**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/).

**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.

**1 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).

**2 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.

**3 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.

**5 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.

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

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

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

Facebook,Netflix

**8 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.

**9 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.

**10 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.

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

NO

**12 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.

**13 what is a Hive Metastore?**

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

**14 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.

**15 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.

**16 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.

**17 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.

**18 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.

**19 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.

**20 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.

**22 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).

**23 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.

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

Maximum size is 2 GB.

**25 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.

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

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

**27 is HQL case sensitive?**

HQL is not case sensitive.

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

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

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

REVERSE function will reverse the characters in a string.

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

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

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

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

**33 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;

**34 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.

**35 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.

**36 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.

**37 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.

**38 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.

**39 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.

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

Using REPLACE column option

ALTER TABLE table\_name REPLACE COLUMNS ……

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

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

TRIM(‘  Hadoop  ‘);

Output: Hadoop.

**42 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.

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

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

**44 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.

**45 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.

**46 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.

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

There is no way you can delete the DBPROPERTY.

**48 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.

**49 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?

**Hadoop Sqoop Interview Questions and Answers**

**1) Compare Sqoop and Flume**

|  |  |
| --- | --- |
| **Sqoop** | **Flume** |
| Used for importing data from structured data sources like RDBMS. | Used for moving bulk streaming data into HDFS. |
| It has a connector based architecture. | It has a agent based architecture. |
| Data import in sqoop is not event driven. | Data load in flume is event driven |
| HDFS is the destination for importing data. | Data flows into HDFS through one or more channels. |

**2) What is the default file format to import data using Apache Sqoop?**

Sqoop allows data to be imported using two file formats

i) Delimited Text File Format

This is the default file format to import data using Sqoop. This file format can be explicitly specified using the –as-textfile argument to the import command in Sqoop. Passing this as an argument to the command will produce the string based representation of all the records to the output files with the delimited characters between rows and columns.

ii) Sequence File Format

It is a binary file format where records are stored in custom record-specific data types which are shown as Java classes. Sqoop automatically creates these data types and manifests them as java classes.

**3) I have around 300 tables in a database. I want to import all the tables from the database except the tables named Table298, Table 123, and Table299. How can I do this without having to import the tables one by one?**

This can be accomplished using the import-all-tables import command in Sqoop and by specifying the exclude-tables option with it as follows-

sqoop import-all-tables

--connect –username –password --exclude-tables Table298, Table 123, Table 299

**4) Does Apache Sqoop have a default database?**

Yes, MySQL is the default database.

**5) How can I import large objects (BLOB and CLOB objects) in Apache Sqoop?**

Apache Sqoop import command does not support direct import of BLOB and CLOB large objects. To import large objects, I Sqoop, JDBC based imports have to be used without the direct argument to the import utility.

**6) How can you execute a free form SQL query in Sqoop to import the rows in a sequential manner?**

This can be accomplished using the –m 1 option in the Sqoop import command. It will create only one MapReduce task which will then import rows serially.

**7) How will you list all the columns of a table using Apache Sqoop?**

Unlike sqoop-list-tables and sqoop-list-databases, there is no direct command like sqoop-list-columns to list all the columns. The indirect way of achieving this is to retrieve the columns of the desired tables and redirect them to a file which can be viewed manually containing the column names of a particular table.

Sqoop import --m 1 --connect 'jdbc: sqlserver: //nameofmyserver; database=nameofmydatabase; username=DeZyre; password=mypassword' --query "SELECT column\_name, DATA\_TYPE FROM INFORMATION\_SCHEMA.Columns WHERE table\_name='mytableofinterest' AND \$CONDITIONS" --target-dir 'mytableofinterest\_column\_name'

**8) What is the difference between Sqoop and DistCP command in Hadoop?**

Both distCP (Distributed Copy in Hadoop) and Sqoop transfer data in parallel but the only difference is that distCP command can transfer any kind of data from one Hadoop cluster to another whereas Sqoop transfers data between RDBMS and other components in the Hadoop ecosystem like HBase, Hive, HDFS, etc.

**9) What is Sqoop metastore?**

Sqoop metastore is a shared metadata repository for remote users to define and execute saved jobs created using sqoop job defined in the metastore. The sqoop –site.xml should be configured to connect to the metastore.

**10) What is the significance of using –split-by clause for running parallel import tasks in Apache Sqoop?**

--Split-by clause is used to specify the columns of the table that are used to generate splits for data imports. This clause specifies the columns that will be used for splitting when importing the data into the Hadoop cluster. —split-by clause helps achieve improved performance through greater parallelism. Apache Sqoop will create splits based on the values present in the columns specified in the –split-by clause of the import command. If the –split-by clause is not specified, then the primary key of the table is used to create the splits while data import. At times the primary key of the table might not have evenly distributed values between the minimum and maximum range. Under such circumstances –split-by clause can be used to specify some other column that has even distribution of data to create splits so that data import is efficient.

**11) You use –split-by clause but it still does not give optimal performance then how will you improve the performance further.**

Using the –boundary-query clause. Generally, sqoop uses the SQL query select min (), max () from to find out the boundary values for creating splits. However, if this query is not optimal then using the –boundary-query argument any random query can be written to generate two numeric columns.

**12) During sqoop import, you use the clause –m or –numb-mappers to specify the number of mappers as 8 so that it can run eight parallel MapReduce tasks, however, sqoop runs only four parallel MapReduce tasks. Why?**

Hadoop MapReduce cluster is configured to run a maximum of 4 parallel MapReduce tasks and the sqoop import can be configured with number of parallel tasks less than or equal to 4 but not more than 4.

**13) You successfully imported a table using Apache Sqoop to HBase but when you query the table it is found that the number of rows is less than expected. What could be the likely reason?**

If the imported records have rows that contain null values for all the columns, then probably those records might have been dropped off during import because HBase does not allow null values in all the columns of a record.

**14) The incoming value from HDFS for a particular column is NULL. How will you load that row into RDBMS in which the columns are defined as NOT NULL?**

Using the –input-null-string parameter, a default value can be specified so that the row gets inserted with the default value for the column that it has a NULL value in HDFS.

**15) If the source data gets updated every now and then, how will you synchronise the data in HDFS that is imported by Sqoop?**

Data can be synchronised using incremental parameter with data import –

--Incremental parameter can be used with one of the two options-

i) append-If the table is getting updated continuously with new rows and increasing row id values then incremental import with append option should be used where values of some of the columns are checked (columns to be checked are specified using –check-column) and if it discovers any modified value for those columns then only a new row will be inserted.

ii) lastmodified – In this kind of incremental import, the source has a date column which is checked for. Any records that have been updated after the last import based on the lastmodifed column in the source, the values would be updated.

**16) Below command is used to specify the connect string that contains hostname to connect MySQL with local host and database name as test\_db –**

**–connect jdbc: mysql: //localhost/test\_db**

**Is the above command the best way to specify the connect string in case I want to use Apache Sqoop with a distributed hadoop cluster?**

When using Sqoop with a distributed Hadoop cluster the URL should not be specified with localhost in the connect string because the connect string will be applied on all the DataNodes with the Hadoop cluster. So, if the literal name localhost is mentioned instead of the IP address or the complete hostname then each node will connect to a different database on their localhosts. It is always suggested to specify the hostname that can be seen by all remote nodes.

**1. Compare Sqoop and Flume**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Sqoop** | **Flume** |
| Application | Importing data from RDBMS | Moving bulk streaming data into HDFS |
| Architecture | Connector  – connecting to respective data | Agent – fetching of the right data |
| Loading of data | Event driven | Not event driven |

**2. Name a few import control commands. How can Sqoop handle large objects?**

Import control commands are used to import RDBMS data

**Append**: Append data to an existing dataset in HDFS. –append

**Columns**: columns to import from the table. –columns  
<col,col……> • Where: where clause to use during import. —

where The common large objects are Blog and Clob.Suppose the object is less than 16 MB, it is stored inline with the rest of the data. If there are big objects, they are temporarily stored in a subdirectory with the name \_lob. Those data are then materialized in memory for processing. If we set lob limit as ZERO (0) then it is stored in external memory.

**3. How can we import data from particular row or column? What is the destination types allowed in Sqoop import command?**

Sqoop allows to Export and Import the data from the data table based on the where clause. The syntax is

--columns

<col1,col2……> --where

--query

Example:

sqoop import –connect jdbc:mysql://db.one.com/corp --table INTELLIPAAT\_EMP --where “start\_date> ’2016-07-20’ ”

sqoopeval --connect jdbc:mysql://db.test.com/corp --query “SELECT \* FROM intellipaat\_emp LIMIT 20”

sqoop import –connect jdbc:mysql://localhost/database --username root --password aaaaa –columns “name,emp\_id,jobtitle”

**4. Role of JDBC driver in sqoop setup? Is the JDBC driver enough to connect the sqoop to the database?**

Sqoop needs a connector to connect the different relational databases. Almost all Database vendors make a JDBC connector available specific to that Database, Sqoop needs a JDBC driver of the database for interaction.  
No, Sqoop needs JDBC and a connector to connect a database.

**5. Using Sqoop command how can we control the number of mappers?.**

We can control the number of mappers by executing the parameter –num-mapers in sqoop command. The –num-mappers arguments control the number of map tasks, which is the degree of parallelism used. Start with a small number of map tasks, then choose a high number of mappers starting the performance may down on the database side.

