BUAN 6346 Big Data Analytics
PHASE 2

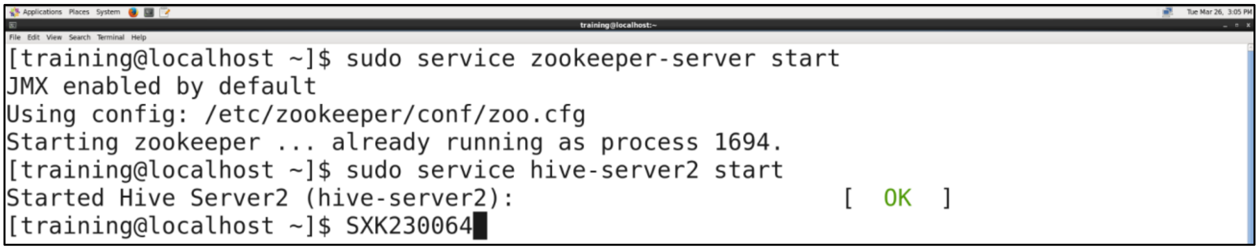
SHIVA KUMAR REDDY KOPPULA
03/27/2024

Table of Contents

CHAPTER 6 - CREATE AND POPULATE TABLES IN IMPALA OR HIVE

Create and Query a Table in Impala or Hive

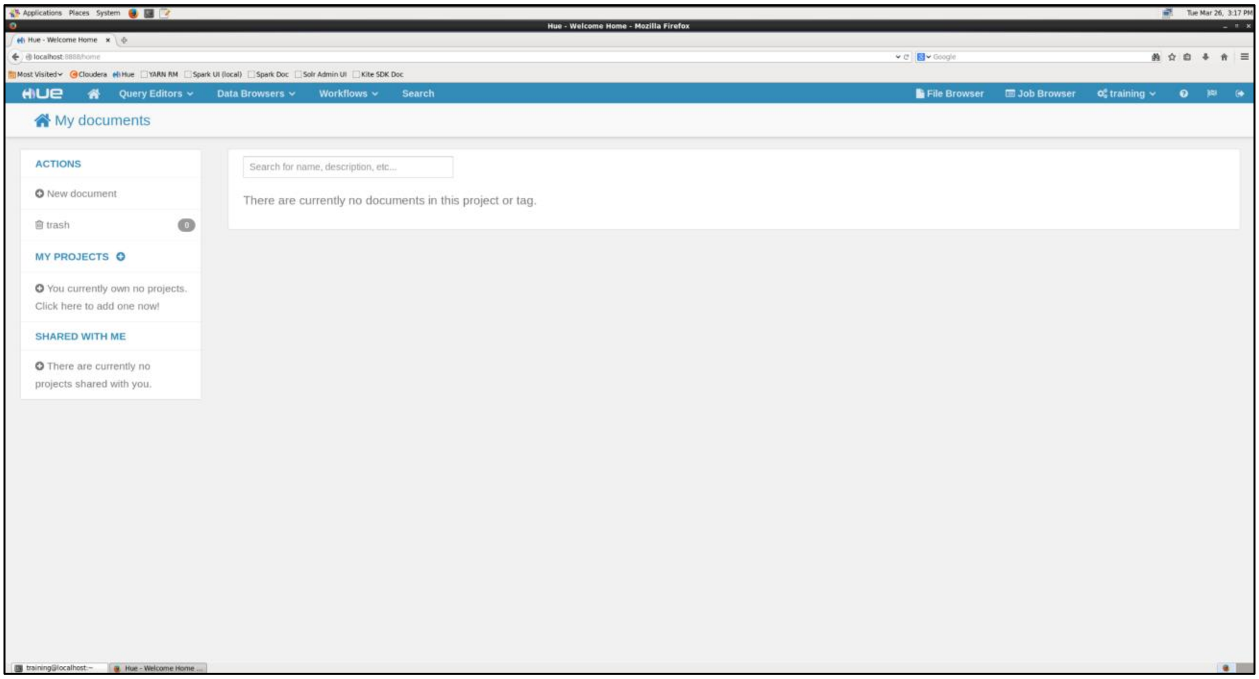
1. Started the Hive server by executing the necessary commands in the terminal.



Applications Places System
Tue Mar 26, 3:05 PM
training@localhost:~
File Edit View Search Terminal Help

```
[training@localhost ~]$ sudo service zookeeper-server start
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Starting zookeeper ... already running as process 1694.
[training@localhost ~]$ sudo service hive-server2 start
Started Hive Server2 (hive-server2):                       [  OK  ]
[training@localhost ~]$ SXK230064
```

2. Accessed the HUE interface in Chrome, clicking on the HUE icon to navigate to the homepage.

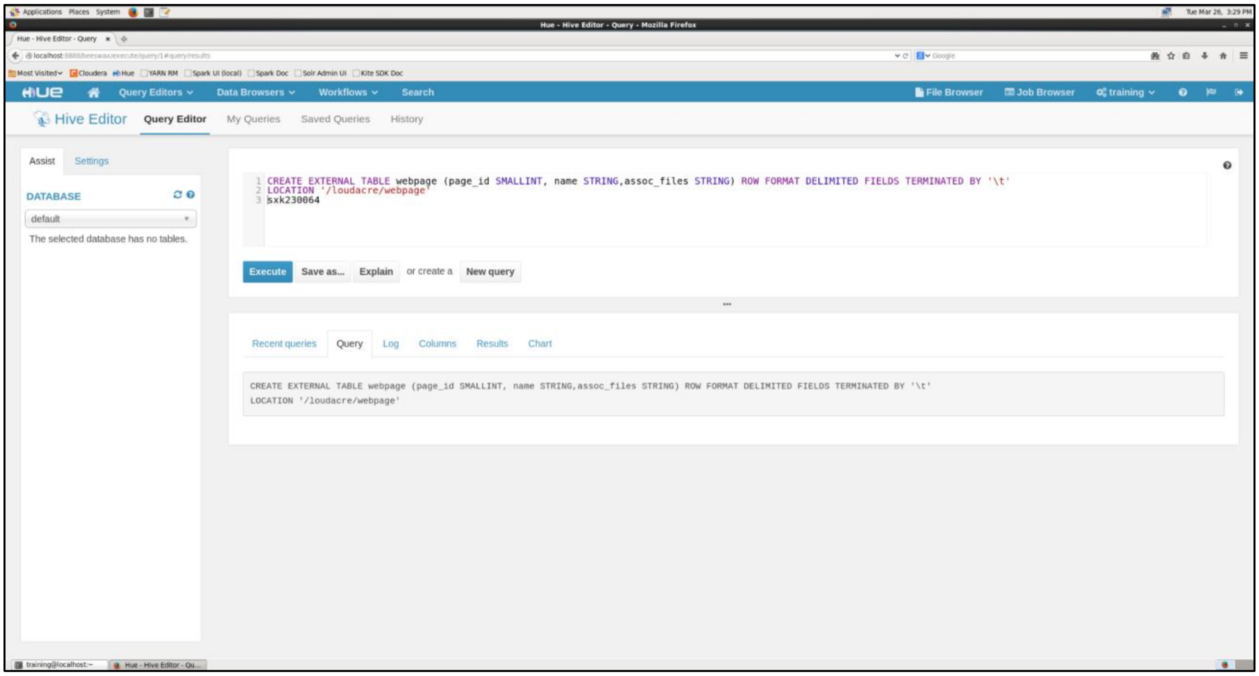


There are currently no documents in this project or tag.

You currently own no projects. Click here to add one now!

There are currently no projects shared with you.

3. Opened the query editor menu, selecting "Hive" from the dropdown, which directed me to the Hive query editor. Then, I entered an SQL command in the query editor pane to create a table for the previously imported webpage data.

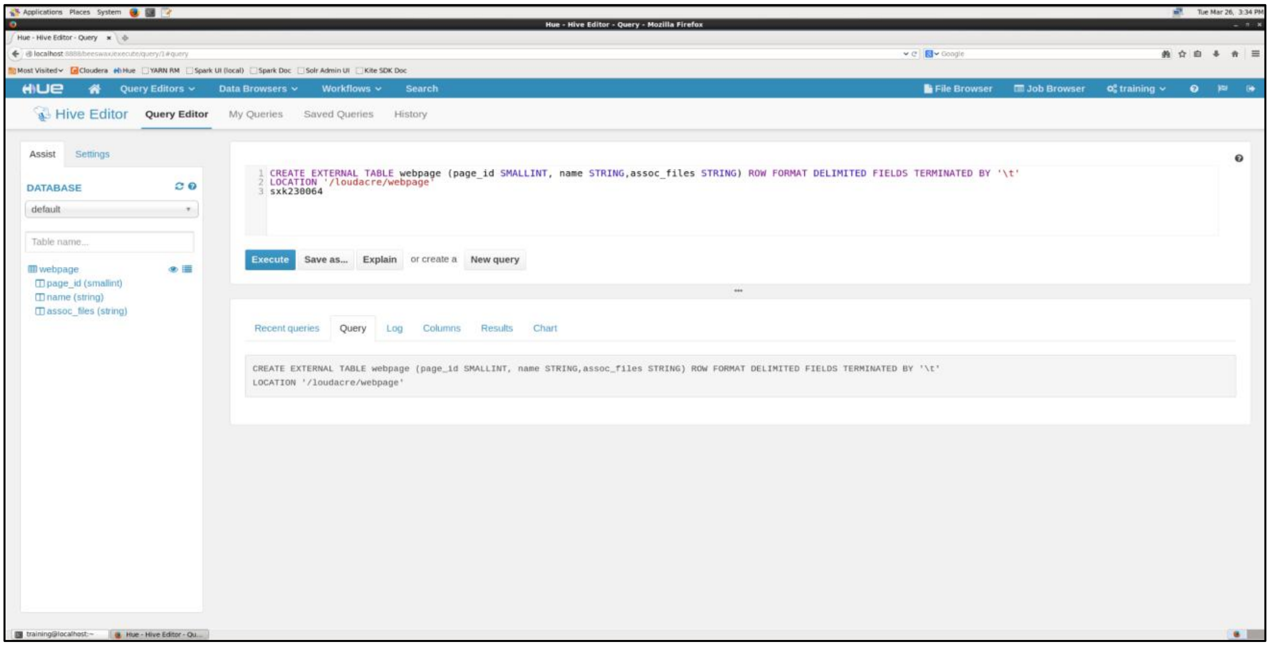


The selected database has no tables.

```
CREATE EXTERNAL TABLE webpage (page_id SMALLINT, name STRING,assoc_files STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/loudacre/webpage'
sxk230064
```

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

4. Executed the query by clicking the "Execute" button. To view the newly created table, I clicked on the 'refresh' button next to 'Database' on the left-hand side of the page, locating the webpage table with columns (page_id), name, and (assoc_files).I clicked on the webpage table to review the column definitions which reside below it.



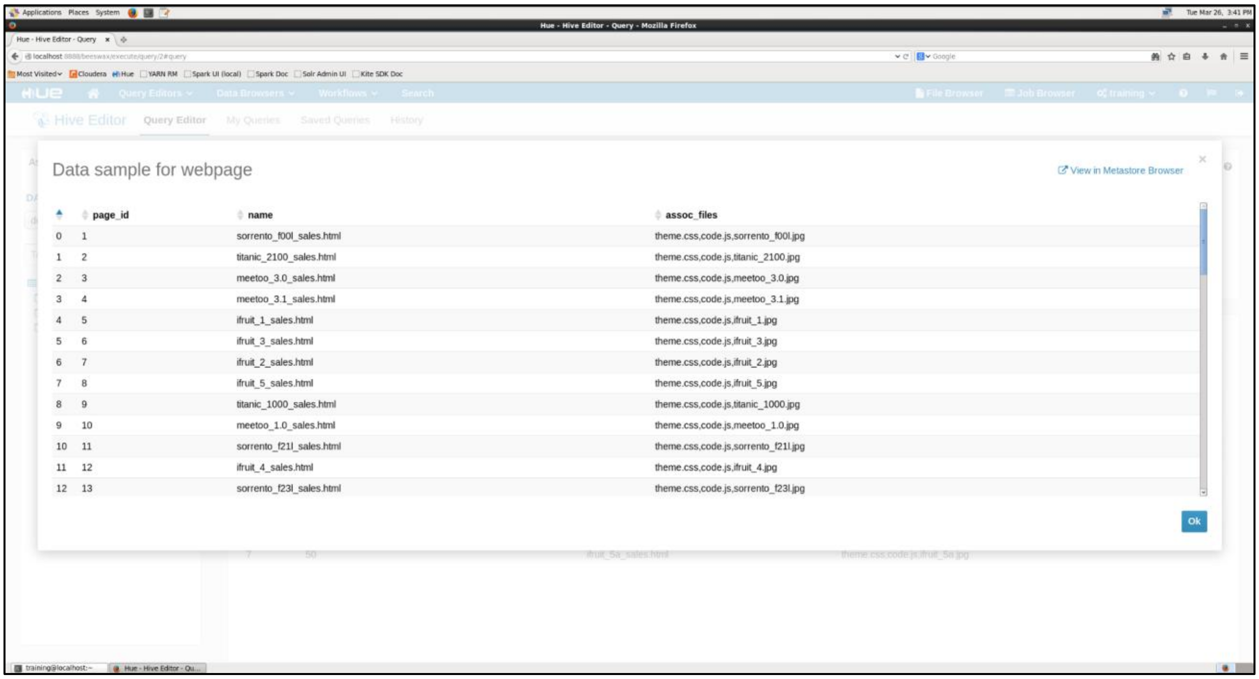
webpage
page_id (smallint)
name (string)
assoc_files (string)

```
CREATE EXTERNAL TABLE webpage (page_id SMALLINT, name STRING,assoc_files STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/loudacre/webpage'
sxk230064
```

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

5. Initiated a test query by clicking the 'New Query' button, observing the findings in the "Results" tab.

6. Previewed sample data by clicking on the preview sample data icon adjacent to the table name.

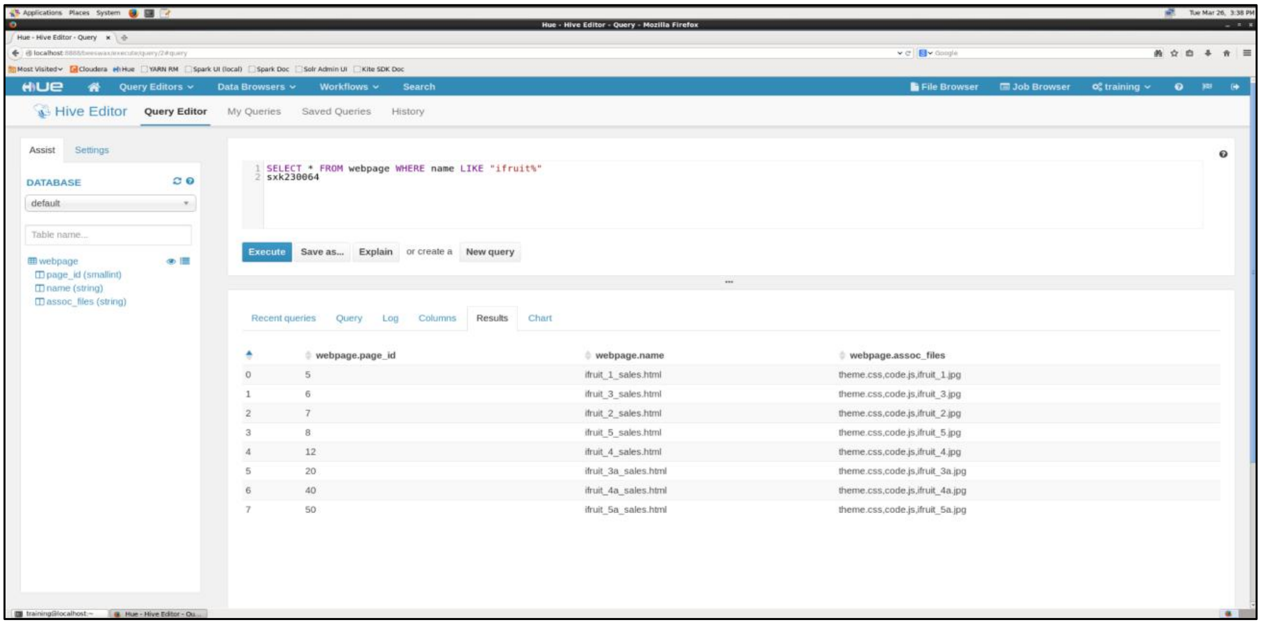


Data sample for webpage

	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

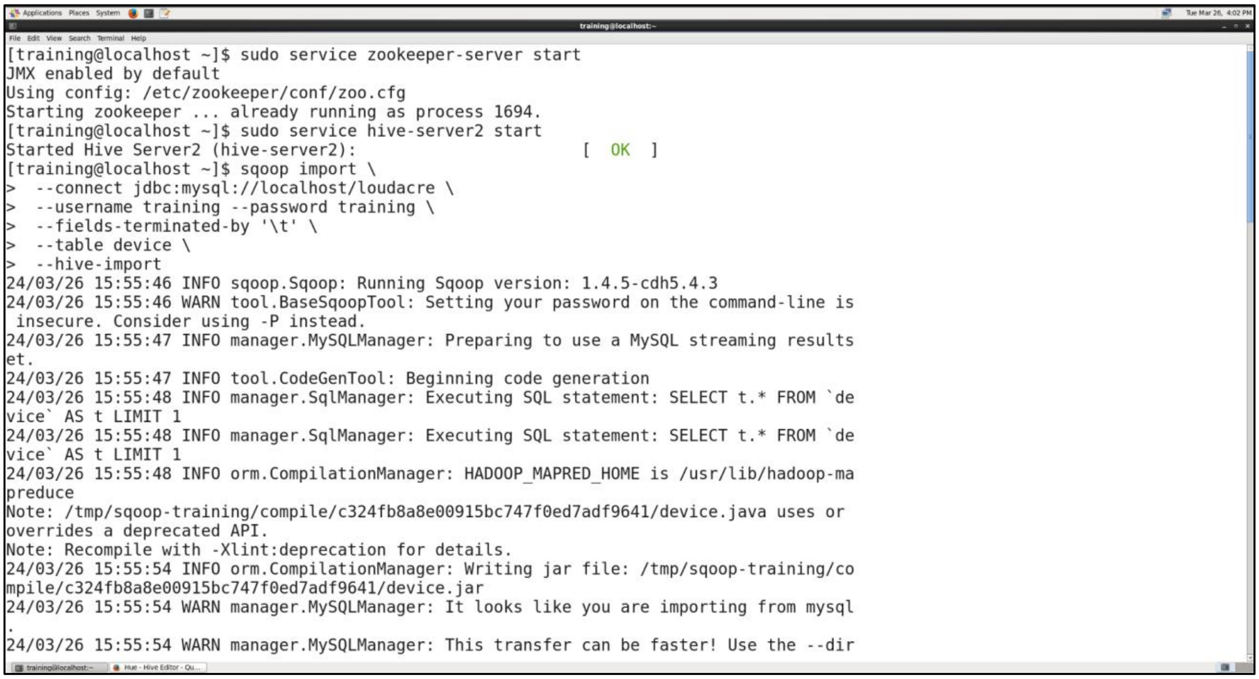
Use Sqoop to Import Directly into Hive and Impala

