Qu 1) Suppose we have a test_db database in mysql. We have an input table Customers inside test_db. (SQL Commands are given)

Cust_Id	Customer_Name	Purchase_Date	Item	City	Price	Cust_Type
100	Rishi	2020-08-16	Mobile	Kanpur	10000	Regular
200	Venu	2019-05-04	Laptop	Bangalore	61000	Premium
300	Priya	2018-06-25	Mobile	Jaipur	20000	Premium
400	Rini	2019-01-30	Handbag	Pune	1000	Regular
700	Deepu	2019-12-12	Appliances	Mumbai	25000	Premium

The table has a Primary key on the Price column (which of course is not the right choice as prices may repeat when data grows).

Do the following: Share Snapshots of the command and Snapshot of the result in each case:

1) Before performing the sqoop import, using the sqoop command display the data present in mysql Customers table.The output of the command should not display on the console,rather should be redirected to log file named 'query.output'. Display the contents of the query.output file , share the Snapshot of the command and the output .

```
File Edit View Search Terminal Tabs Help
hduser@localhost:~                                  hduser@localhost:~
Try --help for usage instructions.
[hduser@localhost ~]$ sqoop-eval --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --query "select * from Customers" 1>
query.out 2>query.err
[hduser@localhost ~]$ cat query.err
22/06/24 19:15:19 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
22/06/24 19:15:20 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/06/24 19:15:20 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/phoenix-4.11.0-HBase-0.98-client.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registere
d via the SPI and manual loading of the driver class is generally unnecessary.
[hduser@localhost ~]$ cat query.out
Warning: /usr/local/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
-----------------------------------------------------------------------------------------------------------------------------
| Cuts_Id    | Customer_Name       | Purchase_Date       | Item                | City                | Price      | Cust_Type           |
-----------------------------------------------------------------------------------------------------------------------------
| 400        | Rini                | 2019-01-30          | Handbag             | Pune                | 1000       | Regular             |
| 100        | Rishi               | 2020-08-16          | Mobile              | Kanpur              | 10000      | Regular             |
| 300        | Priya               | 2018-06-25          | Mobile              | Jaipur              | 20000      | Premium             |
| 700        | Deepu               | 2019-12-12          | Appliances          | Mumbai              | 25000      | Premium             |
| 200        | Venu                | 2019-05-04          | Laptop              | Banglore            | 61000      | Premium             |
-----------------------------------------------------------------------------------------------------------------------------
[hduser@localhost ~]$ sqoop-eval --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --query "select * from Customers" 1>
```

2) Perform a single sqoop import inside the directory in hdfs named sqoop_importdir, considering all the following points:

- Import all the columns except Cust_Type in hdfs.
- Include only the purchases made after 2019-01-01
- The output data generated should have fields separated by | and rows separated by ; (semicolon)
- While importing, Nulls in the data , should be overridden with 'NA'

```
hduser@localhost ~]$ sqoop-import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$
-table Customers --columns Cuts_Id,Customer_Name,Purchase_Date,Item,City,Price --where "Purchase_Date>'2019-01-01
" --null-string "NA" --delete-target-dir --target-dir /user/hduser/sqoop_importdir2 1>log_out1.out 2>log_out2.err
```

- teRedirect the log messages generated on screen to the files log_out1 and log_out2. when sqoop import is successful,share the snapshot of the number of records retrieved.