Syntax: -m, –num-mappers

**6.How will you update the rows that are already exported? Write sqoop command to show all the databases in MySQL server.**

By using the parameter – update-key we can update existing rows. Comma-separated list of columns is used which uniquely identifies a row. All of these columns are used in the WHERE clause generated UPDATE query. All other table columns will be used in the SET part of the query.  
The command below is used to show all the databases in MySQL server.

$ sqoop list –databases –connect jdbc:mysql://database.test.com/

**7. Define Sqoop metastore? What is the purpose of Sqoop-merge?**

Sqoop meta store is a tool for using hosts in a shared metadata repository. Multiple users and remote users can define and execute saved jobs defined in metastore. End users configured to connect the metastore in sqoop-site.xml or with the

–meta-connect argument.

**The purpose of sqoop-merge is:**  
This tool combines 2 datasets where entries in one dataset overwrite entries of an older dataset preserving only the new version of the records between both the data sets.

**8. Explain the saved job process in Sqoop.**

Sqoop allows us to define saved jobs which make this process simple. A saved job records the configuration information required to execute a Sqoop command at a later time. sqoop-job tool describes how to create and work with saved jobs. Job descriptions are saved to a private repository stored in $HOME/.sqoop/.

We can configure Sqoop to instead use a shared metastore, which makes saved jobs offered to multiple users across a shared cluster. Starting the metastore is covered by the section on the sqoop-metastore tool.

**9. How Sqoop word came ? Sqoop is which type of tool and the main use of sqoop?**

Sqoop word came from SQL+HADOOP=SQOOP. And Sqoop is a data transfer tool.  
The main use of Sqoop is to import and export the large amount of data from RDBMS to HDFS and vice versa.

**10. How to enter into Mysql prompt, and explain the command parameter indicates?**

The command for entering into Mysql prompt is “mysql –u root –p”  
-u indicatesthe user  
Root indicates username  
-p indicates password.

**11. I am getting connection failure exception during connecting to Mysql through Sqoop, what is the root cause and fix for this error scenario?**

This will happen when there is lack of permissions to access our Mysql database over the network. We can try the below command to confirm the connect to Mysql database from aSqoop client machine.  
$ mysql –host=MySqlnode> –database=test –user= –password=  
We can grant the permissions with below commands.

mysql> GRANT ALL PRIVILEGES ON \*.\* TO ‘%’@’localhost’;

mysql> GRANT ALL PRIVILEGES ON \*.\* TO ‘ ’@’localhost’;

**12. I am getting java.lang.IllegalArgumentException: during importing tables from oracle database.what might be the root cause and fix for this error scenario?**

Sqoop commands are case- sensitive of table names and user names.  
By specifying the above two values in UPPER case, it will resolve the issue.  
In case, the source table is created under different user namespace,then table name should be like USERNAME.TABLENAME as shown below  
sqoop import  
–connect jdbc:oracle:thin:@intellipaat.testing.com/INTELLIPAAT  
–username SQOOP  
–password sqoop  
–table COMPANY.EMPLOYEES

**13. How can you list all the columns of a table using Apache sqoop?**

There is no straight way to list all the columns of a table in Apache Sqoop like sqoop-list-columns, so first we should retrieve the columns of the particular table and transform to a file containing the column names of particular table.Syntax is:

Sqoop import –m1 –connect ‘jdbc:sqlserver://servername;database=databasename;

Username-DeZyre;password=mypassword’ –query “SELECT column\_name,DATA\_TYPE FROM INFORMATION\_SCHEMA columns WHEREtable\_name=’mytableofinterest’ AND \$CONDITIONS” –target-dir ‘mytableofinterest\_column\_name’.

**14. How to create a table in Mysql and how to insert the values into the table ?**

To create a table in mysql using the below command

mysql> create table tablename( col1 datatype, col2 datatype,…………);

Example –

mysql> create table INTELLIPAAT(emp\_idint,emp\_namevarchar(30),emp\_salint);

Insert the values into the table

mysql> insert into table name(value1,value2,value3,………);

Example-

mysql> insert into INTELLIPAAT(1234,’aaa’,20000);

mysql> insert into INTELLIPAAT(1235,’bbb’,10000);

mysql> insert into INTELLIPAAT(1236,’ccc’,15000);

**15. What are the basic commands in Hadoop Sqoop and its uses?**

The basic commands of HadoopSqoop are

Codegen, Create-hive-table, Eval, Export, Help, Import, Import-all-tables, List-databases, List-tables,Versions.

* Useof HadoopSqoop basic commands
* Codegen- It helps to generate code to interact with database records.
* Create-hive-table- It helps to Import a table definition into a hive
* Eval- It helps to evaluateSQL statement and display the results
* Export-It helps to export an HDFS directory into a database table
* Help- It helps to list the available commands
* Import- It helps to import a table from a database to HDFS
* Import-all-tables- It helps to import tables from a database to HDFS
* List-databases- It helps to list available databases on a server
* List-tables-It helps to list tables in a database
* Version-It helps to display the version information

**16. Is sqoop same as to distcp in hadoop?**

No. Because the only distcp import command is same as Sqoop import command and both the commands submit parallel map-only jobs but both command functions are different. Distcp is used to copy any type of files from Local filesystem to HDFS and Sqoop is used for transferring the data records between RDBMS and Hadoop eco- system service.

**17. For each sqoop copying into HDFS how many MapReduce jobs and tasks will be submitted?**

There are 4 jobs that will be submitted to each Sqoop copying into HDFS and no reduce tasks are scheduled.

**18. How can Sqoop be used in Java programs?**

In the Java code Sqoop jar is included in the classpath. The required parameters are created to Sqoop programmatically like for CLI (command line interface). Sqoop.runTool() method also invoked in Java code.

**19. I am having around 500 tables in a database. I want to import all the tables from the database except the tables named Table 498, Table 323, and Table 199. How can we do this without having to import the tables one by one?**

This can be proficient using the import-all-tables, import command in Sqoop and by specifying the exclude-tables option with it as follows-  
sqoop import-all-tables  
–connect –username –password –exclude-tables Table498, Table 323, Table 199

**20. Explain the significance of using –split-by clause in Apache Sqoop?**

split-by is a clause, it is used to specify the columns of the table which are helping to generate splits for data imports during importing the data into the Hadoop cluster. This clause specifies the columns and helps to improve the performance via greater parallelism. And also it helps to specify the column that has an even distribution of data to create splits,that data is imported.

**1. Mention the best features of Apache Sqoop.**Apache Sqoop is a tool in Hadoop ecosystem have several advantages. Like

1. Parallel import/export
2. Connectors for all major RDBMS Databases
3. Import results of SQL ry
4. Incremental Load
5. Full Load
6. Kerberos Security Integration
7. Load data directly into [**Hive**](https://data-flair.training/blogs/apache-hive-tutorial/) / [**HBase**](https://data-flair.training/blogs/hadoop-hbase-tutorial/)
8. Compression
9. Support for Accumulo

**2. What is Sqoop Import? Explain its purpose.**

While it comes to import tables from RDBMS to HDFS we use Sqoop Import tool. Generally, we can consider that each row in a table is a record in HDFS. Also, when we talk about text files all records are there as text data. However, when we talk about Avro and sence files all records are there as binary data here. To be more specific,  it imports individual tables from RDBMS to HDFS.

**3. What is the default file format to import data using Apache Sqoop?**

By using two file formats Sqoop allows data import. Such as:

i) Delimited Text File Format

Basically, to import data using Sqoop this is the default file format. Moreover, to the import command in Sqoop, this file format can be explicitly specified using the –as-textfile argument. Likewise, passing this argument will produce the string-based representation of all the records to the output files with the delimited characters between rows and columns.

ii) Sence File Format

We can say, Sence file format is a binary file format. Their records are stored in custom record-specific data types which are shown as [**Java classes**](https://data-flair.training/blogs/class-and-object-in-java/). In addition, Sqoop automatically creates these data types and manifests them as java classes.

**4. How can I import large objects (BLOB and CLOB objects) in Apache Sqoop?**  
However, direct import of BLOB and CLOB large objects is not supported by Apache Sqoop import command. So, in order to import large objects like I Sqoop, JDBC based imports have to be used without the direct argument to the import utility.

**5. How can you execute a free-form SQL ry in Sqoop to import the rows in a sential manner?**By using the –m 1 option in the Sqoop import command we can accomplish it. Basically, it will create only one [**MapReduce**](https://data-flair.training/blogs/hadoop-mapreduce-tutorial/) task which will then import rows serially.

**6. Does Apache Sqoop have a default database?**

Yes, MySQL is the default database.

