AWS Glue DataBrew ワークショップ

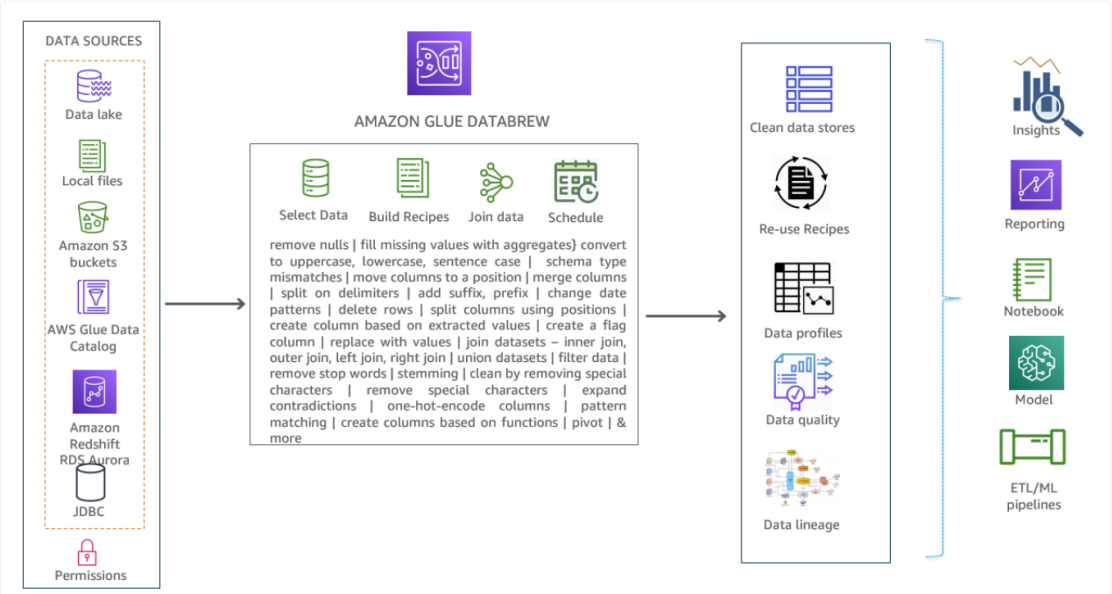
2022/04/20

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

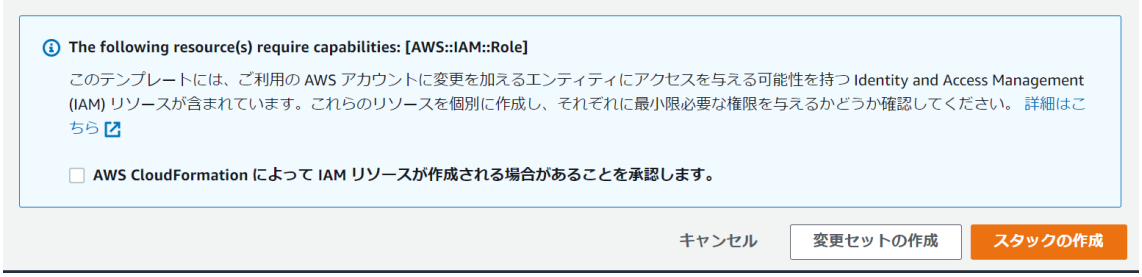
本ワークショップは以下の英語版をベースに手順を一部簡略化し日本語化したものです。動作などに問題がある場合は、英語版を参考にしてください。
https://catalog.us-east-1.prod.workshops.aws/workshops/6532bf37-3ad2-4844-bd26-d775a31ce1fa/



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-

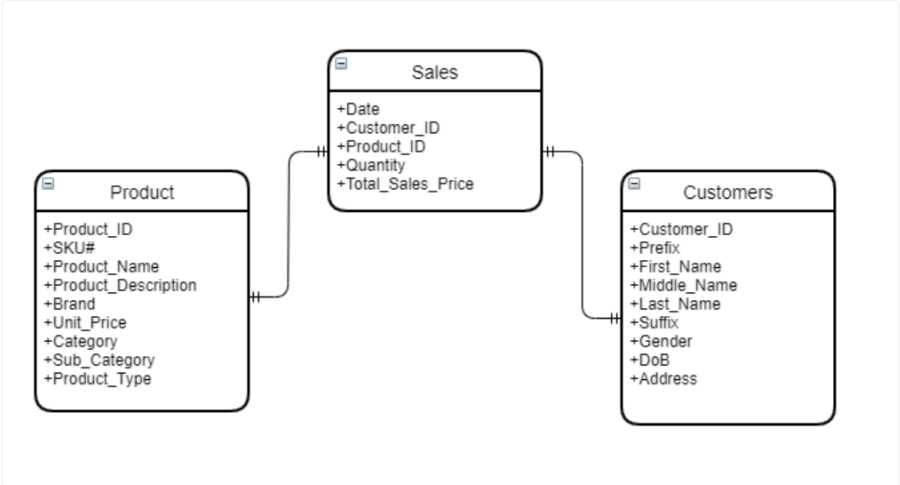
1#/stacks/create/review?templateURL=https://aws-data-analytics-workshops.s3.amazonaws.com/glue-databrew-immersionday-v2/databrew_ID-prod.yaml&stackName=glue-databrew-immersionday

