Hive scenario-based Interview Questions

1. **Suppose, I have a CSV file – 'sample.csv' present in '/temp' directory with the following entries: id first_name last_name email gender ip_address**

   1 Hugh Jackman hughjackman@cam.ac.uk Male 136.90.241.52
   2 David Lawrence dlawrence1@gmail.com Male 101.177.15.130
   3 Andy Hall andyhall2@yahoo.com Female 114.123.153.64
   4 Samuel Jackson samjackson231@sun.com Male 89.60.227.31
   5 Emily Rose rose.emily4@surveymonkey.com Female 119.92.21.19

2. **How will you consume this CSV file into the Hive warehouse using built SerDe?**

SerDe stands for serializer/deserializer. A SerDe allows us to convert the unstructured bytes into a record that we can process using Hive. SerDes are implemented using Java. Hive comes with several built-in SerDes and many other third-party SerDes are also available.

Hive provides a specific SerDe for working with CSV files. We can use this SerDe for the sample.csv by issuing following commands:

```
CREATE EXTERNAL TABLE sample

(id int, first_name string,

last_name string, email string,

gender string, ip_address string)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

STORED AS TEXTFILE LOCATION '/temp';
```

Now, we can perform any query on the table 'sample':

```
SELECT first_name FROM sample WHERE gender = 'male';
```

3. **If we have hive table with 8 cloumns ,but data has only 10 columns,what data will load to hive table?any error?**

   Yes, data will load without any errors, hive load only 8 columns data ,it will ignore remaining to cloumns data

4. **If we have hive table with 10 cloumns ,but data has only 9 columns,what data will load to 10^{th} column in hive table?any error?**

   Yes, data will load without any errors, but 10^{th} column in table is set to defult null value.

5. **Why NameNode goes to safe mode?**

   When namenode is started or restarted, namenode will be in safemode for a period of time. At this time you will not be able to use your Hadoop cluster fully. Write operations to HDFS will fail and because of that your MapReduce jobs will also fail.

   When namenode is in safemode, Merges edits log file on fsimage and results in new file system namespace. It will wait for all the blocks for all the files in HDFS to be available. More importantly, it will wait for the minimum replication for all blocks to be met and available.

6. **Sparklens?**

   Sparklens is an open source Spark profiling tool from Qubole.

   it identifying the potential opportunities for optimizations with respect to driver side computations, lack of parallelism, skew, etc.

   The built-in scheduler simulator can predict how a given spark application will run on any number of executors in a single run.

   Sparklens provides the following information:

   - If the application can run faster with more cores and how to optimize it.
   - If the compute cost can be saved by running the application with less cores and without much increase in wall clock time.
   - The absolute minimum time that the application can take if infinite executors are given.
   - How to run the application below the absolute minimum time.

7. **Let's say a Hive table is created as an external table. If we drop the table, will the data be accessible?**

   The data will be accessible even if the table gets dropped. We can get the data from the table's HDFS location.

8. **A Hive table is created as an external table at location say hdfs://usr/data/table_name. If we dump a data set which are having the data as per the table structure, will we able to fetch the records from the table using a select query?**

   Yes, we will be able to fetch the records from the table after dumping the data set at the hive table external location.

9. **A Hive partition table is created which is partition by a column say yearofexperience. If we create a directory say yearofexperience=3 at the HDFS path of the table and dump the data set which is as per the table structure. Will the data be available if we execute select query on the table?**

   No, the data will not accessible by executing the select query on the table. After dumping the data files at table HDFS location for the partition, you will have to update the metadata using below command:

10. **Let's take the same previous Hive partition table. If we drop the partition, will we able to access the data?**

    If a hive partition created as a managed table, then after dropping the partition, data will also get removed from the path. But in case of an external table, data will be accessible from the same external path of the hive partition table.

