Batch Processing using AWS EMR & Pyspark

In this project, you will learn to do batch processing using AWS services: S3 as storage, EMR as processing cluster, and Athena for querying the processed results.

Business Overview:

In a professional data engineering career, you have various scenarios where data gets collected every day. The data can be processed once a day, i.e., batch processed, and the processed results are stored in a location to derive insights and take appropriate action based on the insights. In this project, we will learn how to implement this using AWS services. We will take a day's worth of data related to Wikipedia, and we will perform batch processing on it.

Data Pipeline:

The sample data will be put into a folder in the S3 bucket. We will have PySpark code that will run on the EMR cluster. This code will fetch the data from the S3 bucket, perform filtering and aggregation on this data, and push the processed data back into S3 in another folder. We will then use Athena to query this processed data present in S3. We will create a table on top of the processed data by providing the relevant schema and then use ANSI SQL to query the data.

Agenda:

Firstly, we will look into the services that we will use in this project. We will understand the problem statement and look into the architecture we will use to solve the problem statement. We will then work on setting up the AWS services that we will use. We will write the PySpark code, which will perform batch processing. We will see how to run this code on the EMR cluster. We will see if we have the desired output in the desired S3 location. Then, in AWS Athena, we will create the tables corresponding to the processed data and use these tables to query the processed data present in S3.

Languages - Python
Package - Pyspark
Services - AWS EMR, AWS S3, AWS Athena.

Amazon S3

Amazon S3 is an object storage service that provides instant scalability, data availability, security, and performance. Users can save and retrieve any amount of data using Amazon S3, irrespective of time and location. AWS charges the user per GB of data stored. It also provides features to customize different roles for accessing the data for various organizational requirements.

Amazon EMR

Amazon EMR (Elastic MapReduce) is an AWS tool for big data processing and analysis. It processes big data across a Hadoop cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3). The user can set up a hassle cluster in a few minutes.

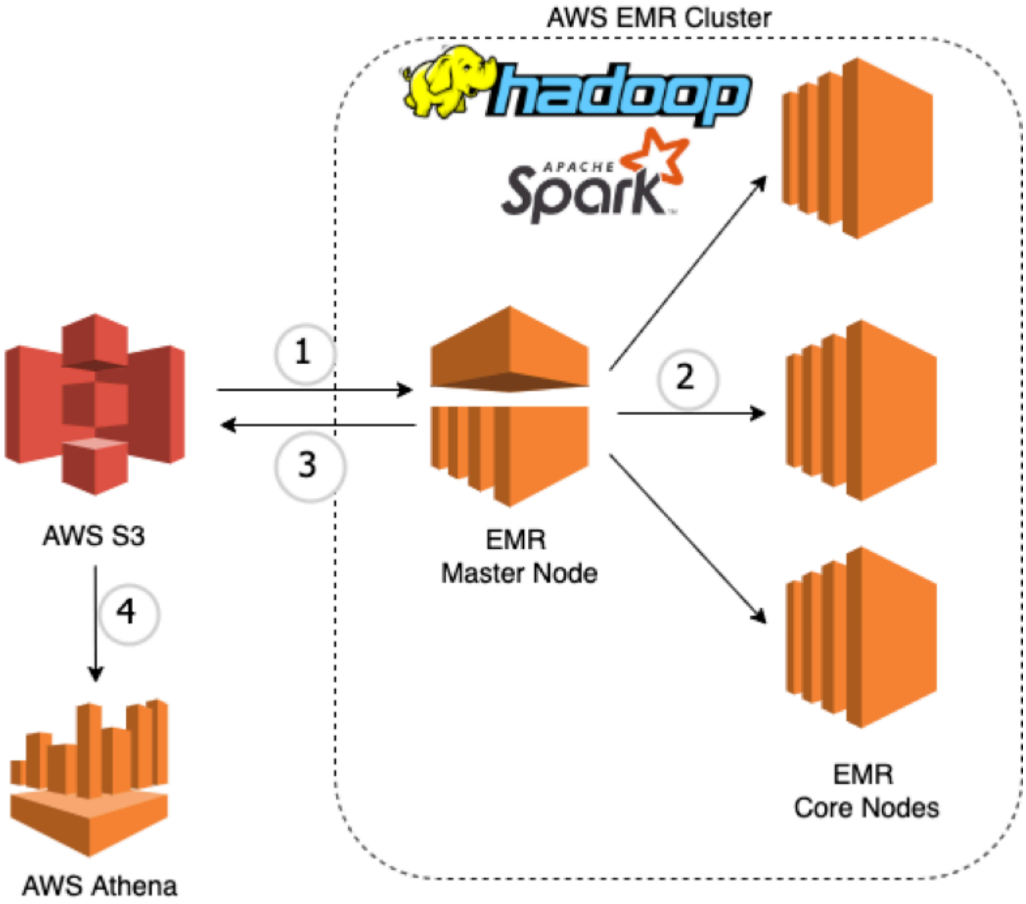
Amazon Athena

Amazon Athena is an interactive query service that uses standard SQL to query data and analyze big data in Amazon S3. Under the hood, it uses Presto, a distributed SQL engine, to run queries. We can query structured, semi-structured, and unstructured data types that might be stored in formats like CSV, AVRO, Parquet, etc. Athena is server-less, so there is no infrastructure to set up or manage, and you pay only for the queries you run.

Key Takeaway

- Understanding the project overview
- Understanding the problem statement
- Understanding the project architecture
- Creating EC2 key-pair
- Creating AWS EMR Cluster
- Creating AWS S3 buckets and folders
- Uploading JSON data files to AWS S3
- Developing and executing the transformation & actions on data
- Performing data analysis using AWS Athena

Architecture :


AWS EMR Cluster
hadoop
APACHE Spark
1
2
3
4
AWS S3
EMR
Master Node
EMR
Core Nodes
AWS Athena