Scripts Execution

Screenshots of the execution of the scripts written

Loading Historical Transactions Data into NoSQL Database

Commands to load the past transactions data into NoSQL database

————————————————HIVE————————————————-

- Creating database in hive

create database capstone_project; use capstone_project;

- Setting some parameters in hive

set hive.auto.convert.join=false;
set hive.stats.autogather=true;
set orc.compress=SNAPPY;
set hive.exec.compress.output=true;
set
mapred.output.compression.codec=org.apache.hadoop.io.com
press.SnappyCod ec; set
mapred.output.compression.type=BLOCK;

set mapreduce.map.java.opts=-Xmx5G;
set mapreduce.reduce.java.opts=-Xmx5G;
set mapred.child.java.opts=-Xmx5G -XX:
+UseConcMarkSweepGC -XX:- UseGCOverheadLimit;

```
[root@ip-172-31-21-211 ~]# hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database capstone_project;
OK
Time taken: 1.513 seconds
hive> use capstone_project;
OK
Time taken: 0.09 seconds
hive> set hive.auto.convert.join=false;
hive> set hive.stats.autogather=true;
hive> set orc.compress=SNAPPY;
hive> set hive.exec.compress.output=true;
hive> set mapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCodec; set mapred.output.compression.type=BLOCK;
hive> set mapreduce.map.java.opts=-Xmx5G;
hive> set mapreduce.reduce.java.opts=-Xmx5G;
hive> set mapred.child.java.opts=-Xmx5G -XX:+UseConcMarkSweepGC -XX:-UseGCOverheadLimit;
hive>
```

- Creating directory in hdfs to store card transaction csv file

hadoop fs -mkdir /user/root/card_transaction

- Uploading csv file from S3 to hdfs

hadoop distcp s3://capstone-aditya/input/ card_transactions.csv hdfs:/user/root/ card_transaction/

```
xMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sou
3://capstone-aditya/input/card_transactions.csv], targetPath=hdfs:/user/root/card_transaction, targetPathExists=true, filtersFile='null'}
INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-21-211.ec2.internal/172.31.21.211:8032
INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
INFO tools.SimpleCopyListing: Build file listing completed.
INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
INFO tools.DistCp: Number of paths in the copy list: 1
INFO tools.DistCp: Number of paths in the copy list: 1
INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-21-211.ec2.internal/172.31.21.211:8032
INFO mapreduce.JobSubmitter: number of splits:1
INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652465715257_0001
INFO impl.YarnClientImpl: Submitted application application_1652465715257_0001
INFO mapreduce.Job: The url to track the job: http://ip-172-31-21-211.ec2.internal:20888/proxy/application_1652465715257_0001/
INFO tools.DistCp: DistCp job-id: job_1652465715257_0001
INFO mapreduce.Job: Running job: job_1652465715257_0001
INFO mapreduce.Job: Job job_1652465715257_0001 running in uber mode : false
INFO mapreduce.Job:  map 0% reduce 0%
INFO mapreduce.Job:  map 100% reduce 0%
INFO mapreduce.Job: Job job_1652465715257_0001 completed successfully
INFO mapreduce.Job: Counters: 38
m Counters
```

- Checking uploaded csv file

hadoop fs -ls /user/root/card_transaction/

```
                Bytes Copied=4829520
                Bytes Expected=4829520
                Files Copied=1
[root@ip-172-31-21-211 ~]# hadoop fs -ls /user/root/card_transaction/
Found 1 items
-rw-r--r--   1 root hadoop    4829520 2022-05-13 19:12 /user/root/card_transaction/card_transactions.csv
[root@ip-172-31-21-211 ~]#
```

hadoop fs -cat /user/root/card_transaction/ card_transactions.csv

```
[root@ip-172-31-21-211 ~]# hadoop fs -cat /user/root/card_transaction/card_transactions.csv
card_id,member_id,amount,postcode,pos_id,transaction_dt,status
348702330256514,000037495066290,9084849,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,330148,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,136052,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,4310362,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,9097094,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,2291118,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,4900011,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,633447,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,6259303,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,369067,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,1193207,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,9335696,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,2241736,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,457701,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,7176668,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,5585098,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,7918756,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,1611089,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,217221,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,2617991,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,6517705,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,5355357,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,3327978,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,5577105,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,1914315,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,4134866,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,4222084,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,1799290,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,2741517,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,8669914,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,6858932,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,290230,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,2148850,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,8644519,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,6118972,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,4748493,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,7595051,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,7574072,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,6949725,33946,614677375609919,11-02-2018 00:00:00,GENUINE
348702330256514,000037495066290,3290936,33946,614677375609919,11-02-2018 00:00:00,GENUINE
```

hadoop fs -cat /user/root/card_transaction/
card_transactions.csv | wc -l

```
[root@ip-172-31-21-211 ~]# hadoop fs -cat /user/root/card_transaction/card_transactions.csv | wc -l
53293
[root@ip-172-31-21-211 ~]#
```

