Cluster con Apache Spark

# Descripción del proyecto

Vamos a crear un cluster en linux con apache spark, nos va a permitir trabajar con grandes conjuntos de datos,nos va a permitir una escalabilidad y es tolerante a fallos, cuantos más servidores haya más potencia tendrá el cluster.

# Requisitos Software:

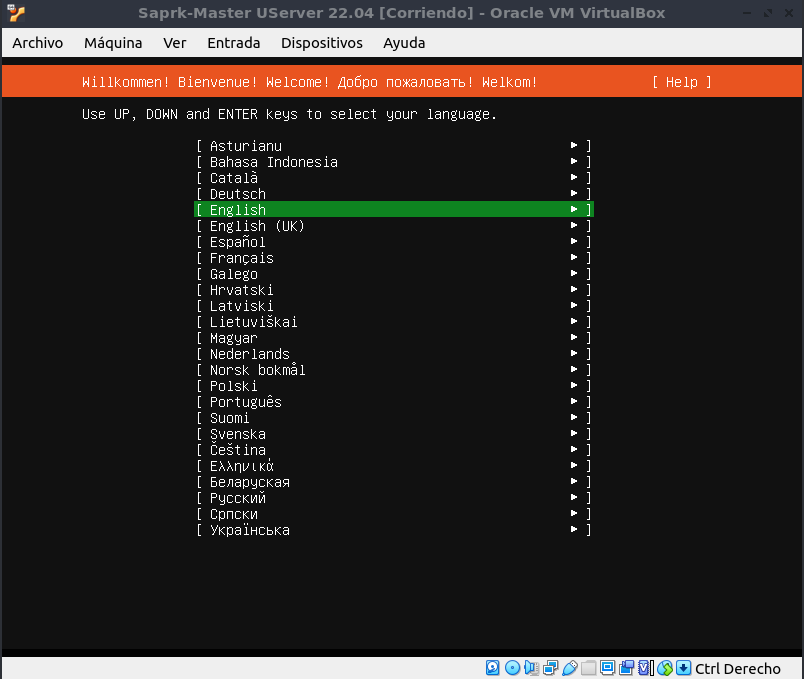
1. Java 7 o Java 8
2. Scala
3. SSH
4. Apache Spark pre-compilado para Hadoop.

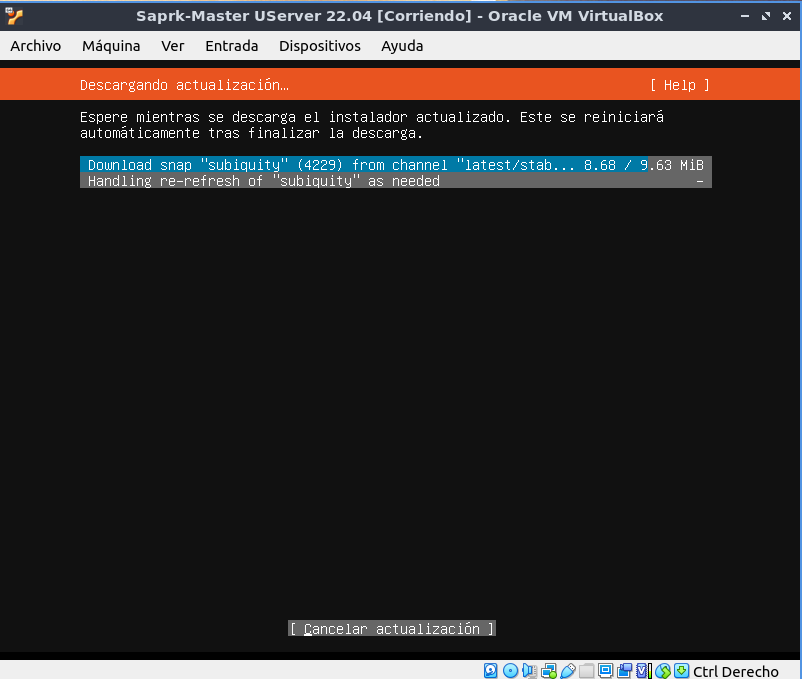
#### Arquitectura de un cluster de Apache Spark

