

Scripts Execution

Screenshots of the execution of the scripts written

The screenshot shows the Amazon EMR console with the cluster 'CCFD_V4' selected. The 'Summary' tab is active. Key details include:

- Cluster info:** Cluster ID: j-37XC3U7XB0M4E, Cluster ARN: arn:aws:elasticmapreduce:us-east-1:586789:cluster/j-37XC3U7XB0M4E, Cluster configuration: Instance groups, Capacity: 1 Primary | 2 Core | 0 Task.
- Applications:** Amazon EMR version: emr-5.36.0, Installed applications: HBase 1.4.13, HCatalog 2.3.9, Hadoop 2.10.1, Hive 2.5.9, Spark 2.4.8, Sqoop 1.4.7, Tez 0.9.2, Zeppelin 0.10.0, ZooKeeper 3.4.14.
- Cluster management:** Log destination in Amazon S3: aws-logs-586789648747-us-east-1/elasticmapreduce, Persistent application UIs: Spark History Server, YARN Timeline Server, Tez UI.
- Status and time:** Status: Waiting, Creation time: September 20, 2025, 15:16 (UTC+05:30), Elapsed time: 13 minutes, 5 seconds.

1: Created EMR with Hadoop, Hive, HBase, HCatalog, Spark, Hue and Sqoop

```
$ scp -i ccfd_emr_key.pem D:/Janan/UpGrad_Projects/CreditCard_FraudTransactions/card_transactions.csv hadoop@ec2-13-221-171-199.compute-1.amazonaws.com:/home/hadoop/
card_transactions.csv
100% 4716KB 1.4MB/s 00:03
```

2: Transferring the card_transactions.csv to Hadoop

```
$ ssh -i ccfd_emr_key.pem hadoop@ec2-13-221-171-199.compute-1.amazonaws.com
Key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-13-221-171-199.compute-1.amazonaws.com' (ED25519) to the list of known hosts.

      #_
~\_ #####_      Amazon Linux 2
~~ \#####\
~~ \###|      AL2 End of Life is 2026-06-30.
~~ \#/ __
~~ \~' '-'>
~~   /      A newer version of Amazon Linux is available!
~~_. /_
~/ /_      Amazon Linux 2023, GA and supported until 2028-03-15.
~/m/'      https://aws.amazon.com/linux/amazon-linux-2023/

21 package(s) needed for security, out of 23 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEE MMMMMMM          MMMMMMM RRRRRRRRRRRRRR
E:::::::::::E M:::::M      M:::::M R:::::R:::::R
EE:::::EEEEEEEEE:::E M:::::M      M:::::M R:::::RRRRRR:::::R
E:::::E      EEEE M:::::M      M:::::M RR:::::R      R:::::R
E:::::E      M::::M:::M      M:::M::::M R:::::R      R:::::R
E:::::EEEEEEEEE M::::M M:::::M M:::::M R:::::RRRRRR:::::R
E:::::EEEEEEEEE M::::M M:::::M M:::::M R:::::RRRRRR:::::R
E:::::E      M::::M M::::M M:::::M R:::::R      R:::::R
E:::::E      EEEE M::::M      MMM  M:::::M R:::::R      R:::::R
EE:::::EEEEEEEEE:::E M::::M      M:::::M R:::::R      R:::::R
E:::::::::::E M::::M      M:::::M RR:::::R      R:::::R
EEEEEEEEEEEEEEEEEE MMMMMMM          MMMMMMM RRRRRRRR

[hadoop@ip-172-31-8-198 ~]$
```

3: Connected to the EMR Instance

```
[hadoop@ip-172-31-8-198 ~]$ ls  
card_transactions.csv  
[hadoop@ip-172-31-8-198 ~]$ █
```

