CAE - 1

Name : Mohnish Devaraj

Reg No: 39110636

Roll No: 195115398

Date : 02 Feb 2022

Subject Name: Big Data Analytics

Subject code: SCSA1603

Total pages: 8

PART - B

(8) Architecture of Hive

Apache Hive is an open source data ware housing system built on top of Hadoop for querying and analyzing large datasets stored in Hadoop files. It process-structured and semi-structured data in Hadoop.

Hive use language called HiveQL (HQL), which is similar to SQL. HiveQL automatically translates SQL-like queries into Map Reduce jobs.

Meta store - It stores meta data for each of the tables like their schema and location

Driver - It acts like a controller which receives the HQL statements.

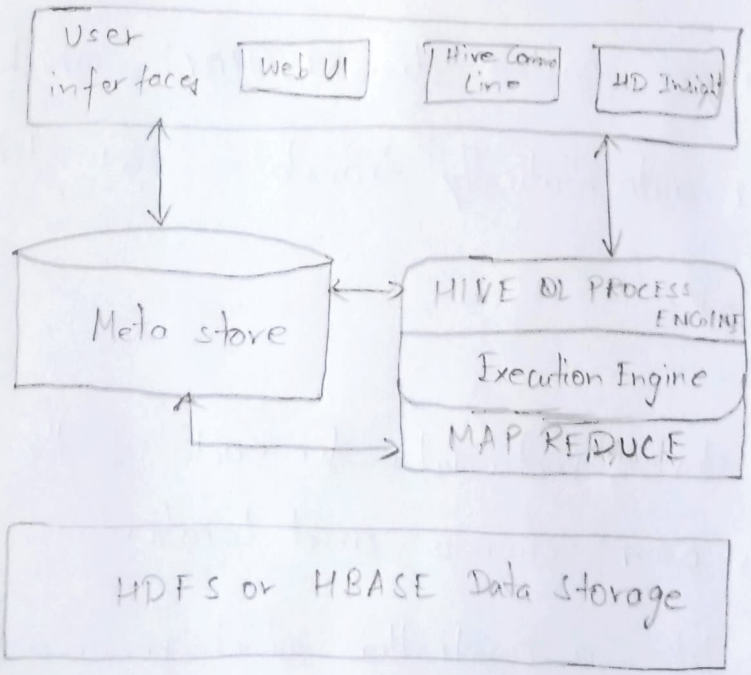
Compiler - It performs the compilation of the Hive QL query.

Optimizer - It performs various transformations on the execution plan provide optimized DAG.

Executor - Once compilation and optimization complete, the executor executes the tasks. Executor takes care of pipelining the tasks.

Features of Hive

- It stores schema in a database and processed data into HDFS.
- It is designed for OLAP.
- It provides SQL type Language for querying called HiveQL
- It is familiar, fast, scalable, and extensible.



User Interface

Hive is a data warehouse infrastructure software that can create interactive between user and HDFS. The user interfaces that Hive supports are Hive Web UI, Hive command Line and Hive HD sight.

Meta store

Hive chooses respective database servers to store the schema or Meta Data of tables, databases, columns in a table, their data types and HDFS mapping.

HiveQL Process Engine

HiveQL is similar to SQL for querying on schema info on the Metastore. It is one of the replacements of traditional approach for MapReduce program. Instead of writing mapreduce job and process it.

Execution Engine

The conjunction part of HiveQL process Engine and MapReduce is Hive Execution Engine. Execution Enigne processes the query and generates results as same as Map reduce Results.

It uses the flavour of Map Reduce.

~~DSS on Hadoop~~

(6) Big Data is defined by the 5Vs of Big data which are also termed as the characteristics of Big data as follows:

a) Volume:

The name 'Big Data' itself is related to a size which is enormous. To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not is dependent upon the volume of Data. Hence while dealing with Big Data, it is necessary to consider a characteristic 'Volume'.

b) Velocity:

Velocity refers to the high speed of accumulation of data. In Big data velocity data flows in from sources like machines, networks, social media, mobile phones, etc. There is a massive and continous flow of data. This determines the

potential of data that how fast the data is generated and procased to meet the demands. sampling data can help in dealing with the issue like 'velocity'.

c) Variety:

It refers to nature of data that is structured, semi-structured and unstructured data. It also refers to heterogenous resource. Variety is basically the arrival of data from new source that are both inside and outside of an enterprise.

-> structured data
-> unstructured data
-> semi-structured data.

d) Veracity:

It refers to inconsistencies and uncertainity in data, that is data which is available can sometimes get messy and quality and accurate are difficult to control. Big data is also variable because of the multim multitude of data dimensions resulting from multiple disparate data types and sources.

e) Value:

After having the 4Vs into account there comes one more V which stands for Value. The bulk of data having no value

is of no good to the company, unless you turn it into something usefull. Data in itself is of no use or importance but it needs to be converted into something valuable to extract information. Hence, you can state Value is the most important V's of all the 5V's.

PART-A

① a) One of the major challenges of conventional system was the uncertainty of Data Management Landscape.

b) Big data is continously expanding, there are new companies and technologies that are being developed every day.

c) They are so large that it is not possible to work on them with traditional analytical tools.

d) These are complex data sets that can be both structured and unstructured.

② There are five characteristics of Big data:

a) Volume

b) Velocity

d) Veracity

d) Variety

e) Value

③ Structured Data

→ It is based on relational database table.

→ Matured transaction & various concurrency techniques.

→ Versioned over tuples, row, tables

→ It is schema dependent & less flexible

Unstructured data

→ It is based on character and binary data

→ No transaction management and no concurrency

→ versioned as a whole

→ It is more flexible & their is absence of schema.

④ () → Tuple Constructor operator

This operator is used to construct a tuple

Eg: (Mohnish, 20)

{} → Bag Constructor Operator

This operator is used to construct a bag

Eg: {(Mohnish, 20), (Hariharan, 20)}

[] - Map Constructor operator

This operator is used to construct a tuple

Eg: [name # Mohnish, age # 20]

(5) There are three types of Complex type of Hive:

Arrays:

It is an ordered collection of elements

Syntax: ARRAY <data_type>

Map:

It is an unordered collection of key-value pairs

syntax: MAP <primitive_type, data_type>

Structs:

It is an collection of elements of different tuples.

syntax: STRUCT < col_name: data_type [COMMENT col_comment], ...>