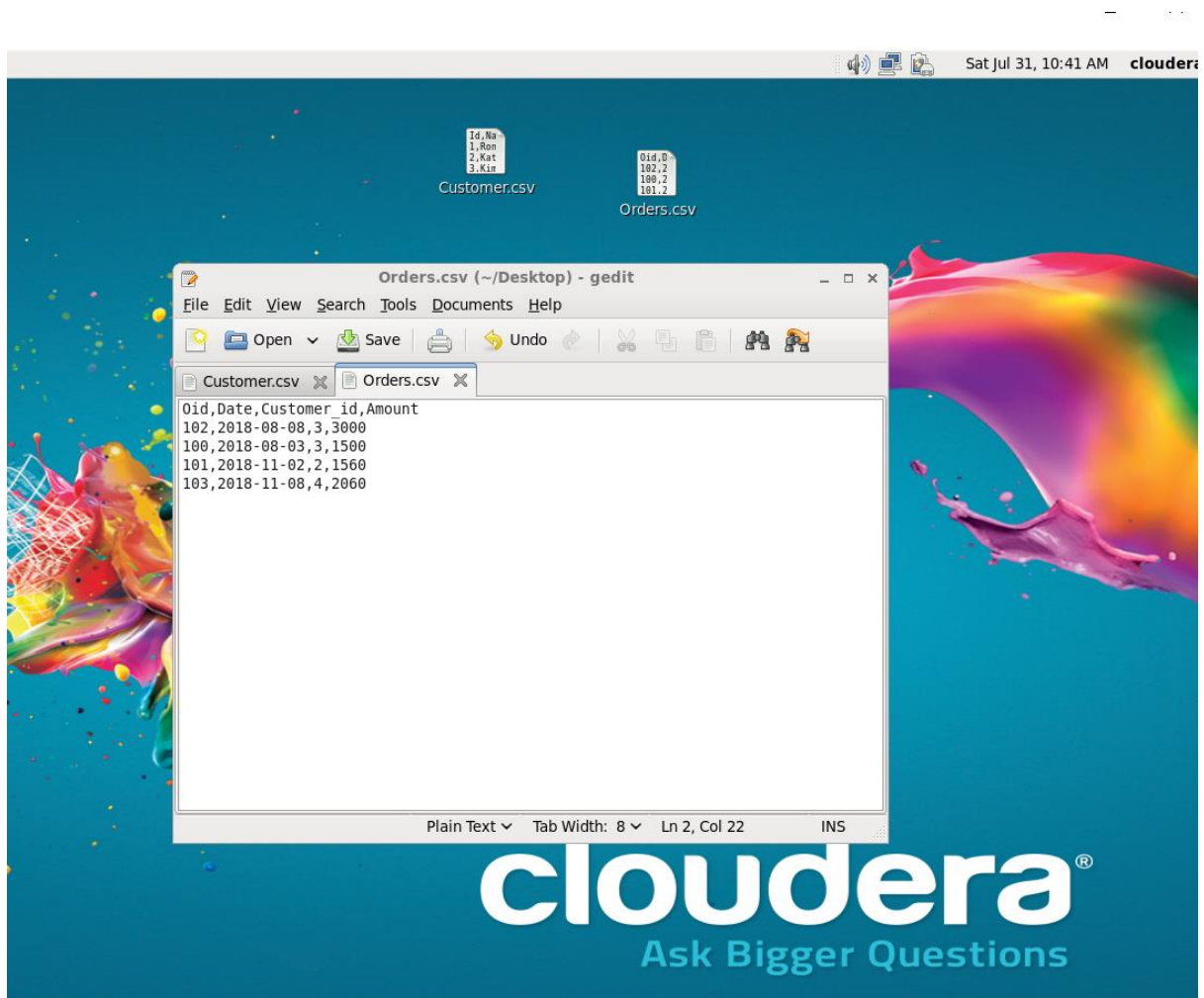


Experiment – 09

Joins Subqueries using HiveQL



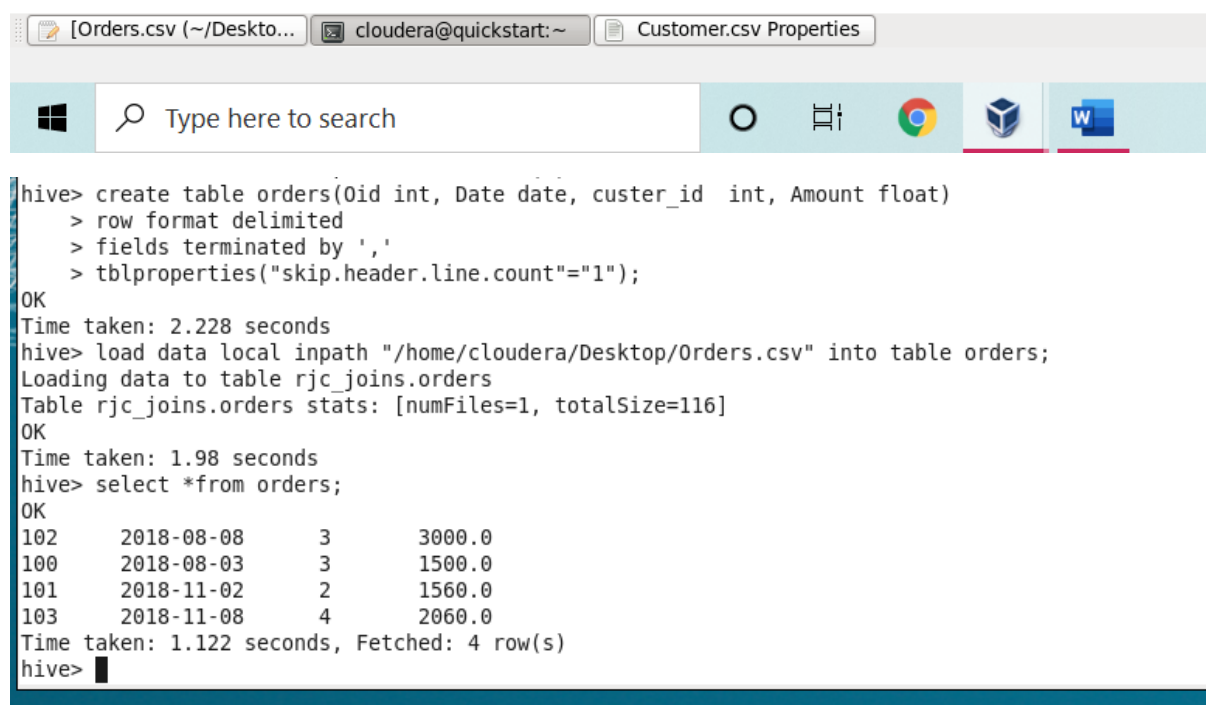
```
cloudera@quickstart ~$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> Create database rjc_joins;
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. Database rjc_joins already exists
hive> use rjc_joins;
OK
Time taken: 0.392 seconds
hive> show databases;
OK
default
hiveql
rjc
rjc_joins
rjcstudent
Time taken: 0.664 seconds, Fetched: 5 row(s)
```

```

hive> create table customers(ID int, Name string, Age int, Address string, Salary float)
> row format delimited
> fields terminated by ','
> tblproperties("skip.header.line.count"="1");
OK
Time taken: 7.503 seconds
hive> load data local inpath "/home/cloudera/Desktop/Customer.csv" into table customers;
Loading data to table rjc_joins.customers
Table rjc_joins.customers stats: [numFiles=1, totalSize=193]
OK
Time taken: 1.258 seconds
hive> select *from customers;
OK
1      Rony    32      New York    2000.0
2      Kate    25      Florida    1500.0
3      Kim     23      Seattle    2000.0
4      Clay    25      Boston     6500.0
5      Henry   27      California  8500.0
6      Kit     22      Chicago    4500.0
7      Muffy   24      New York    10000.0
Time taken: 0.329 seconds, Fetched: 7 row(s)
hive> █