2. 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

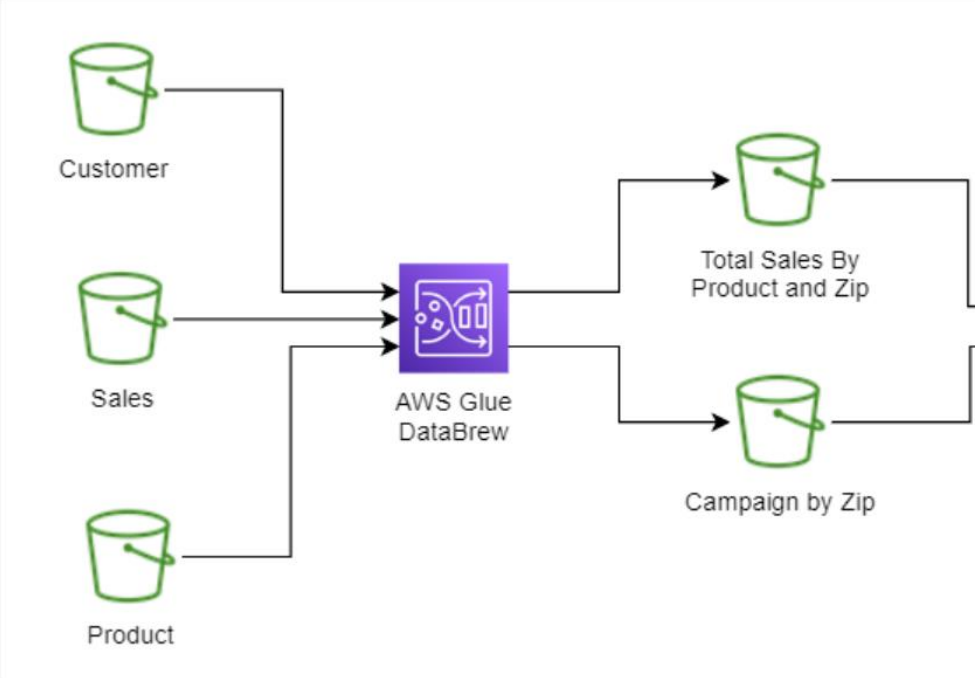


以下のデータが S3 へ展開されます

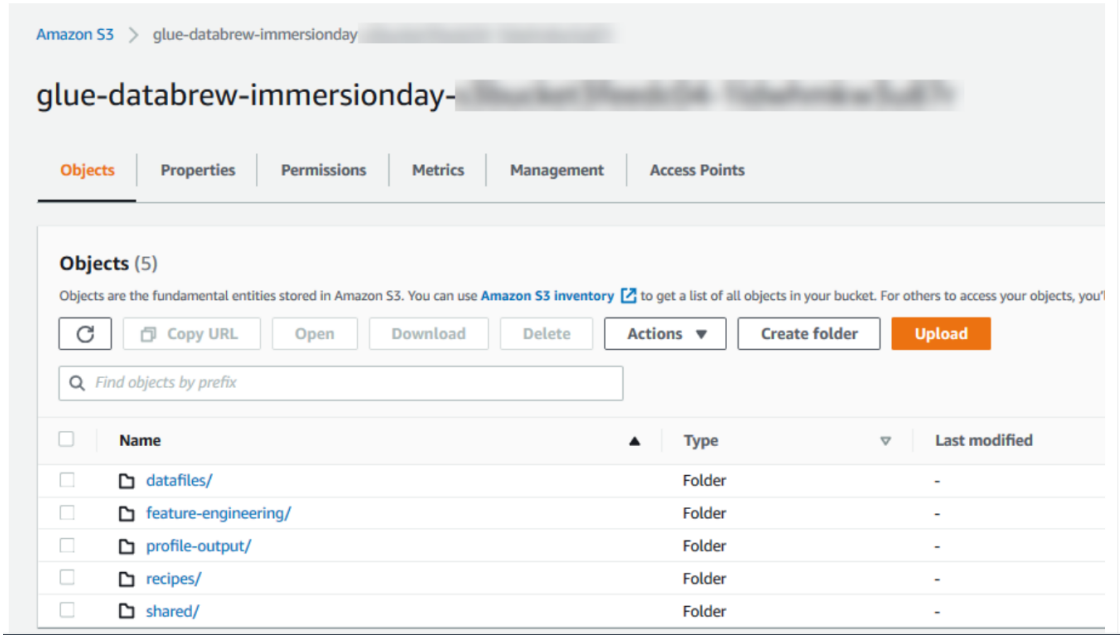
Data Model



Logical data flow



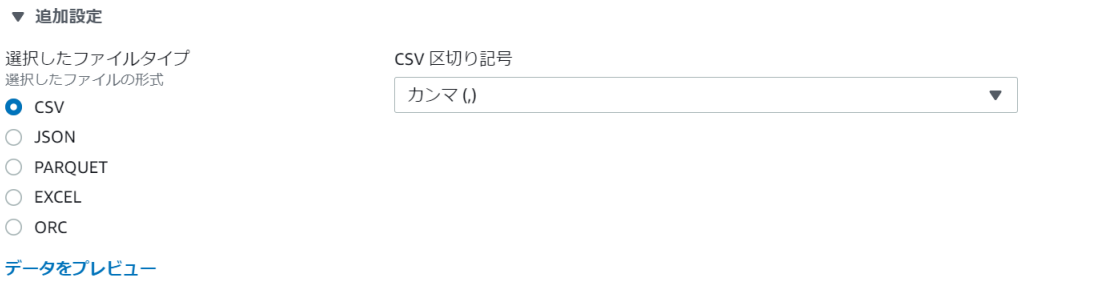
3. 実行が完了すると以下の通り S3 バケットが生成され、フォルダが複数できています。生成された S3 バケットの名前をコピーしておきます（” glue-databrew-immersionday”がバケット名についています）



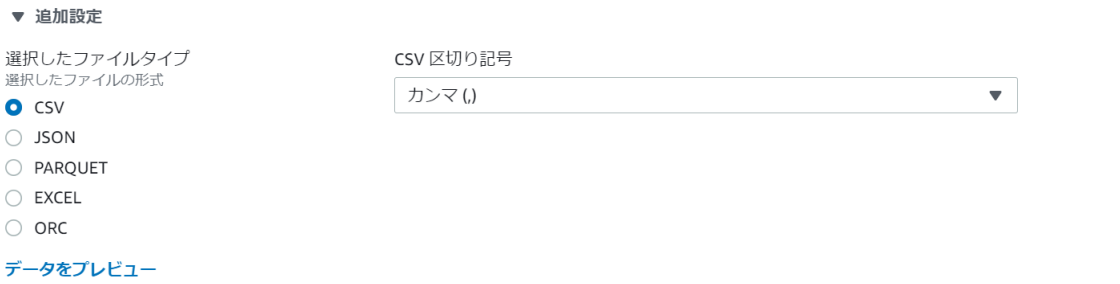
4. Glue DataBrew のマネージメントコンソールへ移動します
5. 左ペインから[データセット]をクリックします

6. [新しいデータセットの接続]をおします
7. [Customers]と名前をつけます
8. 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

9. カンマ区切り CSV を選びます



10. 画面一番下の[データセットを作成]をおします



11. 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

12. 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

13. [データプロファイルを実行]を押します
14. [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
15. [完全なデータセット]を選びます

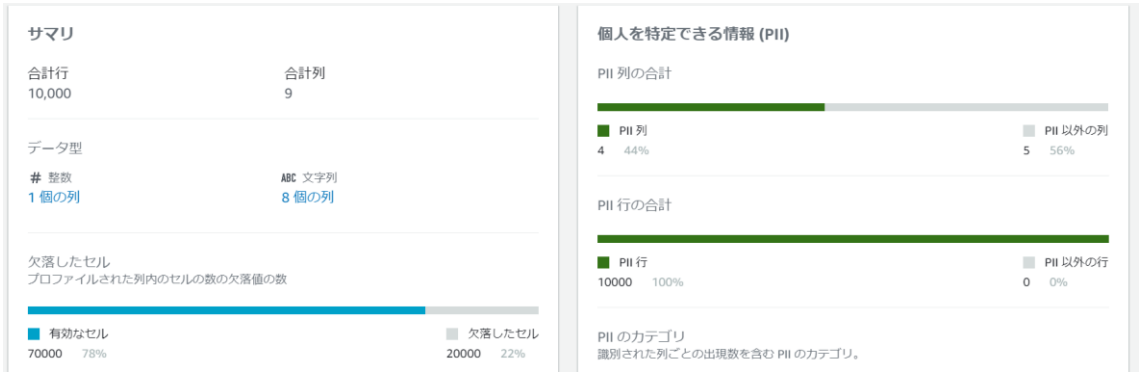
16. 出力先に、先程作成した S3 バケットの[profile-output]フォルダを選びます