11. **Let's take the same previous Hive partition table partitioned by a column named yearofexperience. It is having multiple partitions at the HDFS location. If we drop a partition directory say yearofexperience=3 from the HDFS location, will this partition be listed while querying show partitions on the table?**

If we drop the partition directory say hive/warehouse/bdp.db/partitioned_test_external/yearofexperience=3 from the HDFS location, it will be listed if you query show partitions on the table.

12. **Suppose we have created a Hive partition table which is partitioned by a column named city. We are getting data which are having Empty/Null value for the partition column(city) and have to load these data into the hive table with dynamic partition as it is having multiple city records in the data set. In which partition the records, with an empty value for city column, will be available?**

    While loading the data into a table using dynamic partition if any null or empty value comes for a defined partition column, then it uses to create a default partition named __HIVE_DEFAULT_PARTITION__ at HDFS location and dump those records in that partition.

13. **Let's say we have created a Hive partition table. This table gets updated every day with a huge volume of data. As we already know that the table is having a high volume of data, we want to restrict the query not to do a full scan on the table. How will you achieve this?**

    You can achieve this by setting below properties:

    SET hive.mapred.mode=strict;

    Hive Strict Mode (hive.mapred.mode=strict) enables hive to restrict certain performance intensive operations. Such as – It restricts queries of partitioned tables without a WHERE clause

14. **If we create a table with an EXTERNAL keyword, but not mentioning any location in the create table statement, which kind of table it will be – managed or external?**
    The created table will behave like an external table that means if you drop the table, data will be available at the table HDFS location.

15. **We have created a view on top of a Hive table. If we drop the Hive table, will the View be accessible?**
    The view will not be accessible. It will throw an error like Table not found.

16. **Let's say you want to create a table which is having columns name like Hive keyword (say, timestamp, date, etc.) or column name having space(say "col 50"). How will you create the table in Hive?**
    You can mention the column name enclosed by backticks (`).
    For example- `timestamp` string, `col 50` string

17. **Suppose, I have a lot of small CSV files present in /input directory in HDFS and I want to create a single Hive table corresponding to these files. The data in these files are in the format: {id, name, e-mail, country}. Now, as we know, Hadoop performance degrades when we use lots of small files.**
    So, how will you solve this problem where we want to create a single Hive table for lots of small files without degrading the performance of the system?

One can use the sequencefile format which will group these small files together to form a single sequence file. The steps that will be followed in doing so are as follows:

Create a temporary table:

CREATE TABLE temp_table (id INT, name STRING, e-mail STRING, country STRING)

ROW FORMAT FIELDS DELIMITED TERMINATED BY ',' STORED AS TEXTFILE;

Load the data into temp_table:

LOAD DATA INPATH '/input' INTO TABLE temp_table;

Create a table that will store data in sequencefile format:

CREATE TABLE sample_seqfile (id INT, name STRING, e-mail STRING, country STRING)

ROW FORMAT FIELDS DELIMITED TERMINATED BY ',' STORED AS SEQUENCEFILE;

Transfer the data from the temporary table into the sample_seqfile table:

INSERT OVERWRITE TABLE sample SELECT * FROM temp_table;

Hence, a single sequencefile is generated which contains the data present in all of the input files and therefore, the problem of having lots of small files is finally eliminated.

18. **Scenario: Suppose there are several small CSV files present in /user/input directory in HDFS and you want to create a single Hive table from these files. The data in these files have the following fields: {registration_no, name, email, address}. What will be your approach to solve this, and where will you create a single Hive table for multiple smaller files without degrading the performance of the system?**
    Using sequencefile format and grouping these small files together to form a single sequence file can solve this problem. Below are the steps:

```
> read.table(file = "data.table", header = TRUE)
   name age gender
1  john  23   male
2  mary  21 female
3 jacob  18   male
4 nancy  25 female
```

19. **Scenario: I am inserting data into a table based on partitions dynamically. But, I received an error – FAILED ERROR IN SEMANTIC ANALYSIS: Dynamic partition strict mode requires at least one static partition column. How will you remove this error?**

To remove this error one has to execute following commands:

SET hive.exec.dynamic.partition = true;

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

Things to Remember:

- By default, hive.exec.dynamic.partition configuration property is set to False in case you are using Hive whose version is prior to 0.9.0.

