**SqooP Interview Questions**

**Important Links:-**

<https://data-flair.training/blogs/sqoop-interview-questions/>

<https://intellipaat.com/blog/interview-question/sqoop-interview-questions/>

<https://mindmajix.com/sqoop-interview-questions>

<https://www.dezyre.com/article/sqoop-interview-questions-and-answers-for-2018/274>

<https://www.bigdatatrunk.com/top-50-sqoop-interview-questions/>

<https://www.tutorialspoint.com/sqoop/sqoop_interview_questions.htm>

<http://hadooptutorial.info/sqoop-interview-questions-and-answers-for-experienced/>

<https://www.wisdomjobs.com/e-university/sqoop-interview-questions.html>

**Mention the best features of Apache Sqoop.**

**Ans.** 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 query
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

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

**Ans.**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 sequence files all records are there as binary data here. To be more specific,  it imports individual tables from RDBMS to HDFS.  
Although, there are many more insights of Sqoop Import, to learn all in detail, follow the link:[**Sqoop import**](https://data-flair.training/blogs/sqoop-import/)

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

Ans. 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) Sequence File Format

We can say, Sequence 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.

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

**Ans.**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.  
To learn [**Sqoop Import**](https://data-flair.training/blogs/sqoop-import/) in detail, follow this link.

**Que 5. How can you execute a free-form SQL query in Sqoop to import the rows in a sequential manner?**

**Ans.** 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.

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

**Ans.** Yes, MySQL is the default database.  
To learn [**Sqoop List Databases**](https://data-flair.training/blogs/sqoop-list-databases/) in detail, follow this link.

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

**Ans.** 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’  
To learn [**Sqoop list Table**](https://data-flair.training/blogs/sqoop-list-tables/)in detail, follow this link.

**Que 8.  If the source data gets updated every now and then, how will you synchronize the data in HDFS that is imported by Sqoop?**

**Ans.**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**.**  
To learn [**Sqoop list Table**](https://data-flair.training/blogs/sqoop-list-tables/)in detail, follow this link.

**Que 9. Name a few import control commands. How can Sqoop handle large objects?**

**Ans.**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.  
To learn [**Sqoop Import**](https://data-flair.training/blogs/sqoop-import/)in detail, follow this link.

**Que 10. How can we import data from particular row or column? What is the destination types allowed in Sqoop import command?**

**Ans.** 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  
–query

**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 –query “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:

1. HDFS
2. Hive
3. Hbase
4. Hcatalog
5. Accumulo

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

**Ans**. 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.

**Que 12. What is the process to perform an incremental data load in Sqoop?**

**Ans.**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.

**Que 13. What is the significance of using –compress-codec parameter?**

**Ans.** However, we use the –compress -code parameter to get the out file of a sqoop import in formats other than .gz like .bz2.

**Que 14. Can free-form SQL queries be used with Sqoop import command? If yes, then how can they be used?**

**Ans.**In Sqoop, we can use SQL queries with the import command. Basically, we should use import command with the –e and – query options to execute free-form SQL queries. But note that the –target dir value must be specified While using the –e and –query options with the import command.

**Que 15. What is the importance of eval tool?**

**Ans.** Basically,[**Sqoop Eval**](https://data-flair.training/blogs/sqoop-eval/) helps to run sample SQL queries 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.

**Que 16. How can you import only a subset of rows from a table?**

**Ans.**In the sqoop import statement, by using the WHERE clause we can import only a subset of rows.

**Que 17. What are the limitations of importing RDBMS tables into Hcatalog directly?**

**Ans.** 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-sequencefile, -target-dir , -export-dir.  
To learn [**Sqoop HCatalog**](https://data-flair.training/blogs/sqoop-hcatalog-integration/) in detail, follow this link.

**Que 18. What is the advantage of using –password-file rather than -P option while preventing the display of password in the sqoop import statement?**

**Ans.** Inside a sqoop script, we can use The –password-file option. Whereas the -P option reads from standard input, preventing automation.

**Que 19. What do you mean by Free Form Import in Sqoop?**

**Ans.**By using any SQL Sqoop can import data from a relational database query rather than only using table and column name parameters.

**Que 20. What is the role of JDBC driver in Sqoop?**

**Ans.**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.

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

**Ans.** No. to connect to a database Sqoop needs both JDBC and connector.  
To learn [**Sqoop Connector**](https://data-flair.training/blogs/sqoop-connectors-drivers/) in detail, follow this link.

**Que 22. What is InputSplit in Hadoop?**

**Ans.** 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.

**Que 23. What is the work of Export in Hadoop sqoop?**

**Ans.** Export tool transfer the data from HDFS to RDBMS  
To learn [**Sqoop Export**](https://data-flair.training/blogs/sqoop-export/) in detail, follow this link.

**Que 24. Use of Codegen command in Hadoop sqoop?**

**Ans.**Basically,Codegen command generates code to interact with database records  
To learn [**Sqoop Codegen**](https://data-flair.training/blogs/sqoop-codegen/)in detail, follow this link.

**Que 25. Use of Help command in Hadoop sqoop?**

**Ans.** Help command in Hadoop sqoop generally list available commands

**Que 26.  How can you schedule a sqoop job using Oozie?**

**Ans.** However, Oozie has in-built sqoop actions inside which we can mention the sqoop commands to be executed.  
To learn [**Sqoop Job**](https://data-flair.training/blogs/sqoop-job/)in detail, follow this link.

**Que 27. What is the importance of — the split-by clause in running parallel import tasks in sqoop?**

**Ans.** 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.

**Que 28. What is a sqoop metastore?**

**Ans.** 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.  
To learn [**Sqoop Job**](https://data-flair.training/blogs/sqoop-job/)in detail, follow this link.

**Que 29. What is the purpose of sqoop-merge?**

**Ans.**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.  
To learn [**Sqoop Merge**](https://data-flair.training/blogs/sqoop-merge/)in detail, follow this link.

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

**Ans.** sqoop job –list

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

**Ans.** Basically, on the current machine running sqoop-metastore launches, a shared HSQLDB database instance.

**Que 32. Where can the metastore database be hosted?**

**Ans.** 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/).

**Que 33. Give the sqoop command to see the content of the job named myjob?**

**Ans.** Sqoop job –show myjob

**Que 34. How can you control the mapping between SQL data types and Java types?**

**Ans.**we can configure the mapping between by using the –map-column-java property.  
For example:  
$ sqoop import … –map-column-java id = String, value = Integer  
To learn[**Java Data types**](https://data-flair.training/blogs/java-data-types/)in detail, follow this link.

**Que 35. Is it possible to add a parameter while running a saved job?**

**Ans.**Yes, by using the –exec option we can add an argument to a saved job at runtime.  
sqoop job –exec jobname — — newparameter

**Que 36. What is the usefulness of the options file in sqoop.**

**Ans.** 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.

**Que 37. How can you avoid importing tables one-by-one when importing a large number of tables from a database?**

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

**Ans.** .gz

**Que 38. How can you control the number of mappers used by the sqoop command?**

**Ans.**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.

**Que 40. What is the significance of using –compress-codec parameter?**

**Ans.**We use the –compress -code parameter to get the out file of a sqoop import in formats other than .gz like .bz2.

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

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

**Que 42. How will you update the rows that are already exported?**

**Ans.** Basically, to update existing rows we can use the parameter –update-key. Moreover, in it, a comma-separated list of columns is used which uniquely identifies a row. All of these columns are used in the WHERE clause of the generated UPDATE query. All other table columns will be used in the SET part of the query.

**Que 43. What are the basic commands in Apache Sqoop and its uses?**

**Ans.** 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

**Que 44. How Sqoop word came? Sqoop is which type of tool and the main use of sqoop?**

**Ans.** Sqoop word came from **SQ**L+HAD**OOP**=SQOOP.  
Basically, it is a data transfer tool. We use Sqoop to import and export a large amount of data from RDBMS to HDFS and vice versa.  
Follow this link to know more about [**Sqoop**](https://data-flair.training/blogs/sqoop-introduction/)

**Que 45. What is Sqoop Validation?**

**Ans.** 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.  
Learn all insights of Sqoop Validation, follow the link: [**Sqoop Validation – Interfaces & Limitations of Sqoop Validate**](https://data-flair.training/blogs/sqoop-validation/)

**Que 46. What is Purpose to Validate in Sqoop?**

**Ans.** 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.

**Que 47. What is Sqoop Job?**

**Ans.** 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.  
Learn all insights of Sqoop job, follow the link: [**Sqoop- Introduction to Sqoop Job Tutorial**](https://data-flair.training/blogs/sqoop-job/)

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

**Ans.** Basically, a tool which we use to import all sequential 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.  
Learn all insights of Sqoop Import Mainframe, follow the link: [**Learn Sqoop Import Mainframe Tool – Syntax and Examples**](https://data-flair.training/blogs/sqoop-import-mainframe/)

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

**Ans.** 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.

**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 |

**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.

**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”

### ****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.

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

### ****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’;

Give your career a big boost by going through our Apache [Sqoop Training Course](https://intellipaat.com/big-data-hadoop-training/#sample-video) now!

