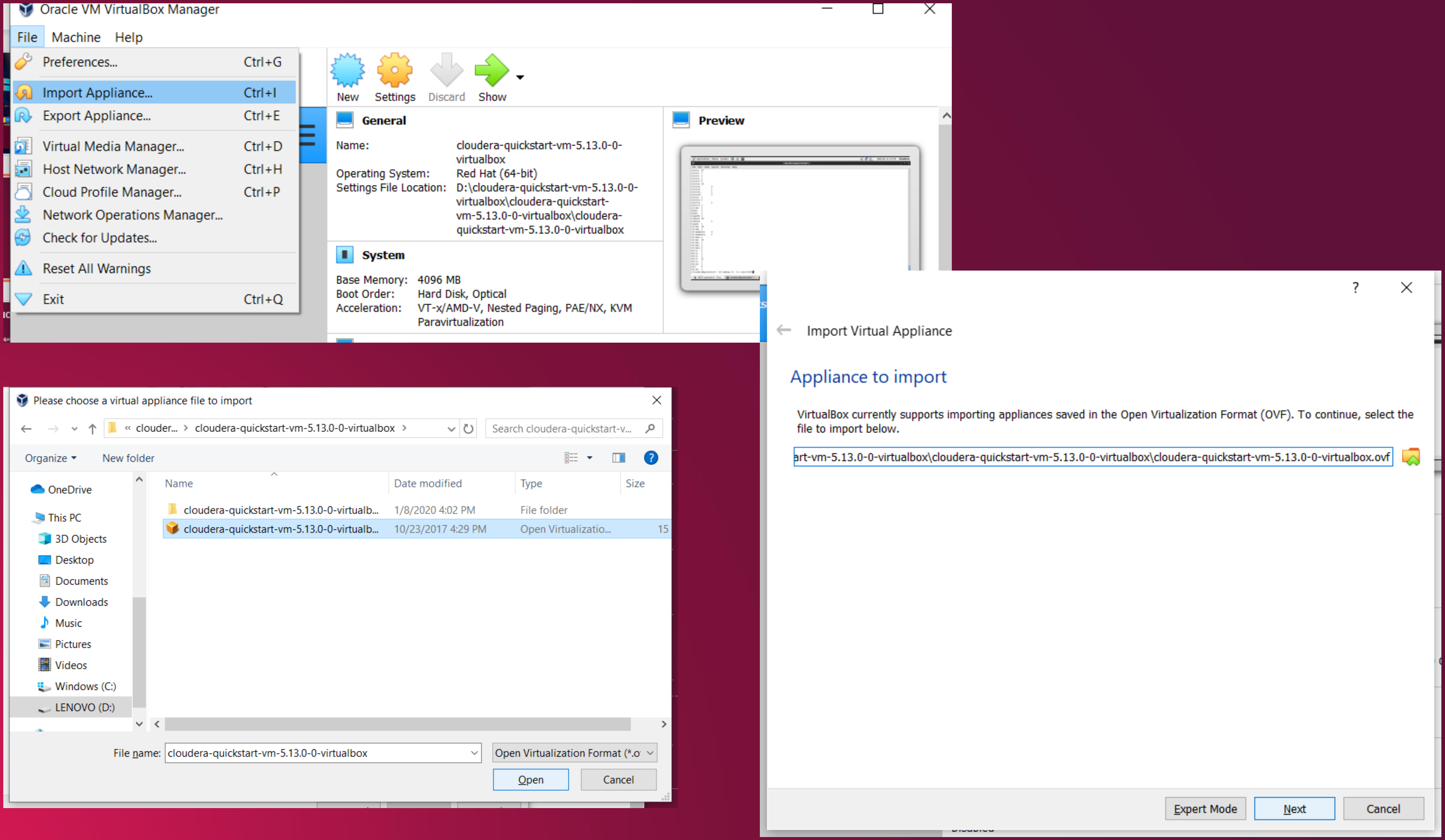
Análisis de Big Data

Cindy lópez


Oracle VM VirtualBox Manager
File
Machine
Help
Preferences... Ctrl+G
Import Appliance... Ctrl+I
Export Appliance... Ctrl+E
Virtual Media Manager... Ctrl+D
Host Network Manager... Ctrl+H
Cloud Profile Manager... Ctrl+P
Network Operations Manager...
Check for Updates...
Reset All Warnings
Exit Ctrl+Q
New
Settings
Discard
Show
Preview

General

Name: cloudera-quickstart-vm-5.13.0-0-virtualbox

Operating System: Red Hat (64-bit)

Settings File Location: D:\cloudera-quickstart-vm-5.13.0-0-virtualbox\cloudera-quickstart-vm-5.13.0-0-virtualbox\cloudera-quickstart-vm-5.13.0-0-virtualbox

System

Base Memory: 4096 MB

Boot Order: Hard Disk, Optical

Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX, KVM Paravirtualization

Please choose a virtual appliance file to import
« clouder... › cloudera-quickstart-vm-5.13.0-0-virtualbox ›
Search cloudera-quickstart-v...
Organize
New folder
OneDrive
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Windows (C:)
LENOVO (D:)

Name	Date modified	Type	Size
cloudera-quickstart-vm-5.13.0-0-virtualb...	1/8/2020 4:02 PM	File folder	
cloudera-quickstart-vm-5.13.0-0-virtualb...	10/23/2017 4:29 PM	Open Virtualizatio...	15

