このコンテンツは公開から3年以上経過しており内容が古い可能性があります
最新情報についてはサービス別資料もしくはサービスのドキュメントをご確認ください

【AWS Black Belt Online Seminar】AWS Lake Formation

サービスカットシリーズ

アマゾン ウェブ サービス ジャパン株式会社
ソリューションアーキテクト 上原 誠
2019/10/01

自己紹介

名前: 上原 誠 (うえはら まこと)

所属: ソリューションアーキテクト

好きな AWS のサービス : AWS Glue, AWS Lake Formation

担当 : メディア, アドテクetc

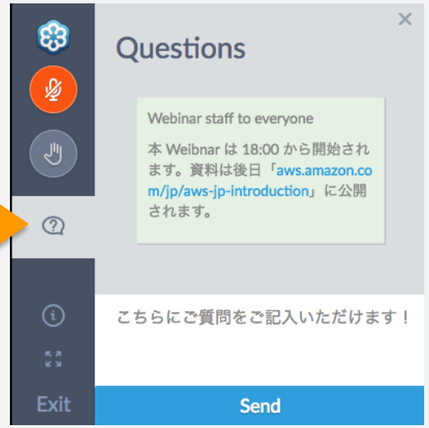
AWS Black Belt Online Seminar とは

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

質問を投げることができます！

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

① 吹き出しをクリック
② 質問を入力
③ Sendをクリック



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

内容についての注意点

- 本資料では2019年10月01日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(http://aws.amazon.com)にてご確認ください。

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

- 価格は税抜表記となっています。日本居住者のお客様が東京リージョンを使用する場合、別途消費税をご請求させていただきます。

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

Agenda

- データレイクとは
- AWS Lake Formation 登場の背景
- AWS Lake Formation とは
- AWS Lake Formation の機能
  - セキュリティ
  - データカタログ
  - ブループリント
- 落ち穂拾い
- 料金
- まとめ

データレイクとは

我々が考える以上に
データは増えている

データ	データプラットフォームに求められるもの	
成長率	使用期間	拡張性
>10倍	15	1,000
5年毎	年間	倍

* IDC, Data Age 20215: The Evolution of Data to Life-Critical Don't Focus on Big Data, Focus on the Data That's Big, April 2017.

データサイエンティスト

ビジネスユーザー

分析

アプリケーション

セキュア　リアルタイム

柔軟性　　拡張性

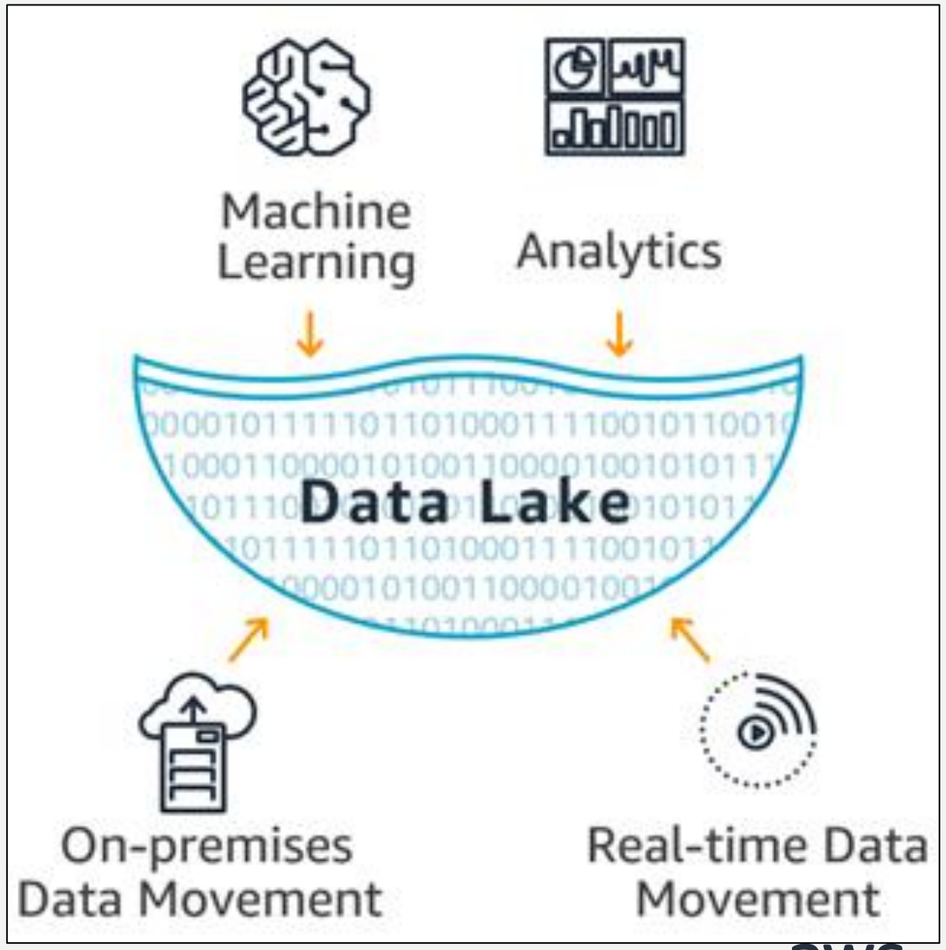
データにアクセスする人々も増えている

分析対象のデータに対する要件も増えている

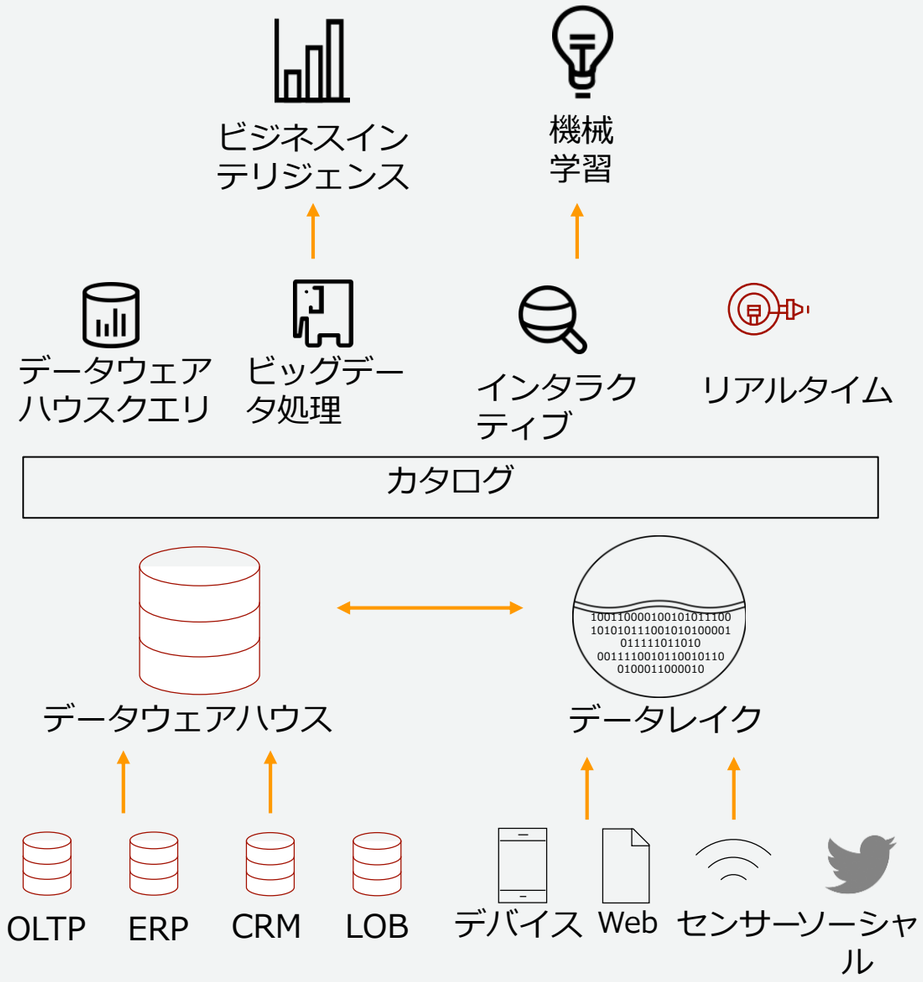
データレイクとは

AWSの公式ドキュメント
https://aws.amazon.com/jp/big-data/datalakes-and-analytics/what-is-a-data-lake/

データレイクは、規模にかかわらず、すべての構造化データと非構造化データを保存できる一元化されたリポジトリです。データをそのままの形で保存できるため、データを構造化しておく必要がありません。また、ダッシュボードや可視化、ビッグデータ処理、リアルタイム分析、機械学習など、さまざまなタイプの分析を実行し、的確な意思決定に役立てることができます。



なぜデータレイクが必要か



データレイクが提供するもの:

構造化、半構造化、非構造化データの取り扱い

ペタバイト、エクサバイトにわたる拡張性

様々な分析および機械学習ツールとの連携

データの移動を伴わずにデータを処理

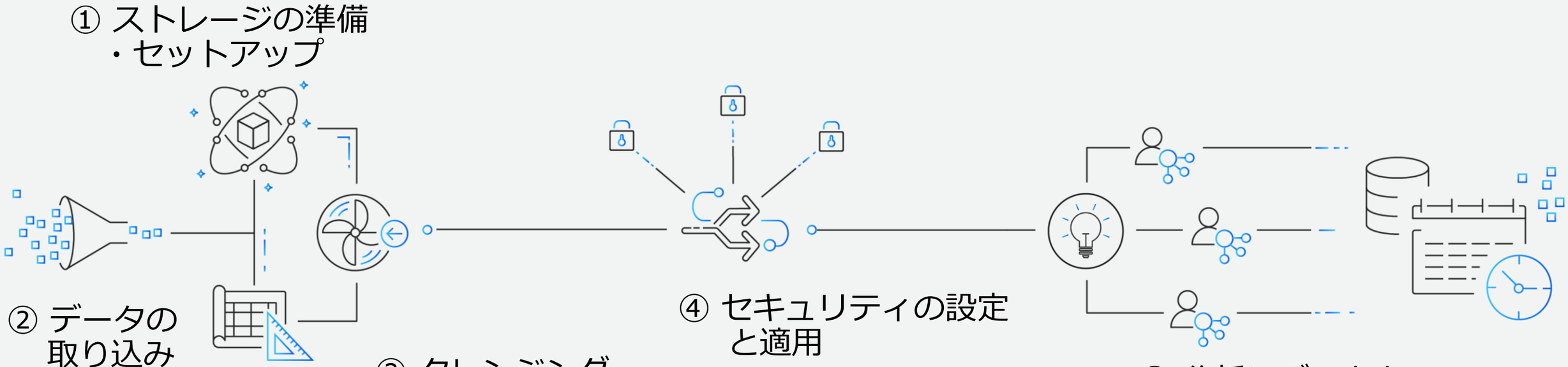
低コストなデータの保存と分析

AWS 上にデータレイク & 分析環境を構築されたお客様

aws

AWS Lake Formation 登場の背景

データレイク構築時の代表的な作業フロー



多くの作業ステップが必要

データレイクの構築には数ヶ月単位の時間が必要

AWS Lake Formation

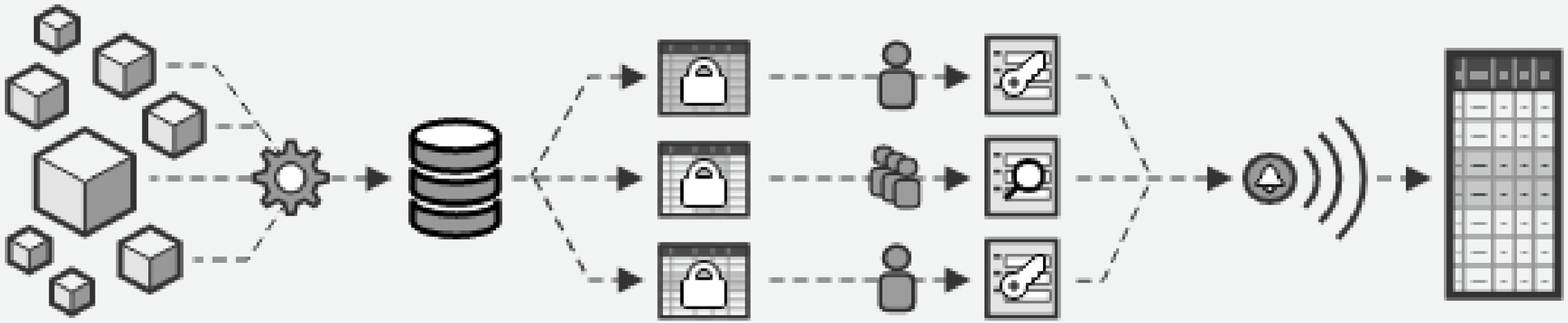
セキュアなデータレイクを短期間で構築

データの認識、取り込み、
整形、変換

複数サービス間でセキュリ
ティポリシーを適用

新たなインサイトを
提供するカタログ管理

AWS Lake Formation 概要



データ取込みと構造化

- 自動的にデータ取込み、成形、暗号化して、既存のAmazon S3バケットに登録

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

協調＆利用

- メタデータカタログを利用した検索と定義確認
- 全てのアクセスはIAMポリシーによりチェック
- 新しいデータが取込まれたり、ツールが変更されてもポリシーにより保護可能

監視＆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

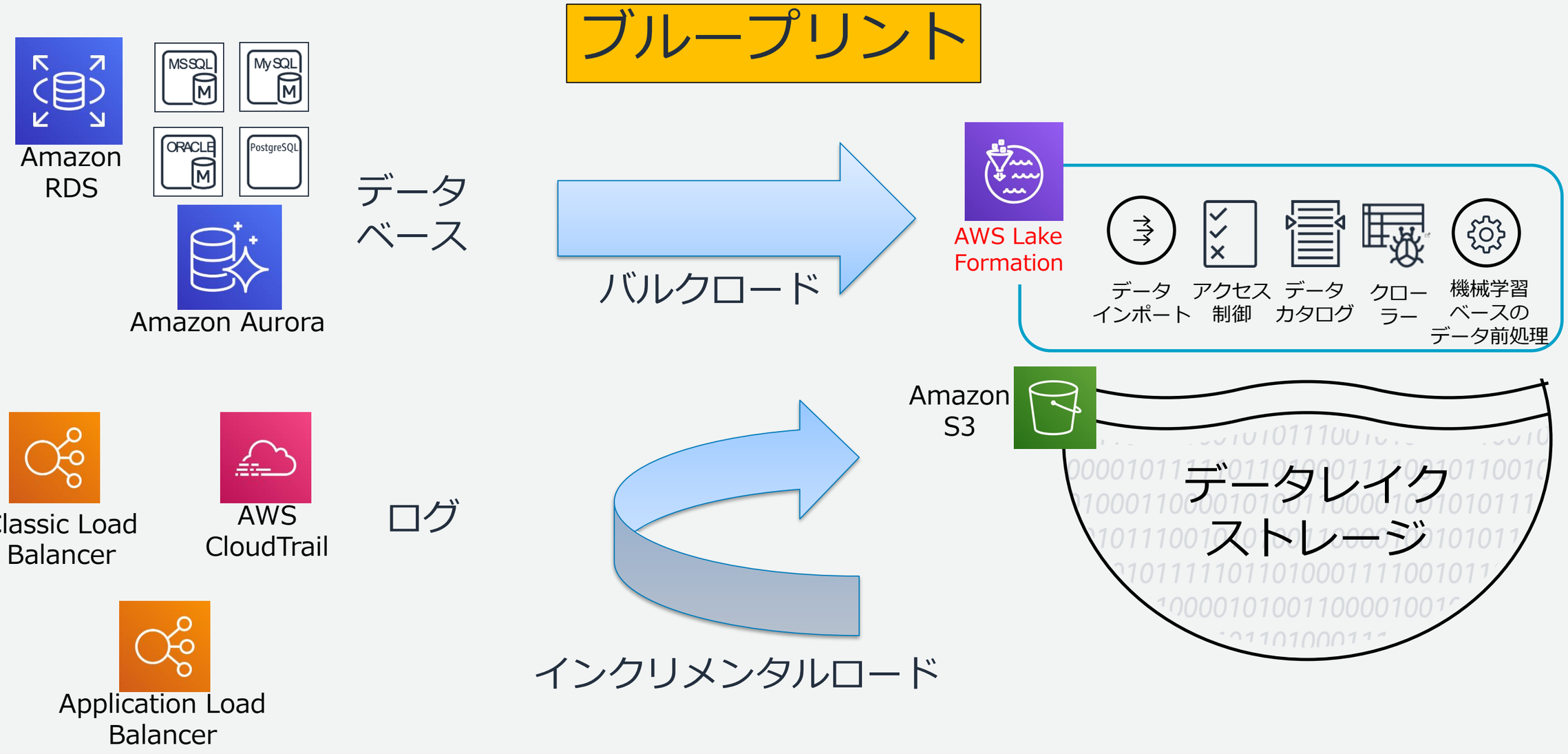
AWS Lake Formation の機能

Lake Formation の各コンポーネント

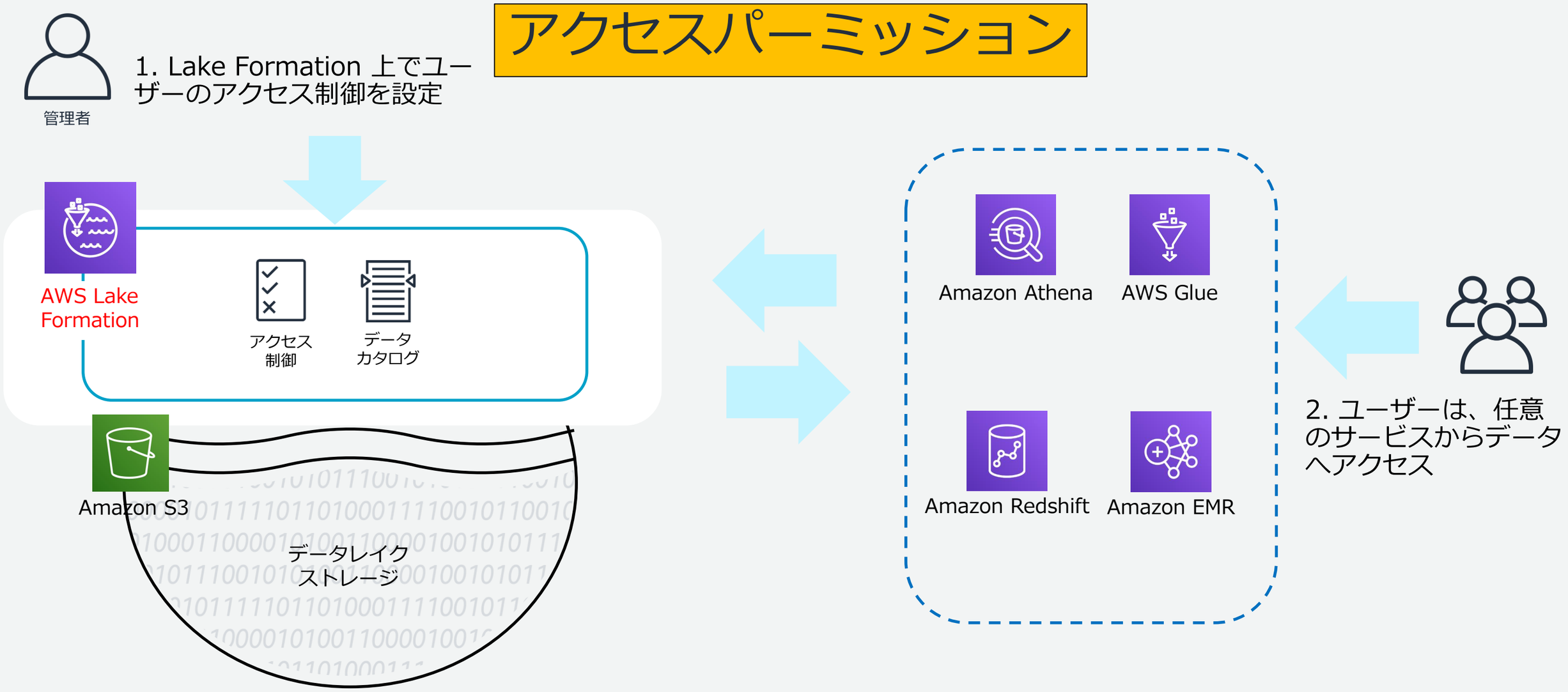
AWS Lake Formation

- データ取込みと構造化 → **ブループリント**：汎用的なデータ取り込みテンプレートを使い、自動的なデータ取り込みと構造化を実現
- セキュリティ＆コントロール → **パーミッション**：SQL ライクな Grant/Revoke でシンプルなアクセス制御を実現
- 協調＆利用 → **データカタログ**：スキーマやロケーションなどのデータのメタ情報を管理し、ファセット検索で探したいデータセットを探しやすく
- 監視＆監査 → **ロギング**：コンソールによる直近のアクティビティの詳細を確認

