[AWS Black Belt Online Seminar]
AWS Glue
-Glue Studioを使ったデータ変換のベストプラクティス-

Amazon Web Service Japan, K. K.
Solutions Architect, Kazutaka Kubo
2021/03/30

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

過去資料
https://amzn.to/JPArchive

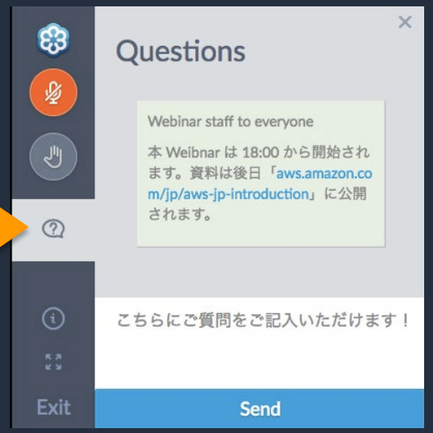
AWS Black Belt Online Seminar とは

「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

内容についての注意点

- 本資料では2021 年 3 月 30 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.

自己紹介

久保　和隆（Kazutaka Kubo）

所属：西日本担当ソリューションアーキテクト

西日本のお客様に対してAWSの技術支援

好きなサービス：

AWS Glue

aws

本日の内容

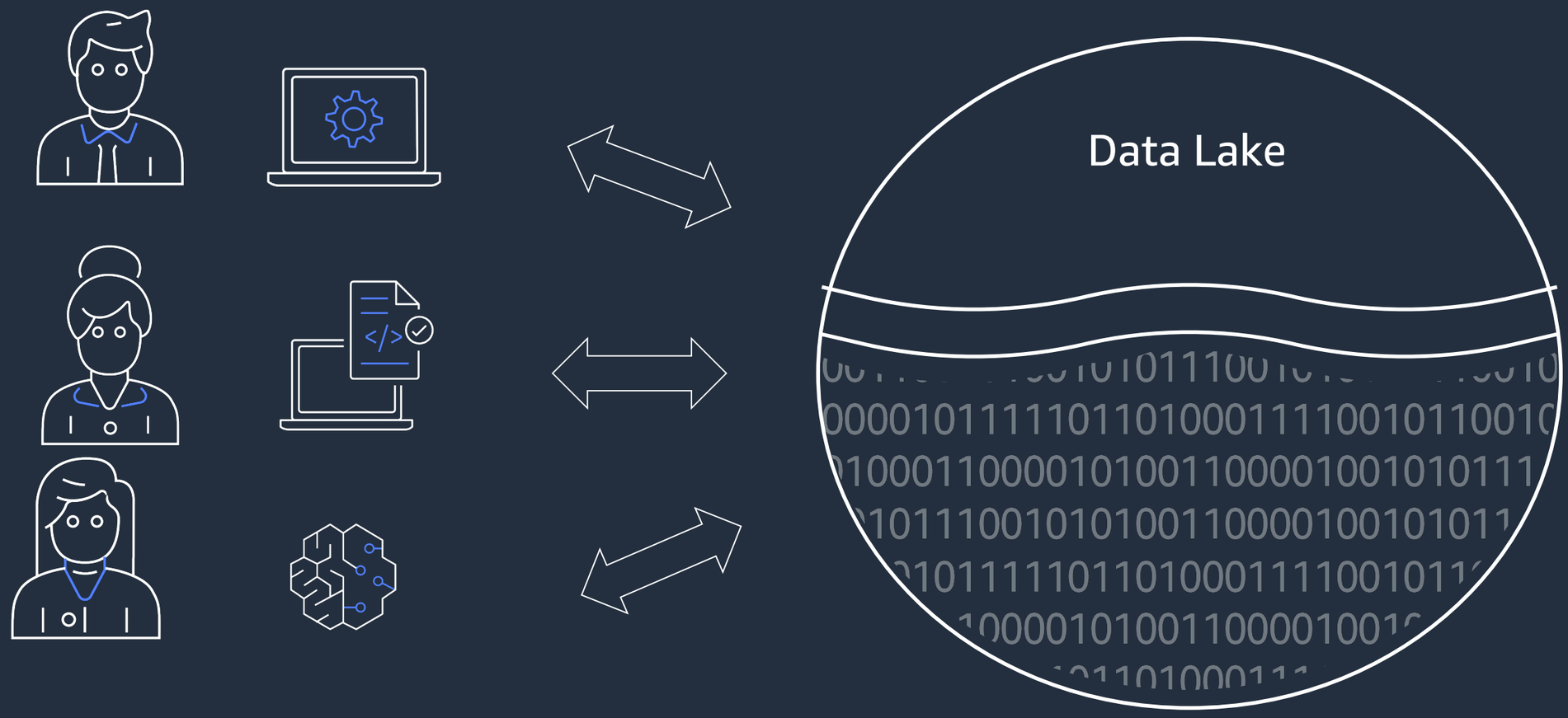
1. AWS Glueの位置付け
2. Glue Studioの概要
3. ユースケースとGlue Studioでの実装
4. その他の主要アップデート
5. まとめ

aws

AWS Glueの位置付け

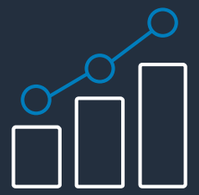
データ分析の現状

・多くの企業がデータレイクを構築し、有効活用を望んでいる。



データ活用を阻む要因

・多くの企業でデータ活用が難航



日々増加するデータ

日々追加される
データソース

日々追加される
データフォーマット

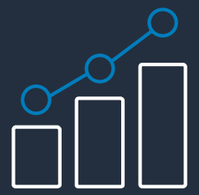
様々な目的のユーザ

データを使いたい
様々なアプリケーション

aws

データ活用を阻む要因

・多くの企業でデータ活用が難航



日々増加するデータ

日々追加される
データソース

日々追加される
データフォーマット

様々な目的のユーザ

データを使いたい
様々なアプリケーション

使いやすい形に整形する「前処理」が、データ活用には重要

aws

前処理における課題

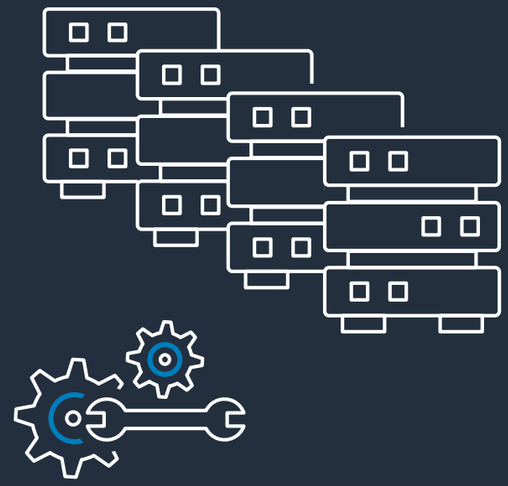
取扱データの増加

- 5年前と比較して、
  10倍以上のデータ量

目的に合わせた
カスタマイズが必要

- ごみデータの排除
- フォーマット変換　等々

基盤の管理負荷



- サーバーのライフサイクル管理
- 監視管理

aws

AWS Glue

サービス間でデータを簡単に**移動**できるようにするための、
サーバーレスデータ統合サービス

aws

AWS Glue の特徴

シンプルで拡張性に優れたサーバ不要のデータ統合

大量データの迅速な統合

データの準備を数ヶ月から数分に短縮

変換処理の自動化

何千ものETLジョブを簡単に実行、管理可能

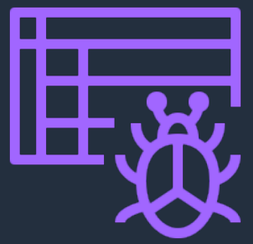
サーバーレス

ジョブ実行で使われたリソースに対してのみ支払い

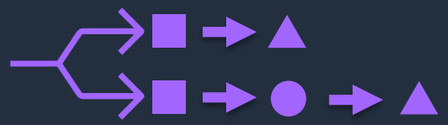
AWS Glue の構成要素と周辺サービス

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog

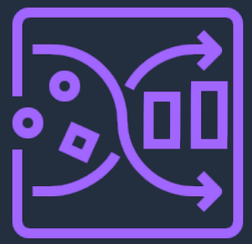


Crawler

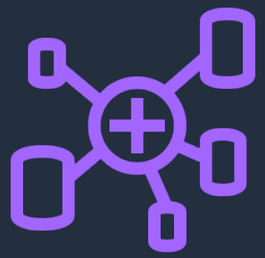


Workflow Management

AWS Glue Studio



AWS Glue DataBrew



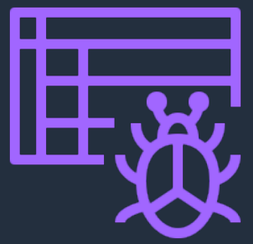
AWS Glue Elastic Views

aws

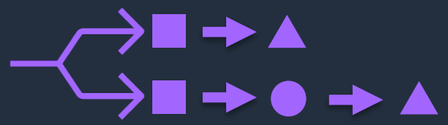
AWS Glue の構成要素と周辺サービス

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog



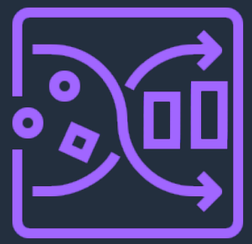
Crawler



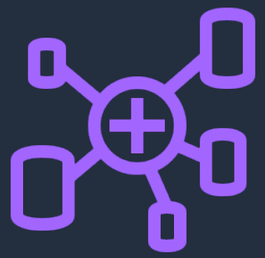
Workflow Management

New

AWS Glue Studio



AWS Glue DataBrew



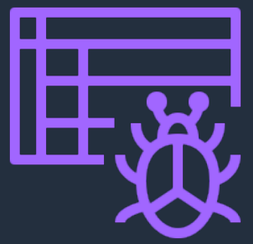
AWS Glue Elastic Views

aws

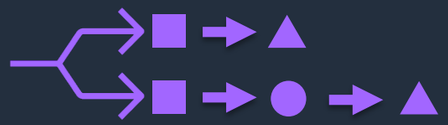
AWS Glue の構成要素と周辺サービス

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog



Crawler



Workflow Management

フォーマット変換したり、データを結合したり、
様々な処理を行うJOBを作成・管理・実行することが可能

AWS Glue Studio

AWS Glue DataBrew

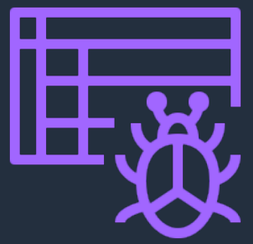
AWS Glue Elastic Views

aws

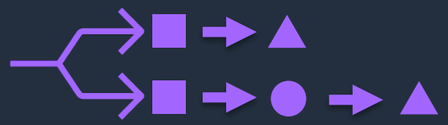
AWS Glue の構成要素と周辺サービス

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog



Crawler



Workflow Management

Apache Hiveメタストア互換のメタデータリポジトリ
データソースにどういったデータが入っているのかをカタログ化して、保存しておくことが可能

AWS Glue Studio

AWS Glue DataBrew

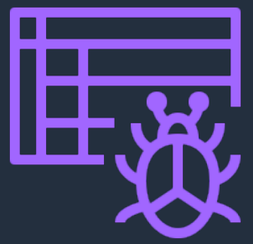
AWS Glue Elastic Views

aws

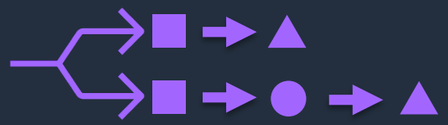
AWS Glue の構成要素と周辺サービス

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog



Crawler



Workflow Management

New

Glueのデータカタログにメタデータを作成するプログラム
分類子の優先度に従って、スキーマ情報を自動で判断し、スキーマを形成

AWS Glue Studio

AWS Glue DataBrew

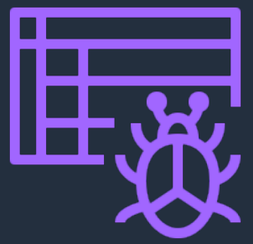
AWS Glue Elastic Views

aws

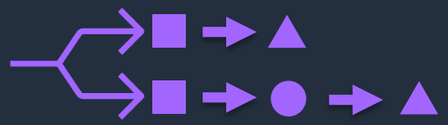
AWS Glue の構成要素と周辺サービス

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog



Crawler



Workflow Management

New

クローラー、トリガー、ジョブのDAGを生成するワークフロー機能
データソースのクロール/データカタログの生成 / JOBの実行、これら一連の処理を自動化

AWS Glue Studio

AWS Glue DataBrew

AWS Glue Elastic Views

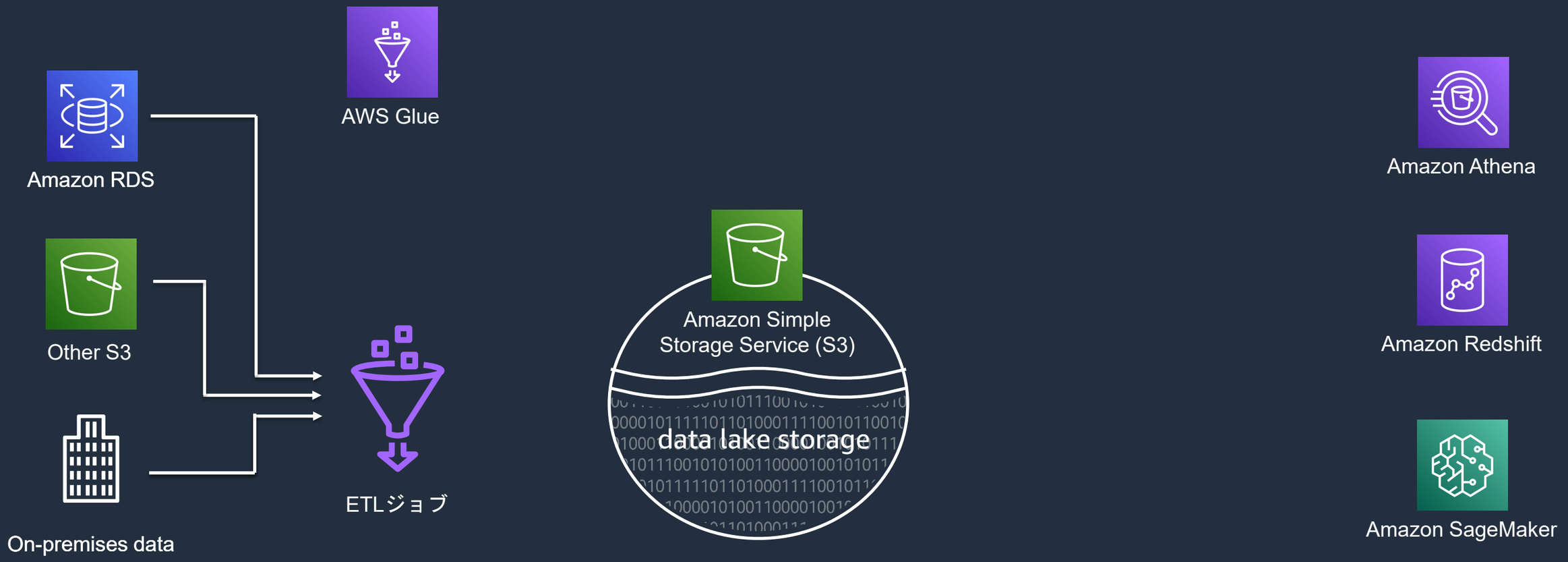
aws

データレイク環境の構成要素

データレイク環境におけるAWS Glueの役割

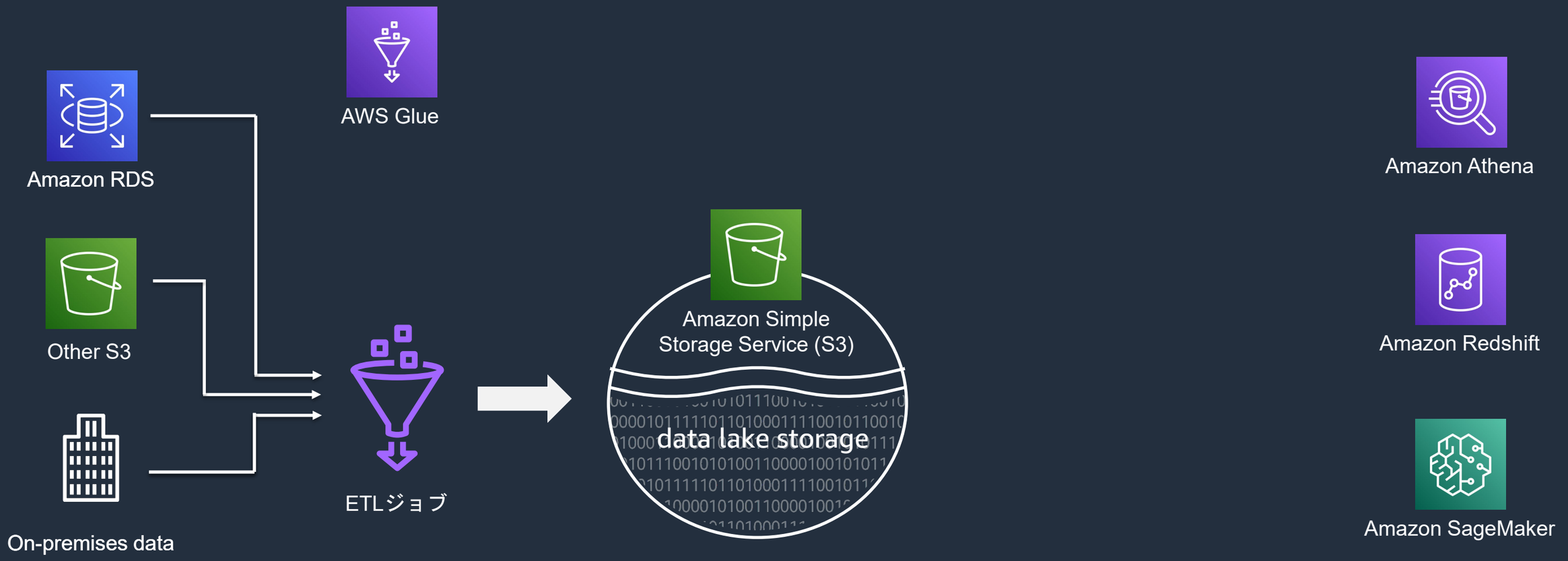
各要素の間を取り持つのがGlueの役割

ETLジョブ



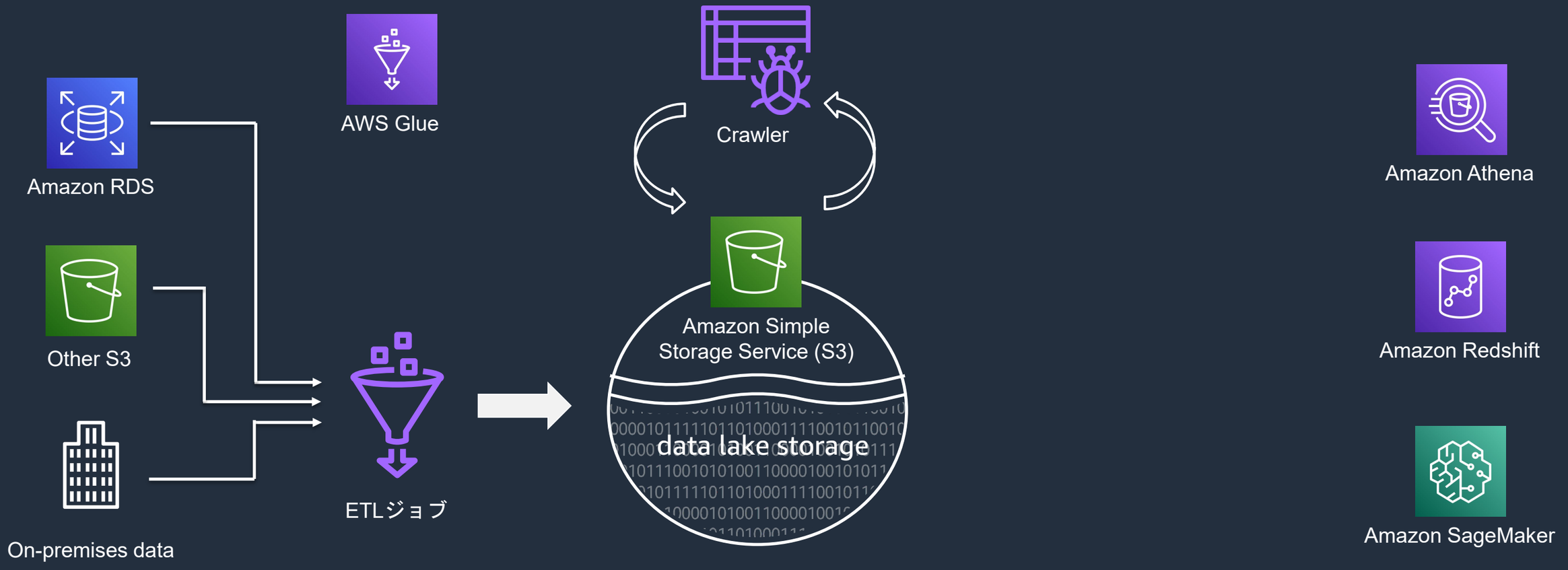
データソースからデータレイクストレージへデータを出力

ETLジョブ



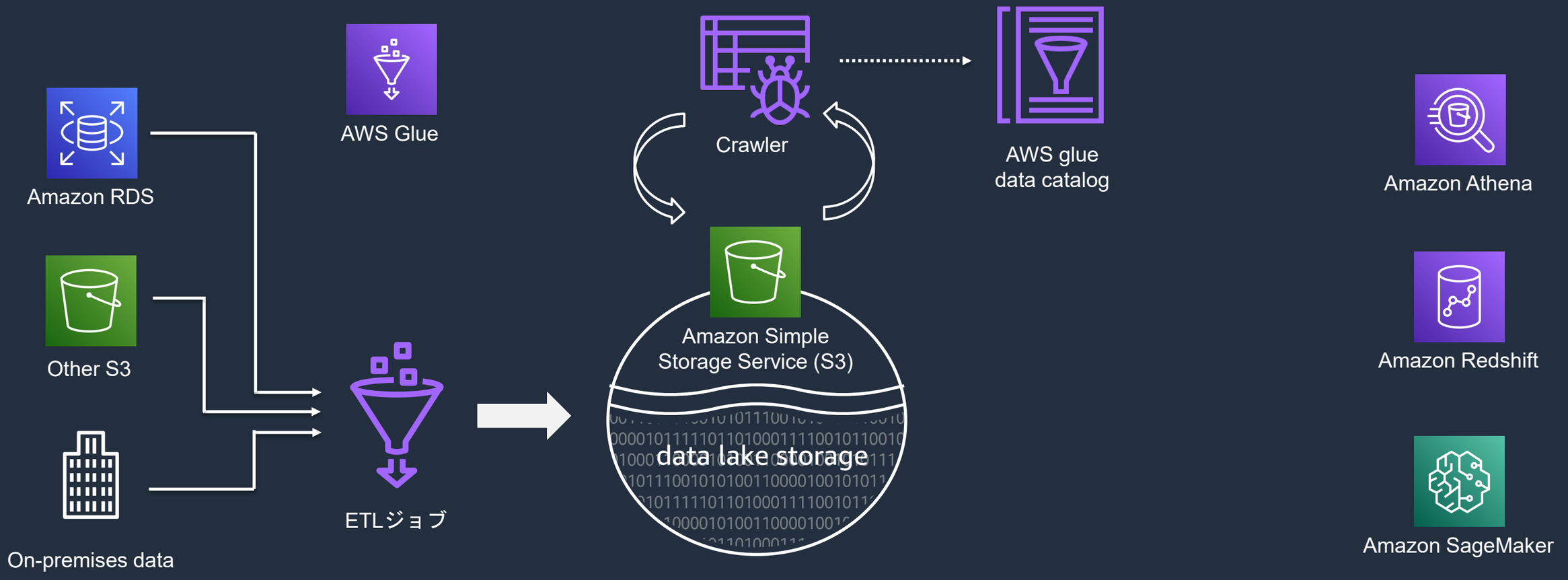
データソースからデータレイクストレージへデータを出力

データのクローリング



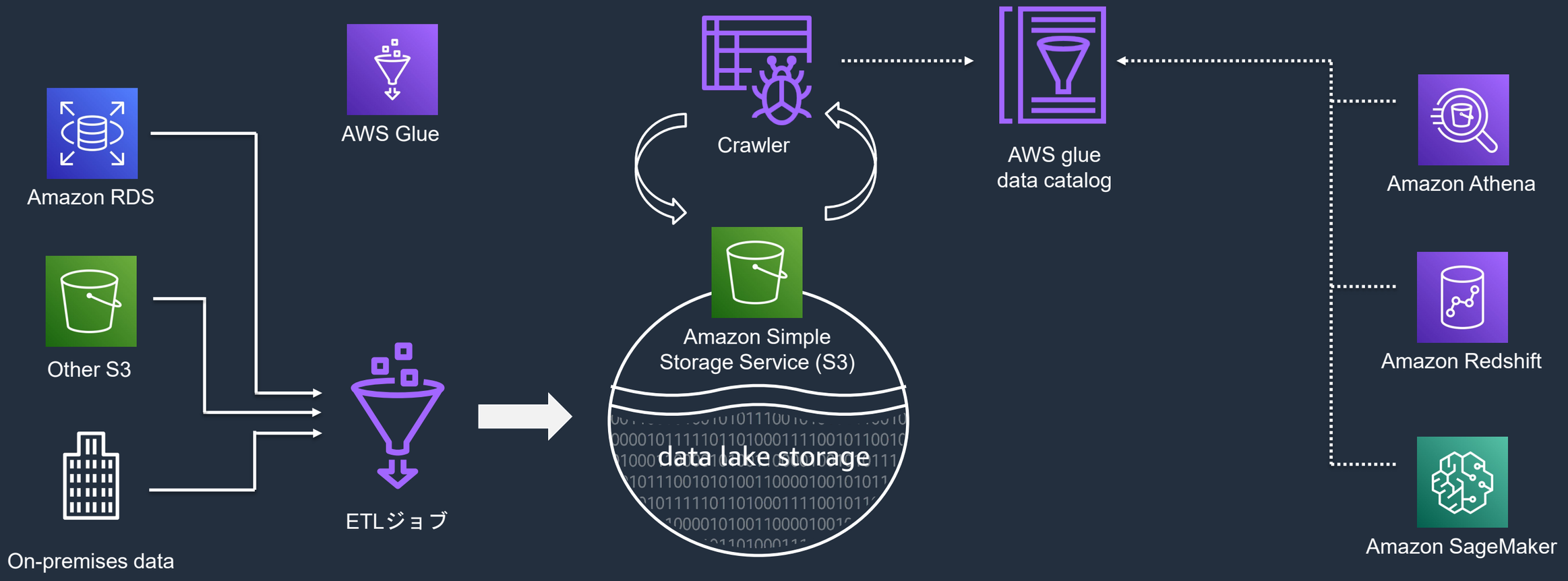
データレイクをクローリングし、データカタログを出力

データのクローリング



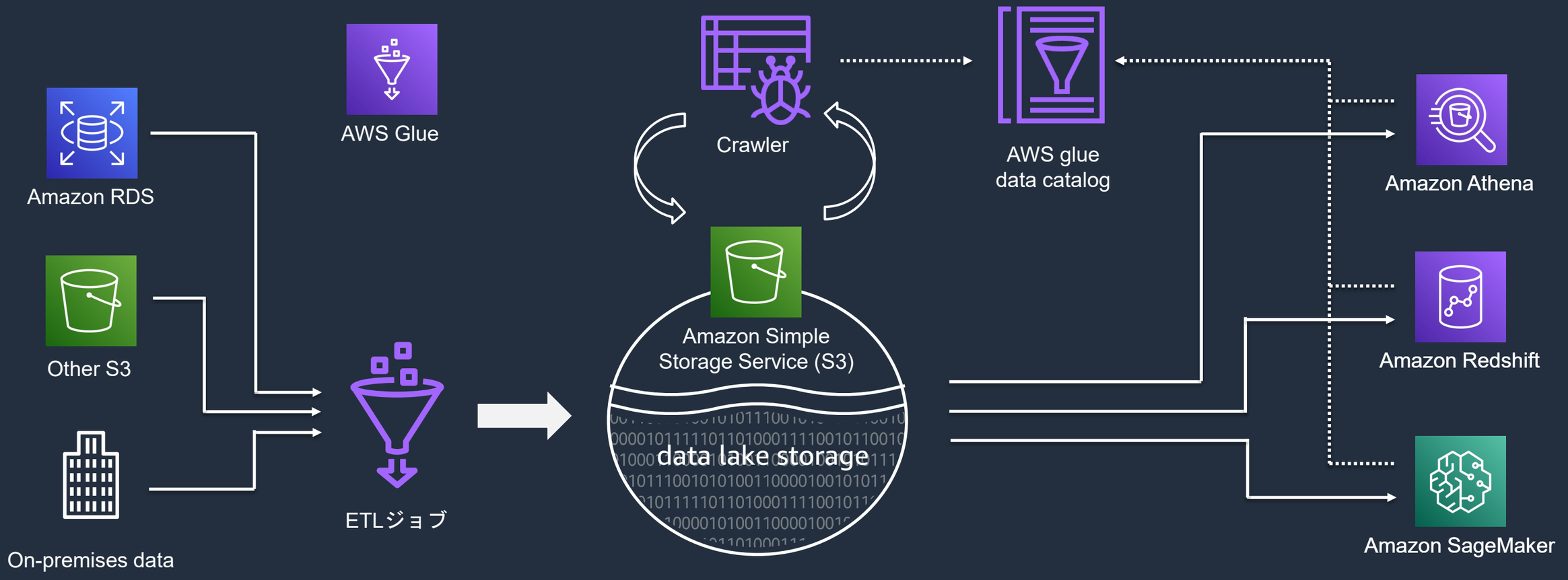
データレイクをクローリングし、データカタログを出力

カタログを使用したデータ参照



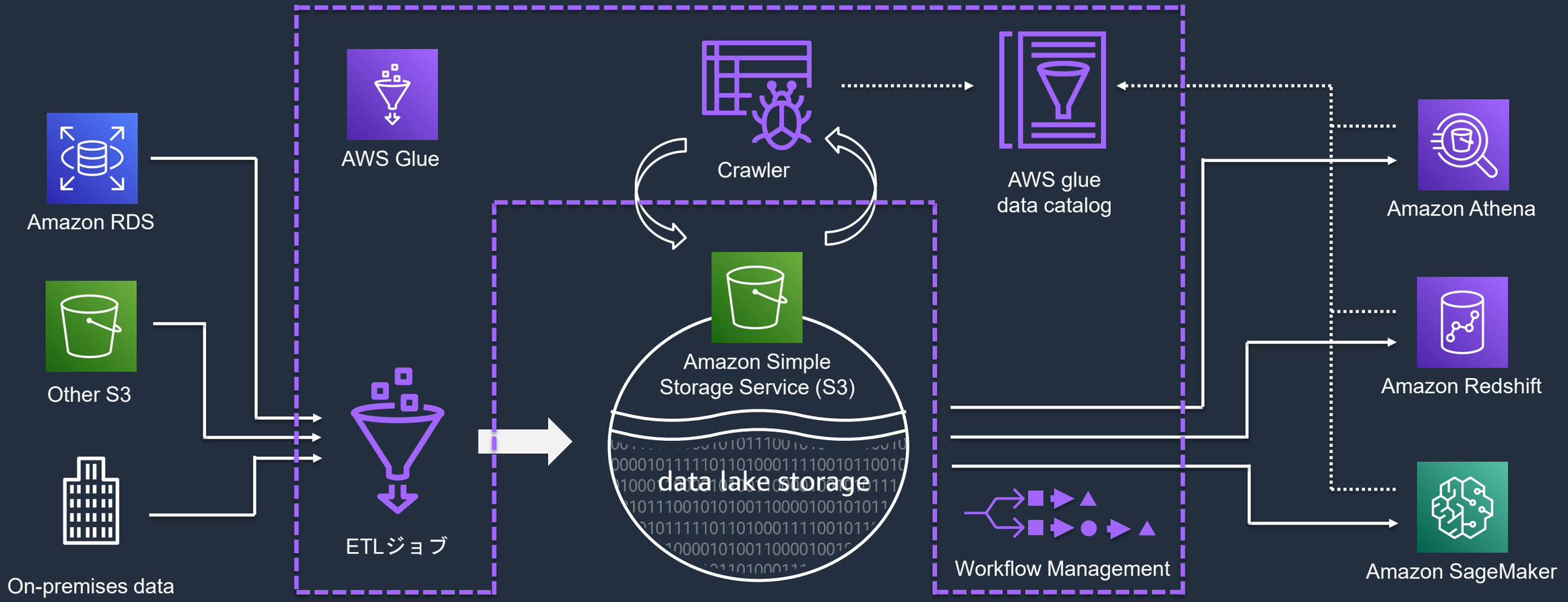
データカタログを参照し、必要データをロード

カタログを使用したデータ参照



データカタログを参照し、必要データをロード

ワークフロー管理

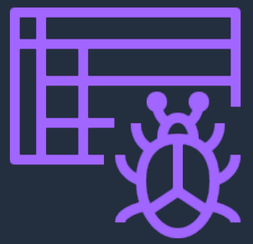


一連の処理をワークフロー化し、自動化可能

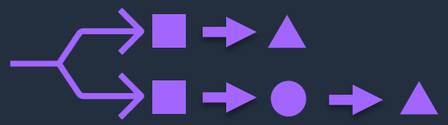
AWS Glue の構成要素と周辺サービス（再掲）

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog



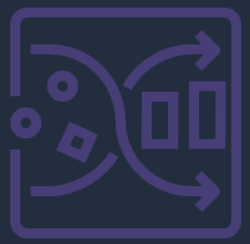
Crawler



Workflow Management

New

AWS Glue Studio



AWS Glue DataBrew

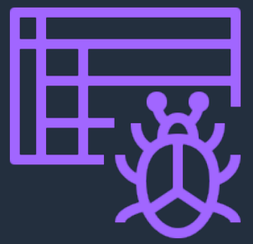
AWS Glue Elastic Views

aws

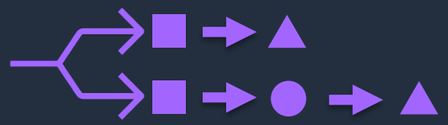
AWS Glue の構成要素と周辺サービス（再掲）

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog



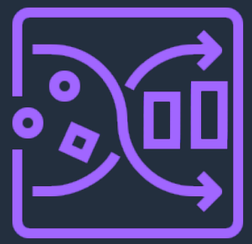
Crawler



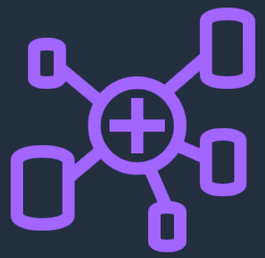
Workflow Management

New

AWS Glue Studio



AWS Glue DataBrew



AWS Glue Elastic Views

aws

AWS Glue の構成要素と周辺サービス（再掲）

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog

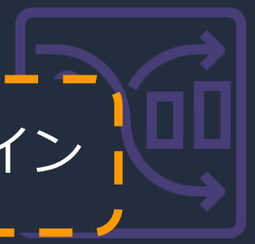
Crawler

Workflow Management

New

AWS Glue Studio

本日のメイン



AWS Glue DataBrew

AWS Glue Elastic Views

aws

AWS Glue Studioの概要

代表的なデータ利用者

ビジネスユーザー：
データ分析についての専門的な知見を持ってはいないが、蓄積したデータを活用する。
組織で活動するあらゆる人が当てはまる。

ETLデベロッパー：
自分の開発・運用しているプロダクトのためにデータを活用する。
システム障害の原因追求、新機能がどう使われているかの分析するロール。

データアナリスト：
データ分析自体が職務で、 マーケティング施策の効果を売り上げデータをもとに検証する。
データによる意思決定の支援を行うロール。

データサイエンティスト：
より高度な手法を使い、複雑な意識決定の支援を実施する。
業務システムに組み込んで使用するための機械学習モデルの開発を行うロール。

aws

代表的なデータ利用者

ビジネスユーザー：
データ分析についての専門的な知見を持ってはいないが、蓄積したデータを活用する。
組織で活動するあらゆる人が当てはまる。

ETLデベロッパー：
自分の開発・運用しているプロダクトのためにデータを活用する。
システム障害の原因追求、新機能がどう使われているかの分析するロール。

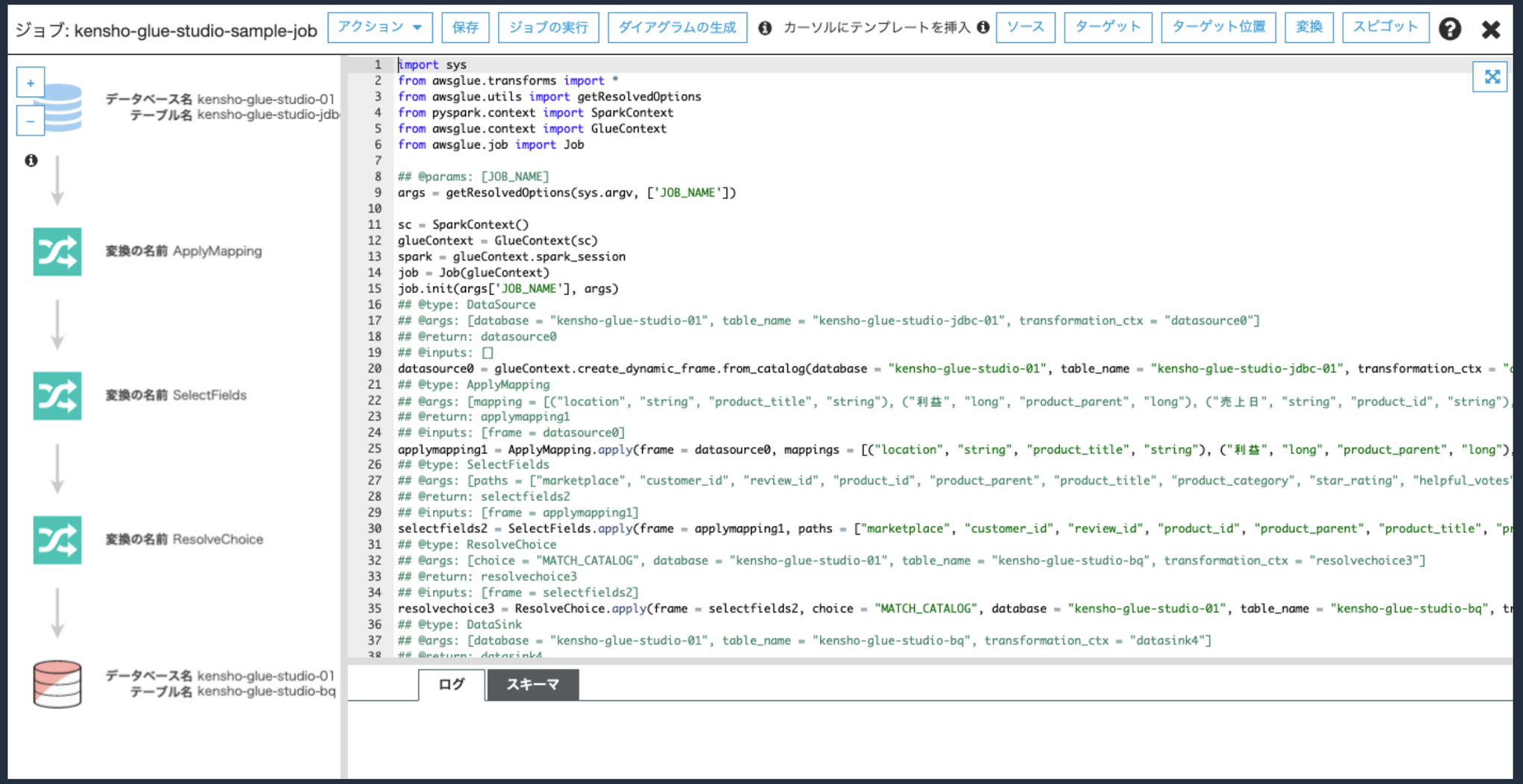
データアナリスト：
データ分析自体が職務で、 マーケティング施策の効果を売り上げデータをもとに検証する。
データによる意思決定の支援を行うロール。

データサイエンティスト：
より高度な手法を使い、複雑な意識決定の支援を実施する。
業務システムに組み込んで使用するための機械学習モデルの開発を行うロール。

aws

従来のAWS Glue でのJOB実装

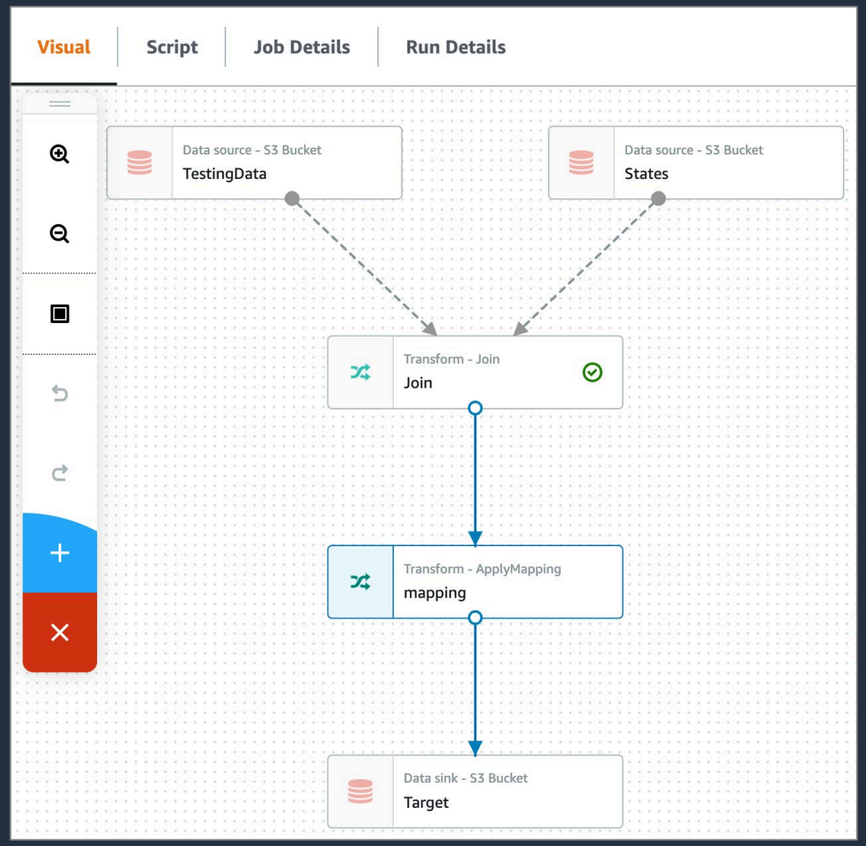
コードベースのJOB作成インターフェース



AWS Glue Studio

NEW

ETL ジョブの作成、実行、監視を容易にする視覚的なインターフェース



ビジュアルオーサリング
コードを書くことなくETLジョブを作成可能

サーバーレス
ビッグデータ処理能力を利用可能

シンプルビュー
単一画面でジョブを管理可能

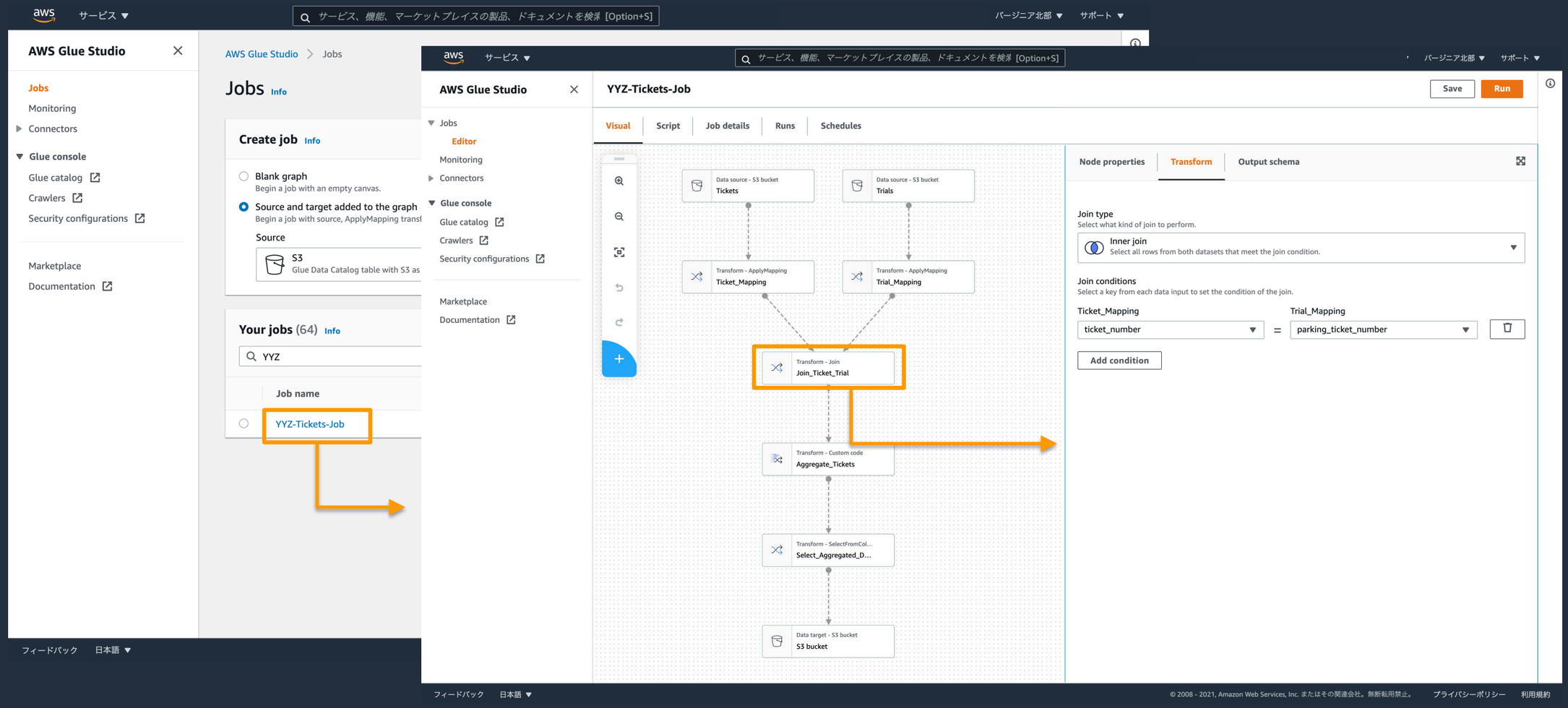
カスタマイズ可能
任意のコード(Python、Scala、Java)で ETL を高度化可能

https://aws.amazon.com/jp/blogs/big-data/making-etl-easier-with-aws-glue-studio/

aws

シンプルビューの詳細

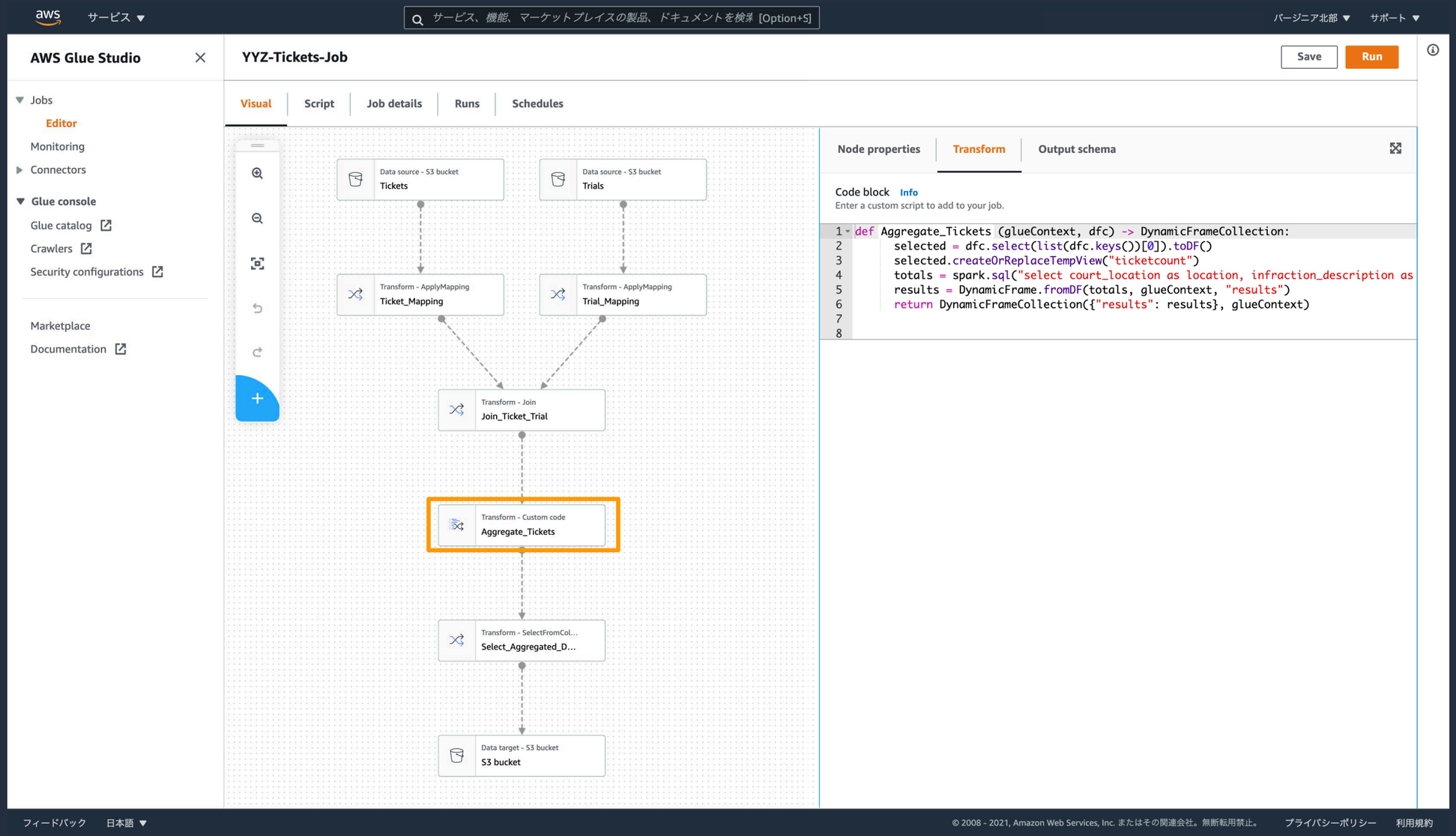
NEW



https://aws.amazon.com/jp/blogs/big-data/making-etl-easier-with-aws-glue-studio/

カスタマイズを可能にするノード

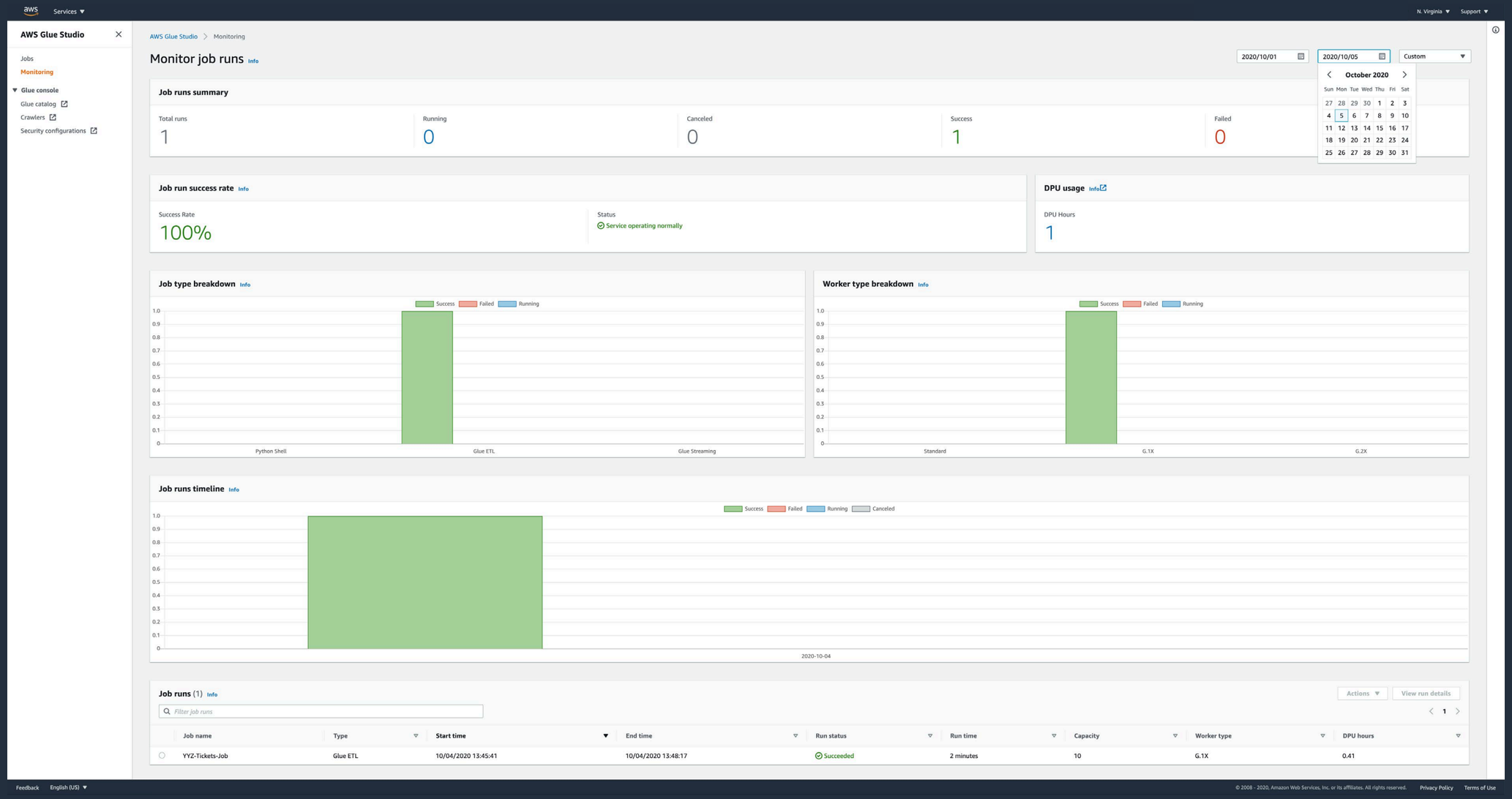
NEW



https://aws.amazon.com/jp/blogs/big-data/making-etl-easier-with-aws-glue-studio/

視覚的なJOBモニタリング

NEW

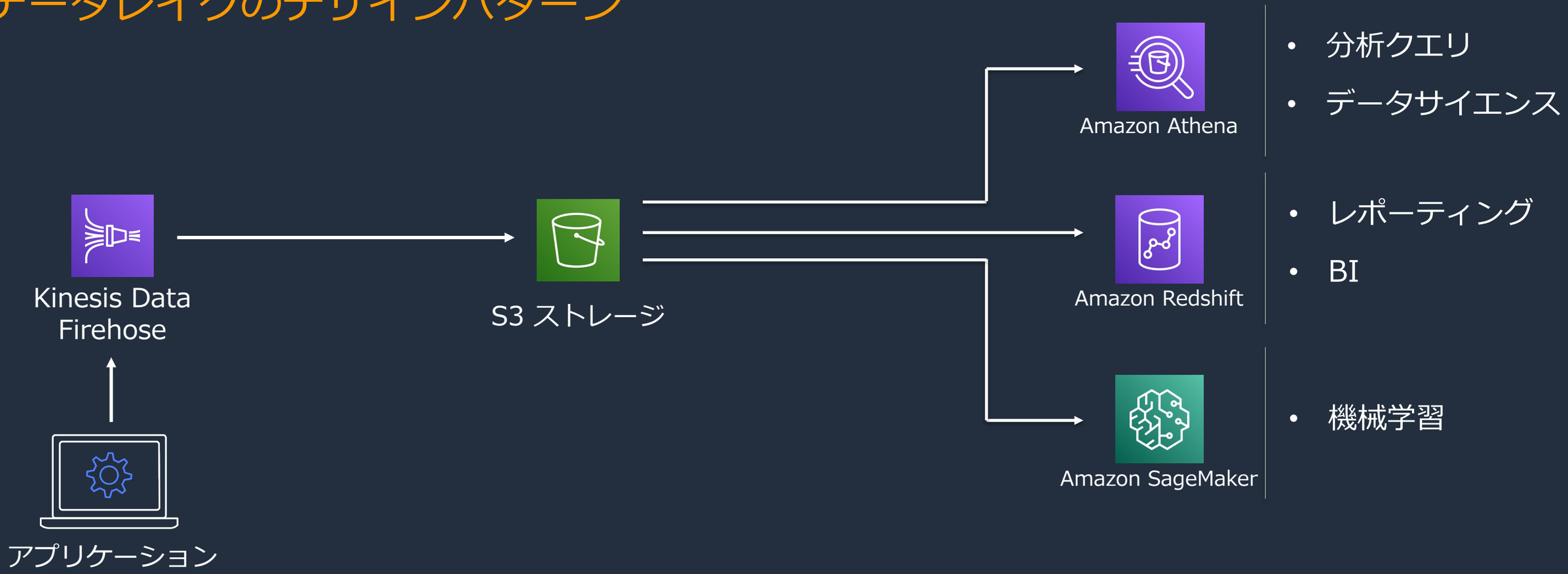


ユースケースとGlue Studioでの実装

ユースケース 1：
ログストリーム/IoTセンサーデータの分析

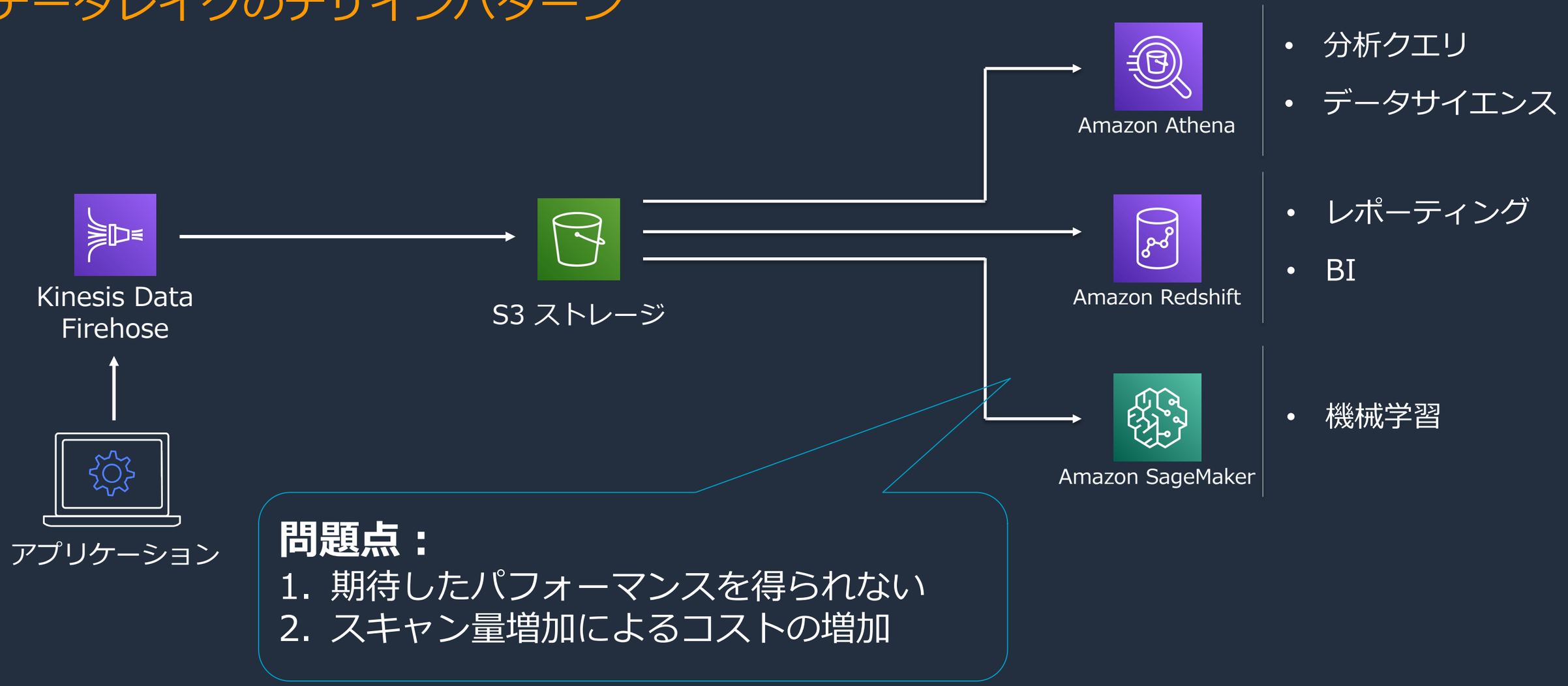
ログストリーム / IoTセンサーデータ分析の課題

リアルタイムで発生するデータを分析するための データレイクのデザインパターン



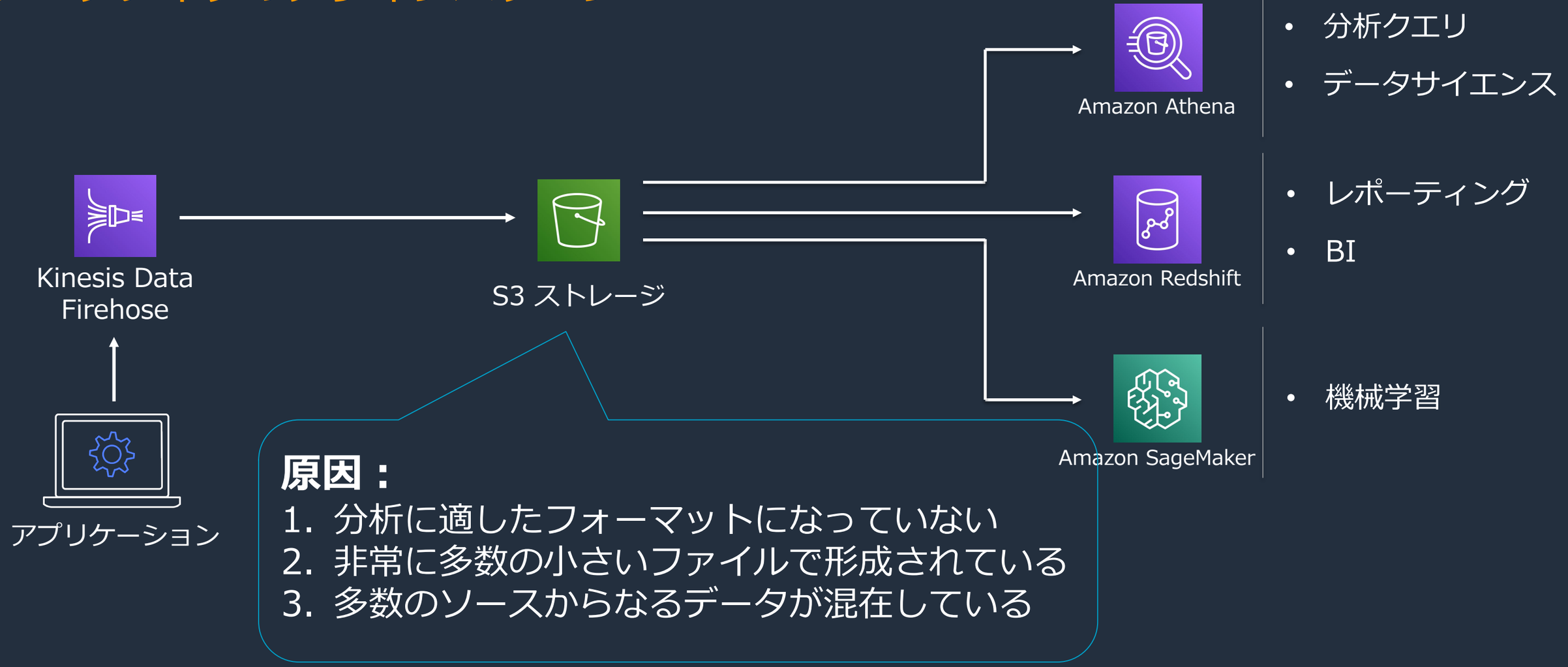
ログストリーム / IoTセンサーデータ分析の課題

リアルタイムで発生するデータを分析するための データレイクのデザインパターン



ログストリーム / IoTセンサーデータ分析の課題

リアルタイムで発生するデータを分析するための データレイクのデザインパターン



具体的なデータイメージ

細かい大量のファイルが定期的に出力

名前	タイプ	最終更新日時	サイズ	ストレージクラス
kensho-glue-studio-01-1-2021-03-02-12-00-31-a12ae31b-c9d5-4fce-bc34-659908b2b585	-	2021/03/02 09:01:33 PM JST	20.3 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-01-34-bdb3957f-ee56-4516-accc-05c1fbd44bab	-	2021/03/02 09:02:36 PM JST	19.4 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-02-35-1978af44-5535-4670-8f29-7807b07017de	-	2021/03/02 09:03:37 PM JST	19.5 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-03-37-0730996d-edf3-425b-85e0-c4e1b1800836	-	2021/03/02 09:04:38 PM JST	19.3 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-04-38-68b18a5e-6208-4294-832e-7cb700db41b9	-	2021/03/02 09:05:40 PM JST	20.5 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-05-40-15c8353e-655a-4cf2-b1e2-cb2df0c863e6	-	2021/03/02 09:06:41 PM JST	18.1 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-06-40-050373b7-f719-4d74-9c7e-b07bb2a81f14	-	2021/03/02 09:07:43 PM JST	20.4 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-07-42-876ccbae-9688-483c-9f5e-6a0d718dc92d	-	2021/03/02 09:08:43 PM JST	18.2 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-08-42-dfe50012-79f0-4bbf-9496-988212cfdb8e	-	2021/03/02 09:09:45 PM JST	19.4 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-09-43-03d4964f-93fd-443c-8c60-168664b38be2	-	2021/03/02 09:10:45 PM JST	19.5 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-10-45-507353e9-0dc3-4173-8bc5-9ed619046633	-	2021/03/02 09:11:46 PM JST	19.3 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-11-46-6d46bb54-37b0-4800-8dd8-a395a2ca07f4	-	2021/03/02 09:12:47 PM JST	19.5 KB	スタンダード
kensho-glue-studio-01-1-2021-03-02-12-12-47-3128179f-24a5-4282-9925-96623035a1ec	-	2021/03/02 09:13:48 PM JST	19.3 KB	スタンダード

中身はJSON形式のテキストファイルかつ、様々なシステムからのアラートが混在


```
{"timestamp":"02/Mar/2021:12:00:29 +0000","alarmlevel":"WARNING","host":"prd-web","user":"SystemA","number":"1001","text":"This is WARNING"}
{"timestamp":"02/Mar/2021:12:00:29 +0000","alarmlevel":"WARNING","host":"prd-db","user":"SystemD","number":"1001","text":"This is WARNING"}
{"timestamp":"02/Mar/2021:12:00:29 +0000","alarmlevel":"WARNING","host":"prd-db","user":"SystemF","number":"1001","text":"This is WARNING"}
{"timestamp":"02/Mar/2021:12:00:29 +0000","alarmlevel":"WARNING","host":"prd-ap","user":"SystemA","number":"1001","text":"This is WARNING"}
```

このデータに対して求められること

解決策：

1. フォーマットを列指向フォーマットに変換する
2. 大量ファイルをコンパクションする
3. 適切な単位でパーティショニングする

上記の変更を加えることで、より効率的な分析が可能

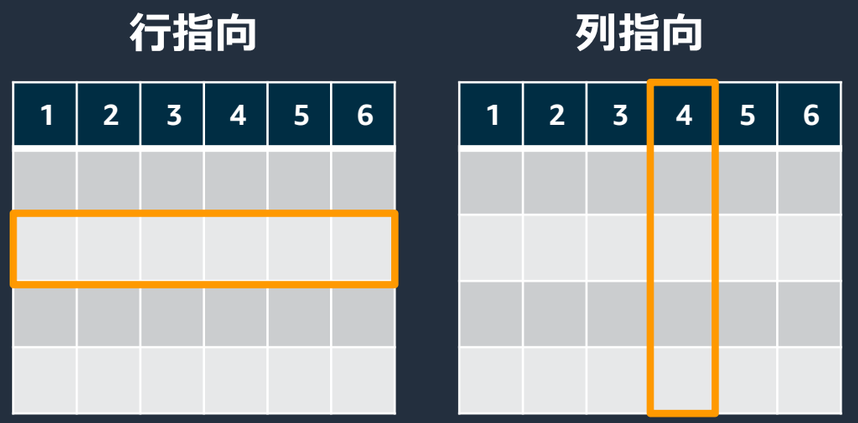
aws

列指向フォーマットとは

カラム（列単位）でデータをまとめて保存するデータフォーマット

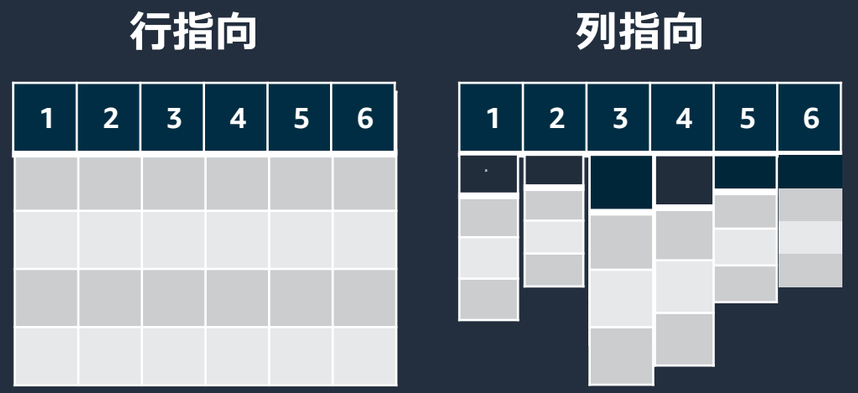
メリット1) OLAP 系の分析クエリを効率的に実行できる

- 多くの分析クエリは、一度のクエリで一部のカラムしか使用しない



メリット2) I/O の効率があがる

- 圧縮と同時に使うことで I/O 効率がさらに向上
- カラムごとに分けられてデータが並んでいる
- 同じカラムは、似たような中身のデータが続くため、圧縮効率がよくなる

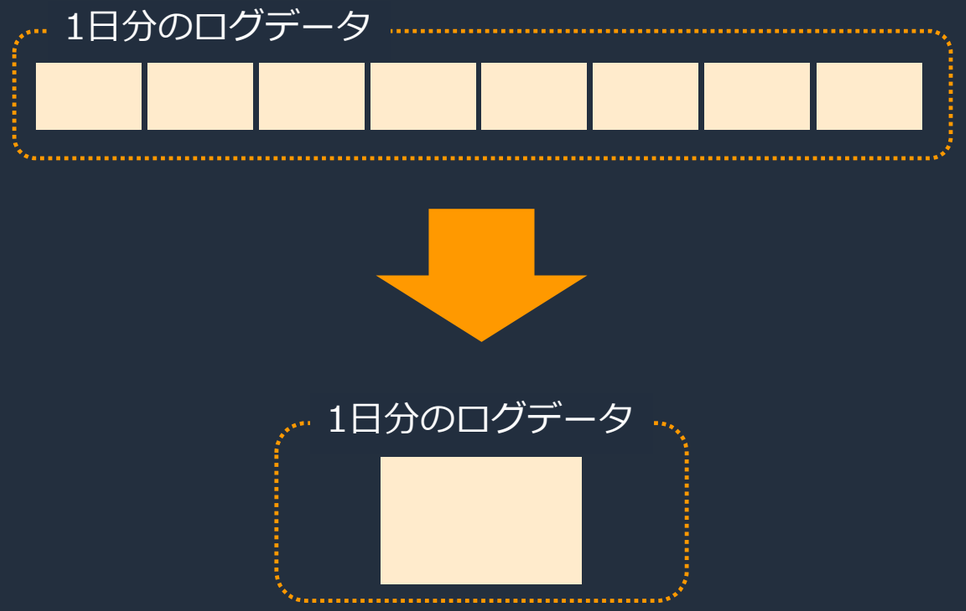


コンパクションとは

サイズの小さい複数ファイルを分析のしやすい単位に集約

ファイル数が多い場合の問題点

- Hadoop系の分散エンジンでは、読み込む必要のあるブロックごとにタスクが生成される。その為、データが少ないブロックへの処理ではオーバーヘッドが発生しパフォーマンスが低下する。



コンパクション後のメリット

- データ分析前にファイルをひとまとめにしておくことで、オーバーヘッドが減少し分析時のパフォーマンスが向上する。

パーティショニングとは

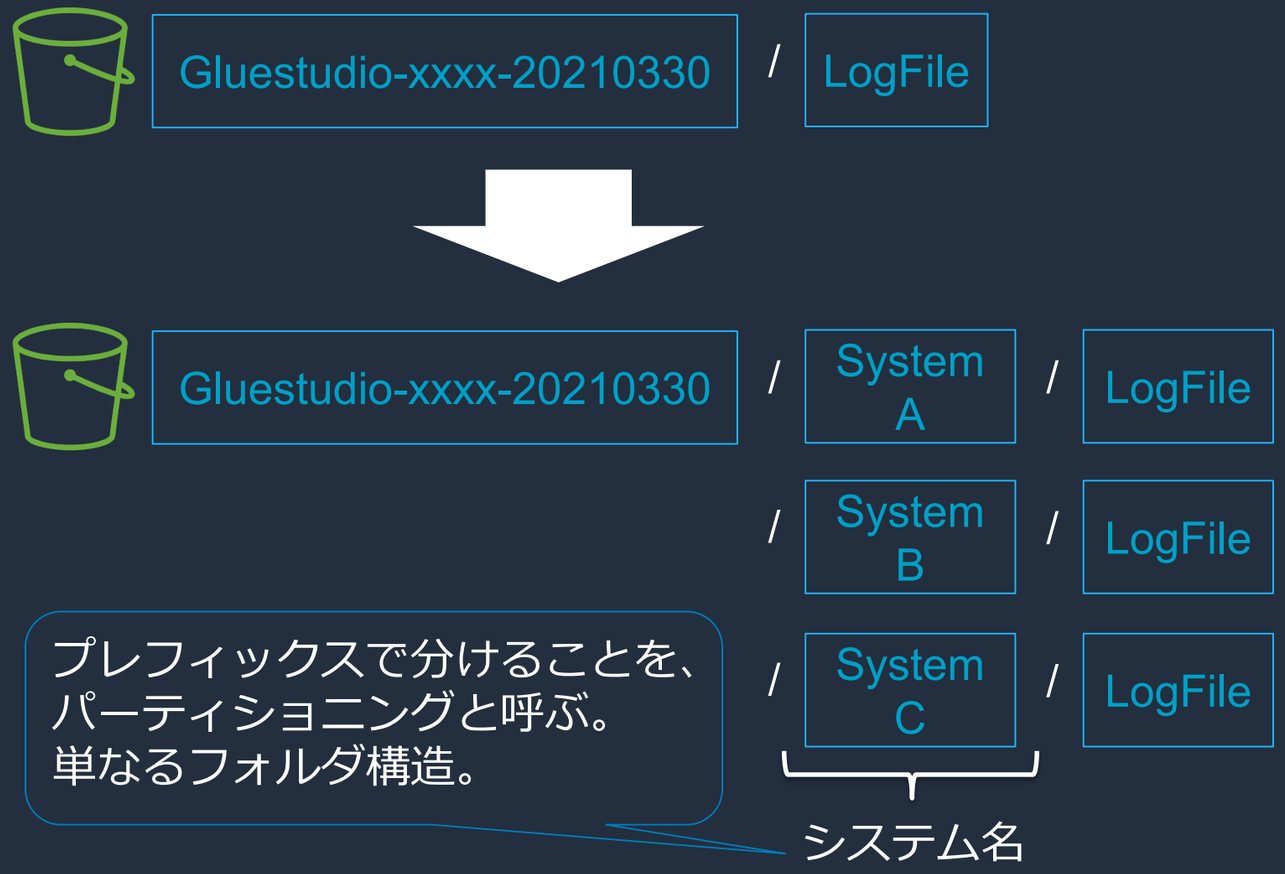
分析によく使う単位でファイルを分析して格納

<u>パーティショニングをしない場合の問題点</u>

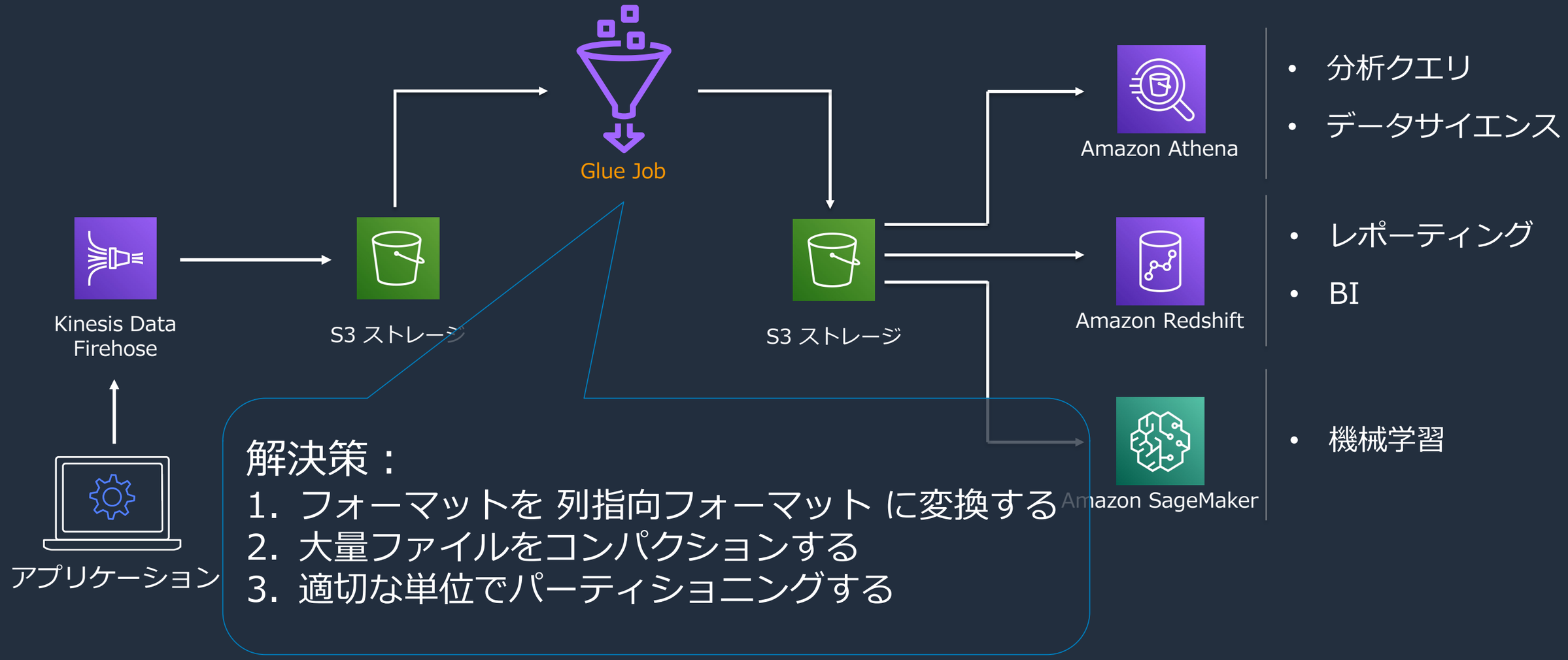
- 特定のシステムのログを検索したい場合、
  全てのLogFileに対してスキャンが掛かる。
  その為、読み込むデータ量が多くなり、
  パフォーマンスに影響がある。

<u>パーティショニング後のメリット</u>

- 検索条件にシステム名を含めることで、
  特定のファイルにのみアクセスする。
  不要データへのアクセスが減る為、
  パフォーマンスが向上する。

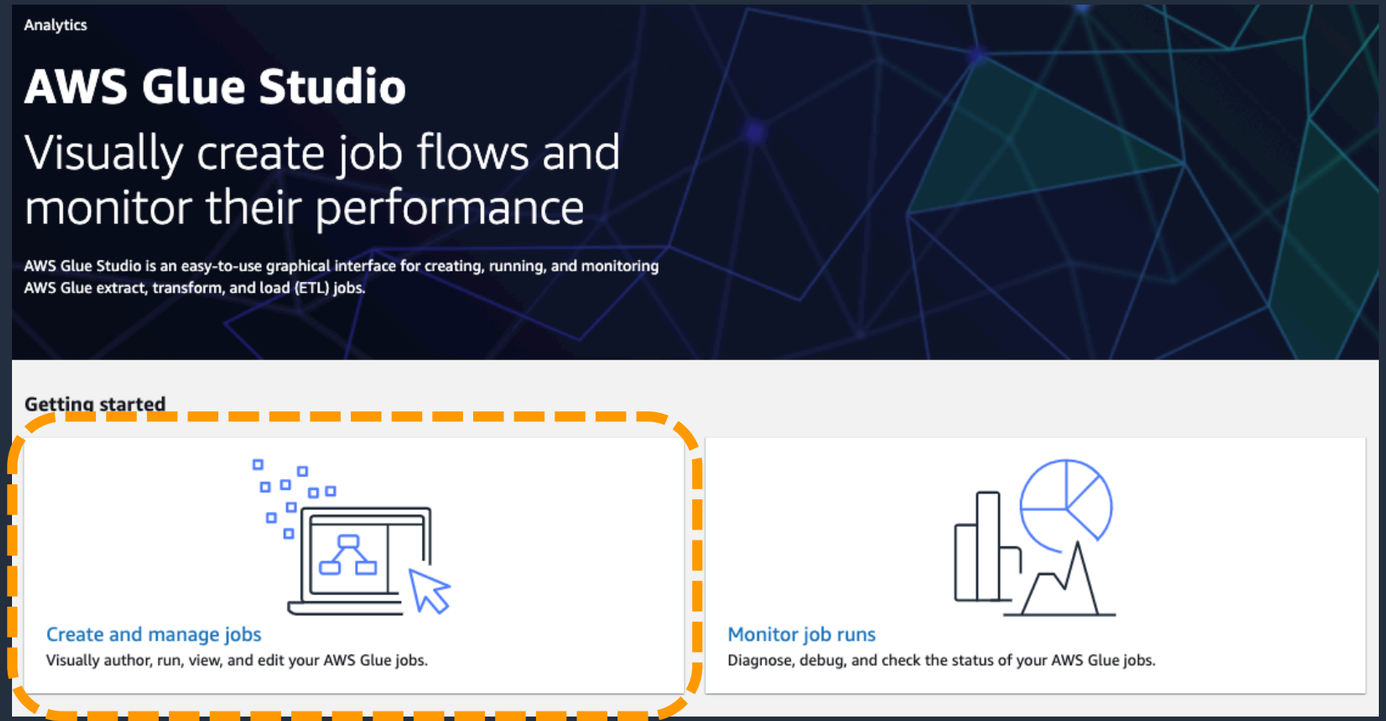


ログストリーム / IoTセンサーデータ分析の解決策



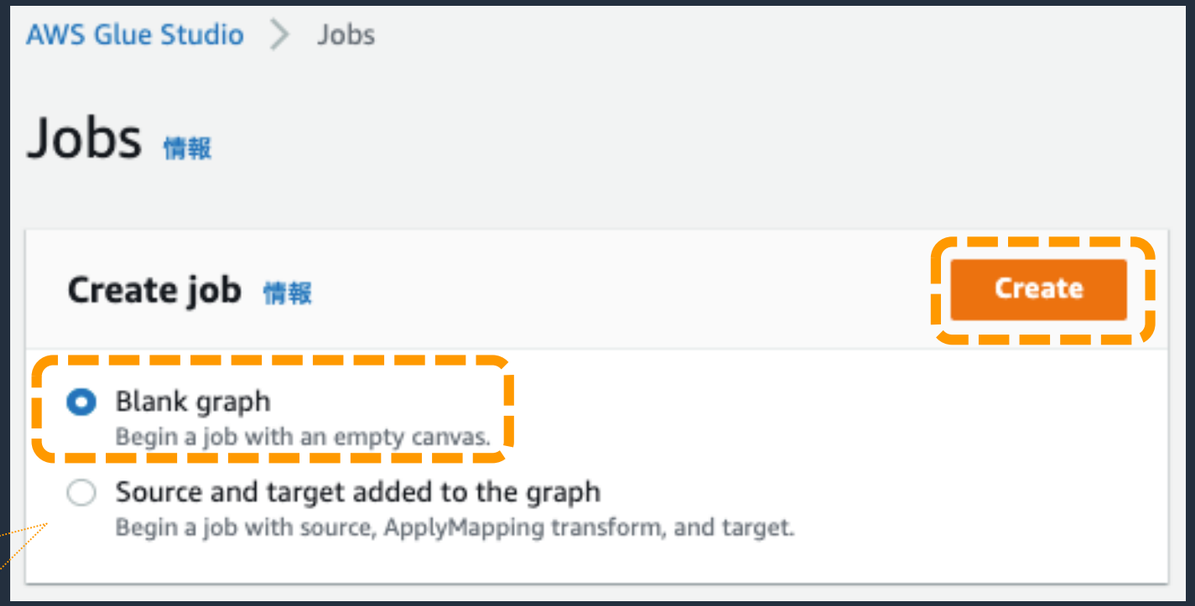
AWS Glue Studioを使用したETL JOBの作成

- マネージメントコンソールより、
  AWS Glue を選択
- AWS Glue コンソールの左ペインより、
  AWS Glue Studioを選択
- 表示されたAWS Glue Studioの画面にて
  Create and manage Jobs を選択



JOBの作成

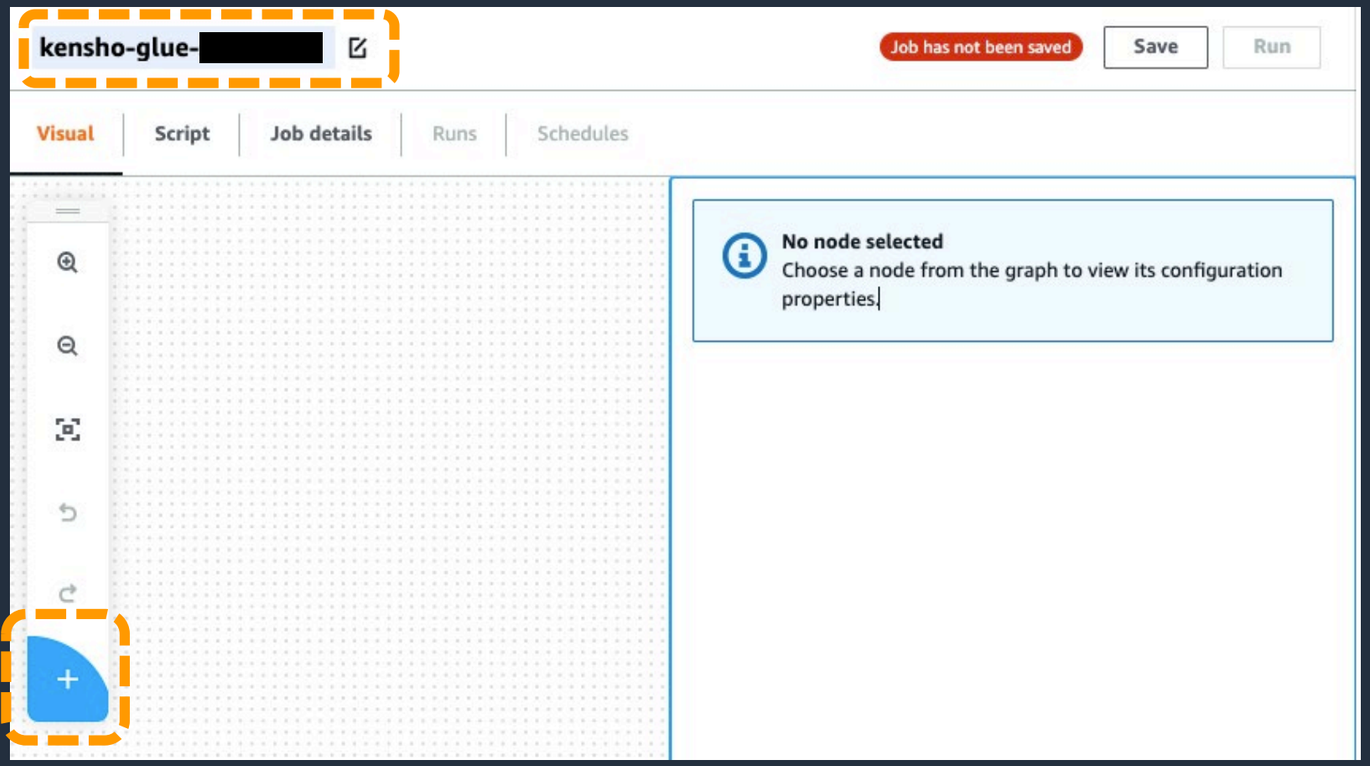
- Create job の項目にて、
  Blank graphを選択しCreateボタンを押下



Source and target added to the graph
指定したデータソースとターゲットノードがあらかじめ作成された状態から開始する。どちらを選んでも問題ない。

データソースの指定①

- JobのEdit画面が表示
- 左上にJob名を入力
- 画面左下の“＋”を押下し、ノードを表示

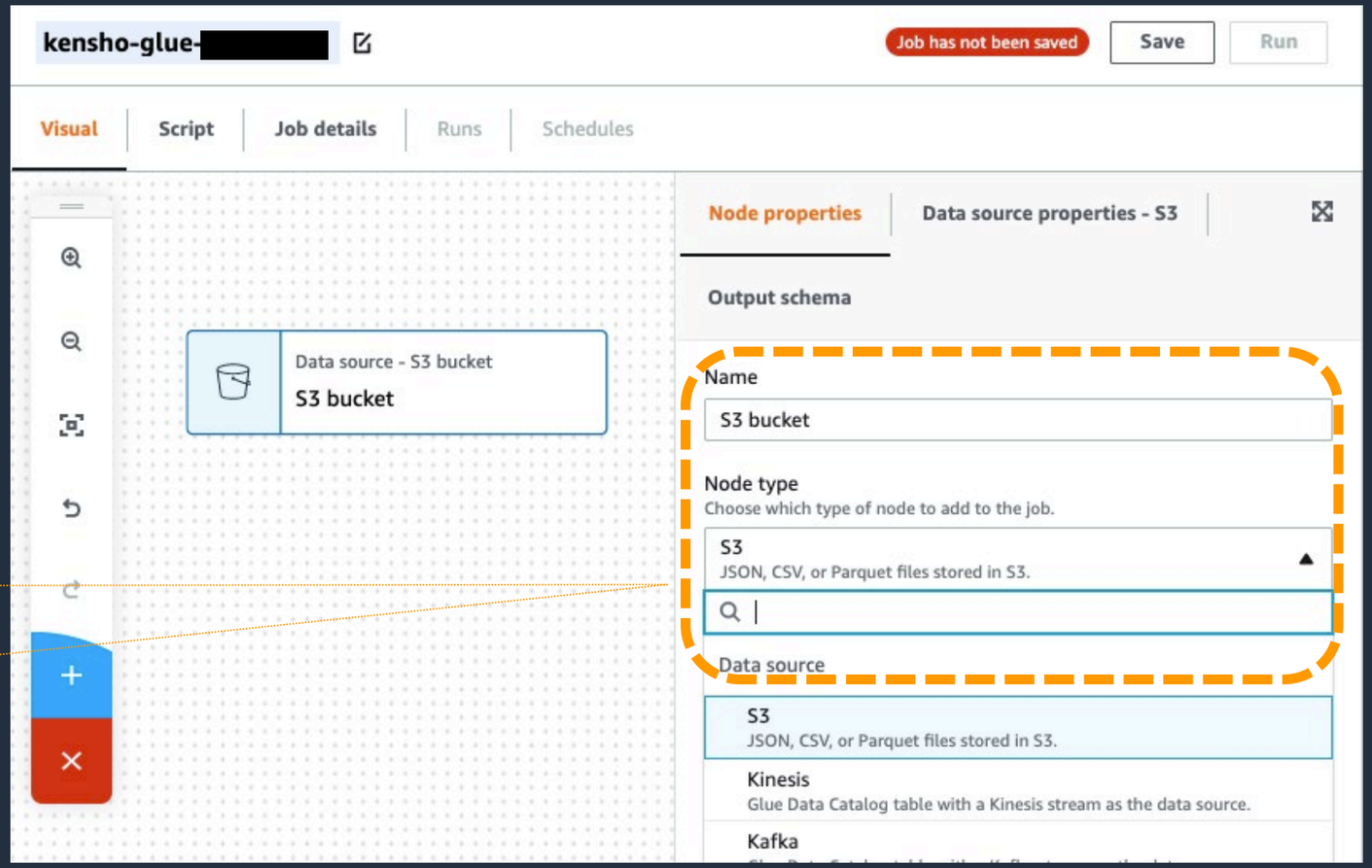


データソースの指定②

- Node typeにS3を選択

選択可能なデータソースは、
Kinesis, Kafka, JDBC, Redshift　等々。

custom connectorを使用することで追加可能。

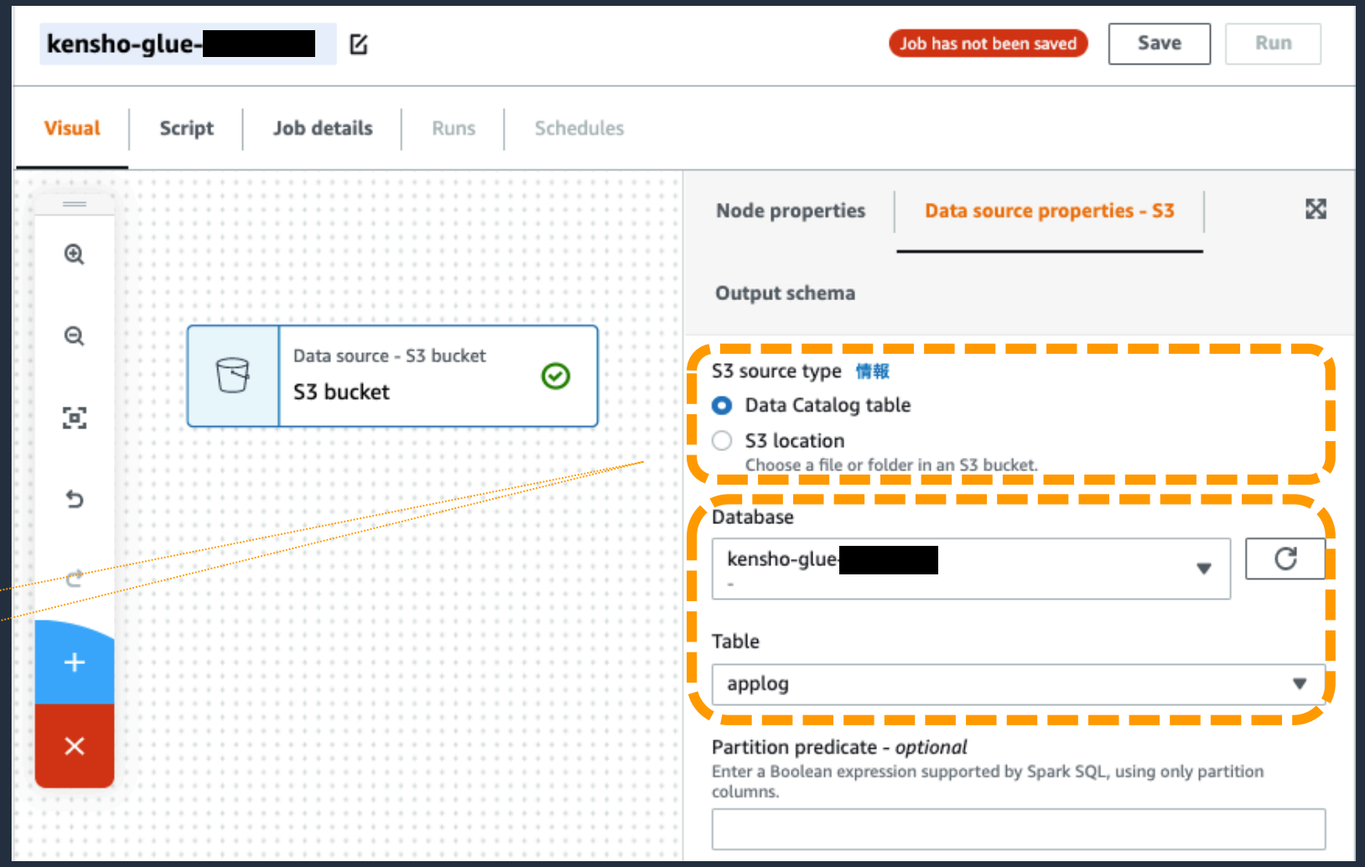


データソースの指定③

- Data Catalog tableを選択

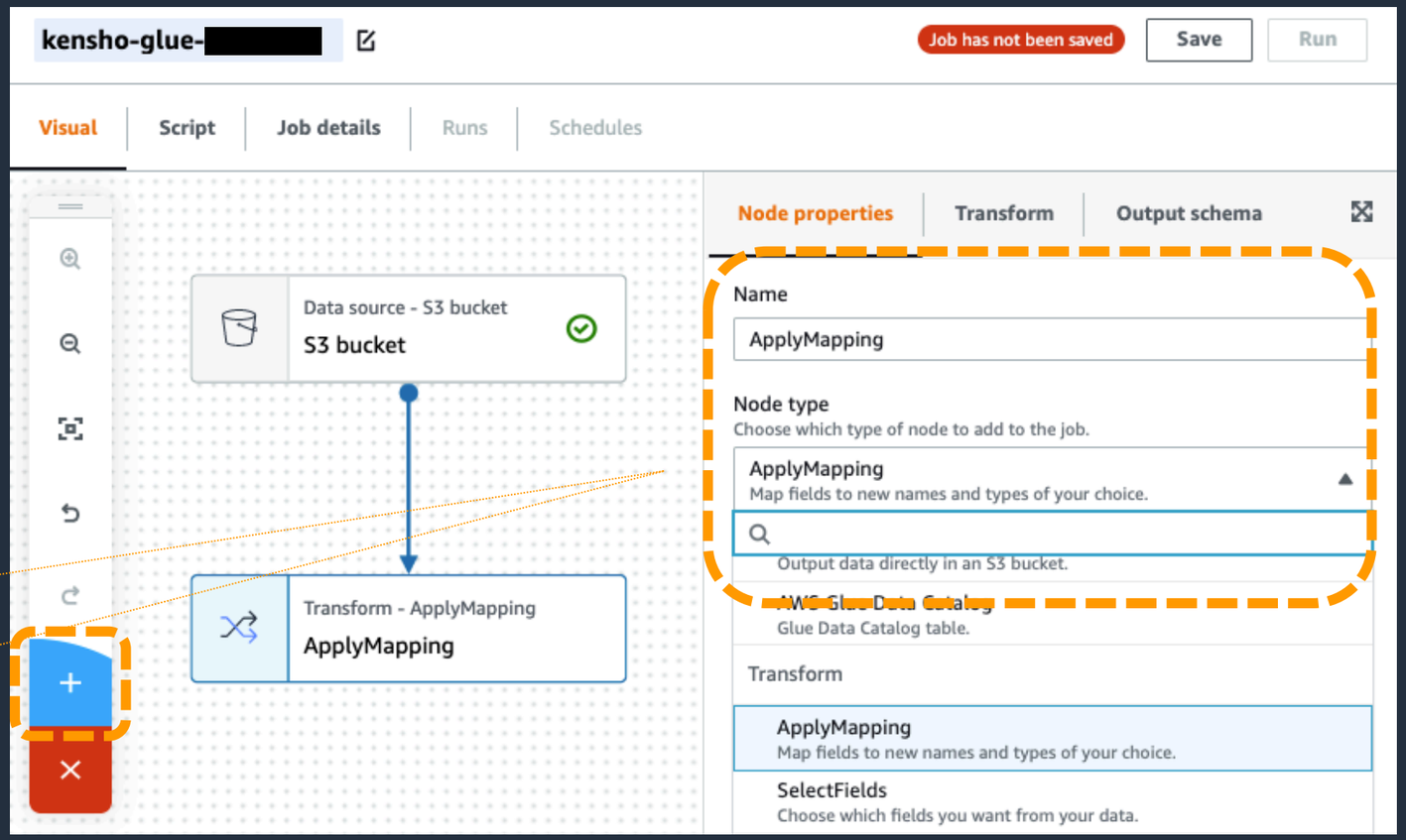
- DatabaseとTableに、
  Glue のクローラによって生成された
  データカタログを選択

S3 locationを選択し、
バケットを直接指定してもOK



データの整形①

- S3 bucket ノードを選択した状態で、画面左下の“＋”を押下
- Node typeにApplyMappingを選択



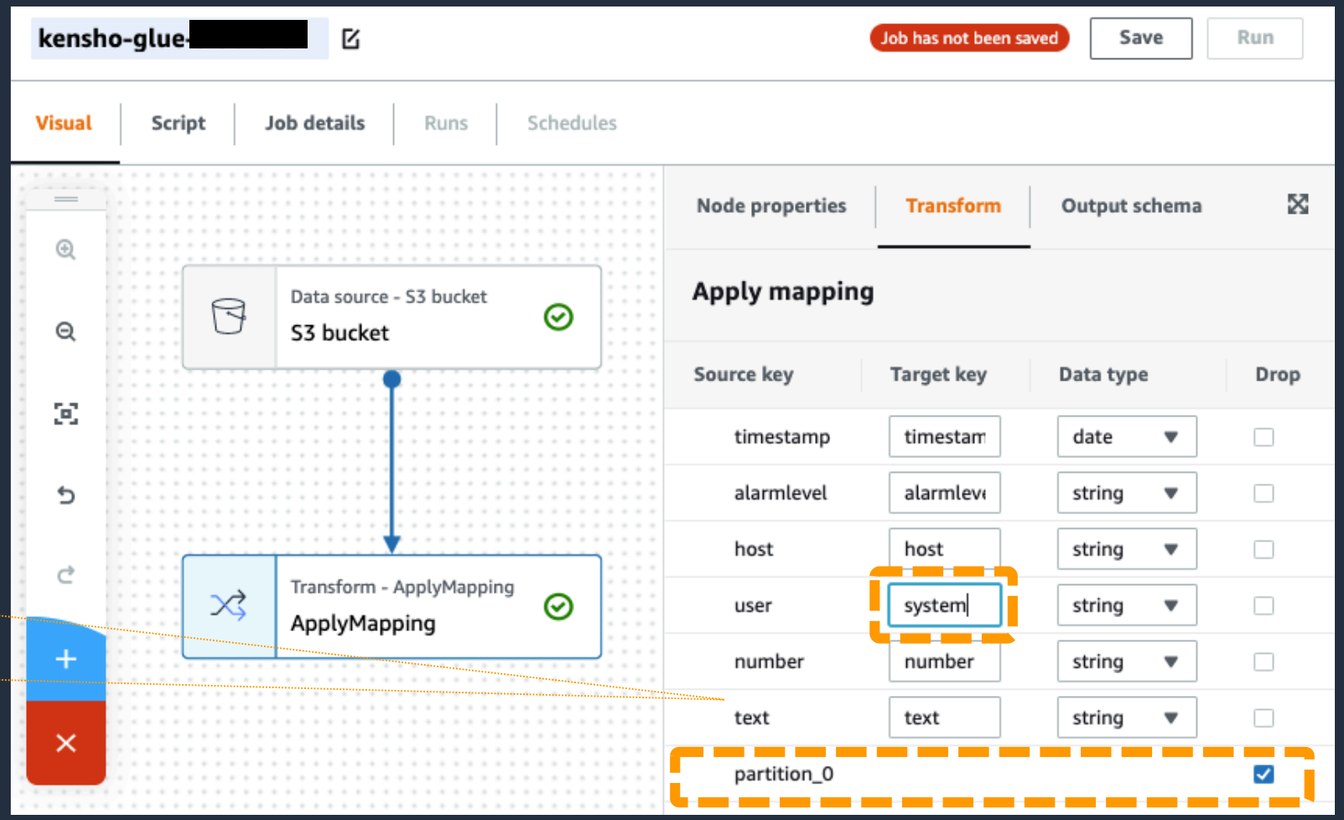
ApplyMappingは、入力データのカラムの修正・削除を行うためのノード。

他にもフィールドの分割や、ソースデータの結合等、様々な変換ノードが用意されている。

データの整形②

- Target key列に入力されている文字列を修正
- Drop列のボックスにチェック

ここでは以下の２点を実施している。
1. user 列の名前をsystem列に変更している。
2. partition_0列を削除している。

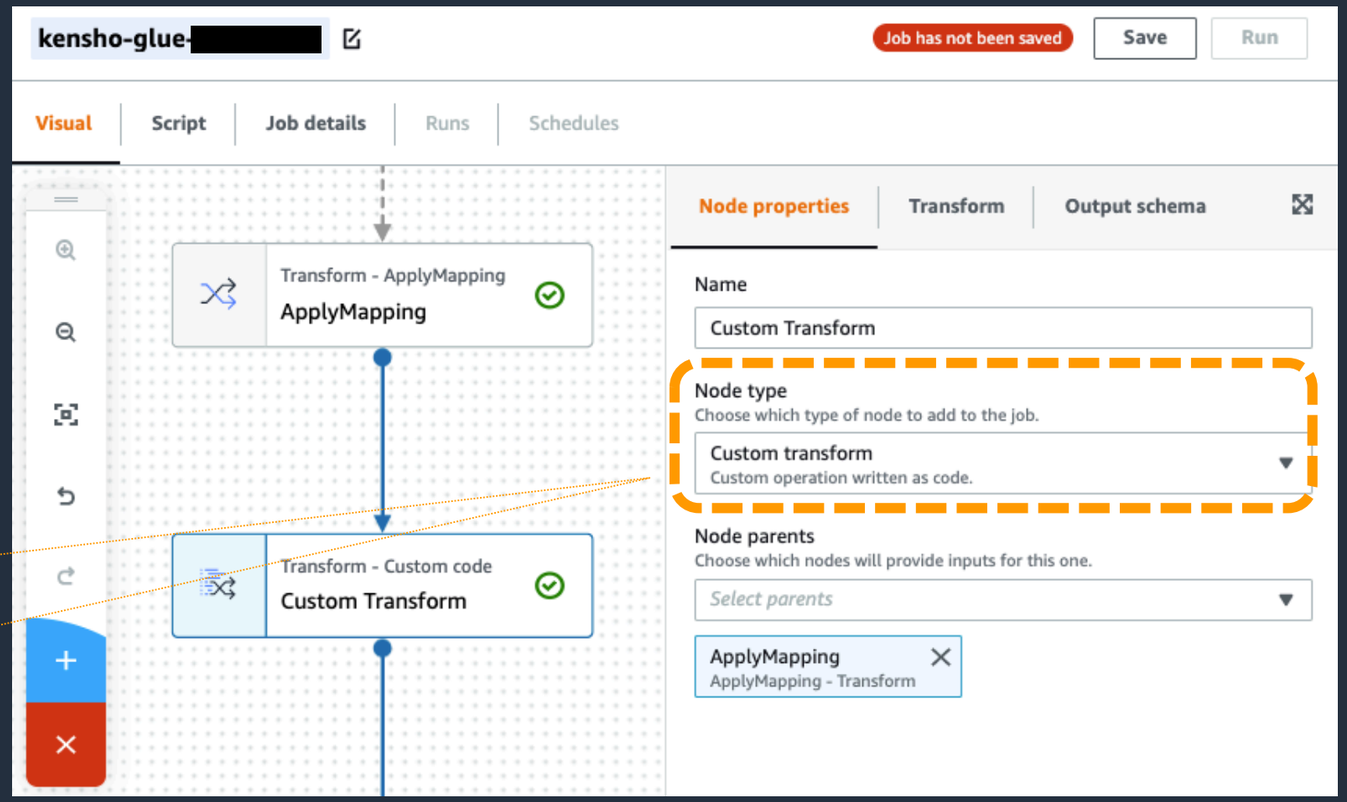


データファイルのコンパクション①

- ApplyMapping ノードを選択した状態で、画面左下の“＋”を押下
- Node typeにCustom transformを選択

Custom transformは、
コードを自由に記述することが可能なノード。

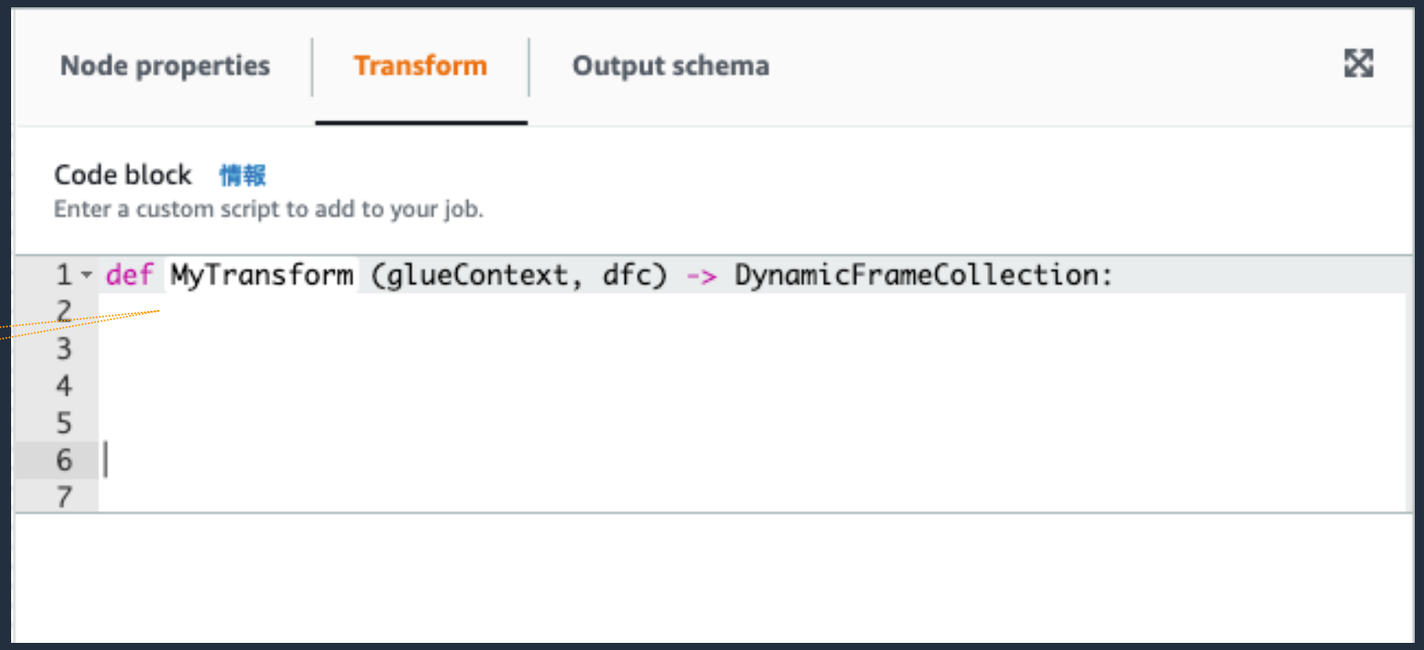
Python, Scala, Javaの、任意のコードを使用可能。



データファイルのコンパクション②

- コードブロックに下記記載の、Pythonコードを入力

```
selected = dfc.select(list(dfc.keys())[0]).toDF()
reprep = selected.repartition(5)
results = DynamicFrame.fromDF(reprep, glueContext, "results")
return DynamicFrameCollection({"results": results}, glueContext)
```



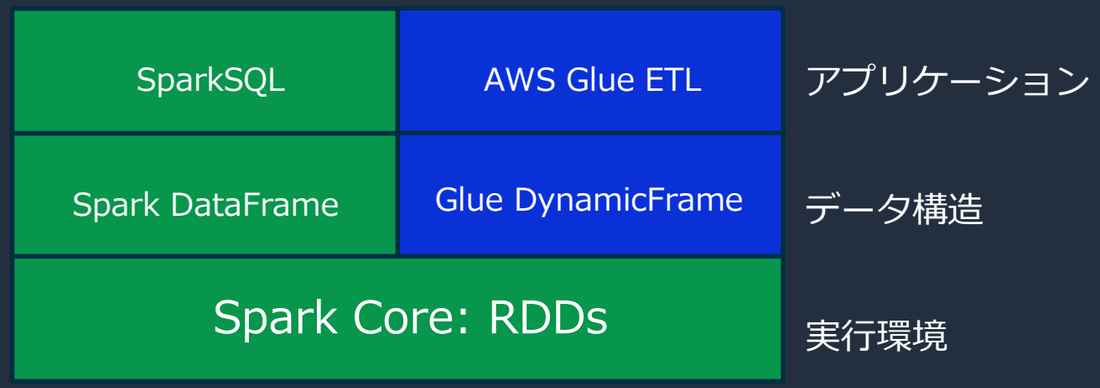
このコードブロックは、DynamicFrameオブジェクトを受け取っている。受け取ったオブジェクトに対し任意の処理を記述可能。

DynamicFrameとは

SparkSQL DataFrameと似たGlue特有の抽象化の概念

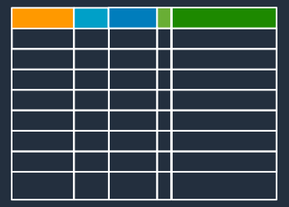
- SparkSQL DataFrameとの違いはETLに特化しているかどうか
  (DynamicFrameはスキーマの不一致を明示的にエンコードする”Schema on the Fly”を採用)
- 複数の型の可能性を残して、後で決定できるようにする (Choice型)
- DynamicFrameはデータ全体を表し、DynamicRecordはデータ１行を指す
- DataFrameとDynamicFrame間でそれぞれ変換することができる(fromDF関数・toDF関数)
- Pythonライブラリ PandasのDataFrameとは異なるので注意

アーキテクチャ：SparkおよびGlueライブラリ



データ構造イメージ

SparkSQL DataFrame



構造テーブルに類似

DynamicFrame

半構造テーブルに類似

aws

コードの内容

Node properties | Transform | Output schema

Code block 情報

Enter a custom script to add to your job.

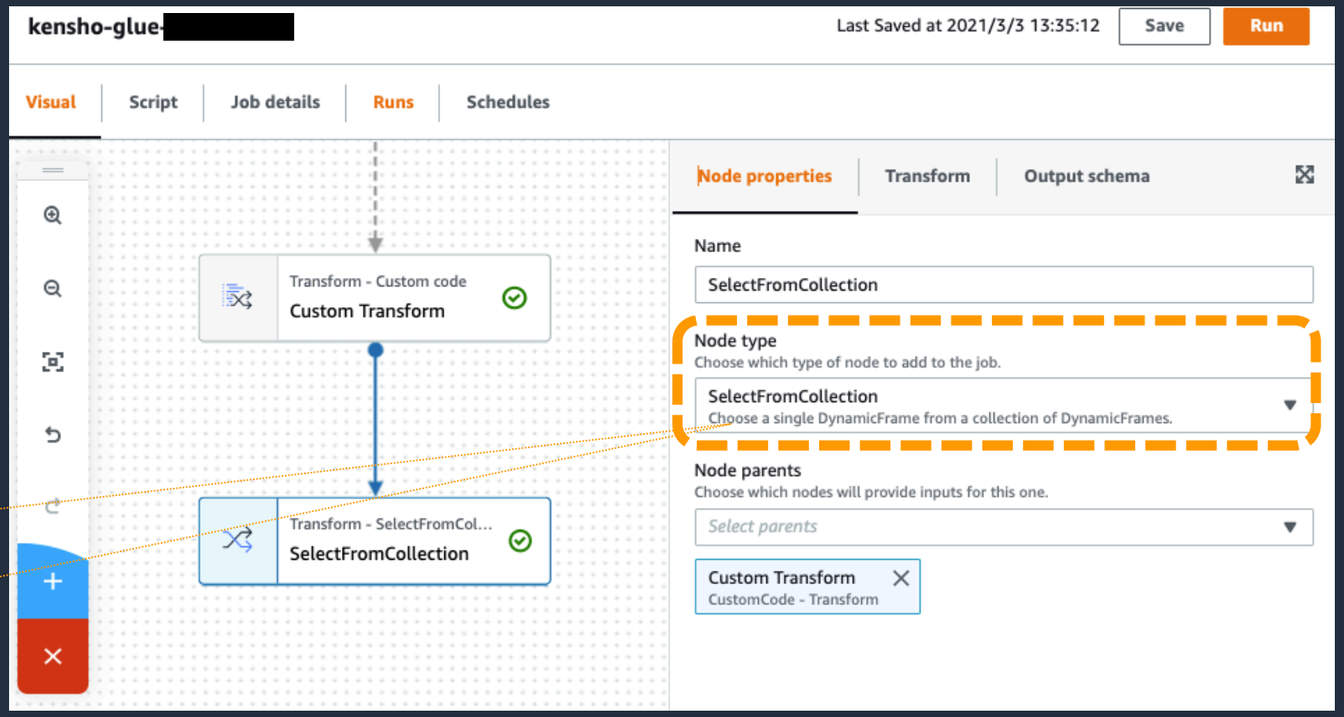
```
1 def MyTransform (glueContext, dfc) -> DynamicFrameCollection:
2     selected = dfc.select(list(dfc.keys())[0]).toDF()
3     reprep = selected.repartition(5)
4     results = DynamicFrame.fromDF(reprep, glueContext, "results")
5     return DynamicFrameCollection({"results": results}, glueContext)
6
```

上記のコードは受け取ったDynamicFrameオブジェクトに以下の内容を実施している。
1. 受け取ったDynamicFrameオブジェクトを、toDF()関数を使用し、SparkSQL DataFrameに変換
2. DataFrameの分割数を変更し、5分割に変更
3. 変更後のDataFrameをfromDF()関数を使用し、DynamicFrameオブジェクトに変換
4. DynamicFrameオブジェクトをリターン

aws

データファイルのコンパクション③

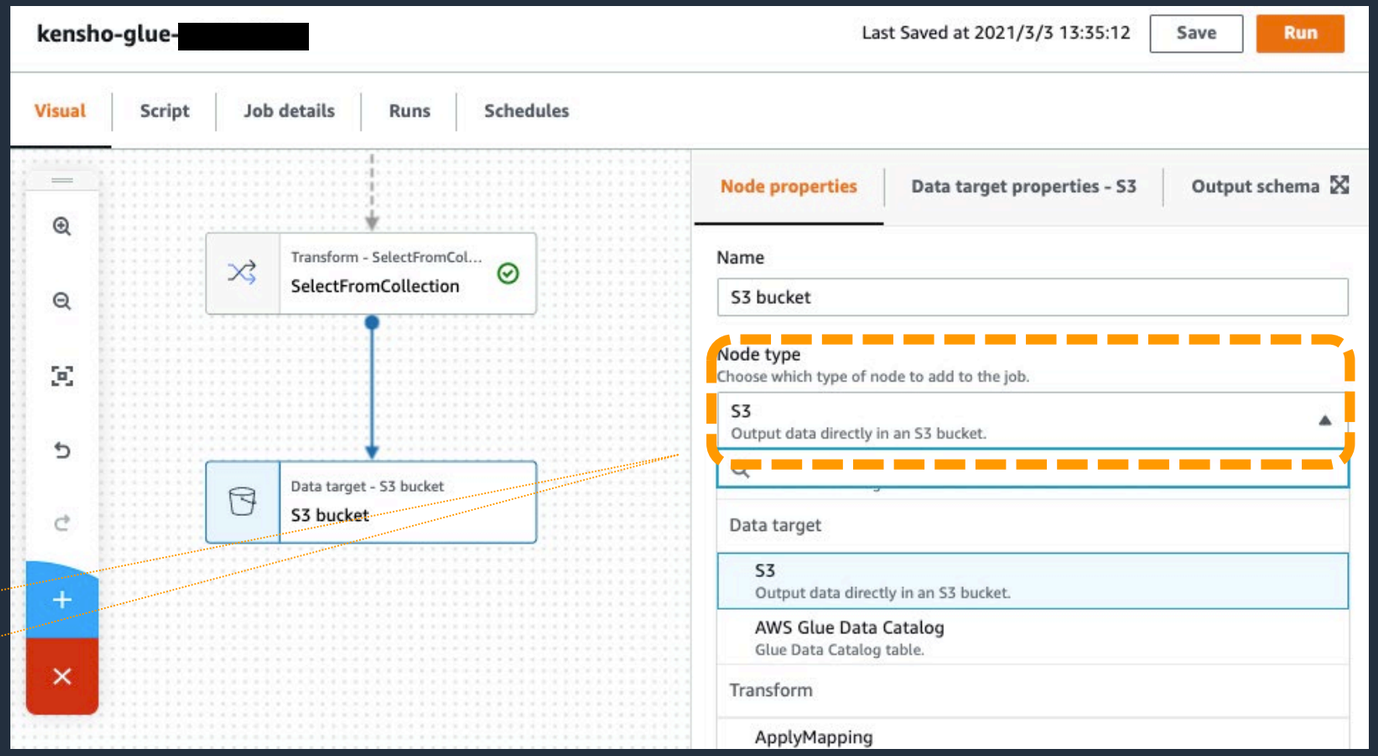
- Custom Transform ノードを選択した状態で、
  画面左下の“＋”を押下
  （自動生成の場合この手順は不要）
- Node typeにSelectFromCollectionを選択



SelectFromCollectionは、Custom Transform
ノードで実行したコードからオブジェクトを受け
取るためのノード。
Custom Transformノードとセットで使用する。

フォーマット変換①

- SelectFromCollection ノードを選択した状態で、画面左下の“＋”を押下
- Node typeにS3を選択



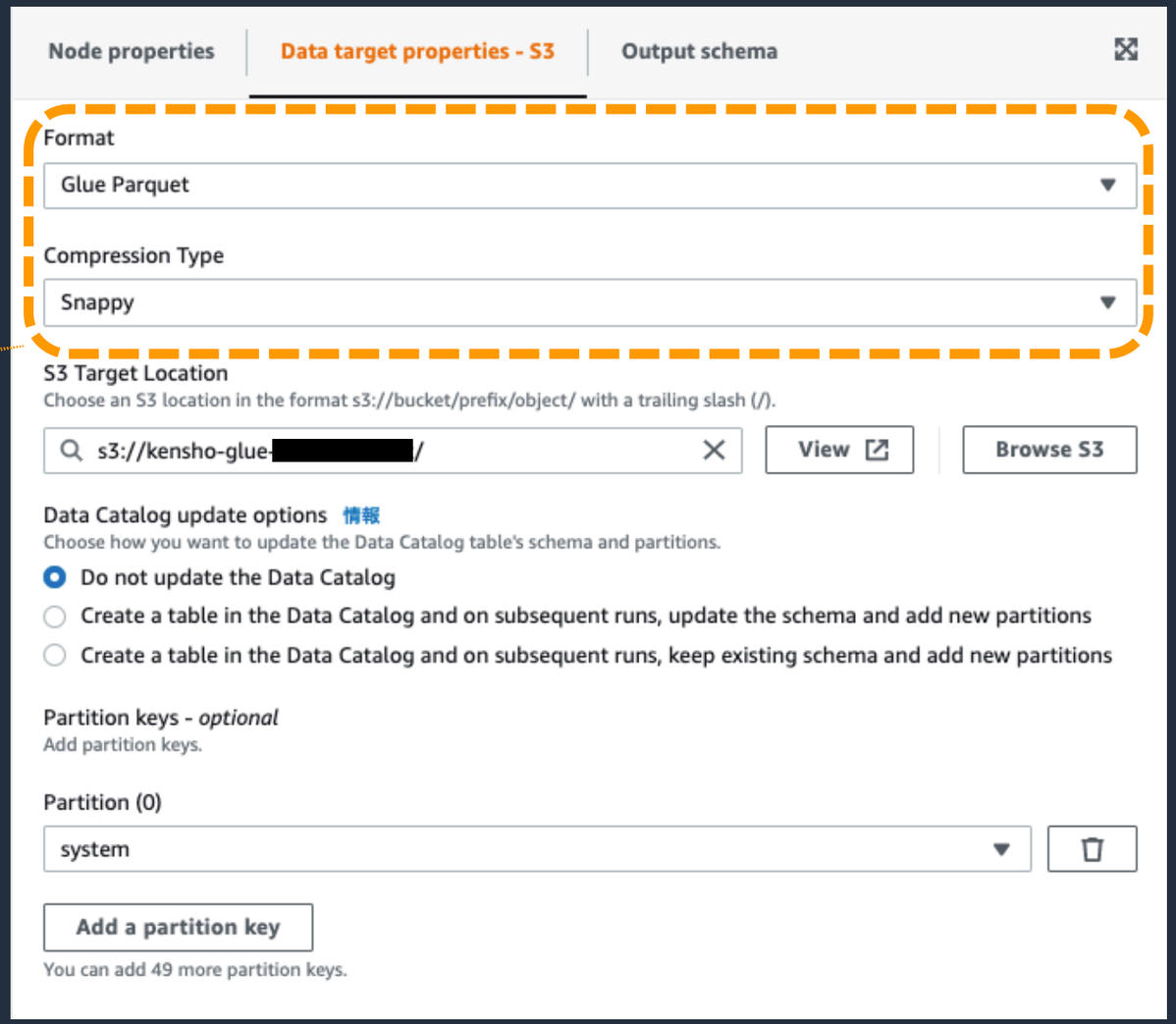
Node typeには、Data targetのS3とData sourceのS3があるので注意する。

フォーマット変換②

- FormatをGlue Parquetを選択
- Compression TypeにSnappyを選択

Parquet とは
Parquetは分析に適した列指向フォーマット

Glue Parquet とは
よりGlueに最適化されたParquetフォーマット
通常のParquetと変わりないが、出力ファイルのスキーマを動的に計算し、高速に”Parquet”ファイルに書き込むことが可能



出力先の指定

- S3 Target Loctionに出力先となる、S3バケットを指定

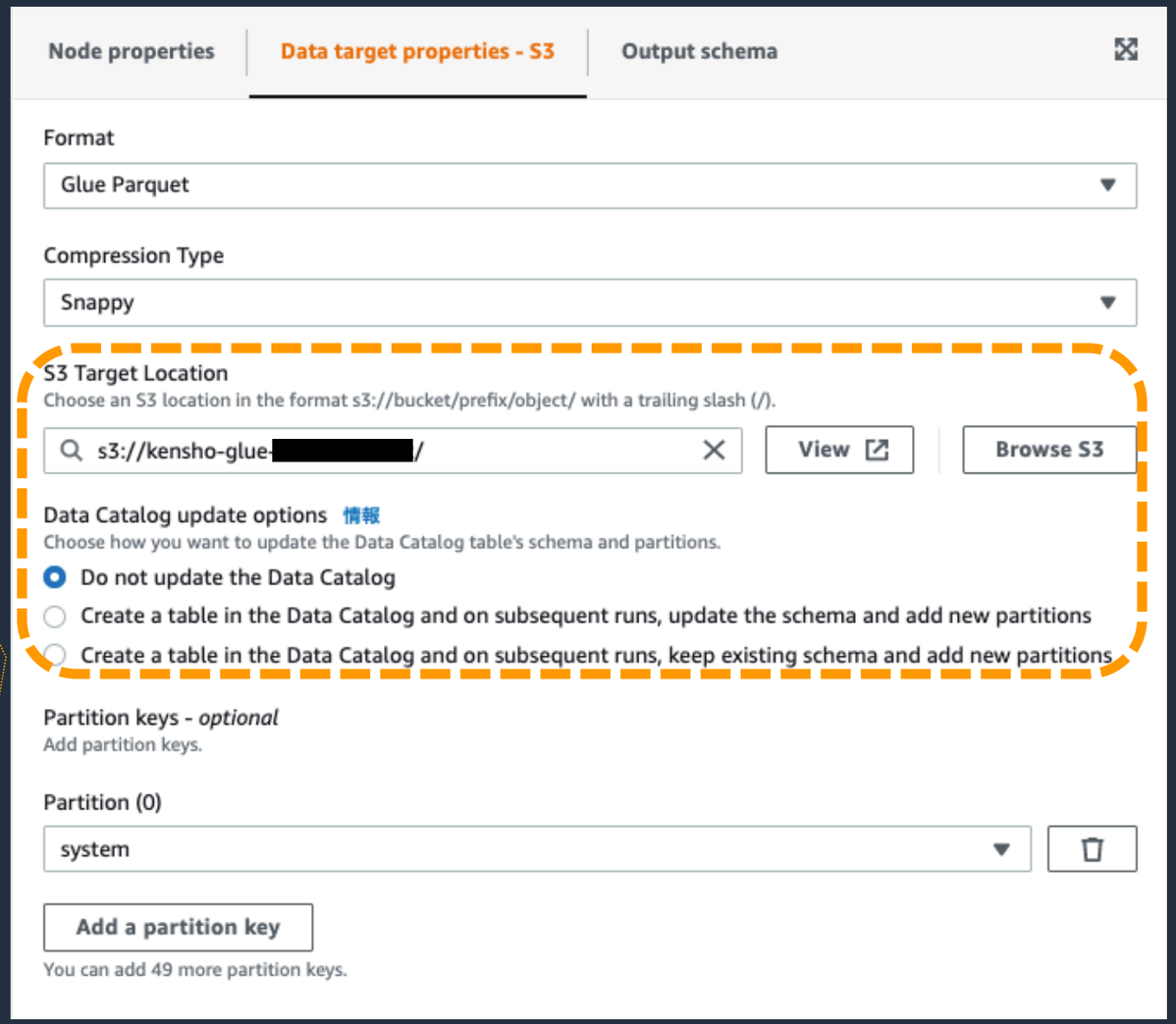
- Data Catalog update optionsにて、Do not update the Data Catelogを選択

Data Catalog update options について

Do not change table definition:
ジョブでDataCatalogの更新を行わない設定。

Update schema and add new partitions :
スキーマの変更または新しいパーティションの追加時にジョブでData Catalogを更新する設定。

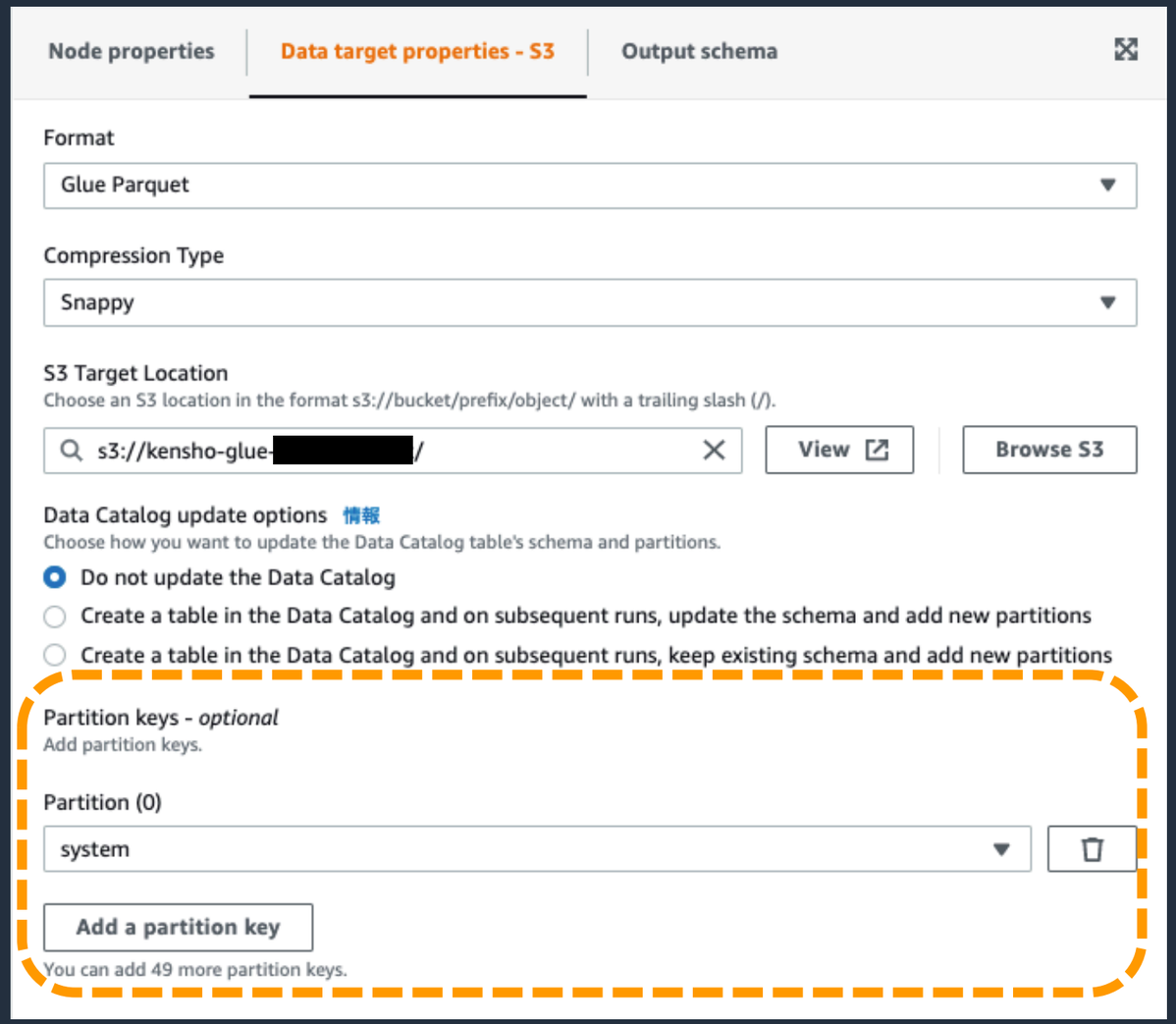
Keep existing schema and add new partitions :
新しいパーティションを追加する目的にのみ、ジョブでDataCatalogを更新する設定。



パーティショニング

- Partition keys にて、
  パーティショニングに使用するカラムを選択

Partitionに使用するカラムは複数選択が可能。
複数選択することで階層構造でファイルが出力される。



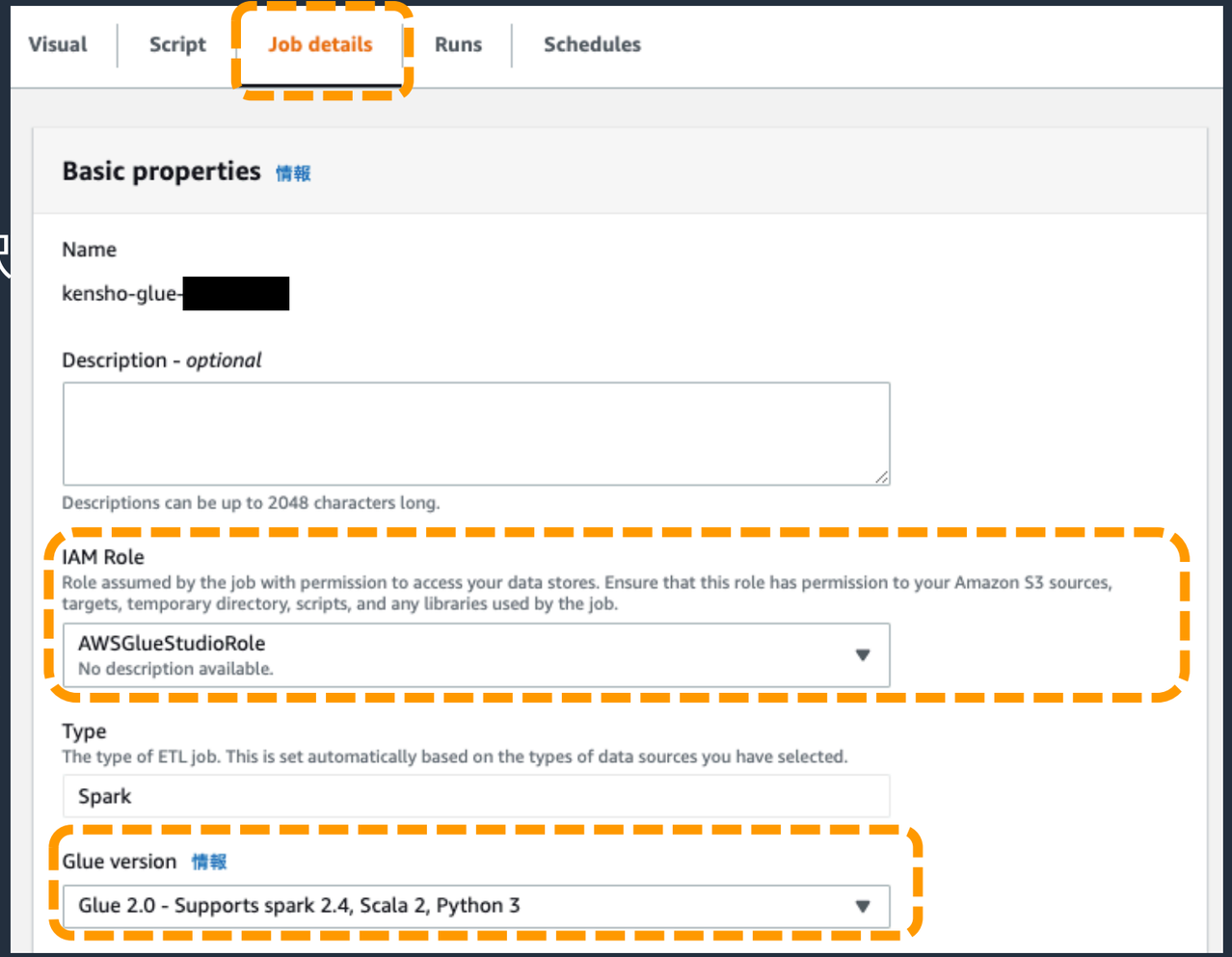
JOBの詳細設定①

- Job detailsタブを選択
- IAM Role に必要な権限を持ったロールを選択
- Glue version にて、Glue 2.0が選択されていることを確認

必要な権限について
ユースケース1で必要となる権限は以下。

- AWS Glue サービスのアクセス許可
- Amazon CloudWatch のアクセス許可
- データソースとデータターゲットへのアクセス許可

これらの権限が付与されたロールを選択する。



JOBの詳細設定②

- Language にPython 3が選択されていることを確認
- Worker type にてG.1Xを選択
- Number of workersをデフォルトの10から2に変更

Worker typeの詳細については次のスライドをご参照

Worker Type

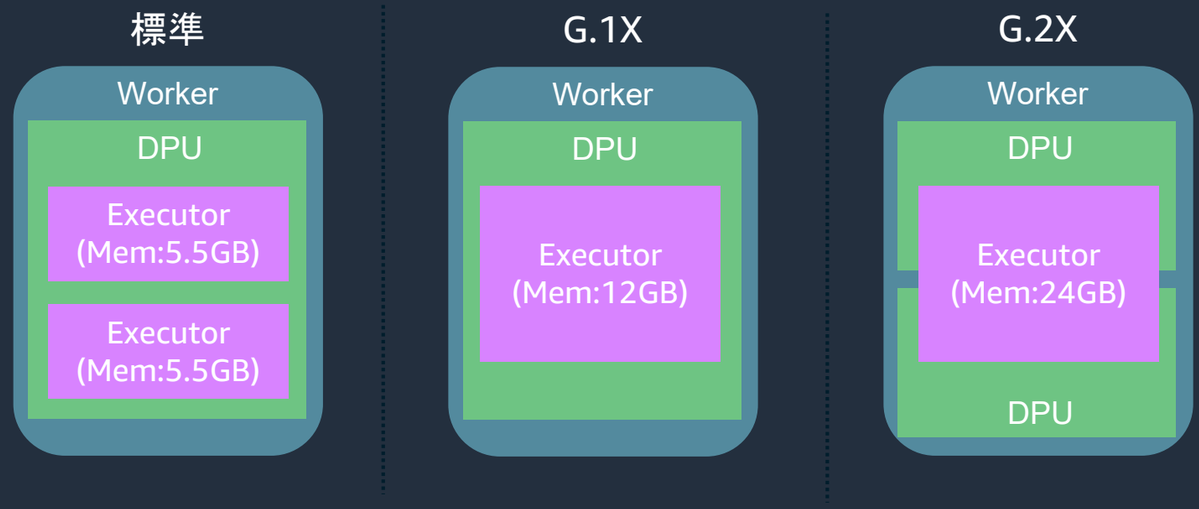
ジョブ実行環境にメモリ大量使用ワークロード向けのWorker Typeを選択可能

- ジョブ実行時に割り当てる処理能力をDPU(Data Processing Unit)という
  1DPU = 4vCPU、16GBメモリ
- Glue Studioでは、G.1xとG.2Xが選択可能

Worker Type一覧

Worker Type	DPU数 /1Worker	Executor数 /1Worker	メモリ数 /1Executor
標準	1	2	5.5GB
G.1X	1	1	12GB
G.2X	2	1	24GB

Worker Type構成イメージ

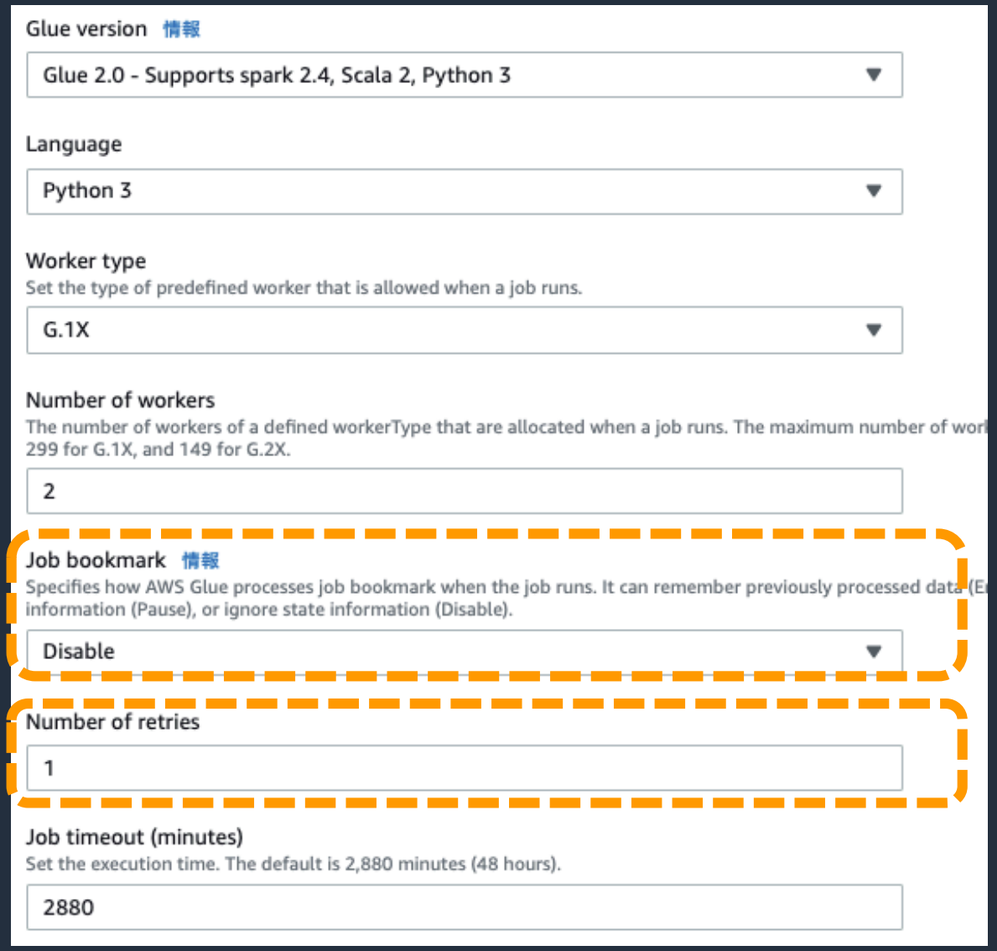


参考URL：Spark Components(https://spark.apache.org/docs/latest/cluster-overview.html)

JOBの詳細設定③

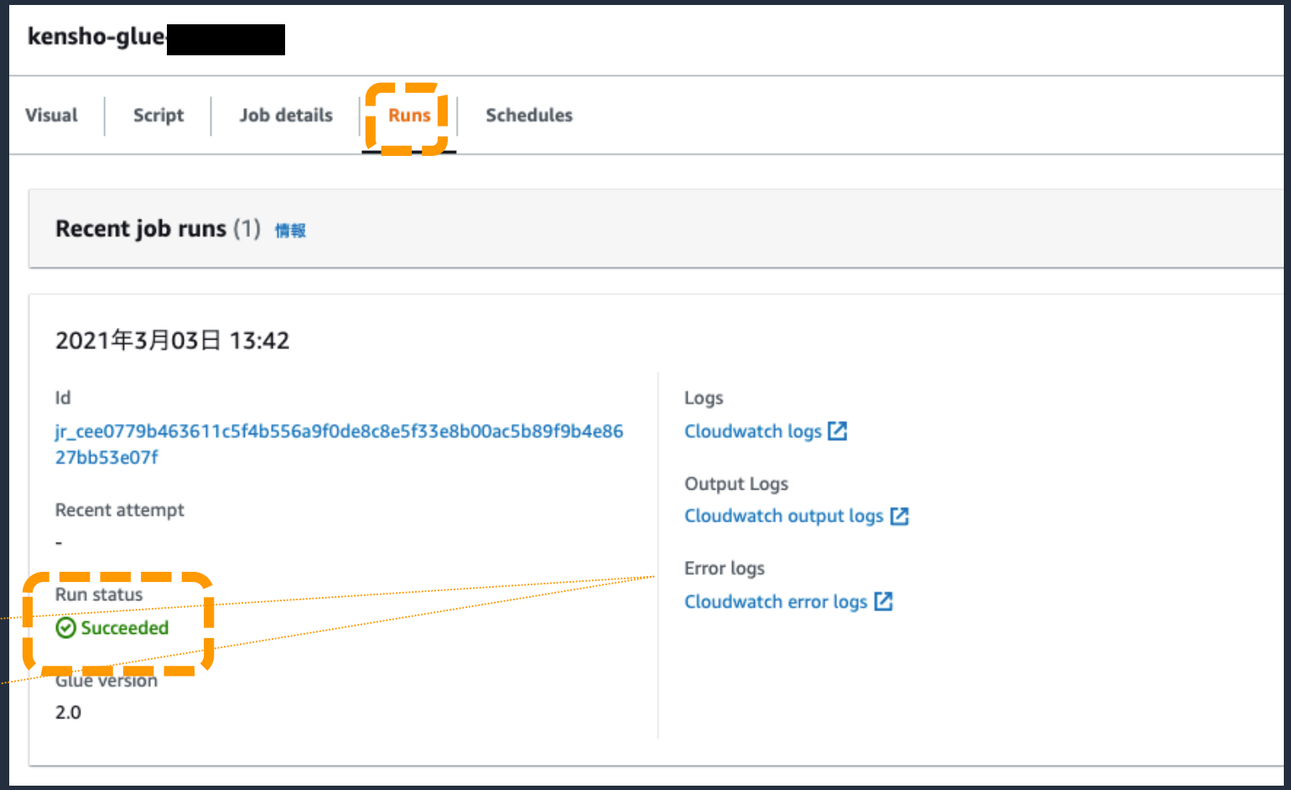
- Job bookmark にてDisableを選択
- Number of retriesをデフォルトの3から1に変更

Job bookmarkについて
ジョブを実行した状態情報を保持することで、古いデータを再処理しないようにする機能。デフォルトのDisableにすることでJobの実行のたびにデータセット全体に対して処理をすることが可能。



結果のモニタリング

- 画面右上のSaveを押下した後、Runボタンを押下しJobを実行
- Runs タブを選択
- Run statusが、Succeededになったことを確認



トラブル対応について

処理が中断した場合や正常終了しなかった場合、Cloudwatchにログが出力されているのでそこで確認が可能。

実行結果

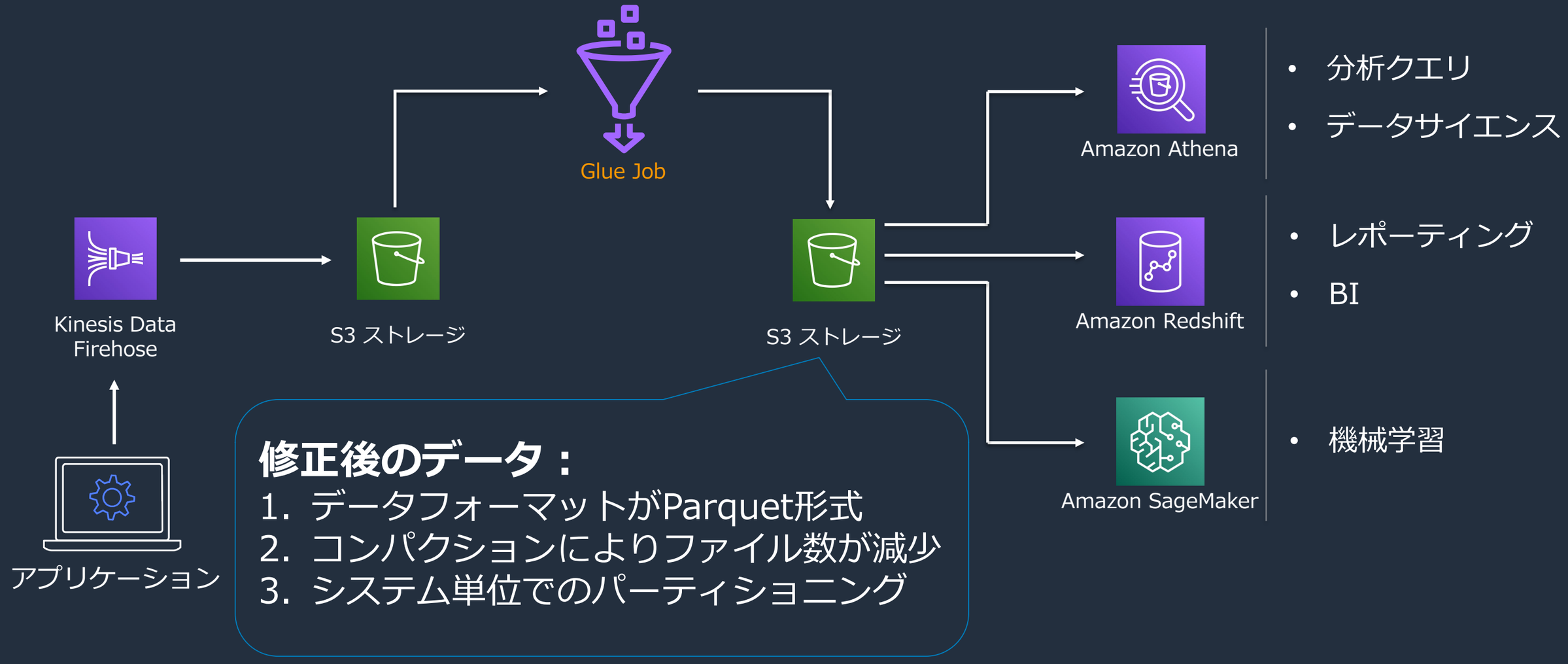
- 出力バケット直下にシステム名でプレフィックスが作成されていることを確認

名前	タイプ	最終更新日時
system=SystemA/	フォルダ	-
system=SystemB/	フォルダ	-
system=SystemC/	フォルダ	-
system=SystemD/	フォルダ	-
system=SystemE/	フォルダ	-
system=SystemF/	フォルダ	-

- 各プレフィックス毎にParquet形式の５つのファイルが出力されていることを確認

名前	タイプ	最終更新日時
run-DataSink0-1-part-block-0-0-r-00000-snappy.parquet	parquet	2021/03/03 01:43:31 PM JST
run-DataSink0-1-part-block-0-0-r-00001-snappy.parquet	parquet	2021/03/03 01:43:31 PM JST
run-DataSink0-1-part-block-0-0-r-00002-snappy.parquet	parquet	2021/03/03 01:43:31 PM JST
run-DataSink0-1-part-block-0-0-r-00003-snappy.parquet	parquet	2021/03/03 01:43:31 PM JST
run-DataSink0-1-part-block-0-0-r-00004-snappy.parquet	parquet	2021/03/03 01:43:31 PM JST

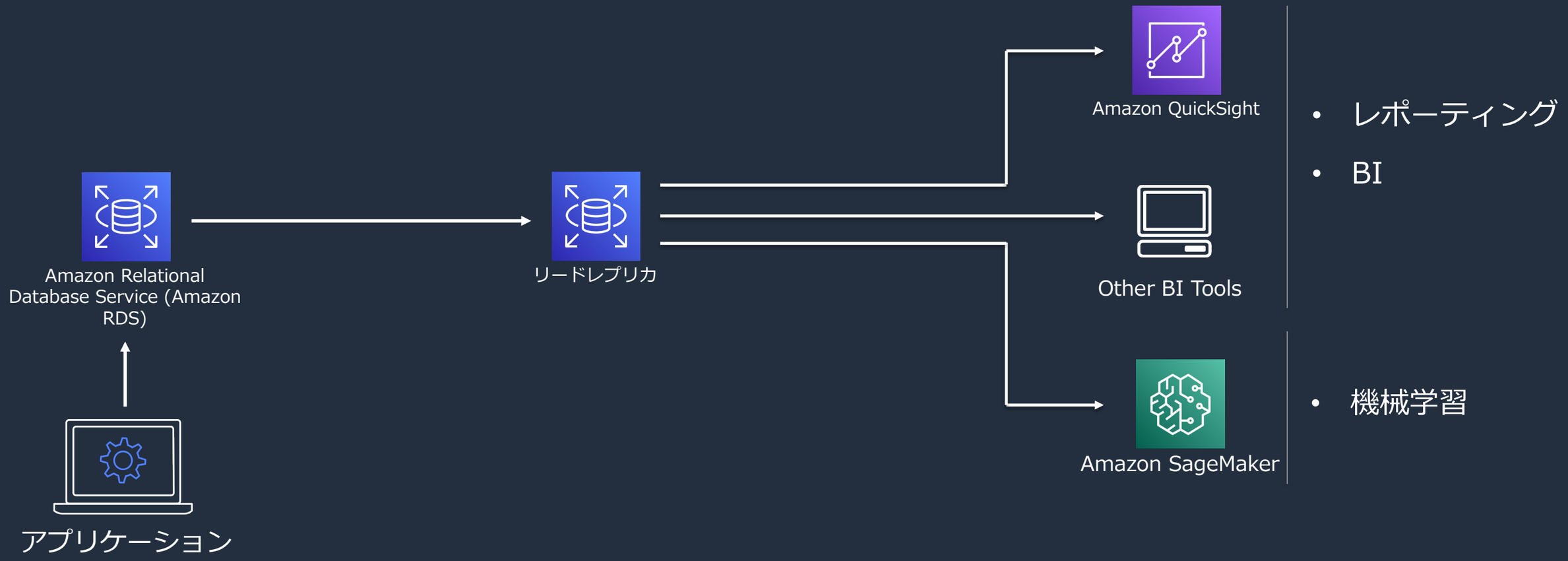
ログストリーム / IoTセンサーデータ分析の解決策



ユースケース 2：
データベースレプリカへの処理をオフロード

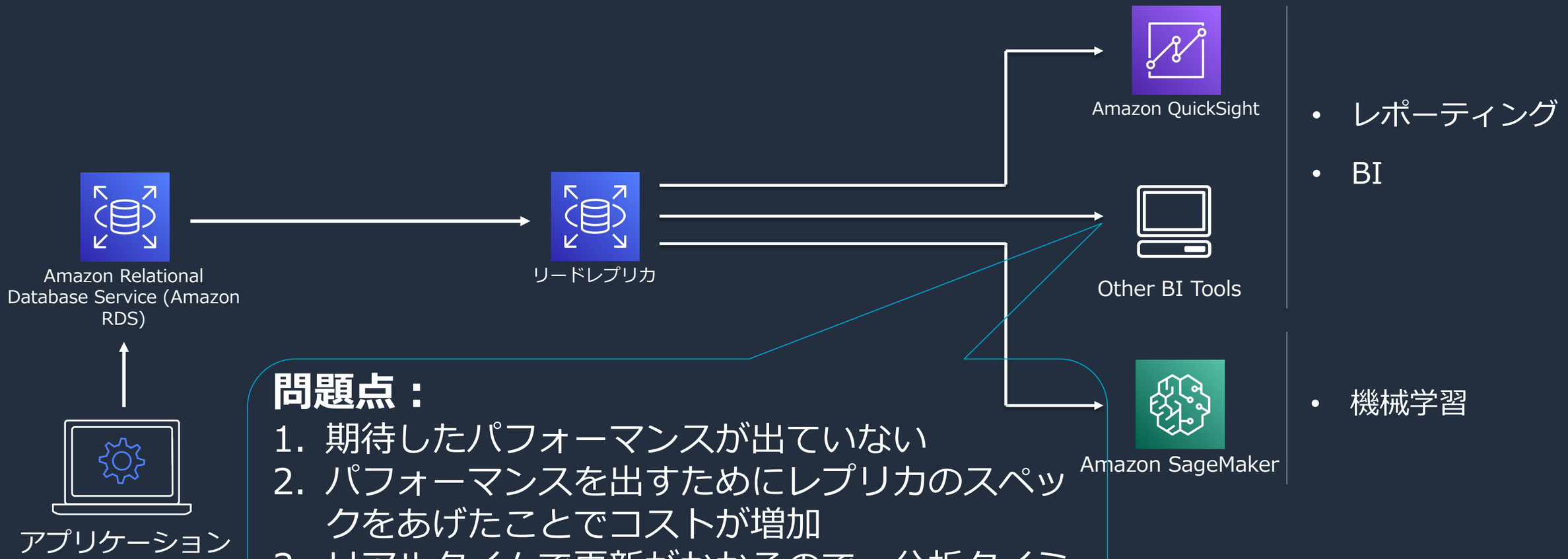
RDSにあるデータを直接分析している場合の課題

分析用のデータベースレプリカにかかる負荷をデータレイクへオフロード



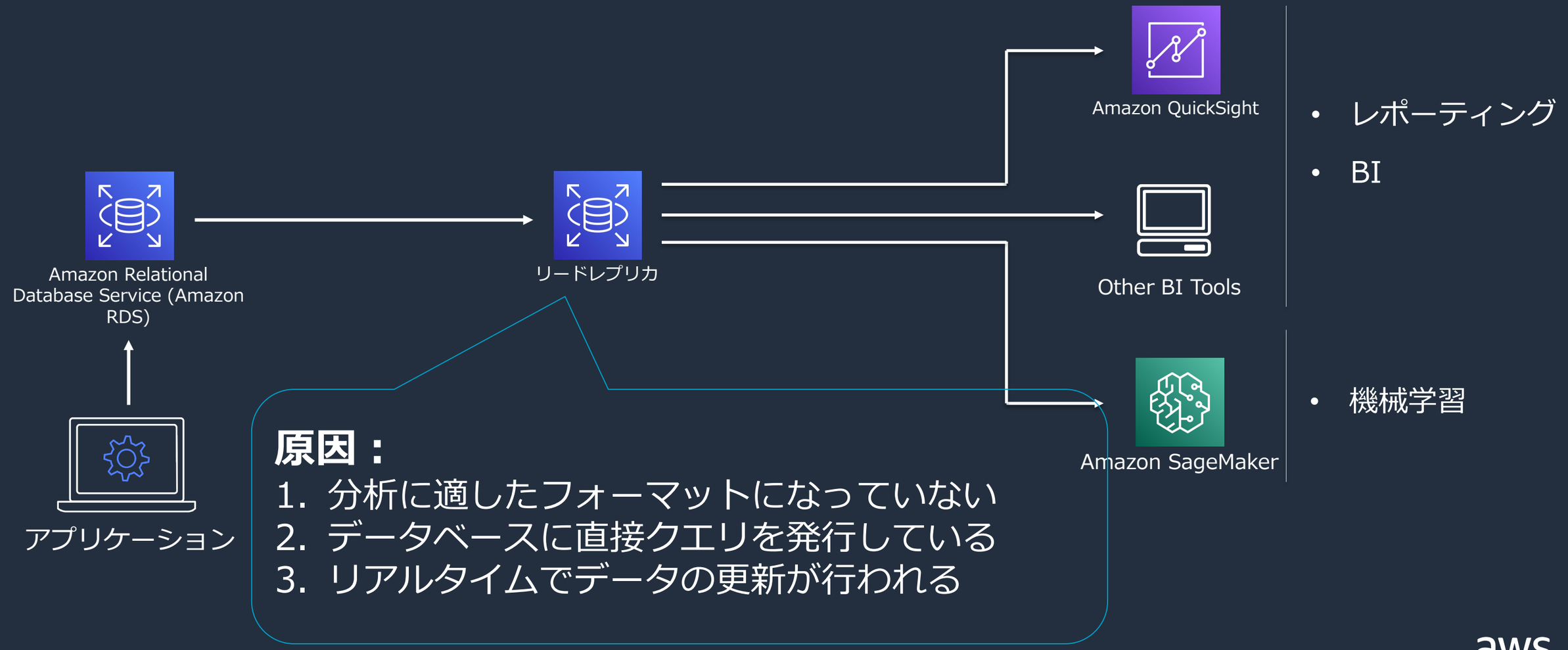
RDSにあるデータを直接分析している場合の課題

分析用のデータベースレプリカにかかる負荷をデータレイクへオフロード

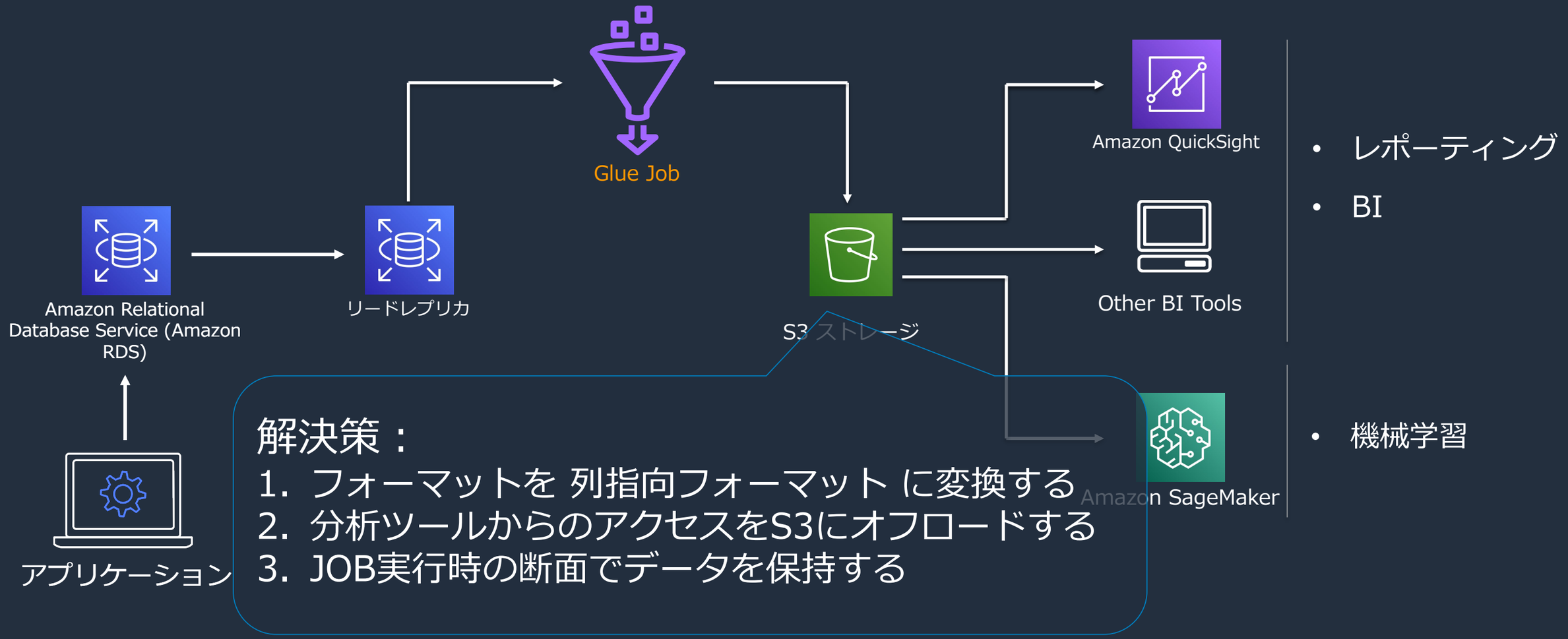


RDSにあるデータを直接分析している場合の課題

分析用のデータベースレプリカにかかる負荷をデータレイクへオフロード

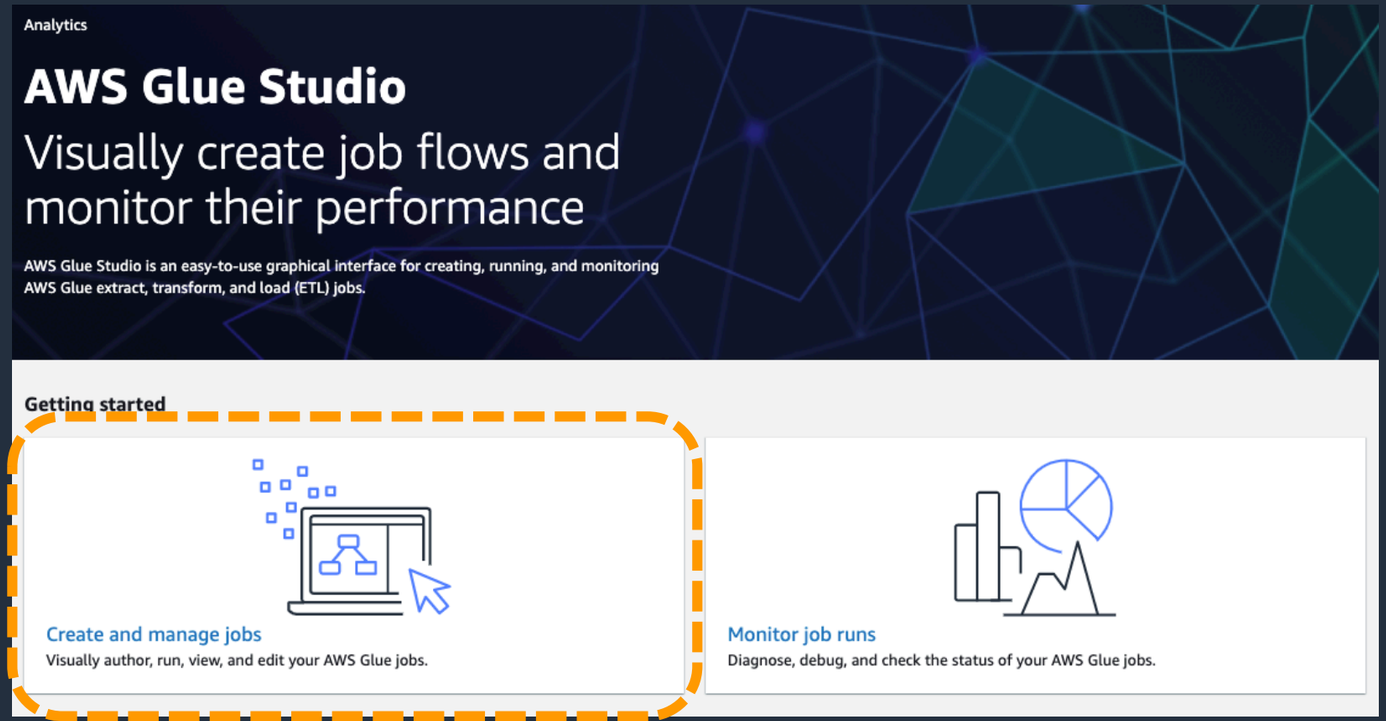


RDSにあるデータを直接分析している場合の解決策



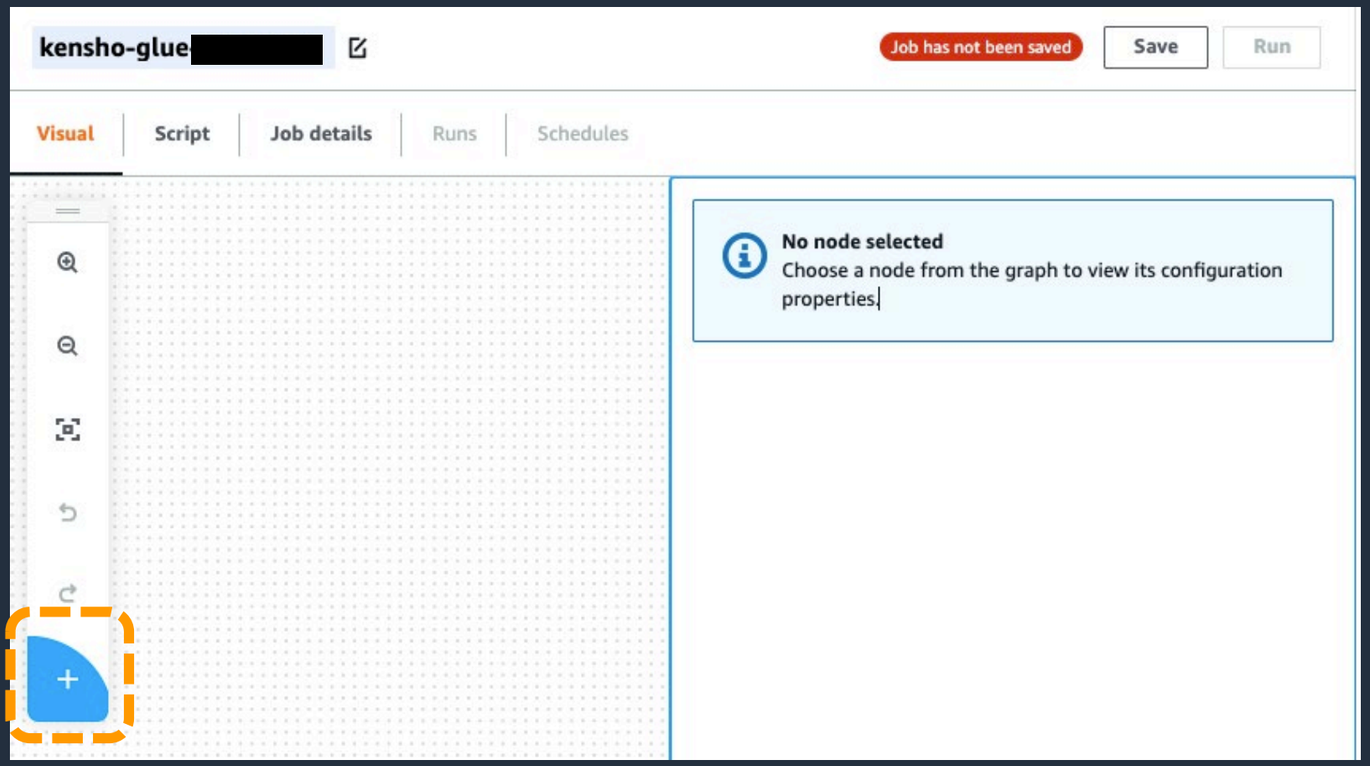
AWS Glue Studioを使用したETL JOBの作成

- マネージメントコンソールより、
  AWS Glue を選択
- AWS Glue コンソールの左ペインより、
  AWS Glue Studioを選択
- 表示されたAWS Glue Studioの画面にて
  Create and manage Jobs を選択



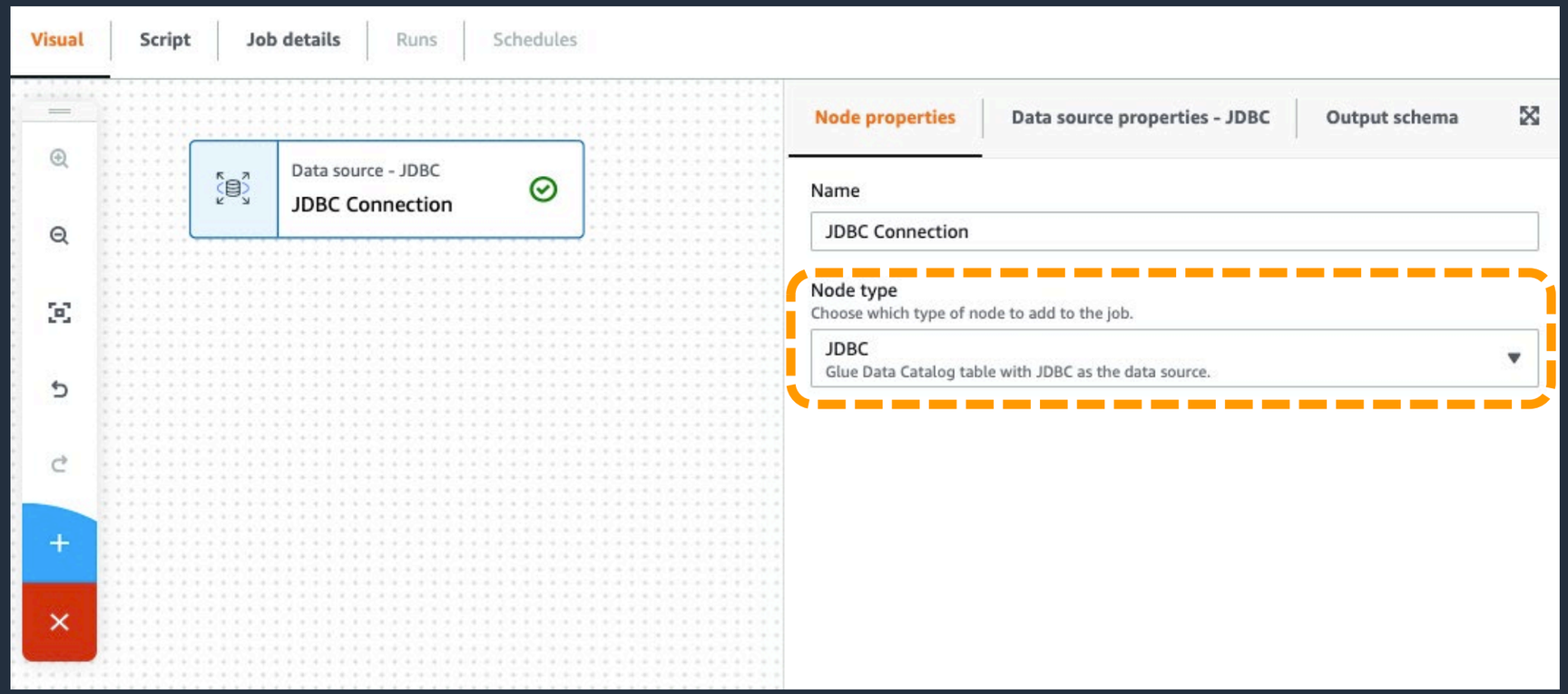
データソースの指定①

- JobのEdit画面が表示
- 左上にJob名を入力
- 画面左下の“＋”を押下し、ノードを表示



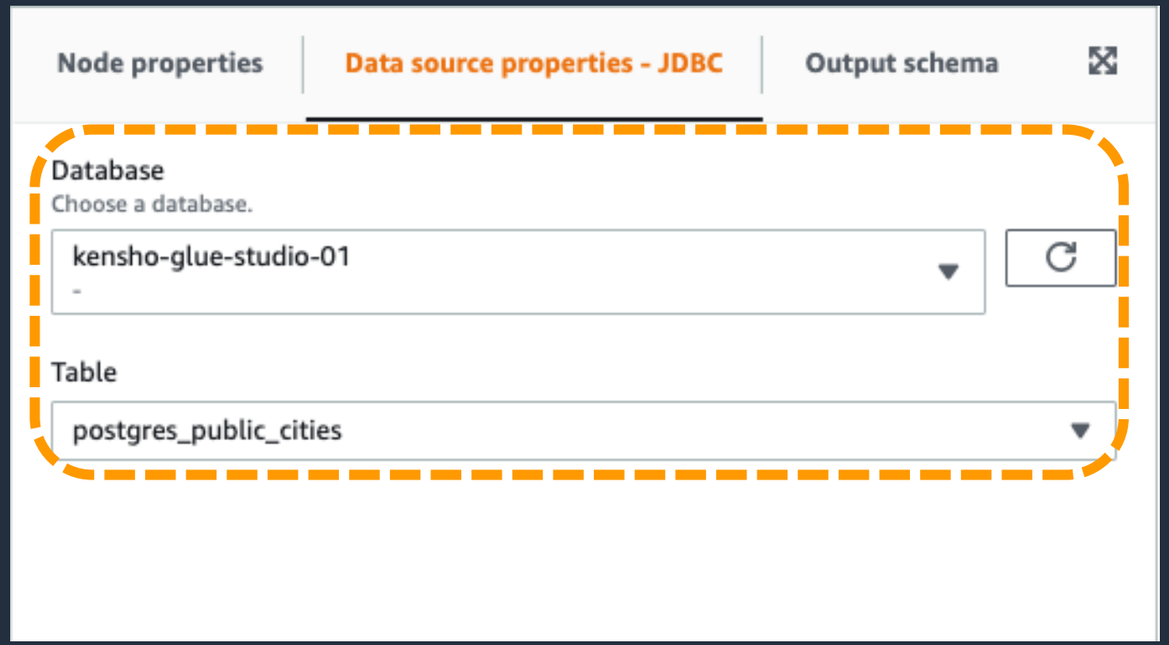
データソースの指定②

- Node typeにJDBCを選択



データソースの指定③

- DatabaseとTableに、
  Glue のクローラによって生成された
  データカタログを選択



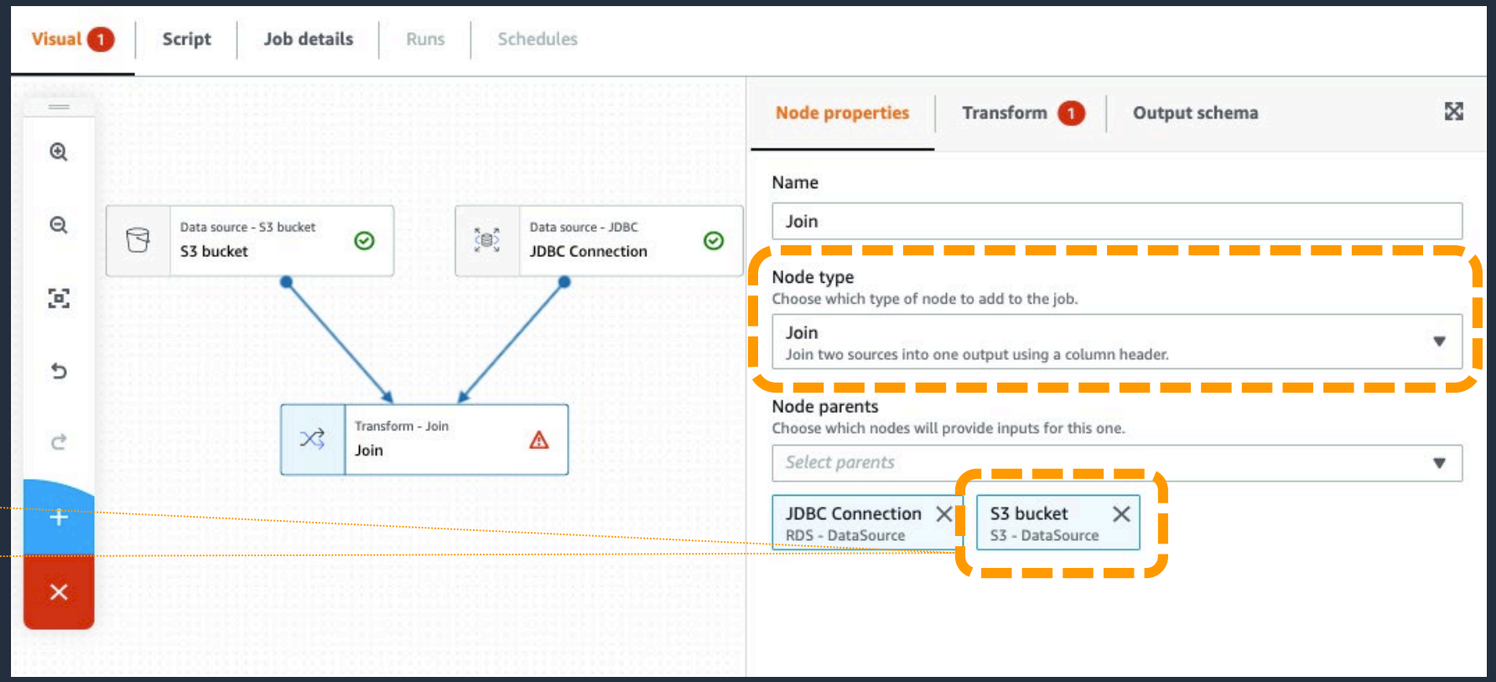
DBとの接続情報について

データカタログにはDBへの接続情報も含まれている。その為、データカタログを指定することでDBへのJDBC接続が可能となる。

外部テーブルの結合①（追加機能のご紹介）

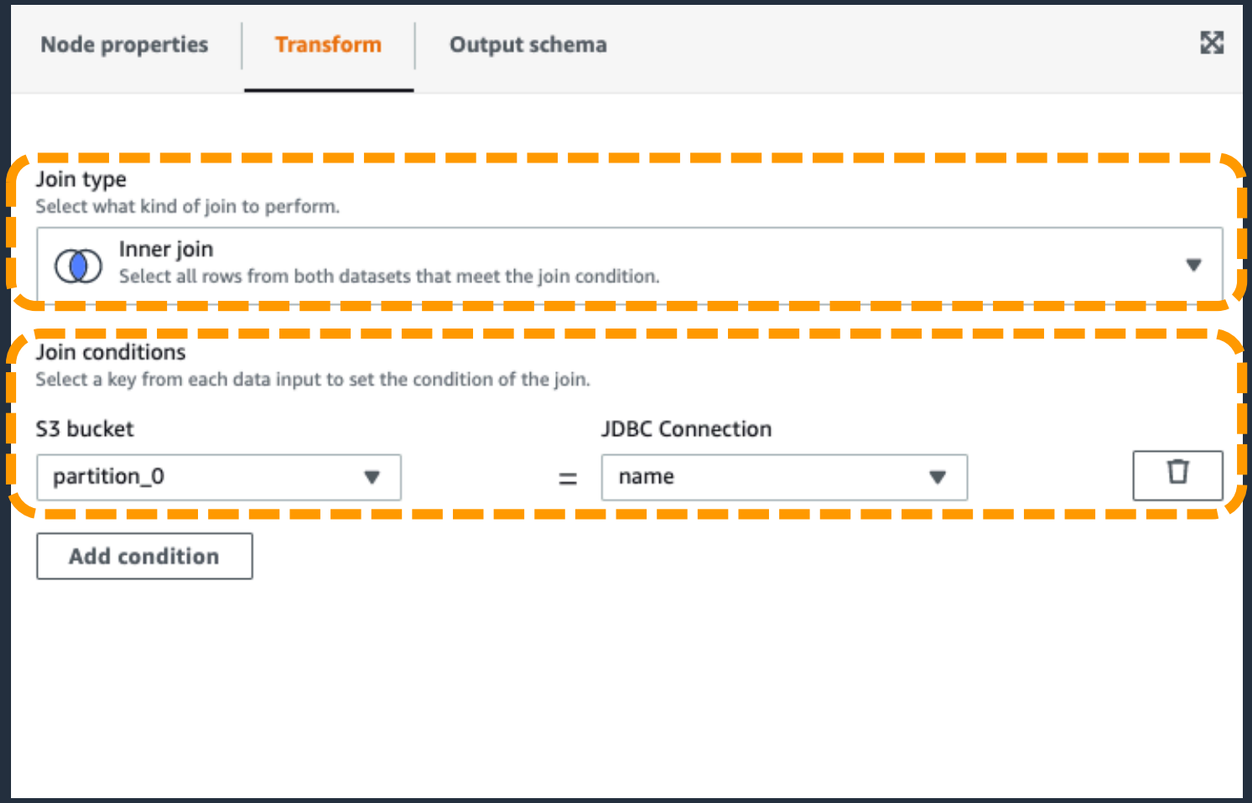
- Node typeにJoinを選択
- Node parentsに結合したい、別のデータソースを選択

結合処理は、JDBCとS3など異なるタイプのデータソース間でも結合が可能。



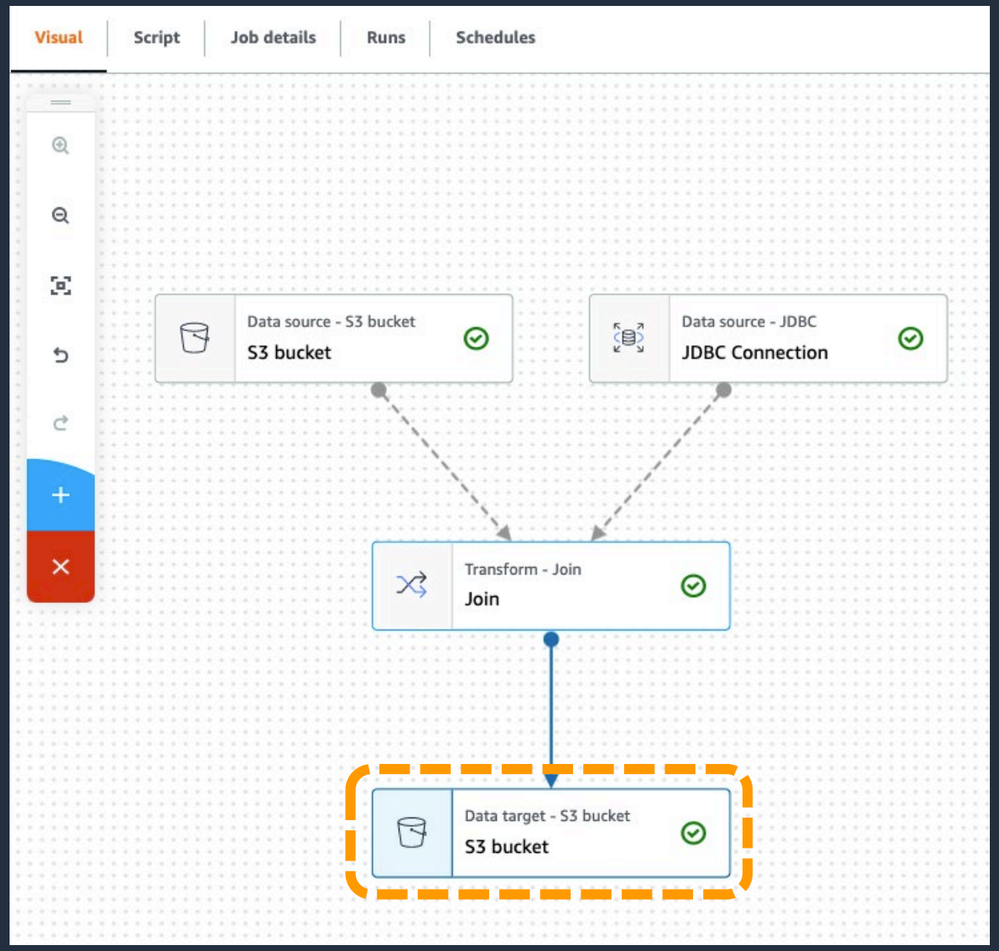
外部テーブルの結合②（追加機能のご紹介）

- Join type にInner join を選択
- Join conditions に結合したいカラムを選択



ターゲットの選択

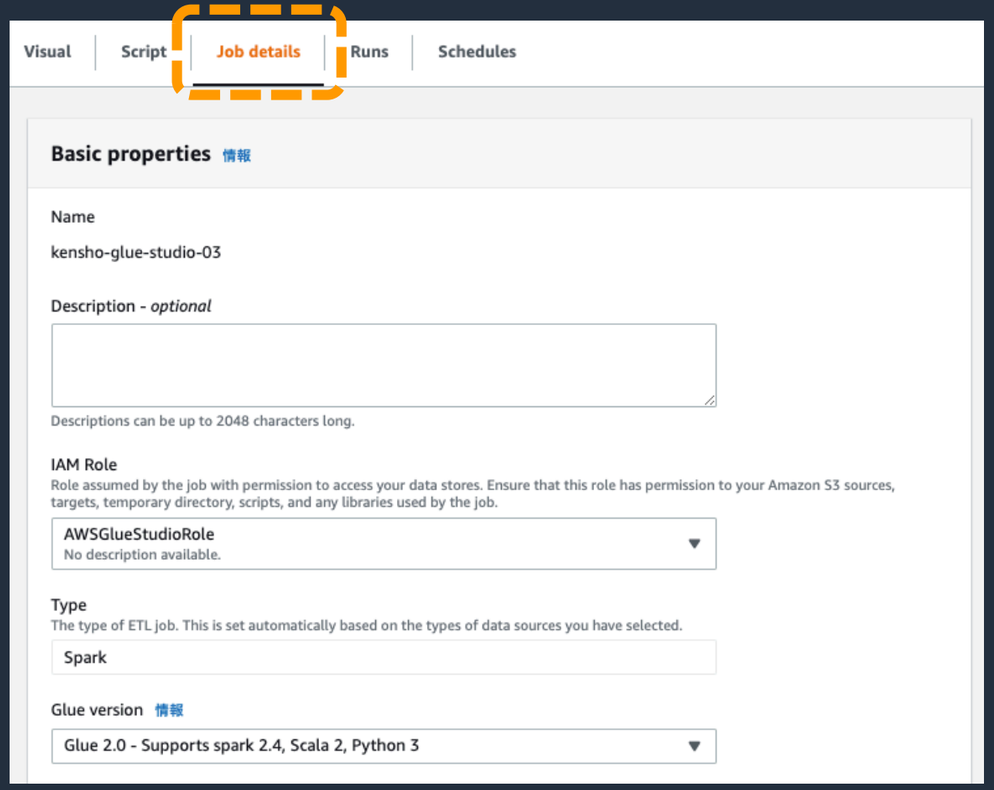
- Node typeにS3を選択
- その他の設定はユースケース1と同様の内容を実施



JOBの詳細設定

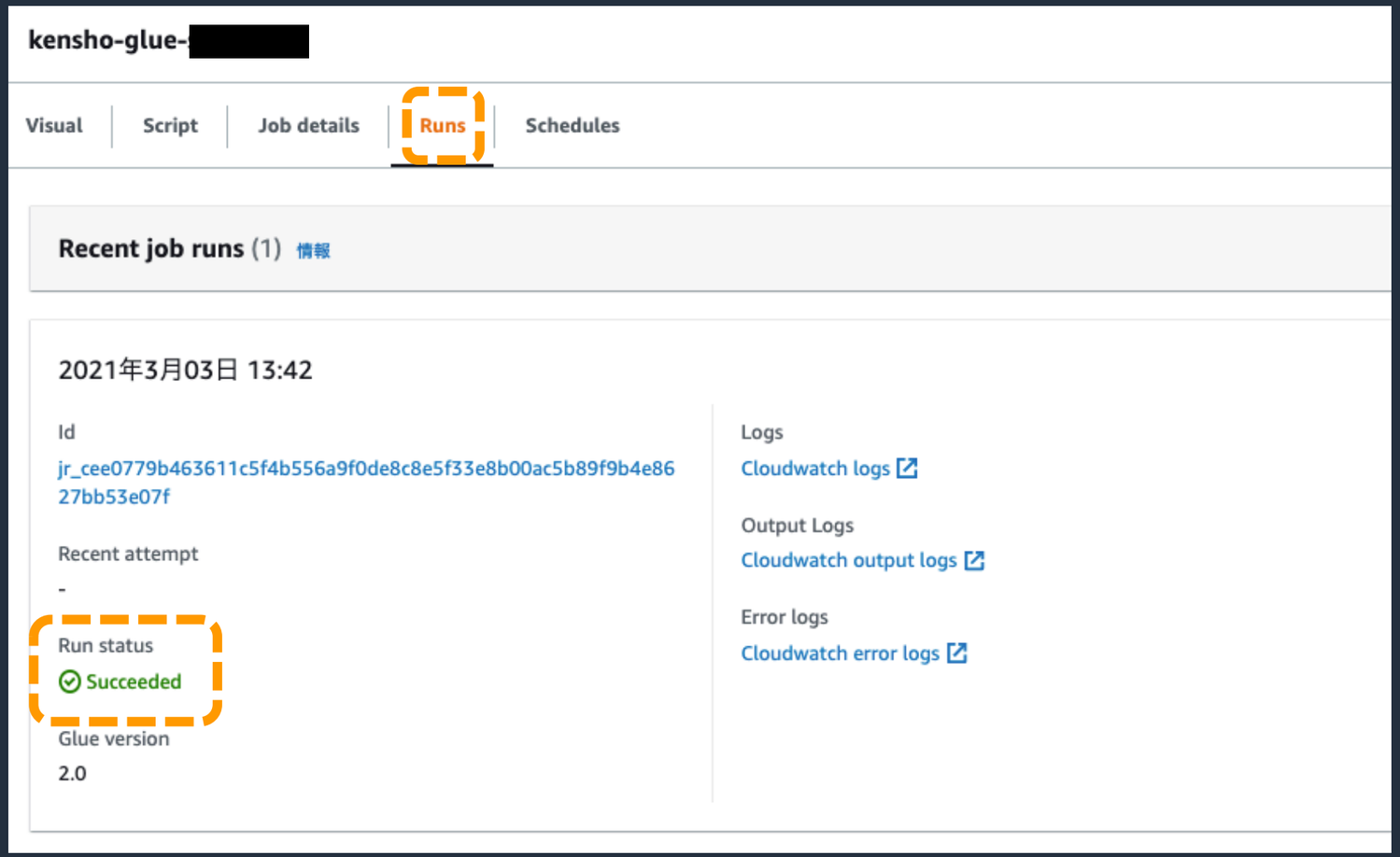
- Job details タブを選択し、
  ユースケース1と同様の内容を設定

- Save ボタンを押下し、
  RunボタンでJobを実行



結果のモニタリング

- Runs タブを選択
- Run statusが、
  Succeededになったことを確認

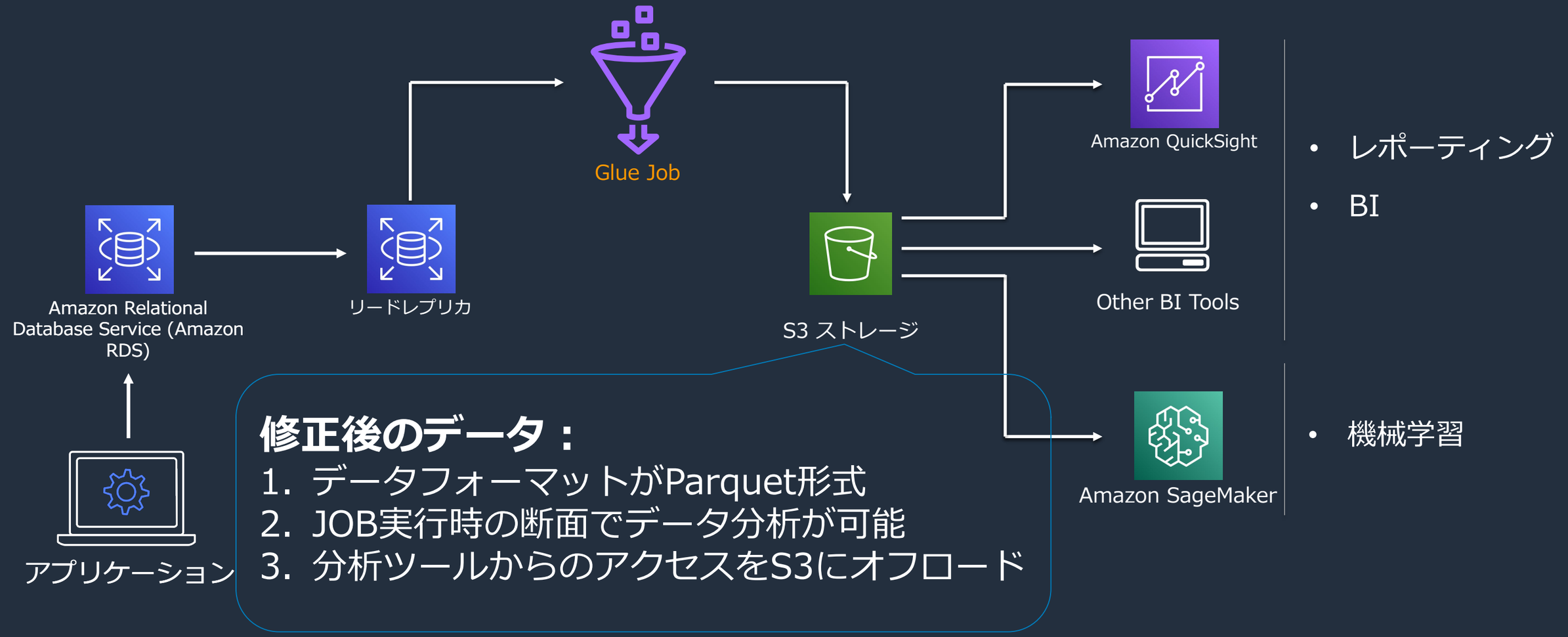


実行結果

- 出力バケットにファイルが作成されていることを確認

名前	タイプ
run-DataSink0-1-part-block-0-0-r-00000-snappy.parquet	parquet
run-DataSink0-1-part-block-0-0-r-00001-snappy.parquet	parquet
run-DataSink0-1-part-block-0-0-r-00002-snappy.parquet	parquet
run-DataSink0-1-part-block-0-0-r-00003-snappy.parquet	parquet
run-DataSink0-1-part-block-0-0-r-00004-snappy.parquet	parquet
run-DataSink0-2-part-block-0-0-r-00000-snappy.parquet	parquet
run-DataSink0-2-part-block-0-0-r-00001-snappy.parquet	parquet
run-DataSink0-2-part-block-0-0-r-00002-snappy.parquet	parquet
run-DataSink0-2-part-block-0-0-r-00003-snappy.parquet	parquet

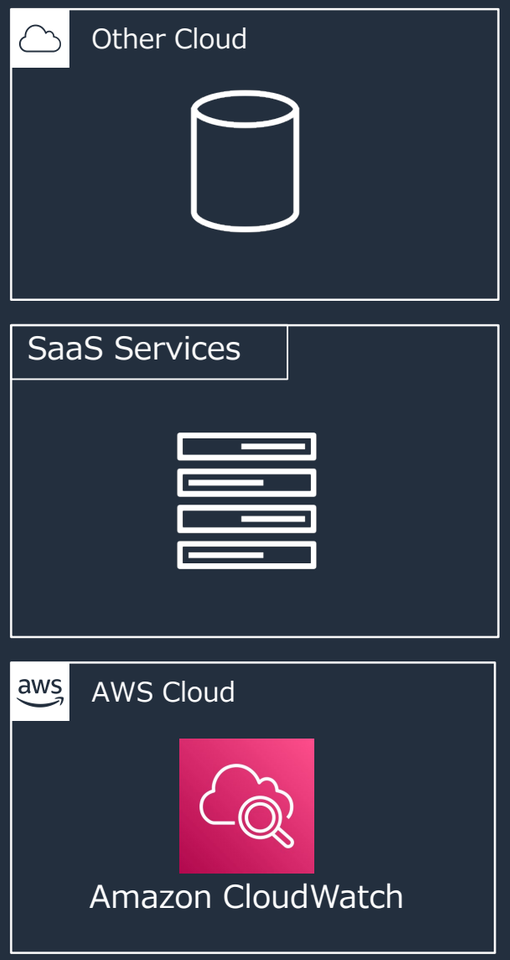
RDSにあるデータを直接分析している場合の解決策

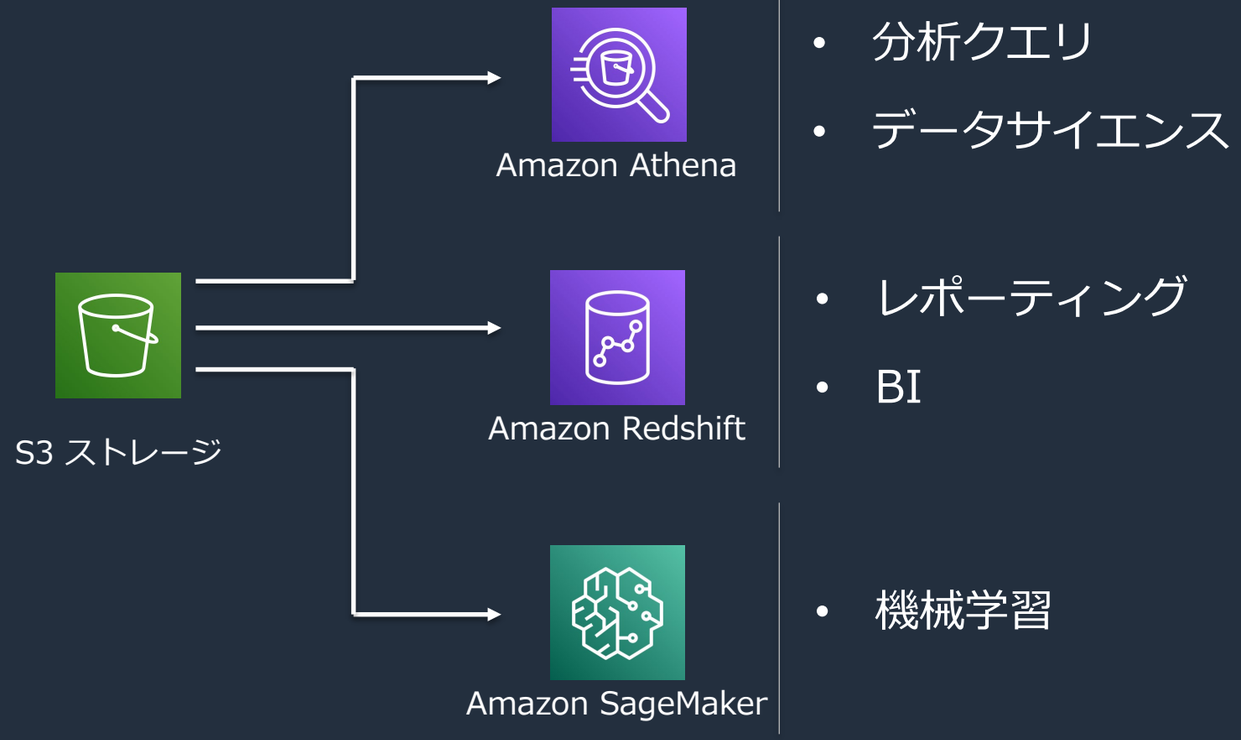


ユースケース 3：
他クラウド/オンプレミス にあるデータの分析

他クラウド/オンプレミス にあるデータを分析する場合の問題点

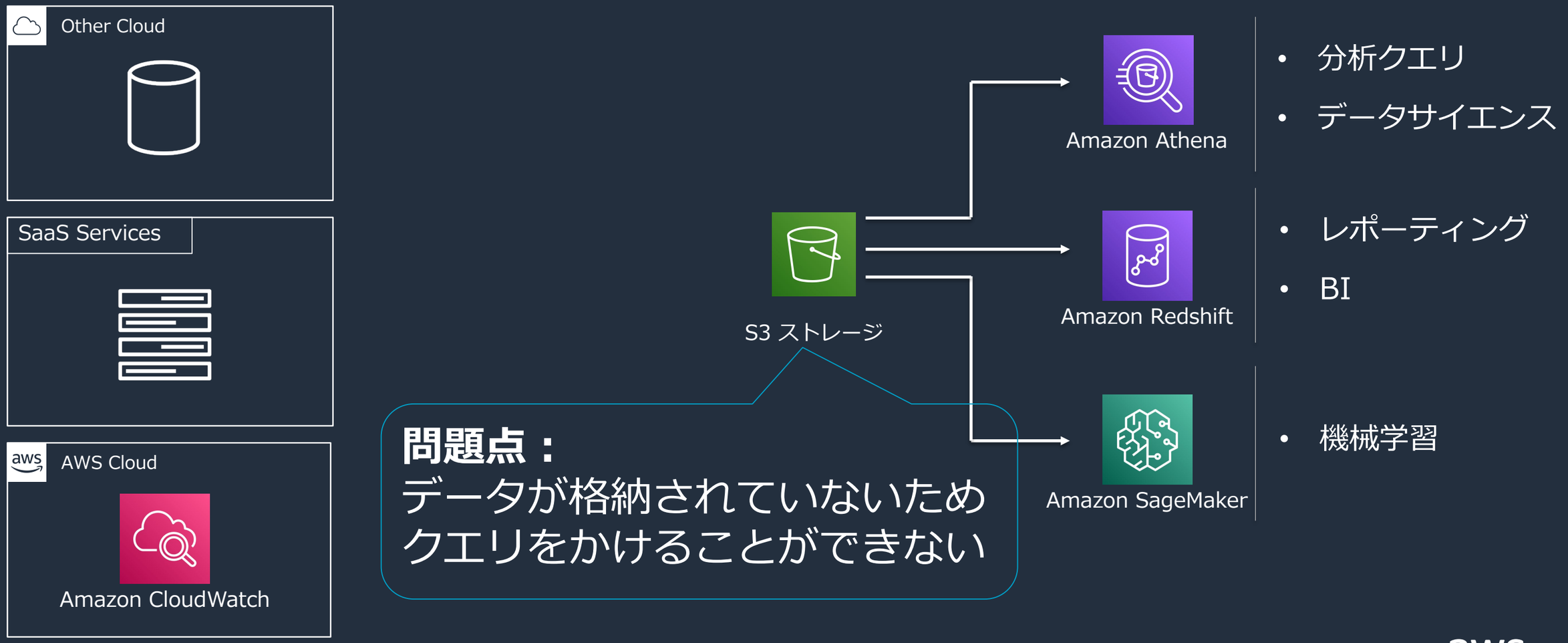
ネイティブでサポートされていないデータソースからのデータ移行





他クラウド/オンプレミス にあるデータを分析する場合の問題点

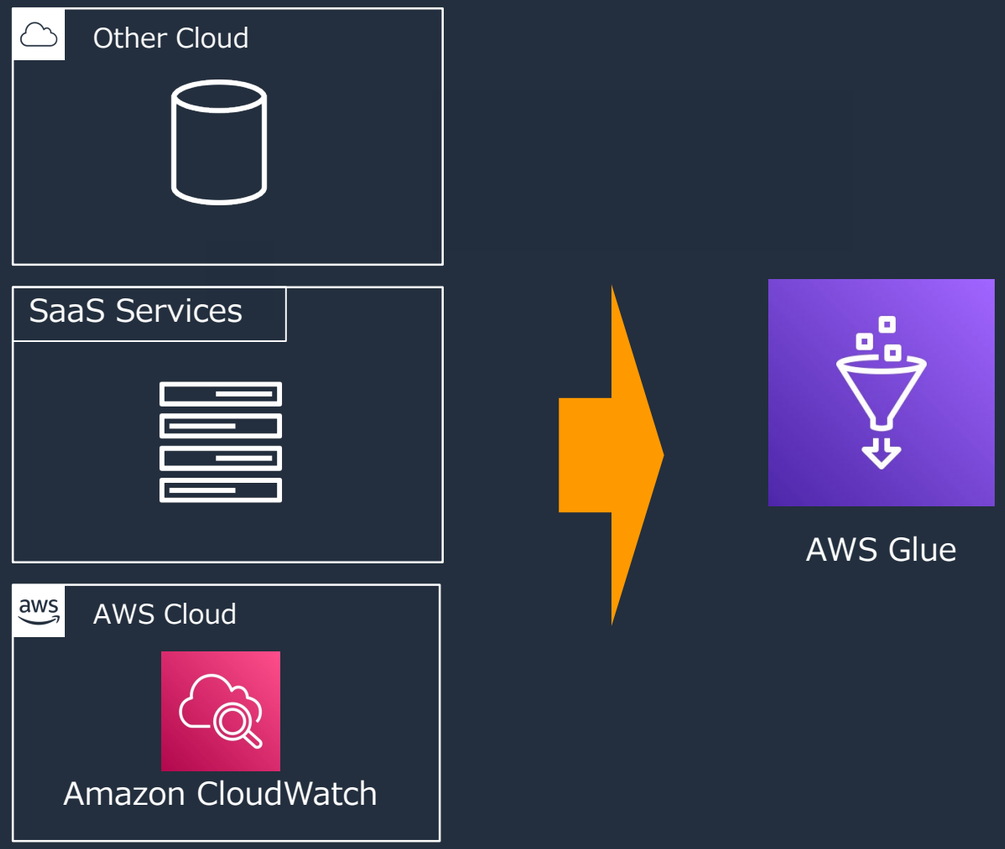
ネイティブでサポートされていないデータソースからのデータ移行



AWS Glue Custom Connectors

NEW

ネイティブにサポートされていないデータストアとの接続



任意のデータソースからデータを転送可能

SaaS アプリケーション等から Amazon S3にデータを転送

多くのコネクタを利用可能

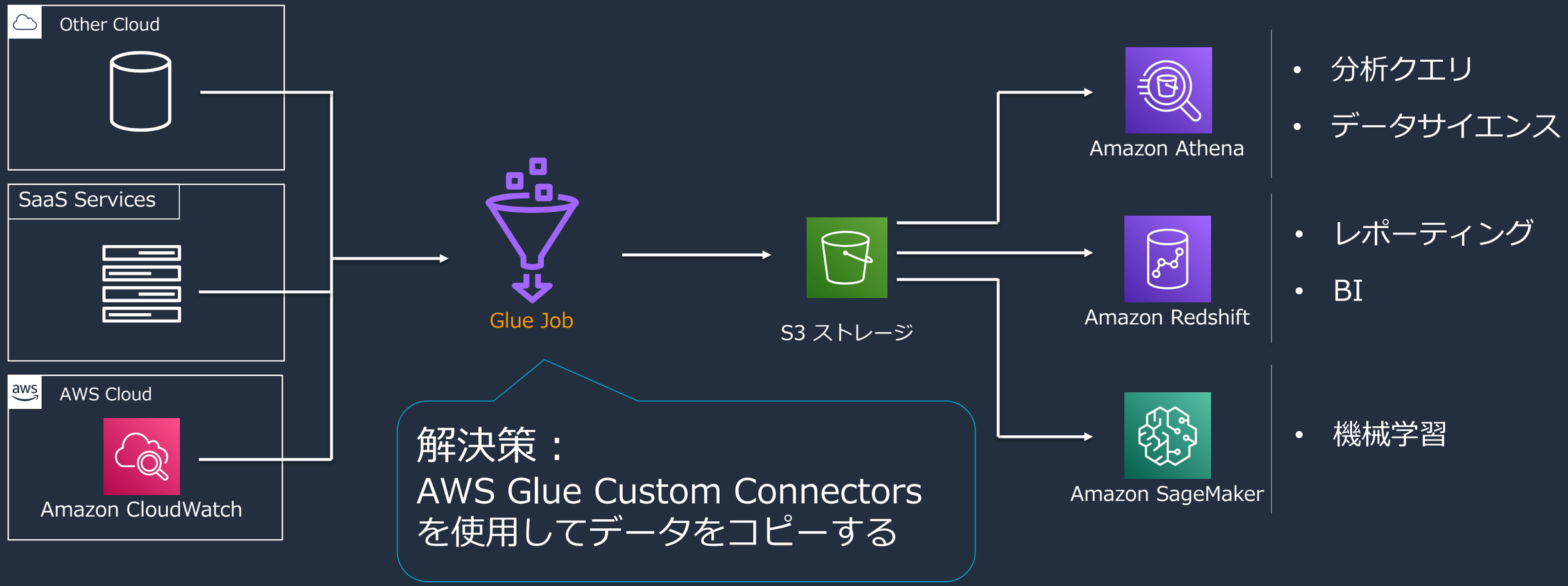
Salesforce、SAP、Snowflake、Google BigQuery 等とも連携

AWS Marketplaceとの統合

開発したカスタムコネクタを AWS Marketplace で共有可能

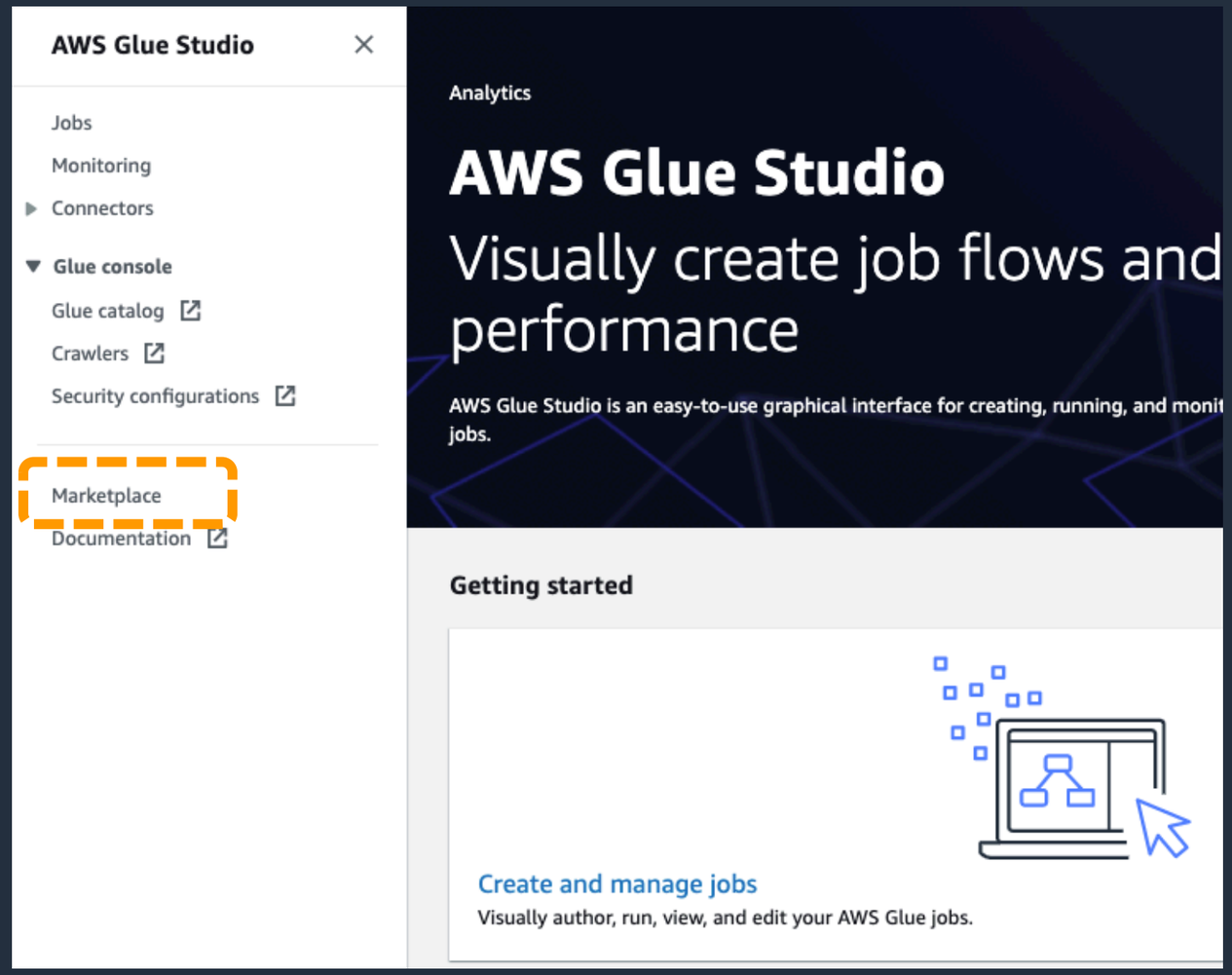
https://aws.amazon.com/jp/about-aws/whats-new/2020/12/aws-glue-launches-aws-glue-custom-connectors/

他クラウド/オンプレミス にあるデータを分析する場合の解決策



AWS Glue Custom Connectors の使用

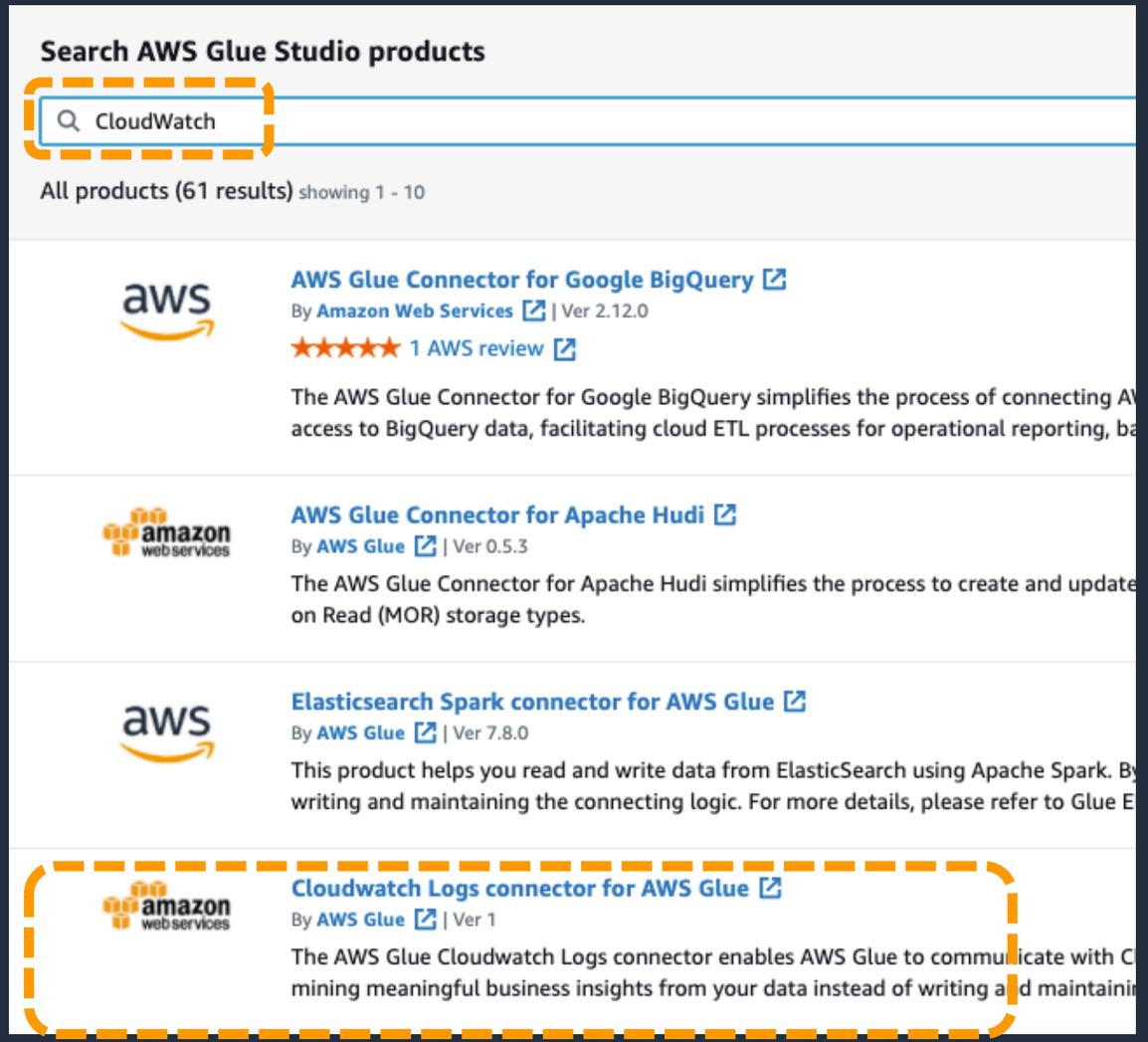
- AWS Glue Studioのホーム画面の左にあるMarketplaceを選択



カスタムコネクターのサブスクライブ①

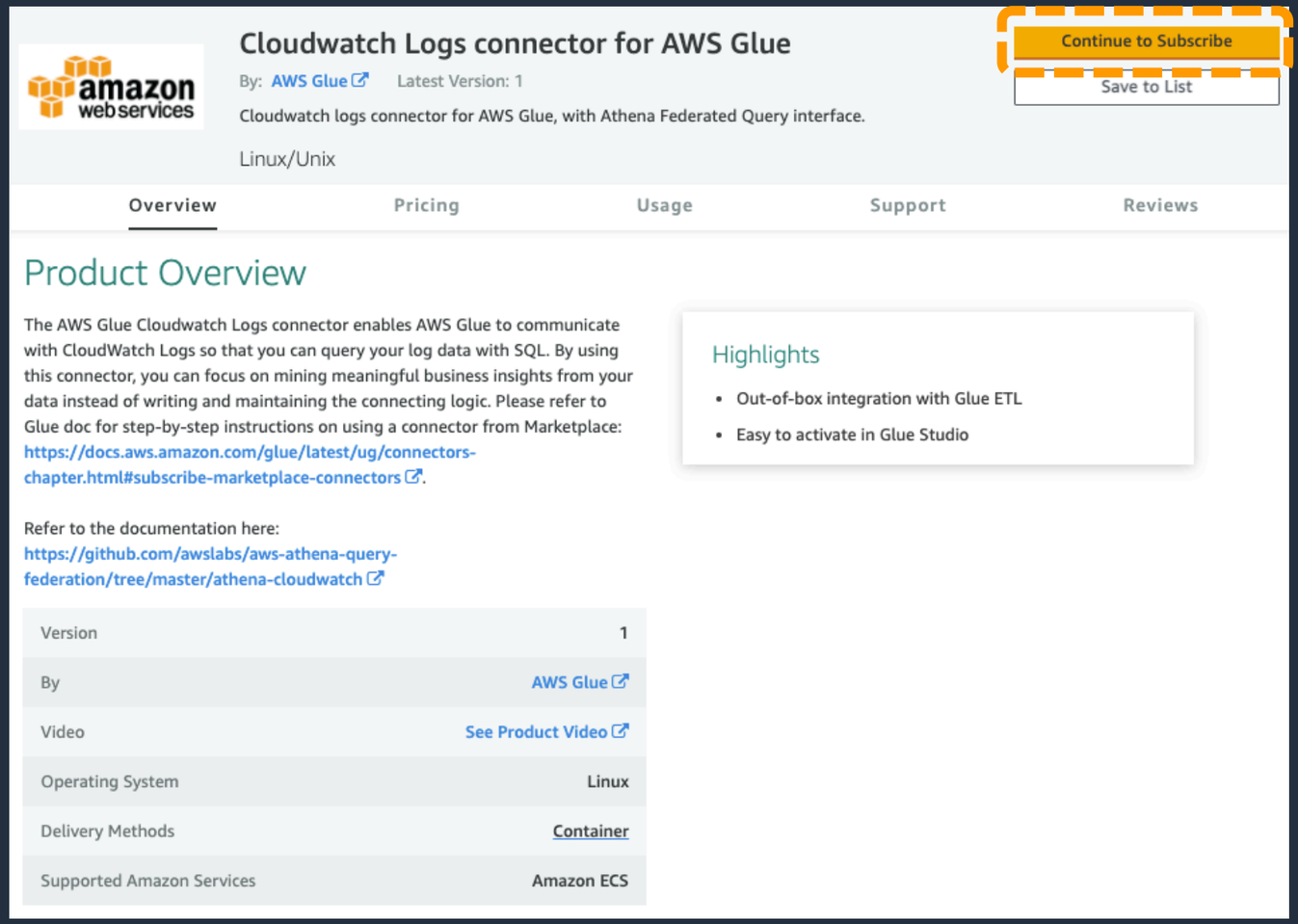
- Topにある検索バーにCloudWatchと入力
- 検索結果に表示される以下を選択

  Cloudwatch Logs connector for AWS Glue



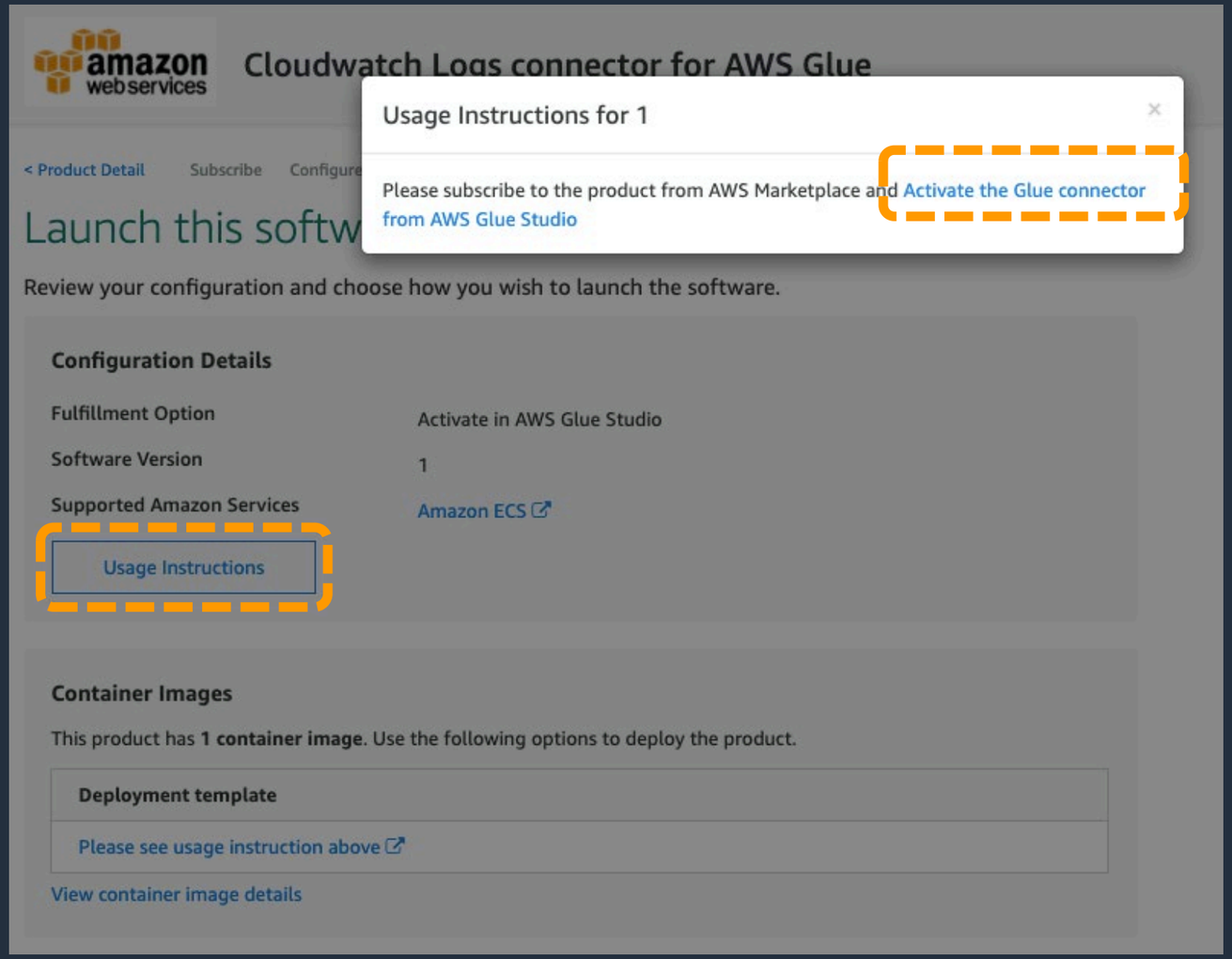
カスタムコネクターのサブスクライブ②

- カスタムコネクターの内容を確認し、
  画面右上のContinue to Subscribe を選択



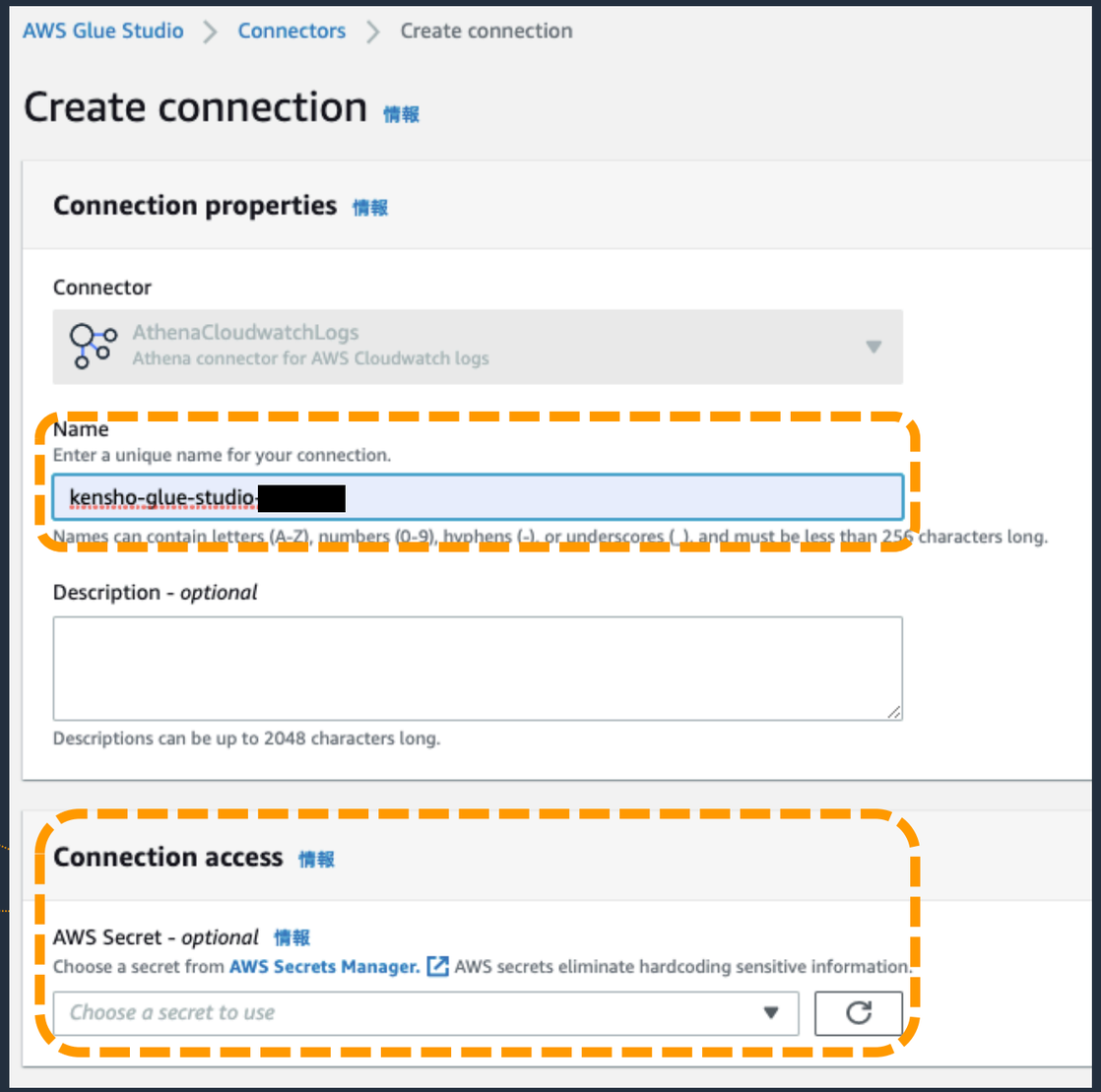
カスタムコネクターのサブスクライブ③

- ソフトウェアのバージョン等の選択後に、出力されるUsage Instructionsを押下
- 出力されるポップアップにて、以下を選択
  Activate the Glue connector in AWS Glue Studio
- Custom connector の設定画面へ自動遷移



コネクターの作成

- Name欄に任意の名前を入力
- Connection access は今回は空欄

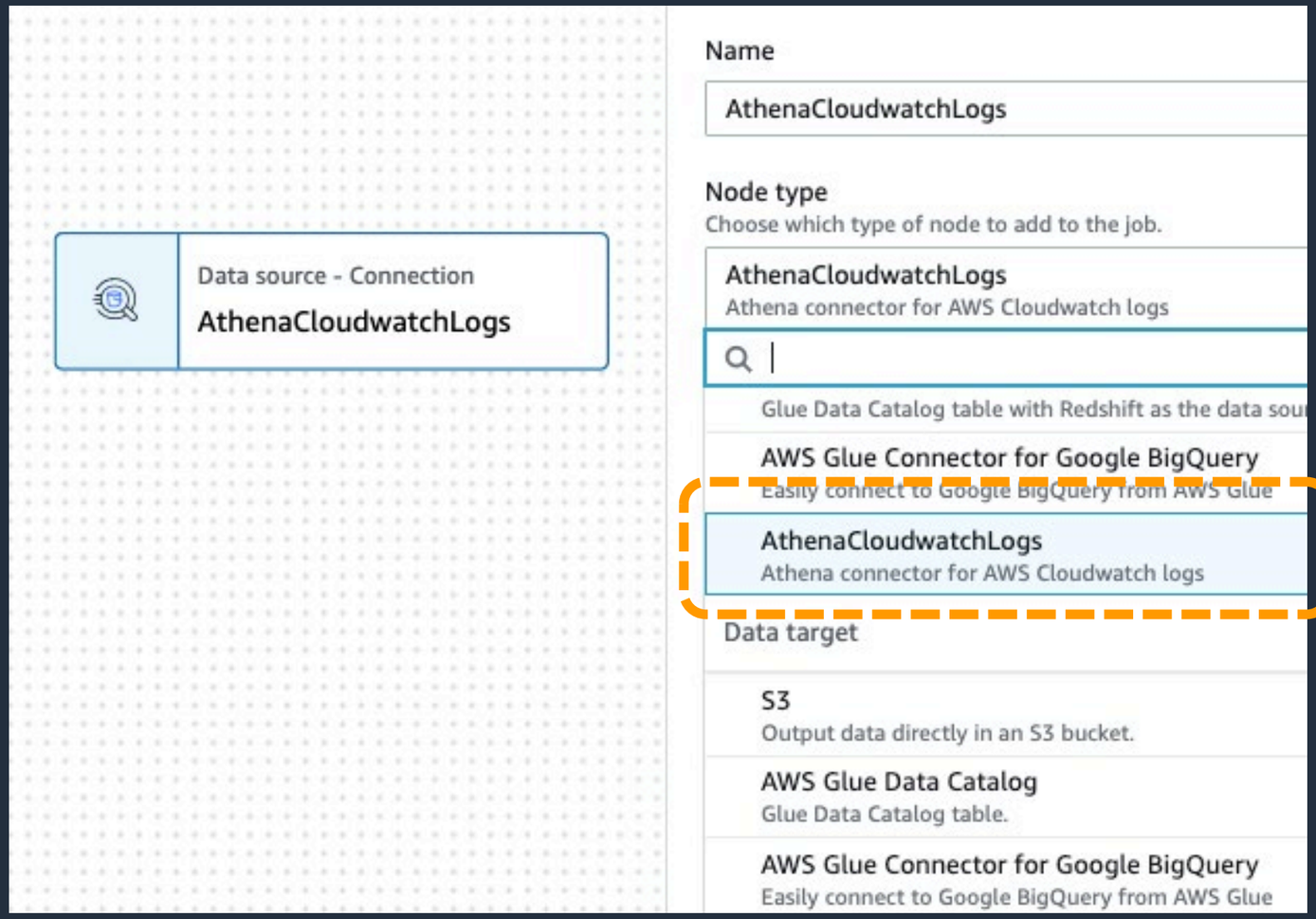


認証情報について

他クラウドやSaaSへの接続を行う際には認証情報はあらかじめ、AWS Secrets Managerに格納する必要がある。
今回は、CloudWatchへの接続のため入力不要。

データソースの指定①

- Node type にAthenaCloudwatchLogsが、追加されていることを確認

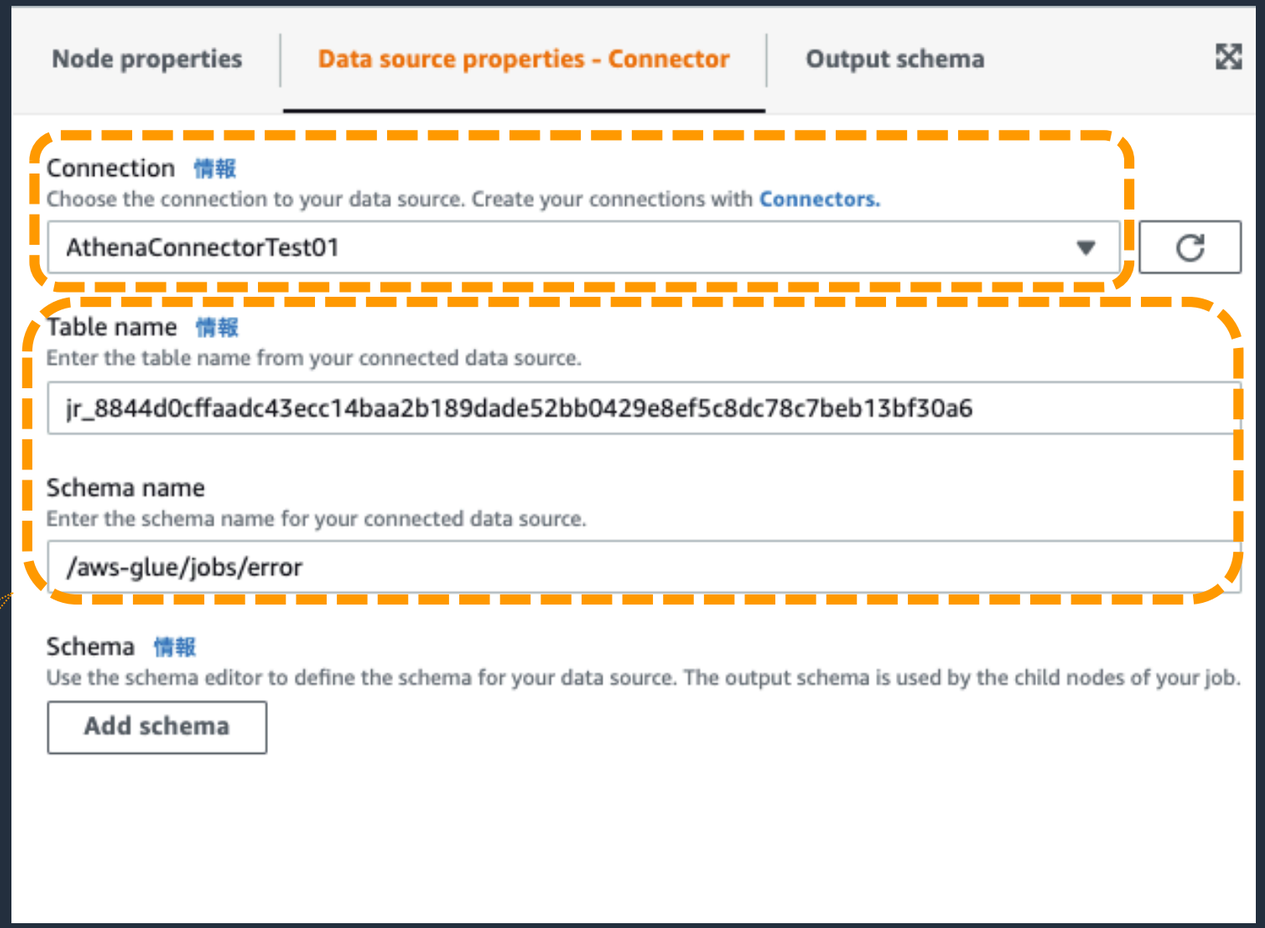


データソースの指定②

- Connectionに先程作成したコネクターを選択
- CloudWatchを確認し取得対象となるLogStreamsとLogGroupsを確認
- Table nameにLogStreamsを入力し、Schema nameにLogGroupsを入力

LogGroup内のすべてのLogStreamsで構成される特別な「all_log_streams (すべてのログストリーム)」などでもログを指定可能。

※詳細は以下リンク先ご参照。

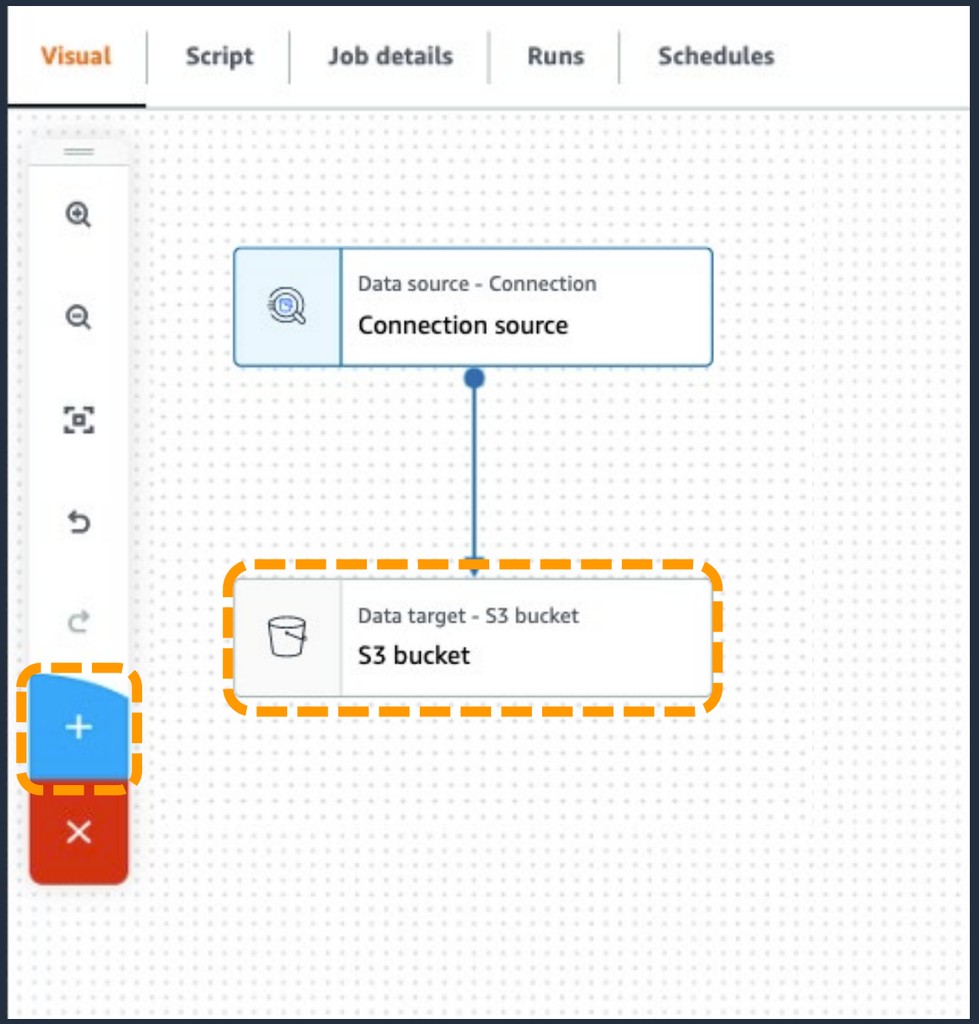


https://github.com/awslabs/aws-athena-query-federation/blob/master/athena-cloudwatch/README.md

ターゲットの指定

- Connection source ノードを選択した状態で、画面左下の“＋”を押下

- Node typeにS3を選択し、ユースケース1と同様のターゲット設定を実施



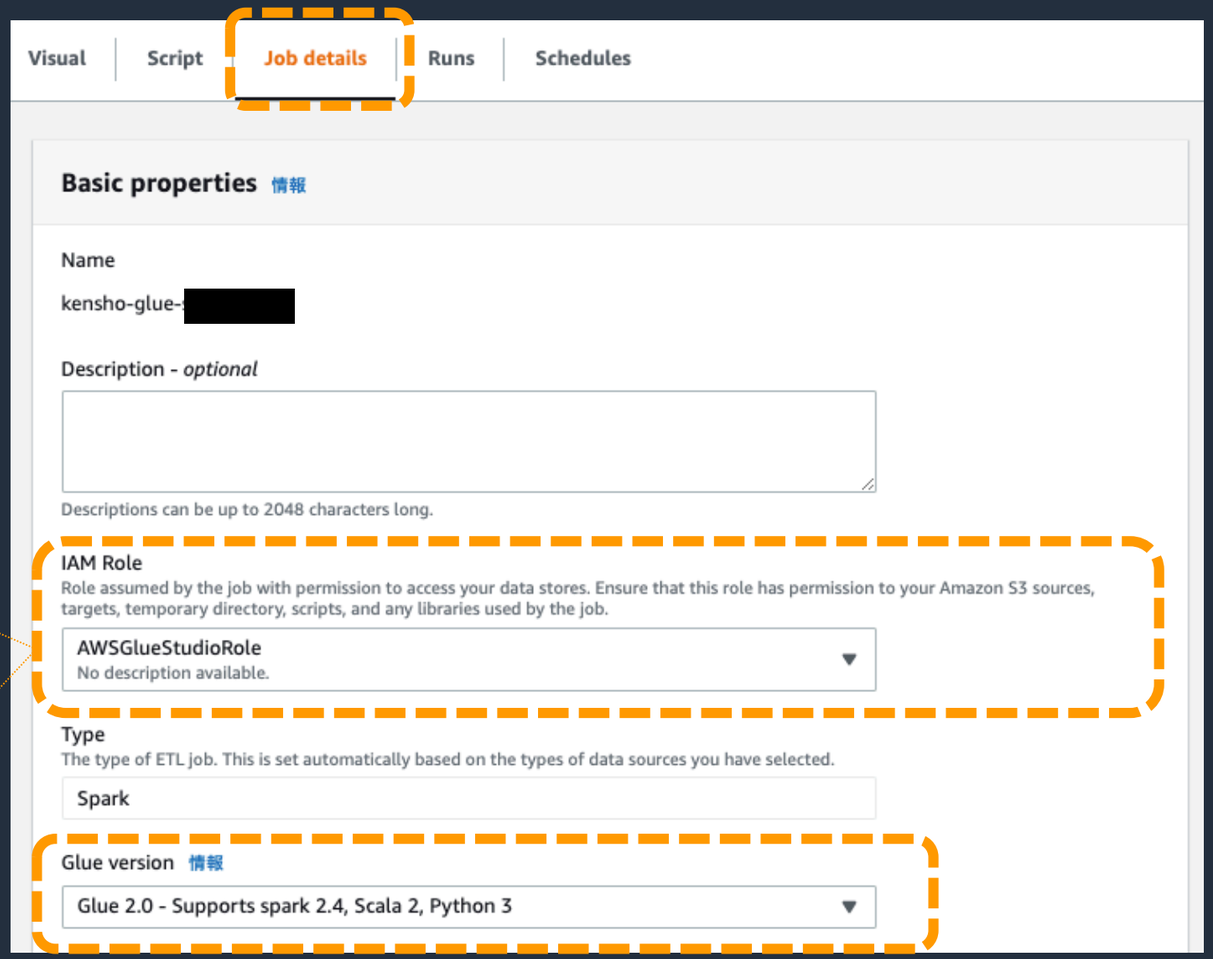
JOBの詳細設定

- Job detailsタブを選択
- IAM Role に必要な権限を持ったロールを選択
- その他はユースケース1と同じよう設定

必要な権限について
ユースケース3にて追加で必要となる権限は以下。

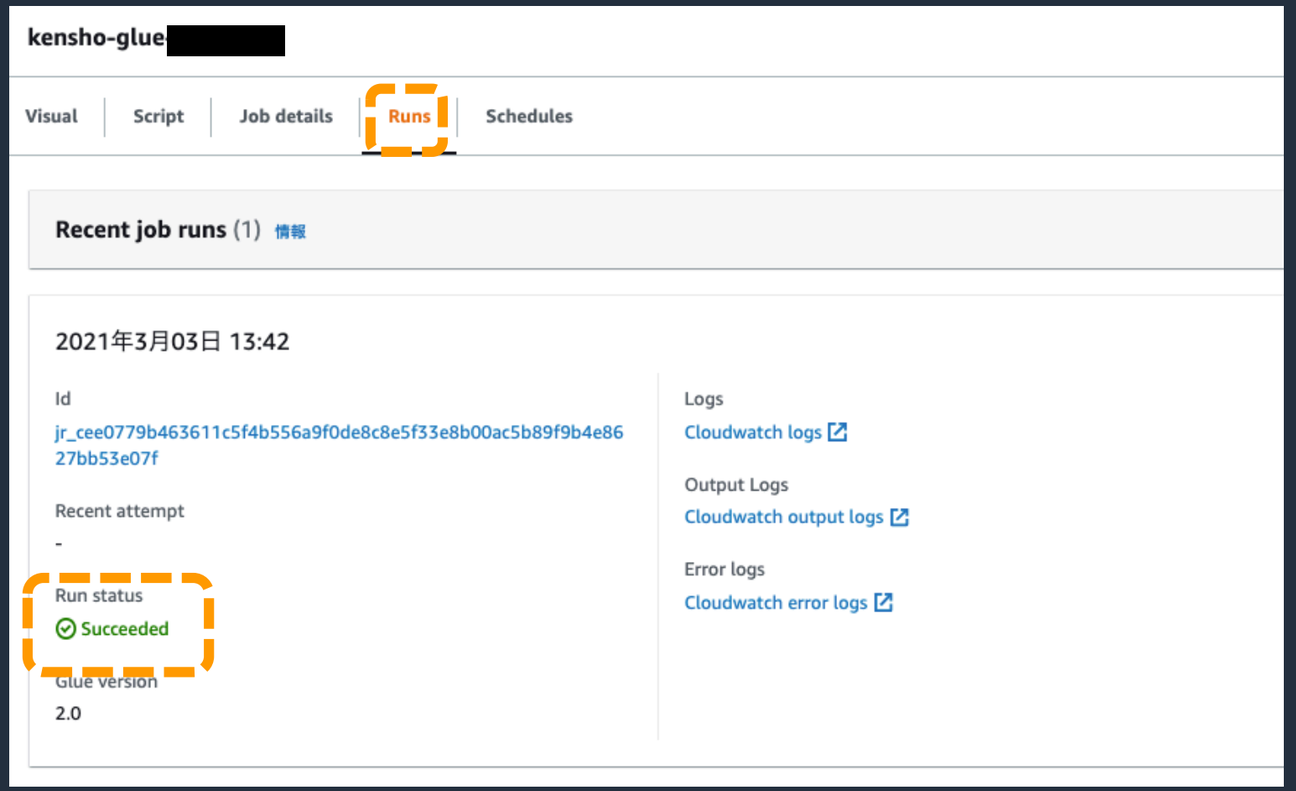
- AWSの管理ポリシーである AmazonEC2ContainerRegistryReadOnly権限
- glue:GetJobおよびglue:GetJobsアクセス許可
- 取得対象となるCloud Watch Logsへのアクセス権限

これらの権限をユースケース1のロールに追加で付与した上でロールを選択する。

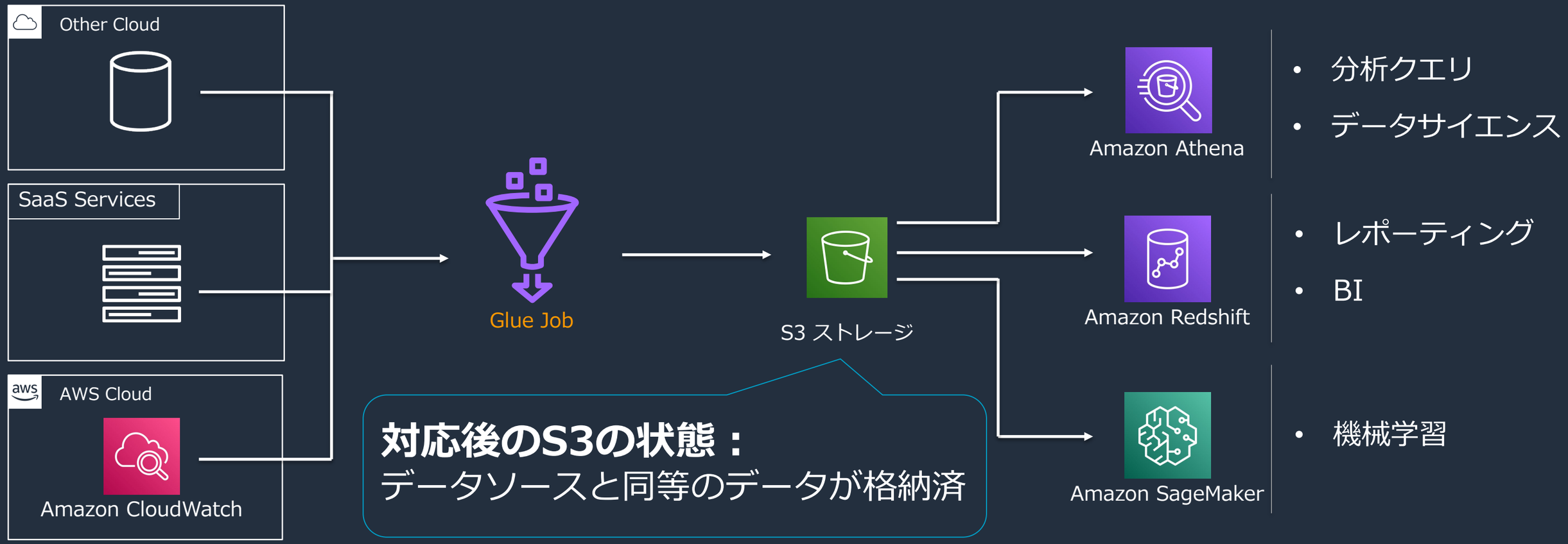


結果のモニタリング

- 画面右上のSaveを押下した後、Runボタンを押下しJobを実行
- Runs タブを選択
- Run statusが、Succeededになったことを確認



他クラウド/オンプレミス にあるデータを分析する場合の解決策



その他の主要アップデート

AWS Glue の構成要素と周辺サービス（再掲）

Extract, Transform, and Load（ETL）ジョブ

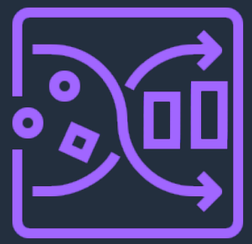
AWS Glue Data Catalog

Crawler

Workflow Management

New

AWS Glue Studio



AWS Glue DataBrew

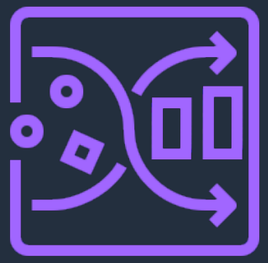
AWS Glue Elastic Views

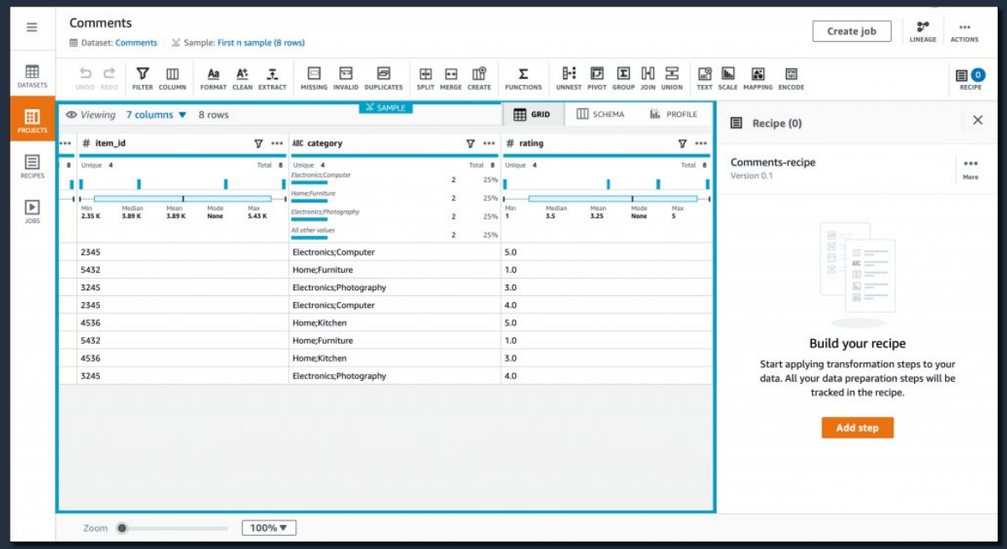
aws

AWS Glue DataBrew

NEW

分析および機械学習のためのビジュアライズツール





データ準備タスクの自動化

データ準備にかかる時間を最大80%短縮

ノンコードでのデータ変換

250を超える既成の組込関数を選択して処理を実施

サーバレス

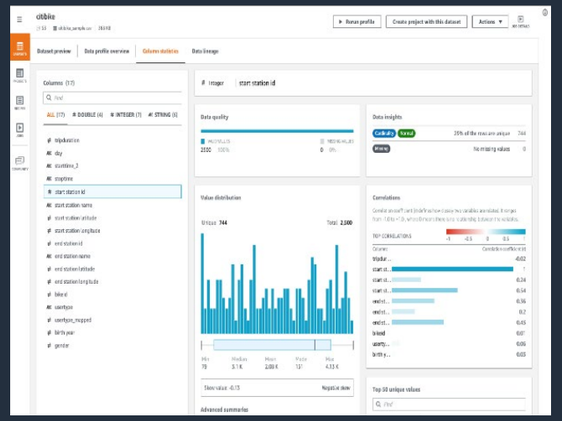
インフラを管理することなく、テラバイト規模のデータを変換

aws

AWS Glue DataBrew の特徴

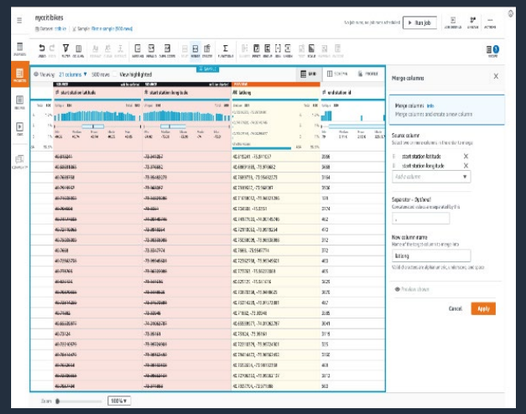
NEW

TB～PBクラスのデータに対して、クレンジングと正規化を直接実行



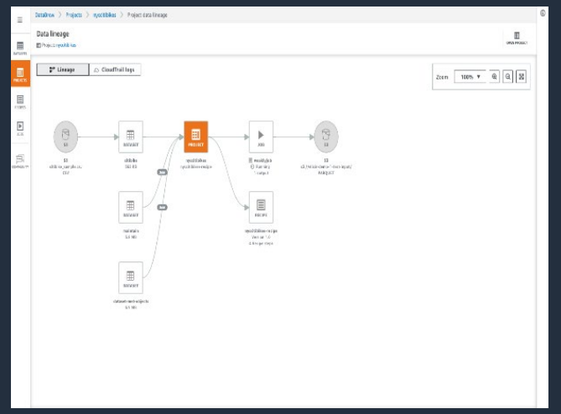
データ品質の理解

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



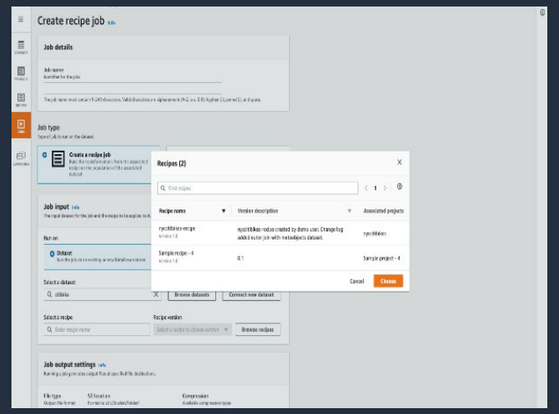
データのクリーンアップと正規化

250種類以上の変換処理から選択して、データの視覚化、クリーンアップ、正規化を実施



データリネージの視覚化

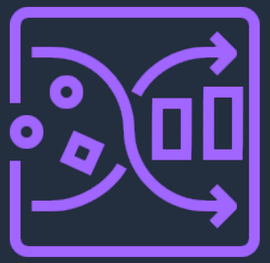
さまざまなデータソースと変換手順を視覚化してトラッキング



自動化

保存された変換手順を新しい入力データに適用

AWS Glue DataBrew と AWS Glue Studioの違い



AWS Glue DataBrew

- GUI でデータを可視化しながらデータ変換したい
- Data Brew で利用可能な 250種類以上の変換処理を活用したい
- データアナリスト & データサイエンティスト向け

AWS Glue Studio

- プログラムコードでデータ変換したい
- GUI でデータ変換を記述しつつ、必要に応じてカスタムコードを加えたい
- カスタムコネクタでさまざまなデータソースと接続したい
- ストリーミング処理を活用したい
- ETLデベロッパー**向け**

aws

AWS Glue の構成要素と周辺サービス（再掲）

Extract, Transform, and Load（ETL）ジョブ

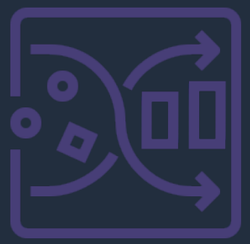
AWS Glue Data Catalog

Crawler

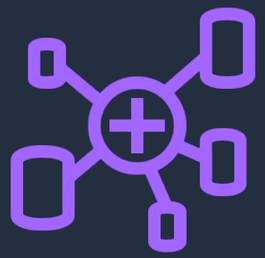
Workflow Management

New

AWS Glue Studio



AWS Glue DataBrew



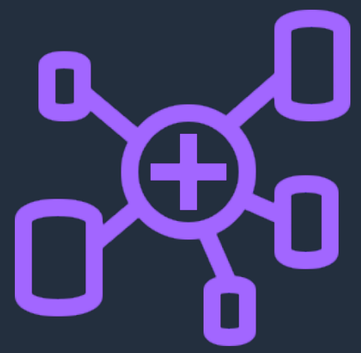
AWS Glue Elastic Views

aws

AWS Glue Elastic Views (Preview)

NEW

複数のデータソースにまたがるマテリアライズドビューを作成



Amazon DynamoDB

Amazon Redshift

Amazon Simple Storage Service (S3)

Amazon Elasticsearch Service

Target Database

SQLを用いてターゲットDBを指定

ターゲットデータベースに対して自動的にデータを複製

データ変更に迅速に対応

データソースをモニタし、データの変更が発生したら迅速で反映

サーバーレス

インフラ管理は不要。キャパシティを自動的に制御

aws

プレビュー時にサポートされるsourceとtarget

NEW

	Source	Target
Amazon DynamoDB	✓	
Amazon Elasticsearch Service		✓
Amazon S3		✓
Amazon Redshift		✓

aws

previewで利用可能なユースケース

NEW

- DynamoDB内のデータの検索インデックスをElasticsearch内に作成

- DynamoDB内の業務データをDataLakeに統合

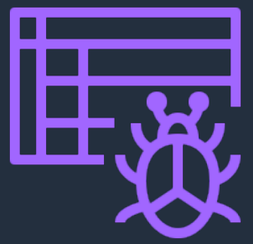
- DynamoDB内のデータをRedshiftで分析

aws

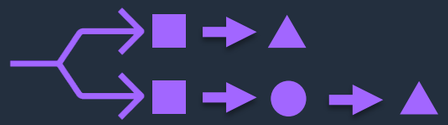
AWS Glue の構成要素と周辺サービス（再掲）

Extract, Transform, and Load（ETL）ジョブ

AWS Glue Data Catalog



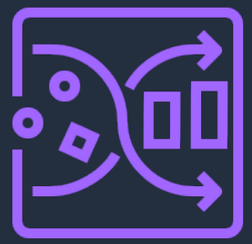
Crawler



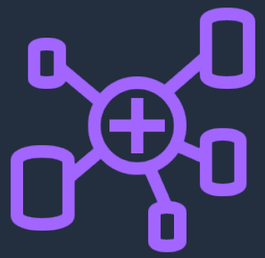
Workflow Management

New

AWS Glue Studio



AWS Glue DataBrew



AWS Elastic Views

aws

まとめ

まとめ

1. AWS Glueとは、
   サービス間でデータを簡単に**移動**するための、**サーバーレスデータ統合サービス**

2. データの前処理が必要な場合、
   AWS Glue Studioを使用することでグラフィカルにETLJobを作成可能

3. データがAWS以外に格納されている場合でも、
   AWS Glue Custom Connectorsを使用することでデータの移動が可能

aws

参考情報

AWS Glue ホームページ

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

AWS Glue 開発者ガイド（公式ドキュメント）

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

AWS Glue custom connectors の設定例

https://aws.amazon.com/jp/blogs/big-data/migrating-data-from-google-bigquery-to-amazon-s3-using-aws-glue-custom-connectors/

AWS Glueの料金

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

AWS Glueのサービス制限

https://docs.aws.amazon.com/ja_jp/general/latest/gr/aws_service_limits.html#limits_glue

aws

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

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]

4月以降のBlack Belt Online Seminarについて

ライブ配信によるBlack Belt Online Seminarは3月一杯で終了し、
今後はオンデマンドによる定期配信に変更いたします。

今後もコンテンツを拡充して行きますので、楽しみにお待ちください。

オンデマンドでの配信スケジュールは、AWS Blog, AWSニュースレターでお知らせいたします（5月17日週に再開を予定しています）

aws

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