

Loading Historical Transactions Data into NoSQL Database

1. Starting hive session and creating a new database named ccfraud_capstone -> switch to ccfraud_capstone database

```
create database ccfraud_capstone;
use ccfraud_capstone;
```

```
hive> create database ccfraud_capstone;
OK
Time taken: 0.481 seconds
hive> use ccfraud_capstone;
OK
Time taken: 0.083 seconds
hive>
```

2. Setting parameters for the hive session

```
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.compress.SnappyCodec; 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;
```

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

3. Creating an external table CARD_TRANSACTIONS_EXT

```
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 '/user/hadoop/upgrad_ccfraud/card_transactions/' TBLPROPERTIES
("skip.header.line.count"="1");
```

```

root@ip-172-31-21-198:~
at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.mkdirs(NameNodeRpcServer.java:1079)
at org.apache.hadoop.hdfs.protocolPB.ClientNameNodeProtocolServerSideTranslatorPB.mkdirs(ClientNameNodeProtocolServerSideTranslatorPB.java:652)
at org.apache.hadoop.hdfs.protocol.proto.ClientNameNodeProtocolProtos$ClientNameNodeProtocol$2.callBlockingMethod(ClientNameNodeProtocolProtos.java)
at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:447)
at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:989)
at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:850)
at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:793)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)
at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2489)
)
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 '/user/hadoop/upgrad_ccfraud/card_transactions/' TBLPROPERTIES
> ("skip.header.line.count"="1");
OK
Time taken: 0.112 seconds
hive>

```

4. Creating table "CC_TRANSACTIONS_ORC" in ORC format for better performance.

CREATE TABLE IF NOT EXISTS CC_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");

```

root@ip-172-31-21-198:~
at org.apache.hadoop.hdfs.protocolPB.ClientNameNodeProtocolProtos.java)
at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:447)
at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:989)
at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:850)
at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:793)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)
at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2489)
)
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 '/user/hadoop/upgrad_ccfraud/card_transactions/' TBLPROPERTIES
> ("skip.header.line.count"="1");
OK
Time taken: 0.112 seconds
hive> CREATE TABLE IF NOT EXISTS CC_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: 0.386 seconds
hive>

```

5. Loading the data in "CC_TRANSACTIONS_ORC" table and type casting transaction_dt column to timestamp format

INSERT OVERWRITE TABLE CC_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;

```

root@ip-172-31-21-198:~
Time taken: 0.112 seconds
hive> CREATE TABLE IF NOT EXISTS CC_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: 0.386 seconds
hive> INSERT OVERWRITE TABLE CC_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 = root_20221219173434_18524e95-22bf-4a3d-9bd8-b57c0d0d6895
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_1671469134332_0002)

-----
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: 7.79 s
-----
Loading data to table ccfraud_capstone.cc_transactions_orc
OK
Time taken: 18.515 seconds
hive>
>

```

6. Verifying **transaction_dt** and **year** columns in " **CC_TRANSACTIONS_ORC** " table.

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

```

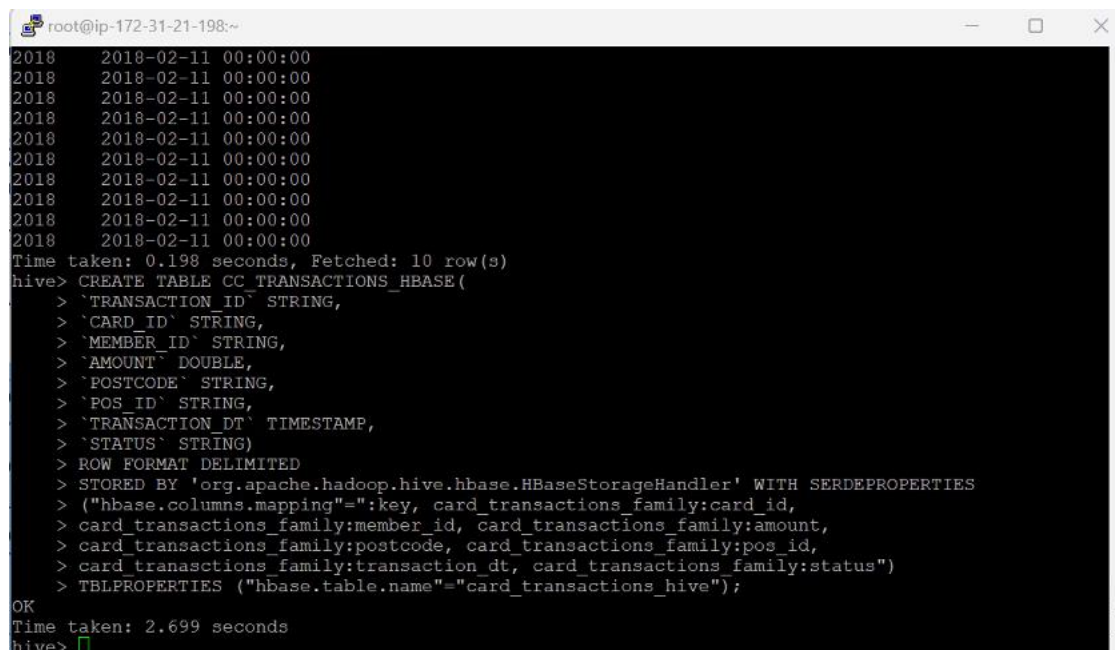
root@ip-172-31-21-198:~
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1671469134332_0002)

-----
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: 7.79 s
-----
Loading data to table ccfraud_capstone.cc_transactions_orc
OK
Time taken: 18.515 seconds
hive>
> select year(transaction_dt), transaction_dt from cc_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: 0.198 seconds, Fetched: 10 row(s)
hive>

```

7. Creating an integrated hive - hbase table that will be visible in HBase as well. " **CC_TRANSACTIONS_HBASE** " table

```
CREATE TABLE CC_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"="cc_transactions_hive");
```



```
root@ip-172-31-21-198:~
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: 0.198 seconds, Fetched: 10 row(s)
hive> CREATE TABLE CC_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");
OK
Time taken: 2.699 seconds
hive>
```

- Loading data in "CC_TRANSACTIONS_HBASE" table that will be visible in HBase as well with table name as "cc_transactions_hive". Using randomUUID to populate TRANSACTION_ID field (row key).

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



```

root@ip-172-31-21-198:~
  at java.lang.ClassLoader.loadClass(ClassLoader.java:351)
  at java.lang.Class.forName0(Native Method)
  at java.lang.Class.forName(Class.java:348)
  at org.apache.hadoop.hive.common.JavaUtils.loadClass(JavaUtils.java:59)
  at org.apache.hadoop.hive.common.JavaUtils.loadClass(JavaUtils.java:55)
  at org.apache.hadoop.hive.ql.udf.generic.GenericUDFReflect.evaluate(GenericUDFReflect.java:
106)
... 26 more
]], Vertex did not succeed due to OWN_TASK_FAILURE, failedTasks:1 killedTasks:0, Vertex vertex_1671
469134332_0002_3_00 [Map 1] killed/failed due to:OWN_TASK_FAILURE]DAG did not succeed due to VERTEX
_FAILURE. failedVertices:1 killedVertices:0
hive> INSERT OVERWRITE TABLE CC_TRANSACTIONS_HBASE SELECT
  > reflect('java.util.UUID', 'randomUUID') as TRANSACTION_ID, CARD_ID, MEMBER_ID, AMOUNT,
  > POSTCODE, POS_ID, TRANSACTION_DT, STATUS
  > FROM CC_TRANSACTIONS_ORC;
Query ID = root_20221219174451_401af4d4-d312-454b-85dc-8cb828792aba
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1671469134332_0002)

-----
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: 11.20 s
-----
OK
Time taken: 15.654 seconds
hive>

