Hive interview Questions

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

It is dataware housing service. It uses HDFS file system.it is a sql like framework and has its own query language. Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. Version is 4.0

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 OLTP system.

Hive supports Online Analytical Processing (OLAP), but not Online Transaction Processing (OLTP).

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

Relational databases, or RDBMS, is a database that stores data in a structured format with rows and columns, a structured form called “tables.” Hive, on the other hand, is a data warehousing system that offers data analysis and queries

Currently, Hive supports ACID transactions on tables that store ORC file format. Transaction tables cannot be accessed from the non-ACID Transaction Manager (Dummy Txn Manager) session. External tables cannot be created to support ACID since the changes on external tables are beyond Hive control.

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

**Thrift clients-**

**Hive server is based on apache thrift and that’s why it can serve requests from theft clients**

**JDBC clients-**

**Hive allows for java applications to connect to it using jdbc driver**

**ODBC clients-hive allows ODBC based on applications**

**User Interface (UI) –**   
As the name describes User interface provide an interface between user and hive

**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.

**Compiler –**   
Queries are parses, semantic analysis on the different query blocks and query expression is done by the compiler.

**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.

**Execution Engine –**   
Execution of the execution plan made by the compiler is performed in the execution engine.

5. Mention what Hive query processor does? And Mention what are the components of a Hive query processor?

Hive enables data summarization, querying, and analysis of data. Hive queries are written in HiveQL, which is a query language similar to SQL. Hive allows you to project structure on largely unstructured data. After you define the structure, you can use HiveQL to query the data without knowledge of Java or MapReduce

The major components of Apache Hive are the Hive clients, Hive services, Processing framework and Resource Management, and the Distributed Storage. The user interacts with the Hive through the user interface by submitting Hive queries

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

We can run Hive in following modes:

* 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. Name Node, Data Node, Job Tracker, Task Tracker 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.

They are:

* Hive doesn't support OLTP. Hive supports Online Analytical Processing (OLAP), but not Online Transaction Processing (OLTP).
* It doesn't support subqueries.
* It has a high latency.
* Hive tables don't support delete or update operations.

Some of the limitations of Apache Hive are:

1. Hive is not designed for the OLTP (Online transaction processing). We can use it for OLAP.

2. It does not offer real-time queries.

3. It provides limited subquery support.

4. Latency of Hive is generally very high.

8. How to create a Database in HIVE?

Create database database\_name;

9. How to create a table in HIVE?

We have 2 type of tables in hive

Internal and external tables

Example

Create table table\_name

{dept\_id int,

Salary int}

Row formatted delimited

Fields termited by”,”;

10. What do you mean by describe and describe extended and describe formatted with respect to database and table

The DESCRIBE DATABASE command is used to get information about the database, such as the name of the database, its comment (if attached during the creation of the database), its location on the filesystem, and its db properties

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?

TBLPROPERTIES ("skip.header.line.count"="1");

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

Apache Hive provides various Built-in operators for data operations to be implemented on the tables present inside Apache Hive warehouse.

Hive operators are used for mathematical operations on operands. It returns specific value as per the logic applied.

Relational Operators

Arithmetic Operators

Logical Operators

String Operators

13. Explain about the Hive Built-In Functions

round(double a) It returns the rounded BIGINT value of the double.

floor(double a) It returns the maximum BIGINT value that is equal or less than the double.

ceil(double a) It returns the minimum BIGINT value that is equal or greater than double

14. Write hive DDL and DML commands.

How to run DDL file in Hive?

DDL Commands on Tables in Hive

DDL means data definition language

DML means data manuplation language

DDL  
create

Drop

Alter

DML

Insert

Update

delete

15. Explain about SORT BY, ORDER BY, DISTRIBUTE BY and CLUSTER BY in Hive.

Cluster By is a short-cut for both Distribute By and Sort By. Hive uses the columns in Distribute By to distribute the rows among reducers. All rows with the same Distribute By columns will go to the same reducer. However, Distribute By does not guarantee clustering or sorting properties on the distributed keys.