```



The screenshot shows a Windows desktop environment. At the top, there is a taskbar with several open applications: 'Orders.csv (~/Desko...', 'cloudera@quickstart:~', and 'Customer.csv Properties'. Below the taskbar is a search bar with the text 'Type here to search'. To the right of the search bar are several icons: a power button, a task view icon, a Google Chrome icon, a Docker icon, and a Microsoft Word icon. Below the taskbar, a terminal window is open, displaying Hive commands and their output. The terminal shows the creation of a table named 'orders', loading data from a local file, and querying the table. The output of the query shows 4 rows of data.

```

hive> create table orders(Oid int, Date date, custer_id int, Amount float)
> row format delimited
> fields terminated by ','
> tblproperties("skip.header.line.count"="1");
OK
Time taken: 2.228 seconds
hive> load data local inpath "/home/cloudera/Desktop/Orders.csv" into table orders;
Loading data to table rjc_joins.orders
Table rjc_joins.orders stats: [numFiles=1, totalSize=116]
OK
Time taken: 1.98 seconds
hive> select *from orders;
OK
102    2018-08-08    3      3000.0
100    2018-08-03    3      1500.0
101    2018-11-02    2      1560.0
103    2018-11-08    4      2060.0
Time taken: 1.122 seconds, Fetched: 4 row(s)
hive> █