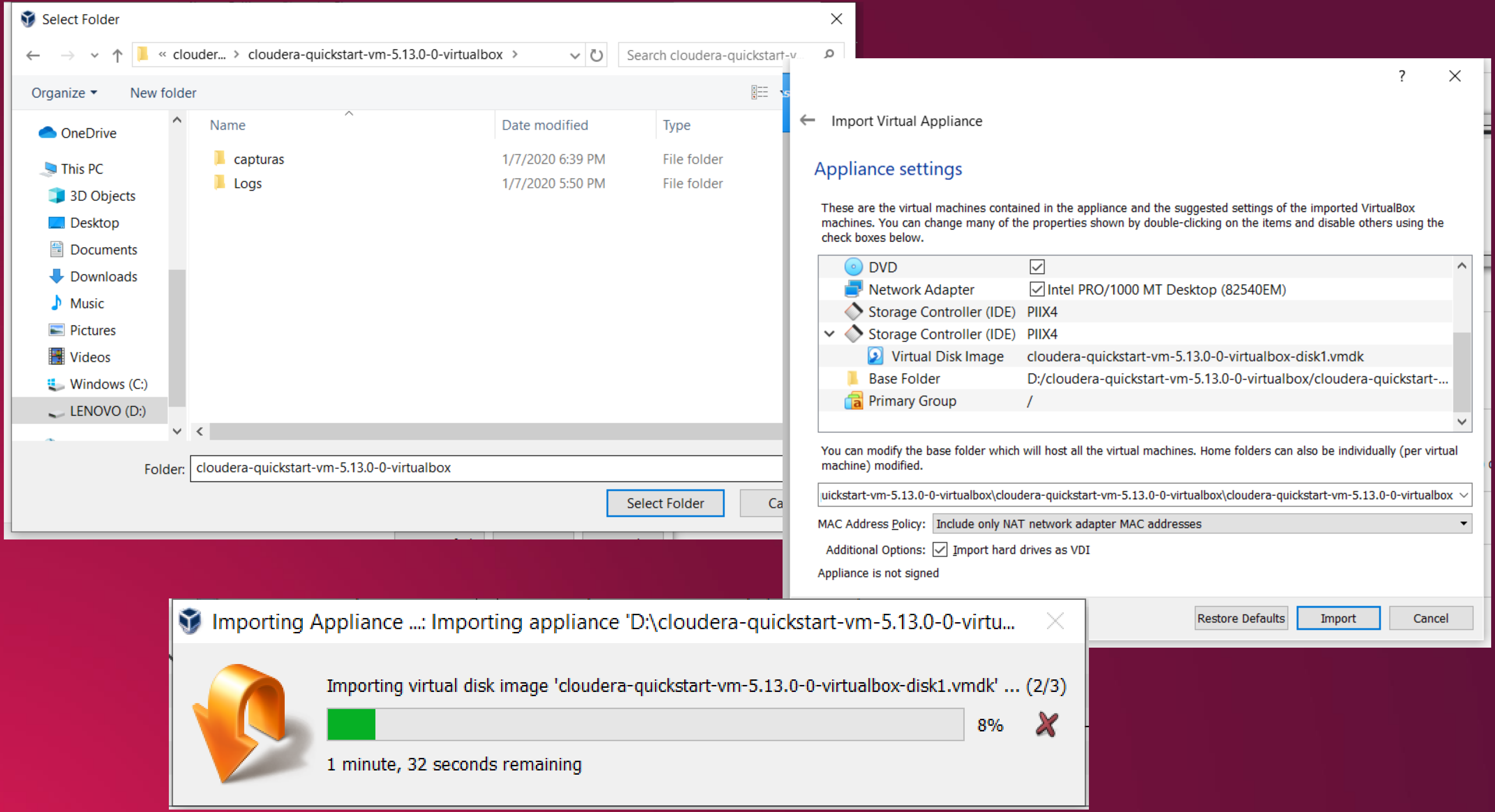
File name: cloudera-quickstart-vm-5.13.0-0-virtualbox

Open Virtualization Format (*.o

Appliance to import

VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.

art-vm-5.13.0-0-virtualbox\cloudera-quickstart-vm-5.13.0-0-virtualbox\cloudera-quickstart-vm-5.13.0-0-virtualbox.ovf



Select Folder

« clouder... › cloudera-quickstart-vm-5.13.0-0-virtualbox ›

Search cloudera-quickstart-v...

Name	Date modified	Type
capturas	1/7/2020 6:39 PM	File folder
Logs	1/7/2020 5:50 PM	File folder

Folder: cloudera-quickstart-vm-5.13.0-0-virtualbox

Import Virtual Appliance

Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

DVD	☑
Network Adapter	☑ Intel PRO/1000 MT Desktop (82540EM)
Storage Controller (IDE)	PIIX4
Storage Controller (IDE)	PIIX4
Virtual Disk Image	cloudera-quickstart-vm-5.13.0-0-virtualbox-disk1.vmdk
Base Folder	D:/cloudera-quickstart-vm-5.13.0-0-virtualbox/cloudera-quickstart-...
Primary Group	/

You can modify the base folder which will host all the virtual machines. Home folders can also be individually (per virtual machine) modified.

uickstart-vm-5.13.0-0-virtualbox\cloudera-quickstart-vm-5.13.0-0-virtualbox\cloudera-quickstart-vm-5.13.0-0-virtualbox

MAC Address Policy: Include only NAT network adapter MAC addresses

Additional Options: ☑ Import hard drives as VDI

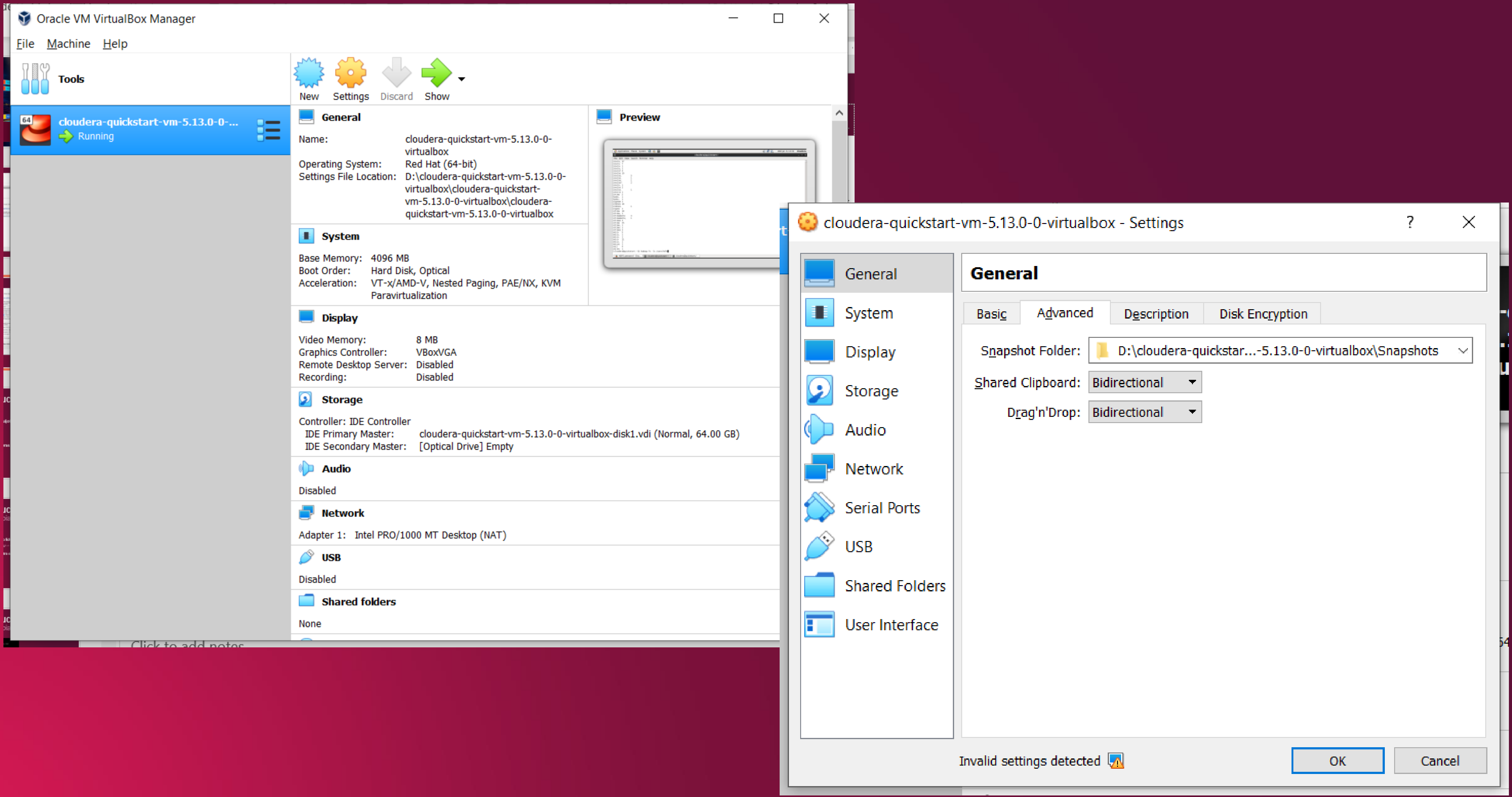
Appliance is not signed

Importing Appliance ...: Importing appliance 'D:\cloudera-quickstart-vm-5.13.0-0-virtu...

Importing virtual disk image 'cloudera-quickstart-vm-5.13.0-0-virtualbox-disk1.vmdk' ... (2/3)

8%

1 minute, 32 seconds remaining



Oracle VM VirtualBox Manager

cloudera-quickstart-vm-5.13.0-0-...
Running

General

Name: cloudera-quickstart-vm-5.13.0-0-virtualbox
Operating System: Red Hat (64-bit)
Settings File Location: D:\cloudera-quickstart-vm-5.13.0-0-virtualbox\cloudera-quickstart-vm-5.13.0-0-virtualbox\cloudera-quickstart-vm-5.13.0-0-virtualbox

Preview

System

Base Memory: 4096 MB
Boot Order: Hard Disk, Optical
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX, KVM Paravirtualization

Display

Video Memory: 8 MB
Graphics Controller: VBoxVGA
Remote Desktop Server: Disabled
Recording: Disabled

Storage

Controller: IDE Controller
IDE Primary Master: cloudera-quickstart-vm-5.13.0-0-virtualbox-disk1.vdi (Normal, 64.00 GB)
IDE Secondary Master: [Optical Drive] Empty

Audio

Disabled

Network

Adapter 1: Intel PRO/1000 MT Desktop (NAT)

USB

Disabled

Shared folders

None

Click to add notes

cloudera-quickstart-vm-5.13.0-0-virtualbox - Settings

General

Snapshot Folder: D:\cloudera-quickstar...-5.13.0-0-virtualbox\Snapshots
Shared Clipboard: Bidirectional
Drag'n'Drop: Bidirectional

Invalid settings detected

Práctica MapReduce

Contando palabras

1. Obtendremos una copia de “El Quijote” en txt.

2. Aplicaremos MapReduce.

3. Obtendremos el número de palabras que contiene nuestra copia del “El Quijote”.

Práctica MapReduce

Descargando nuestra copia.

