

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

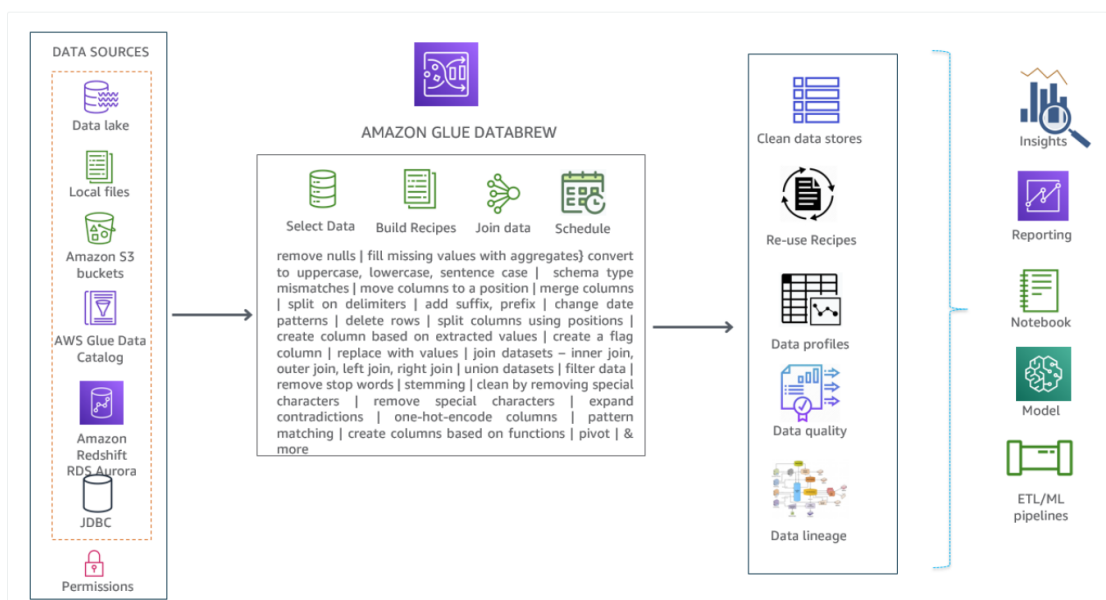
はじめに：AWS Glue DataBrew はデータのクリーニングと正規化を迅速にするビジュアルデータ準備ツールです。このツールでは多くの一般的なデータ変換作業がノーコードで行えるように様々な処理があらかじめ組み込まれています。別途 Glue Studio をという近しいことを実現させるツールもありますが、こちらはローコードで様々な実装を行うとともに、スクリプトを自前で作成し作業を行いたい方向けという違いがありますが、お互いを組み合わせて使うことも可能です。

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](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

④ The following resource(s) require capabilities: [AWS::IAM::Role]

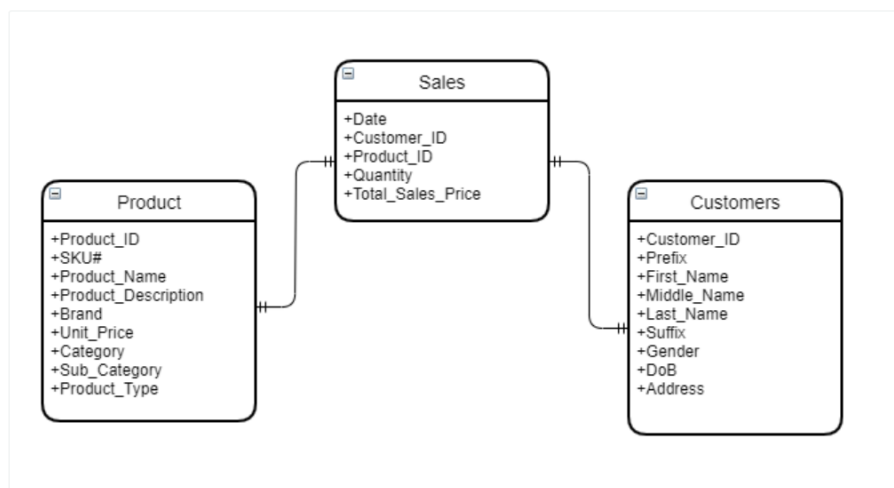
このテンプレートには、ご利用の AWS アカウントに変更を加えるエンティティにアクセスを与える可能性を持つ Identity and Access Management (IAM) リソースが含まれています。これらのリソースを個別に作成し、それぞれに最小限必要な権限を与えるかどうかを確認してください。 [詳細はこちら](#)