**7. How will you list all the columns of a table using Apache Sqoop?**Since to list all the columns we do not have any direct command like sqoop-list-columns. So, indirectly we can achieve this is to retrieve the columns of the desired tables and redirect them to a file that can be viewed manually containing the column names of a particular table.  
Sqoop import –m 1 –connect ‘jdbc: sqlserver: //nameofmyserver; database=nameofmydatabase; username=DeZyre; password=mypassword’ –query “SELECT column\_name, DATA\_TYPE FROM INFORMATION\_SCHEMA.Columns WHERE table\_name=’mytableofinterest’ AND \$CONDITIONS” –target-dir ‘mytableofinterest\_column\_name’

**8.  If the source data gets updated every now and then, how will you synchronize the data in HDFS that is imported by Sqoop?**By using incremental parameter with data import we can synchronize the data–  
–However, with one of the two options, we can use incremental parameter-  
i) append  
Basically, we should use incremental import with append option. Even if the table is getting updated continuously with new rows and increasing row id values then. Especially, where values of some of the columns are checked (columns to be checked are specified using –check-column) and if it discovers any modified value for those columns then only a new row will be inserted.  
ii) lastmodified  
However, in this kind of incremental import, the source has a date column which is checked for. Any records that have been updated after the last import based on the lastmodifed column in the source, the values would be updated**.**

**9. Name a few import control commands. How can Sqoop handle large objects?**To import RDBMS data, we use import control commands  
Append: Append data to an existing dataset in HDFS.   
–append  
Columns: columns to import from the table.   
–columns  
<col,col……> •  
Where: where clause to use during import. —  
Where the common large objects are Blog and Clob. Suppose the object is less than 16 MB, it is stored inline with the rest of the data. If there are big objects, they are temporarily stored in a subdirectory with the name \_lob. Those data are then materialized in memory for processing. If we set lob limit as ZERO (0) then it is stored in external memory.

**10. How can we import data from particular row or column? What is the destination types allowed in Sqoop import command?**Basically, on the basis of where clause, Sqoop allows to Export and Import the data from the data table. So, the syntax is  
–columns  
<col1,col2……> –where  
–ry

**For Example:**  
sqoop import –connect jdbc:mysql://db.one.com/corp –table INTELLIPAAT\_EMP –where “start\_date> ’2016-07-20’ ”  
sqoopeval –connect jdbc:mysql://db.test.com/corp –ry “SELECT \* FROM intellipaat\_emp LIMIT 20”  
sqoop import –connect jdbc:mysql://localhost/database –username root –password aaaaa –columns “name,emp\_id,jobtitle”  
However, into following services Sqoop supports data imported:

HDFS, Hive, Hbase, Hcatalog, Accumulo,

**11. When to use –target-dir and when to use –warehouse-dir while importing data?**Basically, we use –target-dir to specify a particular directory in[**HDFS**](https://data-flair.training/blogs/hadoop-hdfs-tutorial/). Whereas we use –warehouse-dir to specify the parent directory of all the sqoop jobs. So, in this case under the parent directory sqoop will create a directory with the same name as the table.

**12. What is the process to perform an incremental data load in Sqoop?**In Sqoop, the process to perform incremental data load is to synchronize the modified or updated data (often referred as delta data) from RDBMS to [**Hadoop**](https://data-flair.training/blogs/hadoop-2-x-vs-hadoop-3-x-comparison/). Moreover, in Sqoop the delta data can be facilitated through the incremental load command.  
In addition, by using Sqoop import command we can perform incremental load. Also, by loading the data into the hive without overwriting it. However, in Sqoop the different attributes that need to be specified during incremental load are  
1) Mode (incremental)   
It shows how Sqoop will determine what the new rows are. Also, it has value as Append or Last Modified.  
2) Col (Check-column)   
Basically, it specifies the column that should be examined to find out the rows to be imported.  
3) Value (last-value)   
It denotes the maximum value of the check column from the previous import operation.

**13. What is the significance of using –compress-codec parameter?**However, we use the –compress -code parameter to get the out file of a sqoop import in formats other than .gz like .bz2.

**14. Can free-form SQL ries be used with Sqoop import command? If yes, then how can they be used?**In Sqoop, we can use SQL ries with the import command. Basically, we should use import command with the –e and – ry options to execute free-form SQL ries. But note that the –target dir value must be specified While using the –e and –ry options with the import command.

**15. What is the importance of eval tool?**  
Basically,[**Sqoop Eval**](https://data-flair.training/blogs/sqoop-eval/) helps to run sample SQL ries against Database as well as preview the results on the console. Moreover, it helps to know what data we can import or that desired data is imported or not.

**16. How can you import only a subset of rows from a table?**  
In the sqoop import statement, by using the WHERE clause we can import only a subset of rows.

**17. What are the limitations of importing RDBMS tables into Hcatalog directly?**  
By making use of –hcatalog –database option with the –hcatalog –table, we can import RDBMS tables into Hcatalog directly. However, there is one limitation to it is that it does not support several arguments like –as-Avro file, -direct, -as-sencefile, -target-dir , -export-dir.

**18. What is the advantage of using –password-file rather than -P option while preventing the display of password in the sqoop import statement?**  
Inside a sqoop script, we can use The –password-file option. Whereas the -P option reads from standard input, preventing automation.

**19. What do you mean by Free Form Import in Sqoop?**  
By using any SQL Sqoop can import data from a relational database ry rather than only using table and column name parameters.

**20. What is the role of JDBC driver in Sqoop?**  
Basically, sqoop needs a connector to connect to different relational databases. Since, as a JDBC driver, every DB vendor makes this connector available which is specific to that DB. Hence, to interact with Sqoop needs the JDBC driver of each of the database it needs.

**21. Is JDBC driver enough to connect sqoop to the databases?**  
No. to connect to a database Sqoop needs both JDBC and connector.

**22. What is InputSplit in Hadoop?**  
Input Split is defined as while a Hadoop job runs, it splits input files into chunks also assign each split to a mapper to process.

**23. What is the work of Export in Hadoop sqoop?**  
Export tool transfer the data from HDFS to RDBMS

**24. Use of Codegen command in Hadoop sqoop?**  
Basically,Codegen command generates code to interact with database records

**25. Use of Help command in Hadoop sqoop?**  
Help command in Hadoop sqoop generally list available commands

**26.  How can you schedule a sqoop job using Oozie?**  
However, Oozie has in-built sqoop actions inside which we can mention the sqoop commands to be executed.

**27. What is the importance of — the split-by clause in running parallel import tasks in sqoop?**  
In Sqoop, it mentions the column name based on whose value the data will be divided into groups of records. Further, by the [**MapReduce**](https://data-flair.training/blogs/how-hadoop-mapreduce-works/)tasks, these group of records will be read in parallel.

**28. What is a sqoop metastore?**  
A tool that Sqoop hosts a shared metadata repository is what we call sqoop metastore. Moreover, multiple users and/or remote users can define and execute saved jobs (created with the sqoop job) defined in this metastore.  
In addition, with the –meta-connect argument Clients must be configured to connect to the metastore in sqoop-site.xml.

**29. What is the purpose of sqoop-merge?**  
The merge tool combines two datasets where entries in one dataset should overwrite entries of an older dataset preserving only the newest version of the records between both the data sets.

**30. How can you see the list of stored jobs in sqoop metastore?**  
sqoop job –list

**31. Which database the sqoop metastore runs on?**  
Basically, on the current machine running sqoop-metastore launches, a shared HSQLDB database instance.

**32. Where can the metastore database be hosted?**  
Anywhere, it means we can host metastore database within or outside of the [**Hadoop cluster**](https://data-flair.training/blogs/install-hadoop-1-x-on-multi-node-cluster/).

**33. Give the sqoop command to see the content of the job named myjob?**  
Sqoop job –show myjob

**34. How can you control the mapping between SQL data types and Java types?**  
we can configure the mapping between by using the –map-column-java property.  
For example:  
$ sqoop import … –map-column-java id = String, value = Integer

**35. Is it possible to add a parameter while running a saved job?**  
Yes, by using the –exec option we can add an argument to a saved job at runtime.  
sqoop job –exec jobname — — newparameter

**36. What is the usefulness of the options file in sqoop.**  
To specify the command line values in a file and use it in the sqoop commands we use the options file in sqoop.  
For example  
The –connect parameter’s value and –user name value scan be stored in a file and used again and again with different sqoop commands.

**37. How can you avoid importing tables one-by-one when importing a large number of tables from a database?**  
Using the command  
sqoop import-all-tables  
–connect  
–usrename  
–password  
–exclude-tables table1,table2 ..  
Basically, this will import all the tables except the ones mentioned in the exclude-tables clause.

**38. How can you control the number of mappers used by the sqoop command?**  
To control the number of mappers executed by a sqoop command we use the parameter –num-mappers. Moreover, we should start with choosing a small number of map tasks and then gradually scale up as choosing high number of mappers initially may slow down the performance on the database side.

**39. What is the default extension of the files produced from a sqoop import using the –compress parameter?**  
 .gz

**40. What is the significance of using –compress-codec parameter?**  
We use the –compress -code parameter to get the out file of a sqoop import in formats other than .gz like .bz2.

**41. What is a disadvantage of using –direct parameter for faster data load by sqoop?**  
The native utilities used by databases to support faster laod do not work for binary data formats like SenceFile.

**42. How will you update the rows that are already exported?**Basically, to update existing rows we can use the parameter –update-key.Moreover, in it, a comma-separated list of columns is used which unily identifies a row. All of these columns are used in the WHERE clause of the generated UPDATE ry. All other table columns will be used in the SET part of the ry.

