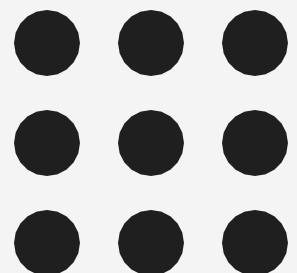# SNS COLLEGE OF ENGINEERING

Department of Artificial Intelligence and Data Science

Course Name – 19AD501 Big Data Analytics
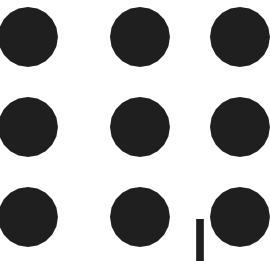
III Year / V Semester

Unit 5 – Big Data Database
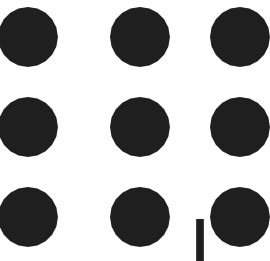
Topic – Hive

Hive

- Apache Hive is a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale., and is used for analyzing structured and semi-structured data.

- It was developed by the Data Infrastructure Team at Facebook. Hive is also one of the technologies that are being used to address the requirements at Facebook.

- A data warehouse provides a central store of information that can easily be analyzed to make informed, data driven decisions.

- Hive allows users to read, write, and manage petabytes of data using SQL.

- Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets.

**Hive /Big Data Analytics /Priyadharshini/ SNSCE**

Hive

- As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data.

- What makes Hive unique is the ability to query large datasets, leveraging Apache Tez or MapReduce, with a SQL-like interface.

- Hive was created to allow non-programmers familiar with SQL to work with petabytes of data, using a SQL-like interface called HiveQL.

- Traditional relational databases are designed for interactive queries on small to medium datasets and do not process huge datasets well.

- Hive instead uses batch processing so that it works quickly across a very large distributed database. Hive transforms HiveQL queries into MapReduce or Tez jobs that run on Apache Hadoop's distributed job scheduling framework, Yet Another Resource Negotiator (YARN).

**Hive /Big Data Analytics /Priyadharshini/ SNSCE**
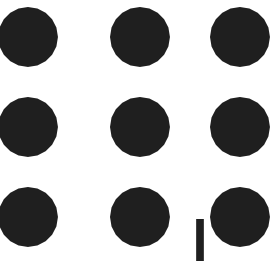
Hive

- It queries data stored in a distributed storage solution, like the Hadoop Distributed File System (HDFS) Hbase, or Amazon S3.

- Hive stores its database and table metadata in a metastore, which is a database or file backed store that enables easy data abstraction and discovery.

- Hive includes HCatalog, which is a table and storage management layer that reads data from the Hive metastore to facilitate seamless integration between Hive, Apache Pig, and MapReduce.

- By using the metastore, HCatalog allows Pig and MapReduce to use the same data structures as Hive, so that the metadata doesn't have to be redefined for each engine.

Hive /Big Data Analytics /Priyadharshini/ SNSCE
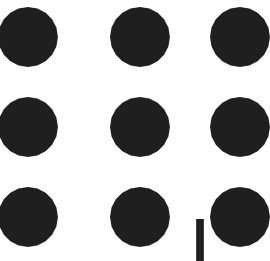
Hive

- Hive includes HCatalog, which is a table and storage management layer that reads data from the Hive metastore to facilitate seamless integration between Hive, Apache Pig, and MapReduce.

- By using the metastore, HCatalog allows Pig and MapReduce to use the same data structures as Hive, so that the metadata doesn't have to be redefined for each engine.

Features

- Allows programmers to plug in custom Mappers and Reducers.
- Has Data Warehouse infrastructure.
- Provides tools to enable easy data ETL.
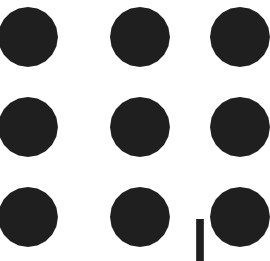- Defines SQL-like query language called QL

Hive /Big Data Analytics /Priyadharshini/ SNSCE

Hive

Hive

Advantages of Hive

- Useful for people who aren't from a programming background as it eliminates the need to write complex MapReduce program.

- Extensible and scalable to cope up with the growing volume and variety of data, without affecting performance of the system.

- It is as an efficient ETL (Extract, Transform, Load) tool.

