Instalación Hadoop en CentOS 7

INSTRUCTIVO

- **`letra normal negrita :`** comandos ingresados
- *`itálicas :`* mensajes desplegados, **notas.**
- normal**:** Explicación principal formato justificado.
  Explicación **extra** alineado a la **derecha**

NOTA: NO *se asume que el usuario es conocedor en el uso e instalación del sistema operativo CentOS. para este manual fue utilizado la versión 7 de 64 bits de forma minimal y Gnome, realizando la instalación de hadoop 2.9.1 la cual fue la última versión estable hasta ahora en Septiembre 2018.*

NOTA*: Las configuraciones de la versión 2.7 a 2.9 son iguales, para la versión 3.0 en adelante cambia mucho.*

1.Instalación de CentOS 7 versión Mínima en maquina nativa por medio de un dvd

A continuación, se mostrarán los pasos necesarios para la instalación del sistema operativo CentOS 7 64 bits modo consola o minimal
Ingresar a la página del sistema operativo de CentOS y seleccione la descarga mínima

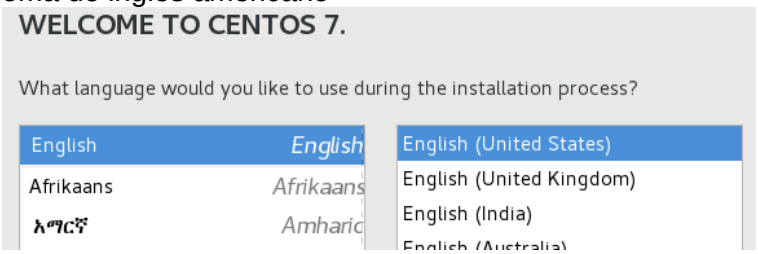
Al ingresar en esta opción, seleccione la descarga de *Actual Country.*

Se recomienda grabar la imagen ISO en un disco de DVD, utilice algún software especializado para grabar la información, en este manual se utilizó el software **ISO Burner**.

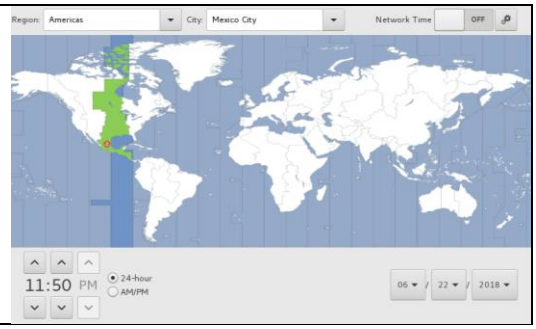
A continuación, introduce el disco de instalación en la computadora.

Seleccione la opción de ***Install CentOS 7***

Seleccione el idioma de inglés americano

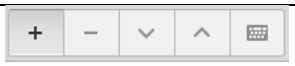


De clic en el botón de continuar. Modifique el lugar y el horario en *Day & Time*

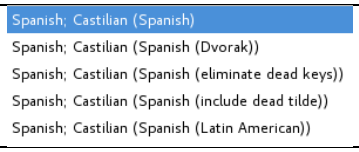


Cambiar la configuración del teclado a idioma español latino, dando clic en la opción de *keyboard*

De clic en el botón **+** para agregar otro idioma de teclado.

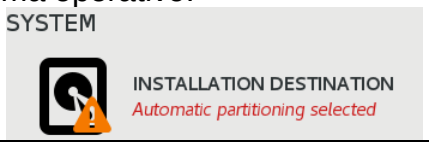


Puede elegir la opción castellano o ***latinoamericano***



Suba la opción para tener como primera elección el idioma español.

Ahora seleccione la unidad de disco en la que deseamos instalar nuestro sistema operativo.



Seleccione su tarjeta de red, ya sea por ETHERNET o por WIFI, también puede no hacerlo en este momento. Más adelante puede configurarlo desde terminal, o en el modo gráfico

Para continuar haga clic en el botón de *Begin Installation*.

Después de haber dado clic en el anterior botón, se desplegarán dos opciones para introducir las contraseñas del usuario *root* y usuarios. Para esto elegiremos la contraseña del administrador **root** la cual será **toor**.

