Guía de instalación de Hadoop y JDK en Ubuntu

Primero se instala Ubuntu con el comando wsl –install.

```
PS C:\Users\matt0> wsl --install
Instalando: Ubuntu
Se ha instalado Ubuntu.
Iniciando Ubuntu...
Installing, this may take a few minutes...
Please create a default UNIX user account. The username does not need to match your Windows username.
For more information visit: https://aka.ms/wslusers
```

Pedirá un nombre de usuario y contraseña, para efectos prácticos le pondremos hadoop a ambos, en caso contrario habrá que crear otro usuario con estas credenciales con los comandos useradd hadoop y passwd hadoop.

```
Enter new UNIX username: matidiaz
New password:
Retype new password:
No password has been supplied.
New password:
Retype new password:
Sorry, passwords do not match.
passwd: Authentication token manipulation error
passwd: password unchanged
Try again? [y/N] y
New password:
Retype new password:
passwd: password updated successfully
Installation successful!
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 5.15.146.1-microsoft-standard-WSL2 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage
```

Para entrar como administrador (root) ponemos el comando sudo -i.

```
matidiaz@LAPTOP-S84FBVUJ: $ sudo -i
[sudo] password for matidiaz:
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 5.15.146.1-microsoft-standard-WSL2 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage


This message is shown once a day. To disable it please create the
/root/.hushlogin file.
root@LAPTOP-S84FBVUJ:~# useradd hadoop
root@LAPTOP-S84FBVUJ:~# passwd hadoop
New password:
Retype new password:
passwd: password updated successfully
root@LAPTOP-S84FBVUJ:~#
```

Una vez estemos en el usuario hadoop, entramos a la carpeta opt y creamos una carpeta llamada hadoop, si nos da error de permisos agregamos sudo al comienzo del comando, luego le damos permiso al usuario hadoop para manipular la carpeta y su contenido.

```
hadoop@LAPTOP-S84FBVUJ:~$ cd /opt
hadoop@LAPTOP-S84FBVUJ:/opt$ mkdir hadoop
mkdir: cannot create directory ‘hadoop’: Permission denied
hadoop@LAPTOP-S84FBVUJ:/opt$ sudo mkdir hadoop
[sudo] password for hadoop:
hadoop@LAPTOP-S84FBVUJ:/opt$ sudo chown -R hadoop:hadoop hadoop
```

Luego descargamos el software de hadoop (Para este desarrollo se ocupó la versión 3.4.0 en binario para Linux), se crea una carpeta en la ruta home/hadoop/ llamada Download y se mueve el software descargado a esta carpeta, posteriormente y devuelta en la ventana de comandos se ingresa el comando sudo tar xvf /home/hadoop/Download/hadoop-3.4.0.tar.gz --strip-components=1 para descomprimir e instalar el software.

```
hadoop@LAPTOP-S84FBVUJ:~$ sudo tar xvf /home/hadoop/Download/hadoop-3.4.0.tar.gz --strip-components=1
```

Hacemos ls -l /opt/ para ver que todo se instaló correctamente

```
hadoop@LAPTOP-S84FBVUJ:/opt$ ls -l /opt/
total 224
-rw-r--r--  1 root   root   31765 Mar  4 02:46 BUILDING.txt
-rw-r--r--  1 root   root   23756 Mar  4 02:44 LICENSE-binary
-rw-r--r--  1 root   root   15696 Mar  4 02:44 LICENSE.txt
-rw-r--r--  1 root   root   27165 Mar  4 02:44 NOTICE-binary
-rw-r--r--  1 root   root    1541 Mar  4 02:44 NOTICE.txt
-rw-r--r--  1 root   root     175 Mar  4 02:44 README.txt
drwxr-xr-x  7 root   root    4096 Apr  3 09:41 dev-support
drwxr-xr-x  2 hadoop hadoop  4096 Apr  3 09:31 hadoop
drwxr-xr-x  3 root   root    4096 Apr  3 09:41 hadoop-assemblies
drwxr-xr-x  3 root   root    4096 Apr  3 09:41 hadoop-build-tools
drwxr-xr-x  9 root   root    4096 Apr  3 09:41 hadoop-client-modules
drwxr-xr-x  5 root   root    4096 Apr  3 09:41 hadoop-cloud-storage-project
drwxr-xr-x 10 root   root    4096 Apr  3 09:41 hadoop-common-project
drwxr-xr-x  2 root   root    4096 Apr  3 09:41 hadoop-dist
drwxr-xr-x  8 root   root    4096 Apr  3 09:41 hadoop-hdfs-project
drwxr-xr-x  9 root   root    4096 Apr  3 09:41 hadoop-mapreduce-project
drwxr-xr-x  3 root   root    4096 Apr  3 09:41 hadoop-maven-plugins
```

Posteriormente se descarga el software de jdk (En versión 11), esta también debe descargarse en su versión en binario para Linux. Se escribe el comando para descomprimir tar xvf /home/hadoop/Download/openjdk…tar.gz -C /opt/hadoop.

