Atelier Hadoop

Application 1 : Manipulation du fichiers sous HDFS

Med EL ASSAD

AU: 2021/2022

En octobre 2018, Cloudera et Hortonworks annonçaient leur fusion. Quelques mois plus tard, en mars 2019, les deux entreprises dévoilaient le fruit de leur alliance : la Cloudera Data Platform, premier **cloud de données d'entreprises (Enterprise Data Cloud)**.

Dans le cadre de l'événement annuel Cloudera Strata qui se déroulait cette semaine à New York, la CDP a enfin été lancée.

Hortonworks et Cloudera, les deux géants du Big Data, annoncent la fusion de leurs entreprises et de leurs plateformes. Ensemble, les deux firmes comptent combiner leurs atouts respectifs pour dominer les marchés du Data Management, du Machine Learning ou encore du Cloud hybride.

Depuis maintenant plusieurs années, trois vendeurs de distributions Hadoop se disputent le marché du Big Data : **MapR, Cloudera et Hortonworks**.

Pour se distinguer, Cloudera et MapR ont ait le choix de proposer des add-on propriétaires à leurs distributions Hadoop.

Hortonworks, de son côté, a préféré rester le plus fidèle possible à la version originale proposée par Apache en open-source.

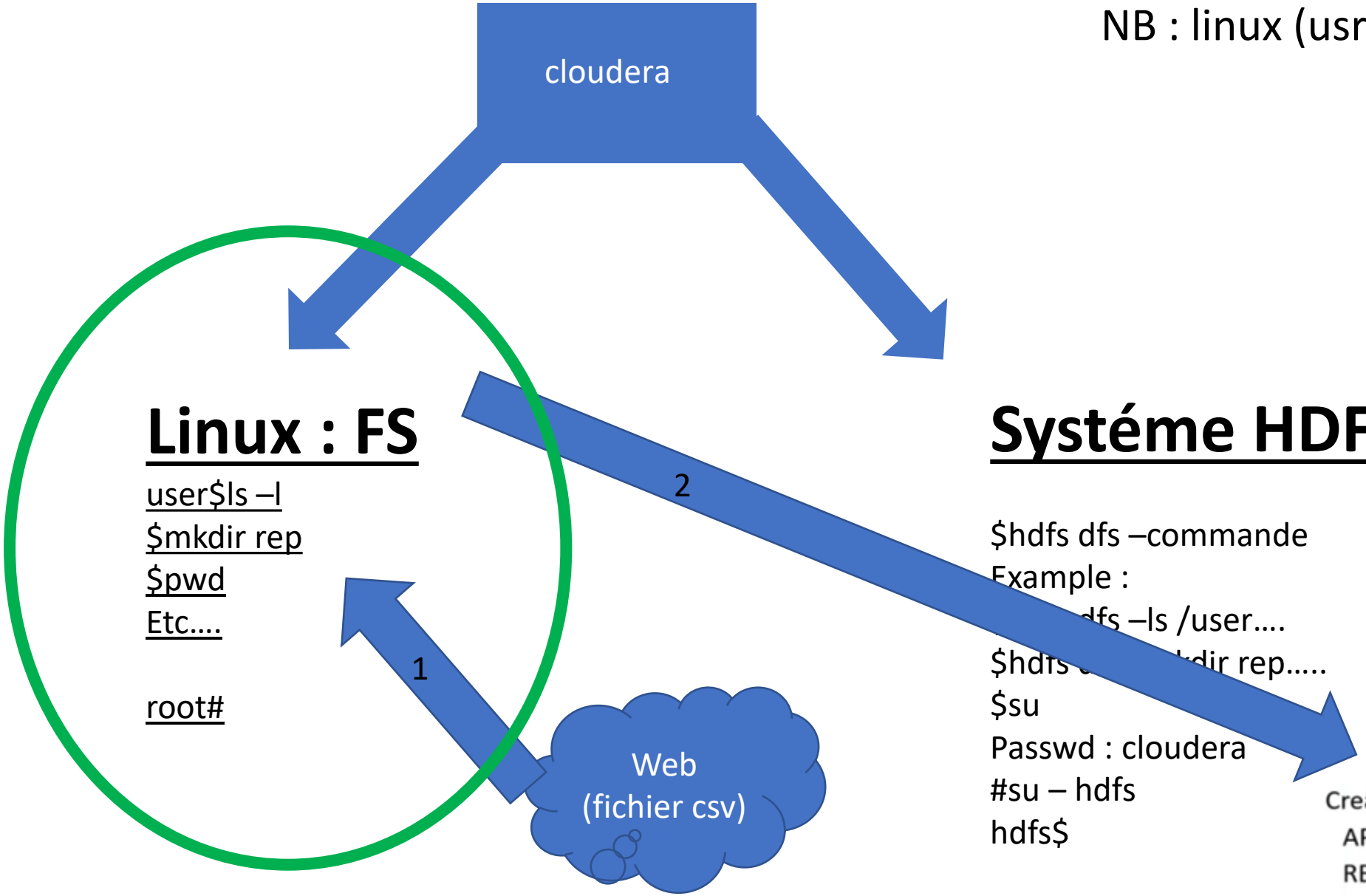
Cependant, si MapR se démarque fortement par son système fichier propriétaire, les différences entre les écosystèmes de Cloudera et Hortonworks ont toujours été minimes. Aujourd’hui, **les deux entreprises annoncent leur fusion**.

Cette décision stratégique a pour but d’accélérer le développement du marché, de stimuler l’innovation, et de produire des bénéfices substantiels pour les clients, les partenaires et la communauté. Selon Tom Reilly, CEO de Cloudera, **les deux entreprises sont très complémentaires**.

Il estime notamment que les investissements d’Hortonworks dans le domaine du Data Management End-to-End vont compléter les investissements de Cloudera dans le Data Warehousing et le Machine Learning.

https://sparkbyexamples.com/apache-hadoop/hadoop-hdfs-dfs-commands-and-starting-hdfs-dfs-services/

NB : linux (usr) vs HDFS (user)

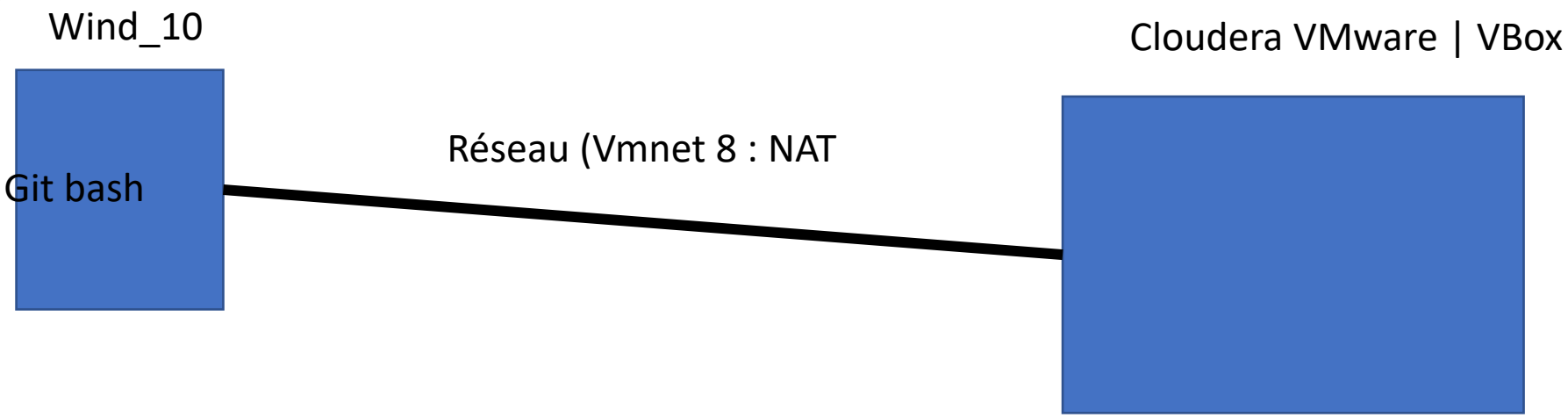


Create the HDFS application's folders :

APP_HOME: /user/industries

RES_DIR: /user/industries/res

Besoins



Vmware ➔

Vmnet 0 : linux (mv) accés l'internet.

