20分で理解する

AWS Lake Formation 概要

2021年11月

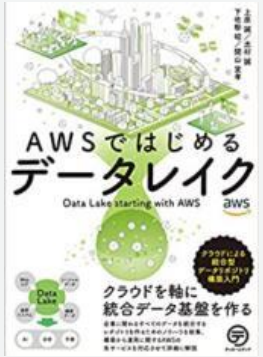
アマゾン ウェブ サービス ジャパン 合同会社
シニアソリューションアーキテクト
下佐粉 昭 (Akira Shimosako)

自己紹介

- 下佐粉 昭（しもさこ あきら）
  アマゾン ウェブ サービス ジャパン
  シニアソリューションアーキテクト (アナリティクス)

  @simosako

- 「AWSではじめるデータレイク」
  https://techiemedia.co.jp/



- 週刊AWS：
  毎週AWSのアップデートをサマリしてお届け

アジェンダ

AWS GlueとAWS Lake Formationの位置づけ

Blueprint - データ取り込み

Lake Formation Permission - 新しいアクセス許可モデル

まとめ

AWS GlueとAWS Lake Formation の位置づけ

データレイクを中心とした分析環境 on AWSクラウド

- データレイクの蓄積部分は Amazon S3
- データを活用するには、蓄積以外にデータカタログと、ETL処理が必要になる

AWS Glue
ETLとデータカタログをサーバレスで提供

AWS Glue

サーバーレス
ETL

スケジューラーと
ワークフロー

コードに集中

カタログ
(Glue data catalog)

