

SQOOP

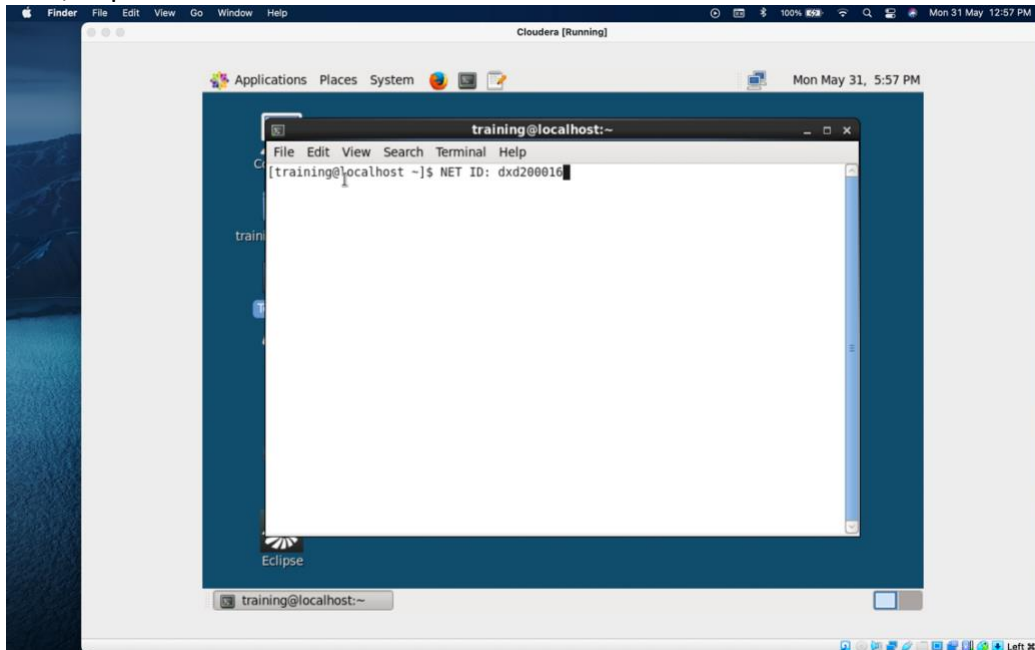
Divyansh Dahiya

LAB Chapter 05: SQOOP

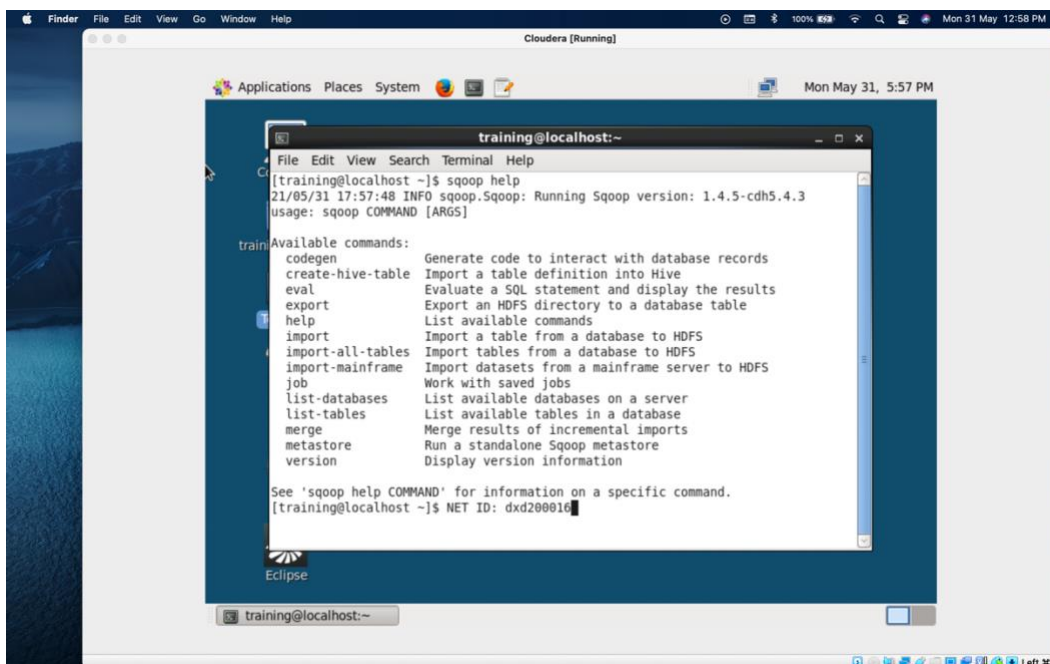
Import data using MySQL using Sqoop

In this lab, I practiced importing MySQL tables into HDFS using Sqoop.

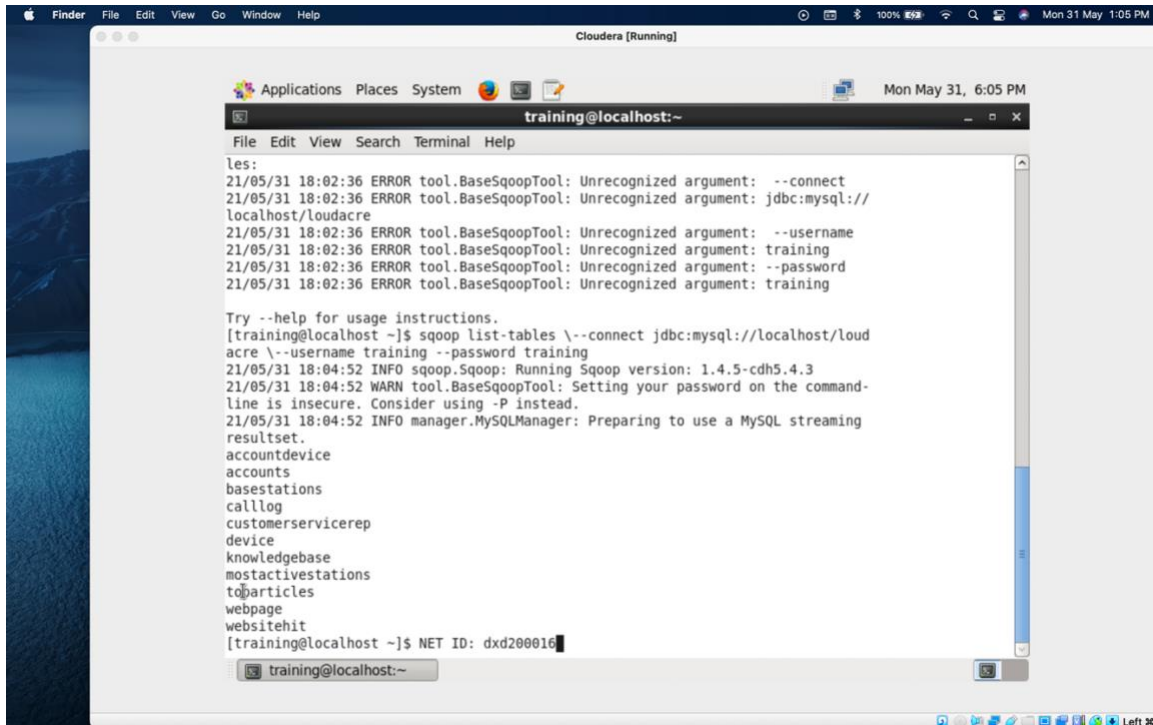
1. First, I open the command line terminal window.



