Apache Sqoop

Apache Sqoop is a command-line interface application for transferring data between relational databases and Hadoop.

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

Answer)Sqoop allows data to be imported using two file formats
i) 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.
ii) 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)How do I resolve a Communications Link Failure when connecting to MySQL?

Answer)Verify that you can connect to the database from the node where you are running Sqoop:
$ mysql --host=IP Address --database=test --user=username --password=password
Add the network port for the server to your my.cnf file
Set up a user account to connect via Sqoop. Grant permissions to the user to access the database over the network:
Log into MySQL as root mysql -u root -p ThisIsMyPassword
Issue the following command: mysql> grant all privileges on test.* to 'testuser'@'%' identified by 'testpassword'

3)How do I resolve an IllegalArgumentException when connecting to Oracle?

Answer)This could be caused a non-owner trying to connect to the table so prefix the table name with the schema, for example SchemaName.OracleTableName.

4)What's causing this Exception in thread main java.lang.IncompatibleClassChangeError when running non-CDH Hadoop with Sqoop?

Answer)Try building Sqoop 1.4.1-incubating with the command line property -Dhadoopversion=20.

5)How do I resolve an ORA-00933 error SQL command not properly ended when connecting to Oracle?

Answer)Omit the option --driver oracle.jdbc.driver.OracleDriver and then re-run the Sqoop command.

6)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?

Answer)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

7)Does Apache Sqoop have a default database?

Answer)Yes, MySQL is the default database.

8)How can I import large objects (BLOB and CLOB objects) in Apache Sqoop?

Answer)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.

9)How can you execute a free form SQL query in Sqoop to import the rows in a sequential manner?

Answer)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.

10)How will you list all the columns of a table using Apache Sqoop?

Answer)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

11)What is the difference between Sqoop and DistCP command in Hadoop?

Answer)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.

12)What is Sqoop metastore?

Answer)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.

13)What is the significance of using –split-by clause for running parallel import tasks in Apache Sqoop?

Answer)--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.

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

Answer)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.

15)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?

Answer)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.

16)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?

Answer)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.

17)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?

Answer)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.

18)If the source data gets updated every now and then, how will you synchronise the data in HDFS that is imported by Sqoop?

Answer)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.

19)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?

Answer)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.

20)What are the relational databases supported in Sqoop?

Answer)Below are the list of RDBMSs that are supported by Sqoop Currently.
MySQL
PostGreSQL
Oracle
Microsoft SQL
IBM's Netezza
Teradata

21)What are the destination types allowed in Sqoop Import command?

Answer)Currently Sqoop Supports data imported into below services.
HDFS
Hive
HBase
HCatalog
Accumulo

22)Is Sqoop similar to distcp in hadoop?

Answer)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.

23)What are the majorly used commands in Sqoop?

Answer)In Sqoop Majorly Import and export commands are used. But below commands are also useful some times.
codegen
eval
import-all-tables
job
list-databases

list-tables
merge
metastore

24)While loading tables 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.

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

Answer)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.

26)What is the importance of eval tool?

Answer)It allow users to run sample SQL queries against Database and preview the result on the console.

27)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.

28)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.

29)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.

30)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.

31)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.

32)How will you update the rows that are already exported?

Answer)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.

33)What is the role of JDBC driver in a Sqoop set up?

Answer)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.

34)When to use --target-dir and when to use --warehouse-dir while importing data?

Answer)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.

35)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.

36)Is it possible to add a parameter while running a saved job?

Answer)Yes, we can add an argument to a saved job at runtime by using the --exec option
sqoop job --exec jobname -- -- newparameter

37)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?

Answer)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.

38)How will you implement all-or-nothing load using sqoop?

Answer)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.

39)How will you update the rows that are already exported?

Answer)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.

40)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.

41)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?

Answer)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.

42)How can you schedule a sqoop job using Oozie?

Answer)Oozie has in-built sqoop actions inside which we can mention the sqoop commands to be executed.

43)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?

Answer)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.

44)How can you force sqoop to execute a free form Sql query only once and import the rows serially.

Answer)By using the –m 1 clause in the import command, sqoop cerates only one mapreduce task which will import the rows sequentially.

45)In a sqoop import command you have mentioned to run 8 parallel Mapreduce task but sqoop runs only 4. What can be the reason?

Answer)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.

46)What happens when a table is imported into a HDFS directory which already exists using the –apend parameter?

Answer)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.

47)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?

Answer)By using the lastmodified mode. Rows where the check column holds a timestamp more recent than the timestamp specified with --last-value are imported.

48)What does the following query do?
**$ sqoop import --connect jdbc:mysql://host/dbname --table EMPLOYEES **
--where start_date > 2012-11-09

Answer)It imports the employees who have joined after 9-Nov-2012.

49)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.

Answer) $ sqoop import --connect jdbc:mysql://db.foo.com/corp --table EMPLOYEES \

--split-by dept_id

**50)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

Answer)It performs an incremental import of new data, after having already imported the first 1000 rows of a table