Hive interview Questions

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

Ans:The Apache Hive™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. The structure can be projected onto data already in storage.3.1.3 is current version of hive.

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

Ans:Hive doesn't support OLTP. Hive supports Online Analytical Processing (OLAP).No Hive does not provide insert and update at row level. So it is not suitable for OLTP system.

3. How is HIVE different from RDBMS? Does hive support ACID transactions. If not then give the proper reason.

Ans:RDBMS iis used for OLTP and hive is used for OLAP.In rdbms sql is used for query but in hive hql is used.Earlier hive doesnt support acid but now it supports acid when table is stored in form of orc.

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

Ans:The major components of Hive and its interaction with the Hadoop is demonstrated in the figure below and all the components are described further:

User Interface (UI) –

As the name describes User interface provide an interface between user and hive. It enables user to submit queries and

other operations to the system. Hive web UI, Hive command line, and Hive HD Insight (In windows server) are supported by the user interface.

Hive Server – It is referred to as Apache Thrift Server. It accepts the request from different clients and provides it to Hive Driver.

Driver –

Queries of the user after the interface are received by the driver within the Hive. Concept of session handles is implemented by driver. Execution and Fetching of APIs modelled on JDBC/ODBC interfaces is provided by the user.

Compiler –

Queries are parses, semantic analysis on the different query blocks and query expression is done by the compiler. Execution plan with the help of the table in the database and partition metadata observed from the metastore are generated by the compiler eventually.

Metastore –

All the structured data or information of the different tables and partition in the warehouse containing attributes and attributes level information are stored in the metastore. Sequences or de-sequences necessary to read and write data and the corresponding HDFS files where the data is stored. Hive selects corresponding database servers to stock the schema or Metadata of databases, tables, attributes in a table, data types of databases, and HDFS mapping.

Execution Engine –

Execution of the execution plan made by the compiler is performed in the execution engine. The plan is a DAG of stages. The dependencies within the various stages of the plan is managed by execution engine as well as it executes these stages on the suitable system components.

5. Mention what Hive query processor does? And Mention what are the components of a Hive query processor?
Ans:The following are the main components of the Hive Query Processor:

Parse and SemanticAnalysis (ql/parse) - This component contains the code for parsing SQL, converting it into Abstract Syntax Trees, converting the Abstract Syntax Trees into Operator Plans and finally converting the operator plans into a directed graph of tasks which are executed by Driver.java.
Optimizer (ql/optimizer) - This component contains some simple rule based optimizations like pruning non referenced columns from table scans (column pruning) that the Hive Query Processor does while converting SQL to a series of map/reduce tasks.
Plan Components (ql/plan) - This component contains the classes (which are called descriptors), that are used by the compiler (Parser, SemanticAnalysis and Optimizer) to pass the information to operator trees that is used by the execution code.
MetaData Layer (ql/metadata) - This component is used by the query processor to interface with the MetaStore in order to retrieve information about tables, partitions and the columns of the table. This information is used by the compiler to compile SQL to a series of map/reduce tasks.
Map/Reduce Execution Engine (ql/exec) - This component contains all the query operators and the framework that is used to invoke those operators from within the map/reduces tasks.
Hadoop Record Readers, Input and Output Formatters for Hive (ql/io) - This component contains the record readers and the input, output formatters that Hive registers with a Hadoop Job.
Sessions (ql/session) - A rudimentary session implementation for Hive.
Type interfaces (ql/typeinfo) - This component provides all the type information for table columns that is retrieved from

the MetaStore and the SerDes.
Hive Function Framework (ql/udf) - Framework and implementation of Hive operators, Functions and Aggregate Functions. This component also contains the interfaces that a user can implement to create user defined functions.
Tools (ql/tools) - Some simple tools provided by the query processing framework. Currently, this component contains the implementation of the lineage tool that can parse the query and show the source and destination tables of the query.

6. What are the three different modes in which we can operate Hive?
Ans:There are three modes for Hive Metastore deployment:
Embedded Metastore.
Local Metastore.
Remote Metastore.

7. Features and Limitations of Hive?
Ans:
Features:
Framework:Hive is a stable batch-processing framework built on top of the Hadoop Distributed File system and can work as a data warehouse.
Easy To Code: Hive uses HIVE query language to query structure data which is easy to code. The 100 lines of java code we use to query a structure data can be minimized to 4 lines with HQL.
Declarative: HQL is a declarative language like SQL means it is non-procedural.
Structure Of Table The table, the structure is similar to the RDBMS. It also supports partitioning and bucketing.
Supported data structures :Partition, Bucket, and tables are the 3 data structures that hive supports.
Supports ETL Apache hive supports ETL i.e. Extract Transform and Load. Before Hive python is used for ETL.
Storage Hive :supports users to access files from HDFS,

Apache HBase, Amazon S3, etc.
Capable Hive: is capable to process very large datasets of Petabytes in size.
Helps in processing unstructured data We can easily embed custom MapReduce code with Hive to process unstructured data.
Drivers: JDBC/ODBC drivers are also available in Hive.
Fault Tolerance Since we store Hive data on HDFS so fault tolerance is provided by Hadoop.

Limitation:
Does not support OLTP Apache Hive doesn't support online transaction processing (OLTP) but Online Analytical Processing(OLAP) is supported.
Doesn't support subqueries Subqueries are not supported.
Latency The latency in the apache hive query is very high.
Only non-real or cold data is supported Hive is not used for real-time data querying since it takes a while to produce a result.
Transaction processing is not supported HQL does not support the Transaction processing feature.

8. How to create a Database in HIVE?
Ans:Create database hivedb;

9. How to create a table in HIVE?
Ans:Create table hive;

10.What do you mean by describe and describe extended and describe formatted with respect to database and table?
Ans:describe extended - This will show table columns, data types, and other details of the table. Other details will be displayed in single line. describe formatted - This will show table columns, data types, and other details of the table. Other details will be displayed into multiple lines.

11.How to skip header rows from a table in Hive?
Ans:By using this property while creating tables.
tblproperties ("skip.header.line.count"="1");

12.What is a hive operator? What are the different types of hive operators?
Ans.Operators are onme which are used to perform operations in hive.
There 4 differenmt types of operators in hive:
Arithmatic operations.+-* etc.
Relational operators.>,=,< etc
logical operator .and ,or ,not
Complex operators

13.Explain about the Hive Built-In Functions
Ans:These functions are the function which are already available in hive and and we can use them by merely calling them.
Their types are as followes:
mathematical inbuilt functions.abs(),round(),ciel()
collection functions.size() etc.
type conversion functions.cast() etc.
conditional functions and string functions

14. Write hive DDL and DML commands.
Ans:DDL:this command is used for defining and creating structures in hive.
ex:create table hivetable(id int,name string)
DML:The command are used for manipulating or changing the values in db objects such as tables etc.
ex:insert into hivetables values(1,'dhiraj')

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

and CLUSTER BY in Hive.
Ans:
Soet by:it sort the records in each reducer based on the columns mentioned
order by:it sorts the records in each reducer as well as globally.
distributed by:The records having same column value are placed in same reducer.
distributed by sort by:it combines the work of both distributed by and sort by.so all recordshaving same column value are sent to same reducer as well as records are sorted in each reducer.

16.Difference between "Internal Table" and "External Table" and Mention when to choose “Internal Table” and “External Table” in Hive?
Ans:While create external table we use keyword external while creating internal we do not need to provide any kehyword.
if we delete external tABLE only metdadata gets deleted not records.but for internal table metdadata as well as data gets deleted.
Hive Internal Table: We can use the internal table in cases: When generating temporary tables. ...
Hive External Table: We can use the external table in cases: When we are not creating the table based on the existing table.

17.Where does the data of a Hive table get stored?
The data of the hive table gets stroed in hadoop inside hdfs.

18.Is it possible to change the default location of a managed table?
Yes, you can do it by using the clause – LOCATION '<hdfs_path>' we can change the default location of a managed table

19.What is a Metastore in Hive?

Metastore – The component that stores all the structure information of the various tables and partitions in the warehouse including column and column type information, the serializers and deserializers necessary to read and write data and the corresponding HDFS files where the data is stored.
Derby is the default database for the embedded metastore. Derby embeded JDBC driver class.

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

21.What is a partition in Hive? And Why do we perform partitioning in
Hive?
The partitioning in Hive means dividing the table into some parts based on the values of a particular column like date, course, city or country. The advantage of partitioning is that since the data is stored in slices, the query response time becomes faster.

