



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=PLzWGOASvSx6FIwIC2X1nObr1KcMCBBBlqY>

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おります。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



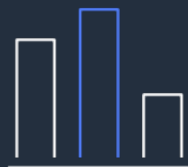
様々な
アプリケーション
による分析

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



レスポンス
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

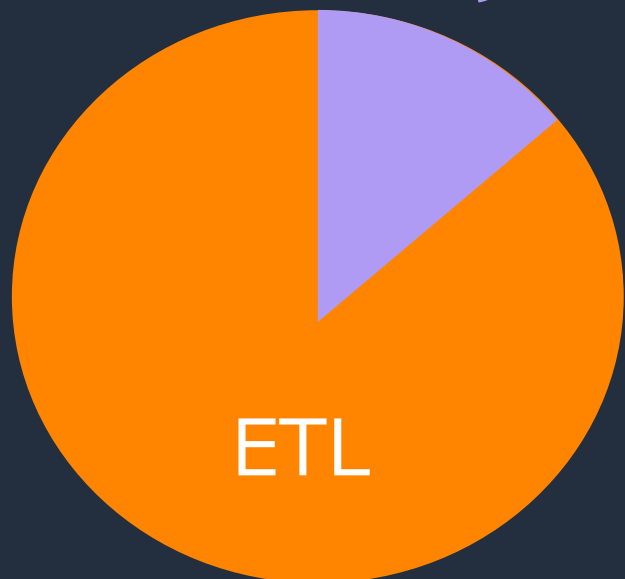


スキルセットの 多様化

データを操作する利用者の技術スキルは様々

データ活用によくある課題

データ活用



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

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



保存

エクサバイト級のデータ量まで



変換と
カタログ化

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



分析と
可視化

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



予測

最も包括的な ML と AI サービスのセット



共有

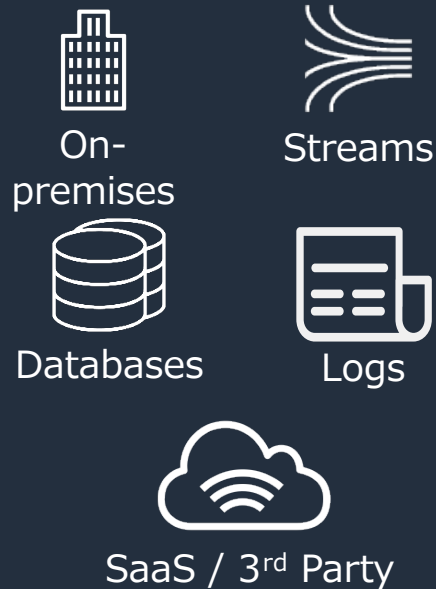
わずか数行のコードで

統合セキュリティ、ガバナンス、データアクセス制御

代表的なデータアーキテクチャ

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。



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



保存

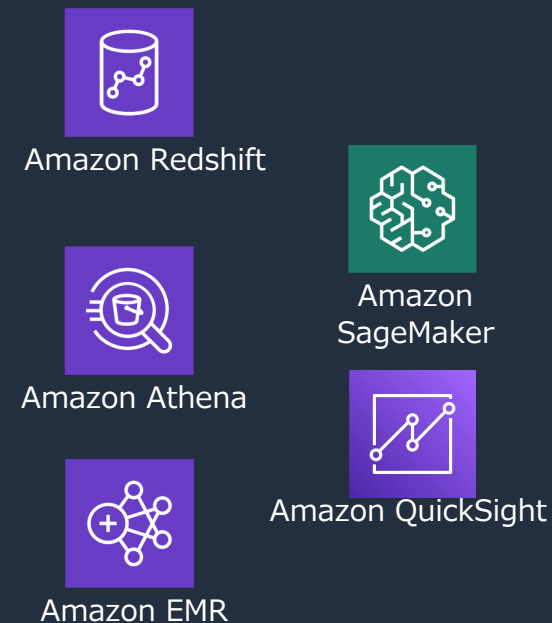
トランザクションデータをデータベースに、分析データをデータウェアハウスとデータレイクにあらゆる規模で保存します。

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

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



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

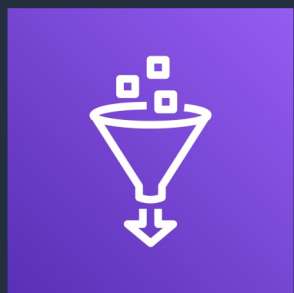


生産性向上



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



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

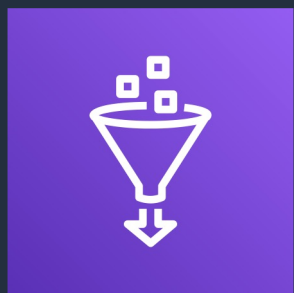


他システム連携の
豊富なコネクタ



スケジューラと
ワークフロー管理

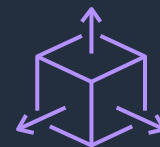
データ収集、カタログ化、データ加工を一気通貫して 行うことができる統合的なデータインテグレーションサービス



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

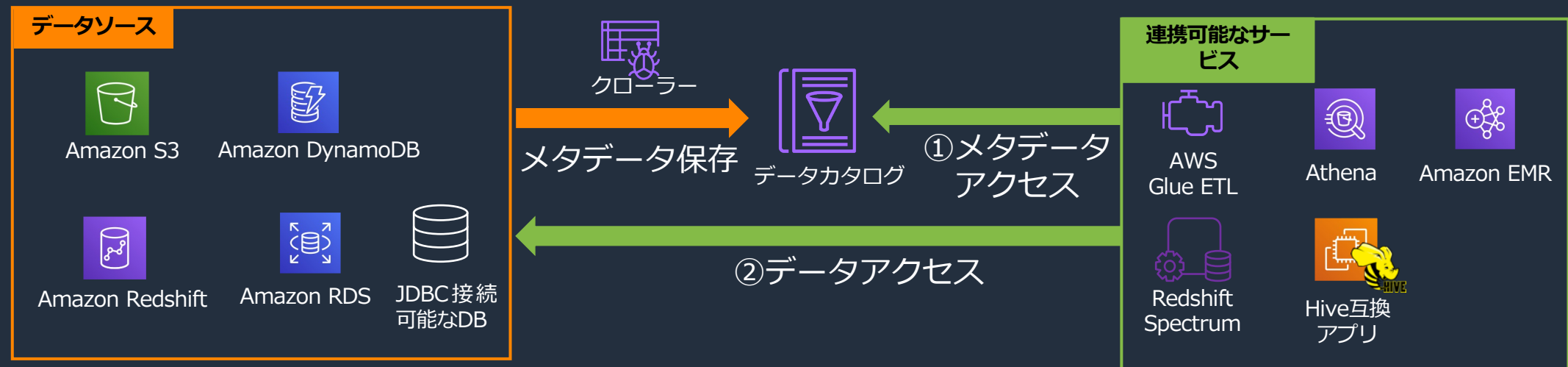


スケジューラと
ワークフロー管理

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

The screenshot displays the AWS Glue console interface. On the left is a navigation sidebar with categories like Data Catalog and Data Integration and ETL. The main content area is divided into two sections: 'Table details' and 'Schema'. The 'Table details' section shows properties for a table named 'bank_additionalssampleddata'. The 'Schema' section shows a list of columns with their names and data types.

Table details

Property	Value
Name	bank_additionalssampledata
Description	サンプルデータ
Database	sampledata
Classification	csv
Location	s3://[REDACTED]
Connection	-
Deprecated	-
Last updated	October 13, 2022 at 07:36:30
Input format	org.apache.hadoop.mapred.TextInputFormat
Output format	org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

Schema (21)

#	Column name	Data type	Partition key	Comment
1	age	bigint	-	年齢
2	job	string	-	職業
3	marital	string	-	-
4	education	string	-	-

テーブル情報

スキーマ情報

AWS Glue データカタログ

文字列検索

Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Last updated: October 17, 2022 at 07:39:52 (UTC)

Tables (10)
View and manage all available tables.

Classification = csv

<input type="checkbox"/>	Name	Database	Location	Classification	Deprecated	View data
<input type="checkbox"/>	assign_csv	qs_handson	s3://[redacted]	csv	-	Table data
<input type="checkbox"/>	bank_additionalssampl	sampledata	s3://[redacted]	csv	-	Table data
<input type="checkbox"/>	patient_info_csv	qs_handson	s3://[redacted]	csv	-	Table data

テーブルや項目へのコメント登録

Table details

Advanced properties

Name bank_additionalssampled	Description サンプルデータ	Database sampledata	Classification csv
Location s3://suhroyu-demo/sampledata/	Connection -	Deprecated -	Last updated October 13, 2022 at 07:36:30
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.q.Lio.HiveIgnoreKeyTextOutputFormat	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema (21)

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	age	bigint	-	年齢
2	job	string	-	職業
3	marital	string	-	-
4	education	string	-	-

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	

BigInt タイプ

列のタイプ

- bigint
- interval
- smallint
- string
- timestamp

更新

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

最終更新日 27 6月 2019 テーブル バージョン (実行バージョン)

