PROJET DE SESSION : HADOOP

Yacine BELHOUL
Abdellilah NAFIA
Cadrick NOUTCHA

Année 2016

BESOIN D'AFFAIRES

Analyser les trajets effectués par les taxis de la ville de New York durant le premier semestre de l'année 2015 :

Q 1 : La journée où il y a plus de trafic

Q 2 : L'heure de pointe

Q3 : La journée de la semaine de pointe

DATASET

Déplacements des taxis de New York City

(Yellow Taxi Trip Data)

- Les coordonnées de départ et d'arrivée, date de départ et d'arrivée, distances, type de payement, nombre de passagers, ….
- Contient 77 080 575 d'enregistrements et qui occupe 10 GO d'espace dans le disque dur.
- Lien de téléchargement: https://data.cityofnewyork.us/view/ba8s-jw6u

CLUSTER

Caractéristiques

- **Host 1 : Master Node**
  - RAM 9 GO
  - Processeur 4 CPU Core TM i7-4770 CPU 3.40 Ghz
  - DDR 25 GO
- **Host 2 : Data Node** :
  - RAM 4 GO
  - Processeur 2 CPU Core TM i7-4770 CPU 3.40 Ghz
  - DDR 25 GO

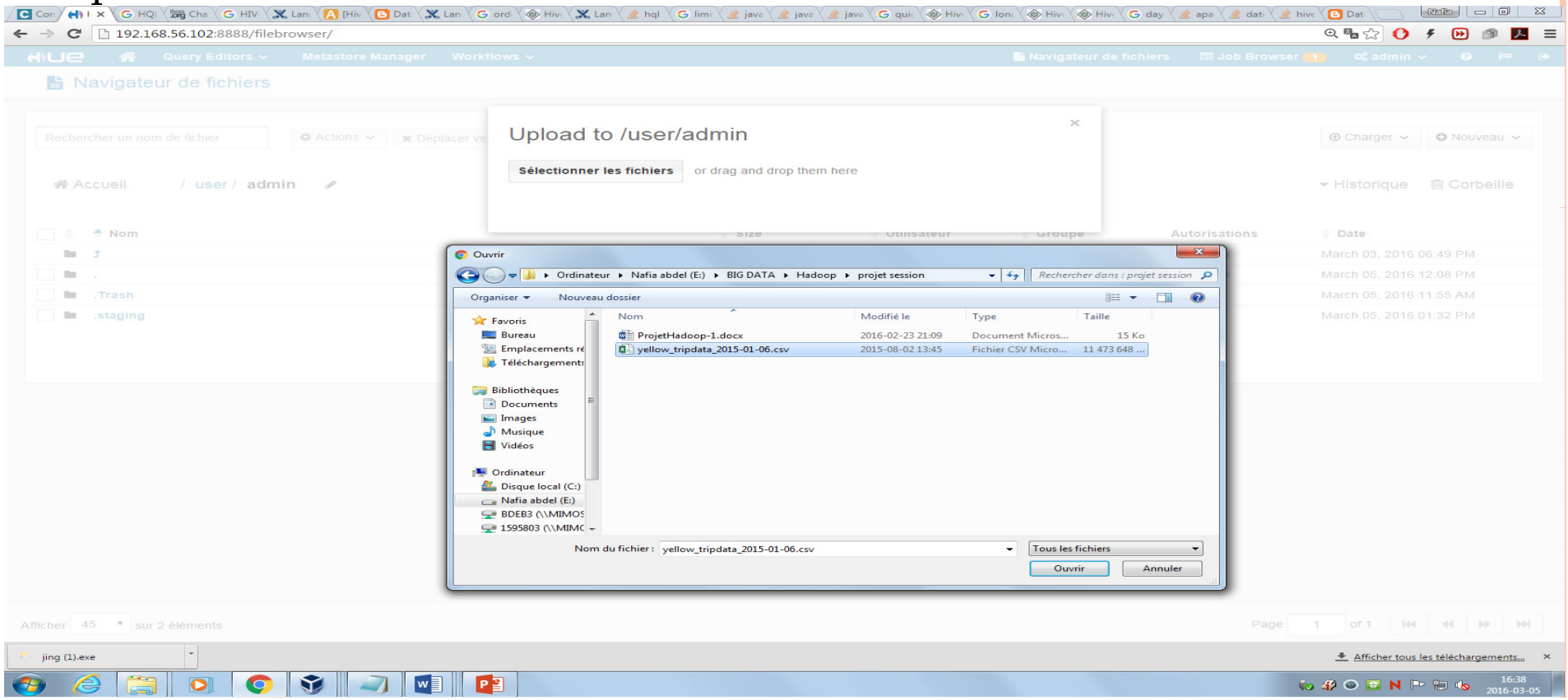
Administration

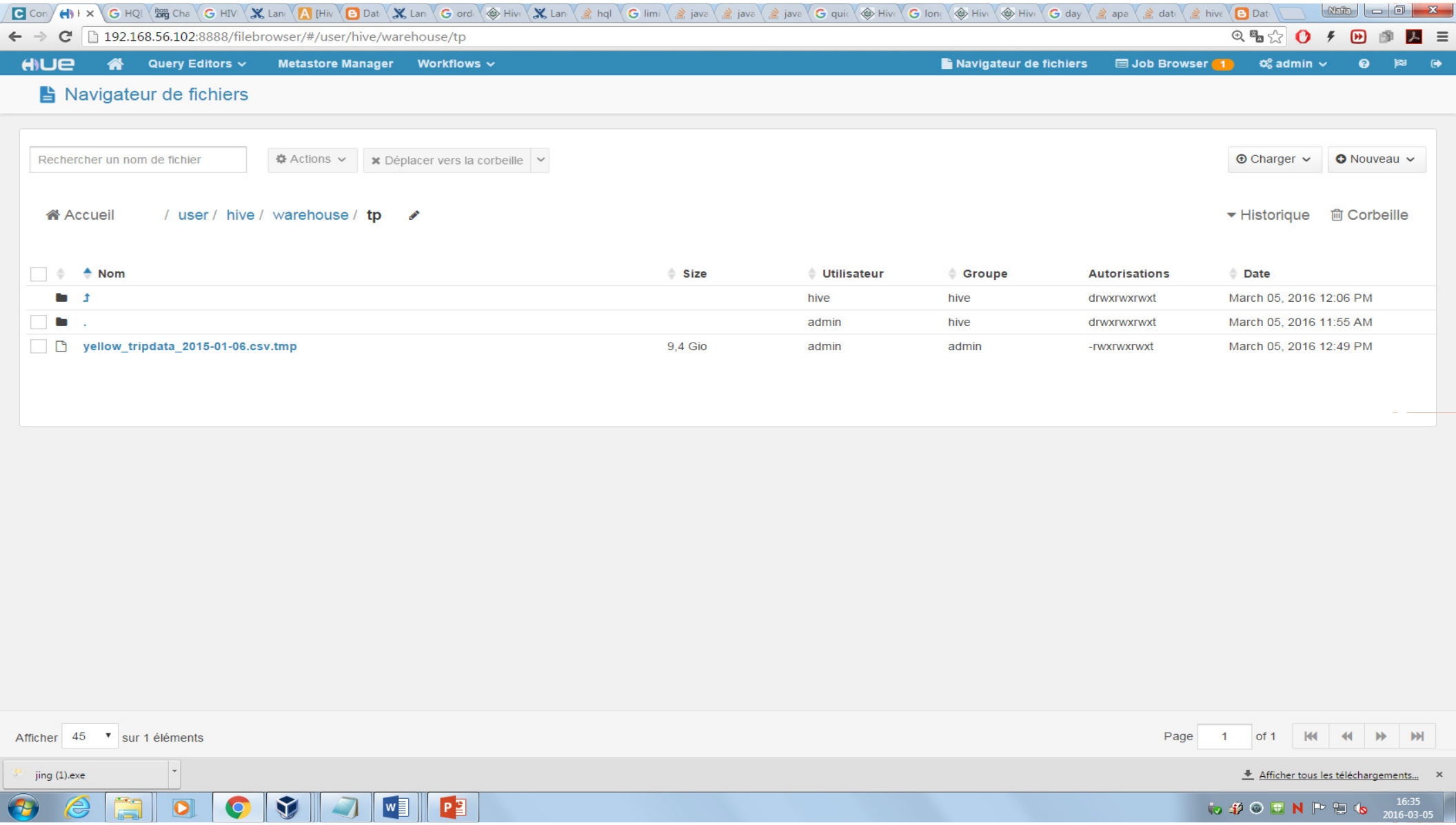
- **Master Node** (NameNode+DataNode) :
  - Centos 6.4 server
  - Cloudera Manager Essai de Cloudera Enterprise Data Hub Edition 5.5.3
  - Rôles : Hdfs, yarn, **Hive**, pig, Hue, Oozie, …
- **Data Node** :
  - Centos 6.4 server
  - Cloudera Agent
  - Rôles: Hdfs, yarn.

Arrêter Cloudera Manager Service

1- PRÉPARATION DE L'ENVIRONNEMENT

- Copier le fichier de données via Hue




192.168.56.102:8888/filebrowser/#/user/hive/warehouse/tp
Query Editors
Metastore Manager
Workflows
Navigateur de fichiers
Job Browser 1
admin
Navigateur de fichiers
Rechercher un nom de fichier
Actions
Déplacer vers la corbeille
Charger
Nouveau
Accueil / user / hive / warehouse / tp
Historique
Corbeille
Nom
Size
Utilisateur
Groupe
Autorisations
Date
hive
hive
drwxrwxrwxt
March 05, 2016 12:06 PM
.
admin
hive
drwxrwxrwxt
March 05, 2016 11:55 AM
yellow_tripdata_2015-01-06.csv.tmp
9,4 Gio
admin
admin
-rwxrwxrwxt
March 05, 2016 12:49 PM
Afficher 45 sur 1 éléments
Page 1 of 1
jing (1).exe
Afficher tous les téléchargements...
16:35
2016-03-05

1- PRÉPARATION DE L'ENVIRONNEMENT

- *Création de la table*:

```
CREATE TABLE TP(VendorID INT ,tpep_pickup_datetime TIMESTAMP,
tpep_dropoff_datetime                                TIMESTAMP,
passenger_countINT,trip_distance    DOUBLE,    pickup_longitude
DOUBLE,        pickup_latitude  DOUBLE,        RateCodeID  INT,
store_and_fwd_flag     STRING     ,dropoff_longitude     DOUBLE,
dropoff_latitude  DOUBLE,  payment_typeINT,fare_amount  DOUBLE,
extra   DOUBLE,   mta_tax   DOUBLE,       tip_amount   DOUBLE,
tolls_amount DOUBLE, total_amount FLOAT)
COMMENT 'TP' ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'  STORED AS TEXTFILE ;
```

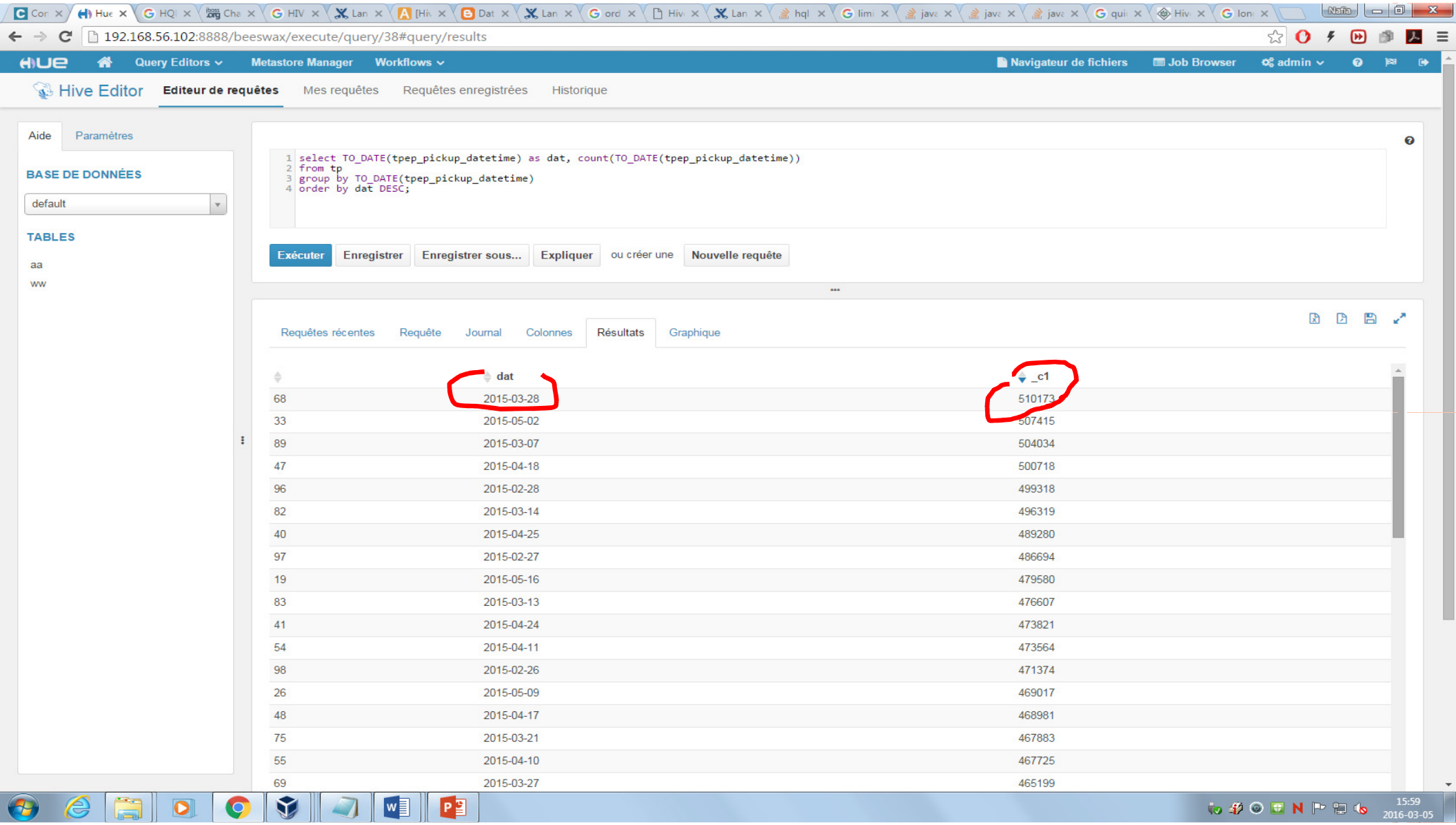
- *Chargement de la table:*

```
LOAD  DATA  INPATH  '/user/admin/yellow_tripdata_2015-01-06.csv'
INTO TABLE tp;
```

2- EXECUTION DES JOBS

- *Q1: La journée où il y a plus de trafic* :

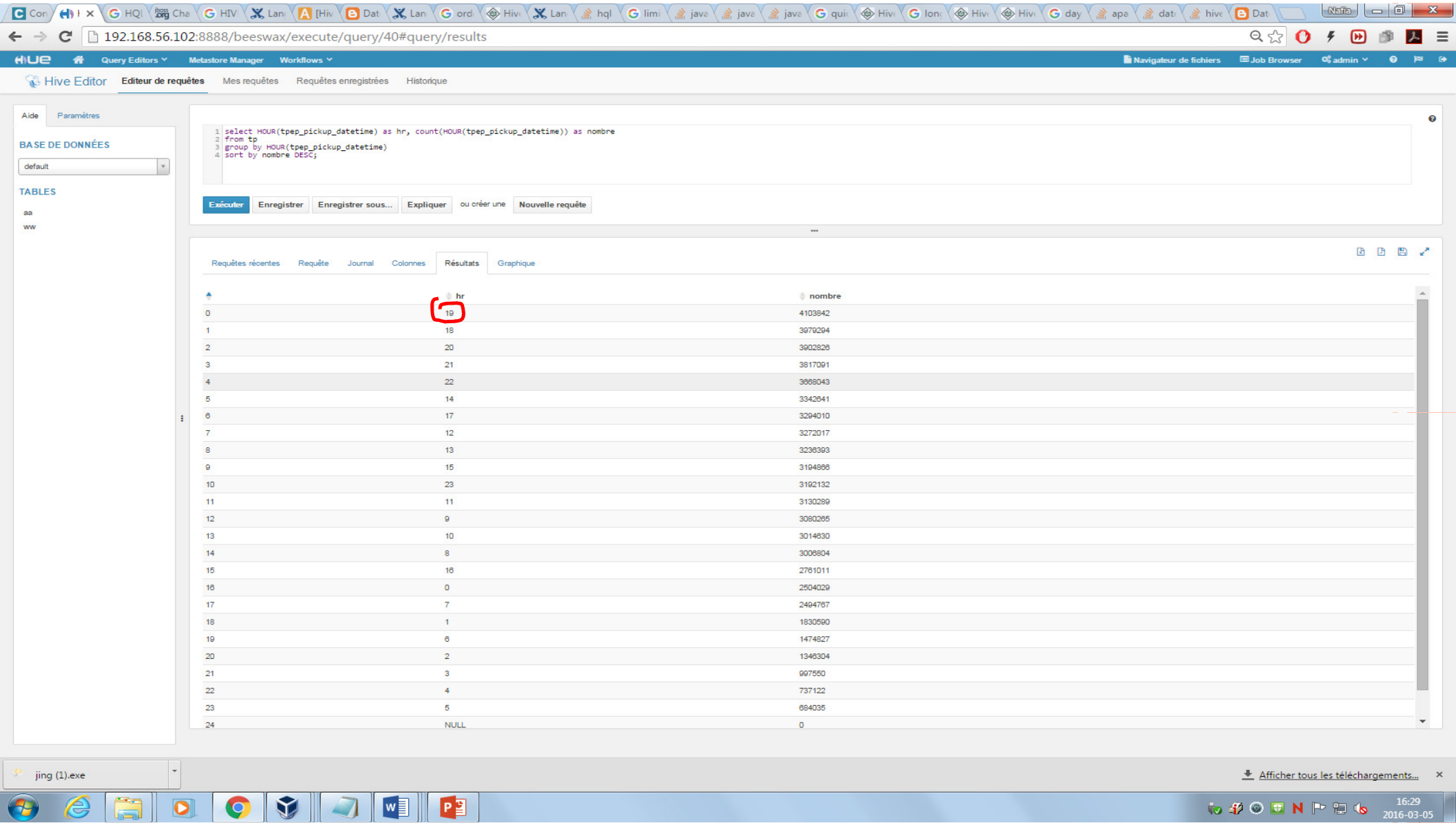
```
select TO_DATE(tpep_pickup_datetime) as dat,
count(TO_DATE(tpep_pickup_datetime)) as nombre
From TP
group by TO_DATE(tpep_pickup_datetime)
order by nombreDESC;
```


192.168.56.102:8888/beeswax/execute/query/38#query/results
Query Editors
Metastore Manager
Workflows
Navigateur de fichiers
Job Browser
admin
Hive Editor
Editeur de requêtes
Mes requêtes
Requêtes enregistrées
Historique
Aide
Paramètres
BASE DE DONNÉES
default
TABLES
aa
ww
1 select TO_DATE(tpep_pickup_datetime) as dat, count(TO_DATE(tpep_pickup_datetime))
2 from tp
3 group by TO_DATE(tpep_pickup_datetime)
4 order by dat DESC;
Exécuter
Enregistrer
Enregistrer sous...
Expliquer
ou créer une
Nouvelle requête
Requêtes récentes
Requête
Journal
Colonnes
Résultats
Graphique
dat
_c1
68 2015-03-28 510173
33 2015-05-02 507415
89 2015-03-07 504034
47 2015-04-18 500718
96 2015-02-28 499318
82 2015-03-14 496319
40 2015-04-25 489280
97 2015-02-27 486694
19 2015-05-16 479580
83 2015-03-13 476607
41 2015-04-24 473821
54 2015-04-11 473564
98 2015-02-26 471374
26 2015-05-09 469017
48 2015-04-17 468981
75 2015-03-21 467883
55 2015-04-10 467725
69 2015-03-27 465199
15:59
2016-03-05

2- EXECUTION DES JOBs

- *Q2: L'heure de pointe* :

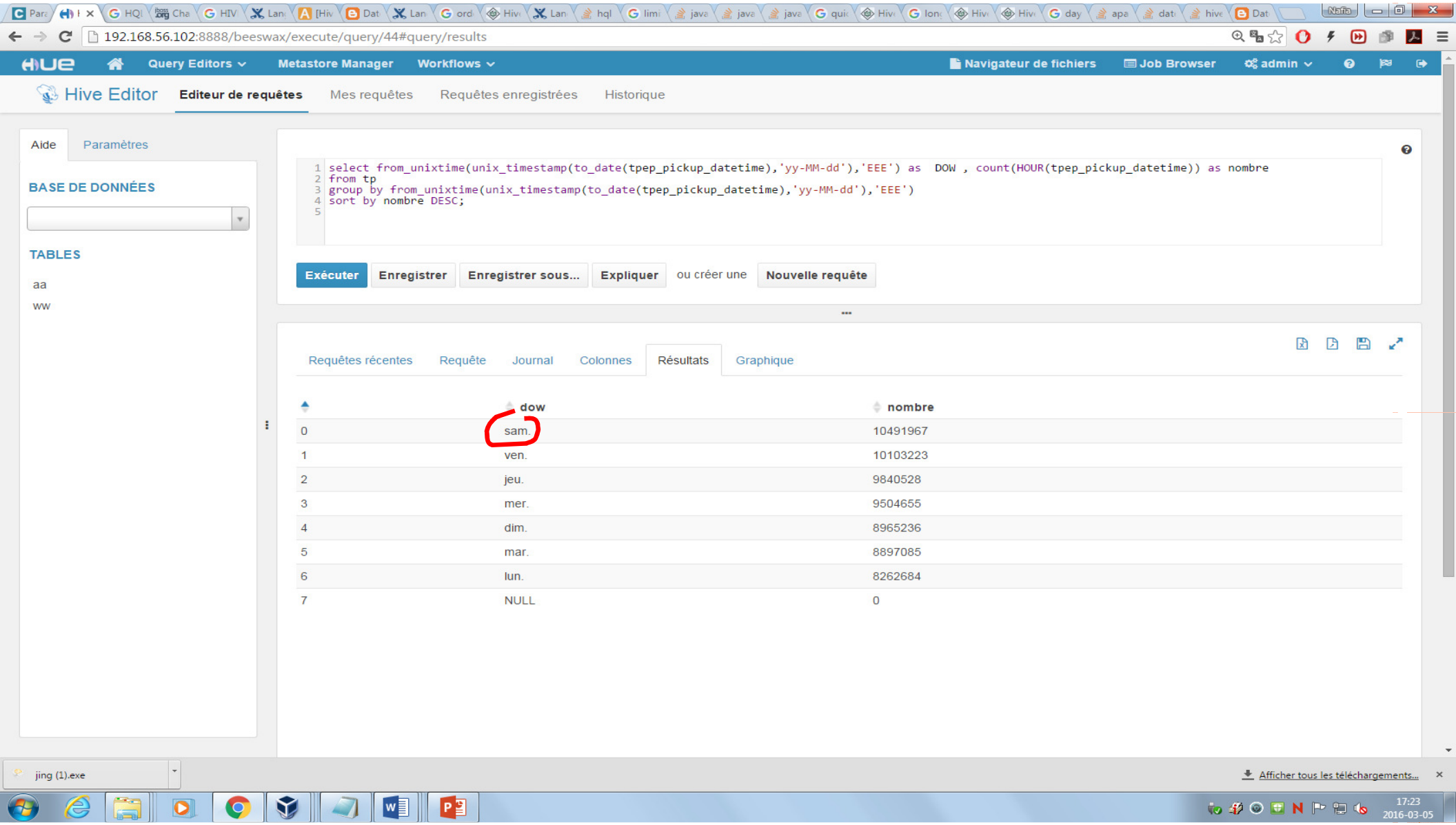
```
select HOUR(tpep_pickup_datetime) as hr,
count(HOUR(tpep_pickup_datetime)) as nombre
fromtp
group by HOUR(tpep_pickup_datetime)
sort by nombre DESC;
```


192.168.56.102:8888/beeswax/execute/query/40#query/results
Query Editors
Metastore Manager
Workflows
Navigateur de fichiers
Job Browser
admin
Hive Editor
Editeur de requêtes
Mes requêtes
Requêtes enregistrées
Historique
Aide
Paramètres
BASE DE DONNÉES
default
TABLES
aa
ww
select HOUR(tpep_pickup_datetime) as hr, count(HOUR(tpep_pickup_datetime)) as nombre
from tp
group by HOUR(tpep_pickup_datetime)
sort by nombre DESC;
Exécuter
Enregistrer
Enregistrer sous...
Expliquer
ou créer une
Nouvelle requête
Requêtes récentes
Requête
Journal
Colonnes
Résultats
Graphique
hr
nombre
0 19 4103842
1 18 3979294
2 20 3902826
3 21 3817091
4 22 3668043
5 14 3342641
6 17 3294010
7 12 3272017
8 13 3236393
9 15 3194866
10 23 3192132
11 11 3130289
12 9 3080265
13 10 3014630
14 8 3006804
15 16 2761011
16 0 2504029
17 7 2494767
18 1 1830590
19 6 1474827
20 2 1346304
21 3 997550
22 4 737122
23 5 684035
24 NULL 0
jing (1).exe
Afficher tous les téléchargements...
16:29
2016-03-05

2- EXECUTION DES JOBs

- *Q3: La journée de la semaine de pointe*

```
Select from_unixtime
(unix_timestamp(to_date(tpep_pickup_datetime), 'yy-MM-
dd'),'EEE') as  DOW , count(HOUR(tpep_pickup_datetime))
as nombre
From TP
group by from_unixtime(unix_timestamp
(to_date(tpep_pickup_datetime), 'yy-MM-dd'),'EEE')
sort by nombre DESC;
```


192.168.56.102:8888/beeswax/execute/query/44#query/results
Query Editors
Metastore Manager
Workflows
Navigateur de fichiers
Job Browser
admin
Hive Editor
Editeur de requêtes
Mes requêtes
Requêtes enregistrées
Historique
Aide
Paramètres
BASE DE DONNÉES
TABLES
aa
ww
1 select from_unixtime(unix_timestamp(to_date(tpep_pickup_datetime),'yy-MM-dd'),'EEE') as DOW , count(HOUR(tpep_pickup_datetime)) as nombre
2 from tp
3 group by from_unixtime(unix_timestamp(to_date(tpep_pickup_datetime),'yy-MM-dd'),'EEE')
4 sort by nombre DESC;
5
Exécuter
Enregistrer
Enregistrer sous...
Expliquer
ou créer une
Nouvelle requête
Requêtes récentes
Requête
Journal
Colonnes
Résultats
Graphique
dow
nombre
0 sam. 10491967
1 ven. 10103223
2 jeu. 9840528
3 mer. 9504655
4 dim. 8965236
5 mar. 8897085
6 lun. 8262684
7 NULL 0
jing (1).exe
Afficher tous les téléchargements...
17:23
2016-03-05

Difficultés rencontrées

- Problèmes d'installation sur les machines
- Problèmes d'instabilité du cluster
- Problèmes des configuration réseaux entre les machines
- Lenteur d'exécution à cause du volume important de données et a cause des ressources mémoire
- Problème d'écriture dans hdfs a partir de ligne de commande
- Interruption de téléchargement du fichier via Hue
- Le fichier est téléchargé juste sur une extention .tmp

Compétences acquises

- Configuration des machines linux
- Installation d'un cluster via Cloudera-manager
- Hdfs
- Hue
- Hive

Conclusion

- Complexité des outils de travail en Big Data
- Domaine en plein essor
- L'avenir est pour les entreprises qui maitrisent les technologies Big Data
- Traiter un dataset volumineux

« Dans le milieu de la difficulté se trouve l'opportunité »