データレイクの活用に欠かせない運用のポイント

Noritaka Sekiyama
Big Data Architect, AWS Glue & Lake Formation
2020/6/11

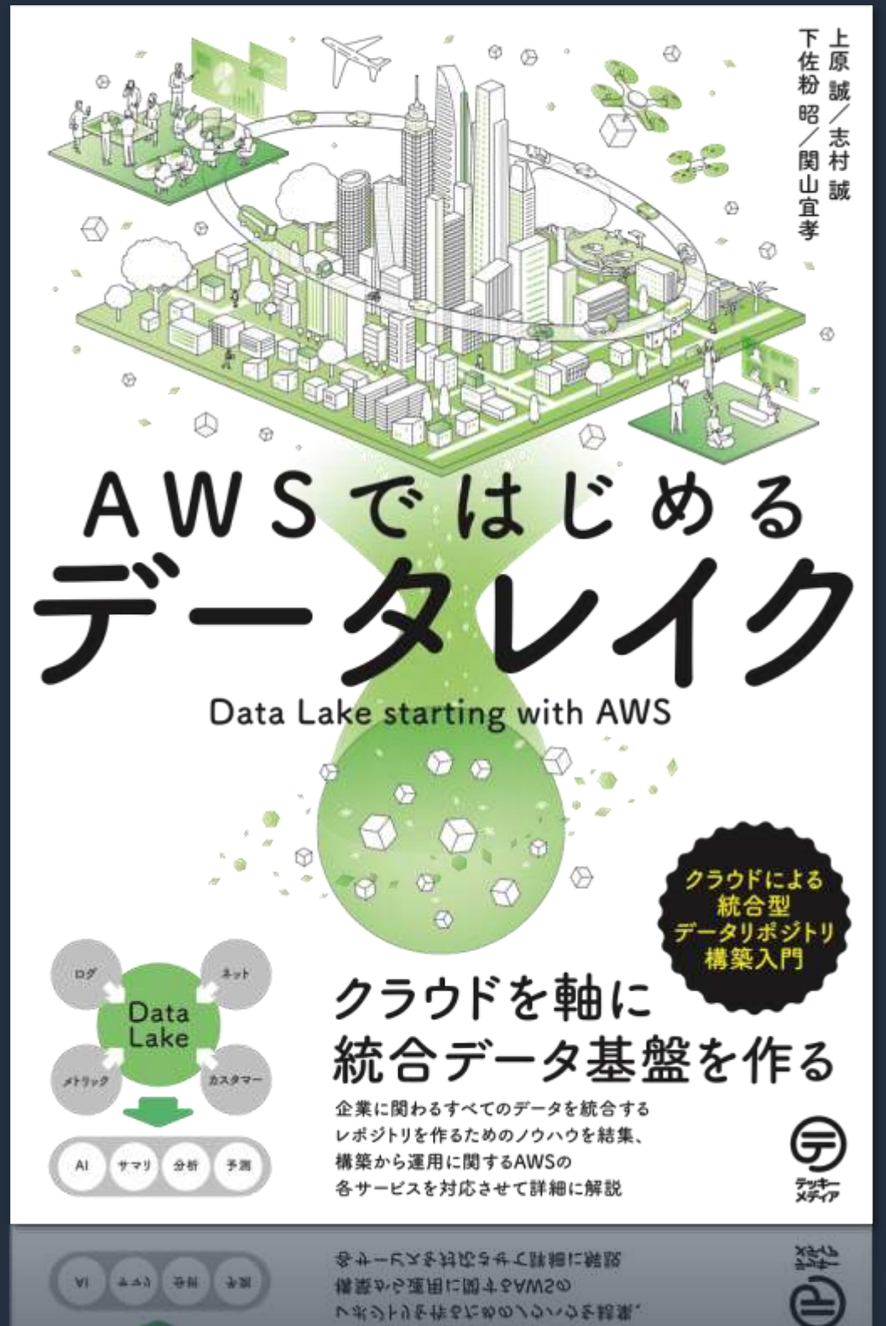
AWSオンラインセミナーへようこそ

お気軽にご質問ください！

- 書き込んだ質問は主催者にしか見えません
- 最後のQ&A時間で、いただいたご質問からピックアップしてご回答をさせていただきます

① 吹き出しをクリック
② 質問を入力
③ Sendをクリック

「AWSではじめるデータレイク」
2020年7月発売予定

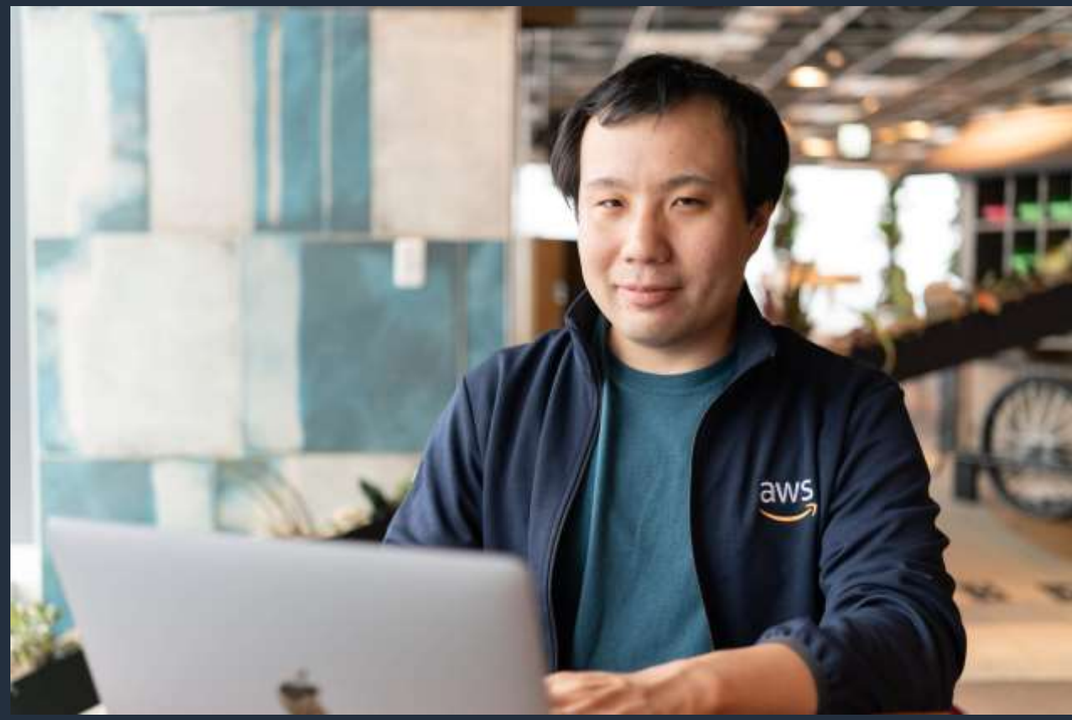


自己紹介

関山 宜孝

Big Data Architect
AWS Glue & Lake Formation

- 5年間 AWS サポートにて技術支援を担当
- 2019年からプロダクト開発チームにジョイン
- GlueとLake Formationに関する
  ユーザーに近い部分の開発を担当



