CS6502: Project Spec for SE cohort

Dr. Andrew Ju

April 12, 2021

1 Project spec

[Note that: class discussion or search on the Internet is permitted. But 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!]

In computing, the three steps of extract, transform, load (ETL) is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s). In this project, you are asked to complete the three steps:

1. Extract, which is to extract the data from the source system(s).
2. Transform, during which a series of rules or functions are applied to the data from the Extract step.
3. The third step, Load, loads the data into the end target (could be any data store including a simple delimited flat file or a data warehouse).

The university has a timetable website (I assume each one of you has already used the timetable website to get your class/module timetable) where you can access your personal timetable as well as timetable for each module that is offered in the university.

In this task you are asked to collect "course timetable" information for all groups of students (e.g., Year 1-4 students for Bachelor of Arts in Law and Accounting, or Year 1 students for MSc in Software Engineering), and store them in a simple delimited flat file.

Once finished the ETL process, you can then use Spark SQL to load the data, and count how many lecture sessions (sessions with type LEC) in total are there for this semester.

Submission requirement

- parsed file (.csv or any format you may prefer)

- your scripts
  - scripts for ETL
  - PySpark scripts
- screenshots of you running scripts locally
  - running scripts for ETL
  - output of PySpark Job

2 Suggested steps

[Note that you may use whatever you see fit for the ETL process (bash, Python, Java, C#, etc). Below are simply for your reference, and you don't have to follow all the steps.]

This project is largely based on Chapter 2 "Ingesting Data into the Cloud" of suggested book (link here). Below are the steps I suggested to follow to complete the project. Since in the book the author has explained each step in great detail, make sure you have read the chapter beforehand.

1. Setup a new project in Google Cloud Console ($\sim$ 2 minutes)
2. Under this new project, use git clone to copy sample source code to your work directory. ($\sim$ 1 minute)
3. Read the README.md file ($\sim$ 5 minutes)
4. Try out the BTS web interface here, and manually download a dataset. (e.g. download the dataset for Dec 2019) ($\sim$ 10 minutes)
5. Read Table 2.1 of Chapter 2 (Selected fields from the airline on-time performance dataset downloaded from the BTS) ($\sim$ 2 minutes)
6. Read sections "Ingesting Data", and "Reverse Engineering a Web Form" and "Dataset Download". Make sure to try out the steps yourself. ($\sim$ 50 minutes)
7. In the previous step you have learned how to reverse engineering a web form, now try the same technique on UL Timetable site here. See if you can figure out a way to automatically download the timetable per course and year. ($\sim$ 30 minutes)
8. Read section "Exploration and Cleanup", and links from the "Useful readings" below. Make sure to try out the steps yourself. ($\sim$ 30 minutes)
9. In the previous step you have been asked to try to reverse engineering the web form on UL Timetable site. Now, in this step - Transform, try to write a parser to convert the raw HTML file into a format that can be loaded into BigQuery. Note that you may define a schema first before this step. (Try to include as many columns as possible) ($\sim$ 60 minutes)

10. Read sections “Uploading Data to Google Cloud Storage”. Make sure to try out the steps yourself. ($\sim$ 30 minutes)

Useful readings

- Bash scripting cheatsheet
- Parsing HTML: A Guide to Select the Right Library
- Python HTML Parser