# Sqoop interview questions

**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 files 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.
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.

$ mysql -u root -p

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

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

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

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 file of first machine.

$ 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';

**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 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.

**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.

sqoop import \

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

--username SQOOP \

--password sqoop \

--table employees

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

This might be due to insufficient permissions to access your MySQL database over the network. To confirm this we can try the below command to connect to MySQL database from Sqoop’s client machine.

$ mysql --host=MySql node> --database=test --user= --password=

If this is the case then we need grant permissions user @ sqoop client machine as per the answer to Question 6 in this post.

**What is Hadoop sqoop ?**

Hadoop sqoop is an open source and sub project of Hadoop.Hadoop sqoop is a tool that designed for efficiently transfer the huge amount of data between Apache hadoop and structure databases such as relational database management systems(RDBMS) like Sql,oracle ,MySql databases.
In other words,Hadoop sqoop is used for import and export the huge amount of data from RDBMS to HDFS and HDFS to RDBMS. RDBMS such as Mysql,oracle,sql. HDFS such as Hive,Hbase

**Hadoop sqoop word came from ?**
Sql + Hadoop = sqoop

**What is the main use of Hadoop sqoop ?**
Hadoop sqoop mainly used for import and export the huge amount of data from RDBMS to HDFS and HDFS to RDBMS

**Hadoop sqoop is which type of Tool ?**
Hadoop sqoop is a Data transfer tool

**What is latest stable version of Hadoop sqoop ?**
Latest stable version of Hadoop sqoop release is 1.4.5

**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 fix 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 is- sue 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 different user namespace, then we need to provide table name as USERNAME.TABLENAME as shown below.

sqoop import \

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

--username SQOOP \

--password sqoop \

--table SIVA.EMPLOYEES

**What are the main methods of data transfering in Hadoop sqoop ?**
Mainly two operations
i) Import
ii) Export

**What is the work of Import in Hadoop sqoop ?**
Import the data from RDBMS to HDFS

**What is the work of Export In Hadoop sqoop ?**
Export the data from HDFS to RDBMS

**What is the default database for Hadoop sqoop ?**
Mysql

**How to install Mysql In Linux Ubuntu Operating System ?**
sudo apt get install mysql server mysql client

**How to Enter Into MySql prompt ?**
mysql -u root -p

**In the above command what -u Indicates ?**
User

**In the above command what root Indicates ?**
Username

**In the above command what -p Indicates ?**
Password

**How to create a Database in Mysql ?**
create database databasename;

**How to show all Databases names in Mysql ?**
show databases;

**How to use a particular database in Mysql ?**
use databasename;

**How to grant all databases Permissions to single user in mysql ?**
Mysql> grant all privileges on databasename.* to '%'@'localhost';

**How to grant all databases Permissions to single user in mysql ?**
Mysql> grant all privileges on databasename.* to ''@'localhost';

**What this '%' symbol Indicates in above command ?**
grant all databases Permissions to single user

**What this '' symbol Indicates in above command ?**
grant all databases Permissions to all users

**How to create a table In Mysql ?**
mysql> create table emp(empId int, eName varchar(30), eSal int);
Query OK, 0 rows affected (0.11 sec)

**How to Insert the values Into the table ?**
mysql> insert into emp values(111,'mahesh', 28000);
Query OK, 1 row affected (0.00 sec)

mysql> insert into emp values(112,'neelesh', 30000);
Query OK, 1 row affected (0.00 sec)

mysql> insert into emp values(113,'rupesh', 26000);
Query OK, 1 row affected (0.00 sec)

mysql> insert into emp values(114,'vijay', 26000);
Query OK, 1 row affected (0.00 sec)

**How to read the entire table In Mysql ?**

select * from tablename;

**How to update the row in a table ?**

mysql> update emp set eSal= 28000 where empId = 114;

**Grant privileges to users and others ?**

mysql> grant all privileges on mahesh.* to '%'@'localhost';
Query OK, 0 rows affected (0.06 sec)

mysql> grant all privileges on mahesh.* to ''@'localhost';
Query OK, 0 rows affected (0.00 sec)

