Loading Historical Transactions Data into NoSQL Database

Commands to load the past transactions data into NoSQL database

1. Setup emr cluster in AWS.

2. Login to emr cluster using putty client with hadoop user

3. Execute below commands to create required directories in ec2 instance.

   sudo su –
   su – hdfs
   hadoop fs -mkdir /capstone_project
   hadoop fs -chown hadoop: hadoop /capstone_project
   hadoop fs -mkdir /capstone_project/card_transactions

4. Download card_transactions.csv upgrad resource section.

5. Copy file from local system to EC2 instance using filezilla.

6. Copy file from EC2 local filesystem to hadoop file system

   Copy card_transaction.csv hadoop fs -put
   card_transactions.csv/capstone_project/card_transactions/.

7. Start hive from command prompt. Create new database namely capstone_project and switch to use capstone_project.

   hive
   create database capstone_project
   use capstone_project

```
-bash-4.2$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
    >
    >
    >
    >
    > create database capstone_project;
OK
```

8. Set below parameters for the hive session

```
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;
hive> 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;
```

9. Create external table card_transactions_ext table which will point to HDFS location created earlier.

```
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");
```

```
hive>
    > 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");
OK
Time taken: 0.441 seconds
```

10. Create table card_transactions_orc. ORC format will help in 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");

```
    >
    > 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");
OK
Time taken: 0.408 seconds
```

11. Load data in card_transactions_orc while casting timestamp format for transaction_dt column.

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;

```
hive>
    > 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;
Query ID = hdfs_20230518065643_3fdbdfb5-2d1a-4164-9a3d-bb08d432cfa9
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_1684390665990_0003)

--------------------------------------------------------------------------------
        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: 6.12 s
--------------------------------------------------------------------------------
Loading data to table capstone_project.card_transactions_orc
OK
Time taken: 16.33 seconds
```

12. Verify transaction_dt and year in card_transactions_orc table.

select year(transaction_dt), transaction_dt from card_transactions_orc limit 10;

```
hive>
    > select year(transaction_dt), transaction_dt from card_transactions_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: 0.236 seconds, Fetched: 10 row(s)
hive>
```

13. Create card_transactions_hbase hive-hbase integrated table which will be visible in HBase as well. - **This table will have all transactions – historical as well as new incoming from streaming layer.**

```
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");
```

```
hive> 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_transa
ctions_family:amount, card_transactions_family:postcode, card_transactions_family:pos_id, card_transactions_family:tran
saction_dt, card_transactions_family:status")
    > TBLPROPERTIES ("hbase.table.name"="card_transactions_hive");
OK
Time taken: 2.645 seconds
```

14. Load data in card_transactions_hbase which will be visible in HBase as well with name as card_transactions_hive. Use randomUUID to populate TRANSACTION_ID field which will become row key in HBase effectively.

```
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;

```
hive> 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;
Query ID = hdfs_20230518094547_e328dfed-48a8-4c16-8dbf-67b94f22f298
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_1684390665990
_0005)

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: 1/1
OK
Time taken: 24.903 seconds
```

15. Check some data in card_transactions_hbase.

select * from card_transactions_hbase limit 10;

```
hive> select * from card_transactions_hbase limit 10;
OK
0000e47c-07b6-46fd-892e-54f2eb4a65cc    4863127030291206        7883348231400969
180153.0        12495   572821677272050 2017-10-18 19:09:36     GENUINE
000360f1-a497-48a8-b471-feb27897ab41    4115056128951163        8405266842701411
590603.0        76566   167237778761295 2017-09-15 22:53:48     GENUINE
0003c2a4-7bb0-4643-b474-02cf06c437e2    341920010925925 185524819205415 4237138.
0       65734   935721714228987 2017-03-31 11:45:13     GENUINE
0003e302-418c-4883-8776-4d74123dc865    345969858843082 035217612427824 4497071.
0       14745   457146580055026 2018-01-06 03:59:03     GENUINE
00056a09-f965-48fa-93e7-ad0aa72a2c72    5137733204831953        0499338382710826
993746.0        58835   282691592931003 2017-02-19 09:41:09     GENUINE
00071039-a80f-4bbf-ac2a-bf41d95dd49e    5397412643360495        9448781983330798
528396.0        80520   770813269233619 2017-06-20 14:42:43     GENUINE
000790d4-365e-4985-a131-f30ecd7f2c90    347816334672492 314841462077900 2098057.
0       10993   693115489753967 2016-02-06 20:27:52     GENUINE
00082e95-5595-4495-a5cd-c98a83a70a14    5342400571435088        0087322675886723
5304.0  14887   202751295246195 2017-03-01 18:58:50     GENUINE
000995ed-fa3c-42f4-aab0-221b2eccadc9    5140973081437202        5060194806436666
577155.0        49918   560182011727900 2017-01-05 18:46:15     GENUINE
00099ab5-76f4-45c6-868e-dd2fcd70dcff    4750699680073601        6772957250877343
398741.0        93204   981461330778216 2017-11-22 06:02:49     GENUINE
Time taken: 0.269 seconds, Fetched: 10 row(s)
hive>
```

16. Start HBase shell from command prompt. In HBase, check details of card_transactions_hive hive-hbase integrated table.

