✦ Member-only story

Building Data Lakes on AWS with Kafka Connect, Debezium, Apicurio Registry, and Apache Hudi

Learn how to build a near real-time transactional data lake on AWS using a combination of Open Source Software (OSS) and AWS Services

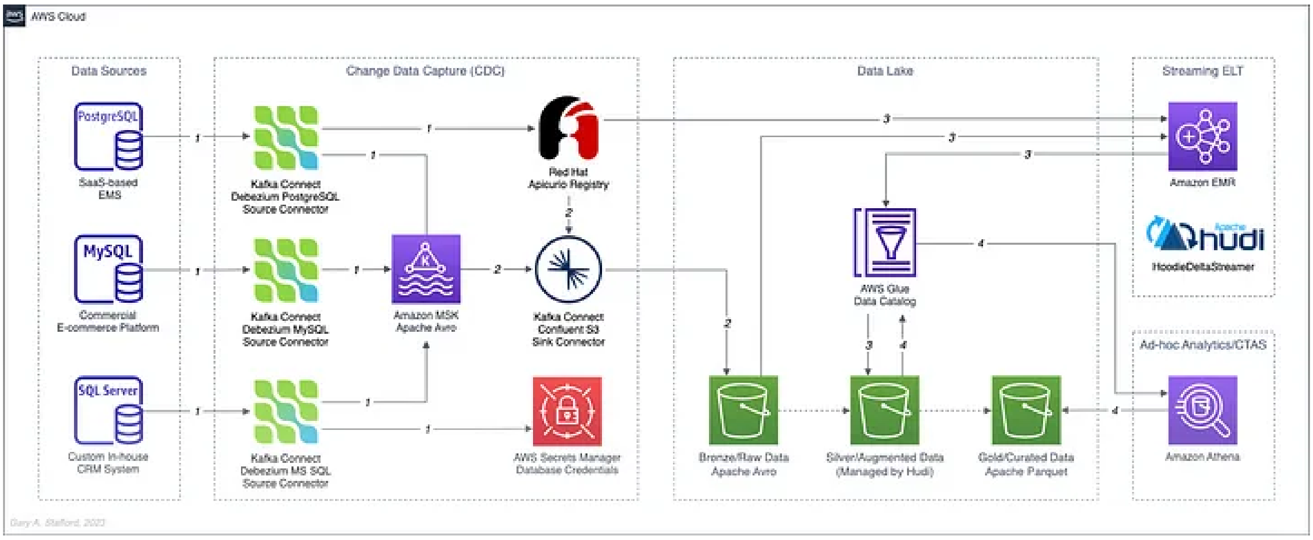
Gary A. Stafford · Follow

Published in ITNEXT

21 min read · Feb 28, 2023

Introduction

In the following post, we will explore one possible architecture for building a near real-time transactional data lake on AWS. The data lake will be built using a combination of open source software (OSS) and fully-managed AWS services. Red Hat's Debezium, Apache Kafka, and Kafka Connect will be used for change data capture (CDC). In addition, Apache Spark, Apache Hudi, and Hudi's DeltaStreamer will be used to manage the data lake. To complete our architecture, we will use several fully-managed AWS services, including Amazon RDS, Amazon MKS, Amazon EKS, AWS Glue, and Amazon EMR.


AWS Cloud
Data Sources
PostgreSQL
SaaS-based EMS
MySQL
Commercial E-commerce Platform
SQL Server
Custom In-house CRM System
Change Data Capture (CDC)
Kafka Connect Debezium PostgreSQL Source Connector
Kafka Connect Debezium MySQL Source Connector
Kafka Connect Debezium MS SQL Source Connector
Amazon MSK Apache Avro
Red Hat Apicurio Registry
Kafka Connect Confluent S3 Sink Connector
AWS Secrets Manager Database Credentials
Data Lake
AWS Glue Data Catalog
Bronze/Raw Data Apache Avro
Silver/Augmented Data (Managed by Hudi)
Gold/Curated Data Apache Parquet
Streaming ELT
Amazon EMR
Apache hudi
HoodieDeltaStreamer
Ad-hoc Analytics/CTAS
Amazon Athena
1
2
3
4
Gary A. Stafford, 2023

