

INSTALANDO SPARK y JUPYTER en UBUNTU EN WINDOWS 10

VIDEO CON TODO EL PROCESO DE INSTALACIÓN:
<https://www.youtube.com/watch?hd=1&v=gcDgqVY0Zz0>

EJECUTANDO UBUNTU en WINDOWS 10

RECOMENDADO: ¡DESINSTALAR UBUNTU Y VOLVER A EMPEZAR SI FALLA LA INSTALACIÓN!

EJECUTAR COMO ADMIN EN POWERSHELL:

```
Enable-WindowsOptionalFeature -Online -FeatureName Microsoft-Windows-Subsystem-Linux
```

2) Obtner e instalar Ubuntu, para eso, en la LUPA buscar Microsoft Store y a continuación buscar: Ubuntu 18.04 LTS, finalmente dar a obtner e inciar.

Se puede usar este enlace: <https://www.microsoft.com/en-us/p/ubuntu-1804-lts/9n9tngvndl3q?activetab=pivot:overviewtab>

3) Para actualizar todas las aplicaciones, dentro de Ubuntu ejecutar el comando:

```
sudo apt update && sudo apt upgrade
```

INSTALANDO SPARK en UBUNTU (En Windows o no)

Observación, si algo abajo falla ejecutar: `sudo apt update && sudo apt upgrade`

Instalar Java, Scala, git y pip3

```
sudo apt install default-jdk scala git python3-pip -y
```

Instalando Spark con Hadoop

Descargar Spark (viene preseleccionado): <https://spark.apache.org/downloads.html>

```
wget https://ftp.cixug.es/apache/spark/spark-3.0.1/spark-3.0.1-bin-hadoop2.7.tgz
```

Abrir el archivo descargado (con el nombre que tenga):

```
tar xvf spark-3.0.1-bin-hadoop2.7.tgz
```

Unzip and move the unzipped directory to a working directory:

```
sudo mv spark-3.0.1-bin-hadoop2.7 /opt/spark-3.0.1
```

Symlink the version of Spark to a spark directory:

```
sudo ln -s /opt/spark-3.0.1/ /opt/spark
```

Edit ~/.bashrc, para definir las variables: SPARK_HOME, PATH

```
nano ~/.bashrc
```

Incluir las siguientes líneas al final de todo:

```
export SPARK_HOME=/opt/spark
export PATH=$SPARK_HOME/bin:$PATH
```

Edit ~/.profile, para definir las variables: SPARK_HOME, PATH y PYSPARK_PYTHON

```
nano ~/.profile
```

Incluir las siguientes líneas al final de todo:

```
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
export PYSPARK_PYTHON=/usr/bin/python3
```

Cerrar y volver a abrir Ubuntu o ejecutar!

```
source ~/.profile
source ~/.bashrc
```

Ya lo tenemos! Ejecutando PySpark:

```
pyspark
```

INSTALANDO JUPYTER NOTEBOOK y CONECTANDO A SPARK

Instalar pip3 (pip para python3):

```
sudo apt install python3-pip
```

Instalar Jupyter Notebook:

```
sudo pip3 install jupyter notebook
```

Configurando Jupyter:

```
jupyter notebook --generate-config
```

Instalando findspark y pyspark (para usar desde Jupyter):

```
pip3 install findspark pyspark
```

Editar ~/.bashrc, para definir las variables: SPARK_HOME como pyspark

```
nano ~/.bashrc
```

cambiar la dirección de SPARK_HOME a el paquete pyspark de Python de tu usuario:

```
export SPARK_HOME=/home/<USUARIO>/local/lib/python3.6/site-packages/pyspark
```

Cerrar y volver a abrir Ubuntu o ejecutar!

```
source ~/.bashrc
```

Lanzando Jupyter Notebook:

```
jupyter notebook
```

Ya lo tenemos! En Jupyter crear un nuevo fichero python3 y ejecutar:

```
SC
sc.parallelize([1,2,3,4]).count()
```