LAB4 REPORT

Please find the images referred in this report in the folder images within the zip archive.

1. Task1

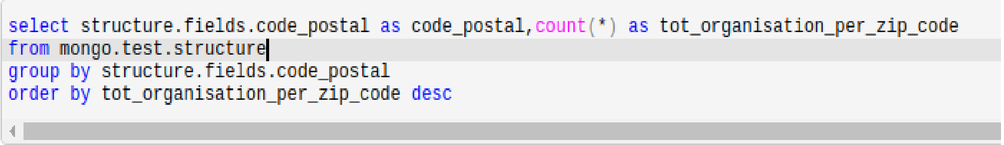
a) SETUP

After lunching the apache Drill server using ./bin/drill-embedded from the installation directory, I then open the user interface from the browser using http://localhost:8087; I then started mongo server,imported the related files and enable the mongo plugin in apache drill interface.

b) Import the structures egalite femmes hommes file in mongo:

Mongoimport -c structure –jsonArray –port 20217 ./filepath

c) The query used is:



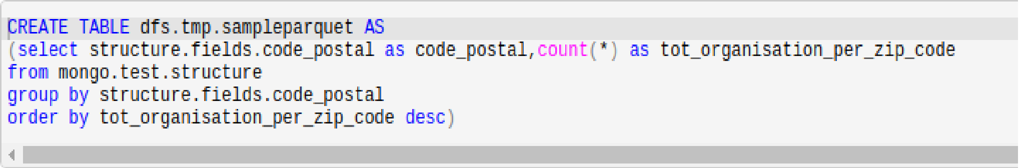
```
select structure.fields.code_postal as code_postal,count(*) as tot_organisation_per_zip_code
from mongo.test.structure
group by structure.fields.code_postal
order by tot_organisation_per_zip_code desc
```

The related picture is **task1_3.**

NB: the command **USE mongo.test** allows to specify the default schema.

d) As we can notice from the picture ,the organisation's zip code data is not complete because we have some missing values substituted with "null" in particular 26 for the attribute code_postal.

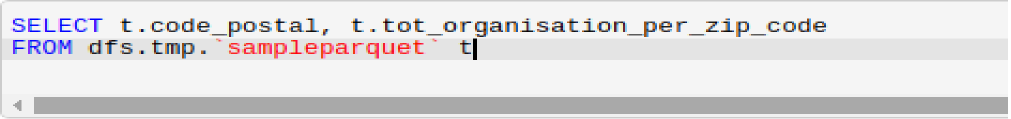
e) The query used to save data into a parquet file is:



```
CREATE TABLE dfs.tmp.sampleparquet AS
(select structure.fields.code_postal as code_postal,count(*) as tot_organisation_per_zip_code
from mongo.test.structure
group by structure.fields.code_postal
order by tot_organisation_per_zip_code desc)
```

And the picture that shows the result is **task1_5**

f) The query used to retrieve data from the parquet file is :



```
SELECT t.code_postal, t.tot_organisation_per_zip_code
FROM dfs.tmp.`sampleparquet` t
```

And the picture that shows the result is **task1_6**

2. Task2

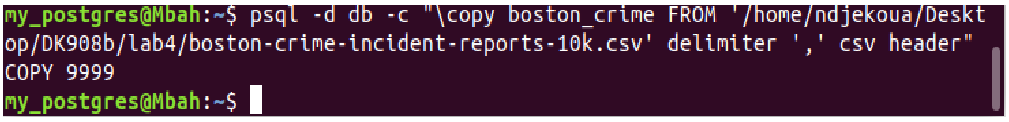
a) To import the Boston crime csv file in postgres, I first created a user named my_postgres, add him to the database and then I created a db named **db** with that user as owner. The commands are the following:

Sudo adduser my_posgres(create a new user)

Su – postgres (login as default posgres user)

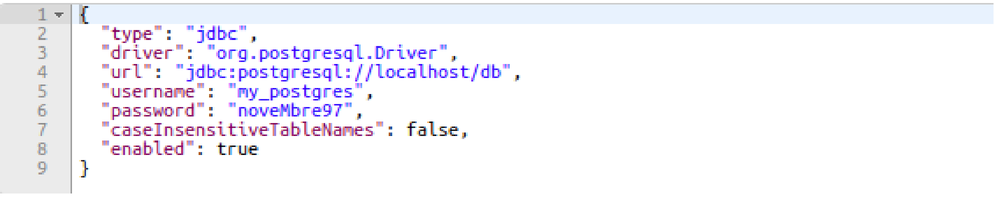
Psql -> then we create the user user in the db, with a database with that user as owner.

In order to load the file, i first login as my_postgres and then i issue the following command:



```
my_postgres@Mbah:~$ psql -d db -c "\copy boston_crime FROM '/home/ndjekoua/Deskt
op/DK908b/lab4/boston-crime-incident-reports-10k.csv' delimiter ',' csv header"
COPY 9999
my_postgres@Mbah:~$ 
```

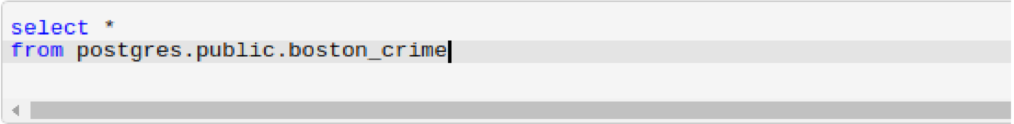
b) The plugin configuration for setting up the connection between drill and the postgres server is the following:



```
{
  "type": "jdbc",
  "driver": "org.postgresql.Driver",
  "url": "jdbc:postgresql://localhost/db",
  "username": "my_postgres",
  "password": "noveMbre97",
  "caseInsensitiveTableNames": false,
  "enabled": true
}
```

NB: we also need to download and add the jdbc driver's jar in the directory <local drill installation folder>/3party/jars

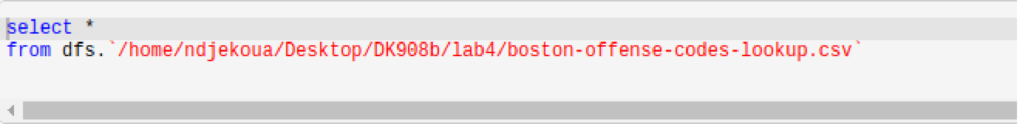
c) Query that shows the content of the postgres database from drill:



```
select *
from postgres.public.boston_crime
```

The result is shown in the picture **task2_3**

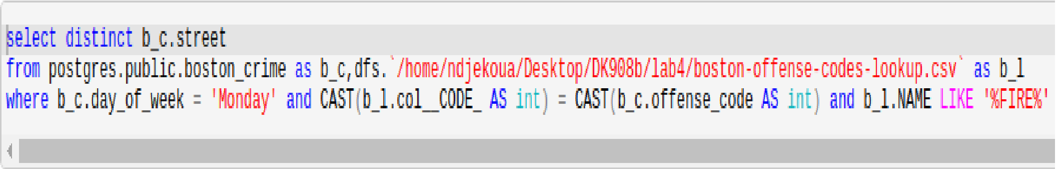
d) The query used to display the content of the file boston crime in apache drill is the following:



```
select *
from dfs.`/home/ndjekoua/Desktop/DK908b/lab4/boston-offense-codes-lookup.csv`
```

The related picture that shows the obtained result is **task2_4**

e) To solve this query we first have to filter each table separately, then we join them together and filter out the ones that does not satisfy the join condition; the corresponding query is the following:



```
select distinct b_c.street
from postgres.public.boston_crime as b_c,dfs.`/home/ndjekoua/Desktop/DK908b/lab4/boston-offense-codes-lookup.csv` as b_l
where b_c.day_of_week = 'Monday' and CAST(b_l.col__CODE_ AS int) = CAST(b_c.offense_code AS int) and b_l.NAME LIKE '%FIRE%'
```

The picture showing the result is: **task2_5.png**