BUAN 6346 BIG DATA

DR. JERRY F. PEREZ

Assignment 1

06/01/2020

Jeffrey Chang / jlc190004

Table of Contents of Assignment 1

Lab-Chapter-02-Setup 2

Setup and General Notes 2

Lab-Chapter-03- Access HDFS with Command-Line and Hue 3

Explore the HDFS Command Line Interface 3

Upload Files to HDFS 4

View HDFS files 5

Use the Hue File Browser to browse, view and manage files 7

Lab-Chapter-03-Run a YARN Job 9

Explore the YARN cluster 9

Submit an application to the YARN cluster 11

View the application in the Hue Job Browser 11

View the application in the YARN UI 12

Lab-Chapter-04-Import Data from MySQL Using Sqoop 14

Import the accounts table from MySQL 14

View the imported data 15

Import incremental updates to accounts 16

Import webpage data using an alternate field delimiter 17

Lab-Chapter-02-Setup - Setup and General Notes

This chapter introduces the basic introduction of homework. In the Cloudera's training courses, we use a Virtual Machine running the CentOS Linux distribution and with one DataNote.

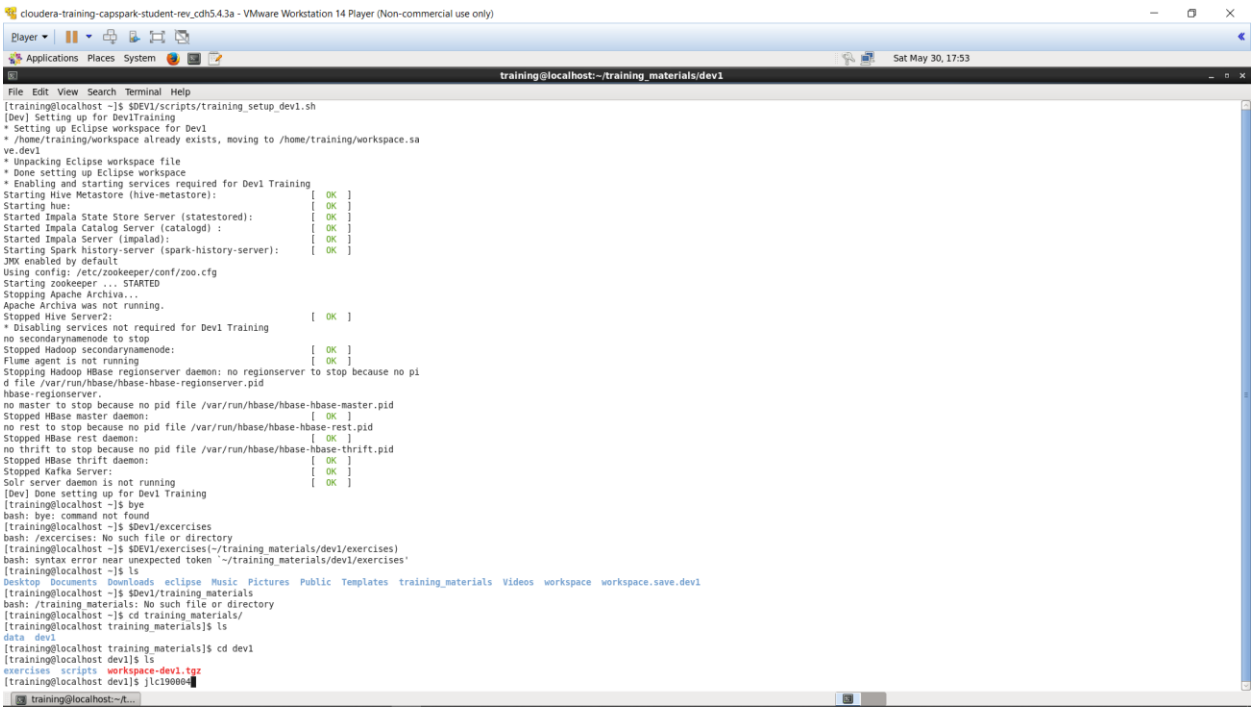
Before starting the homework assignments, run the course setup script in a terminal window:

[$DEV1/scripts/training_setup_dev1.sh]

VM is set to log in as the user training automatically

Find the main directory for the homework for this course by using **"cd"** and **"ls"** commands.

(~/training_materials/dev1/exercises).



Lab-Chapter-03-Access HDFS with Command-Line and Hue

In this chapter, we use Hadoop Distributed File System commands and Hue File Brower web-based to control the files in HDFS.

Explore the HDFS Command Line Interface

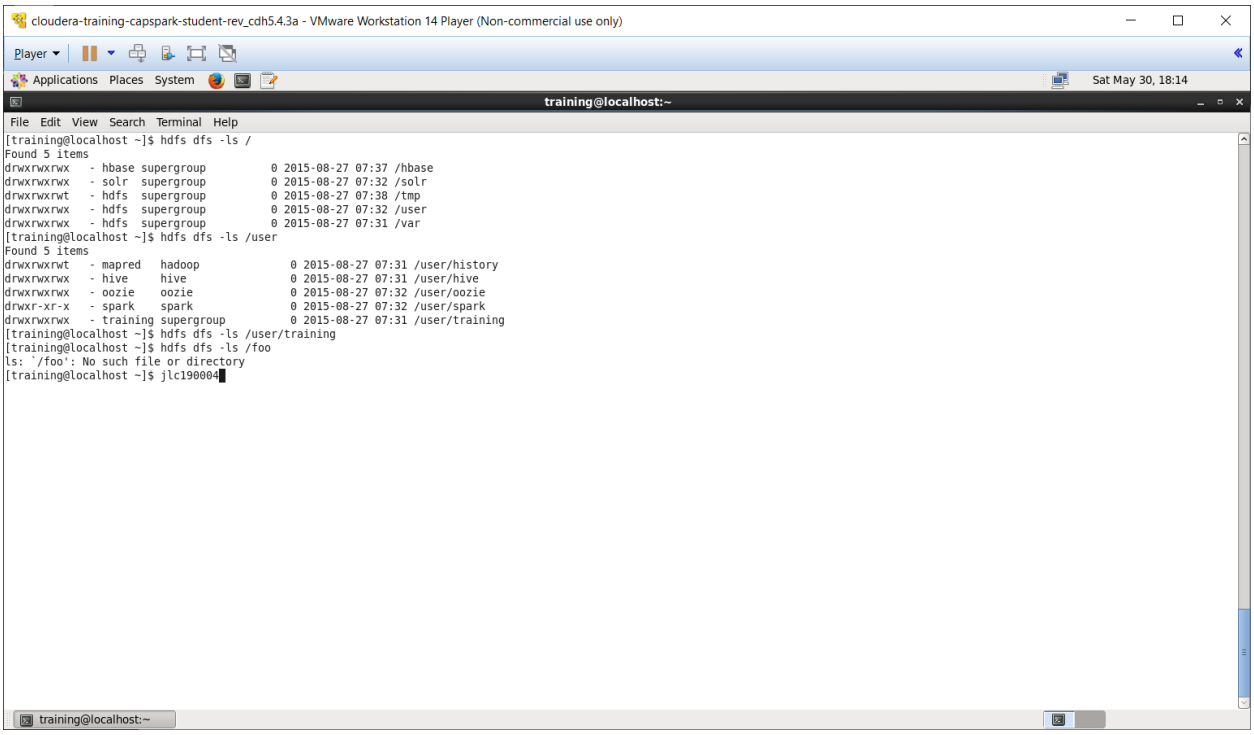
HDFS system: Using the hdfs command to start to interact with HDFS and use the [hdfs dfs] command to execute the file system within HDFS.

[hdfs dfs -ls /]: Shows the contents of the root directory in HDFS, and we found 5 items.

[hdfs dfs -ls /user]: Viewing the contents of the /user directory, and we found 5 items.

[hdfs dfs -ls /user/training]: List the contents of my home directory, there are no files yet, so the command silently exits.

[hdfs dfs -ls /foo]: Refers to a directory that doesn't exist and which would display an error message.



Upload Files to HDFS

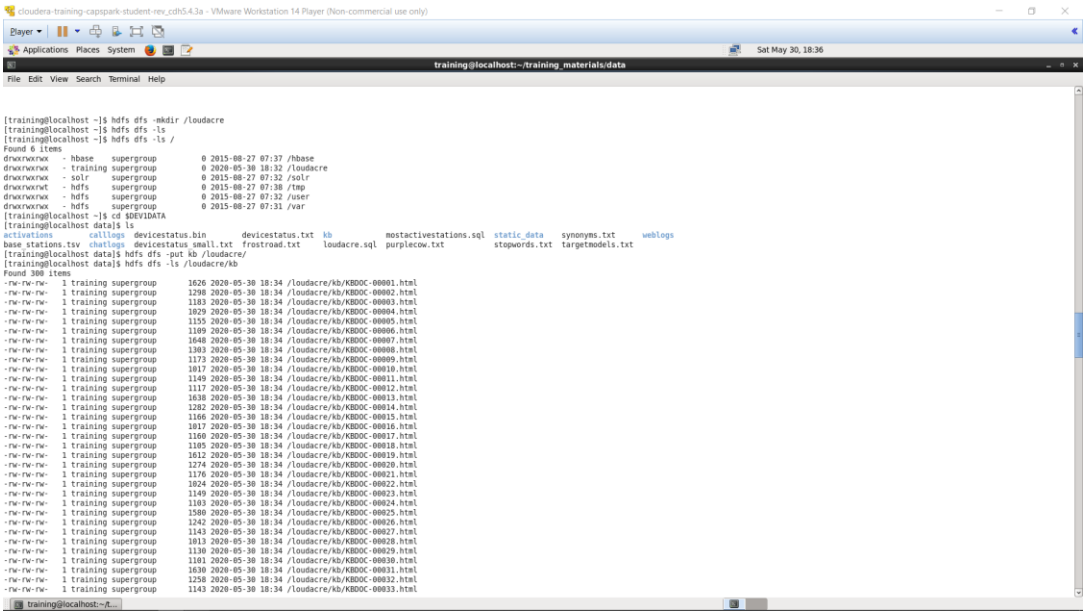
[hdfs dfs -mkdir /loudacre]: Creating a new top-level directory for homework assignments.

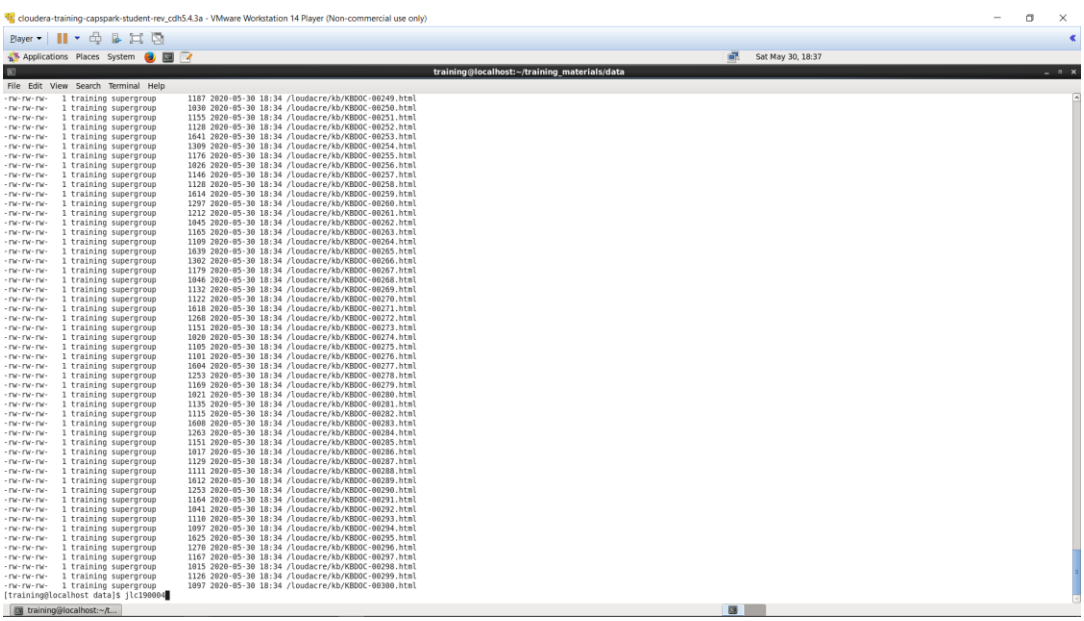
[cd $DEV1DATA]: Change directories to the local filesystem directory containing the sample data.

