Qu 1) Suppose we have a test_db database in mysql. We have an input table

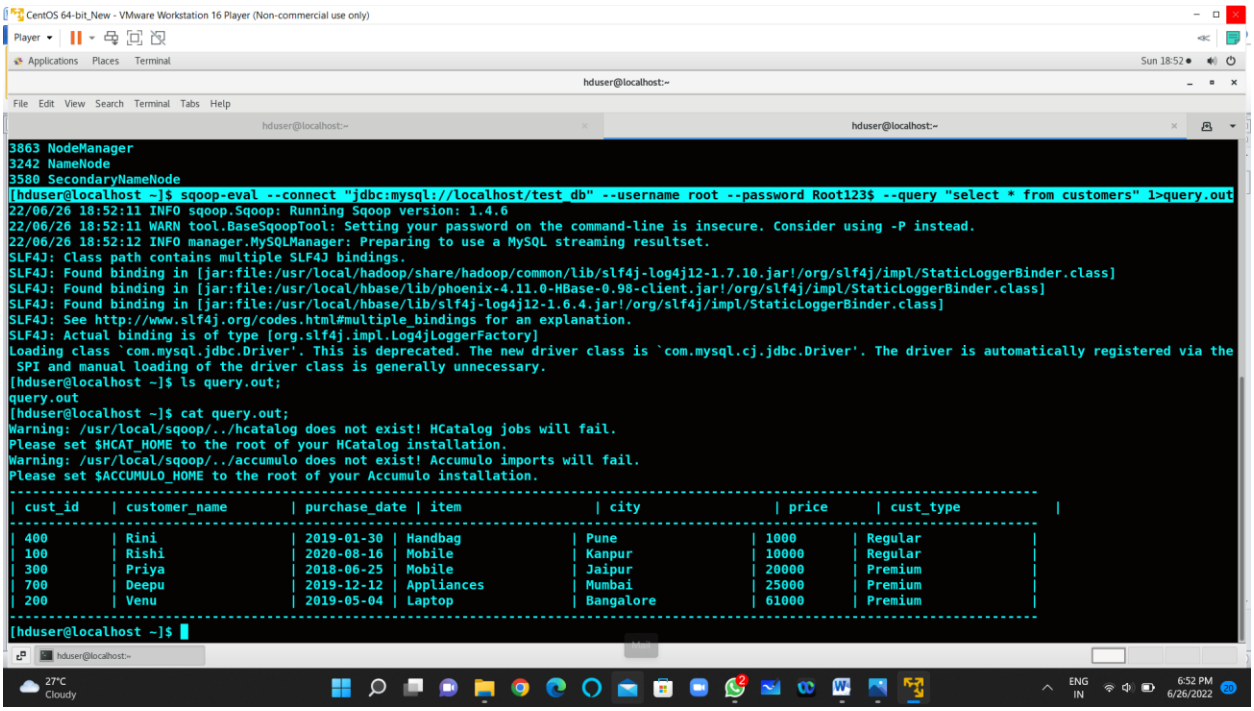
Customers inside test_db. (SQL Commands are given)

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).

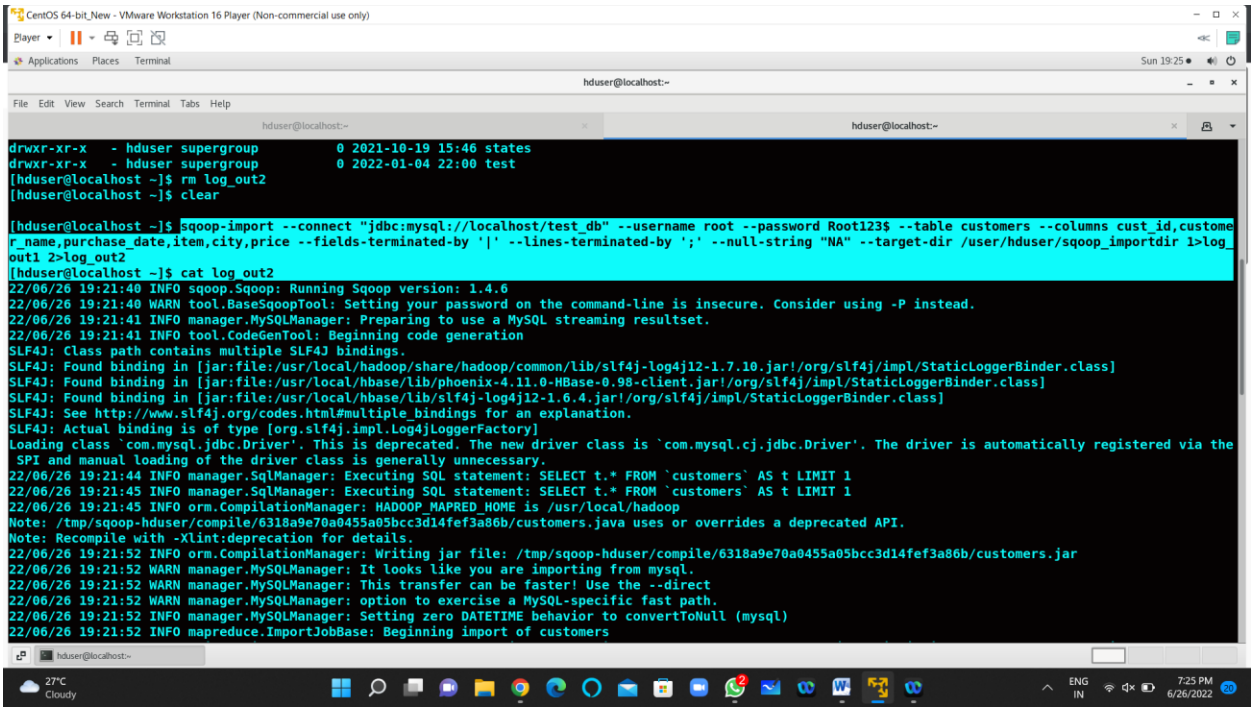
Do the following: Share Snapshots of the command and Snapshot of the result in each case:

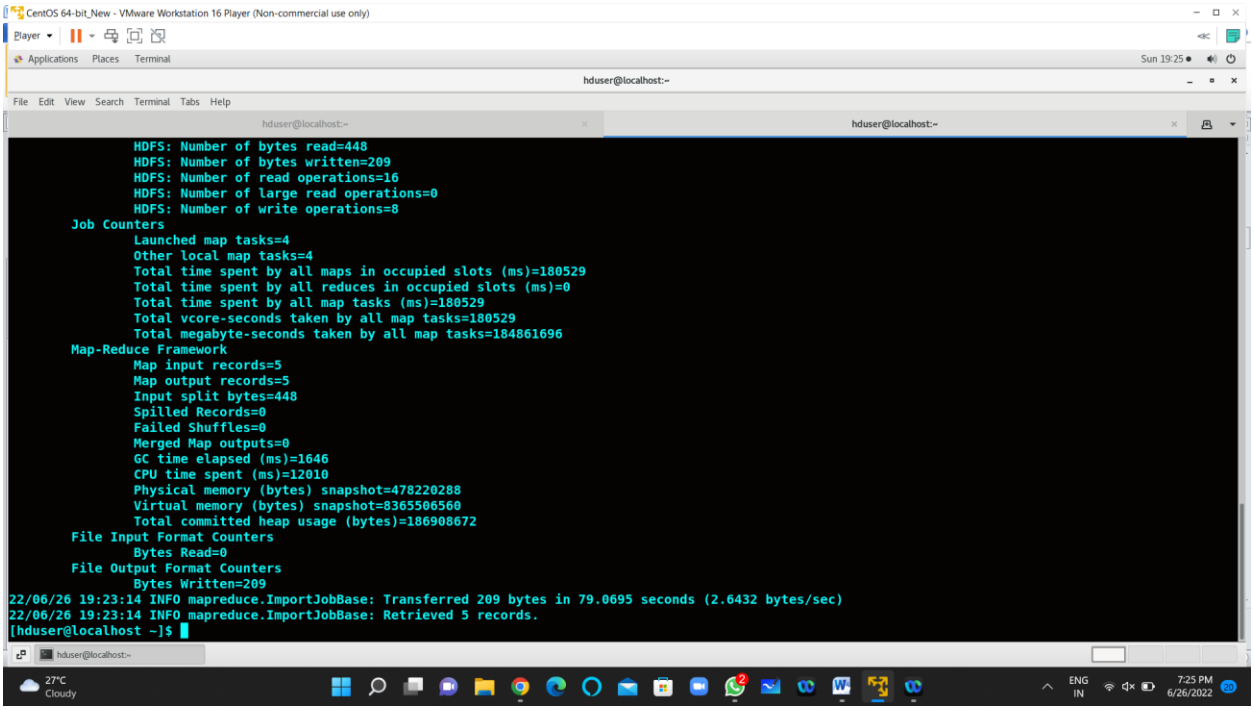
1) Before performing the sqoop import, using the sqoop command display the data present in mysql Customers table.The output of the command should not display on the console,rather should be redirected to log file named ‘query.output’. Display the contents of the query.output file , share the Snapshot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

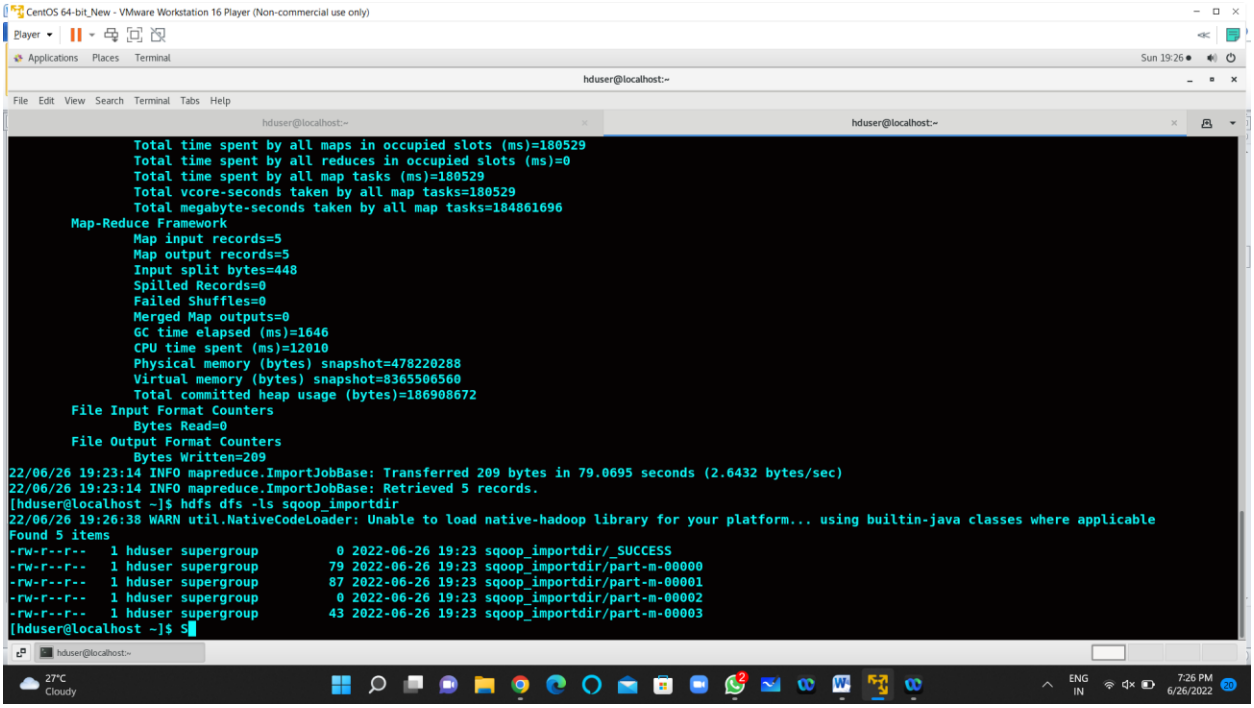


2) Perform a single sqoop import inside the directory in hdfs named sqoop_importdir, considering all the following points:

- 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’
- Redirect the log messages generated on screen to the files log_out1 and log_out2.Display the contents of the log_out2 file , when sqoop import is successful,share the snapshot of the number of records retrieved.
- Display the contents of the sqoop_importdir

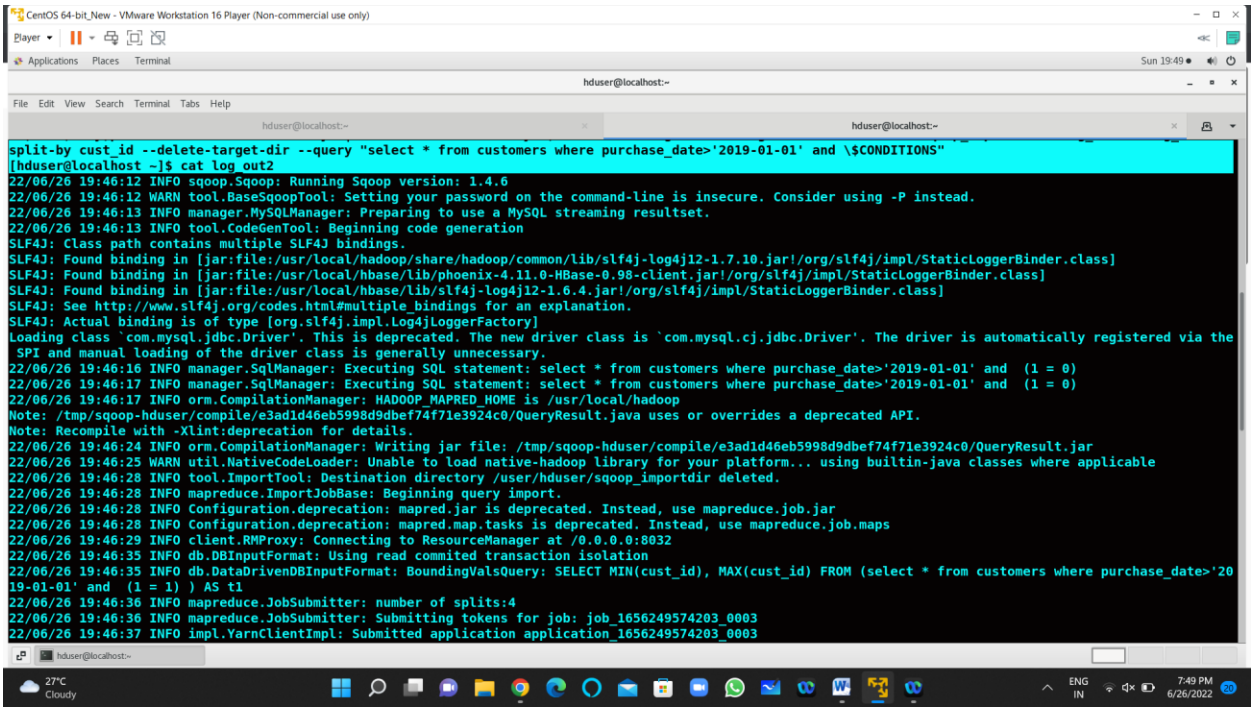

CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 19:25
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
drwxr-xr-x - hduser supergroup 0 2021-10-19 15:46 states
drwxr-xr-x - hduser supergroup 0 2022-01-04 22:00 test
[hduser@localhost ~]$ rm log_out2
[hduser@localhost ~]$ clear
[hduser@localhost ~]$ sqoop-import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --table customers --columns cust_id,customer_name,purchase_date,item,city,price --fields-terminated-by '|' --lines-terminated-by ';' --null-string "NA" --target-dir /user/hduser/sqoop_importdir 1>log_out1 2>log_out2
[hduser@localhost ~]$ cat log_out2
22/06/26 19:21:40 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
22/06/26 19:21:40 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/06/26 19:21:41 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/06/26 19:21:41 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.
22/06/26 19:21:44 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `customers` AS t LIMIT 1
22/06/26 19:21:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `customers` AS t LIMIT 1
22/06/26 19:21:45 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-hduser/compile/6318a9e70a0455a05bcc3d14fef3a86b/customers.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/06/26 19:21:52 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hduser/compile/6318a9e70a0455a05bcc3d14fef3a86b/customers.jar
22/06/26 19:21:52 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/06/26 19:21:52 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/06/26 19:21:52 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/06/26 19:21:52 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/06/26 19:21:52 INFO mapreduce.ImportJobBase: Beginning import of customers
27°C Cloudy
ENG IN
7:25 PM 6/26/2022

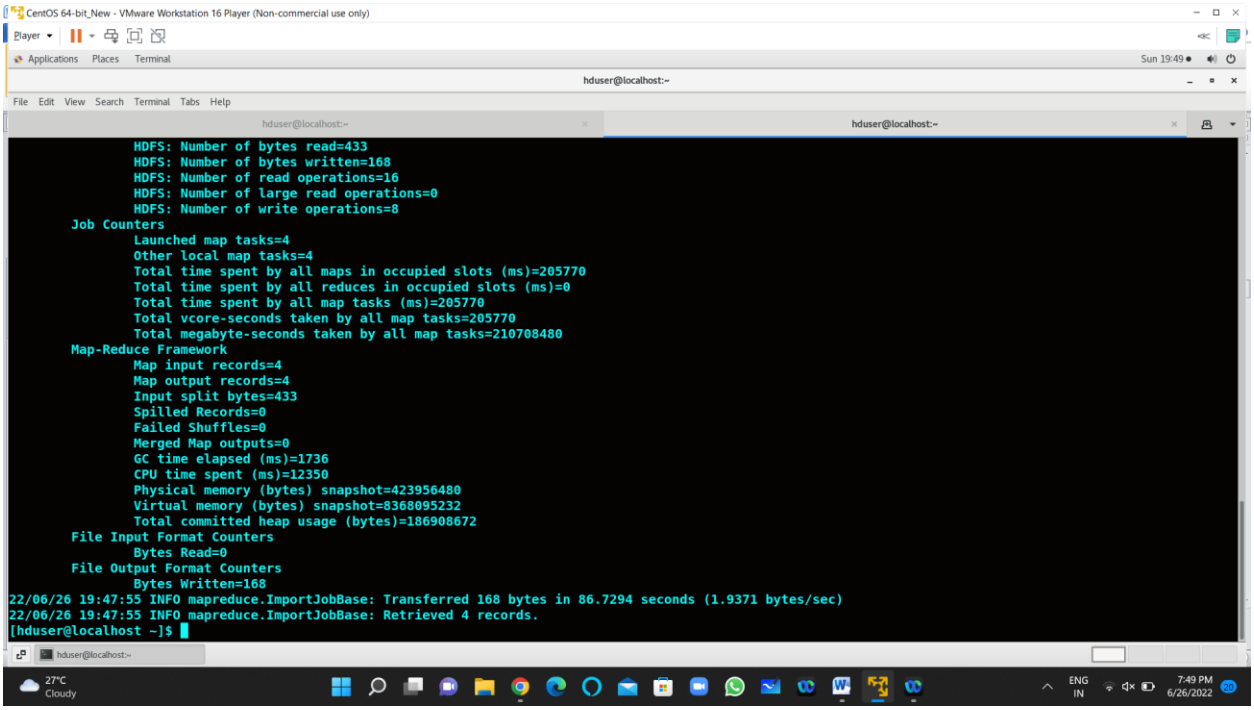

CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 19:25
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
HDFS: Number of bytes read=448
HDFS: Number of bytes written=209
HDFS: Number of read operations=16
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Job Counters
Launched map tasks=4
Other local map tasks=4
Total time spent by all maps in occupied slots (ms)=180529
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=180529
Total vcore-seconds taken by all map tasks=180529
Total megabyte-seconds taken by all map tasks=184861696
Map-Reduce Framework
Map input records=5
Map output records=5
Input split bytes=448
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=1646
CPU time spent (ms)=12010
Physical memory (bytes) snapshot=478220288
Virtual memory (bytes) snapshot=8365506560
Total committed heap usage (bytes)=186908672
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=209
22/06/26 19:23:14 INFO mapreduce.ImportJobBase: Transferred 209 bytes in 79.0695 seconds (2.6432 bytes/sec)
22/06/26 19:23:14 INFO mapreduce.ImportJobBase: Retrieved 5 records.
[hduser@localhost ~]$
27°C Cloudy
ENG IN
7:25 PM 6/26/2022

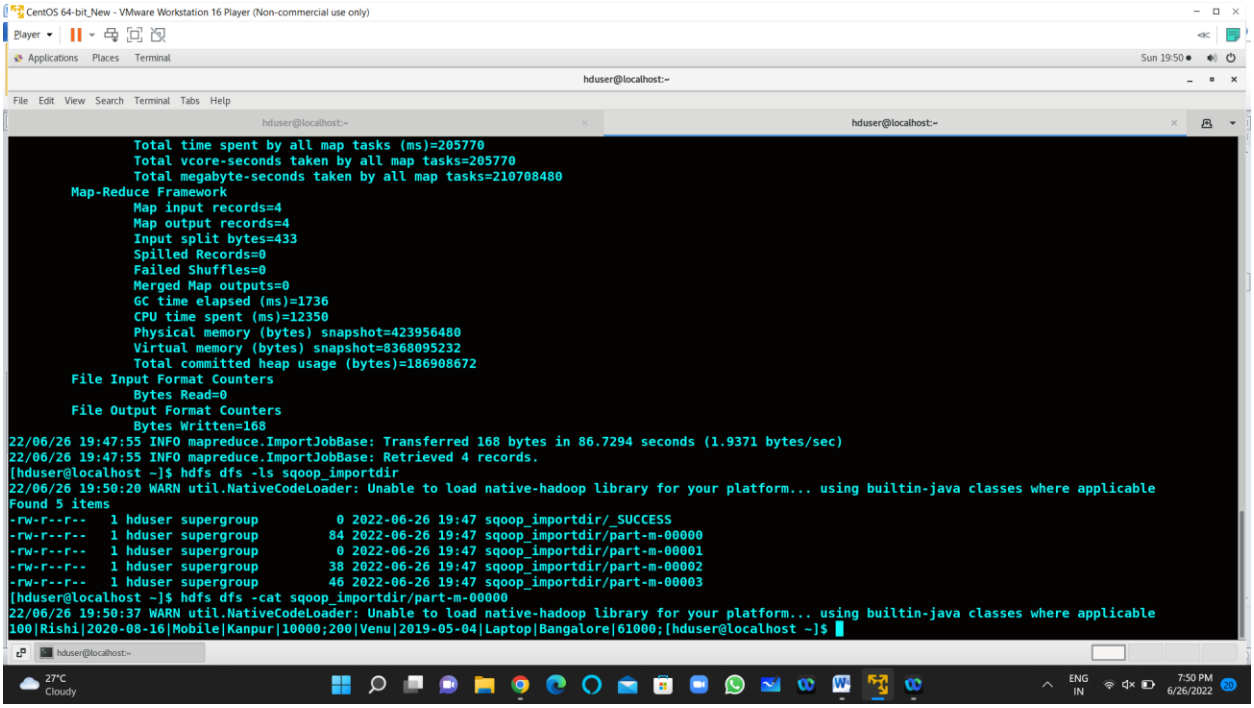


- Now Again modify and run your sqoop import command ,so that cust_id column can be used to decide the input splits, as the Primary key column is not proper. Also ensure that the output directory remains as sqoop_importdir, and the previously imported contents are automatically deleted and new contents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

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


CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 19:49
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
split-by cust_id --delete-target-dir --query "select * from customers where purchase_date>'2019-01-01' and \$CONDITIONS"
[hduser@localhost ~]$ cat log_out2
22/06/26 19:46:12 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
22/06/26 19:46:12 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/06/26 19:46:13 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/06/26 19:46:13 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.
22/06/26 19:46:16 INFO manager.SqlManager: Executing SQL statement: select * from customers where purchase_date>'2019-01-01' and (1 = 0)
22/06/26 19:46:17 INFO manager.SqlManager: Executing SQL statement: select * from customers where purchase_date>'2019-01-01' and (1 = 0)
22/06/26 19:46:17 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-hduser/compile/e3ad1d46eb5998d9dbef74f71e3924c0/QueryResult.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/06/26 19:46:24 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hduser/compile/e3ad1d46eb5998d9dbef74f71e3924c0/QueryResult.jar
22/06/26 19:46:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/06/26 19:46:28 INFO tool.ImportTool: Destination directory /user/hduser/sqoop_importdir deleted.
22/06/26 19:46:28 INFO mapreduce.ImportJobBase: Beginning query import.
22/06/26 19:46:28 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/06/26 19:46:28 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/06/26 19:46:29 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/06/26 19:46:35 INFO db.DBInputFormat: Using read commited transaction isolation
22/06/26 19:46:35 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(cust_id), MAX(cust_id) FROM (select * from customers where purchase_date>'2019-01-01' and (1 = 1)) AS t1
22/06/26 19:46:36 INFO mapreduce.JobSubmitter: number of splits:4
22/06/26 19:46:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1656249574203_0003
22/06/26 19:46:37 INFO impl.YarnClientImpl: Submitted application application_1656249574203_0003
hduser@localhost:~
27°C Cloudy
ENG IN
7:49 PM 6/26/2022


CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 19:49
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
HDFS: Number of bytes read=433
HDFS: Number of bytes written=168
HDFS: Number of read operations=16
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Job Counters
Launched map tasks=4
Other local map tasks=4
Total time spent by all maps in occupied slots (ms)=205770
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=205770
Total vcore-seconds taken by all map tasks=205770
Total megabyte-seconds taken by all map tasks=210708480
Map-Reduce Framework
Map input records=4
Map output records=4
Input split bytes=433
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=1736
CPU time spent (ms)=12350
Physical memory (bytes) snapshot=423956480
Virtual memory (bytes) snapshot=8368095232
Total committed heap usage (bytes)=186908672
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=168
22/06/26 19:47:55 INFO mapreduce.ImportJobBase: Transferred 168 bytes in 86.7294 seconds (1.9371 bytes/sec)
22/06/26 19:47:55 INFO mapreduce.ImportJobBase: Retrieved 4 records.
[hduser@localhost ~]$
hduser@localhost:~
27°C Cloudy
ENG IN
7:49 PM 6/26/2022



==========

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,

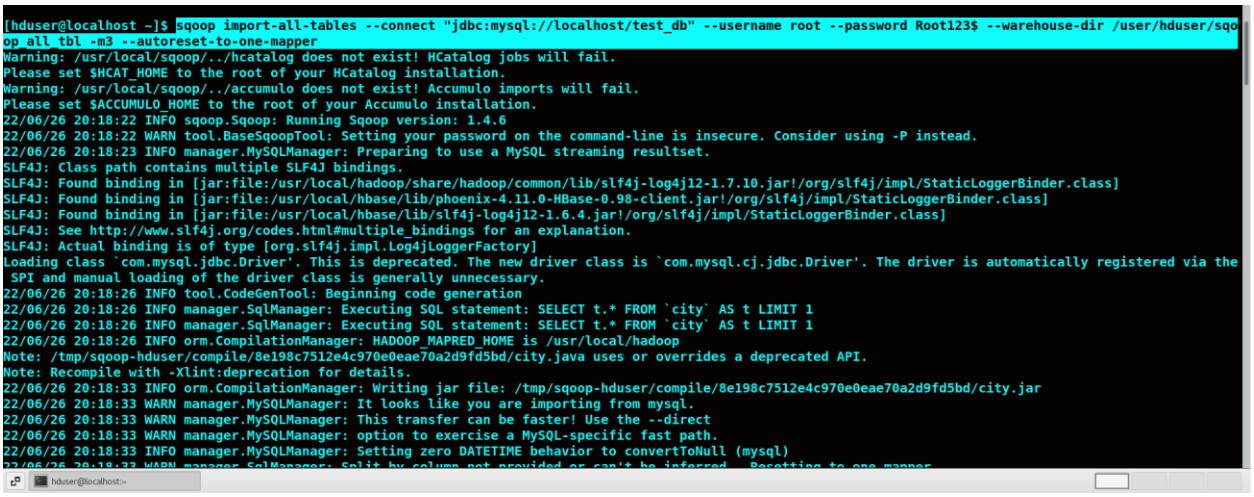
Import all the tables present in test_new_db to hdfs excluding the Country_Tbl .

You have to do it with a single sqoop command.

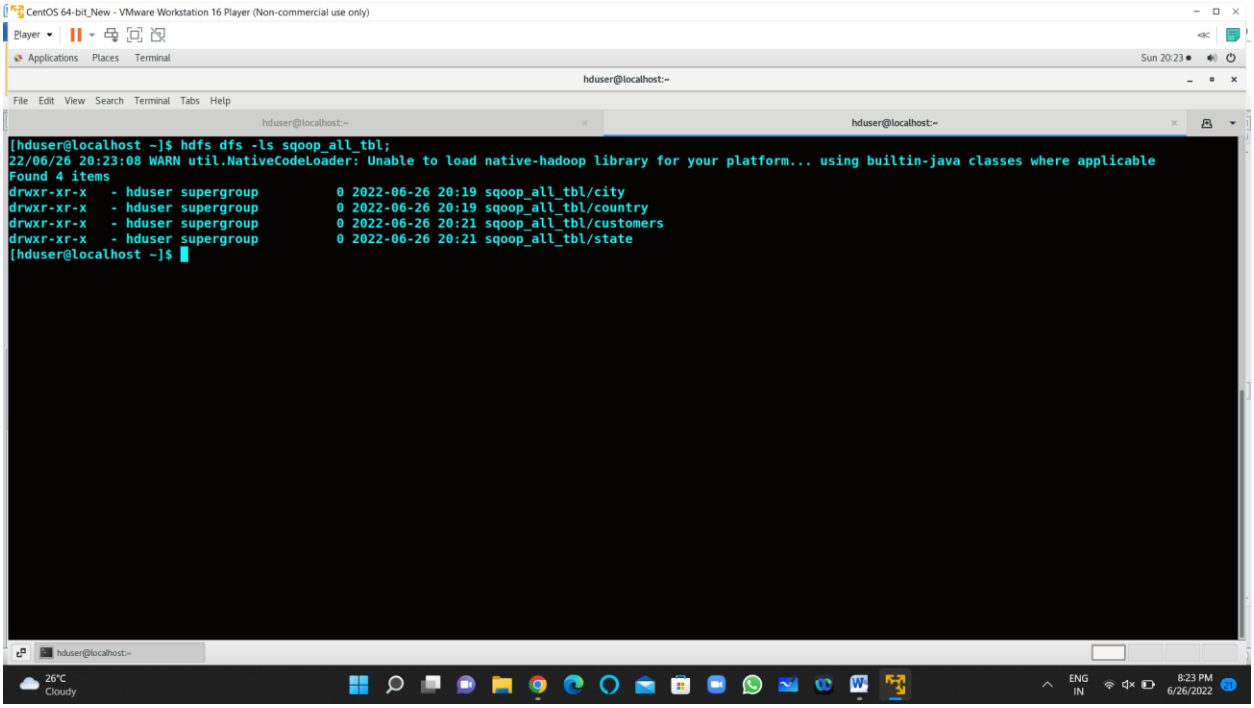
Also, City_Tbl should have 3 output files generated in hdfs. All the output files

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.



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



==========

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.

CREATE TABLE Categories (

category_id INT(11) PRIMARY KEY AUTO_INCREMENT,

category_department_id INT(11),

category_name VARCHAR(45),

inclusion_date datetime NOT NULL);

INSERT INTO Categories values

(1,2,'Football','2020-04-30'),

(2,2,'Handball','2020-05-01'),

(3,2,'Baseball & Softball','2020-05-01'),

(4,2,'Basketball','2020-04-30'),

(5,3 ,'Tennis','2020-04-30'),

(6,3,'Hockey','2020-05-01'),

(7,3,'Swimming','2020-05-01'),

(8,3,'Cardio Equipment','2020-05-01'),

(9,4,'Strength Training','2020-05-01'),

(10,4,'Athletics','2020-05-02'),

(11,null,'Cycling','2020-02-02'),

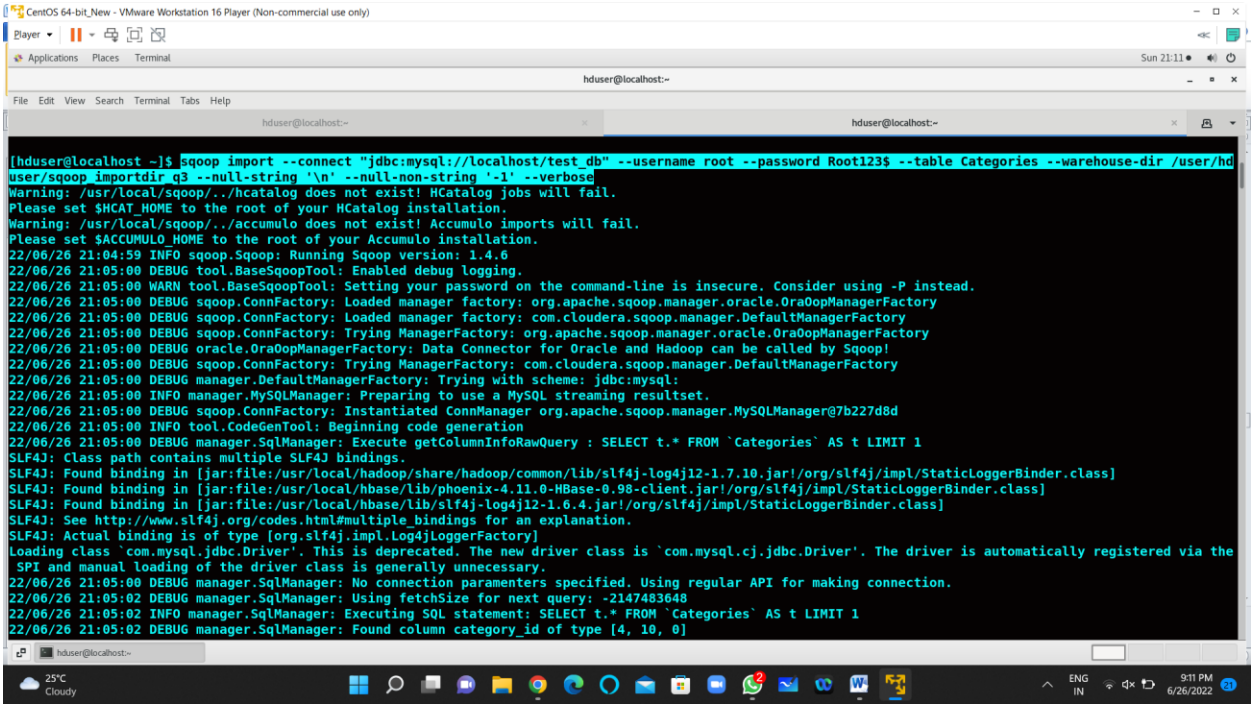
(12,5,null,'2020-01-15');

