたオプションの後ろにリンクされているSnowflakeのドキュメントを参照してください。次の2つのオプションがあります。
- 内部ステージ：データを表にロードする前に、Snowflakeの内部ステージのファイルにデータを書き込みます。この方法は、ダイレクトSQLよりも高速であるため、大規模なデータセットに推奨されます。**ステージの種類：**
  - **臨時：**作成されたステージは、作成されたセッションの終了時に削除されます。ステージはSnowflakeによって管理されるため、追加の設定は必要ありません。
  - **常設：**Snowflakeですでに作成されているステージの名前を指定します。
    - **ステージ名：**Snowflakeの既存の名前付き内部ステージの名前を指定します。Snowflakeの 識別子の構文を参照してください。

- ダイレクトSQLT</>：SQLの挿入ステートメントを使用してデータをエクスポートします。データ量の多いデータセットの場合、このアプローチは内部ステージを使用した場合よりも遅くなります。
  - **エクスポートバッチサイズ：**ダイレクトSQLのエクスポート方法が選択されている場合でデータをエクスポートする際に使用されるバッチサイズ。
- **最大 VARCHAR サイズ：**エクスポート時に許可されるVARCHAR列で使用できる最大サイズです。このサイズより大きい値は、データがSnowflake表にロードされるときにヌル値に置き換えられます。
- **テーブルを自動作成：**エクスポート時に新しいテーブルを自動作成します。有効にした場合、Data Prepは、エクスポートされたデータセットと名前が一致するテーブルを削除し（存在する場合）、エクスポートされたデータセットを使用してテーブルを再作成します。有効になっていない場合、Data Prepは新しい表を作成しませんが、代わりに、エクスポートされたデータを、エクスポートされたデータセットの名前と一致する名前の表にロードします。
- **エクスポート前のSQL：**自動作成が有効になっている場合、テーブルの作成後、エクスポートの開始前に実行するSQLステートメント。
- **エクスポート後のSQL：**エクスポートの完了後に実行するSQLステートメント。

**備考**

「TIMESTAMP_LTZ(9)」タイプを使用してData Prepをエクスポートします。別のタイムスタンプを使用して表が作成された場合、タイムスタンプのタイプが一致しない列にData Prepのデータをエクスポートすると、エラーが発生します。エラーの内容は以下の通りです。「エクスポートの実行中にエラーが発生しました。理由：SQLのコンパイルエラーです。式タイプが列データ型と一致しません。TIMESTAMP_####を期待していましたが、TIMESTAMP_LTZ(9)が列_Column_Nameに含まれていました。」

これを修正するには、次のいずれかを実行します。

- Data Prepに表の作成を許可してからエクスポートを実行するか、または
- TIMESTAMP_LTZ(9)でテーブルを作成してから、エクスポートを実行します。

## 資格情報

ユーザー認証は、共有アカウントまたは個人アカウントを介して行うことができます。個人アカウントでの認証を選択した場合、ユーザーはこのデータソースにアクセスするためにユーザー名とパスワードを入力するように求められます。共有アカウントで認証する場合は、次のフィールドを設定する必要があります。

**ユーザー：**データベースでの認証に使用される共有アカウントのユーザー名。

**パスワード：**データベースへの認証に使用される共有アカウントのパスワード

**ロール：**「ユーザーロール」を使用するセッションに設定するロール。指定するロールは、ユーザーに既に割り当てられた既存のロールでなければなりません。ロールはJDBC URIを使用して指定することもできますが、Roleフィールドで指定した値は、URIで指定したロールよりも優先されます。Snowflakeのロールに関する情報については、https：//docs.snowflake.net/manuals/sql-reference/sql/use-role.htmlを参照してください。

# データインポート情報

## ブラウジング経由

上で選択したデータベース、スキーマ、およびテーブルの可視性設定と、指定されたユーザー資格情報に基づき、ブラウジングエクスペリエンスは異なります。

## SQLクエリー経由

データベース、スキーマ、および表の可視性のセクションに記載したように、ユーザーがクエリを介してインポートできるものの制限は、接続用に提供された認証情報によって決定される認証にのみ制限されます。

クエリは、次で定義されている正当なSQL Selectステートメントを使用して実行できます：https：//docs.snowflake.net/manuals/sql-reference/sql/select.html

例:

SELECT * FROM "SNOWFLAKE_SAMPLE_DATA"."TPCH_SF1"。お客様"

# Data Prep用Tableau Hyperコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクタを使用すると、エクスポート先としてTableauに接続できます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のTableauアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### Tableauの設定

- **TableauサーバーのURL**：Tableauサーバーが稼働するURL（httpまたはhttps）。必要な場合はポートを含めます。Tableau Onlineの場合、URLには10az、10ay、またはus-east-1などのポッド名が含まれている必要があります。
- **サイト**：Data Prepがデータソースを公開するTableauサイト。
- **プロジェクト**：Data Prepがデータソースを公開するTableauプロジェクト。
- **Tableauのデータソースを上書きする**：同じ名前のデータソースが既に存在する場合は、上書きします。

## Tableauの資格情報

- **ユーザー名：**Tableauで認証するためのユーザー名またはメールアドレス。
- **パスワード：**Tableauで認証するためのパスワード。

## Webプロキシ

プロキシサーバーを介してTableauに接続する場合、これらのフィールドはプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」、プロキシサーバー経由でTableau RESTエンドポイントに接続する場合は「プロキシ」。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Webプロキシサーバーのホスト名またはIPアドレス。
- **プロキシポート：**データソースのプロキシサーバー上のポート。
- **プロキシユーザー名：**プロキシサーバーのユーザー名。
- **プロキシパスワード：**プロキシサーバーのパスワード。 認証されていないプロキシ接続では、ユーザー名とパスワードを空白のままにします。

# データエクスポート情報

## ブラウジング経由

ファイルは、Tableauコネクター/データソース設定で定義された指定されたサイトおよびプロジェクトに公開されます。

ブラウジングUIにはディレクトリは表示されません。エクスポート表示の[選択]ボタンをクリックして、指定したサイトとプロジェクトにエクスポートします。

AnswerSetはTableau Hyper Extract (.hyper)ファイルに変換され、ここで定義する指定されたサイトとプロジェクトに公開されることに注意してください。

## SQLクエリー経由

サポートされていません。

# 技術的な仕様

## デプロイシナリオ

Tableau Hyperコネクターは、Tableau SDKのネイティブライブラリを使用して.hyperファイルを書き込みます。これらのネイティブライブラリを使用できるようにするサポートされているデプロイシナリオが2つあります。

1. コネクターにバンドルされているネイティブライブラリを使用します。これが推奨されるアプローチです。
2. Tableau SDKをホストシステムにインストールし、コネクターzipから抽出されたネイティブライブラリを削除します。

追加情報が必要な場合や、Tableau Hyper ConnectorのSDKをインストールする上でサポートが必要な場合は、カスタマーサクセス担当者にお問い合わせください。

# FAQ／トラブルシューティング／一般的な問題

## Tableauコネクターが2つあるのはなぜですか？違いは何ですか？

Tableau「.tde」は古い抽出タイプであり、Tableauの抽出APIのバージョン1.0で実装されました。Tableauの「.hyper」コネクターはより新しく、Tableauの抽出APIのバージョン2.0に基づいています。両方のコネクターがアカウントに共存できるように、「.hyper」コネクターを完全に新しいコネクターとして構築したため、実行しているTableauのバージョンが多数ある顧客は、Tableauの各インスタンスに最適なコネクターを選択することができます。

# Data Prep用のTableau TDEコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、エクスポート先としてTableauサーバーおよびTableau Onlineに接続できます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のSnowflakeアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### Tableau サーバー

- **Tableau サーバーの URL：**Tableau サーバーが実行される URL（http または https）です。必要な場合はポートを含めます。
- **サイト：**Data Prepがデータソースを公開するTableauサイト。
- **プロジェクト：**Data Prepがデータソースを公開するTableauプロジェクト。

### Tableauの資格情報

- **ユーザー名：**これはTableauのユーザー名です。

. **パスワード：**これはTableauのパスワードです。

# データエクスポート情報

## ブラウジング経由

- ブラウジングインターフェイスにはTableauのディレクトリアイテムが表示されないことに注意してください。ファイルは、エクスポート時に選択された宛先ではなく、Tableauコネクター/データソース設定で定義された指定されたサイトおよびプロジェクトに公開されます。
- AnswerSetは、Tableauデータ抽出(.tde)ファイルに変換され、コネクター/データソース設定で定義した指定されたサイトおよびプロジェクトに公開されることに注意してください。

## SQLクエリー経由

- サポートされていません。

# 技術的な仕様

- Tableauサーバーとの通信を可能にし、TDEファイルをエクスポートするには、Tableau SDK(ネイティブライブラリ)をData Prepサーバーホストにインストールする必要があります。
- Tableauコネクター SDK

# FAQ／トラブルシューティング／一般的な問題

- Data PrepからTableauにエクスポートするときにTableauパーミッションの問題が発生することは珍しくありません。

# Data Prep用のTeradataコネクター

**ユーザーペルソナ**：Data Prep管理者またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

Teradataに接続する機能は、Data Prep JDBCコネクターの一部です。このトピックでは、Teradataへの接続のセットアップに固有の詳細を説明します。接続を設定するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

jdbc:teradata://*yourTeradataHost*/TMODE=ANSI,COP=OFF

# 技術的な仕様

## ドライバー仕様

- **Teradata Databaseドライバー名とバージョン：**
  - ドライバーのクラス名：com.teradata.jdbc.TeraDriver
  - バージョン：16.20.00.12
- **サポートされているTeradataデータベースのバージョン：**
  - 16.20、16.10、15.10、15.0、14.10

## ドライバーのドキュメント

- **一般的なドライバーのドキュメント：** https://teradata-docs.s3.amazonaws.com/doc/connectivity/jdbc/reference/current/frameset.html
- **ドライバーに関するよくある質問：** https://teradata-docs.s3.amazonaws.com/doc/connectivity/jdbc/reference/current/faq.html

# Data Prep用のThoughtSpotコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Data Prep AnswerSetをThoughtSpotにエクスポートできます。次のフィールドを使用して、接続パラメーターを定義します。

### 一般

- **名前**：UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepを複数のThoughtSpotインスタンスに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。

### ThoughtSpot サーバーの構成

- **ホスト名：** ThoughtSpot サーバーのホスト名または IP アドレスです。
- **ポート**：ThoughtSpotサーバーポート。標準ポートは22です。

### ThoughtSpot ユーザーの構成

- **ユーザー**：ThoughtSpot サーバーのユーザー。（これはSSH ユーザー名であり、ThoughtSpot Webアプリのログインに使用するユーザー名ではありません。ログイン資格情報の詳細については、Thoughtspotドキュメントを参照してください。）

- **パスワード：** ThoughtSpot サーバーのユーザーのパスワード。

## エクスポート設定

- **既存のテーブルにエクスポートする場合の動作：** エクスポートをする際に、指定したデータベースとスキーマに既に同じ名前のテーブルが既に存在する場合の Data Prepの動作を選択します。
  - **新しいデータを既存の表に追加する：**Data Prepは、データの新しい行を既存の表に追加します。
  - **既存テーブルを削除して新規作成する：** Data Prepはすべての既存の行と列を削除し、新しいテーブルを作成して新規データを追加します。
  - **ターゲット表を空にして新しいデータを追加する：**Data Prepは、既存の表を保持しますが、行の値を削除して新しいデータを入力します。

> **備考**
>
> エクスポートで指定された名前に一致する既存の表がない場合、すべてのオプションは同等です。それらはすべて新しい表を作成し、データセットからすべての行を読み込みます。

# データエクスポート情報

## ブラウジング経由

ThoughtSpot内のデータベースとスキーマのリストを表示します。

## SQLクエリー経由

サポートされていません。

# ベストプラクティス

コネクター構成およびデータソース構成のセットアップ中に、エクスポートの動作を選択するオプションがあります。これらの設定は一度構成され、何度も使用されます。これにより、1つのエクスポートでテーブルを切り捨て、もう1つを追加する場合は、この設定を空白のままにして1つのコネクター構成を作成してから、2つのデータソース構成を作成する必要があります。1つは追加動作を指定し、もう1つは切り捨て動作を指定します。データソースには明確な名前を付けてください。

# Data Prep用のVerticaコネクター

**ユーザーペルソナ**：Data Prep管理者またはデータソース管理者

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

HP Verticaに接続する機能は、Data Prep JDBCコネクターの一部です。このトピックでは、Verticaへの接続のセットアップに固有の詳細を説明します。接続を構成するには、JDBCコネクターのドキュメントも参照してください。

### JDBC URIの例：

`jdbc:vertica://yourVerticaHost:_yourVerticaPort/yourDatabaseName_`

## 技術的な仕様

### ドライバー仕様

- **HP Vertica用ドライバーのバージョン：**9.2.1
- **サポートされているHP Verticaのバージョン：** 9.2.x

### ドライバーのドキュメント

- **一般的なドライバーのドキュメント：** https：//www.vertica.com/docs/9.2.x/HTML/Content/Authoring/ConnectingToVertica/ClientJDBC/ProgrammingJDBCClientApplications.htm
- **SQLリファレンス：** https：//www.vertica.com/docs/9.2.x/HTML/Content/Authoring/SQLReferenceManual/SQLReferenceManual.htm

# Data Prep用のZendeskコネクター

**ユーザーペルソナ**：Data Prepユーザー、Data Prep管理者、データソース管理者、またはIT/DevOps

**備考**

この文書は、コネクターの設定中に利用できるすべての構成フィールドについて論じています。 一部のフィールドは、設定の以前の手順で管理者によってすでに入力されている場合があり、表示されない場合があります。 Data Prepのコネクターフレームワークの詳細については、Data Prepコネクターのセットアップを参照してください。 また、管理者がこのコネクタにデータソースのリストで別の名前を付けている可能性があります。

## Data Prepの設定

このコネクターを使用すると、Zendeskに接続して、使用可能なデータの閲覧とインポートを行うことができます。次のフィールドは、接続パラメーターを定義するために使用されます。

### 一般

- **名前：** UIでユーザーに表示されるデータソースの名前。
- **説明**：UIでユーザーに表示されるデータソースの説明。

**ヒント**

Data Prepは複数のZendeskアカウントに接続できます。わかりやすい名前を使用すると、ユーザーが適切なデータソースを識別する上で非常に役立ちます。Data Prep SaaSを使用している場合は、Data Prep DevOpsにこのセットの希望の旨をお知らせください。

### Webプロキシ

プロキシサーバーを介してZendeskに接続する場合、これらのフィールドはプロキシの詳細を定義します。

- **Webプロキシ：**プロキシが不要な場合は「なし」、プロキシサーバー経由でZendesk RESTエンドポイントに接続する必要がある場合は「プロキシ」にします。Webプロキシサーバーが必要な場合、プロキシ接続を有効にするには以下のフィールドが必要です。
- **プロキシホスト：**Webプロキシ サーバーのホスト名またはIPアドレス。
- **プロキシポート：**データソースのプロキシサーバー上のポート。
- **プロキシ ユーザー名：**プロキシ サーバーのユーザー名です。

- **プロキシ パスワード：**プロキシ サーバーのパスワードです。 *認証されていないプロキシ接続の場合は、ユーザー名とパスワードを空欄にしてください。

## Zendesk の構成

- **Zendesk URL：** [https://your-domain.zendesk.com] の形式をしたZendesk URLです。
- **ユーザー名：** Zendesk に接続するユーザーのメールアドレスです。
- **認証タイプ**: 使用する認証タイプ（パスワードまたは API トークン）です。
  - **パスワード：**Zendeskに接続するためのパスワードです。認証タイプとしてパスワードを選択した場合に提供されます。
  - **API トークン：**Zendesk に接続するための API トークンです。認証タイプとしてAPIトークンを選択した場合に提供されます。
- **タイムアウト：** タイムアウトエラーによって、実行中の操作が取り消されるまでの待機時間（秒単位）です。

# データインポート情報

## ブラウジング経由

テーブルを参照し、インポートするテーブルを「選択」します。

## SQLクエリー経由

正当なSQL選択クエリの使用

# Zendeskオブジェクト

| 名前 | 説明 |
|---|---|
| アカウント設定 | Zendeskのアカウント設定 |
| アクティビティストリーム | Zendeskのアクティビティストリーム。 |
| アプリのロケーション | Zendeskのアプリのロケーション。 |
| 添付ファイル | Zendeskでチケットの添付ファイルを表示します。 |
| 自動化 | Zendeskの自動化。 |

| 名前 | 説明 |
| --- | --- |
| ブランド | Zendeskのブランド。 |
| コラボレーター | Zendeskのコラボレーター。 |
| カスタムエージェントロール | Zendeskのカスタムエージェントロール。 |
| グループメンバーシップ | Zendeskのグループメンバーシップ。 |
| グループ | Zendeskのグループ。 |
| 休日 | Zendeskのスケジュール。 |
| ジョブステータス | 複数のチケットの更新など、誰かがジョブを開始すると、ステータスレコードが作成されます。 * 特定のジョブが作成された後、1時間はジョブステータスのデータにアクセスできますが、その後はデータを使用できなくなります。 |
| ロケール | Zendeskのロケール。 |
| マクロ | Zendeskのマクロ。 |
| 監視対象のTwitterハンドル | Zendeskで監視対象のTwitterハンドル。 |
| 組織フィールド | Zendeskの組織フィールド |
| 組織メンバーシップ | Zendeskの組織メンバーシップ。 |
| 組織サブスクリプション | Zendeskの組織サブスクリプション。 |
| 組織 | Zendeskの組織。 |
| リクエスト | Zendeskでのリクエスト。 |
| 満足度評価 | Zendeskでのリクエスト。 |
| スケジュール | Zendeskのスケジュール。 |

| 名前 | 説明 |
| --- | --- |
| セッション | Zendeskのセッション。 |
| 共有契約 | Zendeskの共有契約。 |
| SLAポリシー | ZendeskのSLAポリシー。 |
| サポートアドレス | Zendeskのサポートアドレス。 |
| 一時停止中のチケット | Zendeskで一時停止中のチケット。 |
| チケット監査 | Zendeskのチケット監査。 |
| チケットコメント | Zendeskの指定されたチケットに属するチケットコメント。 |
| チケットフィールド | Zendeskのチケットフィールド。 |
| チケットフォーム | Zendeskのチケットフォーム。 |
| チケットのメトリックイベント | Zendeskのチケットのメトリックイベント。 |
| チケットメトリクス | Zendeskのチケットメトリクス。 |
| チケット | チケット。 |
| トリガ | Zendeskのトリガ。 |
| ユーザーフィールド | ユーザーフィールド。 |
| ユーザーID | ユーザーID。 |
| ユーザー関連情報 | Zendeskのユーザー関連情報。 |

| 名前 | 説明 |
| --- | --- |
| ユーザー | Zendeskのユーザー。 |
| ビュー | Zendeskのビュー。 |

# データセットの操作

これらのセクションでは、Data Prepでデータセットを管理する方法を説明します。

| トピック | 説明... |
| --- | --- |
| データセットのインポート | データソースまたはローカルファイルからデータをインポートする方法。インポートする前に調整できる設定について説明します。 |
| データセットのエクスポート | データセットとAnswerSetをエクスポートする方法。 エクスポートする前に調整できる設定について説明します。 |
| プロファイルデータセット | データの品質に関する情報を含むデータセットのプロファイルを生成する方法。 |
| 新しいデータでデータセットを更新 | 既存のデータセットのデータを更新する方法。 |
| プロジェクトデータセットの更新 | **ステップ**ツールを使用して、データセットをリフレッシュして置換する方法を説明します。 |

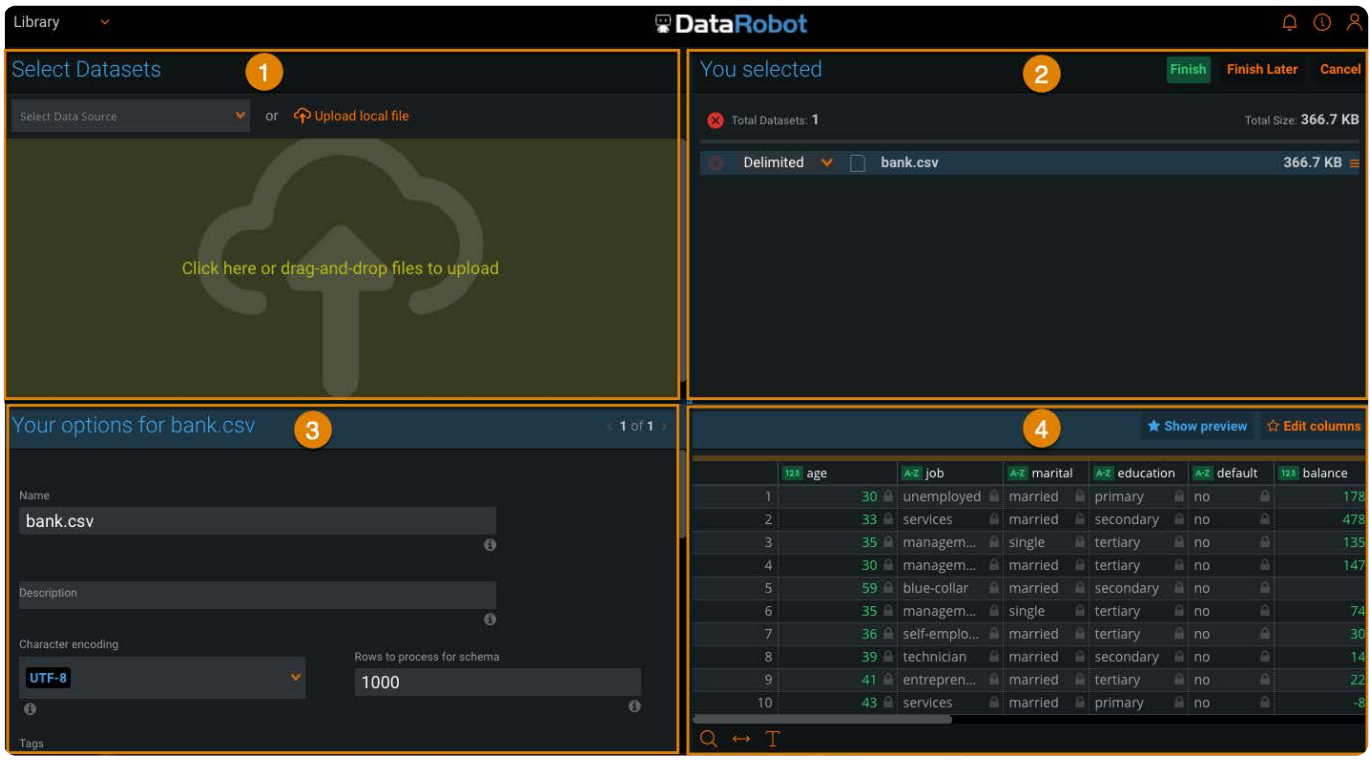
# データセットのインポート

データをData Prepにインポートすることは、データを機械学習用に準備するための最初のステップです。 インポートプロセス中は、次を行うことができます。

- さまざまなデータソースから複数のデータセットを選択します。
- 複数のデータセットを1つのデータセットに結合します。
- インポートするデータセットの列を選択します。
- 拡張子のないファイルを選択します。
- zip形式の（圧縮された）フォルダーのデータセットをインポートします。
- データの分析と構造化に使用する形式を変更します。

## インポートページの使用

インポートするデータセットを選択すると、ページがペインと呼ばれる4つのクアドラントに分割されます。




以下は、**インポート**ページの各ペインの概要です。

| # | 要素 | 説明 |
|---|---|---|
| 1 | **データセットの選択ペイン** | このペインからインポートするデータセットを選択します。以下を実行することが可能です。<br>・ローカルファイルや接続されたデータソースから、複数のデータセットを選択します。<br>・接続されたデータソースを検索し、データセットのクエリを行います。<br>・インポートのために、複数のデータセットを1つのグロブに結合します。 |
| 2 | **選択済みペイン** | データセットを選択すると、このペインにデータセットが表示されます。以下を実行することが可能です。<br>・インポートするために選択したデータセットのリストを表示します。<br>・プレビューするデータセットを選択し、インポートオプションを更新します。<br>・インポートエラーの可能性があるデータセットをすばやく特定します。<br>・インポートするデータの分析と構造化に使用する形式を変更します。<br>・さまざまなインポートオプションを使用して、同じデータセットを複数回インポートします。 |
| 3 | **オプションペイン** | 多くの場合、データはData Prepに簡単にインポートされます。場合によっては、インポートオプションを調整する必要があるか、調整したいことがあります。このペインでは、そのような調整を行います。 |
| 4 | **プレビューペイン** | ここでは、データのプレビューを行うことができます。**選択済み**パネルからデータセットを選択したり、形式を更新したり、インポートオプションを変更したりすると、**プレビュー**パネルには、選択したデータセットがインポートされたときにどのように見えるかが表示されます。このペインから、インポートする列を選択することもできます。 |

# インポートプロセスのスナップショット

以下は、データセットをData Prepにインポートする方法の簡単なスナップショットです。

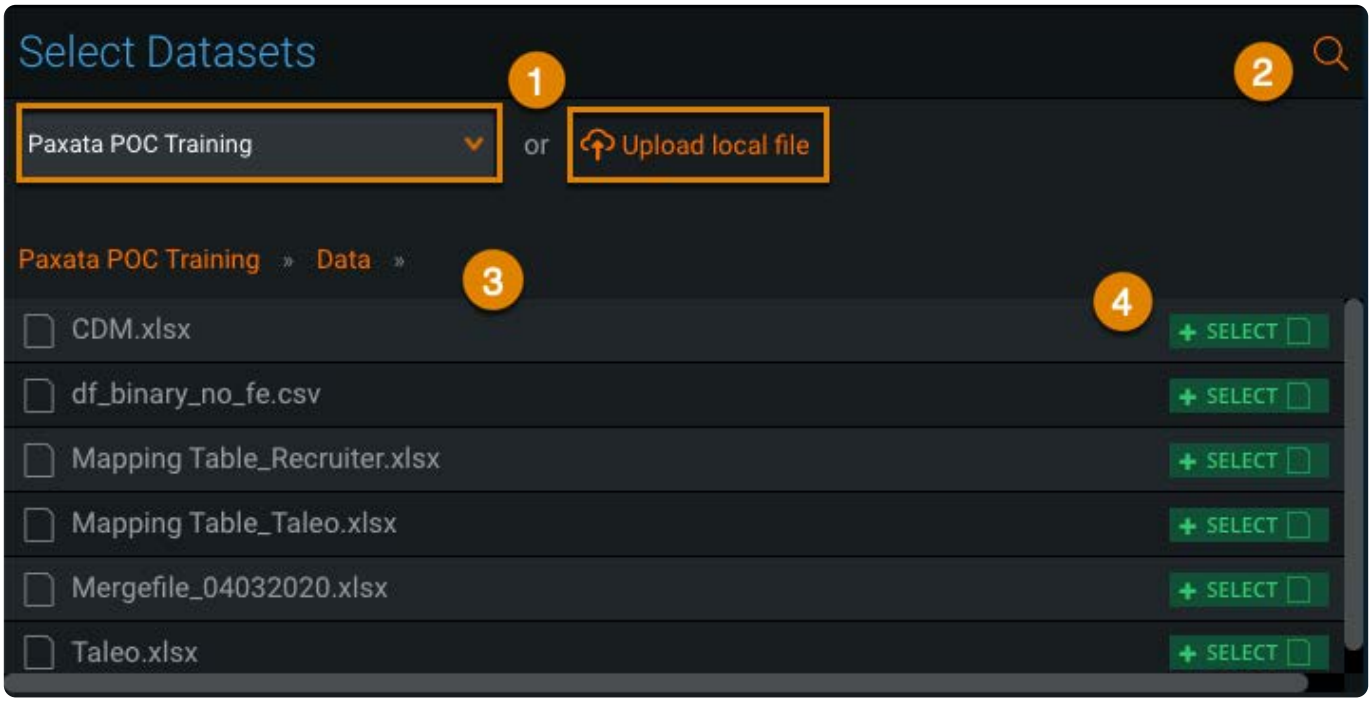
1. **ライブラリ**ページの上で、**インポート**をクリックします。
2. **インポート**ページで、データセットの選択、データセットの検索、またはデータセットの組み合わせを行うことができます。
3. データセットのプレビューを確認します。データは正しく表示されていますか？
   - データが正しければ、データセットの追加を続け、すべてのデータセットを選択してインポートすることができます。
   - データが正しくない場合は、インポート設定の調整を試してみてください。
4. **完了**をクリックします。

   データがデータセットとしてインポートされ、プロジェクトで準備を実行できるようになります。

# データセットを選択

データセットは、ローカル ファイルまたは接続されたデータソースからインポートできます。このセクションでは、インポートする1つ以上のデータセットを選択する方法について、詳しく説明します。

## データセットの選択ペイン




以下は、**データセットの選択**ペインの要素の概要です。

| # | 要素 | 説明 |
|---|---|---|
| 1 | **データ ソースオプション** | Amazon S3、Hadoop、JDBC、またはその他のデータソースからデータセットをインポートする必要がある場合があります。または、単にコンピューターに保存したスプレッドシートをインポートするだけの場合もあります。いずれにしても、ここからが本番です。<br><br>**データソース**リストでは、設定したデータソースを選択できます。管理者は、データソースに接続する必要があります。**ローカルファイルをアップロード**を<br><br>クリックして、コンピュータからデータセットを選択する必要があります。 |
| 2 | **SEARCH** | 特定のデータセット、または類似するデータセットのグループを検索する場合は、検索条件を入力できます。**検索**フィールドではワイルドカード文字を使用できます。これにより、特定の名前付きデータセットと同様の名前のデータセットが見つかります。詳細については、データセットを検索をみてください。 |

| # | 要素 | 説明 |
|---|---|---|
| 3 | **データセットリストペイン** | 選択したデータソースのコンテンツがここに一覧表示されます。この例では、データソースに6つの項目があります。1つのコンマ区切り値（CSV）ファイルと5つのExcelファイルです。 |
| 4 | **選択** | インポートするデータセットが表示されたら、**選択**をクリックします。データセットは**選択済み**ペインにリストされ、**完了**をクリックするとインポートされます。 |

## ローカルファイルからデータセットを選択します

コンピュータや共有ネットワークドライブ上のファイルからデータセットを選択するには：

1. **ファイルをアップロード**ペインをクリックして、データセットを選択するか、ファイルをペインにドラッグします。

   データセットが**選択済み**ペインのリストに追加されます。Data Prepでは、データセットの**オプション**ペインとデータセットのプレビューが表示されます。

2. さらにデータセットを追加するには、インポートに含める追加のデータセットをクリックします。

   追加のデータセットが、**選択済み**ペインのデータセットリストに追加されます。

## データソースからデータセットを選択する

接続したデータソースからデータセットを選択するには：

1. **データソースを選択**をクリックし、使用するデータソースを選択します。
2. インポートするデータセットを検索します。

   検索を実行してデータセットを特定する方法については、データセットの検索をご覧ください。

3. **選択**をクリックしてデータセットを選択します。

   データセットが**選択済み**ペインのリストに追加されます。Data Prepでは、データセットの**オプション**ペインとデータセットのプレビューが表示されます。

4. 現在選択されているデータソースからさらにデータセットを追加するには、インポートに含める追加のデータセットをクリックします。
5. 異なるデータソースからさらにデータセットを追加するには、データソースごとに手順1〜3を繰り返します。

追加のデータセットが、**選択済み**ペインのデータセットリストに追加されます。

# データセットの検索

データセットの名前を入力するか、クエリ文字列を入力すると、データセットを検索できます。検索では大文字と小文字が区別され、検索条件に完全に一致する結果のみが返されます。正確な名前がわからない場合や、同様の名前のデータセットを検索する場合は、ワイルドカード文字を使用してデータセットを検索できます。

# データセットの検索

データセットを検索するには：

1. データソースを選択し、**データセットの選択**ペインの右上にある**検索**アイコンをクリックします。

   **検索**アイコンは、ローカルファイルのアップロード時ではなく、データソースを選択したときにのみ表示されます。

2. **ワイルドカード検索**フィールドに検索条件を入力します。

   検索条件に完全に一致するデータセットが返されます。検索条件の設定については、ワイルドカード文字を参照してください。

# データベースのクエリ

データベースのクエリ：

1. **データソースを選択**をクリックして、使用するデータソースを選択します。
2. **データセットの選択**ペインの右下にある**クエリの作成**をクリックします。
3. **クエリ文字列**フィールドに検索条件を入力します。

   ワイルドカード文字で検索するには、ワイルドカード文字をご覧ください。

   検索条件に完全に一致するデータセットが返されます。

# ワイルドカード文字

以下に、データセットの検索に使用できるワイルドカード文字を示します。

| 文字 | 一致する内容 |
|---|---|
| * | 任意の数（0 を含む）の文字 |
| ? | 単一の文字 |
| [0-9] または [a-z] | かっこ内に指定されている範囲の文字 |
| [123] または [abc] | かっこ内にリストされている文字 |

## ワイルドカードを使用した検索の例

以下に、いくつかの検索例とその結果を示します。

| 検索例 | 戻り値 |
|---|---|
| * | すべてのデータセット |
| *.csv | ファイル拡張子が「.csv」のデータセット |
| a?b.csv | 「aac.csv」、「abc.csv」、...「azc.csv」という名前のデータセット |
| a*z.csv | 文字の種類や文字の数に関係なく、先頭が小文字の「a」で始まり、末尾が「z.csv」となるデータセット |
| a[0-9].csv | 「a0.csv」、「a1.csv」、「a2.csv」、...「a9.csv」という名前のデータセット |
| a[az].csv | 「aa.csv」、「ab.csv」、...「az.csv」という名前のデータセット |
| a[abc].csv | 「aa.csv」、「ab.csv」、「ac.csv」という名前のデータセット |

# データセットの結合

Data Prepでは、複数のデータセットを1つの*グロブ*に結合してインポートすることができます。グロブは、インポート中に複数のデータセットを1つのデータセットに追加した結果です。このセクションでは、インポートの前に複数のデータセットを1つのグロブに結合する方法について詳しく説明します。

## データセットの結合に関するガイドライン

以下は、複数のデータセットを1つのグロブに結合するためのガイドラインです。

- 同一のデータソースからのデータセットのみ、グロブを作成できます。
- ワイルドカード検索を使用する場合のみ、データセットのグロブを作成できます。
- 一緒にグロブが作成される各データセットは、構造（列の数とデータ型）が同じである必要があります。

## グロブの作成をサポートするデータソース

グロブの作成がサポートされているデータソースとファイル形式の一覧については、現在のData Prep（Paxata）リリースノートのプラットフォームサポートマトリックスを確認してください。

## グロブの作成

複数のデータセットを1つのグロブに結合するには：

1. **データソースを選択**をクリックして、データソースを選択します。
2. 検索を使用して、結合するデータセットを特定します。

3. **すべての結果の結合**をクリックします。

   データセットが1つのグロブに結合されます。このグロブは、**選択済み**ペインのデータセットリストに追加されます。グロブの名前は、デフォルトで検索条件になります。Data Prepでは、**グロブのオプション**ペインとグロブのプレビューが表示されます。

# インポート前にデータセットをプレビューする

プレビューでデータセットを変更するには、**選択済み**ペインから、プレビューするデータセットをクリックします。

**プレビュー**ペインに、選択したデータセットが表示されます。

Data Prepのデフォルトでは、最後に選択したデータセットのプレビューが表示されます。

# データセットの再追加

インポート中に、さまざまなインポートオプションを同じデータセットに適用したい場合があります。特に、同じExcelファイルから複数のExcelワークシートをインポートしなければならない場合などです。

さまざまなインポートオプションを含むデータセットを追加するには：

1. **選択済み**ペインから、再追加するデータセットの**その他**ボタン（縦に3つ並んだ点）をクリックします。
2. **再追加**をクリックします。

   データセットが**選択済み**ペインのリストに追加されます。
3. 必要に応じて、インポート設定を調整します。

# インポート設定の調整

データセットが選択されると、 Data Prepはデータを分析して最良の結果を得るための適切な設定を決定します。ただし、データはあらゆるものに適しているわけではありません。場合によっては、設定を適切なデータになるように微調整する必要があります。このセクションでは、インポートする前に、データセットのより一般的な設定のいくつかを調整する方法について説明します。設定に関する具体的な情報については、ヘルプのヒント（疑問符）の上にカーソルを置きます。

以下は、頻繁に使用する調整可能な基本設定の一部です。

| アクション | 手順 |
| --- | --- |
| タグを追加します。 | **オプション**ペインで、**タグ**リストからタグを入力または選択します。 |
| ソースファイルの系統を表示する列を追加します。 | **オプション**ペインで**ソースファイルを表示する列を追加**ボタンを切り替えます。<br><br>新しい**ソースファイル**列がデータセットの末尾に追加され、インポートした各行のソースファイルのパスを表示します。 |

| アクション | 手順 |
| --- | --- |
| データセットの形式を変更します。 | **選択済み**ペインで、**形式**メニューからデータセットに適用する形式を選択します。詳細については、サポートされている形式を参照してください。 |
| データセットの名前を変更します。 | **オプション**ペインで、**名前**フィールドに新しい名前を入力します。**選択済み**ペインでデータセット名を更新します。 |
| インポートから列を除外します。 | **プレビュー**ペインで、以下を実行します。<br>1. **列の編集**をクリックします。<br>2. インポートしない列の選択を解除します。<br>3. **プレビューの表示**をクリックします。<br><br>選択解除された列は**プレビュー**から削除されます。 |
| 同じExcelファイルから追加のワークシートをインポートします。 | 追加のワークシートごとに、次の手順を実行します。<br>1. **選択済み**ペインで、Excelファイルを再度追加します。<br>2. **オプション**ペインで、**ワークシート**メニューからインポート対象のワークシートを選択します。 |
| 列を並べ替えます。 | **プレビュー**ペインで、以下を実行します。<br>1. **列の編集**をクリックします。<br>2. 列が目的の位置にくるまで上矢印または下矢印をクリックします。<br>3. **プレビューの表示**をクリックします。 |
| 列名を変更します。 | **プレビュー**ペインで、以下を実行します。<br>1. **列の編集**をクリックします。<br>2. **編集**（鉛筆アイコン）をクリックして、新しい名前を入力します。<br>3. **プレビューの表示**をクリックします。 |

# サポートされている形式

ファイルベースのコネクターの場合の一般的な形式を次の表に示します。Data Prepのインテリジェント取込みは、ファイル拡張子に依存するのではなく、ファイルのコンテンツを調べることでファイルの形式を識別します。ファイルに拡張子がないか、間違った拡張子が付いている場合でも、Data Prepは形式を正しく識別します。

| 一般的な形式 | ワイルドカードとグロブのインポートサポート |
| --- | --- |
| 区切りファイル（CSV、TSVなど） | はい |

| 一般的な形式 | ワイルドカードとグロブのインポートサポート |
| --- | --- |
| 固定幅の列データ | はい |
| JSON | はい |
| xml | はい |
| Apache Avro | はい |
| Microsoft Excel（XLS、XLSX） | いいえ。ワイルドカード文字およびデータセットを組み合わせるためのガイドラインを参照してください。 |
| SAS BDAT | はい |

Data Prepは、Deflate、LZ4、Snappy、ZIP, Gzip、またはBzipのいずれかで圧縮されたファイルのインポートをサポートしています。一般に、解凍されたファイルは、前の表にリストされている一般的な形式である必要があります。

さらに、Parquetファイルをサポートするコネクターは、Parquetファイルの圧縮バージョンもサポートします。

**備考**

複数のファイルを含むZIPファイルをインポートするとき、圧入セット内の最大のファイルが自動的に識別され、ライブライリにインポートするために選択されます。

# データセットのエクスポート

データセットとAnswerSetは、Data Prepからコンピューター上のファイルまたは接続されているデータソースにエクスポートできます。データをエクスポートすると、他の人や他のシステムとデータを共有できます。

## データセットのエクスポート


データセットまたはAnswerSetをエクスポートするには：

1. **ライブラリ**ページで、エクスポートするデータセットにカーソルを合わせて、**エクスポート**をクリックします。
2. **エクスポート**ページで、**データソース**または**ローカルにダウンロード**をクリックします。
3. **エクスポート設定**ペインで、必要に応じて設定を調整します。
4. **エクスポート**をクリックします。

   データセットが指定した場所にエクスポートされます。

## エクスポート設定の調整

エクスポート設定では、エクスポートするデータセットの構造を定義できます。データセットに使用できる設定は、エクスポートで選択した形式によって異なります。

次は、頻繁に使用する調整可能な基本設定の一部です。

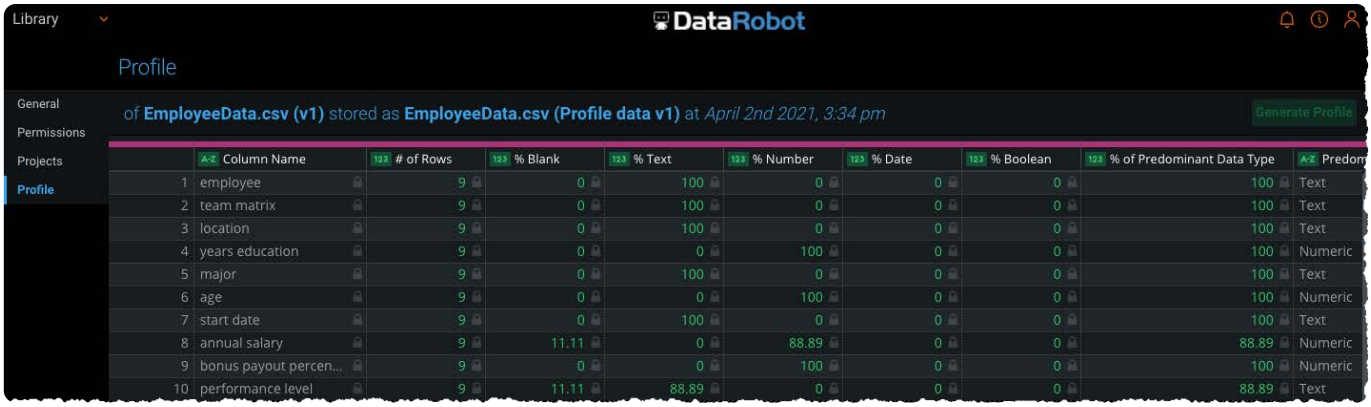
| アクション | 手順 |
|---|---|
| データセットの形式を調整します。 | **形式**メニューからデータセットに適用する形式を選択します。 |
| データセットの名前を更新します。 | **[名前]** フィールドに新しい名前を入力します。 |
| 使用する文字セットを更新します。 | **文字エンコーディング**メニューから別の文字エンコーディングを選択します。 |
| **注意**：英語の文字セットから日本語の文字セットに変更します。 | **備考**：Unicode署名BOMとも呼ばれるバイトオーダーマーク（BOM）は、とりわけUTF-16やUTF-32などの拡張文字セットに適用されます。 |

# プロファイルデータセット

**備考**

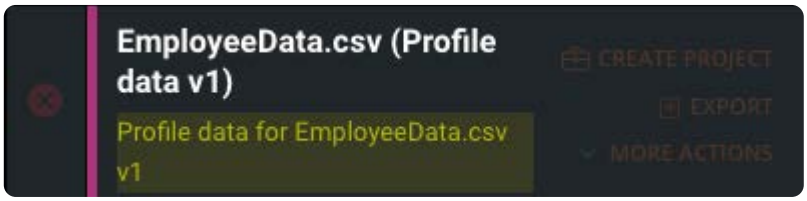
Data Prepの管理者は、アプリケーションでこの特徴量を有効化する必要があります。

データセットのプロファイルを作成すると、そのデータセットのデータに関する統計情報が生成されます。結果は、データセットの**[プロファイル]**ページに表示されます。




| | Column Name | # of Rows | % Blank | % Text | % Number | % Date | % Boolean | % of Predominant Data Type | Predom |
|---|---|---|---|---|---|---|---|---|---|
| 1 | employee | 9 | 0 | 100 | 0 | 0 | 0 | 100 | Text |
| 2 | team matrix | 9 | 0 | 100 | 0 | 0 | 0 | 100 | Text |
| 3 | location | 9 | 0 | 100 | 0 | 0 | 0 | 100 | Text |
| 4 | years education | 9 | 0 | 0 | 100 | 0 | 0 | 100 | Numeric |
| 5 | major | 9 | 0 | 100 | 0 | 0 | 0 | 100 | Text |
| 6 | age | 9 | 0 | 0 | 100 | 0 | 0 | 100 | Numeric |
| 7 | start date | 9 | 0 | 100 | 0 | 0 | 0 | 100 | Text |
| 8 | annual salary | 9 | 11.11 | 0 | 88.89 | 0 | 0 | 88.89 | Numeric |
| 9 | bonus payout percen... | 9 | 0 | 0 | 100 | 0 | 0 | 100 | Numeric |
| 10 | performance level | 9 | 11.11 | 88.89 | 0 | 0 | 0 | 88.89 | Text |

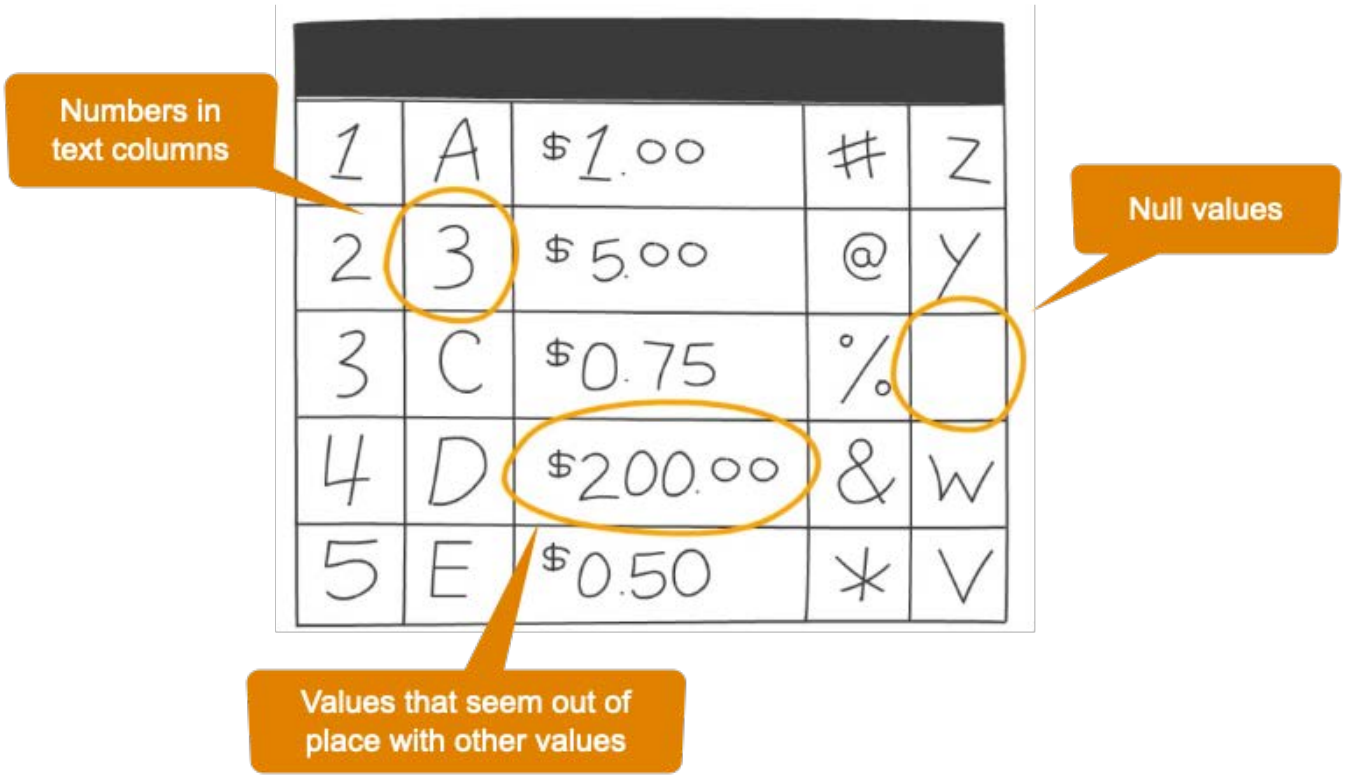
また、作成されたプロファイルは、プロファイルタイプのAnswerSetであることを示す名前でライブラリに自動的に保存されます。




## データのプロファイルの使用方法

データの取得は、パッケージの内容がわからないだけで、パッケージの取得に似ています。パッケージには、パッケージスリップが含まれているので、その内容を知るためにすべてを掘り起こし、解剖する必要はありません。Data Prepがあれば、データのプロファイルを作成することができるので、すぐに理解できます。

データプロファイルは、データセットの中のデータの品質を、そのデータを扱う前に判断する上で欠かせません。たとえば、データ、Null、印刷不可能な文字、例外的なパターンに混合されたタイプがある場合、素早く判定できます。




データプロファイルに基づいて、データをData Prepプロジェクトに取り込むことで品質の問題に対処できます。

ライブラリのデータセットのバージョンを手動または自動のインポートによって更新し続けながら、個々の後続バージョンのプロファイルを継続的に作成できます。このようにして、バージョンごとにデータセットのデータ品質を監視し、必要に応じて修正することができます。

## プロファイルのAnswerSetに表示される各列の意味

データセットのプロファイルを作成すると、データセットの各列を表す行を含むAnswerSetが生成されます。プロファイルAnswerSetの各列は、データセットの列に関する以下の統計情報を提供します。

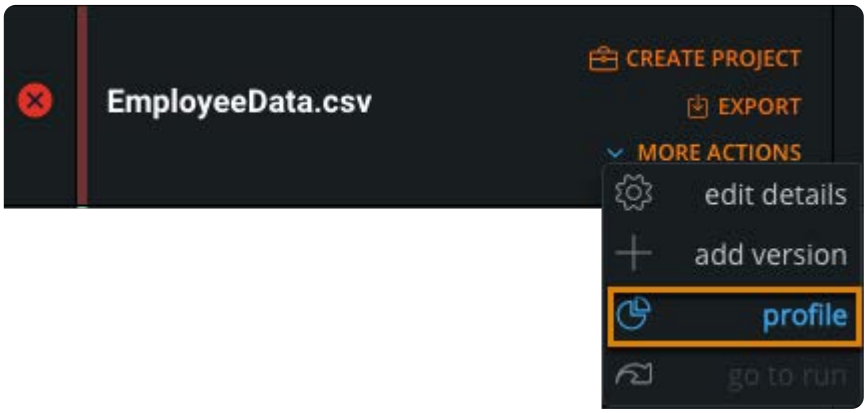
以下は、各列のデータプロファイルに含まれる統計です。

| 列名 | 定義 |
| --- | --- |
| 行数 | データセット内の合計行数 |
| 空白の割合（%） | 列に含まれる空白の割合 |
| テキストの割合（%） | 列に含まれるテキスト値の割合 |
| 数字の割合（%） | 列に含まれる数値の割合 |
| 日付の割合（%） | 列に含まれる日付値の割合 |

| 列名 | 定義 |
| --- | --- |
| ブーリアンの割合（%） | 列に含まれるブール値の割合 |
| 主たるデータ型の割合（%） | 列に含まれる、最も支配的なデータ型を持つ値の割合 |
| 主たるデータ型 | 列内で最も支配的なデータ型 |
| 一意の値の数 | 列内の一意の値の数 |
| 音声学的に一意の値の数（Metaphone） | metaphone（発音近似）アルゴリズムを使用して類似した値をクラスターした後の、列内の一意の値の数。たとえば、「Good Samaritan」と「Good Samertitan」は同じ値としてカウントされます。 |
| 音声学的に音声が重複している割合（Metaphone）（%） | 音声学的に一意の値の数（Metaphone）/一意の値の数。この比率は、列に重複がある可能性を示します。数値が大きいほど、重複する値が含まれる可能性は高くなります。その場合、重複する値を識別するために列に対してCluster + Edit操作を行う必要があります。 |
| トップ5 | 列に最も多く含まれる値の上位5件 |
| 最小文字列長 | 列内の最短文字列の長さ |
| 最大文字列長 | 列内の最長文字列の長さ |
| 平均文字列長 | 列内の文字列の平均文字列の長さ |
| NA、NONE、またはNullの数 | 列に「na」、「none」、「null」が含まれる回数 |
| すべて大文字の割合（%） | すべて大文字の文字列を含むセルの割合 |
| すべて小文字の割合（%） | すべて小文字の文字列を含むセルの割合 |
| 非標準のASCII文字を含む割合（%） | 制御文字などの印字不可能な文字を含むセルの割合 |
| HTMLタグを含む割合（%） | HTMLタグを含むセルの割合 |
| 連続する空白の平均数 | 列に含まれる連続する空白の平均数 |

| 列名 | 定義 |
| --- | --- |
| 負の数の割合（%） | 負の数を含むセルの割合 |
| ゼロの割合（%） | ゼロの値を含むセルの割合 |

# データセットのプロファイルを作成する




データプロファイルを作成するには：

1. **ライブラリ**ページで、データプロファイルを作成するデータセットにマウスを置きます。
2. **その他のアクション**をクリックし、**プロファイル**を選択します。
3. **プロファイル**ページで、右上にある**プロファイルの生成**をクリックします。

   プロファイルが**プロファイル**ペインに表示されます。さらに、作成されたプロファイルはAnswerSetとしてライブラリに自動的に保存されます。

**備考**

AnswerSetのライブラリプレビューは、プロファイルの最初の100行に制限されています。

# 新しいデータでデータセットを更新

データは絶えず変動しています。たった今Data Prepにインポートしたデータであっても、既に鮮度が失われている場合もあります。データセット内のデータを更新することで、データセットを既存のデータセットの最新版としてインポートすることができます。データセットの更新後は、最新版を既存のプロジェクトで使用することができます。

データセットの更新時に、新しい値、構造、および形式を使用して、完全に異なるデータセットに更新することもできます。または、次のデータセットに更新できます。

- 値のみが変更されているデータセット（構造と形式は変更されていないもの）。
- 形式または構造が変更されました。たとえば、列が追加または削除されました。

## 新しいデータのデータセットの更新

新しいデータでデータセットを更新するには：

1. **ライブラリ**ページで、更新するデータセットにカーソルを合わせて、**その他のアクション**をクリックします。
2. **バージョンの追加**を選択します。
3. **データソースを選択** リストからインポートするデータセットを見つけて選択するか、**ローカルファイルをアップロード**をクリックします。

   **備考：**最初のインポートを行った際に SQL ステートメントを使用している場合、この SQL ステートメントは保持されており、データセット内のデータの更新に再利用することができます。

   データセットが**選択済み**ペインのリストに追加されます。Data Prepでは、データセットの**オプション**ペインとデータセットのプレビューが表示されます。
4. データセットのプレビューを確認し、必要に応じてインポート設定を調整します。
5. **完了**をクリックします。

   データが新バージョンとしてインポートされ、プロジェクトで準備できるようになります。

# プロジェクトデータセットの更新

データセットをプロジェクトに追加するとき（ベースのデータセットとして追加するとき、またはルックアップ/追加機能を通じて追加するとき）は、プロジェクトで使用するデータセットの特定のバージョンを選択します。データセットのより新しいバージョンがライブラリで使用可能になっても、プロジェクトでそれらの新しいバージョンが自動的に使用されることはありません。プロジェクトで既に行った作業、およびその後の結果は、最初に選択した特定のデータセットのバージョンによって異なる場合があります。

これは何回も正常に機能します。場合によっては、新しいバージョンでプロジェクトのデータセットを更新しなければならないこともあります。

プロジェクトのデータセットを更新する方法は2通りあります。

- プロジェクトのデータセットを既存のデータセットの最新バージョンにリフレッシュします。
- プロジェクトのデータセットを別のデータセットで置換します。

データセットをリフレッシュすると、データセットの最新バージョンを使用するようにプロジェクトのデータが更新されます。

たとえば、バージョン1のデータセットでプロジェクトを開始し、その後データセットの新しいバージョンが（手動インポートまたは自動化によって）ライブラリにインポートされた場合は、最新バージョンを使用するようにプロジェクトのデータセットをリフレッシュできます。

## データセットのリフレッシュ

データセットを最新のバージョンにリフレッシュするには：

1. プロジェクトで、**ツール**バーの**ステップ**をクリックします。
2. **ステップ**ツールの下部で、**データセットのリフレッシュ**をクリックします。

   **データセットのリフレッシュ**ペインが表示されます。 リフレッシュ可能なすべてのデータセットがデフォルトで選択されています。
3. リフレッシュするデータセットを選択します。**すべて** または、個々のデータセットを選択できます。
4. **保存**をクリックします。

   プロジェクトのデータが、選択したデータセットの最新バージョンに更新されます。

# いつデータセットをリフレッシュできますか？

データセットは以下のときにリフレッシュできます。

- ライブラリに現行のデータセットのより新しいバージョンが存在するとき。
- インタラクティブモード機能が有効になっていて、データセットのインタラクティブ部分のサイズが変更されたとき。

プロジェクトのデータセットがリフレッシュ可能になると、視覚的なヒントが提供されます。

- **データセットのリフレッシュ**ボタンが緑色の場合、プロジェクトで使用されている1つ以上のデータセットの新しいバージョンが検出されたことを意味します。ボタンが灰色の場合は、現行のデータセットより新しいバージョンはありません。
- **[データセットの更新]**ペインにある**最新バージョンを使用**ボタンが緑色の場合は、新しいバージョンのデータセットを利用できます。
- **ファイルの詳細**リンクをクリックすると**バージョン情報**ペインが開き、そこでデータセットの最新バージョンに含まれる新しい行と列の数がすぐにわかります。プロジェクトがインタラクティブモードにあり、そのデータセットにインタラクティブ部分を超える行数が含まれている場合は、プロジェクトに取り込み可能な行数を示す[インタラクティブ]列も表示されます。この数を見ると、インタラクティブ部分の増加または減少がすぐに判り、さらにデータセットのリフレッシュが必要かどうかも判断できるため、この数は重要です。

**備考**

すべてのData Prepプロジェクトには、Data Prepシステム管理者が設定したプロジェクトの最大行数の制限があります。その制限に近づいていて、管理者が上限を増やせない場合は、引き続きプロジェクトの行数制限を超えることなくプロジェクトに新しいデータを取り込めるようにするため、最新バージョンで更新するデータセットを選択的に選ぶことができます。

データセットの選択を解除すると、**最新バージョンを使用**ボタンは濃い灰色に変わります。これは、データセットの新しいバージョンが存在し、なおかつデータセットを更新しないことが選択されていることを示します。

データセットの新しいバージョンがない場合、**最新バージョンを使用**ボタンは薄い灰色になります。

# データセットの置換

データのリフレッシュとは異なり、データセットの置換では、プロジェクトで使用するデータセットまたはデータセットの特定のバージョンを選択できます。たとえば、データセットのバージョン1でプロジェクトを開始し、5つの追加バージョンがインポートされた場合、データセットを置き換えると使用する正確なバージョンを選択できますが、それは最新バージョンではない場合があります。また、データセットの置換では、プロジェクトで使用するデータセットを完全に変更することもできます。

プロジェクトで使用するデータセットを置き換えるには：

1. プロジェクトで、**ツール**バーの**ステップ**をクリックします。
2. **ステップ**ツールで、更新するデータセットのあるステップをクリックし、上部にある**編集**をクリックします。

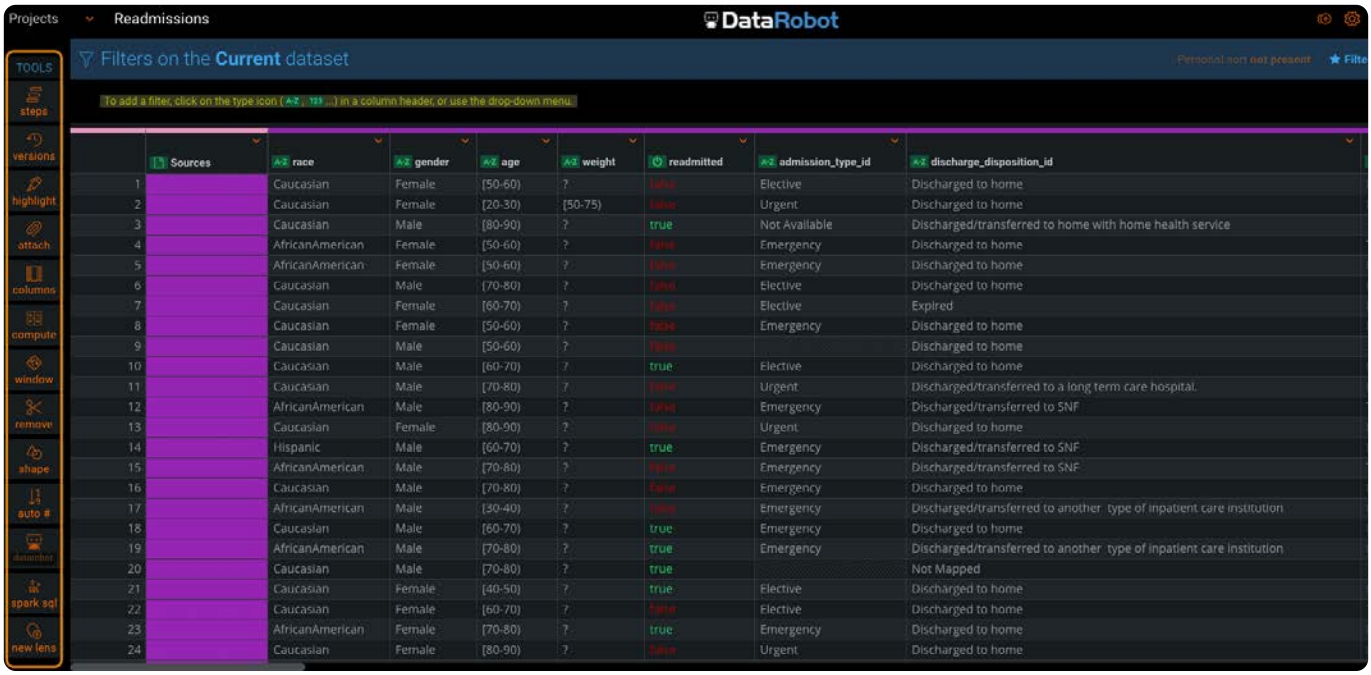
   プロジェクトが、選択したステップが作成されたときの状態に戻ります。

3. **データプレビュー**ペイン上で、更新したいデータセットの名前をクリックします。
4. **データセットを選択**ページで、使用したいデータセットを選択します。
   - データセットの以前のバージョンを選択するには、データセットで**すべてのバージョン**をクリックします。使用するバージョンの**選択**をクリックします。
   - ライブラリ内の別のデータセットを選択するには、データセットで**選択**するをクリックします。
5. **保存**をクリックします。

   プロジェクトデータが更新されます。

# プロジェクトツールの操作

Data Prepは、データのクリーンアップ、整形、組み合わせを行い、最終的に準備するのに役立つ多くのツールを提供します。プロジェクト**ツール**バーで、これらのツールにアクセスします。




これらのページでは、プロジェクトツールについて説明します。

| トピック | 説明... |
|---|---|
| ステップの編集 | プロジェクト内のステップの表示、編集、追加、再配置、および削除を行います。 |
| プロジェクトのバージョンの管理 | プロジェクトのバージョン履歴を確認します。 |
| データの強調表示 | パターン、スペース、および数値範囲を強調表示してから、列操作を使用して値を更新できます。 |
| データセットの結合 | データセットに対してルックアップ、結合、および追加操作を実行できます。 |
| 列の更新 | プロジェクトでの列名、順序、および可用性を編集することができます。 |
| 列の計算 | 新しい列を作成するために関数を使用して列を計算します。 |

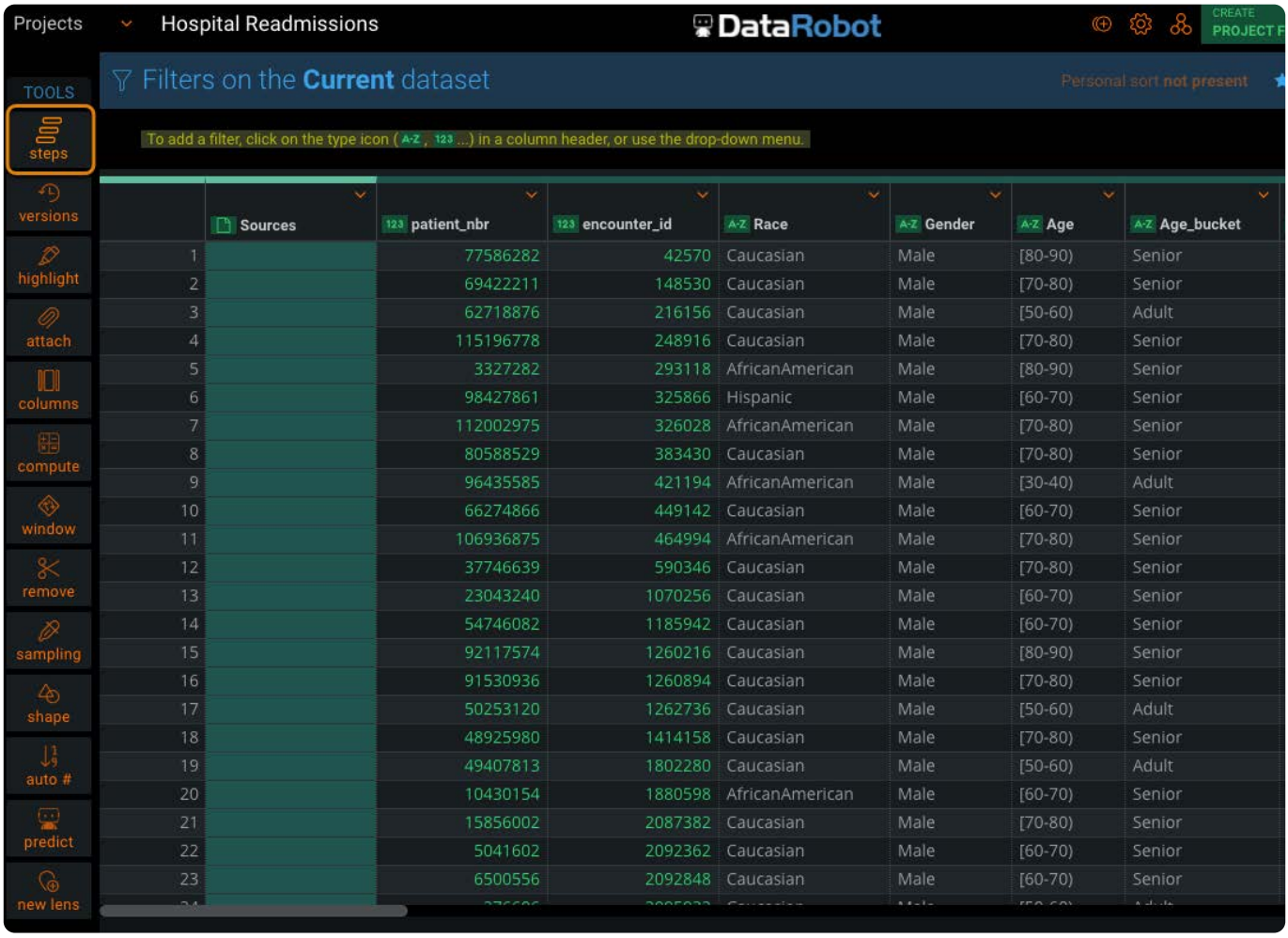
| トピック | 説明... |
| --- | --- |
| ウィンドウでグループ化 | 行をグループ化して、集計、シフト、およびランク付けを実行します。 |
| 行の削除 | ニーズに満たない行を削除します。 |
| サンプルデータ | プロジェクトデータのサンプル。 |
| データの整形 | データ整形ツールを使用して、データの重複排除、グループ化、転置、およびピボットを行います。 |
| 行の自動番号付け | データセットの行に自動番号を付ける新しい列を生成します。 |
| 予測を作成 | 予測ツールを使用してデータのスコアリング方法を学びます。 |
| Spark SQLでのデータの変換 | Spark SQLを使用してデータ変換を実行します。 |
| DRレンズからのDataRobotプロジェクトの作成 | Data Prep構築ツールを使用して、Data Prep DRレンズからDataRobot機械学習プロジェクトを作成します。 |
| レンズを使用して、公開のステップを選択します。 | レンズを使用してAnswerSetに公開するステップを特定します。 |

# ステップの操作

Data Prep**ステップ**ツールを使用すると、プロジェクト内のステップを表示、編集、追加、再配置、および削除できます。**ステップ**ツールから、*AnswerSet*と呼ばれるデータのスナップショットをエクスポートできます。すべてのデータ準備ステップの結果をエクスポートしたり、特定のステップを選択して、選択したステップまでのデータ準備アクティビティの結果であるAnswerSetをエクスポートしたりすることもできます。

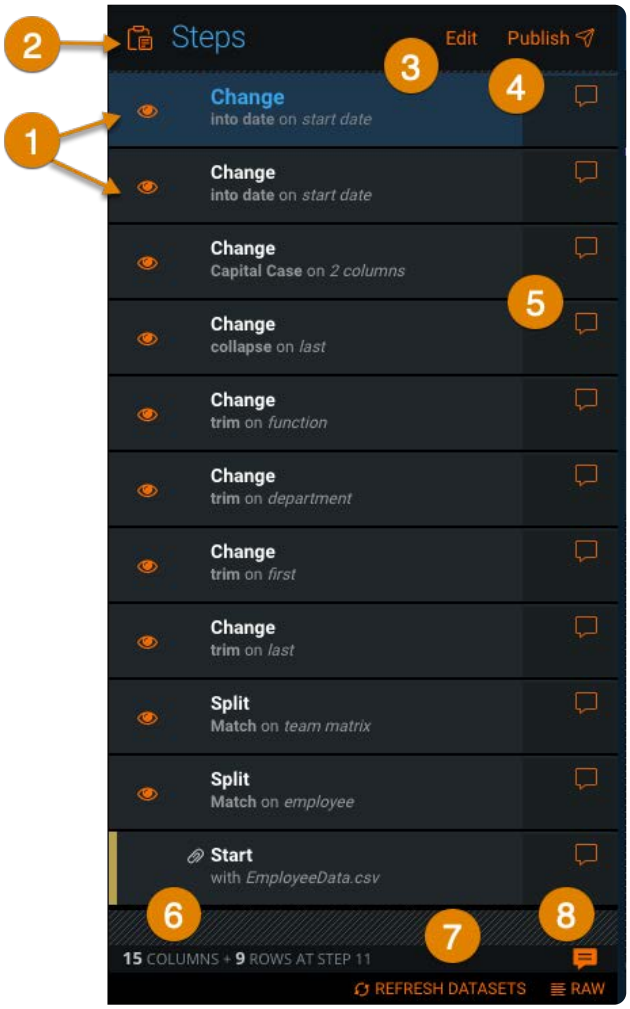
## ステップツールの操作

**ステップ**ツールにアクセスするには、プロジェクト**ツール**バーで**ステップ**をクリックします。




**ステップ**ペインの要素の概要を示します。

ステップが順番に表示されます。最初のステップは一番下にあり、最新のステップは一番上にあります。




| | 要素 | 説明 |
|---|---|---|
| 1 | **ステップの履歴** | 現時点までプロジェクト内で作成されたステップが表示されます。**開始**ステップがパネルの一番下に、最新ステップが一番上に表示されます。 |
| 2 | **ステップをコピー** | 別のプロジェクトでも再利用できるように、プロジェクトステップをクリップボードまたはファイルにコピーします。詳細については、プロジェクトステップの再利用を参照してください。 |
| 3 | **編集** | **ステップ**パネルをエディターモードで開きます。プロジェクト内のステップに変更を加えることができます。 |
| 4 | **公開** | プロジェクトをAnswerSetとしてライブラリに公開します。AnswerSetは、準備したデータを公開した結果です。 |
| 5 | **注釈** | プロジェクト内の任意のステップに注釈を設定します。詳細については、ステップツールでできることを参照してください。 |

| | 要素 | 説明 |
|---|---|---|
| 6 | **データセットのリフレッシュ** | 最新バージョンを使用できるように、プロジェクトのデータセットを更新します。詳細については、プロジェクトのデータセットの更新を参照してください。 |
| 7 | **プロジェクトの統計情報** | プロジェクト内の任意のステップに対する列数や行数など、プロジェクトに関する統計情報を表示します。任意のステップをクリックすると、このステップの統計情報を表示できます。 |
| 8 | **デバッグ** | プロジェクトの生スクリプトをJSON形式で表示します。この機能は、プロジェクトに何らかの問題が生じた場合に、デバッグ目的のみで使用してください。 |

## ステップツールでできること

ここでは、**ステップ**ツールで実行できるアクションを説明します。

| アクション | 説明 |
|---|---|
| **表示** | ステップをクリックして、その特定のステップで表示されたデータを表示します。ステップの横にある目のアイコンをクリックして、実際に削除することなく、プロジェクトからステップをミュート（非表示）にします。目のアイコンをクリックして、そのステップを再度表示します。 |
| **編集** | **ステップ**ツールを使用すると、プロジェクトにすでにコミットされているステップをいつでも編集できます。ステップを選択すると、データが調整され、プロセスの特定のステップでどのように表示されたかが表示されます。**ステップ**ツールの右上にある**編集**をクリックして、現在選択されているステップを編集します。（編集後は必ず右上にある**保存**をクリックしてください。そうしないと、変更が保持されません。）編集モードの場合、**ステップ**ツールに、プロジェクトの個々のステップごとにアクティブなデータFiltergramの数も表示されます。 |
| **ステップの追加** | プロジェクト内でアクションを保存する度に、このアクションは**ステップ**ツールの一番上のステップに追加されます。時間が古いほど下に表示され、最新のアクションが一番上に追加されます。新規ステップを、プロジェクト内の任意の履歴ポイントに追加することもできます。**ステップ**ツールでステップをクリックすると、この特定のステップの完了時点で、追加された履歴順にデータが表示されます。このようにデータが履歴的に表示されている状態で、データに対して新しいアクションを実行すると、この新しいアクション（ステップ）は、**ステップ**ツール内で最初に選択したステップの後に直接追加されます。 |

| アクション | 説明 |
| --- | --- |
| **ステップの並べ替え** | ステップを並べ替えるには、**ステップ**ツール内でステップをクリック、ドラッグ、ドロップします。ステップの順序を変更すると、データは自動的に更新され、再配置による新たな変更が反映されます。ステップの再配置により、いずれかのステップでエラーが生じた場合は、**ステップ**ボタンに警告が表示されます。さらに、エラーの生じた個々のステップにも警告が表示されます。 |
| **ステップの削除** | 目のアイコンをクリックすると、ステップを完全に削除せずに、プロジェクト内のステップをミュート（非表示）できます。**ステップ**ツールの右上隅にある**編集**をクリックすると、目のアイコンが消えます。個々のステップにカーソルを合わせると、代わりに[X]アイコンが表示されます。この[X]アイコンをクリックして、プロジェクトからステップを削除します。 |
| **注釈の追加** | プロジェクト内の任意のステップに注釈を設定できます。ステップの右横に表示される注釈ボタンをクリックすると、テキストを自由に入力するためのフィールドが開きます。注釈は、最大1,000文字まで入力できます。 |
| **再生可能性** | 新規データに対して同じステップを再生できます。AnswerSetの作成に必要なステップを使用してプロジェクトを構築したら、プロジェクトの最初のステップを編集し、ベースデータセットを新しいベースデータセットに置き換えます。これにより、プロジェクトのすべてのステップが新しいベースデータセットで自動的に実行されます。 |

# コピーアンドペーストを使用したプロジェクトステップの再利用

**ステップ**ツールを使用すると、データ準備プロジェクトからステップをコピーして、同じプロジェクトの他の場所で使用したり、別のデータ準備プロジェクトにコピーしたりできます。ステップをコピーする場合は、コンピューターのクリップボードにコピーしておくことで、ワンクリックでペーストできます。オプションとして、ステップをファイルにコピーすることもできます。 これにより、コピーしたステップを後で使用し、他のData Prepユーザーと共有できます。

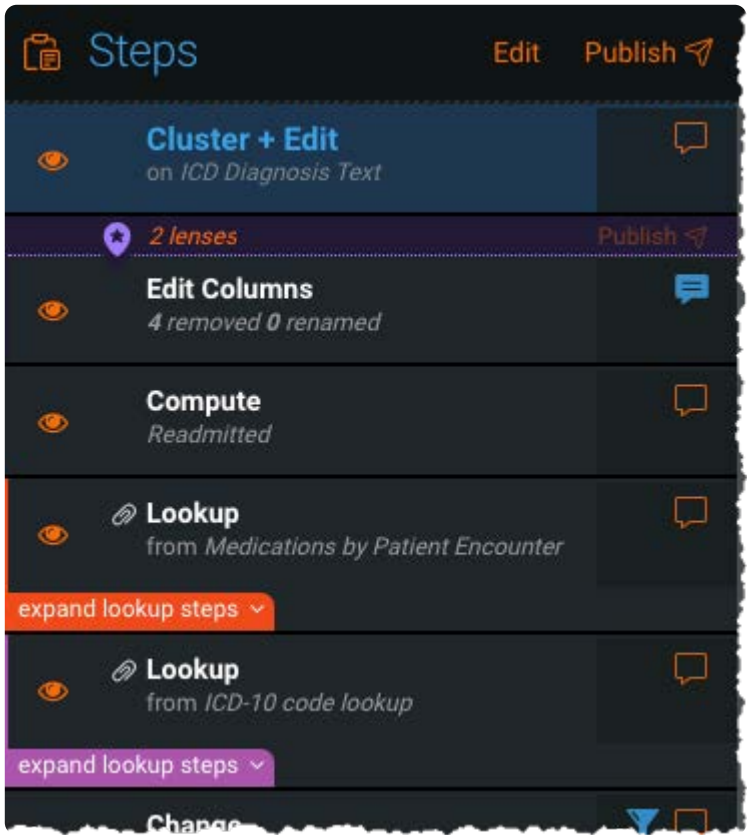
**備考**

- プロジェクト内の基本データセット（最初のステップ）は決してコピーできません。
- ルックアップステップまたは追加ステップを展開した場合、ルックアップまたは追加のインポート中に適用された変換ステップのみを選択できることがわかります。これらの変換ステップは、別のプロジェクトに貼り付けると、個別のステップとして扱われます。宛先プロジェクトのルックアップまたは追加の下にネストすることはできません。

# ステップをコピー

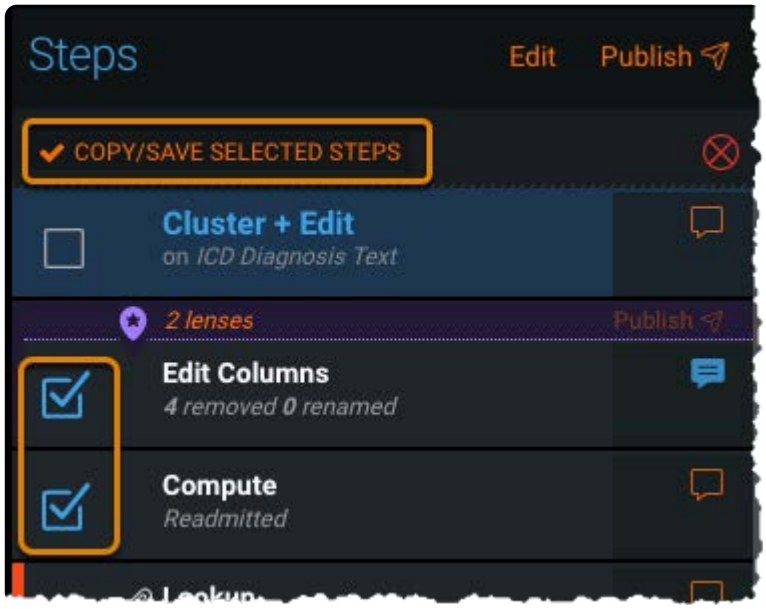
プロジェクトからステップをコピーするには、次の手順に従います：

1. **ステップ**ペインから、**ステップをコピー**（クリップボード）アイコンをクリックします。

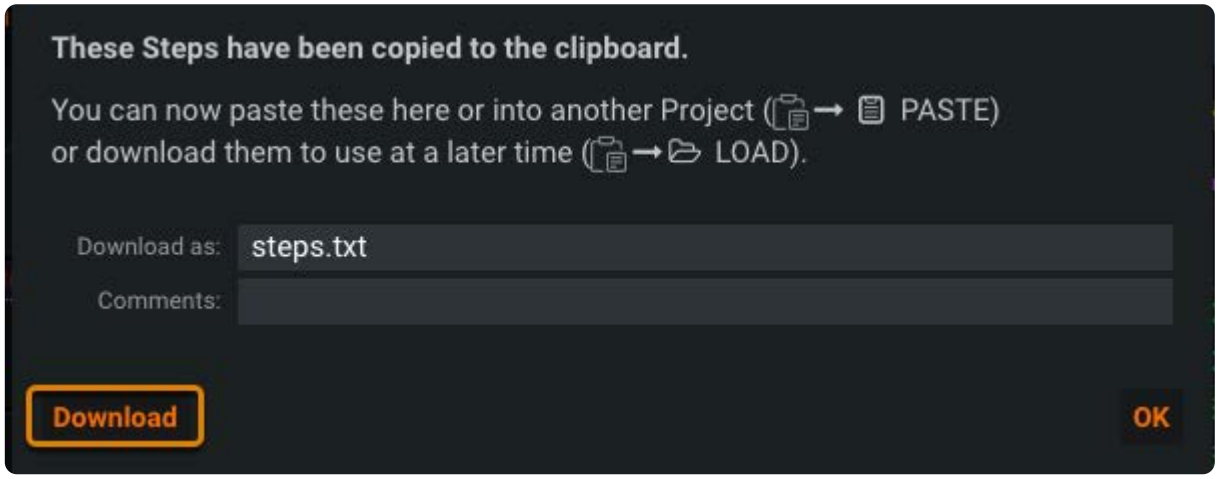



2. **ステップを選択**をクリックします。


3. コピーする各ステップのチェックボックスをクリックして、**選択したステップをコピー／保存**をクリックします。



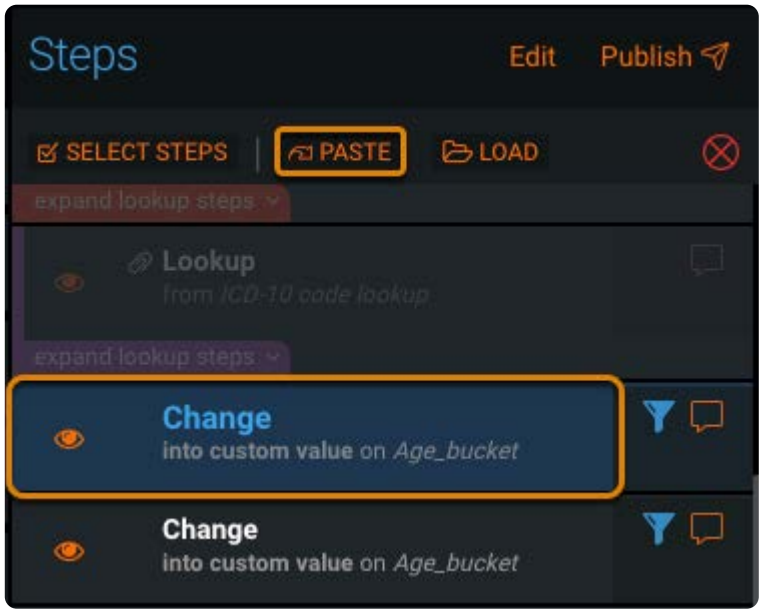

Data Prepは、ステップをクリップボードにコピーします。それらを現在のプロジェクトまたは別のプロジェクトに貼り付けるか、または後で再利用するために .txt ファイルにダウンロードすることができます。




## コピーしたステップの貼り付け

1. **ステップ**ペインから、**ステップをコピー**（クリップボード）アイコンをクリックします。


2. プロジェクト内のステップをクリックしてから、ステップ（この例では丸で囲んだ**変更**ステップ）を貼り付け、**ペースト**をクリックします。

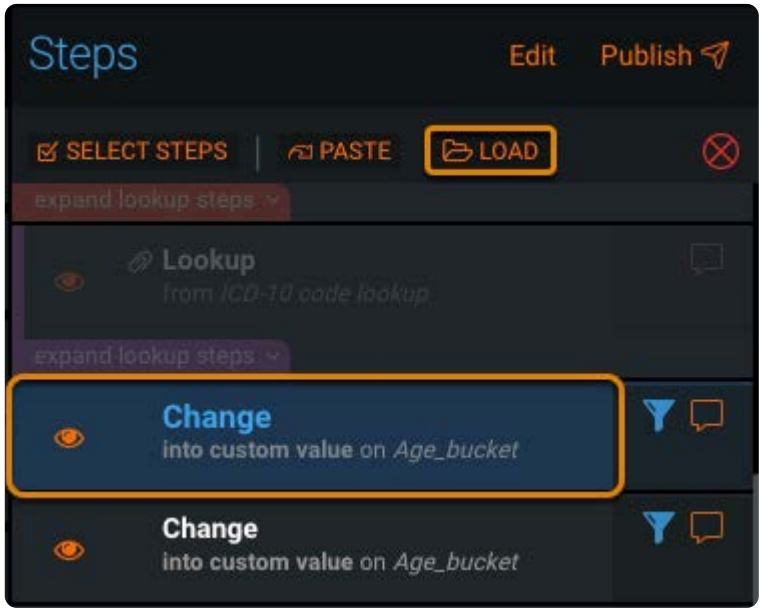



Data Prepは、コピーしたステップを選択したステップの後に貼り付けます。

## コピーしたステップの読み込みと貼り付け

1. **ステップ**ペインから、**ステップをコピー**（クリップボード）アイコンをクリックします。


2. プロジェクト内のステップをクリックしてから、ステップ（この例では丸で囲んだ**変更**ステップ）を貼り付け、**読み込み**をクリックします。




3. 保存したステップファイルに移動し、**開く**をクリックします。

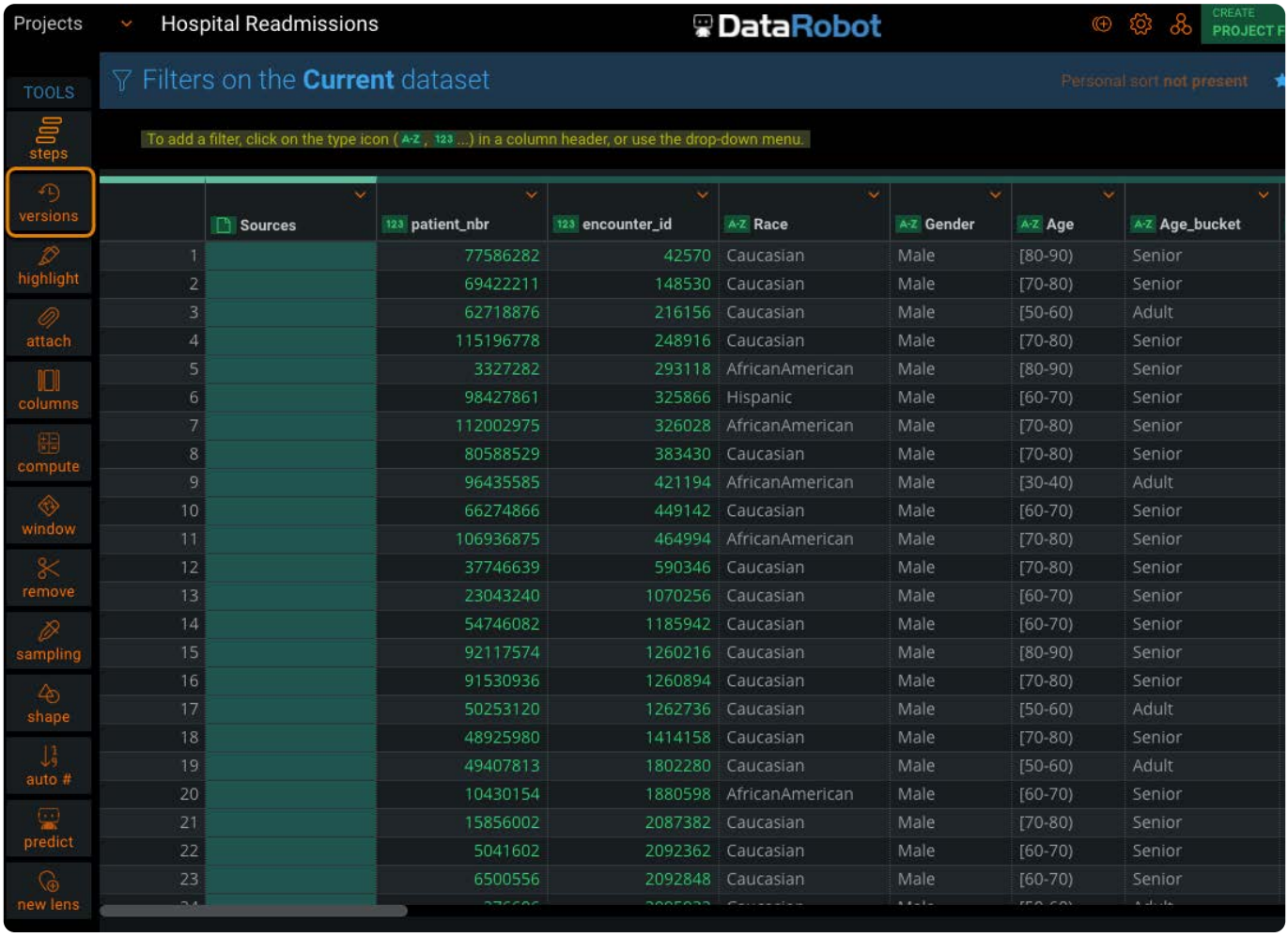
   Data Prepは、コピーしたステップを選択したステップの後に貼り付けます。

# プロジェクトのバージョンの管理

Data Prepプロジェクトでアクション（ステップの追加、ステップの削除、ステップの再配置など）が実行されるたびに、プロジェクトの新しいバージョンが作成されます。

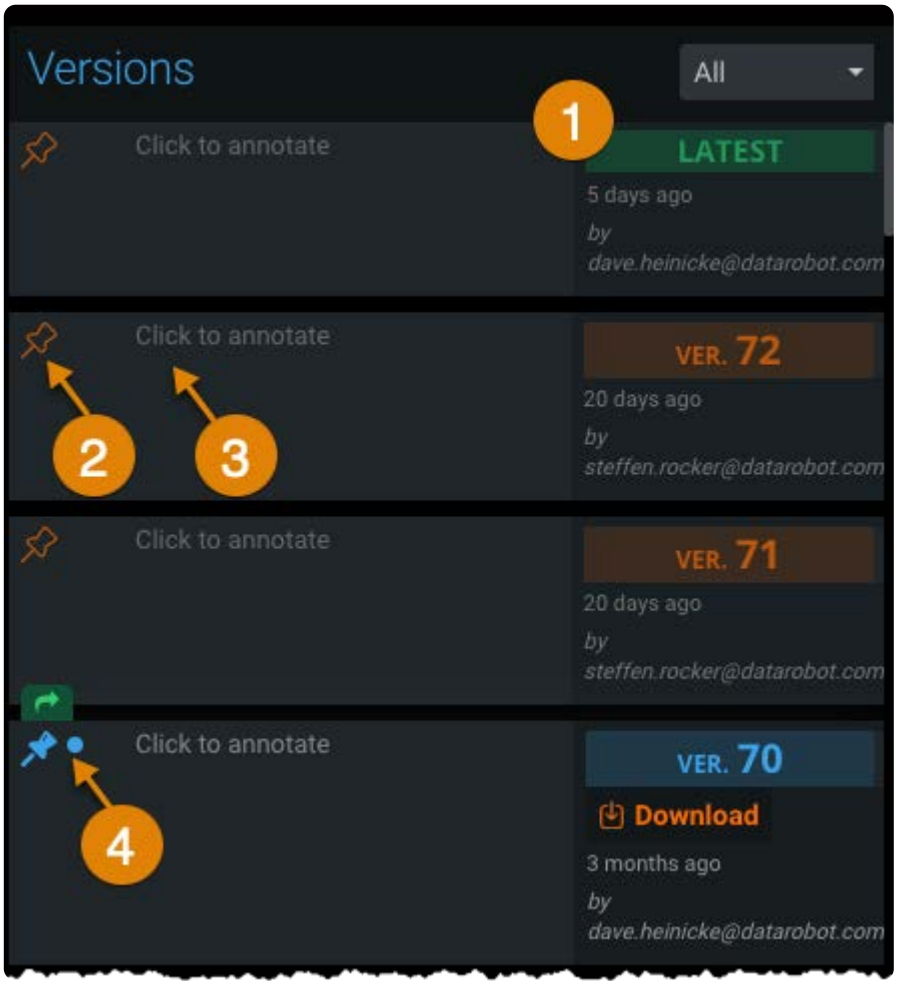
## バージョンツールの操作

**バージョン**ツールアクセスするには、プロジェクト**ツール**バーで**バージョン**をクリックします。




すべてのバージョンは、**バージョン**パネルにリストされます。履歴の中で、あるポイントのプロジェクトを表示するには、ペイン内の任意のバージョンをクリックしてください。

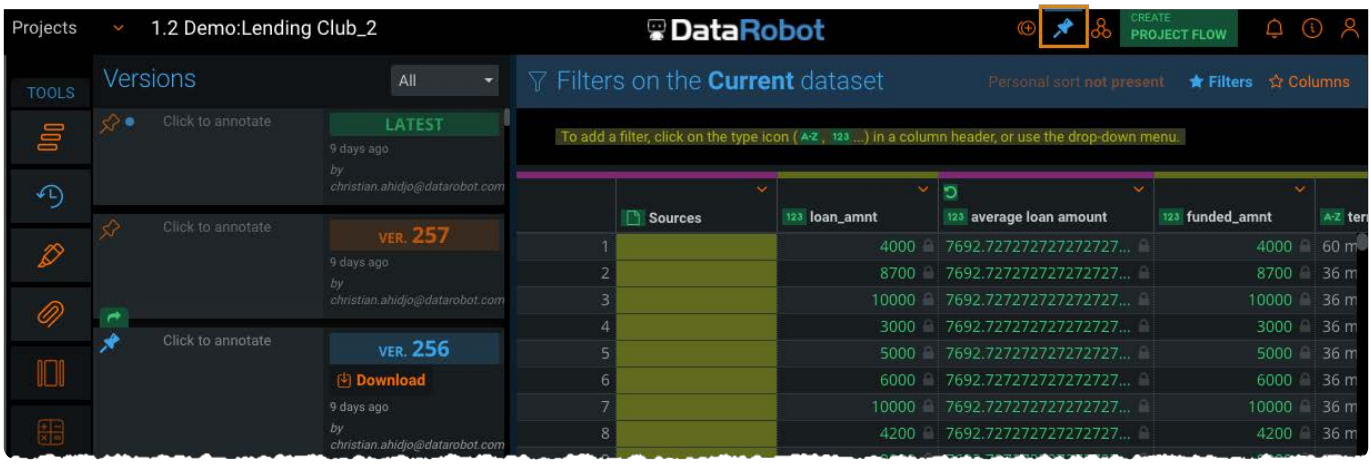
**バージョン**ペインの要素の概要を示します。



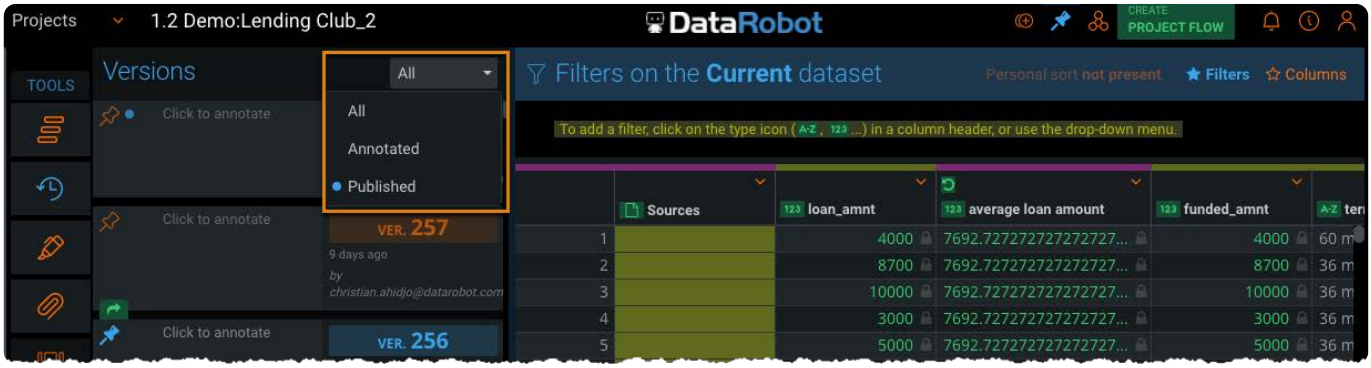

| | 要素 | 説明 |
|---|---|---|
| 1 | 最新バージョン | **最新**バージョンのプロジェクトは、ペインの最上部にリストされます。 |
| 2 | ピンアイコン | ピンのアイコンをクリックして以前のバージョンを表示します。<br>また、以前のバージョンのプロジェクトを最新のバージョンに昇格させることもできます。また、以前のバージョンを最新のバージョンに変更することができます:<br>1. バージョン名の横にあるオレンジ色のピンをクリックします。<br>2. **クリックしてこのバージョンを使用する**を選択します。<br><br>バージョンが**バージョン**ペインで最新のバージョンに変更されます。 |
| 3 | クリックして注釈を付ける | 必要に応じて、プロジェクトのバージョンに注釈を付けて、その特定のバージョンで変更されたステップがわかるようにすると便利です。例：「顧客IDデータを追加します。」<br>バージョンに注釈を付けるには、目的のバージョンの**注釈を設定する場合は、クリック**というテキストを選択し、注釈を指定します。 |

| | 要素 | 説明 |
|---|---|---|
| 4 | AnswerSet | AnswerSet としてデータライブラリに公開されたバージョンには、その横に青いドットが表示されます。 |

プロジェクトの以前のバージョンを表示するときはいつでも、プロジェクト準備ページの上部にピンアイコンが表示され、最新のバージョンが表示されていないことを示します。ピンアイコンをクリックすると、プロジェクトの最新バージョンにすぐに戻ります。




デフォルトでは、プロジェクトのすべてのバージョンがペインに表示されます。ただし、表示されるバージョンのタイプを制限できます。**バージョン**ペイン上部のドロップダウンメニューから、公開されていない**注釈付き**バージョンのみを表示するか、**公開済み**バージョンのみを表示するかを選択します。

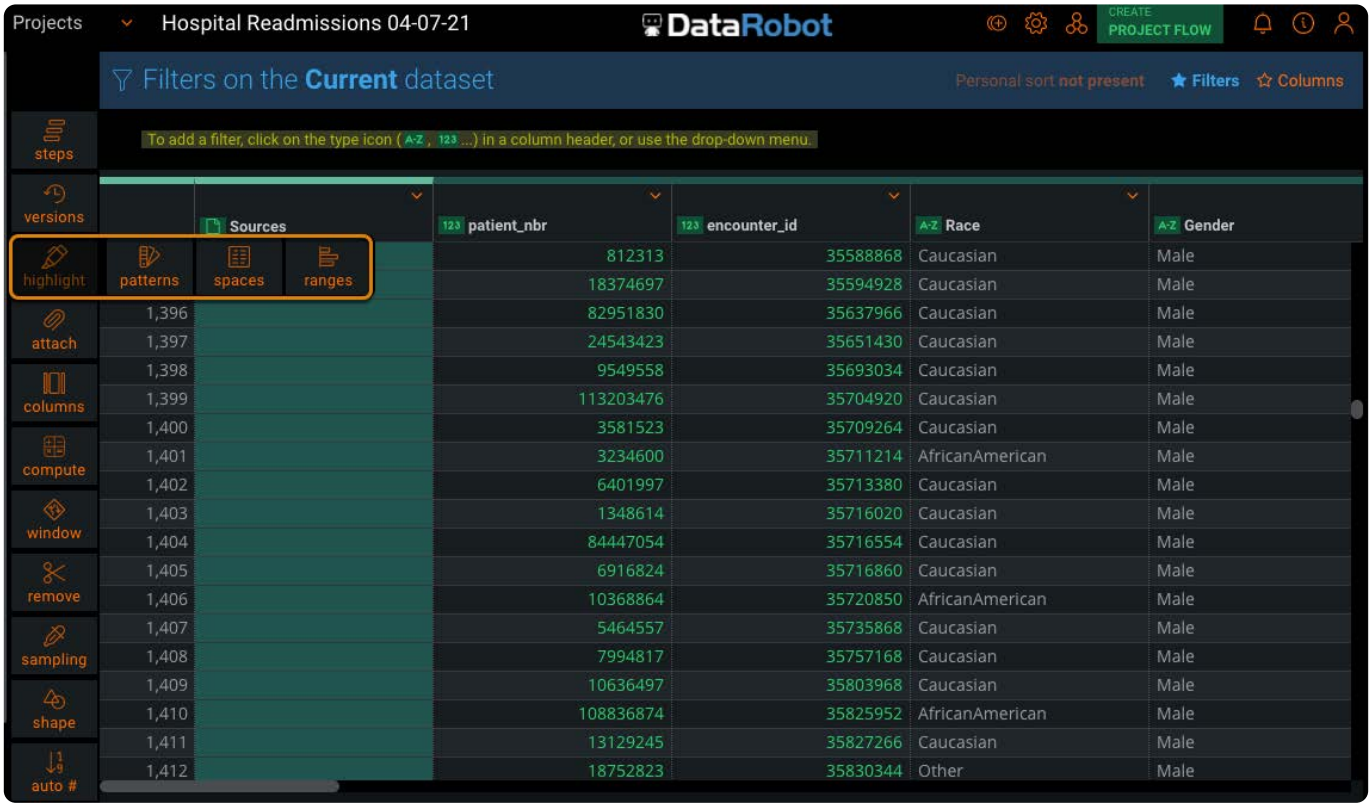


# データの強調表示

Data Prep**ハイライト**ツールは、データをよりよく理解するのに役立つ視覚的なキューを提供します。パターン、スペース、および数値範囲を強調表示してから、列操作を使用して値を更新できます。

## ハイライトツールの操作

ハイライト機能を操作するには、プロジェクト**ツール**バーの**ハイライト**ツールにカーソルを合わせ、パターン、スペース、または 範囲ツールを選択します。




**備考**

すべてのハイライトツールを同時に有効にできます。

## パターンを強調表示

**ハイライトパターン**ツールは、互いに類似している可能性のあるデータセット内のセルを検出し、それらのセルにフラグを立てるためのカラーコードを提供します。

たとえば、一致するデータ値が近くにあるセルは同じ色を共有します。**ハイライトパターン**ツールは、**フィルター**ペインおよび**列**ツールと組み合わせて使用する場合、特に便利です。

# スペースをハイライト

**スペースをハイライト**ツールは、データセット内のスペース文字を自動的に強調表示します。スペース文字は灰色のボックスとして表示されます。スペース文字には、スペース、タブ、キャリッジリターン、改行文字、垂直タブ文字、フォームフィード文字が含まれます。

# 範囲をハイライト

**範囲をハイライト**ツールは、数値列の自動色付けを提供して、値が列内のすべての値の範囲内のどこにあるかを示します。

範囲は、棒グラフと同様の外観で表示されます。

- **すべての正の値**：値の範囲が50から100であるとします。値が75のセルの場合、75は50と100の中間であるため、影付きのバーがセルの途中まで伸びています。セル値100の場合、セルは完全にシェーディングされます。
- **すべての負の値**：すべての負の値は、すべての正の値と同じ基本的な外観で表示されます。-100から-50の範囲の場合、セル値-75のセルは半分シェーディングされ、セル値-50のセルは完全にシェーディングされます。
- **正の値と負の値の両方**：列に負の値と正の値が含まれている場合、負の値の場合、影付きのバーはゼロを表すポイントから左に伸びます。正の値の場合、影付きのバーはゼロを表すポイントから右に伸びます。ゼロが実際の値の範囲の中間点でない場合、ゼロは中央の左または右に比例して配置されます。たとえば、範囲が-10から1000の場合、ゼロを表す点はセルの左端の近くにあります。

**備考**

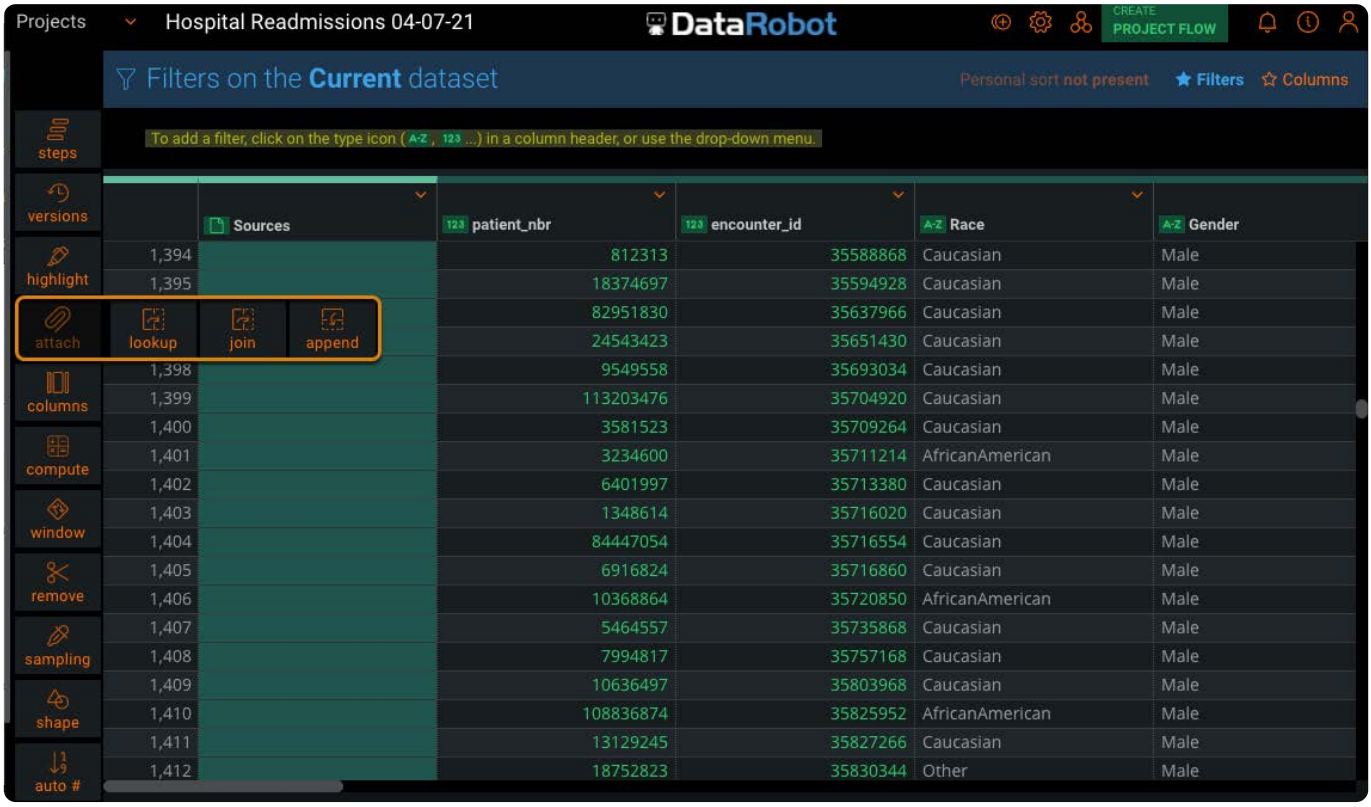
強調表示は、プロジェクトで公開されているAnswerSetには影響しません。AnswerSetについては、データセットをエクスポートするを参照してください。