7. Imported the device table directly into the Hive metastore using the terminal, employing the below command.

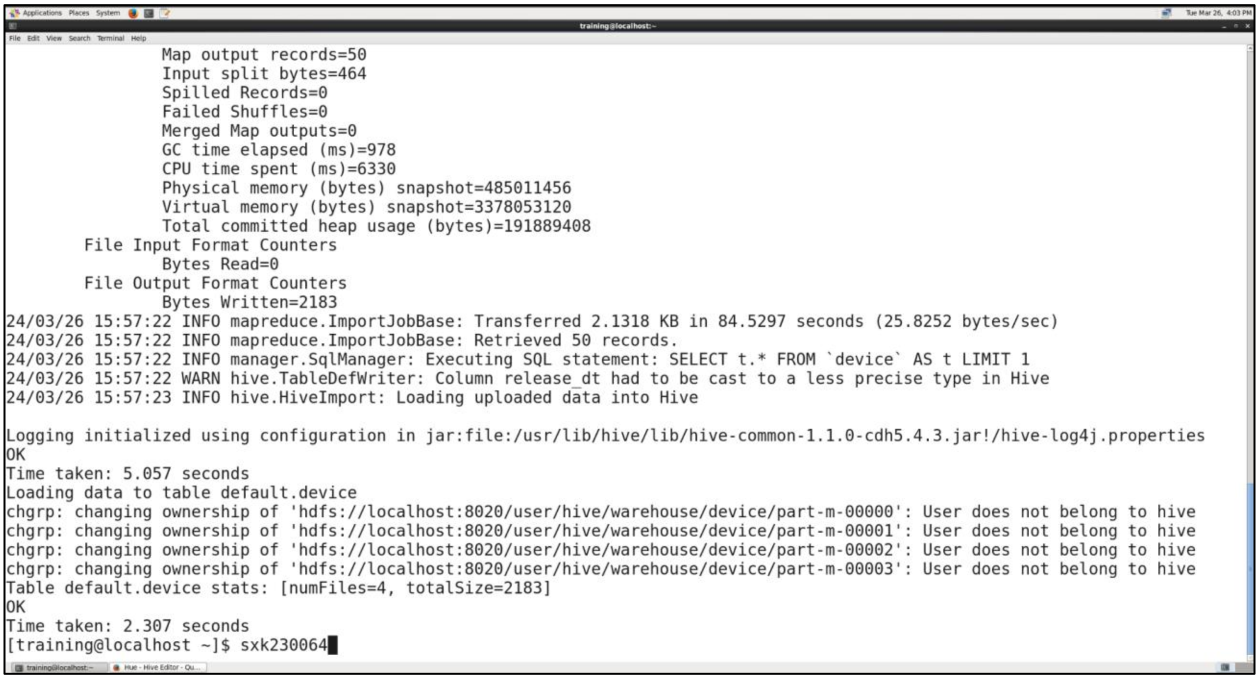


```
SELECT * FROM webpage WHERE name LIKE "ifruit%"
sxk230064
```

	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

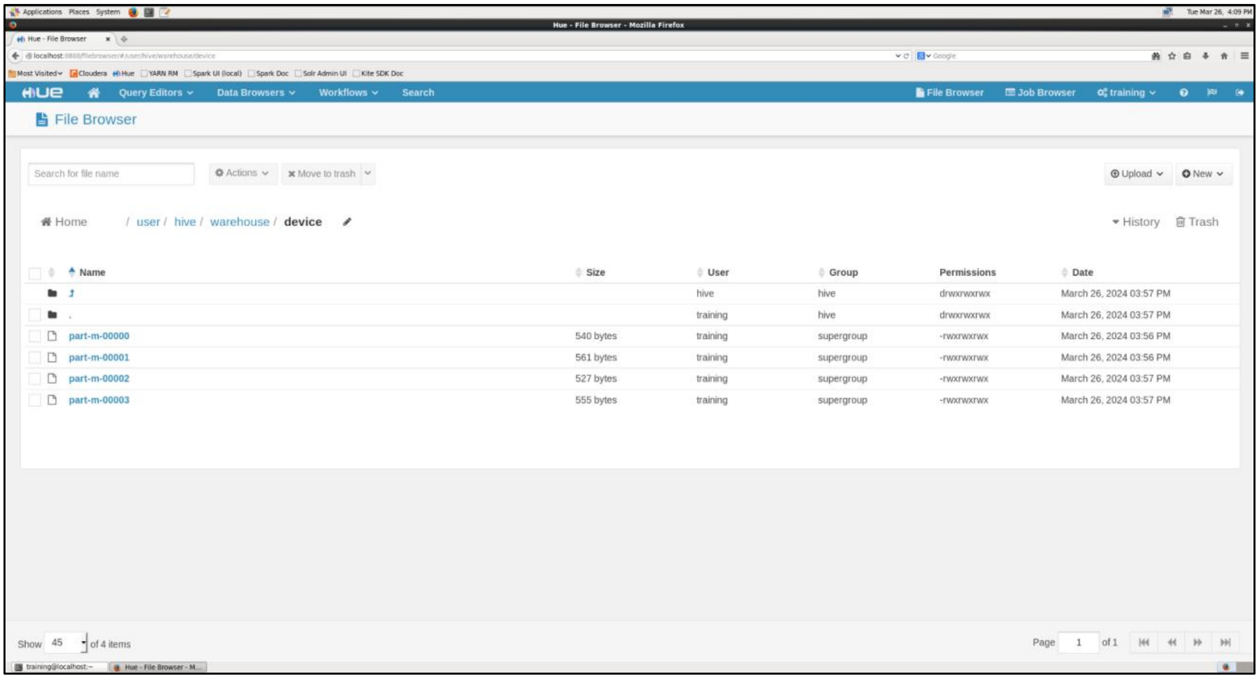

Applications Places System
Tue Mar 26, 4:02 PM
training@localhost:~
File Edit View Search Terminal Help
training@localhost:~
Hue - Hive Editor - Qu...

```
[training@localhost ~]$ sudo service zookeeper-server start
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Starting zookeeper ... already running as process 1694.
[training@localhost ~]$ sudo service hive-server2 start
Started Hive Server2 (hive-server2):                       [  OK  ]
[training@localhost ~]$ sqoop import \
>  --connect jdbc:mysql://localhost/loudacre \
>  --username training --password training \
>  --fields-terminated-by '\t' \
>  --table device \
>  --hive-import
24/03/26 15:55:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/03/26 15:55:46 WARN tool.BaseSqoopTool: Setting your password on the command-line is
 insecure. Consider using -P instead.
24/03/26 15:55:47 INFO manager.MySQLManager: Preparing to use a MySQL streaming results
et.
24/03/26 15:55:47 INFO tool.CodeGenTool: Beginning code generation
24/03/26 15:55:48 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `de
vice` AS t LIMIT 1
24/03/26 15:55:48 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `de
vice` AS t LIMIT 1
24/03/26 15:55:48 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-ma
preduce
Note: /tmp/sqoop-training/compile/c324fb8a8e00915bc747f0ed7adf9641/device.java uses or
overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/03/26 15:55:54 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/co
mpile/c324fb8a8e00915bc747f0ed7adf9641/device.jar
24/03/26 15:55:54 WARN manager.MySQLManager: It looks like you are importing from mysql
.
24/03/26 15:55:54 WARN manager.MySQLManager: This transfer can be faster! Use the --dir
```


Applications Places System
Tue Mar 26, 4:03 PM
training@localhost:~
File Edit View Search Terminal Help
training@localhost:~
Hue - Hive Editor - Qu...

```
                Map output records=50
                Input split bytes=464
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=978
                CPU time spent (ms)=6330
                Physical memory (bytes) snapshot=485011456
                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
24/03/26 15:57:22 INFO mapreduce.ImportJobBase: Transferred 2.1318 KB in 84.5297 seconds (25.8252 bytes/sec)
24/03/26 15:57:22 INFO mapreduce.ImportJobBase: Retrieved 50 records.
24/03/26 15:57:22 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `device` AS t LIMIT 1
24/03/26 15:57:22 WARN hive.TableDefWriter: Column release_dt had to be cast to a less precise type in Hive
24/03/26 15:57:23 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: 5.057 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: 2.307 seconds
[training@localhost ~]$ sxk230064
```

8. Utilized HUE to navigate to the specific data location (/user/hive/warehouse/device), to review the imported data files residing in the default hive warehouse.

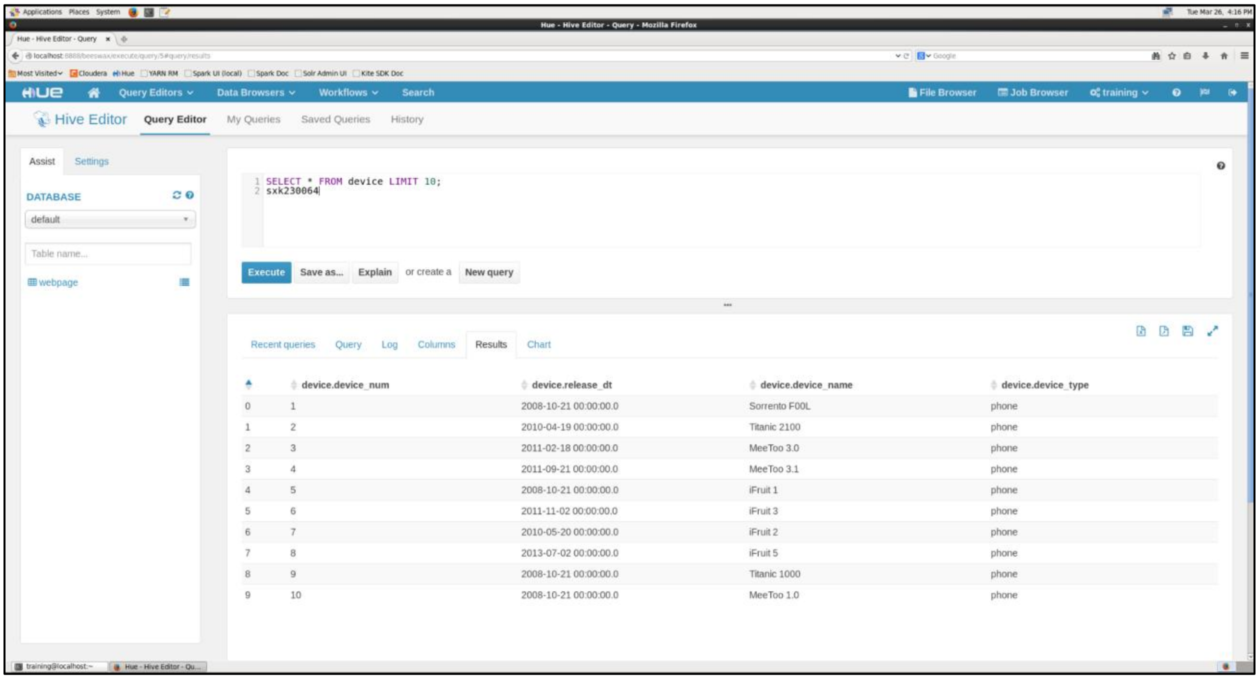


File Browser

Home / user / hive / warehouse / device

Name	Size	User	Group	Permissions	Date
↥		hive	hive	drwxrwxrwx	March 26, 2024 03:57 PM
.		training	hive	drwxrwxrwx	March 26, 2024 03:57 PM
part-m-00000	540 bytes	training	supergroup	-rwxrwxrwx	March 26, 2024 03:56 PM
part-m-00001	561 bytes	training	supergroup	-rwxrwxrwx	March 26, 2024 03:56 PM
part-m-00002	527 bytes	training	supergroup	-rwxrwxrwx	March 26, 2024 03:57 PM
part-m-00003	555 bytes	training	supergroup	-rwxrwxrwx	March 26, 2024 03:57 PM

9. Ran a test query to view all columns of the device table.

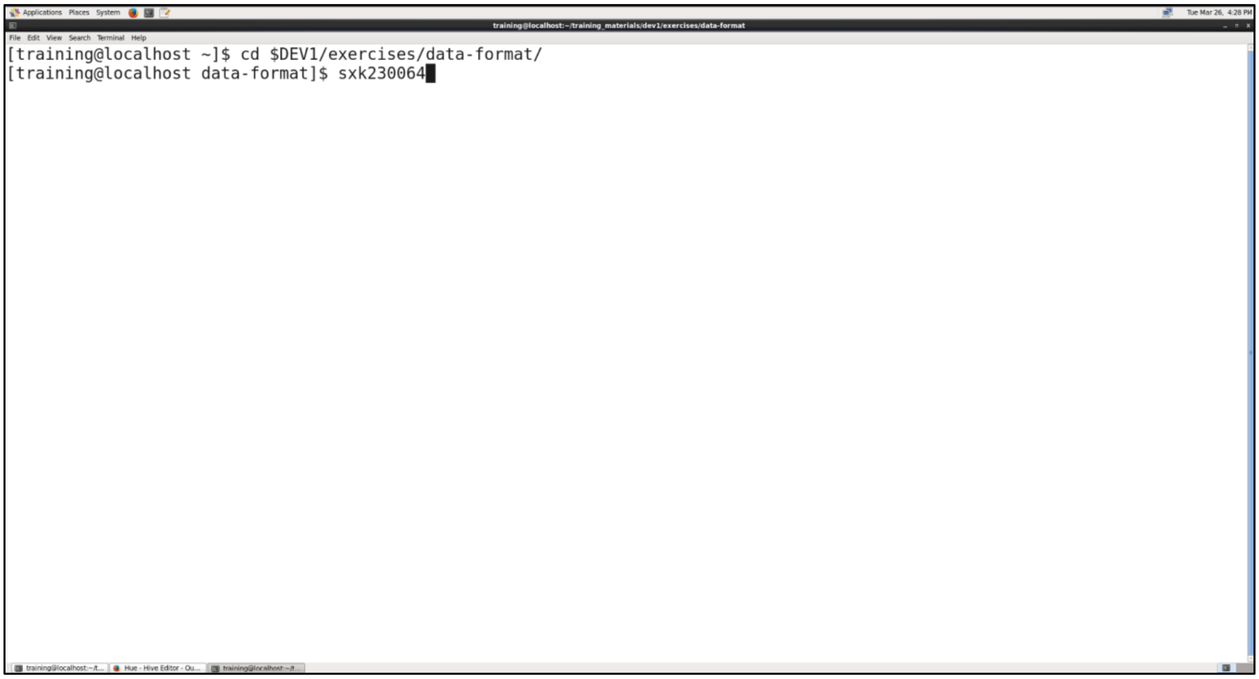


```
SELECT * FROM device LIMIT 10;
sxk230064
```

	device.device_num	device.release_dt	device.device_name	device.device_type
0	1	2008-10-21 00:00:00.0	Sorrento F00L	phone
1	2	2010-04-19 00:00:00.0	Titanic 2100	phone
2	3	2011-02-18 00:00:00.0	MeeToo 3.0	phone
3	4	2011-09-21 00:00:00.0	MeeToo 3.1	phone
4	5	2008-10-21 00:00:00.0	iFruit 1	phone
5	6	2011-11-02 00:00:00.0	iFruit 3	phone
6	7	2010-05-20 00:00:00.0	iFruit 2	phone
7	8	2013-07-02 00:00:00.0	iFruit 5	phone
8	9	2008-10-21 00:00:00.0	Titanic 1000	phone
9	10	2008-10-21 00:00:00.0	MeeToo 1.0	phone