17. [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

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

19. [ジョブを作成し実行する]をおします
20. 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

21. 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

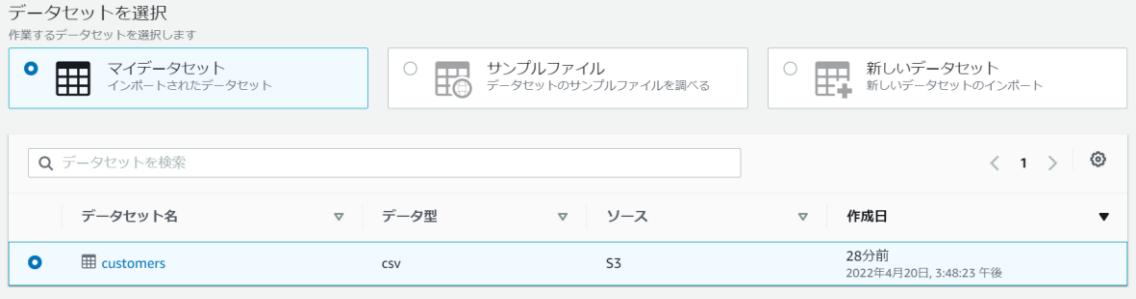


22. [データ系列]タブを見るとどのようにデータが構成されているかが把握できます

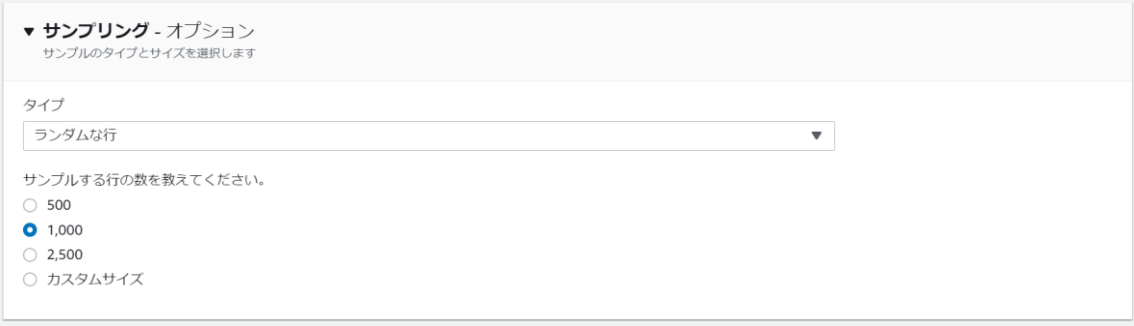
ではこれから、Transform ジョブを作成します。Transform とはデータの変換作業のことです。ジョブは先程のデータセットに対して行うよう作成します

23. 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
24. 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

25. マイデータセットで customers を選びます



26. [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



27. 先程作成した IAM ロールを選びます

28. [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

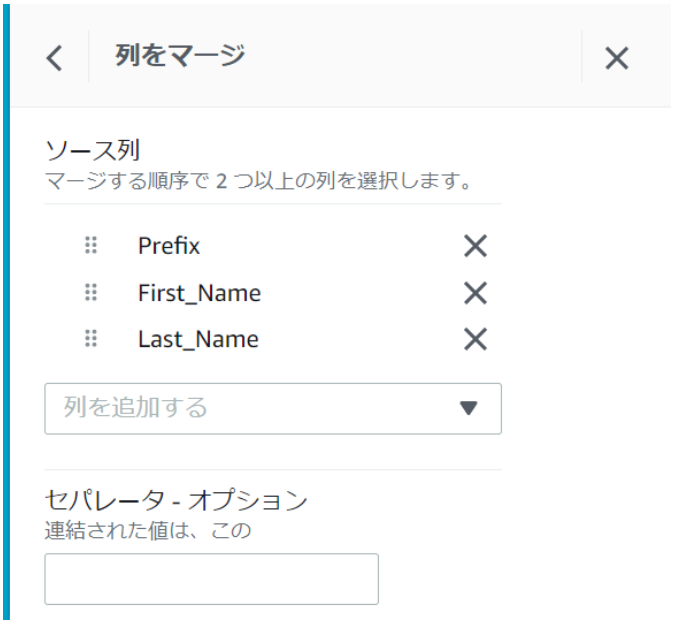
29. はじめて作業される方は画面右、レシピにいくつか設定が表示されますが、全て削除してください。

