BUAN 6346 Big Data Analytics
Phase 1

SHIVA KUMAR REDDY KOPPULA
02/13/2024

Table of Contents

CHAPTER 2 – SET UP 3

Environment Setup 3

CHAPTER 3 - ACCESS HDFS WITH COMMAND LINE AND HUE 3

Exploring the HDFS Command Line Interface 3

Upload Files to HDFS 4

View HDFS files 6

Use the Hue File Browser to browse, view and manage files 8

Chapter 3 – RUN A YARN JOB 12

Explore the YARN cluster 12

Submit an application to the YARN cluster 13

View the application in the Hue Job Browser 15

View the application in the YARN UI 16

CHAPTER 4 - IMPORT DATA FROM MYSQL USING SQOOP 18

Import the accounts table from MySQL 18

View the imported data 21

Import incremental updates to accounts 22

Import webpage data using an alternate field delimiter 24

CHAPTER 2 – SET UP

Environment Setup

1. After installing VMware and configuring my virtual machine, I executed the following setup script in the terminal window to activate services and prepare necessary data for the course and lab exercises.

```
[training@localhost ~]$ $DEV1/scripts/training_setup_dev1.sh
[Dev] Setting up for Dev1Training
* Setting up Eclipse workspace for Dev1
* /home/training/workspace already exists, moving to /home/training/workspace.sa
ve.dev1
* ERROR: /home/training/workspace.save.dev1 already exists. Delete it and retry.
* Enabling and starting services required for Dev1 Training
Started Impala Catalog Server (catalogd) :                 [  OK  ]
Stopping Apache Archiva...
Apache Archiva was not running.
Stopped Hive Server2:                                      [  OK  ]
* Disabling services not required for Dev1 Training
no secondarynamenode to stop
Stopped Hadoop secondarynamenode:                          [  OK  ]
Flume agent is not running                                 [  OK  ]
Stopping Hadoop HBase regionserver daemon: no regionserver to stop because no pi
d file /var/run/hbase/hbase-hbase-regionserver.pid
hbase-regionserver.
no master to stop because no pid file /var/run/hbase/hbase-hbase-master.pid
Stopped HBase master daemon:                               [  OK  ]
no rest to stop because no pid file /var/run/hbase/hbase-hbase-rest.pid
Stopped HBase rest daemon:                                 [  OK  ]
no thrift to stop because no pid file /var/run/hbase/hbase-hbase-thrift.pid
Stopped HBase thrift daemon:                               [  OK  ]
Stopped Kafka Server:                                      [  OK  ]
Solr server daemon is not running                          [  OK  ]
[Dev] Done setting up for Dev1 Training
[training@localhost ~]$ sxk230064
```

CHAPTER 3 - ACCESS HDFS WITH COMMAND LINE AND HUE

Exploring the HDFS Command Line Interface

1. I opened a terminal window and ran the script below to view the contents of the root directory in HDFS, revealing multiple entries.

```
[training@localhost ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx   - hbase    supergroup          0 2015-08-27 05:37 /hbase
drwxrwxrwx   - training supergroup          0 2024-02-08 15:01 /loudacre
drwxrwxrwx   - solr     supergroup          0 2015-08-27 05:32 /solr
drwxrwxrwt   - hdfs     supergroup          0 2024-02-09 15:33 /tmp
drwxrwxrwx   - hdfs     supergroup          0 2024-02-08 14:58 /user
drwxrwxrwx   - hdfs     supergroup          0 2015-08-27 05:31 /var
[training@localhost ~]$ sxk230064
```

2. From the list of entries, I then explored the contents of the /user directory and found the home directory listed.

```
[training@localhost ~]$ hdfs dfs -ls /user
Found 6 items
drwxrwxrwx   - hdfs     supergroup          0 2024-02-08 14:58 /user/hdfs
drwxrwxrwt   - mapred   hadoop              0 2015-08-27 05:31 /user/history
drwxrwxrwx   - hive     hive                0 2015-08-27 05:31 /user/hive
drwxrwxrwx   - oozie    oozie               0 2015-08-27 05:32 /user/oozie
drwxr-xr-x   - spark    spark               0 2015-08-27 05:32 /user/spark
drwxrwxrwx   - training supergroup          0 2015-08-27 05:31 /user/training
[training@localhost ~]$ sxk230064
```

3. Using the 'ls' command, I checked the contents of the home directory (/user/training), finding it empty, which returned no output.

```
[training@localhost ~]$ hdfs dfs -ls /user/training
[training@localhost ~]$ sxk230064
```

Upload Files to HDFS

1. To upload files into HDFS, I first created a new top-level directory (/loudacre/) for homework assignments using the command 'hdfs dfs -mkdir'. Note that upon execution, the command returned "mkdir: '/loudacre' file exists", as I had already performed this step earlier as per the setup virtual machine document.

```
[training@localhost ~]$ hdfs dfs -mkdir /loudacre
mkdir: `/loudacre': File exists
[training@localhost ~]$ sxk230064
```

2. I then changed the directory to $DEV1DATA, a local filesystem directory containing the sample data for the course, using the 'cd' command.

```
[training@localhost ~]$ cd $DEV1DATA
[training@localhost data]$ sxk230064
```

3. Using the 'hdfs dfs -put' command, I transferred the local kb directory, containing knowledge base articles, and its contents into a remote HDFS directory named /loudacre/kb. Similarly, executing the command returned "file exists" due to prior completion of these steps as per the setup virtual machine document.

```
[training@localhost data]$ hdfs dfs -put kb /loudacre/
put: `/loudacre/kb/KBDOC-00001.html': File exists
put: `/loudacre/kb/KBDOC-00002.html': File exists
put: `/loudacre/kb/KBDOC-00003.html': File exists
put: `/loudacre/kb/KBDOC-00004.html': File exists
put: `/loudacre/kb/KBDOC-00005.html': File exists
```

```
put: `/loudacre/kb/KBDOC-00280.html': File exists
put: `/loudacre/kb/KBDOC-00281.html': File exists
put: `/loudacre/kb/KBDOC-00282.html': File exists
put: `/loudacre/kb/KBDOC-00283.html': File exists
put: `/loudacre/kb/KBDOC-00284.html': File exists
put: `/loudacre/kb/KBDOC-00285.html': File exists
put: `/loudacre/kb/KBDOC-00286.html': File exists
put: `/loudacre/kb/KBDOC-00287.html': File exists
put: `/loudacre/kb/KBDOC-00288.html': File exists
put: `/loudacre/kb/KBDOC-00289.html': File exists
put: `/loudacre/kb/KBDOC-00290.html': File exists
put: `/loudacre/kb/KBDOC-00291.html': File exists
put: `/loudacre/kb/KBDOC-00292.html': File exists
put: `/loudacre/kb/KBDOC-00293.html': File exists
put: `/loudacre/kb/KBDOC-00294.html': File exists
put: `/loudacre/kb/KBDOC-00295.html': File exists
put: `/loudacre/kb/KBDOC-00296.html': File exists
put: `/loudacre/kb/KBDOC-00297.html': File exists
put: `/loudacre/kb/KBDOC-00298.html': File exists
put: `/loudacre/kb/KBDOC-00299.html': File exists
put: `/loudacre/kb/KBDOC-00300.html': File exists
[training@localhost data]$ sxk230064
```

4. A subsequent 'ls' command showed the new HDFS directory containing 300 items, all with "read-write" permissions for every user type.

```
[training@localhost data]$ hdfs dfs -ls /loudacre/kb
Found 300 items
-rw-rw-rw-   1 training supergroup       1626 2024-02-08 15:01 /loudacre/kb/KBDOC-00001.html
-rw-rw-rw-   1 training supergroup       1298 2024-02-08 15:01 /loudacre/kb/KBDOC-00002.html
-rw-rw-rw-   1 training supergroup       1183 2024-02-08 15:01 /loudacre/kb/KBDOC-00003.html
-rw-rw-rw-   1 training supergroup       1029 2024-02-08 15:01 /loudacre/kb/KBDOC-00004.html
-rw-rw-rw-   1 training supergroup       1155 2024-02-08 15:01 /loudacre/kb/KBDOC-00005.html
-rw-rw-rw-   1 training supergroup       1109 2024-02-08 15:01 /loudacre/kb/KBDOC-00006.html
-rw-rw-rw-   1 training supergroup       1648 2024-02-08 15:01 /loudacre/kb/KBDOC-00007.html
-rw-rw-rw-   1 training supergroup       1303 2024-02-08 15:01 /loudacre/kb/KBDOC-00008.html
-rw-rw-rw-   1 training supergroup       1173 2024-02-08 15:01 /loudacre/kb/KBDOC-00009.html
-rw-rw-rw-   1 training supergroup       1017 2024-02-08 15:01 /loudacre/kb/KBDOC-00010.html
-rw-rw-rw-   1 training supergroup       1149 2024-02-08 15:01 /loudacre/kb/KBDOC-00011.html
-rw-rw-rw-   1 training supergroup       1117 2024-02-08 15:01 /loudacre/kb/KBDOC-00012.html
-rw-rw-rw-   1 training supergroup       1638 2024-02-08 15:01 /loudacre/kb/KBDOC-00013.html
-rw-rw-rw-   1 training supergroup       1282 2024-02-08 15:01 /loudacre/kb/KBDOC-00014.html
-rw-rw-rw-   1 training supergroup       1166 2024-02-08 15:01 /loudacre/kb/KBDOC-00015.html
-rw-rw-rw-   1 training supergroup       1017 2024-02-08 15:01 /loudacre/kb/KBDOC-00016.html
-rw-rw-rw-   1 training supergroup       1160 2024-02-08 15:01 /loudacre/kb/KBDOC-00017.html
-rw-rw-rw-   1 training supergroup       1105 2024-02-08 15:01 /loudacre/kb/KBDOC-00018.html
-rw-rw-rw-   1 training supergroup       1612 2024-02-08 15:01 /loudacre/kb/KBDOC-00019.html
-rw-rw-rw-   1 training supergroup       1274 2024-02-08 15:01 /loudacre/kb/KBDOC-00020.html
-rw-rw-rw-   1 training supergroup       1176 2024-02-08 15:01 /loudacre/kb/KBDOC-00021.html
-rw-rw-rw-   1 training supergroup       1024 2024-02-08 15:01 /loudacre/kb/KBDOC-00022.html
-rw-rw-rw-   1 training supergroup       1149 2024-02-08 15:01 /loudacre/kb/KBDOC-00023.html
-rw-rw-rw-   1 training supergroup       1103 2024-02-08 15:01 /loudacre/kb/KBDOC-00024.html
-rw-rw-rw-   1 training supergroup       1580 2024-02-08 15:01 /loudacre/kb/KBDOC-00025.html
-rw-rw-rw-   1 training supergroup       1242 2024-02-08 15:01 /loudacre/kb/KBDOC-00026.html
-rw-rw-rw-   1 training supergroup       1143 2024-02-08 15:01 /loudacre/kb/KBDOC-00027.html
-rw-rw-rw-   1 training supergroup       1013 2024-02-08 15:01 /loudacre/kb/KBDOC-00028.html
-rw-rw-rw-   1 training supergroup       1130 2024-02-08 15:01 /loudacre/kb/KBDOC-00029.html
-rw-rw-rw-   1 training supergroup       1101 2024-02-08 15:01 /loudacre/kb/KBDOC-00030.html
-rw-rw-rw-   1 training supergroup       1630 2024-02-08 15:01 /loudacre/kb/KBDOC-00031.html
-rw-rw-rw-   1 training supergroup       1258 2024-02-08 15:01 /loudacre/kb/KBDOC-00032.html
-rw-rw-rw-   1 training supergroup       1143 2024-02-08 15:01 /loudacre/kb/KBDOC-00033.html
-rw-rw-rw-   1 training supergroup       1003 2024-02-08 15:01 /loudacre/kb/KBDOC-00034.html
-rw-rw-rw-   1 training supergroup       1118 2024-02-08 15:01 /loudacre/kb/KBDOC-00035.html
-rw-rw-rw-   1 training supergroup       1096 2024-02-08 15:01 /loudacre/kb/KBDOC-00036.html
-rw-rw-rw-   1 training supergroup       1602 2024-02-08 15:01 /loudacre/kb/KBDOC-00037.html
```

```
-rw-rw-rw-   1 training supergroup       1165 2024-02-08 15:01 /loudacre/kb/KBDOC-00263.html
-rw-rw-rw-   1 training supergroup       1109 2024-02-08 15:01 /loudacre/kb/KBDOC-00264.html
-rw-rw-rw-   1 training supergroup       1639 2024-02-08 15:01 /loudacre/kb/KBDOC-00265.html
-rw-rw-rw-   1 training supergroup       1302 2024-02-08 15:01 /loudacre/kb/KBDOC-00266.html
-rw-rw-rw-   1 training supergroup       1179 2024-02-08 15:01 /loudacre/kb/KBDOC-00267.html
-rw-rw-rw-   1 training supergroup       1046 2024-02-08 15:01 /loudacre/kb/KBDOC-00268.html
-rw-rw-rw-   1 training supergroup       1132 2024-02-08 15:01 /loudacre/kb/KBDOC-00269.html
-rw-rw-rw-   1 training supergroup       1122 2024-02-08 15:01 /loudacre/kb/KBDOC-00270.html
-rw-rw-rw-   1 training supergroup       1618 2024-02-08 15:01 /loudacre/kb/KBDOC-00271.html
-rw-rw-rw-   1 training supergroup       1268 2024-02-08 15:01 /loudacre/kb/KBDOC-00272.html
-rw-rw-rw-   1 training supergroup       1151 2024-02-08 15:01 /loudacre/kb/KBDOC-00273.html
-rw-rw-rw-   1 training supergroup       1020 2024-02-08 15:01 /loudacre/kb/KBDOC-00274.html
-rw-rw-rw-   1 training supergroup       1105 2024-02-08 15:01 /loudacre/kb/KBDOC-00275.html
-rw-rw-rw-   1 training supergroup       1101 2024-02-08 15:01 /loudacre/kb/KBDOC-00276.html
-rw-rw-rw-   1 training supergroup       1604 2024-02-08 15:01 /loudacre/kb/KBDOC-00277.html
-rw-rw-rw-   1 training supergroup       1253 2024-02-08 15:01 /loudacre/kb/KBDOC-00278.html
-rw-rw-rw-   1 training supergroup       1169 2024-02-08 15:01 /loudacre/kb/KBDOC-00279.html
-rw-rw-rw-   1 training supergroup       1021 2024-02-08 15:01 /loudacre/kb/KBDOC-00280.html
-rw-rw-rw-   1 training supergroup       1135 2024-02-08 15:01 /loudacre/kb/KBDOC-00281.html
-rw-rw-rw-   1 training supergroup       1115 2024-02-08 15:01 /loudacre/kb/KBDOC-00282.html
-rw-rw-rw-   1 training supergroup       1608 2024-02-08 15:01 /loudacre/kb/KBDOC-00283.html
-rw-rw-rw-   1 training supergroup       1263 2024-02-08 15:01 /loudacre/kb/KBDOC-00284.html
-rw-rw-rw-   1 training supergroup       1151 2024-02-08 15:01 /loudacre/kb/KBDOC-00285.html
-rw-rw-rw-   1 training supergroup       1017 2024-02-08 15:01 /loudacre/kb/KBDOC-00286.html
-rw-rw-rw-   1 training supergroup       1129 2024-02-08 15:01 /loudacre/kb/KBDOC-00287.html
-rw-rw-rw-   1 training supergroup       1111 2024-02-08 15:01 /loudacre/kb/KBDOC-00288.html
-rw-rw-rw-   1 training supergroup       1612 2024-02-08 15:01 /loudacre/kb/KBDOC-00289.html
-rw-rw-rw-   1 training supergroup       1253 2024-02-08 15:01 /loudacre/kb/KBDOC-00290.html
-rw-rw-rw-   1 training supergroup       1164 2024-02-08 15:01 /loudacre/kb/KBDOC-00291.html
-rw-rw-rw-   1 training supergroup       1041 2024-02-08 15:01 /loudacre/kb/KBDOC-00292.html
-rw-rw-rw-   1 training supergroup       1110 2024-02-08 15:01 /loudacre/kb/KBDOC-00293.html
-rw-rw-rw-   1 training supergroup       1097 2024-02-08 15:01 /loudacre/kb/KBDOC-00294.html
-rw-rw-rw-   1 training supergroup       1625 2024-02-08 15:01 /loudacre/kb/KBDOC-00295.html
-rw-rw-rw-   1 training supergroup       1270 2024-02-08 15:01 /loudacre/kb/KBDOC-00296.html
-rw-rw-rw-   1 training supergroup       1167 2024-02-08 15:01 /loudacre/kb/KBDOC-00297.html
-rw-rw-rw-   1 training supergroup       1015 2024-02-08 15:01 /loudacre/kb/KBDOC-00298.html
-rw-rw-rw-   1 training supergroup       1126 2024-02-08 15:01 /loudacre/kb/KBDOC-00299.html
-rw-rw-rw-   1 training supergroup       1097 2024-02-08 15:01 /loudacre/kb/KBDOC-00300.html
[training@localhost data]$ sxk230064
```

