アーキテクチャ図

サーバーレスデータレイクでの挿入とアップサートの管理

サーバーレスデータレイクでの挿入とアップサートの管理: アーキテクチャ図

Table of Contents

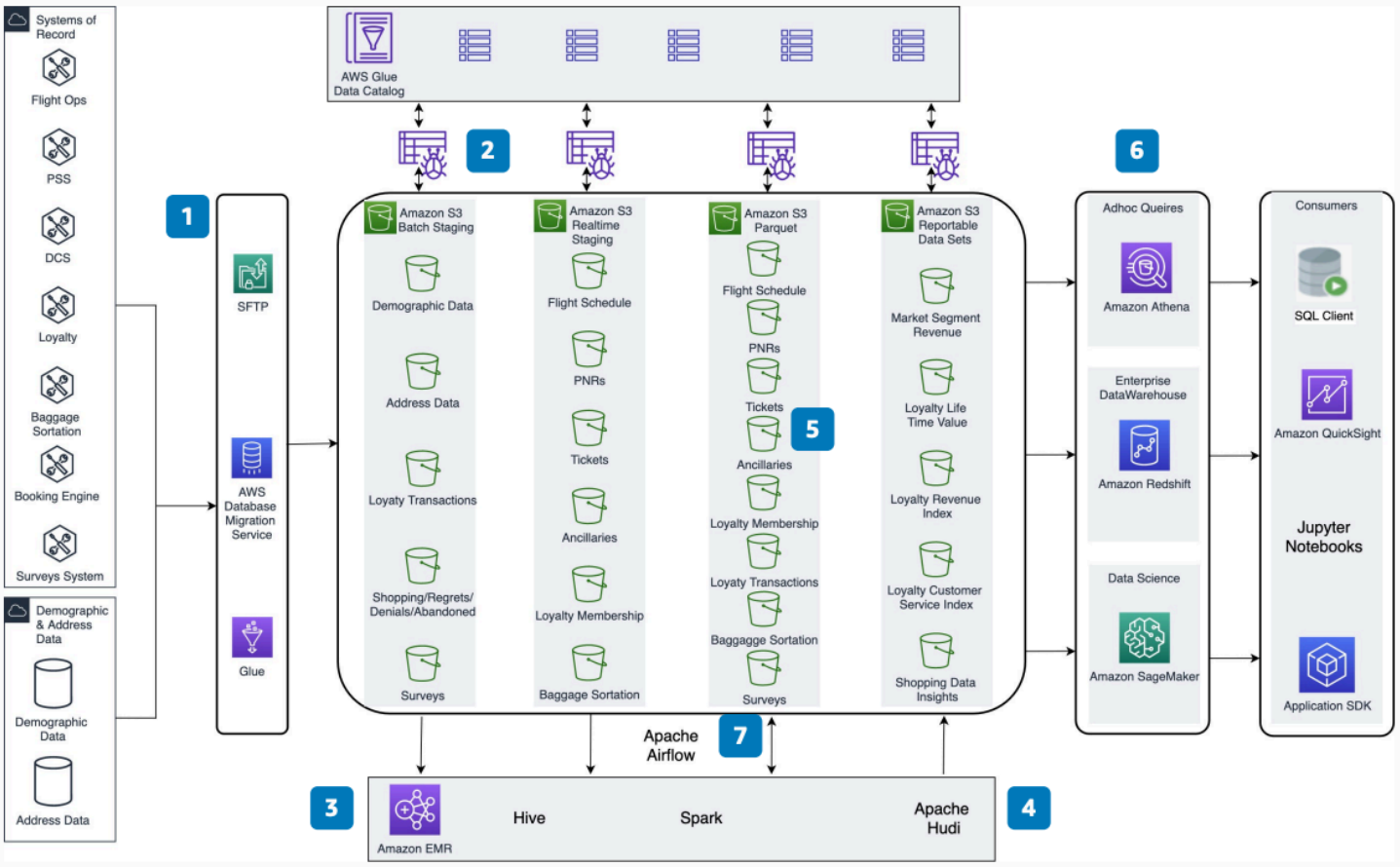
Data Lake の挿入とアップサート i
サーバーレスデータレイクでの挿入とアップサートの管理 1
詳細情報 2
図の履歴 2
..... iii

サーバーレスデータレイクでの挿入とアップサートの管理

公開日: 2021 年 6 月 15 日 (図の履歴)

このアーキテクチャは、Amazon EMR でApache Hudi実行されている を使用して、Amazon Simple Storage Service のデータセットへの挿入と更新を処理する方法を示しています。オンデマンド分析をプロビジョニングし、永続的なデータマートを作成する、費用対効果が高くスケーラブルなデータレイクを構築できます。

サーバーレスデータレイクでの挿入とアップサートの管理



次の手順では、アーキテクチャについて説明します。

1. バッチ、変更データキャプチャ (CDC)、または Amazon S3 の raw レイヤーへのストリーミングを使用して、ソースシステムからデータを取り込みます。
2. Amazon S3 の raw データレイクにデータが保持されたら、データをクロールし、クローラーを使用して AWS Glue データカタログに入力します。

3. raw データを Amazon EMR クラスターにプルし、Hive と Spark を使用して読み取り、クリーニングと変換を行います。
4. Apache Hudi Amazon EMR で を実行すると、Spark APIs を使用してデータが読み取り、必要なデータセットの挿入とアップサートが実行されます。
5. クリーンアップおよび変換されたデータを Amazon S3 で処理および報告可能なバケットに保持します。
6. Athena を使用してレポート対象データをオンデマンドで消費するか、Amazon Redshift にロードします。さまざまなユーザー、ツール、リソースがこのデータを消費できます。
7. 完全なデータ移動を処理し、バッチデータ用にオンデマンド Amazon EMR クラスターをスピンし、Amazon Managed Workflows for Apache Airflow (MWAA) によるワークフローオーケストレーションを使用してデータをロードします。

詳細情報

詳細については、次のリソースを参照してください。

- AWS アーキテクチャアイコン
- AWS アーキテクチャセンター
- AWS Well-Architected

図の履歴

このリファレンスアーキテクチャ図の更新について通知を受け取るには、RSS フィードにサブスクライブします。

変更	説明	日付
初版発行	リファレンスアーキテクチャ図が最初に公開されました。	2021 年 6 月 15 日

> **Note**
>
> RSS 更新をサブスクライブするには、使用しているブラウザで RSS プラグインが有効になっている必要があります。

翻訳は機械翻訳により提供されています。提供された翻訳内容と英語版の間で齟齬、不一致または矛盾がある場合、英語版が優先します。