## Problem Statement:

Perform and explain the code flow and the associated result for the below tasks. Candidates should create and use their own employee dataset for the same. Share the screenshot of the commands used and its associated result.

- Transfer data between Mysql and HDFS (Import and Export) using Sqoop.
- Transfer data between Mysql and Hive (Import and Export only selected columns) using Sqoop.

**Solution:**

- **Transfer data between Mysql and HDFS (Import and Export) using Sqoop.**

--Starting MYSQL , then Creating database 'sqoop' and Using database'sqoop'

```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 15
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> create database sqoop;
Query OK, 1 row affected (0.00 sec)

mysql> use sqoop;
Database changed
```

--Creating Table 'employee' in mysql

```
mysql> create table employee(e_id int,e_name varchar(20),e_unit varchar(20));
Query OK, 0 rows affected (0.03 sec)
```

--Insert Data into 'employee' table in mysql

```
mysql> insert into employee values (1,'Nitisha','CA');
Query OK, 1 row affected (0.01 sec)

mysql> insert into employee values (2,'Rohit','CS');
Query OK, 1 row affected (0.01 sec)

mysql> insert into employee values (3,'Sonal','Engineer');
Query OK, 1 row affected (0.01 sec)
```

--Showing records of 'employee' table from mysql

```
mysql> select * from employee;
+------+---------+----------+
| e_id | e_name  | e_unit   |
+------+---------+----------+
|    1 | Nitisha | CA       |
|    2 | Rohit   | CS       |
|    3 | Sonal   | Engineer |
+------+---------+----------+
3 rows in set (0.00 sec)

mysql>
```

--Importing ‘employee’ table from MYSQL to HDFS using Sqoop

```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/sqoop --username 'root' -P --table 'employee' -m 1;
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/11/20 03:11:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/11/20 03:11:53 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/11/20 03:11:54 INFO tool.CodeGenTool: Beginning code generation
17/11/20 03:11:55 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee` AS t LIMIT 1
17/11/20 03:11:55 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee` AS t LIMIT 1
17/11/20 03:11:55 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/fa3ccb4479cd58ae1b2657a1508e913f/employee.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
17/11/20 03:12:02 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/fa3ccb4479cd58ae1b2657a1508e913f/employee.jar
17/11/20 03:12:02 WARN manager.MySQLManager: It looks like you are importing from mysql.
17/11/20 03:12:02 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
17/11/20 03:12:02 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
17/11/20 03:12:02 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
17/11/20 03:12:02 INFO mapreduce.ImportJobBase: Beginning import of employee
17/11/20 03:12:02 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
17/11/20 03:12:03 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
17/11/20 03:12:04 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
17/11/20 03:12:05 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
17/11/20 03:12:06 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
17/11/20 03:12:06 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
                Map-Reduce Framework
                        Map input records=3
                        Map output records=3
                        Input split bytes=87
                        Spilled Records=0
                        Failed Shuffles=0
                        Merged Map outputs=0
                        GC time elapsed (ms)=410
                        CPU time spent (ms)=3790
                        Physical memory (bytes) snapshot=132292608
                        Virtual memory (bytes) snapshot=1510178816
                        Total committed heap usage (bytes)=60751872
                File Input Format Counters
                        Bytes Read=0
                File Output Format Counters
                        Bytes Written=41
17/11/20 03:13:35 INFO mapreduce.ImportJobBase: Transferred 41 bytes in 90.4124 seconds (0.4535 bytes/sec)
17/11/20 03:13:35 INFO mapreduce.ImportJobBase: Retrieved 3 records.
[cloudera@quickstart ~]$

[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/employee
Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2017-11-20 03:13 /user/cloudera/employee/_SUCCESS
-rw-r--r--   1 cloudera cloudera         41 2017-11-20 03:13 /user/cloudera/employee/part-m-00000
[cloudera@quickstart ~]$
```

--Checking the data imported from MYSQL table ‘employee’ to HDFS file ‘/user/cloudera/employee’

```
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/employee/part-m-00000
1,Nitisha,CA
2,Rohit,CS
3,Sonal,Engineer
[cloudera@quickstart ~]$
```