### ****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.

**Q8) 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.

**Q9) 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.

**Q10) 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.

**Q13) 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)

**Q14) 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.

**Q15) What is the default file format in order to import data with the utilization of Apache Sqoop?**  
At the time of answering this question, you should know that there are two file formats that can be used in the case of importing data. These are as follows:

**Sequencing the file format**

It is a commonly observed fact that a sequence 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.

**Delimiting 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.

**Q17) 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.

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

**Q19) 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.

**Q20) 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 unique fashion. A majority of the columns are used in the Where clause of the update query that has been already been generated. Moreover, all the other types of table columns should be used in the SET portion of the generated query.

**Q21) 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.

**Q23) 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.

**Q24) 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.

### ****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.

### ****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 –num-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.

### ****Sqoop Interview Questions for Experienced****

**1)**I have 20000 records in a table. I want copy them to two separate files( records equally distributed) into HDFS (using Sqoop).   
How do we achieve this, if table does not have primary key or unique key?

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

Answer: 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.

Q2 How Sqoop can be used in a Java program?

Answer: 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.

 Q3 What is the significance of using –compress-codec parameter?

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

Q4 How are large objects handled in Sqoop?

Answer: 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. 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. If you set LOB to 0, those lobs objects placed in external storage.

Q5 What is a disadvantage of using –direct parameter for faster data load by sqoop?

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

Q6 Is it possible to do an incremental import using Sqoop?

Answer: Yes, Sqoop supports two types of incremental imports-

1. Append
2. 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.

 Q7 How can you check all the tables present in a single database using Sqoop?

Answer: 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;

Q8 How can you control the number of mappers used by the sqoop command?

Answer: 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.

Q9 What is the standard location or path for Hadoop Sqoop scripts?

Answer: /usr/bin/Hadoop Sqoop.

Q10 How can we import a subset of rows from a table without using the where clause?

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

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

 Q12 What is a sqoop metastore?

Answer: 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.

 Q13 Can free form SQL queries be used with Sqoop import command? If yes, then how can they be used?

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

Q15 Can free form SQL queries be used with Sqoop import command? If yes, then how can they be used?

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

 Q16 How can you see the list of stored jobs in sqoop metastore?

Answer: sqoop job –list

Q17 What type of databases Sqoop can support?

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

 Q18 What is the purpose of sqoop-merge?

Answer: 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.

Q20 What is the importance of eval tool?

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

Q21 What is the default extension of the files produced from a sqoop import using the –compress parameter?

Answer: .gz

Q22 Can we import the data with “Where” condition?

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

Q25 What is the usefulness of the options file in sqoop.

Answer: 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.

Q26 what are the common delimiters and escape character in sqoop?

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

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

 Q27 What are the two file formats supported by sqoop for import?

Answer: Delimited text and Sequence Files.

Q28 While loading table from MySQL into HDFS, if we need to copy tables with maximum possible speed, what can you do?

Answer: 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.

Q29 How can you sync an exported table with HDFS data in which some rows are deleted?

Answer: Truncate the target table and load it again.

Q31 How can you import only a subset of rows form a table?

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

Q32 How do you clear the data in a staging table before loading it by Sqoop?

Answer: 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.

Q35 How can you export only a subset of columns to a relational table using sqoop?

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

Q36 Which database the sqoop metastore runs on?

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

Q37 How will you update the rows that are already exported?

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

Q38 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?

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

 Q39 Where can the metastore database be hosted?

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

Q41 What is the role of JDBC driver in a Sqoop set up?

Answer: 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.

Q42 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?

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

 Q43 What is InputSplit in Hadoop?

Answer: 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.

Q47 Use of Help command in Hadoop sqoop ?

Answer: List available commands

Q48 How can you schedule a sqoop job using Oozie?

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

 Q49 What are the two file formats supported by sqoop for import?

Answer: Delimited text and Sequence Files.

**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 the table.

**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 --query parameter in place of --table parameter we can specify a sql query. The result of the query 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 query 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 \

--query '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 –query 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 query inside the –boundary-query parameter is executed first and the job is ready with the information on how many mapreduce tasks to create before executing the main query.

**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.

**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

**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 query 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 query 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 1000 rows 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

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

We can control/increase/decrease speed of copying by configuring 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 unique key column or set of columns to form unique values and these should be provided to **-split-by** column argument.

###### **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**.

sqoop import \

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

--username SQOOP \

--password sqoop \

--table employees