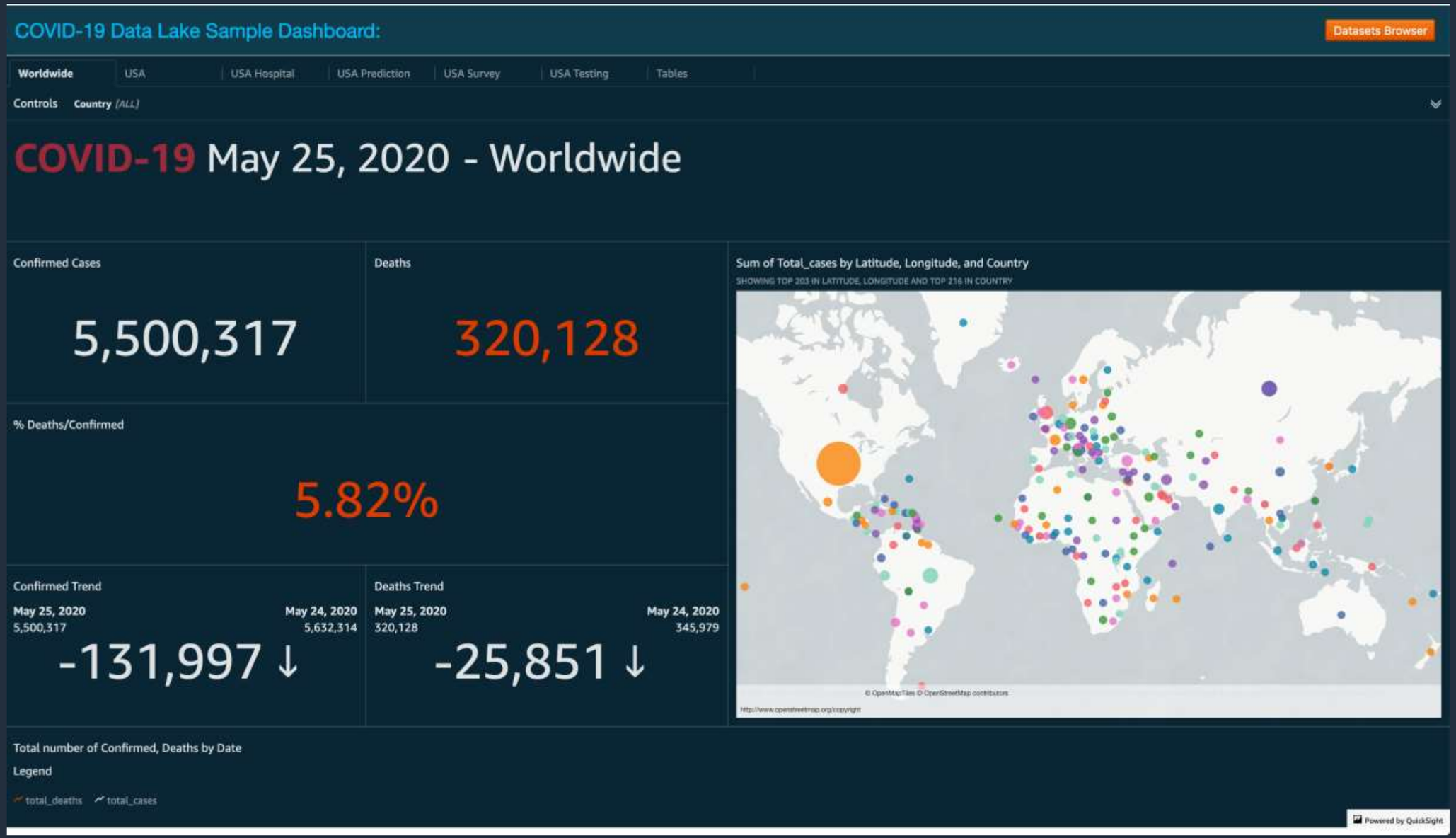
aws Forum **NoritakaS-AWS**

@moomindani

moomindani

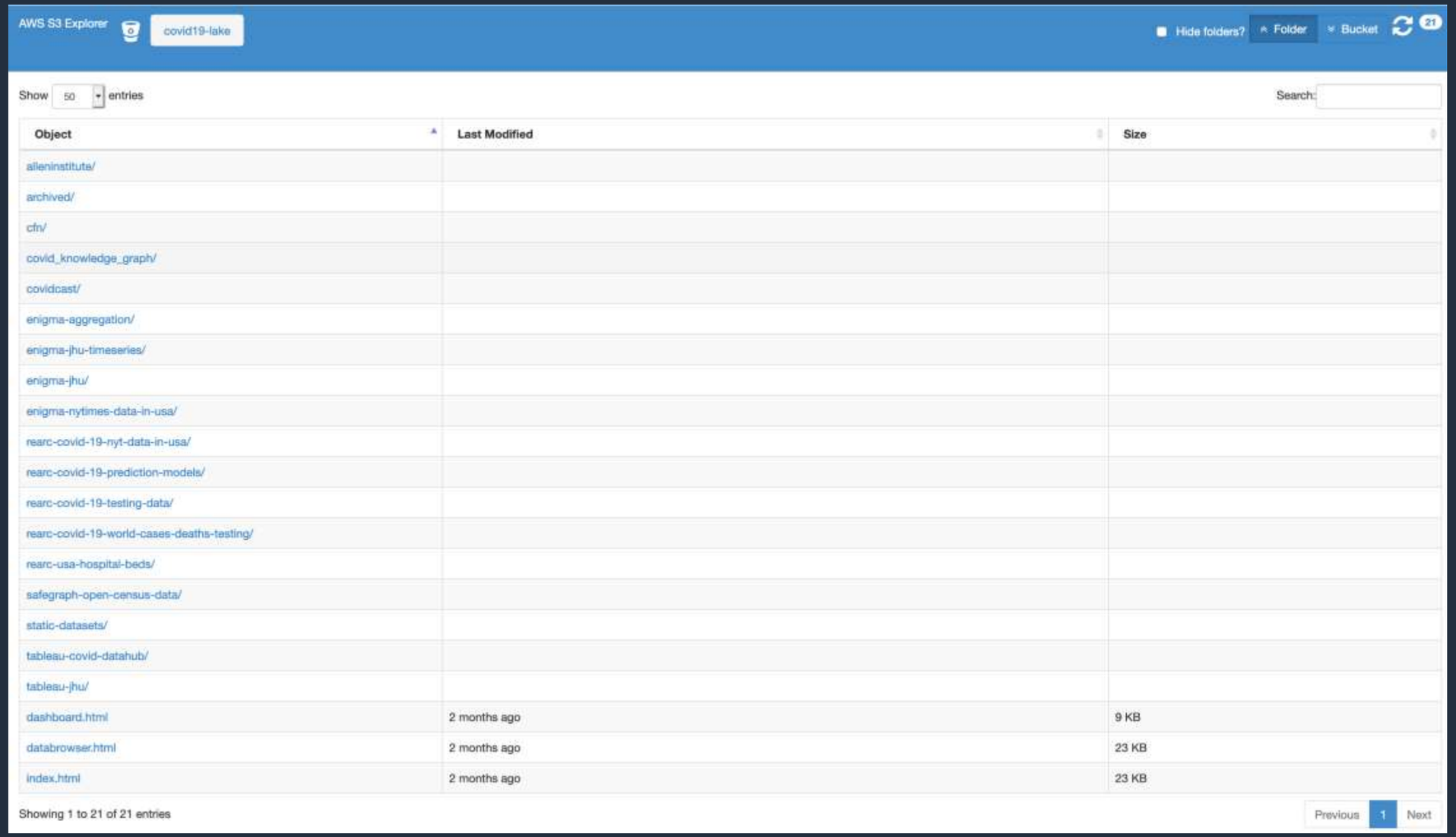
aws

サンプル: COVID-19 データレイク



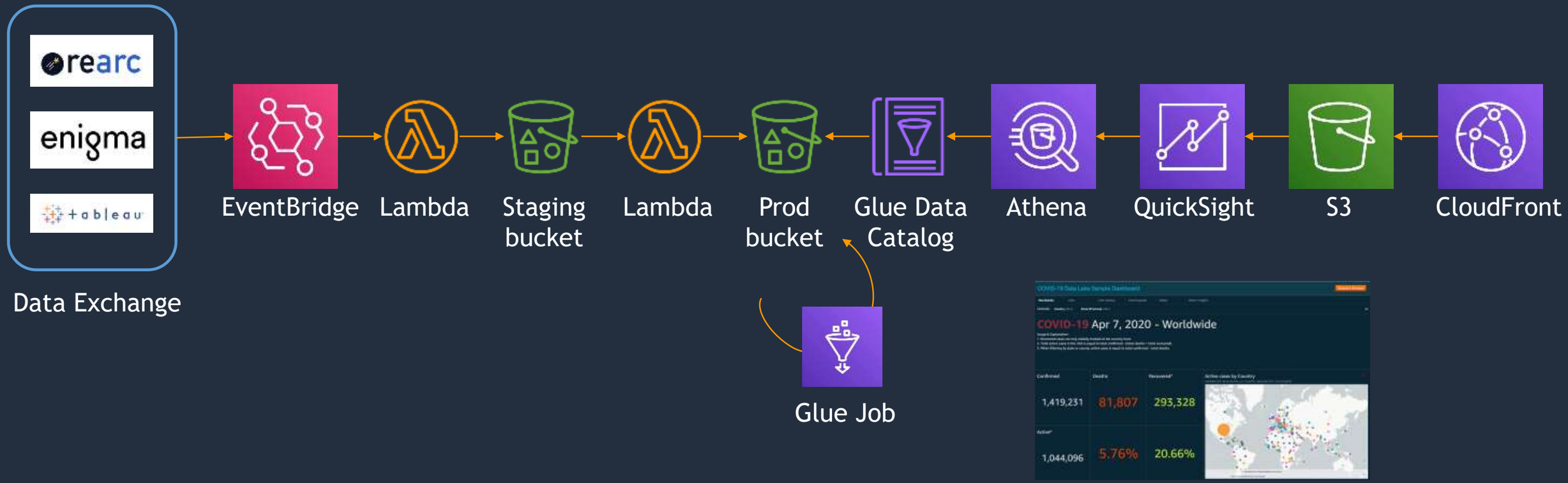
https://go.aws/covid-19-dashboard

サンプル: COVID-19 データレイク



https://go.aws/covid-19-datalake

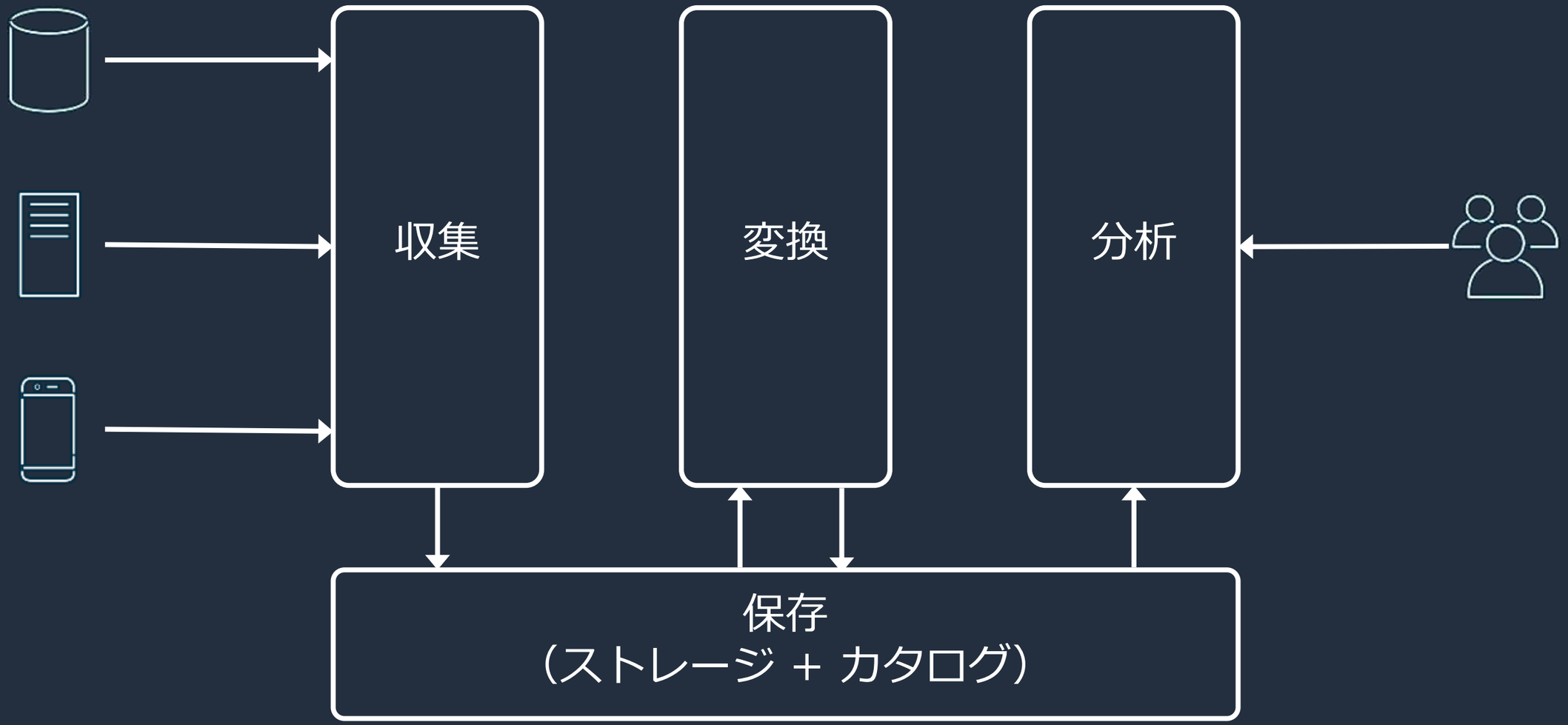
サンプル: COVID-19 データレイクのアーキテクチャ



アジェンダ

- データレイクでよくあるトラブル
- データレイクの運用
- 利用状況の計測とデータレイクの継続的改善
- Q&A

おさらい：データレイクの構築



データレイクでよくあるトラブル

データレイクが完成したら・・・