- Hive supports any client application written in Java, PHP, Python, C++ or Ruby by exposing its Thrift server.

- As the metadata information of Hive is stored in an RDBMS, it significantly reduces the time to perform semantic checks during query execution.

Hive /Big Data Analytics /Priyadharshini/ SNSCE
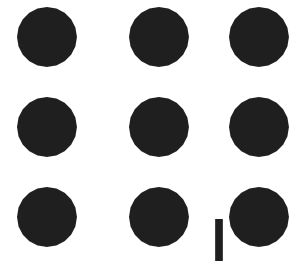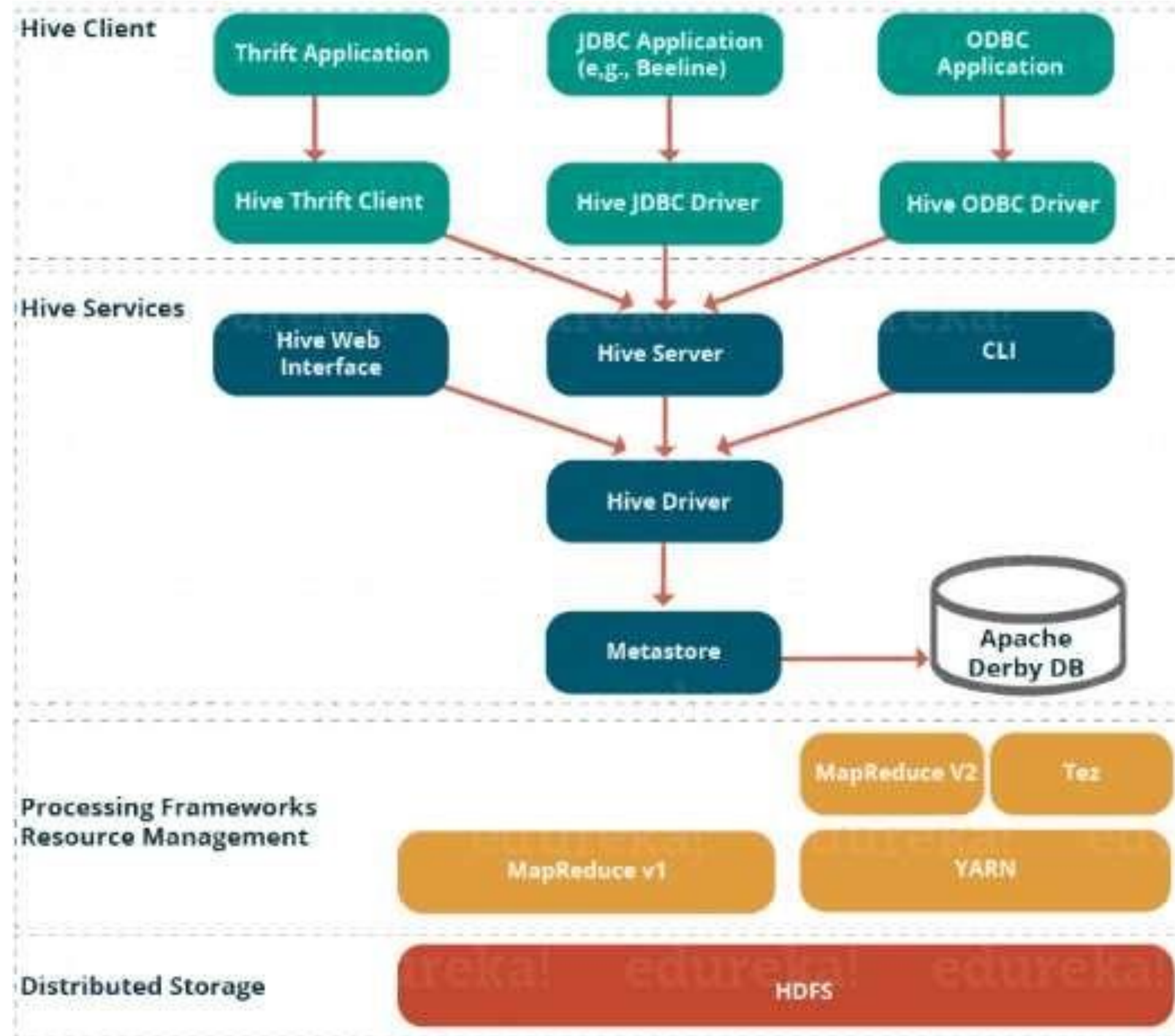
Hive
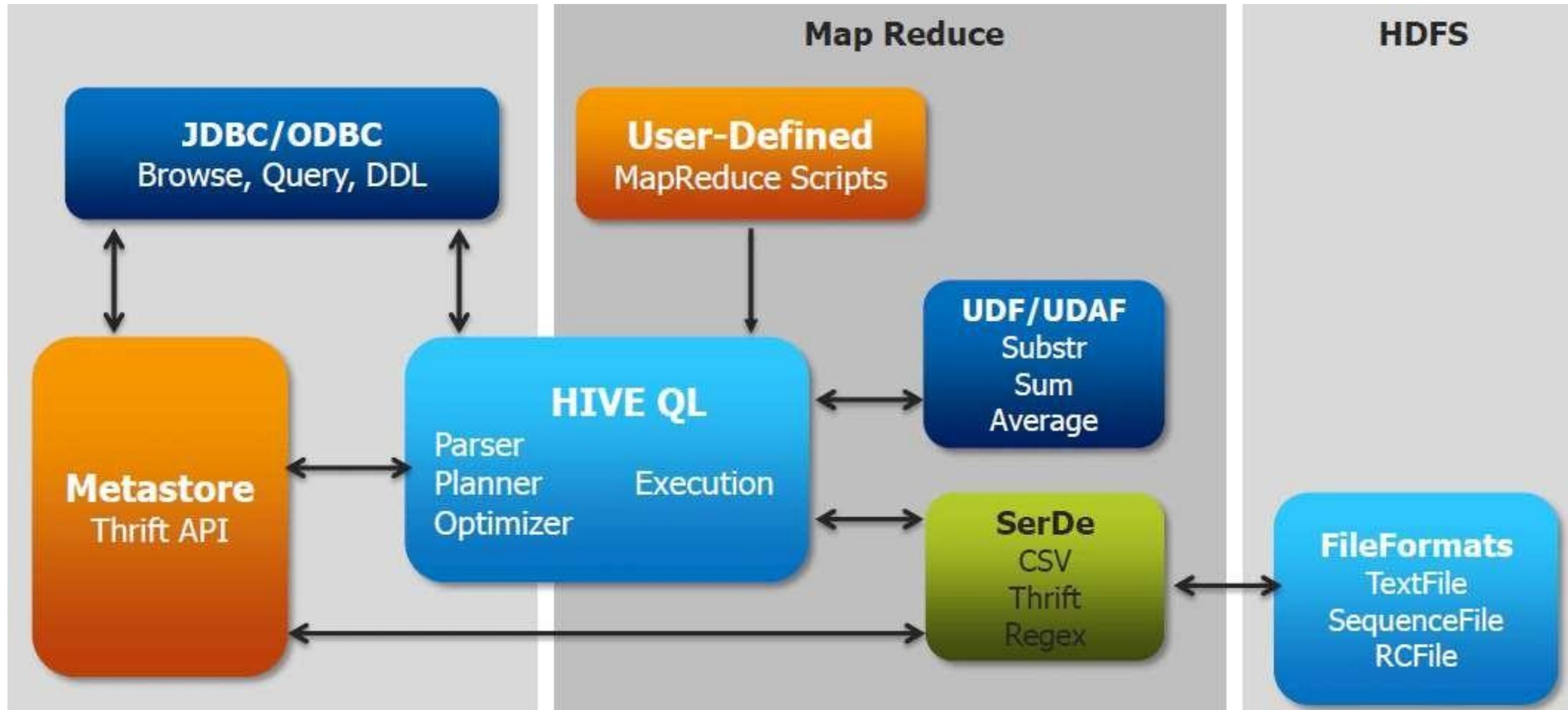
Architecture of Hive