4: Validation of the file present

```
[hadoop@ip-172-31-8-198 ~]$ hadoop fs -mkdir /user/CCFD_project  
[hadoop@ip-172-31-8-198 ~]$ hadoop fs -ls /user/  
Found 11 items  
drwxr-xr-x  - hadoop  hdfsadmingroup      0 2025-09-20 10:03 /user/CCFD_project  
drwxrwxrwx  - hadoop  hdfsadmingroup      0 2025-09-20 09:50 /user/hadoop  
drwxrwxr-x  - hbase   hbase              0 2025-09-20 09:51 /user/hbase  
drwxr-xr-x  - mapred  mapred             0 2025-09-20 09:50 /user/history  
drwxrwxrwx  - hdfs   hdfsadmingroup      0 2025-09-20 09:50 /user/hive  
drwxrwxrwx  - hue    hue                0 2025-09-20 09:50 /user/hue  
drwxrwxrwx  - livy   livy               0 2025-09-20 09:50 /user/livy  
drwxrwxrwx  - oozie  oozie              0 2025-09-20 09:52 /user/oozie  
drwxrwxrwx  - root   hdfsadmingroup      0 2025-09-20 09:50 /user/root  
drwxrwxrwx  - spark   spark              0 2025-09-20 09:50 /user/spark  
drwxrwxrwx  - zeppelin hdfsadmingroup      0 2025-09-20 09:50 /user/zeppelin  
[hadoop@ip-172-31-8-198 ~]$ █
```

```
[hadoop@ip-172-31-8-198 ~]$ hadoop fs -put /home/hadoop/card_transactions.csv /user/CCFD_project/card_transactions.csv  
[hadoop@ip-172-31-8-198 ~]$ hadoop fs -ls /user/CCFD_project/  
Found 1 items  
-rw-r--r--  1 hadoop  hdfsadmingroup  4829520 2025-09-20 10:04 /user/CCFD_project/card_transactions.csv  
[hadoop@ip-172-31-8-198 ~]$ █
```

5: Moving data from EC2 machine to HDFS.

```
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/CCFD_Project/card_transactions.csv'  
> TBLPROPERTIES ("skip.header.line.count"="1");  
OK  
Time taken: 0.348 seconds  
hive> █
```

6 : Creating historical data External table.

```

hive> 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;
OK
Time taken: 0.516 seconds
hive> show tables;
OK
card_transactions_ext
card_transactions_orc
Time taken: 0.049 seconds, Fetched: 2 row(s)
hive> 
```

7: Creating ORC table from external table

```

hive> LOAD DATA INPATH '/user/CCFD_project/card_transactions.csv' INTO TABLE card_transactions_ext;
Loading data to table ccfd.card_transactions_ext
OK
Time taken: 0.687 seconds

```

8: loading data into external table.

```

hive> INSERT OVERWRITE TABLE CARD_TRANSACTIONS_ORC
    > SELECT CARD_ID, MEMBER_ID, AMOUNT, POSTCODE, POS_ID, CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(TRANSACTION_DT,'dd-M
M-yyyy HH:mm:ss')) AS TIMESTAMP), STATUS FROM CARD_TRANSACTIONS_EXT;
Query ID = hadoop_20250920100844_0879aa6e-baec-4687-985e-6ab3eb01d361
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0001)

-----
      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: 7.82 s
-----
Loading data to table ccfd.card_transactions_orc
OK
Time taken: 11.342 seconds
hive> 
```

9: Inserting data into orc table

```

hive> select count(*) from card_transactions_ext;
Query ID = hadoop_20250920101007_29b3c0ee-b554-44df-a69c-6a8f757b8212
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0001)

-----
      VERTICES     MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    1        1        0        0        0        0        0
Reducer 2 ..... container  SUCCEEDED    1        1        0        0        0        0        0
-----
VERTICES: 02/02  [=====>>] 100%  ELAPSED TIME: 3.86 s
-----
OK
53292
Time taken: 6.205 seconds, Fetched: 1 row(s)
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: 0.125 seconds, Fetched: 10 row(s)
hive> 

```

10: Validating external table records & timestamp column in orc table.

```

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_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: 3.226 seconds
hive> 

```

11: Creating card_transactions_hbase hive-hbase integrated table.

```

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 = hadoop_20250920101143_d064a84d-024a-4cc1-a5cb-940645fc81b5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0001)

-----
      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: 8.47 s
-----
OK
Time taken: 13.923 seconds
hive> 

```

12: Inserting data into from orc table.

```

hive> select * from card_transactions_hbase limit 10;
OK
000008b9-7ba0-490e-a858-215a44c9534a  5391723993945313  997128952368160 928092.0  15139
897809  2017-12-26 23:57:23  GENUINE
00006283-b90b-490e-9767-016b1f5b5049  5380072688020054  036571958569825 7789248.0  78349
152480  2018-01-22 18:30:19  GENUINE
00014053-9526-4a4b-91b5-a9bc4542d097  344002520206946  313129704156102 2326643.0  29456  41857856
016-08-15 23:50:35  GENUINE
0001a103-c310-491b-92f6-f4f06a8fa4ef  4540807128933493  241809163782996 4902721.0  76649
440301  2017-04-17 07:32:28  GENUINE
00020a46-e50c-4950-962d-9d2c16d565c5  349143706735646  343824445342591 4813100.0  89439  27472749
017-07-20 18:01:08  GENUINE
0002fe73-ab94-4fff-88fc-f9df935a7180  4314008605559737  846409651699691 845437.0  45880
506070  2017-03-02 23:04:14  GENUINE
000614e8-c98a-4b09-b22f-7374e2930856  6011544071439690  583773644956274 2846891.0  58544
940538  2017-12-31 23:20:13  GENUINE
00072f0d-7a73-4509-8426-e3641c134f93  6011740360743178  066377656985754 1803531.0  65032

```

```

hive> select * from card_transactions_hbase limit 10;
OK
000008b9-7ba0-490e-a858-215a44c9534a    5391723993945313    997128952368160 928092.0    15139    966249108
897809 2017-12-26 23:57:23    GENUINE
00006283-b90b-490e-9767-016b1f5b5049    5380072688020054    036571958569825 7789248.0    78349    750827915
152480 2018-01-22 18:30:19    GENUINE
00014053-9526-4a4b-91b5-a9bc4542d097    344002520206946 313129704156102 2326643.0    29456    418578505560437 2
016-08-15 23:50:35    GENUINE
0001a103-c310-491b-92f6-f4f06a8fa4ef    4540807128933493    241809163782996 4902721.0    76649    170198462
440301 2017-04-17 07:32:28    GENUINE
00020a46-e50c-4950-962d-9d2c16d565c5    349143706735646 343824445342591 4813100.0    89439    274727493822152 2
017-07-20 18:01:08    GENUINE
0002fe73-ab94-4fff-88fc-f9df935a7180    4314008605559737    846409651699691 845437.0    45880    465558985
506070 2017-03-02 23:04:14    GENUINE
000614e8-c98a-4b09-b22f-7374e2930856    6011544071439690    583773644956274 2846891.0    58544    801035095
940538 2017-12-31 23:20:13    GENUINE
00072f0d-7a73-4509-8426-e3641c134f93    6011740360743178    066377656985754 1803531.0    65032    779289002
769653 2017-07-16 08:40:58    GENUINE
0008609c-5062-4abe-b24f-c647c0db80f3    348890647161465 320864375768815 589728.0    49797    710253657063163 2
018-01-16 09:46:05    GENUINE
000960d1-2a48-4066-b8b6-08d64e59fbdf    6228588350544786    669696032320589 3181868.0    19512    448946531
936650 2016-10-07 07:05:44    GENUINE
Time taken: 0.147 seconds, Fetched: 10 row(s)
hive> 

```

13: Validating HBase table 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_family:postcode, lookup_transaction_family:transaction_dt")
   > TBLPROPERTIES ("hbase.table.name" = "lookup_data_hive");
OK
Time taken: 2.308 seconds
hive> Describe lookup_data_hbase;
OK
card_id          string
ucl              double
score            int
postcode         string
transaction_dt  timestamp
Time taken: 0.036 seconds, Fetched: 5 row(s)
hive> 

```

14: Creating Lookup Table & Validating it.

