Hive Interview Questions

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

Hive is a data warehouse system which is used to analyse the structured data. It is built on top of Hadoop. Hive provides the functionality of reading, writing ,and managing large datasets.

It runs SQL like queries called HQL (Hive Query Language) which gets internally converted into Map-Reduce Jobs.

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

No, hive cannot be used for OLTP it can be used only for OLAP.

The reason is hive will not allow us to 'insert' and 'update' at the row level.

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

RDBMS database allows us to store data in a structured manner.

Where hive is a datawarehouse system that offers data analysis and queries.

Yes,hive supports ACID transcations in here it is called as HIVE tables.

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

The Hive architecture consists of

Hive Client:

-JDBC//Hive allows to connect with the JDBC driver

-ODBC//Hive allows to connect with ODBC(Open Database Connectivity)

-Thrift//hive server is based on Apache thrift

Hive Services:

-Hive server2 //It enables clients to submit queries for execution

-Hive Driver//receives hql query from user and creates session handles for query and sends the query to the compiler

-Hive Compiler//It parses the query

-Meta Store//it is the central repository which stores the meta data information about the structure of the table.

Processing & Resource Management

Distributed Storage

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

Hive query processor is responsible for parsing, analysing , optimizing and executing hive queries.

Query Processor components are:

-Parser

-Semantic analyser

-Query optimizer

-Query planner

-Driver

-Exexution engine

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

-Local Mode//it is used for small dataset ,in here one JVM is enough

-MapReduce mode//Hive uses mapreduce method to process data that is stored in HDFS

-Spark mode//hive uses spark method to process data stored in hdfs.

7. Features and Limitations of Hive.

Features:

-Scalabilty

-Flexibility

-Cost effective

-SQL like interface

Limitation

-It is suitable only for batch processing not designed for real time processing

-Performance:While dealing with large data it takes more time.

-Complex queries can be hard to write.

-Limited support for transcation.

8. How to create a Database in HIVE?

We can create it using the command

Create database databaseeg;

To use it use the command

Use databaseeg;

9. How to create a table in HIVE?

Create table barceelona(id int,name string,age int);

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

Describe db// show the name of db and its location in hdfs.

Describe extended db// table type, the input format, the output format, the serialization format, and the partition keys.

Describe formatted db// its show more details including the location of hdfs,storage properties and table paramaters.

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

With the help of

Tblproperties(skip.header.line.count='1')

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

Hive operators are:

Arithmetic operator: +,-,*,/

Logical operator: and , or , not

Comparison operator: >,<,<=,>=

String operator:CONCAT,SUBSTR

Conditional operator: IF, CASE and WHEN

13.Explain about the Hive Built-In Functions

Mathematical fn: exp,log,pow,ceil,abs

String fn: concat,substr,replace

Conditional : IF,CASE,THEN

Aggregate fn: min,max,sum,avg

14. Write hive DDL and DML commands.

DDL:

-create database mydb

-create table(name string,age int)

-alter table tablename add column sex string

-drop table tablename

-show tables

-show databases

DML

-select *from table

-insert into table values('Raj',34)

-delete from tablename where age=34

-select *from tablename limit 10

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

Sort by:

It is used to sort the data within each reducer task

Select *from employee SORT BY age

Order By;

It is used to sort the data globally across all the reducer task.

Select *from employee ORDER BY desc

DISTRIBUTE BY

It is used to distribute the data across reducer task.

Select *from employee DISTRIBUTE BY name;

CLUSTER BY

It is used to cluster the data and sort it within the each cluster.

Select *from employee CLUSTER BY name;

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

Internal table:

It is stored in a directory managed by hive.

When internal table is dropped the data associated with it is also deleted.

External Table:

The data of external table is stored in a location outside of hive.

When external table is dropped the data associated with this will not be deleted.

When to use:

Internal Table

If we want to use the data management capabalites of partitioning and bucketing.

If we have a smaller dataset.

External table

If we want to share data between tools in addition to hive such as spark,pig.

If we want to keep data away from hive.

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

It will be stored either in internal table or external table.

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

Yes,It is possible to change.