AWS Lake Formation – 主要機能

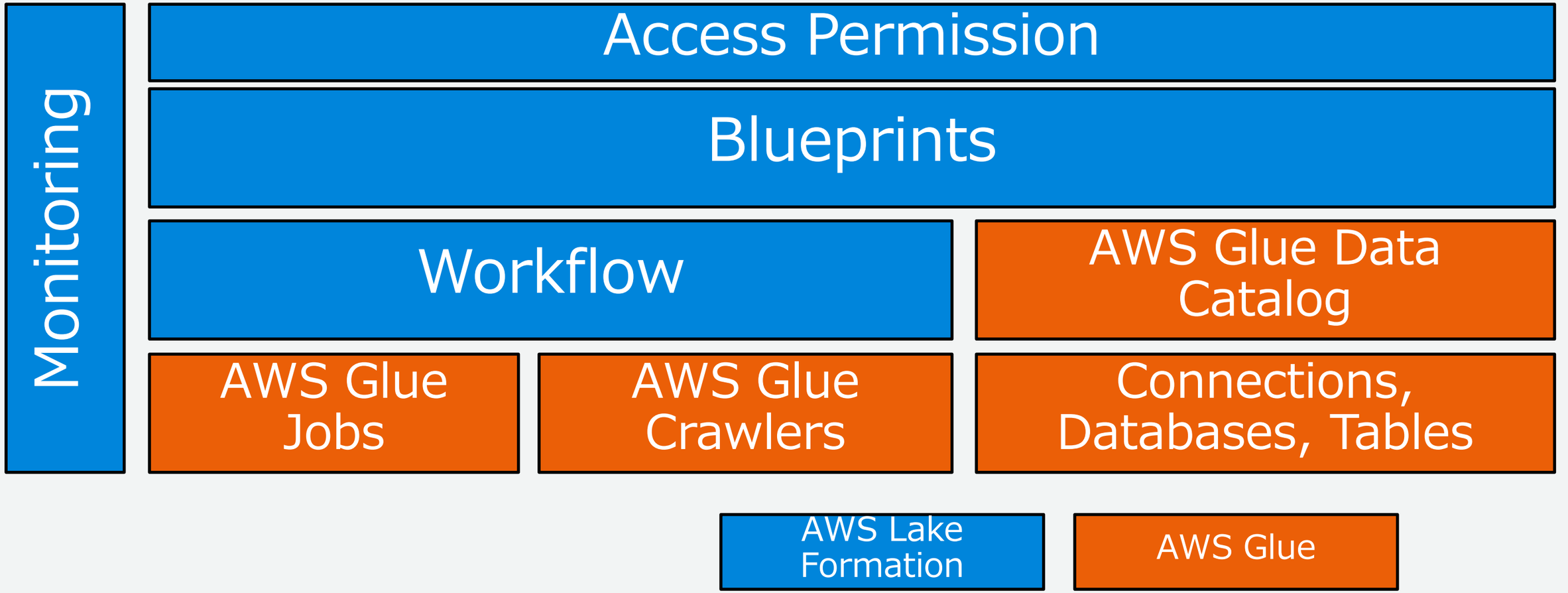


AWS Lake Formation – 主要機能



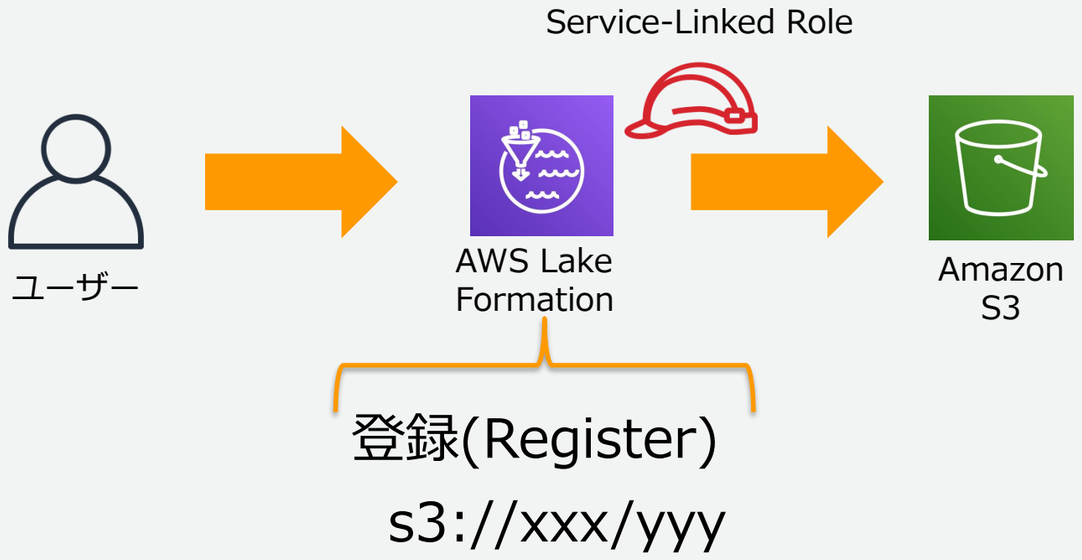
Lake Formation と AWS Glue の関係

Lake Formation は Glue の拡張機能と言え、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などでより便利に Glue の機能を使えるようになっています。



データソースとして S3 を Lake Formation に登録

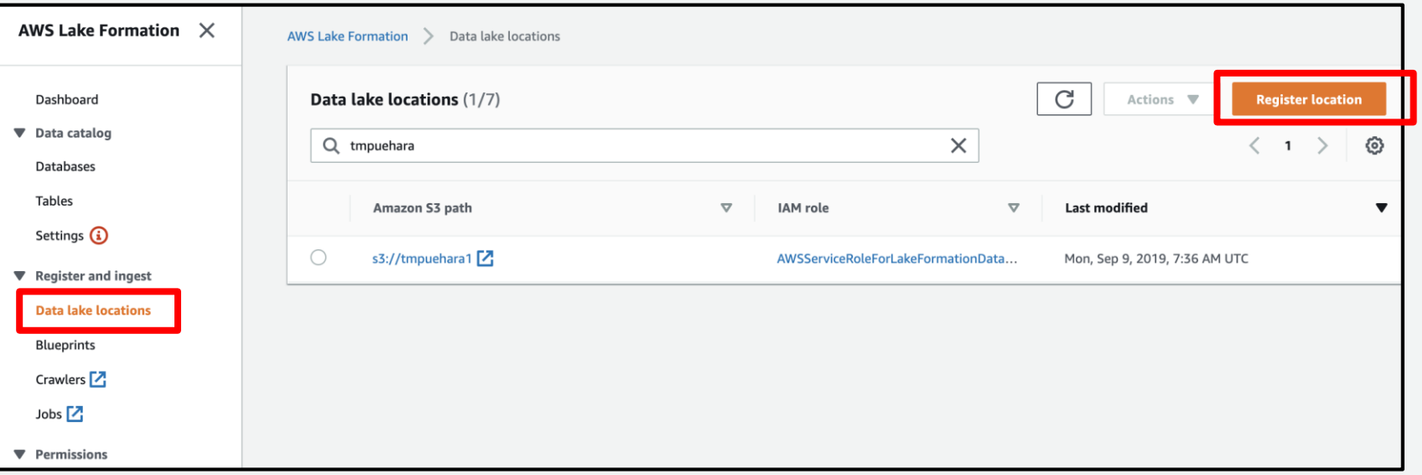
- S3 は Lake Formation のストレージ層
- データソースとして、既存 S3 パスを Lake Formation に登録
- Lake Formation と信頼関係があり S3 パスへの読み取り/書き込み権限を持つ IAM ロールを選択
- S3 パスに格納されたデータに、Lake Formation からアクセスが可能に



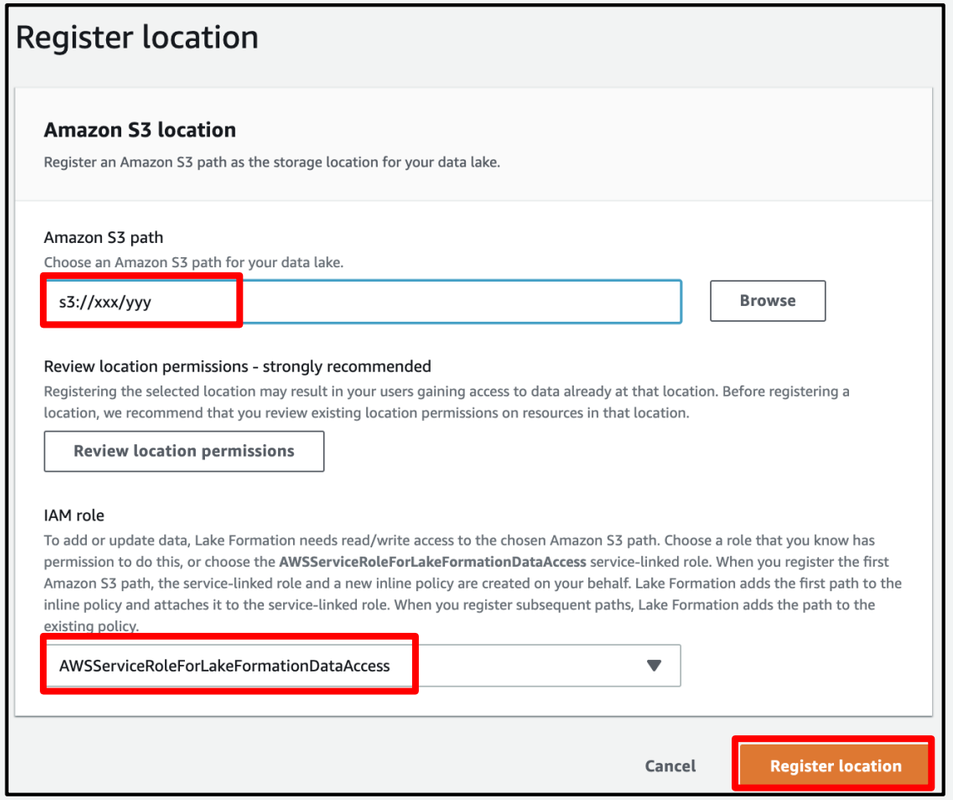
登録することではじめて Lake Formation での権限制御の対象となります

データソースとして S3 を Lake Formation に登録

設定は、Data Lake Location をクリックし、
[Register location]



S3 パスと Service-Linked Role を指定する



セキュリティ

Lake Formation 用語解説

- ✓ Data Lake Location

  Lake Formation でアクセスパーミッションを行う対象の S3 パスを登録します。

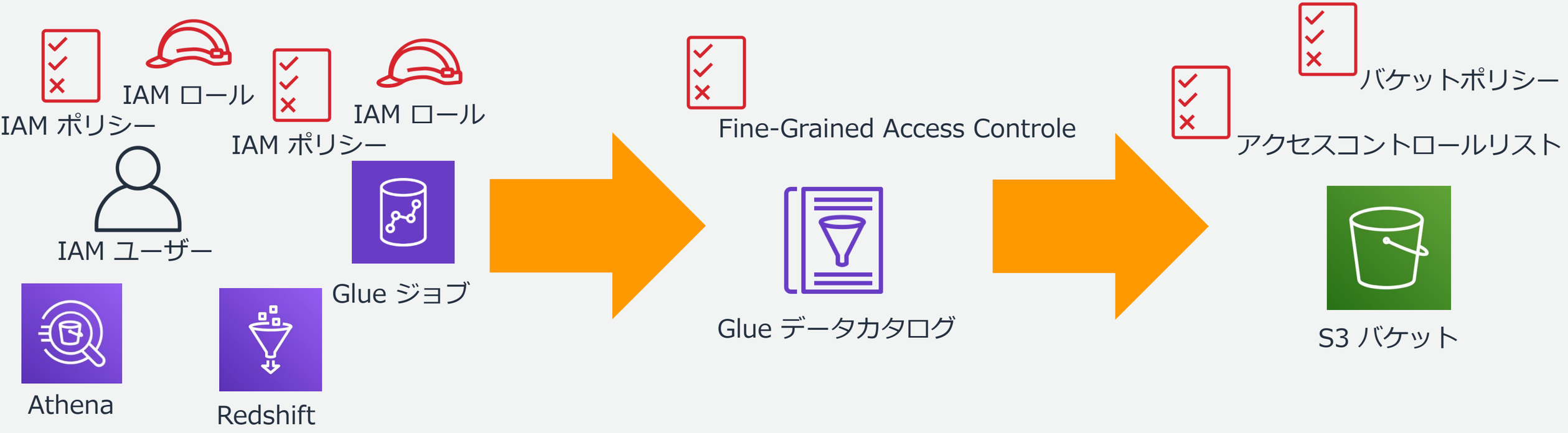
- ✓ Data Location

  プリンシパル※に対して、Data Lake Location で登録した S3 パスにアクセス許可を与えます。これによりデータベースやテーブルの作成ができるようになります。

※注釈：プリンシパルは IAM ユーザーや IAM ロールのこと

Lake Formation 以前のセキュリティ設定

これまで IAM と S3、Glue データカタログを組み合わせて実現していたデータレイクのアクセスパーミッションでは、複数の異なる箇所に別々にポリシーを登録する必要がありました。例えば、IAM ユーザーに特定のバケットへの S3 の権限や特定のデータベースやテーブルへの Glue の権限を与え、S3 バケットポリシーでも特定バケットに対して権限を与えていました。



Lake Formation 以前のセキュ

{"Version": "2012-10-17","Statement": [{"Sid": ”statement1","Effect": "Allow","Action": ["s3:PutAnalyticsConfiguration","s3:GetObjectVersionTagging","s3:GetBucketCORS","s3:GetBucketLocation","s3:GetObjectVersion","s3:ReplicateDelete”,…….],"Resource": "arn:aws:s3:::testbucket"},{"Sid": ”statement2","Effect": "Allow","Action": ["s3:GetAccountPublicAccessBlock","s3:ListAllMyBuckets","s3:ListJobs","s3:CreateJob","s3:HeadBucket”,……],"Resource": "*"}]}

{ "Version": "2012-10-17", "Statement": [{ "Sid": "statement1", "Effect": "Allow", "Principal": { "AWS": "arn:aws:iam::xxxxxxxxxxxx:user/taro" }, "Action": ["s3:GetBucketLocation", "s3:ListBucket"], "Resource": ["arn:aws:s3:::testbucket"] }, { "Sid": "statement2", "Effect": "Allow", "Principal": { "AWS": "arn:aws:iam::xxxxxxxxxxx:user/taro" }, "Action": ["s3:GetObject"], "Resource": ["arn:aws:s3:::testbucket/*"] }] }

グを統
異なる
定のバ
、S3 バケットポリシーでも特
ットに
対して権限
えていました。


IAM ポリシー


IAM ロール


IAM ポリシー


IAM ロール


IAM ユーザー


Glue ジョブ


Fine-Grained Access Controle


e データカタログ


アクセスコントロールリスト


バケットポリシー


S3 バケット

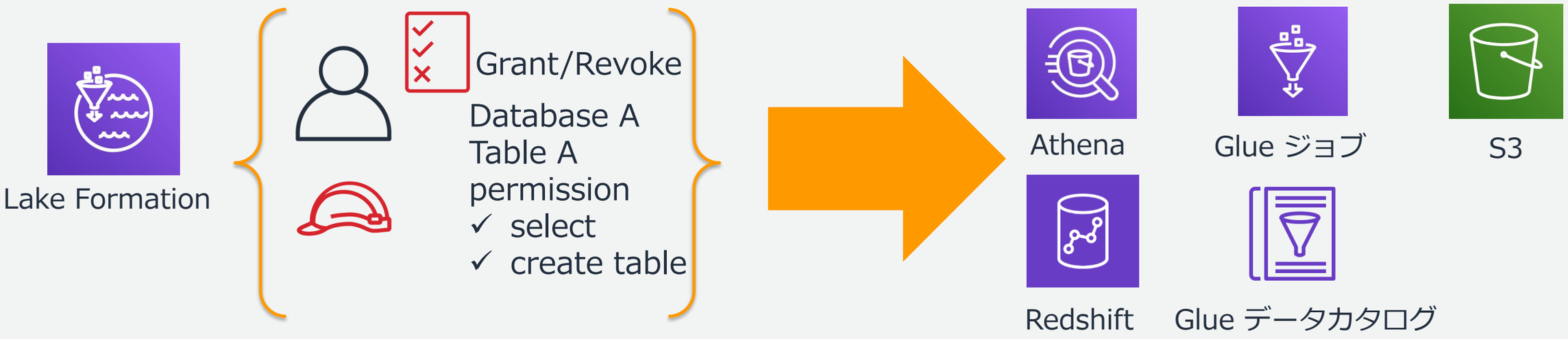

Athena

{ "Version" : "2012-10-17", "Statement" : [{ "Effect" : "Allow", "Principal" : { "AWS" :"arn:aws:iam::xxxxxxxxxxxx:user/taro" }, "Action" : "glue:*", "Resource" : ["arn:aws:glue:ap-northeast-1:xxxxxxxxxxxx:catalog", "arn:aws:glue:ap-northeast-1:xxxxxxxxxxxx:database/dba", "arn:aws:glue:ap-northeast-1:xxxxxxxxxxxx:table/dba/tablea"] }] }

Lake Formation のセキュリティパーミッション

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 IAM ユーザーや IAM ロールに対して、SQL ライクでシンプルな 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。



Lake Formation におけるセキュリティパーミッション

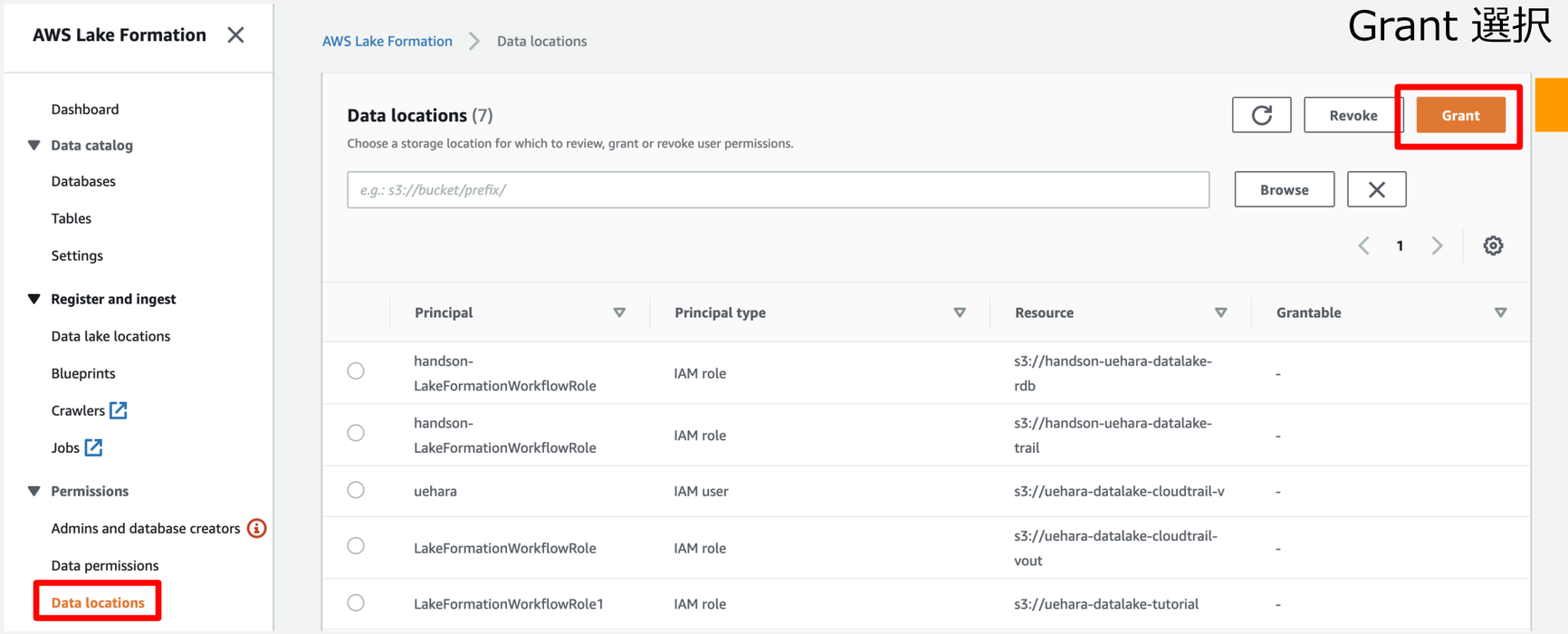
Lake Formation では、データロケーションのアクセス許可、データカタログのアクセス許可、データアクセス許可の3種類に明示的に許可を与えます。さらに明示的な許可またはデータベースやテーブル作成の結果として暗黙的なアクセス許可が与えられます。

- データロケーションのアクセス許可
- データカタログのアクセス許可
- データアクセス許可
- 暗黙的なアクセス許可

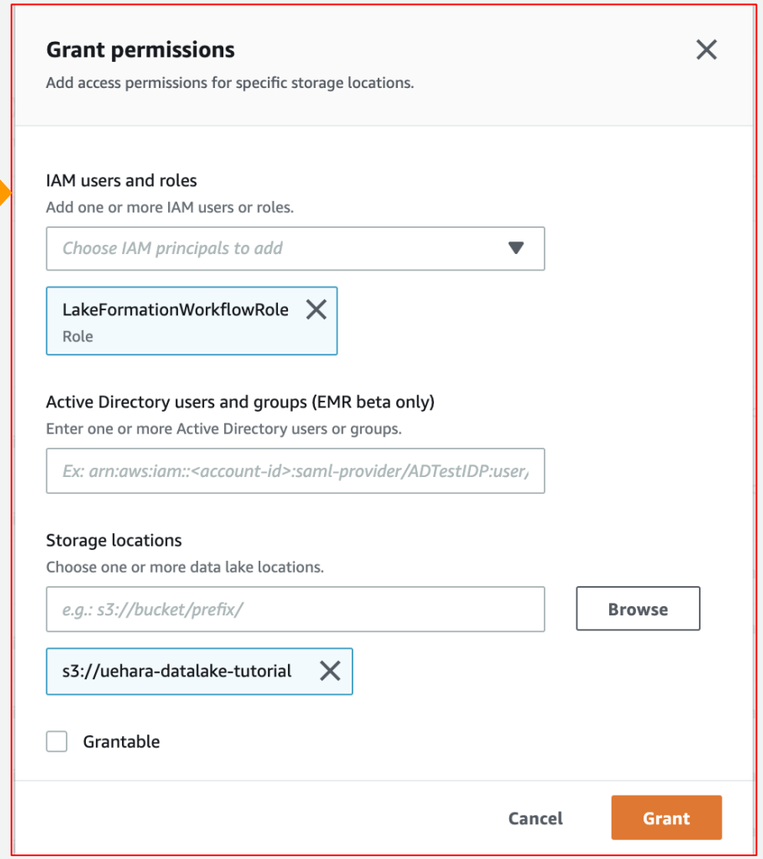
データロケーションのアクセス許可

データロケーションは、データが保存される S3 パスです。登録された S3 のある場所にデータベースまたはテーブルを作成するために、プリンシパルにその場所に対するアクセス許可を与えます。

※ブループリントを使う場合は、S3 パスと、Lake Formation がブループリントにより作成されたワークフローで対象のロケーションに対してテーブルを作成するための IAM ロール を指定します。



Grant 選択

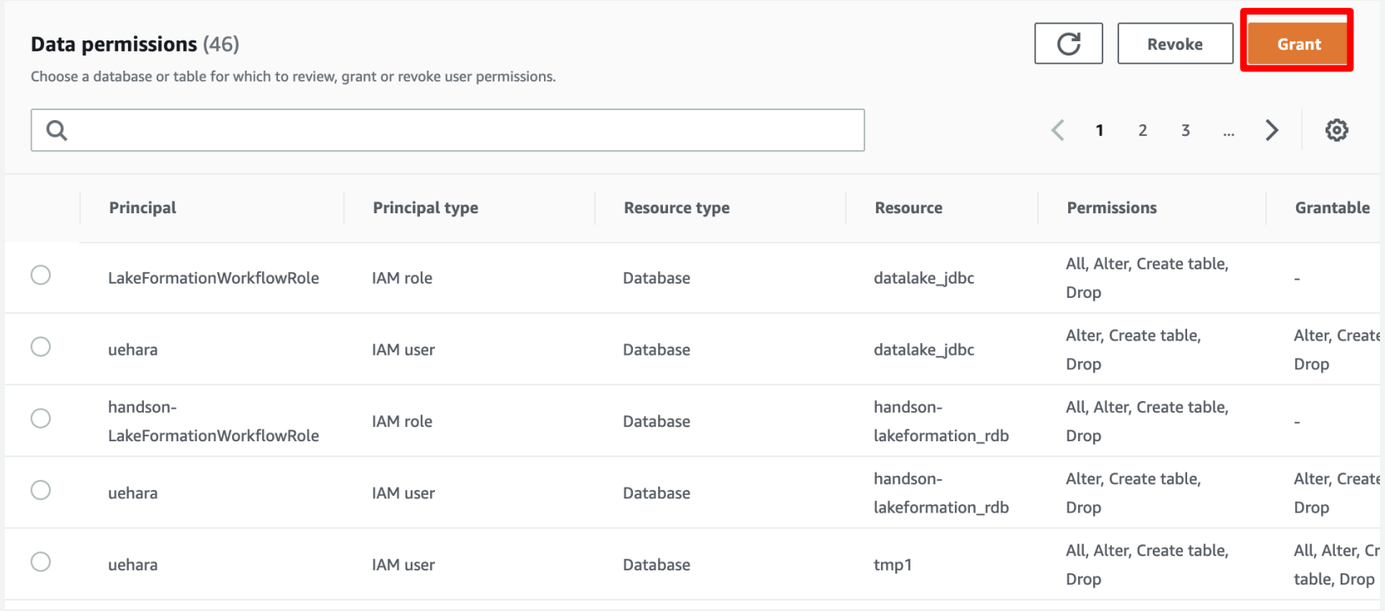


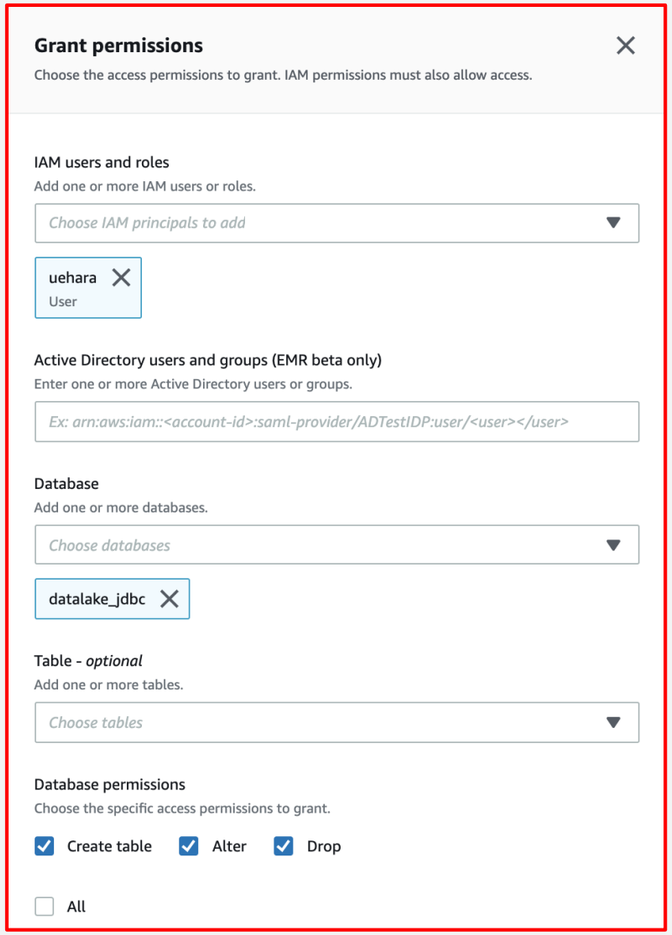
データカタログのアクセス許可

データカタログには、基となるデータに関するメタデータが格納されます。メタデータはデータベースとテーブルとして編成されています。テーブルは S3 に保存されているデータの場所を指します。データベースはテーブルの集合です。データカタログのアクセス許可は、データベースとテーブルを作成、編集、および削除する権限をシンプルな Grant/Revoke で与えます。
(例 CREATE_DATABASE, CREATE_TABLE)

CREATE_TABLEの例

Grant 選択

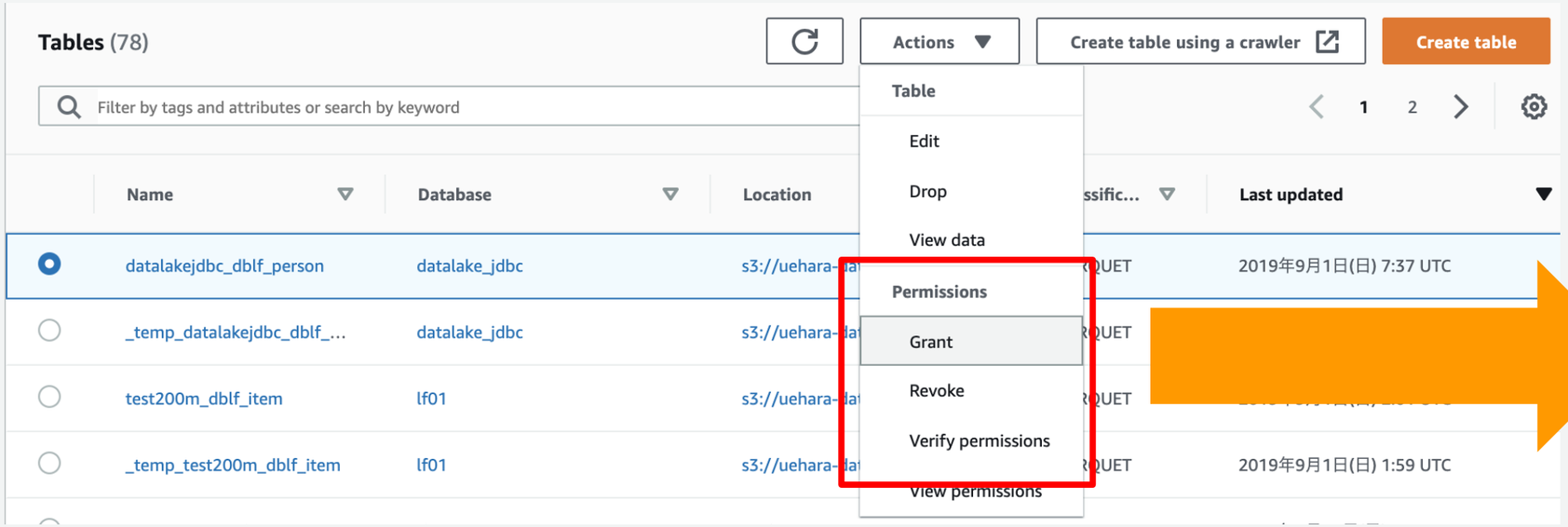




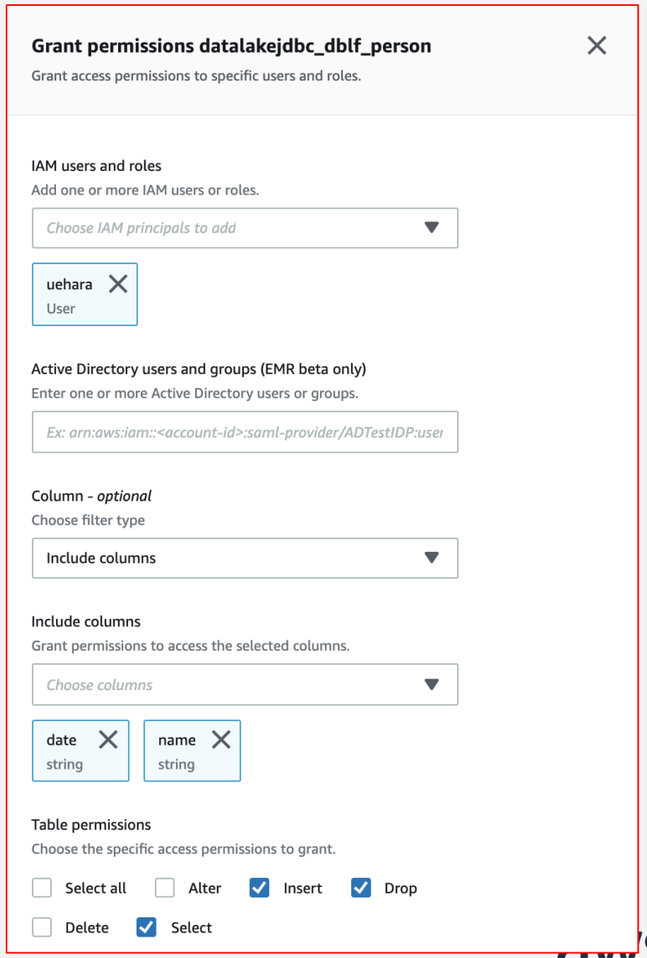
データアクセス許可

テーブルの基になるデータを読み書きするために、プリンシパルにテーブルに対するデータアクセス許可をシンプルな Grant/Revoke で与えます。

(例 SELECT, INSERT)



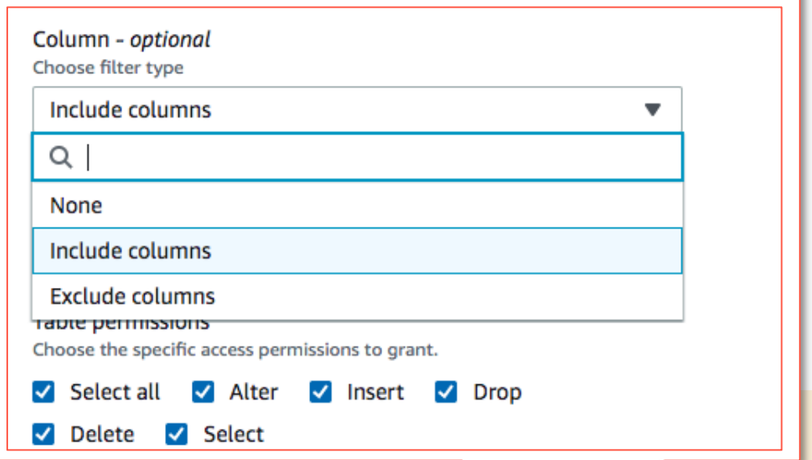
Grant 選択



データアクセス許可

テーブルおよび列レベルの権限付与

列レベルのアクセス制御の指定



ユーザー 1

一部の列のみ
アクセス可能

Column name	Data type
marketplace	string
customer_id	bigint
review_id	string
product_id	string
product_parent	bigint
product_title	string
star_rating	string
helpful_votes	bigint
total_votes	bigint
vine	string
verified_purchase	string
review_headline	string
review_body	string
review_date	string
product_category	string

ユーザー 2

全ての列に
アクセス可能

暗黙的なアクセス許可

データカタログ権限の明示的な付与は、追加の暗黙的な権限の付与をする場合があります。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も行われます。

- 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

aws

データレイク管理者の暗黙的なアクセス許可

Lake Formation では、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。

アクセス許可の一覧

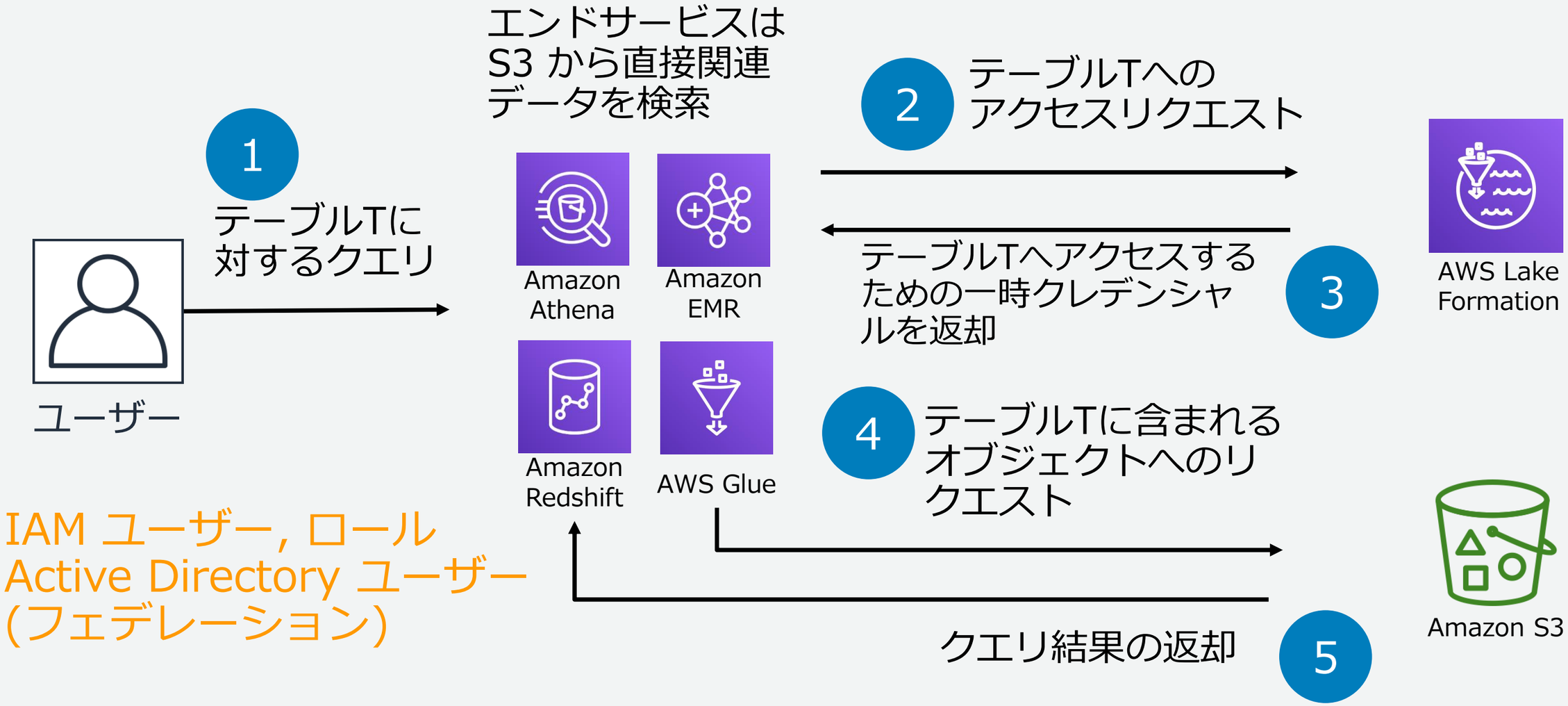
許可の種類	CREATE DATABASE, CREATE TABLE (登録ロケーションを指定) ※1	CREATE DATABASE, CREATE TABLE ※1	SELECT, INSERT	主に作成したリソースに対する権限
データロケーションのアクセス許可	●			
データカタログのアクセス許可	●	●		▲※2
データアクセス許可			●	▲※2
暗黙のアクセス許可				●