[ls]: command under DEV1DATA, we will see several files and directories.

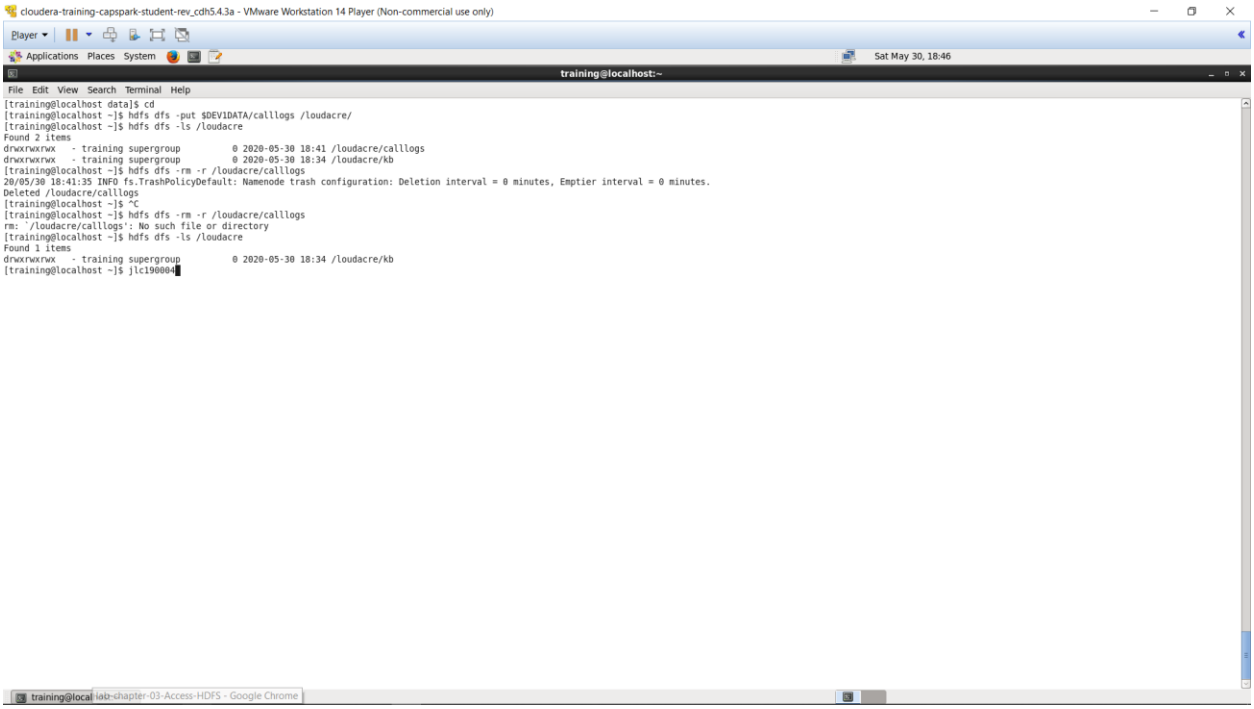
[hdfs dfs -put kb /loudacre/]: This copies the local kb directory and its contents into a remote HDFS directory named /loudacre/kb.

[hdfs dfs -ls /loudacre/kb]: List the contents of the new HDFS directory, check KB articles and found 300 items.



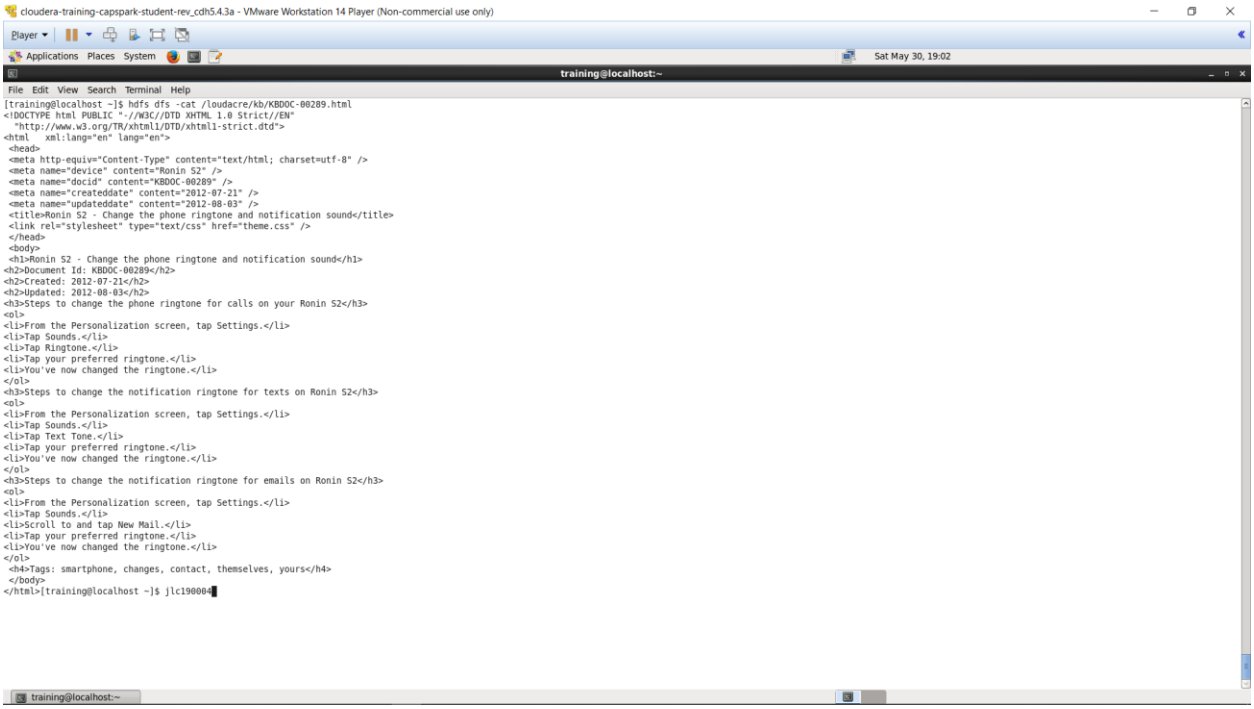


[$ hdfs dfs -put $DEV1DATA/calllogs /loudacre/], [$ hdfs dfs -rm -r /loudacre/calllogs]: Practice uploading a directory, then remove it.



View HDFS files

[$ hdfs dfs -cat /loudacre/kb/KBDOC-00289.html]: All contents in file.

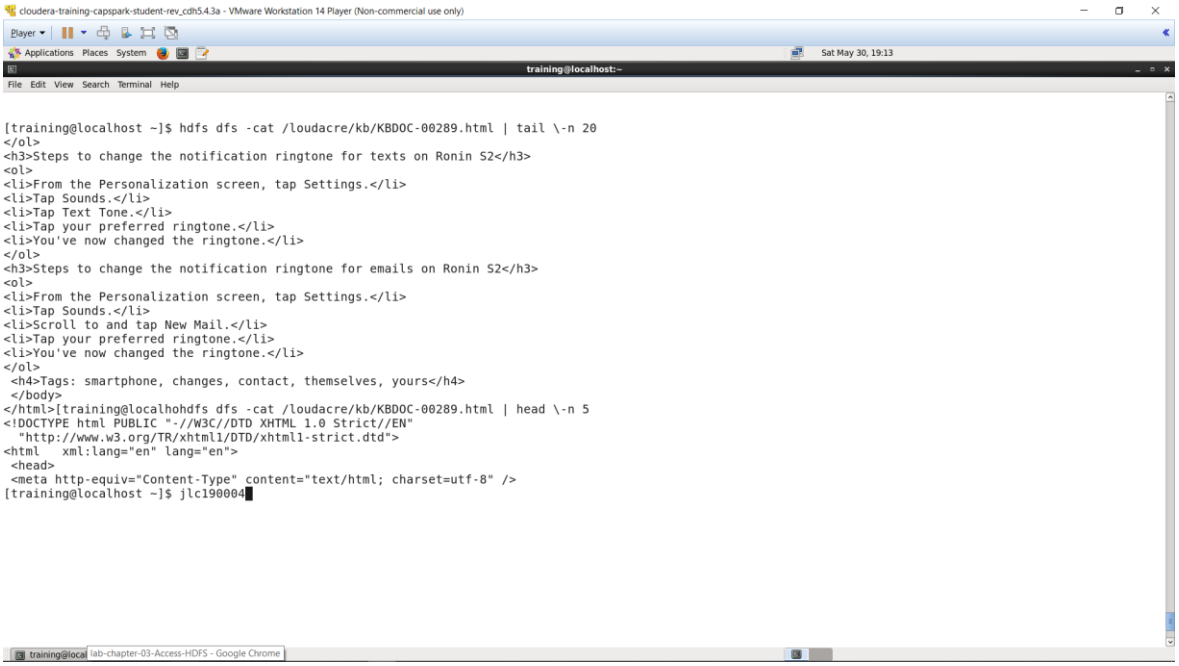


[$ hdfs dfs -cat /loudacre/kb/KBDOC-00289.html | tail \ -n 20]

Show the last 20 lines of the article to your terminal.

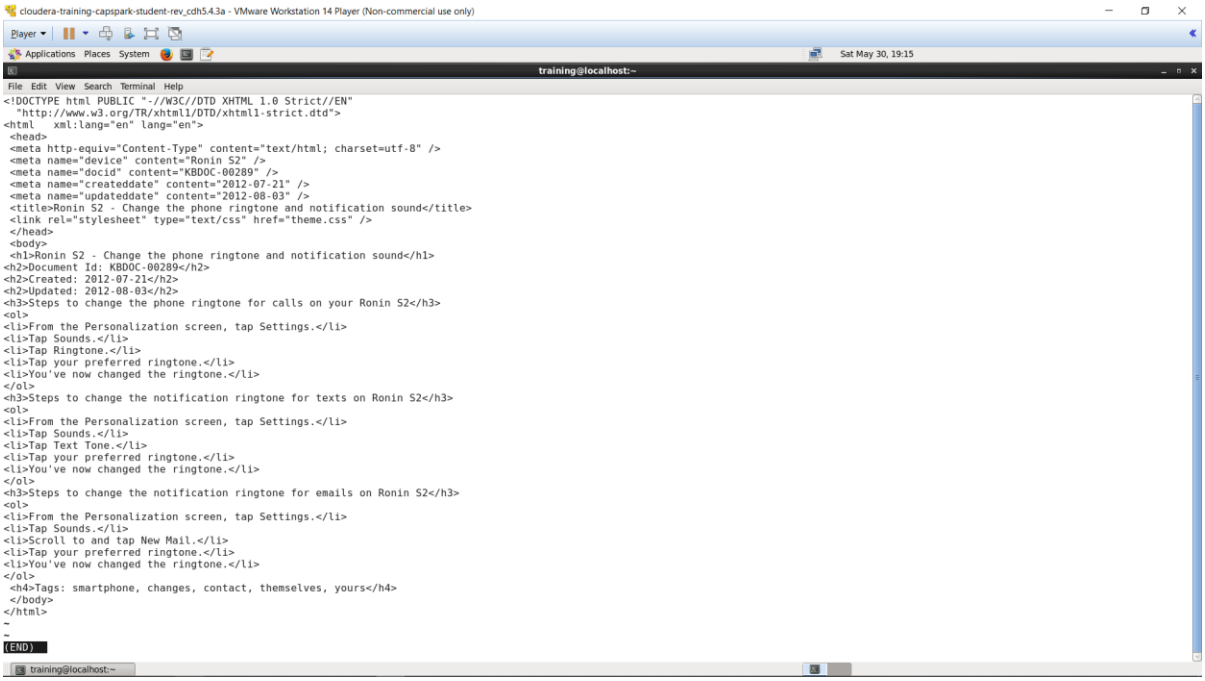
[$ hdfs dfs -cat /loudacre/kb/KBDOC-00289.html | head \ -n 5]

Show the first 5 lines of the article to your terminal.