Agregue un usuario llamado **#maquina**, de clic en usuario administrador, y marque con una flecha la opción de no requerir contraseña para la cuenta de este usuario.

No importa el nombre del usuario que eliga, pero más adelante será importante trabajar con el usuario **hadoop**.

1. Instalación de Hadoop en un nodo desde terminal en CentOS 7

Para ejecutar Hadoop y otras de las herramientas de Big Data se requieren instalar algunos paquetes para poder descargar nuestros archivos ejecutables.

> ***NOTA:*** *Para poder proseguir en los siguientes pasos, usted requiere tener conexión a internet, en las últimas hojas de este manual, existen apartados para la conexión del Wifi desde terminal y conexión a una red alámbrica por medio de una IP estática.*
>
> ***NOTA:*** *Este manual cuenta con los comandos de los editores de texto vi, vim, nano.*
>
> ***RECOMENDACION:*** *Puede saltarse la parte de descarga de los archivos de java y hadoop si los tiene en una usb, en este manual puede encontrar la instrucción para crear un directorio de usb y copiar archivos a una ruta en específico con los respectivos permisos. Este comando podrá saltárselos donde viene indicado* ***INSTALE***

Instalar la herramienta **wget,** para poder descargar java, implemente el siguiente comando:

Ingrese como superusuario (root)

```
sudo su -
yum install wget
```

> ***NOTA:*** *el comando sudo, se utiliza para tener permisos de super usuario.*

Para poder visualizar nuestras interfaces de red en modo terminal

```
yum install net-tools
```

Hadoop utiliza Java, así que **INSTALE** el archivo RPM con el siguiente comando:

```
$ wget --no-cookies --no-check-certificat --header "Cookie: gpw_e24=http%3A%2F%2Fwww.oracle.com%2F;
oraclelicense=accept-securebackup-cookie" "http://download.oracle.com/otn-pub/java/jdk/8u171-
b11/512cd62ec5174c3487ac17c61aaa89e8/jdk-8u171-linux-x64.rpm"
```

> ***NOTA:*** *En el renglón donde es mencionado http://download.oracle.com/otn-pub/java/jdk/8u171- tiene continuación el siguiente relglón sin dar espacio ni enter.*
>
> ***NOTA:*** *La versión jdk varía dependiendo la última versión desarrollada, por lo tanto, algunos parámetros pueden cambiar al momento de introducir el comando anterior, se recomienda buscar como instalar la versión actual de java.*
>
> ***NOTA:*** *No utilice el jre que viene por defecto en algunas instalaciones como son Gnome o KDE debido a que openjdk no tiene las herramientas necesarias para nuestros propósitos. El comando que vienen en algunos tutoriales es* ***yum install java (no lo use).***

Después de terminar la instalación implemente el siguiente comando para la instalación de java:

```
sudo yum localinstall jdk-8u171-linux-x64.rpm
```

> ***Nota:*** *Al igual que el comando anterior, la instalación dependerá de la versión del jdk.*

La versión de java instalada se conoce mediante el comando:

```
java –version
```

Agregue un nuevo usuario

```
sudo adduser hadoop
sudo passwd hadoop
```

Ingresar como usuario: **hadoop**

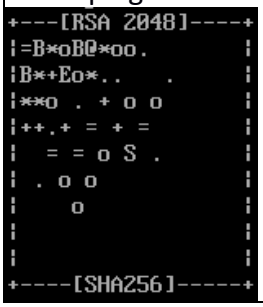
```
su – hadoop
```

Configuración de la clave SSH para iniciar de forma segura sin contraseña.

```
ssh-keygen -t rsa -P “”
```

Este comando le pedirá que ingrese la clave, usted presione **enter** a todas las opciones.

Se desplegará un mensaje como el siguiente.



Ahora con el siguiente comando se generará la clave para que el usuario pueda iniciar sesión con la clave privada y posteriormente con el siguiente comando se le asignarán los permisos para ingesar sin utilizar la contraseña.

