***HIVE INTERVIEW QUESTION***

**1. What is the definition of Hive? What is the present version of Hive?**

Hive is a data warehouse service for hadoop that runs on SQL like queries called HQL(Hive Query Language).Hive is build on top of hadoop.

4.0.0-alpha-2 is the latest and present version of hive.

**2. 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 the OLTP system.

**3. How is HIVE different from RDBMS? Does hive support ACID**

**transactions? If not then give the proper reason.**

***HIVE***

1) Hive is a data warehouse service for Hadoop that runs on SQL-like queries called HQL(Hive Query Language).

2)Hive is built on top of Hadoop. This HQL query gets converted to map-reduce job.

3)Hive can store data in normalised and de-normalised forms.

4)Hive supports various file formats like CSV , json , XML , parquet , ORC etc.

***RDBMS***

1)RDBMS stands for Relational Database Management System. RDBMS is a such type of database management system which is specifically designed for relational databases.

2) A relational database refers to a database that stores data in a structured format using rows and columns and that structured form is known as a table.

Hive supports all ACID properties which enable us to use transactions, create transactional tables, and run queries like Insert, Update, and Delete on tables.

**4. Explain the hive architecture and the different components of a Hive**

**Architecture?**

User Interface: It helps the user to send queries to the Hive system and other operations. The user interface provides hive Web UI, Hive Command-Line and Hive HDInsight.

Driver: It designs a session handle for the query, and then the queries are sent to the compiler for the execution plan.

Metastore: It contains organized data and information on various warehouse tables and partitions.

Compiler: It creates the execution plan for the queries, performs semantic analysis on different query blocks, and generates query expressions.

Execution Engine: It implements the execution plans created by the compiler

**5. Mention what the Hive query processor does? And Mention what are the**

**components of a Hive query processor?**

Hive query processor implements the processing framework for converting SQL to a graph of map/reduce jobs and the execution time framework to run those jobs in the order of dependencies.

The components of a Hive query processor include,

Logical Plan Generation

Physical Plan Generation

Execution Engine

Operators

UDFs and UDAF’s

Optimizer

Parser

Semantic Analyzer

Type Checking

**6. What are the three different modes in which we can operate Hive?**

**Local mode:** In Hive local mode, Map Reduce jobs related to Hive run locally on a user machine. This is the default mode in which Hadoop uses local file system.

**Distributed Mode**: In this mode, Hive as well as Hadoop is running in a fully distributed mode. NameNode, DataNode, JobTracker, TaskTracker etc run on different machines in this mode.

**Pseudo-distributed Mode:** This is the mode used by developers to test the code before deploying to production. In this mode, all the daemons run on same virtual machine. With this mode, we can quickly write scripts and test on limited data sets.

**7. Features and Limitations of Hive.**

***Features* :-**

1) Hive uses HQL which is just like SQL.HQL is easy to write.

2) Hive is capable of analysing large datasets stored in hdfs

3) Hive supports ETL.

4) Hive supports users to access files from hdfs , Hbase etc.

5) Hive is fast and scalable.

***Limitation :-***

1) Hive doesn’t support online transaction processing

2) Subqueries are not supported in hive.

3) The latency in the apache hive query is very high.

**8. How to create a Database in HIVE?**

We can create database in hive by using following command

Create database <database \_name> ;

**9. How to create a table in HIVE?**

We can create table in hive using the following command

Create table <table\_name> ;

**10.What do you mean by describe and describe extended and describe**

**formatted with respect to database and table**

To see table primary info of Hive table, use describe table\_name; command

To see more detailed information about the table, use describe extended table\_name; command

To see code in a clean manner use describe formatted table\_name; command to see all information. also describe all details in a clean manner.

**11.How to skip header rows from a table in Hive?**

We can skip header rows by using the following command

TBLPROPERTIES("skip.header.line.count"="2”);

For example

CREATE EXTERNAL TABLE employee (

name STRING,

job STRING,

dob STRING,

id INT,

salary INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘ ‘

STORED AS TEXTFILE

LOCATION ‘/user/data’

TBLPROPERTIES("skip.header.line.count"="2”);

Here first two lines wont be selected rest will reflect in the table.