```

9. Verify data in "cc_transactions_hbase" table.

select * from cc_transactions_hbase limit 10;

```

root@ip-172-31-21-198:~
Map 1 ..... container      SUCCEEDED      1          1          0          0          0          0
-----
VERTICES: 01/01  [=====>>>] 100% ELAPSED TIME: 11.20 s
-----
OK
Time taken: 15.654 seconds
hive> select * from cc_transactions_hbase limit 10;
OK
0001cef0-4d5b-4d88-b0ce-0da380409a66 5284175522526975 413021843879928 1604268.0 140
20 214696179763291 2016-03-27 05:29:37 GENUINE
0005ec01-c98b-4d41-9d17-b2c0ff5e960 5271320386643574 991784667284668 1590198.0 926
25 225865830430921 2017-05-30 20:07:20 GENUINE
00070787-5f76-4fb6-bb79-45aed2e3a5fd 6460729612153589 266629518896098 4366995.0 975
39 893307976868097 2017-01-12 15:47:50 GENUINE
0007767e-56f7-407f-b2fe-b50306d83dfb 5175735819607028 227686059795799 2082575.0 520
54 215945504043719 2018-02-11 00:00:00 GENUINE
0007e0d3-3b77-4304-b2b6-cdd6e2333937 6223813071182434 736137659378265 1714630.0 913
56 748648404529559 2017-07-29 05:40:05 GENUINE
000916dc-4a82-4068-9062-02bb410fa5c4 5134334388575160 978465390240911 4731739.0 610
80 569335687828707 2016-05-27 03:54:25 GENUINE
000a5c44-d3f7-4650-9e03-55d0747a856b 4389973676463558 295554828848966 6305493.0 567
57 911478839996964 2017-04-17 05:47:08 GENUINE
000b85c8-16ff-4f9a-9d18-f13292fa4005 6011706374151856 199243620982420 9732433.0 538
12 478785382784557 2017-12-14 22:16:57 GENUINE
000baa11-7d7b-4682-94dc-5663885bc269 341722035429601 979218131207765 8325634.0 67741 747
682736002337 2017-02-06 05:40:42 GENUINE
000c16d5-d964-4e69-afdc-09bfd7a4cd2e 6011938409004772 577907767500023 1988126.0 388
20 814499910586436 2016-02-23 00:27:07 GENUINE
Time taken: 0.288 seconds, Fetched: 10 row(s)
hive>

```

10. Starting HBase session and verifying details of "cc_transactions_hive" table (hive-hbase integrated table).

describe 'cc_transactions_hive'

11. Verify count of "cc_transactions_hive" table

count 'cc_transactions_hive'

```
root@ip-172-31-21-198:~  
Current count: 28000, row: 87685a0a-05c3-4a42-a279-210baadf35bc  
Current count: 29000, row: 8c4e959d-c385-453f-8f62-e0d01217173e  
Current count: 30000, row: 90f8731b-3235-4f70-a5cf-3db1156bce4a  
Current count: 31000, row: 95a87f67-0b18-4846-8d67-3368421bcd6  
Current count: 32000, row: 9a3a394f-d232-4d29-b298-52b330bab082  
Current count: 33000, row: 9f16276a-f6ad-4e22-9850-e71f1dc869de  
Current count: 34000, row: a3d39690-2382-41b4-a306-f1f9a429c475  
Current count: 35000, row: a89aad1e-d0d8-4909-a035-1dec9cf09ac6  
Current count: 36000, row: ad239bcb-1dc7-459a-bb26-13c2ceb33d73  
Current count: 37000, row: b1e32188-98c3-4a1c-ae5c-eff3e121327b  
Current count: 38000, row: b6956bfff-e40d-4a07-a245-09887745840f  
Current count: 39000, row: bb3c6641-8dac-4063-8c6c-7ff3c75f98c5  
Current count: 40000, row: c01543a6-6f8c-42ed-a62d-9df2865e9fd6  
Current count: 41000, row: c4be3333-b3bb-422f-a371-085b99f74912  
Current count: 42000, row: c99b7318-3baa-421e-a976-4613d3452172  
Current count: 43000, row: ce61b3f8-9951-4c27-9163-a401fb6875b4  
Current count: 44000, row: d2f91fb9-e32d-44b9-8024-eeee031d71ad  
Current count: 45000, row: d8066d3d-527d-4f3f-b688-4ee6f40d7155  
Current count: 46000, row: dce52057-071e-4ad7-bab3-5bef2f0d247f  
Current count: 47000, row: elae4d3d-bd9c-41c3-9d27-e024a2545a04  
Current count: 48000, row: e6ab685e-7154-4335-99c6-84d021f89795  
Current count: 49000, row: eb793a0e-a874-4908-af1a-0bf507c12c91  
Current count: 50000, row: f059e689-6446-4c34-97e8-2949fe13bec3  
Current count: 51000, row: f50c5b6d-e444-4214-9780-c24131b08f51  
Current count: 52000, row: f9c8db4c-6bc0-4a24-8734-1a4edc19b3dd  
Current count: 53000, row: fe93bb44-8892-4d57-b9af-36050f0a83a3  
53292 row(s) in 3.8990 seconds  
=> 53292  
hbase(main):005:0>
```

Count of the "cc_transactions_hive" table is 53,292 which is matching with given requirement