# データセットの結合

Data Prep **添付**ツールを使用すると、データセットに対してルックアップ、結合、および追加操作を実行できます。

## 添付ツールの操作

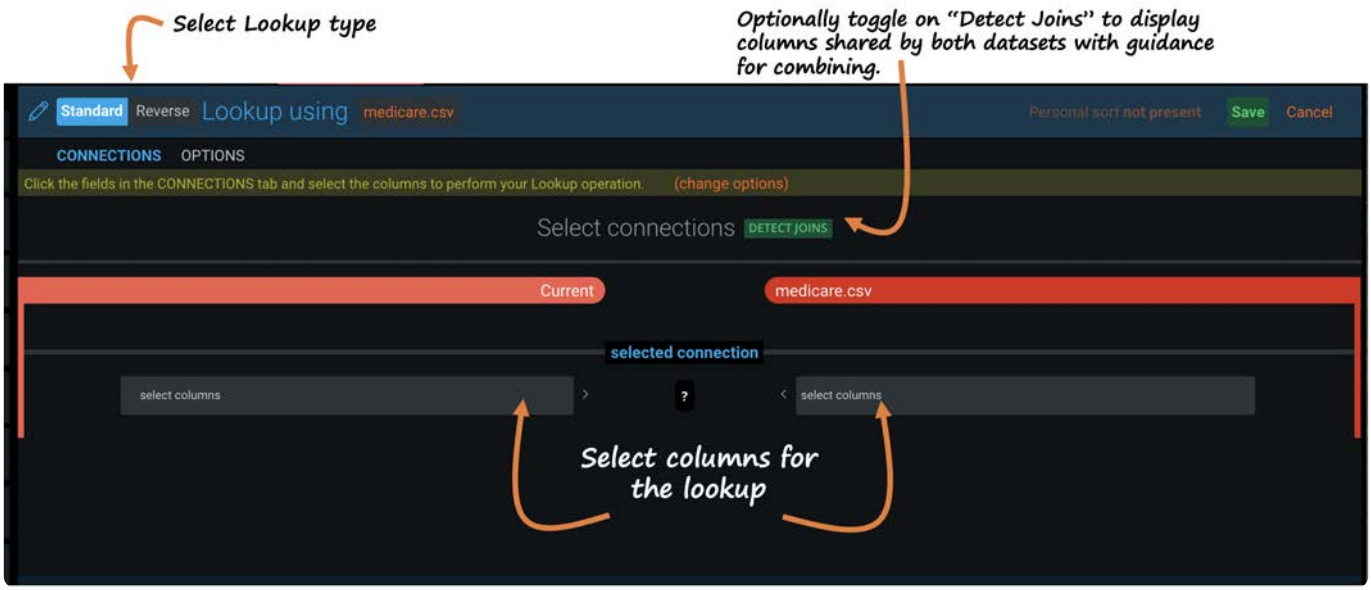
データセットを添付するには、プロジェクト**ツール**バーの**添付**ツールにカーソルを合わせ、ルックアップ、結合、または追加ツールを選択します。




## ルックアップツールの操作

**ルックアップ**ツールは、MS Excel VLOOKUPに相当するルックアップタイプの操作を提供します。逆引きもサポートされています。

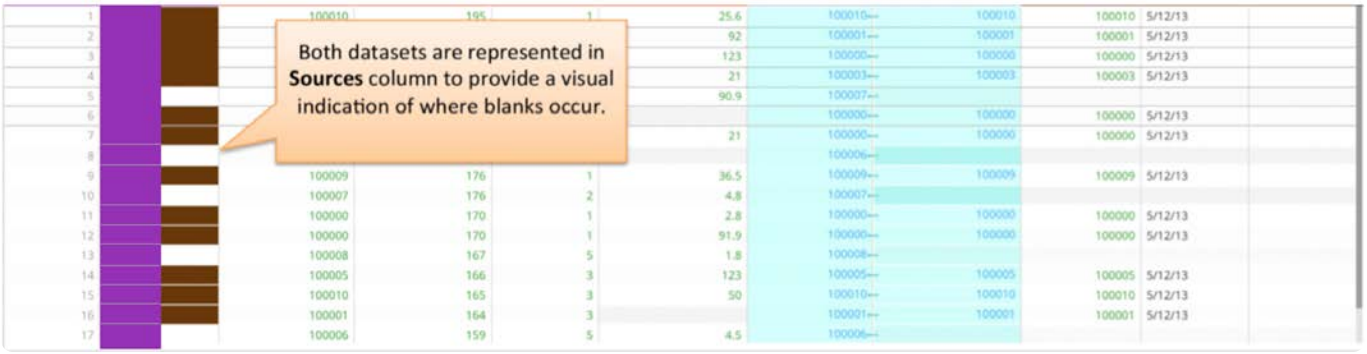
**ルックアップ**ツールを選択し、データライブラリから**ルックアップソース**データセットを選択します。次に、ルックアップ操作の対象列を選択します。**現在**列はベースデータセットを参照し、各列をクリックすると、各データセットで利用可能な列が表示されます。




**備考**

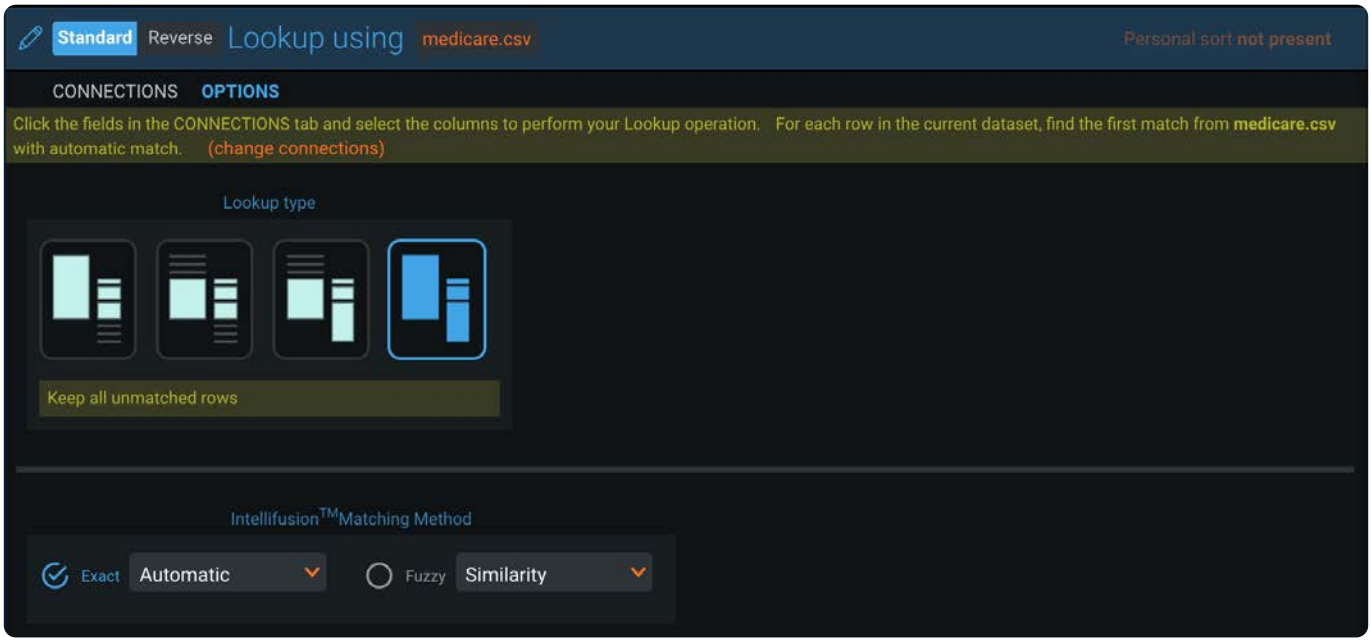
クリックを使用して、緑の**検出結合**オプションを選択すると、**接続**メニューに2つのデータセットが共有される列が表示されます。さらに、パーセンテージスコアは、データセットを組み合わせる最適な方法に関するガイダンスを提供します。**結合を検出**操作のスコアの計算方法については、ルックアップスコアリングの計算を参照してください。

ルックアップの列を選択すると、結合されるデータのプレビューがグリッドに表示されます。




次に、**[オプション]**タブをクリックして、以下の項目を定義します。

- **ルックアップタイプ**一致しない行の処理方法を定義します。
- **一致方法**：ルックアップ操作に使用するアルゴリズム。注意：ファジー一致方法は、標準タイプのルックアップでのみ使用できます。




グリッドでのルックアップのプレビュー方法を確認したら、緑色の**保存**ボタンをクリックしてルックアップ操作を完了します。

## ルックアップスコアリングの計算

DataPrepが**結合を検出**オプションのルックアップスコアを計算する際に、次の2つの要素を考慮します。

- **選択性:** ルックアップ接続の各列にユニークな値が入力される範囲
- **オーバーラップ:** ルックアップ接続で一致する行の割合

これらの2つの要素は、接続の品質を反映するパーセンテージスコアを生成します。パーセンテージスコアが高いほど、接続が良好になります。ただし、100%未満のスコアは、提案された結合操作に問題があることを必ずしも示しているわけではありません。実際、パーセンテージが1%未満の正当なユースケースがあります。

「正しい」か「正しくない」の絶対的な指標というよりは、このパーセンテージは、2つのデータセットに共通するデータ量に関して、ある程度の期待と一致するサニティチェックとして役に立つでしょう。

以下は、低いスコアと高いスコアを生成するデータの例です：

### 低スコア

| ベース（DRIVING）データセット | ルックアップ（SOURCE）データセット | 説明 |
|---|---|---|
| a,b,c | a,a,b,b,c,c | ルックアップの値が重複しています。 |
| a,b,c. | . c,d,e,f | ルックアップには、1つの重複値（「c」）しかありません。 |

### 高スコア

| ベース（DRIVING）データセット | ルックアップ（SOURCE）データセット | 説明 |
|---|---|---|
| a,b,c | a,b,c,d,e,f,g | すべてのルックアップ値は一意です。 |
| a,a,b,b,c,c | a,b,c | すべてのルックアップ値が重複しています。 |

# 結合ツールの操作

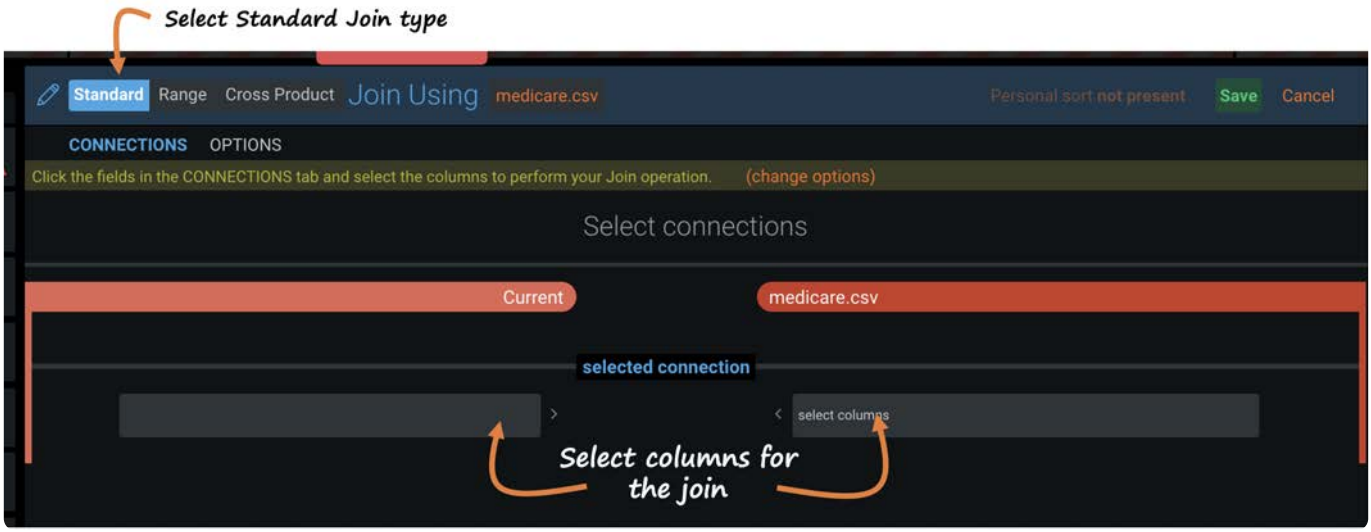
**結合**ツールは、次の結合タイプをサポートしています:

- **標準結合**: 両方のデータセットにまたがるすべての一致を組み合わせます（SQLの結合に相当）。
- **範囲結合**: 結合するデータセットの範囲を表す2つの列と照合されるマーカー列に基づいてデータセットを結合します。
- **外積結合**: 両方のデータセットのすべての行を組み合わせます。重要：結合するデータセットの各行がプロジェクトのベースデータセットの各行に対して追加されるため、外積結合によってプロジェクトに追加される行数は大幅に増加します。

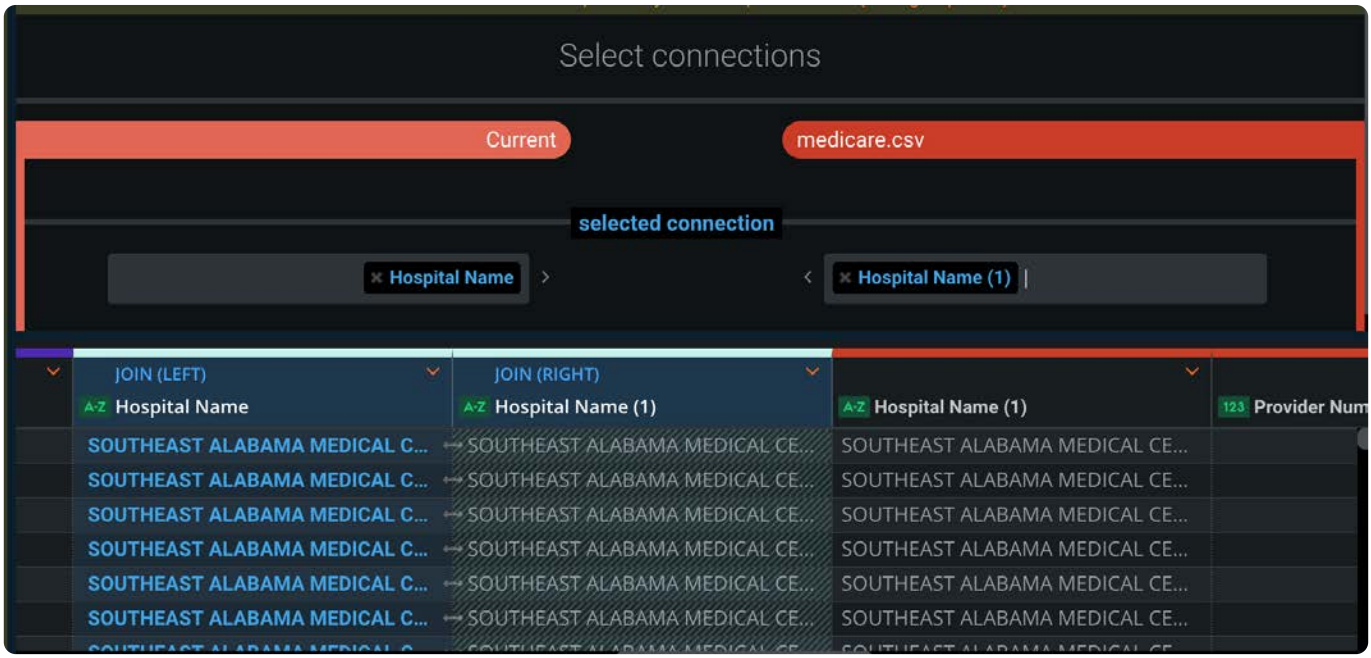
> **備考**
>
> **結合**ツールをUIで使用するには、システム管理者によって有効にする必要があることに注意してください。

## 標準結合

**結合**ツールを選択した後、**標準**タイプを選択し、結合操作を行う列を選択します。**現在**列は、基本データセットを参照します。各列をクリックすると、各データセットで利用可能な列が表示されます。最初に、結合を作成する列を選択します。

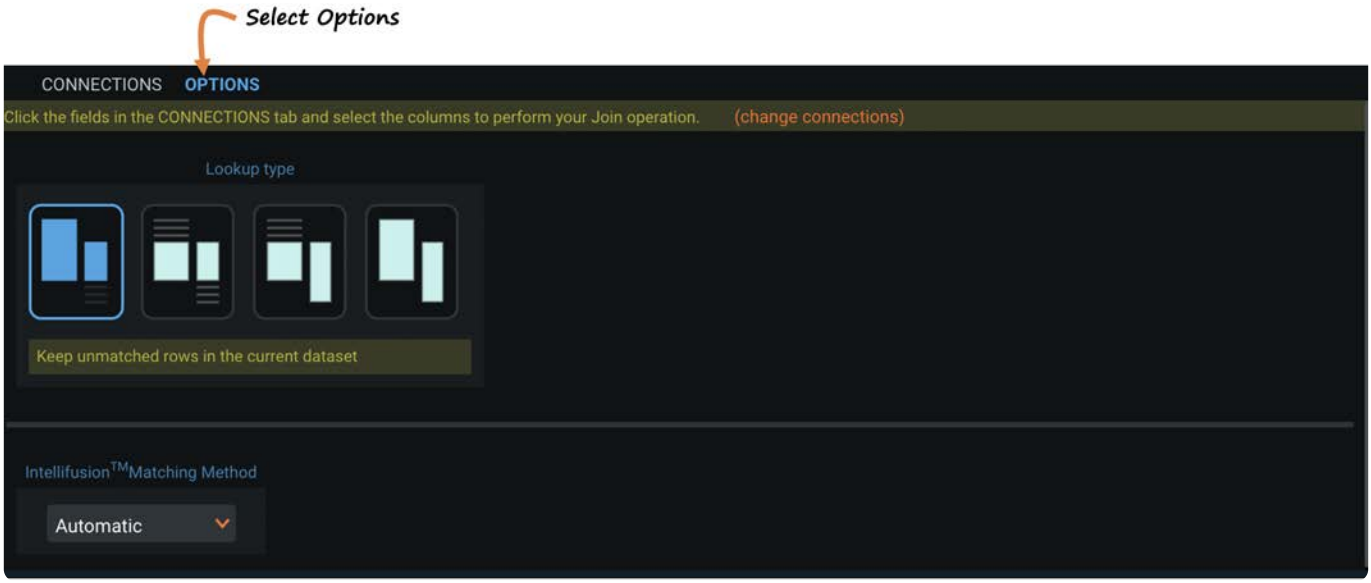


結合する列を選択すると、結合されるデータのプレビューがグリッドに表示されます。




次に、**[オプション]**タブをクリックして、以下の項目を定義します。

- **ルックアップタイプ**一致しない行の処理方法を定義します。
- **一致方法**: 結合操作に使用するアルゴリズム。




グリッドでのルックアップのプレビュー方法を確認したら、緑色の**保存**ボタンをクリックしてルックアップ操作を完了します

## 範囲結合

範囲結合機能を使用すると、隣接するデータセット内の2つの個別「範囲」列と照合されるベースデータセット内の「マーカー」列に基づいてデータベースを結合できます。

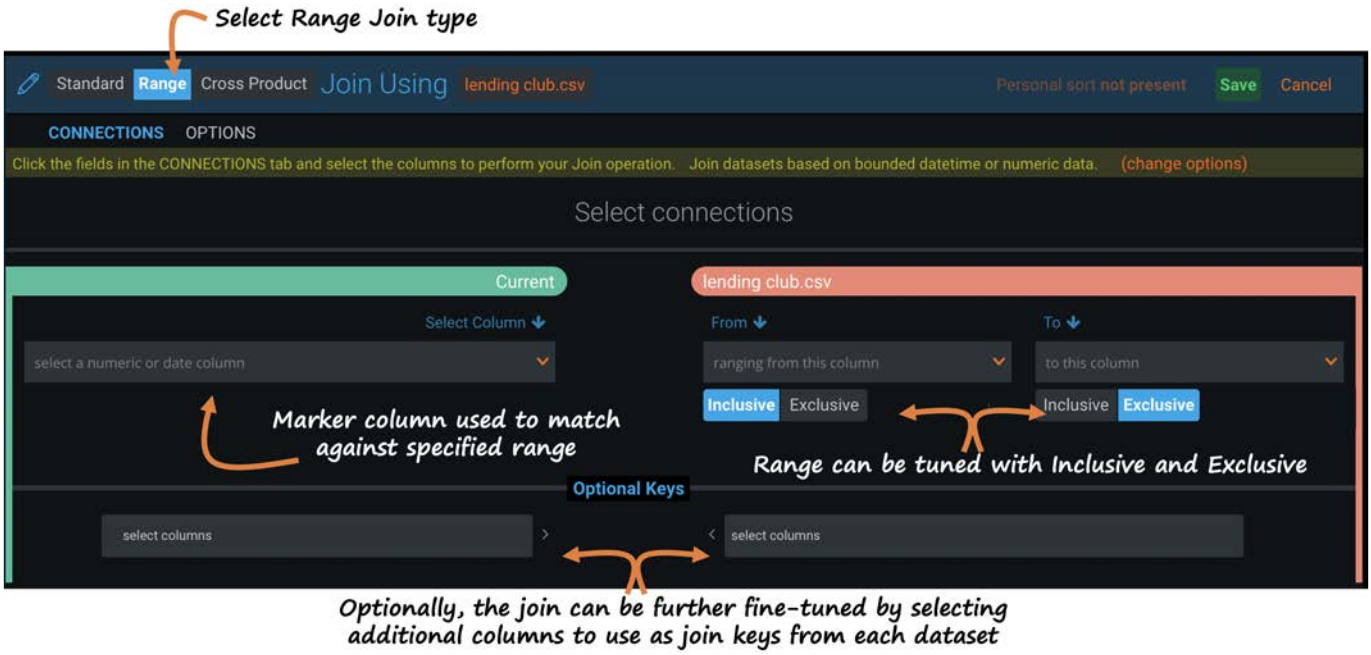
**備考**

範囲結合機能を使用できるのは、列タイプが数値および日付の場合のみです。これらの列のデータを有効に変換できる場合、列メニューの**変更**操作を使用して、列をこれらのタイプに変換できます。

**備考**

範囲結合機能を結合オプションとしてUIに表示するには、システム管理者が範囲結合機能を有効にする必要があります。

**結合**ツールを選択した後、**範囲**オプションを選択し、結合操作を行う列を選択します。**現在**の列は、ベースデータセットのマーカー列を参照します。これは、隣接するデータセットの値の下限と上限に対する照合に使用されます。




範囲結合機能を利用するには2つの方法があります。

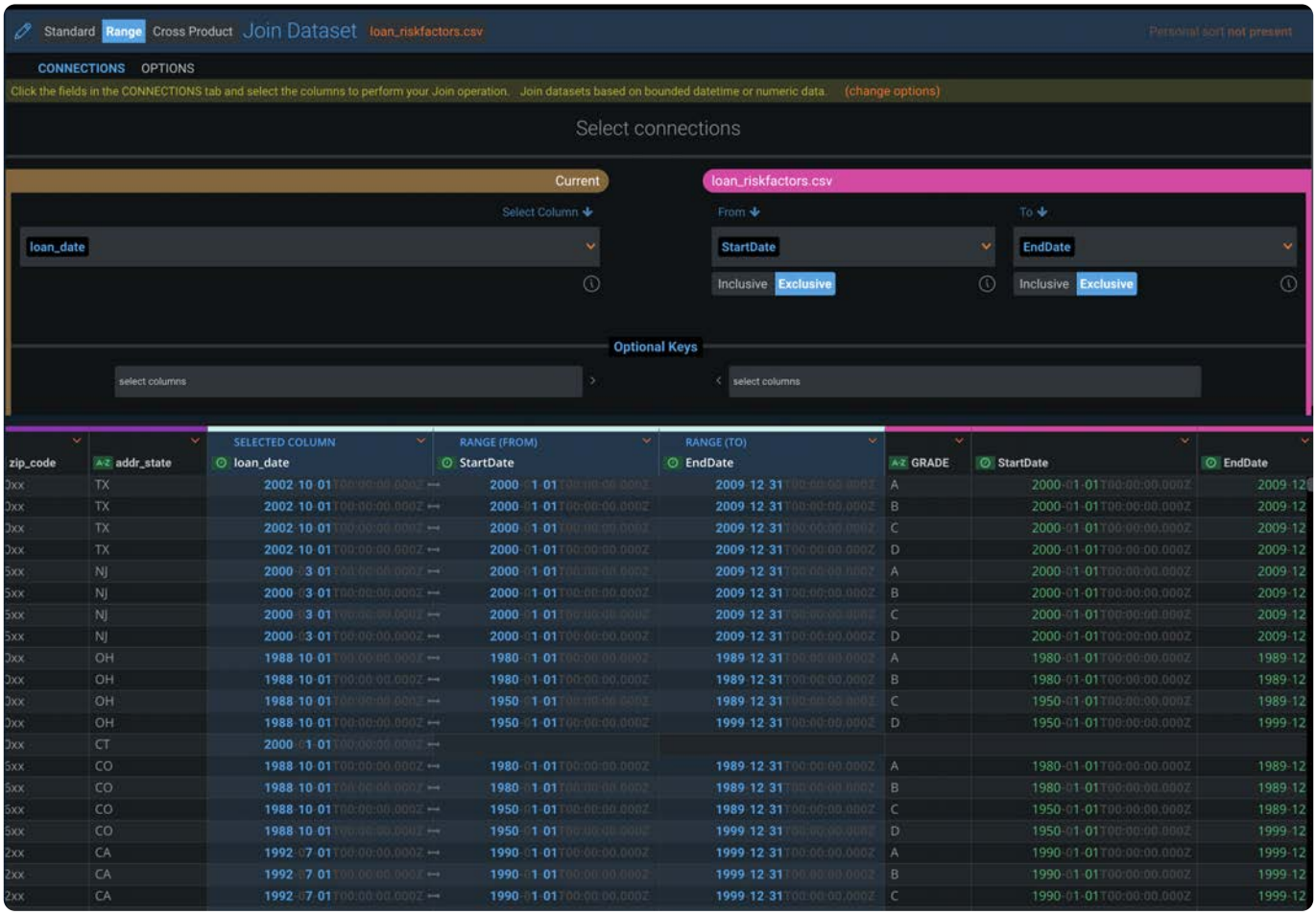
- **キーレス**: キーレス範囲結合は、マーカー値の特定の範囲内にあるエントリをグループ化します。
- **キー付き**: オプションのキーは、結合条件の特異度を慎重に高めるために、ベースデータセットおよび隣接するデータセットから選択できる追加の列です。

結合する列を選択すると、結合されるデータのプレビューがグリッドに表示されます。

# 例

このベースデータセットは、個々のローンに関する典型的な情報を含むレンディングクラブのデータセットです。そのベースデータセットに、グレードと日付に基づいてローンの平均化されたリスク要因を含む `loan_riskfactors.csv` データセットを結合します。このプロジェクトの目標は、リスク要因データセットによって提供される業界平均を使用して、貸出クラブのベースデータセットの各ローンのリスク係数を決定することです。

ここでは、列"loan_date" をレンディングクラブのベースデータセットのマーカー列として `loan_riskfactors.csv` データベースの "StartDate" 列と "EndDate" 列と照合します。グリッドに結合のプレビューが表示されます。

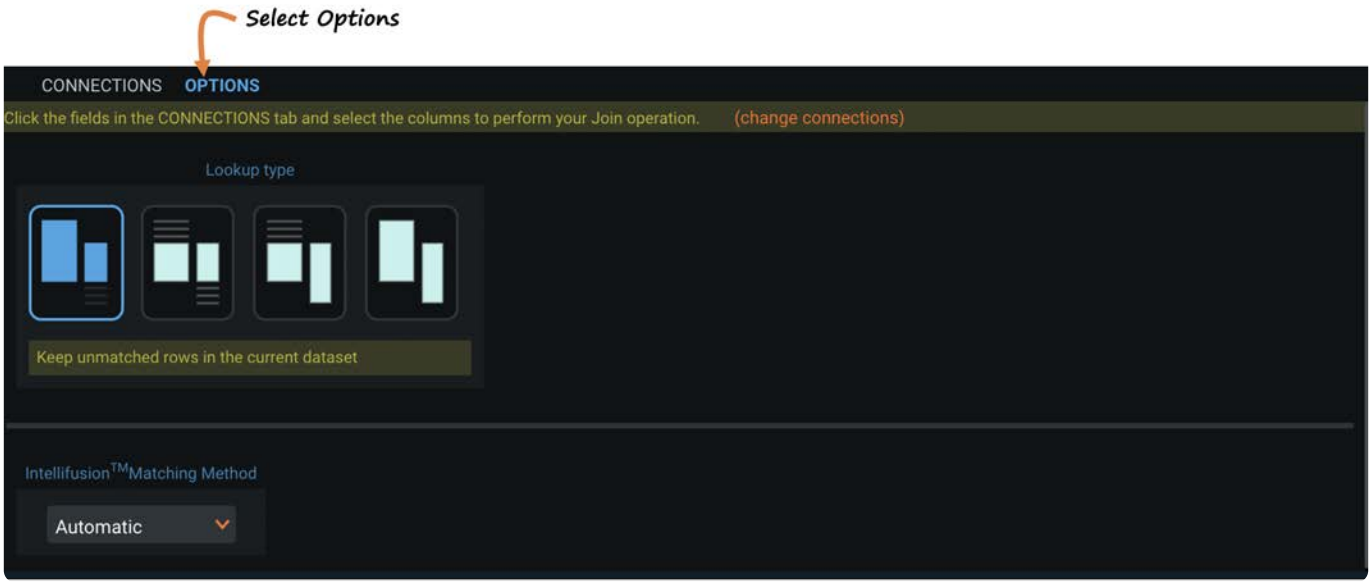



**ヒント**

結合後のデータセットに含まれる行が多すぎる場合は、**オプションのキー**を使用して、結合基準の真陰性率を高め、結果として得られる行の数を減らすキー付き結合を作成できます。

次に、**[オプション]**タブをクリックして、以下の項目を定義します。

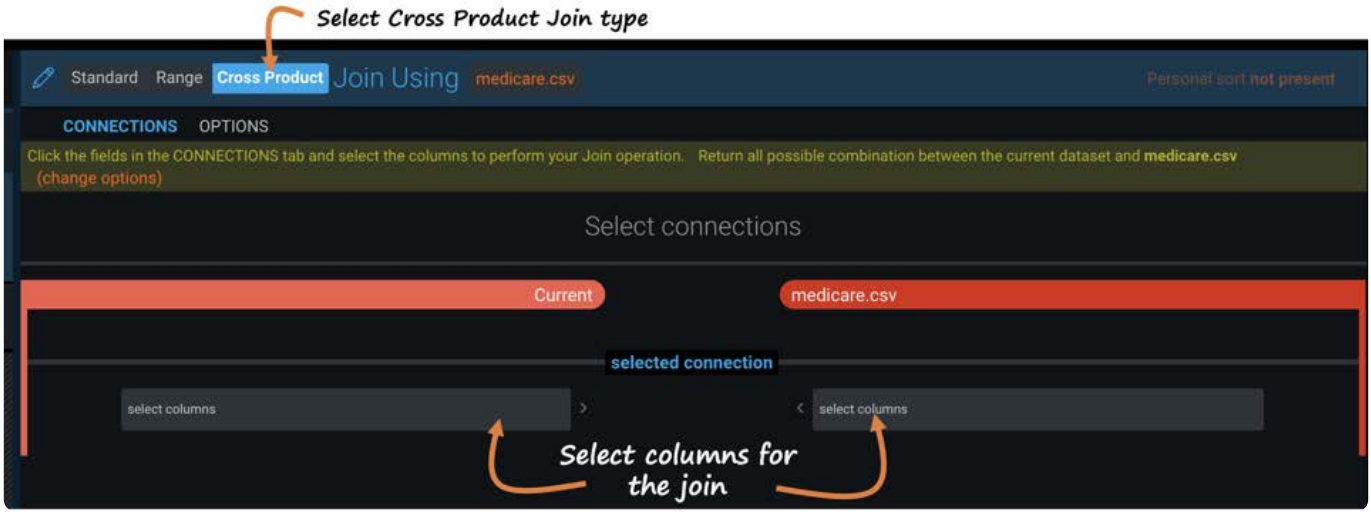
- **ルックアップタイプ**一致しない行の処理方法を定義します。
- **一致方法**: 結合操作に使用するアルゴリズム。



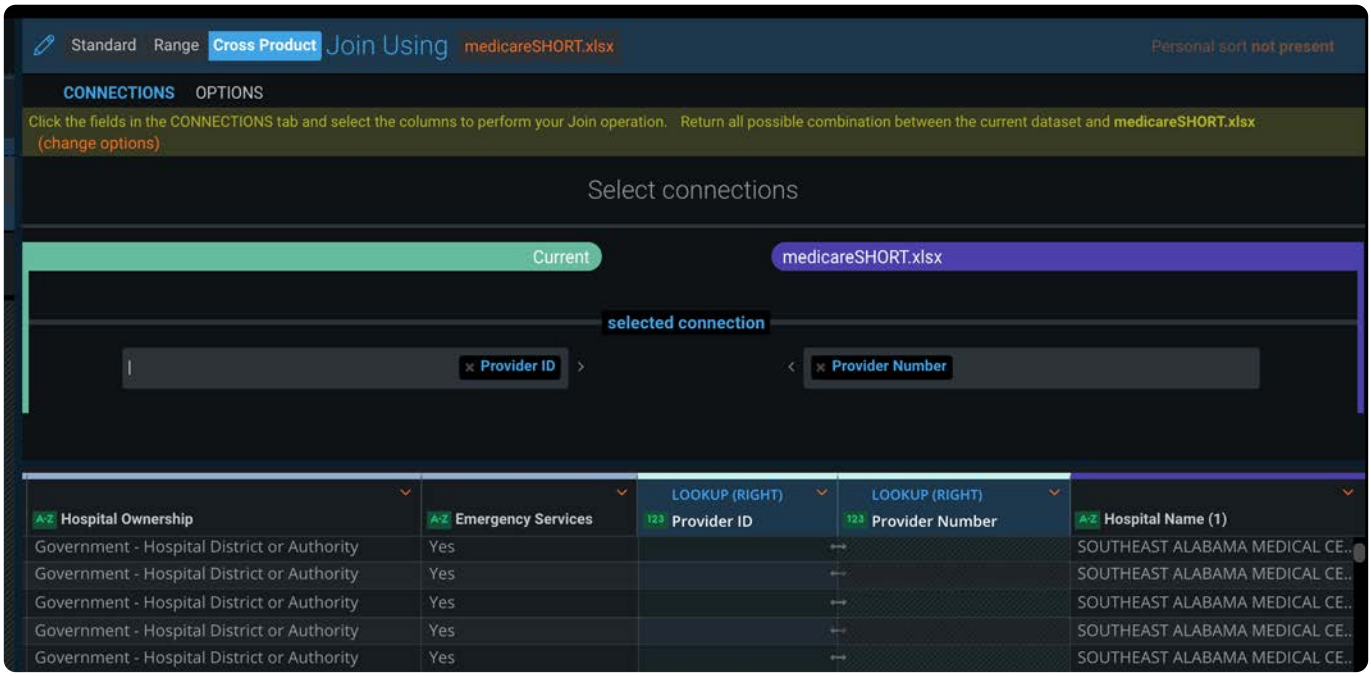

グリッドでのルックアップのプレビュー方法を確認したら、緑色の**保存**ボタンをクリックしてルックアップ操作を完了します。

## 外積結合

**結合**ツールを選択した後、**外積**タイプを選択し、結合操作の対象列を選択します。**現在**列はベースデータセットを参照し、各列をクリックすると、各データセットで利用可能な列が表示されます。最初に、結合を作成する列を選択します。




結合する列を選択すると、結合されるデータのプレビューがグリッドに表示されます。




**備考**

他の結合タイプとは異なり、すべての行がこの操作で一致するので、外積結合のルックアップや一致するオプションはありません。

グリッドでのルックアップのプレビュー方法を確認したら、緑色の**保存**ボタンをクリックしてルックアップ操作を完了します。

**備考**

結合するデータセットの各行がプロジェクトのベースデータセットの各行に対して追加されるため、外積結合によってプロジェクトに追加される行数は大幅に増加します。行数がプロジェクトの行制限を超えると結合操作は失敗し、エラーメッセージが表示されます。この場合、データセットを組み合わせる前にデータセットの行数を減らすか、プロジェクトの行制限についてシステム管理者に相談してください。

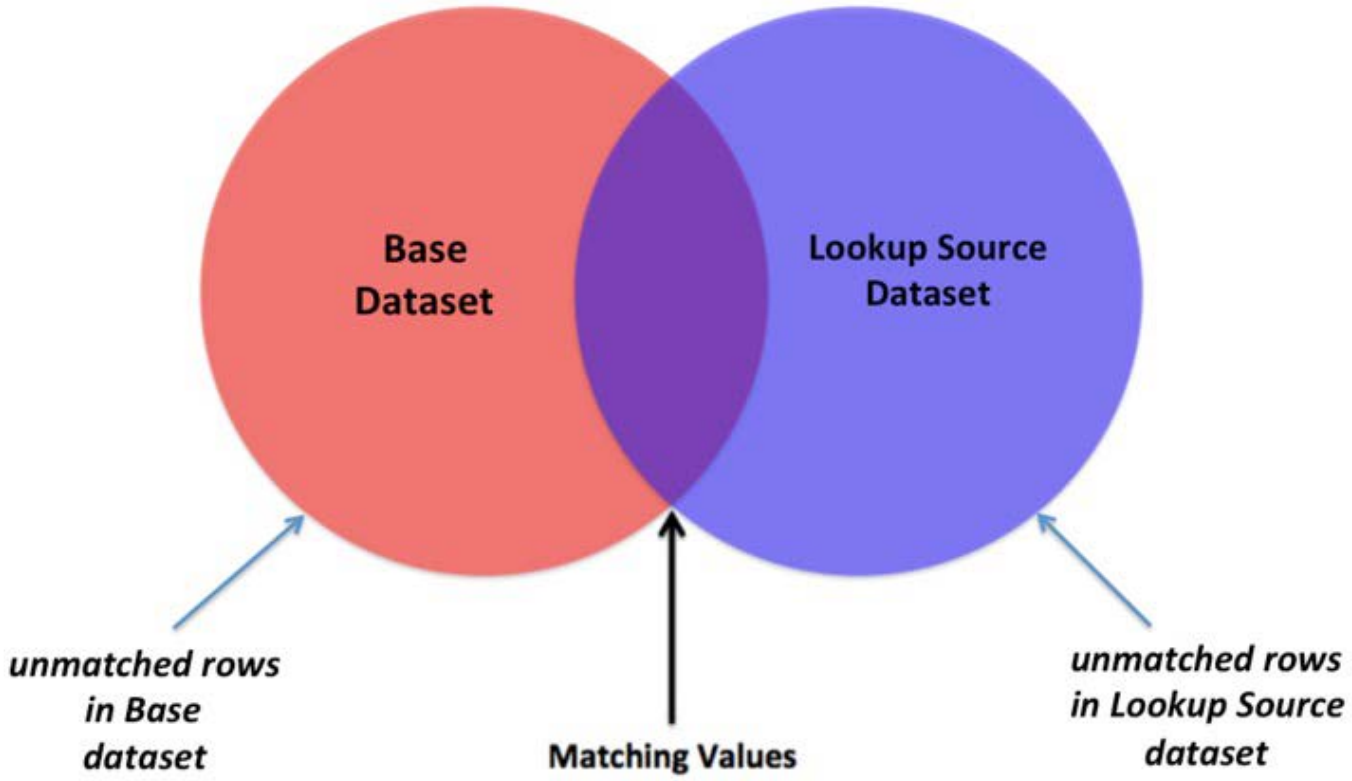
# ルックアップと結合のルックアップタイプオプション

**ルックアップタイプ**オプションでは、一致しない行をどうするかを指定します。4つのオプション（左ルックアップ、内部ルックアップ、右ルックアップ、および外部ルックアップ）があります。

**備考**

各**ルックアップタイプ**オプションのアイコンにカーソルを合わせると、関連付けられるタイプが表示されます。

さまざまなルックアップタイプを理解する最も簡単な方法の1つは、ベン図を使用することです。




## 左ルックアップ

左ルックアップは、一致する値が指定列にあるすべての行を含む結合済みデータセット、および対応する一致する値が右（ルックアップ）データセットになかった左（ベース）データセットのすべての行を返します。

| | |
|---|---|
| AUSTIN | |
| BOSTON | |
| WASHINGTON D.C. | |
| NEW YORK CITY | NEW YORK CITY |
| LOS ANGELES | LOS ANGELES |
| MIAMI | MIAMI |
| DALLAS | DALLAS |
| SAN FRANCISCO | SAN FRANCISCO |

## 内部ルックアップ

内部ルックアップは、指定された列内に一致する値を含む、こうした行のみを含む結合されたデータセットを返します:

| | |
|---|---|
| NEW YORK CITY | NEW YORK CITY |
| LOS ANGELES | LOS ANGELES |
| MIAMI | MIAMI |
| DALLAS | DALLAS |
| SAN FRANCISCO | SAN FRANCISCO |

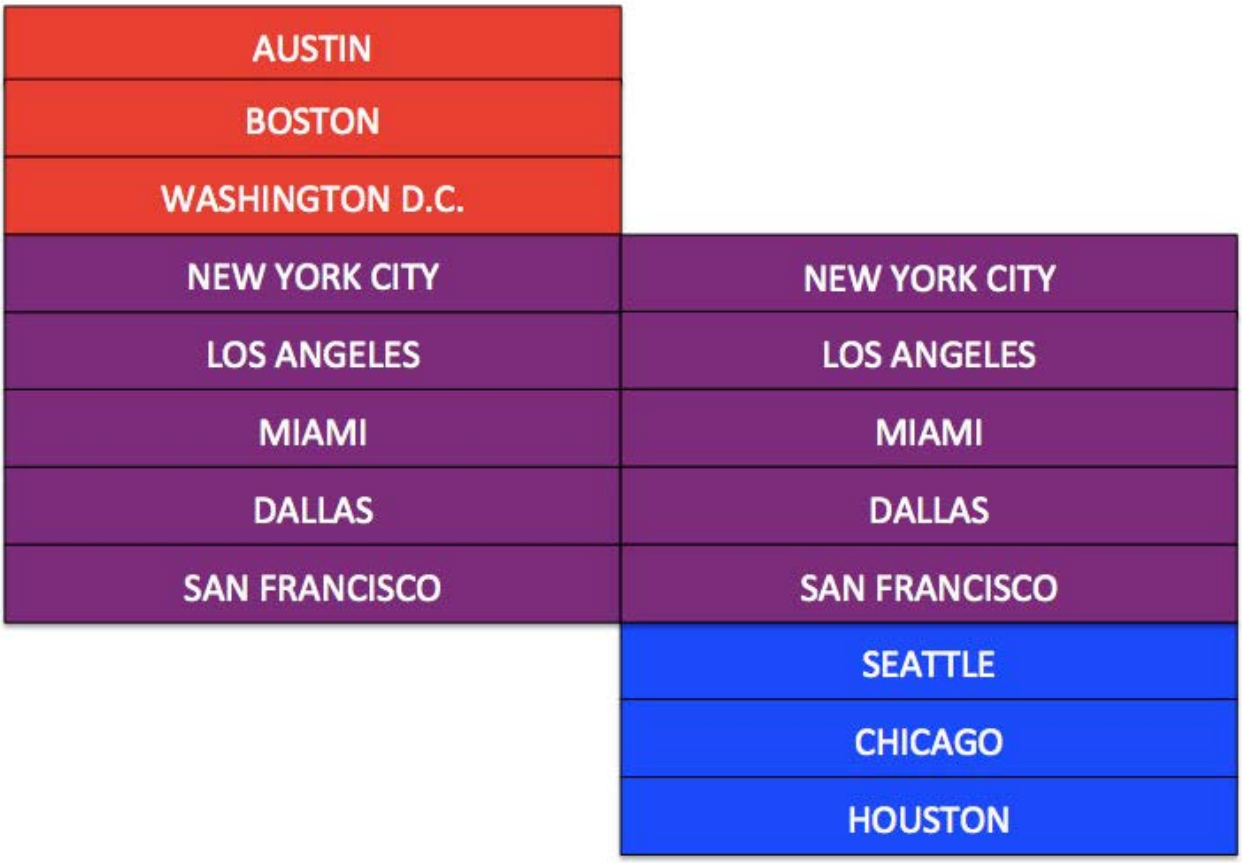
## 右ルックアップ

右ルックアップは、一致する値が指定列にあるすべての行を含む結合済みデータセットを返すという点では左ルックアップに似ていますが、左（ルックアップ）データセットの不一致値を含む行を返すのではなく、右（ベース）データセットのすべての不一致行を返します。

| | |
|---|---|
| NEW YORK CITY | NEW YORK CITY |
| LOS ANGELES | LOS ANGELES |
| MIAMI | MIAMI |
| DALLAS | DALLAS |
| SAN FRANCISCO | SAN FRANCISCO |
| | SEATTLE |
| | CHICAGO |
| | HOUSTON |

## 外部ルックアップ

外部ルックアップは、指定列に一致する値を含むすべての行を含む結合済みデータセット、および一致する値がない両方のデータセットのすべての行を返します:




# ルックアップと結合の一致方法

Intellifusion一致方法アルゴリズムを選択して、**完全**または**ファジー**の関係がある列のテキスト値を照合する方法をカスタマイズします。各オプションについて以下に説明します。

## 完全のオプション

完全のオプションには、自動一致、完全一致、およびカスタム一致の3つの選択があります。

### 自動一致

**自動一致**（デフォルトオプション）：一致するテキスト値の場合、大文字と小文字、語順、単語の前後の句読点が無視されます。一般的に、スペース文字と句読点は単語の境界を定義しますが、重要な例外があります。この照合方法では、単語の前後の句読点が無視されるため、*.ave.*は*AVE*に一致します。ただし、自動一致では、単語内に句読点がある特殊な種類の単語が検出されます。単語内の句読点は単語の境界を定義しないため、句読点は両方のデータセットで正確に一致する必要があります。**自動一致**は、内部の句が適用される次の語が検出されます:

- **ピリオド付きの数値**: 通貨および 12.34 などの浮動小数点付きの数。通貨記号と数字の前後の句読点は無視されます。比較は数値ではなくテキスト比較であるため、「3.0」と「3」は一致しません。ヒント: 問題を減らすには、インポート中に「セルのテキストを数値に解析」オプションを使用するか、列のドロップダウンメニューを使用して列を数値に変換します。
- **Eメールアドレス**: Eメールアドレスは内部のピリオドを含む1つの単語です。

- **ピリオド付きの頭字語**：「U.S.A.」など内側に句読点を含む頭字語は1つの単語として数えます。ただし、単語を比較するときに単語内の句読点が無視されないため、「U.S.A.」は「USA」と一致しません。

## 自動一致の例

次の表は、自動一致の動作を示します。最初の 2 列は値の例を示しています。3番目の列は、これらの値が自動一致と一致するかどうかを示します。完全一致と自動一致の結果が異なる例では、答えが太字で表示されています。

| Val1 | Val2 | Automatic Match |
|---|---|---|
| Mary T | Mary T | yes |
| Jan16 | Jan16 | yes |
| 5,6,7 | 5,6,7 | yes |
| Mary T | Mary M | no |
| U.S. | US | no |
| Mary T | Mary | no |
| Mary T | Marry T | no |
| 1,2,3 | 1 2 3 | no |
| 1.234 | 123.4 | no |
| 5 6 7 | 6 7 5 | **yes** |
| Doe, J | J DOE | **yes** |
| A, B | A B | **yes** |
| A, B, C | C B A | **yes** |
| Main St | MAIN ST. | **yes** |

## 完全一致

**完全一致**オプションを使用する場合は、2つの値のすべての文字が完全に一致する必要があります。

## カスタムマッチ

**カスタムマッチ**オプションを使用して、正確な調整を行います。このオプションで語順、大文字と小文字の区別、ホワイトスペース、および特定の句読点の処理方法を選択できます。語順および大文字と小文字の区別は、ユーザーの選択に応じて無視または保持されます。一般的に、これらのオプションは名前を含むデータに使用されます。語順と大文字と小文字を区別に対する [無視] と [保持] は自由に組み合わせることができます。**[無視]**ボタンと**[保持]**ボタンをクリックすると、その選択に基づいて語順と大文字と小文字の区別の処理例が表示されます。

ホワイトスペースについては、無視、保持、または分割することができます。**[分割]**ボタンを使用すると、データがホワイトスペースで個別の語句に分割されます。ホワイトスペースには、スペースバー、タブキー、およびキャリッジリターンの文字が含まれます。一般的に、ホワイトスペースオプションは住所情報を含むデータの一致を増やすために使用されます。**無視**、**保持**、および**分割**ボタンをクリックすると、その選択に基づいてホワイトスペースの処理例が表示されます。

句読点オプションを使用すると、特定の句読点を無視、保持、または分割できます。デフォルトの句読点の値は、カンマとハイフンです。白い**[その他]**ボタンをクリックすると、デフォルト値に追加できます。句読点を追加するための新しいフィールドが表示されます。デフォルトの句読点または追加した句読点を削除するには、句読点フィールドにカーソルを合わせて、そのフィールドの上部に表示されるオレンジ色の [X] をクリックします。句読点フィールドをすべて削除すると、デフォルトですべての句読点が保持されます。

語順、大文字と小文字、ホワイトスペース、および句読点の設定は、現在のルックアップデータ準備ステップにのみ適用されます。これらの設定は、他のルックアップ手順には適用されません。

## ファジーオプション

**ファジー**オプションは、**標準ルックアップ**でのみ使用可能です。UI のオプションとして表示するには、システム管理者が有効にする必要があります。

**ファジー**オプションでは、編集距離アルゴリズムを使用して、選択した2つの結合キー間の一致の可能性が予測されます。

## ファジーオプションの例

この例では、会社名を含むベースデータセットがあり、そのデータに各会社の住所を追加するとします。

住所情報を含む第2のデータセットがありますが、会社名がベースデータセットにあまり一致していません。

この場合、**ファジー**オプションを使用すると、いずれかのデータセットで「会社名」列のクリーンアップを事前に行わなくても、会社名で結合を容易に作成できます。

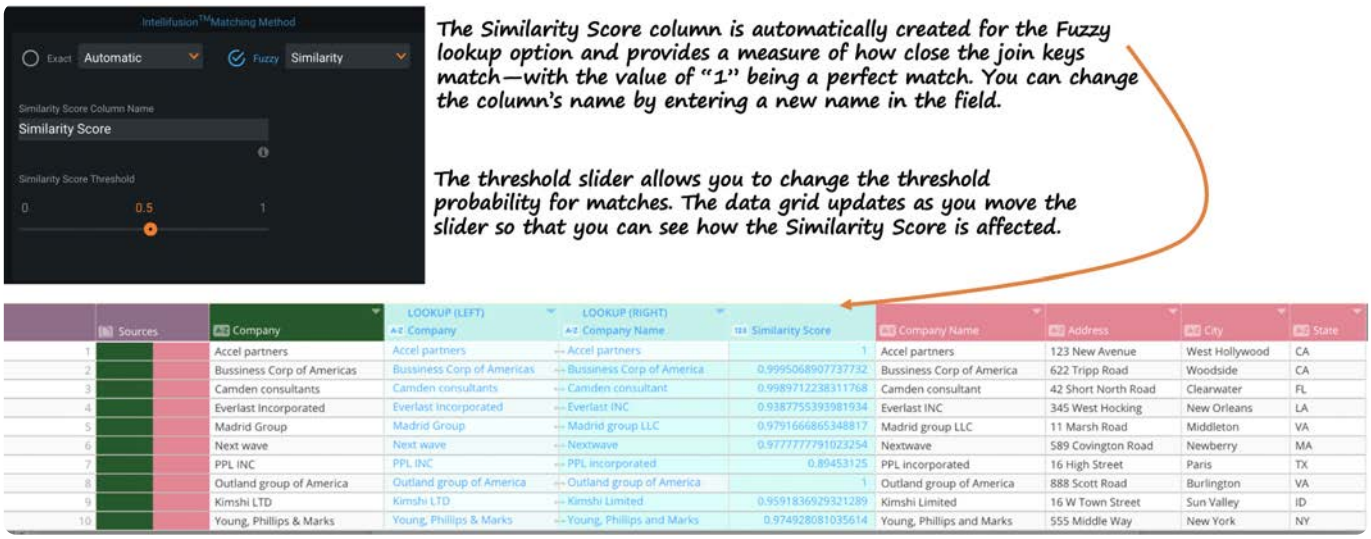
ベースデータセットには会社名が含まれています。

| | Sources | Company |
|---|---|---|
| 1 | | Accel partners |
| 2 | | Bussiness Corp of Americas |
| 3 | | Camden consultants |
| 4 | | Everlast Incorporated |
| 5 | | Madrid Group |
| 6 | | Next wave |
| 7 | | PPL INC |
| 8 | | Outland group of America |
| 9 | | Kimshi LTD |
| 10 | | Young, Phillips & Marks |

アドレスを含む追加データセットには、次のようにリストされた会社名があります。

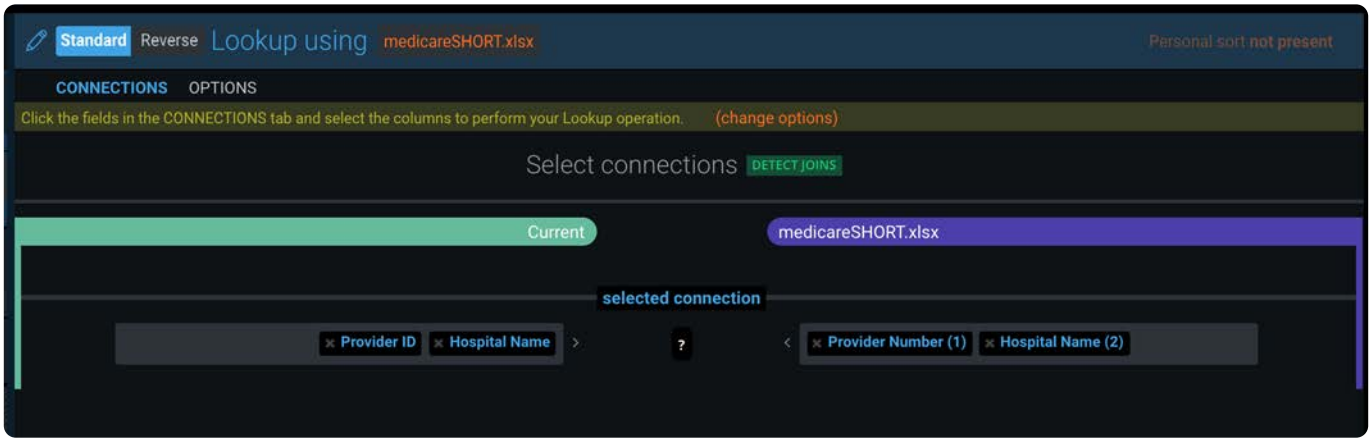
| | Company Name | Address | City | State |
|---|---|---|---|---|
| 1 | Accel partners | 123 New Avenue | West Hollyw... | CA |
| 2 | Bussiness Corp of America | 622 Tripp Road | Woodside | CA |
| 3 | Camden consultant | 42 Short North R... | Clearwater | FL |
| 4 | Everlast INC | 345 West Hocking | New Orleans | LA |
| 5 | Madrid group LLC | 11 Marsh Road | Middleton | VA |
| 6 | Nextwave | 589 Covington Ro... | Newberry | MA |
| 7 | PPL incorporated | 16 High Street | Paris | TX |
| 8 | Outland group of America | 888 Scott Road | Burlington | VA |
| 9 | Kimshi Limited | 16 W Town Street | Sun Valley | ID |
| 10 | Young, Phillips and Marks | 555 Middle Way | New York | NY |

**ファジー**オプションは次のように調整します。




最適なしきい値の一致を見つけて、**保存**をクリックして、データを結合します。

[接続] パネルに列名を追加すると、ファジーオプションの複数の結合キーを同時に作成できることに注意してください。




ただし、両方のデータセットに正確に同じ数の列が含まれている必要があります。そうしないと、エラーが発生して続行できません。

データ結合の後に実行できる追加のオプション：

- 列ツールを使用して、類似性スコア列と会社名列の 1つを削除します。
- 残りの「会社名」列でFiltergramを使用して、名前の形式がすべて正しいことを確認します。会社名を編集する必要がある場合、検索と置換オプションを使用して会社名を正しく更新します。

# ルックアップと結合の例

## 例: 逆ルックアップ

逆引き操作は、現在のデータセットと追加データセットの間で1対多の関係を確立する場合に便利です。この逆引き操作では、現在のベースデータセットがルックアップテーブルとして使用されます。したがって、選択したルックアップデータセットに追加されるのはベースデータセットの最初の一致だけです。

この例では、2つのデータセットがあります。

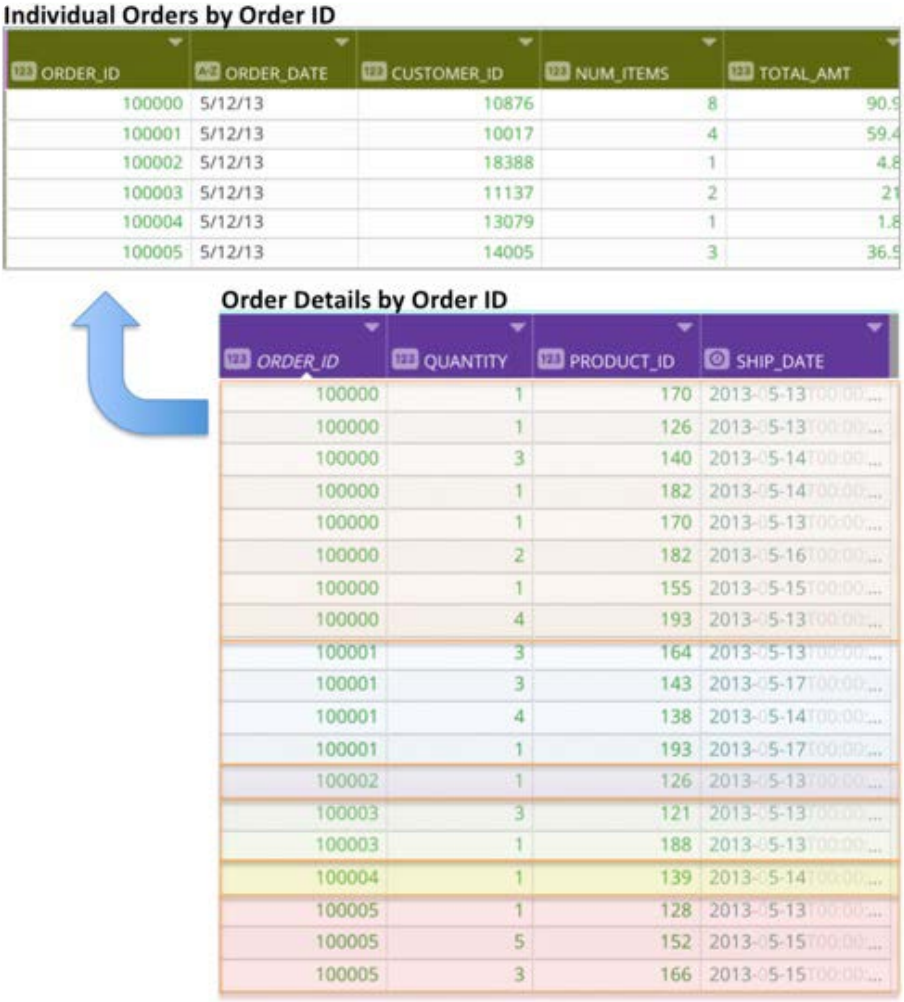
- 一意の**Order ID**を持つ**Orders**データセット

| | ORDER_ID | ORDER_DATE | CUSTOMER_ID | NUM_ITEMS | TOTAL_AMT |
|---|---|---|---|---|---|
| 1 | 100000 | 5/12/13 | 10876 | 8 | 90.9 |
| 2 | 100001 | 5/12/13 | 10017 | 4 | 59.4 |
| 3 | 100002 | 5/12/13 | 18388 | 1 | 4.8 |
| 4 | 100003 | 5/12/13 | 11137 | 2 | 21 |
| 5 | 100004 | 5/12/13 | 13079 | 1 | 1.8 |
| 6 | 100005 | 5/12/13 | 14005 | 3 | 36.5 |
| 7 | 100006 | 5/12/13 | 11037 | 2 | 50 |
| 8 | 100007 | 5/12/13 | 12986 | 4 | 123 |
| 9 | 100008 | 5/12/13 | 12571 | 2 | 25.6 |
| 10 | 100009 | 5/12/13 | 15153 | 1 | 4.5 |
| 11 | 100010 | 5/12/13 | 12196 | 3 | 16.5 |

- **Order ID**に関連付けられたすべてのOrderの詳細を含む**Order Details**データセット

| | ORDER_ID | LINEITEM_ID | PRODUCT_ID | QUANTITY | LIST_PRICE | DISCOUNT | EXT_PRICE | SHIP_DATE |
|---|---|---|---|---|---|---|---|---|
| 1 | 100000 | 1 | 170 | 1 | 5 | 30 | 3.5 | 2013-05-13T00:00:00.000Z |
| 2 | 100000 | 2 | 126 | 1 | 6 | 0 | 6 | 2013-05-13T00:00:00.000Z |
| 3 | 100000 | 3 | 140 | 3 | 5 | 20 | 12 | 2013-05-14T00:00:00.000Z |
| 4 | 100000 | 4 | 182 | 1 | 8 | 20 | 6.4 | 2013-05-14T00:00:00.000Z |
| 5 | 100000 | 5 | 170 | 1 | 5 | 0 | 5 | 2013-05-13T00:00:00.000Z |
| 6 | 100000 | 6 | 182 | 2 | 8 | 0 | 16 | 2013-05-16T00:00:00.000Z |
| 7 | 100000 | 7 | 155 | 1 | 2 | 0 | 2 | 2013-05-15T00:00:00.000Z |
| 8 | 100000 | 8 | 193 | 4 | 10 | 0 | 40 | 2013-05-13T00:00:00.000Z |
| 9 | 100001 | 1 | 164 | 3 | 10 | 0 | 30 | 2013-05-13T00:00:00.000Z |
| 10 | 100001 | 2 | 143 | 3 | 6 | 0 | 18 | 2013-05-17T00:00:00.000Z |
| 11 | 100001 | 3 | 138 | 4 | 1 | 40 | 2.4 | 2013-05-14T00:00:00.000Z |
| 12 | 100001 | 4 | 193 | 1 | 10 | 10 | 9 | 2013-05-17T00:00:00.000Z |
| 13 | 100002 | 1 | 126 | 1 | 6 | 20 | 4.8 | 2013-05-13T00:00:00.000Z |
| 14 | 100003 | 1 | 121 | 3 | 1 | 0 | 3 | 2013-05-13T00:00:00.000Z |
| 15 | 100003 | 2 | 188 | 1 | 20 | 10 | 18 | 2013-05-13T00:00:00.000Z |
| 16 | 100004 | 1 | 139 | 1 | 2 | 10 | 1.8 | 2013-05-14T00:00:00.000Z |
| 17 | 100005 | 1 | 128 | 1 | 5 | 10 | 4.5 | 2013-05-13T00:00:00.000Z |
| 18 | 100005 | 2 | 152 | 5 | 1 | 0 | 5 | 2013-05-15T00:00:00.000Z |
| 19 | 100005 | 3 | 166 | 3 | 10 | 10 | 27 | 2013-05-15T00:00:00.000Z |

**Orders**データセットをベースデータセットとして使用して新しいプロジェクトを作成し、両方のデータベースの"Order ID"列に基づいて**Order Details**のデータを追加します。この場合、**逆引き**オプションを選択し、データベースを照合する列として "Order_ID" を選択します。




**備考**

"Order_ID" を含む不一致（空白）行を保持するには、外部ルックアップまたは左ルックアップのタイプを選択します。内部ルックアップタイプを選択して、これらの行を破棄することもできます。一致しない行の処理方法を定義するルックアップタイプを参照してください。

**重要**

**逆引き**は、ベースデータセットに追加される行の数と、恐らくソート順序に影響します。

## 例: 結合データセット

この選択により、両方のデータセットにわたるすべての一致が結合されます。結合操作は、2つのデータセット間で多対多の関係性を確立する場合に便利です。注意：この選択は、プロジェクトに追加される行の数に影響します。

この例では、2つのデータセットがあります。

- 購入トランザクションIDのセットを含む**Transactions**。各行は購入された書籍を示します。同じ書籍が複数回販売された可能性があるため、同じ書籍で複数のトランザクションIDが存在する場合があることに注意してください。

| | Sources | Transaction ID | Book |
|---|---|---|---|
| 1 | | 1001 | The Elements of Style |
| 2 | | 1002 | All the President's Men |
| 3 | | 1003 | The C Programming Language |
| 4 | | 1004 | The Communist Manifesto |
| 5 | | 1005 | Harry Potter and the Sorcerer's Stone |
| 6 | | 1006 | The Elements of Style |
| 7 | | 1007 | Capitalism and Freedom |
| 8 | | 1008 | Harry Potter and the Sorcerer's Stone |
| 9 | | 1009 | The Elements of Style |
| 10 | | 1010 | Capitalism and Freedom |
| 11 | | 1011 | Harry Potter and the Sorcerer's Stone |
| 12 | | 1012 | The C Programming Language |
| 13 | | 1013 | Harry Potter and the Sorcerer's Stone |
| 14 | | 1014 | Capitalism and Freedom |

- 書籍とその著者のセットを含む**Books**。各書籍には複数の著者がいる場合があるので、1冊の本が複数の行にリストされる場合があります（共著者ごとに1回）。

| | Sources | Book | Author |
|---|---|---|---|
| 1 | | The Elements of Style | Strunk |
| 2 | | The Elements of Style | White |
| 3 | | The C Programming Language | Kernighan |
| 4 | | The C Programming Language | Ritchie |
| 5 | | The Communist Manifesto | Marx |
| 6 | | The Communist Manifesto | Engels |
| 7 | | Capitalism and Freedom | Friedman |
| 8 | | All the President's Men | Woodward |
| 9 | | All the President's Men | Bernstein |
| 10 | | Harry Potter and the Sorcerer's Stone | Rowling |

この例の目的は、著者ごとの購入トランザクション数を決定することです。また、多くの書籍には複数の著者がいるので、この2つのデータベースを結合する場合に選択すべきオプションは、著者の1つの一意の値ではなく、"結合"です。

**Transactions**データセットで新しいプロジェクトを作成します。次に、**Books** データセットで標準の結合操作を行い、データベースを結合する列として "Book" を選択します。

| | Sources | BOOK-TRANSACTIONS Transaction ID | BOOK-TRANSACTIONS Book | BOOKS Book (1) | BOOKS Author |
|---|---|---|---|---|---|
| 1 | | 1003 | The C Programming Language | The C Programming Language | Kernighan |
| 2 | | 1003 | The C Programming Language | The C Programming Language | Ritchie |
| 3 | | 1012 | The C Programming Language | The C Programming Language | Kernighan |
| 4 | | 1012 | The C Programming Language | The C Programming Language | Ritchie |
| 5 | | 1005 | Harry Potter and the Sorcerer's Stone | Harry Potter and the Sorcerer's Stone | Rowling |
| 6 | | 1008 | Harry Potter and the Sorcerer's Stone | Harry Potter and the Sorcerer's Stone | Rowling |
| 7 | | 1011 | Harry Potter and the Sorcerer's Stone | Harry Potter and the Sorcerer's Stone | Rowling |
| 8 | | 1013 | Harry Potter and the Sorcerer's Stone | Harry Potter and the Sorcerer's Stone | Rowling |
| 9 | | 1004 | The Communist Manifesto | The Communist Manifesto | Marx |
| 10 | | 1004 | The Communist Manifesto | The Communist Manifesto | Engels |
| 11 | | 1002 | All the President's Men | All the President's Men | Woodward |
| 12 | | 1002 | All the President's Men | All the President's Men | Bernstein |
| 13 | | 1001 | The Elements of Style | The Elements of Style | Strunk |
| 14 | | 1001 | The Elements of Style | The Elements of Style | White |
| 15 | | 1006 | The Elements of Style | The Elements of Style | Strunk |
| 16 | | 1006 | The Elements of Style | The Elements of Style | White |
| 17 | | 1009 | The Elements of Style | The Elements of Style | Strunk |
| 18 | | 1009 | The Elements of Style | The Elements of Style | White |
| 19 | | 1007 | Capitalism and Freedom | Capitalism and Freedom | Friedman |
| 20 | | 1010 | Capitalism and Freedom | Capitalism and Freedom | Friedman |
| 21 | | 1014 | Capitalism and Freedom | Capitalism and Freedom | Friedman |

**備考**

結合を実行すると、結果のデータセットの行数が大幅に増える可能性があります。行数がプロジェクトの行制限を超えると結合操作は失敗し、エラーメッセージが表示されます。この場合、データセットを結合する前にデータセットの行数を減らすか、プロジェクトの行制限についてシステム管理者に相談してください。

## 例: すべての組み合わせを返す外積結合

これを選択すると、操作で両方のデータセットのすべての行が結合されます。

この例では、3つのデータセットがあります。

- すべての顧客IDを含む**Customer Master**。

| | CUSTOMER_ID |
|---|---|
| 1 | Customer_Alfa |
| 2 | Customer_Kilo |
| 3 | Customer_Oscar |
| 4 | Customer_Echo |
| 5 | Customer_Delta |
| 6 | Customer_Foxtrot |
| 7 | Customer_Tango |
| 8 | Customer_Bravo |
| 9 | Customer_Hotel |

・すべての製品および関連付けられたIDを含む**Products Master**。

| | PRODUCT_ID | Product |
|---|---|---|
| 1 | 120 | ADDE chair |
| 2 | 126 | AINA fabric |
| 3 | 140 | FADO table la... |
| 4 | 150 | FEJS wall clock |
| 5 | 155 | HENSVIK cabi... |
| 6 | 160 | JANINGE armc... |
| 7 | 170 | KLYSA wall clo... |
| 8 | 180 | MAJJE throw |
| 9 | 182 | MUSTIG glass |

・顧客IDおよびそのIDに関連付けられた購入製品と数量を含む**Customer Orders**。

| | CUSTOMER_ID | prod ID | qty |
|---|---|---|---|
| 1 | Customer_Alfa | 120 | 1 |
| 2 | Customer_Kilo | 120 | 1 |
| 3 | Customer_Oscar | 140 | 3 |
| 4 | Customer_Echo | 140 | 1 |
| 5 | Customer_Delta | 150 | 1 |
| 6 | Customer_Foxtrot | 150 | 2 |
| 7 | Customer_Tango | 160 | 1 |
| 8 | Customer_Bravo | 170 | 4 |
| 9 | Customer_Hotel | 180 | 3 |

この例の目的は、すべての顧客とすべての製品を含むマスタープロジェクトを作成し、各顧客が購入していない製品をすべて特定することです。

**Customer Master**をベースデータセットとして新しいプロジェクトを作成します。次に、**外積**結合操作を**Products Master**データセットに対して行います。外積操作の対象列として "Customer_ID" および "Product_ID" を選択します。

これで、すべての顧客とすべての製品を含むマスターデータセットが作成されました。

| | CUSTOMER_ID | prod ID | qty |
|---|---|---|---|
| 1 | Customer_Alfa | 120 | 1 |
| 2 | Customer_Kilo | 120 | 1 |
| 3 | Customer_Oscar | 140 | 3 |
| 4 | Customer_Echo | 140 | 1 |
| 5 | Customer_Delta | 150 | 1 |
| 6 | Customer_Foxtrot | 150 | 2 |
| 7 | Customer_Tango | 160 | 1 |
| 8 | Customer_Bravo | 170 | 4 |
| 9 | Customer_Hotel | 180 | 3 |

次に、**標準のルックアップ**操作を **Customer Orders** データセット実行します。"Customer_ID" と "Product_ID" をルックアップの対象列として選択し、すべての不一致行を保持します。一致しない行の処理方法を定義する**ルックアップタイプ**を参照してください。

結果のデータセットを使用して、各顧客がまだ購入していない製品をすべて簡単に特定できます。

| | Sources | CUSTOMER_MASTER CUSTOMER_ID | PRODUCTS_MASTER PRODUCT_ID | PRODUCTS_MASTER Product |
|---|---|---|---|---|
| 1 | | Customer_Alfa | 120 | ADDE chair |
| 2 | | Customer_Alfa | 126 | AINA fabric |
| 3 | | Customer_Alfa | 140 | FADO table lamp |
| 4 | | Customer_Alfa | 150 | FEJS wall clock |
| 5 | | Customer_Alfa | 155 | HENSVIK cabinet |
| 6 | | Customer_Alfa | 160 | JANINGE armchair |
| 7 | | Customer_Alfa | 170 | KLYSA wall clock |
| 8 | | Customer_Alfa | 180 | MAJJE throw |
| 9 | | Customer_Alfa | 182 | MUSTIG glass |
| 10 | | Customer_Kilo | 120 | ADDE chair |
| 11 | | Customer_Kilo | 126 | AINA fabric |
| 12 | | Customer_Kilo | 140 | FADO table lamp |
| 13 | | Customer_Kilo | 150 | FEJS wall clock |
| 14 | | Customer_Kilo | 155 | HENSVIK cabinet |
| 15 | | Customer_Kilo | 160 | JANINGE armchair |
| 16 | | Customer_Kilo | 170 | KLYSA wall clock |
| 17 | | Customer_Kilo | 180 | MAJJE throw |
| 18 | | Customer_Kilo | 182 | MUSTIG glass |

このプロジェクトの **Customer Orders** データセットを更新するたびにグリッド上のデータが自動的に更新され、新しい購入が反映されます。時間の経過に伴うすべての購入情報を取得するために、プロジェクトレンズを作成して各結果の AnswerSet を公開できます。

**備考**

ルックアップソースの各行がプロジェクトのベースデータセットの各行に追加されるため、**外積**操作によってプロジェクトに追加される行数は大幅に増加します。行数がプロジェクトの行制限を超えると結合操作は失敗し、エラーメッセージが表示されます。この場合、データセットを組み合わせる前にデータセットの行数を減らすか、プロジェクトの行制限についてシステム管理者に相談してください。

# 追加ツールの操作

追加ツールを使用して、ベースデータセットの最後に追加する行を含む追加のデータセットを選択できます。2つのデータセット間の列の一致をカスタマイズできます。追加されたデータセットのすべての列が現在のデータセットの列と一致する場合、元のデータセットの一連の列は結果データセットで変更されません。追加されたデータセットの列を一致させないままにすると、それらの列は結果データセットの新しい列になります。

# 列の更新

Data Prepでデータの準備をするときに、列に変更を加える必要がある場合があります。**列**ツール使用して、プロジェクトでの列名、順序、および可用性を編集することができます。

**備考**

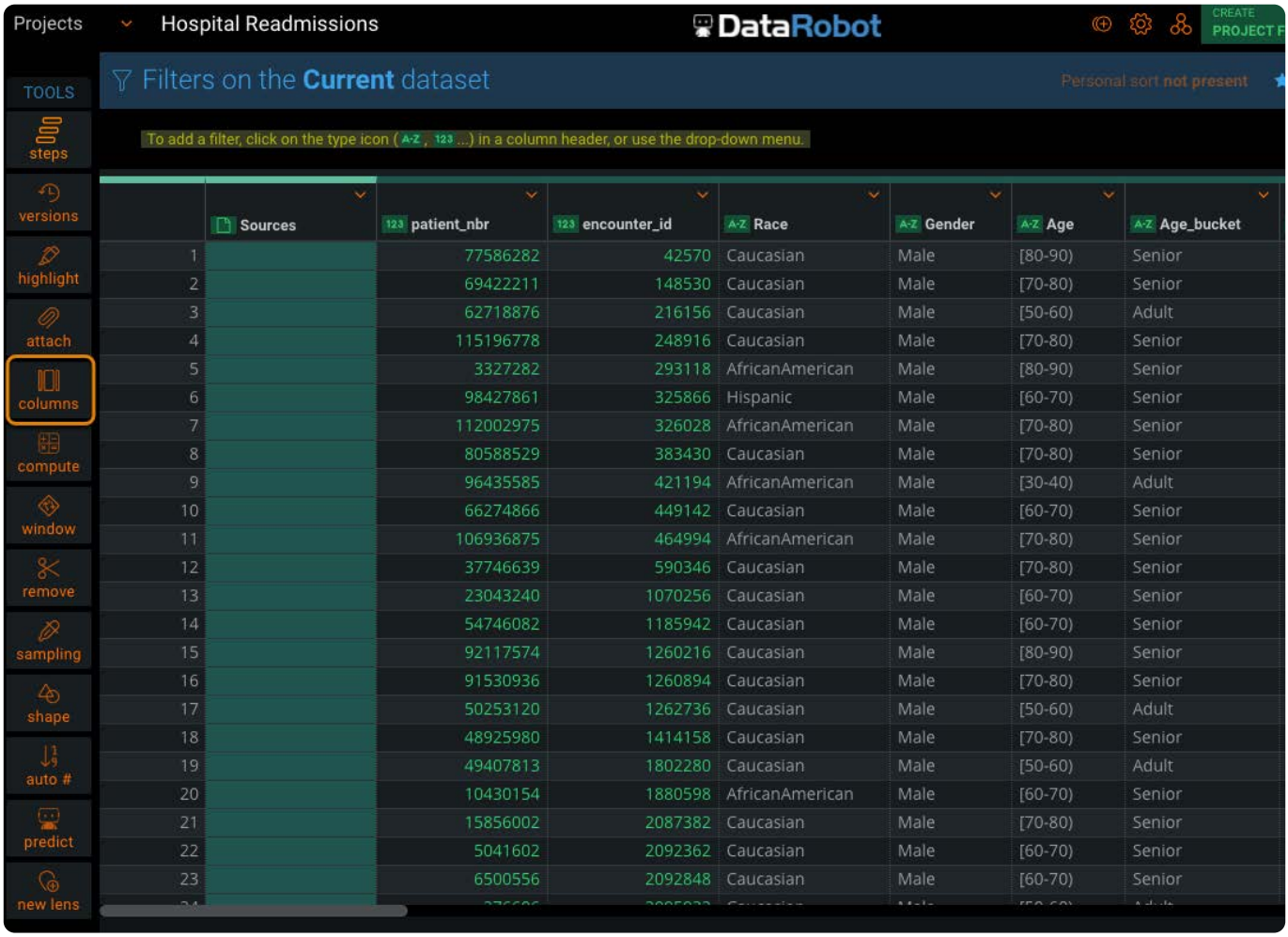
**列**ツールを使用すると、行全体を操作できますが、列データを操作する必要がある場合、列メニュー、および**フィルター**ペインと**列**ペインを使用します。詳細については、列データの操作を参照してください。

## 列ツールの操作

**列**ツールは複数の目的を果たします。プロジェクトに現在ある列と、各列のソースとタイプ文字列、数値、日時が表示されます。**列**ツールには、次の機能もあります。

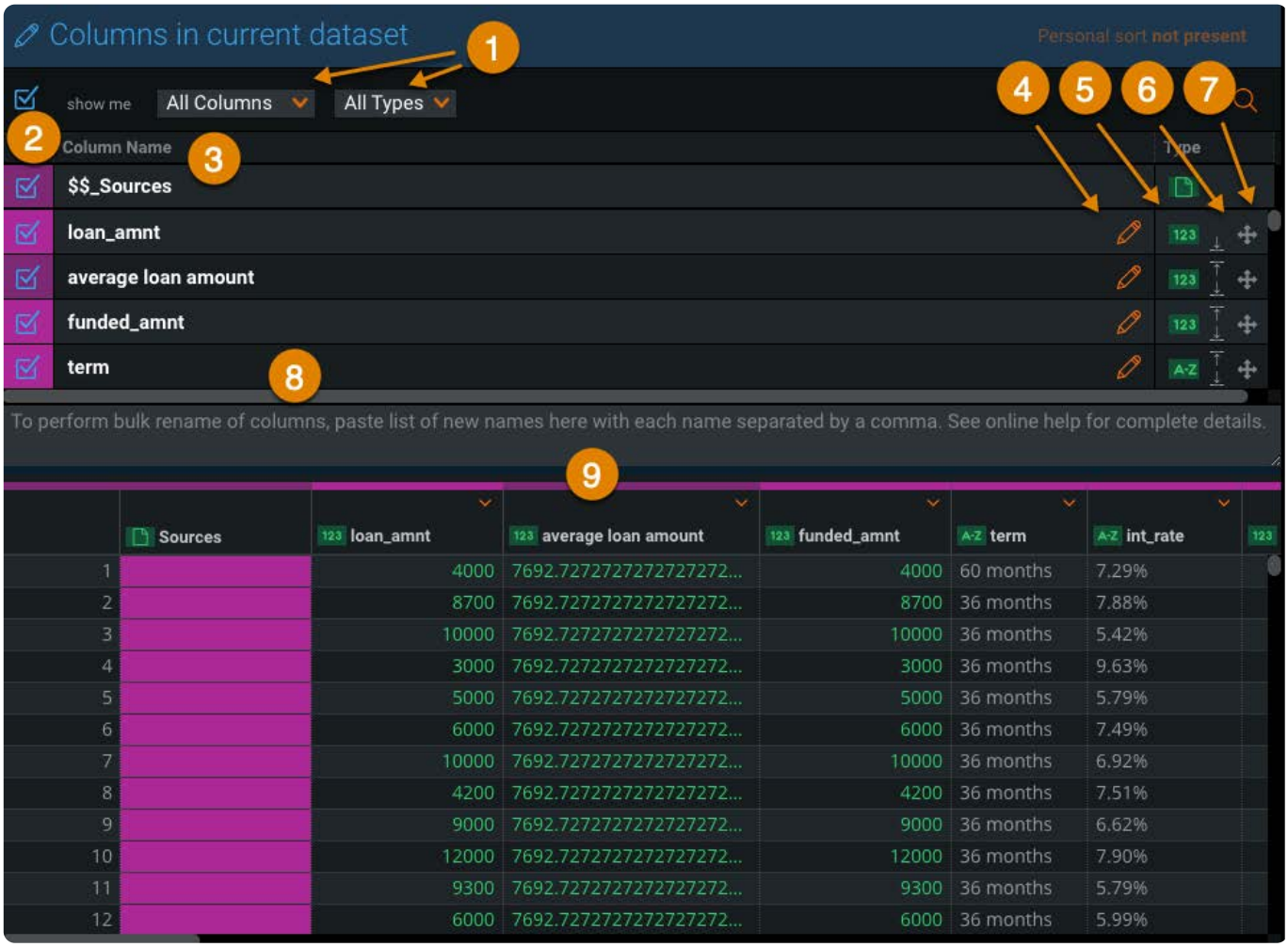
- 列の名前を変更する。
- 列を並べ替える。
- 列を削除する。

**列**ツールにアクセスするには、**ツール**バーで**列**をクリックします。




| | Sources | patient_nbr | encounter_id | Race | Gender | Age | Age_bucket |
|---|---|---|---|---|---|---|---|
| 1 | | 77586282 | 42570 | Caucasian | Male | [80-90) | Senior |
| 2 | | 69422211 | 148530 | Caucasian | Male | [70-80) | Senior |
| 3 | | 62718876 | 216156 | Caucasian | Male | [50-60) | Adult |
| 4 | | 115196778 | 248916 | Caucasian | Male | [70-80) | Senior |
| 5 | | 3327282 | 293118 | AfricanAmerican | Male | [80-90) | Senior |
| 6 | | 98427861 | 325866 | Hispanic | Male | [60-70) | Senior |
| 7 | | 112002975 | 326028 | AfricanAmerican | Male | [70-80) | Senior |
| 8 | | 80588529 | 383430 | Caucasian | Male | [70-80) | Senior |
| 9 | | 96435585 | 421194 | AfricanAmerican | Male | [30-40) | Adult |
| 10 | | 66274866 | 449142 | Caucasian | Male | [60-70) | Senior |
| 11 | | 106936875 | 464994 | AfricanAmerican | Male | [70-80) | Senior |
| 12 | | 37746639 | 590346 | Caucasian | Male | [70-80) | Senior |
| 13 | | 23043240 | 1070256 | Caucasian | Male | [60-70) | Senior |
| 14 | | 54746082 | 1185942 | Caucasian | Male | [60-70) | Senior |
| 15 | | 92117574 | 1260216 | Caucasian | Male | [80-90) | Senior |
| 16 | | 91530936 | 1260894 | Caucasian | Male | [70-80) | Senior |
| 17 | | 50253120 | 1262736 | Caucasian | Male | [50-60) | Adult |
| 18 | | 48925980 | 1414158 | Caucasian | Male | [70-80) | Senior |
| 19 | | 49407813 | 1802280 | Caucasian | Male | [50-60) | Adult |
| 20 | | 10430154 | 1880598 | AfricanAmerican | Male | [60-70) | Senior |
| 21 | | 15856002 | 2087382 | Caucasian | Male | [70-80) | Senior |
| 22 | | 5041602 | 2092362 | Caucasian | Male | [60-70) | Senior |
| 23 | | 6500556 | 2092848 | Caucasian | Male | [60-70) | Senior |

**現在のデータセットの列**ペインが表示されます。




以下は、プロジェクトの列を編集するときに使用する要素の概要です。

| | 要素 | 説明 |
|---|---|---|
| 1 | **列フィルター** | **現在のデータセットの列**リストを次の要素別にフィルターします。<br>・選択された列<br>・名前が変更された列<br>・データ型 |
| 2 | **列セレクター** | 削除する各列のセレクターをクリアします。列を保持するセレクターをチェックします。列を非表示にすると、公開時にAnswerSetから列が削除されます。 |
| 3 | **現在のデータセットリストの列** | 列とそれに含まれるデータの型を表示します。列は、データに表示される順序で一覧表示されます。 |
| 4 | **列名の編集** | 列の名前を更新します。 |

| | 要素 | 説明 |
|---|---|---|
| 5 | **タイプ** | 列のデータ型を表示します。 |
| 6 | **先頭に移動 / 一番下に移動** | 列をデータセットの最初または最後に移動します。 |
| 7 | **移動** | 列を新しいロケーションにドラッグします。 |
| 8 | **一括名前変更** | 単一のコンマ区切り文字列を使用して、すべての列の名前を変更します。 |
| 9 | **データプレビュー**ペイン | プロジェクト内のデータを表示します。データを準備すると、データが変化するのがわかります。 |

# 列の名前の変更

## 個々の列の名前の変更

列の名前を変更するには、次のステップを実行します：

1. **ツール**バーから、**列**をクリックします。

   **現在のデータセットの列**ペインが表示されます。
2. 名前を変更する列の名前をクリックするか、鉛筆アイコンをクリックします。
3. 列の新しい名前を入力して、**Enter**をクリックします。

   **古い列名**セクションが表示され、列の元の名前が表示されます。**データプレビュー**ペインには、更新された列名が表示されます。

   新しい名前を元に戻す場合は、次の手順を実行します。**リセット**をクリックすると、列名が元の名前にリセットされます。
4. 左上にある**保存**をクリックします。

   変更はプロジェクトのステップとして保存されます。**データプレビュー**ペインで、列が更新されます。

## 一括で列の名前を変更

一括名前変更機能を使用すると、すべての列の名前を一度に変更できます。

列のリストの名前を変更するには、次の手順に従います。

- **現在のデータセットの列**リストの下にある一括名前変更フィールドに、新しい列名をコンマで区切って入力します。

  リストの列名は、それに応じて更新されます。コンマで区切られたヘッダーファイルから新しい列名を貼り付けて、データセット内のすべての列の名前をすばやく変更することもできます。

# 列の並べ替え

列の場所を変更するには、次の手順に従います：

1. **ツール**バーから、**列**をクリックします。

   **現在のデータセットの列**ペインが表示されます。

2. **タイプ**セクションで、ポインターを移動する列の移動アイコンの上に置き、列を新しいロケーションにドラッグします。

   **データプレビュー**ペインに、列が新しい位置で表示されます。

   代わりに**先頭に移動**または**一番下に移動**アイコンを使用して、列を最初または最後の位置に移動します。

3. **保存**をクリックします。

   変更はプロジェクトのステップとして保存されます。**データプレビュー**ペインで、列が更新されます。

# 列の削除

次の手順は、プロジェクトから列を削除する方法を示します。

> **警告**
>
> 列を削除すると、プロジェクトでその列を利用できなくなります。後続のステップでその列を利用できなくなるため、後続のステップが削除された列に依存している場合、それらのステップでエラーが発生します。削除された元のステップに戻り、データに含めるために列を再度選択することで、削除された列を再び使用可能にすることができます。列を削除するのではなく、列演算子の非表示を使って、列を非表示にすることができます。

列を削除するには、次のステップを実行します：

1. **ツール**バーから、**列**をクリックします。

   **現在のデータセットの列**ペインが表示されます。

2. 削除する各列の左側にある列セレクターをクリアします。

   **現在のデータセットの列**リストでは、列には陰影が付けられおり、下の**データプレビュー**ペインから削除されています。

3. **保存**をクリックします。

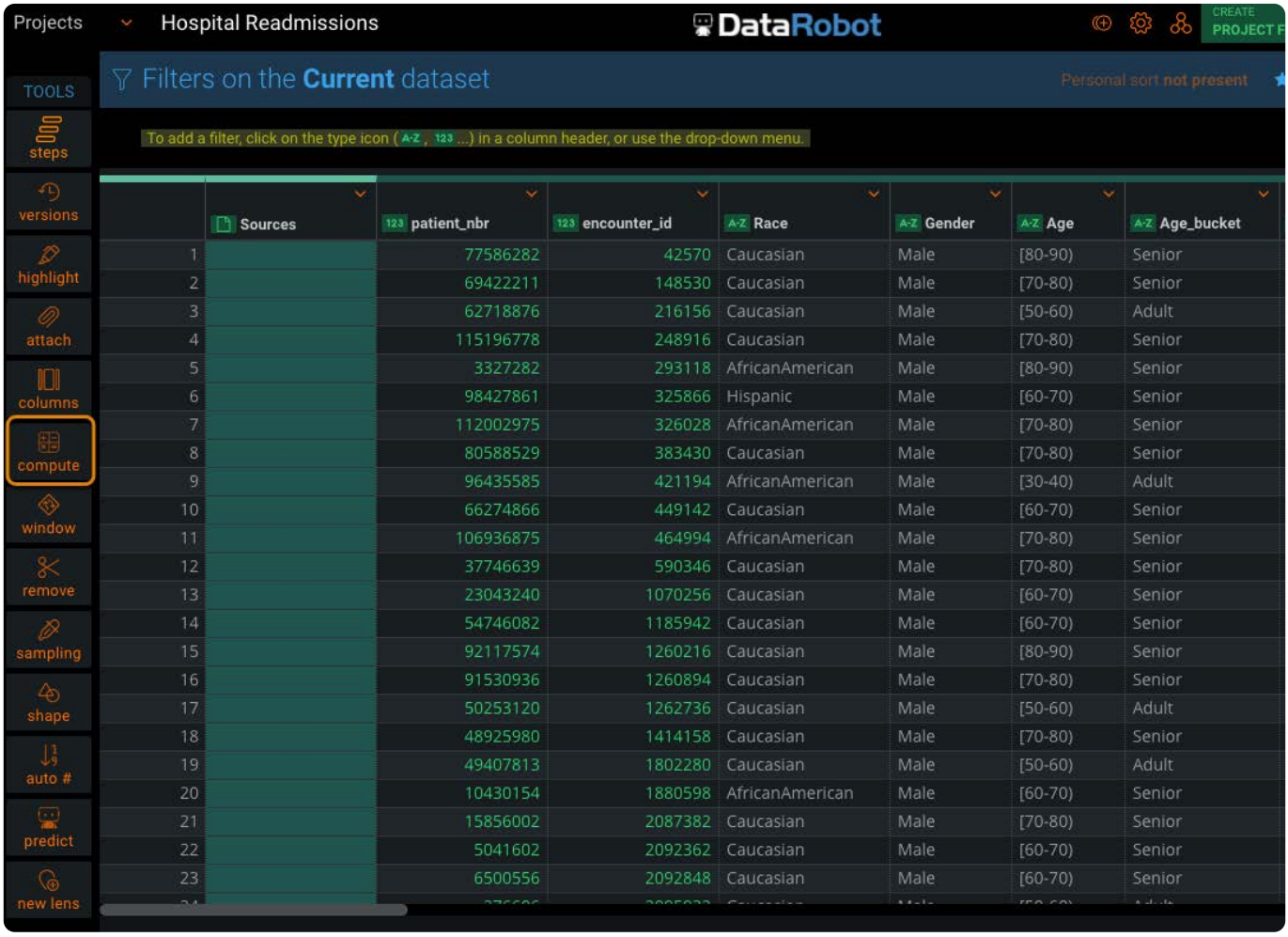
   変更はプロジェクトのステップとして保存されます。

# 列の計算

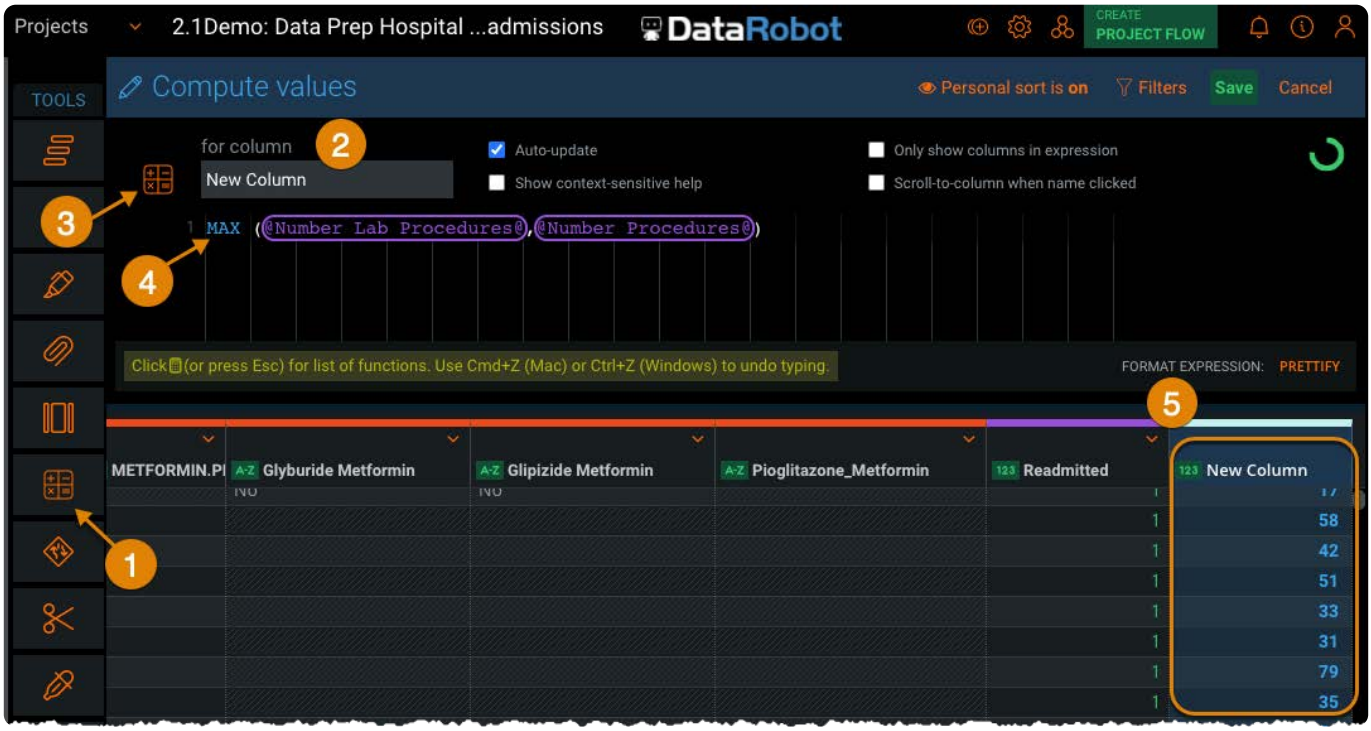
Data Prepを使用すると、データセット内の既存の列に関数を適用して、新しい列を追加できます。

## 計算ツールの操作

Data Prep**計算**ツールにアクセスするには、プロジェクト**ツール**バーで**計算**をクリックします：




ここでは、**値を計算する**ペインの基本的な要素の概要を示します。




| | 要素 | 説明 |
|---|---|---|
| 1 | **計算**ツール | **計算**をクリックして、**値を計算する**ペインにアクセスします。 |
| 2 | **列**フィールドの場合、 | 新しい列の名前を入力します。 |
| 3 | [関数] メニュー | サポートされている関数のリストを含む関数のメニューにアクセスするためにクリックします。詳細については、計算済み列を追加する方法とサポート済み関数を参照してください。 |
| 4 | 式行 | この行を使用して、新しい列の値を計算するために式を構築します。シンプルな式を入力するか、列と関数を使用して高度な式を構築します。この例では、 MAX 関数を示します。列タイトルは 2 つの「アット」（@）記号で囲まれています。数式エラーがある場合、式の行の下にエラーを説明するエラーメッセージが表示されます。<br>**注意:** Data Prepは大文字と小文字が区別されます。式の行に列のタイトルを入力するときは、入力の大文字または小文字と列の名前の大文字または小文字とが一致している必要があります。 |
| 5 | 計算された列 | 数式の結果を含む新しい列です。 |

| | 要素 | 説明 |
|---|---|---|
| 6 | タイマー | タイマーには、データのプレビューが更新されるまでの秒読みが表示されます。タイマーは入力を停止すると表示され、式の入力を再開したりグリッドが更新されたりするとリセットされます。この機能は、大きな計算列式を作成にキー ストロークのたびにグリッドが更新されないようにするときに特に便利です。タイマーを無効にする場合は、新しい列フィールドの隣にある**自動更新**チェックボックスの選択を解除します。タイマーの代わりに**更新**ボタンが利用できるようになり、手動でクリックすることでグリッドを更新できるようになります。 |

# 計算されたカラムの追加方法

計算済み列をプロジェクトに追加するには、次の手順を実行します：

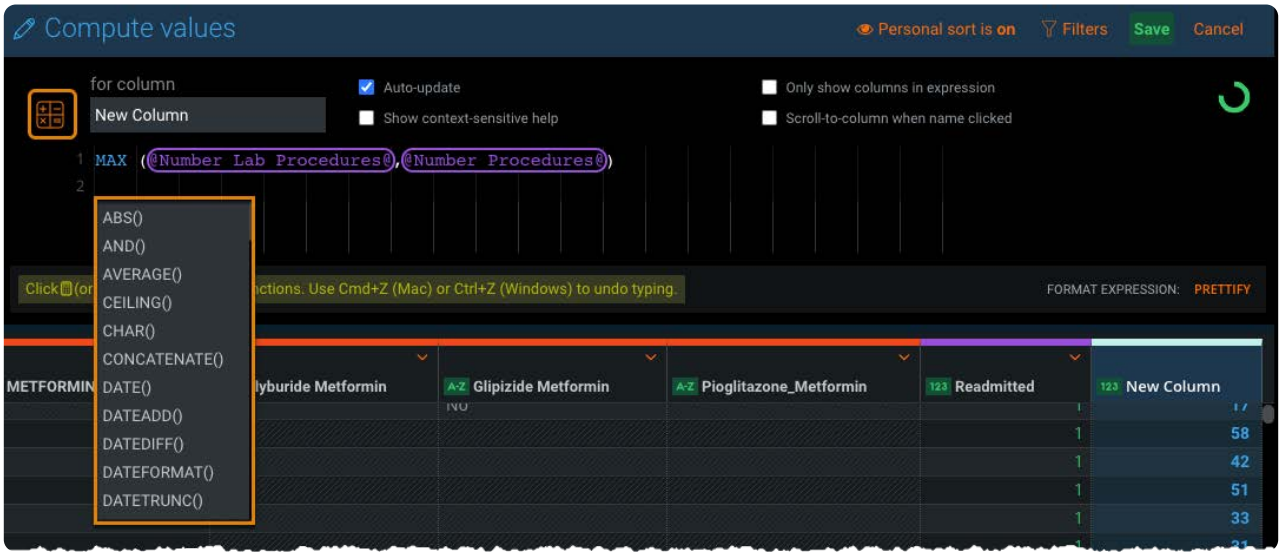
1. **ツール**バーで**計算**をクリックします。

   **値を計算する**ペインが表示されます。

2. **列**フィールドに新しい列の名前を入力します。

3. 数式に関数を追加するには、次のいずれかの方法を使用します。

   - **値を計算する**ペインの左上にある関数アイコンをクリックし、リストから関数を選択します：




   - 式の行で、**ESC**キーを押して、リストから関数を選択します。
   - 式の行に、関数を入力します。

   使用可能な関数の詳細については、サポート済み関数を参照してください。

4. 数式をかっこに列を追加するには、次のいずれかの方法を使用します。

   - [データプレビュー] から列名をクリックします。
   - 式の行で「 @ 」と入力し、**ESC**キーを押して、リストから列を選択します。
   - 式の行に、2 つの @ 文字の間に列の名前を入力します。例: 列のタイトルが Hire Date の場合は、 @Hire_Date@ と入力します。

5. **保存**をクリックします。

計算された値を含む新しい列はプロジェクトにコミットされます。

**備考**

数式エラーがある場合、ステープツールにエラーアイコン（）が表示されます。この場合、レンズを作成して保存できますが、AnswerSetに公開することはできません。

## 構文規則

- 演算子記号の前後には空白文字を入力します（例: `1+1` ではなく、`1 + 1` ）。
- テキストは二重引用符で囲んでください（例: `"Hello"` ）。

  引用符記号を文字として入力する必要がある場合は、引用符記号の前にバックスラッシュを入力します。バックスラッシュを入力するには、バックスラッシュの前にさらにバックスラッシュを入力します（バックスラッシュ文字をバックスラッシュでエスケープする必要があるからです）。 たとえば、以下のテキストを入力する場合を考えてみます： `Go to "C:\windows"` この場合、以下のように入力します: `"Go to \"C:\\windows\""`

# 対応している関数

対応している関数のリストを示します。関数を選択して、説明、構文、および例を表示します。

## 日付/時刻関数

- DATE
- DATEADD
- DATEDIFF
- DATEFORMAT
- DATETRUNC
- DATEVALUE
- DAY
- DAYOFWEEK
- DAYOFYEAR
- ENDOFMONTH
- FROMUNIXTIME
- HOUR
- MAXDATE
- MIDNIGHT
- MINDATE
- MINUTE
- MONTH

- NETWORKDAYS
- NOW
- QUARTER
- SECOND
- SETTIMEZONE
- TODAY
- WEEKOFYEAR
- WORKDAY
- YEAR

## 情報関数

- FIRSTNONBLANK
- ISBLANK
- ISDATE
- ISNULL
- ISNUMBER
- ISTEXT

## 論理関数

- AND
- IF
- IFERROR
- NOT
- OR

## 数学関数

- 列を任意の数で乗算または除算します。
- 列に任意の数を加算または除算します。
- ABS
- CEILING
- EXP
- FACTORIAL
- FLOOR
- INT
- LN

- LOG
- LOG10
- MOD
- POWER
- Round
- ROUNDDOWN
- ROUNDPERC
- ROUNDUP
- Sign
- SQRT
- SUM

## 統計関数

- AVERAGE
- MAX
- MEDIAN
- MIN
- MODE
- STDEV
- STDEVP
- VAR
- VARP

## テキスト関数

- CHAR
- CONCATENATE
- FIND
- HASHVALUE
- LEFT
- LEN
- LOWER
- MID
- PADLEFT
- PADRIGHT
- REGEXP
- REPEAT

- REPLACE
- REVERSE
- RIGHT
- SEARCH
- STR
- SUBSTITUTE
- TRIM
- TRIMLEFT
- TRIMRIGHT
- UPPER
- VALUE

## 比較演算子

- Equal To
- Greater Than
- Greater ThanまたはEqual To
- Less Than
- Less ThanまたはEqual To
- 等しくない