2. Now to get used to all the commands available in the sqoop I ran the following command: sqoop help



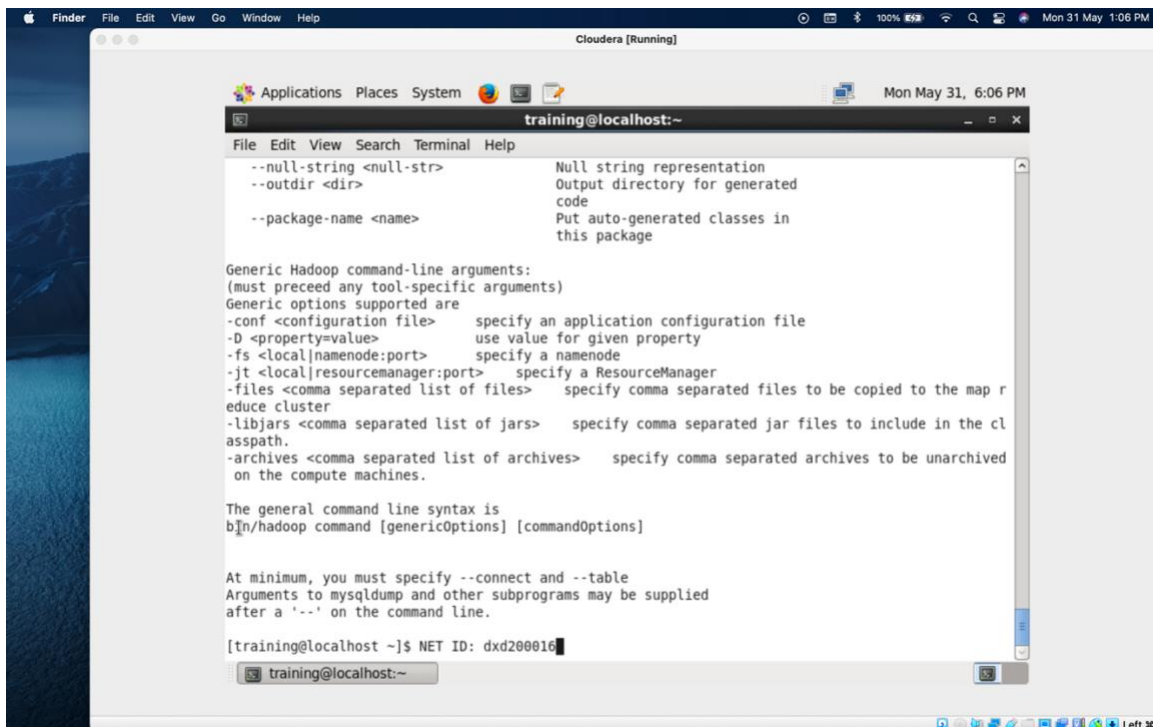
3. Then, to list the tables in the loudacre database I ran the following command:
sqoop list-tables \ --connect jdbc:mysql://localhost/loudacre \ --username training --password training



```
training@localhost:~$ sqoop list-tables \ --connect jdbc:mysql://localhost/loudacre \ --username training --password training
21/05/31 18:02:36 ERROR tool.BaseSqoopTool: Unrecognized argument: --connect
21/05/31 18:02:36 ERROR tool.BaseSqoopTool: Unrecognized argument: jdbc:mysql://localhost/loudacre
21/05/31 18:02:36 ERROR tool.BaseSqoopTool: Unrecognized argument: --username
21/05/31 18:02:36 ERROR tool.BaseSqoopTool: Unrecognized argument: training
21/05/31 18:02:36 ERROR tool.BaseSqoopTool: Unrecognized argument: --password
21/05/31 18:02:36 ERROR tool.BaseSqoopTool: Unrecognized argument: training

Try --help for usage instructions.
[training@localhost ~]$ sqoop list-tables \ --connect jdbc:mysql://localhost/loudacre \ --username training --password training
21/05/31 18:04:52 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
21/05/31 18:04:52 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/05/31 18:04:52 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
accountdevice
accounts
basestations
callog
customerservicerep
device
knowledgebase
mostactivestations
toparticles
webpage
websitehit
[training@localhost ~]$ NET ID: dxd200016
```

4. Now again to see all the command options for importing in sqoop, I ran the following command: sqoop import --help



```
training@localhost:~$ sqoop import --help
--null-string <null-str>      Null string representation
--outdir <dir>                Output directory for generated code
--package-name <name>        Put auto-generated classes in this package

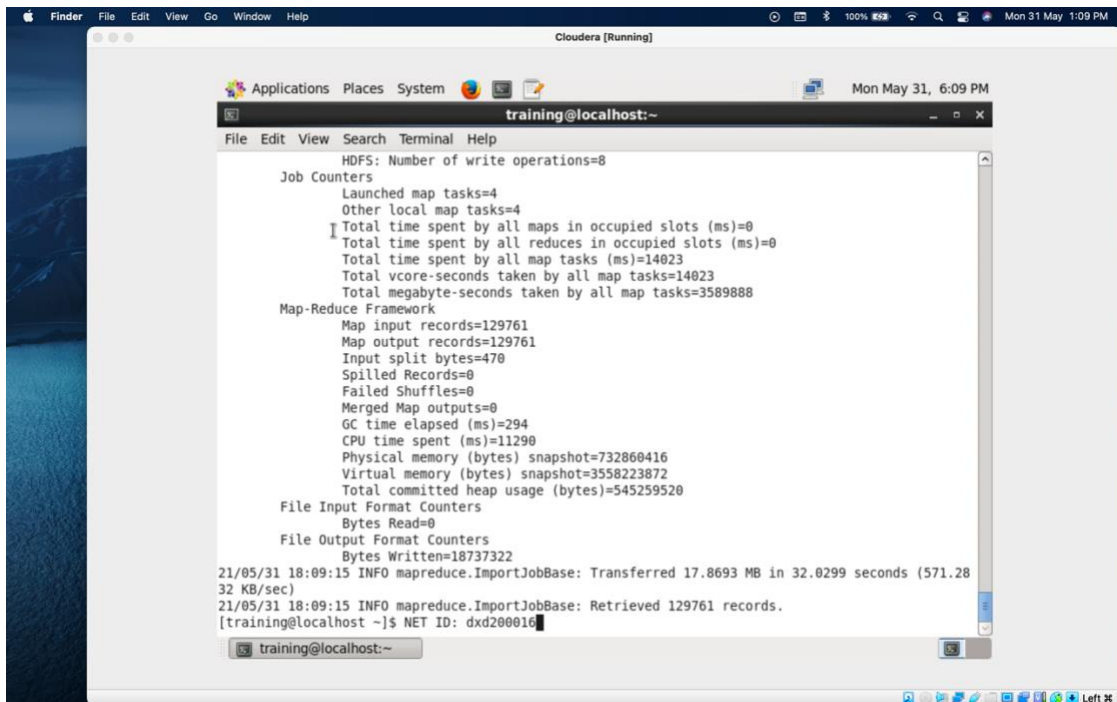
Generic Hadoop command-line arguments:
(must precede any tool-specific arguments)
Generic options supported are
-D <configuration file>      specify an application configuration file
-D <property=value>          use value for given property
-fs <local|namenode:port>    specify a namenode
-jt <local|resourcemanager:port> specify a ResourceManager
-files <comma separated list of files> specify comma separated files to be copied to the map reduce cluster
-libjars <comma separated list of jars> specify comma separated jar files to include in the classpath.
-archives <comma separated list of archives> specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is
bin/hadoop command [genericOptions] [commandOptions]

At minimum, you must specify --connect and --table
Arguments to mysqldump and other subprograms may be supplied after a '--' on the command line.

[training@localhost ~]$ NET ID: dxd200016
```

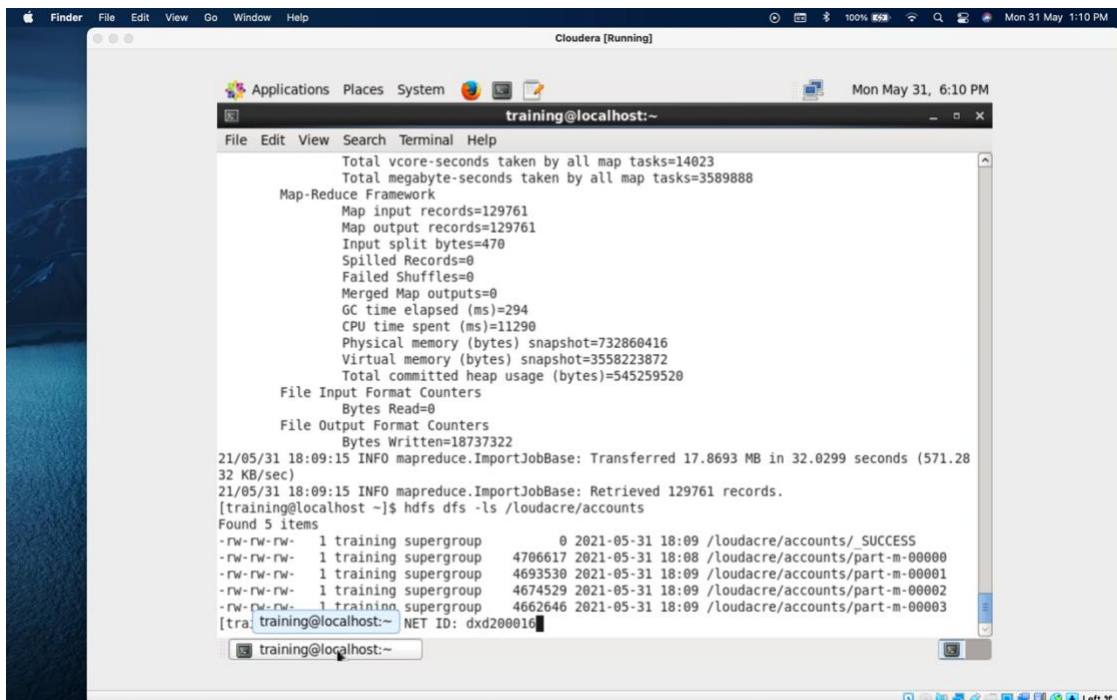
5. Now, finally I imported the accounts table in the loudacre database saved it in the HDFS.
Command: `sqoop import \ --connect jdbc:mysql://localhost/loudacre \ --username training --password training \ --table accounts \ --target-dir /loudacre/accounts \ --null-non-string '\\N'`



```
training@localhost:~$ sqoop import \
--connect jdbc:mysql://localhost/loudacre \
--username training --password training \
--table accounts \
--target-dir /loudacre/accounts \
--null-non-string '\\N'

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)=14023
  Total vcore-seconds taken by all map tasks=14023
  Total megabyte-seconds taken by all map tasks=3589888
Map-Reduce Framework
  Map input records=129761
  Map output records=129761
  Input split bytes=470
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=294
  CPU time spent (ms)=11290
  Physical memory (bytes) snapshot=732860416
  Virtual memory (bytes) snapshot=3558223872
  Total committed heap usage (bytes)=545259520
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=18737322
21/05/31 18:09:15 INFO mapreduce.ImportJobBase: Transferred 17.8693 MB in 32.0299 seconds (571.28
32 KB/sec)
21/05/31 18:09:15 INFO mapreduce.ImportJobBase: Retrieved 129761 records.
[training@localhost ~]$ NET ID: dxd200016
```

