# Sqoop Interview Questions

## What will happen if target directory already exists during sqoop import?

Ans: Sqoop runs a map-only job and if the target directory is present, it will throw an exception.

## What is the use of warehouse directory in Sqoop import?

Ans: warehouse directory is the HDFS parent directory for table destination. If we specify target-directory all our files are stored in that location. But, with warehouse directory, a child directory is created inside it with the name of the table. All the files are stored inside the child directory.

## What is the default number of mappers in a Sqoop job?

Ans: 4

## How to bring data directly into Hive using Sqoop?

Ans: To bring data directly into Hive using Sqoop use –hive-import command.

## We wish to bring data in CSV format in HDFS from RDBMS source. The column in RDBMS table contains ‘,’. How to distinctly import data in this case?

Ans: Use can use the option –optionally-enclosed-by

## How to import data directly to HBase using Sqoop?

Ans: You need to use –hbase-table to import data into HBase using sqoop. Sqoop will import data to the table specified as the argument to –hbase-table. Each row of input table will be transformed into an Hbase put operation to a row of output table.

## What is incremental load in Sqoop?

Ans: To import records which are new. For this, you should specify –last-value parameter so that the sqoop job will import values after the specified value.

## What is the benefit of using a Sqoop job?

In the scenario where you must perform incremental import multiple times, you can create a sqoop job for incremental import and run the job. Whenever you run the sqoop job, it will automatically identify last imported value and then the import will start after the identified value.

## Q 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-

1)Mode (incremental) –The mode defines how Sqoop will determine what the new rows are. The mode can have value as Append or Last Modified.

2)Col (Check-column) –This attribute specifies the column that should be examined to find out the rows to be imported.

3)Value (last-value) –This denotes the maximum value of the check column from the previous import operation.

## Q 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.

## Q 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.

## Q 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-

1)CLOB ‘s – Character Large Objects

2)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.

## Q 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 SequenceFile

## Q 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;

## Q 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.

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

Answer: /usr/bin/Hadoop Sqoop

## Q 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

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

Answer: 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.

## Q 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.

## Q 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.

## Q Tell few import control commands:

Answer: –Append

–Columns

–Where

These command are most frequently used to import RDBMS Data.

## Q 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.

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

Answer: sqoop job –list

## Q What type of databases Sqoop can support?

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

## Q 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.

## Q HOw sqoop can handle large objects?

Answer: 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.

## Q 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?

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

Answer: .gz

## Q Can we import the data with “Where” condition?

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

## Q What are the limitations of importing RDBMS tables into Hcatalog directly?

Answer: 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-sequencefile, -target-dir , -export-dir are not supported.

## Q what are the majorly used commands in sqoop?

Answer: 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.

## Q 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.

## Q 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

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

Answer: Delimited text and Sequence Files.

Q 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.

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

Answer: Truncate the target table and load it again.

## Q Differentiate between Sqoop and distCP.

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

## Q 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.

## Q 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.