--Now for exporting the data from HDFS to MYSQL, First Create a table ‘export_employee’ with same schema

```
mysql> create table export_employee(e_id int,e_name varchar(20),e_unit varchar(20));
Query OK, 0 rows affected (0.02 sec)

mysql>
```

--Exporting data from HDFS file ‘/user/cloudera/employee’ to new table ‘export_employee’ created in MYSQL

```
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/sqoop --username 'root' -P --table 'export_employee' -m 1 --export-dir '/user/cloudera/employee';

Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/11/20 03:28:50 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/11/20 03:42:08 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/11/20 03:42:08 INFO tool.CodeGenTool: Beginning code generation
17/11/20 03:42:10 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `export_employee` AS t LIMIT 1
17/11/20 03:42:10 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `export_employee` AS t LIMIT 1
17/11/20 03:42:10 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/d7315c127107ad079f5500ebb182365d/export_employee.java uses or overrides a deprecated API.

                Physical memory (bytes) snapshot=131395584
                Virtual memory (bytes) snapshot=1508028416
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
17/11/20 03:44:00 INFO mapreduce.ExportJobBase: Transferred 192 bytes in 89.8214 seconds (2.1376 bytes/sec)
17/11/20 03:44:00 INFO mapreduce.ExportJobBase: Exported 3 records.
[cloudera@quickstart ~]$
```

--Checking the data exported from HDFS file ‘/user/cloudera/employee’ to MYSQL new table ‘export_employee’

```
mysql> select * from export_employee;
+------+---------+-----------+
| e_id | e_name  | e_unit    |
+------+---------+-----------+
|    1 | Nitisha | CA        |
|    2 | Rohit   | CS        |
|    3 | Sonal   | Engineer  |
+------+---------+-----------+
3 rows in set (0.00 sec)

mysql>
```

- **Transfer data between Mysql and Hive (Import and Export only selected columns) using Sqoop.**

-- Creating new table ‘employee2’ in MYSQL

```
mysql> create table employee2(e_id int,e_name varchar(20),e_unit varchar(20));
Query OK, 0 rows affected (0.01 sec)
```

--Inserting data into the ‘employee2’ table

```
mysql> insert into employee2 values (1,'Nitisha','CA');
Query OK, 1 row affected (0.01 sec)

mysql> insert into employee2 values (2,'Rohit','CS');
Query OK, 1 row affected (0.00 sec)

mysql> insert into employee2 values (3,'Sonal','Engineer');
Query OK, 1 row affected (0.01 sec)

mysql> insert into employee2 values (4,'Shika','Test Engineer');
Query OK, 1 row affected (0.00 sec)

mysql> insert into employee2 values (5,'Pranshu','Senior Engineer');
Query OK, 1 row affected (0.01 sec)
```

-- Showing records of ‘employee2’ table from mysql

```
mysql> select * from employee2;
+------+---------+-------------------+
| e_id | e_name  | e_unit            |
+------+---------+-------------------+
|    1 | Nitisha | CA                |
|    2 | Rohit   | CS                |
|    3 | Sonal   | Engineer          |
|    4 | Shika   | Test Engineer     |
|    5 | Pranshu | Senior Engineer   |
+------+---------+-------------------+
5 rows in set (0.00 sec)

mysql>
```

Firstly creating database ‘sqoop’ in hive

```
[cloudera@quickstart ~]$ sudo hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
Time taken: 0.96 seconds, Fetched: 1 row(s)
hive> create database sqoop;
OK
Time taken: 5.379 seconds

hive> use sqoop;
OK
Time taken: 0.287 seconds
hive> set hive.cli.print.current.db='true';
```

