Problem Statement 1

1.1. Jimmy, from the healthcare department, has requested a report that shows how the number of treatments each age category of patients has gone through in the year 2022.

The age category is as follows, Children (00-14 years), Youth (15-24 years), Adults (25-64 years), and Seniors (65 years and over).

hive> create view p1 as select count(*) as count,e.category from (select (case when DATEDIFF("2022-12-01",p.dob) / 365.25

> <=14 then "children"

> when DATEDIFF("2022-12-01",p.dob) / 365.25 <=24 then "youth"

> when datediff("2022-12-01",p.dob) /365.25 <= 64 then "Adults"

> else "Seniors" end)

> as category from treatment t join patient p on t.patientID=p.patientID

> where year(t.`date`)=2022) e group by e.category;

OK

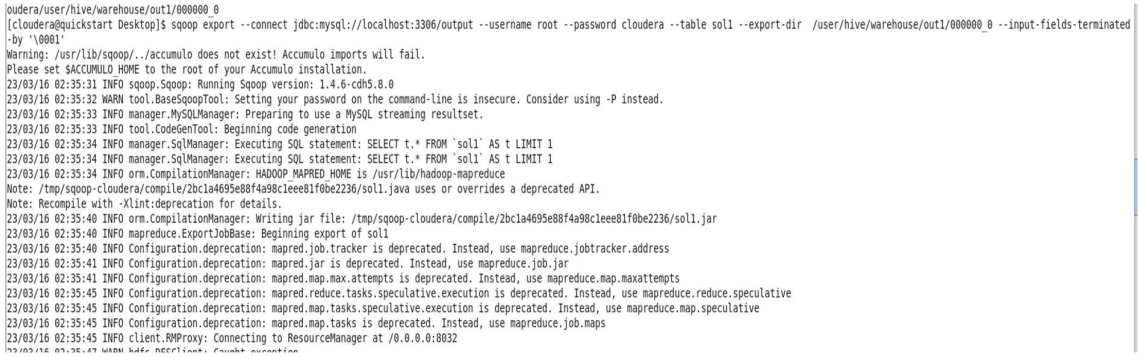
hive> create external table out1(counts int ,category string);

OK

Time taken: 0.049 seconds

hive> insert OVERWRITE table out1 select * from p1;

[training@localhost ~]$ sqoop export --connect jdbc:mysql://localhost:3306/output --username root –password clordera --table sol1 --export-dir /user/hive/warehouse/out1/000000_0 --input-fields-terminated-by '\0001'



Problem Statement : 2

1.2. Jimmy, from the healthcare department, wants to know which disease is infecting people of which gender more often.Assist Jimmy with this purpose by generating a report that shows for each disease the male-to-female ratio. Sort the data in a way that is helpful for Jimmy.

hive> create view p2 as select d.diseasename,p.gender,count(*) as c from person p join treatment t on p.personid=t.patientid join disease d on d.diseaseid=t.diseaseid group by d.diseasename,p.gender;

OK

Time taken: 0.064 seconds

hive> select a.diseasename,a.c as male,b.c as female,(a.c/b.c) as ratio from p2 a join p2 b where a.diseasename=b.diseasename and a.gender='male' and b.gender='female' order by ratio desc desc 10;

Total MapReduce jobs = 8

```
OK
Depression      170     82      2.073170731707317
Multiple sclerosis      173     88      1.9659090909090908
Diabetes mellitus type 1        174     93      1.8709677419354838
Cancer  191     103     1.854368932038835
Anorexia nervosa        177     96      1.84375
Thromboangiitis obliterans      175     96      1.8229166666666667
Alzheimer's disease     173     95      1.8210526315789475
Dementia        162     90      1.8
Diabetes mellitus type 2        178     99      1.797979797979798
Lupus   158     88      1.7954545454545454
Time taken: 87.223 seconds
```

hive> insert OVERWRITE table out2 select a.diseasename,a.c as male,b.c as female,(a.c/b.c) as ratio from p2 a join p2 b where a.diseasename=b.diseasename and a.gender='male' and b.gender='female' order by ratio desc

[training@localhost ~]$ sqoop export --connect jdbc:mysql://localhost:3306/output --username root --table sol2 --export-dir user/hive/warehouse/out2/000000_0 --input-fields-terminated-by '\0001'

```
mysql> select * from sol2 limit 10;
+----------------------+------+--------+--------------+
| dname                | male | female | m_to_f       |
+----------------------+------+--------+--------------+
| Depression           |  170 |     82 | 2.0731707000 |
| Multiple sclerosis   |  173 |     88 | 1.9659091000 |
| Diabetes mellitus ty |  174 |     93 | 1.8709677000 |
| Cancer               |  191 |    103 | 1.8543689000 |
| Anorexia nervosa     |  177 |     96 | 1.8437500000 |
| Thromboangiitis obli |  175 |     96 | 1.8229166000 |
| Alzheimer's disease  |  173 |     95 | 1.8210527000 |
| Dementia             |  162 |     90 | 1.8000000000 |
| Diabetes mellitus ty |  178 |     99 | 1.7979798000 |
| Lupus                |  158 |     88 | 1.7954545000 |
+----------------------+------+--------+--------------+
10 rows in set (0.00 sec)
```

Problem Statement :3

1.3. Jacob, from insurance management, has noticed that insurance claims are not made for all the treatments. He also wants to figure out if the gender of the patient has any impact on the insurance claim. Assist Jacob in this situation by generating a report that finds for each gender the number of treatments, number of claims, and claim-to treatment ratio. And notice if there is a significant difference between the treatment-to-claim ratio of male and female patients.