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/authorized_keys
```

Para agregar el símbolo **~** se pueden utilizar las siguientes teclas

	Windows	Linux
Alt+126	Si	No
Alt Gr + Tecla 4	Si	Si
Alt Gr + tecla ñ	No	Si

Utilice el siguiente comando para saber si puede ingresar al localhost sin necesidad de utilizar contraseña.

```
ssh localhost
```

Para salir del usuario: **hadoop** y regresar como super usuario introduciremos el comando

```
exit
```

Ahora entraremos a la página oficial de *apache hadoop* para saber cuál es la versión más actual.

http://hadoop.apache.org/releases.html#Download

La versión más actual es la 2.9.1, desde terminal introduzca el siguiente comando.
INSTALE.

```
wget http://mirror.rise.ph/apache/hadoop/common/hadoop-2.9.1/hadoop-2.9.1.tar.gz
```

Ahora descomprima el archivo con el siguiente comando.

```
tar xvfz hadoop-2.9.1.tar.gz
```

Para poder mover la siguiente carpeta siendo usuario **hadoop**, permita que tenga permisos de superusuario.

Ingrese como usuario **root**

```
sudo visudo
```

Agregue al usuario hadoop, debajo del usuario root.
Presione **s** para insertar, **Esc** para salir y **:wq** para guardar y salir.

***Recomendación**: Para que las configuraciones funcionen para todos los usuarios; las rutas /**home/hadoop** deben ser sustituidas por **/opt/hadoop**.*

Desplace el contenido de hadoop-2.9.1 al directorio de la cuenta **usuario**

Mueva el comando como usuario **hadoop**

```
sudo mv hadoop-2.9.1/* /home/hadoop/.
```

```
ll
```

***Nota:** Después de copiar los archivos de hadoop-2.9.1 a la carpeta de usuario, no olvide poner los permisos necesarios, ya que usted está copiándolos siendo un superusuario. De no hacerlo tendrá problemas para editar todos los archivos. Si ingresa como usuario **hadoop**, usted no tendrá problemas con los permisos.*

Asignar permisos al archivo

```
chown --recursive hadoop:hadoop /home/hadoop/
```

Otra opción

```
chown -R hadoop:hadoop /home/hadoop/
```

El editor de textos por defecto de CentOS, es **vi**, pero por comodidad se sugiere instalar nano.

```
sudo yum install nano
```

Para encontrar la ruta de instalación de JVM (jre) escribir el comando

```
sudo update-alternatives --config java
```

El sistema desplegará el siguiente mensaje (esto dependerá de la versión instalada de java)

```
/usr/java/jdk1.8.0_171-amd64/jre/bin/java
```

NOTA: *En modo gráfico aparecen openjdk como su versión de java por defecto. En modo minimal no viene instalado ninguna versión de java.*

Edite el archivo **~/.bashrc** con el siguiente comando y escriba la siguiente información

```
nano ~/.bashrc
```

```
export PATH
export HADOOP_PREFIX="/home/hadoop"
export HADOOP_HOME="${HADOOP_PREFIX}"
export HADOOP_COMMON_HOME="${HADOOP_PREFIX}"
export HADOOP_CONF_DIR="${HADOOP_PREFIX}/etc/hadoop"
export HADOOP_HDFS_HOME="${HADOOP_PREFIX}"
export HADOOP_MAPRED_HOME="${HADOOP_PREFIX}"
export HADOOP_YARN_HOME="${HADOOP_PREFIX}"
export
"PATH=${PATH}:${HADOOP_PREFIX}/bin:${HADOOP_PREFIX}/bin"
export
"PATH=${PATH}:${HADOOP_PREFIX}/bin:${HADOOP_PREFIX}/sbin"
export JAVA_HOME=/usr/java/jdk1.8.0_171-amd64/jre
```

Para que tenga efecto estos cambios se debe reiniciar la terminal de comandos

```
source ~/.bashrc
```

Otra opción

```
. ~/.bashrc
```

NOTA*: Error hdfs namenode found es debido a que no está compilado las variables de entorno, utilizar source ~/.bashrc.*

Los archivos del ambiente de Hadoop que se requieren configurar se presentan a continuación, estos se pueden modificar en cualquier orden. Entre al directorio de la ubicación mediante el comando:

```
cd ~/etc/hadoop
```

Edite el archivo **hadoop-env.sh,** usted debe añadir la ruta de la máquina virtual de JAVA para que la utilice el ambiente de Hadoop, así como su directorio HOME y CONF_DIR

```
nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

```
export JAVA_HOME=/usr/java/jdk1.8.0_171-amd64/jre
export HADOOP_HOME=/home/hadoop
export HADOOP_CONF_DIR=/home/hadoop/etc/hadoop
```

Guardamos, confirmamos y salimos del archivo con **CTRL+O, ENTER, CTRL+X**.

Editar archivos de configuración
La descripción de cada configuración que se realice a continuación será la siguiente

Fichero	Formato
hadoop-env.sh	Variables de entorno que son usadas en los scripts que ejecutan Hadoop.
core-site.xml	Configuración para Hadoop Core, como ajustes de I/O que son comunes para HDFS y MapReduce
hdfs-site.xml	Ajustes de configuración para los daemons de HDFS: NameNode, Secundary NameNode y DataNodes
mapred-site.xml	Ajustes de configuración para los daemons MapReduce:Jobtracker, Tasktrackers o Yarn
yarn-site.xml	Ajustes de configuración para YARN
masters	Lista de máquinas (escribir una por línea) que iniciarán un SecundaryNameNode
slaves	Lista de máquinas (escribir una por línea)que iniciarán un DataNode y un NodeManager

NOTA: *ANTES DE INSTALAR UNA VERSION NUEVA, REVISE LOS ARCHIVOS DE CONFIGURACION DE LA VERSION A INSTALAR PARA SUS FUTUROS CAMBIOS EN:*
NOTA: *En la versión 3.0.3 cambia slaves por workers*

http://hadoop.apache.org/docs/r2.9.1/hadoop-project-dist/hadoop-common/DeprecatedProperties.html

En general

http://hadoop.apache.org/docs/r**VERSIÓN**/hadoop-project-dist/hadoop-common/DeprecatedProperties.html

El resto de la configuración requiere que te dirijas a la ruta:

$HADOOP_HOME tiene la ruta de origen, usted puede modificar esta ruta en .bashrc
```
cd $HADOOP_HOME/etc/hadoop
```

NOTA: *Usted puede usar **Tab** para a completar la línea que desea escribir.*

Editar **core-site.xml**

NOTA: *fs.defaultFS Apunta al DATA NODE*
NOTA: *fs.default.name Apunta al NAME NODE*
NOTA: *La variable fs.default.name es obsoleta en la versión 3.0.3*

```
<configuration>
<property>
<name> fs.defaultFS</name>
```

```
<value>hdfs://localhost:54310</value>
</property>
<property>
<name> fs.default.name</name>
<value>hdfs://localhost:54310</value>
</property>
</configuration>
```

Editar **hdfs-site.xml**

```
<configuration>
   <property>
      <name>dfs.replication</name>
      <value>1</value>
   </property>
   <property>
      <name>dfs.blocksize</name>
      <value>4194304</value>
   </property>
   <property>
      <name>dfs.namenode.name.dir</name>
      <value>/home/hadoop/hadoopdata/hdfs/namenode </value>
   </property>
   <property>
      <name>dfs.datanode.data.dir</name>
      <value>/home/hadoop/hadoopdata/hdfs/datanode </value>
   </property>
</configuration>
```

Por el momento se tomará este tamaño de bloque para la creación de archivos.

