AWS Glue DataBrew ワークショップ

2022/04/20

シニアエバンジェリスト 亀田

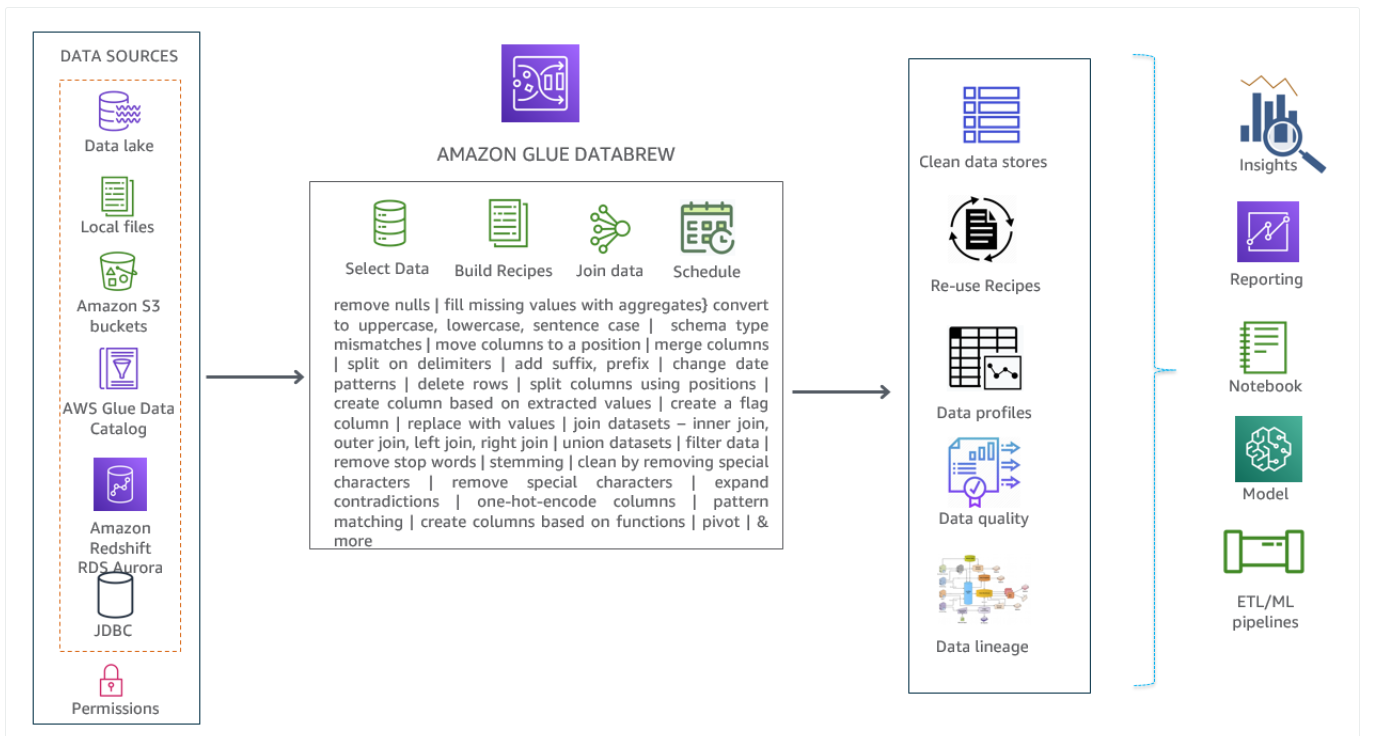
はじめに：AWS Glue DataBrewはデータのクリーニングと正規化を迅速にするビジュアルデータ準備ツールです。このツールでは多くの一般的なデータ変換作業がノーコードで行えるように様々な処理があらかじめ組み込まれています。別途Glue Studioをという近しいことを実現させるツールもありますが、こちらはローコードで様々な実装を行うとともに、スクリプトを自前で作成し作業を行いたい方向けという違いがありますが、お互いを組み合わせて使うことも可能です。

参考：GlueStudio Workshop

<https://github.com/harunobukameda/AWS-Glue-Studio>

本ワークショップは以下の英語版をベースに手順を一部簡略化し日本語化したものです。動作などに問題がある場合は、英語版を参考にしてください。

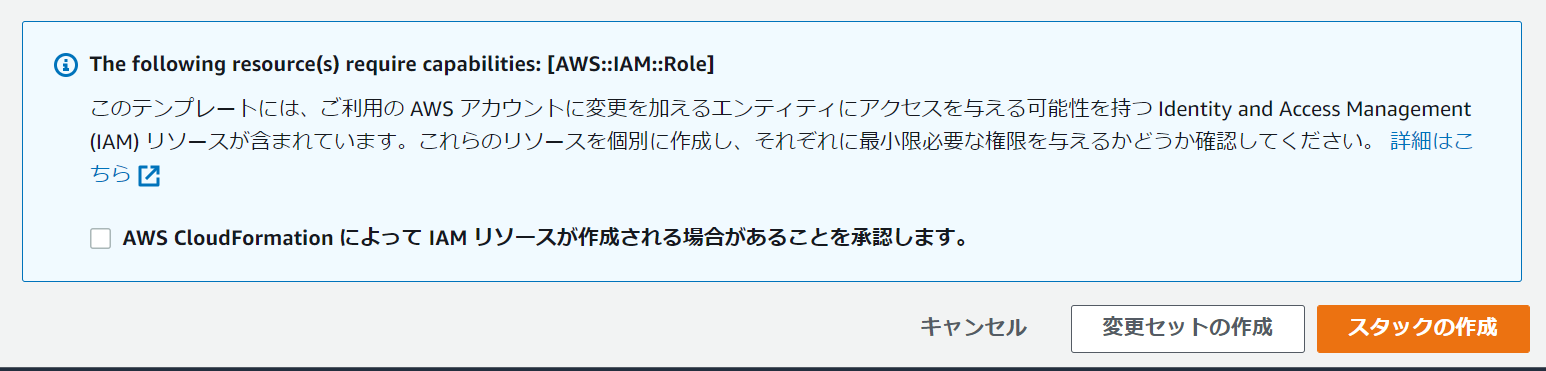
<https://catalog.us-east-1.prod.workshops.aws/workshops/6532bf37-3ad2-4844-bd26-d775a31ce1fa/>



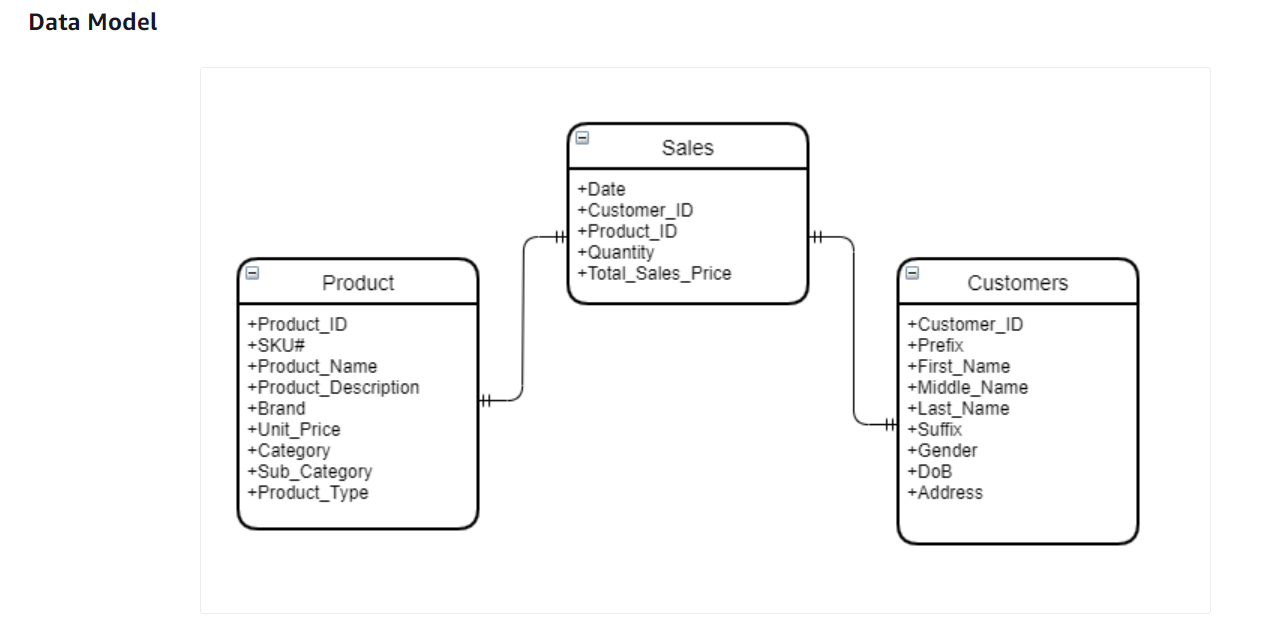
1. 全ての作業はノースバージニアリージョンで行います。以下のURLをクリックしてください