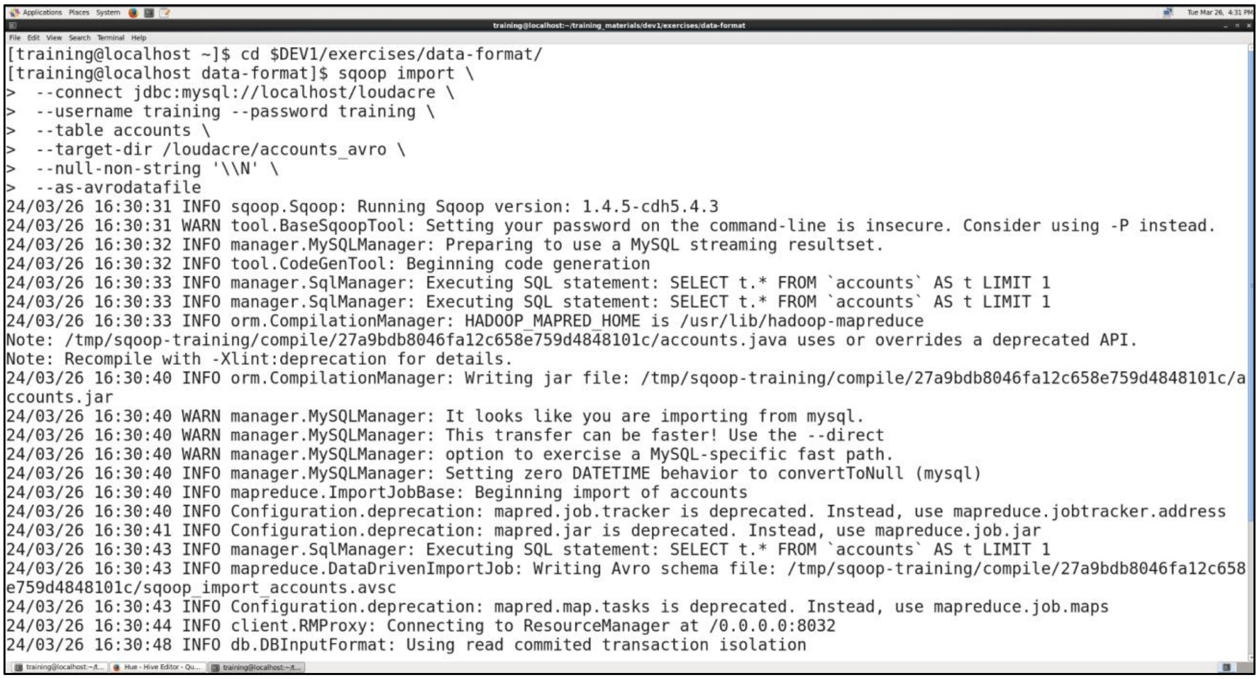
CHAPTER 7 - SELECT A FORMAT FOR A DATA FILE

1. I used the 'cd' command to navigate to the exercise directory.

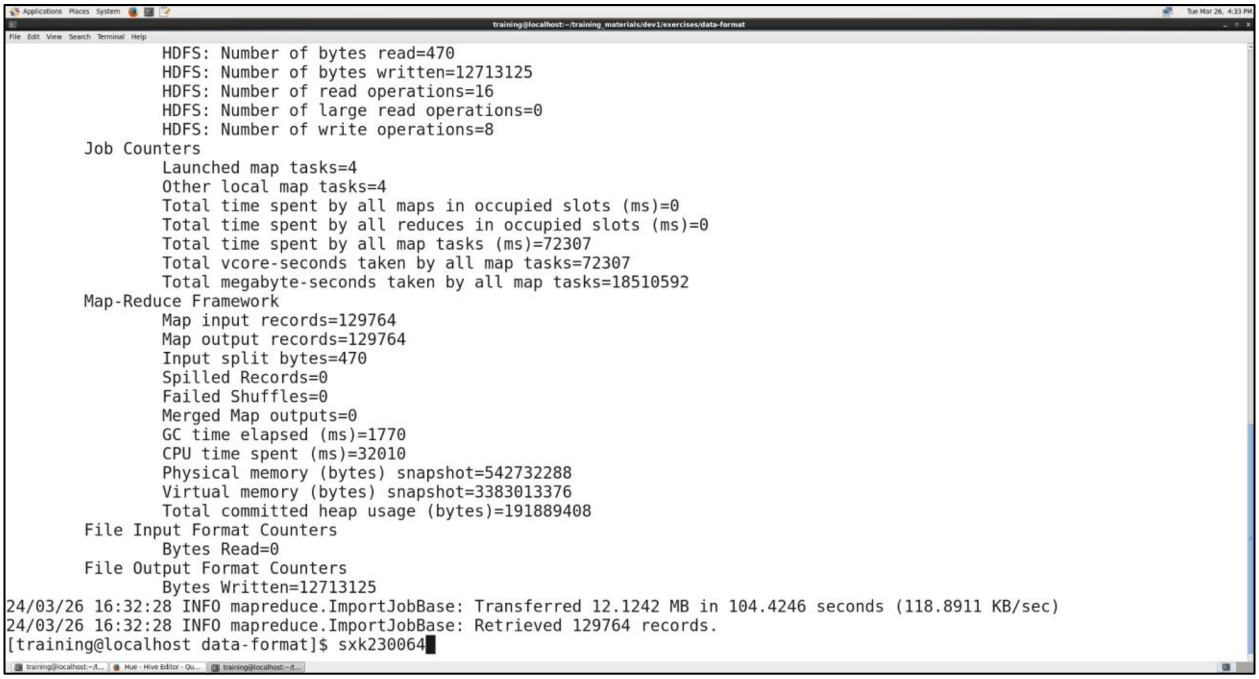

Applications Places System
training@localhost:~/training_materials/dev1/exercises/data-format
File Edit View Search Terminal Help
Tue Mar 26, 4:28 PM

```
[training@localhost ~]$ cd $DEV1/exercises/data-format/
[training@localhost data-format]$ sxk230064
```

2. Utilizing the Sqoop import command, I imported the accounts table into an Avro data format.

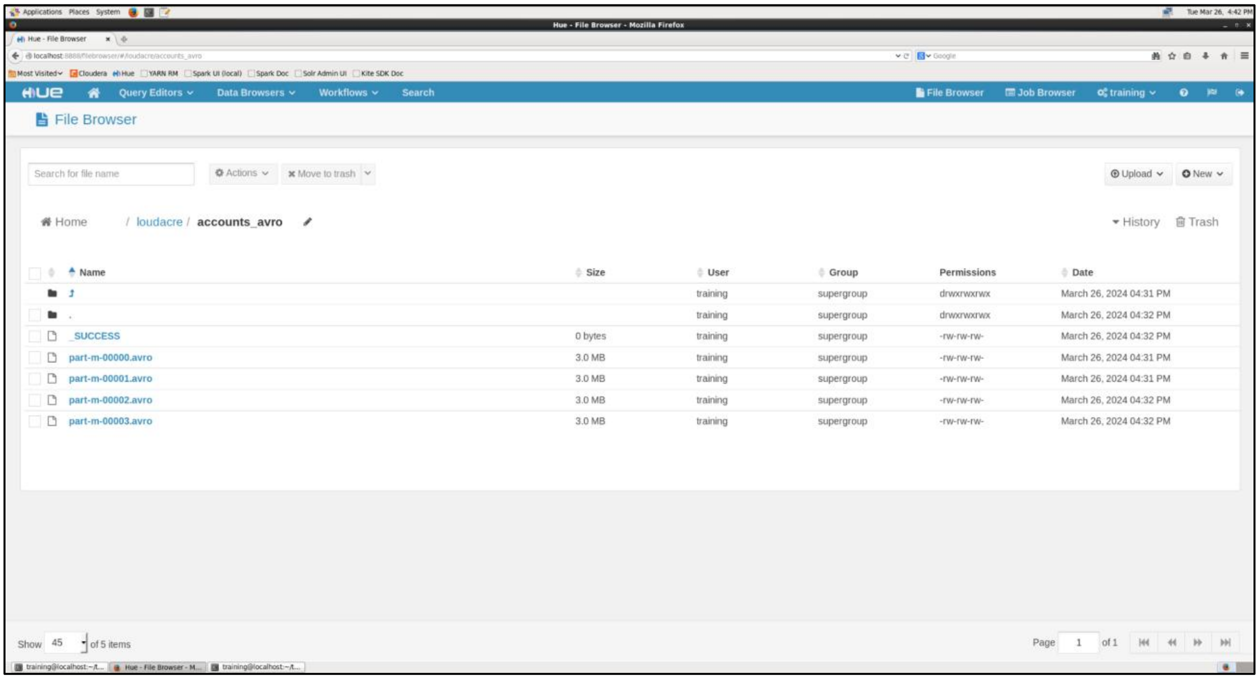

Applications Places System
training@localhost:~/training_materials/dev1/exercises/data-format
File Edit View Search Terminal Help
Tue Mar 26, 4:31 PM

```
[training@localhost ~]$ cd $DEV1/exercises/data-format/
[training@localhost data-format]$ sqoop import \
>  --connect jdbc:mysql://localhost/loudacre \
>  --username training --password training \
>  --table accounts \
>  --target-dir /loudacre/accounts_avro \
>  --null-non-string '\\N' \
>  --as-avrodatafile
24/03/26 16:30:31 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/03/26 16:30:31 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/03/26 16:30:32 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/03/26 16:30:32 INFO tool.CodeGenTool: Beginning code generation
24/03/26 16:30:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
24/03/26 16:30:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
24/03/26 16:30:33 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/27a9bdb8046fa12c658e759d4848101c/accounts.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/03/26 16:30:40 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/27a9bdb8046fa12c658e759d4848101c/a
ccounts.jar
24/03/26 16:30:40 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/03/26 16:30:40 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/03/26 16:30:40 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/03/26 16:30:40 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/03/26 16:30:40 INFO mapreduce.ImportJobBase: Beginning import of accounts
24/03/26 16:30:40 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/03/26 16:30:41 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/03/26 16:30:43 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `accounts` AS t LIMIT 1
24/03/26 16:30:43 INFO mapreduce.DataDrivenImportJob: Writing Avro schema file: /tmp/sqoop-training/compile/27a9bdb8046fa12c658
e759d4848101c/sqoop_import_accounts.avsc
24/03/26 16:30:43 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/03/26 16:30:44 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
24/03/26 16:30:48 INFO db.DBInputFormat: Using read commited transaction isolation
```


Applications Places System
training@localhost:~/training_materials/dev1/exercises/data-format
File Edit View Search Terminal Help
Tue Mar 26, 4:33 PM

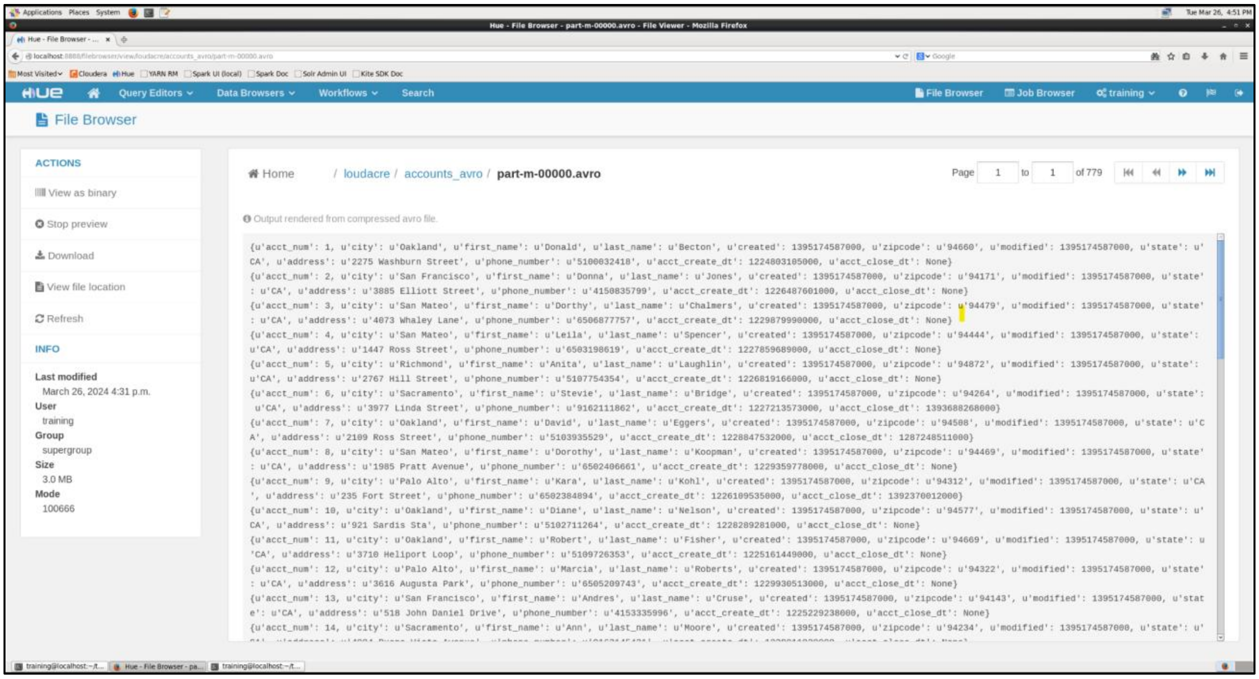
```
                HDFS: Number of bytes read=470
                HDFS: Number of bytes written=12713125
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters
                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)=72307
                Total vcore-seconds taken by all map tasks=72307
                Total megabyte-seconds taken by all map tasks=18510592
        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)=1770
                CPU time spent (ms)=32010
                Physical memory (bytes) snapshot=542732288
                Virtual memory (bytes) snapshot=3383013376
                Total committed heap usage (bytes)=191889408
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=12713125
24/03/26 16:32:28 INFO mapreduce.ImportJobBase: Transferred 12.1242 MB in 104.4246 seconds (118.8911 KB/sec)
24/03/26 16:32:28 INFO mapreduce.ImportJobBase: Retrieved 129764 records.
[training@localhost data-format]$ sxk230064
```

3. Then, I accessed the file browser to view the imported files by Sqoop into HDFS.


Applications Places System
Hue - File Browser - Mozilla Firefox
Hue - File Browser
localhost:8888/filebrowser/#/loudacre/accounts_avro
Most Visited Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc
HUE Query Editors Data Browsers Workflows Search
File Browser Job Browser training
File Browser
Search for file name
Actions
Move to trash
Upload
New
Home / loudacre / accounts_avro
History
Trash
Show 45 of 5 items
Page 1 of 1

Name	Size	User	Group	Permissions	Date
⤴		training	supergroup	drwxrwxrwx	March 26, 2024 04:31 PM
.		training	supergroup	drwxrwxrwx	March 26, 2024 04:32 PM
_SUCCESS	0 bytes	training	supergroup	-rw-rw-rw-	March 26, 2024 04:32 PM
part-m-00000.avro	3.0 MB	training	supergroup	-rw-rw-rw-	March 26, 2024 04:31 PM
part-m-00001.avro	3.0 MB	training	supergroup	-rw-rw-rw-	March 26, 2024 04:31 PM
part-m-00002.avro	3.0 MB	training	supergroup	-rw-rw-rw-	March 26, 2024 04:32 PM
part-m-00003.avro	3.0 MB	training	supergroup	-rw-rw-rw-	March 26, 2024 04:32 PM

4. Upon locating the file part-m-00000.avro, I proceeded to view its contents, experiencing a lengthy loading time which may be due to large number of records.



File Browser

Home / loudacre / accounts_avro / part-m-00000.avro

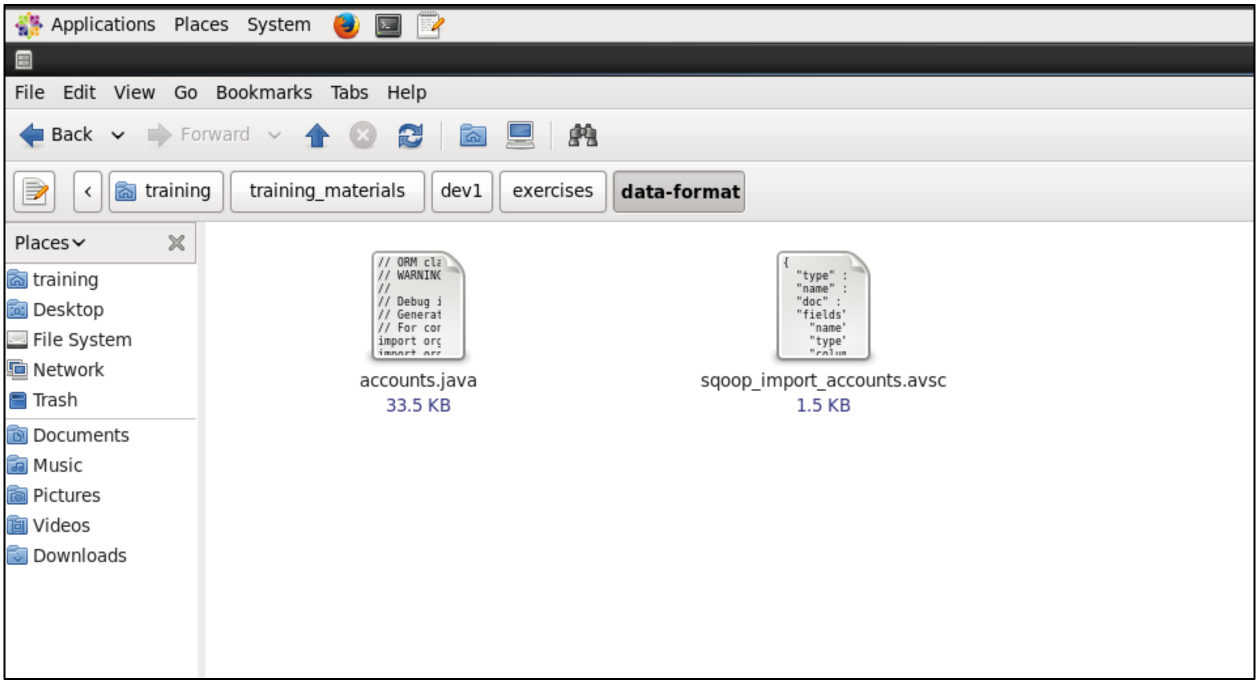
Page 1 to 1 of 779

Last modified
March 26, 2024 4:31 p.m.
User
training
Group
supergroup
Size
3.0 MB
Mode
100666

