1.

Hive should not be used:

Hive use reduces the task of writing mapreduce programs

Custom mappers and reducers can be plugged in hive when writing logic in HiveQL is not easy.

Hive is best suited for OLAP process.

To process structured data in Hadoop, hive is best suited.

Hive should be used for OLTP

It is not used for real time queries.

It provides no support for updation/ deletion of data.

2. Hive was developed so that it can be used by SQL experts to perform MR jobs.

It is easy to write SQL query rather writing Java MR programs.

It take very less time to write Hive Query compared to Map Reduce code. For example, the word count problem which takes around 50 lines of code can be written in 5 lines in Hive. So, you save time.

It has very low maintenance and is very simple to learn & use (low learning curve).

It is very easy to write query involving joins (if there are few joins) in Hive compared to MR programs.

3. Hive metastore is a service that stores metadata information of hive tables in relational databases.

Metadata that metastores have stores the following details:

IDs of Database

IDs of Tables

IDs of Index

The time of creation of an Index

The time of creation of a Table

InputFormat used for a Table

OutputFormat used for a Table

4. Since ORC file format has data divided into columns, it stores statistics like min, max,sum, etc. Data is accessed through index, hive is best suited to use ORC file format. The performance can further be improved by partitioning and bucketing the table

5. Using thrift you can call hive commands from a various programming languages e.g. C++, Java, PHP, Python and Ruby. Thrift HiveServer is a service that allows a remote client to submit requests to Hive, using a variety of programming languages, and retrieve results. JDBC/ODBC driver helps to integrate Hive with other applications.

6. Since hive is used to perform queries on large datasets especially datasets that require full table scan, so to improve performance where to limit the search of data partition in hive is used. Table is paritioned based on column values with low cardinality (which is column having less number of unique values).

In non-partitioned tables, Hive would have to read all the files in a table’s data directory and subsequently apply filters on it. This is slow and expensive—especially in cases of large tables.

In case of partitioned tables, subdirectories are created under the table’s data directory for each unique value of a partition column. So writing queries related to that column data and searching the set of rows with the partitioned data will be easier as it will directly lookup for data in that directory rather than searching in whole table.

7. Like partitioning of data, bucketing is other mechanism to segregate hive table into multiple files. Partitioning gives effective result when there is less number of partition and comparative equal size of partitions. But in real situation this is not possible for equivalent size of partition. One partition csize can be huge while other is small. So to overcome this situation bucket is used which is based on hash function based on values of column. It help in dividing the table in equal size partition. It also helps in efficient map side joins.

8. in static partitioning we need to specify the partition column value in each and every LOAD statement.

suppose we are having partition on column country for table t1(userid, name,occupation, country), so each time we need to provide country value

hive>LOAD DATA INPATH '/hdfs path of the file' INTO TABLE t1 PARTITION(country="US")

hive>LOAD DATA INPATH '/hdfs path of the file' INTO TABLE t1 PARTITION(country="UK")

dynamic partition allow us not to specify partition column value each time. the approach we follows is as below:

i) create a non-partitioned table t2 and insert data into it.

ii) now create a table t1 partitioned on intended column(say country).

iii) load data in t1 from t2 as below:

hive> INSERT INTO TABLE t2 PARTITION(country) SELECT \* from T1;

If we want to use Static partition in hive we should set property set hive.mapred.mode = strict This property set by default in hive-site.xml

Static partition is in Strict Mode

Here is hive dynamic partition properties

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;