SQOOP INTERVIEW QUESTIONS

Dear readers, these **Sqoop Interview Questions** have been designed specially to get you acquainted with the nature of questions you may encounter during your interview for the subject of **Sqoop**. As per my experience good interviewers hardly plan to ask any particular question during your interview, normally questions start with some basic concept of the subject and later they continue based on further discussion and what you answer:

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 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

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 SequenceFile

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 frequently, 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 --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.

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 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.

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 query rather than only using table and column name parameters.

How can you force sqoop to execute a free form Sql query 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 sequentially.

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 Sequence 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 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 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 *createdwithsqoopjob* 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..

What is Next ?

Further you can go through your past assignments you have done with the subject and make sure you are able to speak confidently on them. If you are fresher then interviewer does not expect you will answer very complex questions, rather you have to make your basics concepts very strong.

Second it really doesn't matter much if you could not answer few questions but it matters that whatever you answered, you must have answered with confidence. So just feel confident during your interview. We at tutorialspoint wish you best luck to have a good interviewer and all the very best for your future endeavor. Cheers :-)