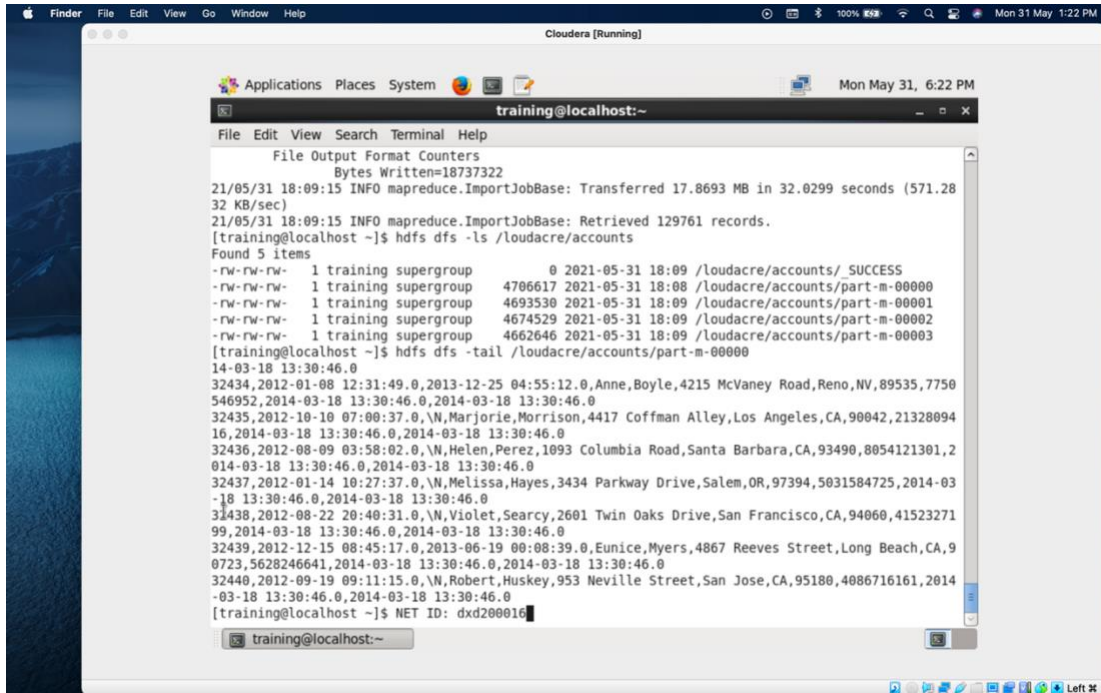
6. Now to view the contents of accounts, I ran the following command: `hdfs dfs -ls /loudacre/accounts`



```
training@localhost:~$ hdfs dfs -ls /loudacre/accounts

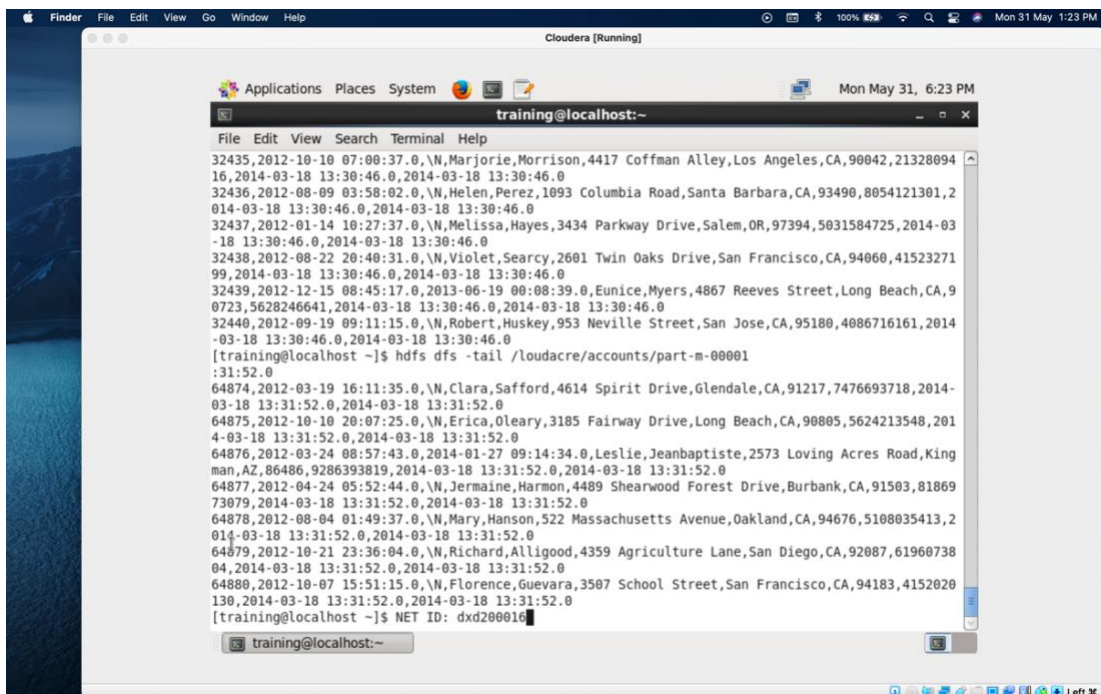
Found 5 items
-rw-rw-rw- 1 training supergroup          0 2021-05-31 18:09 /loudacre/accounts/ SUCCESS
-rw-rw-rw- 1 training supergroup 4706617 2021-05-31 18:08 /loudacre/accounts/part-m-00000
-rw-rw-rw- 1 training supergroup 4693530 2021-05-31 18:09 /loudacre/accounts/part-m-00001
-rw-rw-rw- 1 training supergroup 4674529 2021-05-31 18:09 /loudacre/accounts/part-m-00002
-rw-rw-rw- 1 training supergroup 4662646 2021-05-31 18:09 /loudacre/accounts/part-m-00003
[training@localhost ~]$ NET ID: dxd200016
```

7. Then I ran the tail command to view the last of the file part-m-00000.
Command: `hdfs dfs -tail /loudacre/accounts/part-m-00000`



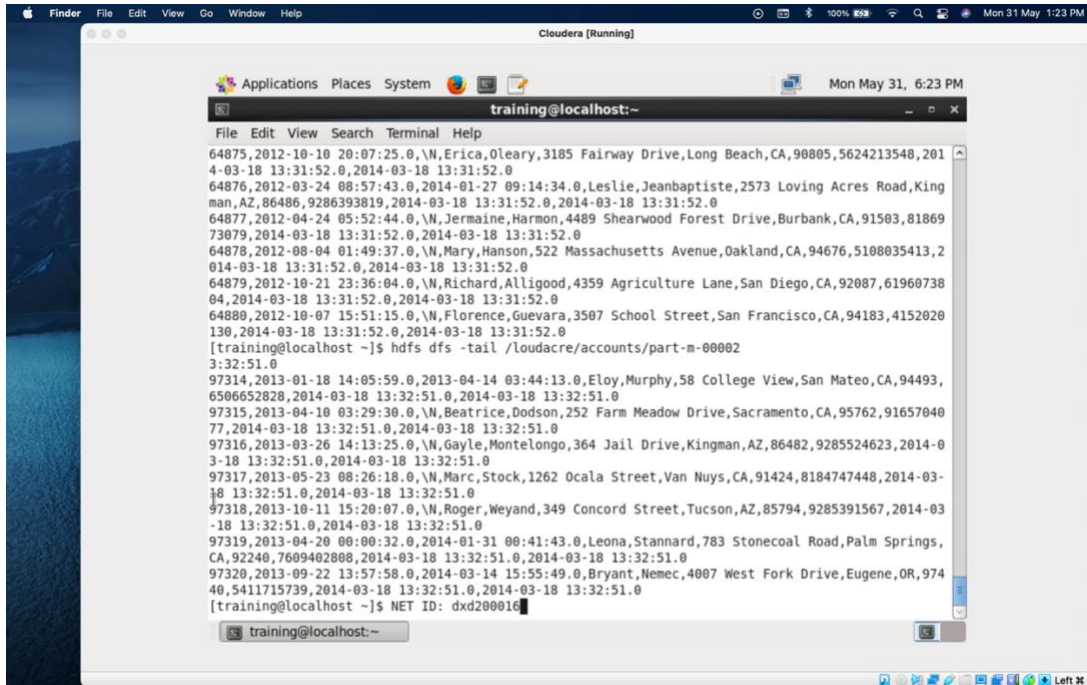
```
training@localhost:~  
File Edit View Search Terminal Help  
File Output Format Counters  
Bytes Written=18737322  
21/05/31 18:09:15 INFO mapreduce.ImportJobBase: Transferred 17.8693 MB in 32.0299 seconds (571.28  
32 KB/sec)  
21/05/31 18:09:15 INFO mapreduce.ImportJobBase: Retrieved 129761 records.  
[training@localhost ~]$ hdfs dfs -ls /loudacre/accounts  
Found 5 items  
-rw-rw-rw- 1 training supergroup 0 2021-05-31 18:09 /loudacre/accounts/_SUCCESS  
-rw-rw-rw- 1 training supergroup 4706617 2021-05-31 18:08 /loudacre/accounts/part-m-00000  
-rw-rw-rw- 1 training supergroup 4693530 2021-05-31 18:09 /loudacre/accounts/part-m-00001  
-rw-rw-rw- 1 training supergroup 4674529 2021-05-31 18:09 /loudacre/accounts/part-m-00002  
-rw-rw-rw- 1 training supergroup 4662646 2021-05-31 18:09 /loudacre/accounts/part-m-00003  
[training@localhost ~]$ hdfs dfs -tail /loudacre/accounts/part-m-00000  
14-03-18 13:30:46.0  
32434,2012-01-08 12:31:49.0,2013-12-25 04:55:12.0,Anne,Boyle,4215 McVane Road,Reno,NV,89535,7750  
546952,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32435,2012-10-10 07:00:37.0,\N,Marjorie,Morrison,4417 Coffman Alley,Los Angeles,CA,90042,21328094  
16,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32436,2012-08-09 03:58:02.0,\N,Helen,Perez,1093 Columbia Road,Santa Barbara,CA,93490,8054121301,2  
014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32437,2012-01-14 10:27:37.0,\N,Melissa,Hayes,3434 Parkway Drive,Salem,OR,97394,5031584725,2014-03-  
-18 13:30:46.0,2014-03-18 13:30:46.0  
32438,2012-08-22 20:40:31.0,\N,Violet,Searcy,2601 Twin Oaks Drive,San Francisco,CA,94060,41523271  
99,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32439,2012-12-15 08:45:17.0,2013-06-19 00:08:39.0,Eunice,Myers,4867 Reeves Street,Long Beach,CA,9  
0723,5628246641,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32440,2012-09-19 09:11:15.0,\N,Robert,Huskey,953 Neville Street,San Jose,CA,95180,4086716161,2014  
-03-18 13:30:46.0,2014-03-18 13:30:46.0  
[training@localhost ~]$ NET ID: dxd200016
```

