Assignment 2

BUAN 6346

Big Data Analytics

Dr. Jerry F. Perez

Yug Jyotirmay Singh
YJS220000

Table of Contents

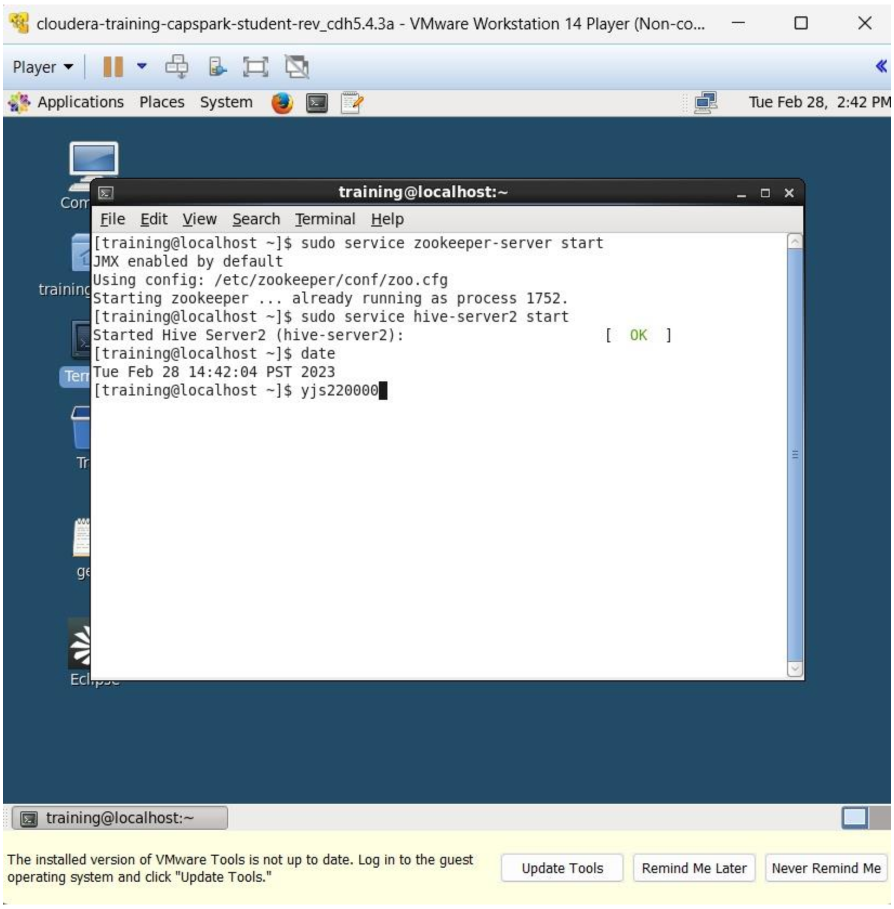
Chapter 6

We know that there are a variety of ways to interact with Impala and Hive. In this particular exercise, you will use the Impala or Hive Query Editor in Hue.

1. If you plan to use Hive rather than Impala for this or subsequent exercises, start the Hive server, which is not started by default, by entering the following two commands in a terminal window:

```
$ sudo service zookeeper-server start
```

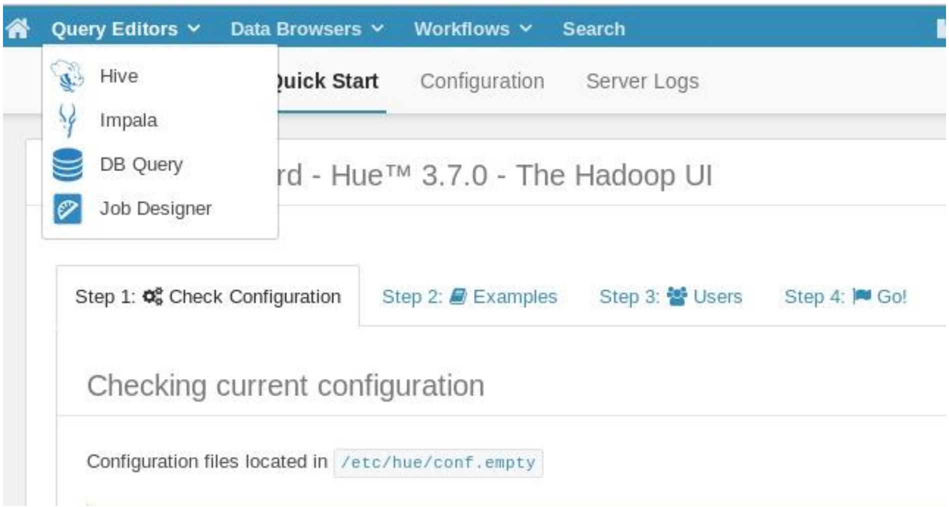
```
$ sudo service hive-server2 start
```


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-co...
Player
Applications Places System
Tue Feb 28, 2:42 PM
training@localhost:~
File Edit View Search Terminal Help
training@localhost:~
The installed version of VMware Tools is not up to date. Log in to the guest operating system and click "Update Tools."
Update Tools
Remind Me Later
Never Remind Me

```
[training@localhost ~]$ sudo service zookeeper-server start
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Starting zookeeper ... already running as process 1752.
[training@localhost ~]$ sudo service hive-server2 start
Started Hive Server2 (hive-server2):                       [  OK  ]
[training@localhost ~]$ date
Tue Feb 28 14:42:04 PST 2023
[training@localhost ~]$ yjs220000
```

2. Now, visit the Hue page in firefox, as described earlier in the “Using HDFS” exercise.

3. Further, open the Impala query editor or Hive query editor, by selecting the editor of your choice from the Query Editors menu.

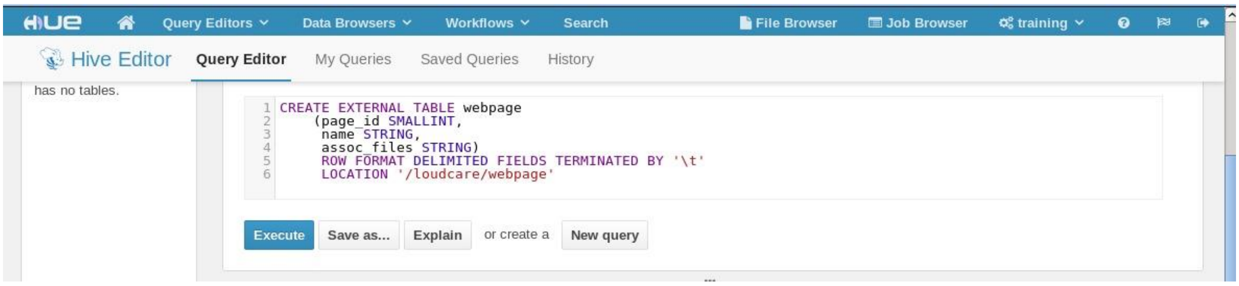


Checking current configuration

Configuration files located in `/etc/hue/conf.empty`

4. In the query editor page on the right hand side, enter a SQL command to create a table for the webpage data imported in the previous exercises:

```
CREATE EXTERNAL TABLE webpage

        (page_id SMALLINT,

        name STRING

        assoc_files STRING)

        ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘\t’

        LOCATION ‘/loudacre/webpage’
```



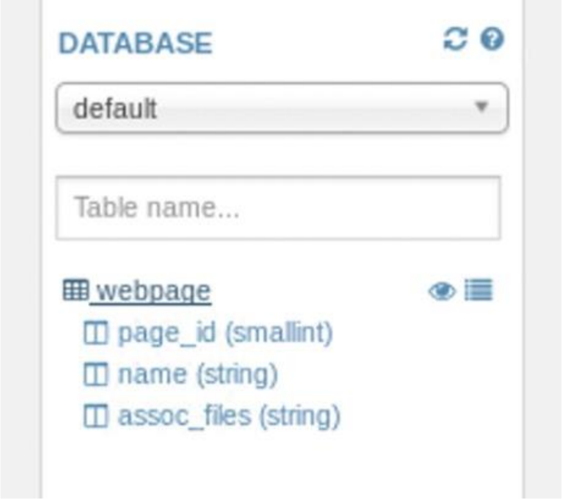
has no tables.

```
CREATE EXTERNAL TABLE webpage
     (page_id SMALLINT,
      name STRING,
      assoc_files STRING)
      ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
      LOCATION '/loudcare/webpage'
```

5. Click the EXECUTE button to execute the command.

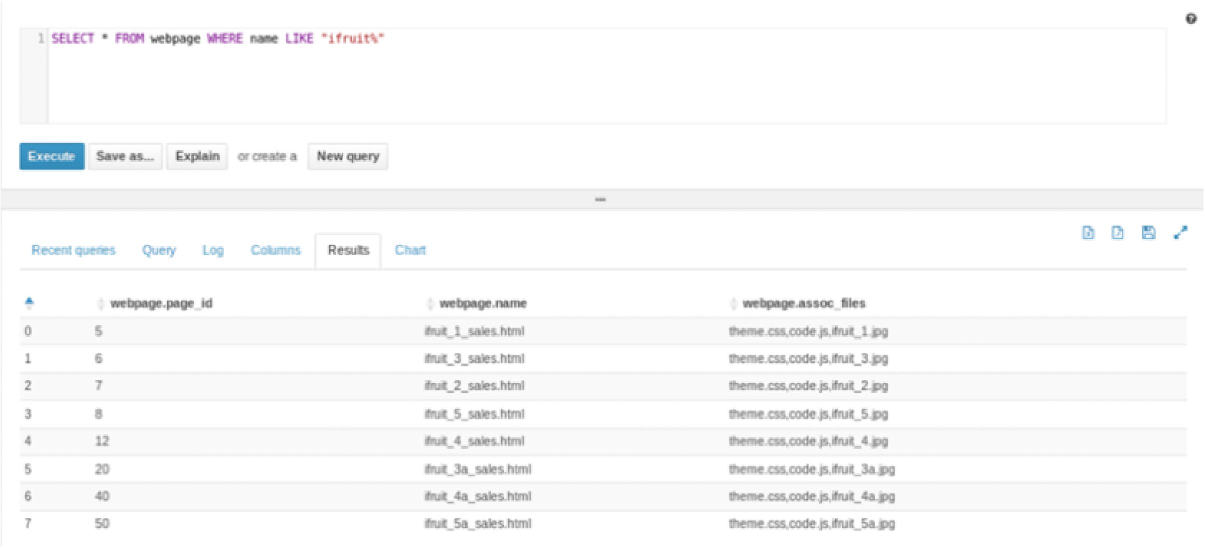
6. To see the table you just created, refresh the table list on the left hand side

7. Now click on the webpage table to see the column definitions



8. Now click the New Query button, then further enter and execute a test query such as:

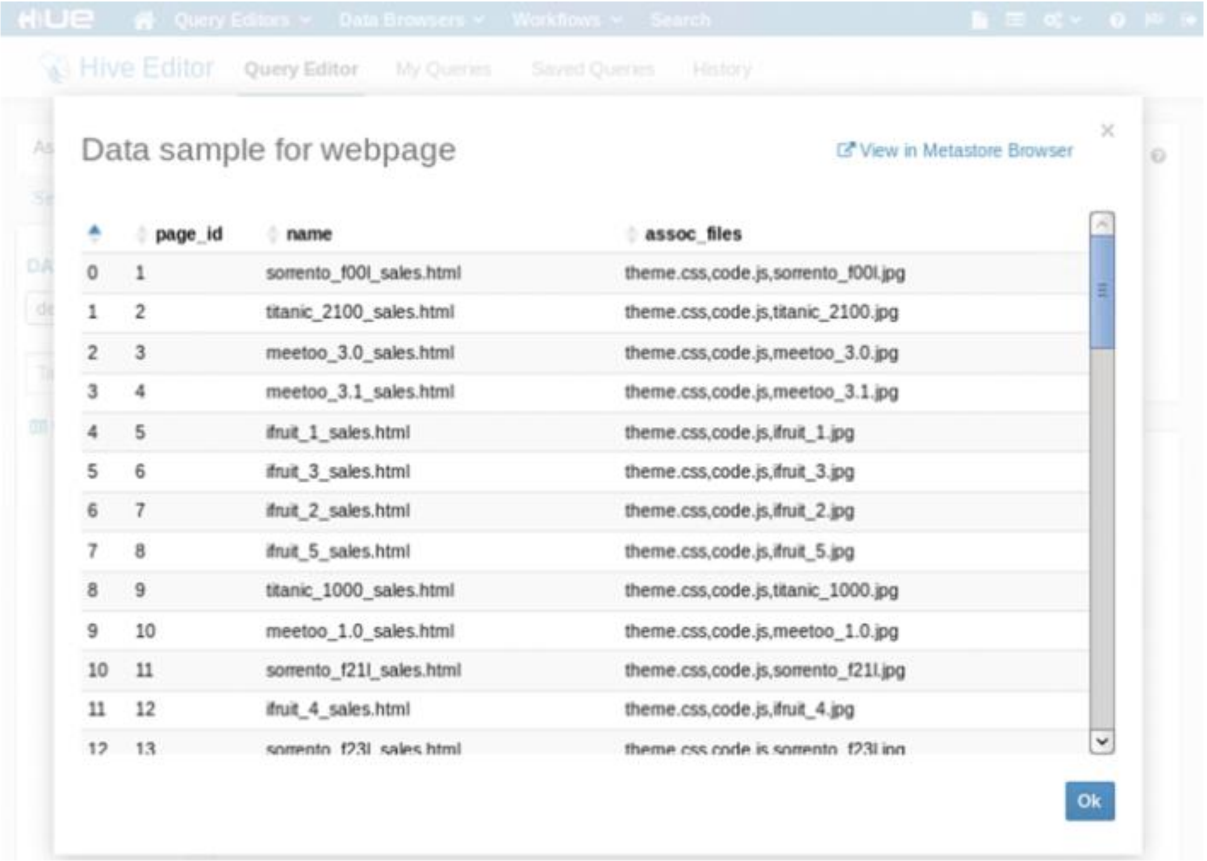
SELECT * FROM webpage WHERE name LIKE “ifruit%”



```
SELECT * FROM webpage WHERE name LIKE "ifruit%"
```

