Explicación previa

Vamos a crear una máquina virtual Ubuntu Server con Spark y sus dependencias para funcionar.

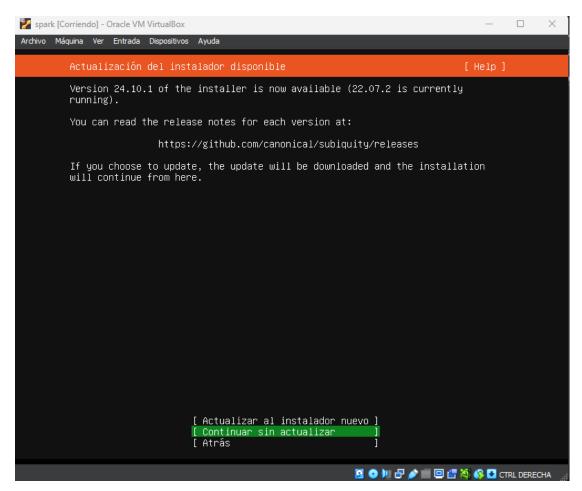
Requisitos:

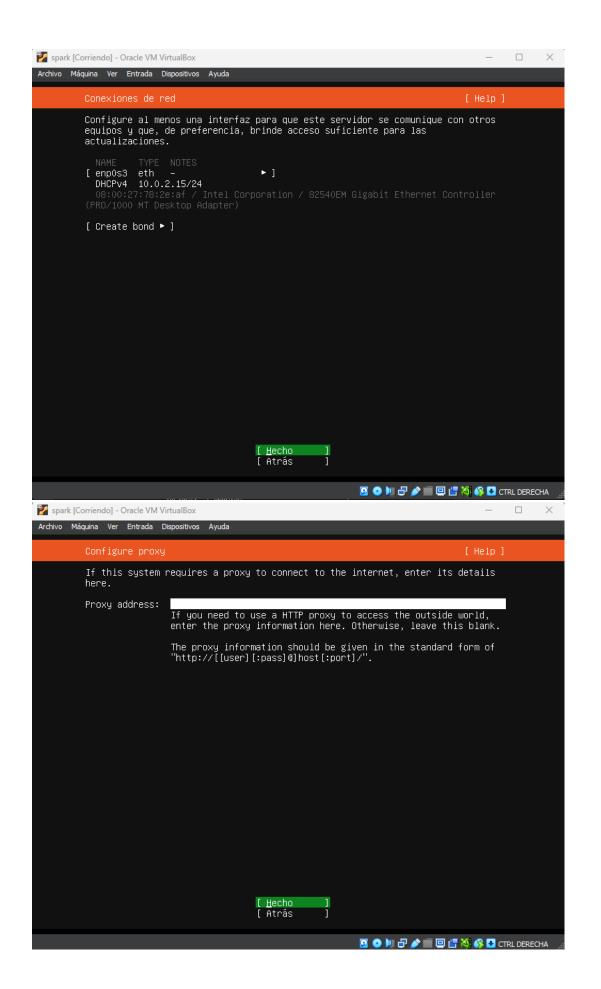
- En host:
 - o Ubuntu desktop [20.04]
 - Virtual Box
- En la máquina virtual:
 - o Spark
 - o Java JDK
 - o Python

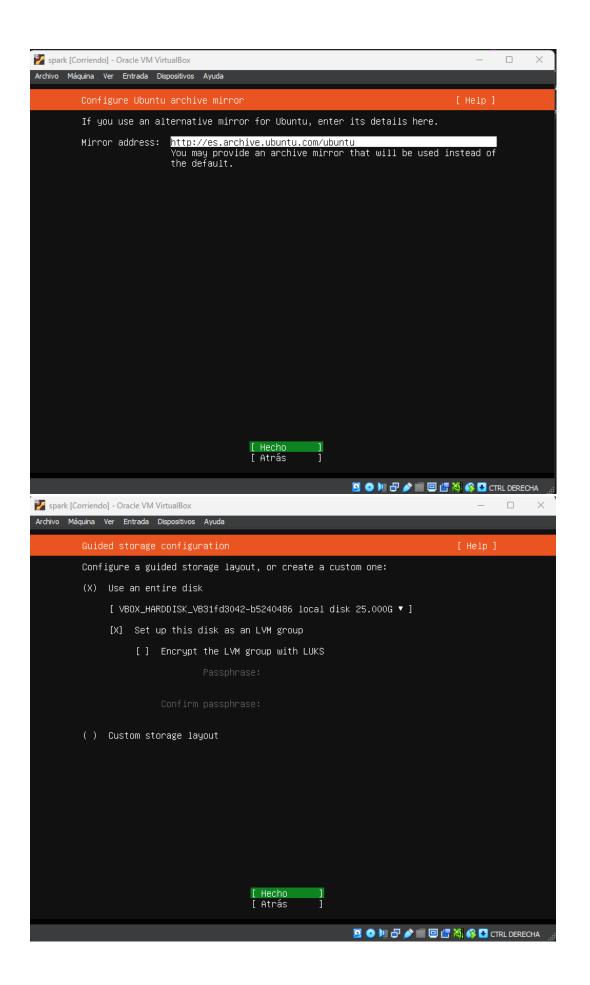
Creación de la máquina virtual en Virtual Box

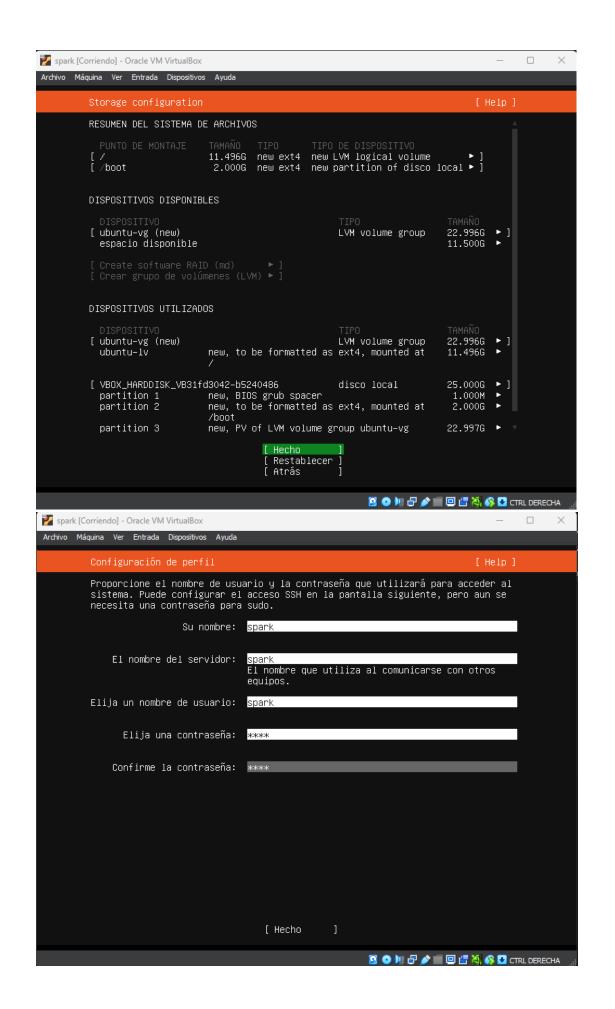
Configuración:

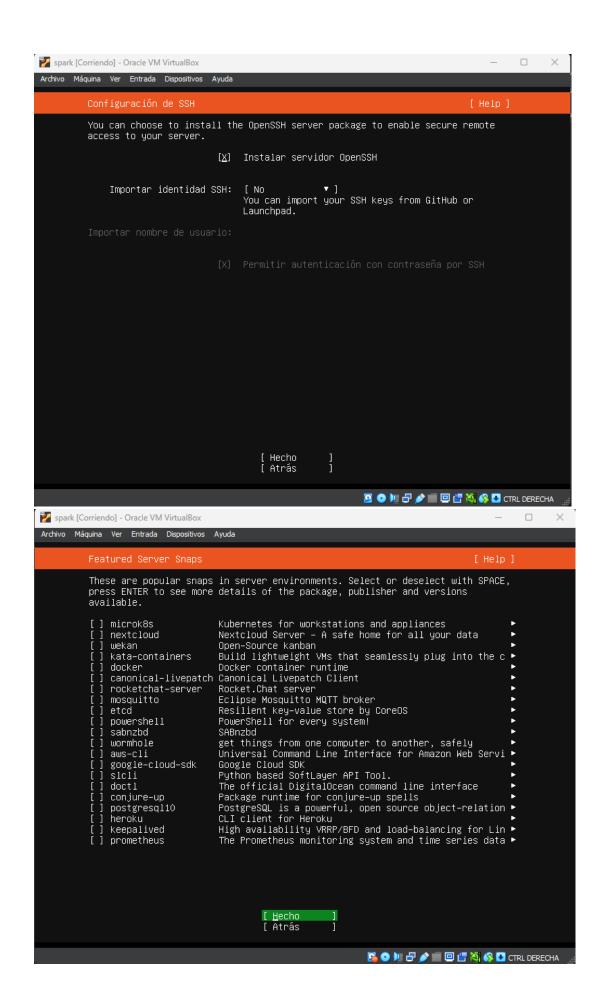
• Memoria 4096, 2 CPUs, 25 Gb de disco duro











Pasos para instalación de Spark en el servidor

```
spark@spark:~$ sudo apt update
[sudo] password for spark: _
spark@spark:~$ sudo apt upgrade_
```

Instalar java

spark@spark:~\$ sudo apt install openjdk–8–jdk_

```
spark@spark:~$ java –version
openjdk version "1.8.0_432"
OpenJDK Runtime Environment (build 1.8.0_432–8u432–ga~us1–Oubuntu2~20.04–ga)
OpenJDK 64–Bit Server VM (build 25.432–bga, mixed mode)
```

Instalar Python3

```
spark@spark:~$ sudo apt install python3
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
python3 ya está en su versión más reciente (3.8.2–Oubuntu2).
fijado python3 como instalado manualmente.
O actualizados, O nuevos se instalarán, O para eliminar y O no actualizados.
```

```
spark@spark:~$ python3 ––version
Python 3.8.10
```

Descargar version de spark

```
spark@spark:~$ wget https://downloads.apache.org/spark/spark–3.5.3/spark–3.5.3–bin–hadoop3.tgz_
```

Descomprimir tar y comprobar archivos

```
spark@spark:~$ tar –xvf spark–3.5.3–bin–hadoop3.tgz_
```

```
spark@spark:~$ ls
spark—3.5.3—bin—hadoop3 spark—3.5.3—bin—hadoop3.tgz
spark@spark:~$ ls spark—3.5.3—bin—hadoop3
bin data jars LICENSE NOTICE R RELEASE yarn
conf examples kubernetes licenses python README.md sbin
```

Mover el contenido de tar a /opt/spark

```
spark@spark:~$ sudo mv spark–3.5.3–bin–hadoop3 /opt/spark
spark@spark:~$ find / .bashrc_
```

Añadir variables de entorno al archivo .bashrc

```
spark@spark:~$ sudo nano .bashrc
```

```
GNU nano 4.8
                                                                                                                       Modified
                                                                .hashno
# some more ls aliases
alias ll='ls -alF'
alias la='ls -A'
alias l='ls -CF'
# Add an "alert" alias for long running commands. Use like so:
# sleep 10; alert
alias alert='notify–send ––urgency=low –i "$([ $? = 0 ] && echo terminal || echo error)" "$(history>
# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash–doc/examples in the bash–doc package.
 if [ -f ~/.bash_aliases ]; then
. ~/.bash_aliases
  enable programmable completion features (you don't need to enable
  this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).

if ! shopt –oq posix; then

if [ -f /usr/share/bash–completion/bash_completion ]; then
  ./usr/share/bash-completion/bash_completionelif [ -f /etc/bash_completion ]; then
     ./etc/bash_completion
 export SPARK_HOME=/opt/spark
 export PATH=
                               /bin:
 export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
 export PATH=$3
                            E/bin:
^G Get Help
^X Exit
                                                                                                Cur Pos M–V Undo
Go To Line M–E Redo
                  ^O Write Out
^R Read File
                                         Where Is
                                                           Cut Text
                                                                              Justify
                                                                                             ^C Cur Pos
                                         Replace
                                                           Paste Text
                                                                              To Spell
```

Iniciar spark