CSCI 5323 Project Proposal: Understanding US Federal Spending

Ahmed Al Hasani, Mohamed Al-Rasbi

Problem Description

Since 2007, a government website launched to provide access to the federal spending data mandated by the Federal Funding Accountability and Transparency Act of 2006. The website, USAspending.gov provides information and data on all spending by the federal government. Politicians and news agencies make claims about spending behaviors and how it impacts US's debt and revenue. These claims change public opinion about different matters, and the change in public opinion leads to administration changes on many levels. As claims change public opinion, changes in public opinions also change policies and future spending. This change can be cyclic.

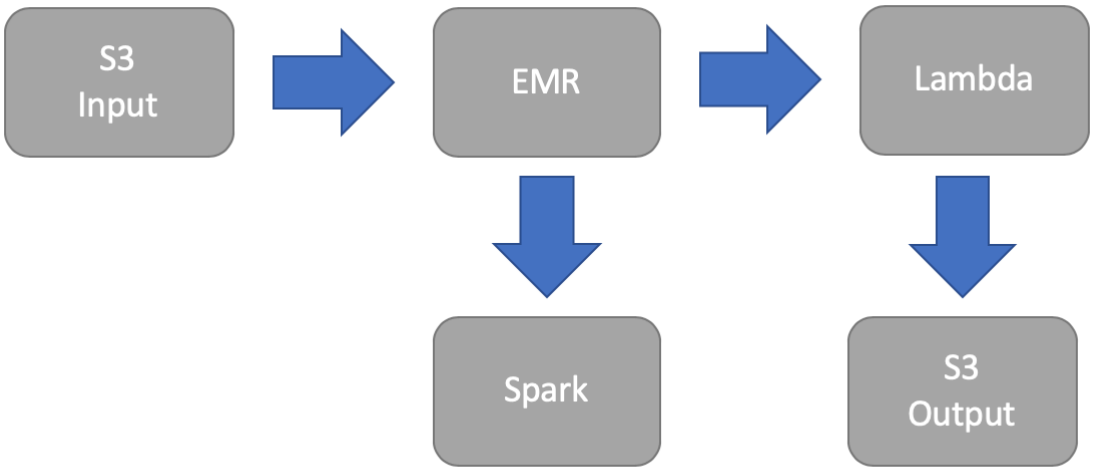
Our group wants to explore the data and extract information to communicate different spending trends along the years in accurate and simple manners. Additionally, the problem is very intriguing because there are numerous and various factors that impact spending. They include:

- Change of administration every 4-8 years
- World trends such as the Oil Crisis and the Financial Crisis in 2008
- Public opinions and reactions
- Debt

Therefore, we are extremely intrigued in understanding spending trends and trying to understand underlying reasons behind changes. Additionally, we want to communicate our findings through visualization the information in compact, simple, and interpretable manner with our peers and instructors.

good

High-Level Solution Architecture



1. S3 (input): We will use the s3 bucket that is publicly available to access the dataset. The data is stored in Amazon and can be accessed through Amazon Relational Database Service (RDS).
2. EMR cluster: Will be used to run spark and access the s3 input bucket easily.
    a. Spark - Three layers:
        i. Data extraction: The data will be extracted, processed and filtered by running computations in parallel so the execution becomes faster.
        ii. MapReduce: Might be used multiple times to count phrases to understand trends and patterns.
        iii. Data aggregation/analysis: once we compile different metrics and trends from MapReduce jobs, we will proceed to aggregate different results together to output in a spreadsheet and perform further analysis to visualize and communicate final results in our report.
3. Lambda function: After applying multiple spark functions on the data, we will apply lambda for row-individual-functions.
4. S3 (output): Finally, the output will be stored on an s3 bucket. Outputs can vary, they can be large text files, or a few CSV files. Text files might contain outputs from MapReduce jobs, whereas CSV files will include data rows with attributes of the final results we want to communicate and share with the class and in our final report.

Dataset

a) Dataset Used

There are various datasets provided through the website and hosted on AWS. Currently, we will rely two datasets provided by USAspending.gov which are:

1. Federal Accounts: Account Balances
2. Treasury Accounts: Account Balances

datasets can be messy, so please try to examine these thoroughly ASAP

There are numerous and various datasets hosted and provided by USAspending.gov, however, we are still exploring datasets that we can use for our project and applicable with our tools. For instance, we accessed a dataset that contained two or three entries per line, which are

1. An index
2. A description of the transaction or account, written in a paragraph
3. A URL or special notes

This particular dataset is not applicable to aggregate, nor can we communicate information that is standardized among all accounts and data lines. One possible approach that we might consider is a word count, but rather than count all the words, we might count phrases and understand what kind of note is frequent among the dataset and communicate it if it is useful and applicable to our solutions for the proposed problem.

b) Data Description

Format

The website has many datasets that can be downloaded by specifying the following:

1. Award level
2. Award type
3. Agency
4. Time range
5. File format (csv, tsv, xml)

need to try to get a more concrete hold on this

Preprocessing

The preprocessing of the data depends on what kind of trends and patterns we could extract. For instance, if we want to analyze what kind of budget function each federal account is spending on, we would preprocess the data such that these attributes would be used, and the rest would be excluded: federal_account_name, agency_name, budget_function.

Streaming/Static, Accessing and Storing Data

The data is available for download from 2001 until present and it is updated every day. We will access the data through the AWS S3 bucket/Amazon RDS that is publicly available and will store the final processed data on another AWS S3 bucket. The data is updated every night, while older data is static, there is a stream of new data at the beginning of each day.

Challenges

Communicating and visualizing spending habits of a huge country like the US can be very difficult. First, as computer science students, we lack the technical background to understand financial spreadsheets, and therefore, we anticipate learning how to read the available spreadsheets and the information the data is conveying. Secondly, as international students, we are unfamiliar with some of the institutions and their roles. We want to communicate interesting and unique spending habits and trends, hence, we will need to understand the different departments included in the datasets.

The challenges we are anticipating include:

1) There are thousands and thousands of listed accounts, aggregating spending data for each account and communicating how much each account spends can be time consuming, especially given the size of the data.
2) We need to understand each account and the background information related to the account and determine which accounts we need to focus on based on importance.
3) There are many attributes included in the data. We need to understand each dimension, and how it impacts our final results. For instance, the difference between 'budget authority unobligated balance brought forward' versus 'adjustments to unobligated balance brought forward'
4) The 40 GB and more size of the data provided offers various datasets. We will spend some time understanding each dataset and if we need to pair and combine various datasets together to communicate the information we gain and learn during the project.

Timeline

To ensure we stay on track to meet and completely fulfill the project's expectations, we aim to meet weekly milestones that we believe are ought to be met each week. Below, is a weekly timeline, the dates, and the task to be completed.

Week	Date	Complete Task by End of Week
Week 1	10/29 – 11/04	Complete Proposal, Access Datasets through AWS
Week 2	11/05 – 11/11	Narrow on datasets related to the problem only. Perform ETL
Week 3	11/12 – 11/18	Data Aggregation using Spark & EMR
Week 4	11/19 – 11/25	Thanksgiving
Week 6	11/26 – 12/02	Data Summary, Graphs, Trends
Week 7	12/03 – 12/09	Prepare Final Report and Presentation

project seems fine, but it will be very challenging to navigate and analyze the large dataset. you may consider tracking only certain budgets (military, science, etc). in other words: try to determine a subset of the data that is more reasonable.