Formas de instalación de PySpark

Windows

> La instalación considera que se realiza con el ambiente Anaconda

- **Instalar Java**

Por lo general Java se utiliza por múltiples programas, lo más probable es que ya exista una versión operativa en su computador.

Para verificar si se encuentra instalado, vayan al Símbolo del Sistema (Command Prompt). Dentro de este, escriban `java -version`.

Si se encuentra instalado, les debería arrojar algo similar a:

```
java version "1.8.0_202"
Java(TM) SE Runtime Environment (build 1.8.0_202-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.202-b08, mixed
mode)
```

Si no obtienen un resultado similar que indique la versión de Java, lo más probable es que necesiten instalarlo. Para ello, pueden descargar la distribución `OpenJDK 8` de la siguiente página: https://adoptopenjdk.net/upstream.html.

- **Instalar Anaconda**: Si no tienen instalado el ambiente Anaconda, pueden instalarlo en https://www.anaconda.com/distribution/, seleccionando su distribución de OS y verificando que sea la versión Python 3.6. Utilizando el gestor de paquetes de Anaconda (`conda`), instalen `findspark` .

- **Instalar Apache Spark**: Diríganse a la página de descarga de Apache Spark http://spark.apache.org/downloads.html. Dentro de esta, seleccionen la versión `2.4.4` de Spark. Seleccionen la opción de tipo de paquete `Pre-built for Apache Hadoop 2.7 and Later` . Descarga el archivo `.tgz` .

Para instalar Spark, no es necesario ejecutar instalador alguno. Simplemente al extraer los archivos contenidos del `.tgz` en una dirección, estaremos listos. Se sugiere utilizar una dirección **sin espacios**.

Si creamos una carpeta llamada `Spark` en nuestro Escritorio, podremos extraer los archivos del `tgz` . Así, tendremos un `SPARK_HOME` con una estructura similar a `C:\Users\<nombre_usuario>\Escritorio\Spark\spark-2.4.4-bin-hadoop2.7` . Esta dirección la vincularemos después, por lo que es necesario saber cómo se llama.

- **Instalar winutils.exe**: Necesitamos instalar `winutils.exe` para finalizar la configuración de Spark. Descargen el contenido desde https://github.com/steveloughran/winutils.

Desde ahí, seleccionen la versión correspondiente a su distribución descargada. Esta se puede encontrar en los últimos dígitos del `tgz` que descargaron de la página de Spark. Dentro de esa carpeta, ubiquen el archivo `winutils.exe` que se encuentra en `bin` y cópienlo en la carpeta `hadoop\bin` dentro de su `SPARK_HOME`

- **Corroborar instalación**: Dentro de un Jupyter Notebook, ejecuten las siguientes líneas:

```
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
df = spark.sql("select 'spark' as hello ")
df.show()
```

MacOSX

> La instalación considera que se realiza con el ambiente Anaconda

- **Instalar Xcode**: Para poder instalar Java, Scala y Apache Spark, es necesario instalar Xcode que entrega una serie de herramientas orientadas al desarrollo. Para instalarlo, diríganse al terminal y ejecuten:

```
xcode-select –install
```

- **Instalar Homebrew**: Homebrew es un gestor y administrador de instalaciones para MacOSX. Para instalarlo, diríganse al terminal y ejecuten:

```
/usr/bin/ruby -e “$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/master/insta
ll)”
```

El terminal debería entregar información similar a la siguiente al finalizar la instalación de Homebrew

```
==> Cleaning up /Library/Caches/Homebrew...
==> Migrating /Library/Caches/Homebrew to
/Users/apple/Library/Caches/Homebrew..
==> Deleting /Library/Caches/Homebrew...
Already up-to-date.
==> Installation successful!

==> Homebrew has enabled anonymous aggregate user behaviour
analytics.
Read the analytics documentation (and how to opt-out) here:
  http://docs.brew.sh/Analytics.html

==> Next steps:
- Run `brew help` to get started
- Further documentation:
    http://docs.brew.sh
```

- **Instalar Java**: Para instalar Java, descargen la versión 8 del Java Development Kit en la siguiente dirección https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
- **Instalar Apache Spark**: Para instalar Apache Spark, diríganse al terminal y ejecuten:

```
brew install apache-spark
```

- **Instalar PySpark y FindSpark**: Instalen `findspark` y `pyspark` desde Anaconda.

No pude instalarlo, ¿Qué hago?

No importa, en el siguiente link encontrará un Jupyter Notebook con las instrucciones de instalacción para Google Colab. Simplemente ejecuta las líneas y tendrás tu entorno listo para trabajar:

https://drive.google.com/open?id=1pfg1y1o32LRFFKBmsgcWYcYd8eC69IFM