30. 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

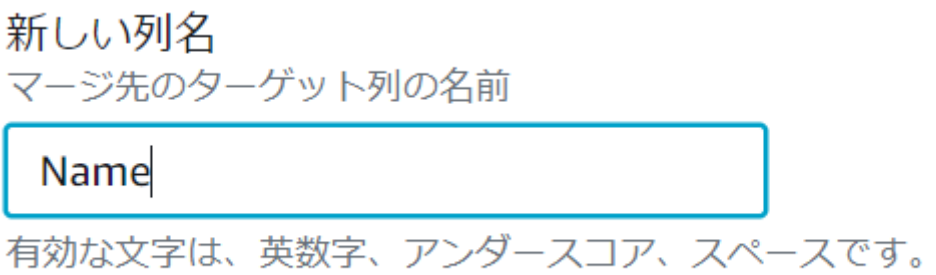
能です）

ではまず、[マージ]をおします

31. 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）



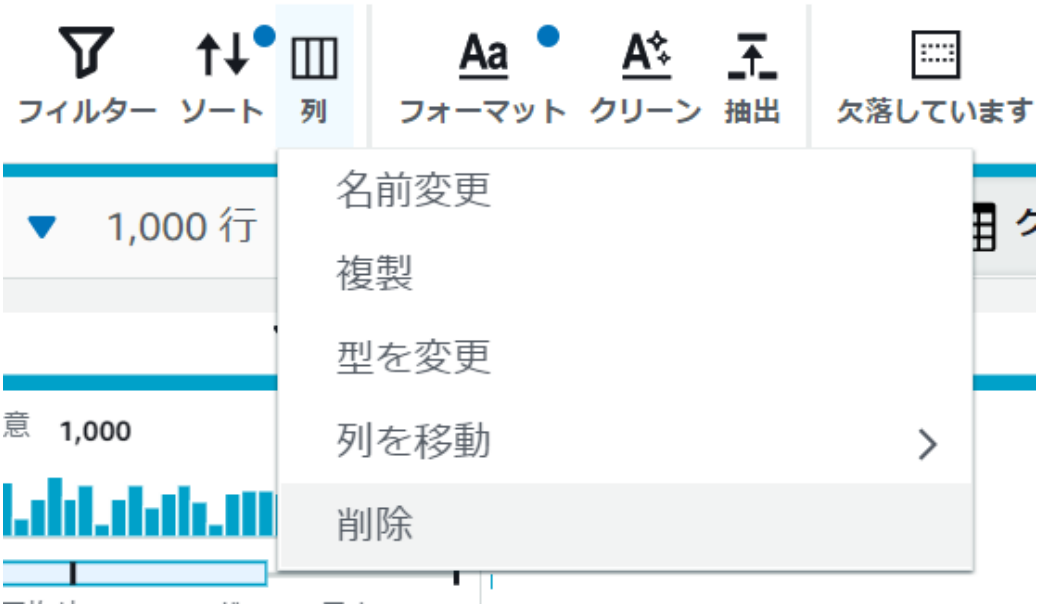
32. [新しい列名]に[Name]と入力します



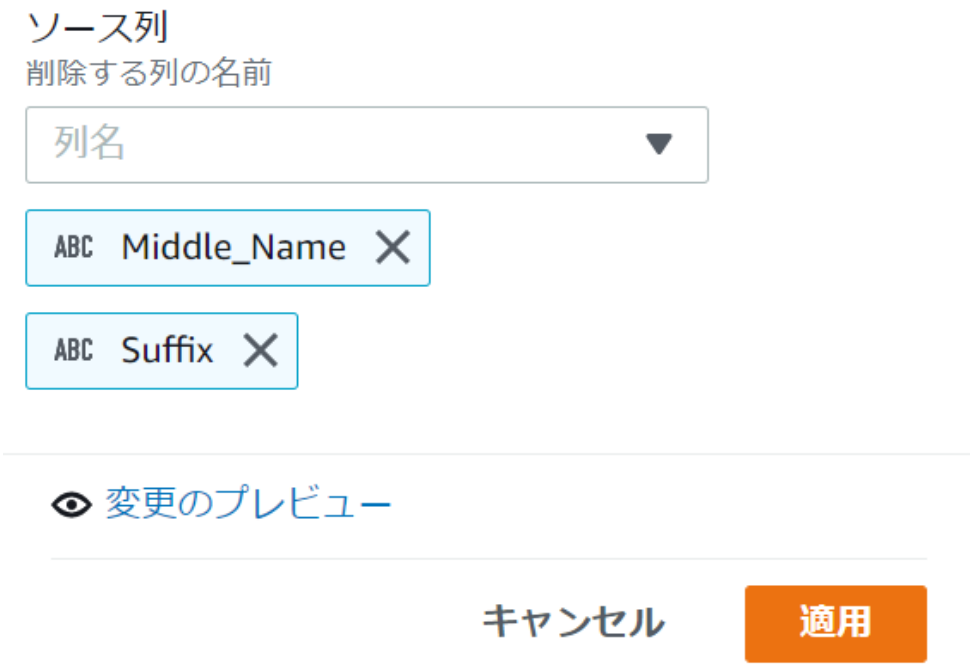
33. [変更のプレビュー]をおしてみてください。3 つの列が Name に結合されています。
34. [適用]をおします
35. レシピにステップが追加されました。この作業を繰り返していきます。

