Solution Design Document

Project 3: USA Spending

Team Lead: Jacob Janz

Team Members: Andrew Vallerie, Bertrand Muhonga, Danni Hu, Jacob Janz, Jayson Smith, Paul Spadaccini, Sean McLaughlin

Created On: 15 January, 2024

Last Modified On: 19 January 2024

Contents

Solution Overview

This solution aims to utilize the data provided through usaspending.gov as an implementation of the DATA act to generate insights into US government spending for the US Department of the Treasury. Key insights to be explored include differences between initial estimates and actual obligation amounts, a breakdown of government spending by the locations of recipients, major agencies involved in government spending, and major recipients of federal obligations. Benefits of this analysis include determining which agencies or recipients may be involved with overestimation or underestimation of planned federal obligations, an enhanced understanding of the main names and industries in government spending, and expanded knowledge of the types of locations most likely to receive federal obligations.

Existing Functionality

The proposed solution will leverage an existing Microsoft Azure subscription through which resources will be reserved. Within the subscription is a pre-existing workspace in Azure Synapse Analytics, as well as a Storage Account with containers representing medallion architecture in Azure Data Lake Storage Gen 2. The solution will also utilize .csv files of raw data provided by usaspending.gov as an implementation of the 2014 DATA act. This data includes Account Spending (all government spending) and Award Spending (a subset of account spending that includes only money the federal government has paid or promised to pay a non-federal recipient through financial assistance or a contract).

Requirements

(F) Functional requirements (actions and tasks the system must be able to perform)
(NF)Non-Functional requirements (performance, scalability, security, and usability)
(U) User requirements (expectations and needs from an end-user perspective)

Requirement Type (F, NF or U)	Description
F	ADLS with bronze, silver, gold architecture to store data from USA Spending 2019-2022 and 5 external datasets
F	Spark Notebooks to perform transformations on USA Spending dataset for bronze->silver and silver->gold that must be able to be automated in a pipeline
F	Dedicated SQL pool in Azure Synapse Analytics using a star or snowflake schema to enable reporting and analytical queries
F	Creation of pipelines in Azure Synapse Analytics and version control and collaboration via Github, pipelines must be reusable with data validation
F	Provide at least 2 dashboards for USA Spending data sets incorporating external datasets and providing interactivity and drill down
F	Final presentation with technical details and high-level summarizations
NF	Provide column level security for PII and row level security to prevent members of an agency from seeing other agencies information
NF	Scalable Spark clusters to be able to handle the transformations of the full dataset
U	As a user, I want to easily understand patterns in the USA's Spending from 2019-2022 via dashboards
U	As a user, I want this project to be completed on time and within the provided $800 budget

Assumptions and Prerequisites

Assumptions:

To ensure the successful execution of this initiative, it is imperative to focus on the following critical elements

- **Data Source Integrity and Accessibility**: The project capitalizes on the comprehensive data provided by USASpending.gov, encompassing Award and Account Data from 2019 to 2022, enriched with five supplementary external datasets. Paramount to this endeavor is the assurance of data quality and relevance, coupled with the necessity for consistent and reliable access to these key data sources.
- **Advanced Infrastructure and Strategic Design**: Anchored by the robust and dependable infrastructure of Microsoft Azure, the project is meticulously planned to include a data pipeline and reporting framework that are not only scalable but also highly adaptable. This strategic design is instrumental in accommodating future data integration needs and efficiently managing escalating data volumes.
- **Collaborative Team Dynamics and Expertise**: The cornerstone of this project's success lies in synergistic collaboration and seamless communication among team members. Harnessing their collective expertise is crucial in navigating the intricacies of the project and in achieving the envisioned goals. This collaborative spirit is vital in surmounting the project's challenges and steering it towards triumphant completion.

Prerequisites:

The successful execution of the project requires the establishment of several foundational elements

1. **Resource Group in Microsoft Azure:**
    - A specifically provisioned Resource Group within Microsoft Azure must be set up. This group will provide the necessary infrastructure and access based on user roles to ensure secure and efficient project progression.
    - Within this group, distinct access levels are to be allocated as follows:
        - **Data Engineers**: Granted full access to the data warehouse, enabling them to manage and manipulate the warehouse structure and contents.