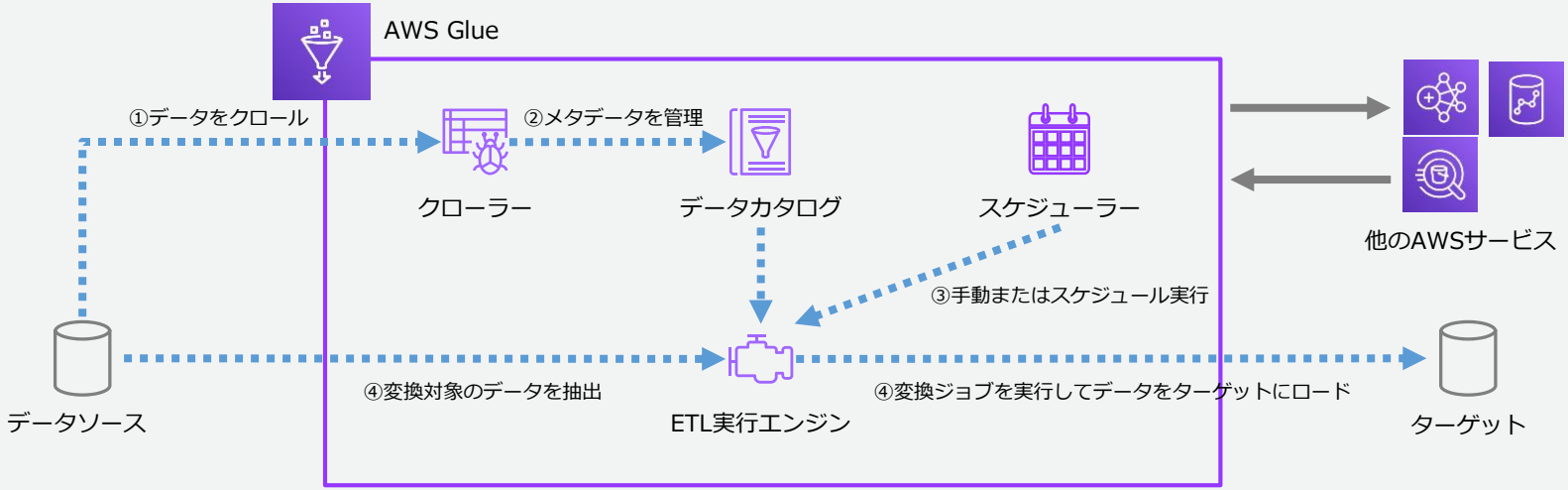
VPC内からのアクセス

他のAWSサービスと
容易に連携

セキュア

Glue Studio

AWS Glueの全体像



概要

①クローラーにてデータソースのメタデータをクロールして、データカタログに登録・更新

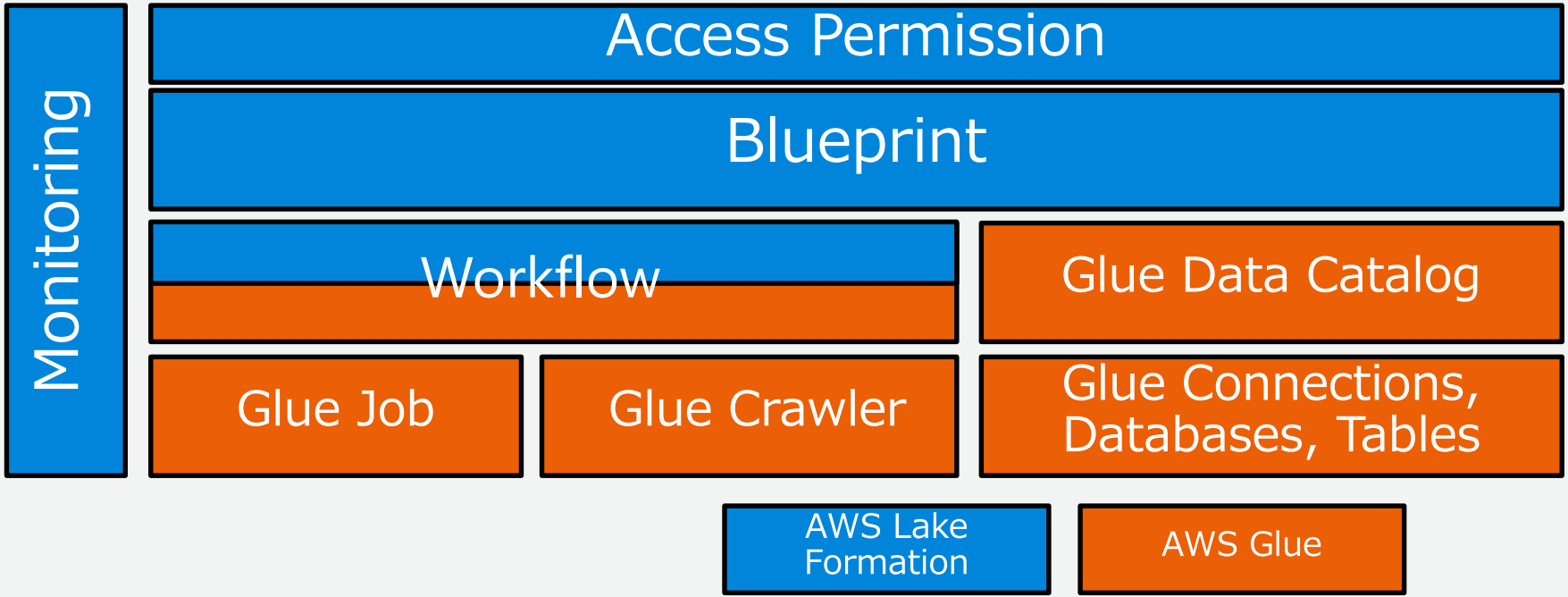
②データカタログにてメタデータを管理

③スケジューラーにてジョブの実行タイミングを定義

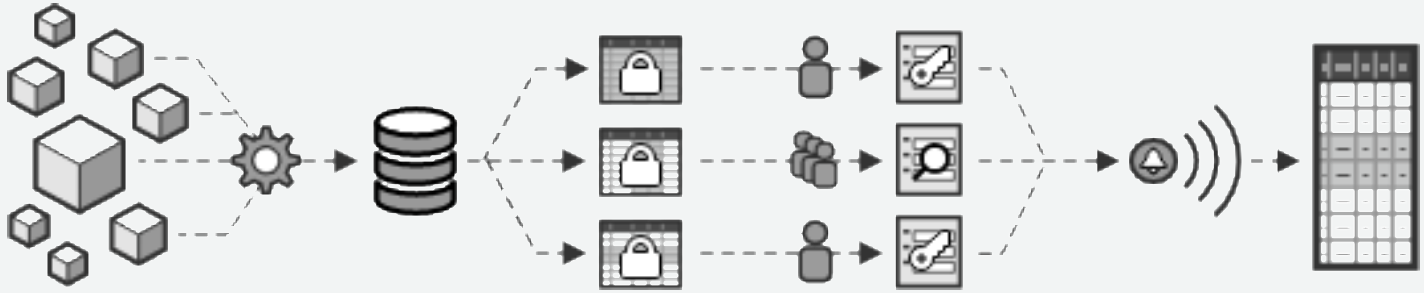
④データソースからデータを抽出し、ETL実行エンジンにてジョブをサーバーレスで実行
(ジョブはSpark(PySpark、Scala)またはPython Shellを選択)

AWS Glueを拡張する形で提供される AWS Lake Formation

Lake Formation は Glue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



AWS Lake Formation 概要



データ取込みと構造化の自動化

- データ取込み、整形、暗号化
- Amazon S3バケットに保存し、登録

セキュリティ & コントロール

- 適切なユーザー、グループに正しいデータへのアクセス制御を定義
- データベース、表、列の単位の粒度で制御可能

利用の促進

- メタデータカタログを利用した検索と定義確認
- 新しいデータ取り込み時のポリシーを定義可能

モニタリングと監査

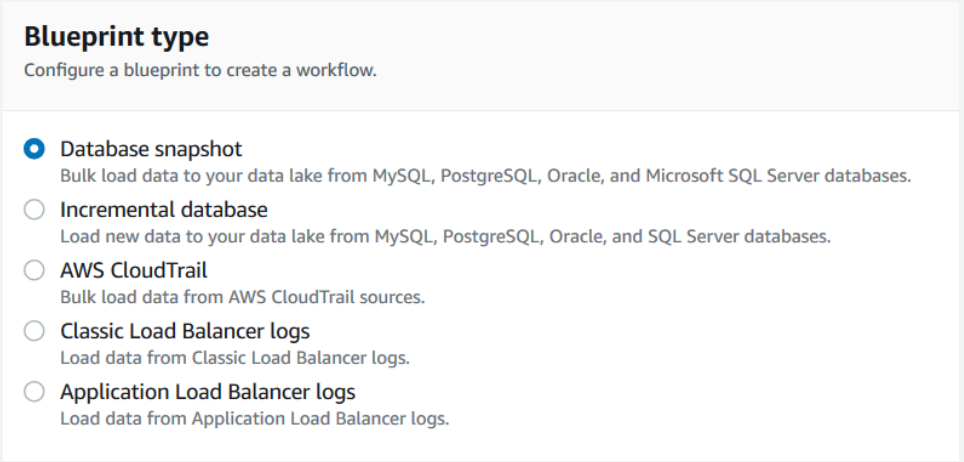
- アクセス要求や発生したポリシー例外を記録
- アクティビティ履歴で詳細に変更ログやデータの入手経路をレビュー