The data lake architecture used in this post's demonstration

Source Code

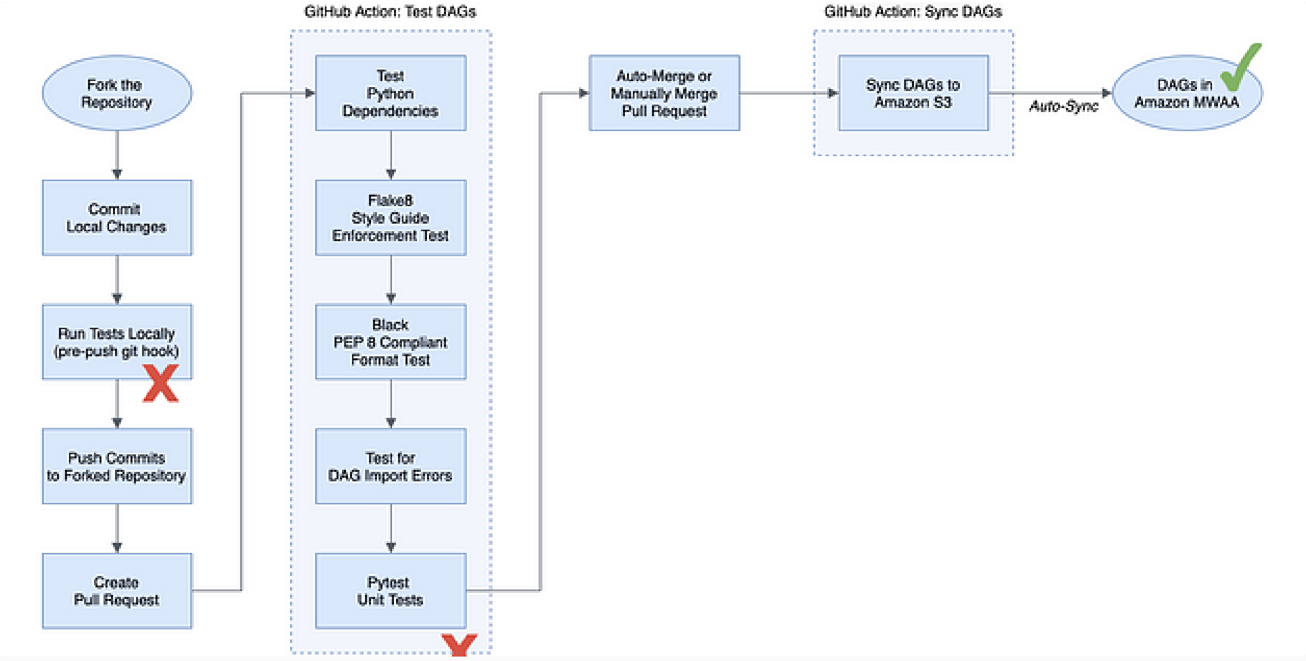
The source code, configuration files, EMR Notebook, and a list of commands shown in this post are open-sourced and available on GitHub.

Written by Gary A. Stafford

4.8K Followers · Writer for ITNEXT

Area Principal Solutions Architect @ AWS | 10x AWS Certified Pro | Polyglot Developer | DataOps | GenAI | Technology consultant, writer, and speaker

More from Gary A. Stafford and ITNEXT


GitHub Action: Test DAGs
GitHub Action: Sync DAGs
Fork the Repository
Commit Local Changes
Run Tests Locally (pre-push git hook)
Push Commits to Forked Repository
Create Pull Request
Test Python Dependencies
Flake8 Style Guide Enforcement Test
Black PEP 8 Compliant Format Test
Test for DAG Import Errors
Pytest Unit Tests
Auto-Merge or Manually Merge Pull Request
Sync DAGs to Amazon S3
Auto-Sync
DAGs in Amazon MWAA

Gary A. Stafford

DevOps for DataOps: Building a CI/CD Pipeline for Apache Airflow DAGs

Build an effective CI/CD pipeline to test and deploy your Apache Airflow DAGs to Amazon MWAA using GitHub Actions

Dec 13, 2021 165

Niels Cautaerts in ITNEXT

Containers: has the pendulum swung too far?

