Data Partitioning

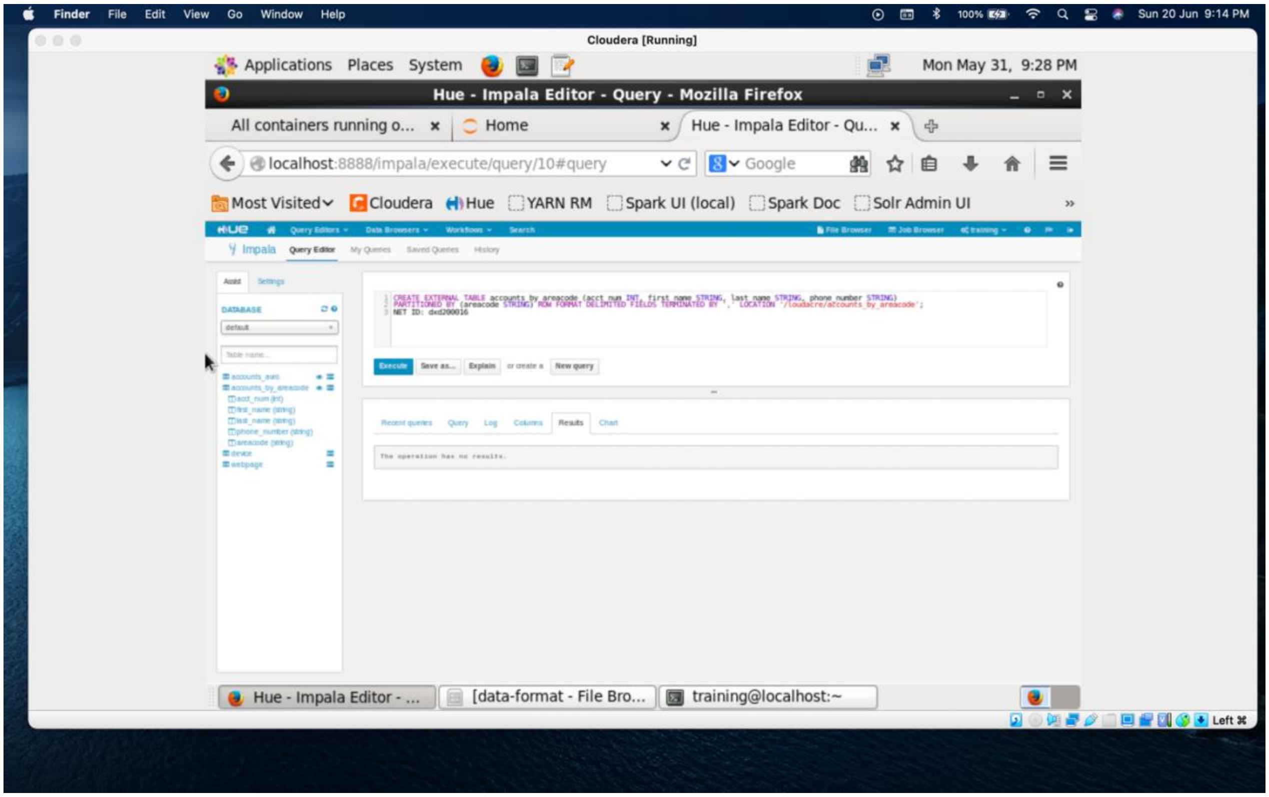
Divyansh Dahiya

LAB Chapter 08: PARTITIONS

Partition Data in Impala or Hive

In this lab, I will create and load an Impala table with account data, partitioned by area code.

