Data Processing with Spark-SQL using Scala In AWS

Agenda

Apache Spark is an open-source distributed processing solution for substantial data workloads. It combines in-memory caching and rapid query execution for quick analytic queries on any amount of data. It includes development APIs in Java, Scala, Python, and R. It allows code reuse across various workloads, including batch processing, interactive queries, real-time analytics, machine learning, and graph processing.
Scala is a multi-paradigm, general-purpose, high-level programming language. It's an object-oriented programming language that also supports functional programming. Scala applications can be converted to bytecodes and run on the Java Virtual Machine (JVM). Scala is a scalable programming language, and Javascript runtimes are also available. This project presents the fundamentals of Scala in an easy-to-understand manner.

Aim

To understand the fundamentals of Scala in an easy-to-understand manner, also creating RDDs and performing transformation operations on them. This project also involves analyzing the Movies dataset using RDD and Spark SQL.

Data Description

In the project, we will use Movies and Rating datasets. The Movies dataset contains movie id, title, release date, etc. The Rating dataset contains customer id, movie id, ratings, and timestamp information.

Tech Stack

➔ Language: SQL, Scala

➔ Services: AWS EC2, Docker, Hive, HDFS, Spark

AWS EC2

Amazon EC2 instance is a virtual server on Amazon's Elastic Compute Cloud (EC2) for executing applications on the Amazon Web Services (AWS) architecture. Corporate customers can use the Amazon Elastic Compute Cloud (EC2) service to run applications in a computer environment. Amazon EC2 eliminates the requirement for upfront hardware investment, allowing customers to design and deploy projects quickly. Users can launch as many or as few virtual servers as they like, configure security and networking, and manage storage on Amazon EC2.

Docker

Docker is a free and open-source containerization platform, and it enables programmers to package programs into containers. These standardized executable components combine application source code with the libraries and dependencies required to run that code in any environment.

Scala

Scala is a multi-paradigm, general-purpose, high-level programming language. It's an object-oriented programming language that also supports functional programming. Scala applications can be converted to bytecodes and run on the Java Virtual Machine (JVM). Scala is a scalable programming language, and Javascript runtimes are also available.

Hive

Apache Hive is a fault-tolerant distributed data warehouse that allows for massive-scale analytics. Using SQL, Hive allows users to read, write, and manage petabytes of data. Hive is built on top of Apache Hadoop, an open-source platform for storing and processing large amounts of data. As a result, Hive is inextricably linked to Hadoop and is designed to process petabytes of data quickly. Hive is distinguished by its ability to query large datasets with a SQL-like interface utilizing Apache Tez or MapReduce.

Approach

- Create an AWS EC2 instance and launch it.
- Create docker images using docker-compose file on EC2 machine via ssh.
- Load data from local machine into Spark container via EC2 machine.
- Perform analysis on Movie and Ratings data.

Project Takeaways

- Understanding various services provided by AWS
- Creating an AWS EC2 instance and launching it
- Connecting to an AWS EC2 instance via SSH
- Dockerization.
- Copying a file from a local machine to an EC2 machine
- Understanding fundamentals of Scala.
- Creating RDDs
- Applying Transformation operations on RDDs
- Difference between RDDs and Dataframes
- Perform analysis using RDDs

- Perform analysis using Dataframe and Spark SQL

Note:

- Postgresql jar file
    1. Download jar file from https://jdbc.postgresql.org/download.html.
    2. Command to move jar file from local machine to ec2 machine =>
       ```
       scp -r -i "demo_hive.pem" postgresql-42.3.1.jar ec2-user@ec2-3-93-63-210.compute-1.amazonaws.com:/home/ec2-user/
       ```
    3. Command to move jar file from ec2 machine to Spark container =>
       ```
       docker cp /home/ec2-user/postgresql-42.3.1.jar hdp_spark-master:/spark/jars
       ```

- Hive-site.xml file
    1. Command to copy hive-site.xml file from Hive container and paste it in ec2 machine =>
       ```
       docker cp ra_hive-server:/opt/hive/conf/hive-site.xml /home/ec2-user
       ```
    2. Command to copy hive-site.xml file from ec2 machine and paste it in Spark container =>
       ```
       docker cp /home/ec2-user/hive-site.xml hdp_spark-master:/spark/conf
       ```