# Installation and Configuration of Apache Spark and Apache Cassandra on Linux Mint

November 20, 2024

## Contents

# 1 Prerequisites

- **Java Development Kit (JDK)**: Install JDK 8 or higher.
- **Python**: Install Python 3.x for Spark (optional if you want Python APIs).
- **Internet Access**: To download packages.

# 2 Step 1: Install Java

1. Open a terminal and run:

```
sudo apt update
sudo apt install openjdk-11-jdk -y
```

2. Verify installation:

```
java -version
```

   You should see a version output like `openjdk version "11.0.x"`.

# 3 Step 2: Install Apache Spark

1. **Download Spark**:
   - Go to Apache Spark Downloads.
   - Choose the latest stable version (e.g., Spark 3.7.0) and a pre-built package for Hadoop.

   Or, use the terminal:

```
wget https://dlcdn.apache.org/spark/spark-3.7.0/spark-3.7.0-bin-hadoop3.tgz
```

2. **Extract Spark**:

```
tar -xvzf spark-3.7.0-bin-hadoop3.tgz
sudo mv spark-3.7.0-bin-hadoop3 /opt/spark
```

3. **Configure Environment Variables**: Add Spark to your `/.bashrc` file:

```
echo "export SPARK_HOME=/opt/spark" >> ~/.bashrc
echo "export PATH=$SPARK_HOME/bin:$PATH" >> ~/.bashrc
source ~/.bashrc
```

4. **Verify Spark Installation**: Run:

```
spark-shell
```

   You should enter the Spark REPL with Scala.

# 4 Step 3: Install Apache Cassandra

1. **Add Cassandra Repository**:

```
echo "deb [arch=amd64] https://www.apache.org/dist/cassandra/debian 41x main" |
    sudo tee -a /etc/apt/sources.list.d/cassandra.sources.list
curl https://downloads.apache.org/cassandra/KEYS | sudo apt-key add -
sudo apt update
```

2. **Install Cassandra**:

```
sudo apt install cassandra -y
```

3. **Start Cassandra**:

```
sudo systemctl start cassandra
sudo systemctl enable cassandra
```

4. **Verify Installation**:

```
nodetool status
```

You should see `UN` (Up and Normal) indicating Cassandra is running.

Access the Cassandra shell:

```
cqlsh
```

If successful, you'll be connected to Cassandra.

# 5 Step 4: Integrate Spark and Cassandra

1. **Download the Cassandra Connector for Spark**:

   - Go to the DataStax Spark-Cassandra Connector.
   - Choose the correct version matching your Spark version.

Example for Spark 3.7:

```
wget
    https://downloads.datastax.com/spark-cassandra-connector/spark-cassandra-connector-assembly
    -P /opt/spark/jars
```

2. **Verify Integration**:

   - Open the `spark-shell` with Cassandra connector:

```
spark-shell --jars
    /opt/spark/jars/spark-cassandra-connector-assembly_2.12-3.3.0.jar
```

   - Test a Cassandra connection:

```
import com.datastax.spark.connector._
val conf = spark.conf.set("spark.cassandra.connection.host", "127.0.0.1")
println("Cassandra connection successful")
```

# 6 Step 5: Create a Keyspace and Tables in Cassandra

1. Access `cqlsh`:

```
cqlsh
```

2. Create a Keyspace:

```
CREATE KEYSPACE social_media_analytics
WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};
```

3. Create Tables:

```
CREATE TABLE social_media_analytics.post_stats (
    post_id uuid PRIMARY KEY,
    comments int,
    likes int,
    shares int
);
```

# 7 Step 6: Troubleshooting

- Check Cassandra service:

```
sudo systemctl status cassandra
```

- Check Spark logs:

```
logs directory in the Spark folder
```