

Ecosystème Hadoop

Atelier 0 : Mise en place de l'environnement Big Data

Objectifs

Dans cet atelier, vous allez :

- Installer des docker et docker-compose sous Ubuntu
- Installation JDK 1.7, JDK 1.8 et NetBeans sous Ubuntu
- Télécharger une image Docker Cloudera/QuickStart
- Déployer et configurer une instance de Cloudera/QuickStart

Installation Docker sous Ubuntu

Docker est un progiciel de plus en plus populaire qui crée un conteneur pour le développement d'applications.

Le développement dans Docker accélère les applications, car il partage le noyau et d'autres ressources, au lieu de nécessiter des ressources de serveur dédiées .

Il existe deux versions de Docker - Docker CE (Community Edition) et Docker EE (Enterprise Edition). Si vous avez un projet à petite échelle ou que vous êtes en train d'apprendre, vous voudrez utiliser Docker CE.

Dans ces étapes, nous vous montrerons comment installer Docker sur Ubuntu 18.04.

Étape 1 : mise à jour des référentiels logiciels

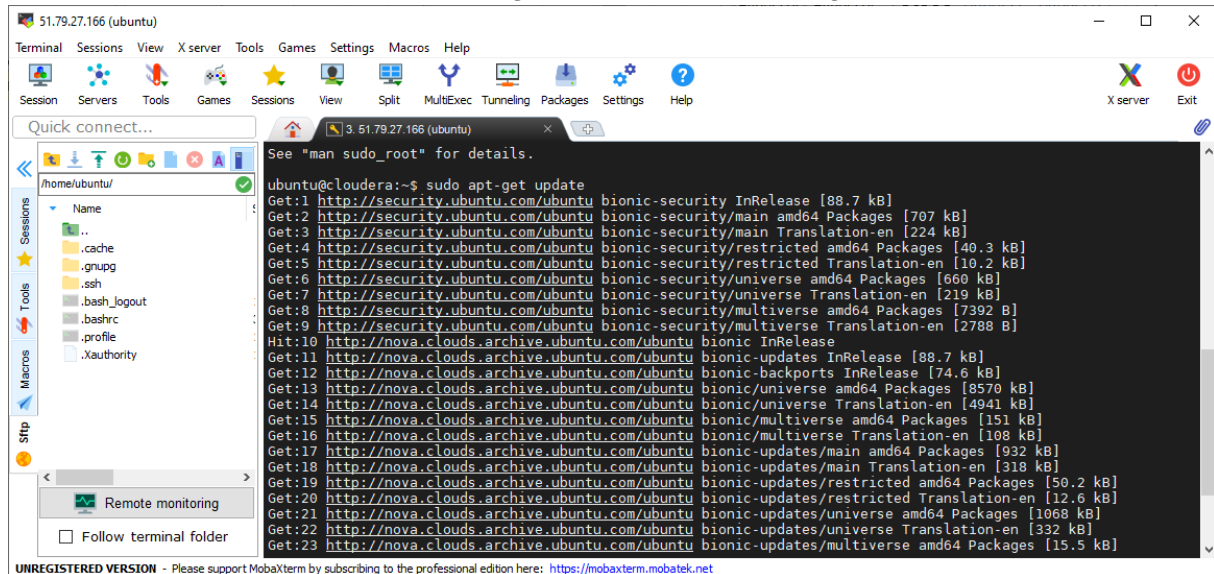
C'est une bonne idée de mettre à jour la base de données locale du logiciel pour vous assurer d'avoir accès aux dernières révisions.

Ouvrez une fenêtre de terminal et saisissez :

```
sudo apt-get update
```

Laissez l'opération se terminer.

Ecosystème Hadoop



```
ubuntu@cloudera:~$ sudo apt-get update
Get:1 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Get:2 http://security.ubuntu.com/ubuntu bionic-security/main amd64 Packages [707 kB]
Get:3 http://security.ubuntu.com/ubuntu bionic-security/main Translation-en [224 kB]
Get:4 http://security.ubuntu.com/ubuntu bionic-security/restricted amd64 Packages [40.3 kB]
Get:5 http://security.ubuntu.com/ubuntu bionic-security/restricted Translation-en [10.2 kB]
Get:6 http://security.ubuntu.com/ubuntu bionic-security/universe amd64 Packages [660 kB]
Get:7 http://security.ubuntu.com/ubuntu bionic-security/universe Translation-en [219 kB]
Get:8 http://security.ubuntu.com/ubuntu bionic-security/multiverse amd64 Packages [7392 B]
Get:9 http://security.ubuntu.com/ubuntu bionic-security/multiverse Translation-en [2788 B]
Hit:10 http://nova.clouds.archive.ubuntu.com/ubuntu bionic InRelease
Get:11 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]
Get:12 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-backports InRelease [74.6 kB]
Get:13 http://nova.clouds.archive.ubuntu.com/ubuntu bionic/universe amd64 Packages [8570 kB]
Get:14 http://nova.clouds.archive.ubuntu.com/ubuntu bionic/universe Translation-en [4941 kB]
Get:15 http://nova.clouds.archive.ubuntu.com/ubuntu bionic/multiverse amd64 Packages [151 kB]
Get:16 http://nova.clouds.archive.ubuntu.com/ubuntu bionic/multiverse Translation-en [108 kB]
Get:17 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/main amd64 Packages [932 kB]
Get:18 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/main Translation-en [318 kB]
Get:19 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/restricted amd64 Packages [50.2 kB]
Get:20 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/restricted Translation-en [12.6 kB]
Get:21 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/universe amd64 Packages [1068 kB]
Get:22 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/universe Translation-en [332 kB]
Get:23 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/multiverse amd64 Packages [15.5 kB]
```

Étape 1: mettre à jour la base de données locale

Mettez à jour la base de données locale avec la commande:

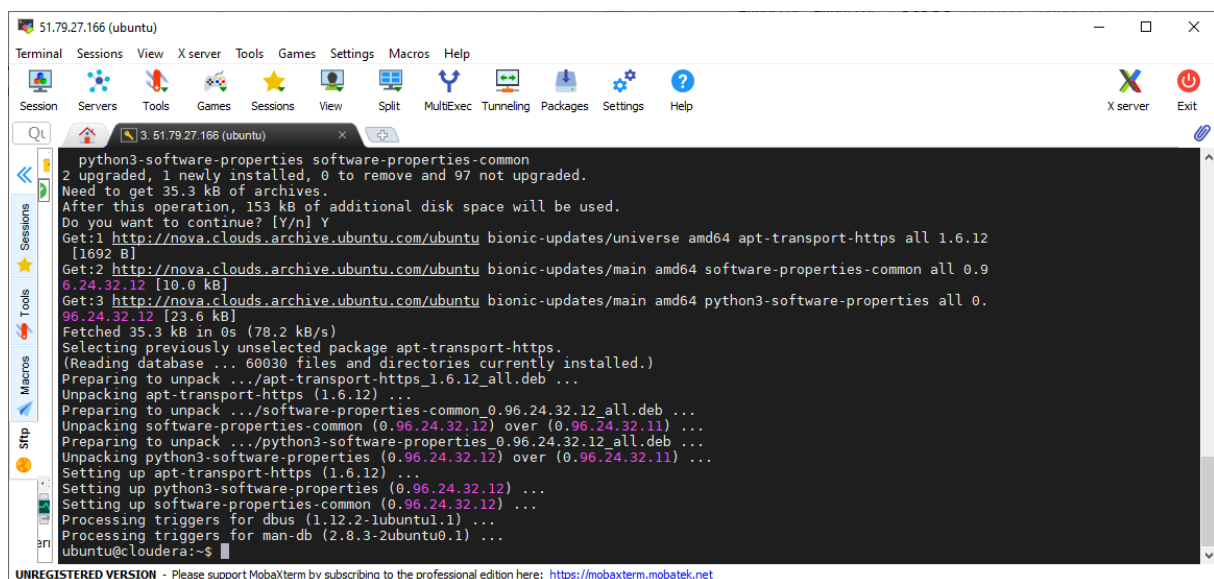
```
sudo apt-get update
```

Étape 2: Télécharger les dépendances

Vous devrez exécuter ces commandes pour permettre à votre système d'exploitation d'accéder aux référentiels Docker via HTTPS.

Dans la fenêtre du terminal, saisissez :

```
sudo apt-get install apt-transport-https ca-certificates curl software-properties-common
```



```
python3-software-properties software-properties-common
2 upgraded, 1 newly installed, 0 to remove and 97 not upgraded.
Need to get 35.3 kB of archives.
After this operation, 153 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/universe amd64 apt-transport-https all 1.6.12 [1692 B]
Get:2 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/main amd64 software-properties-common all 0.96.24.32.12 [10.0 kB]
Get:3 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/main amd64 python3-software-properties all 0.96.24.32.12 [23.6 kB]
Fetched 35.3 kB in 0s (78.2 kB/s)
Selecting previously unselected package apt-transport-https.
(Reading database ... 60030 files and directories currently installed.)
Preparing to unpack .../apt-transport-https_1.6.12_all.deb ...
Unpacking apt-transport-https (1.6.12) ...
Preparing to unpack .../software-properties-common_0.96.24.32.12_all.deb ...
Unpacking software-properties-common (0.96.24.32.12) over (0.96.24.32.11) ...
Preparing to unpack .../python3-software-properties_0.96.24.32.12_all.deb ...
Unpacking python3-software-properties (0.96.24.32.12) over (0.96.24.32.11) ...
Setting up apt-transport-https (1.6.12) ...
Setting up python3-software-properties (0.96.24.32.12) ...
Setting up software-properties-common (0.96.24.32.12) ...
Processing triggers for dbus (1.12.2-1ubuntu1.1) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
ubuntu@cloudera:~$
```

Pour clarifier, voici une brève ventilation de chaque commande:

- **apt-transport-https** : permet au gestionnaire de paquets de transférer des fichiers et des données via https

Ecosystème Hadoop

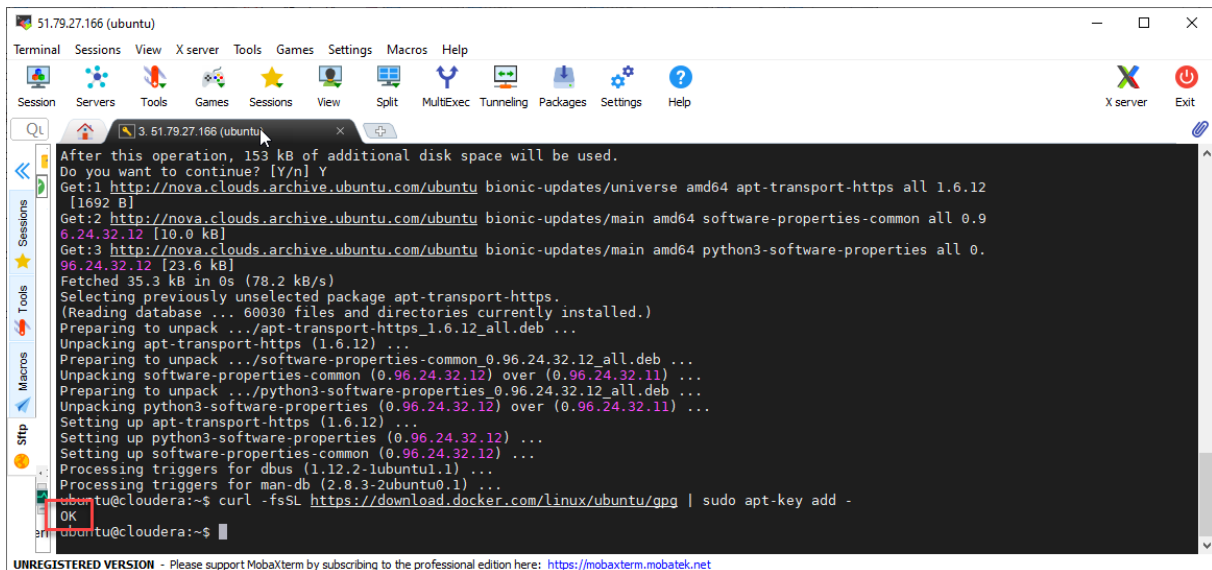
- **ca-certificats** : permet au système (et au navigateur Web) de vérifier les certificats de sécurité
- **curl** : Ceci est un outil pour transférer des données
- **software-properties-common** : ajoute des scripts pour gérer les logiciels

Étape 3: ajouter la clé GPG de Docker

La clé GPG est une fonction de sécurité.

Pour vous assurer que le logiciel que vous installez est authentique, entrez:

```
curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -
```



```
51.79.27.166 (ubuntu)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
3. 51.79.27.166 (ubuntu)
After this operation, 153 kB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/universe amd64 apt-transport-https all 1.6.12
[1692 B]
Get:2 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/main amd64 software-properties-common all 0.9
6.24.32.12 [10.0 kB]
Get:3 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates/main amd64 python3-software-properties all 0.
96.24.32.12 [23.6 kB]
Fetched 35.3 kB in 0s (78.2 kB/s)
Selecting previously unselected package apt-transport-https.
(Reading database ... 60030 files and directories currently installed.)
Preparing to unpack .../apt-transport-https_1.6.12_all.deb ...
Unpacking apt-transport-https (1.6.12) ...
Preparing to unpack .../software-properties-common_0.96.24.32.12_all.deb ...
Unpacking software-properties-common (0.96.24.32.12) over (0.96.24.32.11) ...
Preparing to unpack .../python3-software-properties_0.96.24.32.12_all.deb ...
Unpacking python3-software-properties (0.96.24.32.12) over (0.96.24.32.11) ...
Setting up apt-transport-https (1.6.12) ...
Setting up python3-software-properties (0.96.24.32.12) ...
Setting up software-properties-common (0.96.24.32.12) ...
Processing triggers for dbus (1.12.2-1ubuntu1.1) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
ubuntu@cloudera:~$ curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -
OK
ubuntu@cloudera:~$
```

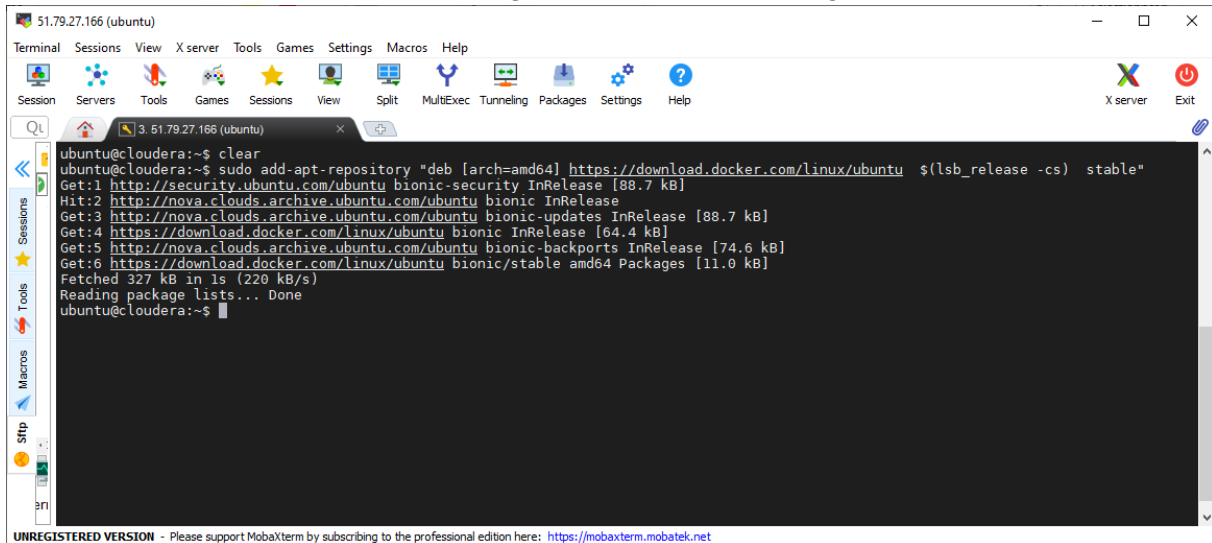
Étape 4: installer le référentiel Docker

Pour installer le référentiel Docker, entrez la commande:

```
sudo add-apt-repository "deb [arch=amd64] https://download.docker.com/linux/ubuntu $(lsb
_release -cs) stable"
```

La commande « **\$(lsb_release -cs)** » scanne et renvoie le nom de code de votre installation Ubuntu - dans ce cas, Bionic. En outre, le dernier mot de la commande - **stable** - est le type de version Docker.

