



Diagramas de arquitectura

Administración de inserciones y actualizaciones en un lago de datos sin servidor



Administración de inserciones y actualizaciones en un lago de datos sin servidor: Diagramas de arquitectura

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

Las marcas comerciales y la imagen comercial de Amazon no se pueden utilizar en relación con ningún producto o servicio que no sea de Amazon, de ninguna manera que pueda causar confusión entre los clientes y que menosprecie o desacredite a Amazon. Todas las demás marcas registradas que no son propiedad de Amazon son propiedad de sus respectivos propietarios, que pueden o no estar afiliados, conectados o patrocinados por Amazon.

Table of Contents

Inserciones y actualizaciones de data lake i

Administración de inserciones y actualizaciones en un lago de datos sin servidor 1

Documentación adicional 2

Historial del diagrama 2

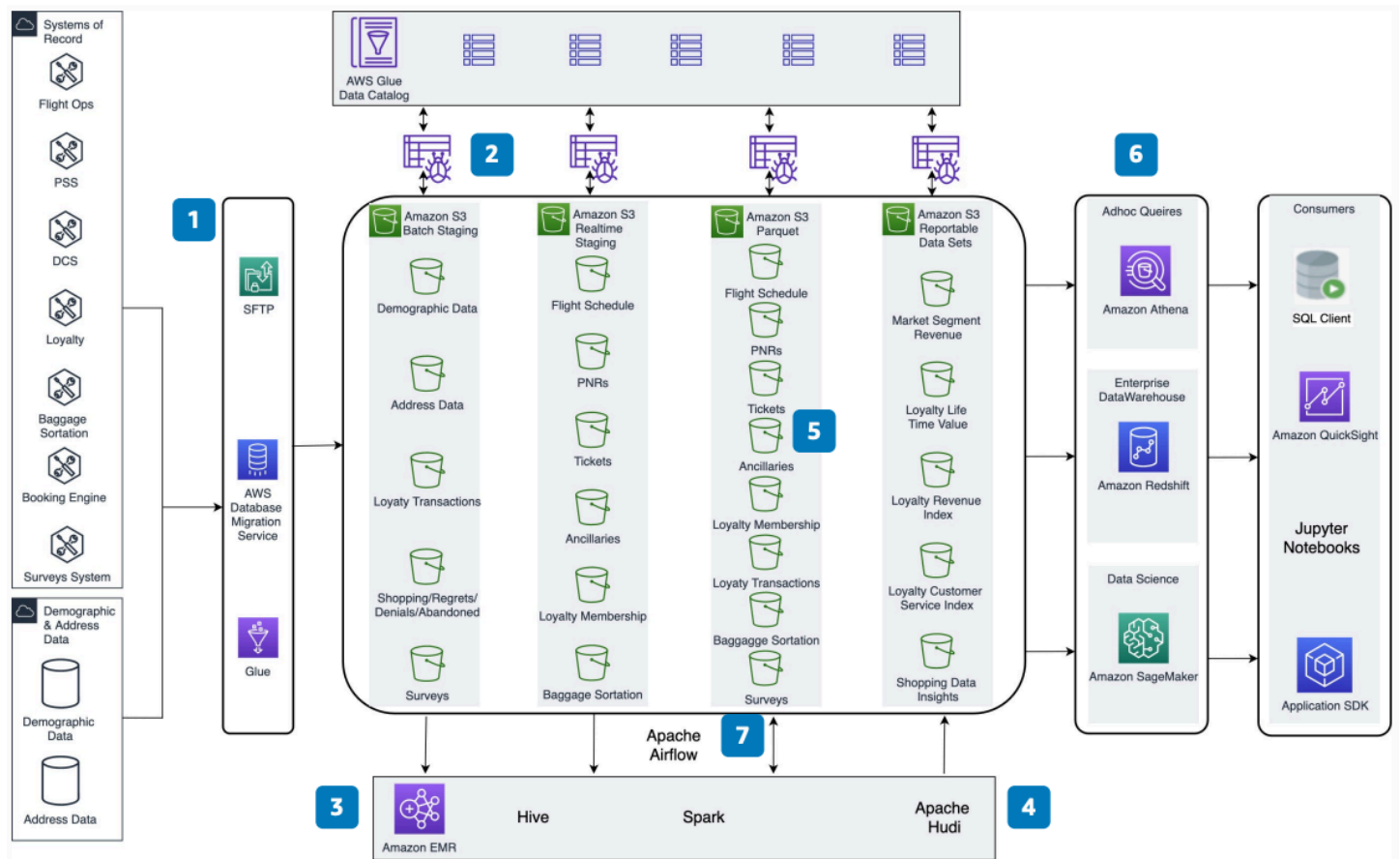
..... iv

Administración de inserciones y actualizaciones en un lago de datos sin servidor

Fecha de publicación: 15 de junio de 2021 () [Historial del diagrama](#)

Esta arquitectura muestra cómo utilizar la Apache Hudi ejecución en Amazon EMR para procesar las inserciones y actualizaciones de los conjuntos de datos en [Amazon Simple Storage Service](#). Puede crear un lago de datos rentable y escalable que suministre análisis bajo demanda y cree mercados de datos persistentes.

Administración de inserciones y actualizaciones en un lago de datos sin servidor



Los siguientes pasos describen la arquitectura:

1. Ingiera datos de los sistemas de origen mediante lotes, captura de datos cambiados (CDC) o transmitiéndolos a la capa sin procesar de Amazon S3.

2. Una vez que los datos persistan en el lago de datos sin procesar de Amazon S3, rastree los datos y rellénelos en el catálogo de [datos de](#) AWS Glue con un rastreador.
3. Extraiga datos sin procesar en un clúster de Amazon EMR y léalos con Hive y Spark para limpiarlos y transformarlos.
4. Apache Hudi se ejecuta en Amazon EMR, lee los datos mediante las API de Spark y realiza inserciones y modificaciones en los conjuntos de datos necesarios.
5. Conserva los datos limpiados y transformados en los buckets procesados y reportables de Amazon S3.
6. Consuma los datos declarables bajo demanda con [Athena](#) o cárguelos en Amazon Redshift. <https://docs.aws.amazon.com/redshift/latest/dg/welcome.html> Diferentes usuarios, herramientas y recursos pueden consumir estos datos.
7. Gestione todo el movimiento de datos, genere clústeres de Amazon EMR bajo demanda para obtener datos por lotes y cargue los datos mediante la organización del flujo de trabajo con Amazon Managed Workflows for Apache Airflow (MWAA).

Documentación adicional

Para obtener información adicional, consulte los siguientes recursos:

- [AWS Iconos de arquitectura](#)
- [AWS Centro de arquitectura](#)
- [AWS Well-Architected](#)

Historial del diagrama

Para recibir notificaciones sobre las actualizaciones de este diagrama de arquitectura de referencia, suscríbase a la fuente RSS.

Cambio	Descripción	Fecha
Publicación inicial	El diagrama de arquitectura de referencia se publicó por primera vez.	15 de junio de 2021

 Note

Para suscribirse a las actualizaciones de RSS, debe tener un complemento de RSS habilitado para el navegador que esté utilizando.

Las traducciones son generadas a través de traducción automática. En caso de conflicto entre la traducción y la version original de inglés, prevalecerá la version en inglés.