**Apache Hive Interview Questions**

**1. Define the difference between Hive and HBase?**

|  |  |
| --- | --- |
| **Hive vs HBase** | |
| **HBase** | **Hive** |
| 1. HBase is built on the top of HDFS | 1. It is a data warehousing infrastructure |
| 2. HBase operations run in a real-time on its database rather | 2. Hive queries are executed as MapReduce jobs internally |
| 3. Provides low latency to single rows from huge datasets | 3. Provides high latency for huge datasets |
| 4. Provides random access to data | 4. Provides random access to data |

**2. What kind of applications is supported by Apache Hive?**

Hive supports all those client applications that are written in Java,PHP,Python,C++,Ruby by exposing its Thrift server.

**3. Where does the data of a Hive table gets stored?**

By default, the Hive table is stored in an HDFS directory – /user/hive/warehouse. One can change it by specifying the desired directory in hive.metastore.warehouse.dir configuration parameter present in the hive-site.xml.

**4. What is a metastore in Hive?**

[Metastore](https://www.edureka.co/blog/hive-tutorial/) in Hive stores the meta data information using RDBMS and an open source ORM (Object Relational Model) layer called Data Nucleus which converts the object representation into relational schema and vice versa.

**5. Why Hive does 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.

**6. What is the difference between local and remote metastore?**

Local Metastore:

In local metastore configuration, the metastore service runs in the same JVM in which the Hive service is running and connects to a database running in a separate JVM, either on the same machine or on a remote machine.

Remote Metastore:

In the remote metastore configuration, the metastore service runs on its own separate JVM and not in the Hive service JVM. Other processes communicate with the metastore server using Thrift Network APIs. You can have one or more metastore servers in this case to provide more availability.

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

By default, Hive provides an embedded Derby database instance backed by the local disk for the metastore. This is called the embedded metastore configuration.

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

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

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

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

a) javax.jdo.option.ConnectionURL property should be set to jdbc:mysql://host/dbname?createDataba  
seIfNotExist=true.

b) javax.jdo.option.ConnectionDriverName property should be set to com.mysql.jdbc.Driver.

c) One should also set the username and password as:

javax.jdo.option.ConnectionUserName is set to desired username.

javax.jdo.option.ConnectionPassword is set to the desired password.

2) The JDBC driver JAR file for MySQL must be on the Hive’s classpath, i.e. The jar file should be copied into the Hive’s lib directory.

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

**9. What is the difference between external table and managed table?**

Here is the key difference between an external table and managed table:

1) In case of managed table, If one drops a managed table, the metadata information along with the table data is deleted from the Hive warehouse directory.

2) On the contrary, in case of an external table, Hive just deletes the metadata information regarding the table and leaves the table data present in HDFS untouched.

Note: I would suggest you to go through the blog on [Hive Tutorial](https://www.edureka.co/blog/hive-tutorial/?#data_model)to learn more about Managed Table and External Table in Hive.

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

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

**11. When should we use SORT BY instead of ORDER BY?**

We should use SORT BY instead of ORDER BY when we have to sort huge datasets because SORT BY clause sorts the data using multiple reducers whereas ORDER BY sorts all of the data together using a single reducer. Therefore, using ORDER BY against a large number of inputs will take a lot of time to execute.

**12. What is a partition in Hive?**

Hive organizes tables into partitions for grouping similar type of data together based on a column or partition key. Each Table can have one or more partition keys to identify a particular partition. Physically, a partition is nothing but a sub-directory in the table directory.

**13. Why do we perform partitioning in Hive?**

Partitioning provides granularity in a Hive table and therefore, reduces the query latency by scanning only relevant partitioned data instead of the whole data set.

For example, we can partition a transaction log of an e – commerce website based on month like Jan, February, etc. So, any analytics regarding a particular month, say Jan, will have to scan the Jan partition (sub – directory) only instead of the whole table data.

**14. What is dynamic partitioning and when is it used?**

In dynamic partitioning values for partition columns are known in the runtime, i.e. It is known during loading of the data into a Hive table.

One may use dynamic partition in following two cases:

1) Loading data from an existing non-partitioned table to improve the sampling and therefore, decrease the query latency.

2) When one does not know all the values of the partitions before hand and therefore, finding these partition values manually from a huge data sets is a tedious task.

**15. Suppose, I create a table that contains details of all the transactions done by the customers of year 2016: CREATE TABLE transaction\_details (cust\_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;**

**Now, after inserting 50,000 tuples in this table, I want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so?**

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

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

Create a partitioned table, say partitioned\_transaction:

CREATE TABLE partitioned\_transaction (cust\_id INT, amount FLOAT, country STRING) PARTITIONED BY (month STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;

2. Enable dynamic partitioning in Hive:

SET hive.exec.dynamic.partition = true;

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

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

INSERT OVERWRITE TABLE partitioned\_transaction PARTITION (month) SELECT cust\_id, amount, country, month FROM transaction\_details;

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

**16. How can you add a new partition for the month December in the above partitioned table?**

For adding a new partition in the above table partitioned\_transaction, we will issue the command give below:

ALTER TABLE partitioned\_transaction ADD PARTITION (month=’Dec’) LOCATION  ‘/partitioned\_transaction’;

Note: I suggest you to go through the dedicated blog on [Hive Commands](https://www.edureka.co/blog/hive-commands-with-examples) where all the commands present in Apache Hive have been explained with an example.

**17. What is the default maximum dynamic partition that can be created by a mapper/reducer? How can you change it?**

By default the number of maximum partition that can be created by a mapper or reducer is set to 100. One can change it by issuing the following command:

SET hive.exec.max.dynamic.partitions.pernode = <value>

Note: You can set the total number of dynamic partitions that can be created by one statement by using: SET hive.exec.max.dynamic.partitions = <value>

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

To remove this error one has to execute following commands:

SET hive.exec.dynamic.partition = true;

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

Things to Remember:

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

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

**19. Why do we need buckets?**

There are two main reasons for performing bucketing to a partition:

1) A [map side join](https://www.edureka.co/blog/map-side-join-vs-join/) requires the data belonging to a unique join key to be present in the same partition. But what about those cases where your partition key differs from that of join key? Therefore, in these cases you can perform a map side join by bucketing the table using the join key.

2) Bucketing makes the sampling process more efficient and therefore, allows us to decrease the query time.

**20. How Hive distributes the rows into buckets?**

Hive determines the bucket number for a row by using the formula: hash\_function (bucketing\_column) modulo (num\_of\_buckets). Here, hash\_function depends on the column data type. For integer data type, the hash\_function will be:

hash\_function (int\_type\_column)= value of int\_type\_column

**21. What will happen in case you have not issued the command:  ‘SET hive.enforce.bucketing=true;’before bucketing a table in Hive in Apache Hive 0.x or 1.x?**

The command:  ‘SET hive.enforce.bucketing=true;’ allows one 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 that will be generated in the table directory to be not equal to the number of buckets. As an alternative, one may also set the number of reducer equal to the number of buckets by using set mapred.reduce.task = num\_bucket.

**22. What is indexing and why do we need it?**

One of the Hive query optimization methods is Hive index. Hive index is used to speed up the access of a column or set of columns in a Hive database because with the use of index the database system does not need to read all rows in the table to find the data that one has selected.

**23. Scenario: Suppose, I have a CSV file – ‘sample.csv’ present in ‘/temp’ directory with the following entries:**

id first\_name last\_name email gender ip\_address

1 Hugh Jackman hughjackman@cam.ac.uk Male 136.90.241.52

2 David Lawrence dlawrence1@gmail.com Male 101.177.15.130

3 Andy Hall andyhall2@yahoo.com Female 114.123.153.64

4 Samuel Jackson samjackson231@sun.com Male 89.60.227.31

5 Emily Rose rose.emily4@surveymonkey.com Female 119.92.21.19

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

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

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

CREATE EXTERNAL TABLE sample

(id int, first\_name string,

last\_name string, email string,

gender string, ip\_address string)

ROW FORMAT SERDE ‘org.apache.hadoop.hive.serde2.OpenCSVSerde’

STORED AS TEXTFILE LOCATION ‘/temp’;

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

SELECT first\_name FROM sample WHERE gender = ‘male’;

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

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

1) Create a temporary table:

CREATE TABLE temp\_table (id INT, name STRING, e-mail STRING, country STRING)

ROW FORMAT FIELDS DELIMITED TERMINATED BY ‘,’ STORED AS TEXTFILE;

2) Load the data into temp\_table:

LOAD DATA INPATH ‘/input’ INTO TABLE temp\_table;

3) Create a table that will store data in SequenceFile format:

CREATE TABLE sample\_seqfile (id INT, name STRING, e-mail STRING, country STRING)

ROW FORMAT FIELDS DELIMITED TERMINATED BY ‘,’ STORED AS SEQUENCEFILE;

4) Transfer the data from the temporary table into the sample\_seqfile table:

INSERT OVERWRITE TABLE sample SELECT \* FROM temp\_table;

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

**26. What is Apache Hive?**

Basically, a tool which we call a data warehousing tool is Hive. However, Hive gives [SQL queries](https://data-flair.training/blogs/sql-subquery/) to perform an analysis and also an abstraction. Although, Hive it is not a database it gives you logical abstraction over the databases and the tables.

**Is Hive suitable to be used for OLTP systems? Why?**

No, it is not suitable for OLTP system since it does not offer insert and update at the row level.

**Where does the data of a Hive table gets stored?**

In an HDFS directory – /user/hive/warehouse, the Hive table is stored, by default only. Moreover, by specifying the desired directory in hive.metastore.warehouse.dir configuration parameter present in the hive-site.xml, one can change it.   
  
**What is a metastore in Hive?**

Basically, to store the metadata information in the Hive we use [Metastore](https://data-flair.training/blogs/apache-hive-metastore/). Though, it is possible by using RDBMS and an open source ORM (Object Relational Model) layer called Data Nucleus. That converts the object representation into the relational schema and vice versa.

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

Using RDBMS instead of HDFS, Hive stores metadata information in the metastore. Basically, to achieve low latency we use RDBMS. Because HDFS read/write operations are time-consuming processes.  
  
**What is the difference between local and remote metastore?**

Local Metastore:

It is the metastore service runs in the same [JVM](https://data-flair.training/blogs/java-virtual-machine-jvm/) in which the Hive service is running and connects to a database running in a separate JVM. Either on the same machine or on a remote machine.  
  
Remote Metastore:  
In this configuration, the metastore service runs on its own separate JVM and not in the Hive service JVM.  
  
**What is the default database provided by Apache Hive for metastore?**

It offers an embedded Derby database instance backed by the local disk for the metastore, by default. It is what we call embedded metastore configuration.

**What is the difference between the external table and managed table?**

Managed table

The metadata information along with the table data is deleted from the Hive warehouse directory if one drops a managed table.\

External table  
Hive just deletes the metadata information regarding the table. Further, it leaves the table data present in HDFS untouched.