※1 CREATE_DATABASE や CREATE_TABLE は留意すべきケースがいくつかあるので詳細は以下のリンクを確認ください。

※2 特定のデータベースや特定のテーブルへのアクセス権限

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

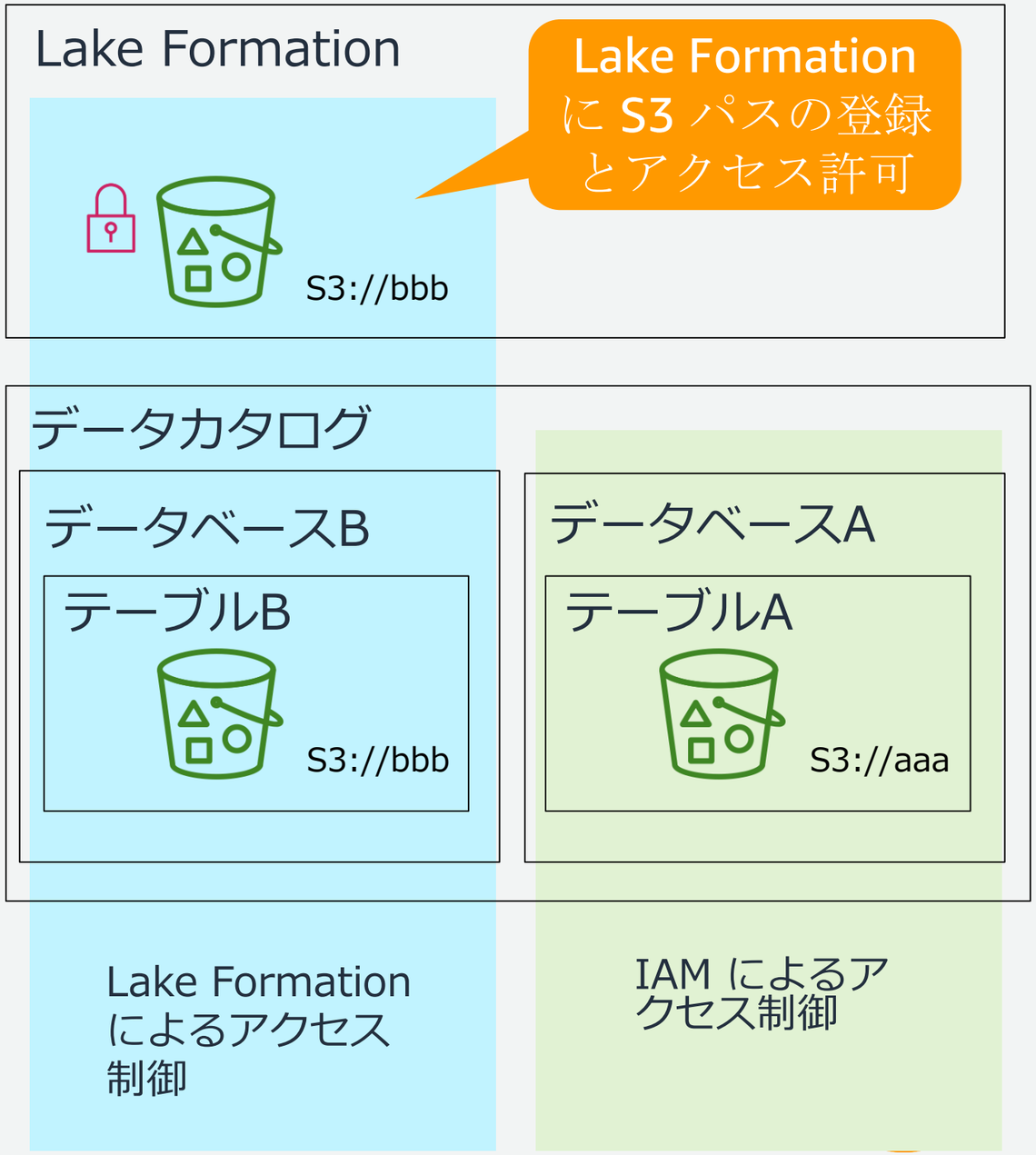
アクセス時の連携イメージ



従来の Glue によるアクセス制御との違い

データベースAは Lake Formation に登録されていない S3 パスのテーブルがあるのでデータカタログのアクセス許可と S3 アクセス許可などの IAM ポリシーにより制御されます。これは従来どおりです。

データベースBは Lake Formation に登録とアクセス許可されている S3 パスのテーブルがあるので Lake Formation の権限により制御されます。



Lake Formation のペルソナ※と権限のリファレンス

ペルソナ	説明
IAM 管理者(AWS管理者)	少なくても IAM と S3 へのフル権限を持つ。または AdministratorAccess のIAM ポリシーを持つ。
データレイク管理者	データカタログへアクセスし、データベースを作成し、他のユーザーに Lake Formation による権限付与が行える。 IAM 管理者より持っている IAM アクセス許可は少ないが、データレイクを管理するのに十分な権限を持つ。
データエンジニア (オプション)	所有するデータベースへの権限を持ち、ブループリントからワークフローを作成し、ワークフローを実行するのに十分な権限を持つ。
データ分析者 (オプション)	Athena などを使用して、データレイクに対してクエリを実行できるユーザー。クエリを実行するための権限のみを持つ。
ワークフローロール	ユーザーに代わってワークフローを実行するロール。 ブループリントからワークフローを作成するときに、このロールを指定します。

※ペルソナ：データレイクを使うユーザータイプ

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

Lake Formation のペルソナ※と権限のリファレンス

ペルソナ	IAM, S3 全権限	Lake Formation の全権限	Lake Formation の一部の権限	Pass Role 権限	Athena 等の分析サービスの権限
IAM 管理者(AWS 管理者)	●				
データレイク管理者		●		●	
データエンジニア (オプション)			●	●	
データ分析者 (オプション)			●		●
ワークフローロール			●	●	

※ペルソナ：データレイクを使うユーザータイプ

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

データレイク管理者の権限

ポリシータイプ	ポリシー
AWSマネージドルール	AWSGlueConsoleFullAccess

インラインポリシー
(basic)
	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Effect": "Allow",
	  "Action": [
	  "lakeformation:*",
	  "cloudtrail:DescribeTrails",
	  "cloudtrail:LookupEvents",
	  "iam:PutRolePolicy",
	  "iam:CreateServiceLinkedRole"
	  ],
	  "Resource": "*"
	  },
	  {
	  "Effect": "Deny",
	  "Action": [
	  "lakeformation:PutDataLakeSettings"
	  ],
	  "Resource": "*"
	  }
	  ]
	}</pre>

インラインポリシー
(pass role)
	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Sid": "PassRolePermissions",
	  "Effect": "Allow",
	  "Action": [
	  "iam:PassRole"
	  ],
	  "Resource": [
	  "arn:aws:iam::*account-id*:role/*workflow_role*"</pre>

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

データエンジニアの権限

ポリシータイプ	ポリシー
AWSマネージドルール	AWSGlueServiceRole

インラインポリシー
(data access)
	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Sid": "Lakeformation",
	  "Effect": "Allow",
	  "Action": [
	  "lakeformation:GetDataAccess",
	  "lakeformation:GrantPermissions"
	  ],
	  "Resource": "*"
	  }
	  ]
	}</pre>

インラインポリシー
(pass role)
	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Sid": "PassRolePermissions",
	  "Effect": "Allow",
	  "Action": [
	  "iam:PassRole"
	  ],
	  "Resource": [
	  "arn:aws:iam::*account-id*:role/*workflow_role*"
	  ]
	  }
	  ]
	}</pre>

インラインポリシー
(for ingesting data outside the data lake, for example Cloudtrail logs)
	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Effect": "Allow",
	  "Action": ["s3:GetObject", "s3:ListBucket"],
	  "Resource": ["arn:aws:s3:::*your-s3-cloudtrail-bucket*/*"]
	  }
	  ]
	}</pre> https://docs.aws.amazon.com/ja_jp/lake-formation/latest/dg/permissions-reference.html

データ分析者の権限

ポリシータイプ	ポリシー
AWSマネージドルール	AmazonAthenaFullAccess

インラインポリシー (basic)	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Effect": "Allow",
	  "Action": [
	  "lakeformation:GetDataAccess",
	  "glue:GetTable",
	  "glue:GetTables",
	  "glue:SearchTables",
	  "glue:GetDatabase",
	  "glue:GetDatabases",
	  "glue:GetPartitions"
	  ],
	  "Resource": "*"
	  }
	  ]
	}</pre>

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

ワークフローロールの権限

ポリシータイプ	ポリシー
AWSマネージドルール	AWSGlueServiceRole

インラインポリシー
(data access)
	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Sid": "Lakeformation",
	  "Effect": "Allow",
	  "Action": [
	  "lakeformation:GetDataAccess",
	  "lakeformation:GrantPermissions"
	  ],
	  "Resource": "*"
	  }
	  ]
	}</pre>

インラインポリシー
(pass role)
	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Sid": "PassRolePermissions",
	  "Effect": "Allow",
	  "Action": [
	  "iam:PassRole"
	  ],
	  "Resource": [
	  "arn:aws:iam::*account-id*:role/*workflow_role*"
	  ]
	  }
	  ]
	}</pre>

インラインポリシー
(for ingesting the data lake, for example CloudTrail logs)
	<pre>{
	  "Version": "2012-10-17",
	  "Statement": [
	  {
	  "Effect": "Allow",
	  "Action": ["s3:GetObject", "s3:ListBucket"],
	  "Resource": ["arn:aws:s3:::*your-s3-cloudtrail-bucket*/*"]
	  }
	  ]
	}</pre> https://docs.aws.amazon.com/ja_jp/lake-formation/latest/dg/permissions-reference.html

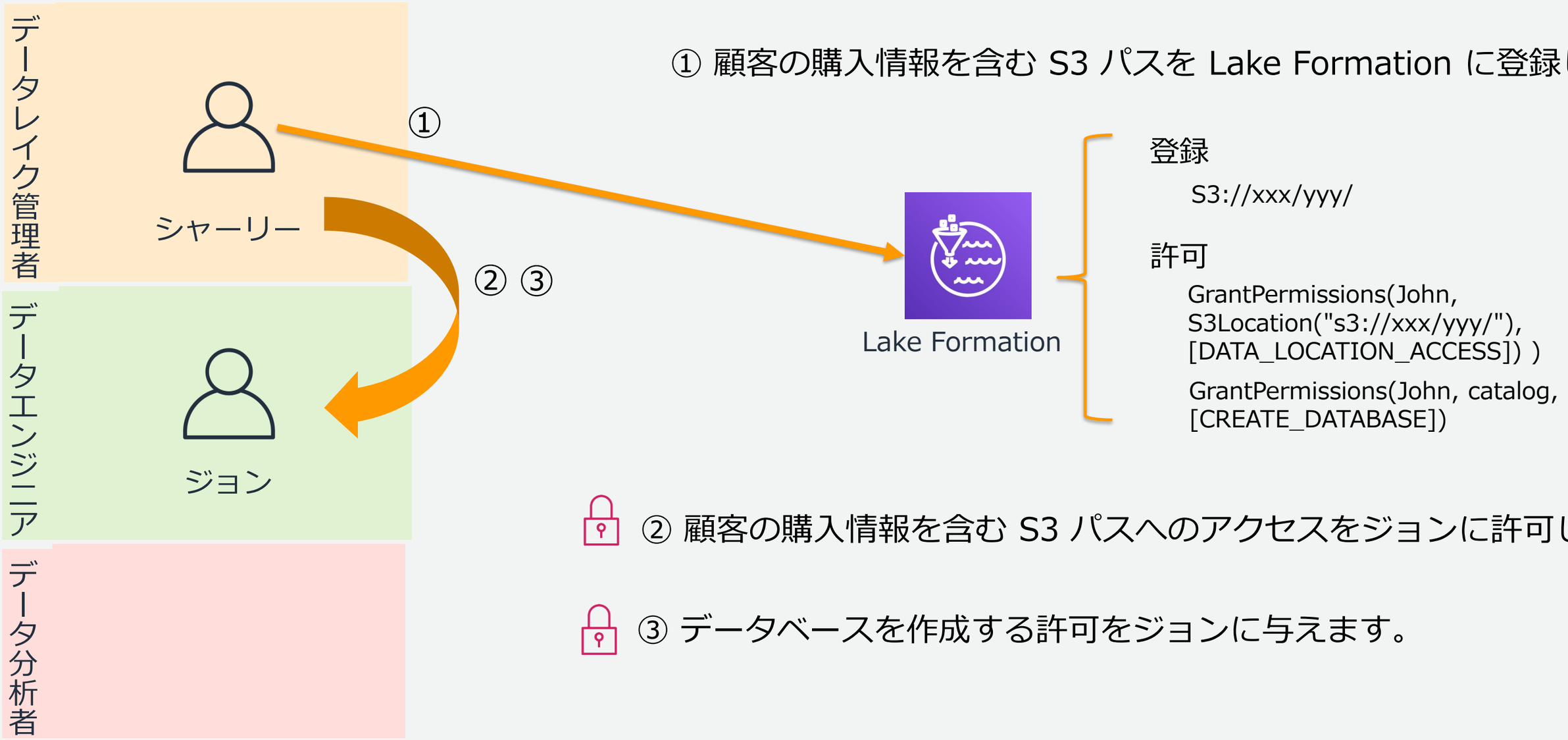
許可を与えるシナリオの例

シャーリーはデータレイク管理者です。彼女は、会社である AnyCompany のデータレイクをセットアップしたいと考えています。現在、すべてのデータは S3 に保存されています。ジョンはマーケティングマネージャーで、顧客の購入情報(s3://customerPurchases)への書き込みアクセスが必要です。この夏マーケティングアナリストのディエゴが加わります。ジョンには、シャーリーが関与することなくデータに対してクエリを実行するためのディエゴへのアクセスを許可することが必要です

ペルソナ	担当者
データレイク管理者	シャーリー
データエンジニア	ジョン
データ分析者	ディエゴ

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

許可を与えるシナリオの例



① 顧客の購入情報を含む S3 パスを Lake Formation に登録します。

② 顧客の購入情報を含む S3 パスへのアクセスをジョンに許可します。

③ データベースを作成する許可をジョンに与えます。

許可を与えるシナリオの例

データレイク管理者

データエンジニア

データ分析者

ジョン

ディエゴ

① ②

③

① データベース John_DB を作成します。ジョンはデータベースを作成したため、自動的にそのデータベースに対する CREATE_TABLE 権限を持ちます(暗黙的な許可)。

Lake Formation

データベース
John_DB

テーブル
John_Table

許可
GrantPermissions(Diego, John_Table, [SELECT])

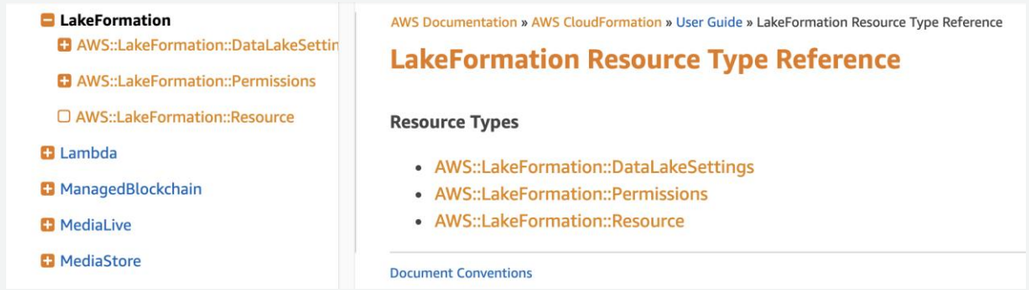
② s3://xxx/yyy/ を指すテーブル John_Table を作成します。ジョンはテーブルを作成したため、このテーブルに対する全ての権限を持ちます(暗黙的な許可)。

③ アナリストのディエゴに、テーブル John_Table にアクセスする許可を与えます。

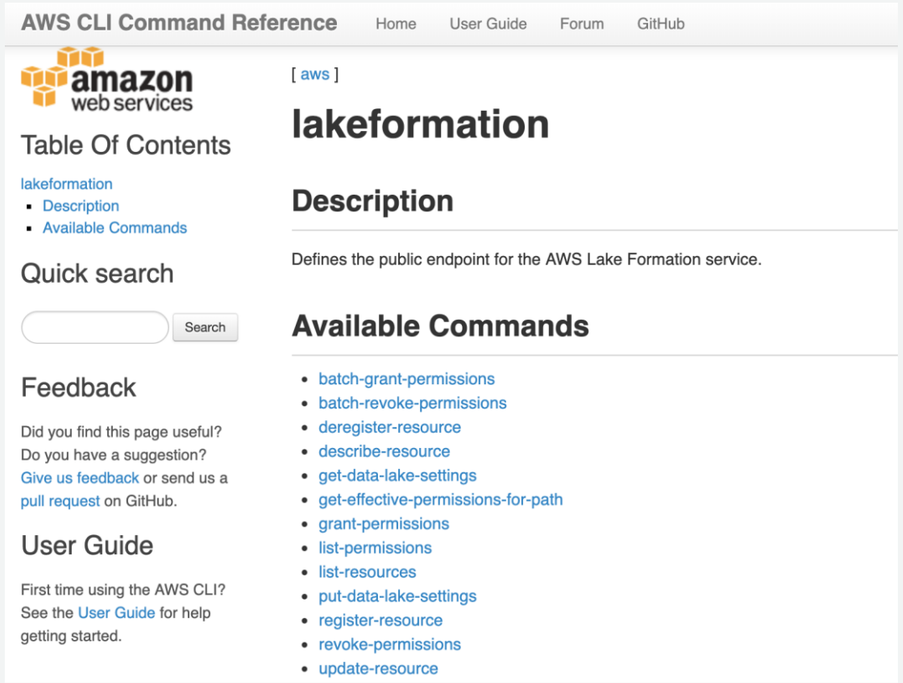
自動化

CloudFormation や AWS CLI で Lake Formation のアクセス許可操作を自動化できます。

- CloudFormation



- AWS CLI



CloudFormation では以下のようなテンプレートでスタックを作成することで、Lake Formation の管理者設定が可能です。

```
{
   "AWSTemplateFormatVersion": "2010-09-09",
   "Resources": {
      "LFDLS4VYBJ": {
         "Type": "AWS::LakeFormation::DataLakeSettings",
         "Properties": {
            "Admins":[{
            "DataLakePrincipalIdentifier": "arn:aws:iam::xxxxxxxxxxxx:user/uehara"
          }]
         }
      }
   }
}
```


https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/AWS_LakeFormation.html