	webpage.page_id	webpage.name	webpage.assoc_files
0	5	ifruit_1_sales.html	theme.css,code.js,ifruit_1.jpg
1	6	ifruit_3_sales.html	theme.css,code.js,ifruit_3.jpg
2	7	ifruit_2_sales.html	theme.css,code.js,ifruit_2.jpg
3	8	ifruit_5_sales.html	theme.css,code.js,ifruit_5.jpg
4	12	ifruit_4_sales.html	theme.css,code.js,ifruit_4.jpg
5	20	ifruit_3a_sales.html	theme.css,code.js,ifruit_3a.jpg
6	40	ifruit_4a_sales.html	theme.css,code.js,ifruit_4a.jpg
7	50	ifruit_5a_sales.html	theme.css,code.js,ifruit_5a.jpg

9. Click on the Preview Sample Data icon to view a sampling of the table data.



Data sample for webpage

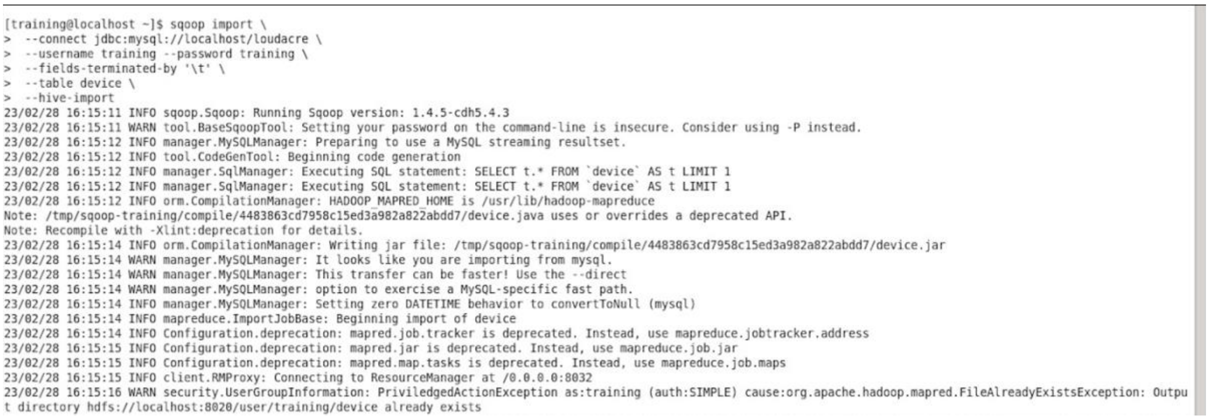
	page_id	name	assoc_files
0	1	sorrento_f00l_sales.html	theme.css,code.js,sorrento_f00l.jpg
1	2	titanic_2100_sales.html	theme.css,code.js,titanic_2100.jpg
2	3	meetoo_3.0_sales.html	theme.css,code.js,meetoo_3.0.jpg
3	4	meetoo_3.1_sales.html	theme.css,code.js,meetoo_3.1.jpg
4	5	ifruit_1_sales.html	theme.css,code.js,ifruit_1.jpg
5	6	ifruit_3_sales.html	theme.css,code.js,ifruit_3.jpg
6	7	ifruit_2_sales.html	theme.css,code.js,ifruit_2.jpg
7	8	ifruit_5_sales.html	theme.css,code.js,ifruit_5.jpg
8	9	titanic_1000_sales.html	theme.css,code.js,titanic_1000.jpg
9	10	meetoo_1.0_sales.html	theme.css,code.js,meetoo_1.0.jpg
10	11	sorrento_f21l_sales.html	theme.css,code.js,sorrento_f21l.jpg
11	12	ifruit_4_sales.html	theme.css,code.js,ifruit_4.jpg
12	13	sorrento_f23l_sales.html	theme.css,code.js,sorrento_f23l.jpg

6.1 Use Sqoop to Import Directly into Hive and Impala

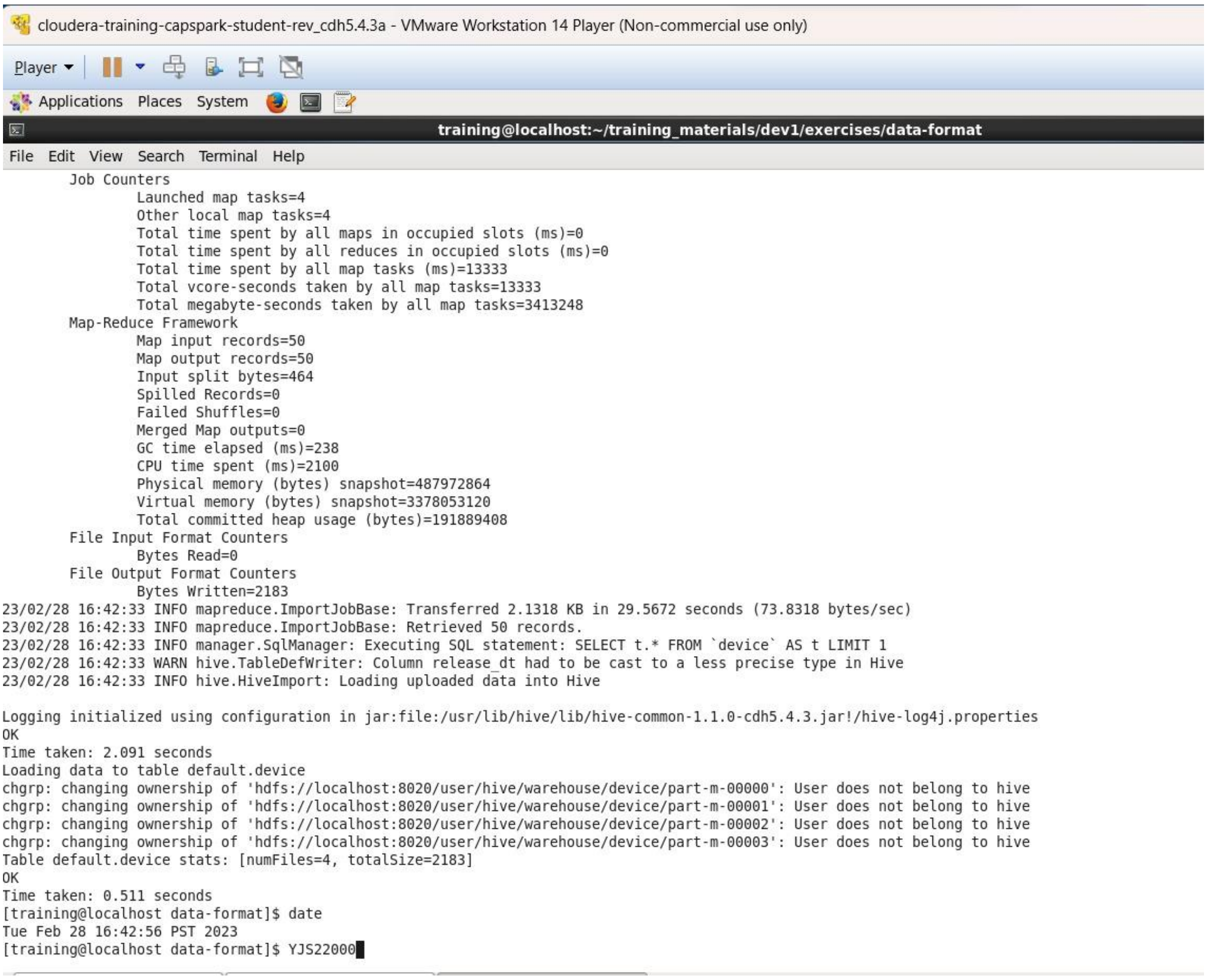
In this particular section, we will use Sqoop to import data from MySQL into HDFS and also automatically create the corresponding table in the Hive Metastore.

10. Now in the terminal window, import the device table directly into the Hive Metastore.

```
$ sqoop import \

--connect jdbc:mysql://localhost/loudacre \

--username training –password training \

--fields-terminated-by ‘\t’ \

--table device \

--hive-import
```



```
[training@localhost ~]$ sqoop import \
>  --connect jdbc:mysql://localhost/loudacre \
>  --username training --password training \
>  --fields-terminated-by '\t' \
>  --table device \
>  --hive-import
23/02/28 16:15:11 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
23/02/28 16:15:11 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/02/28 16:15:12 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/02/28 16:15:12 INFO tool.CodeGenTool: Beginning code generation
23/02/28 16:15:12 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `device` AS t LIMIT 1
23/02/28 16:15:12 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `device` AS t LIMIT 1
23/02/28 16:15:12 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/4483863cd7958c15ed3a982a822abdd7/device.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/02/28 16:15:14 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/4483863cd7958c15ed3a982a822abdd7/device.jar
23/02/28 16:15:14 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/02/28 16:15:14 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/02/28 16:15:14 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/02/28 16:15:14 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/02/28 16:15:14 INFO mapreduce.ImportJobBase: Beginning import of device
23/02/28 16:15:14 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/02/28 16:15:15 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/02/28 16:15:15 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/02/28 16:15:15 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/02/28 16:15:16 WARN security.UserGroupInformation: PriviledgedActionException as:training (auth:SIMPLE) cause:org.apache.hadoop.mapred.FileAlreadyExistsException: Outpu
t directory hdfs://localhost:8020/user/training/device already exists
```


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications
Places
System
training@localhost:~/training_materials/dev1/exercises/data-format
File
Edit
View
Search
Terminal
Help