5. Before viewing HDFS files, I practiced uploading and removing a directory using the '-put' and '-rm' flags of the 'hdfs dfs' command.

```
[training@localhost data]$ hdfs dfs -put $DEV1DATA/calllogs /loudacre/
[training@localhost data]$ hdfs dfs -rm -r /loudacre/calllogs
24/02/10 16:00:19 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /loudacre/calllogs
[training@localhost data]$ sxk230064
```

View HDFS files

6. To view some of the data recently copied into HDFS, I used the command below to print the last 20 lines of the KBDOC-00289.html file to the terminal.

```
[training@localhost data]$ hdfs dfs -cat /loudacre/kb/KBDOC-00289.html | tail \-n 20
</ol>
<h3>Steps to change the notification ringtone for texts on Ronin S2</h3>
<ol>
<li>From the Personalization screen, tap Settings.</li>
<li>Tap Sounds.</li>
<li>Tap Text Tone.</li>
<li>Tap your preferred ringtone.</li>
<li>You've now changed the ringtone.</li>
</ol>
<h3>Steps to change the notification ringtone for emails on Ronin S2</h3>
<ol>
<li>From the Personalization screen, tap Settings.</li>
<li>Tap Sounds.</li>
<li>Scroll to and tap New Mail.</li>
<li>Tap your preferred ringtone.</li>
<li>You've now changed the ringtone.</li>
</ol>
 <h4>Tags: smartphone, changes, contact, themselves, yours</h4>
 </body>
[training@localhost data]$ sxk230064
```

7. I employed the 'hdfs dfs -get' command to retrieve a local copy of a file from HDFS to the local filesystem as article.html, and then used the 'less' command to explore its contents.

```
[training@localhost data]$ hdfs dfs -get \
> /loudacre/kb/KBDOC-00289.html ~/article.html
[training@localhost data]$ sxk230064
```

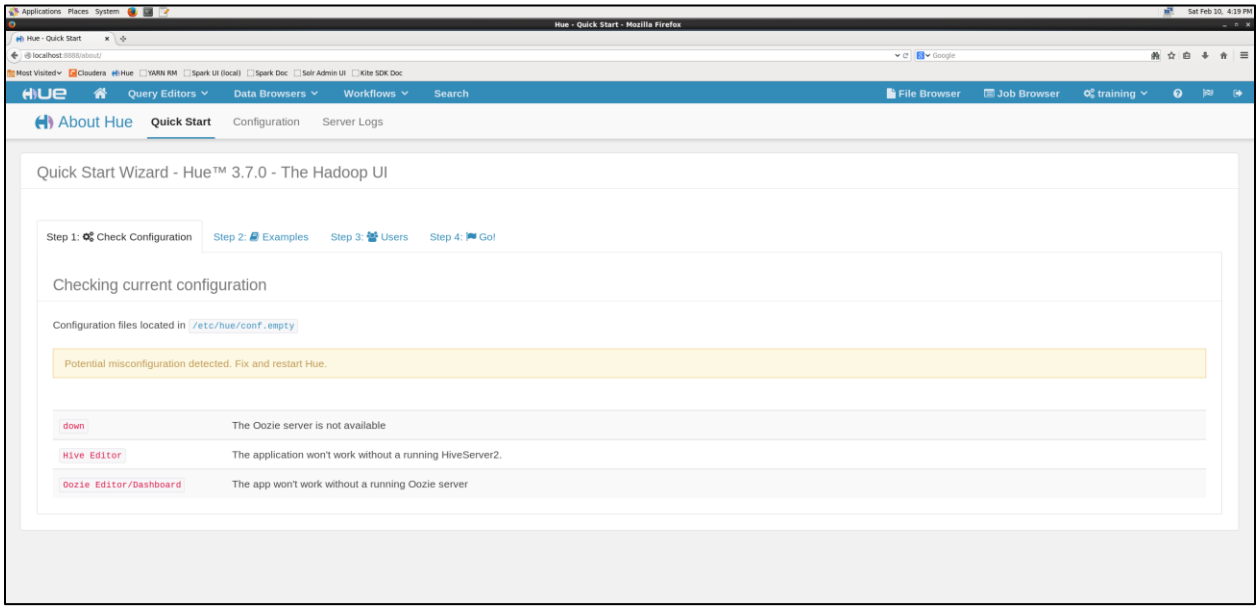
```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html  xml:lang="en" lang="en">
 <head>
 <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
 <meta name="device" content="Ronin S2" />
 <meta name="docid" content="KBDOC-00289" />
 <meta name="createddate" content="2012-07-21" />
 <meta name="updateddate" content="2012-08-03" />
 <title>Ronin S2 - Change the phone ringtone and notification sound</title>
 <link rel="stylesheet" type="text/css" href="theme.css" />
 </head>
 <body>
 <h1>Ronin S2 - Change the phone ringtone and notification sound</h1>
<h2>Document Id: KBDOC-00289</h2>
<h2>Created: 2012-07-21</h2>
<h2>Updated: 2012-08-03</h2>
<h3>Steps to change the phone ringtone for calls on your Ronin S2</h3>
<ol>
<li>From the Personalization screen, tap Settings.</li>
<li>Tap Sounds.</li>
<li>Tap Ringtone.</li>
<li>Tap your preferred ringtone.</li>
<li>You've now changed the ringtone.</li>
</ol>
<h3>Steps to change the notification ringtone for texts on Ronin S2</h3>
<ol>
<li>From the Personalization screen, tap Settings.</li>
<li>Tap Sounds.</li>
<li>Tap Text Tone.</li>
<li>Tap your preferred ringtone.</li>
<li>You've now changed the ringtone.</li>
</ol>
<h3>Steps to change the notification ringtone for emails on Ronin S2</h3>
<ol>
<li>From the Personalization screen, tap Settings.</li>
<li>Tap Sounds.</li>
<li>Scroll to and tap New Mail.</li>
<li>Tap your preferred ringtone.</li>
<li>You've now changed the ringtone.</li>
</ol>
 <h4>Tags: smartphone, changes, contact, themselves, yours</h4>
 </body>
</html>
~
(END)
```

8. To conclude this section of the lab, I explored additional operations available with the 'hdfs dfs' commands by running 'hdfs dfs' in the terminal, revealing various common filesystem manipulations.

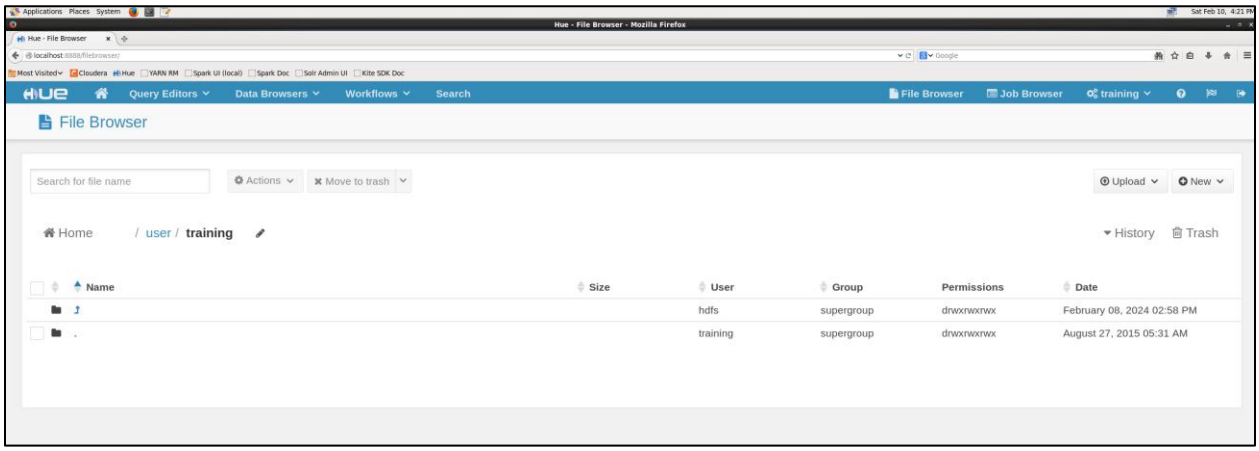
```
[training@localhost data]$  hdfs dfs
Usage: hadoop fs [generic options]
        [-appendToFile <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum <src> ...]
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
        [-chown [-R] [OWNER][:[GROUP]] PATH...]
        [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
        [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-count [-q] [-h] <path> ...]
        [-cp [-f] [-p | -p[topax]] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] <path> ...]
        [-expunge]
        [-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-getfacl [-R] <path>]
        [-getfattr [-R] {-n name | -d} [-e en] <path>]
        [-getmerge [-nl] <src> <localdst>]
        [-help [cmd ...]]
        [-ls [-d] [-h] [-R] [<path> ...]]
        [-mkdir [-p] <path> ...]
        [-moveFromLocal <localsrc> ... <dst>]
        [-moveToLocal <src> <localdst>]
        [-mv <src> ... <dst>]
        [-put [-f] [-p] [-l] <localsrc> ... <dst>]
        [-renameSnapshot <snapshotDir> <oldName> <newName>]
        [-rm [-f] [-r|-R] [-skipTrash] <src> ...]
        [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
        [-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>]|[--set <acl_spec> <path>]]
        [-setfattr {-n name [-v value] | -x name} <path>]
        [-setrep [-R] [-w] <rep> <path> ...]
        [-stat [format] <path> ...]
        [-tail [-f] <file>]
        [-test -[defsz] <path>]
        [-text [-ignoreCrc] <src> ...]
        [-touchz <path> ...]
        [-usage [cmd ...]]

Generic options supported are
-conf <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]

[training@localhost data]$ sxk230064
```

