CES Data Scientist 2016
MongoDB Project

youcef KACER

Summary

Introduction..........2

Data integration..........2

Data queries..........4

1. select number of light industrial in New York..........4
2. select the site with the biggest annual average consumption..........5
3. Select the average consumption during winter for site 766..........6
4. Select the peak consumption among all nothern sites (LAT>37°)..........7
5. Select the peak consumption among all southern sites (LAT<37) 8

Data computation..........9

1. Calculate the sum LD for the 100 sites (timestamp interval : 5 minutes)..........9
2. Calculate the average LD by sector of activity (imestamp interval : 5 minutes)..........11
3. Calculate the total LD for the 100 sites (timestamp interval : a week)..........12
4. Calculate the average LD by sector of activity (timestamp interval : a week)..........15

SQL comparison..........18

Introduction

We present here how to solve MongoDB Project.

One can see whole result by executing bash file **run_mongodb.sh**.

We present three different steps:

- **data integration:** to show how data are put into mongodb collections
- **data queries:** that shows 5 simple queries on collections
- **data computation:** that performs the 4 Load Curve statistics on collections
- **SQL comparison** : that compares time execution between sql and mongodb for the 4 Load Curves statistics.

Data integration

First of all, we get energy data from website and de-tar it into « data » folder :

```
> wget https://open-enernoc-data.s3.amazonaws.com/anon/all-data.tar.gz

> mkdir data

> tar -xzvf all-data.tar.gz -C data
```

Then, we pass all_sites.csv file and all conso .csv files from DOS format to Unix format :

```
> sed -i -e 's/\r/\n/g' data/meta/all_sites.csv

> for conso_csv in `ls data/csv/*csv`

> do

> sed -i -e 's/\r/\n/g' $conso_csv

> done
```

We use mongoimport binary to import all_sites.csv and create « sites » collection into « enernoc » database :

```
> mkdir data/db

> mongoimport --type csv -d enernoc -c sites --file data/meta/all_sites.csv –headerline
```

We use **mongodb_integration.js** file in a bash loop to integrate each xxx.csv consumption to its site document as an array of documents. This array of documents is a new field « CONSO » in « sites » collection

```
> for conso_csv in `ls data/csv/*.csv`

> do

> site_number=$(basename ${conso_csv%%.*}) # example : site_number=14 for data/csv/14.csv

> echo  "currently importing site $site_number ..."

> mongoimport --type csv -d enernoc --file $conso_csv --headerline

> mongo mongodb_1_data_integration.js --eval "var site_id = $site_number"

> done
```

mongodb_1_data_integration.js file is as follow:

```
db = db.getSiblingDB('enernoc')

var n = site_id.toString();

db.getCollection(n).remove({anomaly :{$ne:""}});

db.getCollection(n).find().forEach(function(cs){ cs.iso_date = new Date(cs.timestamp*1000); db.getCollection(n).save(cs);})

db.sites.update({'SITE_ID':site_id},{'$set' : {"CONSO" : db.getCollection(n).find().toArray()}});
```

As one can see, this .js file removes measures that has field « anomaly » non empty and create iso_date field for each consumption measure.

We can summarize fields of our « sites » collection by mongodb shell command :

```
db = db.getSiblingDB('enernoc') ;

var doc = db.sites.findOne();

for (var key in doc) {print(key)} ;
```

```
_id

SITE_ID

INDUSTRY

SUB_INDUSTRY

SQ_FT

LAT

LNG

TIME_ZONE

TZ_OFFSET
```

Data queries

We launch **mongodb_2_data_queries.js** to execute 5 simple queries

```
> mongo mongodb_2_data_queries.js
```

This javascript file containing 5 simple queries, is as follow :