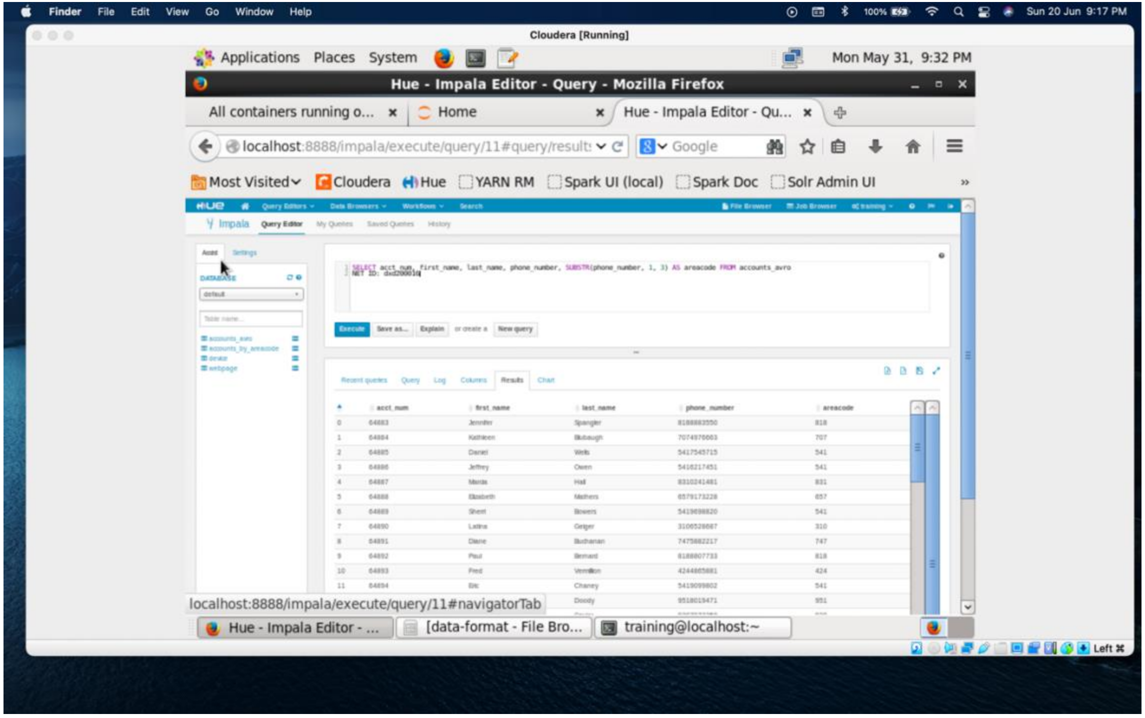
1. First, I opened Impala Query Editor inside Hue Web Interface and ran the following command to create an empty table accounts_by_areacode
Command: CREATE EXTERNAL TABLE accounts_by_areacode (acct_num INT, first_name STRING, last_name STRING, phone_number STRING) PARTITIONED BY (areacode STRING) ROW 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';
NET ID: dxd200016
```

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

2. I ran the following command to select the specific columns and extracted the area code from the phone number.
   Command: SELECT acct_num, first_name, last_name, phone_number, SUBSTR(phone_number,1,3) AS areacode FROM accounts_avro

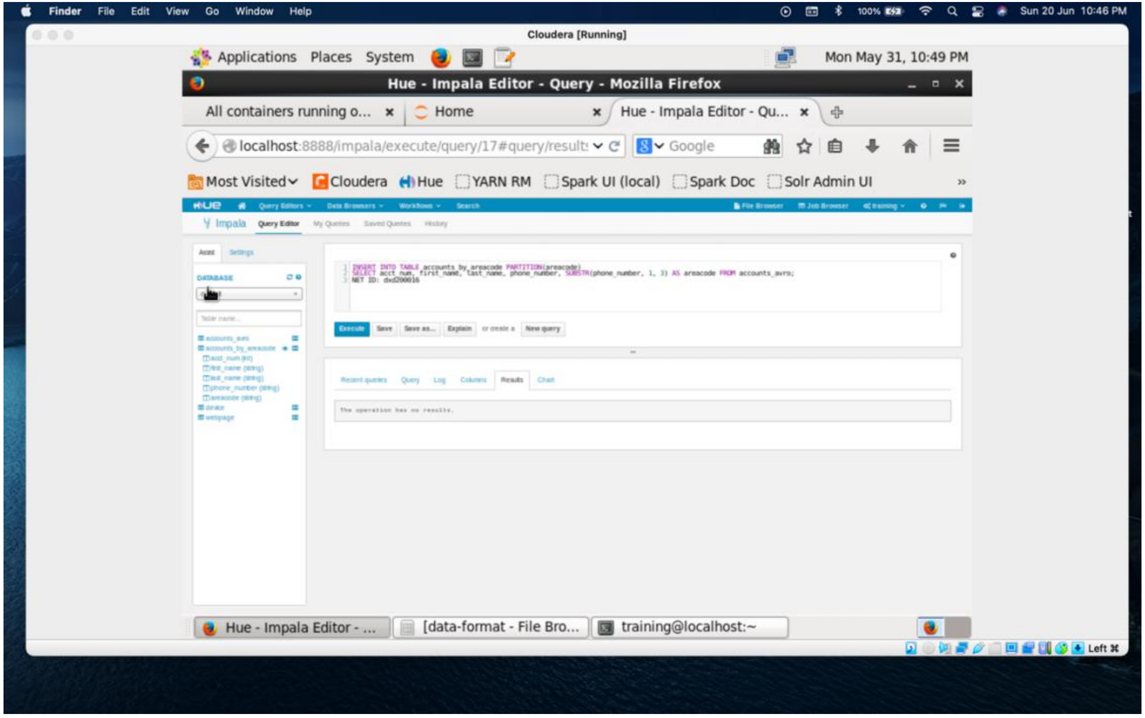