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

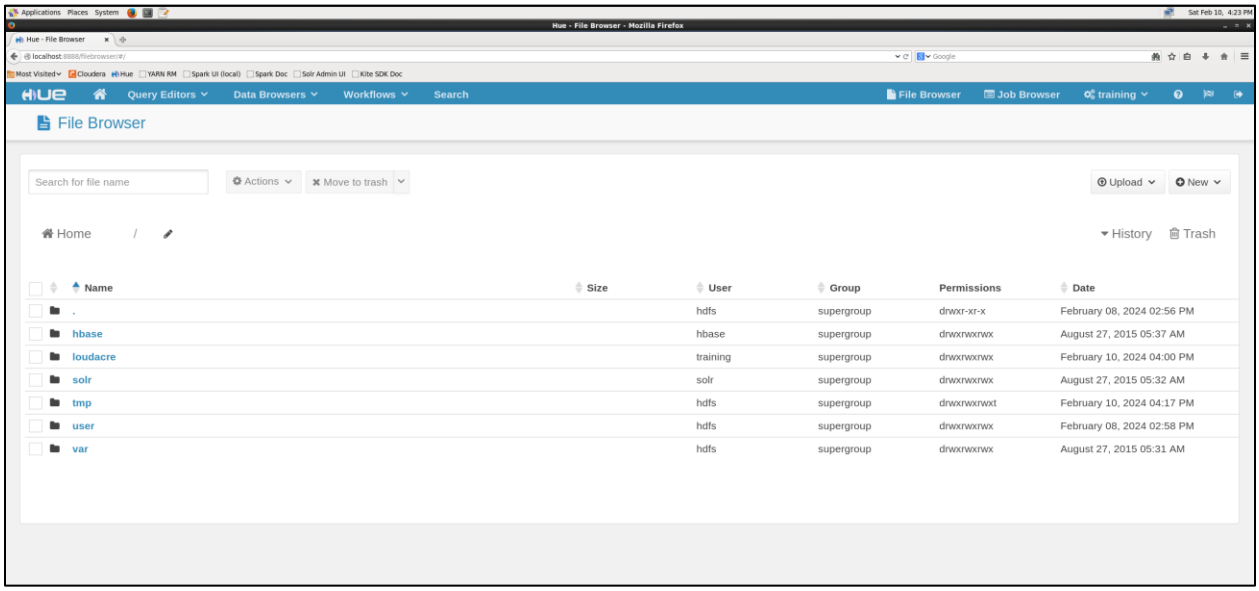
9. I initiated Firefox from the main menu panel of the virtual machine, accessing Cloudera's website as the home page. Navigating to the bookmarks, I clicked on the HUE bookmark to land on the respective page. Notably, account creation had already been completed as instructed in the VM setup document, therefore, I'm unable to produce any screenshots.



10. Within the above screenshot, I accessed the "file browser" located at the top-right of the page transitioning to the below screenshot.

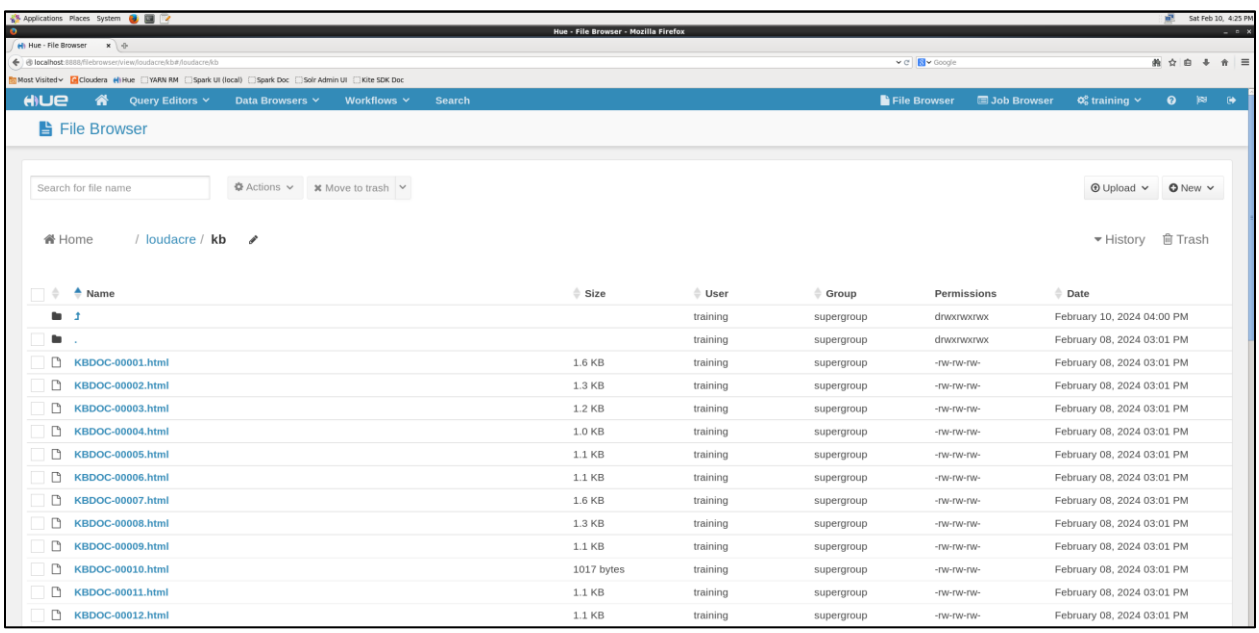


11. Further navigation took me to the root directory by clicking on the forward slash preceding "user" in the file path, leading to the screenshot below.



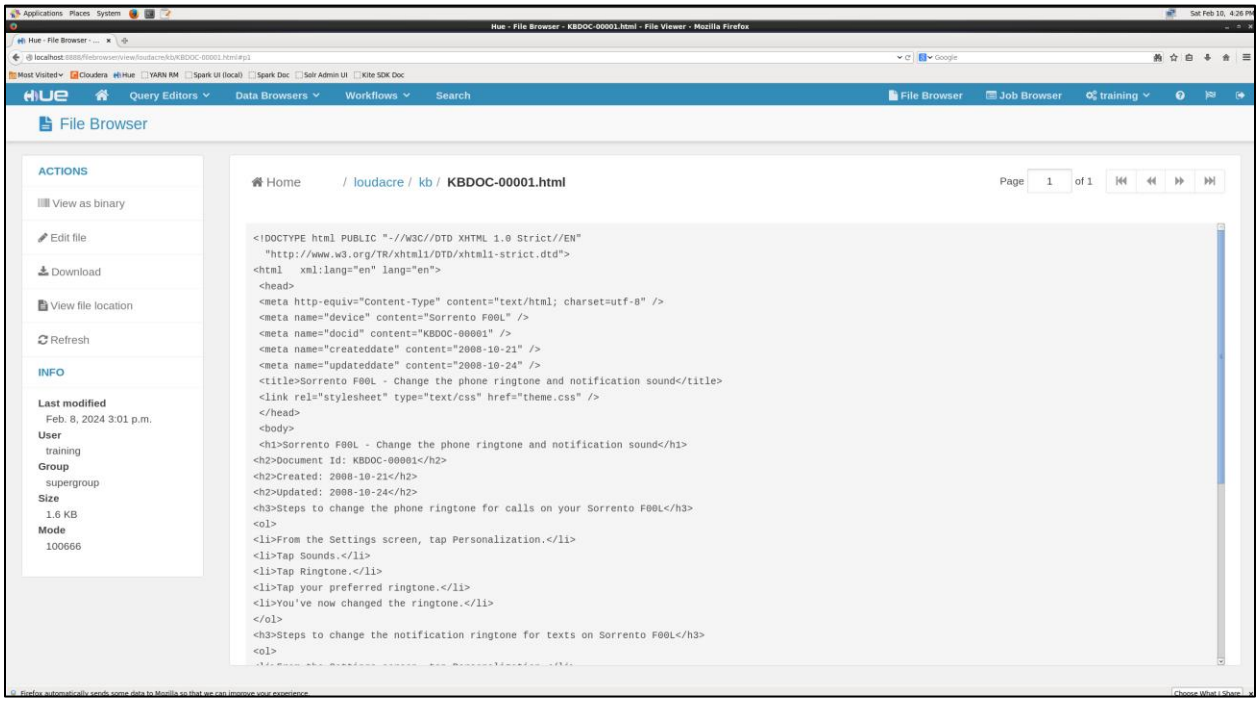
12. The contents of the root directory displayed, including the previously created loudacre directory, which I clicked to view its contents.

13. Subsequently, I accessed the kb directory to view the knowledge base articles uploaded earlier.

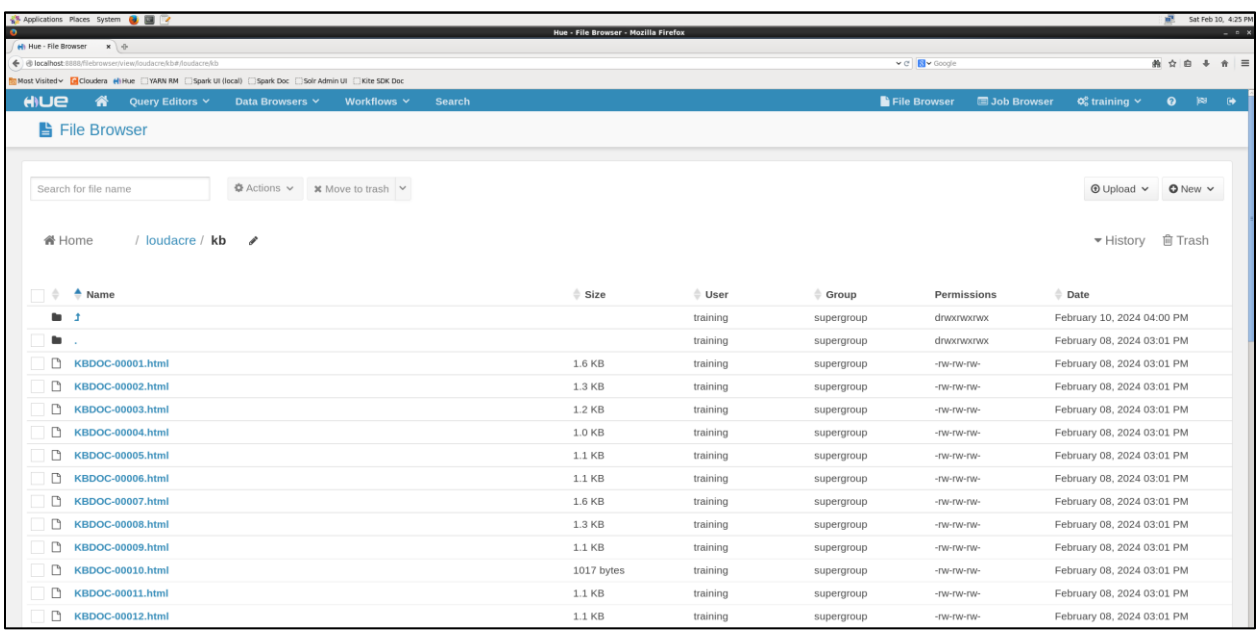


14. Clicking on one of the kb articles, "KBDOC-00001.html", allowed me to view its contents.

15. Observing the available actions and information in the file viewer page, I noted the ability to edit, download, and refresh the file, as well as navigate through pages and view the file as binary code.

