

Capstone Project

Mid Submission

Scripts Execution

By Venkatesh Jagannathan

Transactions History Intermediate Table Load

```
hive> CREATE TABLE CARD_TRANSACTIONS_TEMP(card_id string,member_id string,amount double,postcode
string,pos_id string,transaction_dt string,status string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 2.819 seconds
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/capstone/card_transactions.csv' OVERWRITE INTO TABLE
CARD_TRANSACTIONS_TEMP;LOAD DATA LOCAL INPATH '/home/ec2-user/capstone/card_transactions.csv' OVERWRITE
INTO TABLE CARD_TRANSACTIONS_TEMP;
Loading data to table default.card_transactions_temp
Table default.card_transactions_temp stats: [numFiles=1, totalSize=4829520]
OK
Time taken: 0.842 seconds
Loading data to table default.card_transactions_temp
Table default.card_transactions_temp stats: [numFiles=1, totalSize=4829520]
OK
Time taken: 0.861 seconds
hive> select * from card_transactions_temp limit 5;
OK
card_id member_id      NULL      postcode      pos_id  transaction_dt  status
348702330256514 000037495066290 9084849.0 33946 614677375609919 11-02-2018 00:00:00 GENUINE
348702330256514 000037495066290 330148.0 33946 614677375609919 11-02-2018 00:00:00 GENUINE
348702330256514 000037495066290 136052.0 33946 614677375609919 11-02-2018 00:00:00 GENUINE
348702330256514 000037495066290 4310362.0 33946 614677375609919 11-02-2018 00:00:00 GENUINE
Time taken: 0.362 seconds, Fetched: 5 row(s)
hive>
```

Transactions History Load

```
hive> INSERT INTO card_transactions
> SELECT NAMED_STRUCT('card_id',card_id,'transaction_dt',unix_timestamp(transaction_dt, "DD-MM-YYYY
HH:mm:ss")), card_id, member_id,amount,postcode,pos_id,unix_timestamp(transaction_dt, "DD-MM-YYYY
HH:mm:ss"),status FROM CARD_TRANSACTIONS_TEMP;
Query ID = ec2-user_20190827142626_4329a8db-53f3-4c75-a90b-7e400adf140e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1566905077309_0012, Tracking URL = http://ip-172-31-93-
76.ec2.internal:8088/proxy/application_1566905077309_0012/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill
job_1566905077309_0012
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0
2019-08-27 14:26:48,033 Stage-0 map = 0%, reduce = 0%
2019-08-27 14:27:03,151 Stage-0 map = 100%, reduce = 0%, Cumulative CPU 12.77 sec
MapReduce Total cumulative CPU time: 12 seconds 770 msec
Ended Job = job_1566905077309_0012
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1 Cumulative CPU: 12.77 sec HDFS Read: 4835704 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 770 msec
OK
Time taken: 46.327 seconds
hive> select * from card_transactions limit 5;
OK
{"card_id":"340028465709212","transaction_dt":1451180689} 340028465709212 009250698176266 9230684.0
98588 489939408321423 1451180689 GENUINE
{"card_id":"340028465709212","transaction_dt":1451182163} 340028465709212 009250698176266 4318929.0
97641 566753248680777 1451182163 GENUINE
{"card_id":"340028465709212","transaction_dt":1451182714} 340028465709212 009250698176266 451162.0
71363 633734711060942 1451182714 GENUINE
{"card_id":"340028465709212","transaction_dt":1451186824} 340028465709212 009250698176266 4748901.0
58641 044677148710313 1451186824 GENUINE
{"card_id":"340028465709212","transaction_dt":1451202542} 340028465709212 009250698176266 3438153.0
13051 685692072919031 1451202542 GENUINE
Time taken: 0.117 seconds, Fetched: 5 row(s)
hive>
```

```

hbase(main):014:0> get 'card_transactions', '340028465709212~1451180689'
COLUMN                                CELL
card_details:card_id timestamp=1566916017686, value=340028465709212
card_details:member_ timestamp=1566916017686, value=009250698176266
id
transaction_details: timestamp=1566916017686, value=9230684.0
amount
transaction_details: timestamp=1566916017686, value=489939408321423
pos_id
transaction_details: timestamp=1566916017686, value=98588
postcode
transaction_details: timestamp=1566916017686, value=GENUINE
status
transaction_details: timestamp=1566916017686, value=1451180689
transaction_dt
7 row(s) in 0.2210 seconds

```

Last 10 Genuine Transactions Load

```

hive> INSERT INTO LOOKUP_TXN_MASTER
> SELECT NAMED_STRUCT('card_id',card_id,'transaction_dt',unix_timestamp(transaction_dt, "DD-MM-YYYY
HH:mm:ss")), card_id,member_id,amount,postcode,pos_id,unix_timestamp(transaction_dt, "DD-MM-YYYY
HH:mm:ss"),status FROM (
>   SELECT *, RANK() OVER(PARTITION BY card_id ORDER BY transaction_dt DESC) txrank
>   FROM(
>     SELECT *
>     FROM card_transactions
>     WHERE status = "GENUINE"
>   ) genuine_transactions
> ) ranked_transactions
> WHERE txrank <= 10;
FAILED: SemanticException [Error 10014]: Line 2:55 Wrong arguments '"DD-MM-YYYY HH:mm:ss"': The function
UNIX_TIMESTAMP takes only string/date/timestamp types
hive> INSERT INTO LOOKUP_TXN_MASTER
> SELECT NAMED_STRUCT('card_id',card_id,'transaction_dt',transaction_dt),
card_id,member_id,amount,postcode,pos_id,transaction_dt,status FROM (
>   SELECT *, RANK() OVER(PARTITION BY card_id ORDER BY transaction_dt DESC) txrank
>   FROM(
>     SELECT *
>     FROM card_transactions
>     WHERE status = "GENUINE"
>   ) genuine_transactions
> ) ranked_transactions
> WHERE txrank <= 10;
Query ID = ec2-user_20190827151818_33ce341e-cd57-4b6f-9d23-6ce599d3668b
Total jobs = 1
Launching Job 1 out of 1

```

```

Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1566905077309_0014, Tracking URL = http://ip-172-31-93-
76.ec2.internal:8088/proxy/application_1566905077309_0014/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill
job_1566905077309_0014
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 1
2019-08-27 15:19:19,433 Stage-0 map = 0%, reduce = 0%
2019-08-27 15:19:31,030 Stage-0 map = 100%, reduce = 0%, Cumulative CPU 9.54 sec
2019-08-27 15:19:40,537 Stage-0 map = 100%, reduce = 100%, Cumulative CPU 19.26 sec
MapReduce Total cumulative CPU time: 19 seconds 260 msec
Ended Job = job_1566905077309_0014
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1 Reduce: 1 Cumulative CPU: 19.26 sec HDFS Read: 30315 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 260 msec

```

```

Time taken: 42.965 seconds
hive> select * from lookup_txn_master limit 15;
OK
{"card_id":"340028465709212","transaction_dt":1483268755}      340028465709212 009250698176266 9687739.0
51542 498255481028848 1483268755 GENUINE
{"card_id":"340028465709212","transaction_dt":1483274172}      340028465709212 009250698176266 581323.0
46182 509545401949131 1483274172 GENUINE
{"card_id":"340028465709212","transaction_dt":1483285595}      340028465709212 009250698176266 8851266.0
96738 678279765052404 1483285595 GENUINE
{"card_id":"340028465709212","transaction_dt":1483289096}      340028465709212 009250698176266 5829887.0
24078 485353426643646 1483289096 GENUINE
{"card_id":"340028465709212","transaction_dt":1483289833}      340028465709212 009250698176266 8737772.0
82933 219307586884330 1483289833 GENUINE
{"card_id":"340028465709212","transaction_dt":1483299084}      340028465709212 009250698176266 7381252.0
63951 305578834263571 1483299084 GENUINE
{"card_id":"340028465709212","transaction_dt":1483300540}      340028465709212 009250698176266 2122378.0
95636 653965248170086 1483300540 GENUINE
{"card_id":"340028465709212","transaction_dt":1483309107}      340028465709212 009250698176266 8505019.0
35958 551439667943807 1483309107 GENUINE
{"card_id":"340028465709212","transaction_dt":1483310250}      340028465709212 009250698176266 4917958.0
92250 561945266720889 1483310250 GENUINE
{"card_id":"340028465709212","transaction_dt":1514690735}      340028465709212 009250698176266 8696557.0

```

Analysis – Calculate Average and STDDEV from last 10 transactions into stage table

```

hive> CREATE EXTERNAL TABLE TXN_STATS_STG(
  > card_id string,
  > moving_average double,
  > std_dev double
  > )
  > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.042 seconds
hive> INSERT INTO TXN_STATS_STG
  > SELECT card_id, AVG(amount), stddev_pop(amount) FROM LOOKUP_TXN_MASTER
  > GROUP BY card_id;
Query ID = ec2-user_20190827160707_eb32e836-98ea-4830-94d1-9ea97905dbc1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1566905077309_0017, Tracking URL = http://ip-172-31-93-
76.ec2.internal:8088/proxy/application_1566905077309_0017/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill
job_1566905077309_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

```

```

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-27 16:07:57,897 Stage-1 map = 0%, reduce = 0%
2019-08-27 16:08:07,271 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.11 sec
2019-08-27 16:08:16,695 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 11.03 sec
MapReduce Total cumulative CPU time: 11 seconds 30 msec
Ended Job = job_1566905077309_0017
Loading data to table default.txn_stats_stg
Table default.txn_stats_stg stats: [numFiles=2, numRows=999, totalSize=114336, rawDataSize=44181]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 11.03 sec HDFS Read: 11799 HDFS Write: 45261 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 30 msec
OK
Time taken: 38.383 seconds
hive> select * from txn_stats_stg limit 5;
OK
340028465709212 6531115.1 2957293.743006279
340054675199675 4849223.2 2889521.298857539
340082915339645 7874693.2 1374378.2746957112
340134186926007 6891678.3 2599759.575385234
340265728490548 5369392.0 2989323.112256017
Time taken: 0.039 seconds, Fetched: 5 row(s)
hive>

```

Load Last Transaction Detail into stage table

```

hive>
> CREATE TABLE TXN_LAST_DETAIL_STG(
>   card_id string,
>   postcode string,
>   transaction_dt bigint
> )
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.27 seconds
hive> INSERT INTO TXN_LAST_DETAIL_STG
>   SELECT card_id,postcode,transaction_dt
>   FROM LOOKUP_TXN_MASTER
>   where txrank = 1;
Query ID = ec2-user_20190827161717_48ff4687-0175-4518-ac19-af677bae3c79
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1566905077309_0020, Tracking URL = http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1566905077309_0020/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill
job_1566905077309_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2019-08-27 16:17:32,686 Stage-1 map = 0%, reduce = 0%
2019-08-27 16:17:42,159 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 7.3 sec

```

```

MapReduce Total cumulative CPU time: 7 seconds 300 msec
Ended Job = job_1566905077309_0020
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://ip-172-31-93-76.ec2.internal:8020/user/hive/warehouse/txn_last_detail_stg/.hive-
staging_hive_2019-08-27_16-17-16_937_430952570074537861-1/-ext-10000
Loading data to table default.txn_last_detail_stg
Table default.txn_last_detail_stg stats: [numFiles=1, numRows=999, totalSize=33693, rawDataSize=32694]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 7.3 sec HDFS Read: 6505 HDFS Write: 33780 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 300 msec
OK
Time taken: 26.949 seconds
hive> select * from txn_last_detail_stg limit 5;
OK
340028465709212 24658 1514690735
340054675199675 50140 1514749403
340082915339645 17844 1514747027
340134186926007 67576 1514761970
340265728490548 72435 1514686055
Time taken: 0.045 seconds, Fetched: 5 row(s)
hive>