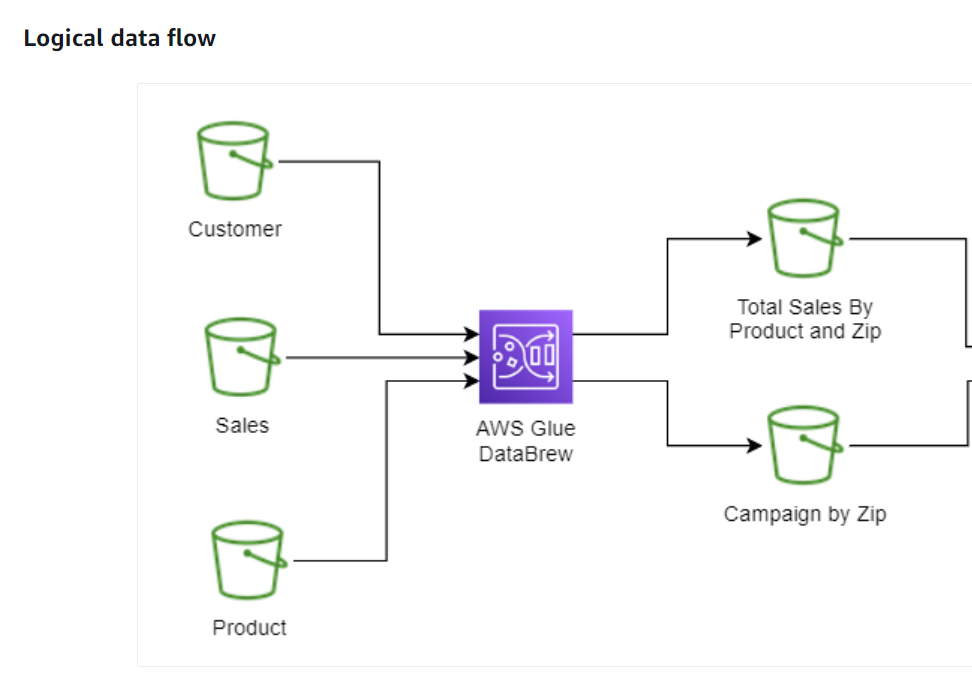
<https://us-east-1.console.aws.amazon.com/cloudformation/home?region=us-east-1#/stacks/create/review?templateURL=https://aws-data-analytics-workshops.s3.amazonaws.com/glue-databrew-immersionday-v2/databrew_ID-prod.yaml&stackName=glue-databrew-immersionday>