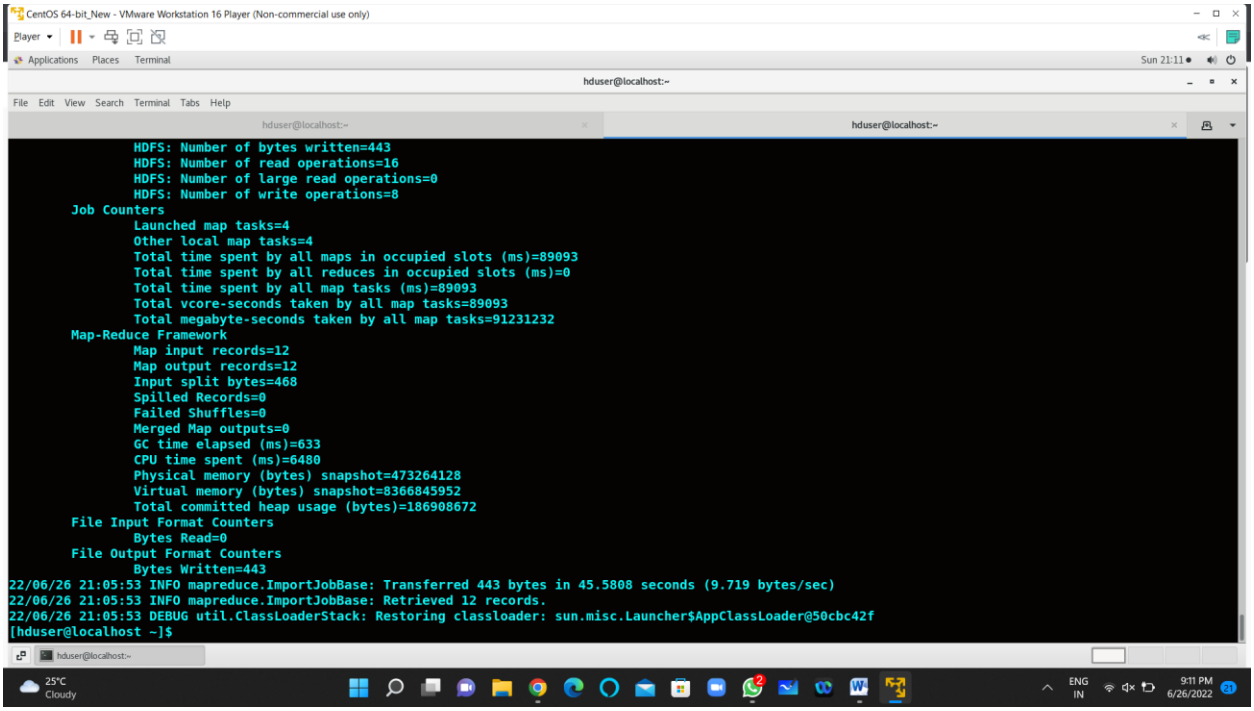
Do the following:

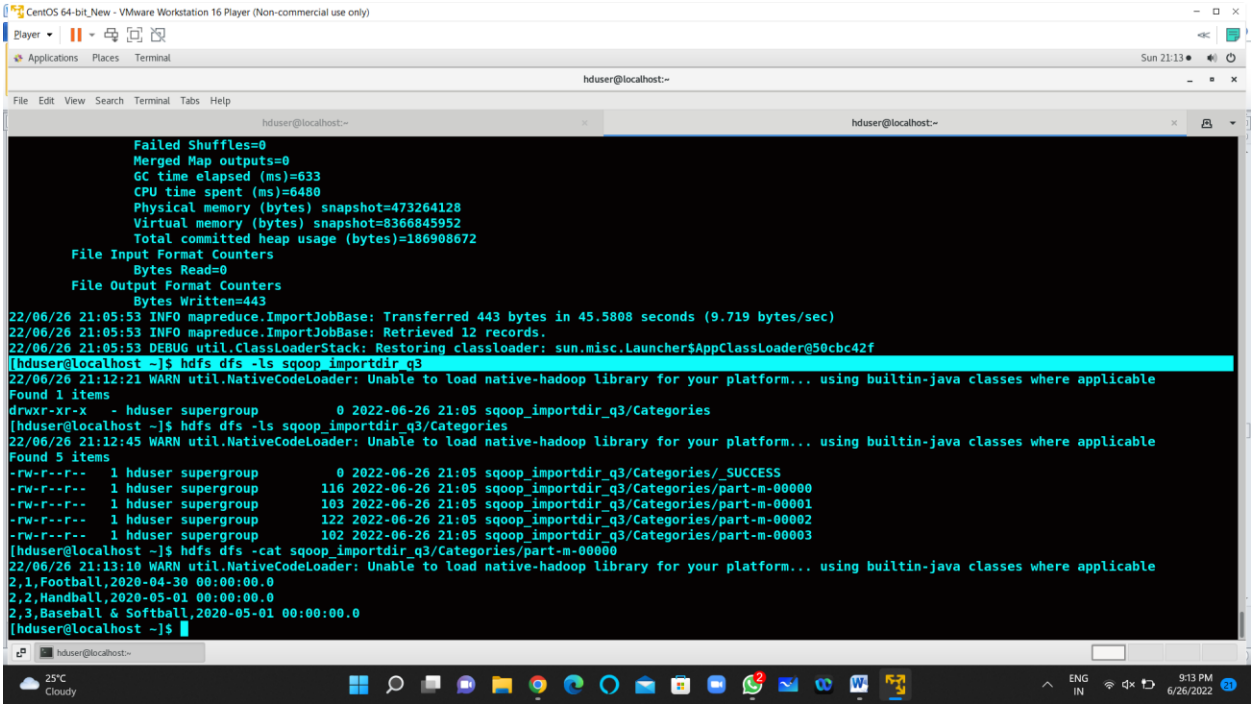
A) Import the Categories table in hdfs but during the import ,do proper Null value handling:

- 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.


CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 21:11
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
[hduser@localhost ~]$ sqoop import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --table Categories --warehouse-dir /user/hduser/sqoop_importdir_q3 --null-string '\n' --null-non-string '-1' --verbose
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.
22/06/26 21:04:59 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
22/06/26 21:05:00 DEBUG tool.BaseSqoopTool: Enabled debug logging.
22/06/26 21:05:00 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/06/26 21:05:00 DEBUG sqoop.ConnFactory: Loaded manager factory: org.apache.sqoop.manager.oracle.OraOopManagerFactory
22/06/26 21:05:00 DEBUG sqoop.ConnFactory: Loaded manager factory: com.cloudera.sqoop.manager.DefaultManagerFactory
22/06/26 21:05:00 DEBUG sqoop.ConnFactory: Trying ManagerFactory: org.apache.sqoop.manager.oracle.OraOopManagerFactory
22/06/26 21:05:00 DEBUG oracle.OraOopManagerFactory: Data Connector for Oracle and Hadoop can be called by Sqoop!
22/06/26 21:05:00 DEBUG sqoop.ConnFactory: Trying ManagerFactory: com.cloudera.sqoop.manager.DefaultManagerFactory
22/06/26 21:05:00 DEBUG manager.DefaultManagerFactory: Trying with scheme: jdbc:mysql:
22/06/26 21:05:00 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/06/26 21:05:00 DEBUG sqoop.ConnFactory: Instantiated ConnManager org.apache.sqoop.manager.MySQLManager@7b227d8d
22/06/26 21:05:00 INFO tool.CodeGenTool: Beginning code generation
22/06/26 21:05:00 DEBUG manager.SqlManager: Execute getColumnInfoRawQuery : SELECT t.* FROM `Categories` AS t LIMIT 1
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.
22/06/26 21:05:00 DEBUG manager.SqlManager: No connection paramenters specified. Using regular API for making connection.
22/06/26 21:05:02 DEBUG manager.SqlManager: Using fetchSize for next query: -2147483648
22/06/26 21:05:02 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Categories` AS t LIMIT 1
22/06/26 21:05:02 DEBUG manager.SqlManager: Found column category_id of type [4, 10, 0]
hduser@localhost:~
25°C Cloudy
ENG IN
9:11 PM 6/26/2022

