Purpose

- Evaluate your design, communication and coding abilities
- Serve as a simplified example of a task one may encounter as a Software Engineer at EQT

Data Enrichment Task

Your goal is to build a pipeline that enriches the information on EQT portfolio companies scraped from our public website with the data from another dataset that is provided. It should export the resulting enriched dataset to a storage location of your choice in a data model that you design.

Portfolio Companies

EQT Funds are listed here, and portfolio companies are here and here. Each company is described by its name, and, optionally, sector, country, and enter / exit information. Some of the portfolio companies have dedicated pages with a detailed description, board information and possibly a link to the company's website.

Organization Reference Data

Stored in GCS as a compressed file with JSON Lines, contains company name, website URL and some additional data such as description, number of employees, etc.

Resulting Dataset

- Should only contain portfolio companies, no need to repeat the entries from the reference dataset that aren't relevant.
- Should contain as much information as possible to obtain from the given data sources.
- Should be in one of the common data formats, such as Avro or JSON.

Evaluation Criteria

- Completeness and correctness of the resulting dataset
- Code quality
    - Tests
    - Documentation
    - Ease of use
    - Use of version control
- Note that the task doesn't include any sort of a graphical user interface

Instructions

- We expect the task to take 4-6 hours depending on your experience with web scraping.
- Use Python or Java as those are the two main programming languages we do data engineering in.
- The solution should be published as a public Git repository which includes all the code necessary to produce the resulting dataset.
    - It could be good to upload the resulting compressed dataset too.
- Feel free to reach out with any questions, however the task has intentionally not been too strictly defined to allow for some degree of freedom.