**43. What are the basic commands in Apache Sqoop and its uses?**  
The basic commands of Apache Sqoop are:  
[**Codegen**](https://data-flair.training/blogs/sqoop-codegen/), Create-hive-table, [**Eval**](https://data-flair.training/blogs/sqoop-eval/), [**Export**](https://data-flair.training/blogs/sqoop-export/), Help,[**Import**](https://data-flair.training/blogs/sqoop-import/), [**Import-all-tables**](https://data-flair.training/blogs/sqoop-import-all-tables/),[**List-databases**](https://data-flair.training/blogs/sqoop-list-databases/), [**List-tables**](https://data-flair.training/blogs/sqoop-list-tables/), Versions.  
Moreover, uses of Apache Sqoop basic commands are:

1. Codegen- It helps to generate code to interact with database records.
2. Create- hive-table- It helps to Import a table definition into a hive
3. Eval- It helps to evaluate SQL statement and display the results
4. Export- It helps to export an HDFS directory into a database table
5. Help- It helps to list the available commands
6. Import- It helps to import a table from a database to HDFS
7. Import-all-tables- It helps to import tables from a database to HDFS
8. List-databases- It helps to list available databases on a server
9. List-tables- It helps to list tables in a database
10. Version- It helps to display the version information

**45. What is Sqoop Validation?**  
It means to validate the data copied. Either import or export by comparing the row counts from the source as well as the target post copy. Likewise, we use this option to compare the row counts between source as well as the target just after data imported into HDFS. Moreover, While during the imports, all the rows are deleted or added, Sqoop tracks this change. Also updates the log file.

**46. What is Purpose to Validate in Sqoop?**  
In Sqoop to validate the data copied isValidation main purpose. Basically, either Sqoop import or Export by comparing the row counts from the source as well as the target post copy.

**47. What is Sqoop Job?**  
To perform an incremental import if a saved job is configured, then state regarding the most recently imported rows is updated in the saved job. Basically, that allows the job to continually import only the newest rows.

**48. What is Sqoop Import Mainframe Tool and its Purpose?**

Basically, a tool which we use to import all sential datasets in a partitioned dataset (PDS) on a mainframe to HDFS is Sqoop Import Mainframe. That tool is what we call import mainframe tool. Also, A PDS is akin to a directory on the open systems. Likewise, in a dataset, the records can only contain character data. Moreover here, records will be stored as a single text field with the entire record.

**49. What is the purpose of Sqoop List Tables?**

Basically, the main purpose of sqoop-list-tables is list tables present in a database. Learn all insights of Sqoop List Tables, follow the link: [Sqoop List Tables – Arguments and Examples](https://data-flair.training/blogs/sqoop-list-tables/)

**50. Difference Between Apache Sqoop vs Flume.**

So, let’s discuss all the differences on the basis of features.

a. Data Flow  
Apache Sqoop – Basically, [Sqoop](http://sqoop.apache.org/) works with any type of relational database system (RDBMS) that has the basic JDBC connectivity. Also, Sqoop can import data from NoSQL databases like MongoDB, Cassandra and along with it. Moreover, it allows data transfer to Apache Hive or HDFS.  
[Apache Flume](https://data-flair.training/blogs/apache-flume-tutorial/)– Likewise, Flume works with streaming data sources those are generated continuously in Hadoop environments. Like log files.

b. Type of Loading  
Apache Sqoop – Basically,  Sqoop load is not driven by events.  
Apache Flume – Here, data loading is completely event-driven.

c. When to use  
Apache Sqoop – However, if the data is being available in Teradata, Oracle, MySQL, PostreSQL or any other JDBC compatible database it is considered an ideal fit.  
Apache Flume – While we move bulk of streaming data from sources likes JMS or spooling directories, it is the best choice.

d. Link to HDFS  
Apache Sqoop – Basically, for importing data in Apache Sqoop, HDFS is the destination  
Apache Flume – In Apache Flume, data generally flow to HDFS through channels

e. Architecture   
Apache Sqoop – Basically, it has connector based architecture. However, that means the connectors know a great deal in connecting with the various data sources. Also to fetch data correspondingly.  
Apache Flume – However, it has agent-based architecture. Basically, it means code written in Flume is we call agent that may responsible for fetching the data.

**1. Shed light on the versatile features of Sqoop**  
It is important to note that Apache Sqoop is also known as a tool in the Hadoop ecosystem which carries with it several benefits. Here is the list of them.Import and export in a parallel manner  
It supports Accumulo Compression of data Full load taking capabilities Incremental load bearing capabilities Security Integration in a proper way  
Can connect a majority of RDBMS databases It can depict results related to the ries of SQL

**2. How can you import large objects like BLOB and CLOB in Sqoop?**  
The direct import function is not supported by Sqoop in case of CLOB and BLOB objects. Hence, if you have to import large purposes, you can use JDBC based imports. This can be done without introducing the direct argument of the import utility.

**3. What is the default database of Apache Sqoop?**  
The default database of Apache Sqoop is MySQL.

**4. Describe the process of executing a free-form SQL ry to import rows**  
To achieve a free-form SQL ry, you have to use the –m1 option. This would create only one Mapreduce task. This would then import the rows directly.

**Related Article:** [Evaluating Performance of Distributed Systems with MapReduce](https://mindmajix.com/mapreduce/evaluating-performance-of-distributed-systems-with-mapreduce)

**5. Describe the importance of using –compress-codec parameter**  
The –compress-codec parameter can be used to get the export file of the Sqoop import in the mentioned formats.

**6. What is the significance of Eval tool?**  
Sqoop Eval would help you to make use of the sample SQL ries. This can be against the database as it can preview the results that are displayed on the console. Interestingly, with the help of the Eval tool, you would be well aware of the fact that the desired data can be imported correctly or not.

**7. What is the meaning of Free form import in Sqoop?**  
With the use of Sqoop, one can import the relational database ry. This can be done using column and table name parameters.

**8. Shed light on the advantage of utilizing –password-file rather than –P option**  
The –password-file option is usually used inside the Sqoop script file. On the other hand, the –P option is able to read the standard input along with the column name parameters.

**9. Is the JDBC driver fully capable to connect Sqoop on the databases?**  
The JDBC driver is not capable to connect Sqoop on the databases. This is the reason that Sqoop requires both the connector and JDBC driver.

**10. What is the meaning of Input Split in Hadoop?**  
Input Split is that kind of a function which is associated with splitting the input files into various chunks. These chunks can also assign each split to a mapper in the ongoing process of data correction.

**11. Illustrate the utility of the Help Command in Sqoop**  
The help command in Sqoop can be utilized to list the various available commands.

**12. Shed light on the service of Codegen command in Sqoop**  
The Codegen command is associated with the generation of code so that it can appropriately interact with the database records.

**13. Describe the procedure involved in executing an incremental data load in Sqoop**  
You should be well aware of the fact that in Sqoop, the process of performing additional data load is to update the uploaded data. This data is often referred to as delta data. In Sqoop, this delta data can be altered with the use of incremental load command. Additionally, it can be said that with the help of Sqoop, the import command can also perform additional load. By loading the data into the hive without overwriting it, its efficiency can be maintained in a significant manner. This is possible only with the help of incremental data load.

It is also essential for you to illustrate the various types of incremental data load. They are as follows:  
Progressive Mode: This variety usually determines the number of new rows. Moreover, it also possesses a value that can best resemble the Append functions.

**Value: This denotes the maximum amount that is derived from the check column from the previous import operation.**

**The Check Column feature:** This function is helpful in specifying the number of columns that should be assessed to determine the number of rows to be imported.

**Related Article:** [An Overview Of Hadoop Hive](https://mindmajix.com/hadoop/an-overview-of-hadoop-hive)

**14. Illustrate on the process of listing all the columns of a table with the help of Apache Sqoop**  
To contain all the columns, you do not have any direct command like the Sqoop indexed columns. However, you can also indirectly achieve this. You can do that by retrieving the columns of the desired tables. After that, you can redirect them to a set of files that can be viewed in a standard manner. This also contains the columns of a particular table.

**15. What is the default file format in order to import data with the utilization of Apache Sqoop?**  
At the time of answering this stion, you should know that there are two file formats that can be used in the case of importing data. These are as follows:

**Senci ng the file format**

It is a commonly observed fact that a sence file format is also known by the name of binary file format. The records of these binary file formats are usually stored in the custom record data types. Moreover, Sqoop can automatically create a varied data types and also manifests them in the form of Java classes.

**Delimiti ng the text file format**

This is the usual file format in importing data. Additionally, it can be said that in order to avail the import command in Sqoop, this file format can be specified. You can specify the file format with the use of text file argument command. On the other hand, when you pass this argument, you would produce a string-based representation of varied types of records. You can also create the output files with the use of delimited characters between columns and rows.

[Check Out Hadoop Tutorials](https://mindmajix.com/hadoop)

**16. List all the basic commands in Apache Sqoop along with their applications**  
The basic controls in Apache Sqoop along with their uses are:

**1. Export:** This function helps to export the HDFS directory into a database table  
**2. List Tables:**This function would help the user to list all tables in a particular database.  
**3. Codegen:** This function would help you to generate code so that you can interact with varied types of database records.   
**4. Create:** This function allows a user to import the table definition within the hive of databases.  
**5. Eval:**This function would always help you to assess the SQL statement and display the results.  
**6. Version:** This function would help you to depict the information related to the text of the database.   
**7. Import all tables:** This function would help a user to import all the tables from a database to HDFS.  
**8. List all the databases:**This function would assist a user to create a list of the available databases on a particular server.

