

# LAB 4 BIG DATA

**Submitted by:**  
**ABHISHEK ARYA**  
**NET ID: axa220149**

**Answer: The desirable outputs of LAB 4 are as follows in the 3 immediate screenshots below:**

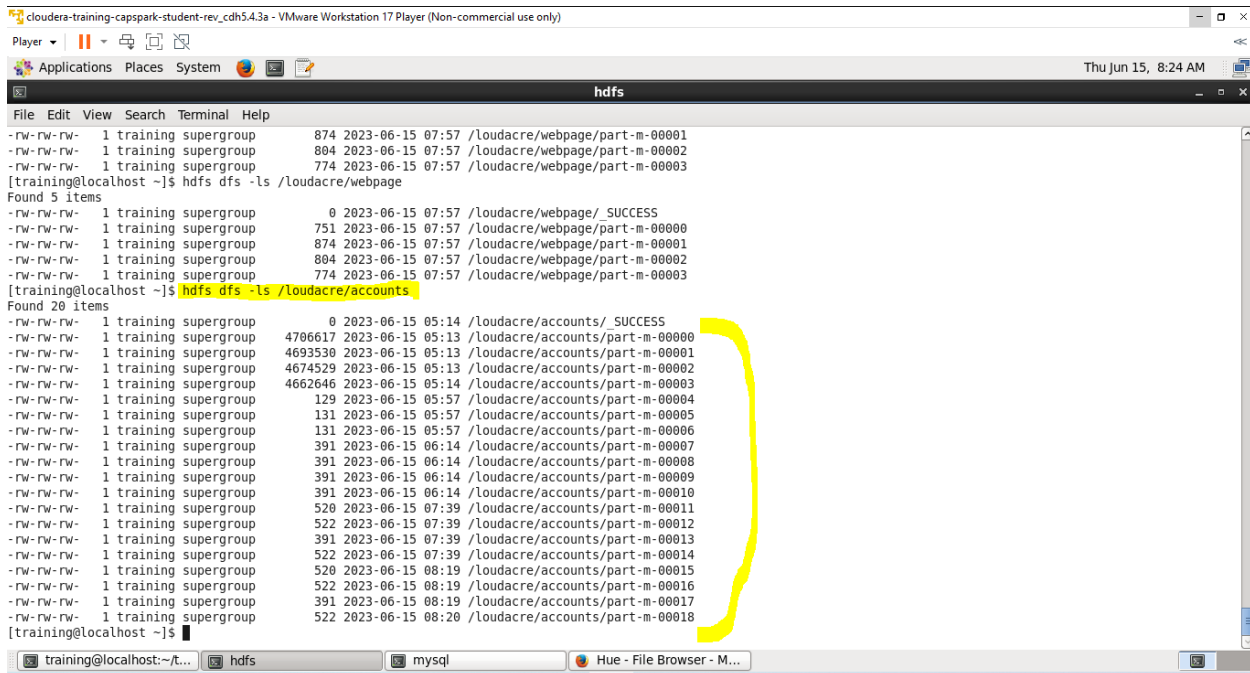
### Screenshot of the code:

The screenshot shows a terminal window with the following command and output:

```
sqoop import \
--connect jdbc:mysql://localhost/loudacre \
--username training --password training \
--incremental append \
--table accounts \
--target-dir /loudacre/accounts \
--null-non-string '\N' \
--check-column acct num \
--last-value 129761
```