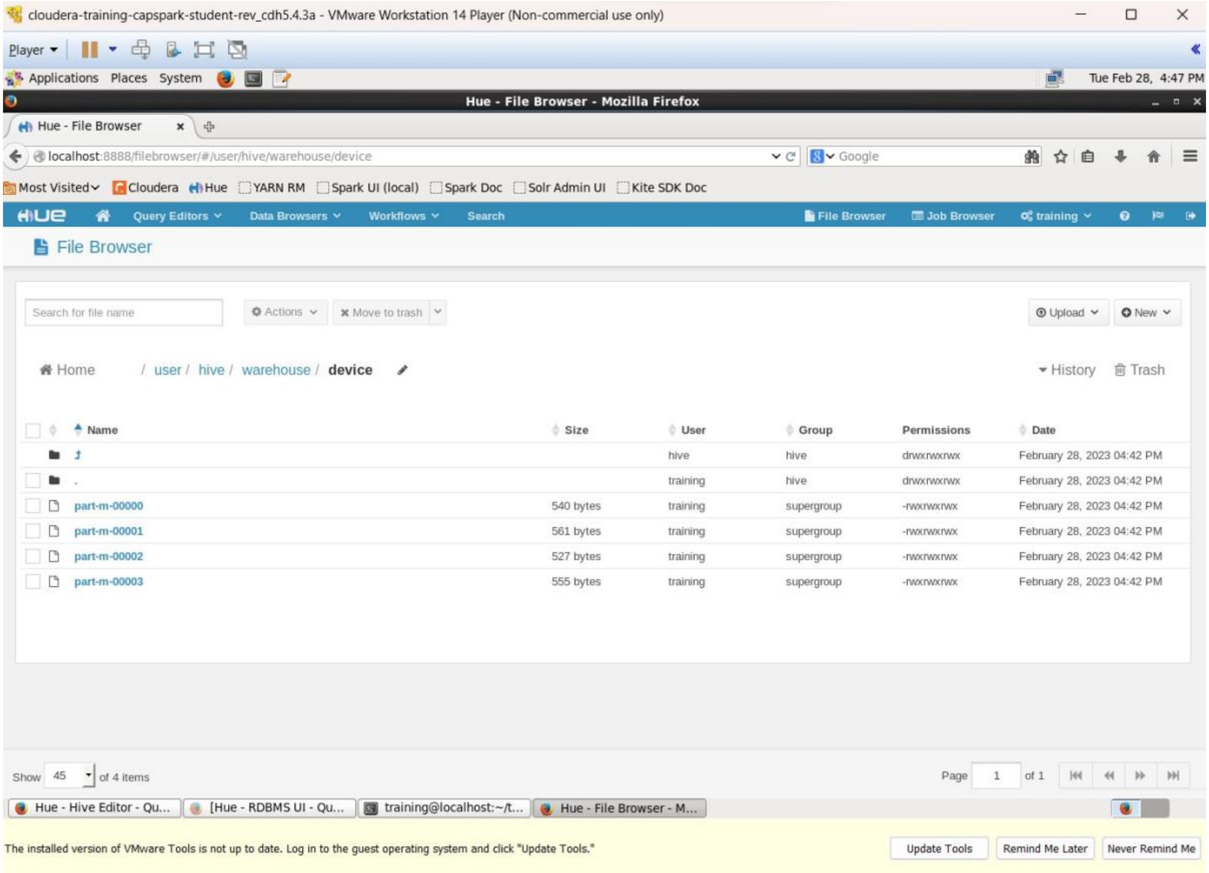
```
        Job Counters 
                Launched map tasks=4
                Other local map tasks=4
                Total time spent by all maps in occupied slots (ms)=0
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=13333
                Total vcore-seconds taken by all map tasks=13333
                Total megabyte-seconds taken by all map tasks=3413248
        Map-Reduce Framework
                Map input records=50
                Map output records=50
                Input split bytes=464
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=238
                CPU time spent (ms)=2100
                Physical memory (bytes) snapshot=487972864
                Virtual memory (bytes) snapshot=3378053120
                Total committed heap usage (bytes)=191889408
        File Input Format Counters 
                Bytes Read=0
        File Output Format Counters 
                Bytes Written=2183
23/02/28 16:42:33 INFO mapreduce.ImportJobBase: Transferred 2.1318 KB in 29.5672 seconds (73.8318 bytes/sec)
23/02/28 16:42:33 INFO mapreduce.ImportJobBase: Retrieved 50 records.
23/02/28 16:42:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `device` AS t LIMIT 1
23/02/28 16:42:33 WARN hive.TableDefWriter: Column release_dt had to be cast to a less precise type in Hive
23/02/28 16:42:33 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.4.3.jar!/hive-log4j.properties
OK
Time taken: 2.091 seconds
Loading data to table default.device
chgrp: changing ownership of 'hdfs://localhost:8020/user/hive/warehouse/device/part-m-00000': User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8020/user/hive/warehouse/device/part-m-00001': User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8020/user/hive/warehouse/device/part-m-00002': User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8020/user/hive/warehouse/device/part-m-00003': User does not belong to hive
Table default.device stats: [numFiles=4, totalSize=2183]
OK
Time taken: 0.511 seconds
[training@localhost data-format]$ date
Tue Feb 28 16:42:56 PST 2023
[training@localhost data-format]$ YJS22000
```

Use –hive-import for either Impala or Hive; this adds metadata to the Metastore, which both tools use.

Note: There might be a possibility that you may get a warning message that chgrp is unable to change the ownership of the generated files; you can disregard the warning, it does not affect the import.

11. Using Hue or the HDFS command line, review the imported data files. The Hive import copies the data to the default Hive warehouse location:

/user/hive/warehouse/device



	Name	Size	User	Group	Permissions	Date
	↥		hive	hive	drwxrwxrwx	February 28, 2023 04:42 PM
□	.		training	hive	drwxrwxrwx	February 28, 2023 04:42 PM
□	part-m-00000	540 bytes	training	supergroup	-rwxrwxrwx	February 28, 2023 04:42 PM
□	part-m-00001	561 bytes	training	supergroup	-rwxrwxrwx	February 28, 2023 04:42 PM
□	part-m-00002	527 bytes	training	supergroup	-rwxrwxrwx	February 28, 2023 04:42 PM
□	part-m-00003	555 bytes	training	supergroup	-rwxrwxrwx	February 28, 2023 04:42 PM

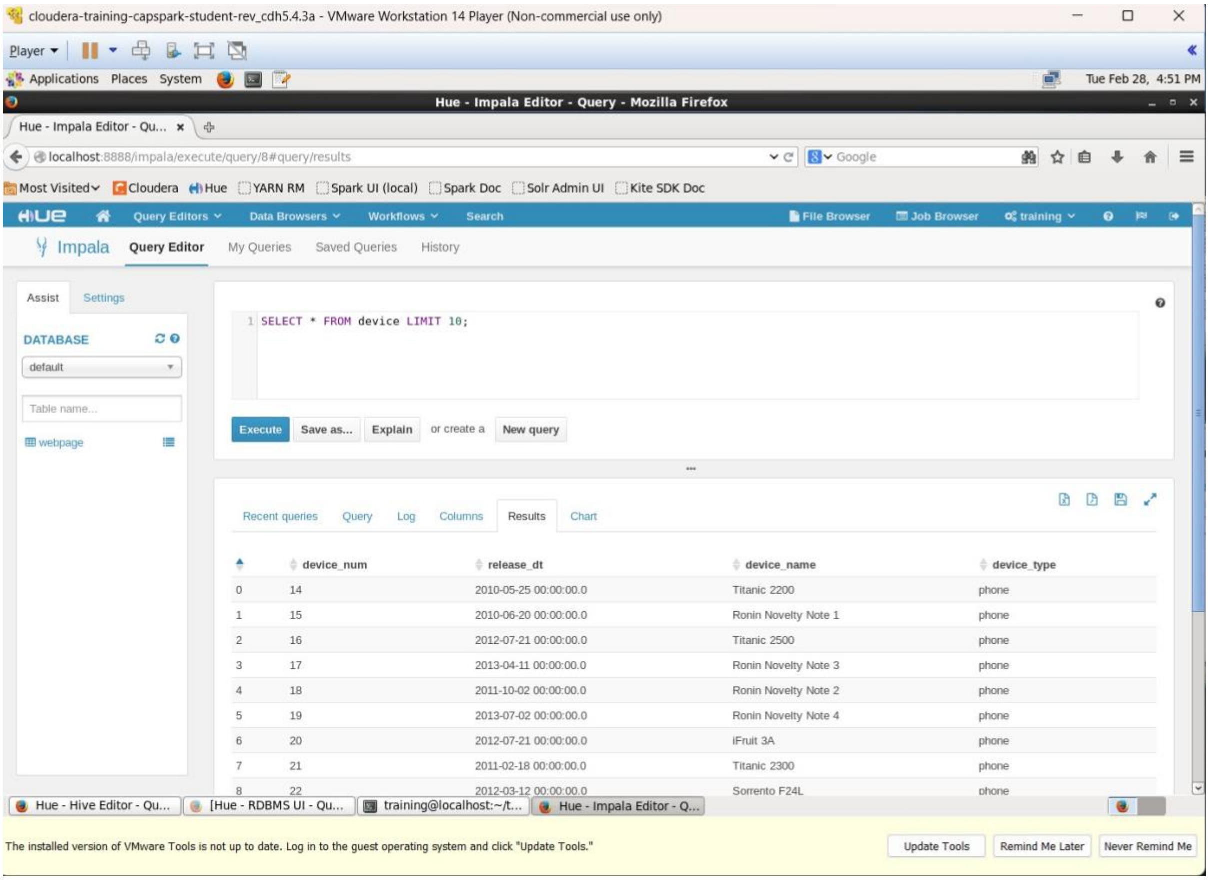
The installed version of VMware Tools is not up to date. Log in to the guest operating system and click "Update Tools."

12. If you are using Impala, make sure you refresh the Impala metadata cache by entering the command in the Hue Impala Query Editor:

```
INVALIDATE METADATA
```

13. As in the previous exercise given, view the columns and execute a test query:

```
SELECT * FROM device LIMIT 10;
```



cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Hue - Impala Editor - Query - Mozilla Firefox
localhost:8888/impala/execute/query/8#query/results
Impala
Query Editor
My Queries
Saved Queries
History
DATABASE
default
webpage
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart

```
SELECT * FROM device LIMIT 10;
```

	device_num	release_dt	device_name	device_type
0	14	2010-05-25 00:00:00.0	Titanic 2200	phone
1	15	2010-06-20 00:00:00.0	Ronin Novelty Note 1	phone
2	16	2012-07-21 00:00:00.0	Titanic 2500	phone
3	17	2013-04-11 00:00:00.0	Ronin Novelty Note 3	phone
4	18	2011-10-02 00:00:00.0	Ronin Novelty Note 2	phone
5	19	2013-07-02 00:00:00.0	Ronin Novelty Note 4	phone
6	20	2012-07-21 00:00:00.0	iFruit 3A	phone
7	21	2011-02-18 00:00:00.0	Titanic 2300	phone
8	22	2012-03-12 00:00:00.0	Sorrento F24L	phone

The installed version of VMware Tools is not up to date. Log in to the guest operating system and click "Update Tools."

Chapter 7

Talking about this exercise, you will you use import data in Avro format and create an Impala/Hive table to access it.

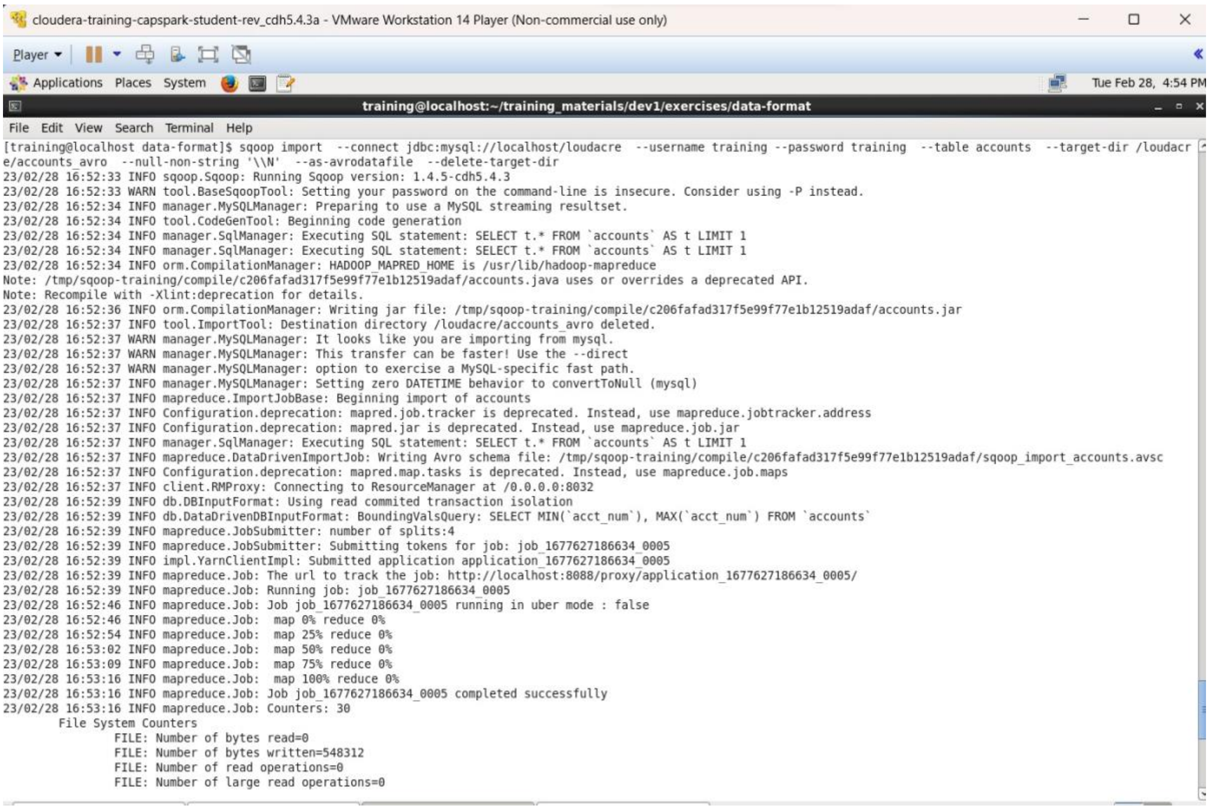
1. Change directories to the exercise directory:

   ```
   $ cd $DEV1/exercises/data-format/
   ```

2. Import the accounts table to an Avro data format.

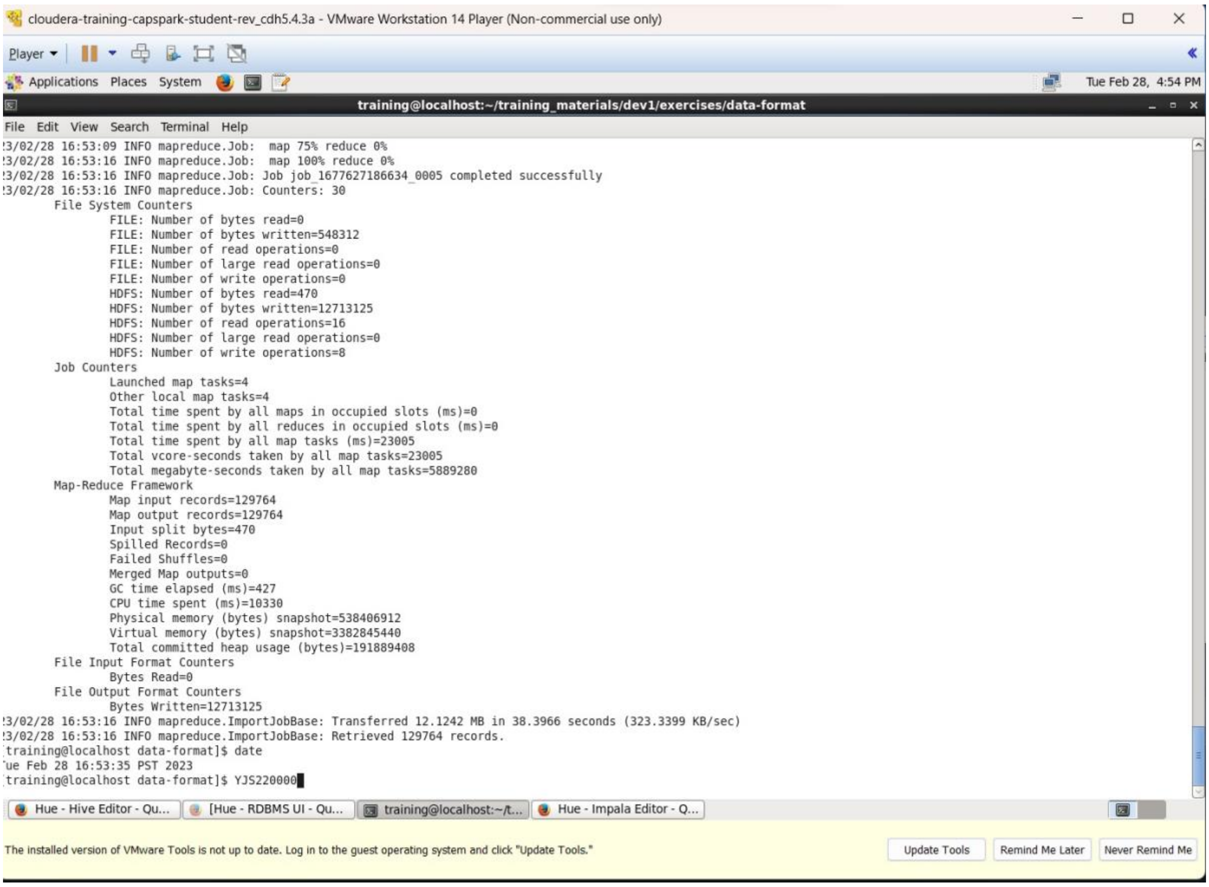
```
$ sqoop import \
--connect jdbc:mysql://localhost/loudacre \
--username training --password training \
--table accounts \
--target-dir /loudacre/accounts_avro \
--null-non-string '\\N' \
```

--as-avrodatafile



cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System
Tue Feb 28, 4:54 PM
training@localhost:~/training_materials/dev1/exercises/data-format
File Edit View Search Terminal Help

```
[training@localhost data-format]$ sqoop import  --connect jdbc:mysql://localhost/loudacre  --username training --password training  --table accounts  --target-dir /loudacr
e/accounts_avro  --null-non-string '\\N'  --as-avrodatafile  --delete-target-dir
23/02/28 16:52:33 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
23/02/28 16:52:33 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/02/28 16:52:34 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/02/28 16:52:34 INFO tool.CodeGenTool: Beginning code generation
23/02/28 16:52:34 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
23/02/28 16:52:34 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
23/02/28 16:52:34 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/c206fafad317f5e99f77e1b12519adaf/accounts.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/02/28 16:52:36 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/c206fafad317f5e99f77e1b12519adaf/accounts.jar
23/02/28 16:52:37 INFO tool.ImportTool: Destination directory /loudacre/accounts_avro deleted.
23/02/28 16:52:37 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/02/28 16:52:37 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/02/28 16:52:37 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/02/28 16:52:37 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/02/28 16:52:37 INFO mapreduce.ImportJobBase: Beginning import of accounts
23/02/28 16:52:37 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/02/28 16:52:37 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/02/28 16:52:37 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
23/02/28 16:52:37 INFO mapreduce.DataDrivenImportJob: Writing Avro schema file: /tmp/sqoop-training/compile/c206fafad317f5e99f77e1b12519adaf/sqoop_import_accounts.avsc
23/02/28 16:52:37 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/02/28 16:52:37 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/02/28 16:52:39 INFO db.DBInputFormat: Using read commited transaction isolation
23/02/28 16:52:39 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`acct_num`), MAX(`acct_num`) FROM `accounts`
23/02/28 16:52:39 INFO mapreduce.JobSubmitter: number of splits:4
23/02/28 16:52:39 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1677627186634_0005
23/02/28 16:52:39 INFO impl.YarnClientImpl: Submitted application application_1677627186634_0005
23/02/28 16:52:39 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1677627186634_0005/
23/02/28 16:52:39 INFO mapreduce.Job: Running job: job_1677627186634_0005
23/02/28 16:52:46 INFO mapreduce.Job: Job job_1677627186634_0005 running in uber mode : false
23/02/28 16:52:46 INFO mapreduce.Job:  map 0% reduce 0%
23/02/28 16:52:54 INFO mapreduce.Job:  map 25% reduce 0%
23/02/28 16:53:02 INFO mapreduce.Job:  map 50% reduce 0%
23/02/28 16:53:09 INFO mapreduce.Job:  map 75% reduce 0%
23/02/28 16:53:16 INFO mapreduce.Job:  map 100% reduce 0%
23/02/28 16:53:16 INFO mapreduce.Job: Job job_1677627186634_0005 completed successfully
23/02/28 16:53:16 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=548312
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```



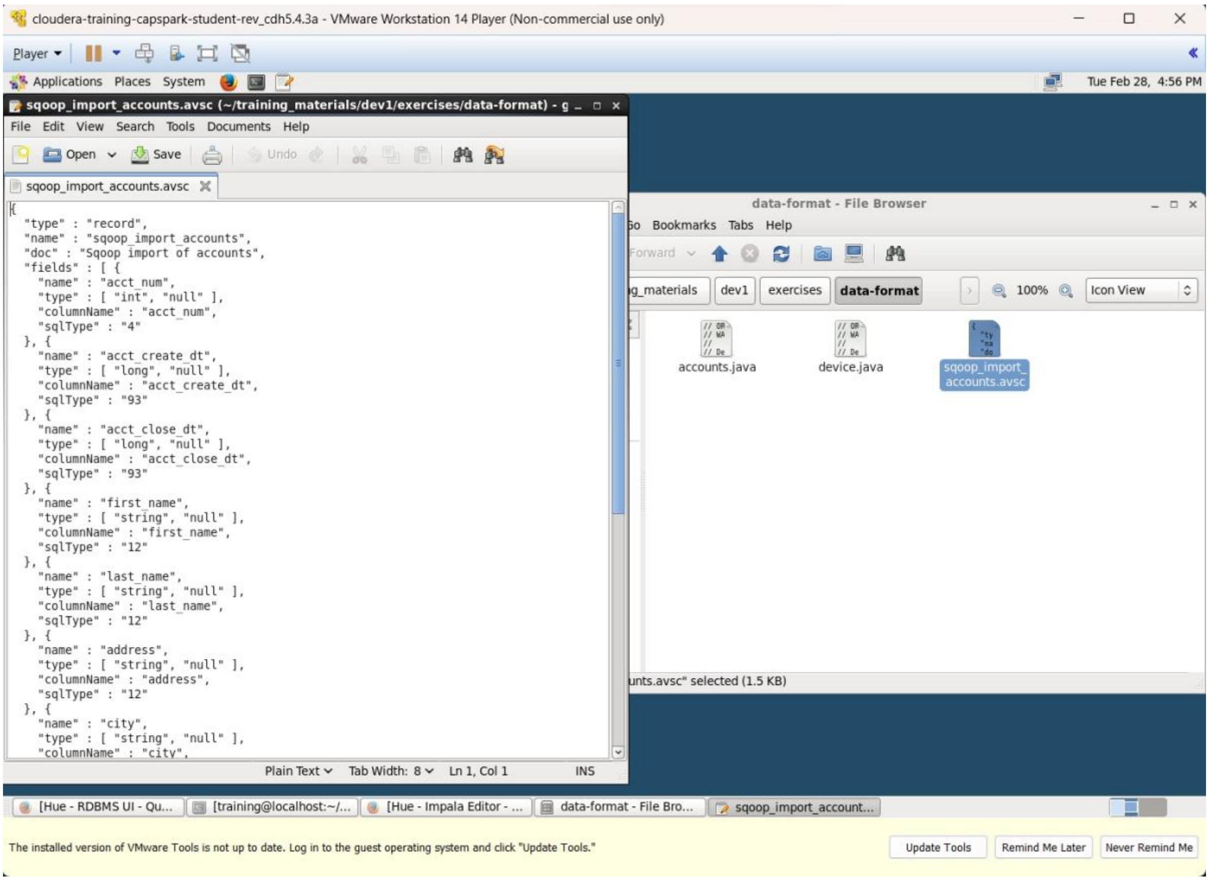
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System
Tue Feb 28, 4:54 PM
training@localhost:~/training_materials/dev1/exercises/data-format
File Edit View Search Terminal Help

