Healthcare Data Analysis using hive

1. Import client data to the distributed filesystem into the warehouse

- The data is available as a SQL file. So first, create a healthcare database.
- Dump the SQL file into the database.

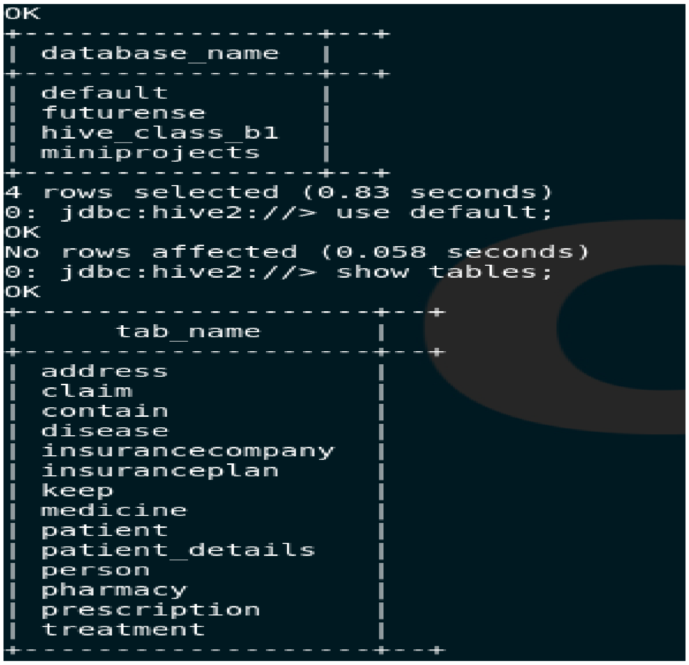
```
sudo mysql -u root -p  -h localhost -D healthcare < healthcare.sql
```

- Using sqoop import all the tables into hive.

```
sqoop import-all-tables
--connect jdbc:mysql://localhost/healthcare
--username root
--password cloudera
--hive-import
--m 1
```

- All the tables will be imported under /user/warehouse
- And we can see them under the default database using the beeline/hive command line interface.

```
beeline --silent=true -u jdbc:hive2://localhost:10000 root cloudera
Use default;
Show tables;
```



```
OK
+-----------------+--+
| database_name   |
+-----------------+--+
| default         |
| futurense       |
| hive_class_b1   |
| miniprojects    |
+-----------------+--+
4 rows selected (0.83 seconds)
0: jdbc:hive2://> use default;
OK
No rows affected (0.058 seconds)
0: jdbc:hive2://> show tables;
OK
+---------------------+--+
|      tab_name       |
+---------------------+--+
| address             |
| claim               |
| contain             |
| disease             |
| insurancecompany    |
| insuranceplan       |
| keep                |
| medicine            |
| patient             |
| patient_details     |
| person              |
| pharmacy            |
| prescription        |
| treatment           |
+---------------------+--+
```

2. Implement data analysis with hive queries on internal tables

- Create external tables to store the results of hive queries performed.
- Run the analytical queries and insert the obtained result into the above external table.

Problem statement -1 :

Jimmy, from the healthcare department, wants to know which disease is infecting people of which gender more often.

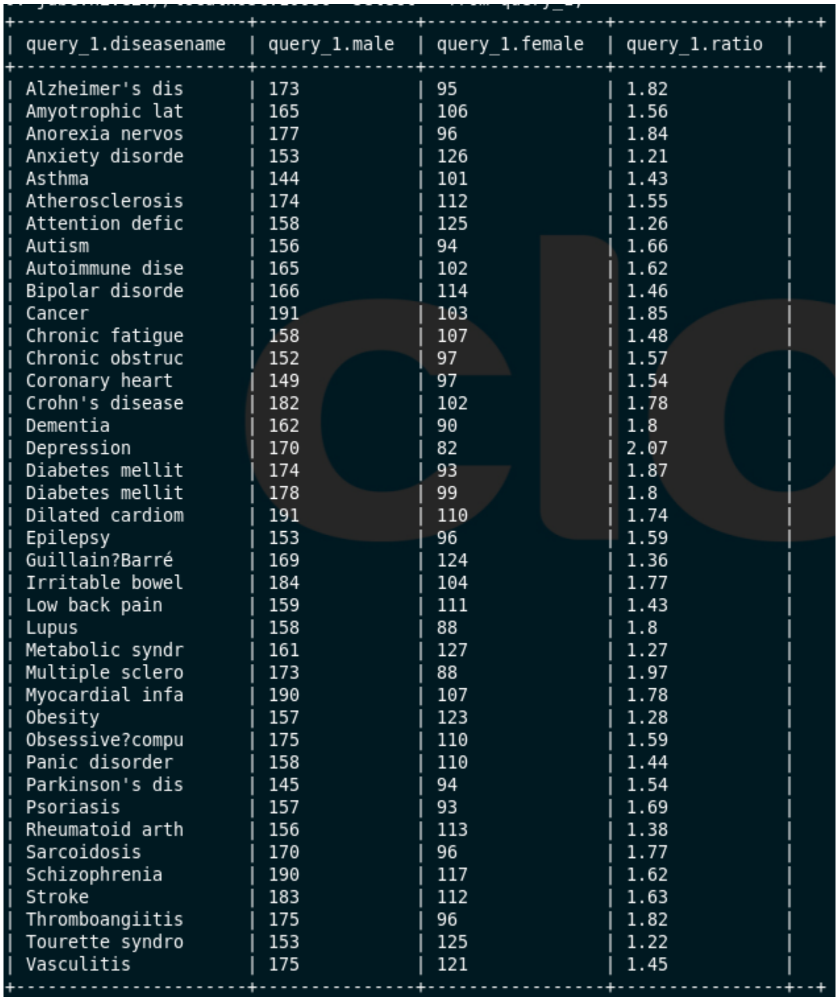
Assist Jimmy with this purpose by generating a report that shows for each disease the male-to-female ratio. Sort the data in a way that is helpful for Jimmy

External table for query-1

```
create external table query_1
(
diseaseName string,
male int,
female int,
ratio double
);
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';
```

Loading output into the external table query-1

```
insert into table query_1
select *, round((male/female),2) as m_f_ratio from
(
select d.diseaseName,
sum(case when gender = 'male' then 1 else 0 end ) as male,
sum(case when gender = 'female' then 1 else 0 end ) as female
from person p inner join treatment t
on p.personid = t.patientid
inner join disease d
on t.diseaseID = d.diseaseID
group by d.diseaseName
) a;
```



```
+----------------------+-------------+-----------------+----------------+--+
| query_1.diseasename  | query_1.male | query_1.female  | query_1.ratio  |
+----------------------+-------------+-----------------+----------------+--+
| Alzheimer's dis      | 173         | 95              | 1.82           |
| Amyotrophic lat      | 165         | 106             | 1.56           |
| Anorexia nervos      | 177         | 96              | 1.84           |
| Anxiety disorde      | 153         | 126             | 1.21           |
| Asthma               | 144         | 101             | 1.43           |
| Atherosclerosis      | 174         | 112             | 1.55           |
| Attention defic      | 158         | 125             | 1.26           |
| Autism               | 156         | 94              | 1.66           |
| Autoimmune dise      | 165         | 102             | 1.62           |
| Bipolar disorde      | 166         | 114             | 1.46           |
| Cancer               | 191         | 103             | 1.85           |
| Chronic fatigue      | 158         | 107             | 1.48           |
| Chronic obstruc      | 152         | 97              | 1.57           |
| Coronary heart       | 149         | 97              | 1.54           |
| Crohn's disease      | 182         | 102             | 1.78           |
| Dementia             | 162         | 90              | 1.8            |
| Depression           | 170         | 82              | 2.07           |
| Diabetes mellit      | 174         | 93              | 1.87           |
| Diabetes mellit      | 178         | 99              | 1.8            |
| Dilated cardiom      | 191         | 110             | 1.74           |
| Epilepsy             | 153         | 96              | 1.59           |
| Guillain?Barré       | 169         | 124             | 1.36           |
| Irritable bowel      | 184         | 104             | 1.77           |
| Low back pain        | 159         | 111             | 1.43           |
| Lupus                | 158         | 88              | 1.8            |
| Metabolic syndr      | 161         | 127             | 1.27           |
| Multiple sclero      | 173         | 88              | 1.97           |
| Myocardial infa      | 190         | 107             | 1.78           |
| Obesity              | 157         | 123             | 1.28           |
| Obsessive?compu      | 175         | 110             | 1.59           |
| Panic disorder       | 158         | 110             | 1.44           |
| Parkinson's dis      | 145         | 94              | 1.54           |
| Psoriasis            | 157         | 93              | 1.69           |
| Rheumatoid arth      | 156         | 113             | 1.38           |
| Sarcoidosis          | 170         | 96              | 1.77           |
| Schizophrenia        | 190         | 117             | 1.62           |
| Stroke               | 183         | 112             | 1.63           |
| Thromboangiitis      | 175         | 96              | 1.82           |
| Tourette syndro      | 153         | 125             | 1.22           |
| Vasculitis           | 175         | 121             | 1.45           |
+----------------------+-------------+-----------------+----------------+--+
```

Problem statement -2 :

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 treatment-to-claim ratio.

And notice if there is a significant difference between the treatment-to-claim ratio of male and female patients.

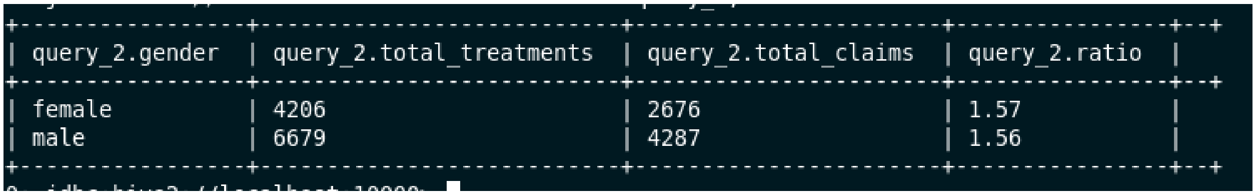
Create an external table for query-2

```
create external table query_2(
gender string,
total_treatments int,
total_claims int,
ratio double)
```

```
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';
```

Load output into the external table query-1

```
insert into table query_2
select *, round(total_treatments/ total_claims, 2) as ratio from
(
select p.gender, count(t.treatmentID) as total_treatments,
count(c.claimID) as total_claims from
person p inner join treatment t
on p.personID = t.patientID
left join claim c
on t.claimID = c.claimID
group by p.gender
)a;
```



```
+------------------+---------------------------+-----------------------+-----------------+--+
| query_2.gender   | query_2.total_treatments  | query_2.total_claims  | query_2.ratio   |
+------------------+---------------------------+-----------------------+-----------------+--+
| female           | 4206                      | 2676                  | 1.57            |
| male             | 6679                      | 4287                  | 1.56            |
+------------------+---------------------------+-----------------------+-----------------+--+
0: jdbc:hive2://localhost:10000>
```

3. Implement partitions and bucketing

4. Implement external tables for results to take into client db

- To perform dynamic partitioning first we need to set some configurations to true.

```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
```

Partitioning on the address table :

- Create a new table address_part by partitioning the address table using the “state” column.

```
create table if not exists address_part
(
addressid int,
address1 string,
city string,
zip int
)
partitioned by (state string)
row format delimited
fields terminated by ','
```

```
stored as textfile;
```

Insert into address_part table :

```
insert into address_part
partition(state)
select addressid,address1, city, zip, state
from address;
```

Partitioning and bucketing on the treatment table:

- Create a new table treatment_part_bkt by partitioning the treatment table by year.
- Create 3 buckets using treatmentid.

```
create table if not exists treatment_part_bkt
(
treatmentid int,
date string,
patientid int,
diseaseid int,
claimid int
)
partitioned by (year string)
clustered by (treatmentid) into 3 buckets
row format delimited
fields terminated by ','
stored as textfile
```

Insert into treatment_part_bkt table :

```
insert into treatment_part_bkt
partition(year)
select treatmentid, date, patientid, diseaseid, claimid,
 year(date) as year from treatment;
```

Problem statement - 3:

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.

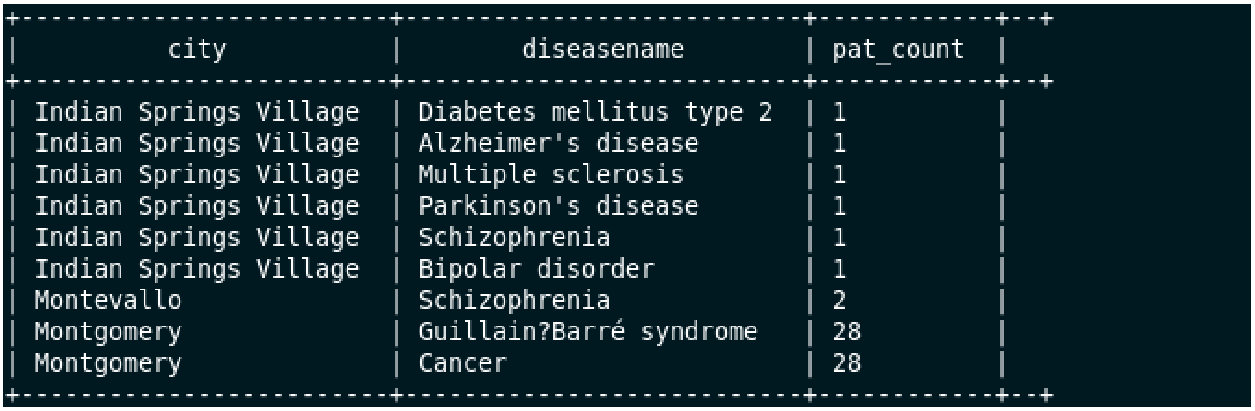
Note: The state of Alabama is represented as AL in Address Table.

Create an external table for query-3

```
create external table query_3
(
city string,
diseasename string,
pat_count int
)
row format delimited
fields terminated by ','
stored as textfile;
```

Insert into query_3 table

```
with cte as
(
select city,d.diseaseName, count(patientID) as pat_count
from (select * from address_part where state='AL') ad inner join person p
on ad.addressID = p.addressID
inner join treatment t
on personID = t.patientID
inner join disease as d
on t.diseaseID = d.diseaseID
group by city,diseaseName
),
cte2 as
(
select  city,diseaseName,pat_count, dense_rank() over(partition by city order by
pat_count desc) as dn from cte
)
insert into query_3 select city,diseaseName,pat_count from cte2 where dn = 1;
```



```
+------------------------+-------------------------+------------+--+
|          city          |       diseasename       | pat_count  |
+------------------------+-------------------------+------------+--+
| Indian Springs Village | Diabetes mellitus type 2 | 1          |
| Indian Springs Village | Alzheimer's disease     | 1          |
| Indian Springs Village | Multiple sclerosis      | 1          |
| Indian Springs Village | Parkinson's disease     | 1          |
| Indian Springs Village | Schizophrenia           | 1          |
| Indian Springs Village | Bipolar disorder        | 1          |
| Montevallo             | Schizophrenia           | 2          |
| Montgomery             | Guillain?Barré syndrome | 28         |
| Montgomery             | Cancer                  | 28         |
+------------------------+-------------------------+------------+--+
```

Problem statement - 4:

Sarah, from the healthcare department, is trying to understand if some diseases are spreading in a particular region. Assist Sarah by creating a report which shows each state the number of the most and least treated diseases by the patients of that state in

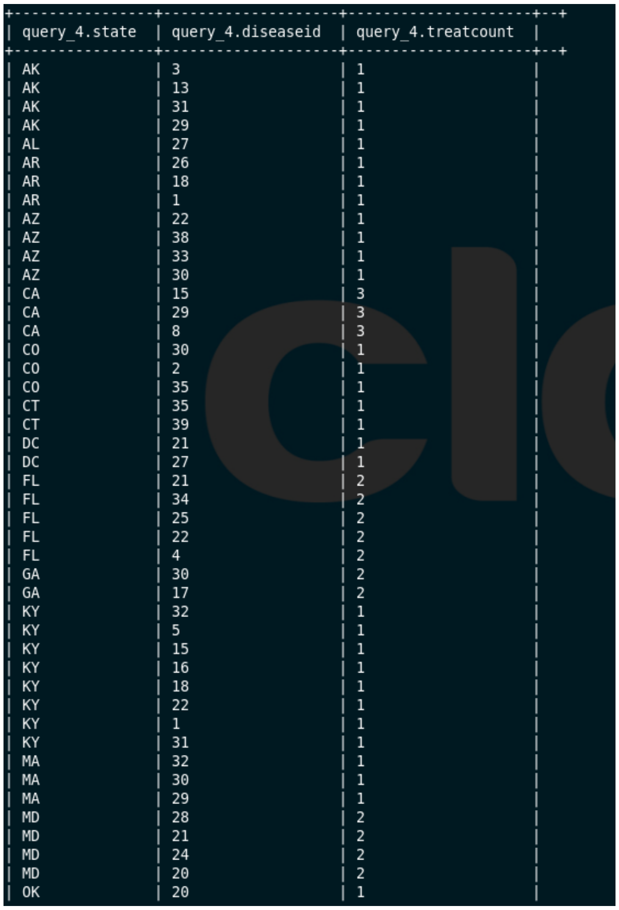
the year 2022. It would be helpful for Sarah if the aggregation for the different combinations is found as well. Assist Sarah to create this report.

External table for query_4

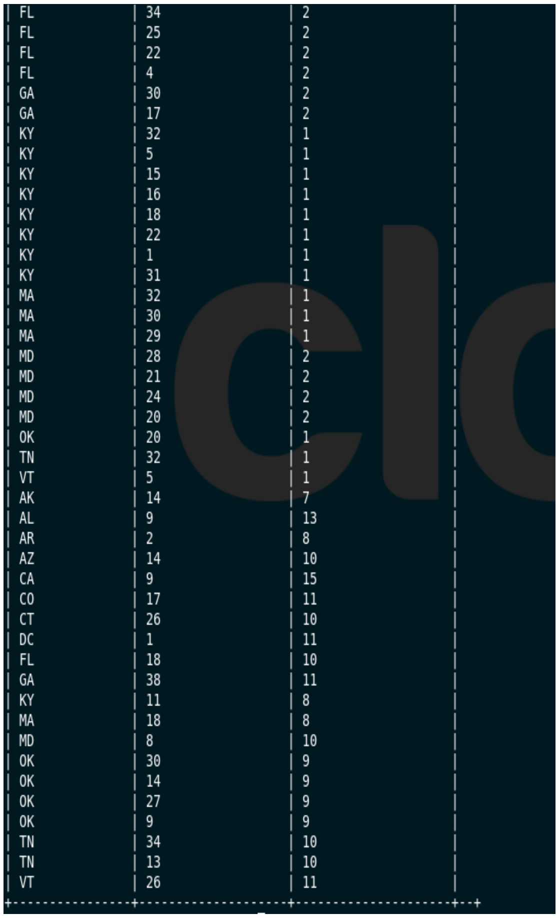
```
create external table query_4
(
state string,
diseaseid int,
treatCount int
)
row format delimited
fields terminated by ','
stored as textfile;
```

Insert into query_4

```
with cte as
(
select ad.state,t.diseaseID, count(t.treatmentID) as treat_count
from address_part ad inner join person p
on ad.addressID = p.addressID
inner join (select * from treatment_part_bkt where year = 2022) t
on p.personID = t.patientID
group by ad.state,t.diseaseID
),
cte_2 as
(
select *, dense_rank() over(partition  by state order by treat_count desc) as dn_desc,
dense_rank() over(partition  by state order by treat_count ) as dn_asc from cte
)
insert into query_4
select state,diseaseID,treat_count from cte_2 where dn_desc = 1
union all
select state,diseaseID,treat_count from cte_2 where dn_asc = 1
order by state;
```



```
+----------------+--------------------+---------------------+--+
| query_4.state  | query_4.diseaseid  | query_4.treatcount  |
+----------------+--------------------+---------------------+--+
| AK             | 3                  | 1                   |
| AK             | 13                 | 1                   |
| AK             | 31                 | 1                   |
| AK             | 29                 | 1                   |
| AL             | 27                 | 1                   |
| AR             | 26                 | 1                   |
| AR             | 18                 | 1                   |
| AR             | 1                  | 1                   |
| AZ             | 22                 | 1                   |
| AZ             | 38                 | 1                   |
| AZ             | 33                 | 1                   |
| AZ             | 30                 | 1                   |
| CA             | 15                 | 3                   |
| CA             | 29                 | 3                   |
| CA             | 8                  | 3                   |
| CO             | 30                 | 1                   |
| CO             | 2                  | 1                   |
| CO             | 35                 | 1                   |
| CT             | 35                 | 1                   |
| CT             | 39                 | 1                   |
| DC             | 21                 | 1                   |
| DC             | 27                 | 1                   |
| FL             | 21                 | 2                   |
| FL             | 34                 | 2                   |
| FL             | 25                 | 2                   |
| FL             | 22                 | 2                   |
| FL             | 4                  | 2                   |
| GA             | 30                 | 2                   |
| GA             | 17                 | 2                   |
| KY             | 32                 | 1                   |
| KY             | 5                  | 1                   |
| KY             | 15                 | 1                   |
| KY             | 16                 | 1                   |
| KY             | 18                 | 1                   |
| KY             | 22                 | 1                   |
| KY             | 1                  | 1                   |
| KY             | 31                 | 1                   |
| MA             | 32                 | 1                   |
| MA             | 30                 | 1                   |
| MA             | 29                 | 1                   |
| MD             | 28                 | 2                   |
| MD             | 21                 | 2                   |
| MD             | 24                 | 2                   |
| MD             | 20                 | 2                   |
| OK             | 20                 | 1                   |
```



```
| FL             | 34                 | 2                   |
| FL             | 25                 | 2                   |
| FL             | 22                 | 2                   |
| FL             | 4                  | 2                   |
| GA             | 30                 | 2                   |
| GA             | 17                 | 2                   |
| KY             | 32                 | 1                   |
| KY             | 5                  | 1                   |
| KY             | 15                 | 1                   |
| KY             | 16                 | 1                   |
| KY             | 18                 | 1                   |
| KY             | 22                 | 1                   |
| KY             | 1                  | 1                   |
| KY             | 31                 | 1                   |
| MA             | 32                 | 1                   |
| MA             | 30                 | 1                   |
| MA             | 29                 | 1                   |
| MD             | 28                 | 2                   |
| MD             | 21                 | 2                   |
| MD             | 24                 | 2                   |
| MD             | 20                 | 2                   |
| OK             | 20                 | 1                   |
| TN             | 32                 | 1                   |
| VT             | 5                  | 1                   |
| AK             | 14                 | 7                   |
| AL             | 9                  | 13                  |
| AR             | 2                  | 8                   |
| AZ             | 14                 | 10                  |
| CA             | 9                  | 15                  |
| CO             | 17                 | 11                  |
| CT             | 26                 | 10                  |
| DC             | 1                  | 11                  |
| FL             | 18                 | 10                  |
| GA             | 38                 | 11                  |
| KY             | 11                 | 8                   |
| MA             | 18                 | 8                   |
| MD             | 8                  | 10                  |
| OK             | 30                 | 9                   |
| OK             | 14                 | 9                   |
| OK             | 27                 | 9                   |
| OK             | 9                  | 9                   |
| TN             | 34                 | 10                  |
| TN             | 13                 | 10                  |
| VT             | 26                 | 11                  |
+----------------+--------------------+---------------------+--+
```

Problem statement - 5:

Brooke is trying to figure out if patients with a particular disease are preferring some pharmacies over others or not,
For this purpose, she has requested a detailed pharmacy report that shows each pharmacy name, and how many prescriptions they have prescribed for each disease in 2021 and 2022,
She expects the number of prescriptions prescribed in 2021 and 2022 be displayed in two separate columns.
Write a query for Brooke's requirement.

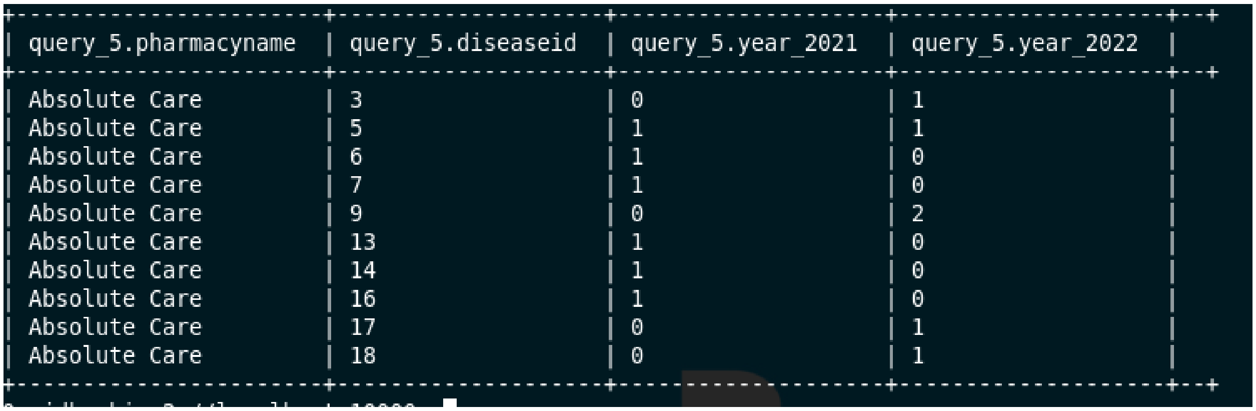
External table for query_5

```
create external table if not exists query_5
(
pharmacyid int,
diseaseid int,
year_2021 int,
year_2022 int
)
```

```
row format delimited
fields terminated by ','
stored as textfile;
```

Insert into query_5

```
select ph.pharmacyName,t.diseaseID,
sum(case when year(t.date) = 2021 then 1 else 0 end) as '2021',
sum(case when year(t.date) = 2022 then 1 else 0 end) as '2022'
from pharmacy ph
inner join prescription pr on ph.pharmacyID = pr.pharmacyID
inner join (select * from treatment_part_bkt where year in (2021,2022)) t on
pr.treatmentID = t.treatmentID
group by ph.pharmacyName,t.diseaseID;
```



```
+-----------------------+--------------------+--------------------+--------------------+--+
| query_5.pharmacyname  | query_5.diseaseid  | query_5.year_2021  | query_5.year_2022  |
+-----------------------+--------------------+--------------------+--------------------+--+
| Absolute Care         | 3                  | 0                  | 1                  |
| Absolute Care         | 5                  | 1                  | 1                  |
| Absolute Care         | 6                  | 1                  | 0                  |
| Absolute Care         | 7                  | 1                  | 0                  |
| Absolute Care         | 9                  | 0                  | 2                  |
| Absolute Care         | 13                 | 1                  | 0                  |
| Absolute Care         | 14                 | 1                  | 0                  |
| Absolute Care         | 16                 | 1                  | 0                  |
| Absolute Care         | 17                 | 0                  | 1                  |
| Absolute Care         | 18                 | 0                  | 1                  |
+-----------------------+--------------------+--------------------+--------------------+--+
```

5. Export external table to SQL or NoSQL using sqoop

- To export tables from hive to MySQL, make sure that the tables with the same schema should be available in the MySQL DB.

Create tables in MySQL

Table - 1

```
create table if not exists query_1
(
diseaseName varchar(200),
male int,
female int,
ratio double
);
```

Table - 2

```
create table if not exists query_2
(
gender varchar(10),
total_treatments int,
total_claims int,
ratio double
);
```

Table - 3

```
create table if not exists query_3
(
city varchar(100),
diseasename varchar(200),
pat_count int
);
```

Table - 4

```
create table if not exists query_4
(
state varchar(20),
diseaseid int,
treatCount int
);
```

Table - 5

```
create table if not exists query_5
(
pharmacyname varchar(100),
diseaseid int,
year_2021 int,
year_2022 int
);
```

Sqoop Commands to export hive tables in MySQL

Export table query_1 :

```
sqoop export --connect jdbc:mysql://localhost:3306/healthcare --username root --password
cloudera --table query_1 --export-dir /user/hive/warehouse/query_1  --input-fields-terminated-by
','
```

Export table query_2 :

sqoop export --connect jdbc:mysql://localhost:3306/healthcare --username root --password cloudera --table query_2 --export-dir /user/hive/warehouse/query_2 --input-fields-terminated-by ','

Export table query_3

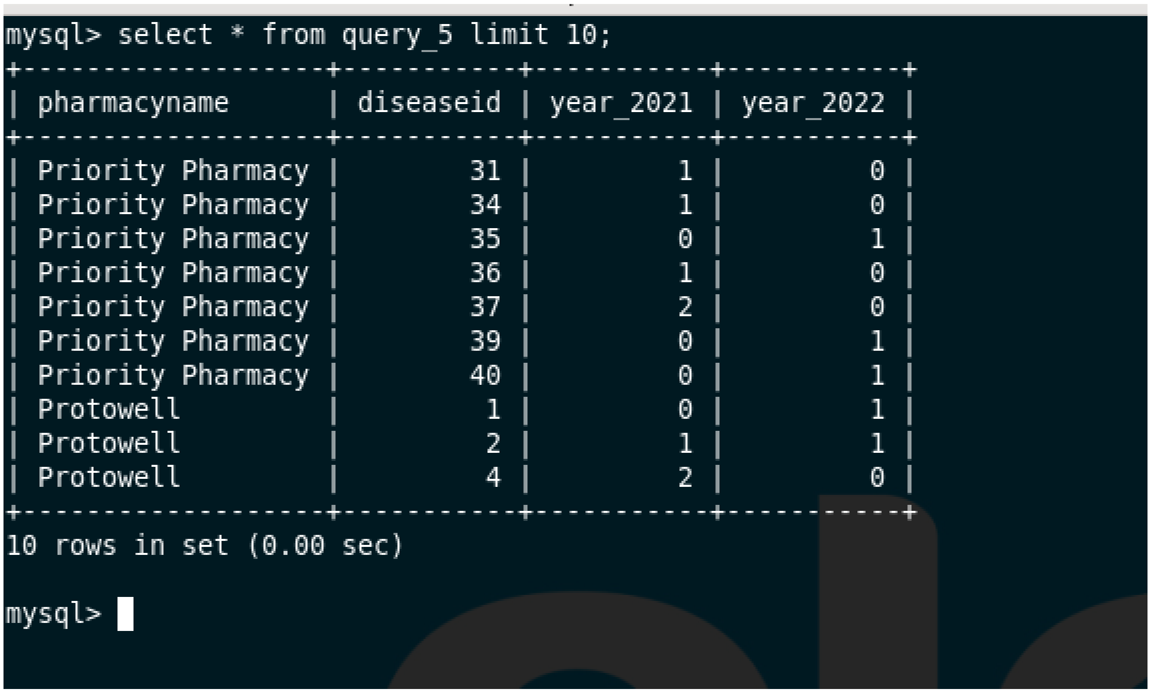
sqoop export --connect jdbc:mysql://localhost:3306/healthcare --username root --password cloudera --table query_3 --export-dir /user/hive/warehouse/query_3 --input-fields-terminated-by ','

Export table query_4

sqoop export --connect jdbc:mysql://localhost:3306/healthcare --username root --password cloudera --table query_4 --export-dir /user/hive/warehouse/query_4 --input-fields-terminated-by ','

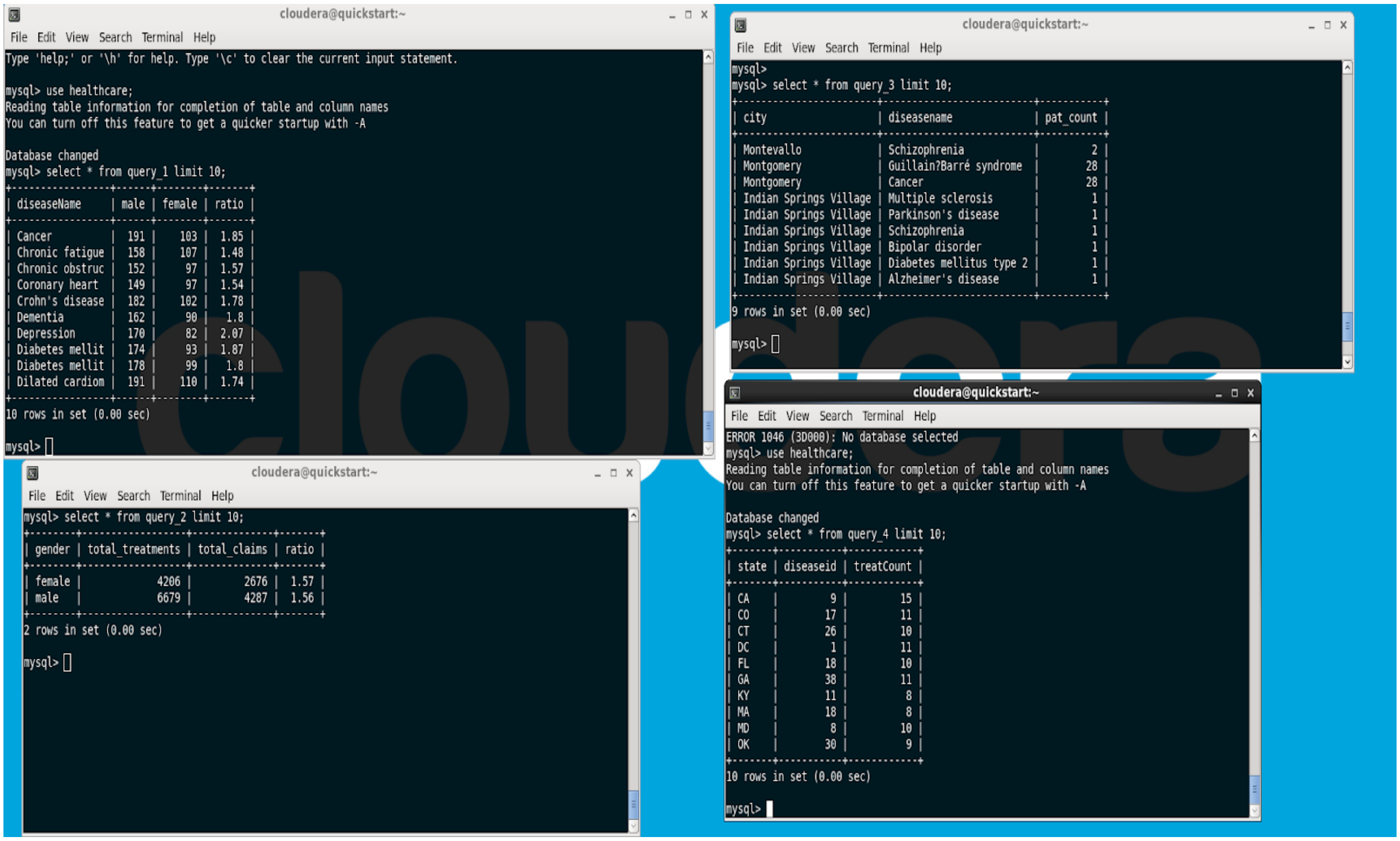
Export table query_5

sqoop export --connect jdbc:mysql://localhost:3306/healthcare --username root --password cloudera --table query_5 --export-dir /user/hive/warehouse/query_5 --input-fields-terminated-by ','



```
mysql> select * from query_5 limit 10;
+-------------------+-----------+-----------+-----------+
| pharmacyname      | diseaseid | year_2021 | year_2022 |
+-------------------+-----------+-----------+-----------+
| Priority Pharmacy |        31 |         1 |         0 |
| Priority Pharmacy |        34 |         1 |         0 |
| Priority Pharmacy |        35 |         0 |         1 |
| Priority Pharmacy |        36 |         1 |         0 |
| Priority Pharmacy |        37 |         2 |         0 |
| Priority Pharmacy |        39 |         0 |         1 |
| Priority Pharmacy |        40 |         0 |         1 |
| Protowell         |         1 |         0 |         1 |
| Protowell         |         2 |         1 |         1 |
| Protowell         |         4 |         2 |         0 |
+-------------------+-----------+-----------+-----------+
10 rows in set (0.00 sec)

mysql> 
```



```
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> use healthcare;
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 query_1 limit 10;
+-----------------+------+--------+-------+
| diseaseName     | male | female | ratio |
+-----------------+------+--------+-------+
| Cancer          |  191 |    103 |  1.85 |
| Chronic fatigue |  158 |    107 |  1.48 |
| Chronic obstruc |  152 |     97 |  1.57 |
| Coronary heart  |  149 |     97 |  1.54 |
| Crohn's disease |  182 |    102 |  1.78 |
| Dementia        |  162 |     90 |   1.8 |
| Depression      |  170 |     82 |  2.07 |
| Diabetes mellit |  174 |     93 |  1.87 |
| Diabetes mellit |  178 |     99 |   1.8 |
| Dilated cardiom |  191 |    110 |  1.74 |
+-----------------+------+--------+-------+
10 rows in set (0.00 sec)

mysql> 
```

```
mysql> select * from query_2 limit 10;
+--------+------------------+--------------+-------+
| gender | total_treatments | total_claims | ratio |
+--------+------------------+--------------+-------+
| female |             4206 |         2676 |  1.57 |
| male   |             6679 |         4287 |  1.56 |
+--------+------------------+--------------+-------+
2 rows in set (0.00 sec)

mysql> 
```

```
mysql>
mysql> select * from query_3 limit 10;
+----------------------+------------------------+-----------+
| city                 | diseasename            | pat_count |
+----------------------+------------------------+-----------+
| Montevallo           | Schizophrenia          |         2 |
| Montgomery           | Guillain?Barré syndrome |        28 |
| Montgomery           | Cancer                 |        28 |
| Indian Springs Village | Multiple sclerosis   |         1 |
| Indian Springs Village | Parkinson's disease  |         1 |
| Indian Springs Village | Schizophrenia        |         1 |
| Indian Springs Village | Bipolar disorder     |         1 |
| Indian Springs Village | Diabetes mellitus type 2 |     1 |
| Indian Springs Village | Alzheimer's disease  |         1 |
+----------------------+------------------------+-----------+
9 rows in set (0.00 sec)

mysql> 
```

```
ERROR 1046 (3D000): No database selected
mysql> use healthcare;
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 query_4 limit 10;
+-------+-----------+-----------+
| state | diseaseid | treatCount |
+-------+-----------+-----------+
| CA    |         9 |        15 |
| CO    |        17 |        11 |
| CT    |        26 |        10 |
| DC    |         1 |        11 |
| FL    |        18 |        10 |
| GA    |        38 |        11 |
| KY    |        11 |         8 |
| MA    |        18 |         8 |
| MD    |         8 |        10 |
| OK    |        30 |         9 |
+-------+-----------+-----------+
10 rows in set (0.00 sec)

mysql> 
```