

Logic Explanation

Mid Submission

STEPS:

1. Please find below submitted task as part of mid submission of capstone project-Credit card fraud detection.
 1. Task1-LoadNoSQL.pdf
 2. Task - SqoopDataIngestion.pdf
 3. Task3 - CreateNoSQL.pdf
 4. Task4 - PreAnalysis.pdf
 5. Task4 - Logic Mid.pdf
 6. Task4 - Scripts Execution.pdf

HBase is used as NoSQL database for this project.

2. There are two sqoop jobs: First for incremental load of card_member table and Second for full load of member_score table.
3. Directory Setup in HDFS for the project.
`hadoop fs -mkdir /capstone_project`
4. All hadoop shell, hive, hbase and sqoop commands are executed via ec2-user.
5. Download card_transactions.csv from Upgrad learning portal in the capstone project section. Upload the file in S3 bucket. Post that create a directory in HDFS and copy card_transactions.csv in that location.

```
hadoop fs -mkdir /capstone_project/card_transactions
hadoop distcp s3://creditcapstone/credit/card_transactions.csv /capstone_project/
card_transactions/
```

Task 1: Load the transactions history data (card_transactions.csv) in a NoSQL database

1. Create new database named capstone_project in EMR cluster console using 'hive' command.

```
CREATE DATABASE capstone_project;
USE capstone_project;
(Set the DB)
```

2. Creation of external table card_transactions_ext table :

```
CREATE EXTERNAL TABLE IF NOT EXISTS card_transactions_ext(`card_id` string, `member_id` string, `amount` DOUBLE, `postcode` string, `pos_id` string, `transaction_dt` string, `status` string) row format delimited fields terminated BY ',' location '/capstone_project/card_transactions' tblproperties ("skip.header.line.count"="1");
```

3. Create table card_transactions_orc for better performance.

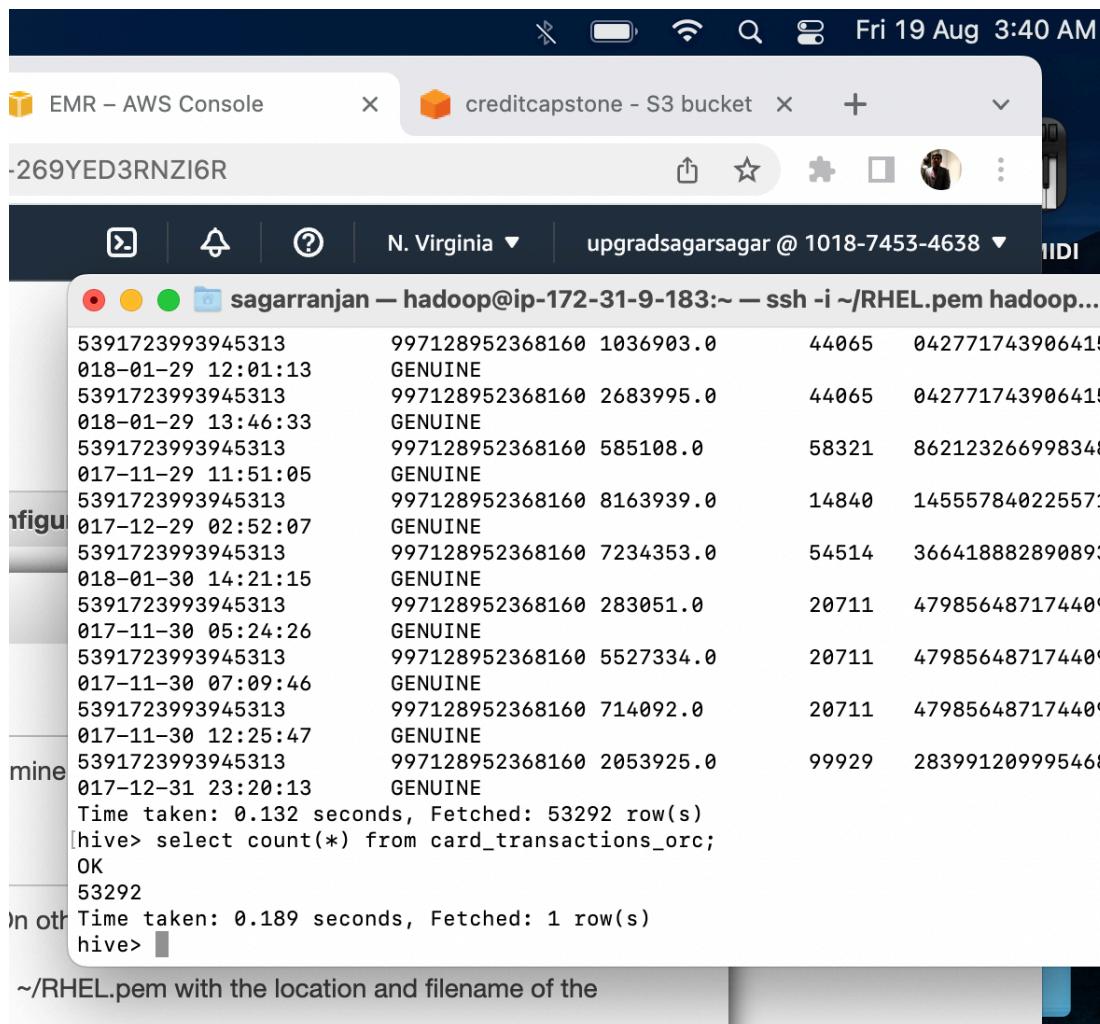
```
CREATE TABLE IF NOT EXISTS card_transactions_orc
(
  `card_id`      STRING,
  `member_id`    STRING,
  `amount`       DOUBLE,
  `postcode`     STRING,
  `pos_id`       STRING,
  `transaction_dt` TIMESTAMP,
  `status`       STRING
) stored AS orc tblproperties ("orc.compress"="SNAPPY");
```

5. Load data in card_transactions_orc:

```
INSERT overwrite TABLE card_transactions_orc
SELECT card_id,
       member_id,
       amount,
       postcode,
       pos_id,
       cast(from_unixtime(unix_timestamp(transaction_dt, 'dd-MM-
yyyy HH:mm:ss')) AS timestamp),
       status
FROM card_transactions_ext;
```

6. Check the number of records from card_transactions_orc table.

```
select count(*) from card_transactions_orc;
```



The screenshot shows a terminal window in the AWS EMR console. The title bar indicates the session is running on the creditcapstone S3 bucket in N. Virginia. The user is sagarranjan, and the session ID is 1018-7453-4638. The command run was `ssh -i ~/RHEL.pem hadoop...`. The output of the `select count(*) from card_transactions_orc;` query is displayed, showing a single row with the value 53292.

```
sagarranjan — hadoop@ip-172-31-9-183:~ — ssh -i ~/RHEL.pem hadoop...
5391723993945313      997128952368160 1036903.0      44065  042771743906415:
018-01-29 12:01:13    GENUINE
5391723993945313      997128952368160 2683995.0      44065  042771743906415:
018-01-29 13:46:33    GENUINE
5391723993945313      997128952368160 585108.0      58321  862123266998348:
017-11-29 11:51:05    GENUINE
5391723993945313      997128952368160 8163939.0     14840  145557840225571:
017-12-29 02:52:07    GENUINE
5391723993945313      997128952368160 7234353.0     54514  366418882890893:
018-01-30 14:21:15    GENUINE
5391723993945313      997128952368160 283051.0      20711  479856487174409:
017-11-30 05:24:26    GENUINE
5391723993945313      997128952368160 5527334.0     20711  479856487174409:
017-11-30 07:09:46    GENUINE
5391723993945313      997128952368160 714092.0      20711  479856487174409:
017-11-30 12:25:47    GENUINE
5391723993945313      997128952368160 2053925.0     99929  283991209995468:
017-12-31 23:20:13    GENUINE
Time taken: 0.132 seconds, Fetched: 53292 row(s)
hive> select count(*) from card_transactions_orc;
OK
53292
Time taken: 0.189 seconds, Fetched: 1 row(s)
hive>
```

~/RHEL.pem with the location and filename of the

7. Table card_transactions_hbase hive-hbase:

```

CREATE TABLE card_transactions_hbase
(
  `transaction_id` STRING,
  `card_id` STRING,
  `member_id` STRING,
  `amount` DOUBLE,
  `postcode` STRING,
  `pos_id` STRING,
  `transaction_dt` TIMESTAMP,
  `status` STRING
) row format delimited stored BY
'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH serdeproperties (
"hbase.columns.mapping"=
":key, card_transactions_family:card_id, card_transactions_family:member_id,
card_transactions_family:amount, card_transactions_family:postcode, card_transactions_family:pos_id, card_transactions_family:transaction_dt, card_transactions_family:status"
) tblproperties ("hbase.table.name"="card_transactions_hive");
  
```

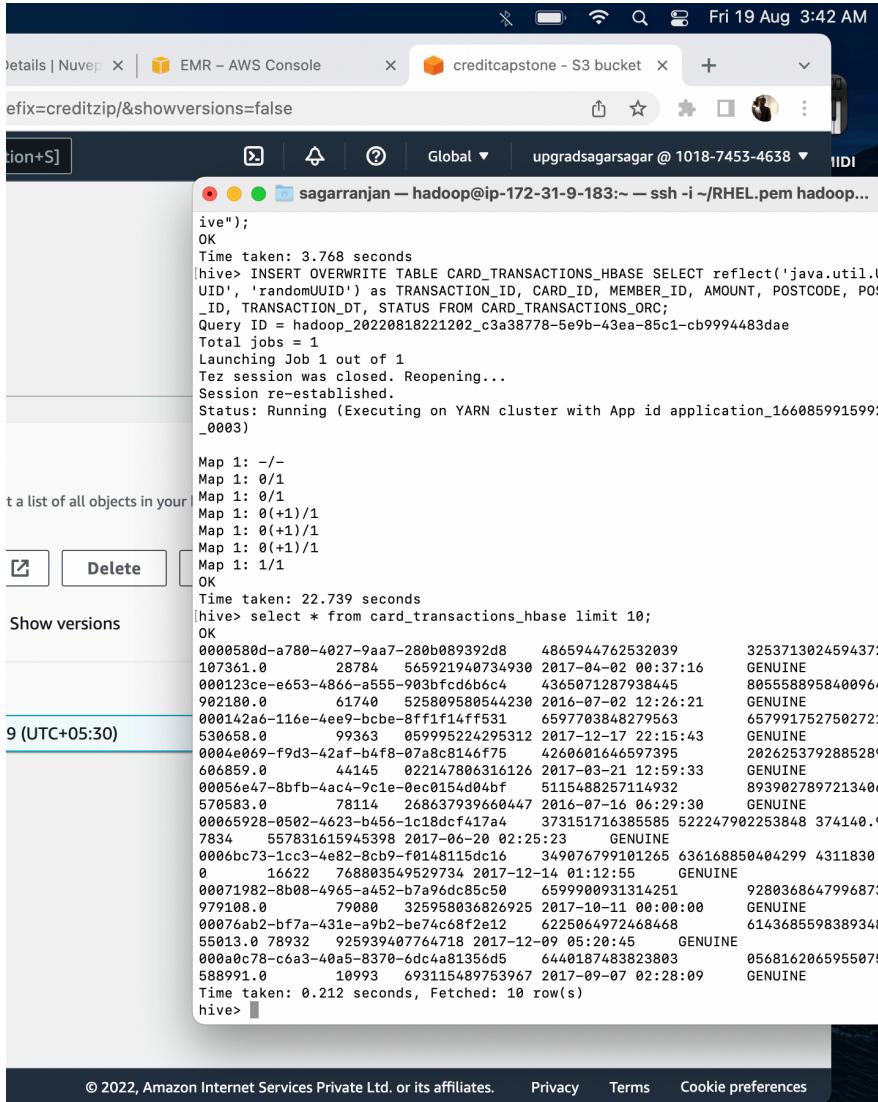
8. Load data into card_transactions_hbase

```

INSERT overwrite TABLE card_transactions_hbase
SELECT reflect('java.util.UUID', 'randomUUID') AS transaction_id,
       card_id,
       member_id,
       amount,
       postcode,
       pos_id,
       transaction_dt,
       status
FROM   card_transactions_orc;
  
```

9. Fetching first ten transactions

```
SELECT * FROM card_transactions_hbase LIMIT 10;
```



```

Fri 19 Aug 3:42 AM
Details | Nuve... X | EMR - AWS Console X | creditcapstone - S3 bucket X + v
efix=creditzip&showversions=false
tion+S] | Global ▾ | upgradsagarsagar@1018-7453-4638 ▾ | IIDI
sagarrajan — hadoop@ip-172-31-9-183:~ — ssh -i ~/RHEL.pem hadoop...
ive");
OK
Time taken: 3.768 seconds
[hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_HBASE SELECT reflect('java.util.U
UID', 'randomUUID') as TRANSACTION_ID, CARD_ID, MEMBER_ID, AMOUNT, POSTCODE, POS
_ID, TRANSACTION_DT, STATUS FROM CARD_TRANSACTIONS_ORC;
Query ID = hadoop_20220818221202_c3a38778-5e9b-43ea-85c1-cb9994483dae
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1660859915992
_0003)

Map 1: -/
Map 1: 0/1
Map 1: 0/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 1/1
OK
Time taken: 22.739 seconds
[hive> select * from card_transactions_hbase limit 10;
OK
0000580d-a780-4027-9aa7-280b089392d8 4865944762532039 3253713024594372
187361.0 28784 565921948734930 2017-04-02 00:37:16 GENUINE
000123ce-e653-4866-a555-903bfcdb66c4 4365071287938445 885588958400964
982180.0 61740 525809580544230 2016-07-07 12:26:21 GENUINE
000142a6-116e-4ee9-bcbe-8ff1f14ff531 6597703848279563 6579917527502721
530658.0 99363 059995224295312 2017-12-17 22:15:43 GENUINE
0004e069-f9d3-42af-b4f8-07a8c8146f75 4260601646597395 2826253792885289
686859.0 44145 022147806316126 2017-03-21 12:59:33 GENUINE
00056e47-8bfb-4ac4-9c1e-0ec0154d04bf 5115488257114932 8939027897213406
570583.0 78114 268637939660447 2016-07-16 06:29:30 GENUINE
00065928-0502-4623-b456-1c18dcf417a4 373151716385585 522247902253848 374140.9
7834 557831615945398 2017-06-20 02:25:23 GENUINE
0006bc73-1cc3-4eb2-8cb9-f0148115dc16 349076799101265 636168850404299 4311830.
0 16622 768803549529734 2017-12-14 01:12:55 GENUINE
00071982-8b08-4965-a452-b7a96dc85c50 65999093131251 9280368647996873
979108.0 79080 325958036826925 2017-10-11 00:00:00 GENUINE
00076ab2-bf7a-431e-a9b2-be74c68f2e12 6225064972468468 6143685598389348
55013.0 78932 925939407764718 2017-12-09 05:20:45 GENUINE
000aac78-c6a3-40a5-8370-0dc4a81356d5 6440187483823803 0568162065955075
588991.0 10993 693115489753967 2017-09-07 02:28:09 GENUINE
Time taken: 0.212 seconds, Fetched: 10 row(s)
hive>

