PROJET DE FIN DE SESSION

PRESENTE PAR

Éric TREMBLAY

&

Raoul KOUANDA

DEMANDE PAR

HAFED BENTEFTIFA

Équipe

- Analyste affaire Éric
- Administrateur Raoul
- Développeur Éric & Raoul

BESOINS

- Notre client est un distributeur de boisson énergétique.
- Une large proportion de sa clientèle sont des cyclistes (40%)
- Clientèle entre 25 et 35 ans (50% ca)

- Donc notre client recherche les endroits opportuns pour ouvrir des kiosques de vente

Besoins (suite)

- Produit: Boisson énergétique
- Clientèle visée : Cycliste de 25 à 35 ans

- But: Ouvrir de nouveaux points de ventes
- Critère principal: Présence de la clientèle cible au moment ou elle a soif.

- Trouver les stations d’arrivée ou notre clientèle est la plus présente

INFRASTRUCTURE

- 1 name node + 1 data node
- CentOS 6.2 avec 8 gigs

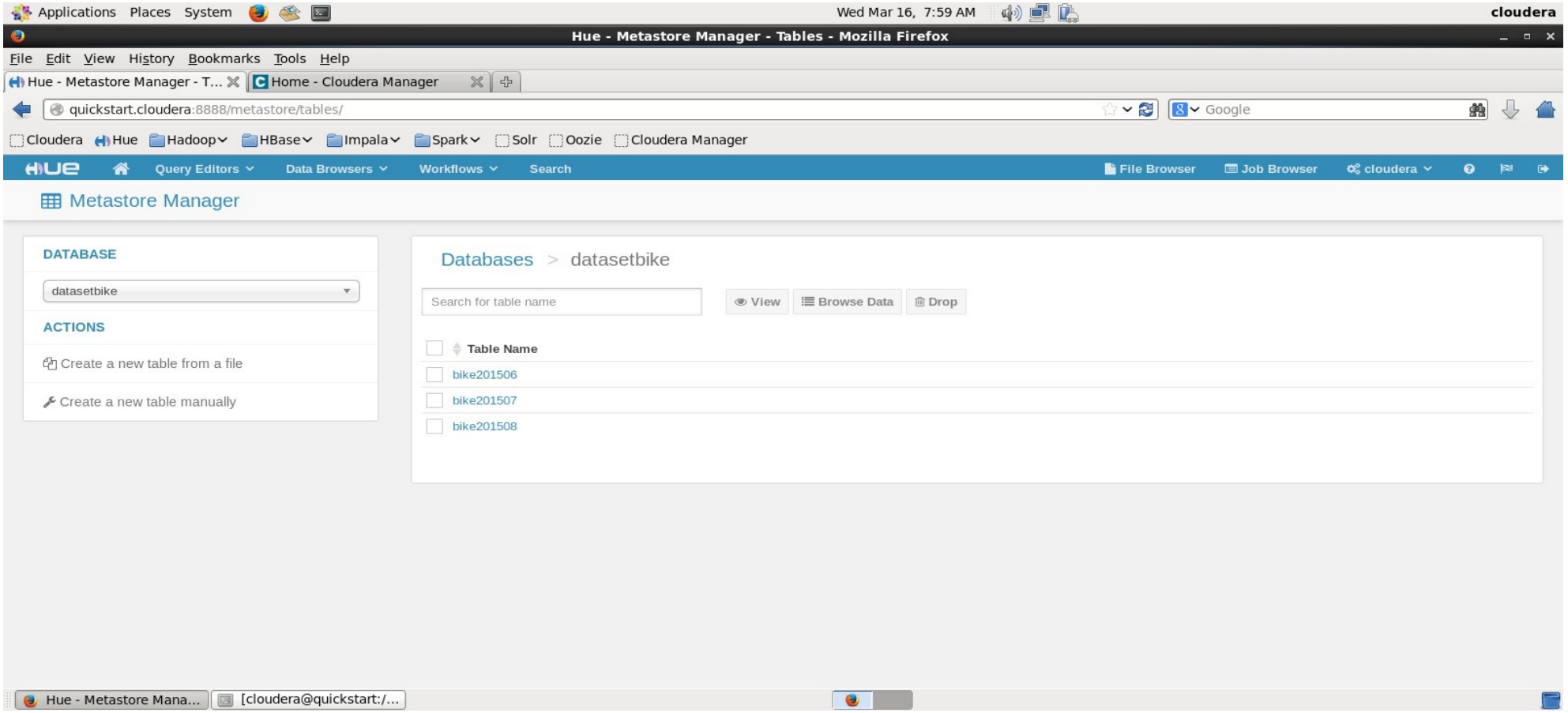
- 1 data node supplémentaire
- CentOS 6.4 avec 2 gigs

Applications

- Cloudera Express 5.1.0
- CDH 5.1.0
- Code utilisé: Requête HIVE
  - Importation des fichiers dans des tables
  - Élaboration des requêtes pour répondre aux besoins du client
- Résultats
  - Liste des stations ou arrive le plus de cycliste de 25 à 35 ans

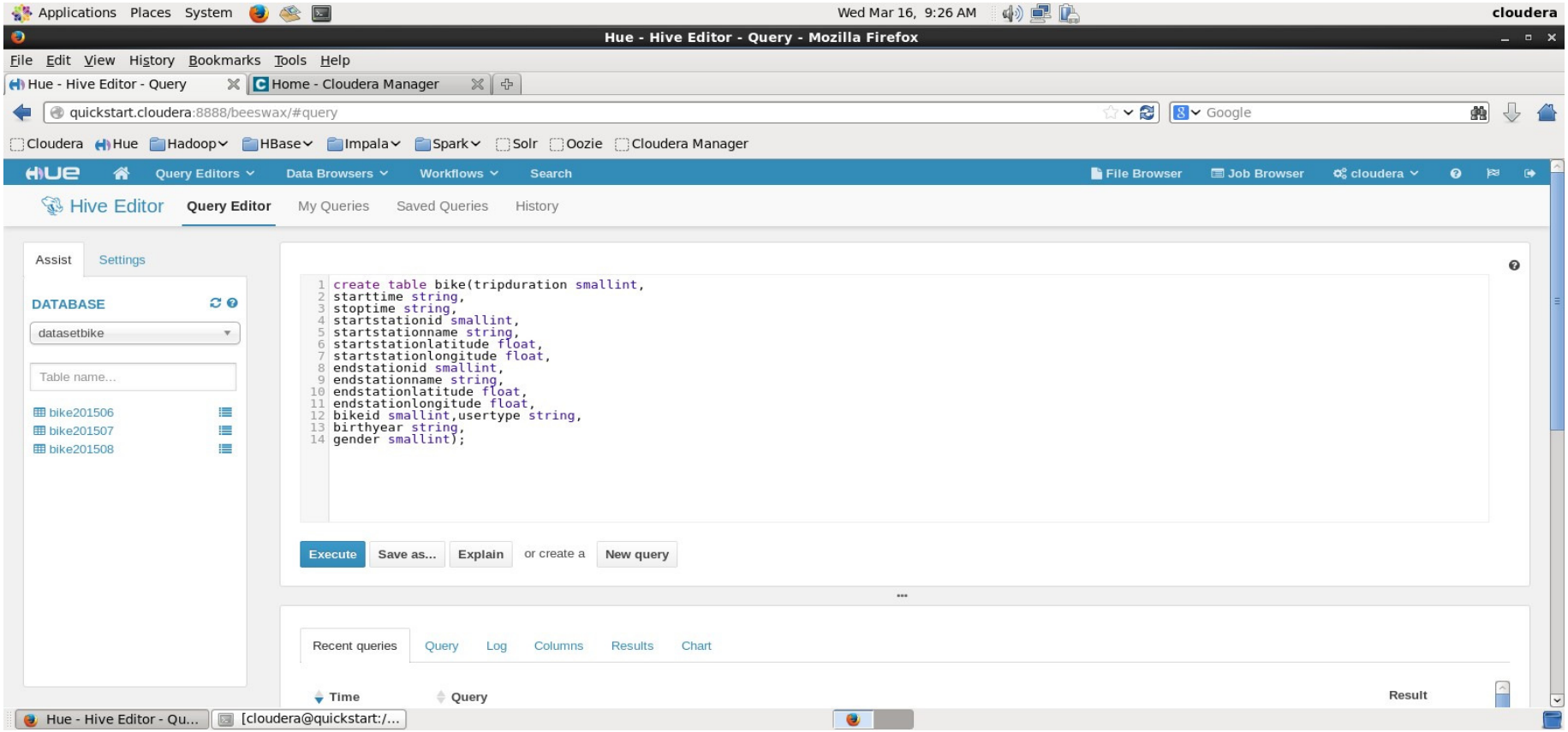
CODE

- Création des tables pour chaque mois (Juin, Juillet et Aout 2015)



CODE(suite)

- Création de la table Bike et insertion des données mensuelles dans la table bike.



CODE(suite)

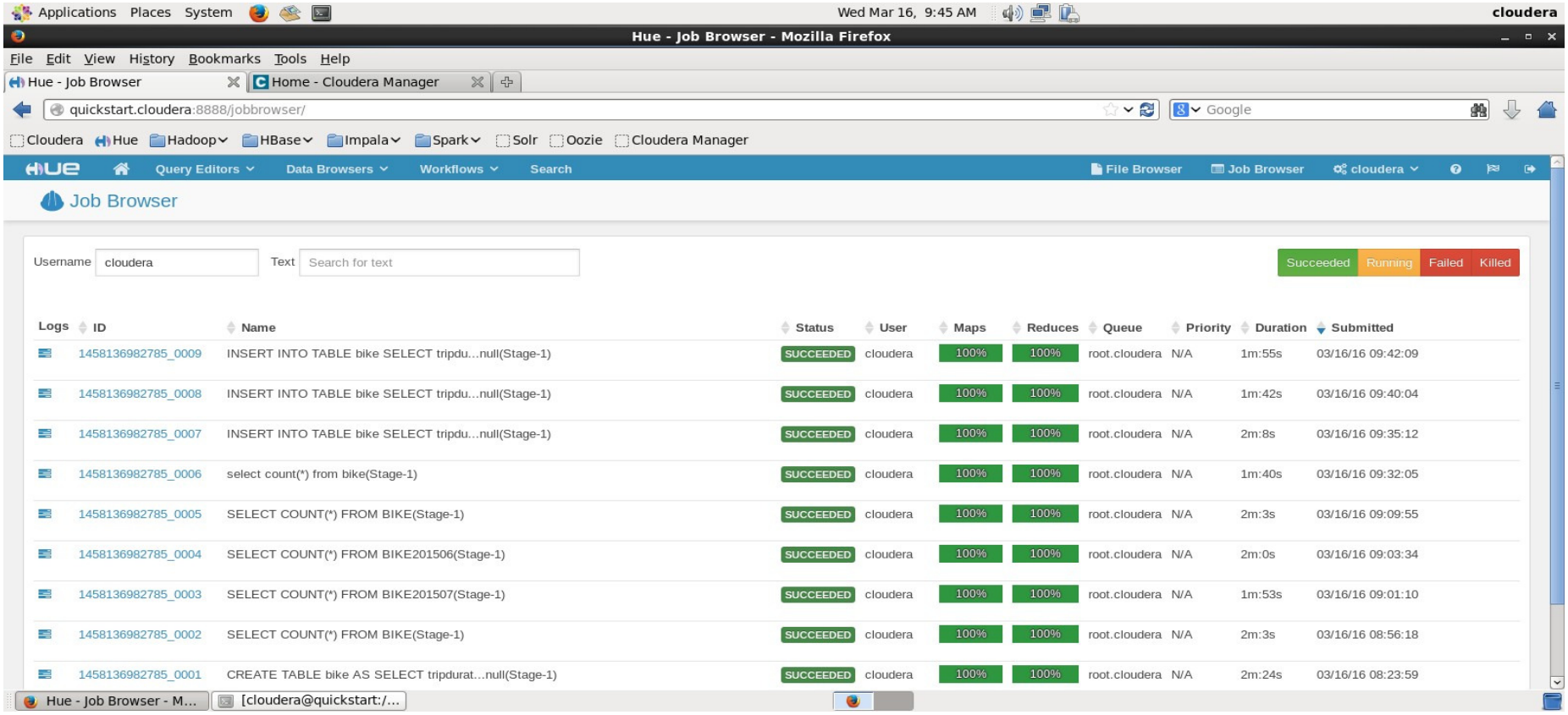
```
INSERT INTO TABLE bike
SELECT tripduration,starttime, stoptime, startstationid, startstationname, startstationlatitude, startstationlongitude,
endstationid,
endstationname, endstationlatitude, endstationlongitude, bikeid, usertype, birthyear, gender
FROM bike201506
WHERE tripduration is not null and startstationid is not null and endstationid is not null and endstationname is not null
and birthyear is not null

INSERT INTO TABLE bike
SELECT tripduration,starttime, stoptime, startstationid, startstationname, startstationlatitude, startstationlongitude,
endstationid,
endstationname, endstationlatitude, endstationlongitude, bikeid, usertype, birthyear, gender
FROM bike201507
WHERE tripduration is not null and startstationid is not null and endstationid is not null and endstationname is not null
and birthyear is not null

INSERT INTO TABLE bike
SELECT tripduration,starttime, stoptime, startstationid, startstationname, startstationlatitude, startstationlongitude,
endstationid,
endstationname, endstationlatitude, endstationlongitude, bikeid, usertype, birthyear, gender
FROM bike201508
WHERE tripduration is not null and startstationid is not null and endstationid is not null and endstationname is not null
and birthyear is not null
```

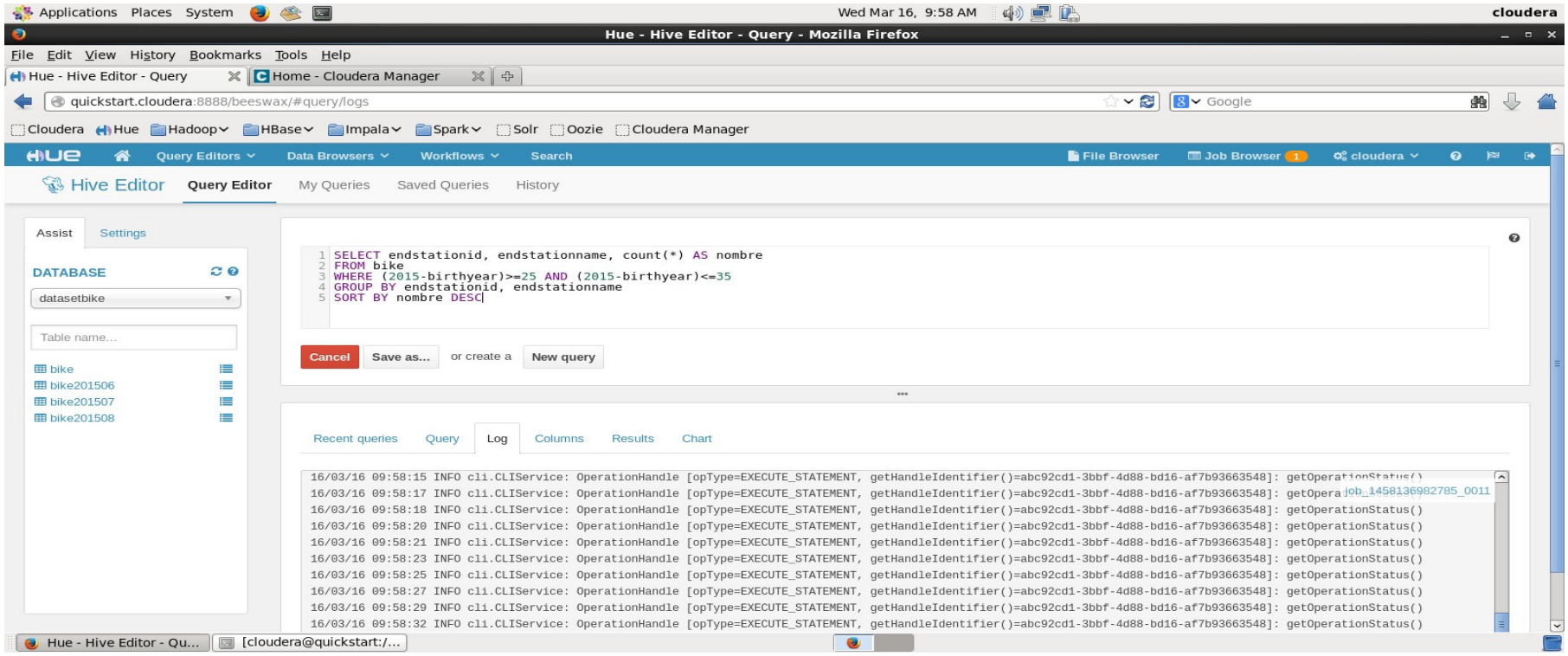
CODE(suite)

Exécution du traitement lors de l'insertion des données dans la table bike

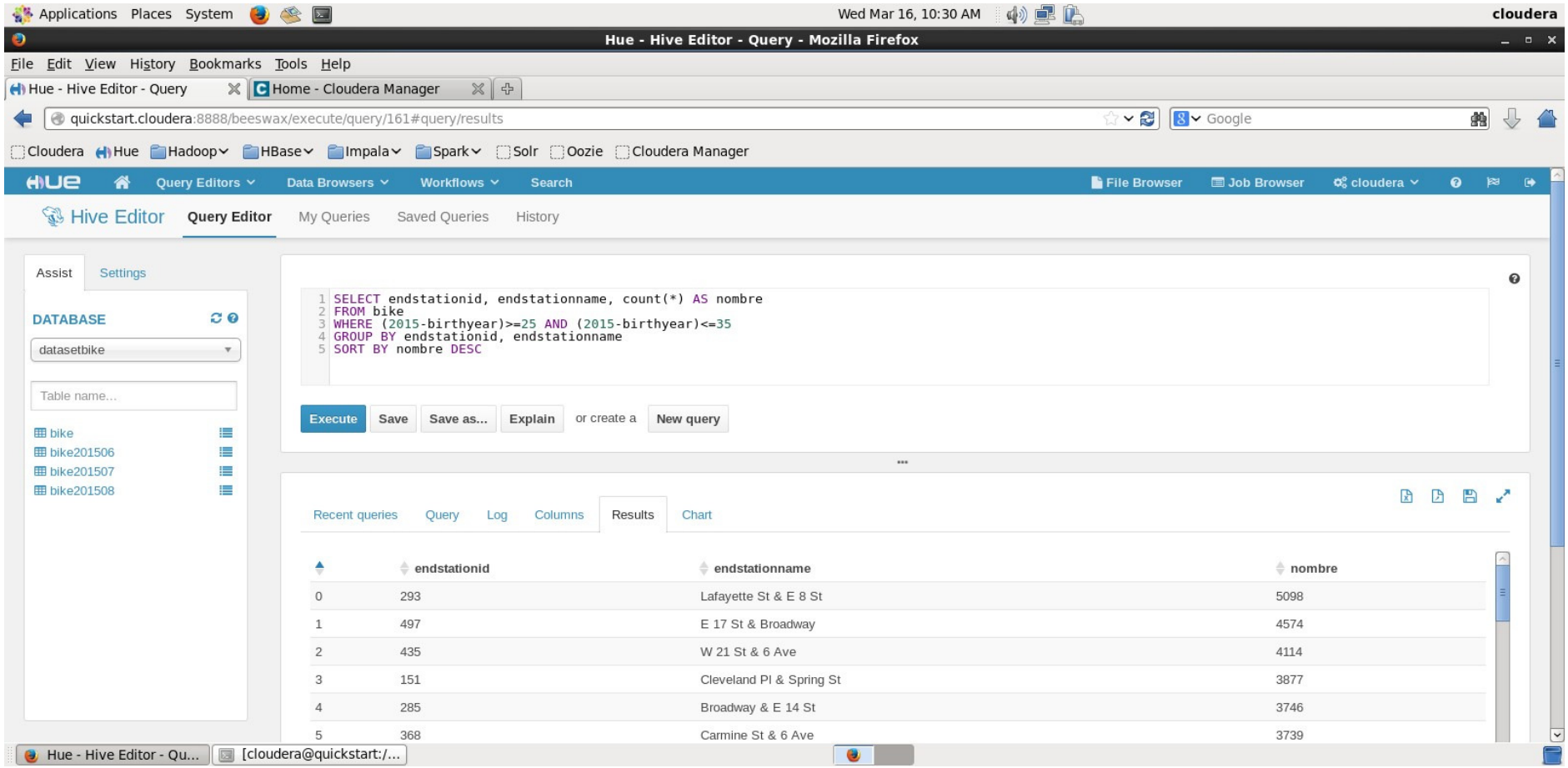


CODE(suite)

Extraction des points de la ville ou arrive le plus de cycliste de 25 à 35 ans



RESULTAT


Applications Places System
Wed Mar 16, 10:30 AM
cloudera
Hue - Hive Editor - Query - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Hue - Hive Editor - Query
Home - Cloudera Manager
quickstart.cloudera:8888/beeswax/execute/query/161#query/results
Google
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager
Query Editors Data Browsers Workflows Search
File Browser Job Browser cloudera
Hive Editor
Query Editor My Queries Saved Queries History
Assist Settings
DATABASE
datasetbike
Table name...
bike
bike201506
bike201507
bike201508
1 SELECT endstationid, endstationname, count(*) AS nombre
2 FROM bike
3 WHERE (2015-birthyear)>=25 AND (2015-birthyear)<=35
4 GROUP BY endstationid, endstationname
5 SORT BY nombre DESC
Execute Save Save as... Explain or create a New query
Recent queries Query Log Columns Results Chart
endstationid endstationname nombre
0 293 Lafayette St & E 8 St 5098
1 497 E 17 St & Broadway 4574
2 435 W 21 St & 6 Ave 4114
3 151 Cleveland Pl & Spring St 3877
4 285 Broadway & E 14 St 3746
5 368 Carmine St & 6 Ave 3739
Hue - Hive Editor - Qu...
[cloudera@quickstart:/...

DIFFICULTÉS RENCONTRÉES

- Configuration du réseau
- Mémoire insuffisante dans nos ordinateurs
- Problème de version de Linux
  - Version graphique/complète trop gourmande
  - Erreur au download utilisation de version 32 bits
- Problème avec réutilisation (clone) d'un datanode déjà installé
- Problème de format au chargement (
  - Numérique avec double quote (format reconnu mais chargement en erreur)
  - Date non reconnu
- Manque de temps pour le travail autant individuel qu'en équipe

COMPÉTENCES ACQUISES

- Nette amélioration de nos connaissances
  - Virtualisation
  - Unix
  - Réseaux
  - Hadoop
  - HDFS
  - HIVE
  - HUE

CONCLUSION

- Nous commençons à comprendre les difficultés à installer un cluster Hadoop.
- Les difficultés sont dans les détails
- Nous croyons que la technologie est encore jeune et manque de stabilité.
- Nous avons encore beaucoup à apprendre.

- Nous sommes loin d'être des pros mais c'est un début.
- Comme dit Hafed, c'est en forgeant qu'on devient forgeron.