Output rendered from compressed avro file.

```
{u'acct_num': 1, u'city': u'Oakland', u'first_name': u'Donald', u'last_name': u'Becton', u'created': 1395174587000, u'zipcode': u'94660', u'modified': 1395174587000, u'state': u'
CA', u'address': u'2275 Washburn Street', u'phone_number': u'5100032418', u'acct_create_dt': 1224803105000, u'acct_close_dt': None}
{u'acct_num': 2, u'city': u'San Francisco', u'first_name': u'Donna', u'last_name': u'Jones', u'created': 1395174587000, u'zipcode': u'94171', u'modified': 1395174587000, u'state'
: u'CA', u'address': u'3885 Elliott Street', u'phone_number': u'4150835799', u'acct_create_dt': 1226487601000, u'acct_close_dt': None}
{u'acct_num': 3, u'city': u'San Mateo', u'first_name': u'Dorthy', u'last_name': u'Chalmers', u'created': 1395174587000, u'zipcode': u'94479', u'modified': 1395174587000, u'state'
: u'CA', u'address': u'4073 Whaley Lane', u'phone_number': u'6506877757', u'acct_create_dt': 1229879990000, u'acct_close_dt': None}
{u'acct_num': 4, u'city': u'San Mateo', u'first_name': u'Leila', u'last_name': u'Spencer', u'created': 1395174587000, u'zipcode': u'94444', u'modified': 1395174587000, u'state':
u'CA', u'address': u'1447 Ross Street', u'phone_number': u'6503198619', u'acct_create_dt': 1227859689000, u'acct_close_dt': None}
{u'acct_num': 5, u'city': u'Richmond', u'first_name': u'Anita', u'last_name': u'Laughlin', u'created': 1395174587000, u'zipcode': u'94872', u'modified': 1395174587000, u'state':
u'CA', u'address': u'2767 Hill Street', u'phone_number': u'5107754354', u'acct_create_dt': 1226819166000, u'acct_close_dt': None}
{u'acct_num': 6, u'city': u'Sacramento', u'first_name': u'Stevie', u'last_name': u'Bridge', u'created': 1395174587000, u'zipcode': u'94264', u'modified': 1395174587000, u'state':
 u'CA', u'address': u'3977 Linda Street', u'phone_number': u'9162111862', u'acct_create_dt': 1227213573000, u'acct_close_dt': 1393688268000}
{u'acct_num': 7, u'city': u'Oakland', u'first_name': u'David', u'last_name': u'Eggers', u'created': 1395174587000, u'zipcode': u'94508', u'modified': 1395174587000, u'state': u'C
A', u'address': u'2109 Ross Street', u'phone_number': u'5103935529', u'acct_create_dt': 1228847532000, u'acct_close_dt': 1287248511000}
{u'acct_num': 8, u'city': u'San Mateo', u'first_name': u'Dorothy', u'last_name': u'Koopman', u'created': 1395174587000, u'zipcode': u'94469', u'modified': 1395174587000, u'state'
: u'CA', u'address': u'1985 Pratt Avenue', u'phone_number': u'6502406661', u'acct_create_dt': 1229359778000, u'acct_close_dt': None}
{u'acct_num': 9, u'city': u'Palo Alto', u'first_name': u'Kara', u'last_name': u'Kohl', u'created': 1395174587000, u'zipcode': u'94312', u'modified': 1395174587000, u'state': u'CA
', u'address': u'235 Fort Street', u'phone_number': u'6502384894', u'acct_create_dt': 1226109535000, u'acct_close_dt': 1392370012000}
{u'acct_num': 10, u'city': u'Oakland', u'first_name': u'Diane', u'last_name': u'Nelson', u'created': 1395174587000, u'zipcode': u'94577', u'modified': 1395174587000, u'state': u'
CA', u'address': u'921 Sardis Sta', u'phone_number': u'5102711264', u'acct_create_dt': 1228289281000, u'acct_close_dt': None}
{u'acct_num': 11, u'city': u'Oakland', u'first_name': u'Robert', u'last_name': u'Fisher', u'created': 1395174587000, u'zipcode': u'94669', u'modified': 1395174587000, u'state': u
'CA', u'address': u'3710 Heliport Loop', u'phone_number': u'5109726353', u'acct_create_dt': 1225161449000, u'acct_close_dt': None}
{u'acct_num': 12, u'city': u'Palo Alto', u'first_name': u'Marcia', u'last_name': u'Roberts', u'created': 1395174587000, u'zipcode': u'94322', u'modified': 1395174587000, u'state'
: u'CA', u'address': u'3616 Augusta Park', u'phone_number': u'6505209743', u'acct_create_dt': 1229930513000, u'acct_close_dt': None}
{u'acct_num': 13, u'city': u'San Francisco', u'first_name': u'Andres', u'last_name': u'Cruse', u'created': 1395174587000, u'zipcode': u'94143', u'modified': 1395174587000, u'stat
e': u'CA', u'address': u'518 John Daniel Drive', u'phone_number': u'4153335996', u'acct_create_dt': 1225229238000, u'acct_close_dt': None}
{u'acct_num': 14, u'city': u'Sacramento', u'first_name': u'Ann', u'last_name': u'Moore', u'created': 1395174587000, u'zipcode': u'94234', u'modified': 1395174587000, u'state': u'
```

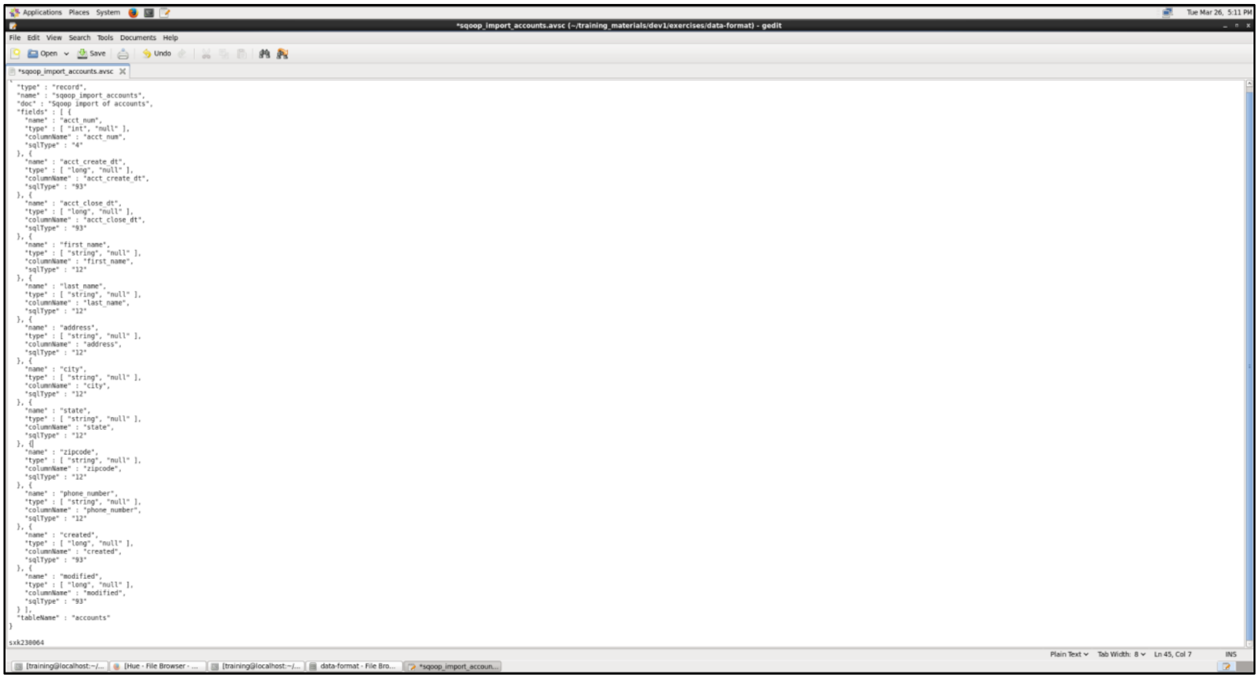
5. Sqoop automatically generated a schema named "sqoop_import_accounts.avsc" in the current directory.



accounts.java
33.5 KB

sqoop_import_accounts.avsc
1.5 KB

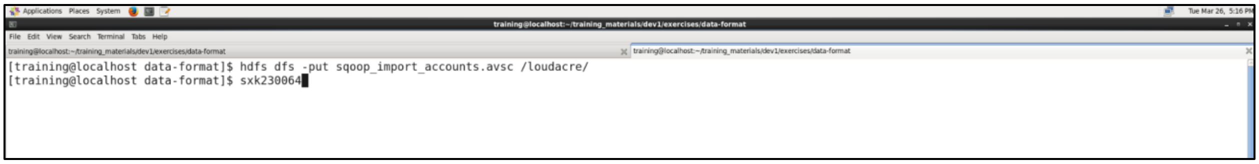
6. To examine the contents of the generated schema file, I accessed and viewed it accordingly.



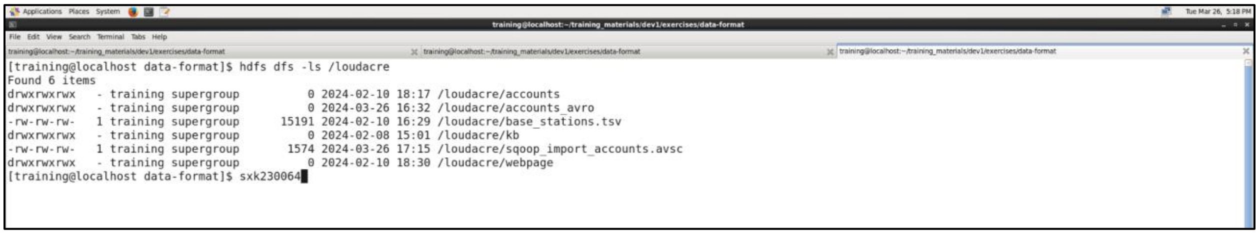
*sqoop_import_accounts.avsc (~/training_materials/dev1/exercises/data-format) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
*sqoop_import_accounts.avsc
Plain Text Tab Width: 8 Ln 45, Col 7 INS

```
{
  "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",
    "sqlType" : "12"
  }, {
    "name" : "state",
    "type" : [ "string", "null" ],
    "columnName" : "state",
    "sqlType" : "12"
  }, {
    "name" : "zipcode",
    "type" : [ "string", "null" ],
    "columnName" : "zipcode",
    "sqlType" : "12"
  }, {
    "name" : "phone_number",
    "type" : [ "string", "null" ],
    "columnName" : "phone_number",
    "sqlType" : "12"
  }, {
    "name" : "created",
    "type" : [ "long", "null" ],
    "columnName" : "created",
    "sqlType" : "93"
  }, {
    "name" : "modified",
    "type" : [ "long", "null" ],
    "columnName" : "modified",
    "sqlType" : "93"
  } ],
  "tableName" : "accounts"
}

sxk230064
```

7. Employing the '-put' command, I copied the schema file to the /loudacre directory in HDFS. Utilizing the '-ls' command, I verified the successful transfer.

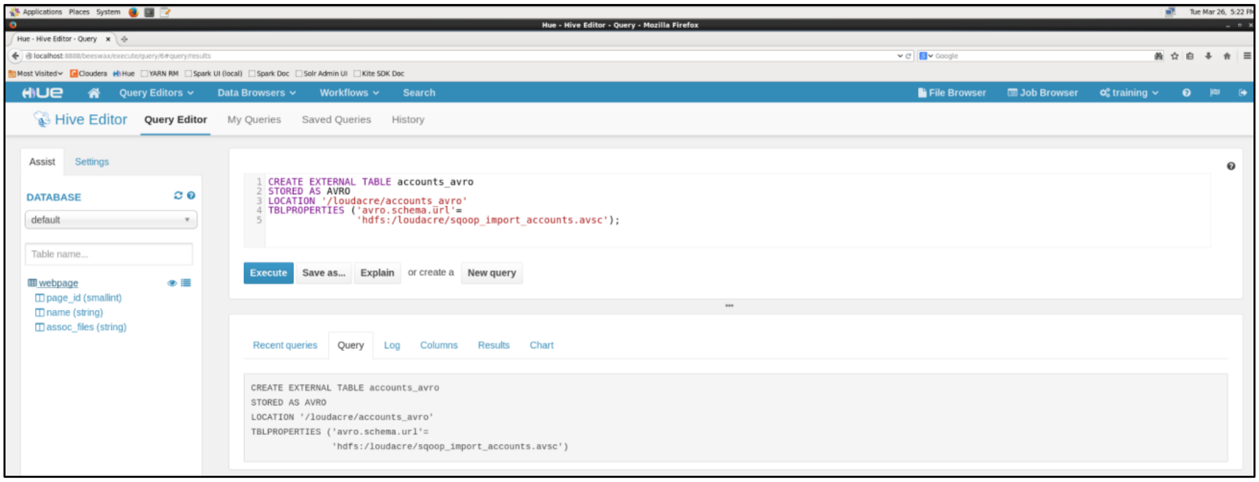


```
[training@localhost data-format]$ hdfs dfs -put sqoop_import_accounts.avsc /loudacre/
[training@localhost data-format]$ sxk230064
```



```
[training@localhost data-format]$ hdfs dfs -ls /loudacre
Found 6 items
drwxrwxrwx   - training supergroup          0 2024-02-10 18:17 /loudacre/accounts
drwxrwxrwx   - training supergroup          0 2024-03-26 16:32 /loudacre/accounts_avro
-rw-rw-rw-   1 training supergroup      15191 2024-02-10 16:29 /loudacre/base_stations.tsv
drwxrwxrwx   - training supergroup          0 2024-02-08 15:01 /loudacre/kb
-rw-rw-rw-   1 training supergroup       1574 2024-03-26 17:15 /loudacre/sqoop_import_accounts.avsc
drwxrwxrwx   - training supergroup          0 2024-02-10 18:30 /loudacre/webpage
[training@localhost data-format]$ sxk230064
```

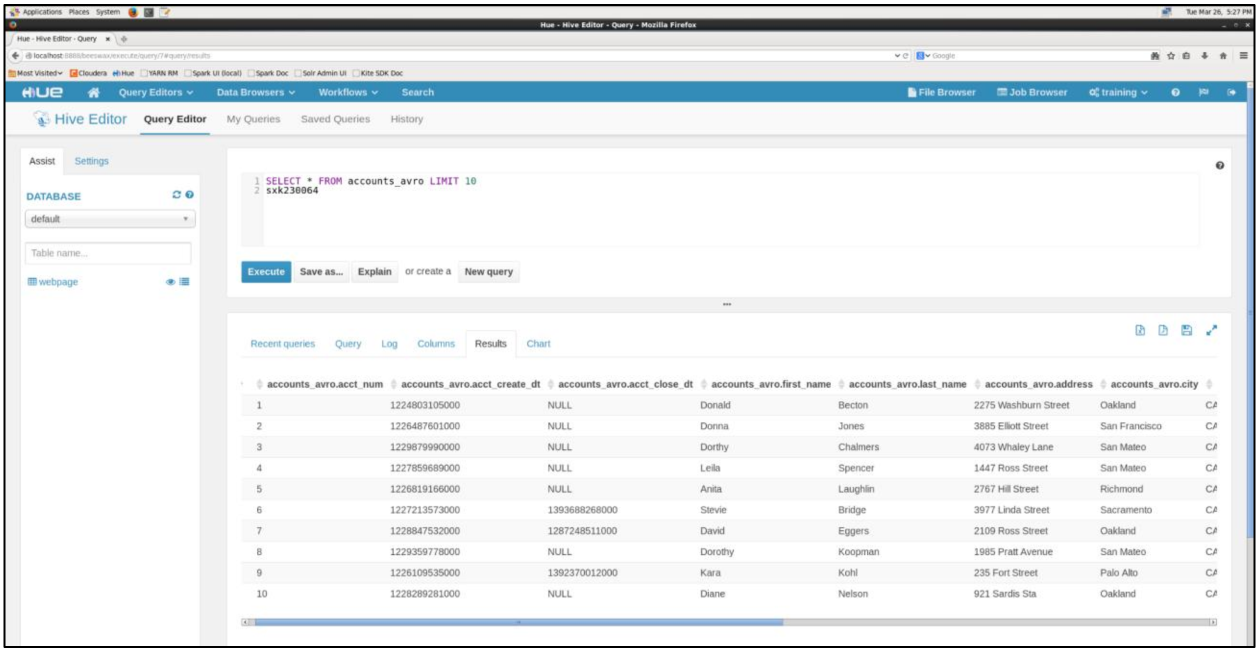
8. Within Hive, I proceeded to create the table "accounts_avro" utilizing the provided Avro schema.



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

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

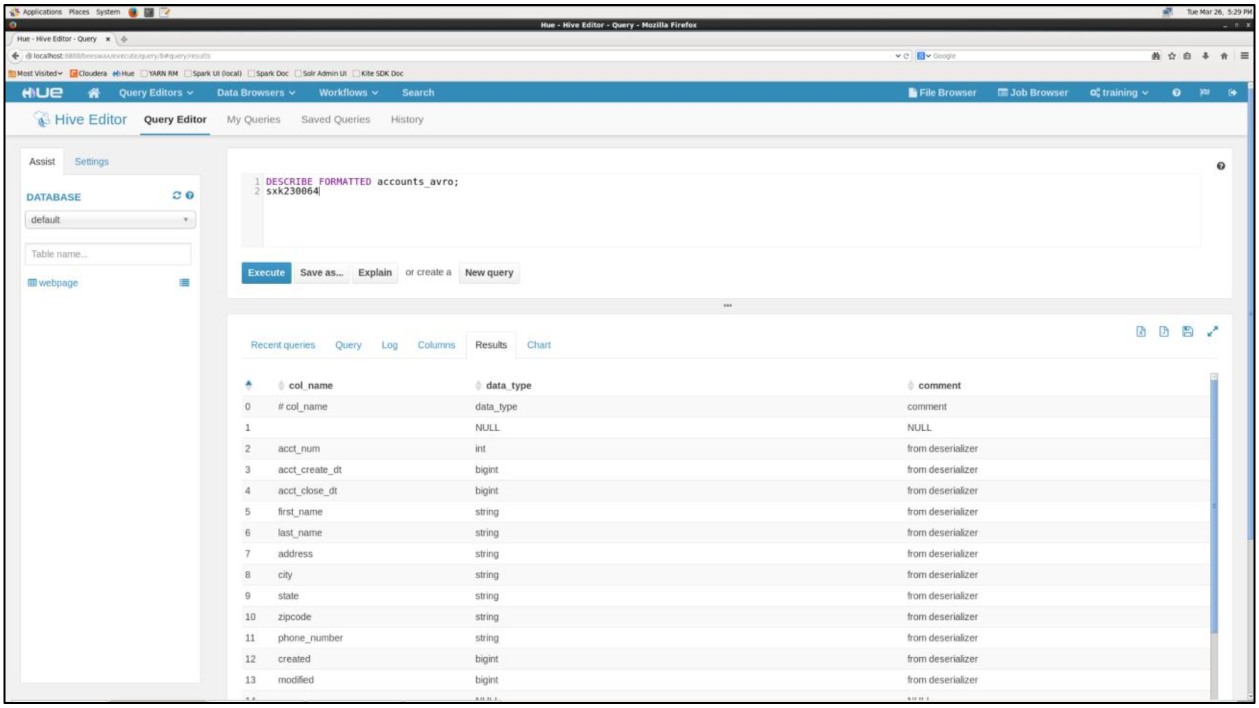
9. After creating the table, I performed a test query using a select statement to ensure its successful created.



```
SELECT * FROM accounts_avro LIMIT 10
sxk230064
```

accounts_avro.acct_num	accounts_avro.acct_create_dt	accounts_avro.acct_close_dt	accounts_avro.first_name	accounts_avro.last_name	accounts_avro.address	accounts_avro.city	
1	1224803105000	NULL	Donald	Becton	2275 Washburn Street	Oakland	CA
2	1226487601000	NULL	Donna	Jones	3885 Elliott Street	San Francisco	CA
3	1229879990000	NULL	Dorthy	Chalmers	4073 Whaley Lane	San Mateo	CA
4	1227859689000	NULL	Leila	Spencer	1447 Ross Street	San Mateo	CA
5	1226819166000	NULL	Anita	Laughlin	2767 Hill Street	Richmond	CA
6	1227213573000	1393688268000	Stevie	Bridge	3977 Linda Street	Sacramento	CA
7	1228847532000	1287248511000	David	Eggers	2109 Ross Street	Oakland	CA
8	1229359778000	NULL	Dorothy	Koopman	1985 Pratt Avenue	San Mateo	CA
9	1226109535000	1392370012000	Kara	Kohl	235 Fort Street	Palo Alto	CA
10	1228289281000	NULL	Diane	Nelson	921 Sardis Sta	Oakland	CA

10. I utilized the 'DESCRIBE FORMATTED' command to list the columns and data types of the accounts_avro table, derived from the Avro schema.