```

© 2022, Amazon Internet Services Private Ltd. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

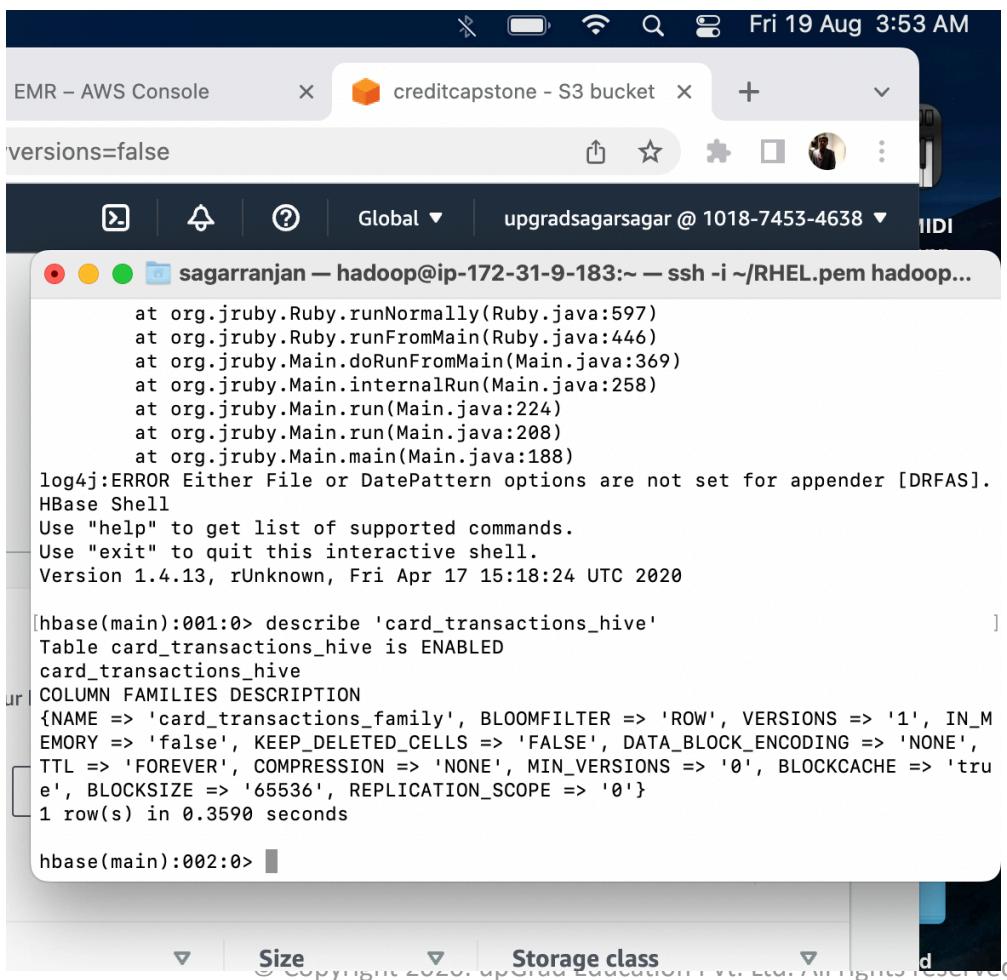
10. `Lookup_data_hbase` : Visible to HBase as well with name as `lookup_data_hive`.

```
CREATE TABLE lookup_data_hbase
(
    `card_id`          STRING,
    `ucl`              DOUBLE,
    `score`            INT,
    `postcode`         STRING,
    `transaction_dt`  TIMESTAMP
) stored BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH
serdeproperties ("hbase.columns.mapping"=
":key, lookup_card_family:ucl, lookup_card_family:score, lookup_transaction_f
amily:postcode, lookup_transaction_family:transaction_dt"
) tblproperties ("hbase.table.name" = "lookup_data_hive");
```

HBase Commands :

1. HBase shell from EMR cluster(Command : `hbase shell`)

```
describe 'card_transactions_hive'
```



The screenshot shows a terminal window titled "sagarranjan — hadoop@ip-172-31-9-183:~ — ssh -i ~/RHEL.pem hadoop..." running on an AWS EMR cluster. The terminal displays the following HBase shell session:

```

at org.jruby.Ruby.runNormally(Ruby.java:597)
at org.jruby.Ruby.runFromMain(Ruby.java:446)
at org.jruby.Main.doRunFromMain(Main.java:369)
at org.jruby.Main.internalRun(Main.java:258)
at org.jruby.Main.run(Main.java:224)
at org.jruby.Main.run(Main.java:208)
at org.jruby.Main.main(Main.java:188)
log4j:ERROR Either File or DatePattern options are not set for appender [DRFAS].
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

[hbase(main):001:0> describe 'card_transactions_hive'
Table card_transactions_hive is ENABLED
card_transactions_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'card_transactions_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_M
EMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE',
TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'tr
ue', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.3590 seconds

hbase(main):002:0>
```

2. Counting Records in card_transactions_hive in HBase

```
COUNT 'card_transactions_hive'
```



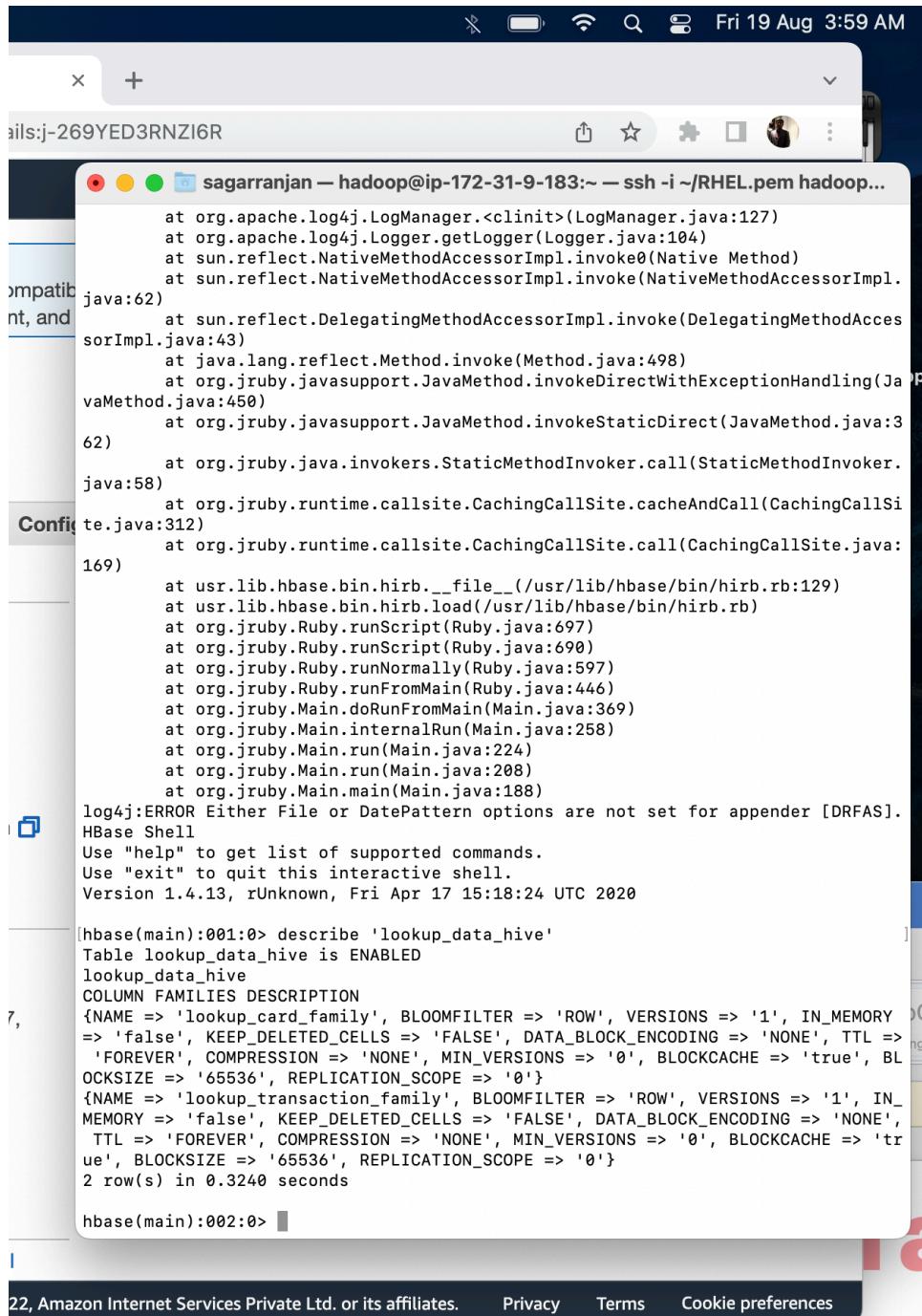
```

Fri 19 Aug 3:57 AM
+ 69YED3RNZI6R
sagarranjan — hadoop@ip-172-31-9-183:~ — ssh -i ~/RHEL.pem hadoop...
Current count: 8000, row: 260b24b4-4ec6-4ef8-a337-c0856e543e7b
Current count: 9000, row: 2b053e0d-0b6a-4370-9239-6a08a7339f6b
Current count: 10000, row: 2fa8dabf-109b-41c3-9690-00bc5306d3d3
Current count: 11000, row: 343fb1e8-3415-4c31-ac7f-81dffbb291075
Current count: 12000, row: 39445efd-882d-41cb-8629-3e38af30b237
Current count: 13000, row: 3e172440-512d-44d6-a7c4-5983969686af
Current count: 14000, row: 430f5f9b-7574-427c-9618-131daea852b9
Current count: 15000, row: 47d117c9-eadd-4ba1-8ab4-9eba2cda03b7
Current count: 16000, row: 4c72a279-9740-4ad9-a4d2-d6a276c8023a
Current count: 17000, row: 515220ed-1c06-457a-89e4-baaa5b344e62
Current count: 18000, row: 5638858a-1e07-4c31-89c2-2fd5fdc33c9
Current count: 19000, row: 5b28e882-bba0-407c-b9da-17e5c88d6d9e
Current count: 20000, row: 600bf5bf-b648-44b4-801c-5eacd702c9d2
Current count: 21000, row: 65364b3c-e0b0-4f11-95b6-e1211ca3223a
Current count: 22000, row: 69e52043-8655-49c9-940f-ae14c7c7e13c
Current count: 23000, row: 6e8897f9-c93d-4ab6-9f8e-257ed1392ed4
Current count: 24000, row: 732ec559-46e4-4850-801c-cf333871b7b6
Current count: 25000, row: 77d39b26-59f2-47da-82ee-69898a7d21e1
Current count: 26000, row: 7ccb8185-5ab2-4676-a5fa-724cc2cc6a24
Current count: 27000, row: 818da37b-39fa-4347-8191-7211af9dd807
Current count: 28000, row: 8663a763-7b11-4ab1-bff5-da5a221cc47a
Current count: 29000, row: 8af509f8-3398-48ef-9b86-728359b7e5fe
Current count: 30000, row: 8fe633b7-592e-4691-9fc5-bc11d7a40068
Current count: 31000, row: 94d250cc-49a9-4d1f-95e6-94f84c19c5f0
Current count: 32000, row: 999609bc-2c4d-43c7-9df9-ffa5ab9aff7
Current count: 33000, row: 9ed4a774-421e-4c03-a620-8c23a5d1d7dd
Current count: 34000, row: a39bd08a-7167-42b3-9c35-e581d042d682
Current count: 35000, row: a87577da-fded-446b-abd9-9c29049d8b68
Current count: 36000, row: ad2e64b4-200c-4eaa-a586-4a55a2b175d9
Current count: 37000, row: b20f9255-a393-4cf6-96ee-87f23f9aec5a
Current count: 38000, row: b6ff5d50-07d5-45ca-b076-4e46359f1640
Current count: 39000, row: bbee76b0-6b0d-4a89-b4cf-4d8578853900
Current count: 40000, row: c0c12f97-54bc-49f6-bb34-e7ba85f63a15
Current count: 41000, row: c59edbe4-d1d6-4a91-8d2a-85144e1fe2a0
Current count: 42000, row: ca7fe4d2-ad13-4119-84f9-700344bd2152
Current count: 43000, row: cf4269ad-1ca3-44d5-8c13-8e74dccdb9a
Current count: 44000, row: d3c6beel-1902-4ac6-990b-92739fb5918
Current count: 45000, row: d8d032ae-013e-4015-8a50-b541f0addc85
Current count: 46000, row: dd7cb0ae-3be5-4c70-83f9-03d37397f8e8
Current count: 47000, row: e24fdfad-3a6f-40aa-8453-47b2e301339b
Current count: 48000, row: e6dc9eb6-0cf3-4114-a53b-7b926f70b2c7
Current count: 49000, row: eba003a2-a14b-4317-8bb2-97aad30b8583
Current count: 50000, row: f06ad594-c60b-483f-b76c-e809c87b3647
Current count: 51000, row: f560aac2-59f9-4be4-b04b-13cb202540ce
Current count: 52000, row: f9ea71de-758b-4d27-a151-0cf1475f3e82
Current count: 53000, row: feba8498-dc06-4271-a388-eb4d2ecad91e
53292 row(s) in 3.5940 seconds

=> 53292
[hadoop@ip-172-31-9-183 ~]$
```

3. Describe lookup_data_hive hive-hbase in hBase

```
describe 'lookup_data_hive'
```



The screenshot shows a terminal window titled 'sagarranjan — hadoop@ip-172-31-9-183:~ — ssh -i ~/RHEL.pem hadoop...'. The window displays the Java stack trace for a log4j error, followed by the HBase shell prompt and the output of the 'describe' command.

```

at org.apache.log4j.LogManager.<clinit>(LogManager.java:127)
at org.apache.log4j.Logger.getLogger(Logger.java:104)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.jruby.javasupport.JavaMethod.invokeDirectWithExceptionHandling(Ja
vaMethod.java:450)
at org.jruby.javasupport.JavaMethod.invokeStaticDirect(JavaMethod.java:3
62)
at org.jruby.java.invokers.StaticMethodInvoker.call(StaticMethodInvoker.
java:58)
at org.jruby.runtime.callsite.CachingCallSite.cacheAndCall(CachingCallsit
e.java:312)
at org.jruby.runtime.callsite.CachingCallSite.call(CachingCallSite.java:
169)
at usr.lib.hbase.bin.hirb.__file__(/usr/lib/hbase/bin/hirb.rb:129)
at usr.lib.hbase.bin.hirb.load(/usr/lib/hbase/bin/hirb.rb)
at org.jruby.Ruby.runScript(Ruby.java:697)
at org.jruby.Ruby.runScript(Ruby.java:690)
at org.jruby.Ruby.runNormally(Ruby.java:597)
at org.jruby.Ruby.runFromMain(Ruby.java:446)
at org.jruby.Main.doRunFromMain(Main.java:369)
at org.jruby.Main.internalRun(Main.java:258)
at org.jruby.Main.run(Main.java:224)
at org.jruby.Main.run(Main.java:208)
at org.jruby.Main.main(Main.java:188)
log4j:ERROR Either File or DatePattern options are not set for appender [DRFAS].
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

[hbase(main):001:0> describe 'lookup_data_hive'
Table lookup_data_hive is ENABLED
lookup_data_hive
COLUMN FAMILIES DESCRIPTION
, {NAME => 'lookup_card_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY
=> 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL =>
'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BL
OCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'lookup_transaction_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_
MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE',
TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'tr
ue', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
2 row(s) in 0.3240 seconds

hbase(main):002:0>
```

22, Amazon Internet Services Private Ltd. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

4.

In HBase, `lookup_data_hive` table has been altered and set VERSIONS to 10 for `lookup_transaction_family`.

Need to store last 10 transactions in lookup table so altering VERSIONS to 10.

I have created two column in lookup table namely `lookup_card_family` and `lookup_transaction_family`.

```
ALTER 'lookup_data_hive', {NAME => 'lookup_transaction_family',  
VERSIONS => 10}
```

For ingesting data from the RDS server into HDFS.

- Run below Sqoop command to import member_score table from RDS into HDFS, from command prompt.

```
sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data --username upgraduser --password upgraduser --table member_score --null-string 'NA' --null-non-string '\\N' --delete-target-dir --target-dir '/capstone_project/member_score' -m1
```

- Run below Sqoop command to import card_member table from RDS into HDFS, from command prompt.

```
sqoop import --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data --username upgraduser --password upgraduser --table card_member --null-string 'NA' --null-non-string '\\N' --delete-target-dir --target-dir '/capstone_project/card_member' -m 1
```

- Hive : Create external table card_member_ext which will point to HDFS location to hold (card_member table) date in RDS.

```
CREATE EXTERNAL TABLE IF NOT EXISTS card_member_ext(`card_id` string, `member_id` string, `member_joining_dt` timestamp, `card_purchase_dt` string, `country` string, `city` string) row format delimited fields terminated BY ',' location '/capstone_project/card_member';
```

- Create external table member_score_ext which will point to HDFS location to hold data from (member_score table) in RDS.

```
CREATE EXTERNAL TABLE IF NOT EXISTS member_score_ext(`member_id` string, `score` int) row format delimited fields terminated BY ',' location '/capstone_project/member_score';
```

- For performance.

```
CREATE TABLE IF NOT EXISTS card_member_orc
(
  `card_id`          STRING,
  `member_id`        STRING,
  `member_joining_dt` TIMESTAMP,
  `card_purchase_dt` STRING,
  `country`          STRING,
  `city`              STRING
) stored AS orc tblproperties ("orc.compress"="SNAPPY");
```

4. For performance.

```
CREATE TABLE IF NOT EXISTS member_score_orc
(
    `member_id` STRING,
    `score`      INT
) stored AS orc tblproperties ("orc.compress"="SNAPPY");
```

5. Load data into card_member_orc.

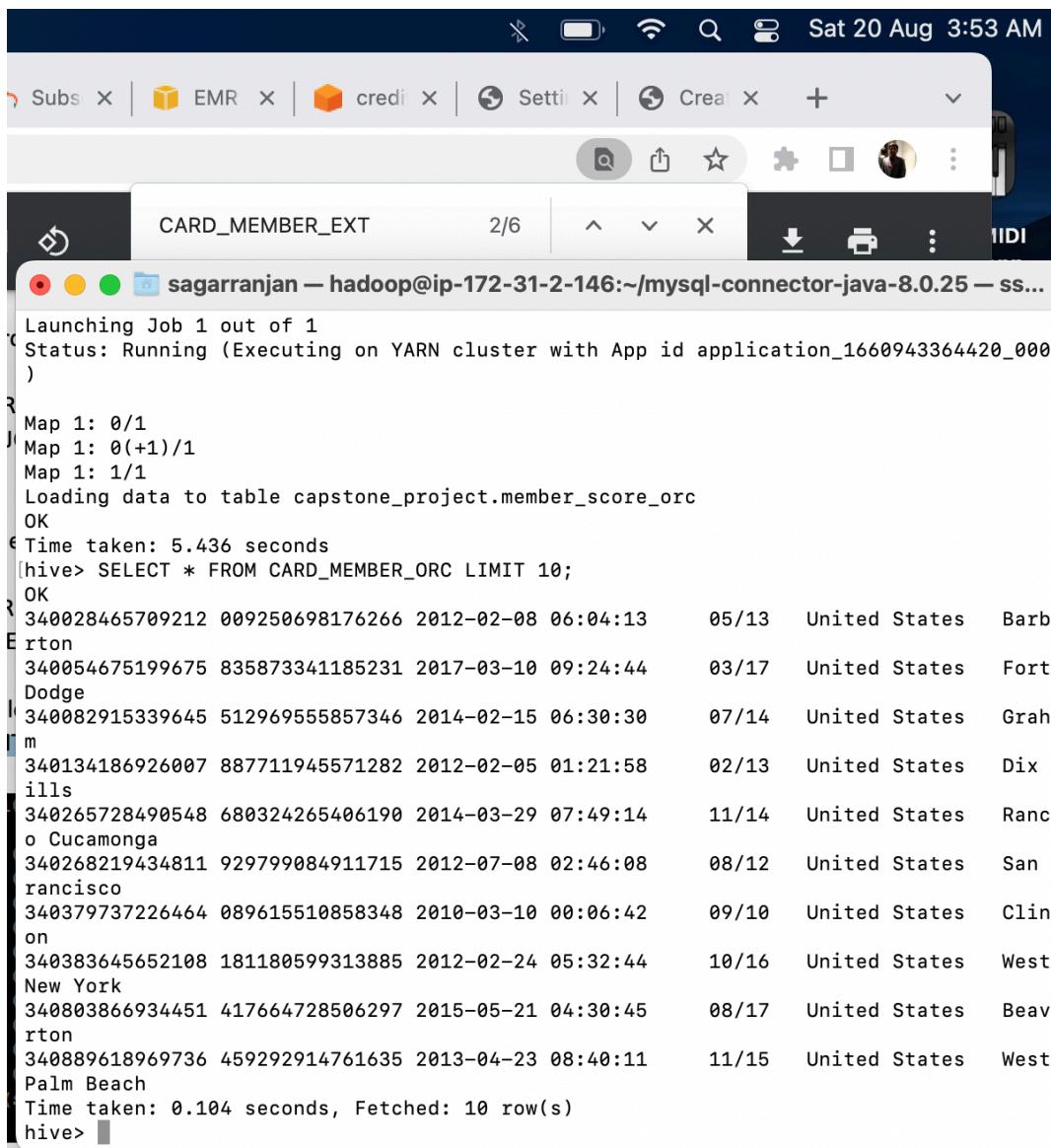
```
INSERT overwrite TABLE card_member_orc
SELECT card_id,
       member_id,
       member_joining_dt,
       card_purchase_dt,
       country,
       city
FROM   card_member_ext;
```

6. Load data into member_score_orc

```
INSERT overwrite TABLE member_score_orc
SELECT member_id,
       score
FROM   member_score_ext;
```

7. Verify data in card_member_orc table.

```
SELECT * FROM card_member_orc LIMIT 10;
```



Sat 20 Aug 3:53 AM

CARD_MEMBER_EXT 2/6

sagarranjan — hadoop@ip-172-31-2-146:~/mysql-connector-java-8.0.25 — ss...

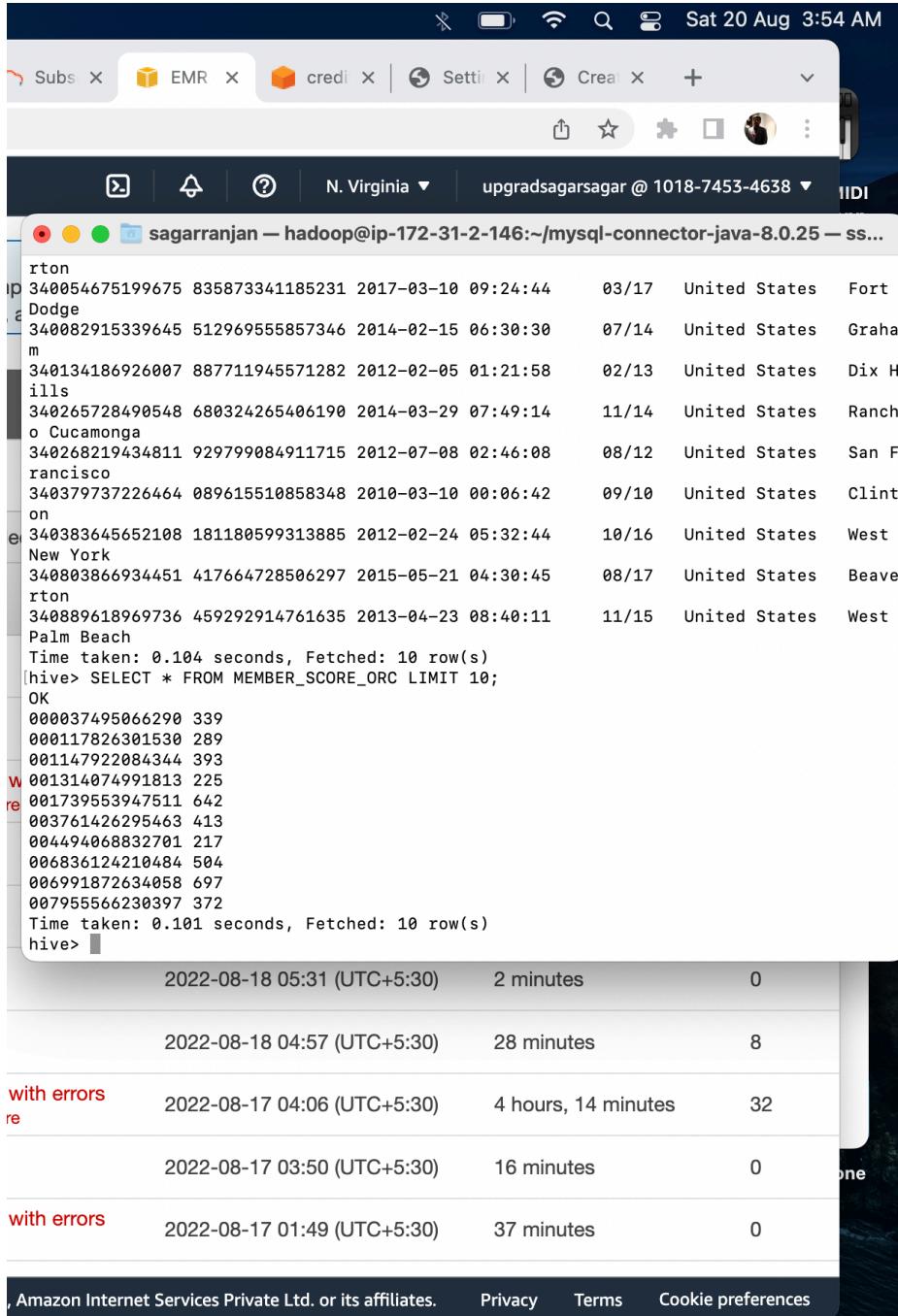
```

Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1660943364420_0008
)

Map 1: 0/1
Map 1: 0(+1)/1
Map 1: 1/1
Loading data to table capstone_project.member_score_orc
OK
Time taken: 5.436 seconds
hive> SELECT * FROM CARD_MEMBER_ORC LIMIT 10;
OK
340028465709212 009250698176266 2012-02-08 06:04:13      05/13  United States  Barbe
rton
340054675199675 835873341185231 2017-03-10 09:24:44      03/17  United States  Fort
Dodge
340082915339645 512969555857346 2014-02-15 06:30:30      07/14  United States  Graha
m
340134186926007 887711945571282 2012-02-05 01:21:58      02/13  United States  Dix F
ills
340265728490548 680324265406190 2014-03-29 07:49:14      11/14  United States  Ranch
o Cucamonga
340268219434811 929799084911715 2012-07-08 02:46:08      08/12  United States  San F
rancisco
340379737226464 089615510858348 2010-03-10 00:06:42      09/10  United States  Clint
on
340383645652108 181180599313885 2012-02-24 05:32:44      10/16  United States  West
New York
340803866934451 417664728506297 2015-05-21 04:30:45      08/17  United States  Beave
rton
340889618969736 459292914761635 2013-04-23 08:40:11      11/15  United States  West
Palm Beach
Time taken: 0.104 seconds, Fetched: 10 row(s)
hive>
```

8. Verify some data in
member_score_orc table.

```
SELECT * FROM member_score_orc LIMIT 10;
```



```
sagarranjan — hadoop@ip-172-31-2-146:~/mysql-connector-java-8.0.25 — ss...
+-----+-----+-----+-----+
| rton | 340054675199675 | 835873341185231 | 2017-03-10 09:24:44 |
| Dodge | 340082915339645 | 512969555857346 | 2014-02-15 06:30:30 |
| m | 340134186926007 | 887711945571282 | 2012-02-05 01:21:58 |
| ills | 340265728490548 | 680324265406190 | 2014-03-29 07:49:14 |
| o Cucamonga | 340268219434811 | 929799084911715 | 2012-07-08 02:46:08 |
| rancisco | 340379737226464 | 089615510858348 | 2010-03-10 00:06:42 |
| on | 340383645652108 | 181180599313885 | 2012-02-24 05:32:44 |
| New York | 340803866934451 | 417664728506297 | 2015-05-21 04:30:45 |
| rton | 340889618969736 | 459292914761635 | 2013-04-23 08:40:11 |
| Palm Beach | Time taken: 0.104 seconds, Fetched: 10 row(s) |
hive> SELECT * FROM MEMBER_SCORE_ORC LIMIT 10;
OK
000037495066290 339
000117826301530 289
001147922084344 393
001314074991813 225
001739553947511 642
003761426295463 413
004494068832701 217
006836124210484 504
006991872634058 697
007955566230397 372
Time taken: 0.101 seconds, Fetched: 10 row(s)
hive> 
```

	Time taken	Fetched
2022-08-18 05:31 (UTC+5:30)	2 minutes	0
2022-08-18 04:57 (UTC+5:30)	28 minutes	8
2022-08-17 04:06 (UTC+5:30)	4 hours, 14 minutes	32
2022-08-17 03:50 (UTC+5:30)	16 minutes	0
2022-08-17 01:49 (UTC+5:30)	37 minutes	0

, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences

Write a script to calculate the moving average and standard deviation of the last 10 transactions for each card_id for the data present in Hadoop and NoSQL database. If the total number of transactions for a particular card_id is less than 10, then calculate the parameters based on the total number of records available for that card_id. The script should be able to extract and feed the other relevant data ('postcode', 'transaction_dt', 'score', etc.) for the look-up table along with card_id and UCL.

Hive :

1. Table ranked_card_transactions_orc to store last 10 transactions for each card_id.

```
CREATE TABLE IF NOT EXISTS ranked_card_transactions_orc
(
  `card_id`      STRING,
  `amount`       DOUBLE,
  `postcode`     STRING,
  `transaction_dt` TIMESTAMP,
  `rank`         INT
) stored AS orc tblproperties ("orc.compress"="SNAPPY");
```

2. Table card_ucl_orc to store UCL values for each card_id.

```
CREATE TABLE IF NOT EXISTS card_ucl_orc
(
  `card_id`  STRING,
  `ucl`       DOUBLE
) stored AS orc tblproperties ("orc.compress"="SNAPPY");
```

3. Load data in ranked_card_transactions_orc table.

```
INSERT overwrite TABLE ranked_card_transactions_orc
SELECT b.card_id,
       b.amount,
       b.postcode,
       b.transaction_dt,
       b.rank
FROM   (
          SELECT    a.card_id,
                    a.amount,
                    a.postcode,
                    a.transaction_dt,
                    rank() OVER(partition BY a.card_id ORDER BY a.transaction_dt DESC, amount DESC) AS rank
          FROM    (
                    SELECT  card_id,
                            amount,
                            postcode,
                            transaction_dt
                    FROM   card_transactions_hbase
                    WHERE  status = 'GENUINE') a ) b
WHERE  b.rank <= 10;
```

4. Load data in card_ucl_orc table

```

INSERT overwrite TABLE card_ucl_orc
SELECT a.card_id,
       (a.average + (3 * a.standard_deviation)) AS ucl
FROM (
      SELECT card_id,
             avg(amount)    AS average,
             stddev(amount) AS standard_deviation
      FROM ranked_card_transactions_orc
      GROUP BY card_id) a;
  
```

5. Load data in lookup_data_hbase table.

```

INSERT overwrite TABLE lookup_data_hbase
SELECT rcto.card_id,
       cuo.ucl,
       cms.score,
       rcto.postcode,
       rcto.transaction_dt
FROM ranked_card_transactions_orc rcto
JOIN card_ucl_orc cuo
ON cuo.card_id = rcto.card_id
JOIN (
      SELECT DISTINCT card.card_id,
                     score.score
      FROM card_member_orc card
      JOIN member_score_orc score
      ON card.member_id = score.member_id) AS c
ms
ON rcto.card_id = cms.card_id
WHERE rcto.rank = 1;
  
```

6. Count in

lookup_data_hbase table.

Sat 20 Aug 3:56 AM

Subs X EMR X credi X Setting X Create X +

N. Virginia ▾ upgradsagarsagar @ 1018-7453-4638 ▾ IIID

sagarranjan — hadoop@ip-172-31-2-146:~/mysql-connector-java-8.0.25 — ss...

```

Map 1: 0(+1)/1  Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 0/2
Map 1: 0(+1)/1  Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 0(+2)/2
Map 1: 0(+1)/1  Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 1(+1)/2
Map 1: 0(+1)/1  Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 2/2
Map 1: 1/1       Map 2: 1/1      Map 3: 1/1      Map 5: 1/1      Reducer 4: 2/2
OK
Time taken: 16.149 seconds
[hive> select count(*) from lookup_data_hbase;
Query ID = hadoop_20220819222630_997ea5fb-e55f-4195-a848-8d4cf2539dea
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1660943364420_0008
)
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 1/1       Reducer 2: 0(+1)/1
Map 1: 1/1       Reducer 2: 1/1
OK
999
Time taken: 2.565 seconds, Fetched: 1 row(s)
[hive> select * from lookup_data_hbase limit 10;
OK
340028465709212 1.6331555548882348E7    233      24658    2018-01-02 03:25:35
340054675199675 1.4156079786189131E7    631      50140    2018-01-15 19:43:23
340082915339645 1.5285685330791473E7    407      17844    2018-01-26 19:03:47
340134186926007 1.5239767522438556E7    614      67576    2018-01-18 23:12:50
340265728490548 1.608491671255562E7    202      72435    2018-01-21 02:07:35
340268219434811 1.2507323937605347E7    415      62513    2018-01-16 04:30:05
340379737226464 1.4198310998368107E7    229      26656    2018-01-27 00:19:47
340383645652108 1.4091750460468251E7    645      34734    2018-01-29 01:29:12
340803866934451 1.0843341196185412E7    502      87525    2018-01-31 04:23:57
340889618969736 1.3217942365515321E7    330      61341    2018-01-31 21:57:18
Time taken: 0.224 seconds, Fetched: 10 row(s)
hive>
```

2022-08-18 05:31 (UTC+5:30)	2 minutes	0	
2022-08-18 04:57 (UTC+5:30)	28 minutes	8	
with errors	2022-08-17 04:06 (UTC+5:30)	4 hours, 14 minutes	32
	2022-08-17 03:50 (UTC+5:30)	16 minutes	0
with errors	2022-08-17 01:49 (UTC+5:30)	37 minutes	0

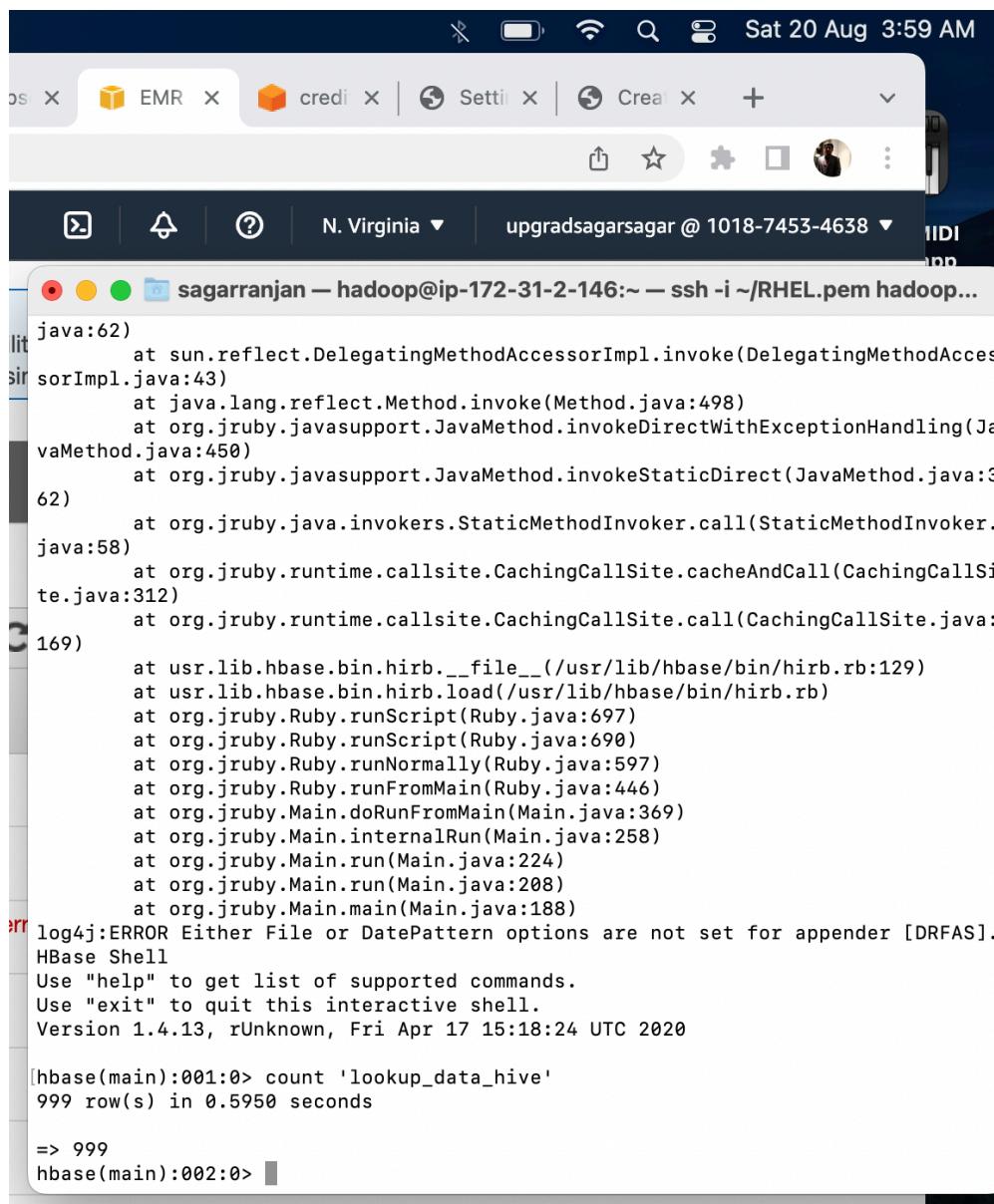
, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences

HBase Commands:

1. Start HBase shell from command prompt. In HBase,

Verify count in lookup_data_hive table.

```
COUNT 'lookup_data_hive'
```



The screenshot shows a terminal window titled "sagarranjan — hadoop@ip-172-31-2-146:~ — ssh -i ~/RHEL.pem hadoop...". The window displays Java stack trace information followed by the HBase shell prompt and command output.

```

java:62)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccess
sorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at org.jruby.javasupport.JavaMethod.invokeDirectWithExceptionHandling(Ja
vaMethod.java:450)
    at org.jruby.javasupport.JavaMethod.invokeStaticDirect(JavaMethod.java:3
62)
    at org.jruby.java.invokers.StaticMethodInvoker.call(StaticMethodInvoker.
java:58)
    at org.jruby.runtime.callsite.CachingCallSite.cacheAndCall(CachingCallsit
e.java:312)
    at org.jruby.runtime.callsite.CachingCallSite.call(CachingCallSite.java:
169)
    at usr.lib.hbase.bin.hirb.__file__(/usr/lib/hbase/bin/hirb.rb:129)
    at usr.lib.hbase.bin.hirb.load(/usr/lib/hbase/bin/hirb.rb)
    at org.jruby.Ruby.runScript(Ruby.java:697)
    at org.jruby.Ruby.runScript(Ruby.java:690)
    at org.jruby.Ruby.runNormally(Ruby.java:597)
    at org.jruby.Ruby.runFromMain(Ruby.java:446)
    at org.jruby.Main.doRunFromMain(Main.java:369)
    at org.jruby.Main.internalRun(Main.java:258)
    at org.jruby.Main.run(Main.java:224)
    at org.jruby.Main.run(Main.java:208)
    at org.jruby.Main.main(Main.java:188)
log4j:ERROR Either File or DatePattern options are not set for appender [DRFAS].
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

[hbase(main):001:0> count 'lookup_data_hive'
999 row(s) in 0.5950 seconds
=> 999
hbase(main):002:0>
]
```

2. In HBase, verify data in

lookup_data_hive table.

```
scan 'lookup_data_hive
```

```
s X EMR X credi X Settings X Create X + v
upgradsagarsagar@1018-7453-4638 N. Virginia
IDI
sagarranjan — hadoop@ip-172-31-2-146:~— ssh -i ~/RHEL.pem hadoop...
=1660947987244, value=2018-01-30 02:03:54
column=lookup_card_family:score, timestamp=1660947987244,
value=412
6595928469079750 column=lookup_card_family:ucl, timestamp=1660947987244, va
lue=1.142797041440079E7
6595928469079750 column=lookup_transaction_family:postcode, timestamp=16609
47987244, value=98349
6595928469079750 column=lookup_transaction_family:transaction_dt, timestamp
=1660947987244, value=2018-01-24 12:38:22
6597703848279563 column=lookup_card_family:score, timestamp=1660947987244,
value=218
6597703848279563 column=lookup_card_family:ucl, timestamp=1660947987244, va
lue=1.4718634149498457E7
6597703848279563 column=lookup_transaction_family:postcode, timestamp=16609
47987244, value=95699
6597703848279563 column=lookup_transaction_family:transaction_dt, timestamp
=1660947987244, value=2018-01-27 10:51:49
6598830758632447 column=lookup_card_family:score, timestamp=1660947987244,
value=293
6598830758632447 column=lookup_card_family:ucl, timestamp=1660947987244, va
lue=1.2227949982601807E7
6598830758632447 column=lookup_transaction_family:postcode, timestamp=16609
47987244, value=19421
6598830758632447 column=lookup_transaction_family:transaction_dt, timestamp
=1660947987244, value=2018-01-30 00:18:34
6599900931314251 column=lookup_card_family:score, timestamp=1660947987244,
value=297
6599900931314251 column=lookup_card_family:ucl, timestamp=1660947987244, va
lue=1.2121408572464656E7
6599900931314251 column=lookup_transaction_family:postcode, timestamp=16609
47987244, value=97423
6599900931314251 column=lookup_transaction_family:transaction_dt, timestamp
=1660947987244, value=2018-01-31 11:25:16
999 row(s) in 1.4300 seconds
hbase(main):003:0>
```