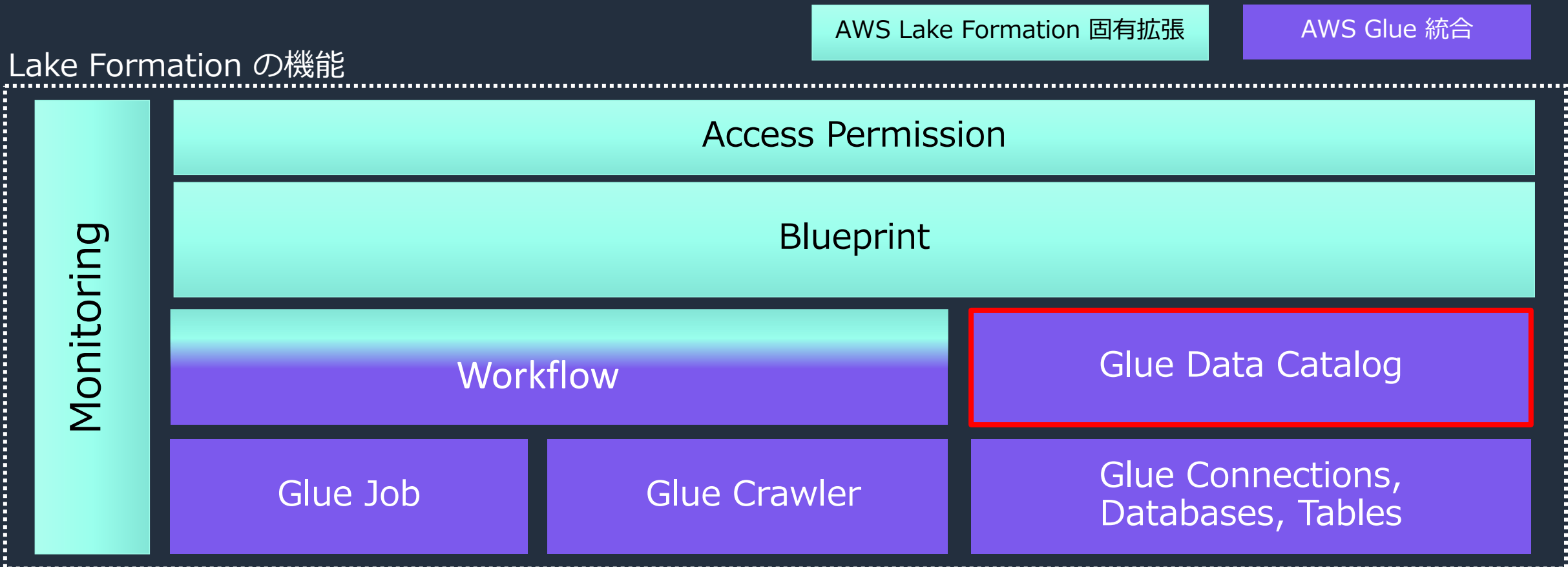
表示中: 1 - 2 < >

バージョン	作成日:	作成者:
1	27 6月 2019 10:1...	
0	27 6月 2019 9:57...	

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



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

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

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

ETL

Jobs

Triggers

Dev endpoints

Crawlers

A crawler connects to a data store, progresses through a prioritized list of objects, and updates the data catalog with the results.

mysql-db1 connected successfully to your instance.

Crawler "mysql-db1-crawler" completed and made the following changes:

[Add crawler](#) [Run crawler](#) [Action](#)

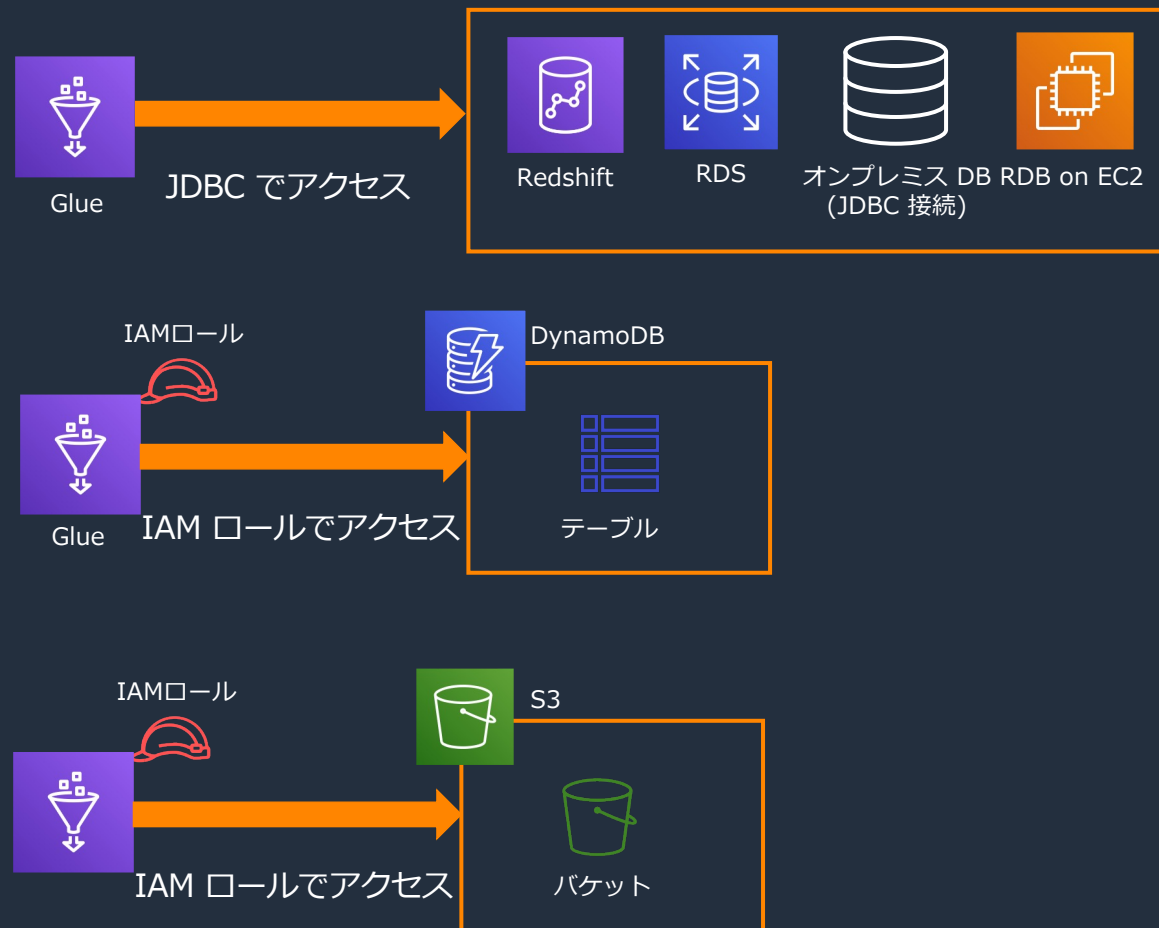
<input type="checkbox"/>	Name	Schedule	Status
<input type="checkbox"/>	mysql-db1-crawler		Ready

データソース・ターゲットへの接続管理

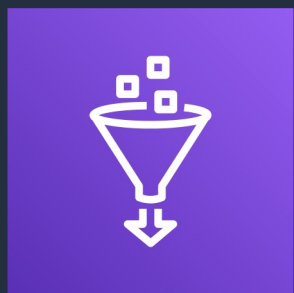
各種データソースへの接続設定を保存して管理/再利用できる

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



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 コネクタ

SaaS / 他クラウド



オンプレミス



AWS Glue Connector

AWS Services



- 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

The screenshot shows the AWS Marketplace search results for 'glue'. The left sidebar contains filters for 'Refine results' with categories like Infrastructure Software (87), Data Products (37), DevOps (7), Business Applications (6), Professional Services (5), Industries (4), Machine Learning (3), and IoT (2). Under 'Delivery methods', there are checkboxes for Container (64), Data Exchange (37), SaaS (12), Amazon Machine Image (9), Professional Services (5), and CloudFormation Template (3). Under 'Publisher', there are checkboxes for CData Software (51), Rearc (19), Investics Data Services (16), Cognosys Inc. (6), Progress Software (5), AWS Glue (3), WireWheel, Inc. (2), and Etleap (2). The main content area shows 129 results for 'glue', displaying the first four items:

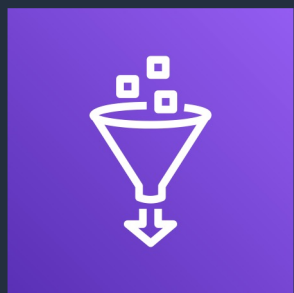
- Elasticsearch Spark connector for AWS Glue** (By AWS Glue | Ver 7.8.0, 32 external reviews): This product helps you read and write data from Elasticsearch using Apache Spark. By using this connector, you can focus on mining meaningful business insights from your data instead of writing and maintaining the connecting logic. For more details, please refer to Glue Elasticsearch connector...
- AWS Glue Connector for Apache Hudi** (By AWS Glue | Ver 0.5.3): The AWS Glue Connector for Apache Hudi simplifies the process to create and update Apache Hudi tables from AWS Glue. This connector can be used for both Copy on Write (COW) and Merge on Read (MOR) storage types.
- AWS Glue Connector for Google BigQuery** (By Amazon Web Services | Ver 2.12.0, 1 AWS review): The AWS Glue Connector for Google BigQuery simplifies the process of connecting AWS Glue jobs to extract and load data from BigQuery. The BigQuery Connector provides comprehensive access to BigQuery data, facilitating cloud ETL processes for operational reporting, backup and disaster recovery, data...
- Cloudwatch Logs connector for AWS Glue** (By AWS Glue | Ver 1): The AWS Glue Cloudwatch Logs connector enables AWS Glue to communicate with CloudWatch Logs so that you can query your log data with SQL. By using this connector, you can focus on mining meaningful business insights from your data instead of writing and maintaining the connecting logic. Please...

