An Insufficient Introduction to Spark

Part 0: Introduction to this Training

Riccardo Murri <riccardo.murri@gmail.com>

2019-02-11

Welcome!

Prerequisites

This course assumes some experience with Python programming.

I try to recall what is needed for exercises, but you will need to be able to write a function on your own and not to be puzzled about lists and tuples, or methods and function calls.

Course outline

1. The Map/Reduce programming model
2. RDDs
3. DataFrames and SQL
4. Window Functions

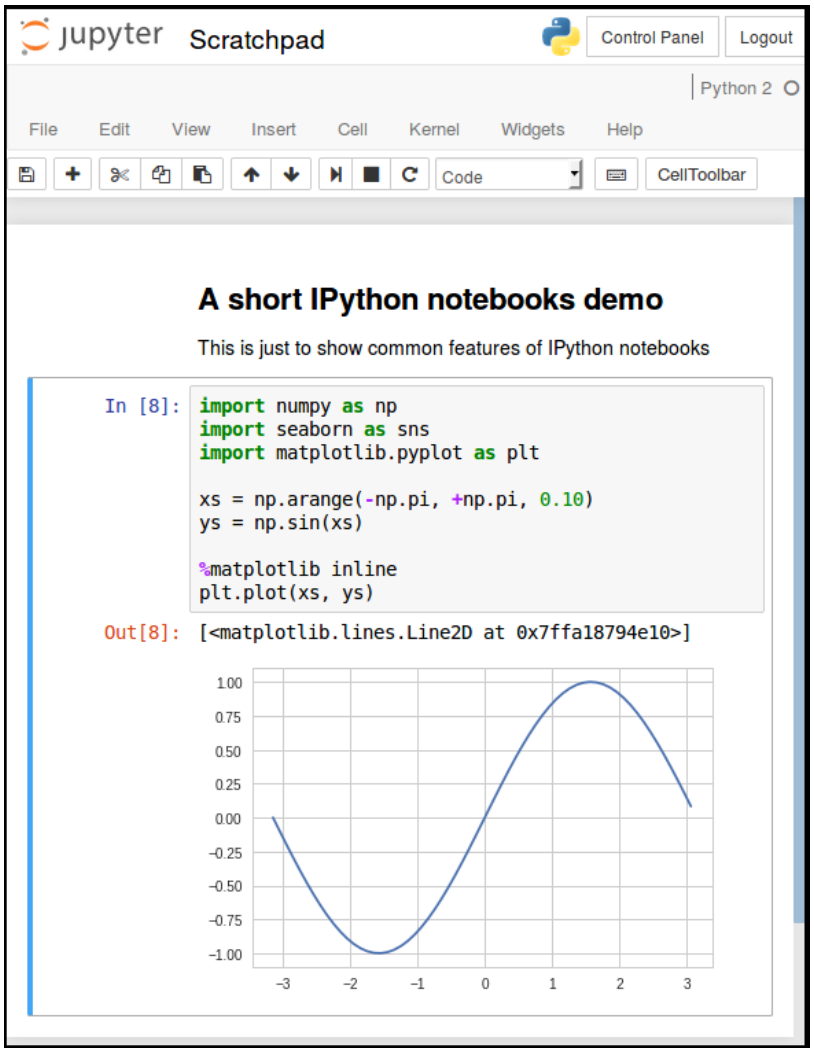
Next steps

The course will be structured as a mixture of slides and hands-on sessions for practicing PySpark programming.

So, the very first step is making sure you can access the Jupyter/IPython server for running the exercise notebooks.

How to run Python code

The IPython notebook, I



An appealing way of interacting with Python is through *IPython notebooks*.

Notebooks are made of "cells", which come in two flavors:

- documentation cells, containing text formatted according to the Markdown conventions;
- code cells, containing arbitrary Python code

The IPython notebook, II

To run Python code in the notebook:

- Type your code in a cell besides the `In [ ]:` (multiple lines are allowed)
- Press **Ctrl+Enter** to evaluate the cell (prompt changes to `In [*]:`) — or press **Alt+Enter** to evaluate the code *and* open a new code cell.
- When the Python kernel has done computing, the result appears *under* the code cell marked with a `Out [ ]:` label.