1. CloudFormationテンプレートが起動しますので、以下にチェックをつけ[スタックの作成]をおします

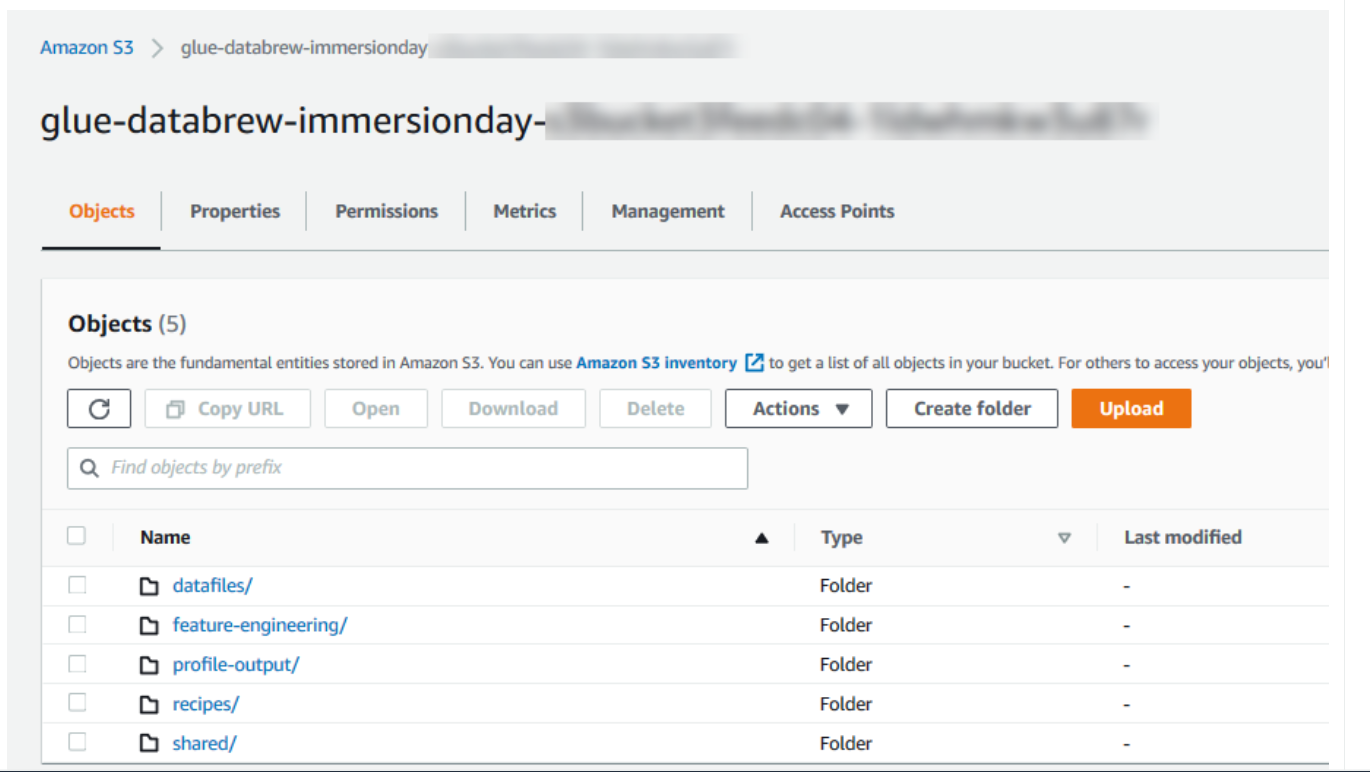


以下のデータがS3へ展開されます

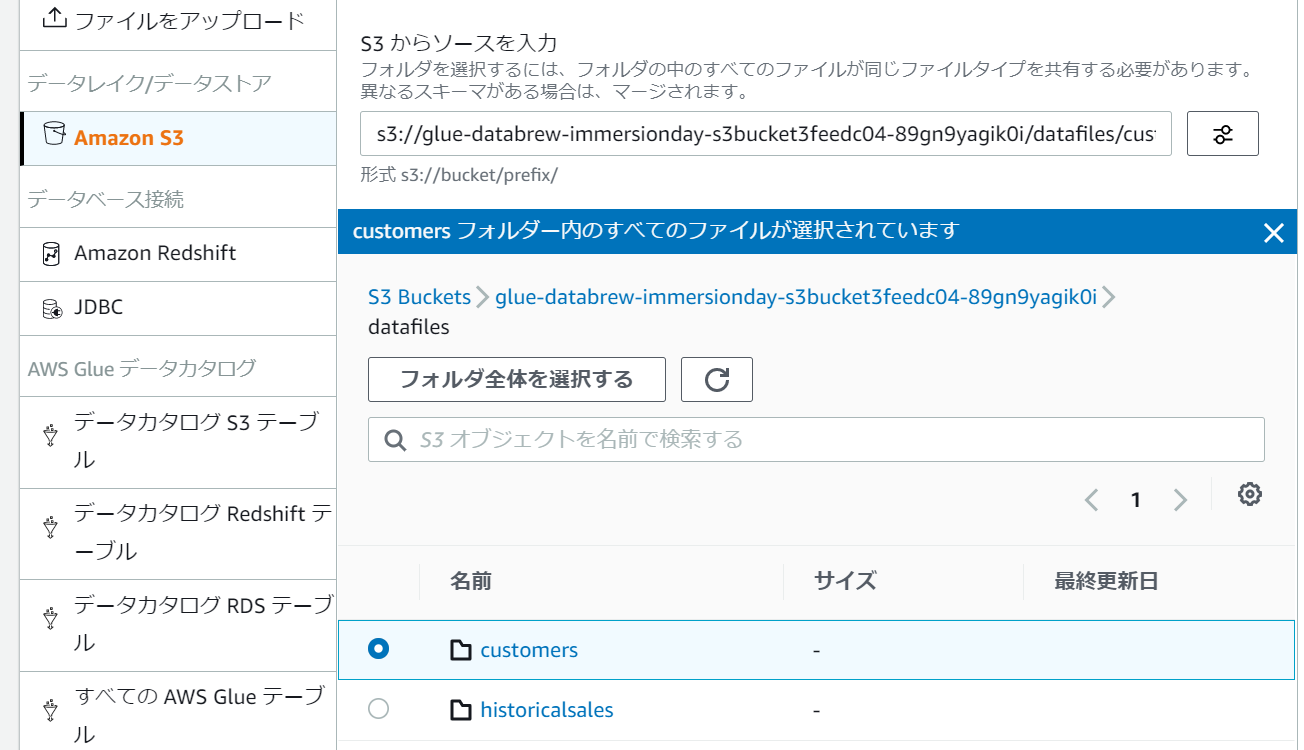




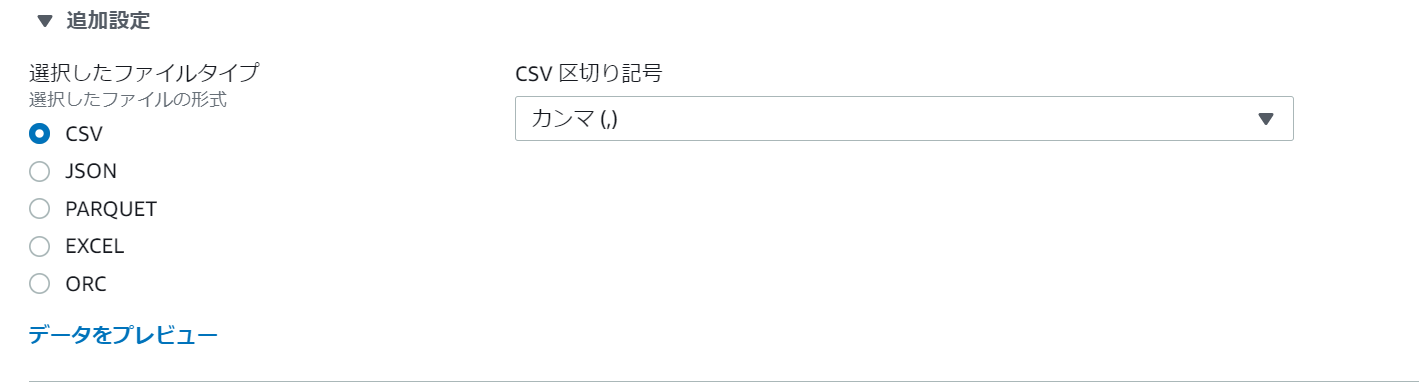
1. 実行が完了すると以下の通りS3バケットが生成され、フォルダが複数できています。生成されたS3バケットの名前をコピーしておきます（” glue-databrew-immersionday”がバケット名についています）



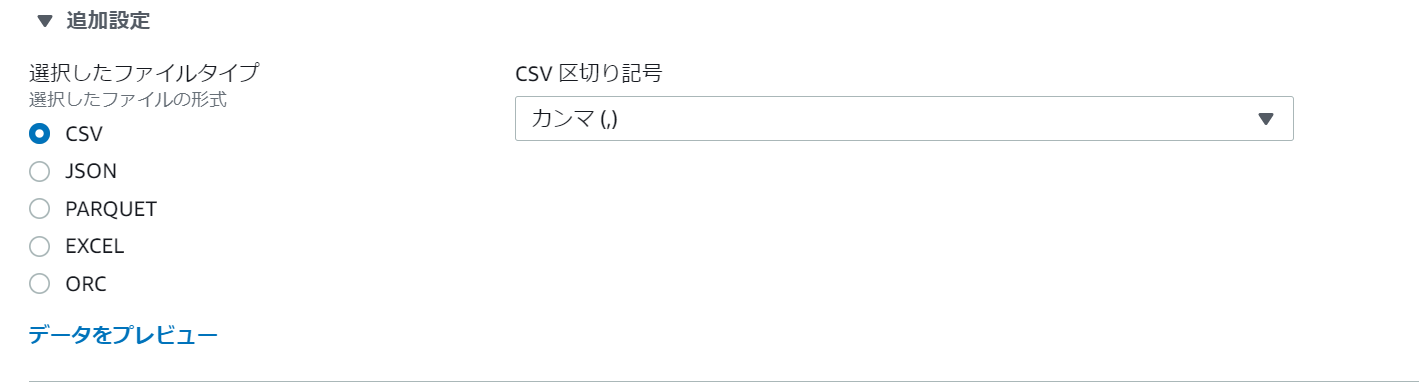
1. Glue DataBrewのマネージメントコンソールへ移動します
2. 左ペインから[データセット]をクリックします
3. [新しいデータセットの接続]をおします
4. [Customers]と名前をつけます
5. 先程作成したS3バケットを指定し[datafiles][customers]フォルダを選び、[customers.csv]を選びます



1. カンマ区切りCSVを選びます



1. 画面一番下の[データセットを作成]をおします



1. 以下の通りデータセットが作成されました。

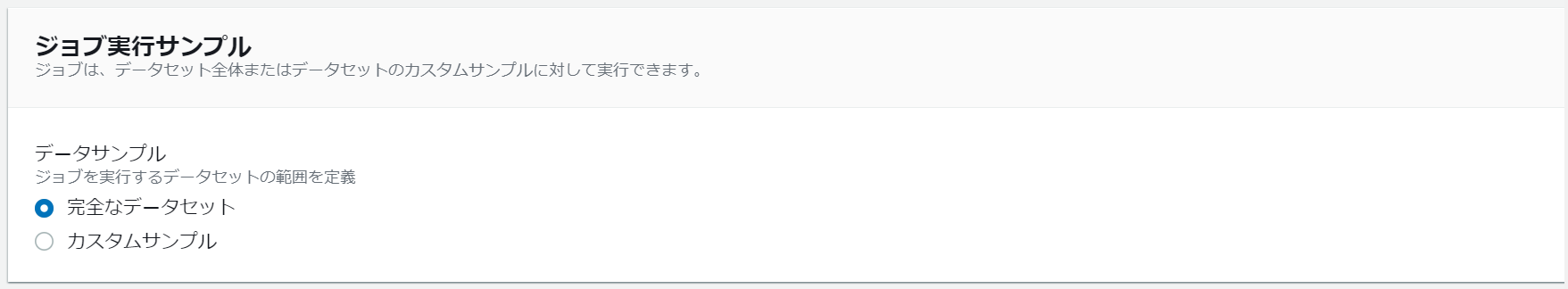


プロジェクトの作成：プロジェクトとはデータセットに対してデータ操作を行う一連の作業をさします。先程作成したcustomersデータセットに対してデータ操作を行うプロジェクトを作成していきます

1. 先程作成したデータセットをクリックし、詳細画面に移動します



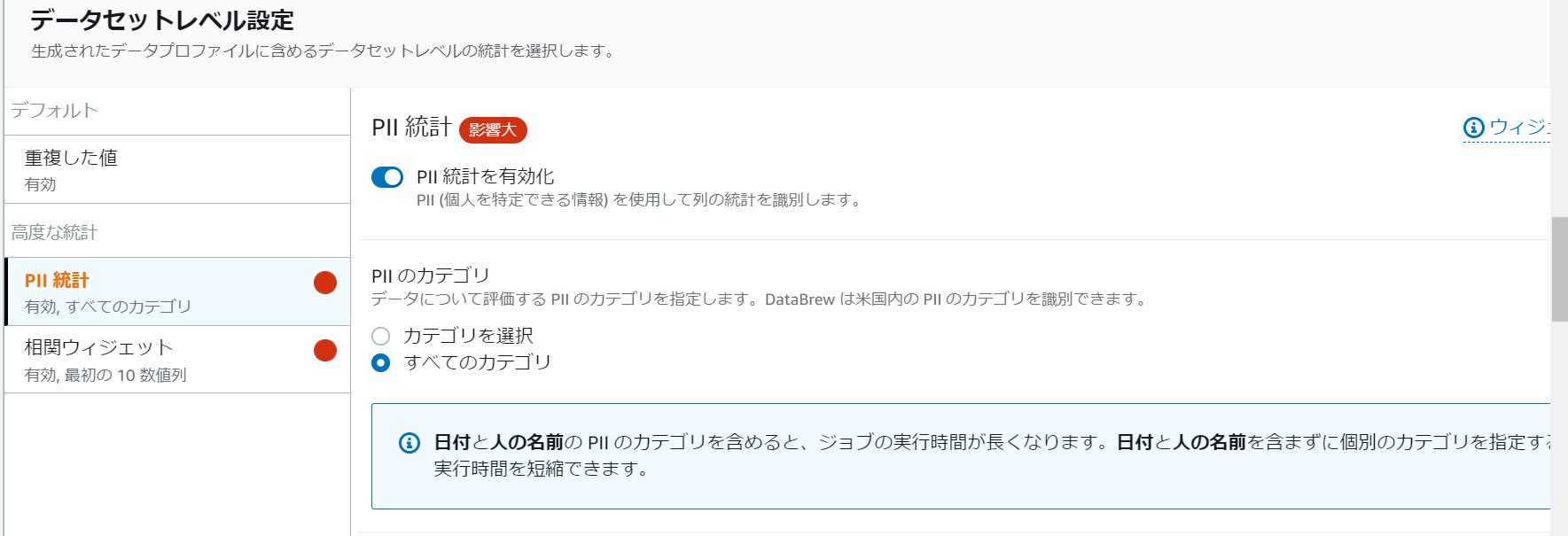
1. [データプロファイルを実行]を押します
2. [customers profile job]と既に名前が入っていますので、そのままで進めます
3. [完全なデータセット]を選びます



