Baruch Big Data Homework E

Title: "Apache Spark"
When issued: Fri, 3rd March 2017
When due: 12 noon, Fri, 10th March 2017
Contact details: andrew.sheppard@baruch.cuny.edu

Introduction

Wall Street is moving on from Hadoop as the Big Data tool of choice to Apache Spark. Why? Because Apache Spark can do everything that Hadoop can, and also do in-memory Big Data streaming in real-time.

The main project of this course will use Apache Spark, so you had better start getting to know it!

Install Apache Spark

For the main course project we will be using Apache Spark in the cloud. But initial development will be done locally on your laptop or desktop.

Search the Internet using the term "Apache Spark download" and find the instructions to download and install Apache Spark on your operating system. Test your installation to make sure it's working.

Assignment

Using Apache Spark on your local machine, write a Spark task (program) in Python that estimates the mathematical constant "pi" using the monte-carlo method.

Push to Github:

1. Your code.
2. An interactive terminal session log (text) showing you running your "pi" program and its output, which should be an approximation for "pi" good to at least decimal places.
3. Also push a full screen snapshot image of both your desktop and a terminal window open showing the interactive session and the result.

Hint, you may well find code on the Internet that does most of the work when writing your program.