TP 4 : Installation de Hadoop simple nœud

1) Télécharger et installer Oracle VM Virtual Box à partir de ce lien https://www.oracle.com/technetwork/server-storage/virtualbox/downloads/index.html.
2) Lancer Oracle VM Virtual Box.
3) Cliquer sur « Nouvelle »
4) Donner un nom à la machine.
5) Choisir le système d'exploitation Ubuntu 64 bits

6) Pour la mémoire RAM, il faut choisir au moins une taille égale à 2048 Mo pour avoir une manipulation acceptable en termes de rapidité.
7) La taille du disque dur recommandé est de 30 GO, choisir « créer un disque dur virtuel maintenant » et cliquer sur « créer ».
8) Choisir l'option VDI pour le type de fichier de disque dur (on n'a pas besoin des autres types puisqu'on n'a pas d'autres logiciels de virtualisations à utiliser.
9) Pour le stockage sur un disque dur physique, choisir l'option « Dynamiquement alloué » (pour optimiser l'utilisation de votre disque selon les données existantes).
10) Ne rien changer pour l'emplacement et choisir 30 GO comme taille limite du disque dur.
11) Sélectionner la machine créer et cliquer sur configuration, passer à l'onglet « affichage » et augmenter la mémoire vidéo à 60 Mo.
12) Démarrer la nouvelle machine.
13) Choisir le disque de démarrage en choisissant l'image iso d'installation de Ubuntu stocker sur votre ordinateur et démarrer.
14) En cas d'échec de démarrage de la nouvelle machine virtuelle, il faut vérifier si le bios de votre machine permet la virtualisation.
15) Choisir la langue.
16) Cliquer sur installer Ubuntu.
17) Cocher les deux options (Téléchargement des mises à jour et l'installation des logiciels tiers) pendant l'installation et cliquer sur continuer.
18) Choisir « Effacer le disque et installer Ubuntu » et cliquer sur continuer.
19) Laisser à Paris et continuer.
20) Pour la disposition du clavier, choisir « Français » et cliquer sur continuer.
21) Pour les options qui êtes-vous ?
    - Choisir nom, (machine1 par exemple)
    - Choisir nom de l'ordinateur, (machine1 aussi par exemple)
    - Choisir nom de l'utilisateur, (u1 également par exemple)
    - Choisir mot de passe : « hadoop » (pour ne pas l'oublier),
    - Cocher « ouvrir la session automatiquement et cliquer sur continuer.
22) Après la fin de l'installation, lancez le terminal.

23) Mettre à jour les fichiers de dépôt avec sudo apt-get update

- apt-get : est un outil logiciel à utiliser en ligne de commande, il permet l'installation et la désinstallation de paquets en provenance d'un dépôt APT, ce dernier est une source (ou un serveur) de logiciels. L'usage de ces outils centralise la gestion des logiciels et la simplifie. Ils permettent également aux distributeurs (ceux qui mettent en place les dépôts) de vous fournir les mises à jour par une voie centralisée.
- apt-get update : l'option « update » met à jour la liste des fichiers disponibles dans les dépôts APT présents dans le fichier de configuration /etc/apt/sources.list. L'exécuter régulièrement est une bonne pratique afin de maintenir à jour votre liste de paquets disponibles.
- apt-get install : permet d'installer un paquet.

24) Installer java avec sudo apt-get install default-jdk.

apt-get install default-jdk : installer le paquet java par défaut (le plus récent).

25) Vérifier l'installation de java avec : java -version

26) Installer ssh avec sudo apt-get install ssh

OpenSSH est une version libre de la suite de protocole de SSH, des outils de connectivité de réseau sur lesquels un nombre, croissant, de personne sur Internet viennent s'appuyer.

Beaucoup d'utilisateurs de Telnet, Rlogin, FTP, ou d'autres programmes de même acabit ne se rendent pas compte que leurs données, et notamment les mots de passe, sont transmis à travers les réseaux en clair ce qui constitue une faille évidente dans la sécurité de leur réseau.