詳細については演算子の比較を参照してください。

## カスタム関数

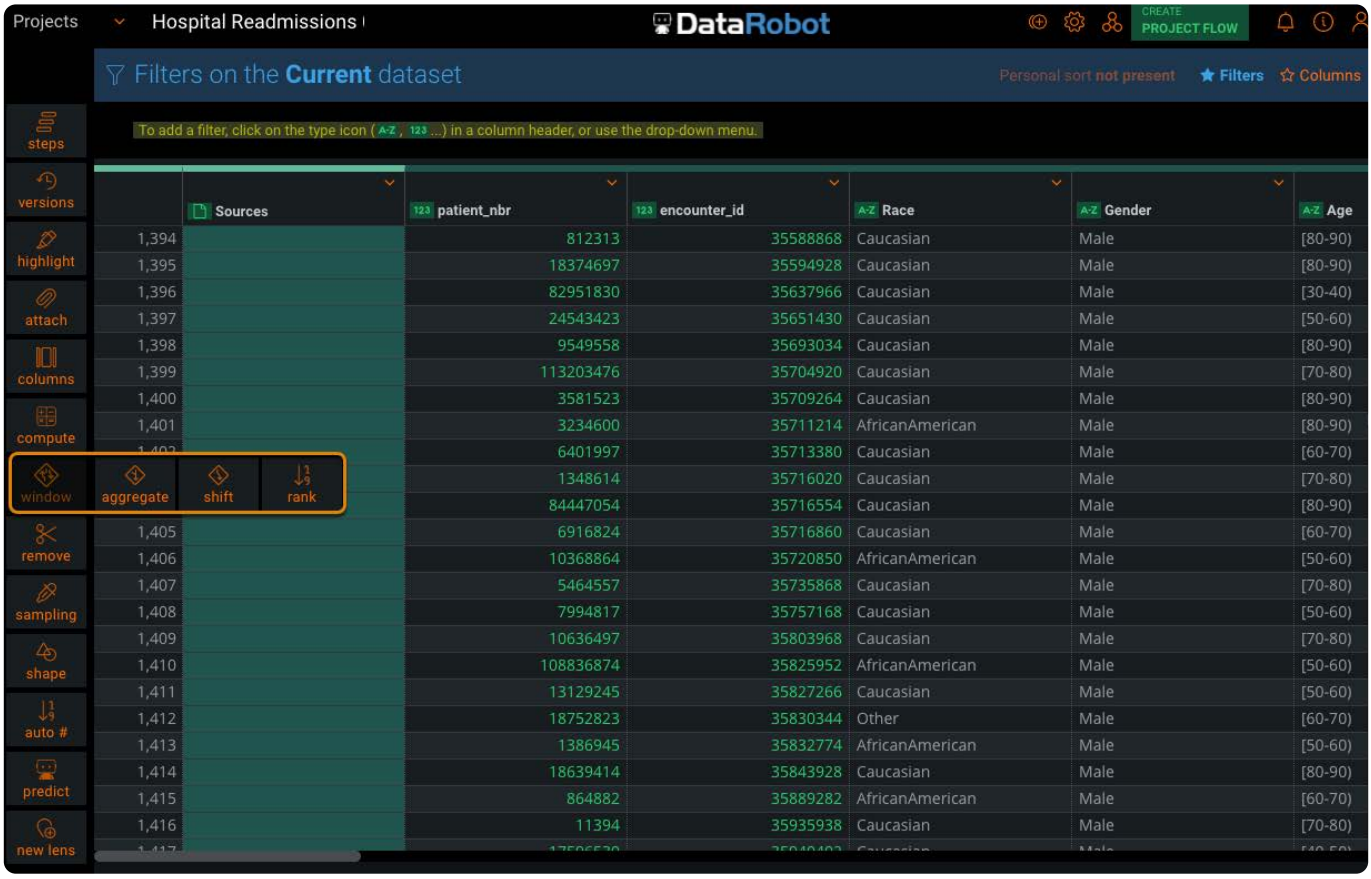
組織がカスタム関数を開発して、インストールしている場合、データセット内の既存の列にカスタム関数を適用して新しい列を追加できます。詳細については、計算されたカスタ列関数を参照してください。

# ウィンドウでグループ化

Data Prepウィンドウ関数は、特定の機能を実行するために、ウィンドウと呼ばれる行のセットをグループ化できるようにするツールのセットです。

## ウィンドウの操作

ウィンドウを操作するには、プロジェクト**ツール**バーの**ウィンドウ**ツールにカーソルを合わせ、集計、シフト、またはランクツールを選択します。




**備考**

**ウィンドウ**ツールが表示されない場合、システム管理者に連絡して、このオプションを有効にするよう依頼してください。

# 集計ツールの操作

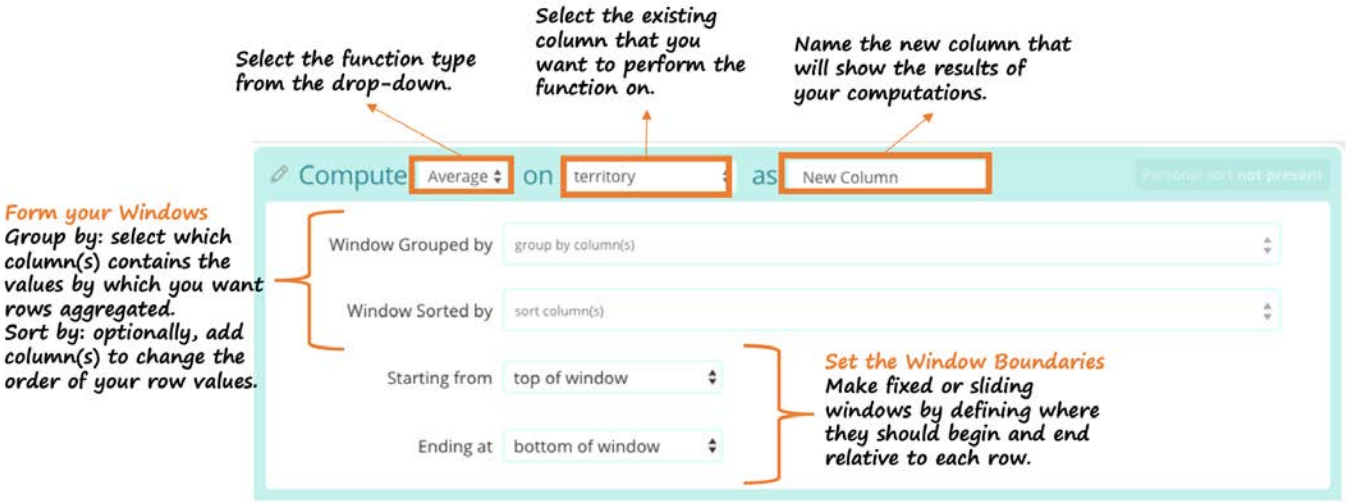
**集計**ツールを使用すると、特定の関数を計算する目的で行のセットをグループ化できます。Data Prepでは、数式を記述する代わりに、ポイントアンドクリック操作で計算が作成されます。

Data Prepでは、ウィンドウは計算に参加する行のセットとして定義されます。ウィンドウは次のように識別できます。

- 固定ウィンドウ：1つ以上の列に共通する近似値に基づく行のグループ化。
- スライディングウィンドウ：現在の行を基準とした行のグループ化（ローリングまたは移動関数）。

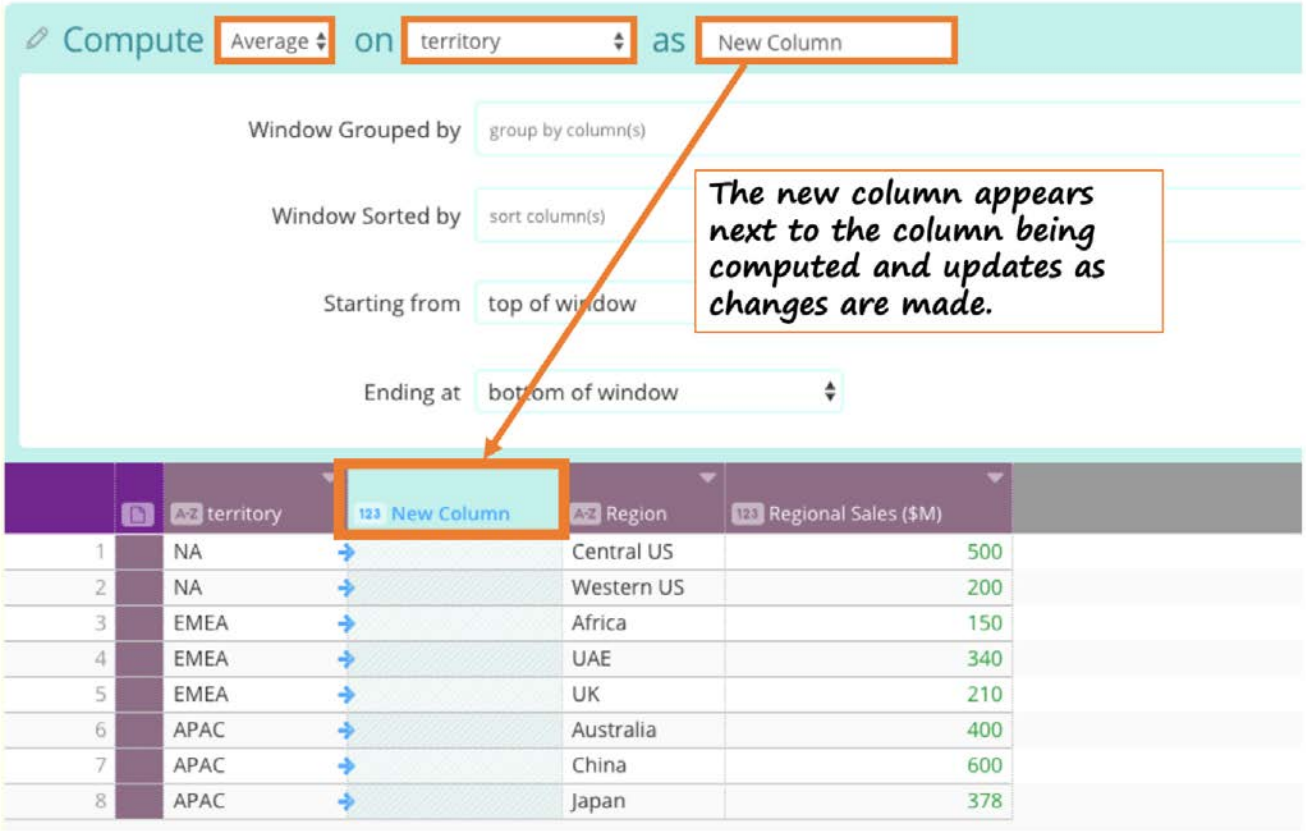
> **備考**
>
> GroupByの集計は、ウィンドウ集計関数とは異なります。主な違いの1つは、GroupByでは行の数が減って集計行の値のみが残りますが、ウィンドウ関数ではデータセットの行ごとに集計されます。

## ウィンドウ集計の定義




1. 出力列を作成します。

   この新しい列のプレビューは、関数の実行対象として選択した列の横に表示されます。




- **関数タイプ：**ドロップダウンに[Average（平均）]、[Sum（合計）]、[Count（カウント）]、[First（最初）]、[Last（最後）]、[Min（最小）]、[Max（最大）]、[Median（中央値）]が表示されます。デフォルトは平均です。
- 計算を行う**列を選択**します。[Average（平均）]、[Sum（合計）]、[Min（最小）]、[Max（最大）]、[Median（中央値）]の列タイプは数値である必要があります。[Count（カウント）]、[First（最初）]、[Last（最後）]の列タイプは任意です。
- 出力列に**名前**を付けます。

2. 行をグループ化してソートします。

- **Group by（グループ化条件）：**ドロップダウンから、行のグループ化基準となる値を含む列を選択します。少なくとも1つの列を選択しますが、列の数を増やしたり減らしたりしてウィンドウのサイズを調整できます。列の集計行は昇順で表示されます。
- **Sort by（ソート基準）**（オプション）**：**ドロップダウンから列を選択して、集計内での行のスタック順序を定義します。選択した列の青い矢印をクリックすると、値の昇順ソートと降順ソートを切り替えることができます。これが特に重要になるのは、ソートの適用後に最初／最後が選択される[First（最初）]関数と[Last（最後）]関数の場合です。

3. 関数の境界を設定します。

- **Top of window（ウィンドウ上限）：**集計の各行で、関数が属するウィンドウの最初の行の値から関数が開始します。
- **Bottom of window（ウィンドウ下限）：**集計の各行で、関数が属するウィンドウの最後の行の値で関数が終了します。
- **Current row with offset of __ rows（現在の行から__行オフセット）**：これを上限、下限、または両方の境界として選択すると、計算対象行（現在行）を基準に変化するローリングウィンドウが作成されます。計算の開始行または終了行は、計算対象行から指定された数だけ上または下にある行になります。たとえば、開始行および終了行として現在行とオフセット0を選択すると、その現在行の単一値を基準に各行が計算されます。オフセットが-1である場合は、現在行と1つ前の行を対象に関数が計算されます。以下に例を示します。

- 選択可能なオプション：
  - ウィンドウ上限からウィンドウ下限まで（固定ウィンドウ）
  - ウィンドウ上限からオフセットありの現在行まで（スライディングウィンドウ）
  - オフセットありの現在行からウィンドウ下限まで（スライディングウィンドウ）
  - オフセットありの現在行からオフセットありの現在行まで（スライディングウィンドウ）

# 例

## 小売会社の売上

以下に示すシンプルなデータセットは、ある小売企業の地域別の売上を示しています。

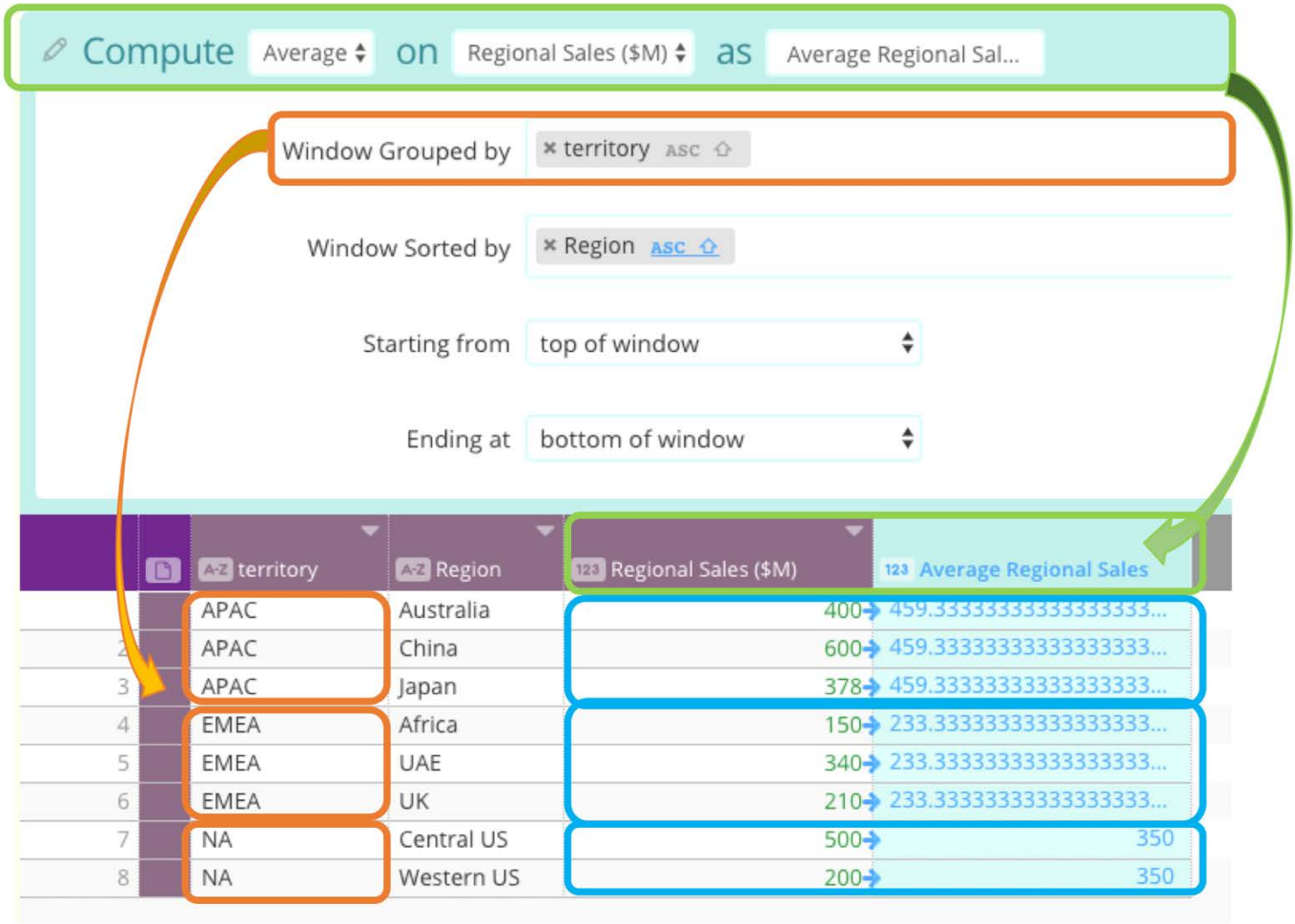
| | territory | Region | Regional Sales ($M) |
|---|---|---|---|
| 1 | NA | Central US | 500 |
| 2 | NA | Western US | 200 |
| 3 | EMEA | Africa | 150 |
| 4 | EMEA | UAE | 340 |
| 5 | EMEA | UK | 210 |
| 6 | APAC | Australia | 400 |
| 7 | APAC | China | 600 |
| 8 | APAC | Japan | 378 |

**ユースケース1：固定ウィンドウ**

この会社は地域間でAverage Regional Sales（地域平均売上）を比較して、業績を比較したいと考えています。

下のイメージの設定では、行が*Territory*（販売区域）を条件にグループ化され、*Region*（地域）を基準にアルファベットの昇順でソートされています。

ウィンドウ上限からウィンドウ下限まで境界が設定されていますので、これは固定ウィンドウです。これは、集計内の*地域販売*列のすべての値を使用して各行の平均が計算されることを意味します。計算に使用される*Regional Sales（地域売上）*列の横に新しいプレビュー列*Average Regional Sales（地域平均売上）*が表示されます。




指定ウィンドウ内のすべての行で、計算された平均値が同じであることがわかります。

したがって、会社は*販売地域*ごとの*地域の平均売上高*を比較して、平均してAPAC販売地域の業績が最高で、EMEA販売地域の業績が最低であることがわかるようになりました。

## 会社の給与

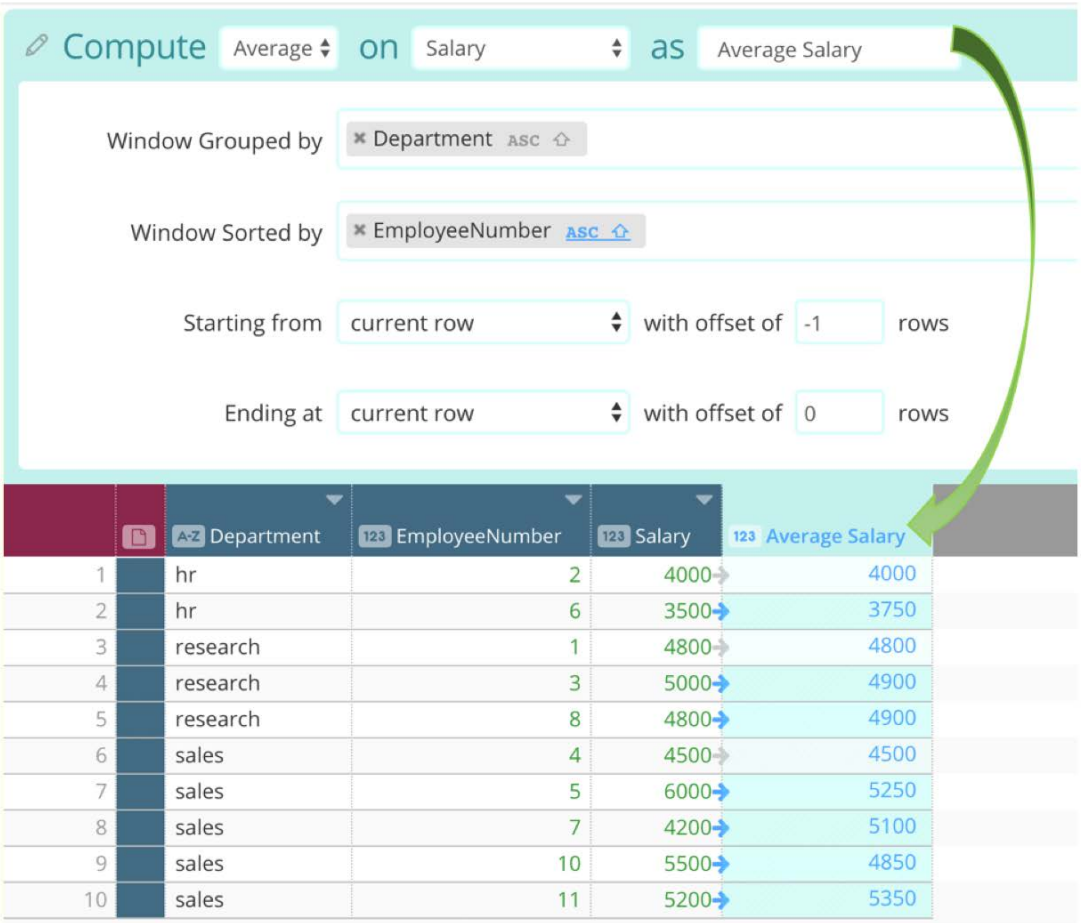
以下に示すシンプルなデータセットは、ある会社の従業員の給与を示しています。

| | Department | EmployeeNumber | Salary |
|---|---|---|---|
| 1 | sales | 11 | 5200 |
| 2 | sales | 7 | 4200 |
| 3 | sales | 4 | 4500 |
| 4 | sales | 5 | 6000 |
| 5 | sales | 10 | 5500 |
| 6 | hr | 2 | 4000 |
| 7 | hr | 6 | 3500 |
| 8 | research | 1 | 4800 |
| 9 | research | 3 | 5000 |
| 10 | research | 8 | 4800 |

**ユースケース2：スライディングウィンドウ**

この会社では営業部門の給与を分析したいと考えており、営業部門で以前に雇用した2人の給与の平均額との比較を行います。

これを行うには、従業員を彼らの*部門*に基づいてグループ化し、*従業員番号*をソート順として使用して、雇用された順序で積み重ねます。営業部門全体の平均を求めるのではなく、各営業員とその直前に雇用した営業員の給与の平均を求めるスライディング（またはローリング）ウィンドウを作成します。ウィンドウの開始は対象従業員の1つ前の行（-1のオフセット）、ウィンドウの終了は対象従業員の行（0のオフセット）に設定します。




営業部門を見ると、ウィンドウの集計値が計算ごとに変化していることがわかります。各平均値は、前の行と現在の行に基づいています。

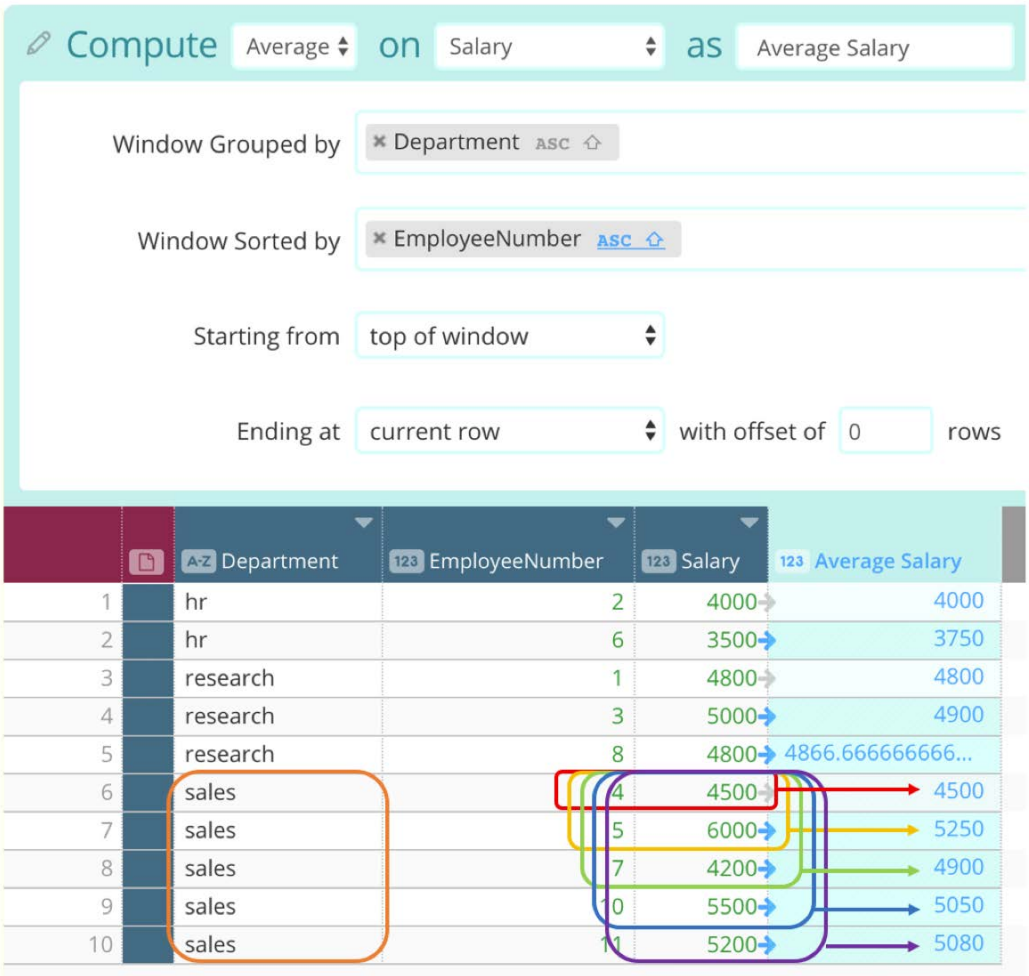

この情報からわかることは、以前に雇用した営業員の平均額より給与が低いのは、従業員番号7の営業員のみであるということです。

#### ユースケース3：スライディングウィンドウ

この会社は、ある部門の給与平均が時間とともにどのように変化するかを分析したいと考えています。

それには、ウィンドウ上限から始まり、その後に続く各行で拡大するスライディングウィンドウを作成します。これは、Excelで移動平均関数を作成するのと似ていますが、絶対セルを設定するために$を含む式を記述する必要はありません。

前の例のように、行を*部門*ごとにグループ化し、*従業員番号*を昇順に並べ替えます。ただし、前の例とは異なり、ウィンドウの下部が計算中の行（つまり現在の行）に関連して変化する間、上部の境界は固定された開始点として残ります。一番上の行から開始し、0のオフセットで現在行で終了するようにウィンドウを設定します。営業部門に関して作成されたウィンドウを以下に示します：

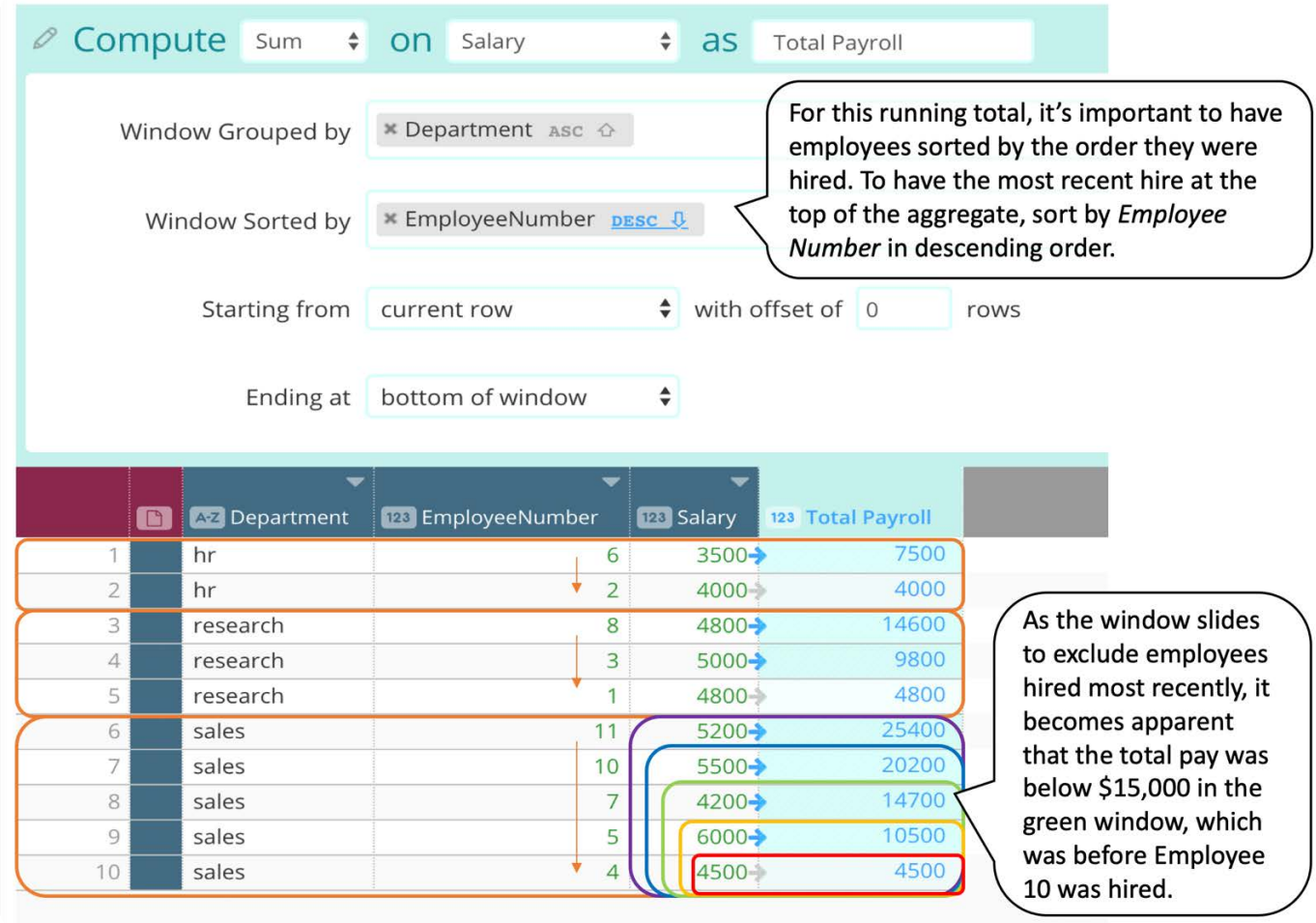



このスライディングウィンドウ関数の出力列を使用すると、時間の経過とともに、営業部門の給与平均が概して増加していることがわかります。また、この計算から興味深い傾向が見られます。これは、時間の経過とともに平均間の分散が小さくなり、給与の一貫性が高まっていることを示しています。

#### ユースケース4：スライディングウィンドウ

この会社は、どの部門の給与総額が15,000ドルを超えているかを分析するために、各部門の給与の累計総額を計算したいと考えています。

*Department*（*部門*）でグループ化し、*Employee Number*（*従業員番号*）で降順にソートすると、最近雇用した従業員が集計行の一番上に表示されます。計算対象の現在行（0でオフセット）で始まり、ウィンドウ下限で終わるスライディングウィンドウを設定します。これは、Excelで累計を作成するのと似ていますが、絶対セルを設定するために$を含む式を記述する必要はありません。

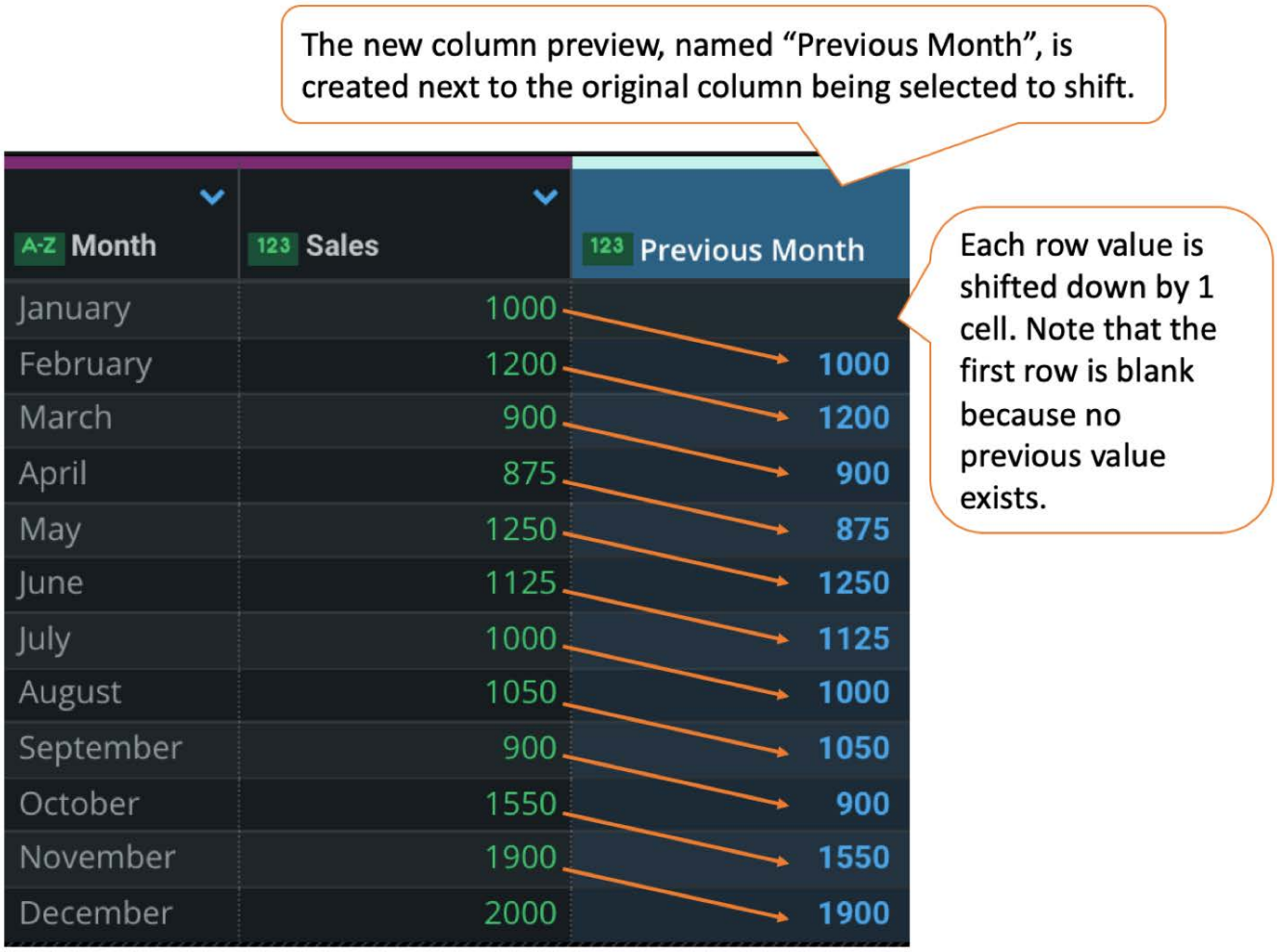



上のイメージに示すように、人事部門と研究部門の累計総額はどちらも分析基準額の15,000ドルを下回っています。しかし、営業部門の現在の給与総額は25,400ドルです。さらに調べると、給与総額がしきい値を超えた時点を確認できます。緑のウィンドウは番号4、5、および7の従業員の給与を含み、しきい値15,000ドルを下回っています。ただし、番号10の従業員が雇用されたときに、青いウィンドウが示すように、しきい値の超過が生じました。

# シフトツールの操作

**シフト**ツールは、参照されている元の列の隣に新しい列を作成し、設定されたセル数だけ行の値を上下にシフトします。このアクションは、SQLのLAGまたはLEADと似ています。**シフト**ツールを使用するとき、シフトする列を指定して新しい名前を付けます。次に、オフセットする行の方向と数を定義します。グループおよびソートのフィールドを使用して、必要に応じて行の値を並べ替えることもできます。

たとえば、ある企業が毎月の売上を前月と比較する場合は、連続する月の売上を横に並べて比較できます。Shift Down（シフトダウン）関数を使用して、元の[Sales（売上）]列に基づいて1のオフセットで新しい"Previous Month"（前月）列を作成します。これで、売上値を横に並べて、計算後の列を簡単に作成して、月ごとの売上の差を定量化できます。




# ランクツールの操作

**ランク**ツールには、データセットに列を追加してランク（Rank）、高密度ランク（Dense Rank）、または行番号（Row Number）を示す関数が含まれています。

- **Rank（ランク）**：同じ値を含む行が複数ある場合、それらの行のランクは同じになります。その次の行には連続しないランクが割り当てられます。たとえば、値が同じ2つのセルのランクは両方とも1になり、次点のセルのランクは3になります。Sort by（ソート基準）オプションを使用すると、ランクはソート順に基づいて割り当てられます。Group by（グループ化条件）オプションを使用すると、グループ内でランクが割り当てられ、次の行グループ内の最上位エントリに対するランクは1にリセットされます。
- **Dense Rank（高密度ランク）**：ランクと同様に、同一値を含む複数の行には同じランクが割り当てられます。ただし、次点の一意の値を含む行に対するランクは単純な連番になり、1つ上のランクを割り当てられた行がいくつあったのかは考慮されません。たとえば、値が同じ2つのセルのランクは両方とも1になり、次点の一意セルのランクは2になります。同じソート基準とグループ化条件のルールがランク操作として適用されます。
- **Row Number（行番号）**：この関数は、ウィンドウのソート順に基づいて、1から始まる各IDを各行に割り当てます（下のイメージ1を参照）。グループまたはソートのフィールドを使用して行の順序を変更すると、作成されるパーティションごとに1からランクが始まります（下のイメージ2を参照）。

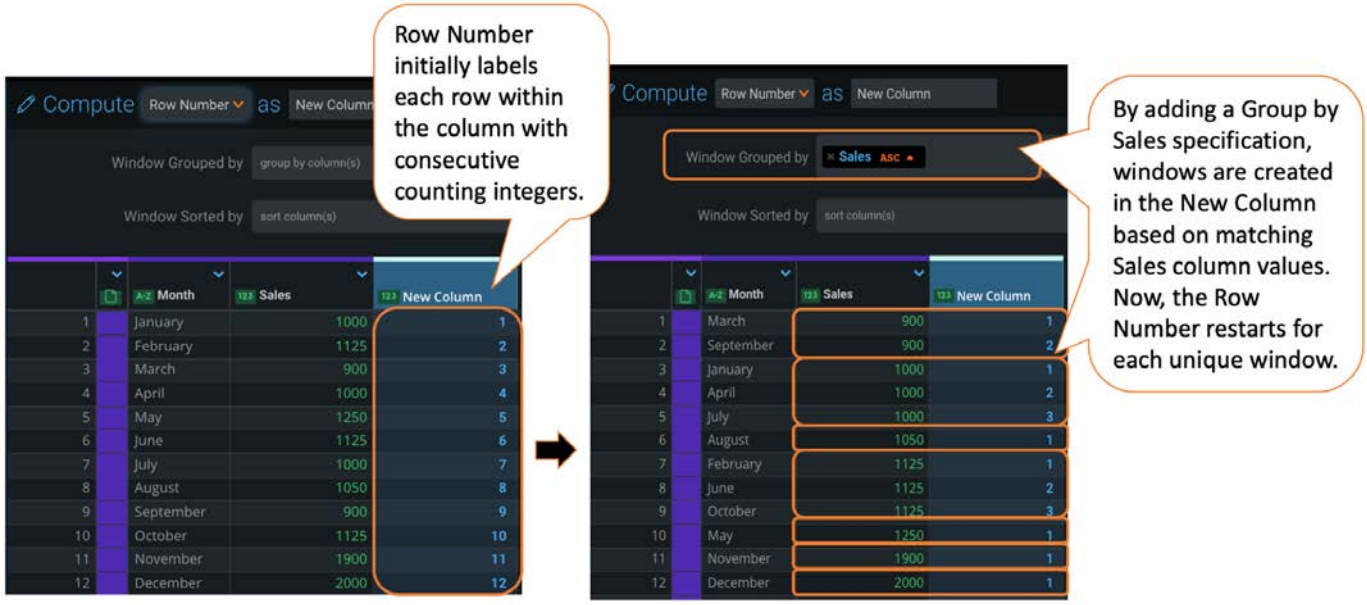



Image 1

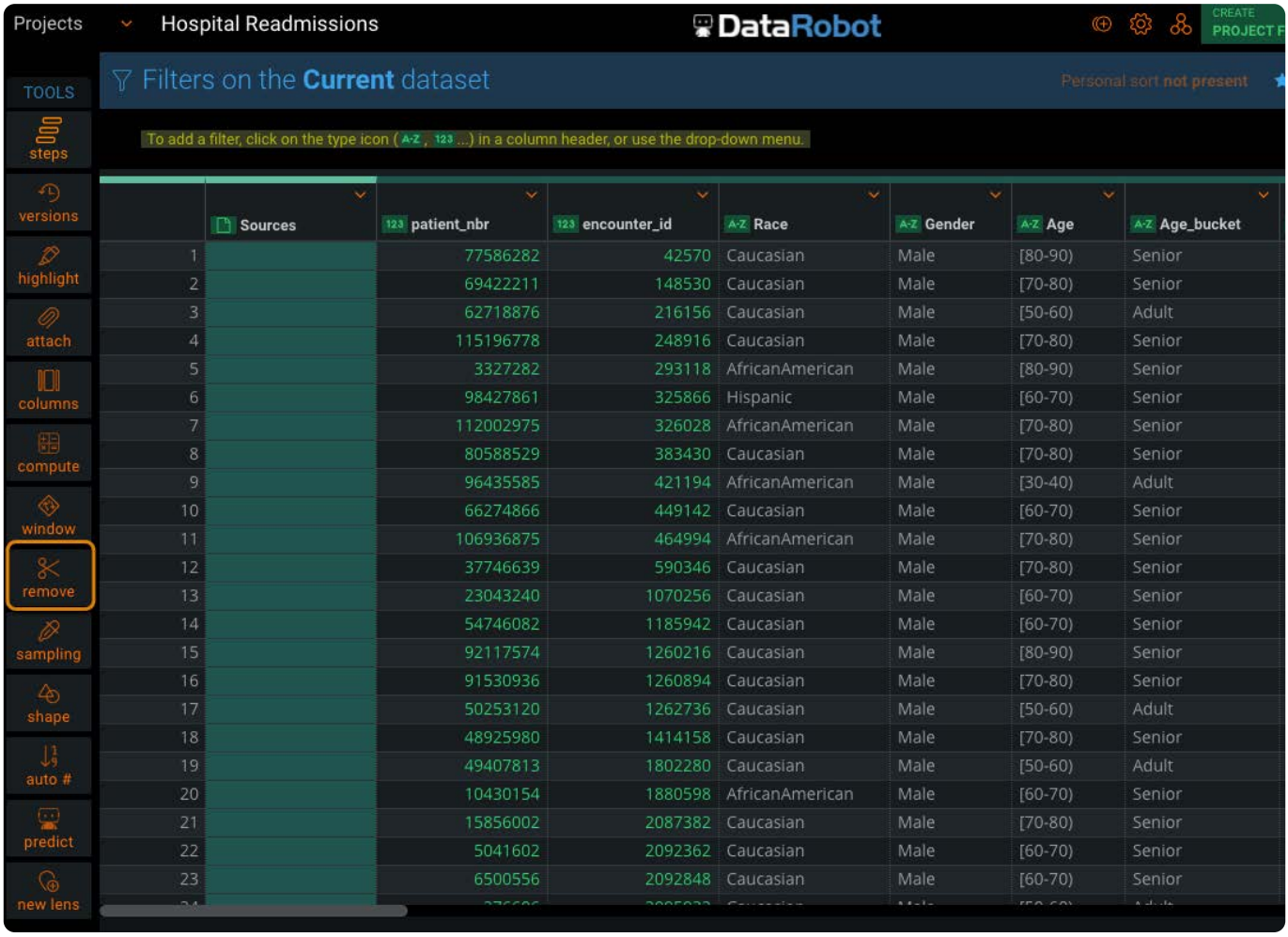
Image 2

# 行の削除

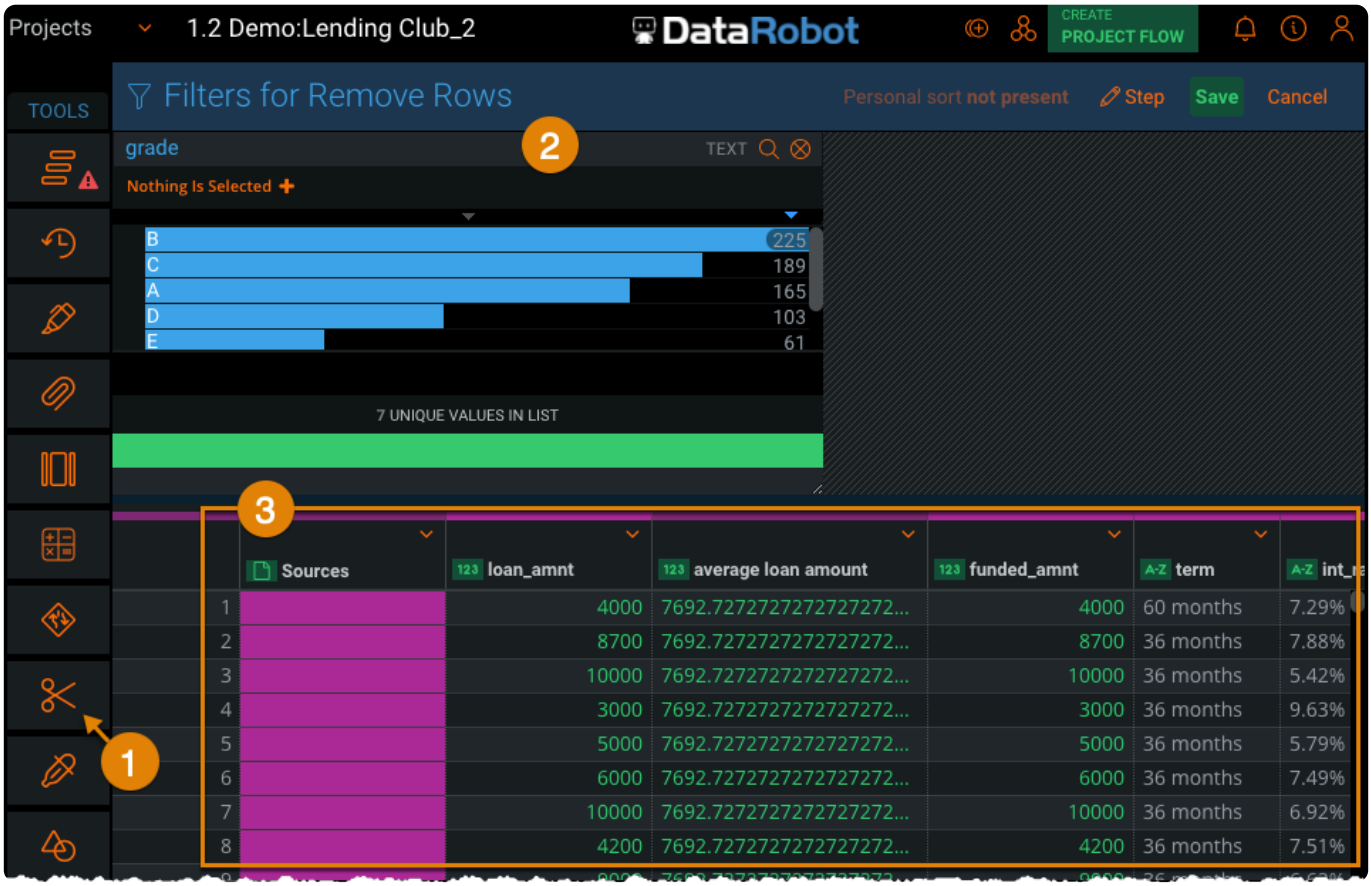
Data Prepでデータを準備すると、データの特定のサブセットを保持したい場合があります。これを実現する最善の方法は、ニーズを満たさないデータの行を削除することです。

## 削除ツールの操作

**削除**ツールにアクセスするには、プロジェクト**ツール**バーの**削除**をクリックします。




次にプロジェクトから行を削除するときに操作する要素の概要を示します：




| | 要素 | 説明 |
|---|---|---|
| 1 | **削除**ツール | **削除**をクリックして、**行の削除**ペインにアクセスし、プロジェクトから行を削除します。 |
| 2 | **行の削除のフィルター**ペイン | データにフィルターを適用して、削除処理のために分離するサブセットのみを表示します。このペインにアクセスするには、右上の**フィルター**をクリックします。 |
| 3 | **データプレビュー**ペイン | プロジェクトにデータを表示し、準備する際にどのように変化するかを表示します。 |

# 行の削除

データから行を削除するには、以下の手順に従います：

1. 右上の**フィルター**リンクをクリックして、Filtergramを追加し、削除する行を隔離します。

   **データプレビュー**には、フィルター基準を満たしているレコードが表示されます。Filtergramの操作方法の詳細については、データ Filtergramを参照してください。

2. **ツール**バーの**削除**をクリックします。

   **現在のデータセットに対するフィルター**ペインが、次に対する**フィルター行の削除**ペインになります。

3. **保存**をクリックします。

   フィルター行は削除されます。使用したフィルターはアクティブなままですが、抽出されたデータは削除されたため、**データのプレビュー**は空白になります。

4. 更新されたデータセットを表示するには、次のいずれかを実行します。

   - Filtergramの**xクリア**をクリックします。
   - [Filtergram] を閉じます。

   Filtergramはクリアされます。**データのプレビュー**に更新されたデータが表示されます。

**ヒント**

後で参照できるように、削除した行のデータは別の AnswerSet に公開しておくと便利です。削除済み行からのデータの取得を参照してください。

**備考**

行を削除した後、データセットの更新または追加した場合、[行の削除] ステップは、次の方法で新しいデータに適用されます：

- ヒストグラムの範囲または個々の値を選択して、行を削除する場合、新しいデータからの行は、正確な条件を満たす場合のみ削除されます。
- 文字列検索または動的なパーセンタイルの範囲で行を選択した場合、ツールは新しいデータに基づいて再計算します。

# 削除された行からのデータのキャプチャ

レンズを追加して、AnswerSetに削除した行を公開します：

1. **ツール**の**ステップ**をクリックします。

   **ステップ**ペインが表示されます。

2. **行の削除**ステップの前のステップをクリックします。

   **データのプレビュー**にフィルターの条件に一致するレコードが表示されます。

3. Filtergram を追加して、削除した行を分離します。

   **データのプレビュー**にフィルターの条件に一致するレコードが表示されます。

4. **ツール**のレンズ > 新しいレンズをクリックします。
5. **保存**をクリックします。

   レンズがプロジェクトに追加されます。

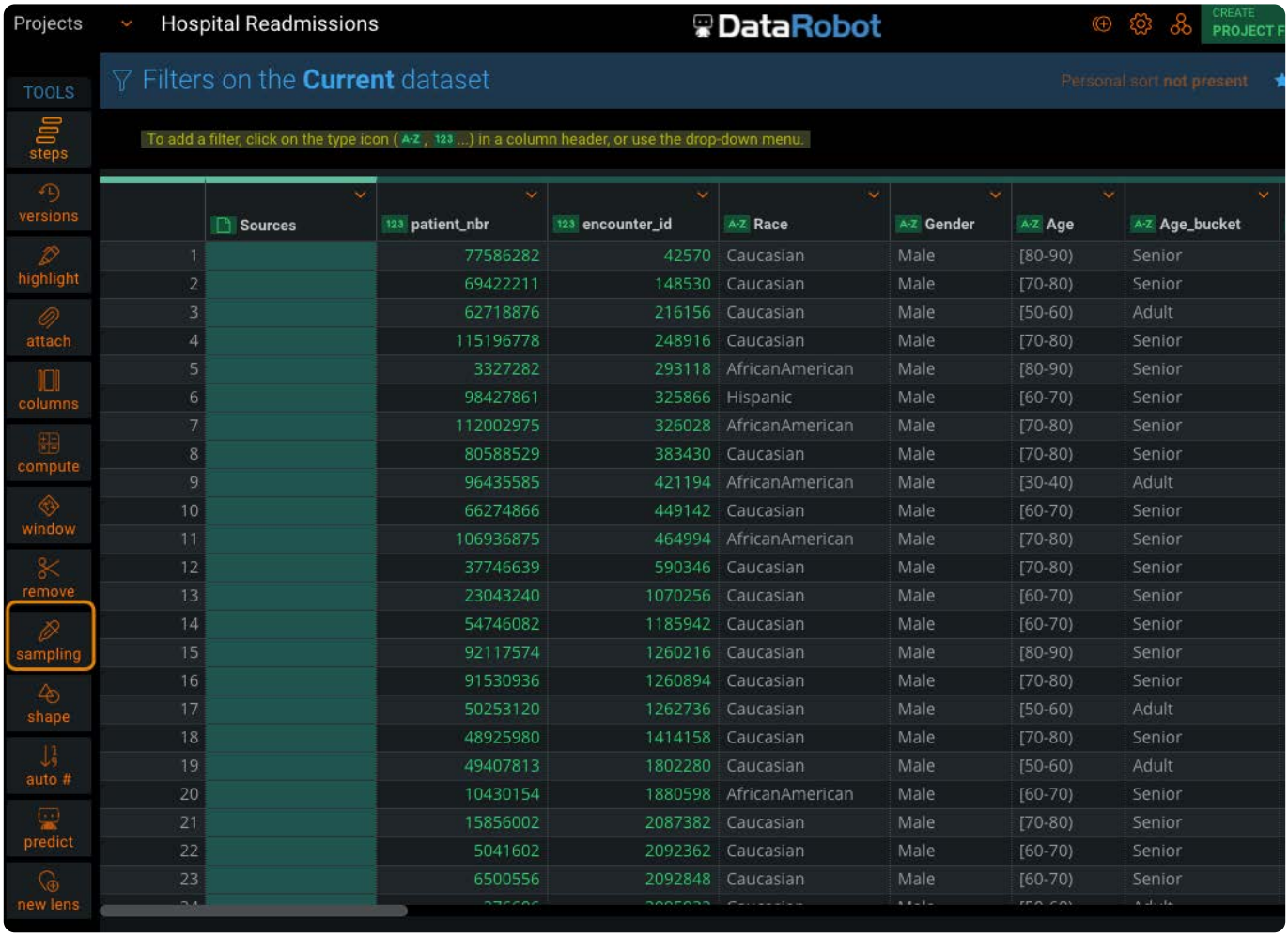
6. レンズを公開するには、レンズから**公開**をクリックします。

# サンプルデータセット

場合によっては、すべてのデータをData Prepプロジェクトに取り込む前に、データセットのサンプリングを行うと便利です。大規模なデータセットでは、これによって初期の探索と検出が簡単になることがあります。**サンプリング**ツールは柔軟性が高く、フィルターを使用してデータ内の特定の行セットに絞り込んだ後で、その結果からサンプリングすることもできます。

## サンプリングツールの操作

**サンプリング**ツールにアクセスするには、プロジェクト**ツール**バーで**サンプリング**バーをクリックします。




すべてのデータをプロジェクトに取込む前に、初期発見のために非常に大きなデータセットをサンプリングすることを推奨します。**サンプリング**ツールは柔軟性が高く、フィルターを使用してデータ内の特定の行セットに絞り込んだ後で、その結果からサンプリングすることもできます。

**備考**

データをサンプリングすることを選択した場合、そのサンプルのパターン、ルックアップの組み合わせ、および集計のみが表示されます。探索が完了したら、**ステップ**ペインでサンプリングをミュートまたは削除することで、サンプリング操作を簡単に削除できます。

# サンプリング方法

サンプリングは、データセットのパーセンテージまたはデータセット内の特定の行数をベースにすることができます。

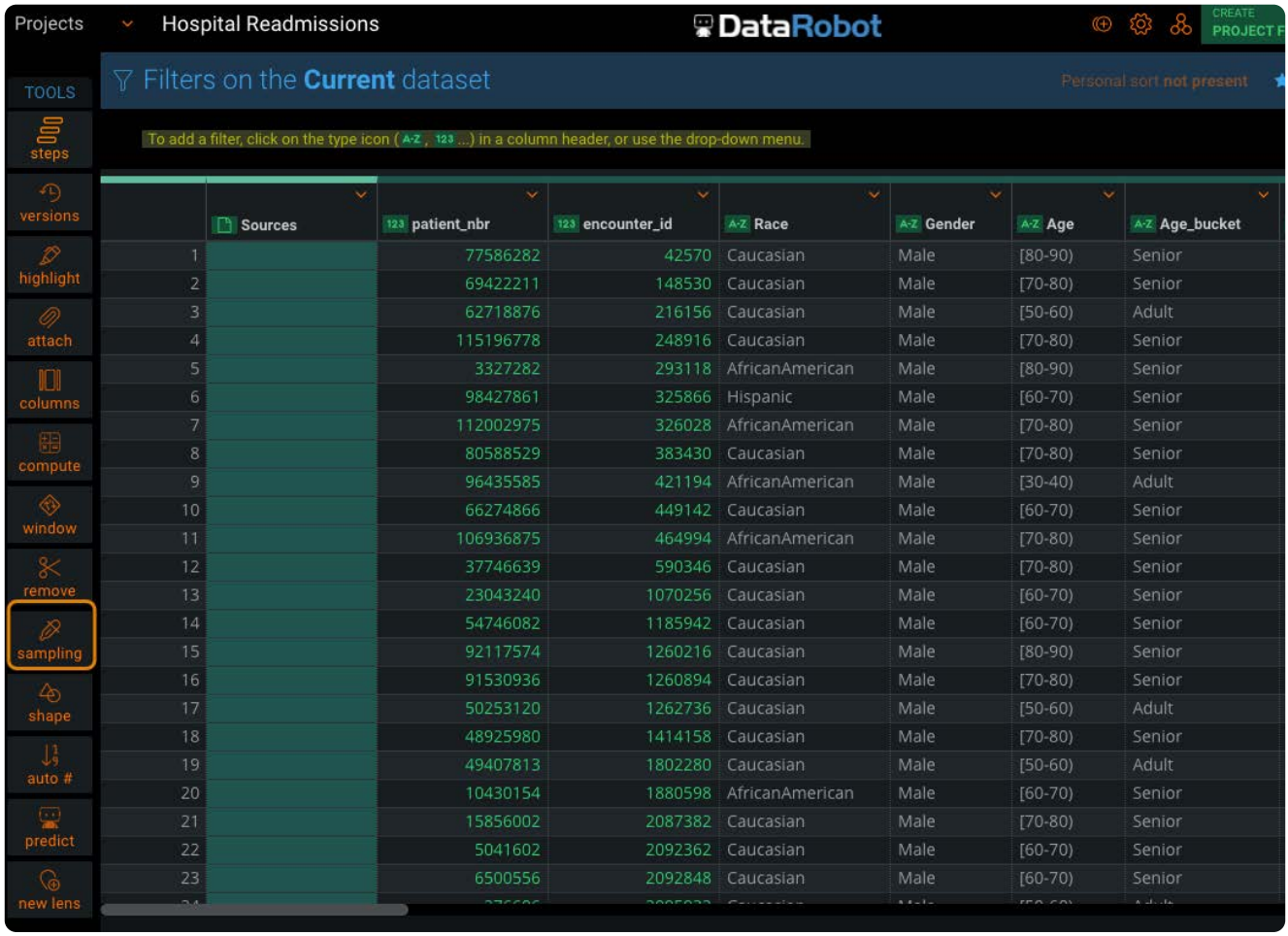
- **パーセンテージベースのサンプリング：**指定したパーセンテージに基づいて、データセット全体でランダムで繰り返し可能なサンプルを実行します。サンプルの生成に使用されるデータセット内の列を指定することもできます。この場合、列のデータのみがサンプルの決定に使用されます。
- **行ベースのサンプリング：**指定した行数に基づいて、データセット全体でランダムで繰り返し可能なサンプルを実行します。指定する行数は、データセット内の行総数で除算されます。データのサブセットサンプルが返されます。プロジェクトのデータ準備ステップとして行ベースのサンプリングを実行している場合、指定した行数は、前のステップのデータセット内の行総数で除算されます。

どちらのタイプのサンプリングでも、「サンプリングシード」番号を保存して、サンプリングしたデータのサブセットを確実に繰り返すことができます。データの別のサブセットサンプルを生成するには、緑の再シードアイコンをクリックすることもできます。最適なサンプルを得るには、データセットが10万行を超える必要があります。

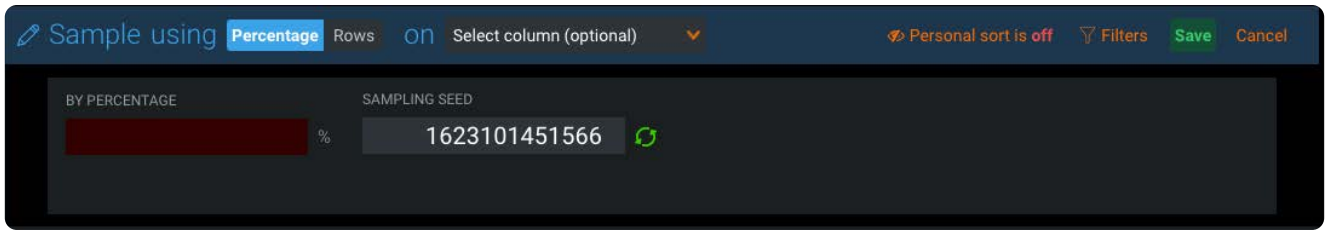
# パーセンテージを使用したサンプル

データセットのパーセンテージをベースにサンプルを作成するには、次の手順に従います：

1. **ツール**バーから、**列**をクリックします。




**使用サンプル**ペインが表示されます。



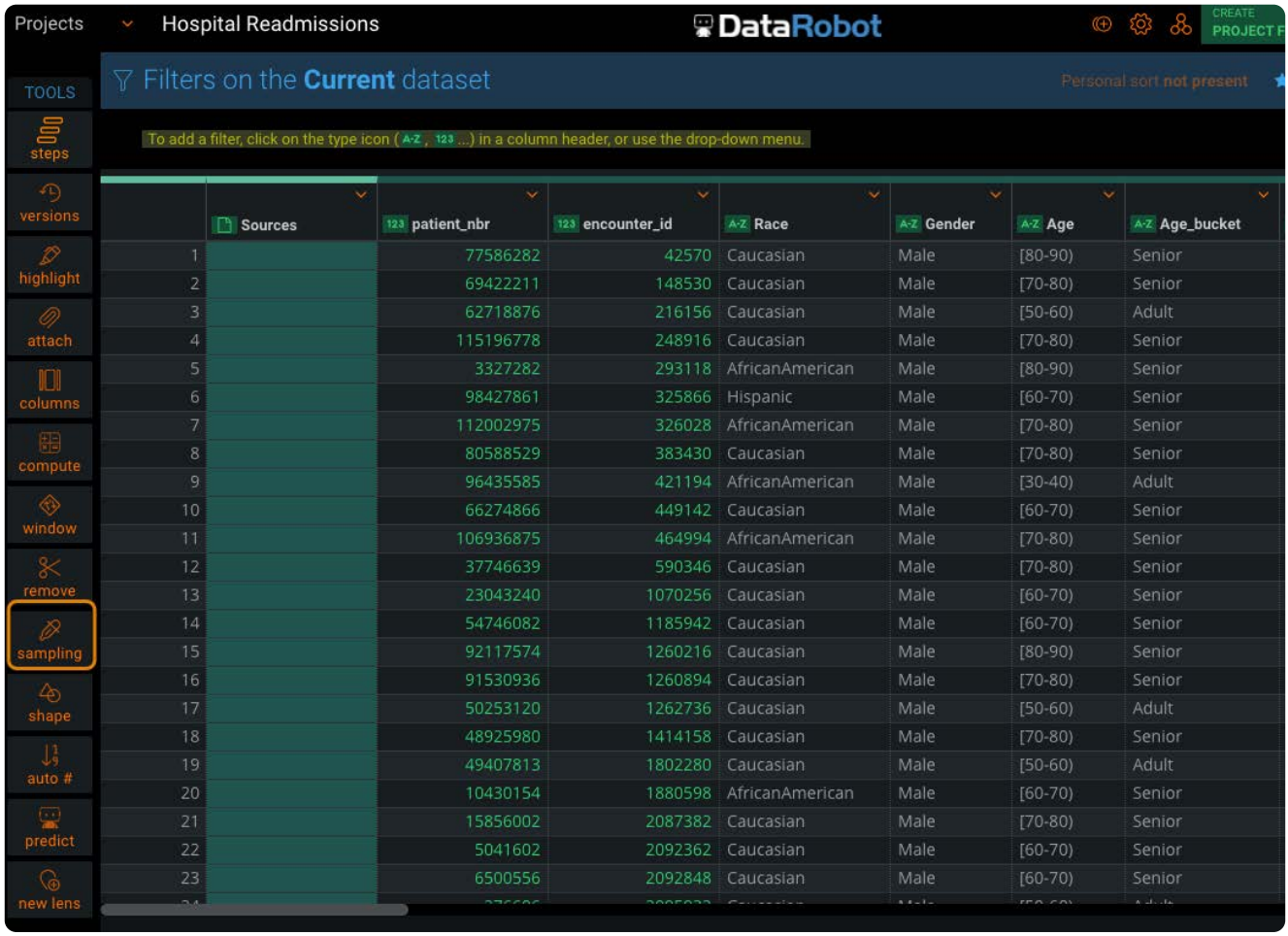

2. まだ選択されていない場合、**パーセンテージ**をクリックします。
3. オプションで列を選択します。

   サンプリングパーセンテージは、選択した列に基づいています
4. **By Percentage**フィールドに、サンプルに含めるデータセットのパーセンテージを入力します。
5. 必要に応じて緑色の**再シード**アイコンをクリックします。
6. **保存**をクリックします。

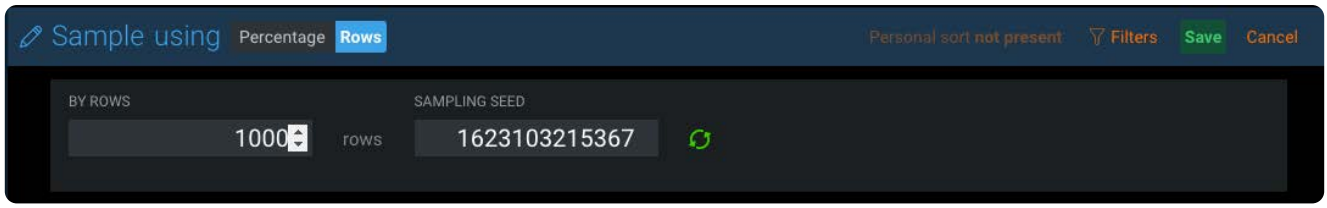
# 行を使用したサンプル

データセットのパーセンテージをベースにサンプルを作成するには、次の手順に従います：

1. **ツール**バーから、**列**をクリックします。




**使用サンプル**ペインが表示されます。




2. まだ選択されていない場合、**パーセンテージ**をクリックします。
3. オプションで列を選択します。

   サンプリングパーセンテージは、選択した列に基づいています
4. **By Percentage**フィールドに、サンプルに含めるデータセットのパーセンテージを入力します。
5. 必要に応じて緑色の**再シード**アイコンをクリックします。
6. **保存**をクリックします。

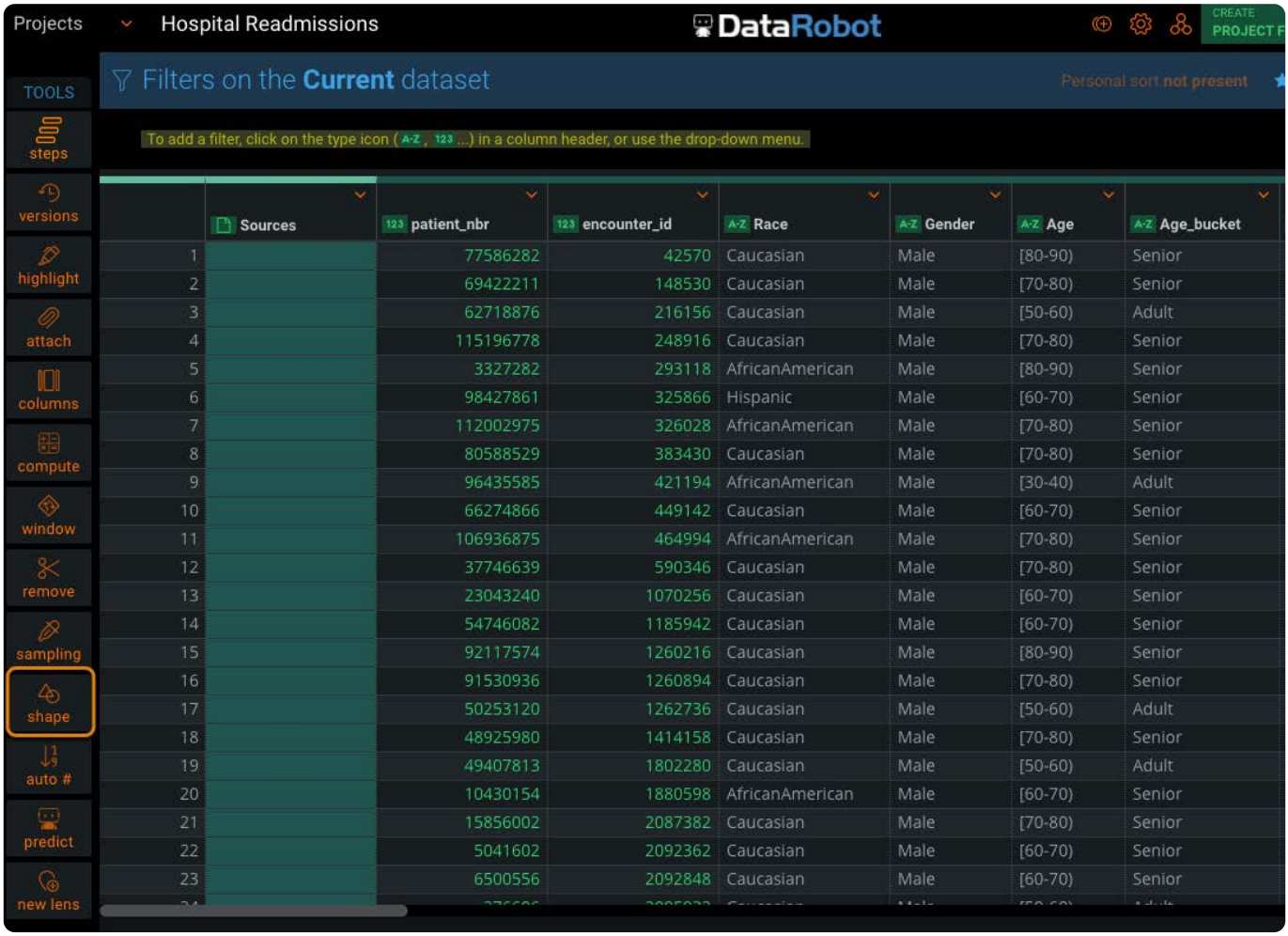
# データの整形

Data Prepでは次のことを可能にする**シェイプ**ツールを用意しています。

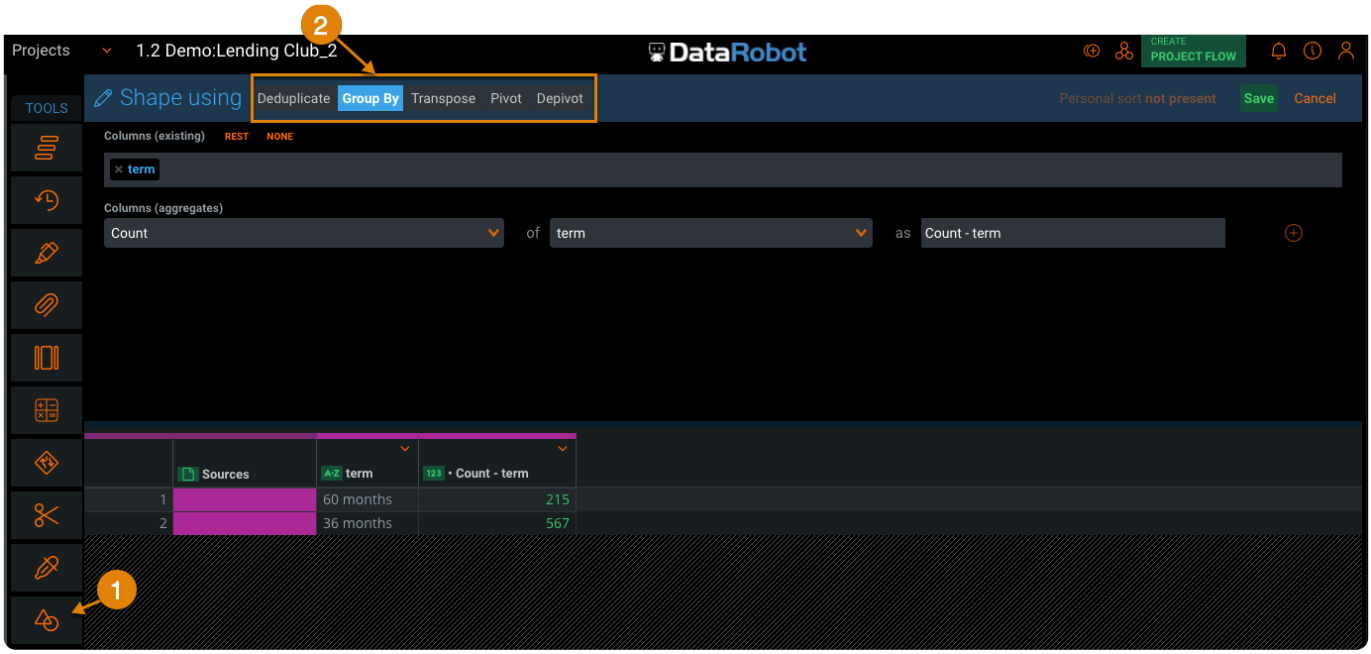
- 重複除去
- グループ化
- 転置
- ピボット
- ピボット解除

## 整形ツールの操作

**シェイプ**ツールにアクセスするには、プロジェクトの**ツール**バーで**シェイプ**をクリックします。



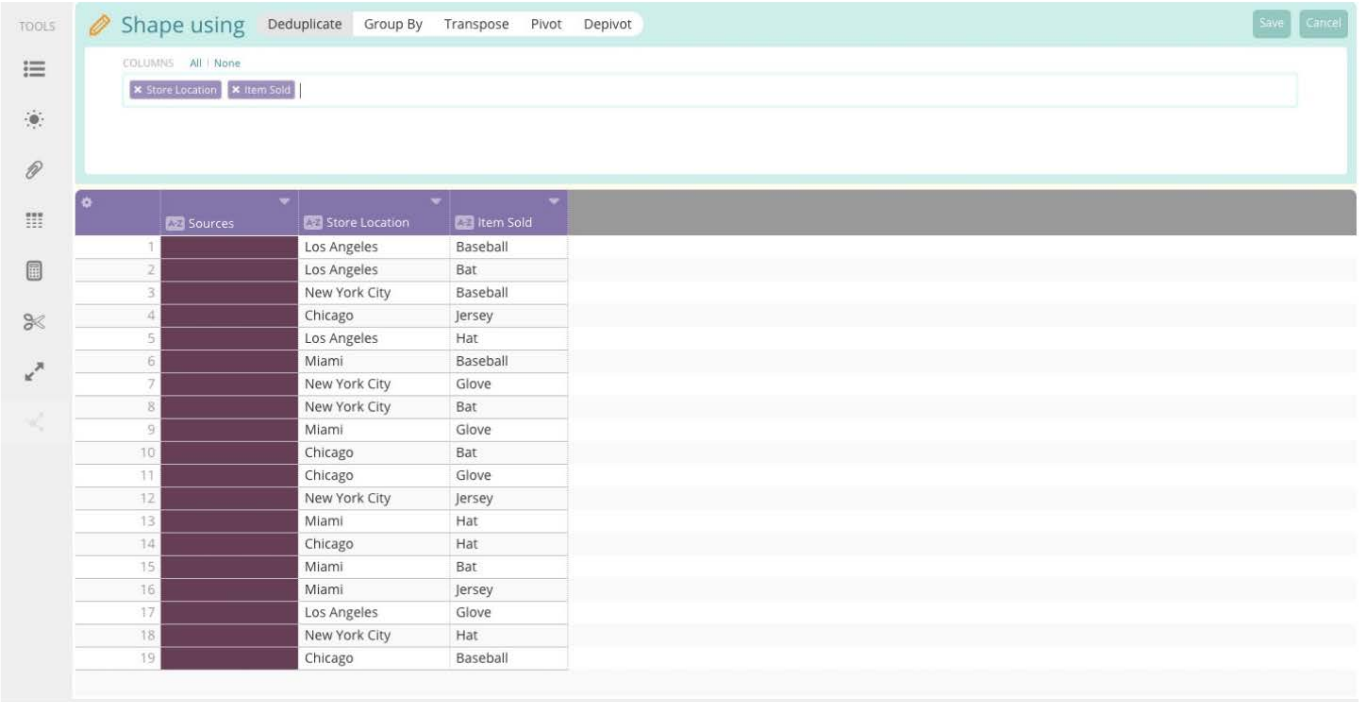

**シェイプ**ウィンドウの要素の概要を示します。




| | 要素 | 説明 |
|---|---|---|
| 1 | **シェイプ**ツール | **シェイプ**をクリックして**シェイプ**ウィンドウにアクセスします。 |
| 2 | **シェイプ**ウィンドウ | **シェイプ**ウィンドウでシェイプツールを選択します：<br>・重複除去<br>・グループ化<br>・転置<br>・ピボット<br>・ピボット解除 |

# 重複除去

**重複除去**機能は、データ内から行単位で互いに正確に一致する値を検索し、これらを単一の行として出力して、重複する値を除去します。

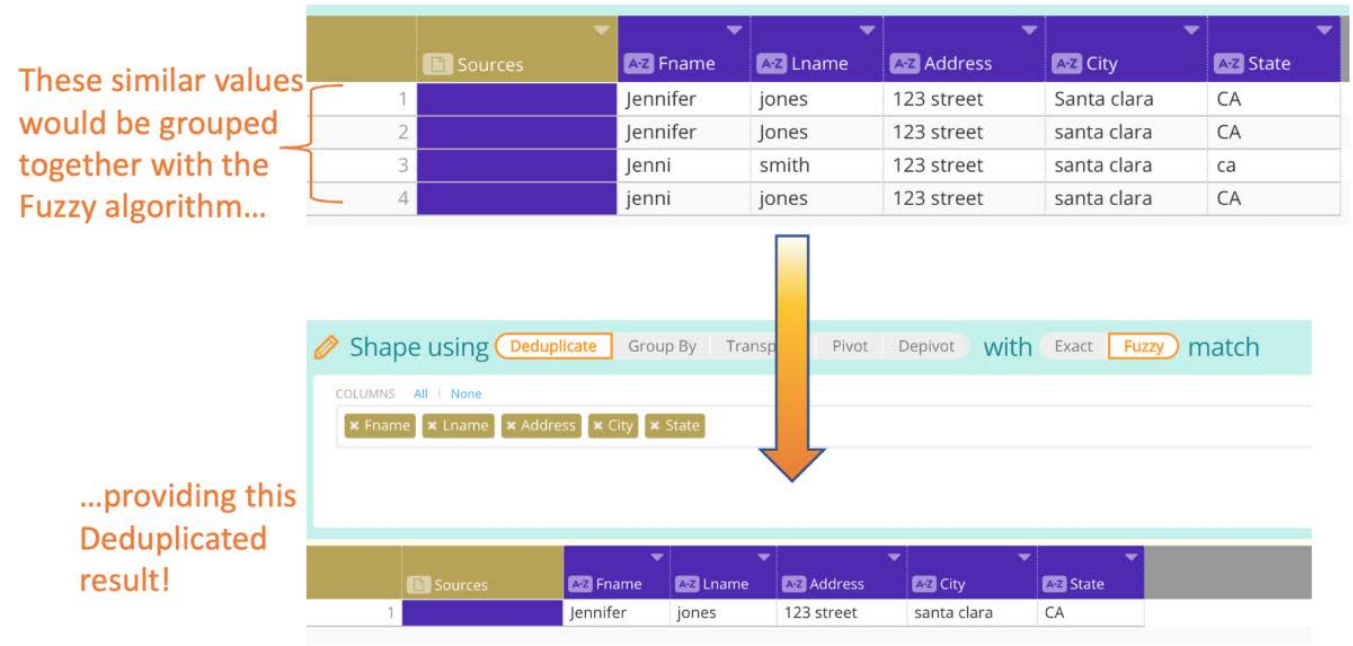



**重複除去**ウィンドウでは、**列**フィールドで列の追加および削除を行うことができます。このフィールドに追加されたすべての列が、重複除去プロセスに含まれます。列を追加すると、**データプレビュー**に表示されます。

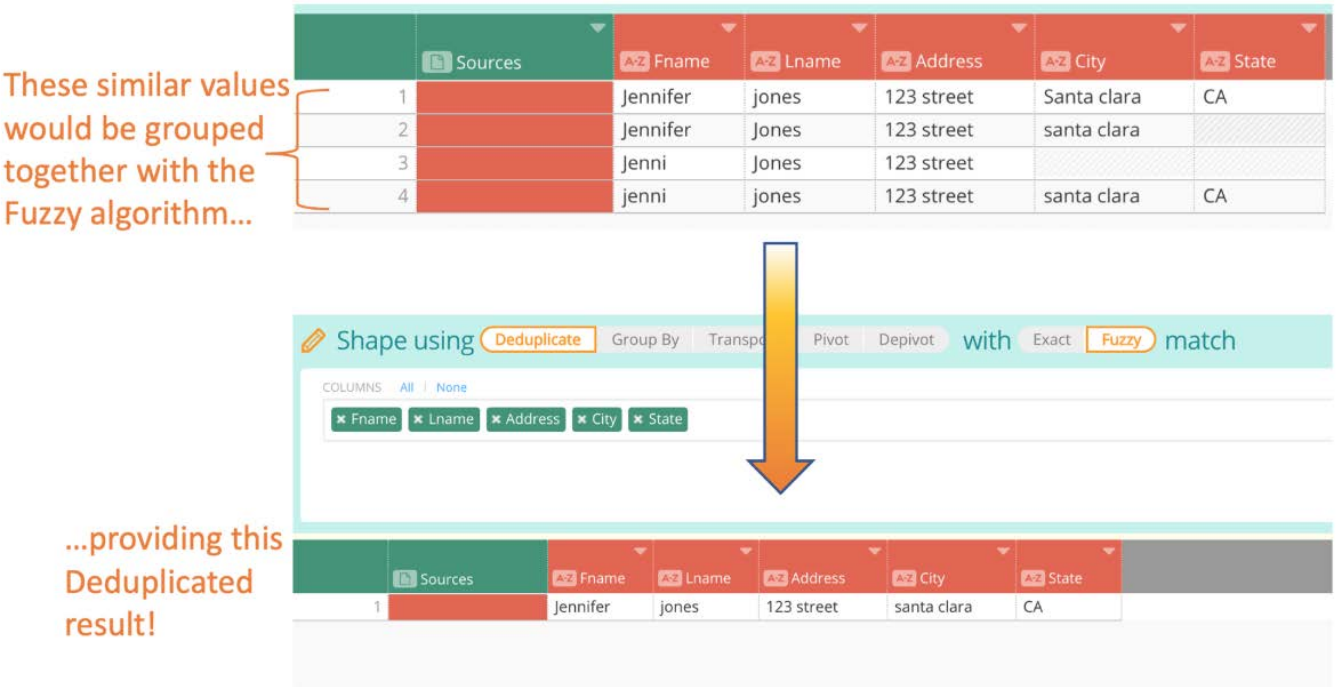
> **備考**
>
> 重複除去プロセスおよび結果出力に含められるのは、**列**フィールドに追加された列だけです。選択しなかった列は重複除去プロセス内で考慮されず、プロセスの完了時にデータから削除されます。重複除去機能の各パラメーターの設定を終えたら、**保存**ボタンをクリックし、プロセスを確定してプロジェクトにコミットします。

**重複除去**機能では、**ファジー**オプションを使用することもできます。ファジー重複除去を有効にすると、**完全一致**の代替手段として使用できます。


ファジー重複除去では、ファジーアルゴリズムを使用して一致行が検出されます。つまり、完全一致ではなく類似した値同士がグループ化され、重複除去されます。以下に例を示します。



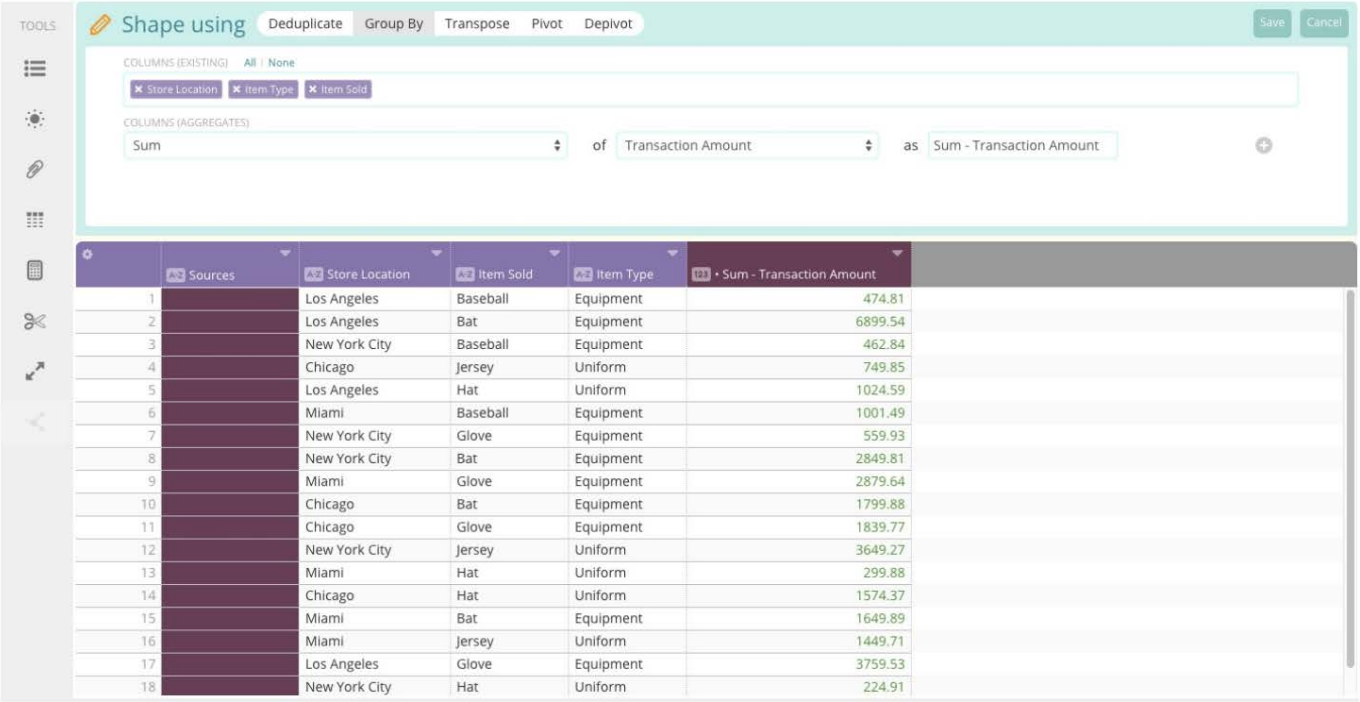

ファジー重複除去では、空白値が含まれている場合でも、類似した項目がグループ化されます。




# グループ化

**グループ化**機能を使用すると、データセット内の任意の既存列に対し、さまざまな種類の集計関数を実行できます。**グループ化**を選択すると、**データプレビュー**の上にウィンドウが表示されます。このウィンドウでは、グループ化プロセスに含める列、集計関数の実行対象とする列、使用する集計関数、新規作成する集計列の名前を指定できます。

ユーザーが選択したデータは**データプレビュー**で青色に強調表示されるので、データがどのように影響されるかを確認できます。重要な点は、**グループ化**機能を実行すると、**列（集計）**フィールドに含まれた列のみがデータ内に残ることです。ここに含まれている列は、グループ化を行う際、重複行を識別するために使用されます。




使用可能な集計関数のリストについては、グループ化集計関数を参照してください。

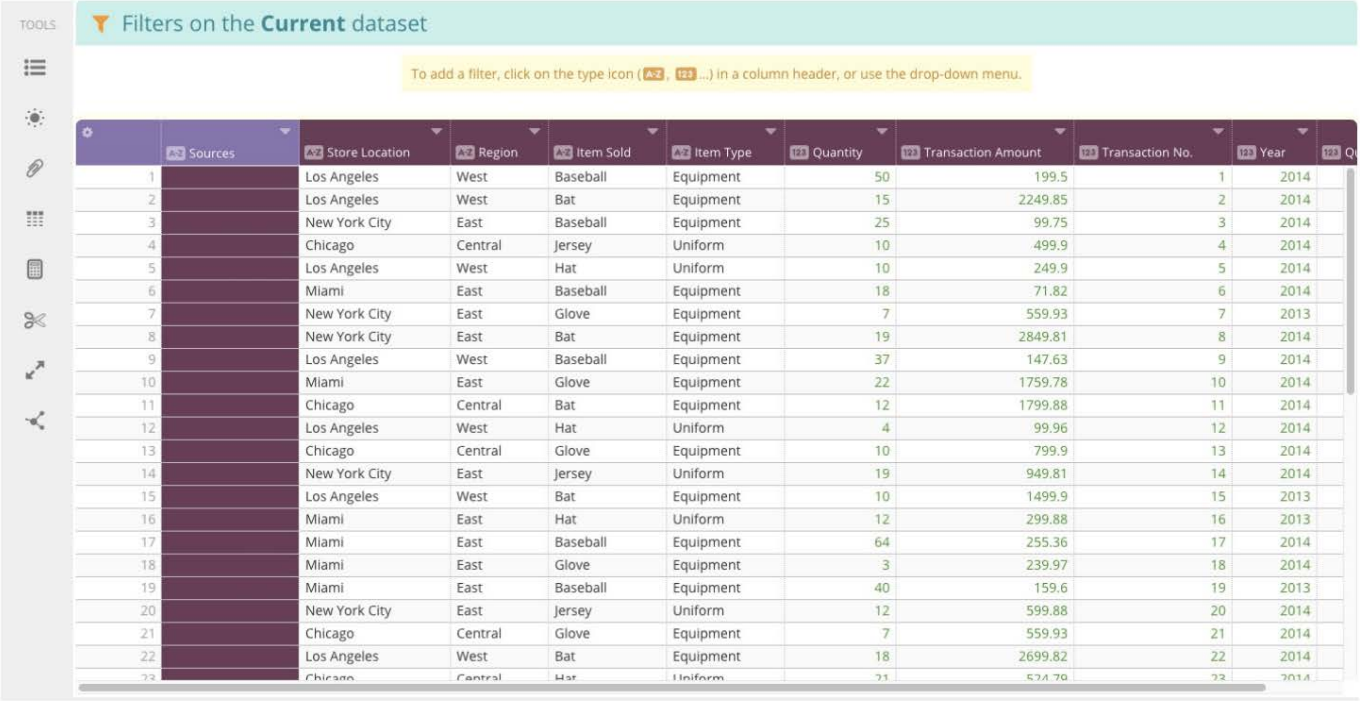
これらの操作は、一致する行をデータセットから検索し、これらを1つの行として結合するため、「集計」と呼ばれます。一致する行とは、列単位で検査したとき（参照列を除く）同じ値を持つ行として定義されます。このカラム単位の検査では、参照カラムは除外されます。 参照カラムの値を集計関数に提供することで、参照カラムの値を単一行の結果として生成するためです。

# 転置

**転置**機能を使用すると、行と列を入れ替えることができます。いわば、データを90度回転させることが可能です。

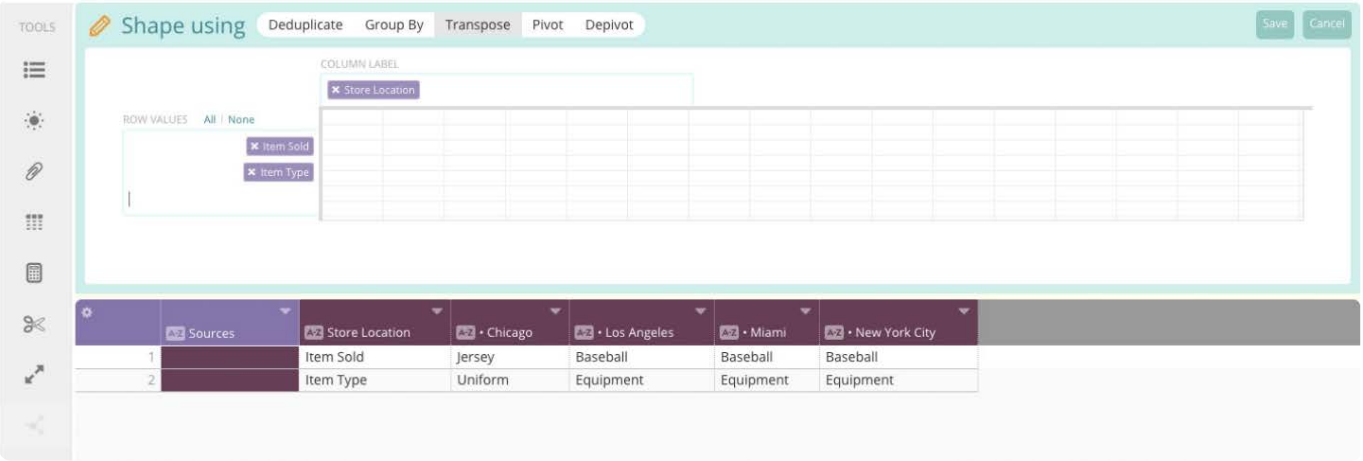
転置機能では、任意の1列を選択して、この列値によって新しい列見出しを作成します。選択した列の値が新規列見出しとなり、他のすべての列見出しは行見出しとして移動されます（転置プロセス中にユーザーが削除しない限り）。この転置プロセスにおいて、特定の見出しセットと一致する値が複数存在する場合は、Data Prepにより、元のデータセット内の最後の有効値が表示されます。

ここに、売上を示す単純なデータセットがあります。




| | Sources | Store Location | Region | Item Sold | Item Type | Quantity | Transaction Amount | Transaction No. | Year |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | Los Angeles | West | Baseball | Equipment | 50 | 199.5 | 1 | 2014 |
| 2 | | Los Angeles | West | Bat | Equipment | 15 | 2249.85 | 2 | 2014 |
| 3 | | New York City | East | Baseball | Equipment | 25 | 99.75 | 3 | 2014 |
| 4 | | Chicago | Central | Jersey | Uniform | 10 | 499.9 | 4 | 2014 |
| 5 | | Los Angeles | West | Hat | Uniform | 10 | 249.9 | 5 | 2014 |
| 6 | | Miami | East | Baseball | Equipment | 18 | 71.82 | 6 | 2014 |
| 7 | | New York City | East | Glove | Equipment | 7 | 559.93 | 7 | 2013 |
| 8 | | New York City | East | Bat | Equipment | 19 | 2849.81 | 8 | 2014 |
| 9 | | Los Angeles | West | Baseball | Equipment | 37 | 147.63 | 9 | 2014 |
| 10 | | Miami | East | Glove | Equipment | 22 | 1759.78 | 10 | 2014 |
| 11 | | Chicago | Central | Bat | Equipment | 12 | 1799.88 | 11 | 2014 |
| 12 | | Los Angeles | West | Hat | Uniform | 4 | 99.96 | 12 | 2014 |
| 13 | | Chicago | Central | Glove | Equipment | 10 | 799.9 | 13 | 2014 |
| 14 | | New York City | East | Jersey | Uniform | 19 | 949.81 | 14 | 2014 |
| 15 | | Los Angeles | West | Bat | Equipment | 10 | 1499.9 | 15 | 2013 |
| 16 | | Miami | East | Hat | Uniform | 12 | 299.88 | 16 | 2013 |
| 17 | | Miami | East | Baseball | Equipment | 64 | 255.36 | 17 | 2014 |
| 18 | | Miami | East | Glove | Equipment | 3 | 239.97 | 18 | 2014 |
| 19 | | Miami | East | Baseball | Equipment | 40 | 159.6 | 19 | 2013 |
| 20 | | New York City | East | Jersey | Uniform | 12 | 599.88 | 20 | 2014 |
| 21 | | Chicago | Central | Glove | Equipment | 7 | 559.93 | 21 | 2014 |
| 22 | | Los Angeles | West | Bat | Equipment | 18 | 2699.82 | 22 | 2014 |

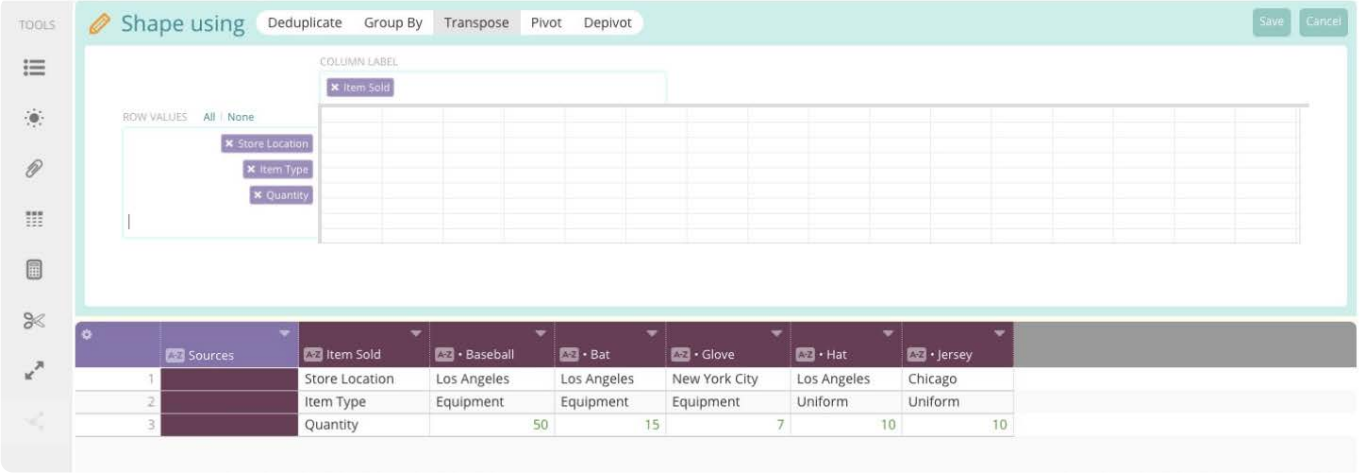
**シェイプ**ツールをクリックしたら、**シェイプ**ウィンドウの上部にある**転置**を選択します。




| | Sources | Store Location | • Chicago | • Los Angeles | • Miami | • New York City |
|---|---|---|---|---|---|---|
| 1 | | Item Sold | Jersey | Baseball | Baseball | Baseball |
| 2 | | Item Type | Uniform | Equipment | Equipment | Equipment |

列見出しとして、1つの列を選択できます（ここで選択した列の値が新しい列見出しになります）。また、新たな転置データに行として含める列を、必要な数だけ選択します。**シェイプ**ウィンドウの下部にあるグリッドには、転置プロセスに対して選択したオプションに基づき、データがどのように出力されるかを示すプレビューが表示されます。

転置の対象として選択した列によっては、結果的に全データが完全には含まれないことがあるので、注意が必要です。ここに示した例で、「Item Sold」の値が「Baseball」となるインスタンスが複数存在することに注目してください。「Item Sold」を新しい列として転置を行うと、Data Prepでは、データセット内の最後の値だけが表示されます（下図参照）。

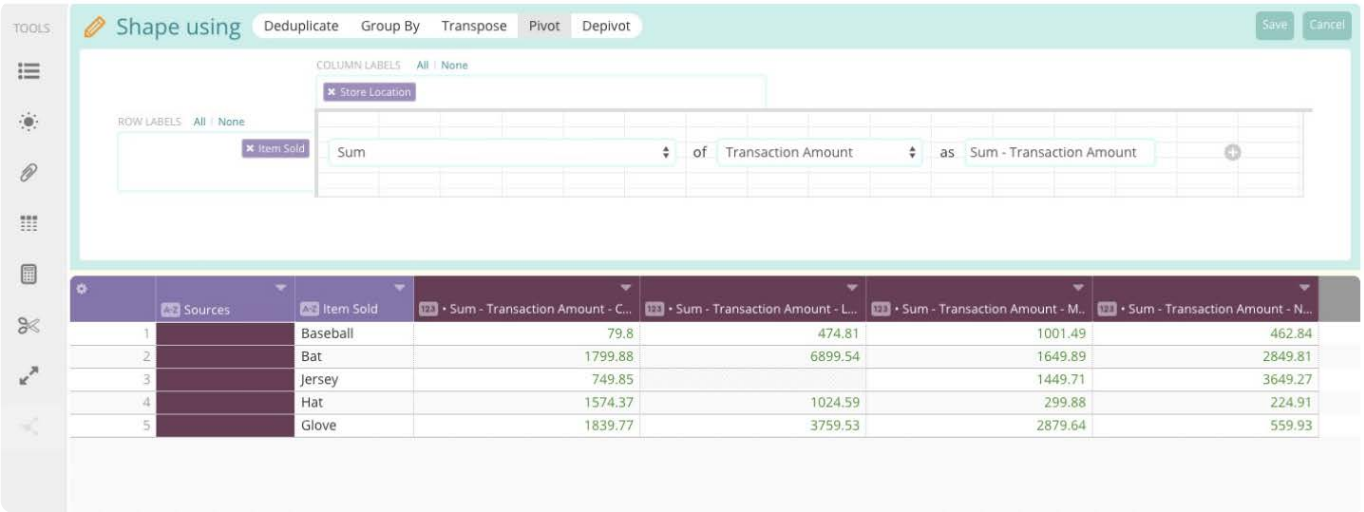



転置プロセスの各オプションを適切に選択したら、画面右上の**保存**ボタンをクリックします。 これで、プロセスが確定し、プロジェクトにコミットされます。

# ピボット

**ピボット**は**転置**と同じように、列見出しを行見出しに置換します。ただし、**ピボット**機能では、選択した列に集計関数を実行して、その結果をピボットテーブル内のデータ本体として表示できます。さらに、**転置**機能とは異なり、**ピボット**では、複数の見出しを列見出しとして選択できます。

**シェイプ**ツールを開き、**ピボット**を選択すると、ピボットオプションが表示されます。列見出しとして使用する列、作成するピボットテーブルに行として含める列をそれぞれ選択できます。最後に、ピボットテーブルのデータ本体として作成する列を選択し、この列に対して実行する集計関数の種類を指定できます。**+**ボタンをクリックすることで、複数の集計関数をピボットテーブルに適用できます。



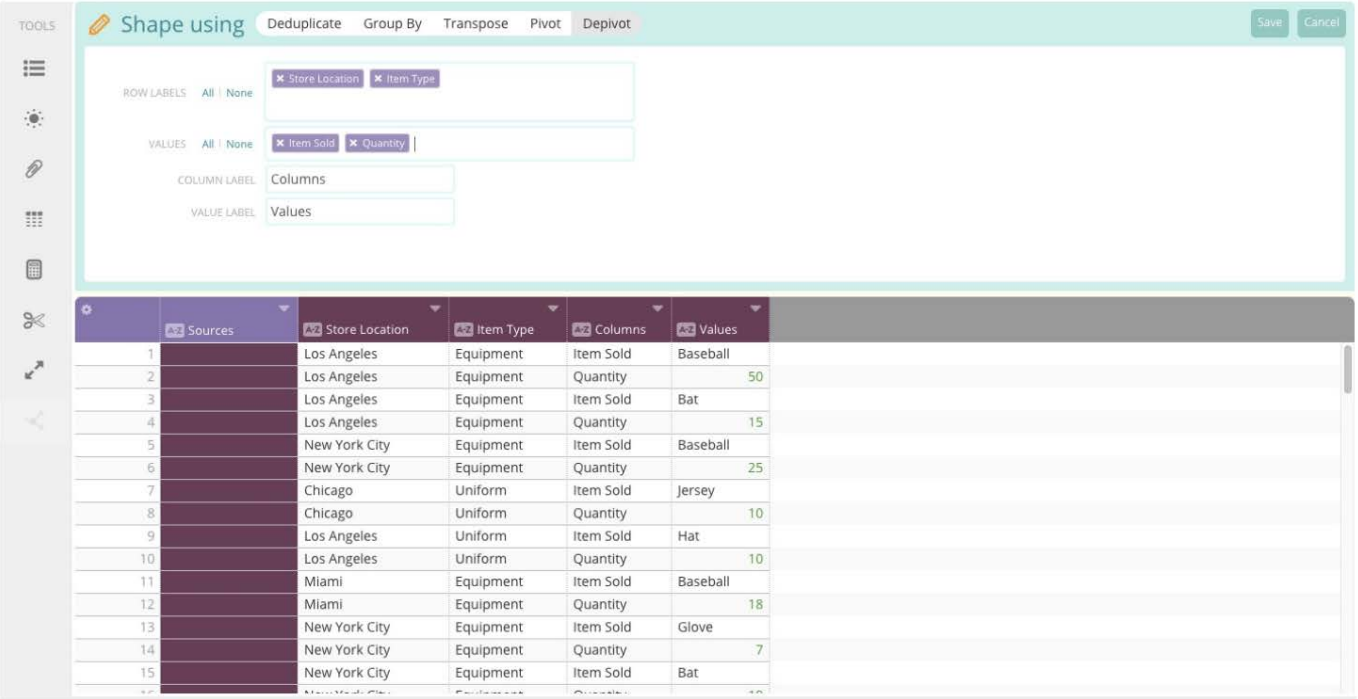

選択した集計関数が、選択した列に対して実行され、選択した列見出しおよび行見出しの交差する位置に基づく合計が表示されます。使用可能な集計関数のリストについては、グループ化集計関数を参照してください。