hive> select p.gender,count(t.treatmentid) as treatments,count(t.claimid) as claims,(count(t.claimid)/count(t.treatmentid)) from treatment t join person p

> on p.personid=t.patientid group by p.gender;

Total MapReduce jobs = 2

```
female  4206    2676    0.6362339514978602
male    6679    4287    0.641862554274592
Time taken: 23.665 seconds
```

hive> create external table out3(gender string,treatments int,claims int,ration float);

OK

Time taken: 0.558 seconds

hive> insert overwrite table out3 select p.gender,count(t.treatmentid) as treatments,count(t.claimid) as claims,(count(t.claimid)/count(t.treatmentid)) from treatment t join person p

> on p.personid=t.patientid group by p.gender;

Total MapReduce jobs = 2

mysql> create table sol3(gender varchar(20),treatments int,claims int,c_to_t numeric(20,10));

Query OK, 0 rows affected (0.00 sec)

[training@localhost ~]$ sqoop export --connect jdbc:mysql://localhost:3306/output --username root --table sol3 --export-dir /user/hive/warehouse/out3/000000_0 --input-fields-terminated-by '\0001'

```
mysql> select * from sol3;
+--------+------------+--------+--------------+
| gender | treatments | claims | c_to_t       |
+--------+------------+--------+--------------+
| female |       4206 |   2676 | 0.6362339000 |
| male   |       6679 |   4287 | 0.6418626000 |
+--------+------------+--------+--------------+
2 rows in set (0.00 sec)
```

Problem Statement :4.

1.5 .The healthcare department suspects that some pharmacies prescribe more medicines than others in a single prescription, for them, generate a report that finds for each pharmacy the maximum, minimum and average number of medicines prescribed in their prescriptions.

hive> create view p5 as select p.pharmacyid,pr.prescriptionid,sum(c.quantity) as c,count(*) as coun from pharmacy p join prescription pr on p.pharmacyid=pr.pharmacyid

> join contain c on c.prescriptionid=pr.prescriptionid group by p.pharmacyid,pr.prescriptionid;

OK

hive> create external table out4(pid int,total int,max_ int,min_ int,avrg float);

OK

Time taken: 1.586 seconds

Insert into External table:

```
hive> insert overwrite table out4 select pharmacyid,sum(s) as total ,max(s),min(s),avg(s) from p5 group by pharmacyid limit 5;
Total MapReduce jobs = 5
Launching Job 1 out of 5
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 mapred.reduce.tasks=<number>
Starting Job = job_202303140233_0024, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202303140233_0024
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202303140233_0024
2023-03-14 03:12:29,965 Stage-6 map = 0%,  reduce = 0%
2023-03-14 03:12:31,988 Stage-6 map = 100%,  reduce = 0%
2023-03-14 03:12:38,042 Stage-6 map = 100%,  reduce = 33%
2023-03-14 03:12:39,051 Stage-6 map = 100%,  reduce = 100%
Ended Job = job_202303140233_0024
Launching Job 2 out of 5
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 mapred.reduce.tasks=<number>
Starting Job = job_202303140233_0025, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202303140233_0025
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202303140233_0025
2023-03-14 03:12:42,294 Stage-1 map = 0%,  reduce = 0%
2023-03-14 03:12:44,305 Stage-1 map = 100%,  reduce = 0%
2023-03-14 03:12:51,336 Stage-1 map = 100%,  reduce = 33%
2023-03-14 03:12:52,343 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202303140233_0025
Launching Job 3 out of 5
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 mapred.reduce.tasks=<number>
Starting Job = job_202303140233_0026, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202303140233_0026
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202303140233_0026
2023-03-14 03:12:53,577 Stage-2 map = 0%,  reduce = 0%
2023-03-14 03:12:55,587 Stage-2 map = 100%,  reduce = 0%
```

SQOOP export:

```
[training@localhost ~]$ sqoop export --connect jdbc:mysql://localhost:3306/output  --username root   --table sol4  --export-dir  /user/hive/warehouse/out4/000000_0 --input-fields-terminated-by '\0001'
23/03/14 03:24:25 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/14 03:24:25 INFO tool.CodeGenTool: Beginning code generation
23/03/14 03:24:26 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `sol4` AS t LIMIT 1
23/03/14 03:24:26 INFO orm.CompilationManager: HADOOP_HOME is /usr/lib/hadoop
23/03/14 03:24:26 INFO orm.CompilationManager: Found hadoop core jar at: /usr/lib/hadoop/hadoop-core.jar
23/03/14 03:24:27 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-training/compile/57d97d791a9b30a2ecd82b6124374348/sol4.java to /home/training/./sol4.java
java.io.IOException: Destination '/home/training/./sol4.java' already exists
        at org.apache.commons.io.FileUtils.moveFile(FileUtils.java:1811)
        at com.cloudera.sqoop.orm.CompilationManager.compile(CompilationManager.java:229)
        at com.cloudera.sqoop.tool.CodeGenTool.generateORM(CodeGenTool.java:85)
        at com.cloudera.sqoop.tool.ExportTool.exportTable(ExportTool.java:66)
        at com.cloudera.sqoop.tool.ExportTool.run(ExportTool.java:99)
        at com.cloudera.sqoop.Sqoop.run(Sqoop.java:146)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:65)
        at com.cloudera.sqoop.Sqoop.runSqoop(Sqoop.java:182)
        at com.cloudera.sqoop.Sqoop.runTool(Sqoop.java:221)
        at com.cloudera.sqoop.Sqoop.runTool(Sqoop.java:230)
        at com.cloudera.sqoop.Sqoop.main(Sqoop.java:239)
23/03/14 03:24:27 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/57d97d791a9b30a2ecd82b6124374348/sol4.jar
23/03/14 03:24:27 INFO mapreduce.ExportJobBase: Beginning export of sol4
23/03/14 03:24:28 INFO input.FileInputFormat: Total input paths to process : 1
23/03/14 03:24:28 INFO input.FileInputFormat: Total input paths to process : 1
23/03/14 03:24:29 INFO mapred.JobClient: Running job: job_202303140233_0034
23/03/14 03:24:30 INFO mapred.JobClient:  map 0% reduce 0%
23/03/14 03:24:34 INFO mapred.JobClient:  map 100% reduce 0%
23/03/14 03:24:35 INFO mapred.JobClient: Job complete: job_202303140233_0034
23/03/14 03:24:35 INFO mapred.JobClient: Counters: 12
23/03/14 03:24:35 INFO mapred.JobClient:   Job Counters
23/03/14 03:24:35 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=4720
23/03/14 03:24:35 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
23/03/14 03:24:35 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
23/03/14 03:24:35 INFO mapred.JobClient:     Launched map tasks=1
23/03/14 03:24:35 INFO mapred.JobClient:     Data-local map tasks=1
23/03/14 03:24:35 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=0
23/03/14 03:24:35 INFO mapred.JobClient:   FileSystemCounters
23/03/14 03:24:35 INFO mapred.JobClient:     HDFS_BYTES_READ=5485
23/03/14 03:24:35 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=65672
23/03/14 03:24:35 INFO mapred.JobClient:   Map-Reduce Framework
23/03/14 03:24:35 INFO mapred.JobClient:     Map input records=213
23/03/14 03:24:35 INFO mapred.JobClient:     Spilled Records=0
23/03/14 03:24:35 INFO mapr training@localhost:~ Map output records=213
```

Exported to client:

```
23/03/14 03:24:35 INFO mapreduce.ExportJobBase: Exported 213 records.
[training@localhost ~]$ mysql -u root
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 12
Server version: 5.0.77 Source distribution

Type 'help;' or '\h' for help. Type '\c' to clear the buffer.

mysql> use output;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from sol4 limit 5;
+------------+-------+---------------+---------------+-----------+
| pharmacyid | total | max_medicines | min_medicines | average   |
+------------+-------+---------------+---------------+-----------+
|       1008 |  2608 |           125 |             6 | 43.466667 |
|       1145 |  2882 |           102 |             1 | 41.768116 |
|       1149 |  2433 |           121 |             8 |  43.44643 |
|       1194 |  1982 |            87 |             1 |  40.44898 |
|       1204 |  2230 |           108 |             8 | 45.510204 |
+------------+-------+---------------+---------------+-----------+
5 rows in set (0.00 sec)
```

Problem Statement: 5.

2.2 The State of Alabama (AL) is trying to manage its healthcare resources more efficiently. For each city in their state, they need to identify the disease for which the maximum number of patients have gone for treatment. Assist the state for this purpose.

Note: The state of Alabama is represented as AL in Address Table.

```
hive> CREATE TABLE IF NOT EXISTS address_PART (addressid int, address1 String, city String,zip int)
   > COMMENT 'address_PART details'
   > PARTITIONED BY (state String)
   > ROW FORMAT DELIMITED
   > FIELDS TERMINATED BY ','
   > LINES TERMINATED BY '\n'
   > STORED AS TEXTFILE;
OK
Time taken: 1.803 seconds
```

hive> create view p6 as select a.city,d.diseasename,count(t.patientid) as coun from address_part a join person p on p.addressid=a.addressid join treatment t on t.patientID=p.personid join disease d on t.diseaseid=d.diseaseid where a.state='AL' group by a.city,d.diseasename;

hive> insert overwrite table out5 select a.c,a.d,a.co from (select city as c,diseasename as d,coun as co,ROW_NUMBER() OVER(partition by city order by coun desc) as rn from p6) as a where a.rn=1;

```
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_1678816063243_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1678816063243_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1678816063243_0007
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 1
2023-03-15 00:18:34,137 Stage-5 map = 0%,  reduce = 0%
2023-03-15 00:18:49,755 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 2.06 sec
2023-03-15 00:19:08,796 Stage-5 map = 100%,  reduce = 100%, Cumulative CPU 6.46 sec
MapReduce Total cumulative CPU time: 6 seconds 460 msec
Ended Job = job_1678816063243_0007
Loading data to table default.out5
Table default.out5 stats: [numFiles=1, numRows=3, totalSize=116, rawDataSize=113]
MapReduce Jobs Launched:
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 7.91 sec   HDFS Read: 137026 HDFS Write: 2700 SUCCESS
Stage-Stage-5: Map: 1  Reduce: 1   Cumulative CPU: 6.46 sec   HDFS Read: 10426 HDFS Write: 185 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 370 msec
OK
Time taken: 137.703 seconds
```

mysql> create table sol5(city varchar(50),diseasename varchar(50),coun numeric(10));

Query OK, 0 rows affected (0.00 sec)

SQOOOP export:

[cloudera@quickstart Desktop]$ sqoop export --connect jdbc:mysql://localhost:3306/output --username root --password cloudera --table sol5 --export-dir /user/hive/warehouse/out5 --input-fields-terminated-by '\0001'

```
23/03/14 08:30:55 INFO mapreduce.JobSubmitter: number of splits:4
23/03/14 08:30:55 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/14 08:30:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1678798225978_0026
23/03/14 08:30:55 INFO impl.YarnClientImpl: Submitted application application_1678798225978_0026
23/03/14 08:30:55 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1678798225978_0026/
23/03/14 08:30:55 INFO mapreduce.Job: Running job: job_1678798225978_0026
23/03/14 08:31:04 INFO mapreduce.Job: Job job_1678798225978_0026 running in uber mode : false
23/03/14 08:31:04 INFO mapreduce.Job:  map 0% reduce 0%
23/03/14 08:31:25 INFO mapreduce.Job:  map 25% reduce 0%
23/03/14 08:31:29 INFO mapreduce.Job:  map 50% reduce 0%
23/03/14 08:31:30 INFO mapreduce.Job:  map 100% reduce 0%
23/03/14 08:31:31 INFO mapreduce.Job: Job job_1678798225978_0026 completed successfully
23/03/14 08:31:31 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=566056
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=886
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=4
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=84953
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=84953
                Total vcore-seconds taken by all map tasks=84953
                Total megabyte-seconds taken by all map tasks=86991872
        Map-Reduce Framework
                Map input records=3
                Map output records=3
                Input split bytes=584
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1373
                CPU time spent (ms)=3350
                Physical memory (bytes) snapshot=499470336
                Virtual memory (bytes) snapshot=6015205376
                Total committed heap usage (bytes)=243531776
```

Output in the client database:

```
mysql> use output;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from sol5;
+-------------------------+---------------------------+------+
| city                    | diseasename               | coun |
+-------------------------+---------------------------+------+
| Indian Springs Village | Diabetes mellitus type 2 |    1 |
| Montevallo              | Schizophrenia             |    2 |
| Montgomery              | Guillain?Barré syndrome  |   28 |
+-------------------------+---------------------------+------+
3 rows in set (0.00 sec)
```

Problem Statement :6.

3.1. Some complaints have been lodged by patients that they have been prescribed hospital-exclusive medicine that they can't find elsewhere and facing problems due to that. Joshua, from the pharmacy management, wants to get a report of which pharmacies have prescribed hospital-exclusive medicines the most in the years 2021 and 2022. Assist Joshua to generate the report so that the pharmacies who prescribe hospital-exclusive medicine more often are advised to avoid such practice if possible.

hive> select ph.pharmacyid,count(c.medicineid) as coun from treatment t join Prescription ph on t.treatmentid=ph.treatmentid

> join contain c on c.prescriptionid=ph.prescriptionid join medicine m on

> c.medicineid=m.medicineid where m.hospitalexclusive ='S' and year(t.date) in (2021,2022)

> group by ph.pharmacyid order by coun desc;

Hive query output writing into external table

```
hive> create table out6(pharmacyid int,hexclusive int);
OK
Time taken: 4.146 seconds
hive> insert overwrite table out6 select ph.pharmacyid,count(c.medicineid) as coun from treatment t join Prescription ph on t.treatmentid=ph.treatmentid
    >   join contain c on c.prescriptionid=ph.prescriptionid join medicine m on
    >   c.medicineid=m.medicineid where m.hospitalexclusive ='S' and year(t.date)  in (2021,2022)
    >   group by ph.pharmacyid order by coun desc;
Query ID = cloudera_20230314111212_1d65b829-fdf7-4eba-a71c-424d41083ab0
Total jobs = 2
Execution log at: /tmp/cloudera/cloudera_20230314111212_1d65b829-fdf7-4eba-a71c-424d41083ab0.log
2023-03-14 11:12:41     Starting to launch local task to process map join;      maximum memory = 1013645312
2023-03-14 11:12:47     Dump the side-table for tag: 1 with group count: 8079 into file: file:/tmp/cloudera/52ae09e5-2671-4dcf-8f65-7d5fd32a4550/hive_2023-03-14_11-12-27_554_3511941460856176477-1/-local-1
hTable-Stage-4/MapJoin-mapfile31--.hashtable
2023-03-14 11:12:47     Uploaded 1 File to: file:/tmp/cloudera/52ae09e5-2671-4dcf-8f65-7d5fd32a4550/hive_2023-03-14_11-12-27_554_3511941460856176477-1/-local-10006/HashTable-Stage-4/MapJoin-mapfile31--.ha
(162611 bytes)
2023-03-14 11:12:47     Dump the side-table for tag: 1 with group count: 13205 into file: file:/tmp/cloudera/52ae09e5-2671-4dcf-8f65-7d5fd32a4550/hive_2023-03-14_11-12-27_554_3511941460856176477-1/-local-
shTable-Stage-4/MapJoin-mapfile41--.hashtable
2023-03-14 11:12:48     Uploaded 1 File to: file:/tmp/cloudera/52ae09e5-2671-4dcf-8f65-7d5fd32a4550/hive_2023-03-14_11-12-27_554_3511941460856176477-1/-local-10006/HashTable-Stage-4/MapJoin-mapfile41--.ha
(687493 bytes)
2023-03-14 11:12:48     Dump the side-table for tag: 1 with group count: 10885 into file: file:/tmp/cloudera/52ae09e5-2671-4dcf-8f65-7d5fd32a4550/hive_2023-03-14_11-12-27_554_3511941460856176477-1/-local-
shTable-Stage-4/MapJoin-mapfile51--.hashtable
2023-03-14 11:12:48     Uploaded 1 File to: file:/tmp/cloudera/52ae09e5-2671-4dcf-8f65-7d5fd32a4550/hive_2023-03-14_11-12-27_554_3511941460856176477-1/-local-10006/HashTable-Stage-4/MapJoin-mapfile51--.ha
(370243 bytes)
2023-03-14 11:12:48     End of local task; Time Taken: 6.377 sec.
Execution completed successfully
MapredLocal task succeeded
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_1678816063243_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1678816063243_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1678816063243_0002
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2023-03-14 11:13:08,937 Stage-4 map = 0%,  reduce = 0%
2023-03-14 11:13:30,547 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 7.46 sec
2023-03-14 11:13:47,553 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 10.5 sec
MapReduce Total cumulative CPU time: 10 seconds 500 msec
Ended Job = job_1678816063243_0002
Launching Job 2 out of 2
 cloudera@quickstart:~/Desktop mined at compile time: 1
```

