

Práctica HDFS - Hadoop 3

Requisitos

1. Verifique que el proxy haya sido configurado de forma correcta para instalar paquetes vía `apt`. Recuerde que este paso solo es necesario si está desde la red de la Universidad Nacional.
2. Instale `openssh-server`

```
$ sudo apt install openssh-server
```

3. Verifique que Java funcione:

```
$ java -version
java version "1.8.0_121"
Java(TM) SE Runtime Environment (build 1.8.0_121-b13)
Java HotSpot(TM) 64-Bit Server VM (build 25.121-b13, mixed mode)
```

Instalación

1. Descargue `hadoop 3.1.2`
2. Ingrese a la carpeta y descomprima el instalador:

```
$ cd Descargas
$ tar -xvzf hadoop-3.1.2.tar.gz
```

3. Regrese a su carpeta home(`/home/mls`), ejecute `nano` `.bashrc` y añada las siguientes líneas al final del archivo:

```
export HADOOP_HOME="/home/mls/Descargas/hadoop-3.1.2"
export PATH=$HADOOP_HOME/bin:$PATH
```

4. Valide que el comando `hadoop` funciona:

```
$ hadoop
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS
]
or hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
where CLASSNAME is a user-provided Java class

...
kms run KMS, the Key Management Server

SUBCOMMAND may print help when invoked w/o parameters
or with -h.
```

5. Abra el archivo `/home/mls/Descargas/hadoop-3.1.2/etc/hadoop/core-site.xml` y actualice el campo `<configuration>` con el siguiente valor:

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

```
</property>  
</configuration>
```

6. Cree una carpeta para el sistema de archivos:

```
$ mkdir -p /home/mls/hadoop_data
```

7. De manera similar, edite el archivo

`/home/mls/Descargas/hadoop-3.0.0/etc/hadoop/hdfs-site.xml` :

```
<configuration>  
  <property>  
    <name>dfs.datanode.data.dir</name>  
    <value>file:/home/mls/hadoop_data/dfs/data/</value>  
  </property>  
  <property>  
    <name>dfs.namenode.name.dir</name>  
    <value>file:/home/mls/hadoop_data/dfs/name/</value>  
  </property>  
  <property>  
    <name>dfs.replication</name>  
    <value>1</value>  
  </property>  
</configuration>
```

8. Edite el archivo `/home/mls/Descargas/hadoop-3.0.0/etc/hadoop/hadoop-env.sh` y añada la siguiente línea:

```
$ export JAVA_HOME="/home/mls/Descargas/jdk1.8.0_121/"
```

9. Luego, en una terminal ejecute las siguientes líneas:

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

10. En la misma terminal, cree el sistema de archivos:

```
$ hdfs namenode -format
```

11. Inicie el servicio `NameNode` y el servicio `DataNode` :

```
$ cd /home/mls/Descargas/hadoop-3.1.2
$ ./sbin/start-dfs.sh
```

12. Puede verificar de dos maneras que se esté ejecutando correctamente. La primera, consiste en verificar que se estén ejecutando los siguientes servicios: `NameNode` , `DataNode` y `SecondaryNameNode` .

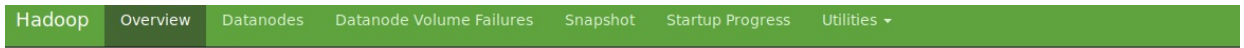
```
$ jps
134304 SecondaryNameNode
```

134487 Jps

133781 NameNode

134014 DataNode

13. También puede monitorear que se haya levantado la dirección <http://localhost:9870/>:



Overview 'localhost:9000' (active)

Started:	Wed Jun 05 09:25:29 -0500 2019
Version:	3.1.2, r1019dde65bcf12e05ef48ac71e84550d589e5d9a
Compiled:	Mon Jan 28 20:39:00 -0500 2019 by sunilg from branch-3.1.2
Cluster ID:	CID-1c2adf3d-5fe2-4f0c-b105-04524902c03c
Block Pool ID:	BP-992850010-127.0.1.1-1559775976670

Monitoreo del sistema de archivos

Interacción con los comandos:

1. Cree una carpeta:

```
$ hdfs dfs -mkdir /carpeta1
```




2. Copie un archivo a la carpeta creada:

```
$ hdfs dfs -put /home/mls/Descargas/hadoop-3.1.2.tar.gz /carpeta1/instalador.tar.gz
```


3. Valide en la interfaz web que el archivo se cargó correctamente en <http://localhost:9870/explorer.html#/carpeta1> :

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/carpeta1 Go!   

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	ssierral	supergroup	317.03 MB	Jun 05 09:34	1	128 MB	inst.tar.gz	

Showing 1 to 1 of 1 entries Previous 1 Next

4. Descargue un archivo del HDFS:

```
$ hdfs dfs -get /carpeta1/instalador.tar.gz copia.tar.gz
```

5. Valide los datos del archivo:

```
$ file copia.tar.gz
copia.tar.gz: gzip compressed data, last modified: Wed Jun 05 08:53:00 2018, from Unix
```

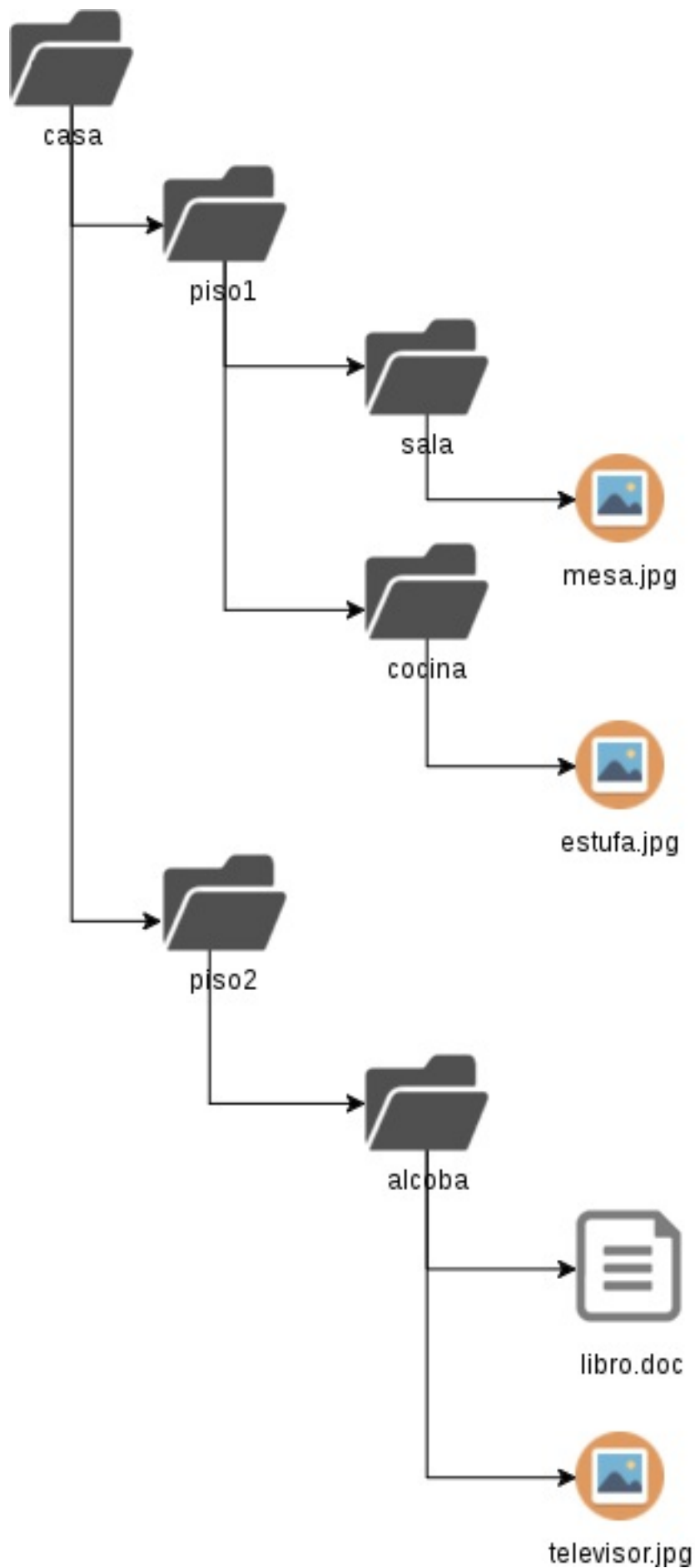
6. Borre el archivo del sistema de archivos:

```
$ hdfs dfs -rm /carpeta1/instalador.tar.gz
```

Práctica

Manejando el sistema de archivos

1. Cree la siguiente estructura en el sistema de archivos hdfs:



2. Realice las siguientes modificaciones:

- Agregue un `estudio` en el `piso2`.
- Agregue una `cama.jpg` en la `alcoba`.
- Mueva el televisor de la `alcoba` a la `sala`.
- Mueva el `libro.doc` al `estudio`.

Ejecución de `wordCount`

1. Descargue el archivo `biblia.txt`.
2. Divida el archivo para que pueda ser procesado en paralelo:

```
$ cd /home/mls/Descargas/  
$ split biblia.txt
```

3. El comando `split` divide el archivo en partes de 1000 líneas cada uno con el prefijo `xa`:

```
$ cd /home/mls/Descargas/  
$ ls xa*  
xaa  xab  xac  xad  xae  xaf  xag
```

4. Cargue los archivos al sistema HDFS en la carpeta `/user/mls/input`. **Nota:** Esta carpeta se va a encontrar realmente en el sistema distribuido de Hadoop:

```
$ mkdir /home/mls/Descargas/archivos_biblia  
$ mv /home/mls/Descargas/xa* /home/mls/Descargas/archivos_biblia  
$ hdfs dfs -mkdir /user
```



```
$ hdfs dfs -mkdir /user/mls
```

```
$ hdfs dfs -put /home/mls/Descargas/archivos_biblia input
```

5. Valide en el explorador de hdfs:

localhost:9870/explorer.html#/user/ssierral/input

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/user/ssierral/input Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	ssierral	supergroup	108.84 KB	Mar 14 16:17	1	128 MB	xaa
-rw-r--r--	ssierral	supergroup	124.59 KB	Mar 14 16:17	1	128 MB	xab
-rw-r--r--	ssierral	supergroup	147.93 KB	Mar 14 16:17	1	128 MB	xac
-rw-r--r--	ssierral	supergroup	185.92 KB	Mar 14 16:17	1	128 MB	xad
-rw-r--r--	ssierral	supergroup	163.49 KB	Mar 14 16:17	1	128 MB	xae
-rw-r--r--	ssierral	supergroup	128.68 KB	Mar 14 16:17	1	128 MB	xaf
-rw-r--r--	ssierral	supergroup	118.85 KB	Mar 14 16:17	1	128 MB	xag

Showing 1 to 7 of 7 entries

Previous 1 Next

6. Ejecute la tarea **wordcount** en el entorno distribuido:

```
$ hadoop jar /home/mls/Descargas/hadoop-3.1.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount input output
```

7. Obtenga los conteos de la carpeta **output** generada por el **wordcount**:

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

...

```
él 592
```

él! 1

él, 215

él. 131

él: 8

él; 43

él? 8

éramos 5

ése 16

ése, 1

ésta 6

ésta, 4

ésta 3

éste 107

éste, 25

...