**シェイプ**ウィンドウ下部に表示されるデータは、選択したオプションに基づいて生成されるピボットテーブルのプレビューであり、よく確認する必要があります。

転置プロセスの各オプションを適切に選択したら、画面右上の**保存**ボタンをクリックします。 これで、プロセスが確定し、プロジェクトにコミットされます。

# ピボット解除

**ピボット解除**機能は、各列のデータを2つの列として積み上げる（スタックする）、非常に強力な機能です。2つの列のうち、1番目の列には、元のデータを保持していた各列見出しが表示され、2番目の列には、この列の値が表示されます。このような機能を使用した経験のないユーザーにとっては、**ピボット解除**の実行によって実際に何が行われるのか、正確に理解することは難しいかもしれません。




ピボット解除ウィンドウでは、いくつものオプションを設定します。

- **行ラベル**: データ内に固定する列を選択します。これらの列は、ピボット解除プロセスではスタックされません。
- **値**：ピボット解除プロセスにおいて、データ内のどの列を含める（スタックする）かを選択します。
- **列ラベル**：**値**フィールドで選択された列のラベルを含む、新しい列の名前を指定します。
- **値ラベル**：値フィールドで選択された列の**値**を含む、新しい列の名前を指定します。

上の例では、**ピボット解除**機能の実行により、**値**フィールドで選択された列が、2 つの新しい列としてスタックされている状態がプレビュー表示されています。次の図は、具体的な動作をわかりやすく示したものです。

| LOCATION | COLUMN | VALUE |
|---|---|---|
| Los Angeles | Q1 SALES | $1,250 |
| Los Angeles | Q2 SALES | $2,700 |
| Los Angeles | Q3 SALES | $2,465 |
| Palo Alto | Q1 SALES | $975 |
| Palo Alto | Q2 SALES | $1,680 |
| Palo Alto | Q3 SALES | $1,800 |
| Seattle | Q1 SALES | $1,100 |
| Seattle | Q2 SALES | $2,355 |
| Seattle | Q3 SALES | $2,890 |

ピボット解除プロセスの実行により、それぞれの値がどこから抽出され、どこに移動されているかを理解しやすくするため、各列は色分けして表示されています。

この例では、**LOCATION**列が**行ラベル**の唯一の値として選択されています。ピボット解除プロセスにより、列値がスタックされるため、それぞれの場所に対して重複行が生成されています。このしくみにより、**値**フィールドで選択された列の個々の値に対し、それぞれの行が生成されます。**ピボット解除**機能を実行することで、すべての売上高が単一列に格納されています。この形式に対してグループ化機能を実行すると、データに対し、普段は気付かない洞察が得られる可能性もあり、たいへん便利です。

# グループ化集計関数

Data Prep集計関数では、行の集合を組み合わせることや、参照された列で特定の関数を計算することができます。

これらの操作は、一致する行をデータセットから検索し、これらを 結合して1つの行にするため、「集計」と呼ばれます。一致する行は、列ごとに同じ値を共有するものとして定義されます。この列単位の検査では、参照列は除外されます。 参照列の値を集計関数に提供することで、参照列の値を単一行の結果として生成するためです。

以下のセクションでは、**グループ化**を使用して操作を形成する場合に使用できる集計関数について説明します:

| 動作 | 目的 |
|---|---|
| array（配列） | 重複する行を凝縮して1つのデータ行にし、コンマで区切られた1つの文字列に参照列データを凝縮します。 |
| average | 重複する行を凝縮して1つのデータ行とし、参照列の値の平均を表示します。 |
| count | 重複する行を凝縮して1つのデータ行とし、参照列中の重複行の数を表示します。 |
| count（数値のみ） | 重複する行を凝縮して1つのデータ行とし、重複する行数を表示しますが、数値のみをカウントします。 |
| count distinct | 重複する行を凝縮して1つのデータ行とし、参照列に一意の値の数を表示します。 |
| first | 重複する行を凝縮して1つのデータ行とし、重複する行に現れた最初の値を表示します。 |
| last | 重複する行を凝縮して1つのデータ行とし、重複する行に表示される最後の値を表示します。 |
| max | 重複する行を凝縮して1つのデータ行とし、参照列中の最大値を表示します。 |
| min | 重複する行を凝縮して1つのデータ行とし、参照列中の最小値を表示します。 |
| median | 重複する行を凝縮して1つのデータ行とし、参照列中の中央値を表示します。 |
| mode | 重複する行を凝縮して1つのデータ行とし、参照列に数字のモードを表示します。 |
| stdev | 重複する行を凝縮して1つのデータ行とし、参照列に数値の標準偏差を表示します。 |

| 動作 | 目的 |
| --- | --- |
| stdevp | 重複する行を凝縮して1つのデータ行とし、参照列内に含まれる母集団の標準偏差を表示します。 |
| sum | 重複する行を凝縮して1つのデータ行とし、参照列に値の合計を表示します。 |
| var | 重複する行を凝縮して1つのデータ行とし、参照列に数値の分散を表示します。 |
| varp | 重複する行を凝縮して1つのデータ行とし、参照列に含まれる母集団分散を表示します。 |

# array（配列）

使用可能な集計関数の中でも、**array** は、テキストと数値の両方に使用できるという点でユニークです。集約行に対して数学演算を実行するのではなく、参照列（**array** が適用される列）のすべての値が一時的に保存されます。ユニークな単一行が作成されると、**array**は、集合の参照列データを列内の単一のカンマ区切り文字列へと組み合わせます。

同一の行を検出するため、すべての行で（**array**が適用される列を除く）すべての行が列単位で照合されます。参照列の見出しの名前は、「*Array of*」に変更されます。

**例**

次の小さなデータセットを使用して、**array**がどのように演算を行うかを示します。

| 列 A | 列 B | 列 C |
| --- | --- | --- |
| 1 | two | 5 |
| 1 | two | 6 |
| two | two | 7 |
| 1 | two | 4 |

**例1**

*列C*に**array**関数を適用すると、行数が4行から2行に減ります。列*Array of Column C*の値は、演算中に折りたたまれた重複行の*列C*の合計を示します。

| 列 A | 列 B | 列 C |
| --- | --- | --- |
| 1 | two | 5.0, 6.0, 4.0 |

| 列 A | 列 B | 列 C |
|---|---|---|
| two | two | 7.0 |

**例2**

*列A*に **array** 関数を適用した場合、行数は減りません。 これは、*列B*と*列C*の各行の値の照合によって、各値がすでに一意であることが分かっているためです。したがって、列*Array of Column A*には、各行に元の値が表示されます。 ただし、数値はテキストに変換されています。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1.0 | two | 5 |
| two | two | 7 |
| 1.0 | two | 6 |
| 1.0 | two | 4 |

# average

この**average**集計関数は、すべての重複した行を一意の1行に折りたたむとともに、参照列（関数が適用される列）の数値の平均値を検出します。同一の行を検出するため、すべての行（**average** が適用される列を除く）が列単位で照合されます。

数学的平均（「算術平均」または単に「平均」とも呼ばれます）は、集合内の数値をすべて加算し、集合に含まれている項目数で結果として得られる合計を除算します。参照列のテキスト値に**average**を適用しようとすると、その行は0になります。参照列のヘッダーの名前は、*「Average of」*に変更されます。

関連する集計関数として、median と mode があります。

**例**

次の小さなデータセットを使用して、**average**がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 5 |
| 1 | two | 6 |

| 列 A | 列 B | 列 C |
|---|---|---|
| two | two | 7 |
| 1 | two | 4 |

**例1**

*列C*に**average**関数を適用すると、行数が4行から2行に減ります。列*Average of Colum*n Cの値は、演算中に折りたたまれた重複行の*列C*の値の平均を示します：(5 + 6 + 4) ÷ 3 = 5と7 ÷ 1 = 7）。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 5 |
| two | two | 7 |

**例2**

*列A*にaverage関数を適用した場合、行数は減りません。 これは、*列B*と*列C*の各行の値の照合によって、各値がすでに一意であることが分かっているためです。**average**関数が数学演算に関わる重複行を検出しなかったため、列*Average of Column A*の値は各行に元の数値（この例では数値1）が表示されます。0は、テキスト「2」を置き換えます。 これは、**average**演算がテキスト値に適用できないためです。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 5 |
| 1 | two | 6 |
| 1 | two | 4 |
| 0 | two | 7 |

# count

この **count**集計関数は、データセットに含まれる重複行の数を返します。適用される列（参照列）を除き、重複する行を検索するため、すべての行が列単位で照合されます。重複データを含む行は、一意の1行に折りたたまれます。参照列の名前は「*Count of*」に変更され、その列に表示される数値は折りたたまれた重複行の数を示します。

**例**

次の小さなデータセットを使用して、**count** がどのように演算を行うかを示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 5 |
| one | two | 6 |
| two | two | 7 |
| one | two | 4 |

**例1**

*列C*に**count**関数を適用すると、行数が4行から2行に減ります。列*Count of Column C*の値は、演算中に折りたたまれた重複行の数（回数）を示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 3 |
| two | two | 1 |

**例2**

*列A*に **count** 関数を適用した場合、行数は減りません。 これは、*列B*と*列C*の各行の値の照合によって、各値がすでに一意であることが分かっているためです。したがって、列*Count of Column A*の値は4行とも1の値になります。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 5 |
| 1 | two | 6 |
| 1 | two | 7 |
| 1 | two | 4 |

# count（数値のみ）

**count （数値のみ）**集計関数の演算は、count 関数とまったく同じになります。ただし、**count （数値のみ）**は数値だけをカウントし、集計プロセス中にテキスト値を無視します。

# count distinct

**count distinct**集計関数は、すべての値をカウントするcount関数とは異なり、カウントされる列の一意の値の数を返します。

# first

**first**集計関数は、（ユーザーが選択した列に基づいて）データの重複行を検索してそれらを1行のデータに凝縮します。**first**関数はその後、重複行に出現したデータ内の最初の値を表示します。最初に出現した値以外の値はすべて、この処理中に失われます。

**例**

次の小さなデータセットを使用して、**first**がどのように演算を行うかを示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 5 |
| one | two | 6 |
| two | two | 7 |
| one | two | 4 |

*列C*に**first**関数を適用すると、行数が4行から2行に減ります。*列First of Column C*の値は、演算中に折りたたまれた重複行の*列C*の最初の値を示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 5 |
| two | two | 7 |

# last

**last**集計関数は、（ユーザーが選択した列に基づいて）データの重複行のデータを検索してそれを1行のデータに凝縮します。**last**関数はその後、重複行に現れたデータ内の最後の値を表示します。最後に出現した値以外の値はすべて、この処理中に失われます。

**例**

次の小さなデータセットを使用して、**last** がどのように演算を行うかを示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 5 |
| one | two | 6 |
| two | two | 7 |
| one | two | 4 |

*列C*にlast関数を適用すると、行数が4行から2行に減ります。列*Last of Column C*の値は、演算中に折りたたまれた重複行の*列C*の最後の値を示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 4 |
| two | two | 7 |

# max

すべての重複行が1つの一意の行に折りたたまれるため、**max**集計関数は、参照列（関数が適用される列）の最大値を返します。同一の行を検出するため、すべての行（**max**が適用される列を除く）が列単位で照合されます。

この関数の逆に相当するのがmin（最小）です。

**例**

次の小さなデータセットを使用して、 **max** がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 5 |
| 1 | two | 6 |
| two | two | 7 |

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 4 |

*列C*に**max**関数を適用すると、行数が4行から2行に減ります。*列Max of Column C*の値は、演算中に折りたたまれた重複行の*列C*の最大値を示します。

**max**関数によって返されたデータセット（下記）では、 最初の行の数値6は、数値の集合{4, 5, 6}から得られた結果です。*列A*と*列B*が照合されたときに、各数値は同一の行の要素であったため、各数値はこの集合に存在しています。（*列C*がこの照合から除外されているのは、この列が参照列であるためです）。この3つの数値を含む集合では、6が最大です。そのため、それが参照列に表示される値となります。

7という数値は、他の数値を拠出する重複行がないため、{7}という1つの数値を含む集合から取得された結果です。7 はこの集合の最小値かつ最大値であるため、この行に対しては 7 が返されます。

# min

すべての重複行が1つの一意の行に折りたたまれるため、**min**集計関数は、参照列（関数が適用される列）の最小値を返します。同一の行を検出するため、すべての行（**min**が適用される列を除く）が列単位で照合されます。

この関数の逆に相当するのがmax（最大）です。

**例**

次の小さなデータセットを使用して、**min**がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 5 |
| 1 | two | 6 |
| two | two | 7 |
| 1 | two | 4 |

*列C*に**min**関数を適用すると、行数が4行から2行に減ります。*列Min of Column C*の値は、演算中に折りたたまれた重複行の*列C*の最小値を示します。

**min**関数によって返されたデータセット（下記）では、 最初の行の数値4は、数値の集合{4, 5, 6}から得られた結果です。*列A*と*列B*が照合されたときに、各数値は同一の行の要素であったため、各数値はこの集合に存在しています。（*列C*がこの照合から除外されているのは、この列が参照列であるためです）。この3つの数値を含む集合では、 4 が最小です。したがって、この値が参照列に表示されます。

7という数値は、他の数値を拠出する重複行がないため、{7}という1つの数値を含む集合から取得された結果です。7 はこの集合の最小値かつ最大値であるため、この行に対しては 7 が返されます。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 4 |
| two | two | 7 |

# median

すべての重複行が1つの一意の行に折りたたまれるため、**median**集計関数は参照列（medianが適用される列）の数値の中央値を検出します。同一の行を検出するため、すべての行（**median**が適用される列を除く）が列単位で照合されます。

中央値とは、ある範囲の数値を小さい順に並べたときに中央に位置する数値のことです。これは、数値の半数は返された値の「右」に含まれ、残りの半数は検出された値の「左」に含まれることを意味します。数値の集合の要素数が偶数の場合（つまり、集合の中央に位置する数値が1つではない場合）は、この関数は、範囲の中央にある数値のペア（つまり、中間点の左右にある2つの数値）の平均値を計算します。

参照列のテキスト値に **median**を適用しようとすると、その行の結果はエラーになります。参照列のヘッダーの名前は、「*Median of*」に変更されます。

関連する集計関数として、average と mode があります。

**例**

次の小さなデータセットを使用して、**median** がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 5 |
| one | two | 6 |
| two | two | 7 |
| one | two | 4 |

*列C*に**median**関数を適用すると、行数が4行から2行に減ります。列*Median of Column C*の値は、演算中に折りたたまれた重複行の*列C*の値の中央値を示します。

**median**関数によって返されるデータセット（下記）では、最初の行の数値5は、順序付けられた数値の集合{4, 5, 6}から得られた結果です。*列A*と*列B*が照合されたときに、各数値は同一の行の要素であったため、各数値はこの集合に存在しています。

（*列C*がこの照合から除外されているのは、この列が参照列であるためです）。この3つの数値の集合では、集合において5は中央値であり、その両側に1つずつ数値が含まれます。

7という数値は、他の数値を拠出する重複行がないため、{7}という1つの数値を含む集合から取得された結果です。7はこの集合の中央値であるため（その左右には数値が 1つもない）、この行に対しては7が返されます。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 5 |
| two | two | 7 |

# mode

モード（最頻値）とは、数値の集合の中で最も頻繁に出現する値です。**mode**集計関数は、同一の行の参照列（**mode**が適用される列）で最も頻繁に発生する同一の数値を検出します。すべての重複行は、一意の1つの行に折り畳まれる前に、列単位の照合を行って検出されます（**mode**が適用されている列を除く）。結果として得られる各行について、基になった重複行から得られた参照列の値は、**mode**関数の演算の対象となる集合の一部となります。

**mode**は参照列にテキストがある行を組み合わせますが、集合の実際のモードを見つける際にはテキストを無視します。参照列のヘッダーの名前は、*「Mode of」*に変更されます。

> **重要**
>
> 集合のモードに「引き分け」がある場合（つまり、複数の数値の発生回数が同数であり、より頻繁に現れる他の数値がない場合）、**mode**の結果は予測できません。適切な行の*Mode of*列に「引き分け」数値の1つが表示されますが、どちらの値が発生するかを決定することはできません。

関連する集計関数としてaverageとmedian があります。

**例**

次の小さなデータセットを使用して、**mode** がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 3 |
| one | two | 6 |
| two | two | 7 |

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 3 |

*列C*に**mode**関数を適用すると、行数が4行から2行に減ります。*列Mode of Column C*の値は、演算中に折りたたまれた重複行の*列C*の値のモードを示します。

結果として得られたデータセット（以下）では、最初の行の数値3は、重複していたが1つの行に折りたたまれた行により拠出された各要素を持つ、順序付けられた数値の集合{3, 6, 3}から得られたものです。この3つの数値を含む集合では、3が最も頻繁に出現しているため（3つ中2つ）、この集合のモードです。

7という数値は、他の数値を拠出する重複行がないため、{7}という1つの数値を含む集合から取得された結果です。集合の要素は 7 だけなので、これが最も頻繁に出現する値になります。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 3 |
| two | two | 7 |

# stdev

**stdev**（標準偏差）集計関数は、データのサンプル集合内に存在する標準偏差（平均値からのばらつきの量）を計算します。この集計関数は、同一である行における参照列（**stdev** が適用される列）の数値の標準偏差を計算します。

すべての重複行は、一意の1つの行に折りたたまれる前に、（**stdev**が適用される列を除き）列単位での照合によって検出されます。結果として得られる各行について、基になった重複行の参照列の値が、標準偏差計算の一部となります。参照列のヘッダーの名前は、*「Stdev of」*に変更されます。

参照列にテキスト値が存在する場合、その値は **stdev** の計算時に無視されます。また、 **stdev**集計関数には、少なくとも2つの値が必要となります。つまり、返された一意の行ごとに、集計に使用可能な同一の行が2行以上存在している必要があります。1回しか発生しない行では、計算に必要な参照列の値が1つしか得られないため、エラーが発生します。

データの標準偏差は、その分散の平方根です。分析対象の集合がすべてのデータポイント（「母集団」と呼びます）を表している場合は、より精度の高い結果を得るために stdevp を使用することをお勧めします。統計的分散を扱う関連関数として、varp があります。

**例**

次のデータセットを使用して、**stdev**がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.2 |
| one | two | 0.1 |
| one | two | 1.1 |
| one | two | 0.2 |
| one | two | 0.6 |
| one | one | 0.2 |
| one | one | 0.27 |
| one | two | 0.2 |
| one | two | 0.4 |

*列C*に**stdev**関数を適用すると、下表に示すように、行数が9行から2行に減ります。列*Stdev of Column C*の値は、演算中に折りたたまれた重複行の*列C*サンプルデータ値の標準偏差を示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.3511884584284246 |
| one | one | 0.049497474683058325 |

# stdevp

**stdevp**（母集団の標準偏差）集計関数は、データの集合全体（母集団）内に存在する標準偏差（平均値からのばらつきの量）を計算します。この集計関数は、同一の行の参照列（**stdevp**が適用される列）内の数値を使用して、母集団の標準偏差を計算します。

すべての重複行は、一意の1つの行に折りたたまれる前に、（**stdevp**が適用される列を除き）列単位での照合によって検出されます。結果として得られる各行について、基になった重複行の参照列から得られた値が、その母集団の標準偏差計算の一部となります。参照列のヘッダーの名前は、「*Stdev of*」に変更されます。

参照列にテキスト値が存在する場合、その値は**stdevp**の計算時に無視されます。また、**stdevp**集計関数には2つ以上の値が必要となります。つまり、返された一意の行ごとに、集計に使用可能な同一の行が2行以上存在している必要があります。1回しか発生しない行では、計算に必要な参照列の値が1つしか得られないため、エラーが発生します。

分析対象集合がデータのサンプルを表している場合は、より精度の高い結果を得るために stdev を使用することをお勧めします。統計的分散を扱うその他の関連関数として、varとvarpがあります。

**例**

次のデータセットを使用して、**stdevp** がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.2 |
| one | two | 0.1 |
| one | two | 1.1 |
| one | two | 0.2 |
| one | two | 0.6 |
| one | one | 0.2 |
| one | one | 0.27 |
| one | two | 0.2 |
| one | two | 0.4 |

*列C*に**stdevp**関数を適用すると、下表に示すように、行数が9行から2行に減ります。列*StdevP of Column C*の値は、演算中に折りたたまれた重複行の*列C*に含まれている母集団の標準偏差を示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.32513733362117264 |
| one | one | 0.034999999999999996 |

# sum

すべての重複行は一意の1つの行に折りたたまれるため、この**sum**集計関数は、参照列（この関数が適用される列）の数値に加算演算を行います。同一の行を検出するため、すべての行が（**sum**が適用される列を除き）列単位で照合されます。参照列のテキスト値に**sum**を適用しようとすると、その行の結果は0になります。参照列のヘッダーの名前は、*「Sum of」*に変更されます。

**例**

次の小さなデータセットを使用して、**sum**がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 5 |
| 1 | two | 6 |
| two | two | 7 |
| 1 | two | 4 |

**例1**

*列C*に**sum**関数を適用すると、行数が4行から2行に減ります。列*Sum of Column C*の値は、演算中に折りたたまれた重複行の*列C*の合計を示します（5 + 6 + 4 = 15と7 + 0 = 7）。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 15 |
| two | two | 7 |

**例2**

*列A* に**sum**関数を適用した場合、行数は減りません。 これは、各行の*列B*と*列C*の値の照合によって、各値がすでに一意であることが証明されているためです。**sum**関数が加算演算に関与している可能性がある重複行を検出しなかったため、*Sum of Column A*列の値は元の数値（この場合、数値 1）で各行を表示します。0は、テキスト「2」を置き換えます。 これは、**sum**演算がテキスト値に適用できないためです。

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 5 |

| 列 A | 列 B | 列 C |
|---|---|---|
| 1 | two | 6 |
| 1 | two | 4 |
| 0 | two | 7 |

# var

**var**（分散）集計関数は、データのサンプル集合内に値がどの程度拡散しているかを推定します。この集計関数は、同一の行の参照列（この関数が適用される列）の数値の分散を計算します。

すべての重複行は、一意の1つの行に折りたたまれる前に、（**var**が適用される列を除いて）列単位の照合によって検出されます。結果として得られる各行について、基になった重複行の参照列の値が、分散計算の一部となります。参照列のヘッダーの名前は、*「Var of」*に変更されます。

参照列にテキスト値がある場合、その値は**var**の計算内で無視されます。また、**var**集計関数には2つ以上の値が必要となります。つまり、返された一意の行ごとに、集計に使用可能な同一の行が2行以上存在している必要があります。1回しか発生しない行では、計算に必要な参照列の値が1つしか得られないため、エラーが発生します。

分析対象の集合がすべてのデータ ポイント（「母集団」と呼びます）を表している場合は、より精度の高い結果を得るためにvarp を使用することをお勧めします。統計的分散を扱う関連関数として、stdevとstdevpがあります。

**例**

次のデータセットを使用して、**var**がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.2 |
| one | two | 0.1 |
| one | two | 1.1 |
| one | two | 0.2 |
| one | two | 0.6 |
| one | one | 0.2 |

| 列 A | 列 B | 列 C |
|---|---|---|
| one | one | 0.27 |
| one | two | 0.2 |
| one | two | 0.4 |

*列C*に**var** 関数を適用すると、下表に示すように、行数が9行から2行に減ります。列*Var of Column C*の値は、演算中に折りたたまれた重複行における*列C*サンプルデータ値の分散を示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.12333333333333334 |
| one | one | 0.00245 |

# varp

**varp**（母集団の分散）集計関数は、データの全体集合（母集団）の値の分散度を計算します。この集計関数は、同一の行の参照列（この関数が適用される列）に含まれている母集団の分散を計算します。

すべての重複行は、一意の1つの行に折りたたまれる前に、（**varp**が適用される列を除いて）列単位での照合によって検出されます。結果として得られる各行について、基になった重複行の参照列の値が、母集団の分散計算の一部となります。参照列のヘッダーの名前は、「*VarP of*」に変更されます。

参照列にテキスト値がある場合、その値は **varp** の計算内で無視されます。また、**varp**関数には2つ以上の値が必要となります。つまり、返された一意の行ごとに、集計に使用可能な同一の行が2行以上存在している必要があります。1回しか発生しない行では、計算に必要な参照列の値が1つしか得られないため、エラーが発生します。

分析対象の集合がデータのサンプル集合を表している場合、より精度の高い結果を得るためには**var**を使用することをお勧めします。統計的分散を扱う関連関数として、stdevとstdevpがあります。

**例**

次のデータセットを使用して、**varp**がどのように演算を行うかを示します。この集計関数の動作を説明するという目的から、以下の表の数字はすべて（テキスト値ではなく）数値と考えてください。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.2 |

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.1 |
| one | two | 1.1 |
| one | two | 0.2 |
| one | two | 0.6 |
| one | one | 0.2 |
| one | one | 0.27 |
| one | two | 0.2 |
| one | two | 0.4 |

*列C* に**varp**関数を適用すると、下表に示すように、行数が9行から2行に減ります。列*VarP of Column C*の値は、演算中に折りたたまれた重複行の*列C*に含まれている母集団の分散を示します。

| 列 A | 列 B | 列 C |
|---|---|---|
| one | two | 0.10571428571428572 |
| one | one | 0.001225 |

# 行の自動番号付け

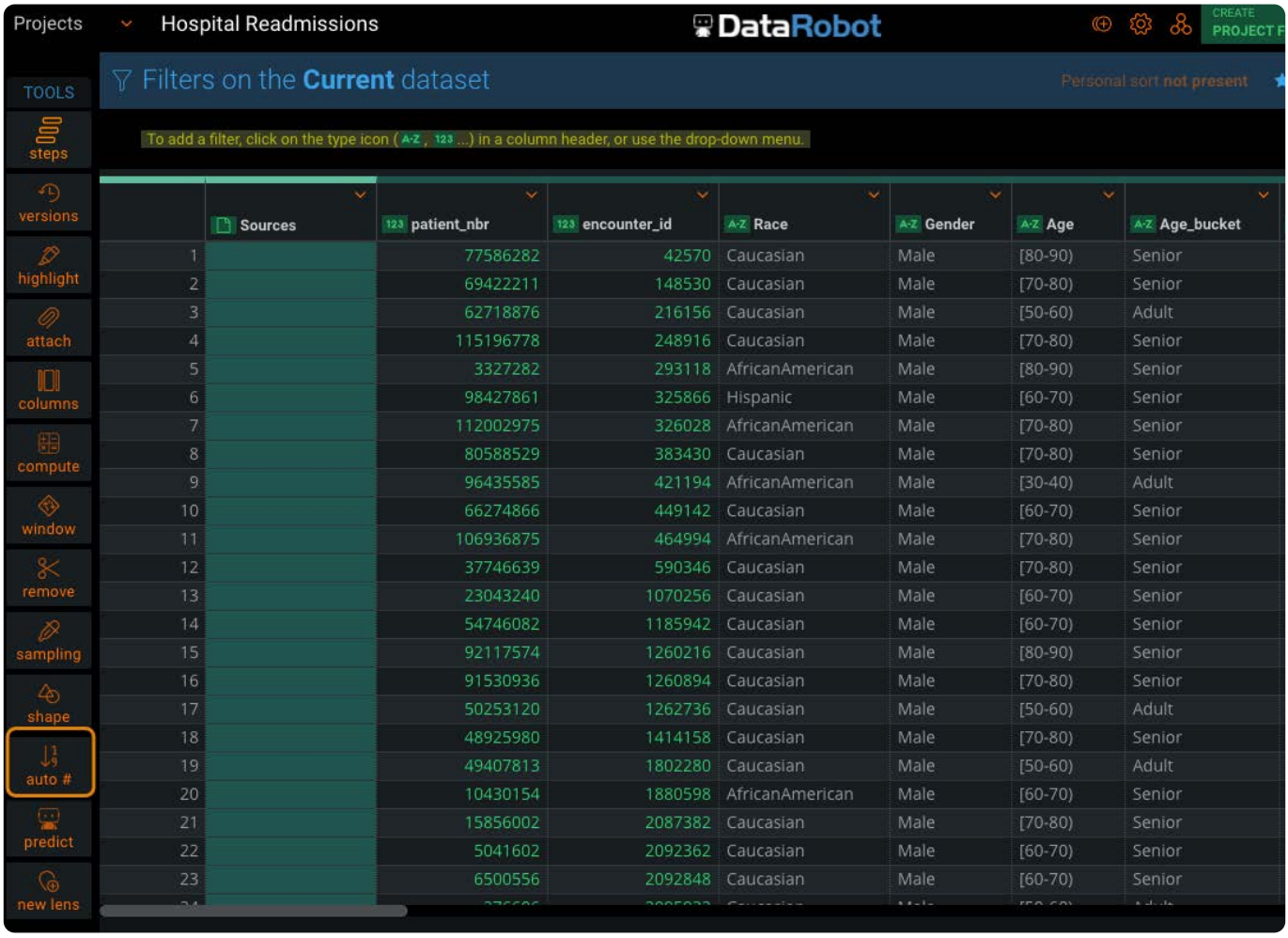
Data Prep **Auto #**ツールは、各行に番号を割り当てます。このツールは、各行に一意の識別子を設定する場合に便利です。**auto #**ツールは、行に自動番号を付ける列を新規作成します。

**auto #**ツールは各行に一意の識別子を設定するため、次のことが必要な場合に役立ちます。

- データセットの元の順序を追跡します。
- データセットに行識別子を割り当てます。

## auto #ツールの操作

**auto #**ツールにアクセスするには、プロジェクト**ツール**バーで**auto #**をクリックします。

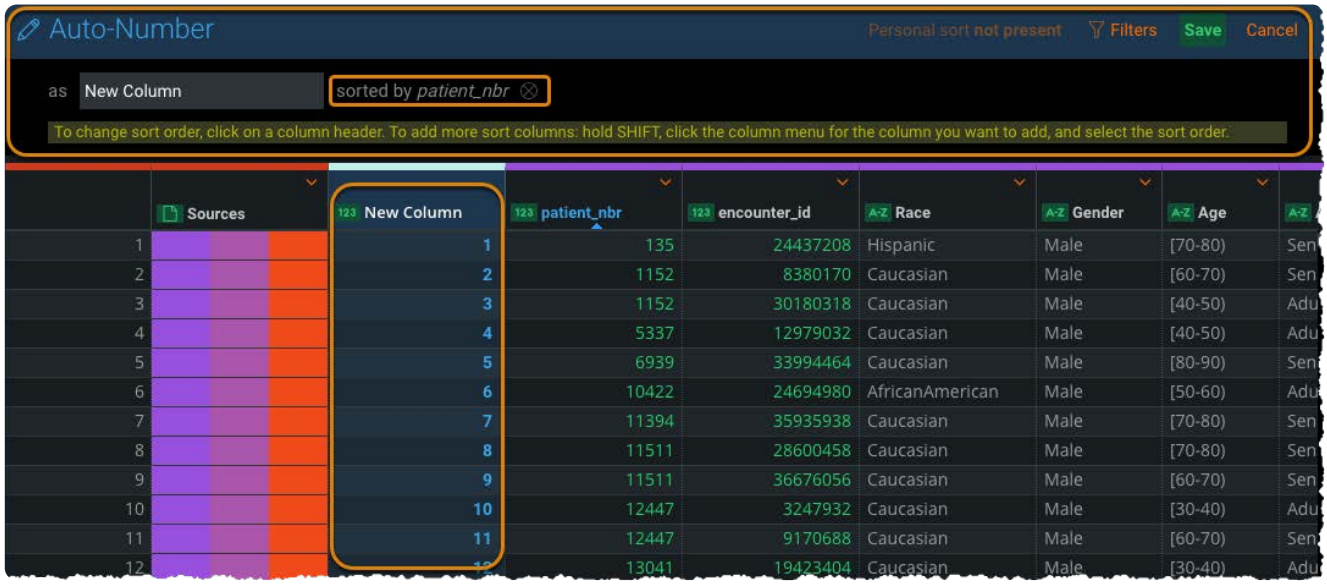


# 自動番号付き列を追加

自動番号付きの列を追加するには、次の手順に従います：

1. **ツール**バーから、**auto #**をクリックします。

   **自動番号**ペインが表示されます。




2. 追加する自動番号付き列の名前を入力します。

   デフォルト名は"New Column(#)"です。ここで、＃は1から始まり、新しい自動番号付けされた列ごとに増加します。

3. 自動番号付けされた列をデータセット内の既存の列の並べ替え順序にバインドするには、選択する列の見出しをクリックします。

   列は**並べ替え**フィールドにリストされています。

4. 並べ替え列をさらに追加するには、Shiftキーを押しながら、追加する列から列メニューにカーソルを合わせ、**昇順で並べ替え**または**降順で並べ替え**をクリックします。

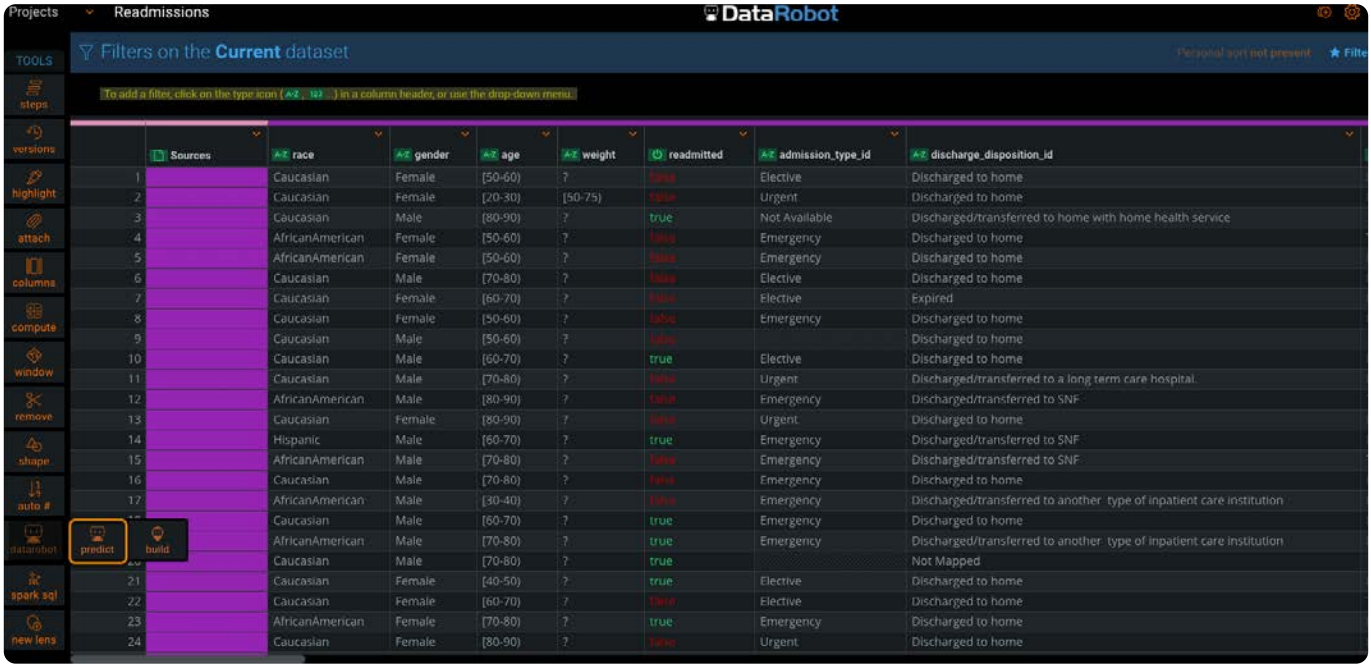
5. 並べ替え用に選択した列を削除するには、**自動番号**ペインの列名の横にある[X]アイコンをクリックします。

# 予測を作成

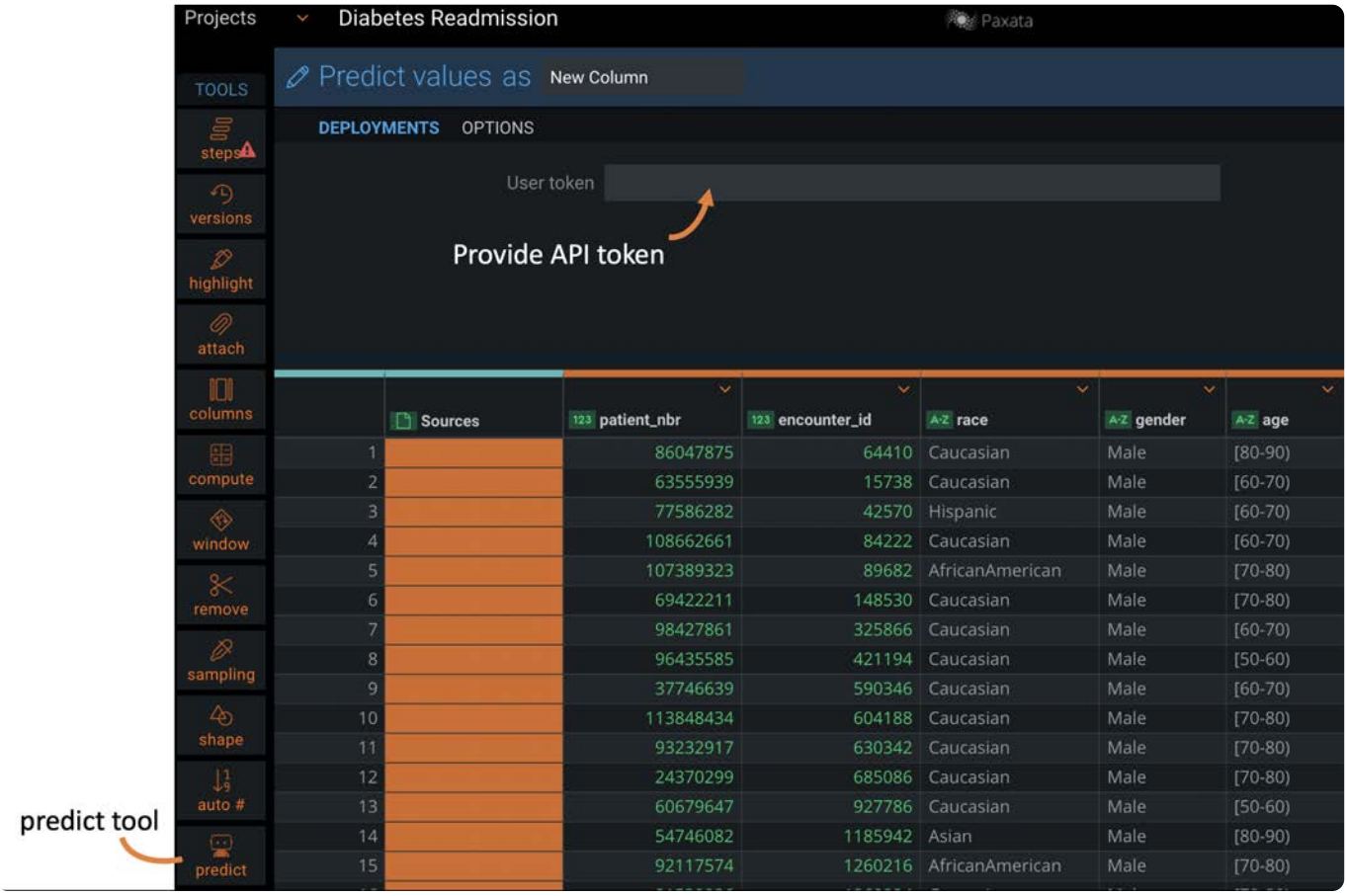
DataRobot内にデプロイされた機械学習（ML）モデルに対してスコアリングする必要のあるデータがある場合、Data Prep**予測**ツールでスコアを生成します。

## 予測ツールの操作

**予測**ツールにアクセスするには、**ツール**バーのDataRobotアイコンをクリックし、**予測**を選択します。




スコアを生成するために、DataRobot APIトークンを提供します。これは、DataRobotデプロイメントリストを取得するために使用されます。




> **備考**
>
> トークンを取得するには、**ユーザー設定** > **開発者ツール** > **APIキー**に移動します。

次に、デプロイメントを選択します。あなたのデータは、このデプロイメントでモデルに対してスコアリングされます。スコアリングに使用するモデルが時系列モデルの場合、**時系列モデル**チェックボックスチェックを選択して、これを指定する必要があります。次に、**オプション**タブで**予測ポイント**指定し、任意で**系列ID**を指定します。詳しくはオプションをご覧ください。

> **備考**
>
> カスタムモデルのデプロイメントは現在サポートされていません。

デフォルトでは、データセットに予測スコアの新しい列が「目標」として作成されます。この名前を変更するには、**オプション**タブをクリックし、**予測列**フィールドに別の名前を入力します。

デプロイメントを選択すると、予測が実行されます。新しい列が作成され、予測スコアが提供されます。さらに、「目標予測値」列も生成され、各スコアに関連する予測値を提供します。マルチクラス予測の場合、予測値は分類ごとに返されます。例えば、画像を"apple"、"orange"、"pear"に分類した場合、さらに3つの列が返され、それぞれの対応するスコアに1つの値があります。

# ユースケース予測値の例

- 病院の患者が退院後に再入院する確率を予測します。予測列には、患者が再入院する可能性が高いか、再入院しないかを示す1または0のバイナリ値が格納されます。
- 画像の集合をオレンジ、梨、リンゴ3つの果物のいずれかに分類します。予測列には、オレンジ、梨、リンゴの3つの値のいずれかが入ります。
- 予測日に基づいて売上を予測します。この場合の予測列には、売上のドル換算額が入ります。

バイナリおよび時系列予測デプロイの場合、**オプション**タブが追加オプションを提供します。詳しくはオプションをご覧ください。

# オプション

時系列予測では、予測ポイントも提供する必要があります。予測ポイントは、予測の作成元となるポイント、つまり「もし今現在だったら」の相対時間です。DataRobotでは、トレーニングデータのすべての潜在的予測ポイントを使用してモデルをトレーニングします。運用環境では、これは一般的に直近の時間です。

> **重要**
>
> この日付のフォーマットは、ISO 2014-08-12T00:00:00Zでなければなりません。

オプションとして、データセットに複数の時系列データが含まれている場合、例えば、複数の店舗の売上を予測するための複数の時系列データが含まれている場合、シリーズIDとして列を指定することで、データをグループ化し、グループごとに別々に予測値を返すことができます。

二値予測の場合、**オプション**タブは、予測が返された理由の理解に役立つ予測の説明を提供します。例えば、「この患者が再入院する可能性のスコアが1である理由」などです。または、「なぜこの画像はリンゴと認識したのでしょうか？」などです。

**説明**を有効にすると、プロジェクトの説明ごとに5つの新しい列が生成されます。

- **特徴量**：予測に貢献する特徴量の名前。
- **特徴量値**：この行に対して特徴量が取った値。
- **力**：この特徴量の値が予測に影響した量。
- **定性**：特徴量が予測に影響した強度を示す人間が読み取ることのできる説明。例えば、以下のようになります。++++; -; +
- **ラベル**：この予測の説明から派生した出力を説明します。連続値プロジェクトの場合、これはターゲット特徴量の名前です。分類プロジェクトでは、その確率が高まれば、この予測説明の正の強さに対応することになるクラスのことです。

さらに、閾値の低い値と高い値を設定することで、閾値を超えたスコアに対してのみ説明を生成することができます。

予測値として返される値の詳細については、予測説明を参照してください。

# Spark SQLでのデータの変換

Data Prepは、Spark SQLを使用したデータ変換を可能にするツールを提供します。Spark SQLは、データの準備、クリーンアップ、変換のための関数のライブラリを提供します。

**備考**

Data Prepの管理者は、アプリケーションでこの特徴量を有効にする必要があります。

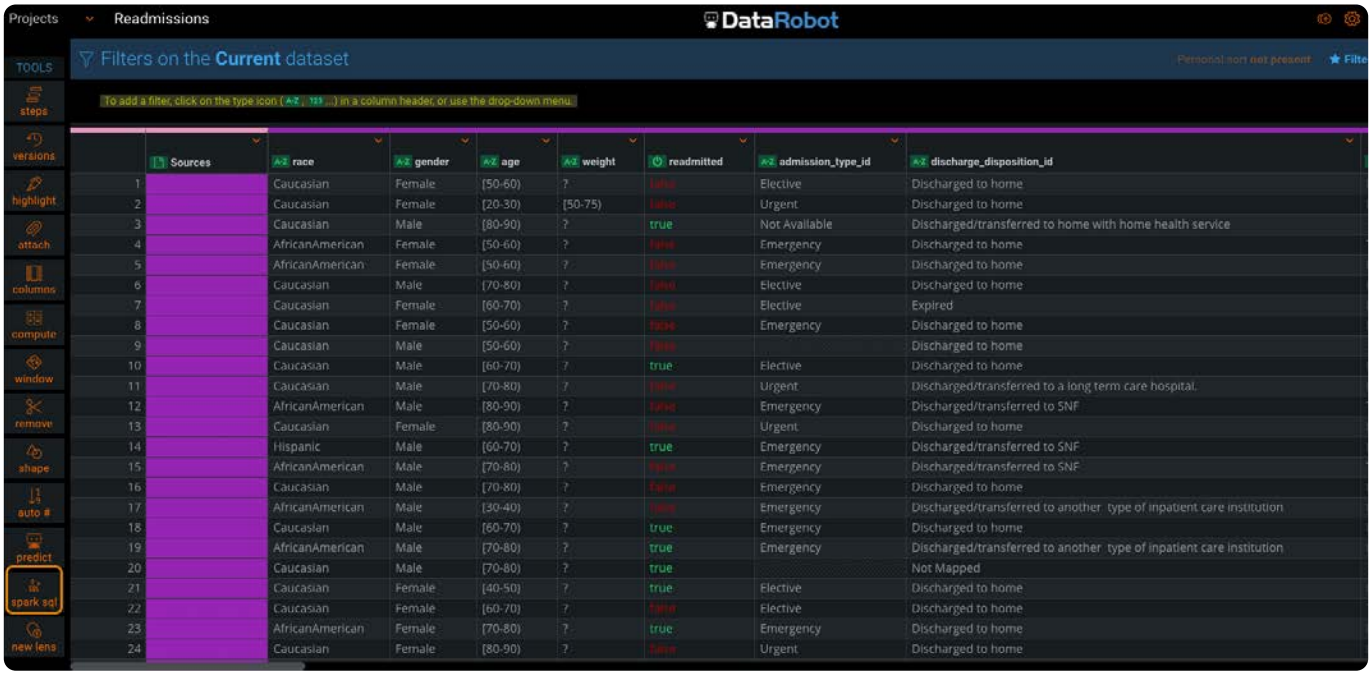
次のセクションでは、Data PrepでのSpark SQLツールの使用方法を説明しています。サポートされているSQLステートメントについては、Data Prep Spark SQLガイドラインを参照してください。

**ヒント**

AIカタログを使用してデータを選択し、変換することもできます。AIカタログでSpark SQLを使用したデータ準備を参照してください。

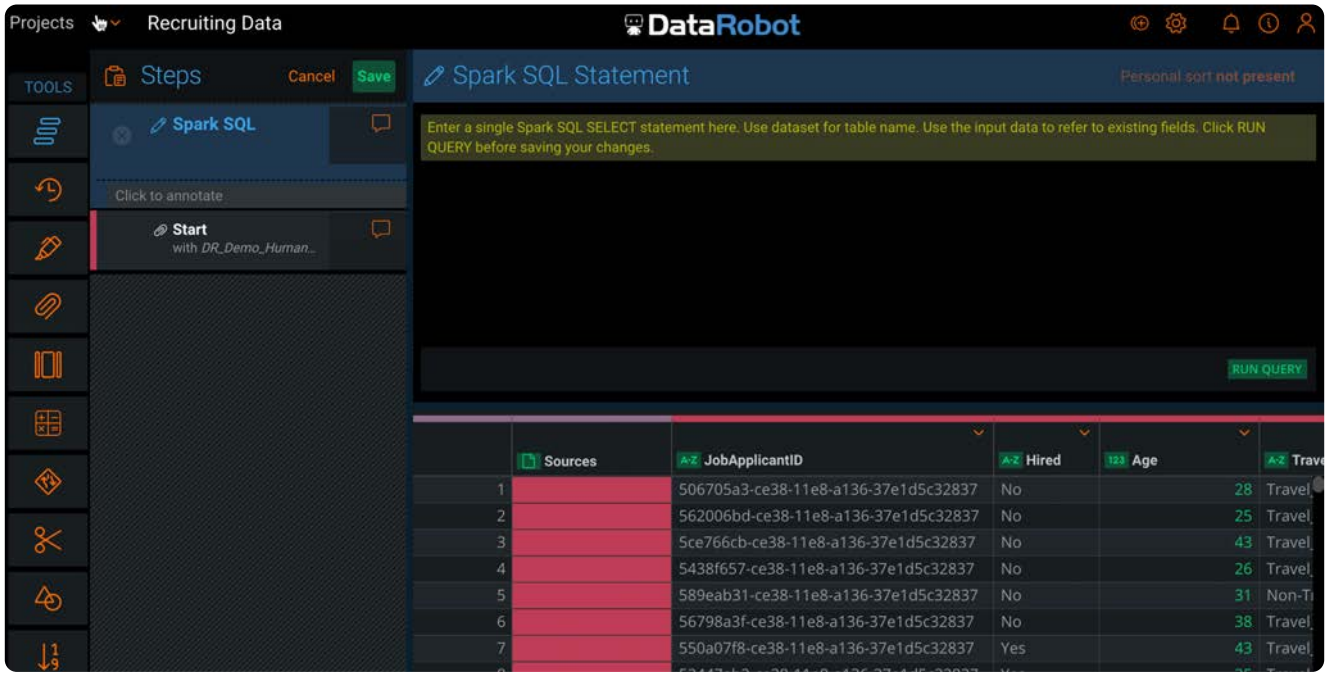
## Spark SQLの操作

**Spark SQL**ツールにアクセスするには、プロジェクトの**ツール**バーで**spark sql**をクリックします。

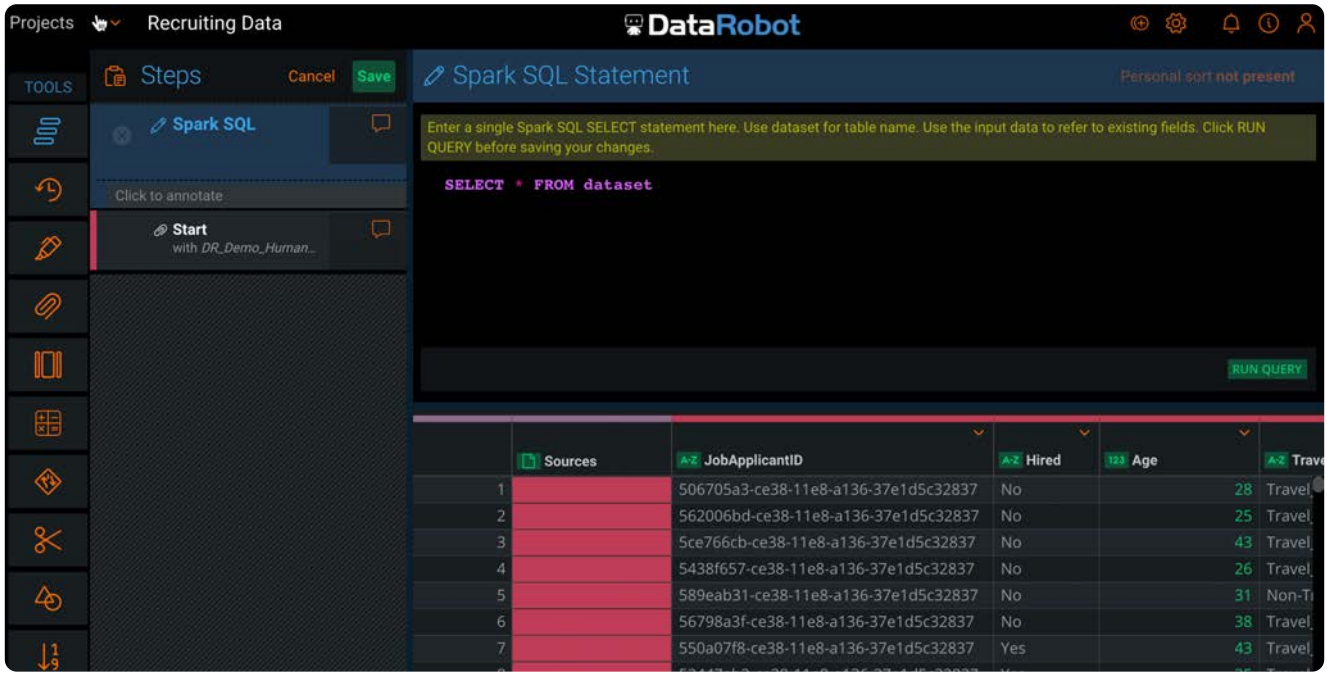


# Spark SQLステートメントの追加

1. **ツール**バーから、**spark sql**をクリックします。**Spark SQLステートメント**ペインが表示されます。

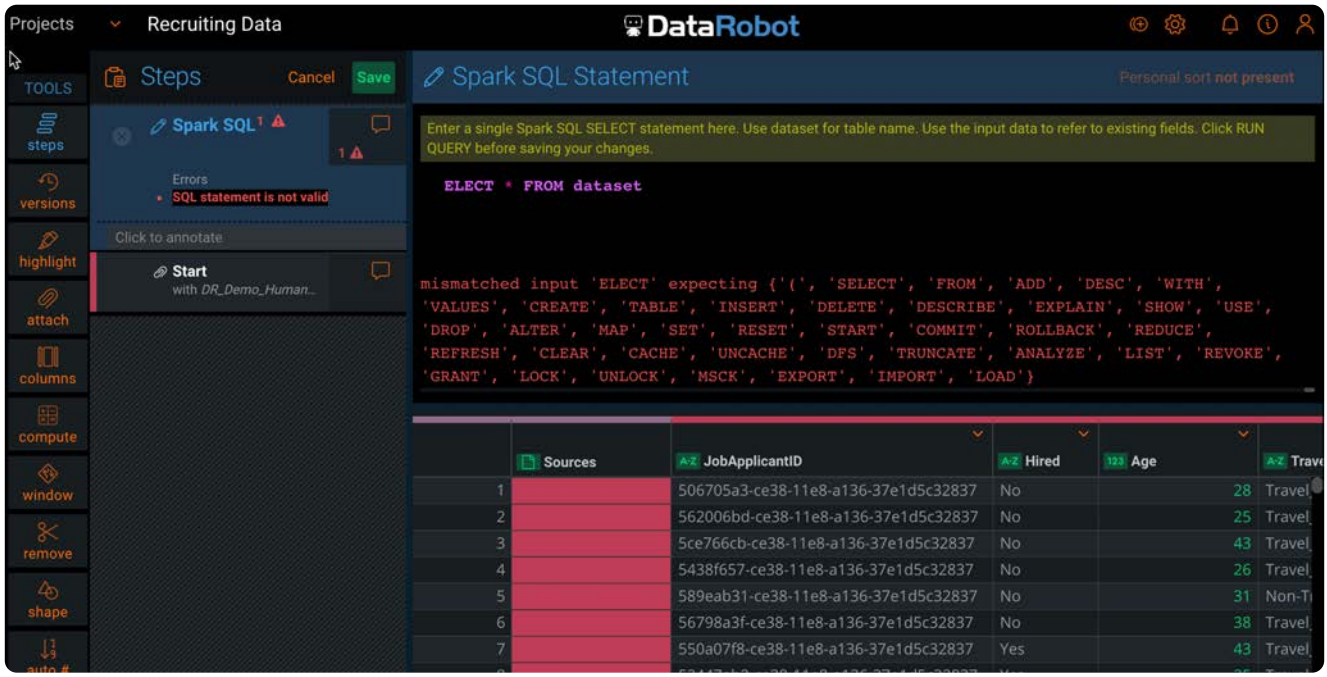



2. Spark SQLステートメントを入力します。使用方法の詳細については[Data Prep Spark SQLガイドライン]を参照してください。




3. **Spark SQLステートメント**ペインの右下にある**クエリーを実行**をクリックして、クエリーを検証します。クエリーが正常に完了した場合、結果が下に表示されます。結果を見て、クエリーが想定通りに機能していることを確認します。

   クエリーが失敗した場合、クエリーの下にエラーメッセージが表示されます。




4. **保存**をクリックすると、クエリーが保存されます。エラーが発生したクエリーを保存し、後で戻ってそのエラーを解決できます。

   SQLクエリーにエラーが含まれている場合、ステップツールのSpark SQLのステップにエラーアイコン（  ）が表示されます。アイコンをクリックするとエラーメッセージが表示されます。

**備考**

ステップツールでSpark SQLステップを保存した後、前のステップに変更を加えたり、Spark SQLステップの前に新しいステップを追加する必要がある場合があります。この場合、Spark SQLステップをクリックして編集してから、**クエリーを実行**をクリックし、クエリーを再度保存します。

# Data Prep Spark SQLガイドライン

構造化照会言語（SQL）は、リレーショナルデータベースに保存されたデータの管理に設計済みの宣言言語です。Spark SQLは、登録されたデータフレーム（名前のついた列に整理されたデータ）に対して、データベースに照会するために使用されるSQLと同じように、クエリを書くことができるSparkのコンポーネントです。Data PrepはSpark SQLツールで使用するSpark SQL関数のライブラリをサポートしています。

## ユースケース

このセクションで示すデータセットの例には、ターゲット特徴量 `Hired` を含むジョブアプリケーションデータのサンプルが含まれています。

| JobApplicantID | Hired | Age | TravelPreference | HiringDepartment | DistanceFro | EducationLev | EducationField | Gender | Role | Internships | Over18 | StandardHours | Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 506705a3-ce38-11e8- | No | 28 | Travel_Frequently | Research & Develo | 4 | 3 | Engineering | Male | Research Scientist | 9 | Y | 80 | Hi there. I am a |
| 562006bd-ce38-11e8- | No | 25 | Travel_Rarely | Research & Develo | 3 | 4 | Medical | Male | Research Scientist | 0 | Y | 80 | Hi there. I am a |
| 5ce766cb-ce38-11e8- | No | 43 | Travel_Rarely | Human Resources | 4 | 3 | Life Sciences | Male | Human Resources | 4 | Y | 80 | Hi there. I am a |
| 5438f657-ce38-11e8- | No | 26 | Travel_Rarely | Sales | 5 | 3 | Other | Male | Sales Representativ | 0 | Y | 80 | Hi there. I am a |
| 589eab31-ce38-11e8- | No | 31 | Non-Travel | Sales | 16 | 3 | Life Sciences | Male | Sales Executive | 1 | Y | 80 | Hi there. I am a |
| 56798a3f-ce38-11e8- | No | 38 | Travel_Rarely | Research & Develo | 25 | 2 | Life Sciences | Male | Research Director | 3 | Y | 80 | I am a Life Scie |
| 550a07f8-ce38-11e8- | Yes | 43 | Travel_Rarely | Research & Develo | 8 | 4 | Other | Male | Research Scientist | 9 | Y | 80 | I am a Other e |
| 52447ab2-ce38-11e8- | Yes | 25 | Travel_Rarely | Research & Develo | 9 | 3 | Medical | Male | Research Scientist | 1 | Y | 80 | Recently I com |
| 5b2194be-ce38-11e8- | No | 34 | Travel_Rarely | Research & Develo | 1 | 4 | Life Sciences | Male | Research Scientist | 2 | Y | 80 | Recently I com |
| 599c1ce1-ce38-11e8- | No | 27 | Travel_Rarely | Research & Develo | 1 | 2 | Medical | Male | Laboratory Technicia | 1 | Y | 80 | Recently I com |
| 585d8dd3-ce38-11e8- | No | 32 | Travel_Frequently | Research & Develo | 2 | 2 | Life Sciences | Male | Laboratory Technicia | 0 | Y | 80 | Hi there. I am a |
| 5cdb0cb8-ce38-11e8- | Yes | 45 | Travel_Rarely | Research & Develo | 1 | 4 | Engineering | Male | Healthcare Represer | 1 | Y | 80 | I am a Enginee |
| 4fd933fa-ce38-11e8- | No | 37 | Travel_Rarely | Research & Develo | 11 | 3 | Medical | Male | Laboratory Technicia | 4 | Y | 80 | Recently I com |
| 536044e5-ce38-11e8- | No | 27 | Non-Travel | Research & Develo | 9 | 3 | Medical | Male | Research Scientist | 6 | Y | 80 | Hi there. I am a |
| 50003dd0-ce38-11e8- | No | 42 | Travel_Rarely | Research & Develo | 28 | 3 | Life Sciences | Male | Research Director | 3 | Y | 80 | Hi there. I am a |
| 5b15fdd2-ce38-11e8- | Yes | 34 | Travel_Rarely | Research & Develo | 29 | 3 | Medical | Male | Laboratory Technicia | 4 | Y | 80 | Recently I com |
| 5628df08-ce38-11e8- | No | 30 | Travel_Rarely | Research & Develo | 6 | 3 | Engineering | Male | Laboratory Technicia | 0 | Y | 80 | Hi there. I am a |
| 5ad72a0a-ce38-11e8- | No | 39 | Travel_Frequently | Sales | 1 | 3 | Marketing | Male | Sales Executive | 0 | Y | 80 | Recently I com |
| 50844cfc-ce38-11e8- | No | 47 | Travel_Frequently | Sales | 27 | 2 | Life Sciences | Male | Sales Executive | 4 | Y | 80 | Hi there. I am a |
| 5b384d75-ce38-11e8- | No | 33 | Travel_Rarely | Research & Develo | 25 | 3 | Engineering | Male | Manufacturing Direc | 3 | Y | 80 | Recently I com |
| 515b5104-ce38-11e8- | No | 42 | Travel_Rarely | Research & Develo | 24 | 3 | Medical | Male | Manufacturing Direc | 1 | Y | 80 | Hi there. I am a |
| 50485ef4-ce38-11e8- | No | 41 | Travel_Rarely | Sales | 9 | 3 | Marketing | Male | Sales Executive | 8 | Y | 80 | Hi there. I am a |
| 5bf05cb9-ce38-11e8- | No | 35 | Travel_Frequently | Research & Develo | 25 | 4 | Life Sciences | Male | Research Scientist | 1 | Y | 80 | Hi there. I am a |
| 5088e020-ce38-11e8- | No | 29 | Non-Travel | Research & Develo | 1 | 4 | Medical | Male | Manufacturing Direc | 0 | Y | 80 | Hi there. I am a |
| 5387c3d4-ce38-11e8- | Yes | 42 | Travel_Frequently | Research & Develo | 9 | 2 | Medical | Male | Laboratory Technicia | 1 | Y | 80 | I am a Medical |
| 50f59a74-ce38-11e8- | No | 36 | Travel_Rarely | Research & Develo | 25 | 2 | Life Sciences | Male | Research Director | 3 | Y | 80 | Hi there. I am a |
| 519bf949-ce38-11e8- | No | 32 | Travel_Rarely | Research & Develo | 1 | 1 | Life Sciences | Male | Research Scientist | 1 | Y | 80 | Recently I com |
| 579be332-ce38-11e8- | No | 27 | Travel_Rarely | Human Resources | 17 | 3 | Other | Male | Human Resources | 1 | Y | 80 | Hi there. I am a |
| 5509e0e1-ce38-11e8- | No | 50 | Travel_Rarely | Research & Develo | 2 | 3 | Medical | Male | Research Director | 5 | Y | 80 | Hi there. I am a |
| 59f3f2f6-ce38-11e8- | No | 33 | Travel_Rarely | Research & Develo | 4 | 4 | Medical | Male | Laboratory Technicia | 8 | Y | 80 | I am a Medic |

## クエリーガイドライン

Data Prepにデータセットが読み込まれると、DataFrameとして登録され、Spark SQLステートメント内の `dataset` エイリアスを使用してクエリーできるようになります。

Spark SQLを使用してデータを形成するには、Spark SQLツールを有効にし、**Spark SQLステートメント**ペインにSQLクエリーを入力します。

Data Prepでは、 `SELECT` クエリーのみ許可されています。列名を使用してクエリーを構築します。例を以下に示します。

```
SELECT
  EducationLevel,
```

```
  Hired
FROM dataset
```

その他の制限については、禁止されているキーワードと関数を参照してください。

# サンプルクエリー

以下に示すのは、Data Prepで使用されるSpark SQLクエリーとその結果の説明です。

> **備考**
>
> Data Prepでは、SQLステートメントは大文字と小文字を区別し、キーワードには大文字のみ、特徴量名には小文字のみを使用するという共通のSQL規則に従います。

**例1**

```
SELECT * FROM dataset
```

更新されたデータセットには、データセットからのすべての列とすべての行が含まれています。

**例2**

```
SELECT * FROM dataset
```

結果は `ParseException` となります。

**例3**

```
SELECT * from doesNotExist
```

結果は `NoSuchTableException` となります。

**例4**

```
SELECT Hired FROM dataset
```

更新されたデータセットには（アプリケーションが成功したかどうかを含む） `Hired` 列のみとすべての行が含まれています。

**例5**

```
SELECT
  EducationLevel,
  Hired
FROM dataset
WHERE EducationLevel = 5
```

更新されたデータセットには、 `Hired` および `EducationLevel` 列と、 `EducationLevel` が5であるこれらの行だけが含まれています。

**例6**

```
SELECT
    EducationLevel,
    CASE WHEN Hired = 'No' THEN 0 ELSE 1 END
AS HiredNum
FROM dataset
WHERE EducationLevel = 5
```

更新されたデータセットには `EducationLevel` と、 `EducationLevel` が5である行の `Hired` 列の数値バージョンが含まれています。

**例7**

```
SELECT
    EducationLevel,
    avg(CASE WHEN Hired = 'No' THEN 0 ELSE 1 END) AS acceptance_rate
FROM dataset
GROUP BY EducationLevel
ORDER BY EducationLevel
```

更新されたデータセットには、 `EducationLevel` を基準に、 `EducationLevel` と、 `EducationLevel` グループ内の平均受入れ率が含まれています。

**例8**

```
SELECT
    EducationLevel,
    avg(CASE WHEN Hired = 'No' THEN 0 ELSE 1 END) AS acceptance_rate,
    std(CASE WHEN Hired = 'No' THEN 0 ELSE 1 END) AS acceptance_rate_std
FROM dataset
GROUP BY EducationLevel
ORDER BY EducationLevel
```

更新されたデータセットには、 `EducationLevel` を基準に、 `EducationLevel` と、 `EducationLevel` グループ内の受入れ率の平均と標準偏差が含まれています。

**例9**

```
SELECT
    EducationLevel,
    length(Summary) AS length_summary
FROM dataset
```

ここには、アプリケーションに沿った `EducationLevel` とカバーレターの長さが表示されます。

**例10**

```
SELECT
    EducationLevel,
```

```
  avg(length(Summary)) AS avg_length_summary,
  std(length(Summary)) AS std_length_summary,
  std(CASE WHEN Hired = 'No' THEN 0 ELSE 1 END) AS acceptance_rate_std,
  avg(CASE WHEN Hired = 'No' THEN 0 ELSE 1 END) AS acceptance_rate
FROM dataset
GROUP BY EducationLevel
ORDER BY EducationLevel
```

この例は、更新されたデータセットには、 `EducationLevel` を基準に、 `EducationLevel` 、 `EducationLevel` グループ内の受け入れ率の平均と標準偏差、およびサマリー長さの平均と標準偏差が含まれます。

# 禁止されたキーワードと関数

一部のSparkSQL 2.4.0関数には、セキュリティリスクがあります。Data Prepは、潜在的なセキュリティリスクを高くするコマンドとメソッドの使用を禁止しています。

以下のタブをクリックして、禁止されたキーワードと関数のリストを表示します。

| キーワード | 関数 |
| --- | --- |
| ALTER (DATABASE\| SCHEMA) | |
| ALTER [TABLE \| VIEW] | |
| ALTER VIEW | |
| CREATE (DATABASE\| SCHEMA) | |
| CREATE FUNCTION | |
| CREATE TABLE | |
| CREATE VIEW | |
| DROP DATABASE | |
| DROP FUNCTION | |
| DROP [TABLE \| VIEW] | |
| MSCK REPAIR TABLE or ALTER TABLE RECOVER PARTITIONS | |
| TRUNCATE TABLE | |
| USE | |
| LOAD DATA | |
| EXPLAIN | |
| ADD FILE | |
| ADD JAR | |
| ANALYZE TABLE | |
| CACHE TABLE | |
| CLEAR CACHE | |
| DESCRIBE DATABASE | |
| DESCRIBE FUNCTION | |
| DESCRIBE TABLE | |
| LIST FILE | |
| LIST JAR | |

REFRESH

REFRESH TABLE

RESET

SET

SHOW

COLUMNS

SHOW CREATE

TABLE

SHOW

(DATABASES|

SCHEMAS)

SHOW

FUNCTIONS

SHOW

PARTITIONS

SHOW TABLES

or SHOW TABLE

EXTENDED

SHOW

TBLPROPERTIES

UNCACHE

TABLE

DESCRIBE

CREATE TEMP

VIEW USING

CREATE TABLE

# 注意事項

INSERT [INTO |

OVERWRITE]

CREATE TABLE

INSERT

OVERWRITE

DIRECTORY

java_method

reflect

- Data Prep Spark SQLツールは、Spark SQL 2.4.0で使用可能なコマンドと関数をサポートしています。
- `SELECT` クエリーのみが許可されています。現在のデータセットを変更する可能性のあるすべてのSQLの操作は、禁止されています（ `INSERT` 、 `UPDATE` 、 `DELETE` など）。
- システムに悪影響を与える可能性がある組み込み関数も禁止されています（ `reflect` 、 `java_method` など）。禁止されたキーワードと関数を参照してください。
- 各SQLステートメントはSparkセッションの下で実行されるため、SQL実行間にデータが漏洩する可能性があります。
- データ型の処理の場合：
  - 無限大がサポートされており、nullとして処理されます。
  - 列にデータ型が混在している場合、変換可能であれば、その値は指定された型に変換されます。変換可能でなければ、nullに設定されます。
  - SQL結果をData Prepデータセットに変換すると、SQL列からのすべてのデータ型がData Prepデータ型に変換されます。型がサポートされていない場合、文字列型に変換されます。

# DRレンズを使用したDataRobotプロジェクトの作成

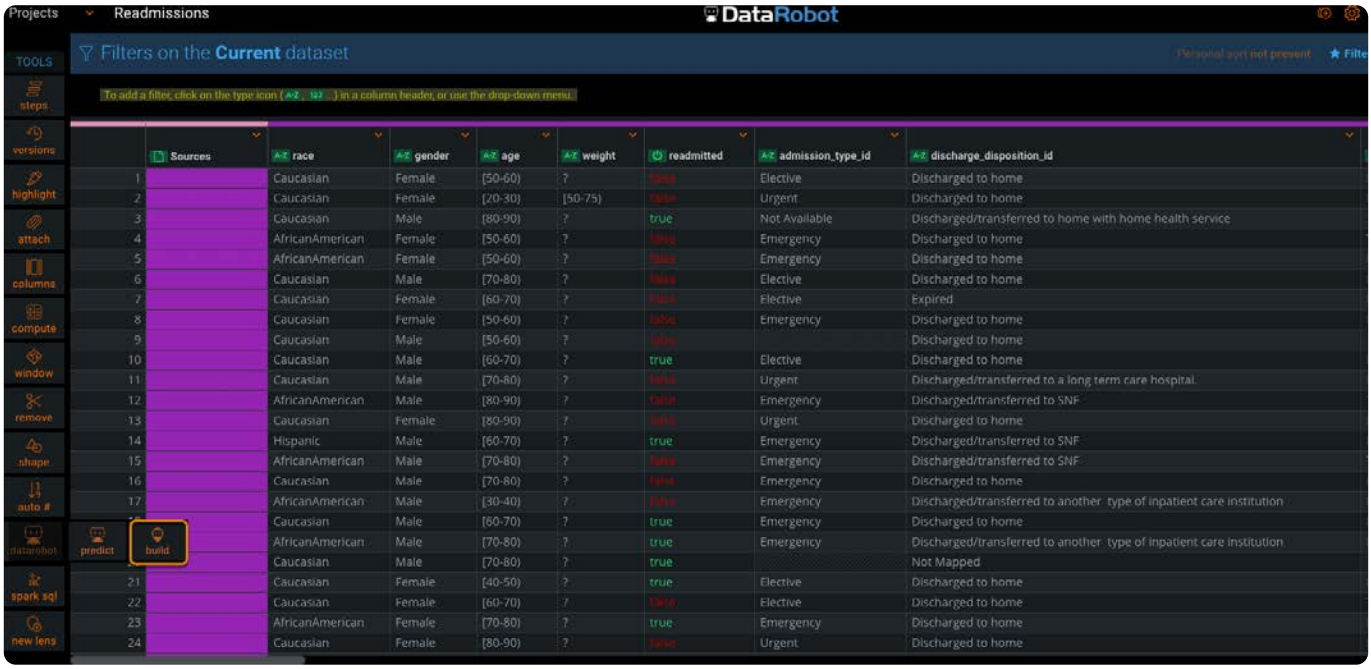
データの準備が完了した後、Data Prepから直接DataRobotプロジェクトを作成し、DataRobotでモデリングを開始することができます。これを行うには、Data Prep**構築**ツールを使用します。

> **備考**
>
> Data PrepアプリケーションからDataRobotプロジェクトを作成するには、DRレンズ機能を有効にする必要があります。**ツール**バーにDataRobot**構築**ツールが表示されない場合は、Data Prepシステム管理者にお問い合わせください。

## 構築ツールの操作

> **備考**
>
> DataRobotコネクターは、**構築**ツールを使用してDataRobotプロジェクトを作成するために設定する必要があります。DataRobotコネクターは、最新バージョンに更新する必要があります。

**構築**ツールにアクセスするには、**ツール**バーのDataRobotアイコンをクリックし、**構築**を選択します。

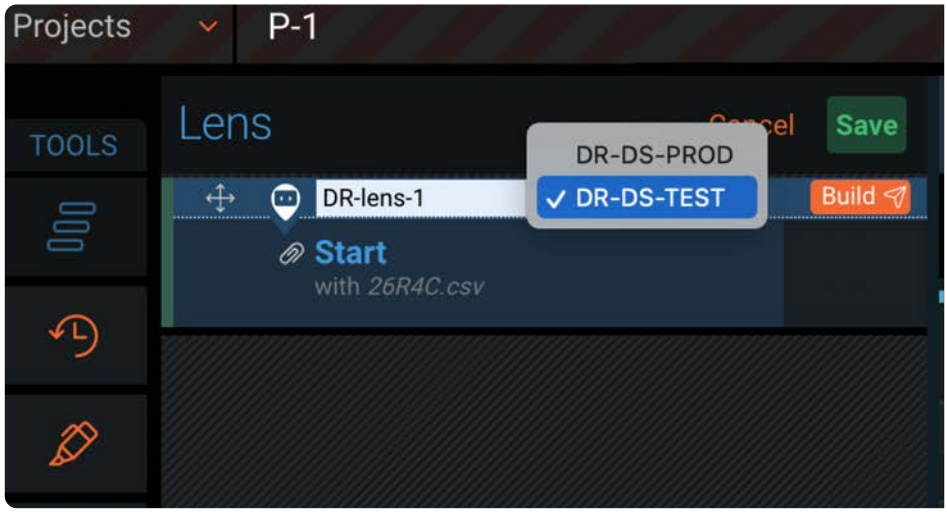


DataRobotプロジェクトを作成するには、最初に*DRレンズ*を作成します。DRレンズは標準レンズと同様にレンズ内のステップに基づきAnswerSetを生成しますが、DRレンズはAnswerSetに基づいてDataRobotプロジェクトを作成します。

> **備考**
>
> DRレンズから自動プロジェクトフロー（APF）を作成することはできません。APFを作成するには、プロジェクトに標準レンズが必要です。

# DRレンズからのDataRobotプロジェクトの作成

1. **ツール**バーのDataRobotアイコンをクリックし、**構築**を選択します。
2. DRレンズの名前を入力し、DataRobotコネクターを選択します。




> **備考**
>
> DataRobotコネクターが設定されていない場合は、コネクタードロップダウンリストは空です。管理者にDataRobotコネクターの設定を依頼してください。

3. **構築**をクリックします。DataRobotプロジェクトが作成中であることを示す通知がウィンドウの上部に表示されます。
4. DataRobotプロジェクトを作成すると、「ここをクリック」リンクを含む成功メッセージがウィンドウの上部に表示されます。DataRobotで作成された機械学習プロジェクトにアクセスするには、このリンクをクリックしてください。DataRobotプロジェクトには、指定したDRレンズ名に基づいて名前が設定されます。

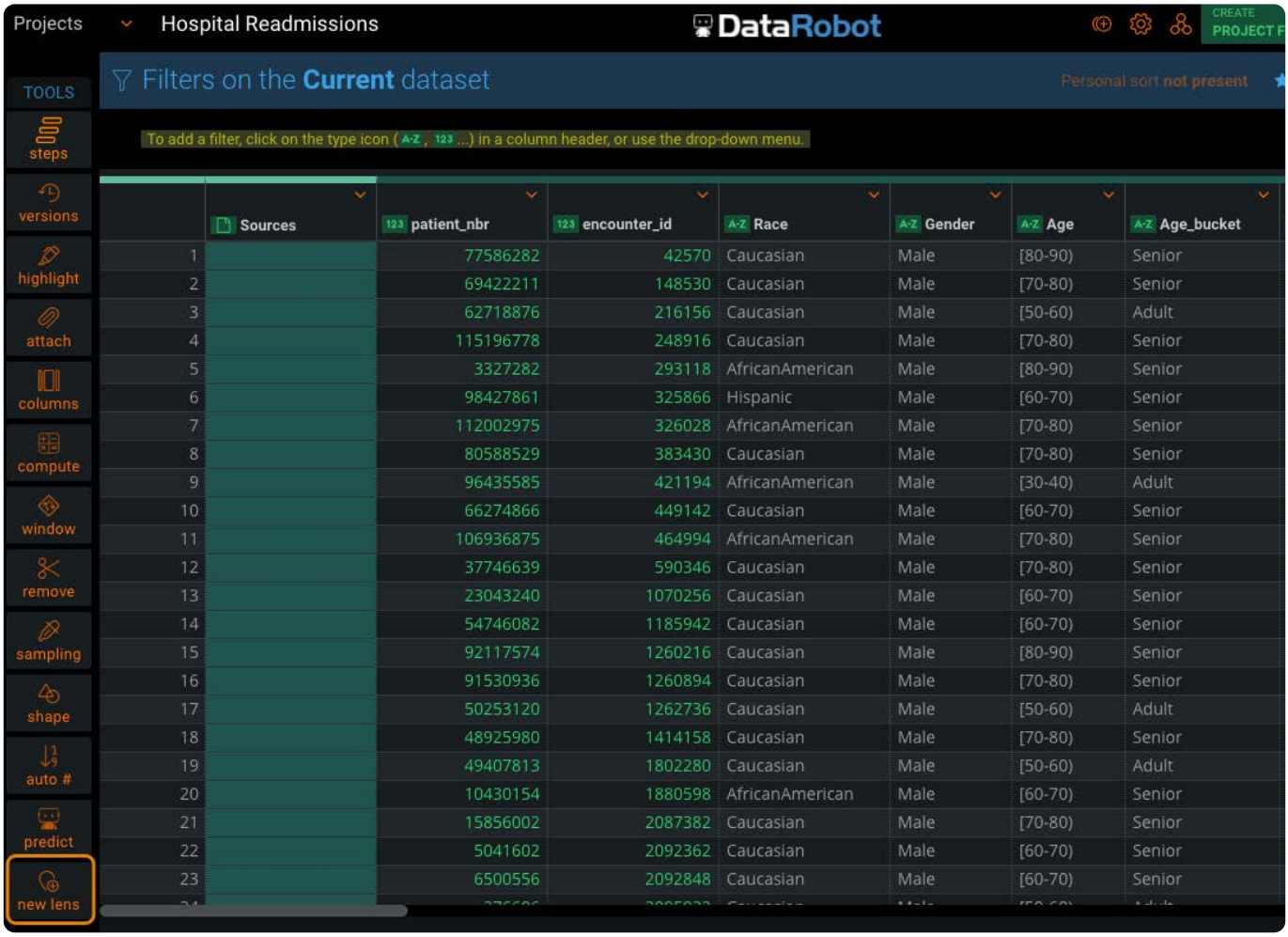
   DataRobotプロジェクトを作成できない場合は、エラーメッセージが表示されます。**ライブラリ > ログのエクスポート**を選択して、エクスポートログの詳細を確認してください。
5. **保存**をクリックすると、ステップツールでDRレンズステップが保存されます。

# 公開にレンズを使用する

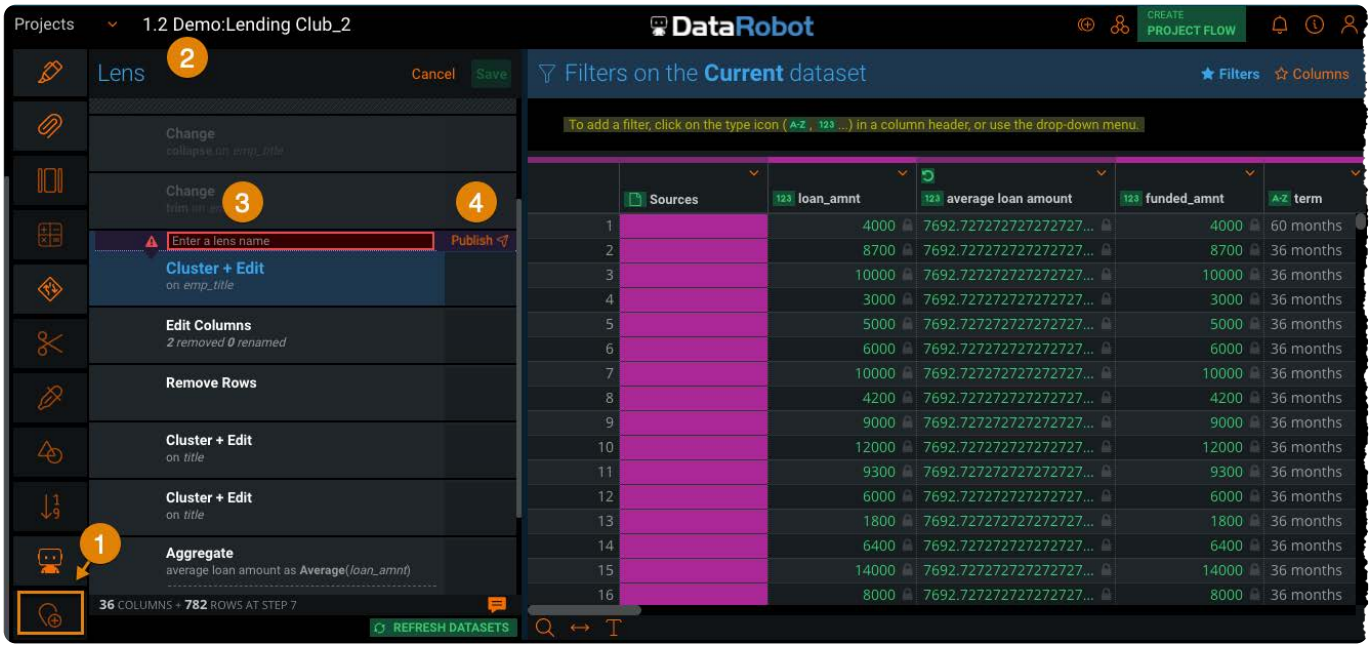
レンズを使用して、Data Prepプロジェクトのステップから公開ポイントを作成します。レンズから公開する場合、結果として得られるAnswerSetは、プロジェクトのこの特定ステップにおけるデータセットのスナップショットとなります。デフォルトでは、AnswerSetはデータライブラリに保存されます。

## 新しいレンズツールの操作

**新しいレンズ**ツールアクセスするには、プロジェクト**ツール**バーで**新しいレンズ**をクリックします。




プロジェクトから行を削除するときに操作する要素の概要を示します：




| | 要素 | 説明 |
|---|---|---|
| 1 | **新しいレンズ**ツール | **ツール**バーで**ステップ**をクリックし、ステップを選択します。次に、**新しいレンズ**をクリックして、**レンズ**ペインにアクセスします。 |
| 2 | **レンズ**ペイン | レンズを作成し、AnswerSetとして公開することができます。 |
| 3 | **レンズ名フィールドを入力します** | レンズの名前を入力し、**保存**をクリックします。 |
| 4 | **公開** | レンズを設定した後、**公開**をクリックして、データの状態をAnswerSetに保存します。 |

# レンズの追加

レンズを追加するには:

1. **ツール**バーの**ステップ**をクリックし、レンズを追加するステップをクリックします。
2. ツールバーの**新しいレンズ**をクリックします。
3. **レンズ**ペインに、一意のレンズ名を入力し、**保存**をクリックします。
4. オプションで、**公開**をクリックして、 Answersetに保存します。

# レンズの使用に関するヒント

- レンズは、プロジェクト内の任意のステップ、または任意のサブステップに追加できます（例: [追加] のインポートステップ）。
- 既存のレンズを任意のステップにドッグするか、既存のレンズを複数回追加できます。
- すべてのレンズは、プロジェクト ステップの一部として保持され、プロジェクトを共有するすべてのユーザーに公開されます。
- レンズ名は、結果の AnswerSet の名前付けに使用されるため、一意である必要があります。
- 計算ツールを使用して作成された数式エラーがある場合、ステープツールにエラーアイコン（）が表示されます。この場合、レンズを作成して保存できますが、AnswerSetに公開することはできません。

作成したレンズはプロジェクトのバージョンをまたいで保持されます。 古いバージョンのプロジェクトのレンズから AnswerSet を公開することもできます。

自動化ジョブで使用する公開ポイントはレンズによって定義されるため、レンズはプロジェクトの自動化にも不可欠です。プロジェクトで自動化を設定するときは、レンズを選択し、AnswerSet を自動的にデータ ライブラリに公開するためのスケジュールを構成します。したがって、プロジェクトを自動化するには、プロジェクト内に少なくとも 1つのレンズが必要です。 プロジェクトの自動化の詳細については、自動化と運用化を参照してください。

レンズを使用する場合の例を示します。

## 行をデータセットから分離する

レンズを使用して、詳細な調査が必要となるデータセットから行を分離する。使用するには、ステップにレンズを追加して、現在のデータセットから分離する行をフィルタリングします。レンズに名前を付け、**公開**をクリックします。結果の AnswerSet がデータライブラリに公開されます。この AnswerSet には分離された行のみが含まれ、後から詳しく調べることができます。この後、現在のデータセットからこれらの行を削除する新しいステップを生成できます。

## 集計の実行前と実行後のデータを比較

集計の実行前と実行後のデータを比較するには、データ シェーピングの前に現在のデータセットを公開するレンズを追加します。レンズに名前を付け、**公開**をクリックします。集計前のデータから成る AnswerSet がデータ ライブラリに公開されます。シェイプ ステップを生成し、結果のデータセットを公開するためのレンズを追加します。これで、集計の実行前と実行後それぞれのデータを表す 2 つの AnswerSet が生成されます。

## プロジェクトの自動化のスケジュールを設定する。

プロジェクトの自動化のスケジュールを設定するには、プロジェクト内の公開ポイントを作成しようとする全ステップにレンズを追加します。各レンズに、それぞれの公開ポイントから生成される出力結果を表す一意の名前を設定します。構成したスケジュールに基づき、レンズを使用して AnswerSet をデータライブラリに公開するための自動化を設定します。詳細については、自動化と運用化を参照してください。

# 計算された列関数

既存の列に関数を適用して、Data Prepプロジェクトに列を追加できます。次のページでは、計算ツールで使用する計算された列関数の構文と例が含まれています。

| トピック | 記述しています... |
| --- | --- |
| 日付/時刻関数 | 日付/時刻関数を既存の列に適用して新しい列を作成します。 |
| 情報関数 | 特定の値を確認するために列データを問い合わせる場合、たとえば、最初の空白以外 |
| の列の値を検索するか、空白とNull値を確認することができます。 | |
| 論理関数 | AND、OR、およびNOTなどの列値の論理関数を評価します。 |
| 数学関数 | 列値に計算関数を評価し、数値データ型として新しい列値を保存します。 |
| 統計関数 | 平均値、最大値、最小値、中央値、最頻値、および標準偏差 |
| などの列値の統計関数を評価します。 | |
| テキスト関数 | ASCII値から文字に変換するためのCHAR、文字列を組み合わせるためのCONCATENATE、文字列の内の文字列検索のためのFIND機能を使用してテキスト列で操作します。 |
| 比較演算子 | =、>、>、=および<>といった関数を使用して論理演算をテストします。 |
| カスタム列関数 | 組織で開発した列関数を列に適用します（オンプレミスでのインストールのみ）。 |

# 日付/時刻が計算された列関数

このセクションでは、Data Prep 計算ツールで使用できる日付/時刻が計算された列関数の構文および例を指定します。

日付/時刻の関数を使用するには、その datetime データ型として値を保存する必要があります。datetimeデータ型は、ヘッダー行のdatetimeデータ型アイコンで識別できます。値が datetime データ型として保存されない場合は、 `DATEVALUE` 関数を使用して値を datetime データ型に変換します。この記事の `DATEVALUE ()` セクションを参照してください。

## DATE

3つの別々の引数を取り、それらを組み合わせて、新しい日付/時刻列で日付を作成します。

**構文**

`DATE(YEAR, MONTH, DATE)`

- `YEAR` は 4 桁の値です。
- `MONTH` は 2 桁の値です。
- `DATE` は 2 桁の値です。

**例**

`DATE(@year@, @month@, @day@)`

| year | month | day | New Column |
|---|---|---|---|
| 1999 | 5 | 8 | 1999-05-08 |
| 1999 | 6 | 8 | 1999-06-08 |
| 1999 | 7 | 8 | 1999-07-08 |
| 1999 | 8 | 8 | 1999-08-08 |

**使用に関する注意**

`MONTH` と `DATE` のリーディングゼロはサポートされません。例:

`DATE(1999,05,08)` は `DATE(1999,5,8)` のように表される必要があります

## DATEADD

期日からの多くの日、週、月を計算します。

**構文**

`DATEADD(DATETIME, INCREMENT, INTERVAL)`

- `DATETIME` は、開始したい日付です。
- `INCREMENT` は、 `DATETIME` に追加するために加える数値です。
- `INTERVAL` は追加する間隔（分、日、年など）です。次の点は `INTERVAL` に認識された値のリストです:
  - 年
  - 月
  - 週
  - 日
  - 時間
  - 分
  - 秒 * ミリ秒

**例**

`DATEADD(@Date Received@, 6, "months")`

| Date Received | New Column |
|---|---|
| 2015-08-24T06:36:33.000Z | 2016-02-24T06:36:33.000Z |
| 2011-09-08T07:38:59.000Z | 2012-03-08T07:38:59.000Z |
| 2012-09-03T07:13:18.000Z | 2013-03-03T07:13:18.000Z |

**使用に関する注意**

提供する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。指定された `INCREMENT` は整数である必要があります。ミリ秒は最大 +/- 2147483647を受け入れます。

# DATEDIFF

2つの日付の間の日、週、月を計算します。

**構文**

`DATEDIFF(DATETIME_1, DATETIME_2, INTERVAL)`

- `DATETIME_1` は、開始したい日付です。
- `DATETIME_2` は終了したい日付です。
- `INTERVAL` は返したい間隔のタイプ（分、日、年など）です。次の点は `INTERVAL` 値に認識された値のリストです:
  - 年
  - 月
  - 週

- 日
- 時間
- 分
- 秒
- ミリ秒

**例**

`DATEDIFF(@Date Received@, @Date Shipped@, “months”)`

| Date Received | Date Shipped | New Column |
|---|---|---|
| 2015-08-24T06:36:33.000Z | 2016-02-24T06:36:33.000Z | 6 |
| 2011-09-08T07:38:59.000Z | 2012-04-08T07:38:59.000Z | 7 |
| 2012-09-03T07:13:18.000Z | 2013-04-03T07:13:18.000Z | 7 |

**使用に関する注意**

提供する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。指定された `INCREMENT` は整数である必要があります。ミリ秒は最大 +/- 2147483647を受け入れます。

`DATETIME_2` に最新の datetime 値を使用することをお勧めします。 `DATETIME_2` 値として最も早い日付を入力すると、 `DATEDIFF` 関数は負の数値を返します。

`DATEDIFF` は常に結果を最も近い整数に切り捨てます。例えば、2つの日付の差が 3 年と 11 ヶ月である場合、 `DATEDIFF` 関数は 3 年として返します。

# DATEFORMAT

datetime データ型として保存された値を、指定した形式でテキスト文字列に変換します。

**構文**

`DATEFORMAT(DATETIME, FORMAT)`

- `DATETIME` は変換したい日付です。
- `FORMAT` は `DATETIME` を変換したい形式です。

**例**

`DATEFORMAT(@Date Received@, “dd-MMM-yyyy HH:mm”)`

| Date Received | New Column |
| --- | --- |
| 2015-08-24T06:36:33.000Z | August 24, 2015 |
| 2011-09-08T07:38:59.000Z | September 08, 2011 |
| 2012-09-03T07:13:18.000Z | September 03, 2012 |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

# DATETRUNC

不要なタイムスタンプの詳細を削除し、望ましい間隔に四捨五入します。これは、SQL `DATE_TRUNC()` 関数と同じ出力を提供します。ユースケース: コミュニティユーザーのサインアップのトレンドを探索し、各イベントが発生した時刻別にサインアップのイベントデータを集計する必要があります。年、月、または日でのサインアップだけではなく、時間、分、ミリ秒にも興味があります。 `DATETRUNC` を使用して、不要なタイムスタンプの部分を削除します。

**構文**

`DATETRUNC(x)` では `x` が次の引数のいずれかとなります。

- 分
- 月
- 週
- 日
- 時間
- 秒

**例**

`DATETRUNC(@DATE@, "months")`

| DATE | New Column |
| --- | --- |
| 2019-01-10T00:00:00.000Z | 2019-01-01T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-01T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-01T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-01T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-01T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-01T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-01T00:00:00.000Z |

# DATEVALUE

datetime テキスト文字列を datetime オブジェクトに変換し、計算に使用できます。

**構文**

`DATEVALUE(DATETIME, FORMAT, TIME_ZONE)`

- `DATETIME` はテキスト文字列としての datetime です。
- `FORMAT` は `DATETIME` の形式です。
- `TIME_ZONE` は datetime オブジェクトに関連付けるタイムゾーンです。

**例**

`DATEVALUE(@Date@, "yyyy-MMM-dd hh:mm a", "GMT-05:00")`

**使用に関する注意**

`DATEVALUE` 関数を使用して、テキスト列を日付列、または入力した日付を日付オブジェクトに変換します。結果のデータオブジェクトにより、Data Prep 日付関数を使用できます。例えば、2つの日付の間の日数または年数を返します（日付操作については次の例を参照してください）。

日付オブジェクトは日付、時刻、または日付と時刻の組み合わせを保存できます。

**テキストを Data Prep 日付オブジェクトに変換するには:**Data Prep 日付形式構文の形式を指定します。繰り返し文字は、このフィールドの長さを示します。例えば、yyyy は 4 桁の年を意味します。

次のように 2012 年 2 月 28 日を指定した入力テキストがある `DateCol` 列:

`2012/28/02`

**日付オブジェクトに変換:**

`DATEVALUE(@DateCol@, "yyyy/dd/MM")`

日付形式は入力データと一致する必要があります。

**2012 年 2 月 28 日の表示が次の場合:**

`2012-15-02`

次の日付形式を使用:

`"yyyy-dd-MM"`

**2012 年 2 月 28 日の表示が次の場合:**

`2-28-12`

次の日付形式を使用:

`"dd-MM-yy"`

**時刻 1:29 pm の表記が次の場合:**

`13:29`

次の時刻形式を使用:

`"HH:mm"`

**時刻 1:29 pm の表記が次の場合:**

`01:29PM`

次の時刻形式を使用:

`"hh:mmaa"`

**高度な例**

入力テキストが文字 T とタイムゾーンで区切られた日付と時刻の場合:

`2012-02-28T09:29:00-05:00`

入力テキストに文字通り表示される文字の場合、直線の引用符で文字を囲みます。次の日付形式を使用:

`"yyyy-MM-dd'T'HH:mm:ssZZ"`

# 日付操作

`DATEDIFF` 関数を使用して、2つの Data Prep の日付オブジェクト間の datatime 値の差を計算します。1998 年 8 月 1 日と日付列間の日数を計算:

`DATEDIFF(DATEVALUE("01-AUG-1998" , "dd-MMM-yyyy"), @MyDate@ , "days")`

# DAY

日付から日を抽出します。

**構文**

`DAY(DATETIME)`

`DATETIME` は日を抽出したい日付です。

**例**

`DAY(@Date@)`

| Date | New Column |
| --- | --- |
| 2011-01-15T06:37:40.000Z | 15 |
| 2011-01-23T07:09:58.000Z | 23 |
| 2011-01-30T07:27:56.000Z | 30 |

**使用に関する注意**

指定する DATETIME は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

戻り値は 1 から 31 までに及びます。

# DAYOFWEEK

日付から 曜日を返します。

**構文**

DAYOFWEEK(DATETIME)

DATETIME は評価したい日付です。

**例**

DAYOFWEEK(@Date@)

| Date | New Column |
| --- | --- |
| 2011-01-15T06:37:40.000Z | 6 |
| 2011-01-23T07:09:58.000Z | 7 |
| 2011-01-30T07:27:56.000Z | 7 |

**使用に関する注意**

指定する DATETIME は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

戻り値は 1（月曜日）から 7（日曜日）までに及びます。

# DAYOFYEAR

日付から通日を返します。

**構文**

DAYOFYEAR(DATETIME)

`DATETIME` は評価したい日付です。

**例**

`DAYOFYEAR(@Date@)`

| Date | New Column |
|---|---|
| 2011-01-15T06:37:40.000Z | 15 |
| 2011-01-23T07:09:58.000Z | 23 |
| 2011-01-30T07:27:56.000Z | 30 |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

戻り値は 1 から 365 までに及びます（うるう年は 366）。

# ENDOFMONTH

新しい Datetime 列で月の最終日の datetime を返します。これは、Excel の `EOMONTH` 関数と同じ出力を提供します。

**構文**

`ENDOFMONTH(DATE_TIME)`

`DATE_TIME` は Datetime オブジェクトです。

**例**

`ENDOFMONTH(@Date@)`

| DATE | New Column |
|---|---|
| 2019-01-10T00:00:00.000Z | 2019-01-31T00:00:00.000 |
| 2019-01-10T00:00:00.000Z | 2019-01-31T00:00:00.000 |
| 2019-01-10T00:00:00.000Z | 2019-01-31T00:00:00.000 |
| 2019-01-10T00:00:00.000Z | 2019-01-31T00:00:00.000 |
| 2019-01-10T00:00:00.000Z | 2019-01-31T00:00:00.000 |
| 2019-01-10T00:00:00.000Z | 2019-01-31T00:00:00.000 |
| 2019-01-10T00:00:00.000Z | 2019-01-31T00:00:00.000 |

# FROMUNIXTIME

Unix タイムスタンプから datetime オブジェクトを返します。これは、MySQL `FROM_UNIXTIME()` 関数と同じ出力を提供します。

**構文**

`FROMUNIXTIME(MILLISECONDS)`

`MILLISECONDS` はミリ秒として表される int 値です。

**例**

`FROMUNIXTIME(@UNIX TIME STAMP@)`

| UNIX TIME STAMP | New Column |
|---|---|
| 831877766 | 1970-01-10T15:04:37.766Z |
| 834556166 | 1970-01-10T15:49:16.166Z |
| 837148166 | 1970-01-10T16:32:28.166Z |

# HOUR

時刻から時間を抽出します。

**構文**

`HOUR(DATETIME)`

`DATETIME` は時間を抽出したい時刻です。

**例**

`HOUR(@Date@)`

| Date | New Column |
|---|---|
| 2011-01-15T06:37:40.000Z | 6 |
| 2011-01-23T07:09:58.000Z | 7 |
| 2011-01-30T07:27:56.000Z | 7 |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

戻り値は、0（12:00 am）から 23（11:00 pm）までに及びます。

# MAXDATE

2 つ以上の日付を比較し、最新の日付を返します。

**構文**

`MAXDATE(DATETIME_1, [DATETIME_2, ...])`

- `DATETIME_1` は最初の日付です。
- `DATETIME_2` 、... [オプション]は追加の日付です。

**例**

`MAXDATE(@Target Ship Date@ ,@Date Shipped@)`

| Target Ship Date | Date Shipped | New Column |
|---|---|---|
| 2015-09-23T06:36:33.000Z | 2015-09-29T06:36:33.000Z | 2015-09-29T06:36... |
| 2011-10-08T07:38:59.000Z | 2011-10-16T07:38:59.000Z | 2011-10-16T07:38... |
| 2012-10-03T07:13:18.000Z | 2012-09-16T07:13:18.000Z | 2012-10-03T07:13... |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

いくつかの一般的なシナリオに対応する `MAXDATE` 関数については、次のようになります:

- 日付が 1 つだけ指定される場合、指定された日付が返されます。
- すべての日付のタイムゾーンは、同じタイムゾーンに一時的に変換され、最新の日付を決定します。変換は、永続的ではなく、視覚の変換でもありません。
- テキスト文字列を含むセルは無視されます。空白のセルは無視されます。
- エラーがあるセルは無視されます。
- datetime オブジェクトが見つからない場合、空白のセルが返されます。

# MIDNIGHT

指定された時刻を真夜（00:00）にリセットします。

**構文**

`MIDNIGHT(DATETIME)`

`DATETIME` はリセットしたい時刻です。

**例**

`MIDNIGHT(@Date@)`

| Date | New Column |
| --- | --- |
| 2011-01-15T06:37:40.000Z | 2011-01-15T00:00:00.000Z |
| 2011-01-23T07:09:58.000Z | 2011-01-23T00:00:00.000Z |
| 2011-01-30T07:27:56.000Z | 2011-01-30T00:00:00.000Z |

**使用に関する注意**

指定する DATETIME は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

タイムゾーンは影響されません。

# MINDATE

2 つ以上の日付を比較し、比較して最も早い日付を返します。

**構文**

MINDATE(DATETIME_1, [DATETIME_2, ...])