Containerization has revolutionized the software industry, but using them blindly for everything without considering their drawbacks or…

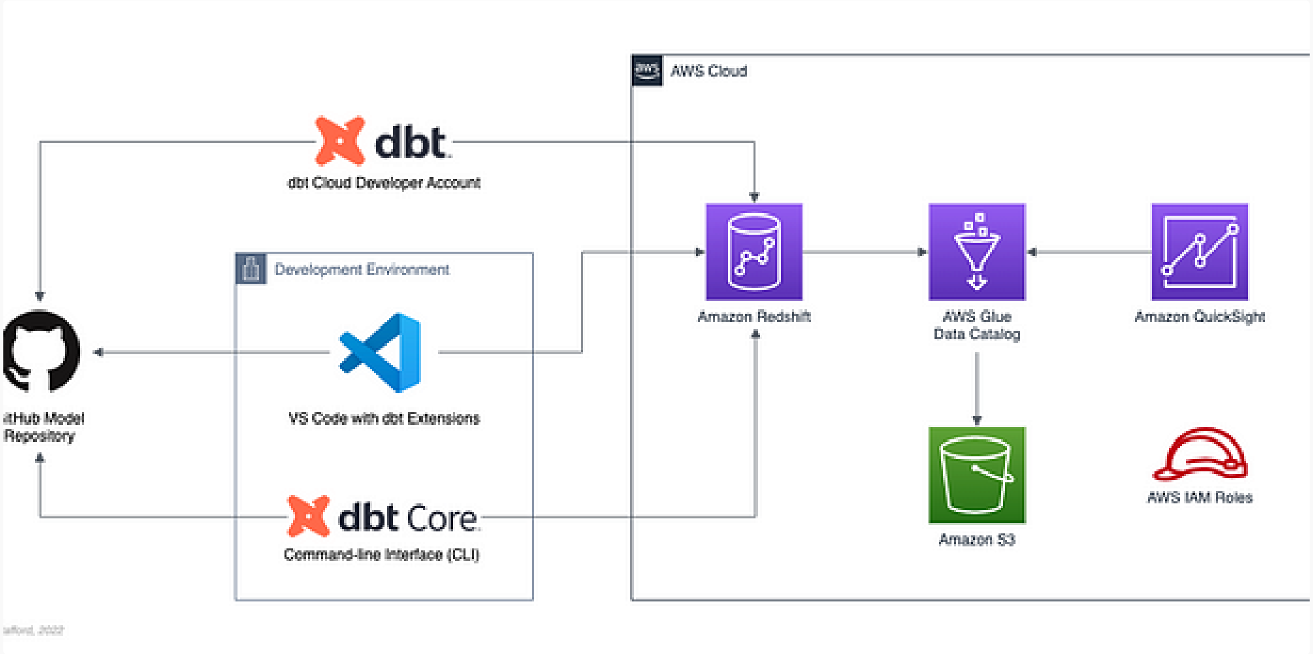
Aug 15 558 7

Mahdi Mallaki in ITNEXT

Dockerfile-less and Daemon-less Build

Building a Docker image without requiring a Dockerfile or Docker Daemon

Sep 1 245 3



Gary A. Stafford in ITNEXT

Lakehouse Data Modeling using dbt, Amazon Redshift, Redshift Spectrum, and AWS Glue

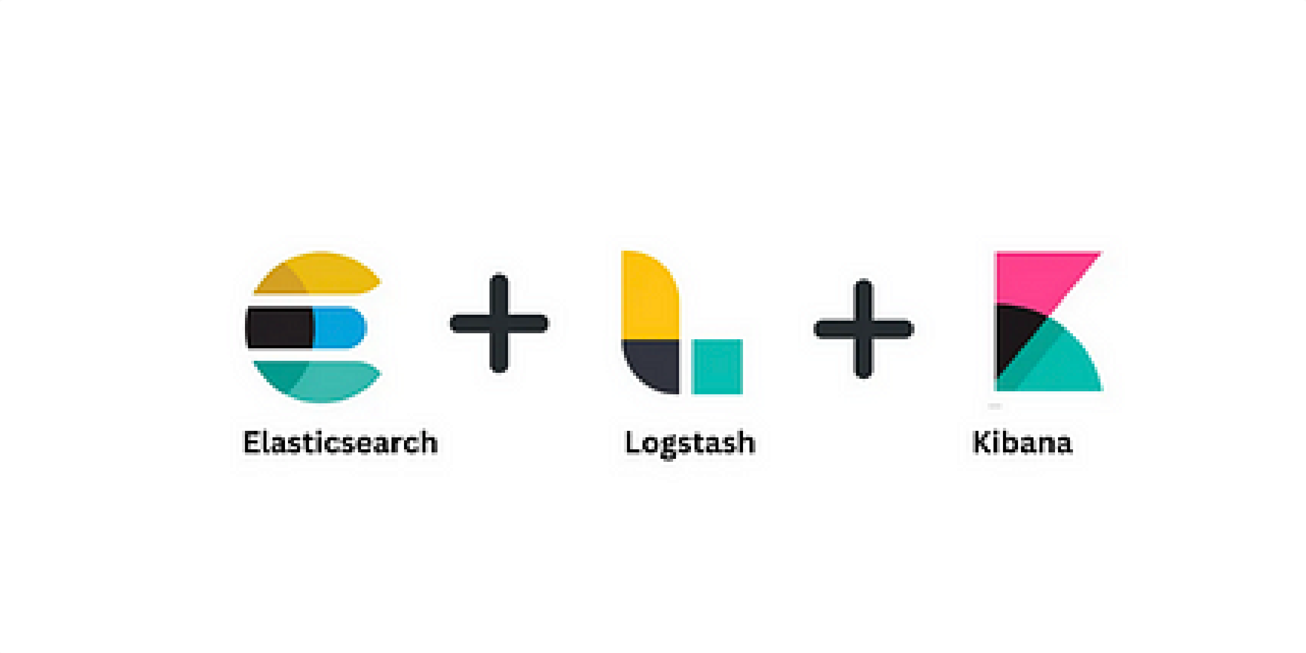
Learn how dbt makes it easy to transform data and materialize models in a modern cloud data lakehouse built on AWS

✦ Aug 19, 2022 👏 197 💬 2

See all from Gary A. Stafford

See all from ITNEXT

Recommended from Medium



Sai Parvathaneni in Towards Dev

Building a Log Analysis Data Pipeline Using Kafka, Elasticsearch, Logstash, and Kibana — ELK Stack

When it comes to analyzing logs, having a real-time, centralized, and automated solution is a game changer. Instead of sifting through...

Sep 11 69

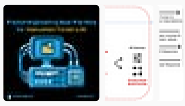
Amit Kumar Manjhi

Handling Schema Evolution in Debezium Kafka Connectors: A Step-by-Step Guide

Introduction:

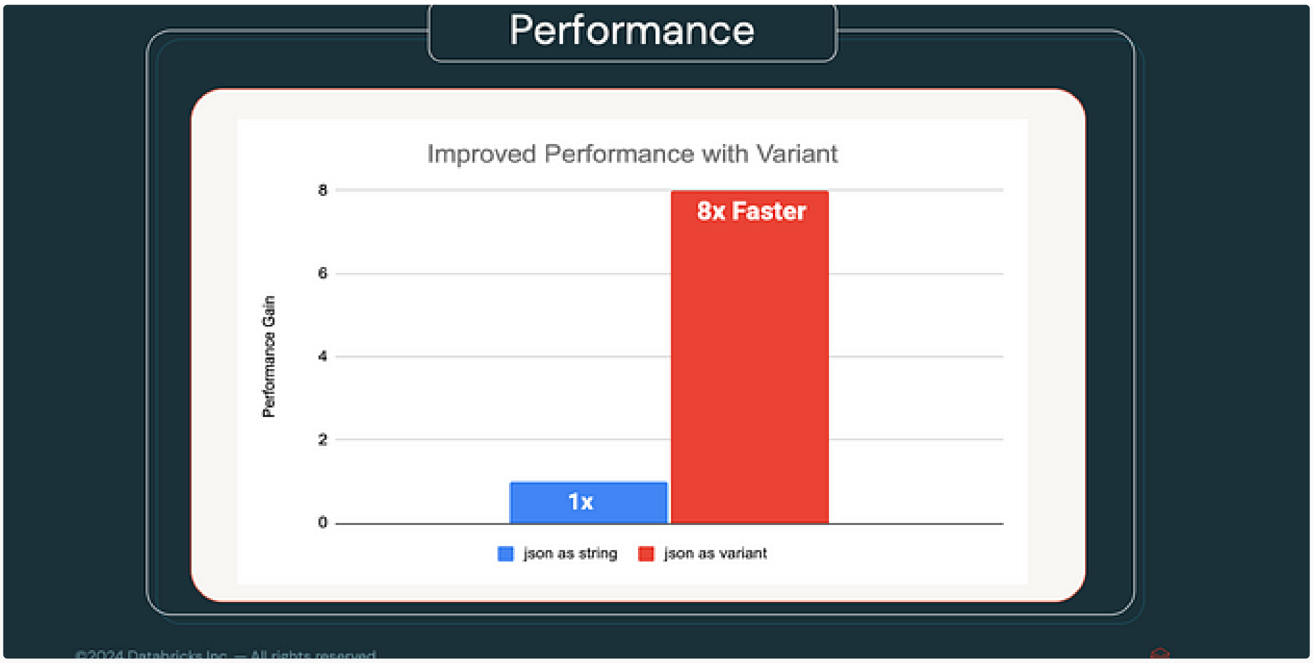
May 4

Lists



Natural Language Processing

1715 stories · 1287 saves



Archana Goyal

What’s Next for Apache Spark 4.0: A Comprehensive Overview with Comparisons to Spark 3.x

My articles are open to everyone; non-member readers can read the full article by clicking this link.

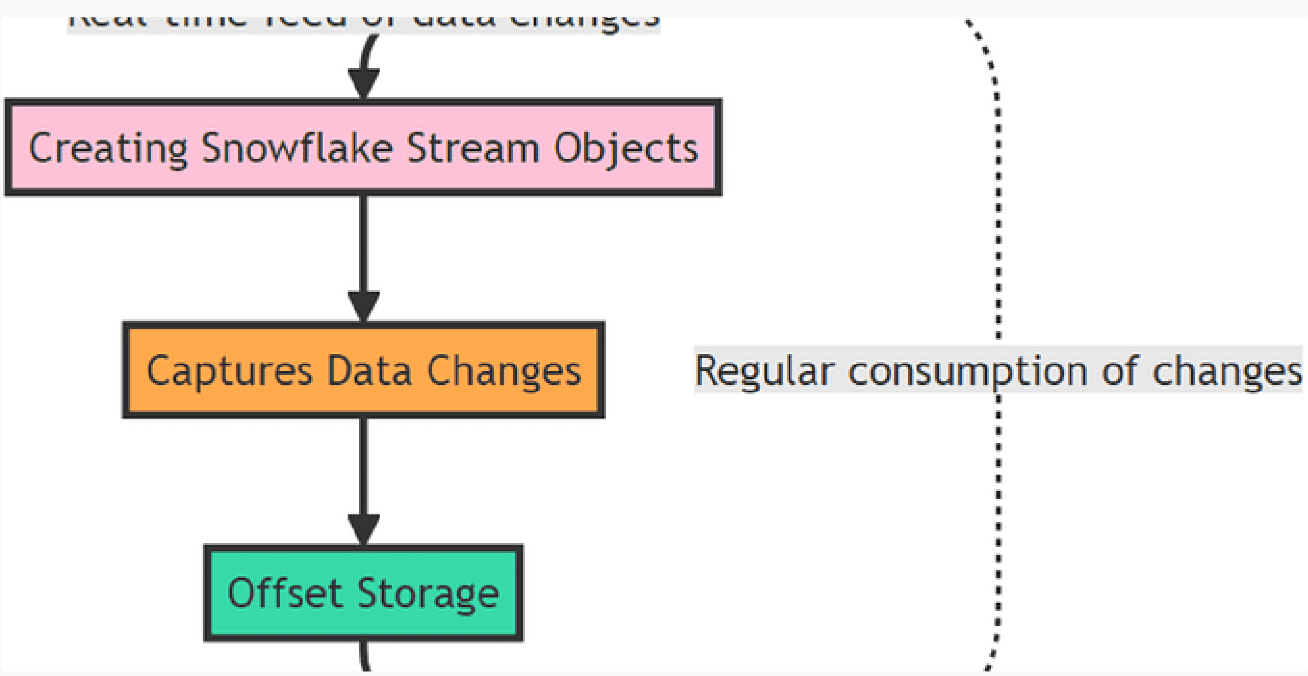
✦ Aug 25 👏 245 💬 2

Hugo Lu

Snowflake vs. Databricks 2024 (actually useful)

Snowflake vs. Databricks is something we've all heard before, so why not take a different approach

Aug 8 205 5


Real-time feed of data changes
Creating Snowflake Stream Objects
Captures Data Changes
Offset Storage
Regular consumption of changes

Hanson Olatunde

Change Data Capture Pipeline Automation in Snowflake

This article offers technical guidance on automating updates and tracking changes of data examples using Snowflake's CDC capabilities. The...

Aug 17 57

Gavin F.

Kafka Streams — How to magically join multiple data streams

Seamless Kafka Streams joining just like SQL table joins

Nov 5, 2023 281 3

See more recommendations