**TP1**

**CLUSTER CLOUDERA ET AWS**

**420-BD3-BB gr.000243**

**Octobre 2018**

**Alvaro Cuervo**

**1895284**

# Introduction

Pendant le cours BD3 on a fait l’installation de Hadoop; une seule machine et ainsi cluster avec trois machines. Malgré que a la fin chaque cluster a fonctionné correctement c’est ne pas dans tous le cas évident et rapide l’installation, le processus a requis la configuration de plusieurs fichiers xml, essaie et erreur de quelques paramètres, permissions, etc.

HortonWorks et Cloudera sont des forts compétiteurs dans le monde Big Data Hadoop, chaque une a des outils de installation et gestion du cluster. Dans le présente travail on va explorer la solution propose par Cloudera, compagnie avec un grand nombre de clients autour du monde dans tous sortes de industries.

# Caractéristiques du cluster

Le cluster Hadoop (distribution Cloudera) est hébergé sur le cloud (Amazon Web Services AWS – Region Canada (Central)), est composé d’un NameNode est trois DataNodes avec les caractéristiques suivantes:

## NameNode (1 instance) [*Amazon EC2 type t3.xlarge*]

### 4 processeurs

### 16 GB RAM

### 30 GB disque dur

## DataNodes (3 instances) [*Amazon EC2 type t3.xlarge*]

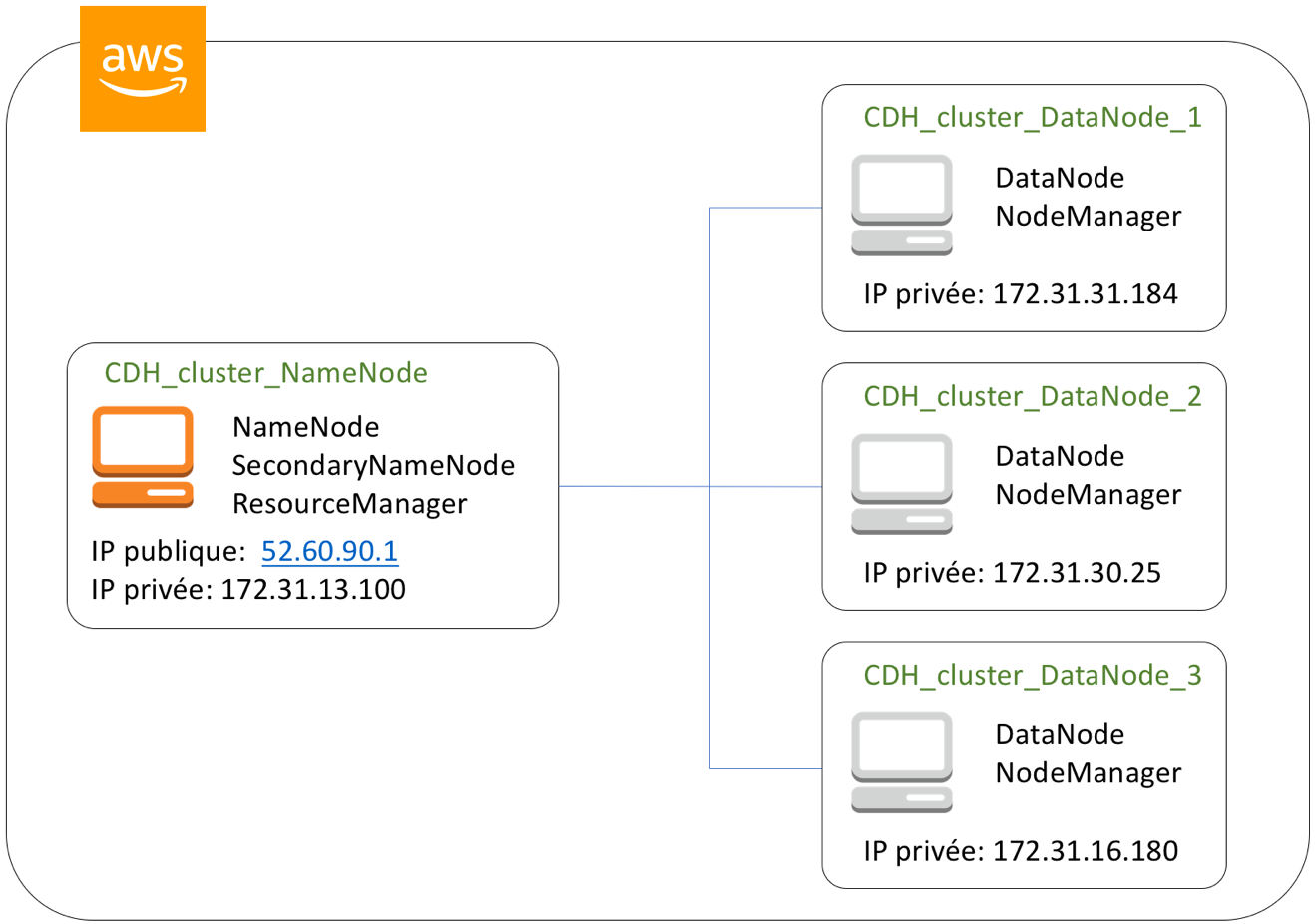
### 2 processeurs (chauqe une)