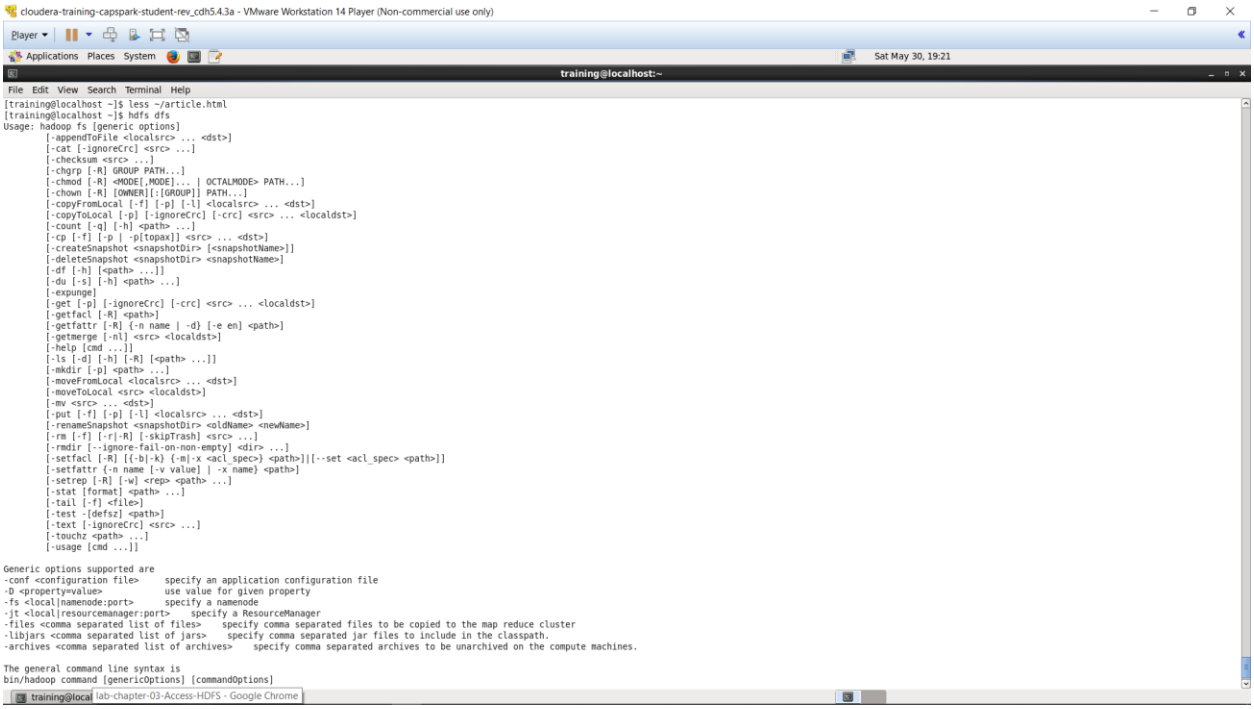


Use the command to retrieve a local copy of a file or directory from HDFS.

[$ hdfs dfs -get \ /loudacre/kb/KBDOC-00289.html ~/article.html], [$ less ~/article.html]

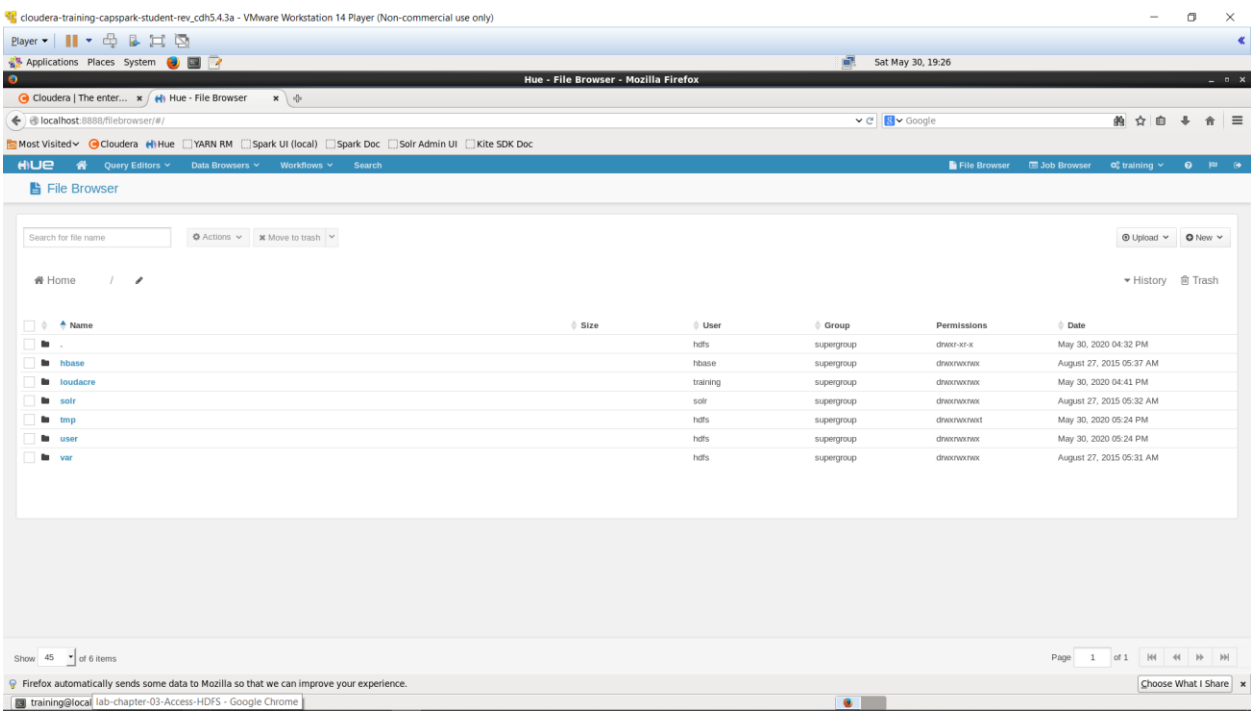


[$ hdfs dfs]: Find a help message describing all the file system commands.

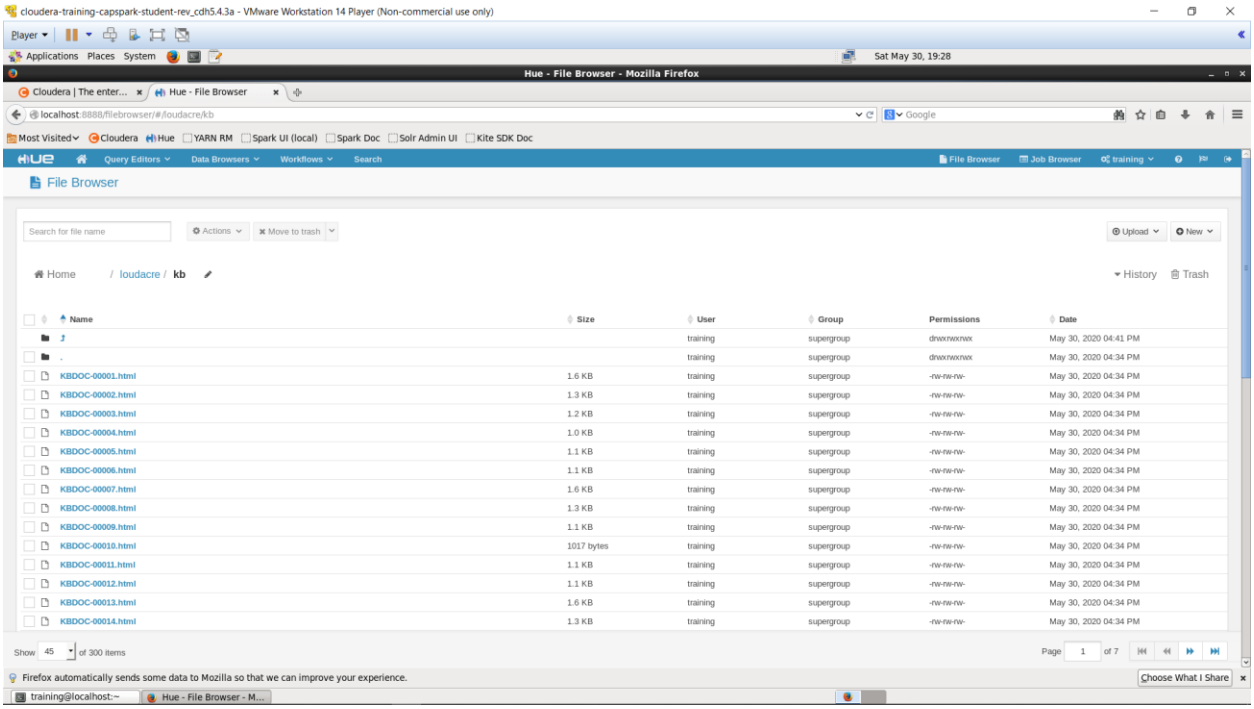


Use the Hue File Browser to browse, view and manage files

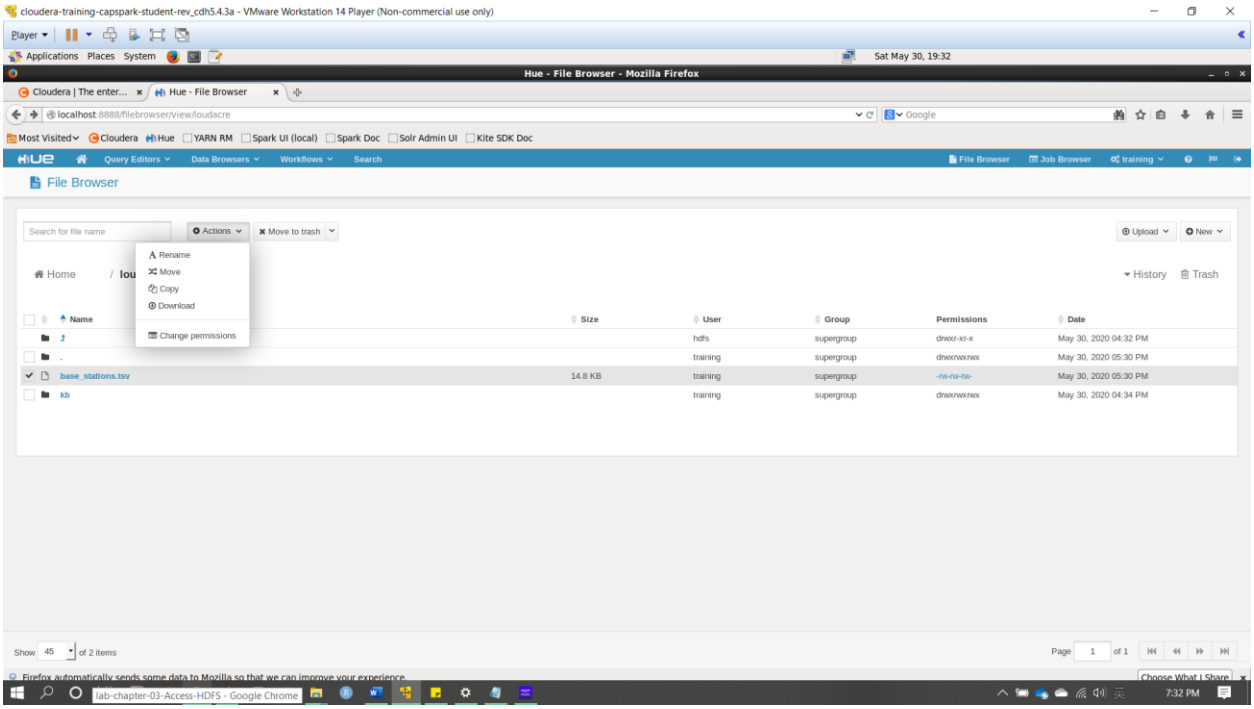
Using Firefox on VM to the Hue File Brower, and use *"training"* to create the account. And then click File Browser to HDFS home directory *(/user/training).* In the directory pathname, click the leading slash (/) to view the HDFS root directory.