Blueprint – データ取り込み

データの取り込み - Blueprint

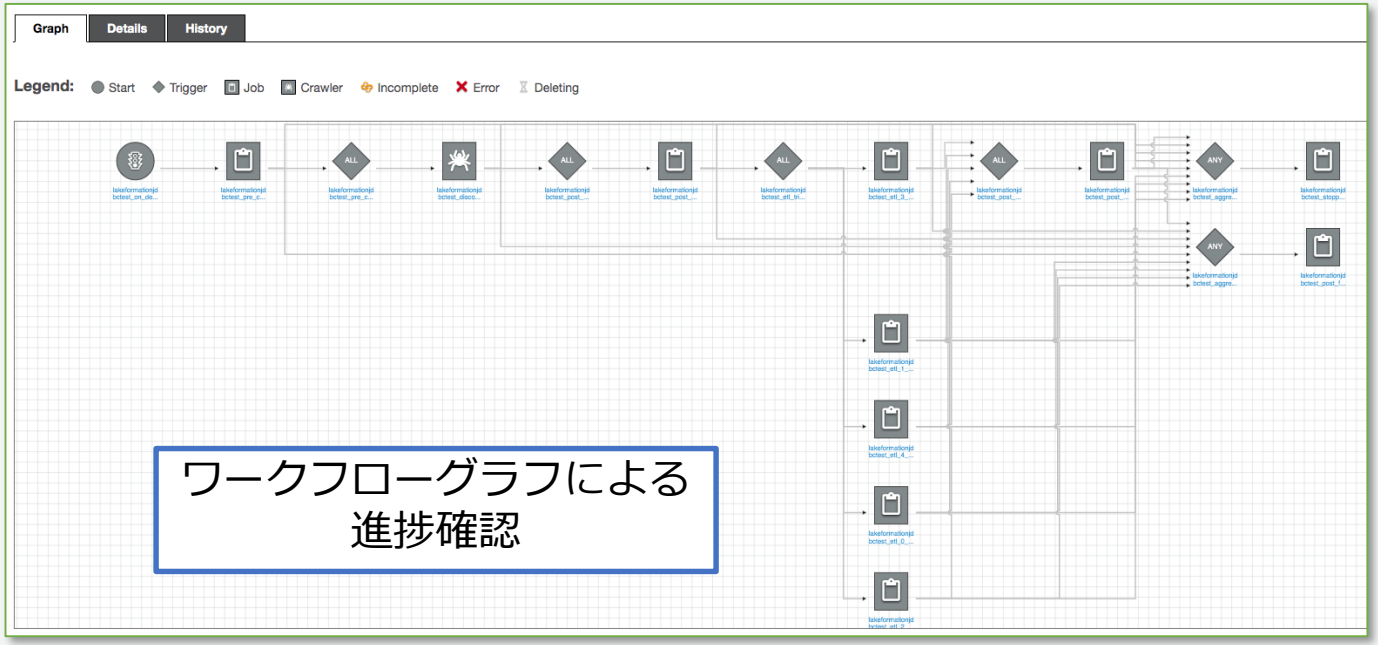
Blueprintは、データレイクにデータを簡単に取込むことを実現する「データ管理のテンプレート」
汎用的なユースケースで必要になるAWS Glueのワークフロー、クローラー、ジョブなどを生成
ファイルフォーマットの変更や暗号化保存の指定も可能

- Database snapshot
  - ✓ JDBCデータベースから表の全データを取り込み
- Incremental database
  - ✓ JDBCデータベースから新データのみ、差分で取り込み
- Log file
  - ✓ AWS CloudTrail、CLB Log、ALB Logの取り込み



ワークフロー：Blueprintから実行される処理フロー

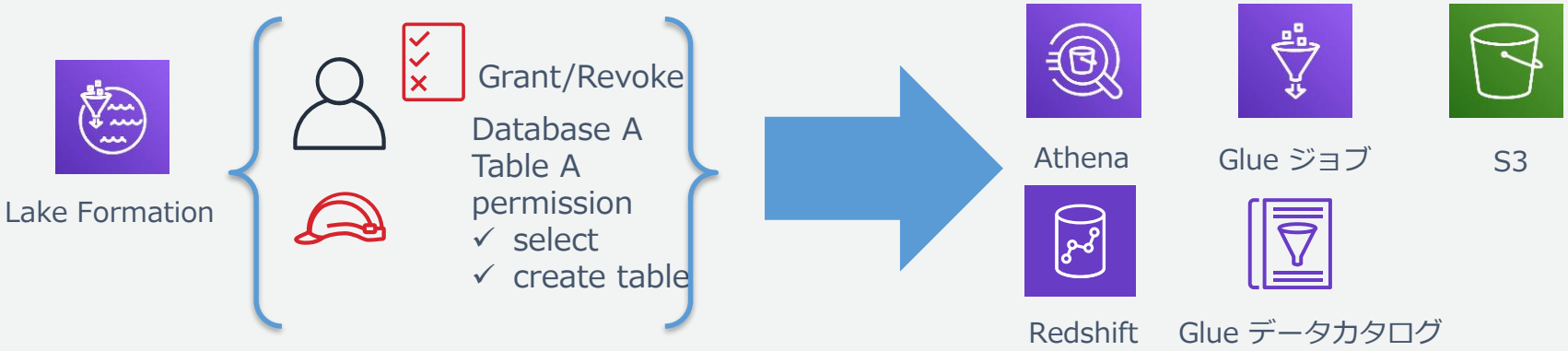
Lake Formationのワークフローは、Blueprintから生成される
一連の AWS Glue ジョブ、クローラー、トリガー等の定義セット
ワークフローグラフにより進捗を可視化することが可能



Lake Formation Permission
- 新しいアクセス許可モデル

Lake Formation Permission – 新しいアクセス許可モデル

- Lake Formation は、IAM を拡張した独自のアクセス許可モデルを持ち、データレイク内のデータへのアクセスを保護します
- Lake Formation のアクセス許可モデルではシンプルな Grant/Revoke により、データレイクに格納されているデータの一元的なきめ細かいアクセス制御が可能です