1. Creamos un directorio llamado /quijote

   ```
   cd ../..
   mkdir quijote
   ```

2. Crearemos un script para descargar nuestro fichero descarga.sh

   ```
   nano descarga.sh
   curl http://www.gutenberg.org/cache/epub/2000/pg2000.txt -o quijote.txt
   ```

3. Establecemos permisos para ejecutar nuestro script

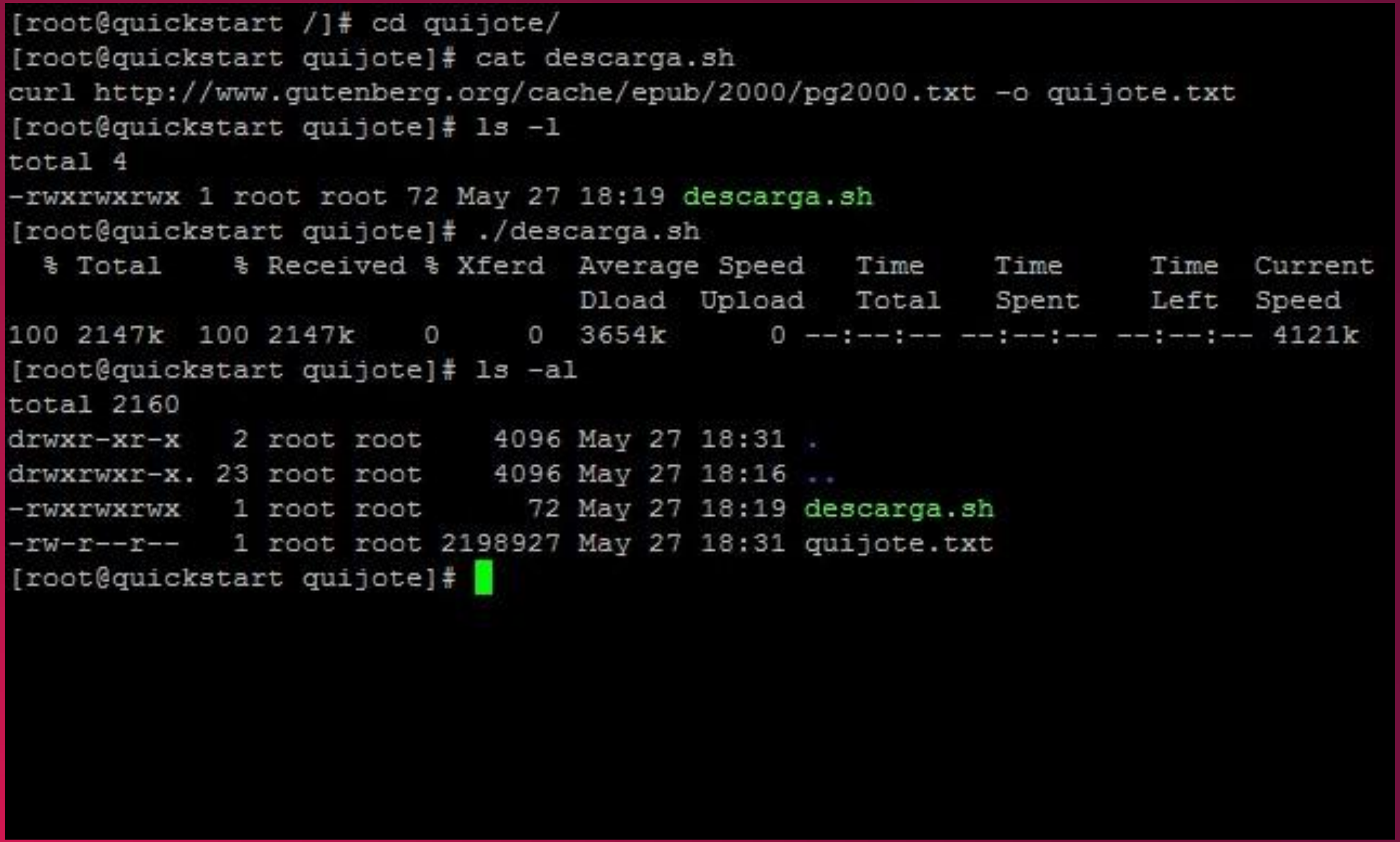
   ```
   chmod 777 descargar.sh
   ```

4. Ejecutamos nuestro script

   ```
   ./descarga.sh
   ```

Práctica MapReduce

Descargando nuestra copia.



```
[root@quickstart /]# cd quijote/
[root@quickstart quijote]# cat descarga.sh
curl http://www.gutenberg.org/cache/epub/2000/pg2000.txt -o quijote.txt
[root@quickstart quijote]# ls -l
total 4
-rwxrwxrwx 1 root root 72 May 27 18:19 descarga.sh
[root@quickstart quijote]# ./descarga.sh
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 2147k  100 2147k    0     0  3654k      0 --:--:-- --:--:-- --:--:-- 4121k
[root@quickstart quijote]# ls -al
total 2160
drwxr-xr-x   2 root root    4096 May 27 18:31 .
drwxrwxr-x. 23 root root    4096 May 27 18:16 ..
-rwxrwxrwx   1 root root      72 May 27 18:19 descarga.sh
-rw-r--r--   1 root root 2198927 May 27 18:31 quijote.txt
[root@quickstart quijote]# 
```

