DATA PIPELINE DOCUMENTATION

Table of Contents

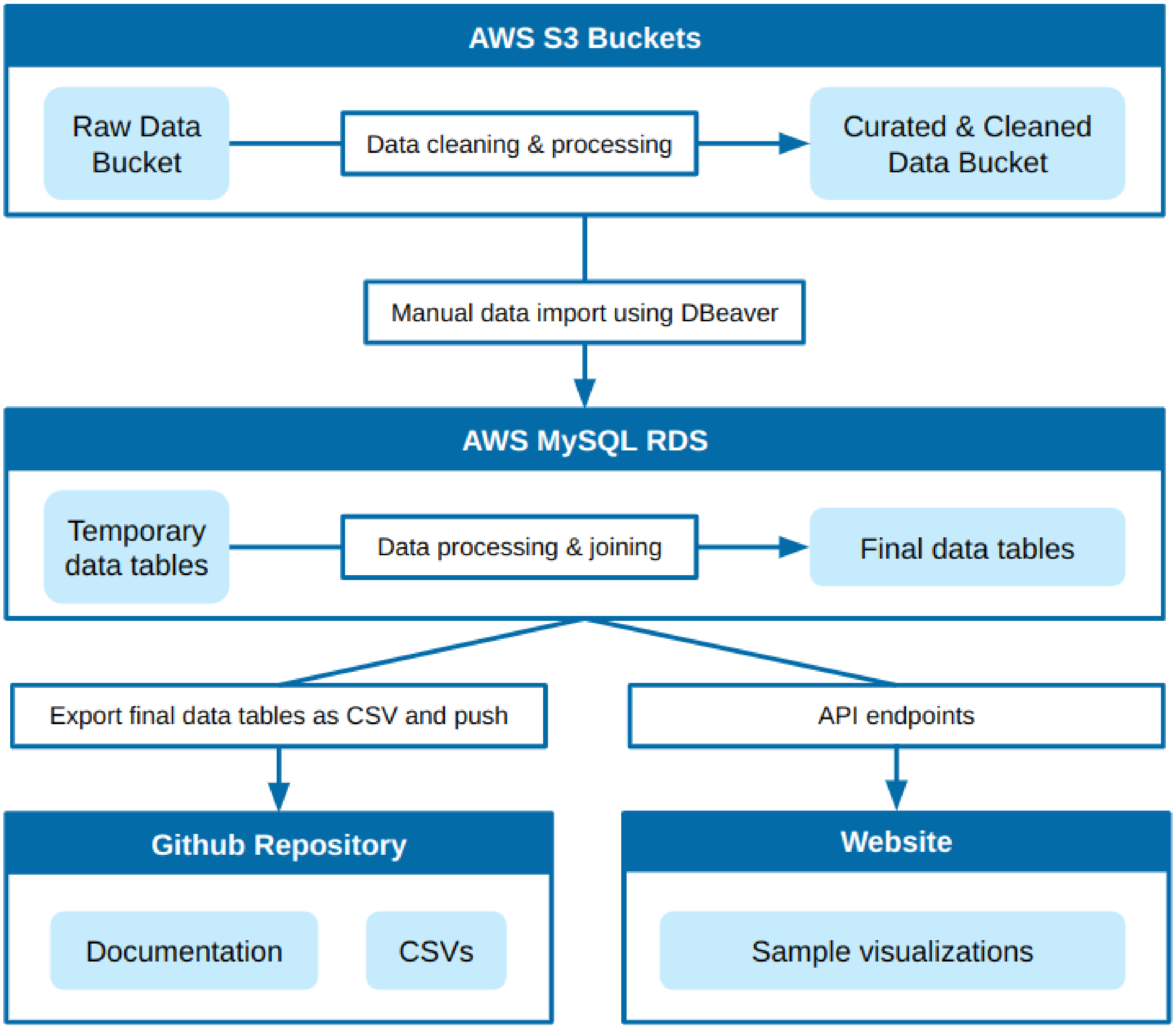
Introduction

This document describes the data pipeline used by the Winter 2023 INFO 498 Capstone class. It provides an overview of the overall pipeline, and then delves deeper into each component, explaining why each tool was chosen and how it is used in the context of this project.

Adapted from Microsoft's Aether Data Documentation Template. More information about this template can be found here.

Data Pipeline Overview

VISUAL OVERVIEW OF THE PIPELINE



The above is a visual representation of the project's data pipeline. Much of the data movement is manual; this is due to the small scale of the project and the need to ensure high quality data.

AWS S3 BUCKETS

Once the raw data has been procured (either via web scraping or download of data files), it is placed in an AWS S3 Bucket. From there, it is then processed and cleaned using scripts, and then placed into a separate AWS S3 Bucket that is exclusively for cleaned & curated data.

AWS RDS - MySQL

Our database is a relational database hosted in AWS, using a MySQL engine. Once the data is available in CSV format in the AWS S3 cleaned & curated data bucket, we manually download it and import it into the database using DBeaver. From there, it is processed and joined so that there are several larger data tables that can be used to create visualizations or cross-reference data.

