

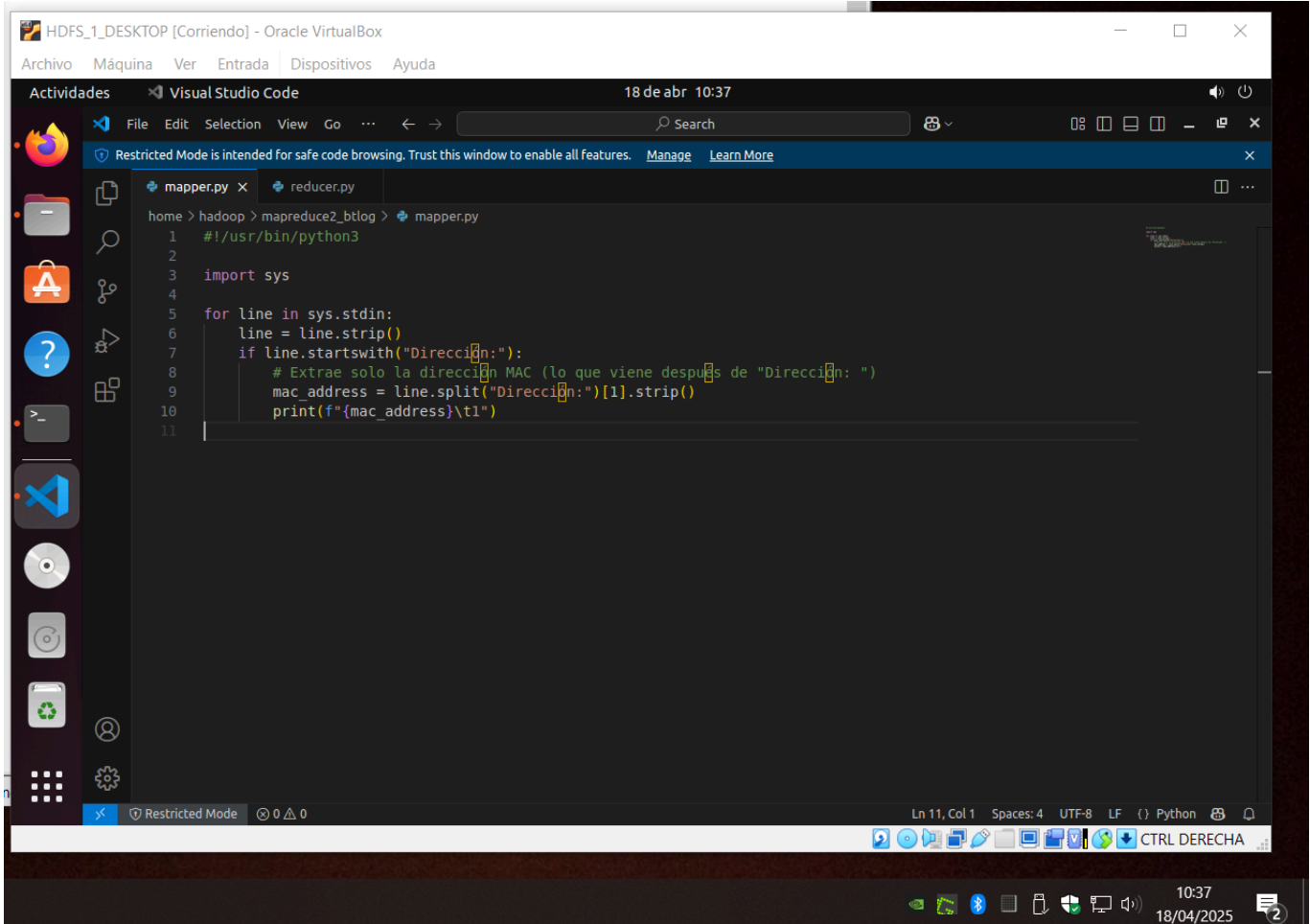
UD4. MapReduce y Spark

"Contador de MACs"

NOMBRE Y APELLIDOS: MARIO REY BULLIDO
DNI: 39459575Q

1.- Crea un archivo mapper.py que procese el archivo de la captura bluetooth que tienes en el aula virtual. Debe funcionar de una manera muy parecida al contador de palabras con la modificación que no todas las líneas son significativas. La única que nos interesa en este problema es la línea en la que aparece la MAC capturada en la que procesaremos única y exclusivamente la MAC, excluyendo la etiqueta "Dirección:". Como respuesta a esta pregunta muestra tu código comentado.