### 8 GB RAM (chaque une)

### 20 GB disque dur (chaque une)

Chaque élément du cluster utilise Ubuntu Server 16.04.

Le diagramme suivant montre en détail les services de chaque composant, ainsi les adresses IP utilisées :



# Étapes principales d’installation

## Prérequis

### Compte AWS

### Création de 4 instances EC2 (Amazon Elastic Compute Cloud) avec les caractéristiques de chaque machine mentionnes dans le point précédent.

#### Chaque machine est liée pendant cette phase avec la clé privée et public (AWS)

#### Configuration de sécurité (ports ouverts afin de faciliter l’installation et exécution –environnement test-)

### Création et association d’un adresse IP fixe à la machine NameNode en utilisant l’option Elastic IPs de AWS

## Étapes principales

### Accéder via terminal au NameNode

### Passer au compte root

#### sudo su -

### Télécharger l’installateur de Cloudera Manager, avec l’instruction :

### wget https://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin

### Exécuter l’installateur:

#### chmod +x cloudera-manager-installer.bin

#### ./cloudera-manager-installer.bin

#### Screen%20Shot%202018-09-09%20at%207.35.35%20PM.png

### Après d’avoir acceptées les licences de Cloudera, java, etc, la installation de quelques composants est exécuté et à la fin le message suivant est affichée (en indiquant que l’installation continue dans le navigateur web):

### 

### 

### En utilisant admin comme utilisateur et mot de passe on doit sélectionner l’édition de Cloudera que on veut installer (express, enterprise trial -60 jours- ou entreprise), dans cette opportunité J’ai décidé selectionner entreprise trial.

### 

### Après on sélectionne le repository, la version cd CDH (Cloduera Distribution Hadoop) et Cloudera manager commence l’installation du cluster…

## 

## 

### A la fin de cette partie l’installateur affiche un résumé avec validations en indiquant que le processus a été effectué correctement.

## 

### Maintenant on arrive à la sélection des services que on veut utiliser dans notre cluster, ainsi comme la configuration de bases de donnes que vont supporter quelques services :

### 

### 

### Dans l’étape suivant on définit des paramètres de notre cluster telles comme le HDFS Block Size, répertoires et ports.

### 

### Finalement Cloduera manager applique les paramètres pour compléter l’installation et démarré les services sélectionnés auparavant :

### 

### 

### Maintenant on a un cluster opérative avec les services, on peut meme acceder a Hue (Hadoop User Experience).

### 

## 

### En plus et en utilisant Hue on a créé un compte pour l’utilisateur ubuntu (notre compte linux) avec superuser statu, ce cette compte que on va utiliser pour tester notre application / cluster.

## 

# Comment tester l'application

# Afin de tester notre cluster on va utiliser le wordcount exemple, le input de notre job sera le livre « The hacker Crackdown », obtenu dans le site du « Project Gutenberg » dans l’address suivant :

# <http://www.gutenberg.org/cache/epub/101/pg101.txt>

# voici les instructions pour y arriver a son exécution dans le cluster proposée :

# cd /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars

# hadoop jar hadoop-examples-2.6.0-mr1-cdh5.15.1.jar wordcount /user/ubuntu/input /user/ubuntu/output

# 

# Plus de détails sont présentes dans le rapport du vidéo, en incluant le résultat et les paramètres d’exécution obtenues.

# Conclusion

La utilisation de distributions commerciaux de Hadoop (Cloudera, Hortonworks, etc) facilitent et guident le processus d’installation et gestion du cluster. Dans l’exercice présentée dans cette travail Cloudera manager a permis une processus d’installation plus conviviale et rapide, ainsi permet un configuration de paramètres à travers de formulaires qui cache la modification directe des fichiers xml.

Cloudera Manager affiche aussi indicateurs de performance de notre cluster d’une façon graphique, plus facile à comprendre pour l’utilisateur finale (administrateur) et plus approprie afin de prendre des décisions sur notre environnement Big data.