Copie **mapred-site.xml.templace** como **mapred-site.xml**

```
cp mapred-site.xml.templace mapred-site.xml
```

Editar **mapred-site.xml**

```
<configuration>
   <property>
      <name>mapreduce.framework.name</name>
      <value>yarn</value>
   </property>
<property>
        <name>mapred.job.tracker</name>
        <value>localhost:54311</value>
    </property>
    <property>
        <name>mapreduce.jobhistory.webapp.address</name>
        <value>localhost:19888</value>
```

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

Editar **yarn-site.xml**

```
<configuration>
   <property>
      <name>yarn.nodemanager.aux-services</name>
      <value>mapreduce_shuffle</value>
   </property>
<property>
  <name>yarn.nodemanager.aux-
services.mapreduce_shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
</configuration>
```

Para el caso de un sistema pseudo distribuido se procede a editar el archivo con el nombre de la máquina maestra33

```
nano masters
```

El contenido de este archivo debe ser:

```
localhost
```

El último archivo que se edita es el que indica las máquinas esclavas, como es un nodo único, solamente es un equipo. El comando de edición es:

```
nano slaves
```

El contenido de este archivo debe ser:

```
localhost
```

NOTA: *Verifica que los archivos contengan las palabras localhost, en versiones futuras slaves cambiara por workers*

Al concluir la edición de estos archivos, se procede a formatear el sistema de archivos de Hadoop mediante el siguiente comando:

```
$ hdfs namenode -format
```

Formatear sirve para iniciar el sistema de archivos distribuido.

Si se ha formateado de la manera correcta, debe aparecerle un mensaje como el siguiente

```
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at localhost/127.0.0.1
************************************************************/
```

NOTA: *Si formateaste como super usuario y quieres volver a intentar el mismo procedimiento como usuario, eliminar la ruta del proceso que incluye* ***/tmp/hadoop-usuario-namenode.pid: Permission denied*** *como usuario* ***root*** *y vuelve a compilar como usuario* ***hadoop****. Este error aparece solo al principio de la ejecución del namenode, así que si se encuentra trabajando en terminal tendrá problema para ver ese error ya que la terminal no le permite visualizar toda la ejecución.*

Para eliminar un archivo **rm –rf <Ruta>**

Puede o no entrar a estar dirección para observar donde se encuentran estos procesos.

```
cd $HADOOP_HOME/sbin/
```

El uso de Hadoop requiere que se activen varios servidores mediante dos scripts:

Se recomienda no usar el anterior comando y usar los siguientes dos comandos:

start-dfs.sh	Namenode,DataNode,SecondaryNameNode
start-yarn.sh	ResourceManager,NodeManager

NOTA: *el comando start-all.sh ha quedado obsoleto.*

Se comprueba si realmente está en funcionamiento, mediante el comando:

```
jps
```

Se podrá visualizar los siguientes servicios.

```
5489 NameNode
5618 DataNode
5860 SecondaryNameNode
6677 Jps
6250 ResourceManager
6458 NodeManager
```

NOTA: *Si encuentra problemas al inicializar los procesos y no encuentra el* ***datanode****, esto se puede solucionar borrando la carpeta* ***datanode****, el cual fue asignado en la ruta que usted asignó en* ***hdfs-site.xml.***

NOTA: *Si al ejecutar* ***start-dfs.sh*** *no aparece el* ***namenode****, problema proviene desde el formateo de* ***hdfs namenode –format****, posiblemente otra solución es que usted debe borrar el* ***namenode*** *creado en la ruta de* ***hdfs-site.xml***

	HADOOP 2.9.1		*HADOOP 3.0.3*	
Hadoop Daemons	**RPC Port**	**WEB - UI**	**RPC Port**	**WEB - UI**
Namenode	54310	50070	9000	9870
SecondaryNameNode		50090		50090
DataNode		50075	50010	50075
Resource Manager		8088	8030	8088

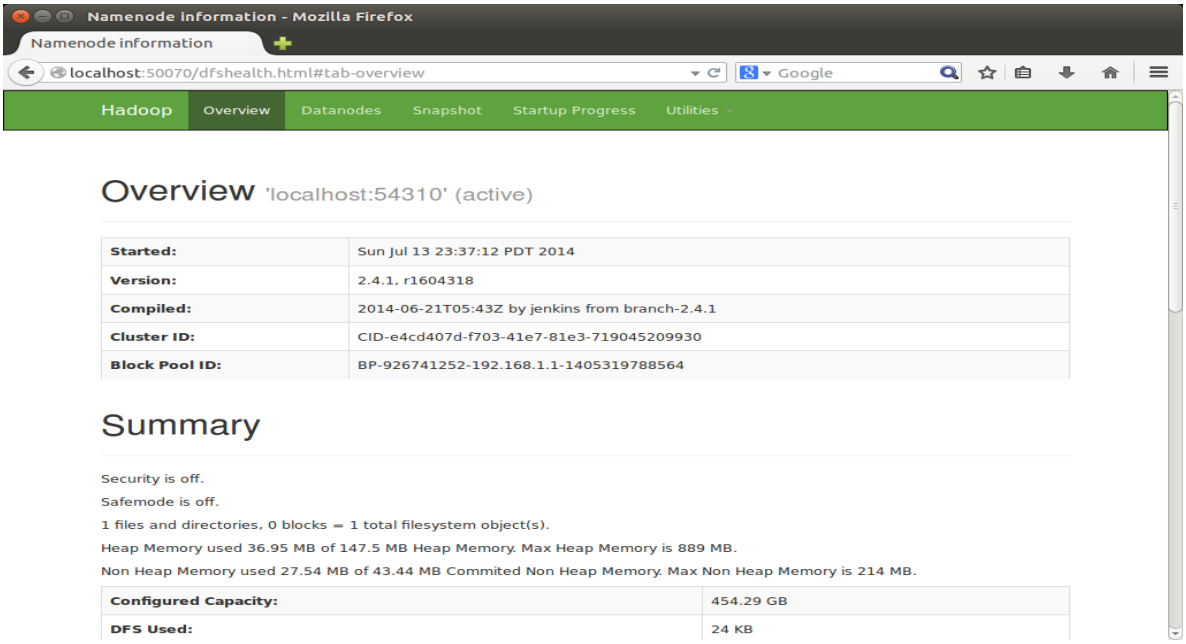
Node Manager		8040	8042

Tabla comparativa entre la versión 2.9.1 y 3.0.3 de Hadoop

Para la opción de CentOS Gnome o KDE puede tener acceso a una interfaz Web con la que se accesa Hadoop en la dirección siguiente:

`http://localhost:50070`

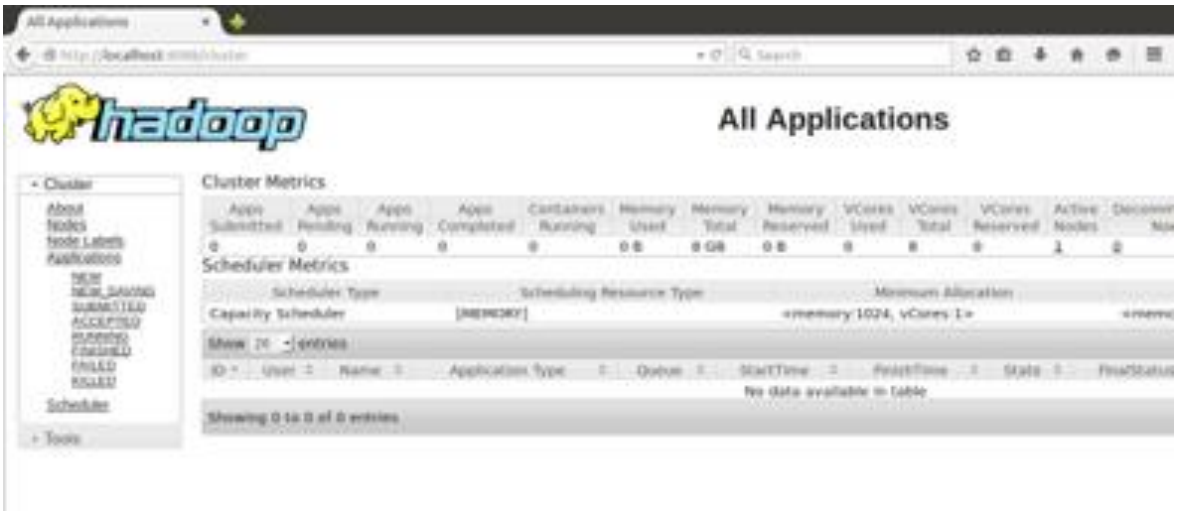
La pantalla que mostrará será algo como la siguiente:



En la liga Utilities se tiene acceso al directorio raíz de Hadoop, ahí se encontrará cada uno de los subdirectorios y archivos.