16. Difference between "Internal Table" and "External Table" and Mention when to choose “Internal Table” and “External Table” in Hive?

For External Tables, Hive stores the data in the LOCATION specified during creation of the table(generally not in warehouse directory). If the external table is dropped, then the table metadata is deleted but not the data. For Internal tables, Hive stores data into its warehouse directory.

1. Hive Internal Table. We can use the internal table in cases: When generating temporary tables. ...
2. 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?

Hive stores its database and table metadata in a metastore, which is a database or file backed store that enables easy data abstraction and discovery.

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? What is the default database provided by Apache Hive for metastore?

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. Unit test data goes in here on your local filesystem. If you want to run Derby as a network server so the metastore can be accessed from multiple nodes, see Hive Using Derby in Server Mode.

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?

The partitioning in Hive means dividing the table into some parts based on the values of a particular column like date, course, city or countr**y**. 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?

Dynamic Partition takes more time in loading data compared to static partition. When you have large data stored in a table then the Dynamic partition is suitable. If you want to partition a number of columns but you don't know how many columns then also dynamic partition is suitable.

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

You can run the HDFS list command to show all partition folders of a table from the Hive data warehouse location

24. How can you stop a partition form being queried?

Alter table table name drop column

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

Bucketing in hive is useful when dealing with large datasets that may need to be segregated into clusters for more efficient management and to be able to perform join queries with other large datasets

Bucketing feature of Hive can be used to distribute/organize the table/partition data into multiple files such that similar records are present in the same file. While creating a Hive table, a user needs to give the columns to be used for bucketing and the number of buckets to store the data into.

26. In Hive, how can you enable buckets?

partition=true property. So, we can enable dynamic bucketing while loading data into hive table By setting this property. ii. Moreover, it will automatically set the number of reduce tasks to be equal to the number of buckets mentioned in the table definition (for example 32 in our case).

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

How does bucketing help in the faster execution of queries?

Bucketing in hive is the concept of breaking data down into ranges, which are known as buckets, to give extra structure to the data so it may be used for more efficient queries. The range for a bucket is determined by the hash value of one or more columns in the dataset (or Hive metastore table)

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

1. Partitioning Tables: Hive partitioning is an effective method to improve the query performance on larger tables. ...
2. De-normalizing data: ...
3. Compress map/reduce output: ...
4. Map join: ...
5. Input Format Selection: ...
6. Parallel execution: ...
7. Vectorization: ...
8. Unit Testing:
9. Sampling:

29. What is the use of Hcatalog?

HCatalog is a tool that allows you to access Hive metastore tables within Pig, Spark SQL, and/or custom MapReduce applications.

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

Moreover, there are several types of Hive join – HiveQL Select Joins: Hive inner join, hive left outer join, hive right outer join, and hive full outer join. We will also learn Hive Join tables in depth.

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

It is a way of joining multiple tables in which all the rows or tuples from one table are paired with the rows and tuples from another table.

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. 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?

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 are more than one reducer, "sort by" may give partially ordered final results.

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

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.

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

1. Ingest the data. You create a single Sqoop import command that imports data from diverse data sources, such as a relational database, into HDFS.
2. Convert the data to ORC format. ...
3. Incrementally update the imported data.

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

When running Hive in embedded mode, it creates a local metastore. When you run the query, it first checks whether a metastore already exists or not.

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

The command: '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?

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)

ALTER table table name add [clomn] column\_definition;

40. What is serde operation in HIVE?

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?

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 in data from a table, and write it back out to HDFS in any custom format.

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

SerDe is short for Serializer/Deserializer. 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.

43. What is the need of custom Serde?

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?

Similar to Spark, Hive also support complex data types which includes Array, Map, Struct and union. Array is used to store the list of elements. Map is used to store key/value pair. Struct is for parent and child assosiations.

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

We can execute Hive queries from the script files using the source command.

**Hive -f command is used to execute one or more hive queries from a file in batch mode.**

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?