Click the "*loudacre*" and click on the name of the "*kb*" directory to see the knowledge base articles.



Return to the /loudacre base directory and upload file(*base_stations.tsv*), and can use Actions button to see a list of actions that can be performed on the selected file(s).



Lab-Chapter-03-Run a YARN Job

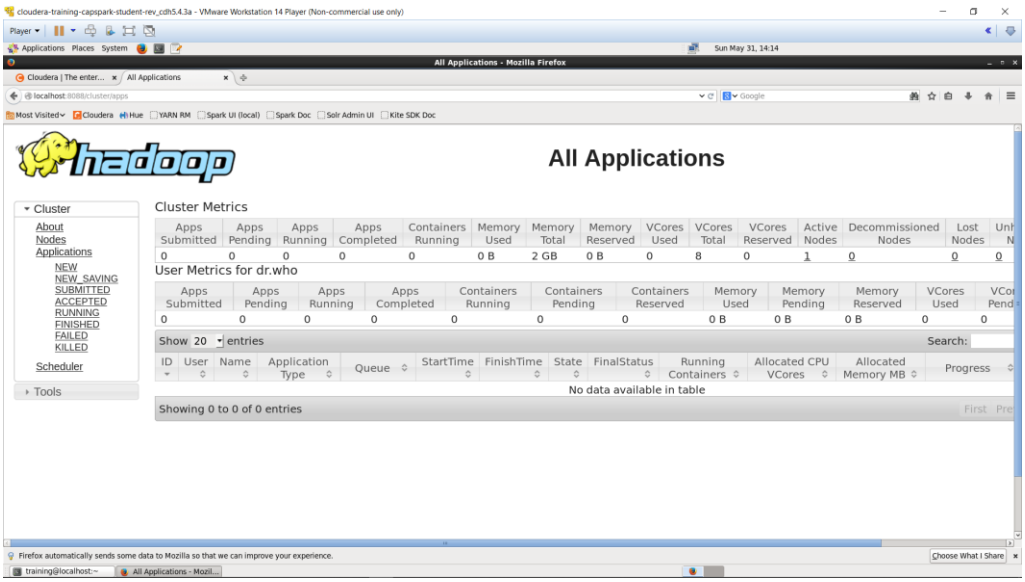
In this assignment, we will present an application to the YARN cluster, and screen the application utilizing both the Hue Job Browser and the YARN Web UI.

Explore the YARN cluster

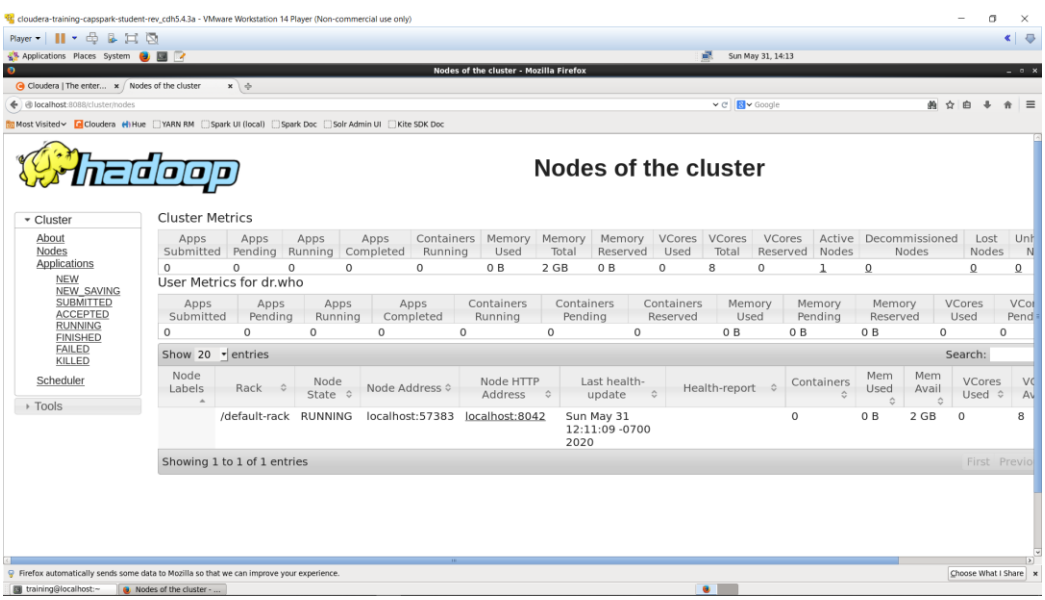
Visit the YARN Resource Manager (RM) UI in Firefox utilizing the gave bookmark, or by going to URL http://localhost:8088

No works are at present running, so the current view shows the group "at rest."

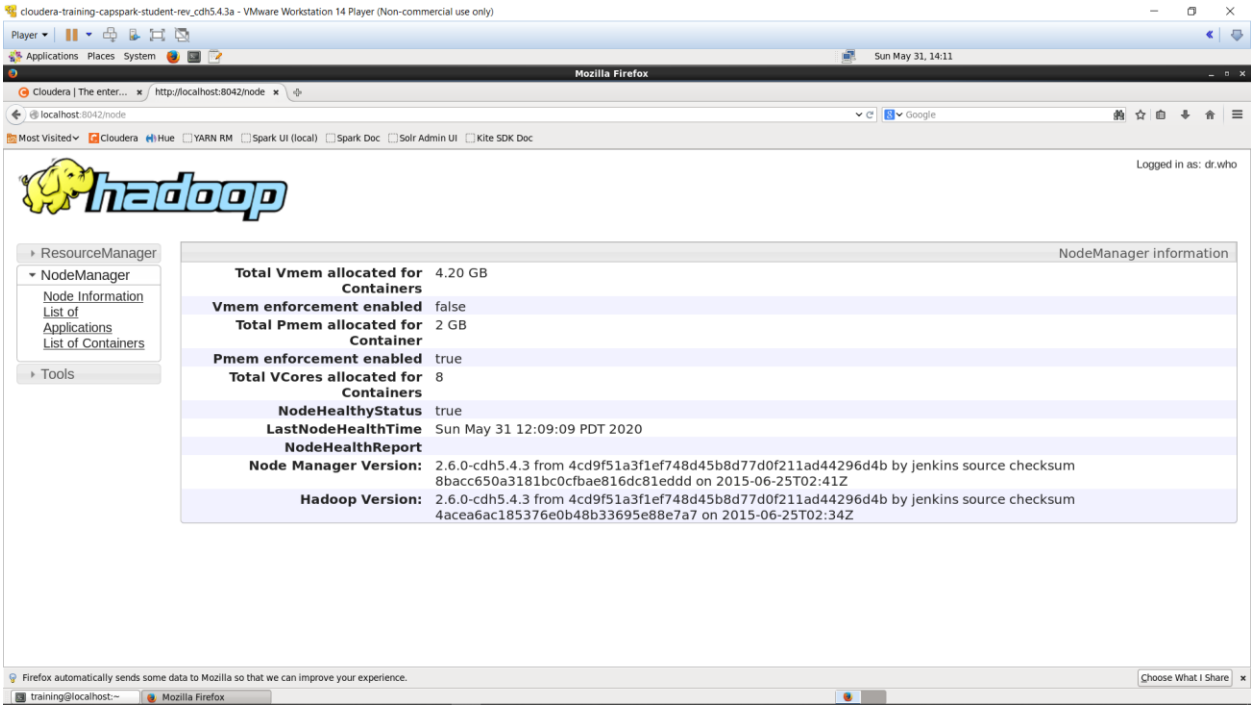
The Cluster Metrics section



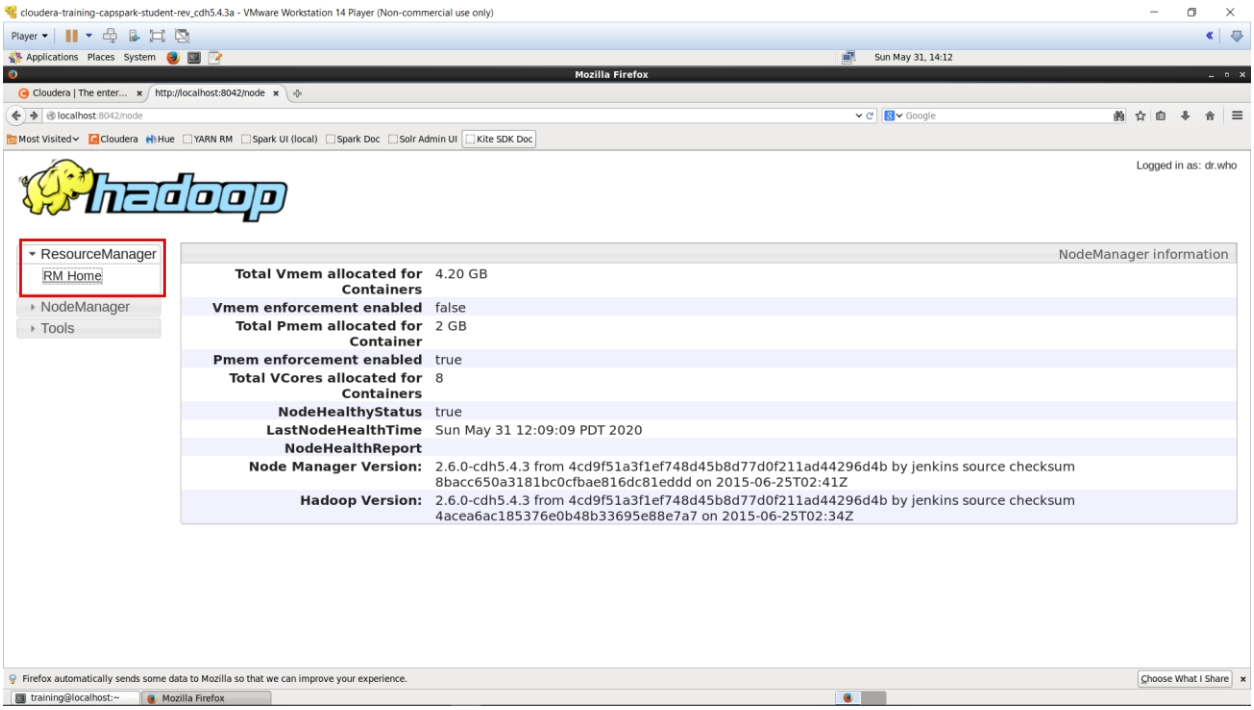
This section will display a list of worker nodes in the cluster.



Click the Node HTTP Address to open the Node Manager UI and will see the information about the chose Node.