Ecosystème Hadoop



```
ubuntu@cloudera:~$ sudo add-apt-repository "deb [arch=amd64] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable"
Get:1 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Hit:2 http://nova.clouds.archive.ubuntu.com/ubuntu bionic InRelease
Get:3 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]
Get:4 https://download.docker.com/linux/ubuntu bionic InRelease [64.4 kB]
Get:5 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-backports InRelease [74.6 kB]
Get:6 https://download.docker.com/linux/ubuntu bionic/stable amd64 Packages [11.0 kB]
Fetched 327 kB in 1s (220 kB/s)
Reading package lists... Done
ubuntu@cloudera:~$
```

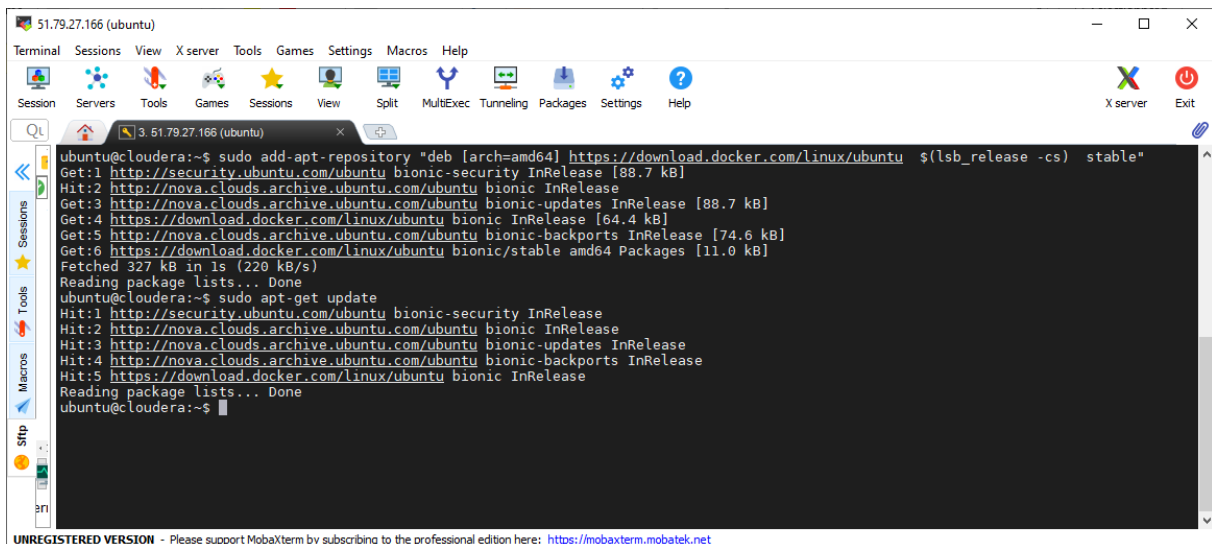
Une version stable est testée et confirmée pour fonctionner, mais les mises à jour sont publiées moins fréquemment.

Une version stable est testée et confirmée pour fonctionner, mais les mises à jour sont publiées moins fréquemment. Vous pouvez remplacer **edge** si vous souhaitez des mises à jour plus fréquentes, au prix d'une instabilité potentielle. Il existe d'autres référentiels, mais ils sont plus risqués - plus d'informations peuvent être trouvées sur la [page Web de Docker](#).

Étape 5: mise à jour des référentiels

Mettez à jour les référentiels que vous venez d'ajouter :

sudo apt-get update



```
ubuntu@cloudera:~$ sudo apt-get update
Hit:1 http://security.ubuntu.com/ubuntu bionic-security InRelease
Hit:2 http://nova.clouds.archive.ubuntu.com/ubuntu bionic InRelease
Hit:3 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-updates InRelease
Hit:4 http://nova.clouds.archive.ubuntu.com/ubuntu bionic-backports InRelease
Hit:5 https://download.docker.com/linux/ubuntu bionic InRelease
Reading package lists... Done
ubuntu@cloudera:~$
```

Étape 6: installer la dernière version de Docker

sudo apt-get install docker-ce docker-ce-cli containerd.io

Ecosystème Hadoop

```
51.79.27.166 (ubuntu)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help

Reading package lists... Done
ubuntu@cloudera:~$ ^C
ubuntu@cloudera:~$ sudo apt-get install docker-ce docker-ce-cli containerd.io
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following package was automatically installed and is no longer required:
  grub-pc-bin
Use 'sudo apt autoremove' to remove it.
The following additional packages will be installed:
  aufs-tools cgroupfs-mount libltdl7 pigz
The following NEW packages will be installed:
  aufs-tools cgroupfs-mount containerd.io docker-ce docker-ce-cli libltdl7 pigz
0 upgraded, 7 newly installed, 0 to remove and 97 not upgraded.
Need to get 85.8 MB of archives.
After this operation, 385 MB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://nova.clouds.archive.ubuntu.com/ubuntu bionic/universe amd64 pigz amd64 2.4-1 [57.4 kB]
Get:2 https://download.docker.com/linux/ubuntu bionic/stable amd64 containerd.io amd64 1.2.13-1 [20.1 MB]
Get:3 http://nova.clouds.archive.ubuntu.com/ubuntu bionic/universe amd64 aufs-tools amd64 1:4.9+20170918-lubuntul [104 kB]
Get:4 http://nova.clouds.archive.ubuntu.com/ubuntu bionic/universe amd64 cgroupfs-mount all 1.4 [6320 B]
Get:5 http://nova.clouds.archive.ubuntu.com/ubuntu bionic/main amd64 libltdl7 amd64 2.4.6-2 [38.8 kB]
Get:6 https://download.docker.com/linux/ubuntu bionic/stable amd64 docker-ce-cli amd64 5:19.03.8~3-0-ubuntu-bionic [42.6 MB]
39% [6 docker-ce-cli 6035 kB/42.6 MB 14%]
```

Installation de Docker Compose sur les systèmes Linux

Sous Linux, vous pouvez télécharger le binaire Docker Compose depuis la [page de publication du référentiel Compose sur GitHub](#). Suivez les instructions du lien, qui impliquent d'exécuter la curl commande dans votre terminal pour télécharger les binaires. Ces instructions étape par étape sont également incluses ci-dessous.

Pour alpine, les paquets de dépendance suivants sont nécessaires: py-pip, python-dev, libffi-dev, openssl-dev, gcc, libc-dev et make.

1. Exécutez cette commande pour télécharger la version stable actuelle de Docker Compose:

```
sudo curl -L "https://github.com/docker/compose/releases/download/1.25.5/docker-compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose
```

Pour installer une version différente de Compose, remplacez-la 1.25.5 par la version de Compose que vous souhaitez utiliser.

Si vous rencontrez des problèmes lors de l'installation avec curl, consultez l'onglet [Options d'installation alternatives](#) ci-dessus.

```
51.79.27.166 (ubuntu)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help

restart  Restart one or more containers
rm       Remove one or more containers
rmi      Remove one or more images
run      Run a command in a new container
save     Save one or more images to a tar archive (streamed to STDOUT by default)
search   Search the Docker Hub for images
start    Start one or more stopped containers
stats    Display a live stream of container(s) resource usage statistics
stop     Stop one or more running containers
tag      Create a tag TARGET_IMAGE that refers to SOURCE_IMAGE
top      Display the running processes of a container
unpause  Unpause all processes within one or more containers
update   Update configuration of one or more containers
version  Show the Docker version information
wait     Block until one or more containers stop, then print their exit codes

Run 'docker COMMAND --help' for more information on a command.
ubuntu@cloudera:~$ sudo curl -L "https://github.com/docker/compose/releases/download/1.25.5/docker-compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total   Spent    Left   Speed
100 638    100 638    0     0    4225      0 --:--:-- --:--:-- --:--:-- 4225
100 16.7M  100 16.7M    0     0   8857k      0  0:00:01  0:00:01 --:--:-- 10.0M
ubuntu@cloudera:~$
```

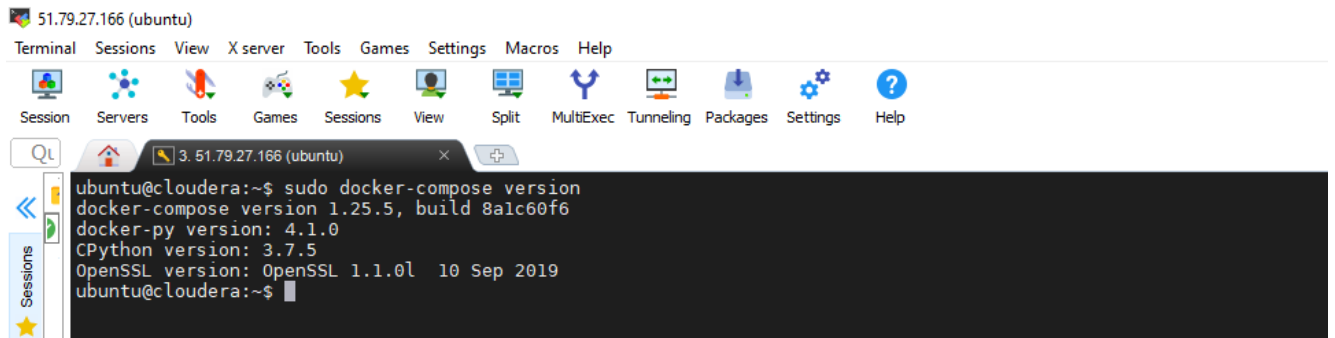
Ecosystème Hadoop

2. Appliquez des autorisations exécutables au binaire:

```
sudo chmod +x /usr/local/bin/docker-compose
```

3. Verifier que la version de docker-compose

```
sudo docker-compose version
```



```
ubuntu@cloudera:~$ sudo docker-compose version
docker-compose version 1.25.5, build 8alc60f6
docker-py version: 4.1.0
CPython version: 3.7.5
OpenSSL version: OpenSSL 1.1.0l 10 Sep 2019
ubuntu@cloudera:~$
```

Remarque : Si la commande docker-compose échoue après l'installation, vérifiez votre chemin. Vous pouvez également créer un lien symbolique vers /usr/bin ou tout autre répertoire de votre chemin.

Installation Netbeans 8.2 avec java 8 et jdk 1.7 sous ubuntu

Installation JDK 1.7

Téléchargez le JDK pour Linux 32 bits ou 64 bits (par exemple: jdk-7u80-linux-x64.tar.gz) via le lien suivant <https://www.oracle.com/java/technologies/javase/javase7-archive-downloads.html>

Java SE Development Kit 7u80		
This software is licensed under the Oracle Binary Code License Agreement for Java SE		
Product / File Description	File Size	Download
Linux x86	130.44 MB	jdk-7u80-linux-i586.rpm
Linux x86	147.68 MB	jdk-7u80-linux-i586.tar.gz
Linux x64	131.69 MB	jdk-7u80-linux-x64.rpm
Linux x64	146.42 MB	jdk-7u80-linux-x64.tar.gz
Mac OS X x64	196.94 MB	jdk-7u80-macosx-x64.dmg
Solaris x86 (SVR4 package)	140.77 MB	jdk-7u80-solaris-i586.tar.Z
Solaris x86	96.41 MB	jdk-7u80-solaris-i586.tar.gz
Solaris x64 (SVR4 package)	24.72 MB	jdk-7u80-solaris-x64.tar.Z

ou en tapant la commande suivante



Ecosystème Hadoop

```
$ wget https://download.oracle.com/otn/java/jdk/7u80-b15/jdk-7u80-linux-x64.tar.gz?AuthParam=1605212725_8dcee86a906e4feffd32e8b294be98a8 -O jdk-7u80-linux-x64.tar.gz
```

Accédez à ~ / Téléchargements :

```
cd /home/"your_user_name"/Downloads
```

Créez un répertoire dans /usr/local où java résidera et copiez l'archive tar ici:

```
sudo mkdir -p /usr/local/java  
sudo cp -r jdk-7u80-linux-x64.tar.gz /usr/local/java/
```

Accédez à /usr /local /java :

```
cd /usr/local/java
```

Extrayez l'archive tar:

```
sudo tar xvfz jdk-7u80-linux-x64.tar.gz
```

Vérifiez si l'archive tar a été extraite avec succès:

```
sudo ls -a #you should see jdk1.7.0_80
```

Ouvrez / etc / profile avec les privilèges sudo:

```
sudo nano /etc/profile
```

Faites défiler jusqu'à la fin du fichier à l'aide des touches fléchées et ajoutez les lignes suivantes ci-dessous à la fin du fichier / etc / profile :

```
JAVA_HOME=/usr/local/java/jdk1.7.0_80  
JRE_HOME=/usr/local/java/jdk1.7.0_80  
PATH=$PATH:$JRE_HOME/bin:$JAVA_HOME/bin
```

Exporter les variables d'environnements en tapant les commandes suivantes sur le Shell

```
export JAVA_HOME  
export JRE_HOME  
export PATH
```

Mettre à jour les alternatives:

```
sudo update-alternatives --install "/usr/bin/java" "java" "/usr/local/java/jdk1.7.0_80/bin/java" 1  
sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/local/java/jdk1.7.0_80/bin/javac" 1
```



Ecosystème Hadoop

```
sudo update-alternatives --install "/usr/bin/javaws" "javaws" "/usr/local/java/jdk1.7.0_80/bin/javaws" 1
sudo update-alternatives --set java /usr/local/java/jdk1.7.0_80/bin/java
sudo update-alternatives --set javac /usr/local/java/jdk1.7.0_80/bin/javac
sudo update-alternatives --set javaws /usr/local/java/jdk1.7.0_80/bin/javaws
```

Recharger le profil:

```
source /etc/profile
```

Vérifiez l'installation:

```
java -version
```

Vous devriez recevoir un message qui affiche :

```
ubuntu@cloudera:/usr/local/java$ java -version
java version "1.7.0_80"
Java(TM) SE Runtime Environment (build 1.7.0_80-b15)
Java HotSpot(TM) 64-Bit Server VM (build 24.80-b11, mixed mode)
ubuntu@cloudera:/usr/local/java$
```

Installation de Netbeans avec JDK 1.8

Télécharger NetBeans avec JDK 1.8 à partir de ce lien

<https://www.oracle.com/technetwork/java/javase/downloads/jdk-netbeans-jsp-3413139-esa.html>

Ecosystème Hadoop

- Java SE
- Java EE
- Java ME
- Java SE Subscription
- Java Embedded
- Java Card
- Java TV
- Community
- Java Magazine

Overview Downloads Documentation Community Technologies Training

JDK 8u111 with NetBeans 8.2

This distribution of the JDK includes the Java SE bundle of NetBeans IDE, which is a powerful integrated development environment for developing applications on the Java platform. [Learn more](#)

You must accept the [JDK 8u111 and NetBeans 8.2 Cobundle License Agreement](#) to download this software.

Thank you for accepting the JDK 8u111 and NetBeans 8.2 Cobundle License Agreement; you may now download this software.

Java SE and NetBeans Cobundle (JDK 8u111 and NB 8.2)		
Product / File Description	File Size	Download
Linux x86	386.72 MB	jdk-8u111-nb-8_2-linux-i586.sh
Linux x64	282.57 MB	jdk-8u111-nb-8_2-linux-x64.sh
Mac OS X x64	342.99 MB	jdk-8u111-nb-8_2-macosx-x64.dmg
Windows x86	317.21 MB	jdk-8u111-nb-8_2-windows-i586.exe
Windows x64	326.03 MB	jdk-8u111-nb-8_2-windows-x64.exe

- License
- Java SE 8 Readme
- NB 8.2 3rd Party Readme
- Installation Instructions
- Java SE Release Notes
- NetBeans Release Notes

Java SE Development Kit 8u111 和 NetBeans IDE 8.2 复合软件包下载 (简体中文)

- 安装说明
- Java SE 发行说明
- NetBeans 发行说明

Ou en tapant la commande suivante

```
wget https://download.oracle.com/otn-pub/java/jdk-nb/8u111-8.2/jdk-8u111-nb-8_2-linux-x64.sh?AuthParam=1605213964_bebb0c8b4e23db77bd1b995d02980ea4
```

Changer les permissions de fichier avec la commande suivante

```
sudo chmod +x jdk-8u111-nb-8_2-linux-x64.sh
```

Installer maintenant netbeans et JDK 1.8 en tapant la commande suivante

```
sudo ./jdk-8u111-nb-8_2-linux-x64.sh
```

Ecosystème Hadoop

Atelier 1 : Configuration et installation de Hadoop Cloudera

Objectifs

Cet atelier a pour but :

- Installation et configuration de l'environnement **Hadoop** via la plateforme **Cloudera**: <http://www.cloudera.com/>.
- Lancement et test de l'environnement Hadoop.

Téléchargement et Configuration de Cloudera. : Cloudera QuickStart VM

Pour utiliser Hadoop deux solutions sont disponibles. La première est d'utiliser la version proposée par la fondation Apache. Cette version est celle de référence et contient le noyau et quelques interfaces d'administration très simplifiée. La seconde solution est d'utiliser les distributions fournies par des entreprises qui font du service autour d'Hadoop. Dans le cadre de cet atelier, nous utilisons la distribution de la compagnie **Cloudera**: <http://www.cloudera.com/>. Cette distribution a l'avantage d'être gratuite pour Cloudera Standard. Elle fournit également des outils d'administration supplémentaires qui facilitent son usage.