☐ AWS CloudFormation によって IAM リソースが作成される場合があることを承認します。

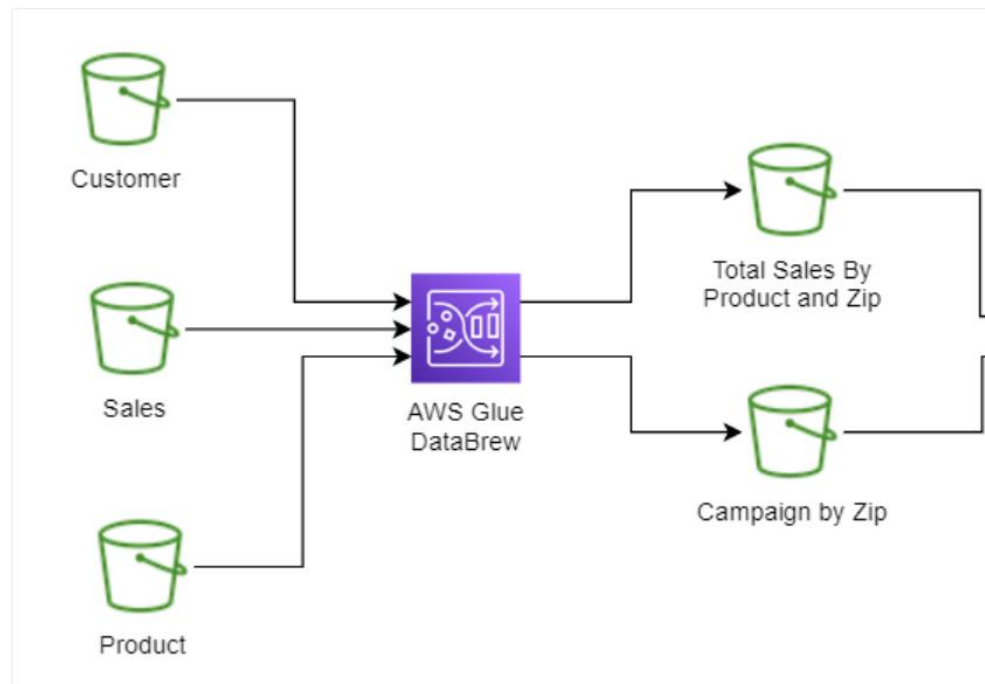
キャンセル 変更セットの作成 **スタックの作成**

以下のデータが S3 へ展開されます

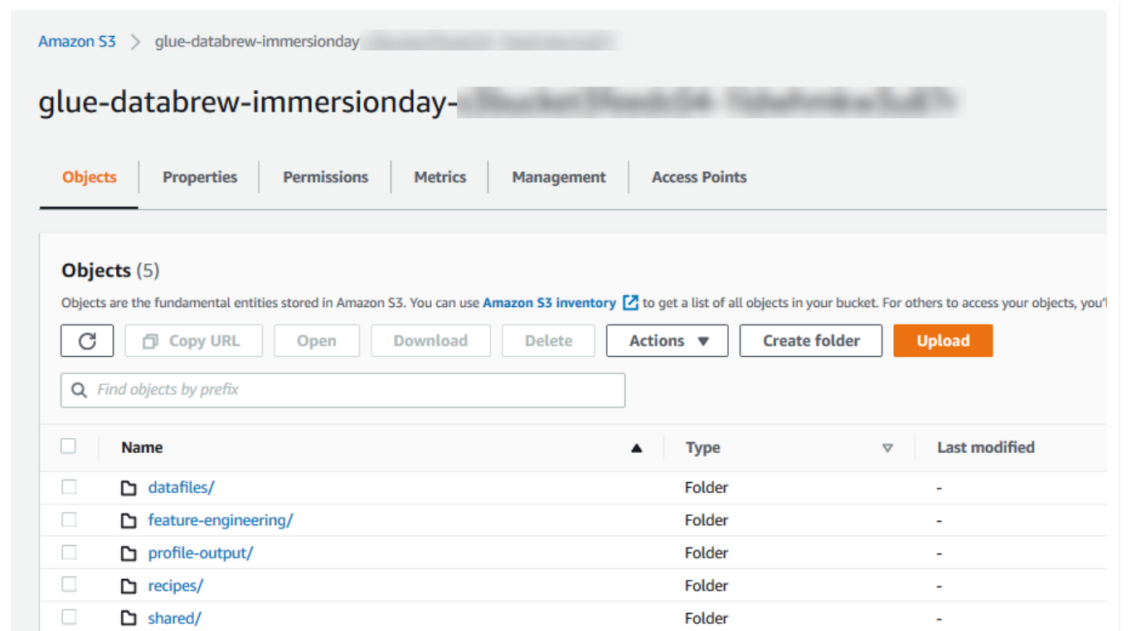
Data Model



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

↑ ファイルをアップロード

データレイク/データストア

Amazon S3

データベース接続

Amazon Redshift

JDBC

AWS Glue データカタログ

データカタログ S3 テーブル

データカタログ Redshift テーブル

データカタログ RDS テーブル

すべての AWS Glue テーブル

S3 からソースを入力
フォルダを選択するには、フォルダの中のすべてのファイルが同じファイルタイプを共有する必要があります。異なるスキーマがある場合は、マージされます。

s3://glue-databrew-immersionday-s3bucket3feedc04-89gn9yagik0i/datafiles/customers

形式 s3://bucket/prefix/

customers フォルダ内のすべてのファイルが選択されています

S3 Buckets > glue-databrew-immersionday-s3bucket3feedc04-89gn9yagik0i > datafiles

フォルダ全体を選択する

S3 オブジェクトを名前を検索する

名前	サイズ	最終更新日
customers	-	
historicalsales	-	

9. カンマ区切り CSV を選びます

▼ 追加設定

選択したファイルタイプ
選択したファイルの形式

☒ CSV

☐ JSON

☐ PARQUET

☐ EXCEL

☐ ORC

データをプレビュー

CSV 区切り記号

カンマ (,)

10. 画面一番下の[データセットを作成]をおします

▼ 追加設定

選択したファイルタイプ
選択したファイルの形式

☒ CSV

☐ JSON

☐ PARQUET

☐ EXCEL

☐ ORC

データをプレビュー

CSV 区切り記号

カンマ (,)

11. 以下の通りデータセットが作成されました。

データセット (1)

Q データセットを検索

<input type="checkbox"/>	データセット名	データ型	データプロファイル	ソース	ロケーション
<input type="checkbox"/>	customers	CSV	-	S3	s3://glue-databrew-immersionday-s3bucket3feedc04-89gn9yagik0i/datafiles/customers/customers.csv

