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

It provides us data **warehousing** facilities on top of an existing Hadoop cluster. Along with that it provides an **SQL like** interface which makes the work easier, in case we are coming from an SQL background. We can create tables in Hive and store data there. And we can even map your existing HBase tables to Hive and operate on them. **Hive** should be used for analytical querying of data collected over a period of time. e.g Calculate trends, summarize website logs . **Hive** can be used for ad-hoc data analysis .

|  |
| --- |
|  |

Hive does not support OLTP transactions, only OLAP transactions are supported by Hive.Incase of low latency we can not use Hive.It can't be used for real time queries and it can't support all un-structured data formats unlike PIG.

2.When should one use Hive over MapReduce?

There are partitions to simplify the data process, bucketing to sample the data,sort the data quickly and simplify the mapreduce process.Partitions and buckets can segment large data sets to improve query performance in Hive.

3.What is Hive metastore?

It is a central repository in Hive which allows to store meta data in external database.Although in Hive 1 by default Hive stores metadata in Derby but we can also use Oracle or Mysql as per our project requirements.

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

Hive can ease out many Hive file format limitations by storing the data in ORC files very efficiently.Through ORC it can improve reading , writing and processing data by running the following commands and creating the table structure.

Set hive.compute.query.using.stats=true;

Set hive.class.dbclass=fs;

CREATE TABLE orc\_table (

Id int ,

Name string

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\;’

LINES TERMINATED BY ‘\n’

STORED AS ORC;

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

Thrift is a cross language RPC framework which generate code and combines a software stack finally to execute the thrift code in remote server. Thrift server works as interpreter between server and client. It allows a remote client to submit request to Hive using different programming languages.

A JDBC driver is a software component to enable a java application to interact with a database.

A ODBC driver is a translation layer between the application and the DBMS.

6.What is the importance of partition in hive?

To analyse a particular set of data ,through partitioning we can analyse it based on particular column.

7.What is the use of bucketing in hive?

We use bucketing to process many chunks of files, to analyse vast amount of data , sometimes to burst the process and time.

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

To prune data during query, partition can minimise the query time and it is created while data is inserted in table. While static partition can insert individual rows , Dynamic partition can process the entire table based on a particular column.One static partition is mandatory for creating any type of partition.In case of large dataset for performing a ETL FLOW dynamic partition is preffered.