

# File Formatting

---

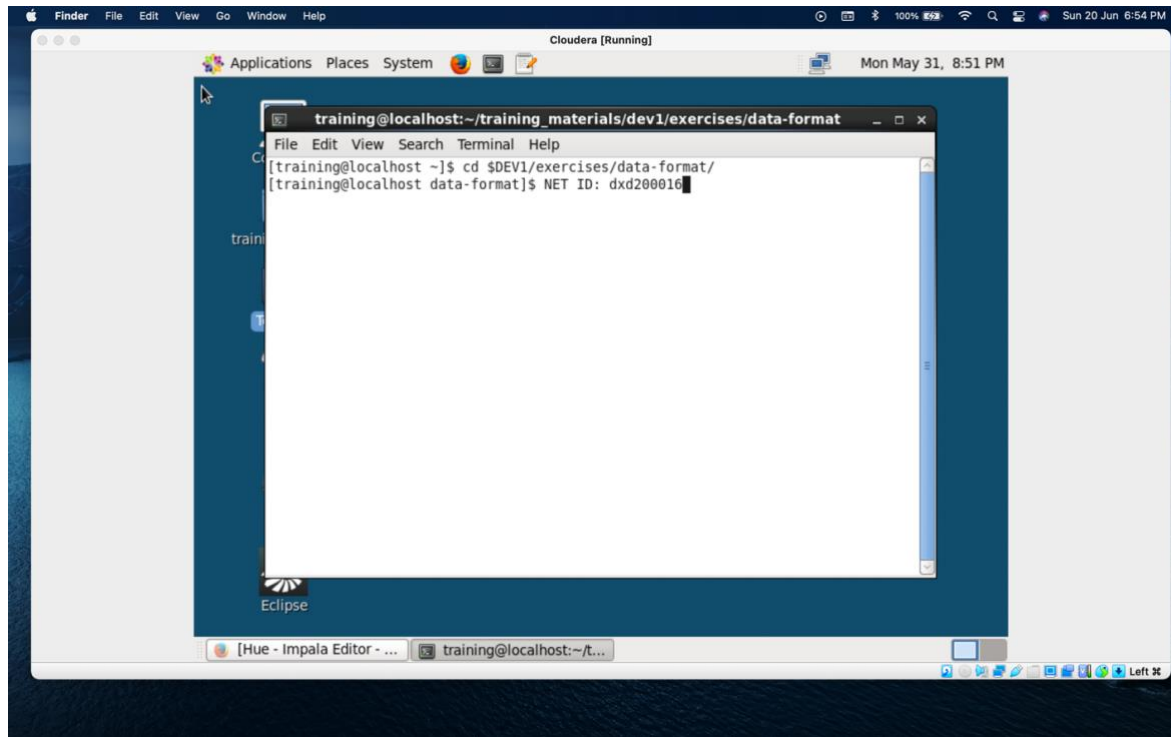
Divyansh Dahiya

## LAB Chapter 07: FILE FORMAT

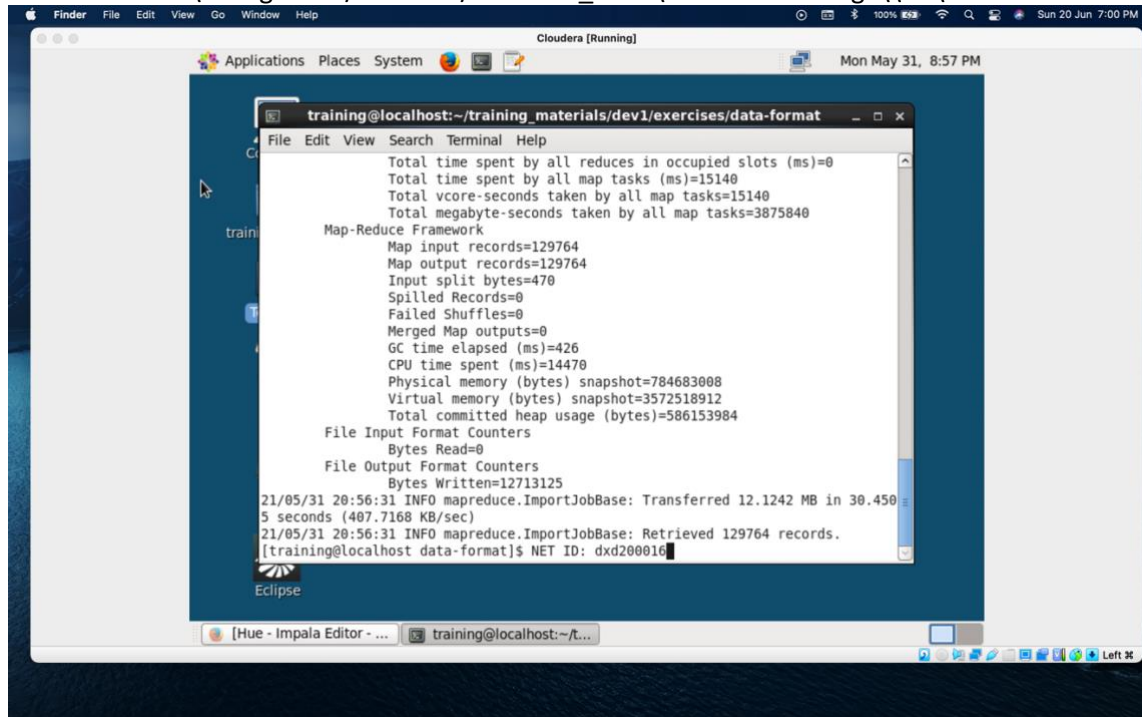
### Select a Format for a Data File

In this lab I will use import data in