```

RDS to Hadoop Incremental Data Ingestion as saved scoop job based on member_joining_dt

card_member

```

[ec2-user@ip-172-31-93-76 ~]$ sqoop job --create import_card_member -- import --connect
jdbc:mysql://upgradawsrds.cpclxrkdvwzmz.us-east-1.rds.amazonaws.com/cred_financials_data --username
upgraduser --P --table card_member --incremental append --check-column member_joining_dt --last-value
'2010-02-02 00:00:00' --warehouse-dir /input/data/tables --hive-import --null-string '\N' --null-non-
string '\N' --map-column-hive member_joining_dt=timestamp
Warning: /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/bin/../lib/sqoop/../accumulo does not exist!
Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
19/08/25 15:40:57 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
Enter password:
19/08/25 15:41:03 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
19/08/25 15:41:03 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
[ec2-user@ip-172-31-93-76 ~]$ sqoop job --exec import_card_member
Warning: /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/bin/../lib/sqoop/../accumulo does not exist!
Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
19/08/25 15:41:11 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
Enter password:
19/08/25 15:41:17 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
19/08/25 15:41:17 INFO tool.CodeGenTool: Beginning code generation
19/08/25 15:41:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t
LIMIT 1
19/08/25 15:41:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t
LIMIT 1
19/08/25 15:41:17 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-
mapreduce

```


Note: /tmp/sqoop-ec2-user/compile/6a41a6977269c083b26901b1174a4e05/card_member.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.

```

19/08/25 15:41:19 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-
user/compile/6a41a6977269c083b26901b1174a4e05/card_member.jar
19/08/25 15:41:20 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX
(`member_joining_dt`) FROM `card_member`
19/08/25 15:41:20 INFO tool.ImportTool: Incremental import based on column `member_joining_dt`
19/08/25 15:41:20 INFO tool.ImportTool: Lower bound value: '2010-02-02 00:00:00'
19/08/25 15:41:20 INFO tool.ImportTool: Upper bound value: '2018-02-01 08:02:34.0'
19/08/25 15:41:20 WARN manager.MySQLManager: It looks like you are importing from mysql.
19/08/25 15:41:20 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
19/08/25 15:41:20 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
19/08/25 15:41:20 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
19/08/25 15:41:20 INFO mapreduce.ImportJobBase: Beginning import of card_member
19/08/25 15:41:20 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/08/25 15:41:20 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use
mapreduce.job.maps
19/08/25 15:41:20 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-93-
76.ec2.internal/172.31.93.76:8032
19/08/25 15:41:27 INFO db.DBInputFormat: Using read committed transaction isolation
19/08/25 15:41:27 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`card_id`), MAX(`card_id`)
FROM `card_member` WHERE ( `member_joining_dt` > '2010-02-02 00:00:00' AND `member_joining_dt` <= '2018-02
-01 08:02:34.0' )

19/08/25 15:41:27 WARN db.TextSplitter: Generating splits for a textual index column.
19/08/25 15:41:27 WARN db.TextSplitter: If your database sorts in a case-insensitive order, this may result
in a partial import or duplicate records.
19/08/25 15:41:27 WARN db.TextSplitter: You are strongly encouraged to choose an integral split column.
19/08/25 15:41:28 INFO mapreduce.JobSubmitter: number of splits:6
19/08/25 15:41:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1566736779983_0004
19/08/25 15:41:28 INFO impl.YarnClientImpl: Submitted application application_1566736779983_0004
19/08/25 15:41:28 INFO mapreduce.Job: The url to track the job: http://ip-172-31-93-
76.ec2.internal:8088/proxy/application_1566736779983_0004/
19/08/25 15:41:28 INFO mapreduce.Job: Running job: job_1566736779983_0004
19/08/25 15:41:38 INFO mapreduce.Job: Job job_1566736779983_0004 running in uber mode : false
19/08/25 15:41:38 INFO mapreduce.Job:  map 0% reduce 0%
19/08/25 15:41:47 INFO mapreduce.Job:  map 33% reduce 0%
19/08/25 15:41:48 INFO mapreduce.Job:  map 50% reduce 0%
19/08/25 15:41:54 INFO mapreduce.Job:  map 67% reduce 0%
19/08/25 15:41:55 INFO mapreduce.Job:  map 83% reduce 0%
19/08/25 15:41:56 INFO mapreduce.Job:  map 100% reduce 0%
19/08/25 15:41:56 INFO mapreduce.Job: Job job_1566736779983_0004 completed successfully
19/08/25 15:41:56 INFO mapreduce.Job: Counters: 30
    File System Counters
        FILE: Number of bytes read=0
        FILE: Number of bytes written=1073352
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0

```

Number of map operations=6
HDFS: Number of bytes read=832
HDFS: Number of bytes written=85082
HDFS: Number of read operations=24
HDFS: Number of large read operations=0
HDFS: Number of write operations=12

Job Counters

Launched map tasks=6
Other local map tasks=6
Total time spent by all maps in occupied slots (ms)=36928
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=36928
Total vcore-milliseconds taken by all map tasks=36928
Total megabyte-milliseconds taken by all map tasks=37814272

Map-Reduce Framework

Map input records=999
Map output records=999
Input split bytes=832
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=773
CPU time spent (ms)=9660

Physical memory (bytes) snapshot=1469808640
Virtual memory (bytes) snapshot=16797999104
Total committed heap usage (bytes)=1438121984

File Input Format Counters

Bytes Read=0

File Output Format Counters

Bytes Written=85082

19/08/25 15:41:56 INFO mapreduce.ImportJobBase: Transferred 83.0879 KB in 36.0901 seconds (2.3022 KB/sec)

19/08/25 15:41:56 INFO mapreduce.ImportJobBase: Retrieved 999 records.