36. 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

37. [適用]をおします
38. 次は、列→削除を選びます



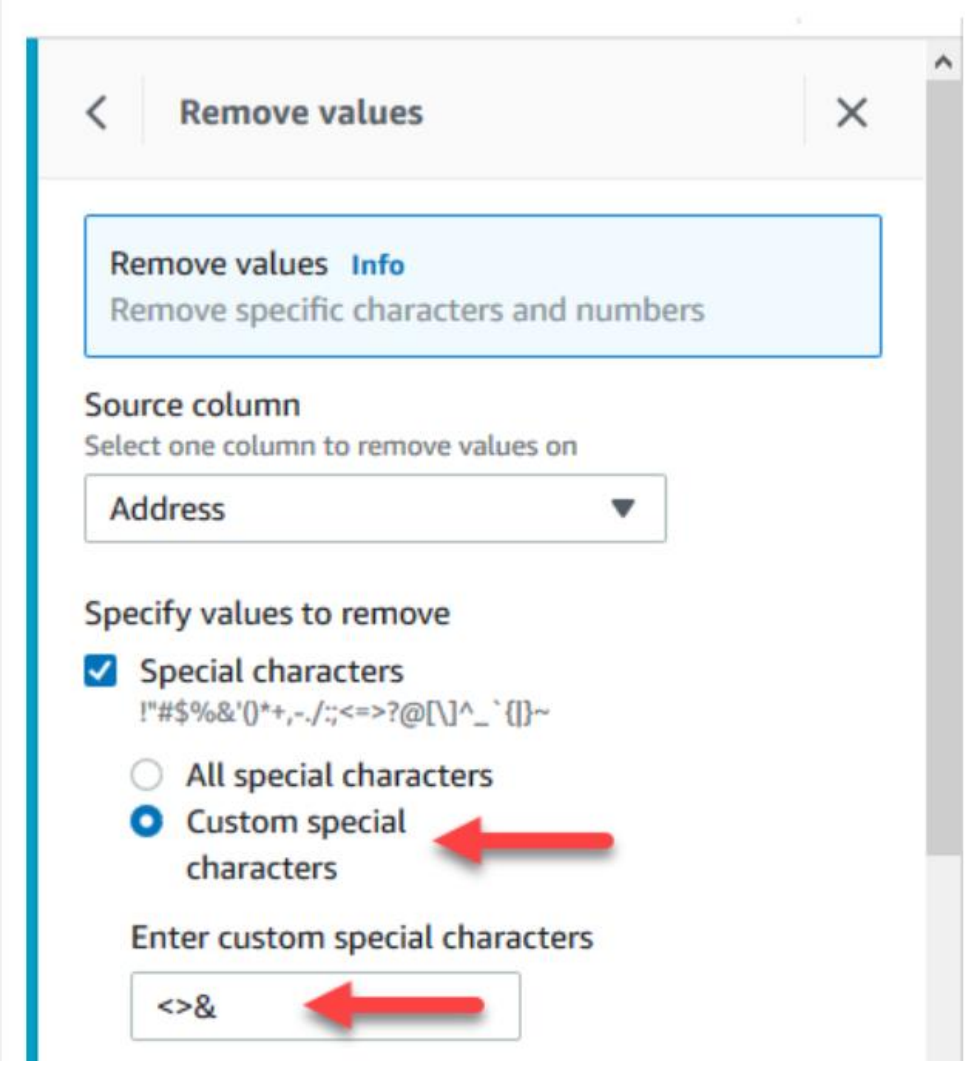
39. 以下 2 列を選んで、適用をおします



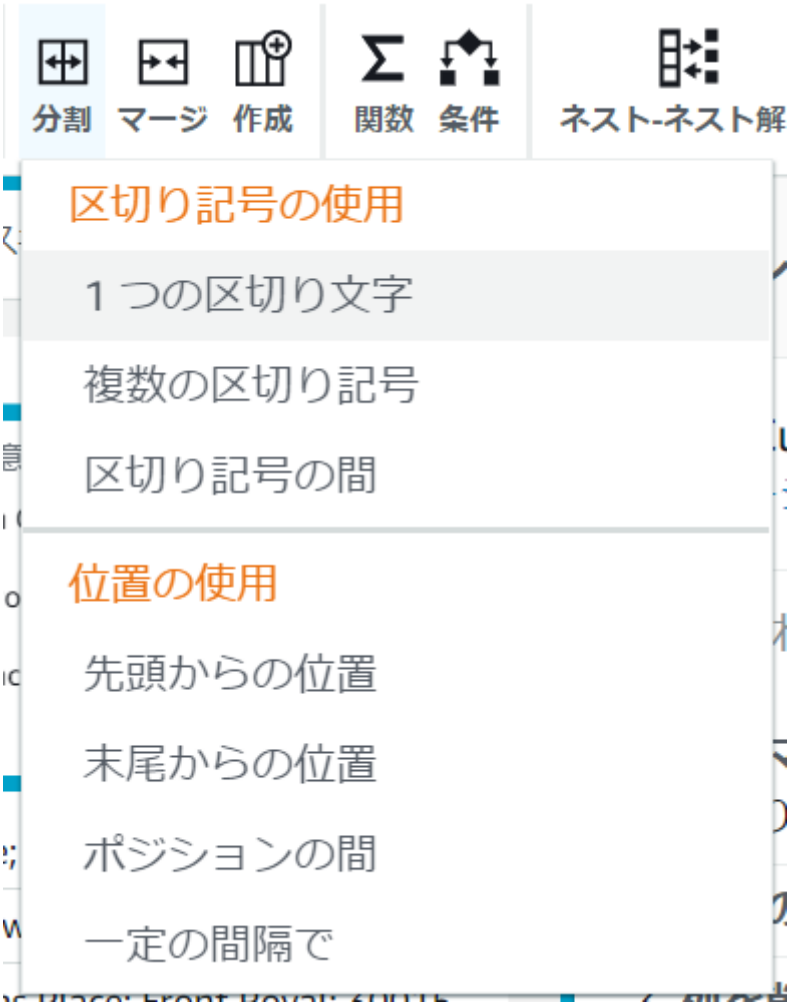
40. 次に DoB（Data of Birth）の[…]から、フォーマット→日付時刻の形式を選び、[mm-dd-yyyy]を選び、[適用]をおします

