Build ETL Data Pipeline on AWS EMR Cluster

Welcome

Build ETL Data Pipeline on AWS EMR Cluster

Introduction to ETL

Project Overview

Problem Description

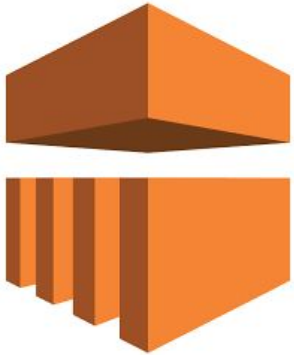
Keep it simple

1. The speed of getting data pipelines up and running
2. Getting new engineers up to speed
3. Spending less time worrying about the service management
4. Complexities
5. Providing business value
6. Overall engineering cost

Tour to existing solution

function	Open source	Managed services
Extract	Debezium or SQL script to pull to data	Stitch or fivetran
Transform	Open source SQL/ Apache Spark	fivetran or dbt cloud
Load	SQL script	Stitch or fivetran
Dashboard	Metabase / graphana	AWS Quicksight or looker or tableau
Monitor	Airflow	dbt cloud
Alert	Airflow with custom logic	dbt cloud
Schedule	Airflow	dbt cloud

Data Infrastructure : Components used

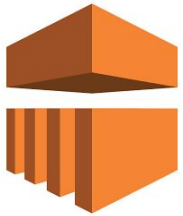


Aws services

Some of the benefits of AWS S3 are:

- Durability: S3 provides 99.999999999 percent durability.
- Low cost: S3 lets you store data in a range of “storage classes.” These classes are based on the frequency and immediacy you require in accessing files.
- Scalability: S3 charges you only for what resources you actually use, and there are no hidden fees or overage charges. You can scale your storage resources to easily meet your organization’s ever-changing demands.
- Availability: S3 offers 99.99 percent availability of objects
- Security: S3 offers an impressive range of access management tools and encryption features that provide top-notch security.
- Flexibility: S3 is ideal for a wide range of uses like data storage, data backup, software delivery, data archiving, disaster recovery, website hosting, mobile applications, IoT devices, and much more.
- Simple data transfer: You don’t have to be an IT genius to execute data transfers on S3. The service revolves around simplicity and ease of use.

Aws services



Benefits of using EMR

- Easy to use
- Low cost
- Elasticity
- Reliability
- Security
- Flexibility

Aws services

Features of Hive

- Hive is fast and scalable.
- It provides SQL-like queries (i.e., HQL) that are implicitly transformed to MapReduce or Spark jobs.
- It is capable of analyzing large datasets stored in HDFS.
- It allows different storage types such as plain text, RCFile, and HBase.
- It uses indexing to accelerate queries.
- It can operate on compressed data stored in the Hadoop ecosystem.
- It supports user-defined functions (UDFs) where user can provide its functionality.

Data Visualization Tools

Solution Description

1. Create an S3 bucket in AWS
2. Upload sales data into the S3
3. Spin an EMR cluster on AWS which has required services
   (1 master node 2 core nodes m5.xlarge)
4. Create Hive external table to point to the data in S3
5. Perform ETLs on Hive table and store it in final Hive table
6. Connect Hive final table in AWS EMR to tableau in local and plot the graphs

Components of a Data Engineering Platform

All data engineering platforms have 3 main sections, they are

1. `Extract` - extracting data from source systems
2. `Transform` - transforming the data according to business/data model requirements
3. `Load` - loading the data into a destination table

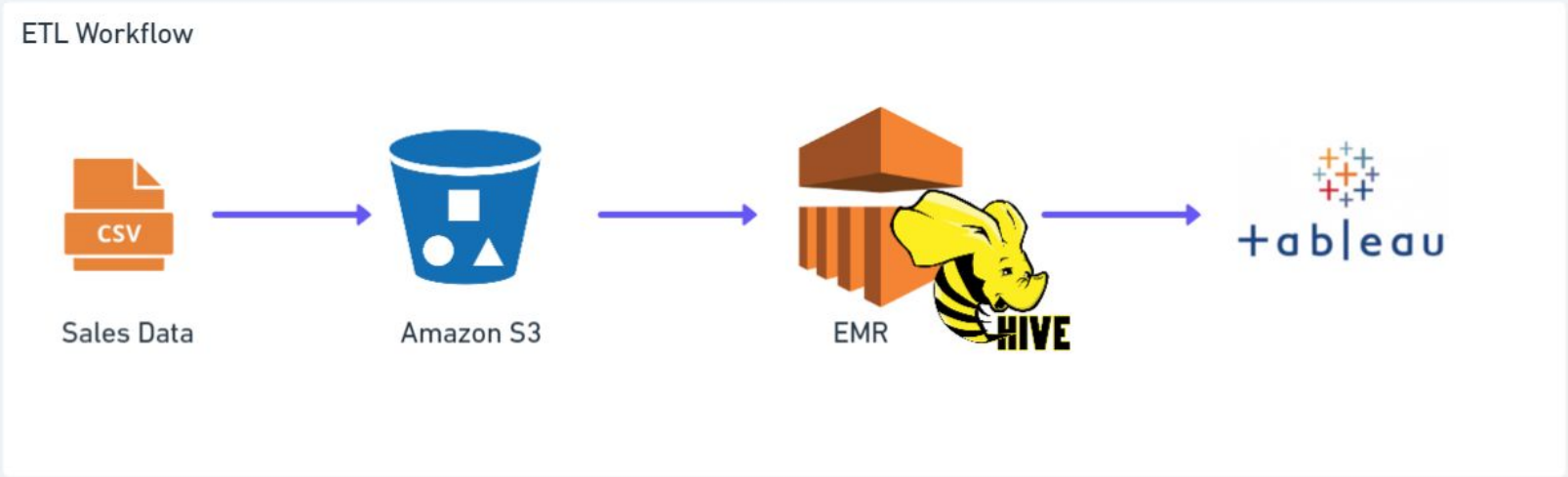
We add another layer, which is what the data users use, the presentation layer(aka Dashboards)

4. `Dashboard` - used by data users to gather insights from the cleaned data

When we operate in production environment, it is crucial to have a monitoring and alerting system to alert you in case something breaks, or data quality tests fail

5. `Monitoring` - used by engineers/analysts to check status of the data pipeline
6. `Alerting` - used to alert engineers/analysts in case of failures
7. `scheduling` - used by engineers to schedule the ETL runs

Tour to Architecture diagram



Cost Involved