19/08/25 15:41:56 INFO util.AppendUtils: Creating missing output directory - card_member

19/08/25 15:41:56 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t
LIMIT 1

19/08/25 15:41:56 WARN hive.TableDefWriter: Column member_joining_dt had to be cast to a less precise type
in Hive

19/08/25 15:41:56 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.15.1-
1.cdh5.15.1.p0.4/jars/hive-common-1.1.0-cdh5.15.1.jar!/hive-log4j.properties

OK

Time taken: 2.949 seconds

Loading data to table default.card_member

Table default.card_member stats: [numFiles=6, totalSize=85082]

OK

Time taken: 0.843 seconds

fer2-user@in-172-31-93-76 ~1\$ hive


```
[ec2-user@ip-172-31-93-76 ~]$ hive
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
Java HotSpot(TM) 64-Bit Server VM warning: Using incremental CMS is deprecated and will likely be removed
in a future release
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/hive-common-1.1.0-cdh5.15.1.jar!/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> describe card_member;
OK
card_id          string
member_id        string
member_joining_dt timestamp
card_purchase_dt string
country           string
city              string
Time taken: 2.007 seconds, Fetched: 6 row(s)
hive> select * from card_member limit 5;
OK
340028465709212 009250698176266 2012-02-08 06:04:13      05/13  United States  Barborton
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
Time taken: 0.296 seconds, Fetched: 5 row(s)
hive>
```

member_score

```
[ec2-user@ip-172-31-93-76 ~]$ sqoop job --create import_member_score -- import - -connect
jdbc:mysql://upgradawsrds.cpclxrkdvmz.us-east-1.rds.amazonaws.com/cred _financials_data --username upgrader --P --
table member_score --delete-target-dir --warehouse-dir /input/data/tables --hive-import --hive-overwrite --null-str
ing '\N' --null-non-string '\N'
Warning: /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/bin/../lib/sqoop/./a ccumulo does not exist! Accumulo
imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
19/08/25 14:24:36 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
Enter password:
19/08/25 14:24:42 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for ou tput. You can override
19/08/25 14:24:42 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-b y, etc.
[ec2-user@ip-172-31-93-76 ~]$ sqoop job --exec import_member_score
Warning: /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/bin/../lib/sqoop/./accumulo does not exist! Accumulo imports will
fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
19/08/25 14:33:59 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
Enter password:
19/08/25 14:34:06 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
19/08/25 14:34:06 INFO tool.CodeGenTool: Beginning code generation
19/08/25 14:34:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
19/08/25 14:34:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
19/08/25 14:34:06 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/dc14ed3445580bac8c7376a5921959dd/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
19/08/25 14:34:07 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-
user/compile/dc14ed3445580bac8c7376a5921959dd/member_score.jar
19/08/25 14:34:09 INFO tool.ImportTool: Destination directory /input/data/tables/member_score is not present, hence not
deleting.
19/08/25 14:34:09 WARN manager.MySQLManager: It looks like you are importing from mysql.
19/08/25 14:34:09 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
19/08/25 14:34:09 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
19/08/25 14:34:09 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
19/08/25 14:34:09 INFO mapreduce.ImportJobBase: Beginning import of member_score
19/08/25 14:34:09 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/08/25 14:34:09 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
19/08/25 14:34:09 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-93-76.ec2.internal/172.31.93.76:8032
19/08/25 14:34:15 INFO db.DBInputFormat: Using read committed transaction isolation
19/08/25 14:34:15 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`member_id`), MAX(`member_id`) FROM
`member_score`
19/08/25 14:34:15 WARN db.TextSplitter: Generating splits for a textual index column.
19/08/25 14:34:15 WARN db.TextSplitter: If your database sorts in a case-insensitive order, this may result in a partial
import or duplicate records.
19/08/25 14:34:15 WARN db.TextSplitter: You are strongly encouraged to choose an integral split column.
19/08/25 14:34:15 INFO mapreduce.JobSubmitter: number of splits:6
```

19/08/25 14:34:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1566736779983_0002
19/08/25 14:34:16 INFO impl.YarnClientImpl: Submitted application application_1566736779983_0002
19/08/25 14:34:16 INFO mapreduce.Job: The url to track the job: http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1566736779983_0002/
19/08/25 14:34:16 INFO mapreduce.Job: Running job: job_1566736779983_0002
19/08/25 14:34:24 INFO mapreduce.Job: Job job_1566736779983_0002 running in uber mode : false
19/08/25 14:34:24 INFO mapreduce.Job: map 0% reduce 0%
19/08/25 14:34:32 INFO mapreduce.Job: map 17% reduce 0%
19/08/25 14:34:33 INFO mapreduce.Job: map 50% reduce 0%
19/08/25 14:34:40 INFO mapreduce.Job: map 100% reduce 0%
19/08/25 14:34:41 INFO mapreduce.Job: Job job_1566736779983_0002 completed successfully
19/08/25 14:34:42 INFO mapreduce.Job: Counters: 30
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=1066554
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=879
HDFS: Number of bytes written=19980
HDFS: Number of read operations=24
HDFS: Number of large read operations=0
HDFS: Number of write operations=12
Job Counters
Launched map tasks=6
Other local map tasks=6
Total time spent by all maps in occupied slots (ms)=36722
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=36722
Total vcore-milliseconds taken by all map tasks=36722
Total megabyte-milliseconds taken by all map tasks=37603328
Map-Reduce Framework
Map input records=999
Map output records=999
Input split bytes=879
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=860
CPU time spent (ms)=8960
Physical memory (bytes) snapshot=1356288000
Virtual memory (bytes) snapshot=16800075776
Total committed heap usage (bytes)=1386217472
File Input Format Counters

```

----- mapreduce -----
      Bytes Read=0
      File Output Format Counters
      Bytes Written=19980
19/08/25 14:34:42 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 33.0224 seconds (605.0434 bytes/sec)
19/08/25 14:34:42 INFO mapreduce.ImportJobBase: Retrieved 999 records.
19/08/25 14:34:42 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
19/08/25 14:34:42 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/hive-common-1.1.0
-cdh5.15.1.jar!/hive-log4j.properties
OK
Time taken: 3.005 seconds
Loading data to table default.member_score
chgrp: changing ownership of 'hdfs://ip-172-31-93-76.ec2.internal:8020/user/hive/warehouse/member_score': User does not
belong to hive
Table default.member_score stats: [numFiles=6, totalSize=19980]
OK
Time taken: 0.608 seconds
[ec2-user@ip-172-31-93-76 ~]$ hive
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
Java HotSpot(TM) 64-Bit Server VM warning: Using incremental CMS is deprecated and will likely be removed in a future release
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/hive-common-1.1.0
-cdh5.15.1.jar!/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> select * from member_score 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: 2.193 seconds, Fetched: 10 row(s)
hive>