```
[root@ip-172-31-92-226 ~]# hbase shell
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 => 'tru
e', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.3950 seconds
```

17. In HBase, check count in card_transactions_hive in HBase

count 'card_transactions_hive'

```
hbase(main):002:0> count 'card_transactions_hive'
Current count: 1000, row: 04d1f3b4-6a0f-4ce6-86d3-4e260481d1de
Current count: 2000, row: 099765ca-3222-4739-b554-8db61ef13a51
Current count: 3000, row: 0e932cba-409b-47f9-bb2c-a61e8891daad
Current count: 4000, row: 1331ade7-9799-4263-87cf-0b58f975e4ad
Current count: 5000, row: 18071f80-c43c-4d3f-a883-2dbda062f203
Current count: 6000, row: 1d09746a-937b-4258-969f-e7990a4c4122
Current count: 7000, row: 21c863aa-fcea-4cf4-adc0-3c8a8497e0ea
Current count: 8000, row: 265d3730-7161-4cbc-8b60-d7589be900f1
Current count: 9000, row: 2b2e8a6b-df9a-43ab-8f2d-5394a26f0583
Current count: 10000, row: 2ffd6027-1797-47d0-9ac2-090146e47604
Current count: 11000, row: 349e205f-e75c-426a-915a-801f1fe0fd1d
Current count: 12000, row: 3966a7e3-0be0-42d9-b6e0-02da867ab4d1
Current count: 13000, row: 3e387af9-7159-4f85-9c17-190d2a42e3b5
Current count: 14000, row: 42d8adeb-5dab-43f5-ae02-46aabe7d3616
Current count: 15000, row: 47d35c83-f072-4b5b-a118-e07e8c6e345b
Current count: 16000, row: 4c6c5981-8948-4125-97a0-366a1091ba28
Current count: 17000, row: 513d62fb-cd92-4483-889c-ea3db022834d
Current count: 18000, row: 5635a79e-72dc-4de2-8a09-b956a2b5e23c
Current count: 19000, row: 5adb2e30-8fc4-4f96-8c01-6eaaf347c21f
Current count: 20000, row: 5f76caa5-72bd-43bd-b3ba-f813580560e3
Current count: 21000, row: 64501917-cc13-45bd-9c60-f30eb6b2e526
Current count: 22000, row: 691f4aa0-1c2a-4ba8-af34-9120e1c724be
Current count: 23000, row: 6e0a21c4-f23c-41d0-b98c-7a86db37eeef
Current count: 24000, row: 7310ff96-ddd2-4ca8-9946-9db56bcd0c14
Current count: 25000, row: 77ddd7b9-c792-44e9-9155-8856aa419d51
Current count: 26000, row: 7cda3499-0382-4804-a1b5-029418153206
Current count: 27000, row: 818964b5-cc56-4990-ab00-edc58f3406e8
Current count: 28000, row: 86568c96-d150-4b78-a213-fbfcd3357bd2
Current count: 29000, row: 8b455c01-aff5-4c2d-acb7-64987f1b7cd5
Current count: 30000, row: 8ffe4acf-6712-4be5-8b42-eb164ea6eeb4
Current count: 31000, row: 94ee05c1-4fda-478e-bcab-9f823e4c46cf
Current count: 32000, row: 99c95dc6-5cfe-4005-a8e7-024628ab5b3a
Current count: 33000, row: 9e70dfe4-8722-45a0-855d-10800de96226
Current count: 34000, row: a2f4900d-a12a-441b-88ff-b84c5513df3a
Current count: 35000, row: a77ff886-964f-4bfc-87ae-82933a6ba0e6
Current count: 36000, row: ac88c76b-dc08-4b4f-84d9-e41e33370dce
Current count: 37000, row: b12a25e3-9608-4919-8318-94a43870709f
Current count: 38000, row: b5fdcd7b-87c0-492d-9b21-6527b72b3f8f
Current count: 39000, row: bab99808-7393-4d14-b32d-7da182f09c99
Current count: 40000, row: bf808783-abdb-4d51-8ddb-f2244489d0de
Current count: 41000, row: c45d17af-c641-4e6e-bc61-94b3ffda6816
Current count: 42000, row: c92ed8ab-35eb-47a4-8832-40c112b53177
Current count: 43000, row: ce27b524-6a9d-401b-ac09-9672d70702e6
Current count: 44000, row: d32f4a39-f997-4839-a6f1-81548fff5716
Current count: 45000, row: d81f5858-2a50-4a2f-b9ac-f792667242d6
Current count: 46000, row: dcfd9f3b-76e5-4792-aef2-c1c2e8a138c4
Current count: 47000, row: e1ff170e-6ed8-4aad-9015-a699c514652d
Current count: 48000, row: e6b6b5fd-612e-49cb-82b3-51e7089915ba
Current count: 49000, row: eb677868-b820-4cf4-b556-962f65ad3b38
Current count: 50000, row: f050703e-3bf8-4b4a-ab34-d4b5817e52da
Current count: 51000, row: f50cd6bc-9e18-466f-b405-ab8445661124
Current count: 52000, row: f9b5cfde-03f7-4bcd-afca-b4040e48cf3d
Current count: 53000, row: fe987fec-642c-4460-bac4-ce8d6c47bb0f
53292 row(s) in 3.4500 seconds

=> 53292
```