```
SELECT acct_num, first_name, last_name, phone_number, SUBSTR(phone_number, 1, 3) AS areacode FROM accounts_avro
NET ID: dxd200016
```

	acct_num	first_name	last_name	phone_number	areacode
0	64883	Jennifer	Spangler	8188883550	818
1	64884	Kathleen	Blubaugh	7074976663	707
2	64885	Daniel	Wells	5417545715	541
3	64886	Jeffrey	Owen	5416217451	541
4	64887	Marda	Hall	8310241481	831
5	64888	Elizabeth	Mathers	6579173228	657
6	64889	Sherri	Bowers	5419698820	541
7	64890	Latina	Geiger	3106528687	310
8	64891	Diane	Buchanan	7475882217	747
9	64892	Paul	Bernard	8188807733	818
10	64893	Fred	Vermillion	4244865881	424
11	64894	Eric	Chaney	5419099802	541
			Doody	9518019471	951

Hue - Impala Editor - Query - Mozilla Firefox
localhost:8888/impala/execute/query/11#navigatorTab
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart

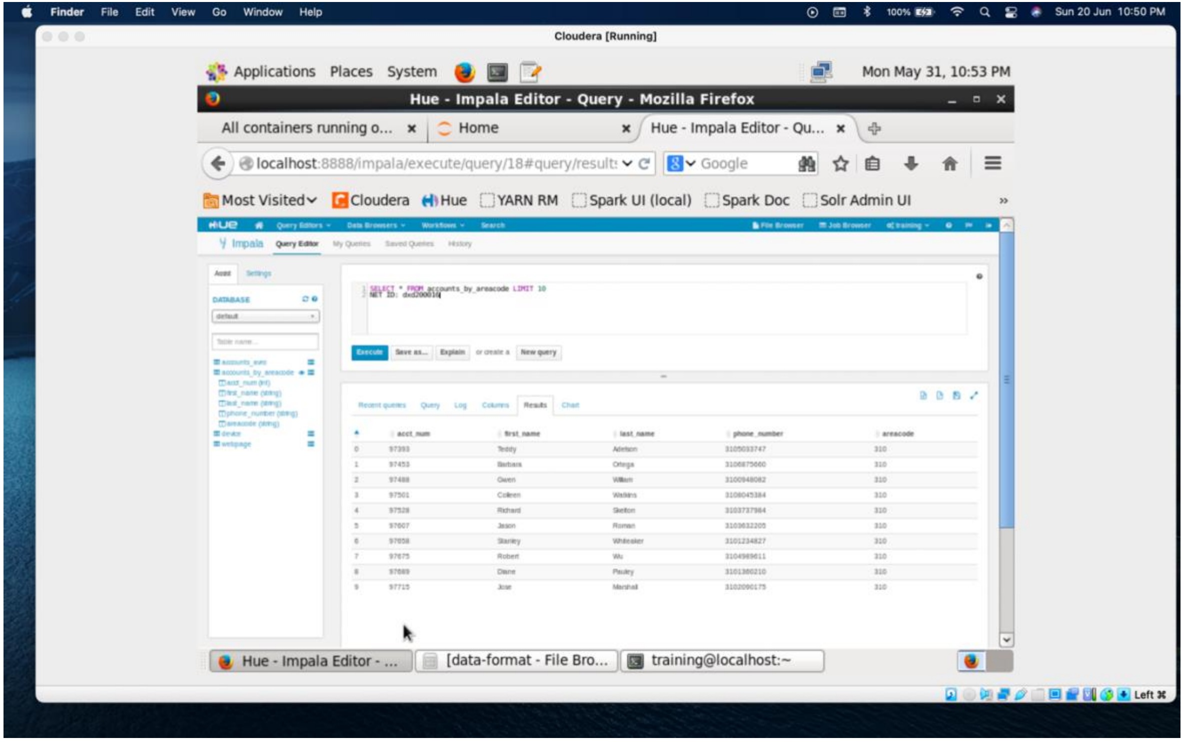
3. Then, I ran the following command to insert the specific columns from accounts_avro table to accounts_by_areacode table, dynamically partitioned by area code.
   Command: INSERT INTO TABLE accounts_by_areacode PARTITION(areacode) SELECT acct_num, first_name, last_name, phone_number, SUBSTR(phone_number,1,3) AS areacode FROM accounts_avro;