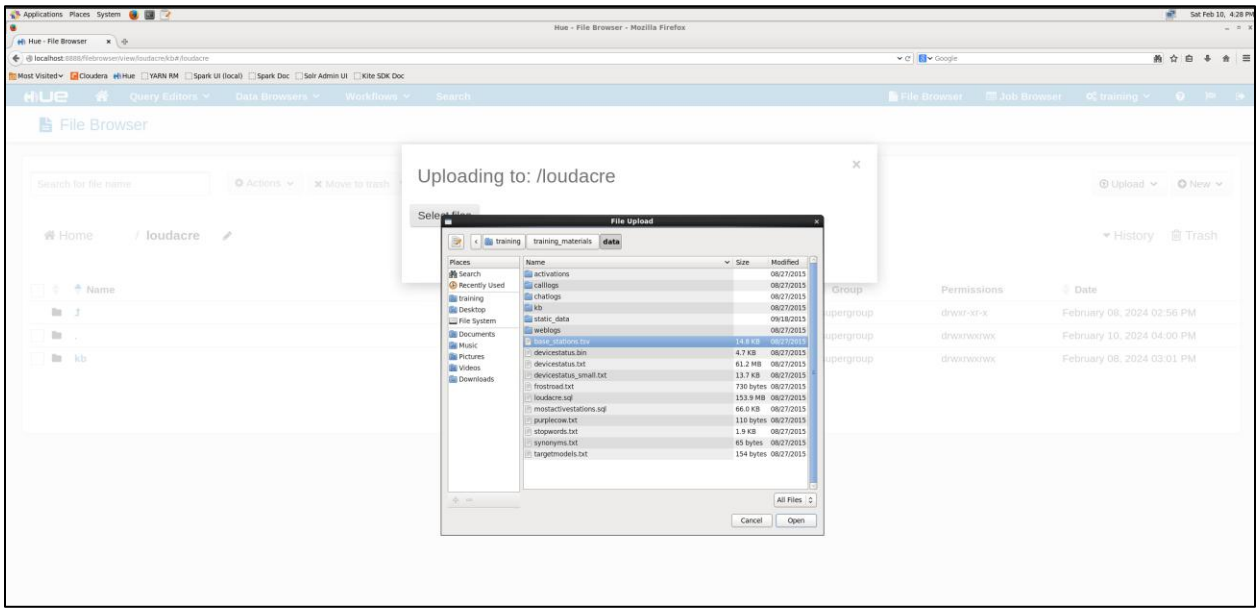


16. Using the "View file location" option in the Action panel, I returned to the parent directory (/loudacre/kb).

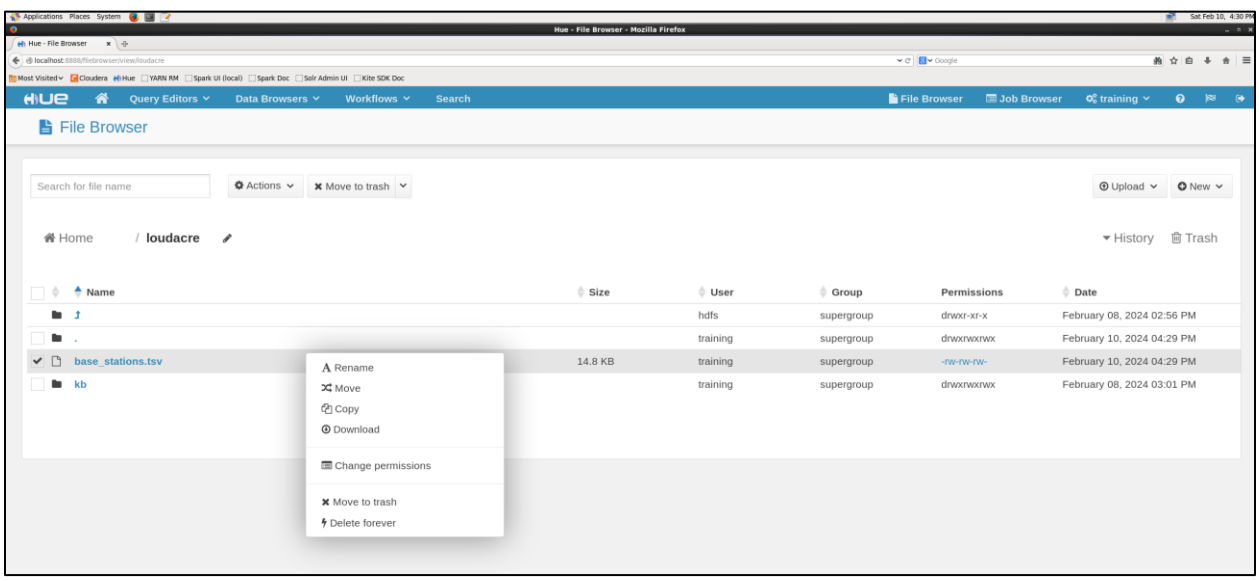


17. Clicking on the upward-facing arrow below "Name" facilitated a return to the /loudacre base directory.

18. Proceeding, I uploaded a file to the loudacre directory by clicking on "upload", then "files", followed by "select files", and navigating to /home/training/training_materials/data. Upon selection of the “base_stations.tsv” file and clicking "open", the upload process was completed.



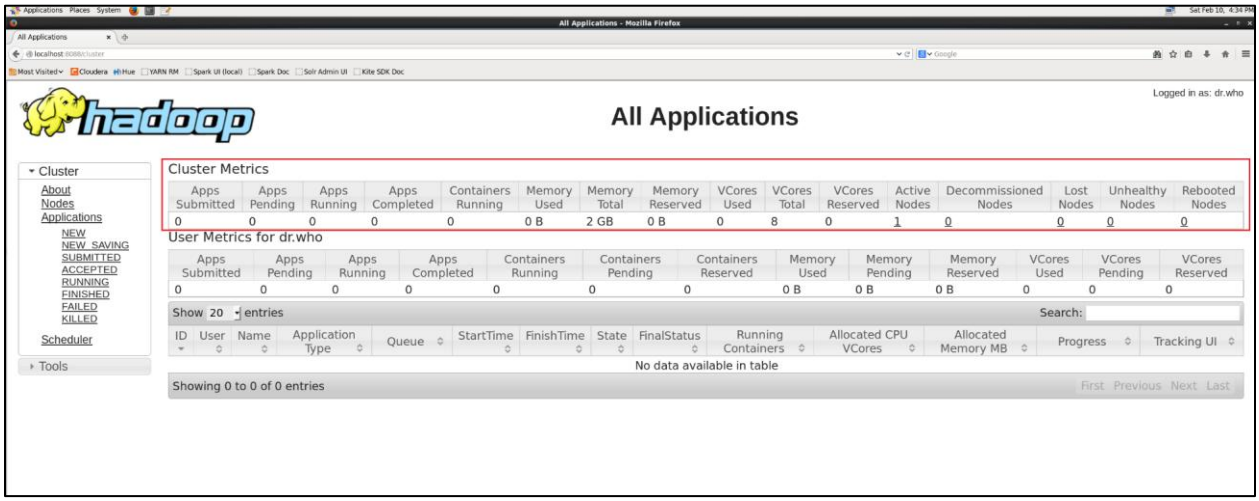
19. Upon confirmation of the upload, the base_stations.tsv file appeared in the /loudacre/ directory. Further actions such as rename, move, copy, download, and changing permissions were explored by selecting the file and clicking on "Actions" from the Hue File Browser interface.



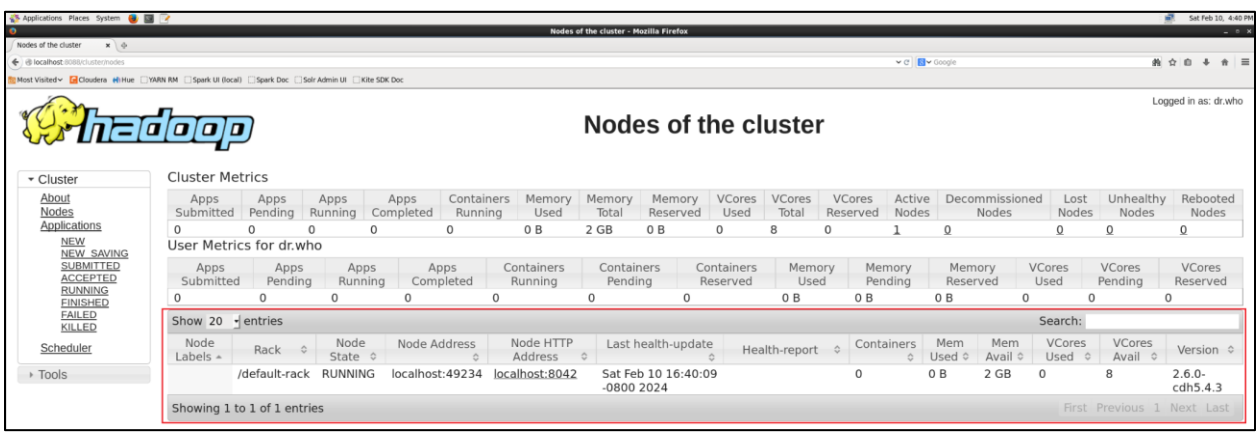
Chapter 3 – RUN A YARN JOB

Explore the YARN cluster

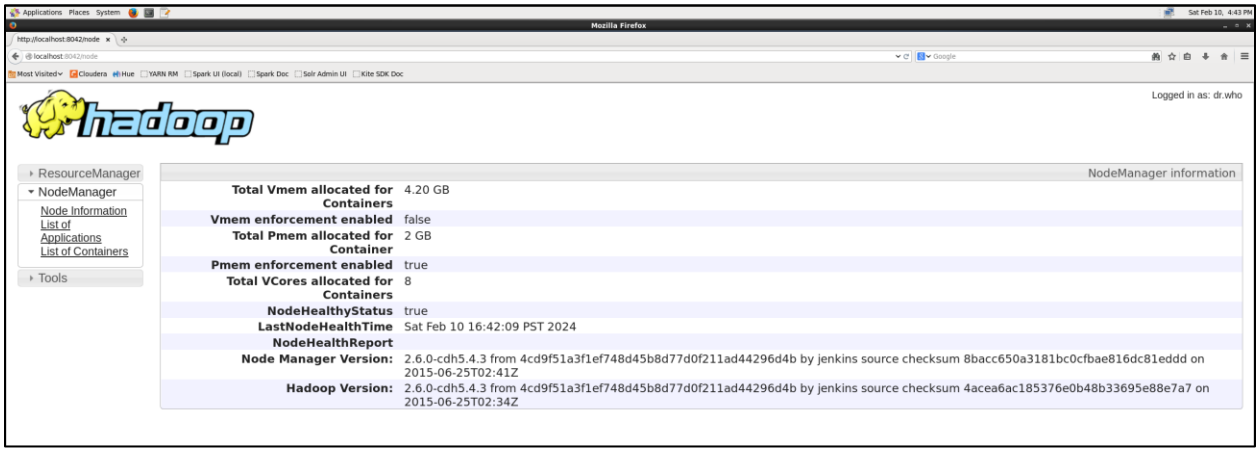
1. Initially, I opened Firefox from the main menu panel and accessed the YARN Resource Manager page via the bookmarks bar.



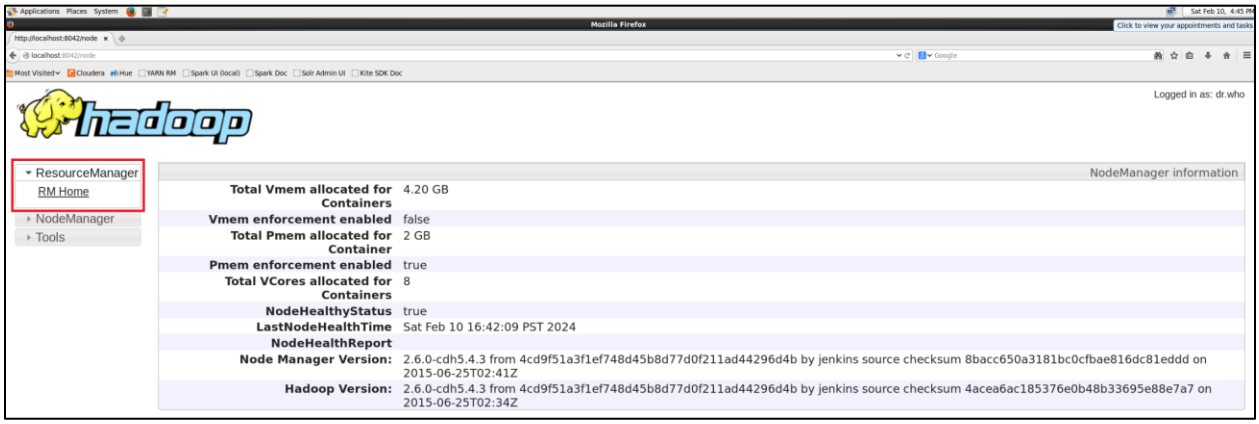
2. I took note of the values presented in the "Cluster Metrics" section, which includes details regarding applications, containers, memory, VCores, and active nodes. Notably, I observed that there is one active node, indicating that the cluster comprises solely the local virtual machine as a node, along with the status of worker nodes within the cluster.



3. Following that, I selected "Nodes" from the "Cluster" drop-down menu situated on the left side of the page in the aforementioned snippet. Subsequently, I clicked on the "Node HTTP Address," leading me to a page presenting diverse statistics concerning the node, as outlined below.

