

3. Configuración modo pseudo distribuido en local

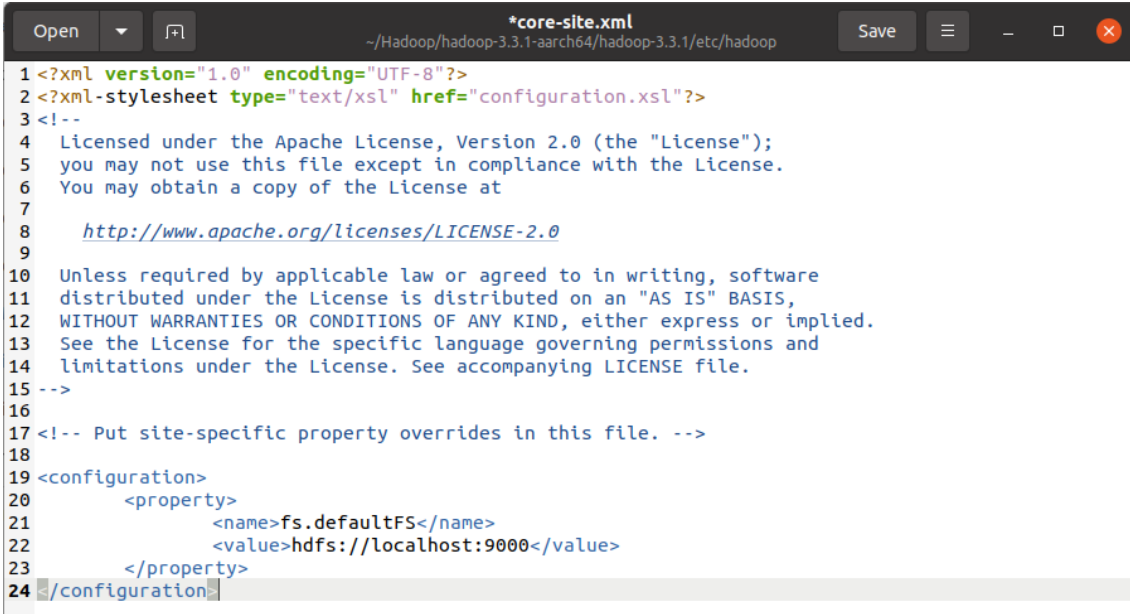
La administración del sistema dfs se puede realizar mediante un entorno web. Es necesario disponer de un servidor apache instalado y en marcha.

Si no está instalado el servicio apache2 seguir los siguientes pasos:

```
$ sudo apt install apache2  
  
$ sudo service start apache2  
  
$ sudo systemctl status apache2
```

3.1. Modificar el archivo de configuración “etc/hadoop/core-site.xml”

```
<configuration>  
  
  <property>  
  
    <name>fs.defaultFS</name>  
  
    <value>hdfs://localhost:9000</value>  
  
  </property>  
  
</configuration>
```



```
1 <?xml version="1.0" encoding="UTF-8"?>  
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>  
3 <!--  
4 Licensed under the Apache License, Version 2.0 (the "License");  
5 you may not use this file except in compliance with the License.  
6 You may obtain a copy of the License at  
7  
8   http://www.apache.org/licenses/LICENSE-2.0  
9  
10 Unless required by applicable law or agreed to in writing, software  
11 distributed under the License is distributed on an "AS IS" BASIS,  
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.  
13 See the License for the specific language governing permissions and  
14 limitations under the License. See accompanying LICENSE file.  
15 -->  
16  
17 <!-- Put site-specific property overrides in this file. -->  
18  
19 <configuration>  
20   <property>  
21     <name>fs.defaultFS</name>  
22     <value>hdfs://localhost:9000</value>  
23   </property>  
24 </configuration>
```

3.2. Modificar el archive de configuración etc/hadoop/hdfs-site.xml

```
<configuration>

  <property>

    <name>dfs.replication</name>

    <value>1</value>

  </property>

</configuration>
```

3.3. Verificar acceso con SSH

Información sobre openSSH en <https://www.openssh.com/>

```
$ ssh localhost
```

En el caso de tener problemas con el acceso ssh

Configurar una clave ssh

```
$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$ chmod 0600 ~/.ssh/authorized_keys
```

Reinstalar el servicio ssh

```
$ sudo apt-get remove openssh-client openssh-server
$ sudo apt-get install openssh-client openssh-server
```

3.4. Verificar la instalación

Acceder con ssh y formatear el sistema de archivos

```
$ bin/hdfs namenode -format
```

Iniciar los procesos de NameNode y DataNode

```
$ sbin/start-dfs.sh
```

(en el caso de error “No puedo crear permisos”) asignar permisos al usuario con sudo chown YOURUSER:YOURUSER -R /home/YOURUSER/hadoop-3.3.4/*