Avro format and create an Impala table to access it.

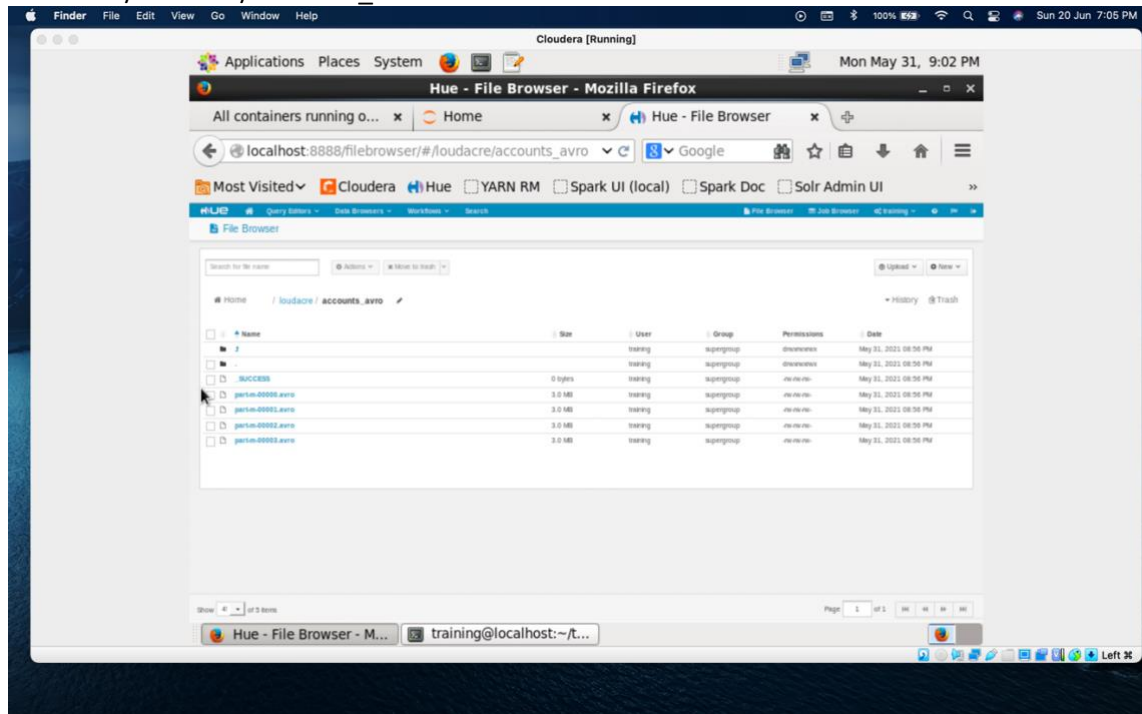
1. First, I opened the terminal and ran the following command to change the directory.  
Location: `cd $DEV1/exercises/data-format/`



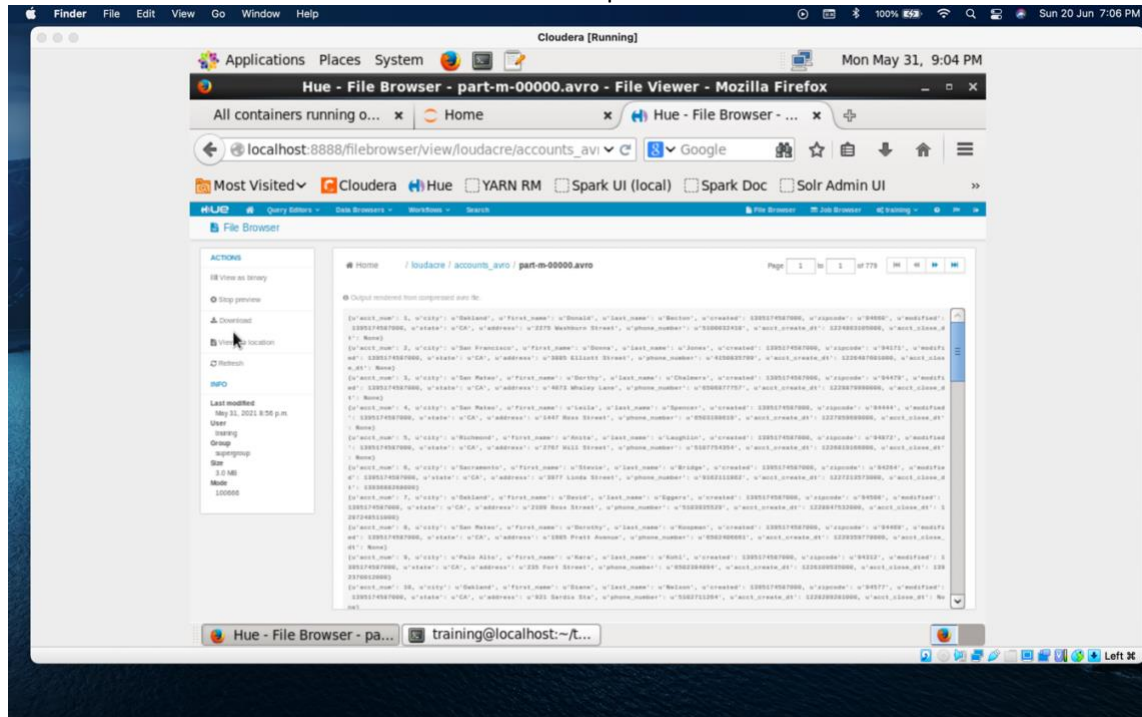
- Then, using Sqoop I ran the following command to import the accounts table in Avro data format.  
Command: `sqoop import \ --connect jdbc:mysql://localhost/loudacre \ --username training --password training \ --table accounts \ --target-dir /loudacre/accounts_avro \ --null-non-string '\\N' \ --as-avrodatafile`



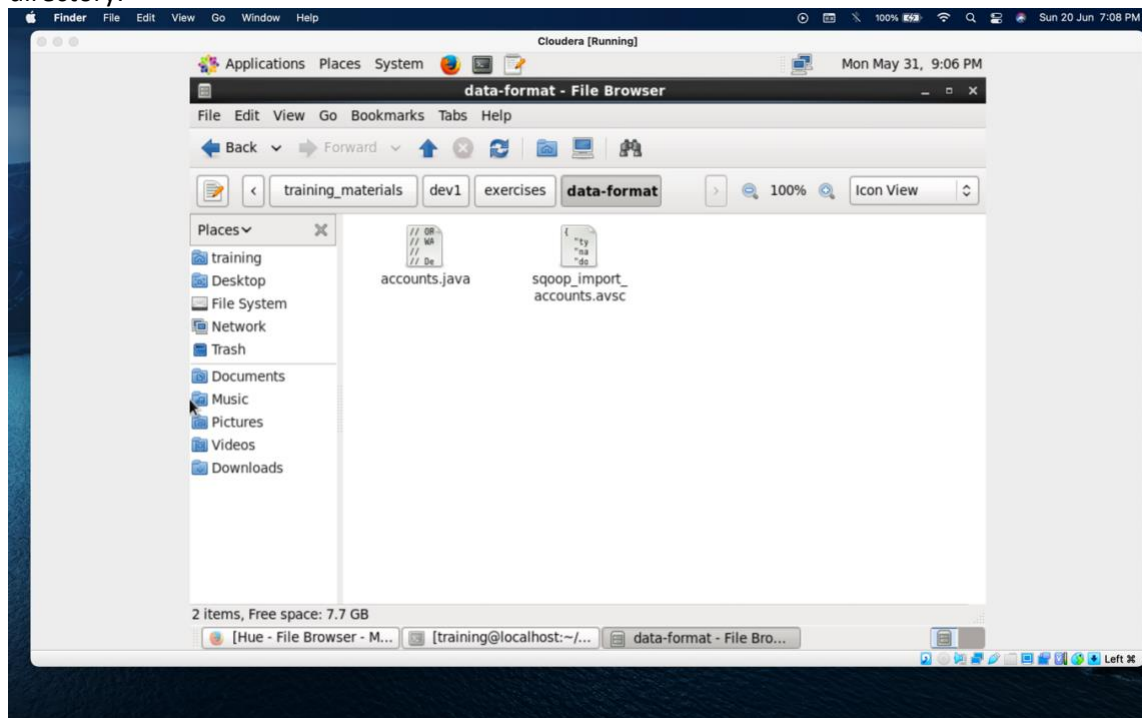
- Then, I opened the Hue Web Interface to view the imported files in the HDFS.  
Location: `/loudacre/accounts_avro`



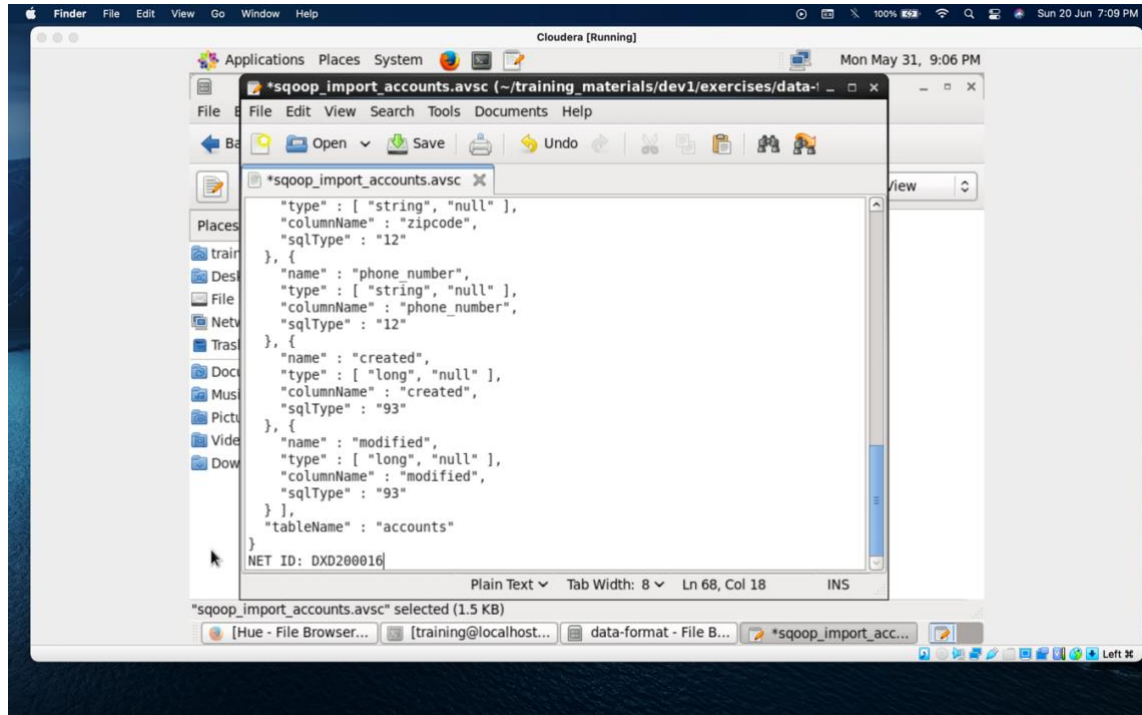
- \_\_\_\_\_



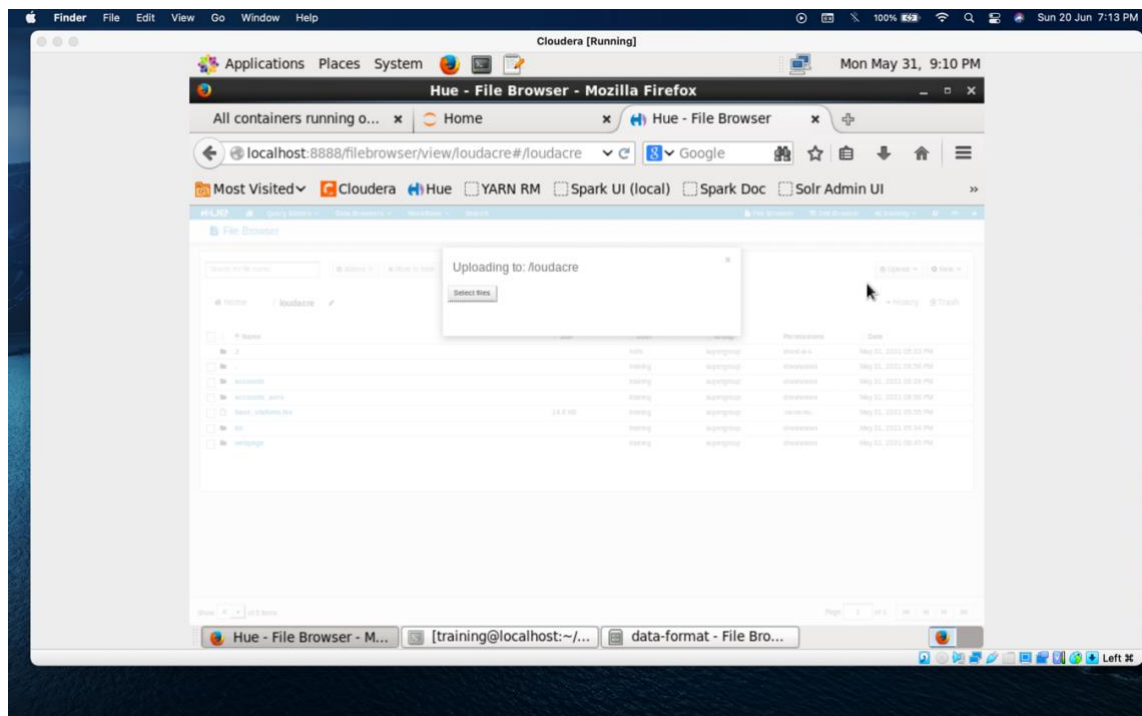
5. Below is the screenshot of the schema generated `sqoop_import_accounts.avsc`, it was stored in the current local directory.



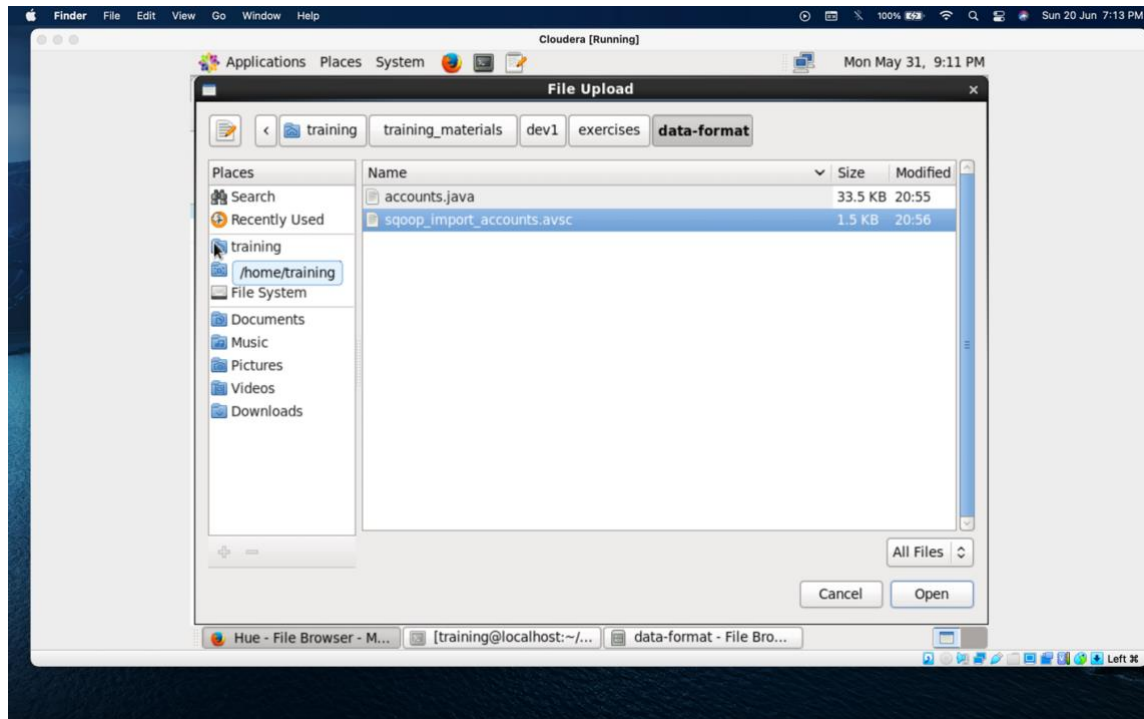
6. Below is the screenshot of the contents inside the sqoop\_import\_accounts.avsc file.



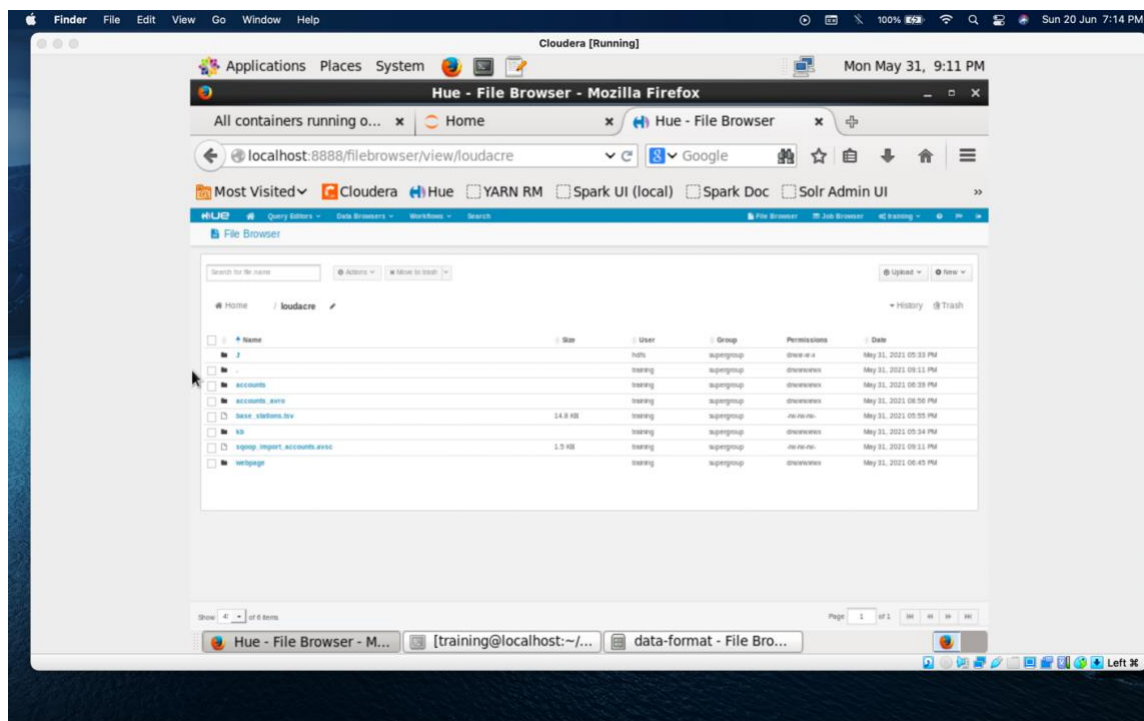
7. Then I opened the Hue Web Interface and opened the /loudacre directory and clicked on the upload button.



8. Then, I selected the schema file and clicked on the open button to upload the file.



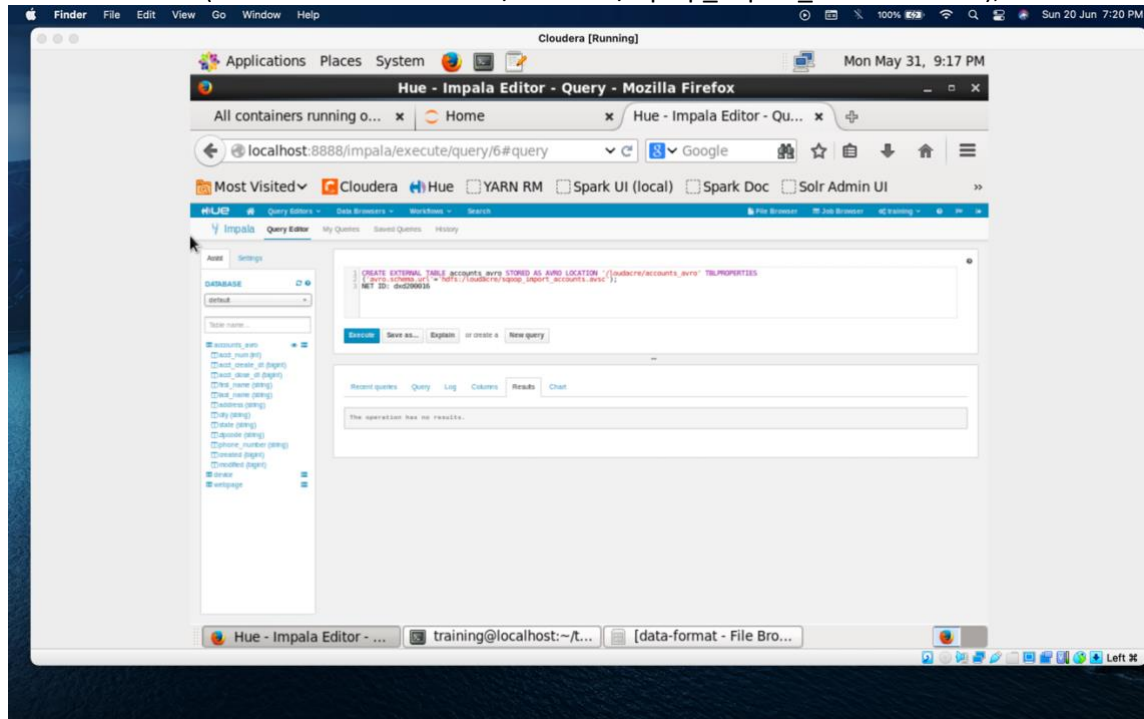
9. Below is the screenshot of the contents inside the /loudacre directory. The sqoop\_import\_accounts.avsc file is successfully imported.





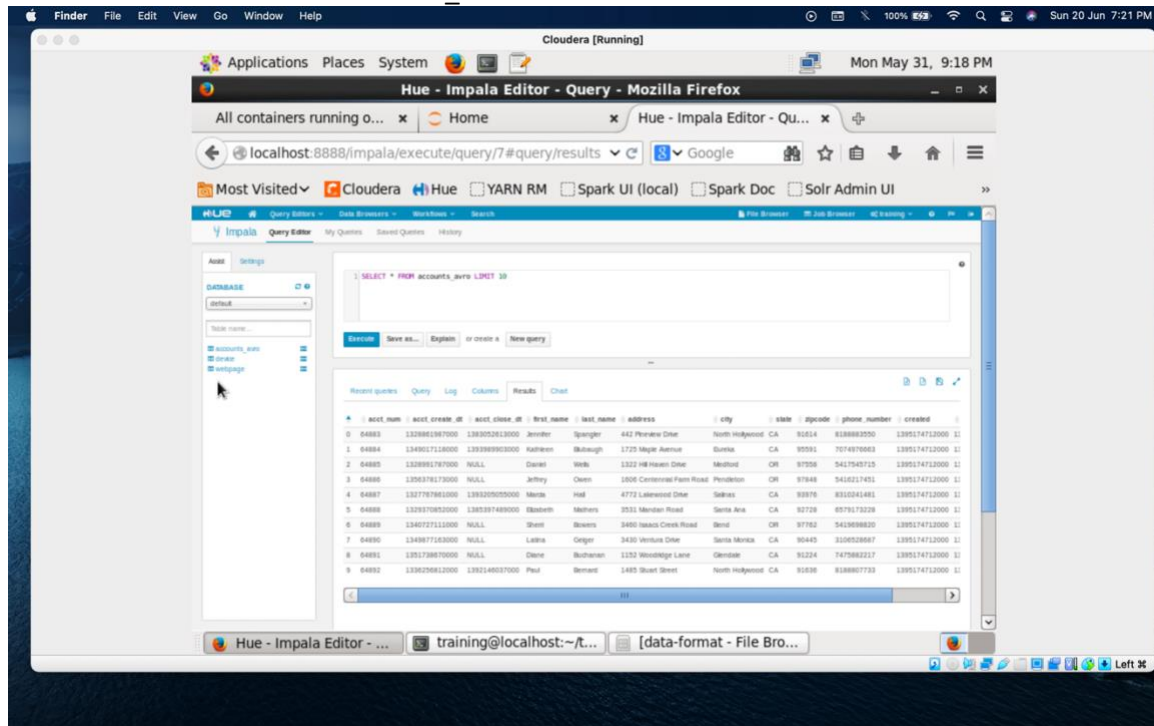
10. Then I opened the Impala Query Editor and ran the following query to create the table using schema imported in the previous step.

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



11. Then, ran the following command to view the first 10 records inside the accounts\_avro table.

Command: `SELECT * FROM accounts_avro LIMIT 10`



12. Finally, I ran the following command to list all the columns and their data types in the accounts\_avro table.  
Command: DESCRIBE FORMATTED accounts\_avro