待望のデータレイク完成！
みなさんたくさん使ってください

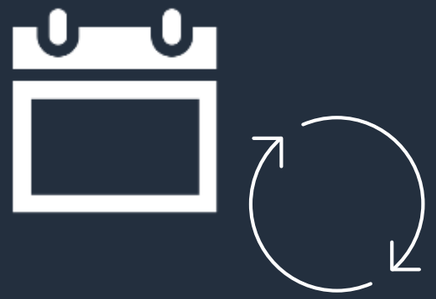
監視？運用？
まだあんまり考えてないけど
そのうちなんとかしますね

データレイク管理者
（初心者）

よくあるトラブル#1. データが古い

3日前からデータが更新されてなくて
新しいデータを分析できず困っています

データレイクユーザー



データレイク管理者
（初心者）

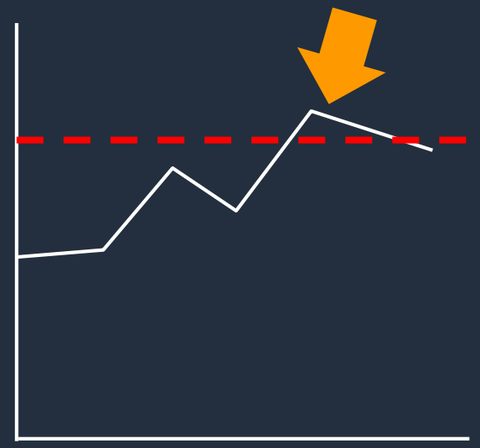
データが届いてない・・？
ジョブが失敗した・・？

aws

よくあるトラブル#2. クエリが遅い

クエリしてみたら以前に比べて
50%長く待たされるようになりました

データレイクユーザー



!?

データレイク管理者
（初心者）

データが予想より増えてた・・・？
でも50%も遅くなる・・・？

aws

よくあるトラブル#3. 想定してない結果が返ってきた

クエリしてみたら本来は日時が格納される
はずのカラムが NULL になっていました

データレイクユーザー

!??

データレイク管理者
（初心者）

ファイルの中身がおかしい・・・？
データの変換に失敗した・・・？

aws

データレイクと運用

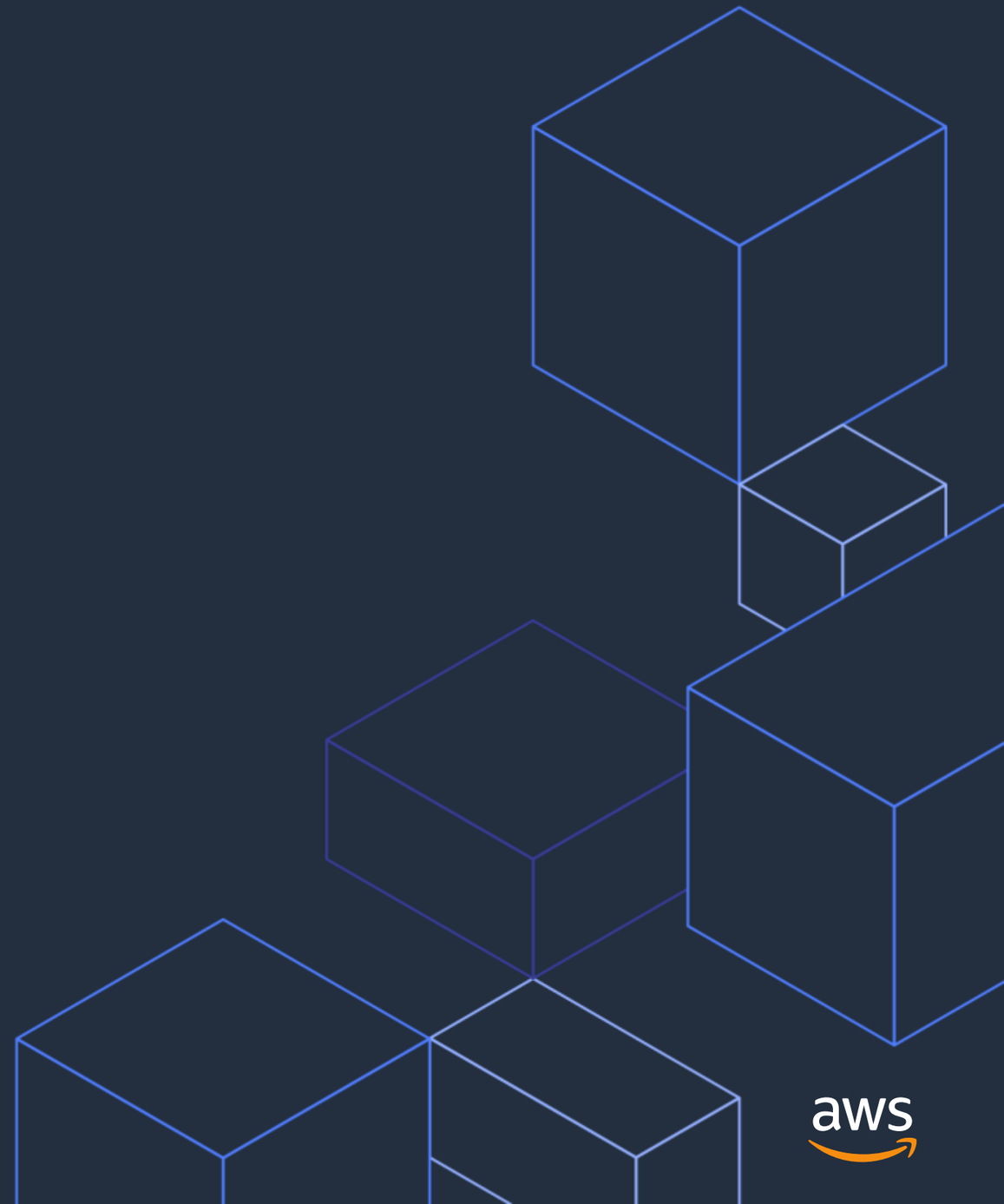
- データレイクは作って終わりではない
  - データレイクは活用されて初めて意味がある

- 適切に運用されていないデータレイクは、まともには使えない
  - 更新されていないデータは価値を失っていく
  - クエリの遅いデータレイクは不便で、急ぎの判断には使えない
  - 信頼性がチェックされていないデータは判断の根拠に使えない

データレイクの活用には適切な運用が欠かせない

aws

データレイクの運用



データレイク、どう運用したらいい？

データレイク、うまいこと運用しといてね

上司

データレイク管理者
（初心者）

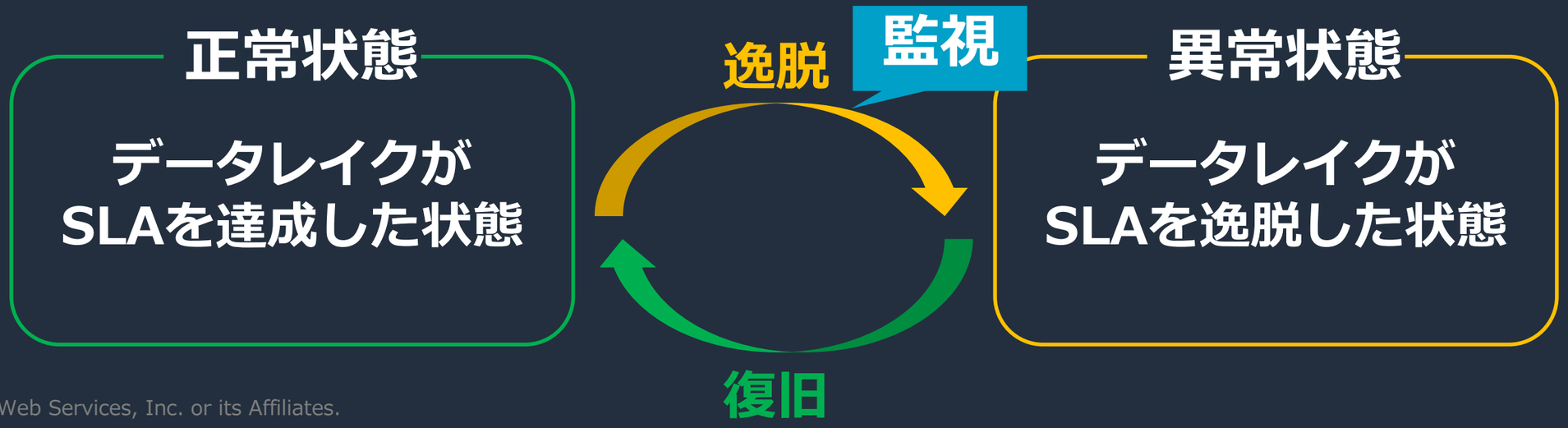
**データレイクといってもITシステムだし
CPU使用率とかディスク使用率とか
従来通りのサーバー監視を
設定しておけばいいよね・・・？**

aws

データレイクの運用のポイント

正常状態を定義し、その状態を維持する

- データレイクがどのような状態であれば正常かを定義する
    - 正常性指標 (SLA: Service Level Agreement)
- データレイクが正常である状態を維持し、逸脱した場合は正常に戻す



データレイクの正常性定義と SLA の例

管理対象	正常状態	SLA の例
データの鮮度	データ発生から利用可能になるまでの時間が規定内	アプリ上で出力されたログデータが**XX時間以内に利用可能**となっている
	所定の日時までに対象データが利用可能となる	レポーティング用のデータの前処理が**午前xx時まで**に完了している
データの信頼性	データ起因のエラーや結果不正の発生率が規定内	分析クエリの**エラーレートがXX%以下**におさまっている
データレイクのコスト	データレイク全体のコストが規定内	ストレージ、ETL、クエリ等の**総コストが$XX以下**におさまっている

COVID-19 データレイクの場合

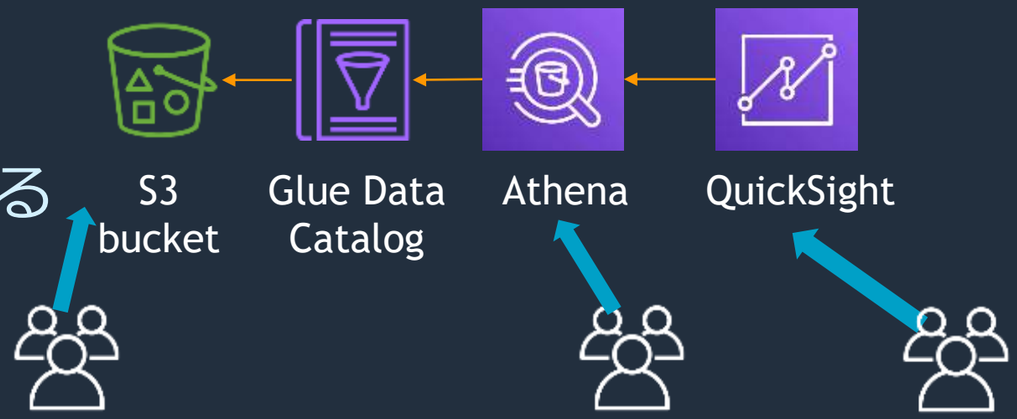
データレイクの各データが
直近 2日以内のデータに更新された状態

データレイクの正常性 SLA の監視

- ユーザーからの End-to-End のアクセスパターンに対するデータレイクの動作の正常性をチェックする
  - **実ユーザーのアクセス**の記録をチェックして判定
  - 監視用にユーザーアクセスを**定期的にシミュレート**して判定