De forma similar es posible acceder al administrador de recursos en el puerto 8088.

`http://localhost:8088`



Si desea detener los procesos de servicio, utilice el siguiente comando:

```
stop-dfs.sh
stop-yarn.sh
```

La versión de Hadoop se verifica con:

```
hadoop version
```

1.1 Comandos de Hadoop

La ejecución de comandos de Hadoop requiere que esté activo, lo que se hace utilizando los comandos **`start-dfs.sh`** aunque para estas primeras pruebas no es necesario utilizar **`start-yarn.sh`** aún.

Para comprobar que está en funcionamiento Hadoop, podemos comprobar que no despliegue ningún archivo, debido a que aún no se ha copiado ningún archivo hacia Hadoop.

```
hdfs dfs –ls /
```

> ***NOTA:*** *El tamaño máximo de bloque de archivos de Hadoop debe ser máximo la mitad de la capacidad de memoria ram o menos debido a que no se ocupa la computadora únicamente para el cluster.*

Para verificar el tamaño disponible de memoria Ram para poder asignar como tamaño de bloque disponible a Hadoop, ingrese los siguientes comandos:

Saber el tamaño disponible de memoria Ram

```
free –m    //MB
free –g    //GB
free       //bytes
```

El tamaño del bloque se modifica indicando el número de bytes, por ejemplo:

1 Mb =	1 * 1024 * 1024 =	1048576
64 Mb =	64 * 1024 * 1024 =	67108864

500 Mb =	500 * 1024 * 1024 =	524288000
1 Gb =	1024 * 1024 * 1024 =	1073741824
2 Gb =	2 * 1024 * 1024 * 1024 =	2147483648

También es importante tener descargado el administrador de procesos **htop**, el cual necesita repositorios de EPEL para poder ser instalado. Esta es una herramienta de monitorización gráfico para terminal Linux, la cual servirá para saber el tamaño de bloque disponible que usted asignará en la configuración de **hdfs-site.xml**.

Para instalar la versión más actual de EPEL que en este momento es la versión 7-11 introduzca el siguiente comando.

```
sudo yum -y install epel-release
sudo yum install htop
htop
```

Otro comando que podría servirle, el cual tiene como función mostrarle información acerca de su computadora.

```
cat /proc/cpuinfo
```

Para más información acerca de los comandos, ingrese en el siguiente link.
https://www.tutorialspoint.com/es/hadoop/hadoop_hdfs_operations.htm

Crear un archivo en la ruta ~

```
nano /home/hadoop/datos.txt
```

Cree un directorio en hadoop

```
hdfs dfs -mkdir /directorio
```

Para comprobar que el archivo esta es esa dirección

```
hdfs dfs -copyFromLocal /home/hadoop/datos.txt /directorio
```

Para borrar los archivos que ya están utiliza

```
hdfs dfs -rm /directorio/datos.txt
```

Para crear otro archivo y no sobrescribirlo, cambiando el tamaño del bloque es:

```
hdfs dfs -D dfs.block.size=1048576 -copyFromLocal /home/hadoop/datos.txt
/directorio/datos1.txt
```

Para crear el número de replicaciones

```
hdfs dfs -D dfs.replication=2 -copyFromLocal /home/hadoop/datos.txt
  /directorio/datos2.txt
```

Cambiar el tamaño del bloque y el número de replicaciones al mismo tiempo

```
hdfs dfs –D dfs.block.size=52428000 –D dfs.replication=3 –put /home/hadoop/datos.txt
  /directorio/datos4.txt
```

Para poder visualizar todos estos comandos agregados, en la parte grafica ingresamos al navegador e introducimos

```
hdfs dfs –df -h
```

Para recuperar los datos de una cierta dirección

```
Hdfs dfs -get /direccion/datos4.txt
```

Para saber el tamaño del bloque en bytes

```
hdfs dfs –du /directorio/datos.txt
```

Remover carpetas en hadoop

```
hdfs dfs –rmdir /
hdfs dfs –rm –r /
hdfs dfs –rmr /
```