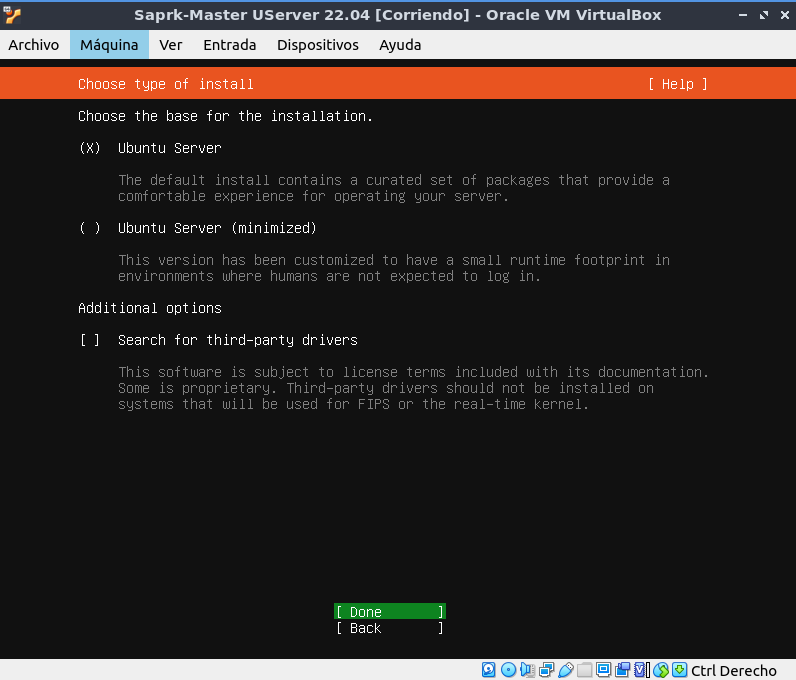
## 

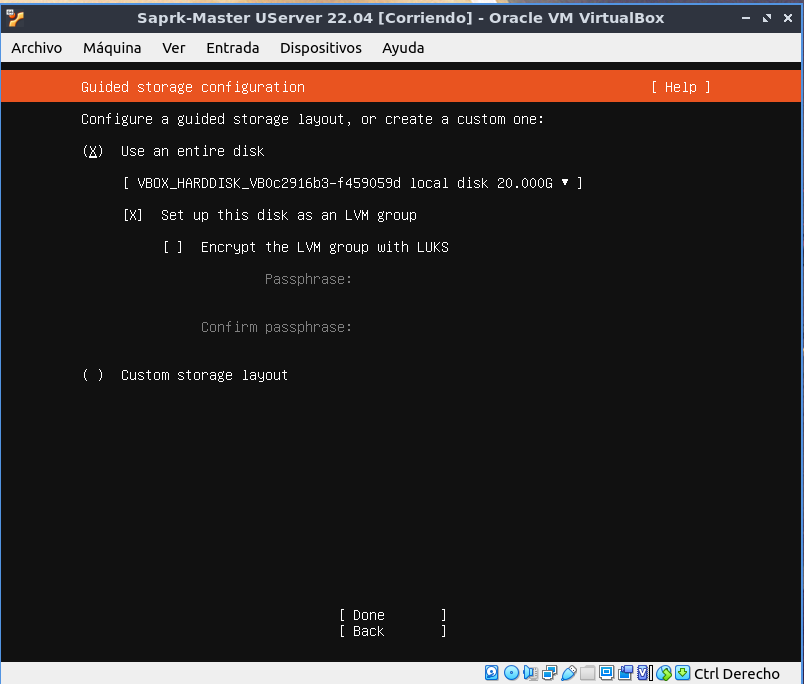
<https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/>→ Apache Spark

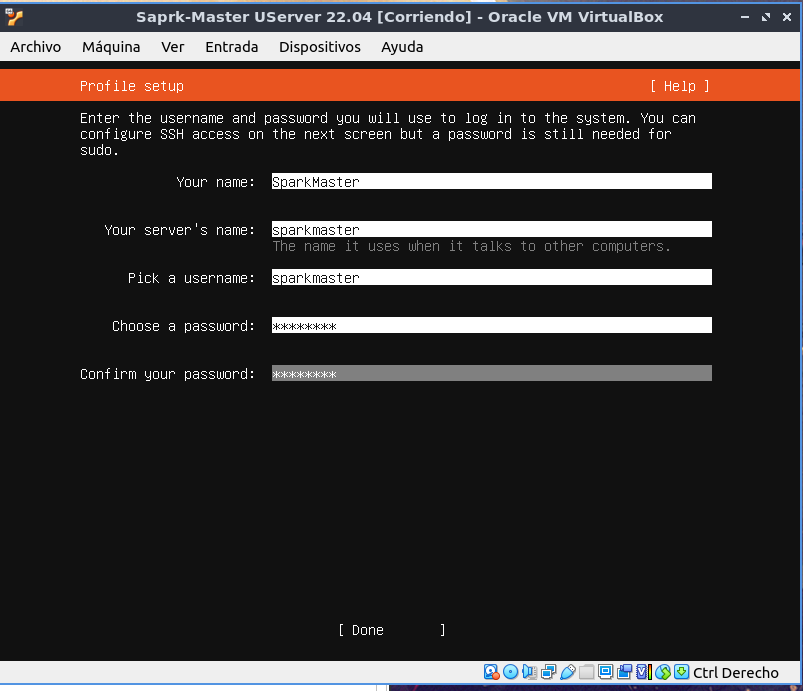
**Instalación de Apache Spark en Ubuntu Server 22.04:**

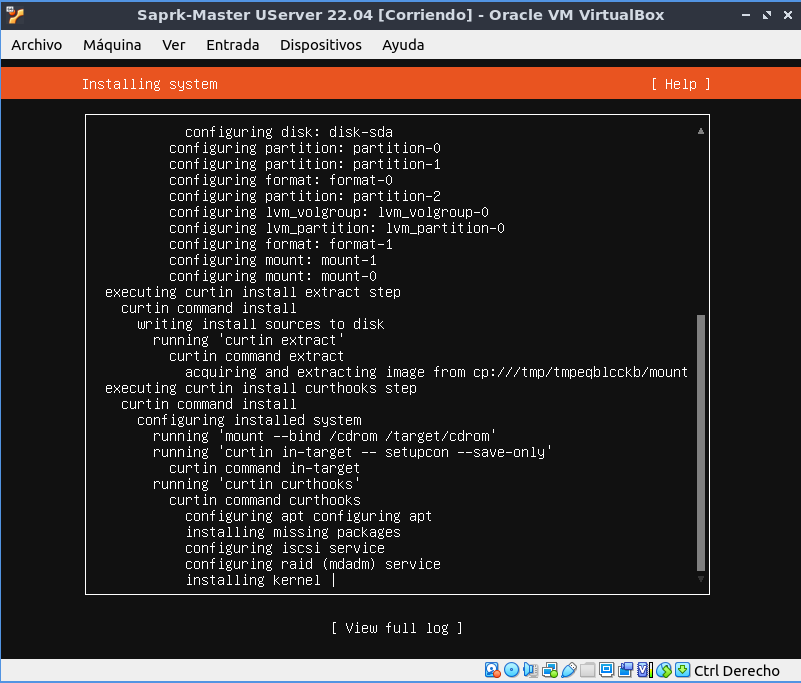






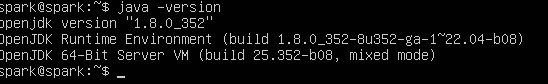






Instalar Java 8:

sudo apt-get -y install openjdk-8-jdk-headless

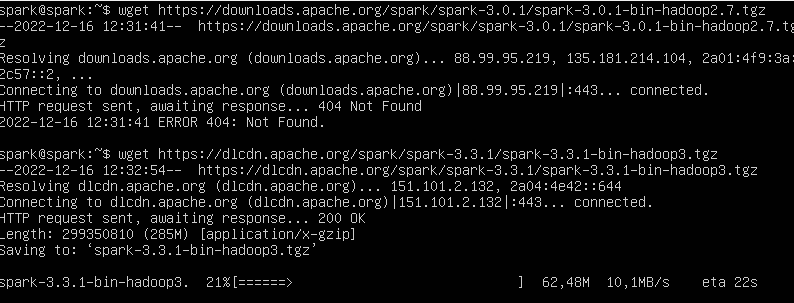


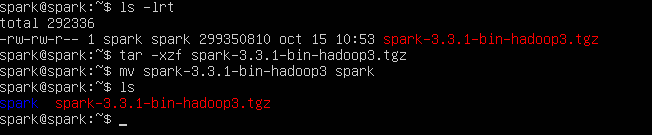
Instalar Python:

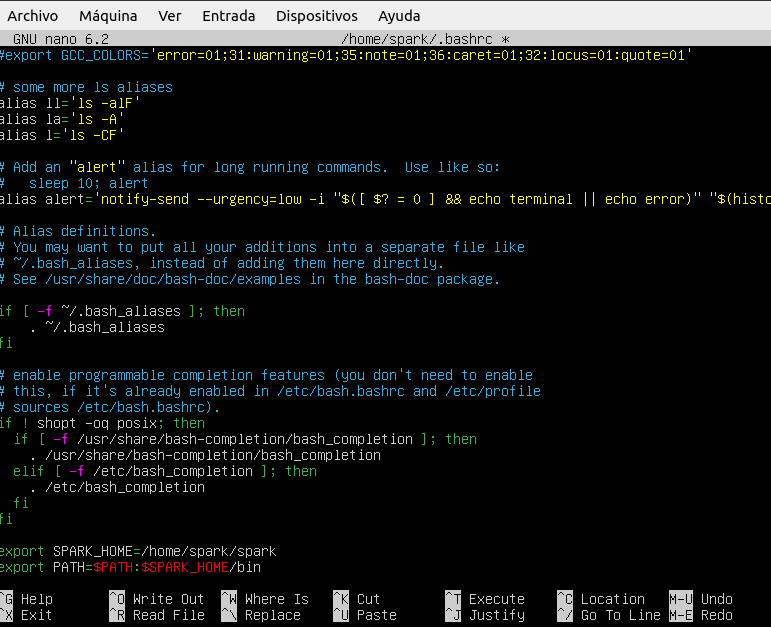
sudo apt install python3



Descargar Apache Spark:

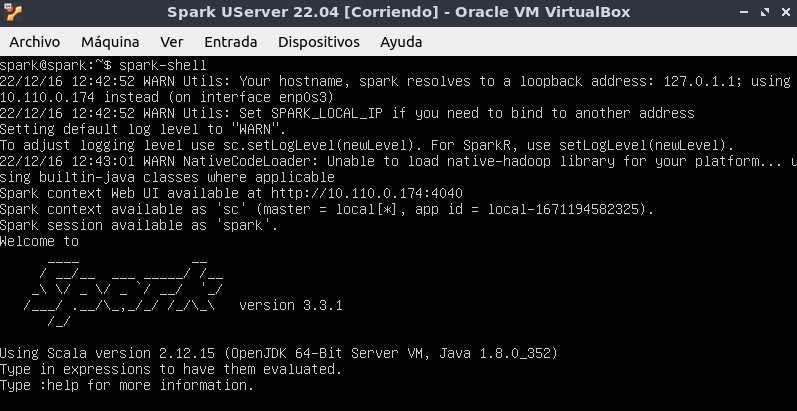


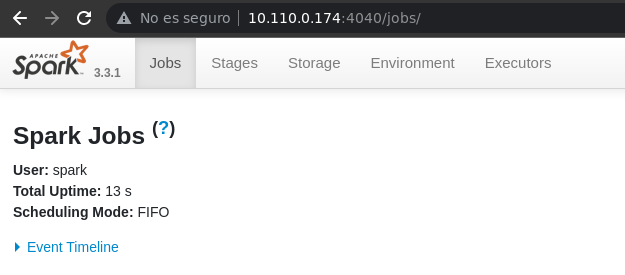








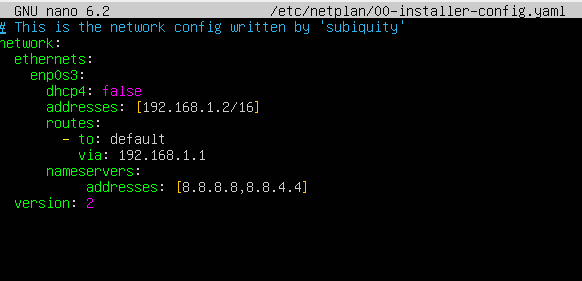


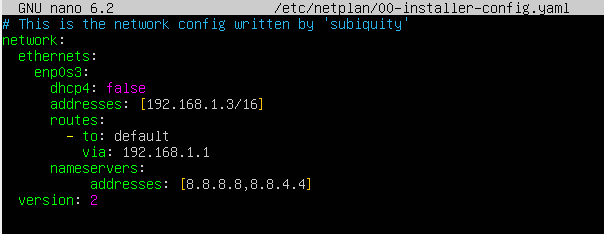


**Clonamos esta máquina y le asignamos a cada máquina una ip fija**

**sudo nano /etc/netplan/00-installer-config.yaml**

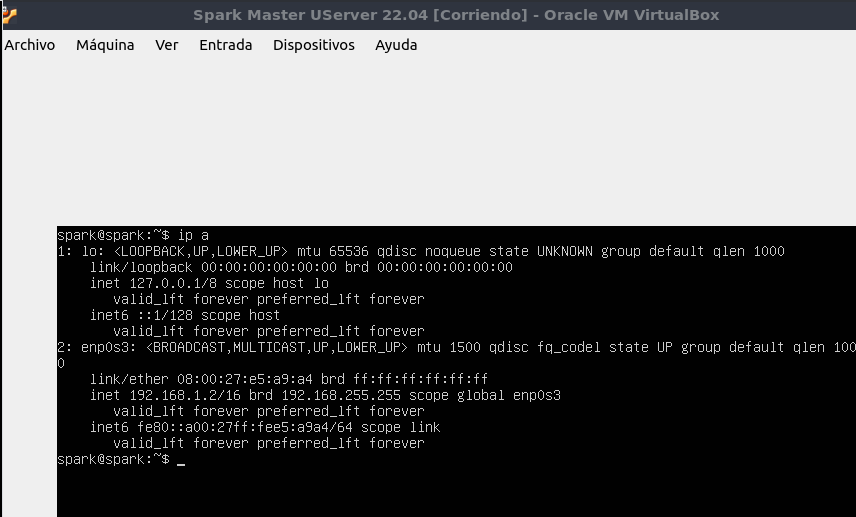
**Spark-Master:**

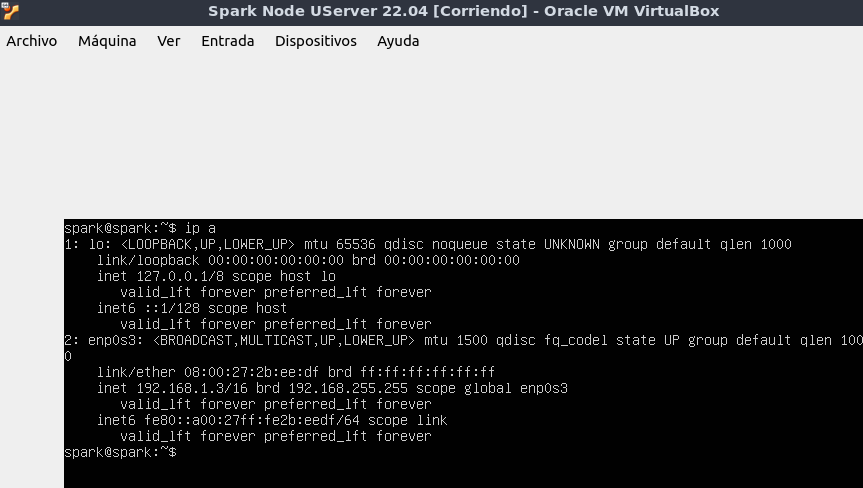
****

****

**sudo netplan apply**

**Verificamos la configuración de red**

****

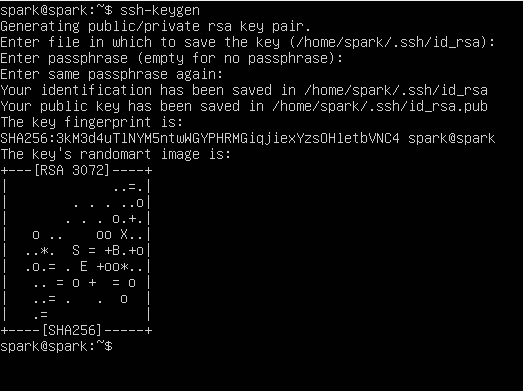
****

**Instalamos SSH en el nodo**

**sudo apt-get install openssh-server**

**Creamos una clave RSA para que la conexión sea segura**

**ssh-keygen**

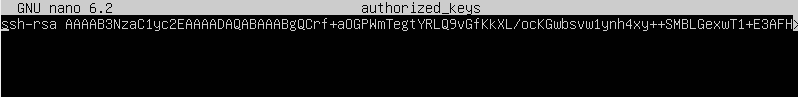