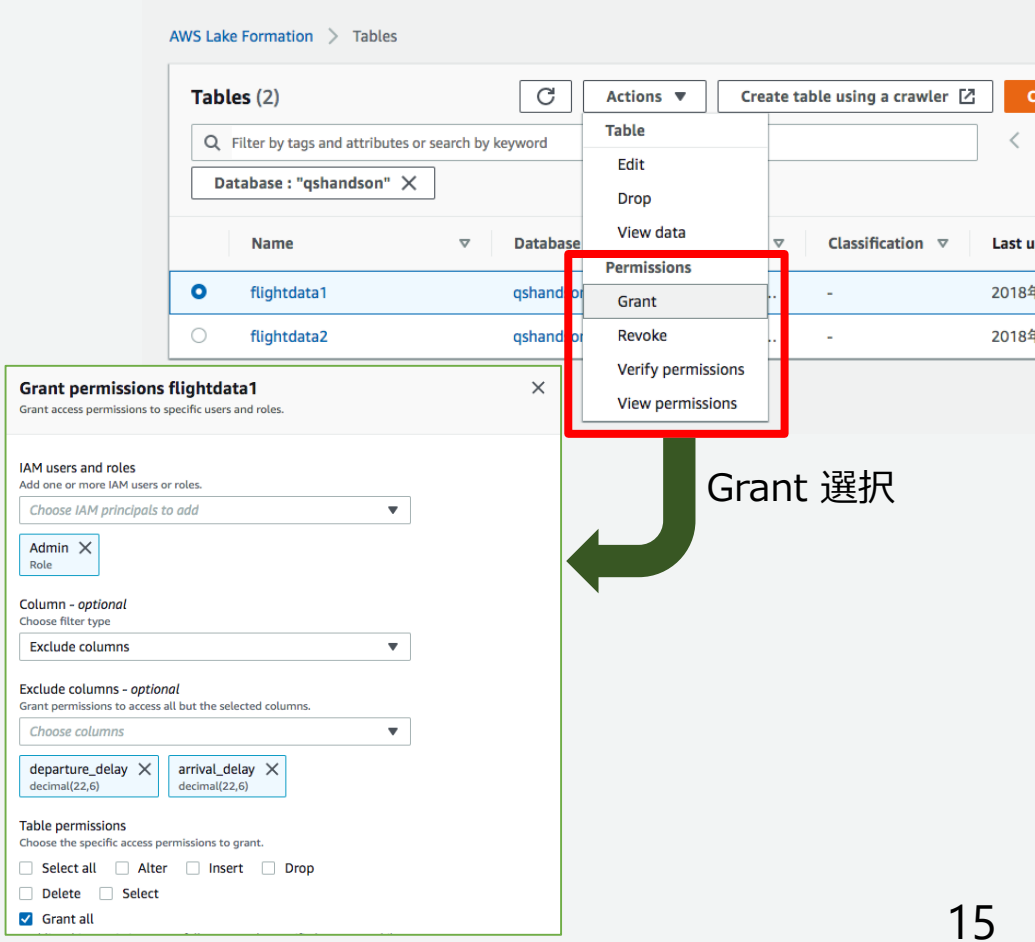


シンプルな集中管理を実現

複数サービスからのデータアクセスを
1箇所で定義／制御
（サービス個別に設定する必要が無い）

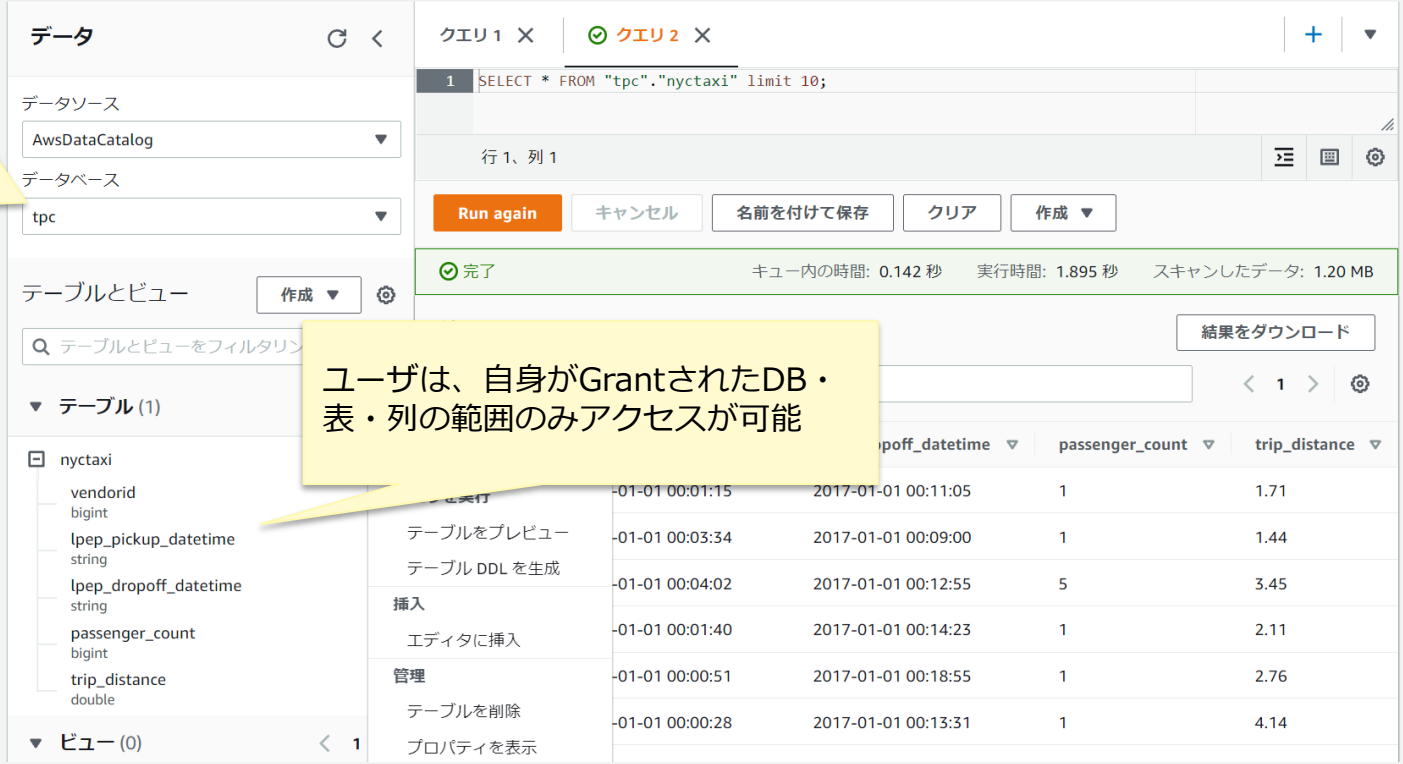
カタログ上でデータベースやテーブルとして登録し、Grant/Revokeスタイルで制御
（実装を意識しなくて良い）

