SNo	HADOOP	RDBMS
1	It can process any kind of data such as Structured, Semi-Structured and Un-Structured data	Mostly Structured data are processed by RDBMS
2	Processing coupled with Data Storage.	Mainly for data storage and limited (or) no data processing.
3	Schema is required on READ. Schema WRITE is just a file copy.	Schema is required on WRITE. Data type is validated while writing data.
4	WRITE Performance is faster than READ. Because schema is validated while reading data.	READ Performance is faster than WRITE. Because schema is validated while writing data. Due to INDEX feature, READ will be very faster in RDBMS environment.
5	Software license is not required as it is OPEN SOURCE. Only pay for support if it is required.	Cost is applicable for Software License.
6	Best fit for processing BIG Data, Unstructured Data and Massive Data Storage\Processing.	Best fit for OLTP environment.

Hive is a data warehouse system which is used to analyze structured data. It is built on the top of Hadoop. It was developed by Facebook.

Hive provides the functionality of reading, writing, and managing large datasets residing in distributed storage.

It runs SQL like queries called HQL (Hive query language) which gets internally converted to MapReduce jobs.

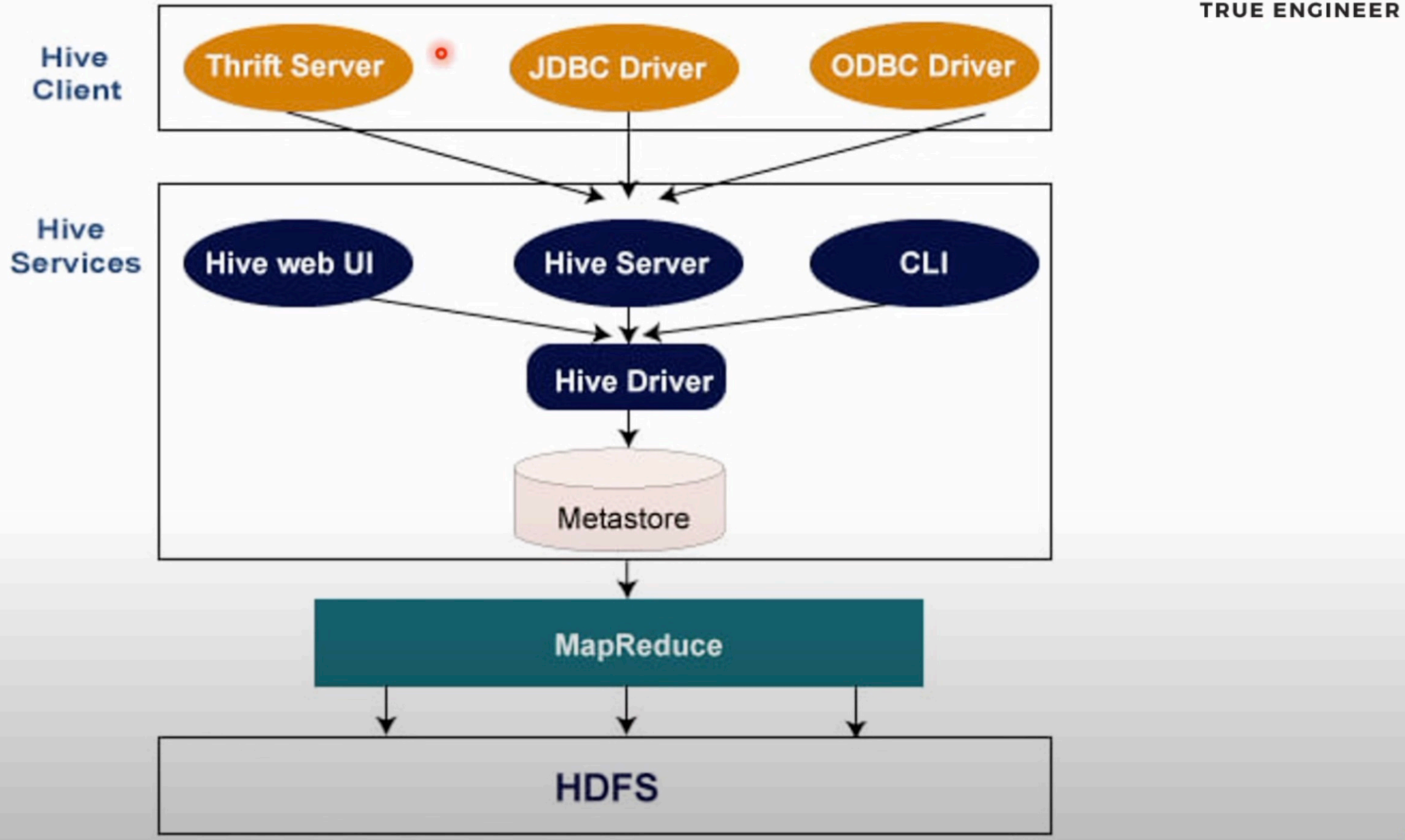
Using Hive, we can skip the requirement of the traditional approach of writing complex MapReduce programs. Hive supports Data Definition Language (DDL), Data Manipulation Language (DML), and User Defined Functions (UDF).

Features of Hive:

- Hive is fast and scalable.
- It provides SQL-like queries (i.e., HQL) that are implicitly transformed to MapReduce or Spark jobs.
- It is capable of analyzing large datasets stored in HDFS.
- It allows different storage types such as plain text, RCFile, and HBase.
- It uses indexing to accelerate queries.
- It can operate on compressed data stored in the Hadoop ecosystem.
- It supports user-defined functions (UDFs) where user can provide its functionality.

Limitations of Hive:

- Hive is not capable of handling real-time data.
- It is not designed for online transaction processing.



Hive Client:

Hive allows writing applications in various languages, including Java, Python, and C++. It supports different types of clients such as:-

•Thrift Server - It is a cross-language service provider platform that serves the request from all those programming languages that supports Thrift.

•JDBC Driver - It is used to establish a connection between hive and Java applications.

•ODBC Driver - It allows the applications that support the ODBC protocol to connect to Hive.

TRUE ENGINEER

Hive Services:

•Hive CLI - The Hive CLI (Command Line Interface) is a shell where we can execute Hive queries and commands.

•Hive Web User Interface - The Hive Web UI is just an alternative of Hive CLI. It provides a web-based GUI for executing Hive queries and commands.

•Hive MetaStore - It is a central repository that stores all the structure information of various tables and partitions in the warehouse. It also includes metadata of column and its type information, the serializers and de-serializers which is used to read and write data and the corresponding HDFS files where the data is stored.

•Hive Server – It is referred to as Apache Thrift Server. It accepts the request from different clients and provides it to Hive Driver.

•Hive Driver – It receives queries from different sources like web UI, CLI, Thrift, and JDBC/ODBC driver. It transfers the queries to the compiler.

•Hive Compiler – The purpose of the compiler is to parse the query and perform semantic analysis on the different query blocks and expressions. It converts HiveQL statements into MapReduce jobs.

TRUE ENGINEER

Pig is a high-level data flow platform for executing Map Reduce programs of Hadoop. It was developed by Yahoo. The language for Pig is pig Latin.

What is Apache Pig?

Apache Pig is a high-level data flow platform for executing MapReduce programs of Hadoop. The language used for Pig is Pig Latin.

The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Apart from that, Pig can also execute its job in Apache Tez or Apache Spark.

Pig can handle any type of data, i.e., structured, semi-structured or unstructured and stores the corresponding results into Hadoop Data File System. Every task which can be achieved using PIG can also be achieved using java used in MapReduce.

MapReduce	Apache Pig
Abstraction is at lower level	Abstraction is at higher level
Needs more lines of code	Needs a less lines of code
Needs more development effort	Needs less development effort
Can be written using language like Java or Python	Can be written using a SQL like scripting language (Pig Latin)

Pig Data Types:

1.Primitive Types:

- **int**: Integer
- **long**: Long Integer
- **float**: Floating-point
- **double**: Double-precision floating-point
- **chararray**: Character array (string)
- **bytearray**: Byte array (binary data)

2.Complex Types:

- **tuple**: Ordered set of fields (similar to a row in a table)
- **bag**: Unordered collection of tuples
- **map**: Collection of key-value pairs

Hive Data Types :

1.Primitive Types:

- **TINYINT**: 8-bit integer
- **SMALLINT**: 16-bit integer
- **INT**: 32-bit integer
- **BIGINT**: 64-bit integer
- **FLOAT**: Single-precision floating-point
- **DOUBLE**: Double-precision floating-point
- **BOOLEAN**: Boolean (true/false)
- **STRING**: Variable-length character string
- **CHAR**: Fixed-length character string
- **VARCHAR**: Variable-length character string

2.Complex Types:

- **ARRAY**: Ordered collection of elements
- **MAP**: Unordered collection of key-value pairs
- **STRUCT**: Ordered set of named fields
- **UNION**: Represents multiple possible types for a column

Hbase :

Apache HBase is an open-source, NoSQL, distributed database.

