Instalacion de Apacha Spark en Ubuntu

Autor: Anderson Tenorio Acha

Paso1: verificamos la versión de jdk y hadoop que tenemos instaladas en nuestra distro, para eso usamos el comando "java -version" y "hadoop version"

```
tenorio@nodo1:~$ java -version
openjdk version "1.8.0_452"
OpenJDK Runtime Environment (build 1.8.0_452-8u452-ga~us1-0ubuntu1~24.04-b09)
OpenJDK 64-Bit Server VM (build 25.452-b09, mixed mode)
tenorio@nodo1:~$ hadoop version
Hadoop 2.10.2
Subversion Unknown -r 965fd380006fa78b2315668fbc7eb432e1d8200f
Compiled by ubuntu on 2022-05-24T22:35Z
Compiled with protoc 2.5.0
From source with checksum d3ab737f7788f05d467784f0a86573fe
This command was run using /opt/hadoop/share/hadoop/common/hadoop-common-2.10.2.
jar
tenorio@nodo1:~$
```

Paso 2: una vez verificado entramos al siguiente enlace https://archive.apache.org/dist/spark/ y descargamos la versión que es mas compatible para nuestra versión de hadoop, en mi caso usare la version 3.3.2

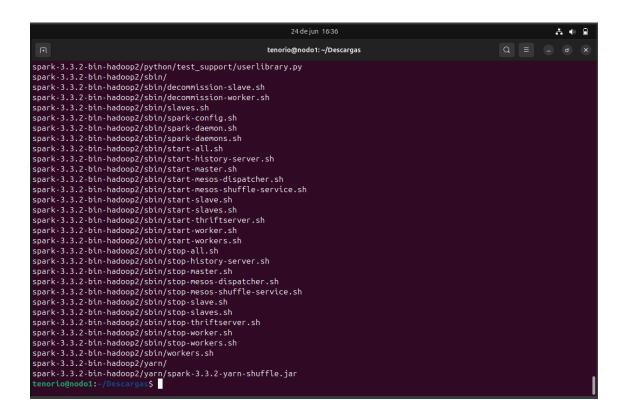
<u>spark-3.2.2/</u>	2022-07-15 14:43 -	
<u>spark-3.2.3/</u>	2022-11-28 18:04 -	
<u>spark-3.2.4/</u>	2023-04-13 13:46 -	
<u>spark-3.3.0/</u>	2022-06-17 11:11 -	
	2022 10 25 07.22	
<u>spark-3.3.2/</u>	2023-02-15 21:15 -	
<u>spark-3.3.3/</u>	2025-08-21 00:25 -	•
<u>spark-3.3.4/</u>	2023-12-16 00:26 -	
<u>spark-3.4.0/</u>	2023-04-13 17:54 -	
<u>spark-3.4.1/</u>	2023-06-23 08:23 -	



Paso 3: una vez descargado el spark, entramos a la carpeta Descargas desde nuestra terminal usando el comando "cd ~/Descargas" y verificamos si está ahí con el comando "ls"

```
tenorio@nodo1:~$ cd ~/Descargas
tenorio@nodo1:~/Descargas$ ls
hadoop-2.10.2.tar.gz prueba_final.txt
tenorio@nodo1:~/Descargas$
spark-3.3.2-bin-hadoop2.tgz
```

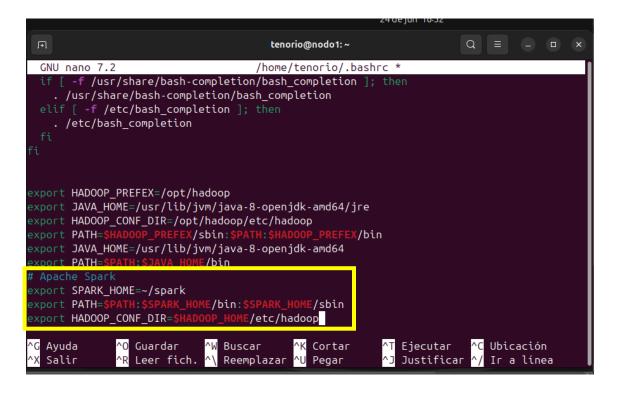
Paso 4: ahí mismo usamos el comando "tar -xvzf spark-3.3.2-bin-hadoop2.tgz" para extraer los archivos dentro de la carpeta, recordemos que el "x" representación la extracción, el "v" muestra en pantalla los archivos que se extraen, el "z" indica que el archivo esta comprimido en gzip y por ultimo el "f" le indica a tar que se usara un archivo.

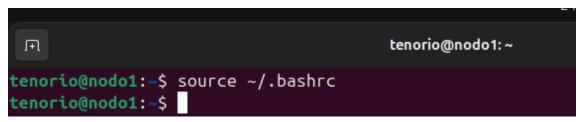


Paso 5: ahora movemos el archivo descomprimido con el comando "mv mv spark-3.3.2-bin-hadoop2 ~/spark" esto cambiara el nombre de la carpeta a "spark" y lo alojara en "home(el inicio de todo el terminal"



Paso 6: editamos el archivo bashrc con el comando "nano ~/.bashrc" y le añadimos lo siguiente en la parte final los siguiente y lo guardamos con el "ctrl + O", una vez guardado usamos el comando "source ~/.bashrc" esto permitirá usar comandos de "spark" desde cualquier parte.





Paso 7: probamos si esta todo correcto usando el comando "spark-shell" y el comando para Python con el comando ""pyspark"