El proceso recibe por el standard input los registros de los ficheros de log y únicamente procesa las líneas que empiezan por la cadena "Dirección:", escribiendo por pantalla y devolviendo el valor de la MAC y un 1 para indicar que se ha encontrado este dato una vez:



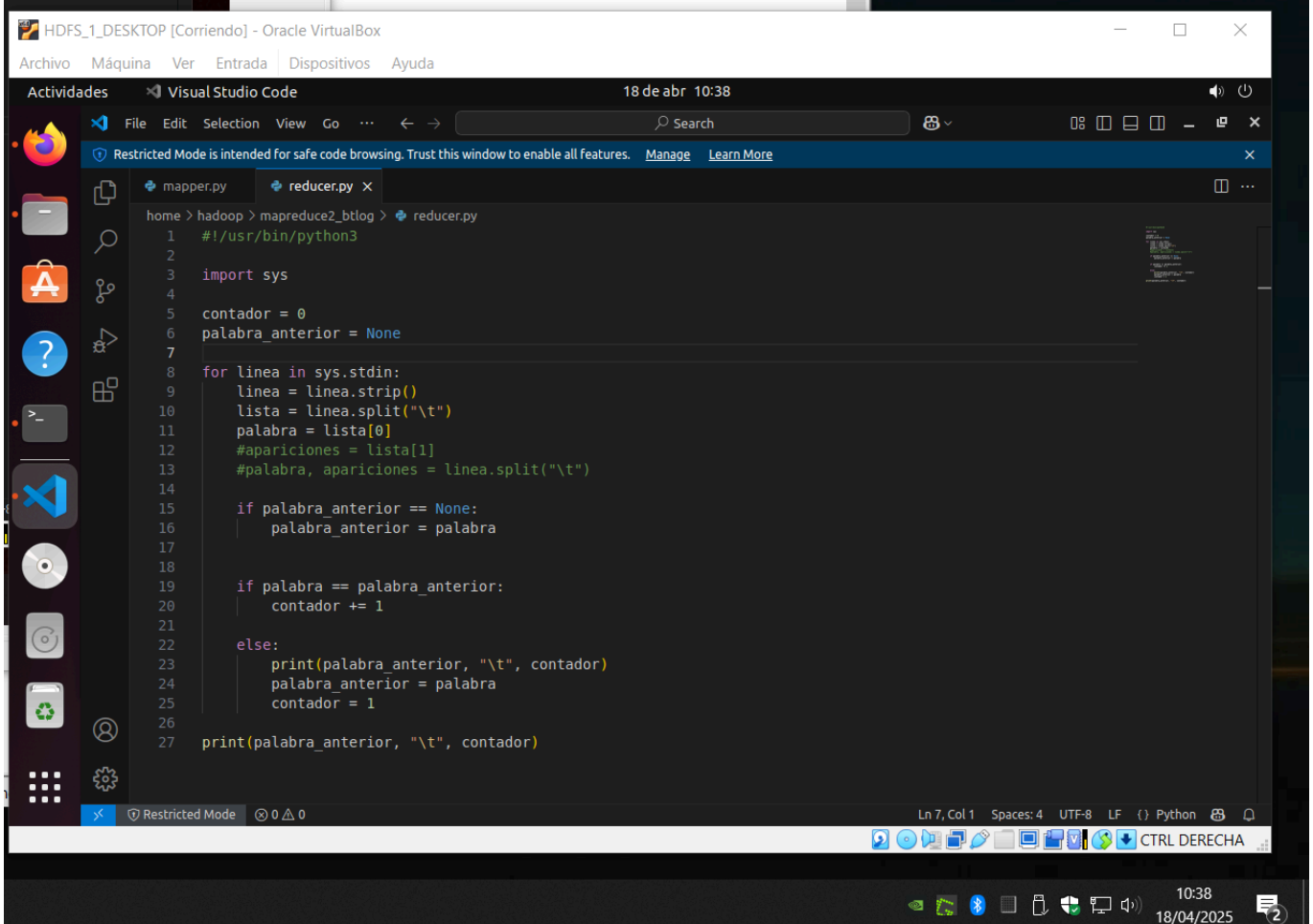
```
home > hadoop > mapreduce2_btlog > mapper.py
1  #!/usr/bin/python3
2
3  import sys
4
5  for line in sys.stdin:
6      line = line.strip()
7      if line.startswith("Dirección:"):
8          # Extrae solo la dirección MAC (lo que viene después de "Dirección: ")
9          mac_address = line.split("Dirección:")[1].strip()
10         print(f"{mac_address}\t1")
11
```

2.- (Opcional) Si tienes Hadoop funcionando no será necesario que pases por esta fase ya que se encargará Hadoop de hacerlo, puedes pasar al siguiente enunciado. Si tienes que ejecutar en local adapta tu código y aplica el archivo Python “ordenar.py” al archivo “salida_mapper.txt” para ordenar los resultados en un nuevo archivo “entrada_reducer.txt”. Muestra la salida del comando “head -n 20 entrada_reducer.txt” para ver las 20 primeras líneas del resultado.

3.- Crea un archivo llamado reducer.py nos devuelva el número de apariciones de cada MAC. Como respuesta muestra tu código comentado.

He reutilizado el script de contador de palabras ya que su lógica es exactamente la que necesitamos en este ejercicio cambiando el intérprete a python3.

Lee los resultados de mapper y va sumando las ocurrencias de la misma palabra (MAC) hasta que aparece una nueva y retorna el valor total de ocurrencias de esa palabra.



```
home > hadoop > mapreduce2_btlag > reducer.py
1  #!/usr/bin/python3
2
3  import sys
4
5  contador = 0
6  palabra_anterior = None
7
8  for linea in sys.stdin:
9      linea = linea.strip()
10     lista = linea.split("\t")
11     palabra = lista[0]
12     #apariciones = lista[1]
13     #palabra, apariciones = linea.split("\t")
14
15     if palabra_anterior == None:
16         palabra_anterior = palabra
17
18
19     if palabra == palabra_anterior:
20         contador += 1
21
22     else:
23         print(palabra_anterior, "\t", contador)
24         palabra_anterior = palabra
25         contador = 1
26
27 print(palabra_anterior, "\t", contador)
```

4.- Prueba tu mapper y reducer en un clúster Hadoop con HDFS y YARN. Indica el comando que usas para ver los resultados y realiza una captura de pantalla de parte del resultado. *Si no tienes el clúster Hadoop con HDFS y YARN funcionando modifica el enunciado de la práctica para usar archivos intermedios en su lugar. Igual que en la primera práctica.

```
hdfs dfs -head /salida_btlog2/part-00000          yarn jar
/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -input /bt_log/* -mapper
mapper.py -file mapper.py -reducer reducer.py -file reducer.py -output /salida_btlog2
```

Inicialmente lancé el proceso tal cual pero la máquina colapsaba porque estaba consumiendo muchos recursos de RAM y de disco duro.

Para probar que el proceso en sí funcionaba, lo ejecuté únicamente para uno de los archivos de log y vi que el proceso finalizaba correctamente así que el problema estaba en la configuración de los recursos de los nodos.

The screenshot shows a Hadoop web interface in a browser window titled 'HDFS_1_DESKTOP [Corriendo] - Oracle VirtualBox'. The browser address bar shows 'localhost:9870/explorer.html#/salida_btlog1'. The page title is 'Browse Directory' and the URL is '/salida_btlog1'. The page displays a table of files with columns: Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The table shows two files: '_SUCCESS' (0 B, Apr 18 11:14, 2, 128 MB) and 'part-00000' (11.1 KB, Apr 18 11:14, 2, 128 MB). A terminal window is overlaid on the bottom right, showing the command: 'yarn jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -input /bt_log/bluetooth_log -mapper mapper.py -file mapper.py -reducer reducer.py -file reducer.py -output /salida_btlog1'.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Apr 18 11:14	2	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	11.1 KB	Apr 18 11:14	2	128 MB	part-00000

```
hadoop@mario1: ~/mapreduce2_btlog
hadoop@mario1:~/mapreduce2_btlog$ yarn jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -input /bt_log/bluetooth_log -mapper mapper.py -file mapper.py -reducer reducer.py -file reducer.py -output /salida_btlog1
```

The screenshot shows a Hadoop web interface in a browser window. The browser address bar shows `localhost:9870/explorer.html#/salida_btlog1`. The page title is "Browse Directory". The search bar contains `/salida_btlog1`. The "Show" dropdown is set to "25" entries. The table below shows the directory contents:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Apr 18 11:14	2	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	11.1 KB	Apr 18 11:14	2	128 MB	part-00000

Overlaid on the bottom right is a terminal window titled `hadoop@mario1: ~/mapreduce2_btlog`. It displays a list of hexadecimal strings and their corresponding replication counts:

```
0E:A8:AD:81:8E:0B 2
0F:29:A4:78:36:5F 1
0F:35:78:D4:4A:8E 5
0F:EF:6B:64:D8:E4 10
0F:FB:03:62:8B:98 10
11:0D:1F:0D:39:4B 2
11:33:85:BC:3A:6A 1
12:24:2C:12:34:BC 12
12:D3:79:EE:C5:02 1
14:13:0B:38:43:87 45
14:49:E4:55:0B:A6 8
14:BF:9D:97:52:74 1
15:0A:97:91:64:80 7
15:18:8C:8C:A9:4A 1
17:C3:59:B5:05:20 1
19:09:99:33:3D:9B 4
1A:A4:A5:47:A3:5E 4
1A:D7:82:A4:D0:68 7
1A:EC:56:B4:67:4D 2
1B:5A:4B:7B:1D:74 3
1D:1C:C2:B4:9B:07 12
1E:1C:93:5C:C9:D7 1
1E:47:C0:B6:D1:A1 8
hadoop@mario1:~/mapreduce2_btlog$
```

Observando el panel WEB de YARN detecté que los nodos estaban configurados para usar 8 GB de RAM y 16 vcores cada uno así que esto estaba produciendo que se consumieran todos los recursos de cada máquina hasta colapsar.

Para solucionar este problema configuré YARN para que usase un poco menos de los recursos disponibles en cada máquina (2GB RAM y 2 vcores) como muestra la siguiente captura:

GNU nano 6.2 hadoop/etc/hadoop/yarn-site.xml

```
<value>mapreduce_shuffle</value>
</property>

<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>marid1</value>
</property>

<property>
  <name>yarn.nodemanager.resource.memory-mb</name>
  <value>1800</value>
</property>

<property>
  <name>yarn.scheduler.maximum-allocation-mb</name>
  <value>1600</value>
</property>

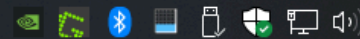
<property>
  <name>yarn.scheduler.minimum-allocation-mb</name>
  <value>512</value>
</property>

<property>
  <name>yarn.nodemanager.resource.cpu-vcores</name>
  <value>2</value>
</property>

<property>
  <name>yarn.scheduler.maximum-allocation-vcores</name>
  <value>2</value>
</property>
```

Help Write Out Where Is Cut Execute Location M-U Undo
Exit Read File Replace Paste Justify Go To Line M-E Redo

CTRL DERECHA

11:47
18/04/2025

De este modo pude ejecutar el proceso para todos los logs y finalizó sin problemas:

The screenshot shows a Hadoop web interface in a browser window titled "HDFS_1_DESKTOP [Corriendo] - Oracle VirtualBox". The browser address bar shows "mario1:8088/proxy/application_1744969378917_0001/". The page title is "MapReduce Application application_1744969378917_0001".

On the left, there is a sidebar with "Cluster" and "Application" sections. The "Application" section is expanded, showing "Active Jobs".

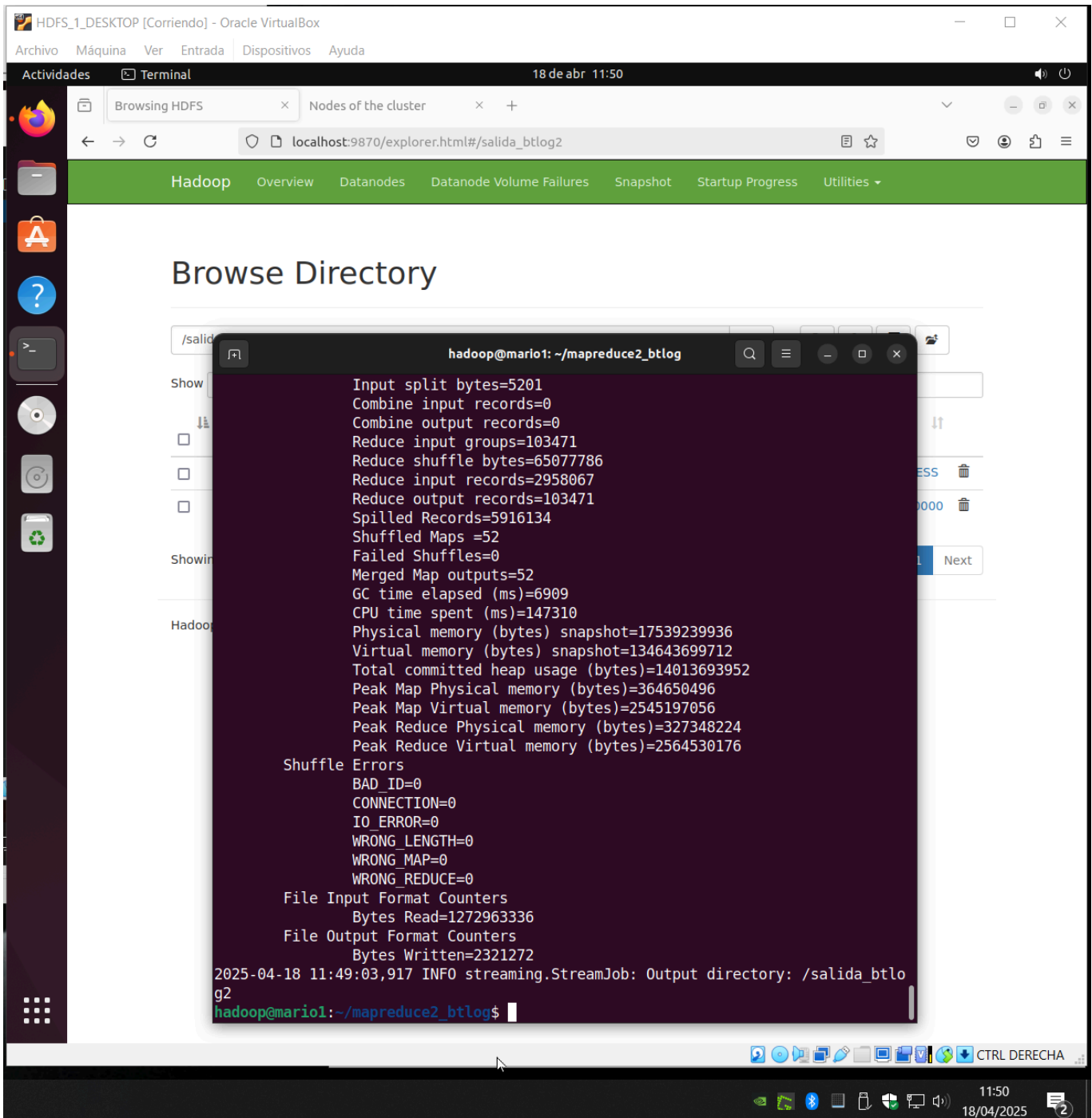
Job ID	Name	State	Map Progress	Maps Total	Maps Completed	Reduce Progress	Reduces Total	Reduces Completed
job_1744969378917_0001	streamjob8985707394889338696.jar	RUNNING	<div></div>	52	18	<div></div>	1	0

Below the table, it says "Showing 1 to 1 of 1 entries".

In the foreground, a terminal window titled "hadoop@mario1: ~/mapreduce2_btlog" displays the following logs:

```
s : 52
2025-04-18 11:43:48,911 INFO mapreduce.JobSubmitter: number of splits:52
2025-04-18 11:43:49,463 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1744969378917_0001
2025-04-18 11:43:49,464 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-18 11:43:49,855 INFO conf.Configuration: resource-types.xml not found
2025-04-18 11:43:49,856 INFO resource.ResourceUtils: Unable to find 'resource-ty
pes.xml'.
2025-04-18 11:43:50,544 INFO impl.YarnClientImpl: Submitted application applicat
ion_1744969378917_0001
2025-04-18 11:43:50,722 INFO mapreduce.Job: The url to track the job: http://mar
io1:8088/proxy/application_1744969378917_0001/
2025-04-18 11:43:50,728 INFO mapreduce.Job: Running job: job_1744969378917_0001
2025-04-18 11:44:02,151 INFO mapreduce.Job: Job job_1744969378917_0001 running i
n uber mode : false
2025-04-18 11:44:02,153 INFO mapreduce.Job: map 0% reduce 0%
2025-04-18 11:44:17,364 INFO mapreduce.Job: map 4% reduce 0%
2025-04-18 11:44:27,500 INFO mapreduce.Job: map 6% reduce 0%
2025-04-18 11:44:28,507 INFO mapreduce.Job: map 8% reduce 0%
2025-04-18 11:44:36,615 INFO mapreduce.Job: map 10% reduce 0%
2025-04-18 11:44:38,632 INFO mapreduce.Job: map 12% reduce 0%
2025-04-18 11:44:46,806 INFO mapreduce.Job: map 13% reduce 0%
2025-04-18 11:44:48,841 INFO mapreduce.Job: map 15% reduce 0%
2025-04-18 11:44:55,942 INFO mapreduce.Job: map 17% reduce 0%
2025-04-18 11:44:58,970 INFO mapreduce.Job: map 19% reduce 0%
2025-04-18 11:45:05,027 INFO mapreduce.Job: map 21% reduce 0%
2025-04-18 11:45:08,119 INFO mapreduce.Job: map 23% reduce 0%
2025-04-18 11:45:13,187 INFO mapreduce.Job: map 25% reduce 0%
2025-04-18 11:45:15,198 INFO mapreduce.Job: map 27% reduce 0%
2025-04-18 11:45:21,294 INFO mapreduce.Job: map 29% reduce 0%
2025-04-18 11:45:22,303 INFO mapreduce.Job: map 31% reduce 0%
2025-04-18 11:45:29,390 INFO mapreduce.Job: map 35% reduce 0%
2025-04-18 11:45:36,443 INFO mapreduce.Job: map 37% reduce 0%
```

The bottom of the screen shows a taskbar with various icons and a system clock indicating 11:45 on 18/04/2025.



HDFS_1_DESKTOP [Corriendo] - Oracle VirtualBox

Archivo Máquina Ver Entrada Dispositivos Ayuda

Actividades Terminal 18 de abr 11:52

Browsing HDFS Nodes of the cluster

localhost:9870/explorer.html#/salida_btlog2

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

hadoop@mario1: ~/mapreduce2_btlog

```
00:0A:9B:9A:9C:51 5
00:0A:9B:9A:9E:D6 1
00:0A:9B:9A:D3:7C 1
00:0A:9B:A1:9D:0D 1
00:0A:9B:A8:17:74 1
00:0A:9B:B1:C6:61 2
00:0A:9B:BC:3C:4A 6
00:0C:76:D2:27:D7 1
00:0C:DE:BB:5F:22 1
00:0E:32:88:14:8E 8
00:0E:E1:72:F3:45 32
00:10:24:E0:D8:36 12
00:11:23:33:4E:F5 40
00:12:6F:07:58:30 18
00:12:6F:08:77:4D 2
00:12:6F:09:A1:34 1
00:12:6F:09:A8:EA 2
00:12:6F:10:23:A4 2
00:12:6F:10:4E:22 24
00:12:6F:10:4F:9D 2
00:12:6F:60:CF:77 1
00:12:6F:60:F3:77 10
00:12:6F:60:FF:1D 4
00:12:6F:60:FF:B2 2
00:12:6F:61:16:91 5
00:12:6F:61:35:38 3
00:12:6F:61:6E:16 4
00:12:6F:61:75:E9 7
00:12:6F:61:B1:38 4
00:12:6F:61:C7:D7 1
00:12:6F:61:CC:15 1
00:12:6F:61:E1:EC 4
00:12:6F:61:E9:2B 11
00:12:6F:61:EF:E0
```

hadoop@mario1:~/mapreduce2_btlog\$

Name

[_SUCCESS](#)

[part-00000](#)

Previous **1** Next

CTRL DERECHA

11:52 18/04/2025