Collège de Bois-de-Boulogne

Une étude des retards des vols aeriens

par

Mohamed Reda Louahala et Albert Zhu

Travail pratique pour le cours 420-BD3

Mars 2016

Collège de Bois-de-Boulogne

Introduction

Le transport aérien est devenu indispensable de nos jours, comprendre les phénomènes qui ont un impact direct sur la fiabilité est devenu une priorité pour établir des plans d'actions et de mesures pour améliorer ce réseau complexe.

Collège de Bois-de-Boulogne

Plan de la présentation

1. Context
2. Creation d'un cluster Hadoop
3. Developpement de l'application
4. Résultat et discussion
5. Conclusion
6. Compétences acquises

Collège de Bois-de-Boulogne

Besoin client

- De quels Airoports les vols sont en retard le plus frequents ?

- A quelles destinations les vols sont fréquement en retard ?

Collège de Bois-de-Boulogne

Source de données

- Un data set qui nous fournit les données concernant les vols aériens, et surtout les données necessaires pour répondre à nos besoins (Aéroport, délais d'attente … etc)

- Contient environ 7009729 d'enregistrements, 673Mo (pour l'année 2008)

Collège de Bois-de-Boulogne

Un échantillon (l'année 2008)

```
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum
2008,1,3,4,2003,1955,2211,2225,WN,335,N712SW,128,150,116,-14,8,IAD,TPA,810,4,8,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,754,735,1002,1000,WN,3231,N772SW,128,145,113,2,19,IAD,TPA,810,5,10,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,628,620,804,750,WN,448,N428WN,96,90,76,14,8,IND,BWI,515,3,17,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,926,930,1054,1100,WN,1746,N612SW,88,90,78,-6,-4,IND,BWI,515,3,7,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1829,1755,1959,1925,WN,3920,N464WN,90,90,77,34,34,IND,BWI,515,3,10,0,,0,2,0,0,0,32
2008,1,3,4,1940,1915,2121,2110,WN,378,N726SW,101,115,87,11,25,IND,JAX,688,4,10,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1937,1830,2037,1940,WN,509,N763SW,240,250,230,57,67,IND,LAS,1591,3,7,0,,0,10,0,0,0,47
2008,1,3,4,1039,1040,1132,1150,WN,535,N428WN,233,250,219,-18,-1,IND,LAS,1591,7,7,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,617,615,652,650,WN,11,N689SW,95,95,70,2,2,IND,MCI,451,6,19,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1620,1620,1639,1655,WN,810,N648SW,79,95,70,-16,0,IND,MCI,451,3,6,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,706,700,916,915,WN,100,N690SW,130,135,106,1,6,IND,MCO,828,5,19,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1644,1510,1845,1725,WN,1333,N334SW,121,135,107,80,94,IND,MCO,828,6,8,0,,0,8,0,0,0,72
2008,1,3,4,1426,1430,1426,1425,WN,829,N476WN,60,55,39,1,-4,IND,MDW,162,9,12,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,715,715,720,710,WN,1016,N765SW,65,55,37,10,0,IND,MDW,162,7,21,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1702,1700,1651,1655,WN,1827,N420WN,49,55,35,-4,2,IND,MDW,162,4,10,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1029,1020,1021,1010,WN,2272,N263WN,52,50,37,11,9,IND,MDW,162,6,9,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1452,1425,1640,1625,WN,675,N286WN,228,240,213,15,27,IND,PHX,1489,7,8,0,,0,3,0,0,0,12
2008,1,3,4,754,745,940,955,WN,1144,N778SW,226,250,205,-15,9,IND,PHX,1489,5,16,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1323,1255,1526,1510,WN,4,N674AA,123,135,110,16,28,IND,TPA,838,4,9,0,,0,0,0,0,0,16
2008,1,3,4,1416,1325,1512,1435,WN,54,N643SW,56,70,49,37,51,ISP,BWI,220,2,5,0,,0,12,0,0,0,25
2008,1,3,4,706,705,807,810,WN,68,N497WN,61,65,51,-3,1,ISP,BWI,220,3,7,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1657,1625,1754,1735,WN,623,N724SW,57,70,47,19,32,ISP,BWI,220,5,5,0,,0,7,0,0,0,12
2008,1,3,4,1900,1840,1956,1950,WN,717,N786SW,56,70,49,6,20,ISP,BWI,220,2,5,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1039,1030,1133,1140,WN,1244,N714CB,54,70,47,-7,9,ISP,BWI,220,2,5,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,801,800,902,910,WN,2101,N222WN,61,70,53,-8,1,ISP,BWI,220,3,5,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1520,1455,1619,1605,WN,2553,N394SW,59,70,50,14,25,ISP,BWI,220,2,7,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,1422,1255,1657,1610,WN,188,N215WN,155,195,143,47,87,ISP,FLL,1093,6,6,0,,0,40,0,0,0,7
2008,1,3,4,1954,1925,2239,2235,WN,1754,N243WN,165,190,155,4,29,ISP,FLL,1093,3,7,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,636,635,921,945,WN,2275,N454WN,165,190,147,-24,1,ISP,FLL,1093,5,13,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,734,730,958,1020,WN,550,N712SW,324,350,314,-22,4,ISP,LAS,2283,2,8,0,,0,NA,NA,NA,NA,NA
```

Collège de Bois-de-Boulogne

Installation du cluster

A cause de la limitation matériel, nous avons fait une installation personnalisée (sans Hbase, Spark, ...)

Spécifications materiels et logiciels

1 - Un namenode (11,4 Go de mémoire et 30 Go de disque dure)

2 - Deux datanodes (1,8 Go de mémoire et 30 Go de disque dure)

3 - Système d'exploitation Linux CentOS 6.6 sur tous les nœuds

Étapes:

1 - Préparation des VMs

2 - Installation d'Apache Ambari et Configuration de la base de données (Postgres)

3 - Installation de Hadoop Hortonworks avec Ambari

4 - déployement des vues (HDFS, HIVE, etc.)

Collège de Bois-de-Boulogne

Problèmes et solutions

1 - Limitation de ressources matériels:

- Nous avons augmenté la mémoire du namenode pour que les services démarrent.

- Nous avons désactivé quelques services qui n'était pas necessaire pour le fonctionnement de HIVE.

2 - Ressource manager ne démarait pas:

- nous avons enlevé deux paramètres, dans la configuration de resource manager.

yarn.scheduler.capacity.root.accessible-node-labels.default.capacity

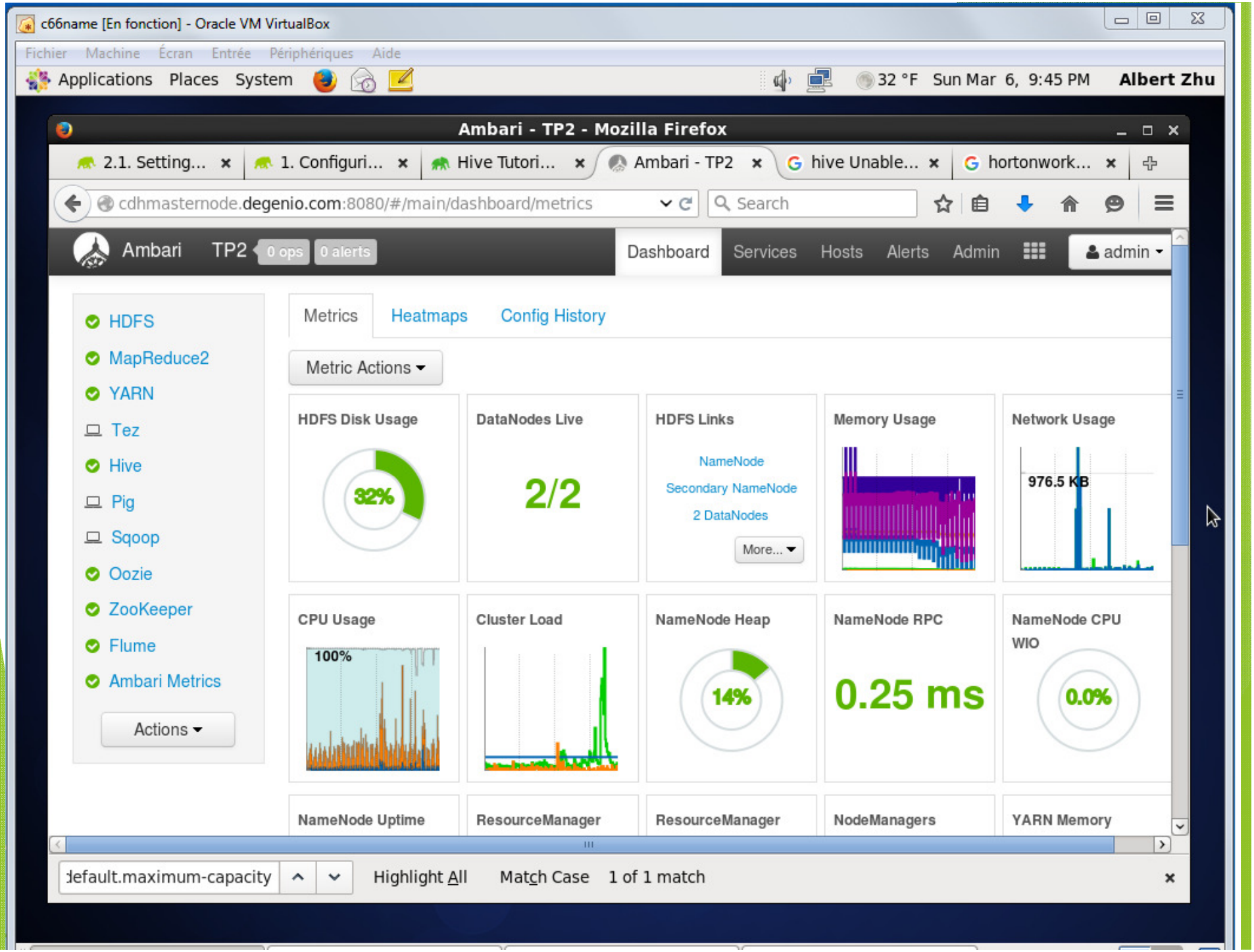
yarn.scheduler.capacity.root.accessible-node-labels.default.maximum-capacity

Collège de Bois-de-Boulogne

Références installation cluster

1 - "Hadoop Cluster setup: Hortonworks Hadoop Installation using Apache Ambari on CentOS6.“

2 - Le cahier du Lab 09 vu en cours


c66name [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide
Applications Places System
32 °F Sun Mar 6, 9:45 PM Albert Zhu
Ambari - TP2 - Mozilla Firefox
2.1. Setting...
1. Configuri...
Hive Tutori...
Ambari - TP2
hive Unable...
hortonwork...
cdhmasternode.degenio.com:8080/#/main/dashboard/metrics
Search
Ambari TP2 0 ops 0 alerts
Dashboard Services Hosts Alerts Admin
admin
HDFS
MapReduce2
YARN
Tez
Hive
Pig
Sqoop
Oozie
ZooKeeper
Flume
Ambari Metrics
Actions
Metrics Heatmaps Config History
Metric Actions
HDFS Disk Usage
32%
DataNodes Live
2/2
HDFS Links
NameNode
Secondary NameNode
2 DataNodes
More...
Memory Usage
Network Usage
976.5 KB
CPU Usage
100%
Cluster Load
NameNode Heap
14%
NameNode RPC
0.25 ms
NameNode CPU WIO
0.0%
NameNode Uptime
ResourceManager
ResourceManager
NodeManagers
YARN Memory
default.maximum-capacity
Highlight All
Match Case
1 of 1 match

Collège de Bois-de-Boulogne

Développement de l'application (1)

- Outil du développement: Hive

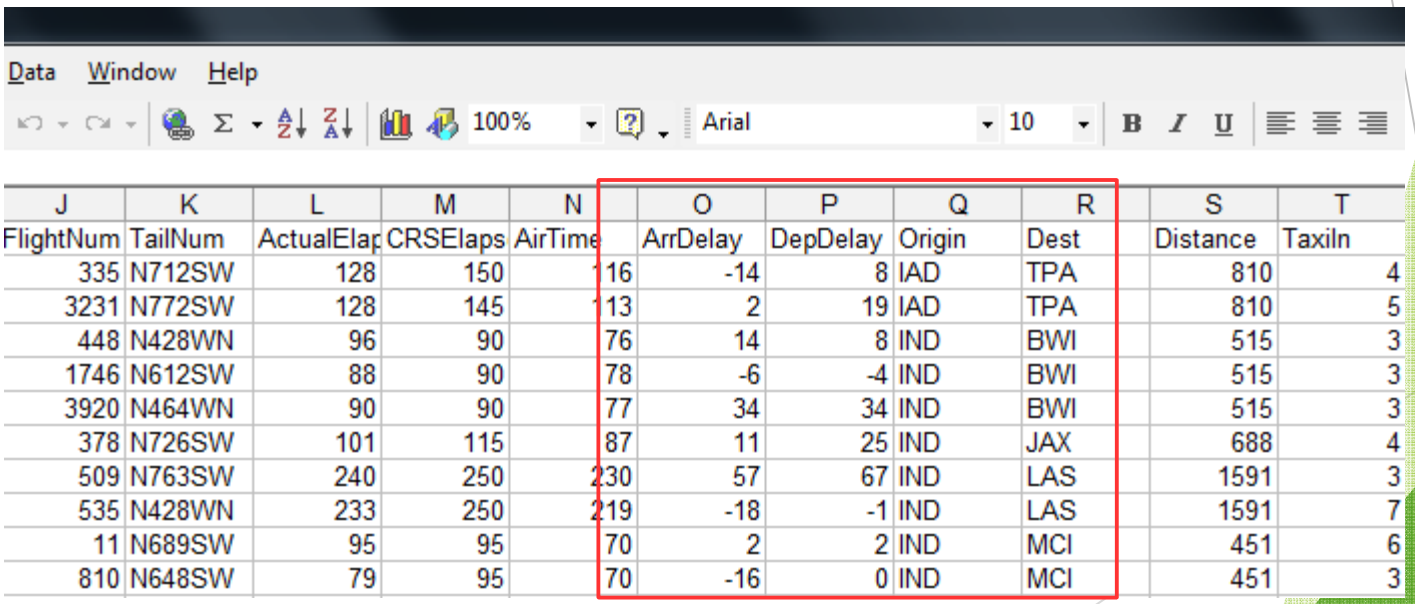
- Langage: HiveQL

- Autres logiciels: MS Excel

Collège de Bois-de-Boulogne

Développement de l'application (2)

Exploration de données avec Excel



J	K	L	M	N	O	P	Q	R	S	T
FlightNum	TailNum	ActualElap	CRSElaps	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn
335	N712SW	128	150	116	-14	8	IAD	TPA	810	4
3231	N772SW	128	145	113	2	19	IAD	TPA	810	5
448	N428WN	96	90	76	14	8	IND	BWI	515	3
1746	N612SW	88	90	78	-6	-4	IND	BWI	515	3
3920	N464WN	90	90	77	34	34	IND	BWI	515	3
378	N726SW	101	115	87	11	25	IND	JAX	688	4
509	N763SW	240	250	230	57	67	IND	LAS	1591	3
535	N428WN	233	250	219	-18	-1	IND	LAS	1591	7
11	N689SW	95	95	70	2	2	IND	MCI	451	6
810	N648SW	79	95	70	-16	0	IND	MCI	451	3

Collège de Bois-de-Boulogne

Développement de l'application (3)

1. Transfert du fichier de données sur HDFS;
2. Création d'une table temporaire (une colonne)
3. Chargement des données dans staging

HiveQL:

- create table temp_2008 (col_value STRING);
- LOAD DATA INPATH '/user/tp/2006.csv' OVERWRITE INTO TABLE temp_2008;

Collège de Bois-de-Boulogne

Développement de l'application (4)

4. Création d'une table de travail (4 champs: Origin, Destination, ArrivalDelay, DepatureDelay)
5. Extraction des données dans cette table

HiveQL:

- create table Delay2008 (Origin STRING, Destination STRING, ArrivalDelay INT, DepatureDelay INT);
- insert overwrite table Delay2008

```
select
regexp_extract(col_value, '^(?:([^,]*)\,?){17}', 1) Origin,
regexp_extract(col_value, '^(?:([^,]*)\,?){18}', 1) Destination,
regexp_extract(col_value, '^(?:([^,]*)\,?){15}', 1) ArrivalDelay,
regexp_extract(col_value, '^(?:([^,]*)\,?){16}', 1) DepatureDelay
from temp_data2008
```

Collège de Bois-de-Boulogne

Problèmes et solutions

1 - Pas d'accès au data set (problème de chargement):

- Nous avons données les droits (permissions d'écriture) sur le data set.

2 – Problème d'insertion de données dans la table:

- nous avons trouvé que Resource Manager était a l'origine de cette exception, donc on est revenu sur la correction de l'erreur au niveau du ressource manager.

Collège de Bois-de-Boulogne

Résultat

Executer le HiveQL suivant

select Origin, count(Origin) CountOrigin from Delay2008 group by Origin order by CountOrigin desc Limit 10;

Nous obtenons 10 aéroports où il y a plus de retard:

ATL 414513
ORD 350380
DFW 281281
DEN 241443
LAX 215608
PHX 199408
IAH 185172
LAS 172876
DTW 161989
SFO 140587

Collège de Bois-de-Boulogne

Discussion

Dans le script précédent, nous mettons en ordre les aéroports selon leurs nombre de fois de retard.

Nous pouvons aussi calculer le somme de la durée du temps de retard et classer les aéroports.

Collège de Bois-de-Boulogne

Conclusion

Nous avons trouvé, de ce projet, la puissance des technologies de BigData (Hadoop, Hive, etc.) dans le traitement des données massives. La taille de nos données est des centaines de gigabytes. Mais la durée du temps pour un de jeu de données (échantillon) se fait en quelques minutes.

Une machine assez puissante pour NameNode est très importante. Ça affecte la performance du cluster. (Nous avons déplacé un cluster à partir d'un ordinateur avec une mémoire de 16 G à un autre ordinateur avec une mémoire de 8 G. le cluster est presque gelé.)

Collège de Bois-de-Boulogne

Compétences acquises

- Installation et maintenance du cluster avec Ambari Apache (Hortonworks) et Cloudera Manager (tous les 2 sont réussites);
- Commandes principales HDFS;
- expriences utiles avec des composants differents de Hadoop : Hive, Pig, Oozie, Hue, etc;
- Développement des scripts HiveQL
- Approfondissement de la compréhension de la théorie de BigData au moyen de pratique;
- et plus d'autres;

Collège de Bois-de-Boulogne

Questions?