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

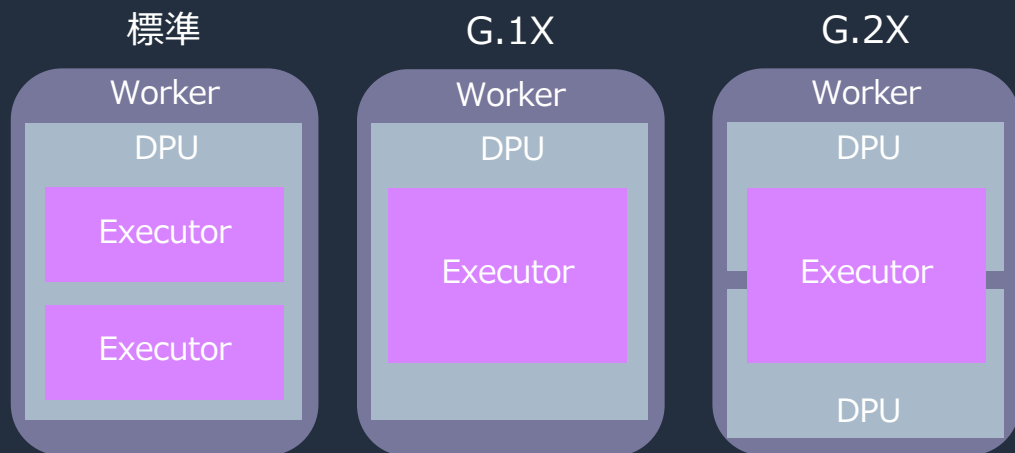
Ray ジョブ

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

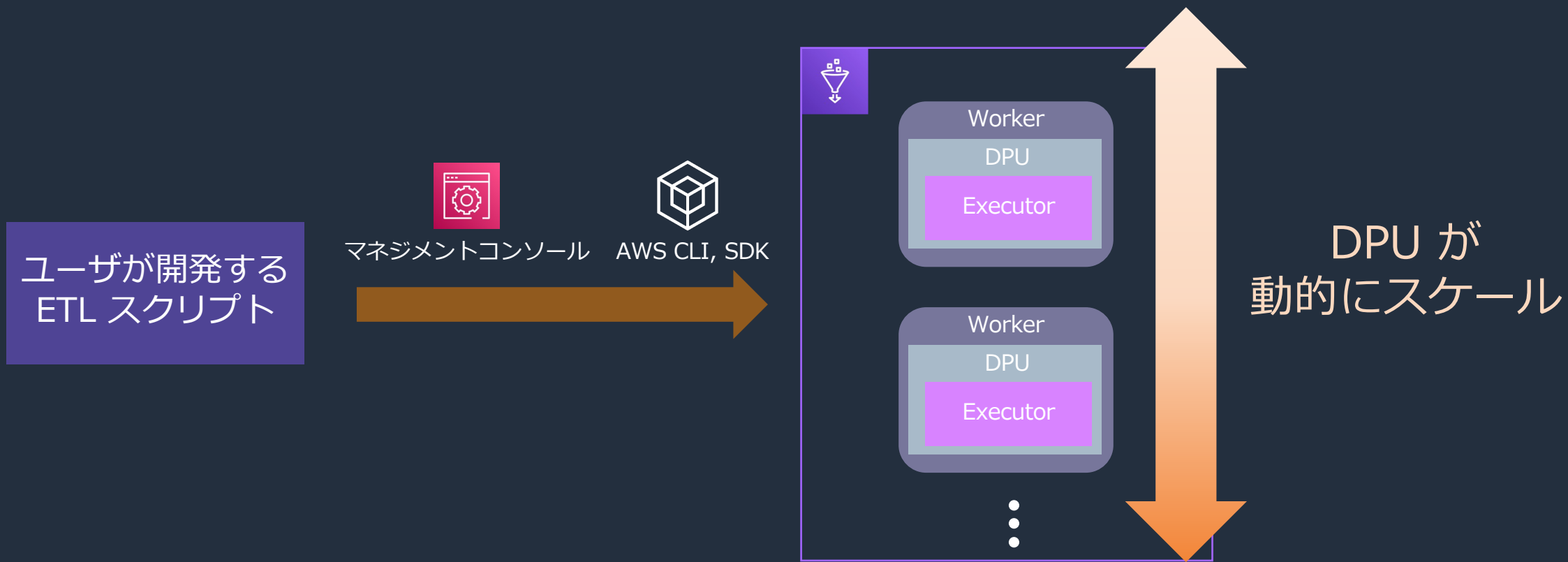
バージョン	Apache Spark のバージョン
Glue 4.0	Spark 3.3.0 (Python 3.10、Scala 2.12)
Glue 3.0	Spark 3.1.1 (Python 3.7、Scala 2.12)
Glue 2.0 ※2024年1月31日サポート終了予定	Spark 2.4.3 (Python 3.7、Scala 2.11)
Glue 1.0 ※ 2022年9月30日サポート終了	Spark 2.4.3 (Python 2.7、Python 3.6、Scala 2.11)
Glue 0.9 ※ 2022年6月1日サポート終了	Spark 2.2.1 (Python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

The screenshot shows the AWS Glue console interface for configuring a job. The top navigation bar includes the AWS logo, a 'サービス' (Services) menu, and a search bar. Below the navigation bar, there are icons for various AWS services: VPC, IAM, EC2, Amazon Redshift, EMR, AWS Glue, AWS Glue DataBrew, Athena, and QuickSight. The main content area is titled 'Untitled job' and has several tabs: 'Visual', 'Script', 'Job details' (which is selected), 'Runs', 'Schedules', and 'Version Control'. Under the 'Job details' tab, there are three main sections: 'Worker type', 'Automatically scale the number of workers', and 'Maximum number of workers'. The 'Worker type' section has a dropdown menu set to 'G 1X (4vCPU and 16GB RAM)'. The 'Automatically scale the number of workers' section has a checked checkbox and a description: 'AWS Glue will optimize costs and resource usage by dynamically scaling the number of workers up and down throughout the job run. Requires Glue 3.0 or later.' The 'Maximum number of workers' section has a text input field containing the number '10'. Two orange callout boxes are overlaid on the image: one pointing to the 'Automatically scale the number of workers' checkbox with the text '有効化' (Enable), and another pointing to the 'Maximum number of workers' input field with the text '最大ワーカー数指定' (Specify maximum number of workers).

有効化

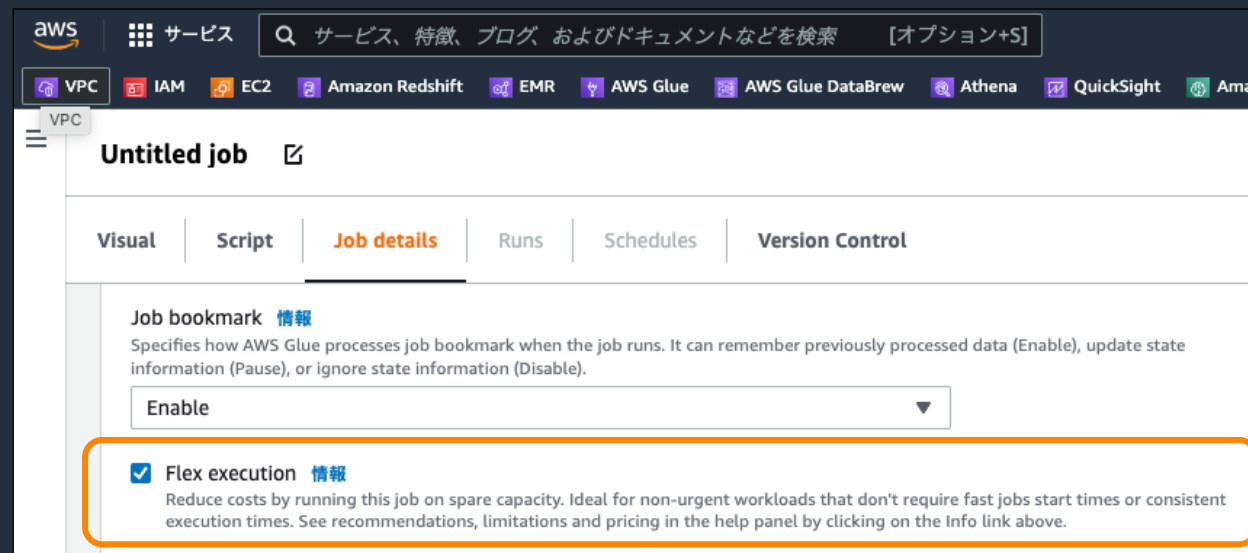
最大ワーカー数
指定

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

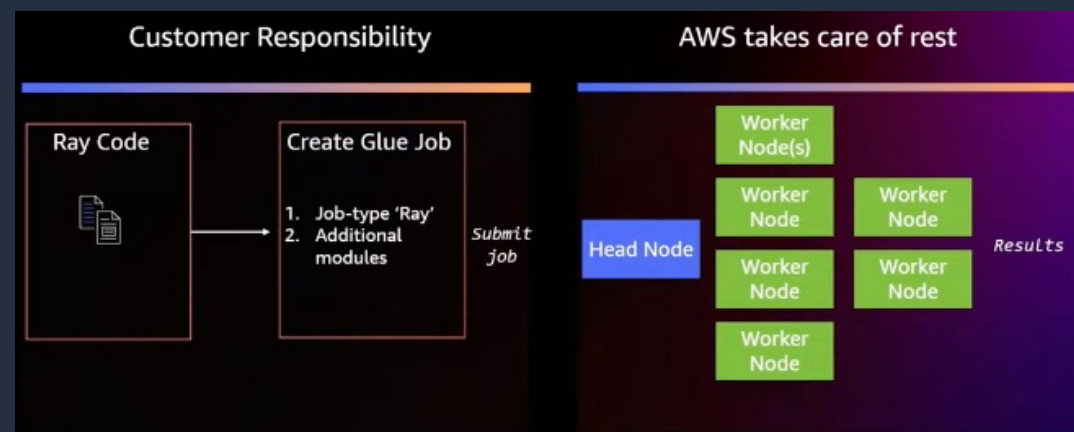
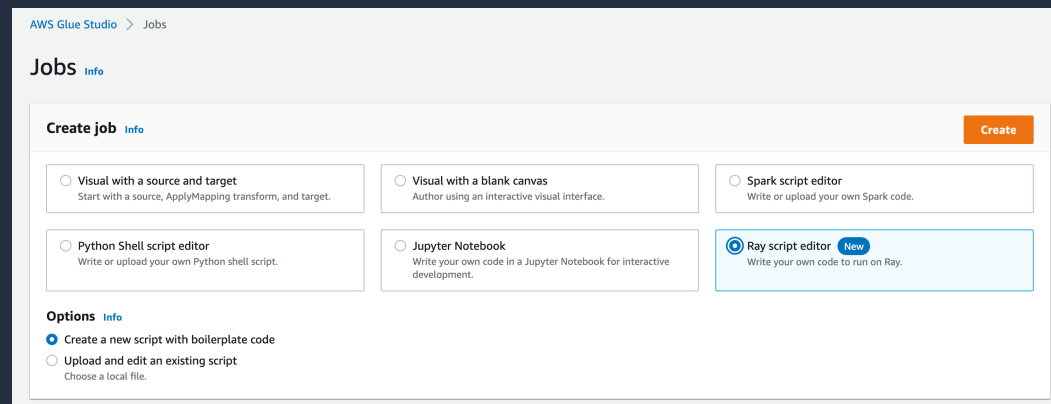
https://docs.aws.amazon.com/ja_jp/glue/latest/dg/add-job-python.html