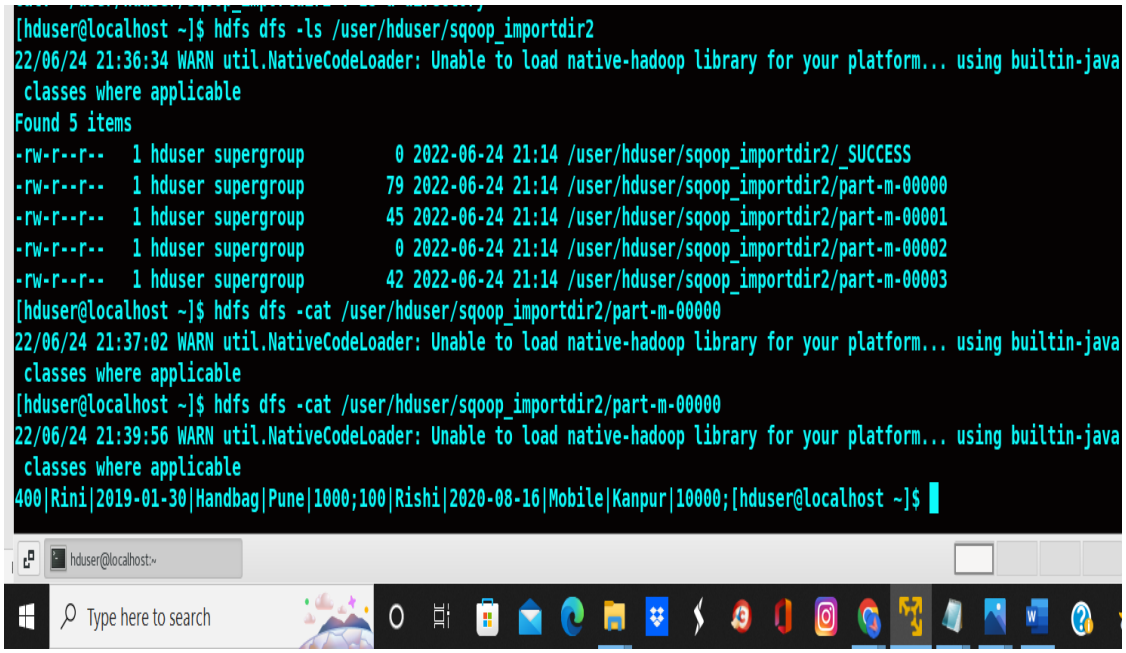
```
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:227)
        at org.apache.sqoop.Sqoop.main(Sqoop.java:236)

[hduser@localhost ~]$ sqoop-import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --table Customers --columns Cuts_Id,
Customer_Name,Purchase_Date,Item,City,Price --where "Purchase_Date>'2019-01-01'" --target-dir /user/hduser/sqoop_importdir2 1>log_out1.out 2>log_out2.
err --fields-terminated-by '|' --lines-terminated-by ';'
[hduser@localhost ~]$ cat log_out2.err
22/06/24 21:13:12 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
22/06/24 21:13:12 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/06/24 21:13:13 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/06/24 21:13:13 INFO tool.CodeGenTool: Beginning code generation
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/phoenix-4.11.0-HBase-0.98-client.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered
via the SPI and manual loading of the driver class is generally unnecessary.
22/06/24 21:13:15 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Customers` AS t LIMIT 1
22/06/24 21:13:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Customers` AS t LIMIT 1
22/06/24 21:13:16 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-hduser/compile/cad8610d58abe708a499e8d1cb191e30/Customers.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/06/24 21:13:20 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hduser/compile/cad8610d58abe708a499e8d1cb191e30/Customers.jar
22/06/24 21:13:20 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/06/24 21:13:20 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/06/24 21:13:20 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/06/24 21:13:20 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/06/24 21:13:21 INFO mapreduce.ImportJobBase: Beginning import of Customers
22/06/24 21:13:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/06/24 21:13:21 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
```

Display the contents of the log_out2 file
//Cat log_out2.err

```
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=177145
                Total vcore-seconds taken by all map tasks=177145
                Total megabyte-seconds taken by all map tasks=181396480
        Map-Reduce Framework
                Map input records=4
                Map output records=4
                Input split bytes=448
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=6515
                CPU time spent (ms)=12580
                Physical memory (bytes) snapshot=476368896
                Virtual memory (bytes) snapshot=8365154304
                Total committed heap usage (bytes)=186908672
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=166
22/06/24 21:14:34 INFO mapreduce.ImportJobBase: Transferred 166 bytes in 71.8213 seconds (2.3113 bytes/sec)
22/06/24 21:14:34 INFO mapreduce.ImportJobBase: Retrieved 4 records.
```