[Read more about Hive internal tables vs External tables](https://data-flair.training/blogs/hive-internal-tables-vs-external-tables-comparison/)

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

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

Hive Interview Questions for Freshers- Q. 1,2,3,4,5,7,8,9,10

Hive Interview Questions for Experience- Q. 6  
  
**When should we use SORT BY instead of ORDER BY?**

Despite [ORDER BY](https://data-flair.training/blogs/hiveql-order-by-query/) we should use SORT BY. Especially while we have to sort huge datasets. The reason is SORT BY clause sorts the data using multiple reducers. ORDER BY sorts all of the data together using a single reducer. Hence, using ORDER BY will take a lot of time to execute a large number of inputs.

**What is a partition in Hive?**

Basically, for the purpose of grouping similar type of data together on the basis of column or partition key, Hive organizes tables into partitions. Moreover, to identify a particular partition each table can have one or more partition keys. On defining [Hive Partition](https://data-flair.training/blogs/apache-hive-partitioning/), in other words, it is a sub-directory in the table directory.

**Why do we perform partitioning in Hive?**

In a Hive table, Partitioning provides granularity. Hence, by scanning only relevant partitioned data instead of the whole dataset it reduces the query latency.  
  
**What is dynamic partitioning and when is it used?**

Dynamic partitioning values for partition columns are known in the runtime. In other words, it is known during loading of the data into a Hive table.

Usage:

While we Load data from an existing non-partitioned table, in order to improve the sampling. Thus it decreases the query latency.

Also, while we do not know all the values of the partitions beforehand. Thus, finding these partition values manually from a huge dataset is a tedious task.

**Why do we need buckets?**

Basically, for [performing bucketing](https://data-flair.training/blogs/bucketing-in-hive/) to a partition there are two main reasons:

A map side join requires the data belonging to a unique join key to be present in the same partition.It allows us to decrease the query time. Also, makes the sampling process more efficient.

**16. How Hive distributes the rows into buckets?**

By using the formula: hash\_function (bucketing\_column) modulo (num\_of\_buckets) Hive determines the bucket number for a row. Basically, hash\_function depends on the column data type. Although, hash\_function for integer data type will be:  
hash\_function (int\_type\_column)= value of int\_type\_column

**17.  What is indexing and why do we need it?**

[Hive index](https://data-flair.training/blogs/hive-view-hive-index/) is a Hive query optimization techniques. Basically, we use it to speed up the access of a column or set of columns in a Hive database. Since, the database system does not need to read all rows in the table to find the data with the use of the index, especially that one has selected.

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

Basically, to share data structures with external systems we use Hcatalog. It offers access to hive metastore to users of other tools on Hadoop. Hence, they can read and write data to hive’s data warehouse.

**19. Where is table data stored in Apache Hive by default?**

hdfs: //namenode\_server/user/hive/warehouse

**20. Are multi-line comments supported in Hive?**

No

**21. What is ObjectInspector functionality?**

To analyze the structure of individual columns and the internal structure of the row objects we use ObjectInspector. Basically, it provides access to complex objects which can be stored in multiple formats in Hive.

**22. Explain about the different types of join in Hive.**

There are  4 different types of [joins in HiveQL](https://data-flair.training/blogs/hive-join/) –

JOIN-  It is very similar to Outer Join in SQL

FULL OUTER JOIN – This join Combines the records of both the left and right outer tables. Basically, that fulfill the join condition.

LEFT OUTER JOIN- Through this Join, All the rows from the left table are returned even if there are no matches in the right table.

RIGHT OUTER JOIN – Here also, all the rows from the right table are returned even if there are no matches in the left table.

**23. How can you configure remote metastore mode in Hive?**Basically, hive-site.xml file has to be configured with the below property, to configure metastore in Hive –  
hive.metastore.uris  
  thrift: //node1 (or IP Address):9083  
  IP address and port of the metastore host

**24. Is it possible to change the default location of Managed Tables in Hive, if so how?**

Yes, by using the LOCATION keyword while creating the managed table, we can change the default location of Managed tables. But the one condition is, the user has to specify the storage path of the managed table as the value of the LOCATION keyword.

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

Basically, the user need not LOAD DATA that moves the files to the /user/hive/warehouse/. But only if data is already present in HDFS. Hence, using the keyword external that creates the table definition in the hive metastore  the user just has to define the table.  
Create external table table\_name (  
 id int,  
 myfields string  
)  
location ‘/my/location/in/hdfs’;

**26. What are the different components of a Hive architecture?**

There are several components of [Hive Architecture](https://data-flair.training/blogs/apache-hive-architecture/). Such as –

User Interface – Basically, it calls the execute interface to the driver. Further, driver creates a session handle to the query. Then sends the query to the compiler to generate an execution plan for it.

Metastore – It is used to Send the metadata to the compiler. Basically, for the execution of the query on receiving the send MetaData request.

Compiler- It generates the execution plan. Especially, that is a DAG of stages where each stage is either a metadata operation, a map or reduce job or an operation on HDFS.

Execute Engine- Basically,  by managing the dependencies for submitting each of these stages to the relevant components we use Execute engine.

**27. Wherever (Different Directory) I run the hive query, it creates 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 already exist or not before creating the metastore. Hence, in 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. Hence, to change the behavior change the location to the absolute path, thus metastore will be used from that location.

**28. Is it possible to use the same metastore by multiple users, in case of the embedded hive?**

No, we cannot use metastore in sharing mode. It is possible to use it in standalone “real” database. Such as MySQL or PostGresSQL.

**29. Usage of Hive.**

Here, we will look at following [Hive](https://hive.apache.org/) usages.  
– We use Hive for Schema flexibility as well as evolution.  
– Moreover, it is possible to portion and bucket, tables in Apache Hive.  
– Also, we can use JDBC/ODBC drivers, since they are available in Hive.

**30. Features and Limitations of Hive.**

Features of Hive

The best feature is it offers data summarization, query, and analysis in much easier manner.

To process data without actually storing in HDFS, Hive supports external tables.

Moreover, it fits the low-level interface requirement of Hadoop perfectly.

Limitation of Hive

We can not perform real-time queries with Hive. Also, it does not offer row-level updates.

 Moreover,  for interactive data browsing Hive offers acceptable latency.

Also, we can say Hive is not the right choice for online transaction processing.

**1. Compare Pig and Hive**

|  |  |  |
| --- | --- | --- |
| Criteria | Pig | Hive |
| Architecture | Procedural data flow language | SQL type declarative language |
| Application | Programming purposes | Report creation |
| Operational field | Client side | Server side |
| Support for avro files | Yes | No |

**2. What is the definition of Hive? What is the present version of Hive and explain about ACID transactions in Hive?**

Hive is an open source data warehouse system. We can use Hive for analyzing and querying in large data sets of Hadoop files. It’s similar to SQL. The present version of hive is 0.13.1. Hive supports ACID transactions: The full form of ACID is Atomicity, Consistency, Isolation, and Durability. ACID transactions are provided at the row levels, there are Insert, Delete, and Update options so that Hive supports ACID transaction.

Insert

Delete

Update

Want to learn more about Hive? Go through this insightful blog “[What is Hive](https://intellipaat.com/blog/what-is-apache-hive/)?”

**3. Explain what is a Hive variable. What do we use it for?**

Hive variable is basically created in the Hive environment that is referenced by Hive scripting languages. It provides to pass some values to the hive queries when the query starts executing. It uses the source command.

**4. What kind of data warehouse application is suitable for Hive? What are the types of tables in Hive?**

Hive is not considered as a full database. The design rules and regulations of Hadoop and HDFS put restrictions on what Hive can do.Hive is most suitable for data warehouse applications.  
Where :

Analyzing the relatively static data.

Less Responsive time.

No rapid changes in data.Hive doesn’t provide fundamental features required for OLTP, Online Transaction Processing.Hive is suitable for data warehouse applications in large data sets.Two types of tables in Hive

Managed table.

External table.

Get a better understanding of Hive by going through this [Hive Tutorial](https://intellipaat.com/tutorial/hadoop-tutorial/apache-hive/) now.

**5. Can We Change settings within Hive Session? If Yes, How?**

Yes we can change the settings within Hive session, using the SET command. It helps to change Hive job settings for an exact query.  
Example: The following commands shows buckets are occupied according to the table definition.

hive> SET hive.enforce.bucketing=true;

We can see the current value of any property by using SET with the property name. SET will list all the properties with their values set by Hive.

hive> SET hive.enforce.bucketing;

hive.enforce.bucketing=true

And this list will not include defaults of Hadoop. So we should use the below like

SET -v

It will list all the properties including the Hadoop defaults in the system.

Interested in learning Hive? Well, we have the comprehensive [Hive Training Course](https://intellipaat.com/big-data-hadoop-training/#curriculum) to give you a head start in your career.

**6. Is it possible to add 100 nodes when we have 100 nodes already in Hive? How?**

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

Take a new system create a new username and password.

Install the SSH and with master node setup ssh connections.

Add ssh public\_rsa id key to the authorized keys file.

Add the new data node host name, IP address and other details in /etc/hosts slaves file  
192.168.1.102 slave3.in slave3.

Start the Data Node on New Node.

Login to the new node like suhadoop or ssh -X hadoop@192.168.1.103.

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

Check the output of jps command on a new node

**7. Explain the concatenation function in Hive with an example .**

Concatenate function will join the input strings.We can specify the  
‘N’ number of strings separated by a comma.  
Example:

CONCAT ('Intellipaat','-','is','-','a','-','eLearning',’-’,’provider’);

Output:

Intellipaat-is-a-eLearning-provider

So, every time we set the limits of the strings by ‘-‘. If it is common for every strings, then Hive provides another command

CONCAT\_WS. In this case,we have to specify the set limits of operator first.

CONCAT\_WS ('-',’Intellipaat’,’is’,’a’,’eLearning’,‘provider’);

Output: Intellipaat-is-a-eLearning-provider.

**8. Trim and Reverse function in Hive with examples.**

Trim function will delete the spaces associated with a string.  
Example:

TRIM(‘ INTELLIPAAT ‘);

Output:

INTELLIPAAT

To remove the Leading space

LTRIM(‘ INTELLIPAAT’);

To remove the trailing space

RTRIM(‘INTELLIPAAT ‘);

In Reverse function, characters are reversed in the string.