1. select number of light industrial in New York

```
db = db.getSiblingDB("enernoc")
print("*** select number of light industrial in New York:");
print(db.sites.find(
                        {INDUSTRY : "Light Industrial"},
                        {TIME_ZONE : "America/New_York"}
                ).count()
);
```

results:

```
*** select number of light industrial in New York:
25
```

2. select the site with the biggest annual average consumption

```
db = db.getSiblingDB("enernoc")
print("*** select the site that has the biggest annual average consumption:");
biggest = db.sites.aggregate([
                    {$project :
                              {
                                        site : "$SITE_ID",
                                        avg_conso : {$avg:"$CONSO.value"}
                              }
                    },
                    {$sort :
                              {
                                        avg_conso :-1
                              }
                    },
                    {$limit : 1}
          ]).next()
print("site",biggest.site,"(annual average consu:",biggest.avg_conso,")");
```

results:

```
*** select the site that has the biggest annual average consumption:
site 766 (annual average consu: 327.6069601453762 )
```

3. Select the average consumption during winter for site 766

```
db = db.getSiblingDB("enernoc")
print("*** select average consumption during winter (between January and Mars) for site 766:");
d = ISODate("2012-04-01T00:00:00Z")
winter = db.getCollection("766").aggregate([
            {$match :
                  {
                        iso_date : {$lt:d}
                  }
            },
            {$group :
                  {
                        _id : null,
                        avg_consu : {$avg:"$value"}
                  }
            }
      ]).next()
print(winter.avg_consu)
```

results :

```
*** select average consumption during winter (between January and Mars) for site 766:
308.8343864349126
```

4. Select the peak consumption among all nothern sites (LAT>37°)

```
db = db.getSiblingDB("enernoc")
print("*** select the peak consumption among all northern sites (LAT>37):");
peak1 = db.sites.aggregate([
                { $match :
                        {
                                LAT : { $gt : 37 }
                        }
                },
                {$project :
                        {
                                SITE_ID : 1,
                                "CONSO.value" : 1,
                                "CONSO.iso_date" : 1
                        }
                },
                {$unwind : "$CONSO"},
                {$sort :
                        {
                                "CONSO.value":-1
                        }
                },
                {$limit : 1}
        ]).next()
print(peak1.CONSO.value,"in site", peak1.SITE_ID,
"on",peak1.CONSO.iso_date.toDateString());
```

results :

```
*** select the peak consumption among all northern sites (LAT>37):
651.5774 in site 14 on Mon Dec 17 2012
```

5. Select the peak consumption among all southern sites (LAT<37°)

```
db = db.getSiblingDB("enernoc")
print("*** select the peak consumption among all southern sites (LAT<37):");
peak2 = db.sites.aggregate([
        {$match:
            {
                LAT : { $lt : 37 }
            }
        },
        {$project :
            {
                SITE_ID : 1,
                "CONSO.value" : 1,
                "CONSO.iso_date" : 1
            }
        },
        {$unwind : "$CONSO"},
        {$sort :
            {
                "CONSO.value" : -1
            }
        },
        {$limit : 1}
    ]).next()
print(peak2.CONSO.value,"in site",peak2.SITE_ID,"on",peak2.CONSO.iso_date.toDateString());
```

results :

```
*** select the peak consumption among all southern sites (LAT<37):
528.0193 in site 45 on Thu Jul 26 2012
```

Data computation

1. Calculate the sum LD for the 100 sites (timestamp interval : 5 minutes)

```
db = db.getSiblingDB("enernoc")
print("*** Calculate the sum LD for the 100 sites (timestamp interval: 5 minutes");
db.getCollection("sites").aggregate([
        {$project :
            {
                site : "$SITE_ID",
                sum_consu : {$sum : "$CONSO.value"}
            }
        },
        {$sort:
            {
                sum_consu : -1
            }
        }
    ]).forEach( function(doc)
        {
            print("site:",doc.site,"\tLD sum:",doc.sum_consu)
        }
    );
```

```
results :
site: 766      LD sum: 34532394.45500381
site: 45       LD sum: 33584845.293399945
site: 10       LD sum: 33544015.780299693
site: 716      LD sum: 33052978.79500011
site: 55       LD sum: 28665910.440900173
site: 786      LD sum: 24303213.122100715
site: 14       LD sum: 22980050.54469993
site: 654      LD sum: 18468253.636890393
site: 718      LD sum: 18050961.867009155
site: 755      LD sum: 12869687.531899346
site: 44       LD sum: 11039816.321199633
site: 690      LD sum: 10931770.223702736
…
site: 92       LD sum: 257495.60190021462
site: 731      LD sum: 245819.5833999371
site: 673      LD sum: 222026.40789988433
site: 648      LD sum: 215914.86370002138
```