```

```

hive> select c.id, c.name, c.age, o.amount
> from customers c JOIN orders o
> on (c.id = o.custer id);
Query ID = cloudera_20210731111515_8a4772f2-f095-4e31-8fa5-764db729fdaa
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20210731111515_8a4772f2-f095-4e31-8fa5-764db729fdaa.log
2021-07-31 11:18:29 Starting to launch local task to process map join; maximum memory = 1013645312
2021-07-31 11:18:41 Dump the side-table for tag: 1 with group count: 3 into file: file:/tmp/cloudera/7f04c3a9-5481-40fa-b19c-7033ef01fc48/hive_2
021-07-31_11-15-06_193_3000295498216232773-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable
2021-07-31 11:18:42 Uploaded 1 File to: file:/tmp/cloudera/7f04c3a9-5481-40fa-b19c-7033ef01fc48/hive_2021-07-31_11-15-06_193_3000295498216232773
-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (338 bytes)
2021-07-31 11:18:42 End of local task; Time Taken: 13.262 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_1627751786351_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1627751786351_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1627751786351_0001
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2021-07-31 11:20:22,585 Stage-3 map = 0%, reduce = 0%
2021-07-31 11:21:23,429 Stage-3 map = 0%, reduce = 0%
2021-07-31 11:21:31,582 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 4.76 sec
MapReduce Total cumulative CPU time: 4 seconds 760 msec
Ended Job = job_1627751786351_0001
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 4.76 sec HDFS Read: 6624 HDFS Write: 66 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 760 msec
OK
2 Kate 25 1560.0
3 Kim 23 3000.0
3 Kim 23 1500.0
4 Clay 25 2060.0
Time taken: 390.645 seconds, Fetched: 4 row(s)
hive>

```

```

hive> select c.id, c.name, c.age, o.amount,o.date
> from customers c LEFT OUTER JOIN orders o
> on (c.id = o.custer id);
Query ID = cloudera_20210731112626_8bf2bccf-a331-40a3-8956-656ce17f43c3
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20210731112626_8bf2bccf-a331-40a3-8956-656ce17f43c3.log
2021-07-31 11:27:01 Starting to launch local task to process map join; maximum memory = 1013645312
2021-07-31 11:27:06 Dump the side-table for tag: 1 with group count: 3 into file: file:/tmp/cloudera/7f04c3a9-5481-40fa-b19c-
le11--.hashtable
2021-07-31 11:27:06 Uploaded 1 File to: file:/tmp/cloudera/7f04c3a9-5481-40fa-b19c-7033ef01fc48/hive_2021-07-31_11-26-30_145
2021-07-31 11:27:06 End of local task; Time Taken: 5.354 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_1627751786351_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1627751786351_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1627751786351_0002
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2021-07-31 11:27:46,913 Stage-3 map = 0%, reduce = 0%
2021-07-31 11:28:21,035 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 2.47 sec
MapReduce Total cumulative CPU time: 2 seconds 470 msec
Ended Job = job_1627751786351_0002
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 2.47 sec HDFS Read: 6640 HDFS Write: 175 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 470 msec
OK
1 Rony 32 NULL NULL
2 Kate 25 1560.0 2018-11-02
3 Kim 23 3000.0 2018-08-08
3 Kim 23 1500.0 2018-08-03
4 Clay 25 2060.0 2018-11-08
5 Henry 27 NULL NULL
6 Kit 22 NULL NULL
7 Muffy 24 NULL NULL
Time taken: 114.311 seconds, Fetched: 8 row(s)
hive>

```

```

hive> select c.id, c.name, c.age, o.amount, o.date
> from customers c RIGHT OUTER JOIN orders o
> on (c.id = o.custer_id);
Query ID = cloudera_20210731113131_756a6c1f-8414-45c4-b806-70891fd49fc8
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20210731113131_756a6c1f-8414-45c4-b806-70891fd49fc8.log
2021-07-31 11:32:09 Starting to launch local task to process map join; maximum memory = 1013645312
2021-07-31 11:32:14 Dump the side-table for tag: 0 with group count: 7 into file: file:/tmp/cloudera/7f04c3e
le20--.hashtable
2021-07-31 11:32:15 Uploaded 1 File to: file:/tmp/cloudera/7f04c3a9-5481-40fa-b19c-7033ef01fc48/hive_2021-07
2021-07-31 11:32:15 End of local task; Time Taken: 5.405 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_1627751786351_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_16277517
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1627751786351_0003
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2021-07-31 11:33:05,950 Stage-3 map = 0%, reduce = 0%
2021-07-31 11:33:58,175 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 3.13 sec
MapReduce Total cumulative CPU time: 3 seconds 130 msec
Ended Job = job_1627751786351_0003
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 3.13 sec HDFS Read: 6523 HDFS Write: 110 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 130 msec
OK
3 Kim 23 3000.0 2018-08-08
3 Kim 23 1500.0 2018-08-03
2 Kate 25 1560.0 2018-11-02
4 Clay 25 2060.0 2018-11-08
Time taken: 143.923 seconds, Fetched: 4 row(s)
hive>