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

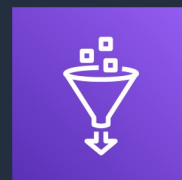
小規模処理



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

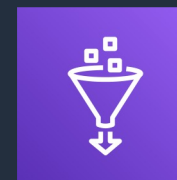
小規模処理



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

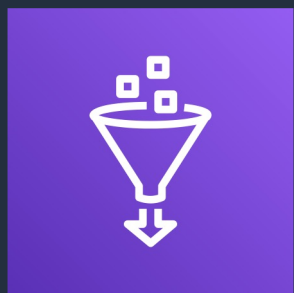


AWS Glue Apache Spark Ray

- 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

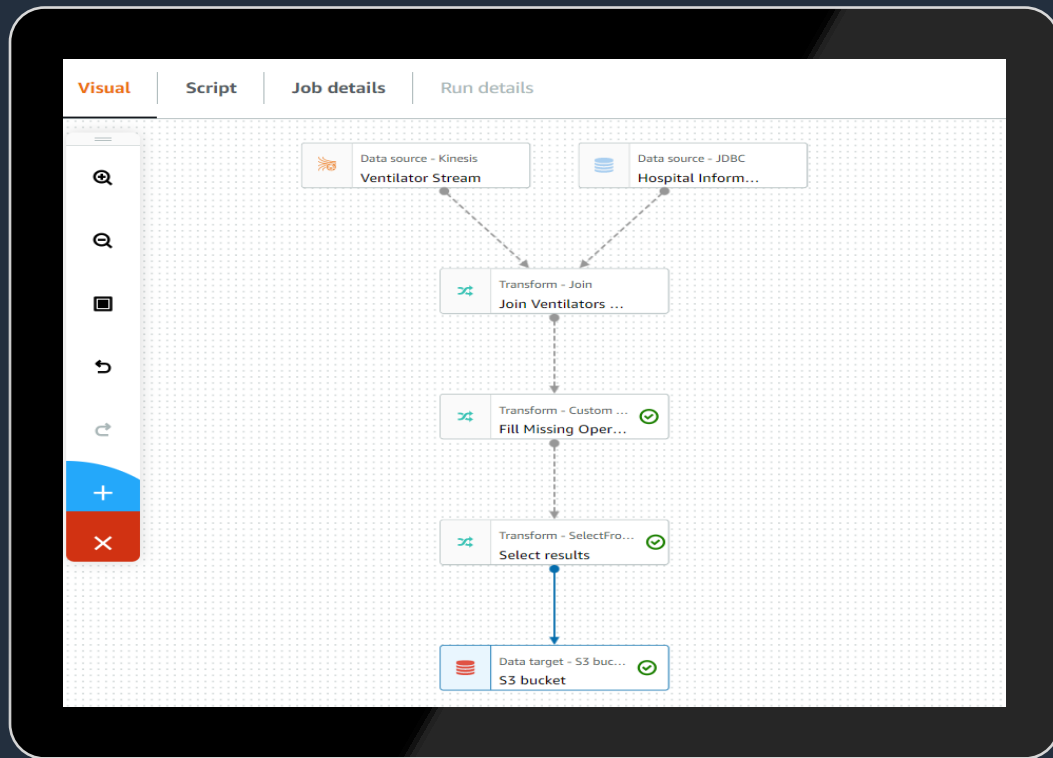


豊富なビルトイン変換と
ローコード開発



スケジューラと
ワークフロー管理

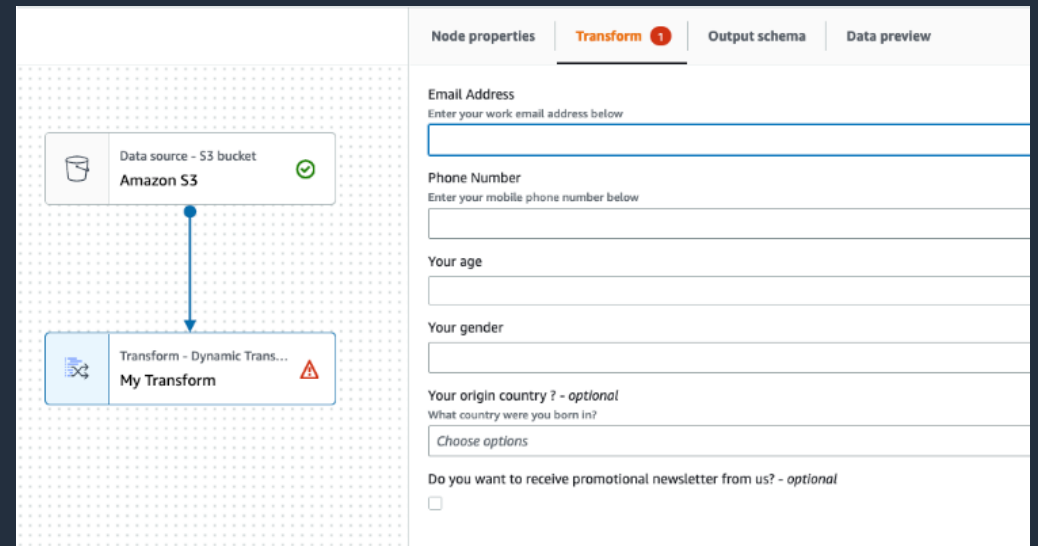
Glue Studio



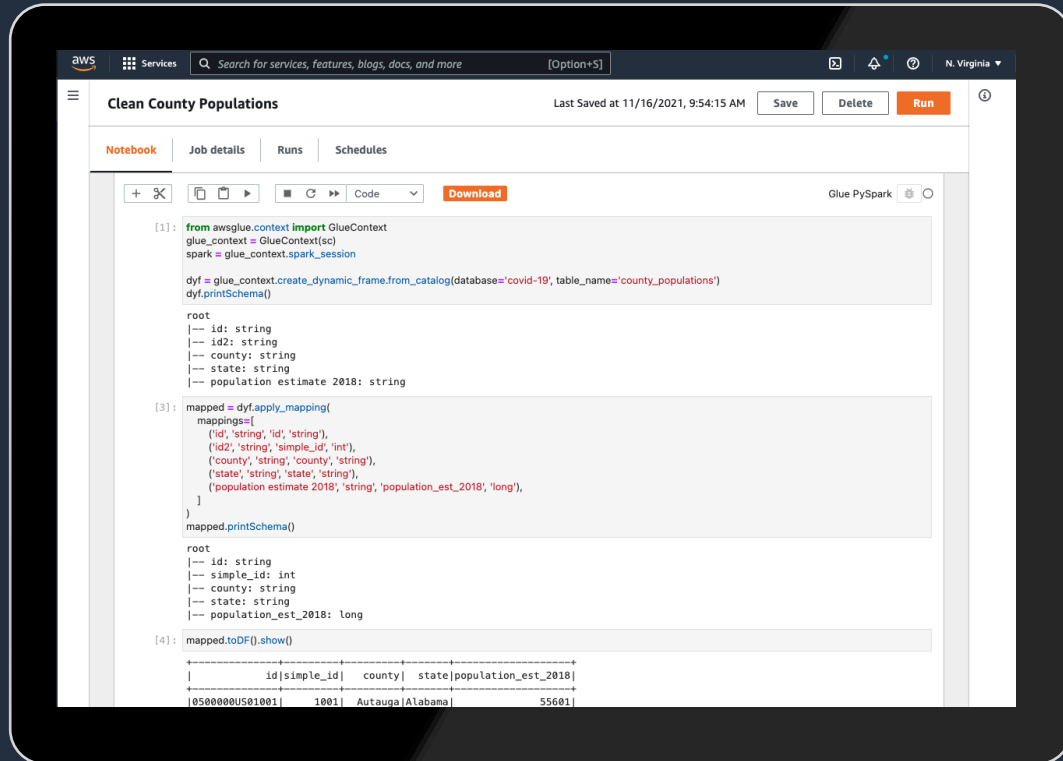
- ETL ジョブの作成、実行、およびモニタリングを簡単にできるようにする新しいグラフィカルインターフェイス
- GUI で作成したフローからソースコードを生成
- データ抽出、出力、ジョイン等の汎用的な処理はビルトインされノンコーディングで実現
- 分散処理を意識せずに開発

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



The screenshot shows the AWS Glue Studio Notebook interface. The notebook is titled "Clean County Populations" and is saved at 11/16/2021, 9:54:15 AM. The code in the notebook is as follows:

```
[1]: from aws glue.context import GlueContext
glue_context = GlueContext(sc)
spark = glue_context.spark_session

dyf = glue_context.create_dynamic_frame_from_catalog(database='covid-19', table_name='county_populations')
dyf.printSchema()

root
|-- id: string
|-- id2: string
|-- county: string
|-- state: string
|-- population estimate 2018: string

[3]: mapped = dyf.apply_mapping(
  mappings=[
    ('id', 'string', 'id', 'string'),
    ('id2', 'string', 'simple_id', 'int'),
    ('county', 'string', 'county', 'string'),
    ('state', 'string', 'state', 'string'),
    ('population estimate 2018', 'string', 'population_est_2018', 'long'),
  ]
)
mapped.printSchema()

root
|-- id: string
|-- simple_id: int
|-- county: string
|-- state: string
|-- population_est_2018: long

[4]: mapped.toDF().show()
```

id	simple_id	county	state	population_est_2018
0500000US01001	1001	Autauga	Alabama	55601

- **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  | <a href="#">Link</a> | <a href="#">Link</a> | ✓                |



SparkSession available as 'spark'.