Show databases like ‘s\*’

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

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

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?

Float=float(51.2)

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

Output in the form of int

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;

it prints all serialized country and serialized state and it is partition by only country and state from staged\_employees serde table

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

INSERT OVERWRITE is used to replace any existing data in the table or partition and insert with the new rows.

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

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?

Hive supports all Avro types. However, Impala does not support complex or nested types with Avro, such as enum, array, fixed, map, union, and record (nested).

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

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.

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

You can make Hive avoid MapReduce to return query results by setting the hive. exec. mode.

58. What is view and indexing in hive?

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?

The name of a view must be unique, and it cannot be the same as any table or database or view's name.

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

Basically, there 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.

SHOW INDEXES ;

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

mapred. supports. subdirectories=true; Once above options are set to true, Hive will recursively access sub-directories of a directory in MapReduce.

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

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?

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?

**ODBC Driver**-This supports the ODBC protocol JDBC Driver- This supports the JDBC protocol

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

Yes, you can do it by using the clause – LOCATION '<hdfs\_path>' we can change the default location of a managed table.

67. What is the Hive ObjectInspector function?

A key concept when working with Generic UDF and UDAF is the ObjectInspector. In generic UDFs, all objects are passed around using the Object type. Hive is structured this way so that all code handling records and cells is generic, and to avoid the costs of instantiating and deserializing objects when it's not needed.

68. What is UDF in Hive?

The Db2 Big SQL environment in Hadoop includes the Hive user-defined functions package. This set of functions is an optional package that you can install to use some of the Hive open source user-defined functions in your Db2 Big SQL queries.

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

INSERT OVERWRITE DIRECTORY "HDFS Path" ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'

SELECT \* FROM XXXX LIMIT 10;

70. What is TextInputFormat and SequenceFileInputFormat in hive

Class TextInputFormat

An InputFormat for plain text files. Files are broken into lines. Either linefeed or carriage-return are used to signal end of line. Keys are the position in the file, and values are the line of text..

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?

This can be achieved by setting the MapReduce jobs to execute in strict mode set hive.

72. When do we use explode in Hive?

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

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?

Therefore, we have to change the behavior of the location to an absolute path so that from that location the metastore can be used.

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

You can replace the column name and the datatype using the ALTER command. Refer to the command below: hive> ALTER TABLE <tablename> REPLACE COLUMNS (<old column name> INT, <new column name> STRING);

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 ?

hive> LOAD DATA INPATH '/user/myname/kv2. txt' OVERWRITE INTO TABLE invites PARTITION (ds='2008-08-15'); The above command will load data from an HDFS file/directory to the table.

77. What is the precedence order in Hive configuration?

We are using a precedence hierarchy for setting properties: The SET command in Hive. The command-line –hiveconf option.

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

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 ?

As text data, JSON data compresses nicely. That's why gzip is our first option to reduce the JSON data size. Moreover, it can be automatically applied in HTTP, the common protocol for sending and receiving JSON. Let's take the JSON produced with the default Jackson options and compress it with gzip

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.

Remote Metastore. Moving further, another metastore configuration called Remote Metastore. In this mode, metastore runs on its own separate JVM, not in the Hive service JVM. If other processes want to communicate with the metastore server they can communicate using Thrift Network APIs.

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

You can use Hadoop archiving to reduce the number of hdfs files in the Hive table partition. Hive has built in functions to convert Hive table partition into Hadoop Archive (HAR). HAR does not compress the files, it is analogous to the Linux tar command.2

82. What is DBPROPERTY in Hive?

The DB properties are nothing but mentioning the details about the database created by the user. Suppose the name of the user, the type of the database and the tables it has, the date on which the database is created etc. This makes the other user easy the recognize the database and use it according to the requirement.

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

Local mode is actually a local simulation of MapReduce in Hadoop's LocalJobRunner class. MapReduce mode (also known as Hadoop mode): Pig is executed on the Hadoop cluster.