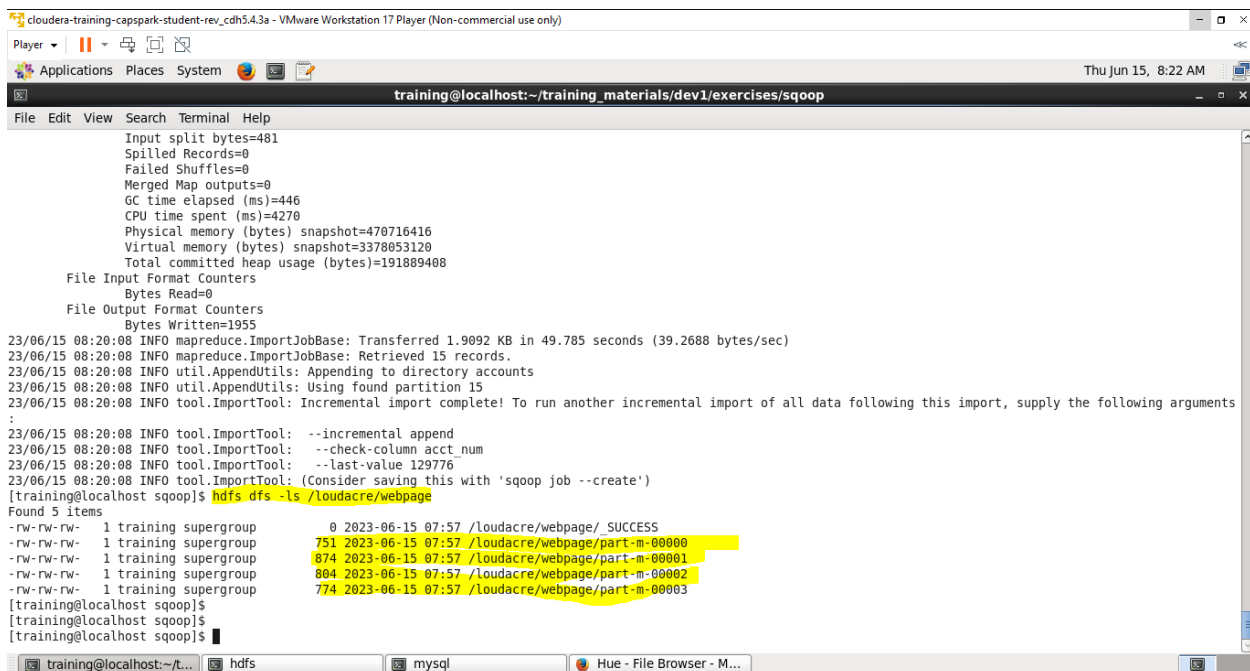
The output shows the last value as 129761.