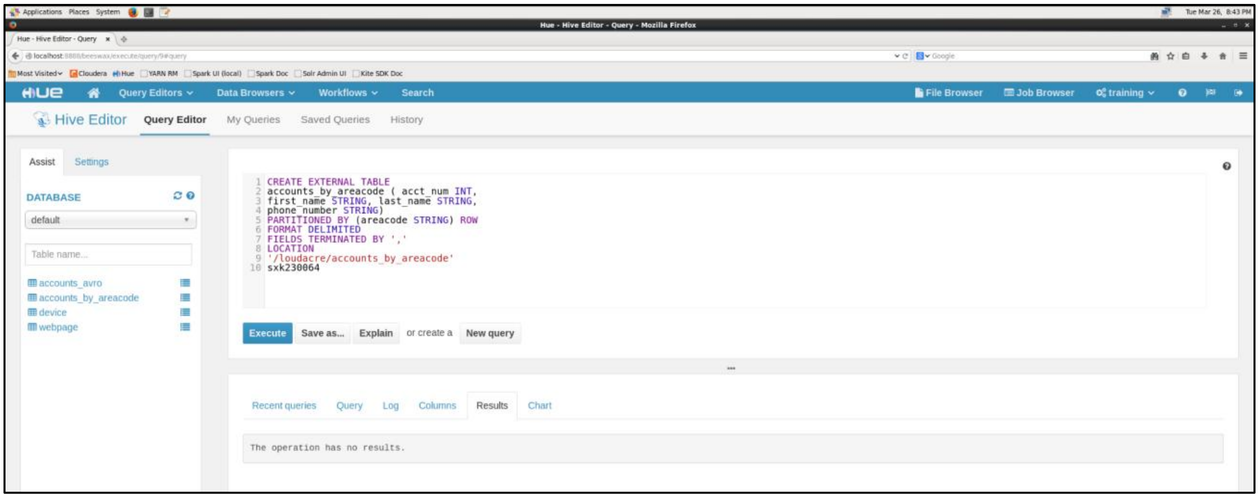


```
DESCRIBE FORMATTED accounts_avro;
sxk230064
```

	col_name	data_type	comment
0	# col_name	data_type	comment
1		NULL	NULL
2	acct_num	int	from deserializer
3	acct_create_dt	bigint	from deserializer
4	acct_close_dt	bigint	from deserializer
5	first_name	string	from deserializer
6	last_name	string	from deserializer
7	address	string	from deserializer
8	city	string	from deserializer
9	state	string	from deserializer
10	zipcode	string	from deserializer
11	phone_number	string	from deserializer
12	created	bigint	from deserializer
13	modified	bigint	from deserializer

CHAPTER 8 - PARTITION DATA IN IMPALA OR HIVE

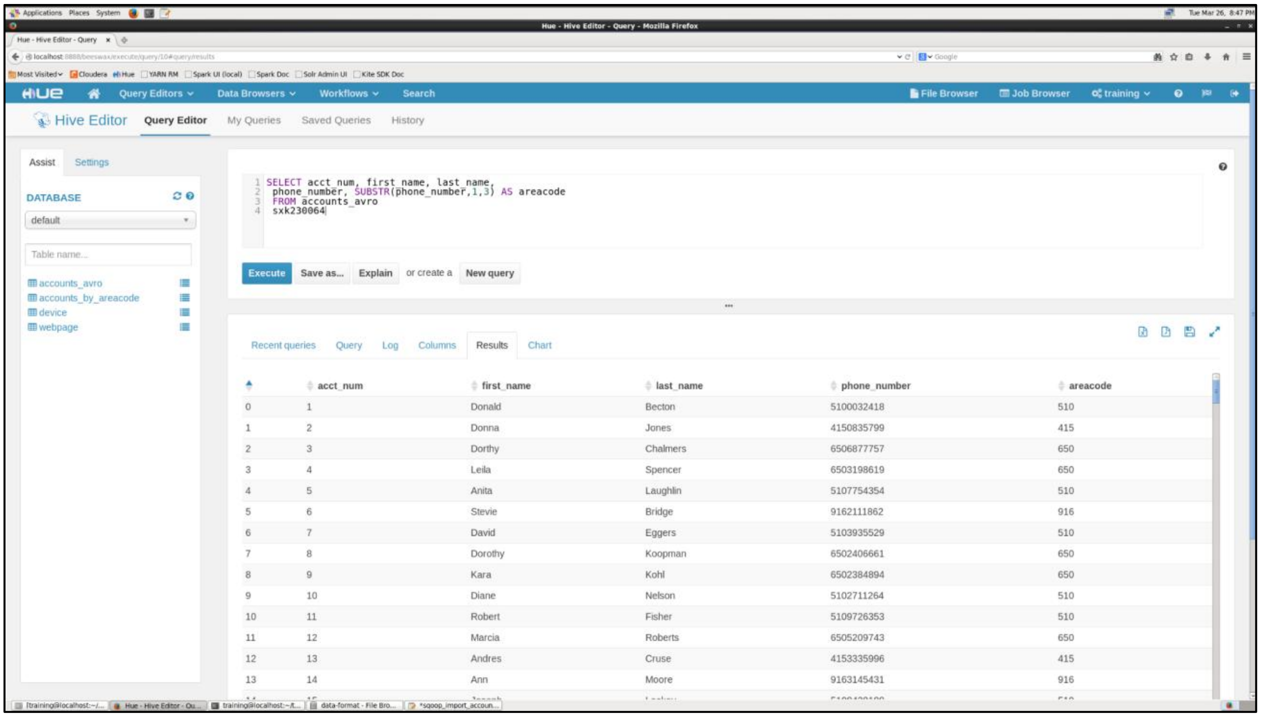
1. Initially, I created a new empty table in Hive using the CREATE EXTERNAL TABLE statement.



```
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'
sxk230064
```

```
The operation has no results.
```

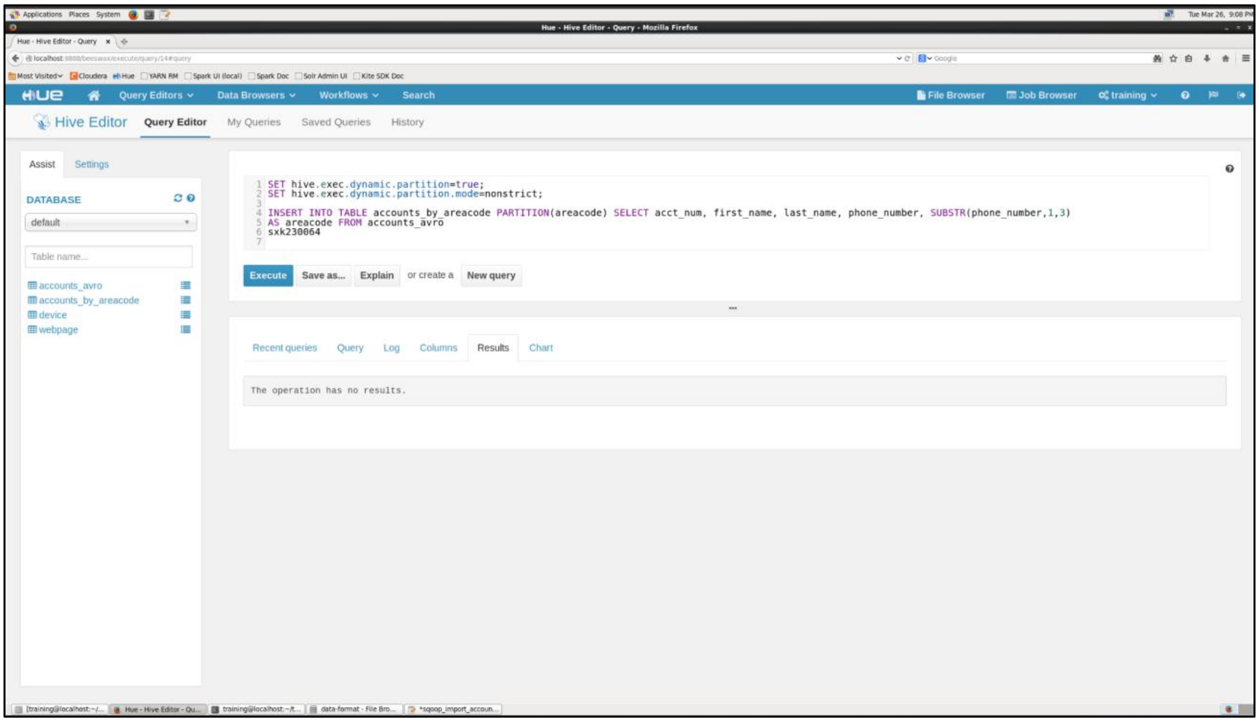
2. To extract the area code from phone numbers, I executed a query as per the below snippet.



```
SELECT acct_num, first_name, last_name,
 phone_number, SUBSTR(phone_number,1,3) AS areacode
 FROM accounts_avro
 sxk230064
```

	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
6	7	David	Eggers	5103935529	510
7	8	Dorothy	Koopman	6502406661	650
8	9	Kara	Kohl	6502384894	650
9	10	Diane	Nelson	5102711264	510
10	11	Robert	Fisher	5109726353	510
11	12	Marcia	Roberts	6505209743	650
12	13	Andres	Cruse	4153335996	415
13	14	Ann	Moore	9163145431	916

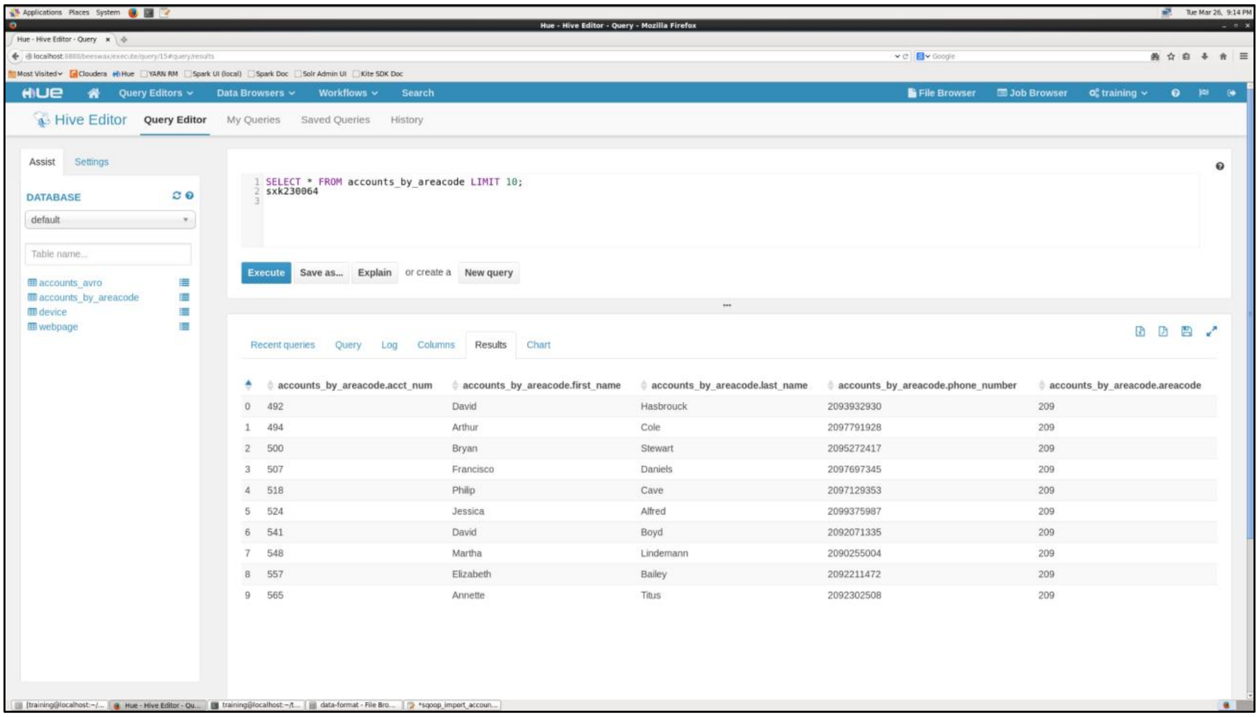
3. Then, I employed the SELECT statement within an INSERT INTO TABLE command to transfer the specified columns into the newly created table. Notably, the process involved dynamic partitioning by area code, as depicted in the following screenshot.



```
SET hive.exec.dynamic.partition=true;
SET hive.exec.dynamic.partition.mode=nonstrict;

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
sxk230064

```

```
The operation has no results.
```

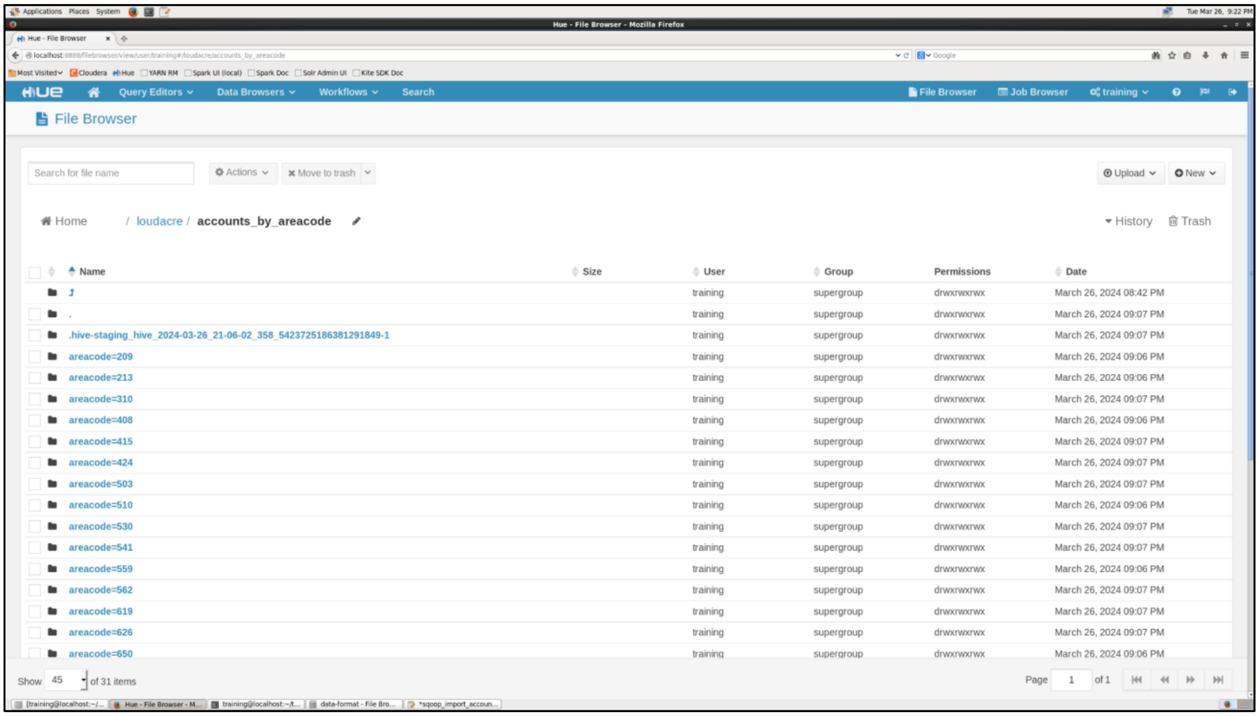
4. I ran a test query to ensure that the table was populated correctly.