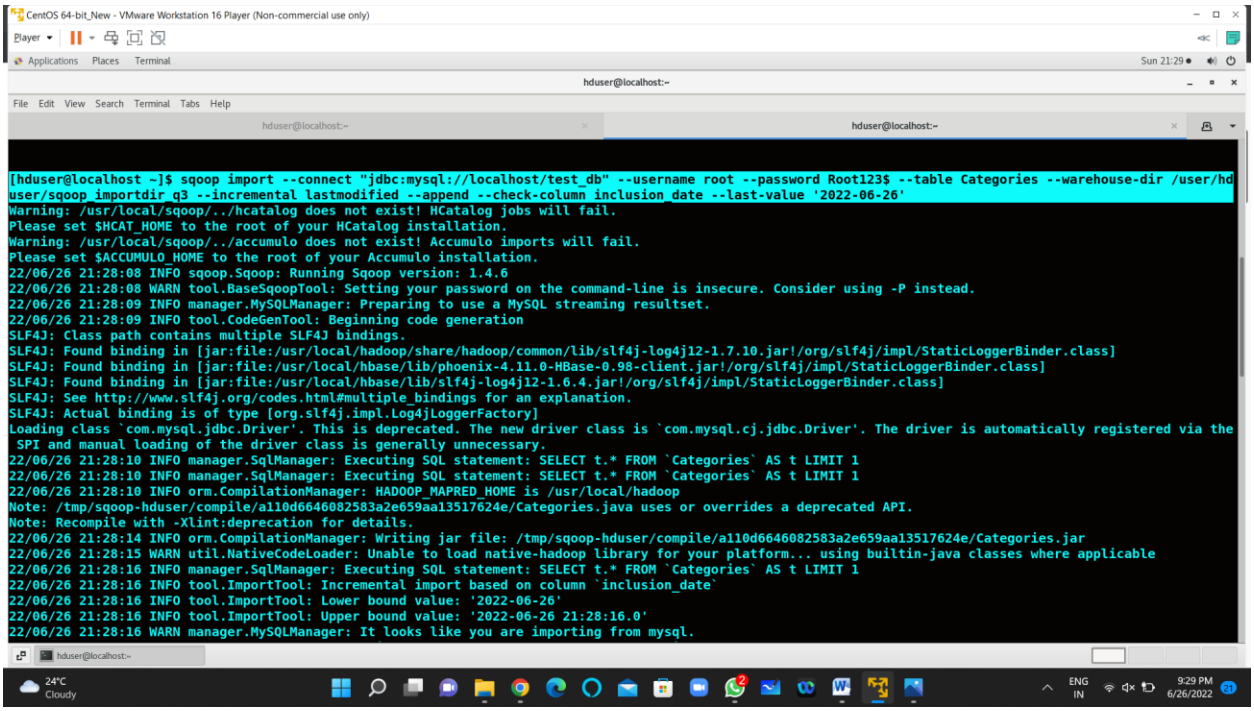

CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 21:11
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
HDFS: Number of bytes written=443
HDFS: Number of read operations=16
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Job Counters
Launched map tasks=4
Other local map tasks=4
Total time spent by all maps in occupied slots (ms)=89093
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=89093
Total vcore-seconds taken by all map tasks=89093
Total megabyte-seconds taken by all map tasks=91231232
Map-Reduce Framework
Map input records=12
Map output records=12
Input split bytes=468
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=633
CPU time spent (ms)=6480
Physical memory (bytes) snapshot=473264128
Virtual memory (bytes) snapshot=8366845952
Total committed heap usage (bytes)=186908672
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=443
22/06/26 21:05:53 INFO mapreduce.ImportJobBase: Transferred 443 bytes in 45.5808 seconds (9.719 bytes/sec)
22/06/26 21:05:53 INFO mapreduce.ImportJobBase: Retrieved 12 records.
22/06/26 21:05:53 DEBUG util.ClassLoaderStack: Restoring classloader: sun.misc.Launcher$AppClassLoader@50cbc42f
[hduser@localhost ~]$
hduser@localhost:~
25°C Cloudy
ENG IN
9:11 PM 6/26/2022

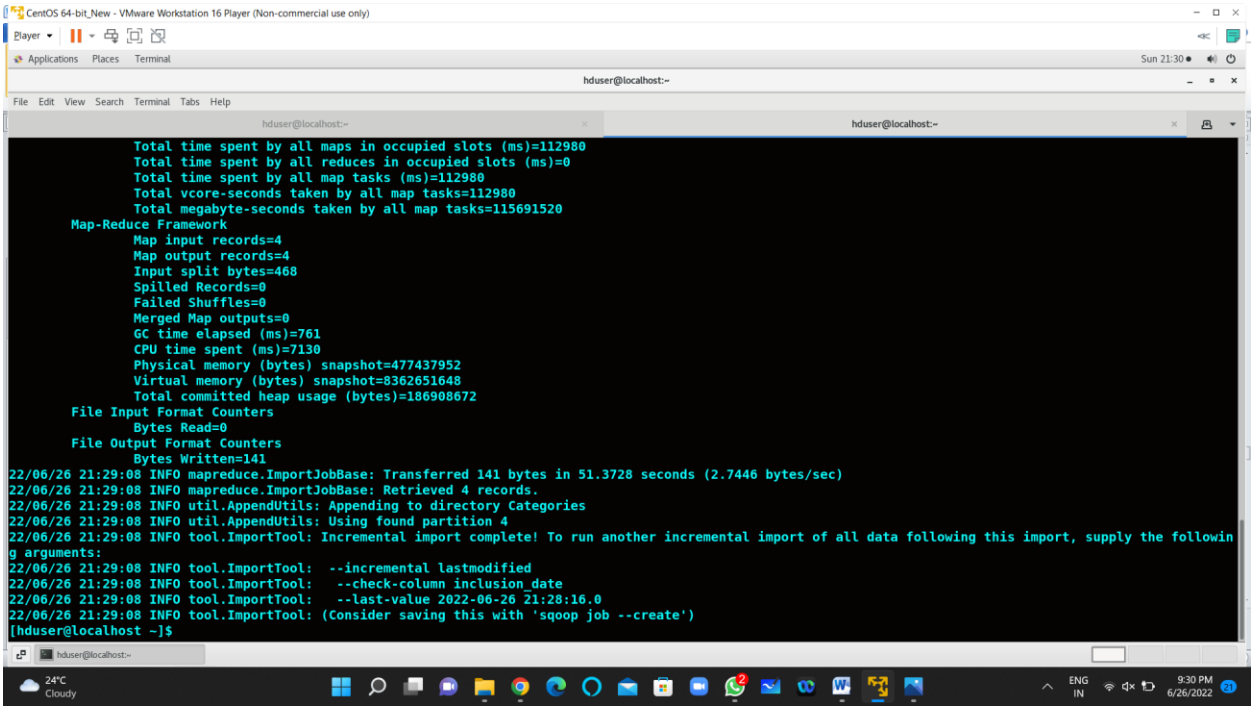


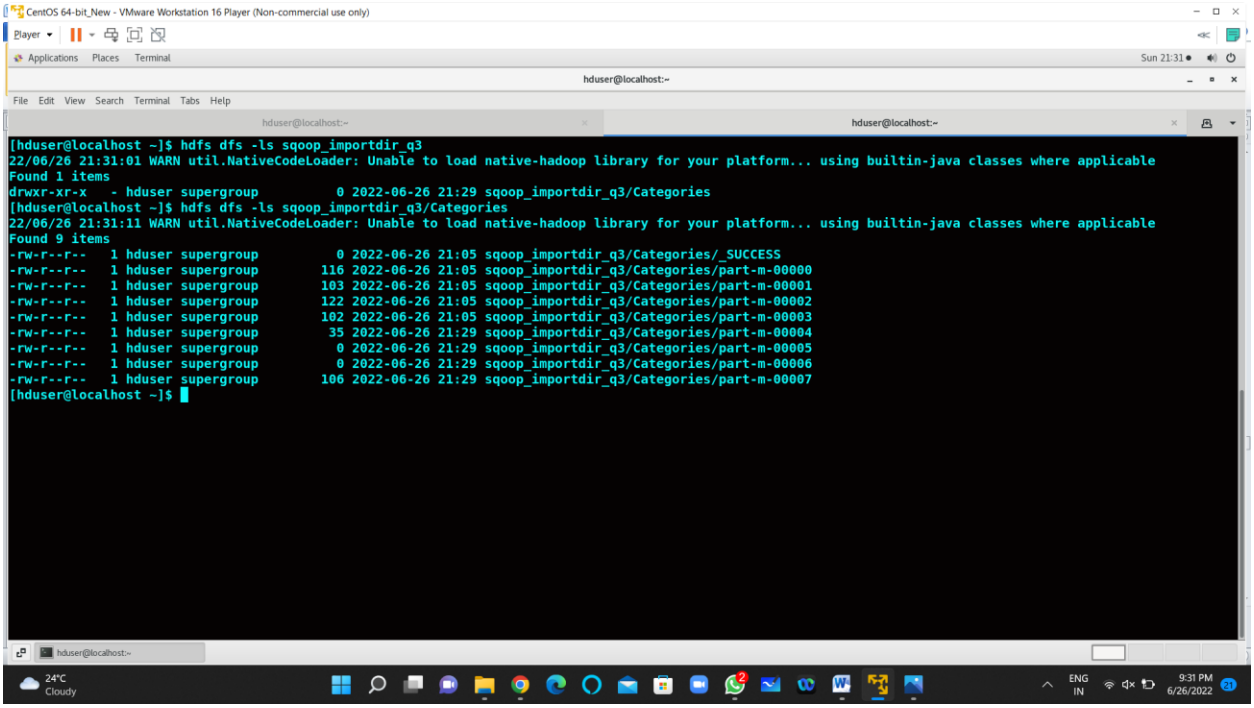
B) New Records are added to the table and also existing records are updated,(refer the mysql_commands text file for the insert and update commands), so import only those newly inserted/updated records from Categories table to hdfs.

The delta records should get appended to existing directory.

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

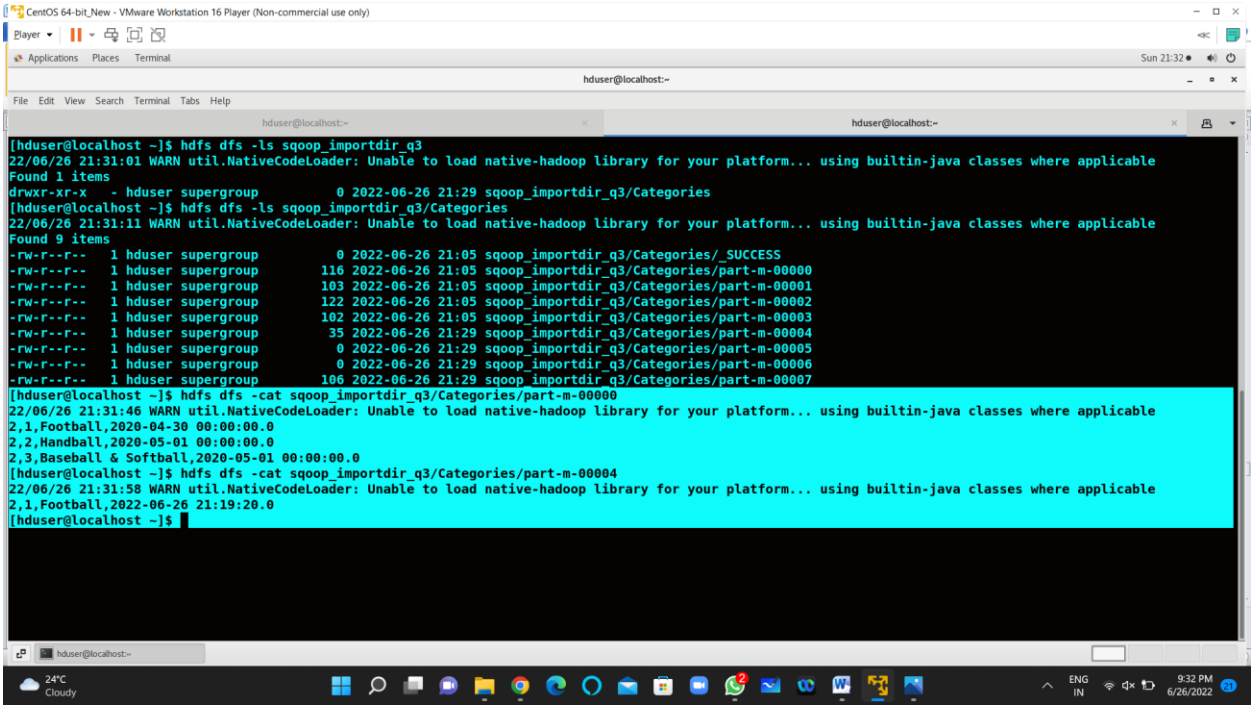

CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 21:29
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
[hduser@localhost ~]$ sqoop import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --table Categories --warehouse-dir /user/hduser/sqoop_importdir_q3 --incremental lastmodified --append --check-column inclusion_date --last-value '2022-06-26'
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.
22/06/26 21:28:08 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
22/06/26 21:28:08 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/06/26 21:28:09 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/06/26 21:28:09 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.
22/06/26 21:28:10 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Categories` AS t LIMIT 1
22/06/26 21:28:10 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Categories` AS t LIMIT 1
22/06/26 21:28:10 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-hduser/compile/a110d6646082583a2e659aa13517624e/Categories.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/06/26 21:28:14 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hduser/compile/a110d6646082583a2e659aa13517624e/Categories.jar
22/06/26 21:28:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/06/26 21:28:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Categories` AS t LIMIT 1
22/06/26 21:28:16 INFO tool.ImportTool: Incremental import based on column `inclusion_date`
22/06/26 21:28:16 INFO tool.ImportTool: Lower bound value: '2022-06-26'
22/06/26 21:28:16 INFO tool.ImportTool: Upper bound value: '2022-06-26 21:28:16.0'
22/06/26 21:28:16 WARN manager.MySQLManager: It looks like you are importing from mysql.
24°C Cloudy
ENG IN
9:29 PM 6/26/2022


CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 21:30
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
Total time spent by all maps in occupied slots (ms)=112980
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=112980
Total vcore-seconds taken by all map tasks=112980
Total megabyte-seconds taken by all map tasks=115691520
Map-Reduce Framework
Map input records=4
Map output records=4
Input split bytes=468
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=761
CPU time spent (ms)=7130
Physical memory (bytes) snapshot=477437952
Virtual memory (bytes) snapshot=8362651648
Total committed heap usage (bytes)=186908672
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=141
22/06/26 21:29:08 INFO mapreduce.ImportJobBase: Transferred 141 bytes in 51.3728 seconds (2.7446 bytes/sec)
22/06/26 21:29:08 INFO mapreduce.ImportJobBase: Retrieved 4 records.
22/06/26 21:29:08 INFO util.AppendUtils: Appending to directory Categories
22/06/26 21:29:08 INFO util.AppendUtils: Using found partition 4
22/06/26 21:29:08 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments:
22/06/26 21:29:08 INFO tool.ImportTool: --incremental lastmodified
22/06/26 21:29:08 INFO tool.ImportTool: --check-column inclusion_date
22/06/26 21:29:08 INFO tool.ImportTool: --last-value 2022-06-26 21:28:16.0
22/06/26 21:29:08 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')
[hduser@localhost ~]$
24°C Cloudy
ENG IN
9:30 PM 6/26/2022



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.

Yes, there are duplicate records.



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.

CREATE TABLE Categories_new (

category_id INT(11) ,

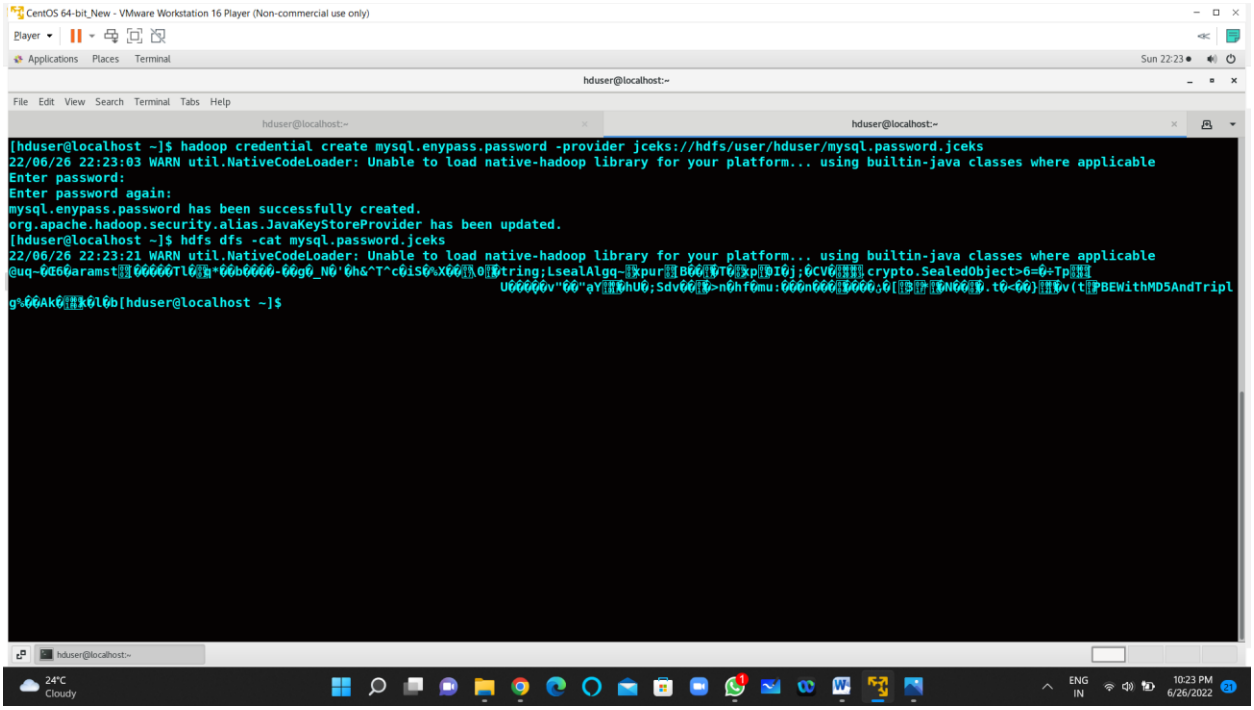
category_department_id INT(11),

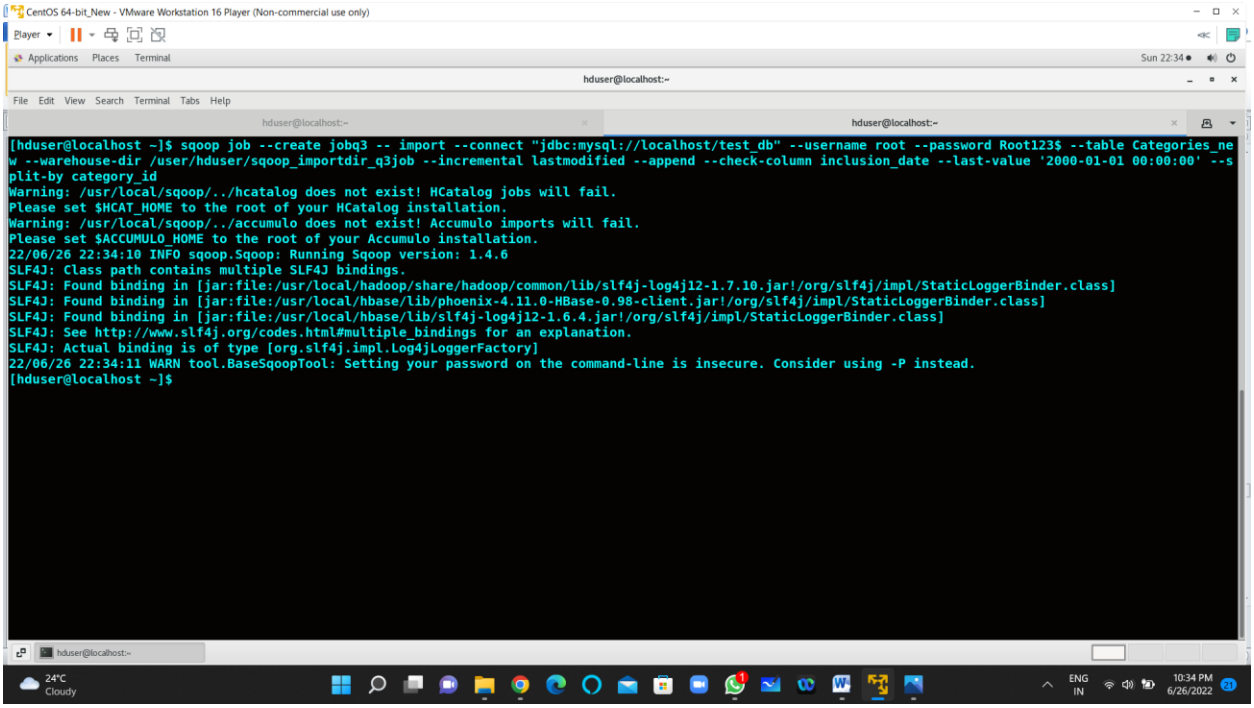
category_name VARCHAR(45),

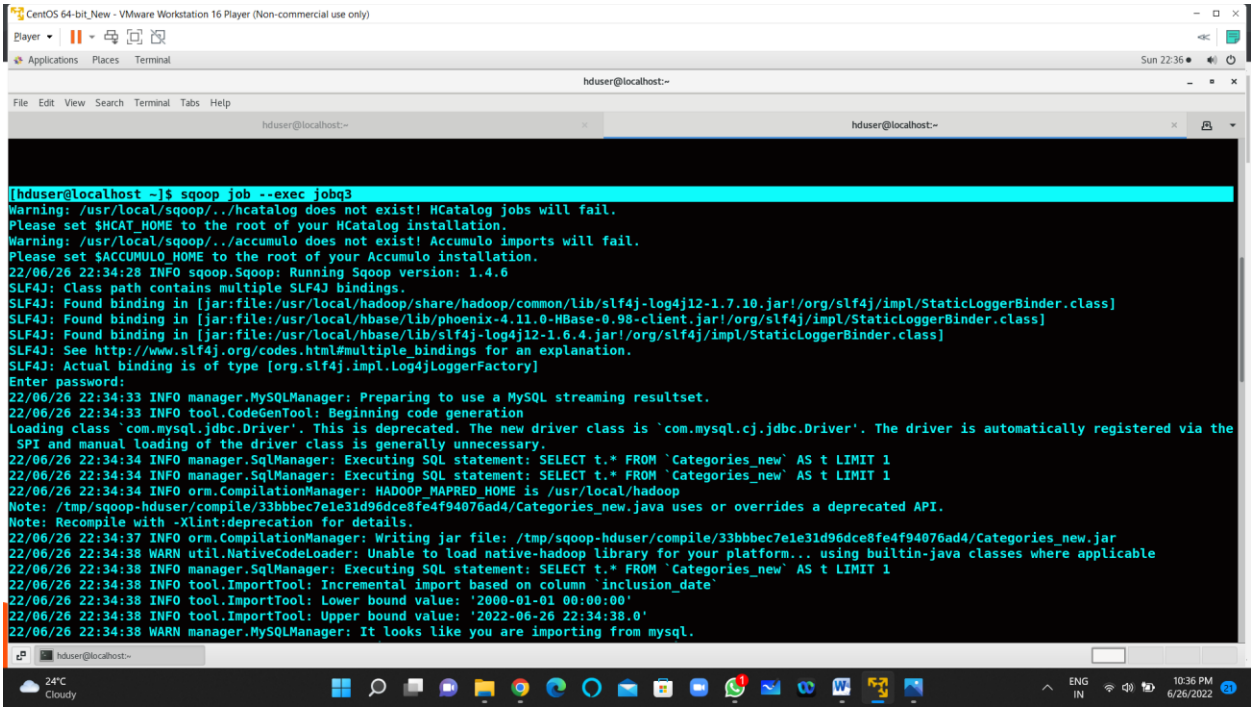
inclusion_date datetime NOT NULL);

