

# INSTALACIÓN DE HADOOP EN UBUNTU

## Antecedentes

Lo primero que tenemos que tener en cuenta es la versión en la cual vamos a realizar la instalación, para obtener esta información debemos tener en cuenta lo siguiente:

Para saber la versión instalada de Linux ejecutamos el siguiente comando:

```
lsb_release -a
```

Si queremos saber información sobre el kernel y el procesador ejecutamos el siguiente comando:

```
uname -a
```

Se mostrará toda la información correspondiente al sistema.

## Instalación y configuración de Java

1.- Para comenzar con la instalación debemos actualizar el repositorio de paquete ejecutando el siguiente comando:

```
sudo apt-get update
```

2.- Procedemos a instalar el jdk:

```
sudo apt-get install default-jdk
```

3.- Verificamos que version de java se instaló:

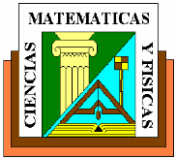
```
java -version
```

## Creación y configuración de usuario para HADOOP

Crearemos un usuario el cual lo usaremos para la instalación de HADOOP, ejecutamos los siguientes comando:

1.- Crear usuario, HADOOP es el nombre que le queramos poner al usuario que vamos a crear:

```
useradd -d /home/hadoop -m hadoop
```



2.- Le otorgamos una contraseña al usuario; debemos ingresar y confirmar la contraseña que deseamos asignar:

```
passwd hadoop
```

3.- Configuramos el usuario y lo agregamos al fichero sudoers:

```
usermod -a -G sudo hadoop
```

```
usermod -s /bin/bash hadoop
```

4.- Le otorgamos permisos al usuario que acabamos de crear:

## Instalación y configuración de SSH:

1.- Para comenzar con la instalación debemos actualizar el repositorio de paquete ejecutando el siguiente comando:

```
sudo apt-get update
```

2.- Procedemos a instalar el SSH:

```
sudo apt-get install ssh
```

3.- Ingresamos con el usuario que acabamos de crear:

```
su - hadoop
```

4.- Asignamos una contraseña vacía para el usuario creado:

```
ssh-keygen -t rsa -P ""
```

5.- Comprobamos el acceso a ssh:

```
ssh localhost
```

6.- Verificamos la configuración del servicio ssh en el siguiente fichero, PubkeyAuthentication (que debe tener el valor **yes**) y AllowUsers (Si está activa, añadir el usuario **hduser**):

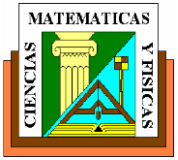
```
sudo cd nano /etc/ssh/sshd_config
```

7.- Si realizamos algún cambio debemos ejecutar el siguiente comando:

```
sudo /etc/init.d/ssh reload
```

8.- Una vez que ya está todo bien configurado, y ha podido conectarse correctamente a localhost, puede volver a salir ejecutando el siguiente comando.

```
exit
```



## Desactivamos IPV6

1.- Tenemos que desactivar IPV6 antes de instalar HADOOP, ingresamos al siguiente fichero y lo modificamos:

```
sudo cd nano /etc/sysctl.conf
```

Agregamos la siguiente configuración:

```
# disable ipv6  
  
net.ipv6.conf.all.disable_ipv6 = 1  
net.ipv6.conf.default.disable_ipv6 = 1  
net.ipv6.conf.lo.disable_ipv6 = 1
```

2.-En este paso tenemos que reiniciar el sistema para que los cambios apliquen, luego podremos verificar con el siguiente comando:

```
cat /proc/sys/net/ipv6/conf/all/disable_ipv6
```

*Esto devuelve uno de estos valores: 0 si IPv6 está activo, o 1 si está desactivado.*

## Instalación de APACHE HADOOP

Para poder instalar apache HADOOP hay que proceder a descargarlo de la siguiente manera:

1.- Ingresamos a la siguiente ruta:

```
# cd /usr/local
```

2.- Descargamos el siguiente archivo:

```
# wget http://apache.claz.org/hadoop/common/hadoop-2.4.1/
```

3.- Obtendremos el siguiente archivo:

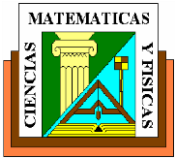
```
hadoop-2.4.1.tar.gz
```

4.- Descomprimos el archivo de la siguiente manera:

```
# tar xzf hadoop-2.4.1.tar.gz
```

5.- Asignamos el archivo al usuario creado para que tenga acceso:

```
# mv hadoop-2.4.1/* to hadoop/  
# exit
```



## Diferentes nodos de configuración

- ❑ **Local/Modo autónomo:** Después de descargar HADOOP en su sistema, por defecto, se configura en modo independiente se puede ejecutar como un solo proceso java.
- ❑ **Pseudo Modo Distribuido:** es una simulación distribuida en una sola máquina. Cada demonio como Hadoop hdfs, hilados, MapReduce, etc. , se ejecute como un proceso java independiente. Este modo es útil para el desarrollo.
- ❑ **Modo Totalmente Distribuida :** Este modo es completamente distribuida con un mínimo de dos o más máquinas como un clúster. Ya hablaremos de este modo en detalle en los próximos capítulos.

## Configuración nodo autónomo

1.-Puede establecer las variables de entorno Hadoop anexar los siguientes comandos para `~/bashrcarchivo`

```
export HADOOP_HOME=/usr/local/hadoop
```

2.- Verificamos que el servicios está instalado correctamente y debe tener la siguiente configuración:

Hadoop 2.4.1

Subversion <https://svn.apache.org/repos/asf/hadoop/common> -r 1529768

Compiled by hortonmu on 2013-10-07T06:28Z

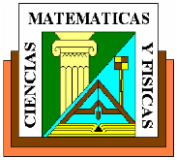
Compiled with protoc 2.5.0

From source with checksum 79e53ce7994d1628b240f09af91e1af4

3.-Verificamos la configuración de los siguientes archivos:

core-site.xml

```
<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/tmp/hadoop/data</value>
<description>Location for HDFS.</description>
</property>
```



```
<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:54310</value>
  <description>The name of the default file system. A URI whose
  scheme and authority determine the FileSystem implementation. </description>
</property>
</configuration>
```

hdfs-site.xml

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
  <description>Default number of block replications.</description>
</property>
</configuration>
```

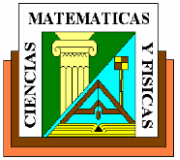
mapred-site.xml

```
<configuration>
<property>
  <name>mapred.job.tracker</name>
  <value>localhost:54311</value>
  <description>The host and port that the MapReduce job tracker runs at.
  </description>
</property>
</configuration>
```

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>
```

vi /root/.bashrc



```
export HADOOP_HOME=/home/ubuntu/serverHadoop/hadoop-2.7.3
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
unalias fs &> /dev/null
alias fs="hadoop fs"
unalias hls &> /dev/null
alias hls="fs -ls"
lzohead () {
  hadoop fs -cat $1 | lzop -dc | head -1000 | less
}
export PATH=$PATH:$HADOOP_HOME/bin
```

## Bibliografía

<http://howto.gandasoftwarefactory.com/desarrollo-software/2014/como-instalar-apache-hadoop-ubuntu-linux-20141209/>  
<https://www.adictosaltrabajo.com/tutoriales/hadoop-first-steps/>  
<http://knowdimension.com/blog/administracion-de-datos/instalando-hadoop-2-7-en-un-sistema-ubuntu/>  
<https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html>  
[https://www.tutorialspoint.com/es/hadoop/hadoop environment setup .htm](https://www.tutorialspoint.com/es/hadoop/hadoop%20environment%20setup.htm)  
<http://dba.mx/2016/08/instalar-hadoop-2-6-en-linux-un-solo-nodo/>  
<https://slimbook.es/tutoriales/linux/86-anadir-usuario-al-fichero-sudoers>  
<https://www.sololinux.es/crear-un-usuario-con-permisos-de-root/>  
<https://www.youtube.com/watch?v=Nb1sinaTlmo>  
<http://telesalud.udea.edu.co/blog/wp-content/uploads/2015/09/Instalaci%C3%B3n-Hadoop-2.pdf>