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

Metastore in hive is the central repository that stores the meta data about tables,partitioning such as partitioning scheme.

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

It doesn't store because HDFS meant to store only the large dataset and it doesn't provide write access.

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

Partition is a way to divide a table into smaller ,manageable parts based on values of one or more columns.

We do partitioning in hive to improve query performance reduce the amount of data that needs to be scanned.

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

Static partitioning is defined at the time of table creation and involves specifying the particular column and their value explicitly

Dynamic Partitioning hive will automatically creates partitioning based on the value in the data as it is loaded into the table.

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

It can be checked using 'show partition' command in hive

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

We can prevent a specific partition by making it 'unavailable' or 'Offline'.

Alter table employee partition set OFFLINE

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

Buckets inn hive is used to physically divide a table into smaller parts based on the hash value of a specified column

26.In Hive, how can you enable buckets?

It can be enabled by

‘set hive.enforce.bucketing=true;’

27.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.

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

By

1.Tune Hive configuration: set hive.exec.dynamic.partiton=nonstrict

2.Bucketing and partitioning

29. What is the use of Hcatalog?

Hcatlog is the storage and table management layer for Hadoop that provides a unified interface for accessing data stored in various Hadoop data stores such as Hadoop.

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

Inner Join

Returns only the rows that have natching values in both tables.

Left Join

Returns all the rows from the left table and matching rows from the right table.

Right Join

Returns all the rows from the right table and the matching rows from the left table.

Full Outer Join

Return all the rows from both the tables with null values in the columns where there are no matching rows.

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

Yes, It is possible to create a Cartesian join between 2 tables.

32.Explain the SMB Join in Hive?

SMB (Sort-Merge-Bucket) join is a type of join optimization technique used in Hive to improve the performance of join operations on large tables

The SMB join consists of three main steps:

Sort: The data in both tables being joined is sorted on the join key. This is done using a MapReduce job that sorts the data within each bucket independently.

Merge: The sorted data from both tables is merged on the join key. This is done using another MapReduce job that combines the matching rows from each table.

Bucket: The result of the merge is then written to a new bucketed table.

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

'Order by ' sorts the data on the specified column and returns the result in asc/desc.

'Sort by' sorts the data only within each reducer.

'SORT BY' is faster and more scalable than 'ORDER BY'

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

The DISTRIBUTE BY clause in Hive is used to specify the columns by which the data is partitioned before the sort is applied. This clause can be useful for improving the performance of sorting operations by reducing the amount of data that needs to be shuffled between reducers.

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

1.Create a table in hive

2.Load data into hive.

3.Optimize data format

4.Paritioning

5.Compression

6.Query Data

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

Basically It creates the local metastore ,while we run the hive in embedded mode .Also it looks whether metastore is already present or notbefore creating the metastore.

Hence in the configuration file hive-site.xml

Property is "javax.jdo.option.ConnectionURL" with default value "jdbc:derby:;databaseName=metastore_db;create=true" this property is defined.

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

The bucketing operation may not work as expected and the tables may not br properly bucketed.

It may write the data to the wrong bucket or may not properly bucket the data at all.

38.Can a table be renamed in Hive?

Yes , it can be renamed with the command

Alter table employee rename to staff;

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 COLUMN (new_col int) BEFORE x_col;

40.What is serde operation in HIVE?

Serde(serializer/deserializer) it is th way of serialize or deserialize the data in a particular format

SERDE operation are used to convert data between hive internal representation and external data formats such as csv,json

41.Explain how Hive Deserializes and serialises the data?

Deserializes means converting the data from an external format(excel,json) into an hive internal format.

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

The built in serde in hive are

-JsonSerde

-OrcSerde

-ParquetHiveSerde

43.What is the need of custom Serde?

The need for custom SerDe in Hive arises when there is a requirement to read and write data in a format that is not supported by the built-in SerDe classes.

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

-Array

-Map

-Struct

-Union

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

Yes , it can be executed

It can be done using CLI(Command Line interface) or using programming language such as python

To execute hive queries from a script file using the Hive CLI

46.What are the default record and field delimiter used for hive text files?

In hive the default record delimiter for text file is the new line character (‘\n’)

This means the each line in a text record is considered as a record.