41. [Address]列からクリーン→特殊文字、を選びます

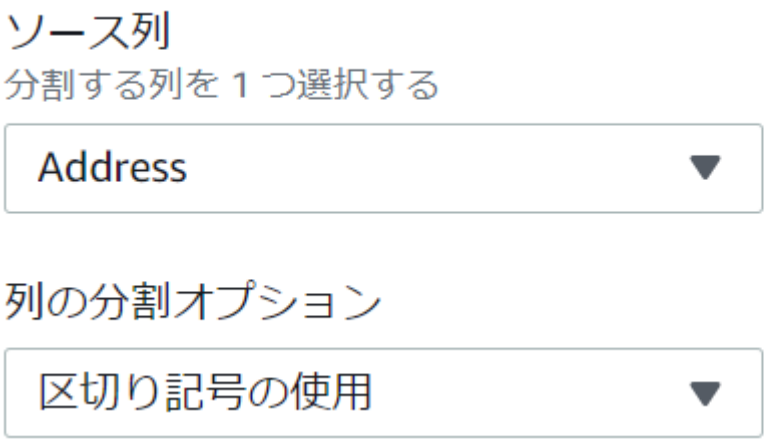
42. 以下の通り設定します

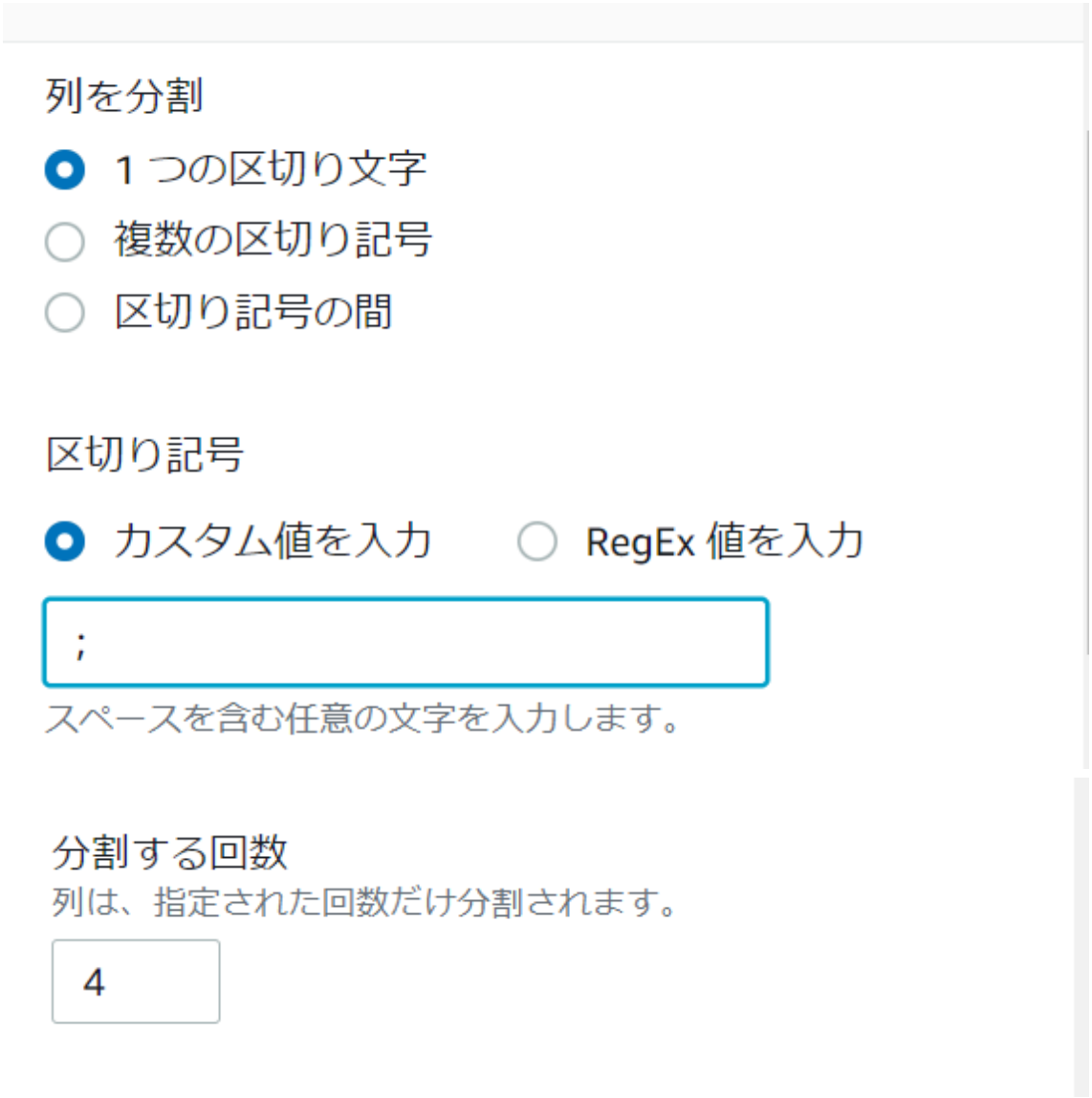


43. [適用]をおします
44. 次に[分割]→[一つの区切り文字]を選びます



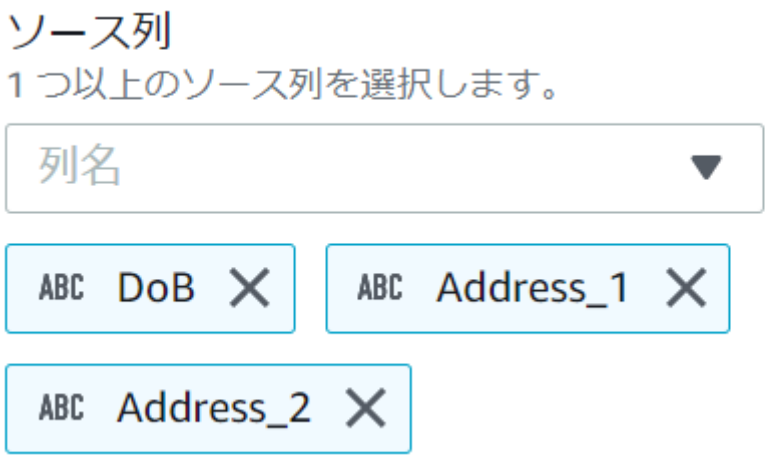
45. Address を選び以下の設定にします。[;]で Address 列を 5 分割します。区切り文字が 4 つなので 4 か所で割ると 5 列になる、という意味です。





46. [適用]をおします
47. [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。
48. 画面右上の詳細から[機密]→[値をマスキング]を選びます

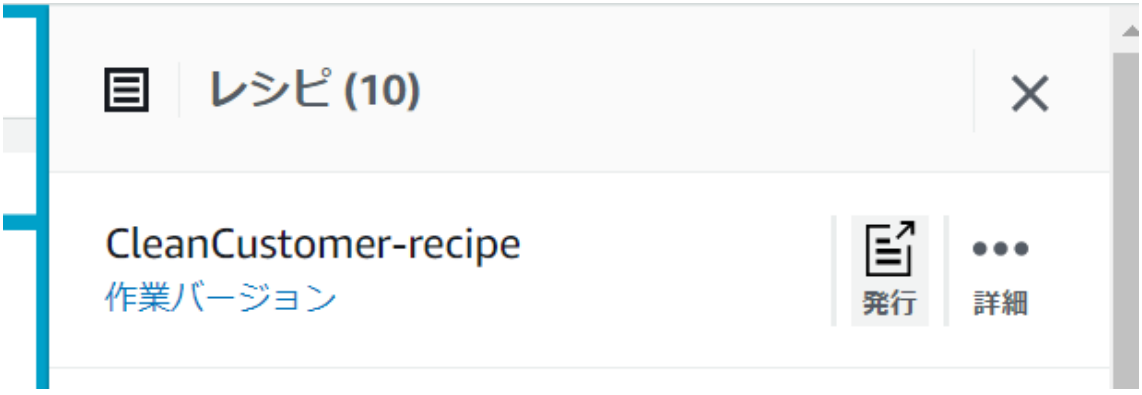
49. 以下の 3 つを選んで[適用]をおしてください。



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。

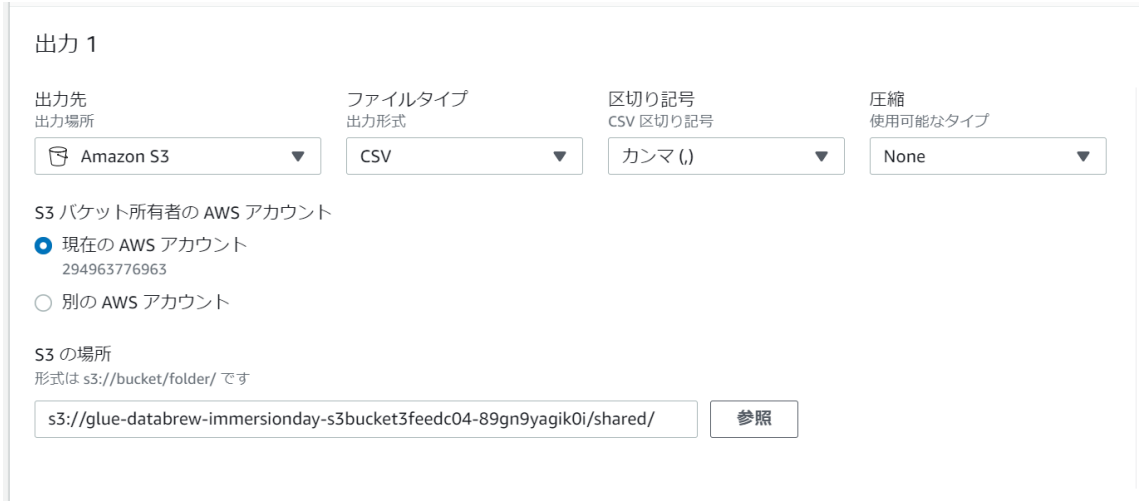
50. ここまでの作業で 10 個のステップがレシピに追加されています。

51. 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



52. 画面右上[ジョブを作成]をおします

53. 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



54. 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
55. 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。
56. 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。

	A	B	C	D	E	F	G	H	I	J
1	Customer	Name	Gender	DoB	Address_1	Address_2	City	Zip	Country	
2	1	MR. PAUL SHELTON	M	##-##-####	###	###### ######### ######	Porterville	30001	US	
3	2	MRS. LISA CLARKE	F	##-##-####	#####	##### ###### #####	Bourbonnais	30080	US	
4	3	DR. LEE HARRISON	M	##-##-####	#####	###### #### ######	Nutley	30024	US	
5	4	MS. CHLOE JONES	F	##-##-####	###	#### ##### #####	Palestine	30020	US	
6	5	MRS. KRISTEN JACKSON	F	##-##-####	###	####### ### #####	Acworth	30060	US	
7	6	MISS MONICA PITTS	F	##-##-####	####	##### ######### ####	Beaumont	30071	US	
8	7	MRS. KATELYN GORDON	F	##-##-####	#####	##### ###### #####	Thomasville	30052	US	

57. 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。

58. 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

59. 名前に[Sales DQ Checks]と入力します
60. 関連付けられたデータセットから[sales]を選びます
61. 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