Installation et exécution d'un cluster simple nœud

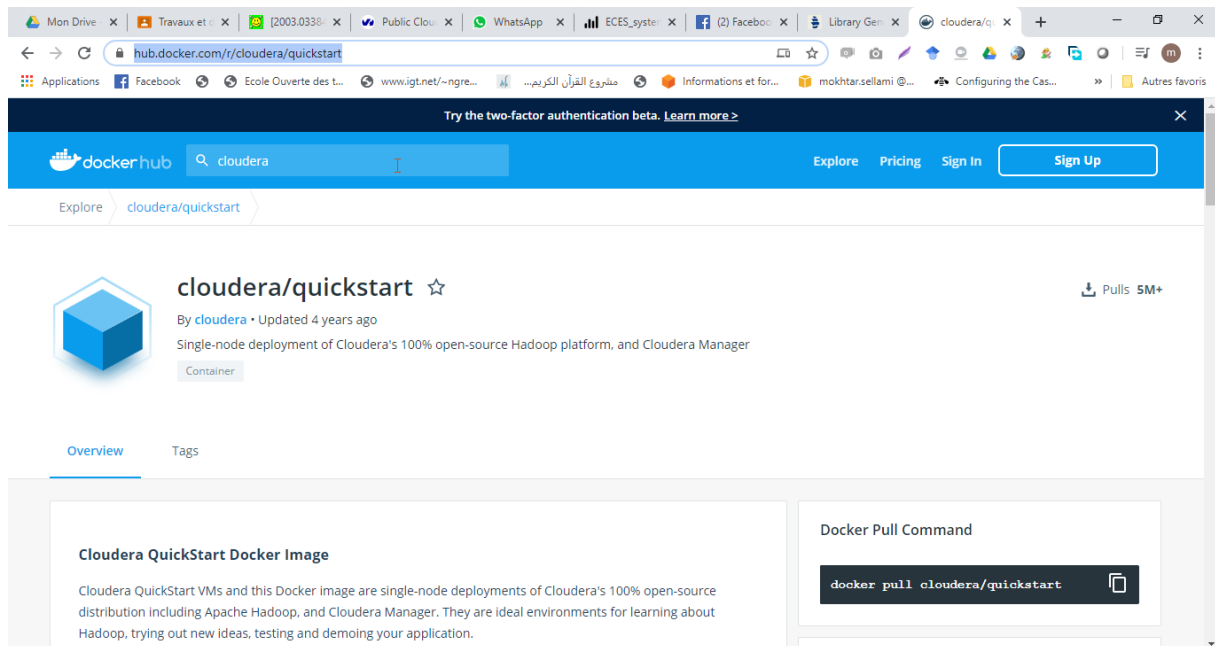
Cloudera fournit des [machines virtuelles](#) prêtes à l'emploi pour VMware, VirtualBox et KVM , docker. Ces machines virtuelles s'exécutent sur Centos . Les machines virtuelles “ Cloudera QuickStart (VMs)” comprennent tout ce que vous devez essayer avec CDH (**C**loudera **D**istributed **H**adoop), Cloudera Manager, Cloudera Impala, et Cloudera Search. Les machines virtuelles utilisent des packages préinstallés de CDC (**C**hange **D**ata **C**apture). Ce qui vous permet de travailler avec ou sans Cloudera Manager. Pour utiliser ces machines virtuelles, vous avez besoins de ces configurations suivantes :

Cloudera QuickStart Docker Image

Cloudera QuickStart VMs and this Docker image are single-node deployments of Cloudera's 100% open-source distribution including Apache Hadoop, and Cloudera Manager. They are ideal environments for learning about Hadoop, trying out new ideas, testing and demoing your application.

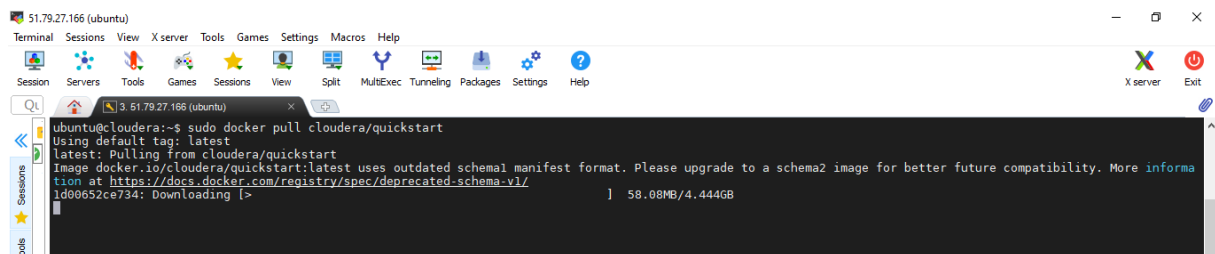
Ecosystème Hadoop

Visiter ce lien <https://hub.docker.com/r/cloudera/quickstart> pour avoir plus de détails sur les prérequis d'installation.



Dans le ssh shell de MobaXterm lancer cette commande

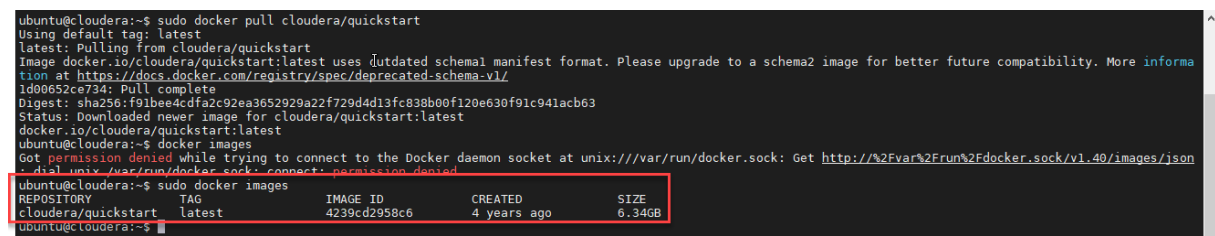
```
sudo docker pull cloudera/quickstart
```



Exécution d'un conteneur Cloudera QuickStart

Pour exécuter un conteneur à l'aide de l'image, vous devez connaître le nom ou le hachage de l'image. Si vous avez suivi les instructions d'importation ci-dessus, le nom pourrait être cloudera / quickstart: latest (ou autre chose si vous avez téléchargé plusieurs versions). Le hachage est également imprimé dans le terminal lorsque vous importez, ou vous pouvez rechercher les hachages de toutes les images importées avec :

```
sudo docker images
```



Ecosystème Hadoop

Une fois que vous connaissez le nom ou le hachage de l'image, vous pouvez l'exécuter:

```
docker run --hostname=quickstart.cloudera --privileged=true -t -i [OPTIONS] [IMAGE]
/usr/bin/docker-quickstart
```

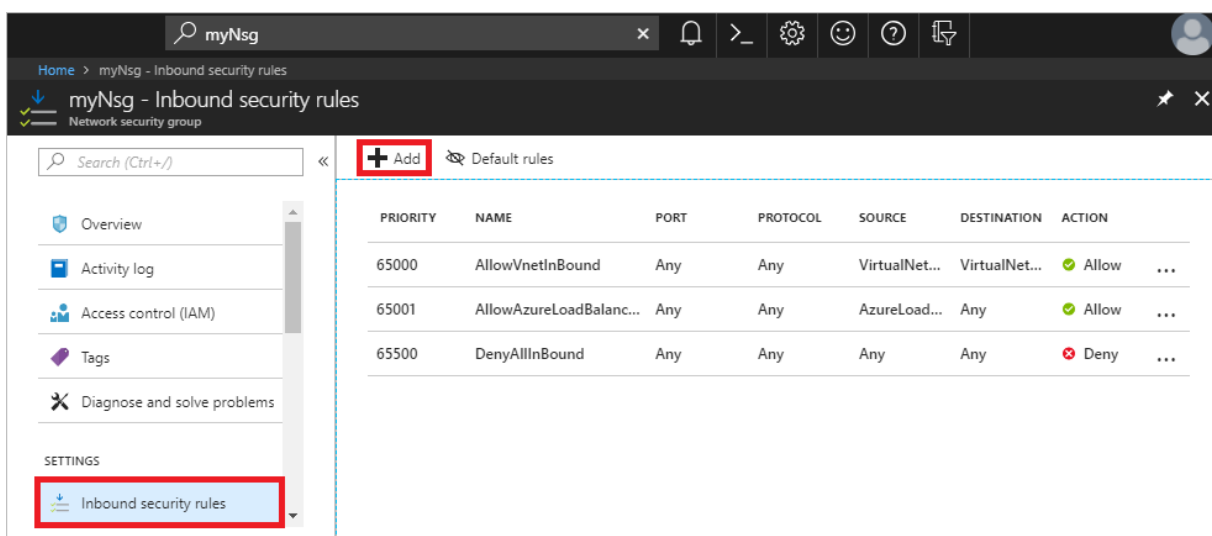
Les explications pour les indicateurs requis et d'autres options sont dans le tableau suivant:

--hostname=quickstart.cloudera	Required: pseudo-distributed configuration assumes this hostname
--privileged=true	Required: for HBase, MySQL-backed Hive metastore, Hue, Oozie, Sentry, and Cloudera Manager, and possibly others
-t	Required: once services are started, a Bash shell takes over and will die without this
-i	Required: if you want to use the terminal, either immediately or attach later
-p 8888	Recommended: maps the Hue port in the guest to another port on the host
-p [PORT]	Optional: map any other ports (e.g. 7180 for Cloudera Manager, 80 for a guided tutorial)
-d	Optional: runs the container in the background

Pour exécuter le conteneur Cloudera/Suickstart tout en assurant que Hue est accessible via le port 8888 en tapant http://IP_Machine_Azure : 8888 dans navigateur. Il faut créer une règle de sécurité de trafic entrant autorisant le trafic et affecter des valeurs aux paramètres suivants :

- Plages de ports de destination : 8888
- Plages de ports sources : * (autorise n'importe quel port source)
- Valeur de priorité : entrez une valeur de priorité inférieure à 65 500 et prioritaire par rapport à la règle fourre-tout par défaut de refus de trafic entrant.
- Associer le groupe de sécurité réseau à l'interface réseau de machine virtuelle ou au sous-réseau.

Accéder au Portail Azure et ajouter la règle.



PRIORITY	NAME	PORT	PROTOCOL	SOURCE	DESTINATION	ACTION
65000	AllowVnetInBound	Any	Any	VirtualNet...	VirtualNet...	Allow
65001	AllowAzureLoadBalanc...	Any	Any	AzureLoad...	Any	Allow
65500	DenyAllInBound	Any	Any	Any	Any	Deny

Une fois terminé la configuration de l'environnement et la machine virtuelle maintenant nous exécutons la commande dans le ssh shell.



```
51.79.27.166 (ubuntu)
Terminal Sessions View Xserver Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split Multitex Tunneling Packages Settings Help
51.79.27.166 (ubuntu)
r/lib/hadoop-mapreduce/lib/commons-io-2.4.jar:/usr/lib/hadoop-mapreduce/lib/guice-3.0.jar:/usr/lib/hadoop-mapreduce/lib/guice-servlet-3.0.jar:/usr/lib/hadoop-mapreduce/lib/hamcrest-core-1.3.jar:/usr/lib/hadoop-mapreduce/lib/jackson-core-asl-1.8.8.jar:/usr/lib/hadoop-mapreduce/lib/jackson-mapper-asl-1.8.8.jar:/usr/lib/hadoop-mapreduce/lib/ivybox.inject-1.jar:/usr/lib/hadoop-mapreduce/lib/jersey-core-1.9.jar:/usr/lib/hadoop-mapreduce/lib/jersey-guice-1.0.jar:/usr/lib/hadoop-mapreduce/lib/jersey-guice-1.0.jar:/usr/lib/hadoop-mapreduce/lib/jetty-3.6.2.Final.jar:/usr/lib/hadoop-mapreduce/lib/leveldbjni-all-1.8.jar:/usr/lib/hadoop-mapreduce/lib/log4j-1.2.17.jar:/usr/lib/hadoop-mapreduce/lib/netty-3.6.2.Final.jar:/usr/lib/hadoop-mapreduce/lib/paramanet-2.3.jar:/usr/lib/hadoop-mapreduce/lib/protobuf-java-2.5.0.jar:/usr/lib/hadoop-mapreduce/lib/snappy-java-1.0.4.1.jar:/usr/lib/hadoop-mapreduce/lib/xz-1.0.jar:/usr/lib/hadoop-mapreduce/modules/*.*jar
STARTUP_MSG: build = http://github.com/cloudera/hadoop -r C00978c67b0d3fe9f3b896b5030741bd40bf541a; compiled by 'jenkins' on 2016-03-23T18:36Z
STARTUP_MSG: build = 1.7.0_67
*****
Starting Hadoop historyserver: [ OK ]
Starting nodemanager, logging to /var/log/hadoop-yarn/yarn-nodemanager-quickstart.cloudera.out
Starting Hadoop nodemanager: [ OK ]
Starting resourcemanager, logging to /var/log/hadoop-yarn/yarn-resource-manager-quickstart.cloudera.out
Starting Hadoop resourcemanager: [ OK ]
Starting master, logging to /var/log/hbase/hbase-hbase-master-quickstart.cloudera.out
Starting HBase master daemon (hbase-master): [ OK ]
Starting rest, logging to /var/log/hbase/hbase-hbase-rest-quickstart.cloudera.out
Starting HBase rest daemon (hbase-rest): [ OK ]
Starting thrift, logging to /var/log/hbase/hbase-hbase-thrift-quickstart.cloudera.out
Starting HBase thrift daemon (hbase-thrift): [ OK ]
Starting Hive Metastore (hive-metastore): [ OK ]
Starting Hive Server2 (hive-server2): [ OK ]
Starting Sqoop Server: [ OK ]
Sqoop home directory: /usr/lib/sqoop2
Setting SQOOP_HTTP_PORT: 12000
Setting SQOOP_ADMIN_PORT: 12001
Setting CATALINA_OPTS: -Xmx2048m
Adding to CATALINA_OPTS: -Dsqoop.http.port=12000 -Dsqoop.admin.port=12001
Using CATALINA_BASE: /var/lib/sqoop2/tomcat-deployment
Using CATALINA_HOME: /usr/lib/bigtop-tomcat
Using CATALINA_TMPDIR: /var/tmp/sqoop2
Using JRE_HOME: /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH: /usr/lib/bigtop-tomcat/bin/bootstrap.jar
Using CATALINA_PID: /var/run/sqoop2/sqoop-server-sqoop2.pid
Starting Spark history-server (spark-history-server): [ OK ]
Starting Hadoop HBase regionserver daemon: starting regionserver, logging to /var/log/hbase/hbase-hbase-regionserver-quickstart.cloudera.out
Starting hbase-regionserver: [ OK ]
*****
UNREGISTERED VERSION - Please support MobalTerm by subscribing to the professional edition here: https://mobalterm.mobal.net
```

The screenshot displays a terminal window titled "51.79.27.166 (ubuntu)" with a standard Ubuntu desktop environment at the top. The terminal output shows the configuration of Oozie services, including setting log directories, keystore paths, and instance IDs. It then proceeds to start the Solr server daemon and the Impala Catalog and Server daemons, with status indicators [OK] and [PW] shown for the latter two. The prompt at the bottom is [root@quickstart ~]#.

```
Using OOOIE LOG: /var/log/oozie
Setting OOOIE LOG4J FILE: oozie-log4j.properties
Setting OOOIE LOG4J RELOAD: 10
Setting OOOIE HTTP HOSTNAME: quickstart.cloudera
Setting OOOIE HTTP PORT: 11000
Setting OOOIE ADMIN PORT: 11001
Using OOOIE HTTPS PORT: 11443
Setting OOOIE BASE URL: http://quickstart.cloudera:11000/oozie
Using CATALINA BASE: /var/lib/oozie/tomcat-deployment
Setting OOOIE HTTPS KEYSTORE FILE: /var/lib/oozie/.keystore
Using OOOIE HTTPS KEYSTORE PASS: password
Setting OOOIE INSTANCE ID: quickstart.cloudera
Setting CATALINA OUT: /var/log/oozie/catalina.out
Using CATALINA PID: /var/run/oozie/oozie.pid

Using CATALINA OPTS: -Doozie.https.port=11443 -Doozie.https.keystore.pass=password -Xmx1024m -Doozie.https.port=11443 -Doozie.https.keystore.pass=password -Dmx1024m -Dderby.stream.error.file=/var/log/oozie/derby.log
Adding to CATALINA OPTS: -Doozie.home.dir=/usr/lib/oozie -Doozie.config.dir=/etc/oozie/conf -Doozie.log.dir=/var/log/oozie -Doozie.data.dir=/var/lib/oozie -Doozie.instance.id=quickstart.cloudera -Doozie.config.file=oozie-site.xml -Doozie.log4j.file=oozie-log4j.properties -Doozie.log4j.reload=10 -Doozie.http.hostname=quickstart.cloudera -Doozie.admin.port=11001 -Doozie.http.port=11000 -Doozie.https.port=11443 -Doozie.base.url=http://quickstart.cloudera:11000/oozie -Doozie.https.keystore.file=/var/lib/oozie/.keystore -Doozie.https.keystore.pass=password -Djava.library.path=/usr/lib/hadoop/lib/native:/usr/lib/hadoop/lib/native

Using CATALINA BASE: /var/lib/oozie/tomcat-deployment
Using CATALINA HOME: /usr/lib/bigtop-tomcat
Using CATALINA TMPDIR: /var/lib/oozie
Using JRE HOME: /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH: /usr/lib/bigtop-tomcat/bin/bootstrap.jar
Using CATALINA PID: /var/run/oozie/oozie.pid
Starting Solr server daemon: [ OK ]

Using CATALINA BASE: /var/lib/solr/tomcat-deployment
Using CATALINA HOME: /usr/lib/solr/..bigtop-tomcat
Using CATALINA TMPDIR: /var/lib/solr/
Using JRE HOME: /usr/java/jdk1.7.0_67-cloudera
Using CLASSPATH: /usr/lib/solr/..bigtop-tomcat/bin/bootstrap.jar
Using CATALINA PID: /var/run/solr/solr.pid
Started Impala Catalog Server (catalogd): [ OK ]
Started Impala Server (imlad): [ PW ]

[ root@quickstart ~ ]#
```