- DATETIME_1 は最初の日付です。
- DATETIME_2 、... [オプション]は追加の日付です。

**例**

MINDATE(@Target Ship Date@ ,@Date Shipped@)

| Target Ship Date | Ship Date | New Column |
| --- | --- | --- |
| 2015-09-23T06:36:33.000Z | 2015-09-29T06:36:33.000Z | 2015-09-23T06:36... |
| 2011-10-08T07:38:59.000Z | 2011-10-16T07:38:59.000Z | 2011-10-08T07:38... |
| 2012-10-03T07:13:18.000Z | 2012-09-16T07:13:18.000Z | 2012-09-16T07:13... |

**使用に関する注意**

指定する DATETIME は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

いくつかの一般的なシナリオに対応する MINDATE 関数については、次のようになります:

- 日付が 1 つだけ指定される場合、指定された日付が返されます。
- すべての日付のタイムゾーンは、同じタイムゾーンに一時的に変換され、最新の日付を決定します。変換は、永続的ではなく、視覚の変換でもありません。
- テキスト文字列を含むセルは無視されます。空白のセルは無視されます。
- エラーがあるセルは無視されます。

- datetime オブジェクトが見つからない場合、空白のセルが返されます。

# MINUTE

時間から分を抽出します。

**構文**

`MINUTE(DATETIME)`

DATETIME は、分を抽出したい時間です。

**例**

`MINUTE(@Date@)`

| Date | New Column |
|---|---|
| 2011-01-15T06:37:40.000Z | 37 |
| 2011-01-23T07:09:58.000Z | 9 |
| 2011-01-30T07:27:56.000Z | 27 |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

戻り値は、0 から 59 までに及びます。

# MONTH

日付から月を抽出します。

**構文**

`MONTH(DATETIME)`

`DATETIME` は月を抽出したい日付です。

**例**

`MONTH(@Date@)`

| Date | New Column |
| --- | --- |
| 2011-03-20T06:03:57.000Z | 3 |
| 2011-06-25T07:32:34.000Z | 6 |
| 2012-08-06T08:23:39.000Z | 8 |

**使用に関する注意**

指定する DATETIME は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

戻り値は、1 月から 12 月までに及びます。

# NETWORKDAYS

2 つの datetime オブジェクト間の営業日の日数を返します。これは、Excel の NETWORKDAYS 関数と同じ出力を提供します。

**構文**

NETWORKDAYS(DATE_TIME_START, DATE_TIME_END)

- DATE_TIME_START は開始日の datatime オブジェクトです。
- DATE_TIME は終了日付の datetime オブジェクトです。

**例**

NETWORKDAYS(@DATE@, DATE(2019,1,12))

| DATE | New Column |
| --- | --- |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |

# NOW

現在の日付と時刻を返します。

**構文**

NOW(TIME_ZONE) は現在の日付と時刻を返します。

オプションの TIME_ZONE はタイムゾーンを設定します。

**例**

`NOW("GMT-03:00")`

| New Column |
| --- |
| 2018-01-22 16:47:49.330 -03:00 |
| 2018-01-22 16:47:49.330 -03:00 |
| 2018-01-22 16:47:49.330 -03:00 |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

関数でタイムゾーンが指定されていない場合、戻り値オブジェクトはデフォルトでグリニッジ標準時（GMT）となります。タイムゾーンのリストとその適切な構文のリストについては、日付と時刻構文の記事を参照してください。

# QUARTER

指定された datetime オブジェクトから四半期を整数として返します。

**構文**

`QUARTER(DDATE_TIME)`

`DATE_TIME` は datetime オブジェクトです。

**例**

`QUARTER(@DATE@)`

| DATE | New Column |
| --- | --- |
| 2019-01-10T00:00:00.000Z | 1 |
| 2019-01-10T00:00:00.000Z | 1 |
| 2019-01-10T00:00:00.000Z | 1 |
| 2019-01-10T00:00:00.000Z | 1 |
| 2019-01-10T00:00:00.000Z | 1 |
| 2019-01-10T00:00:00.000Z | 1 |
| 2019-01-10T00:00:00.000Z | 1 |

# SECOND

時刻から秒を抽出します。

**構文**

`SECOND(DATETIME)`

`DATETIME` は秒を抽出したい時刻です。

**例**

`SECOND(@Date@)`

| Date | New Column |
|---|---|
| 2011-03-20T06:03:57.000Z | 57 |
| 2011-06-25T07:32:34.000Z | 34 |
| 2012-08-06T08:23:39.000Z | 39 |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

戻り値は、0 から 59 までに及びます。

# SETTIMEZONE

指定したタイムゾーンに時刻のタイムゾーンを変更します。

**構文**

`SETTIMEZONE(DATETIME, TIME_ZONE)`

- `DATETIME` はタイムゾーンを設定したい時刻です。
- `TIME_ZONE` は datetime オブジェクトに関連付けるタイムゾーンです。

**例**

`SETTIMEZONE(@Date Received@, "GMT-3:00")`

| Date Received | New Column |
|---|---|
| 2015-08-24T06:36:33.000Z | 2015-08-24T06:36:33.000-03:00 |
| 2011-09-08T07:38:59.000Z | 2011-09-08T07:38:59.000-03:00 |
| 2012-09-03T07:13:18.000Z | 2012-09-03T07:13:18.000-03:00 |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

変換は時刻を変更せず、既存の時刻に新しいタイムゾーンを割り当てます。タイムゾーンのリストとその適切な構文のリストについては、日付と時刻構文の記事を参照してください。

# TODAY

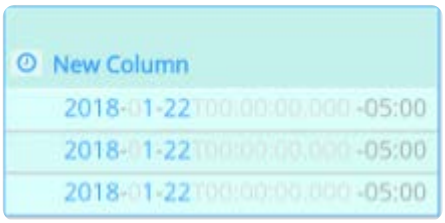
日付を返します。時刻は含まれません。

**構文**

`TODAY()`

**例**

`TODAY()`



| New Column |
| --- |
| 2018-01-22T00:00:00.000 -05:00 |
| 2018-01-22T00:00:00.000 -05:00 |
| 2018-01-22T00:00:00.000 -05:00 |

# WEEKOFYEAR

指定された datetime オブジェクトから週の数を整数として戻します。これは、Excel の `WEEKNUM` 関数と同じ出力を提供します。

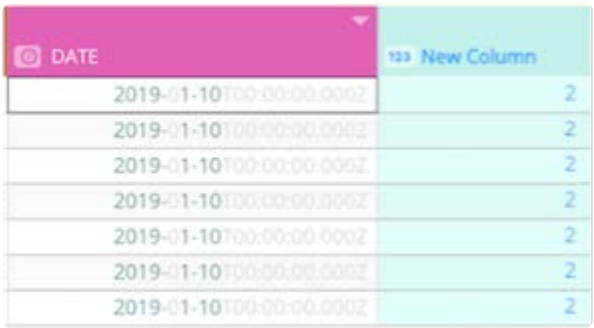
**構文**

`WEEKOFYEAR(DATE_TIME)`

`DATE_TIME` は datetime オブジェクトです。

**例**

`WEEKOFYEAR<@DATE@>`



| DATE | New Column |
| --- | --- |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |
| 2019-01-10T00:00:00.000Z | 2 |

# WORKDAY

日付（開始日）前後の営業日として指定される日付の日数を返します。営業日は、週末と休日として指定される日付を除外します。これは、Excel の `WORKDAY` 関数と同じ出力を提供します。請求書の期日、配達予定日、または稼働日の日数を計算するとき、週末または休日を除外するために `WORKDAY` を使用します。

**構文**

`WORKDAY(STARTDATE, DAYS)`

- `STARTDATE` は開始日付を表す日付です。
- `DAYS` は開始日前後の週末以外および休日以外の日数です。日の正の値は将来の日付を生成します。負の値は過去の日付を生成します。

**例**

`WORKDAY(@DATE@ ,12)`

| DATE | New Column |
|---|---|
| 2019-01-10T00:00:00.000Z | 2019-01-28T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-28T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-28T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-28T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-28T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-28T00:00:00.000Z |
| 2019-01-10T00:00:00.000Z | 2019-01-28T00:00:00.000Z |

# YEAR

日付から年を抽出します。

**構文**

`YEAR(DATETIME)`

`DATETIME` は年を抽出したい日付です。

**例**

`Year()`

| Date | New Column |
|---|---|
| 2012-07-20T07:15:24.000Z | 2012 |
| 2016-03-25T07:06:44.000Z | 2016 |
| 2014-01-30T07:16:25.000Z | 2014 |

**使用に関する注意**

指定する `DATETIME` は datetime オブジェクト、datetime オブジェクトを含む列、または datetime オブジェクトを返す関数である必要があります。

# 情報に関する計算された列関数

このセクションでは、Data Prep計算ツールで使用できる情報が計算された列関数の構文および例を指定します。情報に関する関数では、列値を列に問い合わせ、結果に基づいて新しい列を作成できます。

## FIRSTNONBLANK

2つ以上の列の値を比較し、最初の空白以外の値を返します。この関数は、ExcelのFIRSTNONBLANK関数と同じ出力を提供します。

**構文**

`FIRSTNONBLANK(ARGUMENT_1, [ARGUMENT_2, ...])]()`

- `ARGUMENT_1` は最初の列です。
- `ARGUMENT_2` 、...[オプション]は追加の列です。

**例**

`FIRSTNONBLANK(@Current Employer@, @Previous Employer@, @School@)`

| Current Employer | Previous Employer | School | New Column |
|---|---|---|---|
| Banana Inc | Mermaidhut | Camden College | Banana Inc |
| CloudCo | | | CloudCo |
| | | Greendale Community College | Greendale Community College |
| | Bansheelectronis | University of New York | Bansheelectronis |
| | Pyramidustries | | Pyramidustries |
| | | | |

**使用に関する注意**

1つの列だけが指定された場合、提供された列の値が返されます。

空白以外の値が見つからない場合、 `FIRSTNONBLANK` 関数は空のセル返されます（空白以外の値が見つからない場合、結果に表示する値の最後の引数を含めていない場合）

## ISBLANK

指定された列内の空白またはNull値を確認します。空白またはnull値が見つかった場合、 `TRUE` 値が返されます。

**構文**

`ISBLANK(ARGUMENT)`

`ARGUMENT` は確認する列です。

**例**

`ISBLANK(@Column@)`

| Column | New Column |
|---|---|
| Rufus Daniel | false |
| | true |
| | true |
| 1789 | false |
| Bryant Carr | false |

# ISDATE

**構文**

`ISDATE(ARGUMENT)`

`ARGUMENT` は確認する列です。

**例**

`ISDATE(@Column@)`

| Column | New Column |
|---|---|
| 2013-01-02T00:01:... | true |
| | false |
| 3/6/2014 | false |

**使用に関する注意**

値はdatetimオブジェクトであり、datetime文字列にしない必要があります。Excelスプレッドシートからインポートされたデータセットにより、datetimeオブジェクトとして日付が自動的にインポートされます。その他のソースからの日付は、`DATEVALUE` 関数を使用してdatetimeオブジェクトに変換する必要があります。この記事の `DATEVALUE()` セクションを参照してください。

# ISNULL

指定列内の空白またはnull値を確認します。空白またはnull値が見つかった場合、 `TRUE` 値が返されます。

**構文**

`ISNULL(ARGUMENT)`

`ARGUMENT` は確認する列です。

**例**

`ISNULL(@Column@)`

| Column | New Column |
|---|---|
| Bryant Carr | false |
| | true |
| 2013-01-02T00:01:00.000Z | false |
| Kelli Martinez | false |
| Rufus Daniel | false |
| | true |

# ISNUMBER

指定列内の数値を確認します。数値が見つかった場合、TRUEの値が返されます。

**構文**

`ISNUMBER(ARGUMENT)`

`ARGUMENT` は確認する列です。

**例**

`ISNUMBER(@Column@)`

| Column | New Column |
|---|---|
| 3/6/2014 | false |
| 2013-01-02T00:01:... | false |
| 6 | true |
| 7 | true |

# ISTEXT

指定列内のテキストを確認します。空白またはnullが見つかった場合、 `TRUE` 値が返されます。

**構文**

`ISTEXT(ARGUMENT)`

`ARGUMENT` は確認する列です。

**例**

ISTEXT(@Column@)

| Column | New Column |
| --- | --- |
| 2013-01-02T00:01:... | false |
| Rufus Daniel | true |
| Kelli Martinez | true |
| | false |
| 3/6/2014 | true |

# 論理的に計算された列関数

このセクションでは、Data Prep計算ツールで使用できる論理的に計算された列関数の構文と例を示します。論理関数を使用して列値の論理的関数を評価します。新しい列には、関数結果に応じて、 TRUE または FALSE が含まれています。

## AND

式内のすべての引数が TRUE を評価するかどうか評価します。引数が TRUE を評価する場合、値 TRUE が返されます。

**構文**

AND(ARGUMENT_1, [ARGUMENT_2, ...])

- ARGUMENT_1 は評価する引数です。
- ARGUMENT_2 、...[オプション]は追加の引数です。

**例**

AND(@Column_A@, @Column_B@, @Column_C@)

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| false | true | true | false |
| true | true | true | true |
| false | true | false | false |
| false | false | false | false |

**使用に関する注意**

提供する ARGUMENT は、 TRUE または FALSE 値、値を含む列、または値を返す関数である必要があります。

AND 関数は大文字と小文字を区別しないので、 True 、 TRUE 、および true を同じ方法で処理します。同様に、 False 、 FALSE および false は同じ方法で処理されます。

## IF

与えられたステートメントがTRUEどうかどうかによって、異なる結果を指定できます。

**構文**

`IF(CONDITION, TRUE_VALUE, FALSE_VALUE)`

- `CONDITION` は評価したい式です。
- `CONDITION` がTRUEの場合、 `TRUE_VALUE` は関数が返す値です。
- `CONDITION` がTRUEではない場合、 `FALSE_VALUE` は返される値です。

**例**

`IF(@Current Employer@ = 0, "N/A", @Current Employer@)`

| Current Employer | New Column |
|---|---|
| 0 | N/A |
| Banana Computers Inc | Banana Computers Inc |
| 0 | N/A |
| Banana | Banana |
| CloudCo inc | CloudCo inc |
| Microstuff | Microstuff |
| 0 | N/A |
| 0 | N/A |
| Acme | Acme |
| 0 | N/A |

**使用に関する注意**

`IF` 関数は、1またはその他の列の情報に基づいて値の集を作成することができます。

`CONDITION` は、 `TRUE` または `FALSE` 値のいずれかを指定する必要があります。その他の関数は、 `CONDITION` の一部として組み込むことができます。別の `IF` 関数は、値の1つまたは両方として使用できます。これにより、返された値を非常にきめ細かく管理できます。ほとんどの場合、 `CONDITION` は演算子を含みます。この記事の比較演算子（）セクションを参照してください。

# IFERROR

与えられたステートメントがTRUEであるかどうかどうかによって、別の結果を指定できます。

**構文**

`IFERROR(ARGUMENT, VALUE)`

- `ARGUMENT` は確認したい列です。
- 列のセルがエラーを含む場合、 `VALUE` は返す値です。

**例**

`IFERROR(@New Column@, "N/A" )`

| New Column | New Column (1) |
| --- | --- |
| | N/A |
| Banana Computers Inc | Banana Computers Inc |
| | N/A |
| Banana | Banana |
| CloudCo inc | CloudCo inc |
| Microstuff | Microstuff |
| | N/A |

**使用に関する注意**

指定する VALUEs はテキスト文字列または数値、テキスト文字列または数値を含む列、またはテキスト文字列または数値を返す関数です。

エラーがない場合、セルの元の値が返されます。

# NOT

TRUE または FALSE 値となる式の結果を逆転させます。

**構文**

NOT(ARGUMENT)

ARGUMENT は、逆転させる TRUE または FALSE 値です。

**例**

NOT(@Column@)

| Column | New Column |
| --- | --- |
| false | true |
| true | false |

**使用に関する注意**

提供する ARGUMENT は、 TRUE または FALSE 値、値を含む列、また値を返す関数である必要があります。

NOT 関数は大文字と小文字を区別しないので、 True 、 TRUE 、および true を同じ方法で処理します。同様に、 False 、 FALSE および false は同じ方法で処理されます。

# OR

式内の少なくとも1つの値が TRUE かどうか判断します。1つの値が TRUE である場合、 TRUE 値が返されます。

**構文**

`OR(ARGUMENT_1, [ARGUMENT_2, ...])`

- `ARGUMENT_1` は評価したい最初の引数です。
- `ARGUMENT_2` 、...[オプション]は追加の列です。

**例**

`OR(@Column_A@, @Column_B@, @Column_C@)`

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| false | true | true | true |
| true | true | true | true |
| false | true | false | true |
| false | false | false | false |

**使用に関する注意**

提供する `ARGUMENT` は、 `TRUE` または `FALSE` 値、値を含む列、または値を返す関数である必要があります。

`OR` 関数は大文字と小文字を区別しないので、 `True` 、 `TRUE` 、および `true` を同じ方法で処理します。同様に、 `False` 、 `FALSE` および `false` は同じ方法で処理されます。

# 数学的に計算された列関数

このセクションでは、Data Prep計算ツールで使用できる数学的に計算された列関数の構文および例を指定します。統計関数については、統計的な列関数を参照してください。

数値関数を使用するには、値を数値データ型として保存する必要があります。数値データ型は、ヘッダー行の数値データ型 123 アイコンによって識別されます。数値データ型として保存されていない場合は、VALUE関数を使って数値な形式に変換します。

このセクションに一覧されている数学関数に加えて、以下の標準的な数学演算に対応しています:

- 列を任意の数で乗算または除算します。
- 列に任意の数を加算または除算します。

列名「Revenue」の使用例は以下の通りです。

`@Revenue@ * 100`

`@Revenue@ / 100`

`@Revenue@ + 100`

`@Revenue@ - 100`

## ABS

実数のAbsolute Value (ABS)を返します。

数学式では、Absolute Valueは両側のいずれかのバーで表示されます。たとえば、xのAbsolute Valueは|x|として書き込まれます。

**構文**

`ABS(VALUE)`

`VALUE` は、Absolute Valueを求めたい値です。

**例**

`ABS(@Column@)`

| Column | New Column |
|---|---|
| 6 | 6 |
| 14 | 14 |
| -10 | 10 |
| -11 | 11 |

**使用に関する注意**

提供する VALUE は、実数、実数を含む列、または実数を返す関数でなければなりません。

ABS は、指定した数値が数直線に含む、ゼロからの距離と考えることができます。 ABS の場合、正と負は重要ではありません。ゼロの右側（正）またはゼロの左側（負）であるかどうかに関わらず、数値のゼロからの距離は同じ、または絶対です。数学式では、絶対値は両側のいずれかのバーで表示されます。たとえば、xのAbsolute Valueは|x|として書き込まれます。

# CEILING

与えられた数値を整数に切り上げて返します。

**構文**

CEILING(VALUE)

VALUE は、切り上げたい値です。

**例**

CEILING(@Column A value@)

# EXP

指定された値の指数を返します。

**構文**

EXP(NUMBER)

NUMBER は任意の実数です。

**例**

EXP(@Column A value@)

| Column A value | New Column |
|---|---|
| 6.588 | 726.3267627508812 |
| 9.43 | 12456.526731608414 |
| 11.345 | 84541.68455061226 |
| 14.796 | 2665760.6580085587 |
| 20 | 485165195.4097903 |

# FACTORIAL

整数とその下のすべての整数の積を返します。

**構文**

FACTORIAL(NUMBER)

NUMBER は任意の実数です。

**例**

FACTORIAL(@Column A value@)

| Column A value | New Column |
|---|---|
| 6.588 | 1464.46139524329530184499 2 |
| 9.43 | 468217.34343705300007777749 |
| 11.345 | 36056362.172297317939336454456826 99 |
| 14.796 | 640499653925.3259018035792303107524 |
| 20 | 2.43290200817664E+18 |

# FLOOR

与えられた数値を切り捨てて整数にして返します。

**構文**

FLOOR(VALUE)

VALUE は、切り上げたい値です。

**例**

FLOOR(@Column A value@)

| Column A value | New Column |
|---|---|
| 6.588 | 6 |
| 9.43 | 9 |
| 11.345 | 11 |
| 14.796 | 14 |
| 20 | 20 |

# int

実数を、それよりLess ThanかEqual To次の整数に切り捨てます。

**構文**

`INT(VALUE)`

VALUEは、切り捨てたい実数です。

**例**

`INT(@Column@)`

| Column | New Column |
|---|---|
| 6.1 | 6 |
| 14.11 | 14 |
| -9.88 | -10 |
| -10.88 | -11 |

**使用に関する注意**

提供する `VALUE` は、実数、実数を含む列、または実数を返す関数でなければなりません。

# LN

数値の自然対数を返します。自然対数は、定数e(2.71828182845904)に基づいています。これは、Excelの `LN` 関数と同じ出力を提供します。

**構文**

`LN(NUMBER)`

`NUMBER` は、自然対数を求める正の実数です。

**例**

`LN(@Column A value@)`

| Column A value | New Column |
|---|---|
| 6.588 | 1.8852498123153938 |
| 9.43 | 2.2438960966453663 |
| 11.345 | 2.428777118231805 |
| 14.796 | 2.6943568739702077 |
| 20 | 2.995732273553991 |

# LOG

指定したベースに数値の対数を返します。この関数は、Excelの LOG 関数と同じ出力を提供します。

**構文**

LOG(NUMBER,BASE)

- NUMBER は、自然対数を求める正の実数です。
- BASE は、対数の基数です。

**例**

LOG(@Column A value@,2)

| Column A value | New Column |
|---|---|
| 6.588 | 2.719840555064268 |
| 9.43 | 3.237257770900372 |
| 11.345 | 3.5039847038976126 |
| 14.796 | 3.8871353004619085 |
| 20 | 4.321928094887363 |

# LOG10

数値のベース10対数を返します。この関数は、Excelの LOG10 関数と同じ出力を提供します。

**構文**

LOG10(NUMBER)

NUMBER は、自然対数を求める正の実数です。

**例**

LOG10(@Column A value@)

| Column A value | New Column |
|---|---|
| 6.588 | 0.8187535904977168 |
| 9.43 | 0.9745116927373284 |
| 11.345 | 1.0548045002209547 |
| 14.796 | 1.1701443226433565 |
| 20 | 1.3010299956639813 |

# MOD

数値を除数で割った後の余りを返します。結果は、除数と同じSignです。これは、Excelの `MOD` 関数と同じ出力を提供します。

**構文**

`MOD(MUMBER,DIVISOR)`

- `NUMBER` は任意の実数です。
- `DIVISOR` は任意の実数です。

**例**

`MOD(@Column A value@,3)`

| Column A value | New Column |
|---|---|
| 6.588 | 0.588 |
| 9.43 | 0.43 |
| 11.345 | 2.345 |
| 14.796 | 2.796 |
| 20 | 2 |

# POWER

数値の累乗の結果を計算して返す数学/三角関数です。この関数は、Excelの `POWER` 関数と同じ出力を提供します。

**構文**

`POWER(NUMBER, POWER)`

- `NUMBER` は、任意の実数である基数です。
- `POWER` は指数で、任意の実数であり、基数を上昇させます。

**例**

`POWER(@Column A value@,3)`

| Column A value | New Column |
|---|---|
| 6.588 | 285.930689472 |
| 9.43 | 838.5618069999999 |
| 11.345 | 1460.2038886250002 |
| 14.796 | 3239.1642303359995 |
| 20 | 8000 |

# ROUND

数値を、指定した小数点以下の桁数に丸めることができます。

**構文**

`ROUND(VALUE, PLACES)`

- `VALUE` は、丸めたい実数です。
- `PLACES` は、丸める小数点以下の桁数です。

**例**

`ROUND(@Column@, 2)`

| Column | New Column |
|---|---|
| 6.314159265358979323846264 3 | 6.31 |
| 14.345575191894877256230890 73 | 14.35 |
| -9.623008881569224811384482 84 | -9.62 |
| -10.623008881569224811384482 84 | -10.62 |

**使用に関する注意**

提供する `VALUE` は、実数、実数を含む列、または実数を返す関数でなければなりません。

`PLACES` の値は正の整数でなければなりません（負の数や含まれる10進数は不可です）。小数点以下の値は、 `ROUND` 関数では影響を受けません。

# ROUNDDOWN

数値を指定した小数点以下の桁数に切り捨てます。

**構文**

`ROUNDDOWN(VALUE, PLACES)`

- `VALUE` は、切り捨てたい実数です。
- `PLACES` は、切り捨てを行う小数点以下の桁数です。

**例**

`ROUNDDOWN(@Column@, 3)`

| Column | New Column |
|---|---|
| 6.31415926535897932384626433 | 6.314 |
| 14.34557519189487725623089073 | 14.345 |
| -9.62300888156922481138448284 | -9.623 |
| -10.62300888156922481138448284 | -10.623 |

**使用に関する注意**

`ROUNDDOWN` は、常に切り捨てられることを除けば、 `ROUND` と同様です。

提供する `VALUE` は、実数、実数を含む列、または実数を返す関数でなければなりません。

`PLACES` の値は、正の整数でなければなりません（負の数や小数を含んではいけません）。小数点以下の値は、 `ROUNDDOWN` 関数を使っても影響を受けません。小数点以下の桁数を指定した場合、 `ROUNDDOWN` は `INT` 関数と同様に動作します。

# ROUNDPERC

実際のパーセントの値（-100から100の間）を指定した小数点以下の桁数に丸めます。0%および±100%に近い値は、指定した小数点以下の桁数以上になり、Trueの0%および±100%の値と区別するための小数点以下の桁数はそれ以上にはなりません。

**構文**

`ROUNDPERC(VALUE, PLACES)`

- `VALUE` は、丸めたい実際のパーセントの値です。
- `PLACES` は、丸めるMinimum（最小）の小数点以下の桁数です。

**例**

`ROUNDPERC(@Column@, 2)`

| Column | New Column |
|---|---|
| 6.31415926535897932384626433 | 6.31 |
| 14.34557519189487725623089073 | 14.35 |
| -9.62300888156922481138448284 | -9.62 |
| -10.62300888156922481138448284 | -10.62 |

**使用に関する注意**

提供する `VALUE` は、-100から100までの実際のパーセントの値、実際のパーセントの値を含む列、または実際のパーセントの値を返す関数でなければなりません。

PLACES の値は、整数（小数不可）でなければなりません。小数点以下の値は、 ROUNDPERC 関数を使用しても影響を受けません。

# ROUNDUP

数値を指定した小数点以下の桁数に切り上げます。

**構文**

ROUNDUP(VALUE, PLACES)

- VALUE は、切り上げたい実数です。
- PLACES は、切り上げを行う小数点以下の桁数です。

**例**

ROUNDUP(@Column@, 3)

| Column | New Column |
|---|---|
| 6.314159265358979323846264 3 | 6.315 |
| 14.345575191894877256230890 73 | 14.346 |
| -9.623008881569224811384482 84 | -9.624 |
| -10.623008881569224811384482 84 | -10.624 |

**使用に関する注意**

ROUNDUP は、常に切り上げを行うことを除いて、 ROUND と同様です。

提供する VALUE は、実数、実数を含む列、または実数を返す関数でなければなりません。

PLACES の値は正の整数でなければなりません（負の数や含まれる10進数は不可です）。小数点以下の値は、 ROUNDUP 関数を使用しても影響を受けません。

# SIGN

数値のSignを決定します。 数値が正の場合は1を、0の場合はゼロ（0）を、負の場合は-1を返します。これは、Excelの SIGN 関数と同じ出力を提供します。

**構文**

SIGN(NUMBER)

NUMBER は任意の実数です。

**例**

`SIGN(@Column A value@)`

| Column A value | New Column |
|---|---|
| -4.44 | -1 |
| 9.43 | 1 |
| -11.345 | -1 |
| 14.796 | 1 |
| -20 | -1 |

## SQRT

正の平方根を返します。この関数は、Excelの `SQRT` 関数と同じ出力を提供します。

# 構文

`SQRT(NUMBER)`

`NUMBER` は、平方根を計算したい任意の正の数です。

**例**

`SQRT(@Column A value@)`

| Column A value | New Column |
|---|---|
| 6.588 | 2.5667099563448925 |
| 9.43 | 3.0708305065568173 |
| 11.345 | 3.368233958619858 |
| 14.796 | 3.8465569019579053 |
| 20 | 4.47213595499958 |

## SUM

与えられた数値を加算します。

**構文**

`SUM(VALUE_1, [VALUE_2, ...])`

- `VALUE_1` は最初の値です。
- `VALUE_2` 、...[オプション]は追加の値です。

**例**

`SUM(@Column_A@, @Column_B@, @Column_C@)`

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| 6 | 14 | 36 | 56 |
| 6 | 17 | 36 | 59 |
| 7 | 11 | 24 | 42 |
| 6 | 13 | 24 | 43 |

**使用に関する注意**

提供する VALUE は、数値、数値を含む列、または数値を返す関数でなければなりません。

# 統計的に計算された列関数

このセクションでは、Data Prep計算ツールで使用できる統計的に計算された列関数の構文と例を示します。 統計機能を使用するには、値を数値データ型として保存する必要があります。数値データ型は、ヘッダー行の数値データ型123アイコンによって識別されます。数値データ型として保存されていない場合は、VALUE関数を使って数値形式に変換します。

## AVERAGE

リストのアイテム数で割った数値のリストの合計と等しい値を計算します。

**構文**

`AVERAGE(VALUE_1, [VALUE_2, ...])`

- `VALUE_1` は最初の値です。
- `VALUE_2` 、...[オプション]は追加の値です。

**例**

`AVERAGE(@Column_A@, @Column_B@, @Column_C@)`

| 123 Column_A | 123 Column_B | 123 Column_C | 123 New Column |
|---|---|---|---|
| 5 | 15 | 10 | 10 |
| 7 | 14 | | 10.5 |
| 7 | 21 | 5 | 11 |
| 5 | 10 | 11 | 8.6666666666666... |

**使用に関する注意**

提供する `VALUE` は、数値、数値を含む列、または数値を返す関数でなければなりません。

## Max

値の集合から最大（maximum）値を返します。

**構文**

`MAX(VALUE_1, [VALUE_2, ...])`

- `VALUE_1` は最初の値です。
- `VALUE_2``, ...` [オプション]は追加の値です。

**例**

`MAX(@Column_A@, @Column_B@, @Column_C@)`

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| 5 | 17 | 7 | 17 |
| 6 | 17 | | 17 |
| 6 | 10 | 4 | 10 |
| 7 | 12 | 13 | 13 |

**使用に関する注意**

提供する `VALUE` は、数値、数値を含む列、または数値を返す関数でなければなりません。

# Median

最小値から最高値まで並べられた数値の範囲の中央に存在する数値を返します。

**構文**

`MEDIAN(VALUE_1, [VALUE_2, ...])`

- `VALUE_1` は最初の値です。
- `VALUE_2, ...` [オプション]は追加の値です。

**例**

`MEDIAN(@colum_A@, @colum_B@, @colum_C@, @colum_C@, @colum_E@)`

| Colum A | Column B | Column C | Column D | Column E | New Column |
|---|---|---|---|---|---|
| 3 | 4 | 5 | 2 | 10 | 4 |
| 10 | 4 | 1 | 0 | 0 | 1 |
| 4 | 2 | 0 | 7 | 0 | 2 |
| 6 | 1 | 0 | 4 | 4 | 4 |
| 9 | 3 | 0 | 4 | 3 | 3 |
| 4 | 4 | 7 | 4 | 6 | 4 |
| 2 | 3 | 8 | 9 | 10 | 8 |
| 2 | 3 | 10 | 3 | 1 | 3 |

**使用に関する注意**

提供する `VALUE` は、数値、数値を含む列、または数値を返す関数でなければなりません。

偶数の数値の集合を含む範囲で、中央値は中央の数値です。半分の数値は返される値の右側に、もう半分は返される左側にあります。範囲の中央に単一の数値がない場合）、中央値は中点のいずれかの側にある2つの数値の平均を計算します。

> **備考**
>
> 中央値は平均値とは異なります。平均値は算術平均であり、数値の集合を合計し、集合内の値の数で除して計算されます。中央値は、範囲の中央にある値を取ります。コレクション内の値の分布でバランスを示す数値範囲では、中央値と平均値の計算が偶然に一致することがあります。ひずみのある分布では、値は異なります。

# Min

値の集合から最小（minimum）の値を返します。

**構文**

`MIN(VALUE_1, [VALUE_2, ...])`

- `VALUE_1` は最初の値です。
- `VALUE_2, ...` [オプション]は追加の値です。

**例**

`MIN(@Column_A@, @Column_B@, @Column_C@)`

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| 5 | 15 | 10 | 5 |
| 7 | 14 | | 7 |
| 7 | 21 | 5 | 5 |
| 5 | 10 | 11 | 5 |

**使用に関する注意**

提供する VALUE は、数値、数値を含む列、または数値を返す関数でなければなりません。

# モード

数値の集合で最も頻繁に発生する値を返します。

**構文**

MODE(VALUE_1, VALUE_2, [VALUE_3, ...])

- VALUE_1 は最初の値です。
- VALUE_2 は2番目の値です。
- VALUE_3, ... [オプション]は追加の値です。

**例**

MODE(@Column_A@, @Column_B@, @Column_C@)

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| 3 | 9 | 9 | 9 |
| 8 | 13 | 12 | 8 |
| 7 | 11 | 11 | 11 |
| 6 | 19 | 11 | 6 |

**使用に関する注意**

提供する**値**は、数値、数値を含む列、または数値を返す関数でなければなりません。

複数の数値が1より大きい同等の発生回数を持つ場合、返される値は、集合内で最初に表示される（左から右に読んで）数値（頻度が同等であるもののうち）です。1回を上回る数値が一つも表示されない場合、関数はエラーを返します。

MODE に関連する最も一般的な問題は、提供された数値集合に重複がない場合です。関数が正常に数値を求めるには、少なくとも1つの数値が2回以上表示される必要があります。最小数の引数（2）を使用する場合、各引数が求める数は同じ数値になる必要があり、そうでなければエラーが発生します。予想される通り、より変動に制限のある、より大きな数値集合は、 MODE エラーを返す可能性が小さくなります。

# STDEV

データのサンプル集合に含まれる値の標準偏差（平均からの変動の度合い）を推定します。

**構文**

STDEV(VALUE_1, [VALUE_2, ...])

- VALUE_1 は最初の値です。
- VALUE_2, ... [オプション]は追加の値です。

**例**

STDEV(@Column_A@, @Column_B@, @Column_C@)

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| 0.66 | 0.28 | 0.12 | 0.2773685875028... |
| 0.66 | 0.34 | 0.12 | 0.2715388247255... |
| 0.77 | 0.22 | 0.08 | 0.3647373484212... |
| 0.66 | 0.26 | 0.08 | 0.2968725877094... |

**使用に関する注意**

提供する VALUE は、数値、数値を含む列、または数値を返す関数でなければなりません。

データの標準偏差は、その分散の平方根です。分析中の集合がすべてのデータポイントを表す場合（母集団と呼ばれます）、代わりに STDEVP を使用します。

# STDEVP

データの集合の全体（母集団）に存在する標準偏差（平均からの変動の度合い）を推定します。

**構文**

STDEVP(VALUE_1, [VALUE_2, ...])

- VALUE_1 は最初の値です。
- VALUE_2, ... [オプション]は追加の値です。

**例**

`STDEVP(@Column_A@, @Column_B@, @Column_C@)`

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| 0.66 | 0.28 | 0.12 | 0.2264705033528... |
| 0.66 | 0.34 | 0.12 | 0.2217105219775... |
| 0.77 | 0.22 | 0.08 | 0.2978067979225... |
| 0.66 | 0.26 | 0.08 | 0.2423954528359... |

**使用に関する注意**

提供する `VALUE` は、数値、数値を含む列、または数値を返す関数でなければなりません。

分析中の集合がデータポイントのサンプルのみを表す場合、代わりに `STDEV` を使用します。

# VAR

データのサンプル集合に含まれる値の分散度（値の散らばり度合い）を推定します。

**構文**

`VAR(VALUE_1, [VALUE_2, ...])`

- `VALUE_1` は最初の値です。
- `VALUE_2, ...` [オプション]は追加の値です。

**例**

`VAR(@Column_A@, @Column_B@, @Column_C@)`

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| 0.55 | 0.3 | 0.1 | 0.0508333333333... |
| 0.55 | 0.34 | 0.07 | 0.0579 |
| 0.66 | 0.34 | 0 | 0.1089333333333... |
| 0.66 | 0.2 | 0.04 | 0.1036 |

**使用に関する注意**

提供する `VALUE` は、数値、数値を含む列、または数値を返す関数でなければなりません。

分析中の集合がすべてのデータポイントを表す場合（母集団と呼ばれます）、代わりに VARP を使用します。

# VARP

データの全体集合（母集団）内に存在する分散の程度（値の散らばり度合い）を推定します。

**構文**

VARP(VALUE_1, [VALUE_2, ...])

- VALUE_1 は最初の値です。
- VALUE_2, ... [オプション]は追加の値です。

**例**

VARP(@Column_A@, @Column_B@, @Column_C@)

| Column_A | Column_B | Column_C | New Column |
|---|---|---|---|
| 0.55 | 0.3 | 0.1 | 0.0338888888888... |
| 0.55 | 0.34 | 0.07 | 0.0386 |
| 0.66 | 0.34 | 0 | 0.0726222222222... |
| 0.66 | 0.2 | 0.04 | 0.0690666666666... |

**使用に関する注意**

提供する VALUE は、数値、数値を含む列、または数値を返す関数でなければなりません。

分析中の集合がデータポイントのサンプルのみを表す場合、代わりに VAR を使用します。

# テキストが計算された列関数

このセクションでは、Data Prep 計算ツールで使用できるテキストが計算された列関数の構文と例を示します。

## CHAR

指定された ASCII 値の文字を戻します。

**構文**

`CHAR(INT)`

`INT` は文字が戻る ASCII 値です。

**例**

`CHAR(ASCII)`

| ASCII | New Column |
|---|---|
| 90 | Z |
| 85 | U |
| 70 | F |

## CONCATENATE

一連のテキスト文字列を単一のテキスト文字列に結合するには、[+] 演算子を使用できます。この例では、これらの間で 2 つの列とハイフンを組み合わせます:

`@Last@ +"-" + @First@`

または、 `CONCATENATE` 関数を使用することもできます。

`CONCATENATE(@Last@ , "-" , @First@ )`

**構文**

`CONCATENATE(STRING_1, [STRING_2, ...])`

- `STRING_1` は最初の値です。
- `STRING_2, ...` [オプション]は追加の文字列です。

**例**

`CONCATENATE(@Applicant Last@ ,", ",@Applicant First@ , " of ",@City@)`

| Applicant First | Applicant Last | City | New Column |
|---|---|---|---|
| Maximo | Ehmann | Wolfdale | Ehmann, Maximo of Wolfdale |
| Velia | Goldman | Wolfdale | Goldman, Velia of Wolfdale |
| Nick | Walters | Glennville | Walters, Nick of Glennville |
| Rachel | Ray | Hull | Ray, Rachel of Hull |

**使用に関する注意**

指定する `STRING` はテキスト文字列または数値、テキスト文字列または数値を含む列、またはテキスト文字列または数値を返す関数です。

# FIND

2 番目のテキストで 1 つの単語（またはテキストの文字列）が見つかるかどうかを決定します。見つかった場合、 `FIND` 関数は2番目のテキスト文字列内のテキスト文字列の数値位置を返します。2 番目の文字列の文字がカウントされ、最初のテキストが 2 番目のテキストの文字列で重なる文字を数値が示します。

この関数では、オプションの 3 番目の引数（数値）を指定できます。この数字は、 2番目の文字列で検索を開始したい位置を表示します（文字の個数により）。3番目の引数を省略すると、2番目の文字列の検索が最初の文字から開始されます。

最初の文字列が2番目の文字列中で見つからない場合、この関数は `0` を返します。

**構文**

`FIND(STRING_1, STRING_2, [VALUE])`

- `STRING_1` は見つけたい文字列です。
- `STRING_2` は検索したい文字列です。
- オプションの `VALUE` は、検索を開始したい `STRING_2` 内の数値位置です。

**例**

`FIND("Tech",@School@)`

| School | New Column |
|---|---|
| UC Sunnydale | 0 |
| Pacific Tech | 9 |
| Blue Mountain State | 0 |
| South Harmon Institute of Technology | 27 |

`FIND("the", "The quick sly fox jumped over the lazy brown dog laying next to the other dog.")` は**31**という値を返します。

`FIND("dog", "The quick sly fox jumped over the lazy brown dog laying next to the other dog.")` は **46** という値を返しますが、これは 2 番目の文字列において「dog」の最初の発生に対応します。

`FIND("dog", "The quick sly fox jumped over the lazy brown dog laying next to the other dog.", 47)` は **75** という値を返します。これは、**47** の第 3 引数が、過去の文字 **46**（最初の "dog" が発生する場所）の検索の開始をプッシュし、文字列で 2 番目の "dog" を見つけるように関数を強制するためです。

**使用に関する注意**

指定する `STRING` はテキスト文字列、テキスト文字列を含む列、またはテキスト文字列を返す関数である必要があります。同様に、 `VALUE` は数値、数値を含む列、または数値を返す関数である必要があります。

`STRING_1` が `STRING_2` で複数回発生した場合、 `FIND` は最初に一致した位置のみを示し、ペア内で連続する一致は示しません。

`FIND` 関数は大文字と小文字を区別するため、 `True` 、 `TRUE` 、および `true` を個別に扱います。

この関数は、単なる単語ではなく、テキスト全体で重なる部分を探します。したがって、テキスト "jump" は位置 1での文字列 "jumped" 内にあると判断されます。

単語だけでなく、テキスト文字も検索文字列として使用でき、2番目の文字列で発見されます。

# HASHVALUE

あいまい一致を簡単にするために、テキスト文字列を変換します。

**構文**

`HASHVALUE(STRING, OPTION, [VALUE])`

- `STRING` は変換したい文字列です。
- `OPTION` は変換に使用するアルゴリズムです。使用可能なオプション: METAPHONE
- `NGRAM FINGERPRINT`
- `VALUE` は、 `NGRAM` と一緒に使用すると、使用する Nグラムの数を指定します。

**例**

`HASHVALUE(@Current Employer@, "metaphone")`

| Current Employer | New Column |
|---|---|
| Boogle | PKL |
| CloudCo inc | KLTKNK |
| Self | SLF |
| Banana Inc | PNNNK |
| Acme inc. | AKMNK |

**使用に関する注意**

指定する STRING はテキスト文字列、テキスト文字列を含む列、またはテキスト文字列を返す関数である必要があります。OPTION と VALUE は文字列として扱われ、引用符で、"metaphone" のように囲む必要があります。

HASHVALUE は指定された文字列値に基づいてハッシュを生成するアルゴリズムを使用します。使用するアルゴリズムは、列内の値の間に近い一致を見つけるための列の操作 [Cluster + Edit]　によっても使用できます。METAPHONE、NGRAM、および FINGERPRINT の詳細については、Cluster + Edit を参照してください。

# LEFT

テキスト文字列の左端（最初）の位置から開始する、指定された文字数を返します。

**構文**

LEFT(STRING, VALUE)

- STRING は検索したい文字列です。
- VALUE は返す文字の数です。デフォルト設定では 1 が指定されています。

**例**

LEFT(@School@,4)

| School | New Column |
|---|---|
| UC Sunnydale | UC S |
| Pacific Tech | Paci |
| Blue Mountain State | Blue |
| South Harmon Institute of Technology | Sout |

**使用に関する注意**

指定する STRING はテキスト文字列、テキスト文字列を含む列、またはテキスト文字列を返す関数である必要があります。

# LEN

テキスト文字列に含まれる文字の数をカウントします。

**構文**

LEN(STRING)

STRING は評価したいテキスト文字列です。

**例**

LEN(@School@)

| School | New Column |
| --- | --- |
| Coolidge College | 16 |
| The University of Los Angeles | 29 |
| Camden College | 14 |
| South Harmon Institute of Technology | 36 |

**使用に関する注意**

指定する STRING はテキスト文字列、テキスト文字列を含む列、またはテキスト文字列を返す関数である必要があります。

# LOWER

列のテキストをすべての小文字に変換します。

**構文**

`LOWER(STRING, LOCALE)`

- STRING は小文字に変換したい文字列または列です。
- LOCALE （オプション）はロケールで、小文字に必要な文字を出力するには指定する必要があります。

サポートされているロケール値については、https://www.oracle.com/java/techangues/jdk8-jree8-jreed-loceles.htmlを参照してください。

**例**

`LOWER(@Values@, "tr")`

| Values | New Column |
| --- | --- |
| IAŞLIK | ıaşlık |
| IAŞLIK | ıaşlık |
| IAŞLIK | ıaşlık |

# MID

テキスト文字列の中央から指定された文字の数を返します。

**構文**

`MID(STRING, VALUE_1, VALUE_2)`

- STRING は評価したいテキスト文字列です。
- VALUE_1 は開始位置です。
- VALUE_2 は返す文字の数です。

**例**:

`MID(@School@,4, 5)`

| School | New Column |
| --- | --- |
| Blue Mountain State | e Mou |
| Pennbrook University | nbroo |
| Hillman College | lman |

**使用に関する注意**

指定する `STRING` はテキスト文字列、テキスト文字列を含む列、またはテキスト文字列を返す関数である必要があります。指定する値は数値、数値を含む列、または数値を返す関数である必要があります。

# PADLEFT

指定された回数の間、指定された文字を含む文字列をパッドします。これは、MySQL LPADと同じ出力を提供します。

**構文**

`PADLEFT(STRING, NUMBER, VALUE)`

- `STRING` または列はパッドする値です。
- `NUMBER` はその値に置き換える回数です。
- `VALUE` は文字どおりの置換値です。

**例**

`PADLEFT(@set@, 10, "-")`

| set4 | New Column |
| --- | --- |
| test1 | -----test1 |
| test2 | -----test2 |
| test3 | -----test3 |

# PADRIGHT

指定された回数の間、指定された文字を含む文字列をパッドします。これは、MySQL LPAD および RPAD と同じ出力を提供します。

**構文**

`PADRIGHT(STRING, NUMBER, VALUE)`

- `STRING` または列はパッドする値です。

- NUMBER は、その VALUE に置き換える回数です。
- VALUE は文字どおりの置換値です。

**例**

PADRIGHT(@set@, 10, "-")

| set4 | New Column |
|---|---|
| test1 | test1----- |
| test2 | test2----- |
| test3 | test3----- |

# REGEXP

正規表現を使用してテキスト文字列上で検索と置換を実行します。この関数は、Java Regex に基づいています。

> **ヒント**
>
> 置換を行うことなく、文章内から特定テキストの検索のみを行う場合については、FIND 関数を参照してください。 FIND 関数には、使用可能な構文の数が若干多いという利点がありますが、その代わりに FIND がパターンマッチングの面で若干効果が劣ります。

**構文**

REGEXP(STRING_1, STRING_2, STRING_3)

- STRING_1 は検索したいテキスト文字列です。
- STRING_2 は検索しているテキストです。
- STRING_3 は、STRING_2 を置き換えたいテキストです。

3 つの引数は必須です。 STRING_1 はテキスト文字列、テキスト文字列を含む列、またはテキスト文字列を返す関数である必要があります。 STRING_2 および STRING_3は、検索および置換アクティビティを定義する文字列の組み合わせで構成されます。

> **備考**
>
> 正規表現には、特殊な意味を持つ12文字があります:
> \ ^ $ . | ? * + ( )[ と開き波括弧です。
> こうした実際の文字とその特殊な意味のないものを検索したい場合は、その前にダブルバックスラッシュ （シングルバックスラッシュではなく）を追加します。例えば、正規表現でアスタリスク文字を検索するには「」*ではなく*「\」を入力します。正規表現でバックスラッシュ文字を検索するには、4 つのバックスラッシュ文字を入力します。

**例**

**空白文字をアンダースコアに変換**

`REGEXP(@School@," ", "_")`

| School | New Column |
|---|---|
| Adams College | Adams_College |
| California University | California_University |
| Adams College | Adams_College |
| University of New York | University_of_New_York |

**テキスト文字列を別のテキスト文字列に置き換える**

REGEXP(@ProductID@ , "ABC", "DEF")

**スラッシュをハイフンに変換**

`REGEXP("The/quick/sly/fox." , "/", "-")` は*The-quick-sly-fox* を返します。

**バックスラッシュ（特殊文字）をハイフンに変換**

`REGEXP(@ProductID@ , "\\\\", "-")`

**アスタリスク（特殊文字）をハイフンに変換するには**

`REGEXP(@ProductID@ , "\\*", "-")`

**数値ではない列 1 から文字を削除**

`REGEXP(_column1_ ,"[^0-9]", "")`

**抽出および置換のパターンの例**

| コマンド | 戻り値 |
|---|---|
| `RegexpExtract("replace me", "e m")` | `"e m"` |
| `RegexpExtract("replace me", "e.?m")` | `"e m"` |
| `RegexpExtract("replace me", "r.*c")` | `"replac"` |
| `RegexpExtract("123123456789", "(123)+456(.*)")` | `"123123456789"` |
| `RegexpExtract("123123456789", "(123)+456(.*)", 0)` | `"123123456789"` |
| `RegexpExtract("123123456789", "(123)+456(.*)", 1)` | `"123"` |

| コマンド | 戻り値 |
| --- | --- |
| `RegexpExtract("123123456789", "(123)+456(.*)", 2)` | `"789"` |
| `RegexpExtract("456789", "(123)*456(.*)", 2)` | `"789"` |
| `RegexpReplace("replace me", "e m", "---")` | `"replac---e"` |
| `RegexpReplace("replace me", "e.?m", "---")` | `"replac---e"` |
| `RegexpReplace("replace me", "r.*c", "--")` | `"--e me"` |
| `RegexpReplace("123123456789", "(123)+456(.*)", "---")` | `"---"` |
| `RegexpReplace("123123456789", "abc", "---")` | `"123123456789"` |

**使用に関する注意**

Regex パターンマッチングの詳細については、以下を参照してください:

https://docs.oracle.com/javase/8/docs/api/java/util/regex/Pattern.html

# REPEAT

指定された文字列 N を何度も繰り返します。

**構文**

`REPEAT(VALUE,REPEAT)`

- `VALUE` は検索して繰り返す文字列または列です。
- `REPEAT` は、 `VALUE` を繰り返す回数です。

**例**

`REPEAT(@set4@, 3)`

| set4 | New Column |
| --- | --- |
| test1 | test1test1test1 |
| test2 | test2test2test2 |
| test3 | test3test3test3 |
| test4 | test4test4test4 |

# REPLACE

指定した文字の数に基づいて、テキスト文字列の一部を異なるテキスト文字列と置き換えます。

**構文**

`REPLACE(VALUE, START NUM,NUM CHARS, NEW VALUE)`

- `VALUE` は文字を置き換えたいテキストまたは列です。
- `START NUM` は置き換えたい値にある文字の開始位置です。
- `NUM CHARS` は新しい文字列に置き換えたいテキスト内の文字の数です。
- `NEW VALUE` は置換値です。これは大文字と小文字を区別するので注意してください。

**例**

`REPLACE(@timestamp@,10,5," ")`

| timestamp | New Column (1) |
|---|---|
| 2019-01-23T05:24:56 | 2019-01-2 24:56 |
| 2019-01-23T05:24:56 | 2019-01-2 24:56 |
| 2019-01-23T05:24:56 | 2019-01-2 24:56 |
| 2019-01-23T05:24:56 | 2019-01-2 24:56 |
| 2019-01-23T05:24:56 | 2019-01-2 24:56 |

**使用に関する注意**

テキスト文字列の指定位置で発生するテキストを置き換えたい場合に `REPLACE` を使用します。テキスト文字列で指定したテキストを置き換えたい場合は `SUBSTITUTE` を使用します。例: `REPLACE(@Hospital Name@, Search(@Hospital Name@,"ADVOCATE"), 8, "ALPHA")`

# REVERSE

指定された文字列を反転します。

**構文**

`REVERSE(STRING)`

`STRING` は反転する列の値または文字列です。

**例**

`REVERSE(@set4@)`

# RIGHT

テキスト文字列の右端（終わり）の位置から開始する、指定された文字数を返します。

**構文**

`RIGHT(STRING, VALUE)`

- `STRING` は検索する文字列です。
- `VALUE` は返す文字の数です。デフォルト設定では 1 が指定されています。

**例**

`RIGHT(@School@,4)`

| School | New Column |
|---|---|
| Pacific Tech | Tech |
| Grand Lakes University | sity |
| Coolidge College | lege |
| South Central Louisiana State University | sity |

**使用に関する注意**

指定する `STRING` はテキスト文字列、テキスト文字列を含む列、またはテキスト文字列を返す関数である必要があります。

# SEARCH

指定された文字列を検索し、文字列のインデックスを返します。見つからない場合、-1 の値を返します。

**構文**

`SEARCH(VALUE, STRING)`

- `VALUE` は文字を置換したいテキストまたは列です。
- `STRING` は検索する文字列です。

**例**

`SEARCH(@Hospital Name@, "ADVENTIST")`

| Hospital Name | New Column (1) |
|---|---|
| ADVANCED SURGICAL HOSPITAL | -1 |
| ADVENTIST BOLINGBROOK HOSPITAL | 1 |
| ADVENTIST GLENOAKS | 1 |
| ADVENTIST LA GRANGE MEMORIAL HOSPITAL | 1 |
| ADVENTIST MEDICAL CENTER | 1 |
| ADVENTIST MEDICAL CENTER | 1 |
| ADVENTIST MEDICAL CENTER - REEDLEY | 1 |
| ADVOCATE BROMENN MEDICAL CENTER | -1 |
| ADVOCATE CHRIST HOSPITAL & MEDICAL CEN... | -1 |

**使用に関する注意**

`SEARCH` は `REPLACE` と組み合わせることができます。

**例**

`REPLACE(@Hospital Name@, Search(@Hospital Name@,"ADVOCATE"), 8, "ALPHA")`

# STR

引数内のデータをテキスト文字列に変換します。

**構文**

`STR(VALUE)`

`VALUE` はテキスト文字列に変換したい値です。

**例**

`STR(@Date@)`

| Date | New Column |
| --- | --- |
| 2016-03-19T00:00:00.000Z | 2016-03-19T00:00:00.000Z |
| 2012-06-30T00:00:00.000Z | 2012-06-30T00:00:00.000Z |
| 2013-12-28T00:00:00.000Z | 2013-12-28T00:00:00.000Z |

**使用に関する注意**

指定する値は数値、数値を含む列、または数値を返す関数である必要があります。

`STR` 関数は、数値をテキストに変換する場合や、テキストと数値が混ざった値の列の全体をテキストの列として処理することにより、その他のテキスト関数が正常に実行できることを保証する場合に便利です。

# SUBSTITUTE

テキスト文字列内の古いテキストを新しいテキストに置換します

**構文**

`SUBSTITUTE(VALUE, OLD TEXT, NEW TEXT)`

- `VALUE` は文字を置換したいテキストまたは列です。
- `OLD TEXT` は置き換えたいテキストです。これは大文字と小文字を区別するので注意してください。
- `NEW TEXT` は古いテキストを置き換えるために使用したいテキストです。これは大文字と小文字を区別するので注意してください。

**例**

`SUBSTITUTE(@Hospital Name@ ,"CREIGHTON","Merton")`

| Hospital Name | New Column |
|---|---|
| ALBANY MEMORIAL HOSPITAL | ALBANY MEMORIAL HOSPITAL |
| ALBANY VA MEDICAL CENTER | ALBANY VA MEDICAL CENTER |
| ALBEMARLE HOSPITAL AUTHORITY | ALBEMARLE HOSPITAL AUTHORITY |
| ALBERT EINSTEIN MEDICAL CENTER | ALBERT EINSTEIN MEDICAL CENTER |
| ALEGENT CREIGHTON HEALTH BERGAN MERCY MEDICAL CTR | ALEGENT MERTON HEALTH BERGAN MERCY MEDICAL ... |
| ALEGENT CREIGHTON HEALTH CREIGHTON UNIVERSITY MED | ALEGENT MERTON HEALTH MERTON UNIVERSITY MED |
| ALEGENT CREIGHTON HEALTH IMMANUEL MEDICAL CENTER | ALEGENT MERTON HEALTH IMMANUEL MEDICAL CENT... |
| ALEGENT CREIGHTON HEALTH LAKESIDE HOSPITAL | ALEGENT MERTON HEALTH LAKESIDE HOSPITAL |
| ALEGENT CREIGHTON HEALTH MEMORIAL HOSPITAL, SCHUYL | ALEGENT MERTON HEALTH MEMORIAL HOSPITAL, SCH... |

**使用に関する注意**

テキスト文字列中の特定のテキストを置き換える場合は、 `SUBSTITUTE` を使用します。テキスト文字列の指定位置で発生する任意のテキストを置き換える場合は、 `REPLACE` を使用します。

# TRIM

指定された文字列のすべての先端および末尾のスペースを削除します。

> **備考**
>
> 関 `TRIM` 数は、テキストから 7 ビット の ACCII 空白文字（値32）を削除するように設計されています。Unicode文字セットには、十進値が 160 であるノーブレークスペース文字と呼ばれる追加のスペース文字があります。この文字は、HTML エンティティとして Web ページで一般的に使用されています。TRIM 関数自体はノーブレークスペース文字を削除しません。

**構文**

`TRIM(STRING)`

`STRING` は削除したい値です。

列を次の例の文字列値として指定できます。

**例**

TRIM(@Company@)

| Company | New Column |
|---|---|
| Apple corp | Apple corp |
| Apple corporation | Apple corporation |
| Apple computers | Apple computers |

# TRIMLEFT

文字列の左端からホワイトスペースの削除された文字列を返します。

**構文**

`TRIMLEFT(STRING)`

`STRING` は削除したい列の値です。

**例**

TRIMLEFT(@Company@)

| Company | New Column |
| --- | --- |
| Apple corp | Apple corp |
| Apple corporation | Apple corporation |
| Apple computers | Apple computers |

# TRIMRIGHT

文字列の右端からホワイトスペースの削除された文字列を返します。

**構文**

`TRIMRIGHT(STRING)`

`STRING` は削除したい列の値です。

**例**

`TRIMRIGHT(@Company@)`

| Company | New Column |
| --- | --- |
| Apple corp | Apple corp |
| Apple corporation | Apple corporation |
| Apple computers | Apple computers |

# UPPER

列のテキストをすべて大文字に変換します。

**構文**

`UPPER(STRING,LOCALE)`

- `STRING` は大文字に変換したい文字列または列です。

- LOCALE （オプション）はロケールで、大文字に必要な文字を出力するには指定する必要があります。

サポートされているロケール値については、https://www.oracle.com/java/techangues/jdk8-jree8-jreed-loceles.htmlを参照してください。

**例**

UPPER(@Values@, "tr")

| Values | New Column |
|---|---|
| ıaşlık | IAŞLIK |
| ıaşlık | IAŞLIK |
| ıaşlık | IAŞLIK |

# VALUE

文字列の値として保存された数を数値に変換します。

**構文**

VALUE(STRING)

STRING は数であり、数値に変換したいテキスト文字列として保存されています。

**例**

VALUE(@COLUMN@)

| Column A value | New Column |
|---|---|
| 6.588 | 1464.46139524329530184499 2 |
| 9.43 | 468217.34343705300007777749 |
| 11.345 | 36056362.17229731793933645445682699 |
| 14.796 | 640499653925.32590180357923031075 24 |
| 20 | 2.43290200817664E+18 |

**使用に関する注意**

指定する STRING はテキスト文字列として保存された数値、テキスト文字列として保存された数値を含む列、またはテキスト文字列として保存された数を返す関数である必要があります。

STRING が数以外の文字を含む場合、関数はエラーを返します。実数を作るには、引数内で一つのピリオド（小数点）が許可されます。

VALUE 関数は、テキスト値を数に変換して、それに対して数字ベースの関数が正常に実行できるように、数値の列を数値の列として扱うようにするため、便利です。

# 比較演算子

比較演算子を使用して論理的な条件をテストします。ほとんどの場合、TRUE または FALSE の値を生成するために IF 関数の最初の引数が使用されます。

Data Prepで使用できる演算子は以下の通りです。

| オペレーター | 定義 | TRUE が返される例 |
|---|---|---|
| = | Equal To | 1 + 2 = 3 |
| > | Greater Than | 3 > 2 |
| >= | Greater ThanまたはEqual To | 11 >= 10<br>11 >= 11 |
| < | Less Than | 2 < 3 |
| <= | Less ThanまたはEqual To | 10 <= 11<br>10 <= 10 |
| <> | Equal Toない | 2 <> 3 |

## 数値を含む比較演算子

数値比較のために比較演算子を使用することは簡単です。比較する 2 つの値は、必ずしも同じデータ型を使用しなければなりません。テキスト値“3”は、数値3と同じではありません。

データ型の混合を防ぐために、VALUE 関数を使用して、テキストとして保存されている数字を数値に変換します。たとえば、"3"=3 は FALSE を返しますが、VALUE("3") = 3 は TRUE として評価されます。

## テキスト値を含む比較演算子

テキストに対して最もよく使用される比較演算子は =（等しい）です。この演算子は、2つのテキスト文字列が同じであるかどうかを判定する際に使用されます。マッチングを実行する他の文字列関数（FINDなど）と同様に、大文字と小文字が区別されることに備考してください。つまり、"The" と "the" は異なる文字列として扱われます。true の値になるように比較をするには、大文字と小文字を含めて、すべてのテキストが正確に一致する必要があります。<>（同等ではない）の使用は、=（イコール）の使用と同じパターンに従います。テキスト文字列をチェックするときに、大文字と小文字を区別します。

<=（Less Than or Equals To）および> =（Greater Than or Equals To）を含む<（Less Than）および>（Greater Than）を含む比較でさえ、テキスト値に使用できることに備考してください。 文字は数値で表されます。2つの文字は同じではないため、2つの文字は同じ数値を共有していません。

ユーザーがテキスト比較の結果を予測するには、コンピューターでエンコードされた文字を出力する方法についていくつかの追加の情報が必要です。

# 計算されたカスタム列関数

**備考**

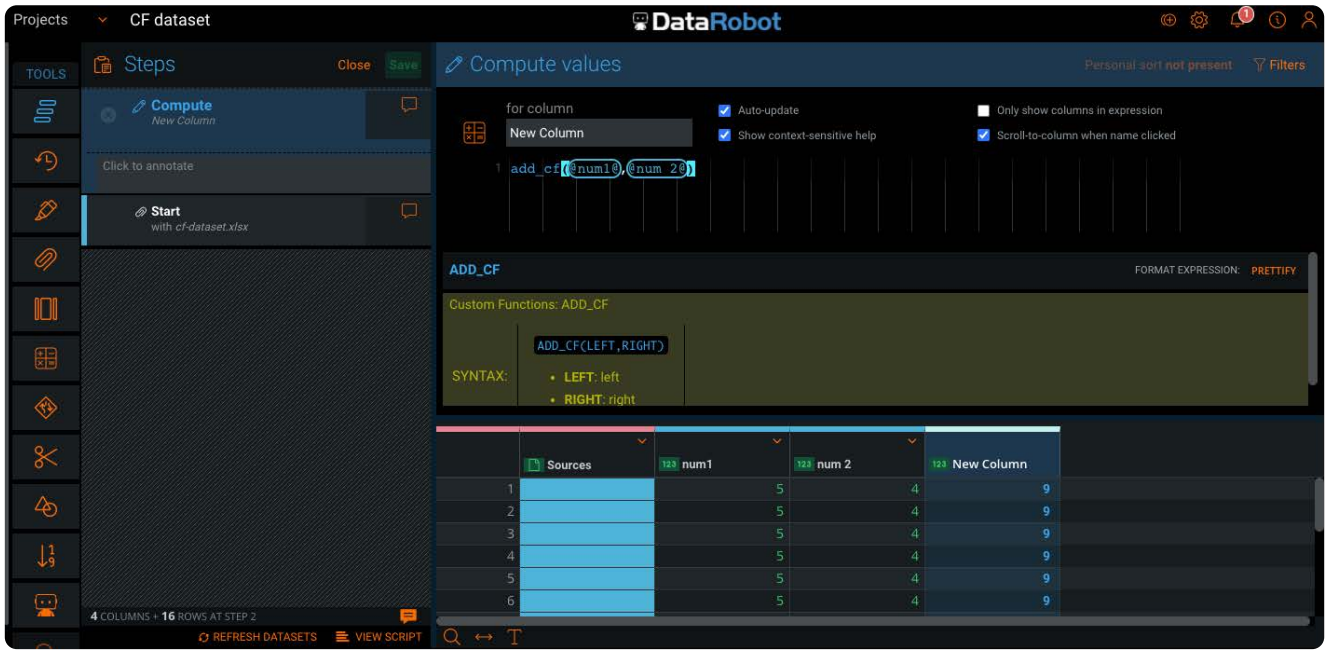
Data Prepカスタム関数は、オンプレミスおよび仮想プライベートクラウドのインストールでのみ利用可能です。Data Prepの管理者は、アプリケーションでこの機能を有効にする必要があります。Data Prepカスタム関数の開発に関するドキュメントについては、DataRobotサポートにお問い合わせください。

組織がカスタム関数を開発して、インストールしている場合、データセット内の既存の列にカスタム関数を適用して新しい列を追加できます。次のセクションでは、カスタム関数で計算ツールを使用する方法を説明します。

## カスタム関数の使用

計算ツールでカスタム関数を使用するには：

1. Data Prepの**ツール**バーで**計算**をクリックします。
2. **値の計算**ペインで新しい列に名前を付けます。
3. 列名の下にある式の行にカスタム関数の名前を入力します。式の行の下に関数の使用詳細が表示されます。
4. 関数で変数として使用する列を選択します。使用ガイドラインに従って表現を作成します。




この例ではカスタム関数の名前は `add_cf` であり、 `num1` と `num2` はデータセット内の列です。 `add_cf(@num1@,@num2@)`

式にエラーが含まれている場合、ビルトインと同様、計算ツールにエラーメッセージが表示されます。エラーは解決されるまでステップツール内の計算ステップに表示されます。 5. 新しい列を表示して関数が意図したとおりに機能していることを確認します。

5. **保存**をクリックして新しい列を保存します。

# 列データを操作する

Data Prepには、列を操作するための多くのメソッドが用意されています。このセクションのトピックでは、値のハイライトと変更、値の検索と置換、データのフィルタリングなど、列データに対して実行できる操作について説明します。これらの作業は、**フィルター**ペイン、**列を表示**ペインで行うことになり、各列の上にあるメニューから列操作を実行します。

> **ヒント**
>
> このセクションでは、列データを使った操作について説明します。列全体を管理するには、プロジェクト**ツール**バーにある**列**ツールを使います。**列**ツールを使用すると、列名を更新したり、列の順序を変更したり、プロジェクトから列を削除したりできます。詳細は列の更新をご覧ください。

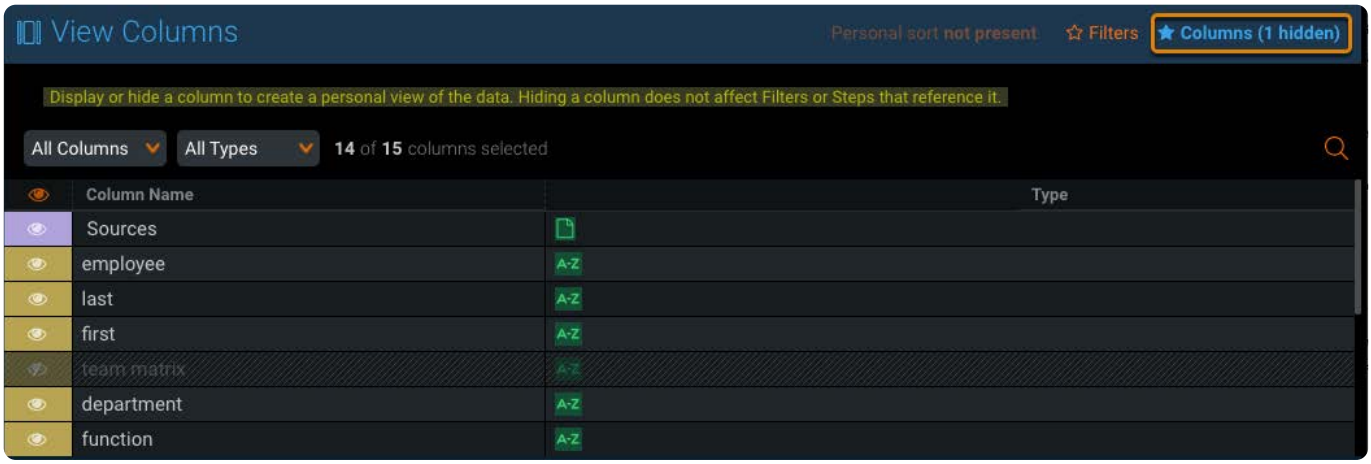
## フィルターペイン

**フィルター**ペインには、選択した列に対するデータ*Filtergram*が表示されます。データ Filtergramは、フィルターとヒストグラムを組み合わせたものです。この例では、*機能*はテキスト列であり、*年俸*は数値列です。


データプレビューペインの右上の**フィルター**をクリックして、**フィルター**ペインを表示します。詳しくは「データフィルタグラム」をご覧ください。

## 列を表示ペイン

**列を表示**ペインでは、列タイプが表示され、列を非表示にできます。この例では、チームマトリックス列が非表示になっています。




データプレビューペインの右上の**列**をクリックして、**列を表示**ペインを表示します。**列を表示**ペインを使用してデータプレビューペインから列を削除する方法については、列を非表示を参照してください。プロジェクトから列を完全に削除するには、列ツールを使用します。

# 列の操作

各列の上にあるメニューで、列データを操作するほとんどのツールにアクセスできます。


これらのページでは、列の操作について説明しています。

| トピック | 説明... |
|---|---|
| データ型の変換 | DataPrepがデータ型を識別して変換する方法を学びます。 |
| 列の値の変更 | たとえば、大文字と小文字の変更、データ型の変更、スペースのトリミング、空白の管理などにより、列の値を変更します。 |

| トピック | 説明... |
| --- | --- |
| 列の検索および置換 | 1列または複数列内のテキストを検索して置換します。 |
| 列の非表示 | 列を非表示にして、個人用ビューを作成したり、AnswerSetを公開する準備をしたりします。実際に列を削除したい場合は、「列の更新」を参照してください。 |
| データのフィルタリング | データ Filtergramを使用してデータを探索し、フィルターを処理します。 |
| 日付形式の検出および変換 | Data Prepで日付の書式設定を操作する方法を学びます。 |
| 列の分割 | 指定された文字列、文字数、または正規表現に基づいて列を分離します。 |
| 列を埋める | 同じ列で空白セルの直前または直後にあるセルの既知の値に基づいて、空白セルにデータを入力できます。 |
| クラスターおよび編集による正規化 | データを正規化し、列の不整合やエラーを特定します。 |
| 列の系統表示 | 選択した列に生じたプロジェクトのステップを特定します。 |

# データ型変換

Data Prepのデータ変換機能は異種データ型をサポートします。異種データ型とは、データセットをData Prepライブラリにインポートするときに、セルレベルでデータ型が自動的に識別されることを意味します。Data Prepプロジェクトの同じ列内で異種データ型をサポートする能力は強力です。この能力により、あらゆるデータをプロジェクトに取り込むことができるからです。その結果、データ品質の課題がつきまとう混在データ型も、以下の「ベストプラクティス」セクションで説明するように、Data Prepでは同質化と調停を容易に実行できます。異種データをサポートしない、強力な型指定を行う他のアプリケーションの場合、データ準備作業を始める前に、別のツールを使用してソースデータを同質化する必要があります。

この記事の目的は、セルと列に対してData Prepがいつどのようにデータ型を特定するかを説明するとともに、Data Prepプロジェクト内で異種データを操作するためのベストプラクティスを説明することです。

## Data Prepはデータタイプをどのように識別するのでしょうか。

ユーザーがデータセットを Data Prepライブラリにインポートするときに、すべてのセル内にあるすべてのデータは、インポートプロセスの一環として以下のデータ型のいずれかを自動的に特定されます。

- 数値
- ブーリアン
- String（文字型）またはText（テキスト型）
- Date Time（日付時刻型）（以下で説明する条件の場合）

Data Prepは、以下のルールに従うアルゴリズムを使用してこの作業を実行します。

1. 値がnullの場合、その値を無視します。
2. 値が文字通り、“true”（真実）または “false”（偽）である場合、その値を Boolean（論理型）として扱います。
3. プログラムにより値を数値として読み取ることが可能な場合、その値をNumeric（数値型）として扱います。
4. その他のすべての値はデフォルトでString（文字型）になります。

たとえば、10列のデータセットと100万行のデータが含まれます。これは、合計1000万個のセルに変換されます。この場合、Data Prepは上記のアルゴリズムのルールに従って、1,000万個のセルのそれぞれに対してデータ型を特定します。

次に、1つの列の各セル内に存在している主要なデータ型に基づいて、*列ごとに*キャスト（データ型変換）を実行します。

## 日付時刻の値

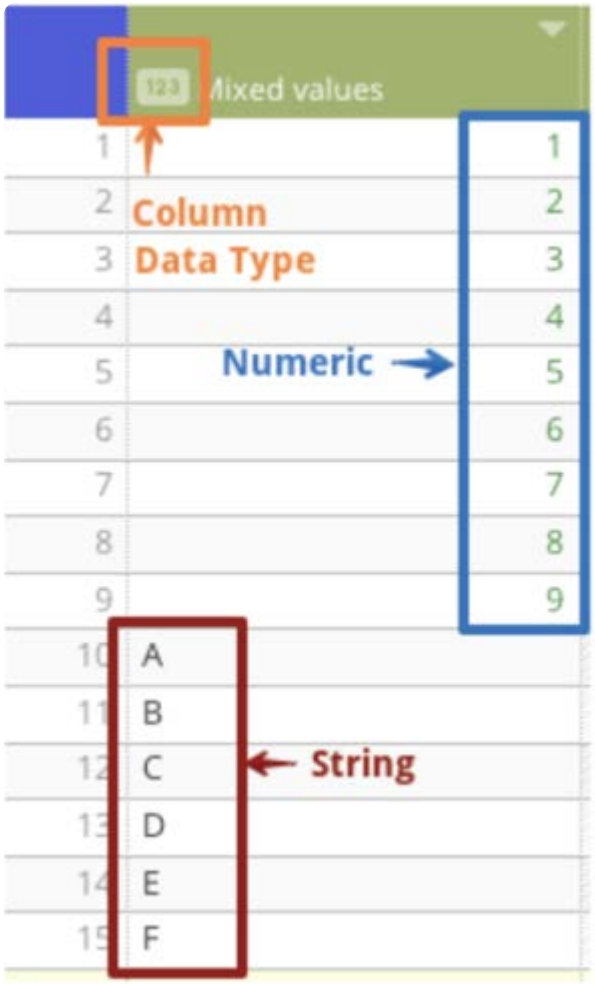
1 つのルールとして、日付と時刻の多様な形式の分析と解決を実行するには複雑な処理が必要なので、Data Prepはフラットファイル内にある date time（日付時刻型）の値を識別しません。ただし、このルールにもいくつかの例外が存在します。以下の条件下でセルのデータをインポートするときに、セルのデータはdate time（日付時刻型）として認識されます。

- データベーステーブル（JDBC、Hiveなど）*および*データベースがスキーマを提供している
- Parquetファイル*および*Parquet形式がスキーマを提供している
- Microsoft Excelファイル*および*Excel形式が各セルのデータ型を指定している

## 異種データ

同じ列内にある異種データに対して、Data Prepはどのように列の型を決定しますか？100万のデータ行が存在している上記の例に戻ると、かなりの確率で*同じ列の中*には別のデータ型に属すデータが存在しています。たとえば、string（文字型）の値とnumeric（数値型）の値が同じ列の中に混在している可能性があります。 この場合、Data Prepには列のデータ型をキャスト（データ型変換）する方法を特定する別のロジックがあります。このロジックについて説明するために、非常に簡潔な別の例を使用しましょう。

ここに、1行につき1列、全部で15行のデータがあります。最初の9行はNumeric（数値型）として識別され、残る6個の値はString（文字型）として識別されています。この列の型が、以下のようにNumeric（数値型）としてキャストされたことに注意してください。




インポートプロセスの際に、アルゴリズムの計算が自動的に実行され、60%の値がNumeric（数値型）、40%がStrings（文字型）であると検出されました。この列内の主要なデータ型に基づいて、Data Prepが列の型をキャストする方法が特定されます。この場合、この列はNumeric（数値型）としてキャストされます。

## 1 つの列内で複数のデータ型が同値（同じ割合）である場合はどうなりますか?

同値、たとえば列の 50% の値が特定の型で、残りの 50% が別の型である場合、この同値を裁定するために、計算ロジックは以下の付加的なルールを提供しています。

| 列データの組み合わせ | 裁定結果 |
|---|---|
| 50%のブール値と50%の日付 | ブーリアン |
| 50%のブール値と50%の数値 | ブーリアン |
| 50%のブール値と50%の文字列 | ブーリアン |
| 50%の日付と50%の数値 | 数値 |

| 列データの組み合わせ | 裁定結果 |
| --- | --- |
| 50%の日付と50%の文字列 | 文字列 |
| 50%の数値と50%の文字列 | 文字列 |

要約すると、優先度の順に並べる場合、同数の場合は以下のように裁定されます。

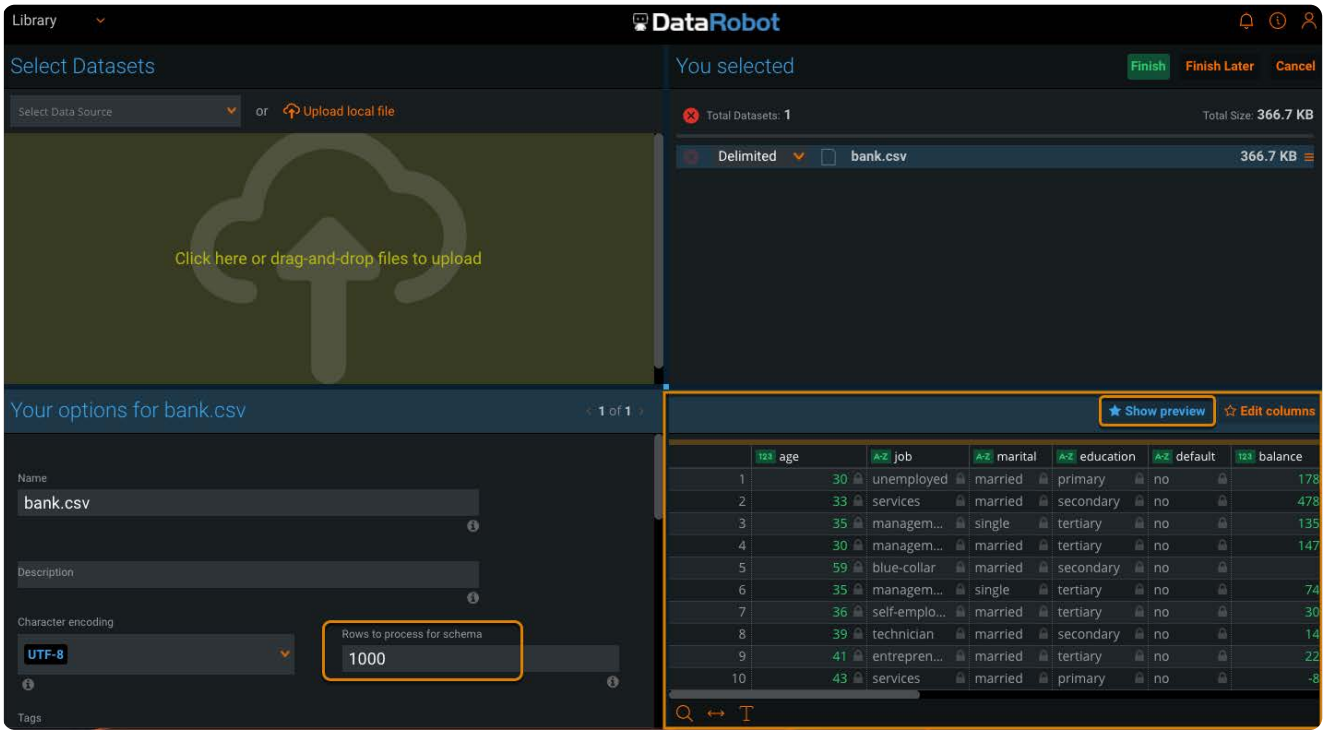
1. ブーリアン
2. 文字列
3. 数値
4. 日付

## データの追加の効果

Data Prepライブラリまたは既存のプロジェクトに対してユーザーが新しいデータを取り込んだ結果、1 つの列内にある主要なデータ型が変化した場合はどうなりますか?カラムの型の推論と、それに伴うキャストが発生するのは、Data Prepライブラリへのインポート プロセスのときだけです。

データ型が、1つの列内にある主要なデータ型を高い精度で反映しない可能性がある2種類のシナリオは、以下のとおりです。

- Data Prepライブラリへのインポートの際に、列の型を推論する目的で最初の1,000行のデータを使用します。一般的に、Data Prepはデータセットの列の型を高精度で推定してキャストを行ううえで、1,000行のデータで十分であるという結論を下しました。なお、この行数はユーザーが設定することも可能です。これら最初の 1,000 行には、「プレビュー」状態という非公式な呼び名があります。これは、最初のロードの場合、または既存のデータセットに対応する更新済みのバージョンをロードする場合のどちらかで、データセットをロードしている間にアプリケーション内でユーザーが目にする状態です。




中には、最初の1,000行を過ぎた後、特定の列の主要なデータ型が変化するという、通常とは異なる状況もあり得ます。この場合、この列はそれら最初の1,000行を使用して引き続きキャストが行われます。プレビューの行数を設定するには、オプションペインの**スキーマ用に処理する行**のフィールドの値を変更します。行数はインポート中に設定できますが、Data Prepのベストプラクティスでは、Filtergramsを使用してデータ品質の問題を特定して対処することを推奨します。詳細については、ベストプラクティスを参照してください。

- 既存のプロジェクト内で参照または追加の操作を実行した後、その操作の結果として列に流入したデータに基づいて、列の主要な型が変化する可能性があります。列の型に関する推論が発生するのはインポートプロセスのときだけなので、当初キャストに使用された列の型は、新しい主要な型にかかわらず、不変のままです。ただしData Prepのベストプラクティスは、データ品質の課題を識別および解決するために、複数のソースから取得したデータをアンサンブルした後、標準的なデータ調停プラクティスの一部として、ユーザーが常にFiltergramsを使用することを推奨しています。

# ベストプラクティス

## 自分のデータ内でデータ型の課題を突き止めて修復する目的で、Data Prepをどのように使用すればよいでしょうか？

Data Prepはごく初期の段階から、データ品質に関係するこの種の課題を突き止めて解決することを目的として構築されています。通常、データセットをライブラリにインポートした直後、またはデータセットをプロジェクトに追加した直後に推奨される次のステップは、データ品質が強化される方法でデータ型を裁定することです。データの調停とは、データ準備のうち重要な要素の1つであり、ユーザーが調停を実行できるように、Data Prepは視覚的なインジケーターとFiltergramsのようなツールを提供しています。

### 例

既存のプロジェクトへの追加を行った後、この列の主要なデータ型が"numeric"（数値型）から"string"（文字型）に変化したとします。インポートプロセスの際に、この列は当初numeric（数値型）として識別され、正しくキャストが行われ、このプロ

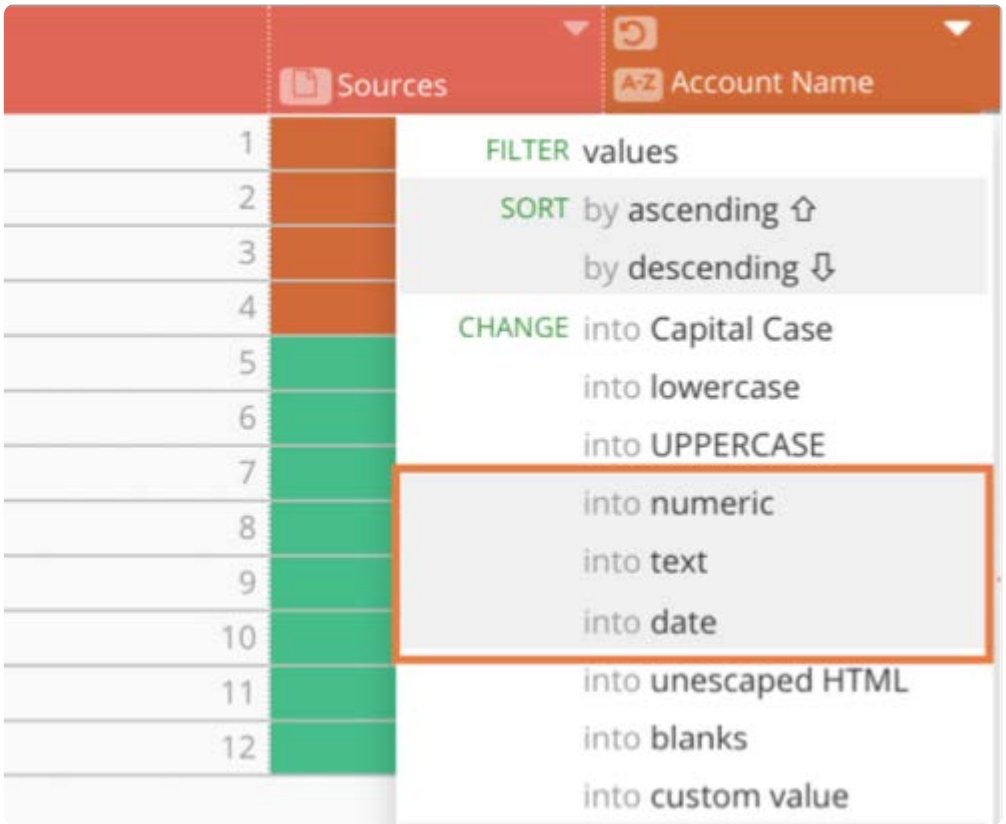
ジェクト内で引き続き使用されています。追加操作の後、主要な型は現在はnumeric（数値型）であるにもかかわらず、列の型がstring（文字型）にとどまっていることに注意してください。ただし、視覚的なインジケーターは、numeric（数値型）の値が右寄せで表示される状態を示しており、この列で型の不一致が発生していることを容易に認識できます。

| | Sources | Account Name |
|---|---|---|
| 1 | | AAA |
| 2 | | AAA rentals |
| 3 | | AAA daily rentals |
| 4 | | AAA montly rentals |
| 5 | | 1234 |
| 6 | | 5678 |
| 7 | | 11223344 |
| 8 | | 11223344 |
| 9 | | 11223344 |
| 10 | | 1234 |
| 11 | | 1234 |
| 12 | | 1234 |

この列に対応するFiltergramを開くと、この列内に存在する値のうち、この列の型に取って「有効」でない値をすぐに特定できます：

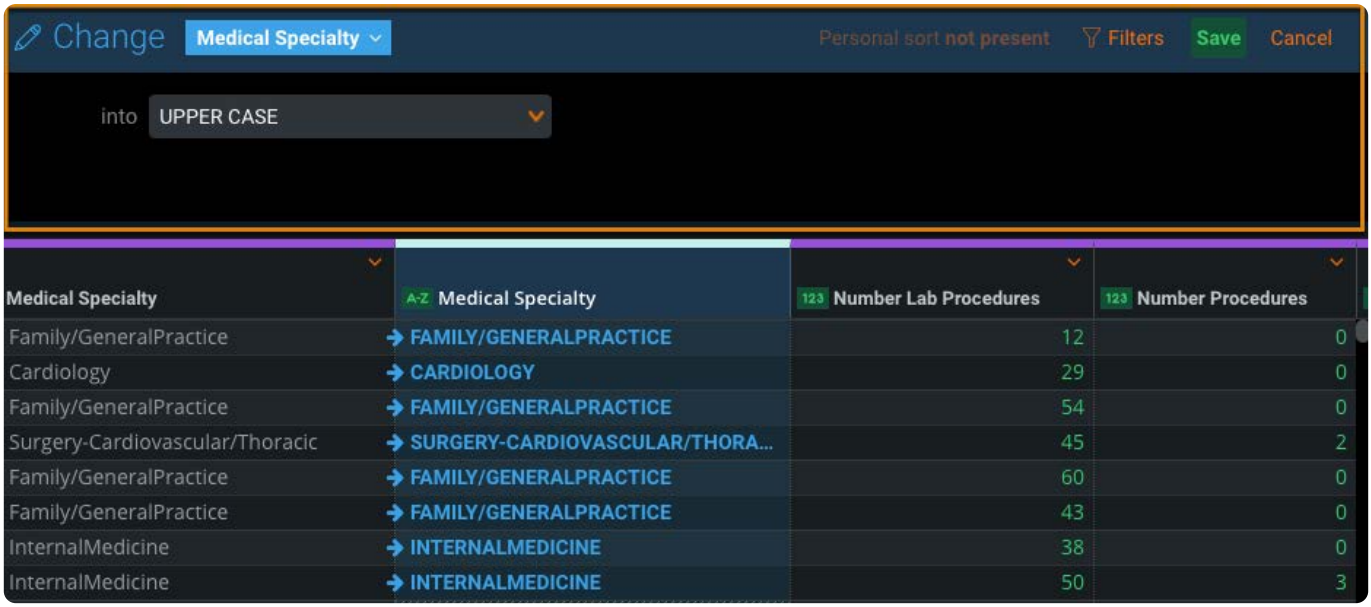

フィルターグラムの赤いバーは、不適合なデータ型があることを示しています。**無効**リンクをクリックして、適合しない値のみを表示します。

「有効ではない」データ型、つまり主要な型とは異なる他のすべての型を識別して、それら有効でない値のみを表示するようにフィルター処理を行った後、プロジェクト内でレンズを作成し、それら不適合の値のみを示す1つのAnswerSetを生成することができます。その後、そのAnswerSetを使用して、それらの値に関する修復プロセスを支援することができます。「無効な」型を確認した後、その列の型を別のデータ型に変換することを考えるはずです。次のように列メニューを使用して、この操作を実行できます。



# 列の値の変更

Data Prepでは、列メニューの**変更**操作を使用してデータ値を変更できます。この例は、テキストを大文字に変更する"Medical Specialty"列の変更操作を示しています。



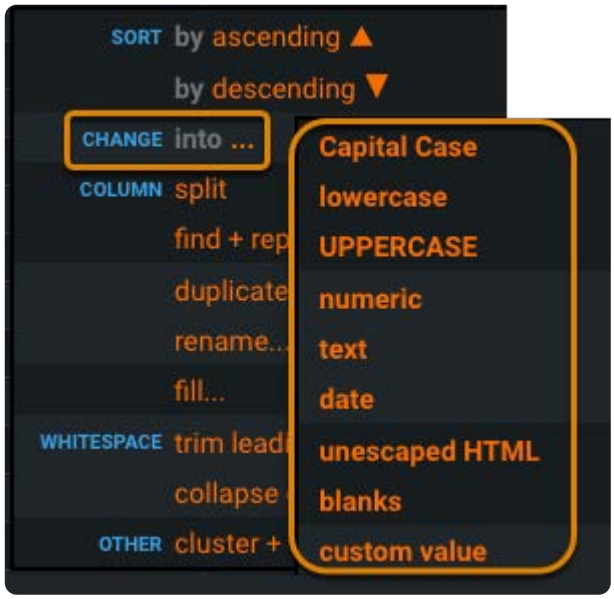

**変更**いずれかの操作を使用して1つ以上の列を選択し、それらの列のデータを次のように変更します。

- 各語の最初の文字のみ大文字
- 小文字
- 大文字
- 数値
- テキスト
- 日付
- アンエスケープ HTML
- 空白
- カスタム値
- 列のセルから前後の空白を削除する
- 連続する複数の空白を1つの空白にする

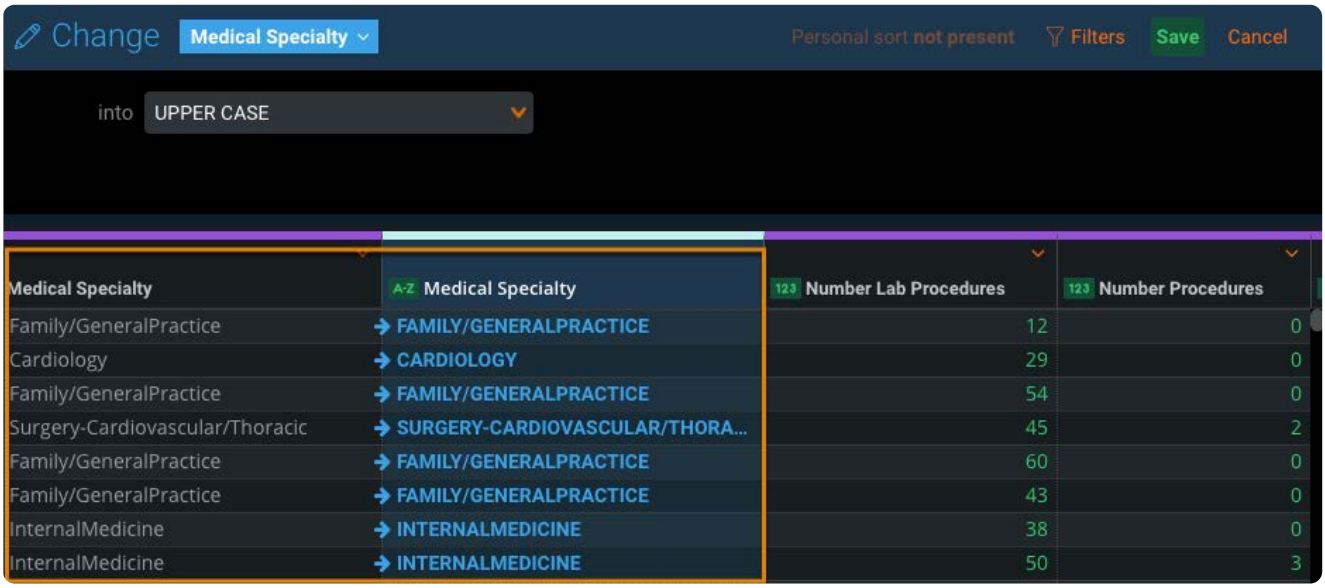
# 単一の列の値を変更する

1つの列でテキストを検索して置換するには：

1. 値を変更する列を見つけます。
2. 列メニューアイコン にカーソルを合わせて、列メニューにアクセスします。**変更**にカーソルを合わせて、必要な変更を選択します。




Data Prepは、変更を反映する元の列のコピーを生成します。次に例を示します。




3. 上部にある**保存**をクリックし 変更を受け入れます。

# 複数の列の値を変更する

データセット全体または特定の列セット全体で値の変更を実行する必要がある場合は、*高度な* **変更**ペインを使用します。高度な機能が役立ついくつかの例を次に示します。

- データセット全体にわたって"incorporated"と"Inc"の両方が存在していて、データセット全体が"Inc"値のみとなるように標準化する場合。データセット全体を標準化して、"Inc"値のみになるようにします。
- データセットはどこにでも「組み込まれて」おり、ほとんどの場合それは精度の高いです。ただし、データセット内の_特定の列_の値を"Inc"に変更する必要があります。
- 2つのデータセットをプロジェクトに取り込みました。1つは"NA"とし、もう1つは適用できない値を表すためにブランクを使用します。すべての"NA"値をブランクに変更したいとします。

複数の列にわたって検索して置換するには：

1. 列のメニューアイコンにカーソルを合わせて、**検索+置換**をクリックします。
2. **検索+置換**ペインに表示される列名をクリックします。


3. 高度な**検索+置換**ペインで、検索と置換の操作に含める各列の横にあるチェックボックスをクリックします。


複数の列にわたる検索と置換の残りの手順は、単一の列を検索して置換する手順と同じです。検索と置換を参照してください。

# 名前または基準による値の変更

高度な**検索+置換**ペインでは、*名前*または*基準*のいずれかで複数の列を選択できます。

## 名前による値の変更：

*名前*で選択して置換は、選択した特定のカラム_のみに_変更が適用されます。

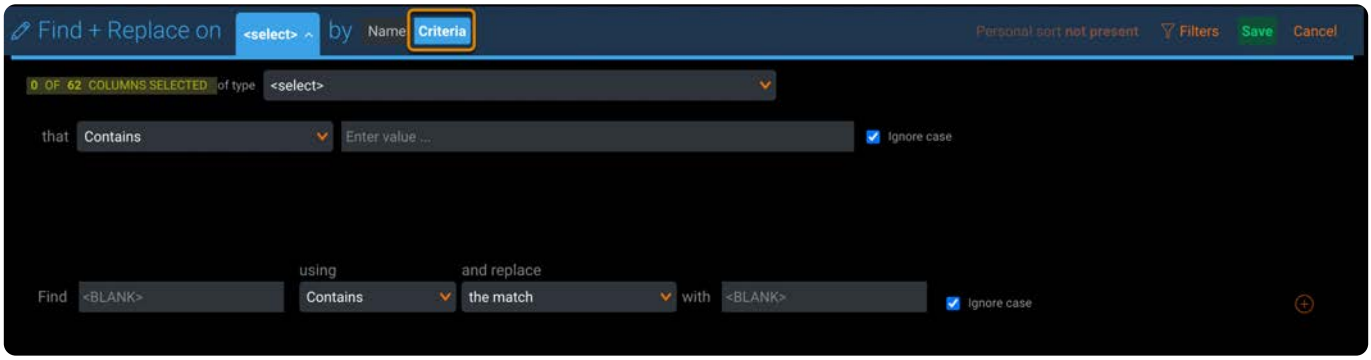
*名前*に基づいて列を選択するには：

- 選択する列の隣にあるチェックボックスをクリックします。

- 最上部のチェックボックスをクリックすると、すべての列が選択されます。
- パネル上部の列フィルターと型フィルターを使用すると、操作対象として選択する列を迅速にフィルタリングできます。
- 検索機能を使用して、列を名前で検索します。

## 基準による値の変更：

基準で検索して置換すると、指定した*基準*を満たす*すべての列*に変更が適用されます。




たとえば、データセットに文字列型の列があり、文字列型の列に対して置換操作を指定した場合、データセット内にあるこの型の既存の列と、このステップの前にデータセットに追加された新しい文字列型の列のすべてが動的に置換されます。

基準に基づいて列を選択するには：

- 必要に応じて列のデータ型（ブール値、日時、数値、文字列）を指定します。
- 必要に応じて列名のパターン（次の値で始まる、次の値を含む、次の値に等しい、次の値で終わる）を指定します。

ヘッダーのメッセージが更新され、その基準に基づいて選択した列の数が表示されます。これより前のステップに新しいデータが取り込まれ、基準を満たした列が追加または削除された場合は、選択した列の数が増加または減少することがあります。

> **備考**
>
> 置換操作を保存する前に**名前**オプションと**基準**オプションを切り替えた場合には、Data Prepは選択内容を記憶します。このとき、**直前の選択を復元**のためのリンクをクリックすると、最初の選択方法に戻ります。

# 例：数値への変更

この列操作は、テキスト文字列として保存されたすべての数字を数値に変換します。これにより、その列内の値に対して数学演算を実行できるようになります。 数字がテキストとして保存されたままの状態では、このようなアクションは無効と見なされます。

文字列として保存されている数字は、左詰めの黒色のテキストとしてセルに表示されます。 数値として保存されている数字は、右詰めの緑色で表示されます。

数値に変換できないセルに対してこの操作を適用しても、効果はありません。1つの列内にテキスト行と数字の行が含まれる場合は、変換可能な行の値だけが変換されます。

変換可能に見える値が正しく変換されない場合は、セル内に数字以外の文字が含まれている可能性があります。

変換を妨げる文字の例を以下に示します。

- 前後の空白。「**数値に変換**」の前に「**前後の空白を**削除」列操作を実行すると、これらの空白を削除できます。

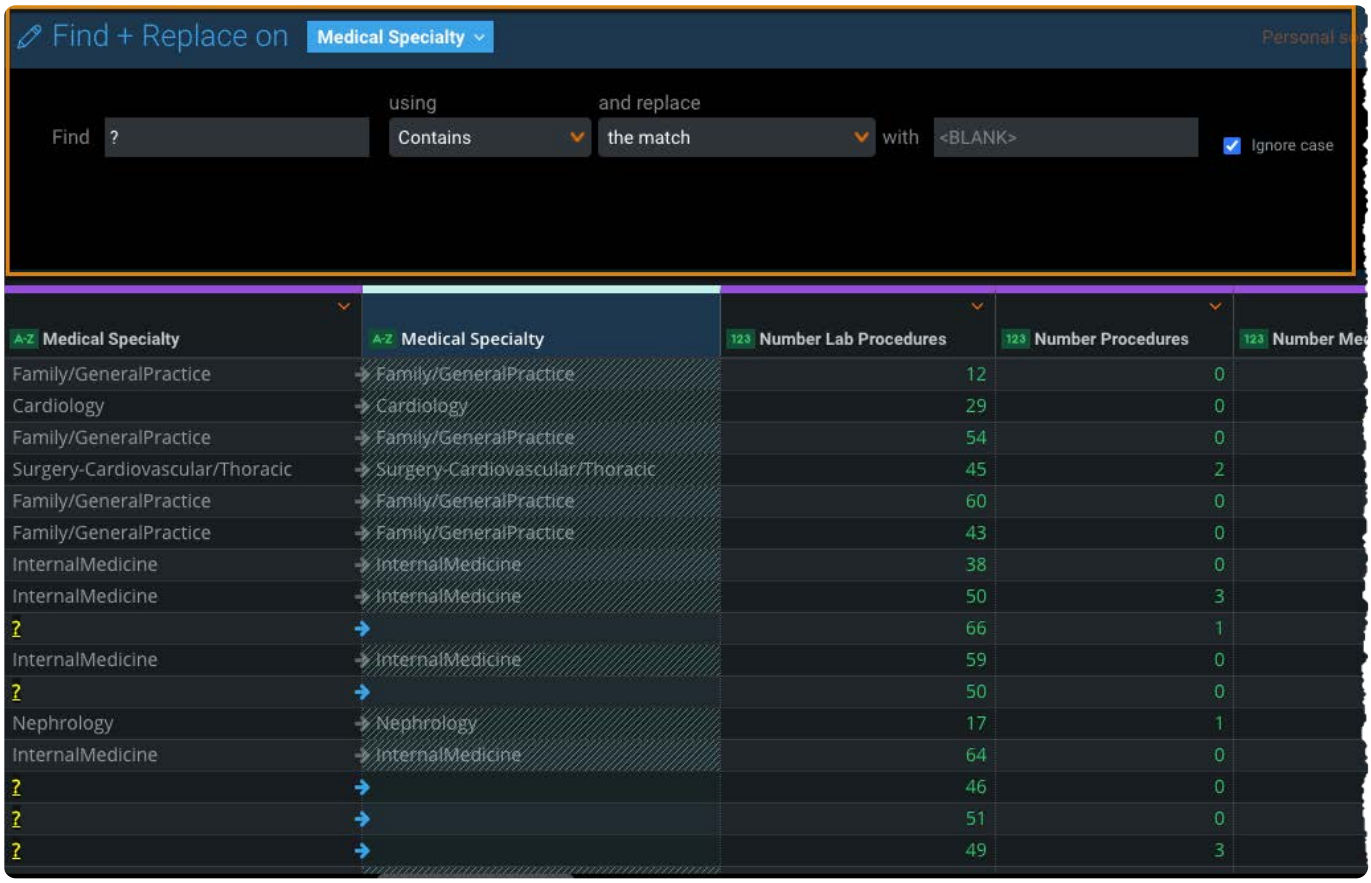
  「前後の空白を削除」機能は、すべての行について、テキスト文字列の先頭と末尾に空白がないかどうかを調べます。見つかった空白は削除され、値だけがセルに残ります。
- 中間文字（カンマ、空白など）数値の列を適切に作成するには、事前に列の分割などの操作を実行したり、REGEXを使用する計算列の使用が必要になったりする場合があります。

**備考**

数値のセルでは、単一のピリオド（"."）は小数点として解釈されます。この特殊文字の場合は、数字型への型変換に影響はありません。

# 列の検索および置換

Data Prep検索と置換操作では、指定された列内にテキストを検索して置換できます。この例は、"Medical Specialty"列での検索と置換の操作を示しています。 ? 値はブランクに置き換えられます。

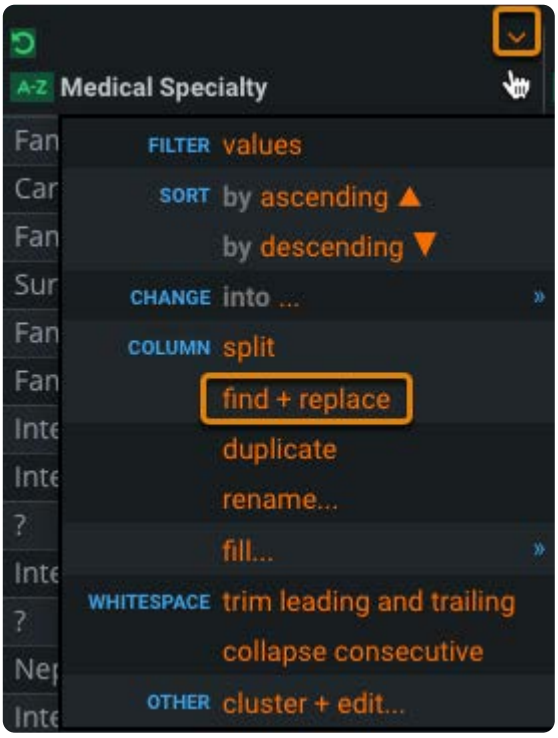



複数またはすべての列にわたって検索操作と置換操作を行うこともできます。詳細については、複数の検索および置換操作を参照してください。

## 検索と置換

1つの列でテキストを検索して置換するには：

1. 値を検索して置き換える列を見つけます。
2. 列メニューアイコンにカーソルを合わせて、**検索 + 置換**をクリックします。




3. **検索**フィールドでは、検索する値を指定します。

   または、検索する値を持つセルをダブルクリックすると、検索と置換のフィールドに、そのセルの値が自動的に設定されます。

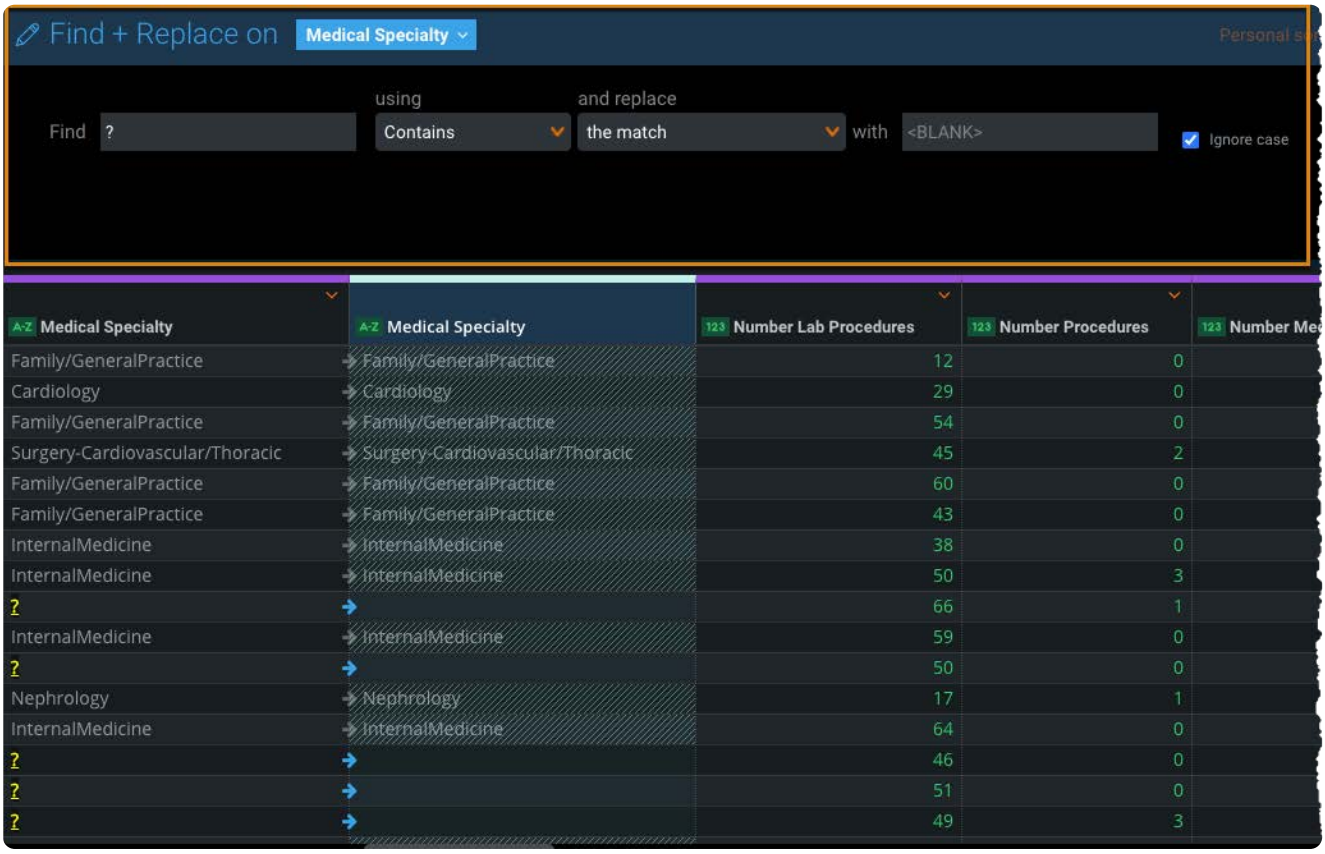
4. **使用**フィールドで、**検索**フィールドで指定されたテキストを一致させる方法を選択します。

   - **次の値を含む：**一致は、セル値のどの部分でもかまいません。
   - **次の値で始まる**: セル値の始まりが一致する必要があります。
   - **イコール：**一致は正確である必要があります。
   - **次の値で終わる:**セル値の終わりが一致する必要があります。