```

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_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.3110 seconds

```

15: Checking details of card_transactions_hive hive-HBase integrated table

```

Current count: 44000, row: d311183a-d418-47d3-8e6c-d8e55dcf3202
Current count: 45000, row: d80ecd9-124e-4dbe-bff9-bf4e1e03b52a
Current count: 46000, row: dce2c8e2-a939-4b14-a8fa-ce82864f26df
Current count: 47000, row: e1c4d097-6d15-40f5-b023-7cb41bee8ca1
Current count: 48000, row: e6726c23-0996-4fe6-b8f1-1d29a712a47a
Current count: 49000, row: eb5e0ffe-0059-47cc-bb98-4dfcb7eaa908
Current count: 50000, row: f023d21b-b943-40cb-afdb-58e379f91703
Current count: 51000, row: f4e5b082-fb9f-4080-a946-9f856c3c0f96
Current count: 52000, row: f9b66a63-e3db-4d8b-97da-1a361d781c36
Current count: 53000, row: fe953f70-9989-4b86-879f-5903c457eddc
53292 row(s) in 2.1400 seconds

=> 53292
hbase(main):003:0> █

```

16: Checking the count of the table.

```

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 VERSIONS => '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 0.0170 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 1.9110 seconds

hbase(main):005:0> 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', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'lookup_transaction_family', BLOOMFILTER => 'ROW', VERSIONS => '10', 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 0.0200 seconds

```

17: Checking the details of lookup_data_hive hive-HBase integrated table. & Altering the lookup_data_hive table and set VERSIONS to 10 for lookup_transaction_family & confirming with describe command.

```
[hadoop@ip-172-31-8-198 ~]$ wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
--2025-09-20 10:18:19-- https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
Resolving de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)... 52.216.220.9, 52.217.228.7
3, 3.5.16.61, ...
Connecting to de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)|52.216.220.9|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4079310 (3.9M) [application/x-gzip]
Saving to: 'mysql-connector-java-8.0.25.tar.gz'

100%[=====] 4,079,310 --.-K/s in 0.02s

2025-09-20 10:18:19 (161 MB/s) - 'mysql-connector-java-8.0.25.tar.gz' saved [4079310/4079310]
```

18: For Scoop Import, we will install MySQL connector before starting with Apache Sqoop.

```
[hadoop@ip-172-31-8-198 ~]$ tar -xvf mysql-connector-java-8.0.25.tar.gz
mysql-connector-java-8.0.25/
mysql-connector-java-8.0.25/src/
mysql-connector-java-8.0.25/src/build/
mysql-connector-java-8.0.25/src/build/java/
mysql-connector-java-8.0.25/src/build/java/documentation/
mysql-connector-java-8.0.25/src/build/java/instrumentation/
mysql-connector-java-8.0.25/src/build/misc/
mysql-connector-java-8.0.25/src/build/misc/debian.in/
mysql-connector-java-8.0.25/src/build/misc/debian.in/source/
mysql-connector-java-8.0.25/src/demo/
mysql-connector-java-8.0.25/src/demo/java/
mysql-connector-java-8.0.25/src/demo/java/demo/
mysql-connector-java-8.0.25/src/demo/java/demo/x/
```

```
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TableSelectTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TableTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TableUpdateTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TransactionTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/package-info.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/InternalXBaseTestCase.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/MysqlxSessionTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/XProtocolAsyncTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/XProtocolAuthTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/XProtocolTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/package-info.java
[hadoop@ip-172-31-8-198 ~]$
```

```
[hadoop@ip-172-31-8-198 ~]$ cd mysql-connector-java-8.0.25/
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ ls /usr/lib/sqoop/lib/
ant-contrib-1.0b3.jar          jackson-core-2.6.7.jar          parquet-column-1.6.0.jar
ant-eclipse-1.0-jvm1.2.jar      jackson-core-asl-1.9.13.jar    parquet-common-1.6.0.jar
avro-1.8.2.jar                 jackson-databind-2.6.7.4.jar  parquet-encoding-1.6.0.jar
avro-mapred-1.8.2-hadoop2.jar   jackson-mapper-asl-1.9.13.jar  parquet-format-2.2.0-rc1.jar
aws-glue-datacatalog-hive2-client.jar kite-data-core-1.1.0.jar    parquet-generator-1.6.0.jar
commons-codec-1.4.jar           kite-data-hive-1.1.0.jar    parquet-hadoop-1.6.0.jar
commons-compress-1.8.1.jar       kite-data-mapreduce-1.1.0.jar  parquet-jackson-1.6.0.jar
commons-io-1.4.jar              kite-hadoop-compatibility-1.1.0.jar postgresql-jdbc.jar
commons-jexl-2.1.1.jar           mariadb-connector-java.jar  RedshiftJDBC.jar
commons-lang3-3.4.jar            mysql-connector-java-8.0.25.jar slf4j-api-1.6.1.jar
commons-logging-1.1.1.jar        opencsv-2.3.jar          snappy-java-1.1.7.3.jar
hsqldb-1.8.0.10.jar             paranamer-2.7.jar         xz-1.5.jar
jackson-annotations-2.6.0.jar    parquet-avro-1.6.0.jar
```

19: Scoop import for card member table

```
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ sqoop import \
> --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --username upgraduser \
> --password upgraduser \
> --table card_member \
> --target-dir /user/CCFD_project/card_member \
> -m 1
Warning: /usr/lib/sqoop/.../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.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]
25/09/20 10:22:14 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
```

```
FILE: Number of write operations=0
HDFS: Number of bytes read=87
HDFS: Number of bytes written=85081
HDFS: Number of read operations=4
HDFS: Number of large read operations=0
HDFS: Number of write operations=2

Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=9547776
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=3108
    Total vcore-milliseconds taken by all map tasks=3108
    Total megabyte-milliseconds taken by all map tasks=9547776

Map-Reduce Framework
    Map input records=999
    Map output records=999
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=72
    CPU time spent (ms)=1490
    Physical memory (bytes) snapshot=343183360
    Virtual memory (bytes) snapshot=4644302848
    Total committed heap usage (bytes)=317194240

File Input Format Counters
    Bytes Read=0
File Output Format Counters
    Bytes Written=85081

25/09/20 10:22:35 INFO mapreduce.ImportJobBase: Transferred 83.0869 KB in 17.7235 seconds (4.688 KB/sec)
25/09/20 10:22:35 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ █
```

20: 999 records imported.

```
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ sqoop import \
> --connect jdbc:mysql://upgradawsrds1.cyaielc9bmnf.us-east-1.rds.amazonaws.com/cred_financials_data \
> --username upgraduser \
> --password upgraduser \
> --table member_score \
> --target-dir /user/CCFD_project/member_score \
> -m 1
Warning: /usr/lib/sqoop/.../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.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]
25/09/20 10:24:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
25/09/20 10:24:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
25/09/20 10:24:38 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
```

21:Scoop import for member score table.

```
FILE: Number of write operations=0
HDFS: Number of bytes read=87
HDFS: Number of bytes written=19980
HDFS: Number of read operations=4
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=9424896
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=3068
    Total vcore-milliseconds taken by all map tasks=3068
    Total megabyte-milliseconds taken by all map tasks=9424896
Map-Reduce Framework
    Map input records=999
    Map output records=999
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=73
    CPU time spent (ms)=1440
    Physical memory (bytes) snapshot=336470016
    Virtual memory (bytes) snapshot=4638531584
    Total committed heap usage (bytes)=317718528
File Input Format Counters
    Bytes Read=0
File Output Format Counters
    Bytes Written=19980
25/09/20 10:24:56 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 15.1455 seconds (1.2883 KB/sec)
25/09/20 10:24:56 INFO mapreduce.ImportJobBase: Retrieved 999 records.
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ ]
```

22: 999 records imported.

```
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ hadoop fs -ls /user/CCFD_project/card_member
Found 2 items
-rw-r--r-- 1 hadoop hdfsadmingroup          0 2025-09-20 10:22 /user/CCFD_project/card_member/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmingroup 85081 2025-09-20 10:22 /user/CCFD_project/card_member/part-m-00000
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ hadoop fs -ls /user/CCFD_project/member_score
Found 2 items
-rw-r--r-- 1 hadoop hdfsadmingroup          0 2025-09-20 10:24 /user/CCFD_project/member_score/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmingroup 19980 2025-09-20 10:24 /user/CCFD_project/member_score/part-m-00000
[hadoop@ip-172-31-8-198 mysql-connector-java-8.0.25]$ █
```

23: Checking the scoop imported files.

```
hive> show tables;
OK
card_transactions_ext
card_transactions_hbase
card_transactions_orc
lookup_data_hbase
Time taken: 0.153 seconds, Fetched: 4 row(s)
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/CCFD_project/card_member';
OK
Time taken: 0.125 seconds
hive> █
```

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS MEMBER_SCORE_EXT(
    > `MEMBER_ID` STRING,
    > `SCORE` INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/user/CCFD_project/member_score';
OK
Time taken: 0.171 seconds
hive> █
```

24:Creating external tables for RDS data.

```

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: 0.613 seconds
hive>

```

```

hive> CREATE TABLE IF NOT EXISTS MEMBER_SCORE_ORC(
    > `MEMBER_ID` STRING,
    > `SCORE` INT)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.081 seconds
hive>

```

25:Creating ORC tables from the external tables.

```

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 = hadoop_20250920102924_2d8be5c6-3538-493d-a949-5de62e3330ef
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_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: 4.56 s
-----
Loading data to table ccfd.card_member_orc
OK
Time taken: 7.814 seconds
hive>

```

```

hive> INSERT OVERWRITE TABLE MEMBER_SCORE_ORC
    > SELECT MEMBER_ID, SCORE FROM MEMBER_SCORE_EXT;
Query ID = hadoop_20250920103014_09505347-6d24-4bae-bb45-94d8690ee863
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_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: 4.68 s
-----
Loading data to table ccfd.member_score_orc
OK
Time taken: 5.514 seconds
hive> █

```

26:Inserting data into ORC tables from external tables.

```

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: 0.051 seconds
hive> █

```

```

hive> CREATE TABLE IF NOT EXISTS CARD_UCL_ORC(
    > `CARD_ID` STRING,
    > `UCL` DOUBLE)
    > STORED AS ORC
    > TBLPROPERTIES ("orc.compress"="SNAPPY");
OK
Time taken: 0.052 seconds
hive> █

```

27: Creating Ranking ORC table of last 10 customer transactions & Creating Card UCL ORC table for each card

```

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_TRANSACTIONS_HBASE WHERE STATUS = 'GENUINE') A ) B
> WHERE B.RANK <= 10;
Query ID = hadoop_20250920103222_609d61a0-b49f-4a31-865a-aac9b12d6226
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0004)

-----
      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: 8.51 s
-----
Loading data to table ccfd.ranked_card_transactions_orc
OK
Time taken: 15.306 seconds
hive> 
```

28: Inserting data into ranked_card_transacations_orc table with conditions.

```

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 = hadoop_20250920103412_02d447e8-5de7-4f3b-b134-dbe6719afcd3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0004)

-----
      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: 4.77 s
-----
Loading data to table ccfd.card_ucl_orc
OK
Time taken: 5.587 seconds
hive> 
```