```

Load Final staging table

```

hive> CREATE TABLE TXN_FINAL_STG(
  > card_id string,
  > moving_average double,
  > std_dev double,
  > postcode string,
  > transaction_dt bigint,
  > score int,
  > member_id string,
  > member_joining_dt string,
  > card_purchase_dt string,
  > country string,
  > city string
  > )
  > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
hive> INSERT INTO TXN_FINAL_STG(card_id,moving_average,std_dev,
  > postcode,transaction_dt,score,member_id,member_joining_dt,card_purchase_dt,country,city)
  > SELECT S.card_id, S.moving_average, S.std_dev,
  >
  > L.postcode,L.transaction_dt,MS.score,CM.member_id,CM.member_joining_dt,CM.card_purchase_dt,CM.country,CM.ci
ty FROM
  > TXN_STATS_STG S LEFT JOIN TXN_LAST_DETAIL_STG L
  > ON S.card_id = L.card_id
  > LEFT JOIN card_member CM
  > ON S.card_id = CM.card_id
  > LEFT JOIN member_score MS ON
  > CM.member_id = MS.member_id
  > ;

```

Query ID = ec2-user_20190827163131_ed9e7ebe-290d-4624-a9c2-12e623de801c
 Total jobs = 1
 Execution log at: /tmp/ec2-user/ec2-user_20190827163131_ed9e7ebe-290d-4624-a9c2-12e623de801c.log
 2019-08-27 04:31:17 Starting to launch local task to process map join; maximum memory = 1908932608
 2019-08-27 04:31:18 Dump the side-table for tag: 1 with group count: 999 into file: file:/tmp/ec2-user/f2ff8006-3a90-49c5-b554-378d9f5d75ac/hive_2019-08-27_16-31-12_827_505800255435740190-1/-local-10003/HashTable-Stage-7/MapJoin-mapfile11--.hashtable
 2019-08-27 04:31:18 Uploaded 1 File to: file:/tmp/ec2-user/f2ff8006-3a90-49c5-b554-378d9f5d75ac/hive_2019-08-27_16-31-12_827_505800255435740190-1/-local-10003/HashTable-Stage-7/MapJoin-mapfile11--.hashtable (37277 bytes)
 2019-08-27 04:31:18 Dump the side-table for tag: 1 with group count: 999 into file: file:/tmp/ec2-user/f2ff8006-3a90-49c5-b554-378d9f5d75ac/hive_2019-08-27_16-31-12_827_505800255435740190-1/-local-10003/HashTable-Stage-7/MapJoin-mapfile21--.hashtable
 2019-08-27 04:31:18 Uploaded 1 File to: file:/tmp/ec2-user/f2ff8006-3a90-49c5-b554-378d9f5d75ac/hive_2019-08-27_16-31-12_827_505800255435740190-1/-local-10003/HashTable-Stage-7/MapJoin-mapfile21--.hashtable (46164 bytes)
 2019-08-27 04:31:18 Dump the side-table for tag: 2 with group count: 999 into file: file:/tmp/ec2-user/f2ff8006-3a90-49c5-b554-378d9f5d75ac/hive_2019-08-27_16-31-12_827_505800255435740190-1/-local-10003/HashTable-Stage-7/MapJoin-mapfile22--.hashtable
 2019-08-27 04:31:18 Uploaded 1 File to: file:/tmp/ec2-user/f2ff8006-3a90-49c5-b554-378d9f5d75ac/hive_2019-08-27_16-31-12_827_505800255435740190-1/-local-10003/HashTable-Stage-7/MapJoin-mapfile22--.hashtable (85757 bytes)

2019-08-27 04:31:18 End of local task; Time Taken: 1.477 sec.
 Execution completed successfully
 MapredLocal task succeeded
 Launching Job 1 out of 1
 Number of reduce tasks is set to 0 since there's no reduce operator
 Starting Job = job_1566905077309_0022, Tracking URL = http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1566905077309_0022/
 Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cd5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill job_1566905077309_0022
 Hadoop job information for Stage-7: number of mappers: 1; number of reducers: 0
 2019-08-27 16:31:33,418 Stage-7 map = 0%, reduce = 0%
 2019-08-27 16:31:40,742 Stage-7 map = 100%, reduce = 0%, Cumulative CPU 3.96 sec
 MapReduce Total cumulative CPU time: 3 seconds 960 msec
 Ended Job = job_1566905077309_0022
 Loading data to table default.txn_final_stg
 Table default.txn_final_stg stats: [numFiles=1, numRows=1998, totalSize=265066, rawDataSize=263068]
 MapReduce Jobs Launched:
 Stage-Stage-7: Map: 1 Cumulative CPU: 3.96 sec HDFS Read: 126382 HDFS Write: 265149 SUCCESS
 Total MapReduce CPU Time Spent: 3 seconds 960 msec
 OK
 Time taken: 30.418 seconds
 hive> select * from txn_final_stg limit 10;

hive> select * from txn_final_stg limit 10;
 OK

340028465709212 6531115.1	2957293.743006279	24658	1514690735	233	009250698176266
2012-02-08 06:04:13 05/13	United States Barberton				
340054675199675 4849223.2	2889521.298857539	50140	1514749403	631	835873341185231
2017-03-10 09:24:44 03/17	United States Fort Dodge				
340082915339645 7874693.2	1374378.2746957112	17844	1514747027	407	512969555857346
2014-02-15 06:30:30 07/14	United States Graham				
340134186926007 6891678.3	2599759.575385234	67576	1514761970	614	887711945571282
2012-02-05 01:21:58 02/13	United States Dix Hills				
340265728490548 5369392.0	2989323.112256017	72435	1514686055	202	680324265406190
2014-03-29 07:49:14 11/14	United States Rancho Cucamonga				
340268219434811 5242562.2	2541444.7580490825	57030	1514724197	415	929799084911715
2012-07-08 02:46:08 08/12	United States San Francisco				
340379737226464 4553363.6	2572990.7968767085	50455	1514764347	229	089615510858348
2010-03-10 00:06:42 09/10	United States Clinton				
340383645652108 3680698.1	2920324.158289503	28641	1514761750	645	181180599313885
2012-02-24 05:32:44 10/16	United States West New York				
340803866934451 3638294.9	2473858.992009951	38581	1514751634	502	417664728506297
2015-05-21 04:30:45 08/17	United States Beaverton				
340889618969736 5076862.2	3100348.8399815205	61341	1514757438	330	459292914761635
2013-04-23 08:40:11 11/15	United States West Palm Beach				

