Data Pipeline

# Summary

The purpose of this data pipeline is to correctly collect and aggregate data from multiple sources, in order to answer the questions mentioned in the initial stakeholder requirement document. This pipeline will regularly be used to deliver raw data from various API and CSV sources from different entities, process, then insert into a MySQL database.

For now this is the best way to continue because of the uncertainty of the longevity of this project (this depends on how useful I deem this project to be, and the current time constraints I face, being at school and it being job hunting season).

# Problems

* Keeping track of last batch collection, per API
* Standardizing formats across all APIs
* Converting CSV style formats to row/column formats
* Routinely and automating collection of download only data

# High-Level Description

Will be done using python scripts that will ideally be run using discretion. The python scripts will initially detect when the last batch was successfully uploaded to the database, and will then proceed to collect the next batches to bring database up to date.

Scripts will call APIs and gather raw data, and will then proceed to transform data into standardized and documented format, and will then insert into MySQL DB.

# Technical Metadata

Data availability: N/A

Data start date: 11/23/23

# Low-Level Metadata

To be determined.

# Comments

# Document Logs

|  |  |
| --- | --- |
| Date | Comment |
| 9/9/24 | Creation and establishing details. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |