AWS Glue

AWS Black Belt Online Seminar

鈴木　浩之

Senior Solutions Architect

2023/03

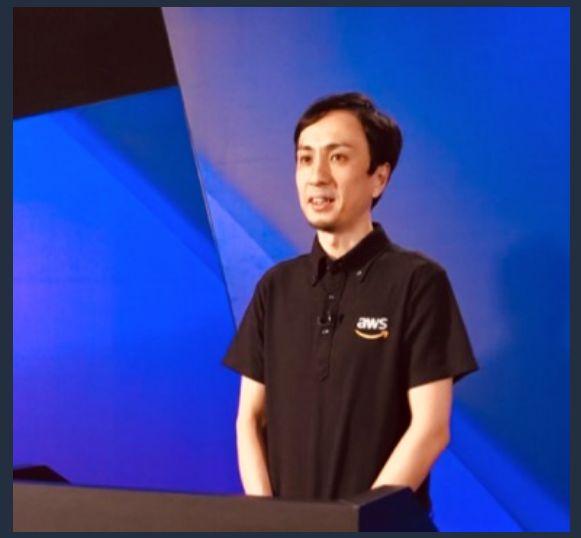
AWS Black Belt Online Seminarとは

- 「サービス別」「ソリューション別」「業種別」などのテーマに分け、アマゾン ウェブ サービス ジャパン合同会社が提供するオンラインセミナーシリーズです
- AWSの技術担当者が、AWSの各サービスやソリューションについてテーマごとに動画を公開します
- 動画を一時停止・スキップすることで、興味がある分野・項目だけの聴講も可能、スキマ時間の学習にもお役立ていただけます
- 以下のURLより、過去のセミナー含めた資料などをダウンロードすることができます
  - https://aws.amazon.com/jp/aws-jp-introduction/aws-jp-webinar-service-cut/
  - https://www.youtube.com/playlist?list=PLzWGOASvSx6FIwIC2X1nObr1KcMCBBlqY

内容についての注意点

- 本資料では2023年03月時点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(https://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

自己紹介



名前： 鈴木　浩之

所属： アマゾン ウェブ サービス ジャパン合同会社
技術統括本部エンタープライズ技術本部
シニアソリューションアーキテクト

経歴：通信事業者で基幹システム開発、新規サービス開発、
ビッグデータ基盤の開発運用などに従事

好きなAWSサービス： Amazon Redshift、AWS Glue

本セミナーの対象者

データ分析基盤を検討されている方

データエンジニア・ETL デベロッパー

AWS Glue 導入を検討中の方

アジェンダ

1. 背景
2. AWS Glue について
3. demo
4. 費用
5. 導入事例の参考情報
6. まとめ

背景

お客様はデータからより多くの価値を求めています

指数関数的に
増えるデータ量

多様な
データソース

多様化する
データの種類

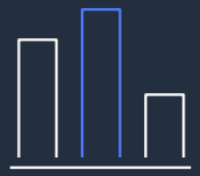
増加する
利用者

様々な
アプリケーション
による分析

より要求の厳しいワークロード

バッチ

DHW



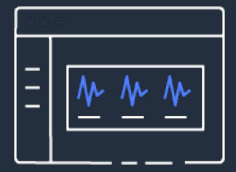
BI

レポーティング

夜間または時間単位の ETL

即時性を必要としないデータの
長時間ジョブ

リアルタイム



レスポンシブ
ダッシュボード

モニターと
アラート

ストリーミング
アプリ

継続的なオペレーション

即時性を必要とするデータの
ストリーム/マイクロバッチジョブ

従来型の ETL ソリューションでは対応が困難

複数ソリューション

処理を実現するには複数のソリューションが必要

拡張性のないインフラストラクチャ

オンプレミスの ETL ツールはインストール、管理、拡張が複雑

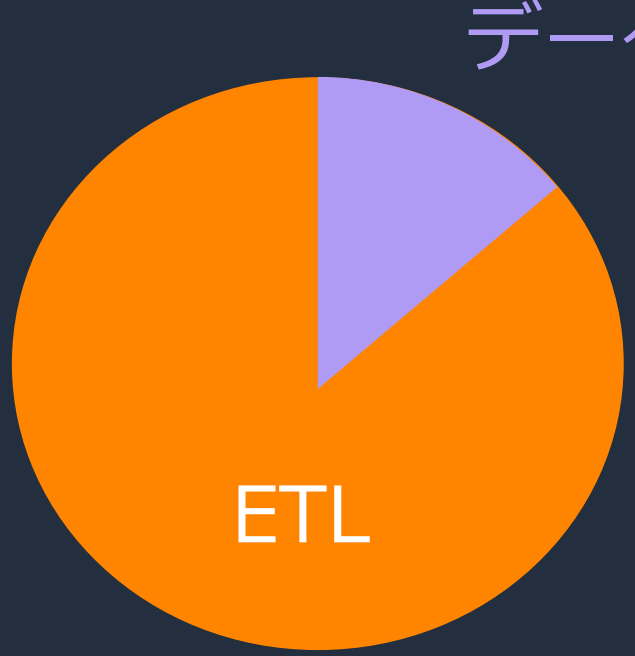
高コスト

一元化されたデータカタログ、ストリーミングデータの処理などの高度な機能は個別にライセンスが必要

スキルセットの多様化

データを操作する利用者の技術スキルは様々

データ活用のよくある課題



ETL：Extract Transfer Load
データの収集・変換・ロード

開発者のリソースの大部分が
ETL に関わるデータパイプラインの管理と
メンテナンスに費やされている

これをシンプルにすることでデータ活用の進展や
コスト削減が期待できる！

AWS Glue のご紹介

AWS Glue で実現できること

スケーラブルなデータ分析のための最新のデータ戦略 (データレイク、データファブリックなど) を実行

高価で従来の ETL ソリューションから移行することで、柔軟性を高め、コストを削減できます

データ資産をカタログ化して、最新のデータアーキテクチャ全体でデータを利用できるようにする

Python と Apache Spark を使用して、数ペタバイトのデータをバッチ、ストリーミング、およびリアルタイムで処理

分析と機械学習用のデータ準備

データライフサイクルと要素






収集	保存	変換とカタログ化	分析と可視化	予測	共有
オンプレミスを含むあらゆるソースからリアルタイムに	エクサバイト級のデータ量まで	データの準備、転送、およびシームレスなデータアクセスの提供	あらゆるユースケースに対応するエンドツーエンドの分析と可視化	最も包括的な ML と AI サービスのセット	わずか数行のコードで

統合セキュリティ、ガバナンス、データアクセス制御

aws

代表的なデータアーキテクチャー

収集

オンプレミスのソースやリアルタイムで生成されるデータなど、あらゆるソースからデータを取り込みます。

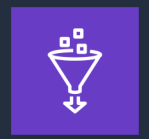

On-premises

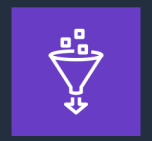

Streams


Databases


Logs


SaaS / 3rd Party


AWS Glue ETL


AWS Glue ETL Crawler / Connector


Amazon S3

保存

トランザクションデータをデータベースに、分析データをデータウェアハウスとデータレイクにあらゆる規模で保存します。

変換とカタログ化

データに簡単にアクセスできるようにし、データがどこにあるかに関係なくデータを同期できるようにします


AWS Glue Data Catalog

分析と可視化

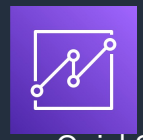
アドホッククエリ、分散フレームワーク、検索エンジンのいずれかを使用してデータを分析し、ダッシュボードでデータを可視化


Amazon Redshift


Amazon Athena


Amazon EMR


Amazon SageMaker


Amazon QuickSight

予測

ML スキルを必要とせずに ML ベースのインテリジェンスをアプリケーションに追加

共有

新しいインサイトを共有して、インリジェントでデータドリブンのアクションにつなげる

AWS Glue の機能

AWS Glue の特徴

サーバーレス データインテグレーションサービス

スケーラブルなデータ インテグレーションエンジン

組み込みデータ変換

多様なジョブ実行エンジン

ジョブのモニタリング

一元化された 統合データガバナンス

データカタログ

データ品質

クローラー

Lake Formation 連携

コネクタとデータ収集

コネクタ

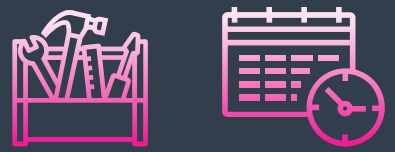
コネクタのマーケットプレイス

多様なインターフェース

生産性の高い データオペレーション

多様なペルソナへの対応

生産性向上



DataOps

データ収集、カタログ化、データ加工を一気通貫して行うことができる統合的なデータインテグレーションサービス

AWS Glue

データカタログ

データクローラー

他システム連携の豊富なコネクタ

サーバーレス ETL エンジン

豊富なビルトイン変換とローコード開発

スケジューラとワークフロー管理

データ収集、カタログ化、データ加工を一気通貫して行うことができる統合的なデータインテグレーションサービス

AWS Glue

データカタログ

サーバーレス ETL エンジン

データクローラー

豊富なビルトイン変換とローコード開発

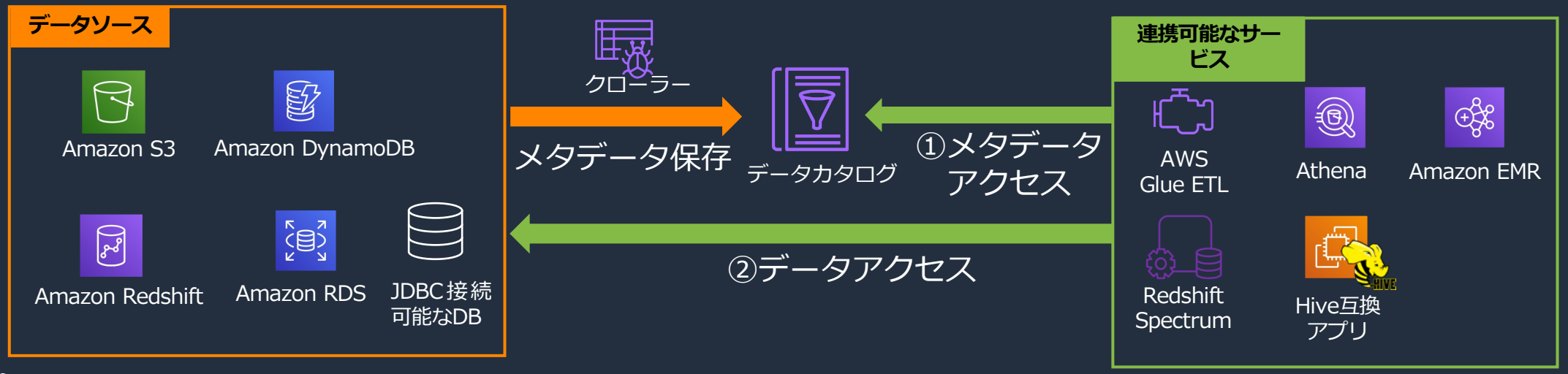
他システム連携の豊富なコネクタ

スケジューラとワークフロー管理

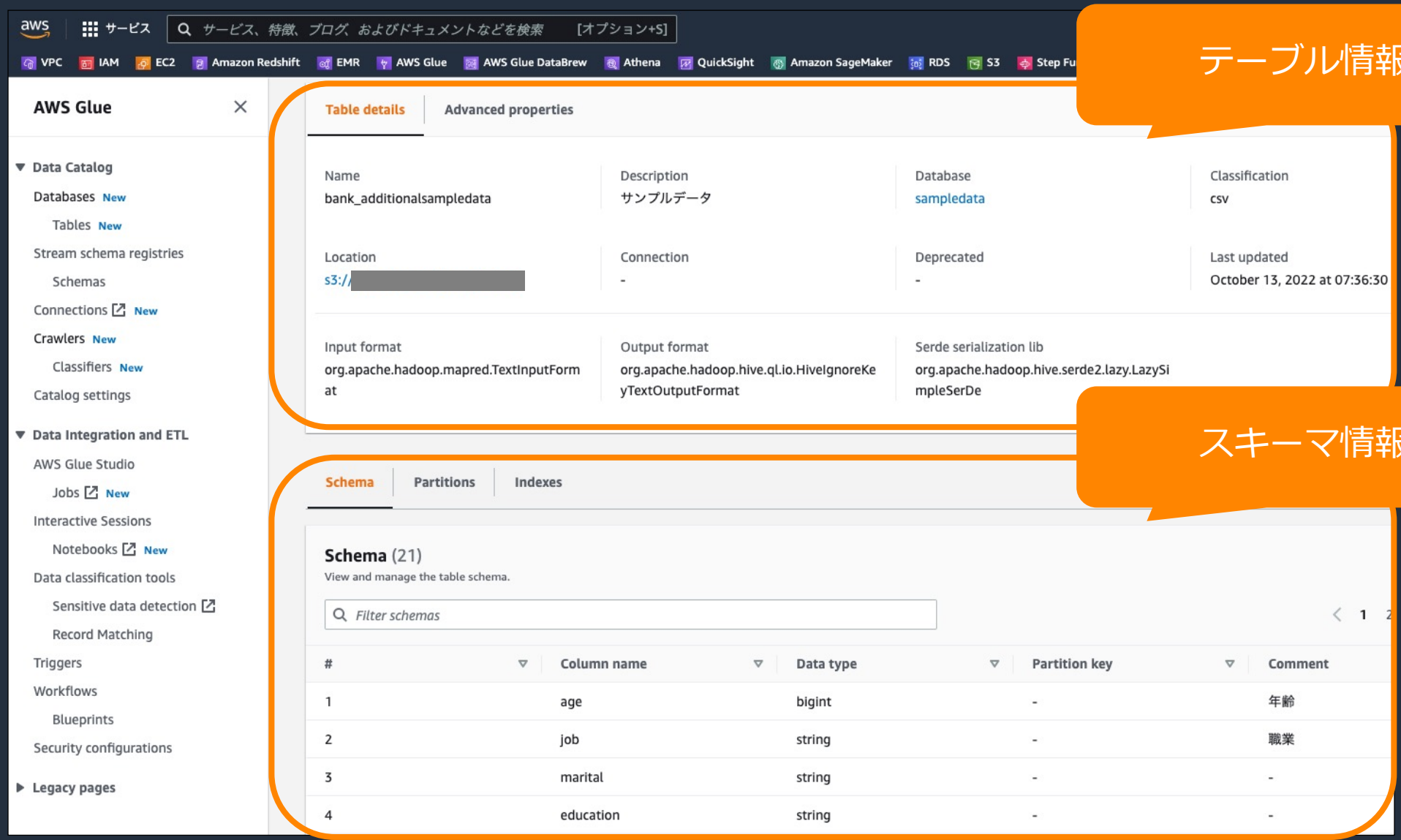
AWS Glue データカタログ

- 様々なデータソースのメタデータを一元管理
- AWS Glue ジョブや Amazon Athena、Amazon Redshift Spectrum などからデータへのアクセスが便利に
- Apache Hive メタストア互換
- クローラーでメタデータを収集し、カタログを自動生成

データカタログの利用イメージ

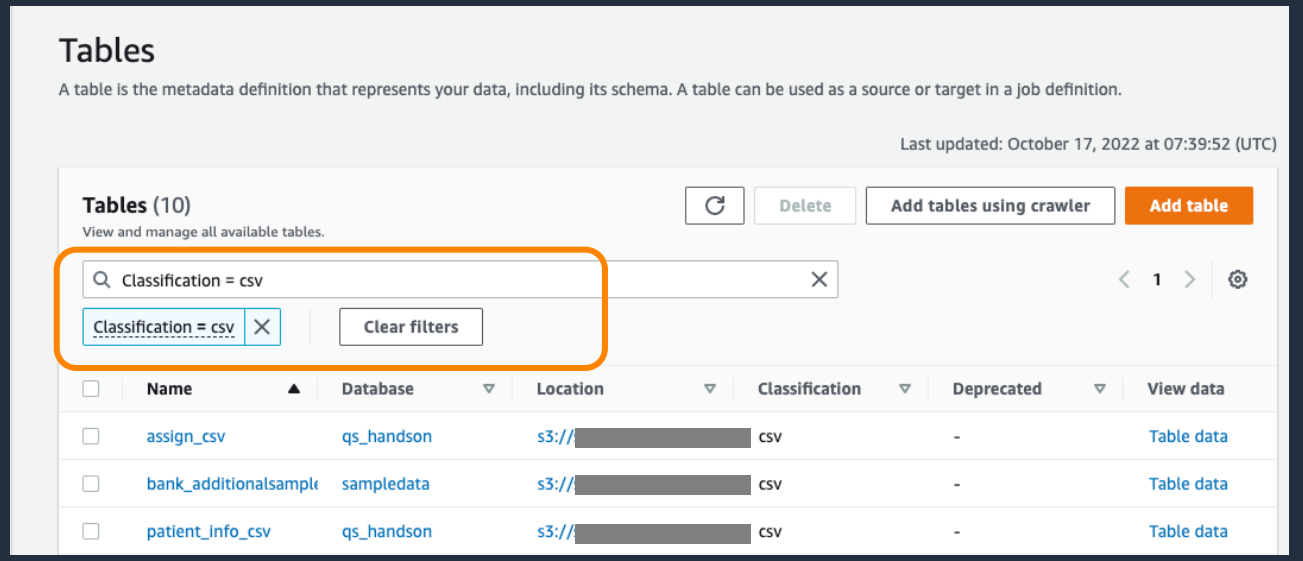


AWS Glue データカタログ

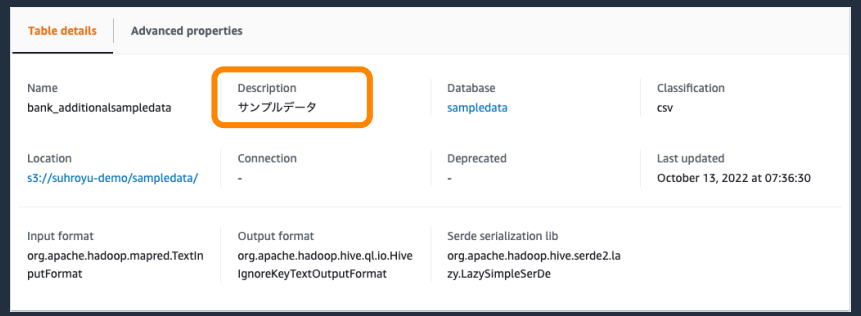


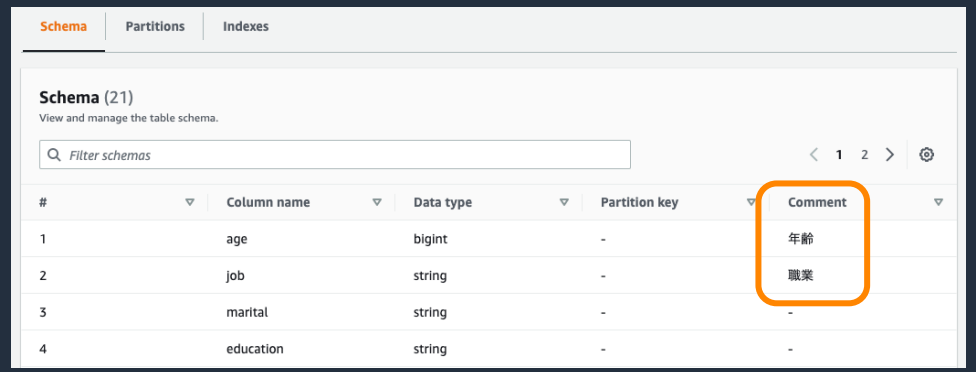
AWS Glue データカタログ

文字列検索



テーブルや項目へのコメント登録





AWS Glue データカタログ

テーブルのスキーマをバージョン管理

- テーブルのスキーマおよびスキーマのバージョンを一覧・比較
- 手動でスキーマ項目を追加、削除、型の変更

スキーマバージョン比較

旧バージョン

変更	列名	データ型	キー
	deviceid	bigint	
	location	string	
	platform	string	
	year	bigint	
	month	bigint	
	day	bigint	

新バージョン

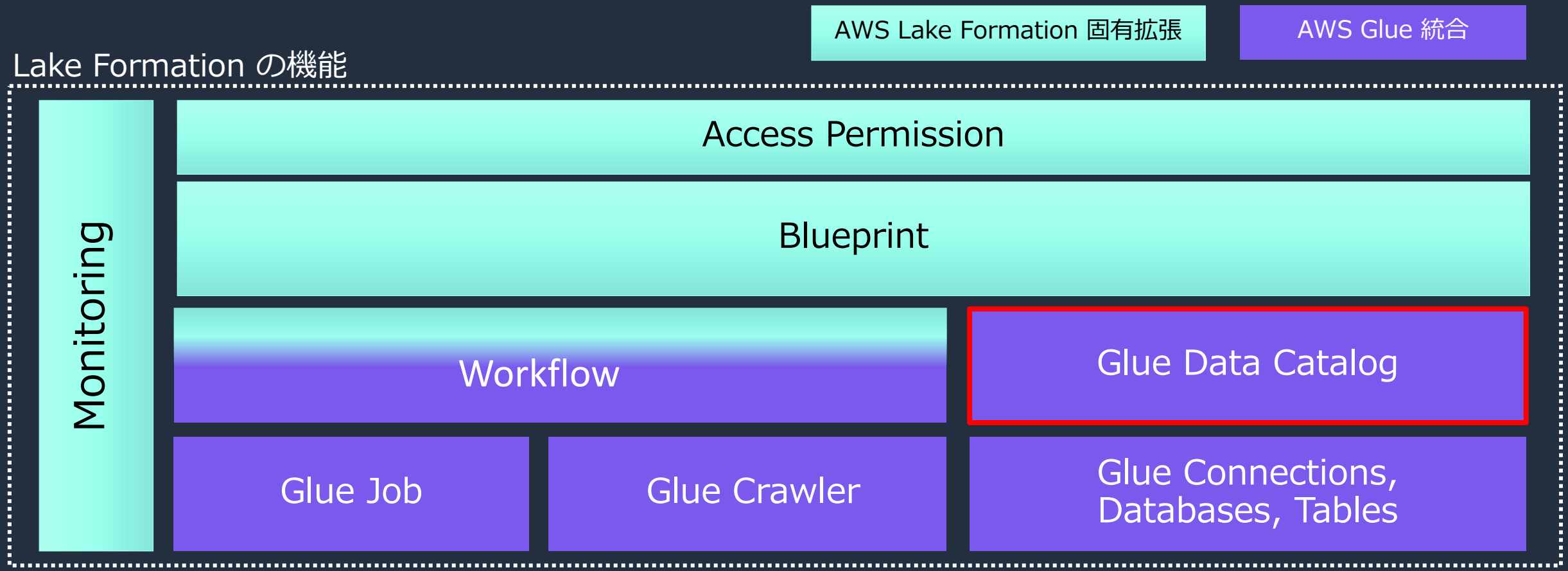
変更	列名	データ型	キー
	deviceid	bigint	
	location	string	
	platform	string	
Added	email	string	
	year	bigint	
	month	bigint	
	day	bigint	

スキーマのバージョンを管理

スキーマ編集画面でデータ型が変更可能

参考）AWS Lake Formationとの関係

Lake Formation は Glue と統合し、機能を拡張する形で、セキュリティ強化や
ブループリントによるデータ取り込み等、**データレイク構築に必要な機能を提供**します



データ収集、カタログ化、データ加工を一気通貫して行うことができる統合的なデータインテグレーションサービス

AWS Glue

データカタログ

データクローラー

他システム連携の
豊富なコネクタ

サーバーレス ETL エンジン

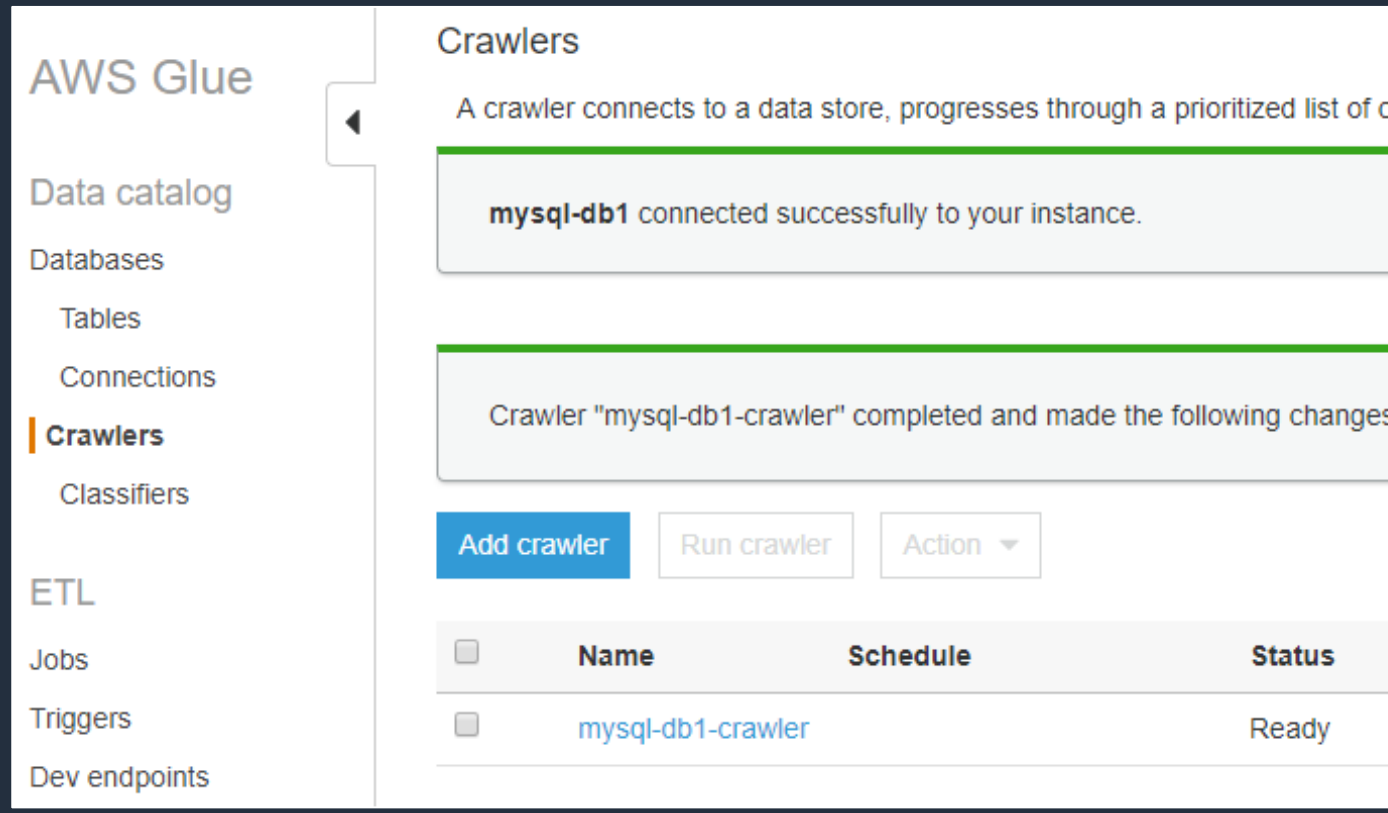
豊富なビルトイン変換と
ローコード開発

スケジューラと
ワークフロー管理

クローラーによるデータカタログの自動更新

- クローラーが自動的にスキーマを推測

  ファイルタイプを識別し、どのような内容が含まれるのかを分類 (Classifier) しスキーマ、ファイルタイプ、パーティションを抽出
- Grok、XML、JSON、CSV フィルタでカスタマイズ可能（カスタム分類子）
- クローラーをスケジュール実行する事で新しいデータやスキーマの変更を発見
- クローラーを使わず手動での登録も可能
- ログは CloudWatch Logs に出力

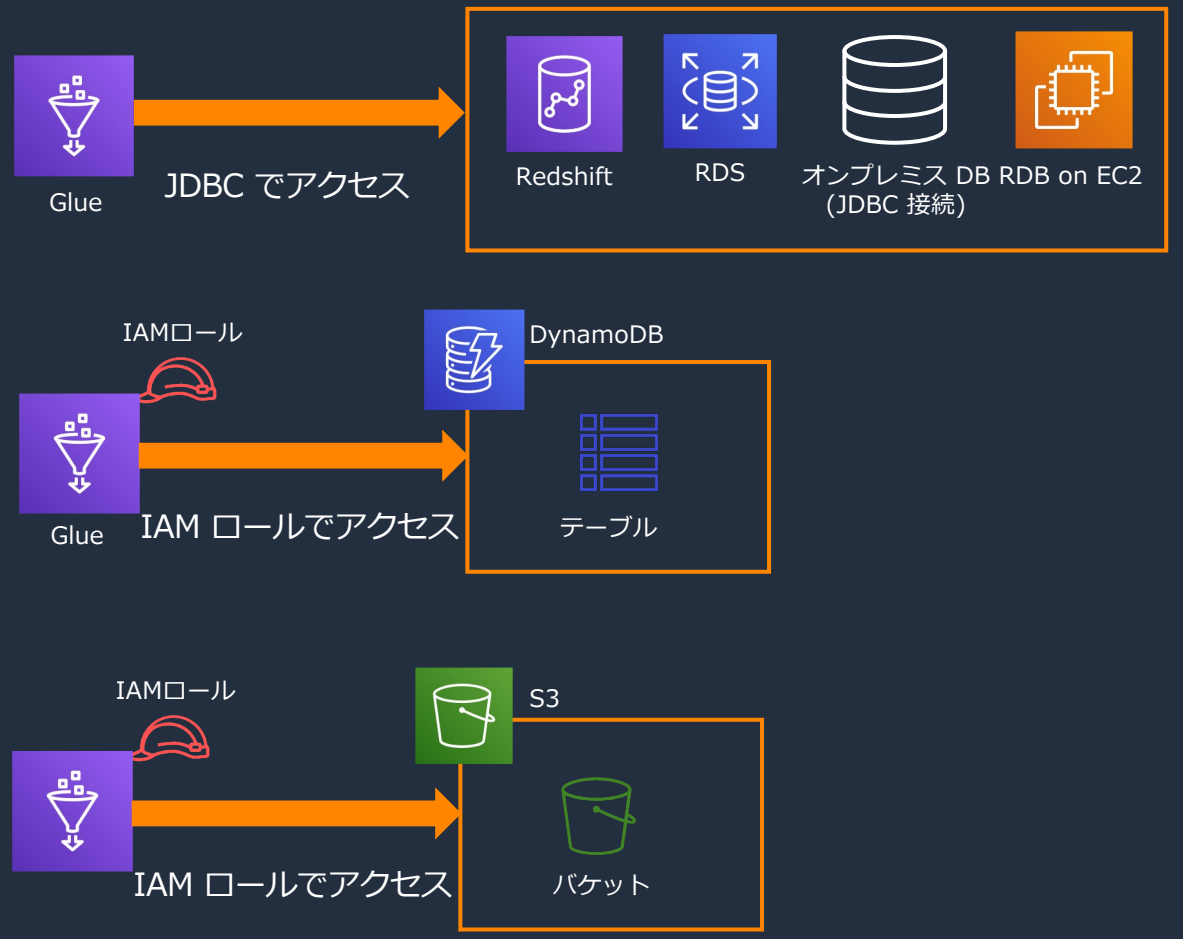


データソース・ターゲットへの接続管理

各種データソースへの接続設定を保存して管理/再利用できる

利用できる ”接続” の種類

- JDBC
- Amazon Redshift
- Amazon RDS
- Amazon DocumentDB
- DynamoDB
- Kafka
- Amazon Kinesis
- MongoDB
- ネットワーク (Amazon VPC 内のデータソースへの接続を指定)
- Amazon S3



データ収集、カタログ化、データ加工を一気通貫して行うことができる統合的なデータインテグレーションサービス

AWS Glue

データカタログ

データクローラー

他システム連携の
豊富なコネクタ

サーバーレス ETL エンジン

豊富なビルトイン変換と
ローコード開発

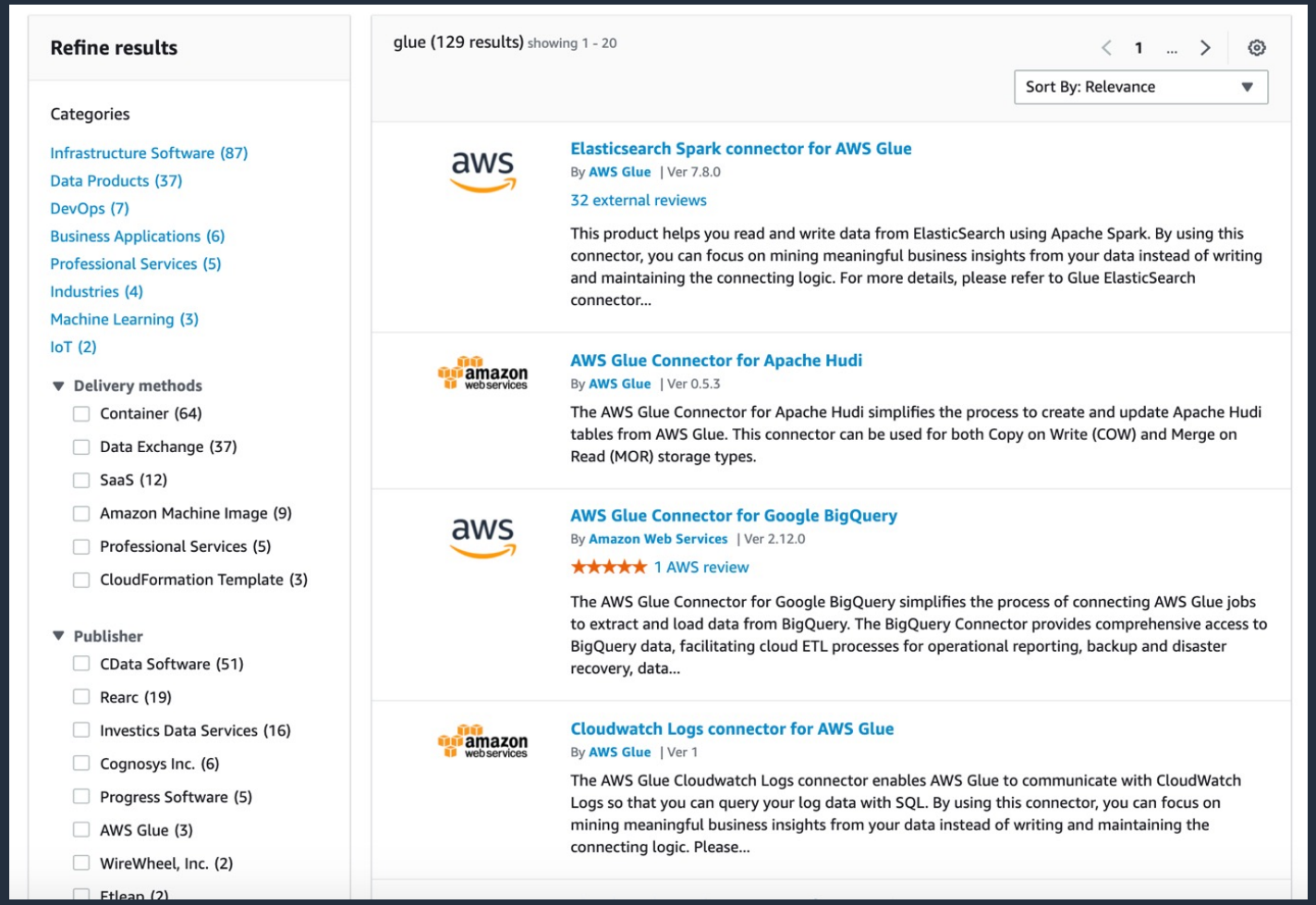
スケジューラと
ワークフロー管理

AWS Glue コネクタ

- 標準で様々なデータソースを選択してデータ連携可能
  - JDBC, Amazon RDS, Amazon Redshift, Amazon DocumentDB, MongoDB, Kafka, Oracle SQL, Microsoft SQL Server, etc.
  - 以下の OSS データレイクフレームワークをネイティブサポート。Amazon S3 に保存するデータに対する読み書きが、一貫性のあるトランザクションにより実行可能
    - Apache Hudi
    - Linux Foundation Delta Lake
    - Apache Iceberg
- 標準搭載していないデータソースは AWS Marketplace からコネクタをサブスクライブして容易に組み込み可能
- 独自のカスタムコネクタを作成可能

AWS Glue コネクタ

AWS Marketplace から subscribe して利用



コネクタ例)

- ✓ Google BigQuery Connector for AWS Glue
- ✓ Snowflake Connector for AWS Glue
- ✓ CData AWS Glue Connector for Salesforce
- ✓ CData AWS Glue Connector for SQL Server
- ✓ CData AWS Glue Connector for SQL Server Analysis Services
- ✓ CData AWS Glue Connector for Google Ads Manager
- ✓ CData AWS Glue Connector for Google Spanner
- ✓ CData AWS Glue Connector for Oracle Sales
- ✓ CData AWS Glue Connector for Couchbase
- ✓ CData AWS Glue Connector for Cassandra
- ✓ CData AWS Glue Connector for SAP Concur
- ✓ DataDirect Cloud Connector for SAP S/4HANA
- ✓ and more.

データ収集、カタログ化、データ加工を一気通貫して行うことができる統合的なデータインテグレーションサービス

AWS Glue

データカタログ

データクローラー

他システム連携の
豊富なコネクタ

サーバーレス ETL エンジン

豊富なビルトイン変換と
ローコード開発

スケジューラと
ワークフロー管理

サーバーレスエンジンとジョブ

ワークロードに合わせた4タイプのサーバーレスジョブ

Spark ジョブ

オープンソースの分散処理システム Apache Spark 環境で実行され、マルチノード環境で大規模なバッチデータ処理を高速に実行できる。

Python シェルジョブ

Python スクリプトをシェルとして実行。このタイプのジョブでは、分散処理を必要としない軽量なタスクを容易に実行できる。

ストリーミング ETL ジョブ

ストリーミングソース* からのデータを、Apache Spark 構造化ストリーミングフレームワークを使用してリアルタイムに処理する。

* 例：Amazon Kinesis Data Streams, Apache Kafka, Amazon MSK

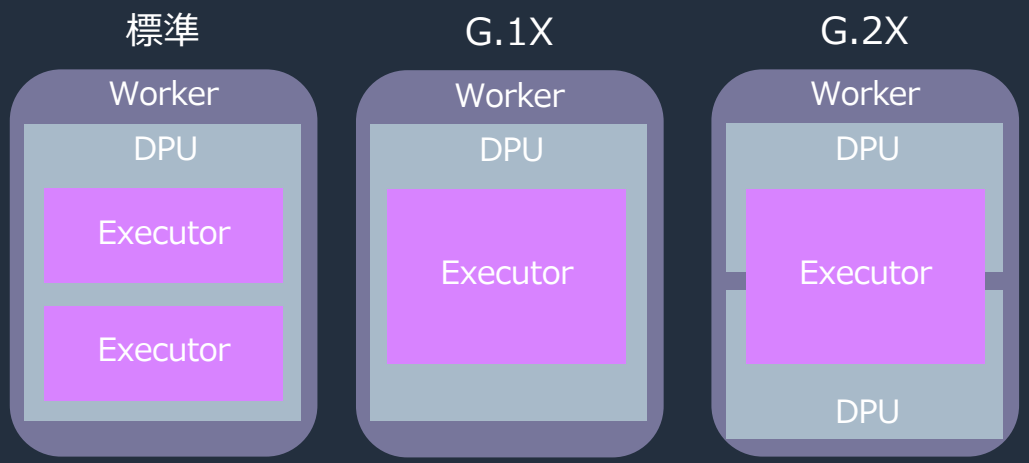
Ray ジョブ

スケーラブルな Python ワークロードを実現可能なオープンソースフレームワーク Ray を用いて、マルチノード環境で高速に実行できる。

Glue Spark ジョブの実行環境

- Worker ノードが複数起動、 Worker 内で「Executor」が起動し、分散処理を行う
- Worker と Executor の性能と数は DPU , Worker Type で決まる
  - DPU (Data Processing Unit) はジョブ実行時に割り当てる処理能力
    - 1DPU = 4 vCPU、16 GB メモリ
  - Worker Type はWorker の種類：標準, G.1X, G.2X, G.025X*

* AWS Glue バージョン 3.0 ストリーミングジョブでのみ利用可能なワーカータイプ



Worker 構成イメージ

Worker Type	DPU数 /1Worker	Executor数 /1Worker	メモリ /1Worker	ディスク /1Worker
標準	1	2	16GB	50GB
G.1X	1	1	16GB	64GB
G.2X	2	1	32GB	128GB
G.025X*	0.25	1	4GB	64GB

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

AWS Glue のバージョン

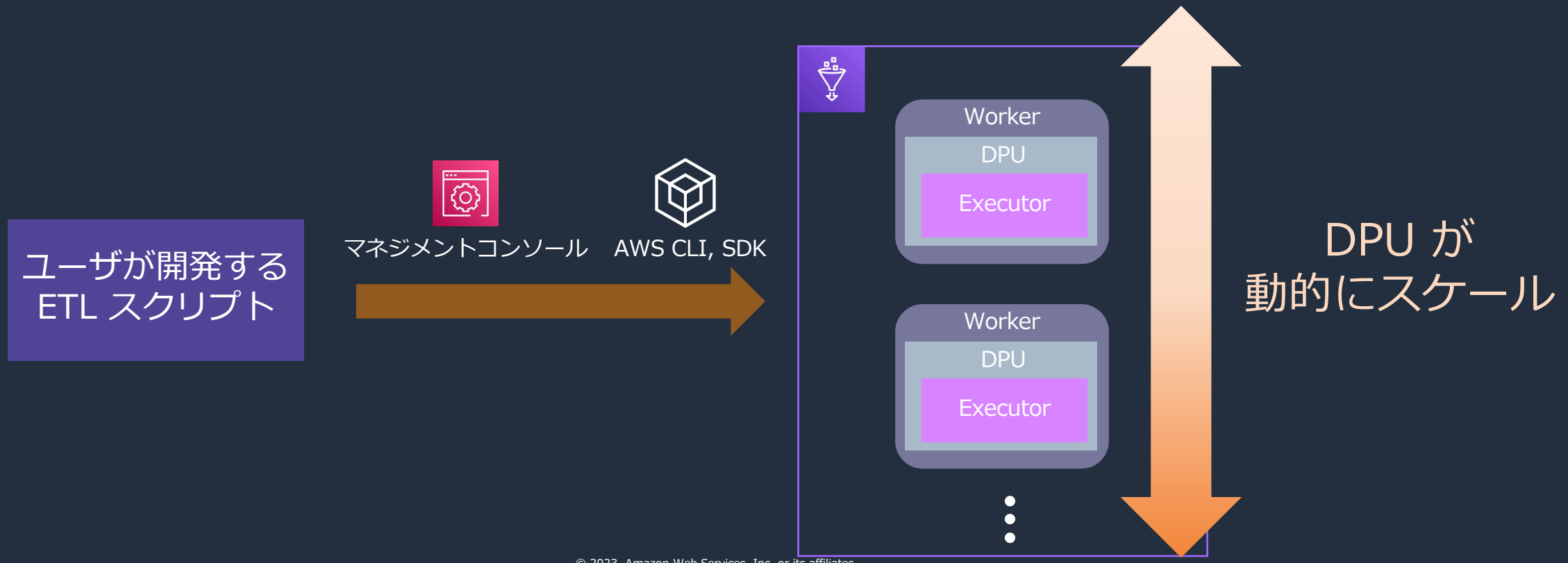
最新のバージョンを選択することで、最新機能や高いパフォーマンスを得られるメリットがあります

バージョン	Apache Spark のバージョン
Glue 4.0	Spark 3.3.0 (Python 3.10、Scala 2.12)
Glue 3.0	Spark 3.1.1 (Python 3.7、Scala 2.12)
Glue 2.0 ※2024年1月31日サポート終了予定	Spark 2.4.3 (Python 3.7、Scala 2.11)
Glue 1.0 ※ 2022年9月30日サポート終了	Spark 2.4.3 (Python 2.7、Python 3.6、Scala 2.11)
Glue 0.9 ※ 2022年6月1日サポート終了	Spark 2.2.1 (Python 2.7、Scala 2.11)

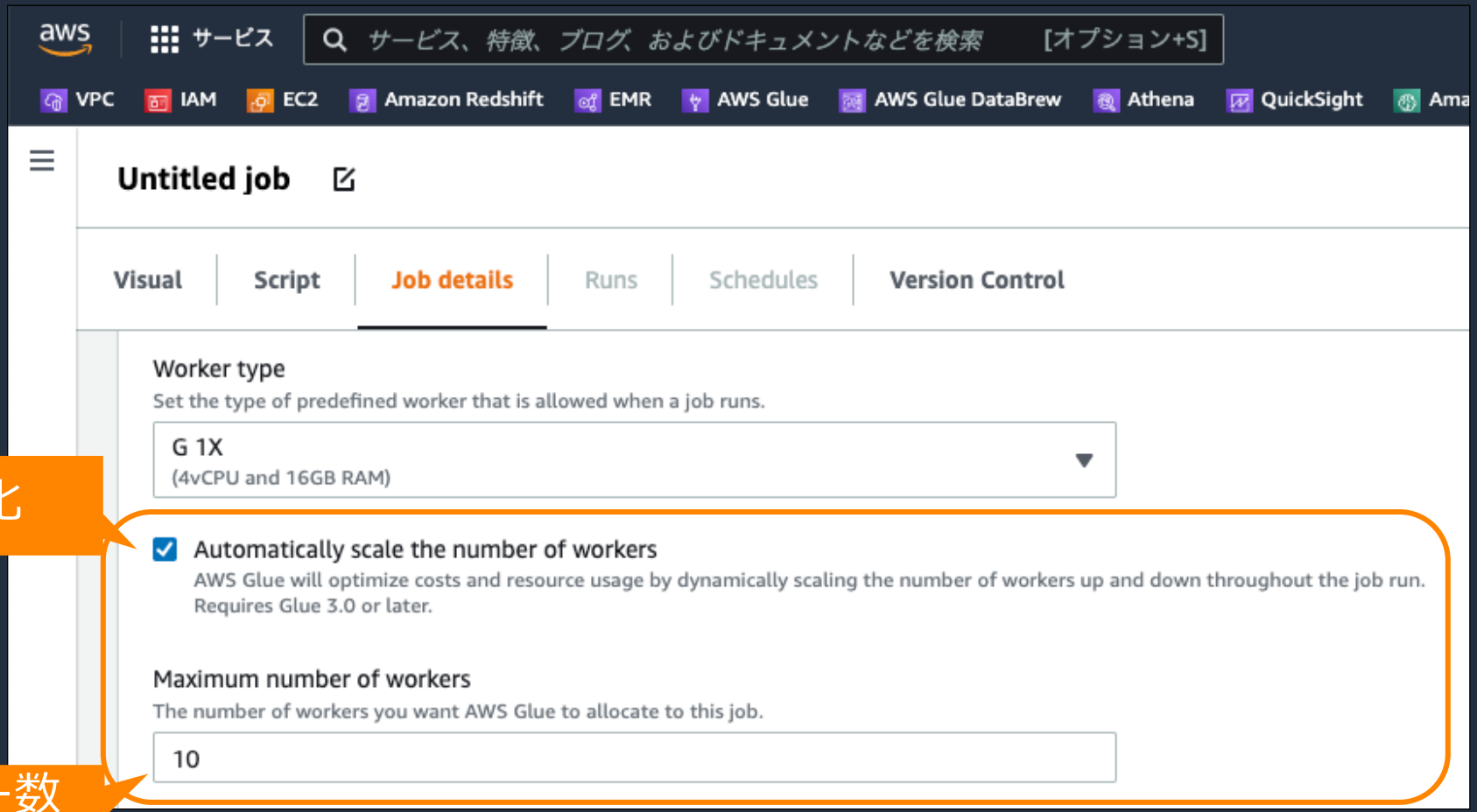
参考URL：
- Glue バージョン　(https://docs.aws.amazon.com/ja_jp/glue/latest/dg/release-notes.html)
- バージョンサポートポリシー（ https://docs.aws.amazon.com/ja_jp/glue/latest/dg/glue-version-support-policy.html ）
- 最新バージョンへの移行方法（https://docs.aws.amazon.com/ja_jp/glue/latest/dg/migrating-version-40.html)

オートスケーリング

- 事前に指定した範囲で負荷に応じて DPU が動的にスケール
- コストとリソースのバランスを自動で最適化
- Glue 3.0 以上の Spark ジョブまたはストリーミング ETL ジョブで対応



オートスケーリング



有効化

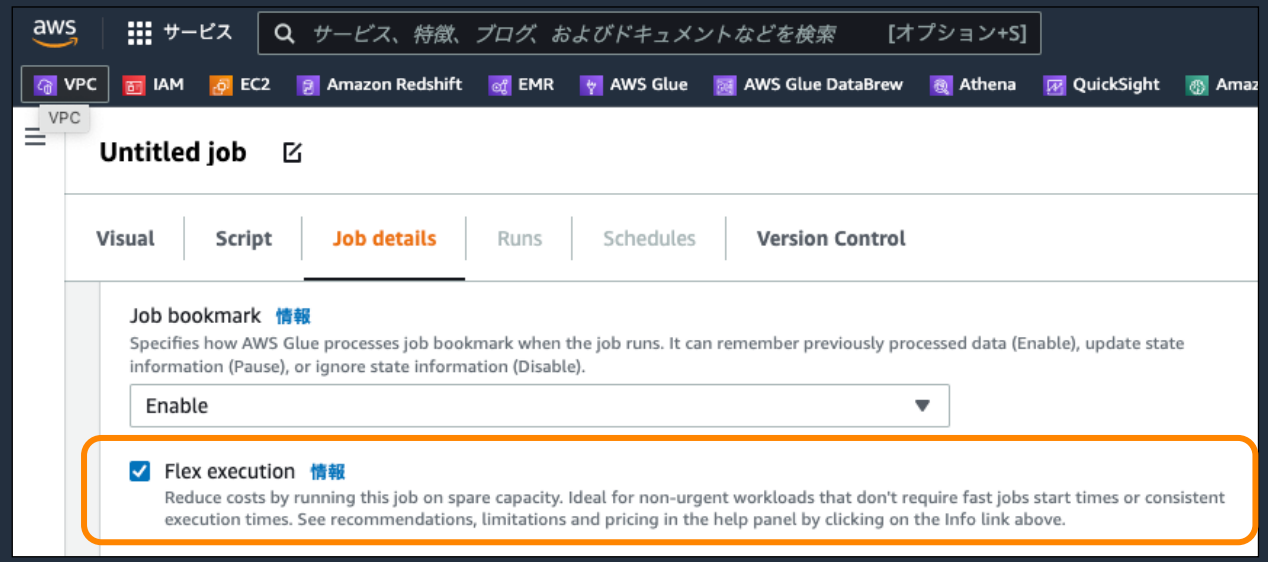
最大ワーカー数指定

Flex 実行オプション

- コストを最大 34% 削減する、新しい実行オプション
- AWS の予備キャパシティで実行される
- 想定ユースケース
  - プリプロダクションデータ
  - テストデータ
  - 緊急度が低いデータの統合ワークロード

考慮事項)

- バージョン3.0 以上の Spark ジョブのみサポート
- 予備キャパシティ状況によってジョブの中断や開始の保留、または実行中リソースの回収 (reclaim) などにより、指定したワーカー数が割り当てられず実行時間を要するリスクあり



https://aws.amazon.com/jp/blogs/big-data/introducing-aws-glue-flex-jobs-cost-savings-on-etl-workloads/

Python Shell ジョブ

- Spark の分散処理を必要としない小～中規模ジョブ向け

  例）
  - ✓ DB や DWH へ SQL 実行
  - ✓ Numpy や pandas で集計処理
  - ✓ 外部サービスの実行

- 右表に示す豊富なライブラリを標準サポート
- PIP、Egg*、Whl ファイルによるライブラリ追加も可能

ライブラリ	Python 3.9
avro	1.11.0
awscli	1.23.5
awswrangler	2.15.1
botocore	1.23.5
boto3	1.22.5
Elasticsearch	8.2.0
numpy	1.22.3
pandas	1.4.2
psycopg2	2.9.3
pyathena	2.5.3
PyGreSQL	
PyMySQL	1.0.2
pyodbc	4.0.32
pyorc	0.6.0
redshift-connector	2.0.907
requests	2.27.1
scikit-learn	1.0.2
scipy	1.8.0
SQLAlchemy	1.4.36
s3fs	2022.3.0

* Python3.9 サポート外

https://docs.aws.amazon.com/ja_jp/glue/latest/dg/add-job-python.html

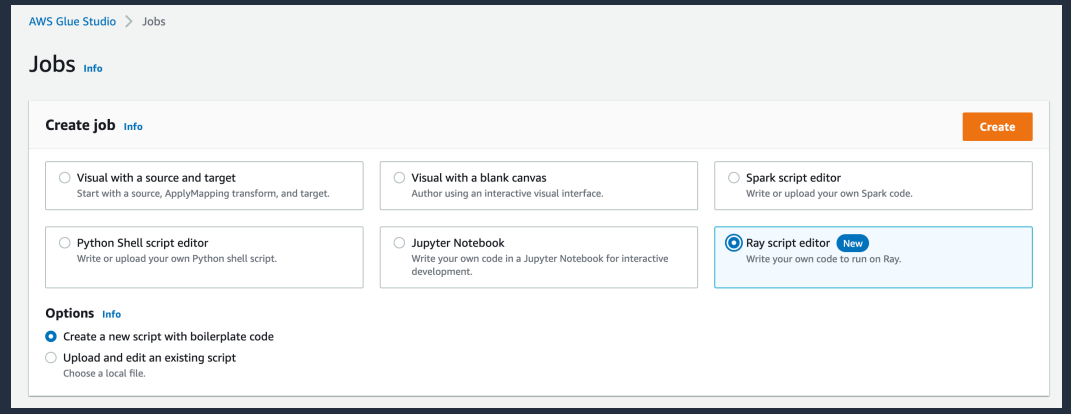
aws

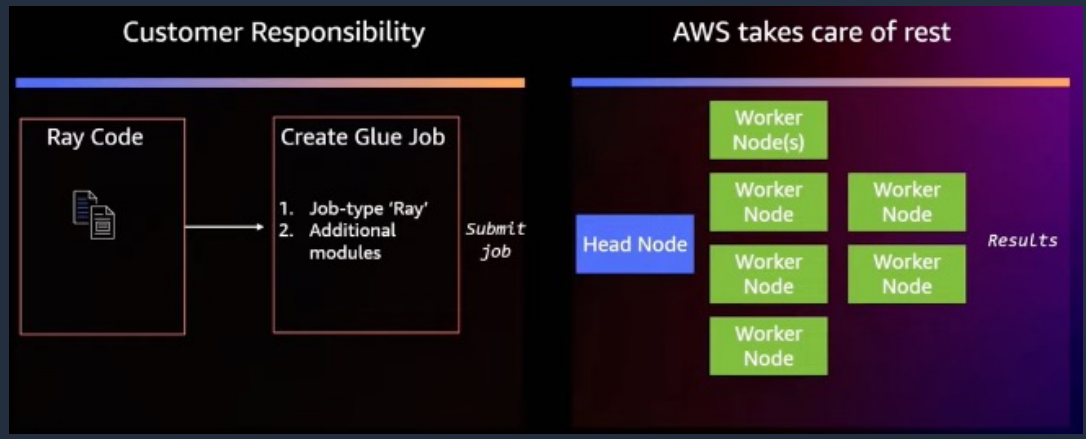
Ray ジョブ

新たな Python の分散並列処理のフレームワークが使用可能

※ 2023年3月時点でプレビュー提供中

- Ray は Python で並列分散処理を実現するためのライブラリ：https://www.ray.io/
- マルチノードクラスターでの Python コードの分散処理をサーバーレスで容易に実現できる
- インタラクティブセッションでの利用も可能
- Pandas や AWS SDK for Pandas など、使い慣れたオープンソースライブラリをそのまま使用できる
- メモリ割り当ての多い M-DPU（ vCPU 4 ， MEM 32 GB ） によりジョブが実行される





公式ドキュメント：https://docs.aws.amazon.com/glue/latest/dg/author-job-ray.html

参考）サーバーレス ETL 処理の使い分け

データの規模やETL処理の中でやりたいことによって使い分ける

小規模処理

AWS Lambda

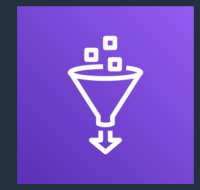
- 15分以内に完了できる処理
- 豊富なトリガー(S3に配置されたタイミングで逐次処理)
- Pandasなどのライブラリが利用可能
- Python以外の豊富な実行環境

小規模処理

AWS Glue
Python Shell

- 実行時間の制限なし
- Lambda に比べてメモリ量が多い(1GBまたは16GB(※1))
- Pandas などのライブラリが利用可能
- Redshift や EMR、Athena に対する SQL ベースの分析

中～大規模処理



AWS Glue
Apache Spark
Ray

- 実行時間の制限なし
- 並列分散処理
- Spark または Ray を利用

(※1) Python Shell では0.0625 (1/16) DPU と 1 DPU が選択でき、その計算結果を記載。

データ収集、カタログ化、データ加工を一気通貫して行うことができる統合的なデータインテグレーションサービス

AWS Glue

データカタログ

データクローラー

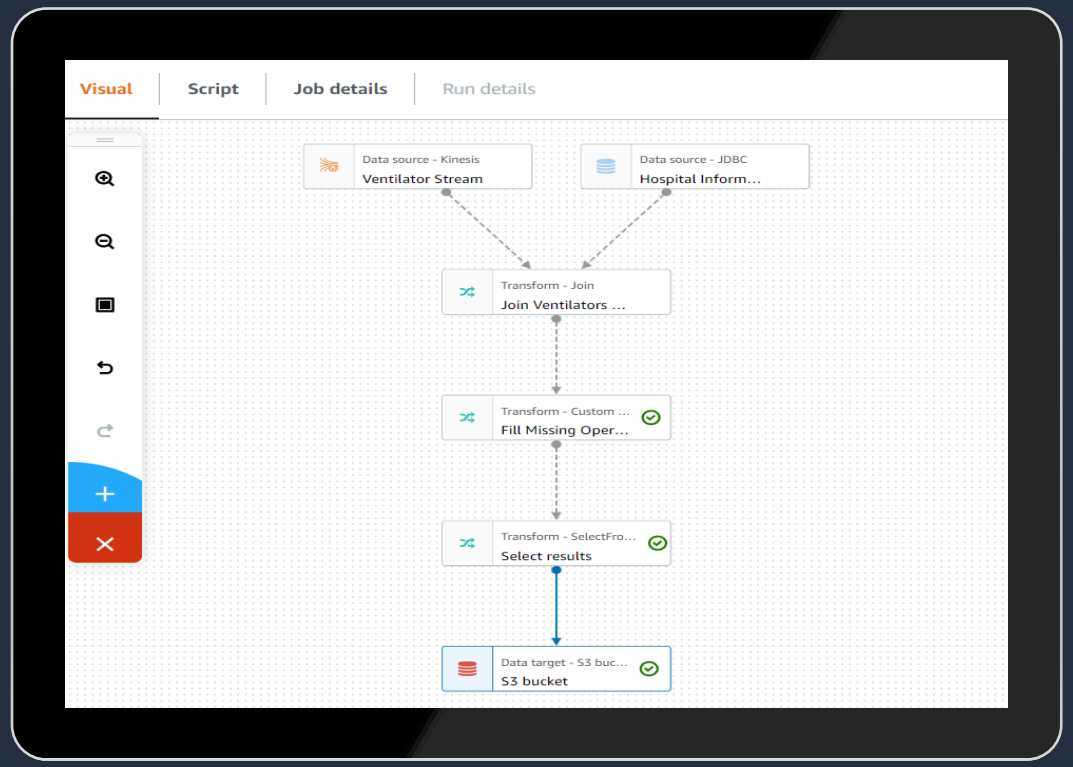
他システム連携の
豊富なコネクタ

サーバーレス ETL エンジン

豊富なビルトイン変換と
ローコード開発

スケジューラと
ワークフロー管理

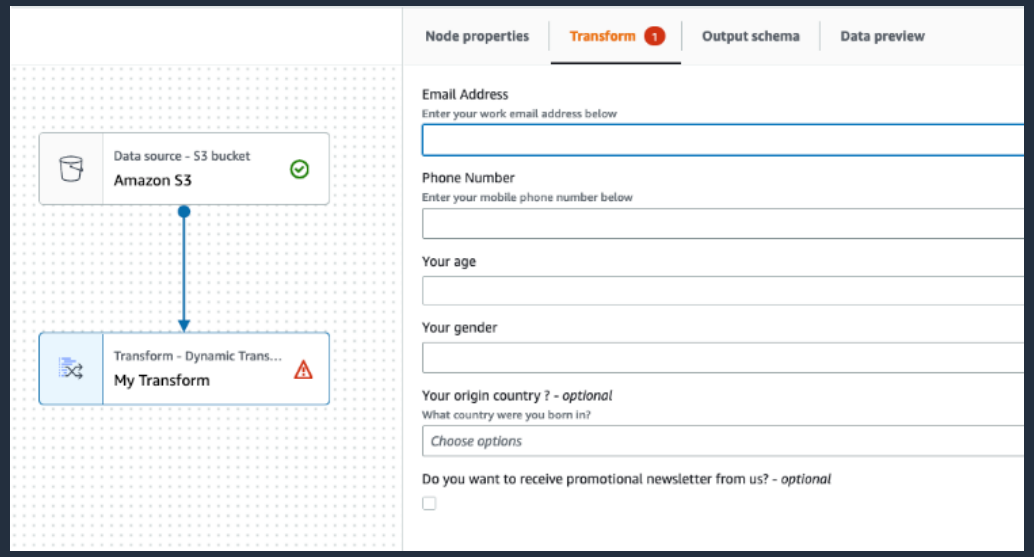
Glue Studio



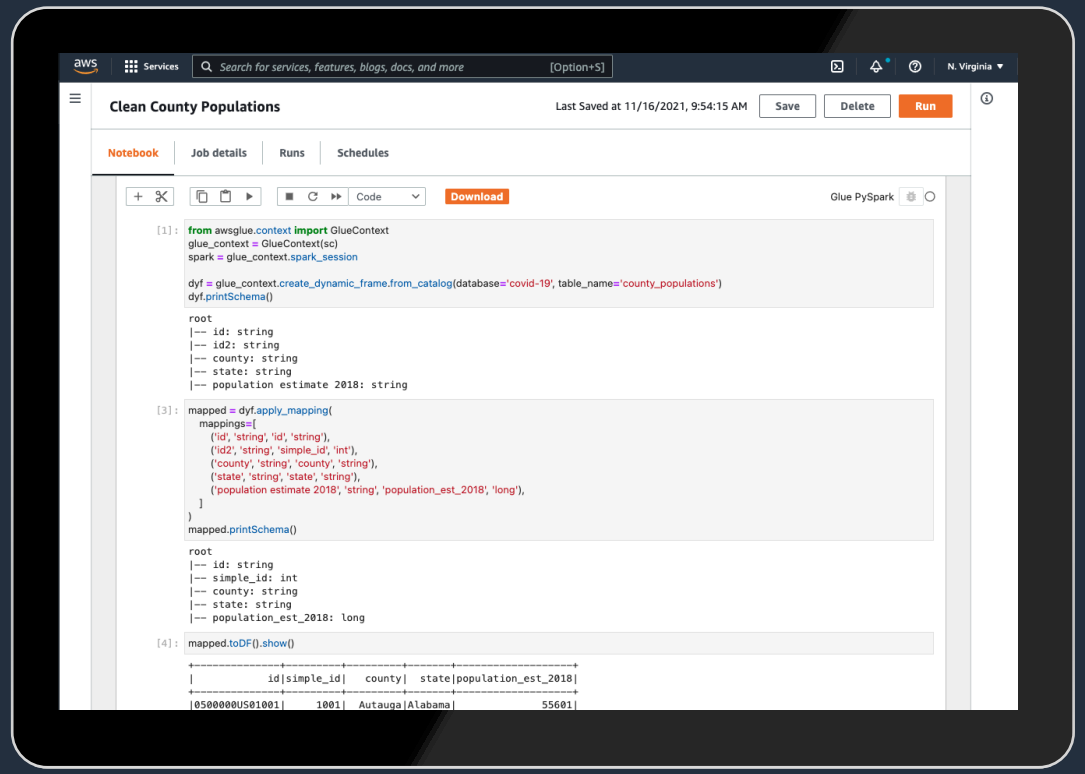
- ETL ジョブの作成、実行、およびモニタリングを簡単にできるようにする新しいグラフィカルインターフェイス
- GUI で作成したフローからソースコードを生成
- データ抽出、出力、ジョイン等の汎用的な処理はビルトインされノンコーディングで実現
- 分散処理を意識せずに開発

Glue Studio：Custom Visual Transform 機能

- 独自の 変換処理 をカスタムビジュアルエディタ向けに自由に定義することができる
  - ✓Apache Spark コードおよびユーザ入力フォームで定義可能
  - ✓登録した変換はドロップダウンリストから選択可能
  - ✓AWS アカウント間の共有も可能
- 再利用可能な変換処理を定義することで、チーム内の作業効率化
- ビジュアルベースの作業範囲を拡張



Glue Studio Notebook



- Jupyter Notebook を利用してインタラクティブな ETL コード開発が可能
- 作成したノートブックを１時間毎、１日毎などでスケジューリング実行可能
- インストールなしでインタラクティブセッションを利用可能

Glue インタラクティブセッション

任意の環境の Jupyter Notebook から Glue ETL 環境を利用して
インタラクティブにコードを開発するためのサーバーレス実行環境
※ Glue 2.0 / 3.0 で対応

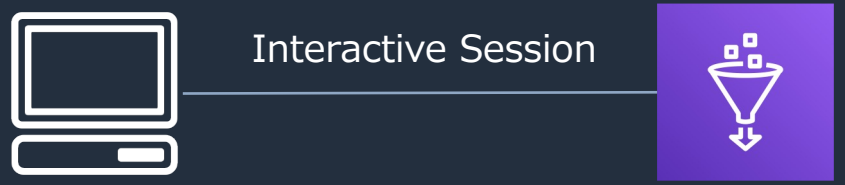
ローカルPC上で Jupyter Notebook を起動
してインタラクティブにジョブを作成

```
In [1]: spark
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1576209965005_0001	pyspark	idle	Link	Link	✓

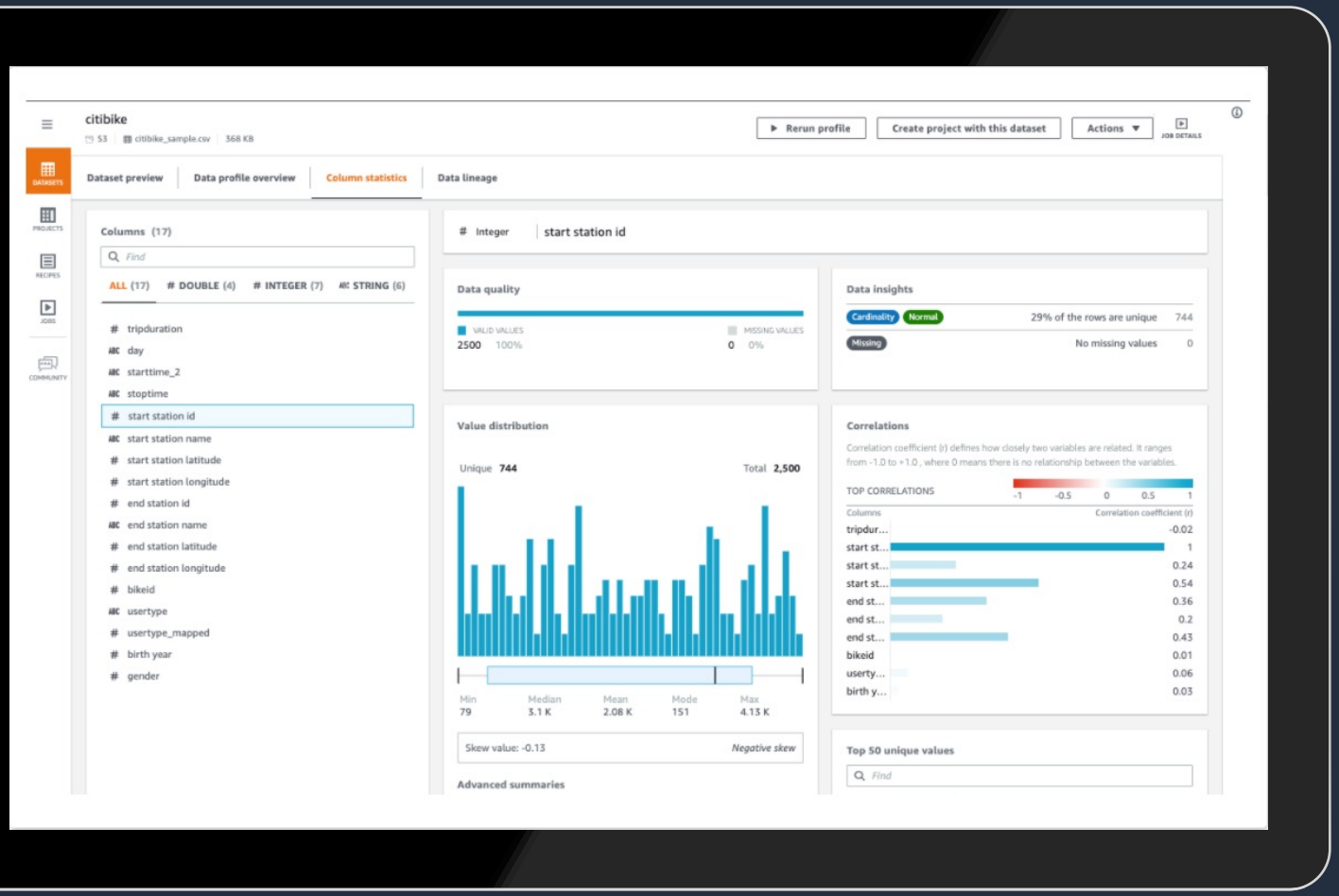
```
SparkSession available as 'spark'.
<pyspark.sql.session.SparkSession object at 0x7f3d54913550>
```

Jupyter Notebook 上でジョブを実行すると
AWS Glue ETL でジョブが実行される



AWS Glue DataBrew

ノーコードデータプレパレーションサービス



Business Analyst
Data Scientist

豊富なビジュアルインターフェイス

250 種類以上の組み込みトランスフォーメーション

データパターンと異常を把握するためのプロファイリングデータ

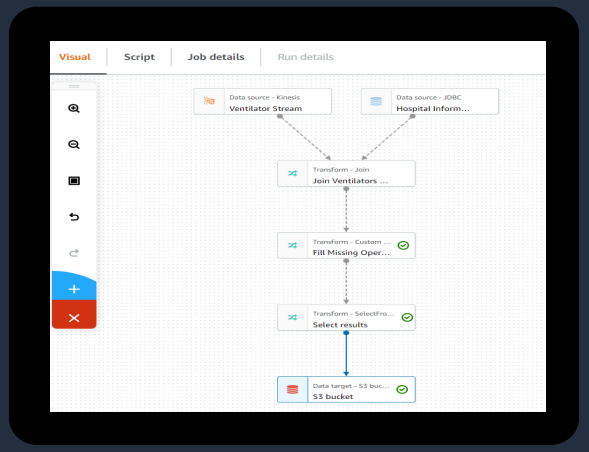
大規模なデータセットを大規模に処理

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

参考）全てのデータ利用者向けにデザインされた選択肢

ETL developer
Data Engineer

AWS Glue Studio



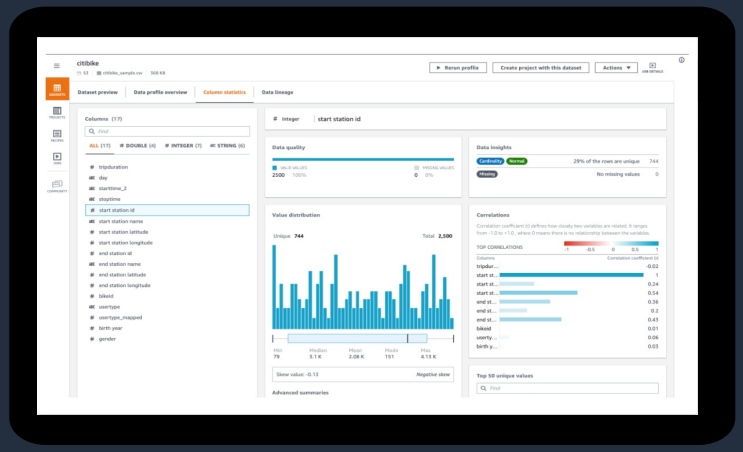
ローコードで迅速に開発

コーディングで柔軟にカスタマイズ

用途別の実行エンジン

Business Analyst
Data scientist

AWS Glue DataBrew



ノーコードでセルフサービス型データ準備

250以上のビルトイン変換機能

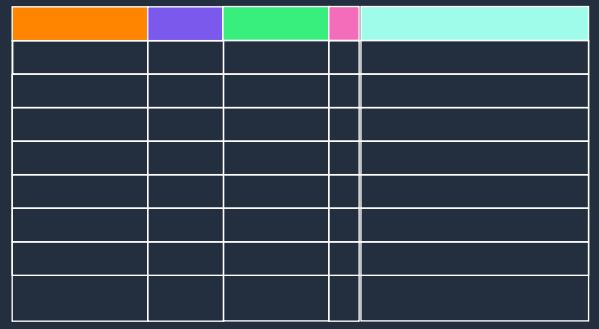
データプロファイル

DynamicFrame

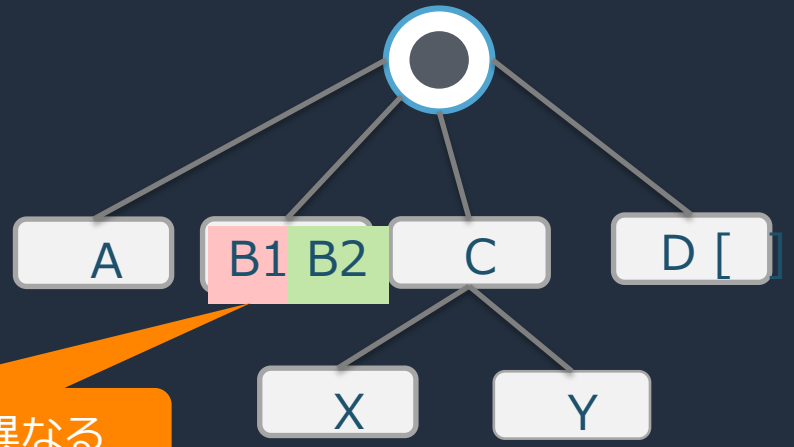
複雑な ETL 処理をシンプルに行うことができるフレームワーク

- Apache Spark の DataFrame に似ているが、JSON、Avro、Apache ログなどの半構造化データのクレンジングや再構成のための機能が豊富
- スキーマの不整合に対して柔軟にデータ取り込みを行うことができる、Schema on the fly の採用により、変換処理をシングルパスで実行可能
- Apache Spark の DataFrame とは fromDF 関数、 toDF 関数を用いて相互にゼロコピーで変換可能

DataFrame Schema



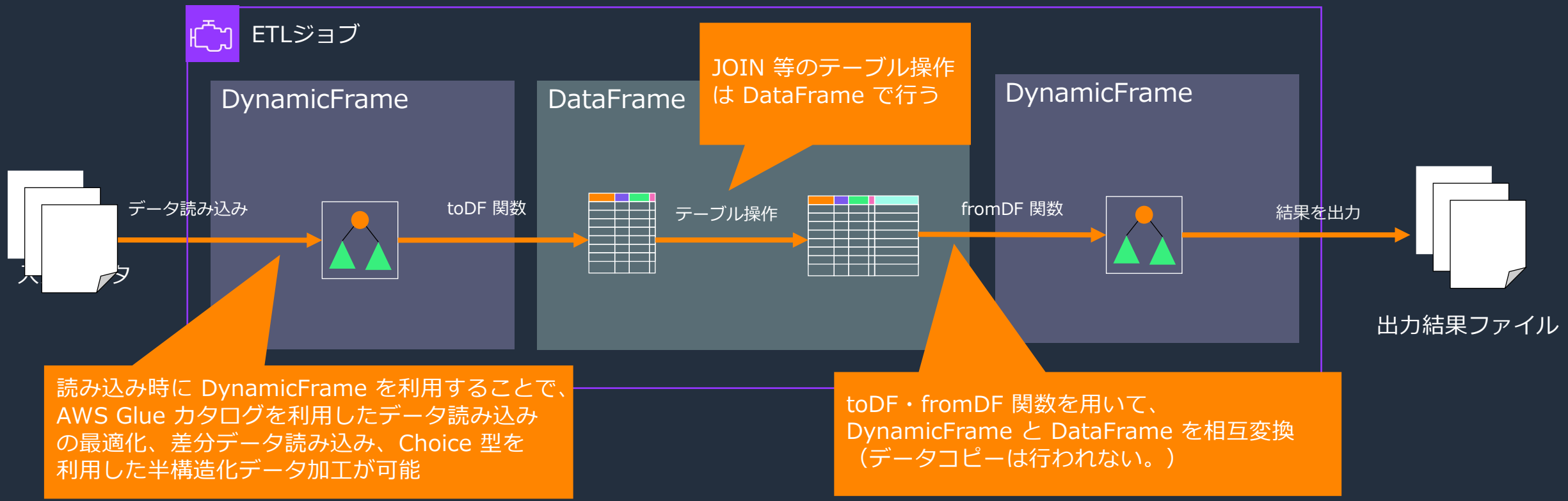
DynamicFrame schema



レコードごとに型が異なる

DynamicFrame と DataFrame の特性を活かした処理

- DynamicFrame はETL処理に強く、DataFrame はテーブル処理に強い
- データ入出力とそれに伴うETL処理は DynamicFrame で行い、テーブル操作は DataFrame で行う



Glue Data Quality

データレイクのデータの品質を自動的に測定・モニタリング・管理

※ 2023年3月時点でプレビュー提供中

- ルールに基づきデータ品質をモニタリング
- データを自動的に分析しデータに関する統計を取得し、適用すべきルールをレコメンデーション
- データ抽出、変換処理、ローディング処理の各ステップでルール定義と警告を設定可能
- オープンソースフレームワークである Deequ を使用して構築されているため、ロックインのない柔軟性と移植性を提供

公式ブログ：https://aws.amazon.com/blogs/aws/join-the-preview-aws-glue-data-quality/

個人識別情報（PII）を自動識別

- データソースをスキャンし、PII を自動識別
- 識別された PII は自動削除または **** など固定文字列へ変換/マスク可能
- ビルトインの検出パターンの他、正規表現でカスタムパターンを定義可能
- 検出可能なビルトインの PII パターン例

Global entities

- PERSON_NAME (人名)
- EMAIL (eメールアドレス)
- IP_ADDRESS (IP アドレス)
- MAC_ADDRESS (MAC アドレス)

※ 2023年3月時点で、日本語表記の氏名の検知は未対応

US data types

- USA_SSN (アメリカの SSN)
- USA_ITIN (アメリカの ITIN)
- PHONE_NUMBER (電話番号)
- USA_PASSPORT_NUMBER (アメリカのパスポート番号)
- USA_DRIVING_LICENSE (アメリカの運転免許番号)
- BANK_ACCOUNT (銀行口座)
- CREDIT_CARD (クレジットカード番号)
- その他

Japan data types

- JAPAN_BANK_ACCOUNT (銀行口座)
- JAPAN_DRIVING_LICENSE (運転免許番号)
- JAPAN_MY_NUMBER (マイナンバー)
- JAPAN_PASSPORT_NUMBER (パスポート番号)

https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-sensitive-data-api-sensitive-data-api-example.html

データ収集、カタログ化、データ加工を一気通貫して行うことができる統合的なデータインテグレーションサービス

AWS Glue

データカタログ

データクローラー

他システム連携の
豊富なコネクタ

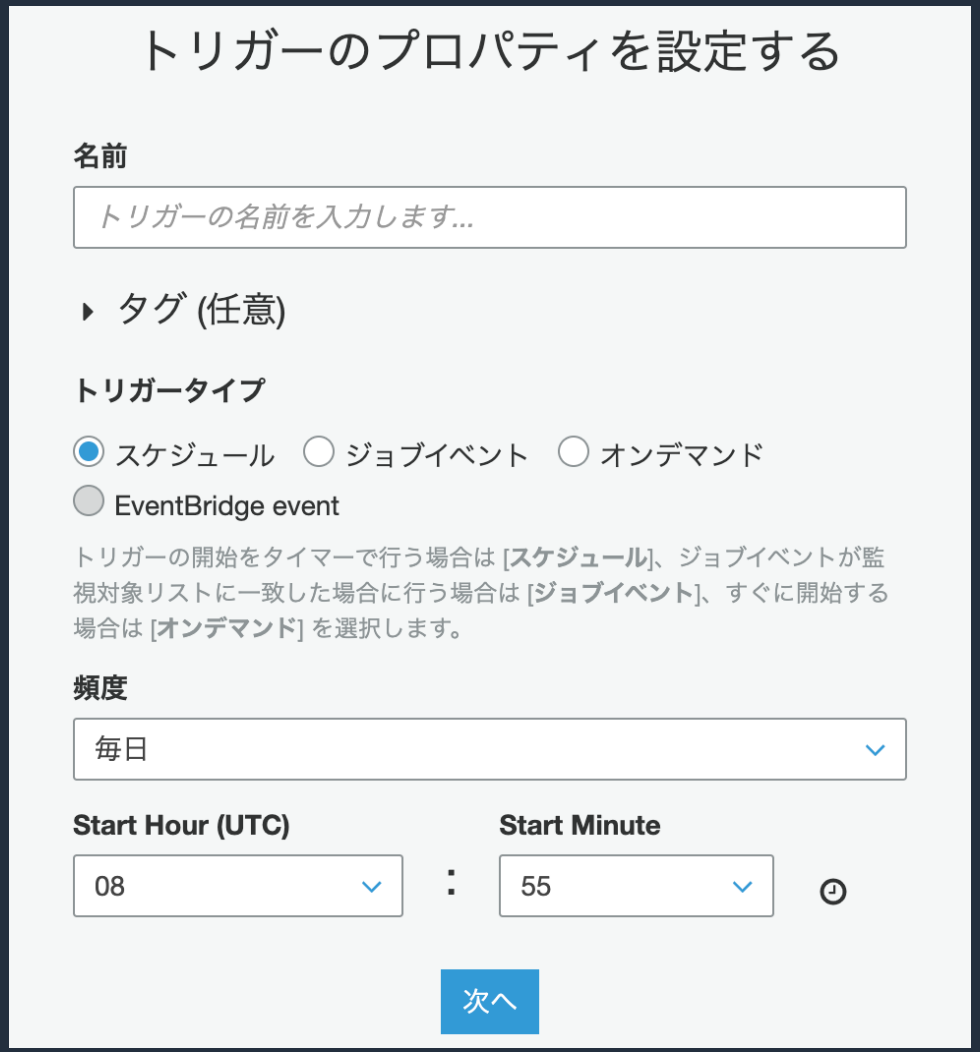
サーバーレス ETL エンジン

豊富なビルトイン変換と
ローコード開発

スケジューラと
ワークフロー管理

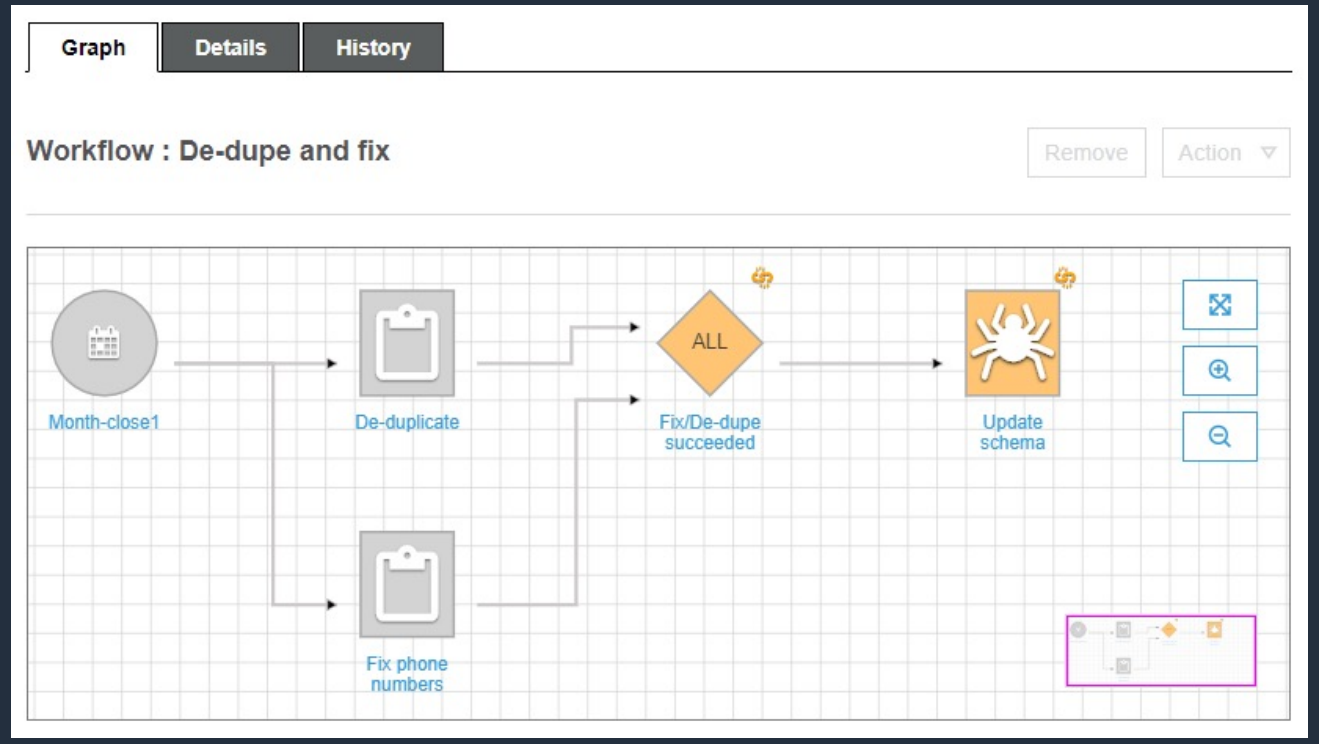
ジョブのスケジュールとトリガー

- ジョブの実行を自動制御機能
- トリガータイプ
  - ✓ スケジュール：cron に基づく時間ベーストリガー
  - ✓ ジョブイベント：対象ジョブまたはクローラーの実行結果に基づきトリガーする
  - ✓ オンデマンド
- EventBridge からのトリガーも可能



ワークフロー管理

- 複数の Glue ETL ジョブの実行タイミングや依存関係を管理
- クローラ、トリガー、ジョブの組み合わせによりワークフローを定義
- トリガーはスケジュール(日時・曜日・cron)、ジョブイベント、Amazon EventBridge、手動(即時実行)で指定可能



参考）ワークフロー制御サービスの選択肢

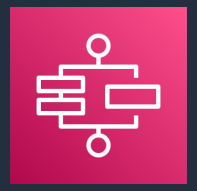
AWS で実践！Analytics Modernization ～ETL 編～
https://aws.amazon.com/jp/blogs/news/aws-analytics-modernization-etl-2021/

Amazon MWAA
（Apache Airflow）

大規模ETL+OSS

Apache Airflow

- ETL 運用に最適化した視認性の高い運用 GUI
- デバッグやリトライ制御の操作性
- Python 言語でパイプライン構築
- AWSに閉じない各種サービスとの連携（ Operator ）
- 完全閉域でのダッシュボード利用
- OSS ベース（ハイブリッド環境）



AWS Step Functions

汎用性＋モダン

- サーバーレス（高い保守性/拡張性/低コスト）
- ETLに限らない高い汎用性
- 各種 AWS サービスとの統合・連携容易性
- GUI ベースの迅速かつ容易なパイプライン構築
  ※または Json/Yaml ベースの定義

AWS Glue
Workflow

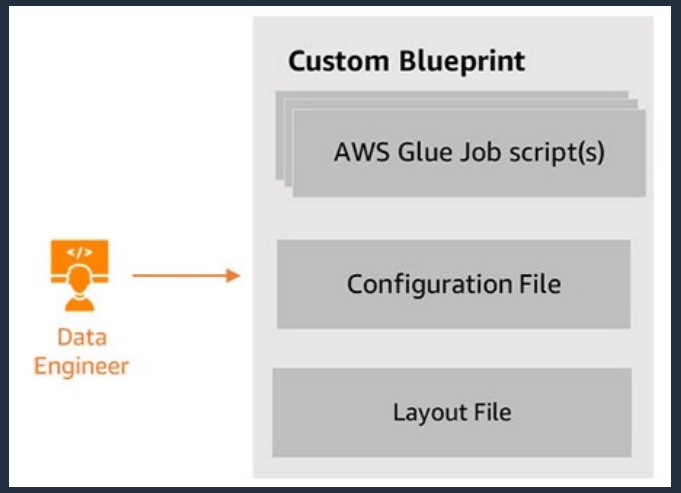
Glue 内で全て完結

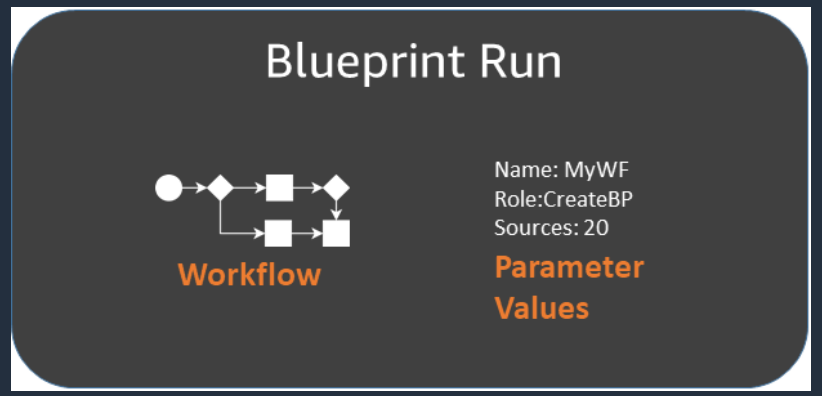
- サーバレス （高い保守性/拡張性/低コスト）
- GUI ベースのパイプライン構築
- Blueprint 機能による共通テンプレート化
- Glue ジョブとの統合・連携容易性

カスタムブループリント

汎用化可能なワークフローをテンプレート化して共有

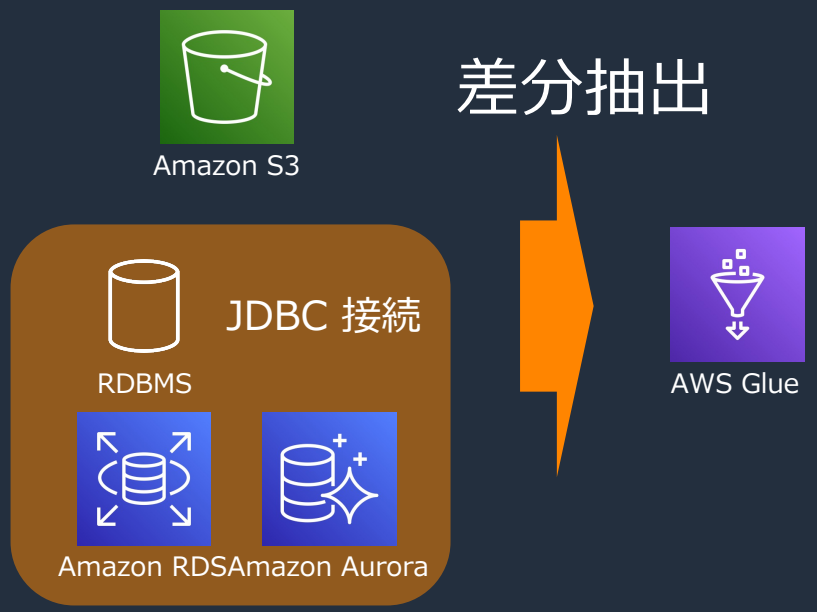
- 入力パラメータを規定でき、例えば入力データソース（S3 パス）やターゲットデータソース（データベース）を可変とし、共通テンプレート（blueprint）を流用
- 例えばデータエンジニアが blueprint を提供し、分析者は任意の blueprint を選択し、セルフサービスで複雑なデータ連携ジョブを実行





https://aws.amazon.com/jp/blogs/news/simplify-data-integration-pipeline-development-using-aws-glue-custom-blueprints/

ブックマーク機能

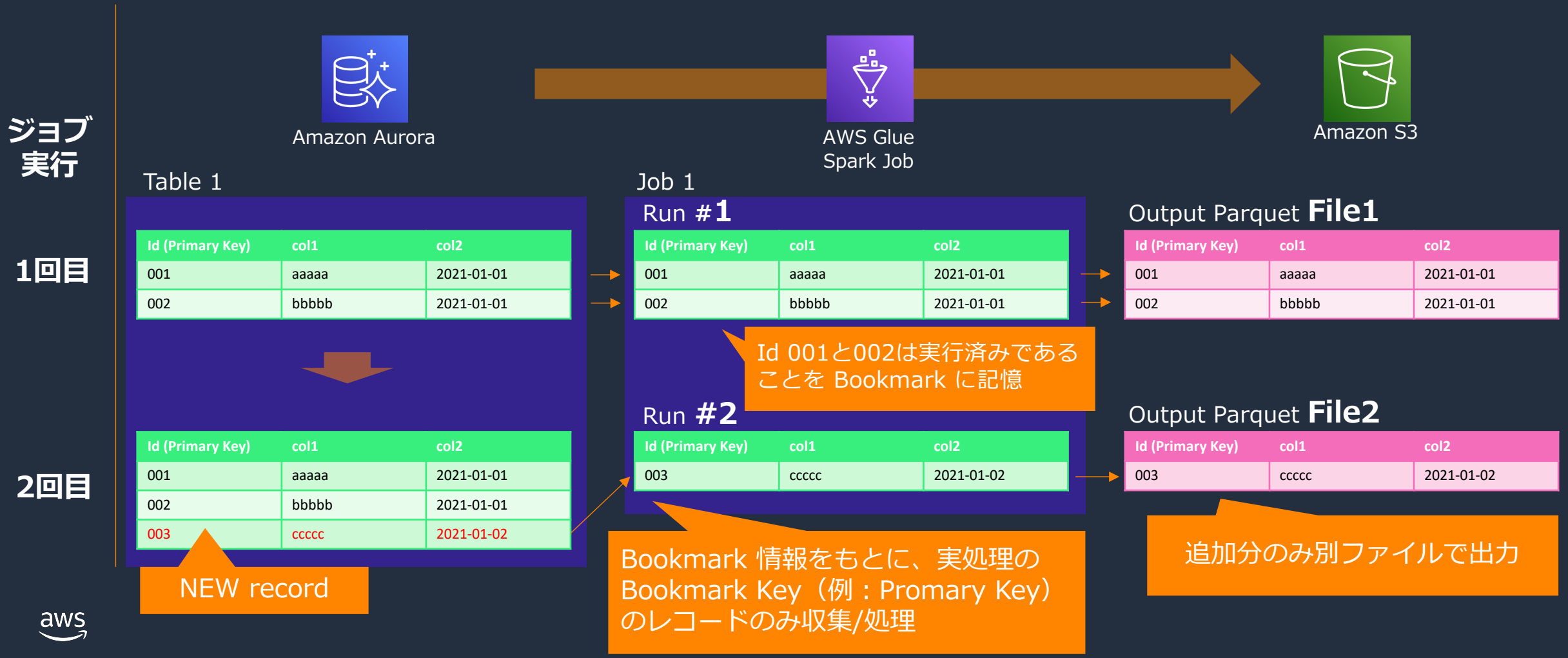


- 各種データソースからの抽出（ Extract ）において、Glue ジョブが処理済みのデータを追跡/記憶し、同じジョブを実行した場合でも処理済みデータをスキップした差分抽出が可能
- ジョブ単位で「有効」「無効」「一時停止」設定可能
- データソース
  - ✓ Amazon S3（オブジェクトファイル）
    - ※JSON/CSV/Avro/XML、Glue version 1.0以降はさらに Parquet/ORC
    - ※オブジェクトの最終更新日時で追跡/判断。
    - ※注意：取得済みファイルが更新された場合も追加再取得対象
  - ✓ JDBC 接続データソース
- Python Shell ジョブ非対応
- ブックマーク情報はジョブ単位で「リセット」「巻き戻し」可能
  - ✓ 巻き戻しの場合は、ジョブ実行IDを指定して任意のポイントへ戻る

データを段階的に読み込み、AWS Glue で最適化された Parquet ライター
https://aws.amazon.com/jp/blogs/news/load-data-incrementally-and-optimized-parquet-writer-with-aws-glue/

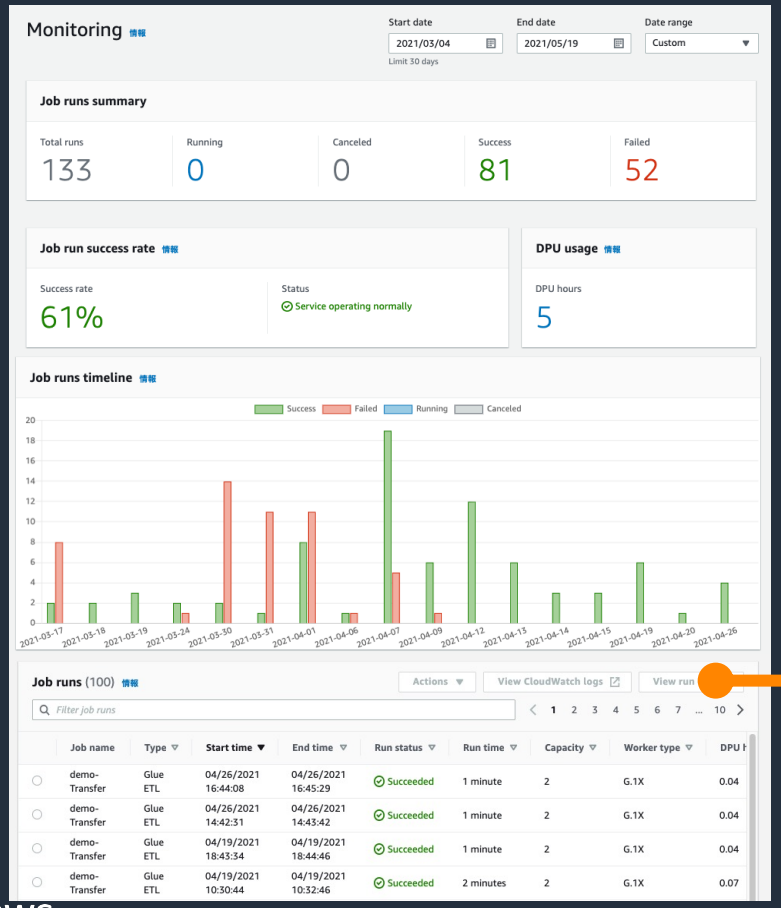
ブックマーク機能

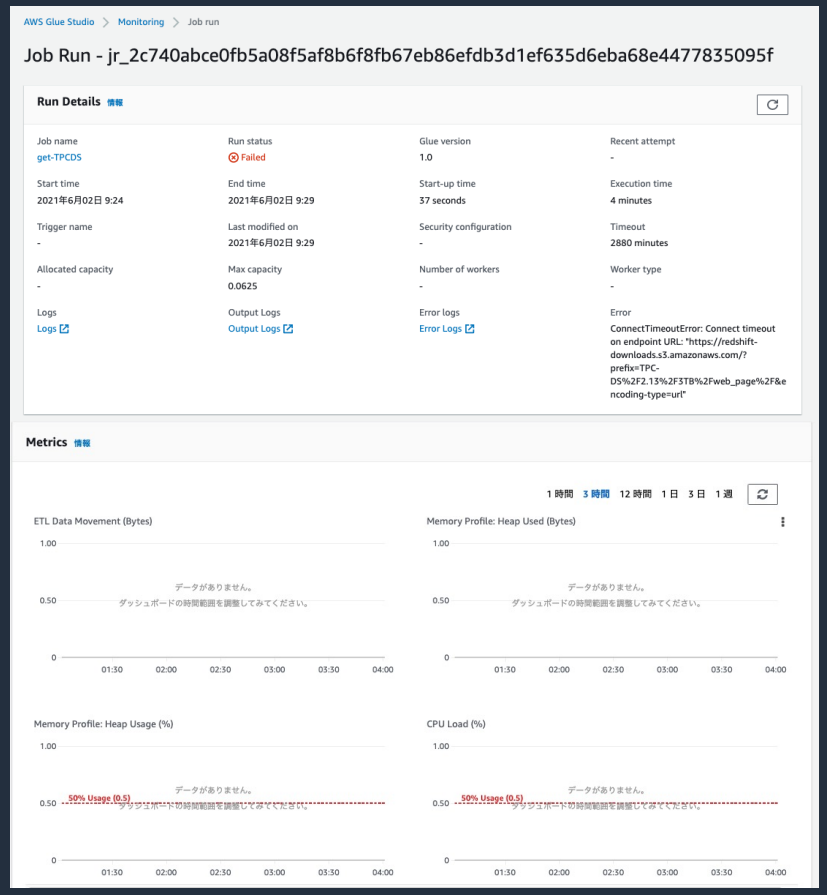
✓ ブックマークを有効化した Glue ジョブで JDBC ソースの Table1 に対する収集と S3 格納の例
✓ 2回目のタイミングでは、ソーステーブルにレコードが追加された差分だけ処理できる



ジョブの実行状況モニタリング

- 実行中/キャンセル済み/成功/失敗のステータスごとにジョブの合計、ジョブの成功率、推定 DPU 使用量、ジョブタイプ、ワーカータイプ、日毎のジョブステータスカウントなどを確認





Amazon CloudWatch

詳細ログへ

Spark History Server

- Add Job の画面で Spark UI を有効化したときに指定したS3パスにログが出力される
- 実行済みのアプリケーションのログはSpark History Server を起動することで可視化できる
- Spark History Server の起動方法は以下などがある。
  - ✓ AWS CloudFormation を利用する
  - ✓ docker でローカル PC や EC2 上で起動する

AWS CloudFormation を使用した Spark 履歴サーバーの起動と Spark UI の表示

AWS CloudFormation テンプレートを使用して Apache Spark 履歴サーバーを起動し、Spark ウェブ UI を表示できます。これらのテンプレートは、要件に応じて変更する必要があるサンプルです。

AWS CloudFormation を使用して Spark 履歴サーバーを起動し、Spark UI を表示するには

1. 次の表に示す**起動スタック**ボタンの 1 つを選択します。これにより、AWS CloudFormation コンソールでスタックが起動されます。

リージョン	Glue 4.0 の起動	Glue 3.0 の起動	Glue 1.0/2.0の起動
米国東部 (オハイオ)	Launch Stack	Launch Stack	Launch Stack
米国東部 (バージニア北部)	Launch Stack	Launch Stack	Launch Stack
米国西部 (北カリフォルニア)	Launch Stack	Launch Stack	Launch Stack
米国西部 (オレゴン)	Launch Stack	Launch Stack	Launch Stack
アフリカ (ケープタウン)	Launch Stack	Launch Stack	最初に、リージョンへのコンソールアクセスを有効にする必要があります。 Launch Stack
アジアパシフィック (香港)	Launch Stack	Launch Stack	最初に、リージョンへのコンソールアクセスを有効にする必要があります。 Launch Stack

History Server

Event log directory: s3a://glueworkshop-chie/sparkeventlogs

There are 990 event log(s) currently being processed which may result in additional applications getting listed on this page. Refresh the page to view updates.

Client local time zone: Asia/Tokyo

Search:

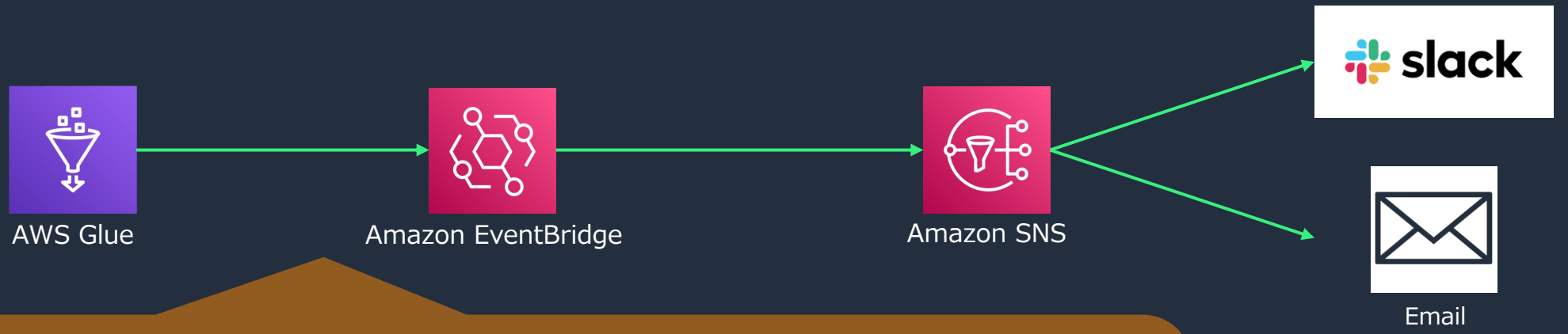
App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
application_1622857946319_0011	livy-session-10	2021-06-15 12:05:21	2021-06-15 13:05:49	1.0 h	livy	2021-06-15 13:06:11	Download
application_1622857946319_0010	livy-session-9	2021-06-09 21:28:16	2021-06-09 22:28:48	1.0 h	livy	2021-06-09 22:29:11	Download
application_1622857946319_0009	livy-session-8	2021-06-09 21:19:27	2021-06-09 21:27:56	8.5 min	livy	2021-06-09 21:28:11	Download
application_1622857946319_0008	livy-session-7	2021-06-09 21:14:36	2021-06-09 21:18:27	3.9 min	livy	2021-06-09 21:18:41	Download
application_1622857946319_0007	livy-session-6	2021-06-09 20:47:25	2021-06-09 21:14:15	27 min	livy	2021-06-09 21:14:41	Download
application_1622857946319_0006	livy-session-5	2021-06-09 20:31:32	2021-06-09 20:47:05	16 min	livy	2021-06-09 20:47:11	Download
application_1622857946319_0005	livy-session-4	2021-06-09 20:04:45	2021-06-09 20:20:06	15 min	livy	2021-06-09 20:20:11	Download
application_1622857946319_0004	livy-session-3	2021-06-09 20:02:11	2021-06-09 20:03:10	59 s	livy	2021-06-09 20:03:11	Download
application_1622857946319_0003	livy-session-2	2021-06-09 15:46:56	2021-06-09 15:49:01	2.1 min	livy	2021-06-09 15:49:11	Download
application_1622857946319_0002	livy-session-1	2021-06-09 13:58:35	2021-06-09 15:46:36	1.8 h	livy	2021-06-09 15:46:41	Download

Showing 1 to 10 of 10 entries

Show incomplete applications

ジョブステータスの通知

- EventBridge を利用して Glue のイベントをトリガーとして通知や他のサービスとの連携が可能

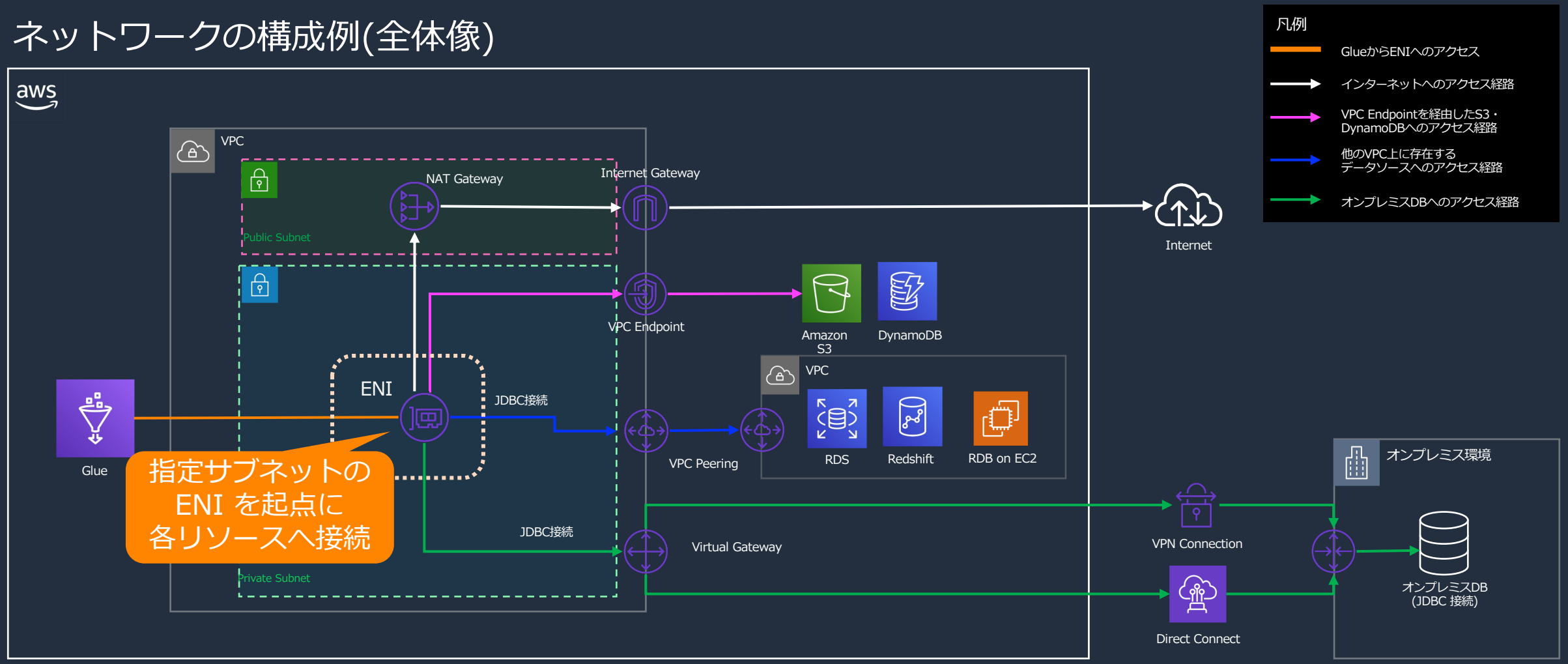


- カタログの状態変化（スキーマ変更など）
- ジョブステータスの変化（ジョブ終了、失敗など）
- ジョブ実行時間閾値超過
- Glue クローラのステータスの変化
- etc

セキュリティとネットワーク

ジョブ・クローラーによる VPC へのアクセス

ネットワークの構成例(全体像)



Glue ジョブやクローラーが接続可能なネットワーク、VPC 内の ENI から到達可能な範囲に限定できるため、閉域網に閉じた構成にも対応可能です

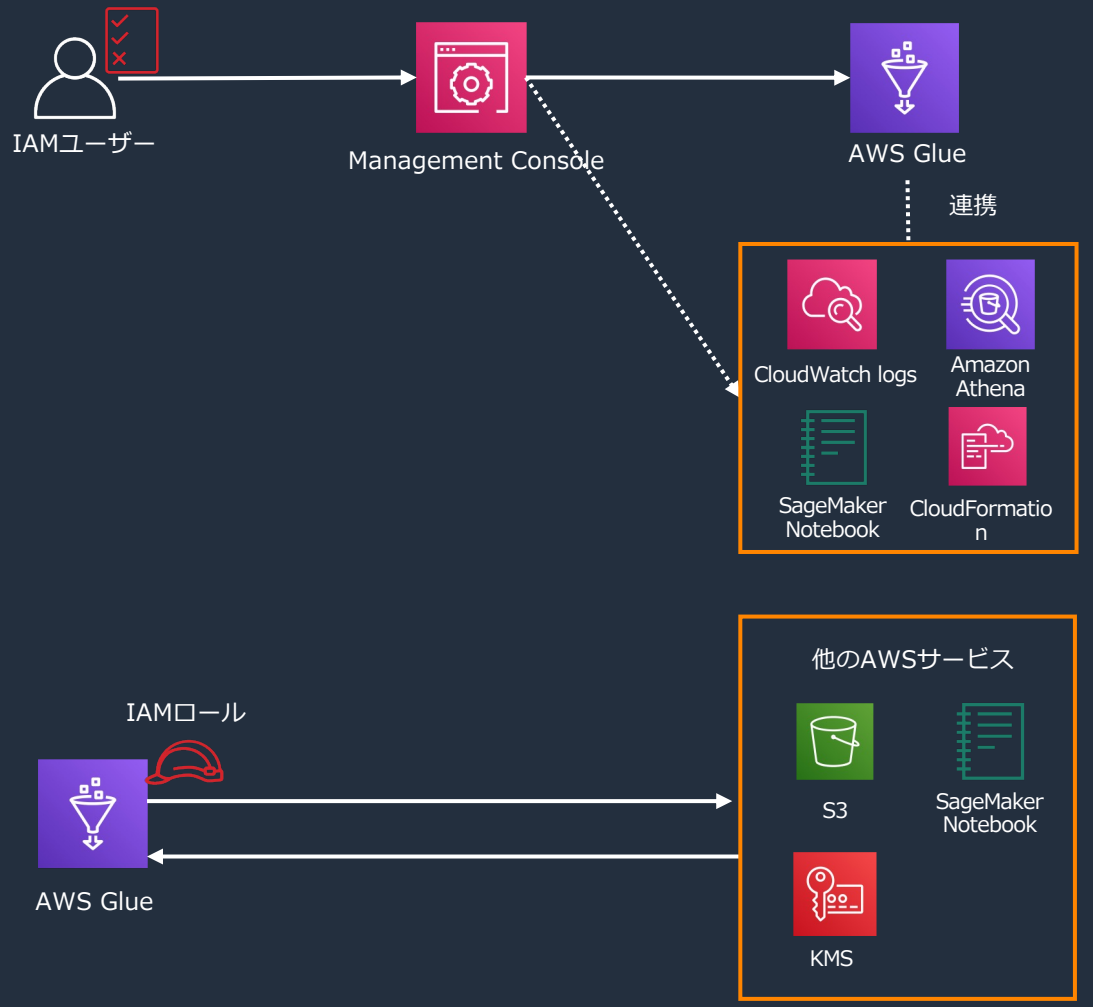
IAM による権限管理

IAM ユーザー

- AWS マネジメントコンソール上で Glue を利用する際に必要な権限
(他サービスはバックグラウンドで使用)

IAM ロール

- ジョブ実行に必要な権限を与えた IAM ロールを付与する
- ジョブ毎にロールを分けることも可能
- S3 上のファイルが暗号化されている場合は復号可能なロールも必要



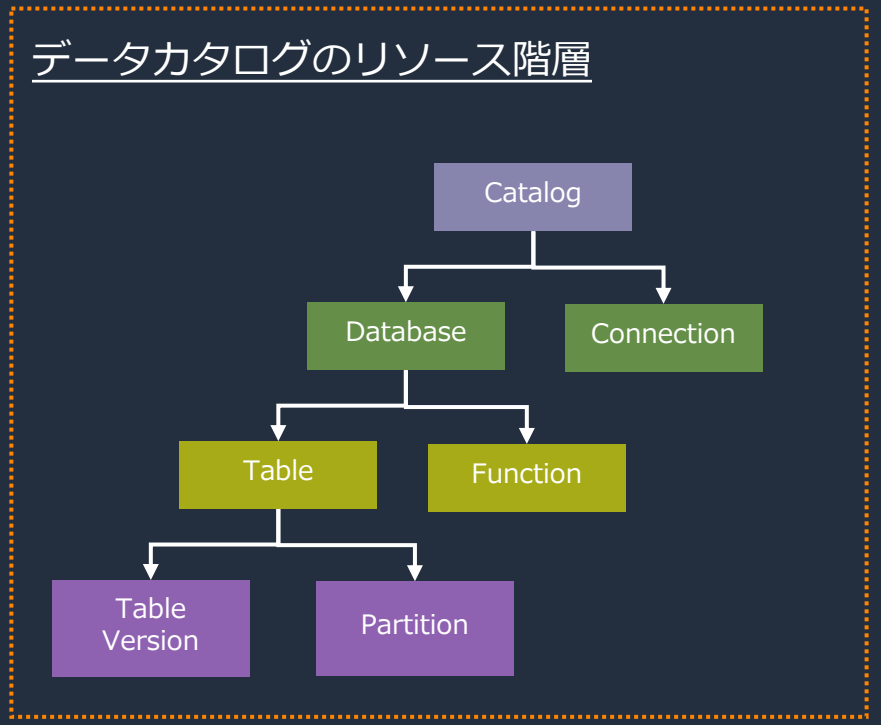
リソースレベルによるポリシーとアクセス制御

AWS Glue 内で管理するアクセスポリシー

- データカタログリソースへのアクセス制御を行う
- データカタログに対するクロスアカウント・クロスリージョンのアクセス制御が可能

リソースポリシーの例

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "glue:CreateTable"
      ],
      "Principal": {"AWS": [
        "arn:aws:iam:: account-A-id:user/dev",
        "arn:aws:iam:: account-B-id:root"
      ]},
      "Resource": [
        "arn:aws:glue:us-east-1:account-A-id:table/db1/*",
        "arn:aws:glue:us-east-1:account-A-id:database/db1",
        "arn:aws:glue:us-east-1:account-A-id:catalog"
      ]
    }
  ]
}
```



※アカウントAのデータベース db1 に任意のテーブルを作成可能なアクセス権。アカウントA の dev ユーザーとアカウントB のルートユーザーに同じアクセスを許可する。
参考URL： リソースベースのアクセス許可(https://docs.aws.amazon.com/ja_jp/glue/latest/dg/using-identity-based-policies.html)

暗号化設定

保管中の暗号化

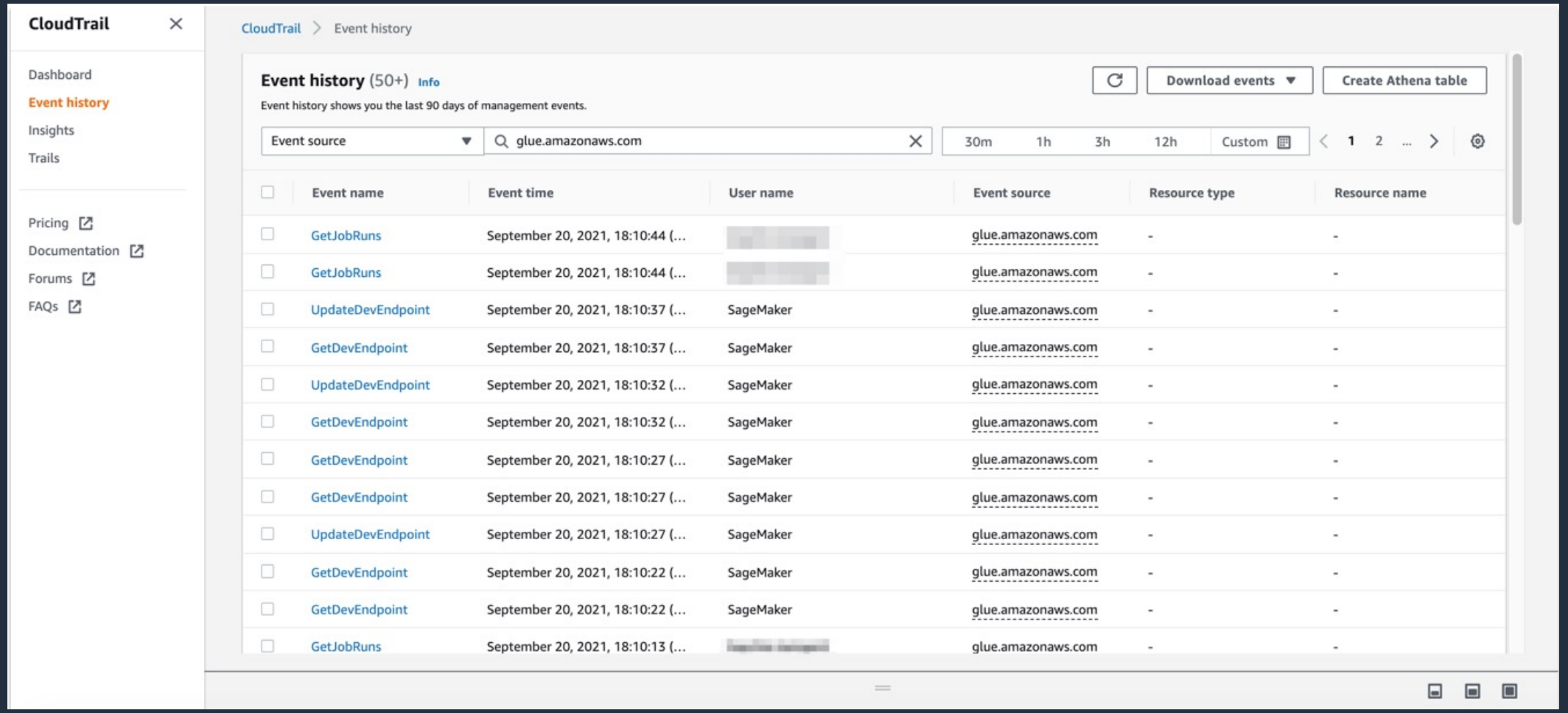
- AWS KMS と統合
- ETLジョブと開発エンドポイントの書き込みデータを暗号化
- AWS Glue Data Catalog に格納されているメタデータと接続パスワード、ジョブのブックマーク、クローラーおよびETL ジョブのログも暗号化可能

転送中の暗号化

- Transport Layer Security (TLS) 暗号化

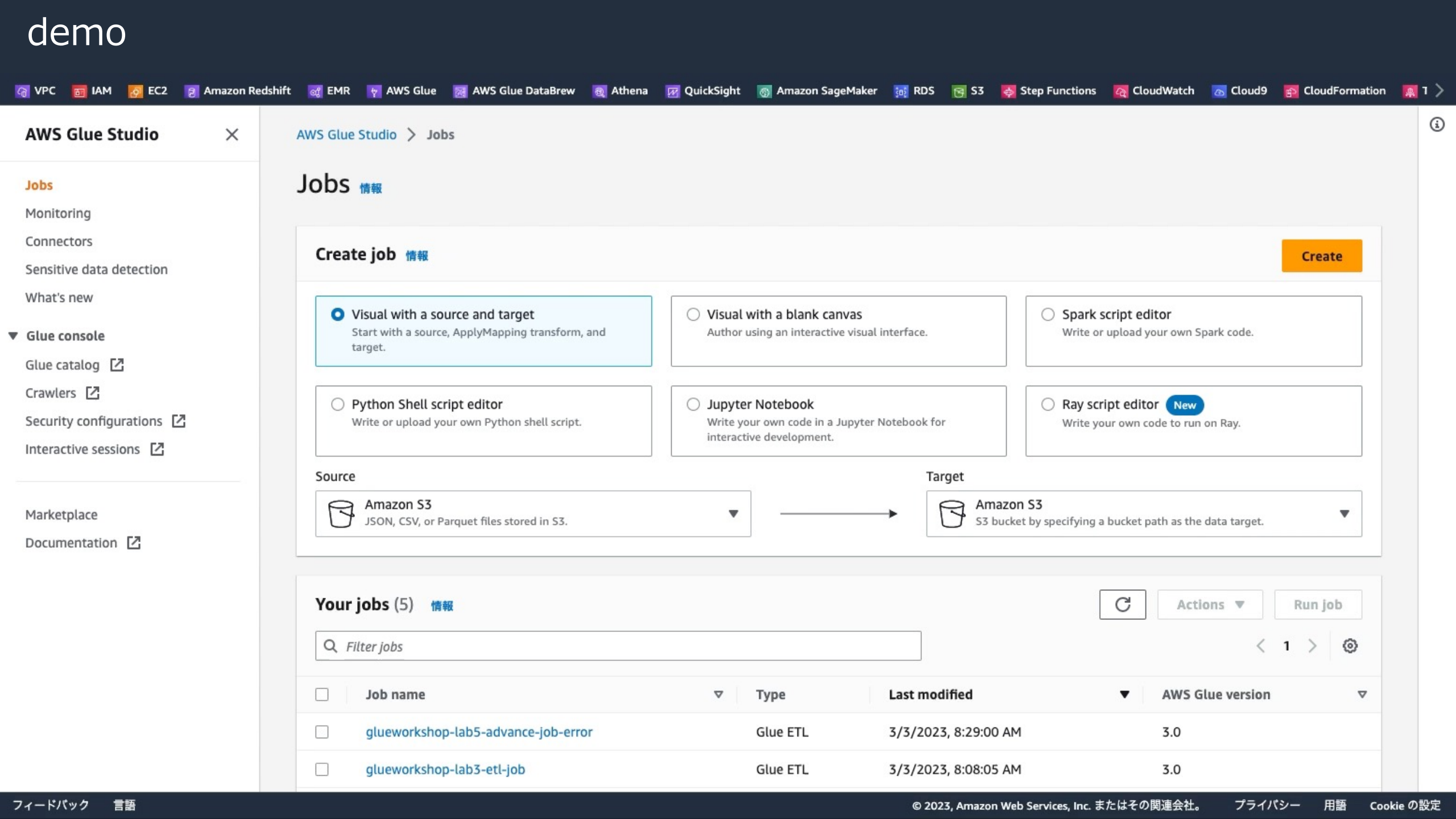
監査

- AWS Glue API の実行履歴を CloudTrail で追跡可能



demo

demo


VPC IAM EC2 Amazon Redshift EMR AWS Glue AWS Glue DataBrew Athena QuickSight Amazon SageMaker RDS S3 Step Functions CloudWatch Cloud9 CloudFormation
AWS Glue Studio
Jobs
Monitoring
Connectors
Sensitive data detection
What's new
Glue console
Glue catalog
Crawlers
Security configurations
Interactive sessions
Marketplace
Documentation
AWS Glue Studio > Jobs
Jobs 情報
Create job 情報
Create
Visual with a source and target
Start with a source, ApplyMapping transform, and target.
Visual with a blank canvas
Author using an interactive visual interface.
Spark script editor
Write or upload your own Spark code.
Python Shell script editor
Write or upload your own Python shell script.
Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.
Ray script editor New
Write your own code to run on Ray.
Source
Amazon S3
JSON, CSV, or Parquet files stored in S3.
Target
Amazon S3
S3 bucket by specifying a bucket path as the data target.
Your jobs (5) 情報
Actions
Run job
Filter jobs
1
Job name
Type
Last modified
AWS Glue version
glueworkshop-lab5-advance-job-error
Glue ETL
3/3/2023, 8:29:00 AM
3.0
glueworkshop-lab3-etl-job
Glue ETL
3/3/2023, 8:08:05 AM
3.0
フィードバック
言語
© 2023, Amazon Web Services, Inc. またはその関連会社。
プライバシー
用語
Cookie の設定

費用

AWS Glue の料金 https://aws.amazon.com/jp/glue/pricing/

1DPU (Data Processing Unit) = 4 vCPU、16 GB

ETL ジョブ/開発エンドポイント

Apache Spark または Spark ストリーミングジョブ

- DPU 時間あたり 0.44 USD が1 秒単位で課金され、最小 1 分 (Glue バージョン 2.0 以降) または最小 10 分 (Glue バージョン 0.9/1.0)

柔軟な実行を伴う Apache Spark（Glue Flex）

- DPU 時間あたり 0.29 USD が1 秒単位で課金され、最小 1 分 (Glue バージョン 3.0 以降)

Python Shell ジョブ

- DPU 時間あたり 0.44 USD が 1 秒単位で課金され、最小 1 分

インタラクティブセッション

- DPU 時間あたり 0.44 USD が 1 秒単位で課金され、最小 1 分です。AWS Glue Studio Job Notebooks はインタラクティブセッションに組み込まれたインターフェイスであり、追加コストなしで提供されます。

AWS Glue Studio データプレビューセッション

- DPU 時間あたり 0.44 USD が、30 分単位で開発エンドポイントとして課金されます

プロビジョニングされた開発エンドポイント

- DPU 時間あたり 0.44 USD が 1 秒単位で課金され、最小 10 分

AWS Glue の料金 https://aws.amazon.com/jp/glue/pricing/

データカタログ

ストレージ：100 万オブジェクトまで無料
(オブジェクト=テーブル、テーブルバージョン、パーティション、データベース)
100万以上保存された場合、10万オブジェクトあたり $1/月

リクエスト：100 万リクエスト/月まで無料
100万以上保存された場合、10万オブジェクトあたり $1/月

クローラー

$0.44 DPU / 時 (課金は秒単位)
- クローラーの実行に使用された DPU の数に応じて時間あたりの課金が発生
- クローラーごとに 10 分の最小期間が設定される。10 分未満の処理は10分として計算される

その他

データ転送量や Amazon S3 の保存データ

導入事例の参考情報

導入事例

AWS Glue を活用した多くの国内事例がございます

- AWSで実践！Analytics Modernization ～事例祭り編～
- https://aws.amazon.com/jp/blogs/news/analytics-modernization-jireimatsuri-20220324/

- AWSで実践！Analytics Modernization ～事例祭り編 第二回～
- https://aws.amazon.com/jp/blogs/news/analytics-modernization-jireimatsuri-20220623/

- AWSで実践！Analytics Modernization ～事例祭り編 第三回～
- https://aws.amazon.com/jp/blogs/news/analytics_jireimatsuri_20220922/

まとめ

データ収集、カタログ化、データ加工を一気通貫して行うことができる統合的なデータインテグレーションサービス

AWS Glue

データカタログ

データクローラー

他システム連携の豊富なコネクタ

サーバーレス ETL エンジン

豊富なビルトイン変換とローコード開発

スケジューラとワークフロー管理

参考資料

AWS Glue 開発者ガイド
https://docs.aws.amazon.com/ja_jp/glue/latest/dg/what-is-glue.html

[AWS Black Belt Online Seminar] AWS Glue DataBrew 資料及び QA 公開
https://aws.amazon.com/jp/blogs/news/webinar-bb-awsgluedatabrew-2021/

[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
https://aws.amazon.com/jp/blogs/news/webinar-bb-awsglue-2021/

[AWS Black Belt Online Seminar] 猫でもわかる、AWS Glue ETLパフォーマンス・チューニング
https://aws.amazon.com/jp/blogs/news/webinar-bb-glueetlperformancetuning-2021/

AWS Glue Immersion day（ハンズオンコンテンツ）
https://catalog.us-east-1.prod.workshops.aws/workshops/ee59d21b-4cb8-4b3d-a629-24537cf37bb5/ja-JP

AWS Glue ETL Code Samples
https://github.com/aws-samples/aws-glue-samples

本資料に関するお問い合わせ・ご感想

技術的な内容に関しましては、有料のAWSサポート窓口へ
お問い合わせください

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

料金面でのお問い合わせに関しましては、カスタマーサポート窓口へ
お問い合わせください（マネジメントコンソールへのログインが必要です）

https://console.aws.amazon.com/support/home#/case/create?issueType=customer-service

具体的な案件に対する構成相談は、後述する個別相談会をご活用ください

ご感想はTwitterへ！ハッシュタグは以下をご利用ください
#awsblackbelt

その他コンテンツのご紹介

ウェビナーなど、AWSのイベントスケジュールをご参照いただけます

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

ハンズオンコンテンツ

https://aws.amazon.com/jp/aws-jp-introduction/aws-jp-webinar-hands-on/

AWS 個別相談会

AWSのソリューションアーキテクトと直接会話いただけます

https://pages.awscloud.com/JAPAN-event-SP-Weekly-Sales-Consulting-Seminar-2021-reg-event.html

Thank you!