My first blog

Monday:

I set up the spark environment on my local device.

* I first installed the Java 8. Many people reported that later Java version like 9 or 10 are not compatible with spark.
* Then I installed Python (the latest version) on my machine. Before installing python, I mostly ran python on Anaconda environment using jupyter notebook, and I ran a few py files using Spyder. I am not very sure about the differences, but before downloading python, I ran command line “python --version” on windows prompt and there was no output. There was an important step during the python installation is to add python to PATH which we can find and edit the system environment variables.
* Then I installed spark from apache spark official website, also the latest version. I learned that checking the software file checksum is a good way to verify the integrity of the download before installing.
* Then I added winutils.exe file for the underlying Hadoop version to my desired folder.
* Then I configured my environment variables. Mainly I added the spark, Hadoop, and java to PATH as I did to python during installation.
* Lastly, I launched spark. By calling spark-shell, I was able to run spark with scala. By calling pyspark, I was able to run spark with python.
* Two bugs took me most of the time. One is that “WARN ProcfsMetricsGetter:Exception when trying to compute pagesize” error, I didn’t solve this problem as I found that this is a minor bug that does not affect the performance and it should be just some metric issues. The second bug occurred when I tried to use pyspark. It is a bug involves JavaError, Python and py4j keywords. I tried a couple methods including downloading py4j, editing the environment variables on PATH, calling different files to launch Python, managing app execution aliases, etc. I solved it by setting the environment variable (PYSPARK\_PYTHON = /path/python) on window prompt before calling the pyspark file. (However, at the moment I am writing this blog. It does not work. I will try to figure out tomorrow.) Anyway, they all worked on Monday.

Tuesday:

Basically I didn’t do anything because the repair took me whole day.

Wednesday:

* I installed Journal. I followed the instruction although I could not understand it completely. What I learned the most from the process is having a better idea about what docker is and how to use docker. I successfully downloaded the Journal image and run the container. However, I don’t know how to use the container. I still don’t know how to run the Journal command line.
* I also implemented fastpages for the my blog. I didn’t change much content and interface of the blog and add additional designs to it yet as docker took me most of the time.