Université Mohammed V de Rabat
Ecole Nationale Supérieure d'Arts et Métiers
Cycle Ingénieur Filière INDIA (3ème année) A.U. : 2023/2024 Pr. A. El Qadi

TP n6 en Mise en œuvre des Framework d'IA et Big data

Exercice 1. Configuration d'un seul cluster Hadoop à l'aide de docker-compose

1. Cloner le référentiel docker-Hadoop **https://github.com/Marcel-Jan/docker-hadoop-spark.git**
   L'exemple de référentiel ci-dessus contient un ensemble Hadoop docker-compose.yml et est prêt à être déployé sur des conteneurs Docker. C'est un environnement multi-conteneurs Docker avec Hadoop, Spark et Hive
   C:\>**git clone https://github.com/Marcel-Jan/docker-hadoop-spark.git**
2. Demarrer Docker desktop
3. Accéder au dossier cloné, puis exécuter la commande suivante pour démarrer le conteneur à l'aide de docker-compose :

   c:**docker-hadoop-spark**>**docker-compose up -d**

   L'indicateur -d définira le conteneur pour qu'il s'exécute dans un modèle détachable, c'est-à-dire en arrière-plan.
   Une fois que tout est terminé, vérifier les conteneurs Hadoop en cours d'exécution à l'aide de la commande suivante ;

   c:**docker-hadoop-spark**>**docker container ls**

   Dans la liste des conteneurs en cours d'exécution, obtenir un port pour le conteneur que vous souhaitez vérifier. Ces numéros de port sont déjà définis dans le référentiel précédemment cloné dans le fichier docker-compose.yml.
   Par exemple :
   Pour datanode, utiliser PORT 9864.
   Pour nodemanager, utiliser PORT 8042.
   Pour namenode, utiliser PORT 9870.
   Pour historyserver, utiliser PORT 8188.
   Pour resourcemanager, utiliser PORT 8088.
   Et l'état actuel peut également être vérifié à l'aide de la page Web http://localhost:9870/explorer.html#/
4. Copier les fichiers *.txt du dossier %HADOOP_HOME% dans le conteneur namenode.
   docker cp %HADOOP_HOME%*.txt namenode:/tmp/
5. Accéder au conteneur namenode et exécuter-le de manière interactive à l'aide de la commande suivante ;

   c:\>**docker-hadoop-spark>docker exec -it namenode bash**

6. Dans le terminal bash résultant, créer un dossier d'entrée pour héberger des fichiers
   Créer le dossier d'entrée input à l'intérieur du conteneur namenode /user/root
   root@c25780d97f98:/# **cd tmp**
7. Copier les fichiers .txt de /tmp vers le dossier d'entrée
8. Visualiser le fichier sur le site : **http://localhost:9870/explorer.html#/user/root/input**
9. Afficher un rapport détaillé sur le dossier HDFS /user/root/input
   - Quel est le nombre de fichier dans hdfs?
   - Quel est le nombre de blocs?
   - Quel est le nombre de blocs corrompus?
   - Quel est le facteur de réplication (Default replication factor)?
   - Quel est le nombre de data-nodes contenant les blocs des fichiers du dossier HDFS?
   - Quel est le nombbre de racks ?
10. Afficher les dernières lignes du fichier notice.txt
11. Copier le fichier « C:\hadoop-2.9.2\share\hadoop\mapreduce**hadoop-mapreduce-examples-2.9.2.jar** » dans le dossier du référentiel cloné **namenode:/tmp/**
12. Exécuter MapReduce, compter le nombre total de mots des fichiers .txt disponibles dans le répertoire d'entrée **user/root/input**, et enregistrer la sortie dans le fichier output / part-r00000,
13. Vérifier les resulatats du mapreduce
14. Arrêter le conteneur.