SQOOP Export to client database:

```
[cloudera@quickstart Desktop]$ sqoop export --connect jdbc:mysql://localhost:3306/output --username root --password cloudera --table sol6 --export-dir  /user/hive/warehouse/out6 --input-fields-terminated-by '\000
1'
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/14 11:20:36 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.8.0
23/03/14 11:20:36 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/03/14 11:20:37 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/14 11:20:37 INFO tool.CodeGenTool: Beginning code generation
23/03/14 11:20:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `sol6` AS t LIMIT 1
23/03/14 11:20:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `sol6` AS t LIMIT 1
23/03/14 11:20:38 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/ed5049a9bfb8185736d7c2ae79ca64af/sol6.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/03/14 11:20:45 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/ed5049a9bfb8185736d7c2ae79ca64af/sol6.jar
23/03/14 11:20:45 INFO mapreduce.ExportJobBase: Beginning export of sol6
23/03/14 11:20:45 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/03/14 11:20:46 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/03/14 11:20:46 INFO Configuration.deprecation: mapred.map.max.attempts is deprecated. Instead, use mapreduce.map.maxattempts
23/03/14 11:20:49 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
23/03/14 11:20:49 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/14 11:20:49 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/03/14 11:20:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/03/14 11:20:52 WARN hdfs.DFSClient: Caught exception
```

Output in the client database:

```
mysql> use output;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from sol6 limit 10;
+------+-----------+
| pid  | exclsuive |
+------+-----------+
| 3673 |        25 |
| 8669 |        25 |
| 1795 |        24 |
| 1724 |        24 |
| 3515 |        24 |
| 4996 |        24 |
| 1478 |        24 |
| 3628 |        24 |
| 6674 |        24 |
| 6576 |        24 |
+------+-----------+
10 rows in set (0.00 sec)
```

Problem Statement 7:

7.5. Anna wants a report on the pricing of the medicine. She wants a list of the most expensive and most affordable medicines only.

Assist anna by creating a report of all the medicines which are pricey and affordable, listing the companyName, productName, description, maxPrice, and the price category of each. Sort the list in descending order of the maxPrice.

Note: A medicine is considered to be “**pricey**” if the max price exceeds 1000 and “**affordable**” if the price is under 5. Write a query to find

hive >select * from (select productname,companyname,description,maxprice m,(case when maxprice>=1000 then "pricy"

when maxprice<=5 then "affordabale" end) as type from medicine) a where a.type is not null order by a.productname,a.m desc ;

hive > CREATE EXTERNAL TABLE IF NOT EXISTS out7 (pname string, cname String, description String, mprice int,typ string)

COMMENT 'Employee details'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

LINES TERMINATED BY '\n';

Sqoop export:

[cloudera@quickstart Desksqoop export --connect jdbc:mysql://localhost:3306/output --username root --password cloudera --table sol7 --export-dir /user/hive/warehouse/out7/000000_0 --input-fields-terminated-by '\t'

```
23/03/15 03:33:15 INFO mapreduce.Job: Job job_1678871603714_0015 running in uber mode : false
23/03/15 03:33:15 INFO mapreduce.Job:  map 0% reduce 0%
l23/03/15 03:33:42 INFO mapreduce.Job:  map 25% reduce 0%
23/03/15 03:33:45 INFO mapreduce.Job:  map 75% reduce 0%
23/03/15 03:33:46 INFO mapreduce.Job:  map 100% reduce 0%
23/03/15 03:33:46 INFO mapreduce.Job: Job job_1678871603714_0015 completed successfully
23/03/15 03:33:46 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=566148
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=601026
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=19
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=4
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=102053
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=102053
                Total vcore-seconds taken by all map tasks=102053
                Total megabyte-seconds taken by all map tasks=104502272
        Map-Reduce Framework
                Map input records=5742
                Map output records=5742
                Input split bytes=666
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1490
                CPU time spent (ms)=5120
                Physical memory (bytes) snapshot=431386624
                Virtual memory (bytes) snapshot=6019416064
                Total committed heap usage (bytes)=243531776
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
23/03/15 03:33:46 INFO mapreduce.ExportJobBase: Transferred 586.9395 KB in 43.8164 seconds (13.3954 KB/sec)
23/03/15 03:33:46 INFO mapreduce.ExportJobBase: Exported 5742 records.
```

Problem Statement 8:

3.3. Johansson is trying to prepare a report on patients who have gone through treatments more than once. Help Johansson prepare a report that shows the patient's name, the number of treatments they have undergone, and their age, Sort the data in a way that the patients who have undergone more treatments appear on top.

hive > CREATE EXTERNAL TABLE IF NOT EXISTS out8 (pname string, coun int,age int)

COMMENT ' details'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

hive> insert overwrite table out8

> select p.patientid as patientid,p.tcount as tcount ,floor(datediff(current_date,t.dob)/365.25)as age from