29: Inserting data into card_ucl_orc table with conditions.

```

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 ccfd@ranked_card_transactions_orc, Columns: postcode, rank, transaction_dt, card_id
No Stats for ccfd@card_ucl_orc, Columns: card_id, ucl
No Stats for ccfd@card_member_orc, Columns: member_id, card_id
No Stats for ccfd@member_score_orc, Columns: member_id, score
Query ID = hadoop_20250920103454_525a5786-991c-4d6a-9d20-4d2402257ca9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_0004)

-----
      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: 8.57 s
-----
```

30: Loading data in lookup_data_hbase table.

```

hive> SELECT COUNT(*) FROM LOOKUP_DATA_HBASE;SELECT COUNT(*) FROM LOOKUP_DATA_HBASE;
Query ID = hadoop_20250920103626_9998c63c-fbc6-4dd8-9734-7abd83bea117
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1758361848317_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: 6.14 s
-----
OK
999
Time taken: 15.01 seconds, Fetched: 1 row(s)
```

```

hive> SELECT * FROM LOOKUP_DATA_HBASE LIMIT 10;
OK
340028465709212 1.633155548882348E7    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: 0.13 seconds, Fetched: 10 row(s)
hive> ■

```

31: Verifying lookup table data.

```

hbase(main):004:0> count 'lookup_data_hive'
999 row(s) in 0.1240 seconds

=> 999
hbase(main):005:0> ■

```

32: Checking count in lookup_data_hive table

```

6595638658736751          column=lookup_card_family:score, timestamp=1758364511884, value=310
6595638658736751          column=lookup_card_family:ucl, timestamp=1758364511884, value=1.356629177577566E7
6595638658736751          column=lookup_transaction_family:postcode, timestamp=1758364511884, value=68328
6595638658736751          column=lookup_transaction_family:transaction_dt, timestamp=1758364511884, value=201
8-01-30 10:50:34
6595814135833988          column=lookup_card_family:score, timestamp=1758364511884, value=210
6595814135833988          column=lookup_card_family:ucl, timestamp=1758364511884, value=1.3926273240525039E7
6595814135833988          column=lookup_transaction_family:postcode, timestamp=1758364511884, value=22508
6595814135833988          column=lookup_transaction_family:transaction_dt, timestamp=1758364511884, value=201
8-01-30 02:03:54
6595928469079750          column=lookup_card_family:score, timestamp=1758364511884, value=412
6595928469079750          column=lookup_card_family:ucl, timestamp=1758364511884, value=1.142797041440079E7
6595928469079750          column=lookup_transaction_family:postcode, timestamp=1758364511884, value=98349
6595928469079750          column=lookup_transaction_family:transaction_dt, timestamp=1758364511884, value=201
8-01-24 12:38:22
6597703848279563          column=lookup_card_family:score, timestamp=1758364511884, value=218
6597703848279563          column=lookup_card_family:ucl, timestamp=1758364511884, value=1.4718634149498457E7
6597703848279563          column=lookup_transaction_family:postcode, timestamp=1758364511884, value=95699
6597703848279563          column=lookup_transaction_family:transaction_dt, timestamp=1758364511884, value=201
8-01-27 10:51:49
6598830758632447          column=lookup_card_family:score, timestamp=1758364511884, value=293
6598830758632447          column=lookup_card_family:ucl, timestamp=1758364511884, value=1.2227949982601807E7
6598830758632447          column=lookup_transaction_family:postcode, timestamp=1758364511884, value=19421
6598830758632447          column=lookup_transaction_family:transaction_dt, timestamp=1758364511884, value=201
8-01-30 00:18:34
6599900931314251          column=lookup_card_family:score, timestamp=1758364511884, value=297
6599900931314251          column=lookup_card_family:ucl, timestamp=1758364511884, value=1.2121408572464656E7
6599900931314251          column=lookup_transaction_family:postcode, timestamp=1758364511884, value=97423
6599900931314251          column=lookup_transaction_family:transaction_dt, timestamp=1758364511884, value=201
8-01-31 11:25:16
999 row(s) in 0.6640 seconds

```

33: Checking data in lookup_data_hive table.