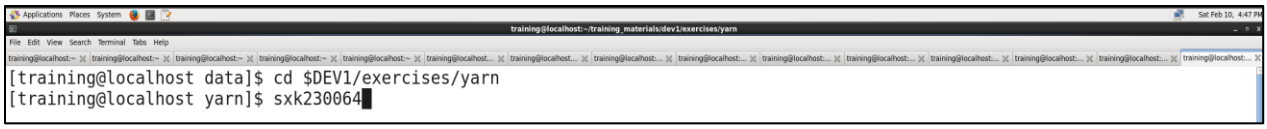


4. Upon reviewing the details presented on the Node Manager UI page, I navigated back to the resource manager interface by selecting "RM Home" from the "ResourceManager" drop-down menu positioned on the left side of the page.



Submit an application to the YARN cluster

5. In order to submit an application to the YARN cluster, I began by switching the current working directory to “DEV1/exercises/yarn” by reopening the terminal window.



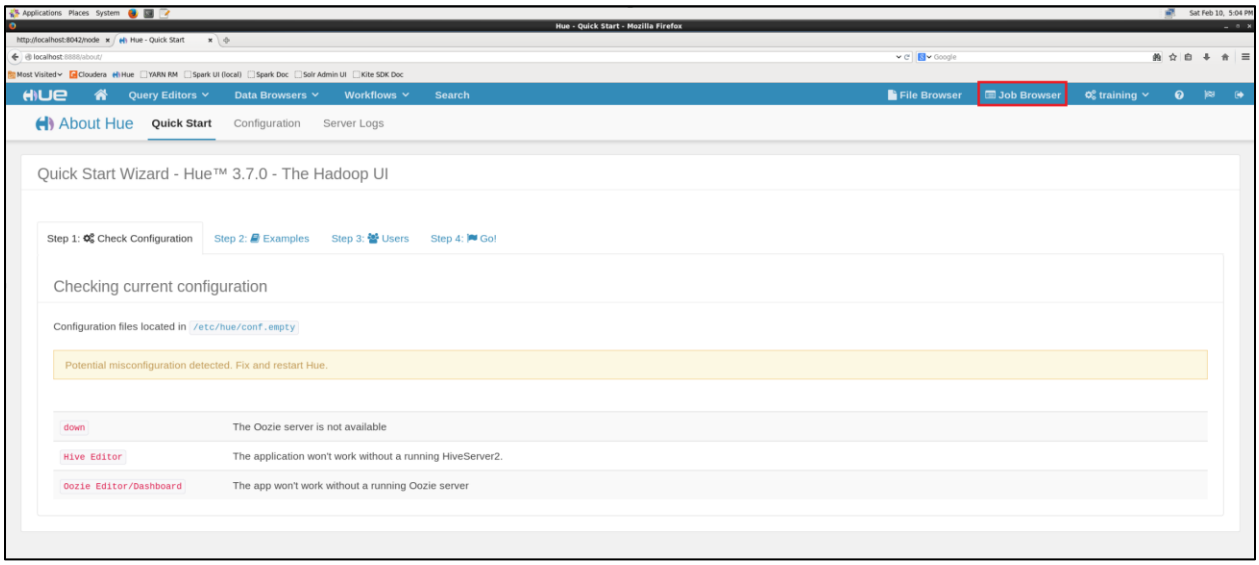
6. Subsequently, I executed the wordcount.py program on the YARN cluster to tally the occurrences of words within the knowledge base articles situated in /loudacre/kb/.

```
[training@localhost yarn]$ spark-submit --master yarn-cluster \
> wordcount.py /loudacre/kb/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/02/10 16:50:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/02/10 16:50:13 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
24/02/10 16:50:14 INFO yarn.Client: Requesting a new application from cluster with 1 NodeManagers
24/02/10 16:50:15 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (1024 MB per container)
24/02/10 16:50:15 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/02/10 16:50:15 INFO yarn.Client: Setting up container launch context for our AM
24/02/10 16:50:15 INFO yarn.Client: Preparing resources for our AM container
24/02/10 16:50:17 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/02/10 16:50:17 INFO yarn.Client: Uploading resource file:/usr/lib/spark/lib/spark-assembly-1.3.0-cdh5.4.3-hadoop2.6.0-cdh5.4.3.jar -> hdfs://localhost:8020/user/training/.sparkSt
24/02/10 16:50:21 INFO yarn.Client: Uploading resource file:/home/training/training_materials/dev1/exercises/yarn/wordcount.py -> hdfs://localhost:8020/user/training/.sparkStaging/a
24/02/10 16:50:21 INFO yarn.Client: Setting up the launch environment for our AM container
24/02/10 16:50:22 INFO spark.SecurityManager: Changing view acls to: training
24/02/10 16:50:22 INFO spark.SecurityManager: Changing modify acls to: training
24/02/10 16:50:22 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(training); users with modify permissions:
24/02/10 16:50:22 INFO yarn.Client: Submitting application 1 to ResourceManager
24/02/10 16:50:22 INFO impl.YarnClientImpl: Submitted application application_1707601577547_0001
24/02/10 16:50:23 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:23 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: root.training
         start time: 1707612622359
         final status: UNDEFINED
         tracking URL: http://localhost:8088/proxy/application_1707601577547_0001/
         user: training
24/02/10 16:50:24 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:25 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:26 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:27 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:28 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:29 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:30 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:31 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:32 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:33 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:34 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
24/02/10 16:50:35 INFO yarn.Client: Application report for application_1707601577547_0001 (state: ACCEPTED)
```

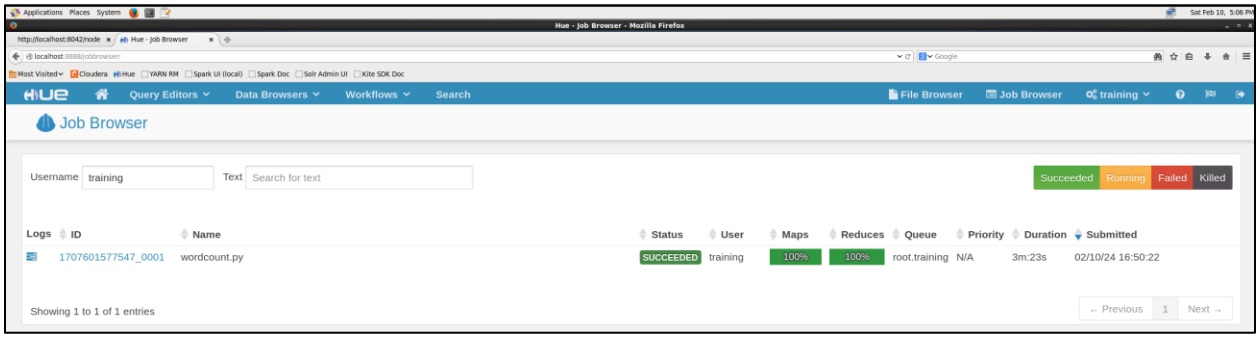
```
24/02/10 16:53:11 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:12 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:13 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:14 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:15 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:16 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:17 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:18 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:19 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:20 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:21 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:22 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:23 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:24 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:25 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:26 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:27 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:28 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:29 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:30 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:31 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:32 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:33 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:34 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:35 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:36 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:37 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:38 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:39 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:40 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:41 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:42 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:43 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:44 INFO yarn.Client: Application report for application_1707601577547_0001 (state: RUNNING)
24/02/10 16:53:45 INFO yarn.Client: Application report for application_1707601577547_0001 (state: FINISHED)
24/02/10 16:53:45 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: 192.168.188.128
         ApplicationMaster RPC port: 0
         queue: root.training
         start time: 1707612622359
         final status: SUCCEEDED
         tracking URL: http://localhost:8088/proxy/application_1707601577547_0001/history/application_1707601577547_0001
         user: training
[training@localhost yarn]$ sxk230064
```

View the application in the Hue Job Browser

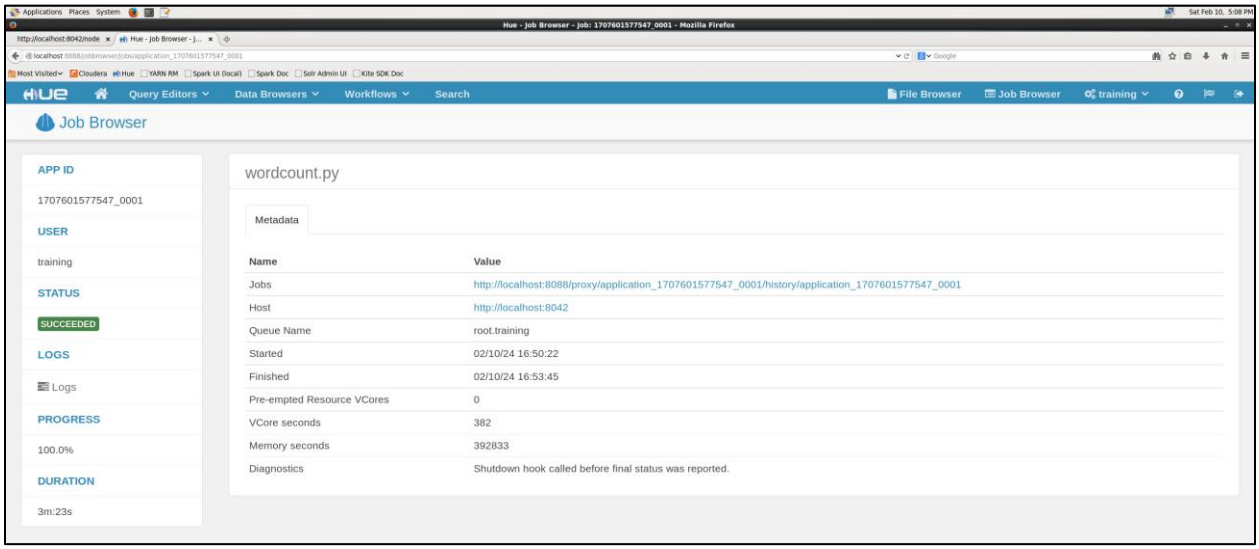
7. In order to access the application through the Hue Job Browser, I initially launched Firefox and selected the Hue bookmark from the bookmarks bar. Following that, I clicked on the job browser icon.



8. The job browser indicated that the earlier-run application had finished successfully. The legend located in the top-right corner provided information on how the job browser displays applications' statuses, indicating whether they have succeeded, are currently running, have failed, or have been terminated. Additionally, the job browser presented details regarding the duration of the job and its submission time.

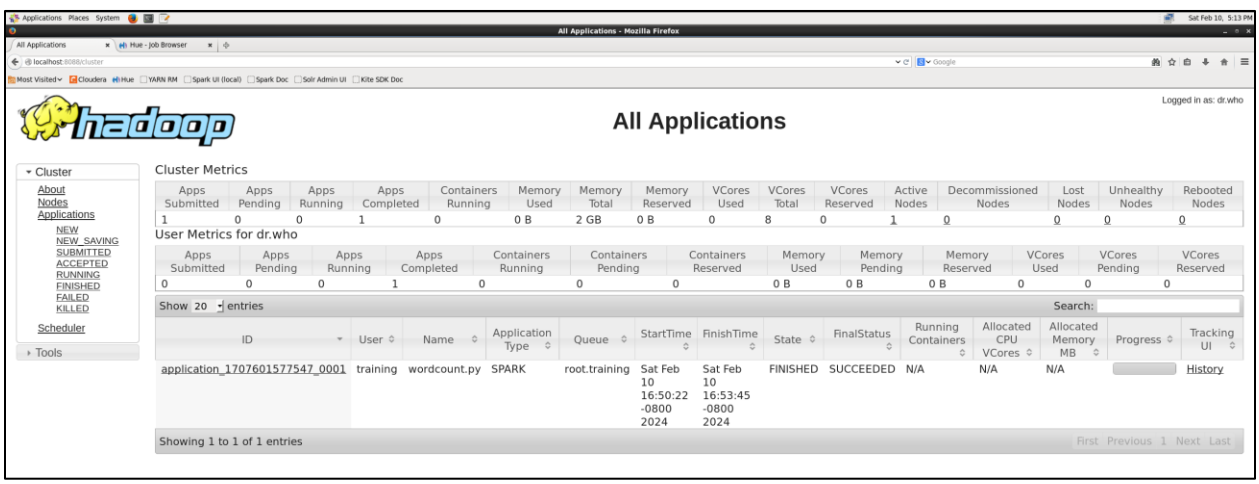


9. Upon noticing, I found that the job ID appears as a clickable link highlighted in blue. Subsequently, I clicked on it. Then, I advanced to the next section.



