Tarea 1 - Big Data

Instalación del ambiente de trabajo para BIG DATA - Mac

Contexto de Spark/PySpark

Apache Spark es una solución de código abierto desarrollado para analizar y procesar datos a gran escala, es un motor de código abierto desarrollado para gestionar y procesar datos en un entorno Big Data.

Spark permite acceder a datos procedentes de diferentes fuentes como puede ser el sistema de archivos distribuidos de Hadoop (HDFS, “Hadoop Distributed File System”), OpenStack Swift, Amazon S3 o Cassandra.

Originalmente Apache Spark se diseño para hacer más fáciles los procesos analíticos en Hadoop. Al tiempo de proporcionar un conjunto completo de herramientas complementarias entre los que se incluyen la biblioteca de aprendizaje automático (MLlib) y el motor de procesamiento de gráficos (GraphX).

PySpark es un API que permite utiliza Apache Spark desde Python. Siendo necesaria la instalación de PySpark en Anaconda si deseamos trabajar con esta herramienta en nuestro sistema.

Instalar Spark/PySpark

Como **prerrequisitos** de esta instalación necesitamos primero haber instalado **Anaconda, Python 3**

La instalación de Spark requiere la versión específica de Java (java 8), que se puede instalar usando Homebrew.

1. Abrir la terminal, ingresar `$ brew install apache-spark`
2. Una vez que sale el siguiente error, ingresar `$ brew cask install caskroom/versions/java8` para instalar Java8

```
apache-spark: Java 1.8 is required to install this formula.
JavaRequirement unsatisfied!
You can install with Homebrew-Cask:
 brew cask install caskroom/versions/java8
You can download from:
 https://www.oracle.com/technetwork/java/javase/downloads/index.html
Error: An unsatisfied requirement failed this build.
```

3. Para verificar que pyspark se instaló apropiadamente al ingresar en la terminal `$ pyspark`, debería verse algo como esto, que significa que todo se instalo correctamente Spark:

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.2.1
      /_/

Using Python version 2.7.6 (default, Sep  9 2014 15:04:36)
```

4. Se instala findspark por medio del comando `$ pip install findspark`

5. Para probar el ambiente abrimos un notebook y agregamos el siguiente código como una pequeña prueba del ambiente instalado

In [22]:
```
import findspark
import random
from pyspark import SparkContext
sc = SparkContext(appName="EstimarPi")
def inside(p):
    x, y = random.random(), random.random()
    return x*x + y*y < 1
NUM_SAMPLES = 1000000
count = sc.parallelize(range(0, NUM_SAMPLES)) \
             .filter(inside).count()
print("Pi es aproximadamente %f" % (4.0 * count / NUM_SAMPLES))
sc.stop()
```

```
Pi es aproximadamente 3.137692
```

Listo!