Exercice 2. Configuration d'un seul cluster Hadoop à l'aide VM Cloudera

Oracle VM VirtualBox est un logiciel de virtualisation disponible en tant qu'hôte sur plusieurs systèmes d'exploitation, notamment Windows, Linux 32 et 64 bits et Mac OS X. Il supporte en tant qu'invité entre autres plusieurs systèmes Windows, Linux et MacOS X. Pour certains systèmes, tels que Windows, il est bien entendu nécessaire de posséder une licence pour installer une machine virtuelle invitée fonctionnant sous ce système, comme pour une machine réell

1. Telecharger et installer la macihne virtuelle : VirtualBox-7.0.12-159484-Win
https://download.virtualbox.org/virtualbox/7.0.12/

2. Telecharger cloudera quickstart vm

La machine virtuelle Cloudera QuickStart utilise une installation basée sur des packages qui permet de travailler avec ou sans Cloudera Manager. Il dispose d'un échantillon de la plate-forme de Cloudera pour le «Big Data ».

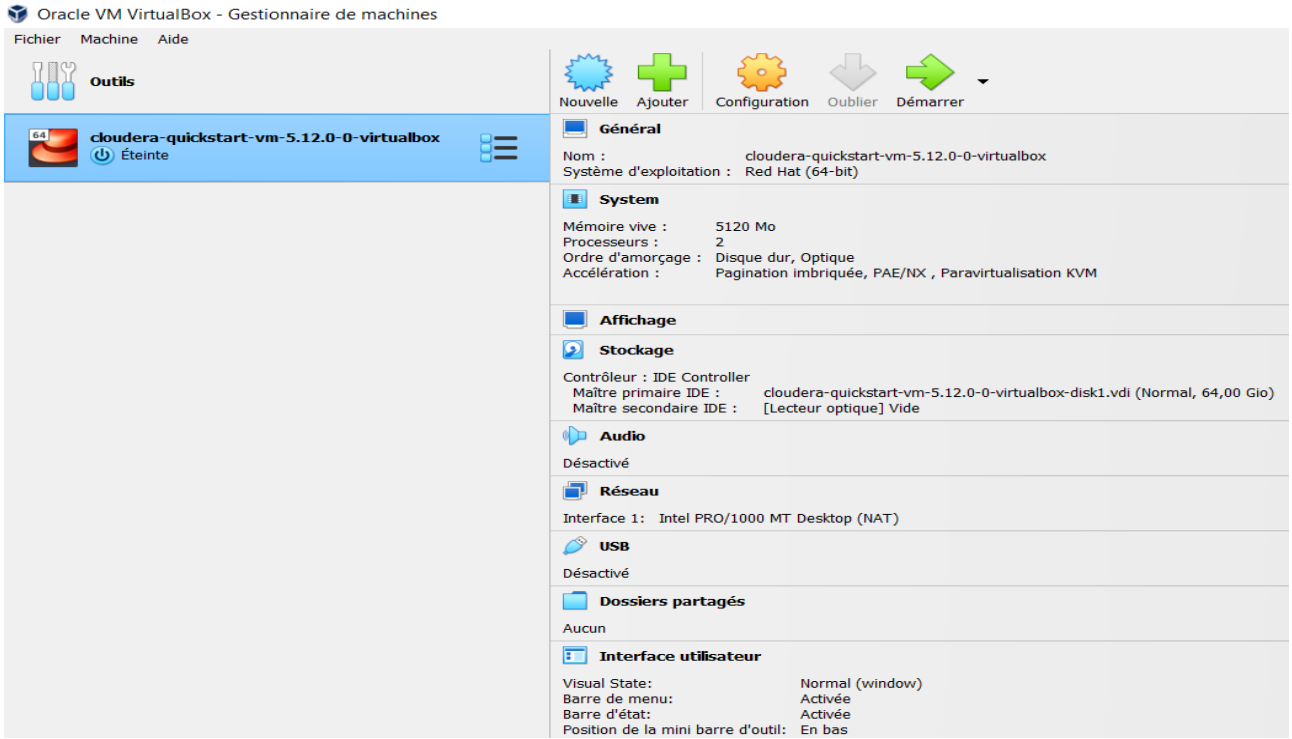
Installation rapide de la machine virtuelle Cloudera - Conditions préalables

1. Une machine virtuelle telle qu'Oracle Virtual Box ou VMWare
2. RAM de 12+ Go. Soit 4+ Go pour le système d'exploitation et 8+ Go pour Cloudera
3. Disque dur de 80 Go

Les machines virtuelles Cloudera QuickStart sont disponibles sous forme d'archives Zip aux formats VirtualBox, VMware et KVM.

