BDA Assignment 7

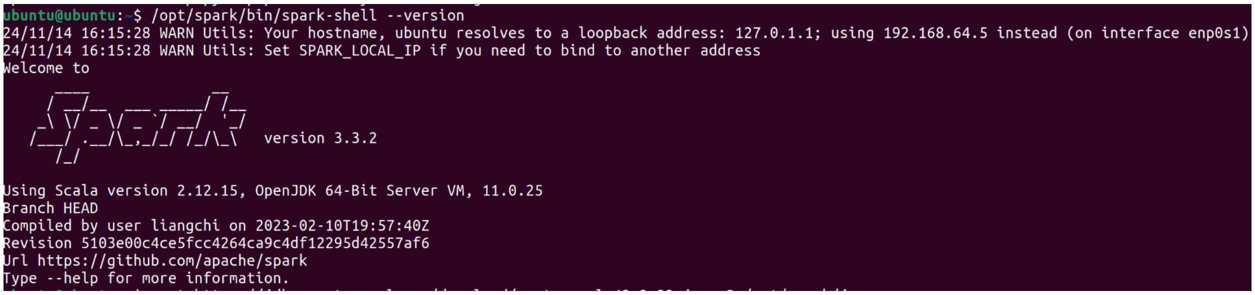
Name: Devansh Thakkar
Srn no: 202201187
Roll no: 25
TY-A

Experiment 7: Access Postgres database tables with Spark SQL

Code:

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

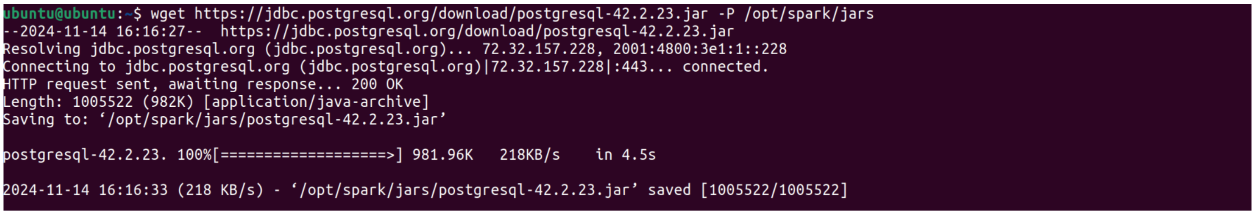
```
sudo apt update
sudo apt install -y default-jdk scala git
wget https://archive.apache.org/dist/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz -P
/tmp
tar xvf /tmp/spark-3.3.2-bin-hadoop3.tgz -C /tmp
sudo mv /tmp/spark-3.3.2-bin-hadoop3 /opt/spark3.tgz
/opt/spark/bin/spark-shell --version
```



```
ubuntu@ubuntu:~$ /opt/spark/bin/spark-shell --version
24/11/14 16:15:28 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 192.168.64.5 instead (on interface enp0s1)
24/11/14 16:15:28 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.2
      /_/

Using Scala version 2.12.15, OpenJDK 64-Bit Server VM, 11.0.25
Branch HEAD
Compiled by user liangchi on 2023-02-10T19:57:40Z
Revision 5103e00c4ce5fcc4264ca9c4df12295d42557af6
Url https://github.com/apache/spark
Type --help for more information.
```

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

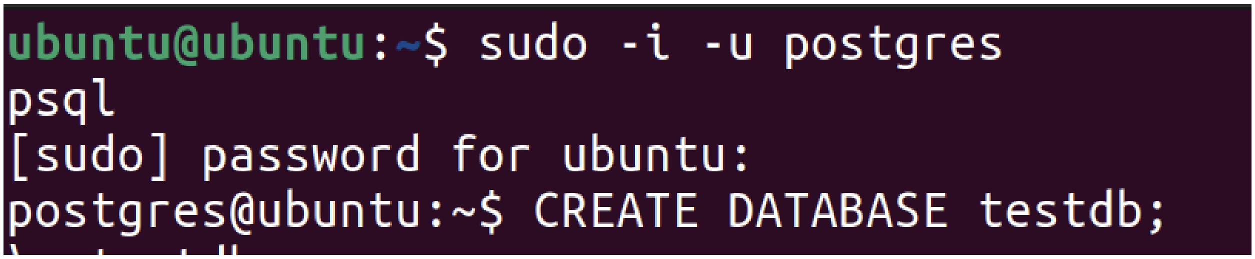
```
wget https://jdbc.postgresql.org/download/postgresql-42.2.23.jar -P /opt/spark/jars
/opt/spark/bin/spark-shell --jars /opt/spark/jars/postgresql-42.2.23.jar
```