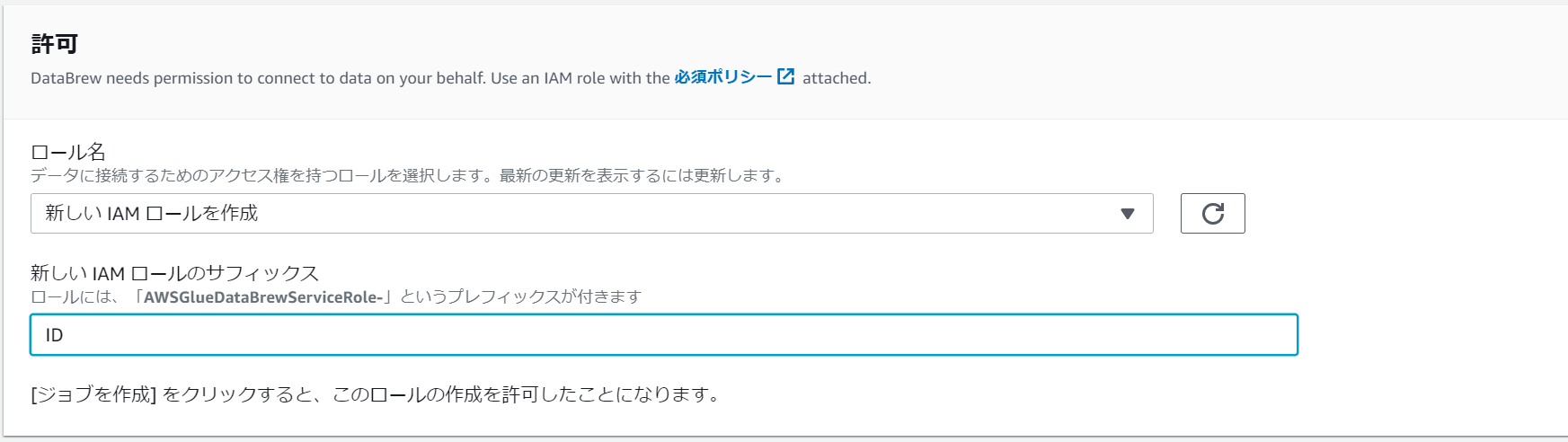
1. 出力先に、先程作成したS3バケットの[profile-output]フォルダを選びます



1. [PII統計]をクリックし、有効化し、すべてのカテゴリを選びます



1. 一番下のロール設定パートで[新しいIAMロ―ルを作成]を選び、IDと入力します

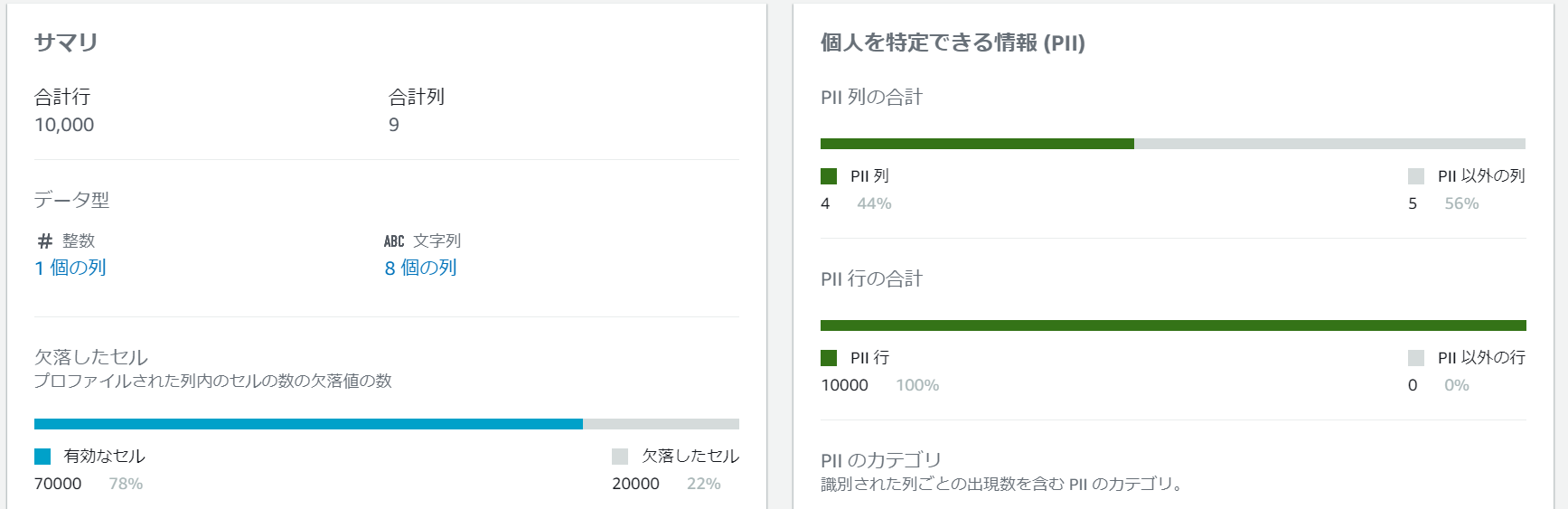


1. [ジョブを作成し実行する]をおします
2. 5分程度待つと、画面が遷移しデータ分析が行われていることがわかります。先程PIIオプションをオンにしたため、個人情報がデータにどのように含まれているか？も併せて精査しているため少し時間がかかります。現時点で対応しているデータ種別一覧はこちらになります。

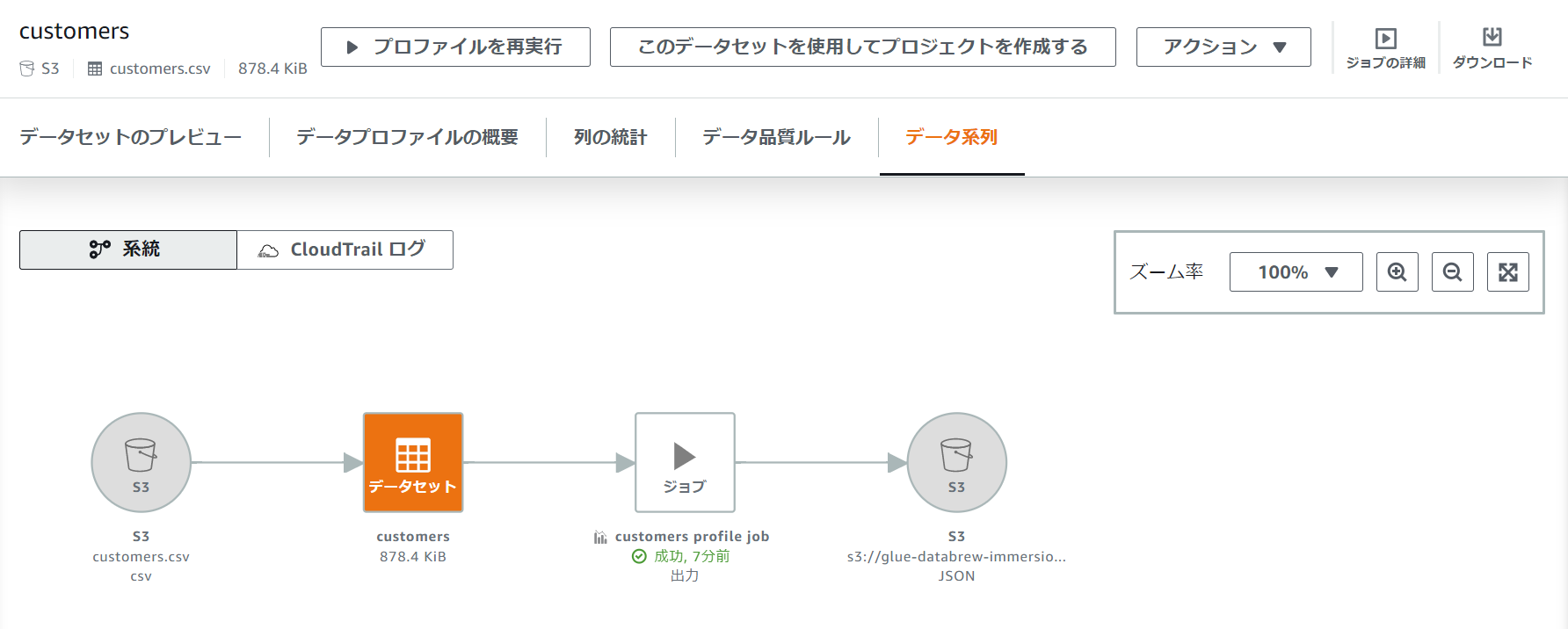
https://docs.aws.amazon.com/databrew/latest/dg/profile.configuration.html