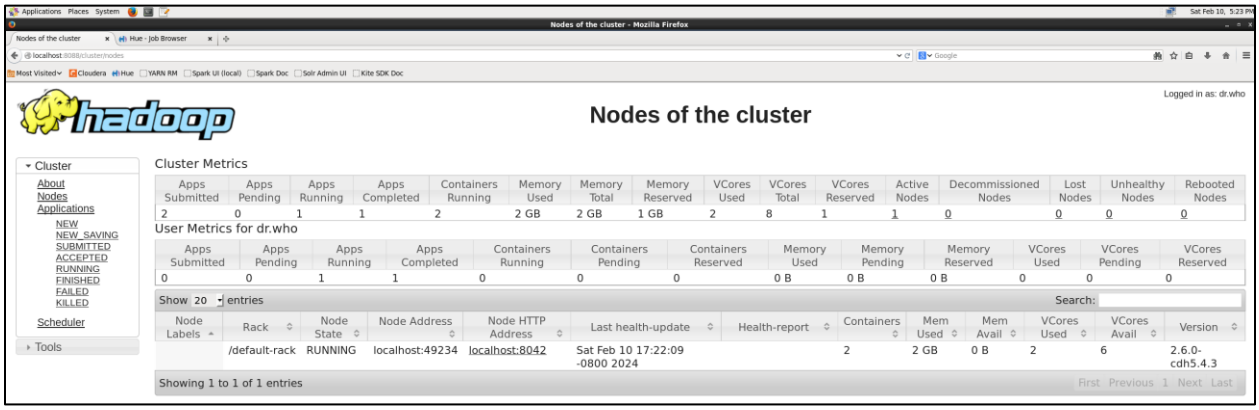
View the application in the YARN UI

10. I accessed the YARN resource manager page by utilizing the bookmark situated in Firefox's bookmarks bar. Upon arrival, I noticed the application I had recently executed positioned at the bottom of the page.

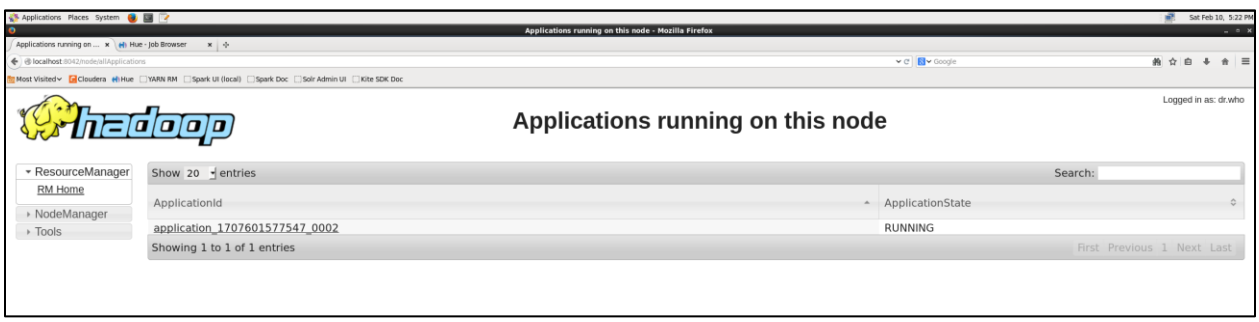


11. Next, I selected "Nodes" from the "Cluster" drop-down menu situated on the left side of the page.

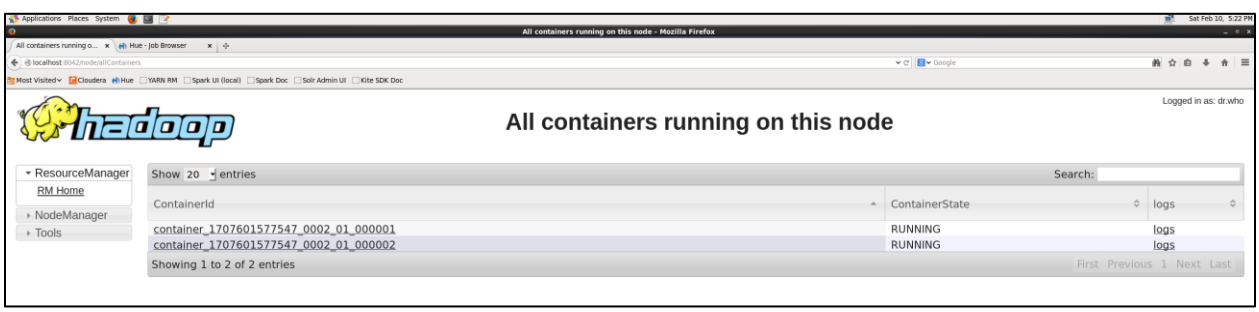
12. Following that, I proceeded to click on the "Node HTTP Address" to access the node manager user interface.



13. With the application currently running, I proceeded to click on "List of Applications" to view the submitted applications.



14. Finally, I navigated to "List of Containers" in the "NodeManager" drop-down menu. Initially, there were no containers visible on the node, as the previously launched application had already concluded. However, upon rerunning the application, I could observe two containers entries on this page.



CHAPTER 4 - IMPORT DATA FROM MYSQL USING SQOOP

Import the accounts table from MySQL

1. Initially, I launched a new terminal window. Following that, I executed the "sqoop help" command to get to know the available options in Sqoop.

```
[training@localhost yarn]$ sqoop help
24/02/10 17:26:48 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
usage: sqoop COMMAND [ARGS]

Available commands:
  codegen            Generate code to interact with database records
  create-hive-table  Import a table definition into Hive
  eval               Evaluate a SQL statement and display the results
  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 yarn]$ sxk230064
```

2. Following that, I employed the "sqoop list-tables" command to display the tables stored in the loudacre database.

```
[training@localhost yarn]$ sqoop list-tables \
> --connect jdbc:mysql://localhost/loudacre \
> --username training --password training
24/02/10 17:29:11 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/02/10 17:29:11 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/02/10 17:29:12 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
accountdevice
accounts
basestations
calllog
customerservicerep
device
knowledgebase
mostactivestations
toparticles
webpage
websitehit
[training@localhost yarn]$ sxk230064
```

3. Upon listing the tables, I proceeded to execute "sqoop import --help" in order to understand the various arguments applicable to the sqoop import command.

```
[training@localhost yarn]$ sqoop import --help
24/02/10 17:30:35 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
usage: sqoop import [GENERIC-ARGS] [TOOL-ARGS]

Common arguments:
   --connect <jdbc-uri>                     Specify JDBC connect
                                            string
   --connection-manager <class-name>        Specify connection manager
                                            class name
   --connection-param-file <properties-file> Specify connection
                                            parameters file
   --driver <class-name>                    Manually specify JDBC
                                            driver class to use
   --hadoop-home <hdir>                     Override
                                            $HADOOP_MAPRED_HOME_ARG
   --hadoop-mapred-home <dir>               Override
                                            $HADOOP_MAPRED_HOME_ARG
   --help                                   Print usage instructions
-P                                          Read password from console
   --password <password>                    Set authentication
                                            password
   --password-alias <password-alias>        Credential provider
                                            password alias
   --password-file <password-file>          Set authentication
                                            password file path
   --relaxed-isolation                      Use read-uncommitted
                                            isolation for imports
   --skip-dist-cache                        Skip copying jars to
                                            distributed cache
   --username <username>                    Set authentication
                                            username
   --verbose                                Print more information
                                            while working

Import control arguments:
   --append                                         Imports data
                                                    in append
                                                    mode
   --as-avrodatafile                                Imports data
                                                    to Avro data
                                                    files
   --as-parquetfile                                 Imports data
                                                    to Parquet
                                                    files
   --as-sequencefile                                Imports data
                                                    to
                                                    SequenceFile
                                                    s
   --as-textfile                                    Imports data
                                                    as plain
                                                    text
                                                    (default)
   --boundary-query <statement>                     Set boundary
                                                    query for
                                                    retrieving
                                                    max and min
```

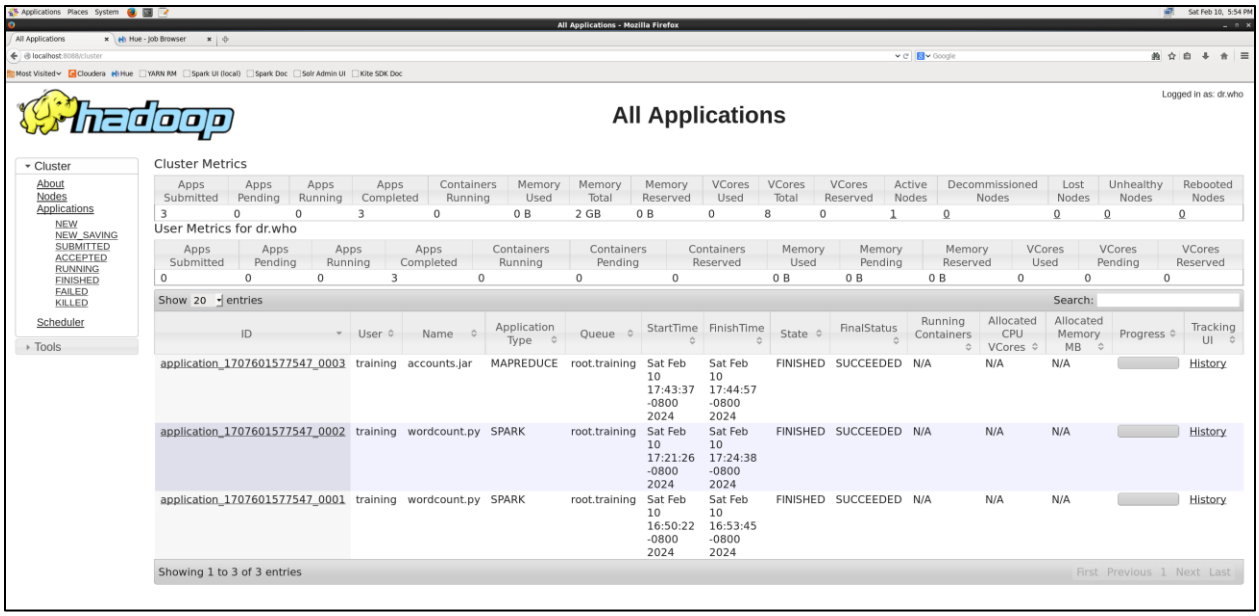
```
                                            the import
   --accumulo-create-table                  If specified, create missing
                                            Accumulo tables
   --accumulo-instance <instance>           Accumulo instance name.
   --accumulo-max-latency <latency>         Max write latency in milliseconds
   --accumulo-password <password>           Accumulo password.
   --accumulo-row-key <col>                 Specifies which input column to
                                            use as the row key
   --accumulo-table <table>                 Import to <table> in Accumulo
   --accumulo-user <user>                   Accumulo user name.
   --accumulo-visibility <vis>              Visibility token to be applied to
                                            all rows imported
   --accumulo-zookeepers <zookeepers>       Comma-separated list of
                                            zookeepers (host:port)

Code generation arguments:
   --bindir <dir>                           Output directory for compiled
                                            objects
   --class-name <name>                      Sets the generated class name.
                                            This overrides --package-name.
                                            When combined with --jar-file,
                                            sets the input class.
   --input-null-non-string <null-str>       Input null non-string
                                            representation
   --input-null-string <null-str>           Input null string representation
   --jar-file <file>                        Disable code generation; use
                                            specified jar
   --map-column-java <arg>                  Override mapping for specific
                                            columns to java types
   --null-non-string <null-str>             Null non-string representation
   --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 preceed any tool-specific arguments)
Generic options supported are
-conf <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 yarn]$ sxk230064
```

4. Subsequently, I used the "sqoop import" command to transfer the accounts table from the loudacre database into HDFS, allocating it under the /loudacre directory.

```
[training@localhost yarn]$ sqoop import \
> --connect jdbc:mysql://localhost/loudacre \
>  --username training --password training \
> --table accounts \
> --target-dir /loudacre/accounts \
>  --null-non-string '\\N'
24/02/10 17:43:20 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/02/10 17:43:20 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/02/10 17:43:21 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/02/10 17:43:21 INFO tool.CodeGenTool: Beginning code generation
24/02/10 17:43:21 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
24/02/10 17:43:22 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
24/02/10 17:43:22 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/c0c62fa27da1c59af65c06593c9a6ed7/accounts.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/02/10 17:43:28 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/c0c62fa27da1c59af65c06593c9a6ed7/accounts.jar
24/02/10 17:43:28 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/02/10 17:43:28 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/02/10 17:43:28 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/02/10 17:43:28 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/02/10 17:43:28 INFO mapreduce.ImportJobBase: Beginning import of accounts
24/02/10 17:43:28 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/02/10 17:43:29 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/02/10 17:43:31 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/02/10 17:43:31 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
24/02/10 17:43:35 INFO db.DBInputFormat: Using read commited transaction isolation
24/02/10 17:43:35 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`acct_num`), MAX(`acct_num`) FROM `accounts`
24/02/10 17:43:35 INFO mapreduce.JobSubmitter: number of splits:4
24/02/10 17:43:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1707601577547_0003
```