****

**Copiamos la clave en cada uno de los nodos de trabajo**

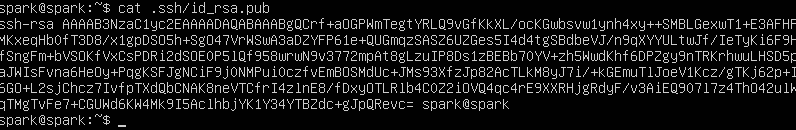
**ssh-copy-id -i ~/.ssh/id\_rsa.pub mario@192.168.107.129**

**Esta clave se guarda en el fichero *authorized\_keys:***

**Nodo de trabajo:**

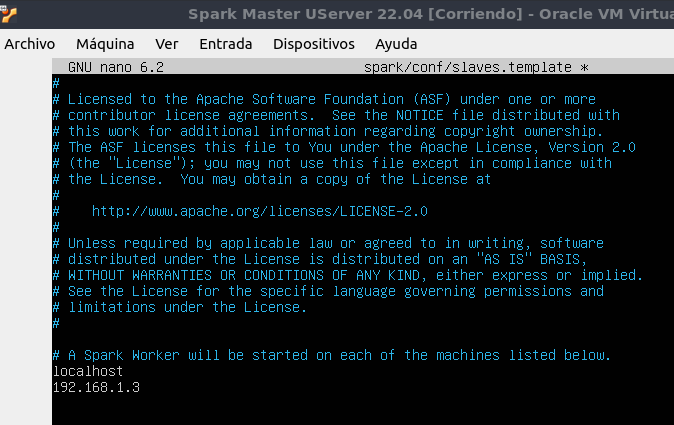
******

**Spark Master:**

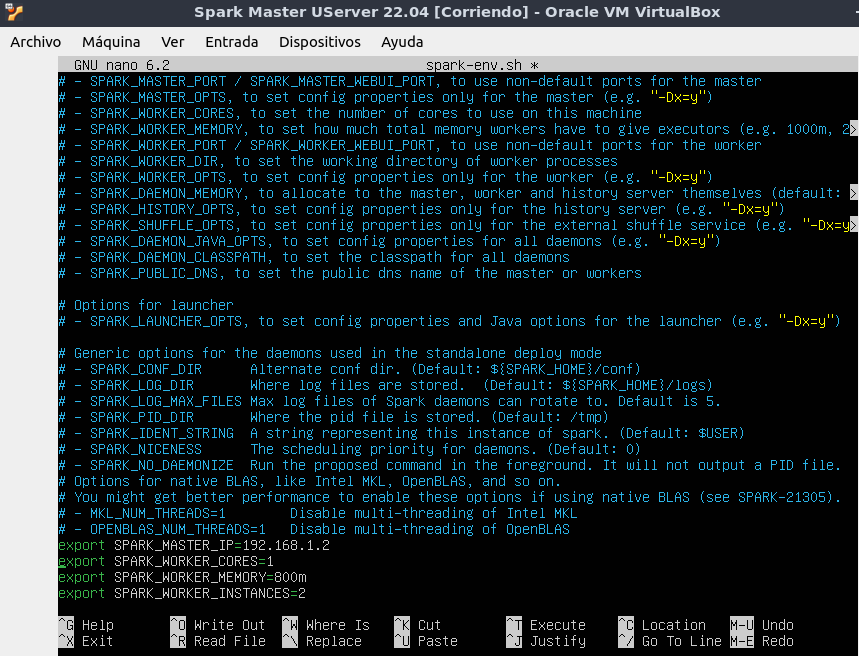
****

**A continuación configuraremos Apache Spark (debe estar instalado en todos los equipos):**

**Creamos un archivo llamado slaves.template para añadir las direcciones IP de los nodos esclavos**

****

**Modificamos el fichero spark-env.sh.template:**

****