22.What is the difference between dynamic partitioning and static
partitioning?
Ans:In static partitioning we need to specify the partition column value in each and every LOAD statement. dynamic partition allow us not to specify partition column value each time.

23.How do you check if a particular partition exists?
Ans:we can use command :show partition tablename
it will show all partition hence we can know from list if a partition exist or not.

24.How can you stop a partition form being queried?
By using the ENABLE OFFLINE clause with ALTER TABLE atatement.

25.Why do we need buckets? How Hive distributes the rows into buckets?
Ans:Buckets are used for storing the records based onm the hash value .the hash value of the record is calculated for example % function and then the record is placed in required bucket.

26.In Hive, how can you enable buckets?
set hive. enforce. bucketing = true;

27.How does bucketing help in the faster execution of queries?
Bucketing has several advantages. The number of buckets is fixed so it does not fluctuate with data. If two tables are bucketed by employee_id , Hive can create a logically correct sampling. Bucketing also aids in doing efficient map-side joins etc.

28.How to optimise Hive Performance? Explain in very detail.
Ans:The hive optimization techniques are as follows:

1.Bucketing
2.partitioning
3.Querying by vector
4.Querying by cost optimization

29. What is the use of Hcatalog?
Ans:HCatalog is a tool that allows you to access Hive metastore tables
30. Explain about the different types of join in Hive.
Ans:

1. Inner join in Hive
2. Left Outer Join in Hive
3. Right Outer Join in Hive
4. Full Outer Join in Hive

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

Ans:Yes cartesian join in hive is possible

32.Explain the SMB Join in Hive?

Ans: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. We need to enable the following properties to use SMB: > SET hive

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

Ans:Order by sorts the record within each reducer as well as globally but sort by sorts the records only reducer wise but not globally.

34.What is the usefulness of the DISTRIBUTED BY clause in

Ans:The distributed by clause feeds the records having same column value to same reducer.

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

Ans:By using command load data inpath sourcepath into table tablename

36.Wherever (Different Directory) I run the hive query, it creates a new metastore_db, please explain the reason for it?

Ans: Basically, it creates the local metastore, while we run the hive in embedded mode. Also, it looks whether metastore already exist or not before creating the metastore.

37. What will happen in case you have not issued the command: ‘SET hive.enforce.bucketing=true;’ before bucketing a table in Hive?

Ans:'SET hive. enforce. bucketing=true;' allows you to have the correct number of reducer while using 'CLUSTER BY' clause for bucketing a column. In case it's not done, one may find the number of files generated in the table directory to be unequal to the number of buckets.

38. Can a table be renamed in Hive?

Ans:table can be renamed in hive

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)

Ans:ALTER TABLE h_table CHANGE COLUMN new_col INT BEFORE x_col

40. What is serde operation in HIVE?

Ans: A SerDe allows Hive to read in data from a table, and write it back out to HDFS in any custom format. Anyone can write their own SerDe for their own data formats. See Hive SerDe for an introduction to SerDes.

41. Explain how Hive Deserializes and serialises the data?

Ans: the concepts serialise and deserialise back to front. Serialise is done on write, the structured data is serialised into a bit/byte stream for storage. On read, the data is deserialised from the bit/byte storage format to the structure required by the reader. eg Hive needs structures that look like rows and columns but hdfs stores the data in bit/byte blocks, so serialise on write, deserialise on read.

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

Ans:Basically, to read and write HDFS files Hive uses these FileFormat classes currently:

AD

TextInputFormat/HiveIgnoreKeyTextOutputFormat

It read/write data in plain text file format.

SequenceFileInputFormat/SequenceFileOutputFormat

It read/write data in Hadoop SequenceFile format.

Moreover, to serialize and deserialize data Hive uses these Hive SerDe classes currently:

MetadataTypedColumnsetSerDe

So, to read/write delimited records we use this Hive SerDe. Such as CSV, tab-separated control-A separated records (sorry, quote is not supported yet).

LazySimpleSerDe

Also, to read the same data format as MetadataTypedColumnsetSerDe and TCTLSeparatedProtocol, we can use this Hive SerDe. Moreover, it creates Objects in a lazy way. Hence, that offers better performance.

AD

Basically, with a specified encode charset starting in Hive 0.14.0, it supports read/write data.