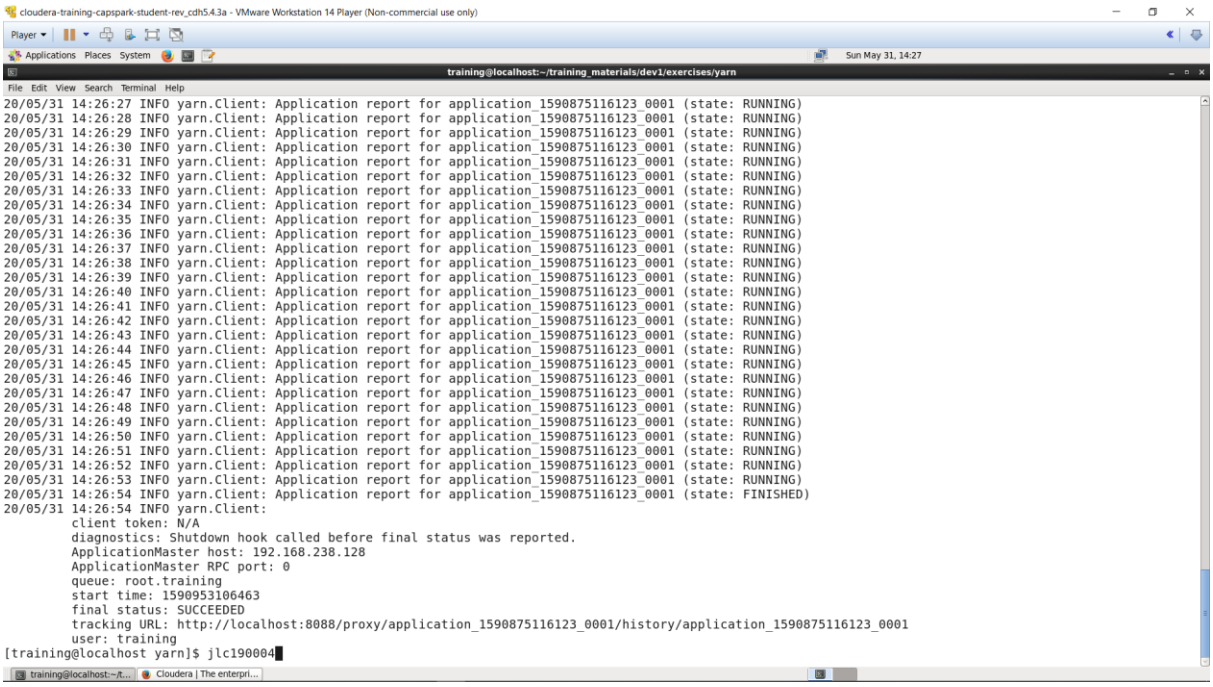
Return to the Resource Manage



Submit an application to the YARN cluster

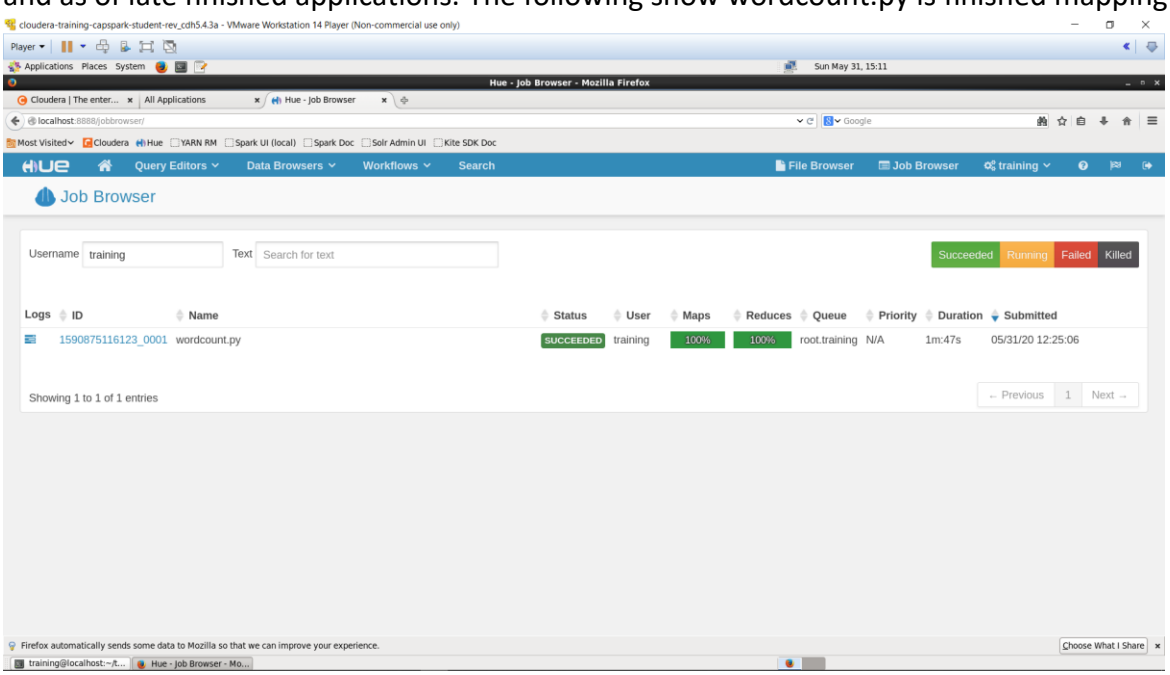
Start the YARN cluster in terminal: [$ cd $DEV1/exercises/yarn]

Run the example wordcount.py program on the YARN cluster to count the frequency of words in the knowledge dataset: [$ spark-submit --master yarn-cluster wordcount.py /loudacre/kb/*]

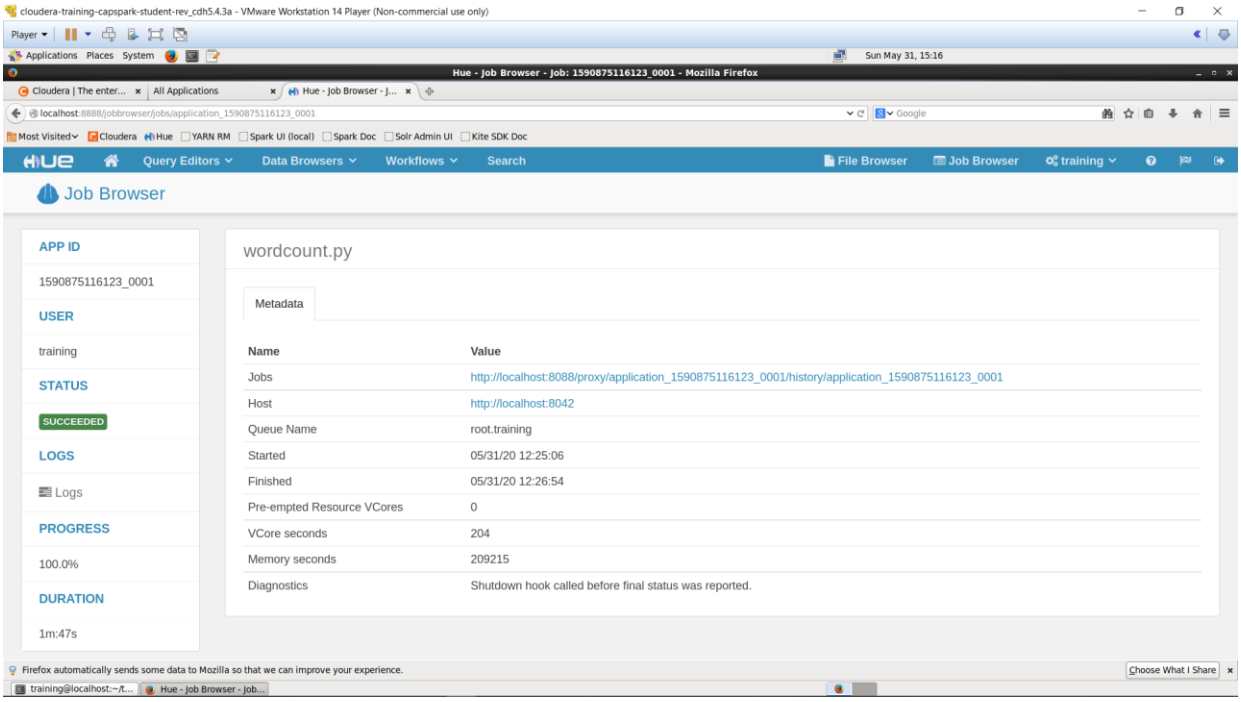


View the application in the Hue Job Browser

Go to Hue, and select the Job Browser, the Job Browser shows a rundown of right now running and as of late finished applications. The following show wordcount.py is finished mapping.

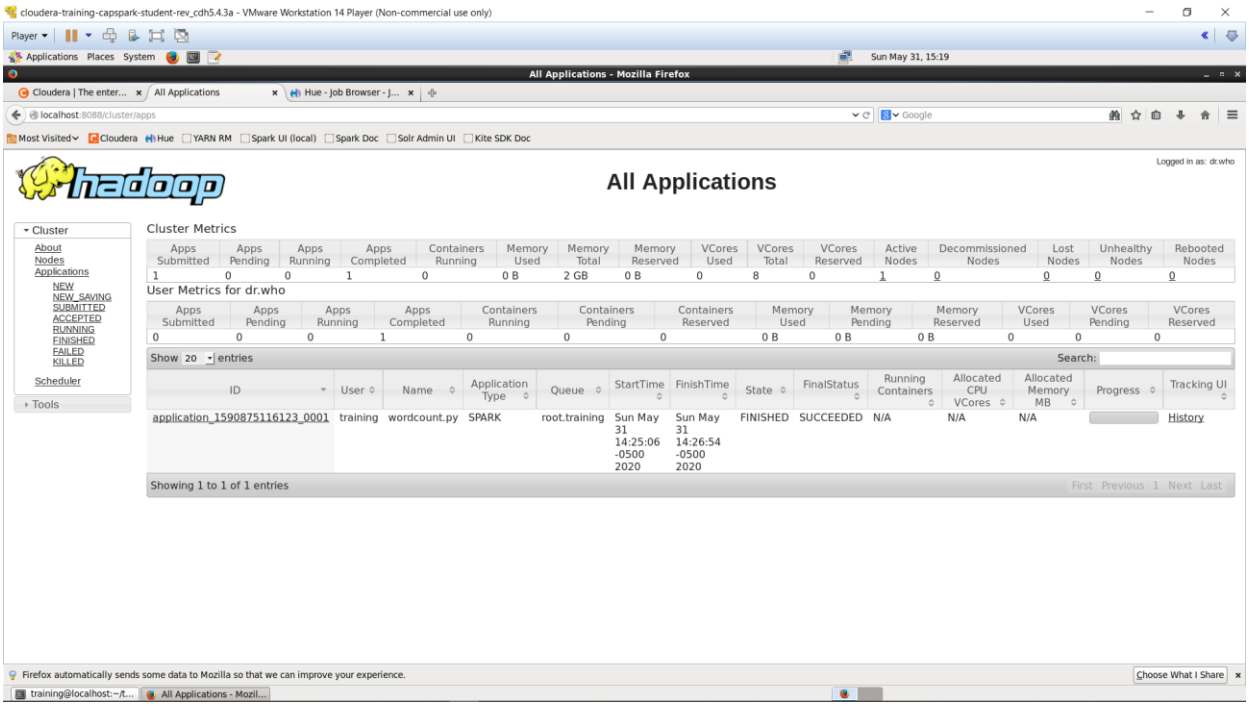


This page permits to tap the application ID to see the information of the running application