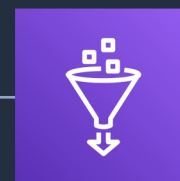
<pyspark.sql.session.SparkSession object at 0x7f3d54913550>
```

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



Client

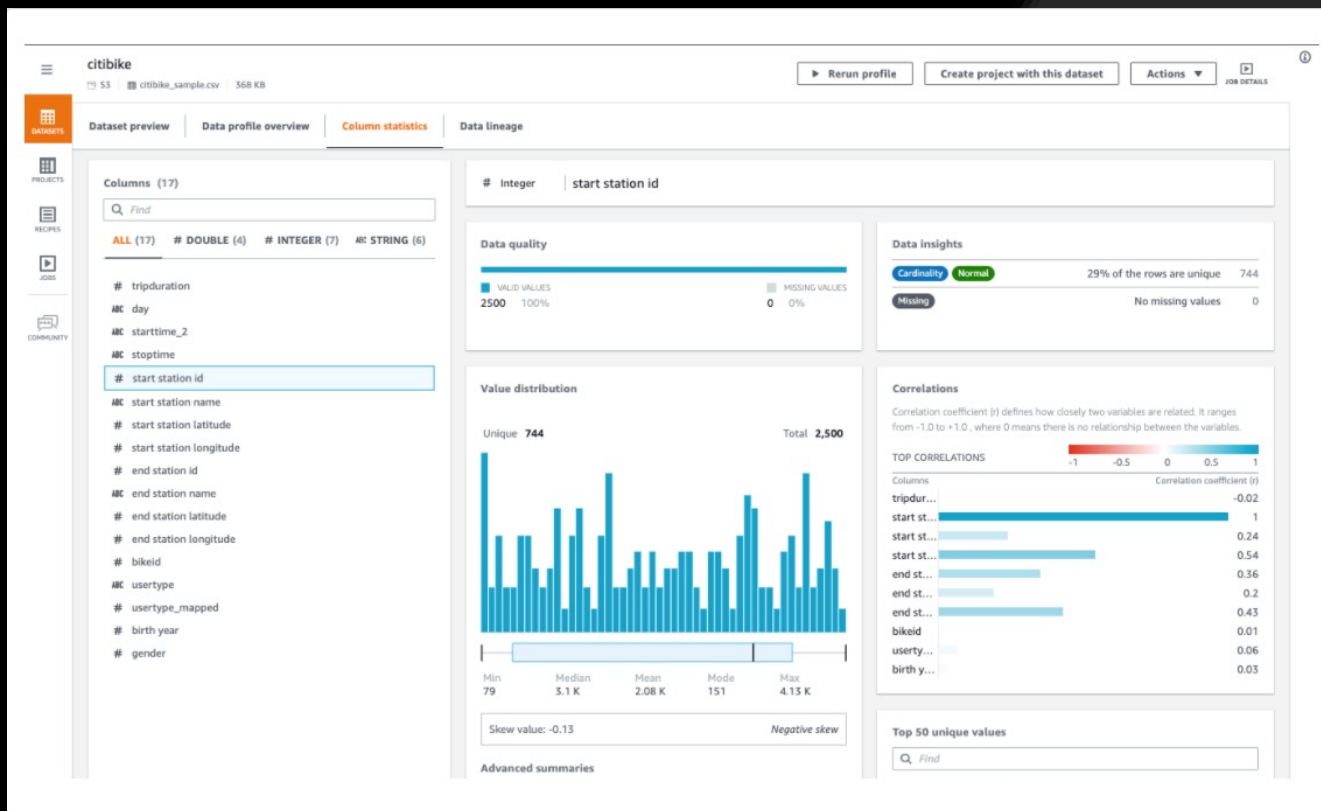
Interactive Session



AWS Glue

AWS Glue DataBrew

ノーコードデータプレ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

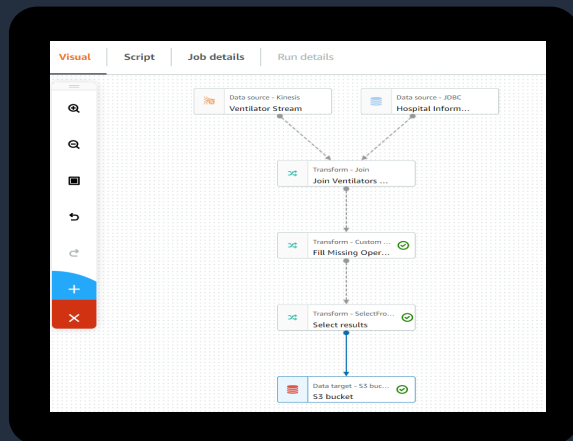


ETL developer
Data Engineer



Business Analyst
Data scientist

AWS Glue Studio

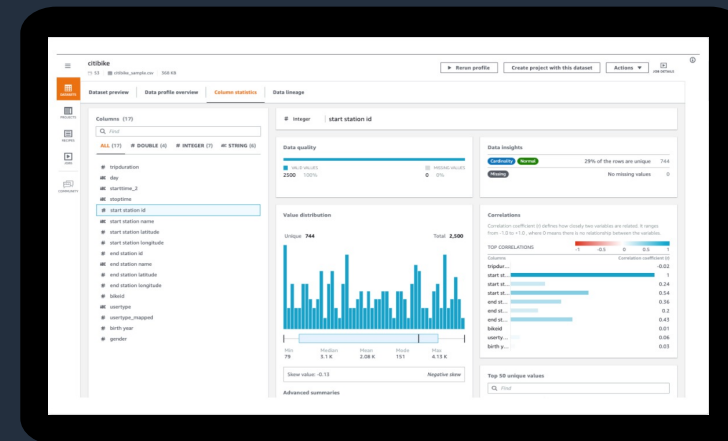


ローコードで迅速に開発

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

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

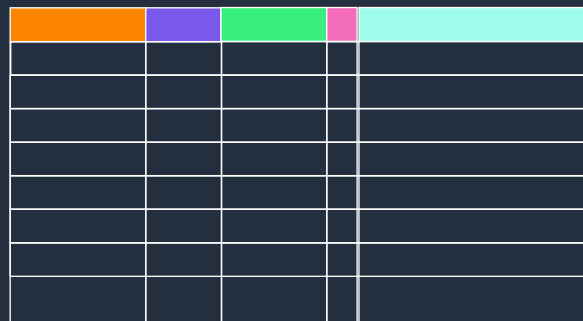
データプロファイル

DynamicFrame

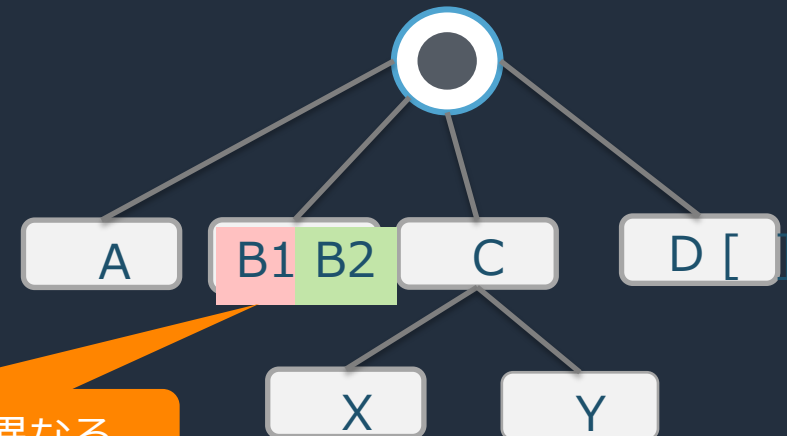
複雑な ETL 処理をシンプルに行うことができるフレームワーク

- 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



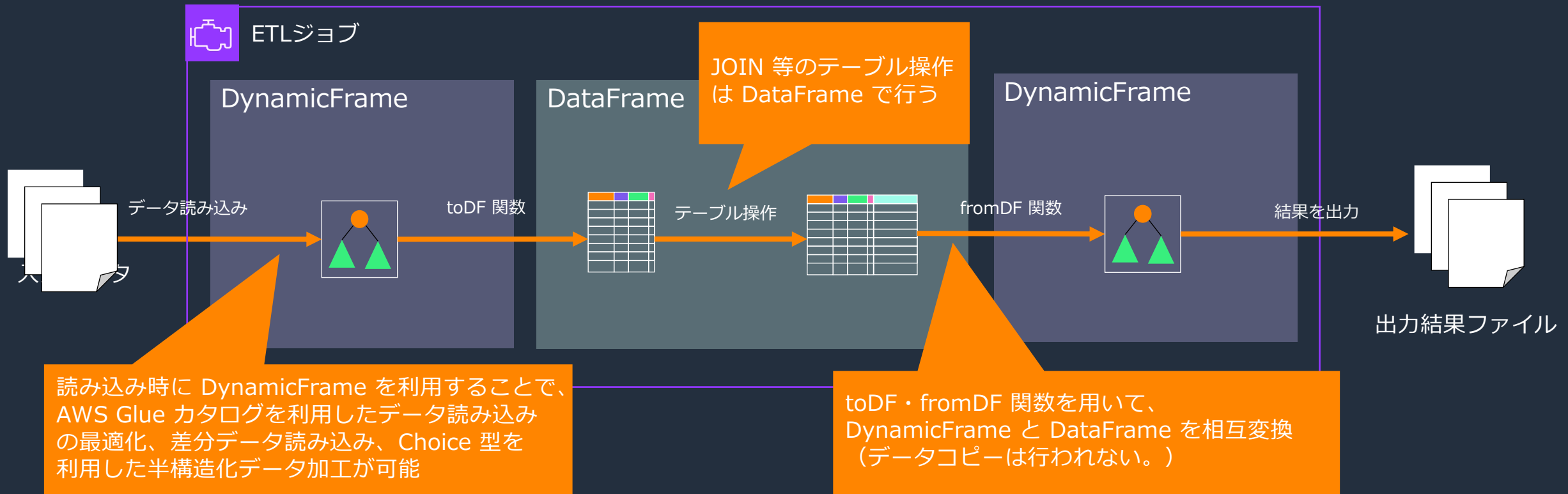
DynamicFrame schema



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

The image shows a screenshot of the AWS Glue Data Quality interface. On the right, a workflow diagram illustrates the process: 'Data source - Data Catalog Data Catalog table' feeds into a 'Transform - Evaluate data quality Evaluate Data Quality' step, which then outputs to 'Data target - S3 bucket S3 bucket'. On the left, a table displays the results of 12 rules for a specific dataset. The table has columns for 'Name', 'Rule', 'Status', and 'Details'. Rules 5 and 6 are marked as 'Rule failed', while the others are 'Rule passed'.

Name	Rule	Status	Details
RS1	Rule_1 RowCount between 1021908 and 4087632	Rule passed	
	Rule_2 IsComplete "TAG_NUMBER_MASKED"	Rule passed	
	Rule_3 ColumnLength "TAG_NUMBER_MASKED" between 6 and 9	Rule passed	
	Rule_4 IsComplete "DATE_OF_INFRACTION"	Rule passed	