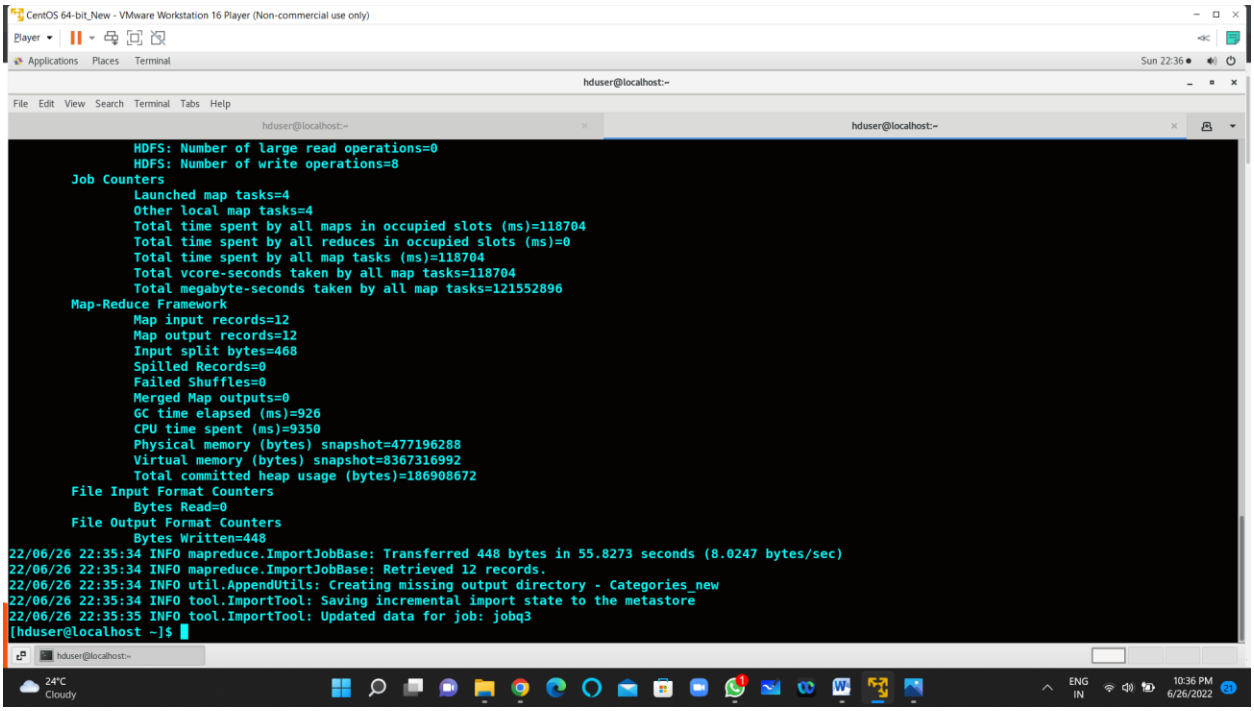
Share the commands you will use when:

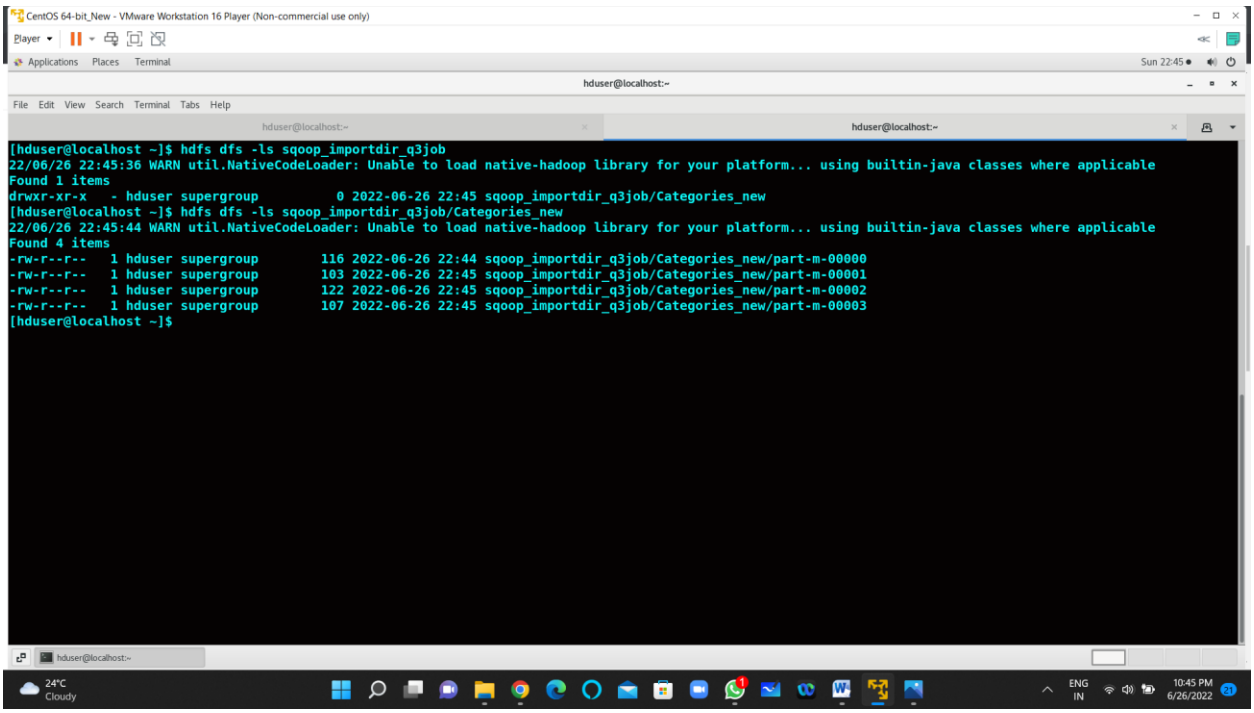
- First time we need to pull all records in hdfs
- Second time to pull only the delta records,but without duplicates in hdfs

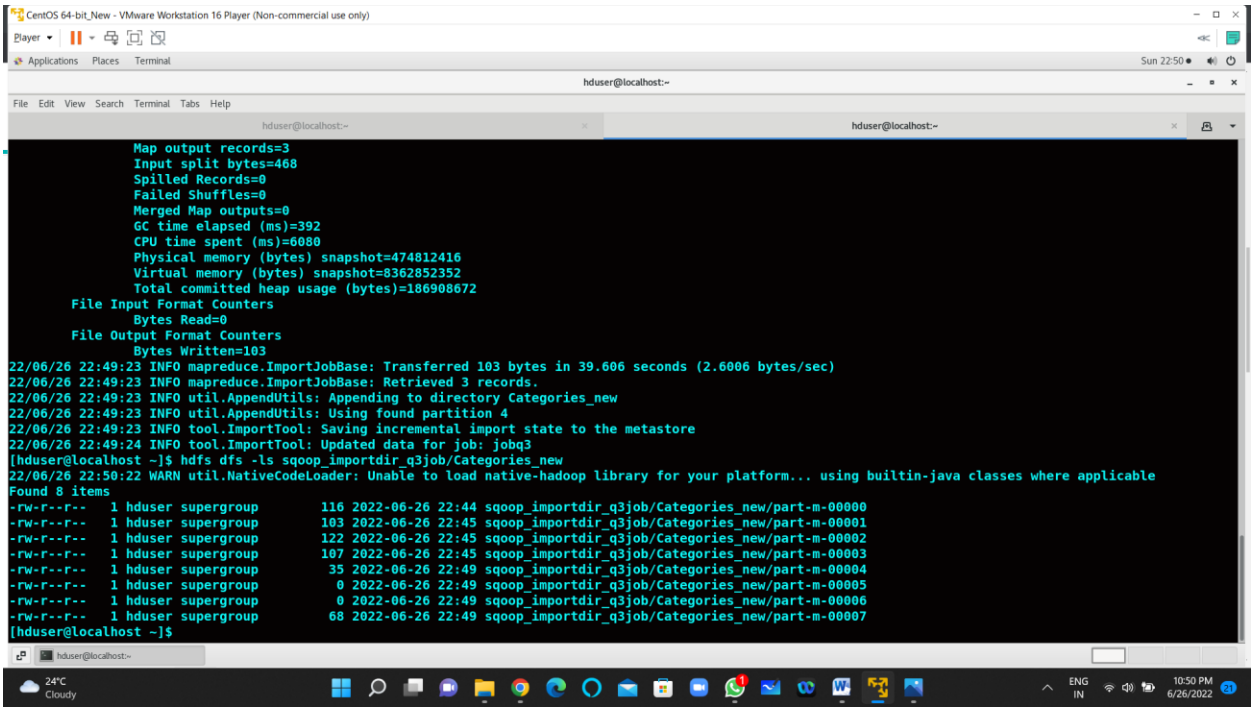


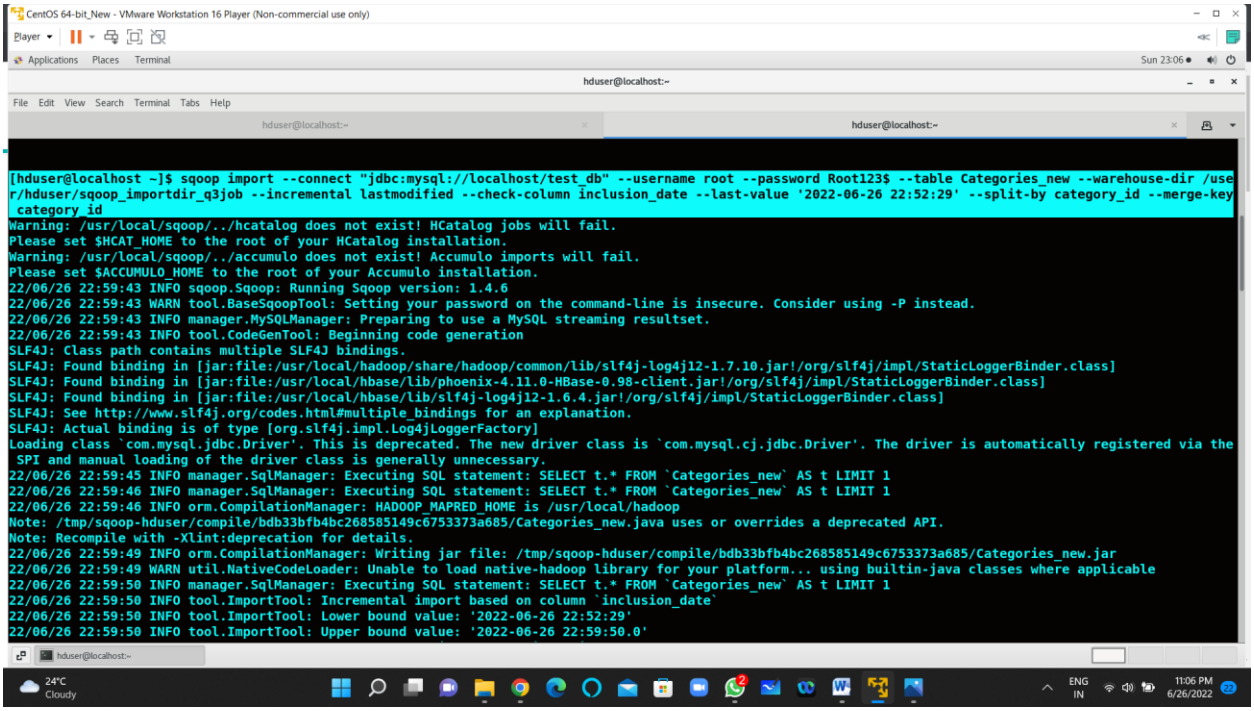

CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 22:34
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
[hduser@localhost ~]$ sqoop job --create jobq3 -- import --connect "jdbc:mysql://localhost/test_db" --username root --password Root123$ --table Categories_new --warehouse-dir /user/hduser/sqoop_importdir_q3job --incremental lastmodified --append --check-column inclusion_date --last-value '2000-01-01 00:00:00' --split-by category_id
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 $ACCUMULO_HOME to the root of your Accumulo installation.
22/06/26 22:34:10 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/phoenix-4.11.0-HBase-0.98-client.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/06/26 22:34:11 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
[hduser@localhost ~]$
hduser@localhost:~
24°C
Cloudy
ENG
IN
10:34 PM
6/26/2022


CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 22:36
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
[hduser@localhost ~]$ sqoop job --exec jobq3
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.
22/06/26 22:34:28 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
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]
Enter password:
22/06/26 22:34:33 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/06/26 22:34:33 INFO tool.CodeGenTool: Beginning code generation
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.
22/06/26 22:34:34 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Categories_new` AS t LIMIT 1
22/06/26 22:34:34 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Categories_new` AS t LIMIT 1
22/06/26 22:34:34 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-hduser/compile/33bbbec7e1e31d96dce8fe4f94076ad4/Categories_new.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/06/26 22:34:37 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hduser/compile/33bbbec7e1e31d96dce8fe4f94076ad4/Categories_new.jar
22/06/26 22:34:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/06/26 22:34:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Categories_new` AS t LIMIT 1
22/06/26 22:34:38 INFO tool.ImportTool: Incremental import based on column `inclusion_date`
22/06/26 22:34:38 INFO tool.ImportTool: Lower bound value: '2000-01-01 00:00:00'
22/06/26 22:34:38 INFO tool.ImportTool: Upper bound value: '2022-06-26 22:34:38.0'
22/06/26 22:34:38 WARN manager.MySQLManager: It looks like you are importing from mysql.
hduser@localhost:~
24°C
Cloudy
ENG
IN
10:36 PM
6/26/2022

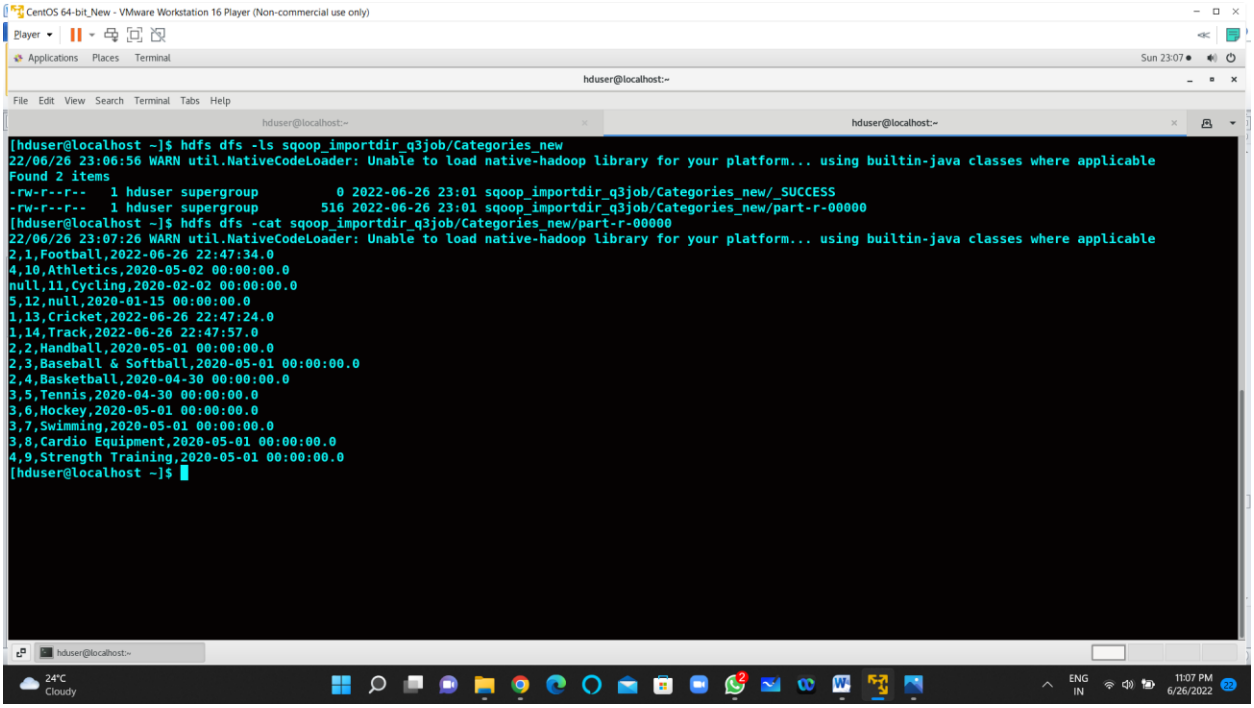

CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 22:36
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Job Counters
Launched map tasks=4
Other local map tasks=4
Total time spent by all maps in occupied slots (ms)=118704
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=118704
Total vcore-seconds taken by all map tasks=118704
Total megabyte-seconds taken by all map tasks=121552896
Map-Reduce Framework
Map input records=12
Map output records=12
Input split bytes=468
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=926
CPU time spent (ms)=9350
Physical memory (bytes) snapshot=477196288
Virtual memory (bytes) snapshot=8367316992
Total committed heap usage (bytes)=186908672
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=448
22/06/26 22:35:34 INFO mapreduce.ImportJobBase: Transferred 448 bytes in 55.8273 seconds (8.0247 bytes/sec)
22/06/26 22:35:34 INFO mapreduce.ImportJobBase: Retrieved 12 records.
22/06/26 22:35:34 INFO util.AppendUtils: Creating missing output directory - Categories_new
22/06/26 22:35:34 INFO tool.ImportTool: Saving incremental import state to the metastore
22/06/26 22:35:35 INFO tool.ImportTool: Updated data for job: jobq3
[hduser@localhost ~]$
hduser@localhost:~
24°C
Cloudy
ENG
IN
10:36 PM
6/26/2022


CentOS 64-bit_New - VMware Workstation 16 Player (Non-commercial use only)
Player
Applications Places Terminal
Sun 22:45
hduser@localhost:~
File Edit View Search Terminal Tabs Help
hduser@localhost:~
hduser@localhost:~
[hduser@localhost ~]$ hdfs dfs -ls sqoop_importdir_q3job
22/06/26 22:45:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x - hduser supergroup 0 2022-06-26 22:45 sqoop_importdir_q3job/Categories_new
[hduser@localhost ~]$ hdfs dfs -ls sqoop_importdir_q3job/Categories_new
22/06/26 22:45:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 4 items
-rw-r--r-- 1 hduser supergroup 116 2022-06-26 22:44 sqoop_importdir_q3job/Categories_new/part-m-00000
-rw-r--r-- 1 hduser supergroup 103 2022-06-26 22:45 sqoop_importdir_q3job/Categories_new/part-m-00001
-rw-r--r-- 1 hduser supergroup 122 2022-06-26 22:45 sqoop_importdir_q3job/Categories_new/part-m-00002
-rw-r--r-- 1 hduser supergroup 107 2022-06-26 22:45 sqoop_importdir_q3job/Categories_new/part-m-00003
[hduser@localhost ~]$
hduser@localhost:~
24°C
Cloudy
ENG
IN
10:45 PM
6/26/2022





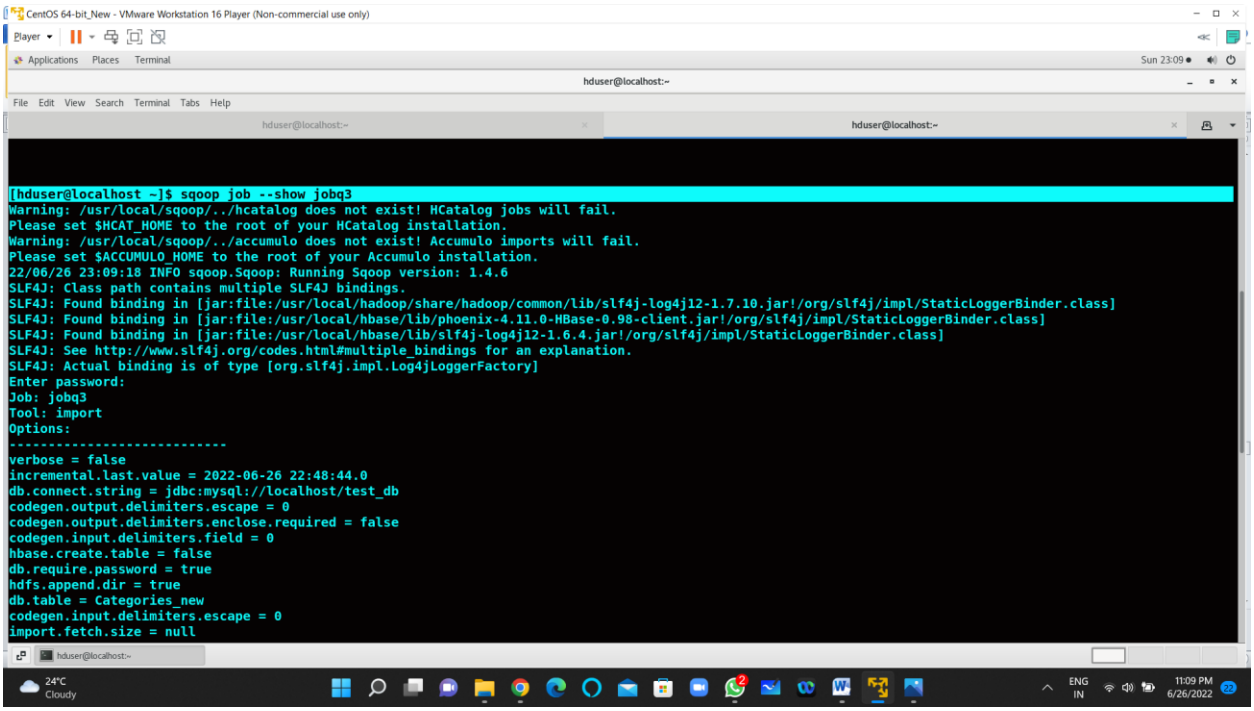
E) How many records do you see this time in hdfs post second import? Do you see any duplicate records now?



F) Are any mapper files generated in hdfs this time after the second import? Explain.

No mapper files are generated since merge operation is performed which is a reduce operation

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



sqoop Quiz

1. Sqoop written in? **C-Java**

A. C

B. C++

C. Java

D. hadoop

2. Sqoop stands for? **A-SQL to Hadoop**

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**

A. TRUE

B. FALSE

C. Can be true or false

D. Can not say

4. Data processed by Scoop can be used for? **C. Mapreduce**

A. Hbase

B. HDFS

C. Mapreduce

D. MahOut

5. _____ tool can list all the available database schemas **B. sqoop-list-databases**

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. **B. FALSE**

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**

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**

A. 0

B. 2

C. 3

D. 1

.

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

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. **B. CREATE TABLE**

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 **B. import-all-tables**

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**

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