Example:

REVERSE(‘INTELLIPAAT’);

Output:

TAAPILLETNI

**9. How to change the column data type in Hive? Explain RLIKE in Hive.**

We can change the column data type by using ALTER and CHANGE.  
The syntax is :

ALTER TABLE table\_name CHANGE column\_namecolumn\_namenew\_datatype;

Example: If we want to change the data type of the salary column from integer to bigint in the employee table.  
ALTER TABLE employee CHANGE salary salary BIGINT;RLIKE: Its full form is Right-Like and it is a special function in the Hive. It helps to examine the two substrings. i.e, if the substring of A matches with B then it evaluates to true.  
Example:

‘Intellipaat’ RLIKE ‘tell’ ◊ True

‘Intellipaat’ RLIKE ‘^I.\*’ ◊ True (this is a regular expression)

**10. What are the components used in Hive query processor?**

The components of a Hive query processor include

* Logical Plan of Generation.
* Physical Plan of Generation.
* Execution Engine.
* Operators.
* UDF’s and UDAF’s.
* Optimizer.
* Parser.
* Semantic Analyzer.
* Type Checking

**11. What is Buckets in Hive?**

The present data is partitioned and divided into different Buckets. This data is divided on the basis of Hash of the particular table columns.

**12. Explain process to access sub directories recursively in Hive queries.**

By using below commands we can access sub directories recursively in Hive

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

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

Hive tables can be pointed to the higher level directory and this is suitable for the directory structure which is like /data/country/state/city/

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

Header records in log files  
System=….  
Version=…  
Sub-version=….  
In the above three lines of headers that we do not want to include in our Hive query. To skip header lines from our tables in the Hive,set a table property that will allow us to skip the header lines.

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”);

**14. What is the maximum size of string data type supported by hive? Mention the Hive support binary formats.**

The maximum size of string data type supported by hive is 2 GB.  
Hive supports the text file format by default and it supports the binary format Sequence files, ORC files, Avro Data files, Parquet files.  
Sequence files: Splittable, compressible and row oriented are the general binary format.  
ORC files: Full form of ORC is optimized row columnar format files. It is a Record columnar file and column oriented storage file. It divides the table in row split. In each split stores that value of the first row in the first column and followed sub subsequently.  
AVRO data files: It is same as a sequence file splittable, compressible and row oriented, but except the support of schema evolution and multilingual binding support.

**15. What is the precedence order of HIVE configuration?**

We are using a precedence hierarchy for setting the properties

SET Command in HIVE

The command line –hiveconf option

Hive-site.XML

Hive-default.xml

Hadoop-site.xml

Hadoop-default.xml

**16. If you run a select \* query in Hive, Why does it not run MapReduce?**

The hive.fetch.task.conversion property of Hive lowers the latency of mapreduce overhead and in effect when executing queries like SELECT, FILTER, LIMIT, etc., it skips mapreduce function

**17. How Hive can improve performance with ORC format tables?**

We can store the hive data in highly efficient manner in the Optimized Row Columnar file format. It can simplify many Hive file format limitations. We can improve the performance by using ORC files while reading, writing and processing the data.

Set hive.compute.query.using.stats-true;

Set hive.stats.dbclass-fs;

CREATE TABLE orc\_table (

idint,

name string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\:’

LINES TERMINATED BY ‘\n’

STORES AS ORC;

Need a reason to learn Apache Hadoop and Hive? Well, go through this blog post to find out why [Hadoop is the new black](https://intellipaat.com/blog/hadoop-is-the-new-black/" \t "_blank).

**18. Explain the functionality of Object-Inspector.**

It helps to analyze the internal structure of row object and individual structure of columns in HIVE. It also provides a uniform way to access complex objects that can be stored in multiple formats in the memory.  
Instance of Java class  
A standard Java object  
A lazily initialized object  
The Object-Inspector tells structure of the object and also ways to access the internal fields inside the object.

**19. Whenever we run hive query, new metastore\_db is created. Why?**

Local metastore is created when we run Hive in embedded mode. And before creating it checks whether the metastore exists or not and this metastore property is defined in the configuration file hive-site.xml. Property is“javax.jdo.option.ConnectionURL” with default value “jdbc:derby:;databaseName=metastore\_db;create=true”.So to change the behavior of the location to an absolute path, so that from that location meta-store will be used.

**20. Differentiate between Hive and HBase**

|  |  |
| --- | --- |
| Hive | HBase |
| Enables most of the SQL queries | This doesn’t allow SQL queries |
| Doesn’t support record level insert, update, and delete operations on table | It supports |
| It is a data warehouse framework | It is NoSQL database |
| Hive run on the top of MapReduce | HBase runs on the top of HDFS |

**21. How can we access the sub directories recursively?**

By using below commands we can access sub directories recursively in Hive

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

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

Hive tables can be pointed to the higher level directory and this is suitable for the directory structure which is like /data/country/state/city/

**22. What are the uses of explode Hive?**

Hadoop developers consider the array as their inputs and convert them into a separate table row. To convert complicate data types into desired table formats Hive is essentially using explode.

**23. What is available mechanism for connecting from applications, when we run hive as a server?**

Thrift Client: Using thrift you can call hive commands from various programming languages. Example: C++, PHP,Java, Python and Ruby.

JDBC Driver: JDBC Driver supports the Type 4 (pure Java) JDBC Driver

ODBC Driver: ODBC Driver supports the ODBC protocol.

**24. How do we write our own custom SerDe?**

End users want to read their own data format instead of writing, so the user wants to write a Deserializer than SerDe.  
Example: The RegexDeserializer will deserialize the data using the configuration parameter ‘regex’, and a list of column names.  
If our SerDe supports DDL, we probably want to implement a protocol based on DynamicSerDe. It’s non-trivial to write a “thrift DDL” parser.

**25. Mention the date data type in Hive. Name the Hive data type collection.**

The TIMESTAMP data type stores date in java.sql.timestamp format.

Three collection data types in Hive

ARRAY

MAP

STRUCT

**26. Can we run UNIX shell commands from Hive? Can Hive queries be executed from script files? How? Give an example.**

Yes, we can run UNIX shell commands from Hive using the! Mark before the command .For example: !pwd at hive prompt will list the current directory.  
We can execute Hive queries from the script files by using the source command.  
Example −

Hive> source /path/to/file/file\_with\_query.hql

LOAD DATA LOCAL INPATH ‘${env:HOME}/country/state/’

OVERWRITE INTO TABLE address;

How do you specify the table creator name when creating a table in Hive?

1**) What is the difference between Pig and Hive ?**

|  |  |  |
| --- | --- | --- |
| Criteria | Pig | Hive |
| Type of Data | Apache Pig is usually used for semi structured data. | Used for Structured Data |
| Schema | Schema is optional. | Hive requires a well-defined Schema. |
| Language | It is a procedural data flow language. | Follows SQL Dialect and is a declarative language. |
| Purpose | Mainly used for programming. | It is mainly used for reporting. |
| General Usage | Usually used on the client side of the hadoop cluster. | Usually used on the server side of the hadoop cluster. |
| Coding Style | Verbose | More like SQL |
| Pig vs Hive | | |

For a detailed answer on the difference between Pig and Hive, refer this link -

https://www.dezyre.com/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79

**2) What is the difference between HBase and Hive ?**

|  |  |
| --- | --- |
| HBase | Hive |
| HBase does not allow execution of SQL queries. | Hive allows execution of most SQL queries. |
| HBase runs on top of HDFS. | Hive runs on top of Hadoop MapReduce. |
| HBase is a NoSQL database. | Hive is a datawarehouse framework. |
| Supports record level insert, updated and delete operations. | Does not support record level insert, update and delete. |
| Hive vs HBase | |

**2) I do not need the index created in the first question anymore. How can I delete the above index named index\_bonuspay?**

DROP INDEX index\_bonuspay ON employee;

Test Your Practical Hadoop Knowledge

**3) Can you list few commonly used Hive services?**

Command Line Interface (cli)

Hive Web Interface (hwi)

HiveServer (hiveserver)

Printing the contents of an RC file using the tool rcfilecat.

Jar

Metastore

For the complete list of big data companies and their salaries- [CLICK HERE](https://docs.google.com/forms/d/1LFuWEKQKCLR231qR9WE5PZakJj77fTDIW6ox5328HFM/viewform)

**4) Suppose that I want to monitor all the open and aborted transactions in the system along with the transaction id and the transaction state. Can this be achieved using Apache Hive?**

Hive 0.13.0 and above version support SHOW TRANSACTIONS command that helps administrators monitor various hive transactions.

FREE eBook on 250 Hadoop Interview Questions and Answers

**5) What is the use of Hcatalog?**

Hcatalog can be used to share data structures with external systems. Hcatalog provides access to hive metastore to users of other tools on Hadoop so that they can read and write data to hive’s data warehouse.

**6) Write a query to rename a table Student to Student\_New.**

Alter Table Student RENAME to Student\_New

**7) Where is table data stored in Apache Hive by default?**

hdfs: //namenode\_server/user/hive/warehouse

**8) Explain the difference between partitioning and bucketing.**

Partitioning and Bucketing of tables is done to improve the query performance. Partitioning helps execute queries faster, only if the partitioning scheme has some common range filtering i.e. either by timestamp ranges, by location, etc. Bucketing does not work by default.

Partitioning helps eliminate data when used in WHERE clause. Bucketing helps organize data inside the partition into multiple files so that same set of data will always be written in the same bucket. Bucketing helps in joining various columns.

In partitioning technique, a partition is created for every unique value of the column and there could be a situation where several tiny partitions may have to be created. However, with bucketing, one can limit it to a specific number and the data can then be decomposed in those buckets.

Basically, a bucket is a file in Hive whereas partition is a directory.

**9) Explain about the different types of partitioning in Hive?**

Partitioning in Hive helps prune the data when executing the queries to speed up processing. Partitions are created when data is inserted into the table. In static partitions, the name of the partition is hardcoded into the insert statement whereas in a dynamic partition, Hive automatically identifies the partition based on the value of the partition field.

Based on how data is loaded into the table, requirements for data and the format in which data is produced at source- static or dynamic partition can be chosen. In dynamic partitions the complete data in the file is read and is partitioned through a MapReduce job based into the tables based on a particular field in the file. Dynamic partitions are usually helpful during ETL flows in the data pipeline.

When loading data from huge files, static partitions are preferred over dynamic partitions as they save time in loading data. The partition is added to the table and then the file is moved into the static partition. The partition column value can be obtained from the file name without having to read the complete file.

**10) When executing Hive queries in different directories, why is metastore\_db created in all** places from where Hive is launched?

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. The property javax.jdo.option.ConnectionURL defined in the hive-site.xml has a default value jdbc: derby: databaseName=metastore\_db; create=true.

The value implies that embedded derby will be used as the Hive metastore and the location of the metastore is metastore\_db which will be created only if it does not exist already. The location metastore\_db is a relative location so when you run queries from different directories it gets created at all places from wherever you launch hive. This property can be altered in the hive-site.xml file to an absolute path so that it can be used from that particular location instead of creating multiple metastore\_db subdirectory multiple times.