```
3/02/28 16:53:09 INFO mapreduce.Job:  map 75% reduce 0%
3/02/28 16:53:16 INFO mapreduce.Job:  map 100% reduce 0%
3/02/28 16:53:16 INFO mapreduce.Job: Job job_1677627186634_0005 completed successfully
3/02/28 16:53:16 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=548312
                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=12713125
                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)=23005
                Total vcore-seconds taken by all map tasks=23005
                Total megabyte-seconds taken by all map tasks=5889280
        Map-Reduce Framework
                Map input records=129764
                Map output records=129764
                Input split bytes=470
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=427
                CPU time spent (ms)=10330
                Physical memory (bytes) snapshot=538406912
                Virtual memory (bytes) snapshot=3382845440
                Total committed heap usage (bytes)=191889408
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=12713125
3/02/28 16:53:16 INFO mapreduce.ImportJobBase: Transferred 12.1242 MB in 38.3966 seconds (323.3399 KB/sec)
3/02/28 16:53:16 INFO mapreduce.ImportJobBase: Retrieved 129764 records.
training@localhost data-format]$ date
ue Feb 28 16:53:35 PST 2023
training@localhost data-format]$ YJS220000
```

Hue - Hive Editor - Qu...
[Hue - RDBMS UI - Qu...
training@localhost:~/t...
Hue - Impala Editor - Q...

The installed version of VMware Tools is not up to date. Log in to the guest operating system and click "Update Tools."

3. Now view the files imported by Sqoop into HDFS. What do you see when you try to view the content of the data files?

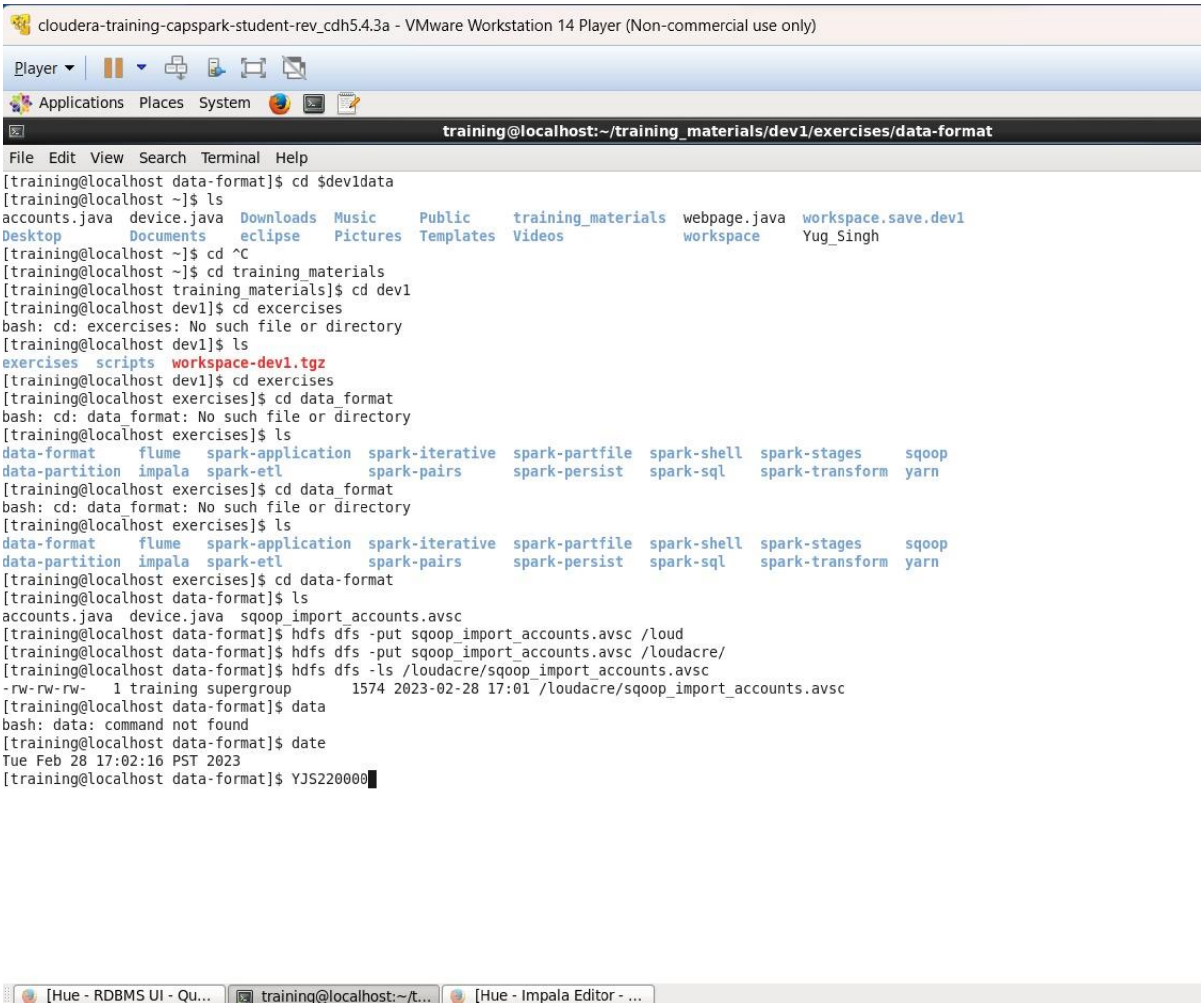

cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System
Tue Feb 28, 4:56 PM
sqoop_import_accounts.avsc (~/training_materials/dev1/exercises/data-format) - g
File Edit View Search Tools Documents Help
Open Save Undo
sqoop_import_accounts.avsc
Plain Text Tab Width: 8 Ln 1, Col 1 INS
data-format - File Browser
Go Bookmarks Tabs Help
Forward
g_materials dev1 exercises data-format
100% Icon View
accounts.java
device.java
sqoop_import_accounts.avsc
[Hue - RDBMS UI - Qu...
[training@localhost:~/...
[Hue - Impala Editor - ...
data-format - File Bro...
sqoop_import_account...
Update Tools
Remind Me Later
Never Remind Me

```
{
  "type" : "record",
  "name" : "sqoop_import_accounts",
  "doc" : "Sqoop import of accounts",
  "fields" : [ {
    "name" : "acct_num",
    "type" : [ "int", "null" ],
    "columnName" : "acct_num",
    "sqlType" : "4"
  }, {
    "name" : "acct_create_dt",
    "type" : [ "long", "null" ],
    "columnName" : "acct_create_dt",
    "sqlType" : "93"
  }, {
    "name" : "acct_close_dt",
    "type" : [ "long", "null" ],
    "columnName" : "acct_close_dt",
    "sqlType" : "93"
  }, {
    "name" : "first_name",
    "type" : [ "string", "null" ],
    "columnName" : "first_name",
    "sqlType" : "12"
  }, {
    "name" : "last_name",
    "type" : [ "string", "null" ],
    "columnName" : "last_name",
    "sqlType" : "12"
  }, {
    "name" : "address",
    "type" : [ "string", "null" ],
    "columnName" : "address",
    "sqlType" : "12"
  }, {
    "name" : "city",
    "type" : [ "string", "null" ],
    "columnName" : "city",
```

unts.avsc" selected (1.5 KB)

The installed version of VMware Tools is not up to date. Log in to the guest operating system and click "Update Tools."

4. Sqoop generated a schema named sqoop_import_accounts.avsc in the current directory. Review this file and then copy it to the /loudacre directory in HDFS.


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System
training@localhost:~/training_materials/dev1/exercises/data-format
File Edit View Search Terminal Help
[Hue - RDBMS UI - Qu...
training@localhost:~/t...
[Hue - Impala Editor - ...

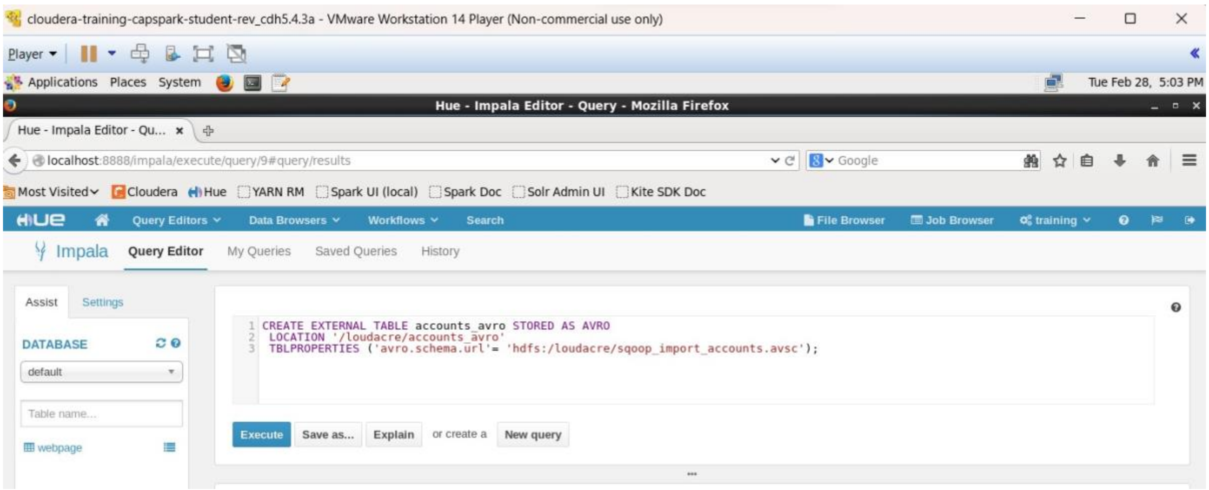
```
[training@localhost data-format]$ cd $dev1data
[training@localhost ~]$ ls
accounts.java  device.java  Downloads  Music     Public     training_materials  webpage.java  workspace.save.dev1
Desktop        Documents    eclipse    Pictures  Templates  Videos              workspace     Yug_Singh
[training@localhost ~]$ cd ^C
[training@localhost ~]$ cd training_materials
[training@localhost training_materials]$ cd dev1
[training@localhost dev1]$ cd excercises
bash: cd: excercises: No such file or directory
[training@localhost dev1]$ ls
exercises  scripts  workspace-dev1.tgz
[training@localhost dev1]$ cd exercises
[training@localhost exercises]$ cd data_format
bash: cd: data_format: No such file or directory
[training@localhost exercises]$ ls
data-format     flume   spark-application  spark-iterative  spark-partfile  spark-shell  spark-stages     sqoop
data-partition  impala  spark-etl          spark-pairs      spark-persist   spark-sql    spark-transform  yarn
[training@localhost exercises]$ cd data_format
bash: cd: data_format: No such file or directory
[training@localhost exercises]$ ls
data-format     flume   spark-application  spark-iterative  spark-partfile  spark-shell  spark-stages     sqoop
data-partition  impala  spark-etl          spark-pairs      spark-persist   spark-sql    spark-transform  yarn
[training@localhost exercises]$ cd data-format
[training@localhost data-format]$ ls
accounts.java  device.java  sqoop_import_accounts.avsc
[training@localhost data-format]$ hdfs dfs -put sqoop_import_accounts.avsc /loud
[training@localhost data-format]$ hdfs dfs -put sqoop_import_accounts.avsc /loudacre/
[training@localhost data-format]$ hdfs dfs -ls /loudacre/sqoop_import_accounts.avsc
-rw-rw-rw-   1 training supergroup       1574 2023-02-28 17:01 /loudacre/sqoop_import_accounts.avsc
[training@localhost data-format]$ data
bash: data: command not found
[training@localhost data-format]$ date
Tue Feb 28 17:02:16 PST 2023
[training@localhost data-format]$ YJS220000
```

5. In Impala or Hive, create a table using this schema:

CREATE EXTERNAL TABLE accounts_avro STORED AS AVRO

LOCATION '/loudacre/accounts_avro'

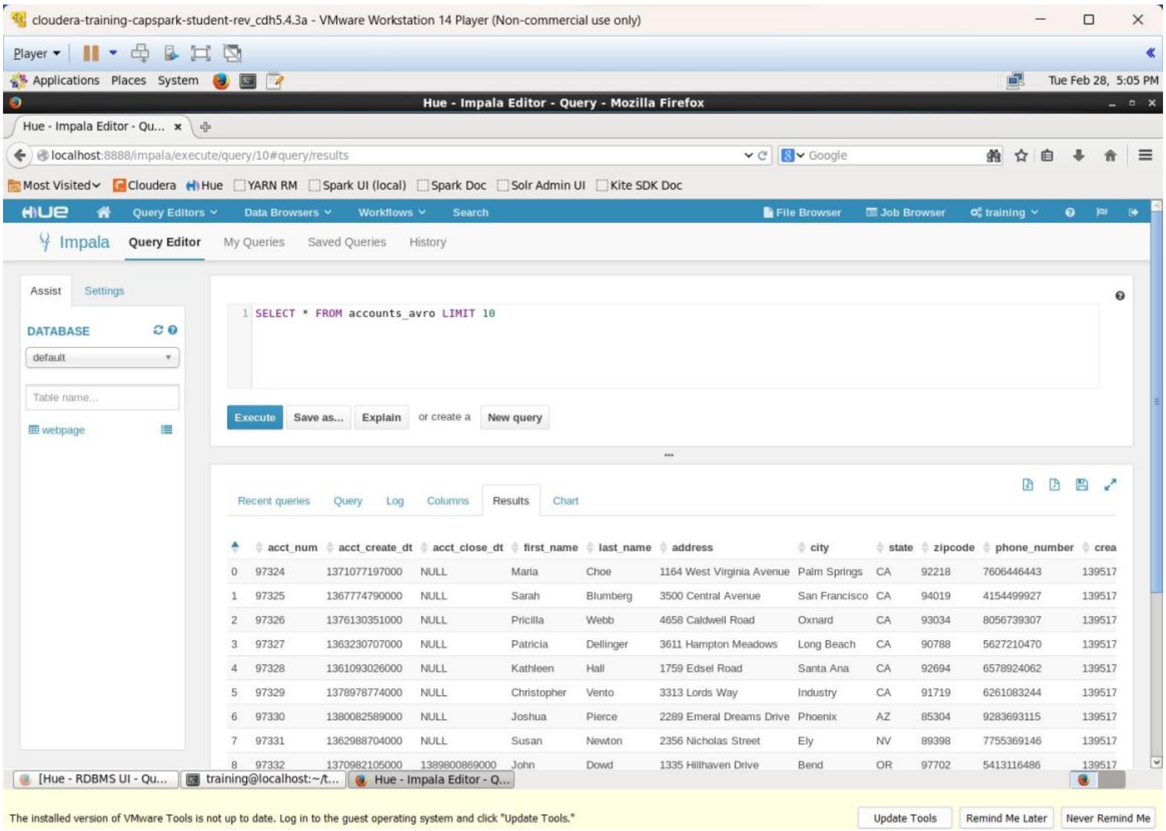
TBLPROPERTIES ('avro.schema.url'=
'hdfs:/loudacre/sqoop_import_accounts.avsc');


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Hue - Impala Editor - Query - Mozilla Firefox
localhost:8888/impala/execute/query/9#query/results
Impala
Query Editor
My Queries
Saved Queries
History
Execute
Save as...
Explain
or create a
New query

