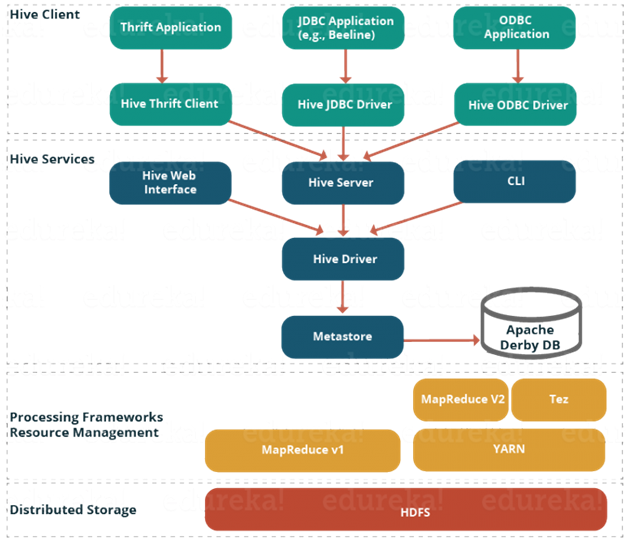
**Assignment 6.3**

**Problem Statement**

● Explain Hive Architecture in Brief.

Hive is a data warehouse solution built on top of Hadoop -by Facebook. It is an essential tool in the Hadoop ecosystem that provides an SQL (Structured Query Language) dialect (called as Hive Query Language) for querying data stored in the Hadoop Distributed File System (HDFS).

As shown in the below image, the Hive Architecture can be categorized into the following components: 

* **Hive Clients:**Hive supports application written in many languages like Java, C++, Python etc. using JDBC, Thrift and ODBC drivers.  Hence one can always write hive client application written in a language of their choice.
* **Hive Services:** Apache Hive provides various services like CLI, Web Interface etc. to perform queries.
* **Processing framework and Resource Management:**Internally, Hive uses Hadoop MapReduce framework as de facto engine to execute the queries.
* **Distributed Storage:** As Hive is installed on top of Hadoop, it uses the underlying HDFS for the distributed storage

● Explain Hive Components in Brief.

**Hive Clients:**

Apache Hive supports different types of client applications for performing queries on the Hive. These clients can be categorized into three types:

1. *Thrift Clients:* As Hive server is based on Apache Thrift, it can serve the request from all those programming language that supports Thrift.
2. *JDBC Clients:* Hive allows Java applications to connect to it using the type 4 (pure Java) JDBC driver which is defined in the class **org.apache.hadoop.hive.jdbc.HiveDriver**.
3. *ODBC Clients:* The Hive ODBC Driver allows applications that support the ODBC protocol to connect to Hive. (Like the JDBC driver, the ODBC driver uses Thrift to communicate with the Hive server.)

**Hive Services:**

Hive provides many services as shown in the image above. Let us have a look at each of them:

* **Hive CLI (Command Line Interface):**This is the default shell provided by the Hive where you can execute your Hive queries and commands directly.
* **Apache Hive Web Interfaces:**Apart from the command line interface, Hive also provides a web based GUI for executing Hive queries and commands.
* **Hive Server:**Hive server is built on Apache Thrift and therefore, is also referred as Thrift Server that allows different clients to submit requests to Hive and retrieve the final result.
* **Apache Hive Driver:**It is responsible for receiving the queries submitted through the CLI, the web UI, Thrift, ODBC or JDBC interfaces by a client. Then, the driver passes the query to the compiler where parsing, type checking and semantic analysis takes place with the help of schema present in the metastore. In the next step, an optimized logical plan is generated in the form of a DAG (Directed Acyclic Graph) of map-reduce tasks and HDFS tasks. Finally, the execution engine executes these tasks in the order of their dependencies, using Hadoop.
* **Metastore:**You can think metastore as a central repository for storing all the Hive metadata information. Hive metadata includes various types of information like structure of tables and the partitions along with the column, column type, serializer and deserializer which is required for Read/Write operation on the data present in HDFS. The metastore comprises of two fundamental units:
  + A service that provides metastore access to other Hive services.
  + Disk storage for the metadata which is separate from HDFS storage.