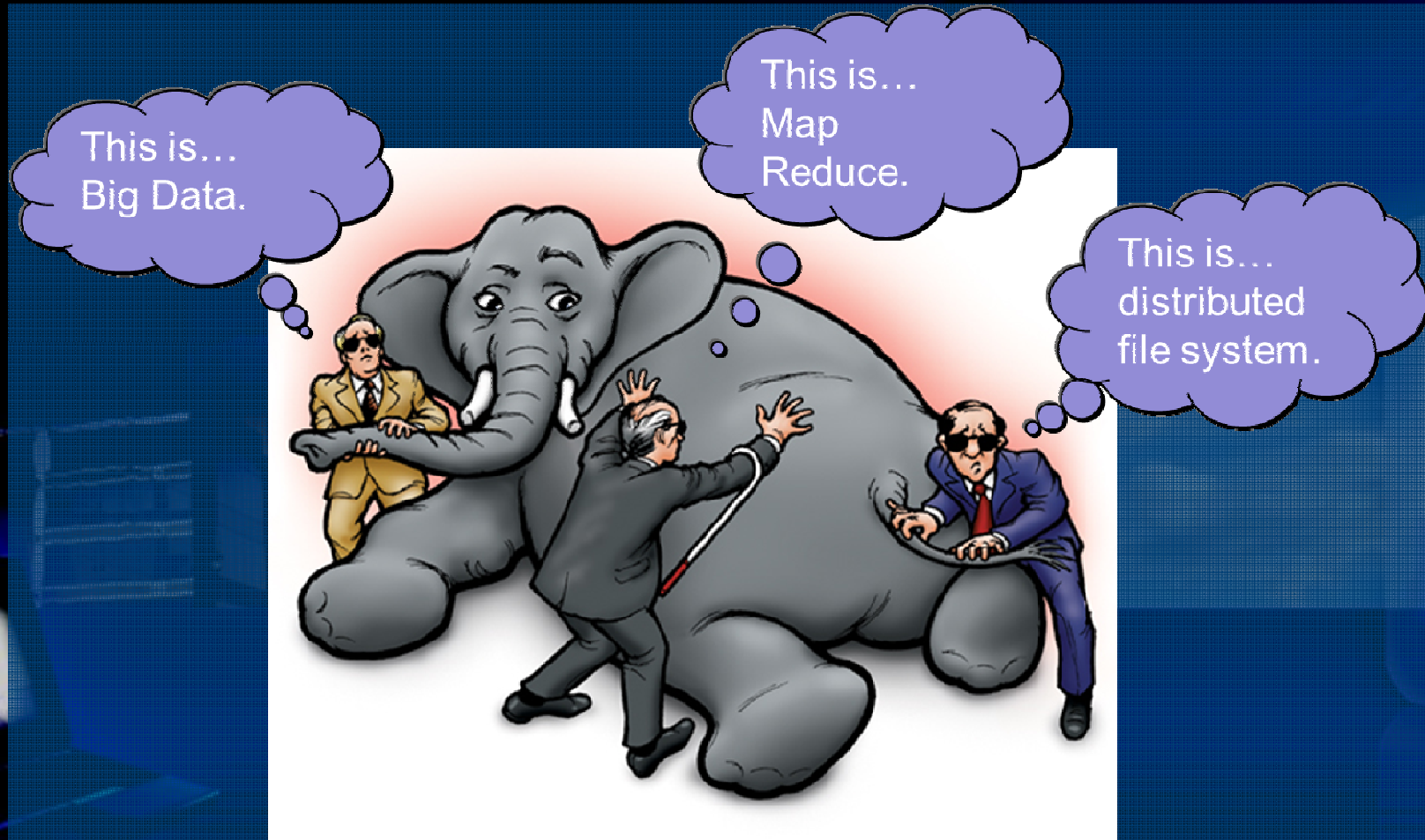


# PROJET DE SESSION



# PROJET HADOOP

METTRE EN PLACE UN CLUSTER POUR  
RÉPONDRE À UN BESOIN D'AFFAIRES

# Présentation de l'équipe

- Client
- Administrateur
- Développeur

# BESOIN

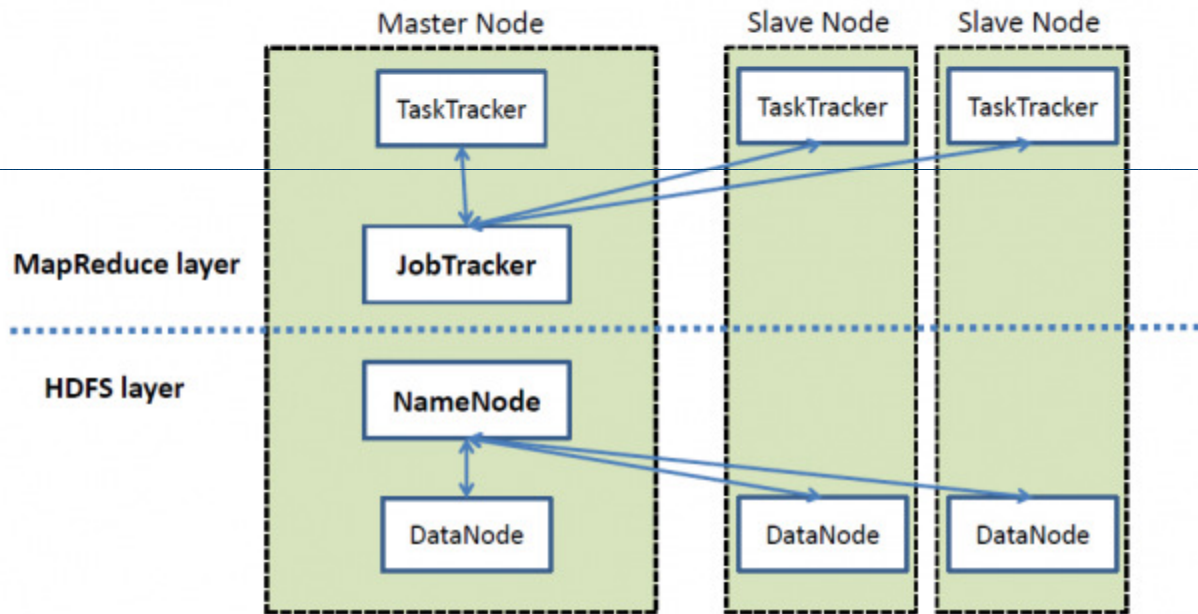
- Établir un diagnostic sur la fertilité à partir d'une série d'observation.
- Notre échantillon

# Installation du cluster

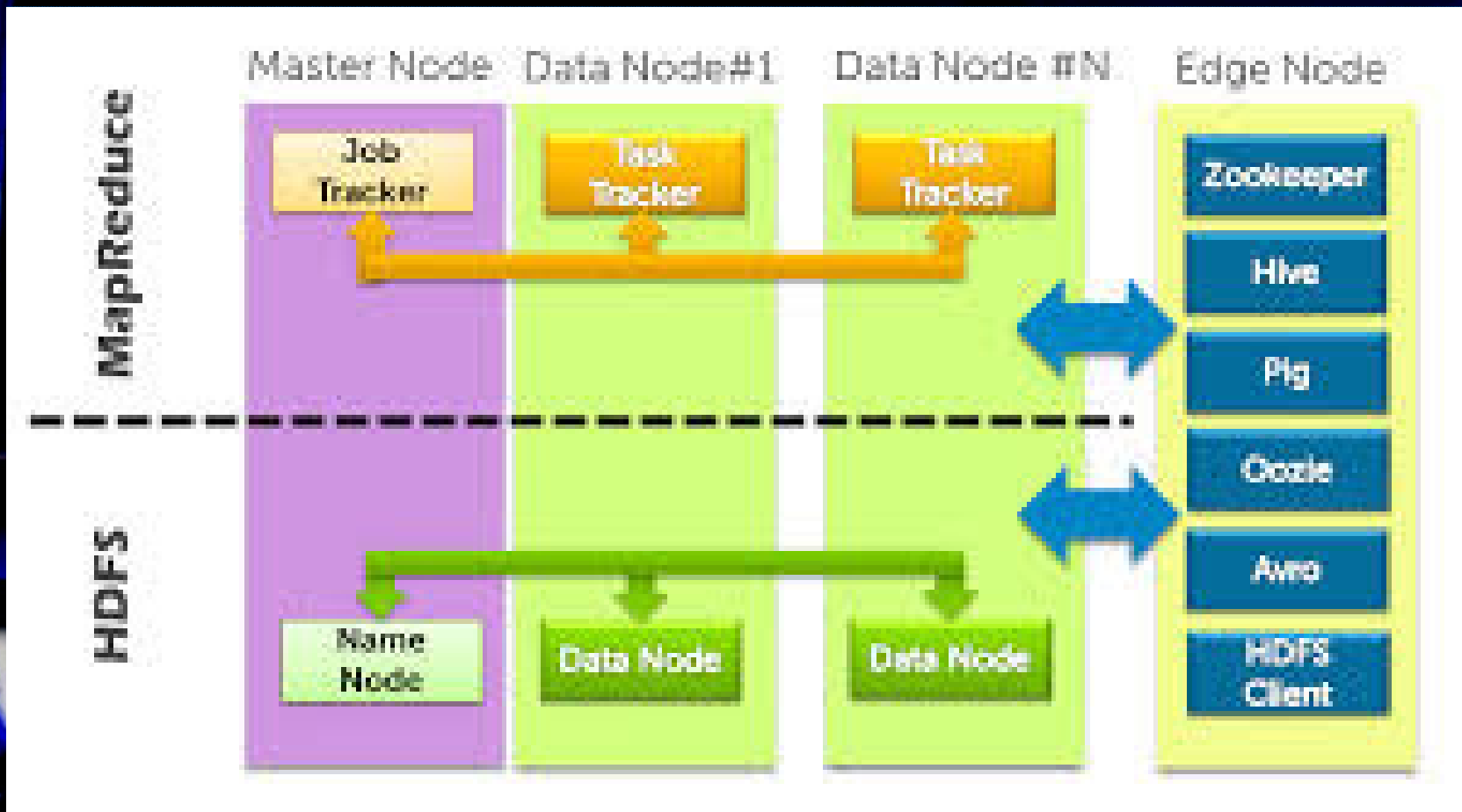
- Les caractéristiques du cluster
- Choix de la distribution: Hortonworks ou Cloudera manager
- Application installée

# Architecture du cluster

## High Level Architecture of Hadoop



# Cloudera Manager



# Notre machine






























## Home

30 minutes preceding March 8, 2016, 1:32 PM PST

Status All Health Issues Configuration  All Recent Commands

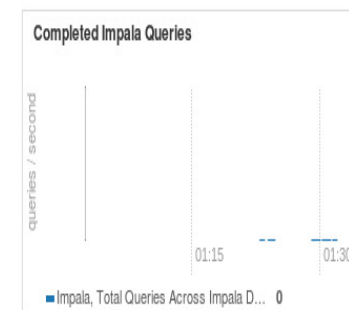
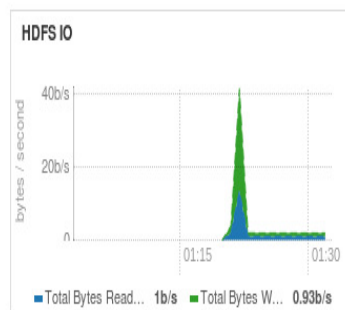
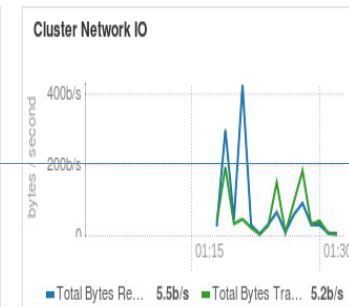
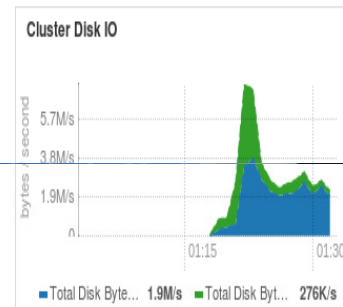
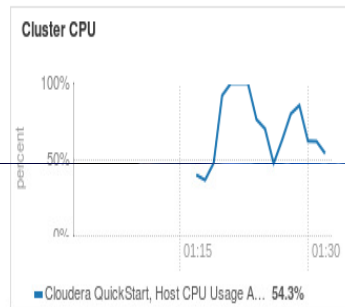
Add Cluster

Try Cloudera Enterprise Data Hub Edition for 60 Days

	<b>Cloudera QuickStart</b> (CDH 5.5.0...)	
	Hosts	
	HBase	
	HDFS	
	Hive	
	Hue	
	Impala	
	Key-Value Store...	
	Oozie	
	Solr	
	Spark	
	Sqoop 1 Client	
	Sqoop 2	
	YARN (MR2 Incl...)	
	ZooKeeper	

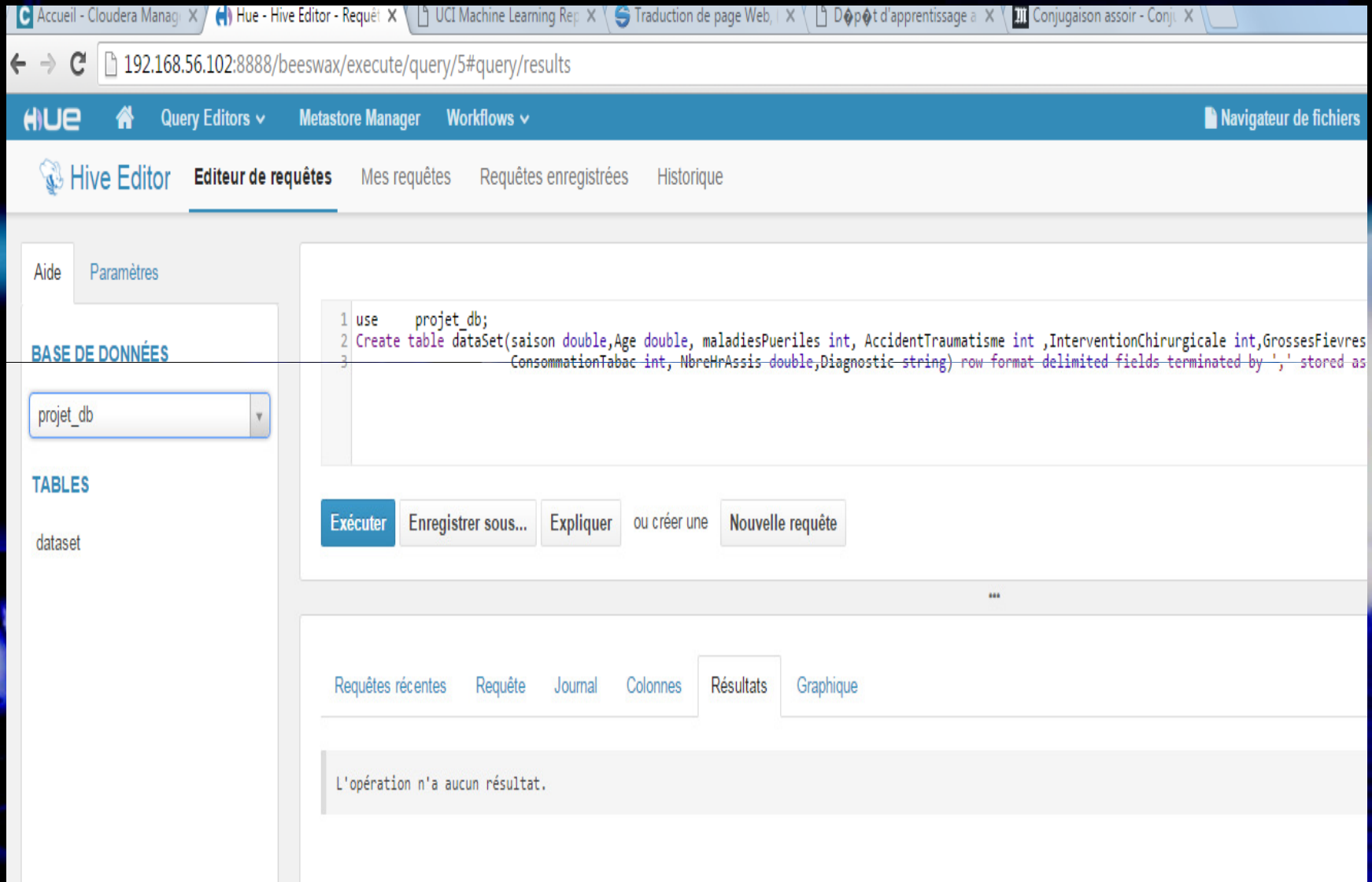
## Charts

30m 1h 2h 6h 12h 1d 7d 30d 





# Déploiement du code – Hive



The screenshot displays the Hue Hive Editor interface. The browser address bar shows the URL `192.168.56.102:8888/beeswax/execute/query/5#query/results`. The interface includes a navigation bar with 'HUE', 'Query Editors', 'Metastore Manager', and 'Workflows'. Below this, the 'Hive Editor' section is active, with tabs for 'Editeur de requêtes', 'Mes requêtes', 'Requêtes enregistrées', and 'Historique'. On the left, a sidebar shows 'Aide' and 'Paramètres' tabs, a 'BASE DE DONNÉES' dropdown menu set to 'projet\_db', and a 'TABLES' section listing 'dataset'. The main area contains a SQL query editor with the following code:

```
1 use projet_db;  
2 Create table dataSet(saison double, Age double, maladiesPueriles int, AccidentTraumatisme int, InterventionChirurgicale int, GrossesFievres  
3 ConsommationTabac int, NbreHrAssis double, Diagnostic string) row format delimited fields terminated by ',' stored as
```

Below the editor are buttons for 'Exécuter', 'Enregistrer sous...', 'Expliquer', and 'Nouvelle requête'. At the bottom, a 'Résultats' tab is selected, showing the message: 'L'opération n'a aucun résultat.'

# Structure de dataset

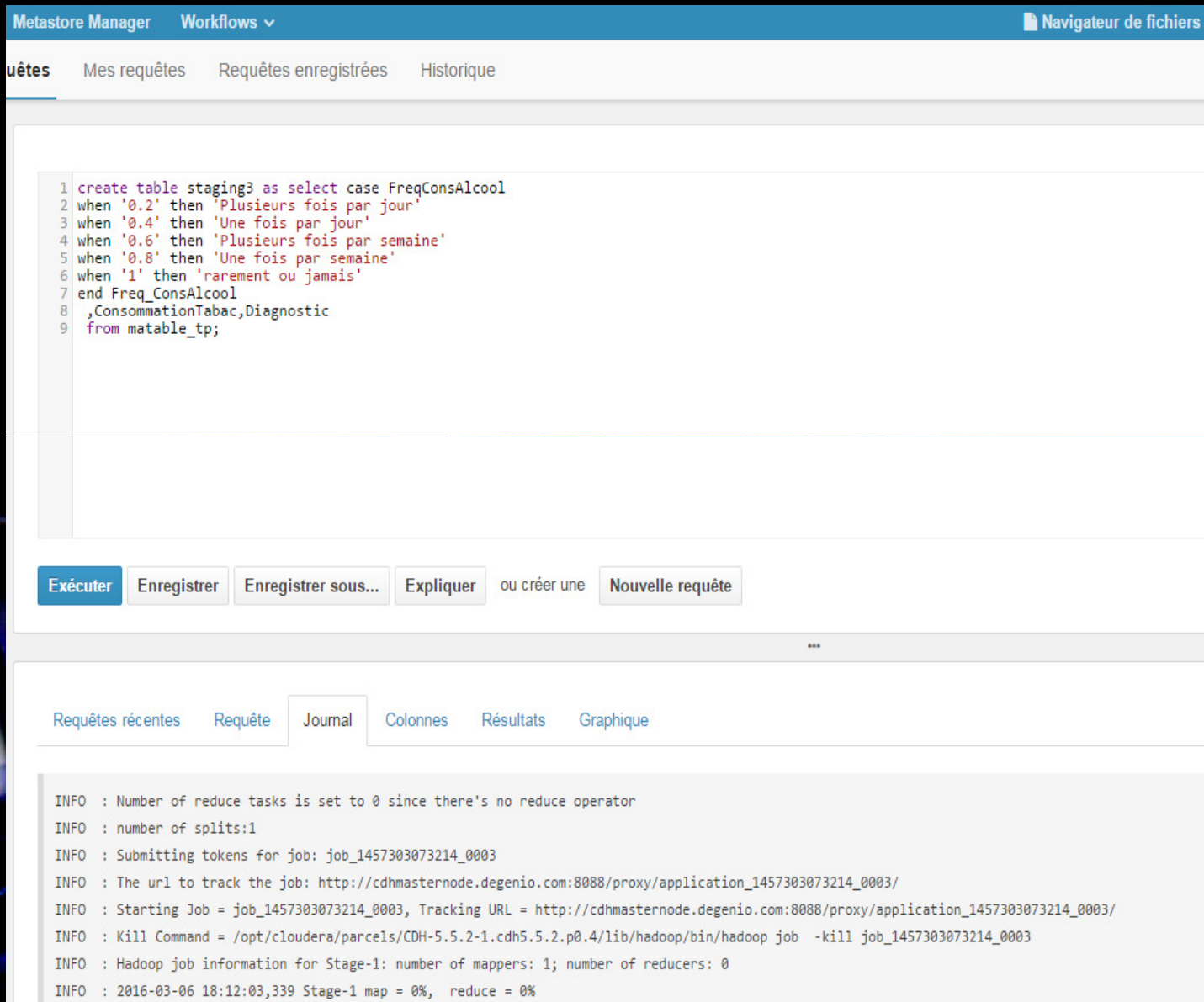
```
1 SELECT * FROM matable_tp;
```

Exécuter Enregistrer Enregistrer sous... Expliquer ou créer une Nouvelle requête

Requêtes récentes Requête Journal Colonnes Résultats Graphique

	matable_tp.saison	matable_tp.age	matable_tp.maladiespueriles	matable_tp.accidenttraumatisme	matable_tp.interventionchirurgicale	matable_tp.grossesfievr	matable_tp.freq
0	-0.33	0.69	0	1	1	0	0.8
1	-0.33	0.94	1	0	1	0	0.8
2	-0.33	0.5	1	0	0	0	1
3	-0.33	0.75	0	1	1	0	1
4	-0.33	0.67	1	1	0	0	0.8
5	-0.33	0.67	1	0	1	0	0.8
6	-0.33	0.67	0	0	0	-1	0.8
7	-0.33	1	1	1	1	0	0.6
8	1	0.64	0	0	1	0	0.8
9	1	0.61	1	0	0	0	1
10	1	0.67	1	1	0	-1	0.8
11	1	0.78	1	1	1	0	0.6
12	1	0.75	1	1	1	0	0.8
13	1	0.81	1	0	0	0	1
14	1	0.94	1	1	1	0	0.2
15	1	0.81	1	1	0	0	1
16	1	0.64	1	0	1	0	1

# Transformation de data



The screenshot displays the Metastore Manager interface. At the top, there is a navigation bar with 'Metastore Manager' and 'Workflows' on the left, and 'Navigateur de fichiers' on the right. Below this, a menu contains 'Requêtes', 'Mes requêtes', 'Requêtes enregistrées', and 'Historique'. The main area shows a SQL query in a text editor:

```
1 create table staging3 as select case FreqConsAlcool
2 when '0.2' then 'Plusieurs fois par jour'
3 when '0.4' then 'Une fois par jour'
4 when '0.6' then 'Plusieurs fois par semaine'
5 when '0.8' then 'Une fois par semaine'
6 when '1' then 'rarement ou jamais'
7 end Freq_ConsAlcool
8 ,ConsommationTabac,Diagnostic
9 from matable_tp;
```

Below the query editor, there are several buttons: 'Exécuter' (highlighted in blue), 'Enregistrer', 'Enregistrer sous...', 'Expliquer', 'ou créer une', and 'Nouvelle requête'. Below these buttons, there is a horizontal menu with 'Requêtes récentes', 'Requête', 'Journal' (selected), 'Colonnes', 'Résultats', and 'Graphique'. The bottom section shows the execution log:

```
INFO : Number of reduce tasks is set to 0 since there's no reduce operator
INFO : number of splits:1
INFO : Submitting tokens for job: job_1457303073214_0003
INFO : The url to track the job: http://cdhmasternode.degenio.com:8088/proxy/application_1457303073214_0003/
INFO : Starting Job = job_1457303073214_0003, Tracking URL = http://cdhmasternode.degenio.com:8088/proxy/application_1457303073214_0003/
INFO : Kill Command = /opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/hadoop/bin/hadoop job -kill job_1457303073214_0003
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
INFO : 2016-03-06 18:12:03,339 Stage-1 map = 0%, reduce = 0%
```

# Résultat

```
1 select * from staging4;
```

Exécuter

Enregistrer

Enregistrer sous...

Expliquer

ou créer une

Nouvelle requête

...

Requêtes récentes

Requête

Journal

Colonnes

Résultats

Graphique

	◆ staging4.freq_consalcool	◆ staging4.consommationtabac	◆ staging4.diagnostic
1	Une fois par semaine	1	O
2	rarement ou jamais	-1	N
3	rarement ou jamais	-1	N
4	Une fois par semaine	-1	O
5	Une fois par semaine	0	N
6	Une fois par semaine	-1	N
7	Plusieurs fois par semaine	-1	N
8	Une fois par semaine	-1	N
9	rarement ou jamais	-1	N
10	Une fois par semaine	0	N
11	Plusieurs fois par semaine	0	N
12	Une fois par semaine	1	N
13	rarement ou jamais	1	N

# Pourcentage des gens ayant une anomalie par rapport à la consommation d'alcool et de tabac

```
1 CREATE TABLE diagno_tabac
2 AS
3
4
5 SELECT concat(Freq_ConsAlcool,'_',ConsommationTabac) My_key ,COUNT(*) percent_altered
6
7 FROM staging_age
8
9 WHERE diagnostic='0'
10
11 GROUP BY Freq_ConsAlcool,ConsommationTabac
12
13 ;
```

Exécuter

Enregistrer

Enregistrer sous...

Expliquer

ou créer une

Nouvelle requête

Requêtes récentes

Requête

Journal

Colonnes

Résultats

Graphique

```
INFO : set mapreduce.job.reduces=<number>
INFO : number of splits:1
INFO : Submitting tokens for job: job_1457303073214_0016
INFO : The url to track the job: http://cdhmasternode.degenio.com:8088/proxy/application_1457303073214_0016/
INFO : Starting Job = job_1457303073214_0016, Tracking URL = http://cdhmasternode.degenio.com:8088/proxy/application_1457303073214_0016/
INFO : Kill Command = /opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/hadoop/bin/hadoop job -kill job_1457303073214_0016
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2016-03-06 21:09:52,058 Stage-1 map = 0%, reduce = 0%
INFO : 2016-03-06 21:09:59,247 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.97 sec
INFO : 2016-03-06 21:10:07,460 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.37 sec
INFO : MapReduce Total cumulative CPU time: 2 seconds 370 msec
INFO : Ended Job = job_1457303073214_0016
INFO : Moving data to: hdfs://cdhmasternode.degenio.com:8020/user/hive/warehouse/projet_db.db/diagno_tabac from hdfs://cdhmasternode.degenio.com:8020/user/hive/warehouse/projet_db.db/diagno_tabac
INFO : Table projet_db.diagno_tabac stats: [numFiles=1, numRows=7, totalSize=186, rawDataSize=179]
```

# Résultat

```
1 SELECT * FROM diagno_tabac;
```

Exécuter

Enregistrer

Enregistrer sous...

Expliquer

ou créer une

Nouvelle requête

Requêtes récentes

Requête

Journal

Colonnes

Résultats

Graphique

	diagno_tabac.my_key	diagno_tabac.percent_altered
0	Plusieurs fois par semaine_-1	2
1	Plusieurs fois par semaine_0	2
2	Une fois par semaine_-1	3
3	Une fois par semaine_0	1
4	Une fois par semaine_1	2
5	rarement ou jamais_-1	1
6	rarement ou jamais_1	1

# Transformation de data

The screenshot displays a web-based interface for data transformation. On the left, a sidebar shows the database structure for 'projet\_db', including tables like 'dataset', 'diagno\_alcool\_tabac', and 'matable\_tp'. The main area contains a SQL query for creating a table 'staging\_age' based on 'matable\_tp'. The query uses a CASE statement to categorize 'FreqConsAlcool' values into age ranges. Below the query are buttons for 'Exécuter', 'Enregistrer', and 'Expliquer'. At the bottom, a 'Journal' tab shows the execution log, which includes job ID, tracking URL, and progress information for Stage-1.

**BASE DE DONNÉES**

projet\_db

**TABLES**

- dataset
- diagno\_alcool\_tabac
- diagno\_alcool\_tabac1
- diagno\_alcool\_tabac2
- diagnocons
- diagnocons2
- diagnoconso
- matable\_tp
- staging
- staging4

```
1 create table staging_age as select case FreqConsAlcool
2
3 when '0.2' then 'Plusieurs fois par jour'
4
5 when '0.4' then 'Une fois par jour'
6
7 when '0.6' then 'Plusieurs fois par semaine'
8
9 when '0.8' then 'Une fois par semaine'
10
11 when '1' then 'rarement ou jamais'
12
13 end Freq_ConsAlcool
14
15 ,ConsommationTabac
16
17 ,case AGE
18
19 when '0.5*' then '18_22'
20
21 when '0.6*' then '22_26'
22
23 when '0.7*' then '25_29'
24
25 when '0.8*' then '29_32'
26
27 else '32_36'
28
29 end TRANCHE_AGE
30 |,Diagnostic
31 from matable_tp;
```

Exécuter Enregistrer Enregistrer sous... Expliquer ou créer une Nouvelle requête

Requêtes récentes Requête Journal Colonnes Résultats Graphique

INFO : Starting Job = job\_1457303073214\_0012, Tracking URL = http://cdhmasternode.degenio.com:8088/proxy/application\_1457303073214\_0012/  
INFO : Kill Command = /opt/cloudera/parcels/CDH-5.5.2-1.cdh5.5.2.p0.4/lib/hadoop/bin/hadoop job -kill job\_1457303073214\_0012  
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0  
INFO : 2016-03-06 20:41:20,602 Stage-1 map = 0%, reduce = 0%  
INFO : 2016-03-06 20:41:27,785 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.08 sec  
INFO : MapReduce Total cumulative CPU time: 1 seconds 80 msec  
INFO : Ended Job = job\_1457303073214\_0012  
INFO : Stage-4 is selected by condition resolver.  
INFO : Stage-3 is filtered out by condition resolver.

# % anomalie par rapport: âge, alcool & tabac

```
1 select * from diagno_age1;
```

Exécuter

Enregistrer

Enregistrer sous...

Expliquer

ou créer une

Nouvelle requête

Requêtes récentes

Requête

Journal

Colonnes

Résultats

Graphique

	◆ diagno_age1.my_key	◆ diagno_age1.percent_altered
0	Plusieurs fois par semaine_-1_32_36	2
1	Plusieurs fois par semaine_0_32_36	2
2	Une fois par semaine_-1_32_36	3
3	Une fois par semaine_0_32_36	1
4	Une fois par semaine_1_32_36	2
5	rarement ou jamais_-1_32_36	1
6	rarement ou jamais_1_32_36	1



# Liste de nos requêtes

Nom	Application Type	Statut	Utilisateur	Maps	Reduces	File d'attente	Priorité	Durée	Envoyé
CREATE TABLE diagno_t...ol,ConsommationTabac(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	20s	03/06/16 18:09:46
CREATE TABLE diagno_a...ionTabac,tranche_age(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	19s	03/06/16 18:01:03
CREATE TABLE diagno_a...ionTabac,tranche_age(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	19s	03/06/16 17:55:06
SELECT concat(Freq_Co...ionTabac,tranche_age(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	19s	03/06/16 17:49:39
create table staging_age as sel...matable_tp(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	12s	03/06/16 17:41:15
CREATE TABLE diagno_a...ol,ConsommationTabac(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	19s	03/06/16 16:44:56
create table diagno_a...'+ConsommationTabac(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	19s	03/06/16 16:30:09
create table diagno_alcool_tabac2 as s...3(Stage-2)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	17s	03/06/16 16:19:36
create table diagno_alcool_tabac2 as s...3(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	19s	03/06/16 16:19:14
create table diagno_alcool_tabac1 as s...3(Stage-2)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	18s	03/06/16 15:38:57
create table diagno_alcool_tabac1 as s...3(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	19s	03/06/16 15:38:35
create table diagno_alcool_tabac as se...3(Stage-2)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	19s	03/06/16 15:20:34
create table diagno_alcool_tabac as se...3(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	20s	03/06/16 15:20:11
create table staging3 as select...matable_tp(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	13s	03/06/16 15:11:57
create table staging4 as select...matable_tp(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	13s	03/06/16 15:02:55
create table staging3 as select...matable_tp(Stage-1)	MAPREDUCE	SUCCEEDED	admin	100%	100%	root.admin	N/A	15s	03/06/16 14:58:04

# Traitement avec Pig

The screenshot displays the Hue Job Browser interface. At the top, there is a navigation bar with 'HUE' logo, home icon, and menu items: 'Query Editors', 'Data Browsers', 'Workflows', and 'Search'. On the right side of the navigation bar are 'File Browser', 'Job Browser', and a user profile dropdown for 'cloudera' with options for 'Edit Profile' and 'Manage Users'. Below the navigation bar, the 'Job Browser' title is shown. There are input fields for 'Username' (set to 'cloudera') and 'Text' (placeholder 'Search for text'). To the right of these fields are four status filters: 'Succeeded' (green), 'Running' (orange), 'Failed' (red), and 'Killed' (grey). The main area contains a table of job entries with the following columns: Logs, ID, Name, Application Type, Status, User, Maps, Reduces, Queue, Priority, Duration, and Submitted. Three jobs are listed, all with a 'SUCCEEDED' status. The first job is 'PigLatin:script.pig' with ID '1457477354887\_0003'. The second and third jobs are 'oozie:launcher:T=pig;W=pig-app-hue-script:A=pig;ID=0000001-160308145025218-oozie-oozi-W' with IDs '1457477354887\_0002' and '1457477354887\_0001' respectively. At the bottom left, it says 'Showing 1 to 3 of 3 entries'. At the bottom right, there are navigation buttons: '- Previous', '1', and 'Next ->'. The URL at the bottom of the browser window is 'http://quickstart.cloudera:8888/obrowser/index.html#'.

Logs	ID	Name	Application Type	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
	1457477354887_0003	PigLatin:script.pig	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	48s	03/08/16 15:17:45
	1457477354887_0002	oozie:launcher:T=pig;W=pig-app-hue-script:A=pig;ID=0000001-160308145025218-oozie-oozi-W	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	1m:50s	03/08/16 15:17:21
	1457477354887_0001	oozie:launcher:T=pig;W=pig-app-hue-script:A=pig;ID=0000000-160308145025218-oozie-oozi-W	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	19s	03/08/16 15:15:23

# Transfert des données à partir de HIVE vers la machine local

- `INSERT OVERWRITE LOCAL DIRECTORY  
`usr/local/example/` ROW FORMAT  
DELIMITED FIELDS TERMINATED BY ``,` select *  
from dataset;`

# Chargement du datas

- `mes_donnees=Load  
'/user/admin/dataset2.txt' using  
pigstorage(',')  
As(Saison:float,Age:float,maladies_Pueriles:int  
,Accident_Traumatisme :int,  
Intervention_Chirurgicale:int,GrossesFievres:i  
nt,FreqConsAlcool:CHARARRAY,Consommation  
Tabac:int,Nbre_Hr_Assis:int,  
Diagnostic:CHARARRAY);`

# Traitement avec PIG

- affiche\_lim=LIMIT mes\_donnees 50;
- DUMP affiche\_lim;
  
- Data\_gp\_age\_al\_tabac=GROUP mes\_donnees
- Data\_gp\_age\_al\_tabac= FOREACH  
mes\_donnees GENERATE  
group(alcool,tabac,age) as my\_key, count(\*) as  
pourcentage;
- Order\_result=ORDER Data\_gp\_age\_al\_tabac  
BY pourcentage DESC;
- DUMP Order\_result;

# Problèmes rencontrés

- Matériels
- Configuration
- Expertise

# Compétences acquises

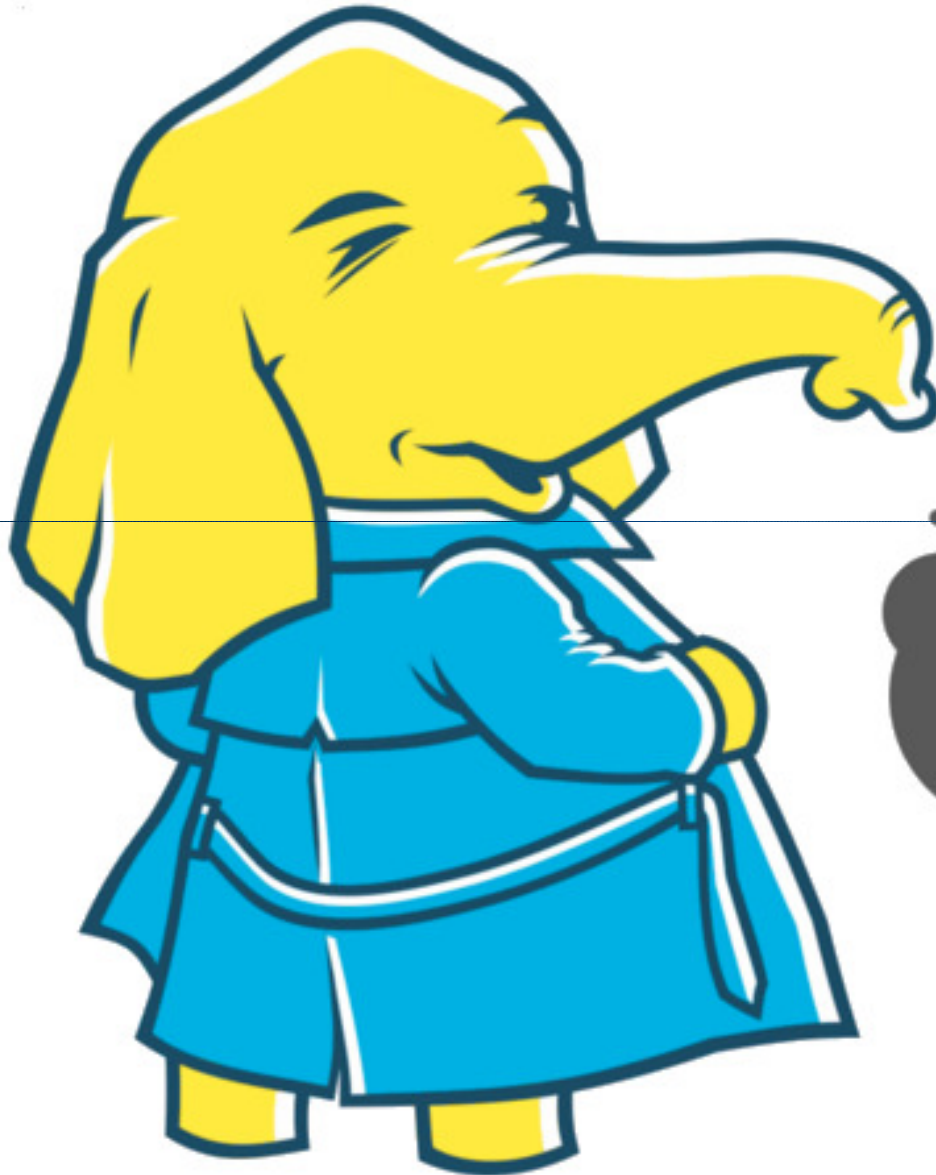
- Mise en place d'un environnement Hadoop
- Cloudera Manager
- Linux

# Conclusion

- Ce projet nous a permis d'appréhender l'environnement Hadoop, Mapreduce, Cloudera Manager
- Familiarisation de l'environnement BIG DATA
- Nuits d'insomnie....
- MERCI



Question ???



MY HADOOP  
IS **BIGGER**  
THAN YOURS...