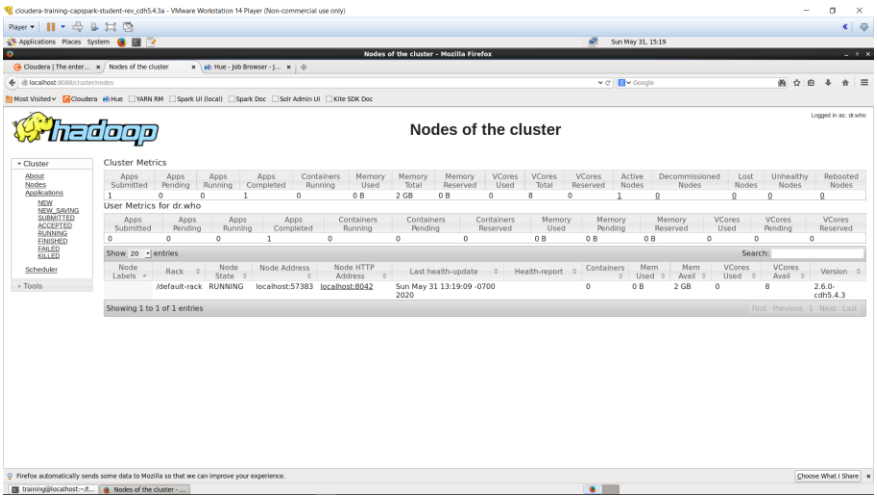


View the application in the YARN UI

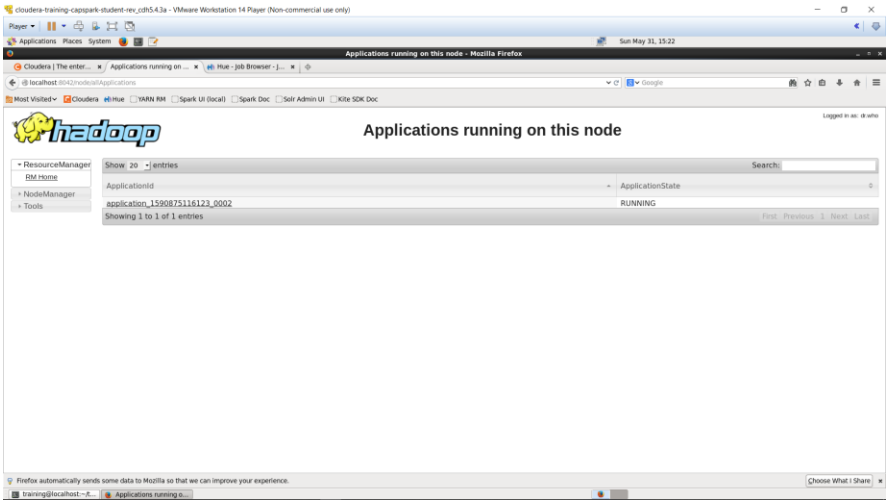
Reload the YARN RM page in Firefox. You will presently observe the application you just began in the base segment of the RM landing page



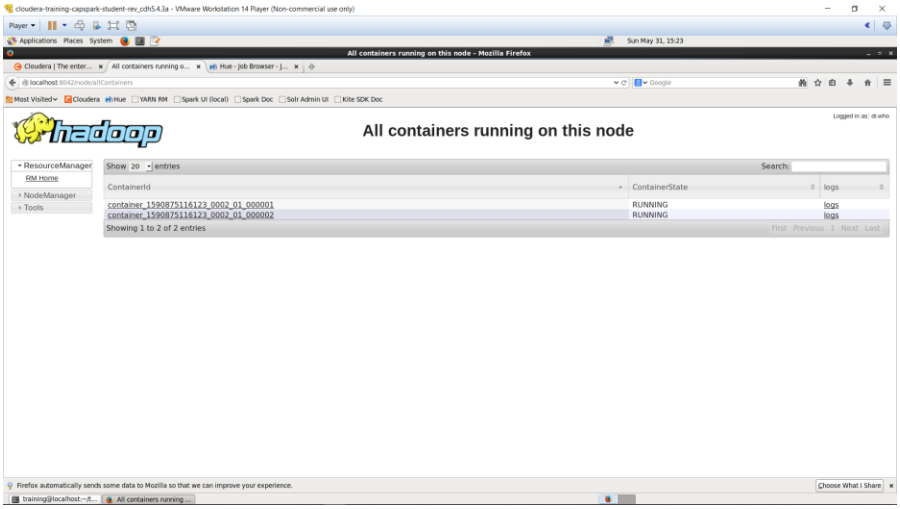
12. Select Nodes and select the Node HTTP Address to open the Node Manager UI.



If the application is running, we can check the List of Applications to see the application you submitted.



If your application is still running, try clicking on the List of Containers.



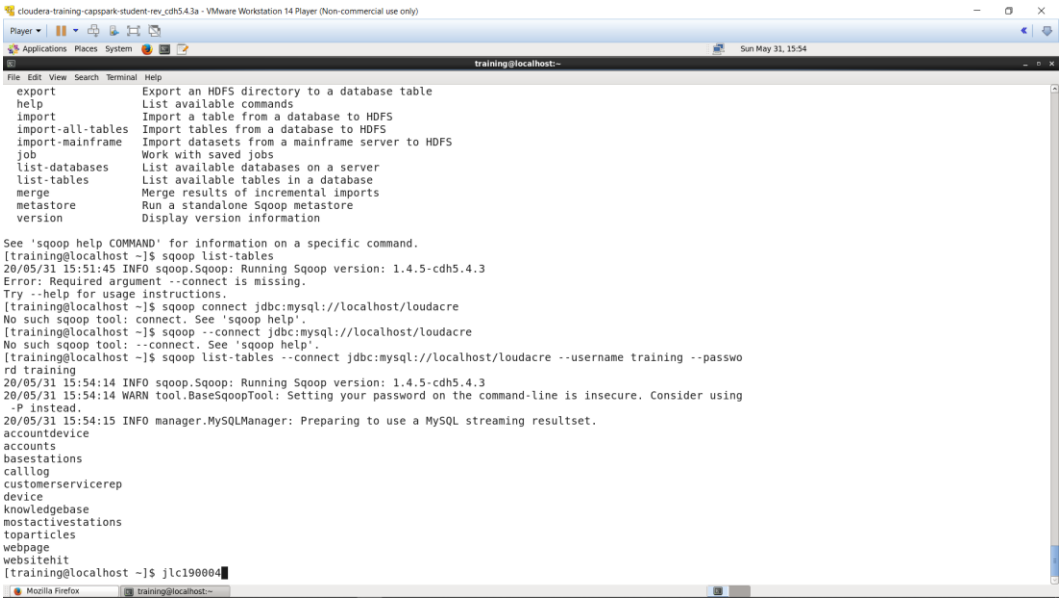
Lab-Chapter-04-Import Data from MySQL Using Sqoop

Import the accounts table from MySQL

Using Sqoop to look at the table layout in MySQL and to import the table from MySQL to HDFS.

[$ sqoop help]: Show the commands we can use under sqoop

[$sqoop list-tables connect jdbc:mysql://localhost/loudacre username training --password training]: List the tables in the loudacre database.

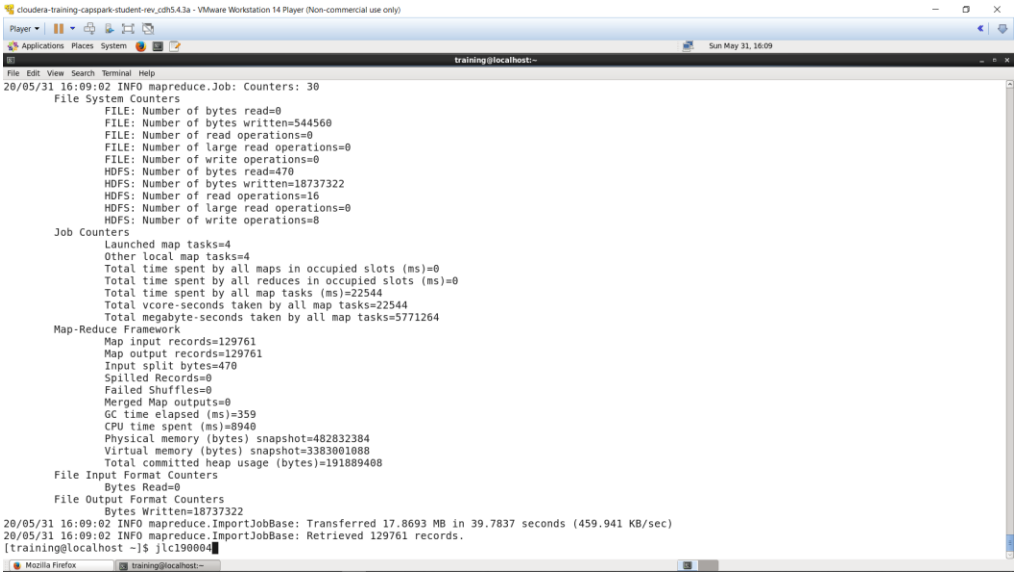


```
  export               Export an HDFS directory to a database table
  help                 List available commands
  import               Import a table from a database to HDFS
  import-all-tables    Import tables from a database to HDFS
  import-mainframe     Import datasets from a mainframe server to HDFS
  job                  Work with saved jobs
  list-databases       List available databases on a server
  list-tables          List available tables in a database
  merge                Merge results of incremental imports
  metastore            Run a standalone Sqoop metastore
  version              Display version information

See 'sqoop help COMMAND' for information on a specific command.
[training@localhost ~]$ sqoop list-tables
20/05/31 15:51:45 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
Error: Required argument --connect is missing.
Try --help for usage instructions.
[training@localhost ~]$ sqoop connect jdbc:mysql://localhost/loudacre
No such sqoop tool: connect. See 'sqoop help'.
[training@localhost ~]$ sqoop --connect jdbc:mysql://localhost/loudacre
No such sqoop tool: --connect. See 'sqoop help'.
[training@localhost ~]$ sqoop list-tables --connect jdbc:mysql://localhost/loudacre --username training --passwo
rd training
20/05/31 15:54:14 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
20/05/31 15:54:14 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using
 -P instead.
20/05/31 15:54:15 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
accountdevice
accounts
basestations
calllog
customerservicerep
device
knowledgebase
mostactivestations
toparticles
webpage
websitehit
[training@localhost ~]$ jlc190004
```

[$ sqoop import --help]: Show the commands we can use under sqoop import.

[sqoop import --connect jdbc:mysql://localhost/loudacre --username training --password training --table accounts --target-dir /loudacre/accounts --null-non-string '\\N']: Import the *accounts* table in the loudacre database and save it in HDFS under /loudacre.



```
20/05/31 16:09:02 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=544560
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=470
                HDFS: Number of bytes written=18737322
                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)=22544
                Total vcore-seconds taken by all map tasks=22544
                Total megabyte-seconds taken by all map tasks=5771264
        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)=359
                CPU time spent (ms)=8940
                Physical memory (bytes) snapshot=482832384
                Virtual memory (bytes) snapshot=3383001088
                Total committed heap usage (bytes)=191889408
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=18737322
20/05/31 16:09:02 INFO mapreduce.ImportJobBase: Transferred 17.8693 MB in 39.7837 seconds (459.941 KB/sec)
20/05/31 16:09:02 INFO mapreduce.ImportJobBase: Retrieved 129761 records.
[training@localhost ~]$ jlc190004
```

View the imported data

[$ hdfs dfs -ls /loudacre/accounts]: List the contents of the *accounts* directory.