1. 分析が完了すると以下の通りデータのサマリが出てきます

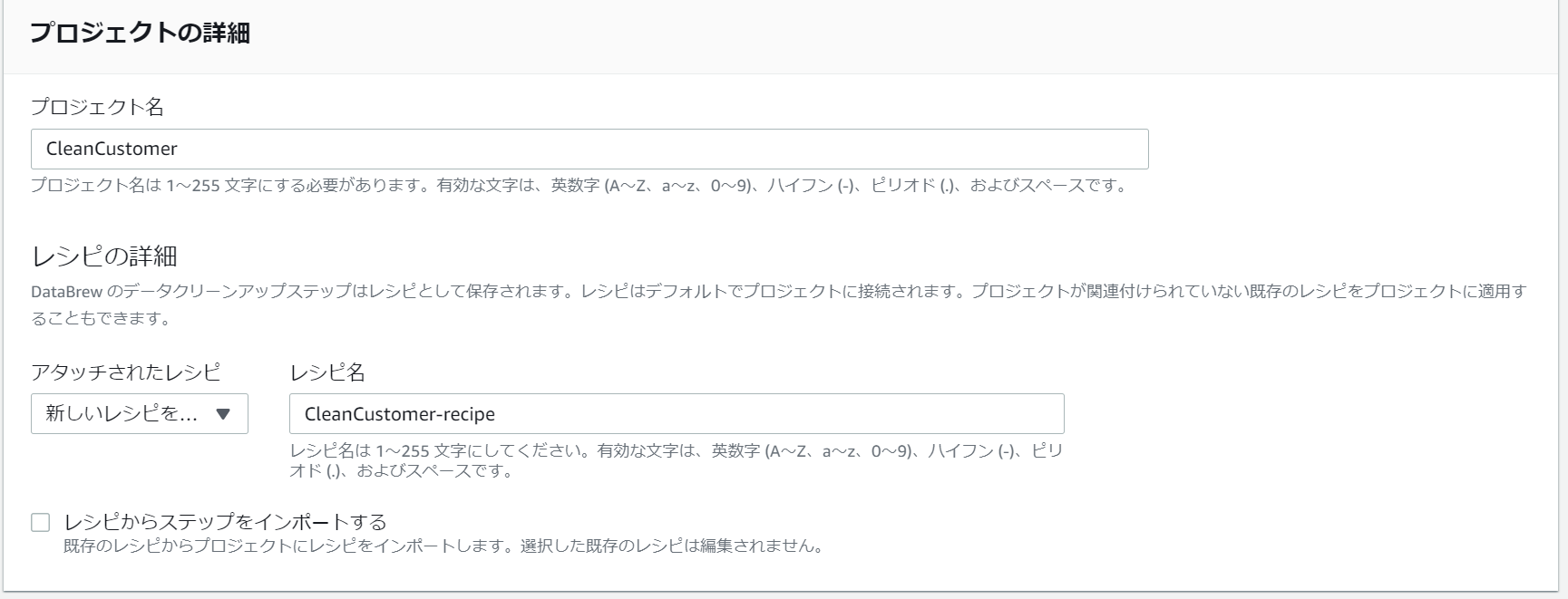


1. [データ系列]タブを見るとどのようにデータが構成されているかが把握できます



ではこれから、Transformジョブを作成します。Transformとはデータの変換作業のことです。ジョブは先程のデータセットに対して行うよう作成します

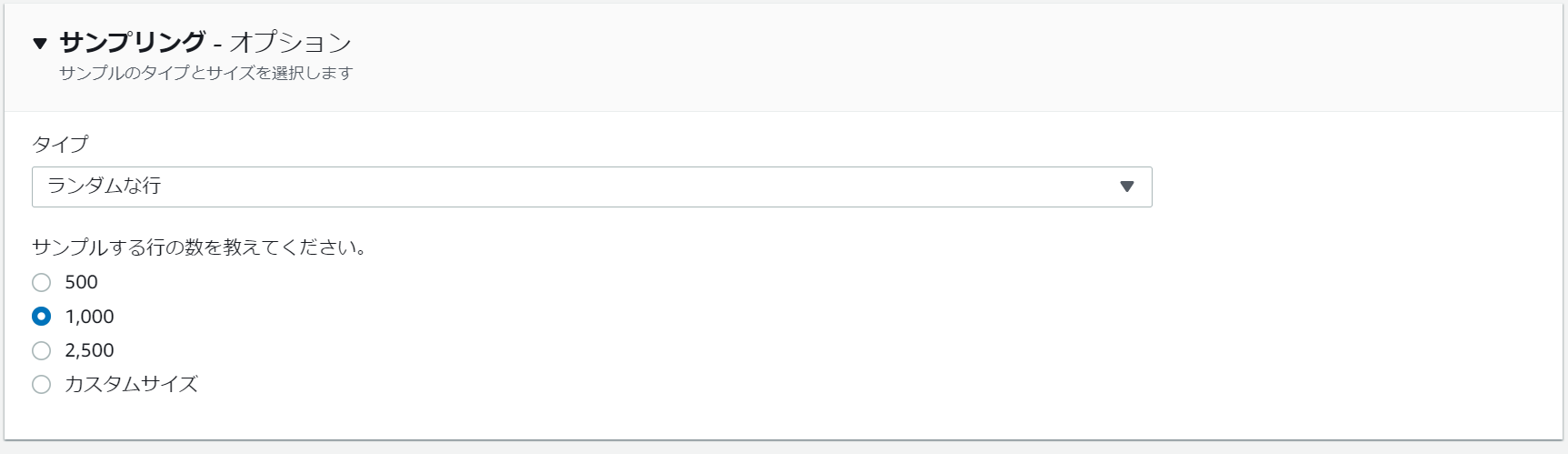
1. 左ペインの[プロジェクト]をクリックし、[プロジェクトを作成]をおします
2. プロジェクト名に[CleanCustomer]と入力します。レシピ名が自動で生成されます。レシピとは、データ変換を行う手順を意味します



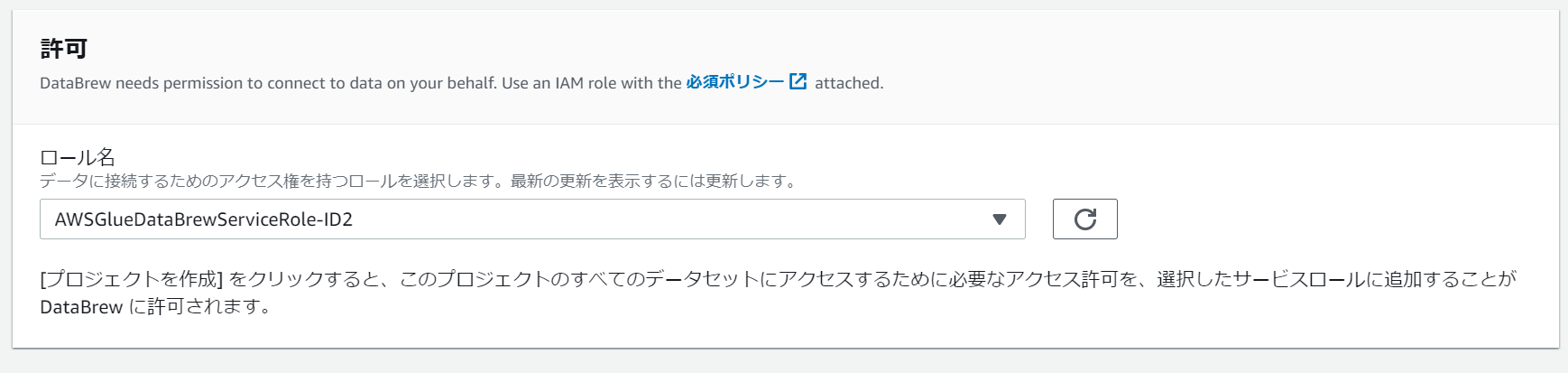
1. マイデータセットでcustomersを選びます



1. [サンプリングセクション]でランダムな行を1000件抽出するように指定します



1. 先程作成したIAMロールを選びます



1. [プロジェクトを作成]をおします。以下の通り初期化が開始されますので少しまちます



1. はじめて作業される方は画面右、レシピにいくつか設定が表示されますが、全て削除してください。

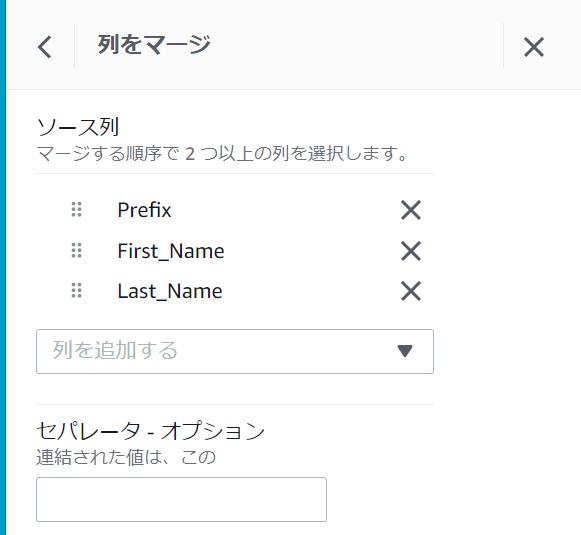


1. 画面上部に様々なデータ操作がノーコードで行えるようなコマンドが備わっています。ここで作成するルールがレシピに追加されていき、最終的に複数の変換ルールを含んだレシピが生成されます。レシピはプロジェクトに紐づき、プロジェクトはデータセットに紐づきます。（当然複数データセットのジョインプロジェクトも作成が可能です）

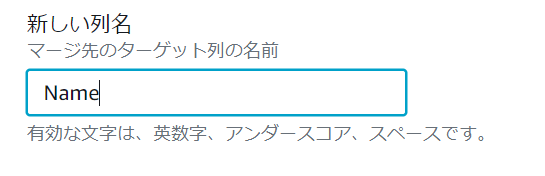


ではまず、[マージ]をおします

1. 以下3つのカラムを選び、セパレーターに半角スペースを入れます。（わかりづらいので注意してください）



1. [新しい列名]に[Name]と入力します



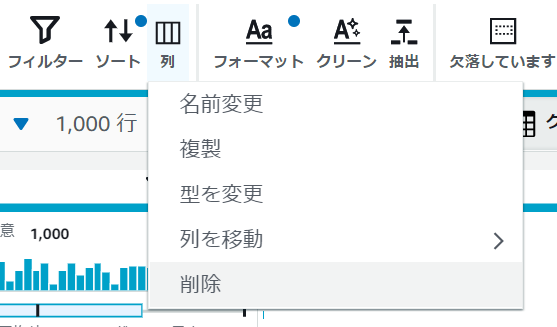
1. [変更のプレビュー]をおしてみてください。3つの列がNameに結合されています。
2. [適用]をおします
3. レシピにステップが追加されました。この作業を繰り返していきます。



