CENFOTEC
Curso Big Data – Especialidad Ciencia de Datos
Estudiante: Sergio Castillo Segura

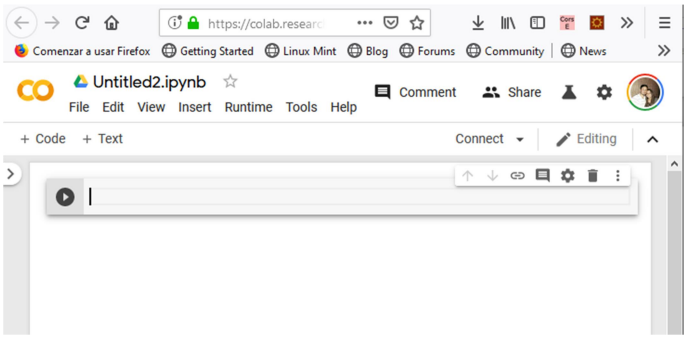
Tarea 1 – 25 de Noviembre, 2019

Guía sobre preparación de ambiente de trabajo para trabajo con Big Data en Spark

- En esta guía explicaremos 2 métodos para trabajar con Pyspark utilizando notebooks de Jupyter.
  El primer método es para iniciación ultra-rápida en ambiente Google Colab, la segunda es un poco más elaborada para instalación local utilizando Condas.

Método 1 – instalando en Google Colab

1. Abrir un nuevo notebook de Python 3 en Google Colab utilizando su propia cuenta de google, el URL es: https://colab.research.google.com



2. Copiar las siguientes líneas de código y pegarlas en el bloque de código que aparece de primero por defecto:

```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://www-us.apache.org/dist/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz
!tar xf spark-2.4.4-bin-hadoop2.7.tgz
!pip install -q findspark
```

3. Al ejecutar el código se instalará java 8 runtime Apache Spark versión 2.4.4 y la librería findspark para Python 3, nótese que el servidor de descarga utilizado para Spark es el de Estados Unidos, esto es para que la descarga sea lo más rápida posible al estar Colab también en USA.
4. Agregar un nuevo bloque de código y agregar las siguientes líneas:

```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-2.4.4-bin-hadoop2.7"
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
```

5. Las líneas anteriores importan el ambiente Java y Spark para crear una nueva sesión en el servidor utilizando los recursos locales.
6. Finalmente agregamos un nuevo bloque de código y ejecutamos las siguientes líneas:

