

# Checkpoint 2: US Federal Spending

Ahmed Al Hasani, Mohamed Al-Rasbi

## Proposal Changes

We made and will make several changes moving forward with the project. We are using different tools to extract information from our datasets, we will add a new dataset, and we will extract information and trends from the dataset to answer questions that will guide our efforts and verify if our assumptions are valid.

First, we tried switching over to PySpark to use the data aggregation tools it provides. We attempted to load our CSV files, but we ran into an error that detailed “Hive is not Instantiated”. We tried loading the CSV files using different commands, such as SQLContext, or SparkContext, but no matter the approach, we faced the same issue. We managed to solve the issue by loading a single CSV file in one Spark Context object as soon as the notebook loads, however, if we try to load another CSV file, we will have the same error. As a result, we decided to use Pandas, because it can conveniently and easily add numerous CSV files, store them in a list of dataframes. talk about this transition in final project report

Secondly, we initially proposed finding trends and valuable information to share in our final submission regarding US’s spending, however, our efforts were not guided. We were randomly exploring trends without purpose. As a result, we decided to formulate questions that reflect a genuine interest from us in learning about spending habits that can be learned and concluded from the datasets. Additionally, we will conduct a brief research that will help us predict what the answer should be to each question. Finally, the datasets will verify whether our assumptions are correct. good

The questions we asked are the following:

1. Predict spending habits in Colorado and Arizona in different areas (health, education) based on priorities set for the nation and state by officials in 2010. Verify whether the actual spending habits met those predictions.
2. Extract spending habits in both states using other datasets from the actual departments and compare them with the predicted results and the results from the initial datasets.
3. What are the spending habits of California, Oregon, Colorado and Arizona based on their geography and population?
4. What are the similarities and differences between California and Oregon and Colorado and Arizona based on their spending habits?

Thirdly, Professor Rozner suggested we find other datasets that will address questions specifically and support our conclusions. We found data related to the Department of Education’s spending across multiple states.

great

## Meetings

We meet three to four times a week because we work in the same lab and we have another class project as a result we discuss the project during each meeting. We meet to specifically work on this project twice a week.

## Timeline

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	Answer questions
Week 8	12/10 – 12/16	Prepare Final Report and Presentation

We are on track to meet week seven's tasks, we started answering the questions we set for the project and they can be reflected in our notebooks that are uploaded in our GitHub repository. We did not yet find the appropriate datasets for verification purposes that we will use to compare with our results, but we hope by the beginning of week 8 to answer each question in detail and have verify whether our predictions and answers hold. We also divided the questions among us so that each person can have goals to meet each week.



Uploaded Ahmed's jupyter notebook for checkpoint 2.

### 1) Ahmed

I am answering the first two questions which will verify if spending habits are according to promises and predictions from politicians and other indicators in 2010. Because the datasets we have show spending habits from 2010 to 2018, we will base of our assumptions to how spending habits should be based on the vision set nationally and at the same state level.

Currently, I explored the following:

- The most frequent award types given in Colorado and Arizona in each year

- The agency that gives out the most awards
- Bar graph that shows the three most agencies that spend money in each state and year.
- A line plot for each agency in each state that shows how much they spent in each year.

Commits on Dec 4, 2018

Added Mohamed's notebook

mohdrasbi committed 2 minutes ago



cb7067f



Uploaded Mohamed's jupyter notebook for checkpoint 2.

## 2) Mohamed

Trying to answer question 3 and 4. Currently, I am comparing the 4 states (Colorado, Arizona, California, Oregon) with each other and try to find correlations between the data and states characteristics.

I explored the following:

- A comparison of total funding between states
- A comparison of unique awards given between states
- A comparison of total awards given between states
- Create histograms of amounts of fundings in each state

And I was able to draw these conclusions:

- The bigger the population, the bigger the total funding.
- Number of unique and total awards given can be affected by the population, but not necessarily.
- The histograms show the amounts of awards given in different states, in different years. They also show the percentage.
- We can see that Colorado and Arizona have higher percentage compared to Arizona and California. The histograms I created show the frequency of the amounts of awards given (0 to 10000 dollars) and the percentage to the total awards.
- I think we need more data to draw a solid conclusion about the amounts of awards given and their relation to population and geography of states.

## Project Cost

The project did not cost us anything so far, because it is not hosted on AWS anymore and in order to get data we need to download it directly from the website. We do not believe we need to add our CSV files to S3 buckets anymore because there isn't value in that anymore.

## Dataset Management

As mentioned earlier, we were not able to use AWS RDS to access the dataset, so instead, we relied on the website to download datasets we need. The website is:

[https://www.usaspending.gov/#/download\\_center/custom\\_award\\_data](https://www.usaspending.gov/#/download_center/custom_award_data).

Datasets are mostly clean, there are empty columns and entries throughout the datasets, but we do not need to convert any columns to cleaner ones that are stripped from any errors. All data is correct and does not contain string characters out of place. They are easy to understand which minimized our ETL processes to data aggregation and pattern exploration mainly.

## **Challenges**

The challenges we mainly faced were the changes to data hosting which forced us to change the tools we wanted to initially utilize to collect the data and extract information. While data access is easier now, we were initially excited to use the tools we learned throughout the semester and apply them in a meaningful manner. In fact, we deliberately choose this dataset because it is huge, hosted on AWS, and will require many other tools in our efforts to meet our goals. When we spoke to Professor Rozner, due to the little time we had in the semester, it was best to continue forward as planned and adjust accordingly.

Other minor challenges include validating our datasets. We believe the datasets may not contain accurate information in the first two years when they were available. We suspect because during the earlier efforts when they were collecting data, some awards were not inclusive, and hence, it could be the reason why spending in the first two years differ significantly from the rest of the datasets. We will address this issue by using another dataset from another source and compare the results.

good workarounds and  
progress  
10/10

# Checkpoint 1: US Federal Spending

*Ahmed Al Hasani, Mohamed Al-Rasbi*

## **Proposal Changes**

The majority of the initial proposal is intact, the only change we are making is how we will access the dataset. Due to an unforeseen change, the DB is not available on AWS RDS anymore, instead, to work with the dataset or access it, we are required to either download the database completely, which is roughly 40 GB, or search and download smaller datasets from the website. The current method in accessing the dataset will be discussed in the Dataset Management section.

Additionally, due to the size of the dataset, we are currently exploring trends in Colorado only and expanding our search for trends to different states (e.g., California, Colorado, Kansas) to compare how awards vary between each state and the reasons behind the variations given different traits and characteristics of these states.

The rest of our initial proposal is the same. We added and modified sections per the feedback we received about the types of trends we want to find, and it is highlighted in yellow.

## Timeline

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

The colored texts include the weeks and tasks completed. The only tasks we did not achieve yet is using Spark and EMR to aggregate data because of the database and datasets issue we faced. We will discuss the work done by each member in the subsections below, but we did not include a GitHub screenshot, instead, we decided it is more efficient and effective to separate the tasks among ourselves and that each member will develop their own scripts and upload them separately to our GitHub page, so that it is clear what each of us completed so far.


### 1) Ahmed

Commits on Nov 15, 2018

Add files via upload

AhmedAlHasani committed 19 hours ago

Verified

 512873f



I extracted information from the original dataset and created new dataframes each with new information containing trends or patterns explored. The dataset I focused on is *All Agency Awards Colorado 2018*. The patterns I explored and printed in my notebook are:

- Counted the top values in each category (according to how many times they received awards):
  - Award Description.
  - Counties and cities in Colorado.

- c. Business types in Colorado.
2. Most frequent awards given in Colorado and mapped them with their total obligations.
3. Counties and cities in Colorado that received the highest obligations.
4. The business types in Colorado with the highest obligations.

Future work and exploration include exploring the same patterns across multiple years, going back to 2010, categorizing awards (e.g. Education, Health), and comparing them with two other states that differ from Colorado.