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

データセット (1)

Q データセットを検索

<input type="checkbox"/>	データセット名	データ型	データプロファイル	ソース	ロケーション
<input type="checkbox"/>	customers	CSV	-	S3	s3://glue-databrew-immersionday-s3bucket3feedc04-89gn9yagik0i/datafiles/customers/customers.csv

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、そのまま進めます

15. [完全なデータセット]を選びます

ジョブ実行サンプル

ジョブは、データセット全体またはデータセットのカスタムサンプルに対して実行できます。

データサンプル
ジョブを実行するデータセットの範囲を定義

☒ 完全なデータセット
☐ カスタムサンプル

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

ジョブ出力設定

ジョブを実行すると、指定したファイルの送信先に出力ファイルが生成されます。

S3 バケット所有者の AWS アカウント
☒ 現在の AWS アカウント
 294963776963
☐ 別の AWS アカウント

ファイルタイプ
 出力形式
 JSON

S3 の場所
 形式は s3://bucket/folder/ です

暗号化
☐ ジョブ出力ファイルの暗号化を有効にする
 SSE-S3 または AWS KMS を使用してジョブ出力ファイルを暗号化する

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

データセットレベル設定
生成されたデータプロファイルに含めるデータセットレベルの統計を選択します。

デフォルト	PII 統計 影響大
重複した値 有効	<input checked="" type="checkbox"/> PII 統計を有効化 PII (個人を特定できる情報) を使用して列の統計を識別します。
高度な統計	PII のカテゴリ データについて評価する PII のカテゴリを指定します。DataBrew は米国内の PII のカテゴリを識別できます。
PII 統計 有効, すべてのカテゴリ	<input type="radio"/> カテゴリを選択 <input checked="" type="radio"/> すべてのカテゴリ
相関ウィジェット 有効, 最初の 10 数値列	

③ 日付と人の名前 の PII のカテゴリを含めると、ジョブの実行時間が長くなります。日付と人の名前 を含まずに個別のカテゴリを指定すると、実行時間を短縮できます。

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

許可
DataBrew needs permission to connect to data on your behalf. Use an IAM role with the [必須ポリシー](#) attached.

ロール名
データに接続するためのアクセス権を持つロールを選択します。最新の更新を表示するには更新します。

新しい IAM ロールを作成

新しい IAM ロールのサフィックス
ロールには、「AWSGlueDataBrewServiceRole-」というプレフィックスが付きます

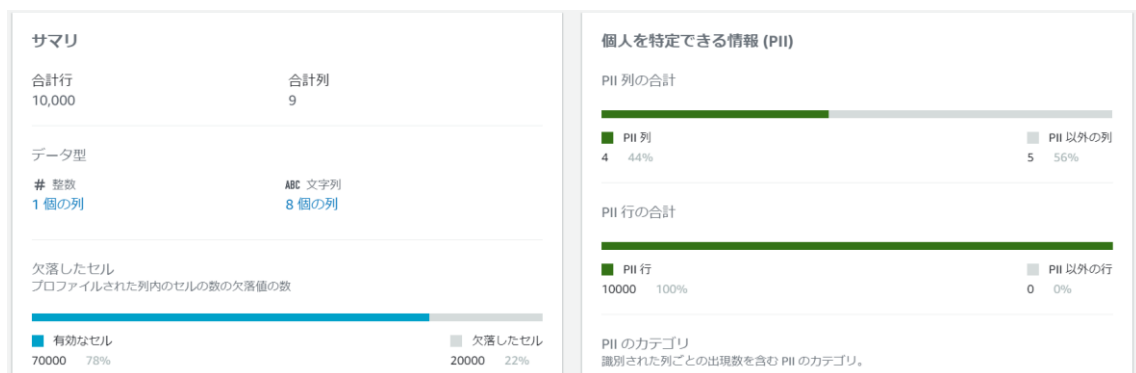
ID

[ジョブを作成] をクリックすると、このロールの作成を許可したことになります。

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

プロジェクトの詳細

プロジェクト名

プロジェクト名は 1~255 文字にする必要があります。有効な文字は、英数字 (A~Z, a~z, 0~9)、ハイフン (-)、ピリオド (.), およびスペースです。

レシピの詳細

DataBrew のデータクリーンアップステップはレシピとして保存されます。レシピはデフォルトでプロジェクトに接続されます。プロジェクトが関連付けられていない既存のレシピをプロジェクトに適用することもできます。

アタッチされたレシピ

新しいレシピを... ▼

レシピ名

レシピ名は 1~255 文字にしてください。有効な文字は、英数字 (A~Z, a~z, 0~9)、ハイフン (-)、ピリオド (.), およびスペースです。

☐ レシピからステップをインポートする
既存のレシピからプロジェクトにレシピをインポートします。選択した既存のレシピは編集されません。

25. マイデータセットで customers を選びます

データセットを選択
作業するデータセットを選択します

☒  **マイデータセット**
インポートされたデータセット

☐  **サンプルファイル**
データセットのサンプルファイルを選ぶ

☐  **新しいデータセット**
新しいデータセットのインポート

Q データセットを検索

データセット名	データ型	ソース	作成日
<input checked="" type="radio"/>  customers	CSV	S3	28分前 2022年4月20日, 3:48:23 午後

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

▼ **サンプリング - オプション**
サンプルのタイプとサイズを選択します

タイプ

ランダムな行

サンプルする行の数を教えてください。

☐ 500

☒ 1,000

☐ 2,500

☐ カスタムサイズ

27. 先程作成した IAM ロールを選びます

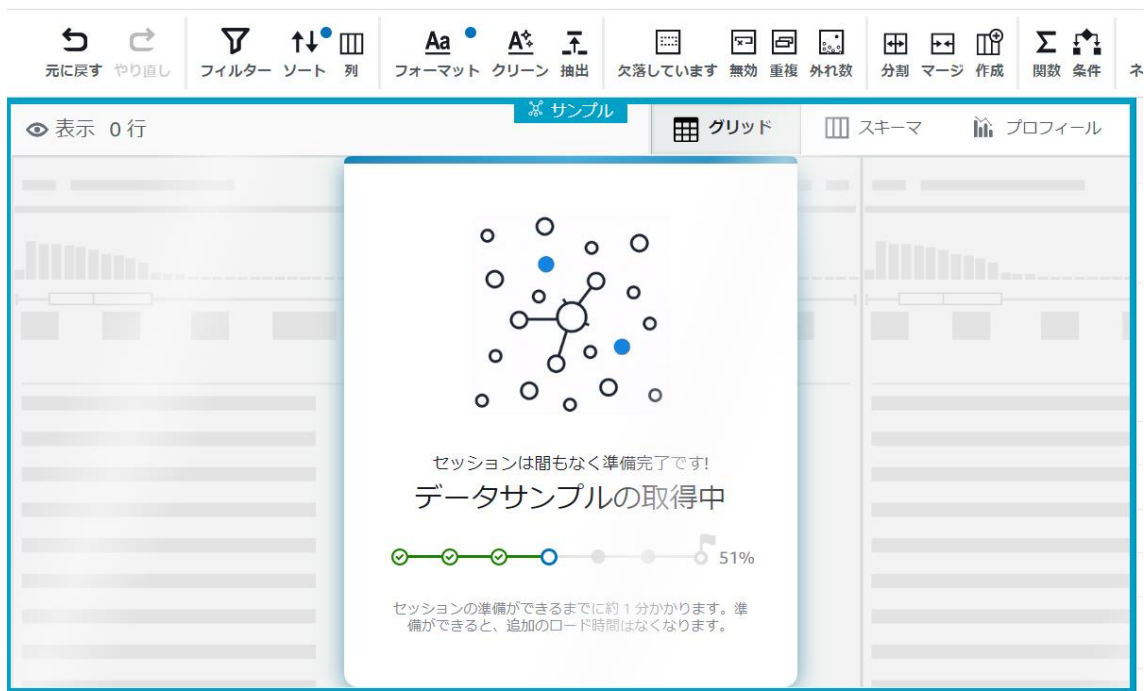
許可
DataBrew needs permission to connect to data on your behalf. Use an IAM role with the [必須ポリシー](#) attached.

ロール名
データに接続するためのアクセス権を持つロールを選択します。最新の更新を表示するには更新します。

AWSGlueDataBrewServiceRole-ID2

[プロジェクトを作成] をクリックすると、このプロジェクトのすべてのデータセットにアクセスするために必要なアクセス許可を、選択したサービスロールに追加することが DataBrew に許可されます。

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待ちます



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)

The screenshot shows the '列をマージ' (Merge Columns) dialog box. It has a title bar with a back arrow, the title '列をマージ', and a close button (X). Below the title bar, it says 'ソース列' (Source Columns) and 'マージする順序で 2 つ以上の列を選択します。' (Select 2 or more columns in the order to merge). There is a list of columns: 'Prefix', 'First_Name', and 'Last_Name'. Each column has a selection icon (three dots) and a close button (X). Below the list is a button '列を追加する' (Add Column) with a dropdown arrow. At the bottom, there is a section 'セパレーター - オプション' (Separator - Option) with the text '連結された値は、この' (The concatenated value is this) and an empty text box.

32. [新しい列名]に[Name]と入力します

新しい列名

マージ先のターゲット列の名前

The screenshot shows the '新しい列名' (New Column Name) input field. It is a text box with a blue border. The text 'Name' is entered in the field.

有効な文字は、英数字、アンダースコア、スペースです。

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。

目

レシピ (1)

×

CleanCustomer-recipe

作業バージョン

発行

詳細

適用されたステップ1

すべてクリア

1. 列をマージ Prefix, First_Name, Last_Name 次へ: Name

e 区切り " "

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

ソース

ABC Name

↑ ↓

...

ABC Middle_Name

ソート

Aa フォーマット

Ac クリーン

抽出

欠落した値の削除または入力

無効な値を削除または置換

重複した値を削除

列を分割

フラグ列を作成する

ネスト-ネスト解除

ワードトークン分割

BC

Name

文字列

大文字に変更

小文字に変更

大文字ケースへの変更

文の大文字と小文字に変更

数値

10 進数の精度

1,000 桁区切り記号

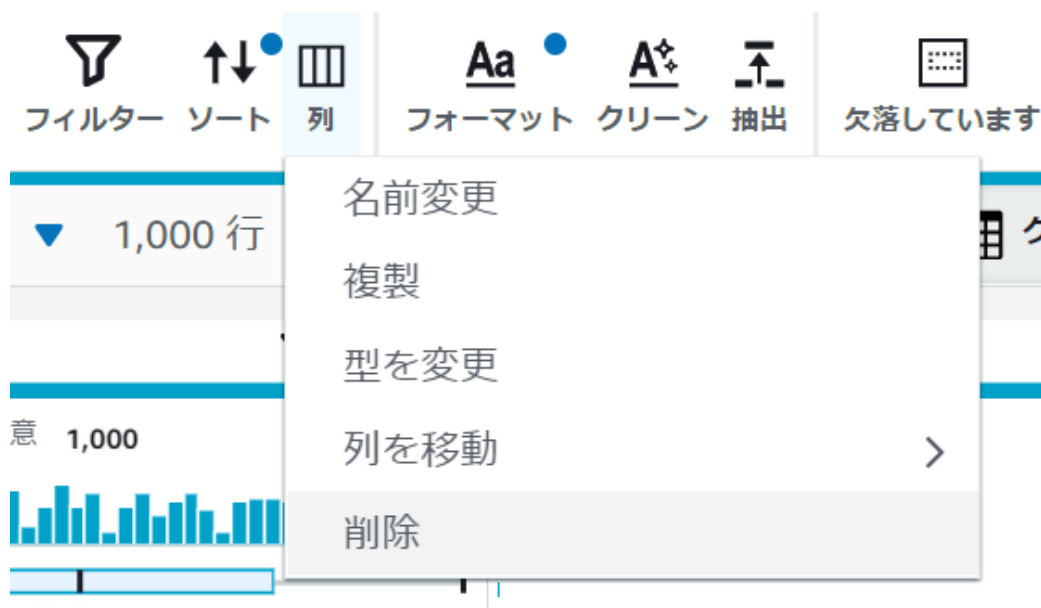
短縮番号

日付時刻の形式

Phone number

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

ソース列

削除する列の名前

列名 ▼

ABC Middle_Name ✕

ABC Suffix ✕

👁 変更のプレビュー

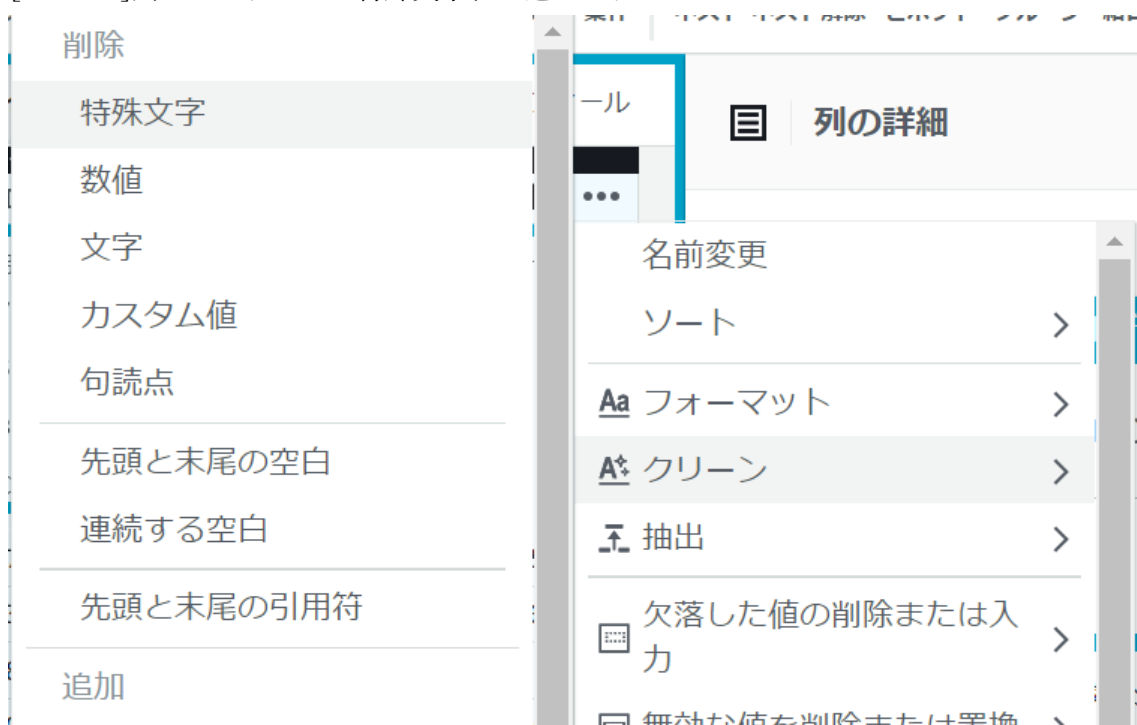
キャンセル

適用

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

< Remove values X

Remove values Info
Remove specific characters and numbers

Source column
Select one column to remove values on

Address ▼

Specify values to remove

☒ Special characters
!"#\$%&'()*+,-./:;<=>?@[\\]^_`{|}~

☐ All special characters

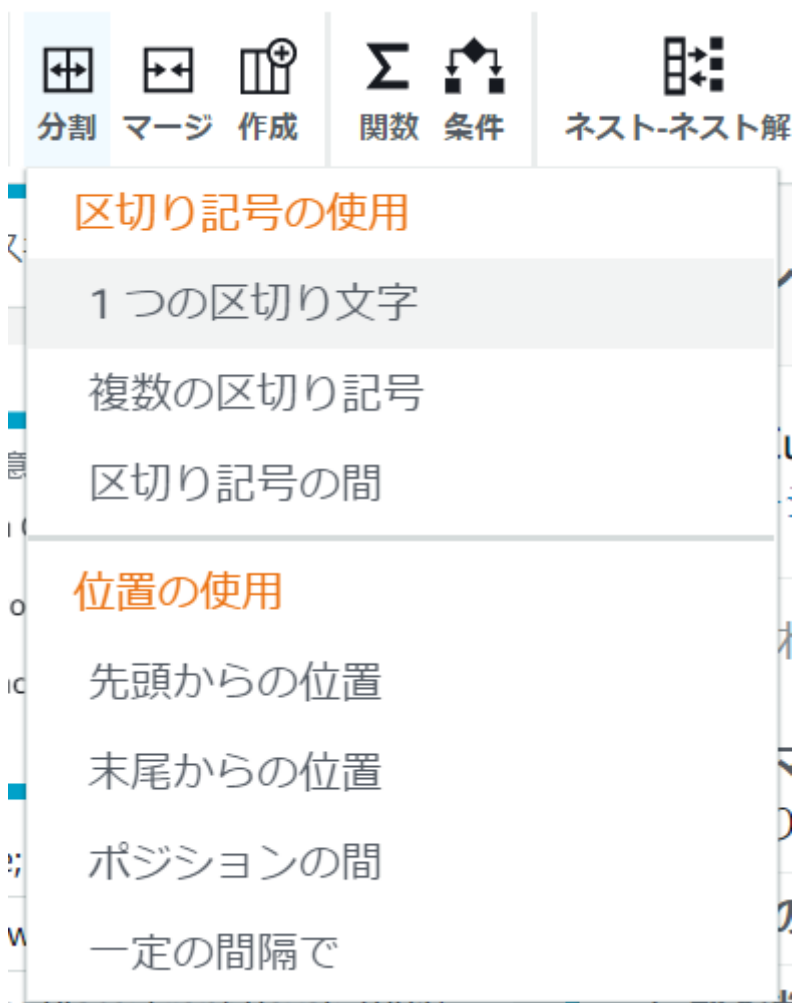
☒ Custom special characters

Enter custom special characters

<>&

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。

ソース列

分割する列を 1 つ選択する

Address ▼

列の分割オプション

区切り記号の使用 ▼

列を分割

- ☒ 1 つの区切り文字
- ☐ 複数の区切り記号
- ☐ 区切り記号の間

区切り記号

- ☒ カスタム値を入力
- ☐ RegEx 値を入力

;

スペースを含む任意の文字を入力します。

分割する回数

列は、指定された回数だけ分割されます。

4

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。

ソース列

1 つ以上のソース列を選択します。

列名 ▼

ABC
DoB
✕

ABC
Address_1
✕

ABC
Address_2
✕

注意：ある程度の一意性が求められる場合は、マスクングではなく、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。

目 レシピ (10)	×
1. 列をマージ Prefix, First_Name, Last_Name 次へ: Name 区切り " "	
2. 形式の変更 / Name 終了 大文字	
3. 列を削除 Middle_Name, Suffix	
4. 形式の変更 / DoB 終了 MM-dd-yyyy	
5. 削除 特殊文字 開始 Address	
6. 1 つの区切り記号で列を分割する ; 次において: Address	
