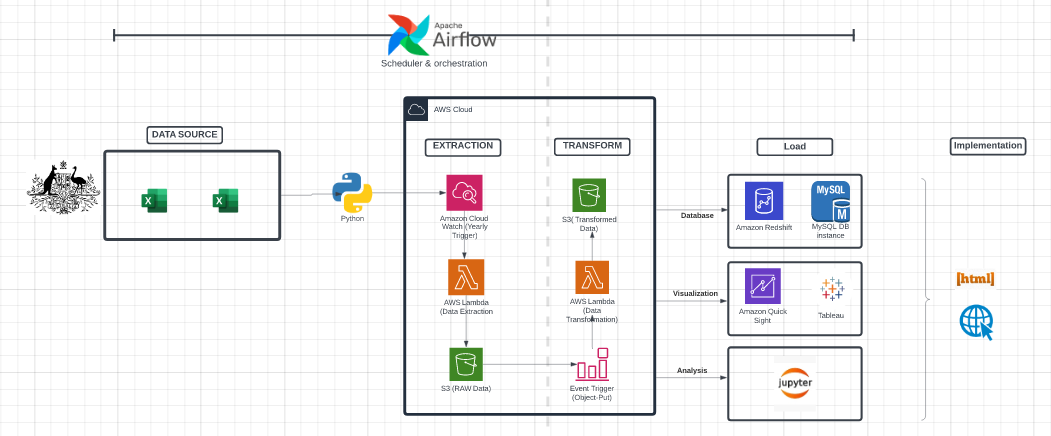
**Gender Inequality ETL Workflow**



* **Data Source:**
  1. Australian Bureau of Statistics – Gender Indicators
  2. Victoria State Government
  3. Lifecycle: Because data is from government, it is not updated frequently, so the function only runs once a year. Past data will be stored in data lake
* **Scheduler:**
  1. 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
  + Data will be cleaned and checked after loading into this application. Process:
    1. Check its completeness: check missing value, outlier, duplicate value
    2. Data cleansing: identify the issue of data and make decision, remove outlier aggregate data and remove missing values
    3. Data transformation: depends on the use of data, transforming and scale the data into desired form.
    4. Validation: Check the data meet the desired standards, use statistical significance and verify the accuracy of calculations.
* **Implementation:**
  1. Dashboard or meaningful information discovered from the data will be write in html and implemented in our website
* **Note:** 
  1. Everything can be done automatically, just need a data engineer to monitor and maintain the whole process
  2. 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
* **Use of Data:** The data will be captured and stored according to the above description. The data is then used for analysis and graphing. The method of use will be strictly in accordance with the copyright regulations of Australia Bureau Statistics. Eg. Cite the website when use their data.
* **Data insight:** 
  1. Gender pays gap: pay rates and working hours.
  2. Education: data on education enrollment rates
  3. Gender-based Violence: Data on sexual assault and domestic violence. Mental health issues

**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 |

**Data Wrangling Code snippet:**

Text

Description automatically generated

**ERD Diagram:**