8. Then I ran the tail command to view the last of the file part-m-00001.
Command: `hdfs dfs -tail /loudacre/accounts/part-m-00001`



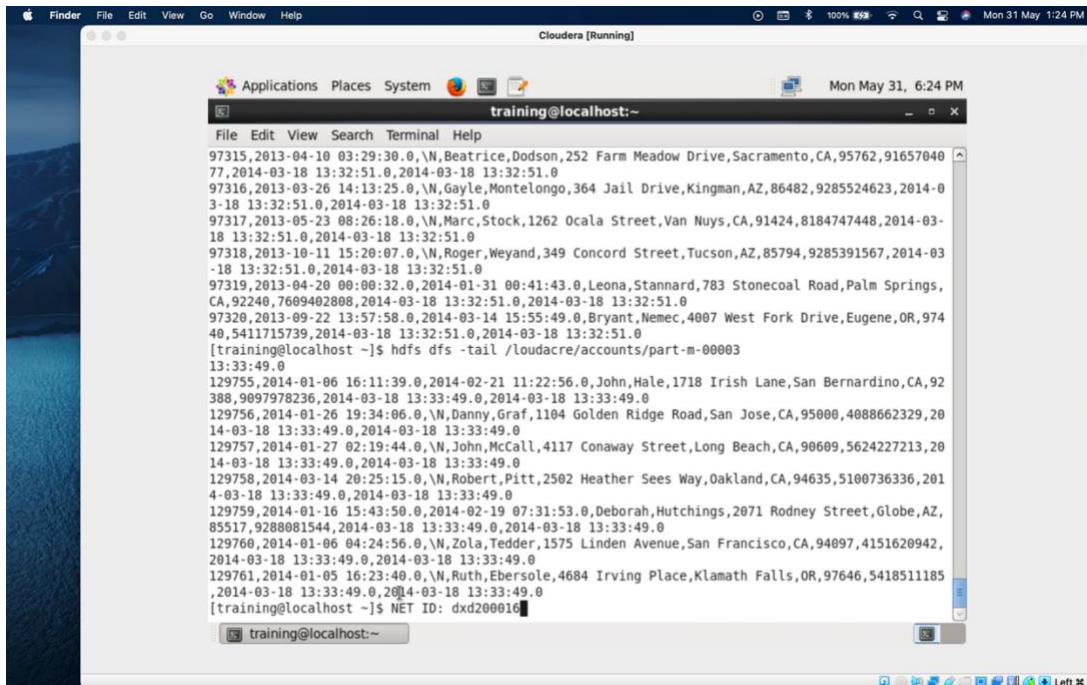
```
training@localhost:~  
File Edit View Search Terminal Help  
32435,2012-10-10 07:00:37.0,\N,Marjorie,Morrison,4417 Coffman Alley,Los Angeles,CA,90042,21328094  
16,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32436,2012-08-09 03:58:02.0,\N,Helen,Perez,1093 Columbia Road,Santa Barbara,CA,93490,8054121301,2  
014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32437,2012-01-14 10:27:37.0,\N,Melissa,Hayes,3434 Parkway Drive,Salem,OR,97394,5031584725,2014-03-  
-18 13:30:46.0,2014-03-18 13:30:46.0  
32438,2012-08-22 20:40:31.0,\N,Violet,Searcy,2601 Twin Oaks Drive,San Francisco,CA,94060,41523271  
99,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32439,2012-12-15 08:45:17.0,2013-06-19 00:08:39.0,Eunice,Myers,4867 Reeves Street,Long Beach,CA,9  
0723,5628246641,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0  
32440,2012-09-19 09:11:15.0,\N,Robert,Huskey,953 Neville Street,San Jose,CA,95180,4086716161,2014  
-03-18 13:30:46.0,2014-03-18 13:30:46.0  
[training@localhost ~]$ hdfs dfs -tail /loudacre/accounts/part-m-00001  
:31:52.0  
64874,2012-03-19 16:11:35.0,\N,Clara,Safford,4614 Spirit Drive,Glendale,CA,91217,7476693718,2014-  
03-18 13:31:52.0,2014-03-18 13:31:52.0  
64875,2012-10-10 20:07:25.0,\N,Erica,Oleary,3185 Fairway Drive,Long Beach,CA,90805,5624213548,201  
4-03-18 13:31:52.0,2014-03-18 13:31:52.0  
64876,2012-03-24 08:57:43.0,2014-01-27 09:14:34.0,Leslie,Jeanbaptiste,2573 Loving Acres Road,King  
man,AZ,86486,9286393819,2014-03-18 13:31:52.0,2014-03-18 13:31:52.0  
64877,2012-04-24 05:52:44.0,\N,Jermaine,Harmon,4489 Shearwood Forest Drive,Burbank,CA,91503,81869  
73079,2014-03-18 13:31:52.0,2014-03-18 13:31:52.0  
64878,2012-08-04 01:49:37.0,\N,Mary,Hanson,522 Massachusetts Avenue,Oakland,CA,94676,5108035413,2  
014-03-18 13:31:52.0,2014-03-18 13:31:52.0  
64879,2012-10-21 23:36:04.0,\N,Richard,Alligood,4359 Agriculture Lane,San Diego,CA,92087,61960738  
04,2014-03-18 13:31:52.0,2014-03-18 13:31:52.0  
64880,2012-10-07 15:51:15.0,\N,Florence,Guevara,3507 School Street,San Francisco,CA,94183,4152020  
130,2014-03-18 13:31:52.0,2014-03-18 13:31:52.0  
[training@localhost ~]$ NET ID: dxd200016
```


9. Then I ran the tail command to view the last of the file part-m-00002.
Command: `hdfs dfs -tail /loudacre/accounts/part-m-00002`