柔軟なアクセスコントロール
- 「読み取りのみ可能」といった制御
- DB、表、列単位のアクセス制御



例：Amazon Athenaからのアクセス

Lake Formationのカタログに定義されたデータベース、テーブルがAthenaから認識される



ユーザは、自身がGrantされたDB・表・列の範囲のみアクセスが可能

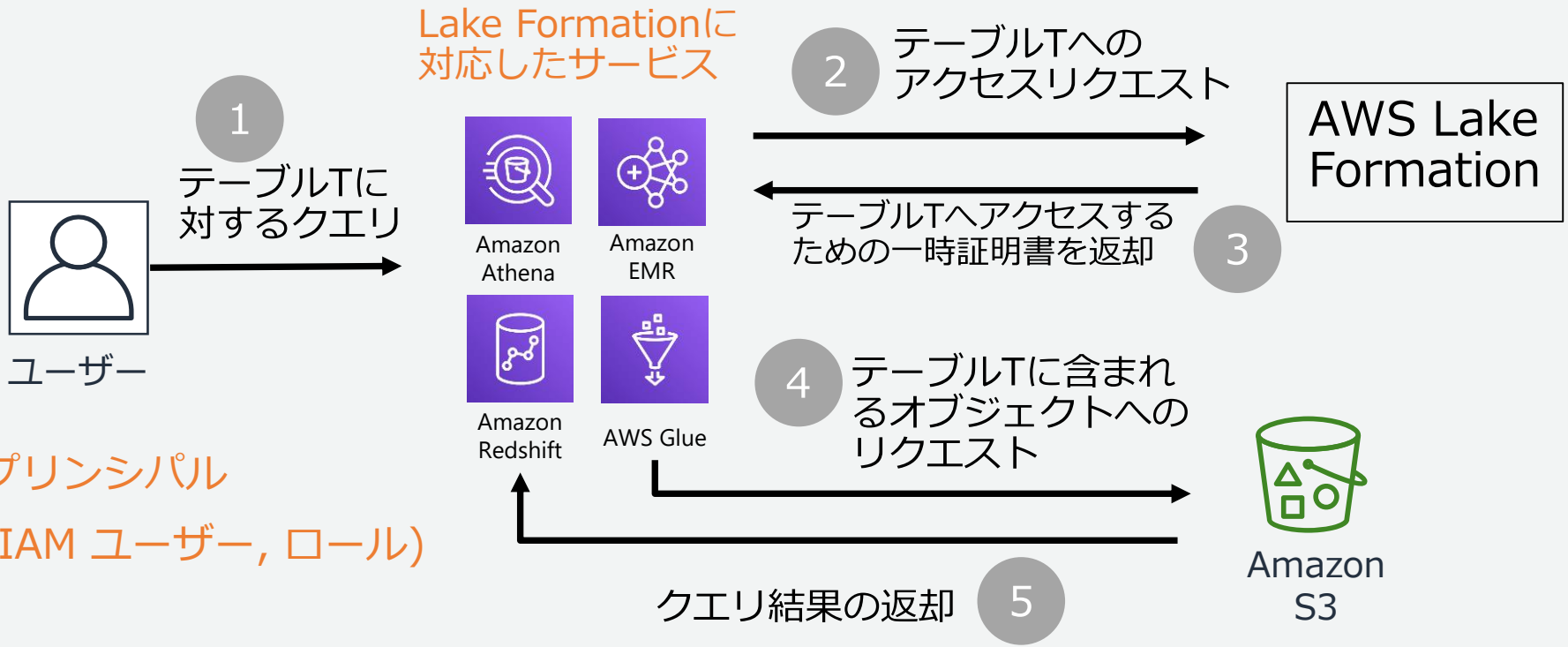
リソース名ベースのアクセス制御と タグ（LF-Tag）ベースのアクセス制御

- リソース名ベースのアクセス制御
  - ✓ リソース（DB、表、列）の名前単位に、プリンシパル（IAMユーザ、ロール）の権限をGrantする
    - ➢ 例： 表Aのコスト列に対して、ユーザZのSELECT権限をGrantする

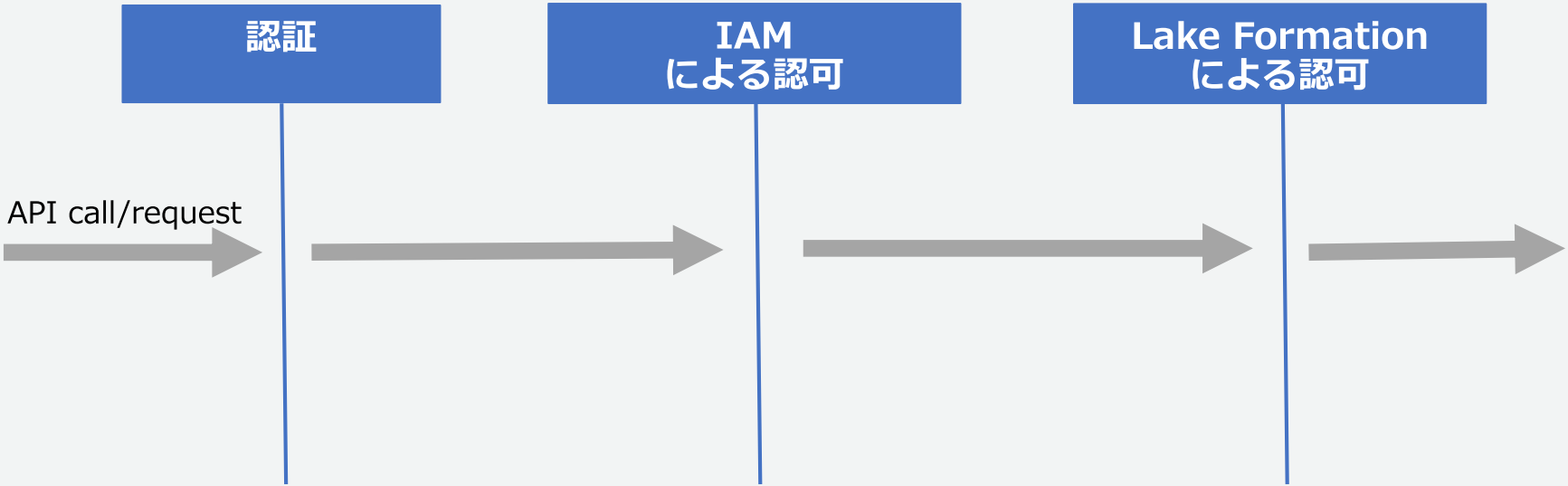
- タグ(LF-Tag)ベースのアクセス制御
  1. タグを作成：キーとバリューの組み合わせを定義
     - ➢ 例： group = [analyist, developer]
  2. カタログ上でタグをリソース（DB、表、列）に付与する
     - ➢ 表Aのコスト列に、 group:analyst を付与
  3. プリンシパルに特定のタグと権限をGrantする
     - ➢ 例：ユーザZはアナリストなので group:analyst タグが付いたリソースへのSELECTを許可

