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

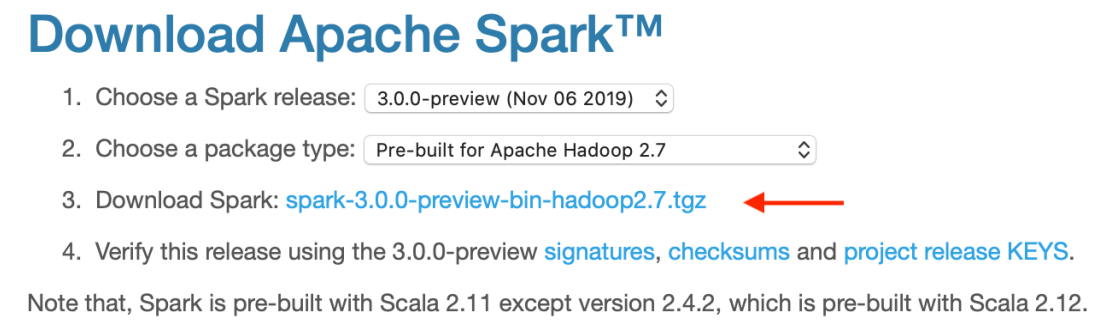
- Anaconda o Conda: esta herramienta es muy sencilla de verificar con el comando en **conda env list**, se pueden observar los environments de anaconda, de no ver una lista, se debe de instalar anaconda
    - Instalación de anaconda:
        - Seguir los pasos: https://docs.anaconda.com/anaconda/install/mac-os/

- Java: en la verificación de la herramienta Java, se puede ejecutar los comandos:
    - Verificar si Java está instalado:
        - Comando: j**ava -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



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

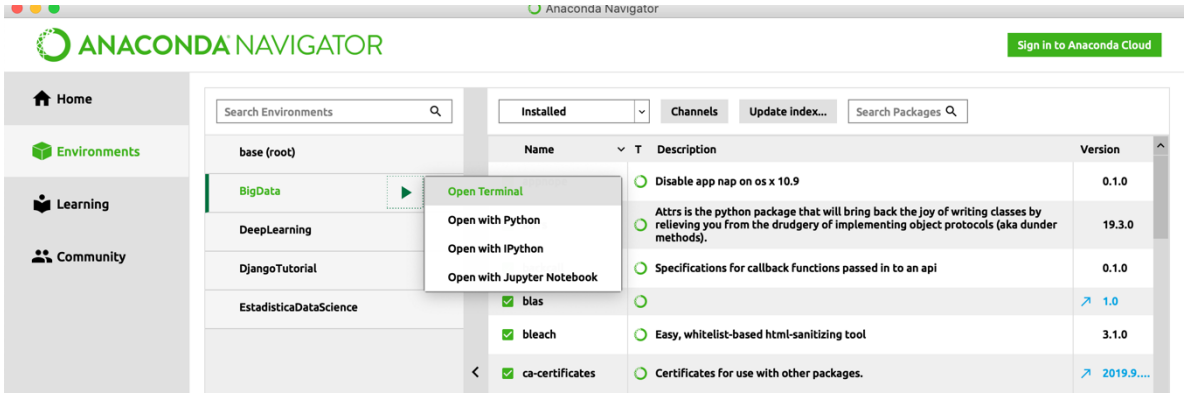
2) Movemos el archivo spark-3.0.0-preview-bin-hadoop2.7.tgz a la carpeta opt.
   a) Creamos la carpeta spark
      `mkdir -p ~/opt/spark`

   b) Descomprimimos 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 enviroment que configuramos para utilizar Spark

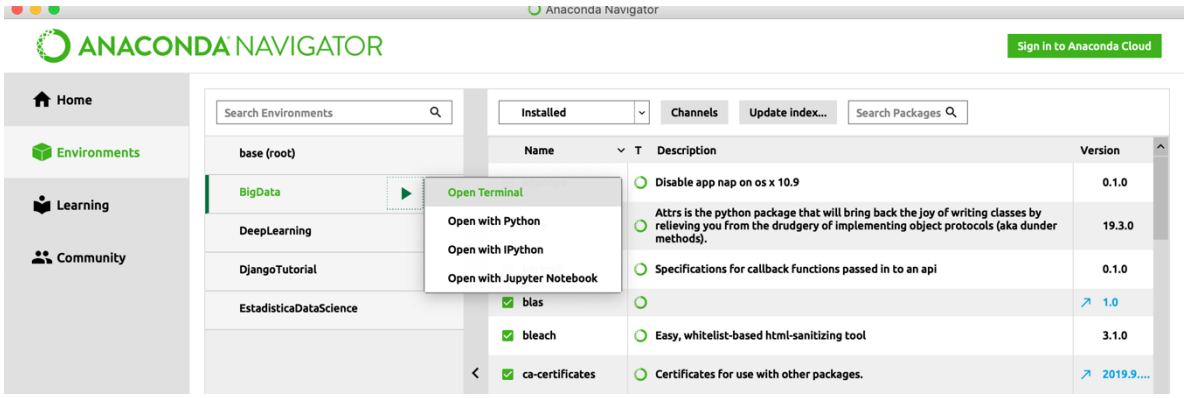
ii) Una vez dentro del enviroment abrimos la terminal



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 enviroment que configuramos para utilizar Spark
      ii) Una vez dentro del enviroment abrimos la terminal



iii) En la terminal escribimos:

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

5) Probamos la instalación
   a) Abrimos jupyternotebook

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
```