	Rule_5 StandardDeviation "DATE_OF_INFRACTION" between 324.19 and 358.31	Rule failed	Expected type of column D
	Rule_6 ColumnValues "DATE_OF_INFRACTION" between 20180100 and 201812	Rule failed	Expected type of column D
	Rule_7 IsComplete "TICKET_DATE"	Rule passed	
	Rule_8 ColumnLength "TICKET_DATE" = 10	Rule passed	
	Rule_9 IsComplete "TICKET_NUMBER"	Rule passed	
	Rule_10 StandardDeviation "TICKET_NUMBER" between 16289083634.63 and 1	Rule passed	
	Rule_11 ColumnValues "TICKET_NUMBER" between 10102501 and 8008108105	Rule passed	
	Rule_12 IsComplete "OFFICER"	Rule passed	

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その他、**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

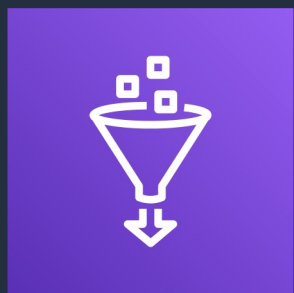
Japan data types

- 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>



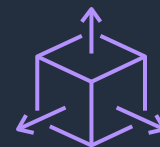
データ収集、カタログ化、データ加工を一気通貫して 行うことができる統合的なデータインテグレーションサービス



AWS Glue



データカタログ



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

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

トリガーのプロパティを設定する

名前

トリガーの名前を入力します...

▶ タグ (任意)

トリガータイプ

スケジュール ジョブイベント オンデマンド

EventBridge event

トリガーの開始をタイマーで行う場合は [スケジュール]、ジョブイベントが監視対象リストに一致した場合に行う場合は [ジョブイベント]、すぐに開始する場合は [オンデマンド] を選択します。

頻度

毎日

Start Hour (UTC)

08

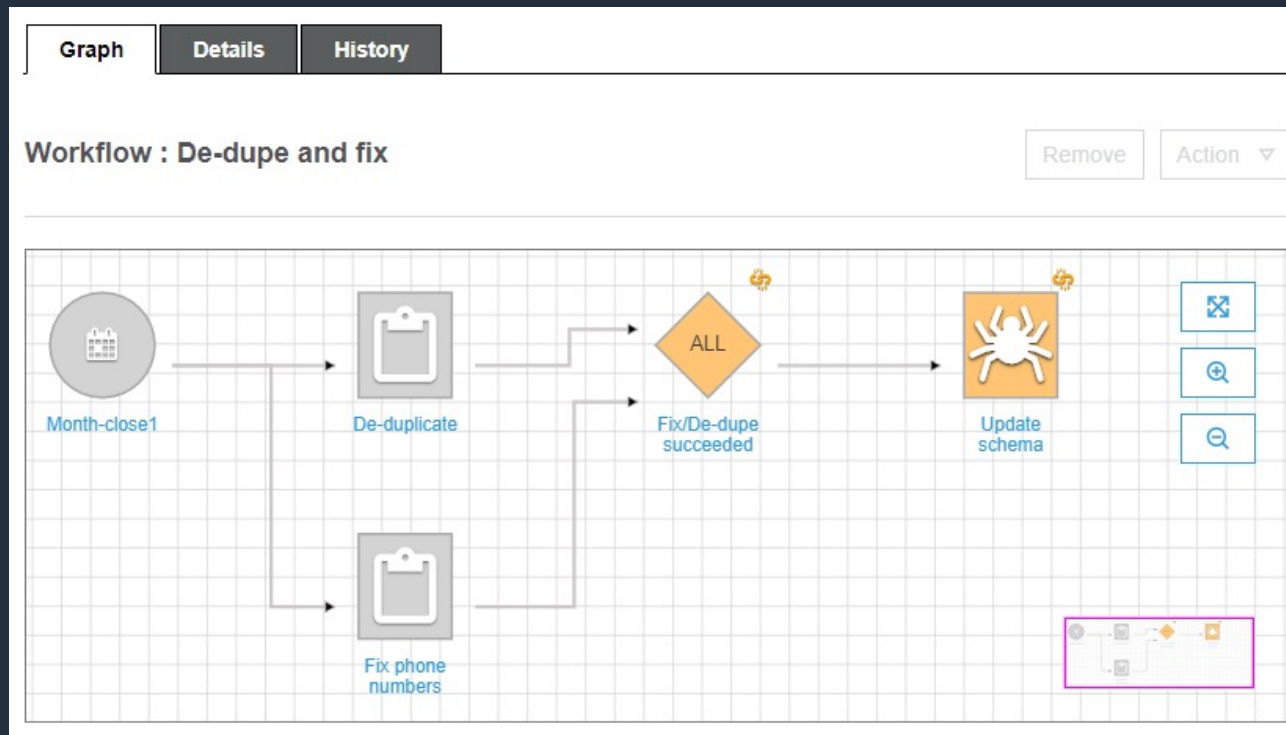
Start Minute

55

次へ

ワークフロー管理

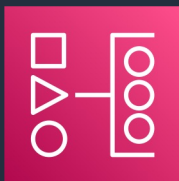
- 複数の Glue ETL ジョブの実行タイミングや依存関係を管理
- クローラ、トリガー、ジョブの組み合わせによりワークフローを定義
- トリガーはスケジュール(日時・曜日・cron)、ジョブイベント、Amazon EventBridge、手動(即時実行)で指定可能



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/>

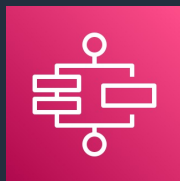


Amazon MWAA
(Apache Airflow)

大規模ETL+OSS



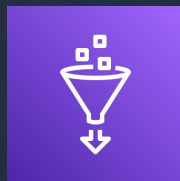
- ETL 運用に最適化した視認性の高い運用 GUI
- デバッグやリトライ制御の操作性
- Python 言語でパイプライン構築
- AWSに閉じない各種サービスとの連携 (Operator)
- 完全閉域でのダッシュボード利用