```
CREATE EXTERNAL TABLE accounts_avro STORED AS AVRO
 LOCATION '/loudacre/accounts_avro'
 TBLPROPERTIES ('avro.schema.url'= 'hdfs:/loudacre/sqoop_import_accounts.avsc');
```

6. Confirm correct creation of the table by issuing a query such as

SELECT * FROM accounts_avro LIMIT 10


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Hue - Impala Editor - Query - Mozilla Firefox
localhost:8888/impala/execute/query/10#query/results
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart

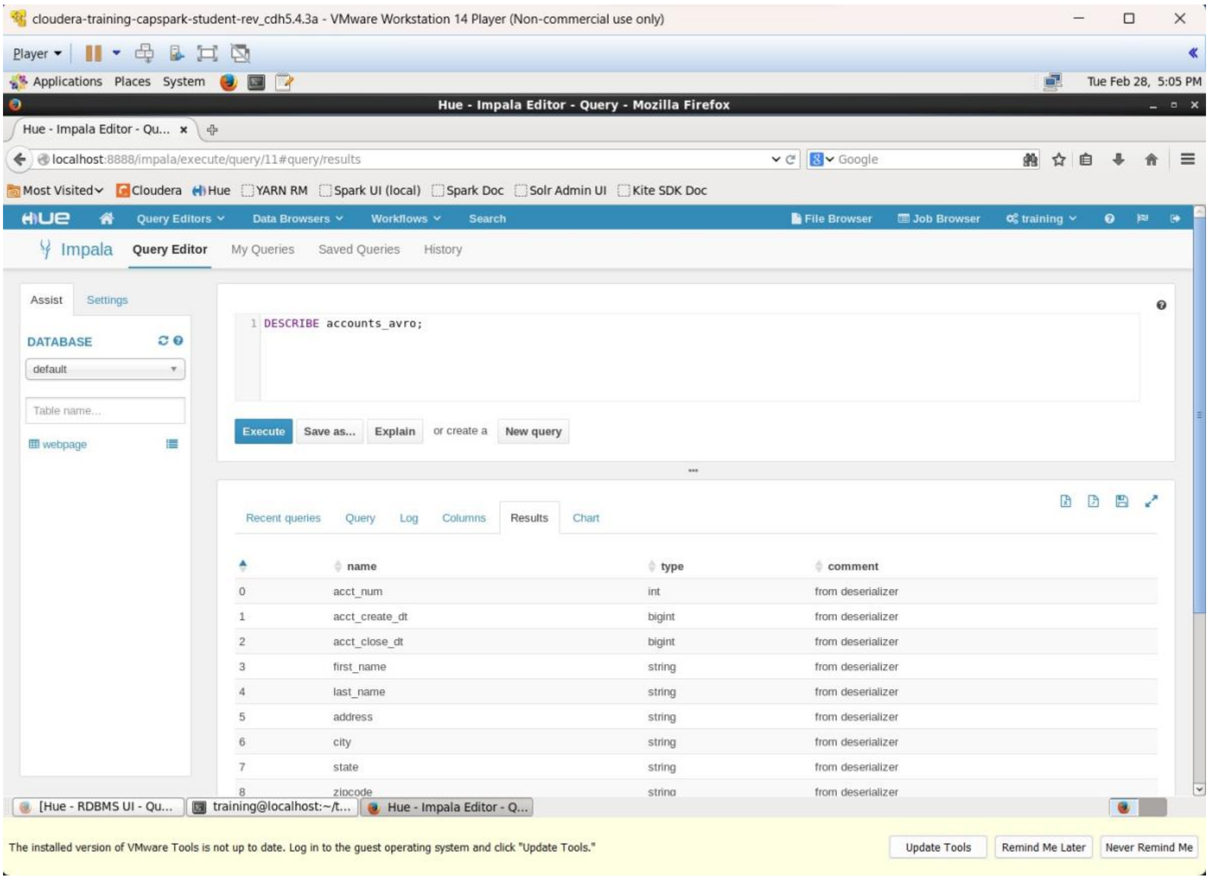
```
SELECT * FROM accounts_avro LIMIT 10
```

	acct_num	acct_create_dt	acct_close_dt	first_name	last_name	address	city	state	zipcode	phone_number	crea
0	97324	1371077197000	NULL	Maria	Choe	1164 West Virginia Avenue	Palm Springs	CA	92218	7606446443	139517
1	97325	1367774790000	NULL	Sarah	Blumberg	3500 Central Avenue	San Francisco	CA	94019	4154499927	139517
2	97326	1376130351000	NULL	Pricilla	Webb	4658 Caldwell Road	Oxnard	CA	93034	8056739307	139517
3	97327	1363230707000	NULL	Patricia	Dellinger	3611 Hampton Meadows	Long Beach	CA	90788	5627210470	139517
4	97328	1361093026000	NULL	Kathleen	Hall	1759 Edsel Road	Santa Ana	CA	92694	6578924062	139517
5	97329	1378978774000	NULL	Christopher	Vento	3313 Lords Way	Industry	CA	91719	6261083244	139517
6	97330	1380082589000	NULL	Joshua	Pierce	2289 Emeral Dreams Drive	Phoenix	AZ	85304	9283693115	139517
7	97331	1362988704000	NULL	Susan	Newton	2356 Nicholas Street	Ely	NV	89398	7755369146	139517
8	97332	1370982105000	1389800869000	John	Dowd	1335 Hillhaven Drive	Bend	OR	97702	5413116486	139517

The installed version of VMware Tools is not up to date. Log in to the guest operating system and click "Update Tools."

7. Use the DESCRIBE or DESCRIBE FORMATTED command to list the columns and data types of the accounts_avro table created from the Avro schema.

DESCRIBE accounts_avro;



cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Hue - Impala Editor - Query - Mozilla Firefox
localhost:8888/impala/execute/query/11#query/results
Impala
Query Editor
My Queries
Saved Queries
History
DATABASE
default
webpage
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart

```
DESCRIBE accounts_avro;
```

	name	type	comment
0	acct_num	int	from deserializer
1	acct_create_dt	bigint	from deserializer
2	acct_close_dt	bigint	from deserializer
3	first_name	string	from deserializer
4	last_name	string	from deserializer
5	address	string	from deserializer
6	city	string	from deserializer
7	state	string	from deserializer
8	zipcode	string	from deserializer

Chapter 8

Talking about this particular exercise, we will create and load an Impala/Hive table with account data, partitioned by area code.

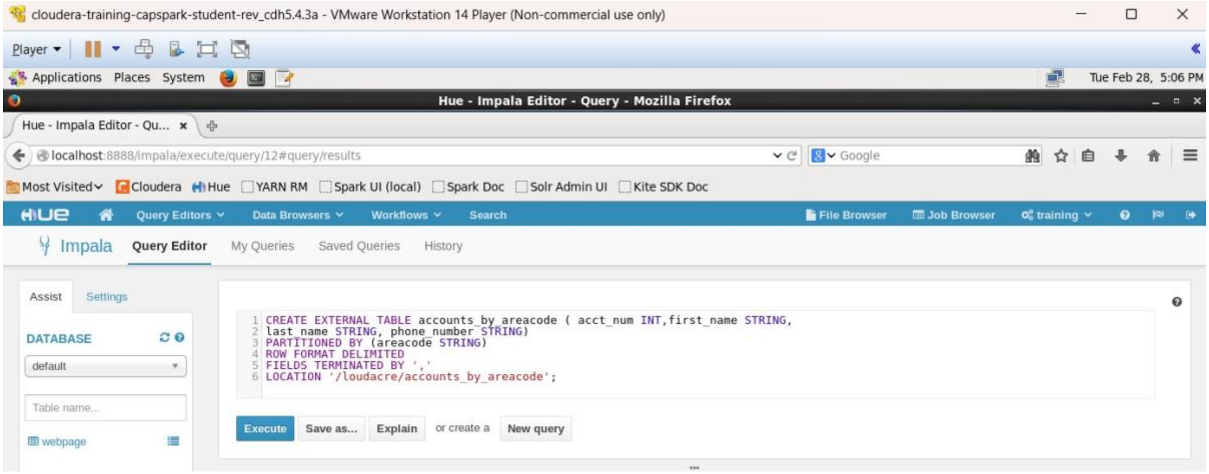
Talking about the previous exercise, we imported data from the accounts table using Sqoop, into a table called accounts_avro. In this exercise, you will create a new table with some of the account data, partitioned by area code (the first three digits of the phone number).

1. Create a new, empty table in Impala or Hive:

```
CREATE EXTERNAL TABLE accounts_by_areacode ( acct_num
INT,first_name STRING,

last_name STRING, phone_number STRING)

PARTITIONED BY (areacode STRING)

ROW FORMAT DELIMITED
```

FIELDS TERMINATED BY ','

LOCATION '/loudacre/accounts_by_areacode';

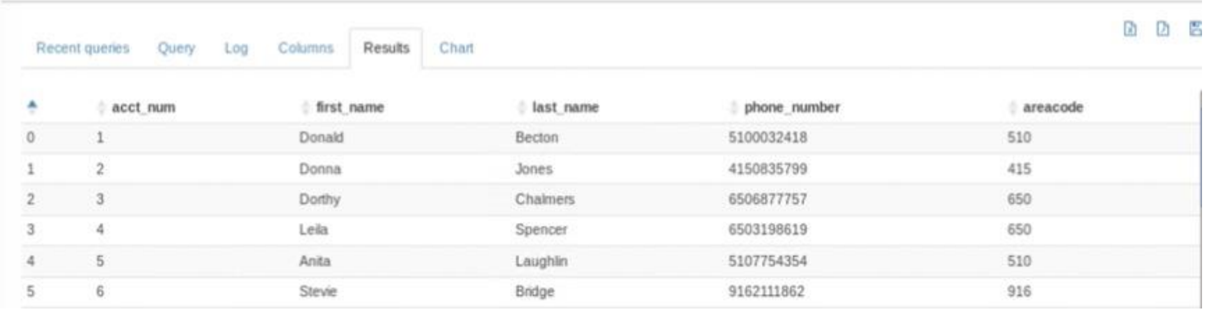


