ID2221: Data Intensive Computing
Lab 2 - Spark and Spark SQL

(Updated 2017-09-23)

In this lab you will first practice the basic operations of Spark (RDDs) and Spark SQL (DataFrames). Next you will use what you learned to do some interactive spark analytics.

1 - Setting up the Work Environment

Docker

We will be using docker in this lab. We assume you already have it installed. If not, refer to the instructions in Lab 1.

Spark Image

We will be using a Spark Docker image that contains all the tools we need. Including, Spark, Jupyter, Python, Scala, and much more. You can see the full description here.

Start the Spark Container

Pull the image:

```
docker pull jupyter/all-spark-notebook
```

Creating a folder that will contain your files. We will mount this folder inside the container:

```
mkdir mywork
```

Now we run the Spark container using the image we just pulled. **Replace** the **red text** with the folder you just created:

```
docker run -it --rm  -p 8888:8888 --name mySpark -v
/home/user/work:/home/jovyan/work jupyter/all-spark-notebook
```

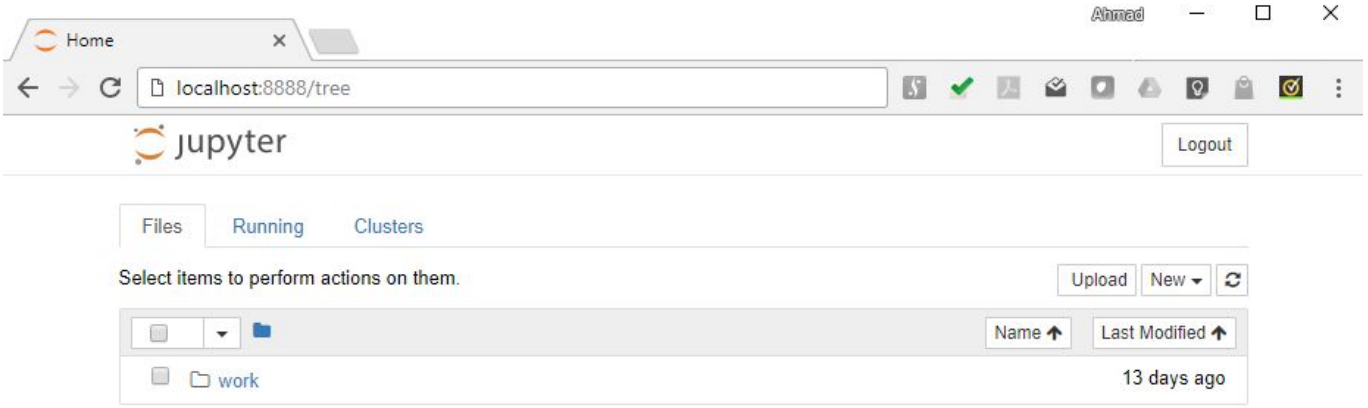
This should start Spark and Jupyter.

Take note of the **authentication token** included in the Jupyter startup log messages. Use this url in your browser to access Jupyter.

```
[I 15:31:11.946 NotebookApp] Writing notebook server cookie secret to /home/jovyan/.local/share/jupyter/runtime/notebook
[W 15:31:11.975 NotebookApp] WARNING: The notebook server is listening on all IP addresses and not using encryption. Thi
ded.
[I 15:31:12.023 NotebookApp] JupyterLab alpha preview extension loaded from /opt/conda/lib/python3.6/site-packages/jupyt
JupyterLab v0.27.0
Known labextensions:
[I 15:31:12.026 NotebookApp] Running the core application with no additional extensions or settings
[I 15:31:12.032 NotebookApp] Serving notebooks from local directory: /home/jovyan
[I 15:31:12.032 NotebookApp] 0 active kernels
[I 15:31:12.032 NotebookApp] The Jupyter Notebook is running at: http://[all ip addresses on your system]:8888/?token=0b
fcf2f2a23ff53c53928cb9c99b861
[I 15:31:12.032 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 15:31:12.032 NotebookApp]

    Copy/paste this URL into your browser when you connect for the first time,
    to login with a token:
        http://localhost:8888/?token=0b9e6a4016cc8cb787dfcf2f2a23ff53c53928cb9c99b861
[I                                                                                  99b861 (81.226.160.135) 0.84ms
```

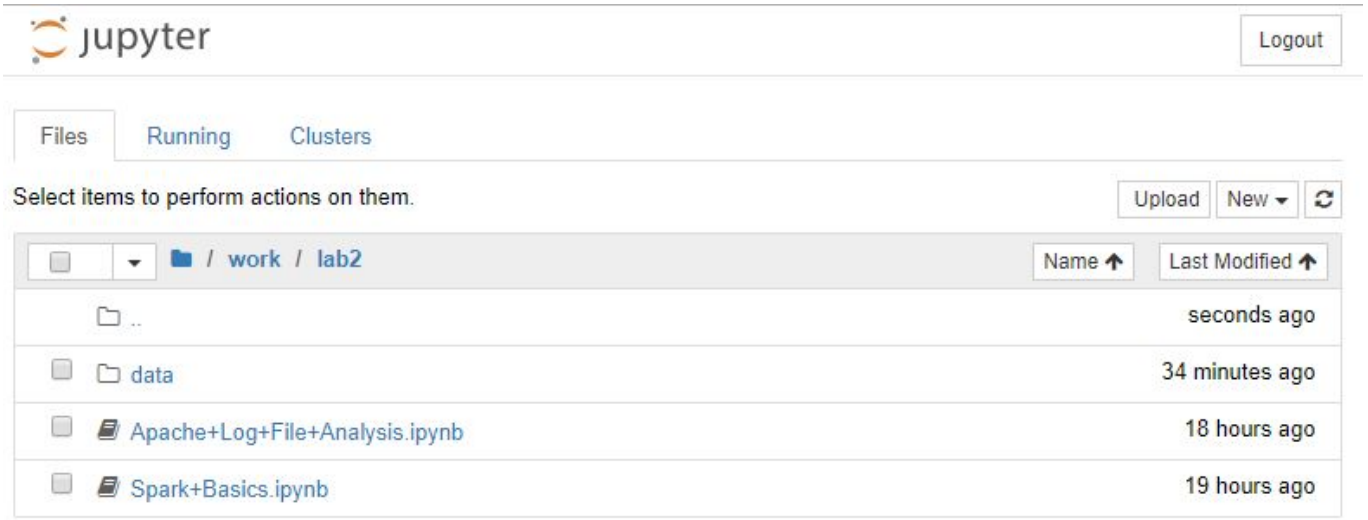
Copy and paste this url in your browser. You should see a page similar to the one below.



2 - Lab 2 Guide

Unzip lab2.zip and copy its contents into the `mywork` folder you just created.

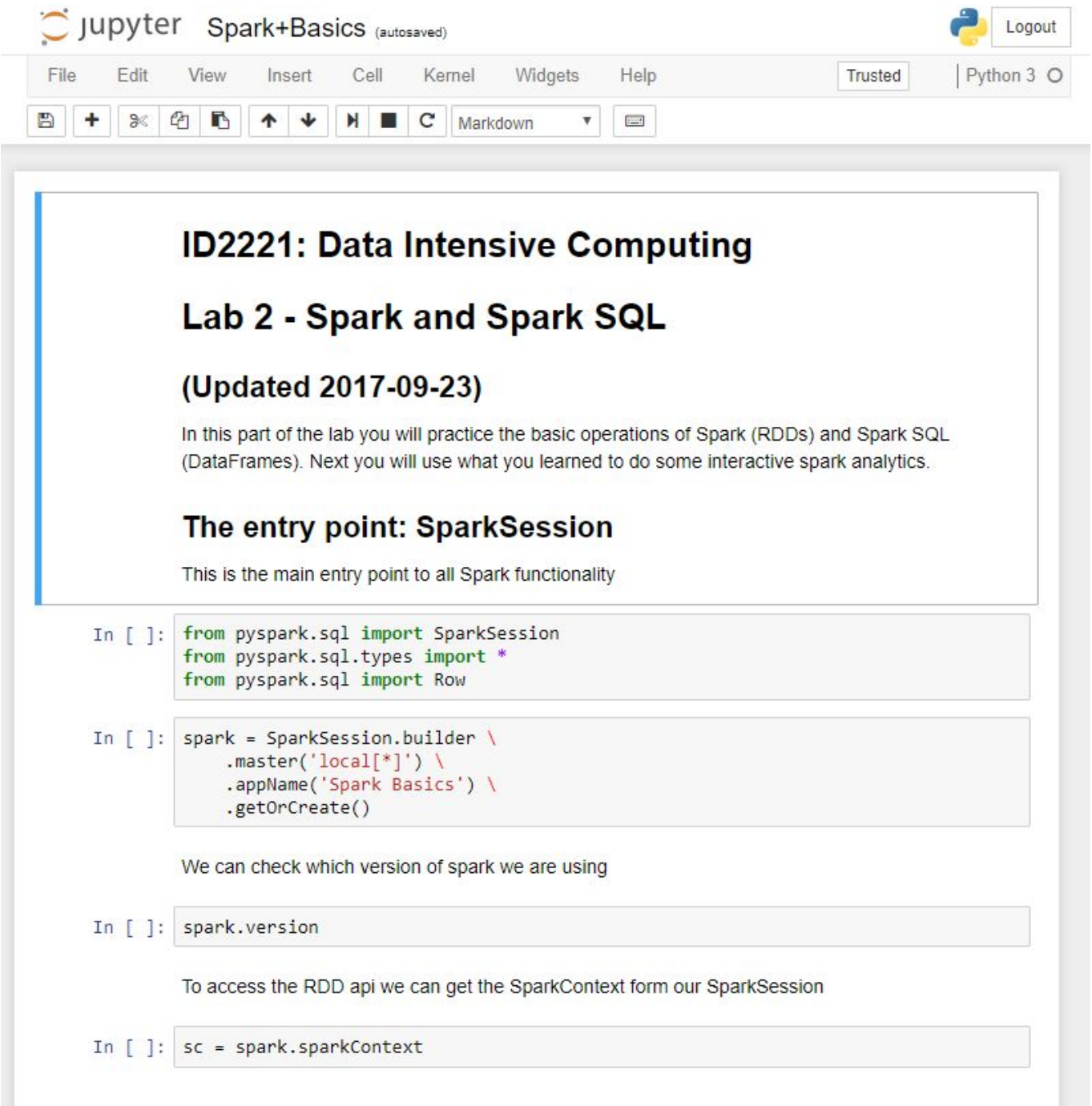
Now the files should be available in Jupyter:



You should see **two notebooks** and a **data folder**

1. **Spark Basics:** This is just for practice. Follow the instructions in the notebook and execute the code in each cell.
2. **Apache Log File Analysis:** You need to complete the code in this notebook and submit it.

When you click on a notebook, a new tab will open in your browser similar to the one below:

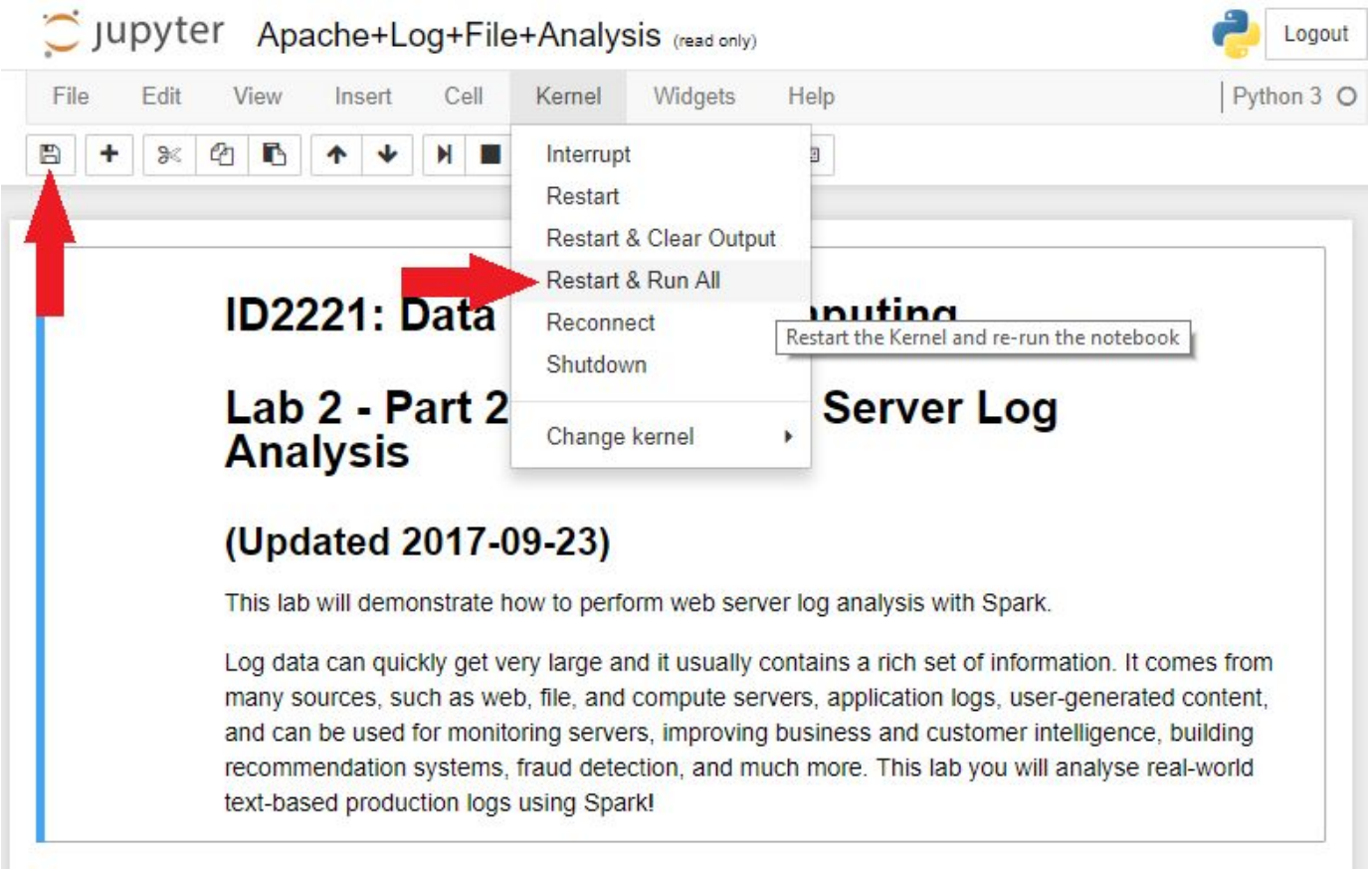


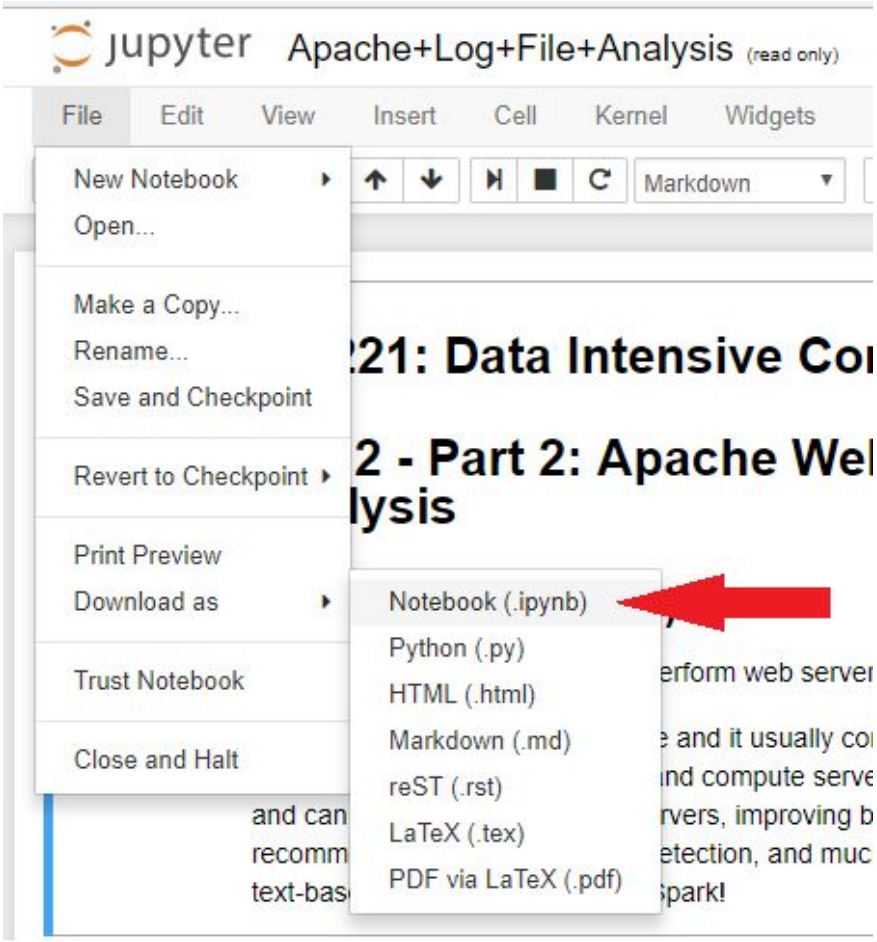
To **execute** the code in a cell, select it then press (**shift+enter**) or click the run button.

3 - Submitting Your Work

You need to complete the code in **Apache Log File Analysis**. When you are done and ready to submit your work, do the following steps:

1. In the Kernel menu, select "**Restart & Run All**".
2. **Wait** till all the instructions are executed.
3. Press the "**Save**" Button.
4. In the File menu, select "**Download as**" → "**Notebook (.ipynb)**"
5. **Upload** this file following the instructions on the course webpage.




jupyter
Apache+Log+File+Analysis (read only)
File
Edit
View
Insert
Cell
Kernel
Widgets
New Notebook
Open...
Make a Copy...
Rename...
Save and Checkpoint
Revert to Checkpoint
Print Preview
Download as
Trust Notebook
Close and Halt
Markdown
Notebook (.ipynb)
Python (.py)
HTML (.html)
Markdown (.md)
reST (.rst)
LaTeX (.tex)
PDF via LaTeX (.pdf)