        - **Data Analysts:** Limited to accessing only reporting data. This restriction ensures that data analysts focus on data interpretation without modifying the underlying data structure.

2. **Azure Services:**
   The following Azure services are critical for the project and should be included within the Resource Group:

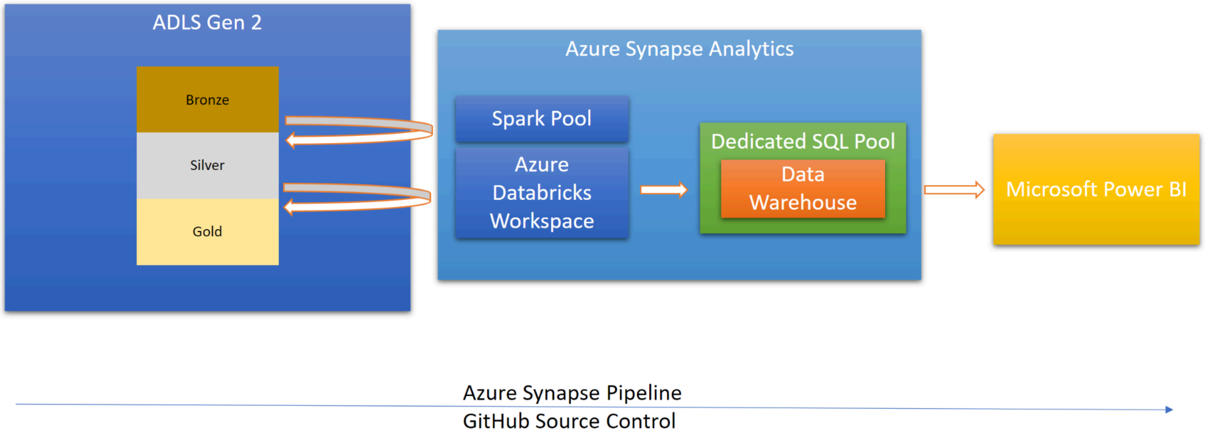
    - **Azure Data Lake Storage (ADLS) Gen2:** For robust data storage and management.
    - **Azure Synapse Analytics:** Serving as the primary platform for data warehousing and large-scale analytics.
    - **Azure Databricks:** Utilized for efficient and scalable data processing, especially for complex ETL operations.
    - **Power BI within the Power Platform:** For advanced data visualization and reporting, enabling the creation of insightful, interactive dashboards.

3. **Budget Allocation:**
   The project is allocated a budget of $800, with an additional $200 set aside as a contingency reserve. This budgetary allocation is designed to cover the costs associated with cloud services, potential unforeseen expenses, and other project-related expenditures.

4. **Data Access:**
   Access to the primary data source, USASpending.gov, is essential for the project. This website is the repository for the U.S. government's spending data, which forms the core dataset for analysis. Additionally, access to five external datasets is required. These datasets will augment the primary data from USASpending.gov, providing a more comprehensive view for analysis and decision-making.

High-Level Design

The high-level design orchestrates data extraction from Azure Data Lake, transformation via Azure Databricks and Synapse Analytics, and visualization using Power BI



The pipeline begins with an Azure Data Lake Storage Gen2 Container that holds directories for the bronze, silver, and gold layers. Transformations are performed on those layers using an Azure Databricks workspace and a Spark pool. The gold layer is stored in a data warehouse within an Azure Synapse Analytics Dedicated SQL pool. Dashboards are then generated from the analytical data using Microsoft Power BI.

As far as transformations are concerned, we will begin actually transforming our data from Bronze / Landing Zone files. We will be selecting a set of columns from each document type in the Bronze layer and transferring it to their respective folder in the Silver layer. From Silver, we will be constructing sets of dimensional tables that connect to a central transactions table. These final tables will be stored in our Gold layer. Included in these transformations, there will also be a cleaning process (high null counts, column formatting, altering data types, and de-duplicating information stored within our dimensional tables).

