HIVE INTERVIEW QUESTIONS

Dear readers, these **Hive Interview Questions** have been designed specially to get you acquainted with the nature of questions you may encounter during your interview for the subject of **Hive**. As per my experience good interviewers hardly plan to ask any particular question during your interview, normally questions start with some basic concept of the subject and later they continue based on further discussion and what you answer −

What are the different types of tables available in HIve?

There are two types. Managed table and external table. In managed table both the data an schema in under control of hive but in external table only the schema is under control of Hive.

Is Hive suitable to be used for OLTP systems? Why?

No Hive does not provide insert and update at row level. So it is not suitable for OLTP system.

Can a table be renamed in Hive?

Alter Table table_name RENAME TO new_name

Can we change the data type of a column in a hive table?

Using REPLACE column option

ALTER TABLE table_name REPLACE COLUMNS ……

What is a metastore in Hive?

It is a relational database storing the metadata of hive tables, partitions, Hive databases etc

What is the need for custom Serde?

Depending on the nature of data the user has, the inbuilt SerDe may not satisfy the format of the data. SO users need to write their own java code to satisfy their data format requirements.

Why do we need Hive?

Hive is a tool in Hadoop ecosystem which provides an interface to organize and query data in a databse like fashion and write SQL like queries. It is suitable for accessing and analyzing data in Hadoop using SQL syntax.

What is the default location where hive stores table data?

hdfs://namenode_server/user/hive/warehouse

What are the three different modes in which hive can be run?

- Local mode
- Distributed mode
- Pseudodistributed mode

Is there a date data type in Hive?

Yes. The TIMESTAMP data types stores date in java.sql.timestamp format

What are collection data types in Hive?

There are three collection data types in Hive.

- ARRAY
- MAP

- STRUCT

Can we run unix shell commands from hive? Give example.

Yes, using the ! mark just before the command.

For example !pwd at hive prompt will list the current directory.

What is a Hive variable? What for we use it?

The hive variable is variable created in the Hive environment that can be referenced by Hive scripts. It is used to pass some values to the hive queries when the query starts executing.

Can hive queries be executed from script files? How?

Using the source command.

Example −

Hive> source /path/to/file/file_with_query.hql

What is the importance of .hiverc file?

It is a file containing list of commands needs to run when the hive CLI starts. For example setting the strict mode to be true etc.

What are the default record and field delimiter used for hive text files?

The default record delimiter is − \n

And the filed delimiters are − \001,\002,\003

What do you mean by schema on read?

The schema is validated with the data when reading the data and not enforced when writing data.

How do you list all databases whose name starts with p?

SHOW DATABASES LIKE ‘p.*’

What does the “USE” command in hive do?

With the use command you fix the database on which all the subsequent hive queries will run.

How can you delete the DBPROPERTY in Hive?

There is no way you can delete the DBPROPERTY.

What is the significance of the line

```
set hive.mapred.mode = strict;
```

It sets the mapreduce jobs to strict mode.By which the queries on partitioned tables can not run without a WHERE clause. This prevents very large job running for long time.

How do you check if a particular partition exists?

This can be done with following query

```
SHOW PARTITIONS table_name PARTITION(partitioned_column='partition_value')
```

Which java class handles the Input record encoding into files which store the tables in Hive?

org.apache.hadoop.mapred.TextInputFormat

Which java class handles the output record encoding into files which result from Hive queries?

org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

What is the significance of ‘IF EXISTS” clause while dropping a table?

When we issue the command DROP TABLE IF EXISTS table_name

Hive throws an error if the table being dropped does not exist in the first place.

When you point a partition of a hive table to a new directory, what happens to the data?

The data stays in the old location. It has to be moved manually.

Write a query to insert a new columnnew_colINT into a hiev table $htab$ at a position before an existing column x_col

```
ALTER TABLE table_name
CHANGE COLUMN new_col  INT
BEFORE x_col
```

Does the archiving of Hive tables give any space saving in HDFS?

No. It only reduces the number of files which becomes easier for namenode to manage.

How can you stop a partition form being queried?

By using the ENABLE OFFLINE clause with ALTER TABLE atatement.