```
CREATE EXTERNAL TABLE accounts_by_areacode ( acct_num INT,first_name STRING,
last_name STRING, phone_number STRING)
PARTITIONED BY (areacode STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LOCATION '/loudacre/accounts_by_areacode';
```

2. In order to populate the new table, we will need to extract the area code from the phone number. We will try executing the following query to demonstrate:

SELECT acct_num, first_name, last_name, phone_number,

SUBSTR(phone_number,1,3) AS areacode FROM accounts_avro



	acct_num	first_name	last_name	phone_number	areacode
0	1	Donald	Becton	5100032418	510
1	2	Donna	Jones	4150835799	415
2	3	Dorthy	Chalmers	6506877757	650
3	4	Leila	Spencer	6503198619	650
4	5	Anita	Laughlin	5107754354	510
5	6	Stevie	Bridge	9162111862	916

3. Use the SELECT statement above in an INSERT INTO TABLE command to copy the specified columns to the new table, dynamically partitioning by area code.

INSERT INTO TABLE accounts_by_areacode PARTITION (areacode)

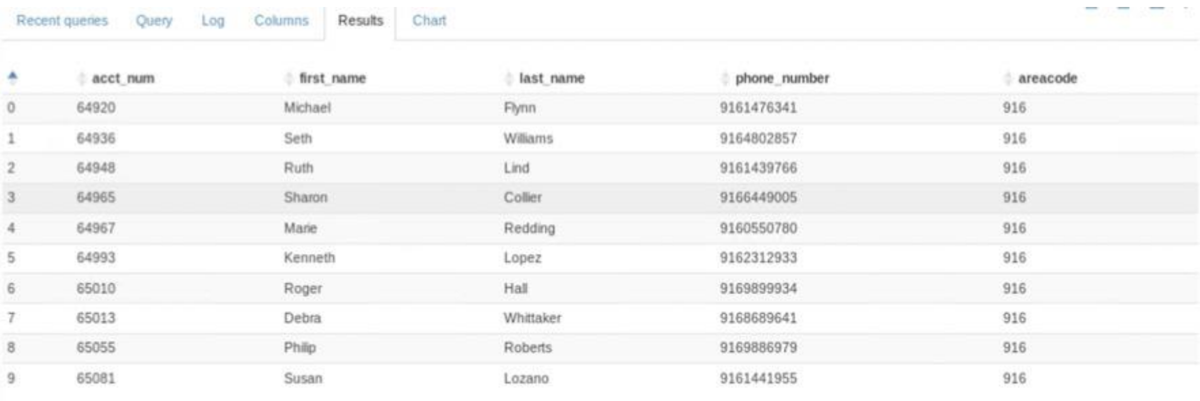
SELECT acct_num, first_name, last_name, phone_number,

SUBSTR(phone_number,1,3) AS areacode

FROM accounts_avro

4. Execute a simple query to confirm that the table was populated correctly, such as

SELECT * FROM accounts_by_areacode LIMIT 10



	acct_num	first_name	last_name	phone_number	areacode
0	64920	Michael	Flynn	9161476341	916
1	64936	Seth	Williams	9164802857	916
2	64948	Ruth	Lind	9161439766	916
3	64965	Sharon	Collier	9166449005	916
4	64967	Marie	Redding	9160550780	916
5	64993	Kenneth	Lopez	9162312933	916
6	65010	Roger	Hall	9169899934	916
7	65013	Debra	Whittaker	9168689641	916
8	65055	Philip	Roberts	9169886979	916
9	65081	Susan	Lozano	9161441955	916

5. Using Hue or the hdfs command-line interface, confirm that the directory structure of the accounts_by_areacode table includes partition directories. Review the data in the directories to verify that the partitioning is correct.

Chapter 9

Talking about this exercise, we will configure Flume to ingest web log data from a local directory to HDFS.

Apache web server logs are generally stored in files on the local machines running the server. In this exercise, we will simulate an Apache server by placing provided web log files into a local spool directory and then using Flume to collect the data. Both the local and HDFS directories must exist before using the spooling directory source.

1. Create a directory in HDFS called /loudacre/weblogs to hold the data files Flume ingests, example:

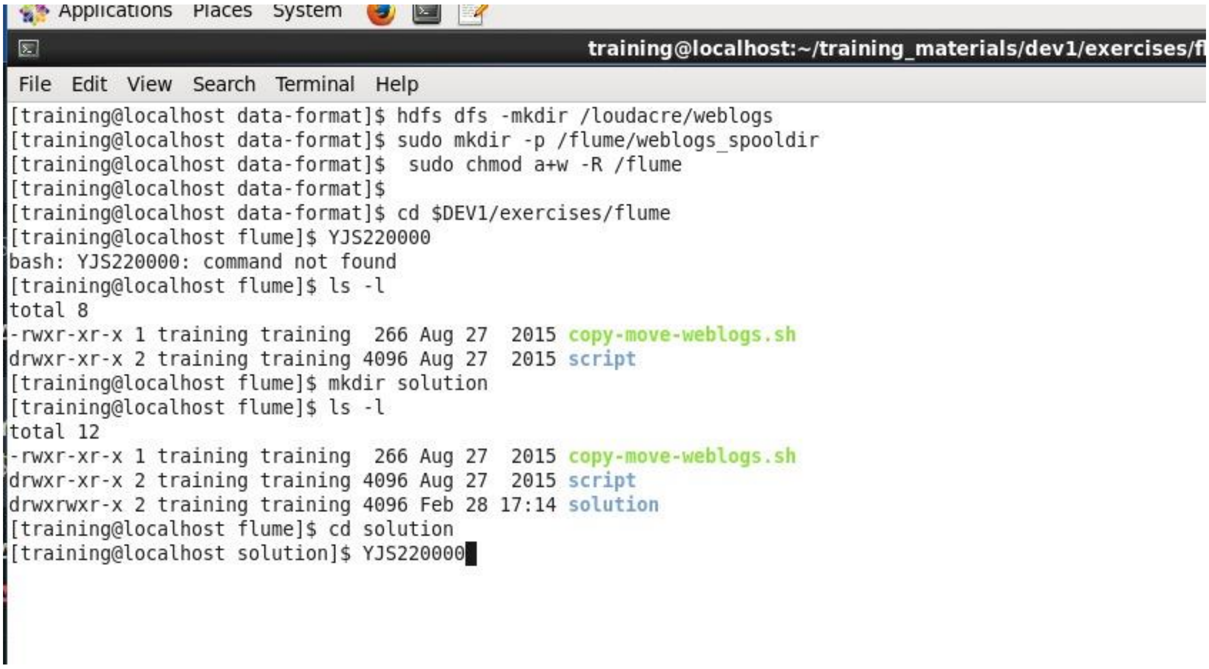
```
$ hdfs dfs -mkdir /loudacre/weblogs
```

2. Create the spool directory into which our weblog simulator will store data files for Flume to ingest. On the local filesystem create */flume/weblogs_spooldir:*

```
$ sudo mkdir -p /flume/weblogs_spooldir
```

3. Give all users the permission to write to the */flume/weblogs_spooldir* directory:

   $ sudo chmod a+w -R /flume


Applications Places System
training@localhost:~/training_materials/dev1/exercises/f
File Edit View Search Terminal Help

```
[training@localhost data-format]$ hdfs dfs -mkdir /loudacre/weblogs
[training@localhost data-format]$ sudo mkdir -p /flume/weblogs_spooldir
[training@localhost data-format]$  sudo chmod a+w -R /flume
[training@localhost data-format]$
[training@localhost data-format]$ cd $DEV1/exercises/flume
[training@localhost flume]$ YJS220000
bash: YJS220000: command not found
[training@localhost flume]$ ls -l
total 8
-rwxr-xr-x 1 training training  266 Aug 27  2015 copy-move-weblogs.sh
drwxr-xr-x 2 training training 4096 Aug 27  2015 script
[training@localhost flume]$ mkdir solution
[training@localhost flume]$ ls -l
total 12
-rwxr-xr-x 1 training training  266 Aug 27  2015 copy-move-weblogs.sh
drwxr-xr-x 2 training training 4096 Aug 27  2015 script
drwxrwxr-x 2 training training 4096 Feb 28 17:14 solution
[training@localhost flume]$ cd solution
[training@localhost solution]$ YJS220000
```

9.1 Configure Flume

In $DEV1/exercises/flume create a Flume configuration file with the characteristics listed below

• The source is a spooling directory source that pulls from

/flume/weblogs_spooldir

• The sink is an HDFS sink that:
- Writes files to the /loudacre/weblogs directory
- Disables time-based file rolling by setting the hdfs.rollInterval property to 0.
- Disables event-based file rolling by setting the hdfs.rollCount property to 0.
- Sets the hdfs.rollSize property as 524288 to enable size-based file rolling at 512KB. - Writes raw text files (instead of SequenceFile format) by setting hdfs.fileType to datastream.

- The channel is a Memory Channel that:

- Can store 10,000 events using the capacity property.
- Has a transaction capacity of 10,000 events using the transactionCapacity property.

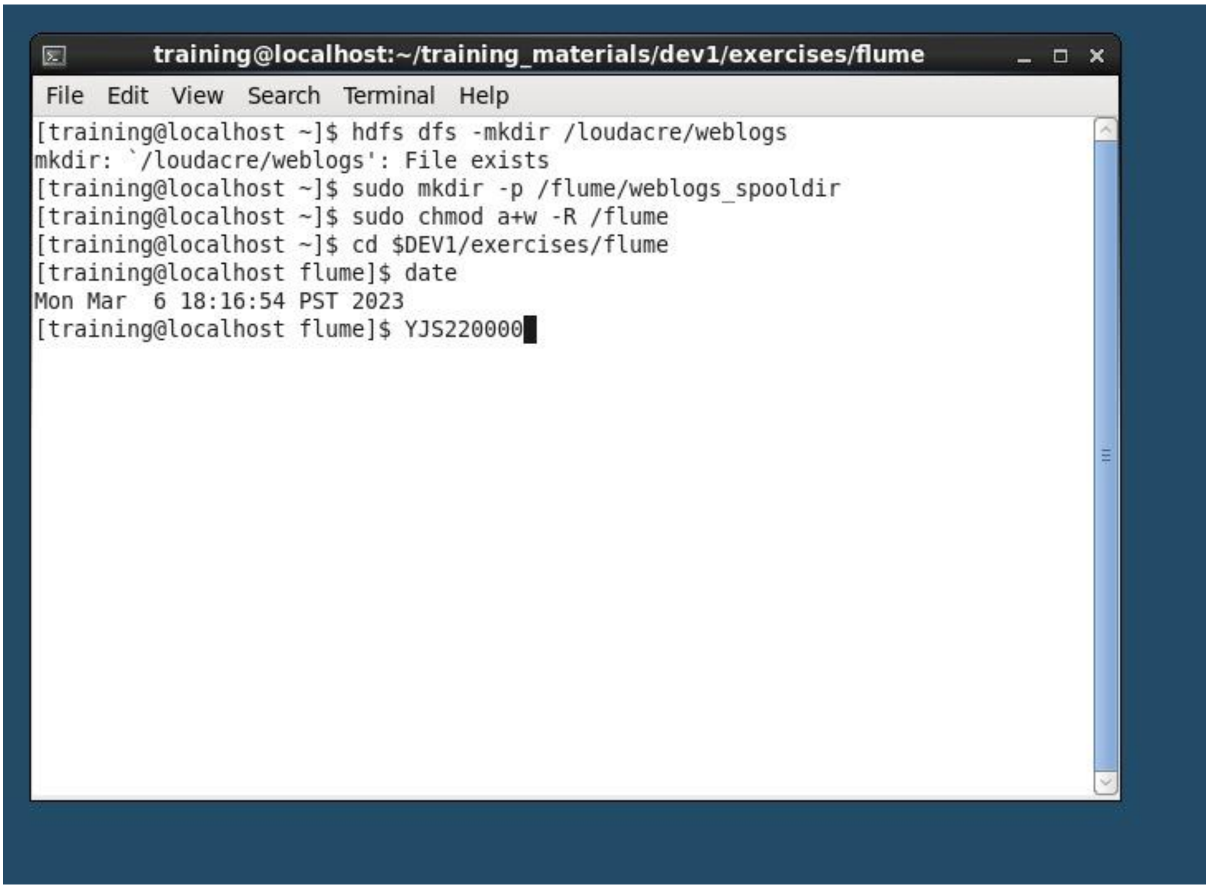
4. I changed the directory to the $DEV1/exercises/flume directory as follows:

   $ cd $DEV1/exercises/flume

- I created solution directory as it didn't exist.

- I then created the spooldir.conf in the solution directory and we can see that in the /solution directory

5. I started the flume agent that I had created in the last step (since it didn't exist I created it).


training@localhost:~/training_materials/dev1/exercises/flume
File Edit View Search Terminal Help

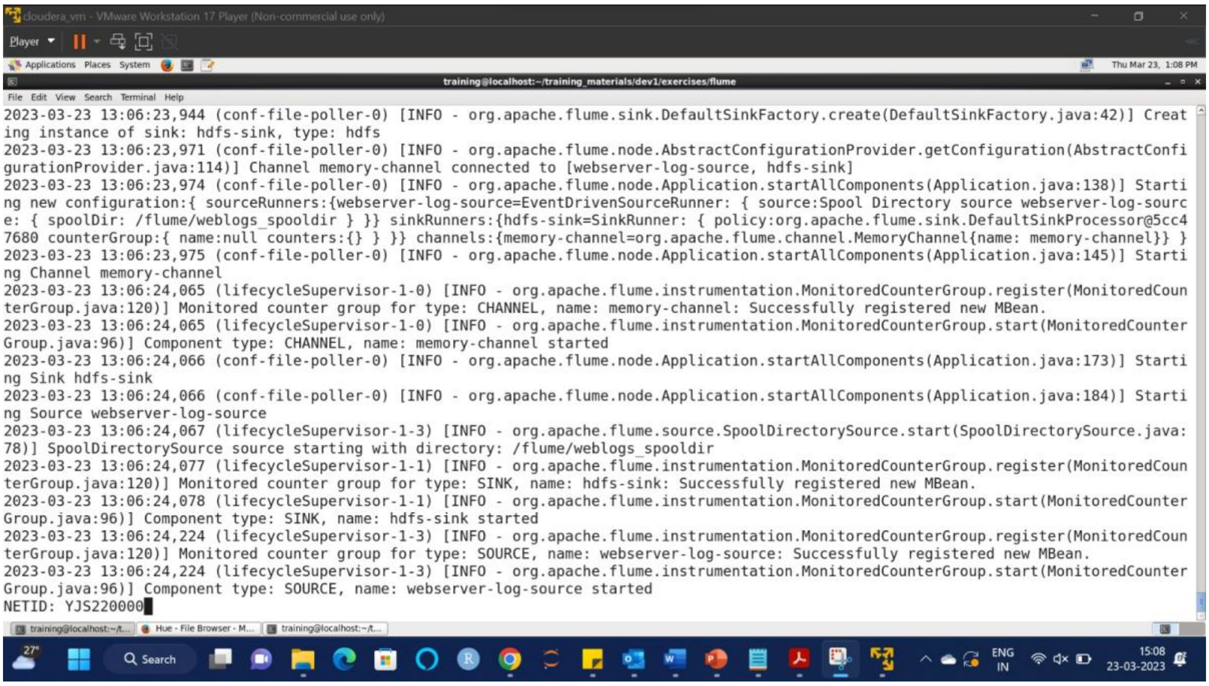
```
[training@localhost ~]$ hdfs dfs -mkdir /loudacre/weblogs
mkdir: `/loudacre/weblogs': File exists
[training@localhost ~]$ sudo mkdir -p /flume/weblogs_spooldir
[training@localhost ~]$ sudo chmod a+w -R /flume
[training@localhost ~]$ cd $DEV1/exercises/flume
[training@localhost flume]$ date
Mon Mar  6 18:16:54 PST 2023
[training@localhost flume]$ YJS220000
```

5.1 Starting the flume agent now using the following command:

flume-ng agent --conf /etc/flume-ng/conf \

```
--conf-file solution/spooldir.conf \
```

```
--name agent1 -Dflume.root.logger=INFO,console
```



doudera_vm - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
Thu Mar 23, 1:08 PM
training@localhost:~/training_materials/dev1/exercises/flume
File Edit View Search Terminal Help

```
2023-03-23 13:06:23,944 (conf-file-poller-0) [INFO - org.apache.flume.sink.DefaultSinkFactory.create(DefaultSinkFactory.java:42)] Creat
ing instance of sink: hdfs-sink, type: hdfs
2023-03-23 13:06:23,971 (conf-file-poller-0) [INFO - org.apache.flume.node.AbstractConfigurationProvider.getConfiguration(AbstractConfi
gurationProvider.java:114)] Channel memory-channel connected to [webserver-log-source, hdfs-sink]
2023-03-23 13:06:23,974 (conf-file-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:138)] Starti
ng new configuration:{ sourceRunners:{webserver-log-source=EventDrivenSourceRunner: { source:Spool Directory source webserver-log-sourc
e: { spoolDir: /flume/weblogs_spooldir } }} sinkRunners:{hdfs-sink=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@5cc4
7680 counterGroup:{ name:null counters:{} } }} channels:{memory-channel=org.apache.flume.channel.MemoryChannel{name: memory-channel}} }
2023-03-23 13:06:23,975 (conf-file-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:145)] Starti
ng Channel memory-channel
2023-03-23 13:06:24,065 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.register(MonitoredCoun
terGroup.java:120)] Monitored counter group for type: CHANNEL, name: memory-channel: Successfully registered new MBean.
2023-03-23 13:06:24,065 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.start(MonitoredCounter
Group.java:96)] Component type: CHANNEL, name: memory-channel started
2023-03-23 13:06:24,066 (conf-file-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:173)] Starti
ng Sink hdfs-sink
2023-03-23 13:06:24,066 (conf-file-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:184)] Starti
ng Source webserver-log-source
2023-03-23 13:06:24,067 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.source.SpoolDirectorySource.start(SpoolDirectorySource.java:
78)] SpoolDirectorySource source starting with directory: /flume/weblogs_spooldir
2023-03-23 13:06:24,077 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.register(MonitoredCoun
terGroup.java:120)] Monitored counter group for type: SINK, name: hdfs-sink: Successfully registered new MBean.
2023-03-23 13:06:24,078 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.start(MonitoredCounter
Group.java:96)] Component type: SINK, name: hdfs-sink started
2023-03-23 13:06:24,224 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.register(MonitoredCoun
terGroup.java:120)] Monitored counter group for type: SOURCE, name: webserver-log-source: Successfully registered new MBean.
2023-03-23 13:06:24,224 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.start(MonitoredCounter
Group.java:96)] Component type: SOURCE, name: webserver-log-source started
NETID: YJS220000
```

training@localhost:~/t...
Hue - File Browser - M...
training@localhost:~/t...
27°
Search
ENG
IN
15:08
23-03-2023

Flume agent to start up. I saw the message like: Component type: SOURCE, name: webserver-log-source started

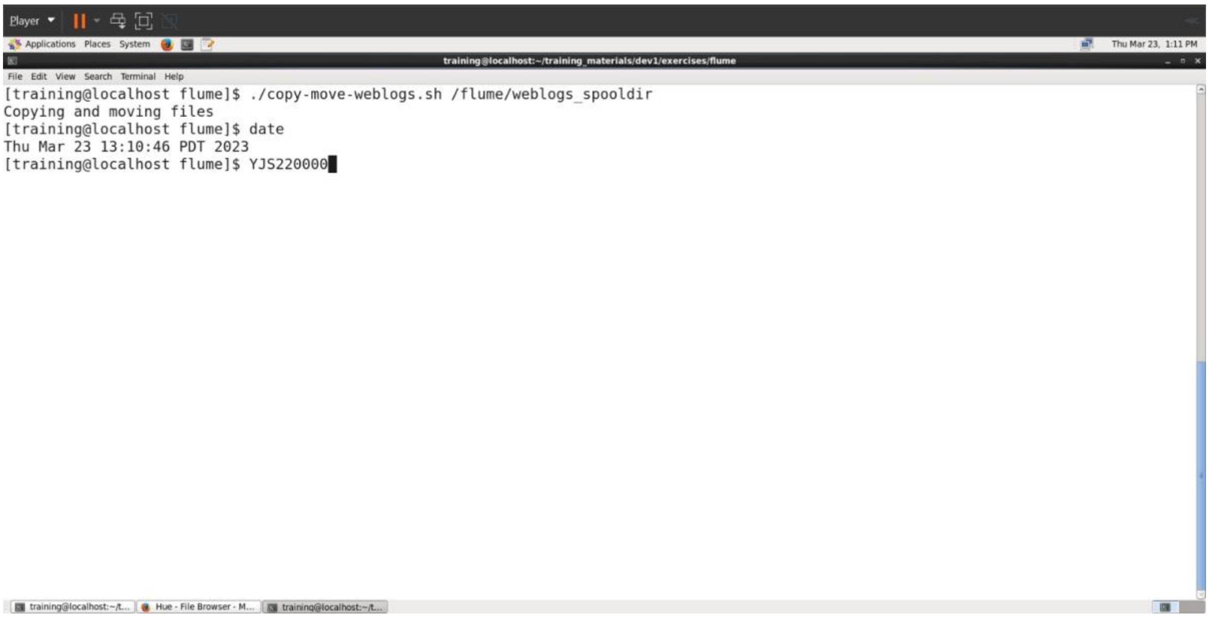
9.2 Simulate Apache Web Server Output

7. Open a separate terminal window, and change to the exercise directory. Run the script to place the web log files in the /flume/weblogs_spooldir directory: This script will create a temporary copy of the web log files and move them to the spooldir directory.

I wasn't able to run the following scripts in VMware because of certain error that was popping on my screen several times.

```
$ cd $DEV1/exercises/flume
```

```
$ ./copy-move-weblogs.sh /flume/weblogs_spooldir
```



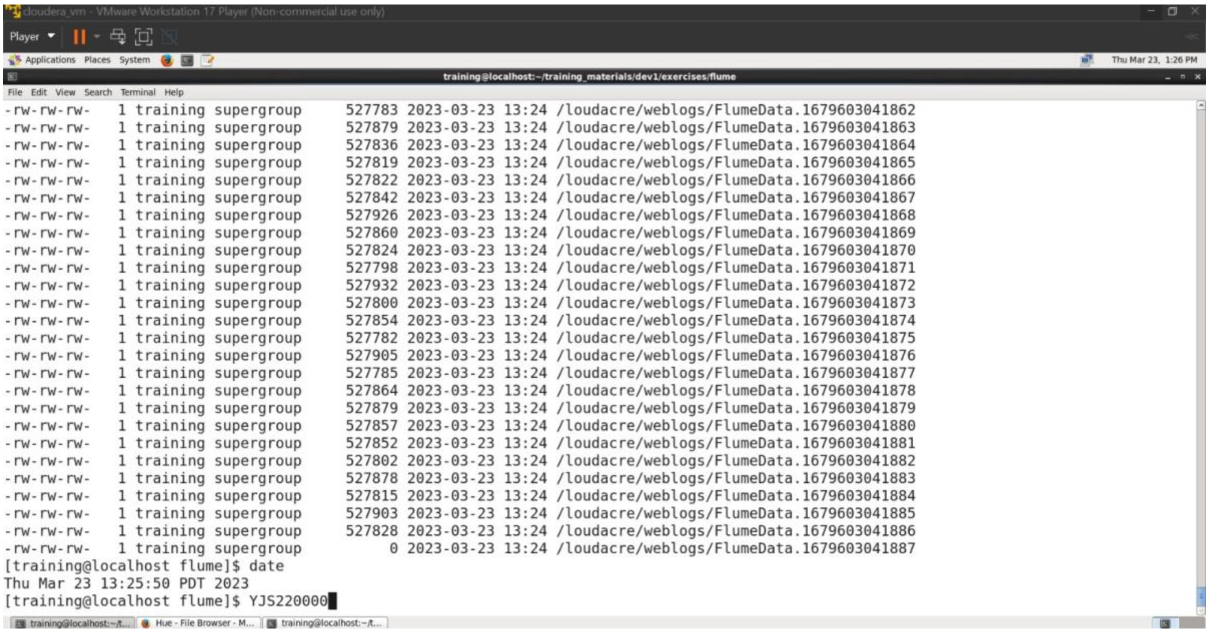
```
[training@localhost flume]$ ./copy-move-weblogs.sh /flume/weblogs_spooldir
Copying and moving files
[training@localhost flume]$ date
Thu Mar 23 13:10:46 PDT 2023
[training@localhost flume]$ YJS220000
```

8. The other terminal that is running the Flume agent and watch the logging output. The output will give information about the files Flume is putting into HDFS.

9. Once the Flume agent has finished, enter CTRL+C to terminate the process.

10. I then listed the files in HDFS that were added by the flume agent.

$ hdfs dfs -ls /loudacre/weblogs



```
-rw-rw-rw-   1 training supergroup     527783 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041862
-rw-rw-rw-   1 training supergroup     527879 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041863
-rw-rw-rw-   1 training supergroup     527836 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041864
-rw-rw-rw-   1 training supergroup     527819 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041865
-rw-rw-rw-   1 training supergroup     527822 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041866
-rw-rw-rw-   1 training supergroup     527842 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041867
-rw-rw-rw-   1 training supergroup     527926 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041868
-rw-rw-rw-   1 training supergroup     527860 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041869
-rw-rw-rw-   1 training supergroup     527824 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041870
-rw-rw-rw-   1 training supergroup     527798 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041871
-rw-rw-rw-   1 training supergroup     527932 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041872
-rw-rw-rw-   1 training supergroup     527800 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041873
-rw-rw-rw-   1 training supergroup     527854 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041874
-rw-rw-rw-   1 training supergroup     527782 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041875
-rw-rw-rw-   1 training supergroup     527905 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041876
-rw-rw-rw-   1 training supergroup     527785 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041877
-rw-rw-rw-   1 training supergroup     527864 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041878
-rw-rw-rw-   1 training supergroup     527879 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041879
-rw-rw-rw-   1 training supergroup     527857 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041880
-rw-rw-rw-   1 training supergroup     527852 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041881
-rw-rw-rw-   1 training supergroup     527802 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041882
-rw-rw-rw-   1 training supergroup     527878 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041883
-rw-rw-rw-   1 training supergroup     527815 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041884
-rw-rw-rw-   1 training supergroup     527903 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041885
-rw-rw-rw-   1 training supergroup     527828 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041886
-rw-rw-rw-   1 training supergroup          0 2023-03-23 13:24 /loudacre/weblogs/FlumeData.1679603041887
[training@localhost flume]$ date
Thu Mar 23 13:25:50 PDT 2023
[training@localhost flume]$ YJS220000
```