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ tar xvf /home/hadoop/Download/openjdk-11_linux-x64_bin.tar.gz -C /opt/hadoop
jdk-11/bin/jaotc
jdk-11/bin/jar
jdk-11/bin/jarsigner
jdk-11/bin/java
jdk-11/bin/javac
jdk-11/bin/javadoc
jdk-11/bin/javap
jdk-11/bin/jcmd
jdk-11/bin/jconsole
jdk-11/bin/jdb
jdk-11/bin/jdeprscan
jdk-11/bin/jdeps
jdk-11/bin/jhsdb
```

Una vez termine, con el comando ls debería mostrarnos una carpeta llamada jdk-11, la cual renombraremos a jdk con el comando mv jdk-11 jdk para efectos prácticos.

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ ls
jdk-11
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ mv jdk-11 jdk
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ ls
jdk
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$
```

Posteriormente iremos a la ruta /home/hadoop y abriremos con un editor de texto el archivo .bashrc, insertaremos las siguientes cuatro líneas al final del documento:

export JAVA_HOME=/opt/hadoop/jdk

export PATH=$PATH:$JAVA_HOME/bin

export HADOOP_HOME=/opt/hadoop/hadoop

export PATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

Nota: La primera y tercera línea tienen las rutas donde fueron instalados hadoop y jdk, si se eligió otra ruta se deben modificar esas líneas. Luego se guarda y se cierra el archivo y se ingresa el comando source /home/hadoop/.bashrc para actualizar dicho archivo en la ventana de comandos, verificamos que todo esté en orden con java -version, si nos entrega las tres líneas de abajo, significa que vamos bien.

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop/jdk/bin$ source /home/hadoop/.bashrc
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop/jdk/bin$ java -version
openjdk version "11" 2018-09-25
OpenJDK Runtime Environment 18.9 (build 11+28)
OpenJDK 64-Bit Server VM 18.9 (build 11+28, mixed mode)
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop/jdk/bin$
```

Devuelta en /opt/hadoop, renombraremos la otra carpeta llamada hadoop-3.4.0 a hadoop con el mismo comando que renombramos la carpeta jdk.

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ ls
hadoop-3.4.0  jdk
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ mv hadoop-3.4.0/ hadoop
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ ls
hadoop  jdk
```

Posteriormente escribimos el comando hadoop -h.

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ hadoop -h
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
 or    hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
  where CLASSNAME is a user-provided Java class

  OPTIONS is none or any of:

--config dir                     Hadoop config directory
--debug                          turn on shell script debug mode
--help                           usage information
buildpaths                       attempt to add class files from build tree
hostnames list[,of,host,names]   hosts to use in worker mode
hosts filename                   list of hosts to use in worker mode
loglevel level                   set the log4j level for this command
workers                          turn on worker mode

  SUBCOMMAND is one of:
```

Si todo se instaló correctamente, ponemos hadoop versión y nos debe entregar la versión de hadoop que está instalada.

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ hadoop version
Hadoop 3.4.0
Source code repository git@github.com:apache/hadoop.git -r bd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled by root on 2024-03-04T06:35Z
Compiled on platform linux-x86_64
Compiled with protoc 3.21.12
From source with checksum f7fe694a3613358b38812ae9c31114e
This command was run using /opt/hadoop/hadoop/share/hadoop/common/hadoop-common-3.4.0.jar
```

Luego entraremos en la carpeta hadoop y escribiremos el comando cp etc/hadoop/*.xml /tmp/input, esto copiará todos los archivos .xml de la carpeta hadoop a la carpeta input.

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop$ cd hadoop/
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop/hadoop$ cp etc/hadoop/*.xml /tmp/input/
```

Luego escribimos el comando hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.0.jar grep /tmp/input/ /tmp/output ‘dfs[a-z.]+’

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop/hadoop$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.0.jar grep /tmp/
input/ /tmp/output 'dfs[a-z.]+'
2024-04-03 11:06:54,962 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-04-03 11:06:55,068 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-04-03 11:06:55,069 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-04-03 11:06:55,282 INFO input.FileInputFormat: Total input files to process : 10
2024-04-03 11:06:55,316 INFO mapreduce.JobSubmitter: number of splits:10
2024-04-03 11:06:55,510 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local139140322_0001
2024-04-03 11:06:55,511 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-03 11:06:55,707 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
```

Para confirmar si procesó todo sin errores haremos una lectura a la carpeta output con el comando ls -l /tmp/output/, el cual debe entregarnos dos archivos llamados _SUCCESS y part-r-00000

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop/hadoop$ ls -l /tmp/output/
total 4
-rw-r--r-- 1 hadoop hadoop  0 Apr  3 11:06 _SUCCESS
-rw-r--r-- 1 hadoop hadoop 11 Apr  3 11:06 part-r-00000
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop/hadoop$
```

Nos centramos en este ultimo archivo, escribiendo el comando cat /tmp/output/part-r-00000 nos entregará una lista que puede contener uno o varios datos.