Carte wifi pont : bridge

Vmnet 1

.

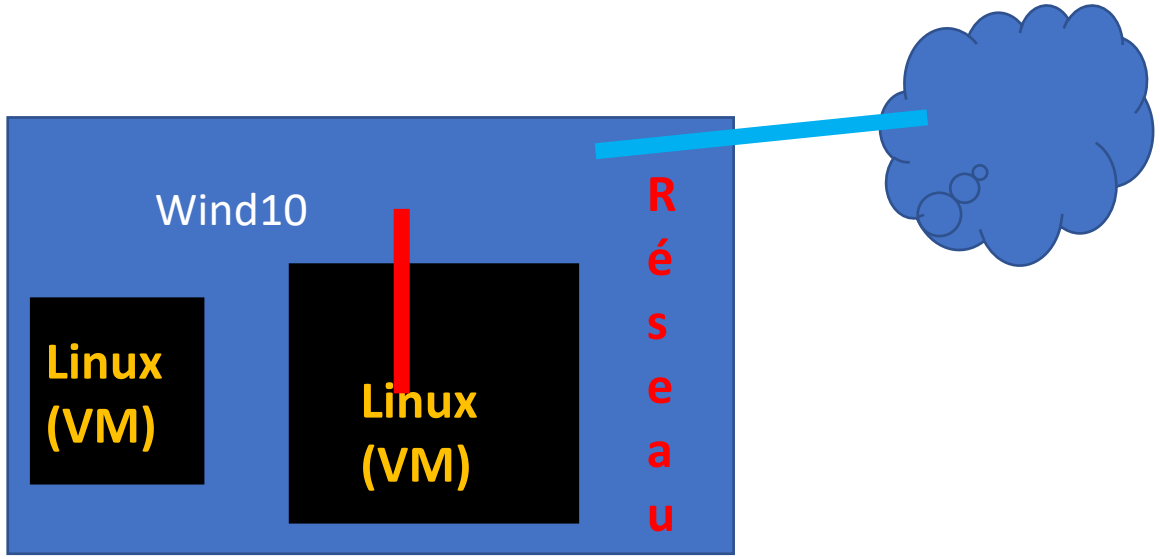
.

.

Vmnet7

Vmnet8 (NAT : Network Adress Translation) :

+ieurs Machines



HDFS Commands Using Java API

The main Hadoop HDFS command-lines:

We can perform Hadoop HDFS file operations like **changing the file permissions**, **viewing the file contents**, **creating files or directories**, **copying file/directory** from the local file system to HDFS or vice-versa, etc

HDFS Commands Using Java API

Additional HDFS command-lines:

Performing additional HDFS file operations like **moving a file**, **deleting a file**, **changing files permissions**, **setting replication factor**, **changing files ownership**, …

HDFS Commands Using Java API

We have seen the main HDFS command lines:

Get hadoop version:

```
Hadoop version
```

Create directories:

```
hdfs dfs –mkdir /hdfs_path/Dir_Name
```

Download file to hdfs:

```
hdfs dfs –put   <localSrc>   <hdfs dest>
Hdfs dfs –copyFromLocal   <localSrc>   <hdfs dest>
```

Upload file from hdfs:

```
Hdfs dfs –get  <hdfs Src>   <Local dest>
Hdfs dfs –copyToLocal  <hdfs Src>   <Local dest>
```

List files and directories:

```
Hdfs dfs  –ls  [-R]  /hdfs_path
```

HDFS Commands Using Java API

Additional HDFS command-lines:

Moves the file or directory from the local filesystem to the destination in Hadoop HDFS:

```
hdfs dfs -moveFromLocal <localsrc> <HDFS dest>
```

Moves the file or directory from the HDFS to the destination in the local filesystem:

```
hdfs dfs -moveToLocal <hdfs src> <localdest>
```

Shows the last 1KB of a file on console or stdout:

```
hdfs dfs -tail [-f] </path_file>        #The -f shows the append data as the file grows.
```

Removes the file or directory present in the specified path:

```
hdfs fs –rm [-R] <path>
```

Makes the trash empty:

```
hdfs dfs -expunge
```

HDFS Commands Using Java API

Additional HDFS command-lines:

Changes the group of the file specified in the path:

```
hdfs dfs -chgrp <group> <path>
```

Changes the replication factor to a specific count for the file specified in the path:

```
hdfs dfs -setrep <rep Factor>   <path>
```

If used for a directory, then it will recursively change the replication factor for all the files residing in the directory.

Prints a summary of the amount of disk usage of all files/directories in the path:

```
hdfs dfs  –du –s /directory/filename
```

Shows the capacity, size, and free space available on the HDFS file system:

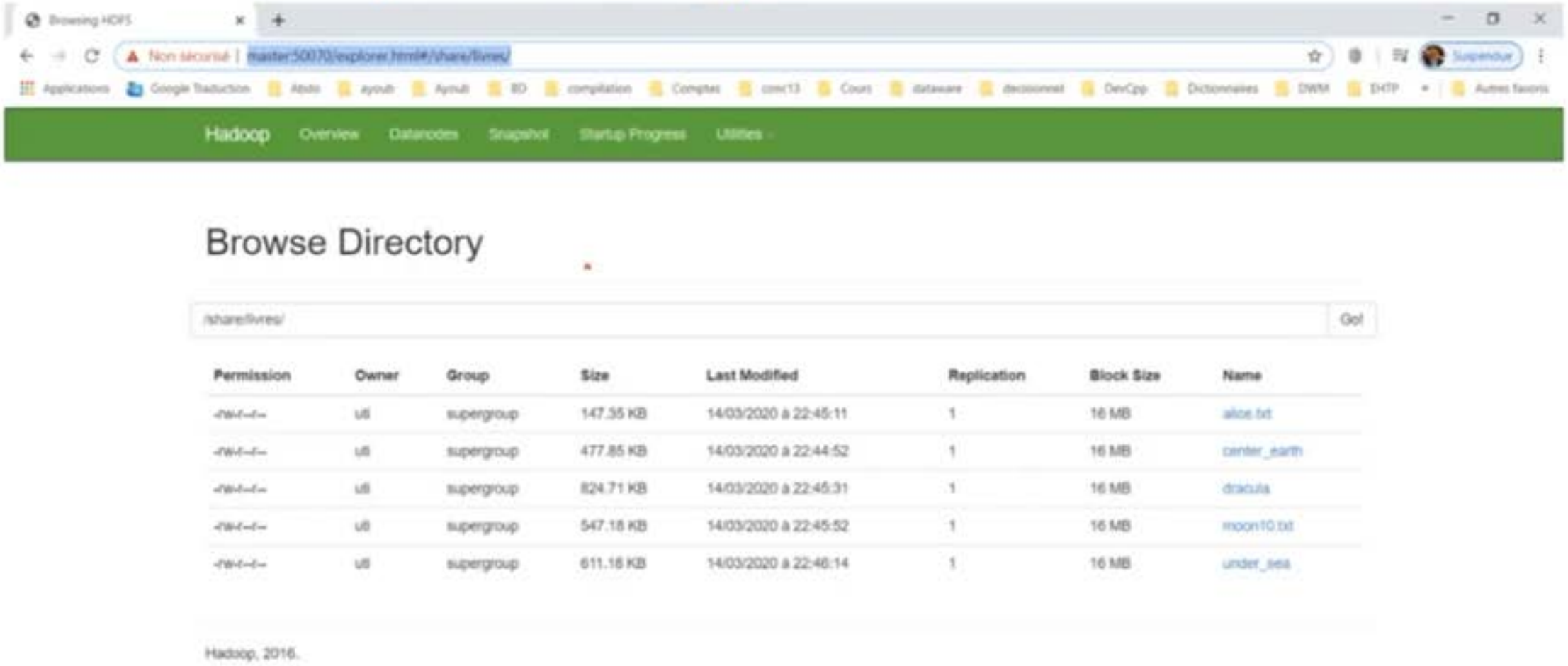
```
Hdfs dfs   -df [-h] <path>
```

Check the health of the HDFS:

```
fsck <path> [ -move | -delete | -openforwrite] [-files [-blocks [-locations | -racks]]]
```

HDFS Commands Using Java API

Web console to explore hdfs storage:



cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation

File Edit View VM Tabs Help

Library

Type here to search

- My Computer
  - redserve
  - ubunto
  - Wind7
  - WS2019
  - Windows7
  - W7
  - kali
  - Red Hat Linux (2)
  - Red Hat Linux (2)
  - ensi_security
  - Ubuntu_21.04_VM_LinuxVMImages.COM
  - cigma_14h
  - Red Hat Linux (2)
  - winxp
  - cloudera-quickstart-vm-5.12.0-0-vmware
  - Windows7
  - W7
  - GNS3 VM
  - IE11-Win7-VMWare
  - cd_hadoop
  - cloudera-quickstart-vm-5.12.0-0-vmware
- Shared VMs (Deprecated)

cloudera-quickstart-vm-5...

cloudera-quickstart-vm-5.12.0-0-vmware

Power on this virtual machine
Edit virtual machine settings
Upgrade this virtual machine

Devices

Memory	4 GB
Processors	1
Hard Disk (SCSI)	64 GB
CD/DVD (IDE)	Auto detect
Network Adapter	NAT
Display	Auto detect

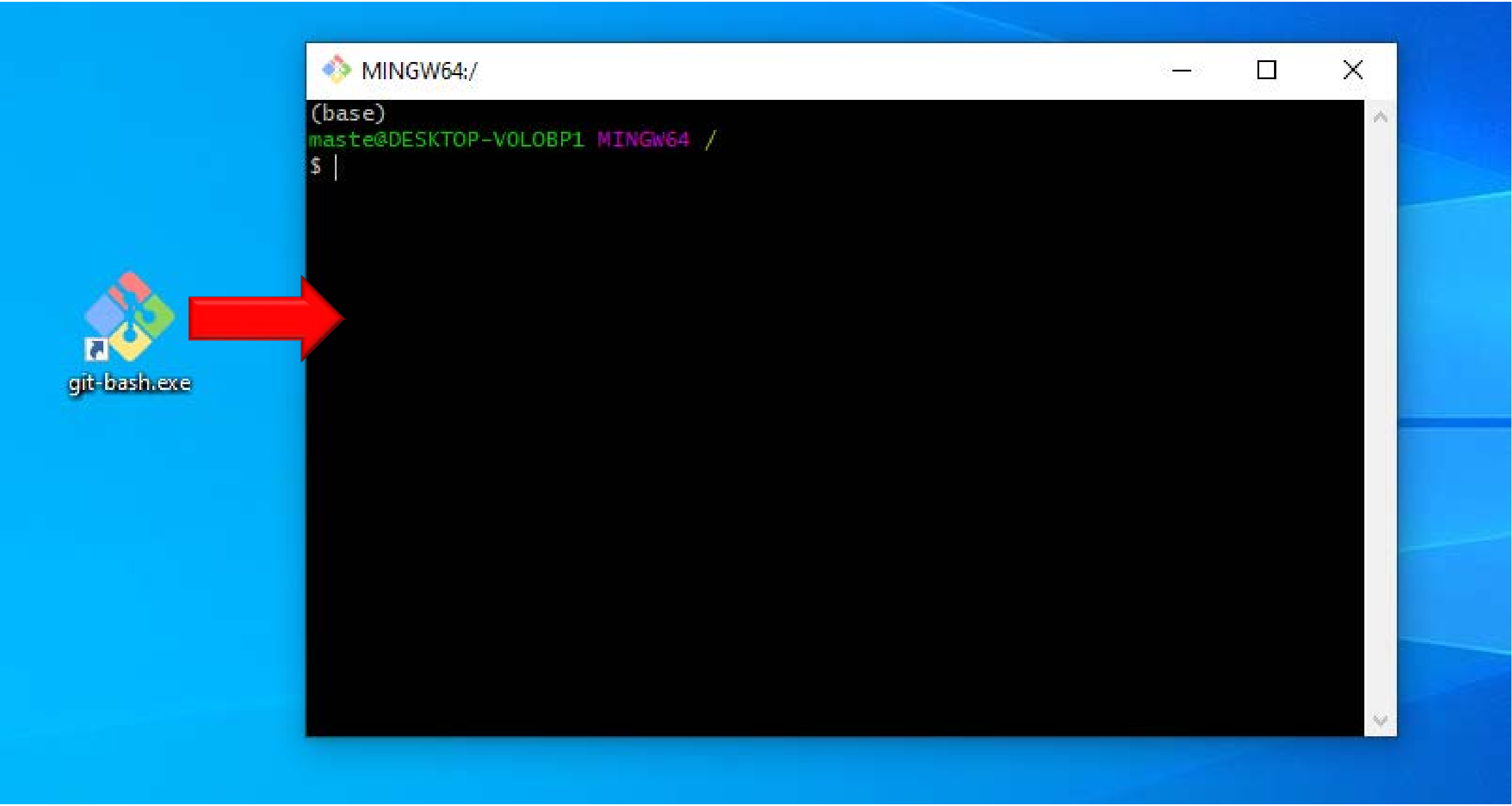
Description

Type here to enter a description of this virtual machine.

Virtual Machine Details

State: Powered off
Configuration file: G:\cloudera-quickstart-vm-5.12.0-0-vmware\cloudera-quickstart-vm-5.12.0-0-vmware.vmx
Hardware compatibility: Workstation 8.x virtual machine
Primary IP address: Network information is not available

CentOS 6.7


git-bash.exe
MINGW64:/
(base)
maste@DESKTOP-VOLOBP1 MINGW64 /
$

cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation

File Edit View VM Tabs Help

cloudera-quickstart-vm-5...

Applications Places System Tue Nov 23, 12:47 AM cloudera

Computer

cloudera's Home

Trash

Migrate to Parcel

Launch Cloudera Express

Launch Cloudera Enterprise (trial)

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

about:sessionrestore | Cloudera Live : Welcom...

quickstart.cloudera/#/

Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

cloudera LIVE Navigation

Welcome to Your Cloudera QuickStart VM!

Your Cluster

Node	Address
Manager Node	192.168.100.129
Worker Node 1	192.168.100.129

Get Started

Cloudera Live : Welco...

To direct input to this VM, click inside or press Ctrl+G.

cloudera-quickstart-vm-5.12.0-0-vmware - VMware Workstation

File Edit View VM Tabs Help

cloudera-quickstart-vm-5...

Applications Places System

Tue Nov 23, 12:48 AM cloudera

Computer

Eclipse

cloudera's Home

Trash

Migrate to Parcels

Launch Cloudera Express

Launch Cloudera Enterprise (trial)

cloudera

[Cloudera Live : Welco...

To direct input to this VM, move the mouse pointer inside or press Ctrl+G.

cloudera@quickstart:~

File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ setxkbmap fr
```

cloudera@quickstart:~

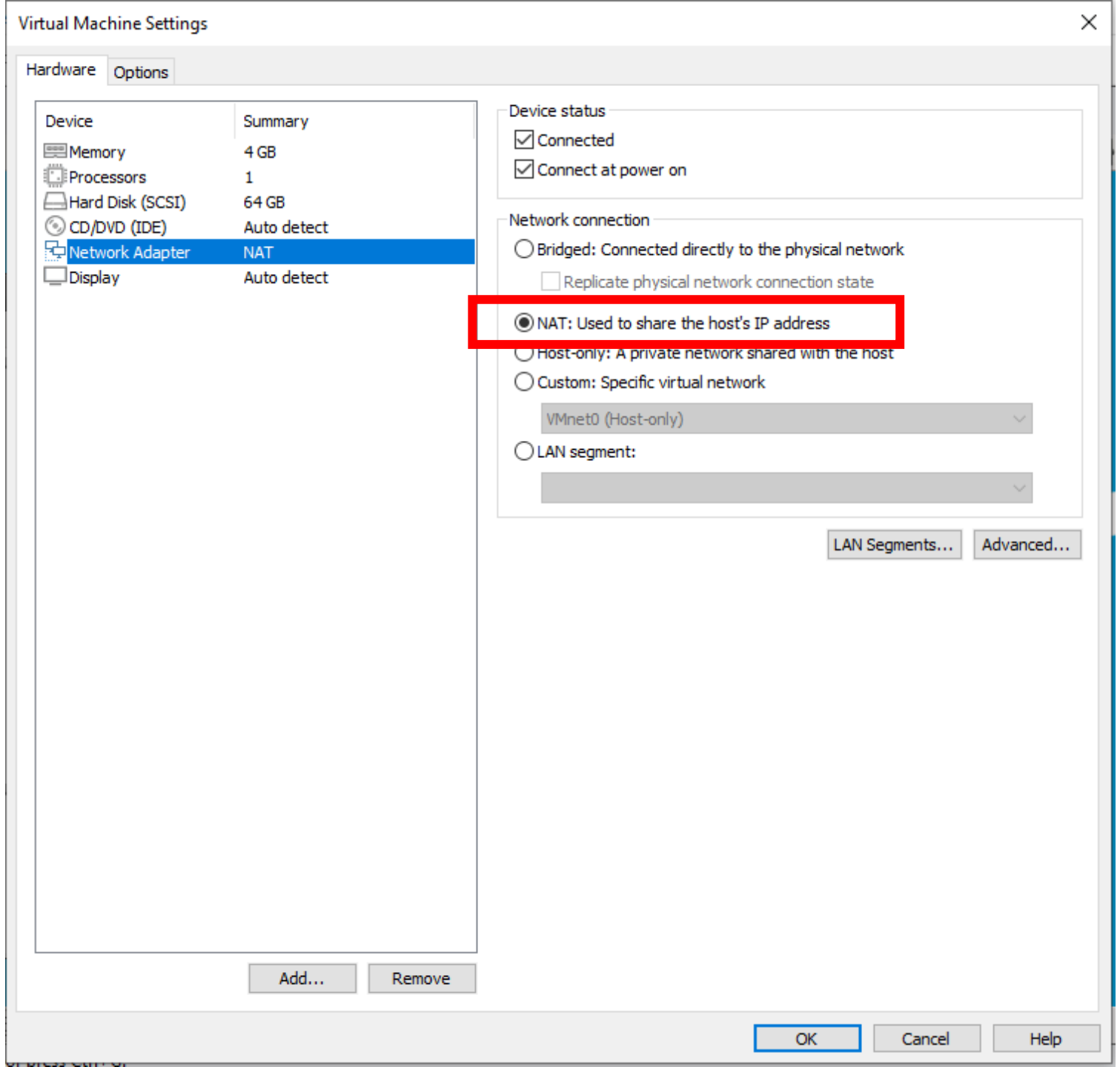
File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ ifconfig
eth1      Link encap:Ethernet  HWaddr 00:0C:29:CE:F7:FE
          inet addr:192.168.100.129  Bcast:192.168.100.255  Mask:255.255.255.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:426 errors:0 dropped:0 overruns:0 frame:0
          TX packets:460 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:96808 (94.5 KiB)  TX bytes:45050 (43.9 KiB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:17452 errors:0 dropped:0 overruns:0 frame:0
          TX packets:17452 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:2474664 (2.3 MiB)  TX bytes:2474664 (2.3 MiB)

[cloudera@quickstart ~]$
```

MINGW64:/

```
(base)
maste@DESKTOP-V0LOBP1 MINGW64 /
$ ipconfig
```

```
Carte Ethernet VMware Network Adapter VMnet8 :

   Suffixe DNS propre ▒ la connexion. . . :
   Adresse IPv6 de liaison locale. . . . . : fe80::9112:7dc6:8e61:cc68%16
   Adresse IPv4. . . . . . . . . . . . . . : 192.168.100.1
   Masque de sous-r▒seau. . . .▒. . . . . : 255.255.255.0
   Passerelle par d▒faut. . . .▒. . . . . :
```


Virtual Machine Settings
Hardware
Options
Device
Summary
Memory
4 GB
Processors
1
Hard Disk (SCSI)
64 GB
CD/DVD (IDE)
Auto detect
Network Adapter
NAT
Display
Auto detect
Add...
Remove
Device status
Connected
Connect at power on
Network connection
Bridged: Connected directly to the physical network
Replicate physical network connection state
NAT: Used to share the host's IP address
Host-only: A private network shared with the host
Custom: Specific virtual network
VMnet0 (Host-only)
LAN segment:
LAN Segments...
Advanced...
OK
Cancel
Help

Virtual Network Editor

Name	Type	External Connection	Host Connection	DHCP	Subnet Address
VMnet0	Host-only	-	Connected	Enabled	192.168.68.0
VMnet1	Custom	-	-	-	192.168.74.0
VMnet2	Host-only	-	Connected	-	192.168.136.0
VMnet8	NAT	NAT	Connected	Enabled	192.168.100.0

Add Network... | Remove Network | Rename Network...

VMnet Information

- ○ Bridged (connect VMs directly to the external network)
  - Bridged to: [ ] Automatic Settings...
- ◉ NAT (shared host's IP address with VMs) — NAT Settings...
- ○ Host-only (connect VMs internally in a private network)

☑ Connect a host virtual adapter to this network

Host virtual adapter name: VMware Network Adapter VMnet8

☑ Use local DHCP service to distribute IP address to VMs — DHCP Settings...