自動化

- AWS CLI では以下のように管理者設定の反映が可能です。

```
$ aws lakeformation put-data-lake-settings --cli-input-json file://lfadmin.json
```

- AWS CLI では以下のように管理者設定の内容をgetすることが可能です。

```
$ aws lakeformation get-data-lake-settings
{
   "DataLakeSettings": {
      "DataLakeAdmins": [
         {
            "DataLakePrincipalIdentifier": "arn:aws:iam::xxxxxxxxxxxx:user/uehara"
         }
      ],
      "CreateDatabaseDefaultPermissions": [
         {
            "Principal": {
               "DataLakePrincipalIdentifier": "EVERYONE"
            },
            "Permissions": [
               "ALL"
            ]
         }
      ],
      "CreateTableDefaultPermissions": [
         {
            "Principal": {
               "DataLakePrincipalIdentifier": "EVERYONE"
            },
            "Permissions": [
               "ALL"
            ]
         }

}
```

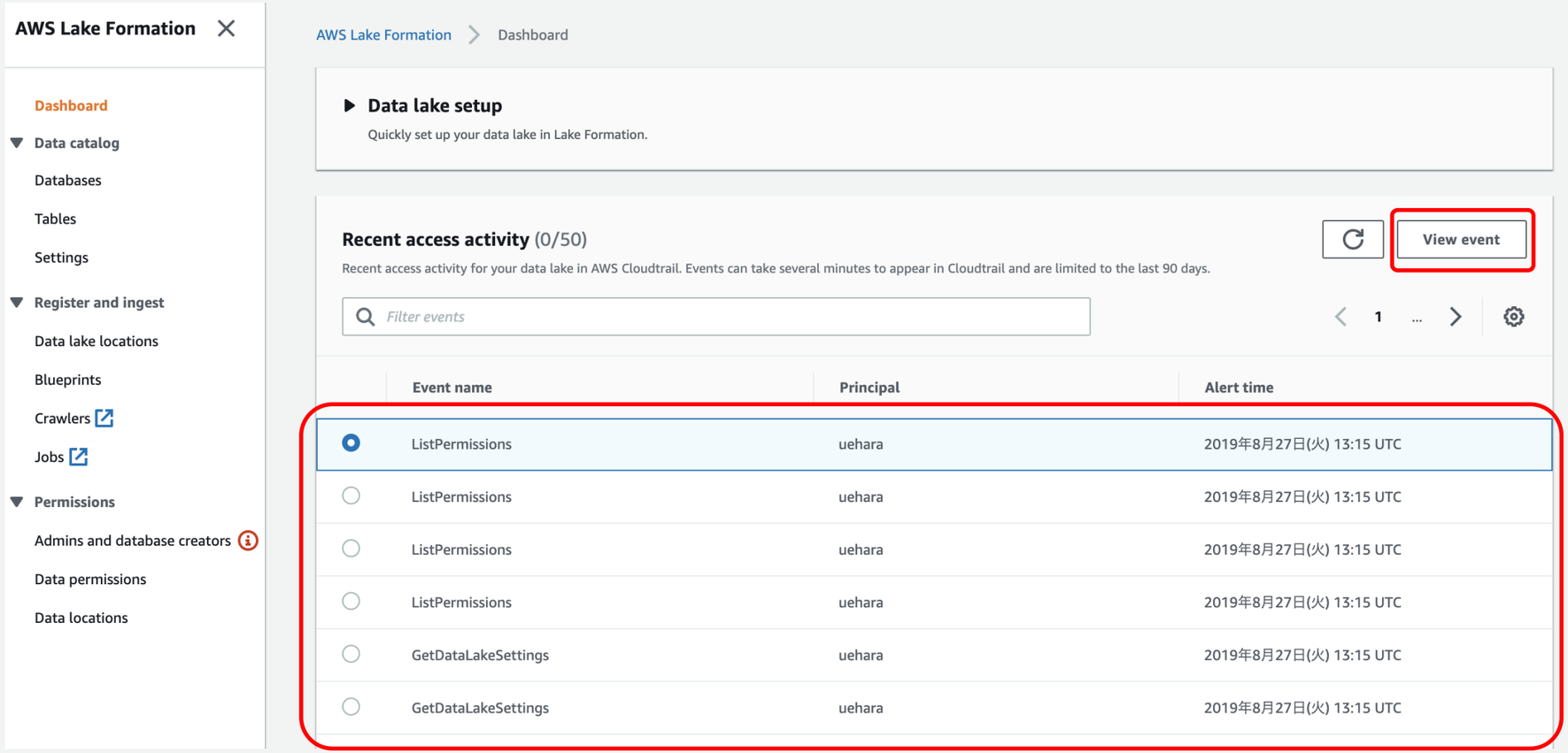
lfadmin.json

```
{
   "DataLakeSettings": {
      "DataLakeAdmins": [
         {
            "DataLakePrincipalIdentifier":
"arn:aws:iam::xxxxxxxxxxxx:user/uehara"
         }
      ],
      "CreateDatabaseDefaultPermissions": [
         {
            "Principal": {
               "DataLakePrincipalIdentifier": "EVERYONE"
            },
            "Permissions": [
               "ALL"
            ]
         }
      ],
      "CreateTableDefaultPermissions": [
         {
            "Principal": {
               "DataLakePrincipalIdentifier": "EVERYONE"
            },
            "Permissions": [
               "ALL"
            ]
         }
      ]
   }
}
```

https://docs.aws.amazon.com/cli/latest/reference/lakeformation/index.html#cli-aws-lakeformation

ニアリアルタイム監査とモニタリング

CloudTrail を有効化しておくことで、Lake Formation のコンソールでデータアクセスを監査できます。[View Event] でどのユーザーまたはロールがどのデータにどのサービスでアクセスしたかなどの詳細を確認できます。

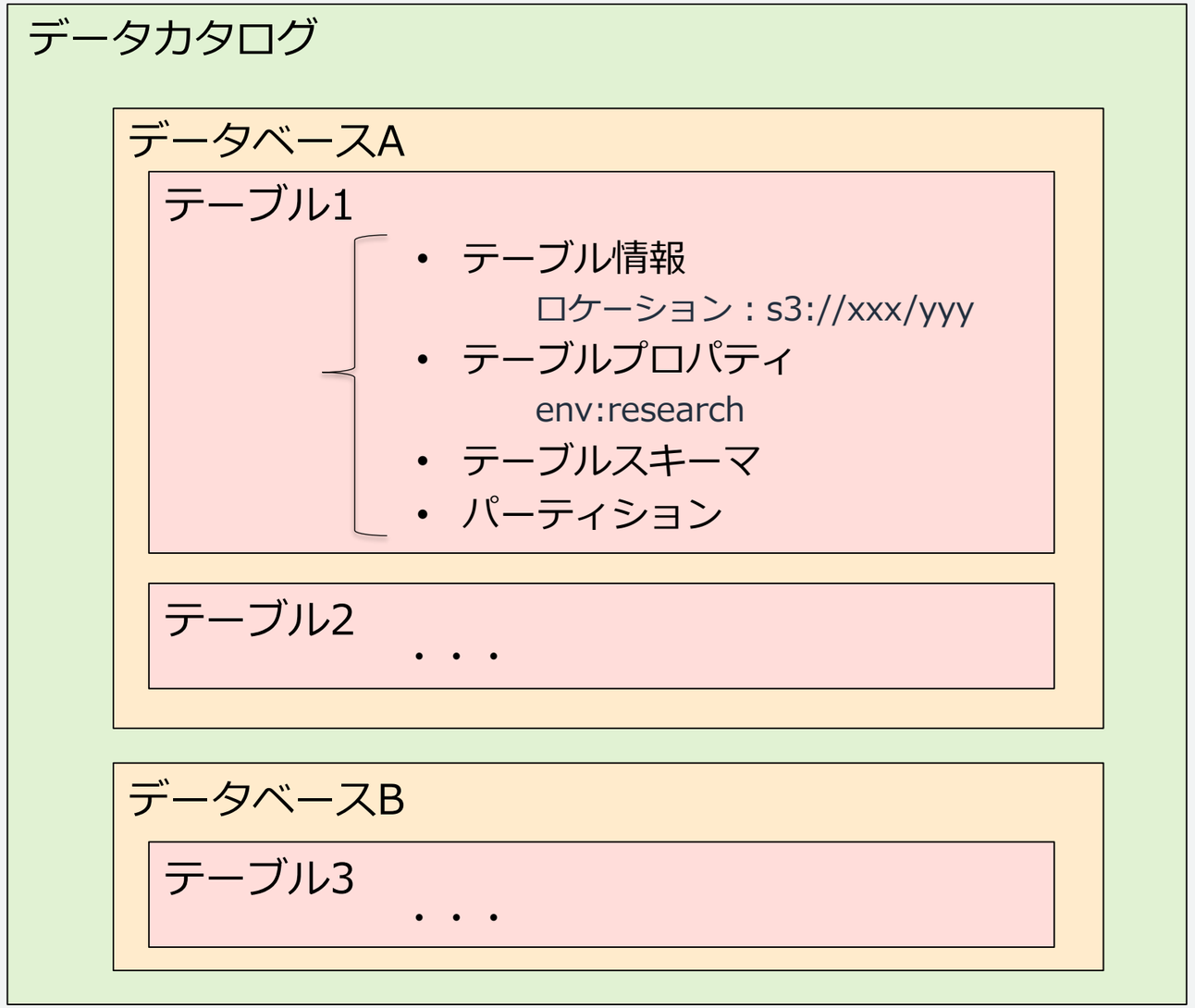


データカタログ機能

Lake Formation データカタログ

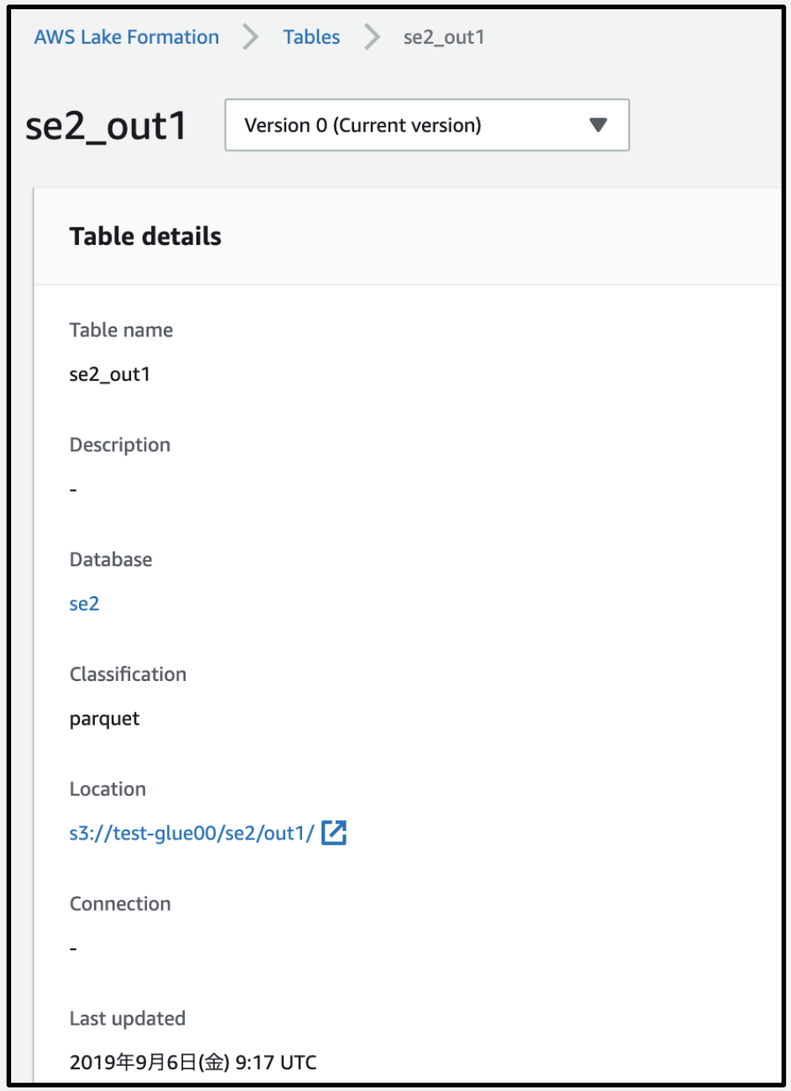
Apache Hive メタストアで行うのと同じ方法で、AWS でメタデータを保存、注釈付け、共有できるマネージドサービスです。

1. データカタログ
   - Glue データカタログと統合
   - Apache Hive メタストア互換
2. 同一 AWS アカウント内で共有可能な統合リポジトリ
3. データベースはテーブルの集合体
4. テーブルはデータのスキーマ情報、ロケーションなどのメタデータを保管



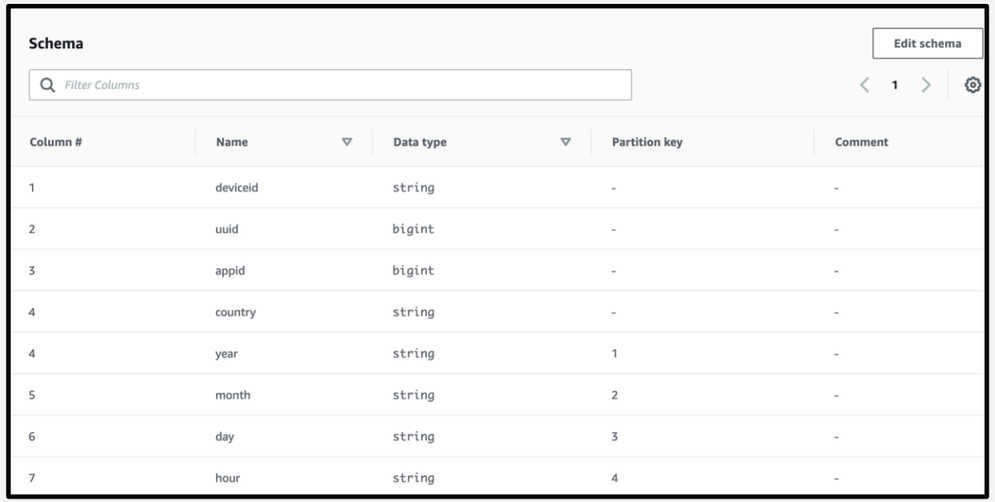
メタデータの構成例

テーブル情報

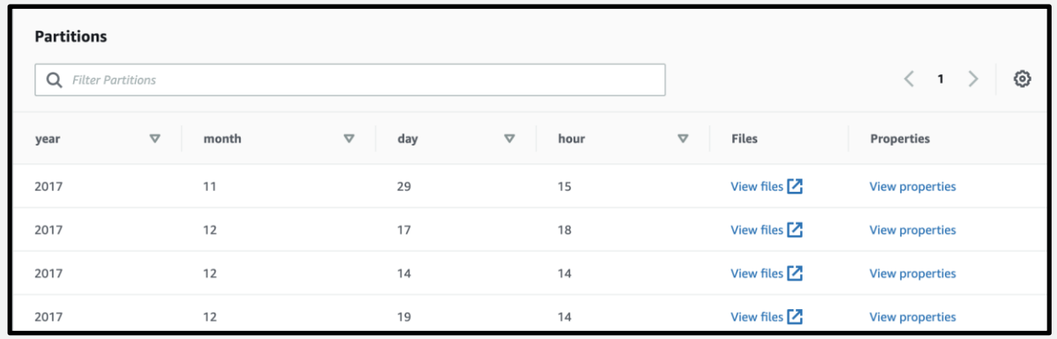


テーブルプロパティ

テーブルスキーマ



テーブルパーティション



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。

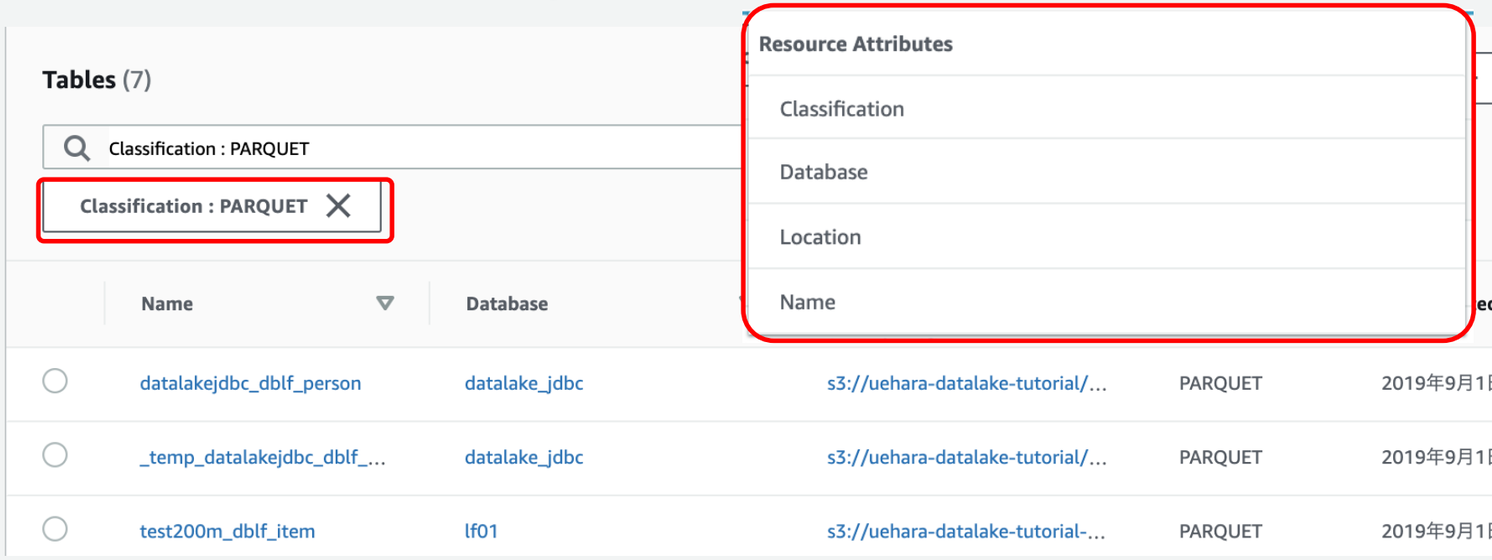
- Resource Attributes 検索
  Classification(例えばPARQUET), Database(例えばDatabaseA), Location(例えばs3),Name(テーブル名で例えばTableA)