While loading data into a hive table using the LOAD DATA clause, how do you specify it is a hdfs file and not a local file ?

By Omitting the LOCAL CLAUSE in the LOAD DATA statement.

If you omit the OVERWRITE clause while creating a hive table,what happens to file which are new and files which already exist?

The new incoming files are just added to the target directory and the existing files are simply overwritten. Other files whose name does not match any of the incoming files will continue to exist.

If you add the OVERWRITE clause then all the existing data in the directory will be deleted before new data is written.

What does the following query do?

```
INSERT OVERWRITE TABLE employees
PARTITION (country, state)
SELECT ..., se.cnty, se.st
FROM staged_employees se;
```

It creates partition on table employees with partition values coming from the columns in the select clause. It is called Dynamic partition insert.

What is a Table generating Function on hive?

A table generating function is a function which takes a single column as argument and expands it to multiple column or rows. Example exploe

How can Hive avoid mapreduce?

If we set the property hive.exec.mode.local.auto to true then hive will avoid mapreduce to fetch query results.

What is the difference between LIKE and RLIKE operators in Hive?

The LIKE operator behaves the same way as the regular SQL operators used in select queries. Example −

street_name like ‘%Chi’

But the RLIKE operator uses more advance regular expressions which are available in java

Example − street_name RLIKE '.*(Chi|Oho).*' which will select any word which has either chi or oho in it.

Is it possible to create Cartesian join between 2 tables, using Hive?

No. As this kind of Join can not be implemented in mapreduce

As part of Optimizing the queries in HIve, what should be the order of table size in a join query?

In a join query the smallest table to be taken in the first position and largest table should be taken in the last position.

What is the usefulness of the DISTRIBUTED BY clause in Hive?

It controls ho wthe map output is reduced among the reducers. It is useful in case of streaming data

How will you convert the string '51.2' to a float value in the price column?

Select cast(price as FLOAT)

What will be the result when you do cast('abc' as INT)?

Hive will return NULL

Can the name of a view be same as the name of a hive table?

No. The name of a view must be unique whne compared to all other tables and views present in the same database.

Can we LOAD data into a view?

No. A view can not be the target of a INSERT or LOAD statement.

What types of costs are associated in creating index on hive tables?

Indexes occupies space and there is a processing cost in arranging the values of the column on which index is cerated.

Give the command to see the indexes on a table.

SHOW INDEX ON table_name

This will list all the indexes created on any of the columns in the table table_name.

What is bucketing ?

The values in a column are hashed into a number of buckets which is defined by user. It is a way to avoid too many partitions or nested partitions while ensuring optimizes query output.

What does /*streamtable($table_name$)*/ do?

It is query hint to stream a table into memory before running the query. It is a query optimization Technique.

Can a partition be archived? What are the advantages and Disadvantages?

Yes. A partition can be archived. Advantage is it decreases the number of files stored in namenode and the archived file can be queried using hive. The disadvantage is it will cause less efficient query and does not offer any space savings.

What is a generic UDF in hive?

It is a UDF which is created using a java program to server some specific need not covered under the existing functions in Hive. It can detect the type of input argument programmatically and provide appropriate response.

The following statement failed to execute. What can be the cause?

```
LOAD DATA LOCAL INPATH '${env:HOME}/country/state/'
OVERWRITE INTO TABLE address;
```

The local inpath should contain a file and not a directory. The $env:HOME is a valid variable available in the hive environment.

How do you specify the table creator name when creating a table in Hive?

The TBLPROPERTIES clause is used to add the creator name while creating a table.

The TBLPROPERTIES is added like −

```
TBLPROPERTIES('creator'= 'Joan')
```

What is Next ?

Further you can go through your past assignments you have done with the subject and make sure you are able to speak confidently on them. If you are fresher then interviewer does not expect you will answer very complex questions, rather you have to make your basics concepts very strong.

Second it really doesn't matter much if you could not answer few questions but it matters that whatever you answered, you must have answered with confidence. So just feel confident during your interview. We at tutorialspoint wish you best luck to have a good interviewer and all the very best for your future endeavor. Cheers :-)