- Display the contents of the sqoop_importdir



```
[hduser@localhost ~]$ hdfs dfs -ls /user/hduser/sqoop_importdir2
22/06/24 21:36:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
Found 5 items
-rw-r--r--   1 hduser supergroup          0 2022-06-24 21:14 /user/hduser/sqoop_importdir2/_SUCCESS
-rw-r--r--   1 hduser supergroup         79 2022-06-24 21:14 /user/hduser/sqoop_importdir2/part-m-00000
-rw-r--r--   1 hduser supergroup         45 2022-06-24 21:14 /user/hduser/sqoop_importdir2/part-m-00001
-rw-r--r--   1 hduser supergroup          0 2022-06-24 21:14 /user/hduser/sqoop_importdir2/part-m-00002
-rw-r--r--   1 hduser supergroup         42 2022-06-24 21:14 /user/hduser/sqoop_importdir2/part-m-00003
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/sqoop_importdir2/part-m-00000
22/06/24 21:37:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/sqoop_importdir2/part-m-00000
22/06/24 21:39:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
400|Rini|2019-01-30|Handbag|Pune|1000;100|Rishi|2020-08-16|Mobile|Kanpur|10000;[hduser@localhost ~]$
```

- Now Again modify and run your sqoop import command ,so that cust_id column can be used to decide the input splits, as the Primary key column is not proper. Also ensure that the output directory remains as sqoop_importdir, and the previously imported contents are automatically deleted and new contents are filled in the output directory.

```
[hduser@localhost ~]$ sqoop-import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --table Customers --
columns Cuts_Id,Customer_Name,Purchase_Date,Item,City,Price --where "Purchase_Date>'2019-01-01'" --delete-target-dir --target-dir /use
r/hduser/sqoop_importdir2 1>log_out1.out 2>log_out2.err --fields-terminated-by '|' --lines-terminated-by ';' --split-by Cuts_Id
^C[hduser@localhost ~]$ sqoop-import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --table Customers
--columns Cuts_Id,Custom_Name,Purchase_Date,Item,City,Price --where "Purchase_Date>'2019-01-01'" --delete-target-dir --target-dir /use
r/hduser/sqoop_importdir2 1>log_out1.out 2>log_out2.err --fields-terminated-by '|' --lines-terminated-by ';' --split-by Cuts_Id
[hduser@localhost ~]$ cat log_out2.err
22/06/24 21:48:50 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
22/06/24 21:48:50 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/06/24 21:48:51 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/06/24 21:48:51 INFO tool.CodeGenTool: Beginning code generation
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinde
r.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/phoenix-4.11.0-HBase-0.98-client.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatic
ally registered via the SPI and manual loading of the driver class is generally unnecessary.
22/06/24 21:48:54 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Customers` AS t LIMIT 1
22/06/24 21:48:54 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Customers` AS t LIMIT 1
22/06/24 21:48:54 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-hduser/compile/821f811db5334ed74b3fe96a77946831/Customers.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/06/24 21:49:01 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hduser/compile/821f811db5334ed74b3fe96a77946831/Customers.
jar
22/06/24 21:49:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes whe
```