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

It can be done using the command

SHOW DATABSE like ‘S*’

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

Both are used for pattern matching in string values.

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?

Alter table employee change column nationality string;

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

Select cast(column_name as float) from table_name;

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

We will get the result as ‘NULL’ coz ‘abc’ cannot be casted into int

52.What does the following query do?

a. INSERT OVERWRITE TABLE employees

This line indicates that the data will be inserted into the table named "employees." The "OVERWRITE" keyword indicates that if the table already exists, any existing data in the table will be overwritten by the new data.

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.

INSERT OVERWRITE TABLE new_table

SELECT * FROM existing_table

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

Explain how Hive supports binary formats.

(between 1 and 65535), which defines the maximum number of characters allowed in the character string.

Hive supports binary formats by providing two types of data types: binary and varbinary.

55. What File Formats and Applications Does Hive Support?

-text files

-sequence files

-Orc files

-parquet files

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

ORC(optimized row columnar) file is a columnar storage format for huve.Hive configuration setting for ORC format tables can improve query perfomance resulting in faster execution and reduced usage of computing resource.

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

Hive can avoid mapreduce while running a query in hive by setting the command as

'hive.exec.mode.local'

58.What is view and indexing in hive?

Views:

Views are generated based on the user requirement.You can save any result set data as a view.

Indexing:

Indexing is nothing but a pointer ok a particular column of a table.

Creating a index means creating a pointer on a particular column of a table.

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

No, the name of a view cannot be same as the nice table name.

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

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 INDEX from employee

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

We need 2 commands as

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?

When we perform a select *.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.

However while using select name from employee ,hive requires a map reduce job.

64.What are the uses of Hive Explode?

Explode is a User defined table generating function.

It takes array as an input and outputs the elements of array as seperate rows.

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

Thrift client: From thrift you can call hive commands from various programming languages such as Java,php

JDBC Driver:

ODBC Driver

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

Yes it can be changed by using the below command

ALTER TABLE employee SET LOCATION 'hdfs://hdjdhh'

67.What is the Hive ObjectInspector function?

ObjectInspector is used to analyze the internal structure of a row object and individual structure of columns

It also offers ways to access complex objects that can be stored in different formats in memory

68.What is UDF in Hive?

UDF(User Defined Function)

It allows us to create custom functions to process records or group of records.

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

INSERT OVERWRITE TABLE <hive_table_name>

SELECT * FROM <hdfs_directory_path>;

70.What is TextInputFormat and SequenceFileInputFormat in hive.

These two are used to read the input data from the hdfs.

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

-Use appropriate query optimization techniques.

-Increase hardware resources.

-Set query limit.

-Break down the query into smaller chunks.

72.When do we use explode in Hive?

The explode() function in hive is used to split a column that contains an array datatype into multiple rows

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

Yes, hive can process any type of data formats

It supports data formats such as Text files,sequence files,orc files , json files

Why:

Coz hive supports various serialization and deseralization techniques

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

When you run a Hive query, Hive stores the metadata about the table and the query execution in a metastore. The metastore is a central repository that stores metadata information about Hive objects like tables, partitions, columns, and their properties. This information is used by Hive to optimize queries, track dependencies, and manage resources.

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

ALTER TABLE employee CHANGE COLUMN age int

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 ?

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

LOAD DATA INPATH 'hdfs://<namenode>:<port>/<path_to_file>' OVERWRITE INTO TABLE <table_name>;

77.What is the precedence order in Hive configuration?

From highest to lowest

-Session level configuration

-Table level configuration

-Query level configuration

-Cluster level configuration

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

WebHCat API web interface is used.

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

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

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?

You can use Hadoop archiving to reduce the number of hdfs files in the Hive table partition

82.What is DBPROPERTY in Hive?

The DB properties are nothing but mentioning the details about the database created by the user.

83.Differentiate between local mode and MapReduce mode in Hive

Both MapReduce mode and local mode seem same to the user but the difference is the way they execute. Local mode: In this mode, Pig script runs on a Single machine without the need of Hadoop cluster or hdfs. Local mode is used for development purpose to see how the script would behave in an actual environments.