Conectar con el navegador a <http://localhost:9870/>

Overview 'localhost:9000' (✓active)

Started:	Mon Oct 04 21:37:35 +0200 2021
Version:	3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2
Compiled:	Tue Jun 15 12:51:00 +0200 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID:	CID-fd0c7dc8-4571-466a-b6f0-0b911184baed
Block Pool ID:	BP-1939965246-127.0.1.1-1633376061230

Summary

Security is off.
Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 118.03 MB of 217.5 MB Heap Memory. Max Heap Memory is 875 MB.
Non Heap Memory used 51.18 MB of 52.65 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	19.07 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	8.31 GB
DFS Remaining:	9.77 GB (51.24%)

Detener los procesos de DataNode y NameNode

```
$ sbin/stop-dfs.sh
```

3.5. Ejemplos incluidos en Hadoop-mapreduce-examples

Relación de comandos incluidos en el fichero de ejemplo

Instrucción	Descripción
aggregatewordcount	Cuenta las palabras de los archivos de entrada.
aggregatewordhist	Calcula el histograma de las palabras de los archivos de entrada.
bbp	Usa una fórmula Bailey-Borwein-Plouffe para calcular los dígitos exactos de Pi.
dbcount	Cuenta los registros de vistas de página almacenados en una base de datos.
distbbp	Usa una fórmula de tipo BBP para calcular los bits exactos de Pi.
grep	Cuenta las coincidencias de una expresión regular en la entrada.
join	Realiza una unión de conjuntos de datos ordenados con particiones equiparables.
multifilewc	Cuenta las palabras de varios archivos.
pentomino	Programa para la colocación de mosaicos con el fin de encontrar soluciones a problemas de pentominó.
pi	Calcula Pi mediante un método cuasi Monte Carlo.
randomtextwriter	Escribe 10 GB de datos de texto aleatorios por nodo.

randomwriter	Escribe 10 GB de datos aleatorios por nodo.
secondarysort	Define una ordenación secundaria para la fase de reducción.
sort	Ordena los datos escritos por el escritor aleatorio.
sudoku	un solucionador de sudokus.
teragen	genera datos para la ordenación de terabytes (terasort).
terasort	ejecuta la ordenación de terabytes (terasort).
teravalidate	comprueba los resultados de la ordenación de terabytes (terasort).
wordcount	Cuenta las palabras de los archivos de entrada.
wordmean	Cuenta la longitud media de las palabras de los archivos de entrada.
wordmedian	Cuenta la mediana de las palabras de los archivos de entrada.
wordstandarddeviation	Cuenta la desviación estándar de la longitud de las palabras de los archivos de entrada.

3.6. Un ejemplo con grep

Iniciar los procesos de DataNode y NameNode según el punto anterior

Crear el directorio HDFS para que se ejecute el trabajo de MapReduce.

Algunos comandos del programa dfs:

Mkdir <nombre carpeta>	Crea la carpeta
Put <origen> <destino>	Carga archivos en el servidor hdfs
Cat <carpeta>	Muestra el contenido de la carpeta
Rmdir <carpeta>	Borra carpeta vacía
Rm -r <carpeta>	Borra carpeta aunque tenga archivos u otras carpetas

La carpeta de trabajo que toma por defecto está en el servidor hdfs/user/<nombre de usuario>. Es necesario crearla antes de lanzar los procesos

```
$ bin/hdfs dfs -mkdir /user
$ bin/hdfs dfs -mkdir /user/<username>
```

Copiar los archivos base para el ejemplo. Son los mismos que en el ejercicio 2

```
$ bin/hdfs dfs -mkdir input
$ bin/hdfs dfs -put etc/hadoop/*.xml input
```

Ejecutar el proceso MapReduce de ejemplo

```
$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar
grep input output 'dfs[a-z.]+'
```

Verificar el resultado

```
$ bin/hdfs dfs -cat output/*
```

Finalizar los procesos NameNode y DataNode

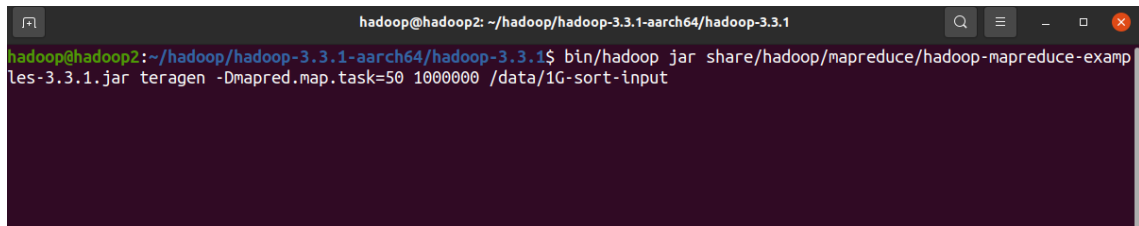
```
$ sbin/stop-dfs.sh
```

3.7. Ejemplo de rendimiento

Teragen, terasort y teravalidate se utilizan para evaluar el rendimiento de sistemas. Permite generar un volumen grande de datos y aplicar una ordenación a los mismos

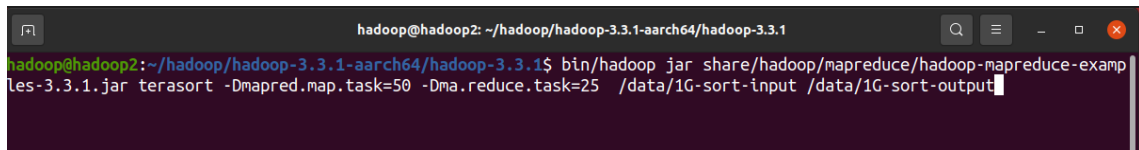
Crear el conjunto de datos

```
$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4. jar teragen -Dmapred.map.task=50 1000000 /data/1G-sort-input
```

A terminal window titled 'hadoop@hadoop2: ~/hadoop/hadoop-3.3.1-aarch64/hadoop-3.3.1' shows the command 'bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.1.jar teragen -Dmapred.map.task=50 1000000 /data/1G-sort-input' being executed. The command is on two lines due to wrapping.

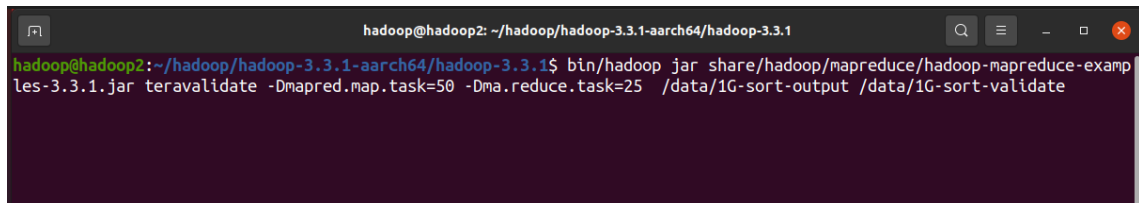
Ordenar los datos

```
$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4. jar terasort -Dmapred.map.task=50 -Dma.reduce.task=25 /data/1G-sort-input /data/1G-sort-output
```

A terminal window titled 'hadoop@hadoop2: ~/hadoop/hadoop-3.3.1-aarch64/hadoop-3.3.1' shows the command 'bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.1.jar terasort -Dmapred.map.task=50 -Dma.reduce.task=25 /data/1G-sort-input /data/1G-sort-output' being executed. The command is on two lines.

Validar resultado

```
$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4. jar teravalidate -Dmapred.map.task=50 -Dma.reduce.task=25 /data/1G-sort-output /data/1G-sort-validate  
$ bin/hdfs dfs -cat /data/1G-sort-validate/*
```

A terminal window titled 'hadoop@hadoop2: ~/hadoop/hadoop-3.3.1-aarch64/hadoop-3.3.1' shows the command 'bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.1.jar teravalidate -Dmapred.map.task=50 -Dma.reduce.task=25 /data/1G-sort-output /data/1G-sort-validate' being executed. The command is on two lines.

Información del nodo de datos

```
bin/hdfs dfs -ls /data  
bin/hdfs dfs -ls /data/1G-sort-input
```

```
hadoop@hadoop2:~/hadoop/hadoop-3.3.1-aarch64/hadoop-3.3.1$ bin/hdfs dfs -ls /data
2021-10-06 20:50:10,553 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
ltin-java classes where applicable
Found 3 items
drwxr-xr-x - hadoop supergroup          0 2021-10-06 18:10 /data/1G-sort-input
drwxr-xr-x - hadoop supergroup          0 2021-10-06 18:12 /data/1G-sort-output
drwxr-xr-x - hadoop supergroup          0 2021-10-06 18:13 /data/1G-sort-validate
hadoop@hadoop2:~/hadoop/hadoop-3.3.1-aarch64/hadoop-3.3.1$ bin/hdfs dfs -ls /data/1G-sort-input
2021-10-06 20:50:40,392 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
ltin-java classes where applicable
Found 2 items
-rw-r--r-- 1 hadoop supergroup          0 2021-10-06 18:10 /data/1G-sort-input/_SUCCESS
-rw-r--r-- 1 hadoop supergroup 1000000000 2021-10-06 18:10 /data/1G-sort-input/part-m-00000
hadoop@hadoop2:~/hadoop/hadoop-3.3.1-aarch64/hadoop-3.3.1$
```