We must consider the design of our Synapse workspace as well as our data warehousing design. For Synapse, we will have a series of pipelines utilizing several methods of data transformation (Spark Notebooks, Staging Tables, etc.) as well as security features to keep personal data / protected information classified. In terms of data warehousing, our final Gold tables will be loaded into a dedicated SQL pool.

For our Power BI dashboards, we will set up a connection to our Synapse dedicated SQL pool tables and query the tables to create appealing and interactive visuals. We will represent the data and explore our dimensional tables as well as the external data that we incorporated to augment the dataset and the effects they have on US spending.

Low-Level Design

At a low level, We can get a better look into the actual technical work that will taking place as well as an Entity Relationship Diagram

The project will leverage the Azure Data lake Storage Gen2 hierarchical namespace, which optimizes files organizations into a structured hierarchy (folder or directory) for efficient data management and access. Complementing this, the data will be stored in the parquet file format, which reduces redundancy and boosts query performance.

The bronze layer will contain files for each year of required award archive data (2019-2022). Since both financial and contract awards will be analyzed, the bronze layer will contain folders for each type of input data, along with folders for the external datasets.

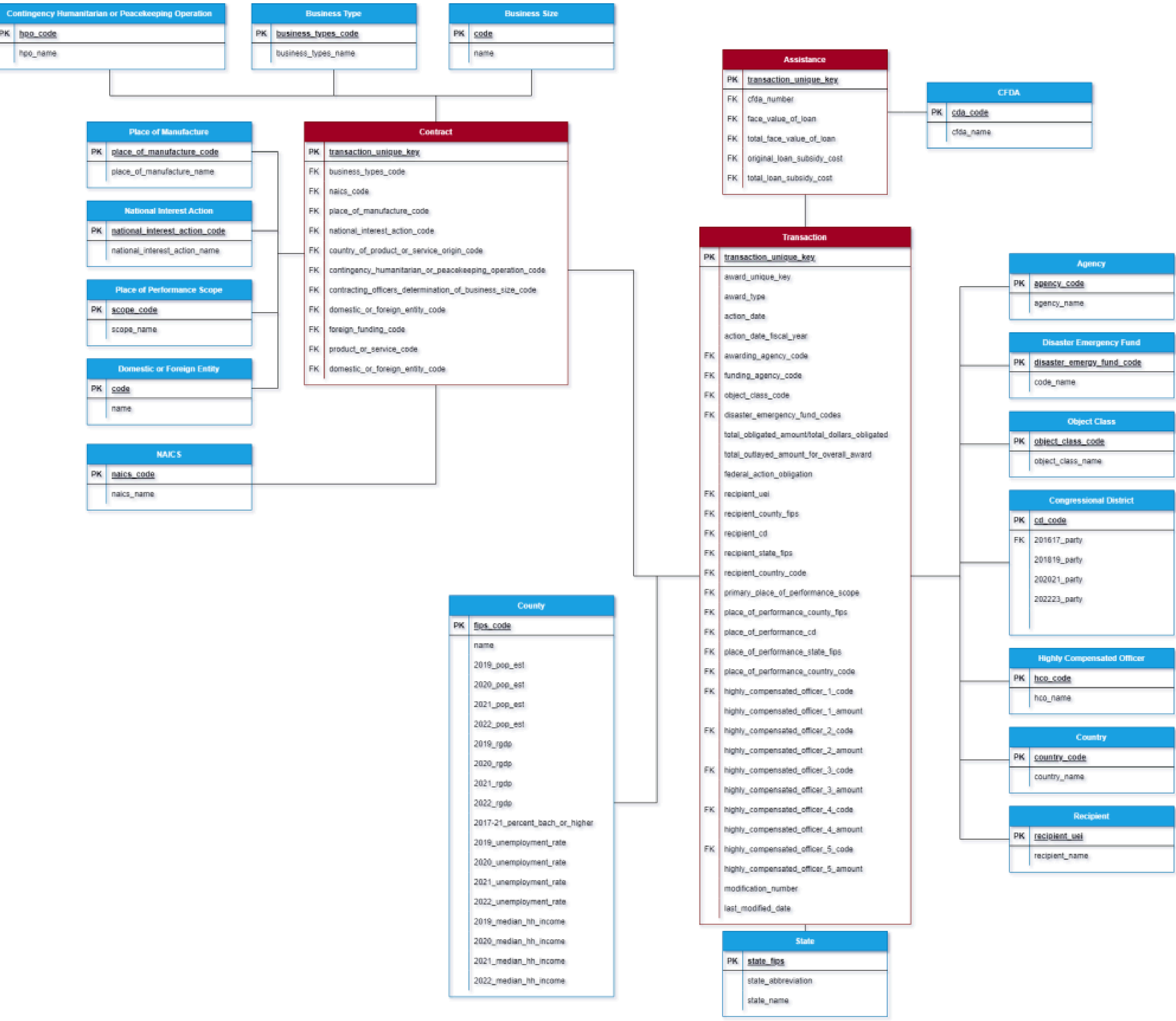
Afterwards, we will write a Spark Notebook that will have the ability to be automated and provide transformations to get both our contracts and financial assistance into our Silver layer. These transformations will include parsing out all of the columns that we plan to use, combining the data into one data frame, and providing necessary cleaning on our data. We will also undergo the same process for cleaning our external datasets and turning them into proper dimension tables ensuring there is a relationship established with the fact table.

