Tarea #1
BigData
Juan Pablo Fernández Delgado

Guia de instalacion de pySpark en JupyterNotebook

Apache Spark y PySpark
Apache Spark es un motor de código abierto desarrollado para gestionar y procesar datos en un entorno Big Data.

Requisitos de instalación

1) Instalación de Java 8

a. Bajar java SDK 8 del sitio oficial: https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

Java SE Development Kit 8u231

You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.
Thank you for accepting the Oracle Technology Network License Agreement for Oracle Java SE; you may now download this software.

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.9 MB	jdk-8u231-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.8 MB	jdk-8u231-linux-arm64-vfp-hflt.tar.gz
Linux x86	170.93 MB	jdk-8u231-linux-i586.rpm
Linux x86	185.75 MB	jdk-8u231-linux-i586.tar.gz
Linux x64	170.32 MB	jdk-8u231-linux-x64.rpm
Linux x64	185.16 MB	jdk-8u231-linux-x64.tar.gz
Mac OS X x64	253.4 MB	jdk-8u231-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	132.98 MB	jdk-8u231-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	94.16 MB	jdk-8u231-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	133.73 MB	jdk-8u231-solaris-x64.tar.Z
Solaris x64	91.96 MB	jdk-8u231-solaris-x64.tar.gz
Windows x86	200.22 MB	jdk-8u231-windows-i586.exe
Windows x64	210.18 MB	jdk-8u231-windows-x64.exe

b. Agregar las variables de entorno

JAVA_HOME = C:\Progra~1\Java\jdk1.8.0_161
PATH += C:\Progra~1\Java\jdk1.8.0_161\bin
Nota: para ver la versión del java: java -version

2) Bajar e Instalar Spark del sitio oficial :http://spark.apache.org/downloads.html

a. Barjar el mas reciente.
b. Extraer el .tgz en C:\Spark
c. Setear las variables de entorno

SPARK_HOME = C:\Spark
PATH += C:\Spark\bin

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.0
      /_/

Using Python version 3.6.6 (default, Jun 28 2018 11:07:29)
SparkSession available as 'spark'.
```

3) Instalación pySpark

```
conda install -c conda-forge pyspark
```

4) Instalacion findSpark

```
conda install -c conda-forge findspark
```

En algunas ocasiones se tiene que agregar el canal, esto debido a que puede haber un proxy de por medio:conda config --add channels conda-forge
En el siguiente enlace se muestra más detalles de la instalación.

https://github.com/conda-forge/findspark-feedstock

5) Probar Spark desde la consola de comando.

Para verificar que la instalación haya sido exitosa probalos las siguientes líneas desde la consola de comando.

```
cmd> pyspark
                >>> nums =
                sc.parallelize([1,2,3,4])
                >>> nums.map(lambda x:
                x*x).collect()
```

Para concluir vimos como cubrir todos los pasos para la utilización de las funcionalidades de pyspark en anaconda y así utilizarlo desde el jupyter notebook.