PySpark Integration with Hive and Cassandra

Agenda

This is the seventh project in the Pyspark series. The sixth project focuses on integrating PySpark with Amazon S3 and MySQL database to perform ETL(Extract-Transform-Load) and ELT(Extract-Load-Transform) operations. This project mainly focuses on integrating PySpark with Apache Cassandra and Apache Hive to perform ETL(Extract-Transform-Load) and ELT(Extract-Load-Transform) operations.

Tech stack:
➔Language: Python
➔Package: Pyspark
➔Services: AWS EC2, Docker, Apache Cassandra, Hive

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.

Apache Cassandra

Apache Cassandra is a distributed data storage system that is free and open-source. It is a column-oriented database. It is fault-tolerant and scalable. Its design enables users to respond to abrupt spikes in demand by allowing them to simply add extra hardware to accommodate more customers and data. Cassandra can handle organized, semi-structured, and unstructured data, allowing users to store data in a variety of ways. Cassandra employs numerous data centers to facilitate data delivery wherever and

wherever it is required. Cassandra supports the ACID properties of atomicity, consistency, isolation, and durability.

Hive

Apache Hive is a fault-tolerant distributed data warehouse that allows for massive-scale analytics. Using SQL, Hive will enable 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.

Key Takeaways:

- Understanding the project overview
- Create an AWS EC2 instance and launch it.
- Create docker images using docker-compose file on EC2 machine via ssh.
- Dockerization
- Introduction to PySpark
- Introduction to Apache Hive
- Introduction to Apache Cassandra
- Need for PySpark integration
- Understanding the concept of ETL
- Difference between ETL and ELT
- PySpark integration with Apache Hive
- PySpark integration with Apache Cassandra

Note:

Prerequisite for Apache Cassandra:

- Python 2
- Java 8

After installing Cassandra, we must set the CASSANDRA_HOME variable with its path in the Environment variable.