2. Calculate the average LD by sector of activity (imestamp interval : 5 minutes)

```
db = db.getSiblingDB("enernoc")
print("*** Calculate the average LD by sector of activity (timestamp interval : 5 minutes)");
db.getCollection("sites").aggregate([
        {$unwind : "$CONSO"},
        {$group :
            {
                _id:"$INDUSTRY",
                avg_consu:{$avg:"$CONSO.value"}
            }
        },
        {$project :
            {
                "_id" : 1,
                "avg_consu" : 1
            }
        }
    ]).forEach( function(doc)
        {
            print("industry:",doc._id,"\tavg LD:",doc.avg_consu) ;
        }
    );
```

results :

```
industry: Light Industrial      avg LD: 80.53685010874234
industry: Education     avg LD: 10.958672942424123
industry: Food Sales & Storage      avg LD: 18.18906005717468
industry: Commercial Property       avg LD: 89.74365052873605
```

3. Calculate the total LD for the 100 sites (timestamp interval : a week)

```
db = db.getSiblingDB("enernoc")
print("*** Calculate the total LD for the 100 sites (timestamp interval: a week)")
db.sites.aggregate([
                {$unwind : "$CONSO"},
                {$group :
                        {
                             _id:
                            {
                                    site_id : "$SITE_ID",
                                    week : {$week : "$CONSO.iso_date"}
                            },
                            sum_consu :
                            {
                                    $sum : "$CONSO.value"
                            }
                         }
                },
                {$sort :
                        {
                             _id:1
                        }
                }
        ]).forEach( function(doc)
                {
                        print("site:",doc._id.site_id,"\tweek:",
                        doc._id.week,"\tsum LD:",doc.sum_consu) ;
                }
        );
```

```
results :
site: 6  week: 0          sum LD: 56.6561
site: 6  week: 1          sum LD: 75468.68400000001
site: 6  week: 2          sum LD: 81014.2788000001
site: 6  week: 3          sum LD: 76065.70359999998
site: 6  week: 4          sum LD: 73427.88070000023
site: 6  week: 5          sum LD: 70004.44240000001
site: 6  week: 6          sum LD: 70106.95649999985
site: 6  week: 7          sum LD: 67114.5973
site: 6  week: 8          sum LD: 69002.77129999992
site: 6  week: 9          sum LD: 72621.84069999975
site: 6  week: 10         sum LD: 63043.608699999924
site: 6  week: 11         sum LD: 57509.87500000002
site: 6  week: 12         sum LD: 62142.37510000003
site: 6  week: 13         sum LD: 56268.33180000011
site: 6  week: 14         sum LD: 57764.874999999905
site: 6  week: 15         sum LD: 55838.3174
site: 6  week: 16         sum LD: 57434.407099999975
site: 6  week: 17         sum LD: 61371.365699999835
site: 6  week: 18         sum LD: 57144.03630000009
site: 6  week: 19         sum LD: 58461.13940000005
site: 6  week: 20         sum LD: 61338.27580000002
site: 6  week: 21         sum LD: 58787.90820000011
site: 6  week: 22         sum LD: 59991.46150000001
site: 6  week: 23         sum LD: 57395.69489999985
site: 6  week: 24         sum LD: 54521.47599999996
site: 6  week: 25         sum LD: 57670.9434
site: 6  week: 26         sum LD: 62107.20630000006
site: 6  week: 27         sum LD: 62821.19369999994
site: 6  week: 28         sum LD: 66124.71800000002
site: 6  week: 29         sum LD: 64931.259799999956
```

```
site: 6  week: 30         sum LD: 65758.94880000006
site: 6  week: 31         sum LD: 64707.46150000002
site: 6  week: 32         sum LD: 67048.9683000001
site: 6  week: 33         sum LD: 67733.46490000004
site: 6  week: 34         sum LD: 62767.059199999974
site: 6  week: 35         sum LD: 65172.5603000001
site: 6  week: 36         sum LD: 60380.02450000004
site: 6  week: 37         sum LD: 59813.32800000012
site: 6  week: 38         sum LD: 58808.78120000002
site: 6  week: 39         sum LD: 58439.149800000065
site: 6  week: 40         sum LD: 56395.40389999992
site: 6  week: 41         sum LD: 55604.10419999997
site: 6  week: 42         sum LD: 54548.572999999946
site: 6  week: 43         sum LD: 52958.85189999996
site: 6  week: 44         sum LD: 51058.407100000026
site: 6  week: 45         sum LD: 53558.36379999997
site: 6  week: 46         sum LD: 62229.47159999993
site: 6  week: 47         sum LD: 59483.10930000006
site: 6  week: 48         sum LD: 57885.16380000005
site: 6  week: 49         sum LD: 59331.08529999999
site: 6  week: 50         sum LD: 64454.69609999995
site: 6  week: 51         sum LD: 72631.36310000008
site: 6  week: 52         sum LD: 72311.03349999996
site: 6  week: 53         sum LD: 23185.245899999998
site: 8  week: 0          sum LD: 27.9357
site: 8  week: 1          sum LD: 165833.21360000226
site: 8  week: 2          sum LD: 199761.07999999935
site: 8  week: 3          sum LD: 193803.78939999978
site: 8  week: 4          sum LD: 216076.93600000197
site: 8  week: 5          sum LD: 212590.53560000192
…..
```

4. Calculate the average LD by sector of activity (timestamp interval : a week)

```
db = db.getSiblingDB("enernoc") ;
print("*** Calculate the average LD by sector of activity (timestamp interval: a week)");
db.sites.aggregate([
        {$unwind : "$CONSO"},
        {$group :
            {
                _id :
                {
                    industry:"$INDUSTRY",week: {$week:"$CONSO.iso_date"}
                },
                avg_consu :
                {
                    $avg:"$CONSO.value"
                }
            }
        },
        {$sort:
            {
                _id:1
            }
        }
    ]).forEach( function(doc)
        {
            print("industry:",doc._id.industry,"\tweek:",doc._id.week,"\tsum LD:",doc.avg_consu)
        }
    );
```

results :

```
industry: Commercial Property    week: 0     sum LD: 69.906064
industry: Commercial Property    week: 1     sum LD: 81.8237099681427
industry: Commercial Property    week: 2     sum LD: 91.4887902052395
industry: Commercial Property    week: 3     sum LD: 94.48737770436513
industry: Commercial Property    week: 4     sum LD: 93.39917154365088
industry: Commercial Property    week: 5     sum LD: 91.20474614880696
industry: Commercial Property    week: 6     sum LD: 91.92764593452262
industry: Commercial Property    week: 7     sum LD: 92.71759324801408
industry: Commercial Property    week: 8     sum LD: 90.72024977579011
industry: Commercial Property    week: 9     sum LD: 91.01131014285522
industry: Commercial Property    week: 10    sum LD: 89.33322925991897
industry: Commercial Property    week: 11    sum LD: 90.10230121626647
industry: Commercial Property    week: 12    sum LD: 93.63770055753429
industry: Commercial Property    week: 13    sum LD: 89.08985976983729
industry: Commercial Property    week: 14    sum LD: 85.16025684920123
industry: Commercial Property    week: 15    sum LD: 83.3563983531696
industry: Commercial Property    week: 16    sum LD: 85.2688094126914
industry: Commercial Property    week: 17    sum LD: 84.9506929781688
industry: Commercial Property    week: 18    sum LD: 90.30943476387927
industry: Commercial Property    week: 19    sum LD: 90.11328439508686
industry: Commercial Property    week: 20    sum LD: 90.52765755158012
industry: Commercial Property    week: 21    sum LD: 91.70135157756431
industry: Commercial Property    week: 22    sum LD: 81.911028029759
industry: Commercial Property    week: 23    sum LD: 89.0749083068501
industry: Commercial Property    week: 24    sum LD: 94.56863999007561
industry: Commercial Property    week: 25    sum LD: 98.6916351726162
industry: Commercial Property    week: 26    sum LD: 97.41820529364593
industry: Commercial Property    week: 27    sum LD: 91.02007079703819
industry: Commercial Property    week: 28    sum LD: 101.13927387499722
industry: Commercial Property    week: 29    sum LD: 101.70688131349178
```

```
industry: Commercial Property       week: 30       sum LD: 101.90584847420102
industry: Commercial Property       week: 31       sum LD: 102.74853501983748
industry: Commercial Property       week: 32       sum LD: 102.19573332340961
industry: Commercial Property       week: 33       sum LD: 99.27649011309177
industry: Commercial Property       week: 34       sum LD: 91.24082630158397
industry: Commercial Property       week: 35       sum LD: 102.21700040872958
industry: Commercial Property       week: 36       sum LD: 91.34941175443164
industry: Commercial Property       week: 37       sum LD: 95.95129207936058
industry: Commercial Property       week: 38       sum LD: 90.06436633332761
industry: Commercial Property       week: 39       sum LD: 84.13700926785428
industry: Commercial Property       week: 40       sum LD: 87.05259087698066
industry: Commercial Property       week: 41       sum LD: 83.74384150793156
industry: Commercial Property       week: 42       sum LD: 85.74106816864568
industry: Commercial Property       week: 43       sum LD: 85.01873010515648
industry: Commercial Property       week: 44       sum LD: 81.48760832738208
industry: Commercial Property       week: 45       sum LD: 86.30356095237909
industry: Commercial Property       week: 46       sum LD: 85.05145403372674
industry: Commercial Property       week: 47       sum LD: 68.50369249167241
industry: Commercial Property       week: 48       sum LD: 84.34925334841124
industry: Commercial Property       week: 49       sum LD: 84.78240724007368
industry: Commercial Property       week: 50       sum LD: 86.72400969444112
industry: Commercial Property       week: 51       sum LD: 87.53539991467996
industry: Commercial Property       week: 52       sum LD: 67.4417095714268
industry: Commercial Property       week: 53       sum LD: 66.934555576387
industry: Education   week: 0      sum LD: 7.777043999999998
industry: Education   week: 1      sum LD: 10.035345357448085
industry: Education   week: 2      sum LD: 11.178326577381965
industry: Education   week: 3      sum LD: 11.035917702381193
industry: Education   week: 4      sum LD: 11.002216700397817
industry: Education   week: 5      sum LD: 11.05387033730232
...
```

SQL comparison

We have performed data integration into mysql database (see **mysql_1_data_integration.sql**) and compute the 4 Load Curve statistics on it (see **mysql_2_data_queries.sql**).

One car performe both by running bash file **run_mysql.sh.**

Here after, the table resumes time comparison between mongoDB and Mysql for the 4 Load Curve queries:

Time execution (s)	mysql	mongodb
Calculate the sum LD for the 100 sites (timestamp interval : 5 minutes)	6	5
Calculate the average LD by sector of activity (imestamp interval : 5 minutes)	9	12
Calculate the total LD for the 100 sites (timestamp interval : a week)	9	14
Calculate the average LD by sector of activity (timestamp interval : a week)	12	14

One can execute :

```
> ./run_mongo.sh > mongodb_results
```

to get time execution in mongodb_results file.

and execute :

```
>./run_mysql.sh >mysql_results
```

to get time execution in mysql_results file.