## Screenshot from the terminal:



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
hdfs
File Edit View Search Terminal Help
-rw-rw-rw- 1 training supergroup 874 2023-06-15 07:57 /loudacre/webpage/part-m-00001
-rw-rw-rw- 1 training supergroup 804 2023-06-15 07:57 /loudacre/webpage/part-m-00002
-rw-rw-rw- 1 training supergroup 774 2023-06-15 07:57 /loudacre/webpage/part-m-00003
[training@localhost ~]$ hdfs dfs -ls /loudacre/webpage
Found 5 items
-rw-rw-rw- 1 training supergroup 0 2023-06-15 07:57 /loudacre/webpage/ SUCCESS
-rw-rw-rw- 1 training supergroup 751 2023-06-15 07:57 /loudacre/webpage/part-m-00000
-rw-rw-rw- 1 training supergroup 874 2023-06-15 07:57 /loudacre/webpage/part-m-00001
-rw-rw-rw- 1 training supergroup 804 2023-06-15 07:57 /loudacre/webpage/part-m-00002
-rw-rw-rw- 1 training supergroup 774 2023-06-15 07:57 /loudacre/webpage/part-m-00003
[training@localhost ~]$ hdfs dfs -ls /loudacre/accounts
Found 20 items
-rw-rw-rw- 1 training supergroup 0 2023-06-15 05:14 /loudacre/accounts/ SUCCESS
-rw-rw-rw- 1 training supergroup 4706617 2023-06-15 05:13 /loudacre/accounts/part-m-00000
-rw-rw-rw- 1 training supergroup 4693530 2023-06-15 05:13 /loudacre/accounts/part-m-00001
-rw-rw-rw- 1 training supergroup 4674529 2023-06-15 05:13 /loudacre/accounts/part-m-00002
-rw-rw-rw- 1 training supergroup 4662646 2023-06-15 05:14 /loudacre/accounts/part-m-00003
-rw-rw-rw- 1 training supergroup 129 2023-06-15 05:57 /loudacre/accounts/part-m-00004
-rw-rw-rw- 1 training supergroup 131 2023-06-15 05:57 /loudacre/accounts/part-m-00005
-rw-rw-rw- 1 training supergroup 131 2023-06-15 05:57 /loudacre/accounts/part-m-00006
-rw-rw-rw- 1 training supergroup 391 2023-06-15 06:14 /loudacre/accounts/part-m-00007
-rw-rw-rw- 1 training supergroup 391 2023-06-15 06:14 /loudacre/accounts/part-m-00008
-rw-rw-rw- 1 training supergroup 391 2023-06-15 06:14 /loudacre/accounts/part-m-00009
-rw-rw-rw- 1 training supergroup 391 2023-06-15 06:14 /loudacre/accounts/part-m-00010
-rw-rw-rw- 1 training supergroup 520 2023-06-15 07:39 /loudacre/accounts/part-m-00011
-rw-rw-rw- 1 training supergroup 522 2023-06-15 07:39 /loudacre/accounts/part-m-00012
-rw-rw-rw- 1 training supergroup 391 2023-06-15 07:39 /loudacre/accounts/part-m-00013
-rw-rw-rw- 1 training supergroup 522 2023-06-15 07:39 /loudacre/accounts/part-m-00014
-rw-rw-rw- 1 training supergroup 520 2023-06-15 08:19 /loudacre/accounts/part-m-00015
-rw-rw-rw- 1 training supergroup 522 2023-06-15 08:19 /loudacre/accounts/part-m-00016
-rw-rw-rw- 1 training supergroup 391 2023-06-15 08:19 /loudacre/accounts/part-m-00017
-rw-rw-rw- 1 training supergroup 522 2023-06-15 08:20 /loudacre/accounts/part-m-00018
[training@localhost ~]$
```

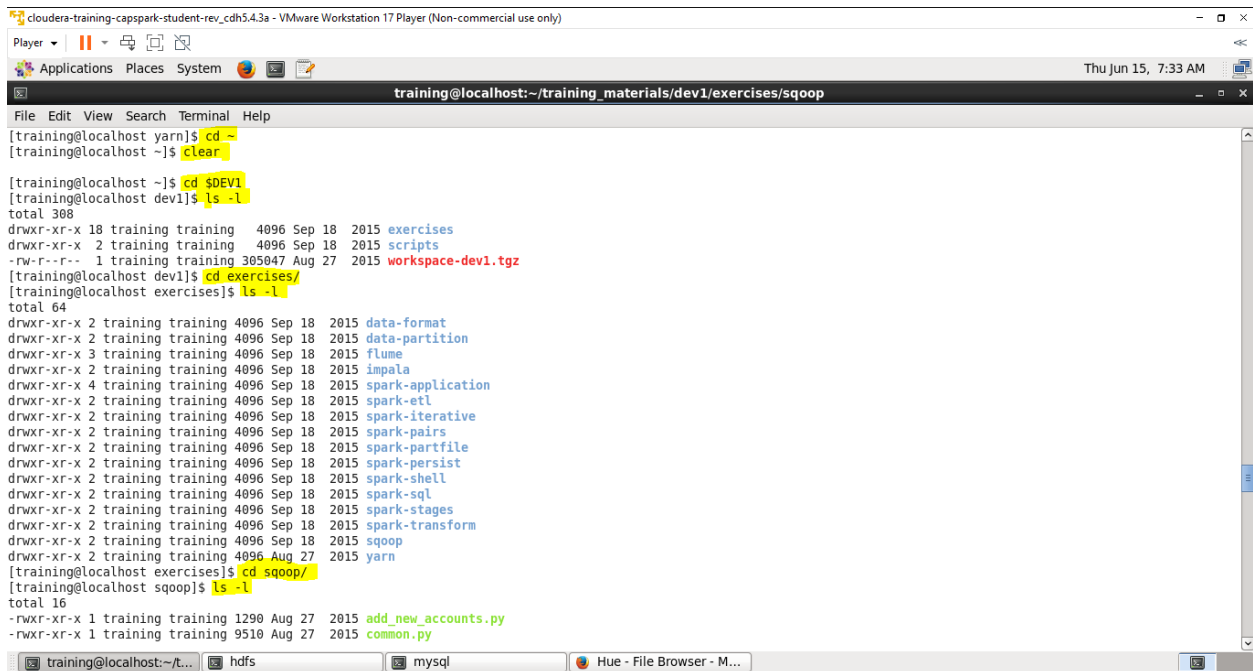
## Screenshot from the terminal:



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
training@localhost:~/training_materials/dev1/exercises/sqoop
File Edit View Search Terminal Help
Input split bytes=481
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=446
CPU time spent (ms)=4270
Physical memory (bytes) snapshot=470716416
Virtual memory (bytes) snapshot=3378053120
Total committed heap usage (bytes)=191809408
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=1955
23/06/15 08:20:08 INFO mapreduce.ImportJobBase: Transferred 1.9092 KB in 49.785 seconds (39.2688 bytes/sec)
23/06/15 08:20:08 INFO mapreduce.ImportJobBase: Retrieved 15 records.
23/06/15 08:20:08 INFO util.AppendUtils: Appending to directory accounts
23/06/15 08:20:08 INFO util.AppendUtils: Using found partition 15
23/06/15 08:20:08 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments
:
23/06/15 08:20:08 INFO tool.ImportTool: --incremental append
23/06/15 08:20:08 INFO tool.ImportTool: --check-column acct_num
23/06/15 08:20:08 INFO tool.ImportTool: --last-value 129776
23/06/15 08:20:08 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')
[training@localhost sqoop]$ hdfs dfs -ls /loudacre/webpage
Found 5 items
-rw-rw-rw- 1 training supergroup 0 2023-06-15 07:57 /loudacre/webpage/ SUCCESS
-rw-rw-rw- 1 training supergroup 751 2023-06-15 07:57 /loudacre/webpage/part-m-00000
-rw-rw-rw- 1 training supergroup 874 2023-06-15 07:57 /loudacre/webpage/part-m-00001
-rw-rw-rw- 1 training supergroup 804 2023-06-15 07:57 /loudacre/webpage/part-m-00002
-rw-rw-rw- 1 training supergroup 774 2023-06-15 07:57 /loudacre/webpage/part-m-00003
[training@localhost sqoop]$
[training@localhost sqoop]$
[training@localhost sqoop]$
```

I am adding all the steps which led me to the above outputs just for the safer side; please find the screenshots of all the steps outputs below:

1.



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
training@localhost:~/training_materials/dev1/exercises/sqoop
File Edit View Search Terminal Help
[training@localhost ~]$ cd -
[training@localhost ~]$ clear
[training@localhost ~]$ cd $DEV1
[training@localhost dev1]$ ls -l
total 308
drwxr-xr-x 18 training training 4096 Sep 18 2015 exercises
drwxr-xr-x 2 training training 4096 Sep 18 2015 scripts
-rw-r--r-- 1 training training 305047 Aug 27 2015 workspace-dev1.tgz
[training@localhost dev1]$ cd exercises/
[training@localhost exercises]$ ls -l
total 64
drwxr-xr-x 2 training training 4096 Sep 18 2015 data-format
drwxr-xr-x 2 training training 4096 Sep 18 2015 data-partition
drwxr-xr-x 3 training training 4096 Sep 18 2015 flume
drwxr-xr-x 2 training training 4096 Sep 18 2015 impala
drwxr-xr-x 4 training training 4096 Sep 18 2015 spark-application
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-etl
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-iterative
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-pairs
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-partfile
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-persist
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-shell
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-sql
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-stages
drwxr-xr-x 2 training training 4096 Sep 18 2015 spark-transform
drwxr-xr-x 2 training training 4096 Sep 18 2015 sqoop
drwxr-xr-x 2 training training 4096 Aug 27 2015 yarn
[training@localhost exercises]$ cd sqoop/
[training@localhost sqoop]$ ls -l
total 16
-rwxr-xr-x 1 training training 1290 Aug 27 2015 add_new_accounts.py
-rwxr-xr-x 1 training training 9510 Aug 27 2015 common.py
```

2.

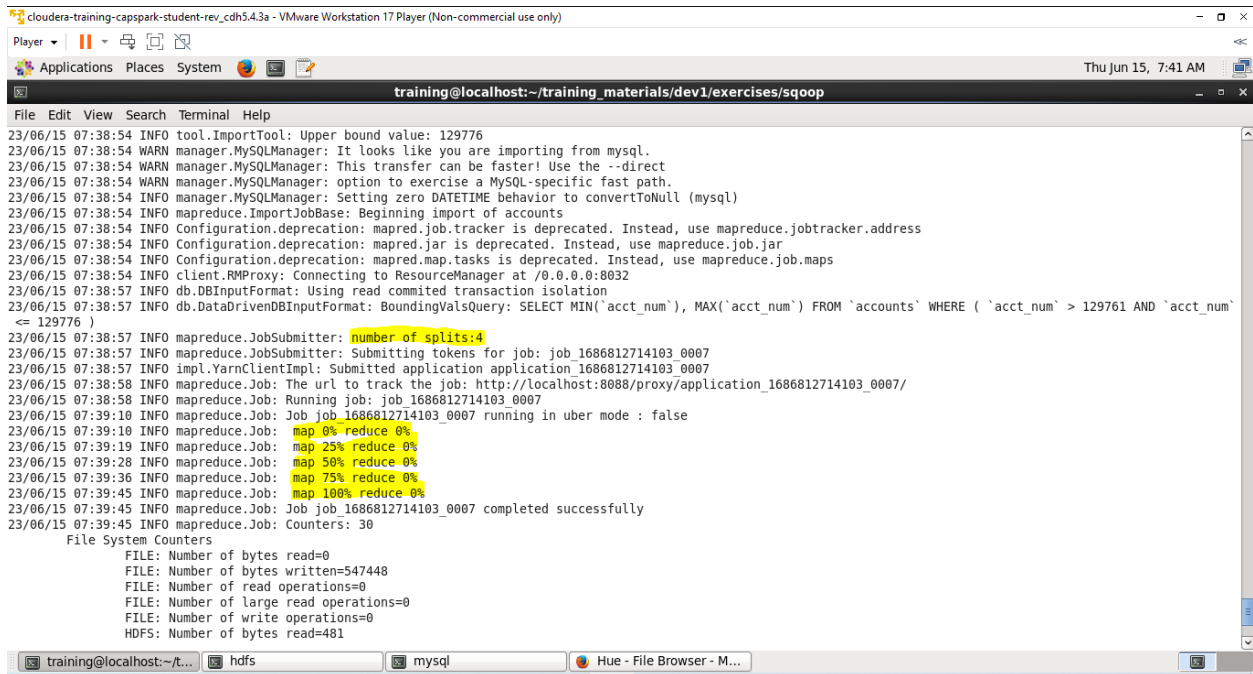
```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
training@localhost:~/training_materials/dev1/exercises/sqoop
File Edit View Search Terminal Help
23/06/15 07:23:42 ERROR tool.BaseSqoopTool: Unrecognized argument: --query
23/06/15 07:23:42 ERROR tool.BaseSqoopTool: Unrecognized argument: select a.acct_num , b.cs_name from accounts a JOIN customerservice b USING(acct_num) where $CONDIT
IONS and a.acct_num > 100
23/06/15 07:23:42 ERROR tool.BaseSqoopTool: Unrecognized argument: -m
23/06/15 07:23:42 ERROR tool.BaseSqoopTool: Unrecognized argument: 8