Hive consists of the following major components:

- Metastore – To store the metadata.

- JDBC/ODBC – Query Compiler and Execution Engine to convert SQL queries to a sequence of MapReduce.

- SerDe and ObjectInspectors – For data formats and types.

- UDF/UDAF – For User Defined Functions.

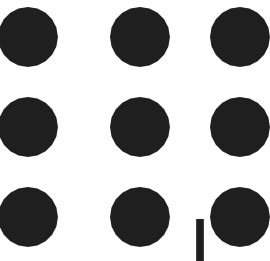- Clients – Similar to MySQL command line and a web UI.

Hive /Big Data Analytics /Priyadharshini / SNSCE
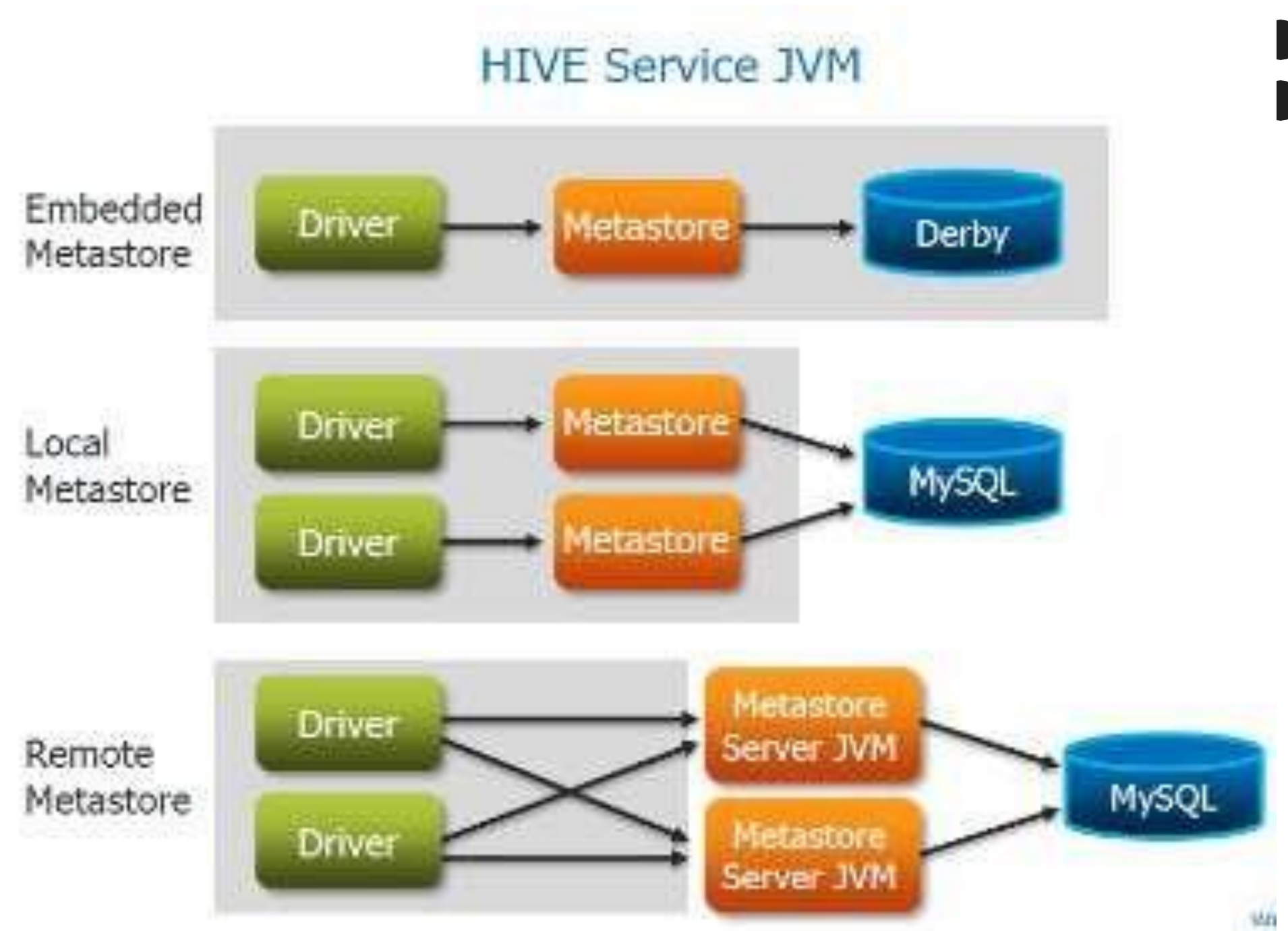
Hive

Architecture of Hive

Hive

Architecture of Hive

- 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 theirchoice.

- 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 asde 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.