UNREGISTERED VERSION - Please support MobaxTerm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

Il faut vérifier que tous les services sont bien démarrés, pour cela il faut taper la commande suivante



Ecosystème Hadoop

```
51.79.27.166 (ubuntu)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultExec Tunneling Packages Settings Help

[root@quickstart ~]# service
Usage: service < option > | [ service_name [ command | --full-restart ] ]
[root@quickstart ~]# service --status-all
atd is stopped

if [ "$1" == "start" ]; then
  if [ "${EC2}" == 'true' ]; then
    FIRST_BOOT_FLAG=/var/lib/cloudera-quickstart/.ec2-key-installed
    if [ ! -f "$FIRST_BOOT_FLAG" ]; then
      METADATA_API=http://169.254.169.254/latest/meta-data
      KEY_URL=${METADATA_API}/public-keys/0/openssh-key
      SSH_DIR=/home/cloudera/.ssh
      mkdir -p ${SSH_DIR}
      chown cloudera:cloudera ${SSH_DIR}
      curl ${KEY_URL} >> ${SSH_DIR}/authorized_keys
      touch ${FIRST_BOOT_FLAG}
    fi
  fi
  if [ "${DOCKER}" != 'true' ]; then
    if [ ! -f /sys/kernel/mm/redhat_transparent_hugepage/defrag ]; then
      echo never > /sys/kernel/mm/redhat_transparent_hugepage/defrag
    fi
  fi
  cloudera-quickstart-ip
  HOSTNAME=quickstart.cloudera
  hostname ${HOSTNAME}
  sed -i -e "s/HOSTNAME=.*/HOSTNAME=${HOSTNAME}/" /etc/sysconfig/network
fi

(
  cd /var/lib/cloudera-quickstart/tutorial;
  nohup python -m SimpleHTTPServer 80 &
)

# TODO: check for expired CM license and update config.js accordingly
fi
+ '[' status == start ']'
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

```
51.79.27.166 (ubuntu)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultExec Tunneling Packages Settings Help

Manager has started all of the services it manages and is ready to accept
connections from clients.

cron is stopped
elume_agent is not running [ FAILED ]
Hadoop datanode is running [ OK ]
Hadoop journalnode is running [ OK ]
Hadoop namenode is running [ OK ]
Hadoop secondarynamenode is running [ OK ]
Hadoop httpfs is running [ OK ]
Hadoop historyserver is running [ OK ]
Hadoop nodemanager is running [ OK ]
Hadoop proxyserver is not running [ FAILED ]
Hadoop resourcemanager is running [ OK ]
HBase master daemon is running [ OK ]
HBase-regionserver is running [ OK ]
HBase-rest-daemon is running [ OK ]
HBase-Solr Indexer is not running [ FAILED ]
HBase-thrift-daemon is running [ OK ]
Hive Metastore is running [ OK ]
Hive Server2 is running [ OK ]
htcacheclean is stopped
httpd is stopped
supervisor (pid 2637) is running...
Impala Catalog Server is running [ OK ]
Impala Server is running [ OK ]
Impala State Store Server is running [ OK ]
iptables: Firewall is not running.
mysqld (pid 169) is running...
netconsole module not loaded
Configured devices:
lo eth0
Currently active devices:
lo eth0if5
ntpd is stopped
running
rdisc is stopped
rpcbind is stopped
rsyslogd is stopped
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

Liste de services qui ont échoué lors de démarrage de la machine

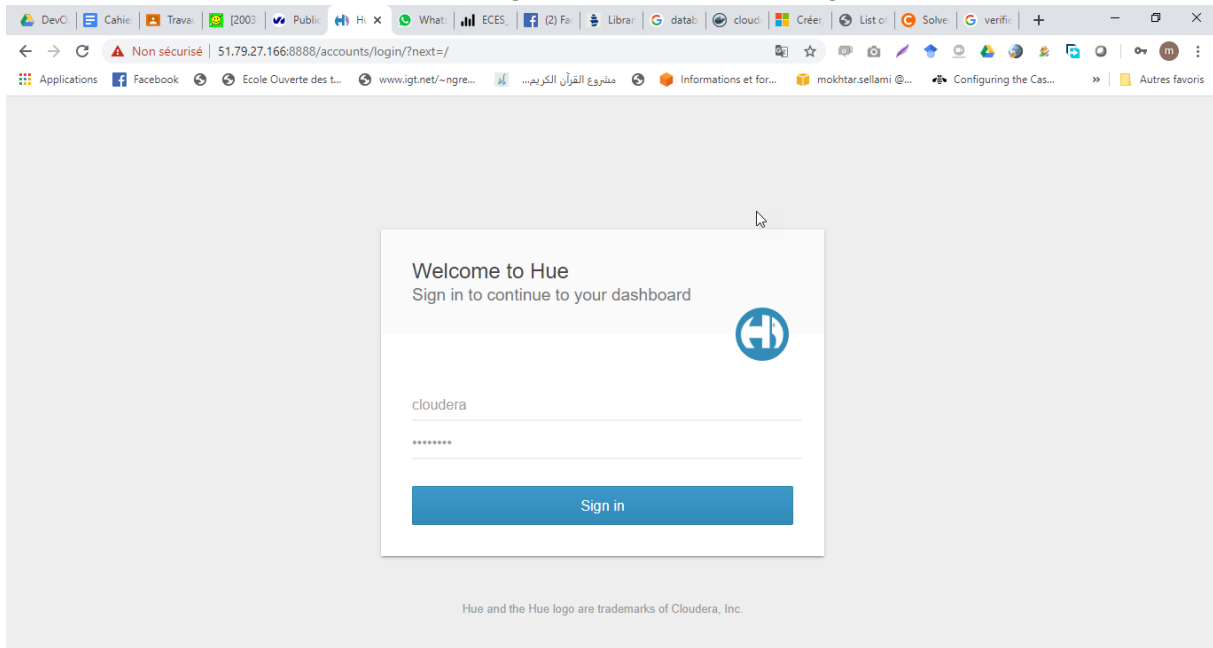
Pour redémarrer ces services il faut se refaire à la documentation fournit par Cloudera via ce lien

https://docs.cloudera.com/documentation/enterprise/5-7-x/topics/cdh_admin_config.html

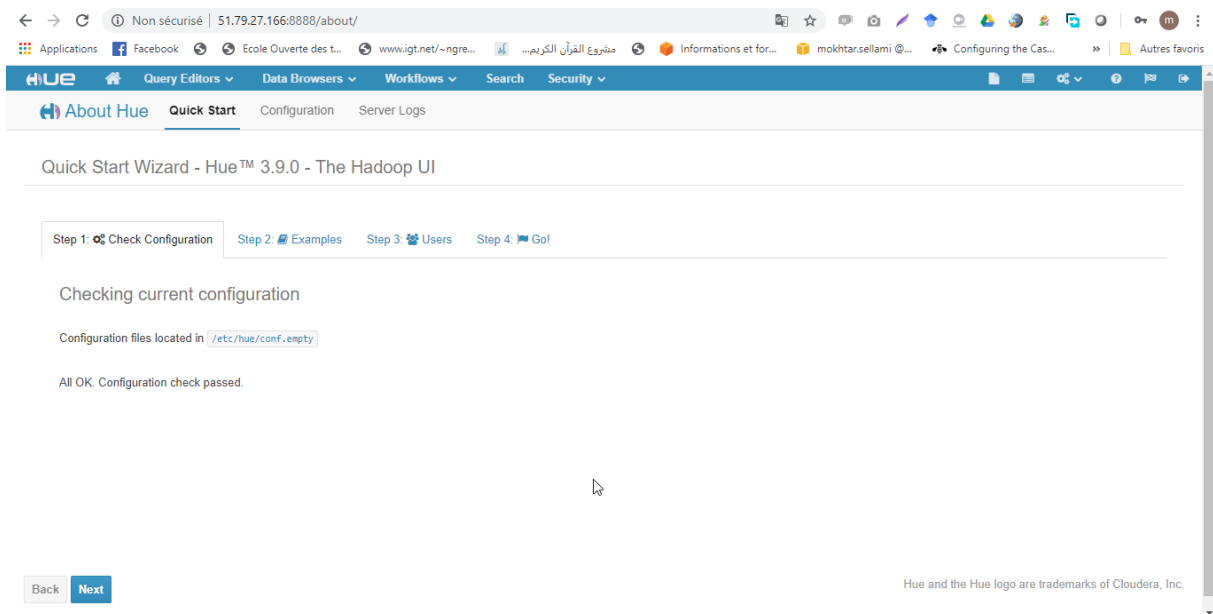
Une fois tous les services sont bien démarrés vous pouvez accéder maintenant Hue en tapant l'adresse

http://IP_VM_Azure:8888

Ecosystème Hadoop



Taper comme Login et password « cloudera » afin de loguer a Hue



Installation d'un éditeur de texte pour éditer les configurations d'Apache Hadoop

Il existe plusieurs éditeurs de texte dans les environnements linux, vous pouvez choisir à installer l'un de ces éditeurs (VIM, Nano). Par exemple vous pouvez installer nano en utilisant cette commande

```
[root@quickstart conf]# sudo yum install nano
```

Ecosystème Hadoop

Atelier 2 : HDFS

Objectifs

Après avoir terminé ces travaux pratiques, vous serez en mesure de :

- ✓ Utiliser les commandes Hadoop pour explorer le HDFS sur le système Hadoop
- ✓ Utiliser la console web Hue pour explorer le HDFS sur le système Hadoop

Configuration requise

Pour compléter cet atelier vous aurez besoin des éléments suivants :

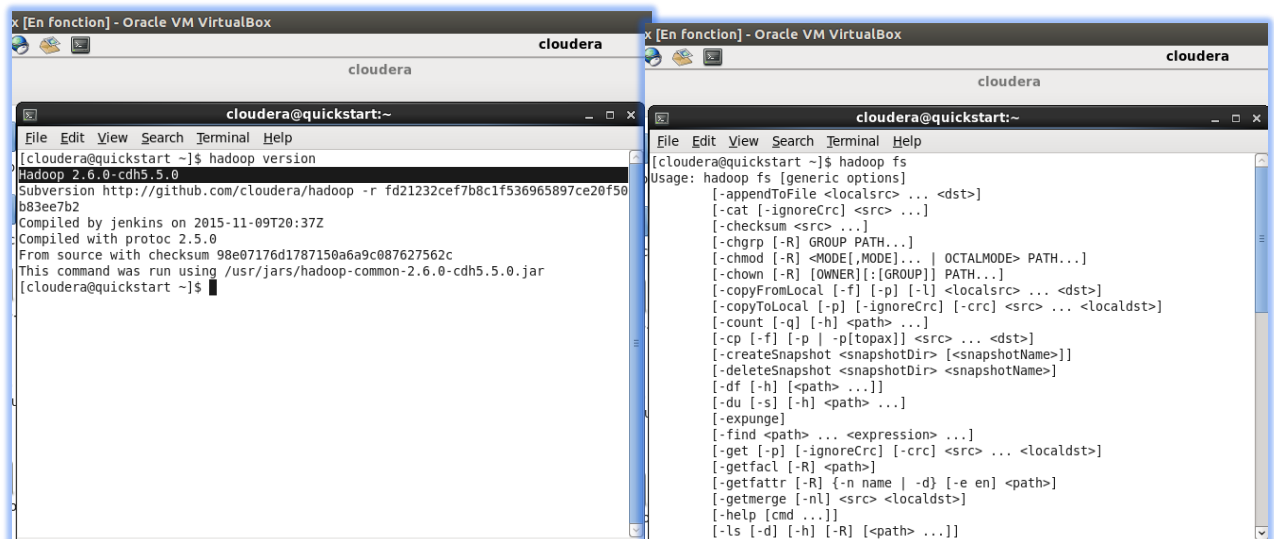
1. Image : Cloudera Quickstart 5.7
2. Accès au Azure VM via SSH et Navigateur web

1. Vérification de Hadoop

Lancer un terminal. Taper la commande suivante

```
[cloudera@quickstart~]$ hadoop version
```

La figure suivant montre la version installée (hadoop 2.6.0-cdh5.5.0) sur la machine virtuelle :



```
[cloudera@quickstart~]$ hadoop version
Hadoop 2.6.0-cdh5.5.0
Subversion http://github.com/cloudera/hadoop -r fd21232cef7b8c1f536965897ce20f50b83ee7b2
Compiled by jenkins on 2015-11-09T20:37Z
Compiled with protoc 2.5.0
From source with checksum 98e07176d1787150a6a9c087627562c
This command was run using /usr/jars/hadoop-common-2.6.0-cdh5.5.0.jar
[cloudera@quickstart ~]$
```

```
[cloudera@quickstart ~]$ hadoop fs
Usage: hadoop fs [generic options]
[-appendToFile <localsrc> ... <dst>]
[-cat [-ignoreCrc] <src> ...]
[-checksum <src> ...]
[-chgrp [-R] GROUP PATH...]
[-chmod [-R] <MODE[,MODE]...> [OCTALMODE] PATH...]
[-chown [-R] [OWNER][:[GROUP]] PATH...]
[-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
[-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
[-count [-q] [-h] <path> ...]
[-cp [-f] [-p | -p[topax]] <src> ... <dst>]
[-createSnapshot <snapshotDir> [<snapshotName>]]
[-deleteSnapshot <snapshotDir> <snapshotName>]
[-df [-h] <path> ...]
[-du [-s] [-h] <path> ...]
[-expunge]
[-find <path> ... <expression> ...]
[-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
[-getfacl [-R] <path>]
[-getfattr [-R] {-n name | -d} [-e en] <path>]
[-getmerge [-nl] <src> <localdst>]
[-help [cmd ...]]
[-ls [-d] [-h] [-R] [<path> ...]]
```

2. Lancement de HDFS

Pour vérifier si les services sont lancés ou pas exécutez la commande :

```
$ service - - status-all
```

Suivant la version du Cloudera et pour lancer HDFS, ainsi que les services NameNode, Secondary NameNode, et DataNode testez les commandes ci-dessous

Ecosystème Hadoop

```
$ sudo service hadoop-hdfs-namenode start
```

Pour le NameNode secondaire

```
$ sudo service hadoop-hdfs-secondarynamenode restart
```

Pour chaque DataNode:

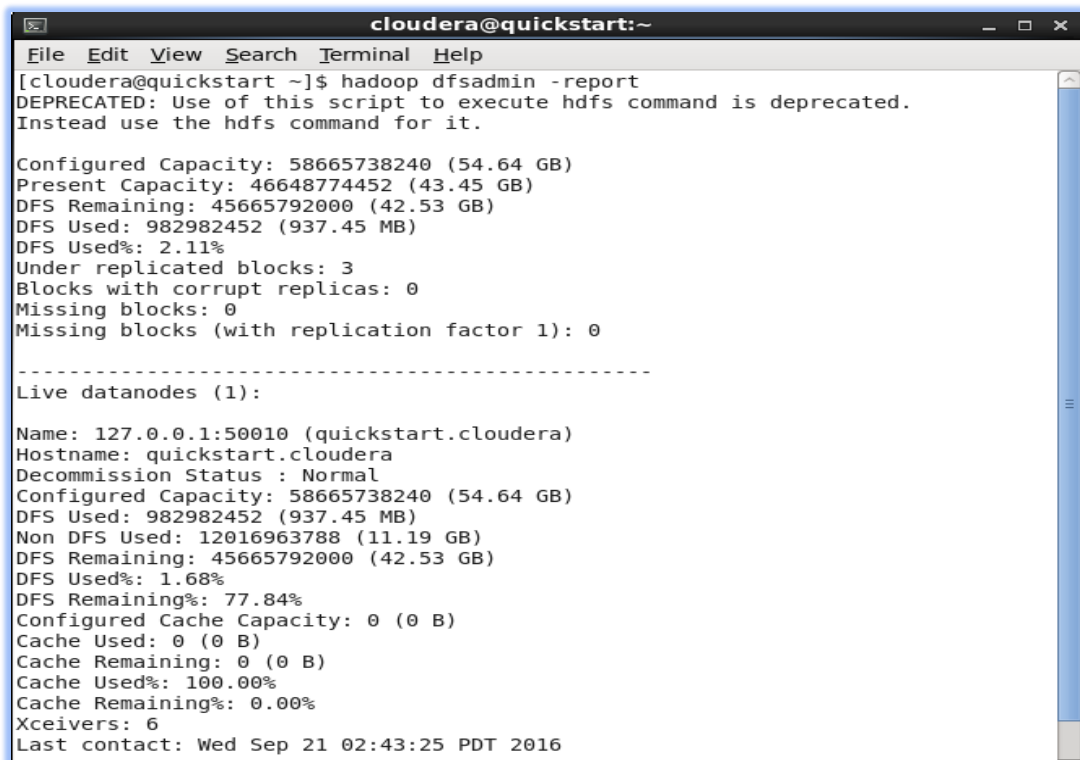
```
$ sudo service hadoop-hdfs-datanode restart
```

3. Vérifier l'état disque de HDFS « HDFS Disk Check »

Plusieurs façons de surveiller l'état du disque HDFS, et cela doit être fait de temps en temps pour éviter des problèmes d'espace qui peut se poser s'il y a un faible stockage du disque restant. Un tel problème peut se produire si le "healthcheck hadoop" ou heartbeat a signalé qu'un nœud est passé en mode hors connexion. Si un nœud est déconnecté pendant un certain laps de temps, les données du nœud déconnecté seront répliquées à d'autres nœuds (car il y a au moins une réplication de 3 nœuds, les données sont toujours disponibles sur les 2 autres nœuds). Si l'espace disque est limité, cela peut rapidement causer un problème. Vous pouvez accéder rapidement au rapport HDFS en exécutant la commande suivante :

```
[cloudrea@quikstart~]$ hadoop dfsadmin -report
```

La figure suivante montre l'état disque du système de fichier HDFS



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ hadoop dfsadmin -report  
DEPRECATED: Use of this script to execute hdfs command is deprecated.  
Instead use the hdfs command for it.  
  