Práctica MapReduce

Estableciendo quijote.txt en HDFS

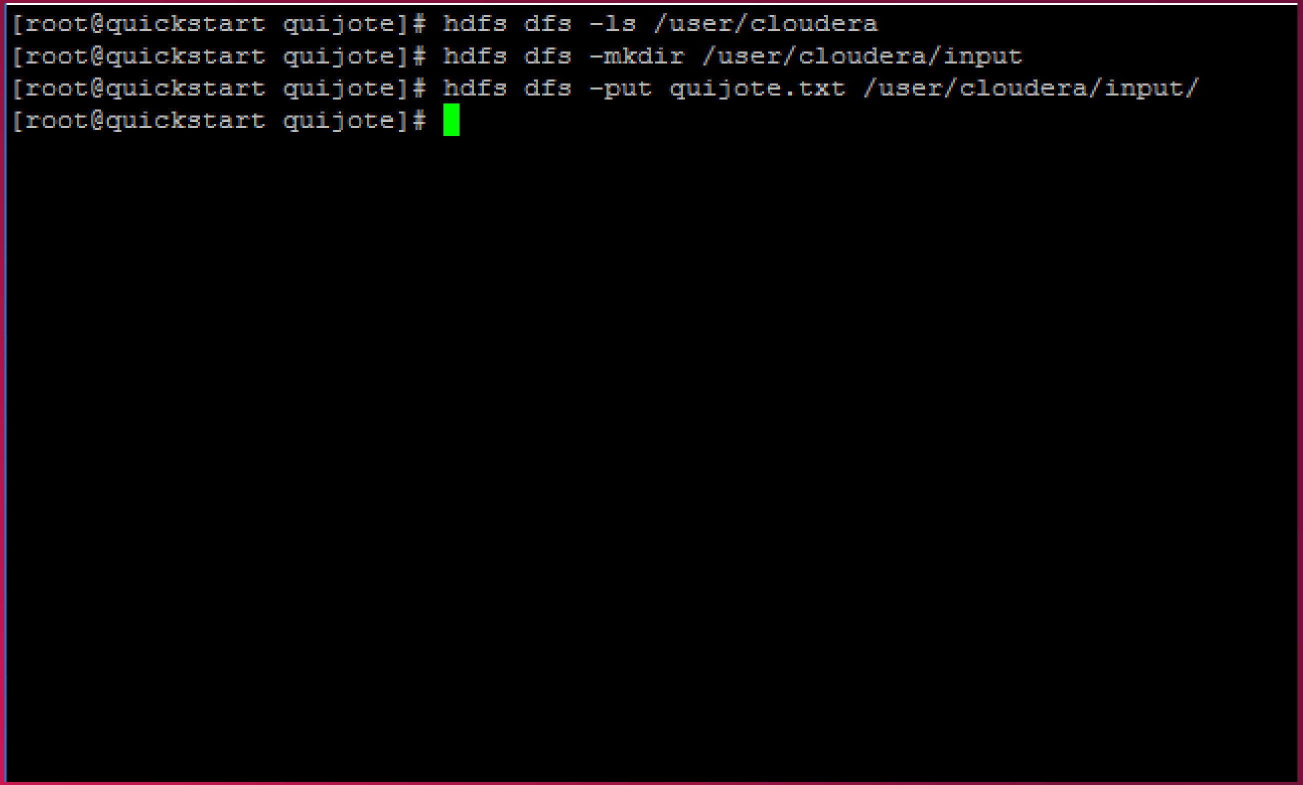
1. hdfs dfs -ls /user/cloudera

2. hdfs dfs -mkdir /user/cloudera/input

3. hdfs dfs -put quijote.txt /user/cloudera/input/

Práctica MapReduce

Estableciendo quijote.txt en HDFS



```
[root@quickstart quijote]# hdfs dfs -ls /user/cloudera
[root@quickstart quijote]# hdfs dfs -mkdir /user/cloudera/input
[root@quickstart quijote]# hdfs dfs -put quijote.txt /user/cloudera/input/
[root@quickstart quijote]#
```

Práctica MapReduce

Crear mapper.py –

nano mapper.py

Guardar: ctrl + o
Salir: ctrl + x

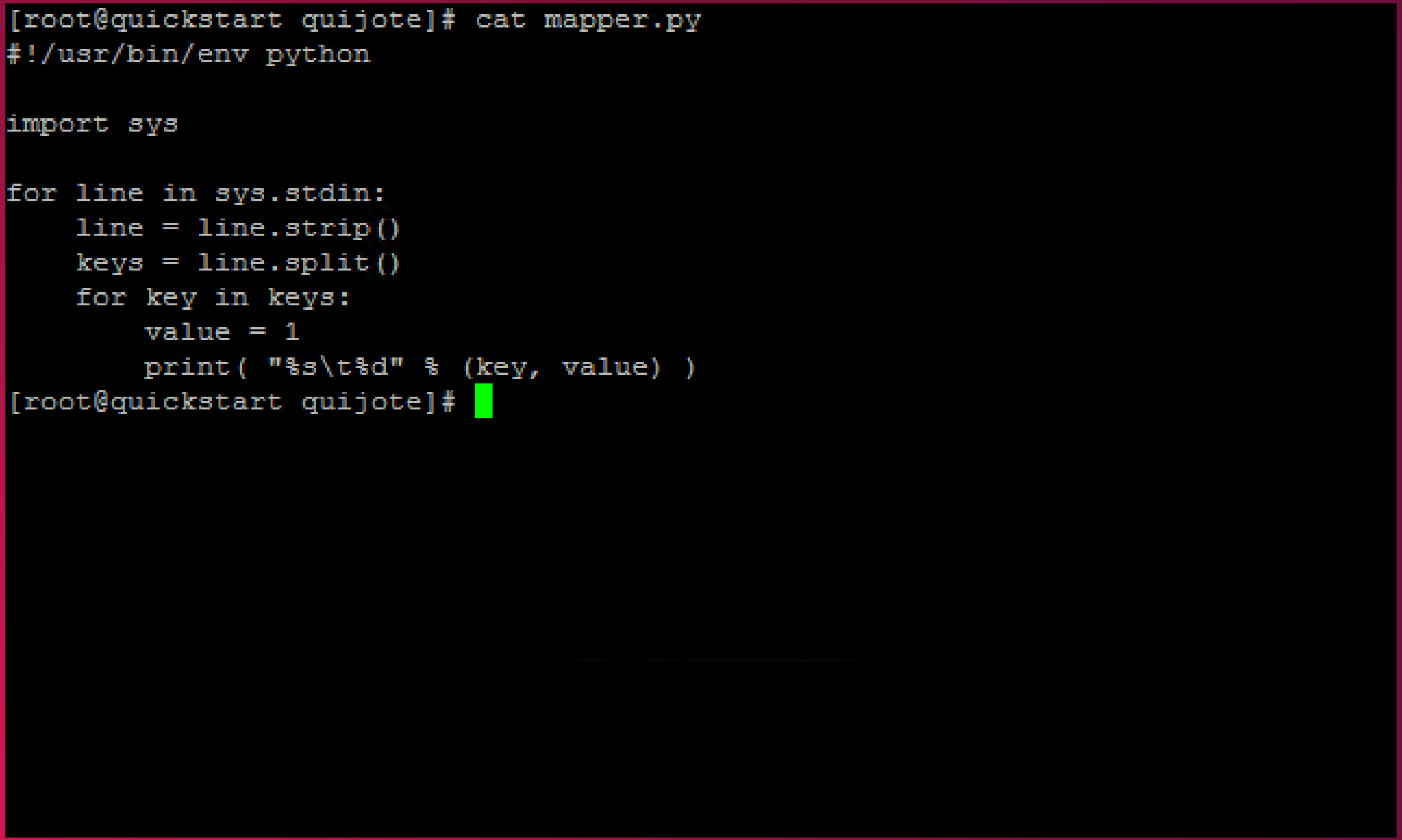
```
#!/usr/bin/env python

import sys

for line in sys.stdin:
    line = line.strip()
    keys = line.split()
    for key in keys:
        value = 1
        print( "%s\t%d" % (key, value) )
```