--Importing ‘employee2’ table from MYSQL to Hive (only few columns) using Sqoop

```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/sqoop --username 'root' -P --table employee2 --split-by e_id --columns e_id,e_name --fields-termi
nated-by "," --target-dir /user/cloudera/employee2 --hive-import --create-hive-table --hive-table sqoop.employees --m 1;
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/11/20 07:46:53 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/11/20 07:47:00 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/11/20 07:47:00 INFO tool.CodeGenTool: Beginning code generation
17/11/20 07:47:04 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee2` AS t LIMIT 1
17/11/20 07:47:04 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee2` AS t LIMIT 1
17/11/20 07:47:04 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/3ac173b2aee18cc05fc8c1f7c01aa4b0/employee2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.

        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=44
17/11/20 07:47:58 INFO mapreduce.ImportJobBase: Transferred 44 bytes in 43.3951 seconds (1.0139 bytes/sec)
17/11/20 07:47:58 INFO mapreduce.ImportJobBase: Retrieved 5 records.
17/11/20 07:47:58 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee2` AS t LIMIT 1
17/11/20 07:47:58 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.12.0.jar!/hive-log4j.properties
OK
Time taken: 3.673 seconds
Loading data to table sqoop.employees
Table sqoop.employees stats: [numFiles=1, totalSize=44]
OK
Time taken: 2.282 seconds
[cloudera@quickstart ~]$
```

--Checking table in hive whether data is imported or not

```
hive> show tables;
OK
employees
Time taken: 0.304 seconds, Fetched: 1 row(s)
hive> select * from employees;
OK
1       Nitisha
2       Rohit
3       Sonal
4       Shika
5       Pranshu
Time taken: 2.339 seconds, Fetched: 5 row(s)
hive>
```

--Creating table ‘export_emp’ inserting data and showing data in Hive

```
hive> create table export_emp(e_id int,e_name varchar(20),e_unit varchar(20))row format delimited fields terminated by ',';
OK
Time taken: 0.486 seconds
hive> insert into export_emp values(1,'ram','tester');
Query ID = root_20171120081717_0b562ced-67ca-4da1-b8ee-3240a53a3190
Total jobs = 3
hive> select * from export_emp;
OK
1       ram     tester
2       sita    engineer
3       sonal   producer
4       sonali  director
5       rita    CA
Time taken: 0.308 seconds, Fetched: 5 row(s)
hive>
```

--Creating table ‘employee_export’ in MYSQL on which the data will be loaded

```
mysql> create table employee_export(e_id int,e_name varchar(20),e_unit varchar(20));
Query OK, 0 rows affected (0.01 sec)
```

--Validating the data in the MySQL ‘employee_export’ table;

```
mysql> select * from employee_export;
Empty set (0.00 sec)
```

--Exporting data (only selected columns) from HIVE to MYSQL

```
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/sqoop --username 'root' -P --table employee_export --columns e_name,e_unit --export-dir /user/hiv
e/warehouse/sqoop.db/export_emp;
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/11/20 08:32:21 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/11/20 08:32:27 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/11/20 08:32:27 INFO tool.CodeGenTool: Beginning code generation
17/11/20 08:32:31 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee_export` AS t LIMIT 1
17/11/20 08:32:31 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee_export` AS t LIMIT 1
17/11/20 08:32:31 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/2216beaef8f0587831725fddbef028ea/employee_export.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
                CPU time spent (ms)=4900
                Physical memory (bytes) snapshot=339099648
                Virtual memory (bytes) snapshot=4524195840
                Total committed heap usage (bytes)=182255616
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
17/11/20 08:34:05 INFO mapreduce.ExportJobBase: Transferred 794 bytes in 87.4924 seconds (9.0751 bytes/sec)
17/11/20 08:34:05 INFO mapreduce.ExportJobBase: Exported 5 records.
[cloudera@quickstart ~]$
```

--Checking whether data exported to  MYSQL

```
mysql> select * from employee_export;
+------+---------+---------+
| e_id | e_name  | e_unit  |
+------+---------+---------+
| NULL | 3       | sonal   |
| NULL | 4       | sonali  |
| NULL | 1       | ram     |
| NULL | 2       | sita    |
| NULL | 5       | rita    |
+------+---------+---------+
5 rows in set (0.04 sec)

mysql>
```