> (select patientid as patientid,count(patientid) as tcount from treatment

> group by patientid having tcount>1 order by tcount desc)p,patient t

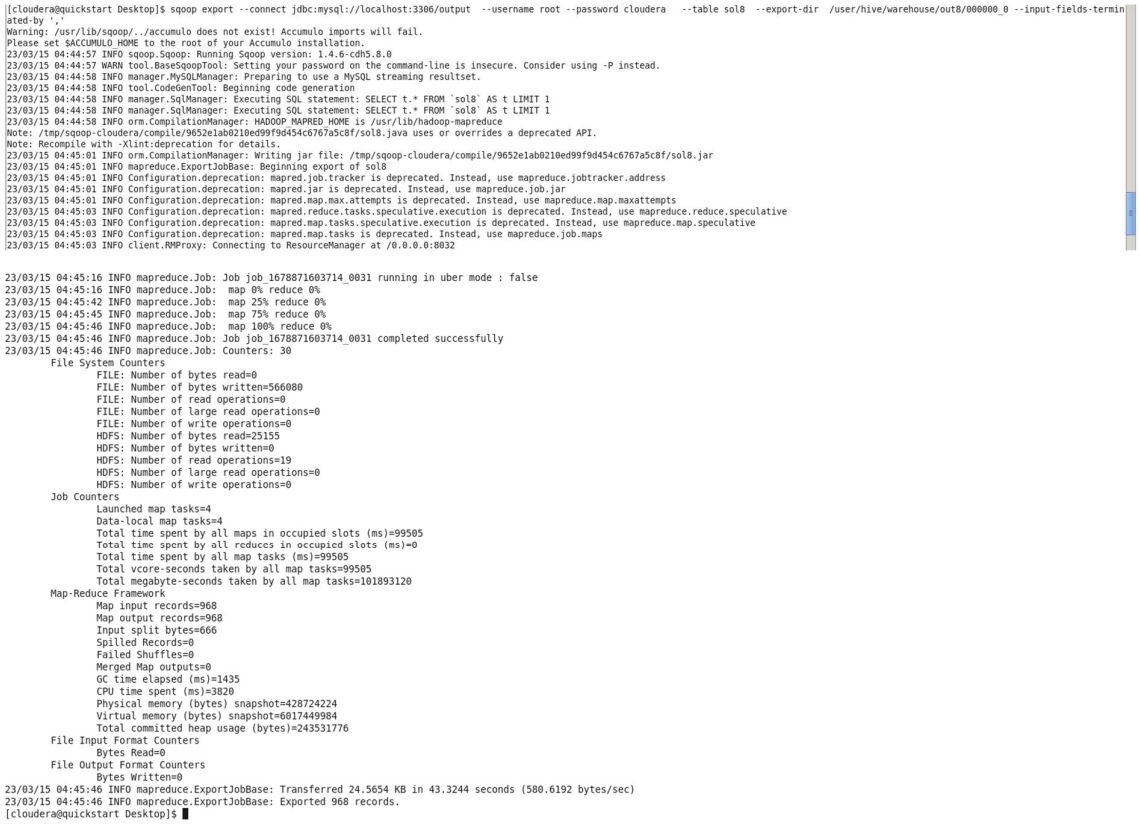
> where p.patientid=t.patientid order by tcount desc;

```
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1678871603714_0029, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1678871603714_0029/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1678871603714_0029
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-03-15 04:37:27,521 Stage-2 map = 0%,  reduce = 0%
2023-03-15 04:37:35,502 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.24 sec
2023-03-15 04:37:44,240 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.66 sec
MapReduce Total cumulative CPU time: 2 seconds 660 msec
Ended Job = job_1678871603714_0029
Execution log at: /tmp/cloudera/cloudera_20230315043636_6529a93e-c4c6-4721-85b6-f180b247f98b.log
2023-03-15 04:37:51     Starting to launch local task to process map join;      maximum memory = 1013645312
2023-03-15 04:37:53     Dump the side-table for tag: 1 with group count: 1126 into file: file:/tmp/cloudera/729f4047-f583-4cb8-8eeb-220c
hTable-Stage-4/MapJoin-mapfile01--.hashtable
2023-03-15 04:37:53     Uploaded 1 File to: file:/tmp/cloudera/729f4047-f583-4cb8-8eeb-220c1588180a/hive_2023-03-15_04-36-42_398_7810705
(37601 bytes)
2023-03-15 04:37:53     End of local task; Time Taken: 1.86 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
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_1678871603714_0030, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1678871603714_0030/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1678871603714_0030
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2023-03-15 04:38:03,410 Stage-4 map = 0%,  reduce = 0%
2023-03-15 04:38:12,334 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 2.05 sec
2023-03-15 04:38:22,196 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 4.12 sec
MapReduce Total cumulative CPU time: 4 seconds 120 msec
Ended Job = job_1678871603714_0030
Loading data to table default.out8
Table default.out8 stats: [numFiles=1, numRows=968, totalSize=11839, rawDataSize=10871]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.45 sec   HDFS Read: 416330 HDFS Write: 21592 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.66 sec   HDFS Read: 25617 HDFS Write: 21592 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 4.12 sec   HDFS Read: 33548 HDFS Write: 11911 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 230 msec
OK
Time taken: 101.61 seconds
```

mysql> create table sol8(pname int,coun int,age int);

Query OK, 0 rows affected (0.01 sec)

