Hola a tod@s!

Espero que ayer disfrutárais del taller y que os llevéis las lecciones aprendidas de costes, Databricks y pensar desde el punto del procesamiento “fila a fila”.

Sé que os quedásteis con muchas dudas pero, como todo programador o iniciado en este arte sabéis mejor que yo que ahora queda profundizar con una expedición stand-alone (que no *forever alone*). Os adjunto el material que tenía preparado para el taller donde hay más cosas incluso de las que nos dio tiempo pero que quizá os sean de ayuda en vuestro propio camino.

**Como recursos indispensables os dejo:**

* El libro: [Learning Spark (2nd Ed)](https://pages.databricks.com/rs/094-YMS-629/images/LearningSpark2.0.pdf), Jules S. Damji et al.
* [Documentación de Databricks](https://docs.databricks.com/)
* Documentación de [PySpark](https://spark.apache.org/docs/latest/api/python/)
* [Spark by examples](https://sparkbyexamples.com/pyspark/) (esta web es casi el stackoverflow de esta librería)
* Copia de los [notebooks](https://drive.google.com/drive/u/0/folders/1POOsjjt-29Q5Q1qdIDa-vddTKlNdPdeK) que preparé
* [Repositorios de Databricks](https://github.com/orgs/databricks-academy/repositories?type=all) con muchísimo material!

**Script para instalar PySpark en local**

Tened en cuenta que al no tener un “cluster” propiamente en local, Spark va a interpretar que tenéis tantos nodos como CPUs en vuestro portátil. La RAM es la que le hayáis comprado a Kingston, pero para jugar está más que bien y así no necesitáis Databricks para seguir jugando con PySpark.

| **#** Aseguraos de tener instalado el JRE de Java instalado **#** Por ejemplo, en ubuntu, para instalar el JRE apt-get update && apt-get install -y default-jre  **#** Luego es tan fácil como pip install pyspark |
| --- |

Luego, para tener el objeto **spark** en vuestro script, sólo hay que inicializarlo (que Databricks eso lo hacía por nosotros).

| *# Create SparkSession from builder* **from** pyspark.sql **import** SparkSession spark = (SparkSession  .builder  .master("local[1]") *# Esto le dice a Spark que la sesión está "enchufada" al driver "local" no a uno de Databricks*  .appName('SessionName')   .getOrCreate() ) *# Mis famosos imports para jugar* **from** pyspark.sql **import** functions **as** F **from** pyspark.sql **import** types **as** T **from** pyspark.sql **import** window **as** W  *# Y a programar!* *# Crear un DataFrame a manubrio* df = spark.createDataFrame(  [("Scala", 25000), ("Spark", 35000), ("PHP", 21000)]) *# Ojo, porque la maravillosa función "display" ya no estará disponible, pero tenemos el método show..* df.show() *# Truco de regalo* df\_pandas = df.toPandas() *# Qué hará esto?* |
| --- |

**Cositas extra!**

* Link al [Meetup](https://www.meetup.com/es-ES/pydatagrx/events/292425237/) para dejar tus good vibes y **comentarios**
* [Twitter](https://twitter.com/nenetto) de Eugenio para dar las gracias, pedir consejo o criticar sus artes oscuras
* Perfil de [contacto](https://www.linkedin.com/in/marinetto/) y [web](https://nenetto.com/) de Eugenio

Si tenéis cualquier duda o probando alguna cosilla os atrancáis me decís en un [correito](mailto:nenetto@gmail.com) **SIN problema, no shame!**. Yo no nací sabiendo nada de todo esto y aún me queda mucho que aprender seguro, pero no tengo por qué hacerlo solo. Mucho ánimo en vuestras carreras profesionales y abrazo muy fuerte!

Eugenio Marinetto

nenetto@gmail.com