Configured Capacity: 58665738240 (54.64 GB)  
Present Capacity: 46648774452 (43.45 GB)  
DFS Remaining: 45665792000 (42.53 GB)  
DFS Used: 982982452 (937.45 MB)  
DFS Used%: 2.11%  
Under replicated blocks: 3  
Blocks with corrupt replicas: 0  
Missing blocks: 0  
Missing blocks (with replication factor 1): 0  
  
-----  
Live datanodes (1):  
  
Name: 127.0.0.1:50010 (quickstart.cloudera)  
Hostname: quickstart.cloudera  
Decommission Status : Normal  
Configured Capacity: 58665738240 (54.64 GB)  
DFS Used: 982982452 (937.45 MB)  
Non DFS Used: 12016963788 (11.19 GB)  
DFS Remaining: 45665792000 (42.53 GB)  
DFS Used%: 1.68%  
DFS Remaining%: 77.84%  
Configured Cache Capacity: 0 (0 B)  
Cache Used: 0 (0 B)  
Cache Remaining: 0 (0 B)  
Cache Used%: 100.00%  
Cache Remaining%: 0.00%  
Xceivers: 6  
Last contact: Wed Sep 21 02:43:25 PDT 2016
```

Ecosystème Hadoop

Filesystem check (fsck)

Hadoop fournit un utilitaire *fsck* pour vérifier les fichiers dans HDFS. L'outil se penche sur les blocs manquants de tous les datanodes, ainsi que sur des blocs répliqués.

Voici un exemple de vérification de l'ensemble du système de fichiers pour un petit cluster:

hdfsfsck /

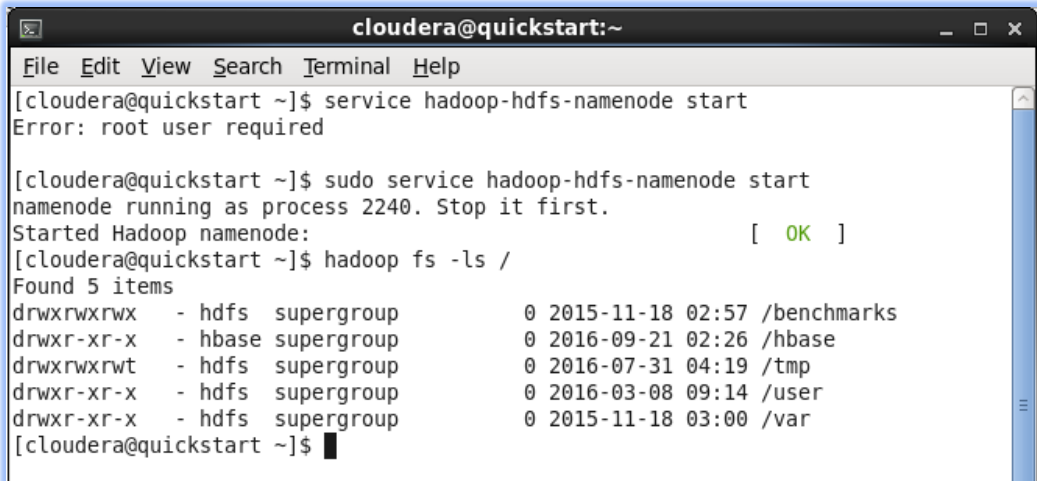
Le *fscktool* fournit un moyen simple de savoir quels blocs se trouvent dans n'importe quel fichier particulier. Par exemple: **hdfsfsck /user/tom/part-00007 -files -blocks -racks**

4. Manipulation De HDFS sous Hadoop (HDFS)

Hadoop Distributed File System (HDFS), permet aux données des utilisateurs d'être organisées sous la forme de fichiers et de répertoires. Il fournit une interface de ligne de commande appelée FS Shell qui permet à un utilisateur d'interagir avec les données dans HDFS et qui sont accessibles aux programmes MapReduce.

4.1. Lister le contenu du répertoire racine.

```
[cloudrea@quikstart~]$ hadoop fs -ls /
```



```
cloudera@quikstart:~  
File Edit View Search Terminal Help  
[cloudera@quikstart ~]$ service hadoop-hdfs-namenode start  
Error: root user required  
  
[cloudera@quikstart ~]$ sudo service hadoop-hdfs-namenode start  
namenode running as process 2240. Stop it first.  
Started Hadoop namenode: [ OK ]  
[cloudera@quikstart ~]$ hadoop fs -ls /  
Found 5 items  
drwxrwxrwx - hdfs supergroup 0 2015-11-18 02:57 /benchmarks  
drwxr-xr-x - hbase supergroup 0 2016-09-21 02:26 /hbase  
drwxrwxrwt - hdfs supergroup 0 2016-07-31 04:19 /tmp  
drwxr-xr-x - hdfs supergroup 0 2016-03-08 09:14 /user  
drwxr-xr-x - hdfs supergroup 0 2015-11-18 03:00 /var  
[cloudera@quikstart ~]$
```

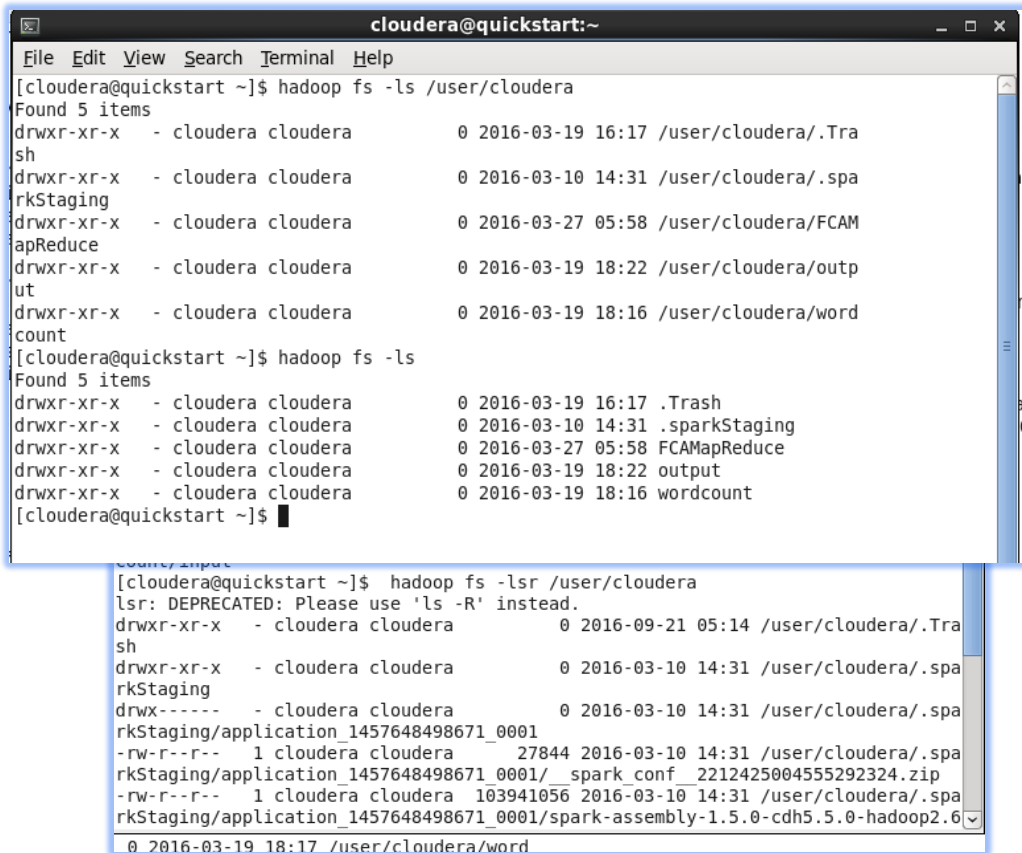
Pour lister le contenu du répertoire /user/cloudera, exécuter :

```
[cloudrea@quikstart~]$ hadoop fs -ls
```

Ou

Ecosystème Hadoop

```
[cloudrea@quikstart~]$ hadoop fs -ls /user/cloudera
```



```
cloudera@quikstart:~  
File Edit View Search Terminal Help  
[cloudera@quikstart ~]$ hadoop fs -ls /user/cloudera  
Found 5 items  
drwxr-xr-x - cloudera cloudera      0 2016-03-19 16:17 /user/cloudera/.Trash  
sh  
drwxr-xr-x - cloudera cloudera      0 2016-03-10 14:31 /user/cloudera/.sparkStaging  
rkStaging  
drwxr-xr-x - cloudera cloudera      0 2016-03-27 05:58 /user/cloudera/FCAMapReduce  
apReduce  
drwxr-xr-x - cloudera cloudera      0 2016-03-19 18:22 /user/cloudera/output  
ut  
drwxr-xr-x - cloudera cloudera      0 2016-03-19 18:16 /user/cloudera/wordcount  
count  
[cloudera@quikstart ~]$ hadoop fs -ls  
Found 5 items  
drwxr-xr-x - cloudera cloudera      0 2016-03-19 16:17 .Trash  
drwxr-xr-x - cloudera cloudera      0 2016-03-10 14:31 .sparkStaging  
drwxr-xr-x - cloudera cloudera      0 2016-03-27 05:58 FCAMapReduce  
drwxr-xr-x - cloudera cloudera      0 2016-03-19 18:22 output  
drwxr-xr-x - cloudera cloudera      0 2016-03-19 18:16 wordcount  
[cloudera@quikstart ~]$  
[cloudera@quikstart ~]$ hadoop fs -lsr /user/cloudera  
lsr: DEPRECATED: Please use 'ls -R' instead.  
drwxr-xr-x - cloudera cloudera      0 2016-09-21 05:14 /user/cloudera/.Trash  
sh  
drwxr-xr-x - cloudera cloudera      0 2016-03-10 14:31 /user/cloudera/.sparkStaging  
rkStaging  
drwx----- - cloudera cloudera      0 2016-03-10 14:31 /user/cloudera/.sparkStaging/application_1457648498671_0001  
-rw-r--r--  1 cloudera cloudera    27844 2016-03-10 14:31 /user/cloudera/.sparkStaging/application_1457648498671_0001/_spark_conf_2212425004555292324.zip  
-rw-r--r--  1 cloudera cloudera 103941056 2016-03-10 14:31 /user/cloudera/.sparkStaging/application_1457648498671_0001/spark-assembly-1.5.0-cdh5.5.0-hadoop2.6  
0 2016-03-19 18:17 /user/cloudera/word
```

Notes : dans la première commande il n'y avait pas le répertoire référencé, mais il est équivalent à

La deuxième commande où `/user/cloudera` est explicitement spécifié. Chaque utilisateur aura son propre répertoire personnel sous `/ utilisateur`. Par exemple, dans le cas de l'utilisateur cloudera, son répertoire est `/user/cloudera`. Toute commande où il n'y a pas de répertoire explicite spécifié sera relatif au répertoire d'accueil de l'utilisateur.

4.2. Création d'un répertoire et affichage de son contenu

Pour créer le répertoire *TestDir* vous pouvez exécuter la commande suivante

```
[cloudrea@quikstart~]$ hadoop fs -mkdir myTestDir
```

Exécutez de nouveau la commande **ls** pour voir le sous-répertoire *myTestDir*

```
[cloudrea@quikstart~]$ hadoop fs -ls
```

Ecosystème Hadoop

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ hadoop fs -mkdir myTestDir  
[cloudera@quickstart ~]$ hadoop fs -ls  
Found 6 items  
drwxr-xr-x - cloudera cloudera 0 2016-03-19 16:17 .Trash  
drwxr-xr-x - cloudera cloudera 0 2016-03-10 14:31 .sparkStaging  
drwxr-xr-x - cloudera cloudera 0 2016-03-27 05:58 FCAMapReduce  
drwxr-xr-x - cloudera cloudera 0 2016-09-21 04:56 myTestDir  
drwxr-xr-x - cloudera cloudera 0 2016-03-19 18:22 output  
drwxr-xr-x - cloudera cloudera 0 2016-03-19 18:16 wordcount  
[cloudera@quickstart ~]$
```

Pour utiliser les commandes HDFS récursive généralement vous ajoutez un "r" à la commande HDFS (Dans le Linux shell ce qui est généralement avec l'argument "-R").

Par exemple, pour faire une liste récursive, nous allons utiliser les **-ls -R** commande plutôt que **-ls** juste, comme les exemples ci-dessous:

```
[cloudera@quickstart~]$ hadoop fs -ls  
[cloudera@quickstart~]$ hadoop fs -ls -R
```

Or

```
[cloudera@quickstart~]$ hadoop fs -lsr
```

Vous pouvez diriger (en utilisant le caractère `|`) toute commande HDFS pour être utilisé avec le shell Linux. Par exemple, vous pouvez facilement utiliser `grep` avec HDFS en procédant comme suit :

```
[cloudera@quickstart~]$ hadoop fs -mkdir /user/cloudera/myTestDir2  
[cloudera@quickstart~]$ hadoop fs -ls /user/cloudera | grep Test
```

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ hadoop fs -mkdir /user/cloudera/myTestDir2  
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera | grep Test  
drwxr-xr-x - cloudera cloudera 0 2016-09-21 04:56 /user/cloudera/myTe  
stDir  
drwxr-xr-x - cloudera cloudera 0 2016-09-21 05:20 /user/cloudera/myTe  
stDir2  
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera | grep Test  
drwxr-xr-x - cloudera cloudera 0 2016-09-21 04:56 /user/cloudera/myTestDir  
drwxr-xr-x - cloudera cloudera 0 2016-09-21 05:20 /user/cloudera/myTestDir2  
[cloudera@quickstart ~]$
```

4.3. Transfert de données vers HDFS

Pour déplacer des fichiers entre votre système de fichiers Linux régulier et HDFS, vous pouvez utiliser la commande **put**. Par exemple, déplacer le fichier texte *README.txt* vers le système de fichiers Hadoop.