Subnet IP: 192 . 168 . 100 . 0 Subnet mask: 255 . 255 . 255 . 0

Restore Defaults | Import... | Export... | OK | Cancel | Apply | Help

cloudera

Windows 10

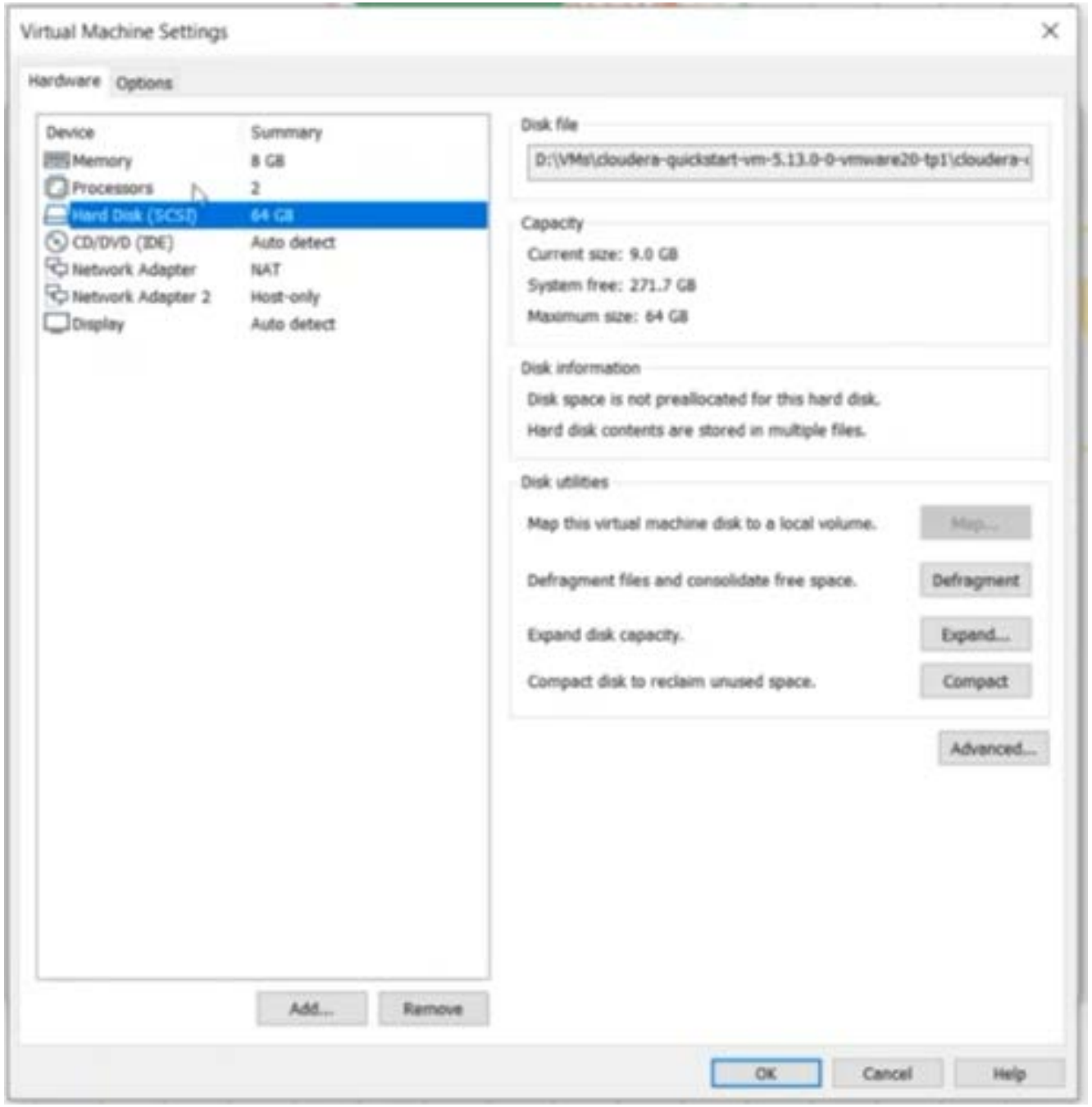
```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ ping 192.168.100.1
PING 192.168.100.1 (192.168.100.1) 56(84) bytes of data.
64 bytes from 192.168.100.1: icmp_seq=1 ttl=128 time=0.668 ms
64 bytes from 192.168.100.1: icmp_seq=2 ttl=128 time=0.305 ms
64 bytes from 192.168.100.1: icmp_seq=3 ttl=128 time=0.311 ms
64 bytes from 192.168.100.1: icmp_seq=4 ttl=128 time=1.27 ms
^C
--- 192.168.100.1 ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3119ms
rtt min/avg/max/mdev = 0.305/0.640/1.277/0.396 ms
[cloudera@quickstart ~]$
```

Windows 10

cloudera

```
MINGW64:/
(base)
maste@DESKTOP-V0LOBP1 MINGW64 /
$ ping 192.168.100.129

Envoi d'une requête 'Ping'  192.168.100.129 avec 32 octets de données:
Réponse de 192.168.100.129: octets=32 temps<1ms TTL=64
Réponse de 192.168.100.129: octets=32 temps<1ms TTL=64
Réponse de 192.168.100.129: octets=32 temps<1ms TTL=64
Réponse de 192.168.100.129: octets=32 temps<1ms TTL=64

Statistiques Ping pour 192.168.100.129:
    Paquets: envoyés = 4, reçus = 4, perdus = 0 (perte 0%),
Durée approximative des boucles en millisecondes :
    Minimum = 0ms, Maximum = 0ms, Moyenne = 0ms
```

NB


Virtual Machine Settings
Hardware
Options
Device
Summary
Memory
8 GB
Processors
2
Hard Disk (SCSI)
64 GB
CD/DVD (IDE)
Auto detect
Network Adapter
NAT
Network Adapter 2
Host-only
Display
Auto detect
Disk file
D:\VMs\cloudera-quickstart-vm-5.13.0-0-vmware20-tp1\cloudera-
Capacity
Current size: 9.0 GB
System free: 271.7 GB
Maximum size: 64 GB
Disk information
Disk space is not preallocated for this hard disk.
Hard disk contents are stored in multiple files.
Disk utilities
Map this virtual machine disk to a local volume.
Map...
Defragment files and consolidate free space.
Defragment
Expand disk capacity.
Expand...
Compact disk to reclaim unused space.
Compact
Advanced...
Add...
Remove
OK
Cancel
Help

etc

Fichier | Accueil | Partage | Affichage

Épingler à Accès rapide | Copier | Coller | Déplacer vers | Copier vers | Supprimer | Renommer | Nouveau dossier | Propriétés | Sélectionner tout | Aucun | Inverser la sélection

Presse-papiers | Organiser | Nouveau | Ouvrir | Sélectionner

Ce PC › Disque local (C:) › Windows › System32 › drivers › etc

Rechercher dans : ...

- Ce PC
- Bureau
- Documents
- Images
- Musique
- Objets 3D
- Téléchargements
- Vidéos
- Disque local (C:)
- Disque local (D:)

Nom	Modifié le	Type	Taille
hosts	19/11/2021 23:06	Fichier	1 Ko
lmhosts.sam	07/12/2019 10:12	Fichier SAM	4 Ko
networks	07/12/2019 10:12	Fichier	1 Ko
protocol	07/12/2019 10:12	Fichier	2 Ko
services	07/12/2019 10:12	Fichier	18 Ko

5 élément(s) | 1 élément sélectionné 873 octet(s)

C:\Windows\System32\drivers\etc\hosts - Notepad++

Fichier Édition Recherche Affichage Encodage Langage Paramètres Outils Macro Exécution Modules d'extension Documents ?

hosts

```
# Copyright (c) 1993-2009 Microsoft Corp.
#
# This is a sample HOSTS file used by Microsoft TCP/IP for Windows.
#
# This file contains the mappings of IP addresses to host names. Each
# entry should be kept on an individual line. The IP address should
# be placed in the first column followed by the corresponding host name.
# The IP address and the host name should be separated by at least one
# space.
#
# Additionally, comments (such as these) may be inserted on individual
# lines or following the machine name denoted by a '#' symbol.
#
# For example:
#
#      102.54.94.97     rhino.acme.com          # source server
#       38.25.63.10     x.acme.com              # x client host

# localhost name resolution is handled within DNS itself.
#   127.0.0.1       localhost
#   ::1             localhost
127.0.0.1 localhost
192.168.100.128 quickstart
```

cloudera@quickstart:~

```
(base)
maste@DESKTOP-V0LOBP1 MINGW64 /
$ ssh cloudera@quickstart
Warning: Permanently added the RSA host key for
 IP address '192.168.100.129' to the list of kn
own hosts.
cloudera@quickstart's password: cloudera
[cloudera@quickstart ~]$
```

Nom de l'utilisateur

Nom de la machine

```
cloudera@quickstart:~
[cloudera@quickstart ~]$ uname -a
Linux quickstart.cloudera 2.6.32-573.el6.x86_64 #1 SMP Thu Jul 23 15:44:03 UTC 2015 x86_64 x86_64 x
86_64 GNU/Linux
```

uname (short for unix name) est une commande Unix qui affiche les informations système sur la machine sur laquelle elle est exécutée

$man uname « plus d’informations »

https://www.ibm.com/docs/en/tnpm/1.4.3?topic=lp-operating-system

ORACLE

– Linux hosts running 64-bit Red Hat Enterprise Linux Version 7.3 and 7.4 with Oracle 12c (12.1.0.2).

To check the version of your operating system, enter:

```
# cat /etc/redhat-release
```

This command must return the output similar to:

```
Red Hat Enterprise Linux Server
release 7.3
```

IBM DB2

– Linux hosts running 64-bit Red Hat Enterprise Linux Version 7.3 and 7.4 with DB2 Version DB2® 11.1.1.1.

To check the version of your operating system, enter:

```
# cat /etc/redhat-release
```

This command must return the output similar to:

```
Red Hat Enterprise Linux Server
release 7.3
```

To verify the processor type, run the following command:

```
uname -p
```

To verify the machine type, run the following command:

```
uname -m
```

To verify the hardware platform, run the following command:

```
uname -i
```

All results should contain the output:

```
x86_64
```

To verify the processor type, run the following command:

```
uname -p
```

To verify the machine type, run the following command:

```
uname -m
```

To verify the hardware platform, run the following command:

```
uname -i
```

All results should contain the output:

```
x86_64
```

cloudera@quickstart:~

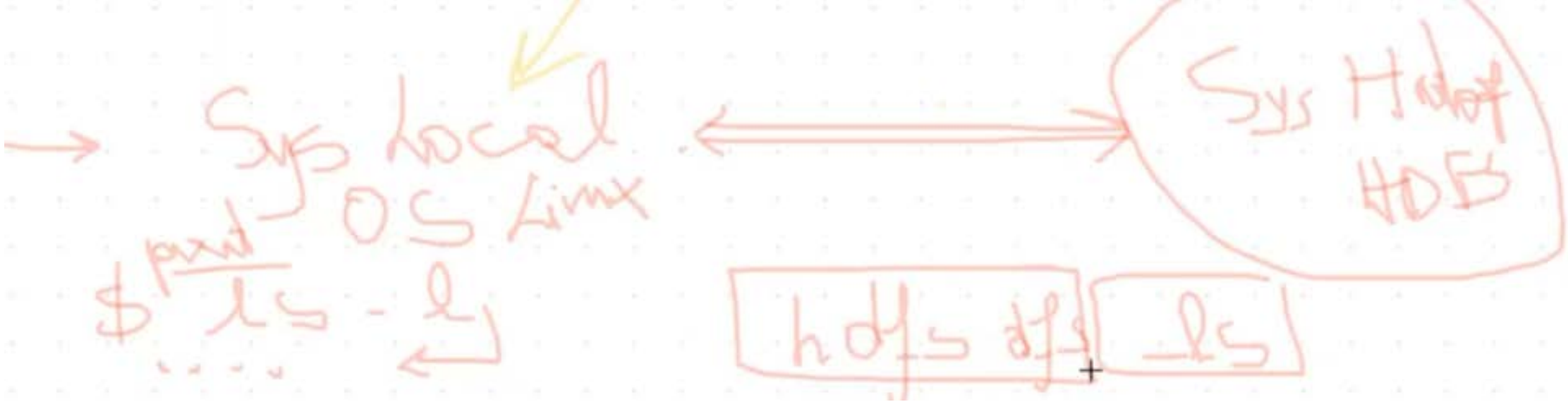
```
[cloudera@quickstart ~]$ hadoop version
Hadoop 2.6.0-cdh5.12.0
Subversion http://github.com/cloudera/hadoop -r dba647c5a8bc5e09b572d76a
8d29481c78d1a0dd
Compiled by jenkins on 2017-06-29T11:32Z
Compiled with protoc 2.5.0
From source with checksum 7c45ae7a4592ce5af86bc4598c5b4
This command was run using /usr/lib/hadoop/hadoop-common-2.6.0-cdh5.12.0
.jar
[cloudera@quickstart ~]$
```

cloudera@quickstart:~

```
[cloudera@quickstart ~]$ pwd
/home/cloudera
```

cloudera@quickstart:~

```
[cloudera@quickstart ~]$ uname -r
2.6.32-573.el6.x86_64
[cloudera@quickstart ~]$ ls -l
total 188
-rwxrwxr-x 1 cloudera cloudera  5387 Jul 19  2017 cloudera-manager
-rwxrwxr-x 1 cloudera cloudera  9964 Jul 19  2017 cm_api.py
drwxrwxr-x 2 cloudera cloudera  4096 Jul 19  2017 Desktop
drwxrwxr-x 4 cloudera cloudera  4096 Jul 19  2017 Documents
drwxr-xr-x 2 cloudera cloudera  4096 Nov 19 14:52 Downloads
drwxrwsr-x 9 cloudera cloudera  4096 Feb 19  2015 eclipse
-rw-rw-r-- 1 cloudera cloudera 53655 Jul 19  2017 enterprise-deployment.json
-rw-rw-r-- 1 cloudera cloudera 50515 Jul 19  2017 express-deployment.json
-rwxrwxr-x 1 cloudera cloudera  5007 Jul 19  2017 kerberos
drwxrwxr-x 2 cloudera cloudera  4096 Jul 19  2017 lib
drwxr-xr-x 2 cloudera cloudera  4096 Nov 19 14:52 Music
-rwxrwxr-x 1 cloudera cloudera  4228 Jul 19  2017 parcels
drwxr-xr-x 2 cloudera cloudera  4096 Nov 19 14:52 Pictures
drwxr-xr-x 2 cloudera cloudera  4096 Nov 19 14:52 Public
drwxr-xr-x 2 cloudera cloudera  4096 Nov 19 14:52 Templates
drwxr-xr-x 2 cloudera cloudera  4096 Nov 19 14:52 Videos
drwxrwxr-x 4 cloudera cloudera  4096 Jul 19  2017 workspace
[cloudera@quickstart ~]$
```



