



Diagrammi di architettura

Gestione di inserti e upsert in un data lake serverless



Gestione di inserti e upsert in un data lake serverless: Diagrammi di architettura

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

I marchi e l'immagine commerciale di Amazon non possono essere utilizzati in relazione a prodotti o servizi che non siano di Amazon, in una qualsiasi modalità che possa causare confusione tra i clienti o in una qualsiasi modalità che denigri o discrediti Amazon. Tutti gli altri marchi non di proprietà di Amazon sono di proprietà dei rispettivi proprietari, che possono o meno essere affiliati, collegati o sponsorizzati da Amazon.

Table of Contents

Inserti e upsert di Data Lake i

Gestione di inserti e upsert in un data lake serverless 1

Approfondimenti 2

Storia del diagramma 2

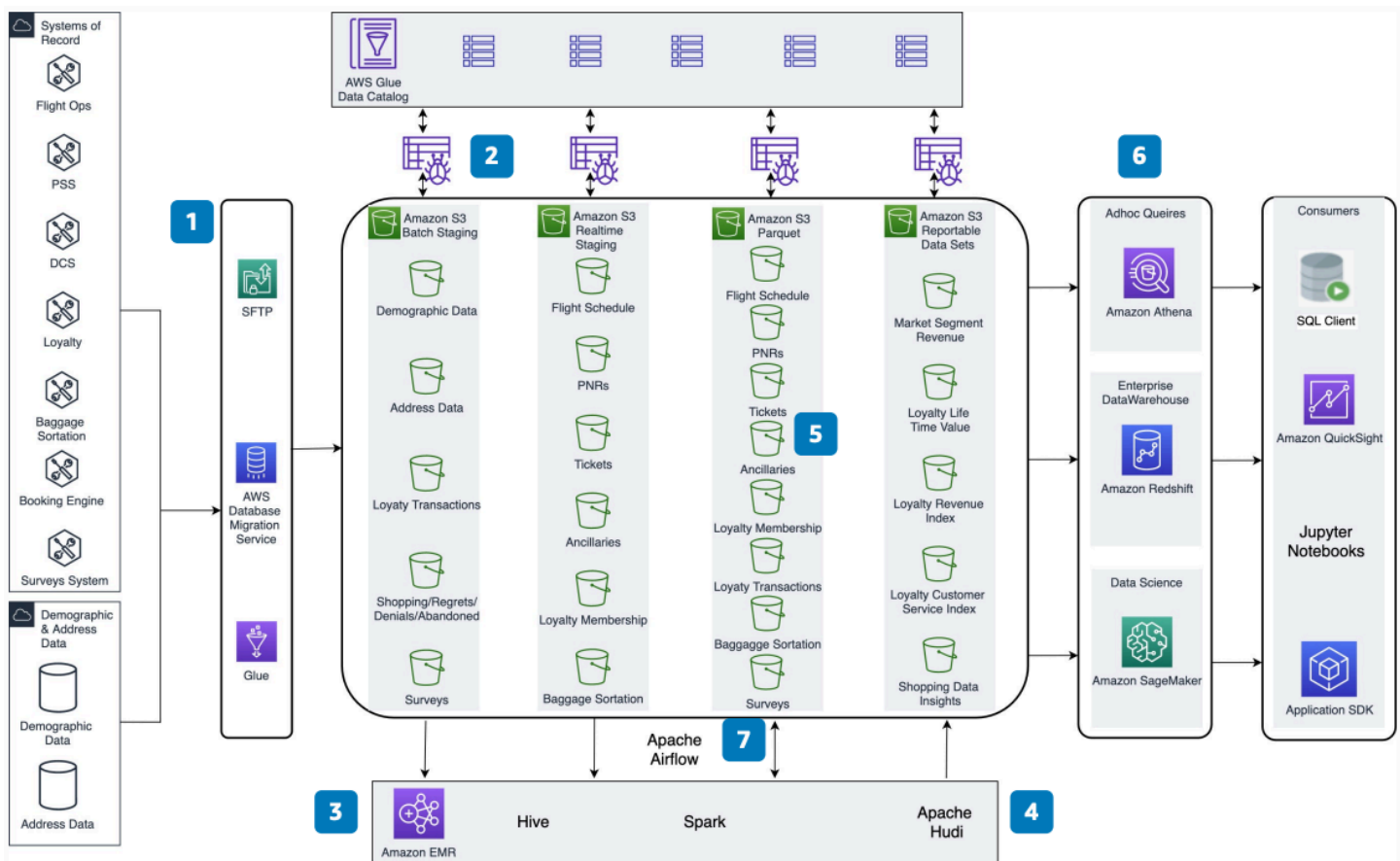
..... iv

Gestione di inserti e upsert in un data lake serverless

Data di pubblicazione: 15 giugno 2021 () [Storia del diagramma](#)

Questa architettura mostra come utilizzare l'Apache Hudi esecuzione su Amazon EMR per elaborare inserimenti e aggiornamenti ai set di dati in [Amazon Simple Storage Service](#). Puoi creare un data lake scalabile e conveniente che fornisce analisi su richiesta e crea data mart persistenti.

Gestione di inserti e upsert in un data lake serverless



I passaggi seguenti descrivono l'architettura:

1. Inserisci dati dai sistemi di origine utilizzando batch, Change Data Capture (CDC) o streaming nel livello raw in Amazon S3.
2. Dopo che i dati sono rimasti nel data lake non elaborato su Amazon S3, esegui la scansione dei dati e inseriscili nel catalogo dati di [AWS Glue utilizzando un crawler](#).

3. Estrai i dati grezzi in un cluster Amazon EMR e leggili utilizzando Hive e Spark per la pulizia e la trasformazione.
4. Apache Hudi in esecuzione su Amazon EMR legge i dati utilizzando le API Spark ed esegue inserimenti e upsert sui set di dati richiesti.
5. Conserva i dati puliti e trasformati nei bucket elaborati e segnalabili di Amazon S3.
6. Utilizza i dati segnalabili su richiesta utilizzando [Athena](#) o caricali in Amazon Redshift. <https://docs.aws.amazon.com/redshift/latest/dg/welcome.html> Utenti, strumenti e risorse diversi possono utilizzare questi dati.
7. Gestisci lo spostamento completo dei dati, crea cluster Amazon EMR on-demand per dati batch e carica i dati utilizzando l'orchestrazione dei flussi di lavoro con Amazon Managed Workflows for (MWAA). Apache Airflow

Approfondimenti

Per ulteriori informazioni, consulta le seguenti risorse:

- [AWS Icone dell'architettura](#)
- [AWS Centro di architettura](#)
- [AWS Well-Architected](#)

Storia del diagramma

Per ricevere notifiche sugli aggiornamenti di questo diagramma di architettura di riferimento, iscriviti al feed RSS.

Modifica	Descrizione	Data
Pubblicazione iniziale	Diagramma dell'architettura di riferimento pubblicato per la prima volta.	15 giugno 2021

 Note

Per iscriverti agli aggiornamenti RSS, devi avere un plugin RSS abilitato per il browser che stai utilizzando.

Le traduzioni sono generate tramite traduzione automatica. In caso di conflitto tra il contenuto di una traduzione e la versione originale in Inglese, quest'ultima prevarrà.