

APLICACIÓN DEL MAPREDUCE EN LA DISTRO UBUNTU POR EL VIRTUAL BOX

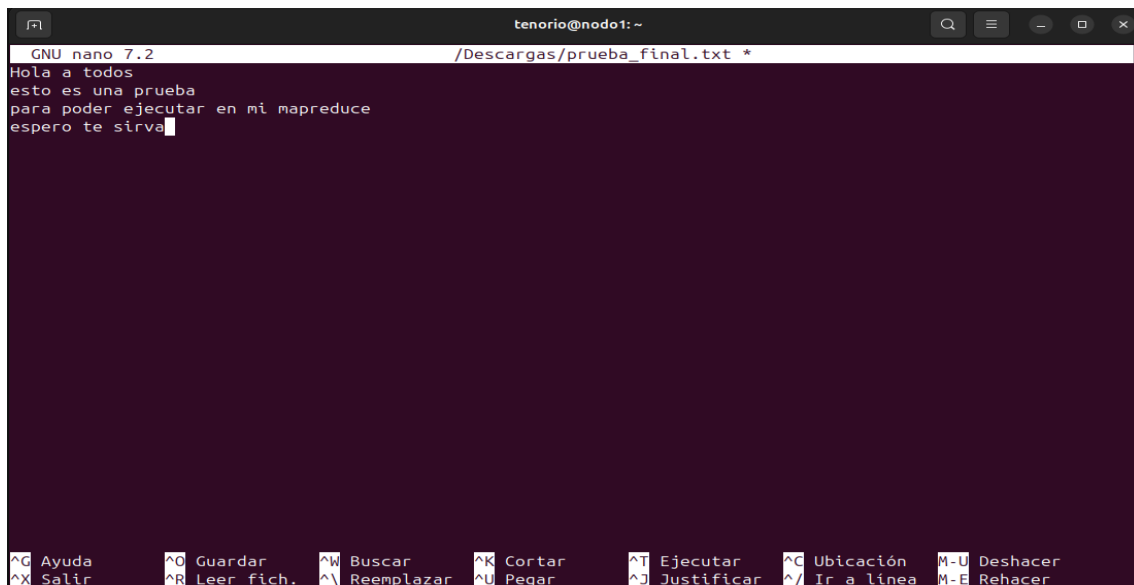
AUTOR: TENORIO ACHA RICHARD ANDERSON

Paso 1: tenemos que tener el dfs y el yarn corriendo en nuestra distro, para eso se usa los comandos “start-dfs.sh” y “start-yarn.sh” (por si un caso no lo tienes corriendo en tu distro) y nos aseguramos que tengamos los daemons con el comando “jps”

```
tenorio@nodo1:~/Descargas$ start-dfs.sh
Starting namenodes on [nodo1]
nodo1: starting namenode, logging to /opt/hadoop/logs/hadoop-tenorio-namenode-nodo1.out
localhost: starting datanode, logging to /opt/hadoop/logs/hadoop-tenorio-datanode-nodo1.out
nodo2: starting datanode, logging to /opt/hadoop/logs/hadoop-tenorio-datanode-nodo2.out
nodo3: starting datanode, logging to /opt/hadoop/logs/hadoop-tenorio-datanode-nodo3.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /opt/hadoop/logs/hadoop-tenorio-secondarynamenode-nodo1.out
```

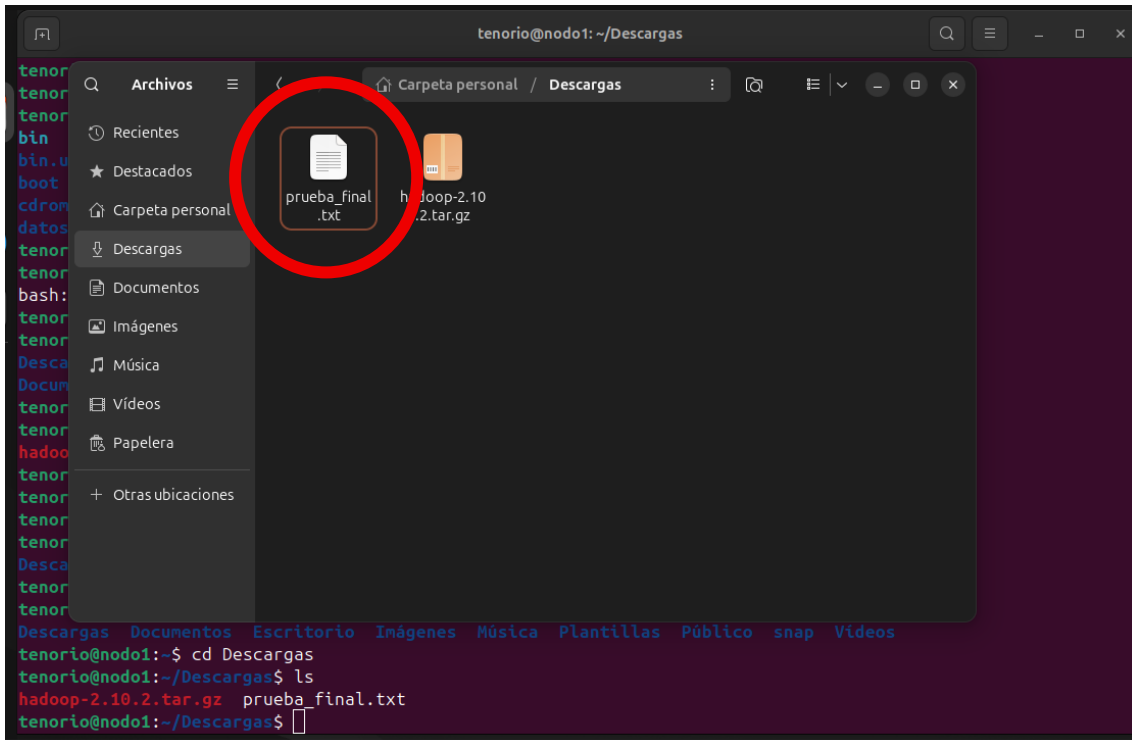
```
tenorio@nodo1:~/Descargas$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/logs/yarn-tenorio-resourcemanager-nodo1.out
nodo2: starting nodemanager, logging to /opt/hadoop/logs/yarn-tenorio-nodemanager-nodo2.out
localhost: starting nodemanager, logging to /opt/hadoop/logs/yarn-tenorio-nodemanager-nodo1.out
nodo3: starting nodemanager, logging to /opt/hadoop/logs/yarn-tenorio-nodemanager-nodo3.out
tenorio@nodo1:~/Descargas$ jps
8085 SecondaryNameNode
8230 ResourceManager
8375 NodeManager
7882 DataNode
7706 NameNode
8699 Jps
```

Paso 2: usamos el comando “nano Descargas/prueba_ultima.txt” saldrá un apartado como en la imagen y escribiremos dentro (en mi caso un ejemplo), con este comando crearemos un texto en nuestra carpeta Descargas, le damos “control + O” para guardar y “control + x” para salir, así crearemos nuestro archivo .txt



```
tenorio@nodo1: ~
GNU nano 7.2 /Descargas/prueba_final.txt *
Hola a todos
esto es una prueba
para poder ejecutar en mi mapreduce
espero te sirva
^G Ayuda      ^O Guardar    ^W Buscar     ^K Cortar     ^T Ejecutar   ^C Ubicación  M-U Deshacer
^X Salir      ^R Leer fich. ^_ Reemplazar  ^U Pegar      ^J Justificar ^/ Ir a línea   M-E Rehacer
```

```
tenorio@nodo1:~/Descargas$ ls
hadoop-2.10.2.tar.gz prueba_final.txt
tenorio@nodo1:~/Descargas$
```



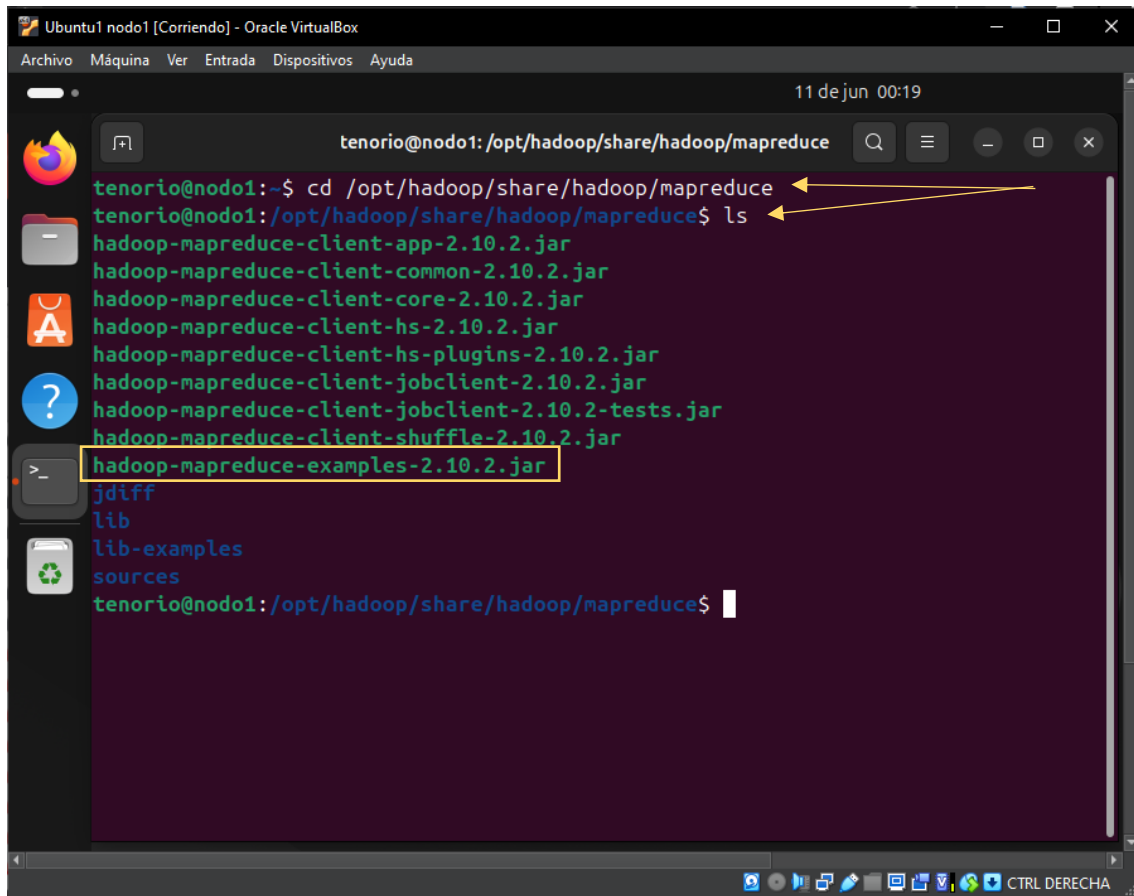
Paso 3: ahora creamos la carpeta “unajma” en la carpeta de Descargas, para eso usaremos el comando “hdfs dfs -mkdir /unajma” (nota: esto crea una carpeta en el sistema de archivos distribuidos de hadoop{HDFS}) y verificamos con el comando “ls -la”

```
tenorio@nodo1:~/Descargas$ ls -la
total 404920
drwxr-xr-x  2 tenorio tenorio    4096 jun 10 23:51 .
drwxr-x--- 16 tenorio tenorio    4096 jun  3 01:39 ..
-rw-rw-r--  1 tenorio tenorio 414624228 may  6 12:43 hadoop-2.10.2.tar.gz
-rw-rw-r--  1 tenorio tenorio      84 jun 10 23:51 prueba_final.txt
tenorio@nodo1:~/Descargas$
```

Paso 4: ahora subimos el archivo que creamos “prueba_final.txt” a la carpeta creada con el comando “hdfs dfs -put prueba_ultima.txt /unajma” Esto sube el archivo local a HDFS dentro de /unajma

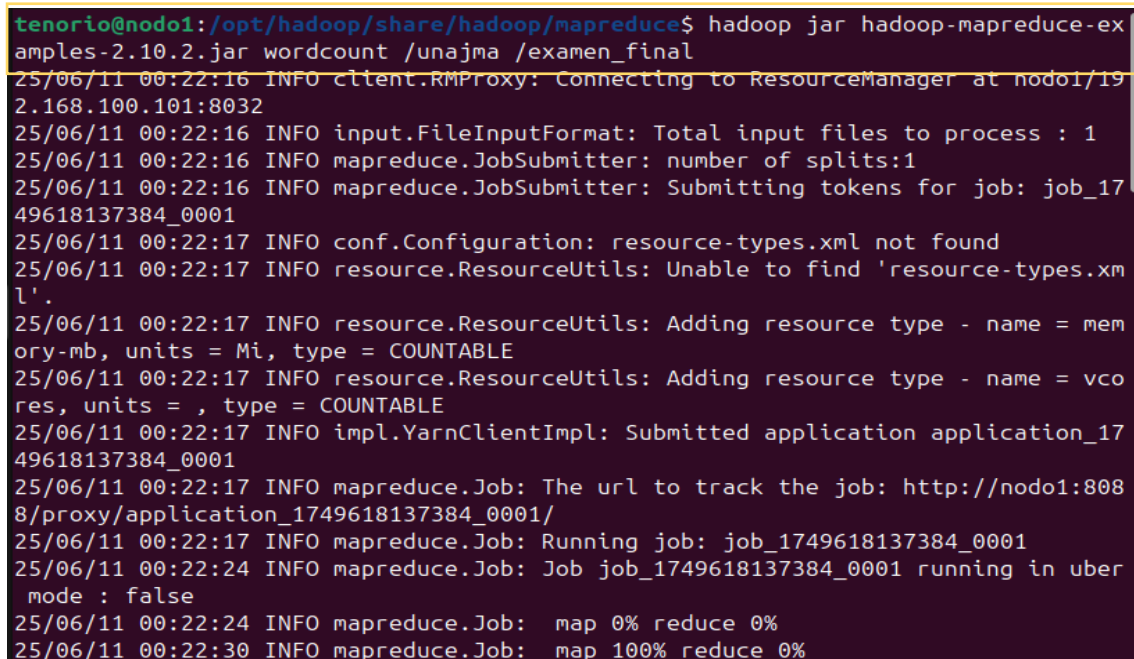
```
tenorio@nodo1:~/Descargas$ hdfs dfs -put prueba_final.txt /unajma
```

Paso 5: entramos al directorio para hacer correr nuestro mapreduce con el siguiente comando “cd /opt/hadoop/share/hadoop/mapreduce” y usamos el comando “ls” para verificar como se ve en la imagen



```
tenorio@nodo1:~$ cd /opt/hadoop/share/hadoop/mapreduce
tenorio@nodo1:/opt/hadoop/share/hadoop/mapreduce$ ls
hadoop-mapreduce-client-app-2.10.2.jar
hadoop-mapreduce-client-common-2.10.2.jar
hadoop-mapreduce-client-core-2.10.2.jar
hadoop-mapreduce-client-hs-2.10.2.jar
hadoop-mapreduce-client-hs-plugins-2.10.2.jar
hadoop-mapreduce-client-jobclient-2.10.2.jar
hadoop-mapreduce-client-jobclient-2.10.2-tests.jar
hadoop-mapreduce-client-shuffle-2.10.2.jar
hadoop-mapreduce-examples-2.10.2.jar
jdiffr
lib
lib-examples
sources
tenorio@nodo1:/opt/hadoop/share/hadoop/mapreduce$
```

Paso 6: usamos el siguiente comando para ejecutar el mapreduce “hadoop jar hadoop-mapreduce-examples-2.10.2.jar wordcount /unajma /examen_final”



```
tenorio@nodo1:/opt/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.10.2.jar wordcount /unajma /examen_final
25/06/11 00:22:16 INFO client.RMProxy: Connecting to ResourceManager at nodo1/192.168.100.101:8032
25/06/11 00:22:16 INFO input.FileInputFormat: Total input files to process : 1
25/06/11 00:22:16 INFO mapreduce.JobSubmitter: number of splits:1
25/06/11 00:22:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1749618137384_0001
25/06/11 00:22:17 INFO conf.Configuration: resource-types.xml not found
25/06/11 00:22:17 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
25/06/11 00:22:17 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
25/06/11 00:22:17 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
25/06/11 00:22:17 INFO impl.YarnClientImpl: Submitted application application_1749618137384_0001
25/06/11 00:22:17 INFO mapreduce.Job: The url to track the job: http://nodo1:8088/proxy/application_1749618137384_0001/
25/06/11 00:22:17 INFO mapreduce.Job: Running job: job_1749618137384_0001
25/06/11 00:22:24 INFO mapreduce.Job: Job job_1749618137384_0001 running in uber mode : false
25/06/11 00:22:24 INFO mapreduce.Job: map 0% reduce 0%
25/06/11 00:22:30 INFO mapreduce.Job: map 100% reduce 0%
```

```
25/06/11 00:22:35 INFO mapreduce.Job: map 100% reduce 100%
25/06/11 00:22:35 INFO mapreduce.Job: Job job_1749618137384_0001 completed successfully
25/06/11 00:22:35 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=186
    FILE: Number of bytes written=420897
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=190
    HDFS: Number of bytes written=116
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=3424
    Total time spent by all reduces in occupied slots (ms)=2331
    Total time spent by all map tasks (ms)=3424
    Total time spent by all reduce tasks (ms)=2331
    Total vcore-milliseconds taken by all map tasks=3424
```

```
Map-Reduce Framework
  Map input records=4
  Map output records=16
  Map output bytes=148
  Map output materialized bytes=186
  Input split bytes=106
  Combine input records=16
  Combine output records=16
  Reduce input groups=16
  Reduce shuffle bytes=186
  Reduce input records=16
  Reduce output records=16
  Spilled Records=32
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=171
  CPU time spent (ms)=1330
  Physical memory (bytes) snapshot=508706816
  Virtual memory (bytes) snapshot=3816669184
  Total committed heap usage (bytes)=319815680
```

Paso 7: Verificamos en el siguiente URL (en este caso en mi maquina es la siguiente)
“htt://nodo1:8088/cluster/apps”

The screenshot shows the Hadoop cluster management interface. The browser address bar displays 'http://nodo1:8088/cluster/apps'. The left sidebar menu has 'Applications' selected. The main content area shows 'Cluster Metrics' and 'Cluster Nodes Metrics'. Below these, there is a table of applications. The first application is highlighted.

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	Final
application_1749618137384_0001	tenorio	word count	MAPREDUCE	default	0	Wed Jun 11 00:22:17 -0500 2025	Wed Jun 11 00:22:17 -0500 2025	Wed Jun 11 00:22:33 -0500 2025	FINISHED	SUCCESS

Paso 8: le damos click en el apartado como indica la imagen

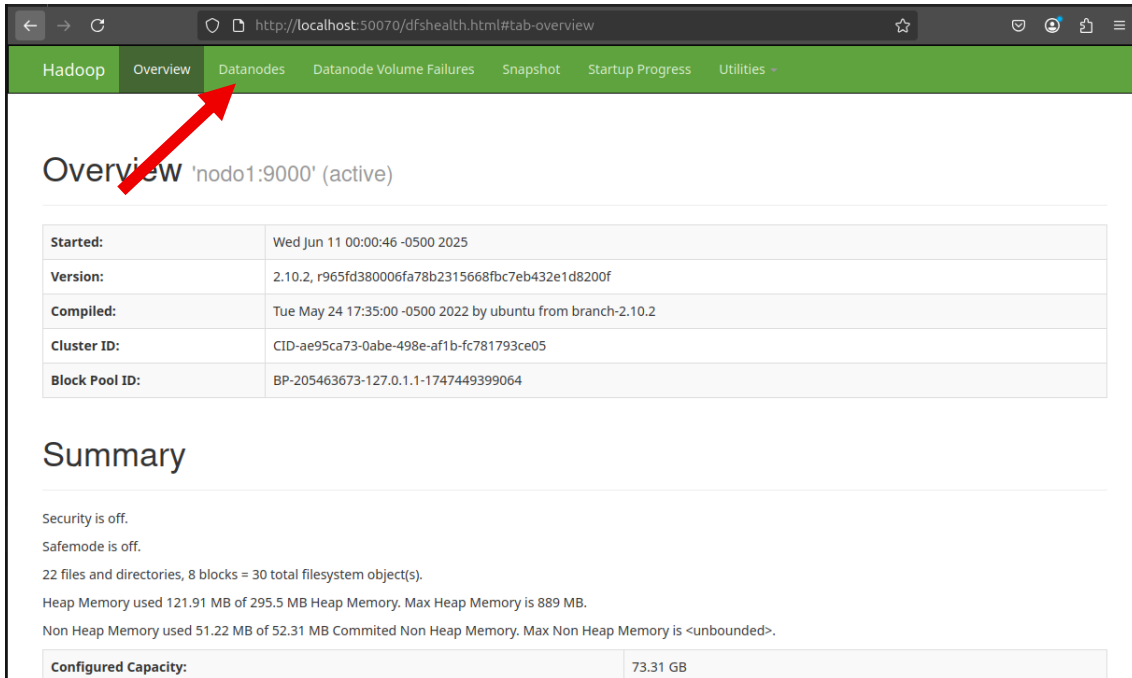
The screenshot shows the Hadoop cluster management interface. A red box highlights the application ID 'application_1749618137384_0001' in the table.

En este punto nos indica que solo fueron 16 segundos el tiempo de rendimiento, esto se debe a que el archivo es muy pequeño y contiene muy pocas letras, lo que hizo que el proceso durase ese tiempo estimado.

The screenshot shows the 'Application Overview' page for the application 'word count'. The 'Elapsed' time is 16sec.

Application Overview	
User:	tenorio
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	mié jun 11 00:22:17 -0500 2025
Launched:	mié jun 11 00:22:17 -0500 2025
Finished:	mié jun 11 00:22:33 -0500 2025
Elapsed:	16sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Paso 9: verificamos ahora en el siguiente URL <http://localhost:50070> y nos vamos al apartado de Datanodes, en este punto como estoy trabajando con 3 nodos (el maestro y 2 esclavos), me aparecerá los 3 en una barra como se muestra en la segunda imagen



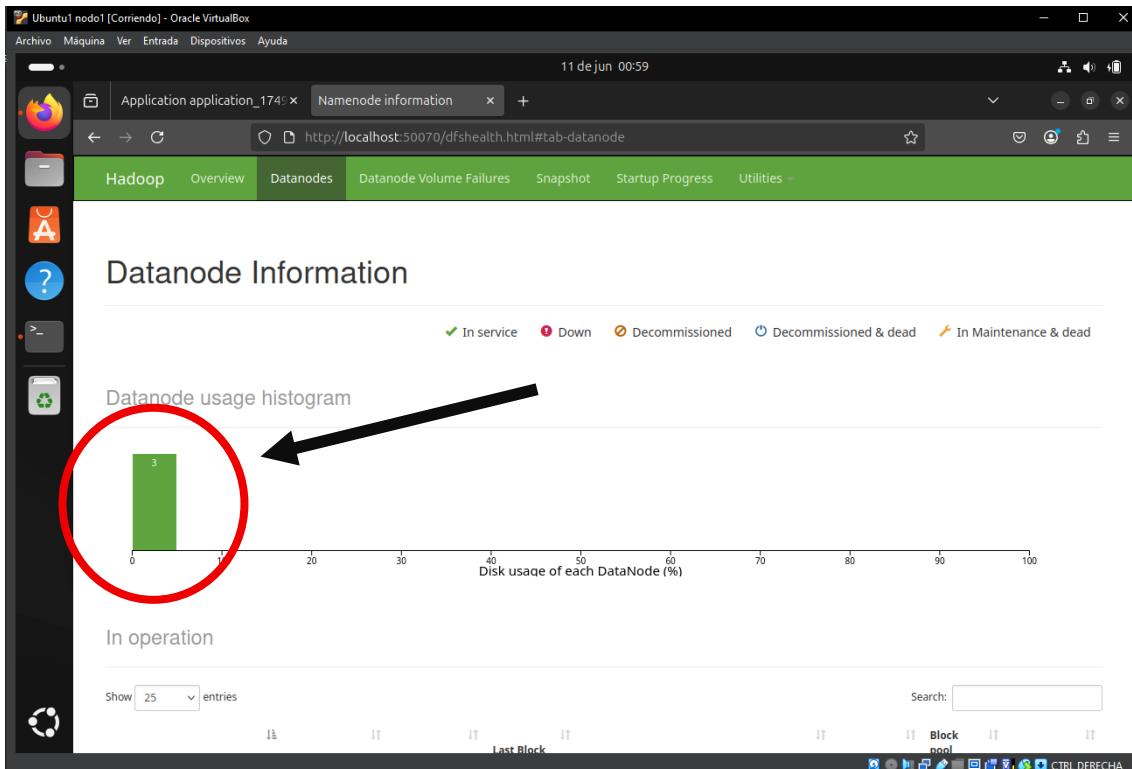
Overview 'nodo1:9000' (active)

Started:	Wed Jun 11 00:00:46 -0500 2025
Version:	2.10.2, r965fd380006fa78b2315668fbc7eb432e1d8200f
Compiled:	Tue May 24 17:35:00 -0500 2022 by ubuntu from branch-2.10.2
Cluster ID:	CID-ae95ca73-0abe-498e-af1b-fc781793ce05
Block Pool ID:	BP-205463673-127.0.1.1-1747449399064

Summary

Security is off.
Safemode is off.
22 files and directories, 8 blocks = 30 total filesystem object(s).
Heap Memory used 121.91 MB of 295.5 MB Heap Memory. Max Heap Memory is 889 MB.
Non Heap Memory used 51.22 MB of 52.31 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	73.31 GB
----------------------	----------



Datanode Information

Legend: ✔ In service ❌ Down ⚠ Decommissioned ⚠ Decommissioned & dead ⚠ In Maintenance & dead

Datanode usage histogram

Disk usage of each DataNode (%)

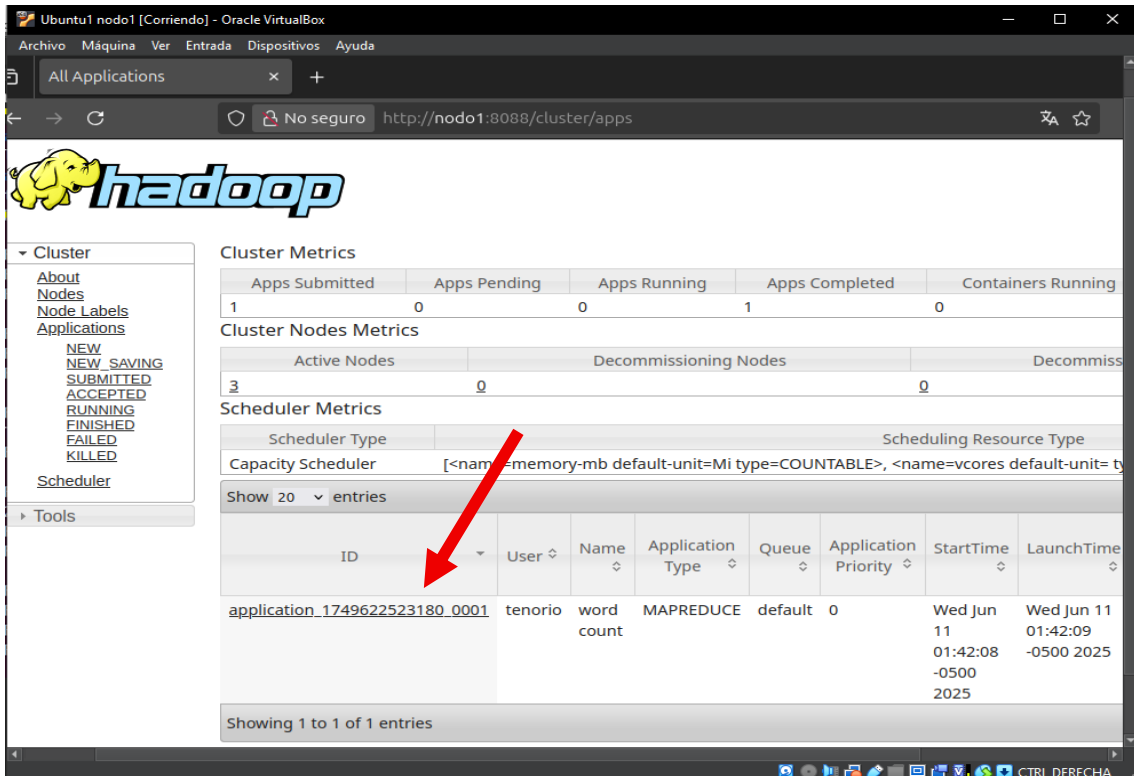
In operation

Show 25 entries

Search:

OPCIONAL: en este punto cuando cierras tu distro, el historial se borra automáticamente ya que no esta configurado para que el historial se guarde, pero si se guarda en la carpeta designada, para poder visualizarlo nuevamente tendrías que eliminar la salida del archivo con el siguiente comando “hdfs dfs -rm -r /examen_final” y volverlo a ejecutar con el comando anterior “hadoop jar hadoop-mapreduce-examples-2.10.2.jar wordcount /unajma /examen_final” y vuelves a visualizar en tu navegador.

```
tenorio@nodo1:/opt/hadoop/share/hadoop/mapreduce$ hdfs dfs -rm -r /examen_final
Deleted /examen_final
tenorio@nodo1:/opt/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.10.2.jar wordcount /unajma /examen_final
25/06/11 01:42:06 INFO client.RMProxy: Connecting to ResourceManager at nodo1/192.168.1.100:100.101:8032
25/06/11 01:42:07 INFO input.FileInputFormat: Total input files to process : 1
25/06/11 01:42:07 INFO mapreduce.JobSubmitter: number of splits:1
25/06/11 01:42:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1749622523180_0001
25/06/11 01:42:08 INFO conf.Configuration: resource-types.xml not found
25/06/11 01:42:08 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
25/06/11 01:42:08 INFO resource.ResourceUtils: Adding resource type - name = memory-mb , units = Mi, type = COUNTABLE
25/06/11 01:42:08 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
25/06/11 01:42:08 INFO impl.YarnClientImpl: Submitted application application_1749622523180_0001
25/06/11 01:42:08 INFO mapreduce.Job: The url to track the job: http://nodo1:8088/proxy/application_1749622523180_0001/
25/06/11 01:42:08 INFO mapreduce.Job: Running job: job_1749622523180_0001
25/06/11 01:42:14 INFO mapreduce.Job: Job job_1749622523180_0001 running in uber mode : false
25/06/11 01:42:14 INFO mapreduce.Job: map 0% reduce 0%
25/06/11 01:42:20 INFO mapreduce.Job: map 100% reduce 0%
25/06/11 01:42:26 INFO mapreduce.Job: map 100% reduce 100%
25/06/11 01:42:26 INFO mapreduce.Job: Job job_1749622523180_0001 completed successfully
```



The screenshot shows the Hadoop web interface in a browser window. The URL is <http://nodo1:8088/cluster/apps>. The interface displays various metrics and a table of running applications.

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
1	0	0	1	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
3	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type
Capacity Scheduler	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit=Mi type=COUNTABLE>]

Applications Table

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime
application_1749622523180_0001	tenorio	word count	MAPREDUCE	default	0	Wed Jun 11 01:42:08 -0500 2025	Wed Jun 11 01:42:09 -0500 2025

Showing 1 to 1 of 1 entries