**Related Article:** [What is Apache Hadoop Sqoop](https://mindmajix.com/hadoop/what-is-apache-hadoop-sqoop)

**17. What is the meaning of Sqoop Validation?**  
It refers to the manner in which data validation happens when it is copied. It can also be executed by either exporting or importing the data. It can also be done with the help of a basic comparison between the row counts from the source. You can also opt to use the option to make sure that you are comparing the row counts between the target as well as the source. During the time of the imports, all the rows can be deleted and added. In this context, it is important to note that during the whole process, Sqoop keeps a tab on the changes that have been affected.

**18. Give a basic introduction to Sqoop**  
When it comes to transfer data between relational database servers and Hadoop, you should know that Sqoop is one of the best tools. In order to be more specific, you should use it in importing data from various types of relational databases. It is important for you to note that you can import data from varied types of databases such as MySQL, HDFS, and Hadoop. It is also interesting to note that you have the option to export data from the Hadoop file with the help of Sqoop. This functionality is being provided by the Apache Software Foundation.

It is also important to mention that Sqoop utilizes two main tools. They are in the form of Sqoop export and Sqoop import. With the help of these two tools, you can now extract data information form varied types of databases.

**Related Article:**[Difference between HBase and RDBMS – Hadoop](https://mindmajix.com/hadoop/difference-between-hbase-rdbms)

**19. What are the limitations of importing the RDBMS tables into the Hcatalog directly?**  
In order to import the tables into the Hcatalog in a direct manner, you have to make sure that you are using the –Hcatalog database option. However, in this process, you would face a limitation of importing the tables. It is in the form of the fact that this option do not supports a plethora of arguments like –direct, –as-Avro file and -export-dir.

**20. Shed light on the procedure of updating the rows that have been directly exported**  
In order to update the existing rows that have been exported, you have to use a particular parameter. This parameter is in the form of update key. You can also opt to use a list of comma-separated commands. This would help you to identify a row in a uni fashion. A majority of the columns are used in the Where clause of the update ry that has been already been generated. Moreover, all the other types of table columns should be used in the SET portion of the generated ry.

**21. What is the significance of the Sqoop Import Mainframe tool? Shed light on its purpose too**  
The Sqoop Import Mainframe tool can also be used to import all the important datasets which lies in a partitioned dataset. The partitioned dataset is also known as PDS. The PDS is also known to a directory on varied types of open systems. It is important for you to note that in a dataset, the various types of records would be stored as a single text field with the help of the entire record. This tool would always help you to make sure that you are importing the right types of data tools and that too in a proper manner.

**22. Define Sqoop metastore**  
It is also known as a shared metadata repository with the help of which the local users can execute and define various types of list tables. In order to connect to the metastore, you have to make changes to the Sqoop –site.xml.

**23. Does Sqoop uses the maps reduce function? If it does then shed light on the reasons**  
Apache Sqoop also uses the Map-Reduce function of Hadoop to obtain data from the relational databases. During the process of importing data, Sqoop controls the mappers and their numbers. The mappers who access RDBMS come across denial of service attacks. Hence, it can be said that with the help of Sqoop, big data can be efficiently managed.

**24. Describe the practicality of opting for Sqoop nowadays**  
Apache Sqoop is regarded as an excellent help for those individuals who face challenges in transferring data out of the data warehouse. It is also used for importing data from RDBMS to HDFS. With the help of Sqoop, the users can also import more than one table. Interestingly, with the use of Apache Sqoop, the data selected columns can be easily exported. Furthermore, Sqoop is also compatible with a majority of JDBC databases. Here is the list of stions which would help you to crack the Sqoop interview.

**What is the role of JDBC driver in a Sqoop set up?**

To connect to different relational databases sqoop needs a connector. Almost every DB vendor makes this connecter available as a JDBC driver which is specific to that DB. So Sqoop needs the JDBC driver of each of the database it needs to inetract with.

**Is JDBC driver enough to connect sqoop to the databases?**

No. Sqoop needs both JDBC and connector to connect to a database.

**When to use --target-dir and when to use --warehouse-dir while importing data?**

To specify a particular directory in HDFS use --target-dir but to specify the parent directory of all the sqoop jobs use --warehouse-dir. In this case under the parent directory sqoop will cerate a directory with the same name as th e table.

**How can you import only a subset of rows form a table?**

By using the WHERE clause in the sqoop import statement we can import only a subset of rows.

**How can we import a subset of rows from a table without using the where clause?**

We can run a filtering ry on the database and save the result to a temporary table in database.

Then use the sqoop import command without using the --where clause

**What is the advantage of using --password-file rather than -P option while preventing the display of password in the sqoop import statement?**

The --password-file option can be used inside a sqoop script while the -P option reads from standard input , preventing automation.

**What is the default extension of the files produced from a sqoop import using the --compress parameter?**

.gz

**What is the significance of using --compress-codec parameter?**

To get the out file of a sqoop import in formats other than .gz like .bz2 we use the --compress -code parameter.

**What is a disadvantage of using --direct parameter for faster data load by sqoop?**

The native utilities used by databases to support faster laod do not work for binary data formats like SenceFile.

**How can you control the number of mappers used by the sqoop command?**

The Parameter --num-mapers is used to control the number of mappers executed by a sqoop command. We should start with choosing a small number of map tasks and then gradually scale up as choosing high number of mappers initially may slow down the performance on the database side.

**How can you avoid importing tables one-by-one when importing a large number of tables from a database?**

Using the command

sqoop import-all-tables

--connect

--usrename

--password

--exclude-tables table1,table2 ..

This will import all the tables except the ones mentioned in the exclude-tables clause.

**When the source data keeps getting updated frently, what is the approach to keep it in sync with the data in HDFS imported by sqoop?**

sqoop can have 2 approaches.

**a** − To use the --incremental parameter with append option where value of some columns are checked and only in case of modified values the row is imported as a new row.

**b** − To use the --incremental parameter with lastmodified option where a date column in the source is checked for records which have been updated after the last import.

**What is the usefulness of the options file in sqoop.**

The options file is used in sqoop to specify the command line values in a file and use it in the sqoop commands.

For example the --connect parameter's value and --user name value scan be stored in a file and used again and again with different sqoop commands.

Is it possible to add a parameter while running a saved job?

Yes, we can add an argument to a saved job at runtime by using the --exec option

sqoop job --exec jobname -- -- newparameter

**How do you fetch data which is the result of join between two tables?**

By using the --ry parameter in place of --table parameter we can specify a sql ry. The result of the ry will be imported.

**How can we slice the data to be imported to multiple parallel tasks?**

Using the --split-by parameter we specify the column name based on which sqoop will divide the data to be imported into multiple chunks to be run in parallel.

**How can you choose a name for the mapreduce job which is created on submitting a free-form ry import?**

By using the --mapreduce-job-name parameter. Below is a example of the command.

sqoop import \

--connect jdbc:mysql://mysql.example.com/sqoop \

--username sqoop \

--password sqoop \

--ry 'SELECT normcities.id, \

countries.country, \

normcities.city \

FROM normcities \

JOIN countries USING(country\_id) \

WHERE $CONDITIONS' \

--split-by id \

--target-dir cities \

--mapreduce-job-name normcities

**Before starting the data transfer using mapreduce job, sqoop takes a long time to retrieve the minimum and maximum values of columns mentioned in –split-by parameter. How can we make it efficient?**

We can use the --boundary –ry parameter in which we specify the min and max value for the column based on which the split can happen into multiple mapreduce tasks. This makes it faster as the ry inside the –boundary-ry parameter is executed first and the job is ready with the information on how many mapreduce tasks to create before executing the main ry.

**What is the difference between the parameters sqoop.export.records.per.statement and sqoop.export.statements.per.transaction**

The parameter “sqoop.export.records.per.statement” specifies the number of records that will be used in each insert statement. But the parameter “sqoop.export.statements.per.transaction” specifies how many insert statements can be processed parallel during a transaction.

**How will you implement all-or-nothing load using sqoop?**

Using the staging-table option we first load the data into a staging table and then load it to the final target table only if the staging load is successful.

**How do you clear the data in a staging table before loading it by Sqoop?**

By specifying the –clear-staging-table option we can clear the staging table before it is loaded. This can be done again and again till we get proper data in staging.

**How will you update the rows that are already exported?**

The parameter --update-key can be used to update existing rows. In it a comma-separated list of columns is used which unily identifies a row. All of these columns is used in the WHERE clause of the generated UPDATE ry. All other table columns will be used in the SET part of the ry.

**How can you sync a exported table with HDFS data in which some rows are deleted?**

Truncate the target table and load it again.

**How can you export only a subset of columns to a relational table using sqoop?**

By using the –column parameter in which we mention the required column names as a comma separated list of values.

**How can we load to a column in a relational table which is not null but the incoming value from HDFS has a null value?**

By using the –input-null-string parameter we can specify a default value and that will allow the row to be inserted into the target table.

**How can you schedule a sqoop job using Oozie?**

Oozie has in-built sqoop actions inside which we can mention the sqoop commands to be executed.

**Sqoop imported a table successfully to HBase but it is found that the number of rows is fewer than expected. What can be the cause?**

Some of the imported records might have null values in all the columns. As Hbase does not allow all null values in a row, those rows get dropped.

**Give a sqoop command to show all the databases in a MySql server.**

$ sqoop list-databases --connect jdbc:mysql://database.example.com/

**What do you mean by Free Form Import in Sqoop?**

Sqoop can import data form a relational database using any SQL ry rather than only using table and column name parameters.

**How can you force sqoop to execute a free form Sql ry only once and import the rows serially.**

By using the –m 1 clause in the import command, sqoop cerates only one mapreduce task which will import the rows sentially.

**In a sqoop import command you have mentioned to run 8 parallel Mapreduce task but sqoop runs only 4. What can be the reason?**

The Mapreduce cluster is configured to run 4 parallel tasks. So the sqoop command must have number of parallel tasks less or equal to that of the MapReduce cluster.

**What is the importance of --split-by clause in running parallel import tasks in sqoop?**

The –split-by clause mentions the column name based on whose value the data will be divided into groups of records. These group of records will be read in parallel by the mapreduce tasks.

**What does this sqoop command achieve?**

$ sqoop import --connnect <connect-str> --table foo --target-dir /dest \

It imports data from a database to a HDFS file named foo located in the directory /dest

**What happens when a table is imported into a HDFS directory which already exists using the –apend parameter?**

Using the --append argument, Sqoop will import data to a temporary directory and then rename the files into the normal target directory in a manner that does not conflict with existing filenames in that directory.

**How can you control the mapping between SQL data types and Java types?**

By using the --map-column-java property we can configure the mapping between.

Below is an example

$ sqoop import ... --map-column-java id = String, value = Integer

**How to import only the updated rows form a table into HDFS using sqoop assuming the source has last update timestamp details for each row?**

By using the lastmodified mode. Rows where the check column holds a timestamp more recent than the timestamp specified with --last-value are imported.

**What are the two file formats supported by sqoop for import?**

Delimited text and Sence Files.

**Give a sqoop command to import the columns employee\_id,first\_name,last\_name from the MySql table Employee**

$ sqoop import --connect jdbc:mysql://host/dbname --table EMPLOYEES \

--columns "employee\_id,first\_name,last\_name"

**Give a sqoop command to run only 8 mapreduce tasks in parallel**

$ sqoop import --connect jdbc:mysql://host/dbname --table table\_name\

-m 8

**What does the following ry do?**

$ sqoop import --connect jdbc:mysql://host/dbname --table EMPLOYEES \

--where "start\_date > '2012-11-09'

It imports the employees who have joined after 9-NOv-2012.

**Give a Sqoop command to import all the records from employee table divided into groups of records by the values in the column department\_id.**

$ sqoop import --connect jdbc:mysql://db.foo.com/corp --table EMPLOYEES \

--split-by dept\_id

**What does the following ry do?**

$ sqoop import --connect jdbc:mysql://db.foo.com/somedb --table sometable \

--where "id > 1000" --target-dir /incremental\_dataset --append

It performs an incremental import of new data, after having already imported the first 100,0rows of a table

**Give a sqoop command to import data from all tables in the MySql DB DB1.**

sqoop import-all-tables --connect jdbc:mysql://host/DB1

**Give a command to execute a stored procedure named proc1 which exports data to from MySQL db named DB1 into a HDFS directory named Dir1.**

$ sqoop export --connect jdbc:mysql://host/DB1 --call proc1 \

--export-dir /Dir1

**What is a sqoop metastore?**

It is a tool using which Sqoop hosts a shared metadata repository. Multiple users and/or remote users can define and execute saved jobs (created with sqoop job) defined in this metastore.

Clients must be configured to connect to the metastore in sqoop-site.xml or with the --meta-connect argument.

**What is the purpose of sqoop-merge?**

The merge tool combines two datasets where entries in one dataset should overwrite entries of an older dataset preserving only the newest version of the records between both the data sets.

**How can you see the list of stored jobs in sqoop metastore?**

sqoop job –list

**Give the sqoop command to see the content of the job named myjob?**

Sqoop job –show myjob

**Which database the sqoop metastore runs on?**

Running sqoop-metastore launches a shared HSQLDB database instance on the current machine.

Where can the metastore database be hosted?

The metastore database can be hosted anywhere within or outside of the Hadoop cluster..

-------------

**1 What is the process to perform an incremental data load in Sqoop?**

The process to perform incremental data load in Sqoop is to synchronize the modified or updated data (often referred as delta data) from RDBMS to Hadoop. The delta data can be facilitated through the incremental load command in Sqoop.

Incremental load can be performed by using Sqoop import command or by loading the data into hive without overwriting it. The different attributes that need to be specified during incremental load in Sqoop are-

* Mode (incremental) –The mode defines how Sqoop will determine what the new rows are. The mode can have value as Append or Last Modified.
* Col (Check-column) –This attribute specifies the column that should be examined to find out the rows to be imported.
* Value (last-value) –This denotes the maximum value of the check column from the previous import operation.

**2 How Sqoop can be used in a Java program?**

The Sqoop jar in classpath should be included in the java code. After this the method Sqoop.runTool () method must be invoked. The necessary parameters should be created to Sqoop programmatically just like for command line.

**3 What is the significance of using –compress-codec parameter?**

To get the out file of a sqoop import in formats other than .gz like .bz2 we use the –compress -code parameter.

**4 How are large objects handled in Sqoop?**

Sqoop provides the capability to store large sized data into a single field based on the type of data. Sqoop supports the ability to store-

* CLOB ‘s – Character Large Objects
* BLOB’s –Binary Large Objects

Large objects in Sqoop are handled by importing the large objects into a file referred as “LobFile” i.e. Large Object File. The LobFile has the ability to store records of huge size, thus each record in the LobFile is a large object.

**5 What is a disadvantage of using –direct parameter for faster data load by sqoop?**

The native utilities used by databases to support faster load do not work for binary data formats like Sence File.

**6 Is it possible to do an incremental import using Sqoop?**

Yes, Sqoop supports two types of incremental imports-

Append and Last Modified

To insert only rows Append should be used in import command and for inserting the rows and also updating Last-Modified should be used in the import command.

**7 How can you check all the tables present in a single database using Sqoop?**

The command to check the list of all tables present in a single database using Sqoop is as follows-

Sqoop list-tables –connect jdbc: mysql: //localhost/user;

**8 How can you control the number of mappers used by the sqoop command?**

The Parameter –num-mappers is used to control the number of mappers executed by a sqoop command. We should start with choosing a small number of map tasks and then gradually scale up as choosing high number of mappers initially may slow down the performance on the database side.

**9 What is the standard location or path for Hadoop Sqoop scripts?**

/usr/bin/Hadoop Sqoop.

**10 How can we import a subset of rows from a table without using the where clause?**

We can run a filtering ry on the database and save the result to a temporary table in database.

Then use the sqoop import command without using the –where clause.

**11 When the source data keeps getting updated frently, what is the approach to keep it in sync with the data in HDFS imported by sqoop?**

qoop can have 2 approaches.

a − To use the –incremental parameter with append option where value of some columns are checked and only in case of modified values the row is imported as a new row.

b − To use the –incremental parameter with lastmodified option where a date column in the source is checked for records which have been updated after the last import.

**12 What is a sqoop metastore?**

It is a tool using which Sqoop hosts a shared metadata repository. Multiple users and/or remote users can define and execute saved jobs (created with sqoop job) defined in this metastore.

Clients must be configured to connect to the metastore in sqoop-site.xml or with the –meta-connect argument.

**13 Can free form SQL ries be used with Sqoop import command? If yes, then how can they be used?**

Sqoop allows us to use free form SQL ries with the import command. The import command should be used with the –e and – ry options to execute free form SQL ries. When using the –e and –ry options with the import command the –target dir value must be specified.

**14 Tell few import control commands:**

–Append

–Columns

–Where

These command are most frently used to import RDBMS Data.

**15 Can free form SQL ries be used with Sqoop import command? If yes, then how can they be used?**

Sqoop allows us to use free form SQL ries with the import command. The import command should be used with the –e and – ry options to execute free form SQL ries. When using the –e and –ry options with the import command the –target dir value must be specified.

**16 How can you see the list of stored jobs in sqoop metastore?**

sqoop job –list

**17 What type of databases Sqoop can support?**

MySQL, Oracle, PostgreSQL, IBM, Netezza and Teradata. Every database connects through jdbc driver.

**18 What is the purpose of sqoop-merge?**

The merge tool combines two datasets where entries in one dataset should overwrite entries of an older dataset preserving only the newest version of the records between both the data sets.

**19 HOw sqoop can handle large objects?**

Blog and Clob columns are common large objects. If the object is less than 16MB, it stored inline with the rest of the data. If large objects, temporary stored in\_lob subdirectory. Those lobs processes in a streaming fashion. Those data materialized in memory for processing. IT you set LOB to 0, those lobs objects placed in external storage.

**20 What is the importance of eval tool?**

It allows user to run sample SQL ries against Database and preview the results on the console. It can help to know what data can import? The desired data imported or not?

**21 What is the default extension of the files produced from a sqoop import using the –compress parameter?**

.gz

**22 Can we import the data with “Where” condition?**

Yes, Sqoop has a special option to export/import a particular data.

