## Experiment No: 01

### Aim: Run simple queries using Impala

### Problem Statement:

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

### Theory:

Apache Impala is an open source massively parallel processing SQL query engine for data stored in a computer cluster running Apache Hadoop. Impala is a MPP (Massive Parallel Processing) SQL query engine for processing huge volumes of data that is stored in a 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 a Hadoop Distributed File System.

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.

**Conclusion:** Hence we have studied, created & performed simple database queries on Imphala.