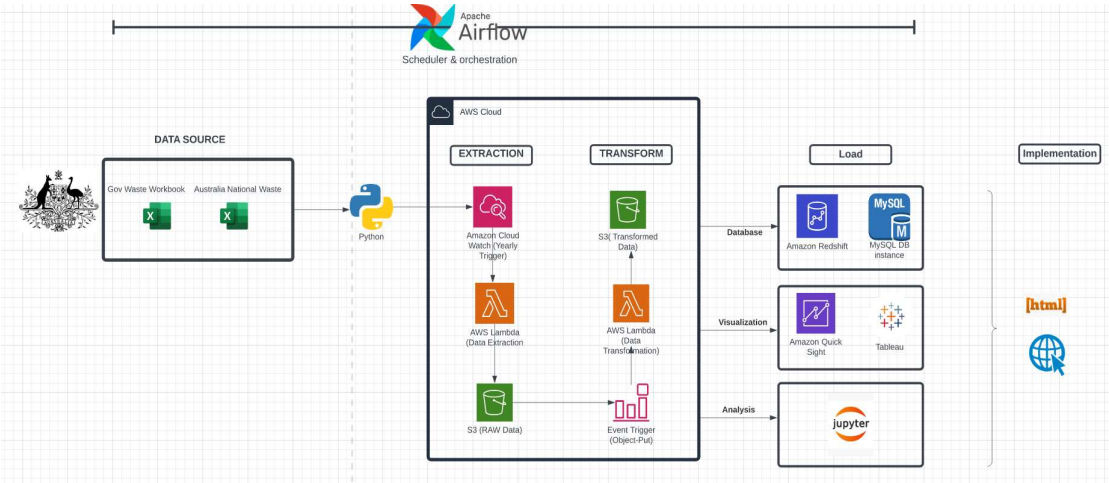
## Waste Insight Planned ETL Workflow




- **Data Source:**
  - Australia Government, Department of Climate change
  - 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

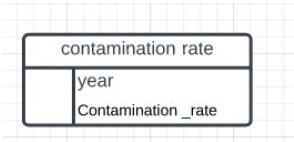
## Open Data Sources

**Data 1:**

| | |
|---|---|
| **Name** | Waste Services Workbook |
| **Link** | https://assets.sustainability.vic.gov.au/susvic/Workbook-Waste-Local-Government-Waste-Services-Workbook-2019-20.xlsx |
| **Physical Access Used** | EXCEL |
| **Frequency of Iteration** | Yearly |
| **Granularity** | Amount of money or tonnes of waste per year |
| **Copyright details** | https://www.sustainability.vic.gov.au/about-us/legal-and-policies/copyright |

**Schema**

| Column | Description |
|---|---|
| Year | Year of record |
| Contamination rate | Rate of wrong items in the recycling bin at Victoria |




**Example Code snippet**

Environment: python 3.8

**Extract Data:**

```
#download data
link = 'https://assets.sustainability.vic.gov.au/susvic/Workbook-Waste-Local-Government-Waste-Services-Workbook-2019-20.x
r = requests.get(link, allow_redirects=True)
with open("Waste-Services-Workbook-2019-20.xlsx",'wb') as f:
    f.write(r.content)
```

**Load:**

```
organics = pd.read_excel('Waste-Services-Workbook-2019-20.xlsx', 'Organics', header = None)
Garbage = pd.read_excel('Waste-Services-Workbook-2019-20.xlsx', 'Garbage', header = None)
Recyclables = pd.read_excel('Waste-Services-Workbook-2019-20.xlsx', 'Recyclables', usecols='A:I', header = None)
Organics = pd.read_excel('Waste-Services-Workbook-2019-20.xlsx', 'Organics', header = None)
```

**Extract table:**

```
Contamination_rate_table = Recyclables.iloc[29:48]
Contamination_rate_table = Contamination_rate_table[[0,1]]
Contamination_rate_table.columns = ['year', 'Contamination_rate']
i = 0
for year in range(2001,2020,1):
    Contamination_rate_table.year.iloc[i] = year
    i = i +1
```

```
Contamination_rate_table.head()
```

| | year | Contamination_rate |
|---|---|---|
| 29 | 2001 | 7.6 |
| 30 | 2002 | 7.2 |
| 31 | 2003 | 10.1 |
| 32 | 2004 | 10.9 |
| 33 | 2005 | 11.5 |

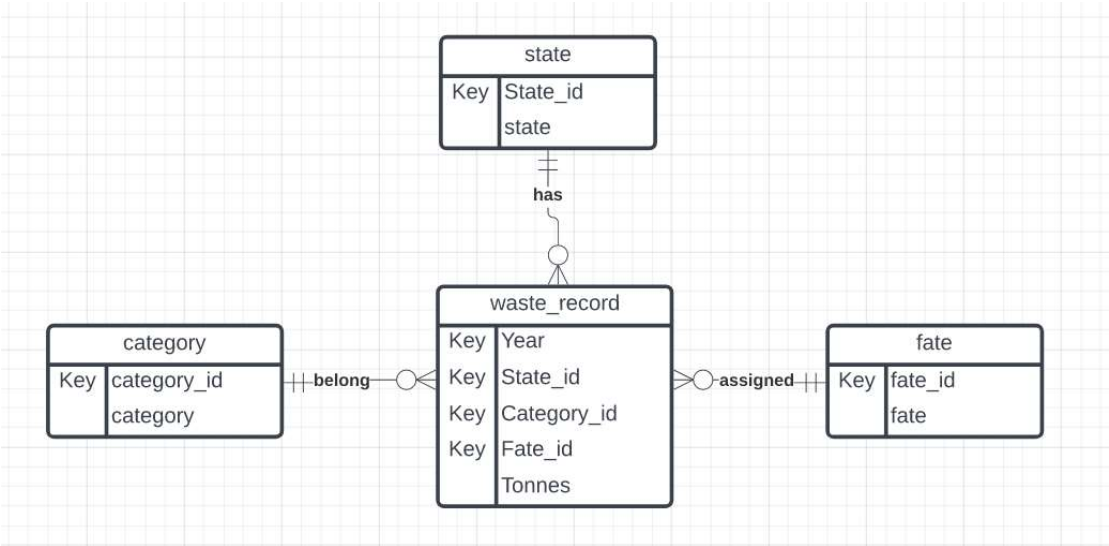
**Data 2:**

| | |
|---|---|
| **Name** | National Wastes Database |
| **Link** | https://www.dcceew.gov.au/sites/default/files/documents/national-waste-database-2022.xlsx |
| **Physical Access Used** | EXCEL |
| **Frequency of Iteration** | Yearly |
| **Granularity** | Amount of money, rates or tonnes of waste per year |
| **Copyright details** | https://www.dcceew.gov.au/about/copyright |

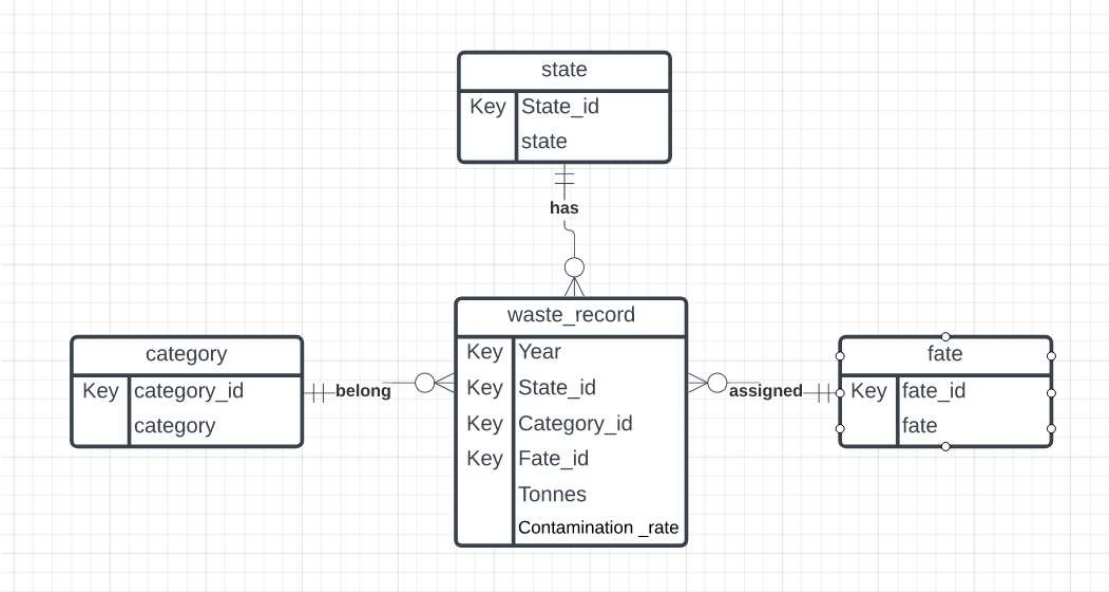
**Schema:**

| Column | Description |
|---|---|
| Year | Year of record |
| Jurisdiction | State, for example: VIC,TAS |
| Category | Waste category, for example, hazard waste, glass, plastic |
| Management | How to manage the waste, for example: recycling, waste reuse |
| Fate | How to dispose the waste, for example: recycling, energy recovery |
| Tonnes | Total tonnes of waste |

**Conceptual Model**




**Data 2 and Data 1 merge:**




**Notes:** will find description for each category and fate and add into database in the future

**Code Snippet**

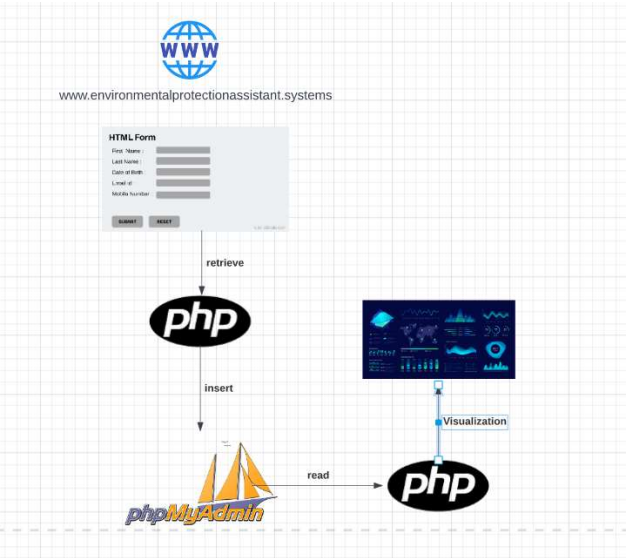
Environment: python 3.8

**Extract Data:**

```
#download data
national_waste_db = 'https://www.dcceew.gov.au/sites/default/files/documents/national-waste-database-2022.xlsx'
r = requests.get(national_waste_db, allow_redirects=True)
with open("national_waste_2022.xlsx",'wb') as f:
    f.write(r.content)
national_waste_db = pd.read_excel('national_waste_2022.xlsx', 'Database 2022', usecols = "A,B,C,H,I,J")
```

- The data is load into Tableau for visualization directly

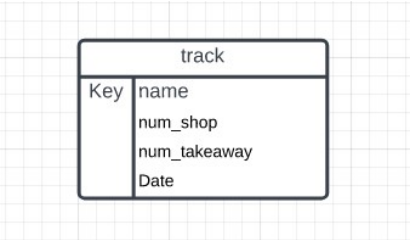
**Website Data Flow**



- **Data Flow:**
  - A Html form is deployed at Track on Waste page, user can enter and record their progress in developing sustainable habits
  - Data will be retrieved by a php function and insert into a table from WordPress database
  - Another php function can read the data from database and plot graphs

**Schema:**

| **Column** | **Description** |
|---|---|
| Name | User's name |
| Num_shop | How many times did the user shop |
| Num_takeaway | How many times did the user order takeaway |
| Date | Date of record |




**Notes:** The current user group is Rosa herself and some of her friends, so the names will not be repeated. When the user group expands in the future, the E-mail will be used as the Primary Key

## Reference

Dashboard, web form and data extraction code link:
https://github.com/JRenMike/FIT5120-Data-Governance