**23 What are the limitations of importing RDBMS tables into Hcatalog directly?**

There is an option to import RDBMS tables into Hcatalog directly by making use of –hcatalog –database option with the –hcatalog –table but the limitation to it is that there are several arguments like –as-avro file , -direct, -as-sencefile, -target-dir , -export-dir are not supported.

**24 what are the majorly used commands in sqoop?**

In Sqoop Majorly Import and export command are used. But below commands are also useful sometimes. codegen, eval, import-all-tables, job, list-database, list-tables, merge, metastore.

**25 What is the usefulness of the options file in sqoop.**

The options file is used in sqoop to specify the command line values in a file and use it in the sqoop commands.

For example the –connect parameter’s value and –user name value scan be stored in a file and used again and again with different sqoop commands.

**26 what are the common delimiters and escape character in sqoop?**

The default delimiters are a comma(,) for fields, a newline(\n) for records

Escape characters are \b,\n,\r,\t,\”, \\’,\o etc

**27 What are the two file formats supported by sqoop for import?**

Delimited text and Sence Files.

**28 while loading table from MySQL into HDFS, if we need to copy tables with maximum possible speed, what can you do?**

We need to use -direct argument in import command to use direct import fast path and this -direct can be used only with MySQL and PostGreSQL as of now.

**29 How can you sync a exported table with HDFS data in which some rows are deleted?**

Truncate the target table and load it again.

**30 Differentiate between Sqoop and distCP.**

DistCP utility can be used to transfer data between clusters whereas Sqoop can be used to transfer data only between Hadoop and RDBMS.

**31 How can you import only a subset of rows form a table?**

By using the WHERE clause in the sqoop import statement we can import only a subset of rows.

**32 How do you clear the data in a staging table before loading it by Sqoop?**

By specifying the –clear-staging-table option we can clear the staging table before it is loaded. This can be done again and again till we get proper data in staging.

**33 What is Sqoop?**

Sqoop is an open source project that enables data transfer from non-hadoop source to hadoop source. It can be remembered as SQL to Hadoop -> SQOOP. It allows user to specify the source and target location inside the Hadoop.

**34 Is it possible to do an incremental import using Sqoop?**

Yes, Sqoop supports two types of incremental imports-

Append and Last Modified

To insert only rows Append should be used in import command and for inserting the rows and also updating Last-Modified should be used in the import command.

**35 How can you export only a subset of columns to a relational table using sqoop?**

By using the –column parameter in which we mention the required column names as a comma separated list of values.

**36 Which database the sqoop metastore runs on?**

Running sqoop-metastore launches a shared HSQLDB database instance on the current machine.

**37 How will you update the rows that are already exported?**

The parameter –update-key can be used to update existing rows. In it a comma-separated list of columns is used which unily identifies a row. All of these columns is used in the WHERE clause of the generated UPDATE ry. All other table columns will be used in the SET part of the ry.

**38 You have a data in HDFS system, if you want to put some more data to into the same table, will it append the data or overwrite?**

No it can’t overwrite, one way to do is copy the new file in HDFS.

**39 Where can the metastore database be hosted?**

The metastore database can be hosted anywhere within or outside of the Hadoop cluster.

**40 Which is used to import data in Sqoop ?**

In SQOOP import command is used to import RDBMS data into HDFS. Using import command we can import a particular table into HDFS.

**41 What is the role of JDBC driver in a Sqoop set up?**

To connect to different relational databases sqoop needs a connector. Almost every DB vendor makes this connecter available as a JDBC driver which is specific to that DB. So Sqoop needs the JDBC driver of each of the database it needs to interact with.

**42 How to import only the updated rows form a table into HDFS using sqoop assuming the source has last update timestamp details for each row?**

By using the lastmodified mode. Rows where the check column holds a timestamp more recent than the timestamp specified with –last-value are imported.

**43 What is InputSplit in Hadoop?**

When a hadoop job is run, it splits input files into chunks and assign each split to a mapper to process. This is called Input Split.

**44 Hadoop sqoop word came from ?**

Sql + Hadoop = sqoop

**45 What is the work of Export In Hadoop sqoop ?**

Export the data from HDFS to RDBMS

**46 Use of Codegen command in Hadoop sqoop ?**

Generate code to interact with database records

**47. Use of Help command in Hadoop sqoop ?**

List available commands

**48. How can you schedule a sqoop job using Oozie?**

Oozie has in-built sqoop actions inside which we can mention the sqoop commands to be executed.

**49. What are the two file formats supported by sqoop for import?**

Delimited text and Sence Files.

**50. What is a sqoop metastore?**

It is a tool using which Sqoop hosts a shared metadata repository. Multiple users and/or remote users can define and execute saved jobs (created with sqoop job) defined in this metastore.Clients must be configured to connect to the metastore in sqoop-site.xml or with the –meta-connect argument.

**1. Define Sqoop and why do we use Sqoop?**

[Sqoop](https://www.educba.com/course/apache-sqoop-training/) is an open source data transfer tool, which is designed for transferring data between [Hadoop Ecosystem](https://www.educba.com/course/hive-concepts-and-managing-data-in-hive/) and Relational Database Servers ([RDBMS](https://www.educba.com/course/er-model-and-rdbms-concepts/)). Sqoop is used for importing the data from Relational Databases such as Oracle, MySQL etc, to Hadoop file system ([HDFS](https://www.educba.com/hdfs-vs-hbase/)), and also forexporting data from Hthe adoop file system to relational databases.

**2.What are the different features of Sqoop?**The below are the different features supported by the Sqoop –

1. Loading capacity
2. Full Loading and Incremental Loading
3. Data Compression techniques
4. Importing the SQL ries results
5. Data Connectors for all the major databases
6. Direct data loading support into Hadoop File Systems
7. Security configurations like Kerberos
8. Concurrent Import or Export functionalities

**3. Name the relational databases and hadoop eco-system sources supported in Sqoop?**

Sqoop currently supports [MySQL](https://www.educba.com/course/php-mysql-training-course/), [PostgreSQL](https://www.educba.com/course/basic-postgresql-tutorials/), [Oracle](https://www.educba.com/course/oracle-sql-output-data-and-table/), MSSQL, [Teradata](https://www.educba.com/course/teradata-and-informatica/) and IBM’s Netezza as part of Relation Databases. Currently supported Hadoop Eco-system destination services are HDFC, [Hive](https://www.educba.com/course/hive-concepts-hands-demonstration/), [HBase](https://www.educba.com/course/hbase-managed-hive-tables/), H Catalog and Accumulo. Sqoop uses MySQL as the default database.

**4. How does Sqoop work?**

This is the common Sqoop Interview stions asked in an interview. To perform data transfer, Sqoop uses [export and import](https://www.educba.com/import-export-management/) commands. Map Reduce program will be used in Sqoop internally for storing dataset to HDFS. Commands will be associated with Map tasks to retrieve data from Relational Databases; Reduce task will take the responsibility of placing the retrieved data into the destinations (HDFS/HBase/Hive)

Let’s see the sample commands below for import and export

A command for connecting to MySQL database for importing data from ‘Log’ table

sqoop import –connect jdbc:mysql://localhost/<databasename> –username <USER\_NAME> –password <PASSWORD> –table <tablename> –m 1  
sqoop import –connect jdbc:mysql://localhost/mytestdb –username root –password admin123 –table log –m 1

A command for exporting data from HDFS to Relational Database

sqoop export –connect jdbc:mysql://localhost/sqoop\_export –table <table\_name> export-dir /sqoop/emp\_last/part-m-00000 –update-key id  
sqoop export –connect jdbc:mysql://localhost/sqoop\_export –table log\_table–export-dir /sqoop/data/foler1/part-m-00000

**5. What is Sqoop Metastore? Explain it?**The Sqoop Metastore is a tool available in the Sqoop which will be used to configure the Sqoop application to enable the hosting of a shared repository in the form of metadata. This Metastore can be used to execute the jobs and manage a number of users based on the user roles and activities. All the multiple users can perform multiple tasks or operations concurrently to achieve the tasks efficiently. The Sqoop Metastore will be implemented as an in-memory representation by default. When a job is created within Sqoop, the job definition is stored inside the Metastore and will be listed using Sqoop jobs if needed.

**6. What file formats does Sqoop support while importing the data?**

Sqoop uses two file formats for data import. They are:- Delimited Test File Format and Sence File Format.

Delimited Text File Format: Delimited Text Format is the default file format for importing. We can still explicitly specify using the –as- textile argument. Likewise passing the argument will set the delimited characters between rows and columns.

Sence File Format: This file format we can say it’s a binary file format. This type of format file records is stored in custom record specific data types which are [exposed as Java Classes](https://www.educba.com/serialization-in-java/).

Let us move to the next Sqoop Interview stions.

**7. Can we control a number of mappers in sqoop? If yes, How?**  
Yes, we can control the number of mappers in Sqoop by specifying the [parameter](https://www.educba.com/how-to-evaluate-an-employers/) “-num-mappers” in the sqoop command. This parameter can control the number of map tasks, that is nothing but the degree of parallelism will be used by sqoop. The number will be decided based on the requirement.