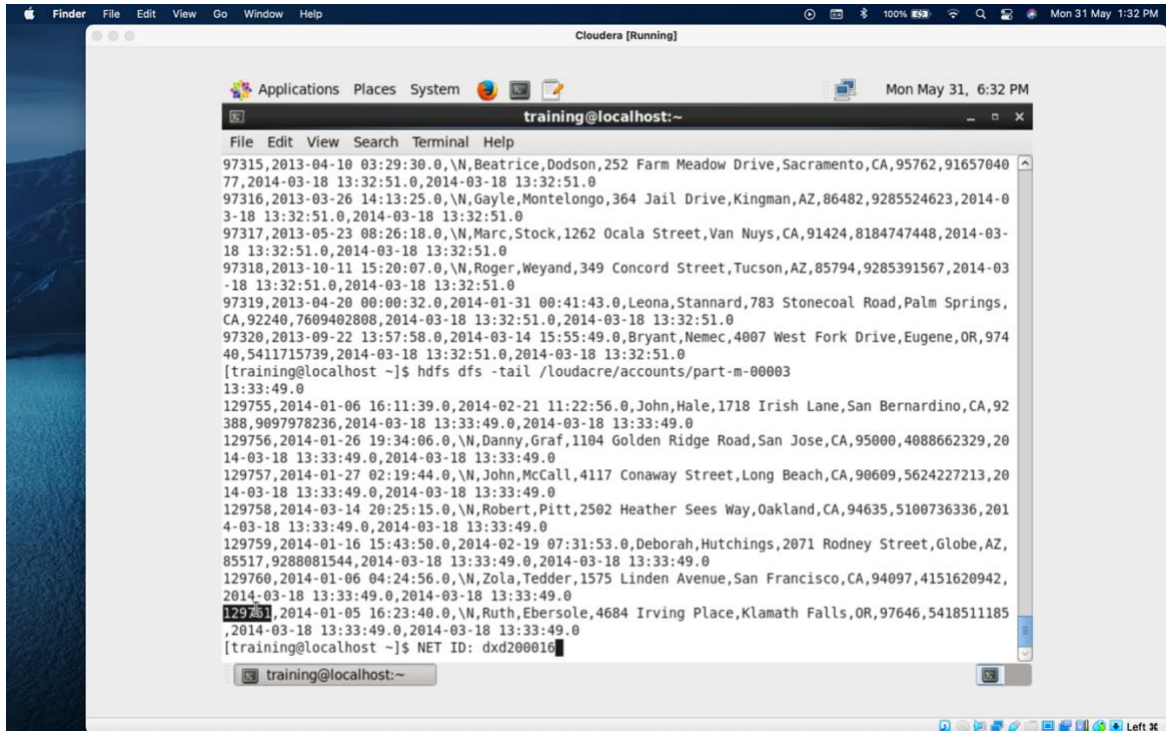
```
training@localhost:~$ hdfs dfs -tail /loudacre/accounts/part-m-00002
64075,2012-10-10 20:07:25.0,\N,Erica,Oleary,3185 Fairway Drive,Long Beach,CA,90805,5624213548,201
4-03-18 13:31:52.0,2014-03-18 13:31:52.0
64076,2012-03-24 08:57:43.0,2014-01-27 09:14:34.0,Leslie,Jeanbaptiste,2573 Loving Acres Road,King
man,AZ,86486,9286393819,2014-03-18 13:31:52.0,2014-03-18 13:31:52.0
64077,2012-04-24 05:52:44.0,\N,Jermaine,Harmon,4489 Shearwood Forest Drive,Burbank,CA,91503,81869
73079,2014-03-18 13:31:52.0,2014-03-18 13:31:52.0
64078,2012-08-04 01:49:37.0,\N,Mary,Hanson,522 Massachusetts Avenue,Oakland,CA,94676,5108035413,2
014-03-18 13:31:52.0,2014-03-18 13:31:52.0
64079,2012-10-21 23:36:04.0,\N,Richard,Alligood,4359 Agriculture Lane,San Diego,CA,92087,61960738
04,2014-03-18 13:31:52.0,2014-03-18 13:31:52.0
64080,2012-10-07 15:51:15.0,\N,Florence,Guevara,3507 School Street,San Francisco,CA,94183,4152020
130,2014-03-18 13:31:52.0,2014-03-18 13:31:52.0
[training@localhost ~]$ hdfs dfs -tail /loudacre/accounts/part-m-00002
3:32:51.0
97314,2013-01-18 14:05:59.0,2013-04-14 03:44:13.0,Eloy,Murphy,58 College View,San Mateo,CA,94493,
6506652828,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0
97315,2013-04-10 03:29:30.0,\N,Beatrice,Dodson,252 Farm Meadow Drive,Sacramento,CA,95762,91657040
77,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0
97316,2013-03-26 14:13:25.0,\N,Gayle,Montelongo,364 Jail Drive,Kingman,AZ,86482,9285524623,2014-0
3-18 13:32:51.0,2014-03-18 13:32:51.0
97317,2013-05-23 08:26:18.0,\N,Marc,Stock,1262 Ocala Street,Van Nuys,CA,91424,8184747448,2014-03-
18 13:32:51.0,2014-03-18 13:32:51.0
97318,2013-10-11 15:20:07.0,\N,Roger,Weyand,349 Concord Street,Tucson,AZ,85794,9285391567,2014-03-
18 13:32:51.0,2014-03-18 13:32:51.0
97319,2013-04-20 00:00:32.0,2014-01-31 00:41:43.0,Leona,Stannard,783 Stonecoal Road,Palm Springs,
CA,92240,7609402808,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0
97320,2013-09-22 13:57:58.0,2014-03-14 15:55:49.0,Bryant,Nemec,4007 West Fork Drive,Eugene,OR,974
40,5411715739,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0
[training@localhost ~]$ NET ID: dxd200016
```

10. Then I ran the tail command to view the last of the file part-m-00003.
Command: `hdfs dfs -tail /loudacre/accounts/part-m-00003`



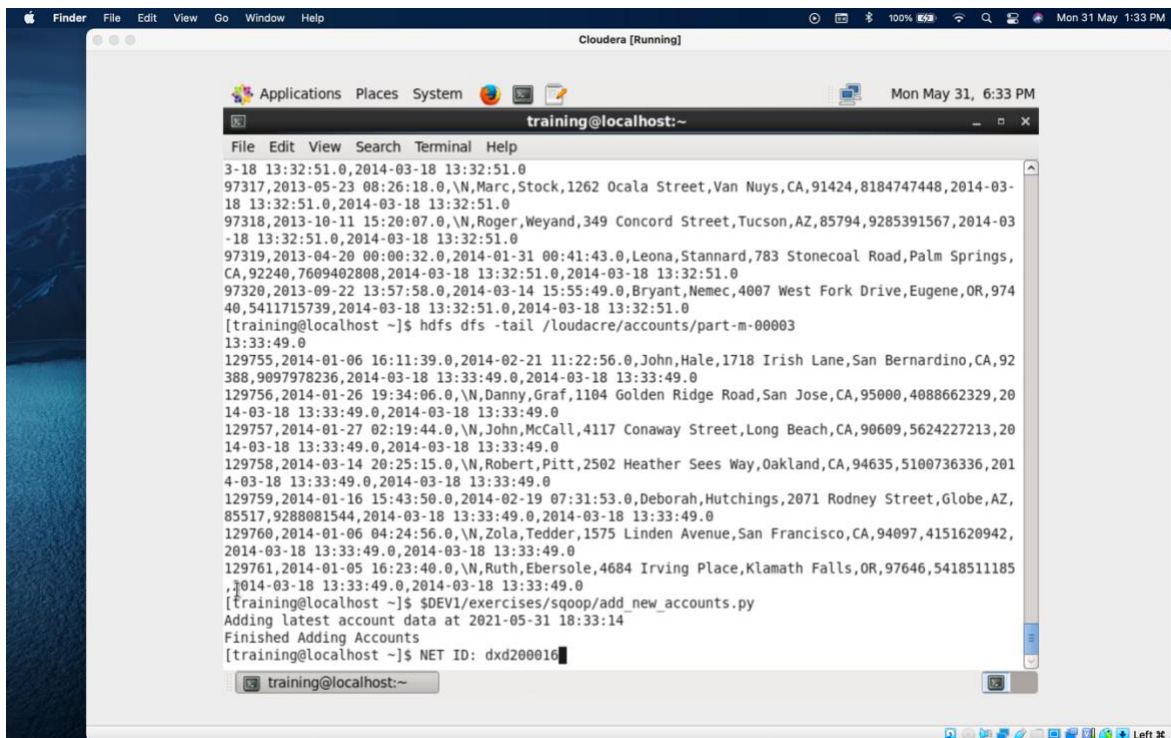
```
training@localhost:~$ hdfs dfs -tail /loudacre/accounts/part-m-00003
97315,2013-04-10 03:29:30.0,\N,Beatrice,Dodson,252 Farm Meadow Drive,Sacramento,CA,95762,91657040
77,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0
97316,2013-03-26 14:13:25.0,\N,Gayle,Montelongo,364 Jail Drive,Kingman,AZ,86482,9285524623,2014-0
3-18 13:32:51.0,2014-03-18 13:32:51.0
97317,2013-05-23 08:26:18.0,\N,Marc,Stock,1262 Ocala Street,Van Nuys,CA,91424,8184747448,2014-03-
18 13:32:51.0,2014-03-18 13:32:51.0
97318,2013-10-11 15:20:07.0,\N,Roger,Weyand,349 Concord Street,Tucson,AZ,85794,9285391567,2014-03-
18 13:32:51.0,2014-03-18 13:32:51.0
97319,2013-04-20 00:00:32.0,2014-01-31 00:41:43.0,Leona,Stannard,783 Stonecoal Road,Palm Springs,
CA,92240,7609402808,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0
97320,2013-09-22 13:57:58.0,2014-03-14 15:55:49.0,Bryant,Nemec,4007 West Fork Drive,Eugene,OR,974
40,5411715739,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0
[training@localhost ~]$ hdfs dfs -tail /loudacre/accounts/part-m-00003
13:33:49.0
129755,2014-01-06 16:11:39.0,2014-02-21 11:22:56.0,John,Hale,1718 Irish Lane,San Bernardino,CA,92
388,9097978236,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0
129756,2014-01-26 19:34:06.0,\N,Danny,Graf,1104 Golden Ridge Road,San Jose,CA,95000,4088662329,20
14-03-18 13:33:49.0,2014-03-18 13:33:49.0
129757,2014-01-27 02:19:44.0,\N,John,McCall,4117 Conaway Street,Long Beach,CA,90609,5624227213,20
14-03-18 13:33:49.0,2014-03-18 13:33:49.0
129758,2014-03-14 20:25:15.0,\N,Robert,Pitt,2502 Heather Sees Way,Oakland,CA,94635,5100736336,201
4-03-18 13:33:49.0,2014-03-18 13:33:49.0
129759,2014-01-16 15:43:50.0,2014-02-19 07:31:53.0,Deborah,Hutchings,2071 Rodney Street,Globe,AZ,
85517,9288081544,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0
129760,2014-01-06 04:24:56.0,\N,Zola,Tedder,1575 Linden Avenue,San Francisco,CA,94097,4151620942,
2014-03-18 13:33:49.0,2014-03-18 13:33:49.0
129761,2014-01-05 16:23:40.0,\N,Ruth,Ebersole,4684 Irving Place,Klamath Falls,OR,97646,5418511185
,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0
[training@localhost ~]$ NET ID: dxd200016
```

11. Then I take a note of the highest account ID in all the files, which was: 129761



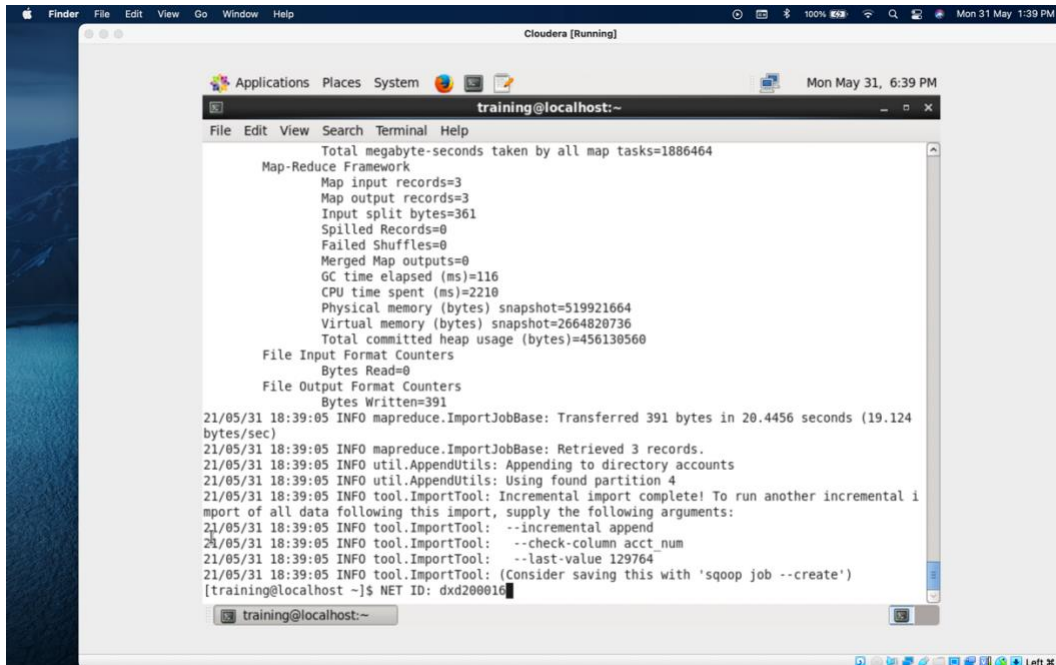
```
training@localhost:~  
File Edit View Search Terminal Help  
97315,2013-04-10 03:29:30.0,\N,Beatrice,Dodson,252 Farm Meadow Drive,Sacramento,CA,95762,91657040  
77,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0  
97316,2013-03-26 14:13:25.0,\N,Gayle,Montelongo,364 Jail Drive,Kingman,AZ,86482,9285524623,2014-0  
3-18 13:32:51.0,2014-03-18 13:32:51.0  
97317,2013-05-23 08:26:18.0,\N,Marc,Stock,1262 Ocala Street,Van Nuys,CA,91424,8184747448,2014-03-  
18 13:32:51.0,2014-03-18 13:32:51.0  
97318,2013-10-11 15:20:07.0,\N,Roger,Weyand,349 Concord Street,Tucson,AZ,85794,9285391567,2014-03-  
18 13:32:51.0,2014-03-18 13:32:51.0  
97319,2013-04-20 00:00:32.0,2014-01-31 00:41:43.0,Leona,Stannard,783 Stonecoal Road,Palm Springs,  
CA,92240,7609402808,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0  
97320,2013-09-22 13:57:58.0,2014-03-14 15:55:49.0,Bryant,Nemec,4007 West Fork Drive,Eugene,OR,974  
40,5411715739,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0  
[training@localhost ~]$ hdfs dfs -tail /loudacre/accounts/part-m-00003  
13:33:49.0  
129755,2014-01-06 16:11:39.0,2014-02-21 11:22:56.0,John,Hale,1718 Irish Lane,San Bernardino,CA,92  
388,9097978236,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129756,2014-01-26 19:34:06.0,\N,Danny,Graf,1104 Golden Ridge Road,San Jose,CA,95000,4088662329,20  
14-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129757,2014-01-27 02:19:44.0,\N,John,McCall,4117 Conaway Street,Long Beach,CA,90609,5624227213,20  
14-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129758,2014-03-14 20:25:15.0,\N,Robert,Pitt,2502 Heather Sees Way,Oakland,CA,94635,5100736336,201  
4-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129759,2014-01-16 15:43:50.0,2014-02-19 07:31:53.0,Deborah,Hutchings,2071 Rodney Street,Globe,AZ,  
85517,9288081544,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129760,2014-01-06 04:24:56.0,\N,Zola,Tedder,1575 Linden Avenue,San Francisco,CA,94097,4151620942,  
2014-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129761,2014-01-05 16:23:40.0,\N,Ruth,Ebersole,4684 Irving Place,Klamath Falls,OR,97646,5418511185  
,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0  
[training@localhost ~]$ NET ID: dxd200016
```

12. Now, I ran the add_new_accounts.py script to add the latest accounts using MySQL
Command: \$DEV1/exercises/sqoop/add_new_accounts.py



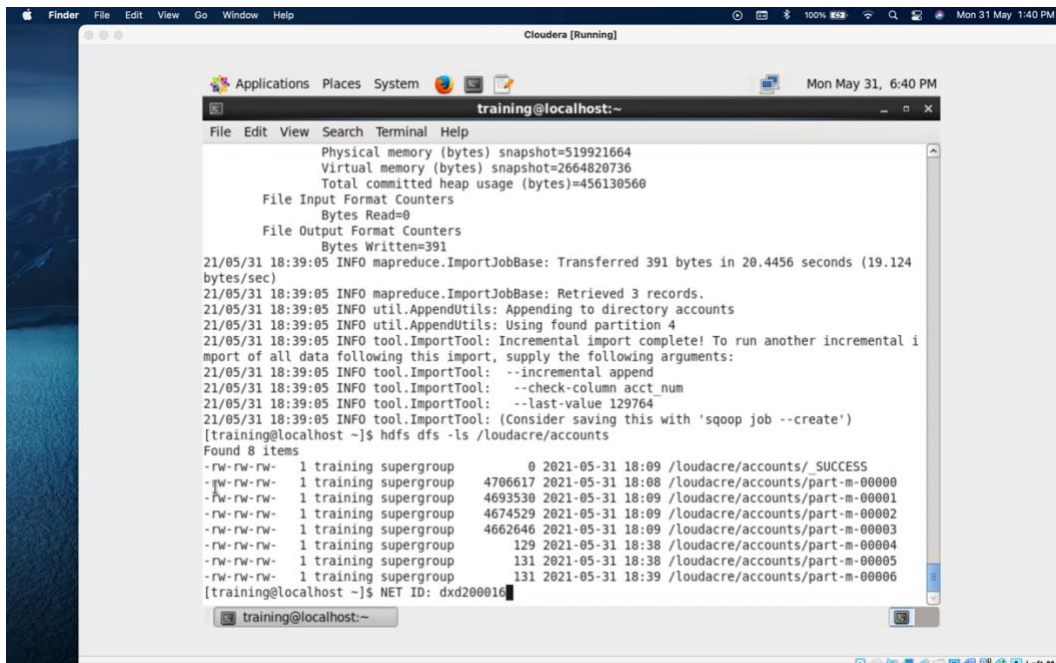
```
training@localhost:~  
File Edit View Search Terminal Help  
3-18 13:32:51.0,2014-03-18 13:32:51.0  
97317,2013-05-23 08:26:18.0,\N,Marc,Stock,1262 Ocala Street,Van Nuys,CA,91424,8184747448,2014-03-  
18 13:32:51.0,2014-03-18 13:32:51.0  
97318,2013-10-11 15:20:07.0,\N,Roger,Weyand,349 Concord Street,Tucson,AZ,85794,9285391567,2014-03-  
18 13:32:51.0,2014-03-18 13:32:51.0  
97319,2013-04-20 00:00:32.0,2014-01-31 00:41:43.0,Leona,Stannard,783 Stonecoal Road,Palm Springs,  
CA,92240,7609402808,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0  
97320,2013-09-22 13:57:58.0,2014-03-14 15:55:49.0,Bryant,Nemec,4007 West Fork Drive,Eugene,OR,974  
40,5411715739,2014-03-18 13:32:51.0,2014-03-18 13:32:51.0  
[training@localhost ~]$ hdfs dfs -tail /loudacre/accounts/part-m-00003  
13:33:49.0  
129755,2014-01-06 16:11:39.0,2014-02-21 11:22:56.0,John,Hale,1718 Irish Lane,San Bernardino,CA,92  
388,9097978236,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129756,2014-01-26 19:34:06.0,\N,Danny,Graf,1104 Golden Ridge Road,San Jose,CA,95000,4088662329,20  
14-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129757,2014-01-27 02:19:44.0,\N,John,McCall,4117 Conaway Street,Long Beach,CA,90609,5624227213,20  
14-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129758,2014-03-14 20:25:15.0,\N,Robert,Pitt,2502 Heather Sees Way,Oakland,CA,94635,5100736336,201  
4-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129759,2014-01-16 15:43:50.0,2014-02-19 07:31:53.0,Deborah,Hutchings,2071 Rodney Street,Globe,AZ,  
85517,9288081544,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129760,2014-01-06 04:24:56.0,\N,Zola,Tedder,1575 Linden Avenue,San Francisco,CA,94097,4151620942,  
2014-03-18 13:33:49.0,2014-03-18 13:33:49.0  
129761,2014-01-05 16:23:40.0,\N,Ruth,Ebersole,4684 Irving Place,Klamath Falls,OR,97646,5418511185  
,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0  
[training@localhost ~]$ $DEV1/exercises/sqoop/add_new_accounts.py  
Adding latest account data at 2021-05-31 18:33:14  
Finished Adding Accounts  
[training@localhost ~]$ NET ID: dxd200016
```

13. Now, to incrementally import and append the newly added accounts to the accounts directory, I used the following command: `sqoop import \ --connect jdbc:mysql://localhost/loudacre \ --username training --password training \ --incremental append \ --null-non-string '\\N' \ --table accounts \ --target-dir /loudacre/accounts \ --check-column acct_num \ --last-value 129761`



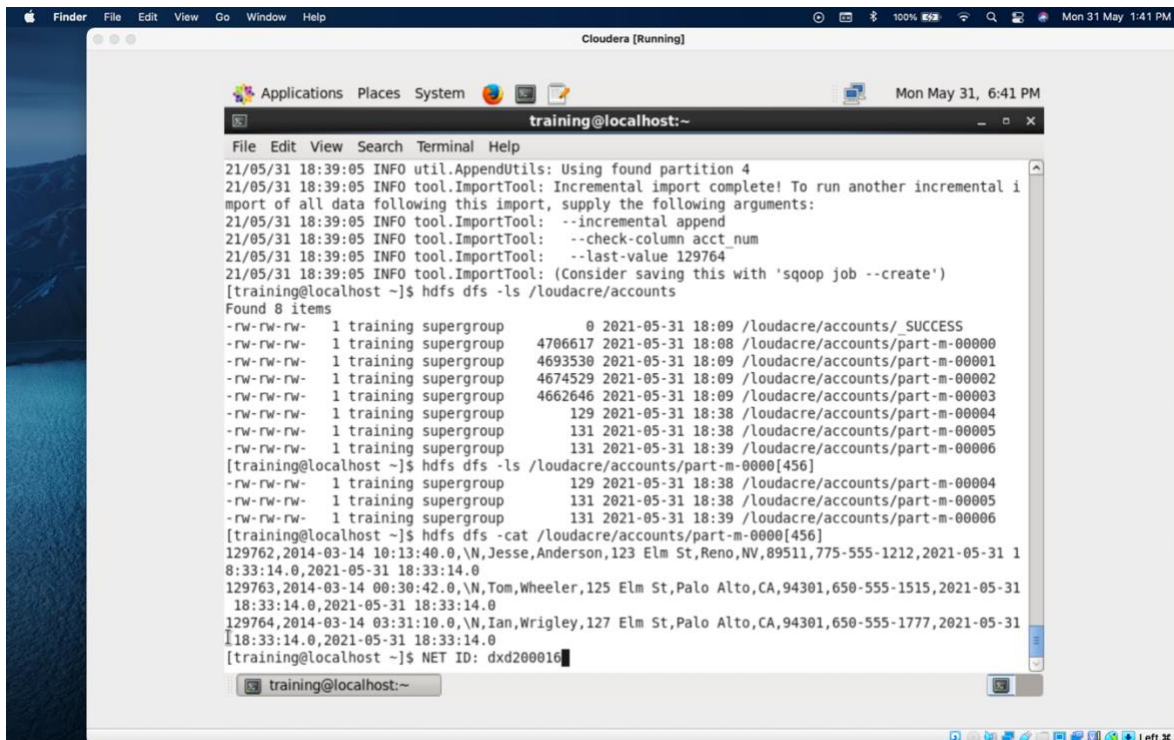
```
training@localhost:~  
File Edit View Search Terminal Help  
Total megabyte-seconds taken by all map tasks=1886464  
Map-Reduce Framework  
  Map input records=3  
  Map output records=3  
  Input split bytes=361  
  Spilled Records=0  
  Failed Shuffles=0  
  Merged Map outputs=0  
  GC time elapsed (ms)=116  
  CPU time spent (ms)=2210  
  Physical memory (bytes) snapshot=519921664  
  Virtual memory (bytes) snapshot=2664820736  
  Total committed heap usage (bytes)=456130560  
File Input Format Counters  
  Bytes Read=0  
File Output Format Counters  
  Bytes Written=391  
21/05/31 18:39:05 INFO mapreduce.ImportJobBase: Transferred 391 bytes in 20.4456 seconds (19.124 bytes/sec)  
21/05/31 18:39:05 INFO mapreduce.ImportJobBase: Retrieved 3 records.  
21/05/31 18:39:05 INFO util.AppendUtils: Appending to directory accounts  
21/05/31 18:39:05 INFO util.AppendUtils: Using found partition 4  
21/05/31 18:39:05 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments:  
21/05/31 18:39:05 INFO tool.ImportTool: --incremental append  
21/05/31 18:39:05 INFO tool.ImportTool: --check-column acct_num  
21/05/31 18:39:05 INFO tool.ImportTool: --last-value 129764  
21/05/31 18:39:05 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')  
[training@localhost ~]$ NET ID: dxd200016
```

14. Now, I listed the contents of the accounts directory to verify the sqoop import.
Command: `hdfs dfs -ls /loudacre/accounts`



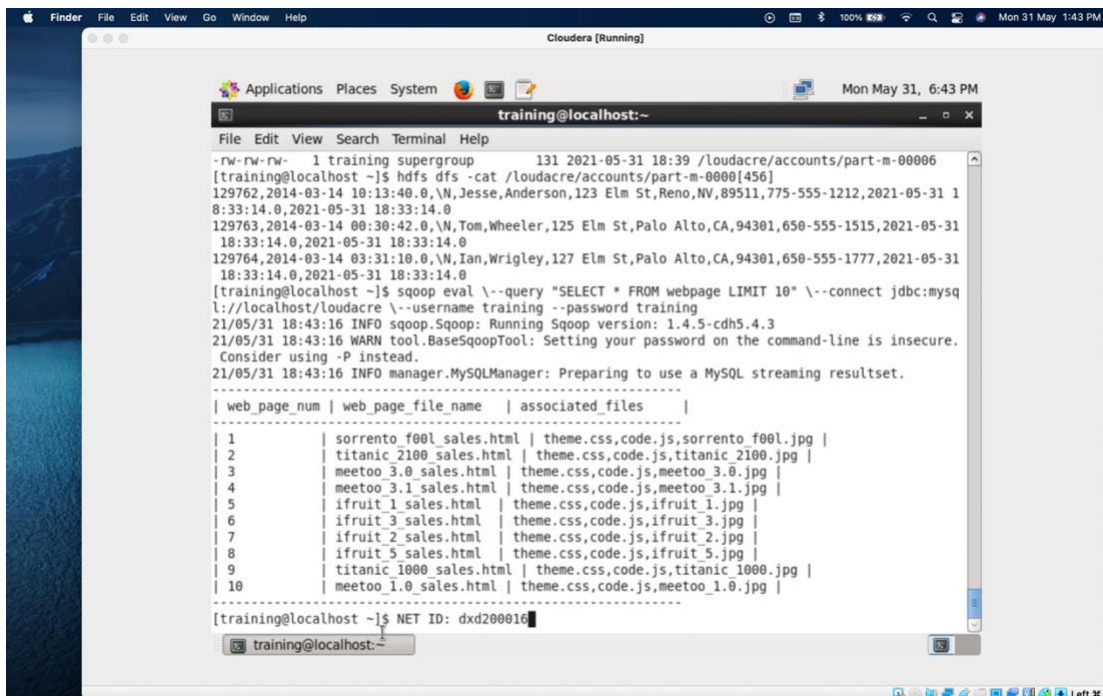
```
training@localhost:~  
File Edit View Search Terminal Help  
Physical memory (bytes) snapshot=519921664  
Virtual memory (bytes) snapshot=2664820736  
Total committed heap usage (bytes)=456130560  
File Input Format Counters  
  Bytes Read=0  
File Output Format Counters  
  Bytes Written=391  
21/05/31 18:39:05 INFO mapreduce.ImportJobBase: Transferred 391 bytes in 20.4456 seconds (19.124 bytes/sec)  
21/05/31 18:39:05 INFO mapreduce.ImportJobBase: Retrieved 3 records.  
21/05/31 18:39:05 INFO util.AppendUtils: Appending to directory accounts  
21/05/31 18:39:05 INFO util.AppendUtils: Using found partition 4  
21/05/31 18:39:05 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments:  
21/05/31 18:39:05 INFO tool.ImportTool: --incremental append  
21/05/31 18:39:05 INFO tool.ImportTool: --check-column acct_num  
21/05/31 18:39:05 INFO tool.ImportTool: --last-value 129764  
21/05/31 18:39:05 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')  
[training@localhost ~]$ hdfs dfs -ls /loudacre/accounts  
Found 8 items  
-rw-rw-rw- 1 training supergroup 0 2021-05-31 18:09 /loudacre/accounts/_SUCCESS  
-rw-rw-rw- 1 training supergroup 4706617 2021-05-31 18:08 /loudacre/accounts/part-m-00000  
-rw-rw-rw- 1 training supergroup 4693530 2021-05-31 18:09 /loudacre/accounts/part-m-00001  
-rw-rw-rw- 1 training supergroup 4674529 2021-05-31 18:09 /loudacre/accounts/part-m-00002  
-rw-rw-rw- 1 training supergroup 4662646 2021-05-31 18:09 /loudacre/accounts/part-m-00003  
-rw-rw-rw- 1 training supergroup 129 2021-05-31 18:38 /loudacre/accounts/part-m-00004  
-rw-rw-rw- 1 training supergroup 131 2021-05-31 18:38 /loudacre/accounts/part-m-00005  
-rw-rw-rw- 1 training supergroup 131 2021-05-31 18:39 /loudacre/accounts/part-m-00006  
[training@localhost ~]$ NET ID: dxd200016
```


15. Now to view the entire contents of these files I ran the following command: `hdfs dfs -cat /loudacre/accounts/part-m-0000[456]`



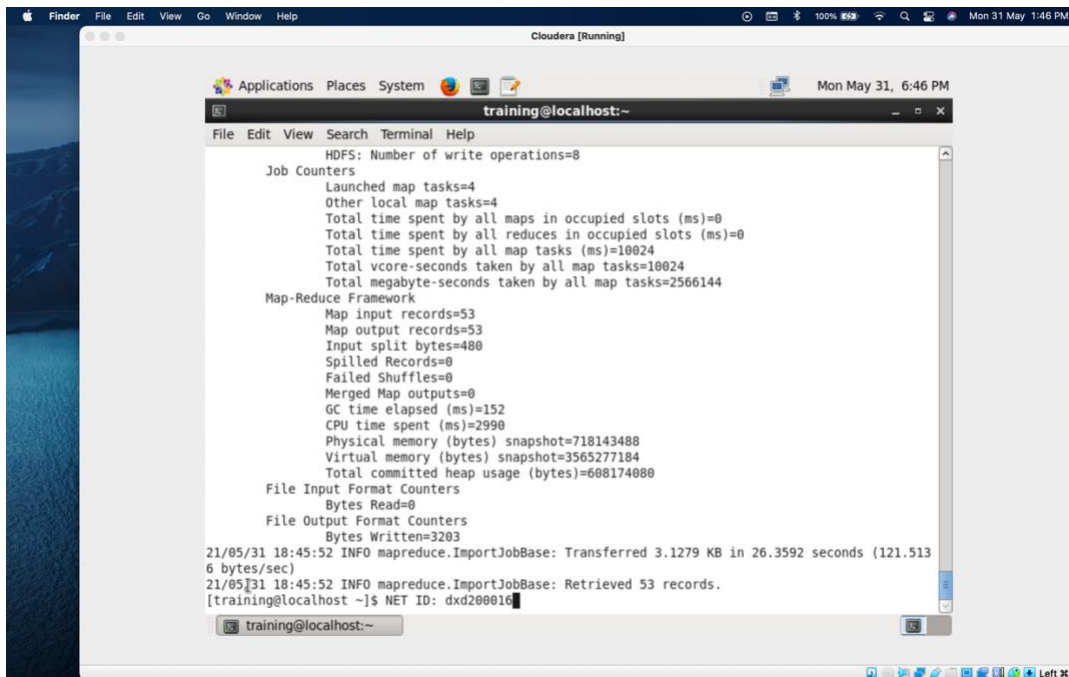
```
training@localhost:~$ hdfs dfs -ls /loudacre/accounts
Found 8 items
-rw-rw-rw- 1 training supergroup          0 2021-05-31 18:09 /loudacre/accounts/ SUCCESS
4706617 2021-05-31 18:08 /loudacre/accounts/part-m-00000
4693530 2021-05-31 18:09 /loudacre/accounts/part-m-00001
4674529 2021-05-31 18:09 /loudacre/accounts/part-m-00002
4662646 2021-05-31 18:09 /loudacre/accounts/part-m-00003
129 2021-05-31 18:38 /loudacre/accounts/part-m-00004
131 2021-05-31 18:38 /loudacre/accounts/part-m-00005
131 2021-05-31 18:39 /loudacre/accounts/part-m-00006
[training@localhost ~]$ hdfs dfs -ls /loudacre/accounts/part-m-0000[456]
-rw-rw-rw- 1 training supergroup          129 2021-05-31 18:38 /loudacre/accounts/part-m-00004
-rw-rw-rw- 1 training supergroup          131 2021-05-31 18:38 /loudacre/accounts/part-m-00005
-rw-rw-rw- 1 training supergroup          131 2021-05-31 18:39 /loudacre/accounts/part-m-00006
[training@localhost ~]$ hdfs dfs -cat /loudacre/accounts/part-m-0000[456]
129762,2014-03-14 10:13:40.0,\N,Jesse,Anderson,123 Elm St,Reno,NV,89511,775-555-1212,2021-05-31 1
8:33:14.0,2021-05-31 18:33:14.0
129763,2014-03-14 00:30:42.0,\N,Tom,Wheeler,125 Elm St,Palo Alto,CA,94301,650-555-1515,2021-05-31
18:33:14.0,2021-05-31 18:33:14.0
129764,2014-03-14 03:31:10.0,\N,Ian,Wrigley,127 Elm St,Palo Alto,CA,94301,650-555-1777,2021-05-31
18:33:14.0,2021-05-31 18:33:14.0
[training@localhost ~]$ NET ID: dxd200016
```

16. Now, to view the records in the webpage table I used the following command: `sqoop eval \ --query "SELECT * FROM webpage LIMIT 10" \ --connect jdbc:mysql://localhost/loudacre \ --username training --password training`



```
training@localhost:~$ sqoop eval \ --query "SELECT * FROM webpage LIMIT 10" \ --connect jdbc:mysql
l://localhost/loudacre \ --username training --password training
21/05/31 18:43:16 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
21/05/31 18:43:16 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure.
Consider using -P instead.
21/05/31 18:43:16 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
+-----+-----+-----+
| web_page_num | web_page_file_name | associated files |
+-----+-----+-----+
| 1 | sorrento f00l sales.html | theme.css,code.js,sorrento f00l.jpg |
| 2 | titanic 2100 sales.html | theme.css,code.js,titanic 2100.jpg |
| 3 | meetoo 3.0 sales.html | theme.css,code.js,meetoo 3.0.jpg |
| 4 | meetoo 3.1 sales.html | theme.css,code.js,meetoo 3.1.jpg |
| 5 | ifruit 1 sales.html | theme.css,code.js,ifruit 1.jpg |
| 6 | ifruit 3 sales.html | theme.css,code.js,ifruit 3.jpg |
| 7 | ifruit 2 sales.html | theme.css,code.js,ifruit 2.jpg |
| 8 | ifruit 5 sales.html | theme.css,code.js,ifruit 5.jpg |
| 9 | titanic 1000 sales.html | theme.css,code.js,titanic 1000.jpg |
| 10 | meetoo 1.0 sales.html | theme.css,code.js,meetoo 1.0.jpg |
+-----+-----+-----+
[training@localhost ~]$ NET ID: dxd200016
```

17. Now, to import the webpage table to HDFS, I ran the following command: `sqoop import \ --connect jdbc:mysql://localhost/loudacre \ --username training --password training \ --table webpage \ --target-dir /loudacre/webpage \ --fields-terminated-by "\t"`



18. Now I opened the HUE to view the data files imported to /loudacre/webpage directory. Files can be visible in the below screenshot.