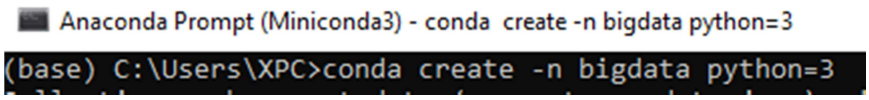
```
from datetime import datetime
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

7. Ahora el ambiente de trabajo está listo para comenzar a trabajar con Spark, como ejemplo se puede utilizar el siguiente código para importar un archivo de datos en un dataframe de Spark:

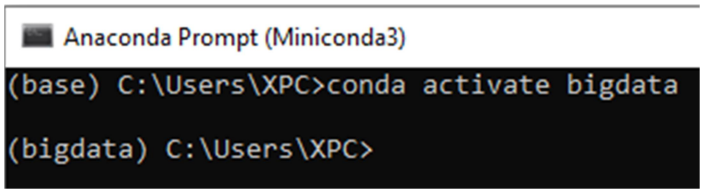
```
from pyspark import SparkFiles
spark.sparkContext.addFile("https://www.myserver.com/mydir/data.csv")
df = spark.read.csv(SparkFiles.get("customer_data.csv"), header=True, inferSchema=True)
df.show()
```

Método 2 – Instalación Local

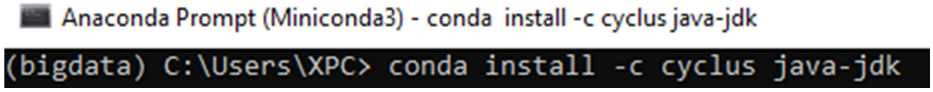
1. Lo primero es instalar Miniconda, el cual se puede obtener de la URL (miniconda python 3.7): https://docs.conda.io/en/latest/miniconda.html
2. Si corre sobre ambiente Windows entonces ejecute el "Anaconda Prompt (Miniconda3)" y ejecute el siguiente comando:



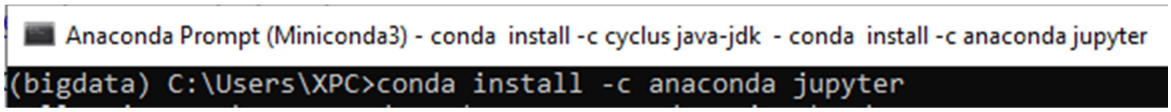
3. Confirme la instalación presionando la tecla "Y" y enter. Luego active el nuevo ambiente utilizando el comando:



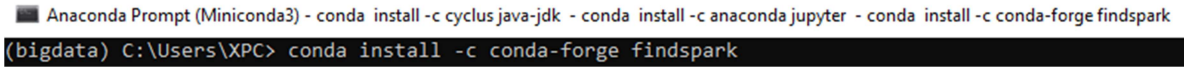
4. Spark corre sobre el ambiente de ejecución de Java, para lo cual ocupamos instalar el JDK de preferencia, para lo cual utilizaremos el siguiente comando (confirmar la instalación):



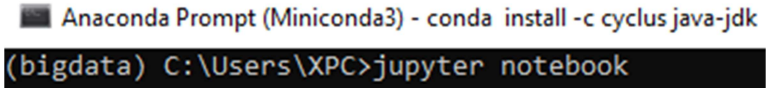
5. También deberemos instalar Jupyter para trabajar con notebooks, el comando:



6. Ahora debemos descargar Apache Spark desde el sitio web oficial: https://spark.apache.org/downloads.html
7. Al día de hoy la última versión disponible (no preview) es la 2.4.4, descargar el archivo .tgz y descomprimirlo en alguna ruta sencilla de acceder como c:\spark (dependiendo de su conexión y ubuicación el mirror de US puede resultar considerablemente más rápido que su mirror local).
8. Finalmente se debe instalar findspark, que es la librería de python utilizada para encontrar y ligar Spark al ambiente Python de forma rápida, para ello el comando:

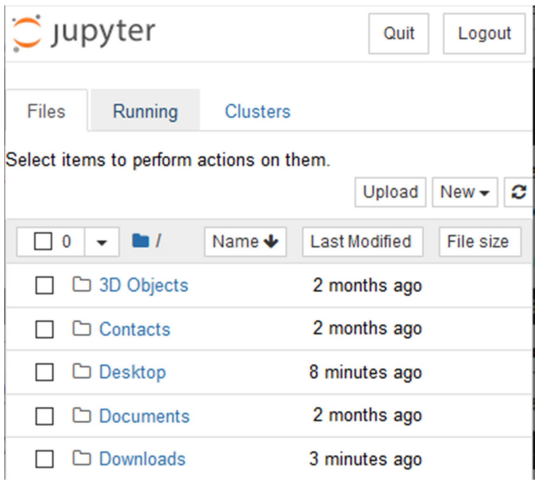


9. Ahora el ambiente está listo para abrir una nueva Notebook de Jupyter y comenzar a trabajar con Spark, para ello corremos el siguiente comando:

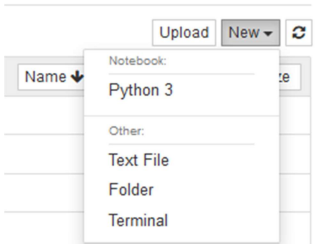


10. Puede que el sistema operativo le pregunte qué navegador desea utilizar para abrir el archivo .html que corresponde a la página de inicio de Jupyter si tiene varios navegadores instalados, Jupyter es compatible con todos los navegadores mayores por lo que utilice el de su preferencia. Una vez cargue tendrá una vista como

la siguiente:



11. Ahora crearemos un nuevo Notebook con Python3, para ello dar clic al botón “New” al lado derecho de la página y luego clic sobre la opción Python 3:



12. En el primer bloque de código puede pegar las siguientes líneas para comenzar a trabajar con Spark en el ambiente de trabajo recién instalado:

```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName('data_processing').getOrCreate()
import pyspark.sql.functions as F
from pyspark.sql.types import *
import findspark
findspark.init(“c:/spark/spark-2.4.4-bin-hadoop2.7”)
from datetime import datetime
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, date_format, udf
from pyspark.sql.types import DateType
```

- Fin -