 Time taken: 0.052 seconds, Fetched: 10 row(s)
 hive>

Card Member Lookup Load

```
ec2-user@ip-172-31-93-76:~/capstone
hive> INSERT INTO LOOKUP_CARD_MEMBER
> SELECT card_id,(3*std_dev + moving_average), postcode,
> transaction_dt,score,member_id,member_joining_dt, card_purchase_dt,country,city FROM
> TXN_FINAL_STG;
Query ID = ec2-user_20190827164444_abd2fd0c-26ef-46fa-9e83-f0d2d2469a57
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1566905077309_0024, Tracking URL = http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1566905077309_0024/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill job_1566905077309_0024
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0
2019-08-27 16:44:22,795 Stage-0 map = 0%, reduce = 0%
2019-08-27 16:44:31,455 Stage-0 map = 100%, reduce = 0%, Cumulative CPU 6.41 sec
MapReduce Total cumulative CPU time: 6 seconds 410 msec
Ended Job = job_1566905077309_0024
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1 Cumulative CPU: 6.41 sec HDFS Read: 272073 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 410 msec
OK
Time taken: 26.846 seconds
hive> select * from lookup_card_member limit 5;
OK
340028465709212 1.5402996329018837E7 24658 1514690735 233 009250698176266 2012-02-08 06:04:13 05/13 Unite
d States Barberton
340054675199675 1.3517787096572615E7 50140 1514749403 631 835873341185231 2017-03-10 09:24:44 03/17 Unite
d States Fort Dodge
340082915339645 1.1997828024087135E7 17844 1514747027 407 512969555857346 2014-02-15 06:30:30 07/14 Unite
d States Graham
340134186926007 1.4690957026155703E7 67576 1514761970 614 887711945571282 2012-02-05 01:21:58 02/13 Unite
d States Dix Hills
340265728490548 1.4337361336768052E7 72435 1514686055 202 680324265406190 2014-03-29 07:49:14 11/14 Unite
d States Rancho Cucamonga
Time taken: 0.282 seconds, Fetched: 5 row(s)
hive>
```

Job Scheduler Output

```
[ec2-user@ip-172-31-93-76 app]$ ./main.sh
truncating lookup tables.....
Java HotSpot(TM) 64-Bit Server VM warning: Using incremental CMS is deprecated and will likely be removed in a future release
19/08/31 21:23:50 INFO Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.0-cdh5.15.1, rUnknown, Thu Aug 9 09:07:41 PDT 2018

truncate 'lookup_txn_master'
Truncating 'lookup_txn_master' table (it may take a while):
- Disabling table...
- Truncating table...
0 row(s) in 4.6640 seconds

Java HotSpot(TM) 64-Bit Server VM warning: Using incremental CMS is deprecated and will likely be removed in a future release
19/08/31 21:23:00 INFO Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.0-cdh5.15.1, rUnknown, Thu Aug 9 09:07:41 PDT 2018

truncate 'lookup_card_member'
Truncating 'lookup_card_member' table (it may take a while):
- Disabling table...
- Truncating table...
0 row(s) in 4.9100 seconds

injecting incremental credit card data.....
Warning: /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/bin/./lib/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
19/08/31 21:23:10 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
Enter password:
19/08/31 21:23:34 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
19/08/31 21:23:34 INFO tool.CodeGenTool: Beginning code generation
19/08/31 21:23:35 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
19/08/31 21:23:35 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
19/08/31 21:23:35 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/44b3e822c22063e99a771352011a58b7/card_member.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
19/08/31 21:23:36 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/44b3e822c22063e99a771352011a58b7
/card_member.jar
19/08/31 21:23:37 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX(`member_joining_dt`) FR
OM `card_member`
19/08/31 21:23:37 INFO tool.ImportTool: Incremental import based on column `member_joining_dt`
19/08/31 21:23:37 INFO tool.ImportTool: Lower bound value: '2010-02-02 00:00:00'
19/08/31 21:23:37 INFO tool.ImportTool: Upper bound value: '2018-02-01 08:02:34.0'
```



```

19/08/31 21:23:37 WARN manager.MySQLManager: It looks like you are importing from mysql.
19/08/31 21:23:37 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
19/08/31 21:23:37 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
19/08/31 21:23:37 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
19/08/31 21:23:37 INFO mapreduce.ImportJobBase: Beginning import of card_member
19/08/31 21:23:37 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/08/31 21:23:37 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
19/08/31 21:23:37 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-93-76.ec2.internal/172.31.93.76:8032
19/08/31 21:23:45 INFO db.DBInputFormat: Using read committed transaction isolation
19/08/31 21:23:45 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`card_id`), MAX(`card_id`) FROM `card_member`
`WHERE ( `member_joining_dt` > '2010-02-02 00:00:00' AND `member_joining_dt` <= '2018-02-01 08:02:34.0' )
19/08/31 21:23:45 WARN db.TextSplitter: Generating splits for a textual index column.
19/08/31 21:23:45 WARN db.TextSplitter: If your database sorts in a case-insensitive order, this may result in a partial impo
rt or duplicate records.
19/08/31 21:23:45 WARN db.TextSplitter: You are strongly encouraged to choose an integral split column.
19/08/31 21:23:45 INFO mapreduce.JobSubmitter: number of splits:6
19/08/31 21:23:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1567284753166_0021
19/08/31 21:23:47 INFO impl.YarnClientImpl: Submitted application application_1567284753166_0021
19/08/31 21:23:47 INFO mapreduce.Job: The url to track the job: http://ip-172-31-93-76.ec2.internal:8088/proxy/application_15
67284753166_0021/
19/08/31 21:23:47 INFO mapreduce.Job: Running job: job_1567284753166_0021
19/08/31 21:23:58 INFO mapreduce.Job: Job job_1567284753166_0021 running in uber mode : false
19/08/31 21:23:58 INFO mapreduce.Job: map 0% reduce 0%
19/08/31 21:24:05 INFO mapreduce.Job: map 17% reduce 0%
19/08/31 21:24:10 INFO mapreduce.Job: map 33% reduce 0%
19/08/31 21:24:14 INFO mapreduce.Job: map 50% reduce 0%
19/08/31 21:24:18 INFO mapreduce.Job: map 67% reduce 0%
19/08/31 21:24:23 INFO mapreduce.Job: map 83% reduce 0%
19/08/31 21:24:27 INFO mapreduce.Job: map 100% reduce 0%
19/08/31 21:24:27 INFO mapreduce.Job: Job job_1567284753166_0021 completed successfully
19/08/31 21:24:27 INFO mapreduce.Job: Counters: 30
File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=1073352
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=832
    HDFS: Number of bytes written=85082
    HDFS: Number of read operations=24
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=12
Job Counters
    Launched map tasks=6
    Other local map tasks=6
    Other local map tasks=6
    Total time spent by all maps in occupied slots (ms)=21046
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=21046
    Total vcore-milliseconds taken by all map tasks=21046
    Total megabyte-milliseconds taken by all map tasks=21551104
Map-Reduce Framework
    Map input records=999
    Map output records=999
    Input split bytes=832
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=377
    CPU time spent (ms)=7890
    Physical memory (bytes) snapshot=1330446336
    Virtual memory (bytes) snapshot=16797433856
    Total committed heap usage (bytes)=1354235904
File Input Format Counters
    Bytes Read=0
File Output Format Counters
    Bytes Written=85082
19/08/31 21:24:27 INFO mapreduce.ImportJobBase: Transferred 83.0879 KB in 49.3887 seconds (1.6823 KB/sec)
19/08/31 21:24:27 INFO mapreduce.ImportJobBase: Retrieved 999 records.
19/08/31 21:24:27 INFO util.AppendUtils: Appending to directory card_member
19/08/31 21:24:27 INFO util.AppendUtils: Using found partition 54
19/08/31 21:24:27 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `card_member` AS t LIMIT 1
19/08/31 21:24:27 WARN hive.TableDefWriter: Column member_joining_dt had to be cast to a less precise type in Hive
19/08/31 21:24:27 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/hive-common-1.1.0-
cdh5.15.1.jar!/hive-log4j.properties
FAILED: SemanticException org.apache.hadoop.hive.ql.metadata.HiveException: java.lang.RuntimeException: Unable to instantiate
org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient
Warning: /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/bin/../lib/sqoop/./accumulo does not exist! Accumulo imports will
fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
19/08/31 21:24:32 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
Enter password:
19/08/31 21:24:47 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
19/08/31 21:24:47 INFO tool.CodeGenTool: Beginning code generation
19/08/31 21:24:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
19/08/31 21:24:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
19/08/31 21:24:47 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/9246d1e5d006b603053479c4564acf40/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
19/08/31 21:24:49 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/9246d1e5d006b603053479c4564acf40