タグベースでの設計が推奨

Lake Formation Permissionの全体イメージ



Lake Formation Permissionを有効にしても IAMは引き続き有効です



これまで通りIAMで認証される
- パスワード
- シークレットキー
- Active Directory / SAML

IAMは引き続きすべてのリクエストを認可する
- IAMポリシー
- リソースポリシー
- タグベースのポリシー

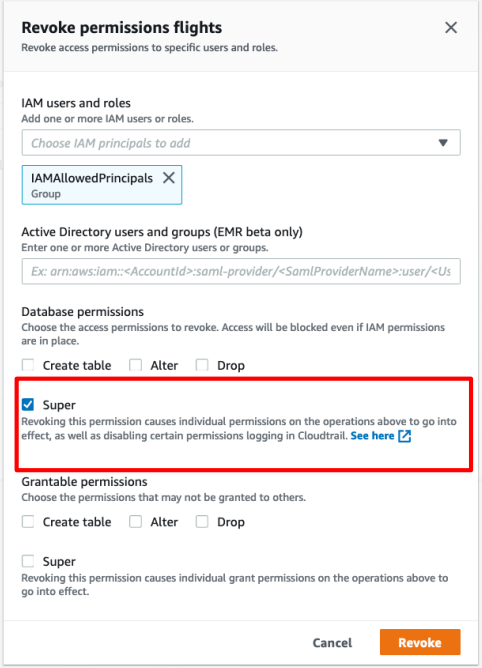
Lake FormationはLake Formation パーミッションへのすべてのアクセスを認可する

初期状態ではIAM許可によるパススルーが有効

- 初期状態ではIAMで認可された場合、Lake Formationの認可をパススルーされる挙動になっています。これは既存のアプリケーションと互換性を保つためです
  - ✓ これはIAMAllowedPrincipalsグループ経由で実行されます

- 新規作成データベースや表にこのデフォルトを適用しないように変更することが可能です
  - ✓ データベースの **“Use only IAM access control for new tables in this database”** のチェックを外す
  - ✓ データカタログのデフォルト設定の **“Use only IAM access control for new tables in this database”** のチェックを外す

- 既存のリソースに対してパススルーを止める場合はIAMAllowedPrincipalsからSuperをRevokeします（右図）



暗黙的なアクセス許可

- データベースの作成など特定の Lake Formation タスクを実行すると、暗黙的な権限の付与が行われます

- 一般的な利用用途に即した暗黙的な権限付与を行うことで、オペレーションを簡素化します

- データベース作成者
  - ✓ 作成するデータベース内のすべてのテーブルに対するすべての権限を持ちます。
- テーブル作成者
  - ✓ 作成するテーブルに対するすべての権限を持ちます。作成するテーブルにアクセス許可を付与できます。
- データレイクユーザー
  - ✓ 権限を持つデータベースまたはテーブルを表示および一覧表示できます。

https://docs.aws.amazon.com/ja_jp/lake-formation/latest/dg/implicit-permissions.html

補足：データレイク管理者の暗黙的なアクセス許可

データレイク管理者※に以下の暗黙的なアクセス許可が与えられます

- データカタログ内のすべてのオブジェクトへの完全なメタデータアクセスがあります。このアクセスは管理者から取り消すことはできません。
- データレイク内のすべての Data Location Permission があります。
- データカタログ内のオブジェクトへのアクセスをプリンシパルに許可できます。このアクセスは管理者から取り消すことはできません。
- データカタログにデータベースを作成できます。
- 別のユーザーにデータベースを作成する権限を付与できます。
- データロケーションに S3 パスを登録できます。

データレイク管理者には、データベースを削除したり、テーブルを変更または削除するための暗黙的な権限はありません。ただしそのための権限を自分で付与できます。

※データレイク管理者は IAM の Administrator 権限を持つユーザーではなく、Lake Formation で定義される管理者です。

Grant/Revokeで設定できるパーミッション

エンティティと、設定可能なパーミッションは右図の通り

- Databaseはメタデータとしての仮想的な定義（実際にRDBを作るわけではない）
- ＊が付いている部分は自動的に実データにも操作権限が及ぶ
- 例えばカタログ上のTable（メタデータ）にSELECT権限を持つと、実データもSELECTできるようになる
- 一方TableのDROP権限ではカタログから定義を削除するだけで、実データには影響がない

Catalog	Database	Table	Column
CREATE_DATABASE	CREATE_TABLE	ALTER	SELECT*
	ALTER	DROP	
	DROP	SELECT*	
		INSERT*	
		DELETE*	

AWS Lake Formation まとめ

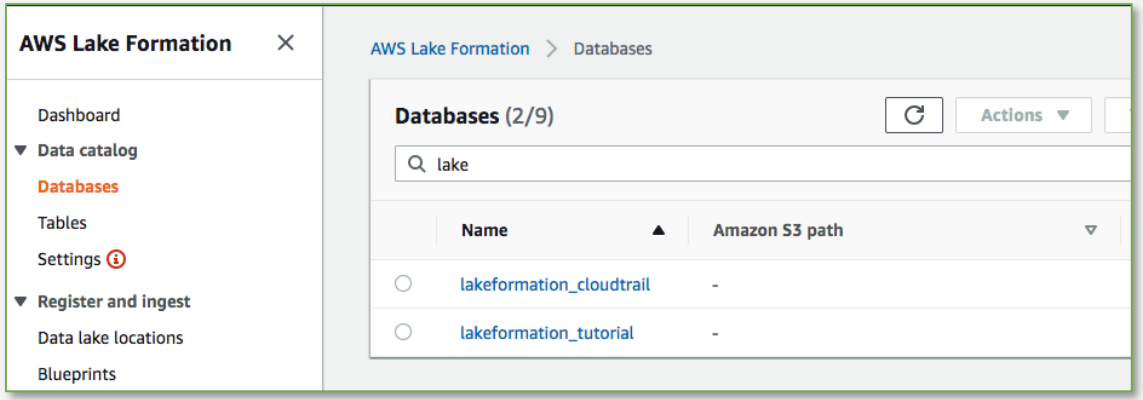
Lake Formationはデータレイク構築・管理を楽にするサービス

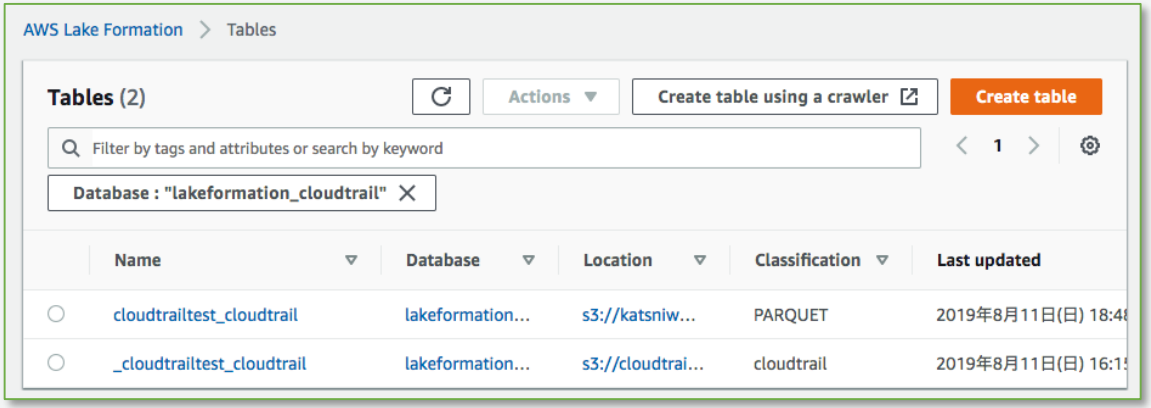
- Blueprint：汎用的なデータ取込み処理のテンプレート化
  - ✓ AWS Glueのジョブとして実装される
- アクセス権限管理をシンプルにする新しいセキュリティモデル
  - ✓ Lake Formationで一元管理
  - ✓ 名前ベースの制御とLF-Tagでの制御が可能
- AWS Lake Formation自体の費用は無償
  - ✓ Lake Formationと連動して動くサービスの費用のみ

補足資料:
AWS Lake Formation
その他の機能

拡張されたデータカタログ

- AWS Glue データカタログを統合・拡張
- クロスアカウント環境での権限管理にも対応
- 管理対象のデータソースをデータベースとテーブルで表現
- データベースはテーブルの集合体、テーブルはデータのスキーマ情報、ロケーションなどを保管





データカタログの検索

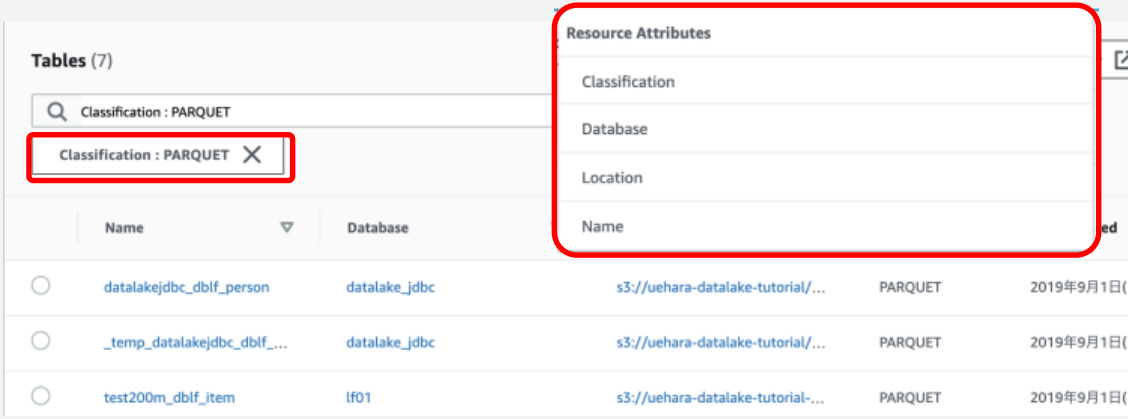
Lake Formation では、テキストベースのファセット検索を全メタデータに対して行えるため、分析に利用可能なデータセットのカタログにセルフサービスでアクセスできます。

- Resource Attributes 検索
  Classification(例えばPARQUET),
  Database(例えばDatabase A),
  Location(S3 URL),Name(テーブル名)

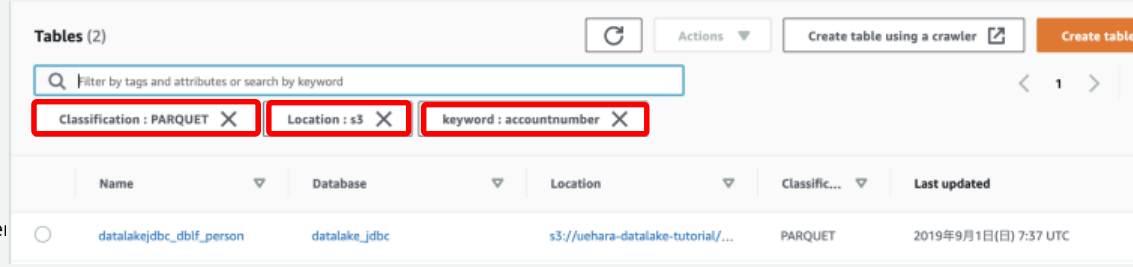
- キーワード検索
  データベース名、テーブル名、列名、Description などのメタデータをキーワード検索