COVID-19 データレイクの場合

異なるアクセスパターンがある
- S3バケットに直接アクセスする
- Athena (EMR/Redshift) 経由でクエリする
- QuickSight ダッシュボードを閲覧する



データレイクの正常性 SLA の監視

- ユーザーカ
  データレイ
  - **実エ**
  - 監視

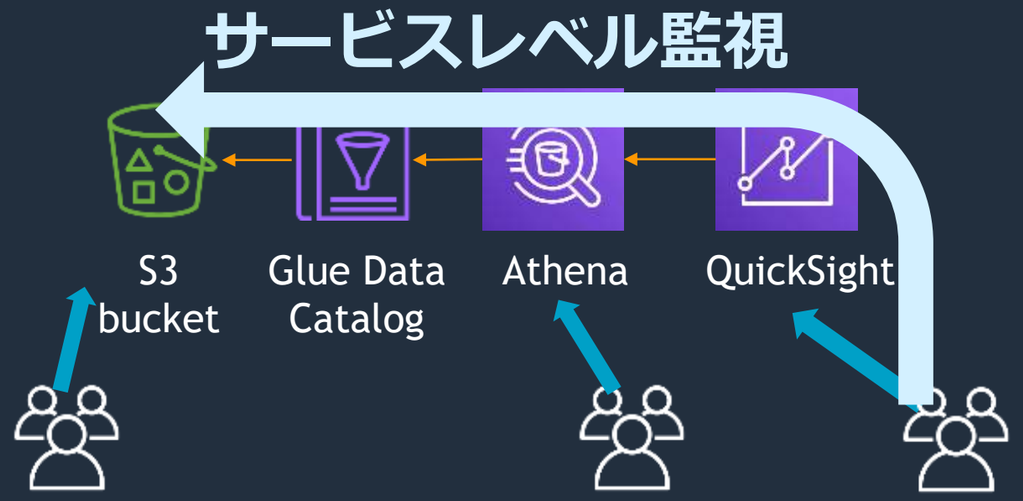
New query 1 +

```
1 SELECT s.date
2 FROM "covid-19"."covid_testing_us_daily" u
3 left outer join "covid-19"."covid_testing_states_daily" s on u.date=s.date
4 left outer join "covid-19"."us_state_abbreviations" abb on s.state=abb.abbreviation
5 GROUP BY s.date
6 ORDER BY s.date DESC
7 LIMIT 1
```

Run query | Save as | Create | (Run time: 1.86 seconds, Data scanned: 1.62 MB) | Format query | Clear