Applications Places System
Hue - Hive Editor - Query - Mozilla Firefox
Hue - Hive Editor - Query
localhost:8888/beeswax/execute/query/15#query/results
Most Visited Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc
HUE Query Editors Data Browsers Workflows Search
File Browser Job Browser training
Hive Editor Query Editor My Queries Saved Queries History
Assist Settings
DATABASE
default
Table name...
accounts_avro
accounts_by_areacode
device
webpage
Execute Save as... Explain or create a New query
Recent queries Query Log Columns Results Chart
Tue Mar 26, 9:14 PM

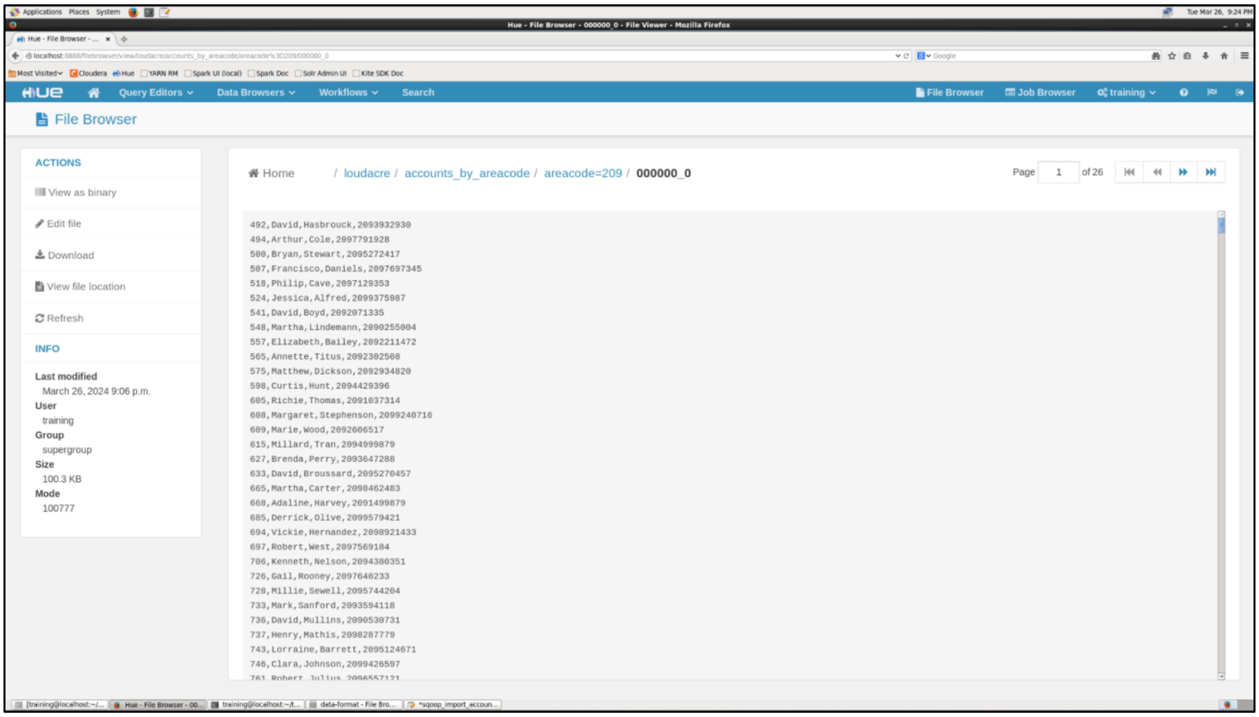
```
SELECT * FROM accounts_by_areacode LIMIT 10;
sxk230064
```

	accounts_by_areacode.acct_num	accounts_by_areacode.first_name	accounts_by_areacode.last_name	accounts_by_areacode.phone_number	accounts_by_areacode.areacode
0	492	David	Hasbrouck	2093932930	209
1	494	Arthur	Cole	2097791928	209
2	500	Bryan	Stewart	2095272417	209
3	507	Francisco	Daniels	2097697345	209
4	518	Philip	Cave	2097129353	209
5	524	Jessica	Alfred	2099375987	209
6	541	David	Boyd	2092071335	209
7	548	Martha	Lindemann	2090255004	209
8	557	Elizabeth	Bailey	2092211472	209
9	565	Annette	Titus	2092302508	209

5. Using Hue, I confirm that the index structure of the accounts_by_areacode table encompassed partition directories. Additionally, I reviewed the data within the directories to affirm that the partitioning is correct.


Applications Places System
Hue - File Browser - Mozilla Firefox
Hue - File Browser
localhost:8888/filebrowser/view/user/training#/loudacre/accounts_by_areacode
Most Visited Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc
HUE Query Editors Data Browsers Workflows Search
File Browser Job Browser training
File Browser
Search for file name
Actions
Move to trash
Upload
New
Home / loudacre / accounts_by_areacode
History Trash
Show 45 of 31 items
Page 1 of 1
Tue Mar 26, 9:22 PM

Name	Size	User	Group	Permissions	Date
↥		training	supergroup	drwxrwxrwx	March 26, 2024 08:42 PM
.		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
.hive-staging_hive_2024-03-26_21-06-02_358_5423725186381291849-1		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=209		training	supergroup	drwxrwxrwx	March 26, 2024 09:06 PM
areacode=213		training	supergroup	drwxrwxrwx	March 26, 2024 09:06 PM
areacode=310		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=408		training	supergroup	drwxrwxrwx	March 26, 2024 09:06 PM
areacode=415		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=424		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=503		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=510		training	supergroup	drwxrwxrwx	March 26, 2024 09:06 PM
areacode=530		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=541		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=559		training	supergroup	drwxrwxrwx	March 26, 2024 09:06 PM
areacode=562		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=619		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=626		training	supergroup	drwxrwxrwx	March 26, 2024 09:07 PM
areacode=650		training	supergroup	drwxrwxrwx	March 26, 2024 09:06 PM



Home / loudacre / accounts_by_areacode / areacode=209 / 000000_0

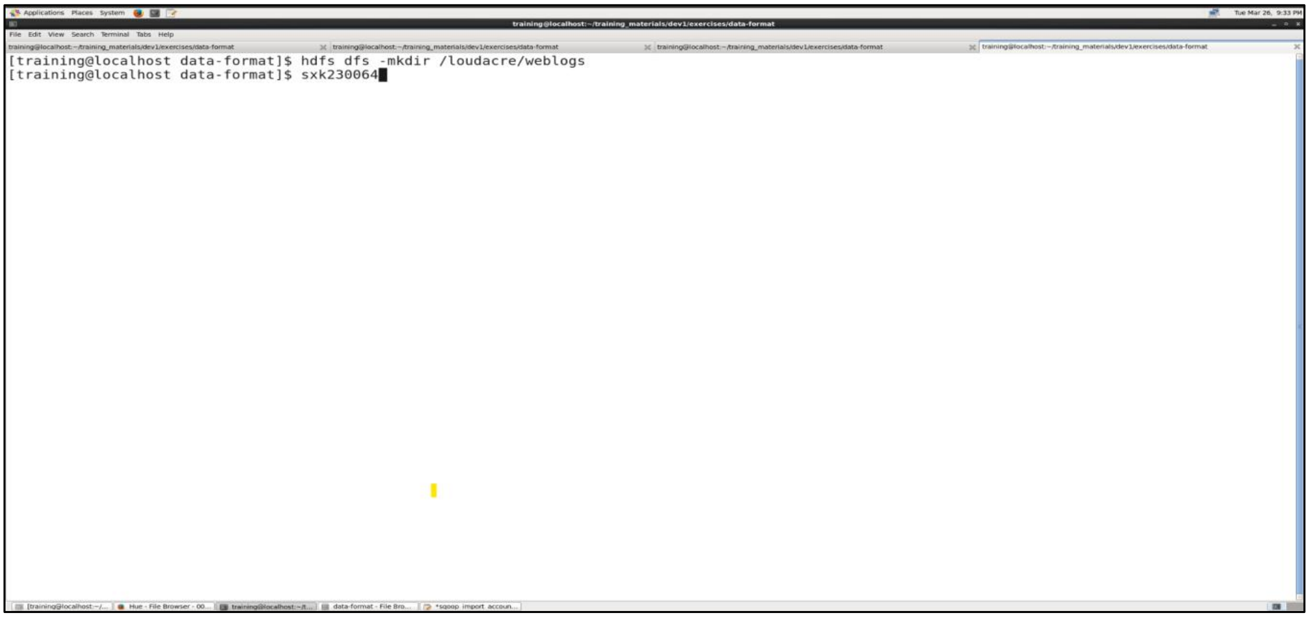
Page 1 of 26

Last modified
March 26, 2024 9:06 p.m.
User
training
Group
supergroup
Size
100.3 KB
Mode
100777

```
492,David,Hasbrouck,2093932930
494,Arthur,Cole,2097791928
500,Bryan,Stewart,2095272417
507,Francisco,Daniels,2097697345
518,Philip,Cave,2097129353
524,Jessica,Alfred,2099375987
541,David,Boyd,2092071335
548,Martha,Lindemann,2090255004
557,Elizabeth,Bailey,2092211472
565,Annette,Titus,2092302508
575,Matthew,Dickson,2092934820
598,Curtis,Hunt,2094429396
605,Richie,Thomas,2091037314
608,Margaret,Stephenson,2099240716
609,Marie,Wood,2092606517
615,Millard,Tran,2094999879
627,Brenda,Perry,2093647288
633,David,Broussard,2095270457
665,Martha,Carter,2098462483
668,Adaline,Harvey,2091499879
685,Derrick,Olive,2099579421
694,Vickie,Hernandez,2098921433
697,Robert,West,2097569184
706,Kenneth,Nelson,2094380351
726,Gail,Rooney,2097640233
728,Millie,Sewell,2095744204
733,Mark,Sanford,2093594118
736,David,Mullins,2090530731
737,Henry,Mathis,2098287779
743,Lorraine,Barrett,2095124671
746,Clara,Johnson,2099426597
761,Robert,Julius,2096557121
```

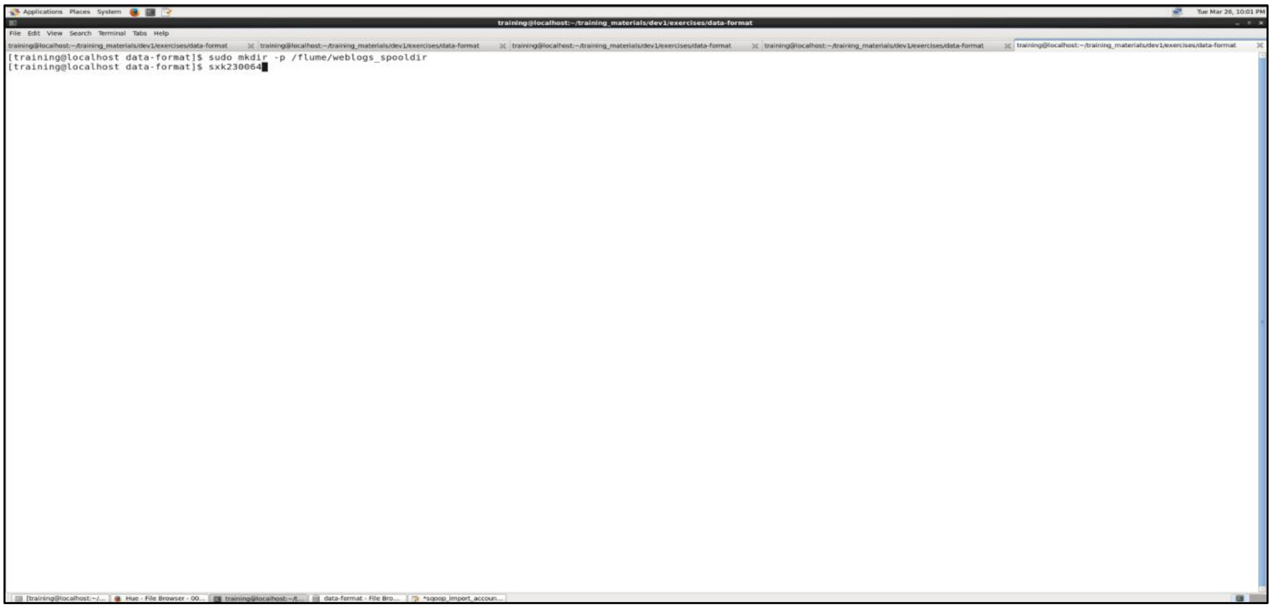
CHAPTER 9 - COLLECT WEB SERVER LOGS WITH FLUME

6. I created a directory named /loudacre/weblogs in HDFS, intending to store the data files ingested by Flume.



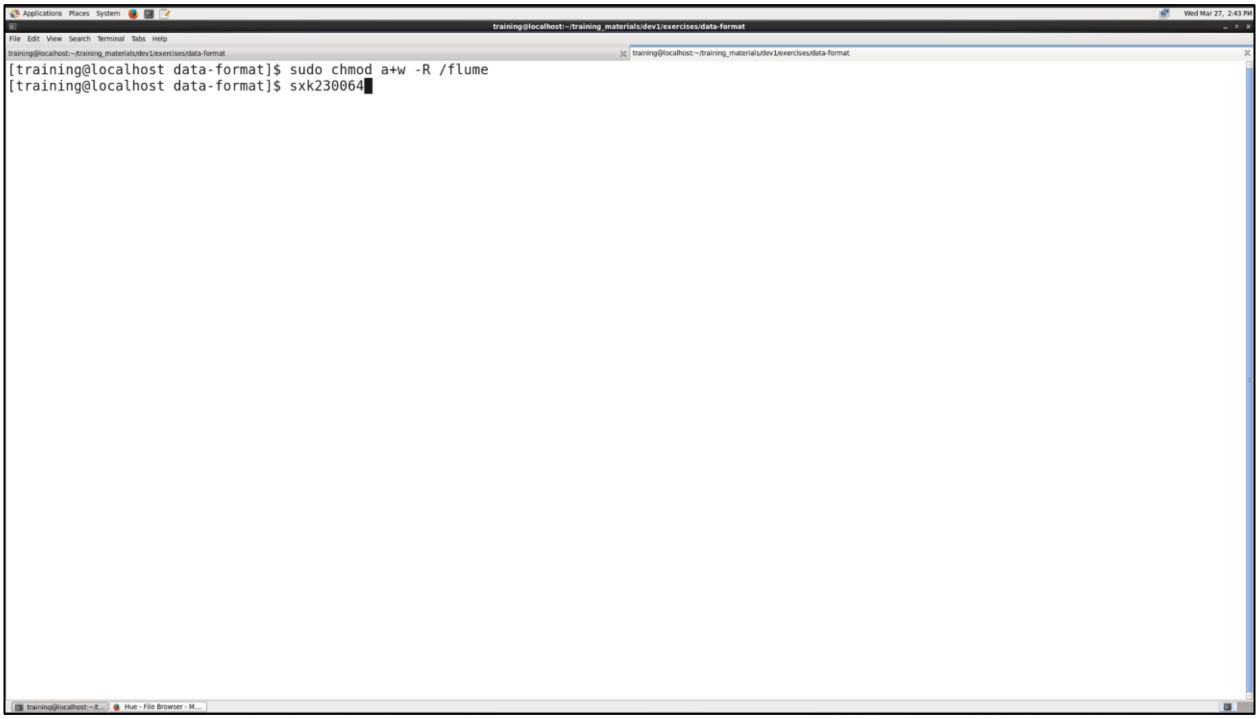
```
[training@localhost data-format]$ hdfs dfs -mkdir /loudacre/weblogs
[training@localhost data-format]$ sxk230064
```

7. Furthermore, I created a spool directory to accommodate the data files that the weblog simulator will generate for Flume ingestion.



```
[training@localhost data-format]$ sudo mkdir -p /flume/weblogs_spooldir
[training@localhost data-format]$ sxk230064
```

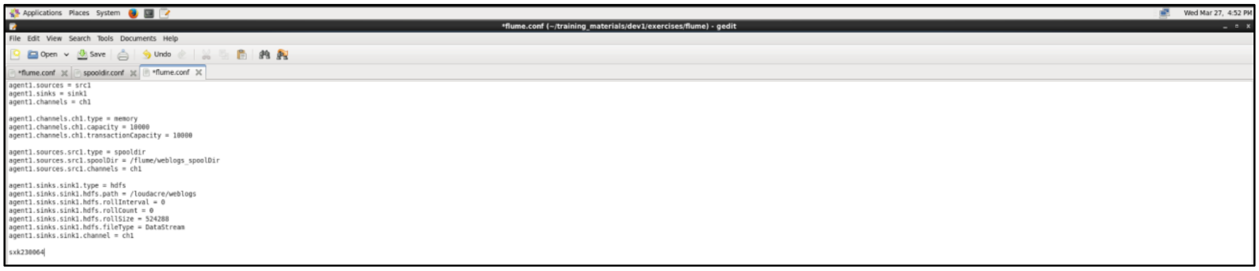
8. To facilitate seamless data ingestion, I granted all users permissions to write to the /flume/weblogs_spooldir directory, executing the command $ sudo chmod a+w -R /flume in the command line.



```
[training@localhost data-format]$ sudo chmod a+w -R /flume
[training@localhost data-format]$ sxk230064
```

Configure Flume

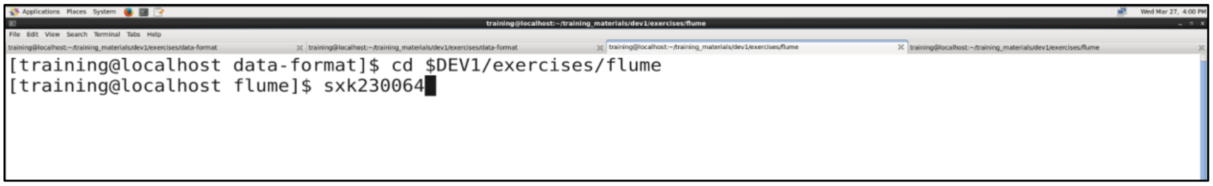
9. I created a flume configuration file with the specified properties for sink, source, and channel.



```
agent1.sources = src1
agent1.sinks = sink1
agent1.channels = ch1

agent1.channels.ch1.type = memory
agent1.channels.ch1.capacity = 10000
agent1.channels.ch1.transactionCapacity = 10000

agent1.sources.src1.type = spooldir
agent1.sources.src1.spoolDir = /flume/weblogs_spoolDir
agent1.sources.src1.channels = ch1

agent1.sinks.sink1.type = hdfs
agent1.sinks.sink1.hdfs.path = /loudacre/weblogs
agent1.sinks.sink1.hdfs.rollInterval = 0
agent1.sinks.sink1.hdfs.rollCount = 0
agent1.sinks.sink1.hdfs.rollSize = 524288
agent1.sinks.sink1.hdfs.fileType = DataStream
agent1.sinks.sink1.channel = ch1

sxk230064
```