hduser@localhost:~

Type here to search

23°C

```
CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications  Places  Terminal                                   Fri 21:58
hduser@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
hduser@localhost:~                hduser@localhost:~
                CPU time spent (ms)=10680
                Physical memory (bytes) snapshot=444858368
                Virtual memory (bytes) snapshot=8365154304
                Total committed heap usage (bytes)=186908672
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=166
22/06/24 21:50:22 INFO mapreduce.ImportJobBase: Transferred 166 bytes in 78.8547 seconds (2.1051 bytes/sec)
22/06/24 21:50:23 INFO mapreduce.ImportJobBase: Retrieved 4 records.
[hduser@localhost ~]$ hdfs dfs -ls /user/hduser/sqoop_importdir2
22/06/24 21:50:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Found 5 items
-rw-r--r--   1 hduser supergroup          0 2022-06-24 21:50 /user/hduser/sqoop_importdir2/_SUCCESS
-rw-r--r--   1 hduser supergroup         83 2022-06-24 21:50 /user/hduser/sqoop_importdir2/part-m-00000
-rw-r--r--   1 hduser supergroup          0 2022-06-24 21:50 /user/hduser/sqoop_importdir2/part-m-00001
-rw-r--r--   1 hduser supergroup         38 2022-06-24 21:50 /user/hduser/sqoop_importdir2/part-m-00002
-rw-r--r--   1 hduser supergroup         45 2022-06-24 21:50 /user/hduser/sqoop_importdir2/part-m-00003
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/sqoop_importdir2/part-m-00000
22/06/24 21:51:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/sqoop_importdir2/part-m-00000
22/06/24 21:53:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/sqoop_importdir2/part-m-00000
22/06/24 21:58:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/sqoop_importdir2/part-m-00000
22/06/24 21:58:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
100|Rishi|2020-08-16|Mobile|Kanpur|10000;200|Venu|2019-05-04|Laptop|Banglore|61000;[hduser@localhost ~]$
hduser@localhost:~
Type here to search                                            23°C
```

- Display the contents of the output directory now and the first 10 records from the mapper output files (hint: use head command)

```
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/sqoop_importdir2/part-m-00000 | head
22/06/25 14:34:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
100|Rishi|2020-08-16|Mobile|Kanpur|10000;200|Venu|2019-05-04|Laptop|Banglore|61000;[hduser@localhost ~]$ hdfs dfs -cat /us
```

- Now Suppose an outlier comes into the mysql table:

The new record inserted is :

Cust_Id	Customer_Name	Purchase_Date	Item	City	Price	Cust_Type
10000	Raman	2019/09/04	Misc	Cochin	9000	Regular

Mention the sqoop import command you will frame from your end to deal with such a situation to ensure even work distribution among mappers, using customized bounding val query.

Note: you got to know that cust_id 10000 is erroneous record and should not be taken care.

```
22/06/25 13:30:46 INFO mapreduce.ImportJobBase: Retrieved 0 records.
[hduser@localhost ~]$ sqoop-import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --table
Customers --warehouse-dir /user/hduser/bound2 --boundary-query "select min(Cuts_id),max(cuts_id) from Customers where Cuts
_Id<10000" --delete-target-dir --split-by 'Cuts_Id'
```

```
                Other local map tasks=4
                Total time spent by all maps in occupied slots (ms)=137178
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=137178
                Total vcore-seconds taken by all map tasks=137178
                Total megabyte-seconds taken by all map tasks=140470272
        Map-Reduce Framework
                Map input records=5
                Map output records=5
                Input split bytes=449
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=783
                CPU time spent (ms)=7680
                Physical memory (bytes) snapshot=473559040
                Virtual memory (bytes) snapshot=8372105216
                Total committed heap usage (bytes)=186908672
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=247
22/06/25 13:38:31 INFO mapreduce.ImportJobBase: Transferred 247 bytes in 63.2237 seconds (3.9068 bytes/sec)
22/06/25 13:38:31 INFO mapreduce.ImportJobBase: Retrieved 5 records.
[hduser@localhost ~]$
```

```
s where applicable
ls: `/bound2': No such file or directory
[hduser@localhost ~]$ hdfs dfs -ls /bound2
22/06/25 13:45:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
^[[Als: `/bound2': No such file or directory
[hduser@localhost ~]$ hdfs dfs -ls /user/hduser/bound2/Customers
22/06/25 13:45:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Found 5 items
-rw-r--r--   1 hduser supergroup          0 2022-06-25 13:38 /user/hduser/bound2/Customers/_SUCCESS
-rw-r--r--   1 hduser supergroup         99 2022-06-25 13:38 /user/hduser/bound2/Customers/part-m-00000
-rw-r--r--   1 hduser supergroup         49 2022-06-25 13:38 /user/hduser/bound2/Customers/part-m-00001
-rw-r--r--   1 hduser supergroup         46 2022-06-25 13:38 /user/hduser/bound2/Customers/part-m-00002
-rw-r--r--   1 hduser supergroup         53 2022-06-25 13:38 /user/hduser/bound2/Customers/part-m-00003
[hduser@localhost ~]$ hdfs dfs -ls /user/hduser/bound2/Customers/part-m-00000
22/06/25 13:45:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
-rw-r--r--   1 hduser supergroup         99 2022-06-25 13:38 /user/hduser/bound2/Customers/part-m-00000
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/bound2/Customers/part-m-00000
22/06/25 13:45:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
100,Rishi,2020-08-16,Mobile,Kanpur,10000,Regular
200,Venu,2019-05-04,Laptop,Banglore,61000,Premium
[hduser@localhost ~]$
```