1. 今度は、新しくできたNameの[…]を選び、フォーマットから大文字に変更を選びます



1. [適用]をおします
2. 次は、列→削除を選びます



1. 以下2列を選んで、適用をおします



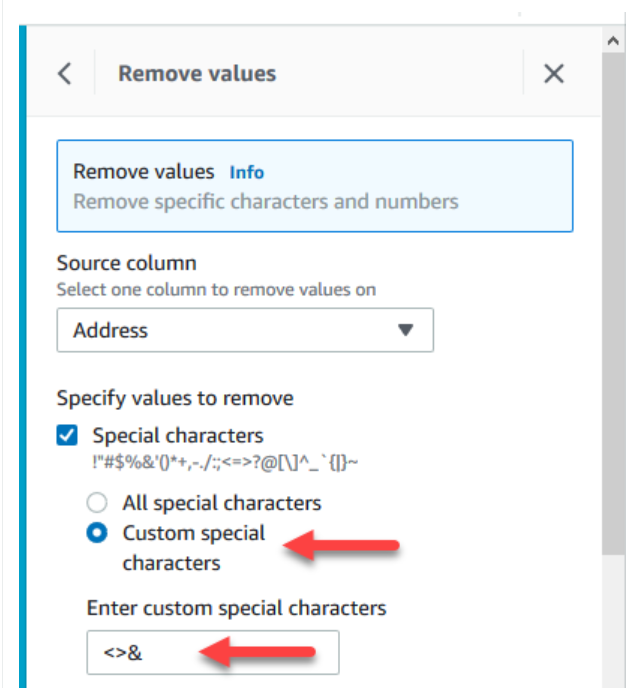
1. 次にDoB（Data of Birth）の[…]から、フォーマット→日付時刻の形式を選び、[mm-dd-yyyy]を選び、[適用]をおします



1. [Address]列からクリーン→特殊文字、を選びます



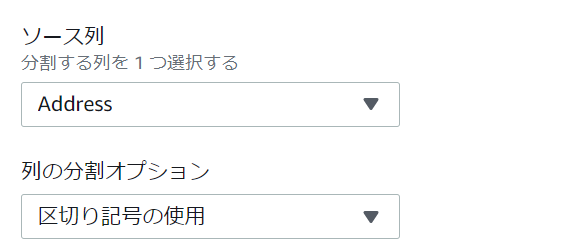
1. 以下の通り設定します



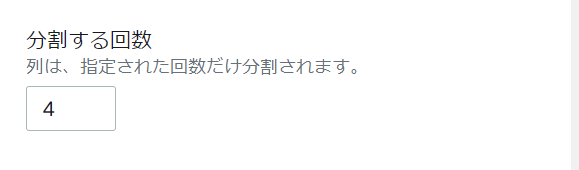
1. [適用]をおします
2. 次に[分割]→[一つの区切り文字]を選びます



1. Addressを選び以下の設定にします。[;]でAddress列を5分割します。区切り文字が4つなので4か所で割ると5列になる、という意味です。



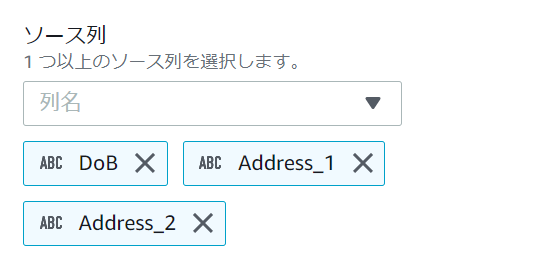




1. [適用]をおします
2. [Address\_3]を[City]、[Address\_4]を[Zip]、[Address\_5]を[Country]にRenameします。それぞれの列の[…]をおして[名前変更]を選び修正してください。
3. 画面右上の詳細から[機密]→[値をマスキング]を選びます



1. 以下の3つを選んで[適用]をおしてください。

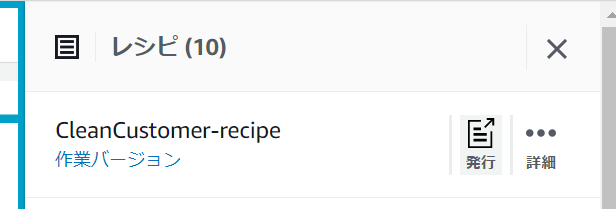


注意：ある程度の一意性が求められる場合は、マスキングではなく、ハッシュ化を選んでください。ハッシュは全レコードにおける一意性を保証していないことに注意してください。オリジナルの値との一意性が担保されているのみです。

1. ここまでの作業で10個のステップがレシピに追加されています。



1. レシピの[発行]をおします。確認ダイアログが出ますので再度[発行]をおします



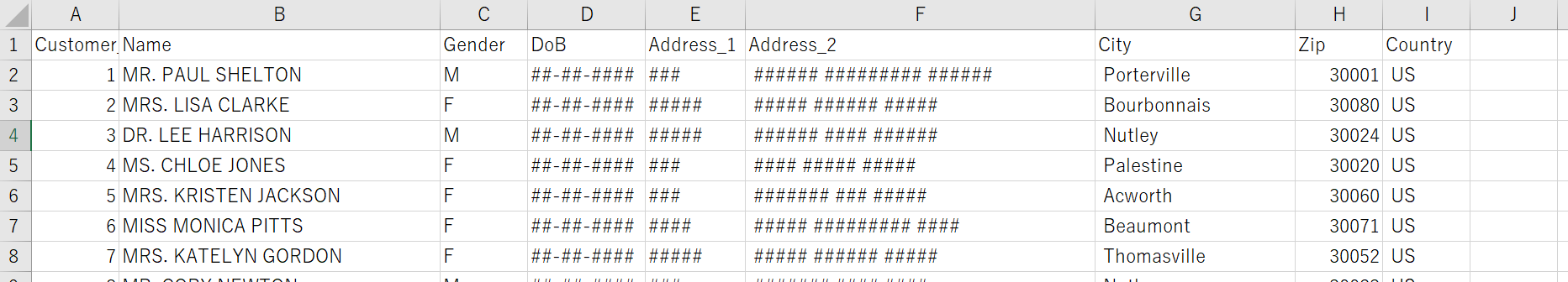
1. 画面右上[ジョブを作成]をおします



1. ジョブ名に[CleanCustomer]を選び、本日作成したS3バケットのsharedフォルダを選びます



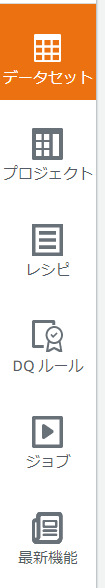
1. 今日作成したIAMロールを選び、[ジョブを作成し実行する]を選びます
2. 以下の通りジョブが進行中です。終わればS3バケットのsharedフォルダに変換されたファイルがcsv形式で生成されます。
3. ジョブが完了したらS3バケットをみてください。csvファイルが生成されています。



1. では次にSalesデータセットを先程と同じ要領で作成します。



1. 左ペインからDQルールをクリックし、[データ品質ルールセットを作成]をおします



1. 名前に[Sales DQ Checks]と入力します
2. 関連付けられたデータセットから[sales]を選びます
3. 画面右側にデータのプレビューが表示されますがQuantityやTotal\_Salesが0となっている行があることがわかります。