7. 名前変更 Address_3 終了 City	
8. 名前変更 Address_4 終了 Zip	
9. 名前変更 Address_5 終了 Country	
10. マスク DoB と #	

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

目 | レシピ (10)

×

CleanCustomer-recipe

作業バージョン

発行

詳細

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

出力 1

出力先
出力場所

ファイルタイプ
出力形式

区切り記号
CSV 区切り記号

圧縮
使用可能なタイプ

S3 バケット所有者の AWS アカウント

S3 の場所
形式は s3://bucket/folder/ です

参照

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

データセットの詳細

🔍 レコメンデーション 2

II

☐ すべて選択

ルールセットに追加

📊 データセットの品質チェック

☐ チェック データセット 次のために: 重複した行

📊 列の品質チェック

☐ チェック すべての列 次のために: 欠落している値 == 0%

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

ルール名

Duplicate rows

データ品質チェックの範囲

各列の個別チェック ▼

ルールの成功条件

すべてのデータ品質チェックが満た... ▼

データ品質チェック

チェック 1

データ品質チェック

重複した行

重複した行の数についてデータセットをチェックします。

▼

条件

次と等しい:

▼

値

0

行 ▼

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

ルール 2

☒ ルールを有効化

削除

ルール名

Quantity and total Sales should be >0

データ品質チェックの範囲

ルール成功条件

選択した列の一般的なチェック ▼

すべてのデータ品質チェックが満た... ▼

選択した列

次のチェックが適用される列のリスト

☐ すべての列

☒ 選択した列

列: Quantity, Total_Sales

消去

正規表現: なし

データ品質チェック

チェック 1

データ品質チェック

数値

条件に基づき、数値について列をチェックします。

▼

条件

次より大きい:

▼

値

☒ カスタム値

☐ 列の値

0

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

データ品質のルールセット (1)

ルールセットを使用してプロファイルジョブを作成

アクション ▼

データ品質ルールセットを作成

Q

ルールセットを検索

<

1

>

⚙

<input checked="" type="checkbox"/>	データ品質ルールセット名 ▼	説明 ▼	関連付けられたデータセット ▼	関連付けられたジョブ ▼	作成日 ▼	作成者	タグ ▼
<input checked="" type="checkbox"/>	Sales DQ Checks 2 個のルール	-	sales	-	数秒前 2022年4月22日, 3:04:46 午後	arn:aws:iam::2949 63776963:root	-