```

```

19/08/31 21:24:47 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-ec2-user/compile/9246d1e5d006b603053479c4564acf40/member_score.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
19/08/31 21:24:49 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ec2-user/compile/9246d1e5d006b603053479c4564acf40
/member_score.jar
19/08/31 21:24:50 INFO tool.ImportTool: Destination directory /input/data/tables/member_score is not present, hence not delet
ing.
19/08/31 21:24:50 WARN manager.MySQLManager: It looks like you are importing from mysql.
19/08/31 21:24:50 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
19/08/31 21:24:50 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
19/08/31 21:24:50 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
19/08/31 21:24:50 INFO mapreduce.ImportJobBase: Beginning import of member_score
19/08/31 21:24:50 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/08/31 21:24:50 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
19/08/31 21:24:51 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-93-76.ec2.internal/172.31.93.76:8032
19/08/31 21:24:58 INFO db.DBInputFormat: Using read committed transaction isolation
19/08/31 21:24:58 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`member_id`), MAX(`member_id`) FROM `member_
score`
19/08/31 21:24:58 WARN db.TextSplitter: Generating splits for a textual index column.
19/08/31 21:24:58 WARN db.TextSplitter: If your database sorts in a case-insensitive order, this may result in a partial impo
rt or duplicate records.
19/08/31 21:24:58 WARN db.TextSplitter: You are strongly encouraged to choose an integral split column.
19/08/31 21:24:58 INFO mapreduce.JobSubmitter: number of splits:6
19/08/31 21:24:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1567284753166_0024
19/08/31 21:24:59 INFO impl.YarnClientImpl: Submitted application application_1567284753166_0024
19/08/31 21:24:59 INFO mapreduce.Job: The url to track the job: http://ip-172-31-93-76.ec2.internal:8088/proxy/application_15
67284753166_0024/
19/08/31 21:24:59 INFO mapreduce.Job: Running job: job_1567284753166_0024
19/08/31 21:25:29 INFO mapreduce.Job: Job job_1567284753166_0024 running in uber mode : false
19/08/31 21:25:29 INFO mapreduce.Job: map 0% reduce 0%
19/08/31 21:25:38 INFO mapreduce.Job: map 33% reduce 0%
19/08/31 21:25:39 INFO mapreduce.Job: map 50% reduce 0%
19/08/31 21:25:46 INFO mapreduce.Job: map 67% reduce 0%
19/08/31 21:25:47 INFO mapreduce.Job: map 100% reduce 0%
19/08/31 21:25:47 INFO mapreduce.Job: Job job_1567284753166_0024 completed successfully
19/08/31 21:25:47 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=1066554
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=879
    HDFS: Number of bytes written=19980
    HDFS: Number of read operations=24
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=12
    HDFS: Number of write operations=12
  Job Counters
    Launched map tasks=6
    Other local map tasks=6
    Total time spent by all maps in occupied slots (ms)=42480
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=42480
    Total vcore-milliseconds taken by all map tasks=42480
    Total megabyte-milliseconds taken by all map tasks=43499520
  Map-Reduce Framework
    Map input records=999
    Map output records=999
    Input split bytes=879
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=737
    CPU time spent (ms)=9120
    Physical memory (bytes) snapshot=1412370432
    Virtual memory (bytes) snapshot=16795832320
    Total committed heap usage (bytes)=1387790336
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=19980
19/08/31 21:25:47 INFO mapreduce.ImportJobBase: Transferred 19.5117 KB in 56.9336 seconds (350.9352 bytes/sec)
19/08/31 21:25:47 INFO mapreduce.ImportJobBase: Retrieved 999 records.
19/08/31 21:25:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `member_score` AS t LIMIT 1
19/08/31 21:25:47 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/hive-common-1.1.0-
cdh5.15.1.jar!/hive-log4j.properties
FAILED: SemanticException org.apache.hadoop.hive.ql.metadata.HiveException: java.lang.RuntimeException: Unable to instantiate
org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient
processing.....
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
Java HotSpot(TM) 64-Bit Server VM warning: Using incremental CMS is deprecated and will likely be removed in a future release
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/hive-common-1.1.0-
cdh5.15.1.jar!/hive-log4j.properties
OK
Time taken: 2.407 seconds
Query ID = ec2-user_20190831212525_07a5c3a1-89ae-4435-ba67-e718a900afc3
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1

