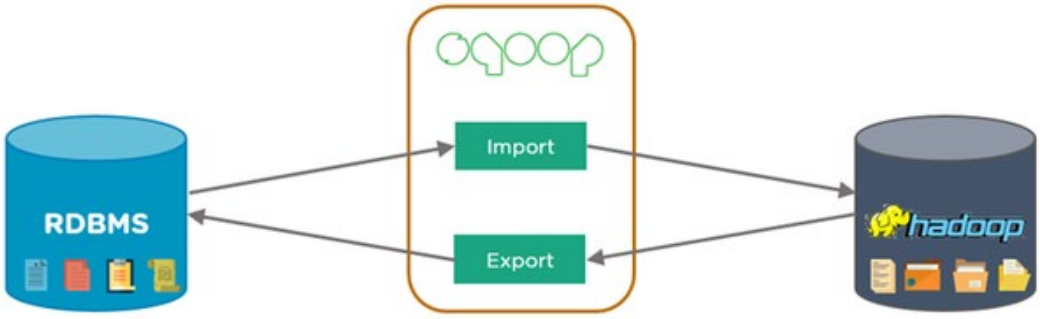
Sqoop

1. **What is Sqoop why it is used?**
   Sqoop word came from SQL+HADOOP=SQOOP. And Sqoop is a data transfer tool.
   Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.



2. **What is the difference between sqoop 1 and sqoop 2?**
   Sqoop:
   - In Sqoop, to access the Hadoop cluster we need to pass the database credentials. It breaches the Security of database.
   - Sqoop cannot integrate with web interface such as hue because it allows only client Architecture.
   - Data transfer from RDBMS to Hive or HBase is supported to Sqoop.
   - Sqoop supports Kerberos Security.

   Sqoop2:
   - In Sqoop2, database credentials will be known to only the admins who manage the cluster. Developers need not know the password.
   - Sqoop 2 runs on client server architecture, Hue can run the sqoop based scripts.
   - In Sqoop2 we need to save the RDBMS data into HDFS after that we need to load to the Hive tables.
   - Sqoop 2 supports Kerberos Security.

3. **What is the difference between warehouse and target directory?**
   - In **target-dir** it will directly import data in the mentioned directory
   - In the case of the **warehouse-dir**, it will create a subfolder named with the table name and import the data.

4. **Can you export the flat file into HDFS using Sqoop?**
   No. There is no direct procedure. Because sqoop is used to do import and export operations between the RDBMS and HDFS.

5. **How many mappers & reducers will run for the Sqoop?**
   By default, the number of mappers is 4. There is no concept of reducers in reducer.

6. **In Sqoop how it works only map function, why not reduce?**
   The Sqoop needs only the mappers to run the import & export operations in parallel. We don't need a reducer because there is no need to use shuffle and sort operations.

7. **How to print the import table data in the console?**
   By using the sqoop eval function we can print the output of the job in the console.
   Ex:- sqoop eval \
   --connect "jdbc:mysql://localhost:3306/goutham" \
   --username root \
   --password hadoop \
   --query "select count(*) from student"

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

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

9. **If the RDBMS table don't have the primary key, how will you import the data into HDFS?**
   By using the concept of **number of mappers equal to one** and using of **split-by** parameter

10. **SQOOP Performance tuning?**

- split-by
- boundary queries
- importing data using fetch-size
- controlling parallelism

11. **What is the use of split-by parameter in sqoop?**
    Whenever the table does not have the primary constraint to ensure the data without loss we use the concept of split-by column name.

12. **You use –split-by clause but it still does not give optimal performance then how will you improve the performance further.**

    --boundary-query : Specifies the range of values that you can import. You can use boundary-query if you do not get the desired results by using the split-by argument alone.

    When you configure the boundary-query argument, you must specify the min(id) and max(id) along with the table name.

13. **importing data using fetch-size?**
    --fetch-size = n : Specifies the number of entries that Sqoop can import at a time. Where <n> represents the number of entries that Sqoop must fetch at a time. Default is 1000. Set the value based on the available memory and bandwidth.

14. **Controlling Parallelism?**
    - Specifies number of map tasks that can run in parallel. Default is 4. To optimize performance, set the number of map tasks to a value lower than the maximum number of connections that the database supports.

- Using more mappers will lead to a higher number of concurrent data transfer tasks, which can result in faster job completion.
- However, it will also increase the load on the database as Sqoop will execute more concurrent queries.

15. How do you check sqoop job is done or not?

By using the submission status –jid <job_id> option is used to inspect or verify particular jobs and their details of job.

16. What is column command in sqoop?

--columns option will enable us to select the particular columns from the table.

17. Don't use insert overwrite command. I want the column data into file ?

By selecting the **--columns** parameter we can get the particular column data.

18. What are the (tools) commands available in sqoop?

Sqoop itself is a tool used to perform the import and export operation between the hdfs and rdbms. The import and export are the two tools in sqoop.

19. When you performing importing and exporting operations if you loose data how you find that loosed records?

By using the concept of staging table if the transaction is dropped middle the data will be stored in the staging table.

20. 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.

21. what are the common delimiters and escape character in sqoop?

- The default delimiters are a comma(,) for fields, a newline(\n) for records
- Escape characters are \b,\n,\r,\t,\", \\',\o etc

22. What is the advantage of using –direct parameter for transferring the data faster way ?

We need to use -direct argument in import command to use direct import in fast rather than normal and this -direct can be used only with MySQL and PostGreSQL as of now.

23. 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.

24. What challenges you faced when you are moving data from RDBMS to Hadoop?

To improve the performance of the Sqoop job we have used the –direct parameter.

- In that case face the problem like it doesn't support to produce text output, binary formats like sequence or Avro.
- In some cases the customization of escape characters, type mapping, column and row delimiters might not supported.

25. What is the use of –validate parameter in sqoop ?

It means to validate the data copied. Either import or export 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.

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

/usr/bin/Hadoop Sqoop

27. What is a sqoop metastore?

Is a shared metadata repository. Multiple users or remote users can define and execute sqoop jobs defined in this metastore.

Clients must be configured to connect to the metastore in sqoop-site.xml or with the –meta-connect argument.

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

sqoop job –list

29. Which database the sqoop metastore runs on?

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

30. 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.

31. what are the majorly used commands in sqoop?

In Sqoop Majorly Import and export command are used. But below commands are also useful sometimes.

- 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 evaluate SQL 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

32. 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 and --password value scan be stored in a file and used again and again with different sqoop commands.

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

Delimited text and Sequence Files.

34. Differentiate between Sqoop and distCP.

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

35. 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 interact with.

36. 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-

EX:

```
sqoop import-all-tables
--connect
–username
–password
--exclude-tables Table298, Table 123, Table 299
```

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

- **--compression-codec:** The compression codec to use when storing the imported data in HDFS. This is set to SnappyCodec. but you can choose a different codec if you prefer.

- The compression codec determines how the data will be compressed and decompressed during the import process.

- It improves performance by reducing the amount of data that needs to be transferred over the network.