- OSS ベース (ハイブリッド環境)



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



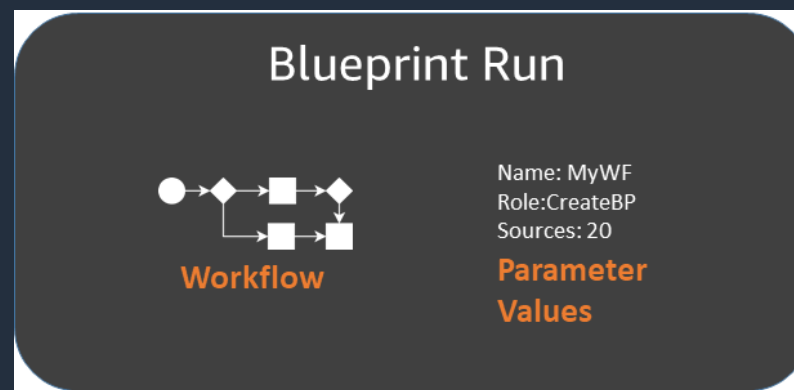
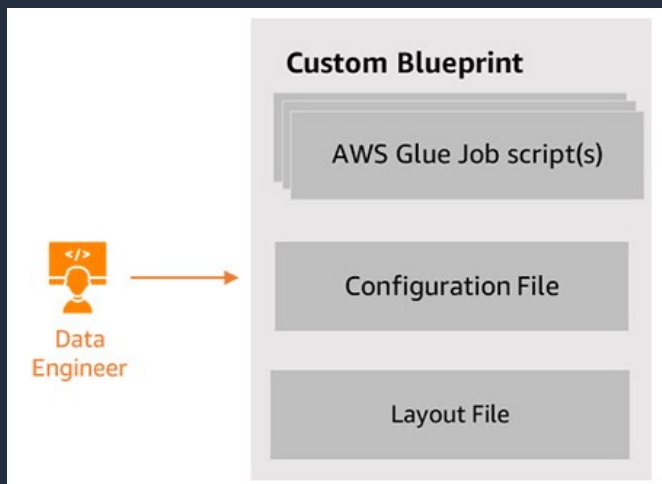
Glue 内で全て完結

- 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

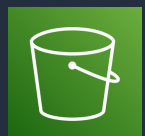
汎用化可能なワークフローをテンプレート化して共有

- 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



ブックマーク機能

Data Source



Amazon S3

差分抽出



RDBMS

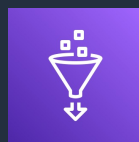
JDBC 接続



Amazon RDS



Amazon Aurora



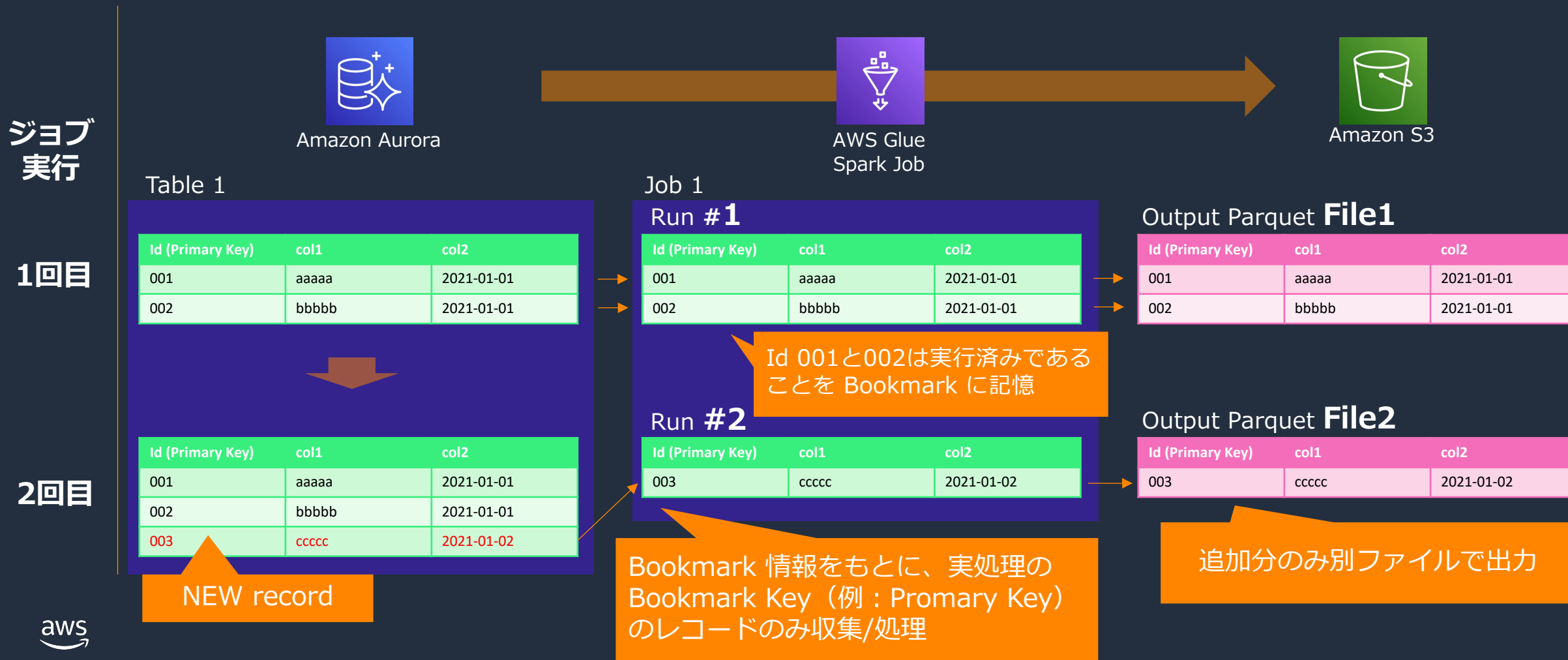
AWS Glue

- 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



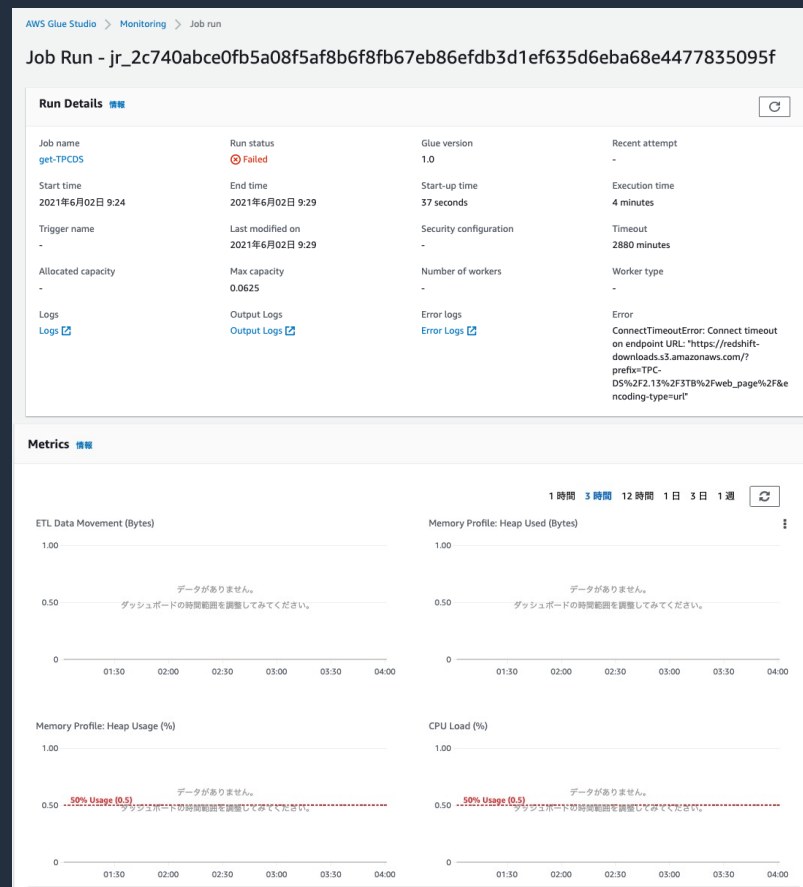
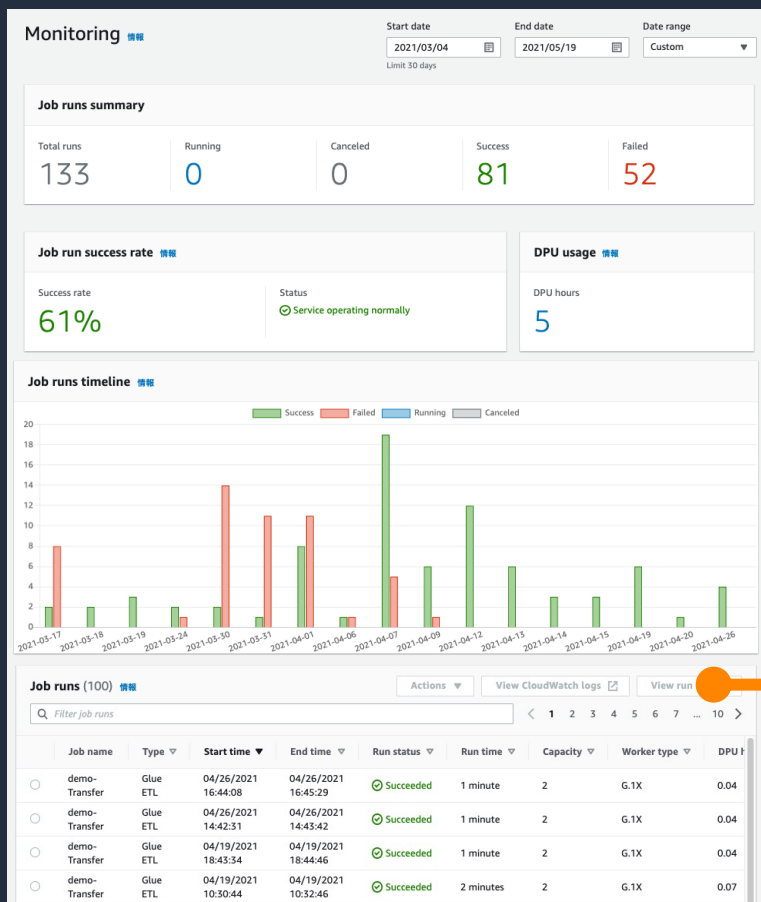
ブックマーク機能

- ✓ 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



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

Spark 2.4.3 History Server

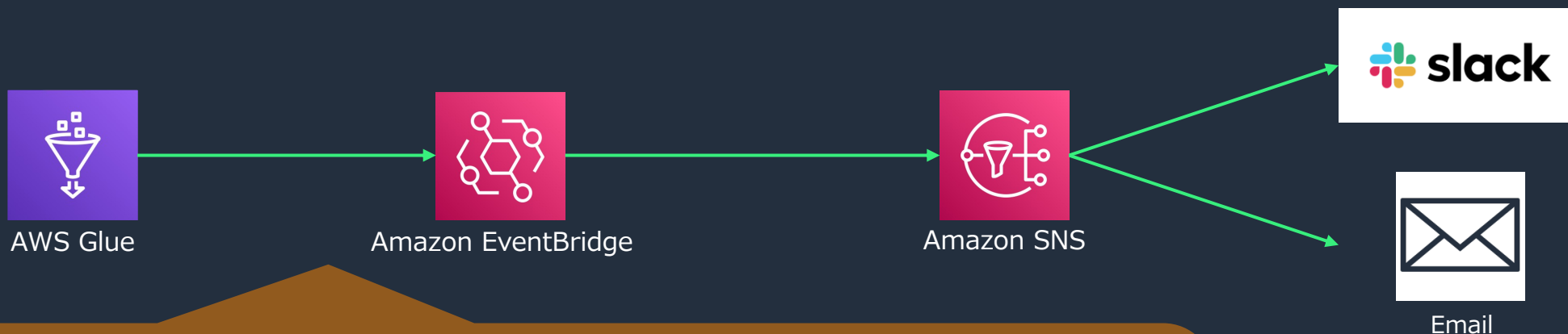
Event log directory: s3a://glueworkshop-chia/sparkeventlogs
There are 990 event log(s) currently being processed which may result in additional applications getting listed on this page. Refresh the page to view updates.
Client local time zone: Asia/Tokyo

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

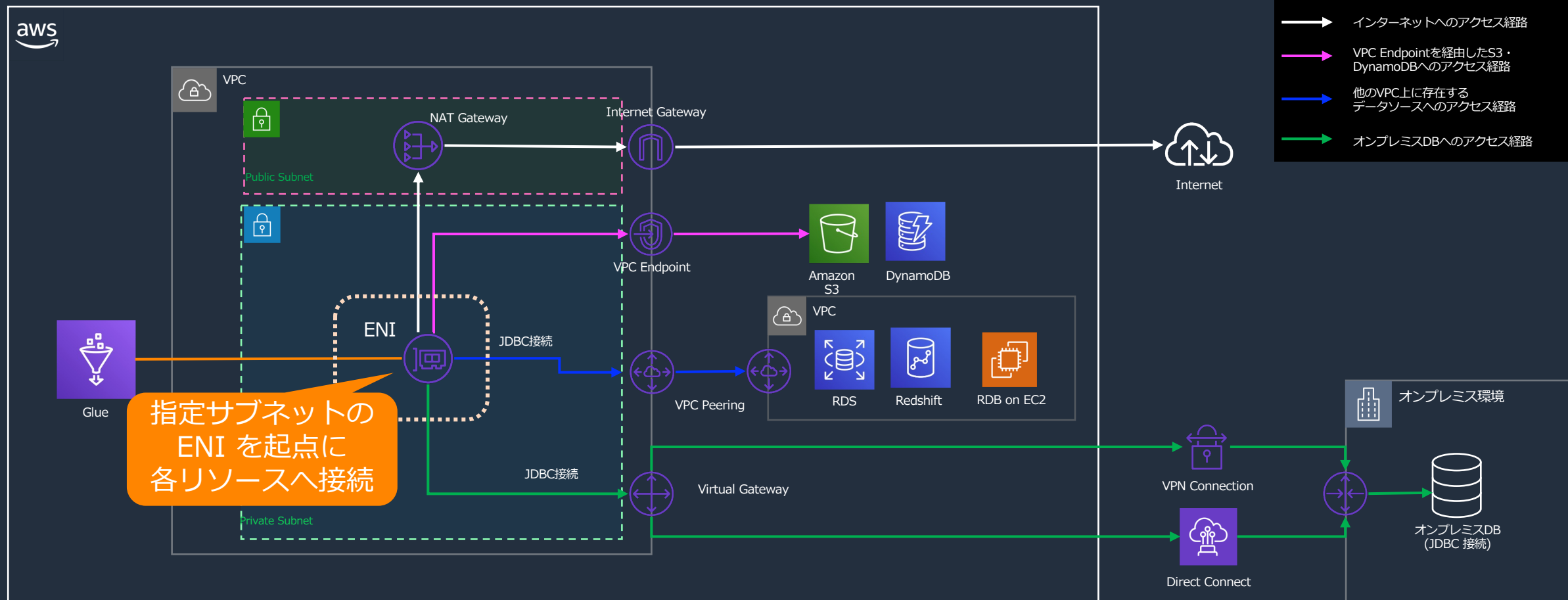


- カタログの状態変化（スキーマ変更など）
- ジョブステータスの変化（ジョブ終了、失敗など）
- ジョブ実行時間閾値超過
- Glue クローラのステータスの変化
- etc

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)

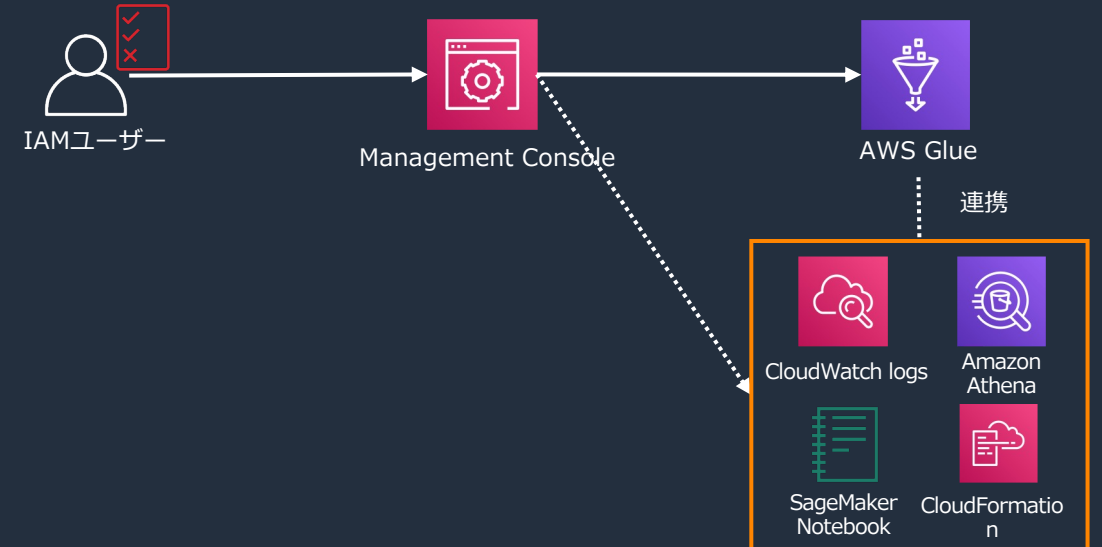


Glue ジョブやクローラーが接続可能なネットワーク、VPC 内の ENI から到達可能な範囲に限定できるため、閉域網に閉じた構成にも対応可能です

IAM による権限管理

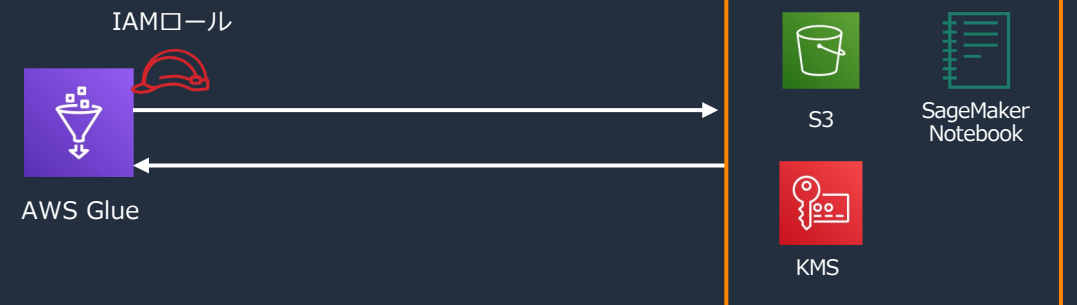
IAM ユーザー

- 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

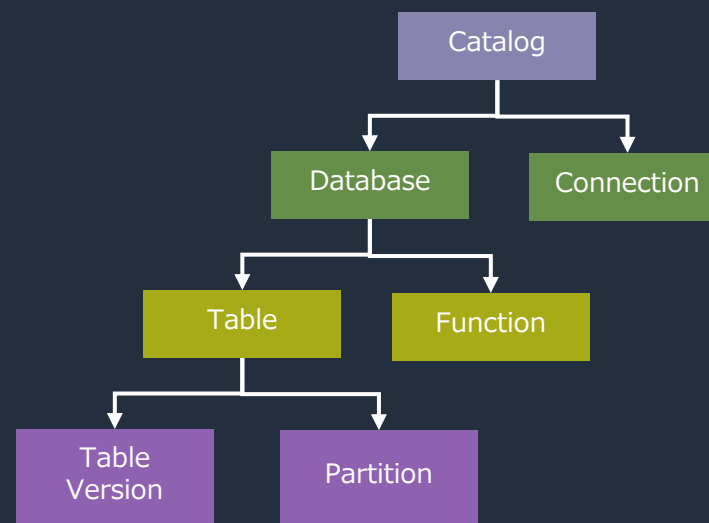
AWS Glue 内で管理するアクセスポリシー

- 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"
      ]
    }
  ]
}
```