```

Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
 set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
 set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
 set mapreduce.job.reduces=<number>
Starting Job = job_1567284753166_0025, Tracking URL = http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1567284753166_0025/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill job_1567284753166_0025
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 1
2019-08-31 21:26:14,628 Stage-0 map = 0%, reduce = 0%
2019-08-31 21:26:24,201 Stage-0 map = 100%, reduce = 0%, Cumulative CPU 9.58 sec
2019-08-31 21:26:32,586 Stage-0 map = 100%, reduce = 100%, Cumulative CPU 19.18 sec
MapReduce Total cumulative CPU time: 19 seconds 180 msec
Ended Job = job_1567284753166_0025
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1 Reduce: 1 Cumulative CPU: 19.18 sec HDFS Read: 30739 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 180 msec
OK
Time taken: 35.695 seconds
OK
Time taken: 0.086 seconds
Query ID = ec2-user_20190831212626_375efb8c-d778-4ca8-b3b2-563ec7331fd2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
 set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
 set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
 set mapreduce.job.reduces=<number>
Starting Job = job_1567284753166_0026, Tracking URL = http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1567284753166_0026/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill job_1567284753166_0026
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-31 21:26:56,754 Stage-1 map = 0%, reduce = 0%
2019-08-31 21:27:04,046 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.15 sec
2019-08-31 21:27:10,310 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 10.07 sec
MapReduce Total cumulative CPU time: 10 seconds 70 msec
Ended Job = job_1567284753166_0026
Loading data to table default.txn_stats_stg
Table default.txn_stats_stg stats: [numFiles=1, numRows=999, totalSize=45180, rawDataSize=44181]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 10.07 sec HDFS Read: 11624 HDFS Write: 45261 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 70 msec

Total MapReduce CPU Time Spent: 10 seconds 70 msec
OK
Time taken: 37.925 seconds
OK
Time taken: 0.158 seconds
Query ID = ec2-user_20190831212727_0066c628-56eb-4fbd-b2a5-be6087732838
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1567284753166_0027, Tracking URL = http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1567284753166_0027/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill job_1567284753166_0027
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2019-08-31 21:27:26,010 Stage-1 map = 0%, reduce = 0%
2019-08-31 21:27:36,419 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 7.35 sec
MapReduce Total cumulative CPU time: 7 seconds 350 msec
Ended Job = job_1567284753166_0027
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://ip-172-31-93-76.ec2.internal:8020/user/hive/warehouse/txn_last_detail_stg/.hive-staging_hive_2019-08-31_21-27-12_831_4613083490417587649-1/-ext-10000
Loading data to table default.txn_last_detail_stg
Table default.txn_last_detail_stg stats: [numFiles=1, numRows=999, totalSize=33693, rawDataSize=32694]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 7.35 sec HDFS Read: 6591 HDFS Write: 33780 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 350 msec
OK
Time taken: 24.857 seconds
OK
Time taken: 0.795 seconds
Query ID = ec2-user_20190831212727_80addf09-e22b-46bd-95c1-c4e3f773398e
Total jobs = 1
Execution log at: /tmp/ec2-user/ec2-user_20190831212727_80addf09-e22b-46bd-95c1-c4e3f773398e.log
2019-08-31 09:27:42 Starting to launch local task to process map join; maximum memory = 1908932608
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1567284753166_0028, Tracking URL = http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1567284753166_0028/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill job_1567284753166_0028
Hadoop job information for Stage-7: number of mappers: 1; number of reducers: 0
2019-08-31 21:27:58,529 Stage-7 map = 0%, reduce = 0%
2019-08-31 21:28:07,952 Stage-7 map = 100%, reduce = 0%, Cumulative CPU 6.98 sec
MapReduce Total cumulative CPU time: 6 seconds 980 msec
Ended Job = job_1567284753166_0028

Ended Job = job_1567284753166_0028
Loading data to table default.txn_final_stg
Table default.txn_final_stg stats: [numFiles=1, numRows=60939, totalSize=8084513, rawDataSize=8023574]
MapReduce Jobs Launched:
Stage-Stage-7: Map: 1 Cumulative CPU: 6.98 sec HDFS Read: 56928 HDFS Write: 8084598 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 980 msec
OK
Time taken: 30.779 seconds
Query ID = ec2-user_20190831212828_70755a4e-a77d-49eb-ade5-76e018a6f9cf
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1567284753166_0029, Tracking URL = http://ip-172-31-93-76.ec2.internal:8088/proxy/application_1567284753166_0029/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job -kill job_1567284753166_0029
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0
2019-08-31 21:28:23,760 Stage-0 map = 0%, reduce = 0%
2019-08-31 21:28:35,261 Stage-0 map = 100%, reduce = 0%, Cumulative CPU 12.14 sec
MapReduce Total cumulative CPU time: 12 seconds 140 msec
Ended Job = job_1567284753166_0029
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1 Cumulative CPU: 12.14 sec HDFS Read: 8091440 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 140 msec
OK
Time taken: 27.043 seconds
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
■