**12.What is a hive operator? What are the different types of hive operators?**

The HiveQL operators facilitate to perform various arithmetic operations such as addition, substraction, multiplication and division etc

HiveQL also supports relational operationssuch as is equal to , not equal to , less than , greater than etc . The relational operators are generally used with clauses like Join and Having to compare the existing records.

**13.Explain about the Hive Built-In Functions**

So, to perform several operations there are some functions built for a specific purpose. Such as Mathematical, arithmetic, logical and relational on the operands of table column names.

There are some built-in functions present in hive they are as follows :-

Hive Date Functions,

Hive Mathematical Functions,

Hive Conditional Functions,

Hive String Functions .

**14. Write hive DDL and DML commands.**

***Hive DDL commands***

**1) To create table**

CREATE TABLE IF NOT EXISTS emp.employee (

id int,

name string,

age int,

gender string )

COMMENT 'Employee Table'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ',' ;

**2) TO get the list of tables**

SHOW TABLES ;

**3) to get the details about tables**

Describe < Table\_name > ;

**4) to delete the content and structure of table**

TRUNCATE TABLE emp;

**5) Rename a Table**

ALTER TABLE < Old\_table\_name > RENAME TO < New\_table\_name > ;

**6) Drop table from hive database**

Drop Table < Table\_name >

**Hive DML commands**

**1) Loading of data**

Load data local inpath ‘file:/// < complete\_file\_path > ’ into table < table\_name >

**2) For retrieving data**

SELECT col1,col2 FROM < table\_name > ;

**3) To insert data**

INSERT INTO TABLE tablename1 select \* FROM table\_name ;

**15.Explain about SORT BY, ORDER BY, DISTRIBUTE BY and**

**CLUSTER BY in Hive.**

**SORT BY :**

The SORT by clause sorts the data per reducer. As a result, if we have N number of reducers, we will have N number of sorted files in the output. These files can have overlapping data ranges.

**ORDER BY :**

ORDER BY clause orders the data globally. Because it ensures the global ordering of the data, all the data need to be passed from a single reducer only. As a result, the order by clause outputs one single file only

**DISTRIBUTE BY :**

DISTRIBUTE BY clause is used to distribute the input rows among reducers. It ensures that all rows for the same key columns are going to the same reducer.

**CLUSTER BY :**

CLUSTER BY clause is a combination of DISTRIBUTE BY and SORT BY clauses together. That means the output of the CLUSTER BY clause is equivalent to the output of DISTRIBUTE BY + SORT BY clauses. The CLUSTER BY clause distributes the data based on the key column and then sorts the output data by putting the same key column values adjacent to each other. So, the output of the CLUSTER BY clause is sorted at the reducer .

**16.Difference between "Internal Table" and "External Table" and Mention**

**when to choose “Internal Table” and “External Table” in Hive?**

***Internal table :-***

It is the default table in Hive. When the user creates a table in Hive without specifying it as external, then by default, an internal table gets created in a specific location in HDFS.

***External table :-***

Hive does not manage the data of the External table.

We create an external table for external use as when we want to use the data outside the Hive.

External tables are stored outside the warehouse directory. They can access data stored in sources such as remote HDFS locations or Azure Storage Volumes.

We can create the external table by specifying the EXTERNAL keyword in the Hive create table statement.

We can use the internal table in cases:

When we want a table for a temporary purpose.

When required that Hive should manage the lifecycle of the table.

And when we don’t want table data after deletion.

We can use the external table in cases:

When we don’t want to delete the table data completely even after DROP.

If we feel the data should not be own by Hive

When required to use data outside of Hive. For example, the data files are read and processed by an existing program that does not lock the files.

**17.Where does the data of a Hive table get stored?**

Hive stores its database and table metadata in a metastore, which is a database.