- Creating external table to store data from csv file

CREATE EXTERNAL TABLE IF NOT EXISTS
CARD_TRANSACTION_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 '/user/root/card_transaction' TBLPROPERTIES
(“skip.header.line.count"="1");

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS CARD_TRANSACTION_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 '/user/root/card_transaction'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.052 seconds
```

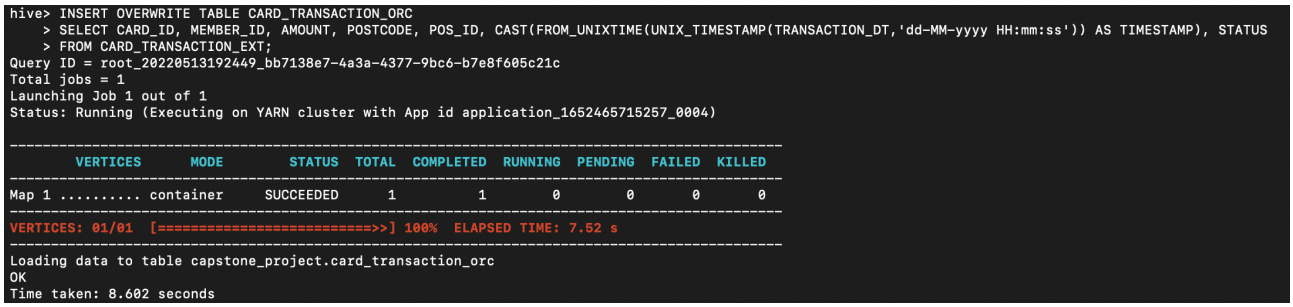
- **Creating ORC format table for better performance**

CREATE TABLE IF NOT EXISTS
CARD_TRANSACTION_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");

```
hive> CREATE TABLE IF NOT EXISTS CARD_TRANSACTION_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");
OK
Time taken: 0.102 seconds
```

- **Inserting data into ORC table**

```
INSERT OVERWRITE TABLE CARD_TRANSACTION_ORC
SELECT CARD_ID, MEMBER_ID, AMOUNT, POSTCODE,
POS_ID,
CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(TRANSACTIO
N_DT,'dd-MM-yyyy HH:mm:ss')) AS TIMESTAMP), STATUS
FROM CARD_TRANSACTION_EXT;
```



- **Creating Hive-HBase integrated table which is available in base as well**

```
CREATE TABLE
CARD_TRANSACTION_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_transaction_hive");

```
hive> CREATE TABLE CARD_TRANSACTION_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:p
os_id, card_transactions_family:transaction_dt, card_transactions_family:status")
    > TBLPROPERTIES ("hbase.table.name"="card_transaction_hive");
OK
Time taken: 3.036 seconds
```

- **Inserting data in table which is visible in base and using random UUID for rowkey**

INSERT OVERWRITE TABLE
CARD_TRANSACTION_HBASE
SELECT
reflect('java.util.UUID', 'randomUUID') as TRANSACTION_ID,
CARD_ID, MEMBER_ID, AMOUNT, POSTCODE, POS_ID,
TRANSACTION_DT, STATUS FROM
CARD_TRANSACTION_ORC;

```
hive> INSERT OVERWRITE TABLE CARD_TRANSACTION_HBASE
    > SELECT
    > reflect('java.util.UUID', 'randomUUID') as TRANSACTION_ID, CARD_ID, MEMBER_ID, AMOUNT, POSTCODE, POS_ID, TRANSACTION_DT, STATUS
    > FROM CARD_TRANSACTION_ORC;
Query ID = root_20220513192642_04f0d74a-214d-4f9d-aa42-a4bd3ff48876
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 11.37 s
----------------------------------------------------------------------------------------------
OK
Time taken: 15.047 seconds
```

Command to list the table in which the data is loaded and the command to get the count of the rows of the table

- **Listing the tables** show tables;

- **Printing the count of rows in table** select count(*) from card_transaction_orc;