Finally, for our Gold layer we will section out the dataset to have a modified star schema design as shown in our ERD. We will do so by extracting each set of columns into its own dataframe. The information common to both assistance and contract awards will remain in a generalized "award" table, while elements unique to either award type will be separated into their corresponding table. Tables will be created for dimensions such as Agency, Recipient, and Location information as well. Each table will correspond to one parquet file in the gold layer, which allows for easy loading into the data warehouse.

The execution of each transformation step will be automated with Azure Synapse Analytics pipelines that ensure that each step is modular and able to be run independently of the others, along with allowing for either ad-hoc or annually scheduled data ingestion. Pipeline parameters will be used to provide correct paths to ADLS.

At the end of the previous pipeline we will create SQL scripts that will automate the process of loading the Gold Layer data into our dedicated SQL pool. These scripts will include the insertion of the data using MERGE statements into already existing tables in our SQL pool. Upon completion of the entire pipeline, we will have the ability to incrementally load data into our data warehouse yearly.

Entity Relationship Diagram


Contingency Humanitarian or Peacekeeping Operation
PK hpo_code
hpo_name
Business Type
PK business_types_code
business_types_name
Business Size
PK code
name
Place of Manufacture
PK place_of_manufacture_code
place_of_manufacture_name
National Interest Action
PK national_interest_action_code
national_interest_action_name
Place of Performance Scope
PK scope_code
scope_name
Domestic or Foreign Entity
PK code
name
NAICS
PK naics_code
naics_name
Contract
PK transaction_unique_key
FK business_types_code
FK naics_code
FK place_of_manufacture_code
FK national_interest_action_code
FK country_of_product_or_service_origin_code
FK contingency_humanitarian_or_peacekeeping_operation_code
FK contracting_officers_determination_of_business_size_code
FK domestic_or_foreign_entity_code
FK foreign_funding_code
FK product_or_service_code
FK domestic_or_foreign_entity_code
Assistance
PK transaction_unique_key
FK cfda_number
FK face_value_of_loan
FK total_face_value_of_loan
FK original_loan_subsidy_cost
FK total_loan_subsidy_cost
CFDA
PK cda_code
cfda_name
Transaction
PK transaction_unique_key
award_unique_key
award_type
action_date
action_date_fiscal_year
FK awarding_agency_code
FK funding_agency_code
FK object_class_code
FK disaster_emergency_fund_codes
total_obligated_amount/total_dollars_obligated
total_outlayed_amount_for_overall_award
federal_action_obligation
FK recipient_uei
FK recipient_county_fips
FK recipient_cd
FK recipient_state_fips
FK recipient_country_code
FK primary_place_of_performance_scope
FK place_of_performance_county_fips
FK place_of_performance_cd
FK place_of_performance_state_fips
FK place_of_performance_country_code
FK highly_compensated_officer_1_code
highly_compensated_officer_1_amount
FK highly_compensated_officer_2_code
highly_compensated_officer_2_amount
FK highly_compensated_officer_3_code
highly_compensated_officer_3_amount
FK highly_compensated_officer_4_code
highly_compensated_officer_4_amount
FK highly_compensated_officer_5_code
highly_compensated_officer_5_amount
modification_number
last_modified_date
Agency
PK agency_code
agency_name
Disaster Emergency Fund
PK disaster_emergy_fund_code
code_name
Object Class
PK object_class_code
object_class_name
Congressional District
PK cd_code
FK 201617_party
201819_party
202021_party
202223_party
Highly Compensated Officer
PK hco_code
hco_name
Country
PK country_code
country_name
Recipient
PK recipient_uei
recipient_name
County
