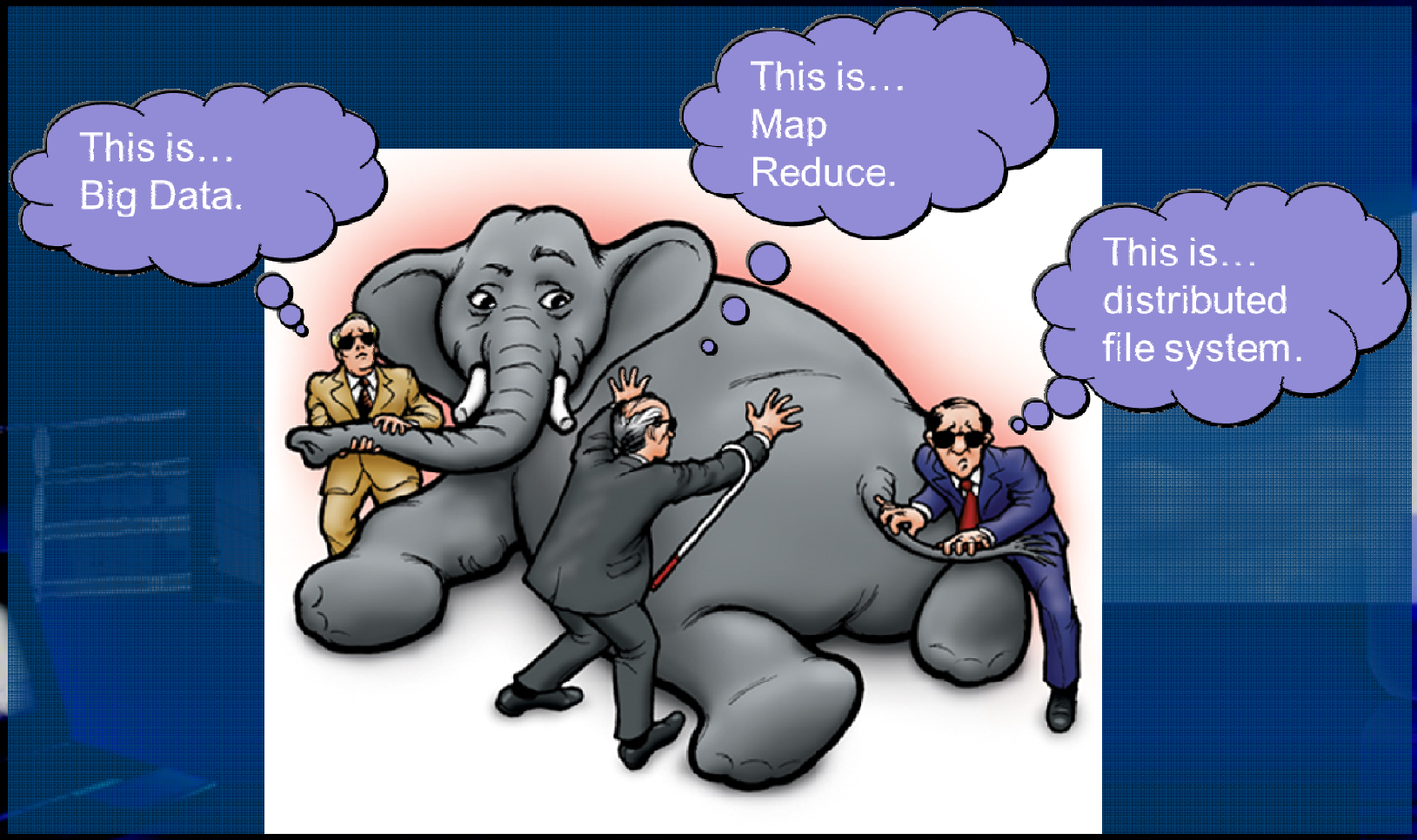
PROJET DE SESSION



PROJET HADOOP

Présentation de l’équipe

- Client
- Administrateur
- Développeur

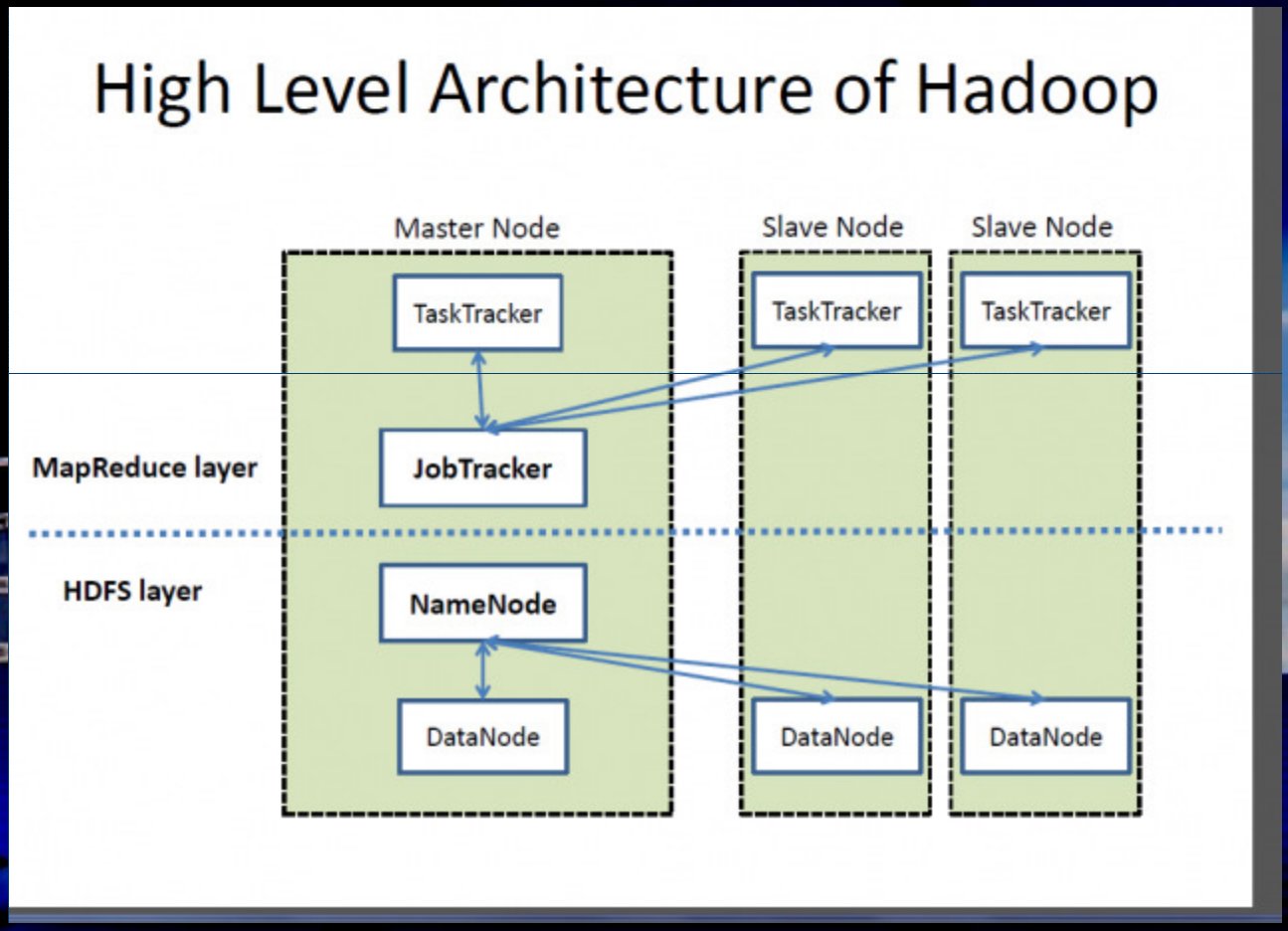
BESOIN

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

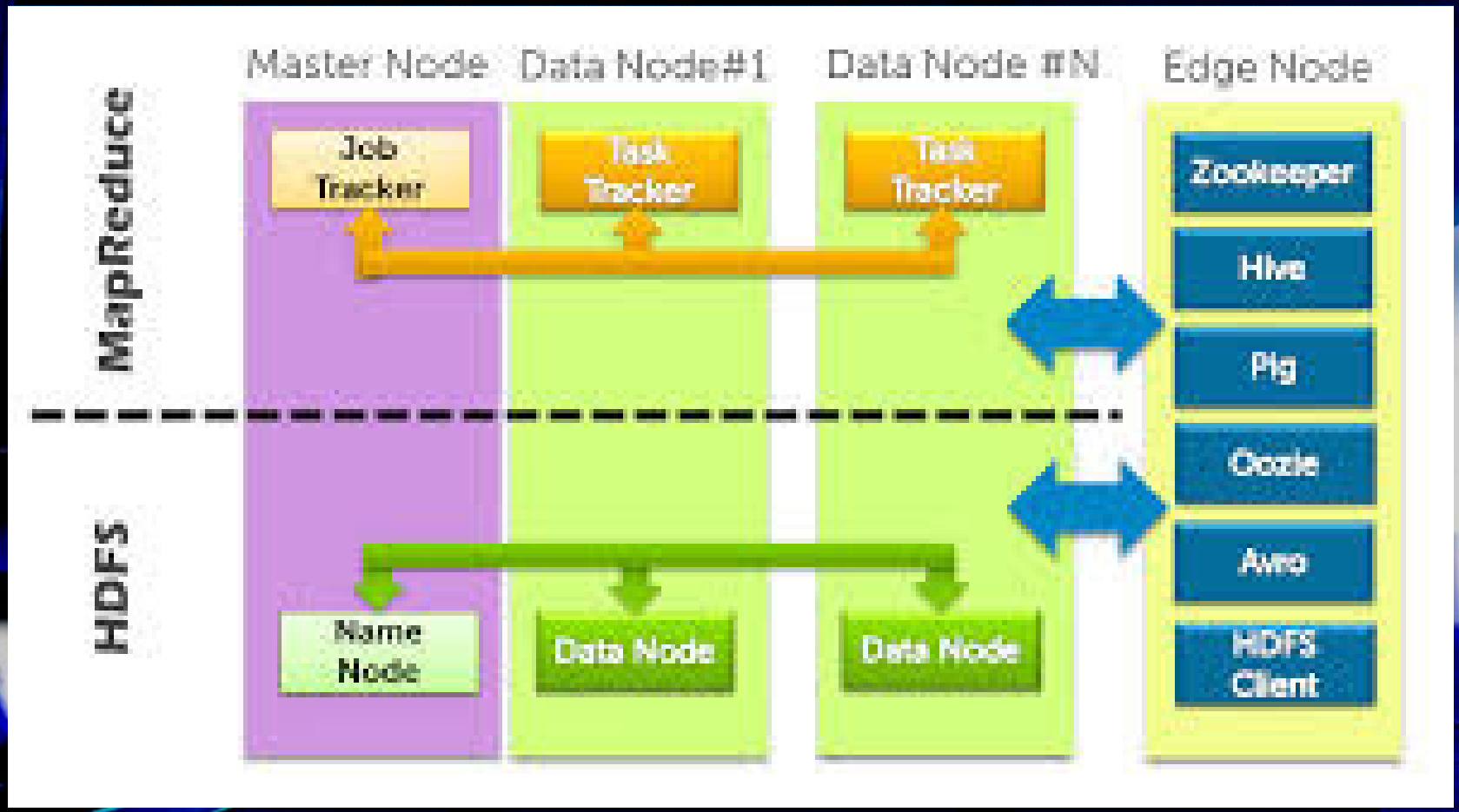
Installation du cluster

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

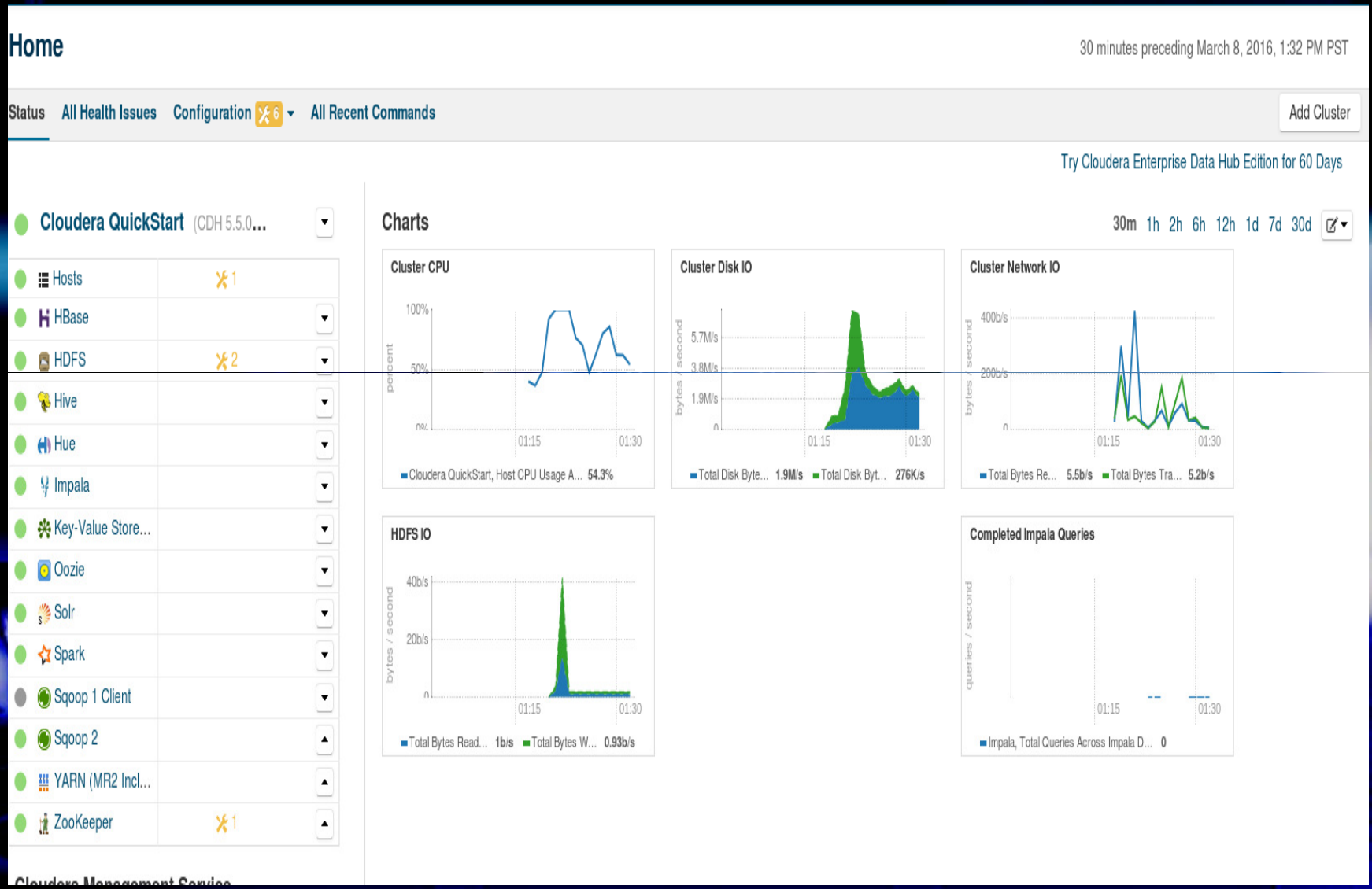
Architecture du cluster



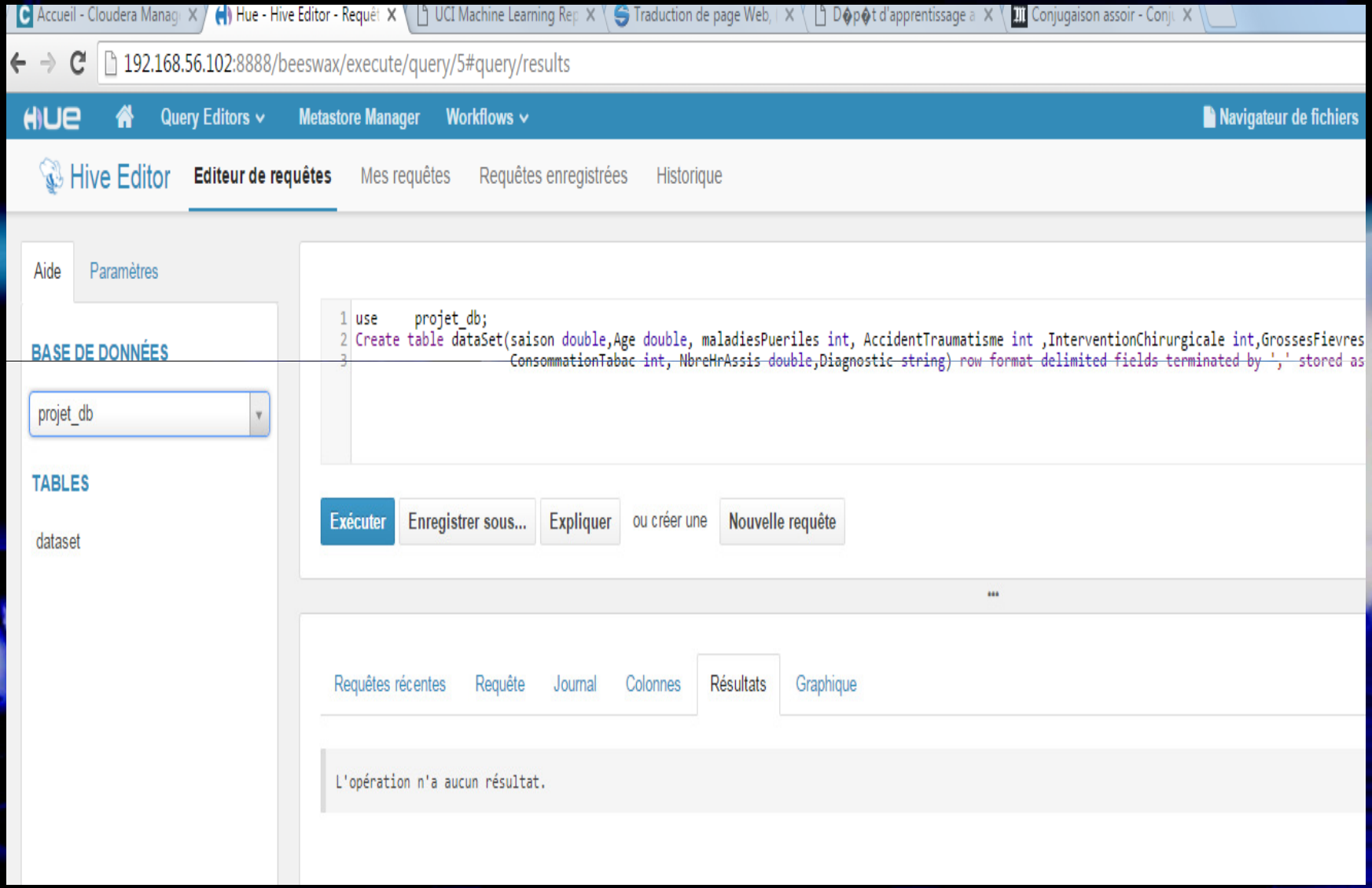
Cloudera Manager



Notre machine



Déploiement du code – Hive



Structure de dataset

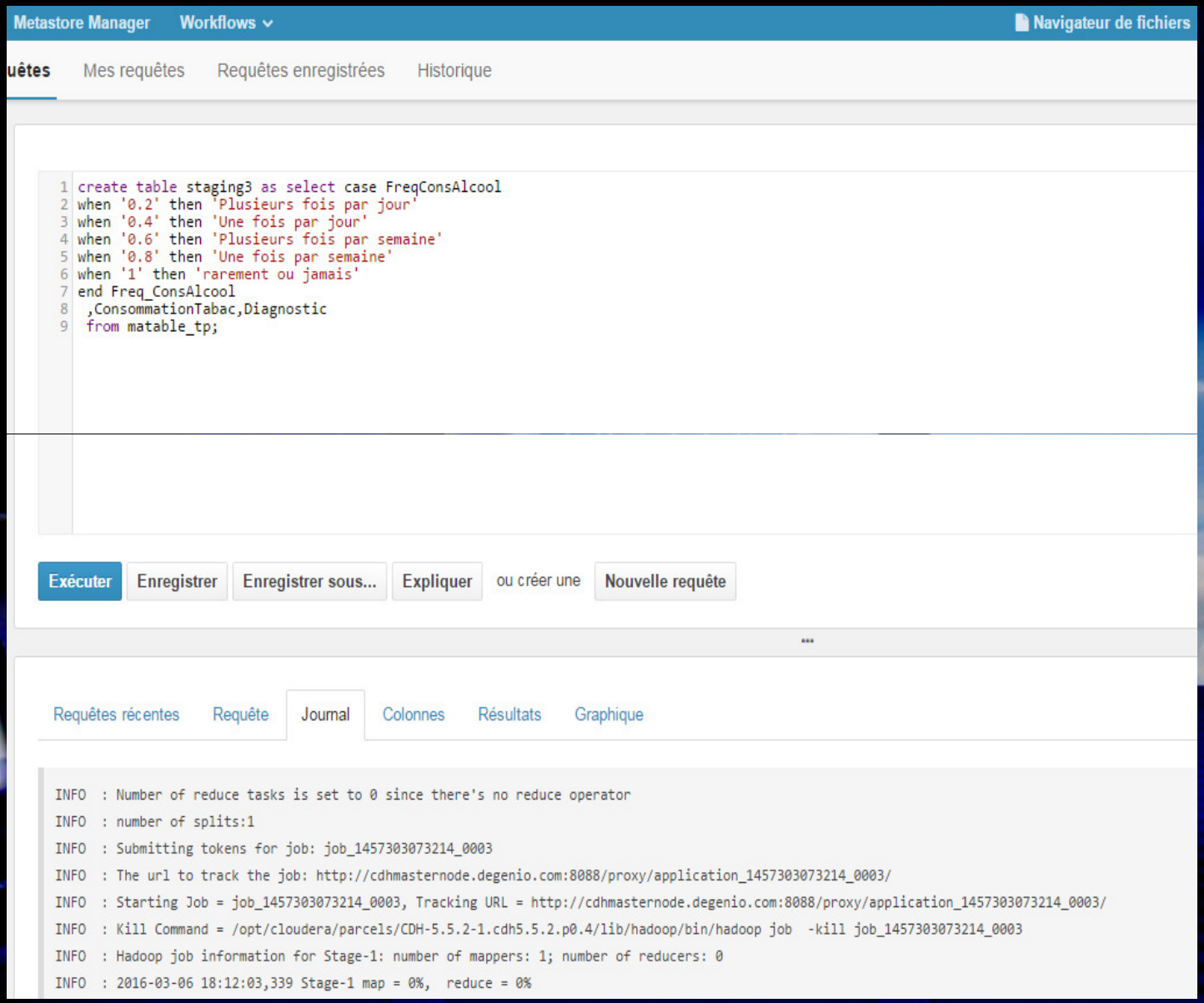
```
1 SELECT * FROM matable_tp;
```

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.grossesfievres	matable_tp.freqc
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



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

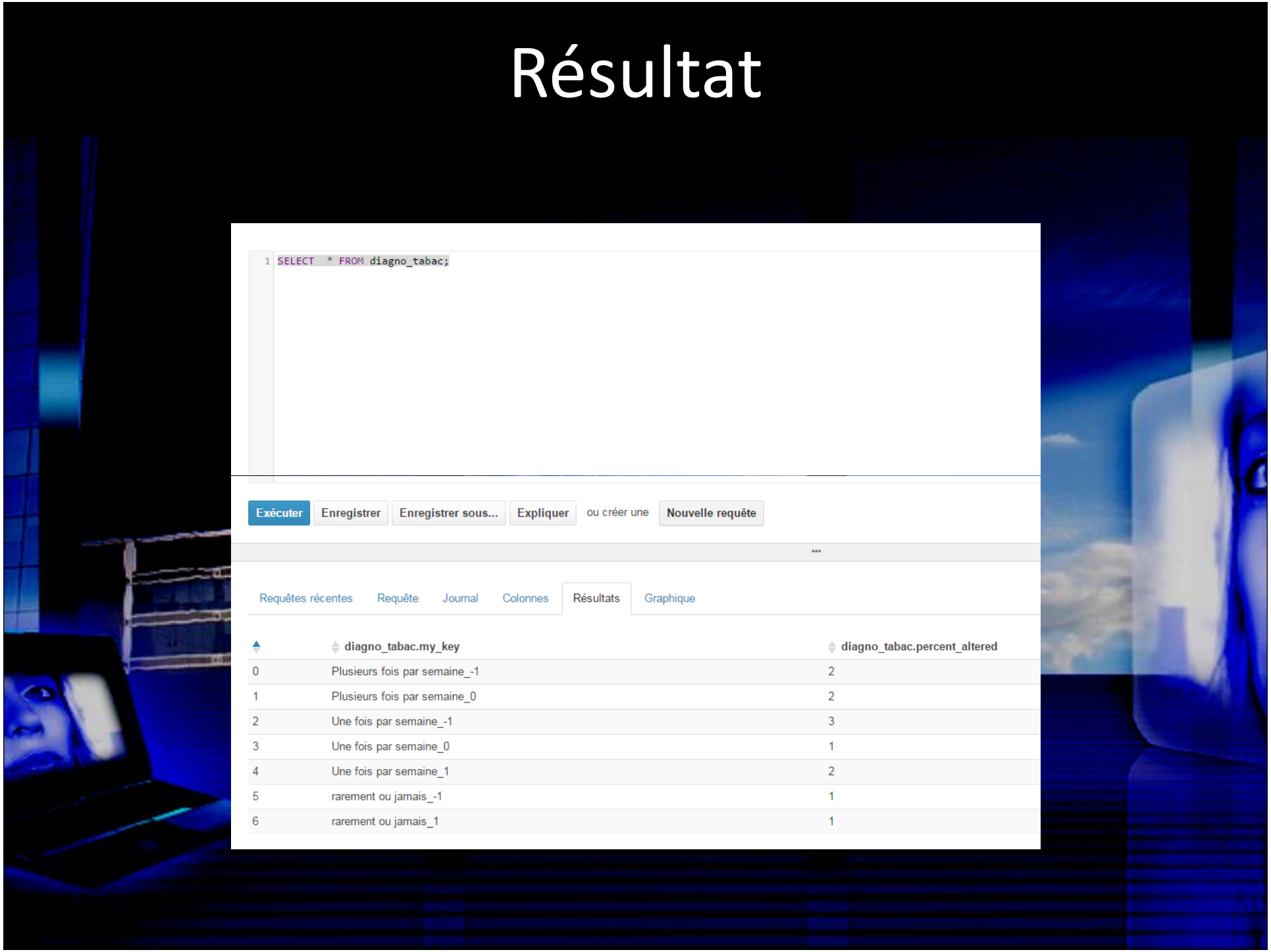
```
1  CREATE TABLE diagno_tabac
2
3  AS
4
5  SELECT concat(Freq_ConsAlcool,'_',ConsommationTabac) My_key ,COUNT(*) percent_altered
6
7  FROM  staging_age
8
9  WHERE diagnostic='O'
10
11 GROUP BY Freq_ConsAlcool,ConsommationTabac
12
13 ;
```

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.deger
aging_hive_2016-03-06_21-09-46_241_4510722684840432510-1/-ext-10001
INFO  : Table projet_db.diagno_tabac stats: [numFiles=1, numRows=7, totalSize=186, rawDataSize=179]
```

Résultat



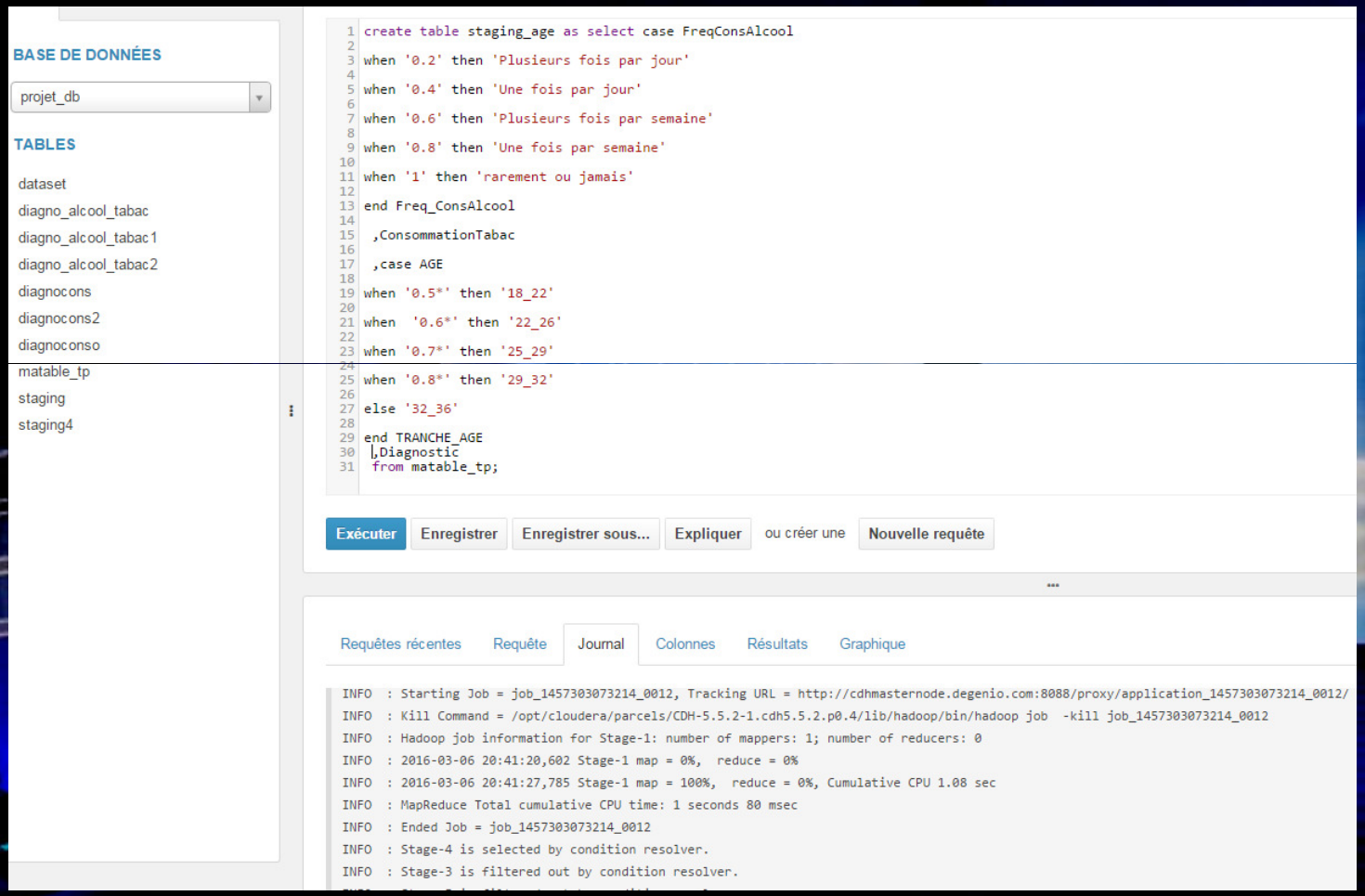
```
1 SELECT  * FROM diagno_tabac;
```

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



% anomalie par rapport: âge, alcool & tabac

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

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

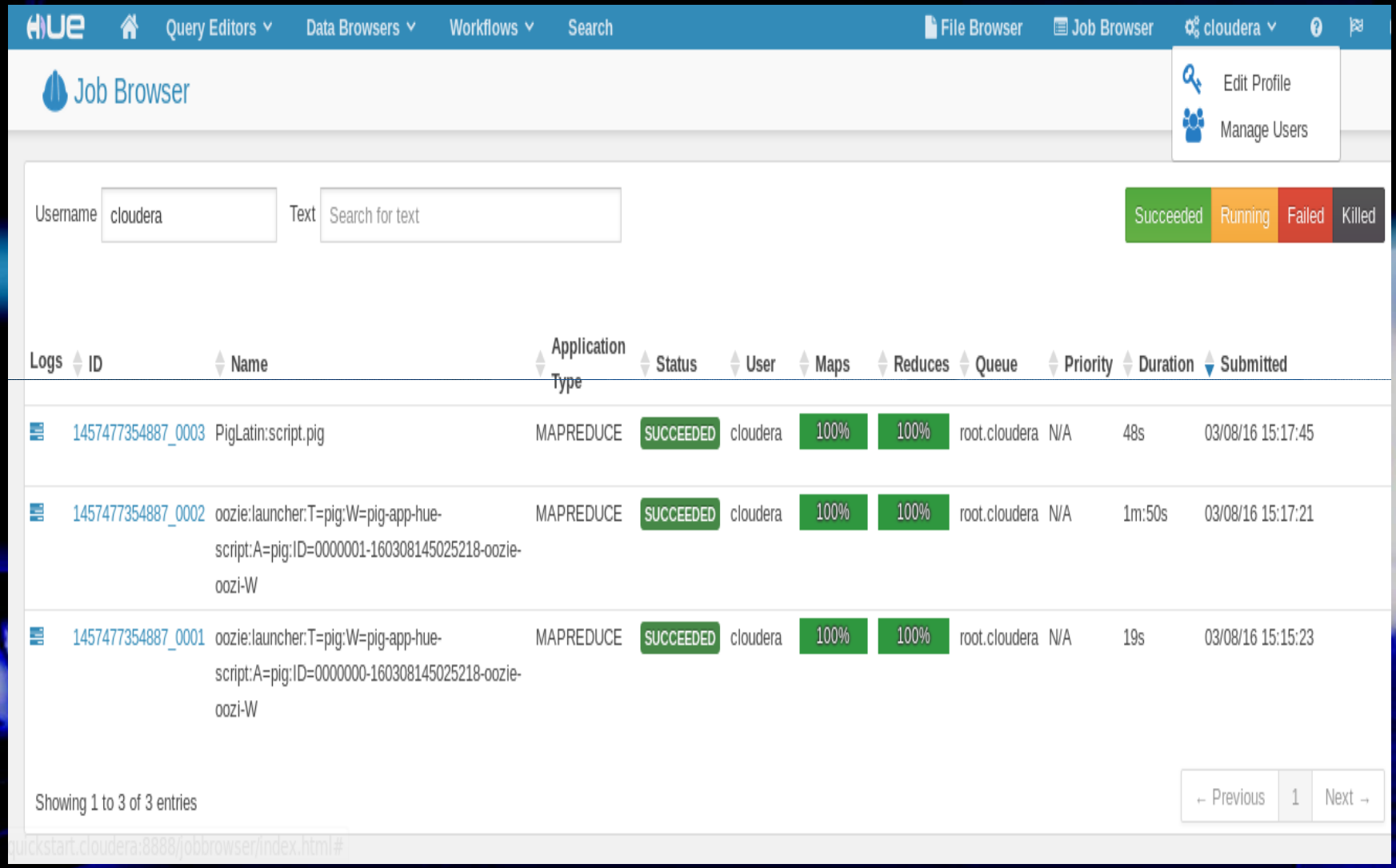
Requêtes récentes | Requête | Journal | Colonnes | Résultats | Graphique

	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



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 ???