- **Printing top 10 transaction date from card transaction table**
select year(transaction_dt), transaction_dt from card_transaction_orc limit 10; show tables;

```
hive> select year(transaction_dt), transaction_dt from card_transaction_orc limit 10;
OK
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
2018    2018-02-11 00:00:00
Time taken: 0.125 seconds, Fetched: 10 row(s)
```

- **Printing top 10 rows from orc table** select * from card_transaction_orc limit 10;

```
hive> select * from card_transaction_orc limit 10;
OK
348702330256514 000037495066290 9084849.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 330148.0        33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 136052.0        33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 4310362.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 9097094.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 2291118.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 4900011.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 633447.0        33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 6259303.0       33946   614677375609919 2018-02-11 00:00:00     GENUINE
348702330256514 000037495066290 369067.0        33946   614677375609919 2018-02-11 00:00:00     GENUINE
Time taken: 0.112 seconds, Fetched: 10 row(s)
```

- **Printing top 10 rows from card transaction table** select * from card_transaction_hbase limit 10;

```
hive> select * from card_transaction_hbase limit 10;
OK
000334c4-1c82-440a-afa6-9f328527772f    6489878454988664        297268311002579 19993.0 81152   071289879033526 2017-12-21 12:17:10     GENUINE
00045bcf-91f6-46c0-835a-414986e05035    372336918813721 064090191685558 1321570.0       17040   332660973087909 2016-03-31 06:03:05     GENUINE
00048586-82a9-4dd1-90d0-85920aedc001    5594912773065625        554119943804330 2099685.0       30442   206609815762524 2017-06-07 00:10:15     GENUINE
000611c8-1baf-46e5-9d75-98d527c9bd19    4256656637307408        987837606639725 5432565.0       88046   633001087829456 2018-01-12 23:15:18     GENUINE
000657d8-c448-4954-880e-41f10b732a6a    6463116552169683        306051887072370 1860859.0       71423   770390341796564 2017-01-10 00:35:44     GENUINE
00066737-511e-4449-adfa-5f6e41807e2e    372367382019172 483278526851211 78395.0 15033   574843866100658 2017-02-11 06:48:52     GENUINE
00069629-4bed-410a-9833-c21d26e90b85    5299801391216322        153755706358463 9017283.0       82520   542929559262350 2017-09-02 18:53:19     GENUINE
0006fb2a-5756-4c9a-93a8-bcc766497e34    4403085735771756        692507849787540 4211434.0       87539   197948297299071 2017-10-18 15:38:56     GENUINE
00076185-d4cd-48dc-8280-119de812f146    4373464339970856        353188612655228 1136988.0       50457   657196939516344 2018-01-11 00:00:00     GENUINE
000802f5-ca66-4e81-9dc6-fcf574566d39    4033232792979330        350825724718484 6221694.0       74745   369247284622271 2017-12-11 00:00:00     GENUINE
Time taken: 0.246 seconds, Fetched: 10 row(s)
```

————————————————HBASE ————————————————-

- Setting up the HBase environment

yum install gcc
sudo yum install python3-devel pip install happybase
hbase thrift start
jps

- Describing card transaction table in HBase

describe ‘card_transaction_hive'

```
hbase(main):001:0> describe 'card_transaction_hive'
Table card_transaction_hive is ENABLED
card_transaction_hive
COLUMN FAMILIES DESCRIPTION
{NAME => 'card_transactions_family', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', M
IN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.3600 seconds
```

- Counting the rows in the table

count 'card_transaction_hive'

```
hbase(main):002:0> count 'card_transaction_hive'
Current count: 1000, row: 04e6bc22-1dba-4a13-beb8-fb3ccd515b43
Current count: 2000, row: 09c5d844-94a3-44ea-8538-5ec7d4df62f5
Current count: 3000, row: 0e5830aa-9381-4466-8c8b-a4ac2fb1b3cf
Current count: 4000, row: 12efb0ee-7c20-450d-abc7-481d772de159
Current count: 5000, row: 17de7c2d-c317-420d-afba-69248d204bb9
Current count: 6000, row: 1cca0c53-db27-4f28-a7e8-9c32bc177be0
Current count: 7000, row: 21d68c38-5c1c-44be-818d-8c67c06b1ee7
Current count: 8000, row: 268b387e-55ee-4fba-88d3-9c1dd8081aa9
Current count: 9000, row: 2b49d905-3bc3-40b4-bedb-4aa4799e8ce4
Current count: 10000, row: 2fee5cee-a40f-435f-bea4-f307af640028
Current count: 11000, row: 34a83398-15d5-4fa4-a1b6-a95210c8aef8
Current count: 12000, row: 399f3089-1e46-4775-aaf9-9ca2d1b29a6f
Current count: 13000, row: 3e93ebbc-8b16-4533-89e4-0fecb0f4b125
Current count: 14000, row: 43633963-82d3-41cc-b596-fa28a0731af6
Current count: 15000, row: 47eae52e-873b-4ecf-93f5-252a18982fdb
Current count: 16000, row: 4cf3fcae-f938-4345-8338-f7a9e801ec44
Current count: 17000, row: 51bd98ea-873e-43ce-9fa7-7632a0dc34ab
Current count: 18000, row: 569c89a1-1794-450b-8420-6f715d20bc6b
Current count: 19000, row: 5b6f58f1-6f37-458b-b3dc-a6c7d764f37f
Current count: 20000, row: 6031295a-5c92-43f9-bf87-e6806a363656
Current count: 21000, row: 64d75e08-aa30-4a76-a5fa-b35fe6d8a80c
Current count: 22000, row: 69b56667-bd6e-4218-83ef-2a3ac2d6e213
Current count: 23000, row: 6e36d1d7-df1d-433d-b047-1e6e5d9e516d
Current count: 24000, row: 7309bfb8-efed-4c47-a29e-e0c9af24c9d7
Current count: 25000, row: 77fa73a0-eee0-477c-b27c-7ddca90199b0
Current count: 26000, row: 7cd34e4c-4c44-4291-a5cd-e5bf0e9e89ff
Current count: 27000, row: 81a38878-73d4-462e-aa95-82afa44663cc
Current count: 28000, row: 8653c0b2-b93d-4d21-8bfc-230e9697249b
Current count: 29000, row: 8af896ac-24dc-4124-8898-aaa81b0182ca
Current count: 30000, row: 8fbbbdd4-760f-4dcb-82ef-d43c6aebeac9
Current count: 31000, row: 943ed5ba-4244-4f71-aee1-a83d9450ad90
Current count: 32000, row: 990723c2-113e-44a4-a3ba-cac96b3fb02c
Current count: 33000, row: 9e194f05-e326-417d-8256-a1f4e14f43bf
Current count: 34000, row: a30fbdcd-17c6-4c01-ba00-4d464e96c51f
Current count: 35000, row: a7b9530a-f709-49e1-a97f-5da6adfe6c9d
Current count: 36000, row: acc51081-96bd-48f8-9eac-7f5d272c1ec3
Current count: 37000, row: b1916c00-eb1a-4f61-ba6c-92f2e1170fe5
Current count: 38000, row: b654e506-a198-4782-bac1-4caa41b38aed
Current count: 39000, row: bb0754af-edf5-4ef9-8c63-ae6632254a1d
Current count: 40000, row: bf9aaf4a-ed2f-4ae5-8372-d46e4d714f41
Current count: 41000, row: c47446d5-cae6-4a83-879f-422fa1cd71fb
Current count: 42000, row: c9049654-1158-4b99-9c28-686792885873
Current count: 43000, row: ce07a5a3-c1af-4ce3-851e-89aeba605ef3
Current count: 44000, row: d2baff0b-d922-4295-976f-a68017d3151c
Current count: 45000, row: d7bf938c-f4f8-4dc7-86c5-74483c6fd0b1
Current count: 46000, row: dcac0768-50e4-481c-9d4c-bca90c2a88c0
Current count: 47000, row: e15c6d4c-46ca-4d35-88a8-349eaf36271f
Current count: 48000, row: e6a83154-7e0d-4622-84e4-ef0b2e3ff93e
Current count: 49000, row: eb57a5e0-4cab-4db6-9848-a718e371eb8f
Current count: 50000, row: f03d1512-cfeb-4325-a889-2cbb7e742edd
Current count: 51000, row: f50dd392-75e7-47e4-852d-dd6308eece7d
Current count: 52000, row: fa05c95c-1d95-420a-9946-f6300fb7a9ea
Current count: 53000, row: febb31c5-515f-489b-8284-026ea275854a
53292 row(s) in 4.1750 seconds
```

Data Ingestion from the RDS to HDFS using Sqoop

Sqoop command used for importing table from RDS to HDFS

————————————————-SQOOP
———————————————— - **Setting up the environment for sqoop**

sudo -i
wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector- java-8.0.25.tar.gz
tar -xvf mysql-connector-java-8.0.25.tar.gz
cd mysql-connector-java-8.0.25/
sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
cd

Command to see the list of imported data in HDFS

- **Importing data from RDS to hdfs using swoop for card member** sqoop import \

--connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \ --table card_member \
--username upgraduser --password upgraduser \ --null-string '\\N' --null-non-string '\\N' \

--target-dir /user/root/card_member \ -m 1 --as-textfile

```
[root@ip-172-31-21-211 ~]# sqoop import \
> --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --table card_member \
> --username upgraduser --password upgraduser \
> --null-string '\\N' --null-non-string '\\N' \
> --target-dir /user/root/card_member \
> -m 1 --as-textfile
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/05/13 19:16:35 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/05/13 19:16:35 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/05/13 19:16:35 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/05/13 19:16:35 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class i
s generally unnecessary.
22/05/13 19:16:36 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
22/05/13 19:16:36 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
22/05/13 19:16:36 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/651ea94beaa0a7b06bf04ccfc0f6935d/card_member.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/05/13 19:16:39 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/651ea94beaa0a7b06bf04ccfc0f6935d/card_member.jar
22/05/13 19:16:39 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/05/13 19:16:39 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/05/13 19:16:39 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/05/13 19:16:39 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/05/13 19:16:39 INFO mapreduce.ImportJobBase: Beginning import of card_member
22/05/13 19:16:39 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/13 19:16:40 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/13 19:16:40 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-21-211.ec2.internal/172.31.21.211:8032
22/05/13 19:16:44 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/13 19:16:44 INFO mapreduce.JobSubmitter: number of splits:1
22/05/13 19:16:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652465715257_0002
22/05/13 19:16:45 INFO impl.YarnClientImpl: Submitted application application_1652465715257_0002
22/05/13 19:16:45 INFO mapreduce.Job: The url to track the job: http://ip-172-31-21-211.ec2.internal:20888/proxy/application_1652465715257_0002/
22/05/13 19:16:45 INFO mapreduce.Job: Running job: job_1652465715257_0002
22/05/13 19:16:55 INFO mapreduce.Job: Job job_1652465715257_0002 running in uber mode : false
22/05/13 19:16:55 INFO mapreduce.Job:  map 0% reduce 0%
22/05/13 19:17:03 INFO mapreduce.Job:  map 100% reduce 0%
22/05/13 19:17:04 INFO mapreduce.Job: Job job_1652465715257_0002 completed successfully
22/05/13 19:17:04 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189492
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

- **Importing data from RDS to hdfs using sqoop for member score** sqoop import \
--connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \

--table member_score \
--username upgraduser --password upgraduser \ --null-string '\\N' --null-non-string '\\N' \ --target-dir /user/root/member_score \
-m 1 --as-textfile

```
[root@ip-172-31-21-211 ~]# sqoop import \
> --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --table member_score \
> --username upgraduser --password upgraduser \
> --null-string '\\N' --null-non-string '\\N' \
> --target-dir /user/root/member_score \
> -m 1 --as-textfile
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/05/13 19:17:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/05/13 19:17:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/05/13 19:17:38 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/05/13 19:17:38 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class i
s generally unnecessary.
22/05/13 19:17:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
22/05/13 19:17:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
22/05/13 19:17:39 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/0617866c7b6995aa5f5b1040c02378c5/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/05/13 19:17:42 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/0617866c7b6995aa5f5b1040c02378c5/member_score.jar
22/05/13 19:17:42 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/05/13 19:17:42 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/05/13 19:17:42 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/05/13 19:17:42 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/05/13 19:17:42 INFO mapreduce.ImportJobBase: Beginning import of member_score
22/05/13 19:17:43 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/05/13 19:17:44 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/05/13 19:17:44 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-21-211.ec2.internal/172.31.21.211:8032
22/05/13 19:17:48 INFO db.DBInputFormat: Using read commited transaction isolation
22/05/13 19:17:48 INFO mapreduce.JobSubmitter: number of splits:1
22/05/13 19:17:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1652465715257_0003
22/05/13 19:17:49 INFO impl.YarnClientImpl: Submitted application application_1652465715257_0003
22/05/13 19:17:49 INFO mapreduce.Job: The url to track the job: http://ip-172-31-21-211.ec2.internal:20888/proxy/application_1652465715257_0003/
22/05/13 19:17:49 INFO mapreduce.Job: Running job: job_1652465715257_0003
22/05/13 19:17:58 INFO mapreduce.Job: Job job_1652465715257_0003 running in uber mode : false
22/05/13 19:17:58 INFO mapreduce.Job:  map 0% reduce 0%
22/05/13 19:18:04 INFO mapreduce.Job:  map 100% reduce 0%
22/05/13 19:18:04 INFO mapreduce.Job: Job job_1652465715257_0003 completed successfully
22/05/13 19:18:04 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189440
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

- **Checking whether importing is successful or not fro card_member?** hadoop fs -ls /user/root/card_member

- **Checking whether importing is successful or not for member_score?**

hadoop fs -ls /user/root/member_score

```
[root@ip-172-31-21-211 ~]# hadoop fs -ls /user/root/card_member
Found 2 items
-rw-r--r--   1 root hadoop          0 2022-05-13 19:17 /user/root/card_member/_SUCCESS
-rw-r--r--   1 root hadoop      85081 2022-05-13 19:17 /user/root/card_member/part-m-00000
[root@ip-172-31-21-211 ~]#
```

```
[root@ip-172-31-21-211 ~]# hadoop fs -ls /user/root/member_score
Found 2 items
-rw-r--r--   1 root hadoop          0 2022-05-13 19:18 /user/root/member_score/_SUCCESS
-rw-r--r--   1 root hadoop      19980 2022-05-13 19:18 /user/root/member_score/part-m-00000
[root@ip-172-31-21-211 ~]#
```

- **Count the number of rows for card member**
hadoop fs -cat /user/root/card_member/part-m-00000 | wc -l

- **Count the number of rows for member score**
hadoop fs -cat /user/root/member_score/part-m-00000 | wc -l

——————————————————HIVE ——————————————————

use capstone_project;

- **Creating table in hive for card member** 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 '/user/root/ card_member';

```
[root@ip-172-31-21-211 ~]# hadoop fs -cat /user/root/card_member/part-m-00000 | wc -l
999
[root@ip-172-31-21-211 ~]#
```

```
[root@ip-172-31-21-211 ~]# hadoop fs -cat /user/root/member_score/part-m-00000 | wc -l
999
[root@ip-172-31-21-211 ~]#
```

```
hive> 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 '/user/root/card_member';
OK
Time taken: 0.471 seconds
```

- **Creating table in hive for member score** CREATE EXTERNAL TABLE IF NOT EXISTS MEMBER_SCORE_EXT(`MEMBER_ID` STRING, `SCORE` INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION '/user/root/ member_score';

- **Creating table in hive for card member orc** 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");

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS MEMBER_SCORE_EXT( `MEMBER_ID` STRING,
    > `SCORE` INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/root/member_score';
OK
Time taken: 0.053 seconds
```

```
hive> 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");
OK
Time taken: 0.331 seconds
```

- **Creating table in hive for member score orc** CREATE TABLE IF NOT EXISTS MEMBER_SCORE_ORC(`MEMBER_ID` STRING, `SCORE` INT)
STORED AS ORC
TBLPROPERTIES (“orc.compress"="SNAPPY");

```
hive> CREATE TABLE IF NOT EXISTS MEMBER_SCORE_ORC(
    > `MEMBER_ID` STRING,
    > `SCORE` INT)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.067 seconds
```

- **Inserting 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;

```
hive> INSERT OVERWRITE TABLE CARD_MEMBER_ORC
    > SELECT CARD_ID, MEMBER_ID, MEMBER_JOINING_DT, CARD_PURCHASE_DT, COUNTRY, CITY FROM CARD_MEMBER_EXT;
Query ID = root_20220513192104_2dcf48de-a67a-43ac-8d06-22e4a6689275
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.20 s
--------------------------------------------------------------------------------
Loading data to table capstone_project.card_member_orc
OK
Time taken: 12.335 seconds
```

- **Inserting into member score orc**

INSERT OVERWRITE TABLE MEMBER_SCORE_ORC SELECT MEMBER_ID, SCORE FROM MEMBER_SCORE_EXT;

```
hive> INSERT OVERWRITE TABLE MEMBER_SCORE_ORC
    > SELECT MEMBER_ID, SCORE FROM MEMBER_SCORE_EXT;
Query ID = root_20220513192121_24fffe79-492b-4c91-ad1f-194d82b33cb5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 0.56 s
--------------------------------------------------------------------------------
Loading data to table capstone_project.member_score_orc
OK
Time taken: 1.945 seconds
```

- **Printing top 10 rows from card member orc** SELECT * FROM CARD_MEMBER_ORC LIMIT 10;

```
hive> SELECT * FROM CARD_MEMBER_ORC LIMIT 10;
OK
340028465709212 009250698176266 2012-02-08 06:04:13     05/13   United States   Barberton
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   Graham
340134186926007 887711945571282 2012-02-05 01:21:58     02/13   United States   Dix Hills
340265728490548 680324265406190 2014-03-29 07:49:14     11/14   United States   Rancho Cucamonga
340268219434811 929799084911715 2012-07-08 02:46:08     08/12   United States   San Francisco
340379737226464 089615510858348 2010-03-10 00:06:42     09/10   United States   Clinton
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   Beaverton
340889618969736 459292914761635 2013-04-23 08:40:11     11/15   United States   West Palm Beach
Time taken: 0.16 seconds, Fetched: 10 row(s)
```

- **Printing top 10 rows from member score orc** SELECT * FROM MEMBER_SCORE_ORC LIMIT 10;

```
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: 0.133 seconds, Fetched: 10 row(s)
```

- **Counting rows from card member orc**
SELECT count(*) FROM CARD_MEMBER_ORC LIMIT 10;

```
hive> SELECT count(*) FROM CARD_MEMBER_ORC LIMIT 10;
Query ID = root_20220513192319_93f614b8-eb0b-480d-9694-fdb010d3f29b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.76 s
--------------------------------------------------------------------------------
OK
999
Time taken: 1.366 seconds, Fetched: 1 row(s)
```

- **Counting rows from member score orc**
SELECT count(*) FROM MEMBER_SCORE_ORC LIMIT 10;

Creating Lookup Table

Command to create the Lookup Table

————————————————-HIVE ————————————————

- **Creating Hive-base integrated lookup table**

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_family:postcode, lookup_transaction_family:transaction_dt") TBLPROPERTIES ("hbase.table.name" = "lookup_data_hive");

Command to see the table created

```
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_family:postcode, lookup_transaction_family:transaction_dt")
    > TBLPROPERTIES ("hbase.table.name" = "lookup_data_hive");
OK
Time taken: 2.39 seconds
```

————————————————HBASE ————————————————

- **Describing lookup table in HBase**

describe ‘lookup_data_hive’

- **Altering lookup table**

alter 'lookup_data_hive', {NAME => 'lookup_transaction_family', VERSIONS => 10}
describe 'lookup_data_hive'

```
hbase(main):003:0> 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_VER
SIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '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 => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
2 row(s) in 0.0240 seconds
```

```
hbase(main):004:0> alter 'lookup_data_hive', {NAME => 'lookup_transaction_family', VERSIONS => 10}
Updating all regions with the new schema...
1/1 regions updated.
Done.
0 row(s) in 1.9100 seconds
```

Loading the Lookup Table

Commands to load the relevant data in the Lookup Table ——————————————HIVE ——————————————-

- **Creating ranked card transaction table to store last 10 transaction** 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");

```
hive> 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");
OK
Time taken: 0.807 seconds
```

- Creating table for UCL

CREATE TABLE IF NOT EXISTS
CARD_UCL_ORC(`CARD_ID` STRING,
`UCL` DOUBLE)
STORED AS ORC

TBLPROPERTIES (“orc.compress"="SNAPPY");

```
hive> CREATE TABLE IF NOT EXISTS CARD_UCL_ORC(
    > `CARD_ID` STRING,
    > `UCL` DOUBLE)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.068 seconds
```

- Inserting last 10 data into ranked card transaction

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_TRANSACTION_HBASE
WHERE
STATUS = 'GENUINE') A) B WHERE B.RANK <= 10;

```
hive> 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_TRANSACTION_HBASE WHERE
    > STATUS = 'GENUINE') A ) B WHERE B.RANK <= 10;
Query ID = root_20220513193100_0b90689b-bff7-4203-81e7-14dc9170b79c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 18.75 s
--------------------------------------------------------------------------------
Loading data to table capstone_project.ranked_card_transactions_orc
OK
Time taken: 25.425 seconds
```

- **Inserting data into UCL**