==========

Qu 2) Suppose we have a database named test_new_db in mysql, We have three tables inside it:

City_Tbl (Consider this is the bigger table)
State_Tbl (Consider this is the smaller table)
Country_Tbl (Smaller Table)

City_Tbl: City_ID is the Primary Key Column

City_Name	City_ID
Bangalore	1000
Mumbai	1001
Chennai	1002
Kolkata	1003
Delhi	1004
Pune	1005
Nagpur	1006
Surat	1007
Kochi	1008

State_Tbl: No Primary Key Column

State_Name	Districts
Karnataka	30
TamilNadu	32
Goa	2
Kerala	14
Assam	33

Country_Tbl: No Primary Key Column

Name	Country_Code
Belgium	32
Brazil	55
France	33
Iran	98
India	91

A) Using a single sqoop import command,
Import all the tables present in test_new_db to hdfs excluding the Country_Tbl .
You have to do it with a single sqoop command.

Also, City_Tbl should have 3 output files generated in hdfs. All the output files
should be stored inside sqoop_all_tbl directory in hdfs, with sub-directories of each table name created inside the main directory. Share the snapshot of the command.

```
[hduser@localhost ~]$ sqoop-import-all-tables --connect jdbc:mysql://localhost/test_new_db --username root --passw
ord Root123$ --exclude-tables Country_tbl --warehouse-dir /user/hduser/sqoop_all_table --autoreset-to-one-mapper;
Warning: /usr/local/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/06/25 16:32:27 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
22/06/25 16:32:27 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -
P instead.
22/06/25 16:32:28 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/im
pl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/phoenix-4.11.0-HBase-0.98-client.jar!/org/slf4j/impl/Static
```

B) Show the contents of the output directory: (Share Snapshot)

```
22/06/25 16:44:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
tin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2022-06-25 16:43 /user/hduser/sqoop_all_table2/City_tbl
drwxr-xr-x   - hduser supergroup          0 2022-06-25 16:43 /user/hduser/sqoop_all_table2/State_tbl
[hduser@localhost ~]$ hdfs dfs -ls /user/hduser/sqoop_all_table2/City_tbl
22/06/25 16:44:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
tin-java classes where applicable
Found 5 items
-rw-r--r--   1 hduser supergroup          0 2022-06-25 16:43 /user/hduser/sqoop_all_table2/City_tbl/_SUCCESS
-rw-r--r--   1 hduser supergroup         26 2022-06-25 16:43 /user/hduser/sqoop_all_table2/City_tbl/part-m-00000
-rw-r--r--   1 hduser supergroup         26 2022-06-25 16:43 /user/hduser/sqoop_all_table2/City_tbl/part-m-00001
-rw-r--r--   1 hduser supergroup         21 2022-06-25 16:43 /user/hduser/sqoop_all_table2/City_tbl/part-m-00002
-rw-r--r--   1 hduser supergroup         34 2022-06-25 16:43 /user/hduser/sqoop_all_table2/City_tbl/part-m-00003
[hduser@localhost ~]$ hdfs dfs -ls /user/hduser/sqoop_all_table2/State_tbl
22/06/25 16:44:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
tin-java classes where applicable
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2022-06-25 16:43 /user/hduser/sqoop_all_table2/State_tbl/_SUCCESS
-rw-r--r--   1 hduser supergroup         48 2022-06-25 16:43 /user/hduser/sqoop_all_table2/State_tbl/part-m-00000
[hduser@localhost ~]$
```

==========
==========

Qu 3) We have a Categories Table in test_db in Mysql. On this table both inserts and updates are performed from time to time.

