CS6502: Week 8 Lab Exercises

Dr. Andrew Ju

March 15, 2021

Exercise 1

The lab in this week is **supervised**. The lecturer/lab assistant will be available between 15:00 – 16:00 for any questions you may have.

The main task for today's lab is to continue with Spark:

- finish practice with Spark (`PySpark shell`)
- practice Spark examples in Spark environment (`spark-submit`)
- get started with Dataproc
- finish lab assignment

In Sulis class page, there is a forum "Labs (week 8)" created. In the forum, there is one topic:

- *Discussion*, use this topic to post your questions and discuss with your fellow classmates.

The deadline for this week's lab assignment is 12am (noon time) of Sunday 21st March.

Below is a list of steps you should follow for today's lab.

- Practice PySpark (based on lecture slide) via PySpark Shell
  - Read though list of transformations/actions
  - try out a few transformations (map, flatMap, filter, etc)
  - try out a few actions (reduce, take, first, collect, etc)
  - try out `persist`, `collect`
  - try out shared variables
- Practice Spark examples using `spark-submit`
  - wordcount
  - pi
  - airline example

- Practice Spark examples using Google Dataproc (please make sure to stop the cluster/instances after your practice)
  - Please figure out how to calculate the cost of running the cluster
- Finish the lab assignment (see spec below)

Lab assignment spec

Met Eieann, the Irish National Meteorological service, is the leading provider of weather information and related services for Ireland. They have made available a large amount of historical weather data via link here.
In this assignment, we are particularly interested in the data from Shannon airport station in County Clare.
To get the data:

- Select "Daily" in Data resolution
- Select "Clare " in County
- Select "Shannon Airport" as the station
- Click "Download the full data series"
- In the link section below, click "Download the full daily data series"
- In the downloaded zipped file, there is a csv file named `dly518.csv`, this is the file for you to work on

Your task for this week's assignment is to write a PySpark application to compute for each year which day has the longest sunshine duration (any day would be fine if there are more than 1)
In the Sulis Assignment section, there is a new assignment "Lab 8" created, when submitting please structure your answer as below

- Answer section
  - Please write down the result you got
- Source code section
  - Please paste your source code here
- Screenshots
  - Please paste your screenshots here
  - you should have one screenshot for your output, one screenshot from your bucket storage (that shows list of files you used), and one screenshot of your job specification (where you specify job type, job name, main python files, etc)

Note that: please complete the task yourself, copy/paste code from Internet or your classmates will result zero mark for the assignment, and in the worst case an F for the entire module!