Lab Assignment 11

Title: Big Data Analytics (Impala)

PROBLEM STATEMENT:

Create databases and tables, insert small amounts of data, and run simple queries using Impala.

THEORY:

What is Impala?

Impala is a Massively Parallel Processing (MPP) SQL engine designed for Apache Hadoop. It provides high-performance, low-latency SQL queries on data stored in HDFS (Hadoop Distributed File System) and HBase. Developed by Cloudera, Impala uses the same metadata, SQL syntax (HiveQL), and ODBC driver as Apache Hive, allowing for better integration and accessibility.

Impala enables users to:
- Perform interactive SQL queries on big data.
- Avoid complex MapReduce jobs.
- Use standard SQL commands.
- Directly query the data without needing to move or transform it.

Installing Impala:

Impala is typically installed as part of Cloudera Distribution of Hadoop (CDH). For educational purposes, one can use the Cloudera QuickStart VM which comes pre-installed with Impala.

Basic Terms:

- Database: Logical collection of tables.
- Table: Structured format to store related data.
- Row: A record in the table.
- Column: A field in the record.

Procedure & Code:

STEP 1: OPEN IMPALA SHELL

impala-shell

STEP 2: CREATE DATABASE

```
CREATE DATABASE student_db;
USE student_db;
```

STEP 3: CREATE TABLE

```
CREATE TABLE students (
   id INT,
   name STRING,
   age INT
);
```

STEP 4: INSERT DATA

```
INSERT INTO students VALUES (1, 'Laukik', 21);
INSERT INTO students VALUES (2, 'Nikhil', 22);
INSERT INTO students VALUES (3, 'Sakshi', 20);
```

STEP 5: RUN SIMPLE QUERIES

```
-- View all data
SELECT * FROM students;

-- Filter by condition
SELECT * FROM students WHERE age > 20;

-- Count total records
SELECT COUNT(*) FROM students;
```

COMPILE AN IMPALA PROGRAM:

There is no separate compilation step in Impala. Queries are executed directly in the Impala shell in real-time.

RUNNING IMPALA PROGRAM:

Launch the shell using:
impala-shell

Type SQL commands directly to execute them.

CONCLUSION:

Students have successfully created databases and tables using Impala, inserted small amounts of data, and run basic SQL queries. This experiment demonstrates fundamental database operations in a big data environment using Apache Impala.