AWS Glue for Ray 入門

Tomoya Okuno

Amazon Web Services Japan
Solutions Architect

アジェンダ

- AWS Glueの概要と課題
- AWS Glue for Ray のご紹介
- デモ

AWS Glueの概要と課題

AWS Glue とは

サービス間でデータを簡単に移動できるようにするための、サーバレス データ統合サービス

大量データの迅速な統合

データの準備を数ヶ月から数分に短縮

変換処理の自動化

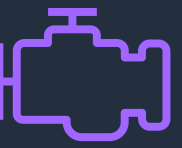
何千ものETLジョブを簡単に実行、管理可能

サーバーレス

処理に使われたリソースに対してのみ支払い

AWS Glue の全体像

Apache Spark / Python Shell Engine の課題



処理エンジンにApache SparkとPython Shellが存在した
分散処理用のApache Spark or シングルノードの Python Shell

Python Shell ではジョブごとにシングルノードでの処理しかできない
V.S.
マルチノード分散クラスターでジョブを実行するにはSparkの学習コストがかかる

AWS Glue for Ray のご紹介

AWS Glue for Ray

Python の分散処理によるサーバレスなデータ統合

数百ノードへ
スケール

高速なスケールアップと
スケールダウン

Python の記法や使い慣れた
フレームワーク (pandasなど)
で処理を記述し
実行はAWS Glueへお任せ

“Python is the preferred coding language for many of our data scientists, but they’ve had challenges with moving from sample datasets to larger datasets. AWS Glue for Ray lets our data scientists use their existing Python code at scale.”

Roberto Figueira

Head of D&A Platform Engineering at Itaú Unibanco

Rayとは

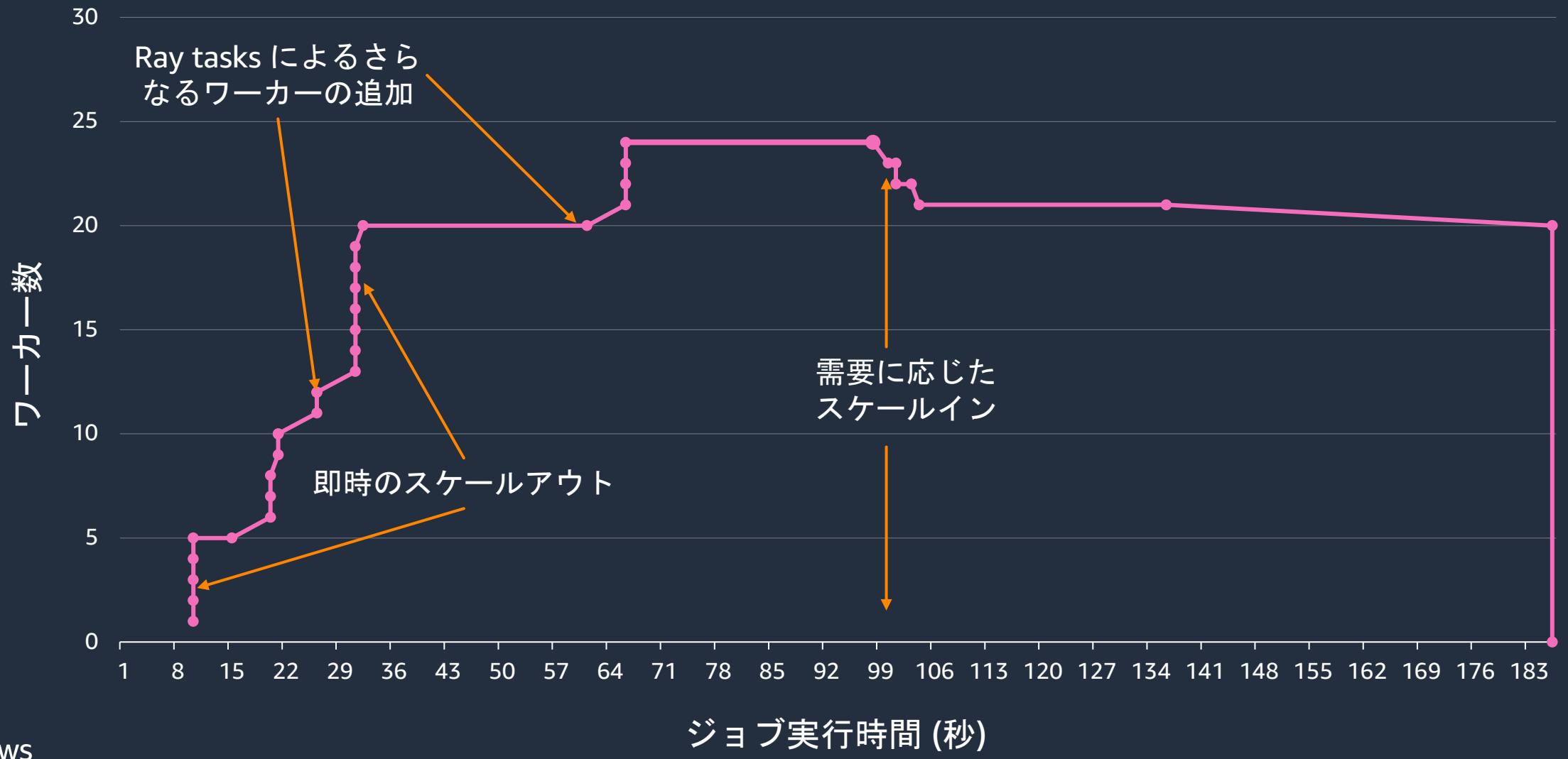
AI および Python アプリケーションの
スケーリングを容易にする
オープンソースのコンピューティングフレームワーク

Ray.io に詳細が記載

AWS Glue for Ray の特徴

- 現在、プレビュー中
- Ray Core (分散処理機能)、Ray Dataset (データの格納) 、Modin (pandas をRayで実行するためのもの) が利用できる環境を提供
  - Ray のその他機能は AWS Glue のサービススコープ外のため未サポート
- 8 vCPU、 64 GB メモリの新しい Graviton2 ベースのワーカー (Z.2x) のみ
- 最小と最大ワーカーの数を設定し自動スケーリングでジョブ実行
- オハイオ、バージニア北部、オレゴン、東京、アイルランドリージョンで利用可能
- 現在 Private Subnet にあるデータベースや Amazon Redshift などには接続不可

Ray ジョブの自動スケーリング



デモ

まとめ

AWS Glue for Ray

Python の分散処理によるサーバレスなデータ統合

数百ノードへ
スケール

高速なスケールアップと
スケールダウン

Python の記法や使い慣れた
フレームワーク (pandasなど)
で処理を記述し
実行はAWS Glueへお任せ