Do the following:

A) Import the Categories table in hdfs but during the import ,do proper Null value handling:

- String Columns nulls should be replaced with '\N' (so that in file it should be read as \n and Non-string column nulls should be replaced with -1
- Use a warehouse directory
- We also want to see the query run by each mapper internally

Share the import command you will use,keeping in mind all of the above. Initially all records to be pulled in.

```
-rw-r--r--   1 hduser supergroup        103 2022-06-25 18:26 /user/hduser/assign3q/Categories/part
-m-00003
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/assign3q/Categories/part-m-00000
22/06/25 18:28:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
rm... using builtin-java classes where applicable
2,1,Football,2020-04-30 00:00:00.0
2,2,Handball,2020-05-01 00:00:00.0
2,3,Baseball & Handball,2020-05-01 00:00:00.0
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/assign3q/Categories/part-m-00001
22/06/25 18:28:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
rm... using builtin-java classes where applicable
2,4,Basketball,2020-04-30 00:00:00.0
3,5,Tennis,2020-04-30 00:00:00.0
3,6,Hockey,2020-05-01 00:00:00.0
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/assign3q/Categories/part-m-00002
22/06/25 18:28:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
rm... using builtin-java classes where applicable
3,7,Swimming,2020-05-01 00:00:00.0
3,8,Cardio Equipment,2020-05-01 00:00:00.0
4,9,Strength Training,2020-05-01 00:00:00.0
[hduser@localhost ~]$ hdfs dfs -cat /user/hduser/assign3q/Categories/part-m-00003
22/06/25 18:28:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
rm... using builtin-java classes where applicable
4,10,Athletics,2020-05-01 00:00:00.0
-1,11,Cycling,2020-02-02 00:00:00.0
5,12,NA,2020-01-15 00:00:00.0
[hduser@localhost ~]$
```

```
[hduser@localhost ~]$ sqoop-import --connect "jdbc:mysql://localhost/test_db" --username root --pa
ssword Root123$ --table Categories --null-string "NA" --null-non-string "-1" --warehouse-dir /user
/hduser/assign3q --verbose
```

B) New Records are added to the table and also existing records are updated,(refer the mysql_commands text file for the insert and update commands), so import only those newly inserted/updated records from Categories table to hdfs.
The delta records should get appended to existing directory.

Share the import command you will use this time, to get only delta records

Inserted and updated table:

```
CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications  Places  Terminal                                   Sat 19:11
hduser@localhost:~
File  Edit  View  Search  Terminal  Help
+-------------+----------------------+--------------------+--------------------+
| category_id | category_department_id | category_name      | inclusion_date     |
+-------------+----------------------+--------------------+--------------------+
           1 |                    2 | Football           | 2020-04-30 00:00:00 |
           2 |                    2 | Handball           | 2020-05-01 00:00:00 |
           3 |                    2 | Baseball & Handball | 2020-05-01 00:00:00 |
           4 |                    2 | shuttle            | 2021-05-01 00:00:00 |
           5 |                    3 | Tennis             | 2020-04-30 00:00:00 |
           6 |                    3 | Hockey             | 2020-05-01 00:00:00 |
           7 |                    3 | Swimming           | 2020-05-01 00:00:00 |
           8 |                    3 | Cardio Equipment   | 2020-05-01 00:00:00 |
           9 |                    4 | Strength Training  | 2020-05-01 00:00:00 |
          10 |                    4 | Athletics          | 2020-05-01 00:00:00 |
          11 |                 NULL | Cycling            | 2020-02-02 00:00:00 |
          12 |                    5 | NULL               | 2020-01-15 00:00:00 |
          13 |                    6 | Throwball          | 2020-06-09 00:00:00 |
+-------------+----------------------+--------------------+--------------------+
```

```
[hduser@localhost ~]$ sqoop-import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$
--table Categories --warehouse-dir /user/hduser/assign3qb --incremental lastmodified --check-column inclusion_date
 --last-value 2020-06-09 --append
```

C) After this second import, how many records do you see in the hdfs folder now? Did you find any duplicate records, give details if any.