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

ジョブ実行サンプル

ジョブは、データセット全体またはデータセットのカスタムサンプルに対して実行できます。

データサンプル

ジョブを実行するデータセットの範囲を定義

☒ 完全なデータセット
☐ カスタムサンプル

67. S3 バケットの[/profile-output/]を選びます

ジョブ出力設定

ジョブを実行すると、指定したファイルの送信先に出力ファイルが生成されます。

S3 バケット所有者の AWS アカウント

☒ 現在の AWS アカウント
294963776963
☐ 別の AWS アカウント

ファイルタイプ

出力形式

JSON

S3 の場所

形式は s3://bucket/folder/ です

暗号化

☐ ジョブ出力ファイルの暗号化を有効にする
SSE-S3 または AWS KMS を使用してジョブ出力ファイルを暗号化する

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。

sales

S3

sales.csv

323.3 KiB

1

進行中のジョブ

▶

プロファイルを再実行

このデータセットを使用してプロジェクトを作成する

アクション ▼

📄

ジョブの詳細

データセットのプレビュー

データプロファイルの概要

列の統計

データ品質ルール

データ系列

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。

sales sales.csv 323.3 KiB

プロフィールを再実行 このデータセットを使用してプロジェクトを作成する アクション

ジョブの詳細 ダウンロード

データセットのプレビュー データプロフィールの概要 列の統計 データ品質ルール データ系列

レコメンデーション 10

データ品質ルール (2)

すべて展開 すべて折りたたむ 検索

すべて (2) 成功 (0) 失敗 (2) エラー (0) 無効 (0)

☒ Sales DQ Checks 2 個のルール 失敗

- ☒ Duplicate rows
次に該当する場合はチェックデータセット 次を持
つ: 重複した行数 == 0
- ☒ Quantity and total Sales should be >0
次に該当する場合はチェックQuantity, Total_Sales 次を持つ: 値 > 0 次のために: 次以
上: 100% (行中)

成功 失敗 エラー
0 個の列 2 個の列 0 個の列

Duplicate rows

次に該当する場合はチェックデータセット 次を持
つ: 重複した行数 == 0

失敗

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

データセットを選択

作業するデータセットを選択します

☒ マイデータセット
インポートされたデータセット

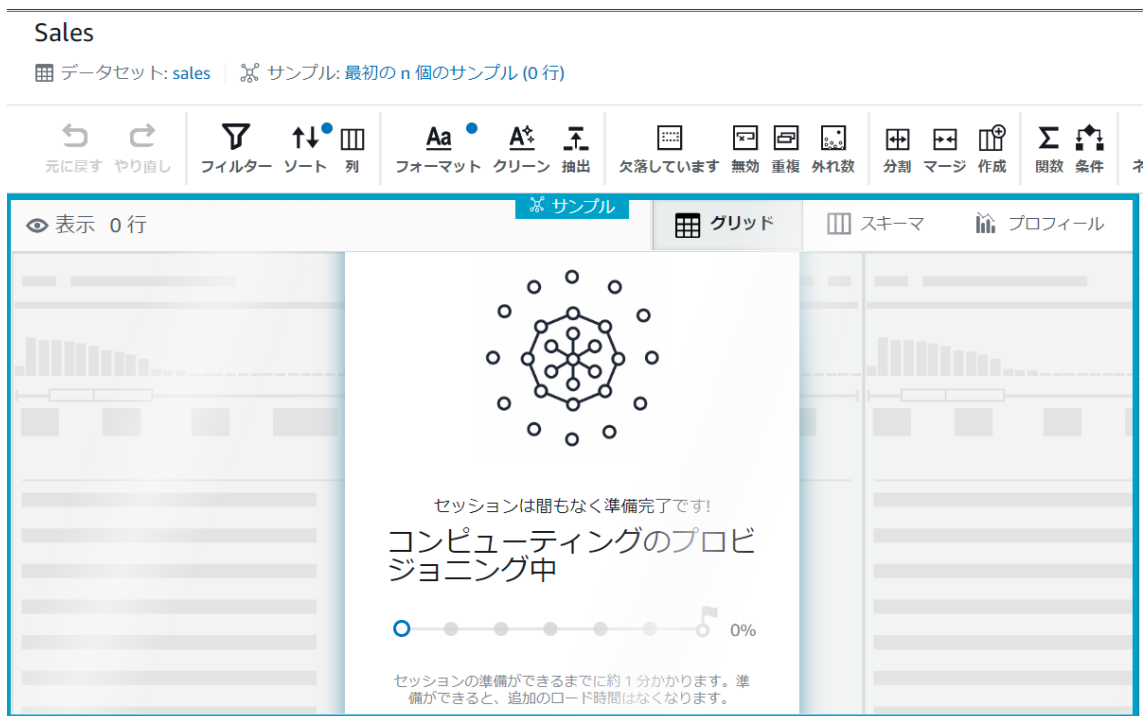
☐ サンプルファイル
データセットのサンプルファイル調べる

☐ 新しいデータセット
新しいデータセットのインポート

Q データセットを検索

データセット名	データ型	ソース	作成日
<input checked="" type="radio"/> sales	csv	S3	29分前 2022年4月22日, 2:52:04 午後
<input type="radio"/> customers	csv	S3	2日前 2022年4月20日, 3:48:23 午後

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

フィルター ソート 列 フォーマット クリーン 抽出 欠落しています 無効 重複 外れ数 分割

適用された条件付きフィルター
フィルターが適用されていません

新しいフィルターを追加

欠落している値
は有効です
条件別 >

サンプル グリッド スキーマ

▽ ↑↓ ... # Quantity

合計 500 個別 3 一意 0

次と等しい:
次ではない:
次を含む:
次を含まない:
次で始まる:
次で終わる:
次未満:
次以下:
次より大きい:

.97 K	1	6	5.86	6
	4			
	8			
	3			
	7			
	1			
	4			

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)