```
[cloudera@quickstart~]$ hadoop fs -put /home/cloudera/ReadMe.txt README.txt  
[cloudera@quickstart~]$ hadoop fs -ls /user/cloudera
```

Ecosystème Hadoop

```
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera
Found 7 items
drwxr-xr-x - cloudera cloudera 0 2016-09-21 05:14 /user/cloudera/.Trash
drwxr-xr-x - cloudera cloudera 0 2016-03-10 14:31 /user/cloudera/.sparkStaging
-rw-r--r-- 1 cloudera cloudera 10 2016-09-21 05:27 /user/cloudera/README
drwxr-xr-x - cloudera cloudera 0 2016-09-21 04:56 /user/cloudera/myTestDir
drwxr-xr-x - cloudera cloudera 0 2016-09-21 05:20 /user/cloudera/myTestDir2
drwxr-xr-x - cloudera cloudera 0 2016-03-19 18:22 /user/cloudera/output
drwxr-xr-x - cloudera cloudera 0 2016-03-19 18:16 /user/cloudera/wordcount
[cloudera@quickstart ~]$
```

Vous devriez maintenant voir un nouveau fichier appelé `/user/cloudera/README.txt` répertorié comme indiqué ci-dessus. Notez qu'il est un «1» mis en évidence dans la figure. Cela représente le nombre de réplcation de ce fichier dans le HDFS.

Pour afficher le contenu de ce fichier utilisez la commande `-cat` comme suit:

```
[cloudera@quickstart~]$ hadoop fs -put /home/cloudera/ReadMe.txt README.txt
[cloudera@quickstart~]$ hadoop fs -ls /user/cloudera
```

Pour afficher la taille du fichier `README`, utilisez la commande suivante :

```
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hadoop fs -cat README
Cloudera provides a scalable, flexible, integrated platform that makes it easy to
Cloudera products and solutions enable you to deploy and manage Apache Hadoop and rel
ected.
Cloudera provides the following products and tools:
CDH-The Cloudera distribution of Apache Hadoop and other related open-source pro
integration with numerous hardware and software solutions.
Cloudera Impala-A massively parallel processing SQL engine for interactive analy
suited for traditional BI-style queries with joins, aggregations, and subqueries
d by MapReduce jobs or loaded into Hive tables. The YARN resource management com
pala SQL queries. You can manage Impala alongside other Hadoop components throug
rization framework.
Cloudera Search-Provides near real-time access to data stored in or ingested int
ext exploration and navigated drill-down, as well as a simple, full-text interfa
ing platform. Search uses the flexible, scalable, and robust storage system incl
res to perform business tasks.
Cloudera Manager-A sophisticated application used to deploy, manage, monitor, an
nsole, a web-based user interface that makes administration of your enterprise d
u can use to obtain cluster health information and metrics, as well as configure
Cloudera Navigator-An end-to-end data management and security tool for the CDH p
explore the large amounts of data in Hadoop, and simplifies the storage and mana
, lifecycle management, and encryption key management in Cloudera Navigator allo
[cloudera@quickstart ~]$
```

```
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hadoop fs -du README
2720 2720 README
[cloudera@quickstart ~]$
```

```
[cloudera@quickstart~]$ hadoop fs -put /home/cloudera/ReadMe.txt README
```

Pour trouver la taille de tous les fichiers individuellement dans le répertoire `user/cloudera` utilisez la commande suivante:

```
[cloudera@quickstart~]$ hadoop fs -du user/cloudera/
```

Ecosystème Hadoop

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ hadoop fs -du user/cloudera/  
du: `user/cloudera/': No such file or directory  
[cloudera@quickstart ~]$ hadoop fs -du /user/cloudera/  
10          10          /user/cloudera/.Trash  
103968900   103968900   /user/cloudera/.sparkStaging  
2720        2720        /user/cloudera/README  
0           0           /user/cloudera/myTestDir  
0           0           /user/cloudera/myTestDir2  
0           0           /user/cloudera/output  
0           0           /user/cloudera/wordcount  
[cloudera@quickstart ~]$
```

Si vous souhaitez obtenir plus d'informations sur les commandes fs Hadoop, invoquer -help comme suit :

```
[cloudera@quickstart~]$ hadoop fs -help
```

4.4. Ingestion des données structurées à partir de web le cas de COVID-19

Pour développer cette application, nous avons besoins de charger, transformer et nettoyer des données prévenant des sources de données disponibles et les stocker sur HDFS. Nous utilisons Apache Pig pour effectuer la transformation des données et développer une Big ETL. Source de données et préparation Les informations COVID-19 les plus fiables proviennent de l'OMS, le Johns Hopkins University Center a commencé à collecter des données auprès de l'OMS et d'autres sources fiables. Ils ont utilisé les données pour créer de très bons tableaux de bord [2] et ont également rendu les ensembles de données disponibles sur un référentiel GitHub [3]. Dans le référentiel, il y a les jeux de données quotidiens bruts et également un rapport de séries chronologiques prétraité. Ce projet utilise les jeux de données quotidiens bruts fournis pour avoir plus de flexibilité pour transformer les données. Pour préparer les données, automatiser la collecte des données et effectuer la transformation des données, nous respectons les différentes étapes suivantes :

4.4.1. Téléchargement des données à partir du web

Lancer le Shell sous Cloudera et supprimer l'ancien répertoire **COVID-19** s'il existe.

```
[cloudera@quickstart~]$ rm -rf COVID-19
```

Cloner le projet en utilisant la commande clone de git

```
[cloudera@quickstart~]$ git clone https://github.com/CSSEGISandData/COVID-19.git
```

Accéder au répertoire Covid-19

```
[cloudera@quickstart~]$ cd COVID-19
```

En cas d'erreur nous utilisons la deuxième solution avec **wget** dont il faut l'installer en utilisant la commande suivante :

Ecosystème Hadoop



Ecosystème Hadoop

Télécharger un fichier zip contenant les sources de données

Décompresser le fichier zip et accéder au contenu de répertoire.

Après avoir téléchargé l'ensemble de données, en tant que fichiers au format CSV à partir de la source d'origine, exécuter les commandes Shell qui exécute quelques petites modifications dans les fichiers CSV d'origine, copie les fichiers dans un répertoire HDFS et suppression des doubles quotes dans quelques champs dans les fichiers csv.

Remplacer les dates du fichier d'origine pour correspondre aux dates incluses dans les noms de fichier.

Déplacer les fichiers dans différent répertoires en fonction des date et formats

Chercher les fichiers ayant comme nom inferieur à la date "2020-03-21" et les déplacer dans ce répertoire

Créer un deuxième répertoire format2



Ecosystème Hadoop

```
[cloudrea@quikstart~]$ mkdir format2
```

Chercher les fichiers ayant comme nom supérieur à la date "2020-03-21" et les déplacer dans ce répertoire

```
[root@quikstart COVID-19-master]# find csse_covid_19_data/csse_covid_19_daily_reports/ -maxdepth 1 -newermt "2020-03-21" -exec b  
asename \{} .po \; | grep csv | sort | xargs -I % mv csse_covid_19_data/csse_covid_19_daily_reports/% format2/  
[root@quikstart COVID-19-master]#
```

```
cloudrea@quikstart~]$ find COVID-  
19/csse_covid_19_data/csse_covid_19_daily_reports/ -maxdepth 1 -newermt  
"2020-03-21" -exec basename \{} .po \; | grep csv | sort | xargs -I % mv  
/csse_covid_19_data/csse_covid_19_daily_reports/% format2/
```

Lister le contenu de répertoire format1, il faut avoir ce résultat.

```
[cloudrea@quikstart~]$ cd format1
```

```
[cloudrea@quikstart~]$ ls
```

```
[root@quikstart format1]# ls  
01-22-2020.csv 01-30-2020.csv 02-07-2020.csv 02-15-2020.csv 02-23-2020.csv 03-02-2020.csv 03-10-2020.csv 03-18-2020.csv  
01-23-2020.csv 01-31-2020.csv 02-08-2020.csv 02-16-2020.csv 02-24-2020.csv 03-03-2020.csv 03-11-2020.csv 03-19-2020.csv  
01-24-2020.csv 02-01-2020.csv 02-09-2020.csv 02-17-2020.csv 02-25-2020.csv 03-04-2020.csv 03-12-2020.csv 03-20-2020.csv  
01-25-2020.csv 02-02-2020.csv 02-10-2020.csv 02-18-2020.csv 02-26-2020.csv 03-05-2020.csv 03-13-2020.csv 03-21-2020.csv  
01-26-2020.csv 02-03-2020.csv 02-11-2020.csv 02-19-2020.csv 02-27-2020.csv 03-06-2020.csv 03-14-2020.csv  
01-27-2020.csv 02-04-2020.csv 02-12-2020.csv 02-20-2020.csv 02-28-2020.csv 03-07-2020.csv 03-15-2020.csv  
01-28-2020.csv 02-05-2020.csv 02-13-2020.csv 02-21-2020.csv 02-29-2020.csv 03-08-2020.csv 03-16-2020.csv  
01-29-2020.csv 02-06-2020.csv 02-14-2020.csv 02-22-2020.csv 03-01-2020.csv 03-09-2020.csv 03-17-2020.csv  
[root@quikstart format1]#
```

Lister le contenu de répertoire format1, il faut avoir ce résultat.

```
cloudrea@quikstart~]$ cd format2
```

```
[cloudrea@quikstart~]$ ls
```

```
[root@quikstart COVID-19-master]# find csse_covid_19_data/csse_covid_19_daily_reports/ -maxdepth 1 -newermt "2020-03-21" -exec b  
asename \{} .po \; | grep csv | sort | xargs -I % mv csse_covid_19_data/csse_covid_19_daily_reports/% format2/  
[root@quikstart COVID-19-master]# cd ..  
[root@quikstart /]# cd COVID-19-master/  
[root@quikstart COVID-19-master]# ls  
archived_data  csse_covid_19_data  format1  format2  fromat1  README.md  who_covid_19_situation_reports  
[root@quikstart COVID-19-master]# cd format2  
[root@quikstart format2]# ls  
03-22-2020.csv 04-21-2020.csv 05-21-2020.csv 06-20-2020.csv 07-20-2020.csv 08-19-2020.csv 09-18-2020.csv 10-18-2020.csv  
03-23-2020.csv 04-22-2020.csv 05-22-2020.csv 06-21-2020.csv 07-21-2020.csv 08-20-2020.csv 09-19-2020.csv 10-19-2020.csv  
03-24-2020.csv 04-23-2020.csv 05-23-2020.csv 06-22-2020.csv 07-22-2020.csv 08-21-2020.csv 09-20-2020.csv 10-20-2020.csv  
03-25-2020.csv 04-24-2020.csv 05-24-2020.csv 06-23-2020.csv 07-23-2020.csv 08-22-2020.csv 09-21-2020.csv 10-21-2020.csv  
03-26-2020.csv 04-25-2020.csv 05-25-2020.csv 06-24-2020.csv 07-24-2020.csv 08-23-2020.csv 09-22-2020.csv 10-22-2020.csv  
03-27-2020.csv 04-26-2020.csv 05-26-2020.csv 06-25-2020.csv 07-25-2020.csv 08-24-2020.csv 09-23-2020.csv 10-23-2020.csv  
03-28-2020.csv 04-27-2020.csv 05-27-2020.csv 06-26-2020.csv 07-26-2020.csv 08-25-2020.csv 09-24-2020.csv 10-24-2020.csv  
03-29-2020.csv 04-28-2020.csv 05-28-2020.csv 06-27-2020.csv 07-27-2020.csv 08-26-2020.csv 09-25-2020.csv 10-25-2020.csv  
03-30-2020.csv 04-29-2020.csv 05-29-2020.csv 06-28-2020.csv 07-28-2020.csv 08-27-2020.csv 09-26-2020.csv 10-26-2020.csv  
03-31-2020.csv 04-30-2020.csv 05-30-2020.csv 06-29-2020.csv 07-29-2020.csv 08-28-2020.csv 09-27-2020.csv 10-27-2020.csv  
04-01-2020.csv 05-01-2020.csv 05-31-2020.csv 06-30-2020.csv 07-30-2020.csv 08-29-2020.csv 09-28-2020.csv 10-28-2020.csv  
04-02-2020.csv 05-02-2020.csv 06-01-2020.csv 07-01-2020.csv 07-31-2020.csv 08-30-2020.csv 09-29-2020.csv 10-29-2020.csv  
04-03-2020.csv 05-03-2020.csv 06-02-2020.csv 07-02-2020.csv 08-01-2020.csv 08-31-2020.csv 09-30-2020.csv 10-30-2020.csv  
04-04-2020.csv 05-04-2020.csv 06-03-2020.csv 07-03-2020.csv 08-02-2020.csv 09-01-2020.csv 10-01-2020.csv 10-31-2020.csv  
04-05-2020.csv 05-05-2020.csv 06-04-2020.csv 07-04-2020.csv 08-03-2020.csv 09-02-2020.csv 10-02-2020.csv 11-01-2020.csv  
04-06-2020.csv 05-06-2020.csv 06-05-2020.csv 07-05-2020.csv 08-04-2020.csv 09-03-2020.csv 10-03-2020.csv 11-02-2020.csv  
04-07-2020.csv 05-07-2020.csv 06-06-2020.csv 07-06-2020.csv 08-05-2020.csv 09-04-2020.csv 10-04-2020.csv 11-03-2020.csv  
04-08-2020.csv 05-08-2020.csv 06-07-2020.csv 07-07-2020.csv 08-06-2020.csv 09-05-2020.csv 10-05-2020.csv 11-04-2020.csv  
04-09-2020.csv 05-09-2020.csv 06-08-2020.csv 07-08-2020.csv 08-07-2020.csv 09-06-2020.csv 10-06-2020.csv 11-05-2020.csv  
04-10-2020.csv 05-10-2020.csv 06-09-2020.csv 07-09-2020.csv 08-08-2020.csv 09-07-2020.csv 10-07-2020.csv 11-06-2020.csv  
04-11-2020.csv 05-11-2020.csv 06-10-2020.csv 07-10-2020.csv 08-09-2020.csv 09-08-2020.csv 10-08-2020.csv 11-07-2020.csv  
04-12-2020.csv 05-12-2020.csv 06-11-2020.csv 07-11-2020.csv 08-10-2020.csv 09-09-2020.csv 10-09-2020.csv 11-08-2020.csv  
04-13-2020.csv 05-13-2020.csv 06-12-2020.csv 07-12-2020.csv 08-11-2020.csv 09-10-2020.csv 10-10-2020.csv 11-09-2020.csv  
04-14-2020.csv 05-14-2020.csv 06-13-2020.csv 07-13-2020.csv 08-12-2020.csv 09-11-2020.csv 10-11-2020.csv 11-10-2020.csv  
04-15-2020.csv 05-15-2020.csv 06-14-2020.csv 07-14-2020.csv 08-13-2020.csv 09-12-2020.csv 10-12-2020.csv  
04-16-2020.csv 05-16-2020.csv 06-15-2020.csv 07-15-2020.csv 08-14-2020.csv 09-13-2020.csv 10-13-2020.csv  
04-17-2020.csv 05-17-2020.csv 06-16-2020.csv 07-16-2020.csv 08-15-2020.csv 09-14-2020.csv 10-14-2020.csv  
04-18-2020.csv 05-18-2020.csv 06-17-2020.csv 07-17-2020.csv 08-16-2020.csv 09-15-2020.csv 10-15-2020.csv  
04-19-2020.csv 05-19-2020.csv 06-18-2020.csv 07-18-2020.csv 08-17-2020.csv 09-16-2020.csv 10-16-2020.csv  
04-20-2020.csv 05-20-2020.csv 06-19-2020.csv 07-19-2020.csv 08-18-2020.csv 09-17-2020.csv 10-17-2020.csv  
[root@quikstart format2]#
```


Ecosystème Hadoop

4.4.2. Créer les répertoires HDFS nécessaires et Ingestion de données

Créer un premier répertoire hdfs nommée /covid/format1/

```
cloudrea@quikstart~]$ hadoop fs -mkdir -p /user/cloudera/covid/format1
```

Créer un deuxième répertoire hdfs nommée /covid/format2/

```
cloudrea@quikstart~]$ hadoop fs -mkdir -p /user/cloudera/covid/format2
```

Copier tous les fichiers locaux vers le premier répertoire hdfs /covid/format2/

```
cloudrea@quikstart~]$ hadoop fs -put format1/*  
/user/cloudera/covid19/format1  
  
cloudrea@quikstart~]$ hadoop fs -copyFromLocal format1/*  
/user/cloudera/covid/format1
```

Copier tous les fichiers locaux vers le deuxième répertoire hdfs /covid/format2/

```
cloudrea@quikstart~]$ hadoop fs -copyFromLocal format2/*  
/user/cloudera/covid/format2
```

