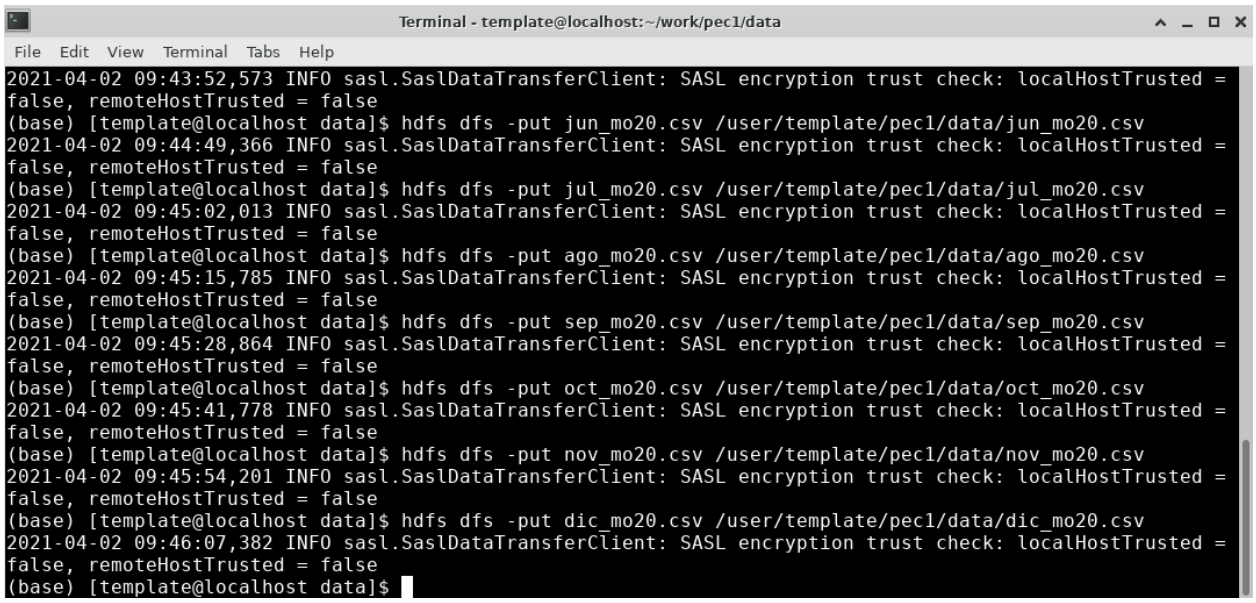
Práctica: PEC1 – Infr. Big Data, tratamiento batch
Nombre: Saulo Valdivia

1. Almacenamiento en HDFS

a) Archivo CSV

b) hdfs dfs -put ene_mo20.csv /user/template/pec1/data/ene_mo20.csv
hdfs dfs -put feb_mo20.csv /user/template/pec1/data/feb_mo20.csv
hdfs dfs -put mar_mo20.csv /user/template/pec1/data/mar_mo20.csv
hdfs dfs -put abr_mo20.csv /user/template/pec1/data/abr_mo20.csv
hdfs dfs -put may_mo20.csv /user/template/pec1/data/may_mo20.csv
hdfs dfs -put jun_mo20.csv /user/template/pec1/data/jun_mo20.csv
hdfs dfs -put jul_mo20.csv /user/template/pec1/data/jul_mo20.csv
hdfs dfs -put ago_mo20.csv /user/template/pec1/data/ago_mo20.csv
hdfs dfs -put sep_mo20.csv /user/template/pec1/data/sep_mo20.csv
hdfs dfs -put oct_mo20.csv /user/template/pec1/data/oct_mo20.csv
hdfs dfs -put nov_mo20.csv /user/template/pec1/data/nov_mo20.csv
hdfs dfs -put dec_mo20.csv /user/template/pec1/data/dec_mo20.csv



```
Terminal - template@localhost:~/work/pec1/data
File  Edit  View  Terminal  Tabs  Help
(base) [template@localhost data]$ hdfs dfs -put oct_mo20.csv /user/template/pec1/data/oct_mo20.csv
2021-04-02 09:45:41,778 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted =
false, remoteHostTrusted = false
(base) [template@localhost data]$ hdfs dfs -put nov_mo20.csv /user/template/pec1/data/nov_mo20.csv
2021-04-02 09:45:54,201 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted =
false, remoteHostTrusted = false
(base) [template@localhost data]$ hdfs dfs -put dic_mo20.csv /user/template/pec1/data/dic_mo20.csv
2021-04-02 09:46:07,382 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted =
false, remoteHostTrusted = false
(base) [template@localhost data]$ hdfs dfs -ls /user/template/pec1/data
Found 12 items
-rw-r--r--   1 template supergroup    1020042 2021-04-02 09:43 /user/template/pec1/data/abr_mo20.csv
-rw-r--r--   1 template supergroup    1084937 2021-04-02 09:45 /user/template/pec1/data/ago_mo20.csv
-rw-r--r--   1 template supergroup    1083812 2021-04-02 09:46 /user/template/pec1/data/dic_mo20.csv
-rw-r--r--   1 template supergroup    1076310 2021-04-02 09:42 /user/template/pec1/data/ene_mo20.csv
-rw-r--r--   1 template supergroup    1017548 2021-04-02 09:43 /user/template/pec1/data/feb_mo20.csv
-rw-r--r--   1 template supergroup    1087706 2021-04-02 09:45 /user/template/pec1/data/jul_mo20.csv
-rw-r--r--   1 template supergroup    1042551 2021-04-02 09:44 /user/template/pec1/data/jun_mo20.csv
-rw-r--r--   1 template supergroup    1054675 2021-04-02 09:43 /user/template/pec1/data/mar_mo20.csv
-rw-r--r--   1 template supergroup    1081063 2021-04-02 09:43 /user/template/pec1/data/may_mo20.csv
-rw-r--r--   1 template supergroup    1051488 2021-04-02 09:45 /user/template/pec1/data/nov_mo20.csv
-rw-r--r--   1 template supergroup    1087706 2021-04-02 09:45 /user/template/pec1/data/oct_mo20.csv
-rw-r--r--   1 template supergroup    1048962 2021-04-02 09:45 /user/template/pec1/data/sep_mo20.csv
(base) [template@localhost data]$
```