Práctica MapReduce

Creando nuestro mapper.py



```
[root@quickstart quijote]# cat mapper.py
#!/usr/bin/env python

import sys

for line in sys.stdin:
    line = line.strip()
    keys = line.split()
    for key in keys:
        value = 1
        print( "%s\t%d" % (key, value) )
[root@quickstart quijote]# 
```

Práctica MapReduce

Crear reducer.py – nano reducer.py

```
#!/usr/bin/env python


import sys


last_key = None
running_total = 0


for input_line in sys.stdin:  input_line =
   input_line.strip()
   this_key, value = input_line.split("\t", 1)
   value = int(value)
```

Práctica MapReduce

Creando nuestro reducer.py

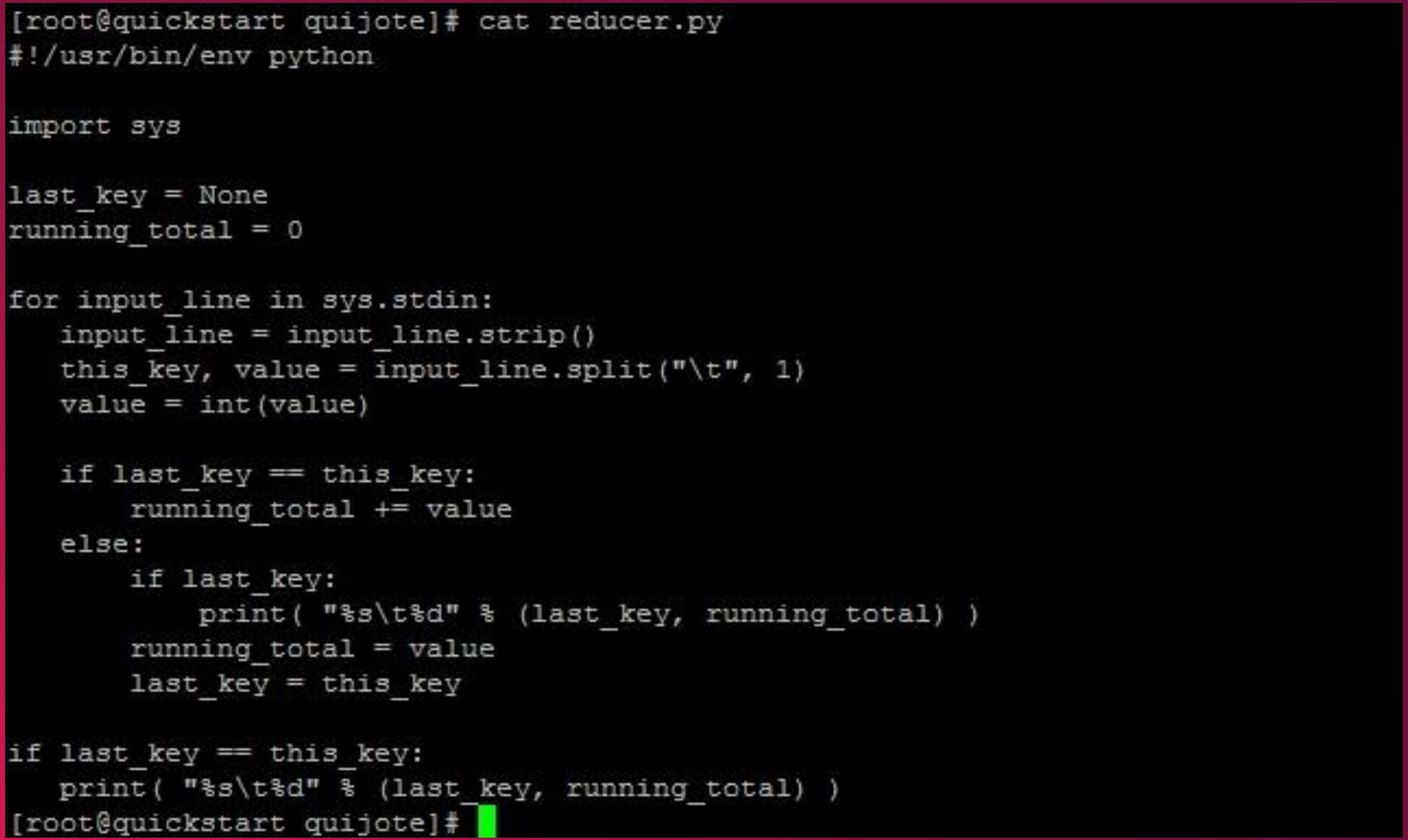
Guardar: ctrl + o
Salir: ctrl + x

```
      if last_key == this_key:
       running_total += value
   else:
       if last_key:
           print( "%s\t%d" % (last_key, running_total) )
       running_total = value
       last_key = this_key

if last_key == this_key:
   print( "%s\t%d" % (last_key, running_total) )
```