Try --help for usage instructions.
[training@localhost sqoop]$ myfirstsqoopcondition.sh
bash: myfirstsqoopcondition.sh: command not found
[training@localhost sqoop]$ vi myfirstsqoopcondition.sh
[training@localhost sqoop]$ vi myfirstsqoopappend.sh
[training@localhost sqoop]$ ./myfirstsqoopappend.sh
23/06/15 07:38:47 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
23/06/15 07:38:47 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/06/15 07:38:47 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/06/15 07:38:47 INFO tool.CodeGenTool: Beginning code generation
23/06/15 07:38:48 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'accounts' AS t LIMIT 1
23/06/15 07:38:48 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'accounts' AS t LIMIT 1
23/06/15 07:38:48 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/970a268045b14b4f9861ab8f7429cd03/accounts.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/06/15 07:38:52 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/970a268045b14b4f9861ab8f7429cd03/accounts.jar
23/06/15 07:38:54 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX('acct_num') FROM 'accounts'
23/06/15 07:38:54 INFO tool.ImportTool: Incremental import based on column 'acct_num'
23/06/15 07:38:54 INFO tool.ImportTool: Lower bound value: 129761
23/06/15 07:38:54 INFO tool.ImportTool: Upper bound value: 129776
23/06/15 07:38:54 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/06/15 07:38:54 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/06/15 07:38:54 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/06/15 07:38:54 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/06/15 07:38:54 INFO mapreduce.ImportJobBase: Beginning import of accounts
23/06/15 07:38:54 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
```

3.

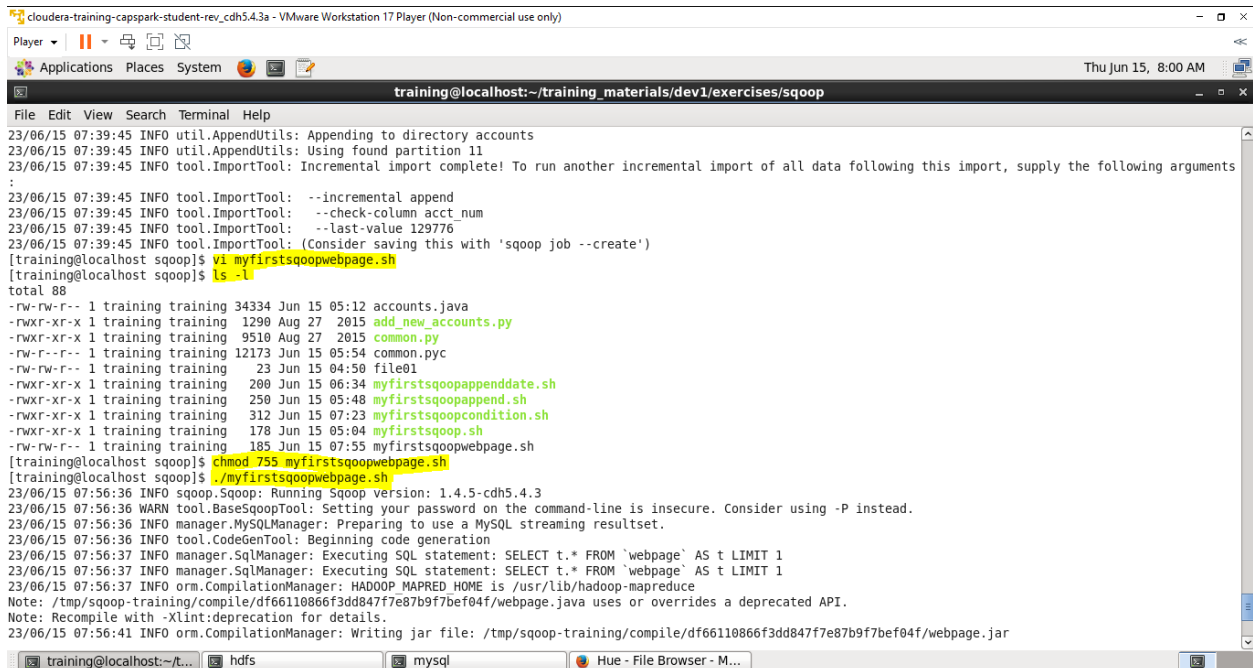
```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
training@localhost:~/training_materials/dev1/exercises/sqoop
File Edit View Search Terminal Help
Other local map tasks=4
Total time spent by all maps in occupied slots (ms)=0
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=27605
Total vcore-seconds taken by all map tasks=27605
Total megabyte-seconds taken by all map tasks=7066880
Map-Reduce Framework
Map input records=15
Map output records=15
Input split bytes=481
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=453
CPU time spent (ms)=4300
Physical memory (bytes) snapshot=474923008
Virtual memory (bytes) snapshot=3378053120
Total committed heap usage (bytes)=191889408
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=1955
23/06/15 07:39:45 INFO mapreduce.ImportJobBase: Transferred 1.9092 KB in 50.9593 seconds (38.364 bytes/sec)
23/06/15 07:39:45 INFO mapreduce.ImportJobBase: Retrieved 15 records.