**11) How will you read and write HDFS files in Hive?**

i) TextInputFormat- This class is used to read data in plain text file format.

ii) HiveIgnoreKeyTextOutputFormat- This class is used to write data in plain text file format.

iii) SequenceFileInputFormat- This class is used to read data in hadoop SequenceFile format.

iv) SequenceFileOutputFormat- This class is used to write data in hadoop SequenceFile format.

**12) What are the components of a Hive query processor?**

Query processor in Apache Hive converts the SQL to a graph of MapReduce jobs with the execution time framework so that the jobs can be executed in the order of dependencies. The various components of a query processor are-

* Parser
* Semantic Analyser
* Type Checking
* Logical Plan Generation
* Optimizer
* Physical Plan Generation
* Execution Engine
* Operators
* UDF’s and UDAF’s.

**13) Differentiate between describe and describe extended.**

Describe database/schema- This query displays the name of the database, the root location on the file system and comments if any.

Describe extended database/schema- Gives the details of the database or schema in a detailed manner.

**14) Is it possible to overwrite Hadoop MapReduce configuration in Hive?**

Yes, hadoop MapReduce configuration can be overwritten by changing the hive conf settings file.

**15) I want to see the present working directory in UNIX from hive. Is it possible to run this command from hive?**

Hive allows execution of UNIX commands with the use of exclamatory (!) symbol. Just use the ! Symbol before the command to be executed at the hive prompt. To see the present working directory in UNIX from hive run !pwd at the hive prompt.

**16)  What is the use of explode in Hive?**

Explode in Hive is used to convert complex data types into desired table formats. explode UDTF basically emits all the elements in an array into multiple rows.

**17) Explain about SORT BY, ORDER BY, DISTRIBUTE BY and CLUSTER BY in Hive.**

SORT BY – Data is ordered at each of ‘N’ reducers where the reducers can have overlapping range of data.

ORDER BY- This is similar to the ORDER BY in SQL where total ordering of data takes place by passing it to a single reducer.

DISTRUBUTE BY – It is used to distribute the rows among the reducers. Rows that have the same distribute by columns will go to the same reducer.

CLUSTER BY- It is a combination of DISTRIBUTE BY and SORT BY where each of the N reducers gets non overlapping range of data which is then sorted by those ranges at the respective reducers.

**18) Difference between HBase and Hive.**

HBase is a NoSQL database whereas Hive is a data warehouse framework to process Hadoop jobs.

HBase runs on top of HDFS whereas Hive runs on top of Hadoop MapReduce.

**19) Write a hive query to view all the databases whose name begins with “db”**

SHOW DATABASES LIKE ‘db.\*’

**20) How can you prevent a large job from running for a long time?**

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

The strict mode ensures that the queries on partitioned tables cannot execute without defining a WHERE clause.

**21) What is a Hive Metastore?**

Hive Metastore is a central repository that stores metadata in external database.

**22) Are multiline comments supported in Hive?**

No

**23) What is ObjectInspector functionality?**

ObjectInspector is used to analyse the structure of individual columns and the internal structure of the row objects. ObjectInspector in Hive provides access to complex objects which can be stored in multiple formats.

**24) Explain about the different types of join in Hive.**

HiveQL has 4 different types of joins –

JOIN- Similar to Outer Join in SQL

FULL OUTER JOIN – Combines the records of both the left and right outer tables that fulfil the join condition.

LEFT OUTER JOIN- All the rows from the left table are returned even if there are no matches in the right table.

RIGHT OUTER JOIN-All the rows from the right table are returned even if there are no matches in the left table.

**25) How can you configure remote metastore mode in Hive?**

To configure metastore in Hive, hive-site.xml file has to be configured with the below property –

 hive.metastore.uris

   thrift: //node1 (or IP Address):9083

   IP address and port of the metastore host

**26) Is it possible to change the default location of Managed Tables in Hive, if so how?**

Yes, we can change the default location of Managed tables using the LOCATION keyword while creating the managed table. The user has to specify the storage path of the managed table as the value to the LOCATION keyword.

**27) How data transfer happens from HDFS to Hive?**

If data is already present in HDFS then the user need not LOAD DATA that moves the files to the /user/hive/warehouse/. So the user just has to define the table using the keyword external that creates the table definition in the hive metastore.

Create external table table\_name (

  id int,

  myfields string

)

location '/my/location/in/hdfs';

**28) In case of embedded Hive, can the same metastore be used by multiple users?**

We cannot use metastore in sharing mode. It is suggested to use standalone real database like PostGreSQL and MySQL.

**29)  The partition of hive table has been modified to point to a new directory location. Do I have to move the data to the new location or the data will be moved automatically to the new location?**

Changing the point of partition will not move the data to the new location. It has to be moved manually to the new location from the old one.

**30)  What will be the output of cast (‘XYZ’ as INT)?**

It will return a NULL value.

[Master Hadoop by working on interesting Hadoop Hive Real-Time Projects](https://www.dezyre.com/projects/big-data-projects/apache-hive-projects%20www.dezyre.com/projects/big-data-projects/apache-hadoop-projects?utm_source=DeZyreBlog&utm_medium=TextLink&utm_campaign=Blog_HiveInte246)

**31) What are the different components of a Hive architecture?**

Hive Architecture consists of a –

User Interface – UI component of the Hive architecture calls the execute interface to the driver.

Driver create a session handle to the query and sends the query to the compiler to generate an execution plan for it.

Metastore - Sends the metadata to the compiler for the execution of the query on receiving the sendMetaData request.

Compiler- Compiler generates the execution plan which is a DAG of stages where each stage is either a metadata operation, a map or reduce job or an operation on HDFS.

Execute Engine- Execution engine is responsible for submitting each of these stages to the relevant components by managing the dependencies between the various stages in the execution plan generated by the compiler.

**32) What happens on executing the below query? After executing the below query, if you modify   the column –how will the changes be tracked?**

Hive> CREATE INDEX index\_bonuspay ON TABLE employee (bonus)

AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler';

The query creates an index named index\_bonuspay which points to the bonus column in the employee table. Whenever the value of bonus is modified it will be stored using an index value.

**33) What is the default database provided by Hive for Metastore ?**

Derby is the default database.

**34) Is it possible to compress json in Hive external table ?**

Yes, you need to gzip your files and put them as is (\*.gz) into the table location.

**How will you optimize Hive performance?**

There are various ways to run Hive queries faster -

Using Apache Tez execution engine

Using vectorization

Using ORCFILE

Do cost based query optimization.

**Will the reducer work or not if you use “Limit 1” in any HiveQL query?**

**Why you should choose Hive instead of Hadoop MapReduce?**

**I create a table which contains transaction details of customers for the year 2018.   
CREATE TABLE transaction\_details (cust\_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;  
I have inserted 60K tuples in this table and now want to know the total revenue that has been generated for each month. However, Hive takes too much time to process this query. List all the steps that you would follow to solve this problem.**

**There is a  Python application that connects to Hive database for extracting data, creating sub tables for data processing, drops temporary tables, etc. 90% of the processing is done through hive queries which are generated from python code and are sent to hive server for execution.Assume that there are 100K rows , would it be faster to fetch 100K rows to python itself into a list of tuples and mimic the join or filter operations hive performs and avoid the executuon of 20-50 queries run against hive or you should look into hive query optimization techniques ? Which one is performance efficient ?**

**1) Explain what is Hive?**

Hive is an ETL and Data warehousing tool developed on top of Hadoop Distributed File System (HDFS). It is a data warehouse framework for querying and analysis of data that is stored in HDFS. Hive is an open-source-software that lets programmers analyze large data sets on Hadoop.

**2) When to use Hive?**

Hive is useful when making data warehouse applications

When you are dealing with static data instead of dynamic data

When application is on high latency (high response time)

When a large data set is maintained

When we are using queries instead of scripting

**3) Mention what are the different modes of Hive?**

Depending on the size of data nodes in Hadoop, Hive can operate in two modes.

These modes are,

Local mode

Map reduce mode

**4) Mention when to use Map reduce mode?**

Map reduce mode is used when, It will perform on large amount of data sets and query going to execute in a parallel way Hadoop has multiple data nodes, and data is distributed across different node we use Hive in this mode Processing large data sets with better performance needs to be achieved

**5) Mention key components of Hive Architecture?**

Key components of Hive Architecture includes,

User Interface,Compiler, Metastore, Driver, Execute Engine

**6) Mention what are the different types of tables available in Hive?**

There are two types of tables available in Hive.

Managed table: In managed table, both the data and schema are under control of Hive

External table: In the external table, only the schema is under the control of Hive.

**7) Explain what is Metastore in Hive?**

Metastore is a central repository in Hive.  It is used for storing schema information or metadata in the external database.

**8) Mention what Hive is composed of ?**

Hive consists of 3 main parts,

Hive Clients

Hive Services

Hive Storage and Computing

**9) Mention what are the type of database does Hive support ?**

For single user metadata storage, Hive uses derby database and for multiple user Metadata or shared Metadata case Hive uses MYSQL.

**10) Mention Hive default read and write classes?**

Hive default read and write classes are

TextInputFormat/HiveIgnoreKeyTextOutputFormat

SequenceFileInputFormat/SequenceFileOutputFormat

**11) Mention what are the different modes of Hive?**

Different modes of Hive depends on the size of data nodes in Hadoop.

These modes are,

Local mode

Map reduce mode

**12) Why is Hive not suitable for OLTP systems?**

Hive is not suitable for OLTP systems because it does not provide insert and update function at the row level.

**13) Mention what is the difference between Hbase and Hive?**

Difference between Hbase and Hive is,

Hive enables most of the SQL queries, but HBase does not allow SQL queries

Hive does not support record level insert, update, and delete operations on table

Hive is a data warehouse framework whereas HBase is NoSQL database

Hive run on the top of MapReduce, HBase runs on the top of HDFS

**14) Explain what is a Hive variable? What for we use it?**

Hive variable is created in the Hive environment that can be referenced by Hive scripts. It is used to pass some values to the hive queries when the query starts executing.

**15) Mention what is ObjectInspector functionality in Hive?**

ObjectInspector functionality in Hive is used to analyze the internal structure of the columns, rows, and complex objects.  It allows to access the internal fields inside the objects.

**16) Mention what is (HS2) HiveServer2?**

It is a server interface that performs following functions.

It allows remote clients to execute queries against Hive

Retrieve the results of mentioned queries

Some advanced features Based on Thrift RPC in its latest version include

Multi-client concurrency

Authentication

**17) Mention what Hive query processor does?**

Hive query processor convert graph of MapReduce jobs with the execution time framework.  So that the jobs can be executed in the order of dependencies.

**18) Mention what are the components of a Hive query processor?**

The components of a Hive query processor include,

Logical Plan Generation

Physical Plan Generation

Execution Engine

Operators

UDF’s and UDAF’s

Optimizer

Parser

Semantic Analyzer

Type Checking

**19) Mention what is Partitions in Hive?**

Hive organizes tables into partitions.