ソース列

Name of the column to filter

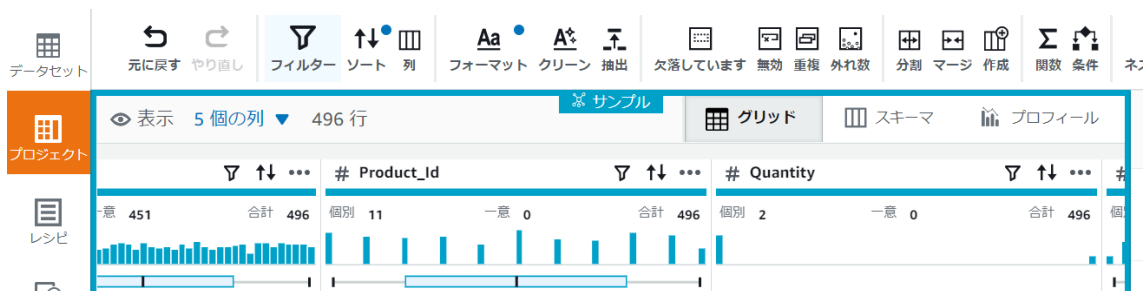
Quantity ▼

フィルター条件

次より大きい: ▼

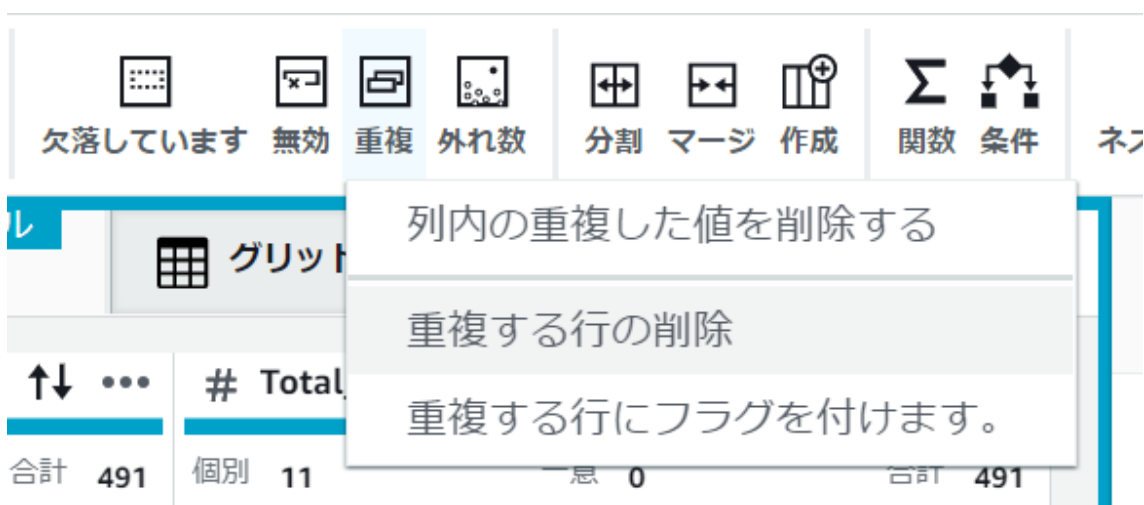
0

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

ステップ 1
データセットを選択

ステップ 2
結合の詳細を指定

結合タイプを選択

☒  内部結合
テーブル A とテーブル B の結合条件を満たすすべての行を選択します。

☐  左結合
テーブル A からすべての行と、テーブル B から結合条件を満たす行を選択します。

88. 以下の画面の通り設定を行います

Join info

Step 1
Select dataset

Step 2
Specify join details

Select join type

☒ Inner join
Select all rows that meet join condition from Table A and Table B.

☐ Left join
Select all rows from Table A and rows that meet join condition from Table B.

☐ Right join
Select all rows from Table B and rows that meet join condition from Table A.

☐ Outer join
Select all rows from Table A and Table B regardless of join condition.

☐ Left excluding join
Select all rows from Table A excluding the rows that meet join condition.

☐ Right excluding join
Select all rows from Table B excluding the rows that meet join condition.

Join keys

Table A (this project)
Sales
Customer_Id

Table B
CleanCustomer
Customer_ID

Add another join key

Column list
Joined table preview

Column list
Select the columns to include in the join

Find columns

Source table	Column name
<input checked="" type="checkbox"/> Table A	#C_Txn_Date
<input checked="" type="checkbox"/> Table A	# Customer_Id
<input checked="" type="checkbox"/> Table A	# Product_Id
<input checked="" type="checkbox"/> Table A	# Quantity
<input checked="" type="checkbox"/> Table A	# Total_Sales
<input type="checkbox"/> Table B	# Customer_ID

☐ Outer join
Select all rows from Table A and Table B regardless of join condition.

☐ Left excluding join
Select all rows from Table A excluding the rows that meet join condition.

☐ Right excluding join
Select all rows from Table B excluding the rows that meet join condition.

☐ Outer excluding join
Select all rows from Table B and Table A excluding the rows that meet join condition.

Find columns

Source table	Column name
<input type="checkbox"/> Table B	# Customer_ID
<input type="checkbox"/> Table B	# Name
<input type="checkbox"/> Table B	# Gender
<input type="checkbox"/> Table B	# DoB
<input type="checkbox"/> Table B	# Address_1
<input type="checkbox"/> Table B	# Address_2
<input type="checkbox"/> Table B	# City
<input checked="" type="checkbox"/> Table B	# Zip
<input type="checkbox"/> Table B	# Country

Cancel Previous Finish

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