**Top 15 Hive Interview Questions**

**1. What is Hive?**

Hive is a data warehousing tool. It is an abstraction and it gives SQL queries to perform an analysis. It gives you logical abstraction over the databases and the tables but it is not a database.

**2. What is Hive a metastore?**

Hive contains two things: data and the metadata. The metadata contains the (column names, partitions information, bucketing information, SerDe etc.) i.e., the data about the actual table this is by default stored in the Derby database, we can also configure it to Oracle or MySQL database.

**3. What is the limitation of Derby database for Hive metastore?**

With derby database, you cannot have multiple connections or multiple sessions instantiated at the same time. Derby database runs in the local mode and it creates a log file so that multiple users cannot access Hive simultaneously.

**4. What are managed and external tables?**

We have got two things, one of which is data present in the HDFS and the other is the metadata, present in some database.

There are two categories of Hive tables i.e., Managed and External Tables.

In the Managed tables, both the data and the metadata are managed by Hive and if you drop the managed table, both data and metadata are deleted.

There are some situations where your data will be controlled by some other application and you want to read that data but you must allow Hive to delete that data.

In such case, you can create an external table in Hive. In the external table, metadata is controlled by Hive but the actual data will be controlled by some other application. So, when you delete a table accidentally, only the metadata will be lost and the actual data will reside wherever it is.

**5. What are the complex data types in Hive?**

* Map – The**Map**contains a key-value pair where you can search for a value using the key.
* Struct – A**Struct**is a collection of elements of different data types. For example, if you take the address, it can have different data types. For example, pin code will be in Integer format.
* Array – An**Array**will have a collection of homogeneous elements. For example, if you take your skillset, you can have N number of skills
* Uniontype – It represents a column which can have a value that can belong to any of the data types of your choice.

**6. How does partitioning help in the faster execution of queries?**

With the help of partitioning, a subdirectory will be created with the name of the partitioned column and when you perform a query using the WHERE clause, only the particular sub-directory will be scanned instead of scanning the whole table. This gives you faster execution of queries.

**7. How to enable dynamic partitioning in Hive?**

Related to partitioning there are two types of partitioning Static and Dynamic. In the static partitioning, you will specify the partition column while loading the data.

Whereas in dynamic partitioning, you push the data into Hive and then Hive decides which value should go into which partition. To enable dynamic partitioning, you have set the below property  
set hive.exec.dynamic.parition.mode = nonstrict;

Example,  
*insert overwrite table emp\_details\_partitioned*  
*partition(location)*  
*select \* from emp\_details;*

**8. How does bucketing help in the faster execution of queries?**

If you have to join two large tables, you can go for reduce side join. But if both the tables have the same number of buckets or same multiples of buckets and also sorted on the same column there is a possibility of SMBMJ in which all the joins take place in the map phase itself by matching the corresponding buckets.

Buckets are basically files that are created inside the HDFS directory.

There are different properties which you need to set for bucket map joins and they are as follows:

* set hive.enforce.sortmergebucketmapjoin = false;
* set hive.auto.convert.sortmerge.join =  false;
* set hive.optimize.bucketmapjoin =  true;
* set hive.optimize.bucketmapjoin.sortedmerge = true;

**9. How to enable bucketing in Hive?**

By default bucketing is disabled in Hive, you can enforce to enable it by setting the below property  
set hive.enforce.bucketing  = true;

**10. Which method has to be overridden when we use custom UDF in Hive?**

Whenever you write a custom UDF in Hive, you have to extend the UDF class and you have to override the **evaluate()**function.

**11. What are the different file formats in Hive?**

There are different file formats supported by Hive

* Text File format
* Sequence File format
* RC file format
* Parquet
* Avro
* ORC

Every file format has its own characteristics and Hive allows you to choose easily the file format which you wanted to use.

**12. How is SerDe different from File format in Hive?**

SerDe stands for Serializer and Deserializer. It determines how to encode and decode the field values or the column values from a record that is: how you serialize and deserialize the values of a column

But file format determines how records are stored in key value format or how do you retrieve the records from the table.

**13. What is RegexSerDe?**

Regex stands for a regular expression. Whenever you want to have a kind of pattern matching, based on the pattern matching, you have to store the fields. RegexSerDe is present in **org.apache.hadoop.hive.contrib.serde2.RegexSerDe.**

In the SerDeproperties, you have to define your input pattern and output fields. For example, you have to get the column values from line **xyz/pq@def** if you want to take xyz, pq and def separately.

To extract the pattern, you can use:  
‘input.regex’ = ‘(.\*)/(.\*)@(.\*)’  
To specify how to store them, you can use  
‘output.format.string’ = ‘%1$s%2$s%3$s’;

**14. How is ORC file format optimised for data storage and analysis?**

ORC stores collections of rows in one file and within the collection the row data will be stored in a columnar format. With columnar format, it is very easy to compress, thus reducing a lot of storage cost.

While querying also, it queries the particular column instead of querying the whole row as the records are stored in columnar format.

ORC has got indexing on every block based on the statistics min, max, sum, count on columns so when you query, it will skip the blocks based on the indexing.

**15. How to access HBase tables from Hive?**

Using Hive-HBase storage handler, you can access the HBase tables from Hive and once you are connected, you can query HBase using the SQL queries from Hive. You can also join multiple tables in HBase from Hive and retrieve the result.