1. **What is the relationship between MapReduce and Hive? or How Mapreduce jobs submit to the cluster?**

Hive provides no additional capabilities to MapReduce. The programs are executed as MapReduce jobs via the interpreter. The Interpreter runs on a client machine which rurns HiveQL queries into MapReduce jobs. Framework submits those jobs onto the cluster.

2. **How Hive can improve performance with ORC format tables?**

Hive can store the data in highly efficient manner in the Optimized Row Columnar (ORC) file format. It can ease many Hive file format limitations. Using ORC files can improves the performance when reading, writing, and processing data. Enable this format by running this command and create table like this.

```
set hive.compute.query.using.stats=true;

set hive.stats.dbclass=fs;

CREATE TABLE orc_table (
```

```
id int,
name string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\;'
LINES TERMINATED BY '\n'
STORED AS ORC;
```

3. **What is the importance of Vectorization in Hive?**

It's a query optimization technique. Instead of processing multiple rows, Vectorization allows to process process a batch of rows as a unit. Consequently it can optimize query performance. The file must be stored in ORC format to enable this Vectorization. It's disabled by default, but enable this property by running this command.

```
set hive.vectorized.execution.enabled=true;
```

4. **Difference between sort by or order by clause in Hive? Which is the fast?** ORDER BY – sort the data in one reducer.

SORT BY – sort the data within each reducer. You can use n number of reducers for sort.

In the first case (order by) maps sends each value to the single reducer and count them all.

In the second case (sort by) maps splits up the values to many reducers and each reduce generates its list and finds the count. So it can sort quickly.

Sort by is much faster than order by.

5. **What are different Hive metastore configurations?**

There are three types of metastores configuration called

1) Embedded metastore

2) Local metastore

3) Remote metastore.

If Hive run any query first it enter into embedded mode, It's default mode. In Command line all operations done in embedded mode only, it can access Hive libraries locally. In the embedded metastore configuration, hive driver, metastore interface and databases use same JVM. It's good for development and testing.

In local metastore the metastore store data in external databases like MYSQL. Here Hive driver and metastore run in the same JVM, but remotely communicate with external Database. For better protection required credentials in Local metastore.

Where as in Remote server, use remote mode to run the queries over Thift server.

In Remote metastore, Hive driver and metastore interface would be running in a different JVM. So for better protection, required credentials such are isolated from Hive users.

6. Can Hive process any type of databases?

Yes, Hive uses the SerDe interface for IO operations. Different SerDe interfaces can read and write any type of data. If normal directly process the data whereas different type of data is in the Hadoop, Hive use different SerDe interface to process such data.

7. **What Is the HWI?**

The Hive Web Interface is an alternative to the Hive command line interface. HWI is a simple graphical interface.

8. **What is the difference between Like and Rlike operators in HIVE?**

Like is used to find the substrings within a main string with regular expression %.

Rlike is a special function which also finds the sub strings within a main string, but return true or false without using regular expression.

9. **What are the Hive default read and write classes?**

Hive uses following classes to read and write the files.

TextInputFormat/HiveIgnoreKeyTextOutputFormat

SequenceFileInputFormat/SequenceFileOutputFormat

First class used to read/write the plain text. Second class used for sequence files.

10. **What is Query processor in Hive?**

It's a core processing unit in Hive framework, it converting SQL to map/reduce jobs and run in the other dependencies. As a result hive can convert the Hive queries into Hive queries.