23/06/15 07:39:45 INFO util.AppendUtils: Appending to directory accounts
23/06/15 07:39:45 INFO util.AppendUtils: Using found partition 11
23/06/15 07:39:45 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments
:
23/06/15 07:39:45 INFO tool.ImportTool: --incremental append
23/06/15 07:39:45 INFO tool.ImportTool: --check-column acct_num
23/06/15 07:39:45 INFO tool.ImportTool: --last-value 129776
23/06/15 07:39:45 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')
[training@localhost sqoop]$
```

4.



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
training@localhost:~/training_materials/dev1/exercises/sqoop
File Edit View Search Terminal Help
23/06/15 07:38:54 INFO tool.ImportTool: Upper bound value: 129776
23/06/15 07:38:54 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/06/15 07:38:54 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/06/15 07:38:54 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/06/15 07:38:54 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/06/15 07:38:54 INFO mapreduce.ImportJobBase: Beginning import of accounts
23/06/15 07:38:54 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/06/15 07:38:54 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/06/15 07:38:54 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/06/15 07:38:54 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/06/15 07:38:57 INFO db.DBInputFormat: Using read committed transaction isolation
23/06/15 07:38:57 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN('acct_num'), MAX('acct_num') FROM 'accounts' WHERE ( 'acct_num' > 129761 AND 'acct_num'
<= 129776 )
23/06/15 07:38:57 INFO mapreduce.JobSubmitter: number of splits:4
23/06/15 07:38:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1686812714103_0007
23/06/15 07:38:57 INFO impl.YarnClientImpl: Submitted application application_1686812714103_0007
23/06/15 07:38:58 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1686812714103_0007/
23/06/15 07:38:58 INFO mapreduce.Job: Running job: job_1686812714103_0007
23/06/15 07:39:10 INFO mapreduce.Job: Job job_1686812714103_0007 running in uber mode : false
23/06/15 07:39:10 INFO mapreduce.Job: map 0% reduce 0%
23/06/15 07:39:19 INFO mapreduce.Job: map 25% reduce 0%
23/06/15 07:39:28 INFO mapreduce.Job: map 50% reduce 0%
23/06/15 07:39:36 INFO mapreduce.Job: map 75% reduce 0%
23/06/15 07:39:45 INFO mapreduce.Job: map 100% reduce 0%
23/06/15 07:39:45 INFO mapreduce.Job: Job job_1686812714103_0007 completed successfully
23/06/15 07:39:45 INFO mapreduce.Job: Counters: 30
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=547448
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=481
training@localhost:~/t... hdfs mysql Hue - File Browser - M...
```

