Session 14: Hive IntroductionAssignment 1

# Problem Statement

**Give brief answers to the below questions:**

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

**Ans:**

Hive is best suited in:

When we are doing data warehouse applications

When we are getting static data instead of dynamic data

When the application on high latency (response time high)

Where a large set is maintained and mined for insights, reports

When we are using queries instead of scripting we use hive.

Hive is not suitable for:

It doesn't provide OLTP transactions support only OLAP transactions

If applications required OLTP, switch to NoSQL database

HQL queries have higher latency, due to the mapreduce

1. **When should one use Hive over MapReduce?**

**Ans:**

One should use hive over mapreduce when there is a need to use 12 or more levels of nested from clause in the query as it will be difficult to implement in mapreduce.

There are partitions to simplify the data process, Bucketing for sampling the data, sort the data quickly and simplify the mapreduce process. Partitions and Buckets can segmenting large data sets to improve Query performance in Hive.

So it is highly recommendable for structure data.

1. **What is Hive metastore?**

**Ans:**

Hive metastore is the central repository to store metadata of hive tables.

Metastore consists of:

* Metastore Service: provides the interface to hive
* Datastore: Stores the data definitions and mapping to the data and others.

The metadata that metastore stores contains things like: ids of database, ids of index, ids of tables, when the table was created etc.

To store the metadata hive can use any of the below 3 strategies:

* Embedded: Mainly use for unit test.

Only one process is allowed to connect to the metastore at a time.

Hive metadata is stored in an embedded Apache Derby database.

In this metastore, both the database and the metastore service run embedded in the main HiveServer process.

* Local: In this mode, we can use an external database which is JDBC compliant like MySQL.

Local metastore allows multiple connections at a time.

Here, the Hive metastore service runs in the main HiveServer process and the metastore database runs in a separate process.

* Remote: In this mode, Hive driver and the metastore interface runs in a different JVM (which can run on different machines also) as shown in the diagram below.

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

**Ans:**

ORC file format provides very efficient way to store the relational data. By using orc files format we can reduce the size of original data up to 75%.ORC files are better than text, sequence or RC files.

ORC files provide bet performance while reading, writing and processing data trough hive. ORC files takes less time to access the data and needs less space to store the data. ORC files improves the performance of Hive query in terms of data access and storing.

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

**Ans:**

Thrift is a cross language RPC framework which generate code and combines a software stack finally execute the thrift code in remote server. Thrift compiler acts as interpreter between server and client. Thrift server allows a remove client to submit request to hive, using different programming languages like Python, Ruby and scala.

JDBC driver : Is a software component enabling a Java application to interact with a database.

ODBC driver: ODBC accomplishes DBMS independence by using an ODBC driver as a transaction layer between the application and the DBMS.

**6. What is the importance of partition in hive?**

**Ans:**

To improve the query performance partition is used in hive.

**7. What is the use of bucketing in hive?**

**Ans:**

Bucketing is helpful for :

1. Enables more efficient queries
2. Makes sampling more efficient
3. Hash(column) MOD(number of buckets) – evenly distributed

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

**Ans:**

To prune data during query, partition can minimize the query time. The partition is created when the data is inserted into table. Static partition can insert individual rows where Dynamic partition can process entire table based on a particular column. At least one static partition is must to create any (static, dynamic) partition.