**18.Is it possible to change the default location of a managed table?**

Yes it is possible to change the location of managed table by using the clause – **LOCATION** **‘<hdfs\_path>’** we can change the default location of a managed table.

**19.What is a metastore in Hive? What is the default database provided by**

**Apache Hive for metastore?**

The Hive metastore is simply a relational database. It stores metadata related to the tables/schemas you create to easily query big data stored in HDFS. When you create a new Hive table, the information related to the schema (column names, data types) is stored in the Hive metastore relational database. Other information like input/output formats, partitions, HDFS locations are all stored in the metastore.

Derby Database is the default database in hive

**20.Why does Hive not store metadata information in HDFS?**

Hive stores metadata information in the metastore using RDBMS instead of HDFS. The reason for choosing RDBMS is to achieve low latency as HDFS read/write operations are time consuming processes.

**21. What is a partition in Hive? And Why do we perform partitioning in**

**Hive?**

Partitioning is a way of dividing a table into related parts based on the values of a particular column. If we perform partitioning on a table it will divide the table on particular column values by doing this we will be able to reduce the query time to a great extent and we will be able to analyse the extremely large dataset in relatively less time.

**22. What is the difference between dynamic partitioning and static**

**Partitioning?**

In Static partitioning value of a partition column will be known to us we load the data into a specific partition. static partition takes less time to load the data.

In dynamic partitioning, the values of partitioned columns exist within the table. So, it is not required to pass the values of partitioned columns manually. Dynamic partition takes more time to load the data.

**23. How do you check if a particular partition exists?**

We can check if a particular partition exists with the help of the following command

SHOW PARTITIONS [db\_name.]table\_name

[PARTITION(partition\_spec)]

**24. How can you stop a partition from being queried?**

By using the ENABLE OFFLINE clause with ALTER TABLE statement.

**25. Why do we need buckets? How Hive distributes the rows into buckets?**

Decomposing the dataset into more manageable parts is known as bucketing.

It is kind of similar to partitioning but added advantage is that we can divide the datasets into more manageable parts known as buckets.

So we can use bucketing in the hive when the implementation of partitioning becomes difficult. we can perform queries more efficiently as compared to partitioning.

The concept of bucketing is based on the hashing technique. Based on the result of the hash function the data is stored in the corresponding bucket.

**26.In Hive, how can you enable buckets?**

By using this command below bucket can be enabled;

set.hive.enforce.bucketing=true;

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

Decomposing the dataset into more manageable parts is known as bucketing.

It is kind of similar to partitioning but added advantage is that we can divide the datasets into more manageable parts known as buckets. partitioned data can be bucketed to separate the data to perform queries more efficiently.

**28. How to optimise Hive Performance? Explain in very detail.**

Optimising the performance means reducing the query time to perform the analytical operation in a speedy manner to gain insights from the datasets. Partitioning and Bucketing are the two ways we can optimise the data also we can use filter conditions as an edge while querying the data.

**29. What is the use of Hcatalog?**

*NOT ABLE TO SOLVE*

**30. Explain the different types of join in Hive.**

Inner join :- with the help of inner join we can retrieve the common data from two tables or multiple table.

Left Outer Join :- even if there are no matches in the right table it returns all the rows from the left table.

Right Outer Join :- Right Outer Join returns all the rows from the right table.

Full Outer Join :- Returns all the records from all of the tables involved.

**31. Is it possible to create a Cartesian join between 2 tables, using Hive?**

*Not able to solve*

**32. Explain the SMB Join in Hive?**

SMB is a join performed on bucket tables that have the same sorted, bucket, and join condition columns. It reads data from both bucket tables and performs common joins (map and reduce triggered) on the bucket tables.

**33. What is the difference between order by and sort by which one we should**

**Use?**

Hive supports SORT BY which sorts the data per reducer. The difference between "order by" and "sort by" is that the former guarantees total order in the output while the latter only guarantees ordering of the rows within a reducer. If there is more than one reducer, "sort by" may give partially ordered final results.