以下の通り推奨されるデータ品質チェックルールが提示されています

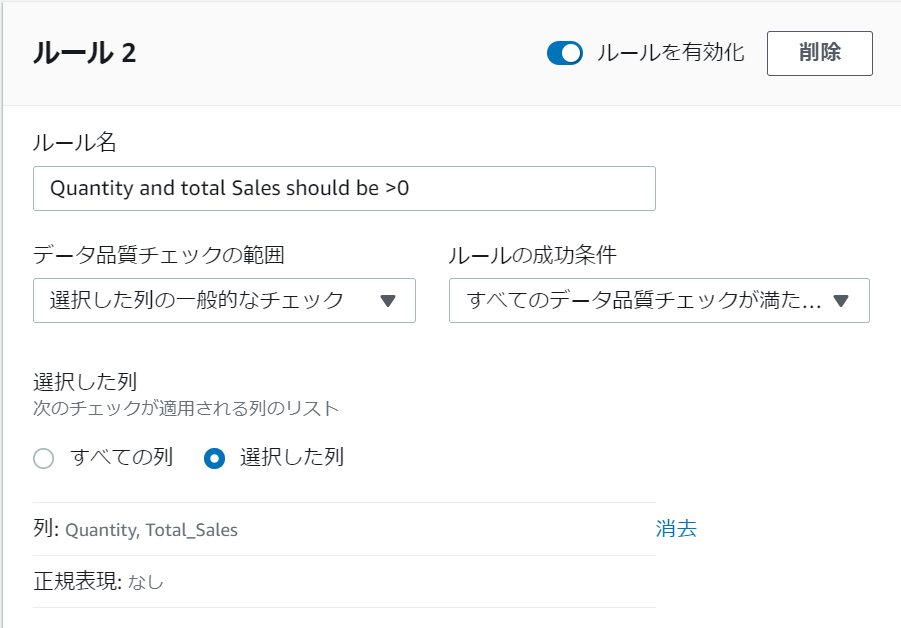


1. 以下の画面の通り、表示されたレコメンデーションを実装します



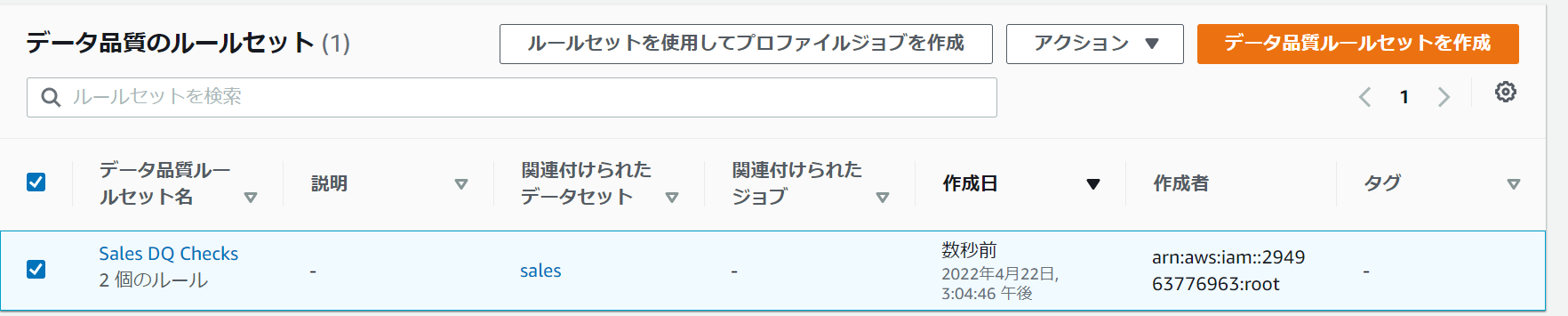
1. [別のルールを追加]をおします

以下の画面を見ながらルールを設定します

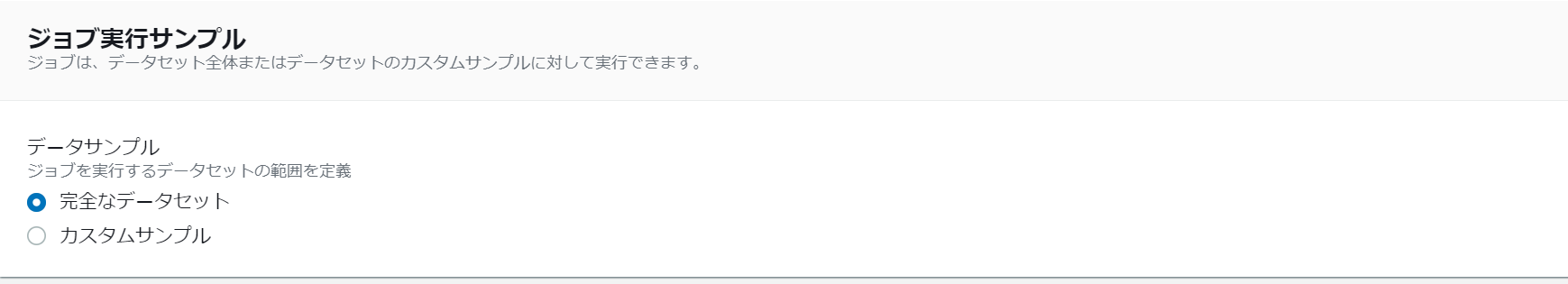




1. [ルールセットを作成]をおします
2. 出来上がったルールセットを選び、[ルールセットを使用してプロファイルジョブを作成]をおします



1. 名前があらかじめついていますのでそのままにして、[ジョブ実行サンプル]で[完全なデータセット]を選びます



1. S3バケットの[/profile-output/]を選びます



1. 先程と同様のIAMロールを選択して、[ジョブを作成し実行する]をおします。以下の通りデータ分析が開始されますので。完了するまで待ちます。



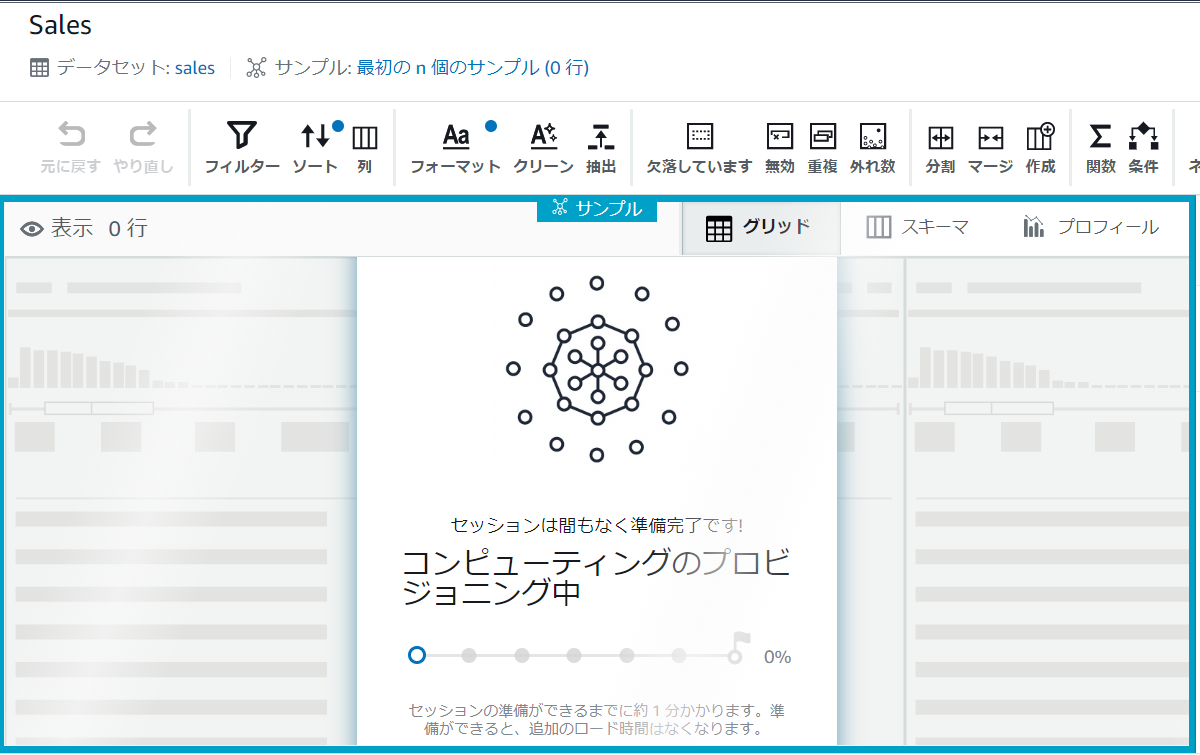
1. 分析が完了すると[データ品質ルールタブ]で品質チェックが失敗（つまり品質が合致していない）していることがわかります。



1. ではこれからSalesデータセットに対して、データ変換の実装と、CustomersデータセットとのJoinを行っていきます。
2. [Sales]という名前のプロジェクトを作成します
3. まずはsalesデータセットを選択します



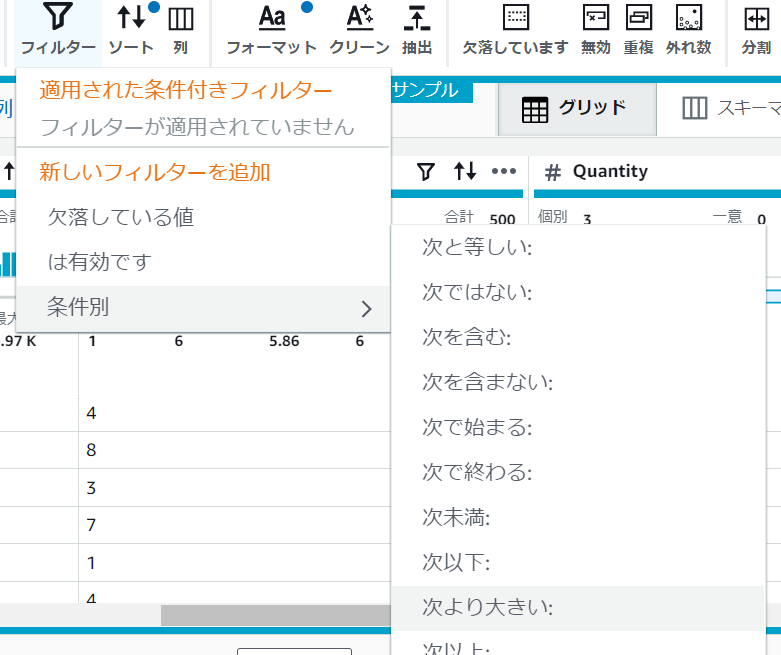
1. IAMロールを選んで[プロジェクトを作成]をおします。以下の通り起動中となるのでしばらく待ちます