Syntax: Use these flags to control the number of mappers: m, -num- mappers

**8. What is Sqoop-merge and explain its use?**

Sqoop merge is a tool which combines two different datasets which maintain the only version by overwriting the entries in an older version of a dataset with new files to make it latest version dataset. There happens a process of flattening while merging the two different datasets which preserves the data without any loss and with efficiency and safety. In order to perform this operation merge key command will be used like “–merge-key”

**9. What are the differences between Sqoop, flume, and distcp?**

Both Distcp and Sqoop are used for transferring the data. Distcp is used for transferring any type of data from one hadoop cluster to another cluster, whereas Sqoop transfers data between Relational databases and[Hadoop ecosystem](https://www.educba.com/hadoop-ecosystem/) such as [Hive](https://www.educba.com/hive-vs-hue/), [HDFS](https://www.educba.com/course/hdfs-tutorials/), and [HBase](https://www.educba.com/hbase-vs-cassandra/) etc. But both the methods use the same approach to copy the data, which is pull/transfer.

Flume has distributed a tool, follows agent-based architecture, for streaming the logs into Hadoop ecosystem. Whereas Sqoop is a connector based architecture.

[Flume](https://www.educba.com/course/apache-flume/) collects and aggregates a huge amount of log data. Flume can collect the data from different type of resources; it doesn’t consider the [schema](https://www.educba.com/course/xsd-xml-schema-definition-beginners-training/) or structured/unstructured data. Flume can pull any type of data. Whereas Sqoop can only import the Relational Database Data, so schema is mandatory for sqoop to process. Generally, for moving bulk workloads, the flume is the best option.

**10. What are the data sources supported by Apache Sqoop?**

The different data sources from various applications supported by the Apache Sqoop are as below:

Hive, HBase, Hadoop Distributed File System (HDFS), HCatalog, Accumulo,

**11. What are the most used commands/functions in Sqoop?**

This is the advanced Sqoop Interview stions asked in an interview. List of basic commands used in Sqoop are as follows:

Codegen -Codegen is used to generate code to communicate with database records.

Eval -Sqoop Eval helps in running [sample SQL ries](https://www.educba.com/is-sql-microsoft/) against the databases and provides the results on the console.

Help -Help list the available commands

Import -Import will import the table into the Hadoop Ecosystem

Export -Export is used to export HDFS Data to Relational Databases.

Create-hive-table -This command is useful for importing table definition into Hive

Import-all-tables -Import-all-tables will import the tables form Relational Databases to HDFS.

List-databases -It will list out all the databases present on a server.

List-tables -It will list out all the tables present in a database.

Versions -It will display the version information.

Functions -Parallel import/export, Full load, Incremental Load, Full load, Comparison, Connectors for [RDBMS](https://www.educba.com/course/er-model-and-rdbms-concepts/) Databases, Kerberos Security [Integration](https://www.educba.com/course/project-management-2/), Load data directly into HDFS (Hive/HBase)

**12. Explain the Best Practices while importing tables from MySQL or any other databases using Sqoop?**

While importing the [tables from MySQL](https://www.educba.com/cheat-sheet-mysql/), we should make sure about few things like authentication and authorization to the target server and databases. We need to make sure that we have granted necessary privileges on the databases, which are to be accessed and also make sure about the hostname resolution when we connect to the source and destination hostnames. If we don’t have necessary permission, we will get connection failure exception while connecting to the database.

**13. How do you update the data or rows already exported?**

To update the rows, that are already exported the destination we can use the parameter “–update-key”. In this, a comma-separated columns list is used which unily identifies a row and all of these columns are used in WHERE clause of the generated UPDATE ry. SET part of the ry will take care of all the other table columns.

Let us move to the next Sqoop Interview stions.

**14. How to configure and install JDBC driver in Apache Sqoop?**

The JDB drivers in Apache Sqoop can be configured based on the Hadoop provider such as Cloudera or Hortonworks where it slightly varies in its configuration based on the Hadoop provider. The JDBC in Cloudera can be configured as by creating a library folder like /var/lib/. This can be done for any third party library required to be configured as per the requirement. In this way, any type of database can be configured using its JDBC driver. Apart from the JDBC driver, Apache Sqoop requires a connector to establish a connection between different relational databases. The main components required to establish a connection with the databases is through Driver and Connector of the particular database provider.

**15. What is the split-by clause and when do we use it?**

A split-by parameter is for slicing the data to be imported into multiple parallel tasks. Using this parameter, we can specify the columns names, these are columns name based on which sqoop will be dividing the data to be imported into multiple chunks and they will be running in a parallel fashion. It is one the technique to tune the performance in Sqoop.

**1. What is Sqoop?**

Sqoop is an open source tool that enables users to transfer bulk data between Hadoop eco system and relational databases.

**2. What are the relational databases supported in Sqoop?**

Below are the list of RDBMSs that are supported by Sqoop Currently.

MySQL, PostGreSQL, Oracle, Microsoft SQL, IBM’s Netezza, Teradata

**3. What are the destination types allowed in Sqoop Import command?**

Currently Sqoop Supports data imported into below services.

HDFS, Hive, HBase, HCatalog, Accumulo

**4. Is Sqoop similar to distcp in hadoop?**

Partially yes, hadoop’s distcp command is similar to Sqoop Import command. Both submits parallel map-only. jobs but distcp is used to copy any type of 􀃕les from Local FS/HDFS to HDFS and Sqoop is for transferring the data records only between RDMBS and Hadoop eco system services, HDFS, Hive and HBase.

**5. What are the majorly used commands in Sqoop?**

In Sqoop Majorly Import and export commands are used. But below commands are also useful some times.

codegen, eval, import-all-tables, job, list-databases, list-tables, merge, metastore

**6. When Importing tables from MySQL to what are the precautions that needs to be taken care w.r.t to access?**

In MySQL, we need to make sure that we have granted all privileges on the databases, that needs to be accessed, should be given to all users at destination hostname. If Sqoop is being run under localhost and MySQL is also present on the same then we can grant the permissions with below two commands from MySQL

shell logged in with ROOT user.

**7. What if my MySQL server is running on MachineA and Sqoop is running on MachineB for the above stion?**

From MachineA login to MySQL shell and perform the below command as root user. If using hostname of second machine, then that should be added to /etc/hosts 􀃕le of 􀃕rst machine.

**8. How Many Mapreduce jobs and Tasks will be submitted for Sqoop copying into HDFS?**

For each sqoop copying into HDFS only one mapreduce job will be submitted with 4 map tasks. There will not be any reduce tasks scheduled.

**9. How can we control the parallel copying of RDBMS tables into hadoop ?**

We can control/increase/decrease speed of copying by con􀃕guring the number of map tasks to be run for each sqoop copying process. We can do this by providing argument -m 10 or –num-mappers 10 argument to sqoop import command. If we specify -m 10 then it will submit 10 map tasks parallel at a time. Based on our requirement we can increase/decrease this number to control the copy speed.

**10. What is the criteria for specifying parallel copying in Sqoop with multiple parallel map tasks?**

To use multiple mappers in Sqoop, RDBMS table must have one primary key column (if present) in a table and the same will be used as split-by column in Sqoop process. If primary key is not present, we need to provide any uni key column or set of columns to form uni values and these should be provided to -split-by column argument.

**11. While loading tables from MySQL into HDFS, if we need to copy tables with maximum possible speed, what can you do ?**

We need to use –direct argument in import command to use direct import fast path and this –direct can be used only with MySQL and PostGreSQL as of now.

**12. What is the example connect string for Oracle database to import tables into HDFS?**

We need to use Oracle JDBC Thin driver while connecting to Oracle database via Sqoop. Below is the sample import command to pull table employees from oracle database testdb.

**13. While connecting to MySQL through Sqoop, I am getting Connection Failure exception what might be the root cause and 􀃕x for this error scenario?**

This might be due to insu􀃞cient permissions to access your MySQL database over the network. To con􀃕rm this we can try the below command to connect to MySQL database from Sqoop’s client machine.

$ mysql -u root -p

mysql> GRANT ALL PRIVILEGES ON \*.\* TO '%'@'localhost';

mysql> GRANT ALL PRIVILEGES ON \*.\* TO ''@'localhost';

$ mysql -u root -p

mysql> GRANT ALL PRIVILEGES ON \*.\* TO '%'@'MachineB hostname or Ip address';

mysql> GRANT ALL PRIVILEGES ON \*.\* TO ''@'MachineB hostname or Ip address';

sqoop import \

--connect jdbc:oracle:thin:@oracle.example.com/testdb \

--username SQOOP \

--password sqoop \

--table employees

If this is the case then we need grant permissions user @ sqoop client machine as per the answer to stion

**14. While importing tables from Oracle database, Sometimes I am getting java.lang.IllegalArgumentException: Attempted to generate class with no**

**columns! or NullPointerException what might be the root cause and 􀃕x for this error scenario?**

While dealing with Oracle database from Sqoop, Case sensitivity of table names and user names matters highly. Most probably by specifying these two values in UPPER case will solve the issue unless actual names are mixed with Lower/Upper cases. If these are mixed, then we need to provide them within double quotes. In case, the source table is created under di􀃠erent user namespace, then we need to provide table name as

**1) I have 20000 records in a table. I want copy them to two separate files( records equally distributed) into HDFS (using Sqoop).**

**2)How do we achieve this, if table does not have primary key or uni key?**