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

Sqoop allows data to be imported using two file formats

#### **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.

1. **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.

**2) 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.

**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.

**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?

Make sure that you are giving --split-by col1 with -m 2 else you will see an error