5.



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
training@localhost:~/training_materials/dev1/exercises/sqoop
File Edit View Search Terminal Help
23/06/15 07:39:45 INFO util.AppendUtils: Appending to directory accounts
23/06/15 07:39:45 INFO util.AppendUtils: Using found partition 11
23/06/15 07:39:45 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments
:
23/06/15 07:39:45 INFO tool.ImportTool: --incremental append
23/06/15 07:39:45 INFO tool.ImportTool: --check-column acct_num
23/06/15 07:39:45 INFO tool.ImportTool: --last-value 129776
23/06/15 07:39:45 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')
[training@localhost sqoop]$ vi myfirstsqoopwebpage.sh
[training@localhost sqoop]$ ls -l
total 88
-rw-rw-r-- 1 training training 34334 Jun 15 05:12 accounts.java
-rwxr-xr-x 1 training training 1290 Aug 27 2015 add_new_accounts.py
-rwxr-xr-x 1 training training 9510 Aug 27 2015 common.py
-rw-r--r-- 1 training training 12173 Jun 15 05:54 common.pyc
-rw-rw-r-- 1 training training 23 Jun 15 04:50 file01
-rwxr-xr-x 1 training training 200 Jun 15 06:34 myfirstsqoopappenddate.sh
-rwxr-xr-x 1 training training 250 Jun 15 05:48 myfirstsqoopappend.sh
-rwxr-xr-x 1 training training 312 Jun 15 07:23 myfirstsqoopcondition.sh
-rwxr-xr-x 1 training training 178 Jun 15 05:04 myfirstsqoop.sh
-rw-rw-r-- 1 training training 185 Jun 15 07:55 myfirstsqoopwebpage.sh
[training@localhost sqoop]$ chmod 755 myfirstsqoopwebpage.sh
[training@localhost sqoop]$ ./myfirstsqoopwebpage.sh
23/06/15 07:56:36 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
23/06/15 07:56:36 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/06/15 07:56:36 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/06/15 07:56:36 INFO tool.CodeGenTool: Beginning code generation
23/06/15 07:56:37 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `webpage` AS t LIMIT 1
23/06/15 07:56:37 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `webpage` AS t LIMIT 1
23/06/15 07:56:37 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/df66110866f3dd847f7e87b9f7bef04f/webpage.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/06/15 07:56:41 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/df66110866f3dd847f7e87b9f7bef04f/webpage.jar
training@localhost:~/t... hdfs mysql Hue - File Browser - M...
```

The screenshot shows a terminal window with the following content:

```

FILE: Number of write operations=0
HDFS: Number of bytes read=480
HDFS: Number of bytes written=3203
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)=0
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=25942
  Total vcore-seconds taken by all map tasks=25942
  Total megabyte-seconds taken by all map tasks=6641152

Map-Reduce Framework
  Map input records=53
  Map output records=53
  Input split bytes=480
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=410
  CPU time spent (ms)=4400
  Physical memory (bytes) snapshot=476438528
  Virtual memory (bytes) snapshot=3378053120
  Total committed heap usage (bytes)=191889408

File Input Format Counters
  Bytes Read=0

File Output Format Counters
  Bytes Written=3203

23/06/15 07:57:33 INFO mapreduce.ImportJobBase: Transferred 3.1279 KB in 49.9131 seconds (64.1715 bytes/sec)
23/06/15 07:57:33 INFO mapreduce.ImportJobBase: Retrieved 53 records.
[training@localhost sqoop]$
  
```



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player  Applications  Places  System  Thu Jun 15, 8:02 AM
training@localhost:~/training_materials/dev1/exercises/sqoop
File Edit View Search Terminal Help
sqoop import \
--connect jdbc:mysql://localhost/loudacre \
--username training --password training \
--table webpage \
--target-dir /loudacre/webpage \
--fields-terminated-by "\t"
6,2
All
training@localhost:~/t... hdfs mysql Hue - File Browser - M...
```