5. **置換**フィールドで、一致のどの部分を置き換えるかを選択します。

   - **セル全体:** セルの内容全体を置換します。
   - **一致:** セルの一致部分のみを置換します。

   Data Prepは、変更を反映する元の列のコピーを生成します。次に例を示します。




6. 上部にある**保存**をクリックし 変更を受け入れます。

## 例

例 1: 元のセル値 = "123456"検索："123"およびMatch"321"で置換し、結果は"321456"です。

例 2: 元のセル値 = "123456""123"を検索し、[セル全体]を指定して"321"に置換すると、結果は"321"です。

> **ヒント**
>
> **ステップ**ツールを無効にした場合、**保存**ボタンは**検索 + 置換**ペインに表示されます。**ステップ**ツールを有効にすると、**保存**ボタンは**ステップ**ペインの上部に表示されます。

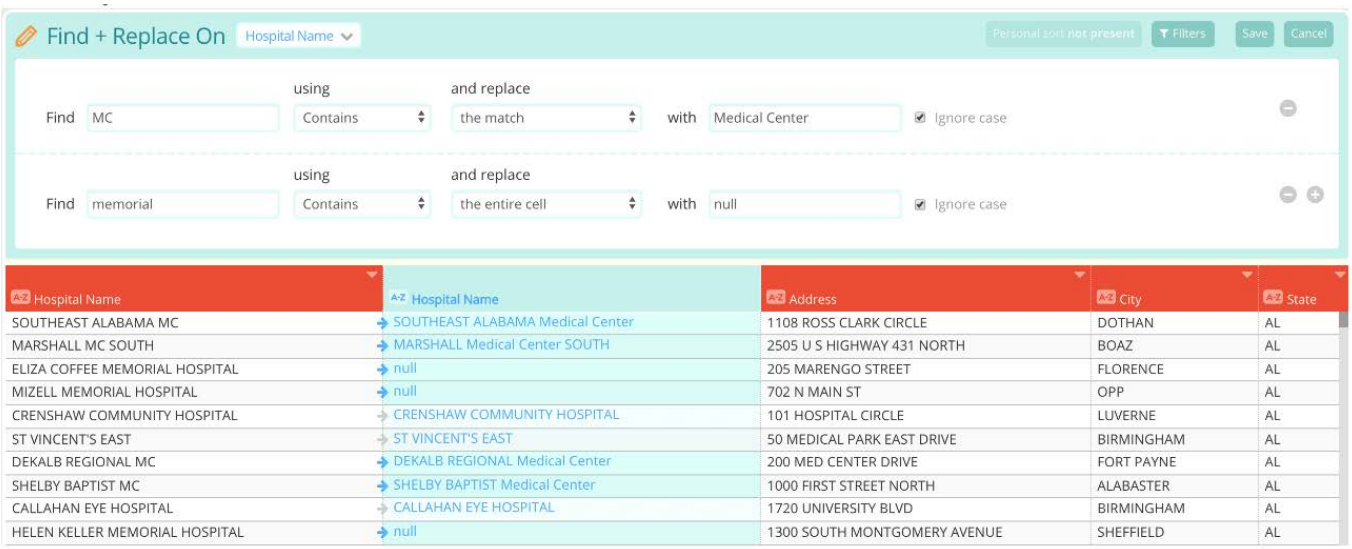
# 複数回の検索と置換操作

置換を継続し、すべての検索と置換による変換を現在のステップに保存する場合は、ここで保存せずに、プラス（+）をクリックして続行します。

追加の検索および置換操作は、それぞれが反復的です。つまり、追加の検索は直前の検索および置換操作の結果に対して実行され、その後にすべての置換が、**検索して置換**ペインに表示されているとおり、上から下に順番に適用されます。

## 複数回の検索と置換操作の例

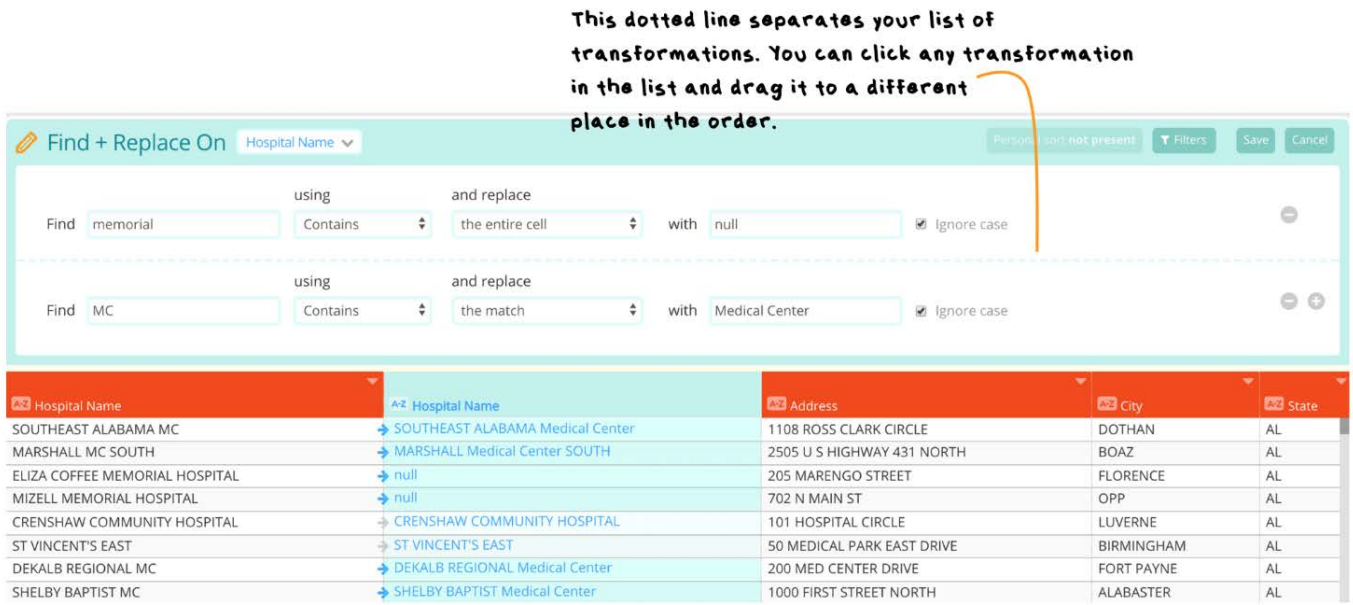
次の例では、最初の操作で"memorial"のすべてのインスタンスが検索され、一致したテキストだけでなくセル全体がnull値に置き換えられます。




以下に、別の例を示します。

> "Detroit"を見つけて"San Francisco"に置換します。"San Francisco"を見つけて"San Jose"に置換します。この結果、"Detroit"が"San Jose"に変換されます。

パネルで検索および置換操作を別の位置にドラッグ＆ドロップすれば、変換の順番をいつでも並べ替えることができます。



このステップのすべての検索および置換変換を保存するには、**保存**ボタンをクリックします。またこのステップの検索して置換の操作を削除するには、その操作の（-）ボタンをクリックします。

# 検索と置換に関する重要な注意

- デフォルトでは、検索で大文字小文字の区別が無視されます。たとえば、テキストcatは、テキストCaTと同じです。大文字の使用に重要な意味がある場合は、[大/小文字の区別を無視する]チェックボックスの選択を解除します。
- デフォルトでは、単一のプロジェクト ステップでの検索および置換操作は、250 回に制限されています。この回数を超えると、**ステップ**ツールにエラーメッセージが表示され、列に対する検索および置換変換を続行できません。この制限回数を多くする必要がある場合は、システム管理者に連絡してください。
- グリッドの**ハイライト**機能は、単一の**検索+置換**変換でのみ使用できます。 別の**検索+置換**変換を追加した場合は、使用できません。

# 複数列にわたる検索と置換

データセット全体または特定の列セット全体で検索と置換を実行する必要がある場合は、高度な**検索+置換**ペインを使用します。 高度な機能が役立ついくつかの例を次に示します。

- データセット全体にわたって"incorporated"と"Inc"の両方が存在していて、データセット全体が"Inc"値のみとなるように標準化する場合。データセット全体を標準化して、"Inc"値のみになるようにします。
- データセットはどこにでも「組み込まれて」おり、ほとんどの場合それは精度の高いです。ただし、データセット内の_特定の列_の値を"Inc"に変更する必要があります。
- 2つのデータセットをプロジェクトに取り込みました。1つは"NA"とし、もう1つは適用できない値を表すためにブランクを使用します。すべての"NA"値をブランクに変更したいとします。

複数の列にわたって検索して置換するには：

1. 列のメニューアイコンにカーソルを合わせて、**検索+置換**をクリックします。
2. **検索+置換**ペインに表示される列名をクリックします。


3. 高度な**検索+置換**ペインで、検索と置換の操作に含める各列の横にあるチェックボックスをクリックします。

複数の列にわたる検索と置換の残りの手順は、単一の列を検索して置換する手順と同じです。検索と置換を参照してください。

# 名前または基準による検索と置換

高度な**検索+置換**ペインでは、*名前*または*基準*のいずれかで複数の列を選択できます。

## 名前で検索して置換

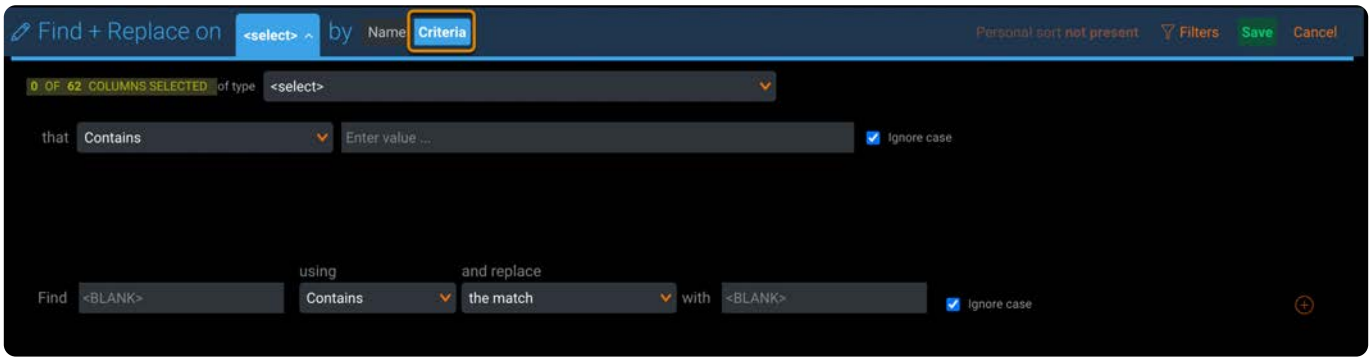
*名前*で選択して置換は、選択した特定の列_のみに_変更が適用されます。

*名前*に基づいて列を選択するには：

- 選択する列の隣にあるチェックボックスをクリックします。
- 最上部のチェックボックスをクリックすると、すべての列が選択されます。
- パネル上部の列フィルターと型フィルターを使用すると、操作対象として選択する列を迅速にフィルタリングできます。
- 検索機能を使用して、列を名前で検索します。

## 基準で検索して置換：

基準で検索して置換すると、指定した*基準*を満たす*すべての列*に変更が適用されます。




たとえば、データセットに文字列型の列があり、文字列型の列に対して置換操作を指定した場合、データセット内にあるこの型の既存の列と、このステップの前にデータセットに追加された新しい文字列型の列のすべてが動的に置換されます。

基準に基づいて列を選択するには：

- 必要に応じて列のデータ型（ブール値、日時、数値、文字列）を指定します。
- 必要に応じて列名のパターン（次の値で始まる、次の値を含む、次の値に等しい、次の値で終わる）を指定します。

ヘッダーのメッセージが更新され、その基準に基づいて選択した列の数が表示されます。これより前のステップに新しいデータが取り込まれ、基準を満たした列が追加または削除された場合は、選択した列の数が増加または減少することがあります。

**備考**

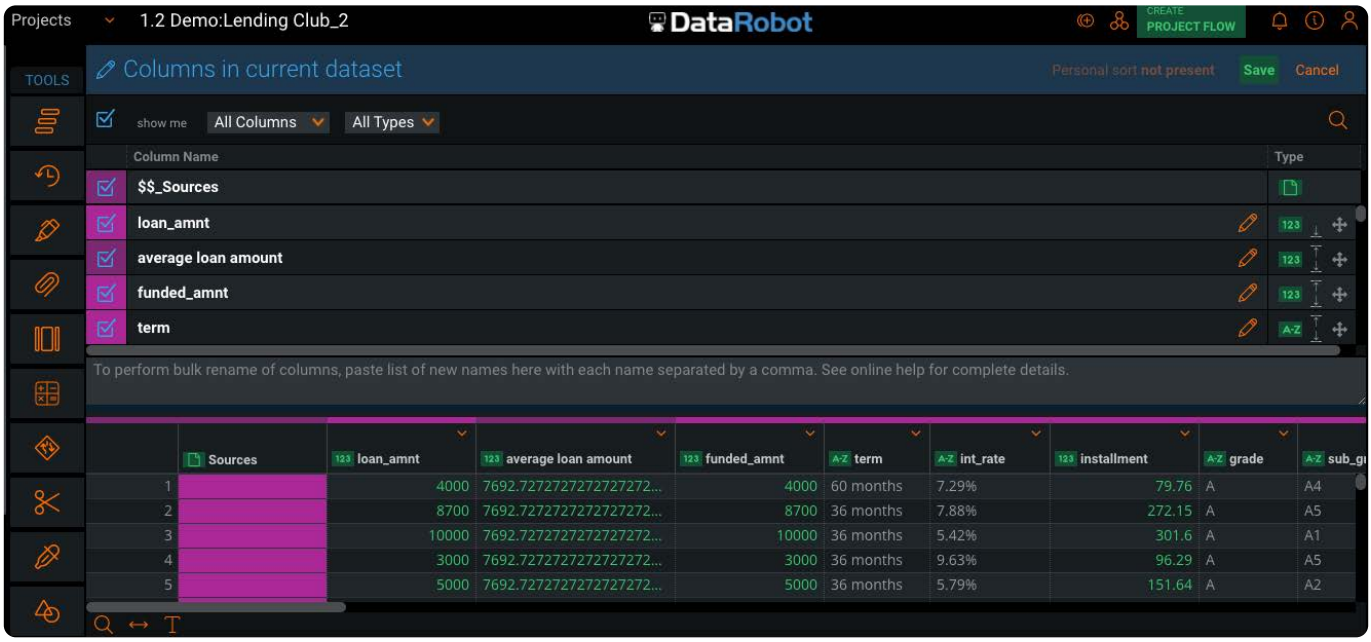
置換操作を保存する前に**名前**オプションと**基準**オプションを切り替えた場合には、Data Prepは選択内容を記憶します。このとき、**直前の選択を復元**のためのリンクをクリックすると、最初の選択方法に戻ります。

# 列を非表示

プロジェクト内に含まれる列を削除するのではなく、非表示にする場合に当機能を使用してください。

- **個人用ビューを作成する場合:** プレビューで表示する必要はないものの、プロジェクトの工程上の計算処理で使用する目的でデータ内に残しておく必要があるカラムを、グリッド上で非表示とすることができます。たとえば、プロジェクト内での作業が主に数値列を対象とした作業である場合、データの個人ビューを作成して、数値以外のすべての列タイプを非表示とします。
- **特定の列のみを含むAnswerSetを公開する場合:** データの一部の列のみを表示するカスタムAnswerSetを公開したい場合があります。まず、AnswerSetを作成するステップにレンズを作成し、次にAnswerSet上に表示したくない列を非表示にします。レンズを保存すると、選択項目が保存されます。レンズから公開をすると、表示を選択した列のみがAnswerSetで公開されます。

この例では、**現在のデータセットの列**ペインで、 `loan_amnt` および `funded_amnt` の列が非表示になっています。




**備考**

列を一時的に非表示にするのではなく、完全に削除する方法については、列の削除をご参照ください。

## 列を非表示

列を非表示にする手順：

1. 右上の**列**をクリックします。

2. 非表示にしたい列の左にある目のアイコンをクリックします。

   選択した目のアイコンがグレーアウトし、その列が**プレビュー画面**で非表示になります。

# 注意事項

列を非表示にしても、プロジェクトのフィルターやステップには影響しません。プロジェクトの編集を開始するとプレビューモードが自動的に終了します。つまり、個人用ビューのために、ここで非表示にした列は、レンズツールを使用する場合を除き、ライブデータに引き続き含まれて表示されます。これはレンズで公開した場合に、表示選択した列のみがAnswerSetに公開されるためです。

# 列の非表示に関するオプション

- 非表示にしたい列にフィルターオプションを設定することができます。**すべての列**および**すべてのタイプ**にフィルターを設定できます。これらのフィルターメニューを使用して、表示あるいは非表示を選択した列のみを表示させたり、**列の表示**ペインに表示させる列のタイプ（文字列、数値、日時）を管理できます。
- 連続する列のグループを選択して非表示にすることができます。そのためには、非表示にしたい列の横にある目のアイコンをクリックし、Shiftキーを押しながら選択したい列のグループを囲んで、目のアイコンをクリックします。
- データ内の列を検索することができます。そのためには、ペインの右上にある虫眼鏡のアイコンをクリックして、列名を入力することから開始します。
- **プレビュー画面**で、特定の列を強調することができます。そのためには、Shiftキーを押しながら列名の上にカーソルを合わせます。次に、その列が表示され、**プレビュー画面**でハイライトされます。

**ヒント**

プロジェクトを共有し、データとして個人用ビューも共有したい場合は、**レンズ**ツールを使用して、ビューをキャプチャします。レンズは、**プレビュー画面**の個人用ビューをキャプチャします。プロジェクトで作業している他の人は、**ステップ**ペインのレンズをクリックすることで、個人用ビューを見ることができます。

# データのフィルタリング

Data Filtergram™ は、データを変換するフィルターの力とデータを視覚化するヒストグラムの機能を組み合わせたツールです。Filtergramsを使用することで、変換前、変換中、および変換後に毎回データを視覚化することができます。フィルターグラムを使用して、次のことができます。

- データの探索。この方法でフィルターを使用する場合、フィルタリングされた選択を**プレビュー画面**に表示することができます。　選択した内容は、プロジェクトのステップに保存されません。むしろ、データを迅速に探索することができます。探索を終了したら、フィルターを削除し、プロジェクトで変換を継続することができます。
- 行の削除.データの準備作業に行の削除が含まれる場合、まずフィルターを使用し、削除したい行を特定する必要があります。
- 特定の行セットに含まれるデータのみを変更します。この場合、フィルターを使用して変更したい行を分割します。次に、これらの行だけに変更を適用することができます。例えば、名前のある列があり、「Anna」のすべてのインスタンスを「Anna」に変更したい場合、列をフィルタリングし、「Anna」の値のみを表示します。次に、**カスタム値に変更**する列操作を適用して "Anna"を "Ann" に変換します。
- 特定の行セットだけを公開します。この場合、フィルターを使用して、公開したい行を分割します。次に、レンズを追加して公開ポイントを作成することができます。

## Data Filtergram の作成

データフィルターグラムを作成する場合:

1. フィルターしたい列を見つけます。
2. 列メニューのアイコンにカーソルを合わせ、**フィルター値**をクリックします。

複数の列に対して Filtergram を開くと、**プレビュー画面**でこれらのフィルターの結果をプレビューすることができます。

各タイプの列データに対応するフィルターグラムには次の5つのタイプがあります:

- テキスト
- 数値
- 日付と時刻
- ブーリアン型
- ソース

以下のセクションでは、各タイプのフィルターを使用する方法について説明します。

# テキスト Filtergram

[テキスト フィルターグラム] ペインには、データセットに表示される、個別のテキスト値のリストが表示されます。左から右へ伸びるバーは、それぞれの値の相対的な出現回数のヒストグラムを示します。ユニーク数の総数がペインの左下に表示されます。リストから、データセットに動的に表示する値を選択できます。


テキスト フィルタグラム で実行できるアクションについては、テキスト フィルタグラム の操作を参照してください。


**a. 現在選択中**（左上）: リストから選択する場合、ボタンのラベルには選択した数が表示されます。ボタンをクリックして、選択されたすべての値を一覧表示する新しいペインを開きます。このペインから、引き続きデータセットに表示するテキスト値を絞り込むことができます。このペインから実行できるアクションは、テキスト フィルタグラム の操作セクションで説明しています。

**b. リスト順序の並べ替え:** デフォルトでは、テキスト値のリストは出現数の多い順に並んでいます。順序を逆にするには、右上隅の出現数の列の上にある三角形をクリックします。リストの上にある三角形をクリックしてアルファベット順に並べ替えることもできます。三角形のオレンジ色は、データセットに数字とアルファベットのどちらの並び順が、現在適用されているかを表しています。

**c. カラーコード付きフィルターバー**: Filtergram ペインにマウスを移動させると、以下のボタンが表示されます:

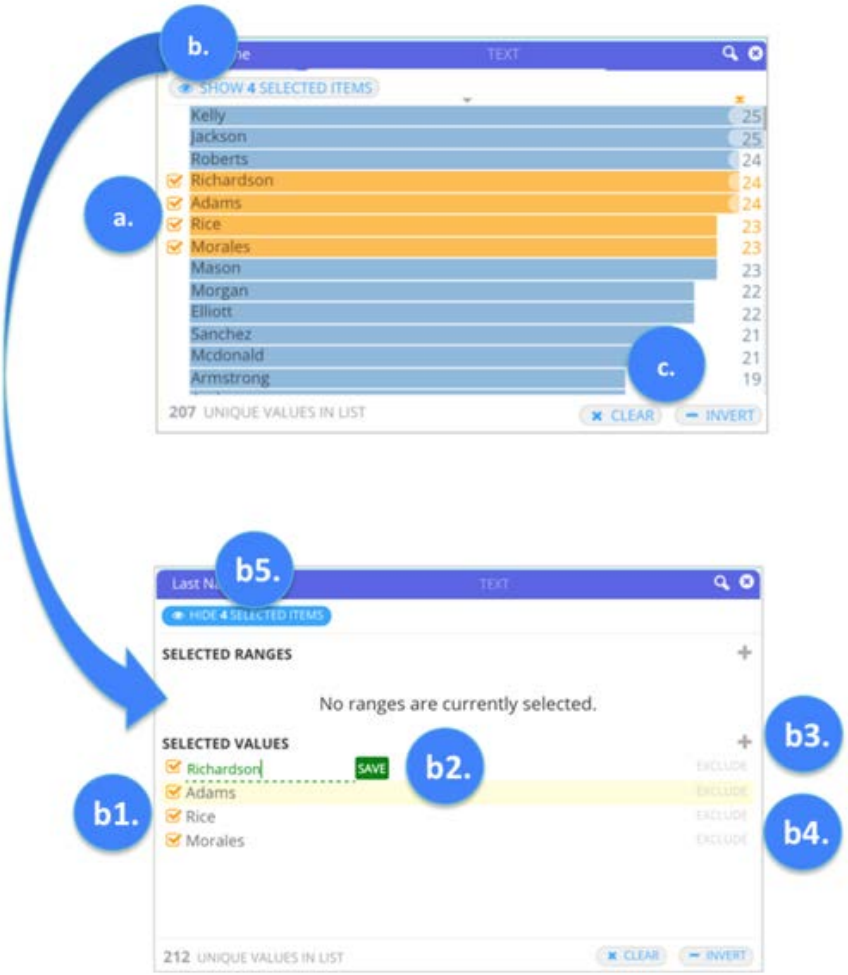
- **タイプ**：このボタンにマウスを置くと、「データセット内のテキスト型の行の総数」に対する「現在選択されているテキスト型の行の数」の比率が表示されます。リストで何も選択されていない場合は、「データセット内の行の総数」に対する「テキスト型の行の総数」の比率が表示されます。このボタンをクリックすると、データセットでこれらのテキスト値が非表示になります。列にブランク、エラー、またはその他のテキスト以外の値が含まれており、これらのデータタイプのみを表示する場合にこれは役立ちます。

- **その他**：列にテキスト以外の値がある場合（数値など）、このボタンが表示されます。このボタンにマウスを置くと、「データセット内の非テキスト値を持つ行の総数」に対する「現在選択されている非テキスト値を持つ行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「非テキスト値の行の総数」の比率が表示されます。**その他**をクリックするとオフになり、現在のデータセットビューから他の値を非表示にします。
- **空白**：列内に空白がある場合は、このボタンが表示されます。このボタンにマウスを置くと、「データセット内の空白行の総数」に対する「現在選択されている空白行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「空白行の総数」の比率が表示されます。**ブランク**をクリックするとオフになり、現在のデータセットビューからブランクセルを非表示にします。
- **エラー**：列内にエラーがある場合は、このボタンが表示されます。このボタンにマウスを置くと、「データセット内のエラーのある行の総数」に対する「現在選択されているセルエラーのある行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「セルエラーのある行の総数」の比率が表示されます。**エラー**をクリックするとオフになり、現在のデータセットビューからセルのエラーを非表示にします。


列内に「その他」の値、空白、またはエラーがある場合は、その値に対応するボタンの横に**+**ボタンも表示されます。**+**ボタンをクリックすると、そのタイプのすべての値 （すべての「その他」の値など）が [選択した項目] リストに追加され、このリストでさらにフィルタリング操作を続けることができます。**選択された項目**から行うフィルター機能については、テキスト Filtergram の操作を覧ください。

## テキスト Filtergram の操作

テキスト フィルタグラムペインには、データを動的に、陽性的中率でフィルタリングするツールが用意されています。このセクションでは、実行可能なアクションについて説明します。




**a.** データセットに表示する値をリストから選択：ペインのいずれかのテキスト値をクリックすると、データセットが動的にフィルタリングされて選択した値のみが表示されます。複数の値を選択するには: Ctrl キー（Windows）またはCommand キー（Mac）を押しながらクリックします。連続した範囲を選択するには、Shift キーを押しながらクリックします。選択を解除するには、Alt キーを押しながらクリックします。

**b. 選択された項目を表示**：フィルターにテキスト値を選択した後、**選択された項目を表示**をクリックします。新しいペインが開き、現在データセットに表示されている選択済みの範囲と値がすべて表示されます。注意: テキストフィールドの範囲は、ASCII のソート順で定義されています。このペインから以下の操作を行うことができます。

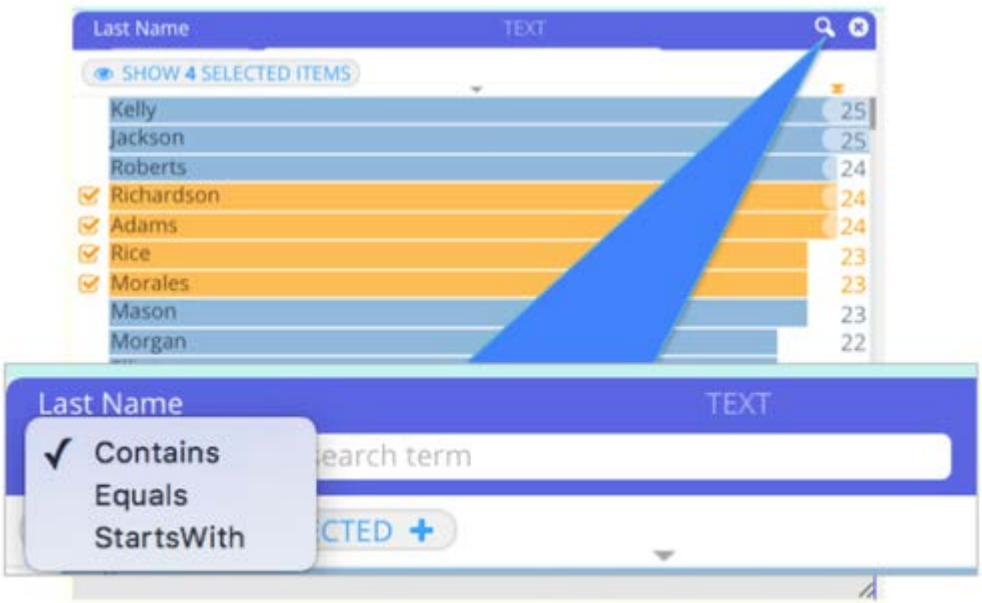
- **b1.** データセットの範囲または値を手動で削除するには、範囲または値の横に隣接する、オレンジ色のチェックマークをクリックして削除します。クリックすると、その範囲または値がフィルタリング済みデータセットに戻されます。注意: 範囲または値の選択を解除すると、ペインの上部にごみ箱のアイコンが表示されます。これを使用して、選択した部分をフィルターから削除できます。
- **b2.** 編集したい値をクリックして、ここにリストされている範囲または値を手動で更新します。それらの値が編集可能になります。新しい値を入力し、**保存**をクリックします。
- **b3. +**ボタンをクリックして、他の範囲または値を手動で追加し、データをフィルタリングします。範囲の最小値と最大値を指定するか、または値を指定して、**保存**をクリックします。データセットが動的に更新され、追加した内容が反映されます。
- **b4.** [除外] をクリックすると、データセットから範囲または値が除外されます。これは、すでに範囲を選択している場合に特に役立ちます。範囲から、現在のデータセットから非表示にする特定の値を（その範囲内から）除外することができま

す。除外対象としてマークした範囲および値は、ヒストグラムではオレンジ色の点線で囲まれ、それらが除外されていることを示していることに注意してください。

- **b5.** このペインでの作業が完了したら、リストビューに戻るために**選択した項目を非表示にする**をクリックします。

**c. クリア**して**反転**：現在のフィルターをすべて削除します。*フィルターするために選択した_もの_を除く*、すべてのデータの表示が反転します。

値を検索することもできます。これを行うには、右上隅の虫眼鏡アイコンをクリックし、検索フィールドを開きます。指定した値を含む、指定した値と一致する、または指定した値で始まる値を検索します。




# 数値 Filtergram

数値 フィルタグラム ペインを開くと、データのフィルタリング操作を実行するために使用することができる2つのビューが表示されます。

- **グラフとして表示**（デフォルトビュー）これはデータセット内の数値の分布を表す数値ヒストグラムです。水平線（x 軸）は、データセットのこの列に出現する値の範囲を表します。各バーの高さは、列に含まれる重複しない値の出現数を表します。縦線（y 軸）の値は、出現回数を示します。ヒストグラムで実行できるフィルタリング操作については、後述の数値 フィルタグラム で実行するセクションで説明します。


・**リストとして表示**このタブをクリックすると、数値ヒストグラムが非表示になり、列に含まれるすべての値のユニーク数が表示されます。リストから、データセットに動的に表示する値を選択します。リストで実行できるフィルタリング操作については、数値 フィルタグムの操作セクションで説明します。


[Filtergram] ペインの上にマウスを動かすと、次のボタンが表示されます。


**a. 現在選択されている**ボタン（左上）: ヒストグラムから選択する場合、ボタンのラベルが変更され、選択した数が表示されます。ボタンをクリックすると、ヒストグラムで現在選択されている、すべての範囲と値が一覧表示されます。このペインから、引き続きデータセットに表示する数値を絞り込むことができます。フィルターによって絞り込む範囲と値が既に分かっている場合は、ヒストグラムを使用する代わりにこのボタンをクリックします。新しいペインから値と範囲を入力して、フィルタリング操作を開始することができます。このペインから実行できるアクションについては、数値 フィルタグラム を使用するセクションで説明しています。

**b. ログ**ボタン（左下）: データの対数スケール（ログ）ビューをオンにします。デフォルトでは、データの線形表示が Filtergram に表示されます。ただし、大半のデータよりも 1 つまたは数個のポイントがはるかに大きくなる巨大な数的範囲がある場合は、ログの表示はデータ内の歪度に合わせて調整されます。

**c. ズーム表示/パン**（左下隅）：数値ヒストグラムでズームインした値と、範囲の相対的な位置を表示する概要ツールのオンとオフを切り替えます。ズームとパニング操作は、後述の数値列をグラフとして表示セクションを参照してください。

**d.** Filtergram ペインの上にマウスを動かすと、次のボタンが表示されます：

- **タイプ**：このボタンにマウスを置くと、「データセット内の数値型の行の総数」に対する「現在選択されている数値型の行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「数値型の行の総数」の比率が表示されます。このボタンをクリックすると、データセットでこれらの数値が非表示になります。これは、この列にブランク、エラー、またはその他の数値以外の値があり、それらのデータタイプのみを表示したい場合に役立ちます。
- **その他**：テキスト値など、列に数値以外の値がある場合にはこのボタンが表示されます。このボタンの上にマウスを置くと、「データセット内の数値以外の値を持つ行の総数」に対する「現在選択されている数値以外の値を持つ行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「非数値の行の総数」の比率が表示されます。**その他**をクリックするとオフになり、現在のデータセットビューから他の値を非表示にします。
- **空白**：列内に空白がある場合は、このボタンが表示されます。このボタンにマウスを置くと、「データセット内の空白行の総数」に対する「現在選択されている空白行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「空白行の総数」の比率が表示されます。このボタンをクリックしてオフに切り替えると、現在のデータセットビューからブランクセルが非表示になります。

- **エラー**：列内にエラーがある場合は、このボタンが表示されます。このボタンにマウスを置くと、「データセット内のエラーのある行の総数」に対する「現在選択されているセルエラーのある行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「セルエラーのある行の総数」の比率が表示されます。このボタンをクリックしてオフに切り替えると、現在のデータセットビューからセルエラーが非表示になります。


列内に「その他」の値、空白、またはエラーがある場合は、その値に対応するボタンの横に**+**ボタンも表示されます。**+**ボタンをクリックすると、そのタイプのすべての値 （すべての「その他」の値など）が [選択した項目] リストに追加され、このリストでさらにフィルタリング操作を続けることができます。**選択した項目**から実行するフィルタリング操作については、数値列をグラフとして表示を参照してください。

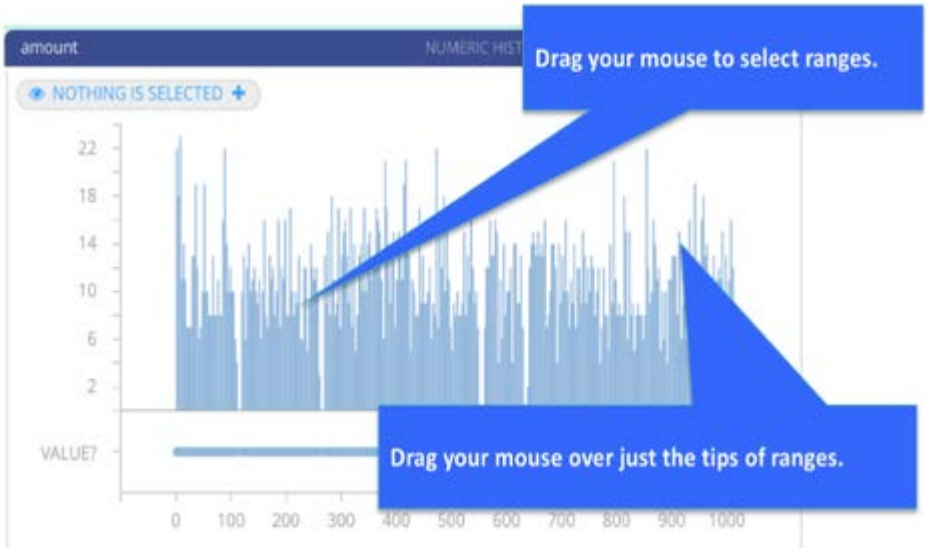
## 数値 Filtergram の操作

数値 フィルタグラム は、高い陽性的中率でデータを動的にフィルタリングするツールを提供します。このセクションでは、実行可能なアクションについて説明します。

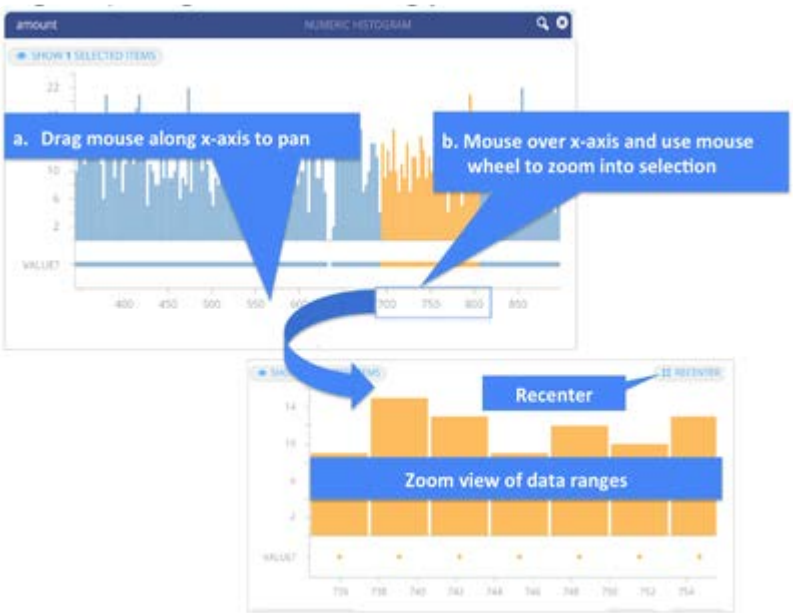
### 数値列をグラフとして表示

- ヒストグラムで表示する範囲を選択: マウスをクリック アンド ドラッグして値の範囲を選択すると、選択した範囲に合わせてデータセットが更新されます。ヒストグラムで連続していない範囲を追加で選択するには、Ctrl キー（Windows）またはCommand キー（Mac）を押しながらクリックしてマウスをドラッグしながら操作します。選択範囲または選択範囲の一部を解除するには、Alt キーを押しながらクリックしてマウスをドラッグします。

  また、ヒストグラムの範囲の先端にマウスでドラッグすると、データセットにそれらの値のみが表示されます。y 軸は、データのピークの相対値を確認するのに役立ちます。

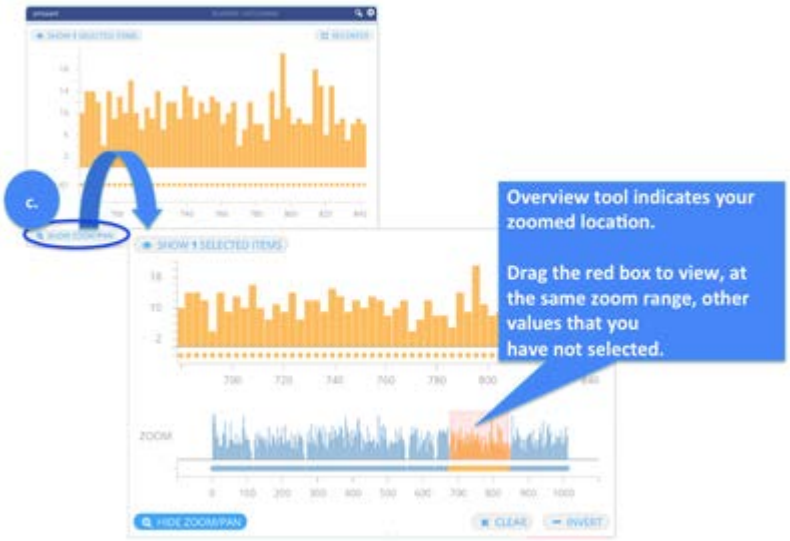



- データの探索と変換を開始:




**a. ヒストグラムのパン**：x 軸の値の上にマウスを置きます。カーソルがポインターからクリック アンド ドラッグの形に変わります。x軸上の値をマウスでクリックし、ヒストグラムおよび選択範囲上のパンする範囲をドラッグします。ヒストグラムをデフォルトの表示に戻すには、**リセンター**をクリックします。選択範囲は維持されていることに注意してください。

**b. 選択範囲のズーム**：x 軸の値の上にマウスを置き、マウスホイールを使用して選択範囲をズームします。Mac では、2 本の指を下にドラッグして選択の範囲をズームインし、2 本の指を上にドラッグしてズームアウトします。ズームインしながら、範囲の選択を続けることができます。ヒストグラムをデフォルトの表示に戻すには、**リセンター**をクリックします。ヒストグラムが初期化された後も選択範囲は維持されていることに注意してください。




**c. ズームした範囲の相対位置の表示**：選択範囲にズームインした後、その範囲がヒストグラム全体のどこに位置しているかを同時に表示できます。**ズーム/パン**をクリックすると、2 つ目の概要ツールのヒストグラムが下に表示されます。概要ツールにはデータの範囲全体が表示され、現在ズームしている範囲が赤のボックスで示されます。概要ツールのヒストグラムから、以下のこともできます:

・赤のボックスをドラッグし、選択していない他の値を同じズーム範囲で表示する。

- 概要ツールとメインのヒストグラムの両方をインタラクティブに操作する。概要ツールのヒストグラムに対する操作はプライマリーヒストグラムにのみ反映されることに注意してください。データを動的にフィルタリングするには、プライマリーのヒストグラムで選択を行う必要があります。


**d. 選択した項目を表示**：選択したデータ範囲のフィルタリングした後、**選択した項目を表示**をクリックします。新しいペインが開き、そこでデータセットに表示する正確な範囲と値を指定します。このペインから以下の操作を行うことができます。

- **d1.)** データセットの範囲または値を手動で削除するには、範囲または値の横に隣接する、オレンジ色のチェックマークをクリックして削除します。クリックすると、その範囲または値がフィルタリング済みデータセットに戻されます。注意: 範囲または値の選択を解除すると、ペインの上部にごみ箱のアイコンが表示されます。これを使用して、選択した部分をフィルターから削除できます。
- **d2.)** 編集したい値をクリックして、ここにリストされている範囲または値を手動で更新します。それらの値が編集可能になります。新しい値を入力し、**保存**をクリックします。
- **d3.) +**ボタンをクリックして、データをフィルタリングする範囲または値を手動に追加します。範囲の最小値と最大値を指定するか、または値を指定して、**保存**をクリックします。その範囲または値のエントリが作成されます。エントリのチェックボックスをクリックすると、データセットが動的に更新され、選択範囲が反映されます。
- **d4.)** データセットから範囲または値を除外します。これは、すでに範囲を選択している場合に特に役立ちます。現在のデータセットから非表示にし、特定の値（その範囲内から）を範囲から除外することができます。例えば、データセットに表示するために1〜2000の範囲を選択します。次に、値195を除外します。データセットは、値 195 を除く 1〜2000 のすべてを表示します。EXCLUDE でマークした範囲と値はオレンジで表示され、その除外を示すために、ヒストグラムにおいて点線のアウトラインで表示されることに注意してください。

. **d5.) 選択の非表示**をクリックすると、ペインをオフにし、選択した部分がヒストグラムで強調表示されるヒストグラムに戻ります。

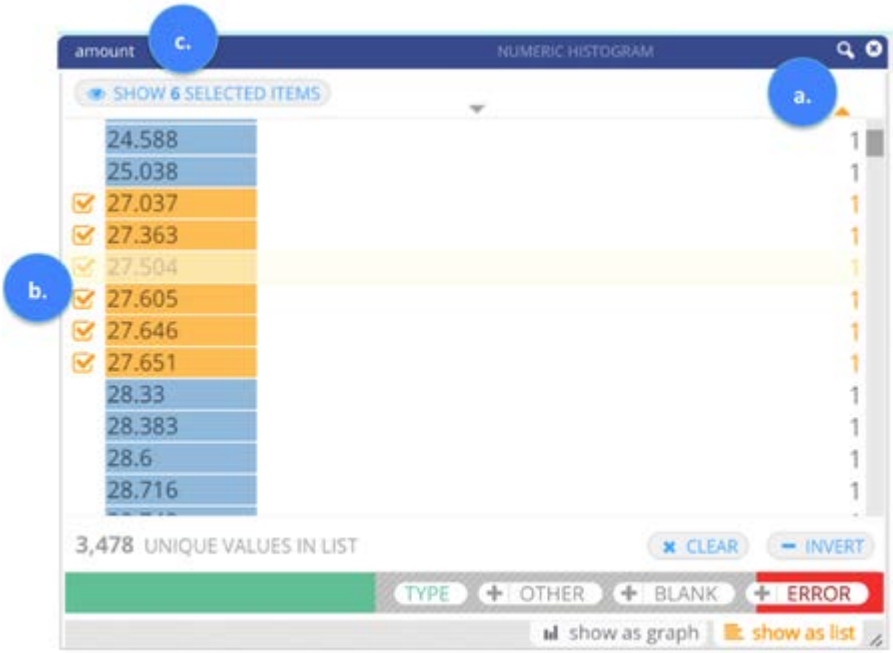

**e. 選択したデータを反転**：フィルタリングするために選択したデータを除くすべてのデータを表示します。

**f. クリア**：現在のフィルターをすべて削除します。

**g. 値の検索**：右上隅にある拡大鏡アイコンをクリックすると、検索フィールドが開きます。指定した値を含む、指定した値と一致する、または指定した値で始まる値を検索できます。

## 数値列をリストとして表示

リストビューでは、カラム内のすべての数値の出現数を確認できます。この形式でデータを表示することは、特定の数値をすばやく選択してフィルタリングしたい場合に特に便利です。**グラフとして表示**タブをクリックした場合、[リスト] 内の選択がヒストグラムに表示されることに注意してください。このセクションでは、実行できるアクションについて説明します。




**a. リスト順の並び替え：** デフォルトでは、値のリストが最小カウントから最高カウントに表示されます。順序を逆にするには、右上隅の出現数の列の上にある三角形をクリックします。リストの上にある三角形をクリックして数値の順に並べ替えることもできます。三角形のオレンジ色は、現在データセットに適用されている並び替え順（発生順または数値順）を示しています。

**b. データセットに動的に表示する値を選択**：クリックして、リストから値を選択します。複数の値を選択するには: Ctrl キー（Windows）またはCommand キー（Mac）を押しながらクリックします。連続した複数行の範囲を選択するには、Shift キーを押しながらクリックします。選択を解除するには、次のキーコマンドを使用して、Alt キーを押しながらクリックします。選択した後、左上隅にある**現在の選択**をクリックすると、フィルタリング操作の精度をより継続できる新しいペインが開きます。データセットのフィルタリングに使用する範囲や値が正確に分かっている場合は、リストから選択する手順をスキップしてもかまいません。代わりに、**選択した項目がない**をクリックすると、正確な値と範囲を入力する新しいペインが開きます。

**c. 選択した項目を再度表示**：選択した後、左上隅にある**選択した項目を表示**をクリックすると、フィルタリング操作の再表示を継続できる新しいペインが開きます。ボタンのラベルには、現在選択されている項目数が反映されます。

データセットのフィルタリングに使用する範囲や値が正確に分かっている場合は、リストから選択する手順をスキップしてもかまいません。この場合、ボタンのラベルは**アイテムが選択されていません**になります。このボタンをクリックして新しいペインを開き、そこで正確な値と範囲を入力します。ペインから実行できるフィルタリング操作については、数値列をグラフとして表示のステップ d1-d5 で説明します。また、選択した部分を**反転**および**クリア**し、リストから特定の値を検索することもできます。詳細については、同じセクションの手順 e-g を参照してください。

# 日付と時刻 Filtergram

日付/時刻 フィルターグラム ペインを開くと、次の 2 つのビューが表示され、 これらのビューを使用して、データに対するフィルタリング操作を行うことができます:

- **グラフとして表示**（デフォルトビュー）


これはデータセット内の日付/時刻の値の分布を表すヒストグラムです。水平線（x 軸）は、データセットのこの列に出現する日付値の範囲を表します。各バーの高さは、列に含まれる重複しない日付値の出現数を表します。縦線（y 軸）の値は、出現回数を示します。ヒストグラムで実行できるフィルタリング操作は、後述の日付と時刻の列をグラフとして表示セクションで説明します。

- **リストとして表示**このタブをクリックすると、日付/時刻のヒストグラムが非表示になり、列に含まれるすべての日付/時刻の値のユニーク数が表示されます。リストから、データセットに動的に表示する値を選択します。リストで実行できるフィルタリング操作は、後述の日付と時刻の列をリストとして表示セクションで説明します。


## 日付と時刻 Filtergram

[Filtergram] ペインの上にマウスを動かすと、次のボタンが表示されます。


**a. 現在選択されている**（左上隅）: ヒストグラムから選択する場合、ボタンのラベルが変更され、選択した数が表示されます。ボタンをクリックすると、ヒストグラムで現在選択されている、すべての範囲と値が一覧表示されます。このペインから、引き続きデータセットに表示する日付/時刻を絞り込むことができます。フィルターによって絞り込む範囲と値が既に分かっている場合は、ヒストグラムを使用する代わりにこのボタンをクリックします。新しいペインから日付と日付の範囲を入力して、フィルタリング操作を開始することができます。このペインから実行できるアクションについては、後述の日付と時刻列をグラフとして表示セクションと、日付と時刻の列をリストとして表示セクションで説明します。

**b. ログ**（左下隅）: データの対数スケール（ログ）ビューをオンにします。デフォルトでは、データの線形表示が Filtergram に表示されます。ただし、大半のデータよりも 1 つまたは数個のポイントがはるかに大きくなる範囲がある場合、ログの表示はデータ内の歪度に合わせて調整されます。

**c. ズーム表示/パン**（左下隅）: ヒストグラムでズームインした値と、範囲の相対的な位置を表示する概要ツールのオンとオフを切り替えます。ズーミングとパニング操作については、日付と時刻列をグラフとして表示セクションで説明します。

**d.** Filtergram ペインの上にマウスを動かすと、次のボタンが表示されます：

- **タイプ**：このボタンにマウスを置くと、「データセット内の日付/時刻型の行の総数」に対する「現在選択されている日付/時刻型の行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「日付/時刻型の行の総数」の比率が表示されます。このボタンをクリックすると、データセットで日付/時刻の値が非表示になります。これは、この列に空白、エラー、またはその他の日付/時刻以外の値があり、それらのデータタイプのみを表示したい場合に役立ちます。
- **その他**：列に日付/時刻以外の値がある場合は、このボタンが表示されます。このボタンにマウスを置くと、「データセット内の日付/時刻以外の値を持つ行の総数」に対する「現在選択されている日付/時刻以外の値を持つ行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「日付/時刻以外の値を持つ行の総数」の比率が表示されます。**その他**をクリックするとオフになり、現在のデータセットビューから他の値を非表示にします。

- **空白**：列内に空白がある場合は、このボタンが表示されます。このボタンにマウスを置くと、「データセット内の空白行の総数」に対する「現在選択されている空白行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「空白行の総数」の比率が表示されます。**ブランク**をクリックするとオフになり、現在のデータセットビューからブランクセルを非表示にします。
- **エラー**：列内にエラーがある場合は、このボタンが表示されます。このボタンにマウスを置くと、「データセット内のエラーのある行の総数」に対する「現在選択されているセルエラーのある行の数」の比率が表示されます。ヒストグラムで何も選択されていない場合は、「データセット内の行の総数」に対する「セルエラーのある行の総数」の比率が表示されます。エラーをクリックするとオフになり、現在のデータセットビューからセルのエラーを非表示にします。


列内に「その他」の値、空白、またはエラーがある場合は、その値に対応するボタンの横に**+**ボタンも表示されます。**+**ボタンをクリックすると、そのタイプのすべての値 （すべての「その他」の値など）が [選択した項目] リストに追加され、このリストでさらにフィルタリング操作を続けることができます。**選択した項目**から実行するフィルタリング操作については、以下を参照してください。

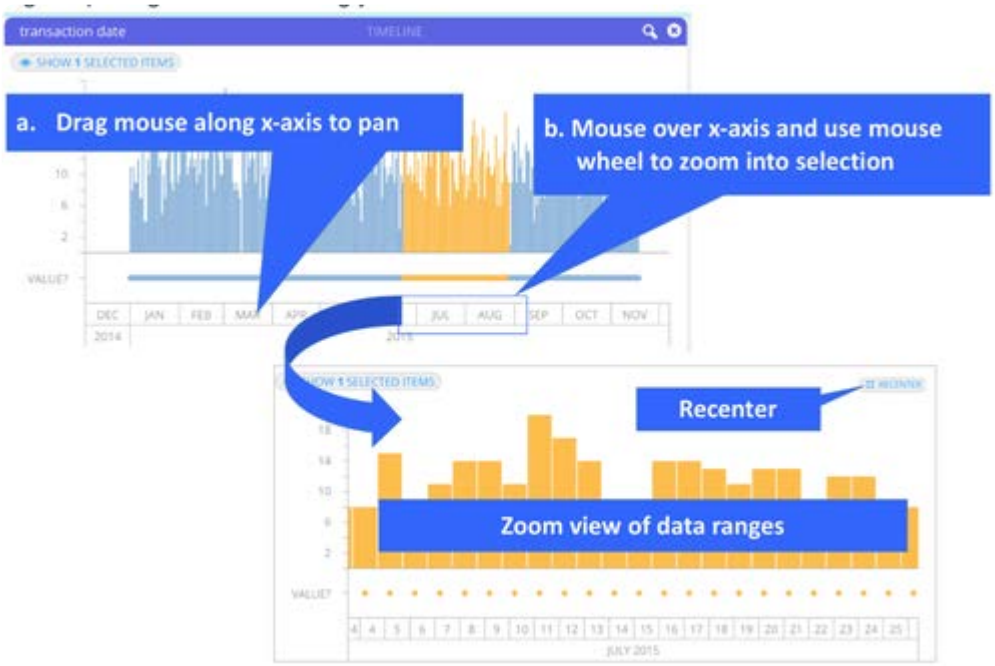
日付/時刻のヒストグラムには、データを高い精度で動的にフィルタリングできる強力なツールが用意されています。このセクションでは、実行可能なアクションについて説明します。

## 日付と時刻列をグラフとして表示

- ヒストグラムで表示する範囲を選択: マウスをクリック アンド ドラッグして値の範囲を選択すると、選択した範囲に合わせてデータセットが更新されます。ヒストグラムで連続していない範囲を追加で選択するには、Ctrl キー（Windows）またはCommand キー（Mac）を押しながらクリックしてマウスをドラッグしながら操作します。選択範囲または選択範囲の一部を解除するには、Alt キーを押しながらクリックしてマウスをドラッグします。

  また、一定範囲のバーの先端部のみをマウスでドラッグすると、データセットにそれらの値のみが表示されます。y 軸は、データのピークの相対値を確認するのに役立ちます。


- データの探索と変換を開始:




**a. ヒストグラムのパン**：x 軸の値の上にマウスを置きます。カーソルがポインターからクリック アンド ドラッグの形に変わります。x軸上の値をマウスでクリックし、ヒストグラムおよび選択範囲上のパンする範囲をドラッグします。ヒストグラムをデフォルトの表示に戻すには、**リセンターボタン**をクリックします。選択範囲は維持されていることに注意してください。

**b. 選択範囲のズーム**：x 軸の値の上にマウスを置き、マウスホイールを使用して選択範囲をズームします。Mac では、2 本の指を下にドラッグして選択の範囲をズームインし、2 本の指を上にドラッグしてズームアウトします。ズームインしながら、範囲の選択を続けることができます。ヒストグラムをデフォルトの表示に戻すには、**リセンター**をクリックします。ヒストグラムが初期化された後も選択範囲は維持されていることに注意してください。


**c. ズームした範囲の相対位置の表示：**選択範囲にズームインした後、その範囲がヒストグラム全体のどこに位置しているかを同時に表示できます。[ズーム/パン] ボタンをクリックすると、2 つ目の概要ツールのヒストグラムが下に開きます。概要ツー

ルにはデータの範囲全体が表示され、現在ズームしている範囲が赤のボックスで示されます。概要ツールのヒストグラムから、以下のこともできます。

- 赤のボックスをドラッグし、選択していない他の値を同じズーム範囲で表示する。
- 概要ツールとメインのヒストグラムの両方をインタラクティブに操作する。概要ツールのヒストグラムに対する操作はプライマリーヒストグラムにのみ反映されることに注意してください。データを動的にフィルタリングするには、メインのヒストグラムで選択を行う必要があります。

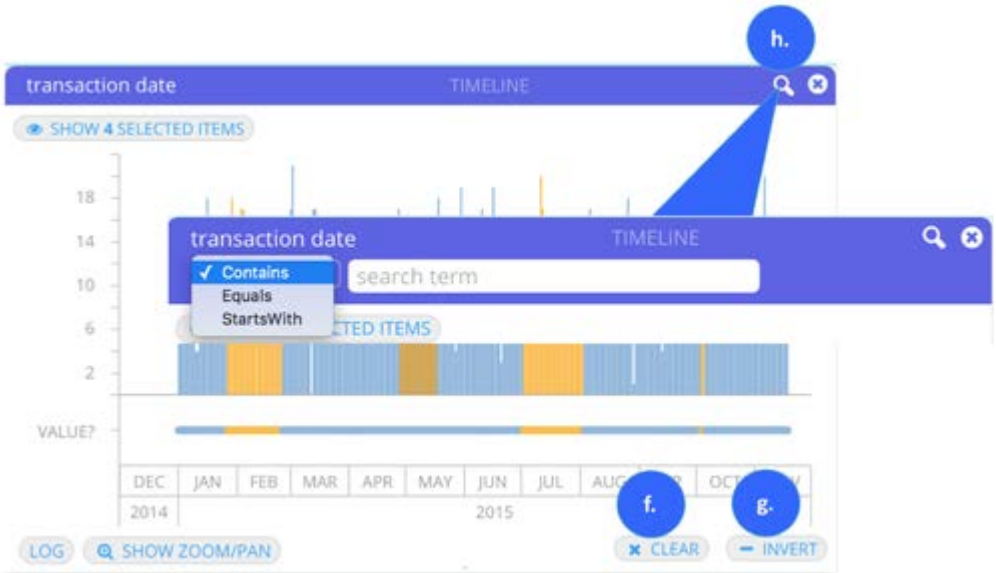

**d. 選択した項目を表示：** 選択した範囲をフィルタリングした後、**選択した項目を表示**します新しいペインが開き、現在データセットに表示されている選択済みの範囲と値がすべて表示されます。このペインから、データセットから範囲または値を除外することができます。これは、すでに範囲を選択している場合に特に役立ちます。現在のデータセットから非表示にし、特定の値（その範囲内から）を範囲から除外することができます。例えば、データセットに表示する日付範囲を次のように選択しました:2015年3月1日～2015年3月15日。 次に、日付 2015年3月10日を除外します。データセットには、2015年3月10日を除く範囲内のすべてが表示されます。

除外対象としてマークした範囲および値は、ヒストグラムではオレンジ色の点線で囲まれ、それらが除外されていることを示していることに注意してください。

範囲または値の選択を解除すると、ペインの上部にごみ箱のアイコンが表示されます。これを使用して、選択した部分をフィルターから削除できます。

**選択の非表示**をクリックすると、ペインをオフにし、選択した部分と除外した部分がヒストグラムで強調表示されるヒストグラムに戻ります。


**e. 利用可能なチャート：**このタブをクリックすると、4つの追加フィルター（年月、月、週、日）から選択して、陽性的中率で日付/時刻データをフィルタリングできます。フィルターを選択した後、[使用可能な 5 つのチャート] タブをもう一度クリックすると、そのフィルターがビューに固定されます。これらのフィルターで作業している間、選択した内容を反映してデータセットが動的に更新されます。



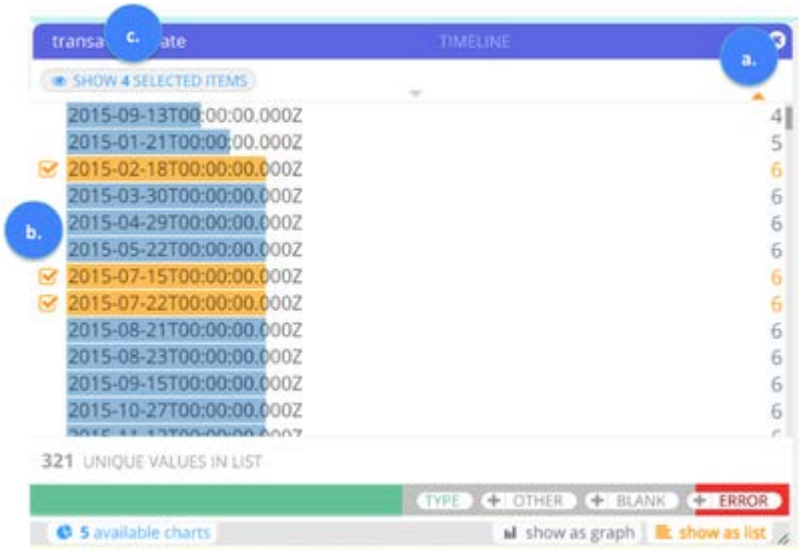

**f. クリア：**現在のフィルターをすべて削除します。

**g. 選択したデータを反転**：フィルタリングするために選択したデータを除くすべてのデータを表示します。

**h. 値の検索**：右上隅にある拡大鏡アイコンをクリックすると、検索フィールドが開きます。指定した値を含む、指定した値と一致する、または指定した値で始まる値を検索できます。

## 日付と時刻列をリストとして表示

リスト形式のフィルタービューでは、カラム内のすべての日付/時刻の値の出現数を確認できます。この形式でデータを表示することは、特定の日付をすばやく選択してフィルタリングしたい場合に特に便利です。**グラフとして表示**タブをクリックした場合、[リスト] 内の選択がヒストグラムに表示されることに注意してください。このセクションでは、実行できるアクションについて説明します。




**a. リスト順の並び替え：** デフォルトでは、データのリストが最小カウントから最大出現に表示されます。順序を逆にするには、右上隅の出現数の列の上にある三角形をクリックします。日付/時刻の値の上にある三角形をクリックして時系列順に並べ替えることもできます。三角形のオレンジ色は、現在データセットに適用されている並べ替え順序（発生または時系列）を表示します。

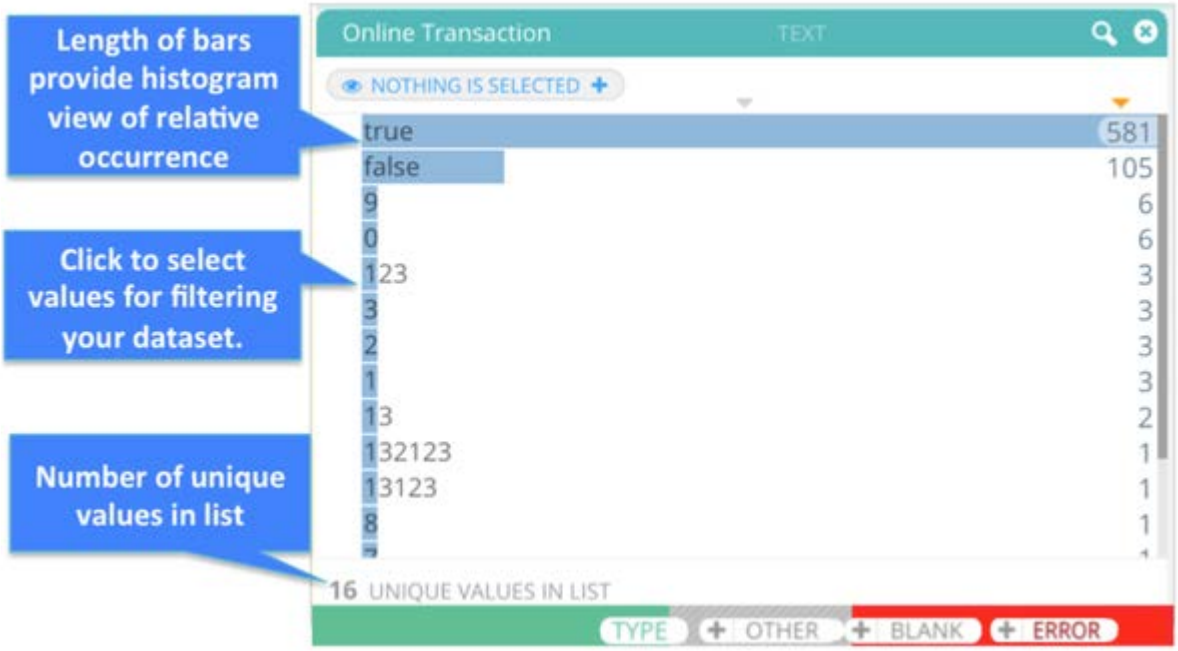
**bデータセットに動的に表示する値を選択：** クリックして、リストから値を選択します。複数の値を選択するには: Ctrl キー（Windows）またはCommand キー（Mac）を押しながらクリックします。連続した複数行範囲を選択するには、Shift キーを押しながらクリックします。選択を解除するには、次のキーコマンドを使用して、Alt キーを押しながらクリックします。選択した後、左上隅にある**現在の選択**をクリックすると、フィルタリング操作の精度をより継続できる新しいペインが開きます。データセットのフィルタリングに使用する範囲や日付が正確にわかっている場合は、リストから選択する手順をスキップしてもかまいません。代わりに、**選択した項目がない**をクリックすると、正確な値と範囲を入力する新しいペインが開きます。

**c. 選択した項目を再度表示：**選択した後、左上隅にある**選択した項目を表示**をクリックすると、フィルタリング操作の再表示を継続できる新しいペインが開きます。ボタンのラベルには、現在選択されている項目数が反映されます。

ペインから実行するフィルタリング操作は、日付と時刻列をグラフとして表示セクションのステップ d で説明します。また、選択した部分を**反転**および**クリア**し、リストから特定の値を検索することもできます。同じセクションの手順 f-g を参照してください。

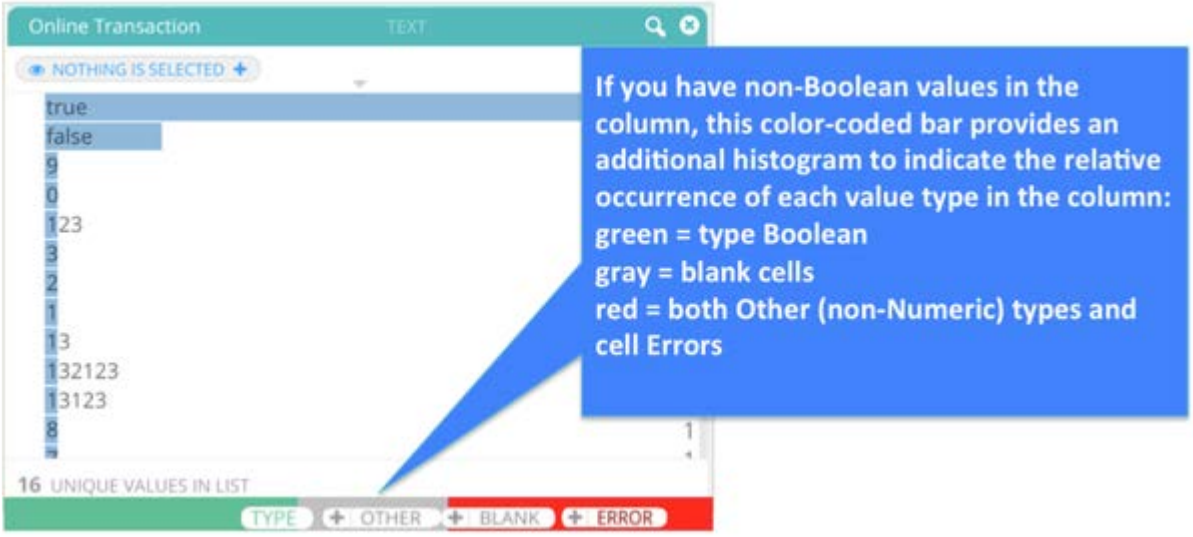
# ブール値 Filtergram

ブール値の フィルタグラムでは、データセット内のブール値の出現数を表示し、その他の値をデータセットからフィルターで除外できます。




左から右へ伸びるバーは、それぞれの値の相対的な出現回数のヒストグラムを示します。ユニーク数を総数に表示する値の総数がペインの左下に表示されます。リストから、データセットに動的に表示する値を選択できます。

ブール値 フィルタグラム 表示の概要を以下に示します:



## ブール値 Filtergram の操作

[Filtergram] ペインの上にマウスを動かすと、次のボタンが表示されます。

- **タイプ**：このボタンにマウスを置くと、この列でのブール値の出現数が表示されます。このボタンをクリックすると、列内のブール値が非表示になります。
- **その他**：列にブール値以外の値が含まれている場合は、このボタンが表示されます。このボタンにマウスを置くと、この列のブール値以外の値の出現数が表示されます。このボタンをクリックすると、この列のブール値以外の値が非表示になります。また、**+**ボタンをクリックすると、「その他」のすべての値が Filtergram リストに追加されます。その後、現在のデータセットビューから特定の「その他」の値を除外して非表示にすることができます。そのためには、Alt+CRTL キー（ウィンドウズ）またはAlt+Command キー（Mac）を押しながら、非表示にする「その他」の値をクリックします。
- **空白**：列内に空白がある場合は、このボタンが表示されます。このボタンにマウスを置くと、この列の空白値の出現数が表示されます。このボタンをクリックすると、列内の空白値が非表示になります。
- **エラー**：列内にエラーがある場合は、このボタンが表示されます。このボタンにマウスを置くと、この列のセルエラーの出現数が表示されます。このボタンをクリックすると、列内のセルエラーが非表示になります。

**クリア**と**反転**を使用して、Filtergram リストで選択を管理します。**反転**はリストで選択した値を_除く_、現在のデータセットビューにある値 すべてを表示していることに注意してください。

# ソース Filtergram

ソース フィルタグラム では、ルックアップ データセットおよび追加したデータセットの行がプロジェクトの基本データセットにどのように関係しているかを表示できます。

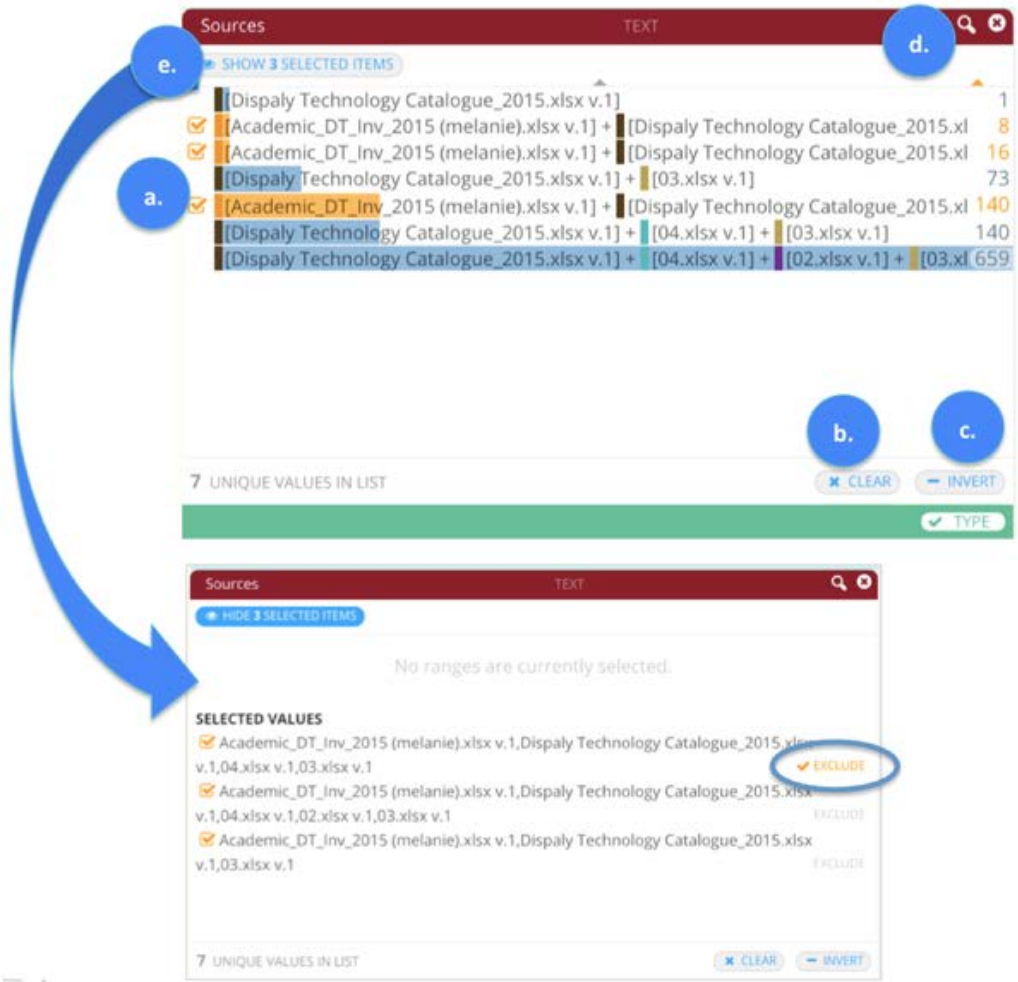
## ソース Filtergram の例

基本データセットと一致しないルックアップ データセットの行を含む外部結合があるとします。ソース Filtergram を使用すると、この結合に関係している基本データセットの行数とルックアップ データセットの行数がわかります。さらに、各データセットソースの一致していない行数も表示されます。

以下にソース フィルタグラム 表示の概要例を示します:


デフォルトでは、ソースのリストは出現数の多い順に並んでいます。順序を逆にするには、右上隅の出現数の列の上にあるオレンジ色の三角形をクリックします。リストの上にある三角形をクリックしてアルファベット順に並べ替えることもできます。三角形のオレンジ色は、現在リストに数字とアルファベットのどちらの並び順が適用されているかを示します。

# ソース Filtergram の操作




**a.** データセットに表示するソースを選択：いずれかのソースをクリックすると、データセットが動的にフィルタリングされてそのソースのみが表示されます。複数のソース を選択するには: Ctrl キー（Windows）またはCommand キー（Mac）を押しながらクリックします。連続した複数行範囲を選択するには、Shift キーを押しながらクリックします。

**b. クリア**：すべてのソースフィルターの選択を削除します。

**c.** 選択を反転：選択したものを除くすべてのソースを表示します。

**d.** ソースファイルの検索: 右上隅にある虫眼鏡アイコンをクリックすると、検索フィールドが開きます。指定したテキスト値を含む、指定したテキスト値と一致する、または指定したテキスト値で始まるソースファイルを検索できます。

**e.** フィルター選択の調整：**選択した項目を表示**をクリックします。新しいペインが開き、現在選択されているソースが表示されます。このペインから以下の操作を行うことができます。

- ソースの除外:これは、再度除外をクリックするまで、関連付けられたソースをデータセットから非表示にするトグルです。除外対象としてマークされたソースは、最初の Filtergram ペインではオレンジ色の点線で囲まれ、それらが除外されていることを示していることに注意してください。
- ソースの選択を解除すると、ペインの上部にごみ箱のアイコンが表示されます。これを使用して、ソースをフィルターから削除できます。

- このペインでの作業が完了したら、最初の Filtergram ビューに戻るために**選択した項目を非表示にする**をクリックします。

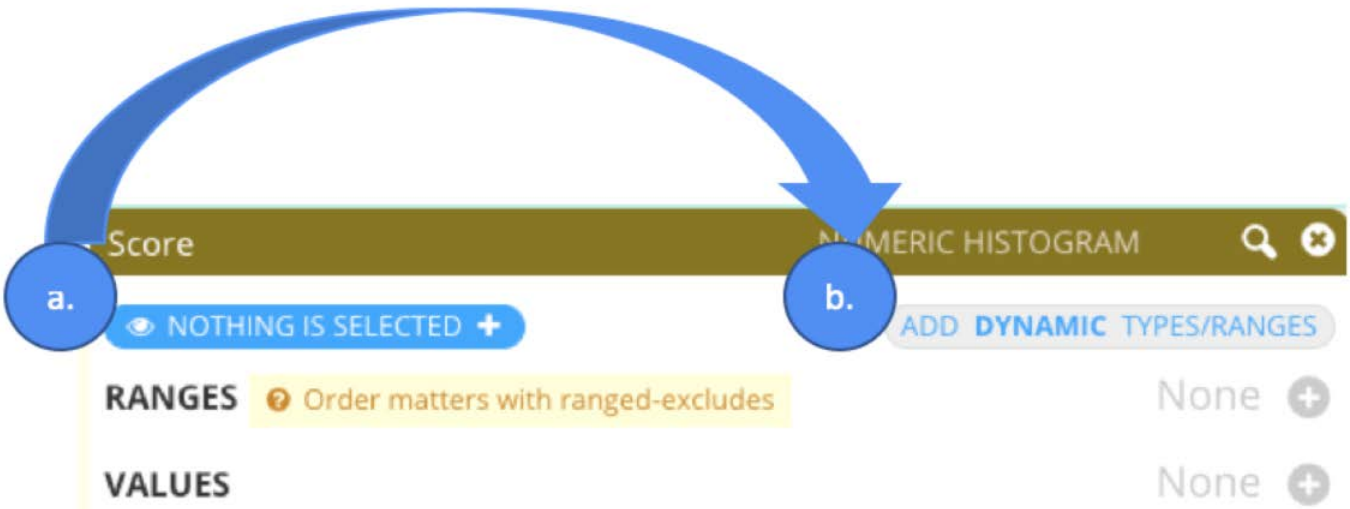
# ダイナミック レンジ

Filtergram の動的百分位数機能では、選択値の百分位数を指定する強力なオプションを利用できます。たとえば、販売された製品の地域および週単位の列があるインベントリデータセットがあれば、動的百分位数を使用して、地域ごとに販売されたトップ 5% の製品をフィルタリングして選択することができます。パーセンタイルの選択範囲は最新バージョンのデータセットにも動的に適用され、それらは自動プロジェクトフロー（APF）を通してライブラリで自動的に更新されます。たとえば、動的百分位数と APF 機能を使用して、AnswerSet を毎週自動的に生成し、地域および週単位で販売された上位 5% の製品を確認することができます。

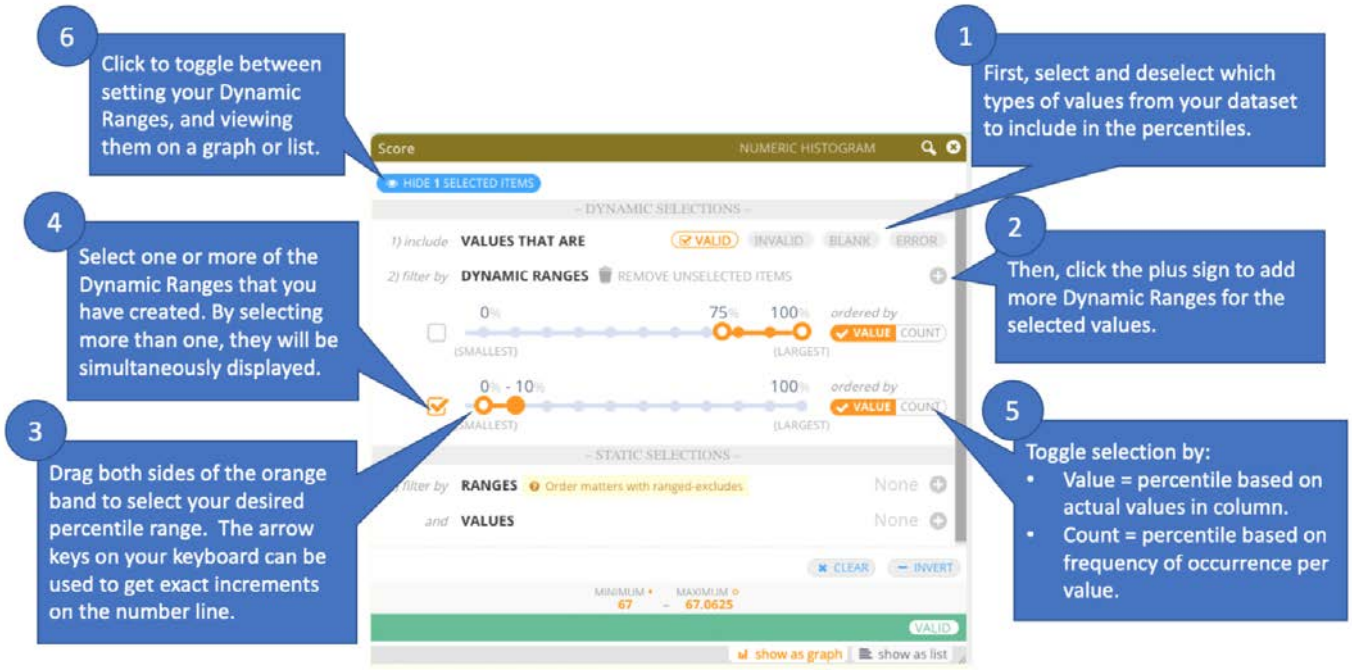
動的フィルタリングオプションは、日付/時刻型、文字列型、数値型の列に適用できます。動的フィルタリングペインを開くには:

**(a.)** Filtergram の左上隅にある**現在の選択**をクリックします。

**(b.) ダイナミックタイプ/レンジを追加**をクリックします。




ダイナミクスペインが開きます:




## 動的選択の操作

**1.** パーセンタイル範囲に含める値のタイプを選択または選択解除：検証、無効、空白、エラー:

- 検証：数値タイプの列など、列のタイプと同じタイプの値。
- 無効：数値タイプの列のアルファベット文字など、列タイプと同じタイプではない値。
- 空白: 列に空白がある場合は使用します。
- エラー: 列内にエラーがある場合は使用します。

**2. +**ボタンをクリックして、希望の数のダイナミックレンジを追加します。

> **備考**
>
> これらの範囲はそれぞれ手順1で選択した値の型にのみ適用されます。

**3.** ダイナミックレンジを設定: 数直線上の境界をドラッグし、希望の値を設定します。

> **ヒント**
>
> キーボードの左と右の矢印キーを使用して、数値線の正確な値を調整します。

**4.** （オプション）: 範囲選択をさらに追加します（上記の手順 1 から 3 を行います）。複数の範囲選択を作成すると、各選択範囲は AND 演算として扱われます。たとえば、2 つの選択範囲（上位 10 番目のパーセンタイルを選択する範囲と下位 10 番目のパーセンタイルを選択する範囲）を作成した場合は、これらのパーセンタイルと一致する値が Filtergram 上でハイライト表示され、それに応じてデータグリッドに表示されます。

**備考**

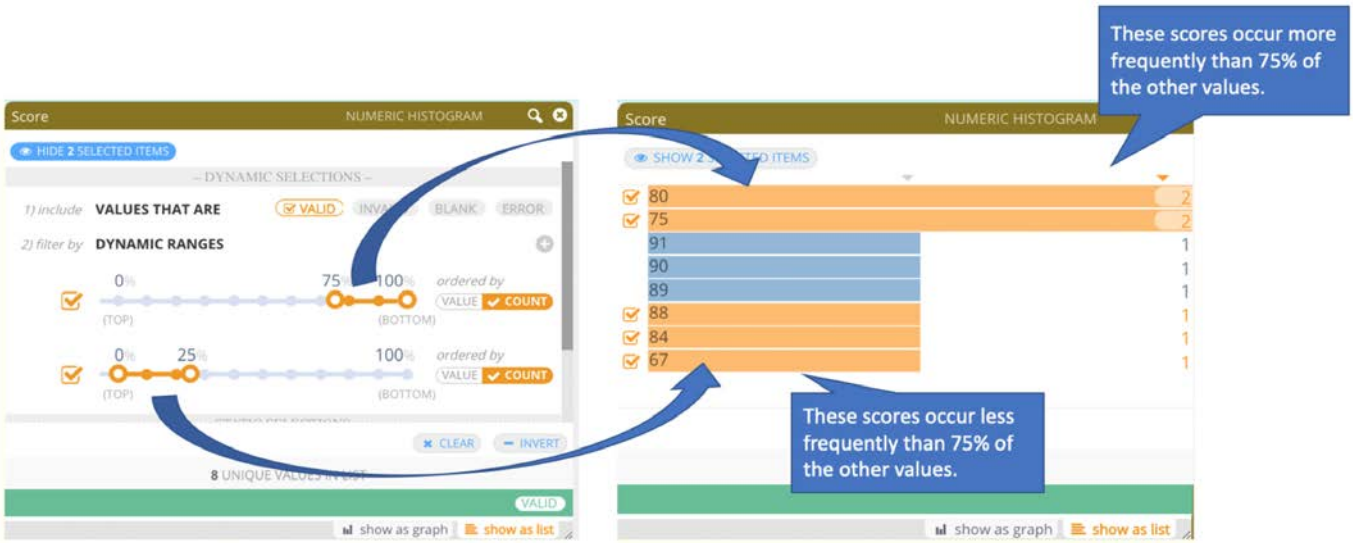
ダイナミックフィルタリングオプションは、列タイプごとに、このドキュメントで説明されている他のフィルタリング操作と常に連動するように設定することができます。

**5. Value（値）**または**Count（カウント）**順に並べる：**Value（値）** 順に並べた場合は、パーセンタイルは列内の*実測値*に基づいて算出されます。データを **Count（カウント）** 順に並べた場合は、パーセンタイルは値ごとの*出現頻度*に基づいて算出されます。たとえば、試験の参加者の年齢とスコアを含む、以下のようなデータセットがあるとします。

| | Sources | Name | Age | Score |
|---|---|---|---|---|
| 1 | | Dennis | 32 | 91 |
| 2 | | Sara | 30 | 90 |
| 3 | | Jess | 35 | 89 |
| 4 | | Mark | 25 | 88 |
| 5 | | Rick | 36 | 84 |
| 6 | | Grace | 31 | 80 |
| 7 | | Jack | 35 | 80 |
| 8 | | Yana | 34 | 75 |
| 9 | | Erin | 26 | 75 |
| 10 | | Max | 36 | 67 |

各参加者が相対的にどのようにスコアを得ているかを確認するには、**Value（値）** 順に並べます。

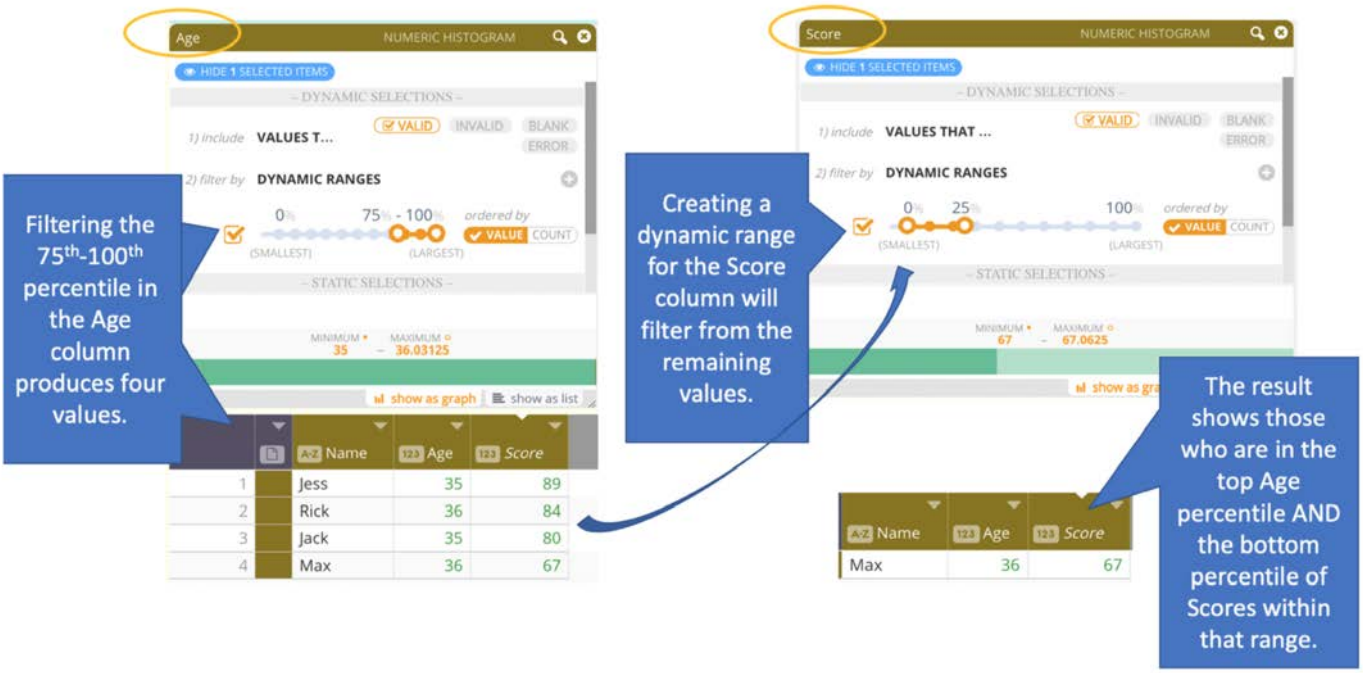

検査自体の妥当性と有用性を確認したり、検査結果のパターンを表示させたりするには、**Count（カウント）** 順に並べます。




**6. 現在の選択**をクリックすると、グラフまたはリストビューで強調表示されたパーセンタイルが表示されます。

## 列全体の複数の Filtergram パターン

複数列の Filtergram は同時に開いて動的にフィルタリングできます。複数の Filtergram 解像度は左から右となっています。左側にダイナミックレンジが設定されている場合は、右側の隣接した Filtergram によって選択される結果値に影響することに注意してください。




左の Filtergram を抜けると、その列に適用されているダイナミックレンジは削除されます。

**備考**

- もしインタラクティブモード機能が有効になっている場合は、選択範囲はデータセット全体に適用されます。
- ダイナミック レンジは、有効にする必要がある機能です。このボタンがプロジェクトに表示されない場合は、Data Prep システム管理者に連絡してください。

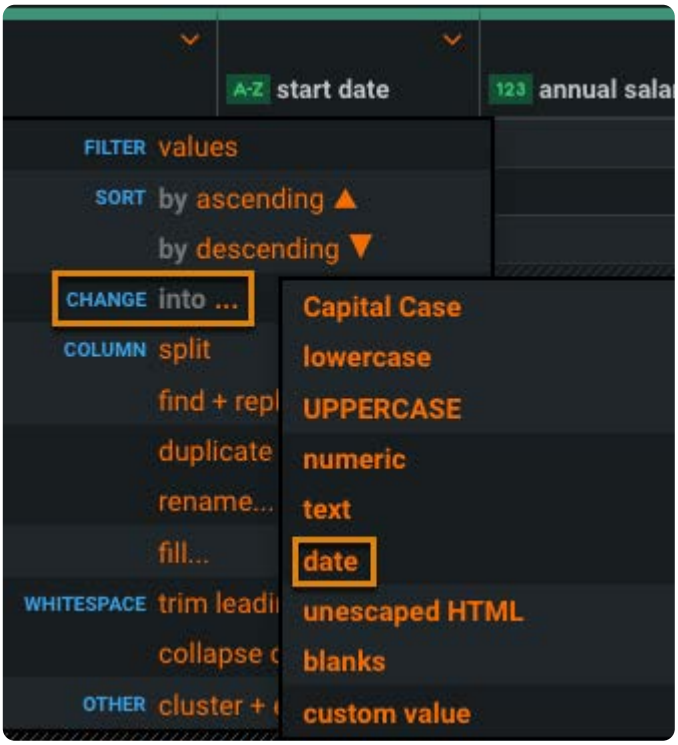
# 日付形式の検出および変換

Data Prepは、列内の日付形式を検出し、列内のセルをISOの日付形式に変換することができます。日付形式の問題は、トレーニングデータセットの準備に必要な元のデータで多く見られます。多くの場合、値は文字列や数値形式であり、それらを日付形式としてキャストする必要があります。例えば、20210101のような形式を2021年1月1日のような形式に変換したいとします。

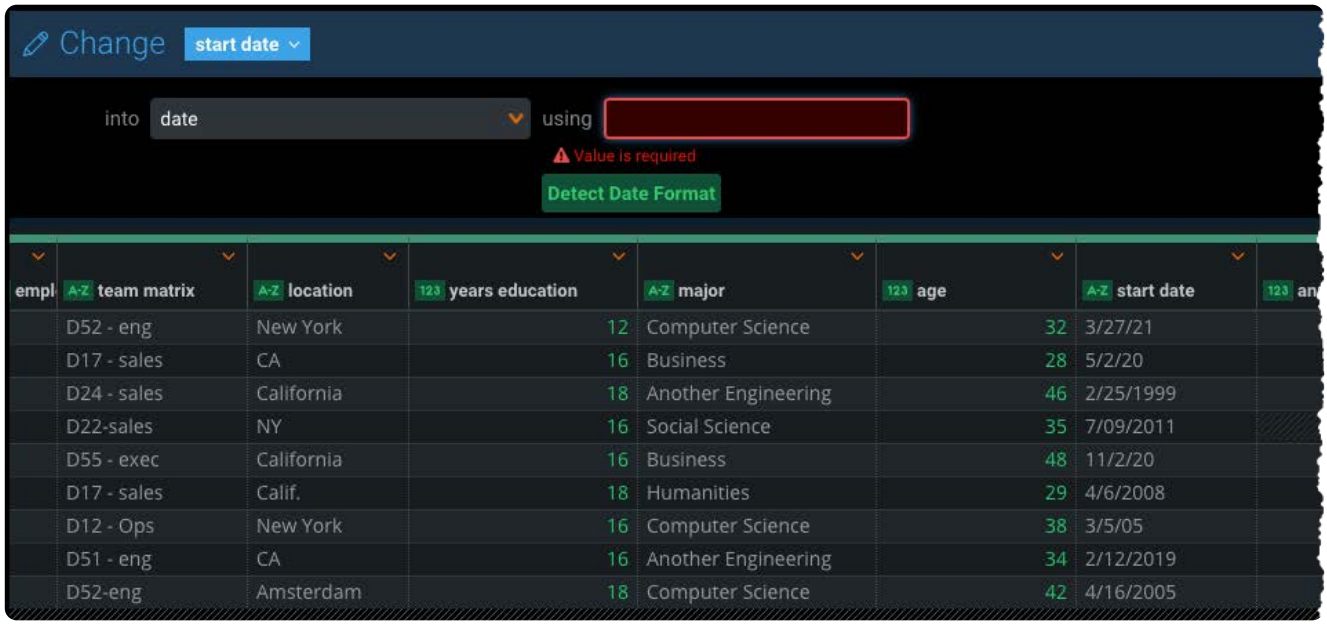
## 日付の形式化

データセットの列において、日付を形式化する手順：

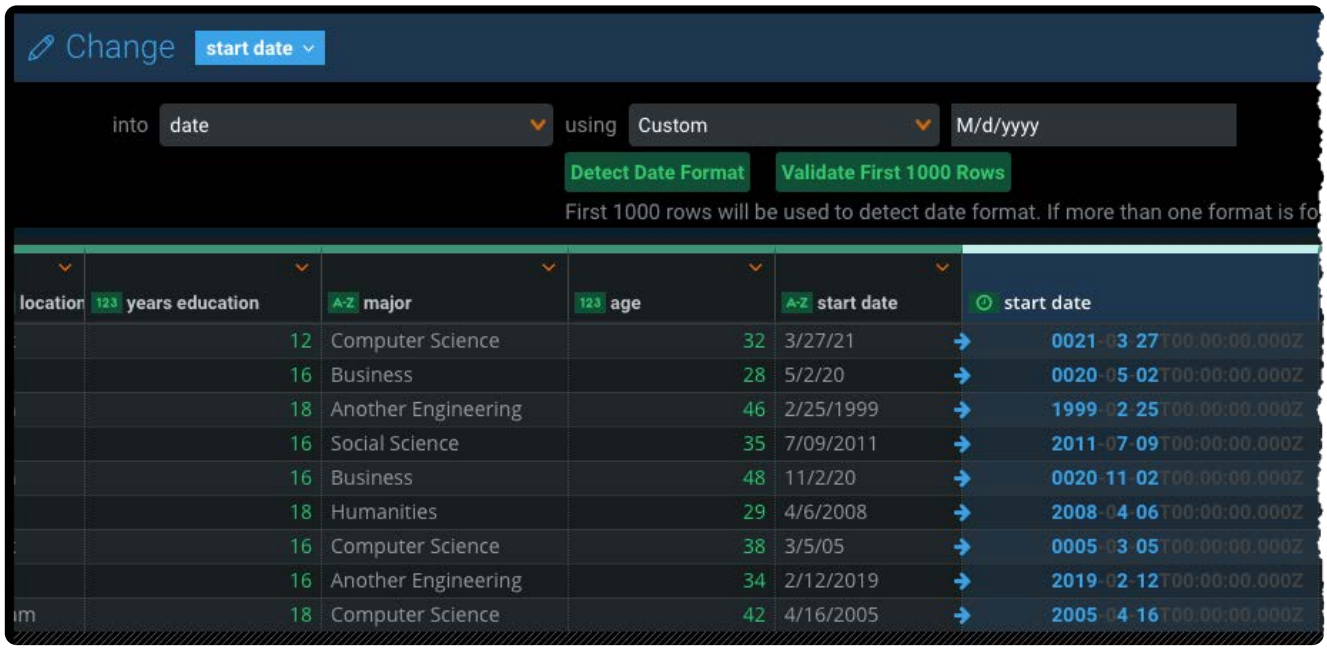
1. 日付を形式化したい列を検索します。
2. 列メニューのアイコンにカーソルを合わせて、**変換 > 日付**をクリックします。




3. **日付形式検出**をクリックします。




最も一般的な形式は、カスタム形式として表示されます。




有効な形式が検出されない場合、エラーメッセージが表示されます。

4. **最初の1000行を検定**をクリックします。

   Data Prepは、指定された形式に対してデータを検定し、不整合があるかを確認します。列内のいずれかのセルで検定が失敗した場合、どの列で失敗したかを示すエラーメッセージが表示されます。

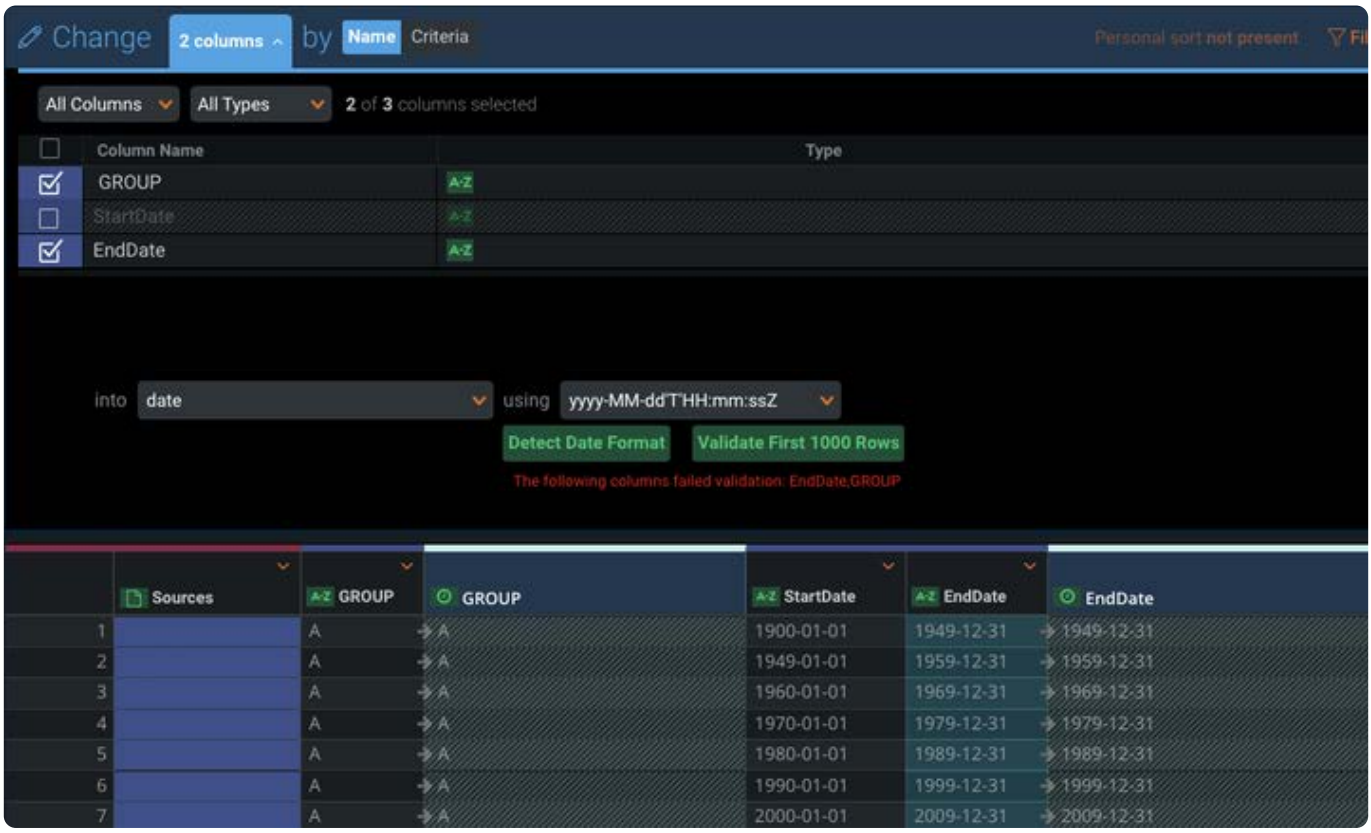
   検定エラーがある場合は、無効な日付の値ごとに日付変換を行ってください。すべての値が指定した日付形式に変換されるまで、この操作を繰り返します。

5. 右上の**保存**をクリックします。

   選択内容を保存すると、列の形式が日付形式に変換され、タイマーのマークが表示される点に注意してください。

| location | years education | major | age | start date | annual salary |
|---|---|---|---|---|---|
| New York | 12 | Computer Science | 32 | 0021-03-27T00:00:00.000Z | 98324 |
| CA | 16 | Business | 28 | 0020-05-02T00:00:00.000Z | 45467 |
| California | 18 | Another Engineering | 46 | 1999-02-25T00:00:00.000Z | 113384 |
| NY | 16 | Social Science | 35 | 2011-07-09T00:00:00.000Z | |
| California | 16 | Business | 48 | 0020-11-02T00:00:00.000Z | 356738 |
| Calif. | 18 | Humanities | 29 | 2008-04-06T00:00:00.000Z | 45467 |
| New York | 16 | Computer Science | 38 | 0005-03-05T00:00:00.000Z | 147098 |
| CA | 16 | Another Engineering | 34 | 2019-02-12T00:00:00.000Z | 168232 |
| Amsterdam | 18 | Computer Science | 42 | 2005-04-16T00:00:00.000Z | 134000 |

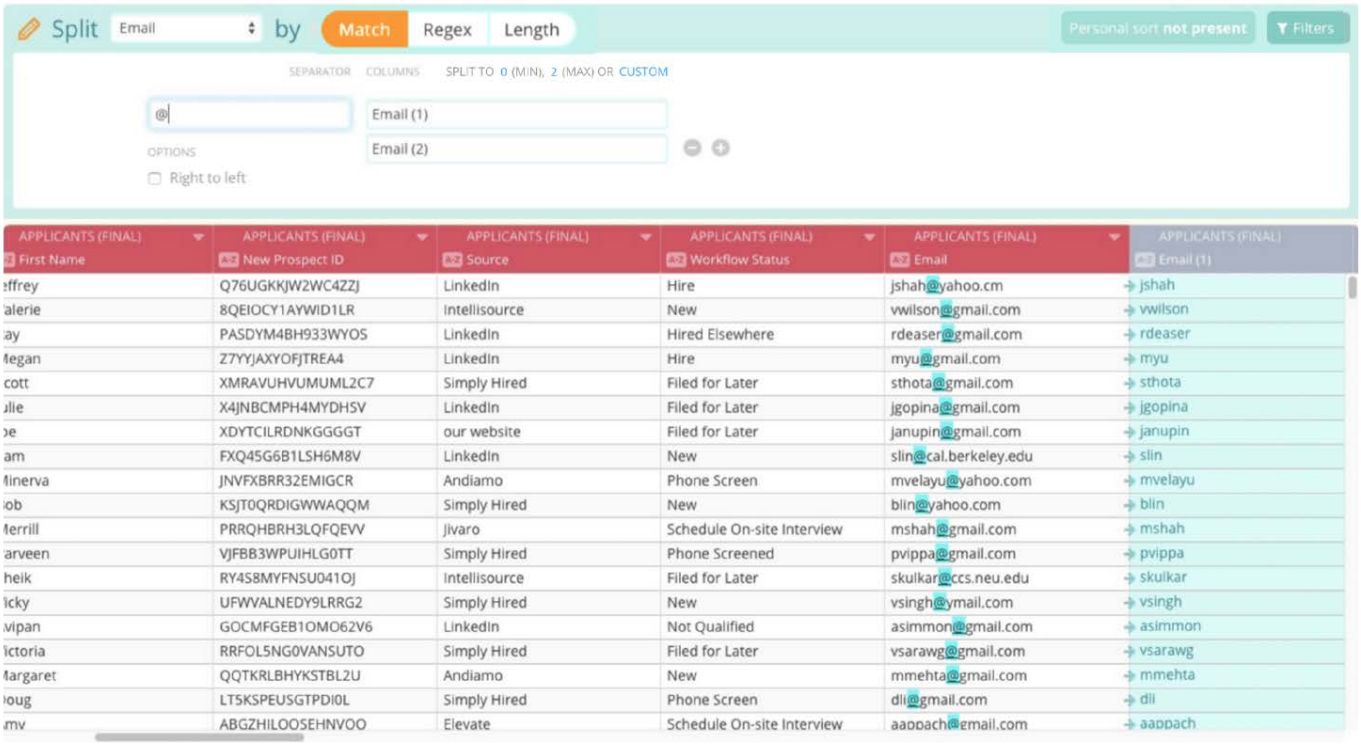
また、複数の列を検出して検定することもできます。



# 列の分割

Data Prepには、データの準備中に列を分割する機能があります。列の分割とは、1つの列の値を、同じ行の1つまたは複数の新しい列に分散させることです。列の分離は列を分離するための文字列を入力するか、文字の長さを入力するいずれかによって動作します。