1. It is one of the ways of dividing tables into different parts based on partition keys.
2. Partition is helpful when the table has one or more Partition keys.
3. Partition keys are basic elements for determining how the data is stored in the table.

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

In Hive you can choose internal table,

1) If the processing data available in local file system

2) If we want Hive to manage the complete lifecycle of data including the deletion

You can choose External table,

1) If processing data available in HDFS

2) Useful when the files are being used outside of Hive

**21) Mention if we can name view same as the name of a Hive table?**

No. The name of a view must be unique compared to all other tables and as views present in the same database.

**22) Mention what are views in Hive?**

In Hive, Views are Similar to tables. They are generated based on the requirements.

1) We can save any result set data as a view in Hive

2) Usage is similar to as views used in SQL

3) All type of DML operations can be performed on a view

**23) Explain how Hive Deserialize and serialize the data?**

Usually, while read/write the data, the user first communicate with inputformat. Then it connects with Record reader to read/write record.  To serialize the data, the data goes to row. Here deserialized custom serde use object inspector to deserialize the data in fields.

**24) What is Buckets in Hive?**

* The data present in the partitions can be divided further into Buckets
* The division is performed based on Hash of particular columns that is selected in the table.

**25) In Hive, how can you enable buckets?**

In Hive, you can enable buckets by using the following command,

set.hive.enforce.bucketing=true;

**26) In Hive, can you overwrite Hadoop MapReduce configuration in Hive?**

Yes, you can overwrite Hadoop MapReduce configuration in Hive.

**27) Explain how can you change a column data type in Hive?**

You can change a column data type in Hive by using command,

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

**28) Mention what is the difference between order by and sort by in Hive?**

* SORT BY will sort the data within each reducer. You can use any number of reducers for SORT BY operation.
* ORDER BY will sort all of the data together, which has to pass through one reducer. Thus, ORDER BY in hive uses a single

**29) Explain when to use explode in Hive?**

Hadoop developers sometimes take an array as input and convert into a separate table row. To convert complex data types into desired table formats, Hive use explode.

**30) Mention how can you stop a partition form being queried?**

You can stop a partition form being queried by using the ENABLE OFFLINE clause with ALTER TABLE statement.

**Q1 What is Hive ?**

Hive is a data warehouse software which is used for facilitates querying and managing large data sets residing in distributed storage.Hive language almost look like SQL language called HiveQL.Hive also allows traditional map reduce programs to customize mappers and reducers when it is inconvenient or inefficient to execute the logic in HiveQL (User Defined Functions UDFS).

**Q2 How Facebook Uses Hadoop,Hive and Hbase ?**

Facebook data stored on HDFS,everyday millions of photos uploaded into facebook with the help of Hadoop Facebook Messages,Likes and statues updates running on top of Hbase Hive to generate reports for third-party developers and advertisers who need to track the success of their applications or campaigns.

**Q3 What is the difference between HBase and Hive?**

Both hive and hbase can be used in different technologies that are based on Hadoop. Hive happens to be a infrastructure warehouse of data which is used on Hadoop whereas HBase is NoSQL. The key value stores which run on Hadoop themselves. Hive will also help those who know about SQL run a few jobs in MapReduce when Hbase will also support 4 of the operations such as put, get, scan and delete. The Hbase happens to be good for querying for data but Hive on the other hand is good for querying data is analytical and is collected over a while.

**Q5 What is Hive Metastore ?**

Hive Meta store is a database that stores metadata of your hive tables like table name,column name,data types,table location,number of buckets in the table etc.

**Q6 Hive new version supported Hadoop Versions ?**

This release works with Hadoop 0.20.x, 0.23.x.y, 1.x.y, 2.x.y

**Q7 Which companies are mostly using Hive ?**

Facebook,Netflix

**Q8 Wherever (Different Directory) I run hive query, it creates new metastore\_db, please explain the reason for it?**

Whenever you run the hive in embedded mode, it creates the local metastore. And before creating the metastore it looks whether metastore already exist or not. This property is defined in configuration file hive – site.xml. Property is “javax.jdo.option.ConnectionURL” with default value “jdbc:derby:;databaseName=metastore\_db;create=true”. So to change the behavior change the location to absolute path, so metastore will be used from that location.

**Q9 Is it possible to use same metastore by multiple users, in case of embedded hive?**

No, it is not possible to use metastore in sharing mode. It is recommended to use standalone “real” database like MySQL or PostGresSQL.

**Q10 What is the functionality of Query Processor in Apached Hive ?**

This component 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.

**Q11 Is multi line comment supported in HIVE Script?**

NO

**Q12 What is the functionality of Query Processor in Apache Hive?**

This components implements the processing framework for converting SQL to graph of map/reduce jobs and the execution time framework to run those jobs in the order od dependencies.

**Q13 what is a Hive Metastore?**

Hive Metastore is a central repository that stores metadata in external database.

**Q14 Explain about the SMB Join in Hive.**

In SMB join in Hive, each mapper reads a bucket from the first table and the corresponding bucket from the second table and then a merge sort join is performed. Sort Merge Bucket (SMB) join in hive is mainly used as there is no limit on file or partition or table join. SMB join can best be used when the tables are large. In SMB join the columns are bucketed and sorted using the join columns. All tables should have the same number of buckets in SMB join.

**Q15 What is ObjectInspector functionality?**

ObjectInspector is used to analyze the structure of individual columns and the internal structure of the row objects. ObjectInspector in Hive provides access to complex objects which can be stored in multiple formats.

**Q16 Is it possible to use same metastore by multiple users, in case of embedded hive?**

No, it is not possible to use metastore in sharing mode. It is recomended to use standalone “real” database like MySQL or PostGreSQL.

**Q17 Explain about the different types of join in Hive.**

HiveQL has 4 different types of joins – JOIN- Similar to Outer Join in SQL

1. FULL OUTER JOIN – Combines the records of both the left and right outer tables that fulfil the join condition.
2. LEFT OUTER JOIN- All the rows from the left table are returned even if there are no matches in the right table.
3. RIGHT OUTER JOIN-All the rows from the right table are returned even if there are no matches in the left table.

**Q18 Is it possible to change the default location of Managed Tables in Hive, if so how?**

Yes, we can change the default location of Managed tables using the LOCATION keyword while creating the managed table. The user has to specify the storage path of the managed table as the value to the LOCATION keyword.

**Q19 How can you connect an application, if you run Hive as a server?**

When running Hive as a server, the application can be connected in one of the 3 ways-

* ODBC Driver-This supports the ODBC protocol
* JDBC Driver- This supports the JDBC protocol
* Thrift Client- This client can be used to make calls to all hive commands using different programming language like PHP, Python, Java, C++ and Ruby.

**Q20 Which classes are used by the Hive to Read and Write HDFS Files**

Following classes are used by Hive to read and write HDFS files:

* TextInputFormat/HiveIgnoreKeyTextOutputFormat: These 2 classes read/write data in plain text file format.
* SequenceFileInputFormat/SequenceFileOutputFormat: These 2 classes read/write data in hadoop SequenceFile format.

**Q22 Is it possible to create multiple table in hive for same data?**

As hive creates schema and append on top of an existing data file. One can have multiple schema for one data file, schema will be saved in hive’s metastore and data will not be parsed or serialized to disk in given schema. When we will try to retrieve data, schema will be used. For example if we have 5 column (name, job, dob, id, salary) in the data file present in hive metastore then, we can have multiple schema by choosing any number of columns from the above list. (Table with 3 columns or 5 columns or 6 columns).

**Q23 What kind of datawarehouse application is suitable for Hive?**

Hive is not a full database. The design constraints and limitations of Hadoop and HDFS impose limits on what Hive can do. Hive is most suited for data warehouse applications, where

1) Relatively static data is analyzed,

2) Fast response times are not required, and

3) When the data is not changing rapidly.

Hive doesn’t provide crucial features required for OLTP, Online Transaction Processing. It’s closer to being an OLAP tool, Online Analytic Processing.So, Hive is best suited for data warehouse applications, where a large data set is maintained and mined for insights, reports, etc.

**Q24 What is the maximum size of string data type supported by Hive?**

Maximum size is 2 GB.

**Q25 What are the Binary Storage formats supported in Hive**?

By default Hive supports text file format, however hive also supports below binary formats.

Sequence Files, Avro Data files, RCFiles, ORC files, Parquet files

* Sequence files: General binary format. splittable, compressible and row oriented. a typical example can be. if we have lots of small file, we may use sequence file as a container, where file name can be a key and content could stored as value. it support compression which enables huge gain in performance.
* Avro datafiles: Same as Sequence file splittable, compressible and row oriented except support of schema evolution and multilingual binding support.
* RCFiles: Record columnar file, it’s a column oriented storage file. it breaks table in row split. in each split stores that value of first row in first column and followed sub subsequently.
* ORC Files: Optimized Record Columnar files.

**Q26 CONCAT function in Hive with Example?**

CONCAT function will concat the input strings. You can specify any number of strings separated by comma.

**Q27 is HQL case sensitive?**

HQL is not case sensitive.

**Q28 REPEAT function in Hive with example?**

REPEAT function will repeat the input string n times specified in the command.

**Q29 Describe REVERSE function in Hive with example?**

REVERSE function will reverse the characters in a string.

**Q30 LOWER or LCASE function in Hive with example?**

LOWER or LCASE function will convert the input string to lower case characters.

**Q31 UPPER or UCASE function in Hive with example?**

UPPER or UCASE function will convert the input string to upper case characters.

**Q33 Rename a table in Hive – How to do it?**

Using ALTER command, we can rename a table in Hive.

ALTER TABLE hive\_table\_name RENAME  TO new\_name;

**Q34 Difference between order by and sort by in hive?**

SORT BY will sort the data within each reducer. You can use any number of reducers for SORT BY operation.

ORDER BY will sort all of the data together, which has to pass through one reducer. Thus, ORDER BY in hive uses single reducer.

ORDER BY guarantees total order in the output while SORT BY only guarantees ordering of the rows within a reducer. If there is more than one reducer, SORT BY may give partially ordered final results.

**Q35 RLIKE in Hive?**

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.

**Q36 Difference between external table and internal table in HIVE ?**

Hive has a relational database on the master node it uses to keep track of state. For instance, when you CREATE TABLE FOO(foo string) LOCATION ‘hdfs://tmp/’;, this table schema is stored in the database. If you have a partitioned table, the partitions are stored in the database(this allows hive to use lists of partitions without going to the filesystem and finding them, etc). These sorts of things are the ‘metadata’.

When you drop an internal table, it drops the data, and it also drops the metadata. When you drop an external table, it only drops the meta data. That means hive is ignorant of that data now. It does not touch the data itself.