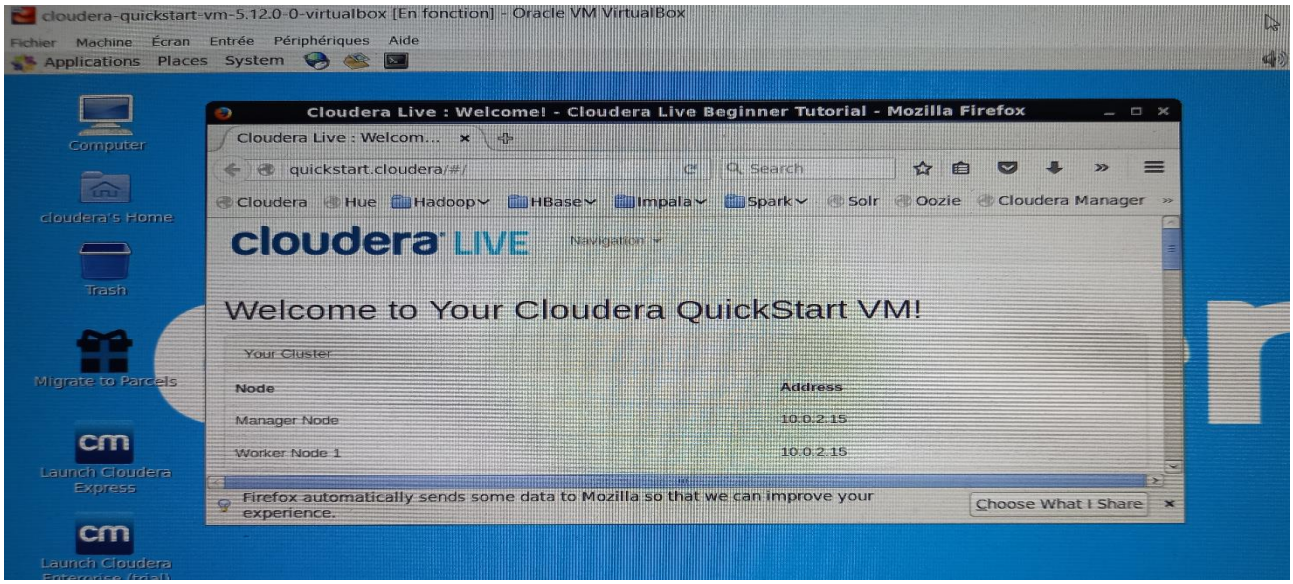
Lien de telechargement : https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.12.0-0-virtualbox.zip

- Une fois le fichier téléchargé, aller dans le dossier de téléchargement et décompresser ces fichiers. Il peut ensuite être utilisé pour configurer un cluster Cloudera à nœud unique.
- Pour configurer la machine virtuelle Cloudera QuickStart dans Oracle VirtualBox Manager, cliquer sur « Fichier », puis sélectionner « Importer ».
- Choisir l'image de machine virtuelle de démarrage rapide. Cliquer sur 'Ouvrir' puis sur 'Suivant', puis cliquer sur « Importer ». Cela commencera à importer **le fichier. vmdk** de l'image de disque virtuel dans la machine virtuelle.
- L'étape suivante consiste à configurer une machine virtuelle Cloudera QuickStart pour la pratique. Une fois l'importation terminée, pour donner plus de RAM et de cœurs de processeur, cliquer sur « Paramètres », puis sur « Système », et augmenter la RAM à 8 Go. Cliquer sur le processeur et attribuer 2 cœurs de processeur. Ensuite, sélectionner « Réseau ». Les paramètres de l'adaptateur 1 doivent être NAT par défaut. Cliquer ensuite sur « OK ».



- L'étape suivante consistera à démarrer la machine en cliquant sur le symbole « Démarrer » en haut.

- Une fois que la machine s'allume, elle ressemblera à ceci :

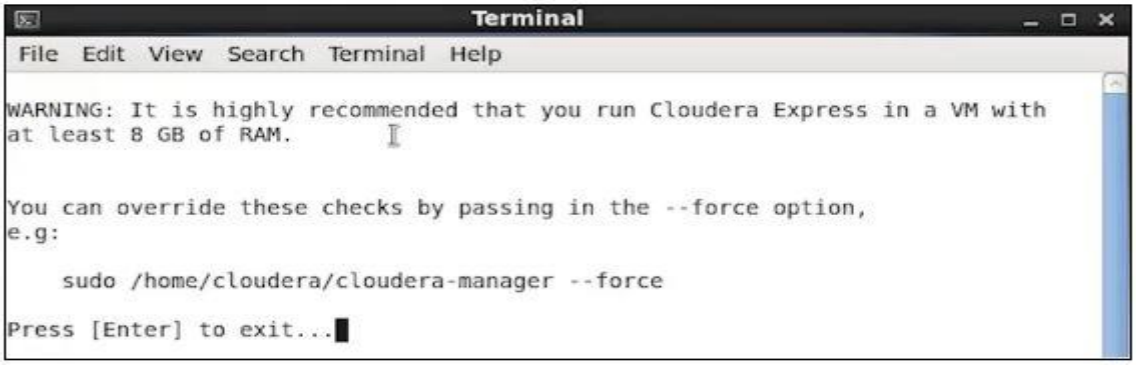


- Ensuite, nous devons suivre quelques étapes pour obtenir l'accès à la console d'administration. Cliquer sur le terminal présent en haut de l'écran du bureau et taper ce qui suit :

```
hostname # Ceci montre le nom d'hôte qui sera quickstart.cloudera
ls / #  affiche ce qui existe sur l'emplacement HDFS par défaut
su # Indique la commande à taper pour utiliser cloudera express free
su  root- #Login comme racine
service cloudera-scm-server status # Le mot de passe pour root est cloudera
```

- Ensuite, Lancer Cloudera Express

- L'écran apparaîtra avec la commande suivante :



Copier la commande et l'exécuter sur un terminal distinct.

- Maintenant que notre déploiement a été configuré, les configurations client ont également été déployées. De plus, il a redémarré le service de gestion Cloudera, qui donne accès à la console d'administration Cloudera QuickStart à l'aide d'un nom d'utilisateur et d'un mot de passe.
  root@quickstart cloudera]# su root
  Password : cloudera

3. Utiliser l'outil **jps** pour lister les services en cours d'exécution :
   [root@quickstart cloudera]# jps

   Remarque : pour arrêter les processus Hadoop, il faut exécuter les commandes stop-dfs.sh et stop-yarn.sh (et éventuellement mr-jobhistory-daemon.sh stop historyserver). La commande jps doit alors seulement afficher comme liste le processus Jps

4. Lister le contenu de la racine
5. Transférer des fichiers entre le système hôte et la VM.
   a. Lister le contenu de la racine HDFS
   b. Créer un dossier HDFS input dans /user
   c. Avec l'editeur vi, créer le fichier texte « demo.txt »
   d. Copier le fichier demo.txt vers le dossier HDFS /user/input
   e. Afficher le contenu du dossier HDFS /user/input
   f. Afficher les dernières lignes du fichier demo.txt dans HDFS
   g. Afficher la liste des fichiers dans le navigator hdfs : Quickstart.cloudera :50070/explorer.html
   h. Récupérer la taille d'un block HDFS: hdfs getconf -confKey dfs.blocksize
   i. Récupérer le facteur de réplication: hdfs getconf -confKey dfs.replicatio
   j. Avec l'editeur vi, créer le fichier texte « myfile.txt »
   k. Copier le fichier local file.txt vers le dossier HDFS /user/input
   l. Utiliser la commande hdfs fsck pour afficher un rapport détaillé sur le fichier myfile.txt dans HDFS.
      - Quel est le nombre de blocs?
      - Quelle est la taille moyenne de chacun?
      - Quel est le facteur de réplication (Default replication factor)?
      - Quel est le nombre de data-nodes contenant les blocs des fichiers du dossier HDFS?
      - Quel est le nombre de blocs corrompus?
   m. Modifier le facteur de réplication, valeur 2, du fichier **myfile.xt** dans HDFS
6. Essayer d'exécuter un programme qui fait seulement un traitement parallèle sans manipulation de fichier:
   hadoop jar /…/hadoop/mapreduce/hadoop-mapreduce-examples-2.X.Y.jar pi 2 5