Hive /Big Data Analytics /Priyadharshini / SNSCE
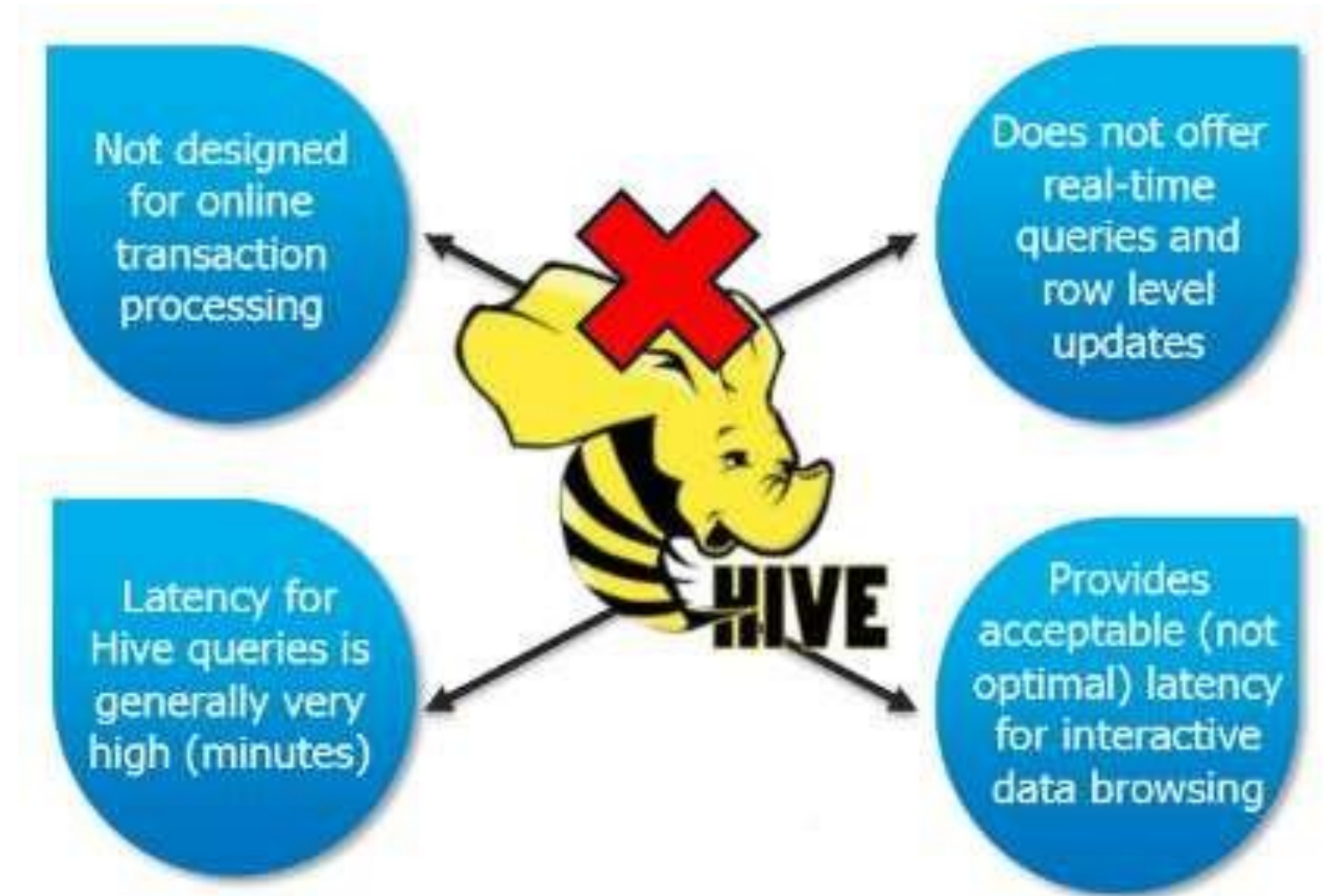
Hive

Metastore:

- The Metastore stores the information about the tables, partitions, the columns within the tables.

- There are 3 ways of storing in Metastore: **Embedded Metastore, Local Metastore and Remote Metastore**.

- Mostly, Remote Metastore will be used in production mode.

Hive /Big Data Analytics /Priyadharshini / SNSCE

Hive

Hive has the following limitations and cannot be used under such circumstances:

- Not designed for online transaction processing.

- Provides acceptable latency for interactive data browsing.

- Does not offer real-time queries and row level updates.

- Latency for Hive queries is generally very high.

Hive /Big Data Analytics /Priyadharshini / SNSCE

THANK YOU

Hive /Big Data Analytics /Priyadharshini / SNSCE