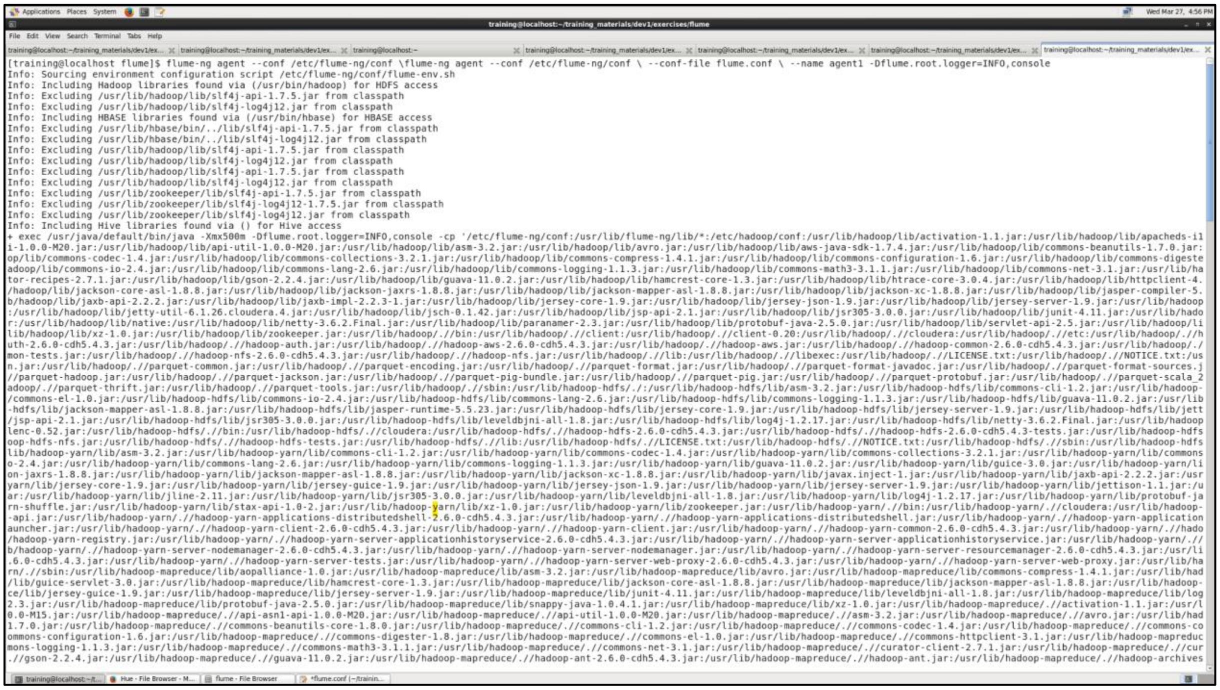
Run the Agent

10. Following configuration setup, I navigated to the /training_materials/dev1/exercises/flume directory.



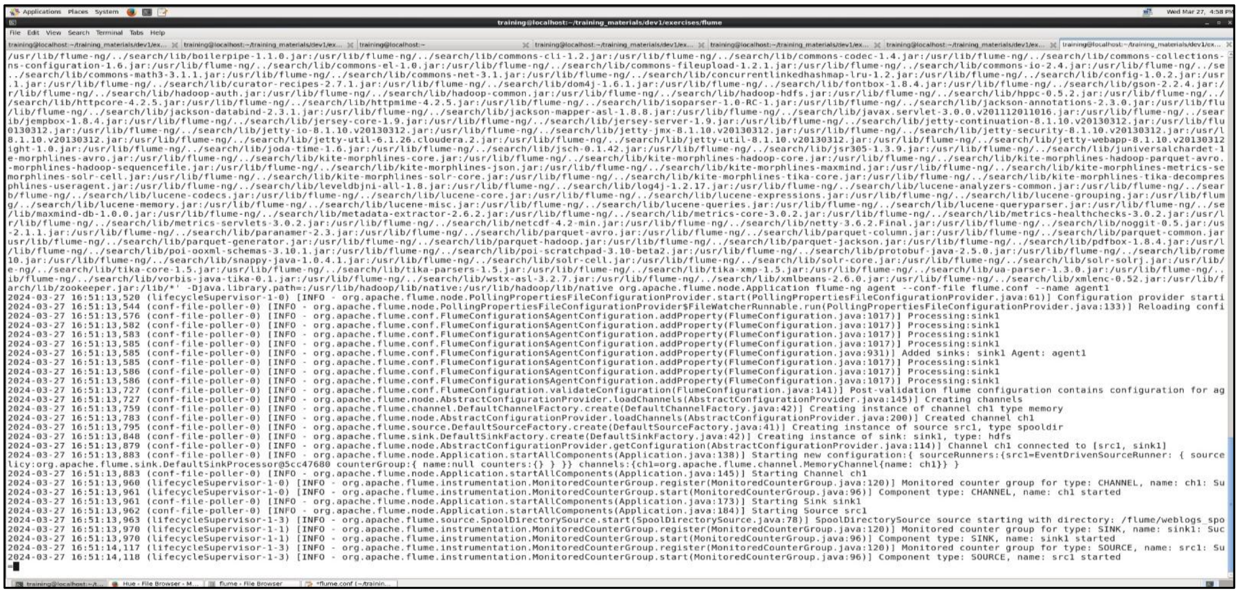
```
[training@localhost data-format]$ cd $DEV1/exercises/flume
[training@localhost flume]$ sxk230064
```

11. I launched the Flume agent utilizing the previously created configuration.



```
[training@localhost flume]$ flume-ng agent --conf /etc/flume-ng/conf \flume-ng agent --conf /etc/flume-ng/conf \ --conf-file flume.conf \ --name agent1 -Dflume.root.logger=INFO,console
Info: Sourcing environment configuration script /etc/flume-ng/conf/flume-env.sh
Info: Including Hadoop libraries found via (/usr/bin/hadoop) for HDFS access
Info: Excluding /usr/lib/hadoop/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-log4j12.jar from classpath
Info: Including HBASE libraries found via (/usr/bin/hbase) for HBASE access
Info: Excluding /usr/lib/hbase/bin/../lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/hbase/bin/../lib/slf4j-log4j12.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-log4j12.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-log4j12.jar from classpath
Info: Excluding /usr/lib/zookeeper/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Excluding /usr/lib/zookeeper/lib/slf4j-log4j12.jar from classpath
Info: Including Hive libraries found via () for Hive access
+ exec /usr/java/default/bin/java -Xmx500m -Dflume.root.logger=INFO,console -cp '/etc/flume-ng/conf:/usr/lib/flume-ng/lib/*:/etc/hadoop/conf:/usr/lib/hadoop/lib/activation-1.1.jar:/usr/lib/hadoop/lib/apacheds-i1
i-1.0.0-M20.jar:/usr/lib/hadoop/lib/api-util-1.0.0-M20.jar:/usr/lib/hadoop/lib/asm-3.2.jar:/usr/lib/hadoop/lib/avro.jar:/usr/lib/hadoop/lib/aws-java-sdk-1.7.4.jar:/usr/lib/hadoop/lib/commons-beanutils-1.7.0.jar:
op/lib/commons-codec-1.4.jar:/usr/lib/hadoop/lib/commons-collections-3.2.1.jar:/usr/lib/hadoop/lib/commons-compress-1.4.1.jar:/usr/lib/hadoop/lib/commons-configuration-1.6.jar:/usr/lib/hadoop/lib/commons-digeste
adoop/lib/commons-io-2.4.jar:/usr/lib/hadoop/lib/commons-lang-2.6.jar:/usr/lib/hadoop/lib/commons-logging-1.1.3.jar:/usr/lib/hadoop/lib/commons-math3-3.1.1.jar:/usr/lib/hadoop/lib/commons-net-3.1.jar:/usr/lib/ha
tor-recipes-2.7.1.jar:/usr/lib/hadoop/lib/gson-2.2.4.jar:/usr/lib/hadoop/lib/guava-11.0.2.jar:/usr/lib/hadoop/lib/hamcrest-core-1.3.jar:/usr/lib/hadoop/lib/htrace-core-3.0.4.jar:/usr/lib/hadoop/lib/httpclient-4.
/hadoop/lib/jackson-core-asl-1.8.8.jar:/usr/lib/hadoop/lib/jackson-jaxrs-1.8.8.jar:/usr/lib/hadoop/lib/jackson-mapper-asl-1.8.8.jar:/usr/lib/hadoop/lib/jackson-xc-1.8.8.jar:/usr/lib/hadoop/lib/jasper-compiler-5.
b/hadoop/lib/jaxb-api-2.2.2.jar:/usr/lib/hadoop/lib/jaxb-impl-2.2.3-1.jar:/usr/lib/hadoop/lib/jersey-core-1.9.jar:/usr/lib/hadoop/lib/jersey-json-1.9.jar:/usr/lib/hadoop/lib/jersey-server-1.9.jar:/usr/lib/hadoop
:/usr/lib/hadoop/lib/jetty-util-6.1.26.cloudera.4.jar:/usr/lib/hadoop/lib/jsch-0.1.42.jar:/usr/lib/hadoop/lib/jsp-api-2.1.jar:/usr/lib/hadoop/lib/jsr305-3.0.0.jar:/usr/lib/hadoop/lib/junit-4.11.jar:/usr/lib/hado
r:/usr/lib/hadoop/lib/native:/usr/lib/hadoop/lib/netty-3.6.2.Final.jar:/usr/lib/hadoop/lib/paranamer-2.3.jar:/usr/lib/hadoop/lib/protobuf-java-2.5.0.jar:/usr/lib/hadoop/lib/servlet-api-2.5.jar:/usr/lib/hadoop/li
lib/hadoop/lib/xz-1.0.jar:/usr/lib/hadoop/lib/zookeeper.jar:/usr/lib/hadoop/.//bin:/usr/lib/hadoop/.//client:/usr/lib/hadoop/.//client-0.20:/usr/lib/hadoop/.//cloudera:/usr/lib/hadoop/.//etc:/usr/lib/hadoop/.//h
uth-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop/.//hadoop-auth.jar:/usr/lib/hadoop/.//hadoop-aws-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop/.//hadoop-aws.jar:/usr/lib/hadoop/.//hadoop-common-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop/./
mon-tests.jar:/usr/lib/hadoop/.//hadoop-nfs-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop/.//hadoop-nfs.jar:/usr/lib/hadoop/.//lib:/usr/lib/hadoop/.//libexec:/usr/lib/hadoop/.//LICENSE.txt:/usr/lib/hadoop/.//NOTICE.txt:/us
n.jar:/usr/lib/hadoop/.//parquet-common.jar:/usr/lib/hadoop/.//parquet-encoding.jar:/usr/lib/hadoop/.//parquet-format.jar:/usr/lib/hadoop/.//parquet-format-javadoc.jar:/usr/lib/hadoop/.//parquet-format-sources.j
//parquet-hadoop.jar:/usr/lib/hadoop/.//parquet-jackson.jar:/usr/lib/hadoop/.//parquet-pig-bundle.jar:/usr/lib/hadoop/.//parquet-pig.jar:/usr/lib/hadoop/.//parquet-protobuf.jar:/usr/lib/hadoop/.//parquet-scala_2
adoop/.//parquet-thrift.jar:/usr/lib/hadoop/.//parquet-tools.jar:/usr/lib/hadoop/.//sbin:/usr/lib/hadoop-hdfs/./:/usr/lib/hadoop-hdfs/lib/asm-3.2.jar:/usr/lib/hadoop-hdfs/lib/commons-cli-1.2.jar:/usr/lib/hadoop-
/commons-el-1.0.jar:/usr/lib/hadoop-hdfs/lib/commons-io-2.4.jar:/usr/lib/hadoop-hdfs/lib/commons-lang-2.6.jar:/usr/lib/hadoop-hdfs/lib/commons-logging-1.1.3.jar:/usr/lib/hadoop-hdfs/lib/guava-11.0.2.jar:/usr/lib
-hdfs/lib/jackson-mapper-asl-1.8.8.jar:/usr/lib/hadoop-hdfs/lib/jasper-runtime-5.5.23.jar:/usr/lib/hadoop-hdfs/lib/jersey-core-1.9.jar:/usr/lib/hadoop-hdfs/lib/jersey-server-1.9.jar:/usr/lib/hadoop-hdfs/lib/jett
/jsp-api-2.1.jar:/usr/lib/hadoop-hdfs/lib/jsr305-3.0.0.jar:/usr/lib/hadoop-hdfs/lib/leveldbjni-all-1.8.jar:/usr/lib/hadoop-hdfs/lib/log4j-1.2.17.jar:/usr/lib/hadoop-hdfs/lib/netty-3.6.2.Final.jar:/usr/lib/hadoop
lenc-0.52.jar:/usr/lib/hadoop-hdfs/.//bin:/usr/lib/hadoop-hdfs/.//cloudera:/usr/lib/hadoop-hdfs/.//hadoop-hdfs-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop-hdfs/.//hadoop-hdfs-2.6.0-cdh5.4.3-tests.jar:/usr/lib/hadoop-hdfs
oop-hdfs-nfs.jar:/usr/lib/hadoop-hdfs/.//hadoop-hdfs-tests.jar:/usr/lib/hadoop-hdfs/.//lib:/usr/lib/hadoop-hdfs/.//LICENSE.txt:/usr/lib/hadoop-hdfs/.//NOTICE.txt:/usr/lib/hadoop-hdfs/.//sbin:/usr/lib/hadoop-hdfs
lib/hadoop-yarn/lib/asm-3.2.jar:/usr/lib/hadoop-yarn/lib/commons-cli-1.2.jar:/usr/lib/hadoop-yarn/lib/commons-codec-1.4.jar:/usr/lib/hadoop-yarn/lib/commons-collections-3.2.1.jar:/usr/lib/hadoop-yarn/lib/commons
o-2.4.jar:/usr/lib/hadoop-yarn/lib/commons-lang-2.6.jar:/usr/lib/hadoop-yarn/lib/commons-logging-1.1.3.jar:/usr/lib/hadoop-yarn/lib/guava-11.0.2.jar:/usr/lib/hadoop-yarn/lib/guice-3.0.jar:/usr/lib/hadoop-yarn/li
on-jaxrs-1.8.8.jar:/usr/lib/hadoop-yarn/lib/jackson-mapper-asl-1.8.8.jar:/usr/lib/hadoop-yarn/lib/jackson-xc-1.8.8.jar:/usr/lib/hadoop-yarn/lib/javax.inject-1.jar:/usr/lib/hadoop-yarn/lib/jaxb-api-2.2.2.jar:/usr
yarn/lib/jersey-core-1.9.jar:/usr/lib/hadoop-yarn/lib/jersey-guice-1.9.jar:/usr/lib/hadoop-yarn/lib/jersey-json-1.9.jar:/usr/lib/hadoop-yarn/lib/jersey-server-1.9.jar:/usr/lib/hadoop-yarn/lib/jettison-1.1.jar:/u
ar:/usr/lib/hadoop-yarn/lib/jline-2.11.jar:/usr/lib/hadoop-yarn/lib/jsr305-3.0.0.jar:/usr/lib/hadoop-yarn/lib/leveldbjni-all-1.8.jar:/usr/lib/hadoop-yarn/lib/log4j-1.2.17.jar:/usr/lib/hadoop-yarn/lib/protobuf-ja
rn-shuffle.jar:/usr/lib/hadoop-yarn/lib/stax-api-1.0-2.jar:/usr/lib/hadoop-yarn/lib/xz-1.0.jar:/usr/lib/hadoop-yarn/lib/zookeeper.jar:/usr/lib/hadoop-yarn/.//bin:/usr/lib/hadoop-yarn/.//cloudera:/usr/lib/hadoop-
-api.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-applications-distributedshell-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-applications-distributedshell.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-application
auncher.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-client-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-client.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-common-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop-yarn/.//hado
/hadoop-yarn-registry.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-server-applicationhistoryservice-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-server-applicationhistoryservice.jar:/usr/lib/hadoop-yarn/.//
b/hadoop-yarn/.//hadoop-yarn-server-nodemanager-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-server-nodemanager.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-server-resourcemanager-2.6.0-cdh5.4.3.jar:/usr/li
.6.0-cdh5.4.3.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-server-tests.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-server-web-proxy-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop-yarn/.//hadoop-yarn-server-web-proxy.jar:/usr/lib/ha
rn/.//sbin:/usr/lib/hadoop-mapreduce/lib/aopalliance-1.0.jar:/usr/lib/hadoop-mapreduce/lib/asm-3.2.jar:/usr/lib/hadoop-mapreduce/lib/avro.jar:/usr/lib/hadoop-mapreduce/lib/commons-compress-1.4.1.jar:/usr/lib/had
/lib/guice-servlet-3.0.jar:/usr/lib/hadoop-mapreduce/lib/hamcrest-core-1.3.jar:/usr/lib/hadoop-mapreduce/lib/jackson-core-asl-1.8.8.jar:/usr/lib/hadoop-mapreduce/lib/jackson-mapper-asl-1.8.8.jar:/usr/lib/hadoop-
ce/lib/jersey-guice-1.9.jar:/usr/lib/hadoop-mapreduce/lib/jersey-server-1.9.jar:/usr/lib/hadoop-mapreduce/lib/junit-4.11.jar:/usr/lib/hadoop-mapreduce/lib/leveldbjni-all-1.8.jar:/usr/lib/hadoop-mapreduce/lib/log
2.3.jar:/usr/lib/hadoop-mapreduce/lib/protobuf-java-2.5.0.jar:/usr/lib/hadoop-mapreduce/lib/snappy-java-1.0.4.1.jar:/usr/lib/hadoop-mapreduce/lib/xz-1.0.jar:/usr/lib/hadoop-mapreduce/.//activation-1.1.jar:/usr/l
0.0-M15.jar:/usr/lib/hadoop-mapreduce/.//api-asn1-api-1.0.0-M20.jar:/usr/lib/hadoop-mapreduce/.//api-util-1.0.0-M20.jar:/usr/lib/hadoop-mapreduce/.//asm-3.2.jar:/usr/lib/hadoop-mapreduce/.//avro.jar:/usr/lib/had
1.7.0.jar:/usr/lib/hadoop-mapreduce/.//commons-beanutils-core-1.8.0.jar:/usr/lib/hadoop-mapreduce/.//commons-cli-1.2.jar:/usr/lib/hadoop-mapreduce/.//commons-codec-1.4.jar:/usr/lib/hadoop-mapreduce/.//commons-co
ommons-configuration-1.6.jar:/usr/lib/hadoop-mapreduce/.//commons-digester-1.8.jar:/usr/lib/hadoop-mapreduce/.//commons-el-1.0.jar:/usr/lib/hadoop-mapreduce/.//commons-httpclient-3.1.jar:/usr/lib/hadoop-mapreduc
mons-logging-1.1.3.jar:/usr/lib/hadoop-mapreduce/.//commons-math3-3.1.1.jar:/usr/lib/hadoop-mapreduce/.//commons-net-3.1.jar:/usr/lib/hadoop-mapreduce/.//curator-client-2.7.1.jar:/usr/lib/hadoop-mapreduce/.//cur
.//gson-2.2.4.jar:/usr/lib/hadoop-mapreduce/.//guava-11.0.2.jar:/usr/lib/hadoop-mapreduce/.//hadoop-ant-2.6.0-cdh5.4.3.jar:/usr/lib/hadoop-mapreduce/.//hadoop-ant.jar:/usr/lib/hadoop-mapreduce/.//hadoop-archives
```

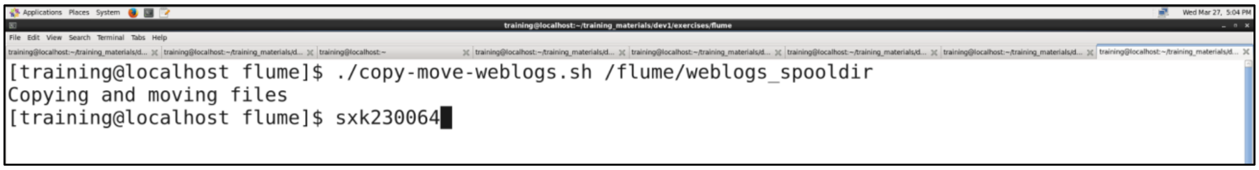
12. Towards the end of the lines, it's confirmed that the agent is successfully running with the configuration named src1.


Applications Places System
Wed Mar 27, 4:58 PM
training@localhost:~/training_materials/dev1/exercises/flume
File Edit View Search Terminal Tabs Help
training@localhost:~/training_materials/dev1/ex...
training@localhost:~
Hue - File Browser - M...
flume - File Browser
*flume.conf (~/trainin...

```
/usr/lib/flume-ng/../search/lib/boilerpipe-1.1.0.jar:/usr/lib/flume-ng/../search/lib/commons-cli-1.2.jar:/usr/lib/flume-ng/../search/lib/commons-codec-1.4.jar:/usr/lib/flume-ng/../search/lib/commons-collections-
ns-configuration-1.6.jar:/usr/lib/flume-ng/../search/lib/commons-el-1.0.jar:/usr/lib/flume-ng/../search/lib/commons-fileupload-1.2.1.jar:/usr/lib/flume-ng/../search/lib/commons-io-2.4.jar:/usr/lib/flume-ng/../se
../search/lib/commons-math3-3.1.1.jar:/usr/lib/flume-ng/../search/lib/commons-net-3.1.jar:/usr/lib/flume-ng/../search/lib/concurrentlinkedhashmap-lru-1.2.jar:/usr/lib/flume-ng/../search/lib/config-1.0.2.jar:/usr
.1.jar:/usr/lib/flume-ng/../search/lib/curator-recipes-2.7.1.jar:/usr/lib/flume-ng/../search/lib/dom4j-1.6.1.jar:/usr/lib/flume-ng/../search/lib/fontbox-1.8.4.jar:/usr/lib/flume-ng/../search/lib/gson-2.2.4.jar:/
r/lib/flume-ng/../search/lib/hadoop-auth.jar:/usr/lib/flume-ng/../search/lib/hadoop-common.jar:/usr/lib/flume-ng/../search/lib/hadoop-hdfs.jar:/usr/lib/flume-ng/../search/lib/hppc-0.5.2.jar:/usr/lib/flume-ng/../
/search/lib/httpcore-4.2.5.jar:/usr/lib/flume-ng/../search/lib/httpmime-4.2.5.jar:/usr/lib/flume-ng/../search/lib/isoparser-1.0-RC-1.jar:/usr/lib/flume-ng/../search/lib/jackson-annotations-2.3.0.jar:/usr/lib/flu
/lib/flume-ng/../search/lib/jackson-databind-2.3.1.jar:/usr/lib/flume-ng/../search/lib/jackson-mapper-asl-1.8.8.jar:/usr/lib/flume-ng/../search/lib/javax.servlet-3.0.0.v201112011016.jar:/usr/lib/flume-ng/../sear
ib/jempbox-1.8.4.jar:/usr/lib/flume-ng/../search/lib/jersey-core-1.9.jar:/usr/lib/flume-ng/../search/lib/jersey-server-1.9.jar:/usr/lib/flume-ng/../search/lib/jetty-continuation-8.1.10.v20130312.jar:/usr/lib/flu
0130312.jar:/usr/lib/flume-ng/../search/lib/jetty-io-8.1.10.v20130312.jar:/usr/lib/flume-ng/../search/lib/jetty-jmx-8.1.10.v20130312.jar:/usr/lib/flume-ng/../search/lib/jetty-security-8.1.10.v20130312.jar:/usr/l
8.1.10.v20130312.jar:/usr/lib/flume-ng/../search/lib/jetty-util-6.1.26.cloudera.2.jar:/usr/lib/flume-ng/../search/lib/jetty-util-8.1.10.v20130312.jar:/usr/lib/flume-ng/../search/lib/jetty-webapp-8.1.10.v20130312
ight-1.0.jar:/usr/lib/flume-ng/../search/lib/joda-time-1.6.jar:/usr/lib/flume-ng/../search/lib/jsch-0.1.42.jar:/usr/lib/flume-ng/../search/lib/jsr305-1.3.9.jar:/usr/lib/flume-ng/../search/lib/juniversalchardet-1
e-morphlines-avro.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-core.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-hadoop-core.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-hadoop-parquet-avro.
-morphlines-hadoop-sequencefile.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-json.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-maxmind.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-metrics-se
morphlines-solr-cell.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-solr-core.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-tika-core.jar:/usr/lib/flume-ng/../search/lib/kite-morphlines-tika-decompres
phlines-useragent.jar:/usr/lib/flume-ng/../search/lib/leveldbjni-all-1.8.jar:/usr/lib/flume-ng/../search/lib/log4j-1.2.17.jar:/usr/lib/flume-ng/../search/lib/lucene-analyzers-common.jar:/usr/lib/flume-ng/../sear
b/flume-ng/../search/lib/lucene-codecs.jar:/usr/lib/flume-ng/../search/lib/lucene-core.jar:/usr/lib/flume-ng/../search/lib/lucene-expressions.jar:/usr/lib/flume-ng/../search/lib/lucene-grouping.jar:/usr/lib/flum
g/../search/lib/lucene-memory.jar:/usr/lib/flume-ng/../search/lib/lucene-misc.jar:/usr/lib/flume-ng/../search/lib/lucene-queries.jar:/usr/lib/flume-ng/../search/lib/lucene-queryparser.jar:/usr/lib/flume-ng/../se
/lib/maxmind-db-1.0.0.jar:/usr/lib/flume-ng/../search/lib/metadata-extractor-2.6.2.jar:/usr/lib/flume-ng/../search/lib/metrics-core-3.0.2.jar:/usr/lib/flume-ng/../search/lib/metrics-healthchecks-3.0.2.jar:/usr/l
r/lib/flume-ng/../search/lib/metrics-servlets-3.0.2.jar:/usr/lib/flume-ng/../search/lib/netcdf-4.2-min.jar:/usr/lib/flume-ng/../search/lib/netty-3.6.2.Final.jar:/usr/lib/flume-ng/../search/lib/noggit-0.5.jar:/us
-2.1.1.jar:/usr/lib/flume-ng/../search/lib/paranamer-2.3.jar:/usr/lib/flume-ng/../search/lib/parquet-avro.jar:/usr/lib/flume-ng/../search/lib/parquet-column.jar:/usr/lib/flume-ng/../search/lib/parquet-common.jar
usr/lib/flume-ng/../search/lib/parquet-generator.jar:/usr/lib/flume-ng/../search/lib/parquet-hadoop.jar:/usr/lib/flume-ng/../search/lib/parquet-jackson.jar:/usr/lib/flume-ng/../search/lib/pdfbox-1.8.4.jar:/usr/l
/lib/flume-ng/../search/lib/poi-ooxml-schemas-3.10.1.jar:/usr/lib/flume-ng/../search/lib/poi-scratchpad-3.10-beta2.jar:/usr/lib/flume-ng/../search/lib/protobuf-java-2.5.0.jar:/usr/lib/flume-ng/../search/lib/rome
10.jar:/usr/lib/flume-ng/../search/lib/snappy-java-1.0.4.1.jar:/usr/lib/flume-ng/../search/lib/solr-cell.jar:/usr/lib/flume-ng/../search/lib/solr-core.jar:/usr/lib/flume-ng/../search/lib/solr-solrj.jar:/usr/lib/
e-ng/../search/lib/tika-core-1.5.jar:/usr/lib/flume-ng/../search/lib/tika-parsers-1.5.jar:/usr/lib/flume-ng/../search/lib/tika-xmp-1.5.jar:/usr/lib/flume-ng/../search/lib/ua-parser-1.3.0.jar:/usr/lib/flume-ng/..
ib/flume-ng/../search/lib/vorbis-java-tika-0.1.jar:/usr/lib/flume-ng/../search/lib/wstx-asl-3.2.7.jar:/usr/lib/flume-ng/../search/lib/xmlbeans-2.6.0.jar:/usr/lib/flume-ng/../search/lib/xmlenc-0.52.jar:/usr/lib/f
arch/lib/zookeeper.jar:/lib/*' -Djava.library.path=:/usr/lib/hadoop/lib/native:/usr/lib/hadoop/lib/native org.apache.flume.node.Application flume-ng agent --conf-file flume.conf --name agent1
2024-03-27 16:51:13,520 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.node.PollingPropertiesFileConfigurationProvider.start(PollingPropertiesFileConfigurationProvider.java:61)] Configuration provider starti
2024-03-27 16:51:13,544 (conf-file-poller-0) [INFO - org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java:133)] Reloading confi
2024-03-27 16:51:13,576 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:sink1
2024-03-27 16:51:13,582 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:sink1
2024-03-27 16:51:13,583 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:sink1
2024-03-27 16:51:13,585 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:sink1
2024-03-27 16:51:13,585 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:931)] Added sinks: sink1 Agent: agent1
2024-03-27 16:51:13,585 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:sink1
2024-03-27 16:51:13,586 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:sink1
2024-03-27 16:51:13,586 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:sink1
2024-03-27 16:51:13,727 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration.validateConfiguration(FlumeConfiguration.java:141)] Post-validation flume configuration contains configuration for ag
2024-03-27 16:51:13,727 (conf-file-poller-0) [INFO - org.apache.flume.node.AbstractConfigurationProvider.loadChannels(AbstractConfigurationProvider.java:145)] Creating channels
2024-03-27 16:51:13,759 (conf-file-poller-0) [INFO - org.apache.flume.channel.DefaultChannelFactory.create(DefaultChannelFactory.java:42)] Creating instance of channel ch1 type memory
2024-03-27 16:51:13,783 (conf-file-poller-0) [INFO - org.apache.flume.node.AbstractConfigurationProvider.loadChannels(AbstractConfigurationProvider.java:200)] Created channel ch1
2024-03-27 16:51:13,795 (conf-file-poller-0) [INFO - org.apache.flume.source.DefaultSourceFactory.create(DefaultSourceFactory.java:41)] Creating instance of source src1, type spooldir
2024-03-27 16:51:13,848 (conf-file-poller-0) [INFO - org.apache.flume.sink.DefaultSinkFactory.create(DefaultSinkFactory.java:42)] Creating instance of sink: sink1, type: hdfs
2024-03-27 16:51:13,879 (conf-file-poller-0) [INFO - org.apache.flume.node.AbstractConfigurationProvider.getConfiguration(AbstractConfigurationProvider.java:114)] Channel ch1 connected to [src1, sink1]
2024-03-27 16:51:13,883 (conf-file-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:138)] Starting new configuration:{ sourceRunners:{src1=EventDrivenSourceRunner: { source
licy:org.apache.flume.sink.DefaultSinkProcessor@5cc47680 counterGroup:{ name:null counters:{} } }} channels:{ch1=org.apache.flume.channel.MemoryChannel{name: ch1}} }
2024-03-27 16:51:13,883 (conf-file-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:145)] Starting Channel ch1
2024-03-27 16:51:13,960 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.register(MonitoredCounterGroup.java:120)] Monitored counter group for type: CHANNEL, name: ch1: Su
2024-03-27 16:51:13,961 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:96)] Component type: CHANNEL, name: ch1 started
2024-03-27 16:51:13,961 (conf-file-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:173)] Starting Sink sink1
2024-03-27 16:51:13,962 (conf-file-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:184)] Starting Source src1
2024-03-27 16:51:13,963 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.source.SpoolDirectorySource.start(SpoolDirectorySource.java:78)] SpoolDirectorySource source starting with directory: /flume/weblogs_spo
2024-03-27 16:51:13,970 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.register(MonitoredCounterGroup.java:120)] Monitored counter group for type: SINK, name: sink1: Suc
2024-03-27 16:51:13,970 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:96)] Component type: SINK, name: sink1 started
2024-03-27 16:51:14,117 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.register(MonitoredCounterGroup.java:120)] Monitored counter group for type: SOURCE, name: src1: Su
2024-03-27 16:51:14,118 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:96)] Component type: SOURCE, name: src1 started
```

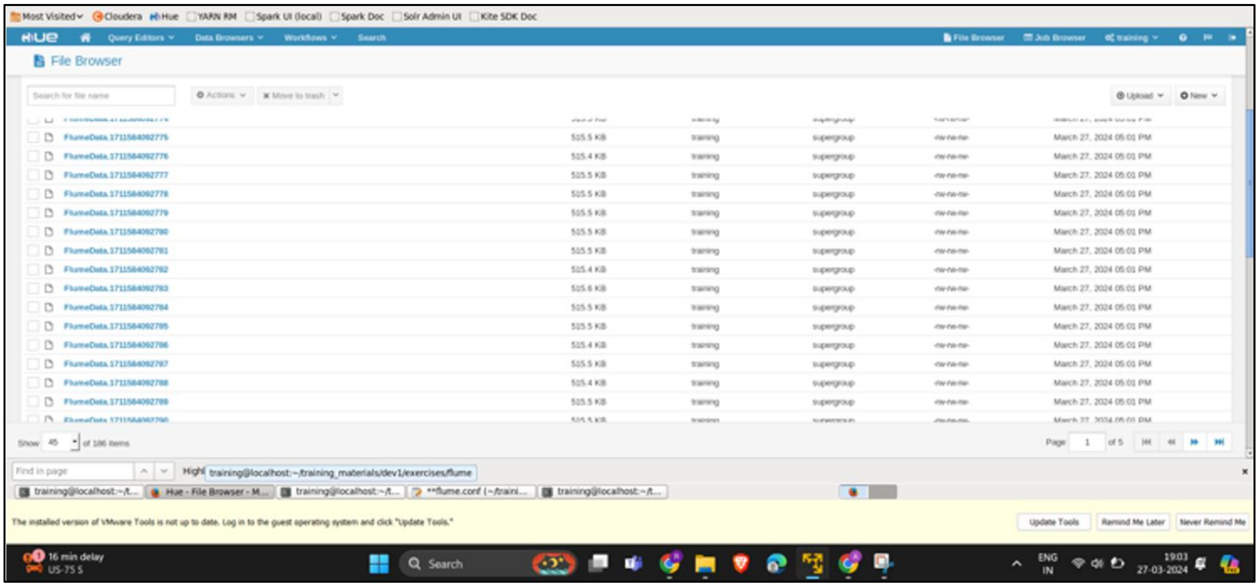
Simulate Apache web server output

13. I opened a new terminal window, changed to the exercise directory, and executed the script to place the web log files in the /flume/weblogs_spooldir directory.


Applications Places System
Wed Mar 27, 5:04 PM
training@localhost:~/training_materials/dev1/exercises/flume
File Edit View Search Terminal Tabs Help

```
[training@localhost flume]$ ./copy-move-weblogs.sh /flume/weblogs_spooldir
Copying and moving files
[training@localhost flume]$ sxk230064
```

14. Upon completion, I terminated the process by clicking CTRL + C. I used the HUE file browser to validate that Flume successfully copied the weblogs into the weblogs directory. Each imported file is tagged with a Unix timestamp corresponding to the time it was imported.



File Browser

FlumeData.1711584092775	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092776	515.4 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092777	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092778	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092779	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092780	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092781	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092782	515.4 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092783	515.6 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092784	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092785	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092786	515.4 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092787	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092788	515.4 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092789	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM
FlumeData.1711584092790	515.5 KB	training	supergroup	-rw-rw-rw-	March 27, 2024 05:01 PM

Show 45 of 186 items

Page 1 of 5

Find in page
Highl
training@localhost:~/training_materials/dev1/exercises/flume
training@localhost:~/t...
Hue - File Browser - M...
training@localhost:~/t...
**flume.conf (~/traini...
training@localhost:~/t...

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