**Q37 Does Hive support record level Insert, delete or update?**

Hive does not provide record-level update, insert, or delete. Henceforth, Hive does not provide transactions too. However, users can go with CASE statements and built in functions of Hive to satisfy the above DML operations. Thus, a complex update query in a RDBMS may need many lines of code in Hive.

**Q38 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.

**Q39 What kind of datawarehouse application is suitable for Hive?**

Hive is not a full database. The design constraints and limitations of Hadoop and HDFS impose limits on what Hive can do.

Hive is most suited for data warehouse applications, where

 Relatively static data is analyzed,

 Fast response times are not required, and

 When the data is not changing rapidly.

Hive doesn’t provide crucial features required for OLTP, Online Transaction Processing. It’s closer to being an OLAP tool, Online Analytic Processing.So, Hive is best suited for data warehouse applications, where a large data set is maintained and mined for insights, reports, etc.

**Q40 Can we change the data type of a column in a hive table?**

Using REPLACE column option

ALTER TABLE table\_name REPLACE COLUMNS ……

**Q41 TRIM function in Hive with example?**

TRIM function will remove the spaces associated with a string. xample:

TRIM(‘  Hadoop  ‘);

Output: Hadoop.

**Q42 Why do we need Hive?**

Hive is a tool in Hadoop ecosystem which provides an interface to organize and query data in a databse like fashion and write SQL like queries. It is suitable for accessing and analyzing data in Hadoop using SQL syntax.

**Q43 Is there a date data type in Hive?**

Yes. The TIMESTAMP data types stores date in java.sql.timestamp format

**Q44 What is a Hive variable? What for we use it?**

The hive variable is variable created in the Hive environment that can be referenced by Hive scripts. It is used to pass some values to the hive queries when the query starts executing.

**Q45 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 ?**

By Omitting the LOCAL CLAUSE in the LOAD DATA statement.

**Q46 What does the “USE” command in hive do?**

With the use command you fix the database on which all the subsequent hive queries will run.

**Q47 How can you delete the DBPROPERTY in Hive?**

There is no way you can delete the DBPROPERTY.

**Q48 Does the archiving of Hive tables give any space saving in HDFS?**

No. It only reduces the number of files which becomes easier for namenode to manage.

**Q49 What is the usefulness of the DISTRIBUTED BY clause in Hive?**

It controls ho wthe map output is reduced among the reducers. It is useful in case of streaming data.

**50. Can a partition be archived? What are the advantages and Disadvantages?**

Yes. A partition can be archived. Advantage is it decreases the number of files stored in namenode and the archived file can be queried using hive. The disadvantage is it will cause less efficient query and does not offer any space savings.

**1. What is Hive?**

Hive is a data warehousing tool. It is an abstraction and it gives SQL queries to perform an analysis. It gives you logical abstraction over the databases and the tables but it is not a database.

**2. What is Hive a metastore?**

Hive contains two things: data and the metadata. The metadata contains the (column names, partitions information, bucketing information, SerDe etc.) i.e., the data about the actual table this is by default stored in the Derby database, we can also configure it to Oracle or MySQL database.

**3. What is the limitation of Derby database for Hive metastore?**

With derby database, you cannot have multiple connections or multiple sessions instantiated at the same time. Derby database runs in the local mode and it creates a log file so that multiple users cannot access Hive simultaneously.

**4. What are managed and external tables?**

We have got two things, one of which is data present in the HDFS and the other is the metadata, present in some database.

There are two categories of Hive tables i.e., Managed and External Tables.

In the Managed tables, both the data and the metadata are managed by Hive and if you drop the managed table, both data and metadata are deleted.

There are some situations where your data will be controlled by some other application and you want to read that data but you must allow Hive to delete that data.

In such case, you can create an external table in Hive. In the external table, metadata is controlled by Hive but the actual data will be controlled by some other application. So, when you delete a table accidentally, only the metadata will be lost and the actual data will reside wherever it is.

**5. What are the complex data types in Hive?**

**Map** – The Map contains a key-value pair where you can search for a value using the key.

**Struct** – A Struct is a collection of elements of different data types. For example, if you take the address, it can have different data types. For example, pin code will be in Integer format.

**Array** – An Array will have a collection of homogeneous elements. For example, if you take your skillset, you can have N number of skills

**Uniontype** – It represents a column which can have a value that can belong to any of the data types of your choice.

**6. How does partitioning help in the faster execution of queries?**

With the help of partitioning, a subdirectory will be created with the name of the partitioned column and when you perform a query using the WHERE clause, only the particular sub-directory will be scanned instead of scanning the whole table. This gives you faster execution of queries.

**7. How to enable dynamic partitioning in Hive?**

Related to partitioning there are two types of partitioning Static and Dynamic. In the static partitioning, you will specify the partition column while loading the data.

Whereas in dynamic partitioning, you push the data into Hive and then Hive decides which value should go into which partition. To enable dynamic partitioning, you have set the below property  
set hive.exec.dynamic.parition.mode = nonstrict;

Example,  
insert overwrite table emp\_details\_partitioned  
partition(location)  
select \* from emp\_details;

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

If you have to join two large tables, you can go for reduce side join. But if both the tables have the same number of buckets or same multiples of buckets and also sorted on the same column there is a possibility of SMBMJ in which all the joins take place in the map phase itself by matching the corresponding buckets.

Buckets are basically files that are created inside the HDFS directory.

There are different properties which you need to set for bucket map joins and they are as follows:

**set hive.enforce.sortmergebucketmapjoin = false;**

**set hive.auto.convert.sortmerge.join =  false;**

**set hive.optimize.bucketmapjoin =  true;**

**set hive.optimize.bucketmapjoin.sortedmerge = true;**

**9. How to enable bucketing in Hive?**

By default bucketing is disabled in Hive, you can enforce to enable it by setting the below property  
set hive.enforce.bucketing  = true;

**10. Which method has to be overridden when we use custom UDF in Hive?**

Whenever you write a custom UDF in Hive, you have to extend the UDF class and you have to override the evaluate() function.

**11. What are the different file formats in Hive?**

There are different file formats supported by Hive

Text File format, Sequence File format, RC file format, Parquet, Avro, ORC.

Every file format has its own characteristics and Hive allows you to choose easily the file format which you wanted to use.

**12. How is SerDe different from File format in Hive?**

SerDe stands for Serializer and Deserializer. It determines how to encode and decode the field values or the column values from a record that is: how you serialize and deserialize the values of a column,But file format determines how records are stored in key value format or how do you retrieve the records from the table.

**13. What is RegexSerDe?**

Regex stands for a regular expression. Whenever you want to have a kind of pattern matching, based on the pattern matching, you have to store the fields. RegexSerDe is present in **org.apache.hadoop.hive.contrib.serde2.RegexSerDe.**

In the SerDeproperties, you have to define your input pattern and output fields. For example, you have to get the column values from line xyz/pq@def if you want to take xyz, pq and def separately.

To extract the pattern, you can use:  
‘input.regex’ = ‘(.\*)/(.\*)@(.\*)’  
To specify how to store them, you can use  
‘output.format.string’ = ‘%1$s%2$s%3$s’;

**14. How is ORC file format optimised for data storage and analysis?**

ORC stores collections of rows in one file and within the collection the row data will be stored in a columnar format. With columnar format, it is very easy to compress, thus reducing a lot of storage cost.While querying also, it queries the particular column instead of querying the whole row as the records are stored in columnar format.ORC has got indexing on every block based on the statistics min, max, sum, count on columns so when you query, it will skip the blocks based on the indexing.

**15. How to access HBase tables from Hive?**

Using Hive-HBase storage handler, you can access the HBase tables from Hive and once you are connected, you can query HBase using the SQL queries from Hive. You can also join multiple tables in HBase from Hive and retrieve the result.

Along with the above Hive interview questions, also check our [Hbase tutorial](https://acadgild.com/blog/hbase-tutorial-beginners-guide) . These 2 articles would be helpful to attend a Hadoop interview. Enroll for [Big Data and Hadoop Training](https://acadgild.com/big-data/big-data-development-training-certification) with ACADGILD and become a successful Hadoop developer.

**List out the different components of Hive architecture?**

There are five core components in Hive architecture are listed below:  
•User Interface (UI): It acts as a communicator between users and drivers when the user writes the queries the UI accepts it and runs it on the driver, there are two types of interface available they are Command line and GUI interface.  
•Driver: It maintains the life cycle of the HiveQL query. It receives the queries from the user interface and creates the session to process the query.  
•Compiler: It receives the query plans from the driver and gets the required information from Metastore in order to execute the plan.  
•Metastore: It stores the information about the data like a table; it can be of an internal or external table. It sends the metadata information to the compiler to execute the query.  
•Execute Engine: Hive service will execute the result in execution engine; it executes the query in MapReduce to process the data. It is responsible for controlling each stage for all these components.

**2. Which are the different types of modes that Hive can operate?**  
This is the common Hive Interview Questions asked in an interview. Hive can operate on two modes based on the size of data,  
These modes are:  
•Map reduce Mode  
•Local Mode

**3. Which are the scenarios where Hive can be used and cannot be used?**

When you’re creating Data warehouse applications when your data is [Static](https://www.educba.com/course/static-and-dynamic-routing/) when your application does not need high response time, when the data volume is huge, when the data is not changing rapidly and when you are using queries instead of scripting. Hive supports only OLAP transaction it is not suitable for OLTP transactions.

**4. What are the file formats that Hive supports? List the type of applications that are supported by HIVE?**By default, Hive supports Text File format and it also supports the binary file format such as Sequence file, ORC files, Parquet files, Avro Data files.  
•Sequence file: It is generally a binary format file, which can be compressed and is splittable.  
•ORC file: Optimized Row Columnar file is recorded column-based file and column-oriented storage file.  
•Parquet file: It is a column-oriented binary file it is highly efficient for large-scale queries.  
•Avro Data file: It is same as sequence file format which is a splittable, compressible and row-oriented file.The maximum size of string data type allowed in Hive is 2 GB.

Hive is a data warehouse framework that is suitable for those applications that are written in [Java](https://www.educba.com/course/java-8-tutorials/), [C++](https://www.educba.com/c-programming-language-basics/), [PHP](https://www.educba.com/course/php-mysql-for-beginners/), [Python](https://www.educba.com/course/advanced-python-iot-iot-based-data-analysis/) or [Ruby](https://www.educba.com/course/ruby-programming-training/).

**6. What is a Metastore in Hive? List and explain the different types of Hive Metastores configuration?**Metastore in Hive is used to store the metadata information, it is a central repository in Hive. It allows storing the metadata information in an external database. By default, Hive stores Metadata information in Derby database but it can also be stored in other databases such as [Oracle](https://www.educba.com/course/oracle-application-express-online-training/), [MySql](https://www.educba.com/course/mysql-training/)etc.  
There are three types of Metastore configuration, they are:  
•Embedded metastore: It is a default mode; it can locally access the Hive library, all the command line operations are done in embedded mode. The Hive service, the metastore service, and the database run in same JVM.  
•Local metastore: It stores data in an external database such as MySql or Oracle. The Hive service and metastore service runs in same JVM, it connects to the database that is running in separate JVM.  
•Remote metastore: It uses the remote mode to run queries, here the metastore service and hive service runs in a separate JVM. You can have multiple metastore servers to increase the availability.

**7. What is a Hive Query Processor? What are the different components of the Hive Query Processor?**

This is the frequently asked Hive Interview Questions in an interview. Hive Query Processor is used to convert SQL to MapReduce jobs. Based on the order of dependencies the jobs are executed.  
The components of Hive Query Processor are listed below:  
•Semantic Analyser •UDF’s and UDAF’s •Optimizer •Operator •Parser •Execution Engine •Type Checking •Logical Plan Generation •Physical Plan Generation

**8. What is the functionality of Object-Inspector in Hive?**  
It is composed of Hive that is used to identify the structure of the individual columns and internal structure of row objects. The complex objects that are stored in multiple formats can be accessed using Object-Inspector in Hive.  
Object-Inspector will identify the structure of an object and ways to access the internal fields inside the object.

**9. What are the different ways to connect the applications to Hive Server?**

There are three ways to connect the applications to the Hive server, they are:  
1) Thrift Client: This is used to run all the hive commands using a different [programming language](https://www.educba.com/course/using-programming-languages-ubuntu/) such as [Java](https://www.educba.com/course/java/), [C++](https://www.educba.com/c-programming-language-basics/), [PHP](https://www.educba.com/course/php-database-basic-tutorials/), [Python](https://www.educba.com/course/web-application-development-pyramid-micro-framework-python/) or [Ruby](https://www.educba.com/course/ruby-rails-training/).  
2) ODBC Driver: This will support the ODBC protocol  
3) JDBC Driver: This will support the JDBC protocol

**10. What is the default read and write classes in Hive?**  
Below is the read and write classes available in Hive:

1) TextInputFormat – This class is used to read data in plain text format.  
2) HiveIgnoreKeyTextOutputFormat – This class is used to write data in plain text format.  
3) SequenceFileInputFormat – This class is used to read data in Hadoop Sequence file format.  
4) SequenceFileOutputFormat – This class is used to write data in Hadoop Sequence file format.

**What is Hive?**

It’s an open source project under the Apache Software Foundation, it’s a data warehouse software ecosystem in Hadoop. Which manage vast amount of structured data sets, by using HQl language; it’s similar to SQL.

**Where hive is the best suitable?**

When you are doing data warehouse applications,

Where you are getting static data instead of dynamic data,

when the application on high latency (response time high).

where a large data set is maintained and mined for insights, reports.

When we are using queries instead of scripting we use Hive.

**When hive is not suitable?**

It doesn’t provide OLTP transactions supports only OLAP transactions.

If application required OLTP, switch to NoSQL databases.

HQL queries have higher latency, due to the mapreduce.

**Hive Support Acid Transactions?**

By default it doesn’t support record-level update, insert and delete, but recent Hive 1.4 later versions supporting insert, update and delete operations. So hive support ACID transactions. To achieve updates & deletion transactions in 1.4 version, you must change given default values.

**hive.support.concurrency – true**

**hive.enforce.bucketing – true**

**hive.exec.dynamic.partition.mode – nonstrict**

**hive.txn.manager – org.apache.hadoop.hive.ql.lockmgr.DbTxnManager**

**hive.compactor.initiator.on – true (for exactly one instance of the Thrift metastore service)**

**hive.compactor.worker.threads – a positive number on at least one instance of the Thrift metastore service**

**What is Hive MetaStore?**

MetaStore is a central repository of Hive, that allows to store meta data in external database. By default Hive

store meta data in Derby database, but you can store in MySql, Oracle depends on project.

**Why I choose Hive instead of MapReduce?**

There are Partitions to simplify the data process, Bucketing for sampling the data, sort the data quickly, and

simplify the mapreduce process. Partitions and Buckets can segmenting large data sets to improve Query performance in Hive. So It is highly recommendable for structure data.

**Can I access Hive without Hadoop?**

Hive store and process the data on the top of Hadoop, but it’s possible to run in Other data storage systems like Amazon S3, GPFS (IBM) and MapR file systems.

**What is the relationship between MapReduce and Hive?** or **How Mapreduce jobs submits on the cluster?**

Hive provides no additional capabilities to MapReduce. The programs are executed as MapReduce jobs via the

interpreter. The Interpreter runs on a client machine which rurns HiveQL queries into MapReduce jobs.

Framework submits those jobs onto the cluster.

**If you run select \* query in Hive, why it’s not run Mpareduce?**

It’s an optimization technique. hive.fetch.task.conversion property can (FETCH task) minimize latency of

mapreduce overhead. When queried SELECT, FILTER, LIMIT queries, this property skip mapreduce and using

FETCH task. As a result Hive can execute query without run mapreduce task.

By default it’s value “minimal”. Which optimize: SELECT STAR, FILTER on partition columns, LIMIT queries

only, where as another value is “more” which optimize : SELECT, FILTER, LIMIT only (+TABLESAMPLE,

virtual columns).

**How Hive can improve performance with ORC format tables?**

Hive can store the data in highly efficient manner in the Optimized Row Columnar (ORC) file format. It can

ease many Hive file format limitations. Using ORC files can improves the performance when reading, writing,

and processing data. Enable this format by run this command and create table like this.

set hive.compute.query.using.stats=true;

set hive.stats.dbclass=fs;

CREATE TABLE orc\_table (

id int,

name string

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\;’

LINES TERMINATED BY ‘\n’

STORED AS ORC;

**What is the importance of Vectorization in Hive?**

It’s a query optimization technique. Instead of processing multiple rows, Vectorization allows to process process

a batch of rows as a unit. Consequently it can optimize query performance. The file must be stored in ORC

format to enable this Vectorization. It’s disabled by default, but enable this property by run this command.

set hive.vectorized.execution.enabled=true;

**Difference between sort by or order by clause in Hive? Which is the fast?**

ORDER BY – sort the data in one reducer. Sort by much faster than order by.

SORT BY – sort the data within each reducer. You can use n number of reducers for sort.

In the first case (order by) maps sends each value to the single reducer and count them all.

In the second case (sort by) maps splits up the values to many reducers and each reduce generates its list and

finds the count. So it can sort quickly.

**Example:**

SELECT name, id, cell FROM user\_table ORDER BY id, name;

SELECT name, id, cell FROM user\_table DISTRIBUTE BY id SORT BY name;

**Wherever you run hive query, first it creates new metastore\_db, why? What is the importance of Metastore\_db?**

When we run the hive query, first it creates a local metastore, before creates the metastore first Hive checks

whether metastore is already exists or not? If presents shows error, else the process goes on. This configuration

is set in hive-site.xml like this.

<property>

<name>javax.jdo.option.ConnectionURL</name>

<value>jdbc:derby:;databaseName=metastore\_db;create=true</value>

<description>JDBC connect string for a JDBC metastore</description>

</property>

**Tell me different Hive metastore configuration.**

There are three types of metastores configuration called

1) Embedded metastore

2) Local metastore

3) Remote metastore.

If Hive run any query first it enter into embedded mode, It’s default mode. In Command line all operations done

in embedded mode only, it can access Hive libraries locally. In the embedded metastore configuration, hive

driver, metastore interface and databases use same JVM. It’s good for development and testing.

In **local metastore** the metastore store data in external databases like MYSQL. Here Hive driver and metastore

run in the same JVM, but remotely communicate with external Database. For better protection required

credentials in Local metastore.

Where as in **Remote server**, use remote mode to run the queries over Thift server.

In Remote metastore, Hive driver and metastore interface would be running in a different JVM. So for better

protection, required credentials such are isolated from Hive users.

3/22/2019 Hive Interview Question Answers For Freshers & Experienced

https://www.bigdataanalyst.in/hive-interview-questions/ 4/14

**Hive can process any type of data formats?**

Yes, Hive uses the SerDe interface for IO operations. Different SerDe interfaces can read and write any type of

data. If normal directly process the data where as different type of data is in the Hadoop, Hive use different

SerDe interface to process such data.

**Example:**

MetadataTypedColumnsetSerDe: used to read/write CSV format data.

JsonSerDe: process Json data.

RejexSerDe: process weblog data.

AvroSerde: Avro format data.

**What Is the HWI?**

The Hive Web Interface is an alternative to the command line interface. HWI is a simple graphical interface, It’s

hive web interface. The HWI allows start at database level directly. you can get all SerDe, column names and

types and simplifies the hive steps. It’s seccession based interface, so you can run multiple hive queries

simultaneously. There is no local metastore mode in HWI.

**What is the difference between Like and Rlike operators in HIVE?**

Like: used to find the substrings within a main string with regular expression %.

Rlike is a special fuction which also finds the sub strings within a main string, but return true or false without

using regular expression.

**Example:** Tablename is table, column is name.

name=VenuKatragadda, venkatesh, venkateswarlu

Select \* from table where name like “venu%. //VenuKatragadda.

select \* from table where name rlike “venk%”. // false, true, true.

**What are the Hive default read and write classes?**

Hive use 2+2 classes to read and write the files.

1)TextInputFormat/HiveIgnoreKeyTextOutputFormat.

2) SequenceFileInputFormat/SequenceFileOutputFormat:

First class used to read/write the plain text. Second class used for sequence files.

**What is Query processor in Hive?**

It’s a core processing unit in Hive framework, it converting SQL to map/reduce jobs and run in the other

dependencies. As a result hive can convert the Hive queries into Hive queries.

**What are Views in Hive?**

Based on user requirement create and manage view. You can set data as view. It’s a logical construct. It’s used

where query is more complicated and to hide complexity of query and make easy to the users.

**Example:**

Create view table\_name as select \* from employee where salary>10000;

3/22/2019 Hive Interview Question Answers For Freshers & Experienced

https://www.bigdataanalyst.in/hive-interview-questions/ 5/14

**What is different between database and data-warehouse?**

Typically database is designed for OLTP transactional operations. Where as Data-warehouse is implemented for

OLAP (analysis) operations.

OLTP can constrained to a single application. OLAP resists as a layer on the top of several databases.

OlTP process current, streaming and dynamic data where as OLAP process Retired, historic and static data only.

Database completely has normalization concept. DWH is De-normalization concept.

Hive Interview Question and Answers

**What is the different between Internal and external tables in Hive?**

Hive will create a database on the master node to store meta data to keep data in safe. Let example, If you

partition table, table schema stores data in the external table.

In **Managed table**, Schema stored in the local system, but in External table MetaStore separate from the node

and stored in a secure database. In Internal Table, Hive reads and loads entire file as it is to process, but in

External simply loads depends on the query logic.

If user drop the table, Hive drop original data and MetaStore, but in External table, just drop MetaStore, but not

original data. Hive by default store in internal table, but it’s not recommendable. Store the data in external table.

**How to write single and multiple line commands in Hive?**

To write single line commands we use –followed by commands.

Hive doesn’t supports multiple comments now.

**What is Thrift server & client, JDBC and ODBC driver importance in Hive?**

**Thrift** is a cross language RPC framework which generate code and combines a software stack finally execute the Thrift code in remote server. Thrift compiler acts as interpreter between server and client. Thrift server allows a remove client to submit request to Hive, using different programming languages like Python, Ruby and scala.

**JDBC driver**: A JDBC driveris a software component enabling a Java application to interact with a database.

**ODBC driver:** ODBC accomplishes DBMS independence by using an *ODBC driver* as a translation layer between the application and the DBMS

**Does Hive support 100% SQL Quries like Insert, Delete and Updates?**

Hive doesn’t support Updates in record level. To update, It integrate with Hbase.

**When you are use Hive?**

When the data is structured data, Static data, Low density is not a problem, If the data processed based on the queries, Hive is the best option. Most often data warehouse data processed in the Hive.

**What is the use of partition in hive?**

To analyze a particular set of data, not required to load entire data, desired data partition is a good approach. To achieve this goal, Hive allows to partition the data based on particular column. Static partition and Dynamic partition, both can optimize the Hive performance. For Instant, required a particular year information, partition based on year.

**Is is mandatory Schema in Hive?**

yes, It’s mandatory to create a table in Database. Hive is schema oriented modal. It store schema information in external database.

**How Hive Serialize and DeSerialize the data?**

In Hive language, SerDe also called Serialization and DeSerialization. Usually when read/write the data, user first communicate with inputformat, then it connect with Record reader to read/write record.The data is stored in Serialized (binary) format in Record. To serialize the data dat goes to row, here deserialized custem serde use object inspector to deserialize the data in fields. now user see the data in the fields, that deliver to the end user.

**How Hive use Java in SerDe?**

To insert data into table, Hive create an object by using Java. To transfer java objects over network, the data

should be serialized. Each field serialized by using Object inspector and finally serialized data stored in Hive

table.

**Does Hive Support Insert, delete, or updation?**

As of now, Hive doesn’t support record level updadation, insert and deletion queries. HQL is subset of SQL, but

not equalto SQL. To update Hive integrate with Hbase.

**Tell me few function names in Hive**

CONTACT(‘Venu’-‘Bigdata’-‘analyst’); // Venu-Bigdata-analyst

CONTACT\_WS(‘-‘, ‘venu’, ‘bigdata’, ‘analyst’); //venu-bigdata-analyst

REPEAT(‘venu’,3);

TRIM(‘ VENU ‘); //VENU (without spaces)

LTRim(‘ venu ‘); //venu (trim leftside, but not rightside)

RTRIM(‘ venu ‘); // venu(trim rightside only, but not leftside)

REVERSE(‘venu’); //unev

LOWER(‘Venu’); //venu

LCASE “”

UPPER OR UCASE(‘Venu’); //VENU

RLIKE .. return T/F for sub string.

‘Venu’ RLIKE ‘en’ //True

‘Venu’ RLIKE ‘^V.\*’ //T

**Difference between order by and sort by in hive?**

SORT BY -use number of reducers, so it can process quickly.

ORDER BY – use single reducer. If data is too large, it’s take a long time to sort the data.

**Difference between Internal and External Table?**

External table: Schema is stored in Database. Actual data stored in Hive tables. If data lost in External table, it

lost only metastore, but not actual data.

Internal table: MetaStore and actual data both stored in local system. If any situation, data lost, both actual data

and meta store will be lost.

**What is the difference between Hive and Hbase?**

Hive allows most of the SQL queries, but Hbase not allows SQL queries directly.

Hive doesn’t support record level update, insert, and deletion operations on table, but Hbase can do it.

Hive is a Data warehouse framework where as Hbase is a NoSQL database.

Hive run on the top of Mapreduce, Hbase run on the top of HDFS.

**How many ways you can run Hive?**

In CLI mode (By using command line inerface).

By using JDBC or ODBC.

By Called Hive Thift client. It allows java, PHP, Python, Ruby and C++ to write commands to run in Hive.

**Can you explain different type of SerDe?**

By default Hive used Lazy Serde also allows Jeson Serde and most often used RegexSerde to be Serialized and DeSerialized Data.

**Why we are using buckets in Hive?**

To process many chunks of files, to analyze vast amount of data, sometime burst the process and time.Bucketing is a sampling concept to analyze the data, by using hashing algorithm. set

hive.enforce.bucketing=true; can enable the process.

**How Hive Organize the data?**

Hive organize in three ways such as Tables, Partitions and Buckets. Tables organize based on Arrays, Maps,

primitive column types. Partitions has one or more partition keys based on project requirements.

Buckets used for analyze the data for sampling purpose. It’s good approach to process a pinch of data in the form

of buckets instead of process all data.

**Can you explain about Hive Architecture?**

There are 5 core components there in Hive such as: UI, Driver, Compiler, Metastore, Execute Engine.

**What is User Interface (UI)?**

UI: This interface is interpreter between users and Driver, which accept queries from User and execute on the

Driver. Now two types of interfaces available in Hive such as command line interface and GUI interface.

3/22/2019 Hive Interview Question Answers For Freshers & Experienced

https://www.bigdataanalyst.in/hive-interview-questions/ 8/14

Hadoop provides Thrift interface and JDBC/ODBC for integrating other applications.

**What is importance of Driver in Hive?**

**Driver:** It manages life cycle of HiveQL queries. Driver receives the queries from User Interface and fetch on the ODBC/JDBC interfaces to process the query. Driver create separate independent section to handle each query.

**Compiler:** Compiler accept plans from Drivers and gets the required metadata from MetaStore, to execute Plan.

**MetaStore:** Hive Store meta data in the table. It means information about data is stored in MetaStore in the form of table, it may be internal or external table. Hive compiler get the meta data information from metastore table.

**Execute Engine:**

Hive Driver execute the output in the execution Engine. Here, execute engine executes the queries in the MapReduce JobTracker. Based on Required information, Hive queries run in the MapReduce to process the data.

**When we are use explode in Hive?**

Sometime Hadoop developer takes array as input and convert into a separate table row. To achieve this goal,

Hive use explode, it acts as interpreter to convert complex data-types into desired table formats.

Syntax:

SELECT explode (arrayName) AS newCol FROM TableName;

SELECT explode(map) AS newCol1, NewCol2 From TableName;

**What is ObjectInspector functionality in Hive?**

Hive uses ObjectInspector to analyze the internal structure of the rows, columns and complex objects.

Additionally gives us ways to access the internal fields inside the object. It not only process common data-types

like int, bigint, STRING, but also process complex data-types like arrays, maps, structs and union.

**Can you overwrite Hadoop Mapreduce configuration in Hive?**

Yes, You can overwrite Hive map, reduce steps in hive conf settings. Hive allows to overwrite Hadoop configuration files.

**How to display the present database name in the terminal?**

There are two ways to know the current database. One temporary in cli and second one is persistently.

1) in CLI just enter this command: **set hive.cli.print.current.db=true;**

2) In hive-site.xml paste this code:

<property>

<name>hive.cli.print.current.db</name>

<value>true</value>

</property>

In second scenario, you can automatically display the Hive database name when you open terminal.

**Is a job split into map?**

No, Hadoop framework can split the data-file, but not Job. This chunks of data stored in blocks. Each split need

a map to process. Where as Job is a configurable unit to control execution of the plan/logic. Job is not a physical

data-set to split, it’s a logical configuration API to process those split.

**What is the difference between Describe and describe extended?**

To see table definition in Hive, use ***describe <table name>;*** *command*

Where as

To see more detailed information about the table, use **describe extended <tablename>;** command

Another important command **describe formatted <tablename>**; also describe all details in a clean manner.

**What is difference between static and dynamic partition of a table?**

To prune data during query, partition can minimize the query time. The partition is created when the data is inserted into table. Static partition can insert individual rows where as Dynamic partition can process entire table based on a particular column. At least one static partition is must to create any (static, dynamic) partition. If you are partitioning a large datasets, doing sort of a ETL flow Dynamic partition recommendable.

**What is the difference between partition and bucketing?**

The main aim of both Partitioning and Bucketing is execute the query more efficiently. When you are creating a table the slices are fixed in the partitioning the table.

Bucketing follows Hash algorithm. Based on number of buckets, randomly the data inserted into the bucket to sampling of the data

**Questions**

What are the different types of tables available in HIve?

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

Can a table be renamed in Hive?

Can we change the data type of a column in a hive table?

What is a metastore in Hive?

What is the need for custom Serde?

Why do we need Hive?

What is the default location where hive stores table data?

What are the three different modes in which hive can be run?

Is there a date data type in Hive?

What are collection data types in Hive?

Can we run unix shell commands from hive?

Give example.What is a Hive variable?

What for we use it?

Can hive queries be executed from script files? How?

What is the importance of .hiverc file?

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

What do you mean by schema on read?

How do you list all databases whose name starts with p?

What does the “USE” command in hive do?

How can you delete the DBPROPERTY in Hive?

What is the significance of the line

set hive.mapred.mode = strict;

How do you check if a particular partition exists?

Which java class handles the Input record encoding into files which store the tables in Hive?

Which java class handles the output record encoding into files which result from Hive queries?

What is the significance of ‘IF EXISTS” clause while dropping a table?

When you point a partition of a hive table to a new directory, what happens to the data?

Write a query to insert a new column(new\_col INT) into a hiev table (htab) at a position before an existing column (x\_col)Does the archiving of Hive tables give any space saving in HDFS?  
How can you stop a partition form being queried?

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 ?

If you omit the OVERWRITE clause while creating a hive table,

what happens to file which are new and files which already exist?

What does the following query do?

INSERT OVERWRITE TABLE employees

PARTITION (country, state)

SELECT ..., se.cnty, se.st

FROM staged\_employees se;

What is a Table generating Function on hive?

How can Hive avoid mapreduce?

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

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

As part of Optimizing the queries in HIve, what should be the order of table size in a join query?

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

How will you convert the string ’51.2’ to a float value in the price column?

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

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

Can we LOAD data into a view?

What types of costs are associated in creating index on hive tables?

Give the command to see the indexes on a table.What is bucketing ?

What does /\*streamtable(table\_name)\*/ do?

Can a partition be archived? What are the advantages and Disadvantages?

What is a generic UDF in hive?

The following statement failed to execute. What can be the cause?

Explain the difference between SQL and Apache Hive.

Why mapreduce will not run if you run select \* from table in hive?