```
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;
NET ID: dxd200016
```

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

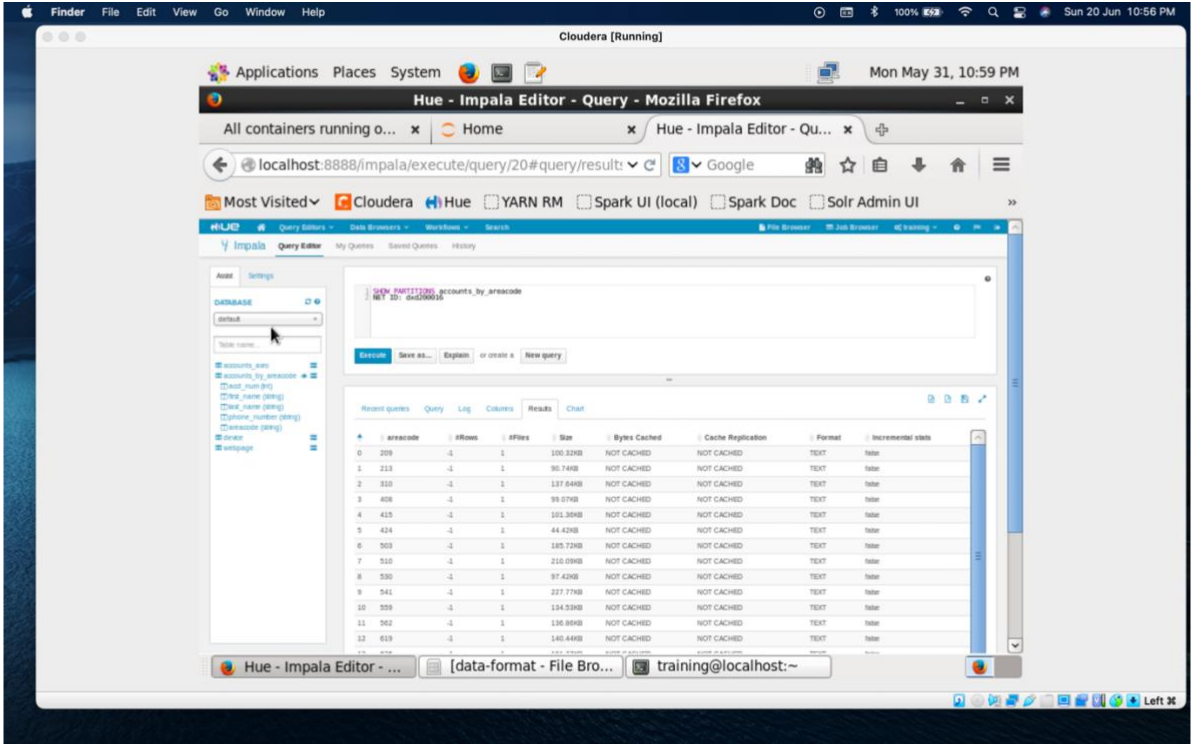
4. Then, to verify If the records are inserted successfully, I ran the following command to view the first 10 records from the accounts_by_areacode table.
Command: SELECT * FROM accounts_by_areacode LIMIT 10