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System                                                    Sun May 31, 16:11
training@localhost:~
File Edit View Search Terminal Help
                FILE: Number of write operations=0
                HDFS: Number of bytes read=470
                HDFS: Number of bytes written=18737322
                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)=22544
                Total vcore-seconds taken by all map tasks=22544
                Total megabyte-seconds taken by all map tasks=5771264
        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)=359
                CPU time spent (ms)=8940
                Physical memory (bytes) snapshot=482832384
                Virtual memory (bytes) snapshot=3383001088
                Total committed heap usage (bytes)=191889408
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=18737322
20/05/31 16:09:02 INFO mapreduce.ImportJobBase: Transferred 17.8693 MB in 39.7837 seconds (459.941 KB/sec)
20/05/31 16:09:02 INFO mapreduce.ImportJobBase: Retrieved 129761 records.
[training@localhost ~]$ hdfs dfs -ls /loudacre /accounts
Found 3 items
drwxrwxrwx   - training supergroup          0 2020-05-31 16:09 /loudacre/accounts
-rw-rw-rw-   1 training supergroup      15191 2020-05-30 19:30 /loudacre/base_stations.tsv
drwxrwxrwx   - training supergroup          0 2020-05-30 18:34 /loudacre/kb
ls: `/accounts': No such file or directory
[training@localhost ~]$ jlc190004
Mozilla Firefox   training@localhost:~
```

Use HDFS tail command to view the last part of the file for each of the MapReduce partition files, from part-m-00000 ~part-m-00003.

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System                                                    Sun May 31, 16:13
training@localhost:~
File Edit View Search Terminal Help
[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 McVaney Road,Reno,NV,89535,7750546952,2014-03-18 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,90042,2132809416,2014-03-18 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,93490,8054121301,2014-03-18 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,94060,4152327199,2014-03-18 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,90723,5628246641,2014-03-18 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,90805,5624213548,2014-03-18 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,Kingman,AZ,86486,9286393819,2014-03-18 13:31:52.0,2014-03-18 13:3
1:52.0
64877,2012-04-24 05:52:44.0,\N,Jermaine,Harmon,4489 Shearwood Forest Drive,Burbank,CA,91503,8186973079,2014-03-18 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,94676,5108035413,2014-03-18 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,92087,6196073804,2014-03-18 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,94183,4152020130,2014-03-18 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,95762,9165704077,2014-03-18 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,86482,9285524623,2014-03-18 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,97440,5411715739,2014-03-18 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,92388,9097978236,2014-03-18 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,95000,4088662329,2014-03-18 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,90609,5624227213,2014-03-18 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,94635,5100736336,2014-03-18 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 ~]$ jlc190004
Mozilla Firefox   training@localhost:~
```

Import incremental updates to accounts

Use Sqoop to append these new records, incrementally import and append the newly added accounts to the accounts directory and list the contents of the *accounts* directory to verify the Sqoop import.

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System                                                    Sun May 31, 16:21
training@localhost:~
File Edit View Search Terminal Help
129756,2014-01-26 19:34:06.0,\N,Danny,Graf,1104 Golden Ridge Road,San Jose,CA,95000,4088662329,2014-03-18 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,90609,5624227213,2014-03-18 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,94635,5100736336,2014-03-18 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 ~]$ $DEV1/exercises/sqoop/add_new_accounts.py
Adding latest account data at 2020-05-31 16:15:10
Finished Adding Accounts
[training@localhost ~]$ 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 <largest_acct_num>
bash: syntax error near unexpected token `newline'
[training@localhost ~]$ 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 <largest_acct_num>
bash: syntax error near unexpected token `newline'
[training@localhost ~]$ 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 <largest_acct_num>
bash: syntax error near unexpected token `newline'
[training@localhost ~]$ $DEV1/exercises/sqoop/add_new_accounts.py
Adding latest account data at 2020-05-31 16:19:25
Finished Adding Accounts
[training@localhost ~]$ 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 <largest_acct_num>
bash: syntax error near unexpected token `newline'
[training@localhost ~]$ hdfs dfs -ls /loudacre/accounts
Found 5 items
-rw-rw-rw-   1 training supergroup          0 2020-05-31 16:09 /loudacre/accounts/_SUCCESS
-rw-rw-rw-   1 training supergroup    4706617 2020-05-31 16:08 /loudacre/accounts/part-m-00000
-rw-rw-rw-   1 training supergroup    4693530 2020-05-31 16:08 /loudacre/accounts/part-m-00001
-rw-rw-rw-   1 training supergroup    4674529 2020-05-31 16:08 /loudacre/accounts/part-m-00002
-rw-rw-rw-   1 training supergroup    4662646 2020-05-31 16:09 /loudacre/accounts/part-m-00003
[training@localhost ~]$ jlc190004
Mozilla Firefox   training@localhost:~
```

Use Hadoop's *cat* command to see all the contents of files.

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System                                                    Sun May 31, 16:22
training@localhost:~
File Edit View Search Terminal Help
32404,2012-07-11 02:11:21.0,\N,Tami,Ketcham,2105 Alfred Drive,Palo Alto,CA,94304,6506245243,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32405,2012-01-30 21:21:21.0,\N,Suzanne,Alexander,3326 Daylene Drive,Riverside,CA,92565,9511066991,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32406,2012-04-21 03:20:16.0,\N,Anita,Byrd,1034 Hillside Street,Torrance,CA,90553,4241475405,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32407,2012-06-17 13:13:17.0,2014-01-25 20:20:00.0,Kenneth,Khalil,1878 Sand Fork Road,Bend,OR,97761,5410280348,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32408,2012-04-28 10:01:32.0,2014-01-23 18:29:20.0,Jennifer,Merida,251 Walnut Hill Drive,San Francisco,CA,94041,4150531415,2014-03-18 13:30:46.0,2014-03-18 13:
30:46.0
32409,2012-05-30 01:58:17.0,2013-11-18 07:28:52.0,Kevin,Gibson,3637 Smith Road,Oxnard,CA,93006,8054174381,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32410,2012-07-15 01:07:38.0,\N,Sheila,Morgan,2652 Columbia Mine Road,Eugene,OR,97404,5417364093,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32411,2012-10-02 04:32:42.0,\N,Kenny,Conger,10 Hart Street,Santa Rosa,CA,95447,7071518273,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32412,2012-07-24 10:51:13.0,\N,Angel,Scoggins,454 Powder House Road,Carson City,NV,89787,7755123461,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32413,2012-03-19 16:50:13.0,2013-11-09 15:53:02.0,Ida,Castro,3387 Carson Street,Pendleton,OR,97814,5418307107,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32414,2012-05-30 00:19:37.0,2013-11-03 06:04:17.0,Tammy,Buxton,4073 Jefferson Street,San Jose,CA,95090,4086231163,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32415,2012-09-20 09:32:21.0,2014-01-12 03:39:49.0,James,Ward,2746 Essex Court,Burbank,CA,91568,8185549957,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32416,2012-10-20 16:51:46.0,\N,Emma,Carmichael,981 Michael Street,Kingman,AZ,86459,9286901080,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32417,2012-01-13 21:45:25.0,\N,Carmen,Lane,3617 Glendale Avenue,San Francisco,CA,94082,4150050343,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32418,2012-10-19 09:07:29.0,\N,Andrea,Holt,290 Star Trek Drive,Ely,NV,89347,7754947784,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32419,2012-07-12 20:54:02.0,2014-02-26 08:42:34.0,Catherine,Roberts,1252 Oakmound Road,Salem,OR,97345,5034265086,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32420,2012-08-17 12:14:35.0,\N,Erin,Little,2247 Marie Street,Riverside,CA,92532,9518136429,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32421,2012-11-10 13:38:41.0,\N,Johnny,Graves,4447 Farland Avenue,Berkeley,CA,94768,5108561558,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32422,2012-07-14 14:33:30.0,2014-02-10 19:16:36.0,Leonard,Lamm,900 Clover Drive,Sacramento,CA,95619,9169092101,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32423,2012-03-19 20:12:43.0,\N,Dennis,Hodge,1584 Lost Creek Road,Portland,OR,97145,5035626144,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32424,2012-05-24 07:12:59.0,2013-12-17 22:57:57.0,Nadine,Frias,1526 Sardis Station,Prescott,AZ,86340,9282869702,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32425,2012-08-15 22:49:15.0,\N,Juan,Mendel,2573 McDowell Street,Riverside,CA,92515,9514481522,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32426,2012-04-10 05:59:47.0,\N,Dorothy,Thompson,4717 Jerome Avenue,San Bernardino,CA,92412,9095440672,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32427,2012-01-22 22:05:47.0,2014-02-10 20:51:28.0,Tara,Jenkins,3803 Sunrise Road,Pendleton,OR,97809,5414274958,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32428,2012-11-29 18:42:37.0,\N,Philip,Gamble,2306 Simpson Street,San Francisco,CA,94158,4152626049,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32429,2012-03-03 23:47:39.0,\N,Marvin,Potts,4086 Concord Street,Fresno,CA,93620,5592544766,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32430,2012-11-25 08:37:15.0,\N,Stephanie,Wootton,2634 Mulberry Lane,Klamath Falls,OR,97647,5418698398,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32431,2012-04-05 19:16:54.0,2013-11-13 09:18:32.0,Judith,Lindgren,1890 Tanglewood Road,Anaheim,CA,92847,7141354975,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32432,2012-04-08 09:52:52.0,2014-03-09 13:01:18.0,Darlene,Gonzales,2777 Highland Drive,Mojave,CA,93560,6619388480,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32433,2012-10-19 19:47:32.0,2013-06-09 16:03:44.0,Nicholas,Monger,3172 Old Dear Lane,Portland,OR,97045,5034669567,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32434,2012-01-08 12:31:49.0,2013-12-25 04:55:12.0,Anne,Boyle,4215 McVaney Road,Reno,NV,89535,7750546952,2014-03-18 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,90042,2132809416,2014-03-18 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,93490,8054121301,2014-03-18 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,94060,4152327199,2014-03-18 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,90723,5628246641,2014-03-18 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 ~]$ jlc190004
Mozilla Firefox   training@localhost:~
```

Import webpage data using an alternate field delimiter

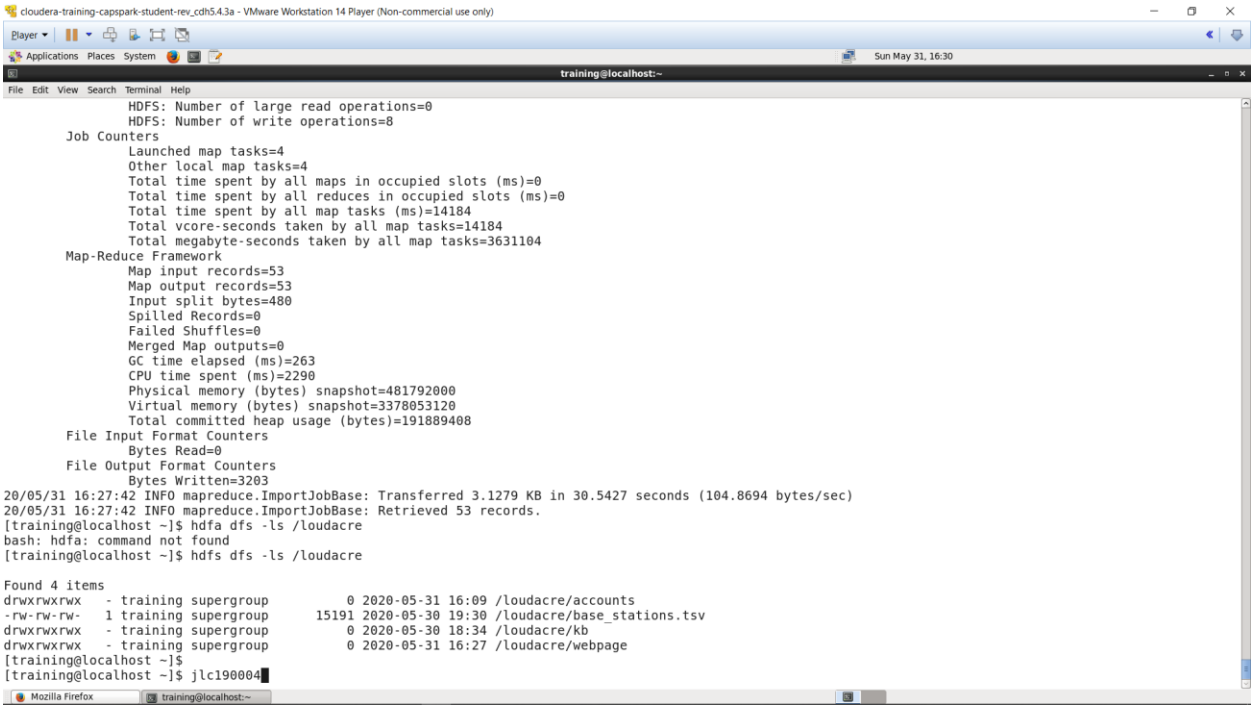
We can use *sqoop eval* command to import the webpage table to HDFS, but use the tab character (\t) instead of the default (comma) as the field terminator.

```
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 ~]$ clear

[training@localhost ~]$ sqoop eval --query "SELECT * FROM webpage LIMIT 10" --connect jdbc:mysql://localhost/loudacre --username training --password training
20/05/31 16:26:06 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
20/05/31 16:26:06 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/05/31 16:26:06 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 ~]$ sqoop import --connect jdbc:mysql://localhost/loudacre --username training --password training --table webpage --target-dir /loudacre/
webpage --fields-terminated-by "\t"
20/05/31 16:27:08 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
20/05/31 16:27:08 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/05/31 16:27:08 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
20/05/31 16:27:08 INFO tool.CodeGenTool: Beginning code generation
20/05/31 16:27:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `webpage` AS t LIMIT 1
20/05/31 16:27:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `webpage` AS t LIMIT 1
20/05/31 16:27:08 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/6465c802cf5852de91f948730d6afd0e/webpage.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/05/31 16:27:10 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/6465c802cf5852de91f948730d6afd0e/webpage.jar
20/05/31 16:27:10 WARN manager.MySQLManager: It looks like you are importing from mysql.
20/05/31 16:27:10 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
20/05/31 16:27:10 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
20/05/31 16:27:10 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
20/05/31 16:27:10 INFO mapreduce.ImportJobBase: Beginning import of webpage
20/05/31 16:27:10 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
```

```
20/05/31 16:27:42 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=543672
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                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)=14184
                Total vcore-seconds taken by all map tasks=14184
                Total megabyte-seconds taken by all map tasks=3631104
        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)=263
                CPU time spent (ms)=2290
                Physical memory (bytes) snapshot=481792000
                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
20/05/31 16:27:42 INFO mapreduce.ImportJobBase: Transferred 3.1279 KB in 30.5427 seconds (104.8694 bytes/sec)
20/05/31 16:27:42 INFO mapreduce.ImportJobBase: Retrieved 53 records.
[training@localhost ~]$ jlc190004
```