Práctica MapReduce

Creando nuestro reducer.py



```
[root@quickstart quijote]# cat reducer.py
#!/usr/bin/env python

import sys

last_key = None
running_total = 0

for input_line in sys.stdin:
   input_line = input_line.strip()
   this_key, value = input_line.split("\t", 1)
   value = int(value)

   if last_key == this_key:
       running_total += value
   else:
       if last_key:
           print( "%s\t%d" % (last_key, running_total) )
       running_total = value
       last_key = this_key

if last_key == this_key:
   print( "%s\t%d" % (last_key, running_total) )
[root@quickstart quijote]#
```

Práctica MapReduce

Ejecutando nuestros archivos

1.
   ```
   chmod 777 *.py
   ```

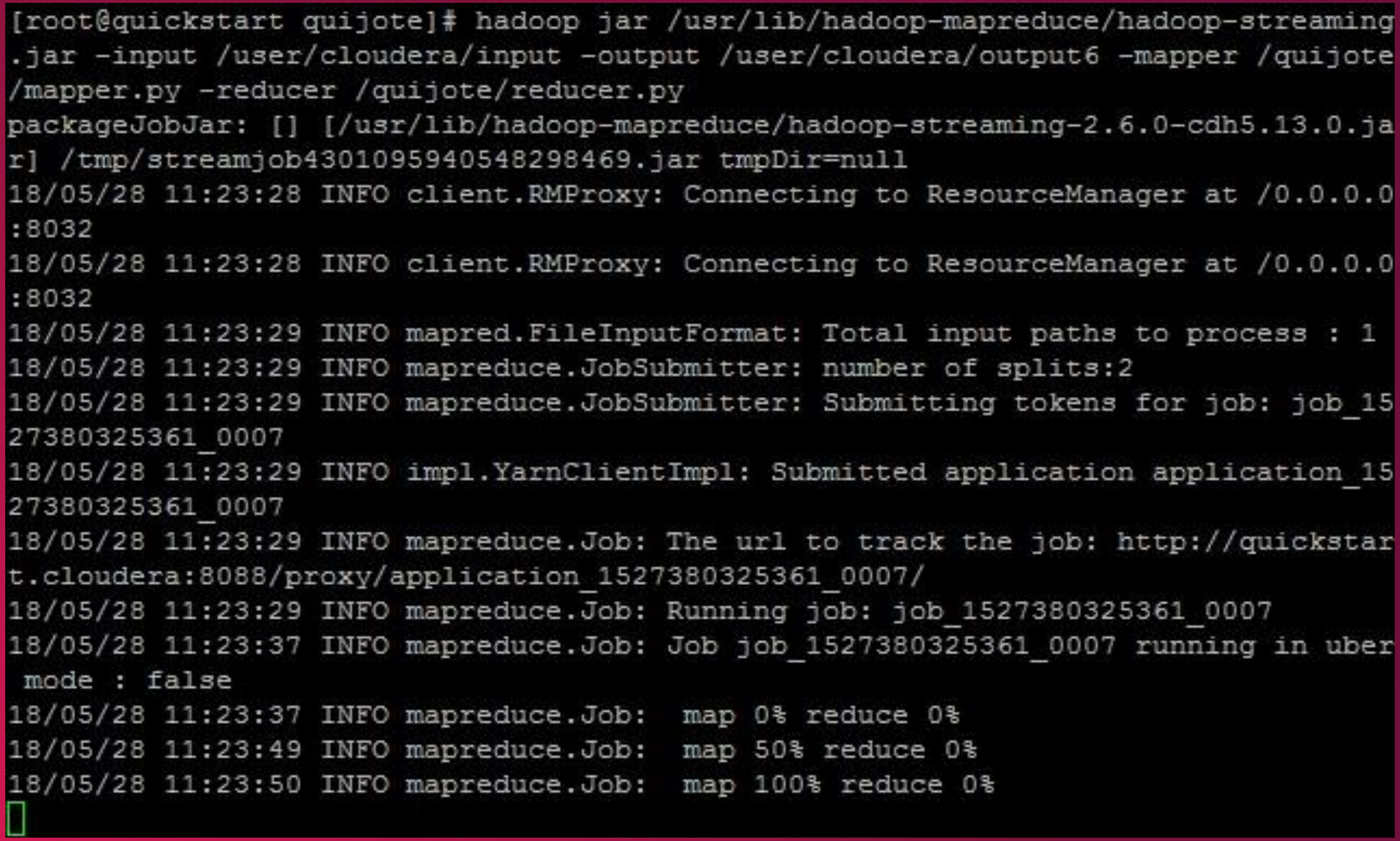
2.
   ```
   hdfs dfs -mkdir /user/cloudera/input
   ```

3.
   ```
   hdfs dfs -put quijote.txt /user/cloudera/input/
   ```

4.
   ```
   hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -input /user/cloudera/input -output /user/cloudera/output -mapper /quijote/mapper.py -reducer /quijote/reducer.py
   ```

Práctica MapReduce

Ejecutando nuestros archivos

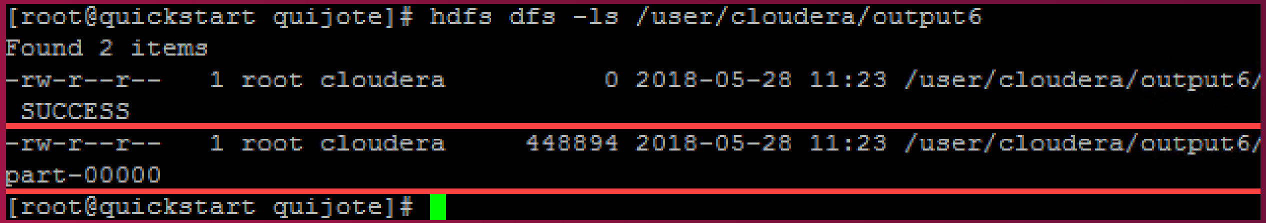


```
[root@quickstart quijote]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming
.jar -input /user/cloudera/input -output /user/cloudera/output6 -mapper /quijote
/mapper.py -reducer /quijote/reducer.py
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.13.0.ja
r] /tmp/streamjob4301095940548298469.jar tmpDir=null
18/05/28 11:23:28 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
18/05/28 11:23:28 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
18/05/28 11:23:29 INFO mapred.FileInputFormat: Total input paths to process : 1
18/05/28 11:23:29 INFO mapreduce.JobSubmitter: number of splits:2
18/05/28 11:23:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
27380325361_0007
18/05/28 11:23:29 INFO impl.YarnClientImpl: Submitted application application_15
27380325361_0007
18/05/28 11:23:29 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1527380325361_0007/
18/05/28 11:23:29 INFO mapreduce.Job: Running job: job_1527380325361_0007
18/05/28 11:23:37 INFO mapreduce.Job: Job job_1527380325361_0007 running in uber
 mode : false
18/05/28 11:23:37 INFO mapreduce.Job:  map 0% reduce 0%
18/05/28 11:23:49 INFO mapreduce.Job:  map 50% reduce 0%
18/05/28 11:23:50 INFO mapreduce.Job:  map 100% reduce 0%
```

Práctica MapReduce

Visualizando resultados

1. hdfs dfs -ls /user/cloudera/output



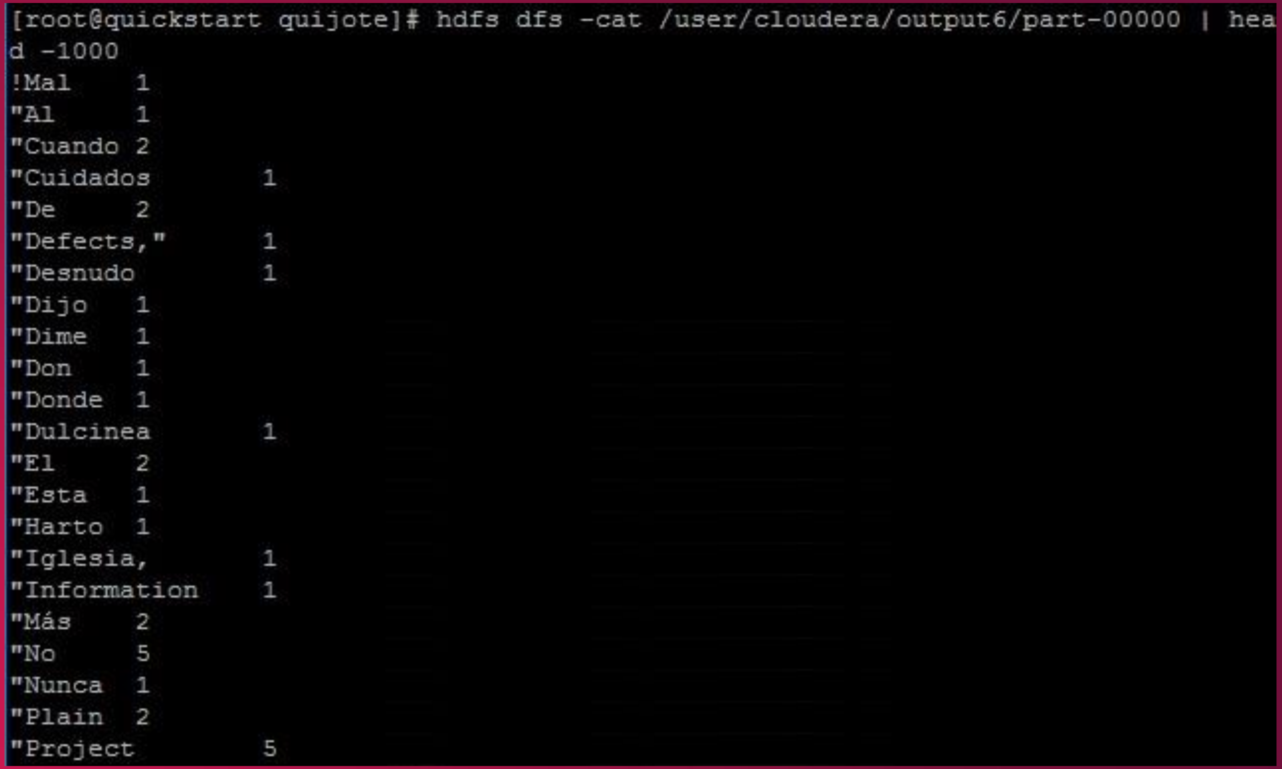
```
[root@quickstart quijote]# hdfs dfs -ls /user/cloudera/output6
Found 2 items
-rw-r--r--   1 root cloudera          0 2018-05-28 11:23 /user/cloudera/output6/
 SUCCESS
-rw-r--r--   1 root cloudera     448894 2018-05-28 11:23 /user/cloudera/output6/
part-00000
[root@quickstart quijote]#
```

Práctica MapReduce

Visualizando resultados

1. hdfs dfs -cat /user/cloudera/output/part-00000 | head -1000

2. hdfs dfs -cat /user/cloudera/output/*



```
[root@quickstart quijote]# hdfs dfs -cat /user/cloudera/output6/part-00000 | hea
d -1000
!Mal    1
"Al     1
"Cuando 2
"Cuidados       1
"De     2
"Defects,"      1
"Desnudo        1
"Dijo   1
"Dime   1
"Don    1
"Donde  1
"Dulcinea       1
"El     2
"Esta   1
"Harto  1
"Iglesia,       1
"Information    1
"Más    2
"No     5
"Nunca  1
"Plain  2
"Project        5
```