**What is Hadoop sqoop scripts standard location?**
/usr/bin/Hadoop sqoop

**How we can check Hadoop sqoop installed or not in a system ?**
Just type the Hadoop sqoop help command
Hadoop sqoop help

**What are the basic available commands in Hadoop sqoop ?**

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

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

Generate code to interact with database records

**Use of Create-hive-table command in Hadoop sqoop ?**

Import a table definition into Hive

**Use of Eval command in Hadoop sqoop ?**

Evaluate a SQL statement and display the results

**Use of Export command in Hadoop sqoop ?**

**Export an HDFS directory to a database table**

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

List available commands

**Use of Import command in Hadoop sqoop ?**

Import a table from a database to HDFS

**Use of Import-all-tables command in Hadoop sqoop ?**

Import tables from a database to HDFS

**Use of list-databases command in Hadoop sqoop ?**

List available databases on a server

**Use of list-tables command in Hadoop sqoop ?**

List available tables in a database

**Use of version command in Hadoop sqoop ?**

Display version information

**How to see a information about specific command in Hadoop sqoop ?**

sqoop help COMMAND
sqoop help import
sqoop import –help

**command aliases in Hadoop sqoop ?**

sqoop-(toolname)
sqoop-import,sqoop-export

**How to check List of Tables in single database by using sqoop ?**

root@ubuntu:/home/mahesh/sqoop-related# sqoop list-tables –connect jdbc:mysql://localhost/mahesh;

13/11/07 18:58:21 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.

**How to check List of Databases in RDBMS by using sqoop ?**

root@ubuntu:/home/mahesh/sqoop-related# sqoop list-databases –connect jdbc:mysql://localhost;

13/11/07 18:55:47 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.

O/P:-
information_schema
Gopal_Lab
NewYearDB
RK
batch18
bhargav
chandu
kelly
manoj
mahesh
sivanag

**How to use SQOOP in Java program?**

You can run sqoop from inside your java code by including the sqoop jar in your classpath and calling the Sqoop.runTool() method. You would have to create the required parameters to sqoop programmatically as if it were the command line (e.g. "--connect" etc.).
Please pay attention to the following:

- Make sure that the sqoop tool name (e.g. import/export etc.) is the first parameter.
- Pay attention to classpath ordering - The execution might fail because sqoop requires version X of a library and you use a different version. Ensure that the libraries that sqoop requires are not overshadowed by your own dependencies. I've encountered such a problem with commons-io (sqoop requires v1.4) and had a NoSuchMethod exception since I was using commons-io v1.2.
- Each argument needs to be on a separate array element. For example, "--connect jdbc:mysql:..." should be passed as two separate elements in the array, not one.
- The sqoop parser knows how to accept double-quoted parameters, so use double quotes if you need to (I suggest always). The only exception is the fields-delimited-by parameter which expects a single char, so don't double-quote it.
- I'd suggest splitting the command-line-arguments creation logic and the actual execution so your logic can be tested properly without actually running the tool.
- It would be better to use the --hadoop-home parameter, in order to prevent dependency on the environment.
- The advantage of Sqoop.runTool() as opposed to Sqoop.Main() is the fact that runTool() return the error code of the execution.

http://stackoverflow.com/questions/9229611/how-to-use-sqoop-in-java-program?lq=1

```
        SqoopOptions options = new SqoopOptions();
        options.setConnectString("jdbc:mysql://HOSTNAME:PORT/DATABASE_NAME");
        //options.setTableName("TABLE_NAME");
        //options.setWhereClause("id>10");     // this where clause works when importing
whole table, ie when setTableName() is used
        options.setUsername("USERNAME");
        options.setPassword("PASSWORD");
        //options.setDirectMode(true);    // Make sure the direct mode is off when importing
data to HBase
        options.setNumMappers(8);         // Default value is 4
        options.setSqlQuery("SELECT * FROM user_logs WHERE $CONDITIONS limit 10");
        options.setSplitByCol("log_id");

        // HBase options
        options.setHBaseTable("HBASE_TABLE_NAME");
        options.setHBaseColFamily("colFamily");
        options.setCreateHBaseTable(true);    // Create HBase table, if it does not exist
        options.setHBaseRowKeyColumn("log_id");

        int ret = new ImportTool().run(options);
```