c) hdfs dfs -du -s -h /user/template/pec1/data

```
Terminal - template@localhost:~/work/pec1/data
File  Edit  View  Terminal  Tabs  Help
(base) [template@localhost data]$ hdfs dfs -du -s -h /user/template/pec1/data
12.1 M  12.1 M  /user/template/pec1/data
(base) [template@localhost data]$
```

d) hdfs fsck /user/template/pec1/data/ene_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/feb_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/mar_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/abr_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/may_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/jun_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/jul_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/ago_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/sep_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/oct_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/nov_mo20.csv -files -blocks
hdfs fsck /user/template/pec1/data/dic_mo20.csv -files -blocks

```
Terminal - template@localhost:~/work/pec1/data
File  Edit  View  Terminal  Tabs  Help
(base) [template@localhost data]$ hdfs fsck /user/template/pec1/data/ene_mo20.csv -files -blocks
Connecting to namenode via http://localhost:9870/fsck?ugi=template&files=1&blocks=1&path=%2Fuser%2Fte
mplate%2Fpec1%2Fdata%2Fene_mo20.csv
FSCK started by template (auth:SIMPLE) from /127.0.0.1 for path /user/template/pec1/data/ene_mo20.csv
 at Fri Apr 02 09:50:16 UTC 2021
/user/template/pec1/data/ene_mo20.csv 1076310 bytes, replicated: replication=1, 1 block(s):  OK
0. BP-1122312928-127.0.0.1-1595758942714:blk_1073742447_1626 len=1076310 Live_repl=1


Status: HEALTHY
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    0
 Total symlinks:                0

Replicated Blocks:
 Total size:    1076310 B
 Total files:   1
 Total blocks (validated):      1 (avg. block size 1076310 B)
 Minimally replicated blocks:   1 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
 Total block groups (validated):        0
 Minimally erasure-coded block groups:  0
 Over-erasure-coded block groups:       0
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
FSCK ended at Fri Apr 02 09:50:16 UTC 2021 in 6 milliseconds


The filesystem under path '/user/template/pec1/data/ene_mo20.csv' is HEALTHY
```

2. Tratamiento con Hive

a) Creamos una Tabla Externa que permita mapear los datos HDFS

```
create external table MadridData(
PROVINCIA STRING, MUNICIPIO STRING, ESTACION STRING, MAGNITUD STRING,
PUNTO_MUESTREO STRING, ANO STRING, MES STRING, DIA STRING,
H01 INT, V01 STRING, H02 INT, V02 STRING, H03 INT, V03 STRING, H04 INT, V04 STRING,
H05 INT, V05 STRING, H06 INT, V06 STRING, H07 INT, V07 STRING, H08 INT, V08 STRING,
H09 INT, V09 STRING, H10 INT, V10 STRING, H11 INT, V11 STRING, H12 INT, V12 STRING,
H13 INT, V13 STRING, H14 INT, V14 STRING, H15 INT, V15 STRING, H16 INT, V16 STRING,
H17 INT, V17 STRING, H18 INT, V18 STRING, H19 INT, V19 STRING, H20 INT, V20 STRING,
H21 INT, V21 STRING, H22 INT, V22 STRING, H23 INT, V23 STRING, H24 INT, V24 STRING
)
row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
with serdeproperties (
        "separatorChar" = ";"
);
```

```
0: jdbc:hive2://> create external table MadridData(
. . . . . > PROVINCIA STRING,
. . . . . > MUNICIPIO STRING,
. . . . . > ESTACION STRING,
. . . . . > MAGNITUD STRING,
. . . . . > PUNTO_MUESTREO STRING,
. . . . . > ANO STRING,
. . . . . > MES STRING,
. . . . . > DIA STRING,
. . . . . > H01 INT,
. . . . . > V01 STRING,
. . . . . > H02 INT,
. . . . . > V02 STRING,
. . . . . > H03 INT,
. . . . . > V03 STRING,
. . . . . > H04 INT,
. . . . . > V04 STRING,
. . . . . > H05 INT,
. . . . . > V05 STRING,
. . . . . > H06 INT,
. . . . . > V06 STRING,
. . . . . > H07 INT,
. . . . . > V07 STRING,
. . . . . > H08 INT,
. . . . . > V08 STRING,
. . . . . > H09 INT,
. . . . . > V09 STRING,
. . . . . > H10 INT,
. . . . . > V10 STRING,
. . . . . > H11 INT,
. . . . . > V11 STRING,
. . . . . > H12 INT,
. . . . . > V12 STRING,
. . . . . > H13 INT,
. . . . . > V13 STRING,
. . . . . > H14 INT,
. . . . . > V14 STRING,
. . . . . > H15 INT,
. . . . . > V15 STRING,
. . . . . > H16 INT,
. . . . . > V16 STRING,
. . . . . > H17 INT,
. . . . . > V17 STRING,
. . . . . > H18 INT,
. . . . . > V18 STRING,
. . . . . > H19 INT,
. . . . . > V19 STRING,
. . . . . > H20 INT,
. . . . . > V20 STRING,
```

Realizamos la carga de la tabla con los ficheros .csv

```
hdfs dfs -put ene_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put feb_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put mar_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put abr_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put may_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put jun_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put jul_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put ago_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put sep_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put oct_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put nov_mo20.csv /user/hive/warehouse/madriddata
hdfs dfs -put dic_mo20.csv /user/hive/warehouse/madriddata
```

b) Select ANO, MES, DIA, AVG(H12) as S02 from madriddata group by ANO, MES, DIA;

```
0: jdbc:hive2://> select ANO, MES, DIA, AVG(H12) as S02 from madriddata group by ANO, MES, DIA;
21/04/01 16:44:02 [HiveServer2-Background-Pool: Thread-75]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2
g a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = template_20210401164402_6a7ee4b4-1251-40dd-9e90-5ab56cb236b6
Total jobs = 1
Launching Job 1 out of 1
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>
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a di
eleases.
21/04/01 16:44:03 [HiveServer2-Background-Pool: Thread-75]: WARN mapreduce.JobResourceUploader: Hadoop command-
e and execute your application with ToolRunner to remedy this.
Starting Job = job_1617295184041_0002, Tracking URL = http://localhost:8088/proxy/application_1617295184041_000
Kill Command = /usr/local/hadoop-3.2.1/bin/mapred job  -kill job_1617295184041_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
21/04/01 16:44:15 [HiveServer2-Background-Pool: Thread-75]: WARN mapreduce.Counters: Group org.apache.hadoop.ma
ce.TaskCounter instead
2021-04-01 16:44:15,436 Stage-1 map = 0%,  reduce = 0%
```

```
File  Edit  View  Terminal  Tabs  Help
OK
+-------+------+------+---------------------+
|  ano  | mes  | dia  |         s02         |
+-------+------+------+---------------------+
| 2020  | 01   | 01   | 28.972287581699348  |
| 2020  | 01   | 02   | 52.253006535947726  |
| 2020  | 01   | 03   | 60.904509803921556  |
| 2020  | 01   | 04   | 19.488104575163405  |
| 2020  | 01   | 05   | 25.80869281045752   |
| 2020  | 01   | 06   | 28.921699346405227  |
| 2020  | 01   | 07   | 67.73091503267972   |
| 2020  | 01   | 08   | 72.93699346405229   |
| 2020  | 01   | 09   | 79.97960784313725   |
| 2020  | 01   | 10   | 24.06843137254902   |
| 2020  | 01   | 11   | 21.77509803921569   |
| 2020  | 01   | 12   | 30.438310810810812  |
| 2020  | 01   | 13   | 74.75470588235295   |
| 2020  | 01   | 14   | 53.505490196078426  |
| 2020  | 01   | 15   | 57.63313725490197   |
| 2020  | 01   | 16   | 44.75601307189541   |
| 2020  | 01   | 17   | 18.63098039215686   |
| 2020  | 01   | 18   | 29.4581045751634    |
| 2020  | 01   | 19   | 12.561111111111112  |
| 2020  | 01   | 20   | 18.49117647058823   |
| 2020  | 01   | 21   | 15.324324324324328  |
| 2020  | 01   | 22   | 25.229999999999997  |
| 2020  | 01   | 23   | 31.71831081081082   |
| 2020  | 01   | 24   | 25.127499999999998  |
| 2020  | 01   | 25   | 19.02662162162162   |
| 2020  | 01   | 26   | 28.36878378378379   |
| 2020  | 01   | 27   | 40.87695945945946   |
| 2020  | 01   | 28   | 17.039527027027027  |
```

c)

Habilitamos la partición dinámica

set hive.exec.dynamic.partition=true;

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

```
0: jdbc:hive2://> set hive.exec.dynamic.partition=true;
No rows affected (0.142 seconds)
0: jdbc:hive2://> set hive.exec.dynamic.partition.mode=nonstrict;
No rows affected (0.011 seconds)
0: jdbc:hive2://>
```

Creamos la tabla con partición dinámica
create table medidasAire_part1(
PROVINCIA STRING, MUNICIPIO STRING, ESTACION STRING, MAGNITUD STRING,
PUNTO_MUESTREO STRING, DIA STRING,
H01 INT, V01 STRING, H02 INT, V02 STRING, H03 INT, V03 STRING, H04 INT, V04 STRING,
H05 INT, V05 STRING, H06 INT, V06 STRING, H07 INT, V07 STRING, H08 INT, V08 STRING,
H09 INT, V09 STRING, H10 INT, V10 STRING, H11 INT, V11 STRING, H12 INT, V12 STRING,
H13 INT, V13 STRING, H14 INT, V14 STRING, H15 INT, V15 STRING, H16 INT, V16 STRING,
H17 INT, V17 STRING, H18 INT, V18 STRING, H19 INT, V19 STRING, H20 INT, V20 STRING,
H21 INT, V21 STRING, H22 INT, V22 STRING, H23 INT, V23 STRING, H24 INT, V24 STRING)
PARTITIONED BY(ANO STRING, MES STRING);

```
0: jdbc:hive2://> create table medidasAire_part1(
. . . . . . . . > PROVINCIA STRING, MUNICIPIO STRING, ESTACION STRING, MAGNITUD STRING,
. . . . . . . . > PUNTO_MUESTREO STRING, DIA STRING,
. . . . . . . . > H01 INT, V01 STRING, H02 INT, V02 STRING, H03 INT, V03 STRING, H04 INT, V04 STRING,
. . . . . . . . > H05 INT, V05 STRING, H06 INT, V06 STRING, H07 INT, V07 STRING, H08 INT, V08 STRING,
. . . . . . . . > H09 INT, V09 STRING, H10 INT, V10 STRING, H11 INT, V11 STRING, H12 INT, V12 STRING,
. . . . . . . . > H13 INT, V13 STRING, H14 INT, V14 STRING, H15 INT, V15 STRING, H16 INT, V16 STRING,
. . . . . . . . > H17 INT, V17 STRING, H18 INT, V18 STRING, H19 INT, V19 STRING, H20 INT, V20 STRING,
. . . . . . . . > H21 INT, V21 STRING, H22 INT, V22 STRING, H23 INT, V23 STRING, H24 INT, V24 STRING)
. . . . . . . . > PARTITIONED BY(ANO STRING, MES STRING);
OK
No rows affected (0.301 seconds)
0: jdbc:hive2://>
```

Insertamos los datos a la tabla
INSERT OVERWRITE TABLE medidasAire_part1
PARTITION(ANO, MES)
SELECT PROVINCIA,MUNICIPIO,ESTACION,MAGNITUD,PUNTO_MUESTREO,DIA,
H01,V01,H02,V02,H03,V03,H04,V04,H05,V05,H06,V06,H07,V07,H08,V08,H09,V09,H10,
V10,H11,V11,H12,V12,H13,V13,H14,V14,H15,V15,H16,V16,H17,V17,H18,V18,H19,V19,
H20,V20,H21,V21,H22,V22,H23,V23,H24,V24,ANO,MES from madriddata;

```
0: jdbc:hive2://> INSERT OVERWRITE TABLE medidasAire_part1
. . . . . . . . > PARTITION(ANO, MES)
. . . . . . . . > SELECT PROVINCIA,MUNICIPIO,ESTACION,MAGNITUD,PUNTO_MUESTREO,DIA,
. . . . . . . . > H01,V01,H02,V02,H03,V03,H04,V04,H05,V05,H06,V06,H07,V07,H08,V08,H09,V09,H10,
. . . . . . . . > V10,H11,V11,H12,V12,H13,V13,H14,V14,H15,V15,H16,V16,H17,V17,H18,V18,H19,V19,
. . . . . . . . > H20,V20,H21,V21,H22,V22,H23,V23,H24,V24,ANO,MES from madriddata;
21/04/02 17:06:00 [9aa4c70c-058a-4814-a0cc-a86950cb6dc9 main]: WARN parse.BaseSemanticAnalyzer: Dynamic partitioning is used;
 only validating 0 columns
```

Podemos ver que las particiones se crearon exitosamente.

```
0: jdbc:hive2://> show partitions medidasAire_part1;
OK
+------------------+
|    partition     |
+------------------+
| ano=2020/mes=01  |
| ano=2020/mes=02  |
| ano=2020/mes=03  |
| ano=2020/mes=04  |
| ano=2020/mes=05  |
| ano=2020/mes=06  |
| ano=2020/mes=07  |
| ano=2020/mes=08  |
| ano=2020/mes=09  |
| ano=2020/mes=10  |
| ano=2020/mes=11  |
| ano=2020/mes=12  |
| ano=ANO/mes=MES  |
+------------------+
13 rows selected (0.436 seconds)
```

d) Realizamos la creación de una tabla para alojar los resultados del query.

```
create table queryHive ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED as textfile AS SELECT
'ANO' as ANO ,'MES' as MES ,'DIA' as DIA ,'MAGNITUD' as MAGNITUD
,'H01' as H01,'H02' as H02,'H03' as H03,'H04' as H04,'H05' as H05,'H06' as H06
,'H07' as H07,'H08' as H08,'H09' as H09,'H10' as H10,'H11' as H11,'H12' as H12
,'H13' as H13,'H14' as H14,'H15' as H15,'H16' as H16,'H17' as H17,'H18' as H18
,'H19' as H19,'H20' as H20,'H21' as H21,'H22' as H22,'H23' as H23,'H24' as H24;
```

```
0: jdbc:hive2://> create table queryHive ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
. . . . . . . . > LINES TERMINATED BY '\n'
. . . . . . . . > STORED as textfile AS SELECT
. . . . . . . . > 'ANO' as ANO ,'MES' as MES ,'DIA' as DIA ,'MAGNITUD' as MAGNITUD
. . . . . . . . > ,'H01' as H01,'H02' as H02,'H03' as H03,'H04' as H04,'H05' as H05,'H06' as H06
. . . . . . . . > ,'H07' as H07,'H08' as H08,'H09' as H09,'H10' as H10,'H11' as H11,'H12' as H12
. . . . . . . . > ,'H13' as H13,'H14' as H14,'H15' as H15,'H16' as H16,'H17' as H17,'H18' as H18
. . . . . . . . > ,'H19' as H19,'H20' as H20,'H21' as H21,'H22' as H22,'H23' as H23,'H24' as H24;
21/04/02 17:17:25 [HiveServer2-Background-Pool: Thread-130]: WARN ql.Driver: Hive-on-MR is deprecated
e available in the future versions. Consider using a different execution engine (i.e. spark, tez) or u
Query ID = template_20210402171724_07dc6b76-9a63-4481-95ca-41f05dd2a36a
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider u
ion engine (i.e. spark, tez) or using Hive 1.X releases.
21/04/02 17:17:26 [HiveServer2-Background-Pool: Thread-130]: WARN mapreduce.JobResourceUploader: Hadoo
arsing not performed. Implement the Tool interface and execute your application with ToolRunner to rem
```

Hacemos el insert a esta tabla con los resultados del query.

INSERT INTO queryHive
SELECT ANO, MES, DIA, MAGNITUD,
avg(H01) AS H01, avg(H02) AS H02, avg(H03) AS H03, avg(H04) AS H04,
avg(H05) AS H05, avg(H06) AS H06, avg(H07) AS H07, avg(H08) AS H08,
avg(H09) AS H09, avg(H10) AS H10, avg(H11) AS H11, avg(H12) AS H12,
avg(H13) AS H13, avg(H14) AS H14, avg(H15) AS H15, avg(H16) AS H16,
avg(H17) AS H17, avg(H18) AS H18, avg(H19) AS H19, avg(H20) AS H20,
avg(H21) AS H21, avg(H22) AS H22, avg(H23) AS H23, avg(H24) AS H24
from madriddata group by ANO, MES, DIA, MAGNITUD;

```
0: jdbc:hive2://> INSERT INTO queryHive
. . . . . . . . > SELECT ANO, MES, DIA, MAGNITUD,
. . . . . . . . > avg(H01) AS H01, avg(H02) AS H02, avg(H03) AS H03, avg(H04) AS H04,
. . . . . . . . > avg(H05) AS H05, avg(H06) AS H06, avg(H07) AS H07, avg(H08) AS H08,
. . . . . . . . > avg(H09) AS H09, avg(H10) AS H10, avg(H11) AS H11, avg(H12) AS H12,
. . . . . . . . > avg(H13) AS H13, avg(H14) AS H14, avg(H15) AS H15, avg(H16) AS H16,
. . . . . . . . > avg(H17) AS H17, avg(H18) AS H18, avg(H19) AS H19, avg(H20) AS H20,
. . . . . . . . > avg(H21) AS H21, avg(H22) AS H22, avg(H23) AS H23, avg(H24) AS H24
. . . . . . . . > from madriddata group by ANO, MES, DIA, MAGNITUD;
21/04/02 17:19:24 [9aa4c70c-058a-4814-a0cc-a86950cb6dc9 main]: WARN metastore.ObjectStore:
 is set to unsupported value null . Setting it to value: ignored
21/04/02 17:19:25 [HiveServer2-Background-Pool: Thread-155]: WARN ql.Driver: Hive-on-MR is
e available in the future versions. Consider using a different execution engine (i.e. spark
Query ID = template_20210402171924_22913168-a2fa-46b8-85cd-fdd3f359bfa6
Total jobs = 2
Launching Job 1 out of 2
```

Exportamos el contenido de la tabla a un archivo .csv

hdfs dfs -cat /user/hive/warehouse/queryhive/* > ~/queryHive.csv

```
(base) [template@localhost ~]$ hdfs dfs -cat /user/hive/warehouse/queryhive/* > ~/queryHive.csv
2021-04-02 17:23:42,457 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTr
usted = false, remoteHostTrusted = false
```

Terminal - template@localhost:~

File Edit View Terminal Tabs Help

```
(base) [template@localhost ~]$ ls
codpostales.csv      Downloads          miniconda.sh          output.csv     prueba.txt
 spark-warehouse  work-bigdata
comandos_usados.txt  gente.csv          mongodata             pec1.csv       Public
 Templates
derby.log            log-generator.py   Music                 pesadilla.txt  pueblos2.csv
 users.json
Desktop              metastore_db       network_wordcount.py  Pictures       pueblos.csv
 Videos
Documents            miniconda3         output1.csv                          queryHive.csv
 work
(base) [template@localhost ~]$
```

3. Tratamiento con MongoDB

a) Para almacenar los datos procesados en el punto anterior necesitamos una estructura en formato JSON. Específicamente una Colección/Array de elementos. Esto nos permitirá importar los datos a MongoDB.


```
Viewer | Text
[] JSON
  {} 0
  {} 1
    PROVINCIA : "28"
    MUNICIPIO : "079"
    ESTACION : "4"
    MAGNITUD : "1"
    PUNTO_MUESTREO : "28079004_1_38"
    ANO : "2020"
    MES : "01"
    DIA : "01"
    H01 : 1
    V01 : "V"
    H02 : 2
    V02 : "V"
    H03 : 3
    V03 : "V"
    H04 : 4
    V04 : "V"
    H05 : 5
    V05 : "V"
    H06 : 6
    V06 : "V"
    H07 : 7
    V07 : "V"
    H08 : 8
    V08 : "V"
    H09 : 9
    V09 : "V"
    H10 : 10
    V10 : "V"
    H11 : 11
```

b) Paso1. Exportar datos de Hive a CSV.

- Creamos una tabla para guardar todos los datos

```
Terminal - template@localhost:~
File  Edit  View  Terminal  Tabs  Help
0: jdbc:hive2://> CREATE TABLE table_csv_export_data_pec1
. . . . . . . . > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
. . . . . . . . > LINES TERMINATED BY '\n'
. . . . . . . . > STORED as textfile
. . . . . . . . > AS
. . . . . . . . > select
. . . . . . . . > 'PROVINCIA' as PROVINCIA
. . . . . . . . > ,'MUNICIPIO' as MUNICIPIO
. . . . . . . . > ,'ESTACION' as ESTACION
. . . . . . . . > ,'ANO' as ANO
. . . . . . . . > ,'MES' as MES
. . . . . . . . > ,'DIA' as DIA
. . . . . . . . > ,'H01' as H01
. . . . . . . . > ,'V01' as V01
. . . . . . . . > ,'H02' as H02
. . . . . . . . > ,'V02' as V02
. . . . . . . . > ,'H03' as H03
. . . . . . . . > ,'V03' as V03
. . . . . . . . > ,'H04' as H04
. . . . . . . . > ,'V04' as V04
. . . . . . . . > ,'H05' as H05
. . . . . . . . > ,'V05' as V05
. . . . . . . . > ,'H06' as H06
. . . . . . . . > ,'V06' as V06
. . . . . . . . > ,'H07' as H07
. . . . . . . . > ,'V07' as V07
. . . . . . . . > ,'H08' as H08
. . . . . . . . > ,'V08' as V08
. . . . . . . . > ,'H09' as H09
. . . . . . . . > ,'V09' as V09
. . . . . . . . > ,'H10' as H10
. . . . . . . . > ,'V10' as V10
. . . . . . . . > ,'H11' as H11
. . . . . . . . > ,'V11' as V11
. . . . . . . . > ,'H12' as H12
. . . . . . . . > ,'V12' as V12
. . . . . . . . > ,'H13' as H13
. . . . . . . . > ,'V13' as V13
. . . . . . . . > ,'H14' as H14
. . . . . . . . > ,'V14' as V14
. . . . . . . . > ,'H15' as H15
. . . . . . . . > ,'V15' as V15
. . . . . . . . > ,'H16' as H16
. . . . . . . . > ,'V16' as V16
. . . . . . . . > ,'H17' as H17
```

- Realizamos el Insert

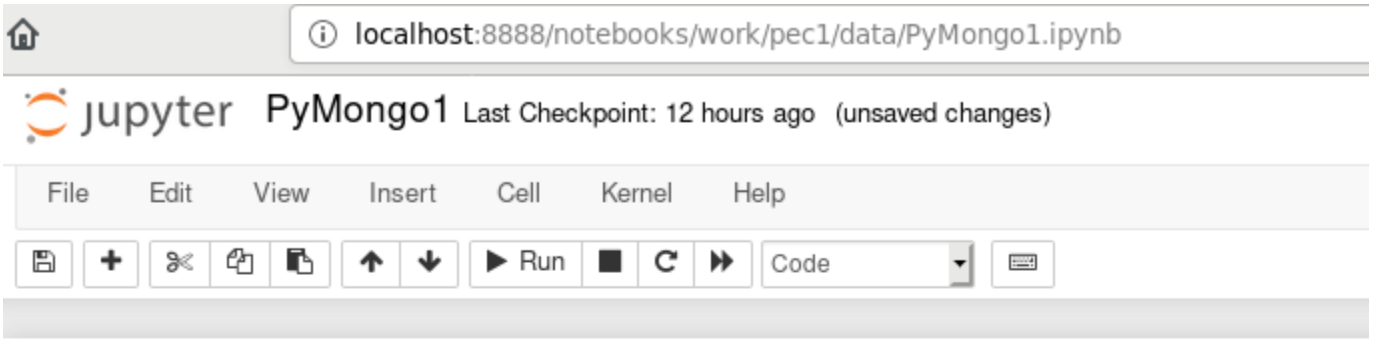
```
Terminal - template@localhost:~
File  Edit  View  Terminal  Tabs  Help
0: jdbc:hive2://> INSERT INTO table_csv_export_data_pec1
. . . . . . . . . > SELECT
. . . . . . . . . >  PROVINCIA
. . . . . . . . . >  ,MUNICIPIO
. . . . . . . . . >  ,ESTACION
. . . . . . . . . >  ,ANO
. . . . . . . . . >  ,MES
. . . . . . . . . >  ,DIA
. . . . . . . . . >  ,H01
. . . . . . . . . >  ,V01
. . . . . . . . . >  ,H02
. . . . . . . . . >  ,V02
. . . . . . . . . >  ,H03
. . . . . . . . . >  ,V03
. . . . . . . . . >  ,H04
. . . . . . . . . >  ,V04
. . . . . . . . . >  ,H05
. . . . . . . . . >  ,V05
. . . . . . . . . >  ,H06
. . . . . . . . . >  ,V06
. . . . . . . . . >  ,H07
. . . . . . . . . >  ,V07
. . . . . . . . . >  ,H08
. . . . . . . . . >  ,V08
. . . . . . . . . >  ,H09
. . . . . . . . . >  ,V09
. . . . . . . . . >  ,H10
. . . . . . . . . >  ,V10
. . . . . . . . . >  ,H11
. . . . . . . . . >  ,V11
. . . . . . . . . >  ,H12
. . . . . . . . . >  ,V12
. . . . . . . . . >  ,H13
. . . . . . . . . >  ,V13
. . . . . . . . . >  ,H14
. . . . . . . . . >  ,V14
. . . . . . . . . >  ,H15
. . . . . . . . . >  ,V15
. . . . . . . . . >  ,H16
. . . . . . . . . >  ,V16
. . . . . . . . . >  ,H17
. . . . . . . . . >  ,V17
. . . . . . . . . >  ,H18
. . . . . . . . . >  ,V18
. . . . . . . . . >  ,H19
```

- Realizamos la exportación
  hdfs dfs -cat /user/hive/warehouse/table_csv_export_data_pec1/* > ~/pec1.csv

```
Terminal - template@localhost:~
File  Edit  View  Terminal  Tabs  Help
(base) [template@localhost ~]$ hdfs dfs -cat /user/hive/warehouse/table_csv_export_data_pec1/* > ~/pec1.cs
2021-04-02 10:28:43,099 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted =
remoteHostTrusted = false
(base) [template@localhost ~]$ ls
codpostales.csv      Downloads         miniconda.sh          output.csv      prueba.txt       Templates
comandos_usados.txt  gente.csv         mongodata             pec1.csv        Public           users.json
derby.log            log-generator.py  Music                 pesadilla.txt   pueblos2.csv     Videos
Desktop              metastore_db      network_wordcount.py  Pictures        pueblos.csv      work
Documents            miniconda3        output1.csv           prueba2.txt     spark-warehouse  work-bigdata
(base) [template@localhost ~]$
```

Paso 2.

- Realizamos un script con PyMongo



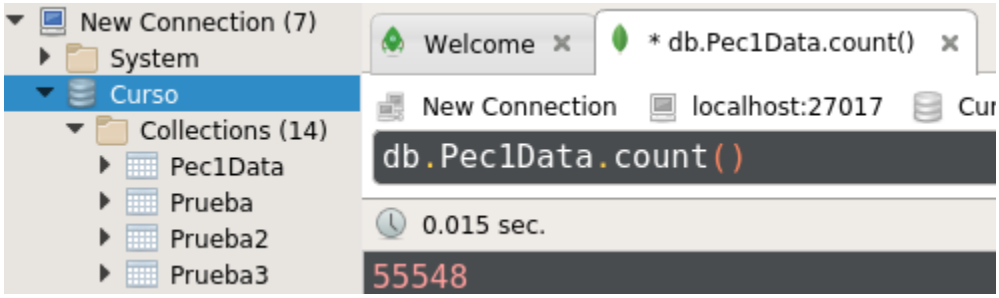
In [5]:
```
from pymongo import MongoClient
import pandas as pd
import csv

# DB connectivity
client = MongoClient('localhost', 27017)
database = client['Curso']
collection = database['Pec1Data']

# Function to parse csv to dictionary
def csv_to_json(filename):
    data = pd.read_csv(filename)
    return data.to_dict('records')

# Final insert statement
collection.insert_many(csv_to_json('pec1.csv'))
```

Out[5]:
```
<pymongo.results.InsertManyResult at 0x7fd99d7ff640>
```



c)

Realizamos la Query en MongoDB

```
db.Pec1.aggregate(
   [
     {"$match":{"MES":{"$gte":1, "$lte":12}}},
     {
       $group:
         {
           _id: {
            "ANO": "$ANO",
            "MES": "$MES",
            "DIA": "$DIA",
            "MAGNITUD": "$MAGNITUD"
            },
            Media: { $avg:{ $add: [ "$H01", "$H02", "$H03", "$H04", "$H05", "$H06", "$H07", "$H08", "$H09", "$H10", "$H11", "$H12",
                                    "$H13", "$H14", "$H15", "$H16", "$H17", "$H18", "$H19", "$H20", "$H21", "$H22", "$H23", "$H24"] } }
         }
     }
   ])
```

```
/* 1 */
{
    "_id" : {
        "ANO" : 2020,
        "MES" : 5,
        "DIA" : 10
    },
    "Media" : 221.501895424837
}

/* 2 */
{
    "_id" : {
        "ANO" : 2020,
        "MES" : 11,
        "DIA" : 8
    },
    "Media" : 447.855098039216
```

4. Arquitectura Big Data

1. La arquitectura que mejor se adapta es la arquitectura Lambda. Debido a que no existe una gran sobrecarga de información que requiera una arquitectura específica para Streaming. Otro motivo por el cual se justifica la arquitectura lambda es debido a que tenemos un repositorio central con todos los datos históricos al menos de un año.

2. Diagrama de Arquitectura