--->2 records retrieved

--->No

D) Create a new table in test_db named

Categories_new. This newly created table does not have

a Primary key.

We want to do periodic imports and updates in this mysql table. But we do not want any duplicate records in the hdfs post import.
Also we want to automate the process of import & want a good way to manage the password. Choose a different warehouse directory this time.

```
Root123$[hduser@localhsqoop job --create jobpass3 -- import --connect "jdbc:mysql://localhost/test_db" --username root --password
-file file:///home/hduser/.passwordfile --table Categories_new --warehouse-dir /user/hduser/Keerthupass --incremental lastmodifie
d  --check-column inclusion_date --last-value 2020-01-14 -m 1 --append
```

Share the commands you will use when:

- First time we need to pull all records in hdfs.

```
[hduser@localhost ~]$ sqoop job --create jobassignment -- import --connect "jdbc:mysql://localhost/test_db" --username root
--table Categories_new --warehouse-dir /user/hduser/Keerassign --incremental lastmodified  --check-column inclusion_date --l
ast-value 2020-01-14 -m 1 --append
```

- Second time to pull only the delta records,but without duplicates in hdfs

```
[hduser@localhost ~]$ sqoop job --create jobkeer3 -- import --connect "jdbc:mysql://localhost/test_db" --username root --pas
sword Root123$ --table Categories_new --null-string "NA" --null-non-string -1 --warehouse-dir /user/hduser/Assignkeerthu1 --
incremental lastmodified  --check-column inclusion_date --last-value 2021-06-01 -m 1 --merge-key category_id
```

E) How many records do you see this time in hdfs post second import? Do you see any duplicate records now? 1,No
F) Are any mapper files generated in hdfs this time after the second import? Explain. No

G) Share the command you will use to see the last value of a Saved Sqoop Job.

```
[hduser@localhost ~]$ sqoop job --show jobkeer4
```

==========

sqoop Quiz

1. Sqoop written in?

A. C
B. C++
C. Java
D. hadoop

2. Sqoop stands for?

A. SQL to Hadoop
B. SQL to Hbase
C. MySQL to Hadoop
D. SQL Hadoop

3. Is Apache Sqoop is an open-source tool?

A. TRUE
B. FALSE
C. Can be true or false
D. Can not say

4. Data processed by Scoop can be used for?

A. Hbase
B. HDFS
C. Mapreduce
D. MahOut

5. _____tool can list all the available database schemas

A. sqoop-list-tables
B. sqoop-list-databases
C. sqoop-list-schema
D. sqoop-list-columns
.

6. The active Hadoop configuration is loaded from $HADOOP_HOME/conf/, unless the $HADOOP_CONF_DIR environment variable is unset.

A. TRUE
B. FALSE
C. Can be true or false
D. Can not say
.

7. Data can be imported in maximum_____file formats.

A. 2
B. 3
C. 4
D. 5
.

8. If you set the inline LOB limit to_____all large objects will be placed in external storage.

A. 0
B. 2
C. 3
D. 1
.

9. The import-tables tool imports a set of tables from an RDBMS to?

A. Hive
B. Sqoop
C. HDFS
D. Mapreduce
.

10. Sqoop can also import the data into Hive by generating and executing a___statement to define the data's layout in Hive.

A. SET TABLE
B. CREATE TABLE
C. INSERT TABLE
D. All of the above

11. The following tool imports a set of tables from an RDBMS to HDFS
A. export-all-tables
B. import-all-tables
C. import-tables
D. none of the mentioned

12. With the –staging-table parameter, the data is moved from staging to final table

A. Automatically if staging load is successful
B. Has to be done by user after verifying the data in staging
C. Depends on the data size
D. Depends on the memory available to move the data