API

In order to provide the web-team with the visualization data, we create API endpoints using AWS API Gateway service. Each API endpoint in the gateway service is powered by an AWS Lambda function where we call the queries to get the required data from our AWS RDS Database instance. The API endpoint is called on the client-side through Axios, a JS library used to make HTTP requests Node.js. From there, the web-team utilizes the data from the API endpoint to create visualizations on the client side.

Github Repository

Once the data is ready in the RDS, it is exported in CSV format and pushed to the Github repository, which is open source.

AWS S3 Buckets

OVERVIEW

TEXT HERE

RAW DATA BUCKET

TEXT HERE

DATA CLEANING & PROCESSING

TEXT HERE

CLEANED & CURATED DATA BUCKET

TEXT HERE

AWS MySQL RDS

OVERVIEW

TEXT HERE

DATA IMPORTING

We use a manual process to transfer the data from the AWS S2 Bucket to our database. First, the data is downloaded as a CSV (or several CSVs) locally onto a computer via the AWS S3 interface. Then, using DBeaver, import tables are created to host the data. These tables have to be created in order to import the data; this is a DBeaver requirement. We chose to use DBeaver because we ran into issues with MySQL Community Workbench.

Once the import tables are ready, we follow this process to import the data: https://dbeaver.com/2022/06/23/import-data-with-dbeaver/

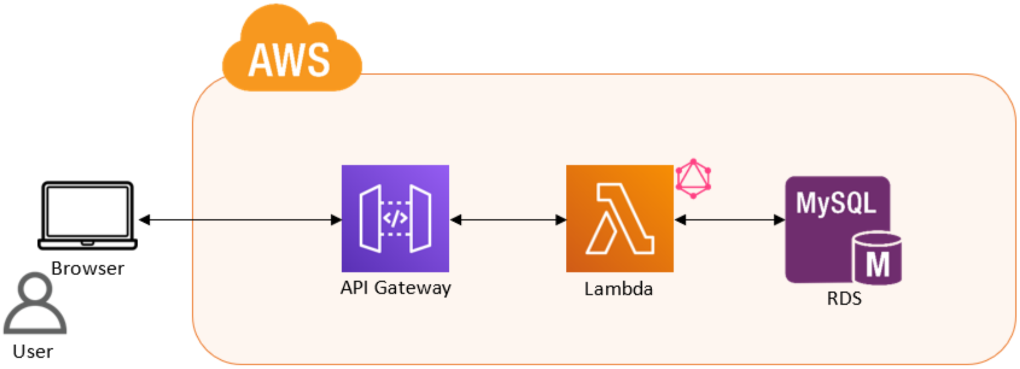
FINAL DATA TABLES

Secondary Data ERD

Data Table Processing & Joining

API

OVERVIEW



ENDPOINTS

GET

https://rprwae53w2.execute-api.us-west-2.amazonaws.com/v-1/cases-by-county

Requires a query parameter 'County' where users can get data for total covid cases, expected covid cases, and 7 day average of covid cases by County.

Here's a comprehensive list of 'County' names the endpoint accepts:

```
countyNames = ['Adams', 'Asotin', 'Benton', 'Chelan', 'Clallam', 'Clark',
'Columbia', 'Cowlitz', 'Douglas', 'Ferry', 'Franklin', 'Garfield', 'Grant',
'Grays Harbor', 'Island', 'Jefferson', 'King', 'Kitsap', 'Kittitas',
'Klickitat', 'Lewis', 'Lincoln', 'Mason', 'Okanogan', 'Pacific', 'Pend
Oreille', 'Pierce', 'San Juan', 'Skagit', 'Skamania', 'Snohomish', 'Spokane',
'Stevens', 'Thurston', 'Wahkiakum', 'Walla Walla', 'Whatcom', 'Whitman',
'Yakima', 'Unassigned', 'Western Washington', 'Unassigned Region', 'Better
Health Together', 'Elevate Health', 'Greater Columbia', 'Healthier Here',
'North Sound', 'Olympic Community of Health', 'Southwest Washington','Cascade
Pacific Action Alliance', 'North Central', 'Unassigned ACH']
```

Example API Call

If I wanted COVID-19 data for King county, here's what my endpoint request would look like:

https://rprwae53w2.execute-api.us-west-2.amazonaws.com/v-1/cases-by-county?County=King

Sample JSON Data from the Endpoint

```
{
  "EarliestSpecCollectDate": "2020-01-10",
  "County": "King",
  "TotalCases": 0,
  "ConfirmedCases": 0,
  "ProbableCases": 0,
  "TotalCases_7DAvg": "4.00000",
  "SevDayCaseCount": "28.00000",
  "SevDayCaseRate": "1.20000",
  "FourteenDayCaseCount": "42.00000",
  "FourteenDayCaseRate": "1.90000",
  "DateTimeUpdated": "2023-02-02 06:28:14"
}
```

GET

https://rprwae53w2.execute-api.us-west-2.amazonaws.com/v-1/latino-populations

Returns the