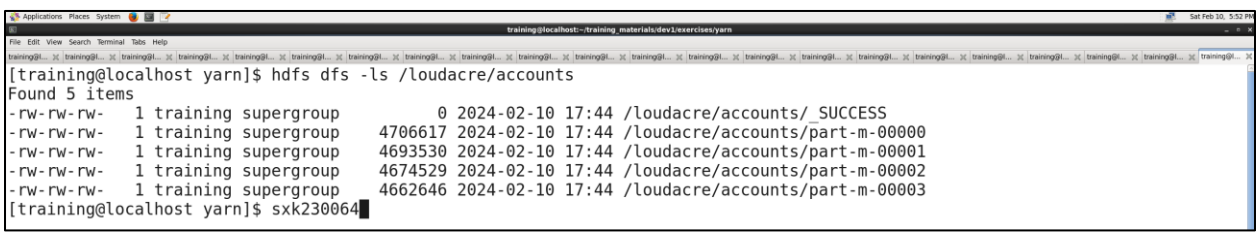
```
24/02/10 17:43:54 INFO mapreduce.Job: Job job_1707601577547_0003 running in uber mode : false
24/02/10 17:43:54 INFO mapreduce.Job:  map 0% reduce 0%
24/02/10 17:44:10 INFO mapreduce.Job:  map 25% reduce 0%
24/02/10 17:44:26 INFO mapreduce.Job:  map 50% reduce 0%
24/02/10 17:44:41 INFO mapreduce.Job:  map 75% reduce 0%
24/02/10 17:44:57 INFO mapreduce.Job:  map 100% reduce 0%
24/02/10 17:44:58 INFO mapreduce.Job: Job job_1707601577547_0003 completed successfully
24/02/10 17:44:59 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)=56367
                Total vcore-seconds taken by all map tasks=56367
                Total megabyte-seconds taken by all map tasks=14429952
        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)=902
                CPU time spent (ms)=20340
                Physical memory (bytes) snapshot=507002880
                Virtual memory (bytes) snapshot=3383017472
                Total committed heap usage (bytes)=191889408
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=18737322
24/02/10 17:44:59 INFO mapreduce.ImportJobBase: Transferred 17.8693 MB in 87.6751 seconds (208.7043 KB/sec)
24/02/10 17:44:59 INFO mapreduce.ImportJobBase: Retrieved 129761 records.
[training@localhost yarn]$ sxk230064
```

5. Upon completing the application execution, I could locate it within the YARN Web UI interface.



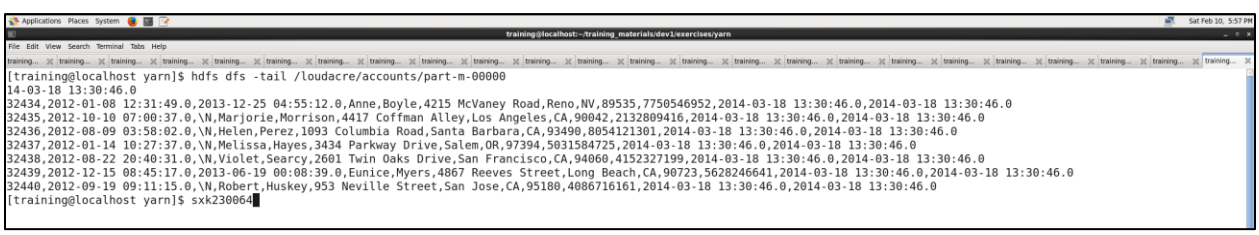
View the imported data

6. In order to check the data recently imported into HDFS, I employed the "hdfs dfs -ls" command to review the contents of the /loudacre/accounts directory. Upon execution, I observed multiple MapReduce partition files displayed in the terminal.

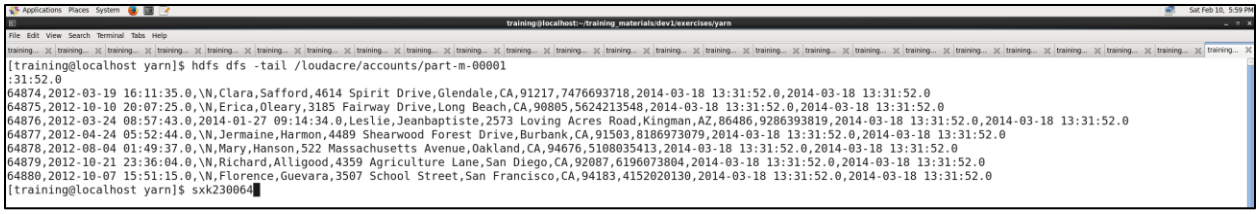


```
[training@localhost yarn]$ hdfs dfs -ls /loudacre/accounts
Found 5 items
-rw-rw-rw-   1 training supergroup          0 2024-02-10 17:44 /loudacre/accounts/_SUCCESS
-rw-rw-rw-   1 training supergroup    4706617 2024-02-10 17:44 /loudacre/accounts/part-m-00000
-rw-rw-rw-   1 training supergroup    4693530 2024-02-10 17:44 /loudacre/accounts/part-m-00001
-rw-rw-rw-   1 training supergroup    4674529 2024-02-10 17:44 /loudacre/accounts/part-m-00002
-rw-rw-rw-   1 training supergroup    4662646 2024-02-10 17:44 /loudacre/accounts/part-m-00003
[training@localhost yarn]$ sxk230064
```

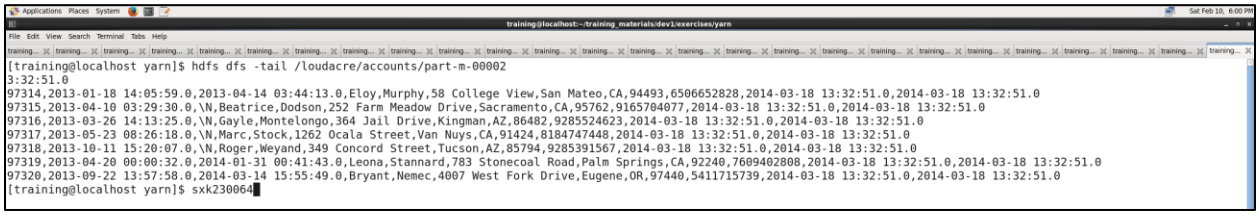
7. Then, I used the "hdfs dfs -tail" command to view the last part of each partition file.



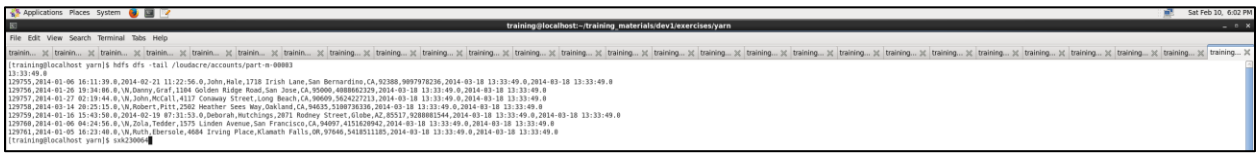
```
[training@localhost yarn]$ hdfs dfs -tail /loudacre/accounts/part-m-00000
14-03-18 13:30:46.0
32434,2012-01-08 12:31:49.0,2013-12-25 04:55:12.0,Anne,Boyle,4215 McVaney Road,Reno,NV,89535,7750546952,2014-03-18 13:30:46.0,2014-03-18 13:30:46.0
32435,2012-10-10 07:00:37.0,\N,Marjorie,Morrison,4417 Coffman 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 yarn]$ sxk230064
```



```
[training@localhost yarn]$ 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 13:31:52.0
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 yarn]$ sxk230064
```



```
[training@localhost yarn]$ 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 yarn]$ sxk230064
```

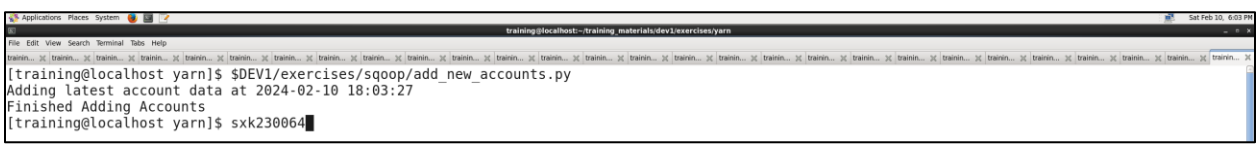


```
[training@localhost yarn]$ hdfs dfs -tail /loudacre/accounts/part-m-00003
13:33:49.0
129755,2014-01-06 16:11:39.0,2014-02-21 11:22:56.0,John,Hale,1718 Irish Lane,San Bernardino,CA,92388,9097978236,2014-03-18 13:33:49.0,2014-03-18 13:33:49.0
129756,2014-01-26 19:34:06.0,\N,Danny,Graf,1104 Golden Ridge 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,85517,9280081544,2014-03-18 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 yarn]$ sxk230064
```

8. Following that, I made note of the highest account ID, which was 129761, discerned from the initial six digits of output for each record.

Import incremental updates to accounts

9. After viewing the imported accounts data, I proceeded to import incremental updates to the accounts directory by executing the add_new_accounts.py script in the terminal.



```
[training@localhost yarn]$ $DEV1/exercises/sqoop/add_new_accounts.py
Adding latest account data at 2024-02-10 18:03:27
Finished Adding Accounts
[training@localhost yarn]$ sxk230064
```

10. To import and append the recently added accounts to the existing accounts directory, I utilized Sqoop to import based on the last value found in the acct_num column, corresponding to the largest account ID.

```
[training@localhost yarn]$ 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
24/02/10 18:16:31 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/02/10 18:16:31 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/02/10 18:16:32 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/02/10 18:16:32 INFO tool.CodeGenTool: Beginning code generation
24/02/10 18:16:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
24/02/10 18:16:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
24/02/10 18:16:33 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/1cfb6edcad8434ac8185f63ef2594880/accounts.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/02/10 18:16:39 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/1cfb6edcad8434ac8185f63ef2594880/accounts.jar
24/02/10 18:16:42 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX(`acct_num`) FROM `accounts`
24/02/10 18:16:42 INFO tool.ImportTool: Incremental import based on column `acct_num`
24/02/10 18:16:42 INFO tool.ImportTool: Lower bound value: 129761
24/02/10 18:16:42 INFO tool.ImportTool: Upper bound value: 129764
24/02/10 18:16:42 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/02/10 18:16:42 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/02/10 18:16:42 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/02/10 18:16:42 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/02/10 18:16:42 INFO mapreduce.ImportJobBase: Beginning import of accounts
24/02/10 18:16:42 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/02/10 18:16:42 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/02/10 18:16:42 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/02/10 18:16:42 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
24/02/10 18:16:46 INFO db.DBInputFormat: Using read commited transaction isolation
24/02/10 18:16:46 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`acct_num`), MAX(`acct_num`) FROM `accounts` WHERE ( `acct_num` > 129761 AND `acct_num` <= 129764 )
24/02/10 18:16:46 INFO mapreduce.JobSubmitter: number of splits:3
24/02/10 18:16:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1707601577547_0004
24/02/10 18:16:47 INFO impl.YarnClientImpl: Submitted application application_1707601577547_0004
24/02/10 18:16:47 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1707601577547_0004/
24/02/10 18:16:47 INFO mapreduce.Job: Running job: job_1707601577547_0004
24/02/10 18:17:04 INFO mapreduce.Job: Job job_1707601577547_0004 running in uber mode : false
24/02/10 18:17:04 INFO mapreduce.Job:  map 0% reduce 0%
24/02/10 18:17:15 INFO mapreduce.Job:  map 33% reduce 0%
24/02/10 18:17:27 INFO mapreduce.Job:  map 67% reduce 0%
24/02/10 18:17:39 INFO mapreduce.Job:  map 100% reduce 0%
24/02/10 18:17:39 INFO mapreduce.Job: Job job_1707601577547_0004 completed successfully
24/02/10 18:17:40 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=410586
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=361
                HDFS: Number of bytes written=391
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
```

```
24/02/10 18:16:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1707601577547_0004
24/02/10 18:16:47 INFO impl.YarnClientImpl: Submitted application application_1707601577547_0004
24/02/10 18:16:47 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1707601577547_0004/
24/02/10 18:16:47 INFO mapreduce.Job: Running job: job_1707601577547_0004
24/02/10 18:17:04 INFO mapreduce.Job: Job job_1707601577547_0004 running in uber mode : false
24/02/10 18:17:04 INFO mapreduce.Job:  map 0% reduce 0%
24/02/10 18:17:15 INFO mapreduce.Job:  map 33% reduce 0%
24/02/10 18:17:27 INFO mapreduce.Job:  map 67% reduce 0%
24/02/10 18:17:39 INFO mapreduce.Job:  map 100% reduce 0%
24/02/10 18:17:39 INFO mapreduce.Job: Job job_1707601577547_0004 completed successfully
24/02/10 18:17:40 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=410586
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=361
                HDFS: Number of bytes written=391
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=6
        Job Counters
                Launched map tasks=3
                Other local map tasks=3
                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)=30462
                Total vcore-seconds taken by all map tasks=30462
                Total megabyte-seconds taken by all map tasks=7798272
        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)=554
                CPU time spent (ms)=4530
                Physical memory (bytes) snapshot=380145664
                Virtual memory (bytes) snapshot=2533539840
                Total committed heap usage (bytes)=143917056
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=391
24/02/10 18:17:40 INFO mapreduce.ImportJobBase: Transferred 391 bytes in 58.0865 seconds (6.7313 bytes/sec)
24/02/10 18:17:40 INFO mapreduce.ImportJobBase: Retrieved 3 records.
24/02/10 18:17:40 INFO util.AppendUtils: Appending to directory accounts
24/02/10 18:17:40 INFO util.AppendUtils: Using found partition 4
24/02/10 18:17:40 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments:
24/02/10 18:17:40 INFO tool.ImportTool:  --incremental append
24/02/10 18:17:40 INFO tool.ImportTool:   --check-column acct_num
24/02/10 18:17:40 INFO tool.ImportTool:   --last-value 129764
24/02/10 18:17:40 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')
[training@localhost yarn]$ ssk230064
```

