

Instalación de Spark

Estudiante : Jason Solano

Breve Explicación

En este tutorial indicaremos la instalación en el sistema operativo Mac OS.

Para instalar Spark en Python son necesarias las siguientes herramientas:

- Anaconda o Conda
- Java

Verificación si las herramientas están instaladas en nuestro equipo

- Anaconda o Conda: esta herramienta es muy sencilla de verificar con el comando en **conda env list**, se pueden observar los environments de anaconda, de no ver una lista, se debe de instalar anaconda
 - Instalación de anaconda:
 - Seguir los pasos: <https://docs.anaconda.com/anaconda/install/mac-os/>
- Java: en la verificación de la herramienta Java, se puede ejecutar los comandos:
 - Verificar si Java está instalado:
 - Comando: **java -version**
Debe mostrar una información similar a la siguiente:
java version "1.8.0_171"
Java(TM) SE Runtime Environment (build 1.8.0_171-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.171-b11, mixed mode)
 - Verificar si el JDK de Java está instalado
 - Comando: **javac**
Debe mostrar instrucciones disponibles para el comando **javac**, en caso de mostrar un mensaje de error, significa que no está bien instalado.
 - Para instalar el JDK y Java:
 - <https://docs.oracle.com/javase/10/install/installation-jdk-and-jre-macos.htm#JSJIG-GUID-F575EB4A-70D3-4AB4-A20E-DBE95171AB5F>

Pasos para instalar Spark con Python

1. Descargamos el archivo Spark
 - a) Abrimos la siguiente dirección: <https://spark.apache.org/downloads.html>, y abrimos el link señalado por la línea

Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-3.0.0-preview-bin-hadoop2.7.tgz](#) ←
4. Verify this release using the 3.0.0-preview [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12.

- b) Seleccionamos el mirror de preferencia para la descarga, en nuestro caso seleccionamos el de http UCR.



We suggest the following mirror site for your download:

<http://mirrors.ucr.ac.cr/apache/spark/spark-3.0.0-preview/spark-3.0.0-preview-bin-hadoop2.7.tgz>

Other mirror sites are suggested below.

It is essential that you [verify the integrity](#) of the downloaded file using the PGP signature (`.asc` file) or a hash (`.md5` or `.sha*` file).

Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA* etc) -- or if no other mirrors are working.

HTTP

<http://mirrors.ucr.ac.cr/apache/spark/spark-3.0.0-preview/spark-3.0.0-preview-bin-hadoop2.7.tgz> ←

FTP

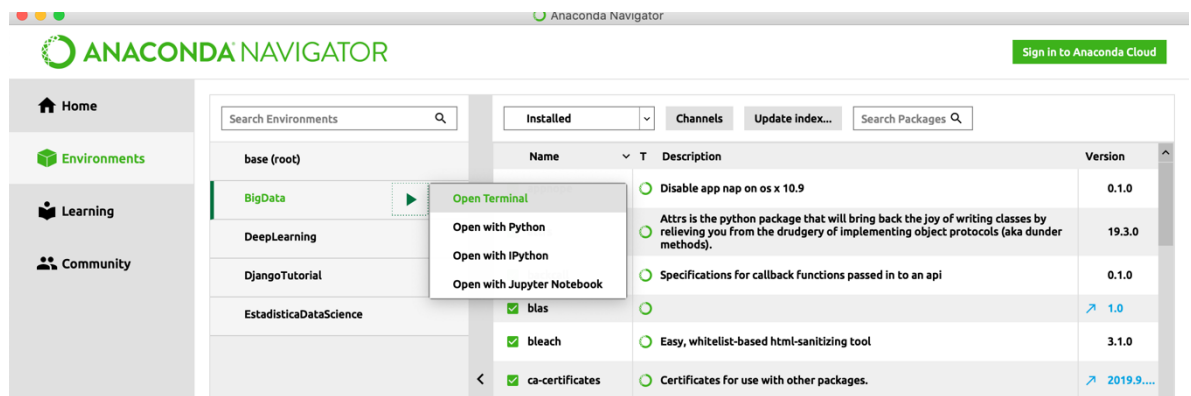
<ftp://mirrors.ucr.ac.cr/apache/spark/spark-3.0.0-preview/spark-3.0.0-preview-bin-hadoop2.7.tgz>

- 2) Movemos el archivo `spark-3.0.0-preview-bin-hadoop2.7.tgz` a la carpeta `opt`.
 - a) Creamos la carpeta `spark`

```
mkdir -p ~/opt/spark
```
 - b) Descomprimos el archivo
 - c) Movemos el archivo `spark-3.0.0-preview-bin-hadoop2.7`

```
mv ~/Downloads/spark-3.0.0-preview-bin-hadoop2.7 ~/opt/spark/spark-3.0.0-preview-bin-hadoop2.7
```
- 3) Instalar `find-spark`
 - a) Abrimos la aplicación `anaconda-navigator`
 - i) Seleccionamos el environment que configuramos para utilizar Spark

ii) Una vez dentro del environment abrimos la terminal



iii) En la terminal escribimos:

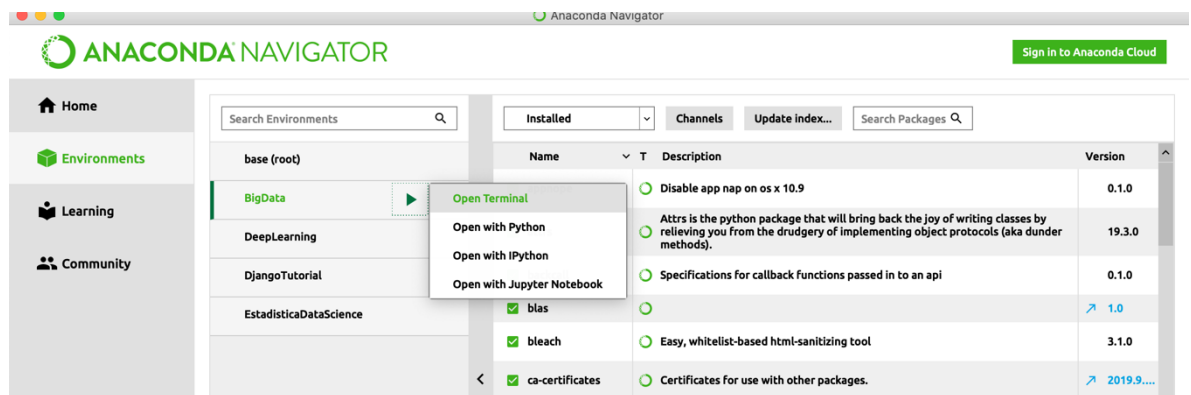
```
(1) conda config --add channels conda-forge  
(2) conda install findspark
```

4) Instalar pyspark

a) Abrimos la aplicación anaconda-navigator

i) Seleccionamos el environment que configuramos para utilizar Spark

ii) Una vez dentro del environment abrimos la terminal



iii) En la terminal escribimos:

```
(1) conda install findspark
```

5) Probamos la instalación

a) Abrimos jupyter notebook

b) Probamos los siguiente scripts

Comenzamos con la importación de librerías y demás parámetros necesarios...

```
In [4]: from pyspark.sql import SparkSession
        spark=SparkSession.builder.appName('data_processing').getOrCreate()

        import pyspark.sql.functions as F
        from pyspark.sql.types import *
```

```
In [5]: import findspark
        findspark.init('/opt/spark')

        from datetime import datetime
        from pyspark.sql import SparkSession
        from pyspark.sql.functions import col, date_format, udf
        from pyspark.sql.types import DateType
```