OpenSSH chiffre tout le trafic (mots de passe y compris), via une combinaison astucieuse de chiffrement symétrique et asymétrique. OpenSSH fournit également d'autres méthodes d'authentification alternatives au traditionnel mot de passe.

Comme son nom l'indique, OpenSSH est développé dans le cadre du projet OpenBSD

SSH remplace de manière sécurisée :

- Telnet: Vous pouvez exécuter des commandes depuis un Réseau Local ou Internet via SSH,
- FTP: Si vous ne souhaitez qu'ajouter ou modifier des fichiers sur un serveur, SSH est bien plus adapté que FTP,
- Et d'autres, vous pouvez donc sécuriser n'importe quel protocole grâce à SSH.

SSH permet de faire, en usage de base :

-Accès à distance sur la console en ligne commande (shell), ce qui permet, entre autres, d'effectuer la totalité des opérations courantes et/ou d'administration sur la machine distante,

Déporter l'affichage graphique de la machine distante,

- Transferts de fichiers en ligne de commande,
- Montage ponctuel de répertoires distants,
- Montage automatique de répertoires distants.
-

27) Installer rsync avec sudo apt-get install rsync

- rsync (pour remote synchronization ou synchronisation à distance), est un logiciel de synchronisation de fichiers. Il est fréquemment utilisé pour mettre en place des systèmes de sauvegarde distante,
- rsync travaille de manière unidirectionnelle c'est-à-dire qu'il synchronise, copie ou actualise les données d'une source (locale ou distante) vers une destination (locale ou distante) en ne transférant que les octets des fichiers qui ont été modifiés.

28) Générer une clé de cryptage avec ssh-keygen -t rsa -P ""
    - Le chiffrement RSA (nommé par les initiales de ses trois inventeurs) est un algorithme de cryptographie asymétrique, très utilisé dans le commerce électronique, et plus généralement pour échanger des données confidentielles sur Internet.
    - -P "": c'est à dire sans mot de passe.
29) Copier cette clé dans « authorized_keys » avec cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
30) Vérifier l'installation avec ssh localhost.
31) Télécharger une version de Hadoop.
32) Accéder au répertoire Téléchargement et extraire le dossier zippé avec sudo tar -zxvf hadoop-…...tar.gz ou tout simplement avec clique bouton droit et extraire ici.
33) Créer le dossier hadoop dans /usr/local avec sudo mkdir /usr/local/hadoop
34) Déplacez le dossier source de hadoop avec sudo mv hadoop-…….. /usr/local/hadoop.
35) Récupérer le non et le chemin de java jdk, /usr/lib/jvm/java-8-….
36) Ajouter les alias ci-dessous au fichier .bashrc

    Remarque : Les alias sont des substitutions abrégées de commandes répétitives et/ou longues à taper dans la console. Il est possible de définir vos alias dans un fichier caché, appelé .bashrc, qui se trouve dans votre Dossier Personnel.

    avec sudo gedit (ou nano) .bashrc (fichier caché se trouvant dans /home/<nom>/.bashrc) et écrire ces lignes en bas du fichier :

    ```
    #HADOOP VARIABLES START
    export JAVA_HOME=/….
    export HADOOP_INSTALL=/….
    export PATH=$PATH:$HADOOP_INSTALL/bin
    export PATH=$PATH:$HADOOP_INSTALL/sbin
    export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
    export HADOOP_COMMON_HOME=$HADOOP_INSTALL
    export HADOOP_HDFS_HOME=$HADOOP_INSTALL
    export YARN_HOME=$HADOOP_INSTALL
    export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
    export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
    #HADOOP VARIABLES END
    ```
37) Pour que les alias puissent entrer en vigueur, il faut redémarrer le terminal ou taper la commande source .bashrc.
38) Modifier le fichier de démarrage de hadoop « hadoop-env.sh » en ajoutant le chemin du JAVA_HOME:

    export JAVA_HOME=/…

avec la commande sudo nano $HADOOP_INSTALL/etc/hadoop/hadoop-env.sh
ou bien sudo nano /usr/local/hadoop/hadoop<version>/etc/hadoop/hadoop-env.sh
ou bien sudo gedit /usr/local/hadoop/hadoop<version>/etc/hadoop/hadoop-env.sh
Remarque : le fichier hadoop-env.sh démarre les démons, ces derniers, en termes de programmation, sont des processus s'exécutant en arrière-plan (background). Hadoop possède cinq à savoir :

