[AWS Black Belt Online Seminar]
AWS Glue DataBrew

Junpei Ozono, Solutions Architect
2021.2.17

AWS 公式 Webinar
https://amzn.to/JPWebinar

過去資料
https://amzn.to/JPArchive

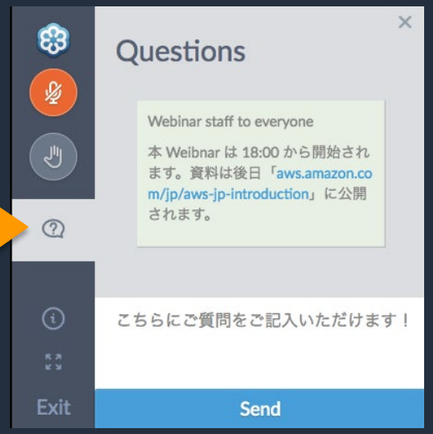
AWS Black Belt Online Seminar とは

「サービス別」「ソリューション別」「業種別」のそれぞれのテーマに分かれて、アマゾン ウェブ サービス ジャパン株式会社が主催するオンラインセミナーシリーズです。

質問を投げることができます！

- 書き込んだ質問は、主催者にしか見えません
- 今後のロードマップに関するご質問は
  お答えできませんのでご了承下さい

① 吹き出しをクリック
② 質問を入力
③ Sendをクリック



Twitter ハッシュタグは以下をご利用ください
#awsblackbelt

自己紹介

大薗 純平 (おおぞの じゅんぺい)

@jostandard

アマゾン ウェブ サービス ジャパン

アナリティクスソリューションアーキテクト

内容についての注意点

- 本資料では2021 年 2 月 17 日現在のサービス内容および価格についてご説明しています。最新の情報はAWS公式ウェブサイト(http://aws.amazon.com)にてご確認ください。

- 資料作成には十分注意しておりますが、資料内の価格とAWS公式ウェブサイト記載の価格に相違があった場合、AWS公式ウェブサイトの価格を優先とさせていただきます。

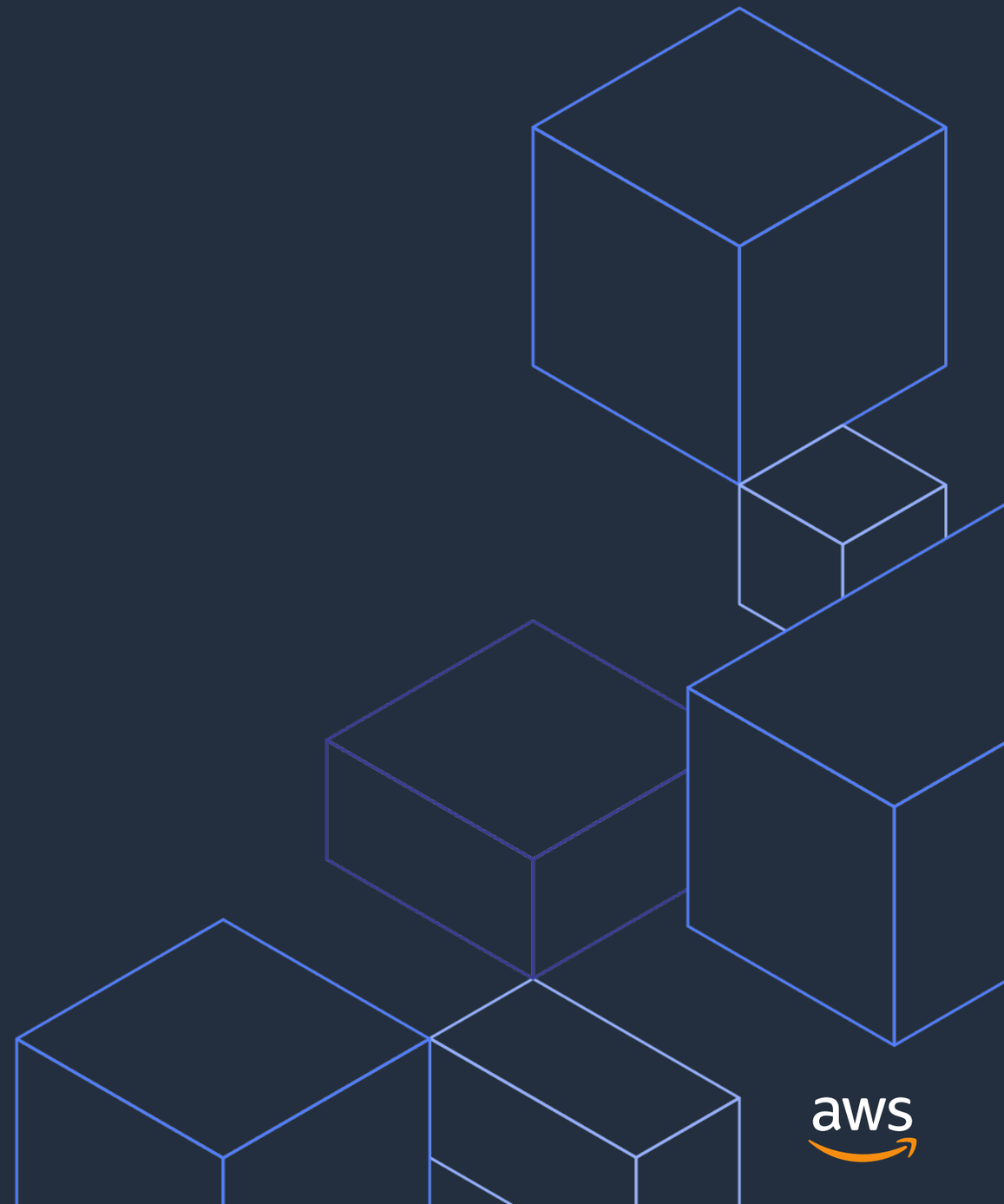
- 価格は税抜表記となっています。日本居住者のお客様には別途消費税をご請求させていただきます。

- AWS does not offer binding price quotes. AWS pricing is publicly available and is subject to change in accordance with the AWS Customer Agreement available at http://aws.amazon.com/agreement/. Any pricing information included in this document is provided only as an estimate of usage charges for AWS services based on certain information that you have provided. Monthly charges will be based on your actual use of AWS services, and may vary from the estimates provided.

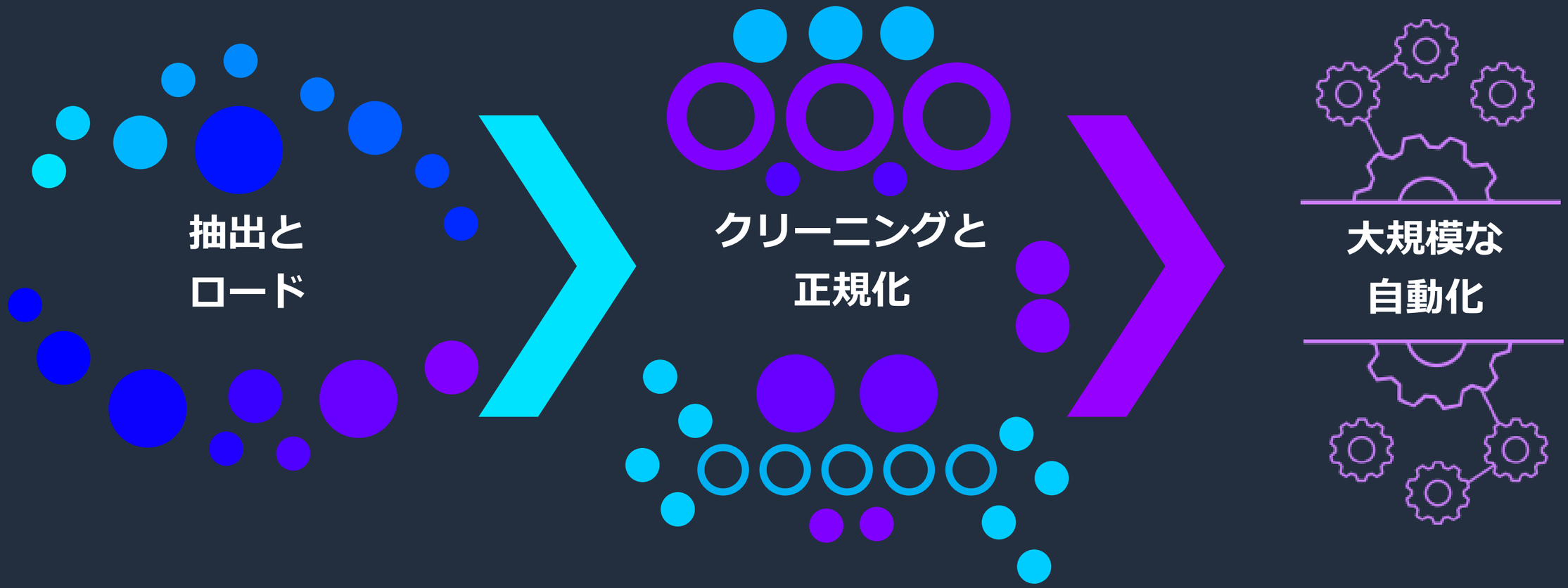
本日のアジェンダ

- データ準備の課題
- AWS Glue DataBrew 概要
- AWS Glue DataBrew の使い方
- AWS Glue DataBrew のユースケース
- AWS Glue DataBrew の料金
- まとめ

データ準備の課題



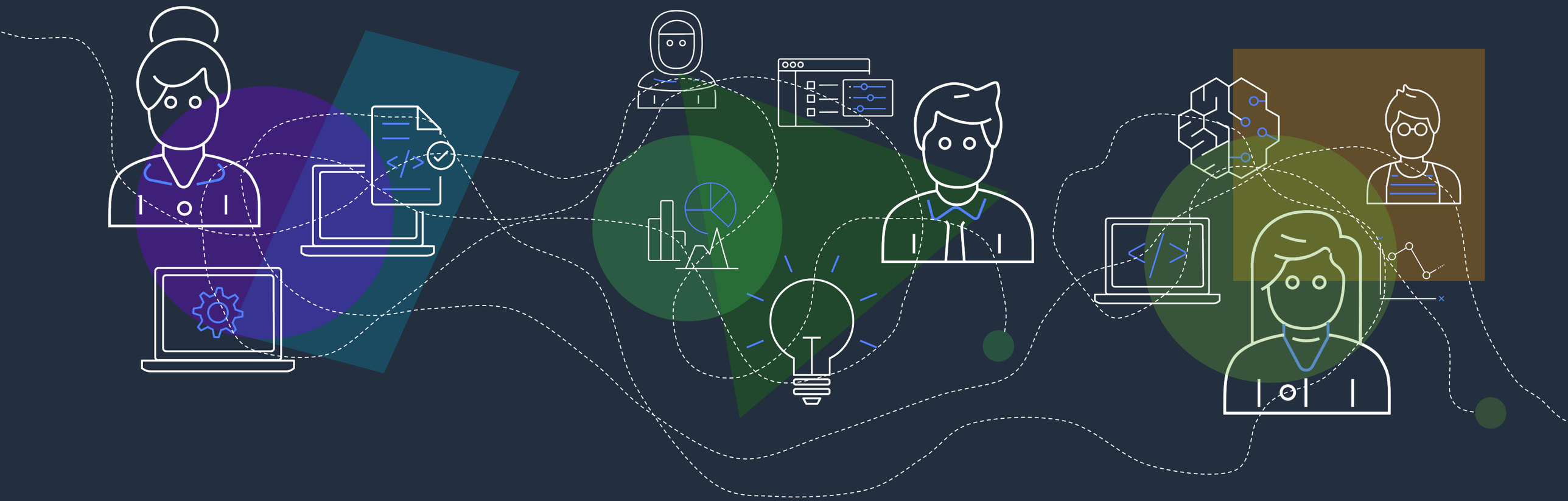
データ準備 (Data prep) には複雑なタスクを伴う



大規模に活用するためには複雑な ETL パイプラインの実装が必要

80% の時間がデータ準備に費やされている

データエンジニア | ETL 開発者 | データアナリスト | データサイエンティスト



利用ユーザーに合った適切なツールが必要

典型的なデータ準備における課題

手動

繰り返しのワークフローを構築・運用するのは困難
スケールさせるには大規模コーディングが必要

大容量データの移動

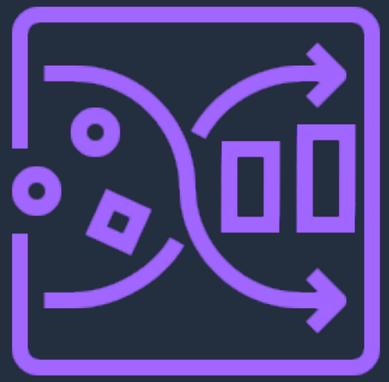
組織間やシステム間での繰り返しのデータ移動が発生

時間がかかる

大規模なデータの抽出、クレンジング、正規化、ロードを
マルチステップで行う必要がある

AWS Glue DataBrew 概要

AWS Glue DataBrew



データのクリーンアップおよび正規化を
最大 80% 高速化するビジュアルデータ準備ツール

データアナリストとデータサイエンティストのためのツール

高度なデータ準備機能をノンコーディングで利用可能

データ品質の理解

データパターンを理解し異常を検出するためにプロファイリングを行いデータの品質を評価

データのクリーンアップと正規化

250 種類以上の組み込みの変換処理から選択し、データの視覚化、クリーニング、正規化を実施

データリネージの視覚化

データソースと変換手順を視覚化してトラッキング

自動化

保存された変換手順を使いまわしたり自動実行する

AWS Glue DataBrew の使い方

AWS Glue DataBrew 用語の紹介

プロジェクト

データセットのクリーンアップや正規化などの変換に関するステップをまとめた**レシピ**を作成するためのワークスペース

データセット

AWS Glue DataBrew が接続する、フィールド (列) を持つデータの集合

レシピ

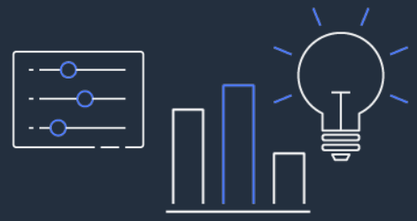
データ変換ステップの一連のセット

ジョブ

データセットに対してレシピを適用して変換処理を行うもの (レシピジョブ)

データセットの統計に関するプロファイルを作成するもの (プロファイルジョブ)

AWS Glue DataBrew の使い方



事前準備 (IAM*)

- IAM ユーザー/グループ
- IAM ロール
- IAM ポリシー

データ変換処理の作成

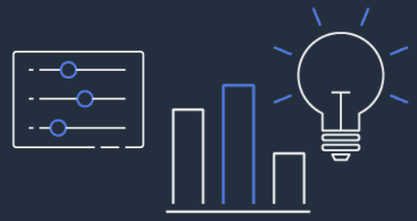
- プロジェクトの作成
- データセットへの接続
- レシピの作成

ジョブの実行

- レシピジョブ
- プロファイルジョブ

* AWS Identity and Access Management

AWS Glue DataBrew の使い方



事前準備 (IAM*)

- IAM ユーザー/グループ
- IAM ロール
- IAM ポリシー

データ変換処理の作成

- プロジェクトの作成
- データセットへの接続
- レシピの作成

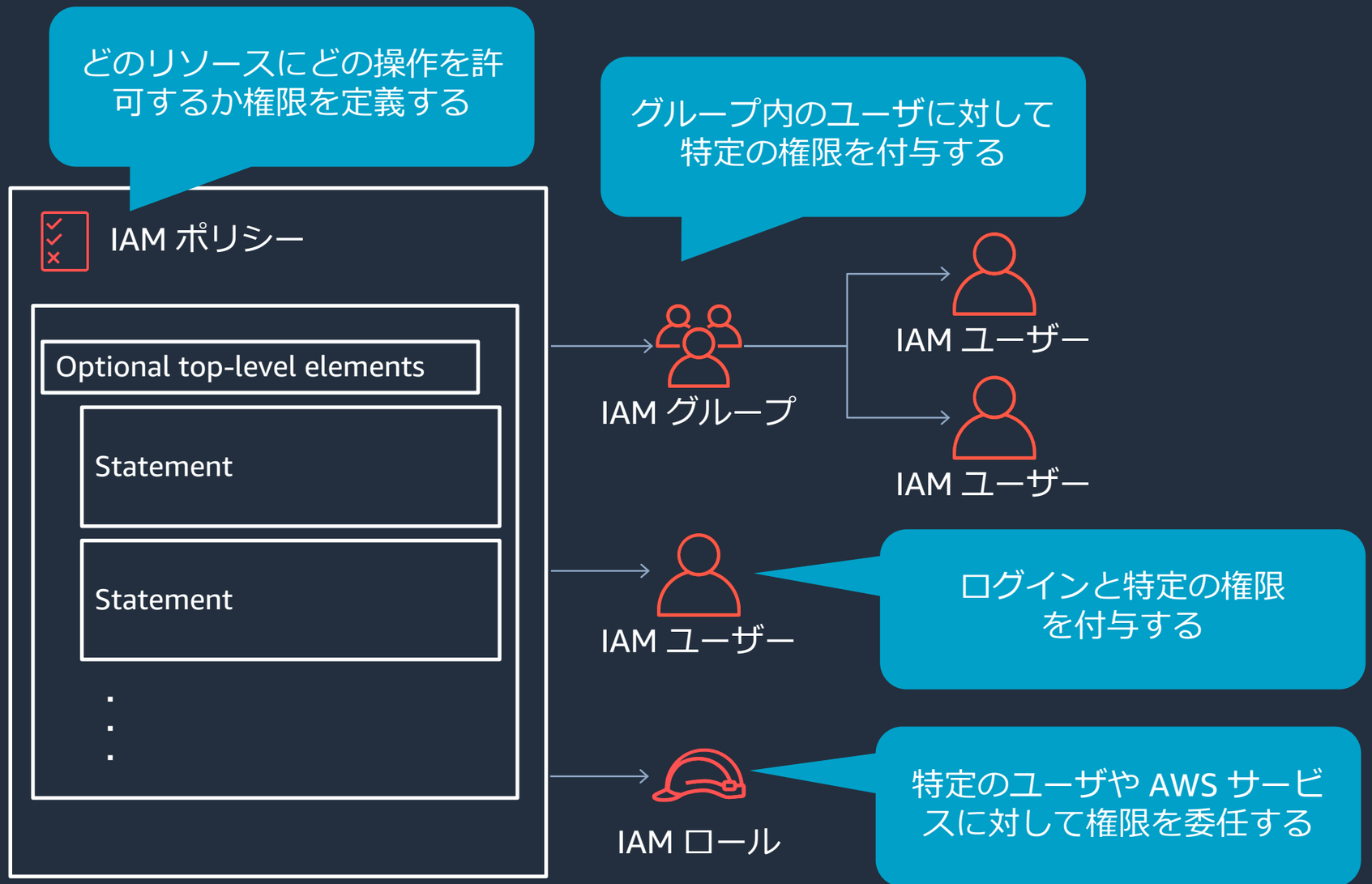
ジョブの実行

- レシピジョブ
- プロファイルジョブ

* AWS Identity and Access Management

IAM おさらい

事前準備 | データ変換処理の作成 | ジョブの実行



https://www.slideshare.net/AmazonWebServicesJapan/20190129-aws-black-belt-online-seminar-aws-identity-and-access-management-iam-part1
https://www.slideshare.net/AmazonWebServicesJapan/20190130-aws-black-belt-online-seminar-aws-identity-and-access-management-aws-iam-part2

事前準備として必要なもの

1. AWS Glue DataBrew の利用者が
   認証に使用するための IAM ユーザー/グループ
   および IAM ユーザー/グループにアタッチする IAM ポリシー

2. AWS Glue DataBrew サービス自体が
   他の AWS サービスにアクセスする際に使用する IAM ロール
   および IAM ロールにアタッチする IAM ポリシー

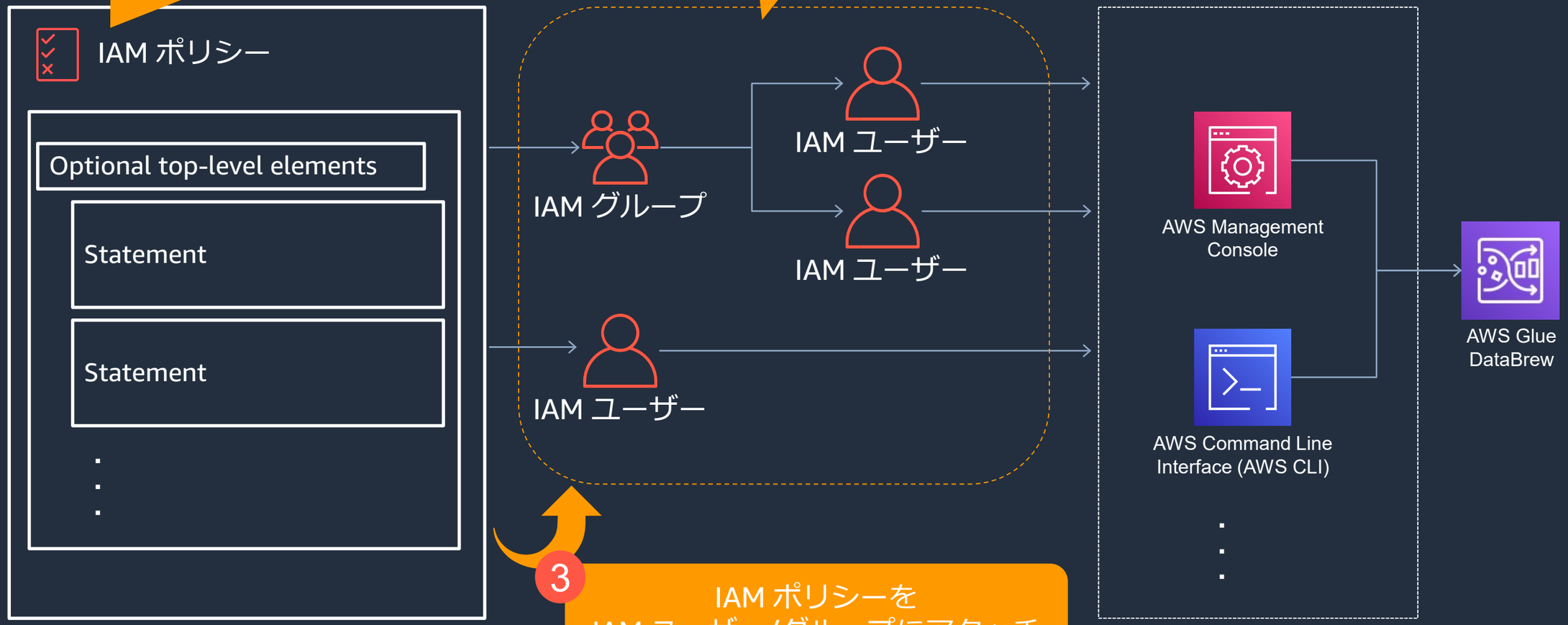
https://docs.aws.amazon.com/ja_jp/databrew/latest/dg/setting-up.html

IAM ユーザー/グループ準備

事前準備 | データ変換処理の作成 | ジョブの実行

1 IAM ユーザー/グループにアタッチする必要な権限が揃った IAM ポリシーを作成

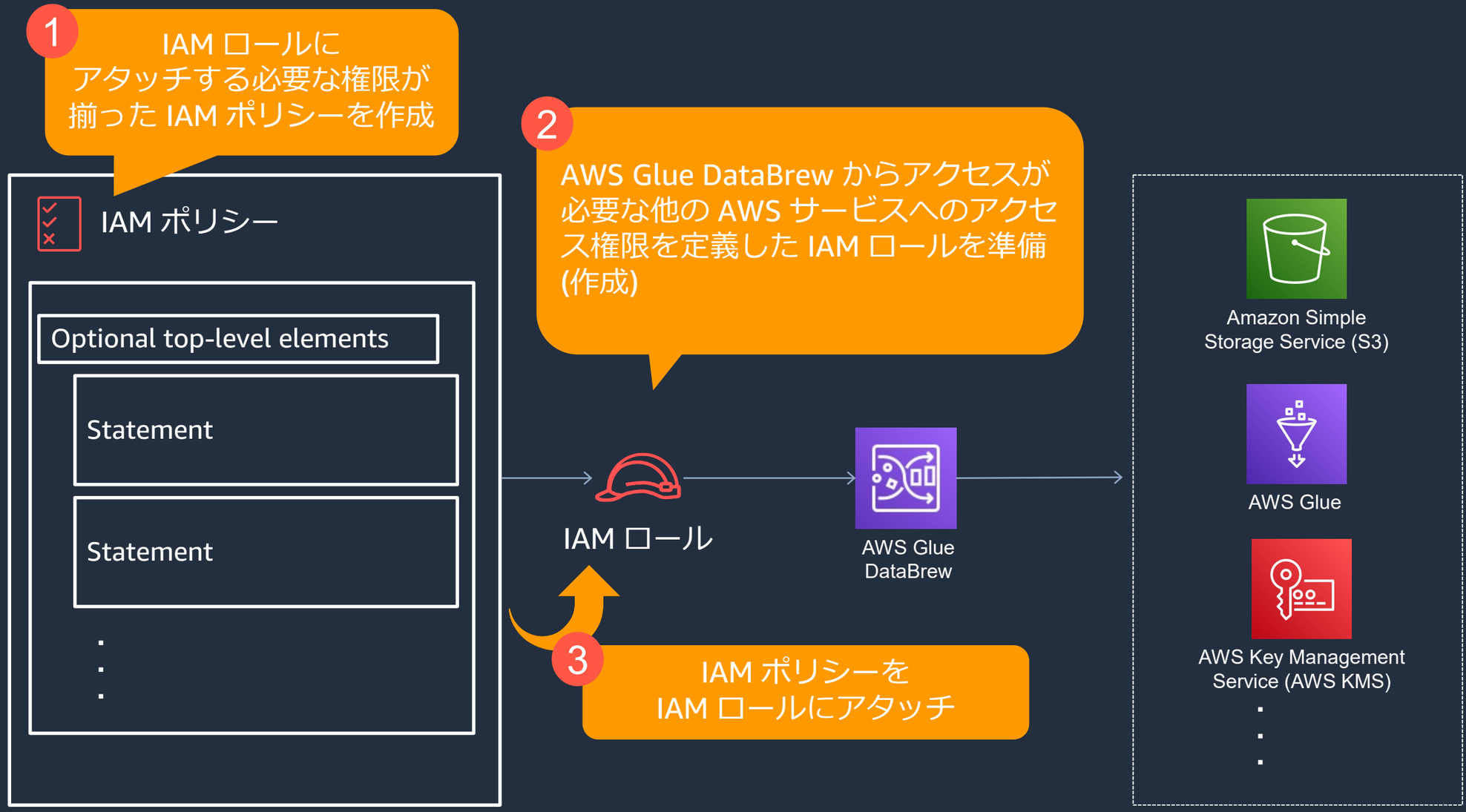
2 AWS マネージメントコンソールやコマンドラインインターフェース (CLI) から AWS Glue DataBrew に接続するための IAM ユーザー/グループを準備(作成)



3 IAM ポリシーを IAM ユーザー/グループにアタッチ

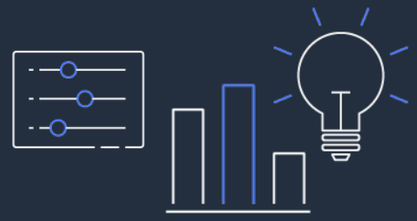
https://docs.aws.amazon.com/ja_jp/databrew/latest/dg/setting-up.html

IAM ロール準備



https://docs.aws.amazon.com/ja_jp/databrew/latest/dg/setting-up.html

AWS Glue DataBrew の使い方



事前準備 (IAM*)

- IAM ユーザー/グループ
- IAM ロール
- IAM ポリシー

データ変換処理の作成

- プロジェクトの作成
- データセットへの接続
- レシピの作成

ジョブの実行

- レシピジョブ
- プロファイルジョブ

* AWS Identity and Access Management

プロジェクトの作成

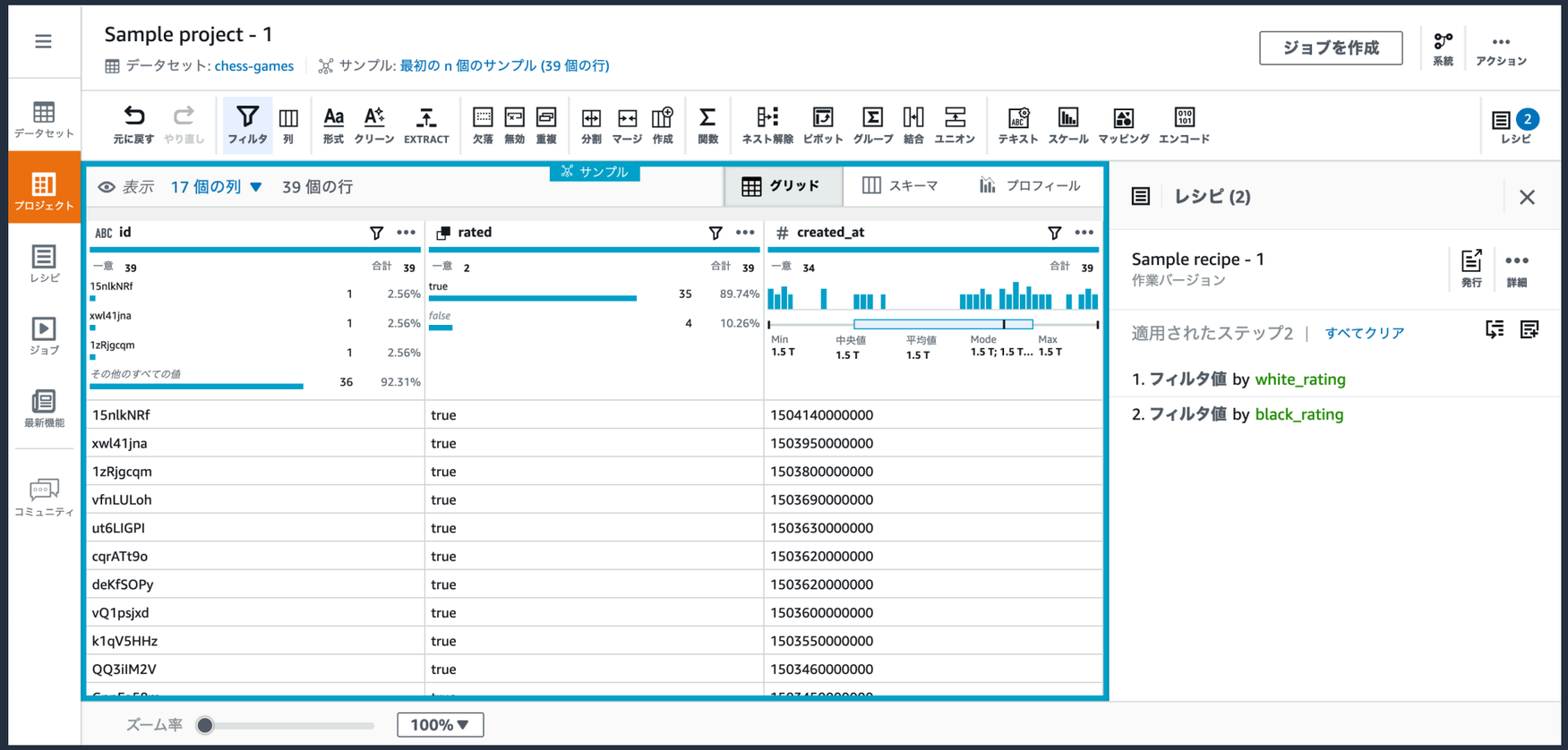
事前準備 | データ変換処理の作成 | ジョブの実行

マネージメントコンソールに AWS Glue DataBrew の操作権限を持った IAM ユーザーでアクセスしてプロジェクトを作成する

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

プロジェクトとは

プロジェクトは、特定のデータセットに対する変換ステップを定義する"レシピ"を作成するためのワークスペース



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

プロジェクト作成手順①

レシピとデータセット

事前準備 > データ変換処理の作成 > ジョブの実行

レシピ

- 新しいレシピを作成
- 既存のレシピを編集
- レシピからステップをインポート

データセット

- マイデータセット
- サンプルファイル
- 新しいデータセット

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

プロジェクト作成手順②

データセットの選択

事前準備 | データ変換処理の作成 | ジョブの実行

加工/変換したいデータを以下の中から選択

- ローカルファイル
- Amazon S3 上のファイル
- AWS Glue データカタログ
- AWS Data Exchange

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

プロジェクト作成手順③
行サンプリング/アクセス許可

プロジェクト内で操作するデータセットの行数をサンプリング可能

データセットに対するアクセス許可を指定

- 新しい IAM ロールを作成
- 既存の IAM ロールを選択

「事前準備」で IAM ロールを作成した場合はこちらを選択

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

サポートされるデータセット

インプットファイルのサポートフォーマットと拡張子

フォーマット	拡張子 (非圧縮) *1	拡張子 (圧縮) *1
CSV *2	.csv	.csv.gz, .csv.snappy, .csv.lz4, .csv.bz2, .csv.deflate
TSV *2	.tsv	.tsv.gz, .tsv.snappy, .tsv.lz4, .tsv.bz2, .tsv.deflate
Microsoft Excel ワークブック	.xlsx	非サポート
JSON	.json	.json.gz, .json.snappy, .json.lz4, .json.bz2, .json.deflate
JSON lines	.jsonl	.jsonl.gz, .jsonl.snappy, .jsonl.lz4, .jsonl.bz2, .jsonl.deflate
Apache Parquet	.parquet	.parquet.gz, .gz.parquet, .parquet.snappy, .snappy.parquet, .parquet.lz4, .lz4.parquet

*1 DataBrew は拡張子でファイルフォーマットを判断するため、必ず上記拡張子を使用する

*2 区切り文字は Comma (,), Colon (:), Semi-colon (;), Pipe (|), Tab (¥t), Caret (^), Space に対応

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

Amazon S3 上のデータ

Amazon S3 上の特定ファイル/フォルダを示す「S3 パス」または
正規表現を用いた「パラメータ化された S3 パス」を指定可能

例

- ある特定のファイルを指定
  s3://bucket-name/inventory-data.csv
- ある特定のフォルダ配下にあるすべてのファイルを指定
  s3://*bucket-name/folder-name/*
- “2021” を名称に含むフォルダ配下にあるすべてのファイルを指定
  s3://bucket-name/<.>2021<.*>/*

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

他 AWS サービス上のデータ

事前準備 | データ変換処理の作成 | ジョブの実行

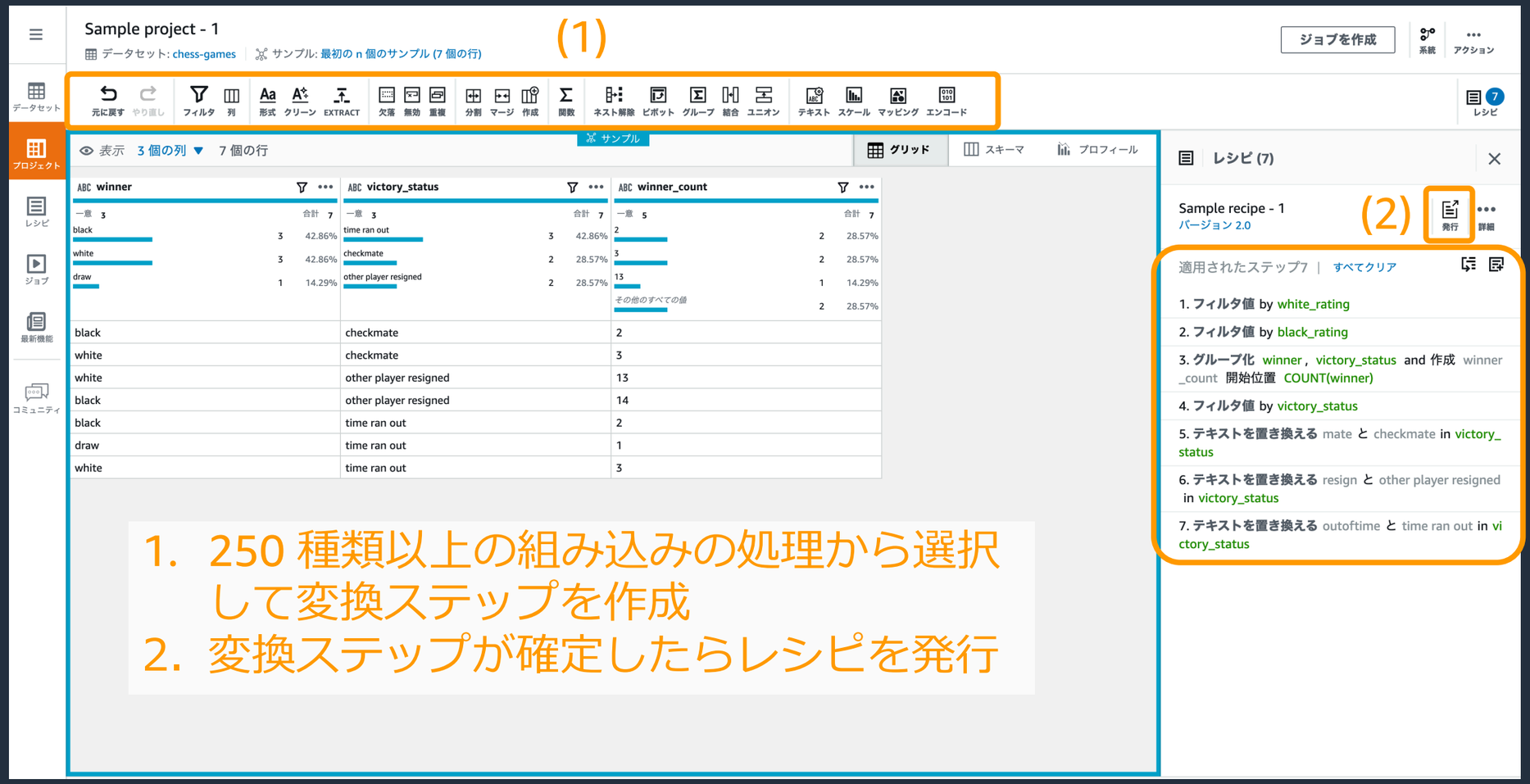
AWS Glue データカタログ経由で以下サービス上のデータと接続可能

- Amazon Redshift
- Amazon Aurora MySQL
- Amazon Aurora PostgreSQL
- Amazon RDS for MySQL
- Amazon RDS for PostgreSQL

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

レシピの作成

プロジェクト内でインタラクティブに変換イメージを確認しながら、
データセットに対する変換ステップのコレクションであるレシピを作成

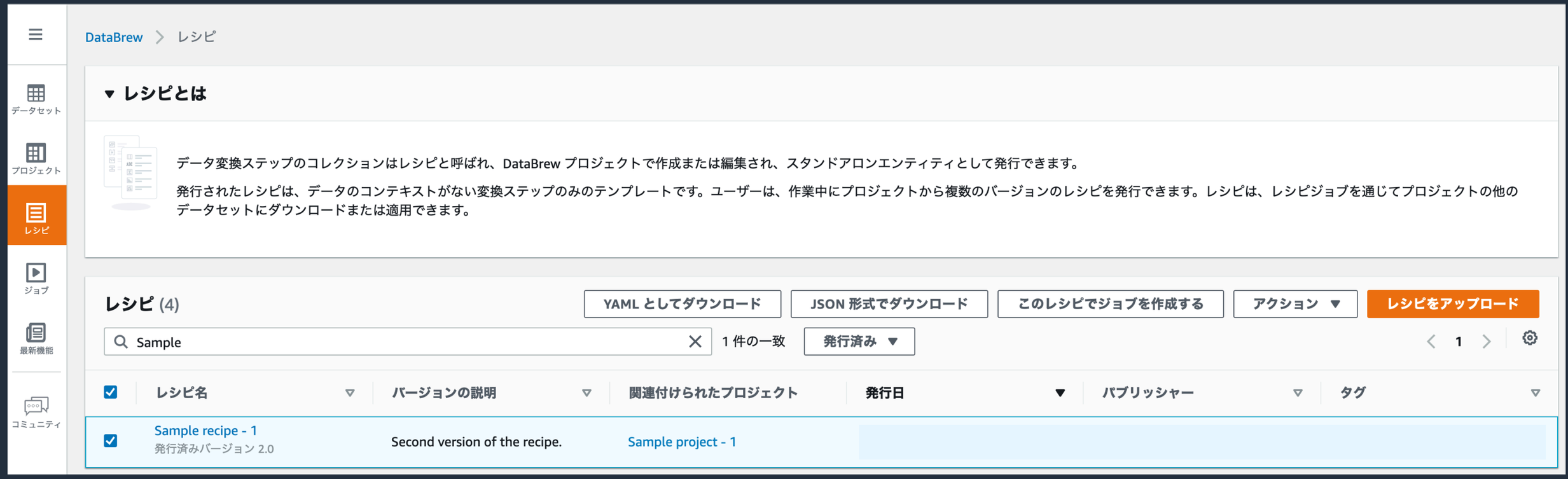


https://docs.aws.amazon.com/databrew/latest/dg/recipes.html

レシピの管理

事前準備 | データ変換処理の作成 | ジョブの実行

- 作成したレシピは編集・削除でき、バージョン管理も可能
- レシピは YAML/JSON でのダウンロード, JSON のアップロードも可能

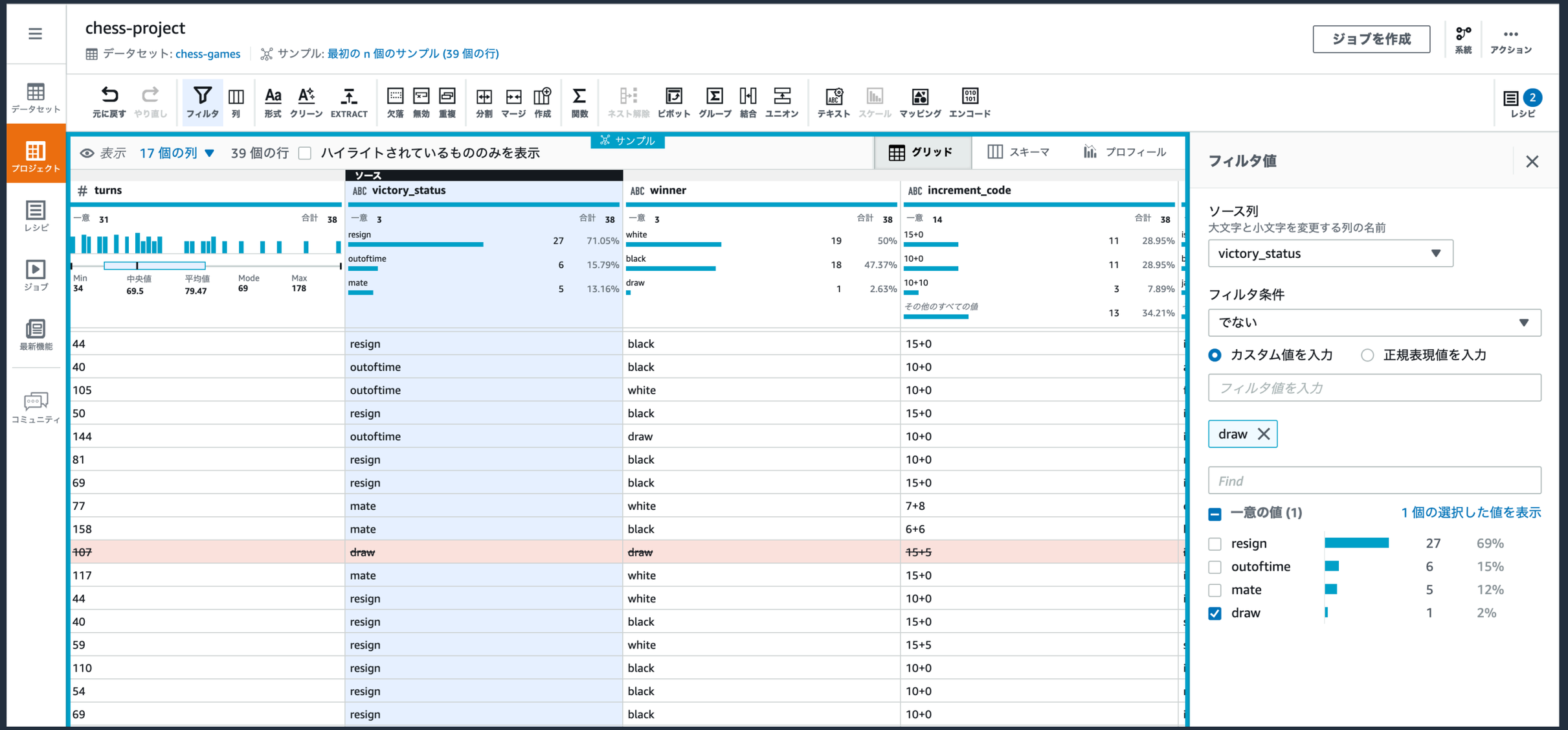


https://docs.aws.amazon.com/databrew/latest/dg/recipes.html

代表的な変換処理

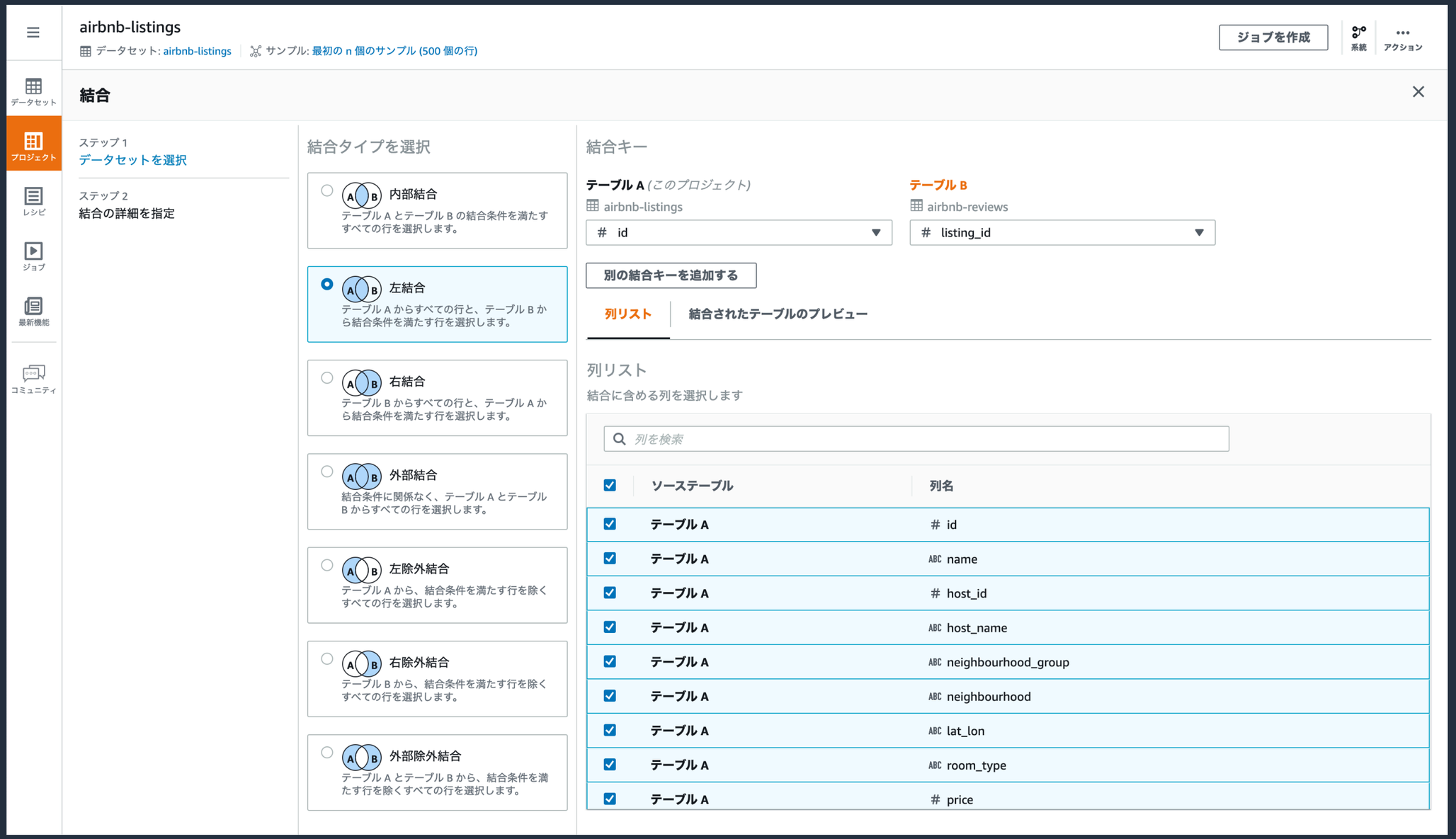
データのフィルタリング

事前準備 | データ変換処理の作成 | ジョブの実行



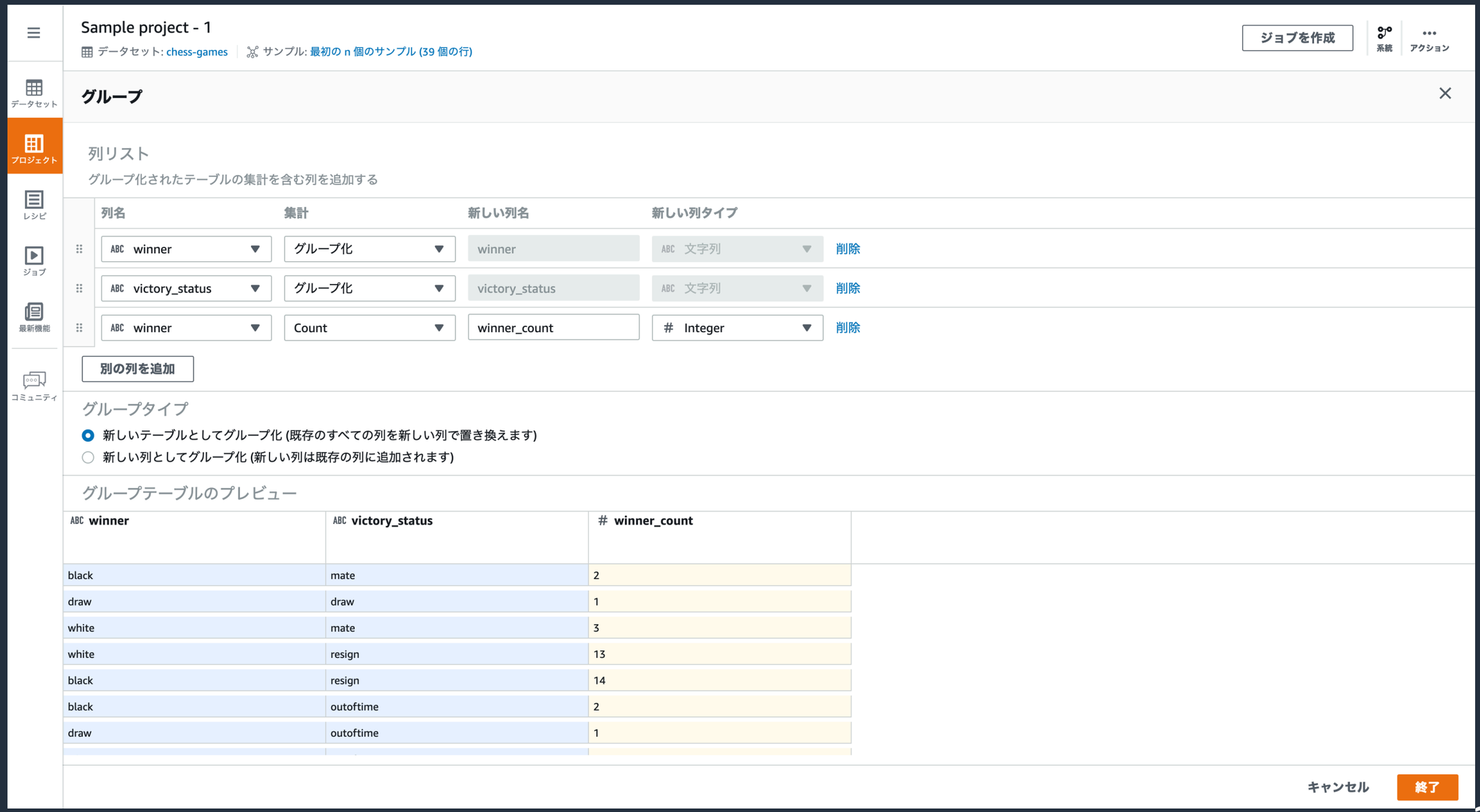
データの結合

事前準備 | データ変換処理の作成 | ジョブの実行



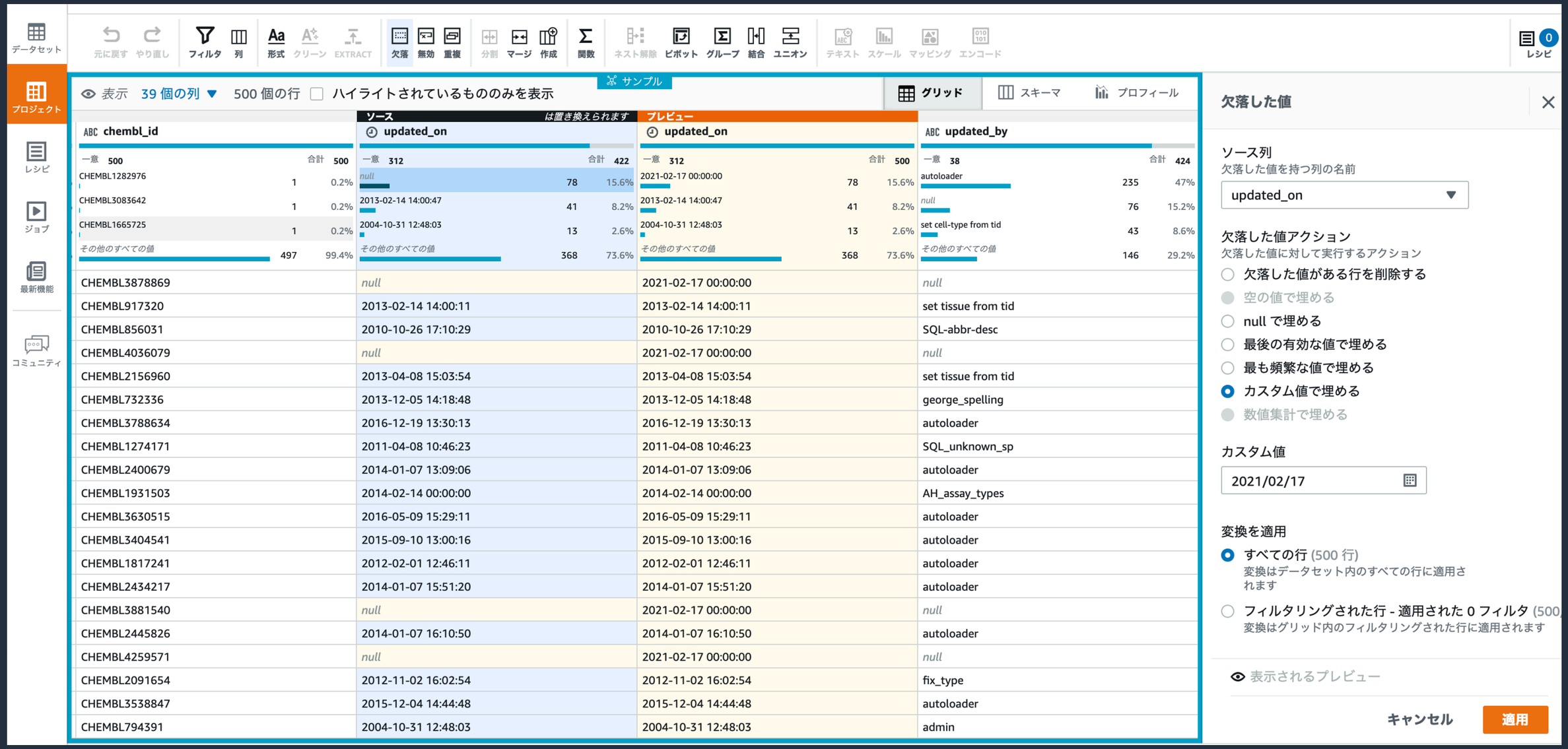
データの集計

事前準備 | データ変換処理の作成 | ジョブの実行



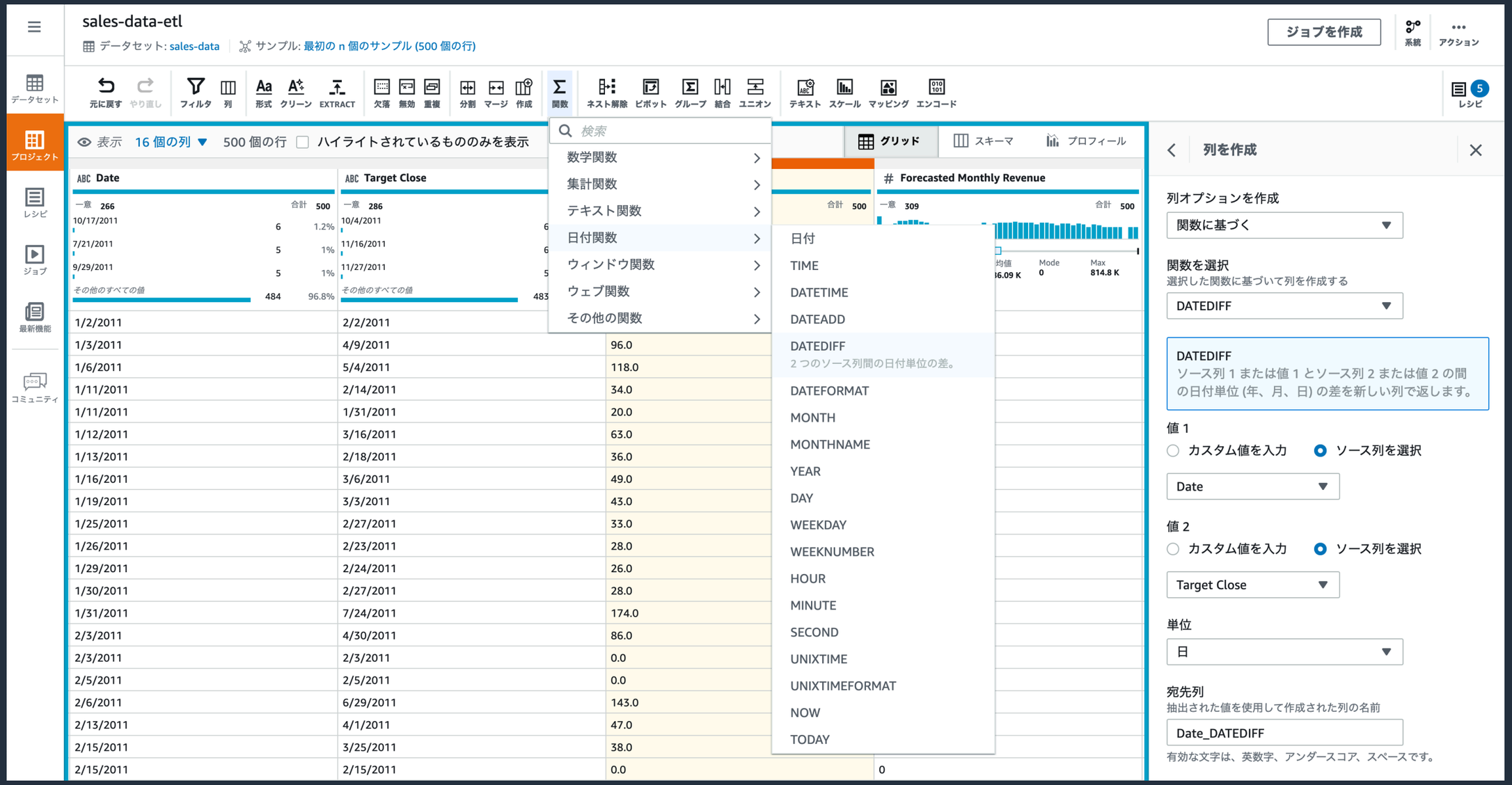
欠損値の補完

事前準備 | データ変換処理の作成 | ジョブの実行



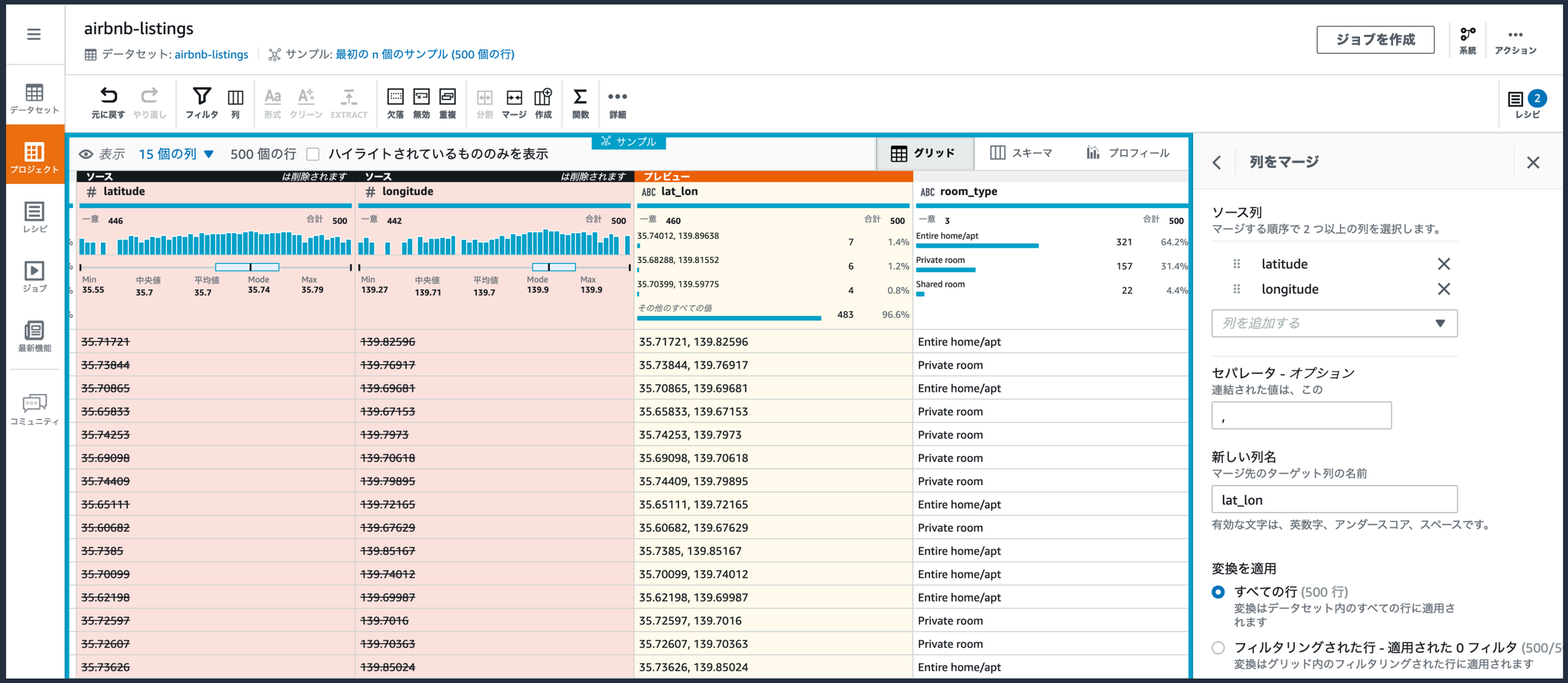
関数を使った新たな列の作成

事前準備 | データ変換処理の作成 | ジョブの実行



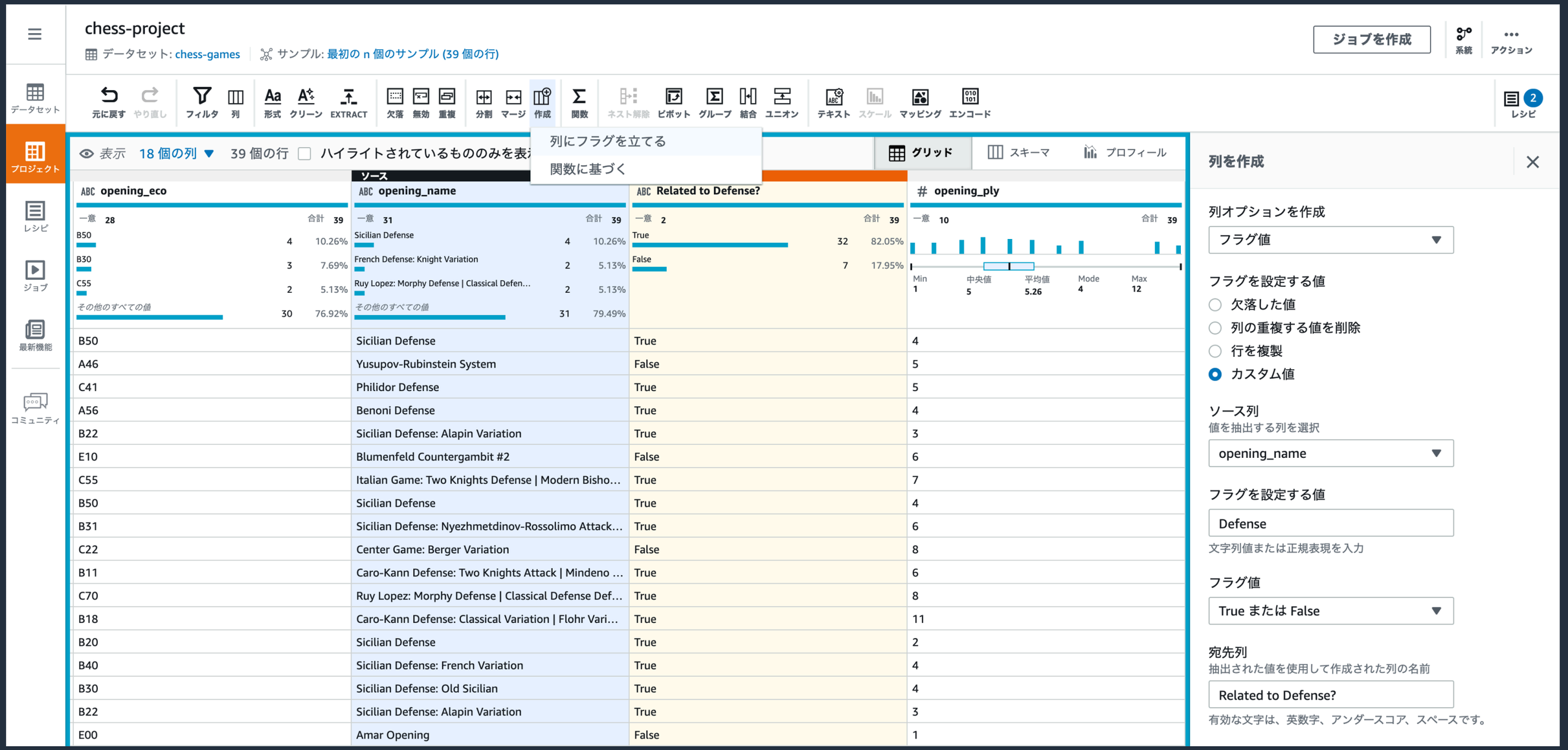
複数列の統合

事前準備 | データ変換処理の作成 | ジョブの実行



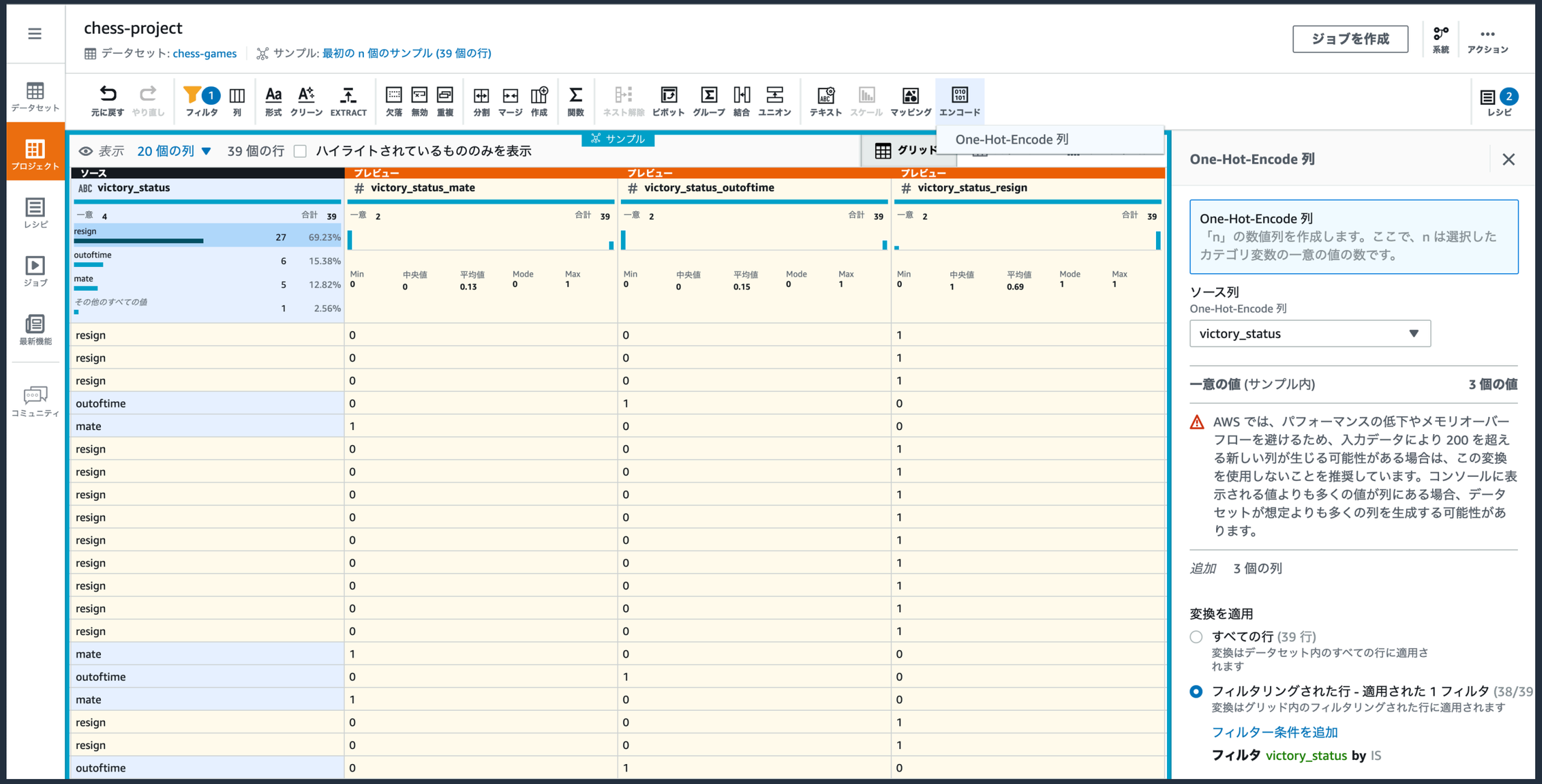
フラグ値の作成

事前準備 | データ変換処理の作成 | ジョブの実行



One-hot エンコーディング

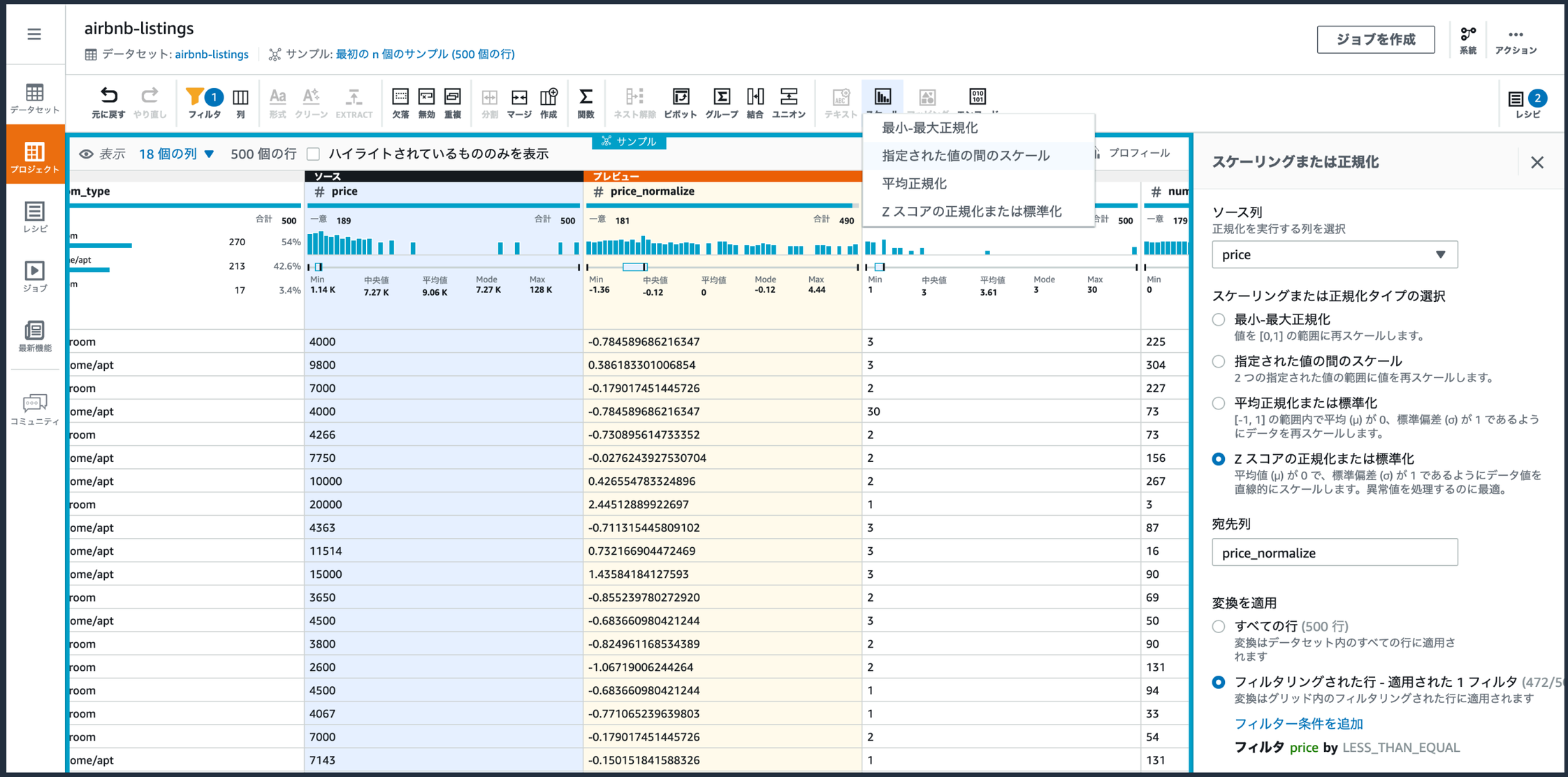
事前準備 | データ変換処理の作成 | ジョブの実行



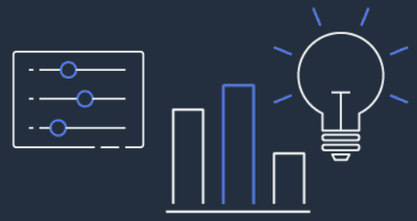
victory_status	victory_status_mate	victory_status_outoftime	victory_status_resign
resign	0	0	1
resign	0	0	1
resign	0	0	1
outoftime	0	1	0
mate	1	0	0
resign	0	0	1
resign	0	0	1
resign	0	0	1
resign	0	0	1
resign	0	0	1
resign	0	0	1
resign	0	0	1
resign	0	0	1
resign	0	0	1
resign	0	0	1
mate	1	0	0
outoftime	0	1	0
mate	1	0	0
resign	0	0	1
resign	0	0	1
outoftime	0	1	0

数値データの正規化

事前準備 | データ変換処理の作成 | ジョブの実行



AWS Glue DataBrew の使い方



事前準備 (IAM*)

- IAM ユーザー/グループ
- IAM ロール
- IAM ポリシー

データ変換処理の作成

- プロジェクトの作成
- データセットへの接続
- レシピの作成

ジョブの実行

- レシピジョブ
- プロファイルジョブ

* AWS Identity and Access Management

ジョブ

AWS Glue DataBrew には 2 種類のジョブがある

- **レシピジョブ**
  - データセットに対してレシピを適用して変換処理を行うもの
- **プロファイルジョブ**
  - データセットの統計に関するプロファイルを作成するもの

ジョブを実行しても対象のデータセットを書き換えることはせず、
指定した Amazon S3 上に結果を書き出す

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

レシピジョブ

レシピジョブの作成手順①

データセットに対してレシピを適用して変換処理

事前準備 > データ変換処理の作成 > ジョブの実行

ジョブタイプとして
「レシピジョブを作成」を選択

ジョブを実行する対象の
- データセット
- プロジェクト
- レシピ

を選択

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

レシピジョブの作成手順②

データセットに対してレシピを適用して変換処理

事前準備 > データ変換処理の作成 > ジョブの実行

ジョブの出力結果のアウトプット
ファイルフォーマットや書き出し
先の S3 プレフィックスを指定

パーティションの設定や
ファイルの上書きオプション、
暗号化設定も指定することが可能

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

サポートされるデータセット

アウトプットファイルのサポートフォーマットと拡張子

フォーマット	拡張子 (非圧縮)	拡張子 (圧縮)
CSV *1	.csv	.csv.snappy, .csv.gz, .csv.lz4, csv.bz2, .csv.deflate
Apache Parquet	非サポート	.parquet.snappy, .parquet.gz, .parquet.lz4, .parquet.lzo
AWS Glue Parquet	非サポート	.glue.parquet.snappy
Apache Avro	.avro	.avro.snappy, .avro.gz, .avro.lz4, .avro.bz2, .avro.deflate
Apache Orc	非サポート	.orc.snappy, .orc.lzo, .orc.zlib
XML	.xml	.xml.snappy, .xml.gz, .xml.lz4, .xml.bz2, .xml.deflate
JSON (JSON Lines format only)	.json	.json.snappy, .json.gz, .json.lz4, json.bz2, .json.deflate

*1 区切り文字は Comma (,), Colon (:), Semi-colon (;), Pipe (|), Tab (¥t), Caret (^), Space に対応

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

レシピジョブの作成手順②

データセットに対してレシピを適用して変換処理

事前準備 > データ変換処理の作成 > ジョブの実行

ジョブに割り当てるノード数やタイムアウト、リトライ回数を指定し、パフォーマンスを調整することが可能

- ノード数はデフォルト 5, 最大 149
- 1 ノード 4 vCPUs, 16GB メモリ

データセットに対するアクセス許可を指定

- 新しい IAM ロールを作成
- 既存の IAM ロールを選択

プロジェクト作成時と同じものを選択

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

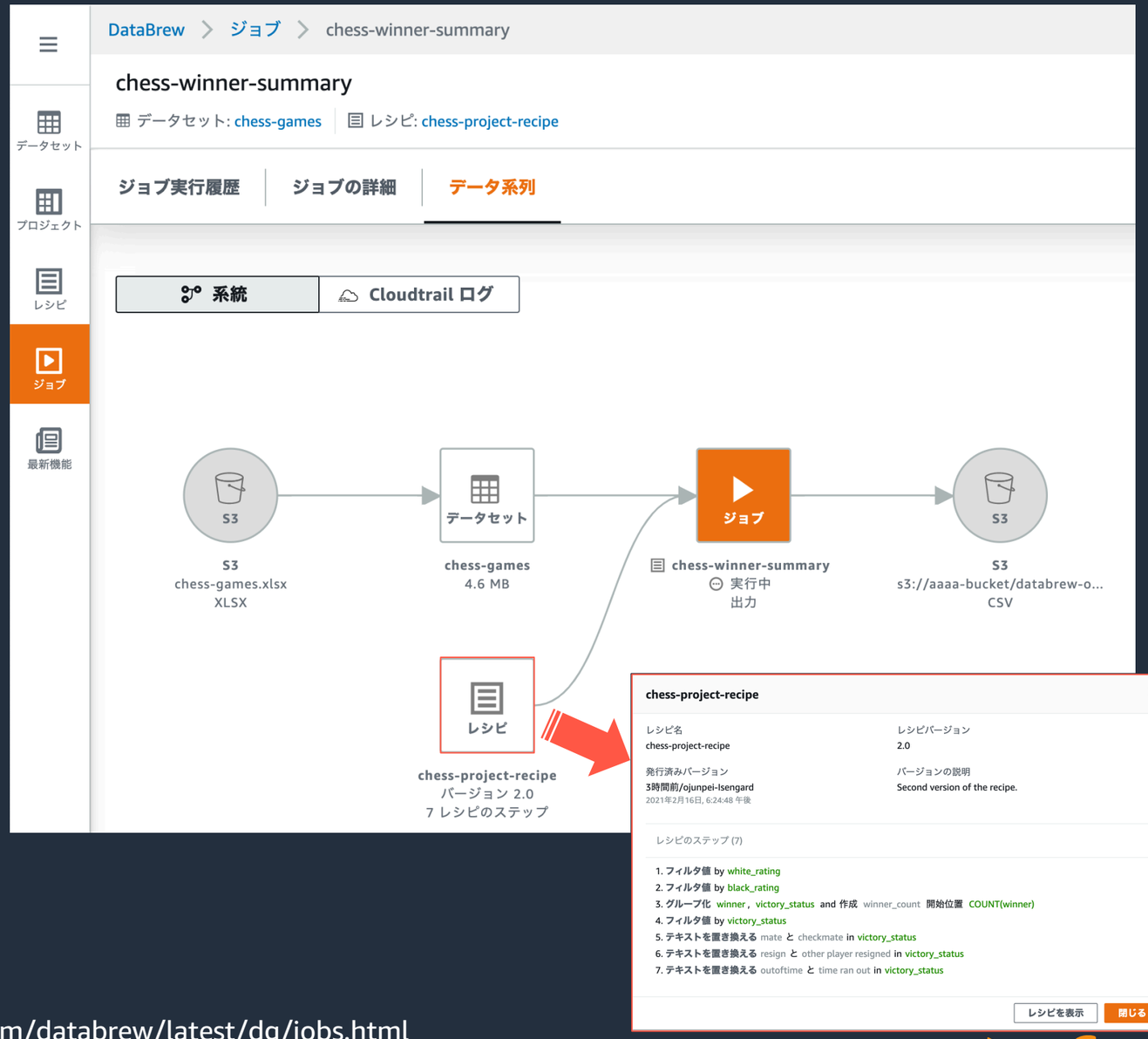
データリネージ

事前準備 > データ変換処理の作成 > ジョブの実行

データのインプットから
アウトプットまでの流れを可視化

ジョブの実行状況も確認可能

各アイコンをクリックすることで
詳細情報の確認も可能



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

プロファイルジョブ

プロファイルジョブの作成手順①

事前準備 > データ変換処理の作成 > ジョブの実行

データセットの統計に関するプロファイルを作成

ジョブタイプとして
「プロファイルジョブを作成する」
を選択

ジョブを実行する対象の
データセットを選択

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

プロファイルジョブの作成手順②

事前準備 > データ変換処理の作成 > ジョブの実行

データセットの統計に関するプロファイルを作成

データセットの
サンプリング件数を指定

- 全件
- 件数指定

ジョブの出力結果の
アウトプットファイルの
書き出し先の S3 プレフィックス
を指定

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

プロファイルジョブの作成手順③

事前準備 > データ変換処理の作成 > ジョブの実行

データセットの統計に関するプロファイルを作成

ジョブに割り当てるノード数やタイムアウト、リトライ回数を指定し、パフォーマンスを調整することが可能

- ノード数はデフォルト 5, 最大 149
- 1 ノード 4 vCPUs, 16GB メモリ

データセットに対するアクセス許可を指定

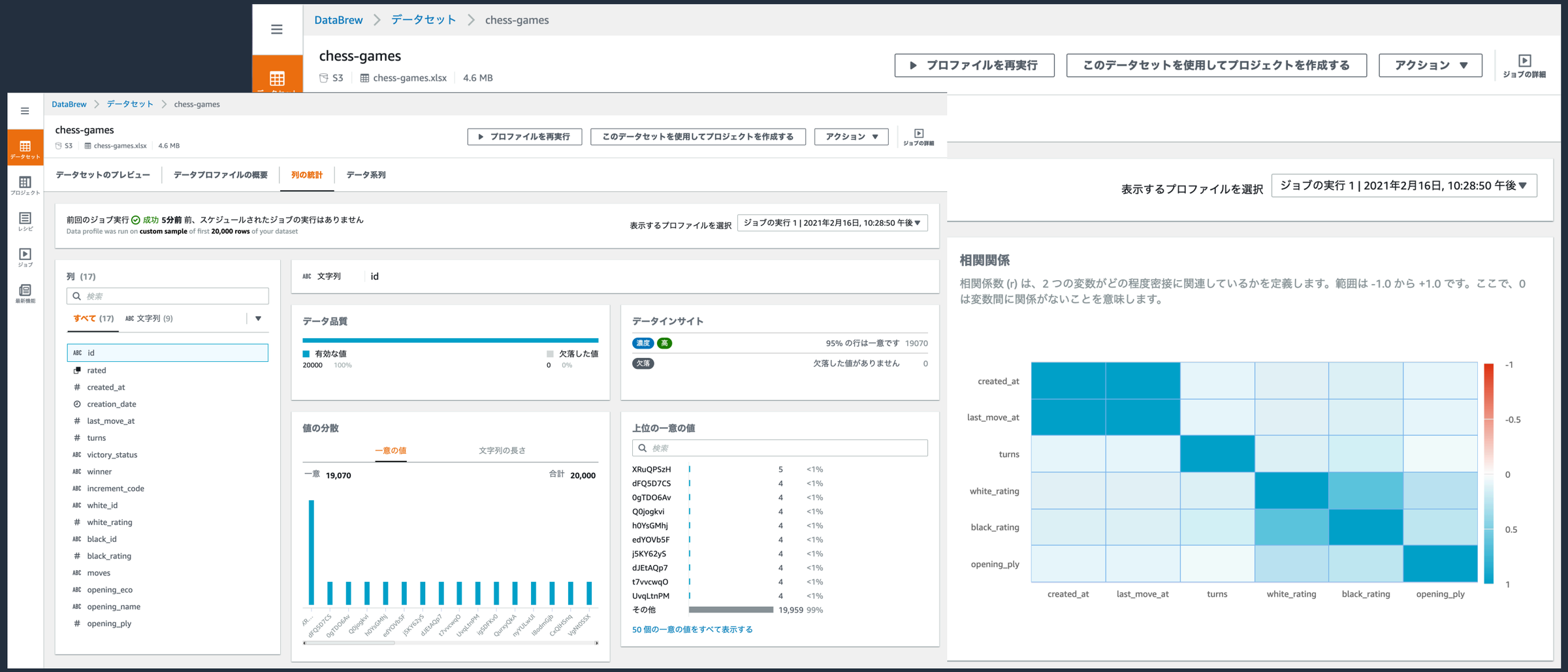
- 新しい IAM ロールを作成
- 既存の IAM ロールを選択

プロジェクト作成時と同じものを選択

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

データプロファイルの生成

事前準備 > データ変換処理の作成 > ジョブの実行



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

AWS Glue DataBrew のユースケース

AWS Glue DataBrew のユースケース

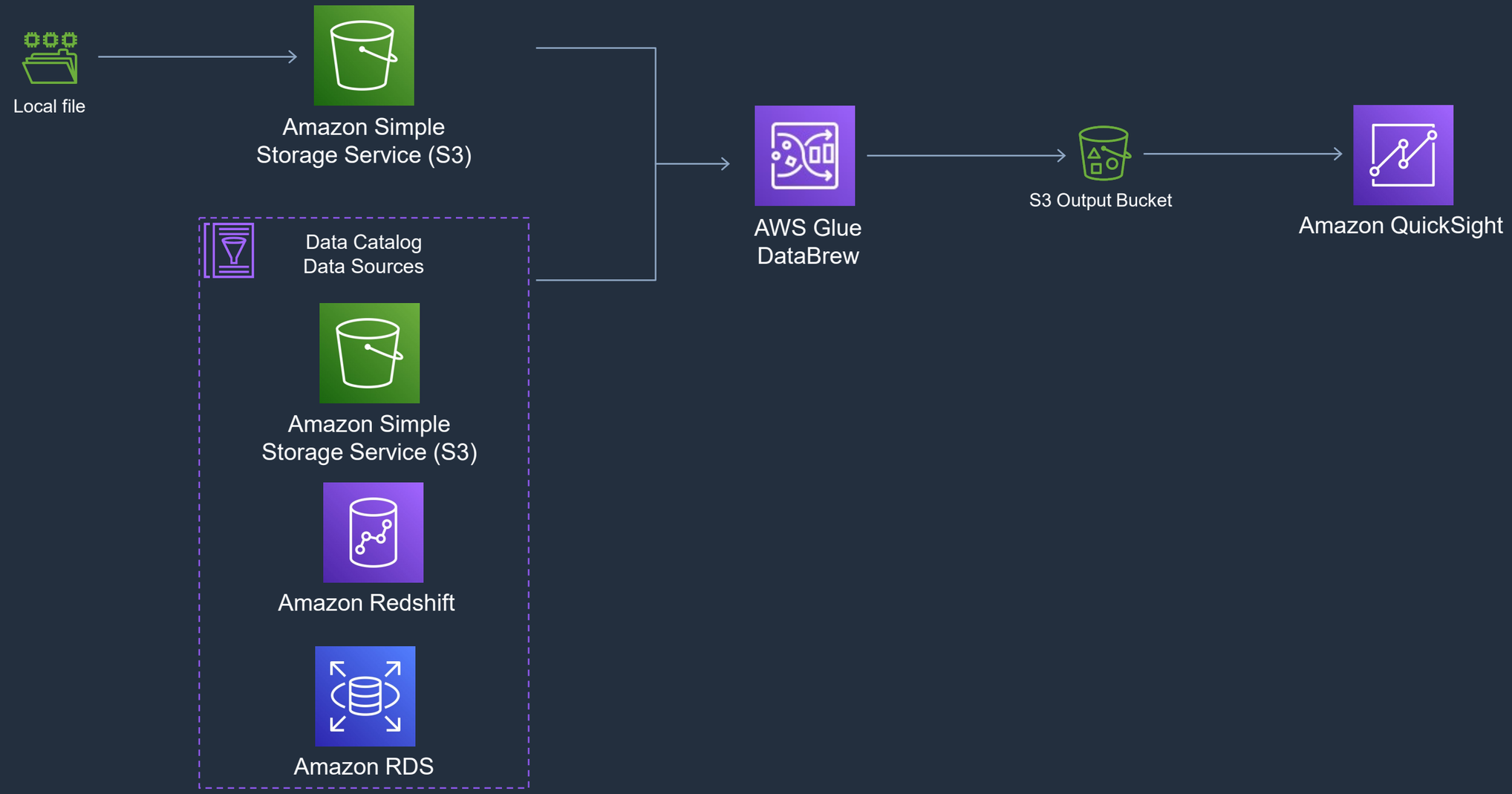
1. Amazon Redshift や Amazon RDS など他の AWS サービス内のデータをアドホックに探索して整備し、BI レポーティング環境を構築する

2. 定常的に生成するデータのプロファイルチェックを自動化し通知する

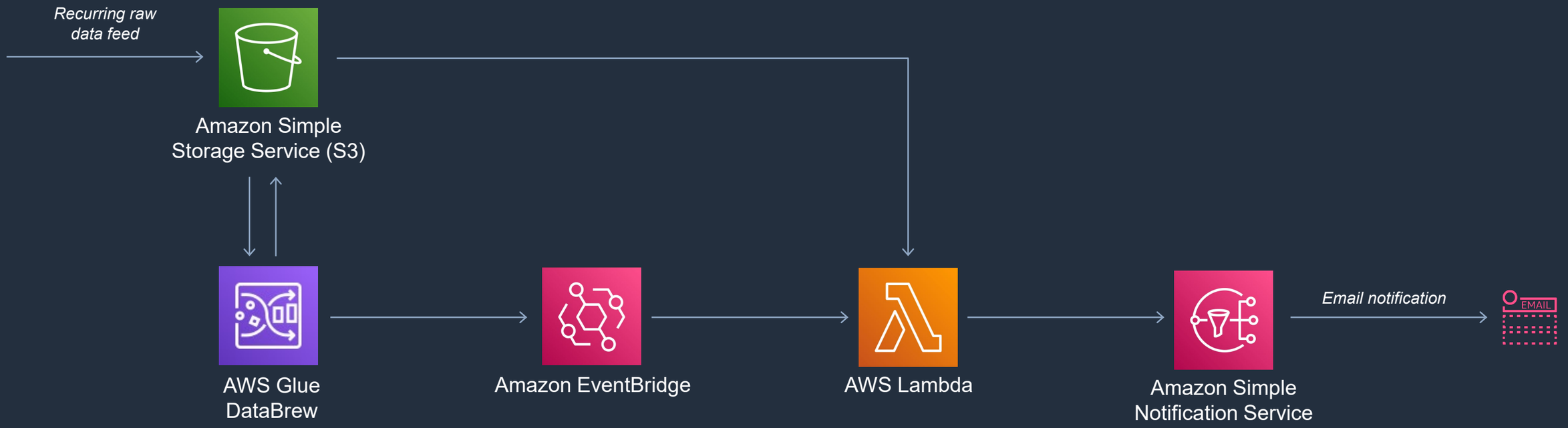
3. 機械学習モデルを構築するためのデータを準備する

4. Amazon Athena にクエリして Amazon QuickSight で可視化するためのデータを準備するパイプラインをコーディングレスで構築する

1. BI レポーティングのためのアドホックデータ分析

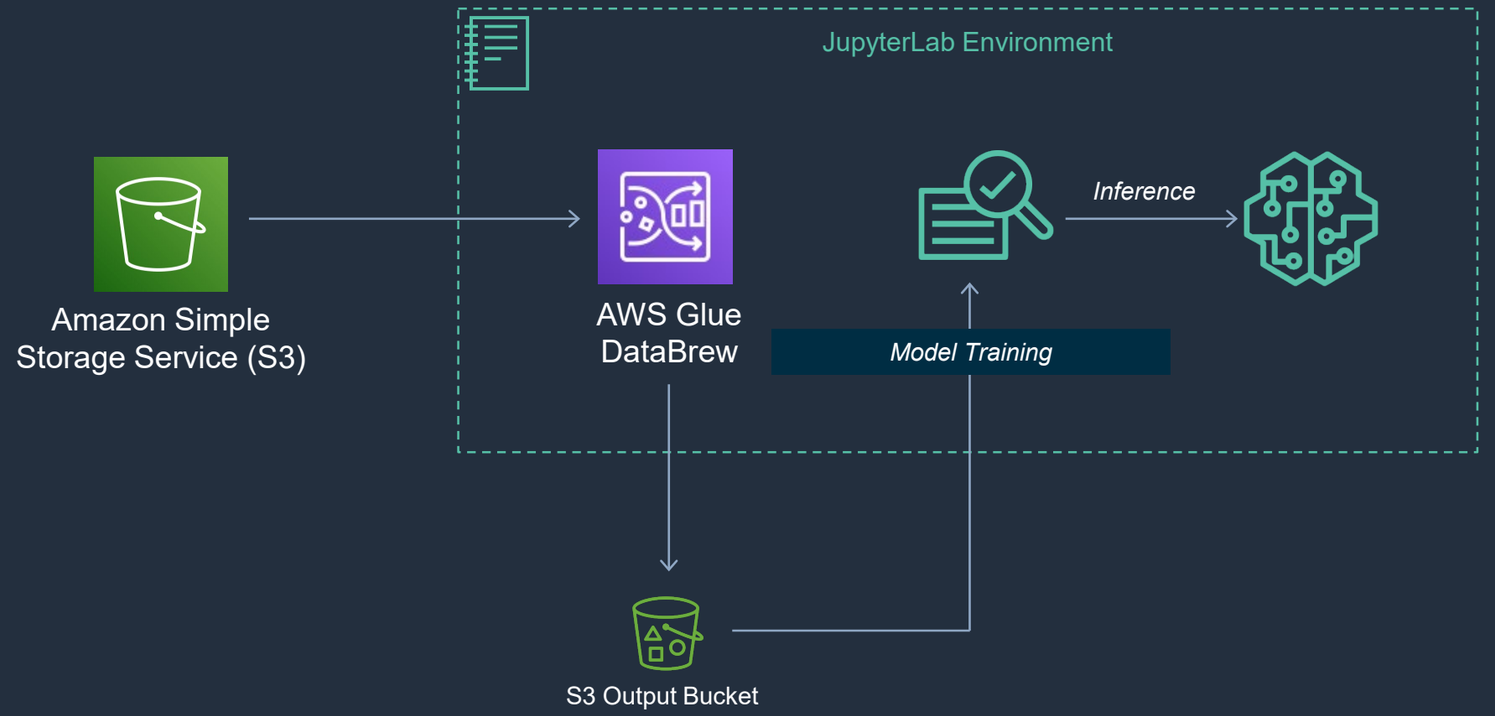


2. AWS Lambda でデータ品質ルールを設定



https://aws.amazon.com/jp/blogs/big-data/setting-up-automated-data-quality-workflows-and-alerts-using-aws-glue-databrew-and-aws-lambda/

3. 機械学習のためのデータ前処理

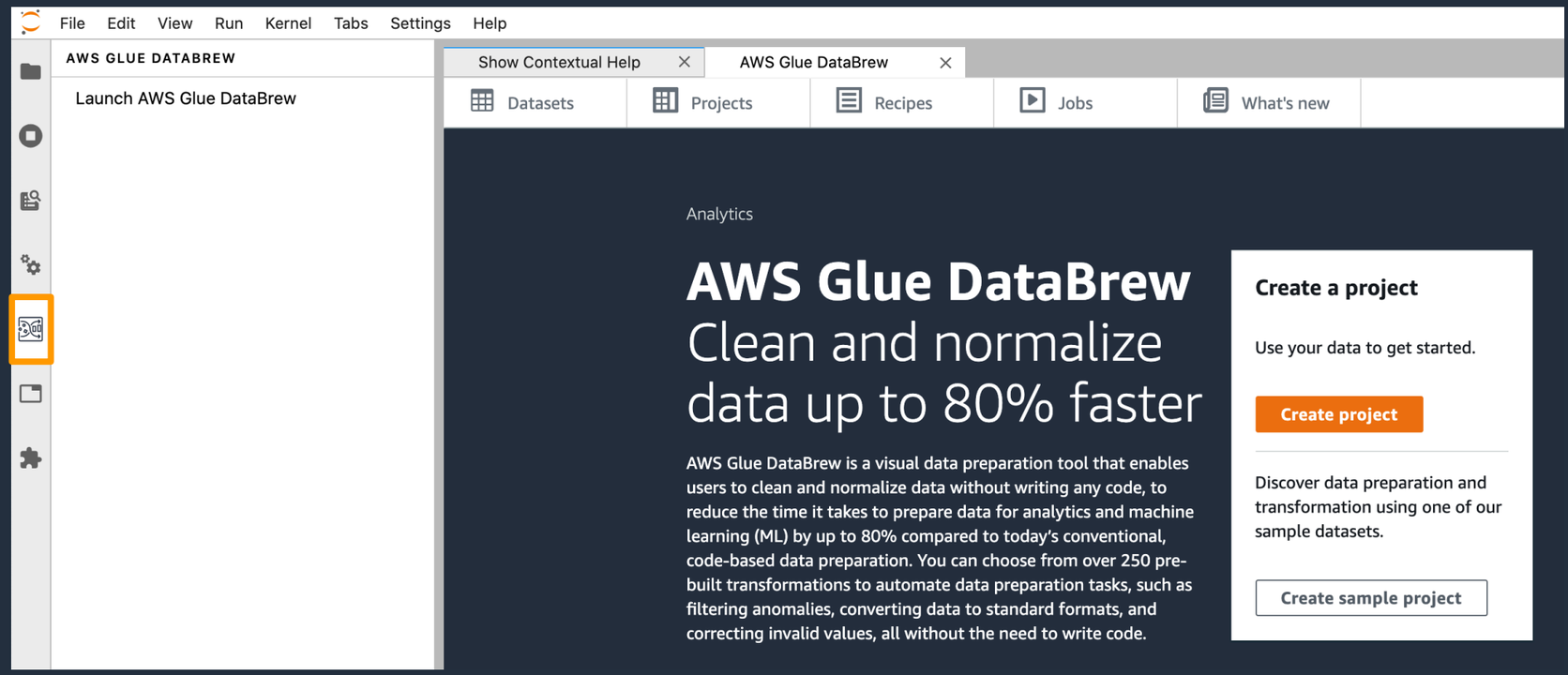


JupyterLab Extension

JupyterLab 経由で AWS Glue DataBrew に接続可能

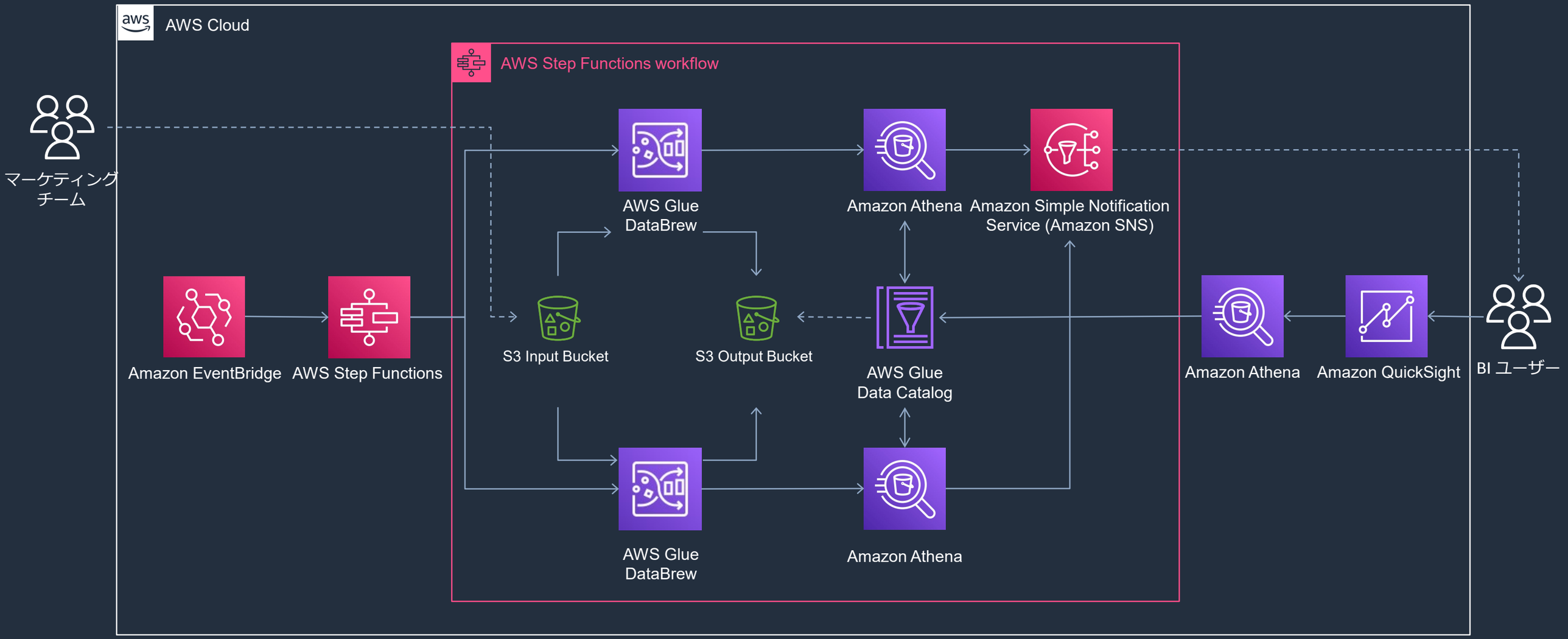
- AWS CLI, AWS Jupyter proxy をセットアップ
- JupyterLab (v.2.2.6 以降) に aws_glue_databrew_jupyter Extension をインストール
- JupyterLab から AWS Glue DataBrew の接続経路はパブリック通信となる

Extension インストールで
JupyterLab から
AWS Glue DataBrew に
直接接続できるようになる



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

4. ワークフロー内でデータ準備をオーケストレーション

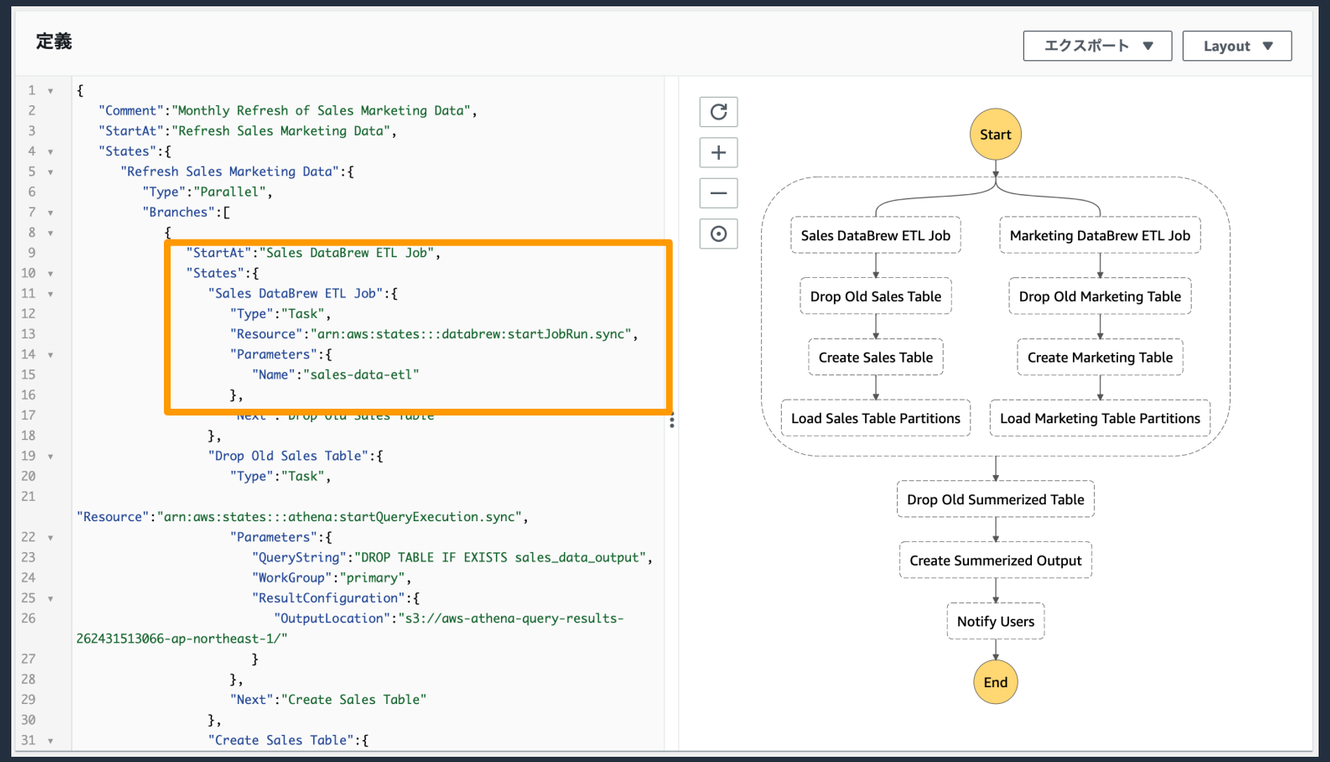


https://aws.amazon.com/jp/blogs/big-data/orchestrating-an-aws-glue-databrew-job-and-amazon-athena-query-with-aws-step-functions/

AWS Step Functions との連携

AWS Glue DataBrew ジョブを
AWS Step Functions の
ワークフローに統合可能

データのクリーニングや正規化の
ステップを、分析や機械学習の
ワークフローの一部として
オーケストレーション可能に



 https://docs.aws.amazon.com/step-functions/latest/dg/connect-databrew.html


ビジネスアナリスト

データサイエンティスト

ETL デベロッパー

AWS Glue DataBrew

リッチなビジュアルインターフェースによりデータを整形・正規化

250 以上の組み込みの変換機能を選択し、タスクを自動化

データパターンや異常値を把握するためのデータプロファイル機能

大規模なデータセットを操作可能

AWS Glue Studio

コードを記述せずに ETL ジョブを視覚的にオーサリング

コンソールから数千のジョブを監視

学習コストなしに分散処理を活用

再利用可能なコードを使った高度な変換

AWS Glue DataBrew の料金

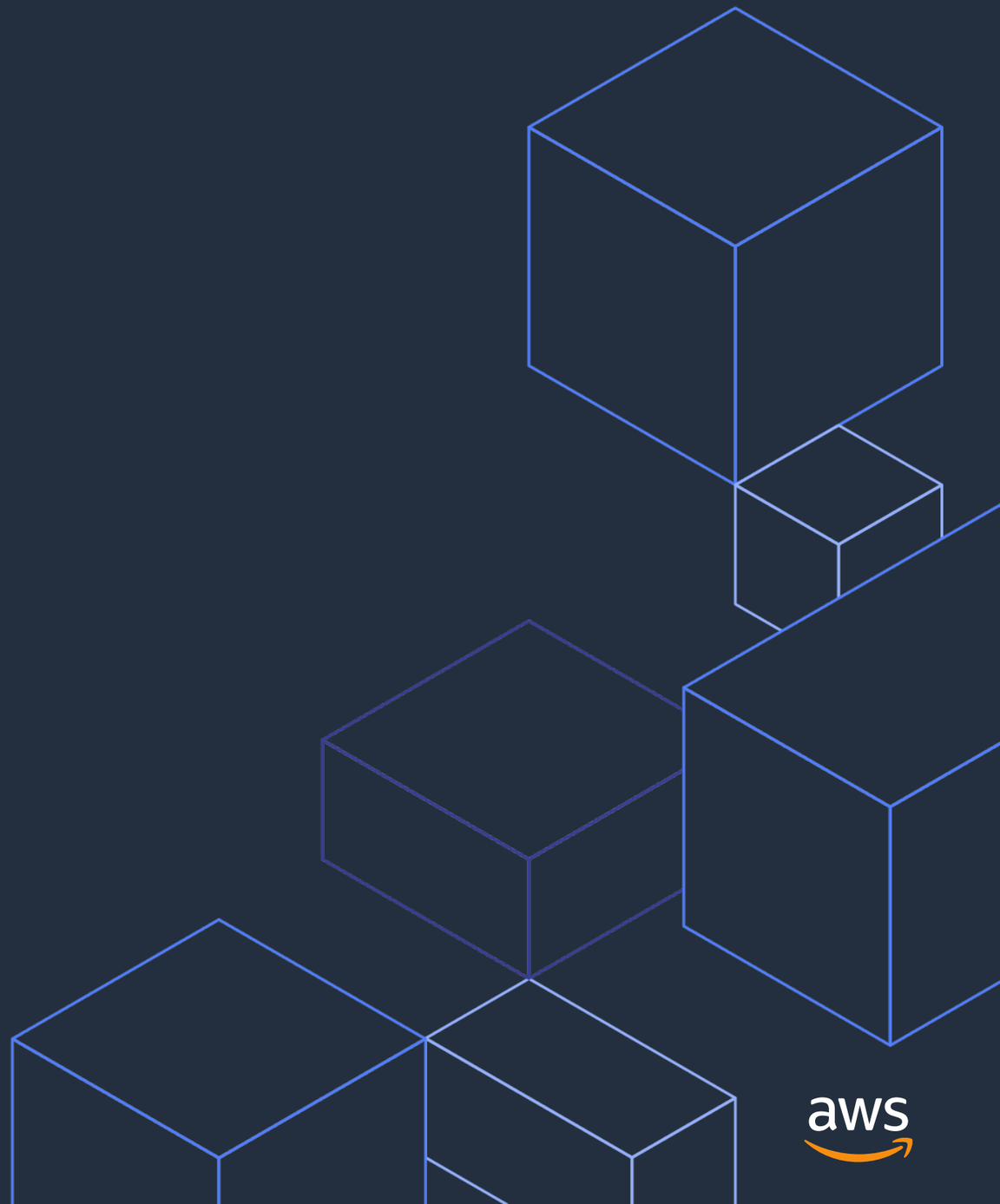
AWS Glue DataBrew の料金

- DataBrew ジョブ - $0.48/node/hour
  - ジョブの実行に使用された AWS Glue DataBrew ノードの数に基づいて 1 時間ごとの料金が発生
  - デフォルトでは各ジョブに 5 ノード 割り当てられる
  - 1 ノード 4 vCPUs, 16GB メモリ

- DataBrew インタラクティブセッション - $1/30分
  - DataBrew プロジェクトを開くとセッションが開始され、未操作の時間が続いた場合自動的にサスペンド
  - はじめて DataBrew にアクセスする場合、最初の 40 セッションは無償

https://aws.amazon.com/jp/glue/pricing/

まとめ



まとめ

- AWS Glue DataBrew は、データのクリーンアップおよび正規化を最大 80% 高速化するビジュアルデータ準備ツール

- データアナリストやサイエンティストがコーディングを行うことなしに、250 種類以上の組み込みの変換処理を使ってデータを分析に必要な形に簡単に整形することが可能

- アドホックなデータ探索、データの品質チェック、機械学習モデル構築の前処理、データ分析パイプライン構築などさまざまなユースケースに活用することが可能

Q&A

お答えできなかったご質問については
AWS Japan Blog https://aws.amazon.com/jp/blogs/news/ に
後日掲載します。

AWS の日本語資料の場所「AWS 資料」で検索

https://amzn.to/JPArchive

AWS Well-Architected 個別技術相談会

毎週”W-A個別技術相談会”を実施中

- AWSのソリューションアーキテクト(SA)に対策などを相談することも可能
- 申込みはイベント告知サイトから

(https://aws.amazon.com/jp/about-aws/events/)

AWS イベント で[検索]

ご視聴ありがとうございました

AWS 公式 Webinar
https://amzn.to/JPWebinar

過去資料
https://amzn.to/JPArchive