11. To verify the Sqoop import, I used the "hdfs dfs -ls" command to list the contents of the accounts directory.

```
[training@localhost yarn]$ hdfs dfs -ls /loudacre/accounts
Found 8 items
-rw-rw-rw-   1 training supergroup          0 2024-02-10 17:44 /loudacre/accounts/_SUCCESS
-rw-rw-rw-   1 training supergroup    4706617 2024-02-10 17:44 /loudacre/accounts/part-m-00000
-rw-rw-rw-   1 training supergroup    4693530 2024-02-10 17:44 /loudacre/accounts/part-m-00001
-rw-rw-rw-   1 training supergroup    4674529 2024-02-10 17:44 /loudacre/accounts/part-m-00002
-rw-rw-rw-   1 training supergroup    4662646 2024-02-10 17:44 /loudacre/accounts/part-m-00003
-rw-rw-rw-   1 training supergroup        129 2024-02-10 18:17 /loudacre/accounts/part-m-00004
-rw-rw-rw-   1 training supergroup        131 2024-02-10 18:17 /loudacre/accounts/part-m-00005
-rw-rw-rw-   1 training supergroup        131 2024-02-10 18:17 /loudacre/accounts/part-m-00006
[training@localhost yarn]$ sxk230064
```

12. I ran Hadoop's "cat" command to display the complete contents of these files.

```
[training@localhost yarn]$ hdfs dfs -cat /loudacre/accounts/part-m-0000[456]
129762,2014-03-14 22:41:27.0,\N,Jesse,Anderson,123 Elm St,Reno,NV,89511,775-555-1212,2024-02-10 18:03:27.0,2024-02-10 18:03:27.0
129763,2014-03-14 23:52:03.0,\N,Tom,Wheeler,125 Elm St,Palo Alto,CA,94301,650-555-1515,2024-02-10 18:03:27.0,2024-02-10 18:03:27.0
129764,2014-03-14 15:36:50.0,\N,Ian,Wrigley,127 Elm St,Palo Alto,CA,94301,650-555-1777,2024-02-10 18:03:27.0,2024-02-10 18:03:27.0
[training@localhost yarn]$ sxk230064
```

Import webpage data using an alternate field delimiter

13. We additionally aim to transfer the webpage table to HDFS. However, before proceeding, examine a few records in that table utilizing the sqoop eval command.

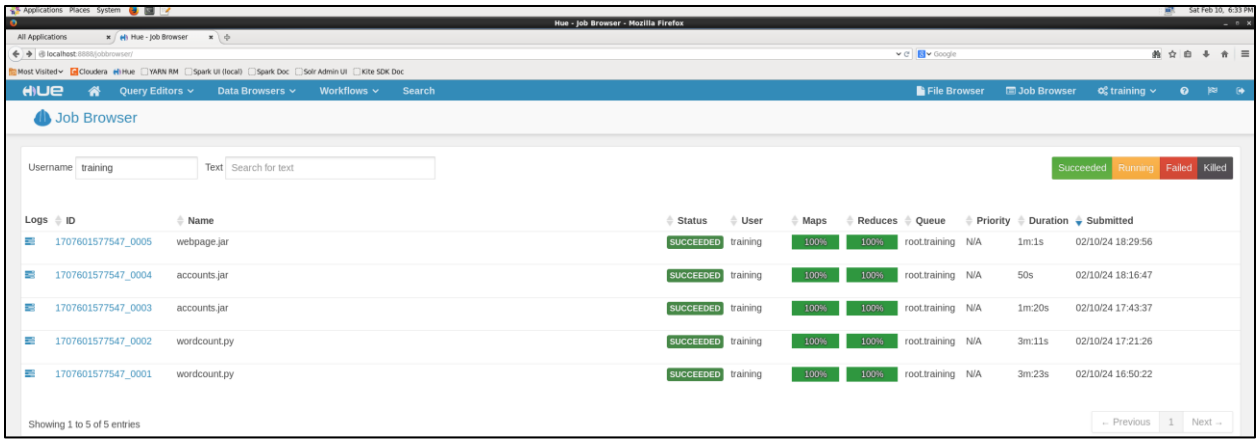
```
[training@localhost yarn]$ sqoop eval \
> --query "SELECT * FROM webpage LIMIT 10" \
> --connect jdbc:mysql://localhost/loudacre \
> --username training --password training
24/02/10 18:27:26 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/02/10 18:27:26 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/02/10 18:27:27 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 yarn]$ sxk230064
```

14. To import the webpage table using sqoop, specify the tab character (\t) as the field terminator instead of the default comma.

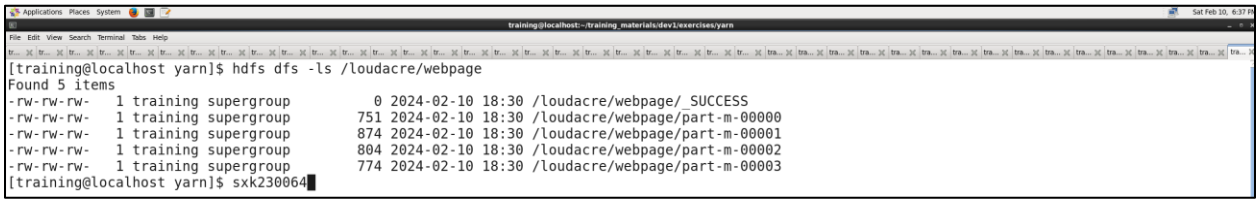
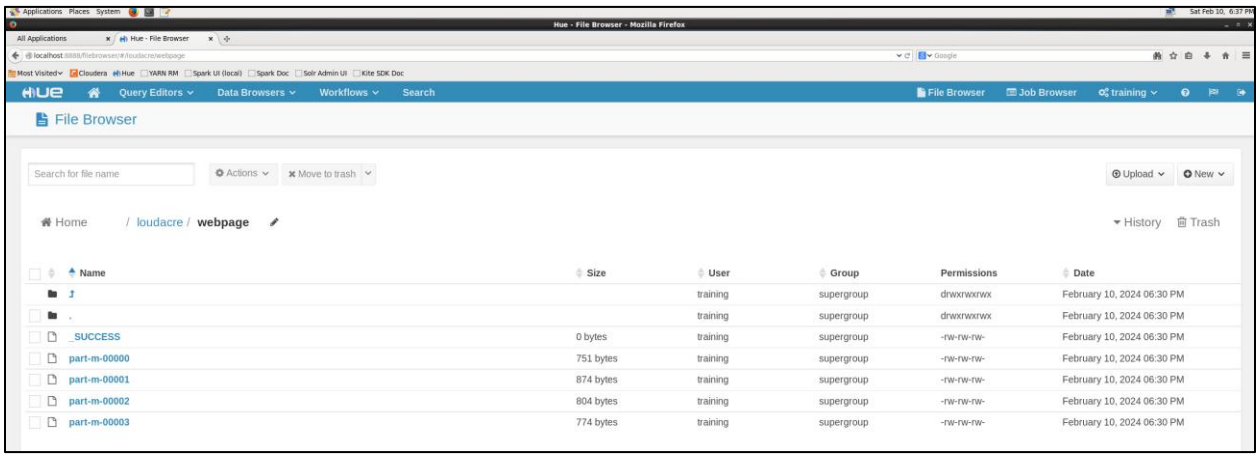
```
[training@localhost yarn]$ sqoop import \
> --connect jdbc:mysql://localhost/loudacre \
> --username training --password training \
> --table webpage \
> --target-dir /loudacre/webpage \
> --fields-terminated-by "\t"
24/02/10 18:29:41 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/02/10 18:29:41 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/02/10 18:29:42 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/02/10 18:29:42 INFO tool.CodeGenTool: Beginning code generation
24/02/10 18:29:42 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `webpage` AS t LIMIT 1
24/02/10 18:29:42 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `webpage` AS t LIMIT 1
24/02/10 18:29:42 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/7236c7f3a6a6f94b3c723da27cf5e316/webpage.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/02/10 18:29:48 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/7236c7f3a6a6f94b3c723da27cf5e316/webpage.jar
24/02/10 18:29:48 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/02/10 18:29:48 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/02/10 18:29:48 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/02/10 18:29:48 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/02/10 18:29:48 INFO mapreduce.ImportJobBase: Beginning import of webpage
24/02/10 18:29:48 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/02/10 18:29:49 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/02/10 18:29:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/02/10 18:29:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
24/02/10 18:29:54 INFO db.DBInputFormat: Using read commited transaction isolation
24/02/10 18:29:54 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`web_page_num`), MAX(`web_page_num`) FROM `webpage`
24/02/10 18:29:54 INFO mapreduce.JobSubmitter: number of splits:4
24/02/10 18:29:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1707601577547_0005
24/02/10 18:29:56 INFO impl.YarnClientImpl: Submitted application application_1707601577547_0005
24/02/10 18:29:56 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1707601577547_0005/
24/02/10 18:29:56 INFO mapreduce.Job: Running job: job_1707601577547_0005
24/02/10 18:30:12 INFO mapreduce.Job: Job job_1707601577547_0005 running in uber mode : false
24/02/10 18:30:12 INFO mapreduce.Job:  map 0% reduce 0%
24/02/10 18:30:25 INFO mapreduce.Job:  map 25% reduce 0%
24/02/10 18:30:36 INFO mapreduce.Job:  map 50% reduce 0%
24/02/10 18:30:47 INFO mapreduce.Job:  map 75% reduce 0%
24/02/10 18:30:58 INFO mapreduce.Job:  map 100% reduce 0%
24/02/10 18:30:58 INFO mapreduce.Job: Job job_1707601577547_0005 completed successfully
24/02/10 18:30:58 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
```

```
24/02/10 18:30:12 INFO mapreduce.Job: Job job_1707601577547_0005 running in uber mode : false
24/02/10 18:30:12 INFO mapreduce.Job:  map 0% reduce 0%
24/02/10 18:30:25 INFO mapreduce.Job:  map 25% reduce 0%
24/02/10 18:30:36 INFO mapreduce.Job:  map 50% reduce 0%
24/02/10 18:30:47 INFO mapreduce.Job:  map 75% reduce 0%
24/02/10 18:30:58 INFO mapreduce.Job:  map 100% reduce 0%
24/02/10 18:30:58 INFO mapreduce.Job: Job job_1707601577547_0005 completed successfully
24/02/10 18:30:58 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)=39986
                Total vcore-seconds taken by all map tasks=39986
                Total megabyte-seconds taken by all map tasks=10236416
        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)=754
                CPU time spent (ms)=5850
                Physical memory (bytes) snapshot=513482752
                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
24/02/10 18:30:58 INFO mapreduce.ImportJobBase: Transferred 3.1279 KB in 67.1389 seconds (47.7071 bytes/sec)
24/02/10 18:30:58 INFO mapreduce.ImportJobBase: Retrieved 53 records.
[training@localhost yarn]$ sxk230064
```

15. I used Hue to inspect the data files imported into the loudacre directory.



16. I also employed the hdfs command line to check the same data files that were imported into the /loudacre/webpage directory, and I took note of the structure of the data.



```
[training@localhost yarn]$ hdfs dfs -ls /loudacre/webpage
Found 5 items
-rw-rw-rw-   1 training supergroup          0 2024-02-10 18:30 /loudacre/webpage/_SUCCESS
-rw-rw-rw-   1 training supergroup        751 2024-02-10 18:30 /loudacre/webpage/part-m-00000
-rw-rw-rw-   1 training supergroup        874 2024-02-10 18:30 /loudacre/webpage/part-m-00001
-rw-rw-rw-   1 training supergroup        804 2024-02-10 18:30 /loudacre/webpage/part-m-00002
-rw-rw-rw-   1 training supergroup        774 2024-02-10 18:30 /loudacre/webpage/part-m-00003
[training@localhost yarn]$ sxk230064
```