CS 696 Intro to Big Data
Fall Semester, 2020
Assignment 2

© 2020, All Rights Reserved, SDSU & Roger Whitney

San Diego State University -- This page last updated 3/5/20

Due Mar 19 23:59

Performance of Tools

The goal of this assignment is to study the relative performance of a set of tools: pandas, dask, vaex, HDF5 and PySpark. We will follow the article "How to analyze 100 GB of data on your laptop with Python" which is at https://towardsdatascience.com/how-to-analyse-100s-of-gbs-of-data-on-your-laptop-with-python-f83363dda94. You should read that article before starting this assignment.

In that article Jovan Veljanoski analysis 100 GB of Yellow Taxi data from New York City. Since some students may not have 100 GB of free disk space we will use the data from the Green taxi company, which is only 10GB of data. You can download the data from the assignment page for the class. Note that the data dictionary for the Green Taxi data has slightly different labels for two fields. The article has Jupiter notebook of the analysis done in the article so you might want to look at that.

You are going to measure the time it take to perform the following tasks using pandas, dask, vaex and PySpark.

- Plot the number of unique trips with certain number of passengers in the entire green taxi data. (Green taxis can not operate in parts of NYC.)
- Filter out the trips longer than 100 miles and trips that claim more than 10 people.
- On the filtered data produce a heat map by pick up location color coded by average fare amount,
- Compute the arc distance for the filtered data and plot the distribution number of trips of trip distance and arc distance.

The article contains the code to perform these operations. The goal of this assignment is to compare using pandas, dask, vaex and PySpark to perform the above task. Runtime performance of each to perform the task is one metric to report. But also discuss any issues the you had with using any of the 4 systems.

You also need to convert the data from a collection of CSV field into a single HDF5 file. You are to use that file as input.

Your jupiter notebook(s) for this assignment should run the 4 tasks for each of the four different tools sets (pandas, dask, vaex and PySpark) measuring the time they take. It also should contain a comparison of the performance of the four systems, a discussion of the performance differences and a discussion of the issues in using them.

Instructions

You are free to use any IDE to write your code.. However you are to turn in a Jupyter Python notebook. Your jupyter notebook should be self contained. All calculations and answers to the questions are to be in one notebook. This assignment requires you to use files, some of which are provided. You notebook needs to read the unmodified files, including names. Any needed modification to the files needs to be done in the notebook.

At the beginning of your notebook you should create variable that hold the path (plus name) of any input files that you use. It is likely that for grading purposes those paths will need to change. I should be able to run your notebook using my input files by just changing the path to files at the top of your notebook.

Notebooks can contain text, code and output. Use text to indicate what problem you are solving. The code used to answer the problem need to be complete.

Grading

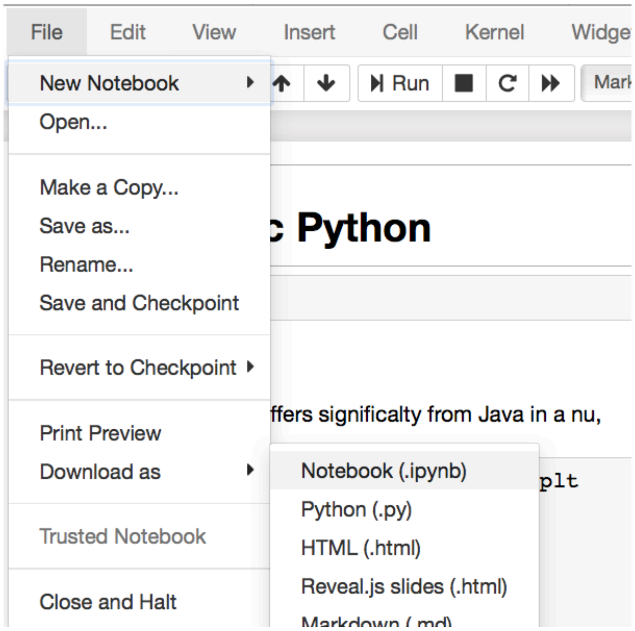
Each problem is worth 10 points.

What to turn in

To turn in your assignment download your Jupyter notebook as an IPython Notebook (.ipynb). See image below. This will allow me to run your assignment in Jupyter. Note that when you download your assignment it will create a file with the extension .ipynb.json. I will remove the .json extension.

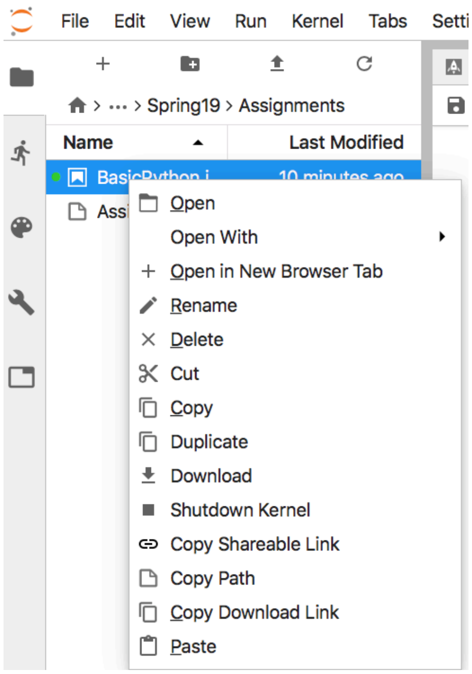
Once you have downloaded the assignment zip it up and then upload the zip file to the course portal.

Using Classic Juptyer Notebook



Using JupterLab

Right-click on the Notebook name in the file browser and select download.



Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.