Vérifier de l'ingestion des données vers HDFS en tapant les commandes suivantes

```
cloudrea@quikstart~]$ hadoop fs -ls /user/cloudera/covid/format1
```

```
[root@quikstart COVID-19-master]# hadoop fs -ls /user/cloudera/covid/format2  
Found 234 items  
-rw-r--r-- 1 root cloudera 328044 2020-11-12 18:42 /user/cloudera/covid/format2/03-22-2020.csv  
-rw-r--r-- 1 root cloudera 348883 2020-11-12 18:42 /user/cloudera/covid/format2/03-23-2020.csv  
-rw-r--r-- 1 root cloudera 349057 2020-11-12 18:42 /user/cloudera/covid/format2/03-24-2020.csv  
-rw-r--r-- 1 root cloudera 349499 2020-11-12 18:42 /user/cloudera/covid/format2/03-25-2020.csv  
-rw-r--r-- 1 root cloudera 349724 2020-11-12 18:42 /user/cloudera/covid/format2/03-26-2020.csv  
-rw-r--r-- 1 root cloudera 350496 2020-11-12 18:42 /user/cloudera/covid/format2/03-27-2020.csv  
-rw-r--r-- 1 root cloudera 329477 2020-11-12 18:42 /user/cloudera/covid/format2/03-28-2020.csv  
-rw-r--r-- 1 root cloudera 329924 2020-11-12 18:42 /user/cloudera/covid/format2/03-29-2020.csv  
-rw-r--r-- 1 root cloudera 330453 2020-11-12 18:42 /user/cloudera/covid/format2/03-30-2020.csv  
-rw-r--r-- 1 root cloudera 248163 2020-11-12 18:42 /user/cloudera/covid/format2/03-31-2020.csv  
-rw-r--r-- 1 root cloudera 253369 2020-11-12 18:42 /user/cloudera/covid/format2/04-01-2020.csv  
-rw-r--r-- 1 root cloudera 243956 2020-11-12 18:42 /user/cloudera/covid/format2/04-02-2020.csv  
-rw-r--r-- 1 root cloudera 268036 2020-11-12 18:42 /user/cloudera/covid/format2/04-03-2020.csv  
-rw-r--r-- 1 root cloudera 254724 2020-11-12 18:42 /user/cloudera/covid/format2/04-04-2020.csv  
-rw-r--r-- 1 root cloudera 282506 2020-11-12 18:42 /user/cloudera/covid/format2/04-05-2020.csv  
-rw-r--r-- 1 root cloudera 267360 2020-11-12 18:42 /user/cloudera/covid/format2/04-06-2020.csv  
-rw-r--r-- 1 root cloudera 292252 2020-11-12 18:42 /user/cloudera/covid/format2/04-07-2020.csv  
-rw-r--r-- 1 root cloudera 294973 2020-11-12 18:42 /user/cloudera/covid/format2/04-08-2020.csv  
-rw-r--r-- 1 root cloudera 298012 2020-11-12 18:42 /user/cloudera/covid/format2/04-09-2020.csv  
-rw-r--r-- 1 root cloudera 301216 2020-11-12 18:42 /user/cloudera/covid/format2/04-10-2020.csv  
-rw-r--r-- 1 root cloudera 303921 2020-11-12 18:42 /user/cloudera/covid/format2/04-11-2020.csv  
-rw-r--r-- 1 root cloudera 305548 2020-11-12 18:42 /user/cloudera/covid/format2/04-12-2020.csv  
-rw-r--r-- 1 root cloudera 309742 2020-11-12 18:42 /user/cloudera/covid/format2/04-13-2020.csv  
-rw-r--r-- 1 root cloudera 311068 2020-11-12 18:42 /user/cloudera/covid/format2/04-14-2020.csv  
-rw-r--r-- 1 root cloudera 312551 2020-11-12 18:42 /user/cloudera/covid/format2/04-15-2020.csv  
-rw-r--r-- 1 root cloudera 314226 2020-11-12 18:42 /user/cloudera/covid/format2/04-16-2020.csv  
-rw-r--r-- 1 root cloudera 314848 2020-11-12 18:42 /user/cloudera/covid/format2/04-17-2020.csv  
-rw-r--r-- 1 root cloudera 315926 2020-11-12 18:42 /user/cloudera/covid/format2/04-18-2020.csv  
-rw-r--r-- 1 root cloudera 317954 2020-11-12 18:42 /user/cloudera/covid/format2/04-19-2020.csv  
-rw-r--r-- 1 root cloudera 317177 2020-11-12 18:42 /user/cloudera/covid/format2/04-20-2020.csv  
-rw-r--r-- 1 root cloudera 318374 2020-11-12 18:42 /user/cloudera/covid/format2/04-21-2020.csv  
-rw-r--r-- 1 root cloudera 319302 2020-11-12 18:42 /user/cloudera/covid/format2/04-22-2020.csv  
-rw-r--r-- 1 root cloudera 321448 2020-11-12 18:42 /user/cloudera/covid/format2/04-23-2020.csv  
-rw-r--r-- 1 root cloudera 322416 2020-11-12 18:42 /user/cloudera/covid/format2/04-24-2020.csv  
-rw-r--r-- 1 root cloudera 323140 2020-11-12 18:42 /user/cloudera/covid/format2/04-25-2020.csv  
-rw-r--r-- 1 root cloudera 324105 2020-11-12 18:42 /user/cloudera/covid/format2/04-26-2020.csv  
-rw-r--r-- 1 root cloudera 325150 2020-11-12 18:42 /user/cloudera/covid/format2/04-27-2020.csv
```

```
cloudrea@quikstart~]$ hadoop fs -ls /user/cloudera/covid/format2
```

Ecosystème Hadoop

```
[root@quickstart COVID-19-master]# hadoop fs -ls /user/cloudera/covid/format1
Found 60 items
-rw-r--r-- 1 root cloudera 1675 2020-11-12 18:41 /user/cloudera/covid/format1/01-22-2020.csv
-rw-r--r-- 1 root cloudera 1832 2020-11-12 18:41 /user/cloudera/covid/format1/01-23-2020.csv
-rw-r--r-- 1 root cloudera 1695 2020-11-12 18:41 /user/cloudera/covid/format1/01-24-2020.csv
-rw-r--r-- 1 root cloudera 1790 2020-11-12 18:41 /user/cloudera/covid/format1/01-25-2020.csv
-rw-r--r-- 1 root cloudera 1896 2020-11-12 18:41 /user/cloudera/covid/format1/01-26-2020.csv
-rw-r--r-- 1 root cloudera 2049 2020-11-12 18:41 /user/cloudera/covid/format1/01-27-2020.csv
-rw-r--r-- 1 root cloudera 2102 2020-11-12 18:41 /user/cloudera/covid/format1/01-28-2020.csv
-rw-r--r-- 1 root cloudera 2184 2020-11-12 18:41 /user/cloudera/covid/format1/01-29-2020.csv
-rw-r--r-- 1 root cloudera 2334 2020-11-12 18:41 /user/cloudera/covid/format1/01-30-2020.csv
-rw-r--r-- 1 root cloudera 2569 2020-11-12 18:41 /user/cloudera/covid/format1/01-31-2020.csv
-rw-r--r-- 1 root cloudera 2785 2020-11-12 18:41 /user/cloudera/covid/format1/02-01-2020.csv
-rw-r--r-- 1 root cloudera 3151 2020-11-12 18:41 /user/cloudera/covid/format1/02-02-2020.csv
-rw-r--r-- 1 root cloudera 3201 2020-11-12 18:41 /user/cloudera/covid/format1/02-03-2020.csv
-rw-r--r-- 1 root cloudera 3295 2020-11-12 18:41 /user/cloudera/covid/format1/02-04-2020.csv
-rw-r--r-- 1 root cloudera 3341 2020-11-12 18:41 /user/cloudera/covid/format1/02-05-2020.csv
-rw-r--r-- 1 root cloudera 3345 2020-11-12 18:41 /user/cloudera/covid/format1/02-06-2020.csv
-rw-r--r-- 1 root cloudera 3402 2020-11-12 18:41 /user/cloudera/covid/format1/02-07-2020.csv
-rw-r--r-- 1 root cloudera 3409 2020-11-12 18:41 /user/cloudera/covid/format1/02-08-2020.csv
-rw-r--r-- 1 root cloudera 3429 2020-11-12 18:41 /user/cloudera/covid/format1/02-09-2020.csv
-rw-r--r-- 1 root cloudera 3433 2020-11-12 18:41 /user/cloudera/covid/format1/02-10-2020.csv
-rw-r--r-- 1 root cloudera 3490 2020-11-12 18:41 /user/cloudera/covid/format1/02-11-2020.csv
-rw-r--r-- 1 root cloudera 3493 2020-11-12 18:41 /user/cloudera/covid/format1/02-12-2020.csv
-rw-r--r-- 1 root cloudera 3545 2020-11-12 18:41 /user/cloudera/covid/format1/02-13-2020.csv
-rw-r--r-- 1 root cloudera 3504 2020-11-12 18:41 /user/cloudera/covid/format1/02-14-2020.csv
-rw-r--r-- 1 root cloudera 3509 2020-11-12 18:41 /user/cloudera/covid/format1/02-15-2020.csv
-rw-r--r-- 1 root cloudera 3510 2020-11-12 18:41 /user/cloudera/covid/format1/02-16-2020.csv
-rw-r--r-- 1 root cloudera 3585 2020-11-12 18:41 /user/cloudera/covid/format1/02-17-2020.csv
-rw-r--r-- 1 root cloudera 3588 2020-11-12 18:41 /user/cloudera/covid/format1/02-18-2020.csv
-rw-r--r-- 1 root cloudera 3550 2020-11-12 18:41 /user/cloudera/covid/format1/02-19-2020.csv
-rw-r--r-- 1 root cloudera 3551 2020-11-12 18:41 /user/cloudera/covid/format1/02-20-2020.csv
-rw-r--r-- 1 root cloudera 3941 2020-11-12 18:41 /user/cloudera/covid/format1/02-21-2020.csv
-rw-r--r-- 1 root cloudera 4011 2020-11-12 18:41 /user/cloudera/covid/format1/02-22-2020.csv
-rw-r--r-- 1 root cloudera 4050 2020-11-12 18:41 /user/cloudera/covid/format1/02-23-2020.csv
-rw-r--r-- 1 root cloudera 4261 2020-11-12 18:41 /user/cloudera/covid/format1/02-24-2020.csv
-rw-r--r-- 1 root cloudera 4415 2020-11-12 18:41 /user/cloudera/covid/format1/02-25-2020.csv
-rw-r--r-- 1 root cloudera 4647 2020-11-12 18:41 /user/cloudera/covid/format1/02-26-2020.csv
-rw-r--r-- 1 root cloudera 4796 2020-11-12 18:41 /user/cloudera/covid/format1/02-27-2020.csv
```

Afficher le contenu d'un fichier sous HDFS avec la commande cat

```
cloudrea@quikstart~]$ hadoop fs -cat /user/cloudera/covid/format2/11-10-2020.csv
```

```
[root@quickstart COVID-19-master]# hadoop fs -cat /user/cloudera/covid/format2/11-10-2020.csv
FIPS,Admin2,Province_State,Country_Region,Last_Update,Lat,Long,Confirmed,Deaths,Recovered,Active,Combined_Key,Incident_Rate,Case_Fatality_Ratio
,,,Afghanistan,2020-11-11 05:25:30,33.93911,67.709953,42463,1577,34954,5932,Afghanistan,109.07991172806463,3.7138214445517272
,,,Albania,2020-11-11 05:25:30,41.1533,20.1683,25294,579,12353,12362,Albania,878.9352977969282,2.2890804143275085
,,,Algeria,2020-11-11 05:25:30,28.0339,1.6596,63446,2077,42626,18743,Algeria,144.6852700858221,3.2736500330990133
,,,Andorra,2020-11-11 05:25:30,42.5063,1.5218,5477,75,4405,997,Andorra,7088.5912120623825,1.3693627898484573
,,,Angola,2020-11-11 05:25:30,-11.2027,17.8739,12816,308,6036,6472,Angola,38.99438780210762,2.403245942571785
,,,Antigua and Barbuda,2020-11-11 05:25:30,17.0608,-61.7964,131,3,122,6,Antigua and Barbuda,133.77175067396453,2.2900763358778624
,,,Argentina,2020-11-11 05:25:30,-38.4161,-63.6167,1262476,34183,1081897,146396,Argentina,2793.349475991972,2.707615827944452
,,,Armenia,2020-11-11 05:25:30,40.0691,45.0382,108687,1609,66835,40243,Armenia,3667.850733354167,1.4803978396680375
,,,Australian Capital Territory,Australia,2020-11-11 05:25:30,-35.4735,149.0124,114,3,111,0,"Australian Capital Territory Australia",26.62929221443592,2.6315789473684212
,,,New South Wales,Australia,2020-11-11 05:25:30,-33.8688,151.2093,4469,53,3156,1260,"New South Wales Australia",55.05050505050505
45,1.1859476392929067
,,,Northern Territory,Australia,2020-11-11 05:25:30,-12.4634,130.8456,41,0,33,8,"Northern Territory Australia",16.693811074918568
,0.0
,,,Queensland,Australia,2020-11-11 05:25:30,-27.4698,153.0251,1179,6,1163,10,"Queensland Australia",23.04760043006549,0.508905852
4173028
,,,South Australia,Australia,2020-11-11 05:25:30,-34.9285,138.6007,517,4,495,18,"South Australia Australia",29.433532593225163,0.
7736943907156673
,,,Tasmania,Australia,2020-11-11 05:25:30,-42.8821,147.3272,230,13,217,0,"Tasmania Australia",42.95051353874883,5.652173913043478
5
,,,Victoria,Australia,2020-11-11 05:25:30,-37.8136,144.9631,20345,819,19522,4,"Victoria Australia",306.8673735652121,4.0255591054
3131
,,,Western Australia,Australia,2020-11-11 05:25:30,-31.9505,115.8605,776,9,757,10,"Western Australia Australia",29.49897361818596
4,1.1597938144329898
,,,Austria,2020-11-11 05:25:30,47.5162,14.5501,164866,1499,98663,64704,Austria,1830.54272517321,0.9092232479710796
,,,Azerbaijan,2020-11-11 05:25:30,40.1431,47.5769,67392,867,50009,16516,Azerbaijan,664.669462752147,1.286502849002849
,,,Bahamas,2020-11-11 05:25:30,25.025885,-78.035889,7012,154,5035,1823,Bahamas,1783.0987061599806,2.1962350256702794
,,,Bahrain,2020-11-11 05:25:30,26.0275,50.55,83811,331,81415,2065,Bahrain,4925.472339580261,0.39493622555511804
,,,Bangladesh,2020-11-11 05:25:30,23.685,90.3563,423620,6108,341416,76096,Bangladesh,257.2236244275686,1.4418582692035316
,,,Barbados,2020-11-11 05:25:30,13.1939,-59.5432,243,7,231,5,Barbados,84.55968069151028,2.880658436213992
,,,Belarus,2020-11-11 05:25:30,53.7098,27.9534,108300,1016,91646,15638,Belarus,1146.1140964520093,0.938134810710988
,,,Belgium,2020-11-11 05:25:30,50.8333,4.469936,507475,13561,30504,463410,Belgium,4378.704177946879,2.6722498645253463
,,,Belize,2020-11-11 05:25:30,17.1899,-88.4976,4414,73,2440,1901,Belize,1110.1023336292599,1.6538287267784322
,,,Benin,2020-11-11 05:25:30,9.3077,2.3158,2781,43,2515,223,Benin,22.93949170837596,1.5462064005753327
```

Comptez le nombre de répertoires, fichiers et octets sous les chemins qui correspondent au modèle de fichier spécifié.

```
cloudrea@quikstart~]$ hadoop fs -count /user/cloudera/covid
```

```
[root@quickstart COVID-19-master]# hadoop fs -count /user/cloudera/covid
3          294          110589144 /user/cloudera/covid
```

Ecosystème Hadoop

Affiche la taille des fichiers et des répertoires contenus dans le répertoire `/user/cloudera/covid /` ou la longueur d'un fichier au cas où il ne s'agirait que d'un fichier.

```
[root@quickstart COVID-19-master]# hadoop fs -du /user/cloudera/covid
413664      413664      /user/cloudera/covid/format1
110175480  110175480  /user/cloudera/covid/format2
[root@quickstart COVID-19-master]#
```

4.5. Configuration des paramètres par défaut Hadoop

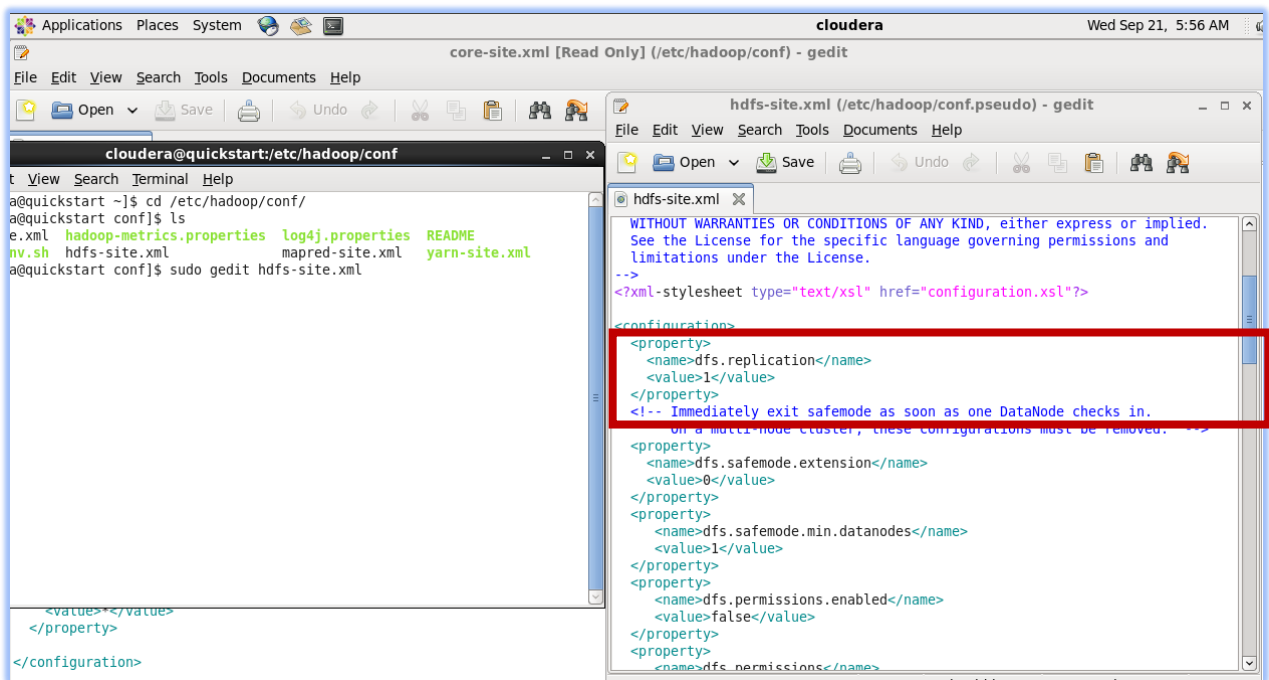
L'augmentation de la taille de bloc de stockage

Le fichier `/etc/hadoop/conf/hdfs-site.xml` contient les paramètres spécifiques au système de fichiers HDFS. Pour accéder à ce fichier pointez-vous sous le répertoire `/etc/hadoop/conf/`

```
[cloudrea@quikstart~]$ cd /etc/hadoop/conf/
```

Éditer le fichier avec la commande suivante :

```
[cloudrea@quikstart~]$ sudo nano hdfs-site.xml
```



