Apache Spark Installation and Configuration

Before installing Cassandra, we need to make sure that we have Java installed on our machine.

To do that, run the following command;

```
sudo apt install openjdk-8-jdk -y
or
sudo apt install openjdk-11-jdk -y
```

Verify if Java has been installed with the following command;

```
java --version
```

Now since we are using PySpark, we need to install python as well,

```
sudo apt install python3
```

Now download Apache Spark from the official website, using the following command,

```
wget https://downloads.apache.org/spark/spark-3.3.2/spark-3.3.2-bin-hadoop2.7.tgz
```

Once your download is complete, untar the archive file contents using tar command, tar is a file archiving tool. Once untar complete, rename the folder to spark.

```
tar -xzf spark-3.3.2-bin-hadoop2.7.tgz
mv spark-3.3.2-bin-hadoop2.7 spark
```

Now we need to tell Ubuntu about the spark location,

```
bhavyom@bhavyom:~$ vi ~/.bashrc
# Add below lines at the end of the .bashrc file.
export SPARK_HOME=/home/sparkuser/spark
export PATH=$PATH:$SPARK_HOME/bin
```

Now load the environment variables to the opened session by running below command

```
bhavyom@bhavyom:~$ source ~/.bashrc
```

To check if spark is installed, open a now terminal, and enter command and press enter,

```
spark-shell
```

If a scala terminal opens up, then spark has been installed correctly.

Now to run the spark for our project, enter the following command,

```
spark-submit --packages "com.datastax.spark:spark-cassandra-
connector_2.12:3.3.0","org.apache.spark:spark-sql-kafka-0-
10_2.12:3.3.2"
/home/bhavyom/Projects/AdvDBWebsite/AdvDB_backend/app/sparkPro
cessor.py
```

This will start the spark consumer.

Note: The version of connector is very crucial. Use the compatible version only.

Use the below website to find the appropriate connector,

https://mvnrepository.com/artifact/com.datastax.spark/spark-cassandra-connector