- Le Namenode,
- Le Secondary Namenode,
- Le Datanode,
- Le JobTracker,
- Le TaskTraker.

39) Modifier le fichier core-site.xml en ajoutant :

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

avec la commande sudo nano $HADOOP_INSTALL/etc/hadoop/core-site.xml
Remarque : Le fichier core-site.xml informe les démons de hadoop qu'un namenode s'exécute sur le cluster en mentionnant son adresse.

40) Modifier le fichier hdfs-site.xml en ajoutant :

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.name.dir</name>
<value>/home/…/name</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/home/…/data</value>
</property>
</configuration>
```

Remarque : le fichier hdfs-site.xml informe hadoop et son système hdfs du nombre de machine au sein du cluster (nombre de réplication).

41) Faire une copie du fichier mapred-site.xml.template sous le nom mapred-site.xml avec la commande : sudo cp /usr/local/hadoop/hadoop-<version>/etc/hadoop/mapred-site.xml.template /usr/local/hadoop/hadoop-<version>/etc/hadoop/mapred-site.xml
Remarque : le fichier mapred-site.xml informe le package MapReduce qu'il va s'exécuter en tant que application yarn (séparation entre la gestion des ressources et la gestion des travaux).

42) Modifier le fichier mapred-site.xml en ajoutant :

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
```

</configuration>

43) Modifier le fichier yarn-site.xml en ajoutant :

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
</configuration>
```

Remarque : le fichier yarn-site.xml indique au nodemanager qu'il aura un service auxiliaire indiquant au MapReduce comment faire son shuffling.

44) Formater le système hdfs avec : hdfs namenode -format.
45) Vérifier les services actifs (avant de démarrer hadoop) avec jps.
46) Démarrer le système hadoop (hdfs aussi) avec start-all.sh.
47) Vérifier les services actifs après le démarrage.
48) Avec votre navigateur web, vous pouvez accéder à l'interface web NameNode via http://localhost:50070/
49) Créer un répertoire dans hdfs pour mettre les fichiers d'entrées et de résultats avec hdfs dfs -mkdir /user/
50) Créer un répertoire datainput dans votre dossier personnel avec mkdir ~/datainput. Ce dossier va servir comme un bassin ou on va mettre les fichiers à analyser.
51) Créer un fichier texte.txt (à remplir avec plusieurs mots) dans ce dossier datainput avec cat > ~/datainput/texte.txt suivi de la saisie de quelques mots et appuyer sur Ctrl+D à la fin.
52) Transférer les fichiers à traiter avec hdfs dfs -put /home/<nom>/datainput /user/input
53) Lancez un jar exemple (wordcount) avec hadoop jar /usr/local/hadoop/hadoop-…/share/hadoop/mapreduce/hadoop-mapreduce-examples-….jar wordcount /user/input /user/output
54) Récupérer le résultat avec hdfs dfs -cat /user/output/*
55) Stopper hadoop avec stop-all.sh
56) Cloner cette machine pour avoir une machine prête au cas où celle d'origine aurait des problèmes suite à l'installation de nouveaux logiciels. Pour effectuer cette action, passer à l'interface de virtual box, cliquer avec le bouton droit de la souris sur la machine configurer et choisir « cloner ».

Remarques :

- En cas de problème désactiver l'IP6 avec iptables -A INPUT -p tcp -dport 50070 -j ACCEPT
- S'il y a des problèmes de permission avec safemode : hdfs dfsadmin -safemode leave