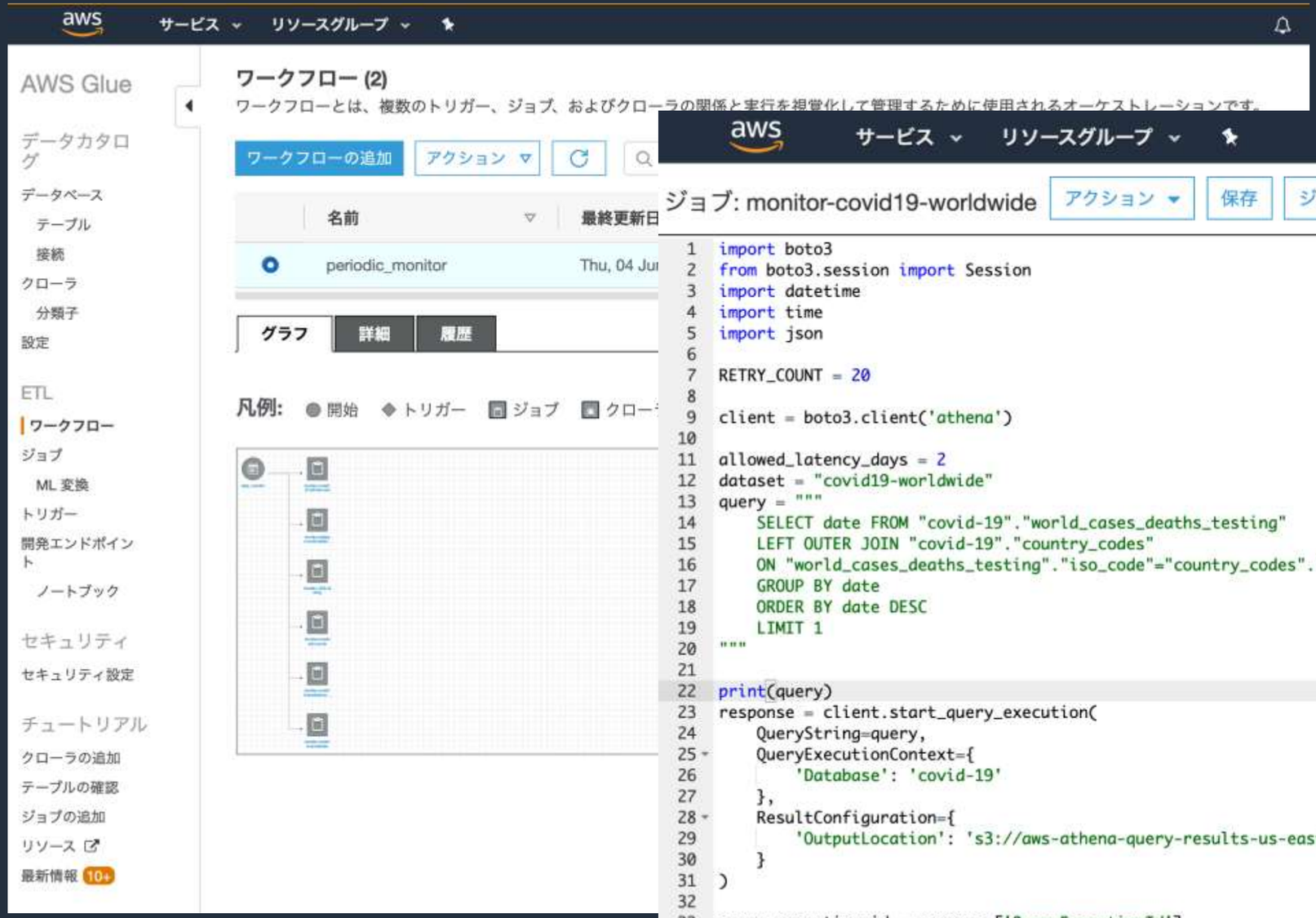
COVID-19 データレイクの場合

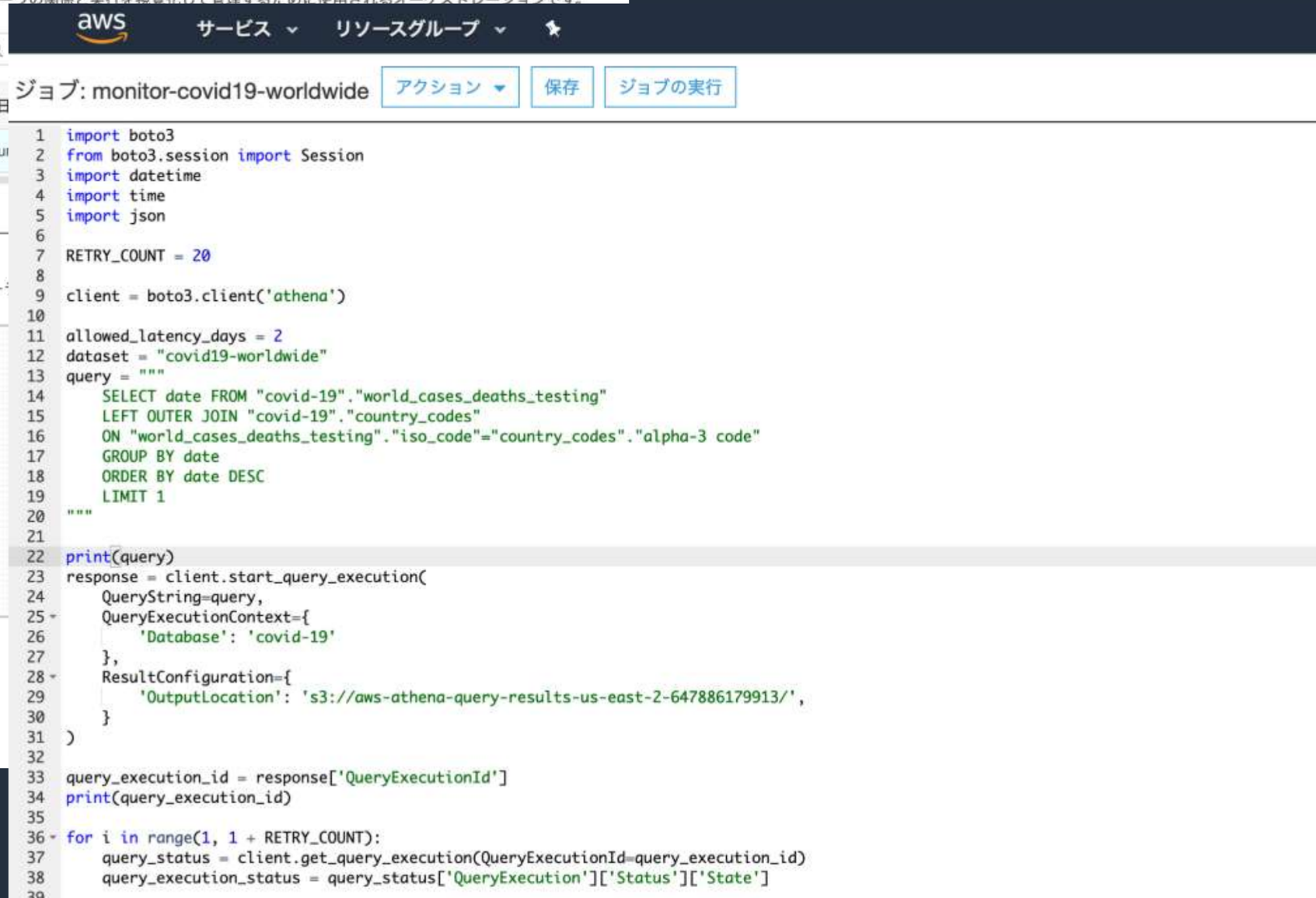
QuickSight へのデータロード時に
Athena に対して発行されるクエリと
同等のクエリをシミュレート実行して
データの更新日時と現在日時を比較



aws

Tips：複数の Athena クエリの定期実行





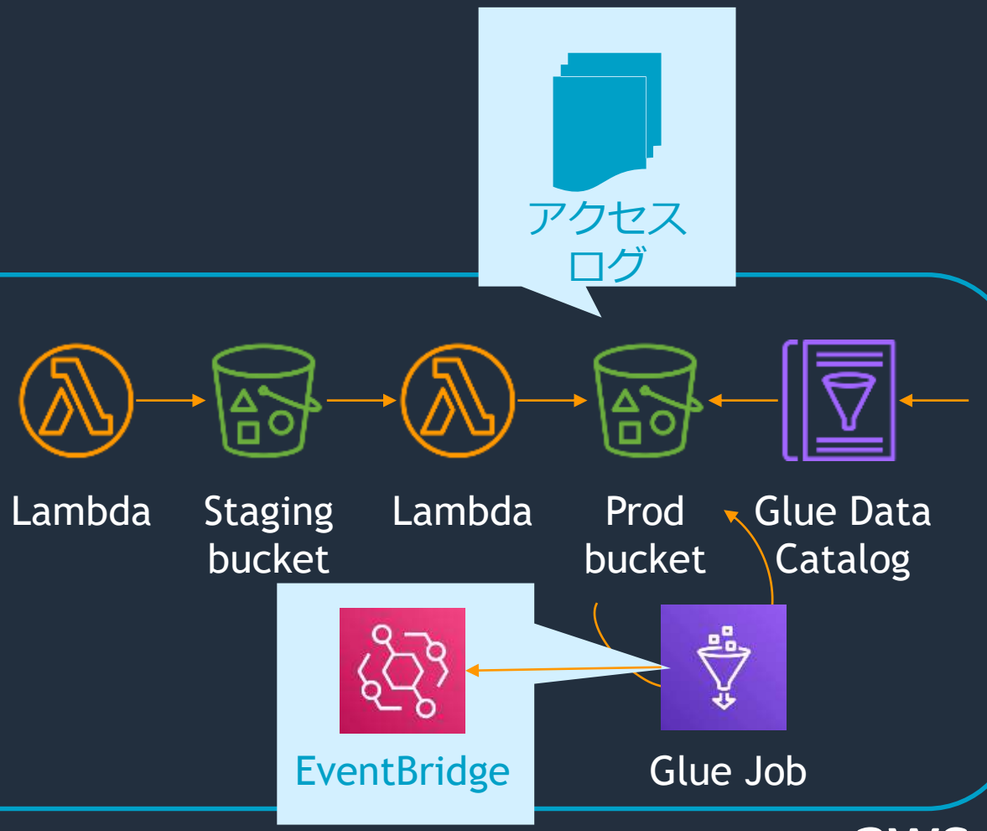

```python
import boto3
from boto3.session import Session
import datetime
import time
import json

RETRY_COUNT = 20

client = boto3.client('athena')

allowed_latency_days = 2
dataset = "covid19-worldwide"
query = """
    SELECT date FROM "covid-19"."world_cases_deaths_testing"
    LEFT OUTER JOIN "covid-19"."country_codes"
    ON "world_cases_deaths_testing"."iso_code"="country_codes"."alpha-3 code"
    GROUP BY date
    ORDER BY date DESC
    LIMIT 1
"""

print(query)
response = client.start_query_execution(
    QueryString=query,
    QueryExecutionContext={
        'Database': 'covid-19'
    },
    ResultConfiguration={
        'OutputLocation': 's3://aws-athena-query-results-us-east-2-647886179913/',
    }
)

query_execution_id = response['QueryExecutionId']
print(query_execution_id)

for i in range(1, 1 + RETRY_COUNT):
    query_status = client.get_query_execution(QueryExecutionId=query_execution_id)
    query_execution_status = query_status['QueryExecution']['Status']['State']
```

データレイクの副次的な正常性監視：異常傾向の検知

- ユーザーからのアクセスの異常を示すイベントを検知する
  - アクセス数の異常低下
  - データ収集や ETL ジョブの失敗

COVID-19 データレイクの場合

- S3 バケットへのアクセス数を S3 サーバーアクセスログで計測、異常検知
- ジョブの失敗を EventBridge で検知



異常状態から正常状態への復旧

- 正常状態を逸脱した場合、何らかの手段で正常状態に戻す
  - リトライ
  - キャパシティ追加
  - 管理者による操作 (データ収集、ETL パイプラインの修正など)

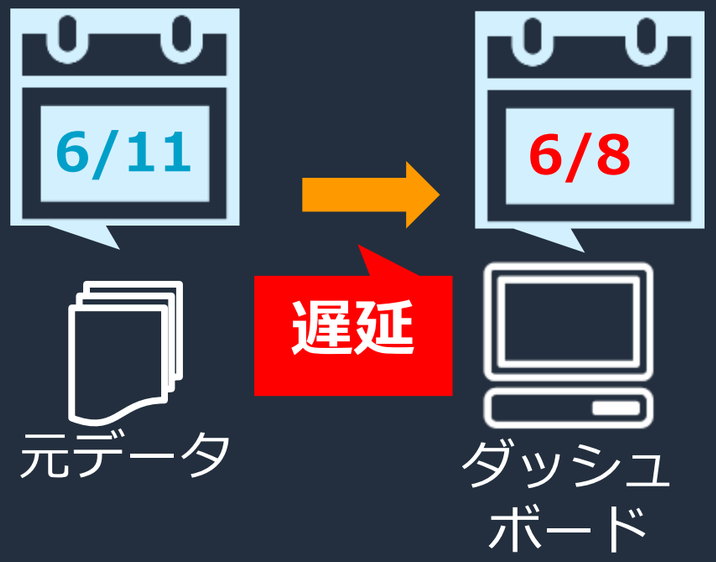
COVID-19 データレイクの場合

- データが利用可能になるまでの時間の SLA に抵触した場合にメール通知し管理者により対応

[covid19-lake-monitor]
Data latency increase detected

aws

例1：データ鮮度の低下



原因

- データ収集・配信の遅延
- ETL パイプラインの遅延・失敗

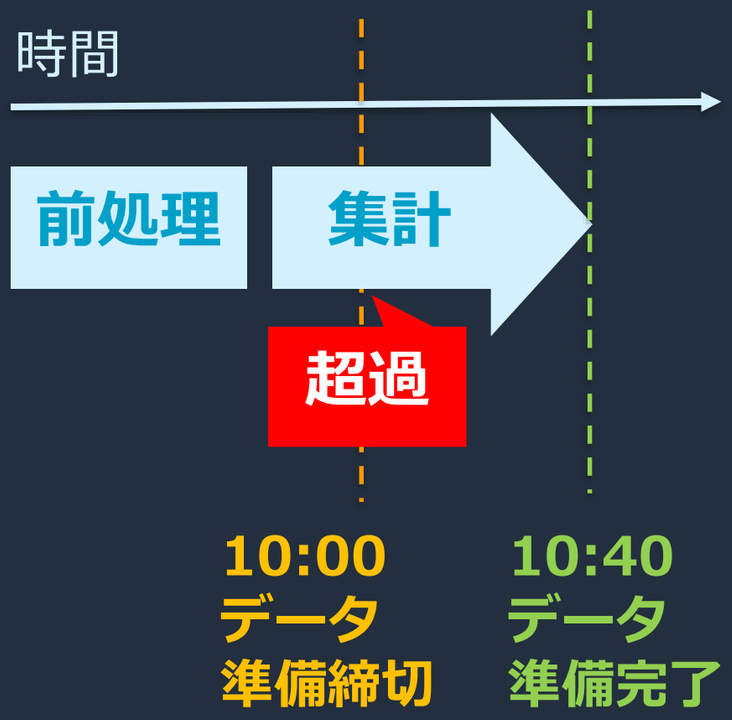
対応方法

- データ収集・配信の問題の修正
- ETL パイプラインの高速化・安定化

COVID-19 データレイクの場合

- 同時実行数超過による ETL ジョブ実行失敗 →**ETLパイプラインの修正**
- データ提供元からのファイル名の変更により、更新データが ETL から漏れる →**ETLパイプラインの修正**
- データ提供元からのデータ配信の停止 →**データ提供元との交渉**

例2：規定時間の超過



原因

- ETL パイプラインの遅延・失敗
- データ量、種類の増大

対応方法

- ETL パイプライン、データのチューニング
  - データフォーマットの変換
  - データ配置の最適化
  - データ量の削減 (保存期間の調整, アーカイブ化)
  - キャパシティ増強

COVID-19 データレイクの場合

→データの最新リビジョンのみを保持
(データ量がそこまで大きくないためこの種の問題は起きていない)

Tips：ETL パイプラインの最適化

小規模処理

AWS Lambda

- 実行時間に最大15分間の制限あり
- 豊富なトリガーを持ち、S3に配置されたタイミングで逐次処理することも可能
- Python以外の言語も選択可能

中規模処理

AWS Glue
Python Shell

データを加工するコードのみを実装
Python であれば NumPy, Pandas などのライブラリも利用可能

- 実行時間の制限なし
- Lambda に比べて利用できるメモリ量が多い（1GBまたは16GB）
- Athena、Redshift、EMR に対するSQLベースの処理も可能

大規模処理

AWS Glue
Spark

PySpark や Scalaで実装
必要なものはコードのみ

- 実行時間の制限なし
- 複数のワーカーで並列分散処理
- 数100GB以上の大量データ処理も可能
- DPU 設定によりスケーラビリティを調整可能

aws

Tips：データのパフォーマンス最適化

- ファイルフォーマット変換・圧縮
  - テキストベースのファイル (JSON, CSV等) を、分析に適したカラムナフォーマット (Parquet, ORC) に変換
  - 非圧縮ファイルを、分析に適した形式で圧縮
- 集約 (コンパクション)
  - ファイル処理のオーバーヘッドを小さくするために、大量の小さいファイルを 128MB程度のまとまりに集約
- パーティショニング
  - 単一ディレクトリにフラットに配置したファイルを、日付で区切ったディレクトリに保存

aws

Tips：保存期間の調整、アーカイブ化

- S3 ライフサイクル設定例
  - **1年経過**
    →ストレージクラスを
    Standard-IA に変更
  - **5年経過**
    →削除

例3：意図せぬクエリ結果

ダッシュボード

原因

- データ収集・配信・ETL パイプラインの遅延・失敗
- データ内容の変化、意図せぬ削除
- アクセスパターンの変化

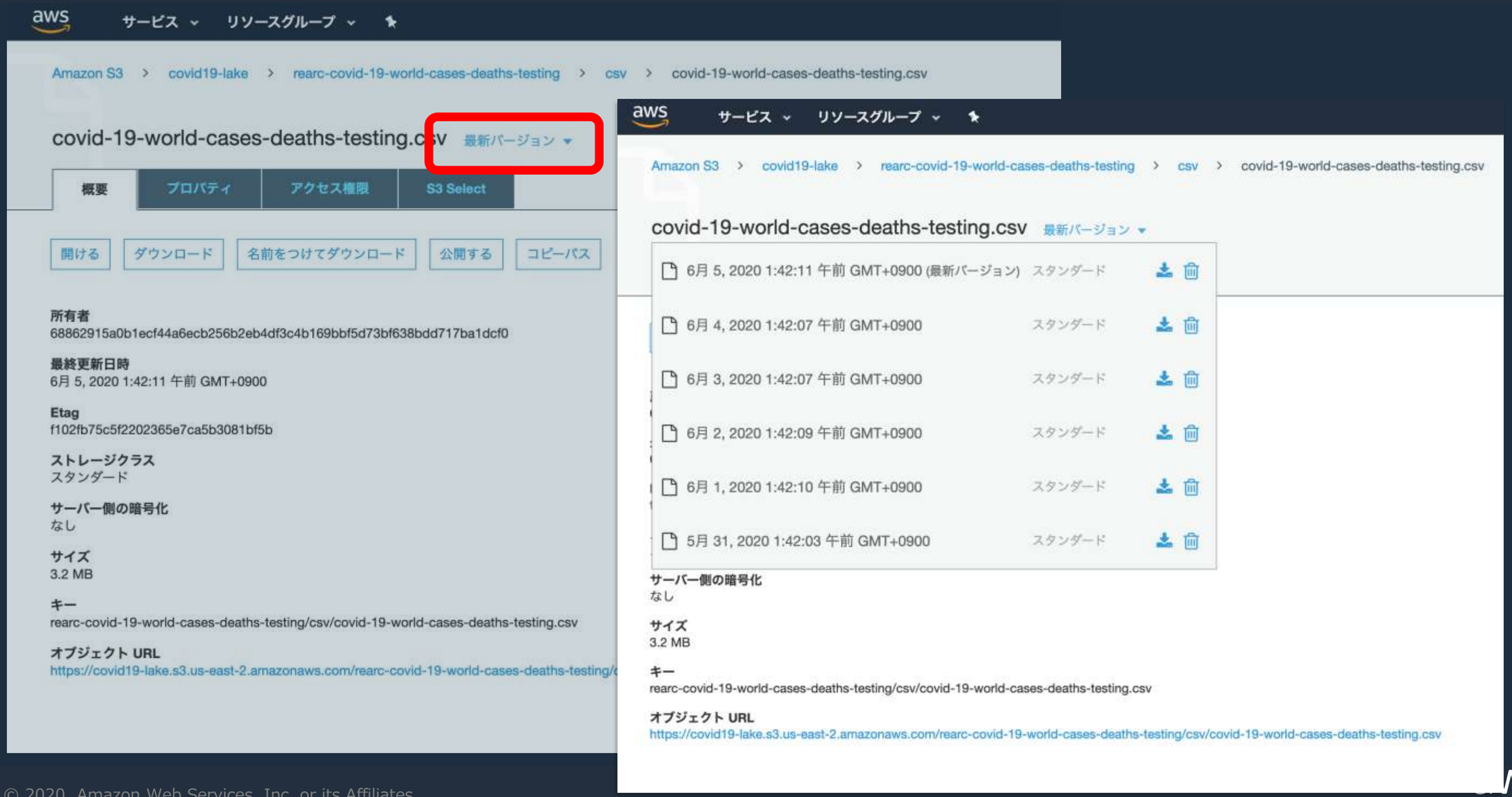
対応方法

- データ収集・配信の問題の修正
- ETL パイプラインの修正
- データのリストア

COVID-19 データレイクの場合

- 収集したデータ内容の変更により想定したデータが含まれなくなった
  →**データ収集方法の修正 (API, データソース等)**
- S3 バケットに保存される過去バージョンのファイルからリストア

Tips：S3 バケットのデータリストア



利用状況の計測と
データレイクの継続的改善

データレイク、誰がどのくらい使ってる？

直近の1週間で
利用量がトップ20のユニークユーザーと
テーブルごとのアクセス数を
分析しといてくれる？

上司

!?

データレイク管理者
（初心者）

困った・・
何も記録してないぞ・・・

aws

利用状況の計測

継続的改善にはユーザー利用状況の計測が必須

アクティビティの種類と計測方法

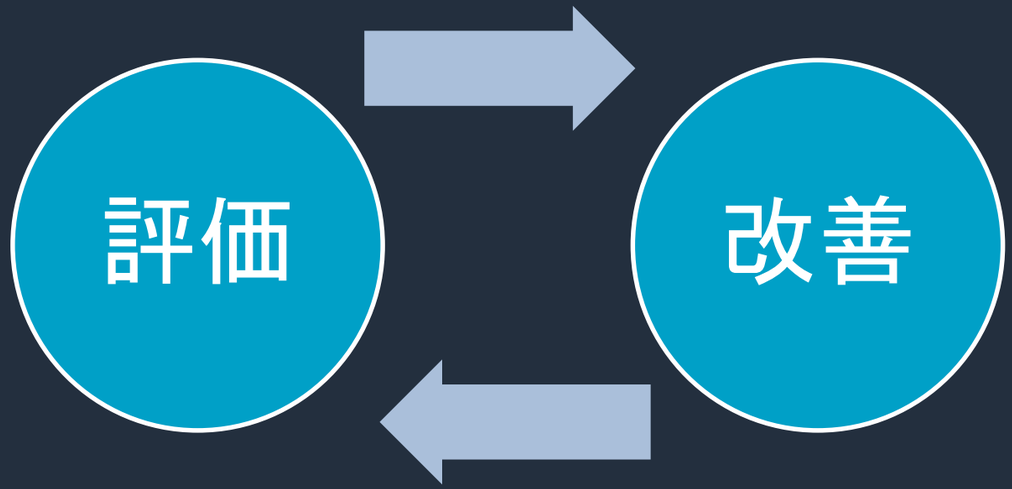
- データ参照: S3 サーバーアクセスログ、CloudTrail 等
- クエリ実行: クエリログ

COVID-19 データレイクの場合

- S3 サーバーアクセスログを整形
- ユーザーアクセスログ分析用の QuickSight ダッシュボードを作成

データレイクの継続的改善

データレイクは塩漬けにせず継続的に進化させることが大切



- 利用状況の分析
- フィードバックの収集

- 収集対象データの拡充
- 利用エンジンの拡充
- ドキュメント整備
- 利用ポリシーの改善

aws

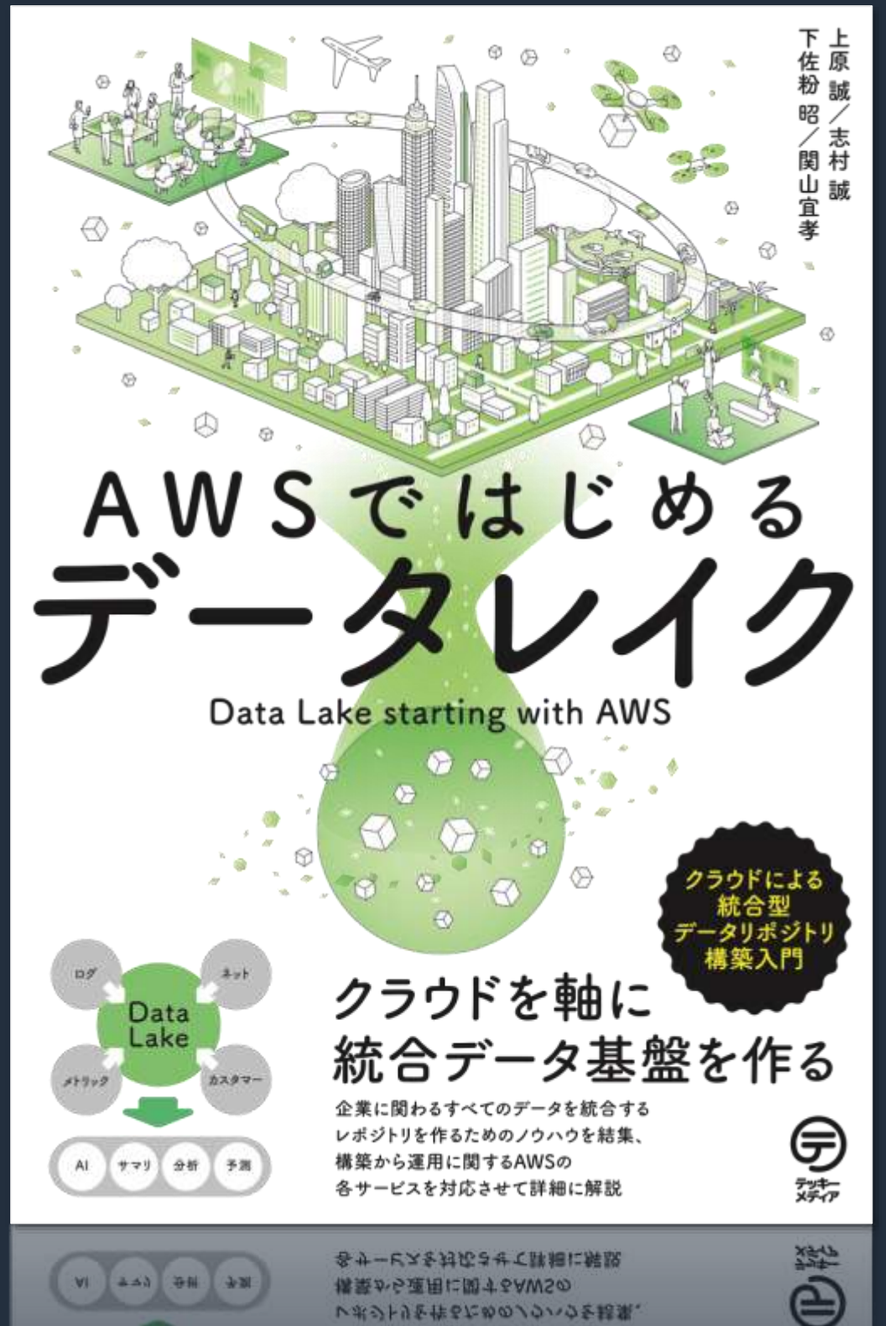
まとめ

- データレイクの運用
  - 正常状態を定義し、SLA を遵守する
  - ユーザーからの End-to-End のアクセスパターンに対するデータレイクの動作の正常性をチェックする

- 利用状況の計測とデータレイクの継続的改善
  - 継続的な改善には利用状況の計測が必須
  - データレイクは塩漬けにせず継続的に進化させることが大切

aws

「AWSではじめるデータレイク」
2020年7月発売予定



内容についての注意点

- 本資料では2020年6月11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には別途消費税をご請求させていただきます。

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

Q&A

関連リンク

AWSではじめるデータレイク 出版記念 オンラインセミナー

- 第１回 データレイクはじめの一歩
    - 資料: https://go.aws/3df1Ebo

- 第２回 データレイクの構築と蓄積されたデータの活用方法
    - 資料: https://go.aws/3cxGvbe

関連セミナーのご案内

AWSではじめるデータレイク 出版記念 オンラインセミナー

- 第4回 Glue, Lake Formation, Athena, EMR 最新アップデート
  - 2020 年 6 月 18 日（木）11:00 ～ 12:00
  - https://aws-seminar.smktg.jp/public/seminar/view/2674