For example:

ALTER TABLE person SET SERDEPROPERTIES (‘serialization.encoding’=’GBK’)

Since, the configuration property hive.lazysimple.extended_boolean_literal is set to true (Hive 0.14.0 and later) LazySimpleSerDe can treat ‘T’, ‘t’, ‘F’, ‘f’, ‘1’, and ‘0’ as extended, legal boolean literals.

However, the default is false. Hence it means only ‘TRUE’ and ‘FALSE’ are treated as legal boolean literals.

Thrift SerDe in Hive

To read/write Thrift serialized objects, we use this Hive SerDe. However, make sure, for the Thrift object the class file must be loaded first.

Dynamic SerDe in Hive

To read/write Thrift serialized objects we use this Hive SerDe. Although, it understands Thrift DDL so the schema of the object can be provided at runtime.

Also, it supports a lot of different protocols, including TBinaryProtocol, TJSONProtocol, TCTLSeparatedProtocol (which writes data in delimited records).

Also:

For JSON files, JsonSerDe was added in Hive 0.12.0. An Amazon SerDe is

available at s3://elasticmapreduce/samples/hive-ads/libs/jsonserde.jar for releases prior to 0.12.0.

In Hive 0.9.1 an Avro SerDe was added. Starting in Hive 0.14.0 its specification is implicit with the STORED AS AVRO clause.

Afterward, in Hive 0.11.0, a SerDe for the ORC file format was added.

Further, in Hive 0.10 and natively in Hive 0.13.0 a SerDe for Parquet was added via the plug-in.

Then, in Hive 0.14, a SerDe for CSV was added.

43. What is the need of custom Serde?

Ans:

A SerDe allows Hive to read in data from a table, and write it back out to HDFS in any custom format. Anyone can write their own SerDe for their own data formats. See Hive SerDe for an introduction to SerDes.

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

Ans:Different complex types are as followes

Arrays	ARRAY<data_type>
Maps	MAP<primitive_type, data_type>
Structs	`STRUCT<col_name : data_type [COMMENT col_comment], ...>`
Union	UNIONTYPE<data_type, data_type, …>

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

Ans:it is possible using source command

```
For example –
Hive> source /path/to/file/file_with_query.hql
```

46. 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.

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

Ans:show databases like 's%',using this command we can get required result

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

Ans:

LIKE is an operator similar to LIKE in SQL. We use LIKE to search for string with similar text.

E.g. user_name LIKE '%Smith'

RLIKE (Right-Like) is a special function in Hive where if any substring of A matches with B then it evaluates to true. It also obeys Java regular expression pattern. Users don't need to put % symbol for a simple match in RLIKE.

Hive provides RLIKE operator that can be used for searching advanced Regular Expressions in Java.

E.g. user_name RLIKE '.(Smith|Sam).'

This will return user_name that has Smith or Sam in it.

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

Ans: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?

Ans:

select cast('1.78' as float);

This is how we can convert

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

Ans:it will give error

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;

Ans:This will insert the records from table staged_employees into table employees but into their respective partition file based on country and state

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

Ans:Insert overwrite table tablename1 select * from table2

54. What is the maximum size of a string data type supported by Hive? Explain how Hive supports binary formats.

Ans:

The 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. Sequence file: It is a splittable, compressible, and row-oriented file with a general binary format.

55. What File Formats and Applications Does Hive Support?
Ans:
Hive facilitates managing large data sets supporting multiple data formats, including comma-separated value (. csv) TextFile, RCFile, ORC, and Parquet. The PXF Hive connector reads data stored in a Hive table.

56.How do ORC format tables help Hive to enhance its performance?
Ans:Using the ORC format leads to a reduction in the size of the data stored, as this file format has high compression ratios. As the data size is reduced, the time to read and write the data is also reduced. The ORC format improves query performance also by the way it stores data in a file

57.How can Hive avoid mapreduce while processing the query?
Ans:You can make Hive avoid MapReduce to return query results by setting the hive. exec. Mode.

58.What is view and indexing in hive?
Ans:Indexing is a relatively new feature in Hive. In Hive, the index table is different than the main table. Indexes facilitate in making query execution or search operation faster. However, storing indexes require disk space and creating an index involves cost.