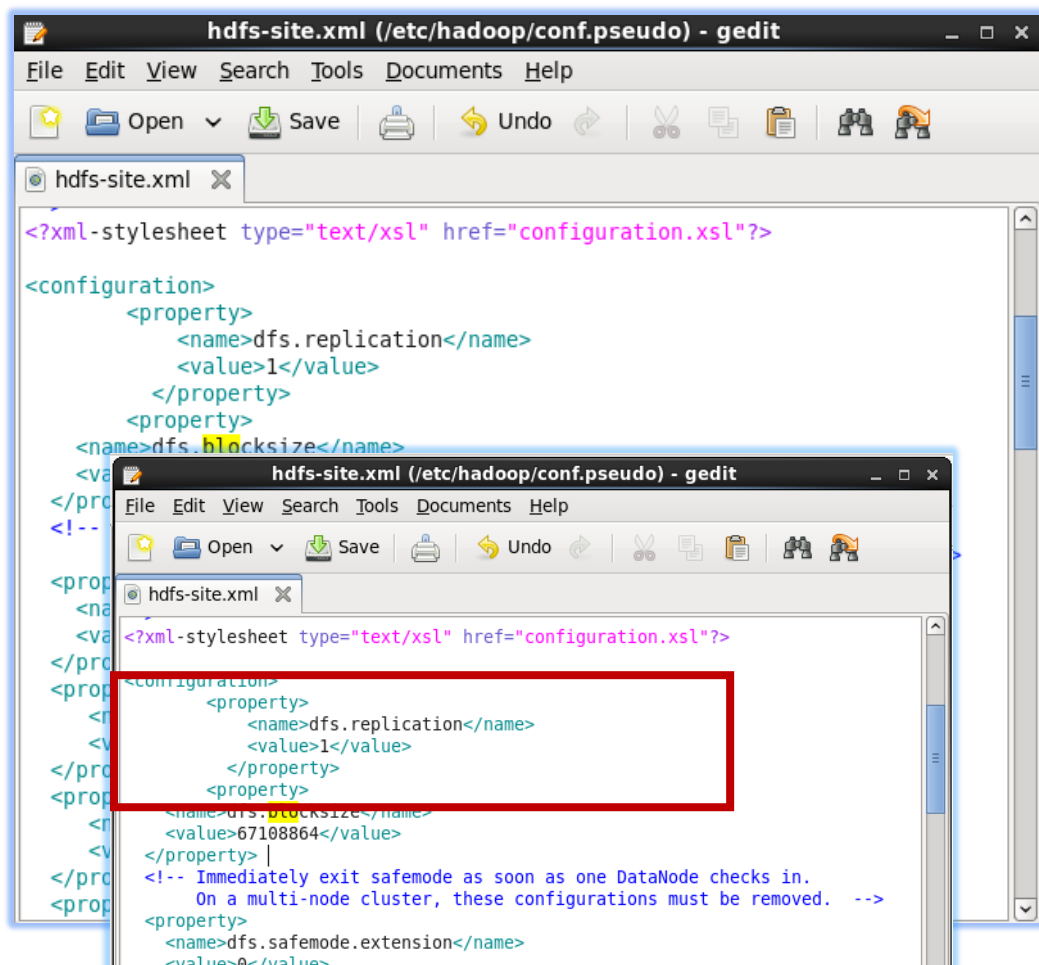
Ecosystème Hadoop

Ajouter les balises suivantes aux fichiers et enregistrer pour qu'il soit pris en considération

Configuration du facteur de réplication

La valeur actuelle du facteur de réplication par défaut est 1.

Vous pouvez remplacer la valeur par défaut en ajoutant les lignes suivantes dans ce fichier (hdfs-site.xml). La valeur sera le numéro de votre choix.



4.6. Commande de l'utilitaire de vérification du système de fichiers HDFS

Comme mentionné dans la documentation, la commande `hdfs fsck` est conçue pour signaler des problèmes avec divers fichiers, par exemple, des blocs manquants pour un fichier ou des blocs sous-répliqués. Mais nous pouvons également l'utiliser pour vérifier la distribution des données.

La commande `hdfs fsck` affiche des informations sur un chemin HDFS donné :

- Statut
- Taille totale
- Nombre de fichiers dans le référentiel
- Liste des blocs HDFS pour chaque fichier
- Facteur de réplication de chaque fichier

Ecosystème Hadoop

- Taille de chaque bloc HDFS
- Emplacement de chaque bloc HDFS

Taper la commande `hdfs fsck` comme ceci pour dumper le résultat dans un fichier `stats.txt`

```
hadoop fsck /user/cloudera/covid/ -files -blocks -locations > stats.txt
```

Taper la commande `hdfs fsck` comme ceci pour dumper le résultat sous la console

```
hadoop fsck /user/cloudera/covid/ -files -blocks -locations
```

Quelques explications des arguments :

- `-files`: imprimer les fichiers en cours de vérification
- `-blocks`: imprimer le rapport de bloc
- `-locations`: affiche les emplacements pour chaque bloc

L'argument `-racks` est également disponible, vous pouvez l'utiliser pour voir comment les données sont équilibrées dans les différents racks de votre cluster. J'ai choisi de ne pas inclure cet argument dans mon script car tout mon cluster fonctionne sur un seul rack.

Voici la sortie de la commande

```
[root@quickstart COVID-19-master]# hadoop fs -fsck /user/cloudera/covid/ -files -blocks -locations
-fsck: Unknown command
[root@quickstart COVID-19-master]# hadoop fsck /user/cloudera/covid/ -files -blocks -locations
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Connecting to namenode via http://quickstart.cloudera:50070
FSCK started by root (auth:SIMPLE) from /172.17.0.2 for path /user/cloudera/covid/ at Thu Nov 12 19:33:27 UTC 2020
/user/cloudera/covid/ <dir>
/user/cloudera/covid/format1/ <dir>
/user/cloudera/covid/format1/01-22-2020.csv 1675 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073742842_2018 len=1675 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-47c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format1/01-23-2020.csv 1832 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073742843_2019 len=1832 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-47c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format1/01-24-2020.csv 1695 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073742844_2020 len=1695 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-47c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format1/01-25-2020.csv 1790 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073742845_2021 len=1790 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-47c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format1/01-26-2020.csv 1896 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073742846_2022 len=1896 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-47c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format1/01-27-2020.csv 2049 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073742847_2023 len=2049 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-47c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format1/01-28-2020.csv 2102 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073742848_2024 len=2102 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-47c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format1/01-29-2020.csv 2184 bytes, 1 block(s): OK
```

En regardant la sortie, nous pouvons faire une analyse très simple. Le référentiel HDFS que nous avons utilisé comme exemple contient chaque fichier occupe un bloc puisque sa taille est inférieure à 64MB :

Ecosystème Hadoop

```
7c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format2/11-07-2020.csv 558164 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073743132_2308 len=558164 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-4
7c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format2/11-08-2020.csv 558458 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073743133_2309 len=558458 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-4
7c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format2/11-09-2020.csv 558365 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073743134_2310 len=558365 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-4
7c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

/user/cloudera/covid/format2/11-10-2020.csv 558561 bytes, 1 block(s): OK
0. BP-1120155954-10.0.0.1-1459909528739:blk_1073743135_2311 len=558561 Live_repl=1 [DatanodeInfoWithStorage[172.17.0.2:50010,DS-4
7c0146f-ef5a-4325-b9f1-3e6fadb9f897,DISK]]

Status: HEALTHY
Total size: 110589144 B
Total dirs: 3
Total files: 294
Total symlinks: 0
Total blocks (validated): 294 (avg. block size 376153 B)
Minimally replicated blocks: 294 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Thu Nov 12 19:46:59 UTC 2020 in 171 milliseconds

The filesystem under path '/user/cloudera/covid/' is HEALTHY
[root@quickstart COVID-19-master]#
```

```
Status: HEALTHY
Total size: 110589144 B
Total dirs: 3
Total files: 294
Total symlinks: 0
Total blocks (validated): 294 (avg. block size 376153 B)
Minimally replicated blocks: 294 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Thu Nov 12 19:46:59 UTC 2020 in 171 milliseconds

The filesystem under path '/user/cloudera/covid/' is HEALTHY
[root@quickstart COVID-19-master]#
```

Nombre de fichiers

Nombre de blocks

Facteur de réplication

5. Gestion de HDFS via le Console Web Hue

5.1. Lancement de la console Web Hue.

Lancer le navigateur web installé sous Cloudera.

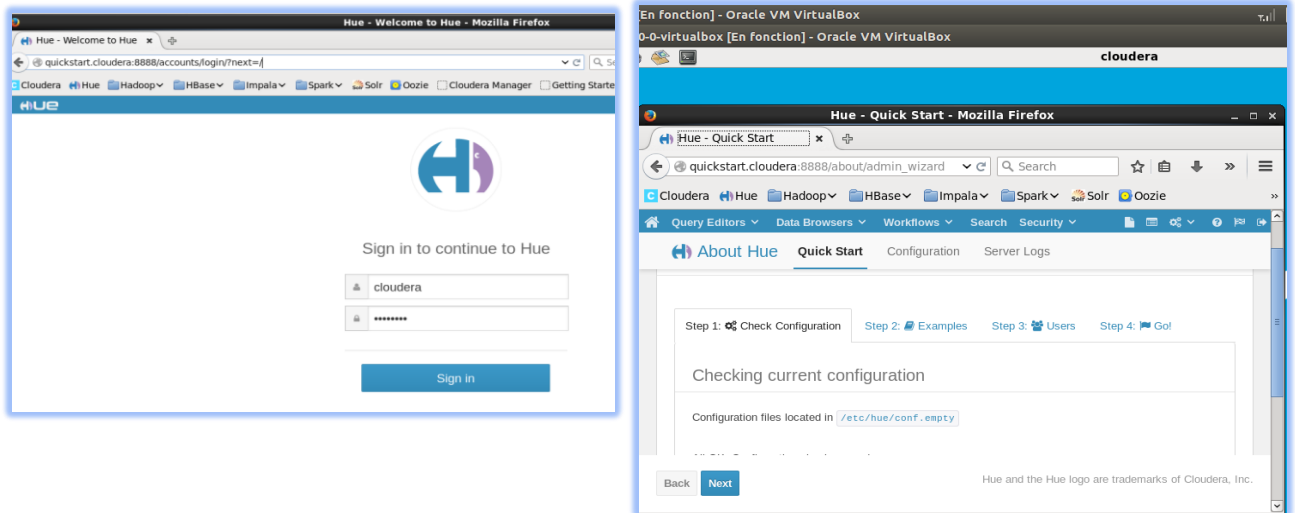
Taper l'adresse URL http://IP_AzureVM:8888/

Saisir les paramètres de connections

User : cloudera

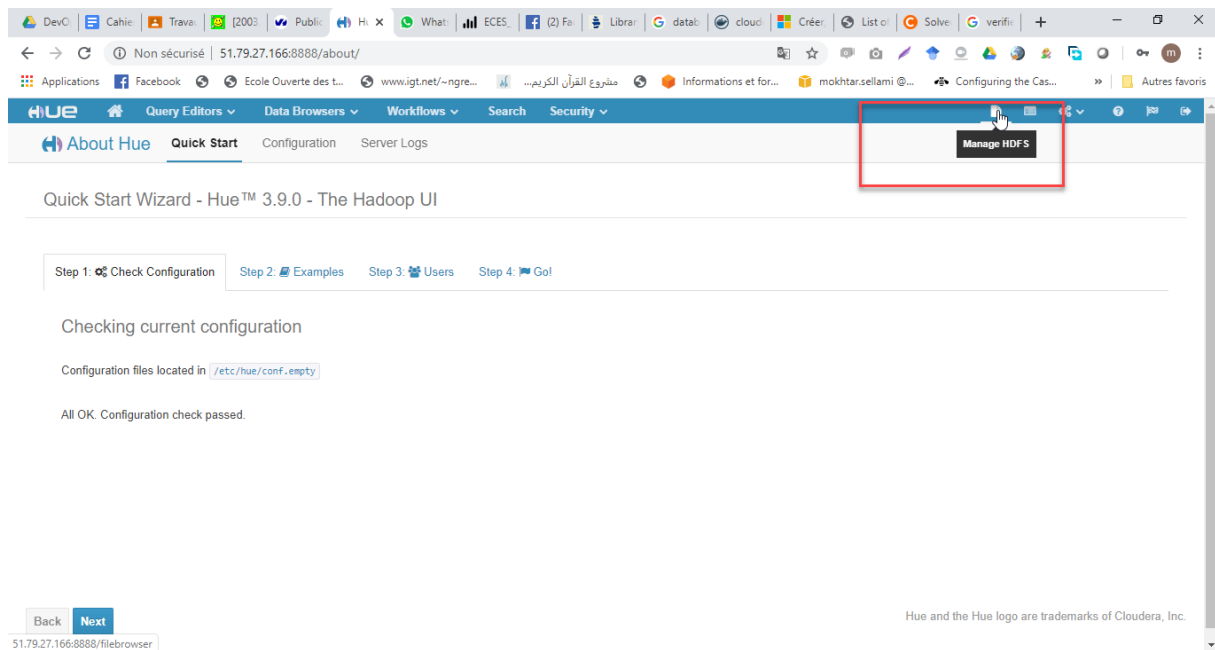
password :cloudera

Ecosystème Hadoop



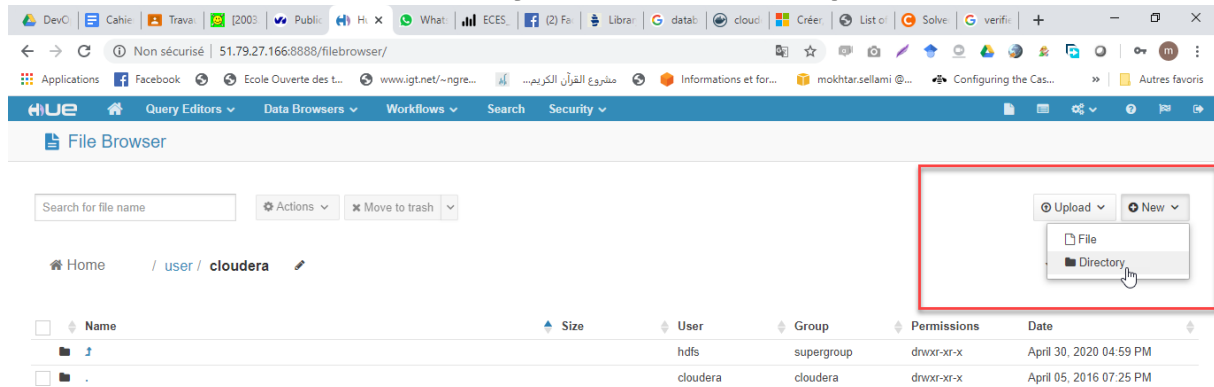
5.2. Manipulation des fichiers HDFS dans la console Web

Accéder maintenant au gestionnaire de fichier HDFS



Créer un répertoire HDFS « abdata »

Ecosystème Hadoop



Search for file name

Actions Move to trash

Home / user / cloudera

Name	Size	User	Group	Permissions	Date
.		hdfs	supergroup	drwxr-xr-x	April 30, 2020 04:59 PM
.		cloudera	cloudera	drwxr-xr-x	April 05, 2016 07:25 PM

Show 45 of 0 items

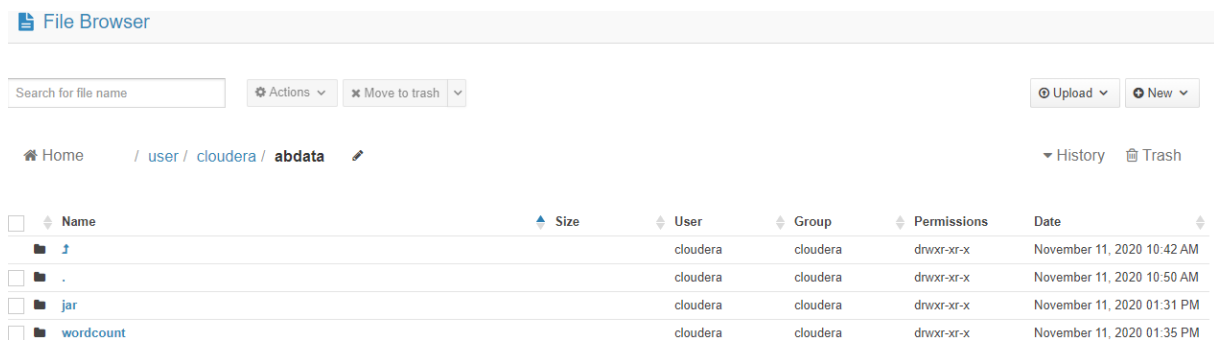
Page 1 of 1

Create Directory

Directory Name

Cancel Create

Créer un fichier texte contenant des phrases par exemple et essayer de l'uploader par la suite.



Search for file name

Actions Move to trash

Upload New

Home / user / cloudera / abdata

History Trash

Name	Size	User	Group	Permissions	Date
.		cloudera	cloudera	drwxr-xr-x	November 11, 2020 10:42 AM
.		cloudera	cloudera	drwxr-xr-x	November 11, 2020 10:50 AM
jar		cloudera	cloudera	drwxr-xr-x	November 11, 2020 01:31 PM
wordcount		cloudera	cloudera	drwxr-xr-x	November 11, 2020 01:35 PM