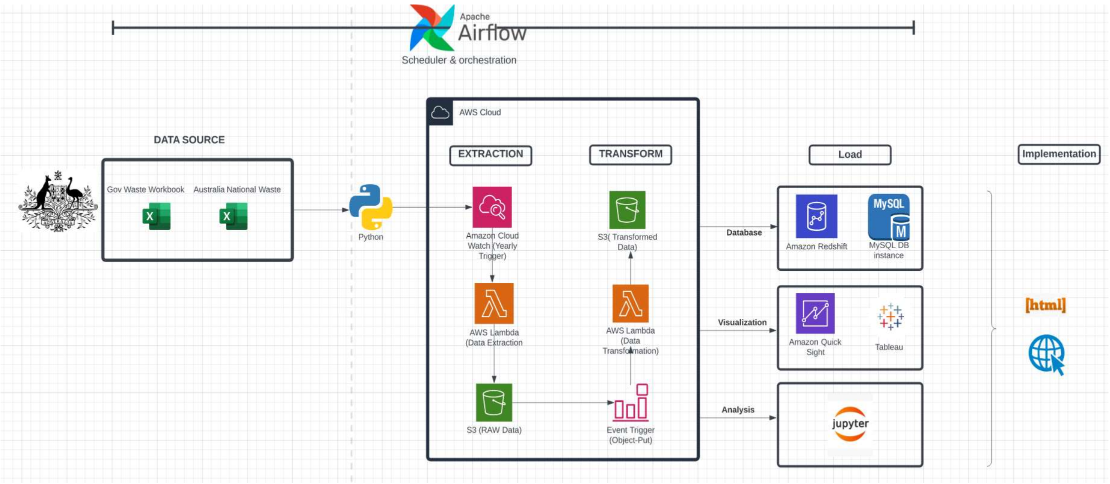
Waste Insight Planned ETL Workflow



- **Data Source:**
  - ➢ Australian Bureau of Statistics – Gender Indicators
  - ➢ Victoria State Government
- **Scheduler:**
  - ➢ Airflow: it creates, schedules and monitor data workflow, very helpful when managing data pipeline.
- **Extraction:**
  1. Create a Lambda function to extract data from open sources.
  2. Set triggers via CouldWatch. Once a year, the lambda function will extract data from the open data source
  3. The extracted raw data will be sent to an AWS S3 bucket/Data lake
- **Transform:**
  1. Create a Lambda function to transform the data in the AWS S3 bucket containing the raw data. And send the transformed data to a new bucket
  2. Set up a trigger so Lambda will run whenever a file is added to the raw bucket.
- **Load:**
  - ➢ Load to database/management; eg. Redshift, Mysql
  - ➢ Load for visualization; eg. QuickSight, Tableau
  - ➢ Load for Analysis; eg. JupyterNote Book
- **Implementation:**
  - ➢ Dashboard or meaningful information discovered from the data will be write in html and implemented in our website
- **Note:**
  - ➢ Everything can be done automatically, just need a data engineer to monitor and maintain the whole process
  - ➢ To extract newest file, just change the year of file. For example 2020.xlsx to 2021.xlsx, which can be done by python string manipulation
- **Other Plan:** Build unit test for ETL pipeline

Open Data Sources

Data 1:

Name	Gender indicators
Link	https://www.abs.gov.au/statistics/people/people-and-communities/gender-indicators
Physical Access Used	EXCEL
Frequency of Iteration	Yearly
Granularity	Salary per hour; weekly hours worked, etc
Copyright details	https://www.abs.gov.au/website-privacy-copyright-and-disclaimer#copyright-and-creative-commons

Schema: To be decided

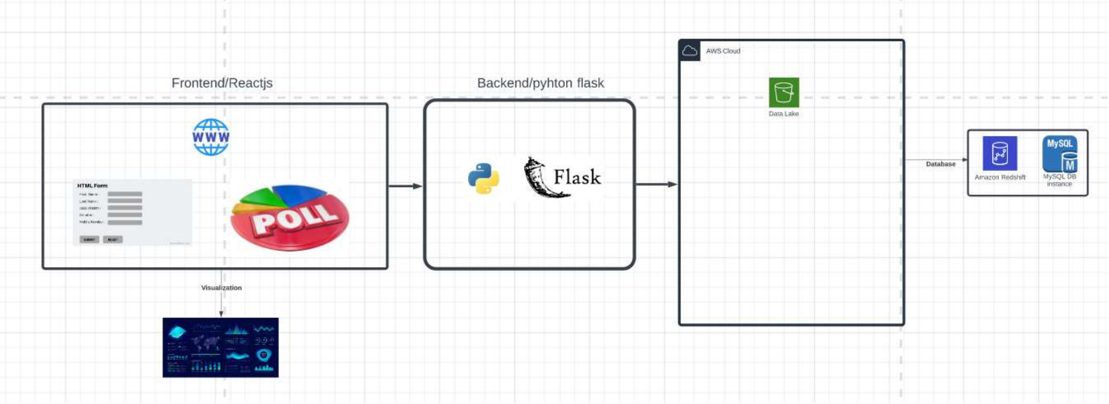
Column	Description

Data Model: To be decided

Example Code snippet: To be decided

Environment: python 3.8

Website Data Flow



- **Data Flow:**
  - A Html form or poll is deployed, audience share their opinion on gender inequalities and visualize it
  - May use Reactjs as frontend, python flask as backend, data from website will be stored in AWS S3 and load into Mysql database

Schema: To be decided

Column	Description

Data Model: To be decided

Example Code snippet: To be decided

Environment: python 3.8