cloudera@quickstart:~

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx   - hdfs   supergroup          0 2017-07-19 06:29 /benchmarks
drwxr-xr-x   - hbase  supergroup          0 2021-11-23 00:47 /hbase
drwxr-xr-x   - solr   solr                0 2017-07-19 06:31 /solr
drwxrwxrwt   - hdfs   supergroup          0 2021-11-19 14:53 /tmp
drwxr-xr-x   - hdfs   supergroup          0 2017-07-19 06:31 /user
drwxr-xr-x   - hdfs   supergroup          0 2017-07-19 06:31 /var
[cloudera@quickstart ~]$
```

Système local : OS = linux

ls -l

Système : hfds (hadoop)

Notre Système de stockage

hdfs dfs –ls /

Fixe pour le système HDFS

Module

“hdfs dfs -ls /” ou “hadoop hdfs -ls /”

hadoop hdfs .... : obsolète : l'utilisation de ce script pour exécuter la commande hdfs est obsolète. Utilisez plutôt la commande hdfs pour cela.

Usr (linux) diff user (hadoop)

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user
Found 8 items
drwxr-xr-x   - cloudera cloudera          0 2017-07-19 06:28 /user/cloudera
drwxr-xr-x   - mapred   hadoop            0 2017-07-19 06:29 /user/history
drwxrwxrwx   - hive     supergroup        0 2017-07-19 06:31 /user/hive
drwxrwxrwx   - hue      supergroup        0 2017-07-19 06:30 /user/hue
drwxrwxrwx   - jenkins  supergroup        0 2017-07-19 06:29 /user/jenkins
drwxrwxrwx   - oozie    supergroup        0 2017-07-19 06:30 /user/oozie
drwxrwxrwx   - root     supergroup        0 2017-07-19 06:29 /user/root
drwxr-xr-x   - hdfs     supergroup        0 2017-07-19 06:31 /user/spark
[cloudera@quickstart ~]$
```

Remarque : supposons que nous travaillons beaucoup avec ce chemin : /user/hive/warehouse

cloudera@quickstart:~

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/hive
Found 1 items
drwxrwxrwx   - hive supergroup          0 2017-07-19 06:31 /user/hive/warehouse
[cloudera@quickstart ~]$ hdfs dfs -ls /user/hive/warehouse
[cloudera@quickstart ~]$ MYPath=/user/hive/warehouse
[cloudera@quickstart ~]$ echo $MYPath
/user/hive/warehouse
[cloudera@quickstart ~]$ hdfs dfs -ls MYPath
ls: `MYPath': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -ls $MYPath
[cloudera@quickstart ~]$
```

```
[cloudera@quickstart ~]$ MYPath=/user/hive
[cloudera@quickstart ~]$ hdfs dfs -ls $MYPath
```

cloudera@quickstart:~

```
[cloudera@quickstart ~]$ sudo jps
2475 NameNode
3337 RESTServer
4369 HRegionServer
5121
5159
3047 ResourceManager
2736 Bootstrap
2556 SecondaryNameNode
3475 ThriftServer
5098 Bootstrap
9992 Jps
3570 RunJar
4222 Bootstrap
2278 DataNode
4243 HistoryServer
2363 JournalNode
2209 QuorumPeerMain
4877 Bootstrap
2873 NodeManager
2790 JobHistoryServer
3684 RunJar
3245 HMaster
[cloudera@quickstart ~]$
```

2-Pour gérer le Stockage de données

3-Pour planifier des exécution des taches

1-Pour effectuer le Stockage de données

4-Pour exécuter ces taches

Les 4 services de base : qui veut dire hadoop en exécution

Travail à faire

To do so, we have to run the following commands using a terminal console:

1. Create the HDFS application's folders :

   APP_HOME: /user/industries

   RES_DIR: /user/industries/res

2. **Download a csv file using its URL** and upload it to HDFS folder RES_DIR under the name "indicCommercExter1.csv"

   FILE_URL: https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-volumes/Overseas-trade-indexes-prices-and-volumes-December-2019-quarter-provisional/Download-data/overseas-trade-indexes-december-2019-quarter-provisional.csv

1) Hadoop version

2) Hdfs dfs -ls /...

3) Hdfs dfs -mkdir /.../nmDir

cloudera@quickstart:/home/cloudera

```
[cloudera@quickstart ~]$ APP_HOME=/user/industries
[cloudera@quickstart ~]$ RES_DIR=$APP_HOME/res
[cloudera@quickstart ~]$ su
Password: cloudera
[root@quickstart cloudera]#
```

cloudera@quickstart:/home/cloudera

```
[cloudera@quickstart ~]$ APP_HOME=/user/industries
[cloudera@quickstart ~]$ RES_DIR=$APP_HOME/res
[cloudera@quickstart ~]$ su
Password:
[root@quickstart cloudera]# su - hdfs
-bash-4.1$
```

Basculer vers un autre utilisateur HDFS

```
[cloudera@quickstart ~]$ APP_HOME=/user/industries
[cloudera@quickstart ~]$ RES_DIR=$APP_HOME/res
[cloudera@quickstart ~]$ su
Password:
[root@quickstart cloudera]# su - hdfs
[root@quickstart cloudera]# su - hdfs
-bash-4.1$ APP_HOME=/user/industries
-bash-4.1$ RES_DIR=$APP_HOME/res
-bash-4.1$
```

Retapez les deux commande pour l'utilisateur HDFS

cloudera@quickstart:/home/cloudera

```
[root@quickstart cloudera]# su - hdfs
-bash-4.1$ APP_HOME=/user/industries
-bash-4.1$ RES_DIR=$APP_HOME/res
-bash-4.1$ echo $APP_HOME
/user/industries
-bash-4.1$ echo $RES_DIR
/user/industries/res
-bash-4.1$
```

cloudera@quickstart/home/cloudera

```
[root@quickstart cloudera]# su - hdfs
-bash-4.1$ echo $APP_HOME

-bash-4.1$ APP_HOME=/user/industries
-bash-4.1$  RES_DIR=$APP_HOME/res
-bash-4.1$ echo $APP_HOME
/user/industries
-bash-4.1$ echo $RES_DIR
/user/industries/res
-bash-4.1$ hdfs dfs -mkdir $APP_HOME
-bash-4.1$ hdfs dfs -mkdir $RES_DIR
-bash-4.1$ hdfs dfs -ls /user
Found 9 items
drwxr-xr-x   - cloudera cloudera          0 2017-10-23 10:28 /user/cloudera
drwxr-xr-x   - mapred   hadoop            0 2017-10-23 10:29 /user/history
drwxrwxrwx   - hive     supergroup        0 2017-10-23 10:31 /user/hive
drwxrwxrwx   - hue      supergroup        0 2017-10-23 10:30 /user/hue
drwxr-xr-x   - hdfs     supergroup        0 2020-03-24 02:10 /user/industries
drwxrwxrwx   - jenkins  supergroup        0 2017-10-23 10:30 /user/jenkins
drwxrwxrwx   - oozie    supergroup        0 2017-10-23 10:30 /user/oozie
drwxrwxrwx   - root     supergroup        0 2017-10-23 10:30 /user/root
drwxr-xr-x   - hdfs     supergroup        0 2017-10-23 10:31 /user/spark
-bash-4.1$ hdfs dfs -ls -R /user
```

