Apache HIVE

Notes By Mannan Ul Haq

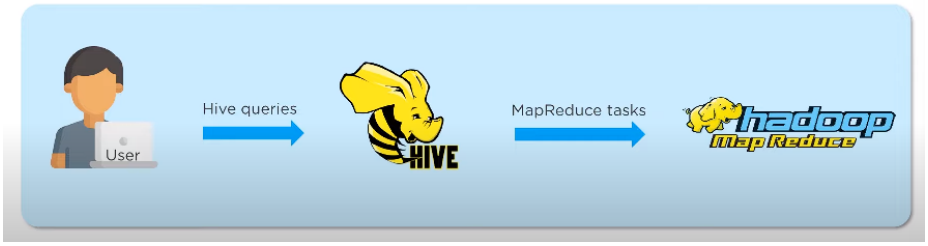
Introduction

- Facebook used Hive as a solution to handle the growing big data.
- As we know, Hadoop uses MapReduce for processing data. MapReduce required users to write long codes.
- Not all users were well versed with Java and other coding languages. This proved to be a disadvantage for them.
- Users were comfortable with writing queries in SQL.
- Hive was developed with a vision to incorporate the concepts of tables, columns just like SQL.

What is Hive?

Hive is a data warehouse system which is used for querying and analyzing large datasets stored in HDFS.

Hive uses a query language call HiveQL which is similar to SQL.



Working

1. **Handling Large Data**: Hive is designed to handle petabytes of data efficiently. It provides a familiar SQL interface for users to interact with massive datasets.
2. **Batch Processing**: It employs batch processing techniques, enabling it to operate swiftly even across extensive distributed databases. This ensures that operations can be performed efficiently on large volumes of data.
3. **Transformation to MapReduce**: Hive translates queries written in its SQL-like language, HiveQL, into MapReduce jobs. These jobs are then executed on Apache Hadoop's distributed processing framework, leveraging its parallel computing capabilities.
4. **Utilization of YARN**: Apache Hive utilizes Apache Hadoop's YARN (Yet Another Resource Negotiator) framework for job scheduling and resource management. YARN efficiently allocates resources across the cluster, ensuring optimal performance during query execution.

5. **Data Source Compatibility**: Hive can query data stored in various distributed storage solutions, such as the Hadoop Distributed File System (HDFS) or cloud storage services like Amazon S3. This flexibility allows users to analyze data regardless of its storage location.
6. **Metadata Management**: Hive maintains metadata about its databases and tables in a centralized repository known as the metastore. This metastore stores essential information about the schema, partitioning, and location of the data, enabling efficient query processing and optimization.

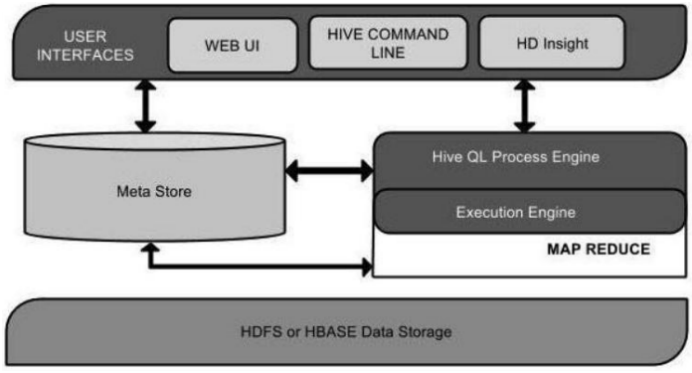
Features and Benefits

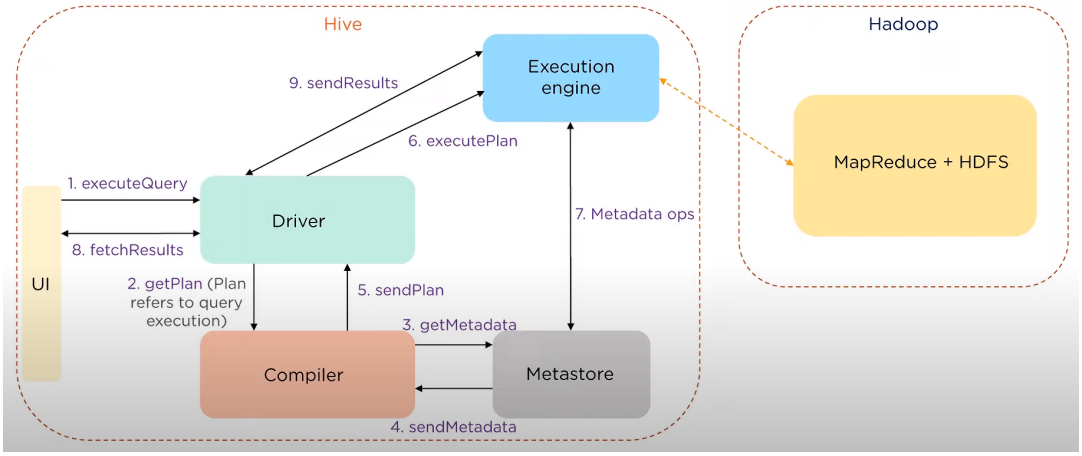
1. Uses Data Stores to create tables similar to SQL.
2. Easy to code and scalable.
3. Supports SQL, Join, Group By, and Order by clauses.
4. Supports custom types and custom functions.
5. Translates queries into MapReduce jobs.
6. Provides rich data types, Structs, Map and Array.
7. Supports web interfaces as well.

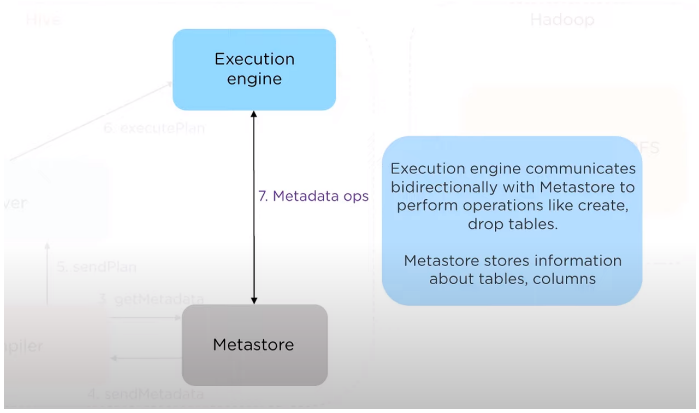
Limitations

1. Not a full database. HiveQL doesn't support the full range of SQL features found in traditional databases.
2. Not developed for unstructured data.
3. Performs the partitions always from the last column.
4. Not used for real time queries as it takes a bit of time to give the results.
5. Support for Updates and Deletion is very minimal.

Architecture



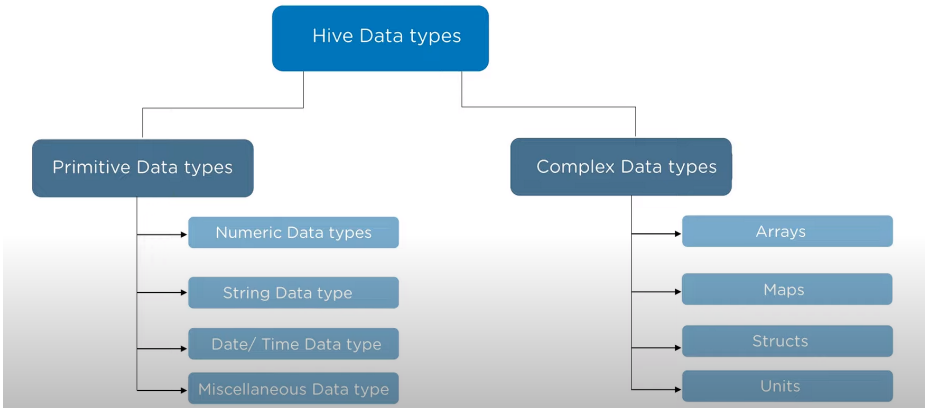




Difference between Hive and RDBMS

Characteristics	Hive	RDBMS
Record level queries	No Update and Delete	Insert, Update and Delete
Transaction support	No	Yes
Latency	Minutes or more	In fractions of a second
Data size	Petabytes	Terabytes
Data per query	Petabytes	Gigabytes
Query language	HiveQL	SQL
Support JDBC/ODBC	Limited	Full

Hive Data Types



Name	Description
STRUCT	Similar to 'C' struc, a collection of fields of different data types. An access to field uses dot notation. For example, struct ('a', 'b')
MAP	A collection of key-value pairs. Fields access using [] notation. For example, map ('key1', 'a', 'key2', 'b')
ARRAY	Ordered sequence of same types. Accesses to fields using array index. For example, array ('a', 'b')

Hive Data Model

Name	Description
Database	Namespace for tables
Tables	Similar to tables in RDBMS Support filter, projection, join and union operations The table data stores in a directory in HDFS
Partitions	Table can have one or more partition keys that tell how the data stores
Buckets	Data in each partition further divides into buckets based on hash of a column in the table. Stored as a file in the partition directory.