SQOOP export:



```
[cloudera@quickstart Desktop]$ sqoop export --connect jdbc:mysql://localhost:3306/output  --username root --password cloudera   --table sol8  --export-dir  /user/hive/warehouse/out8/000000_0 --input-fields-terminated-by ','
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/15 04:44:57 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.8.0
23/03/15 04:44:57 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/03/15 04:44:58 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/15 04:44:58 INFO tool.CodeGenTool: Beginning code generation
23/03/15 04:44:58 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `sol8` AS t LIMIT 1
23/03/15 04:44:58 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `sol8` AS t LIMIT 1
23/03/15 04:44:58 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/9652e1ab0210ed99f9d454c6767a5c8f/sol8.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/03/15 04:45:01 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/9652e1ab0210ed99f9d454c6767a5c8f/sol8.jar
23/03/15 04:45:01 INFO mapreduce.ExportJobBase: Beginning export of sol8
23/03/15 04:45:01 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/03/15 04:45:01 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/03/15 04:45:01 INFO Configuration.deprecation: mapred.map.max.attempts is deprecated. Instead, use mapreduce.map.maxattempts
23/03/15 04:45:03 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
23/03/15 04:45:03 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/15 04:45:03 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/03/15 04:45:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032


23/03/15 04:45:16 INFO mapreduce.Job: Job job_1678871603714_0031 running in uber mode : false
23/03/15 04:45:16 INFO mapreduce.Job:  map 0% reduce 0%
23/03/15 04:45:42 INFO mapreduce.Job:  map 25% reduce 0%
23/03/15 04:45:45 INFO mapreduce.Job:  map 75% reduce 0%
23/03/15 04:45:46 INFO mapreduce.Job:  map 100% reduce 0%
23/03/15 04:45:46 INFO mapreduce.Job: Job job_1678871603714_0031 completed successfully
23/03/15 04:45:46 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=566080
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=25155
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=19
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=4
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=99505
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=99505
                Total vcore-seconds taken by all map tasks=99505
                Total megabyte-seconds taken by all map tasks=101893120
        Map-Reduce Framework
                Map input records=968
                Map output records=968
                Input split bytes=666
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1435
                CPU time spent (ms)=3820
                Physical memory (bytes) snapshot=428724224
                Virtual memory (bytes) snapshot=6017449984
                Total committed heap usage (bytes)=243531776
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
23/03/15 04:45:46 INFO mapreduce.ExportJobBase: Transferred 24.5654 KB in 43.3244 seconds (580.6192 bytes/sec)
23/03/15 04:45:46 INFO mapreduce.ExportJobBase: Exported 968 records.
[cloudera@quickstart Desktop]$
```

Problem Statement 9:

6.2.Sarah, from the healthcare department, has noticed many people do not claim insurance for their treatment. She has requested a state-wise report of the percentage of treatments that took place without claiming insurance. Assist Sarah by creating a report as per her requirement.

hive> create view p9 as with cte as (select ad.state as state,t.treatmentid as tid,t.claimid as cid

> from address ad join person p on ad.addressid=p.addressid

> join treatment t on t.patientid=p.personid)

> select a.state,a.count,(a.count/b.total) from

> (select state as state,count(tid) as count from cte where cid is null group by state)a

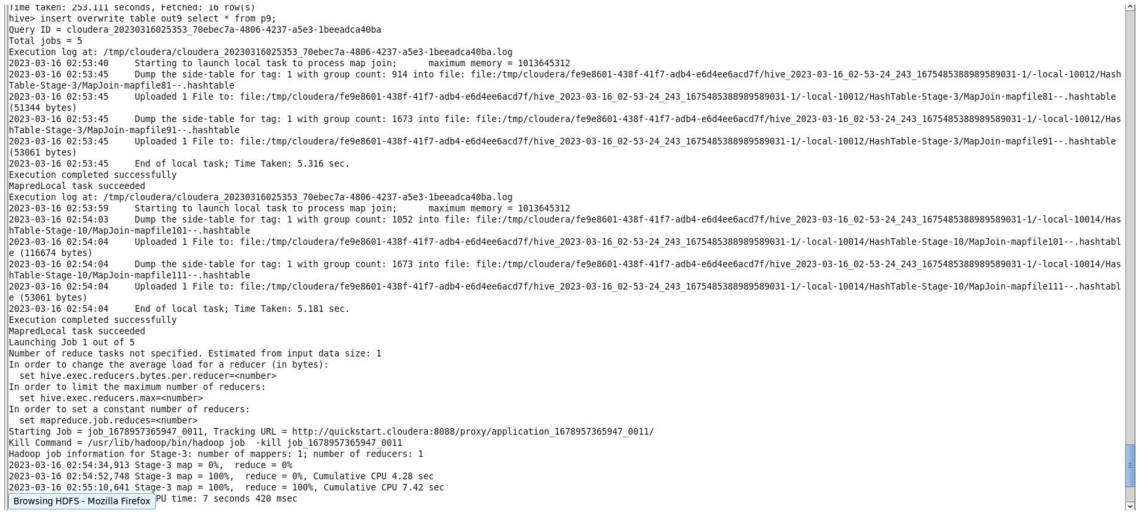
> join

> (select state as state,count(tid) as total from cte group by state)b on a.state=b.state;

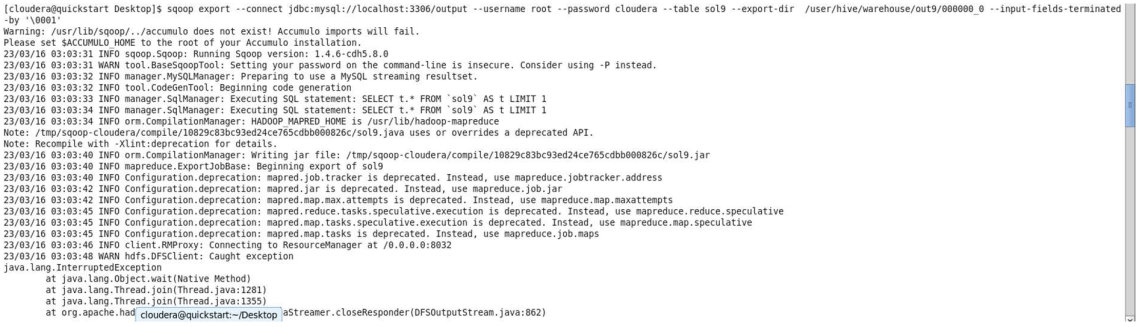
OK

Time taken: 0.996 seconds

Create an external table out9 and store the result into it.



