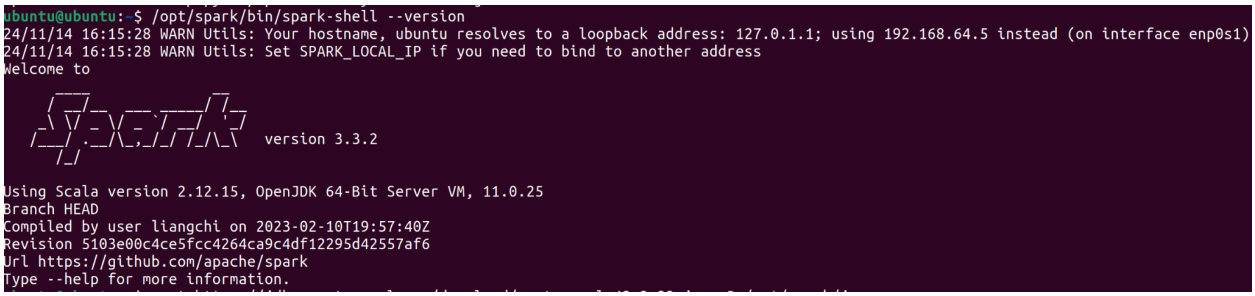
BDA
Assignment 7
Aakanksha Darekar
202200733 09 A1

Experiment 7: Access Postgres database tables with Spark SQL

Code:

1. **Install Apache Spark on your Ubuntu system if it's not already installed:**

   sudo apt update
   sudo apt install -y default-jdk scala git
   wget https://archive.apache.org/dist/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz -P /tmp
   tar xvf /tmp/spark-3.3.2-bin-hadoop3.tgz -C /tmp
   sudo mv /tmp/spark-3.3.2-bin-hadoop3 /opt/spark
   /opt/spark/bin/spark-shell --version



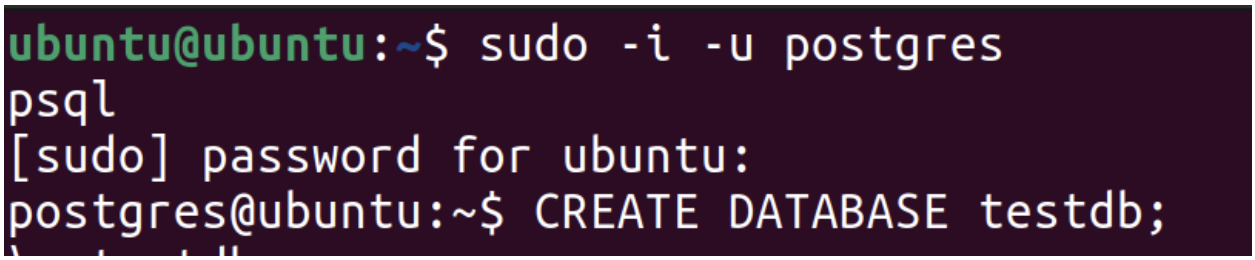
2. **Download the PostgreSQL JDBC Driver:**

   wget https://jdbc.postgresql.org/download/postgresql-42.2.23.jar -P /opt/spark/jars
   /opt/spark/bin/spark-shell --jars /opt/spark/jars/postgresql-42.2.23.jar

3. **Start PostgreSQL and Create a Database:**

```
sudo apt install -y postgresql postgresql-contrib
sudo systemctl start postgresql
sudo -i -u postgres
psql

createdb testdb;
\c testdb;
CREATE TABLE example_table (
    id SERIAL PRIMARY KEY,
    name VARCHAR(50),
    age INT
);
INSERT INTO example_table (name, age) VALUES ('Alice', 30), ('Bob', 25);
```



