**Session 14: Hive Introduction**

**Assignment 1**

1. When Hive is best suited and when is it not?

**Answer:** Hive is best suited for Data Warehousing Applications where data is stored, mined and reporting is done based on processing. It is used to process batch jobs on huge data which is immutable.

Hive is not suited for small amount of data as at end it’s code get converted into mapreduce and mapreduce is best suited for large amount of data .

2. When should one use Hive over MapReduce?

**Answer:** Hive should be used at the time of ad-hoc querying and analysis of Big data on Hadoop. And at the time of joins hive is best suited.

3. What is Hive metastore?

**Answer:** stores the metadata for Hive tables and partitions in a relational database, and provides clients (including Hive) access to this information via the metastore service API. The metastore service runs in the same JVM as the Hive service and contains an embedded Derby database instance

4. How can Hive improve performance with orc file format tables?

**Answer:** Optimized Row Columnar File Format is better:

a) can reduce the size of original data up to 75%.Comparing to Text,Sequence,Rc file formats ORC is better .

b) ORC takes less time to access the data

c) ORC takes Less space to store data.

d) a single file as the output of each task, which reduces the NameNode's load

e) concurrent reads of the same file using separate RecordReaders

5. What is thrift server and client, jdbc and odbc driver importance in hive?

**Answer :**

**Thift Server:-** is an independent entity. In production environment to provide various interfaces that enables remote clients to execute queries against Hive and retrieve the results. Client pass command in string format and thrift server converted it into driver understand format for interaction

**JDBC Driver:**

|  |
| --- |
| Using jdbc driver a Java application will connect to a Hive server running in a separate process at the given host and port. |

**ODBC Driver:**

|  |
| --- |
| An ODBC driver allows applications that support the ODBC protocol (such as business intelligence software) to connect to Hive. |

6. What is the importance of partition in hive?

**Answer:**

a) Partitioning is used for distributing execution load horizontally.

b) As the data is stored as slices/parts, query response time is faster to process the small part of the data instead of looking for a search in the entire data set, thus enhance query performance.

c) For example, In a large user table where the table is partitioned by country, then selecting users of country ‘IN’ will just scan one directory ‘country=IN’ instead of all the directories.

7. What is the use of bucketing in hive?

**Answer:** If we want to create a lot of tiny partitions than we use bucketing. We can decompose data into more manageable parts known as buckets. It is based on the value of a hash function of some column of the Table

8. What is the difference between static partitioning and dynamic partitioning in hive?

**Answer:**

**Static Partitioning:** When we know about the data and it’s partition and separate hive statement for every partitions.

In DML/DDL involving multiple partitioning columns, the columns whose values are known at COMPILE TIME (given by user).

Usually when loading files (big files) into Hive tables static partitions are preferred. That saves your time in loading data compared to dynamic partition

**Dynamic Partitioning:** As many partitions with single hive statement. Used when the values for partition columns are known only during loading of the data into a **Hive** table.

In case of dynamic partition whole big file i.e. every row of the data is read and data is partitioned through a MR job into the destination tables depending on certain field in file.