## 一致



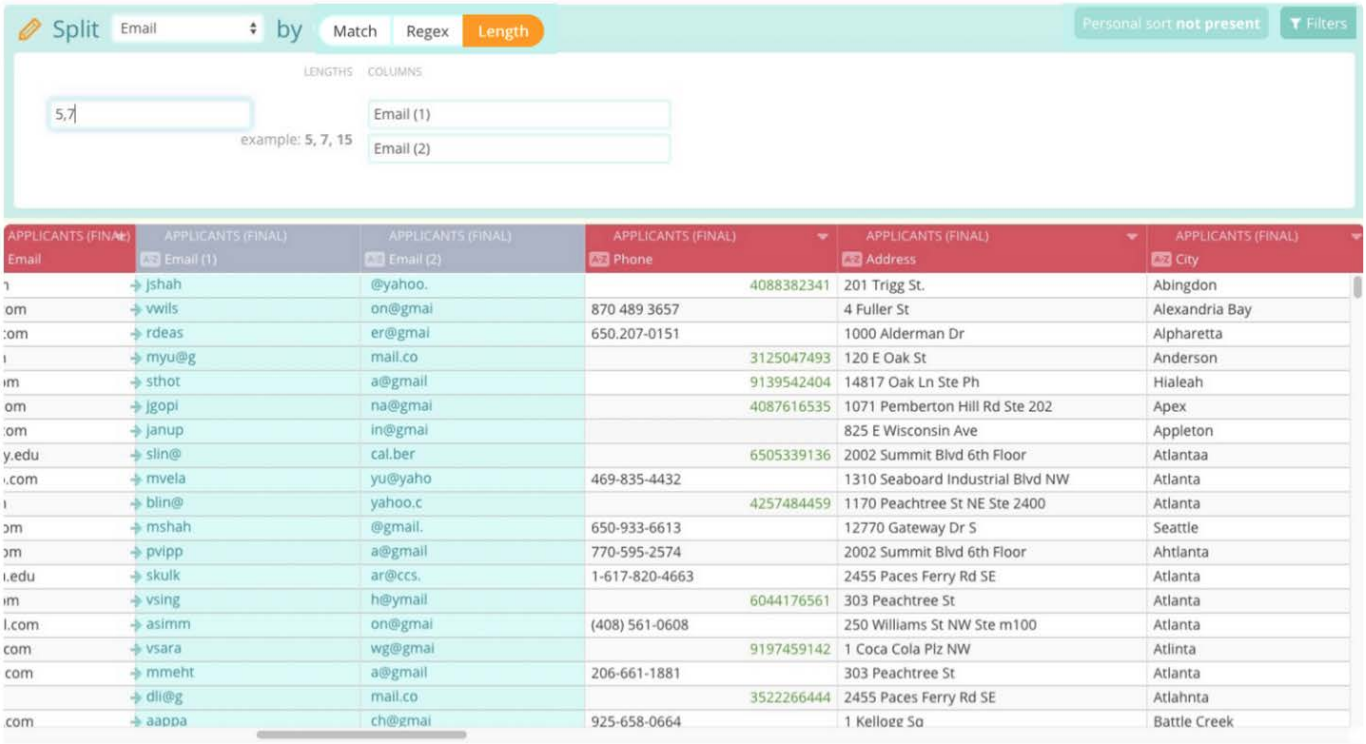

列の分離に使用する**区切り文字**フィールドに1つまたは複数の文字を入力します。入力された文字は、パターンとして認識され、文字が一致する部分ごとに値を分離します。指定される文字は、新たに生成される結果の列には含まれません。分離に使われた文字は、分離された新しい列からは完全に除外されます。

> **備考**
>
> テキストセパレータの値は、大文字と小文字を区別します。

**最小**リンクと**最大**リンクを使用すると、分離操作用に選択した区切り文字と、その区切り文字がこの列のセル全体に出現する回数に基づいて、分割後に作成する新しいカラムの数をすばやく選択できます。また、**カスタム**リンクでは、生成する列の数を正確に指定できます。各フィールドの近くにある [+] ボタンと [-] ボタンをクリックして、新しいカラムフィールドを手動で追加または削除することもできます。

分割を**右から左に**選択するオプションもあります。カラムの分離関数は、デフォルトの設定では、ユーザーが入力したセパレータを使用して、左から右に分析します。**右から左に**のオプションでは、列のテキストを右から解析することができます。これは特に、テキスト文字列を分離する場合に役立ちます。 たとえば、ファイル名をディレクトリ パスから切り離すことができます。この場合、スラッシュ「/」を区切り文字として指定し、**右から左に**オプションを選択します。

## Length



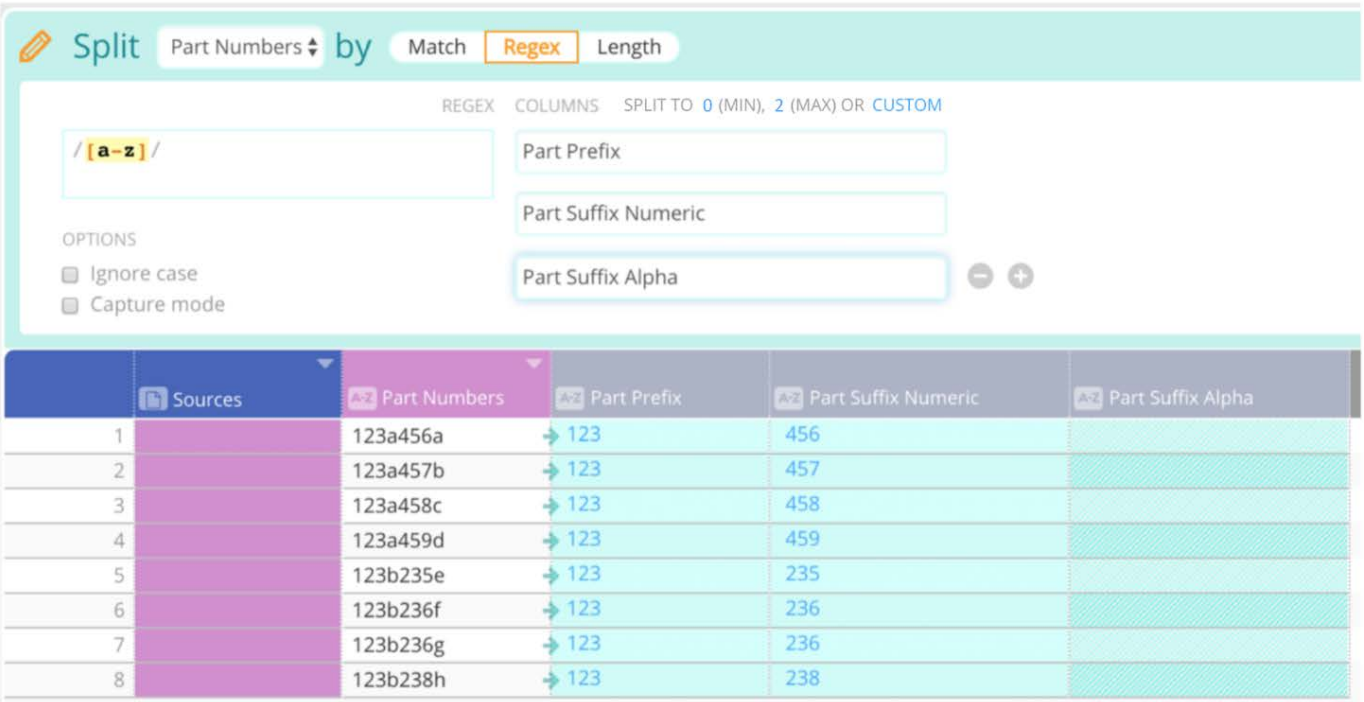

**長さ**フィールドには、カンマで区切られた1つ以上の数字が必要です。**長さ**フィールドは、新たに生成する各列の文字の数を決定するために使用されます。このように、フィールドに「2,3,2」という値を入力すると、最初の2文字で1番目の列が作成され、その後の3文字で2番目の列が作成されます。そして、元の列の中にある次の2文字で3番目の列が作成されます。

**区切り文字での分離**とは異なり、この分離タイプでは、列の途中で文字が除外されることはありません。ただし、不明な（残っている）文字が最後の列にまとめて配置されるのではなく、新しい列から完全に除外されます。思わず、カラムを切り捨てることを回避するために、最後の数字は残っている列の長さを考慮するのに十分な大きさで指定することをお勧めします。

ただし、この方法では、この値が利用可能な文字値の長さよりも大きいことを考慮した「空白」バッファは作成されません。このように大きな値を使用して分離すると、新しく生成された最後のカラムに任意の文字がすべて含まれているものと考えることができます。**長さ**パラメータで指定された余分なスペースが列の文字数を超えた場合は、単に無視されます。

# 正規表現（Regex）

正規表現（Regex）に慣れている場合は、このオプションを使用して、分離を強制したい位置の文字列を特定する検索パターンを定義できます。次の画面は、正規表現を使用して、文字列のアルファベット文字で分離して、新しい 2 つの数値型カラムを作成する例を示します。分離の結果、2 つの数値型カラムが作成されます。




**キャプチャモード**オプションを使用すると、Regexパターンに一致する文字列を抽出できます。

次の例では、Regex のキャプチャモードを使用して、文字列内の 2 番目の数字のセットだけを取り出しています。分離の結果、1 つの数値型カラムが作成されます。

# 列を埋める

Data Prep入力操作を使用して、空白の直前にまたは次の既知の値に基づいて、列内に空白のセルを入力します。列に空白補完関数を適用するには、列の操作メニューにカーソルを合わせ、ドロップダウンで[Fill（空白補完）]までスクロールして、アクションを選択します。Group by（グループ化条件）オプションとSort by（ソート条件）オプションは、すべての空白補完操作に適用できます。**[保存]**をクリックして、入力値をプロジェクトに適用します。

- **後の値で補完：**データポイントが欠損した場合、次に使用可能な空白以外の値を使用して入力されます。
- **前の値で補完：**データポイントが欠損している場合は、最後に表示された空白以外の値を使用して入力されます。
- **平均で補完：**データポイントが欠損した場合、パーティション内で最も近い前と次の空白以外の値の平均値が入力されます。入力値が同じでなので、連続する空白セルは同じ値で埋められます。
- **線形補完：**データポイントがない一連のセルには、それらの空白セルの前後にあるセルのそれぞれの値を結ぶ直線に適合する値が入力されます。すべての欠損値に同じ平均値が割り当てられる「平均で補完」とは異なることに注意してください。線形補完では線形平均が計算されて、欠損値の数に基づいて値が調整されます。


**備考**

平均補完と線形補完は、数値列タイプにのみ適用できます。

# クラスタリングによる正規化

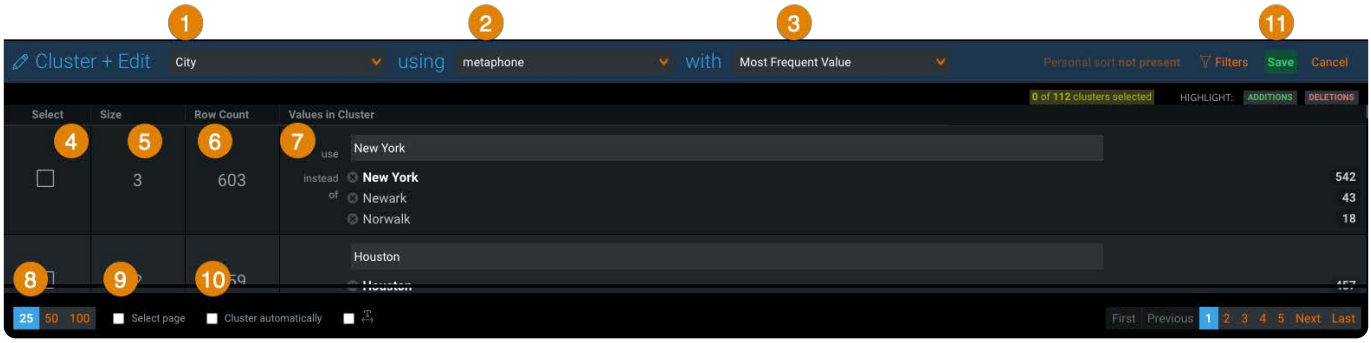
Data Prepの**クラスター + 編集**操作を使うと、列データを迅速に正規化できます。これは、列の一貫性の欠如やエラーの発見に特に役立ちます。

データセットの列で**クラスター+編集**を実行すると：

- Data Prepはすべての列の値を検索し、密接に一致する値をひとつのクラスターにまとめます。
- 各クラスターは、クラスターサイズ（クラスター内のユニークな値の数）およびクラスターの行数（列内で各ユニーク値が発生する回数）とともに**クラスター+編集**ペインに表示されます。

クラスター化されたデータに基づいて、Data Prepはクラスターのすべての値を正規化するための単一の置換値を提案します。提案された値を適用するか、または、別の値の設定してクラスターを標準化することもできます。

以下に、**クラスター+編集**ペインとそのコンポーネントの説明を示します。




| | 要素 | アクション |
|---|---|---|
| 1 | 列フィールド | 操作を実行する列。 |
| 2 | **使用** | ドロップダウンメニューを使って、クラスタリング操作用に使用するアルゴリズムを選択します。詳細については、クラスタリングアルゴリズムを参照してください。 |
| 3 | **対象** | ドロップダウンメニューを使って、出力オプション用に使用するアルゴリズムの1つを選択します。選択したアルゴリズムによってクラスターに対する値の提案が変わります。詳細については、出力アルゴリズムを参照してください。 |
| 4 | **選択** | クリックして、更新するクラスターを選択します。 |

| | 要素 | アクション |
|---|---|---|
| 5 | **サイズ** | クラスター内のユニークな値の数。 |
| 6 | **行数** | クラスター内の行数。 |
| 7 | **クラスター内の値** | Data Prepがクラスター内のすべての値の代替として提案する値を表示します。Data Prepの提案は、選択したクラスタリングアルゴリズムに基づいています。Data Prepの提案をオーバーライドするには、別の値を入力します。値の横の「X」をクリックして、更新されないようにします。 |
| 8 | **25 / 50 / 100** | 一括編集用に、ページごとにクラスター数を選択します。 |
| 9 | **ページを選択** | 一括編集用のクラスターを選択できます。バルク編集では、1ページのクラスター上で操作します。**ページサイズ**フィールドを使用して、ページあたりのクラスター数を指定します。 |
| 10 | **クラスター化を自動的に行う** | **保存**をクリックした後、データセットの全行に一括編集を実行します。 |
| 11 | **保存** | 変更を保存します。 |

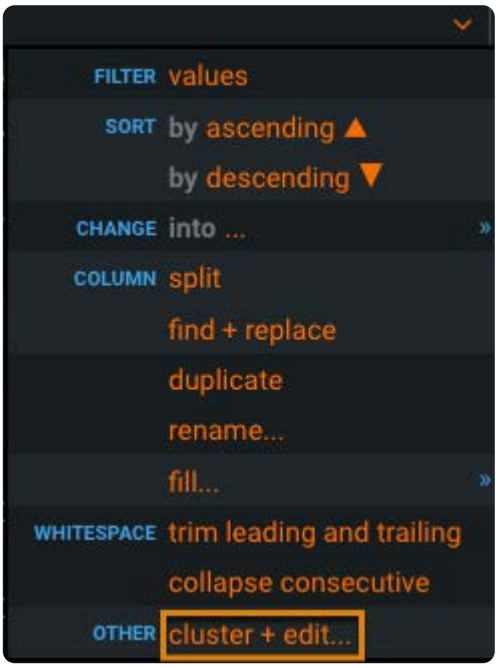
# クラスター+編集を使用するタイミング

次の表は、**クラスター+編集**を使用する一般的なシナリオを示します。

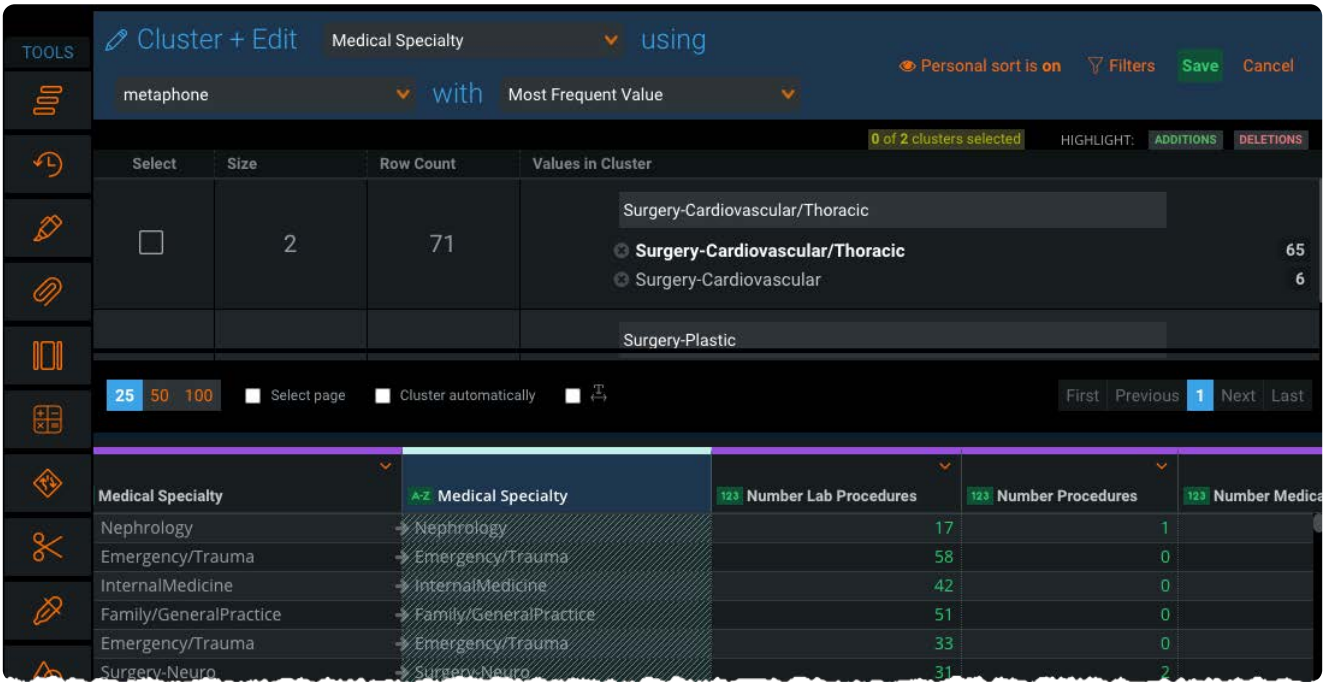
| シナリオ | 説明 | 例 |
|---|---|---|
| 修正と一貫性の欠如 | データ入力の誤り、スペルミス、異なった省略形や略記の一括修正。 | Acme Co.、Acme Company、Acme Comp. |
| 再分類 | 詳細値を集計値に再分類します。 | 「12オンスのソーダ」と「8オンスのソーダ」の両方が「ソーダ」になります。 |
| 統合 | 異なるシステムからのデータが 1 つのカラムに組み込まれたときに生じる、一貫性があるが異なる値を結合する。 | あるデータソースでは一貫して「ソーダ」と表記され、別のデータソースでは一貫して「トニック」と表記されている場合。 |

# 列でクラスター+編集を実行する

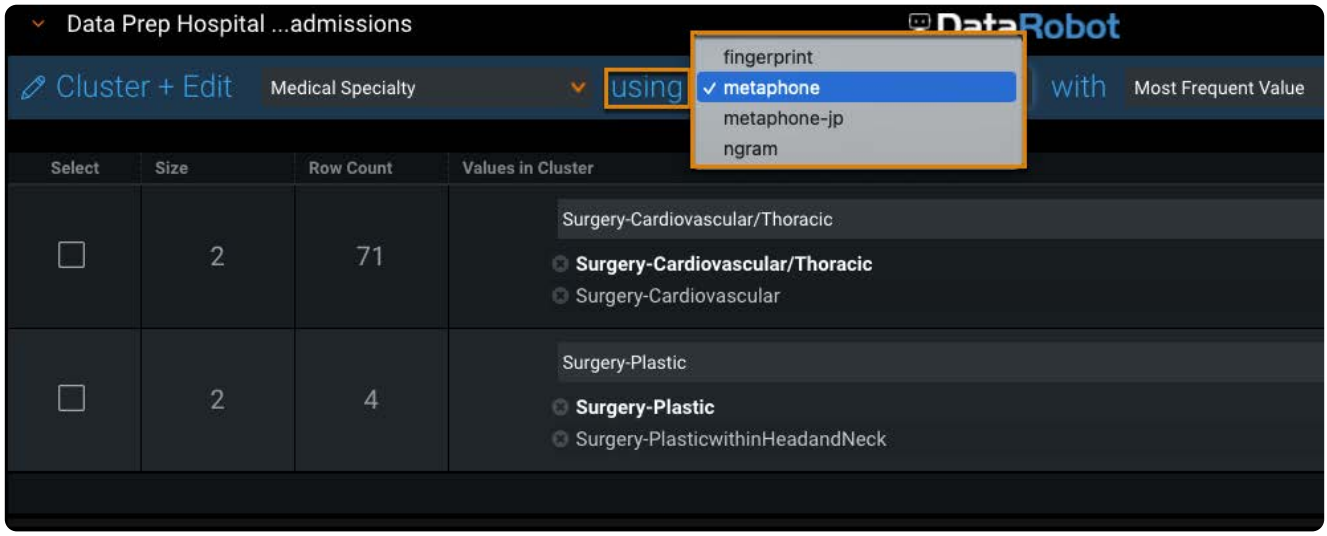
1. 正規化する列を検索します。
2. 列メニューアイコンにカーソルを合わせ、**その他 > クラスター+編集**をクリックします。




3. **クラスター+編集**ペインが開きます。

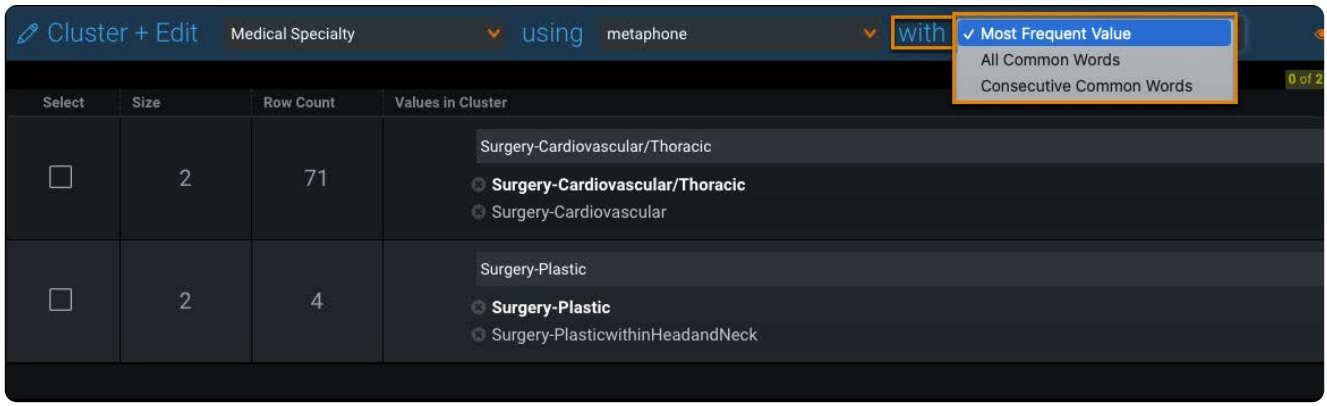



4. **使用**フィールドで、ドロップダウンメニューを使用して、クラスタリング操作に使用するアルゴリズムを選択します。



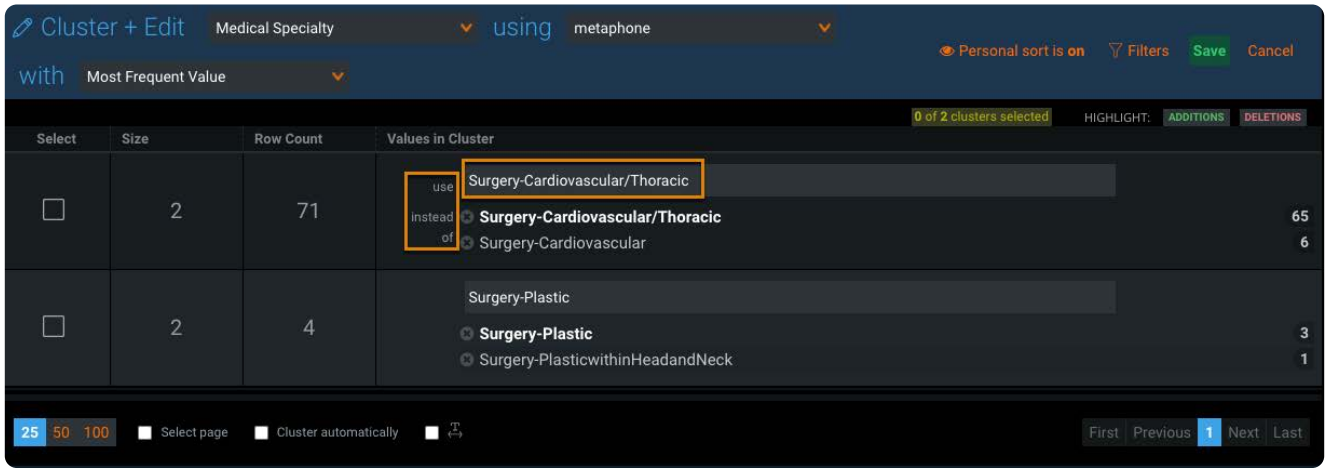

詳細については、クラスタリングアルゴリズムを参照してください。

5. **対象**フィールドで、ドロップダウンメニューを使って、出力オプションに対して使用するアルゴリズムを選択します。選択したアルゴリズムによってクラスターに対する値の提案が変わります。

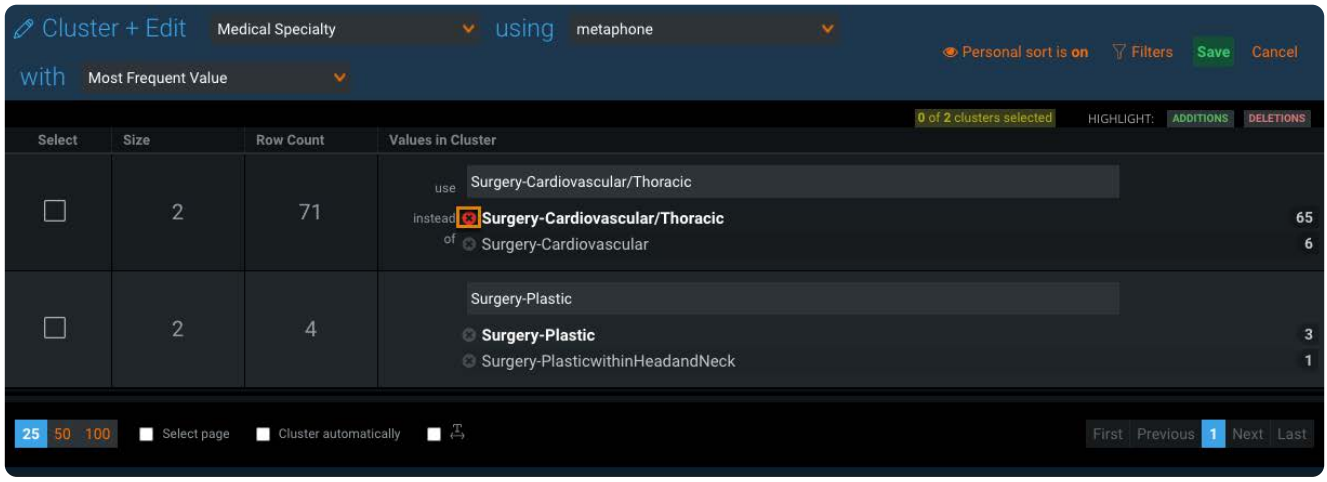



詳細については、出力アルゴリズムを参照してください。

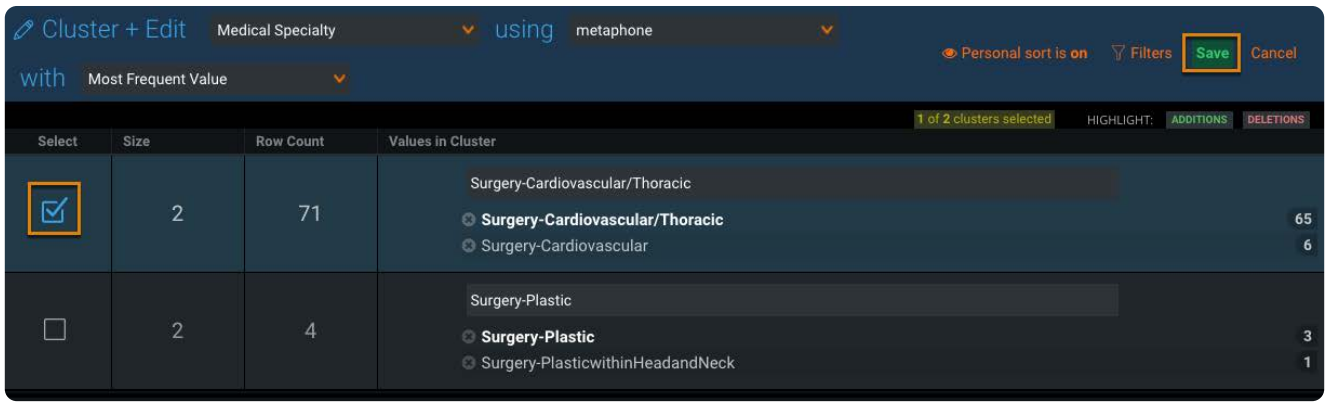
6. Data Prepがクラスターに提案した値を変更するには、**使用**フィールドに用語を入力します。




7. Data Prepで値が置換されないようにするには、値の横の**X**をクリックします。




8. **選択**列で、更新するクラスターの横のチェックボックスをクリックし、**保存**をクリックします。

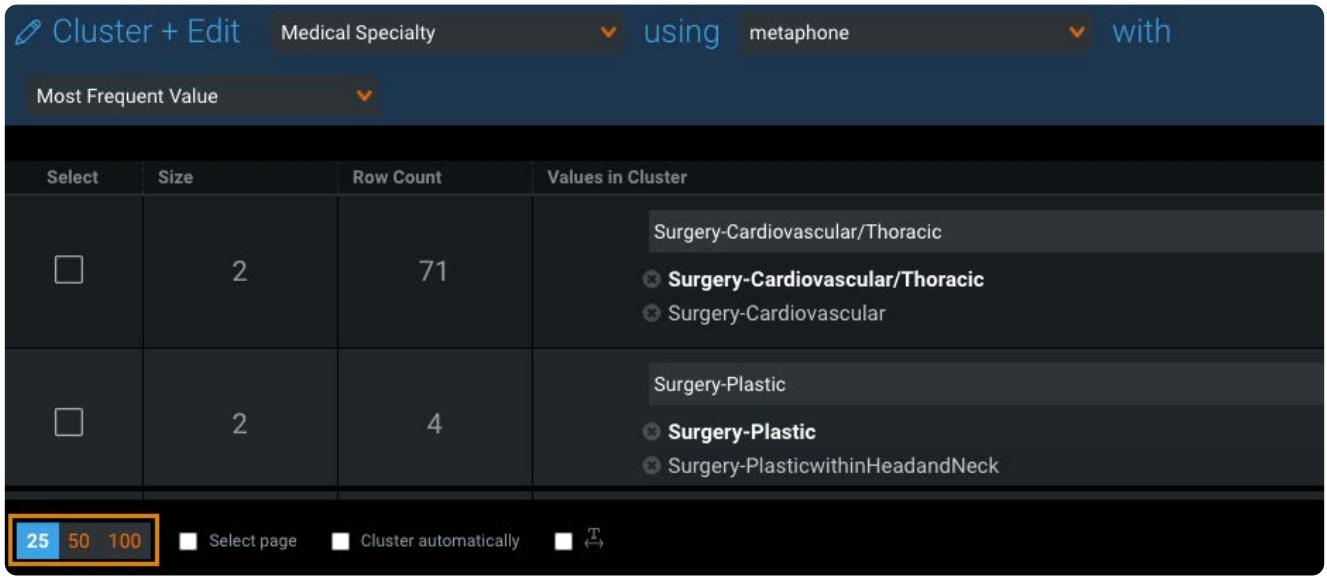



9. 引き続き個別のクラスター編集を行います。

# クラスターの一括編集を実行する

一括編集を使用すると、ページのすべてのクラスターをスピーディに正規化できます。

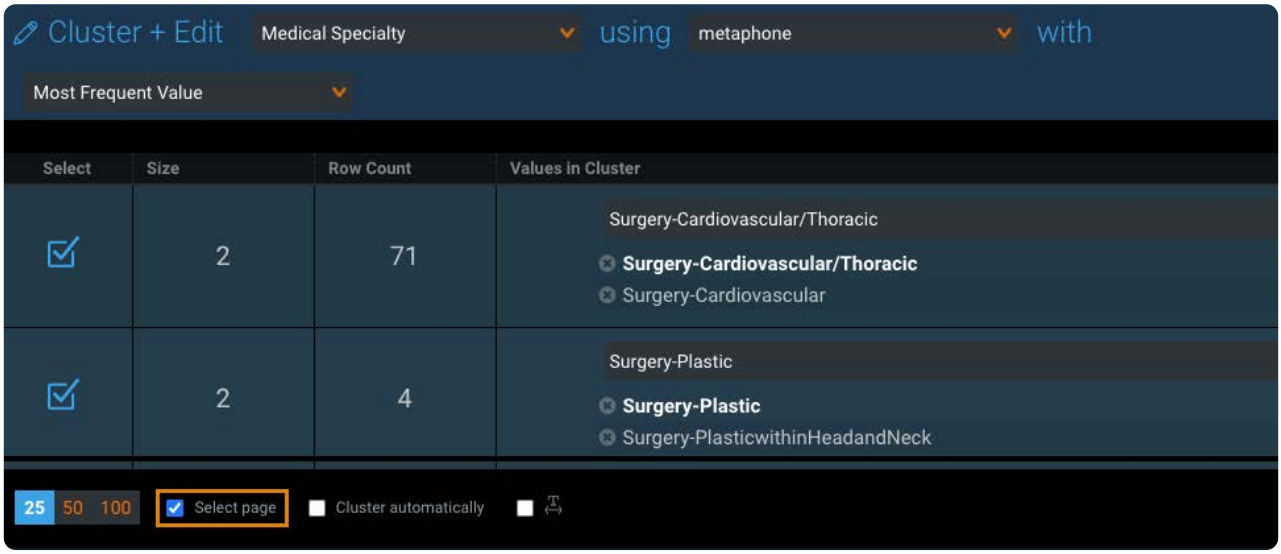
1. ページごとに25、50、または100のクラスターを選択します。




一括編集操作は、クラスターの1ページに制限されています。

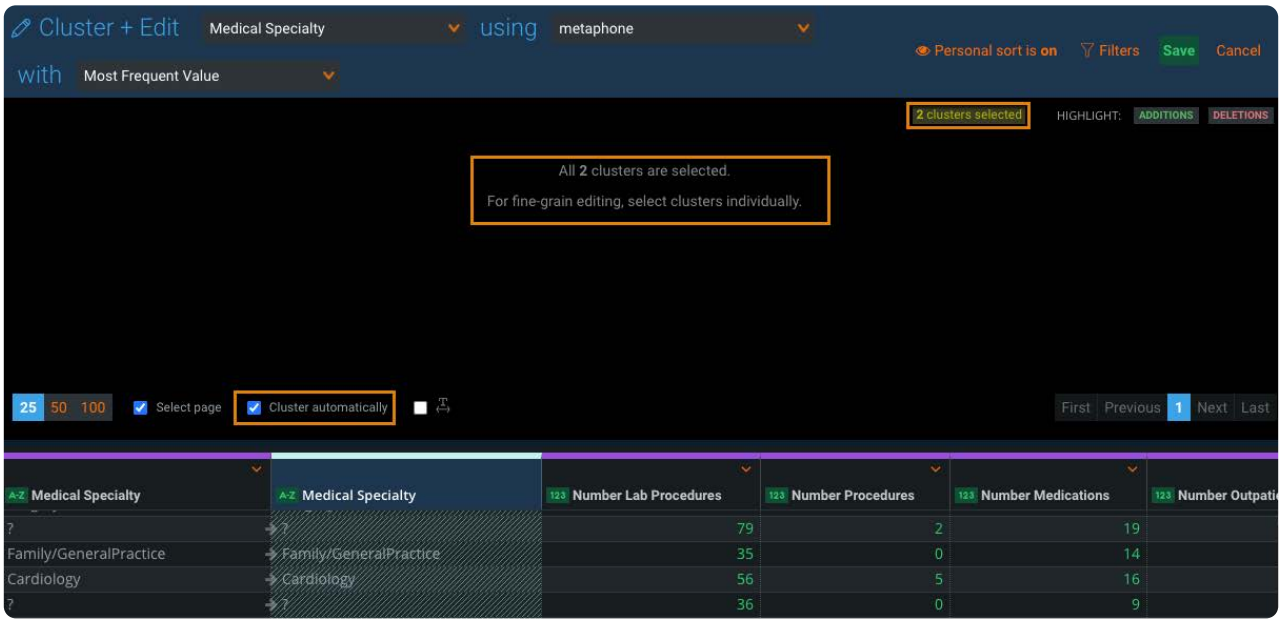
2. 一括編集を実行するには、次のうちいずれかを実行します。

   - **ページを選択**をクリックして、ページのすべてのクラスターを選択します。




   保存する前に提案された置換値を確認して編集を行いたい場合には、この方法を使用します。

   - **クラスター化を自動的に行う**をクリックして、データセットのすべてのクラスターを選択します。




   提案した置換値をすべて受け入れることが確定している場合は、この方法を使用します。

3. クラスターを更新するには、**保存**ボタンをクリックします。

   各クラスターで、すべての値が提案された値に変更されます。

# クラスターの操作に使用するツール

次のツールは、クラスター用に提案された値がどのように派生したかをよりよく認識できるように、視覚的なキューを提供します。


| | 要素 | 説明 |
|---|---|---|
| 1 | 固定幅フォント | デフォルトでは、クラスター値が可変幅フォントで表示されます。固定幅フォントでクラスターの値を表示するには、このオプションをクリックします。固定幅のオプションを使用すると、すべての文字が揃えられ、クラスター間の空白や他の方式の文字も簡単に比較できます。 |
| 2 | ハイライトツール | ハイライトすることで、提案されたクラスター置換値がどのように派生したかを認識できます。**追加**ツールは、すべての一般的な文字に対して追加された文字がハイライトされます。**削除**ツールでは、一般的な文字を派生させるためにどこが削除されるかが示されます。削除箇所は赤いXに凝縮されます。**追加**と**削除**ツールは同時に有効化できます。 |

# クラスタリングアルゴリズム

クラスタリングアルゴリズム**を使用**すると、一緒にグループ化するべき値を定義できます。

> **備考**
>
> すべてのクラスタリングアルゴリズムでは、クラスターを構築する際に空白や Null は含まれません。

アプリケーションでは次のアルゴリズムを利用できます。

## metaphone

metaphoneアルゴリズムは、デフォルト選択で、英語の発音に基づいて単語をグループ化します。これは、テキストを発話したときの音がどれだけ似ているか異なっているかに基づいているため、「音声的」アルゴリズムに分類されます。このアルゴリズムは、手動で入力したデータ（ミススペルが含まれている可能性があるデータ）や、複数のソースシステムから追加されたデータ（細かい差異が含まれている可能性があるデータ）を操作する場合に役立ちます。

## ngram

ngramアルゴリズムは、列内のデータを指定された文字数（n）に分割します。テキストのこれらの「チャンク」（すなわちgrams）は、その後に続く可能性のある確率に基づいて比較されます。ngramアルゴリズムは検索エンジンでよく使用されます。ユーザーが検索バーに文字を入力すると、エンジンが最終的な検索語が取る可能性のある形式の確率を調べ、ユーザーが入力すると同時に候補を表示します。

## fingerprint

fingerprintアルゴリズムは、類似した値を、句読点、語順、大文字化のみが異なるクラスターにグループ化します。fingerprintアルゴリズムは、たとえば「Adèle Smith」と「SMITH, ADELE」という名前を一致させるためによく使用されます。

# outputアルゴリズム

**with**出力オプションは、クラスターの値に対するデフォルト置換値を決定します。出力オプションは、**新しい値**に対する最良の推奨を行うことを試みます。置換値はユーザーの具体的なビジネス要件に合わせて手動で編集することができます。

アプリケーションは次のアルゴリズムを提供します。

## 最頻値

「最頻値」出力アルゴリズム（デフォルト選択）は、クラスター内で最も頻繁に発生する値を使用してクラスターを構築します。

## すべての一般的な単語

「すべての一般的な単語」出力アルゴリズムは、一致する単語の文字列を使用して、順序に関係なく、文字列の先頭からクラスターを構築します。その次に、各文字列が発生する頻度により**新しい値**が決定されます。

**例**

Apple Computer Corporation
Apple Computer Inc
Apple Corporation Computer
Apple Computer
Apple Corp Computer
**新しい値**：Apple Computer

クラスターを構築する際に使用するアルゴリズムは、提案値に影響を及ぼします。

- **metaphone**はクラスター内の単語の語義的な意味を維持しようと試みるため、提案値の一部がクラスター内のすべての一般的な単語を厳密に反映していない場合があります。クラスターに句読点が含まれている場合がこれに該当するかもしれません。
- クラスター内の非連続の共通の語を含めるには、ngram アルゴリズムを使用する必要があります。

# 連続した一般的な単語

「連続した一般的な単語」出力アルゴリズムは、文字列の先頭から始まり、一致する連続する単語の最長のシーケンスを使用してクラスターを構築します。**新しい値**の推奨を決定する際に、クラスターの10%未満で発生する値は含まれません。ほとんどの句読点は、一致のシーケンスを妨げません。

**例**

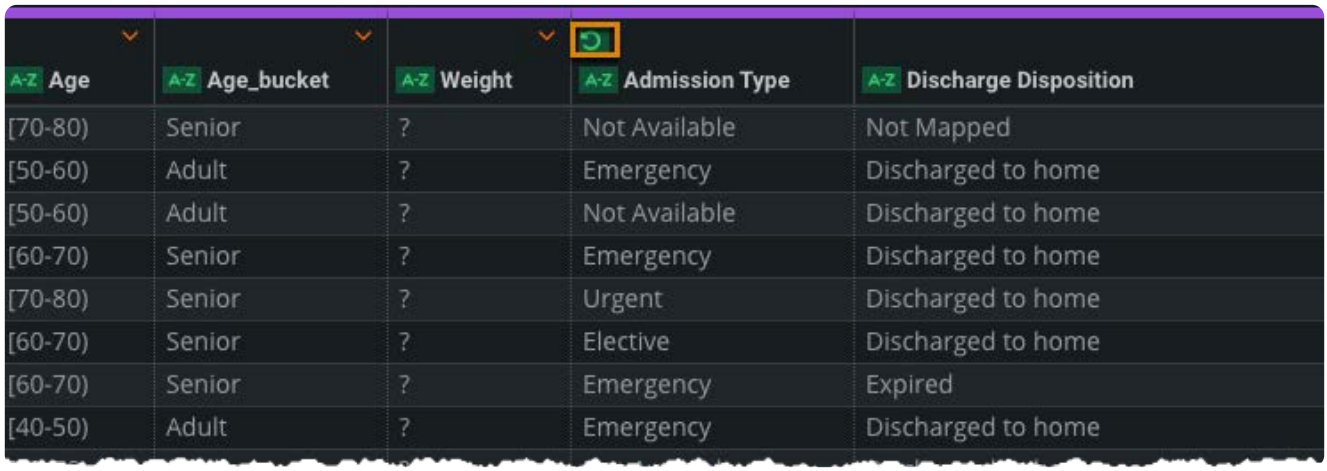
Apple-Computer
Apple Computer
Apple ComputerAG
Apple Computer Corp
Apple Computer Corporation
Apple Computer Inc
**新しい値：**Apple Computer

# 列の系統表示

**ステップ**ペインで列系統モードを有効にし、選択した列を作成したプロジェクトステップを特定できます。

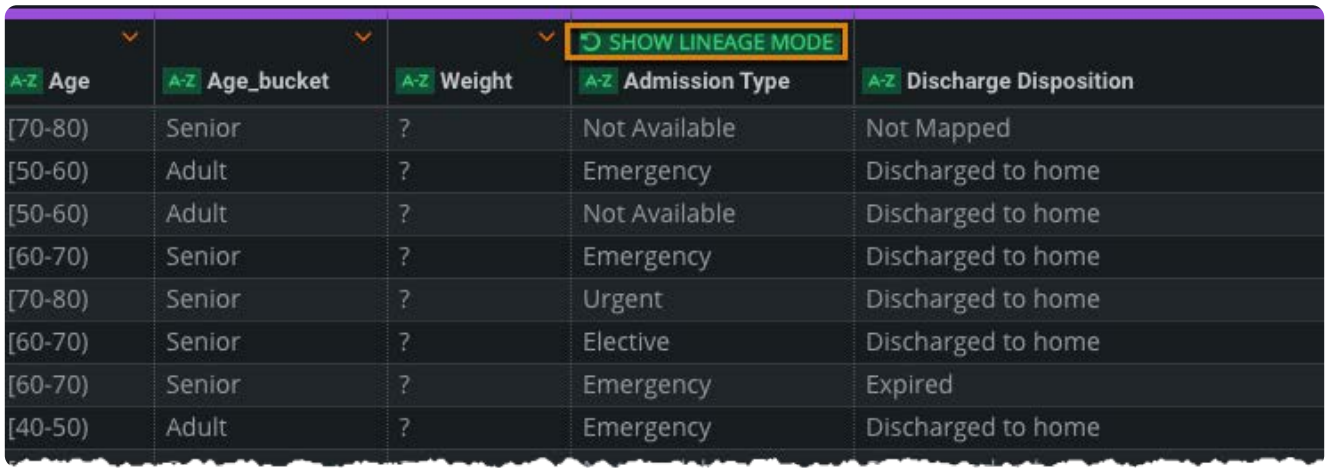
列の系統を表示するには：

1. 列アイコンにカーソルを合わせます。



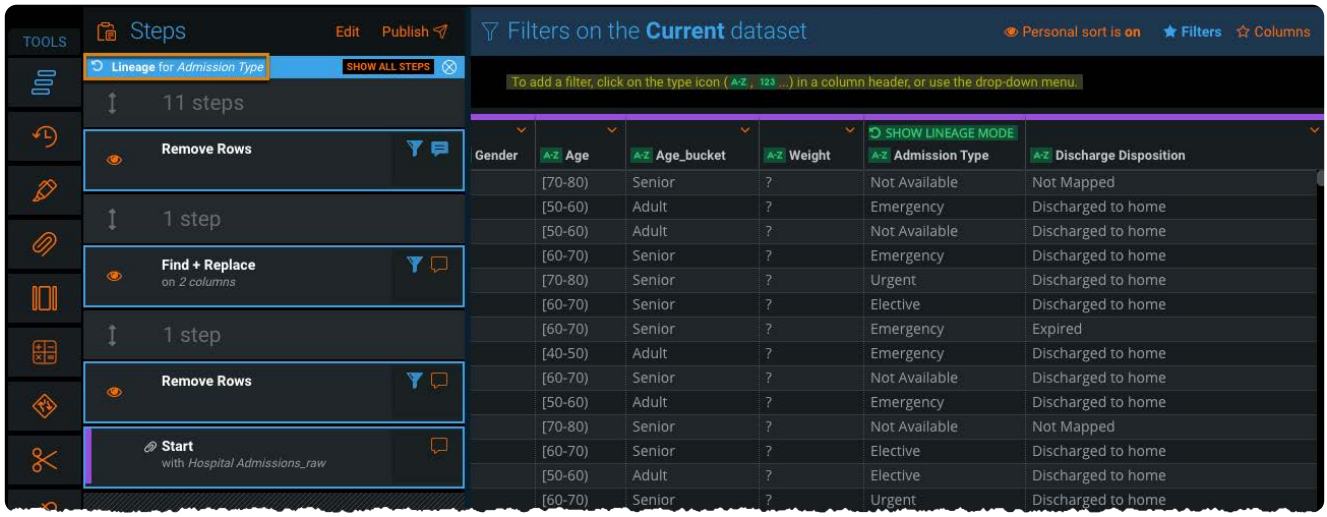
| Age | Age_bucket | Weight | Admission Type | Discharge Disposition |
|---|---|---|---|---|
| [70-80) | Senior | ? | Not Available | Not Mapped |
| [50-60) | Adult | ? | Emergency | Discharged to home |
| [50-60) | Adult | ? | Not Available | Discharged to home |
| [60-70) | Senior | ? | Emergency | Discharged to home |
| [70-80) | Senior | ? | Urgent | Discharged to home |
| [60-70) | Senior | ? | Elective | Discharged to home |
| [60-70) | Senior | ? | Emergency | Expired |
| [40-50) | Adult | ? | Emergency | Discharged to home |

2. 表示された**系統モードを表示**リンクをクリックします。




| Age | Age_bucket | Weight | Admission Type | Discharge Disposition |
|---|---|---|---|---|
| [70-80) | Senior | ? | Not Available | Not Mapped |
| [50-60) | Adult | ? | Emergency | Discharged to home |
| [50-60) | Adult | ? | Not Available | Discharged to home |
| [60-70) | Senior | ? | Emergency | Discharged to home |
| [70-80) | Senior | ? | Urgent | Discharged to home |
| [60-70) | Senior | ? | Elective | Discharged to home |
| [60-70) | Senior | ? | Emergency | Expired |
| [40-50) | Adult | ? | Emergency | Discharged to home |

Data Prepは、選択した列の状態に寄与したステップレベル変換のアウトラインを表示します。

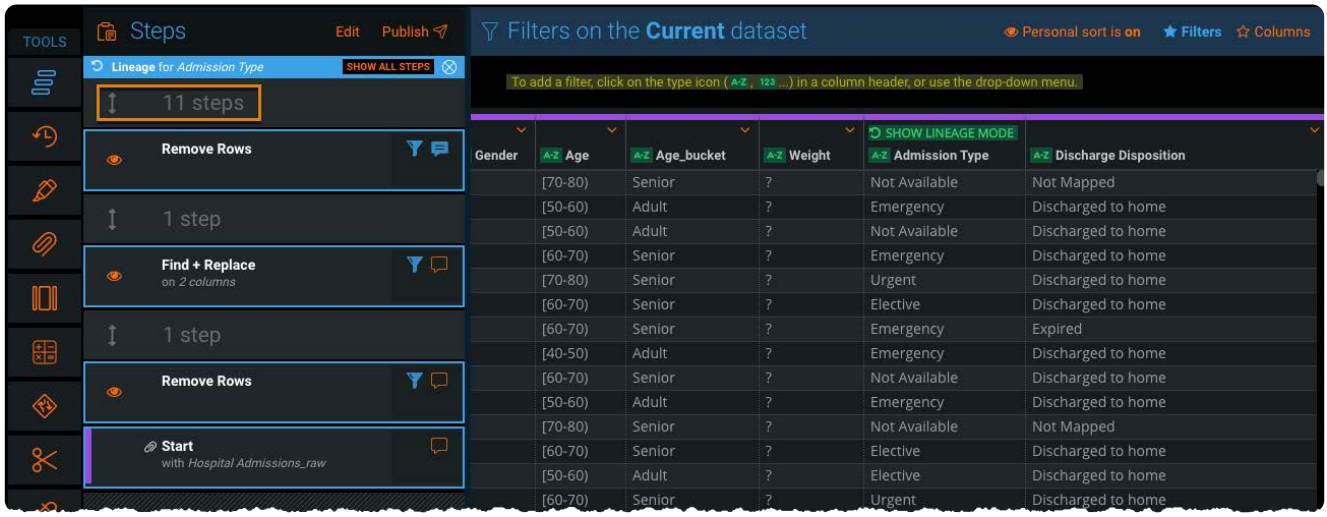



そのアウトラインを使用して、列に影響を与えたステップやデータを変更したステップを確認します。エディター内に列に影響を与えなかったステップがある場合、それらのステップはグレーアウトされ、折りたたまれ、折りたたまれたステップの数を示すラベルが貼られます。

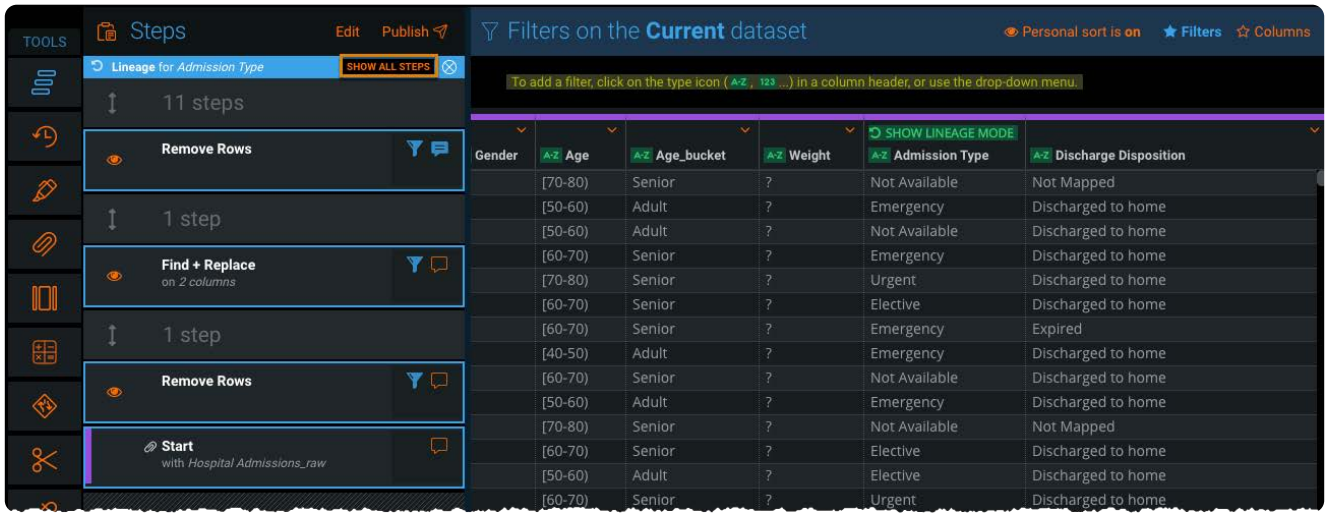
# 系統モードオプション

系統モードで使用できるオプションは以下の通りです。

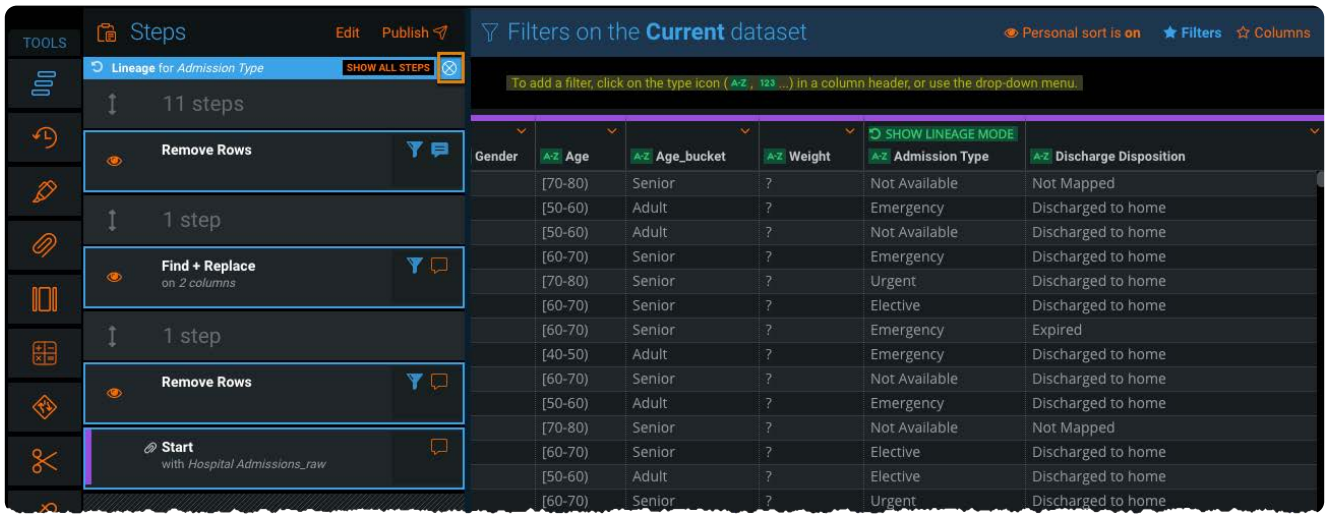
- グレーアウトされたステップをクリックすると、関連付けられ、折りたたまれたステップが展開されます。




- オレンジの系統モードヘッダーで**すべてのステップを表示**をクリックして、スクリプト内のすべての折りたたまれたステップを展開します。




・系統モードヘッダーで**X**をクリックすると、系統モードが閉じます。




**備考**

ステップエディターペインでステップをミュートしたり、プロジェクトで新しい変換を作成したりした場合、系統モードが自動的に閉じます。

# 例

プロジェクトには次のステップがあります。

1. *int'lセル番号*の列で顧客の連絡先情報の基本データセットをインポートします。その列では、すべての数値は*+44-2071838750*という形式に従います。
2. *int'lセル番号*内のダッシュで分割操作を実行して、2つの新しい列（*国コード*と*セル番号*）を作成します。
3. 最初の新規作成列（*国コード*）の名前を変更します。
4. *国コード*列で**検索+置換**操作を実行して、前の**+**文字を削除します。
5. 2番目の新規作成列（*セル番号*）の名前を変更します。
6. 列ツールを使用して、元の*int'lセル番号*列を非表示にします。

*セル番号*列の列系統モードを有効にすると、上記の2番目と5番目のステップがステップエディターページでハイライトされます。これは、それらのステップは*セル番号*列内のデータに直接影響するためです。2番目のステップはデータの元であり、5番目のステップは新しい列名です。その他のステップは列に影響しないため、すべてのステップはグレーアウトされ、折りたたまれます。

**備考**

系統モードに加えて、列のヘッダーの色は、その列のデータの元となるデータソースを示すクイックリファレンスを提供します。データソースの入力ステップの色は、そのソースに由来するすべてのカラムを識別するために使用されます。列の入力データソースがない場合（例えば、列の計算操作の結果として列が作成された場合）、その列はプロジェクトの色で色分けされます。

# 自動化と運用化

Data Prepを使用して、プロジェクトとデータセットを自動化および運用化できます。自動プロジェクトフロー (APFs)を使用すると、複数のData Prepプロジェクト、データセット、AnswerSetにまたがるデータ準備ステップのシーケンス全体を計算し、データに対応するエンドツーエンド（完結型）の自動化出力フローを生成します。

このページでは以下の内容について説明します。

| トピック | 説明... |
| --- | --- |
| 自動プロジェクトフロー | キュレートされたデータフローをインテリジェントに運用します。 |
| 自動化 | 自動化は、個々のプロジェクトとデータセットを自動化できるようにするレガシー機能です。 |

# 自動化

**備考**

自動化は、個々のプロジェクトとデータセットを自動化するオプションを提供する従来の特徴量です。2019年1月リリースで導入された自動プロジェクトフロー（APF）を使用すると、キュレートされたデータフローをインテリジェントに運用できます。新しいAPF 特徴量は、Data Prepプロジェクト、データセット、およびAnswerSetにまたがるデータ準備ステップのシーケンス全体を計算し、データに対応するエンドツーエンドの自動化された出力フローを生成します。現在2018年2月の自動化 特徴量を使用していて、自動化されたジョブをAPFにアップグレードする準備ができているユーザーは、DataRobotの担当者までお問い合わせください。

AnswerSetの作成にかかる反復タスクの数を減らすワークロード自動化には、ライブラリ自動化とプロジェクト自動化の2種類があります。

## ライブラリの自動化

データライブラリデータセットを自動化する場合、定義したスケジュールに基づいてソースから更新を自動的に取得するようにスケジュールが作成されます。自動化プロセス中に、ファイルが最初にデータライブラリにアップロードされたときに指定された、インポートおよび解析オプションを使用して、データセットが新しいバージョンのデータで更新されます。ただし、データセットの自動化を設定するときにこれらの解析オプションを変更することもできます。

**備考**

ローカル システム上では、データセットを自動化できないことに注意してください。

## プロジェクトの自動化

プロジェクトの自動化をスケジュールするときは、定義したスケジュールとパラメータに基づいて AnswerSet が自動的にデータ ライブラリに公開されるように設定します。AnswerSet を AWS S3 などの外部データ ソースにエクスポートすることもできます。

**備考**

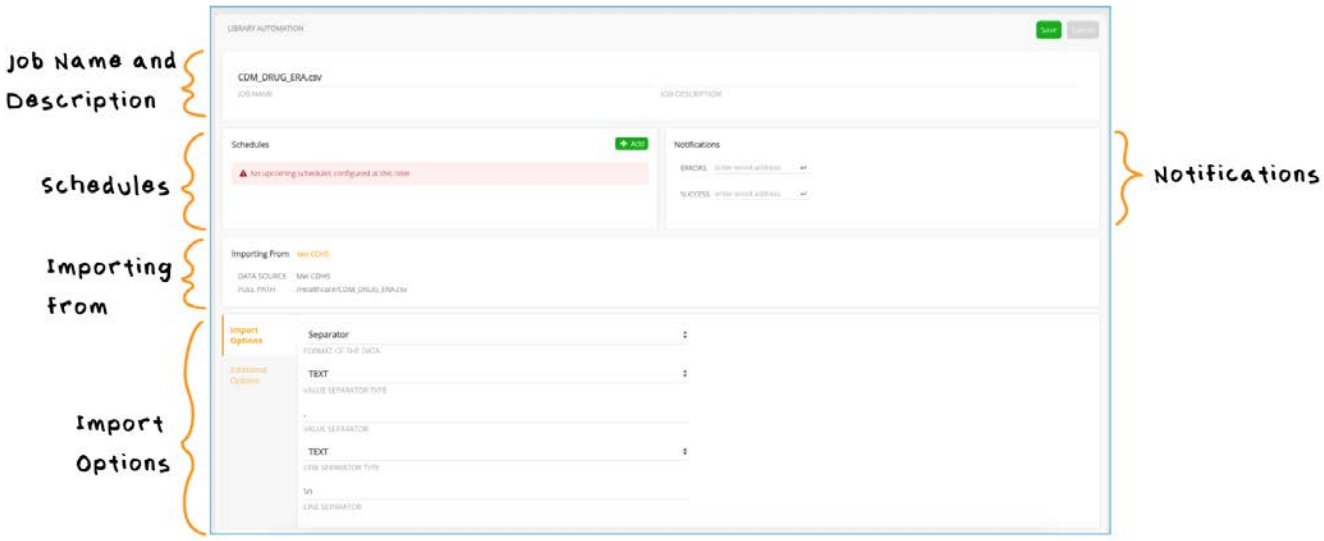
プロジェクトレンズは、自動化されたジョブに使用する公開ポイントを定義するため、プロジェクトの自動化には不可欠です。プロジェクトを自動化するには、データを公開するプロジェクトの各ポイントにレンズを定義する必要があります。プロジェクトで少なくとも1つのレンズを定義する必要があり、そうしないと、データを公開できません。レンズの詳細については、プロジェクトレンズ の記事を参照してください。

データ ライブラリのファイルやプロジェクトの自動化スケジュールを設定した後は、両方まとめて自動化 「ジョブ」と呼ばれます自動化ダッシュボードでは、すべての自動化スケジュールの詳細と、すべての自動化ジョブの状態を確認できます。

# データライブラリの自動設定ページ

データ ライブラリの自動化設定ページを開くには:

1. データ ライブラリを開きます。
2. 自動化するファイルを特定します。
3. 表示される**その他のアクション**ボタンをクリックし、**自動化**オプションを選択します。設定ページが開きます。




## ジョブ名とジョブの説明

これらのフィールドには、データセットの名前と説明が表示されます。これらは、ファイルが最初にデータ ライブラリにインポートされたときの初期のデフォルト値になっています。ここで各フィールドに新しい情報を入力して、値を変更できます。

> **備考**
>
> **この自動化されたスケジュールのオーナーとして設定する**チェックボックスオプションも表示されます。このオプションは、現在のユーザーが、このデータセットで最初に自動化を設定したユーザーではない場合、またはその自動化スケジュールの現在の所有者でない場合にのみ表示されます。所有権は重要な意味を持ち、所有者には、システムで自動化ジョブを実行しているユーザーを特定して監査する手段が提供されます。通常、このオプションは、組織内で自動化に関する責任者が交代した場合に使用されます。自動化ジョブの所有権を取得すると、自動化によって実行される_すべての_操作の実行に必要な、あらゆる権限がなければなりません。

## スケジュール

ここには、データセットに設定されている今後のスケジュールがすべて表示されます。**追加**ボタンを使用して新しいスケジュールを設定できます。このペインの**無効にする**リンクでは、戻って**再有効化する**ボタンをクリックするまで、このデータセット用

のすべてのスケジュールジョブを無期限に停止することができます。スケジュールの設定については、自動化用のデータライブラリの設定を参照してください。

## 通知

データ ライブラリへのアップロードの成功やエラーの発生をユーザーに知らせるために、メール通知を送信できます。通知の設定については、自動化のためのデータライブラリの設定を参照してください。

## インポート元

これらは、データセットの前回のアップロードから引き継がれた接続パラメータです。これらの接続パラメータを変更するには、新しいパラメータを指定して、新しいバージョンのファイルを手動でデータ ライブラリにアップロードします。これにより、スケジュールされている次回の更新時に新しい接続パラメータが使用されるようになります。

## インポートの解析オプション

ファイルベースのデータセットでは、接続の詳細の下にインポートの解析オプションが表示されます。インポート オプションは最新バージョンのデータセットから引き継がれますが、ここで変更することもできます。

> **備考**
>
> このデータセットの別のバージョンを*手動*でデータライブラリにインポートすると、手動アップロード向けに選択した解析オプションが自動バージョンから継承されること*はありません*。

# 自動化用のデータライブラリデータセットを設定

自動化用のデータセットを設定します。

- スケジュールの設定
- 通知の設定
- 自動化設定の保存

## スケジュールの設定

**追加**をクリックして、自動化によってデータセットを更新する時刻を新しく設定します。データセットの自動化頻度のデフォルトの設定は、指定した日時に繰り返すことです。**リピート**トグルボタンは指定したときに**一度**だけ自動化に切り替えることが出来ます。

定期的な更新を設定するには：

1. 上下の矢印を使用して時刻を調整します。
2. **PM**または**AM**のボタンを切り替えて、適切な期間を選択します。

3. 頻度として、**週**、**日**、または**月**を選択します。デフォルトは**週**です。選択を変更するには、フィールドをクリックします。
4. 選択した頻度に応じて、曜日または月の日付を指定します。
5. **オッケー**をクリックしてスケジュールを追加します。新しく追加したスケジュールが表示されます。編集するには鉛筆アイコンをクリックし、削除するには**X**ボタンをクリックします。

**備考**

- 選択する時刻は、現在のタイムゾーンに基づいています。
- ここで選択する時刻、曜日、日付は、これから先のものである必要があります。たとえば、現在時刻が月曜日の午後 1 時だとして、毎週月曜日の午前 10 時に自動化が実行するように設定した場合、今日、このファイルの自動化は実行されません。
- ローカル システム上のデータセットは自動化できません。

単一の更新をスケジュールするには：

1. 日付フィールドをクリックして、カレンダー選択ツールを開きます。
2. 上下の矢印を使用して時刻を調整します。
3. **PM**または**AM**のボタンを切り替えて、適切な期間を選択します。
4. **オッケー**をクリックしてスケジュールを追加します。新しく追加したスケジュールが表示されます。編集するには鉛筆アイコンをクリックし、削除するには**X**ボタンをクリックします。

**備考**

- スケジュールに選択する時刻は、現在のタイムゾーンに基づきます。
- 一度だけ実行する場合、現在の時間に近いジョブの開始時刻を設定しないでください。ローカル コンピューターの時計は、ジョブを処理する Web サーバーと、正確に同期していない可能性があります。ローカルコンピューターの時計が遅れて実行されている場合、ジョブに指定した時刻は、すでにウェブサーバー上では経過している可能性があります。この場合、ジョブは開始されません。
- このデータセットを自動で1回だけテスト実行したい場合は、**1度のみ**の実行を設定する代わりに、**キューに追加**機能を使用します。この機能の詳細については、オートメーション設定の保存を参照してください。
- ローカル システム上のデータセットは自動化できません。

自動化のためのデータセットを設定する際は、以下の注意事項を確認してください:

- 自動化されたプロジェクトでこのデータセットを入力として使用する場合は、プロジェクトの自動実行が開始される前にデータ ライブラリでデータセットの更新のアップロードが完了するように、十分な時間的余裕を確保する必要があります。
- ここで指定する時刻は、このジョブがアップロード用のキューに追加される時刻です。 その時刻に必ずしも自動インポートが開始されるとは限りません。

## 通知の設定

データ ライブラリへのアップロードの成功やエラーの発生をユーザーに知らせるために、メール通知を送信できます。エラーメールには、エラーの原因を特定できる、ファイルのログファイルのためのリンクを貼っています。

通知を設定するには:

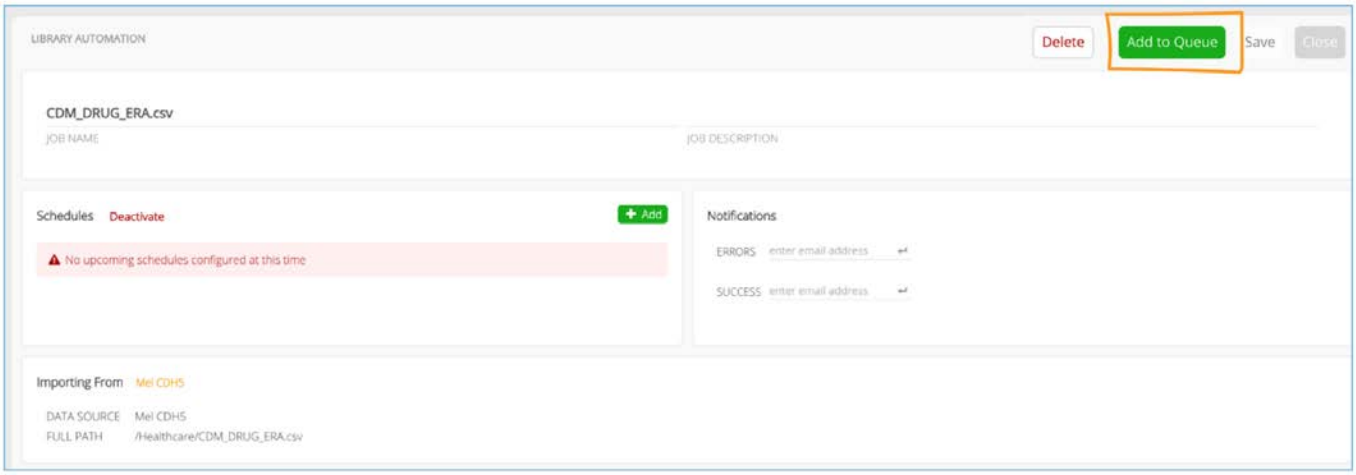
1. ドロップダウン メニューをクリックして、送信するメール通知の種類（「エラー」 または「正常」）をを選択します。
2. メールアドレスを追加し、エンターキーを押します。

重要な考慮事項:

- メール アドレスは、通知の種類ごとに1度だけ追加できます。
- 受信者には、自動化の結果を表示するためのシステム権限が必要です。

## 自動化設定の保存

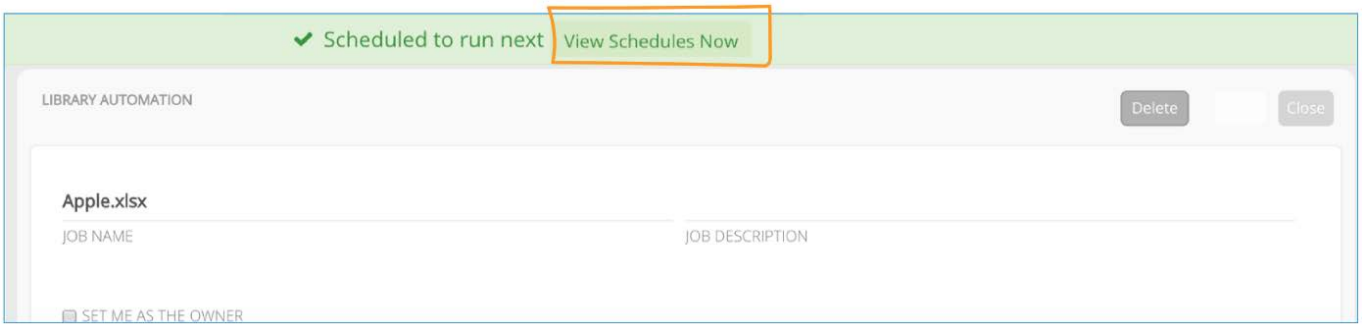
上部にある**保存**をクリックして、すべての設定を保存します。保存した後、**キューに追加**ボタンが表示されます。




このボタンを使用すると、次回の自動化の開始時に実行されるジョブのキューに、この自動化ジョブを追加できます。このオプションは、スケジュールされている実行時刻まで待たずに自動化設定をテストする場合に役立ちます。

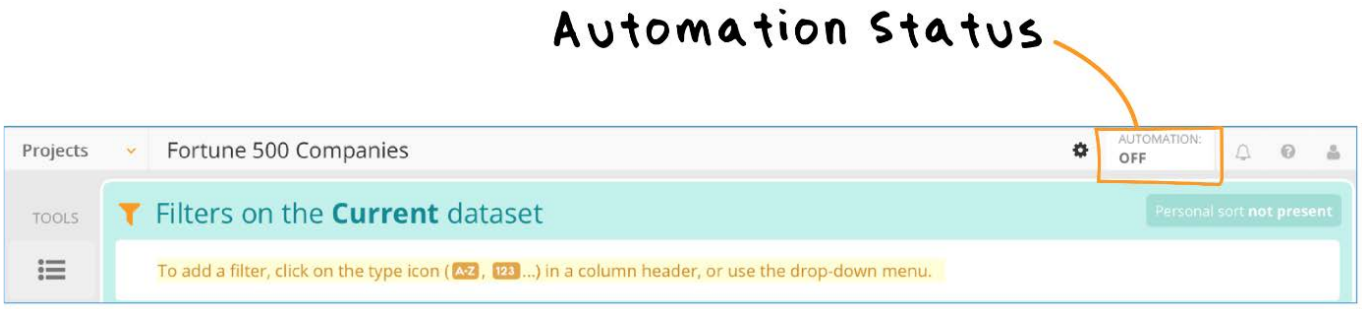
**ヒント**

自動化ペインには、次の自動実行の開始がスケジュールされている時刻の詳細が表示されます。**キューに追加**ボタンをクリックした後にヘッダーに表示される**今すぐスケジュールを表示**リンクをクリックすると、**スケジュール**ペインに素早く移動できます。

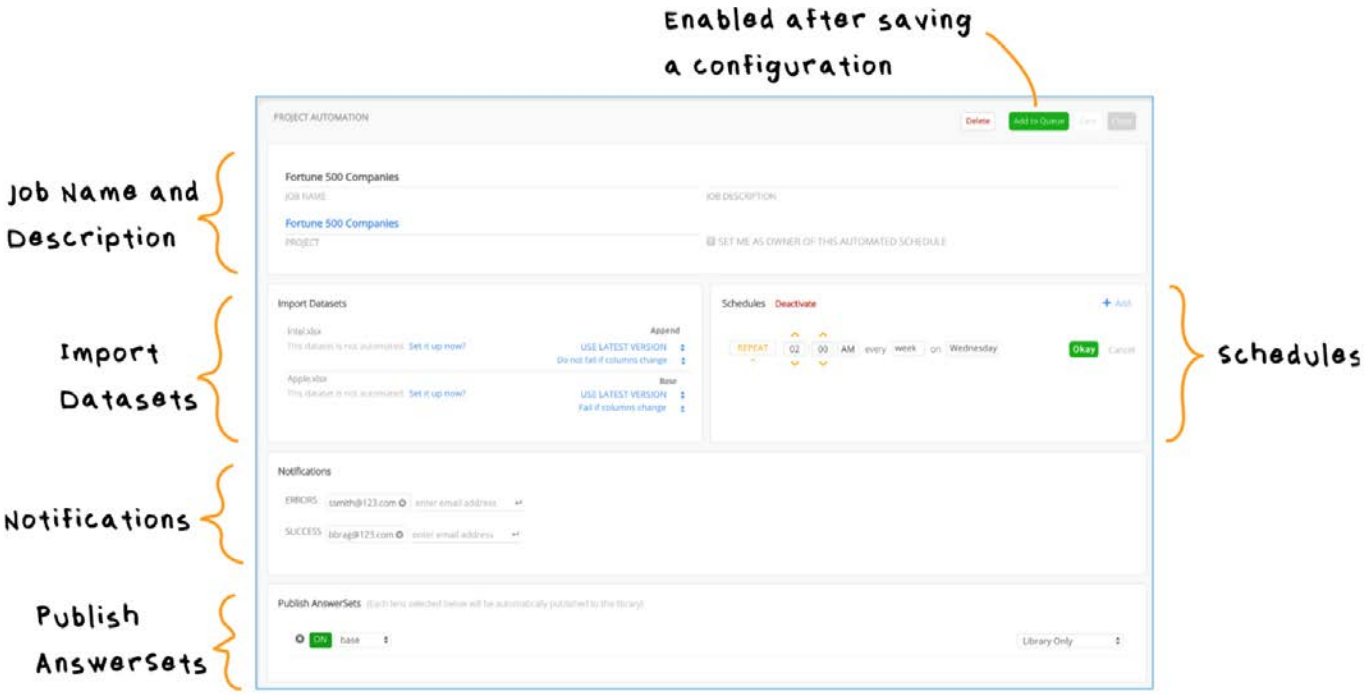



# プロジェクト自動化設定ページ

プロジェクトの自動化ページを開くには、プロジェクトを開き、自動化の状態ボタンをクリックします。




**自動化の設定**ページが開きます。



## ジョブ名、ジョブの説明、およびプロジェクト

これらのフィールドには、ジョブの名前と説明が表示されます。これらの名前と説明は、プロジェクトが最初に作成されたときに指定されたものです。ここでフィールドに新しい情報を入力すると、自動化されたバージョンのプロジェクトに対する値を変更できます。

[プロジェクト] フィールドには、自動化の設定対象のプロジェクトを自動的に開くリンクが表示されます。このリンクは、プロジェクトに複数のバージョンがあり、その中の特定の 1 バージョンをこの設定フォームで自動化している場合に特に便利です。

**備考**

**この自動化されたスケジュールのオーナーとして設定する**チェックボックスオプションも表示されます。このオプションは、現在のユーザーが、このプロジェクトで最初に自動化を設定したユーザーではない場合、またはその自動化スケジュールの現在の所有者でない場合にのみ表示されます。所有権は重要な意味を持ち、所有者には、システムで自動化ジョブを実行しているユーザーを特定して監査する手段が提供されます。通常、このオプションは、組織内で自動化に関する責任者が交代した場合に使用されます。自動化ジョブの所有権を取得すると、自動化によって実行される すべての操作の実行に必要な、あらゆる権限がなければなりません。

## データセットのインポート

ここには、プロジェクトに既にインポートしたデータセットが表示されます。データセットが表示されていない場合は、プロジェクトのステップにて最新の変更が保存されていることを確認してください。

「ベース」 データセットが表示され、その上にプロジェクトの 「ルックアップ」 または「追加」 データセットが一覧表示されます。

データ ライブラリで自動化が設定されたデータセットをこのプロジェクトで使用している場合は、ここにスケジュールが表示されます。自動化されたデータセットを使用する場合は、その自動化スケジュールを考慮し、新しいバージョンがデータ ライブラリに公開されるように十分な時間的余裕を確保してください。

このプロジェクトの自動化を設定する前に自動化のデータセットを設定するには、**今すぐ設定しますか？**をクリックしてください。というデータセットの名前に隣接するリンクをクリックします。データ ライブラリのスケジュール設定ページが表示され、自動化のパラメータとスケジュールを設定できます。自動化のためのデータライブラリデータセットの設定を参照してください。

**備考**

**最新のバージョン**の使用は、この自動化設定で使用されるデータセットのバージョンを参照します。このプロジェクトの自動化に選択するバージョンの詳細については、自動化のプロジェクトを設定を参照してください。

## スケジュール

ここには、プロジェクトに設定されている今後のスケジュールがすべて表示されます。**追加**ボタンを使用して新しいスケジュールを設定できます。このペインの**無効にする**リンクでは、戻って**再有効化する**ボタンをクリックするまで、このデータセット用のすべてのスケジュールジョブを無期限に停止することができます。スケジュールの設定については、自動化のためのプロジェクトの設定を参照してください。

## 通知

自動化されたプロジェクトの更新終了やエラーの発生をユーザーに知らせるために、メールを送信できます。通知の設定については、自動化のプロジェクトの設定を参照してください。

## AnswerSetの公開

AnswerSet を公開するレンズを選択します。レンズはプロジェクト内のステップに固定され、自動化によって AnswerSet を公開するときに使用できる公開ポイントを作成します。プロジェクトの自動化設定はレンズを選択しなくても保存できますが、プロジェクトの自動実行は、レンズを選択するまで正常に作動しません。

自動化されたプロジェクトは、データ ライブラリに自動的に公開されます。ただし、公開された出力を外部データ ソースにエクスポートするように自動化を設定することもできます。自動化のためのプロジェクトの設定を参照してください。

# 自動化用のプロジェクトの設定

自動化用のデータセットを次の方法で設定します：

- データセットのインポート
- スケジュールの設定
- 通知の設定
- レンズの選択と公開先
- 自動化設定の保存

## データセットのインポート

プロジェクトで使用する各データセットについて、入力に**最新のバージョン**を使用するか、**現在のバージョン**を使用するかを選択します。

- **最新のバージョン**を選択すると、自動化ジョブが実行されるときにデータセットが最新のバージョンが使用されます。

**備考**

- **最新のバージョン**を使用すると、この自動化された設定が実行されるたびにプロジェクトの新しいバージョンが発生します。最新バージョンを選択すると、新しい列の追加、プロジェクトのステップで使用されていない列の削除、既存の列の列型の変更、順序の変更など、データセットの最新バージョンのレイアウト（スキーマ）が異なる場合、自動化の実行を失敗させるかどうかを指定できる追加オプションを使用できます。
- このプロジェクトの自動化に使用するデータセットの少なくとも1つが**最新のバージョン**である必要があります。そうしないと、プロジェクトを自動実行した後に入力データセット内に変更が発生しなかった場合に、プラットフォームでこのジョブが再度実行されなくなります。

- **現在のバージョン**は、すべての将来の自動化された実行においてデータセットを現在の状態に固定します。**現在のバージョン**は、静的なデータセットがプロジェクトの参照テーブルとして使用されている場合に便利です。

**備考**

前回のプロジェクトの自動化の実行以降、入力データセットに変更がない場合、プロジェクトに変更が加えられるまで、プロジェクトの自動化は再度実行されません。したがって、プロジェクトの自動化に使用されるデータセットのうち、少なくとも1つに対して**最新バージョン**を選択する必要があります。

## スケジュールの設定

**追加**をクリックして、このプロジェクトを実行する時刻を新しく設定します。プロジェクトの自動化のデフォルト設定は、ここで指定した時刻と曜日に繰り返し行われます。**繰り返し**ボタンをクリックすると、指定した日時に自動化が**1度のみ**実行されるように切り替わります。

定期的な実行を設定するには：

1. 上下の矢印を使用するか、フィールドに値を入力して、時刻を調整します。
2. **PM**または**AM**のボタンを切り替えて、適切な期間を選択します。
3. 頻度として、**週**、**日**、または**月**を選択します。デフォルトは**週**です。選択を変更するには、フィールドをクリックします。
4. 選択した頻度に応じて、曜日または月の日付を指定します。
5. **オッケー**をクリックしてスケジュールを追加します。
6. 新しく追加したスケジュールが表示されます。編集するには鉛筆アイコンをクリックし、削除するには**X**ボタンをクリックします。

**備考**

- ここで選択する時刻は、現在のタイムゾーンに基づきます。
- ここで選択する時刻、曜日、日付は、これから先のものである必要があります。たとえば、現在月曜日の午後1時で、毎週月曜日の午前10時に自動化を実行するように設定した場合、このファイルの自動化は、今日は実行されません。

単一の実行を設定するには：

1. 日付フィールドをクリックして、カレンダー選択ツールを開きます。
2. 上下の矢印を使用して時刻を調整します。
3. **PM**または**AM**のボタンを切り替えて、適切な期間を選択します。
4. **オッケー**をクリックしてスケジュールを追加します。
5. 新しく追加したスケジュールが表示されます。編集するには鉛筆アイコンをクリックし、削除するには**X**ボタンをクリックします。

**備考**

- ここで選択する時刻は、現在のタイムゾーンに基づきます。
- プロジェクトの自動化を一度だけ実行する場合、現在の時間に近すぎるジョブの開始時刻を設定しないでください。ローカル コンピューターの時計は、ジョブを処理する Web サーバーと、正確に同期していない可能性があります。ローカルコンピューターの時計が遅れて実行されている場合、ジョブに指定した時刻は、すでにウェブサーバー上では経過している可能性があります。この場合、ジョブは開始されません。
- このプロジェクトの自動化を1回だけテスト実行したい場合は、**1度のみ**の実行を設定する代わりに、**キューに追加**機能を使用します。詳細については、プロジェクト自動化設定の保存を参照してください。

プロジェクトの自動化を設定する場合の重要な考慮事項:

- プロジェクトの自動化が、自動化されたデータ ライブラリ ファイルからの入力、または他の自動化されたプロジェクトから公開される AnswerSet からの入力に依存している場合は、プロジェクトの自動実行が開始される前にすべての入力の更新が完了するように、十分な時間的余裕を確保してください。
- 自動化設定で指定した時刻は、このプロジェクトがAnswerSetを公開するためのキューに追加される時刻であり、必ずしも公開開始時刻ではありません。

## 通知の設定

自動化されたプロジェクトの更新終了やエラーの発生をユーザーに知らせるために、メールを送信できます。エラーメールには、エラーの原因を特定できるプロジェクトのログファイルへのリンクが貼られています。

通知を設定するには:

1. ドロップダウン メニューをクリックして、送信するメール通知の種類（「エラー」 または「正常」）をを選択します。
2. メールアドレスを追加し、エンターキーを押します。

重要な考慮事項:

- メール アドレスは、通知の種類ごとに1度だけ追加できます。
- 受信者には、自動化の結果を表示するためのシステム権限が必要です。

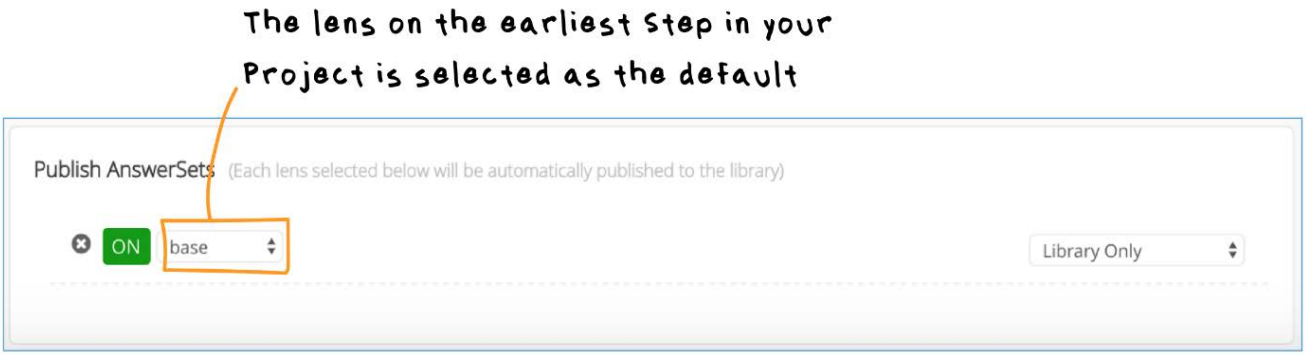
# レンズおよび公開先の選択

レンズを追加するには:

1. 緑色の**追加**ボタンをクリックします。


   プロジェクトのレンズがこの自動化設定に追加されます。デフォルトでは、プロジェクト ステップ内にある最初のレンズが選択されます。

2. デフォルトの選択を変更するには、ドロップダウンメニューをクリックして、プロジェクトに現在存在している別のレンズを選択します。
3. **追加**ボタンをクリックして、プロジェクトのこの自動実行で使用するレンズを追加します。その後、レンズの選択を続行します。




レンズを無効にするには：

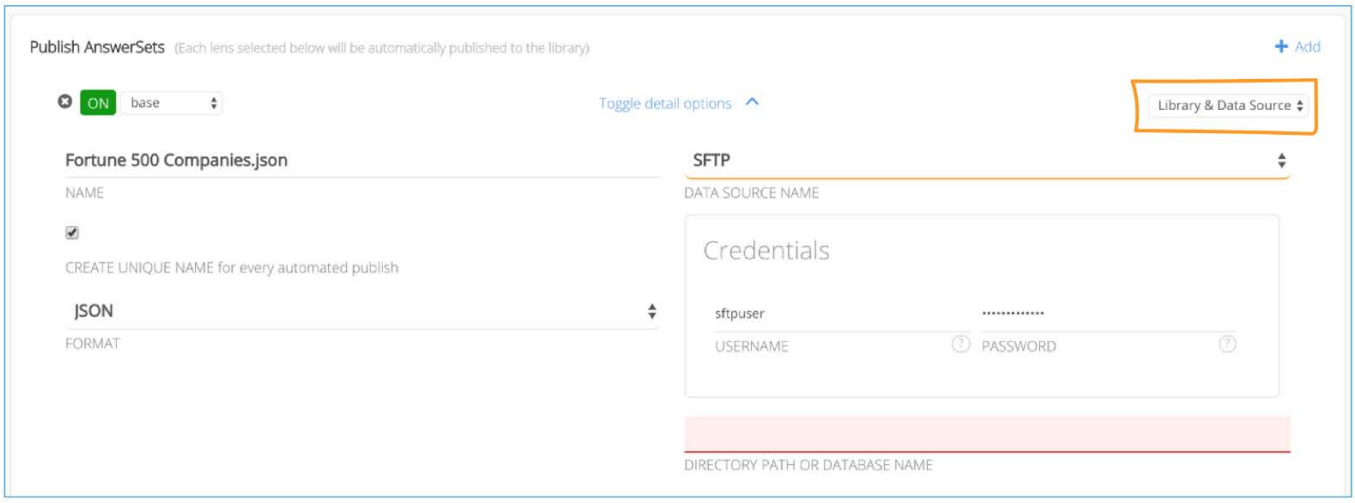
- 緑色の**オン**ボタンをクリックして、オフに切り替えます

レンズを取り外すには：

- レンズの**X**ボタンをクリックします。

  このプロジェクトの自動化用に新しいレンズを追加する場合は、プロジェクトを開き、適切なステップにレンズを追加する必要があります。プロジェクトへのレンズの追加の詳細については、公開のためにレンズを使用する、を参照してください。

自動化の実行時にこのプロジェクトが公開される場所は、デフォルトでは**ライブラリのみ**になります。データライブラリへの公開に加えて、公開された出力を外部データソースにエクスポートするには、ドロップダウンメニューをクリックし、**ライブラリ & データソース**を選択します。




- **名前**: このプロジェクトの自動化されたバージョンで使用される名前です。
- **データソース名**：ドロップダウンメニューをクリックして、使用できるデータソースを選択します。

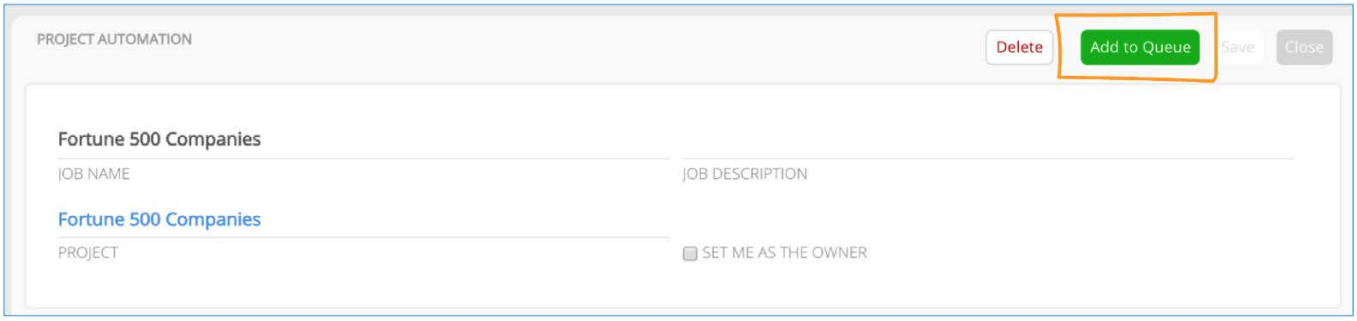
> **備考**
>
> エクスポート用に設定され、アクセス権限を持っているデータソースだけが、**データソース名**のドロップダウンメニューに表示されます。エクスポート先として指定するデータ ソースが表示されない場合は、システム管理者にお問い合わせください。

- **ディレクトリパスまたはデータベース名**: エクスポートが書き込まれるデータソース上のパスまたはデータベースを示します。
- **フォーマット**: エクスポート用に選択したデータソースに応じて、ファイル形式を選択するオプションも使用できます。適用できる解析オプションがあれば、それらも提示されます。
- **資格情報**: ここには、選択されたデータソースに書き込むためのユーザーの資格情報が表示されます。ここで資格情報を編集できます。
- **ユニーク数名を作成**： この機能を有効にすると、自動化されたエクスポートを連続して行うたびに、ファイル名またはテーブル名にアンダースコアとタイムスタンプが付加され、このプロジェクトの以前のエクスポートがデータソース上で上書きされないようになります。

> **備考**
>
> JDBCデータソースでこのオプションを有効にすると、システム管理がJDBCコネクター形式**でテースを自動的に作成する**オプションも有効にします。そうしないと、このプロジェクトの自動化が失敗します。

## プロジェクトの自動化設定の保存

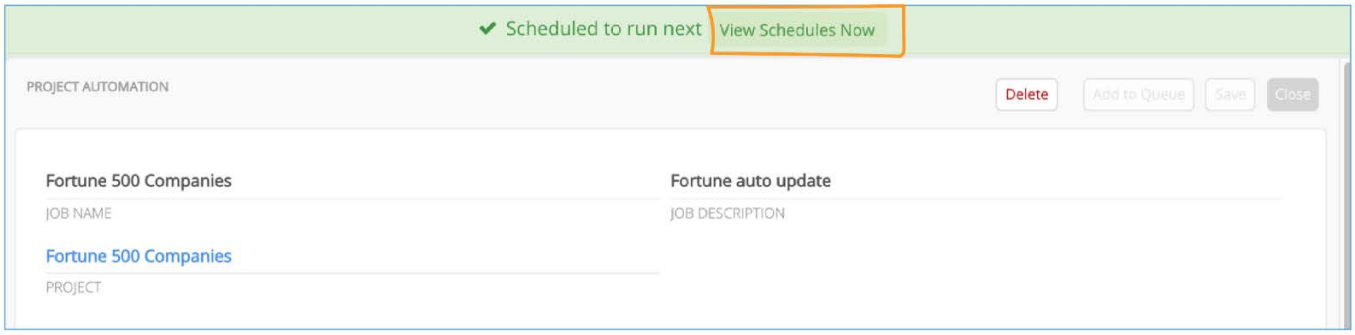
右上にある**保存**ボタンをクリックすると、このプロジェクトの自動化用のすべての設定を保存します。自動化スケジュールを保存すると、**キューに追加**ボタンが表示されます。




このボタンを使用すると、次回の自動化の開始時に実行されるジョブのキューに、この自動化ジョブを追加できます。このオプションは、スケジュールされている実行時刻まで待たずに自動化設定をテストする場合に役立ちます。

**備考**

自動化ペインには、次回の自動実行が開始されるスケジュールの詳細が表示されます。**キューに追加**ボタンをクリックすると、ヘッダーに**今すぐスケジュールを表示**というリンクが表示されます。このリンクをクリックすると、すばやくスケジュールペインに移動できます。




# 自動化ダッシュボード

自動化ダッシュボードは、自動化するように設定されているすべてのデータライブラリーファイルとプロジェクトの詳細と履歴を確認できます。ここでは:

- 自動化タスクの詳細情報を確認。
- 自動化ジョブのスケジュールを確認および管理する。
- ジョブの実行履歴と状態を確認する。
- 失敗したジョブを再実行する。

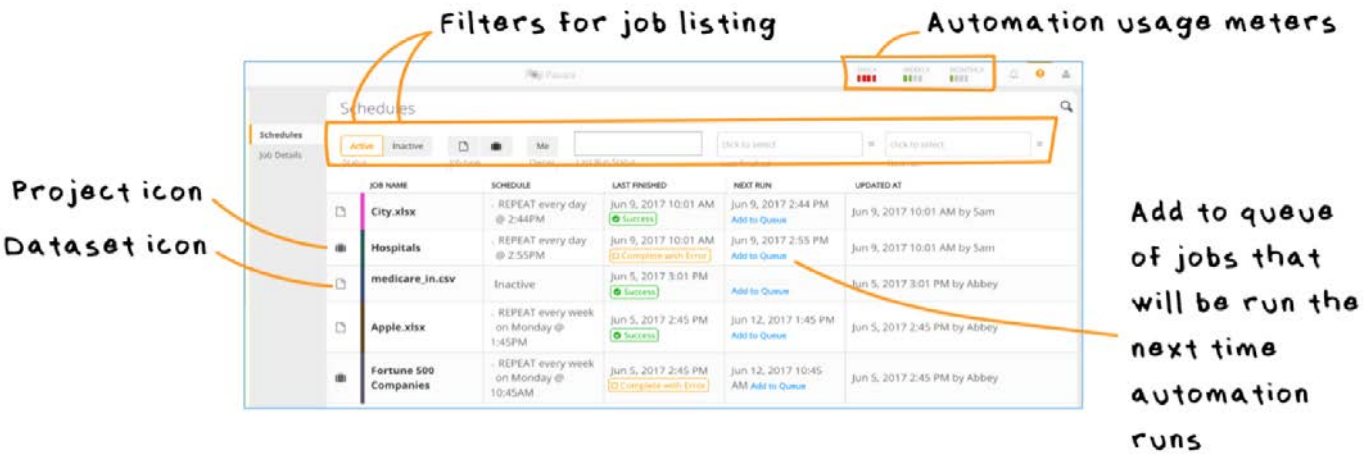
ダッシュボードは、 スケジュール とジョブの詳細で構成されています。

# スケジュール

**スケジュール**ページには、現在の自動化で構成されたデータライブラリファイルとプロジェクトのリストが表示されます。自動化の使用状況の詳細情報を確認するには、メーターの上にマウスを置くと、すでに完了した自動化ジョブの数や、日、週、月に実行できるMaximum（最大）数に関する追加情報が表示されます。

**スケジュール**ページに次のようなフィルターを適用して、表示方法をさまざまに変えることもできます。

- **自動化**あるいは**無効化**されたジョブは自動スケジュールが無効となっているものです。
- ジョブのタイプ\–**プロジェクトのみ**または**ライブラリのみ**。
- 自分が所有している自動化ジョブ。
- ジョブの状態\–**成功**、**エラーで完了**、**エラー**、または**制限超過**。各状態の意味については、ジョブ状態の定義を参照してください。
- 指定した日付範囲の間に前回の実行が終了したジョブ、または指定した日付範囲の間に次回の自動実行で実行されるジョブ。

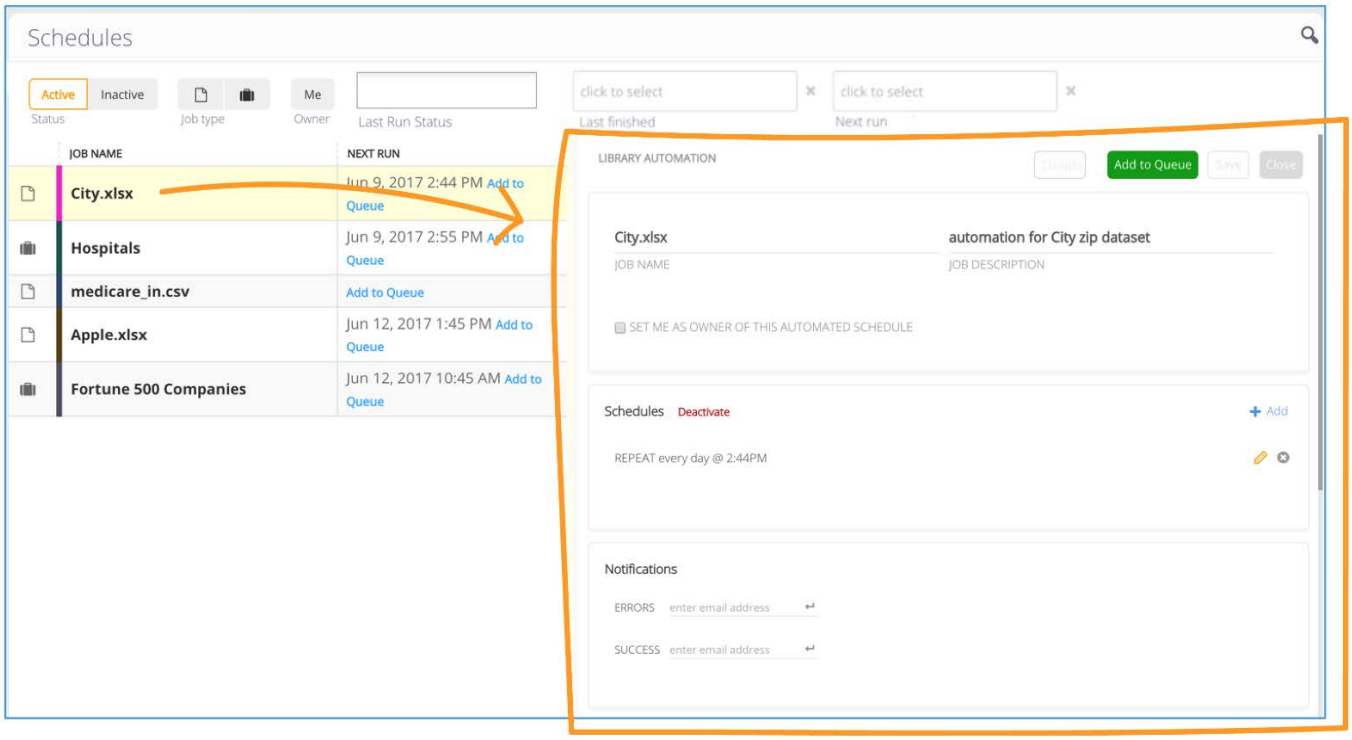



**キューに追加**リンクをクリックして、ジョブを再実行できます。これにより、ジョブに対する内部スケジュールがオンザフライで作成され、定期的にスケジュールされているジョブが自動化サービスによって次回開始されるときに、そのジョブの実行も含められます。キューに追加されたジョブを実行するためには、システムのリソースが利用可能である必要があることに注意してください。たとえば、自動化ジョブの実行に十分な数のスレッドが割り当てられている必要があります。そうでない場合は、ジョブを実行できるだけのリソースが利用可能になるまで、ジョブはキューに追加された状態のまま残ります。

**備考**

- ジョブを再実行する前にエラーを確認するには、**ジョブ詳細**タブに移動して、目的のジョブ実行の**結果**ページを開きます。
- エラーが発生したジョブを再実行するために**キューに追加**を使用する場合は、既存の自動化guardrail制限のカウントには含められません。

ジョブを無効化する場合など、ジョブの構成設定に変更を加えるには、ジョブの名前をクリックして設定ページを開きます。次に、設定の変更を行って保存します。

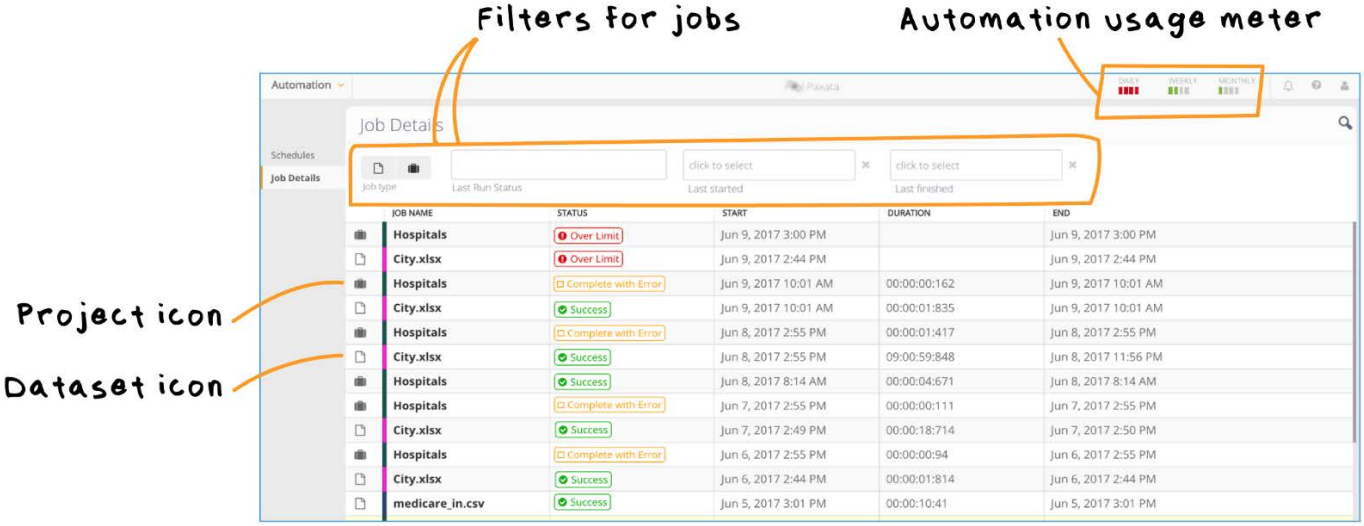



## タスクの詳細

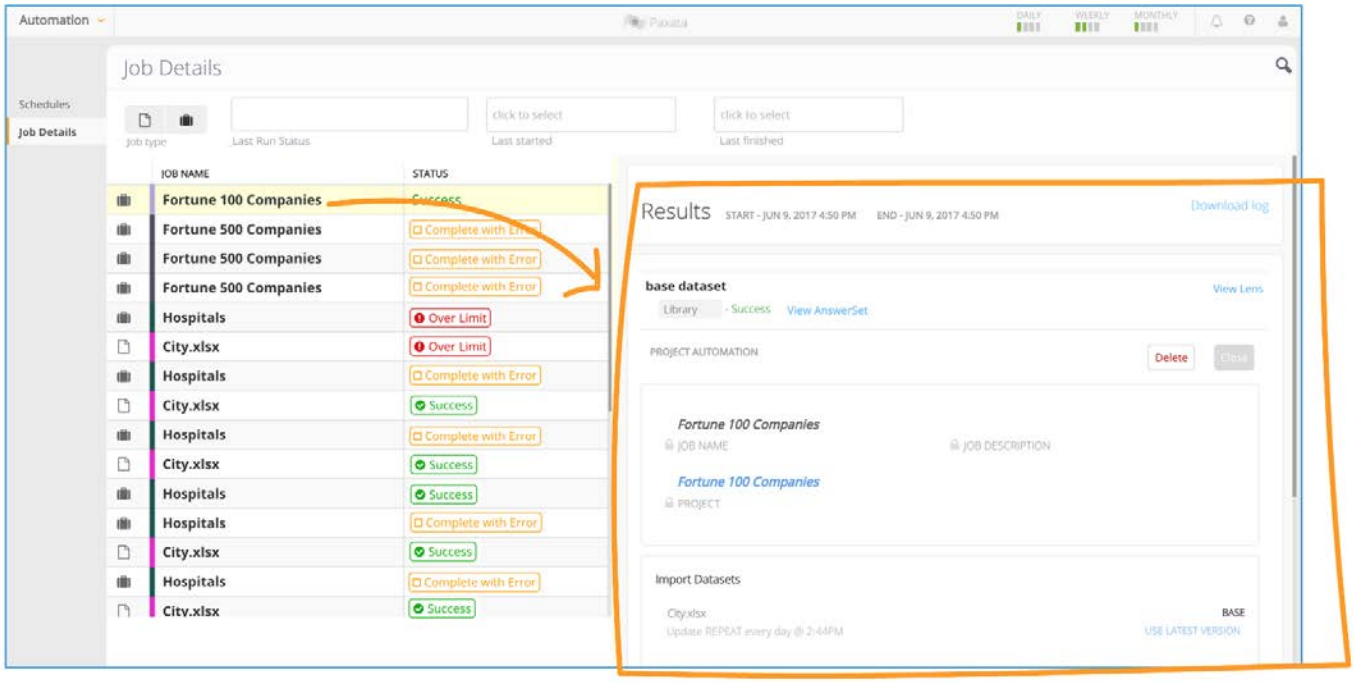
**ジョブ詳細**ページでは、削除済みの自動化ジョブを含めて、実行されたすべての自動実行の監査証跡を確認できます。自動化の使用状況の詳細情報を確認するには、メーターの上にマウスを置くと、すでに完了した自動化ジョブの数や、日、週、月に実行できるMaximum（最大）数に関する追加情報が表示されます。

**ジョブの詳細**ページをさまざまな方法でフィルターできます：

- ジョブのタイプ\–**プロジェクトのみ**または**ライブラリのみ**。
- ジョブの状態\–**成功**、**エラーで完了**、**エラー**、**制限超過**、**キュー**、**実行中**。各状態の意味については、ジョブ状態の定義を参照してください。
- 指定した日付範囲の間に前回の実行が開始または終了したジョブ。



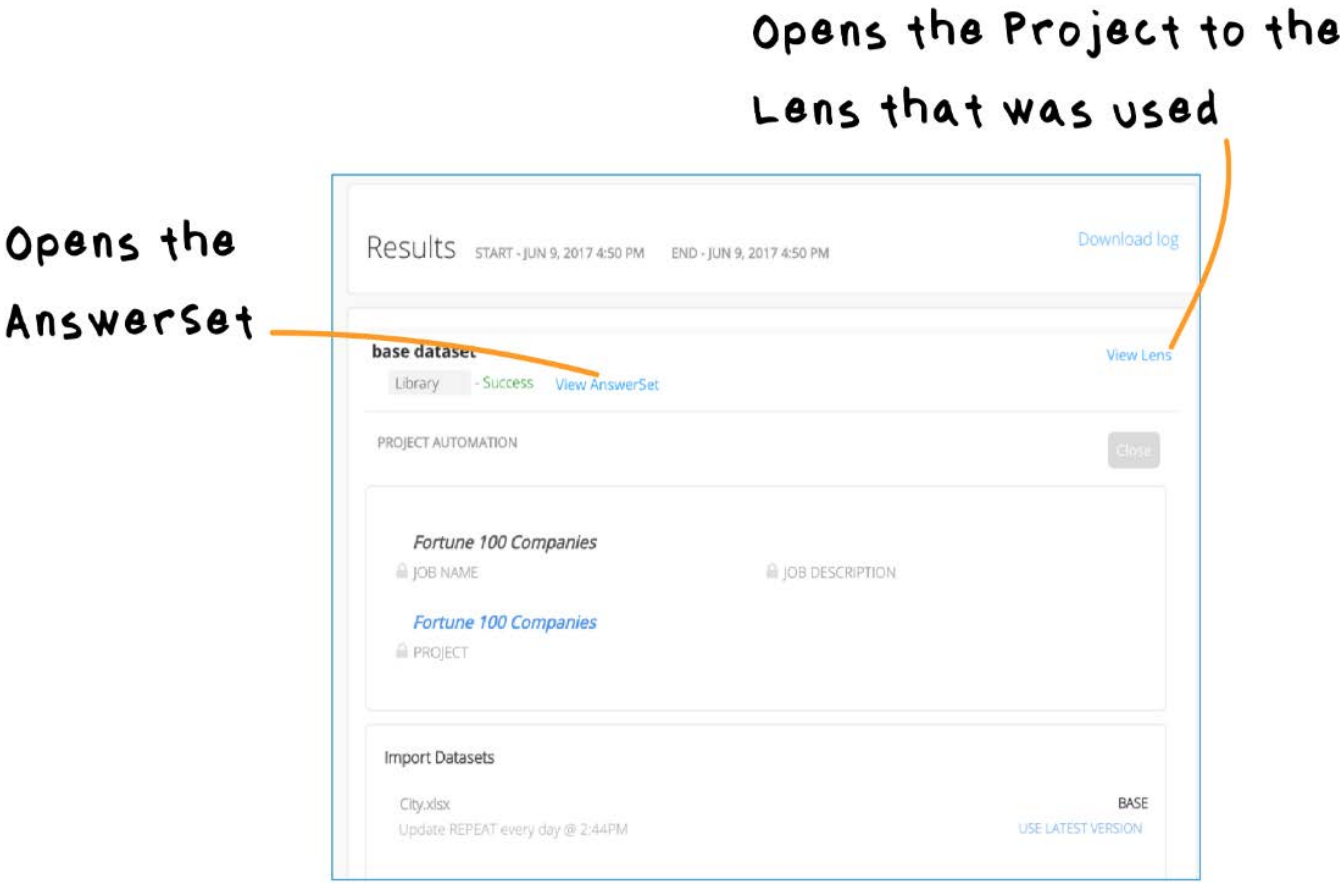
ジョブ実行の詳細を表示するには、そのジョブの行をクリックします。ジョブの**結果**ページが開き、そのジョブ実行のインスタンスに使用された構成設定のスナップショットが表示されます。



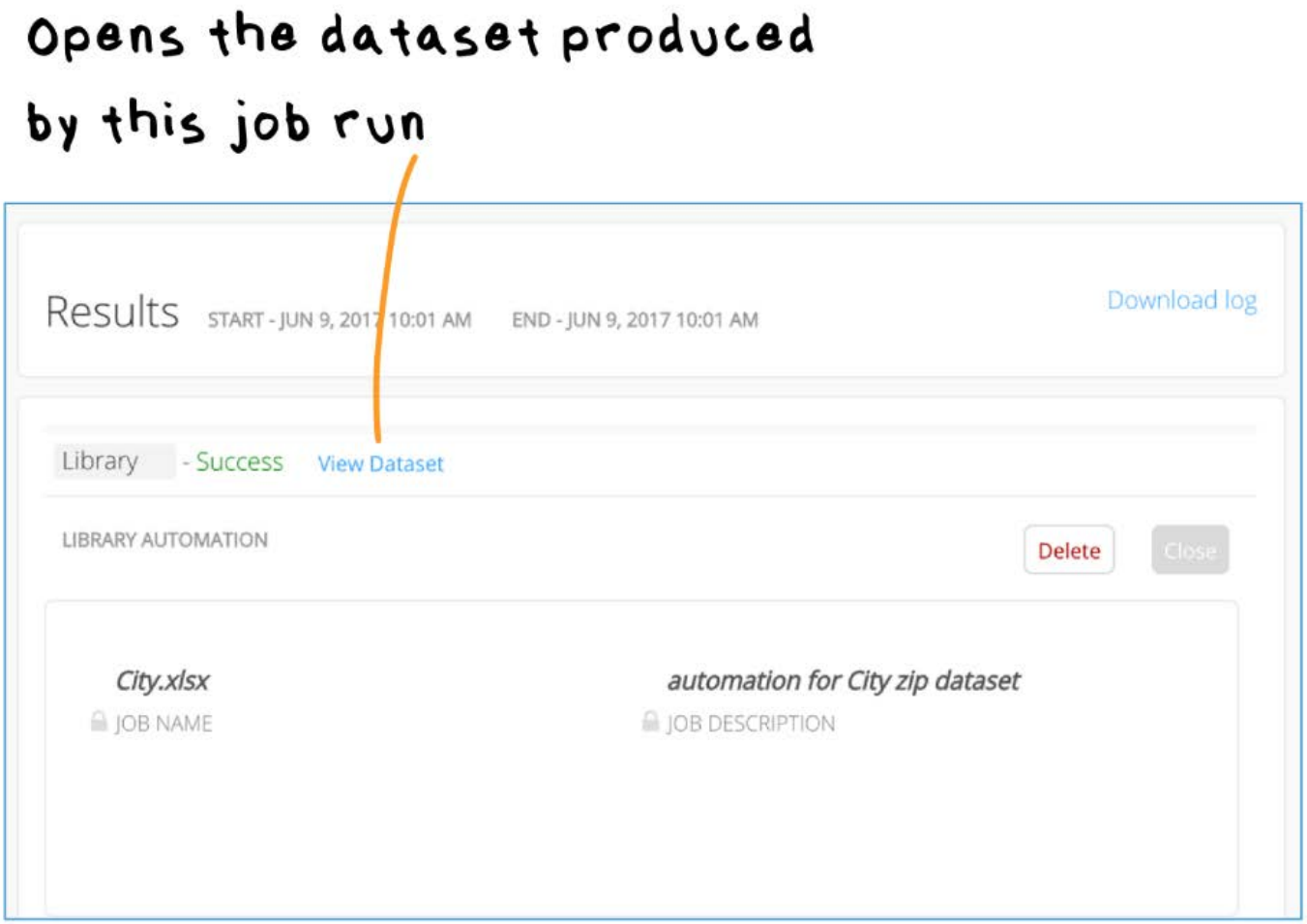

**備考**

こレはスナップショットであるため、この自動実行以降にジョブ設定が変更されている場合があります。

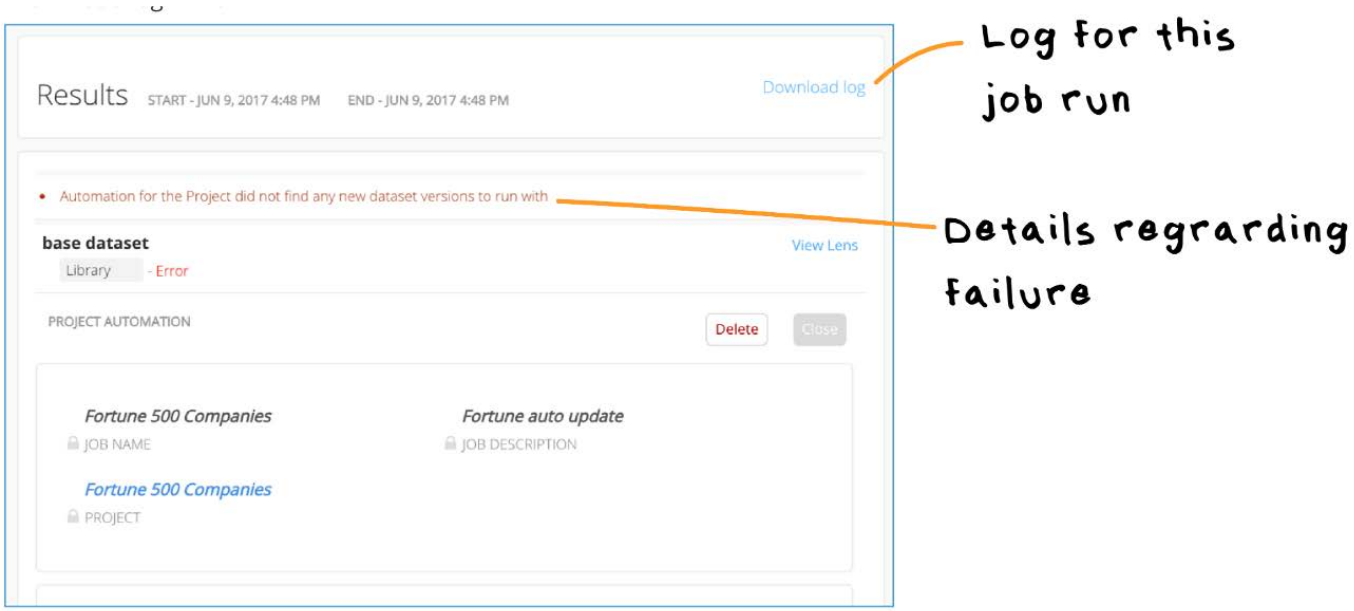
これがプロジェクトジョブの場合、**レンズの表示**リンクをクリックして、この実行にアンサーセットを公開するために使用されたレンズにプロジェクトを開きます。**AnswerSetを表示**リンクをクリックすると、この実行で公開されたAnswerSetが表示されます。




これがライブラリのジョブである場合、**データセットの表示**リンクをクリックして、データライブラリ内のファイルを開きます。




実行中にエラーが発生した場合は、ここに表示されます。**ログのダウンロード**リンクをクリックして、ジョブのログファイルをダウンロードできます。



# ジョブ状態の定義

自動化されたジョブの可能な状態は次のとおりです。

- **実行中**: ジョブの実行が現在進行中です。
- **成功**: ジョブはエラーなしで正常に終了しました。
- **エラー**: ジョブの実行に失敗しました。
- **エラーで完了**: ジョブの実行は完了しましたが、エラーが発生したため、完全には実行できませんでした。たとえば、データライブラリには正常に公開されたものの、指定されたデータ ソースにエクスポートできなかったジョブは、このタイプのエラーで完了します。
- **待機状態**: ジョブがキューに追加されると、内部スケジュールがオンザフライで作成され、定期的にスケジュールされているジョブが自動化サービスによって次回実行されるときに、自動化によるそのジョブの実行もトリガーされます。ただし、キューに追加されたジョブを実行するためには、システムのリソースが利用可能である必要があることに注意してください。たとえば、自動化ジョブの実行に十分な数のスレッドが割り当てられている必要があります。そうでない場合は、ジョブを実行できるだけのリソースが利用可能になるまで、ジョブはキューに追加された状態のまま残ります。
- **制限超過**: ジョブの実行が1日、1か月、または1週間あたりのguardrailの制限を超過すると、ジョブは**制限超過**エラーで失敗します。重要:
  - 自動化の guardrails はテナント レベルで適用されます。
  - *1週間*の自動化制限は、月曜日の00:00から日曜日の23:59までとして定義されます。
  - エラーで終了したジョブは制限に対してカウントされますが、失敗したジョブの再試行（**キューに追加**を使用した場合）はカウント*されません*。

# 自動プロジェクトフロー

**注意**

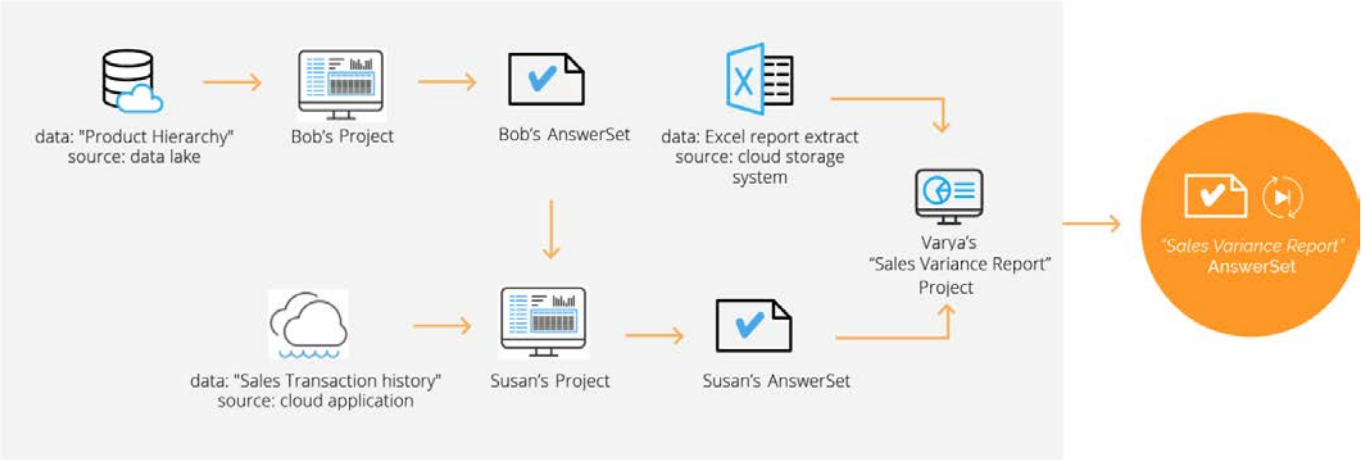
APF機能は有効化される必要があります。プロジェクトページの上部に**プロジェクトフローを作成**ボタンが表示されない場合は、Data Prepのシステム管理者にお問い合わせください。

Data Prep自動プロジェクトフロー（APF）機能を使用すれば、キュレーションされたデータの流れをインテリジェントに運用できます。APF は複数のプロジェクト、データセット、AnswerSetにまたがるデータ準備ステップのシーケンス全体を計算し、データに対応するエンドツーエンド（完結型）の自動化出力フローを生成します。時刻に基づくスケジュールでフローを繰り返し実行するか、または 1 度だけ実行して最終的結果の AnswerSet を生成するように設定してください。次に、APFの監視機能を使用してすべての実行を管理します。ビジネスアナリストやデータエンジニアは、複雑なデータの流れをData Prepプロジェクトの小さいグループに分割して、簡素化するためAPFを使用します。

APFでは、データフローの運用を可能にします。各プロジェクトが関連する、または結合度の高い一連のステップを実行することで、読みやすさを向上させ、複雑さを抑えます。プロジェクトを作成した後、シーケンスの最終プロジェクトをターゲットプロジェクトとして選択できます。残りの作業、つまりエンドツーエンドのフロー全体のシーケンシング、準備、自動化については、APFが実行し、手動の調整などは必要ありません。

APFは、チームがデータを共有し、ビジネスリーダーやITリーダーから情報を収集することに役立ちます。チームメンバーは、他の方が作成した出力AnswerSetに応じてData Prepプロジェクトを構築できます。メンバーは、自分のData PrepプロジェクトでData Prep作業を完了し、単一のターゲットプロジェクトからシーケンス全体の運用を可能にします。APF は、プロジェクトや AnswerSetの作成者にも所有者にも関係なく、手動の調整を必要とせずに作業を実行します。チームのメンバーはフローを監視し、グラフを見てプロジェクトとAnswerSetがフローの最終出力にどのように寄与しているかを確認できます。

## APFの例



この例では、APFは、複数の人が作成した一連のData PrepプロジェクトとAnswerSetから、最終状態の「売上変動レポート」を生成します。

Bobは、自身の 「製品階層」データのためにデータレイクに接続し、準備して、クラウドアプリから「販売取引履歴」のデータを取得するSusanと共有されるAnswerSetを生成します。

Susanはこのデータを準備し、AnswerSetを作成し、Varyaが管理する売上変動プロジェクトのために彼女と共有します。SusanからのAnswerSetに加えて、Varyaは、クラウドストレージからプルしたExcelレポートからのデータも結合します。

Varyaがデータの準備を終了すると、彼女は「売上変動レポート」のAnswerSetを作成します。彼女は、毎週このレポートを作成する必要があります。彼女は、売上変動プロジェクトで**プロジェクトフローを作成**をクリックし、フローを実行するための時間ベースのトリガーを設定します。最終状態のAnswerSetの生成に必要な依存関係のチェーンを作成するために、APFは関連のプロジェクト、AnswerSet、およびデータセットのフローを遡って通過します。Varyaは次に、フローの後続のすべての実行を管理するためにAPF監視インターフェースを使用します。

# APFの要件

- 関係者は、フローを作成する前に、フロー内のすべてのデータセットとすべてのプロジェクトに対する権限を持っている必要があります。持っていないと正常に実行されません。

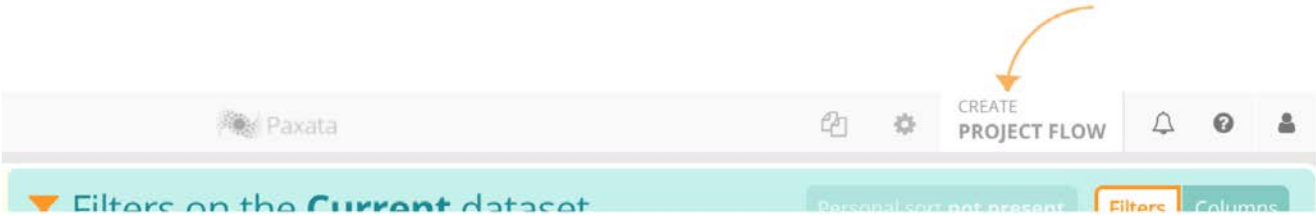
> **備考**
>
> 関係者が AnswerSet の権限を所有していれば、AnswerSet が生成された元となるプロジェクトの権限を有していない場合でも、読み取り権限が終了するまではフローを作成することができます。フロー作成におけるこの柔軟性により、関係者は独立したアクセス権限がある部分のフローの運用化を管理できます。

- 関係者は、監視インターフェースからそれらを管理するために、フロー内のすべてのデータセットおよびプロジェクトに対する権限も持っている必要があります。Data Prepのシステムの管理者は、これらの権限を提供します。
- ターゲットプロジェクトには、定義されたフローの下流で生成されたものは含まれません。前の例では、「売上変動レポート」AnswerSetがプロジェクトを使用した場合は、プロジェクトがフローに含まれません。ターゲットプロジェクトが常にフローのエンドポイントとなります。

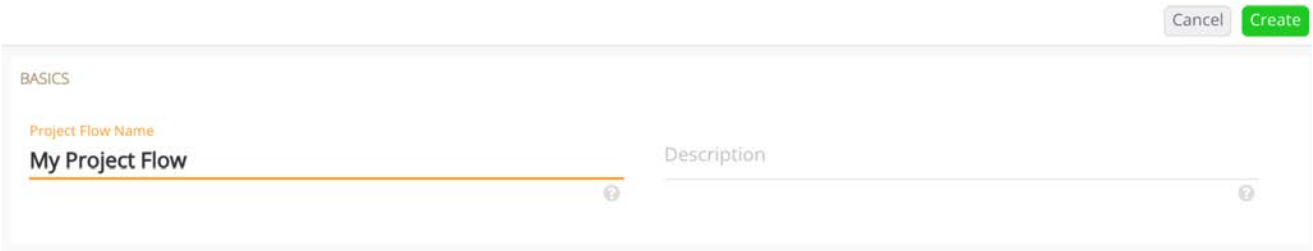
# プロジェクトフローを設定する

プロジェクトフローを作成するには：

1. ターゲットプロジェクト、つまり最終状態のAnswerSetを生成するプロジェクトを開いてください。
2. **プロジェクト**ページの右上にある**プロジェクトフローの作成**をクリックします。



3. フローの名前とオプションの説明を入力し、**作成**をクリックします。




インテリジェント自動化エンジンはフロー依存関係を計算し、APFを設定する**プロジェクトフロー**ページがAPFに表示されます。既存のプロジェクトフローを編集する際は、**プロジェクトフロー**ページにアクセスすることもできます。

すべてのフローに対して実行できる一般的なアクションについては、フローの管理を参照してください。

# APFの設定

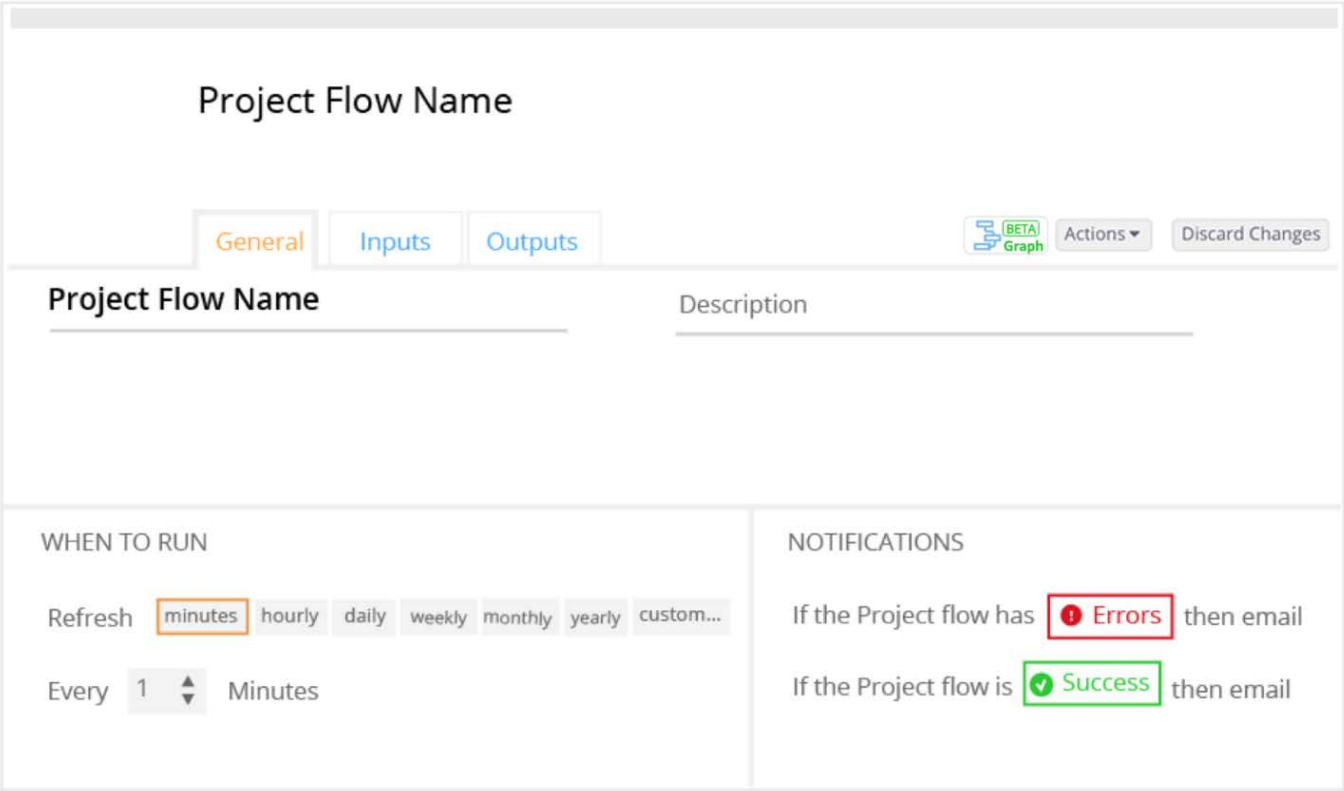
APFを設定するには、**プロジェクトフロー**ページでトリガーと通知を設定します。フローの入力および出力のデータセットの設定を調整することもできます。

**プロジェクトフロー**ページには、フロー設定用に3つのタブが表示されます。

- **一般**タブ
- **入力**タブ
- **出力**タブ

## 一般タブ

**一般**タブは、プロジェクトフローの詳細を更新し、トリガーを追加するのに使用します。



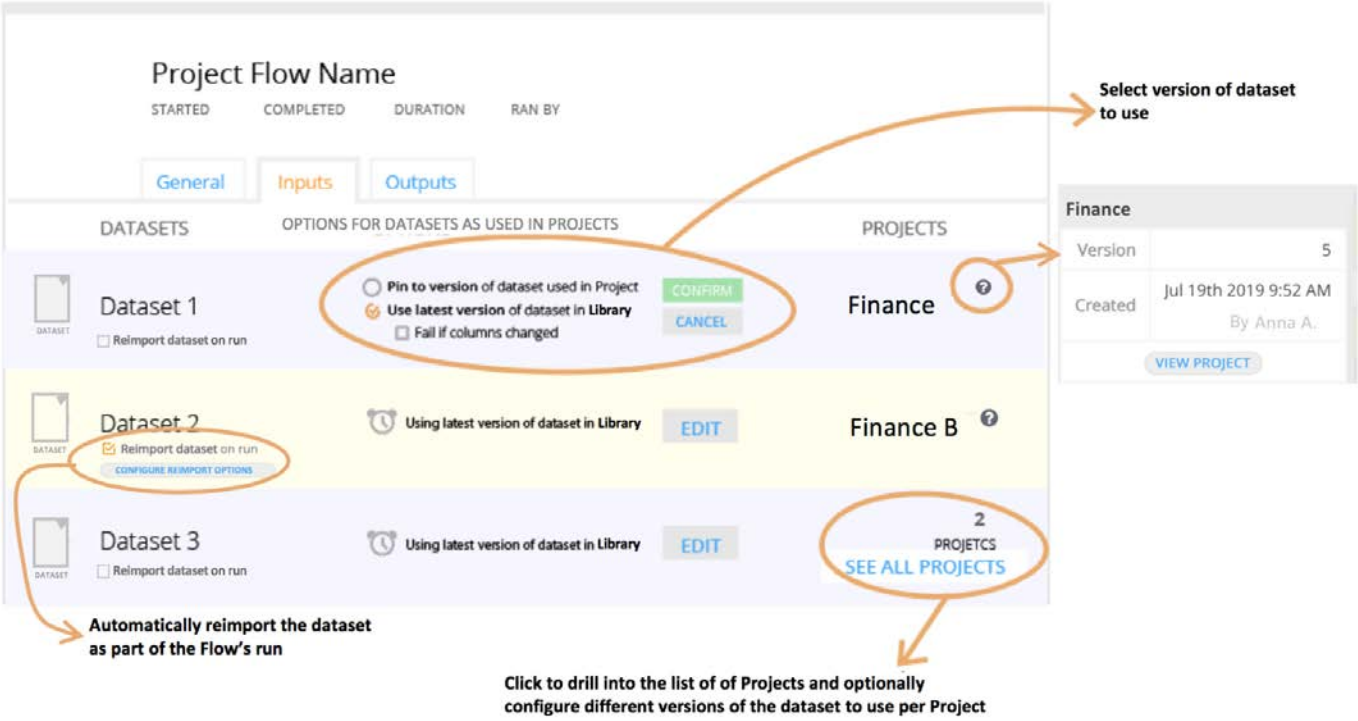

**一般**タブでは、以下の操作ができます。

- 作成したフローの **名前**と**説明**を更新します。
- フローを実行するためのトリガーを指定します。トリガーは時間に基づくものと頻度に基づくものがあります。また、**カスタム**オプションを使用し、トリガー用の cron 式を入力することもできます。
- 実行ステータスの通知先メールアドレスを入力します。各アドレスはコンマで区切ってください。

> **備考**
>
> フローが作成されると、**一般**タブに**プロジェクトIDフロー**が表示されます。このIDは、REST APIコールのフローを識別し、フローのトラブルシューティングを行うために使用されます。

## 入力タブ

**入力**タブには、フローで使用するデータセットのリスト、フローの作成に使用するデータセットのバージョン、および各データセットが使用されるプロジェクトが表示されます。




**入力**タブでは以下の操作ができます。

- フローが実行されるたびにデータセットが自動的に再インポートされるように指定します。

  デフォルトでは、すべてのプロジェクトは、ライブラリに保存されているデータセットの最新のバージョンを使用するように設定されています。ただし、新しいバージョンのデータセットが、その新バージョンがData Prepライブラリに手動でインポートされる前に、元のデータソースから利用可能となる可能性もあります。この場合、フローが実行されるたびに、元のデータソースから自動的に再インポートされるように、データセットを設定することができます。その後、この最新バージョンが常にライブラリに保存されます。この自動更新を有効にするには、**Reimport dataset on run**（実行時にデータセットを再インポート）をクリックします。オプションが有効にされている場合、**再インポートオプションの設定**ボタンも表示されます。このボタンは、データソースのパスの変更、クエリの実行、エクスポート解析オプションの入力を行うことができるライブラリのインポートペインを開きます。これらのオプションは、データセットとともにライブラリに保存されます。オプションの設定は、現在の設定を変更したい場合にだけ行う必要があります。

- プロジェクトに使用するデータセットのバージョンを設定します。

  デフォルトでは、すべてのプロジェクトは、ライブラリに保存されいるデータセットの最新バージョンを使用するように設定されてます。ただし、このデフォルト動作を変更しようとする場合、**データセットをプロジェクト内で使用する場合のオプション**列の**編集**をクリックすることによって変更できます。

  - **特定バージョンに固定**：データセットは、プロジェクトが現在使用している正確なバージョンを指定し続けます。
  - **列が変更されている場合は失敗する**：ライブラリから入ってくる最新バージョンのレイアウト（スキーマ）が異なる場合は、プロジェクトへのデータセットのインポートが失敗します。例えば、新しい列が追加された場合、プロジェクトのステップで使用されていない列が削除されない場合、列タイプが変更された場合、列の順序が変更された場合などです。

- 複数のプロジェクトがフローの入力として同じデータセットを使用している場合、これは、プロジェクトの列に記述されます。

  データセットを使用するすべてのプロジェクトを表示し、オプションで、プロジェクトごとに使用するデータセットの異なるバージョンを設定するには、**すべてのプロジェクトを表示**をクリックします。例えば、あるプロジェクトがライブラリの

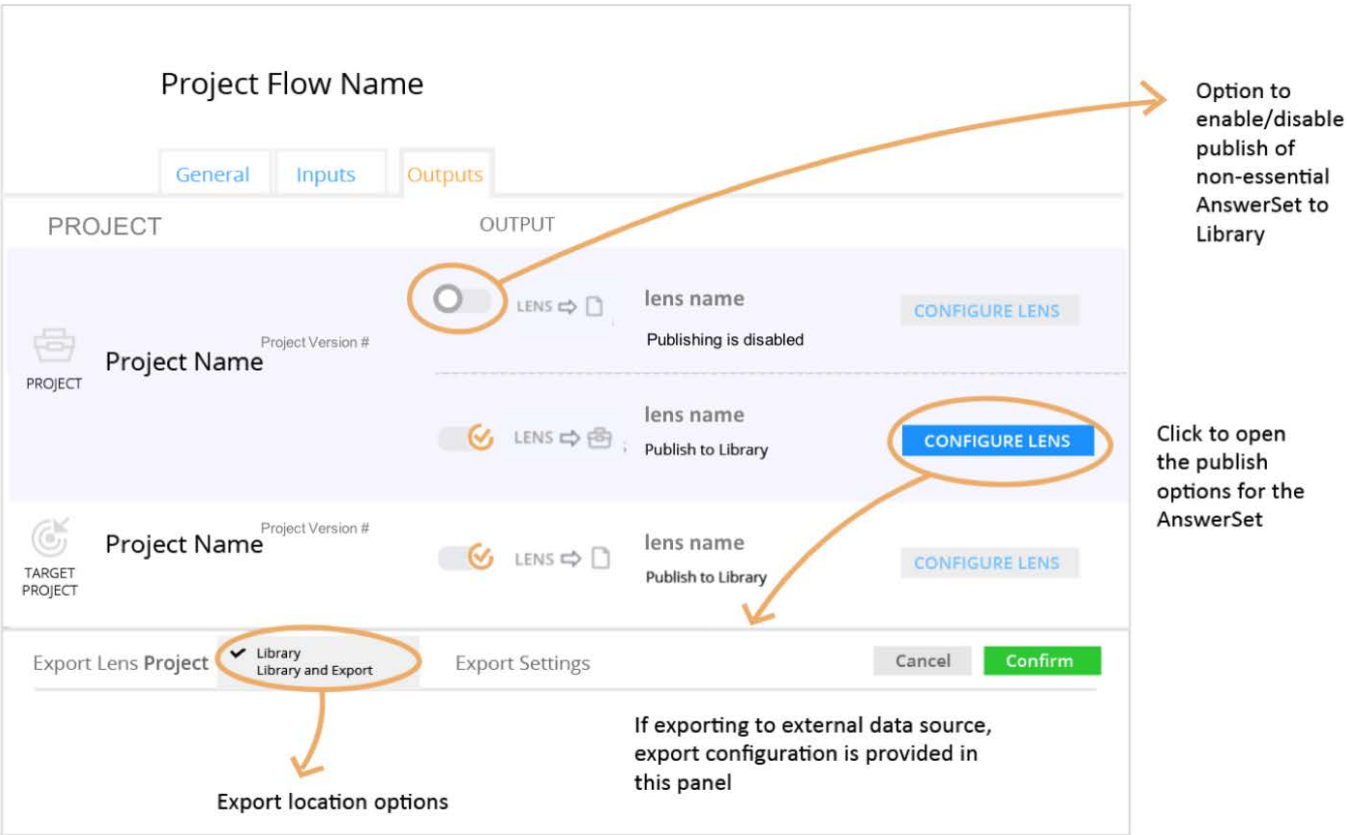
最新バージョンのデータセットを使用し、その一方で、別のプロジェクトが、関連するバージョンのプロジェクトに現在保存されている、データセットの正確なバージョンを使用するように指定できます。

**備考**

**データセット**列のデータセット名にカーソルを合わせると、データセット入力のメタデータセット統計が表示されます。データセットのバージョン、作成の日付、ライブラリにデータセットを追加したユーザー、および列と行の数がポップアップウインドウに表示されます。

## 出力タブ

**出力**タブは、フローから公開されているすべての出力 AnswerSet のリストを表示します。




Data Prepプロジェクトから公開ポイントを作成するには、公開するレンズが常に必要なので、すべての出力はレンズレベルで設定されます。

フローに複数のレンズがあるプロジェクトが含まれる場合があります。出力AnswerSetを生成するために、これらすべてのレンズが必要なわけではありません。デフォルトでは、必要なレンズだけが ライブラリ内に保存されたAnswerSet を自動的に公開します。フローに必要とされていない場合でも AnswerSetsを公開したい場合、**出力**タブで有効にできます。

**備考**

出力AnswerSetを生成し、フローに必要であるレンズは、決して無効化できません。

必須でないAnswerSetの公開オプションを調整することに加えて、任意のレンズの出力AnswerSetを、例えばデータベースやクラウドストレージシステムなどの外部データソースに公開することができます。Data Prepライブラリに加えて公開場所を指定するには、**レンズを設定**をクリックして、**エクスポート**ペインを開きます。

**出力**タブで、次の操作を実行できます。

- ブリッジ機能を果たさないレンズを無効化して、AnswerSetをライブラリに公開させないようにします。

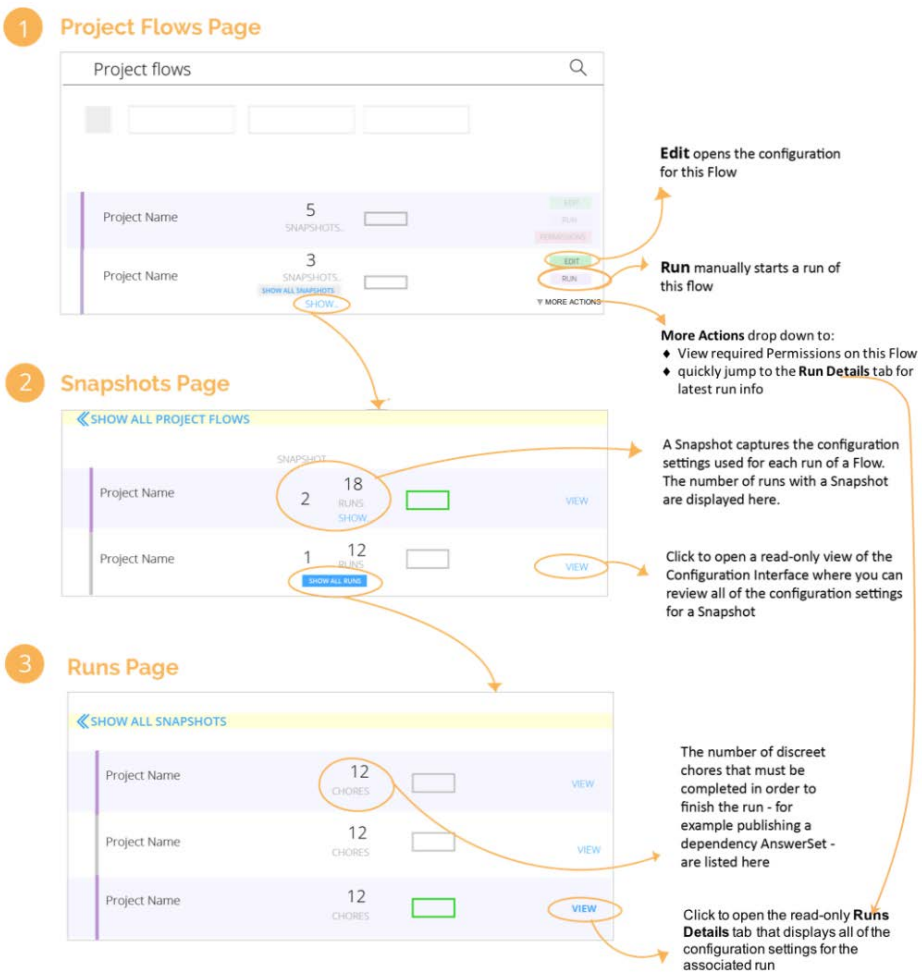
  無効化するにはレンズに隣接するスライダーをクリックしてください。

- 公開済みの AnswerSet を、（デフォルトのライブラリ設定に加えて）データソースにエクスポートします。

  レンズに対して**レンズの設定**をクリックします。**エクスポート**ペインがページの下部に開きます。デフォルトでは、Data Prepは、Data PrepライブラリにAnswerSetsを公開します。外部データソースに公開するには、**レンズのエクスポート**フィールドのドロップダウンメニューをクリックし、**ライブラリとエクスポート**を選択します。その後、そのAnswerSetに関する出力場所の詳細と、エクスポート解析オプションを指定できます。

# フローを監視する

APFでは、フローのステータスを監視できます。フローの出力を生成するための主要な構成要素は、スナップショット、実行、および操作です。次の図は、これらの構成要素がフローを監視する方法を示しています。詳細については、次のセクションを参照してください。




# プロジェクトフローページ

**プロジェクトフロー**ページには、表示および編集する権限があるフロー、およびそれぞれに関して最新の実行の現在のステータスが表示されます。このページでは、次のことができます：

- フローの構成の詳細を編集します。**編集**をクリックすると、APF 構成インターフェースが開き、ここで構成に関する調整を行うことができます。APFの設定を参照してください。
- **実行**をクリックして、フローを手動で実行します。フローを手動で開始することは、新しいフロー、またはフローの設定変更を確かめる必要がある場合や時間に基づくトリガーが開始するのを待ちたくない場合に、特に役立ちます。
- フローのスナップショットを表示します。**すべてのスナップショットを表示**をクリックして、[スナップショット] ペインを開きます。
- **その他のアクション > 権限**をクリックして、このフローを他の人と共有できるように、権限の設定を更新します。これらの権限を表示できるのが、そのフローを作成したユーザー、または作成者がすべての権限を共有したユーザーのみであることに注意してください。
- **その他のアクション > 最新の結果を表示**をクリックして、最新のフローに移動します。これは、フローが少なくとも一回実行されるまで表示されません。

## スナップショットページ

**スナップショット**ページには、フローのスナップショットが表示されます。 フローが実行されるたびに（フローの「実行」と呼ばれます）、スナップショットが作成され、実行の出力を作成するために使用される構成設定がキャプチャされます。フローの設定が変更（スケジュール、通知、入力、出力設定の変更など）されるまでは、このスナップショットで実行が継続します。その後、フローの新しいスナップショットが作成されます。新しいスナップショットは、変更された構成の設定で実行される実行をキャプチャします。スナップショットは、実行ごとにプロジェクトフローの正確な状態の監査を可能にします。

> **備考**
>
> データセットがライブラリの最新バージョンを使用するように設定されている場合、APFは新しいスナップショットを作成しません。データセットの設定オプションについては入力タブを参照してください。

このページでは、次のことができます：

- **表示**をクリックして、スナップショットのAPF設定の読み取り専用表示を開きます。
- **すべての実行を表示**をクリックして、**実行リスト**ページを開きます。ここには、スナップショットの各実行の詳細が表示されます。

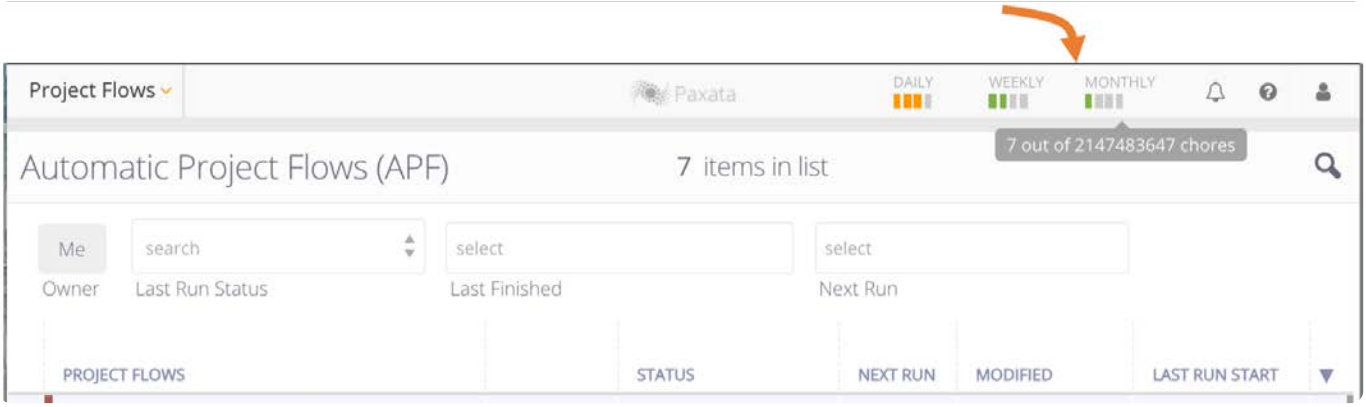
## 実行ページ

**実行リスト**ページでは、スナップショットの下の各実行ごとに、すべての詳細がキャプチャされます。実行を終了するために完了する必要があるその他の操作の数（依存関係AnswerSetの公開など）がページに表示されます。フローが実行されるたびに、新しい実行エントリーがこのページに表示されます。

実行に関連したAPF構成設定の読み取り専用表示を開くには、**表示**をクリックします。

> **備考**
>
> フローの作成に使用されたデータに変更がない場合（例えば、フローに使用されたすべてのデータセットが前回の実行時に使用されたものと全く同じバージョンである場合）、APFエンジンはリソースを節約し、新しいデータの入力が可能になるまでフローを再実行しません。

APF割り当てメーターは、使用状況を知らせるためにフローページの上部に表示されます。日次、週次、月次のいずれかにカーソルを合わせます。ツールチップは、現在の使用状況と制限の詳細を提供します。



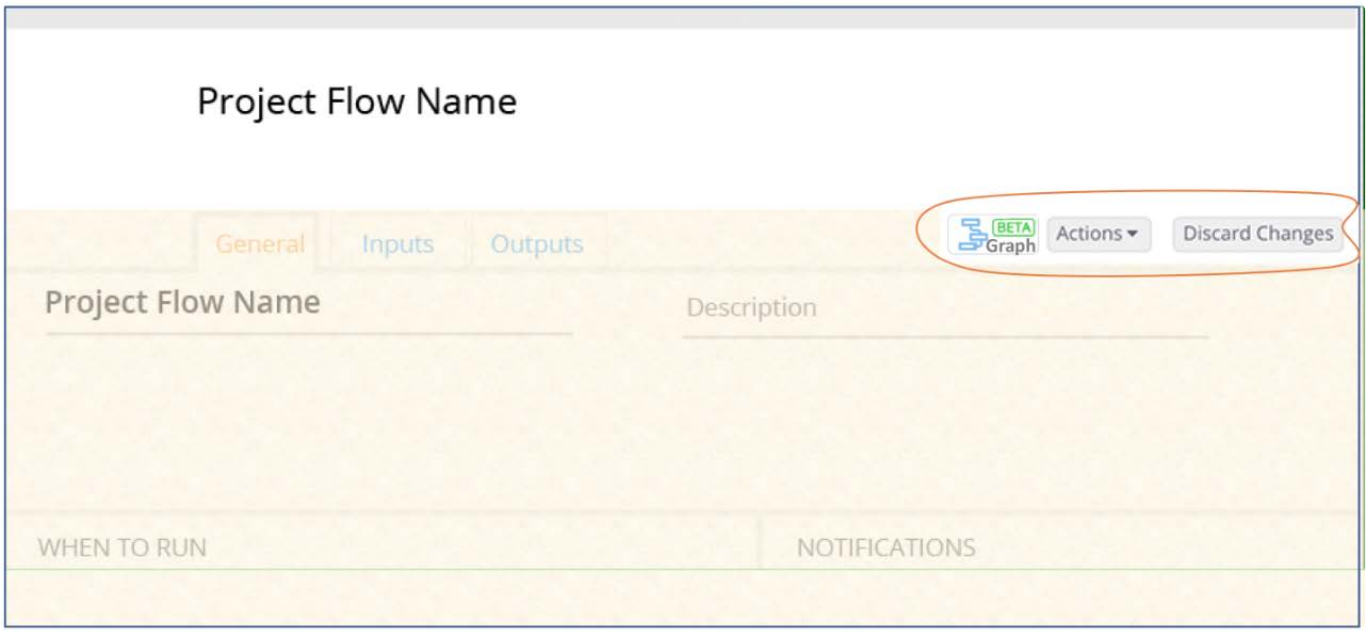

割り当ては操作の数に基づいており、操作は次のように定義されます。

- フローを生成するために個々のプロジェクトの実行が必要です。
- フローの生成に必要なデータセットまたはAnswerSetのインポート（公開ではなく）。

すべての操作は最終的にフローの出力を生成します。フローの実行中に、フローのページの割り当てメーターを更新するためにブラウザの表示を更新してください。操作の数の割り当てを増やす必要がある場合は、 DataRobot Data Prep管理者またはDataRobotカスタマーサクセスに連絡してください。

# フローの管理

**プロジェクトフロー**ページの右上にあるフロー管理用のツールにアクセスします。



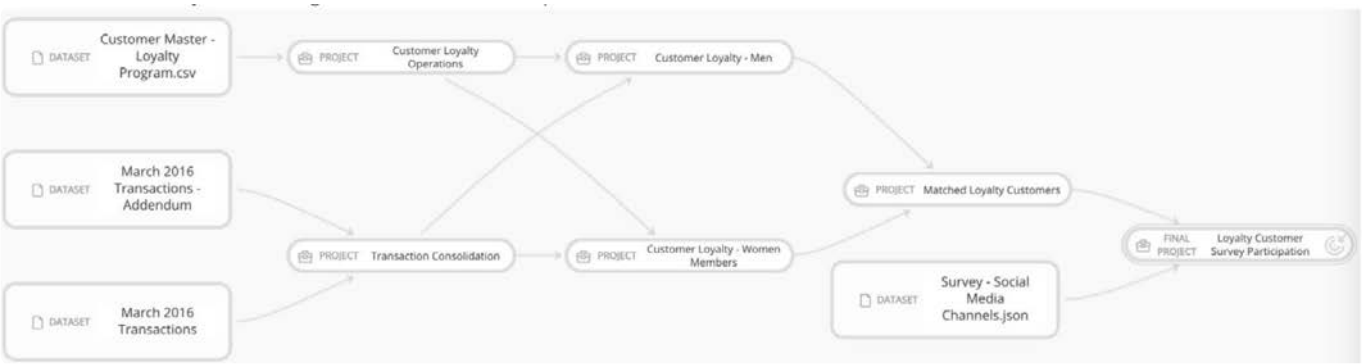

保存されたフローは、次の方法で管理することができます。

- フローの視覚的なグラフを生成する
- フローを手動で実行する
- フローを削除する

- 最新のプロジェクトのバージョンを使用するためにフローを更新する

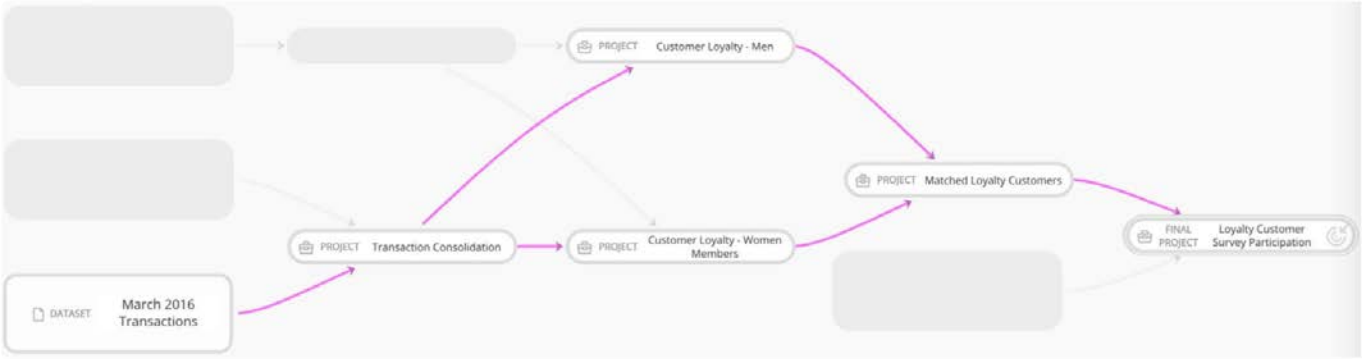
## フローに対応するビジュアル グラフの生成

**グラフ**ボタンをクリックすると、新しいブラウザーウィンドウ内でAPFグラフが生成されます。 このグラフは、データセットを表示するほか、フローの最終出力AnswerSetを生成するために使用される個別プロジェクトに対してこれらのデータセットがどのように流入するかを示します。

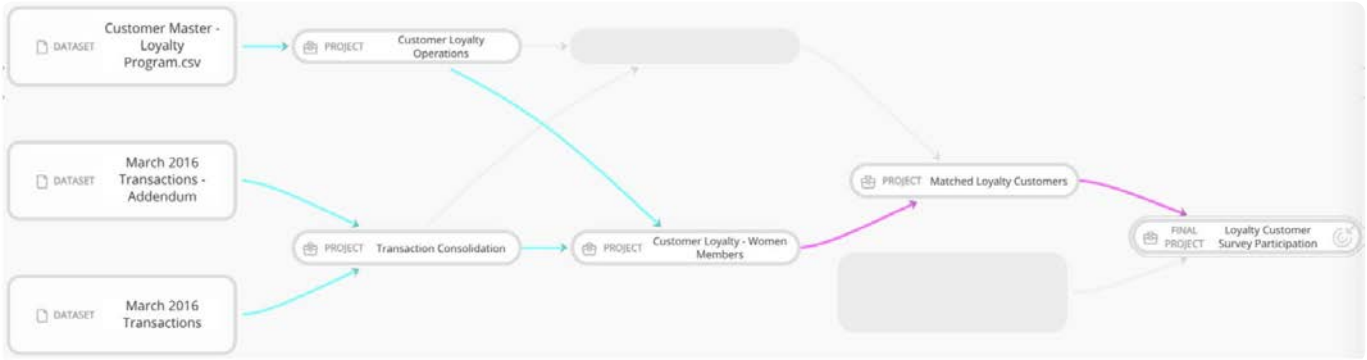



フロー内のデータセットまたはプロジェクトにカーソルを合わせると、対応するダウンストリーム系統（ピンク）とアップストリーム依存関係（青）が表示されます。

たとえば、**2016年3月トランザクション**のデータセットにカーソルを合わせると、次のように表示されます。




フローの中間プロジェクトにカーソルを合わせると（この例では**顧客忠実度 - 女性メンバー**です）上流の依存関係はブルーで表示し、下流の系統はピンクで表示されます。



両方の例において、データセットとプロジェクトが選択したフローの部分に関与していない場合、グラフ内でグレーアウトされていることがわかります。

一部のフローのグラフ内に点線が表示されることもあります。点線は、フロー内のプロジェクトからAnswerSetが公開され、その後、その AnswerSet が同じフロー内の同じプロジェクトまたは別のプロジェクトによって使用されたことを示します。この状況をループ入力と呼び、点線で表現されます。

## フローを手動で実行

時には、スケジュールした開始時刻を待つ必要なしで、フローの実行を手動で開始しようと考えることがあります。これは、**アクション**ドロップダウンメニューから実行できます。**今すぐ実行**をクリックします。

## フローの削除

保存されたフローを保持しない場合は削除できます。**アクション > 削除**をクリックします。選択を確認するように求められます。このフローを削除した結果、このフローを実行した結果としてライブラリに公開されたどの AnswerSetも削除されないことに注意してください。

## 最新プロジェクトバージョンでフローを更新する

プロジェクトでアクション（ステップの追加、ステップの削除、ステップの再配置など）が実行されるたびに、プロジェクトの新しいバージョンが作成されます。各バージョンは、データ準備作業中にユーザーがデータに対して実施した変更の監査証跡を提供します。プロジェクトフローを作成する場合、フローは、そのフローの作成時点で特定のプロジェクトバージョンに常時固定されます。しかし、フローがすべてのプロジェクトの最新バージョンを使用するように更新することもできます。これは、**アクション**ドロップダウンメニューから実行できます。**プロジェクトのバージョンを更新**を選択すると、選択の確認を求めるプロンプトが表示されます。


既存のAPFを上書きすることも、新しいAPFを作成することも選択できます。新しいAPFを作成することにした場合、すべてのトリガーは新しいAPFにコピーされますが、デフォルトで無効化されています。

**備考**

既存のAPFを更新する機能を有効化する必要があります。**すべてのプロジェクトバージョンの更新**ウィンドウが表示されない場合は、Data Prepシステム管理者に連絡してこの機能を有効にする必要があります。この機能を有効にしていない場合、注意メッセージが表示され、プロジェクトに重要な変更がなかった場合（レンズのプロジェクトデータセットに変更がない場合など）にのみバージョンを更新できます。

フローのすべてのプロジェクトではなく、特定のプロジェクトのバージョンを更新するには、**出力**タブで、バージョンを更新するプロジェクトにカーソルを合わせてから、右側の列で**プロジェクトバージョンの更新**をクリックします。

# APFの用語

以下は、APFに対応する固有の用語です。

| 用語 | 定義 |
| --- | --- |
| 操作 | データセットのインポートあるいはプロジェクトの実行。データセットのインポート操作は、データソースを介してデータセットの再インポートを実行します。プロジェクトを実行する操作は、ライブラリへのAnswerSetの公開や、AnswerSetのエクスポートなど、フローに必要なその他すべてのタスクに対応します。 |

| 用語 | 定義 |
| --- | --- |
| フロー | 1 つの単位として実行できる一連のプロジェクトを意味します。1 つのフローに対して、頻度をベースとする 1 つ以上のスケジュールを関連付けることができ、その結果、1 つのフローを繰り返し実行する ことが可能になります。 |
| 入力 | フローの実行に必要な、ライブラリから取得されるデータセットです。 |
| 出力 | フローを実行して生成されたライブラリに書き込まれる複数の AnswerSet を意味します。 |
| 実行 | ターゲットプロジェクトにとって必要である、各プロジェクトの実行です。実行は、アップストリームの依存関係プロジェクトからすべての段階を実行し、結果のAnswerSetをライブラリに書き込みます。 |
| スナップショット | フローの実行ごとにキャプチャされる構成の設定。Data Prepの管理者は、アプリケーションでこの特徴量を有効にする必要があります。 |
| ターゲット プロジェクト | フローが作成される元となるData Prepプロジェクト。フローを作成すると、すべてのアップストリームの依存性はAPF エンジンによって自動的に計算されます。 |

# Data Prepの高度なトピック

Data Prepは、データを保護し、インタラクティブな処理をサポートする機能を提供します。また、ビジネスインテリジェンス（BI）およびデータ視覚化ツールからData PrepステップおよびFiltgramsへの*ClicktoPrep*リンクを生成します。

このページでは以下の内容について説明します。

| トピック | 説明... |
| --- | --- |
| ClicktoPrepリンク | ビジネスインテリジェンス（BI）およびデータ視覚化ツールに含めることができるData PrepステップまたはFiltergramsへのリンクを生成します。 |
| インフラストラクチャとセキュリティ | Data Prepのインフラストラクチャとセキュリティ機能について説明します。 |
| 交互作用モード | データの一部に対してデータ準備を実行できるインタラクティブモードを有効にして、大量のデータをData Prepにインポートする必要がないようにします。 |

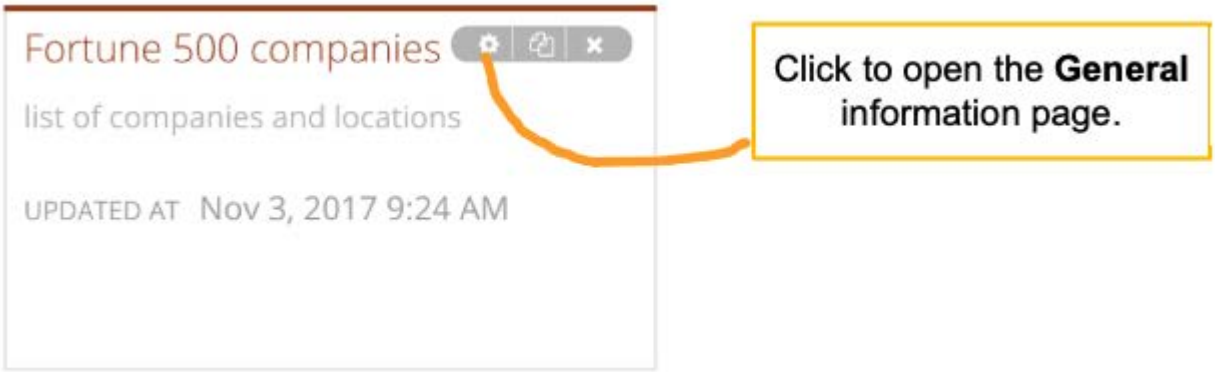
# ClicktoPrepリンク

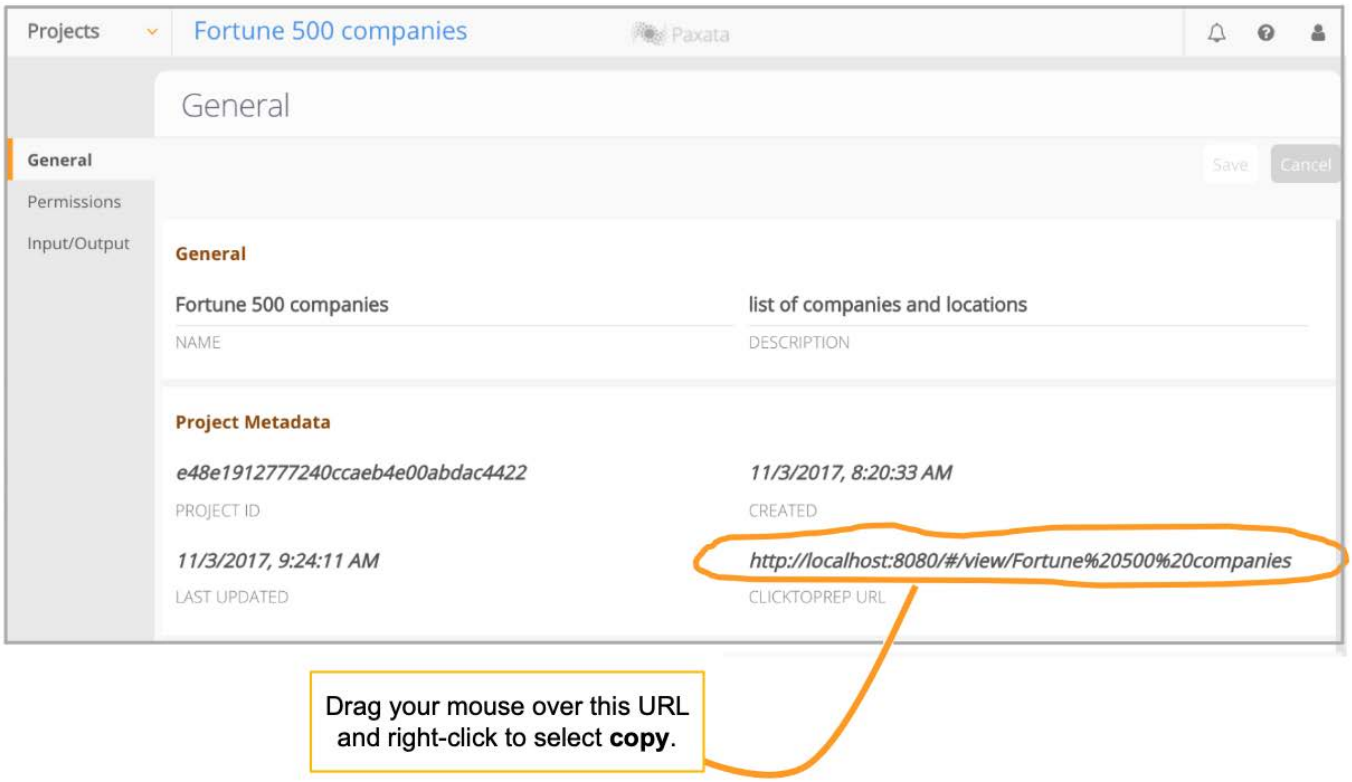
Data Prepはデータ準備製品として、頻繁にTableau®、Power BI®、Qlik®、MicroStrategy®などのビジネスインテリジェンス（BI）やデータ視覚化ツールと組み合わせて使用されます。ハイパーリンクまたはURLに対応したツールの場合は、ツールから直接次の場所に移動するData Prep ClicktoPrepのリンクを生成できます。

- Data Prepプロジェクトの最後のステップ
- Data PrepプロジェクトのFiltergram
- データを生成または変更した特定のData Prepプロジェクトステップ

BIツールや視覚化ツールからリンクを開いたり、Data Prepデータを変更したり、再発行したり、視覚化やレポートを更新して更新データを反映させることができます。

まず、プロジェクトのベース*URL*が必要です。 これにより、Data Prepプロジェクトの最後のステップに移動します。プロジェクトの**一般**ページからプロジェクトのベースURLをコピーします。

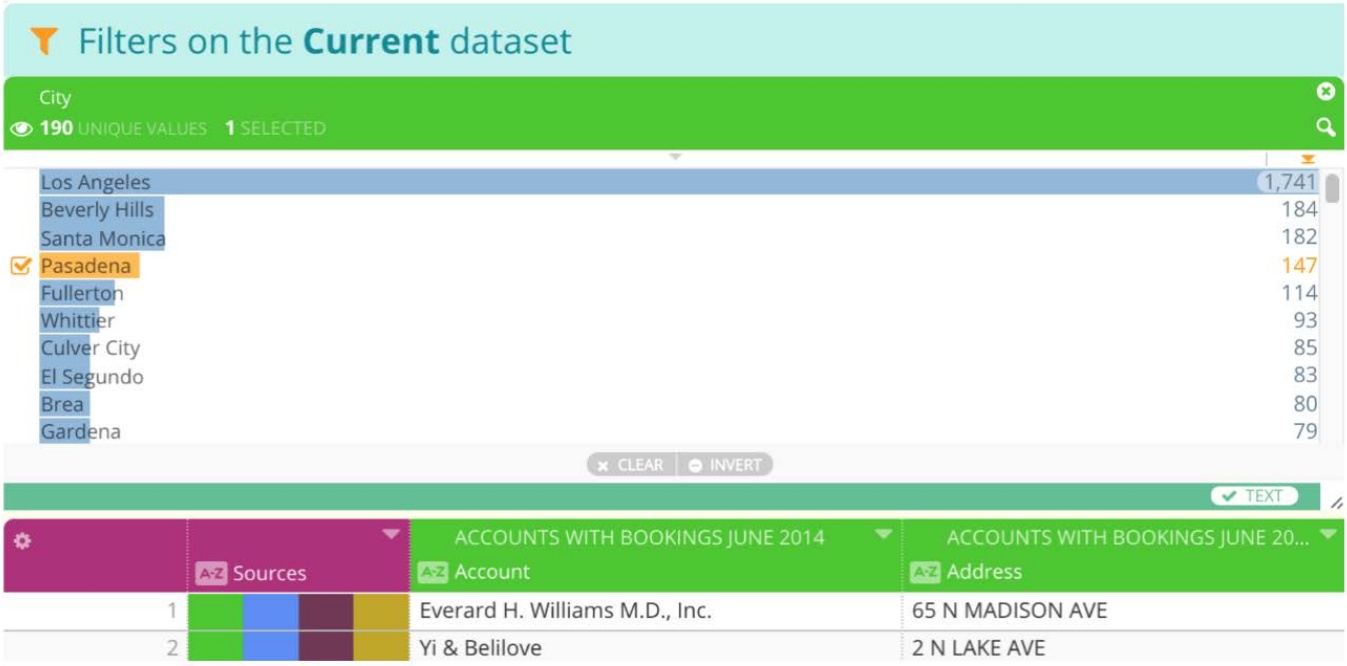





このベースURLを使用して、視覚化ツールからプロジェクトの最後のステップを指すことができます。または、以下に説明するパラメーターを追加して、Data Prepプロジェクトの特定のFiltergramまたはステップを指すようにすることもできます。

# プロジェクトのFiltergramへのClicktoPrepリンク

このタイプのリンクは、データセットで指定した列と値に関するData PrepプロジェクトのFiltergramを開きます。たとえば、次のClicktoPrepリンクは**Customers**プロジェクトに移動し、**Pasadena**にフィルターされた**City**の列でFiltergramを開きます。



このタイプのリンクを作成するには、プロジェクトのFiltergramへのリンクの設定を参照してください。

# プロジェクトステップへのClicktoPrepリンク

このタイプのリンクは、次のいずれかに設定された編集モードでData Prepプロジェクトを開きます。

- 列を編集または変更する最後のステップ。
- 列を編集または変更する最後の[StepType]（列の [検索と置換] ステップなど）。

ClicktoPrepで対応している [StepTypes]は次のとおりです。

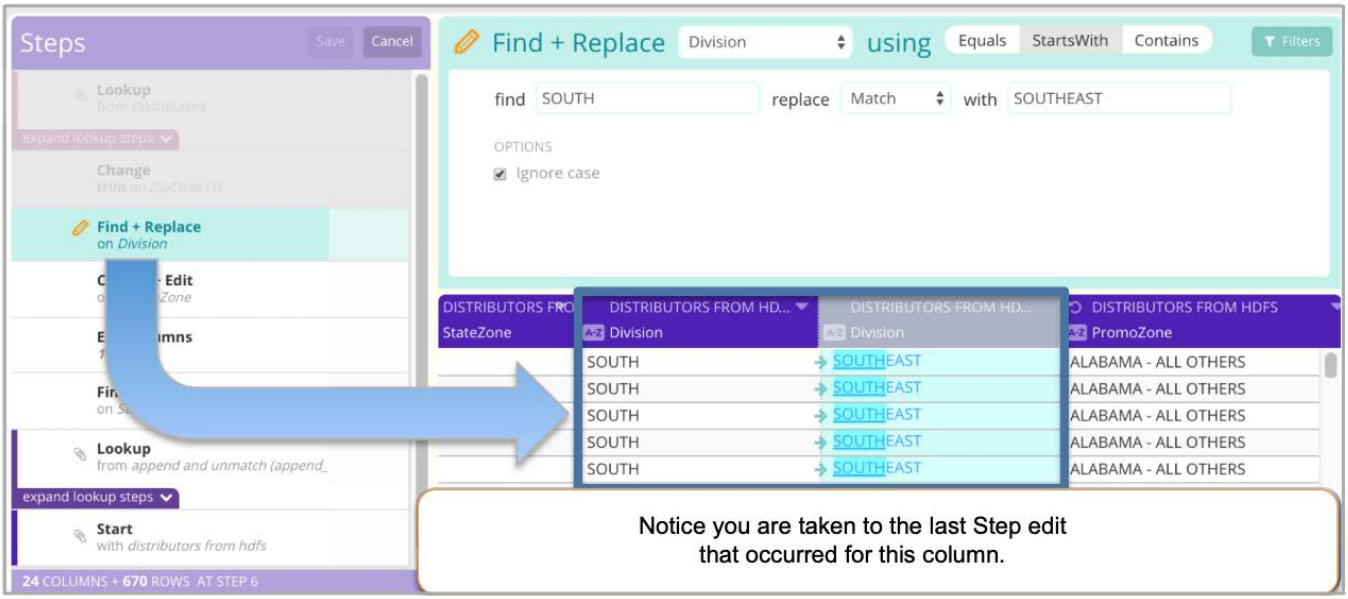
| StepType | Syntax |
|---|---|
| Import | `AnchorTableStep` |
| Append | `AppendStep` |
| Find and Replace | `BulkEditStep` |
| Cluster and Edit | `ClusterEditStep` |
| Duplicate Column | `DuplicateColumnStep` |
| Computed Column | `ExpressionStep` |
| Manage Columns—includes all of the following operations: Hide, Reorder, Delete columns (***see note below*) | `EditColumnsStep` |
| Pivot—includes all of the following shaping operations: Deduplicate, Depivot, Transpose, Group and Pivot (***see note below*) | `PivotStep` |
| Transform—includes all of the following transformations: capital case, lowercase, upper case, unescaped HTML, blanks, custom value, rename, whitespace for both numeric and string values (***see note below*) | `TransformStep` |

**備考**

これらの [StepTypes]がプロジェクト内に複数ある場合、URLは**Steps**ペインに表示されている最後のものを指します。

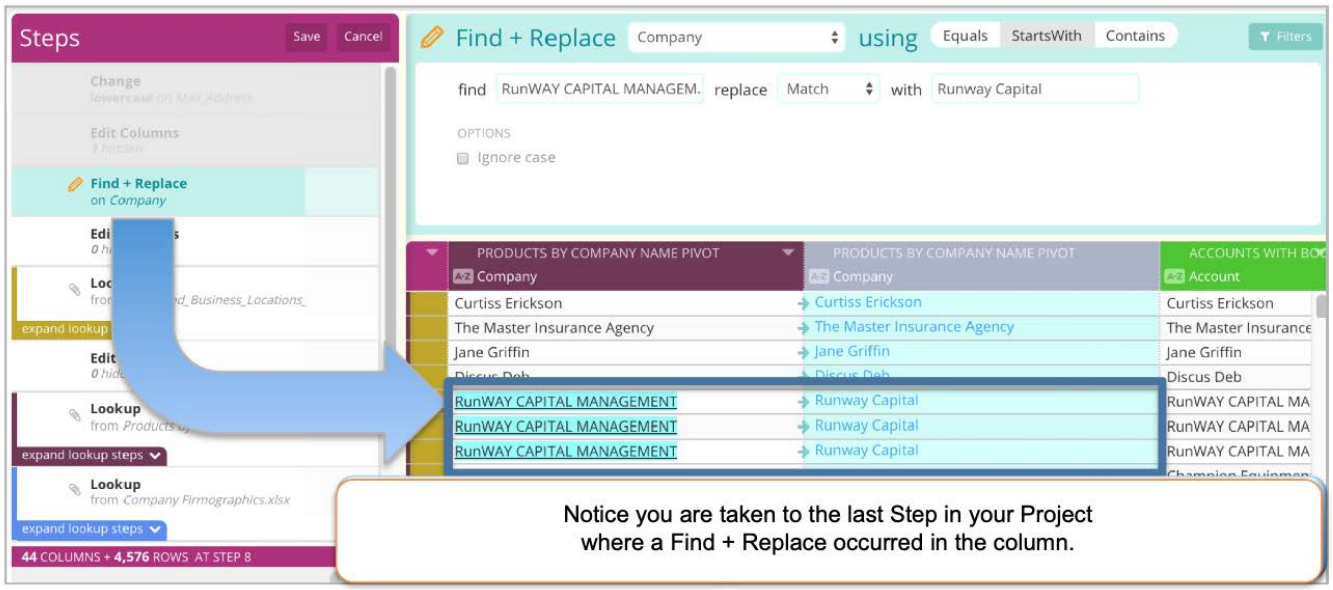
## 例1

次のClicktoPrepリンクでは、**Division**列のデータに影響を与えたり、変更した**Customer**プロジェクトの最後のステップに移動します。




## 例2

次のClicktoPrepリンクは、前回の「検索と置換」[**BulkEditStep**]ステップが **Company** 列で行われた**Customer**プロジェクトに移動します。




このタイプのリンクを作成するには、プロジェクトステップへのリンクの設定を参照してください。

# 要件および考慮事項

Data Prep ClicktoPrepリンクを作成する前に、次の要件と考慮事項に留意してください。

- 使用する視覚化ツールがURLに対応している必要があります。
- Data Prepプロジェクトに誘導するURLを開くには、必要なData Prepリソースレベルの権限が必要です。

# プロジェクトのFiltergramへのリンクの設定

このセクションでは、ClicktoPrepプロジェクトでのFiltergramリンクの形式と設定パラメーターについて説明します。（Tableauをお使いの場合は、Tableauでプロジェクトの Filtergramへのリンクを作成するを参照してください。）

形式は次のとおりです。

`https://[server]/#/view/[projectname]?filtercolumn=[column]&filtervalue=[value]`

**ヒント**

疑問符の前の文字がベースURLになります。**一般**ページの**ClicktoPrep URL**フィールドからベースとなるURLをコピーします。

プロジェクトのFiltergramへのリンクの要件は次のとおりです：

- URLの#以降はすべて大文字と小文字が区別されます。
- Data PrepプロジェクトのFiltergramへのリンクを作成するには、視覚化ツールが動的URLに対応している必要があります。
- プロジェクト名または列名に1つ以上のスペースが含まれていて、それを**一般**ページからコピーしていない場合は、URLの各スペースを%20（スペースのHTMLエンコーディング値）に置き換えて解決してください。例： `https://<server>/#/edit/Web%20Campaigns%20demo/Phone%20Number`

**備考**

視覚化ツールには自動的にスペースをエンコードするオプションが実装されている場合があります。エンコードが必要なその他の特殊文字については、HTML URLのエンコーディングリファレンスを参照してください。

# プロジェクトステップへのリンクの設定

このセクションでは、ClicktoPrepのプロジェクトステップへのリンクの形式と設定パラメーターについて説明します。（Tableauをお使いの場合は、TableauでData Prepプロジェクトのステップへのリンクを作成するを参照してください。）

列を編集または変更する最後のステップでは、形式は次のとおりです。

`https://[server]/#/edit/[projectname]/[columnname]`

列に対する [検索と置換] ステップなど、列を編集または変更する最後の[StepType]では、形式は次のとおりです。

`https://[server]/#/edit/[projectname]/[columnname]?filter=[StepType]`

**ヒント**

疑問符の前の文字がベースURLになります。**一般**ページの**ClicktoPrep URL**フィールドからベースとなるURLをコピーします。

次の[StepTypes]に対応しています：

- AnchorTableStep
- AppendStep
- BulkEditStep
- ClusterEditStep
- DuplicateColumnStep
- EditColumnsStep
- ExpressionStep
- PivotStep
- TransformStep

それぞれの[StepTypes]の説明については、プロジェクトステップへのClicktoPrepリンクを参照してください。

プロジェクトステップへのリンクの要件は次のとおりです：

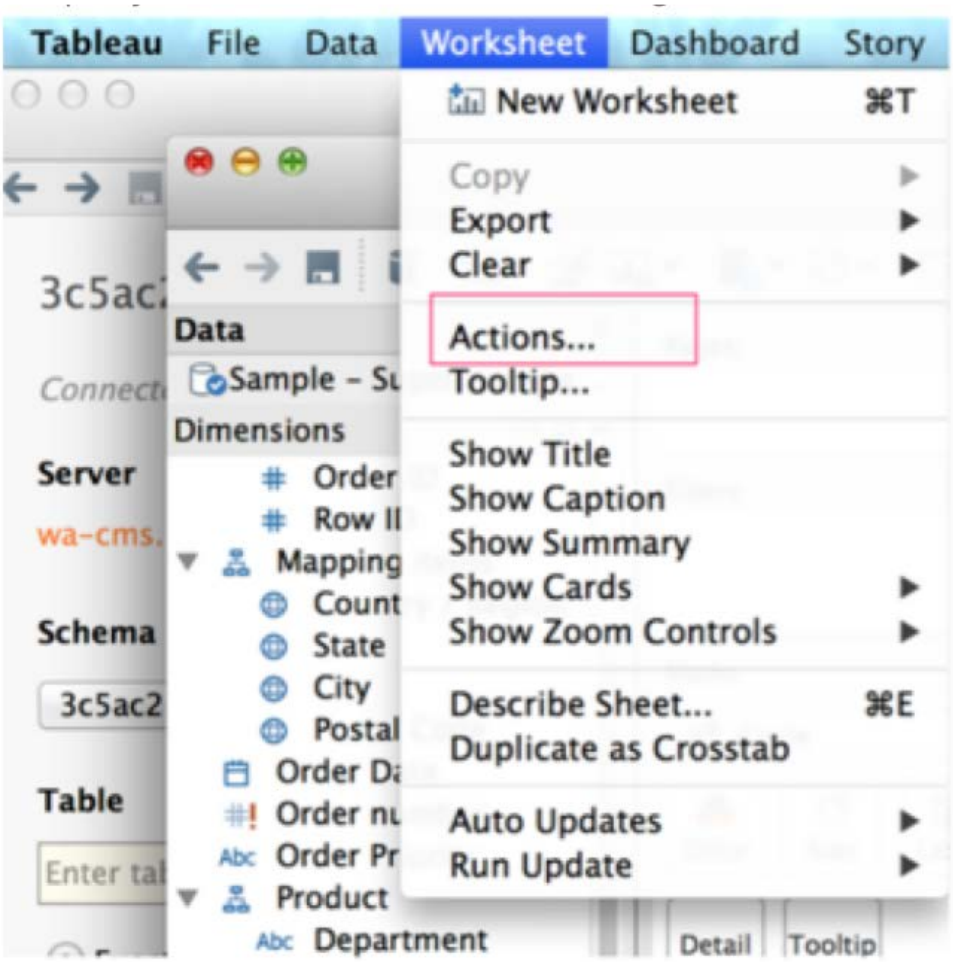
- URL の # 以降はすべて大文字と小文字が区別されます。
- プロジェクト名または列名に1つ以上のスペースが含まれていて、それを上記のように**一般**ページからコピーしていない場合は、URLの各スペースを%20（スペースのHTMLエンコーディング値）に置き換えて解決する必要があります。例: `https://<server>/#/edit/Web%20Campaigns%20demo/Phone%20Number`

> **備考**
>
> 視覚化ツールには自動的にスペースをエンコードするオプションが実装されている場合があります。エンコードが必要なその他の特殊文字については、HTML URLのエンコーディングリファレンスを参照してください。

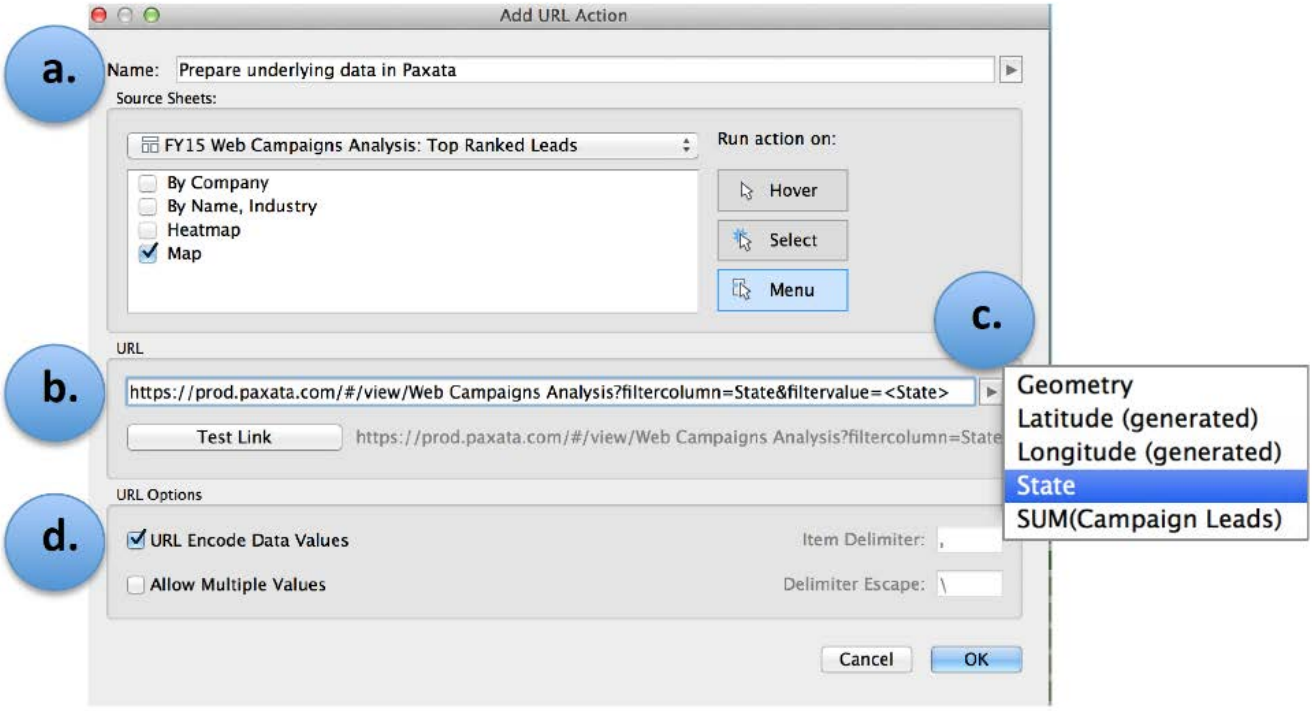
# TableauでプロジェクトのFiltergramへのリンクを作成する

動的ハイパーリンクまたはURLに対応しているツールでは、選択ツールに基づいて適用されるフィルターを使用して、Data Prepプロジェクトに直接リンクすることができます。このセクションでは、Tableauでこれを行う具体的な方法について説明します。

1. Tableauを開き、**[ワークシート] > [アクション...]**に移動します。




2. **[URLアクションの追加]**を選択して、以下の設定情報を入力します。




a. Tableauに表示されるリンクの名前。

b. URLはこの形式です。 `https://[server]/#/view/[projectname]?filtercolumn=[column]&filtervalue=[value]` 疑問符の前の文字列がベースURLになります。**一般**ページの**ClicktoPrep URL**フィールドからベースとなるURLをコピーします。URLの#以降はすべて大文字と小文字が区別されます。

c. URLからデータを動的に受け取るTableauの列をクリックして選択します。

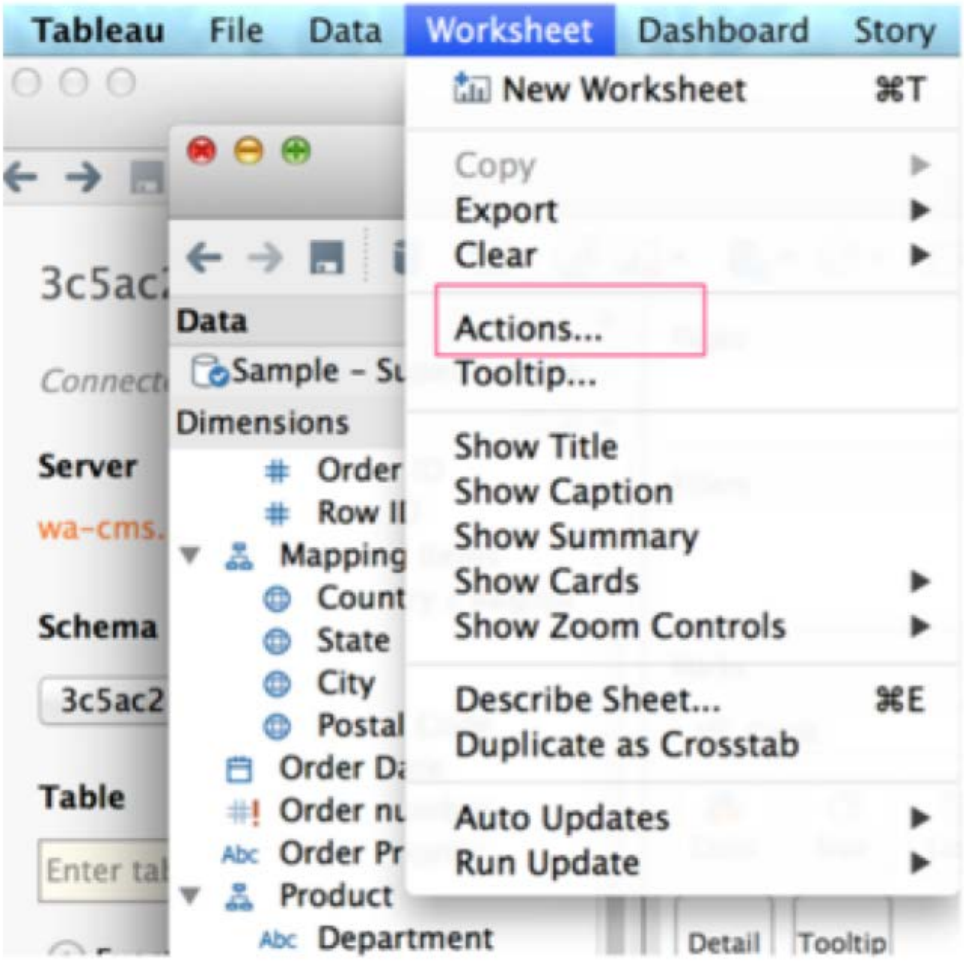
d. **[URLエンコードデータ値]**がチェックされていることを確認します。

3. **[リンクのテスト]**をクリックして、リンクが機能していることを確認します。
4. **[OK]** をクリックして保存します。

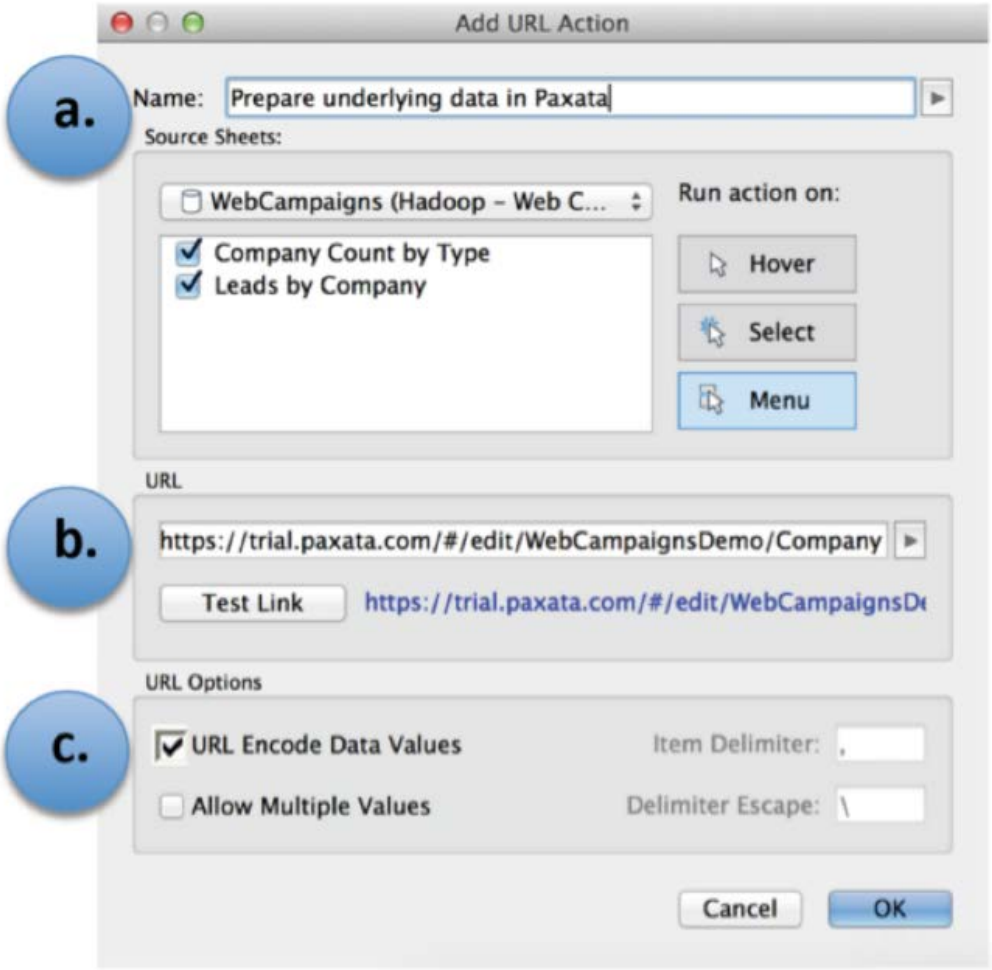
# TableauでData Prepプロジェクトのステップへのリンクを作成する

動的ハイパーリンクまたはURLに対応しているツールでは、選択ツールに基づいて適用されるフィルターを使用して、Data Prepプロジェクトに直接リンクすることができます。このセクションでは、Tableauでこれを行う具体的な方法について説明します。

1. Tableauを開き、**[ワークシート] > [アクション...]**に移動します。




2. **[URLアクションの追加]**を選択して、以下の設定情報を入力します。




a. Tableauに表示されるリンクの名前を入力します。

b. 以下のいずれかのフォーマットでURLを指定します：

- 列を編集または変更する最後のステップ： https://[server]/#/edit/[projectname]/[columnname]
- 列を編集または変更する最後の[StepType]： https://[server]/#/edit/[projectname]/[columnname]?filter=[StepType]

疑問符の前の文字がベースURLになります。**一般**ページの**ClicktoPrep URL**フィールドからベースとなるURLをコピーします。URLの#以降はすべて大文字と小文字が区別されます。

それぞれの[StepTypes]の説明については、プロジェクトステップへのClicktoPrepリンクを参照してください。

c. URLからデータを動的に受け取るTableauの列をクリックして選択します。

d. **[URLエンコードデータ値]**がチェックされていることを確認します。

3. **[リンクのテスト]**をクリックして、リンクが機能していることを確認します。

4. **[OK]** をクリックして保存します。

# Data Prepのインフラストラクチャとセキュリティ

このトピックでは、DataRobot DataPrepインフラストラクチャのセキュリティとアプリケーションのセキュリティについて詳しく説明します。

## インフラストラクチャのセキュリティ

### アクセス制御

内部インターフェイスまたは外部インターフェイスに関係なく、すべての入力ポートはセキュリティグループによって保護されます。セキュリティグループは、Data Prepのシステム設定ツールによって自動的に構成されます。顧客/パブリックに直面するポートはTCP 80/443です。

Data Prepは、本番インフラストラクチャへのSSHアクセスにジャンプホストを利用し、多要素認証を使用してすべての本番管理者をアクセス制御します。

本番アカウントは厳密なIAMロールを使用し、ビジネス上の必要性が確認された主要な従業員のみが管理アクセスを取得します。

## 評価

DataRobotでは、お客様が要求したセキュリティスキャンエージェントを本番SaaS環境にインストールすることはできません。Data Prepは、オンデマンドのクラウドコンピューティングプラットフォームを活用して、環境に対して脆弱性スキャンを実行します。Data Prepの侵入テストは、資格のあるサードパーティの評価者によって実行され、その結果は優先度に基づいて開発ワークフローに統合されます。リクエストに応じて、DataRobotはSaaSオファリングの脆弱性スキャンをスケジュールし、クラウドコンピューティングプラットフォームサービスとリクエストを調整できます。

## データ保護

Data Prepは、TLSとHTTPSを利用して、転送中にデータを暗号化します。Data Prepは、データが保存されているときに暗号化された形式でデータを保存し、第三者による不正アクセスを防ぎます。

# アプリケーションのセキュリティ

## パスワードポリシー

Native Data Prepアカウント（LDAPまたはSAMLを使用しないアカウントとして定義）は、次のパスワード要件に準拠しています。パスワードには、少なくとも1つの数字、1つの小文字、1つの大文字、1つの特殊文字、および少なくとも8文字が含まれている必要があります。例：(!@#$%^&*+=)

Data Prepは、アカウントロックアウトポリシーを適用することも、ネイティブアカウントのアカウントロックアウトポリシー管理機能を有することもありません。

SAML認証の場合、顧客のSAMLアイデンティティープロバイダーで設定されたアカウントポリシーとパスワード要件が適用されます。

管理者またはユーザーは、本番サービスアカウントをログインに使用できません。このアカウントは、Data Prepアプリケーションの起動と実行にのみ厳密に使用されます。アカウントには、Data Prep内の顧客データまたは権限へのアクセス権はありません。

## セキュリティアップデート

オペレーティングシステムのセキュリティパッチは、セキュリティの脅威の評価/レビュー後に本番SaaS環境に適用されます。アプリケーションやサービスの整合性が損なわれないように、セキュリティ更新プログラムを適用する前に慎重なテストが実行されます。SaaSオファリングのアプリケーションセキュリティアップデートは、修正が利用可能になるとすぐに適用されます。

# 交互作用モード

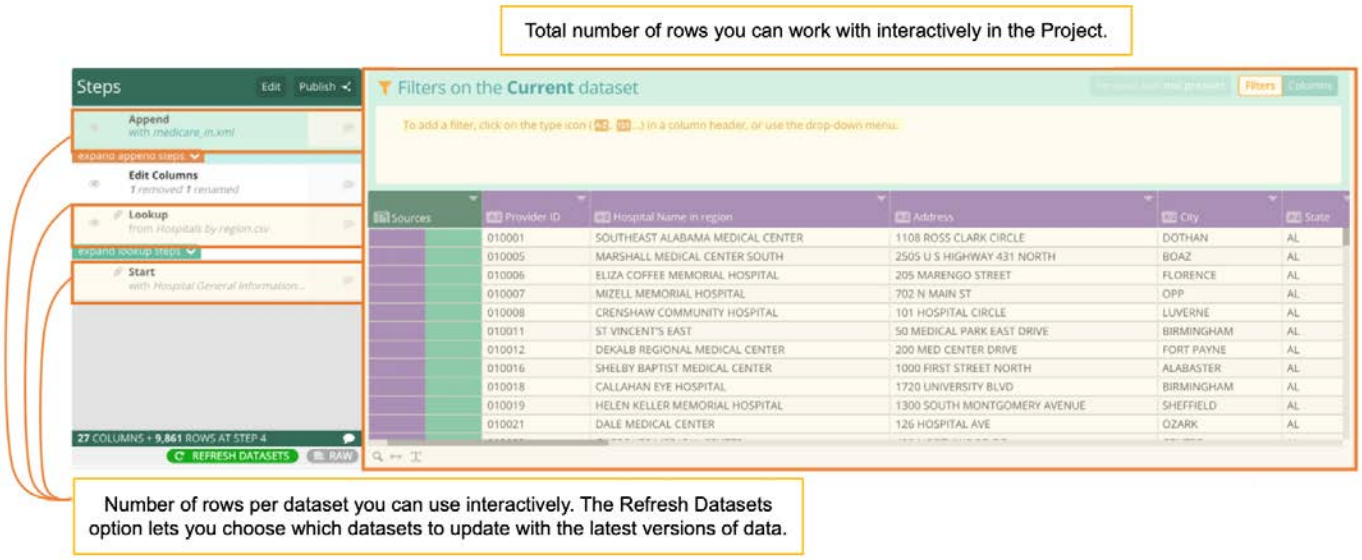
ソースデータセットが大きくなると、データ準備ツールでそのデータセットを効率的にインポートして操作する能力に影響を及ぼす可能性があります。このようなデータ量増大の問題に対処するため、Data Prepにはインタラクティブ モード機能が用意されています。これは、データの一部のみ（読み込み量はプロジェクトのニーズに合わせてユーザーが決定できる）をより迅速に扱えるようにするモードです。このモードを使用すると、Data Prepプロジェクトでデータの一部分だけを使って（そのすべてのデータをプロジェクトに取り込まずに）、効率的かつインタラクティブに準備作業を行うことができます。

**備考**

Data Prepの管理者は、アプリケーションでこの特徴量を有効にする必要があります。

インタラクティブ モードの機能の主要なメリットには、次のようなものがあります：

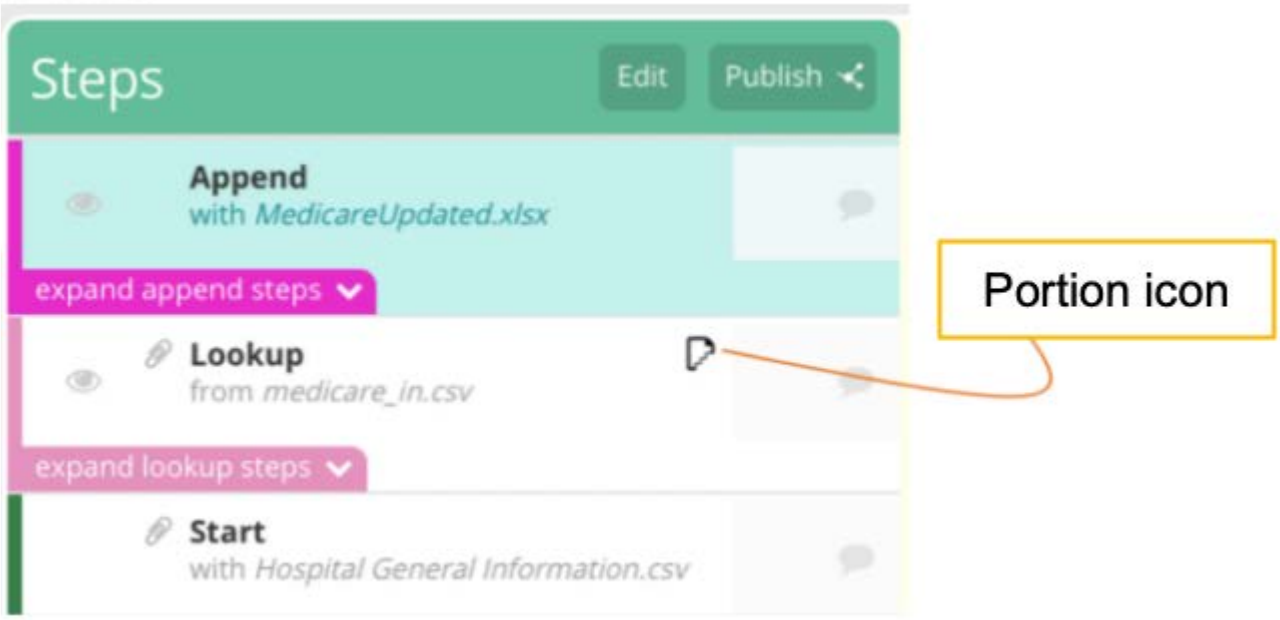
- データセット全体がライブラリに読み込まれるまで待たずに、Data Prepプロジェクトでデータセットの操作を開始できます。または、データセットの読み込み量を定義し、ユーザーが定義した読み込み量に達すると、そのデータがプロジェクトの準備に使用できるようにすることもできます。残りのデータセットは引き続きライブラリに読み込まれます。
- プロジェクトでのデータの準備が完了した後、自動プロジェクトフロー機能を通じて、変換をネイティブデータセットの**すべてのデータ**に簡単に適用できます。
- インタラクティブ モードで扱うデータセットの読み込み量は、いつでも再設定できます。たとえば、読み込み量の制限をデータセットあたり 50,000行に設定して作業した後、各データセットの読み込み量をもっと増やした方がよいと気づく場合があります。読み込み量の変更は、Data Prep管理者がワンステップで行うことができます。読み込み量の制限を変更すると、その変更がプロジェクトで動的に認識され、新しいデータを取り込むためにデータセットをリフレッシュするオプションがユーザーに提供されます。
- ユーザーはプロジェクト内のデータセットの定義された部分を処理するだけで済むため、Data Prepプロジェクトでの対話的なエクスペリエンスが最適化されます。
- 大規模なプロジェクトをより柔軟に扱うことができます。インタラクティブ モードのData Prepプロジェクトでは、プロジェクト内で準備できる行の最大数を定義する行が制限されています。この制限はData Prep管理者によって設定されており、使用可能なシステムリソースに基づいてユーザーが最適な対話的エクスペリエンスを得られるように管理者が確認できるため、便利です。




デフォルトでは、インタラクティブ モードはData Prepプロジェクトに対して無効になっており、Data Prep管理者に連絡して有効にしてもらう必要があります。インタラクティブ モードを有効にする前に、以下の点を検討してください：

- 既存のプロジェクトの場合、これらのプロジェクトのデータセットのプロファイリング機能を使用します。データセットのプロファイルを作成すると、データの詳細なインサイトが得られ、データセットに最適な読み込み量を決定する参考になります。
- 定義した読み込み量を超える行数のデータセットが含まれている既存のプロジェクトでは、動的な更新によって行が削除されることはありません。その代わりに、各データセットに読み込み量として定義されている行数を適用できるように、該当するプロジェクトを開いたときにデータセットのリフレッシュ機能を使用するオプションが提供されます。データセットのリフレッシュを選択しない限り、読み込み量は適用されません。

インタラクティブ モードを有効にすると、現在データセットの一部を操作していることを示すために、次のアイコンが表示されます。



アイコンにカーソルを合わせると、読み込み量として定義されている行数が表示されるため、すべてのデータセットに適用されている行数がすぐにわかります。

**ヒント**

データセットに含まれる_行の総数_を確認するには、各データセットの行の総数が表示されるライブラリ ページに移動します。ライブラリ ページには、各データセットの行の総数に加えて、インタラクティブ モード機能に固有の次のような情報も表示されます。

- データセットの読み込み状態と、そのインタラクティブ部分がプロジェクトでいつ使用可能になったか。
- インタラクティブ モードでプロジェクトから公開されたAnswerSet。

詳細については、Data Prepライブラリを参照してください。