1. まず先程のDQ（データ品質）ルールに応じたデータとなるように、フィルターを選びます



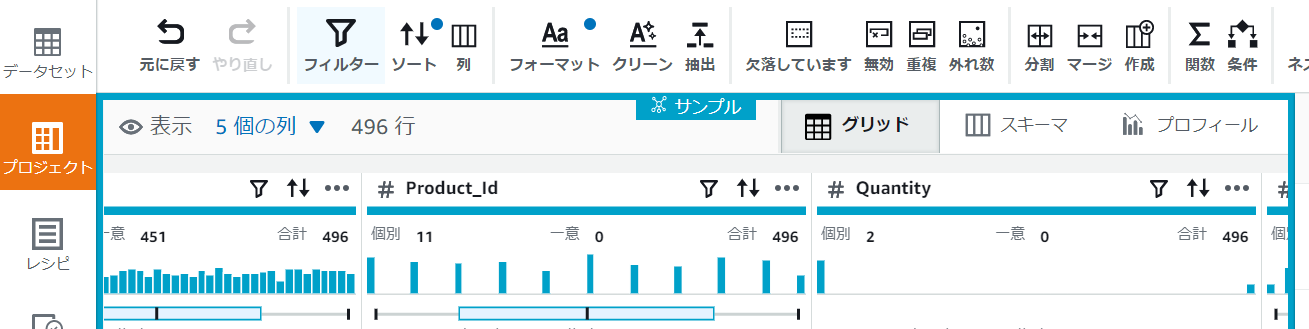
1. [条件別]→[次より大きい]を選びます



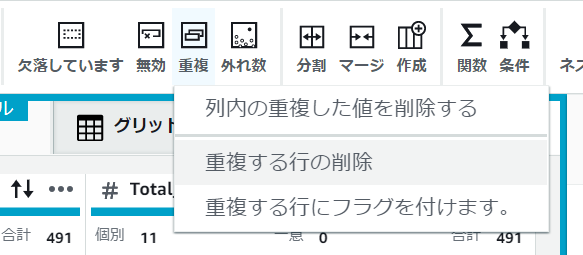
1. 以下のように0より大きいものを設定します。（注意：フィルタールールでは、残すものを指定します）



1. [適用]をおすと、データ総数が500から496に減っています



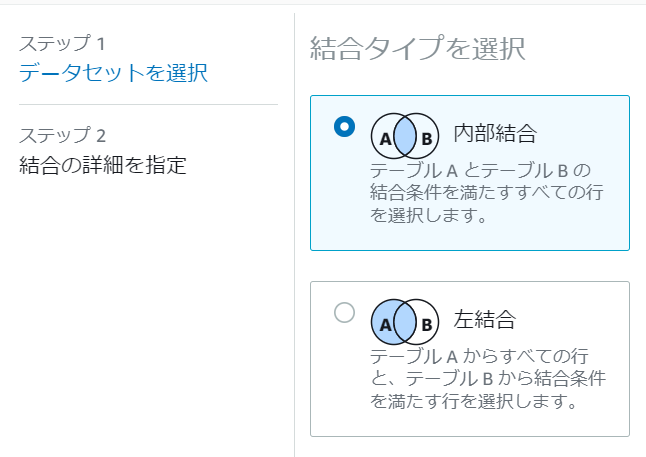
1. 同じように、もう1個[Total\_Sales]に対しても設定をしてください。データ総数は491になります。
2. 今度は[重複する行の削除]を選びます



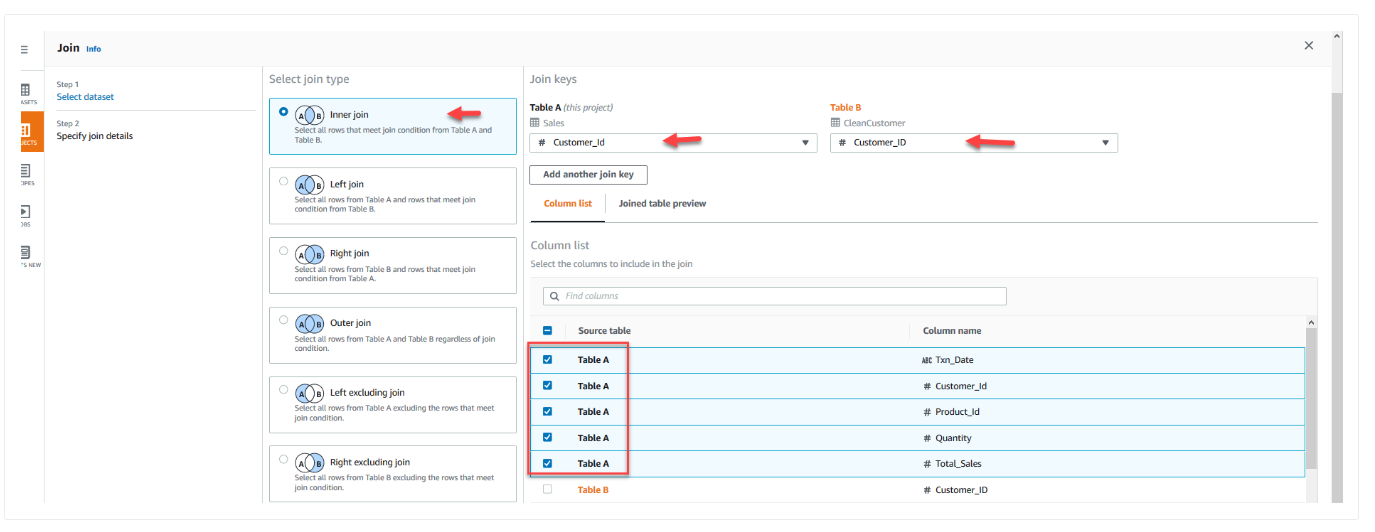
1. [適用]をおすと、データ総数は481になります。
2. 次にcustomerデータセットとのjoinを行います。[結合]をおします（注意：すでにあるcustomerデータセットではなく、前段で作業を行った変換されたcustomerデータを用いるため、新たにデータセットを作成します）

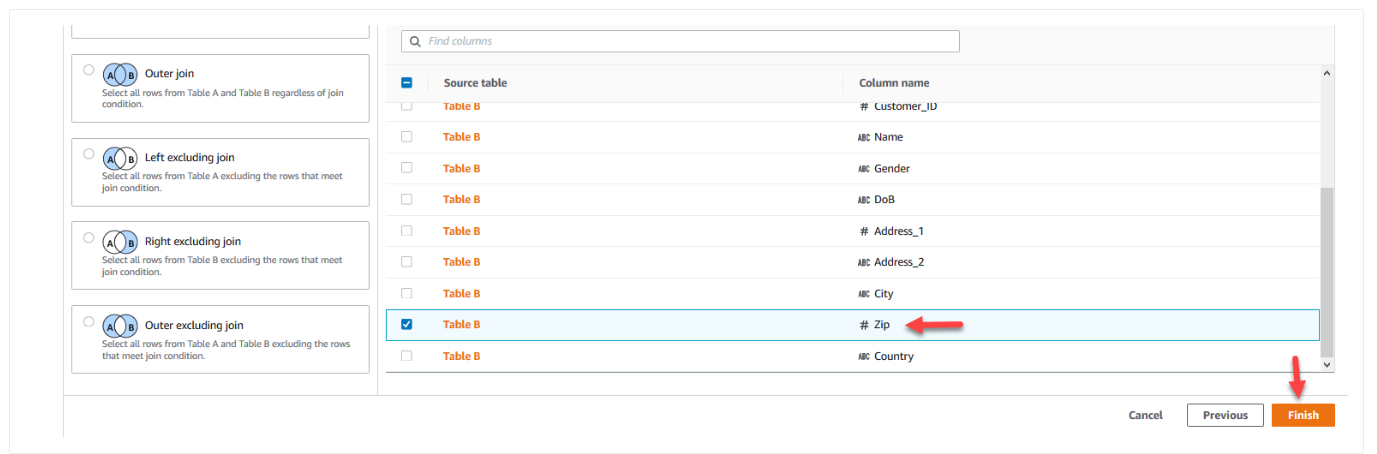


1. [新しいデータセットへの接続]をおします
2. 名前に[CleanCustomer]と入力します
3. S3バケットの[/shared/]に出力されている変換済データセット(csvファイル)を選びます
4. [データセットを作成]をおします
5. [次へ]をおします
6. [内部結合]を選びます



1. 以下の画面の通り設定を行います





1. [終了]をおします
2. レシピを[発行]し、[ジョブを実行]をおします。
3. S3バケットには[/shared/]フォルダを指定します
4. ジョブが完了したらS3バケットにファイルが作成されています。ダウンロードして眺めてみてください。

おつかれさまでした！：

オリジナルシナリオ（英語）にはまだ続きがありますので、興味がある方は挑戦してみてください

<https://catalog.us-east-1.prod.workshops.aws/workshops/6532bf37-3ad2-4844-bd26-d775a31ce1fa/en-US/>

削除は以下をお願いします

・プロジェクト

・レシピ

・ジョブ

・DQルール

・データセット

・IAMロール

・CFnスタック