Data Engineering Assignment

Given the Github API (https://docs.github.com/en/free-pro-team@latest/rest), create a simple ETL pipeline.

For a particular repository, example Airflow (https://github.com/apache/airflow), pull in the commits over the last 6 months. (plus points if this window of time can be varied)

With the data ingested, address the follow queries:

- For the ingested commits, determine the top 5 committers by count of commits and the number of commits.
- For the ingested commits, determine the committer with the longest commit streak.
- For the ingested commits, generate a heatmap of number of commits count by all users by day of the week and by 3 hour blocks.

Sample heatmap

	00-03	03-06	06-09	09-12	12-15	15-18	18-21	21-00
Mon								
Tues								
Wed								
Thurs								
Fri								
Sat								
Sun								

The exercise can be done with a series of scripts and simple local database. Candidate to use any tools and language that they are comfortable with. We expect a commitment of between 4 to 6 hours for this assignment. Candidates can commit more time if it is fun.

We will be looking at the scripts, documentation on how to setup and run the scripts.

Good luck and have fun!

Hints:

To get a project, the request need to include the organsiation and the project (https://docs.github.com/en/free-pro-team@latest/rest/reference/projects#list-organization-projects)

Commits API: https://docs.github.com/en/free-pro-team@latest/rest/reference/repos#list-commits

There are 2 user concepts, an author and a committer. We will use the author object.