```

```

*Orders.csv (~/Desktop...) cloudera@quickstart:~
hive> select max(salary) from customers where customers.salary not in (select max(salary) from customers);
Warning: Map Join MAPJOIN[131][bigTable=customers] in task 'Stage-8:MAPRED' is a cross product
Warning: Shuffle Join JOIN[24][tables = [customers, sq_1notin_nullcheck]] in Stage 'Stage-1:MAPRED' is a cross product
Query ID = cloudera_20210731113838_dfbdcfda-d9f9-403d-b715-006385454bb9
Total jobs = 7
Launching Job 1 out of 7
Number of reduce tasks determined at compile time: 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_1627751786351_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1627751786351_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1627751786351_0004
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2021-07-31 11:38:48,792 Stage-4 map = 0%, reduce = 0%
2021-07-31 11:39:27,350 Stage-4 map = 100%, reduce = 0%, Cumulative CPU 2.75 sec
2021-07-31 11:39:58,191 Stage-4 map = 100%, reduce = 100%, Cumulative CPU 5.93 sec
MapReduce Total cumulative CPU time: 5 seconds 930 msec
Ended Job = job_1627751786351_0004

MapReduce Jobs Launched:
Stage-Stage-4: Map: 1 Reduce: 1 Cumulative CPU: 5.93 sec HDFS Read: 6383 HDFS Write: 117 SUCCESS
Stage-Stage-5: Map: 1 Reduce: 1 Cumulative CPU: 7.14 sec HDFS Read: 7483 HDFS Write: 114 SUCCESS
Stage-Stage-6: Map: 1 Cumulative CPU: 2.19 sec HDFS Read: 5392 HDFS Write: 243 SUCCESS
Stage-Stage-6: Map: 1 Cumulative CPU: 3.52 sec HDFS Read: 5305 HDFS Write: 117 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 5.01 sec HDFS Read: 4649 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 23 seconds 790 msec
OK
8500.0
Time taken: 448.106 seconds, Fetched: 1 row(s)
hive>

```

```

*Orders.csv (~/Desktop...) cloudera@quickstart:~

```

```

hive> select salary from customers sort by salary desc limit 4;
Query ID = cloudera_20210731114848_52ale783-1484-473c-8c61-5b4d31fc6bd8
Total jobs = 2
Launching Job 1 out of 2
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_1627751786351_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1627751786351_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1627751786351_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-31 11:49:03,085 Stage-1 map = 0%, reduce = 0%
2021-07-31 11:49:29,496 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.22 sec
2021-07-31 11:49:58,988 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.22 sec
MapReduce Total cumulative CPU time: 2 seconds 220 msec
Ended Job = job_1627751786351_0009
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 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_1627751786351_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1627751786351_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1627751786351_0010
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-07-31 11:50:26,911 Stage-2 map = 0%, reduce = 0%
2021-07-31 11:50:48,242 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.98 sec
2021-07-31 11:51:09,751 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.72 sec
MapReduce Total cumulative CPU time: 4 seconds 720 msec
Ended Job = job_1627751786351_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.65 sec HDFS Read: 5396 HDFS Write: 180 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 4.72 sec HDFS Read: 4388 HDFS Write: 29 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 370 msec
OK
10000.0
8500.0
6500.0
4500.0
Time taken: 152.386 seconds, Fetched: 4 row(s)
hive>

```

*Orders.csv (~/.Desкто... cloudera@quickstart:~

```

Time taken: 228.500 seconds, Fetched: 1 row(s)
hive> select salary from (select salary from customers sort by salary desc limit 4) result sort by salary asc limit 1;
Query ID = cloudera_20210731115454_bdc07047-0228-4a7d-b807-7c631eec34dd
Total jobs = 4
Launching Job 1 out of 4
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_1627751786351_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1627751786351_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1627751786351_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-31 11:54:59,120 Stage-1 map = 0%, reduce = 0%
2021-07-31 11:55:23,580 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.34 sec
2021-07-31 11:55:56,055 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.37 sec
MapReduce Total cumulative CPU time: 5 seconds 370 msec
Ended Job = job_1627751786351_0011
Launching Job 2 out of 4

Ended Job = job_1627751786351_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.37 sec HDFS Read: 5413 HDFS Write: 180 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 5.15 sec HDFS Read: 3837 HDFS Write: 180 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 4.7 sec HDFS Read: 3849 HDFS Write: 117 SUCCESS
Stage-Stage-4: Map: 1 Reduce: 1 Cumulative CPU: 5.87 sec HDFS Read: 4330 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 21 seconds 90 msec
OK
4500.0
Time taken: 298.257 seconds, Fetched: 1 row(s)
hive>

```