Experiment No. 11

Aim:

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

Objective:

To create database and perform different operation on database using Cloudera Impala.

Theory:

What is Impala?
Impala is a MPP (Massive Parallel Processing) SQL query engine for processing huge volumes of data that is stored in Hadoop cluster. It is an open source software which is written in C++ and Java. It provides high performance and low latency compared to other SQL engines for Hadoop.

In other words, Impala is the highest performing SQL engine (giving RDBMS-like experience) which provides the fastest way to access data that is stored in Hadoop Distributed File System.
Why Impala?
Impala combines the SQL support and multi-user performance of a traditional analytic database with the scalability and flexibility of Apache Hadoop, by utilizing standard components such as HDFS, HBase, Metastore, YARN, and Sentry.

With Impala, users can communicate with HDFS or HBase using SQL queries in a faster way compared to other SQL engines like Hive.

Impala can read almost all the file formats such as Parquet, Avro, RCFile used by Hadoop.

Impala uses the same metadata, SQL syntax (Hive SQL), ODBC driver, and user interface (Hue Beeswax) as Apache Hive, providing a familiar and unified platform for batch-oriented or real-time queries.

Unlike Apache Hive, Impala is not based on MapReduce algorithms. It implements a distributed architecture based on daemon processes that are responsible for all the aspects of query execution that run on the same machines.

Thus, it reduces the latency of utilizing MapReduce and this makes Impala faster than Apache Hive.

Advantages of Impala

- Using impala, you can process data that is stored in HDFS at lightning-fast speed with traditional SQL knowledge.
- Since the data processing is carried where the data resides (on Hadoop cluster), data transformation and data movement is not required for data stored on Hadoop, while working with Impala.
- Using Impala, you can access the data that is stored in HDFS, HBase, and Amazon s3 without the knowledge of Java (MapReduce jobs). You can access them with a basic idea of SQL queries.

Features of Impala

- Impala is available freely as open source under the Apache license.
- Impala supports in-memory data processing, i.e., it accesses/analyzes data that is stored on Hadoop data nodes without data movement.

- You can access data using Impala using SQL-like queries.
- Impala provides faster access for the data in HDFS when compared to other SQL engines.
- Using Impala, you can store data in storage systems like HDFS, Apache HBase, and Amazon s3.
- You can integrate Impala with business intelligence tools like Tableau, Pentaho, Micro strategy, and Zoom data.

Impala Environment:

1. **Downloading Cloudera Quick Start VM**

Step 1:Open the homepage of cloudera website http://www.cloudera.com/. You will get the page as shown below.

Step 2: Click the Sign in link on the cloudera homepage, which will redirect you to the Sign in page as shown below.

Step 3:After signing in, open the download page of cloudera website by clicking on the Downloads link highlighted in the following snapshot.

Step 4 - Download QuickStartVM

Download the cloudera QuickStartVM by clicking on the Download Now button, as highlighted in the following snapshot

2. Importing the Cloudera QuickStartVM

After downloading the cloudera-quickstart-vm-5.5.0-0-virtualbox.ovf file, we need to import it using virtual box. For that, first of all, you need to install virtual box in your system. Follow the steps given below to import the downloaded image file.

Step 1:Download virtual box from the following link and install it https://www.virtualbox.org/

Step 2: Open the virtual box software. Click File and choose Import Appliance, as shown below.

Step 3: On clicking Import Appliance, you will get the Import Virtual Appliance window. Select the location of the downloaded image file as shown below.

After importing Cloudera QuickStartVM image, start the virtual machine. This virtual machine has Hadoop, cloudera Impala, and all the required software installed.

Starting Impala Shell

To start Impala, open the terminal and execute the following command.

```
[cloudera@quickstart ~] $ impala-shell
```

This will start the Impala Shell, displaying the following message.

```
Starting Impala Shell without Kerberos authentication
Connected to quickstart.cloudera:21000
Server version: impalad version 2.3.0-cdh5.5.0 RELEASE (build
0c891d79aa38f297d244855a32f1e17280e2129b)
***********************************************************
 Welcome to the Impala shell. Copyright (c) 2015 Cloudera, Inc. All rights
(Impala Shell v2.3.0-cdh5.5.0 (0c891d7) built on Mon Nov 9 12:18:12 PS

Press TAB twice to see a list of available commands.
***********************************************************
[quickstart.cloudera:21000] >
```

Impala Query editor
In addition to Impala shell, you can communicate with Impala using the Hue browser. After installing CDH5 and starting Impala, if you open your browser, you will get the cloudera homepage.
click the bookmark Hue to open the Hue browser. On clicking, you can see the login page of the Hue Browser, logging with the credentials cloudera and cloudera.
As soon as you log on to the Hue browser, you can see the Quick Start Wizard of Hue browser.
On clicking the Query Editors drop-down menu, you will get the list of editors Impala supports
On clicking Impala in the drop-down menu, you will get the Impala query editor

- **Algorithm:**

  In Impala, a database is a construct which holds related tables, views, and functions within their namespaces. It is represented as a directory tree in HDFS; it contains tables partitions, and data files.

Step 1: **CREATE DATABASE**
Step 2: **SHOW DATABASES**.
Step 3: **CREATE TABLE** with schema
Step 4: Perform insertion operation in the database
Step 5: Show the result

- **Input:**
- **Output:**
- **Conclusion:**
-
- **Outcome:**

  Upon completion of this experiment, students will be able to: