

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

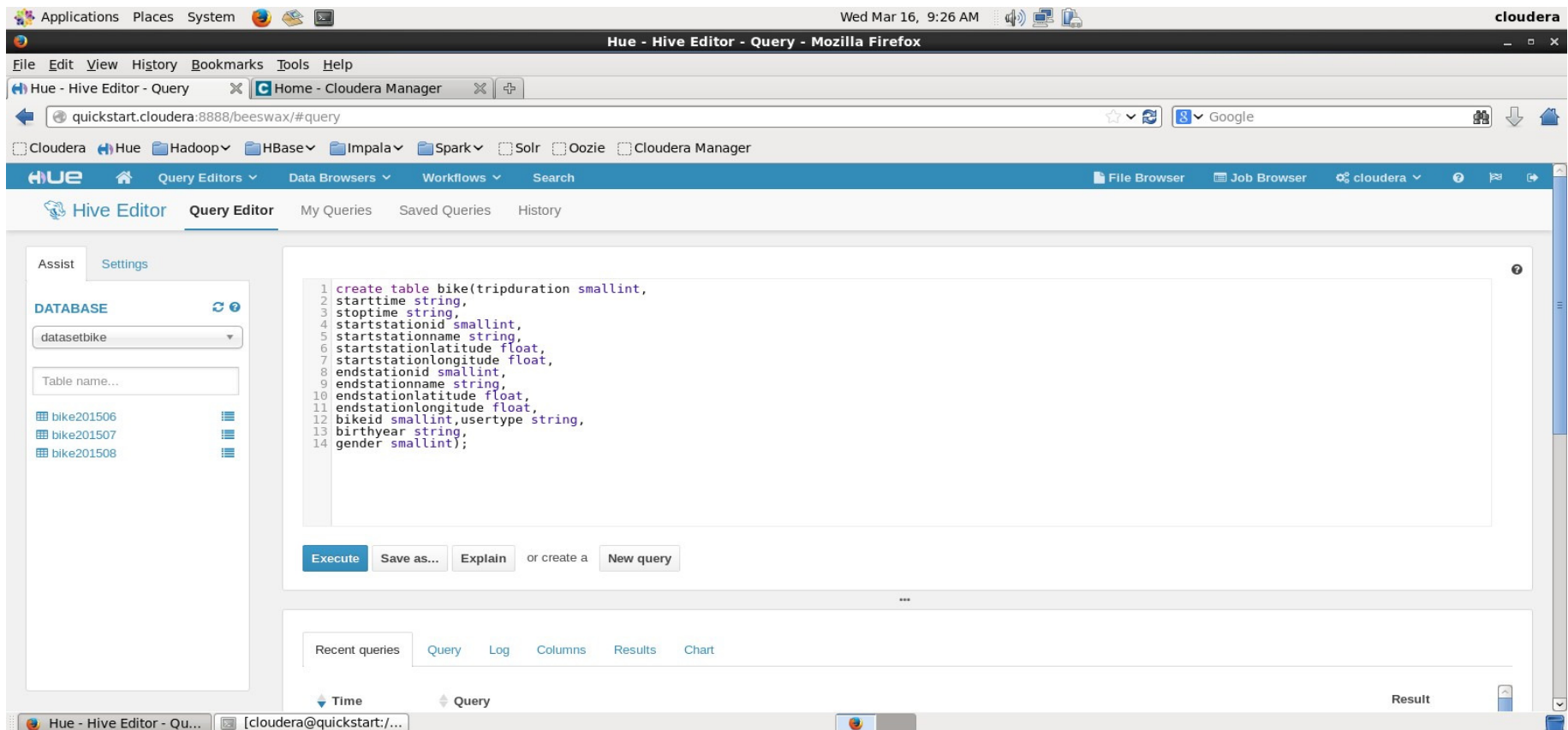
CODE

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

The screenshot displays the Hue Metastore Manager interface within a Mozilla Firefox browser window. The browser's address bar shows the URL `quickstart.cloudera:8888/metastore/tables/`. The Hue application's navigation bar includes the Hue logo and menu items for Query Editors, Data Browsers, Workflows, and Search. The main content area is titled "Metastore Manager" and shows the "Databases > datasetbike" view. On the left, a sidebar contains a "DATABASE" dropdown menu set to "datasetbike" and "ACTIONS" for creating tables from a file or manually. The main table lists three tables: "bike201506", "bike201507", and "bike201508", each with a checkbox and a "Table Name" label. The system tray at the bottom shows the Hue application icon and the user's terminal session `[cloudera@quickstart:/...]`.

CODE(suite)

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



The screenshot shows the Hue Hive Editor interface in a Mozilla Firefox browser window. The browser's address bar displays the URL `quickstart.cloudera:8888/ beeswax/#query`. The Hue interface includes a top navigation bar with options like 'Query Editors', 'Data Browsers', and 'Workflows'. The main workspace is titled 'Hive Editor - Query Editor' and contains a SQL query editor with the following code:

```
1 create table bike(tripduration smallint,  
2 starttime string,  
3 stoptime string,  
4 startstationid smallint,  
5 startstationname string,  
6 startstationlatitude float,  
7 startstationlongitude float,  
8 endstationid smallint,  
9 endstationname string,  
10 endstationlatitude float,  
11 endstationlongitude float,  
12 bikeid smallint,usertype string,  
13 birthyear string,  
14 gender smallint);
```

Below the query editor, there are buttons for 'Execute', 'Save as...', 'Explain', and 'New query'. On the left side, a sidebar shows the 'DATABASE' section with a dropdown menu set to 'datasetbike' and a list of tables: 'bike201506', 'bike201507', and 'bike201508'. The bottom of the interface features a 'Recent queries' section with tabs for 'Query', 'Log', 'Columns', 'Results', and 'Chart'.

CODE(suite)

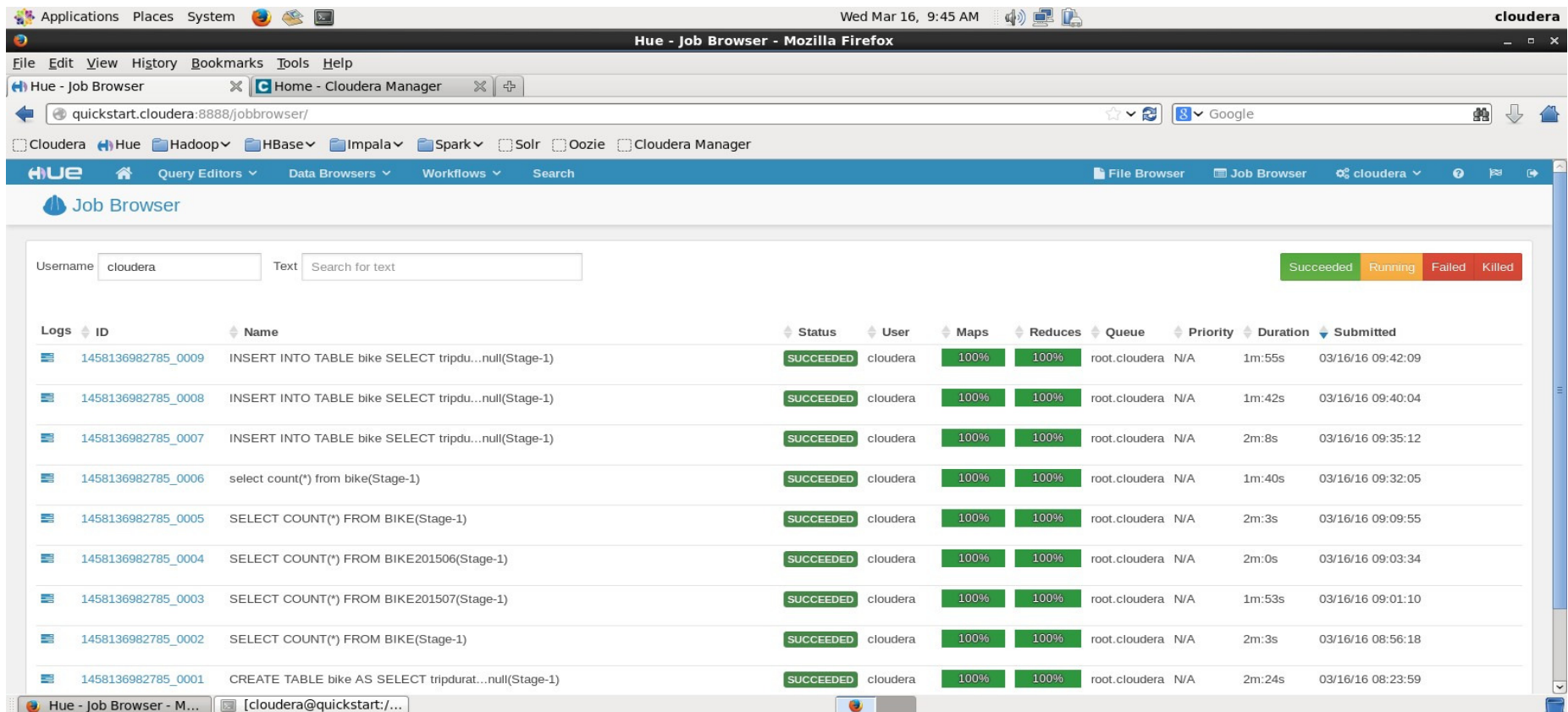
```
INSERT INTO TABLE bike
SELECT tripduration,starttime, stoptime, startstationid, startstationname, startstationlatitude, startstationlongitude,
endstationid,
endstationname, endstationlatitude, endstationlongitude, bikeid, usertype, birthyear, gender
FROM bike201506
WHERE tripduration is not null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



The screenshot displays the Hue Job Browser interface in a Mozilla Firefox browser window. The page title is "Hue - Job Browser - Mozilla Firefox". The browser address bar shows "quickstart.cloudera:8888/jobbrowser/". The Hue interface includes a navigation bar with "HUE" logo, "Query Editors", "Data Browsers", "Workflows", "Search", "File Browser", and "Job Browser". Below the navigation bar, there is a search section with "Username" set to "cloudera" and a "Text" search box. A legend indicates job statuses: Succeeded (green), Running (orange), Failed (red), and Killed (black). The main content area is a table of jobs with the following columns: Logs, ID, Name, Status, User, Maps, Reduces, Queue, Priority, Duration, and Submitted. All jobs shown have a "SUCCEEDED" status.

Logs	ID	Name	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
	1458136982785_0009	INSERT INTO TABLE bike SELECT tripdu...null(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	1m:55s	03/16/16 09:42:09
	1458136982785_0008	INSERT INTO TABLE bike SELECT tripdu...null(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	1m:42s	03/16/16 09:40:04
	1458136982785_0007	INSERT INTO TABLE bike SELECT tripdu...null(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	2m:8s	03/16/16 09:35:12
	1458136982785_0006	select count(*) from bike(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	1m:40s	03/16/16 09:32:05
	1458136982785_0005	SELECT COUNT(*) FROM BIKE(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	2m:3s	03/16/16 09:09:55
	1458136982785_0004	SELECT COUNT(*) FROM BIKE201506(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	2m:0s	03/16/16 09:03:34
	1458136982785_0003	SELECT COUNT(*) FROM BIKE201507(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	1m:53s	03/16/16 09:01:10
	1458136982785_0002	SELECT COUNT(*) FROM BIKE(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	2m:3s	03/16/16 08:56:18
	1458136982785_0001	CREATE TABLE bike AS SELECT tripdurat...null(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	2m:24s	03/16/16 08:23:59

CODE(suite)

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

The screenshot displays the Hue Hive Editor interface within a Mozilla Firefox browser window. The browser's address bar shows the URL `quickstart.cloudera:8888/ beeswax/#query/logs`. The interface includes a top navigation bar with options like 'Query Editors', 'Data Browsers', and 'Workflows'. On the left, a sidebar shows the 'DATABASE' section with a dropdown menu set to 'datasetbike' and a list of tables: 'bike', 'bike201506', 'bike201507', and 'bike201508'. The main area contains a SQL query editor with the following code:

```
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
```

Below the query editor are buttons for 'Cancel', 'Save as...', and 'New query'. The bottom section of the interface shows a 'Log' tab with a list of execution logs. The logs consist of multiple entries with timestamps and status information, such as:

```
16/03/16 09:58:15 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:17 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:18 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:20 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:21 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:23 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:25 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:27 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:29 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
16/03/16 09:58:32 INFO c11.CLIService: OperationHandle [opType=EXECUTE_STATEMENT, getHandleIdentifier()=abc92cd1-3bbf-4d88-bd16-af7b93663548]: getOperationStatus()
```

RESULTAT

The screenshot shows the Hue Hive Editor interface in a Mozilla Firefox browser window. The browser address bar shows the URL: `quickstart.cloudera:8888/beeswax/execute/query/161#query/results`. The interface includes a top navigation bar with 'HUE' and various tool categories like 'Query Editors', 'Data Browsers', 'Workflows', and 'Search'. Below this, the 'Hive Editor' section is active, displaying a SQL query in a text editor. The query is as follows:

```
1 SELECT endstationid, endstationname, count(*) AS nombre
2 FROM bike
3 WHERE (2015-birtheyar)>=25 AND (2015-birtheyar)<=35
4 GROUP BY endstationid, endstationname
5 SORT BY nombre DESC
```

Below the query editor are buttons for 'Execute', 'Save', 'Save as...', 'Explain', and 'New query'. The 'Execute' button has been clicked, and the results are displayed in a table below. The table has columns for 'endstationid', 'endstationname', and 'nombre'. The results are sorted in descending order of 'nombre'.

	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

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.