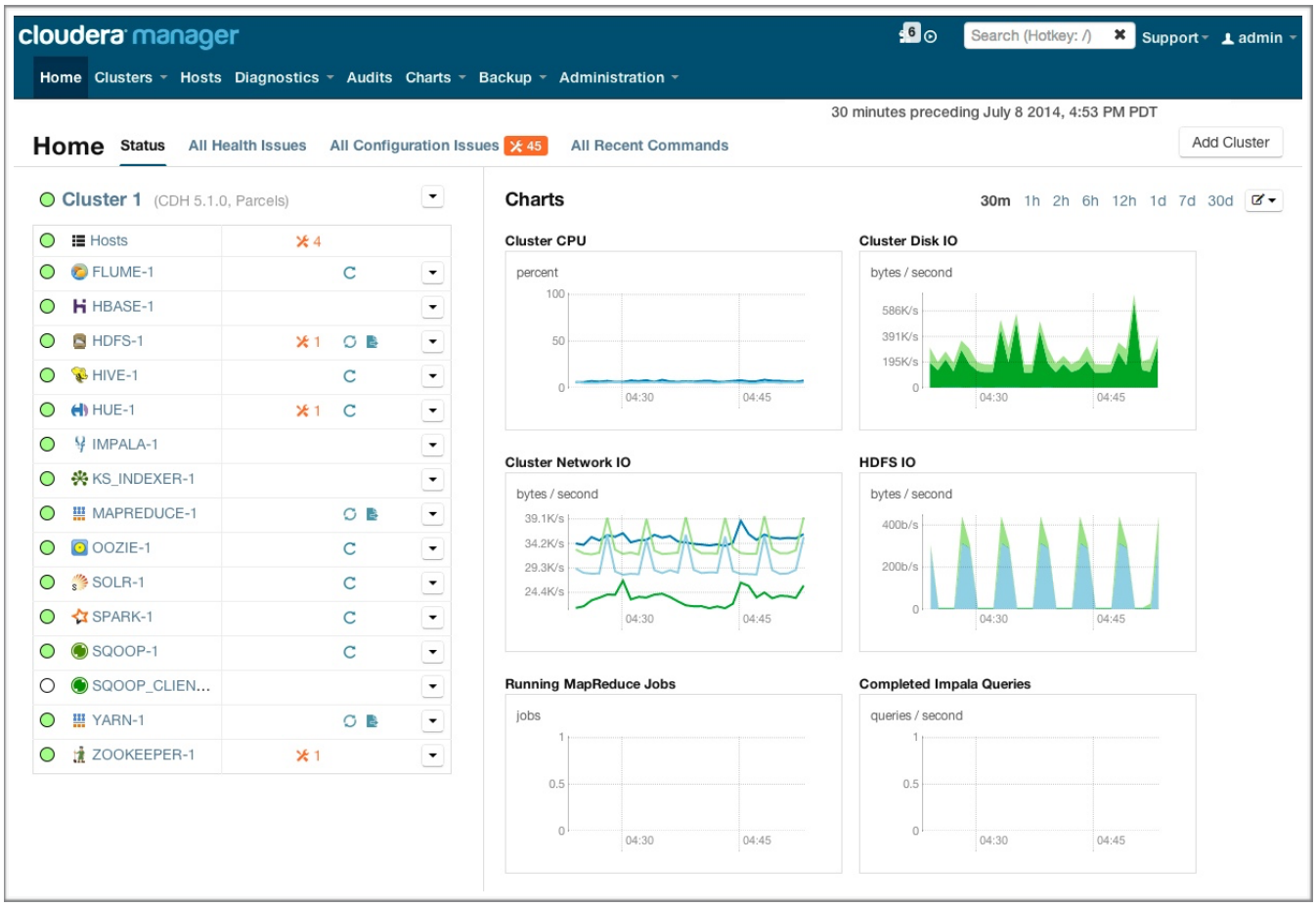
TP1: Mise en place d'un cluster Cloudera

Collège Bois-de-Boulogne

Cours: 420-BD3-BB - Session Automne 2015

Professeur: Hafed Benteftifa

Mario Nadon (1595857)



Introduction

L'objectif du travail est de choisir une distribution Cloudera ou Hortonworks, de procéder à l'installation d'un cluster de 2 machines et de le tester a l'aide d'une application de type Yarn MapReduce.

Le système d'exploitation devra être sélectionné selon la préférence de l'étudiant (CentOS, Debian, Ubuntu ou autre)

Caractéristiques du cluster

- Le système d'exploitation sélectionné est Linux CentOs 6.5
- La distribution Hadoop sélectionné est Cloudera 5.4.3
- Trois (3) machines virtuelles (Oracle VM VirtualBox) configuré selon les caractéristiques suivantes:

Configuration	Cloudera-Master	Cloudera-Datanode1	Cloudera-Datanode2
Système exploitation	CentOS 6.5 Software Development Workstation	CentOs 6.5 Minimal desktop	CentOs 6.5 minimal desktop
Mémoire vive	9GB	2,5GB	2,5GB
Espace disque	20GB	20Gb	20Gb
Carte réseaux 1	NAT	NAT	NAT
Carte réseau 2	Réseau privé hôte 192.168.56.102	Réseau privé hôte 192.168.56.103	Réseau privé hôte 192.168.56.104

Principales étapes d'installation

L'outil sélectionné pour la mise en place du cluster Hadoop et son écosystème est :

❖ **Sur chaque machine virtuelle, effectuer des modifications ou faire des ajouts à certains fichiers:**

- Ajouter au contenu du fichier: **/etc/resolv.conf** sur chacune des machines:
  search neuronia.com

- Modifier le fichier: **/etc/sysconfig/network**
  Sur le Master: HOSTNAME=master.neuronia.com
  Sur le Datanode1: HOSTNAME=datanode1.neuronia.com
  Sur le Datanode2: HOSTNAME=datanode2.neuronia.com

- Modifier le fichier: **/etc/selinux/config** sur chacune des machines:
  SELINUX=disabled

- Modifier le contenu du fichier: **/etc/hosts** sur chacune des machines:
  127.0.0.1 localhost
  192.168.56.102 master.neuronia.com
  192.168.56.103 datanode1.neuronia.com
  192.168.56.104 datanode2.neuronia.com

- **Utilisation de la procédure décrite sur le blog suivant:**

  https://weidongzhou.wordpress.com/2015/09/17/install-cloudera-hadoop-cluster-using-cloudera-manager/

Avec quelques exceptions et ajout à la procédure:

Étapes préparatoires à l'installation

Étapes 3 & 4:

Ajouter au contenu du fichier /etc/sysctl.conf sur le master:
vm.swappiness=10
net.ipv6.conf.default.disable_ipv6=1
net.ipv6.conf.all.disable_ipv6=1

Étape 5:

yum -y install perl openssl-clients
ssh-keygen (enter, enter, enter)
cd ~/.ssh
cp cp id_rsa.pub authorized_keys

Étapes de l'installation

Étape 1:

Le lien est wget http://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin

Étape 5:

Compléter la case avec les donnés suivantes, et pressez sur le bouton New Search
192.168.56.102
192.168.56.103
192.168.56.104

Étape 13:

Choisir All services

- **Après avoir complété la procédure, j'ai consulter les différents messages reliés à des problèmes de configuration, et j'ai dû:**

  - Installer manuellement NTP sur toutes les machines virtuelles:
    yum install ntp
    service ntpd restart

  - Installer java jdk 1.7.0 sur toutes les machines virtuelles:
    apt-get install openjdk-7-jdk
    java -version

  - J'ai vérifier et enlever (supress) tous les messages d'avertissement relié à la mémoire, etc…

Test du cluster

Exécution d'une application Yarn type sur le cluster

On exécuteras une application Yarn-Mapreduce WordCount version 1.0 qui permet de compte le nombre de fois que chaque mot apparait dans des fichiers et liste dans le fichier de résultat chaque mot avec son nombre d'occurrences dans l'ensemble des fichiers contenues dans un répertoire.

Le répertoire d'entrée qui contient les fichiers à analyser est le : /user/root/testtp/input

Le répertoire de sortie qui contiendra un fichier indiquant le succès et un fichier contenant les résultats est le : /user/root/testtp/output.

La commande pour exécuter le Wordcount est la suivante:
[root@master WordCount1]# hadoop jar wordcount.jar org.myorg.WordCount /user/root/testtp/input /user/root/testtp/output

Paramètres d'exécution obtenus à partir de l'outil de gestion

```
16/02/06 22:46:45 INFO client.RMProxy: Connecting to ResourceManager at master.neuronia.com/192.168.56.102:8032
16/02/06 22:46:46 INFO input.FileInputFormat: Total input paths to process : 2
16/02/06 22:46:46 INFO mapreduce.JobSubmitter: number of splits:2
16/02/06 22:46:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1454646394534_0006
16/02/06 22:46:46 INFO impl.YarnClientImpl: Submitted application application_1454646394534_0006
16/02/06 22:46:46 INFO mapreduce.Job: The url to track the job: http://master.neuronia.com:8088/proxy/application_1454646394534_0006/
16/02/06 22:46:46 INFO mapreduce.Job: Running job: job_1454646394534_0006
16/02/06 22:46:51 INFO mapreduce.Job: Job job_1454646394534_0006 running in uber mode : false
16/02/06 22:46:51 INFO mapreduce.Job:  map 0% reduce 0%
16/02/06 22:46:56 INFO mapreduce.Job:  map 50% reduce 0%
16/02/06 22:47:01 INFO mapreduce.Job:  map 100% reduce 0%
16/02/06 22:47:08 INFO mapreduce.Job:  map 100% reduce 100%
16/02/06 22:47:08 INFO mapreduce.Job: Job job_1454646394534_0006 completed successfully
16/02/06 22:47:09 INFO mapreduce.Job: Counters: 49
    File System Counters
        FILE: Number of bytes read=14019
        FILE: Number of bytes written=375450
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=44607
        HDFS: Number of bytes written=6997
        HDFS: Number of read operations=9
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=6436
        Total time spent by all reduces in occupied slots (ms)=3550
        Total time spent by all map tasks (ms)=6436
        Total time spent by all reduce tasks (ms)=3550
        Total vcore-seconds taken by all map tasks=6436
        Total vcore-seconds taken by all reduce tasks=3550
        Total megabyte-seconds taken by all map tasks=6590464
        Total megabyte-seconds taken by all reduce tasks=3635200
    Map-Reduce Framework
        Map input records=1053
        Map output records=12638
        Map output bytes=104305
        Map output materialized bytes=19258
        Input split bytes=250
        Combine input records=0
        Combine output records=0
        Reduce input groups=800
        Reduce shuffle bytes=19258
        Reduce input records=12638
        Reduce output records=800
        Spilled Records=25276
```

```
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=127
        CPU time spent (ms)=1840
        Physical memory (bytes) snapshot=890281984
        Virtual memory (bytes) snapshot=4819791872
        Total committed heap usage (bytes)=647634944
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=44357
    File Output Format Counters
        Bytes Written=6997
[root@master WordCount1]#
```

Conclusion

L'installation d'un cluster Hadoop et son écosystème est grandement facilité avec l'outil Cloudera Manager. Également, Cloudera manager permet une gestion facilitée et visuelle des différents composants de l'écosystème Hadoop.