SQOOP:



Client database

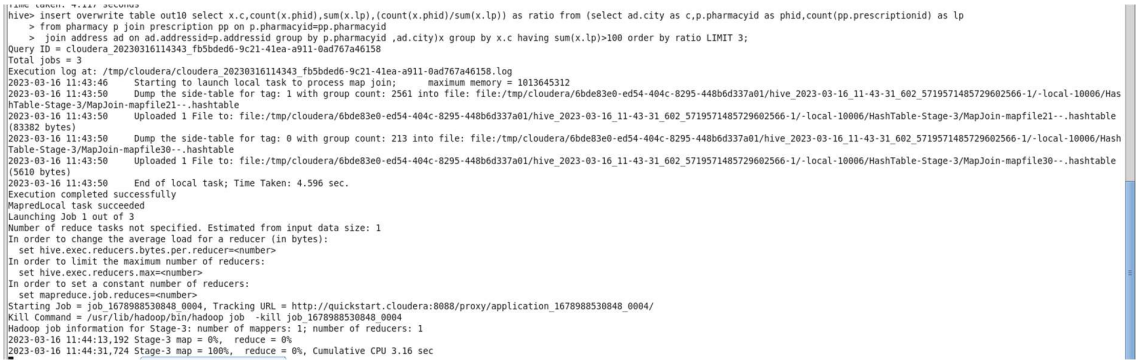
```
mysql> use output;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from sol9;
+-------+------+------------+
| state | coun | rat        |
+-------+------+------------+
| MD    |  220 | 0.34920636 |
| AK    |  150 |  0.3504673 |
| AL    |  280 | 0.33816424 |
| AR    |  216 | 0.36548224 |
| AZ    |  212 | 0.37192982 |
| OK    |  314 | 0.39847717 |
| TN    |  307 |  0.3886076 |
| VT    |  219 | 0.37308347 |
| CA    |  363 | 0.33241758 |
| CO    |  253 |  0.3523677 |
| CT    |  256 |  0.3667622 |
| DC    |  243 |  0.3379694 |
| FL    |  281 | 0.37921727 |
| GA    |  256 | 0.36209336 |
| KY    |  169 | 0.36034116 |
| MA    |  183 | 0.34593573 |
+-------+------+------------+
16 rows in set (0.00 sec)
```

Problem statement 10:

2.1. A company needs to set up 3 new pharmacies, they have come up with an idea that the pharmacy can be set up in cities where the pharmacy-to-prescription ratio is the lowest and the number of prescriptions should exceed 100. Assist the company to identify those cities where the pharmacy can be set up.

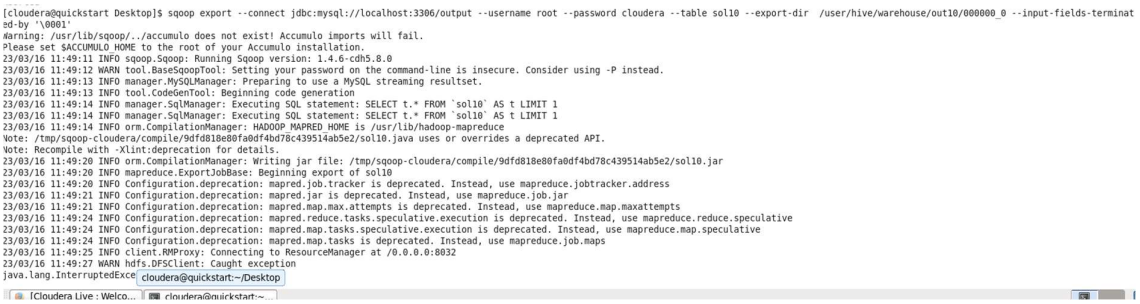
Insert result into external table:



Create table sol10 in client database and export data to client using sqoop

mysql> create table sol10(city varchar(20),coun int,summ int,rat double);

Query OK, 0 rows affected (0.02 sec)



```
23/03/16 11:49:49 INFO mapreduce.Job: Job job_1678988530848_0007 running in uber mode : false
23/03/16 11:49:49 INFO mapreduce.Job:  map 0% reduce 0%
23/03/16 11:50:49 INFO mapreduce.Job:  map 50% reduce 0%
23/03/16 11:50:51 INFO mapreduce.Job:  map 100% reduce 0%
23/03/16 11:50:51 INFO mapreduce.Job: Job job_1678988530848_0007 completed successfully
23/03/16 11:50:51 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=566148
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=932
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=19
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=4
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=228603
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=228603
                Total vcore-seconds taken by all map tasks=228603
                Total megabyte-seconds taken by all map tasks=234089472
        Map-Reduce Framework
                Map input records=3
                Map output records=3
                Input split bytes=671
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=4247
                CPU time spent (ms)=6540
                Physical memory (bytes) snapshot=426065920
                Virtual memory (bytes) snapshot=6015533056
                Total committed heap usage (bytes)=243531776
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
23/03/16 11:50:51 INFO mapreduce.ExportJobBase: Transferred 932 bytes in 86.8376 seconds (10.7327 bytes/sec)
Cloudera Live : Welcome! - Cloudera ExportJobBase: Exported 3 records.
```