```
hadoop@LAPTOP-S84FBVUJ:/opt/hadoop/hadoop$ cat /tmp/output/part-r-00000
1       dfsadmin
```

Procederemos a configurar el SSH, salimos a la ruta base (Para salir usamos el comando cd .. las veces que sean necesarias), y ponemos el comando ssh-keygen, nos preguntará algunas cosas, pero dejamos todo en blanco y presionamos enter.

```
hadoop@LAPTOP-S84FBVUJ:/$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:b+K2a+AdHCaBWph0BmNL4/at6o9T4tHV3eXYQICRWEQ hadoop@LAPTOP-S84FBVUJ
The key's randomart image is:
+---[RSA 3072]----+
| .B+o. =E+.o.    |
| +o*o o o   . .  |
|  +o   o . . *   |
| ... .o + . o o  |
|   ....+S.       |
|  o o.. o.       |
| . +.. o..o      |
|  oo  ..+o       |
| .+o.  o+o       |
+----[SHA256]-----+
hadoop@LAPTOP-S84FBVUJ:/$
```

Listamos el contenido de la carpeta .ssh con el comando ls -l .ssh y encontraremos dos archivos.

```
hadoop@LAPTOP-S84FBVUJ:~$ ls -l .ssh
total 8
-rw------- 1 hadoop hadoop 2610 Apr  3 12:51 id_rsa
-rw-r--r-- 1 hadoop hadoop  576 Apr  3 12:51 id_rsa.pub
hadoop@LAPTOP-S84FBVUJ:~$
```

Entramos a la carpeta .ssh y utilizando uno de los archivos escribimos el comando cp id_rsa.pub authorized_keys, posteriormente escribimos ssh NOMBRE_DE_LA_MAQUINA (El nombre debiera salir siempre al comienzo de la línea de código, en el texto que sale en verde después del símbolo @, en mi caso se llama LAPTOP-S84FBVUJ, por lo que debo ingresar el comando ssh LAPTOP-S84FBVUJ). Si rechaza la conexión significa que no hemos instalado lo necesario para conectarnos, por lo que deberemos

escribir sudo apt update.

```
hadoop@LAPTOP-S84FBVUJ:~$ cd .ssh
hadoop@LAPTOP-S84FBVUJ:~/.ssh$ cp id_rsa.pub authorized_keys
hadoop@LAPTOP-S84FBVUJ:~/.ssh$ ssh LAPTOP-S84FBVUJ
ssh: connect to host laptop-s84fbvuj port 22: Connection refused
hadoop@LAPTOP-S84FBVUJ:~/.ssh$ sudo apt update
[sudo] password for hadoop:
Get:1 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Hit:2 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [119 kB]
Get:4 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [1303 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [109 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy/universe amd64 Packages [14.1 MB]
Get:7 http://security.ubuntu.com/ubuntu jammy-security/main Translation-en [233 kB]
Get:8 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages [1616 kB]
Get:9 http://security.ubuntu.com/ubuntu jammy-security/restricted Translation-en [271 kB]
Get:10 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [852 kB]
Get:11 http://security.ubuntu.com/ubuntu jammy-security/universe Translation-en [163 kB]
```

Luego escribiremos sudo apt install openssh-server y nos preguntará si queremos continuar, a lo que escribiremos la letra ‘Y’ y luego apretamos enter.

```
hadoop@LAPTOP-S84FBVUJ:~/.ssh$ sudo apt install openssh-server
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  libwrap0 ncurses-term openssh-client openssh-sftp-server ssh-import-id
Suggested packages:
  keychain libpam-ssh monkeysphere ssh-askpass molly-guard
The following NEW packages will be installed:
  libwrap0 ncurses-term openssh-server openssh-sftp-server ssh-import-id
The following packages will be upgraded:
  openssh-client
1 upgraded, 5 newly installed, 0 to remove and 97 not upgraded.
Need to get 1705 kB of archives.
After this operation, 6161 kB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-client amd64 1:8.9p1-3ubuntu0.6 [906 kB]
Get:2 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-sftp-server amd64 1:8.9p1-3ubuntu0.6 [38.7 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy/main amd64 libwrap0 amd64 7.6.q-31build2 [47.9 kB]
```

Con esto ya instalado, volvemos a escribir el comando anterior y ahora debería iniciar la conexión, nos preguntará si estamos seguros de querer continuar con la conexión, aquí escribiremos yes, y confirmamos con enter. Si muestra lo mismo que en el pantallazo de abajo, habremos realizado una conexión exitosa.

```
hadoop@LAPTOP-S84FBVUJ:~/.ssh$ ssh LAPTOP-S84FBVUJ
The authenticity of host 'laptop-s84fbvuj (127.0.1.1)' can't be established.
ED25519 key fingerprint is SHA256:6i7Bc0dQ4GFUxANru19VQpmXs2MOoYGGV1PPfDQ1k2k.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'laptop-s84fbvuj' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 5.15.146.1-microsoft-standard-WSL2 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge
Last login: Wed Apr  3 09:28:32 2024
hadoop@LAPTOP-S84FBVUJ:~$
```