62. 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

63. [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

64. [ルールセットを作成]をおします
65. 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

66. 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

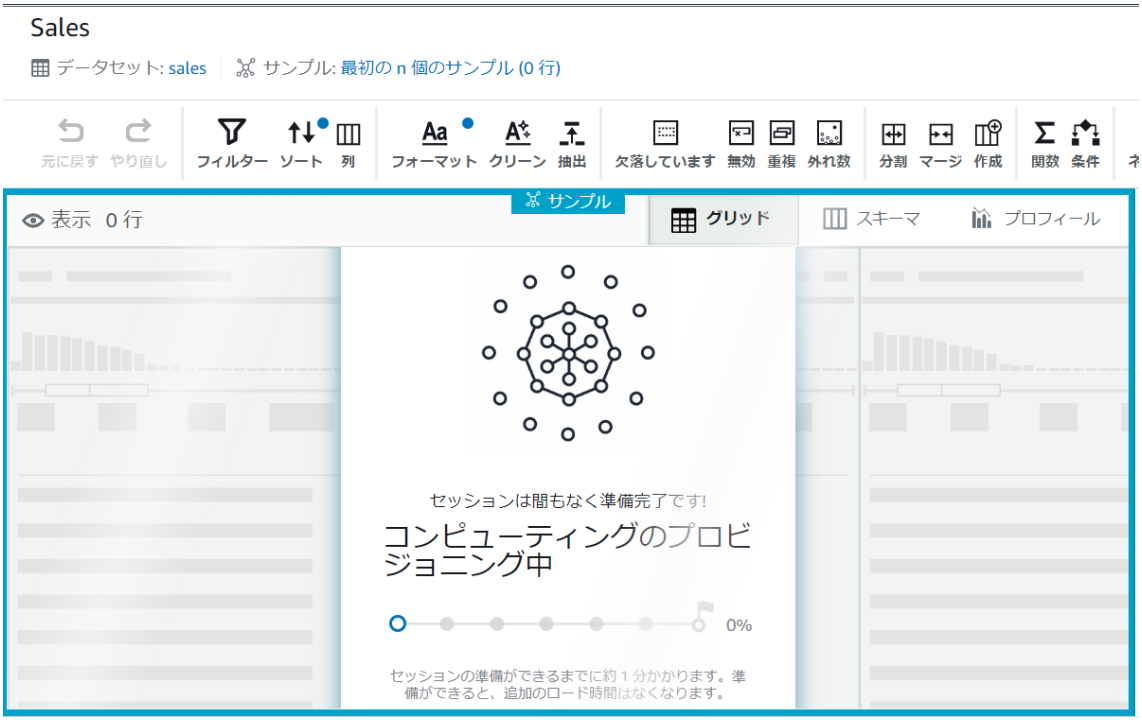
67. S3 バケットの[/profile-output/]を選びます

68. 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。

69. 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。

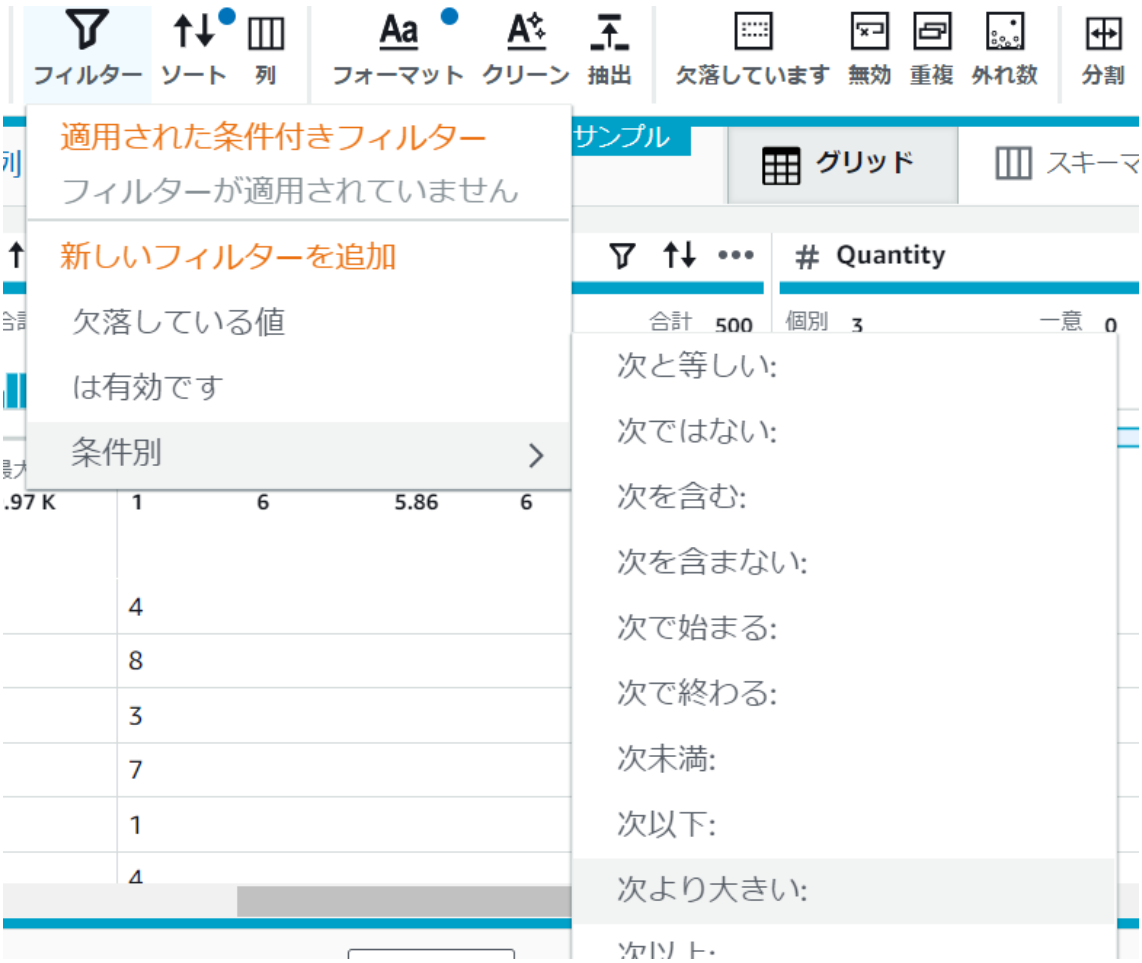
70. ではこれから Sales データセットに対して、データ変換の実装と、Customers データセットとの Join を行っていきます。
71. [Sales]という名前のプロジェクトを作成します
72. まずは sales データセットを選択します

73. 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



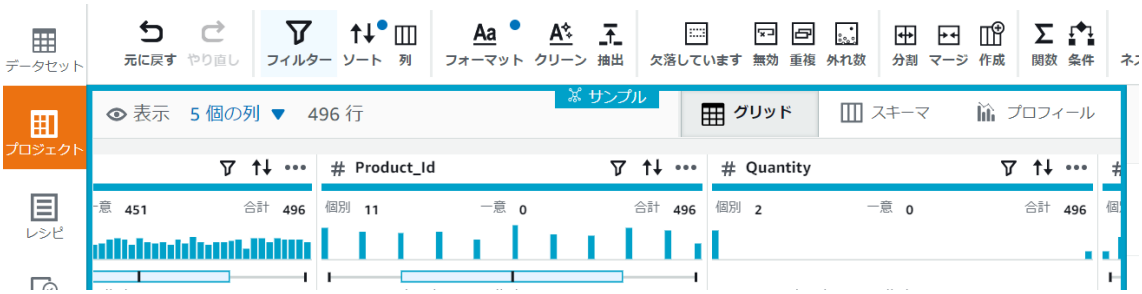
74. まず先程の DQ（データ品質）ルールに応じたデータとなるように、フィルターを選びます

75. [条件別]→[次より大きい]を選びます

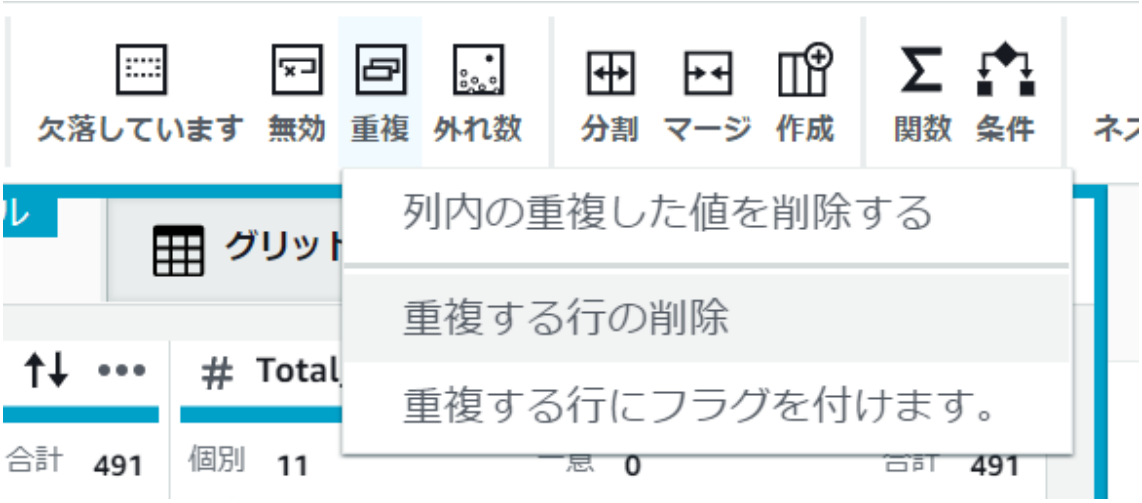


76. 以下のように 0 より大きいものを設定します。(注意：フィルタールールでは、残すものを指定します)

77. [適用]をおすと、データ総数が 500 から 496 に減っています



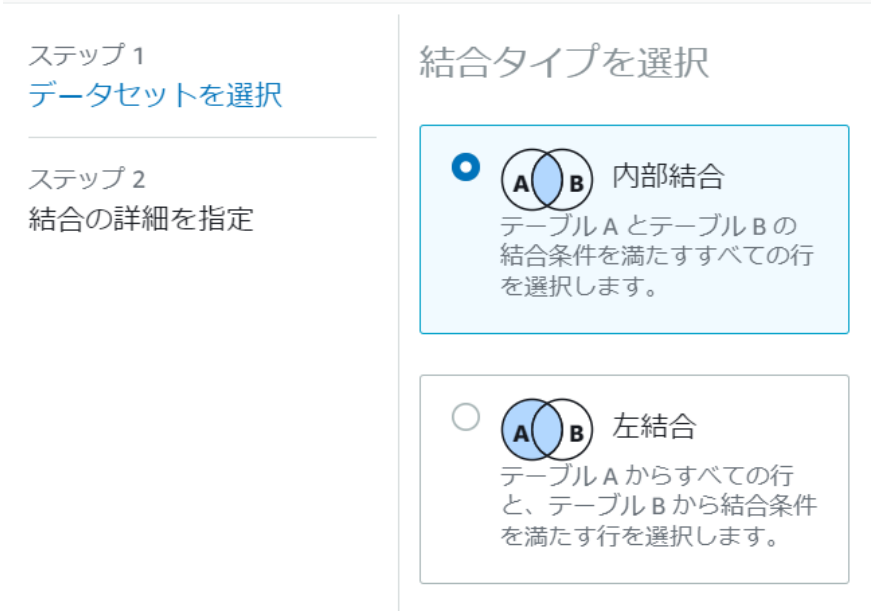
78. 同じように、もう 1 個[Total_Sales]に対しても設定をしてください。データ総数は 491 になります。
79. 今度は[重複する行の削除]を選びます



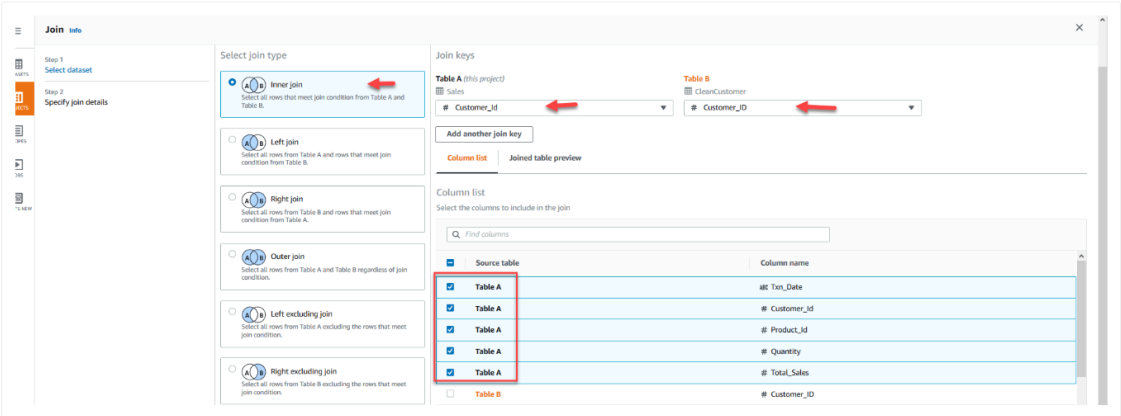
80. [適用]をおすと、データ総数は 481 になります。
81. 次に customer データセットとの join を行います。[結合]をおします（注意：すでにある customer データセットではなく、前段で作業を行った変換された customer データを用いるため、新たにデータセットを作成します）

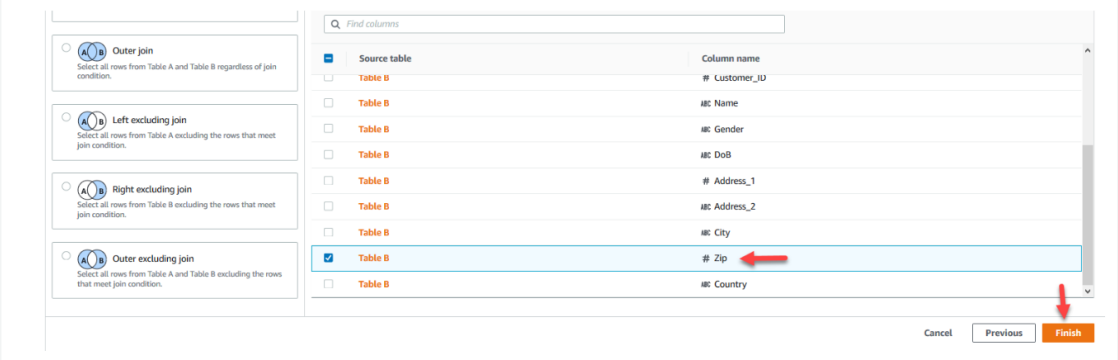
82. [新しいデータセットへの接続]をおします
83. 名前に[CleanCustomer]と入力します
84. S3 バケットの[/shared/]に出力されている変換済データセット(csv ファイル)を選びます

85. [データセットを作成]をおします
86. [次へ]をおします
87. [内部結合]を選びます



88. 以下の画面の通り設定を行います





89. [終了]をおします

90. レシピを[発行]し、[ジョブを実行]をおします。
91. S3 バケットには[/shared/]フォルダを指定します
92. 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