Review the data files imported to the */loudacre/webpage* directory.



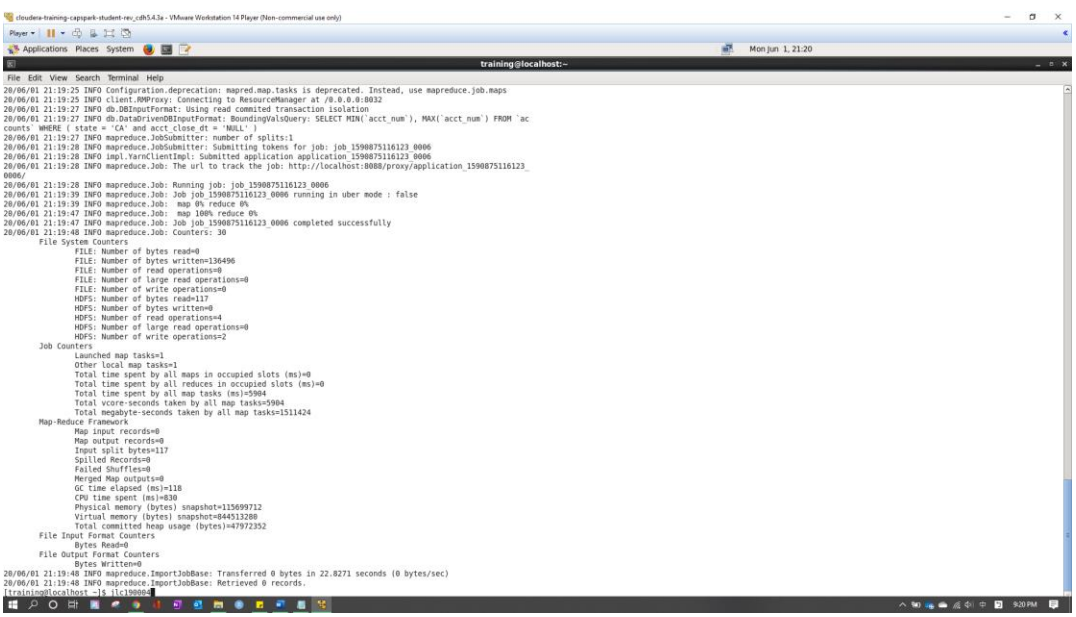
```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System                                   Sun May 31, 16:30
training@localhost:~
File Edit View Search Terminal Help
                HDFS: Number of large read operations=0
                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)=14184
                Total vcore-seconds taken by all map tasks=14184
                Total megabyte-seconds taken by all map tasks=3631104
        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)=263
                CPU time spent (ms)=2290
                Physical memory (bytes) snapshot=481792000
                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
20/05/31 16:27:42 INFO mapreduce.ImportJobBase: Transferred 3.1279 KB in 30.5427 seconds (104.8694 bytes/sec)
20/05/31 16:27:42 INFO mapreduce.ImportJobBase: Retrieved 53 records.
[training@localhost ~]$ hdfa dfs -ls /loudacre
bash: hdfa: command not found
[training@localhost ~]$ hdfs dfs -ls /loudacre

Found 4 items
drwxrwxrwx   - training supergroup          0 2020-05-31 16:09 /loudacre/accounts
-rw-rw-rw-   1 training supergroup      15191 2020-05-30 19:30 /loudacre/base_stations.tsv
drwxrwxrwx   - training supergroup          0 2020-05-30 18:34 /loudacre/kb
drwxrwxrwx   - training supergroup          0 2020-05-31 16:27 /loudacre/webpage
[training@localhost ~]$
[training@localhost ~]$ jlc190004
Mozilla Firefox   training@localhost:~
```

Optional Homework

Use Sqoop to import only accounts where the person lives in California (state = ”CA”) and has an active account (acct_close_dt IS NULL)

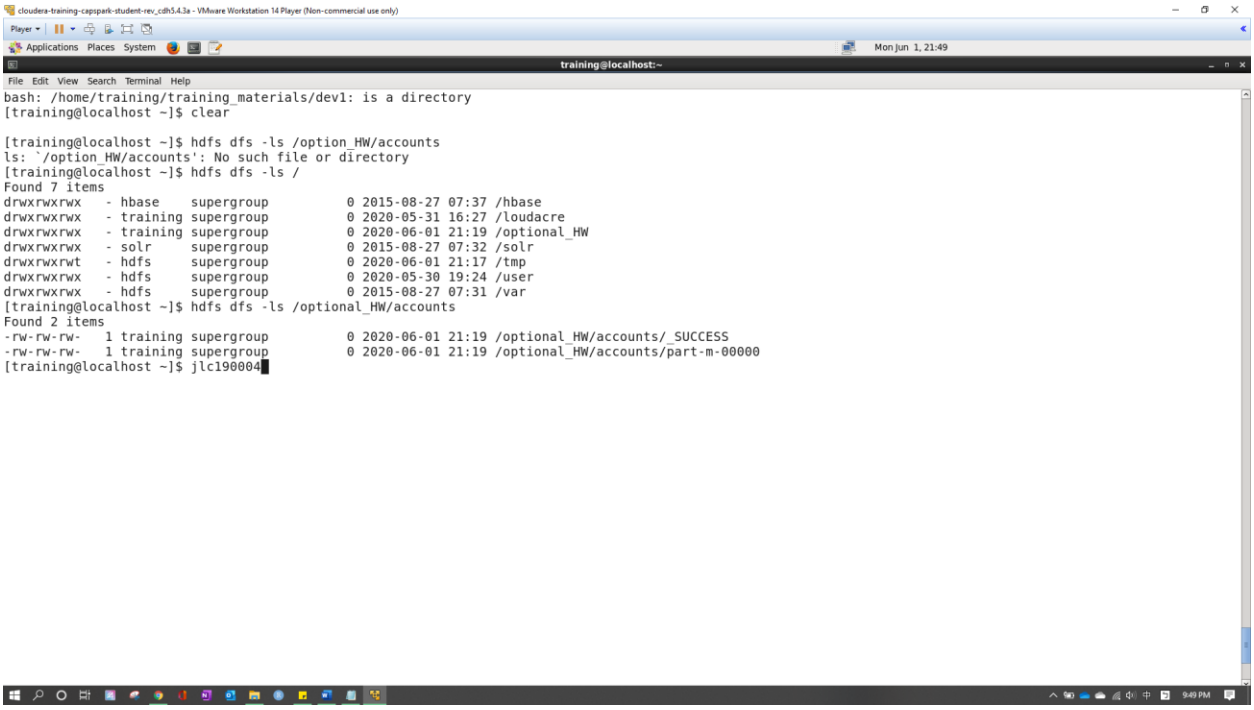
[sqoop import --connect jdbc:mysql://localhost/loudacre --username training --password training --table accounts --where "state = 'CA' and acct_close_dt = 'NULL'" --target-dir /optional_HW/accounts --null-non-string '\\N']



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System                                   Mon Jun 1, 21:20
training@localhost:~
File Edit View Search Terminal Help
20/06/01 21:19:25 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
20/06/01 21:19:25 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/06/01 21:19:27 INFO db.DBInputFormat: Using read commited transaction isolation
20/06/01 21:19:27 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`acct_num`), MAX(`acct_num`) FROM `ac
counts` WHERE ( state = 'CA' and acct_close_dt = 'NULL' )
20/06/01 21:19:27 INFO mapreduce.JobSubmitter: number of splits:1
20/06/01 21:19:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1590875116123_0006
20/06/01 21:19:28 INFO impl.YarnClientImpl: Submitted application application_1590875116123_0006
20/06/01 21:19:28 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1590875116123_
0006/
20/06/01 21:19:28 INFO mapreduce.Job: Running job: job_1590875116123_0006
20/06/01 21:19:39 INFO mapreduce.Job: Job job_1590875116123_0006 running in uber mode : false
20/06/01 21:19:39 INFO mapreduce.Job:  map 0% reduce 0%
20/06/01 21:19:47 INFO mapreduce.Job:  map 100% reduce 0%
20/06/01 21:19:47 INFO mapreduce.Job: Job job_1590875116123_0006 completed successfully
20/06/01 21:19:48 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=136496
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=117
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                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)=5904
                Total vcore-seconds taken by all map tasks=5904
                Total megabyte-seconds taken by all map tasks=1511424
        Map-Reduce Framework
                Map input records=0
                Map output records=0
                Input split bytes=117
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=118
                CPU time spent (ms)=830
                Physical memory (bytes) snapshot=115699712
                Virtual memory (bytes) snapshot=844513280
                Total committed heap usage (bytes)=47972352
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
20/06/01 21:19:48 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 22.8271 seconds (0 bytes/sec)
20/06/01 21:19:48 INFO mapreduce.ImportJobBase: Retrieved 0 records.
[training@localhost ~]$ jlc190004
```

Check the files is whether import correct.

[$ hdfs dfs -ls /optional_HW/accounts]



```
bash: /home/training/training_materials/dev1: is a directory
[training@localhost ~]$ clear

[training@localhost ~]$ hdfs dfs -ls /option_HW/accounts
ls: `/option_HW/accounts': No such file or directory
[training@localhost ~]$ hdfs dfs -ls /
Found 7 items
drwxrwxrwx   - hbase    supergroup          0 2015-08-27 07:37 /hbase
drwxrwxrwx   - training supergroup          0 2020-05-31 16:27 /loudacre
drwxrwxrwx   - training supergroup          0 2020-06-01 21:19 /optional_HW
drwxrwxrwx   - solr     supergroup          0 2015-08-27 07:32 /solr
drwxrwxrwt   - hdfs     supergroup          0 2020-06-01 21:17 /tmp
drwxrwxrwx   - hdfs     supergroup          0 2020-05-30 19:24 /user
drwxrwxrwx   - hdfs     supergroup          0 2015-08-27 07:31 /var
[training@localhost ~]$ hdfs dfs -ls /optional_HW/accounts
Found 2 items
-rw-rw-rw-   1 training supergroup          0 2020-06-01 21:19 /optional_HW/accounts/_SUCCESS
-rw-rw-rw-   1 training supergroup          0 2020-06-01 21:19 /optional_HW/accounts/part-m-00000
[training@localhost ~]$ jlc190004
```