## 2) Mohamed

Commits on Nov 14, 2018



Created two plots from which we can get an idea about the dataset and could be a start to finding other patterns: (1.png and 2.png in initial\_findings directory)

1. The first plot shows the number of awards every fiscal year (2010-2018) that was granted by each agency. Each subplot shows the number of awards in a certain FY. After looking at this plot, we can get an idea of how many awards each agency granted to recipients throughout the years.
2. The second plot shows the total funding every fiscal year for each agency. We can use this plot with the first one to make a relationship between the number of awards and the total funding. The total funding attribute in the dataset is interesting because the values could be negative. We need to pay more attention to that in the future and make sure what exactly a negative value means in this context.

Future work will include creating plots for other states (for comparison), finding a relationship between total funding and the number of awards for each agency, and use some of the tools we used in class to help solve these problems.

## 3) Future Work

Per the timeline, by week 6, we will aggregate all these patterns and across all years and states we are interested each of us is exploring and summarize them.

## Project Cost

We did not utilize AWS so far because we are able to download smaller datasets from USA Spending's website directly. We suspect that we will add all the downloaded datasets eventually to an S3 bucket and utilize AWS EMR to run Spark jobs on these datasets as they grow with time.

## **Dataset Management**

As mentioned earlier, we were not able to use AWS RDS to access the dataset, so instead, we relied on the website to download datasets we need. The website is:

[https://www.usaspending.gov/#/download\\_center/custom\\_award\\_data](https://www.usaspending.gov/#/download_center/custom_award_data).

In regards to data processing, the data are not noisy, because we restricted our work to datasets that are in .csv files only as opposed to .dat files which mainly contain long text paragraphs and indices and are not helpful for our project. Additionally, our datasets are very clean and easy to understand which minimized our ETL processes to data aggregation and pattern exploration mainly as explained in the Timeline section in more details and we did not have to clean our data.

# 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](http://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.

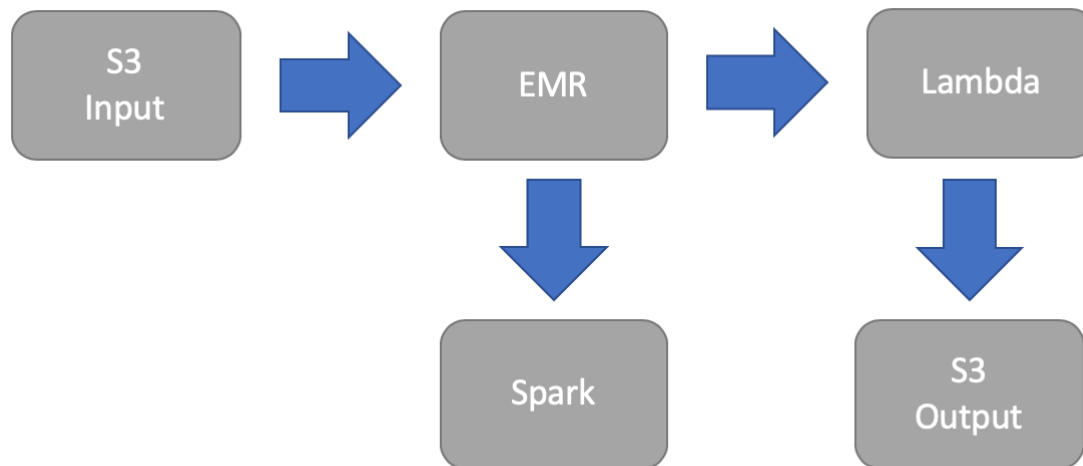
Our group wants to explore the data and extract information to communicate different spending trends along the years in accurate and simple manners. Additionally, the problem is very intriguing because there are numerous and various factors that impact spending. They include:

- 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.



## 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

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)

#### **Preprocessing**

The preprocessing of the data depends on what kind of trends and patterns we could extract.

We want to analyze and collect information on how much each state spends, the description of that spending, the county/cities that receive the most awards, and to plot these trends over the years as far back as the data allows. Additionally, 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.

<b>Week</b>	<b>Date</b>	<b>Complete Task by End of Week</b>
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