- 複数のフィルタ検索
  上記のいくつかを組み合わせて検索

✓ Classification を PARQUETで検索



✓ Classification を PARQUET、Location を s3、Keyword を accountnumber(列名) で複数フィルタ検索

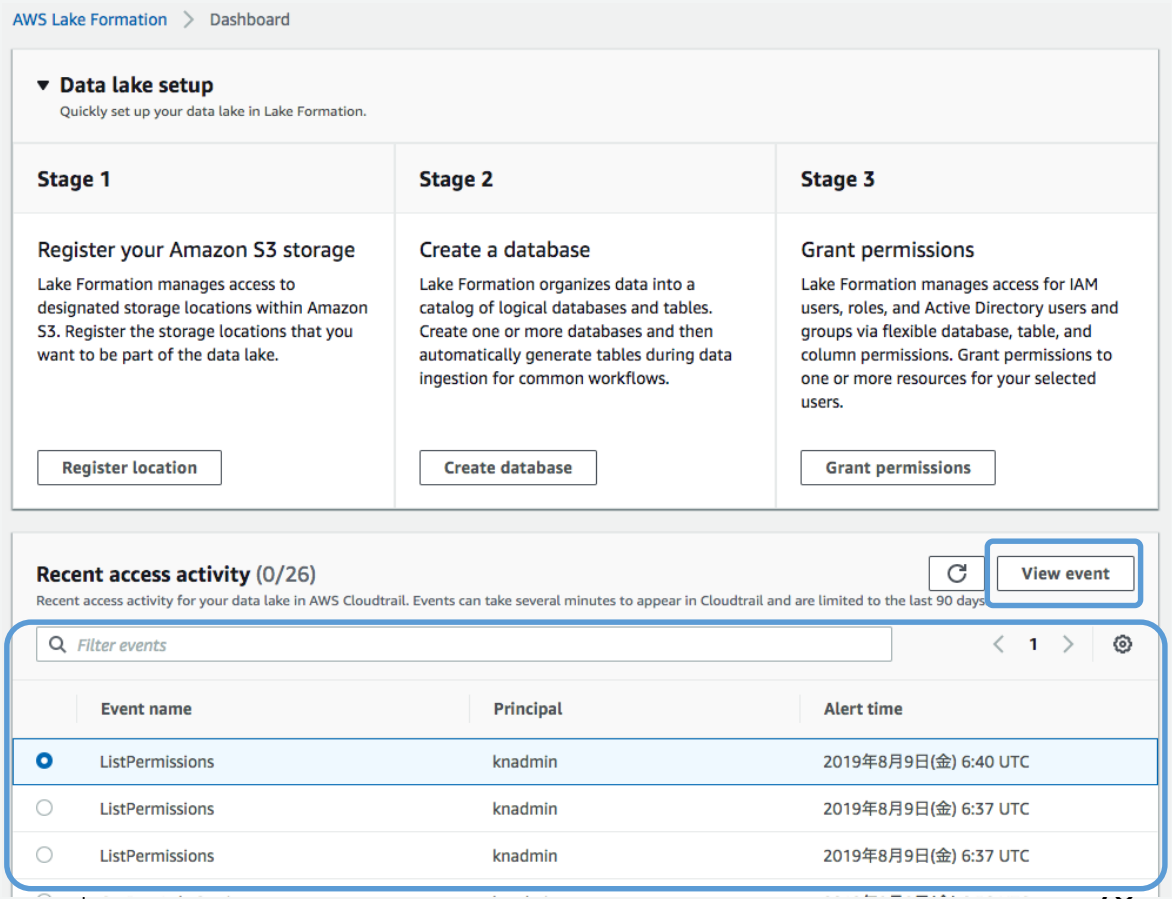


モニタリングと監査

集中的なモニタリングと監査

Lake Formationのコンソールで直近のアクティビティを確認可能

詳細な情報を「View Event」で確認可能（監査）



参考資料

- AWS Lake Formation ホームページ
  https://aws.amazon.com/jp/lake-formation/
- AWS Lake Formation ドキュメンテーション (開発者ガイド)
  https://docs.aws.amazon.com/ja_jp/lake-formation/
- AWS Lake Formation FAQ
  https://aws.amazon.com/jp/lake-formation/faqs/
- データレイクとは
  https://aws.amazon.com/jp/big-data/datalakes-and-analytics/what-is-a-data-lake/
- AWS Lake Formation API
  https://docs.aws.amazon.com/lake-formation/latest/dg/aws-lake-formation-api.html
- CLI Command Reference
  https://docs.aws.amazon.com/cli/latest/reference/lakeformation/index.html#cli-aws-lakeformation
- Cloud Formation Lake Formation Resource Type Reference
  https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/AWS_LakeFormation.html

参考資料(公式ブログ)

- AWS Lake Formation の開始方法
  https://aws.amazon.com/jp/blogs/news/getting-started-with-aws-lake-formation/
- AWS Lake Formation でデータレイクを構築、保護、管理
  https://aws.amazon.com/jp/blogs/news/building-securing-and-managing-data-lakes-with-aws-lake-formation/
- AWS Lake Formation でメタデータを見つける: パート 1
  https://aws.amazon.com/jp/blogs/news/discovering-metadata-with-aws-lake-formation-part-1/
- AWS Lake Formation でメタデータを見つける: パート 2
  https://aws.amazon.com/jp/blogs/news/discover-metadata-with-aws-lake-formation-part-2/
- AWS Lake Formation Preview中の新機能解説 (全5回)
  https://aws.amazon.com/jp/blogs/news/blogs-big-data-part-1-effective-data-lakes-using-aws-lake-formation-part-1-getting-started-with-governed-tables/
- AWS Lake Formation FindMatches を使用してデータセットの統合および重複の削除を実施
  https://aws.amazon.com/jp/blogs/news/integrate-and-deduplicate-datasets-using-aws-lake-formation-findmatches/