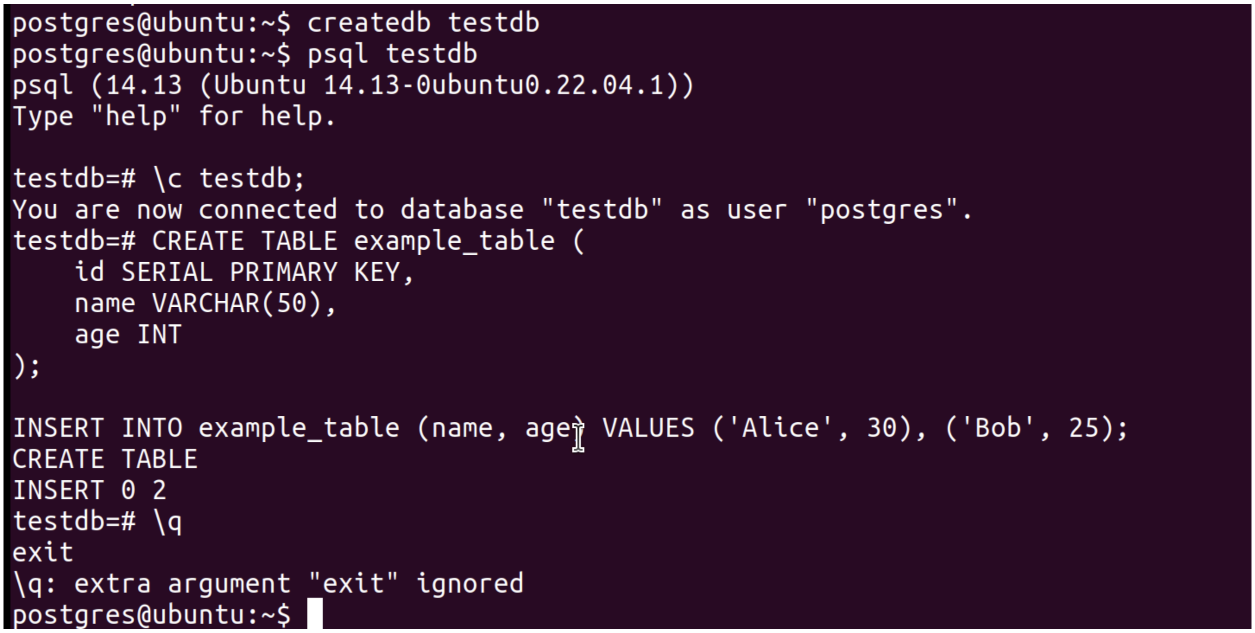
```
ubuntu@ubuntu:~$ wget https://jdbc.postgresql.org/download/postgresql-42.2.23.jar -P /opt/spark/jars
--2024-11-14 16:16:27--  https://jdbc.postgresql.org/download/postgresql-42.2.23.jar
Resolving jdbc.postgresql.org (jdbc.postgresql.org)... 72.32.157.228, 2001:4800:3e1:1::228
Connecting to jdbc.postgresql.org (jdbc.postgresql.org)|72.32.157.228|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1005522 (982K) [application/java-archive]
Saving to: ‘/opt/spark/jars/postgresql-42.2.23.jar’

postgresql-42.2.23. 100%[===================>] 981.96K   218KB/s    in 4.5s

2024-11-14 16:16:33 (218 KB/s) - ‘/opt/spark/jars/postgresql-42.2.23.jar’ saved [1005522/1005522]
```

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);
```



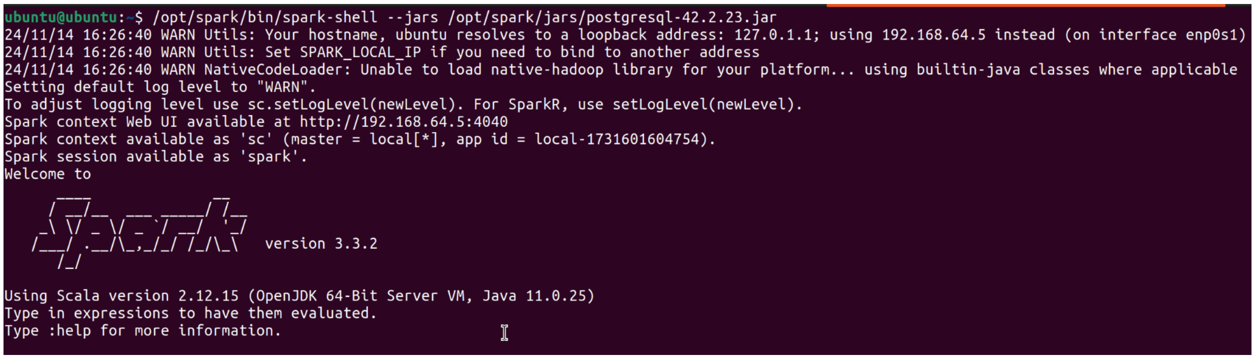
```
ubuntu@ubuntu:~$ sudo -i -u postgres
psql
[sudo] password for ubuntu:
postgres@ubuntu:~$ CREATE DATABASE testdb;
```



```
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

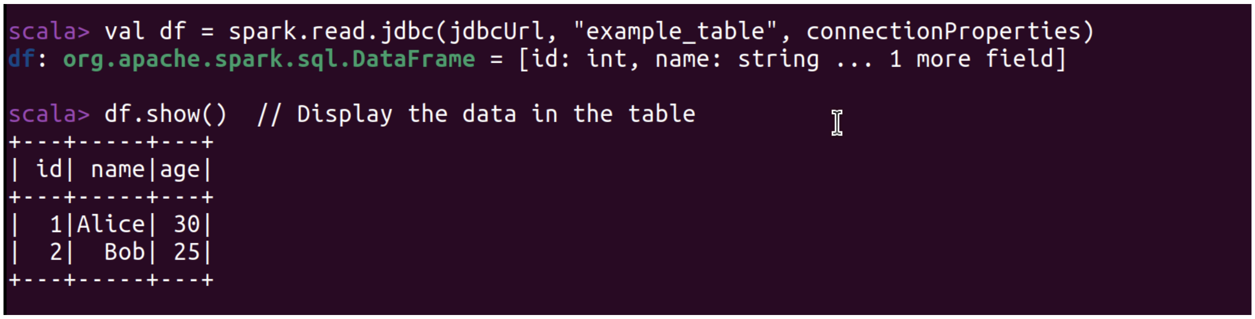


```
ubuntu@ubuntu:~$ /opt/spark/bin/spark-shell --jars /opt/spark/jars/postgresql-42.2.23.jar
24/11/14 16:26:40 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 192.168.64.5 instead (on interface enp0s1)
24/11/14 16:26:40 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/11/14 16:26:40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://192.168.64.5:4040
Spark context available as 'sc' (master = local[*], app id = local-1731601604754).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.2
      /_/

Using Scala version 2.12.15 (OpenJDK 64-Bit Server VM, Java 11.0.25)
Type in expressions to have them evaluated.
Type :help for more information.
```

5. Configure the connection to PostgreSQL:

val jdbcUrl = "jdbc:postgresql://localhost:5432/testdb"
val connectionProperties = new java.util.Properties()
connectionProperties.put("user", "postgres")
connectionProperties.put("password", "12345678")

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

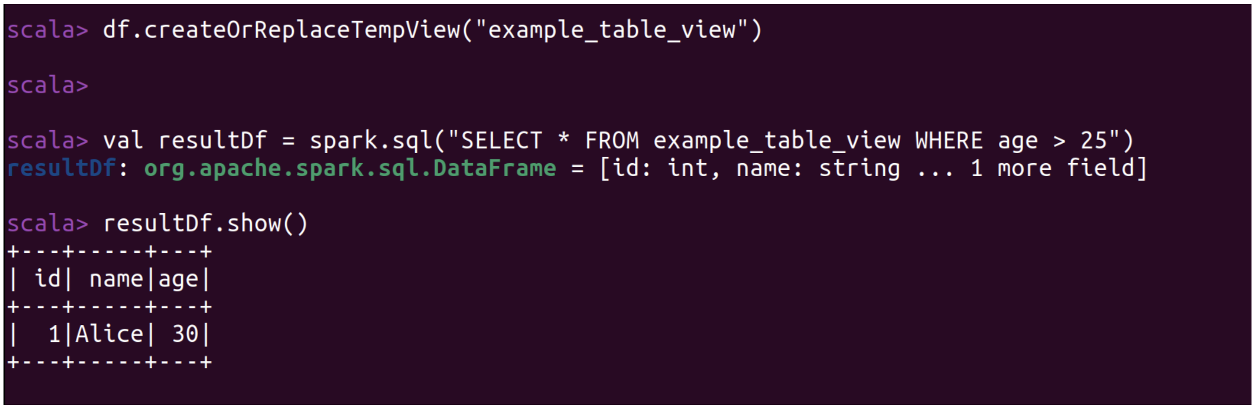


```
scala> val df = spark.read.jdbc(jdbcUrl, "example_table", connectionProperties)
df: org.apache.spark.sql.DataFrame = [id: int, name: string ... 1 more field]

scala> df.show()  // Display the data in the table
+---+-----+---+
| id| name|age|
+---+-----+---+
|  1|Alice| 30|
|  2|  Bob| 25|
+---+-----+---+
```

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|
+---+-----+---+
```