- 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

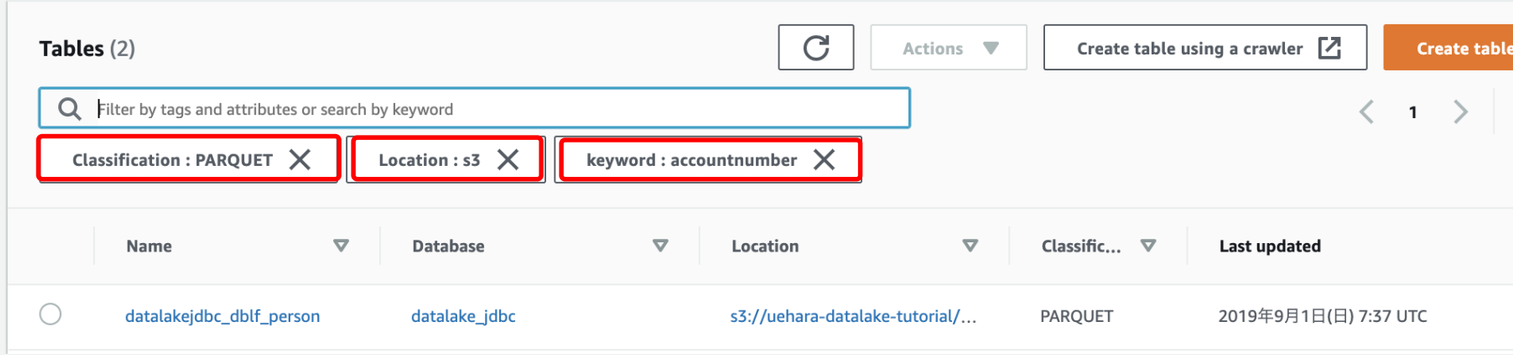
- 複数のフィルタ検索
  上記のいくつかを組み合わせて検索

※複数のキーワードは組み合わせられない

✓ 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



データカタログの検索

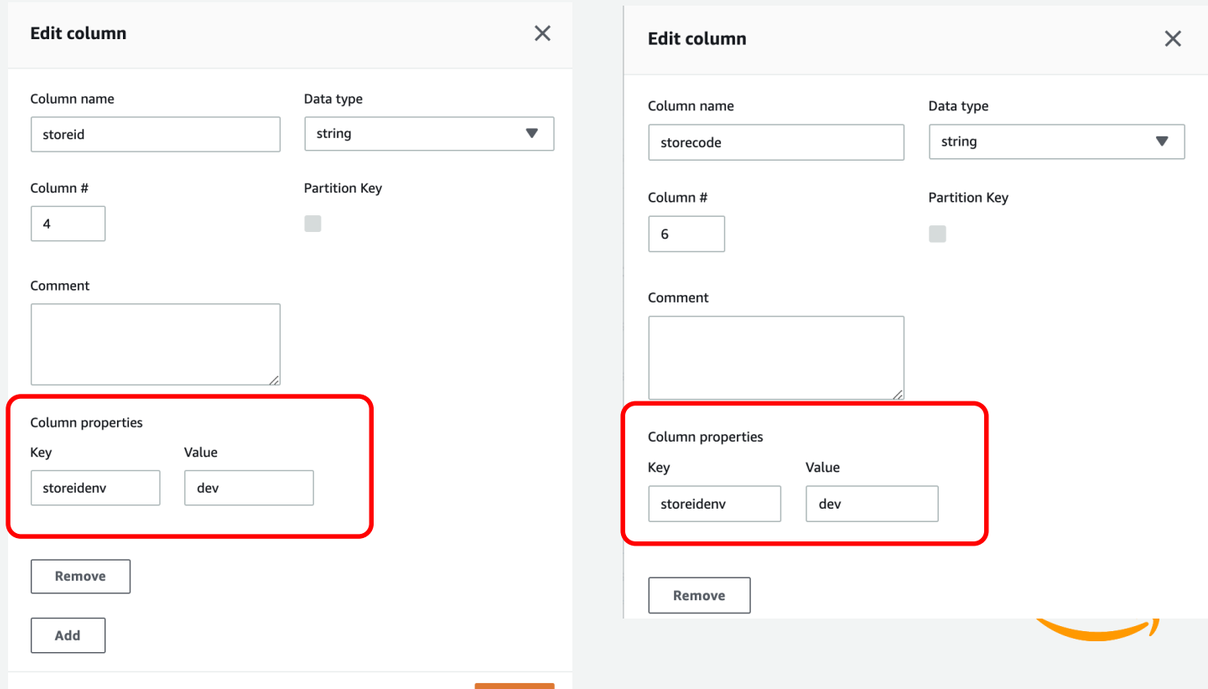
ビジネス固有の属性をテーブルプロパティやカラムプロパティとして追加できます。

タグを設定

- テーブル datalakejdbc_dblf_person のテーブルプロパティに env:research を追加
- テーブル datalakejdbc_dblf_person のカラム storeid とテーブル test_person のカラム storecode に、カラムプロパティ storeidenv:dev を追加

✓ テーブルプロパティ (テーブル datalakejdbc_dblf_person)

✓ カラムプロパティ (テーブル datalakejdbc_dblf_person の storied 、テーブル test_person の storecode)



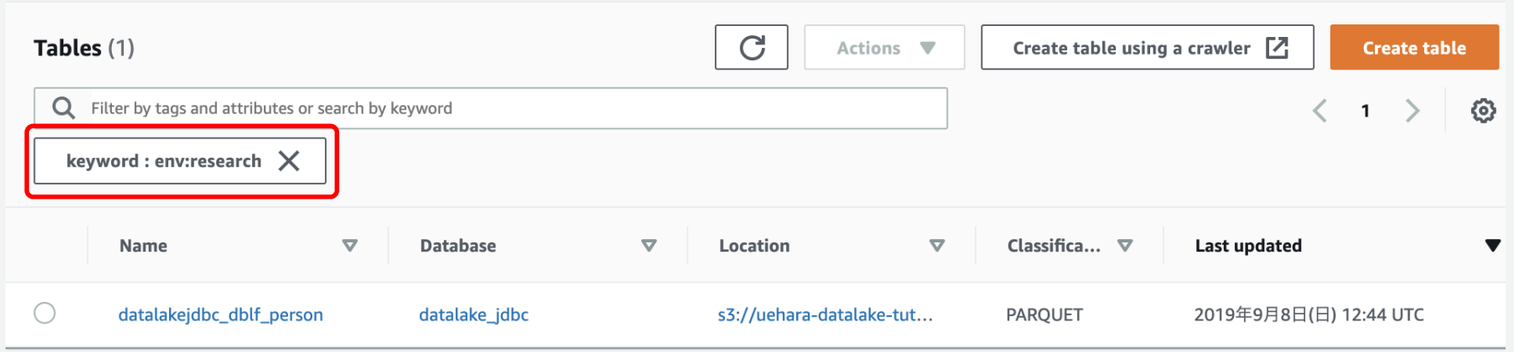
データカタログの検索

追加したビジネス固有の属性(テーブルプロパティやカラムプロパティ)で検索できます。マーケティングやリサーチャーなどデータを扱いたい人が欲しいデータを探しやすくなります。

タグによる検索

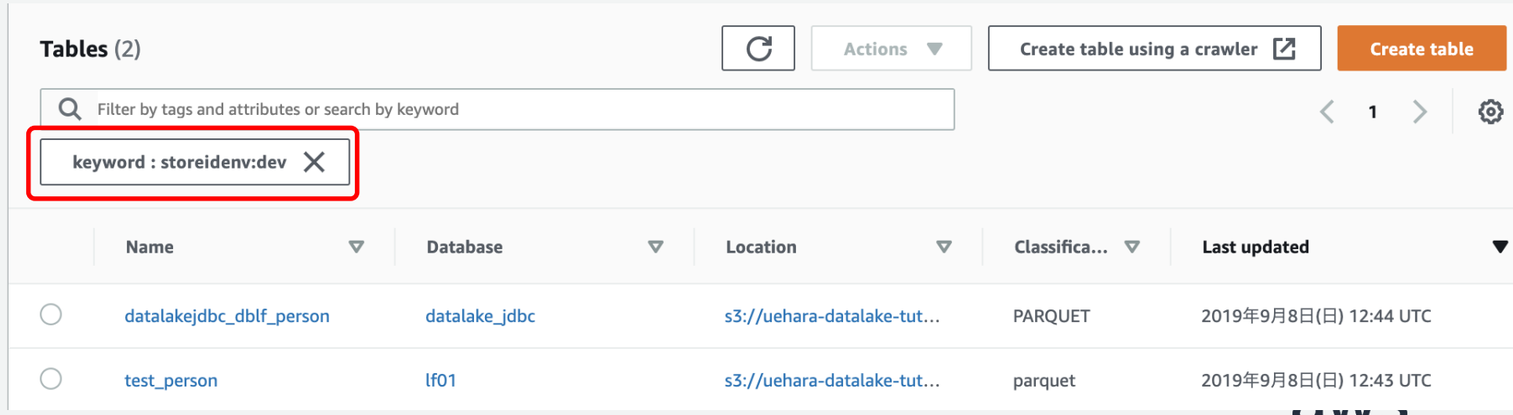
- テーブルプロパティで“env:research”で検索。datalakejdbc_dblf_person がヒット

✓ テーブル datalakejdbc_dblf_person (テーブルプロパティ env:research)



- カラムプロパティで”storeidenv:dev”による検索。test_person, datalakejdbc_dblf_person の2つのテーブルがヒット

✓ テーブル datalakejdbc_dblf_person (storeidカラムプロパティ storeidenv:dev)
✓ テーブル test_person (storecodeカラムプロパティ storeidenv:dev)



データカタログのアクセス許可モデルの移行

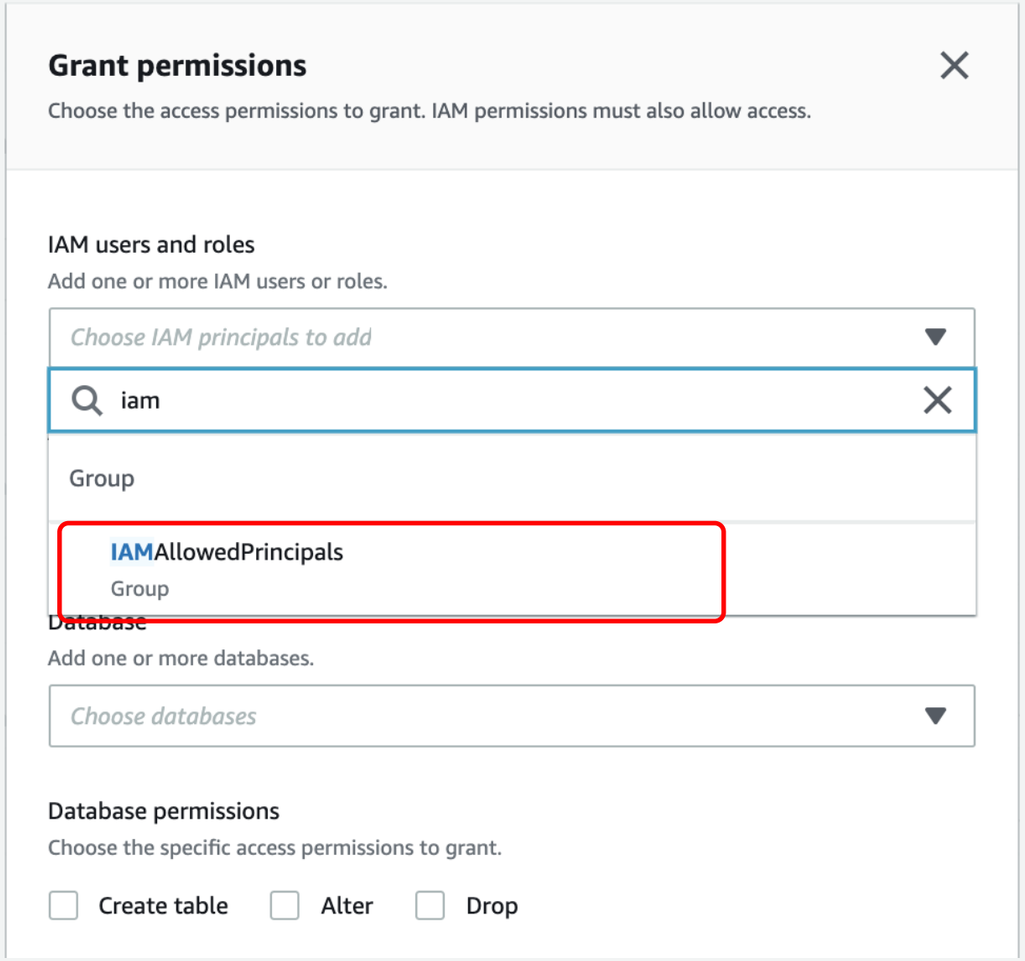
Lake Formation により広範な IAM を使ったアクセス制御を簡単に設定することが出来ます。これに対して Glue のアクセス許可モデルは IAM を介してアクセスを許可します。

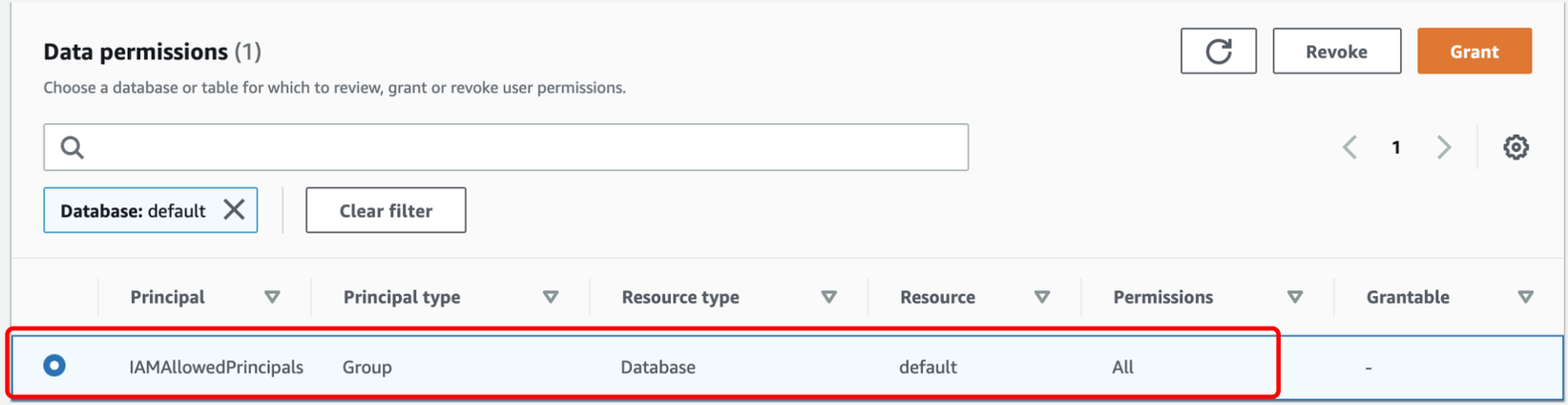
従来から Glue をご利用頂いているお客様は Lake Formation のアクセス許可モデルへの移行が必要になります。

IAMAllowedPrincipals

Lake Formation は、既存の Glue データカタログとの互換性のために有効化された"grant All to IAMAllowedPrincipals"で利用を開始します。これには、既存の Glue アクセス許可を引き続き有効にするデフォルトのアクセス許可が含まれています。これは「All to IAMAllowedPrincipals」と呼ばれる特別な許可を介して行われます。

Lake Formation 権限の使用に移行した後、この設定を無効にすることをお勧めします。





データカタログのアクセス許可モデルの移行

公式ドキュメントによる手順

https://docs.aws.amazon.com/ja_jp/lake-formation/latest/dg/upgrade-glue-lake-formation.html

- Step1：既存のアクセス許可を確認
- Step2：Lake Formation でこれらの権限を作成
- Step3：IAM ユーザーに Lake Formation のアクセス許可を付与
- Step4：データストアを Lake Formation アクセス許可モデルに切り替える
- Step5：IAM ユーザーに新しい追加の IAM ポリシーを付与
- Step6：IAM ポリシーをクリーンアップ

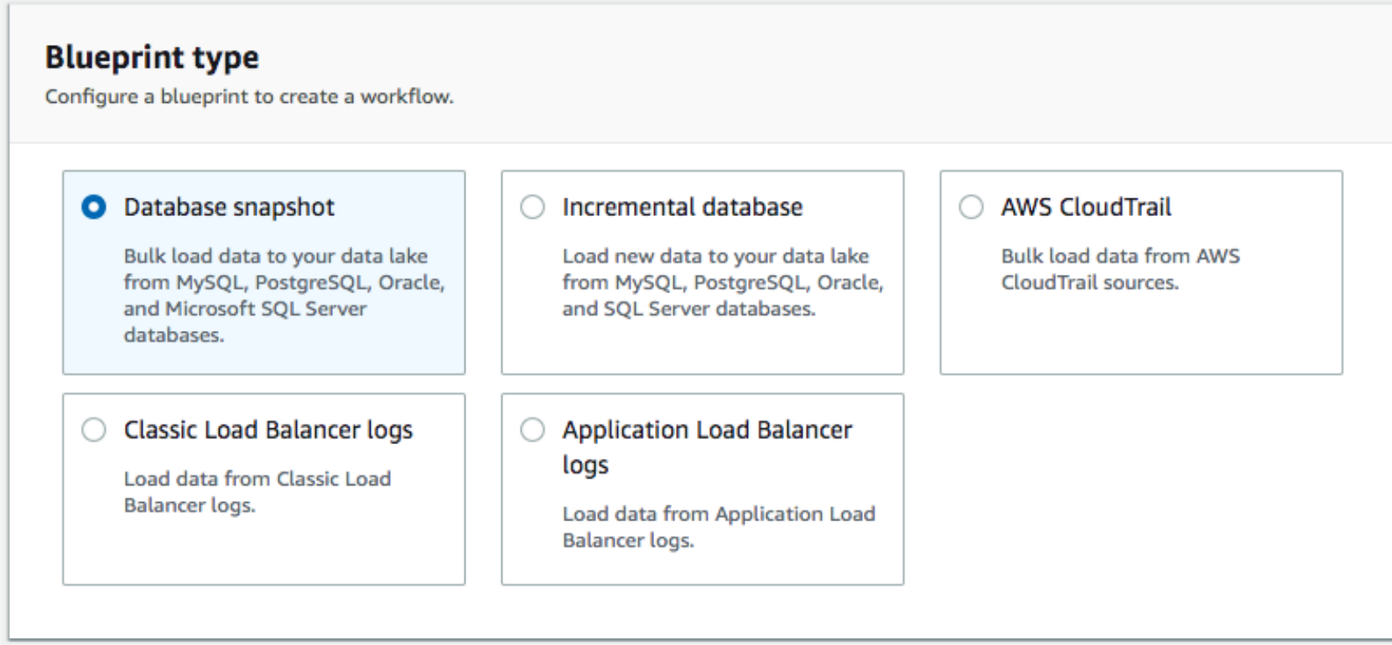
ブループリント機能

ブループリントとは

ブループリントは、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」です。汎用的なユースケースのブループリントを複数用意しています。

2 種類の ブループリント (2019/10/01時点)　※ブループリントは随時追加予定

- データベース用 ブループリント
  - MySQL、PostgreSQL、Oracle、SQL Server からのデータの取込み
- ログファイル用 ブループリント
  - 以下の AWS 主要ログの取込み
    - AWS CloudTrail
    - Application Load Balancer
    - Classic Load Balancer