```
-bash-4.1$ FILE_URL=https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-volumes/Over
seas-trade-indexes-prices-and-volumes-December-2019-quarter-provisional/Download-data/overseas-trade-indexe
s-december-2019-quarter-provisional.csv
-bash-4.1$ echo $RES_DIR
/user/industries/res
-bash-4.1$ echo $FILE_URL
https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-volumes/Overseas-trade-indexes-p
rices-and-volumes-December-2019-quarter-provisional/Download-data/overseas-trade-indexes-december-2019-quar
ter-provisional.csv
-bash-4.1$ FILE_BASENAME=indComExt1.csv
-bash-4.1$ FILE_NAME=$RES_DIR/$FILE_BASENAME
-bash-4.1$
```

cloudera@quickstart:/home/cloudera

```
drwxr-xr-x   - mapred   hadoop              0 2017-07-19 06:29 /user/history
drwxrwxrwx   - hive     supergroup          0 2017-07-19 06:31 /user/hive
drwxrwxrwx   - hue      supergroup          0 2017-07-19 06:30 /user/hue
-bash-4.1$ FILE_URL=https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-
volumes/Overseas-trade-indexes-prices-and-volumes-December-2019-quarter-provisional/Download-da
ta/overseas-trade-indexes-december-2019-quarter-provisional.csv
-bash-4.1$
-bash-4.1$ echo $FILE_URL
https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-volumes/Overseas-tra
de-indexes-prices-and-volumes-December-2019-quarter-provisional/Download-data/overseas-trade-in
dexes-december-2019-quarter-provisional.csv
-bash-4.1$
```

cloudera@quickstart:/home/cloudera

```
-bash-4.1$ FILE_BASENAME=indComExt1.csv
-bash-4.1$ FILE_NAME=$FILE_BASENAME
-bash-4.1$ FILE_NAME=$RES_DIR/$FILE_BASENAME
-bash-4.1$ echo $FILE_NAME Pour tester
/user/industries/res/indComExt1.csv
-bash-4.1$
```

cloudera@quickstart:/home/cloudera

```
-bash-4.1$ echo $FILE_NAME
/user/industries/res/indComExt1.csv
-bash-4.1$ echo $FILE_URL
https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-volumes/Overseas-trade-indexes-p
rices-and-volumes-December-2019-quarter-provisional/Download-data/overseas-trade-indexes-december-2019-quar
ter-provisional.csv
-bash-4.1$ wget -O - $FILE_URL|hdfs dfs -put - $FILE_NAME
--2020-03-24 02:23:26--  https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-volumes
/Overseas-trade-indexes-prices-and-volumes-December-2019-quarter-provisional/Download-data/overseas-trade-i
ndexes-december-2019-quarter-provisional.csv
Resolving www.stats.govt.nz... 45.60.13.104
Connecting to www.stats.govt.nz|45.60.13.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 20711303 (20M) [text/csv]
Saving to: "STDOUT"

100%[=================================================================>] 20,711,303  1.25M/s   in 17s

2020-03-24 02:23:47 (1.18 MB/s) - written to stdout [20711303/20711303]

-bash-4.1$ hdfs dfs -ls -R $APP_HOME
```

cloudera@quickstart:/home/cloudera

```
-bash-4.1$ wget -O - $FILE_URL | hdfs dfs -put - /user/industries
--2021-11-23 08:30:54--  https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-vo
lumes/Overseas-trade-indexes-prices-and-volumes-December-2019-quarter-provisional/Download-data/overse
as-trade-indexes-december-2019-quarter-provisional.csv
Resolving www.stats.govt.nz... 45.60.13.104
Connecting to www.stats.govt.nz|45.60.13.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 20711303 (20M) [text/csv]
Saving to: “STDOUT”

100%[===================================================================>] 20,711,303  1.01M/s   in 32s

2021-11-23 08:31:29 (627 KB/s) - written to stdout [20711303/20711303]

-bash-4.1$
```

```
-bash-4.1$ hdfs dfs -ls -R $APP_HOME
drwxr-xr-x   - hdfs supergroup          0 2021-11-23 08:38 /user/industries/res
-rw-r--r--   1 hdfs supergroup   20711303 2021-11-23 08:35 /user/industries/res/indComExt1.csv
-bash-4.1$
```

On vérifie la présence du fichier

cloudera@quickstart:/home/cloudera

```
-bash-4.1$ hdfs getconf -confkey dfs.blocksize
134217728
-bash-4.1$
```

La taille des blocks

La taille d'un bloc est : 134MB

cloudera@quickstart:/home/cloudera

```
-bash-4.1$ vim /etc/hadoop/conf/hdfs-site.xml
```

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
```

```
-bash-4.1$ cat /etc/hadoop/conf/hdfs-site.xml|grep blocksize
```

cloudera@quickstart:/home/cloudera

```xml
<?xml version="1.0"?>
<!--
  Licensed to the Apache Software Foundation (ASF) under one or more
  contributor license agreements.  See the NOTICE file distributed with
  this work for additional information regarding copyright ownership.
  The ASF licenses this file to You under the Apache License, Version 2.0
  (the "License"); you may not use this file except in compliance with
  the License.  You may obtain a copy of the License at

      http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License.
-->
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <!-- Immediately exit safemode as soon as one DataNode checks in.
       On a multi-node cluster, these configurations must be removed.  -->
  <property>
    <name>dfs.safemode.extension</name>
    <value>0</value>
  </property>
  <property>
     <name>dfs.safemode.min.datanodes</name>
```

cloudera@quickstart:/home/cloudera

```
-bash-4.1$ hdfs getconf -confkey dfs.replication
1
-bash-4.1$
```

cloudera@quickstart:/home/cloudera

```
maste@DESKTOP-V0LOBP1 MINGW64 /
$
(base)
maste@DESKTOP-V0LOBP1 MINGW64 /
$ ssh cloudera@quickstart
cloudera@quickstart's password:
Last login: Tue Nov 23 09:13:41 2021 from 192.168.100.1
[cloudera@quickstart ~]$ su
Password:
[root@quickstart cloudera]# su - hdfs
-bash-4.1$
-bash-4.1$
-bash-4.1$ hdfs fsck $RES_DIR -files -blocks
```

cloudera@quickstart:/home/cloudera

```
-bash-4.1$ hdfs getconf -confKey dfs.replication
1
-bash-4.1$ hdfs dfs -ls -R $APP_HOME
drwxr-xr-x   - hdfs supergroup          0 2020-03-24 02:23 /user/industries/res
-rw-r--r--   1 hdfs supergroup   20711303 2020-03-24 02:23 /user/industries/res/indComExt1.csv
-bash-4.1$ hdfs fsck $RES_DIR  -files -blocks
Connecting to namenode via http://quickstart.cloudera:50070/fsck?ugi=hdfs&files=1&blocks=1&path=%2Fuser%2Fi
ndustries%2Fres
FSCK started by hdfs (auth:SIMPLE) from /127.0.0.1 for path /user/industries/res at Tue Mar 24 02:36:28 PDT
 2020
/user/industries/res <dir>
/user/industries/res/indComExt1.csv 20711303 bytes, 1 block(s):  OK
0. BP-333635372-127.0.0.1-1508779710286:blk_1073742764_1942 len=20711303 Live_repl=1

Status: HEALTHY
 Total size:    20711303 B
 Total dirs:    1
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      1 (avg. block size 20711303 B)
 Minimally replicated blocks:   1 (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 Tue Mar 24 02:36:28 PDT 2020 in 2 milliseconds


The filesystem under path '/user/industries/res' is HEALTHY
```

https://www.stats.govt.nz/assets/Uploads/Overseas-trade-indexes-prices-and-volumes/Overseas-trade-indexes-prices-and-volumes-December-2019-quarter-provisional/Download-data/overseas-trade-indexes-december-2019-quarter-provisional.csv