- hive.exec.dynamic.partition.mode is set to strict by default. Only in non – strict mode Hive allows all partitions to be dynamic.

20. **Scenario:Suppose, I create a table that contains details of all the transactions done by the customers of year 2016: CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;Now, after inserting 50,000 tuples in this table, I want to know the total revenue generated for**

each month. But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so?

We can solve this problem of query latency by partitioning the table according to each month. So, for each month we will be scanning only the partitioned data instead of whole data sets.

As we know, we can't partition an existing non-partitioned table directly. So, we will be taking following steps to solve the very problem:

Create a partitioned table, say partitioned_transaction:

CREATE TABLE partitioned_transaction (cust_id INT, amount FLOAT, country STRING) PARTITIONED BY (month STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ;

Enable dynamic partitioning in Hive:

SET hive.exec.dynamic.partition = true;

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

Transfer the data from the non – partitioned table into the newly created partitioned table:

INSERT OVERWRITE TABLE partitioned_transaction PARTITION (month) SELECT cust_id, amount, country, month FROM transaction_details;

Now, we can perform the query using each partition and therefore, decrease the query time.

21. **Scenario: Suppose I have installed Apache Hive on top of my Hadoop cluster using default metastore configuration. Then, what will happen if we have multiple clients trying to access Hive at the same time?**

The default metastore configuration allows only one Hive session to be opened at a time for accessing the metastore. Therefore, if multiple clients try to access the metastore at the same time, they will get an error. One has to use a standalone metastore, i.e. Local or remote metastore configuration in Apache Hive for allowing access to multiple clients concurrently.

Following are the steps to configure mysql database as the local metastore in Apache Hive:

One should make the following changes in hive-site.xml:

Javax.jdo.option.connectionurl property should be set to jdbc:mysql://host/dbname?Createdataba

Seifnotexist=true.

Javax.jdo.option.connectiondrivername property should be set to com.mysql.jdbc.Driver.

One should also set the username and password as:

Javax.jdo.option.connectionusername is set to desired username.

Javax.jdo.option.connectionpassword is set to the desired password.

The JDBC driver JAR file for mysql must be on the Hive's classpath, i.e. The jar file should be copied into the Hive's lib directory.

Now, after restarting the Hive shell, it will automatically connect to the mysql database which is running as a standalone metastore.

22. Is it possible to add 100 nodes when we already have 100 nodes in Hive? If yes, how?

Yes, we can add the nodes by following the below steps:

Step 1: Take a new system; create a new username and password

Step 2: Install SSH and with the master node setup SSH connections

Step 3: Add ssh public_rsa id key to the authorized keys file

Step 4: Add the new datanode hostname, IP address, and other details in /etc/hosts slaves file:

192.168.1.102 slave3.in slave3

Step 5: Start the datanode on a new node

Step 6: Login to the new node like suhadoop or: ssh -X hadoop@192.168.1.103

Step 7: Start HDFS of the newly added slave node by using the following command:./bin/hadoop-daemon.sh start data node

Step 8: Check the output of the jps command on the new node

23. **Define the difference between Hive and hbase?**

Hbase	Hive
1. Hbase is nosql database	1. It is a data warehousing infrastructure
2. Hbase is built on the top of HDFS	2. Hive queries are executed as mapreduce jobs internally
3. Provides low latency to single rows from huge datasets	3. Provides high latency for huge datasets
4. Provides random access to data	4. Provides random access to data
Hbase does not allow SQL queries	Hive enables most of the SQL queries
	Hive does not support record level insert, update, and delete operations on table

24. Difference between Hive and Impala?

• **Hive**	• **Impala**
Basically, in Hive every query has the common problem of a "cold start".	Impala avoids any possible startup overheads, being a native query language. However, that are very frequently and commonly observed in mapreduce based jobs. Moreover, to process a query always Impala daemon processes are started at the boot time itself, making it ready.`
Basically, Hive materializes all intermediate results. Hence, it enables enabling better scalability and fault tolerance. However, that has an adverse effect on slowing down the data processing.	However, it's streaming intermediate results between executors. Although, that trades off scalability as such.
At Compile time, Hive generates query expressions.	During the Runtime, Impala generates code for "big loops".