```
SELECT * FROM accounts_by_areacode LIMIT 10
NET ID: dxd200016
```

	acct_num	first_name	last_name	phone_number	areacode
0	97393	Teddy	Adelson	3105033747	310
1	97453	Barbara	Ortega	3106875660	310
2	97488	Owen	William	3100948082	310
3	97501	Colleen	Watkins	3108045384	310
4	97528	Richard	Skelton	3103737984	310
5	97607	Jason	Roman	3103632205	310
6	97658	Stanley	Whiteaker	3101234827	310
7	97675	Robert	Wu	3104989611	310
8	97689	Diane	Pauley	3101380210	310
9	97715	Jose	Marshall	3102090175	310

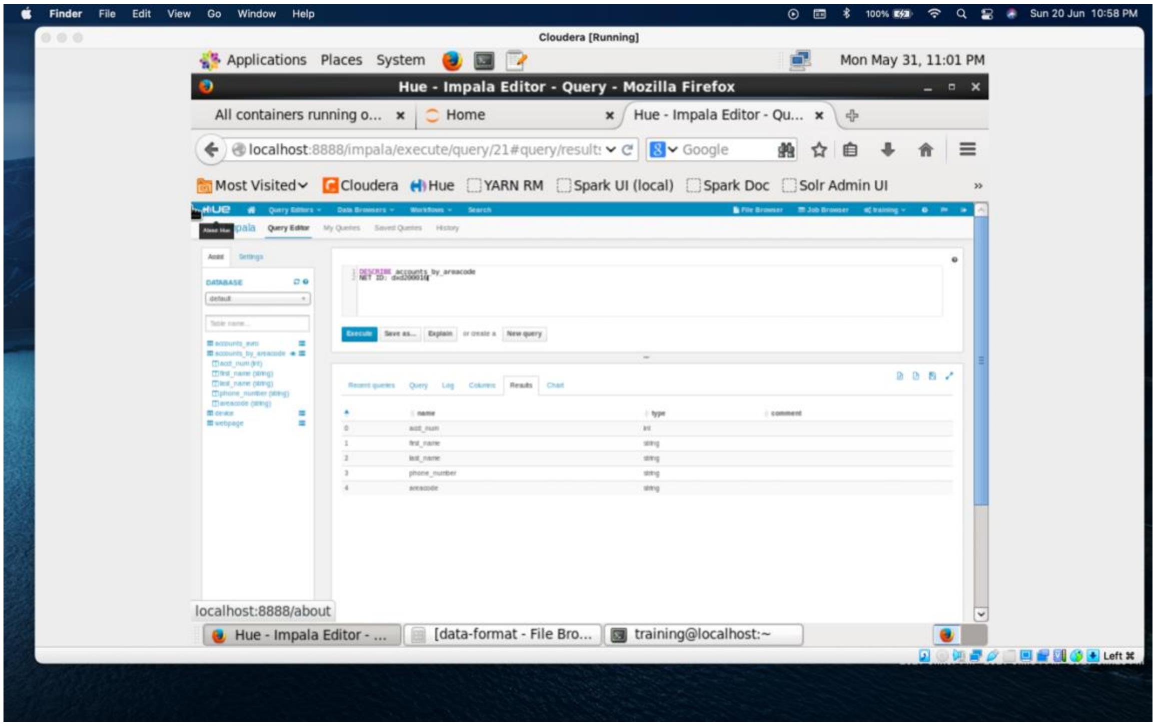
5. Then, I ran the following command to view the current partitions in the accounts_by_areacode table.
Command: SHOW PARTITIONS accounts_by_areacode



```
SHOW PARTITIONS accounts_by_areacode
NET ID: dxd200016
```

	areacode	#Rows	#Files	Size	Bytes Cached	Cache Replication	Format	Incremental stats
0	209	-1	1	100.32KB	NOT CACHED	NOT CACHED	TEXT	false
1	213	-1	1	90.74KB	NOT CACHED	NOT CACHED	TEXT	false
2	310	-1	1	137.64KB	NOT CACHED	NOT CACHED	TEXT	false
3	408	-1	1	99.07KB	NOT CACHED	NOT CACHED	TEXT	false
4	415	-1	1	101.36KB	NOT CACHED	NOT CACHED	TEXT	false
5	424	-1	1	44.42KB	NOT CACHED	NOT CACHED	TEXT	false
6	503	-1	1	185.72KB	NOT CACHED	NOT CACHED	TEXT	false
7	510	-1	1	210.09KB	NOT CACHED	NOT CACHED	TEXT	false
8	530	-1	1	97.42KB	NOT CACHED	NOT CACHED	TEXT	false
9	541	-1	1	227.77KB	NOT CACHED	NOT CACHED	TEXT	false
10	559	-1	1	134.53KB	NOT CACHED	NOT CACHED	TEXT	false
11	562	-1	1	136.86KB	NOT CACHED	NOT CACHED	TEXT	false
12	619	-1	1	140.44KB	NOT CACHED	NOT CACHED	TEXT	false

6. Finally, I ran the following command to view all the columns and their data types of the accounts_by_areacode table. Command: DESCRIBE accounts_by_areacode



```
DESCRIBE accounts_by_areacode
NET ID: dxd200010
```

	name	type	comment
0	acct_num	int	
1	first_name	string	
2	last_name	string	
3	phone_number	string	
4	areacode	string	

Hue - Impala Editor - Query - Mozilla Firefox
localhost:8888/impala/execute/query/21#query/results
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart