****

**Projet de session**

**420-BD3-BB**

**Samir LAKEHAL**

**(1895291)**

**TP1: mise en place d’un cluster Cloudera/Hortonworks**

**Mai 2018**

**Introduction :**

L’objectif est de créer dans un premier temps un cluster Hadoop sur des machines virtuelles créées à partir d’une distribution CentOS, et en utilisant Cloudera Manager au final, et ensuite tester le cluster en exécutant un job MapReduce.

**Caractéristiques du cluster :**

J’ai opté pour 3 datanodes comme minimum recommandé par Cloudera Manager :

Masternode

CPU : 4 Cores i7, 3.4 Ghz

RAM : 6 Go

HDD : 32 Go

Datanode 3

CPU : 4 Cores i7, 3.4 Ghz

RAM : 2 Go

HDD : 32 Go

Datanode 2

CPU : 4 Cores i7, 3.4 Ghz

RAM : 2 Go

HDD : 32 Go

Datanode 1

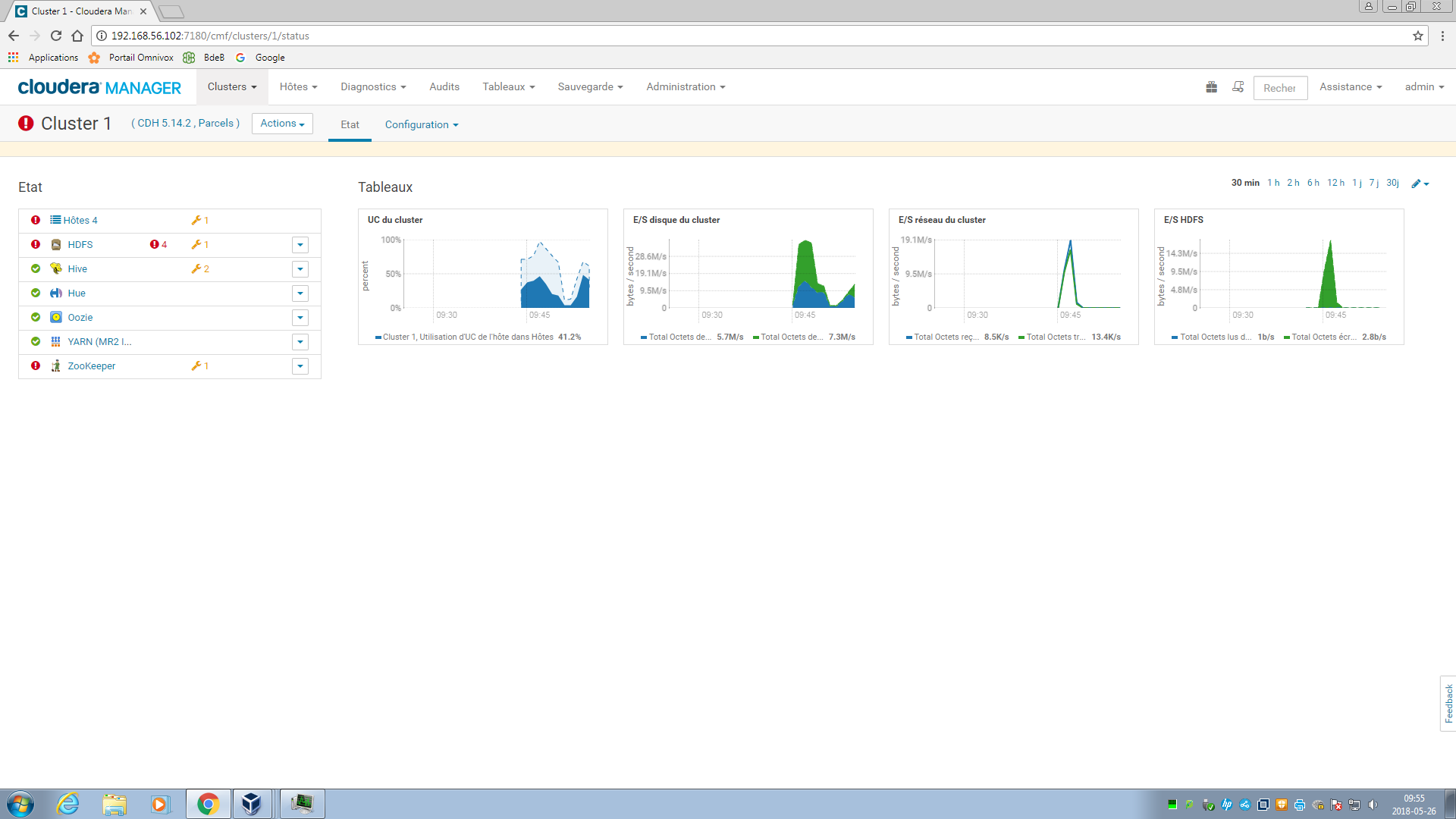
CPU : 4 Cores i7, 3.4 Ghz

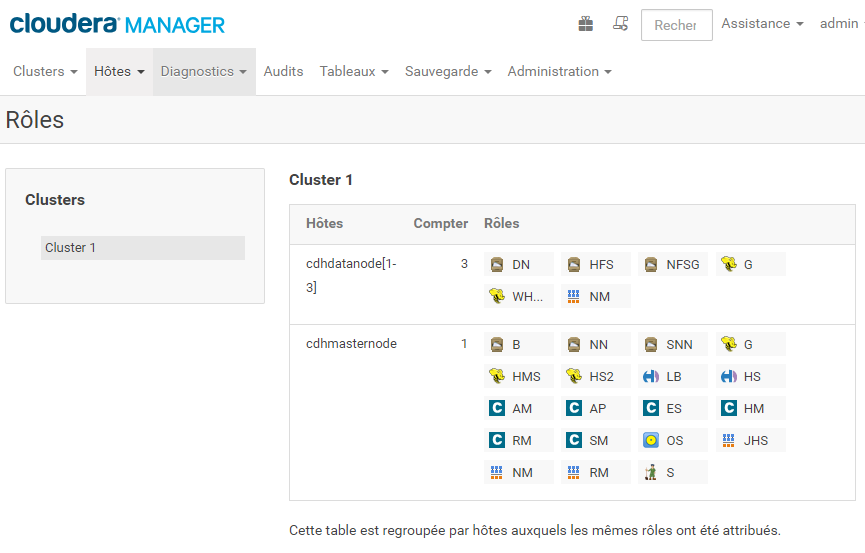
RAM : 2 Go

HDD : 32 Go

**Étapes principales d'installation :**

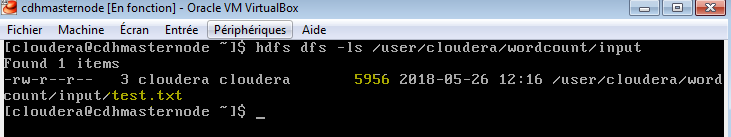
* Création d’une machine virtuelle CentOS 4.6
* Configuration taille mémoire et espace disque
* Configuration des cartes réseaux
* Désactivation du système SELINUX
* Désactivation du firewall
* Installation de Perl et Openssh
* Exécution de SSH pour générer les clés
* Clonage de la machine virtuelle (masternode) pour créer 3 datanodes
* Paramétrages des clones en changeant les adresses IP et le hostname pour chaque datanode
* Téléchargement et installation de Cloudera Manager sur le masternode
* Lancement de l’interface graphique de Cloudera Manager (sur le port 7180) et suivre les instructions pour l’installation du cluster (voir ci-dessous résultat installation)



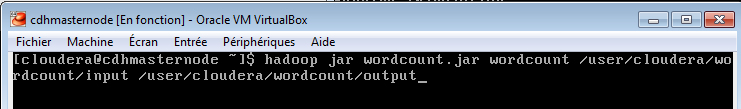


**Test de l’application :**

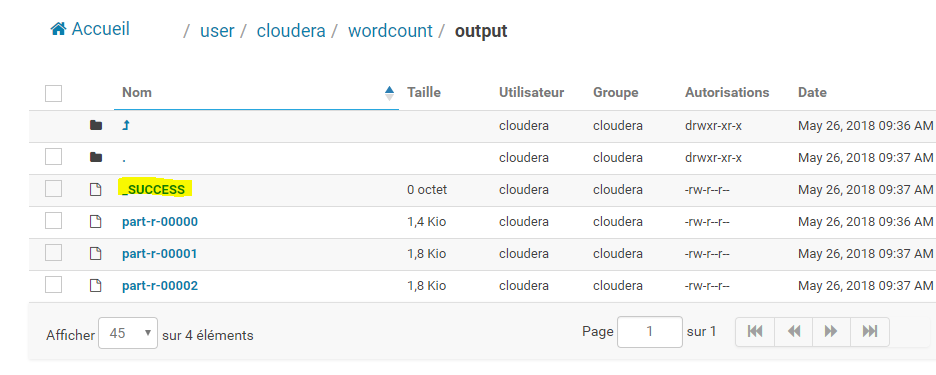
J’ai créé un fichier test.txt dans le folder /user/cloudera/wordcount/input

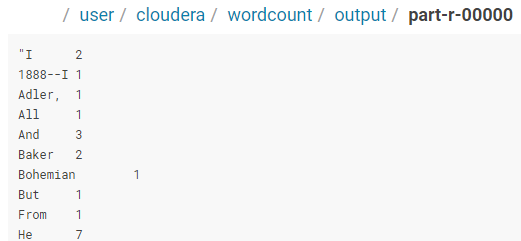


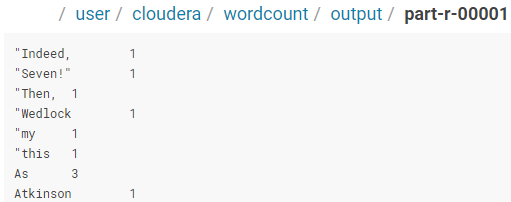
Ensuite j’ai lancé le job mapreduce (jar) :

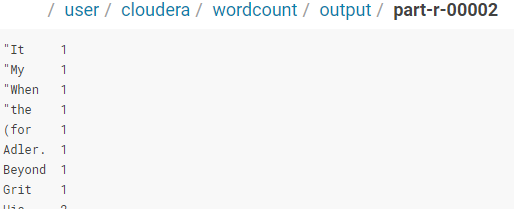


Résultat :









**Conclusion :**

J’ai pu apprendre beaucoup de choses avec cet exercice, et surtout surmonter beaucoup de problèmes liés aux ressources matérielles, notamment la plus exigeante qui est la taille mémoire.