データカタログのリソース階層



※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

The screenshot displays the AWS CloudTrail console's 'Event history' page. The left sidebar contains navigation options: Dashboard, Event history (selected), Insights, Trails, Pricing, Documentation, Forums, and FAQs. The main content area shows the 'Event history (50+)' section with a search filter for 'glue.amazonaws.com' and a time range of '30m'. A table lists the following events:

Event name	Event time	User name	Event source	Resource type	Resource name
GetJobRuns	September 20, 2021, 18:10:44 (...)	[Redacted]	glue.amazonaws.com	-	-
GetJobRuns	September 20, 2021, 18:10:44 (...)	[Redacted]	glue.amazonaws.com	-	-
UpdateDevEndpoint	September 20, 2021, 18:10:37 (...)	SageMaker	glue.amazonaws.com	-	-
GetDevEndpoint	September 20, 2021, 18:10:37 (...)	SageMaker	glue.amazonaws.com	-	-
UpdateDevEndpoint	September 20, 2021, 18:10:32 (...)	SageMaker	glue.amazonaws.com	-	-
GetDevEndpoint	September 20, 2021, 18:10:32 (...)	SageMaker	glue.amazonaws.com	-	-
GetDevEndpoint	September 20, 2021, 18:10:27 (...)	SageMaker	glue.amazonaws.com	-	-
GetDevEndpoint	September 20, 2021, 18:10:27 (...)	SageMaker	glue.amazonaws.com	-	-
UpdateDevEndpoint	September 20, 2021, 18:10:27 (...)	SageMaker	glue.amazonaws.com	-	-
GetDevEndpoint	September 20, 2021, 18:10:22 (...)	SageMaker	glue.amazonaws.com	-	-
GetDevEndpoint	September 20, 2021, 18:10:22 (...)	SageMaker	glue.amazonaws.com	-	-
GetJobRuns	September 20, 2021, 18:10:13 (...)	[Redacted]	glue.amazonaws.com	-	-

demo



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 > ⚙

<input type="checkbox"/>	Job name	Type	Last modified	AWS Glue version
<input type="checkbox"/>	glueworkshop-lab5-advance-job-error	Glue ETL	3/3/2023, 8:29:00 AM	3.0
<input type="checkbox"/>	glueworkshop-lab3-etl-job	Glue ETL	3/3/2023, 8:08:05 AM	3.0

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

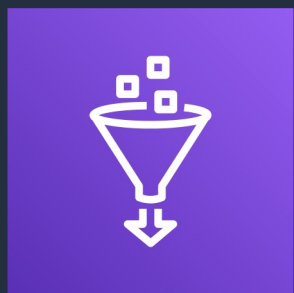
導入事例

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

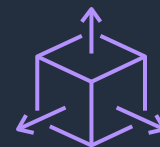
データ収集、カタログ化、データ加工を一気通貫して 行うことができる統合的なデータインテグレーションサービス



AWS Glue



データカタログ



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

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!