ORDER BY works on a single reducer and it causes a performance bottleneck. But, SORT BY orders the data only within each reducer and performs a local ordering where each reducer’s output will be sorted ensuring better performance.

**34.What is the usefulness of the DISTRIBUTED BY clause in Hive?**

*Not able to solve*

**35. How does data transfer happen from HDFS to Hive?**

To query data in HDFS in Hive, you apply a schema to the data and then store data in ORC format.

**36.Wherever (Different Directory) I run the hive query, it creates a new**

**metastore\_db, please explain the reason for it?**

*Not able to solve*

**37. What will happen in case you have not issued the command: ‘SET**

**hive.enforce.bucketing=true;’ before bucketing a table in Hive?**

If we dont issue the command Number of files that will be generated in the table directory will not be equal to the number of buckets.

**38.Can a table be renamed in Hive?**

we can use ALTER TABLE to rename the name of the table with help of following command

ALTER TABLE table\_name

RENAME TO new\_table\_name;

**39.Write a query to insert a new column(new\_col INT) into a hive table at a**

**position before an existing column (x\_col)**

*Not able to solve*

**40. What is serde operation in HIVE?**

Serialization : It is the processes of converting the data into byte that can be stored in HDFS or it can be sent over network.

Deserialization : It is the process of converting the bytes into the data. Reverse of the serialization.

**41. Explain how Hive Deserializes and serialises the data?**

Hive uses the SerDe interface for IO. The interface handles both serialization and deserialization and also interpreting the results of serialization as individual fields for processing.

A SerDe allows Hive to read data from a table, and write it back out to HDFS in any custom format. Avro, ORC, RegEx, CSV Parquet, jsonSerDe etc are built in serde present in hive

**42.Write the name of the built-in serde in the hive.**

A SerDe allows Hive to read in data from a table, and write it back out to HDFS in any custom format.

Avro, ORC, RegEx, CSV Parquet, jsonSerDe etc are built in serde present in hive.

**43. What is the need of custom Serde?**

**44. Can you write the name of a complex data type(collection data types) in**

**Hive?**

*Not able to solve*

**45. Can hive queries be executed from script files? How?**

*Not able to solve*

**46. What is 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

**47. How do you list all databases in Hive whose name starts with s?**

show databases like ' s\* ' ;

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

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

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

**49. How to change the column data type in Hive?**

ALTER TABLE table\_name CHANGE column\_name column\_name new\_datatype ;

**50. How will you convert the string ’51.2’ to a float value in the particular**

**Column?**

SELECT CAST (‘ 51.2 ’ AS float);

**51. What will be the result when you cast ‘abc’ (string) as INT?**

Hive will return NULL as an output

**52. What does the following query do?**

**a. INSERT OVERWRITE TABLE employees**

**b. PARTITION (country, state)**

**c. SELECT ..., se.cnty, se.st**

**d. FROM staged\_employees se;**

**53. Write a query where you can overwrite data in a new table from the**

**existing table.**

**FROM existing\_table INSERT OVERWRITE TABLE New\_table select\* ;**

**54. What is the maximum size of a string data type supported by Hive?**

**Explain how Hive supports binary formats.**

Maximum size of a string data type supported by Hive is 2 GB. Hive supports the text file format by default, and it also supports the binary format sequence files, ORC files, Avro data files, and Parquet files to read or support this kind of file system hive have serde libraries.

**55. What File Formats and Applications Does Hive Support?**

There are some specific file formats which Hive can handle such as:

TEXTFILE, SEQUENCEFILE, RCFILE, ORCFILE

**56. How do ORC format tables help Hive to enhance its performance?**

ORC standas for optimized row columnar format. When we load data into ORC file format it gets serialised means it gets converted to bytes, this bytes acquire less space than the plain text file or csv file hence hive engine will have to scan less amount of memory to fetch the data.

Column based file format have quite low seek time as compared to row based file format.