It enables random, strictly consistent, real-time access to petabytes of data. HBase is very effective for handling large, sparse datasets.

HBase integrates seamlessly with Apache Hadoop and the Hadoop ecosystem and runs on top of the Hadoop Distributed File System (HDFS) or Amazon S3 using Amazon Elastic MapReduce (EMR) file system, or EMRFS. HBase serves as a direct input and output to the Apache MapReduce framework for Hadoop, and works with Apache Phoenix to enable SQL-like queries over HBase tables.

How does HBase work?

HBase is a column-oriented, non-relational database. This means that data is stored in individual columns, and indexed by a unique row key. This architecture allows for rapid retrieval of individual rows and columns and efficient scans over individual columns within a table. Both data and requests are distributed across all servers in an HBase cluster, allowing you to query results on petabytes of data within milliseconds. HBase is most effectively used to store non-relational data, accessed via the HBase API. Apache Phoenix is commonly used as a SQL layer on top of HBase allowing you to use familiar SQL syntax to insert, delete, and query data stored in HBase.

Benefits of HBase?

Scalable

HBase is designed to handle scaling across thousands of servers and managing access to petabytes of data. With the elasticity of Amazon EC2, and the scalability of Amazon S3, HBase is able to handle online access to massive data sets.

Fast

HBase provides low latency random read and write access to petabytes of data by distributing requests from applications across a cluster of hosts. Each host has access to data in HDFS and S3, and serves read and write requests in milliseconds.

Fault-Tolerant

HBase splits data stored in tables across multiple hosts in the cluster and is built to withstand individual host failures. Because data is stored on HDFS or S3, healthy hosts will automatically be chosen to host the data once served by the failed host, and data is brought online automatically.

HBASE	RDMS
Open source, Column oriented, Distributed database storage system	Row oriented, Tabular structured Data only
HBASE Handle Structured and Semi-Structured Data	RDBMS only handles Structured Data
Cannot be used for Transaction processing	Can be used for Transaction processing
Multidimensional database	Two dimensional database
Flexible Schema – as column can be easily added for specific column	Fixed Schema

TRUE ENGINEER

Apache Hadoop ecosystem refers to the various components of the Apache Hadoop software library

It includes open source projects as well as a complete range of complementary tools.

Some of the most well-known tools of the Hadoop ecosystem include HDFS, Hive, Pig, YARN, MapReduce, Spark, Hbase *etc.*

What is HDFS?

Hadoop Distributed File System (HDFS), is one of the largest Apache projects and primary storage system of Hadoop. It employs a NameNode and DataNode architecture. It is a distributed file system able to store large files running over the cluster of commodity hardware.

What is Hive?

Hive is an ETL and Data warehousing tool used to query or analyze large datasets stored within the Hadoop ecosystem. Hive has three main functions: data summarization, query, and analysis of unstructured and semi-structured data in Hadoop. It features a SQL-like interface, HQL language that works similar to SQL and automatically translates queries into MapReduce jobs.

TRUE ENGINEER

What is Apache Pig?

This is a high-level scripting language used to execute queries for larger datasets that are used within Hadoop. Pig's simple SQL-like scripting language is known as Pig Latin and its main objective is to perform the required operations and arrange the final output in the desired format.

What is MapReduce?

This is another data processing layer of Hadoop. It has the capability to process large structured and unstructured data as well as to manage very large data files in parallel by dividing the job into a set of independent tasks (sub-job).

TRUE ENGINEER

What is YARN?

YARN stands for Yet Another Resource Negotiator. It is one of the core components in open source Apache Hadoop suitable for resource management. It is responsible for managing workloads, monitoring, and security controls implementation. It also allocates system resources to the various applications running in a Hadoop cluster while assigning which tasks should be executed by each cluster nodes.

YARN has two main components:

- Resource Manager
- Node Manager

What is Apache Spark?

Apache Spark is a fast, in-memory data processing engine suitable for use in a wide range of circumstances. Spark can be deployed in several ways, it features Java, Python, Scala, and R programming languages, and supports SQL, streaming data, machine learning, and graph processing, which can be used together in an application.