25. **Differentiate between Pig and Hive?**

Apache Pig	Apache Hive
Uses a high-level procedural (data flow language) language called Pig Latin for programming	It uses a declarative language, called hiveql,
Used for programming	Used for report creation
Researchers and programmers	Mainly Data Analysts
Operates on the client-side of the cluster and allows both structured and semi-structured data	Operates on the server-side of the cluster and allows structured data.
Not as fast as hiveql	Faster with in-built features
Always defined in the script itself	Stored in the local database
Takes little extra time and effort to master	Easy to learn from database experts
Supports Avro file format by default.	It does not support the Avro file format by default. This can be done using **"Org.Apache.Hadoop.Hive.serde 2.Avro"**
Yahoo developed it, and it does not support partition	Facebook developed it and it supports partition

26. Difference between SQL and hiveql?

1. **SQL**	2. HQL
It is based on a relational database model.	It is a combination of object-oriented programming with relational database concepts.
It manipulates data stored in tables and modifies its rows and columns.	It is concerned about objects and its properties.
It consider the relationship that exists between two tables	It considers the relation between two objects.

27. **Hive vs Spark SQL**

1. Apache Hive	2. Spark SQL
Basically, the hive was first released in the year 2012.	Whereas, Spark SQL was first released in the year 2014.
Currently released on 18 November 2017: version 2.3.2	Currently released on 09 October 2017: version 2.1.2
Although, Facebook developed it originally. Further donated to the Apache Software Foundation, that has maintained it since.	However, Apache Software Foundation developed it originally.

28. WHAT IS UDF IN HIVE?

UDF is a user-defined function created with a Java program to address a specific function that is not part of the existing Hive functions.

29. How to Write a UDF function in Hive?

Create a Java class for the User Defined Function which extends ora.apache.hadoop.hive.sq.exec.UDF and implements more than one evaluate() methods. Put in your desired logic and you are almost there.

Package your Java class into a JAR file

Go to Hive CLI, add your JAR, and verify your jars is in the Hive CLI classpath

CREATE TEMPORARY FUNCTION in Hive which points to your Java class

30. How do you write your own serde?

However, following are the ways:

Despite serde users want to write a deserializer in most cases. It is because users just want to read their own data format instead of writing to it.

By using the configuration parameter ‘regex’, the regexdeserializer will deserialize the data, and possibly a list of column names (see serde2.metadatatypedcolumnsetserde).

31. Unable to instantiate org.apache.hadoop.hive.metastore.hivemetastoreclient

Ans. There is a possibility that because of following reasons above error may occur:

1. While we use derby metastore, Then lock file would be there in case of the abnormal exit.

Hence, do remove the lock file

Rm metastore_db/*.lck

1. Moreover, Run hive in Debug mode

Hive -hiveconf hive.root.logger=DEBUG,console

32.

33. What is the output of regexp_replace("pqrser", "qr|er", "") ?
Ps

34. Which of the following function will remove duplicates?
COLLECT_SET()

35. Which of the following is NOT a window function?
SPLIT()

36. What will be the output of CONCAT_WS('|','hey','coder','how','are','you')
hey|coder|how|are|you

37. What is the extension of hive query file?
.hql

38. How would you delete the data of hive table without deleting the table?
truncate <table_name>;

39. Which of the following function will return the size of string?
length()

40. You have one column in hive table named as "my_ts" having datatype as string and sample value like "2018-02-24 17:22:35". how would you extract only day from it i.e. 24 ?
day(myts)

41. Which of the following will cast a column “a” having value 3.2 to 3 ?

CAST(a as INT)

42. Which of the following method will remove the spaces from both the ends of ” bigdata ” ?

trim()