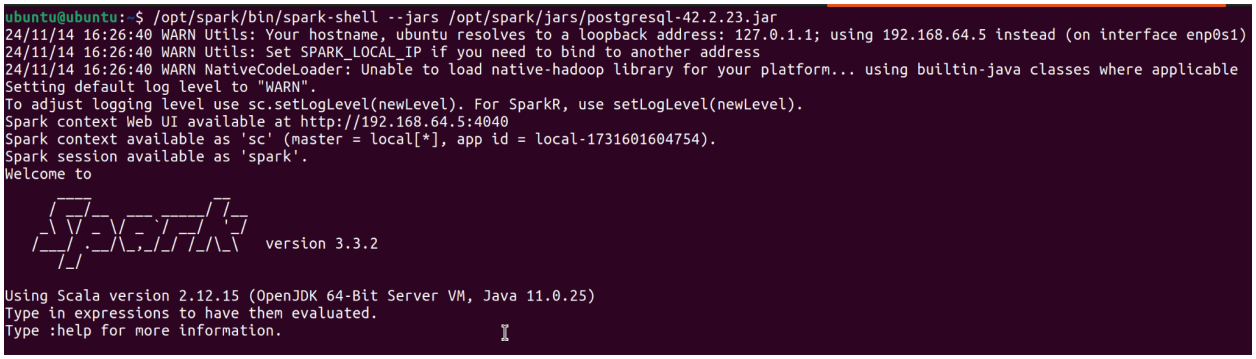

```
postgres@ubuntu:~$ createdb testdb
postgres@ubuntu:~$ psql testdb
psql (14.13 (Ubuntu 14.13-0ubuntu0.22.04.1))
Type "help" for help.

testdb=# \c testdb;
You are now connected to database "testdb" as user "postgres".
testdb=# CREATE TABLE example_table (
    id SERIAL PRIMARY KEY,
    name VARCHAR(50),
    age INT
);

INSERT INTO example_table (name, age) VALUES ('Alice', 30), ('Bob', 25);
CREATE TABLE
INSERT 0 2
testdb=# \q
exit
\q: extra argument "exit" ignored
postgres@ubuntu:~$
```

4. **Connect Spark SQL to PostgreSQL:**

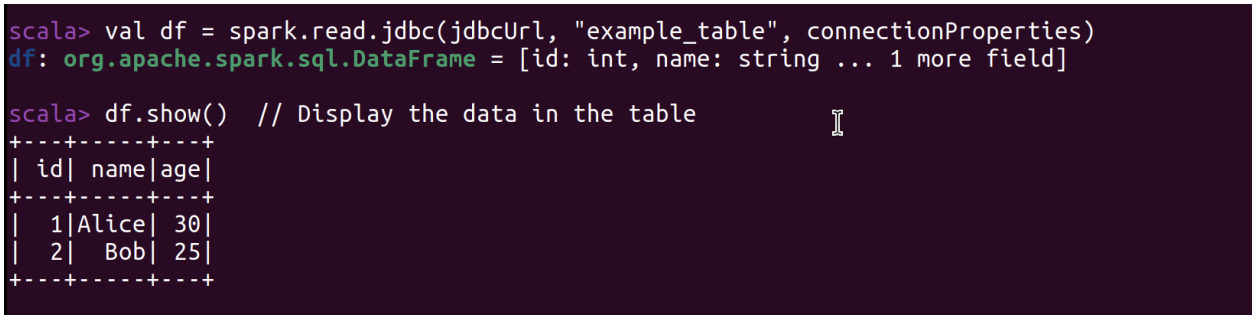
/opt/spark/bin/spark-shell --jars /opt/spark/jars/postgresql-42.2.23.jar



5. **Configure the connection to PostgreSQL:**

val jdbcUrl = "jdbc:postgresql://localhost:5432/testdb"
val connectionProperties = new java.util.Properties()
connectionProperties.put("user", "postgres") // replace with your PostgreSQL username
connectionProperties.put("password", "ubuntu") // replace with your PostgreSQL password

val df = spark.read.jdbc(jdbcUrl, "example_table", connectionProperties)
df.show() // Display the data in the table



6. **View query on PostgreSQL table with Spark SQL:**

df.createOrReplaceTempView("example_table_view")

val resultDf = spark.sql("SELECT * FROM example_table_view WHERE age > 25")
resultDf.show()

```
scala> df.createOrReplaceTempView("example_table_view")

scala>

scala> val resultDf = spark.sql("SELECT * FROM example_table_view WHERE age > 25")
resultDf: org.apache.spark.sql.DataFrame = [id: int, name: string ... 1 more field]

scala> resultDf.show()
+---+-----+---+
| id| name|age|
+---+-----+---+
|  1|Alice| 30|
+---+-----+---+
```