59.Can the name of a view be the same as the name of a hive table?
Ans:The name of a view must be unique, and it cannot be the same as any table or database or view's name. CREATE TABLE… LIKE clause can be used to copy a view into another

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

Ans:here is a processing cost in arranging the values of the column on which index is created since Indexes occupies.

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

Ans:`SHOW INDEX ON table_name`

This command we can use to get all index on the table

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

Ans:hive> Set mapred.input.dir.recursive=true;

hive> Set hive.mapred.supports.subdirectories=true;

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

Ans:When you perform a "select * from <tablename>", Hive fetches the whole data from file as a FetchTask rather than a mapreduce task which just dumps the data as it is without doing anything on it. This is similar to "hadoop dfs -text <filename>"

However, while using "select <column> from <tablename>", Hive requires a map-reduce job since it needs to extract the 'column' from each row by parsing it from the file it loads.

64.What are the uses of Hive Explode?

Ans:

The explode function explodes an array to multiple rows. Returns a row-set with a single column (col), one row for each element from the array

65. What is the available mechanism for connecting applications when we run Hive as a server?

Ans:Thrift Client: Using Thrift, we can call Hive commands from various programming languages, such as C++, PHP, Java, Python, and Ruby

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

Ans: Absolutely, by using the LOCATION keyword, we can change the default location of Managed tables while creating the managed table in Hive.

67.What is the Hive ObjectInspector function?

Ans:Hive ObjectInspector is a group of flexible APIs to inspect value in different data representation, and developers can extend those API as needed, so technically, object inspector supports arbitrary data type in java

68.What is UDF in Hive?

Ans:A user-defined function (UDF) is **a function you define so you can call it from SQL**. As with built-in functions you can call from SQL, a UDF's logic typically extends or enhances SQL with functionality that SQL doesn't have or doesn't do well

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

Ans:load data inpath from hdfslocation into table tablename

70.What is TextInputFormat and SequenceFileInputFormat in hive.

Ans:

This is very familiar input format in the Hadoop. The input will be given as key and value to Mapper, where key and value are generated in record reader. The record reader is just like a multiplexing.

For TextInputFormat, No need to create external record reader.

Key generated-----LongWritable(position ,generally "\n" or offset)

Value generated-----Text of each line.

The important point in the TextInputFormat is

a) Number of maps created is equal to number of files given in input path.

b) For each line, map method in mapper will be called.

Sequence files are in the binary format which can be split and the main use of these files is to club two or more smaller files and make them as a one sequence file. In Hive we can create a sequence file by specifying

STORED AS SEQUENCEFILE in the end of a CREATE TABLE statement.

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

Ans:setting the MapReduce jobs to execute in strict mode set hive.

72.When do we use explode in Hive?

Ans:The explode function explodes an array to multiple rows. Returns a row-set with a single column (col), one row for each element from the array.

73.Can Hive process any type of data formats? Why? Explain in very detail
Ans:Hive supports four file formats those are TEXTFILE, SEQUENCEFILE, ORC and RCFILE (Record Columnar File). For single user metadata storage, Hive uses derby database and for multiple user Metadata or shared Metadata case Hive uses MYSQL.

74.Whenever we run a Hive query, a new metastore_db is created. Why?

Ans: Since you use Embedded derby mode. To use single metastore_db location. you need to change following properties.

```
<property>
  <name>javax.jdo.option.ConnectionURL</name>
    <value>jdbc:derby:;databaseName=/<file-location>/metastore_db;create=true</value>
  <description>JDBC connect string for a JDBC metastore</description>
</property>
```

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

Ans: ALTER TABLE table_name CHANGE column_name column_name new_datatype

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 ?

Ans:when we use local keyword in load data inpath it assumes data need to be taken from local otherwise from hdfs.

77. What is the precedence order in Hive configuration?

Ans:Hive SET command has the highest priority

-hiveconf option from Hive Command Line

hive-site.xml file

hive-default.xml file

hadoop-site.xml file

hadoop-default.xml file

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

Ans:WebHCat API web interface can be used for Hive commands. It is a REST API that allows applications to make HTTP requests to access the Hive metastore (HCatalog DDL).

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

Ans:Just gzip your files and put them as is (*.gz) into the table location

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

Ans: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?

Ans:The in-built support in Hive to convert files in existing partitions to Hadoop Archive (HAR) is one approach to reducing the number of files in sections, as the number of files in the filesystem directly affects the memory consumption in the Namenode.