**57. How can Hive avoid mapreduce while processing the query?**

*Not able to solve*

**58. What are view and indexing in hive?**

Hive indexing is a query optimization technique to reduce the time needed to access a column or a set of columns within a Hive database

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

No, the name of the view must be unique .

**60. What types of costs are associated with creating indexes on hive tables?**

*Not able to solve*

**61. Give the command to see the indexes on a table.**

SHOW INDEX ON table\_name

**62. Explain the process to access subdirectories recursively in Hive queries.**

We can access subdirectories recursively by using the following command:

Set mapred.input.dir.recursive=true;

Set hive.mapred.supports.subdirectories=true;

**63. If you run a select \* query in Hive, why doesn't it run MapReduce?**

Because Select \* query dont involve the arithmetic , logical or relational operation it just fetch the data from the database and reflects as an output.

**64. What are the uses of Hive Explode?**

When we want to give array as an input and convert it into a separate table row hence we need to convert complicated data types into desired table formats, to do this we need to use explode.

**65. What is the available mechanism for connecting applications when we**

**run Hive as a server?**

**66. Can the default location of a managed table be changed in Hive?**

Yes we can do it by using the following clause

LOCATION ‘<hdfs\_path>’

we can change the default location of a managed table.

**67. What is the Hive ObjectInspector function?**

*Not able to solve*

**68. What is UDF in Hive?**

User Defined Functions (UDFs) in hive are used to plug in our own logic in terms of code into hive when we are not able to get the desired result from hive's built in functions. We can invoke the UDFs from hive query.

**69. Write a query to extract data from hdfs to hive.**

**Create a folder on HDFS**

hadoop fs -mkdir hive

**Move the text file from local file system into newly created folder**

hadoop fs -put ~/Desktop/student.txt hive/

**Create Empty table STUDENT in HIVE**

create table student

( std\_id int,

std\_name string,

std\_grade string,

std\_addres string)

partitioned by (country string)

row format delimited

fields terminated by ',' ;

**Load Data from HDFS path into HIVE TABLE.**

load data inpath 'hive/student.txt' into table student

**70. What are TextInputFormat and SequenceFileInputFormat in hive.**

*Not able to solve*

**71. How can you prevent a large job from running for a long time in a hive?**

We can increase the number of reducers to execute and complete a large task in less time but by doing so we are using the resources of the system because of that others will not be able to able to use the system properly or it will crate burden on the system.

Example : Set mapreduce.job.reduces = 10

More the number of reducers we have faster will be an operation.

**72. When do we use explode in Hive?**

We use explode when we want to return a row-set with a single column (col), one row for each element from the array.

**73. Can Hive process any type of data format? Why? Explain in very detail**

Hive can process the following type of file format

Hive supports four file formats those are TEXTFILE, SEQUENCEFILE, ORC and RCFILE (Record Columnar File).

Hive does have serialisation and deserialisation to del with various file format.

**74. Whenever we run a Hive query, a new metastore\_db is created. Why?**

*Unable to answer*

**75. Can we change the data type of a column in a hive table? Write a**

**complete query.**

Yes, we can change the datatype of a column in a hive table

ALTER TABLE <table-name> CHANGE <old-col-name> <new-col-name> <data-type>;

**76. 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 ?**

We just need to give a complete HDFS file path if we want to load the data from HDFS

**77. What is the precedence order in Hive configuration?**

*Unable to solve*

**78. Which interface is used for accessing the Hive metastore?**

WebHCat API web interface can be used for Hive commands

**79. Is it possible to compress json in the Hive external table ?**

*Not able to solve*

**80. What is the difference between local and remote metastores?**

Local Metastore:- Here metastore service still runs in the same JVM as Hive but it connects to a database running in a separate process either on same machine or on a remote machine.

Remote Metastore:- Metastore runs in its own separate JVM not on hive service JVM.

**81. What is the purpose of archiving tables in Hive?**

*Not able to solve*

**82. What is DBPROPERTY in Hive?**

It means mentioning the details about the database created by user.

**83. Differentiate between local mode and MapReduce mode in Hive.**

*Need help to answe this questiom*