```
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;
```

```
hive> 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;
Query ID = root_20220513193127_e9034fa4-4877-4c7d-99a9-90acccd2b472
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 1.87 s
----------------------------------------------------------------------------------------------
Loading data to table capstone_project.card_ucl_orc
OK
Time taken: 3.204 seconds
```

- **Inserting data into lookup 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 CMS
ON RCTO.CARD_ID = CMS.CARD_ID
WHERE RCTO.RANK = 1;
```

```
hive> 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 CMS
    > ON RCTO.CARD_ID = CMS.CARD_ID
    > WHERE RCTO.RANK = 1;
No Stats for capstone_project@ranked_card_transactions_orc, Columns: postcode, rank, transaction_dt, card_id
No Stats for capstone_project@card_ucl_orc, Columns: card_id, ucl
No Stats for capstone_project@card_member_orc, Columns: member_id, card_id
No Stats for capstone_project@member_score_orc, Columns: member_id, score
Query ID = root_20220513193236_913ab00a-12d6-4a3a-aa80-4d1a4264998e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 14.40 s
--------------------------------------------------------------------------------
OK
Time taken: 21.529 seconds
```

Command to see the table created and it's content

- Counting rows in lookup table

select count(*) from lookup_data_hbase;

```
hive> select count(*) from lookup_data_hbase;
Query ID = root_20220513193343_a2db2ecd-3433-4898-adb5-291a0f7eccdc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1652465715257_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.20 s
--------------------------------------------------------------------------------
OK
999
```

- **Printing top 10 rows of lookup table** select * from lookup_data_hbase limit 10;

```
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 1.528568530791473E7     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: 0.291 seconds, Fetched: 10 row(s)
hive>
```

————————————————————HBASE ————————————————————— - Counting rows in lookup table in HBASE

count 'lookup_data_hive'

```
hbase(main):001:0> count 'lookup_data_hive'
999 row(s) in 0.6450 seconds

=> 999
hbase(main):002:0>
```

- **Printing top 10 rows of lookup table in HBASE** scan ‘lookup_data_hive'

```
hbase(main):002:0> scan 'lookup_data_hive'
ROW                                    COLUMN+CELL
 340028465709212                       column=lookup_card_family:score, timestamp=1652470377779, value=233
 340028465709212                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.6331555548882348E7
 340028465709212                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=24658
 340028465709212                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-02 03:25:35
 340054675199675                       column=lookup_card_family:score, timestamp=1652470377779, value=631
 340054675199675                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.4156079786189131E7
 340054675199675                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=50140
 340054675199675                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-15 19:43:23
 340082915339645                       column=lookup_card_family:score, timestamp=1652470377779, value=407
 340082915339645                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.5285685330791473E7
 340082915339645                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=17844
 340082915339645                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-26 19:03:47
 340134186926007                       column=lookup_card_family:score, timestamp=1652470377779, value=614
 340134186926007                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.5239767522438556E7
 340134186926007                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=67576
 340134186926007                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-18 23:12:50
 340265728490548                       column=lookup_card_family:score, timestamp=1652470377779, value=202
 340265728490548                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.608491671255562E7
 340265728490548                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=72435
 340265728490548                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-21 02:07:35
 340268219434811                       column=lookup_card_family:score, timestamp=1652470377779, value=415
 340268219434811                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.2507323937605347E7
 340268219434811                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=62513
 340268219434811                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-16 04:30:05
 340379737226464                       column=lookup_card_family:score, timestamp=1652470377779, value=229
 340379737226464                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.4198310998368107E7
 340379737226464                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=26656
 340379737226464                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-27 00:19:47
 340383645652108                       column=lookup_card_family:score, timestamp=1652470377779, value=645
 340383645652108                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.4091750460468251E7
 340383645652108                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=34734
 340383645652108                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-29 01:29:12
 340803866934451                       column=lookup_card_family:score, timestamp=1652470377779, value=502
 340803866934451                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.0843341196185412E7
 340803866934451                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=87525
 340803866934451                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-31 04:23:57
 340889618969736                       column=lookup_card_family:score, timestamp=1652470377779, value=330
 340889618969736                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.3217942365515321E7
 340889618969736                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=61341
 340889618969736                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-31 21:57:18
 340924125838453                       column=lookup_card_family:score, timestamp=1652470377779, value=644
 340924125838453                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.1712496575647112E7
 340924125838453                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=46392
 340924125838453                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2017-10-12 15:41:23
 341005627432127                       column=lookup_card_family:score, timestamp=1652470377779, value=629
 341005627432127                       column=lookup_card_family:ucl, timestamp=1652470377779, value=1.2843150457220368E7
 341005627432127                       column=lookup_transaction_family:postcode, timestamp=1652470377779, value=23306
 341005627432127                       column=lookup_transaction_family:transaction_dt, timestamp=1652470377779, value=2018-01-22 14:59:38
```