**Problem Statement**: Explore which federal government agencies have the highest web traffic and which have low traffic and how this varies over time. The federal government provides a vast array of services and information to the American people through its websites. Websites that have little traffic may be poorly designed or not well-publicized. When websites have very high traffic, this may imply that the government must be prepared with resources to meet the high demand for both web traffic and downstream service delivery. In this project, I build a data pipeline, ingesting data in batches from an API weekly, transforming the data with Spark, storing the data in a mySQL database, and visualizing the output with python’s altair package. I orchestrate the workflow with Apache Airflow.

**High Level Overview of Steps**

1. Install required software
2. Use python to extract historical website visits data from the analyicts.usa.gov API (Jan 1, 2020 through May 1, 2023) and store the raw data in a mySQL database.
3. Setup directed acyclic graph in Apache Airflow to do the following steps **once a week:**
   1. Task 1: Extract the latest website visits data
   2. Task 2: Transform and store the data so it is ready for analysis.
   3. Task 3: Load transformed data into mysql into python. Visualize using the python altair package.

**Big Dataset Source**: [analytics.usa.gov API](https://open.gsa.gov/api/dap/)

**Hardware**: MacOS Monterey Version 12, M1 chip, 16GB

**Software:** Python 3.9, Spark 3.3.1, Apache Airflow 2.6.0

**Benefits**: Airflow has a user-friendly visual display of the pipeline.

**Challenges**: I encountered a few hiccups as I was new to the airflow technology (1) airflow uses the UTC timezone for scheduling jobs (2) if you need to update the configuration file, afterwards, be sure to reset the airflow database, webserver, and scheduler. (3) Your scheduler’s first run is sensitive to the start\_date setting. For example, if you want to run a weekly job that starts today, be sure your start date is more than a week prior to today.

**YouTube URL:** [Short Video](https://youtu.be/AYaaRbIt7I4) | [Long Video](https://studio.youtube.com/video/CKeJR5OuxSQ/edit). **Git Repo** [Link](https://github.com/jhsmith22/e63finalproject.git).

**DATASET SIZE** (Jan 1, 2019 – May 7, 2023): 3,510,428 rows

**MOST VISITED AGENCIES IN 2023:**

1. Department of Health and Human Services
2. Postal Service
3. Commerce
4. Treasury

**LEAST VISITED AGENCIES IN 2023:**

1. Nuclear Regulatory Commission
2. U.S. International Development
3. National Science Foundation
4. Housing and Urban Development

**NEXT STEPS:**

* Additional data from API: device use (mobile, desktop)
* Analyze individual website domain trends
* Integrate weekly/monthly economic data into the pipeline
* Bring in weekly twitter data for each agency to highlight topics of conversation, key influencers
* Serve on website