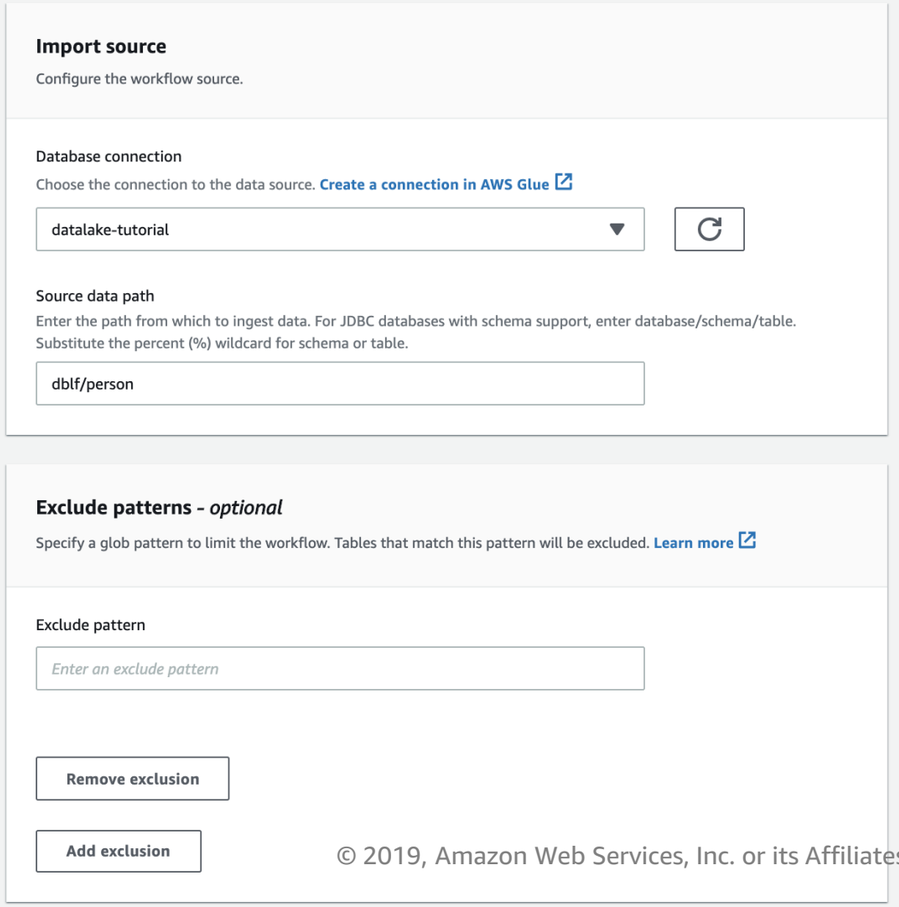


バルクロード or 増分ロード

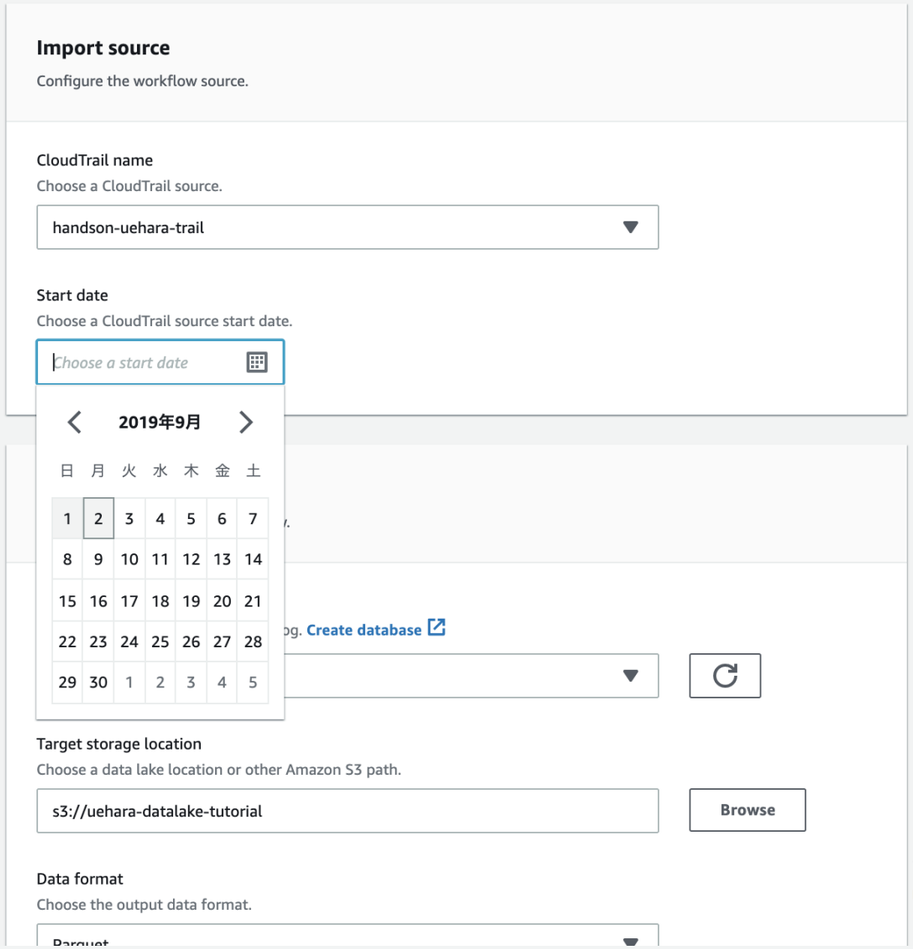
ブループリントとは

ブループリント上で必要なパラメータを入力し Lake Formation のワークフローを作成することで、Glue のトリガー、ワークフロー、クローラー、ジョブを自動で生成します。

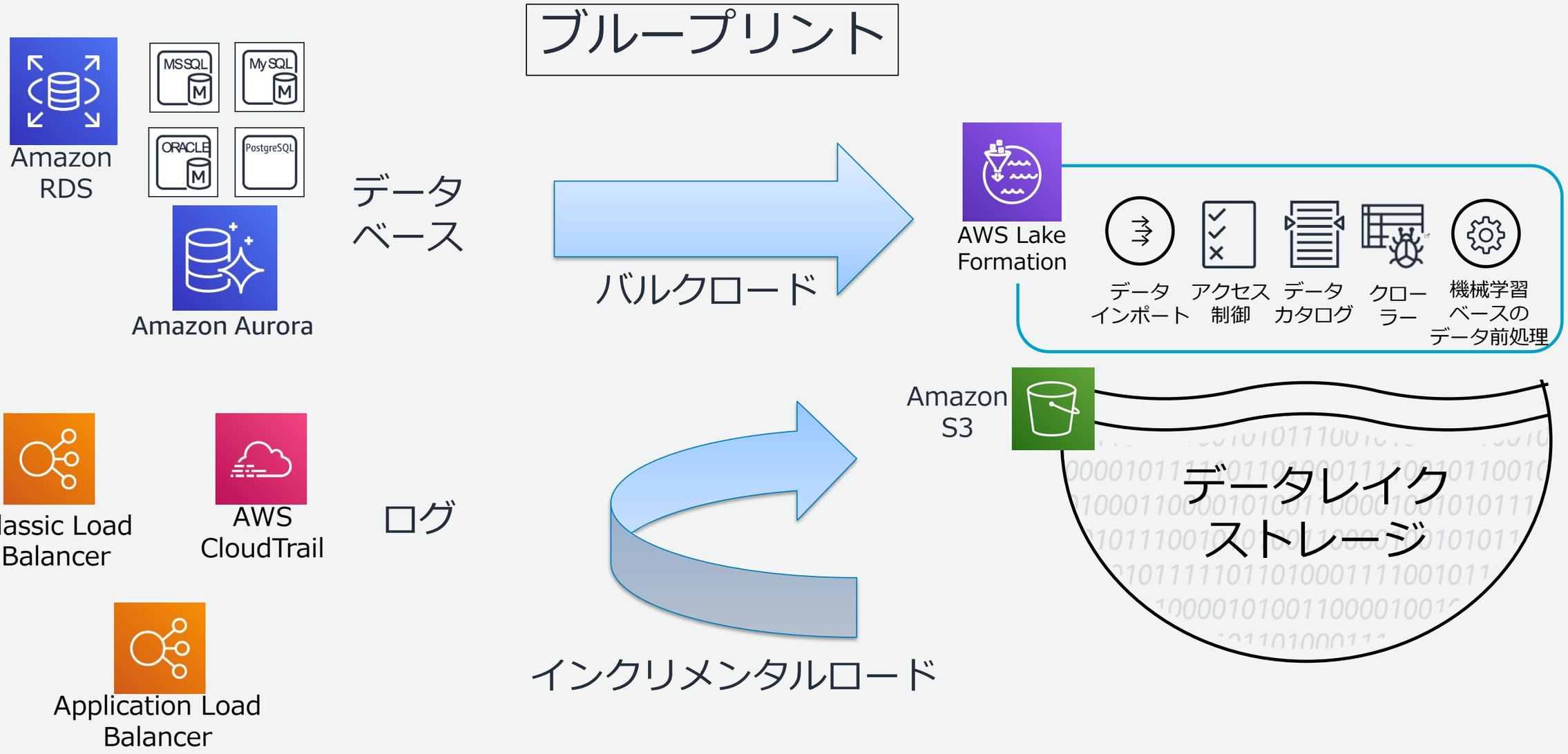
“Database snapshot” の例



“AWS CloudTrail” の例

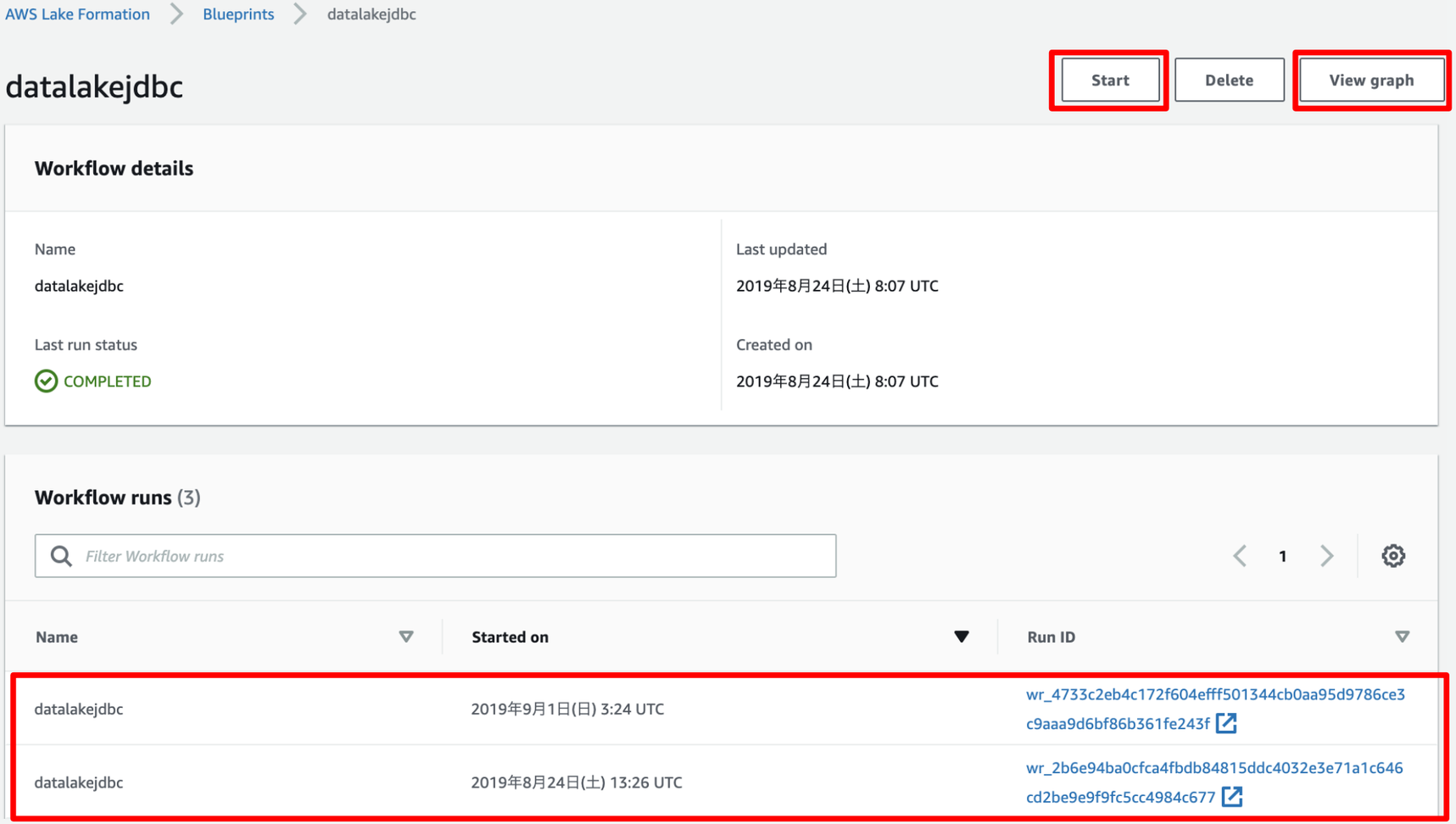


ブループリントによるデータレイクへの容易なデータロード



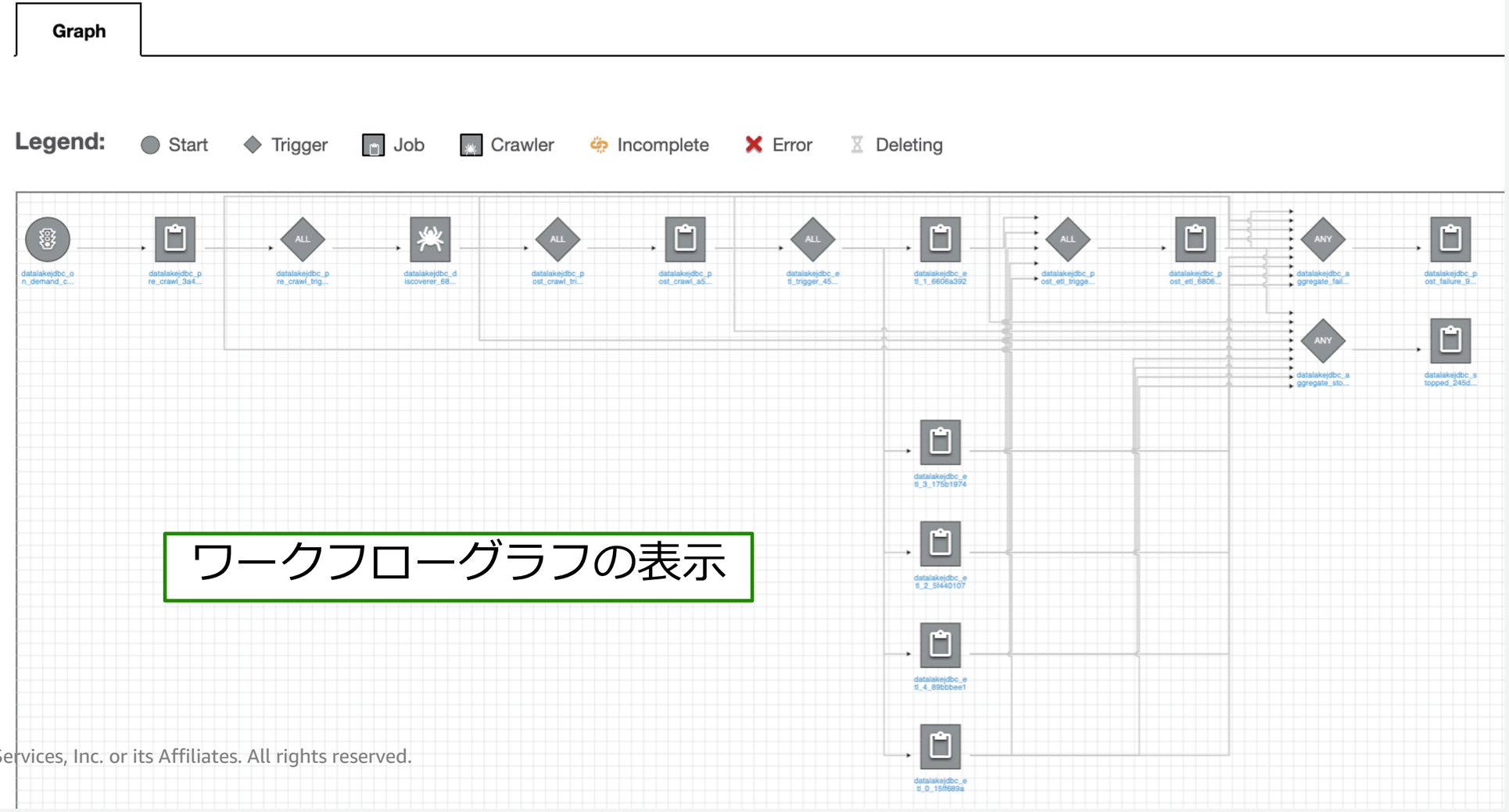
ワークフロー：ブループリントから作成される処理フロー

ブループリントを使い作成された Lake Formation のワークフローは、[Start] で実行することができます。ワークフローの画面では実行履歴の確認や [View graph] でワークフローの詳細を確認することができます。



ワークフロー：ブループリントから作成される処理フロー

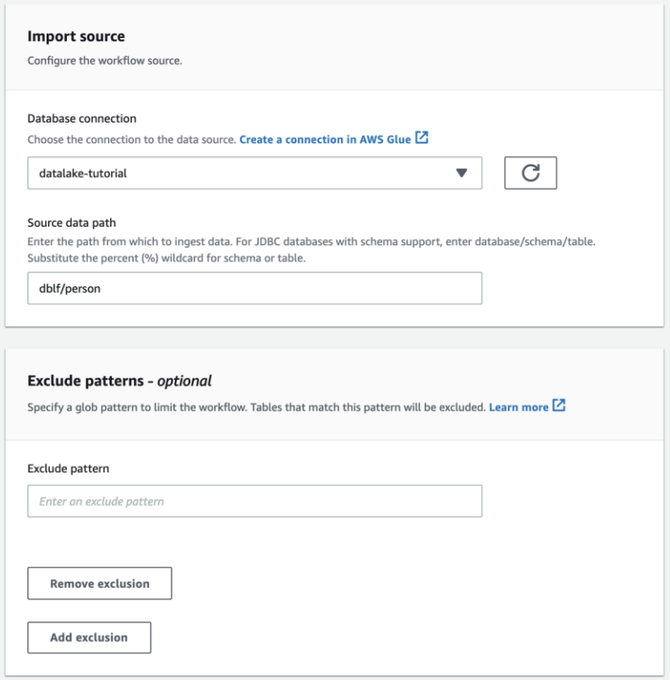
Lake Formation のワークフローの実態は、Lake Formation のブループリントから生成される一連の Glue ジョブ、クローラー、トリガーの定義セットである Glue ワークフローです。ワークフローによって複雑な ETL ジョブアクティビティを隠蔽し、下記のようなワークフローグラフにより進捗を可視化することが可能です。



ユースケース：Database snapshot のブループリント

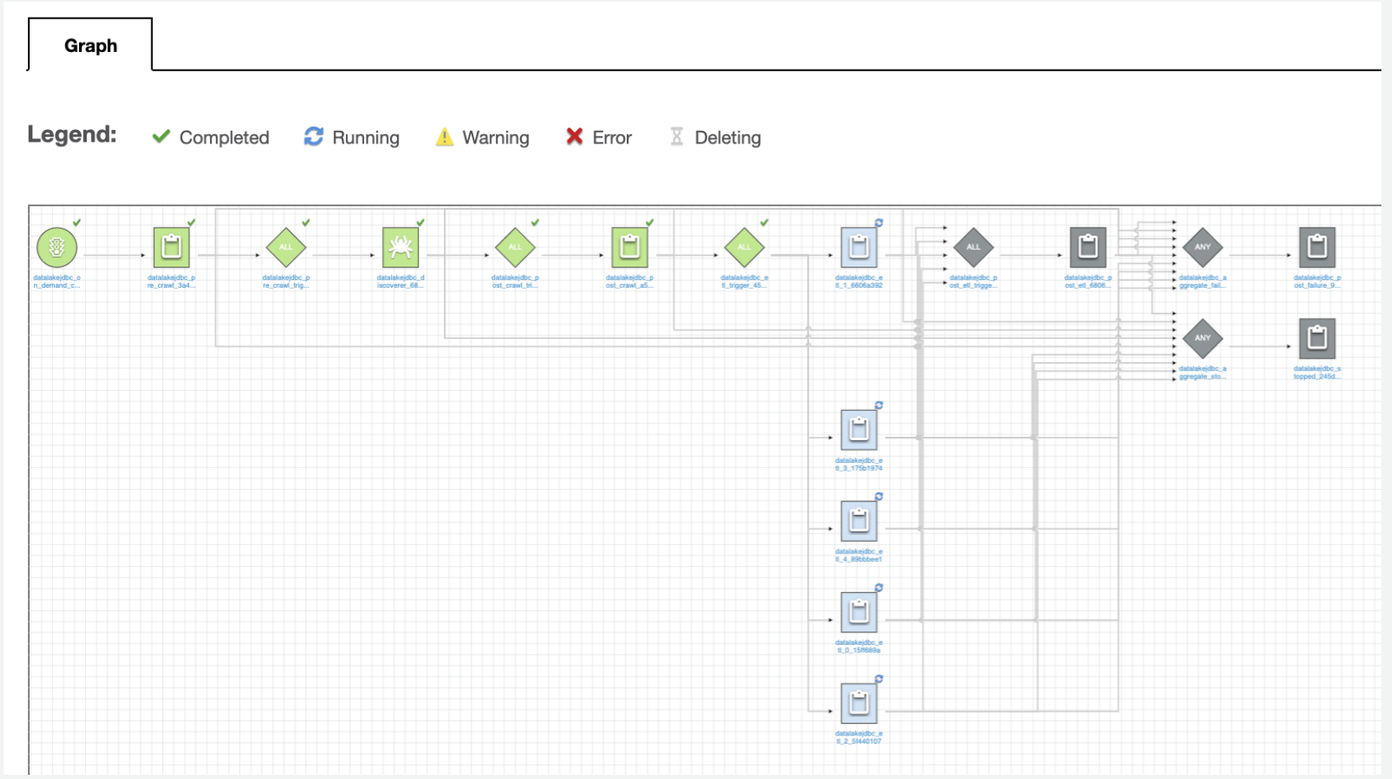
Database snapshot のブループリントを使うと、RDB のデータを分析に適したパーティションの最適化やフォーマット(parquetなど)変換を行い S3 に出力し Athena 等で分析が行えます。ブループリントに必要なパラメータを入力し、作成された Lake Formation のワークフローを実行することでデータセットやテーブルの作成は完了します。

Database snapshot のブループリントによりワークフロー作成



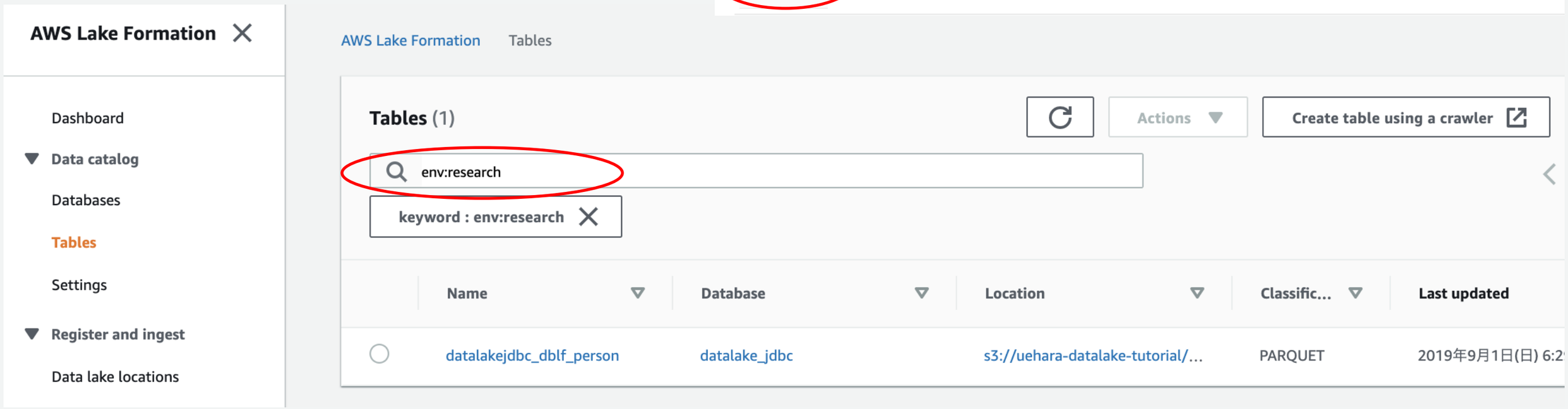
ワークフロー実行

ワークフローの進捗状況



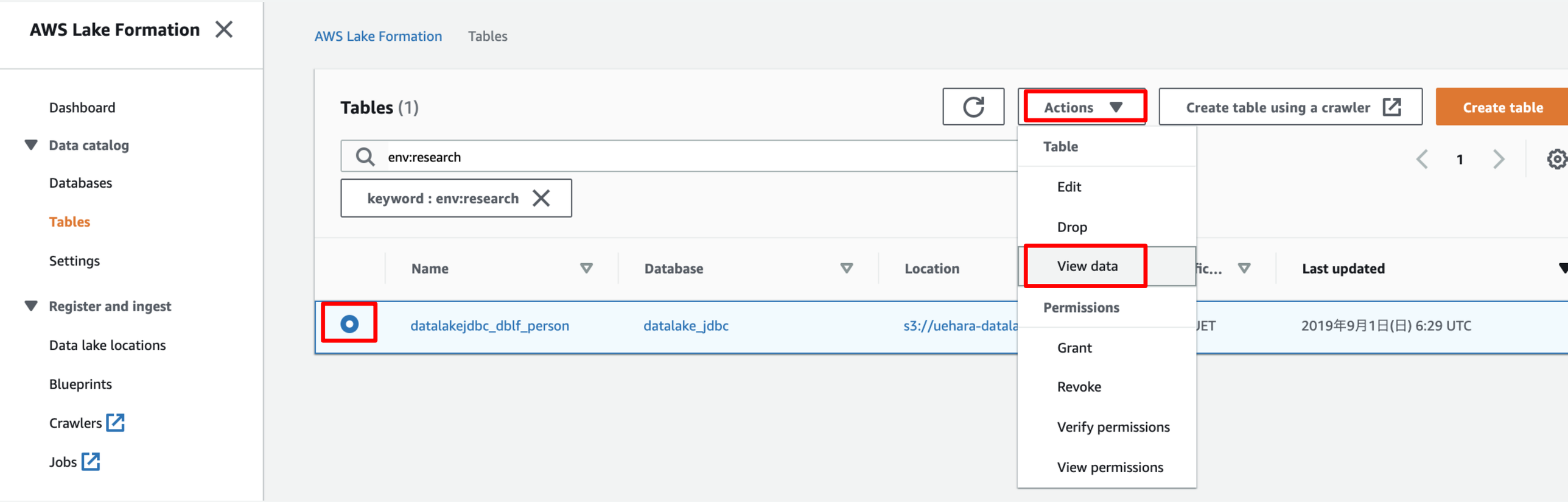
ユースケース：Database snapshot のブループリント

利用者は見たいテーブルを簡単に探せます。テーブルにテーブルプロパティを キー:バリュー (例、env:research)で設定しておくことで、利用者が env:research で検索し閲覧権限のある該当のテーブルを見つけることができます。



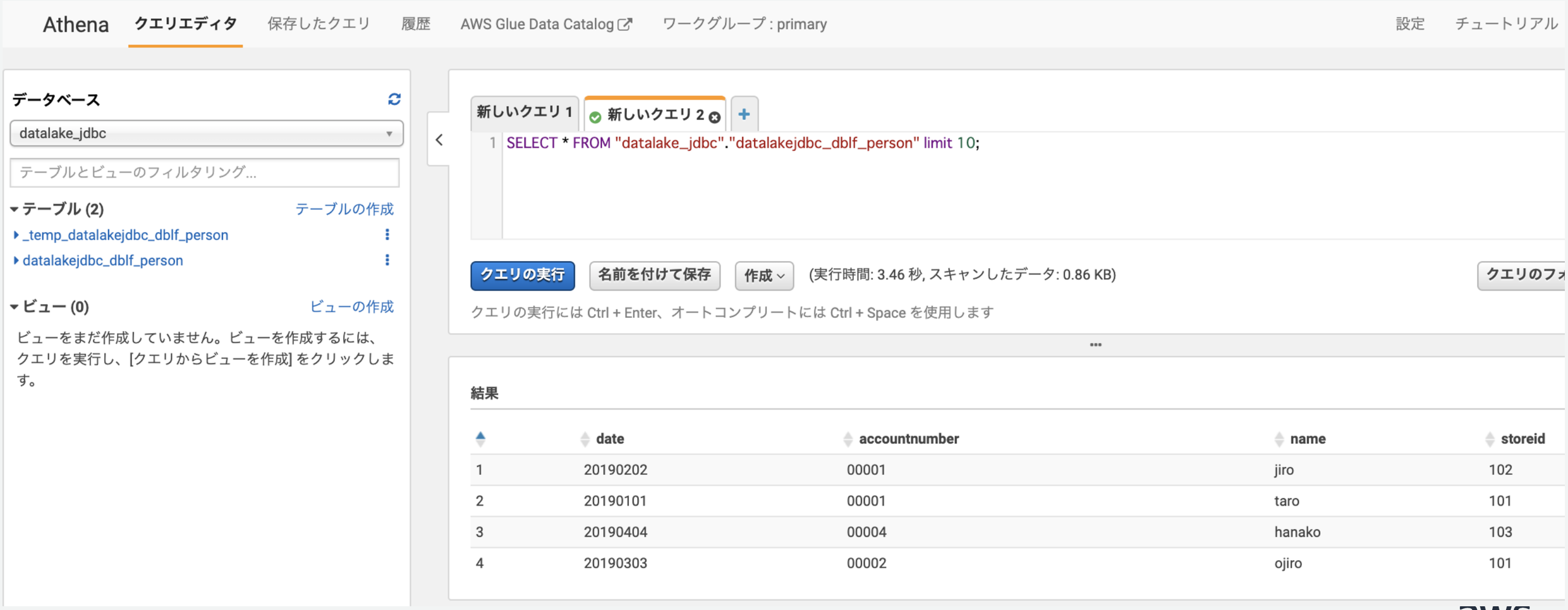
ユースケース：Database snapshot のブループリント

該当のテーブルにチェックを入れ、[Actions] から [View data] をクリックします。



ユースケース：Database snapshot のブループリント

Athena の画面に遷移し Athena で分析を行うことができます。

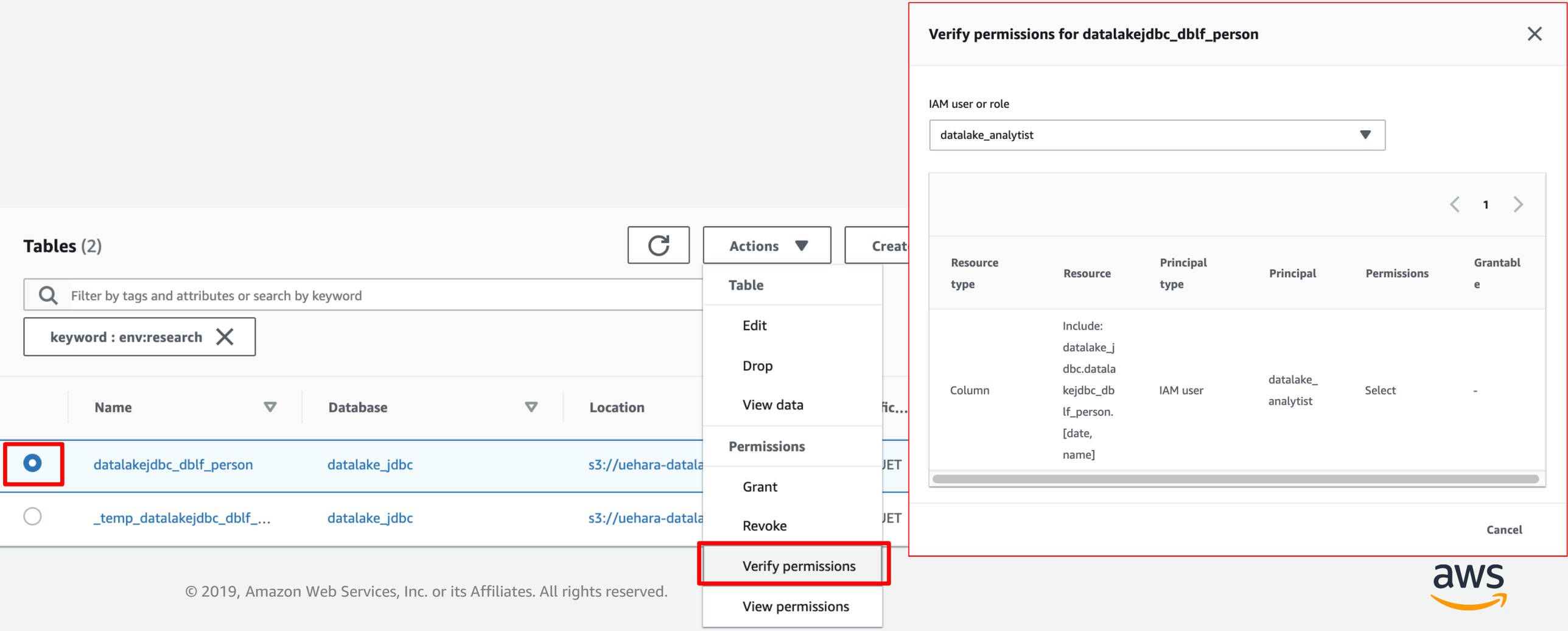


ユースケース：Database snapshot のブループリント

ユーザーごとに列レベルの制御を行い、アクセスできる列を絞ることで特定のユーザーに見せるデータを制御することも可能です。
以下の例は datalake_analytist というユーザーに date と name の列のみ select を許可しています。

ユースケース：Database snapshot のブループリント

該当のテーブルの [Verify permissions] で datalake_analytist に与えられている権限を確認することができます。date と name の列のみ select が許可されていることがわかります。



落ち穂拾い

Lake Formation と Glue の関係

Glue は抽出、変換、ロード (ETL) を行うジョブとデータカタログを持つマネージドサービスです。

- Lake Formation は Glue の拡張機能であり、Glue に大きく依存しています。
- Lake Formation は Glue とデータカタログを共有しています。
- Lake Formation のジョブとクローラーは Glue のジョブとクローラーを呼び出します。
- Lake Formation のブループリントで生成されるワークフローは Glue のワークフローです。Lake Formation と Glue の両方でワークフローの表示と管理ができます。

https://www.slideshare.net/AmazonWebServicesJapan/20190806-aws-black-belt-online-seminar-aws-glue

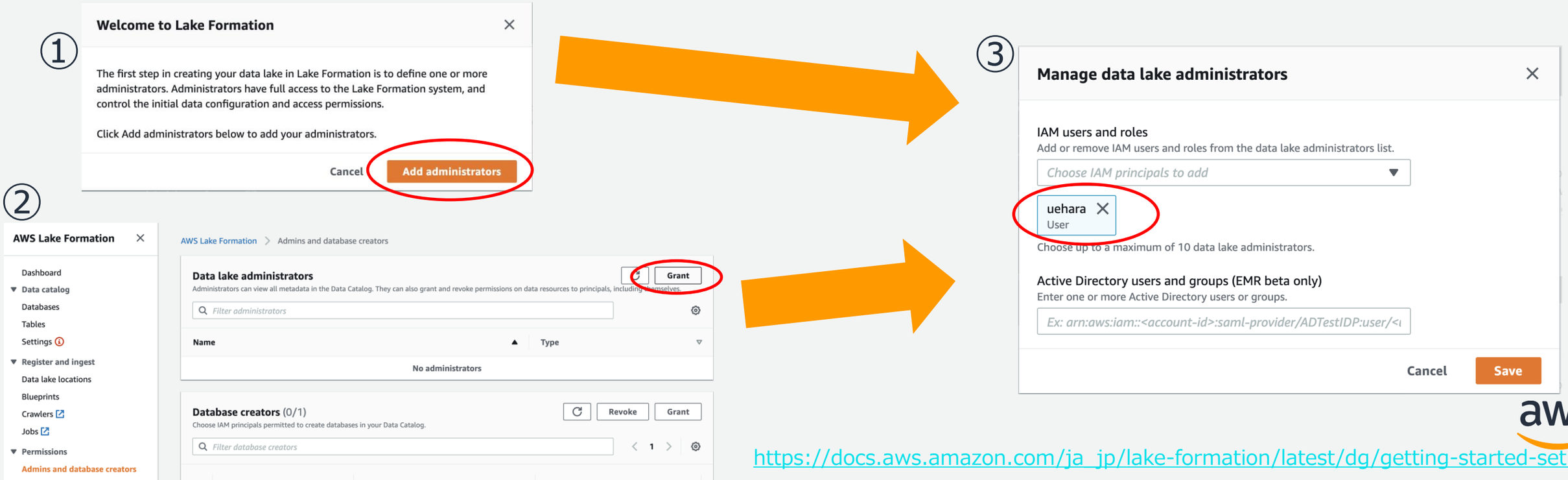
初めてデータレイク管理者を作成する

任意の IAM 管理ユーザーとしてサインインします。①または②を行った後③を実行します。

① Lake Formation に初めてアクセスするとウェルカムメッセージが表示されます。この画面で [Add administrator]を選択します。

② ナビゲーションペインで、[Admins and database creators]を選択します。次に、”Data lake administrators”で、[Grant]を選択します。

③ [Manage data lake administrators]ダイアログボックスの”IAM users and roles”で、 データレイク管理者になるユーザーを選択し、[Save]を選択します。



https://docs.aws.amazon.com/ja_jp/lake-formation/latest/dg/getting-started-setup.html

料金

AWS Lake Formation 費用

AWS Lake Formation 自体の利用は無償

配下で活用するサービスに

かかる費用のみのお支払い

料金

例として、ブループリント”Database snapshot”をデフォルト設定で RDS から200MBのテーブルを S3 にロードし、Athena で100回フルスキャンクエリを利用した場合

- Glue クローラー1つを数分実行
  最小時間の10分の料金： $0.44 / 6 = $0.0734
- Glue Spark ジョブ 5つを数分実行
  最小時間の10分の料金* 5 ジョブ： ($0.44 / 6)*5 = $0.367
- Glue Python Shell ジョブ5つを数秒実行
  最小時間の1分の料金* 5 ジョブ： ($0.44 / 60)*5 = $0.0367
- S3 に200MBのファイルを保存
  1GBあたり$0.025なので200MB換算： $0.005
- Athena で100回フルスキャンクエリを実行
  1 TBで$5なので、0.0002TB(200MB)を掛けて回数100を掛ける：($5 * 0.0002)*100 = $0.1

✓**合計で$ 0.5821**

※本日時点の東京リージョンの料金の例です。正確には公式の料金のページを確認ください
※実際の処理時間はデータの内容により変わります
※Glue データカタログ、S3 のリクエスト料金は少額なので割愛しています
※Lake Formation によって作成されるリソースの数や内容は今後変更される可能性はあります

https://aws.amazon.com/jp/athena/pricing/
https://aws.amazon.com/jp/s3/pricing/
https://aws.amazon.com/jp/glue/pricing/

まとめ

AWS Lake Formation まとめ

- ブループリントと呼ばれる汎用的なデータ取込み処理のテンプレート化
  - Glue のワークフロー、クローラー、ジョブと連携

- データのアクセス権限管理をシンプルに直感的に設定可能
  - データカタログのデータベース、テーブル、列に対して、Grant/Revoke 設定
  - メタデータ、S3 のデータレイクストアへのアクセス制御の一元管理
  - Athena、Redshift Spectrum、EMR(今後予定)からのアクセスを容易に
  - 既存の Glue データカタログからは移行が必要

- Lake Formation 自体の費用は無償
  - 配下で使用するサービスの費用のみ

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 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/

Q&A

お答えできなかったご質問については
AWS Japan Blog 「https://aws.amazon.com/jp/blogs/news/」にて
資料公開と併せて、後日掲載します。

AWS の日本語資料の場所「AWS 資料」で検索

https://amzn.to/JPArchive

ご視聴ありがとうございました