PK fips_code
name
2019_pop_est
2020_pop_est
2021_pop_est
2022_pop_est
2019_rgdp
2020_rgdp
2021_rgdp
2022_rgdp
2017-21_percent_bach_or_higher
2019_unemployment_rate
2020_unemployment_rate
2021_unemployment_rate
2022_unemployment_rate
2019_median_hh_income
2020_median_hh_income
2021_median_hh_income
2022_median_hh_income
State
PK state_fips
state_abbreviation
state_name

Out of Scope

To maintain a clear focus and ensure efficient utilization of resources, certain aspects and features fall outside the purview of this project

- **Real-Time Data Processing:** The project will not incorporate real-time data processing or streaming analytics. The focus will remain on the batch of data for the U.S. government's spending from fiscal years 2019 to 2022 and any external data sources used to augment that data.
- **Custom Analytical Tools:** The use of custom analytical tools or software outside of the Azure ecosystem is beyond scope. For successful completion of the project, only existing tools within the Azure ecosystem will be utilized.
- **Integration with the Company Systems:** Updating or integrating legacy systems to work with the developed infrastructure is not included. The project's scope is confined to the use of Microsoft Azure and its associated services.
- **Detailed training of the solution:** The project's solution will not include a detailed training documentation on how to use the developed infrastructure. The focus will be solely on the development and deployment of the pipeline and reporting tools.
- **Unlimited Scalability:** The project's scalability is constrained by budgetary limitations, especially when data volumes exceed initial projections.
- **Analytical Solutions:** The analytical data that will be loaded into our data warehouse will only lend itself to the analysis pertaining to our use case. It will consist of data from the given time period and any external datasets used to augment the analysis. From the USASpending data, it will include award and assistance information, their amount values, award descriptions, recipients and their respective information, and agencies.

Risks and Mitigation

The risks and mitigation section highlights potential challenges in project cost, scalability, data security, and technical issues

- **Cost:** Given the budget and contingency reserve allocated for the project, possible risks may include incurring additional charges based on the usage of computational power and memory in the transformation and load process. We will mitigate this by doing testing on mock/sample data before full transformations. We will also look to use scheduling triggers instead of event triggers to avoid running any unnecessary pipelines.

- **Scalability:** The risks in scalability consist of any newly loaded data that may fall out of bounds in size which would require additional resources to compute. We will mitigate this risk by scaling the workloads based on the prerequisite data rather than any new data of varying sizes.

- **Data security:** The risks in data security will consist of access permissions in the data warehouse and the visibility to sensitive information or information pertaining to certain agencies. We will mitigate these by assigning role permissions so that Data Engineers will have full access to the data warehouse and Data Analysts will only have access to the reporting data in the Gold Layer. Additionally, column level security will be applied for all sensitive information and agencies will only be able to see information relevant by implementing a security policy on the row level.

- **Technical challenges:** There may be risks in completing an end-to-end solution with a single service in the event of outages or failures to Azure. We will mitigate this by ensuring our storage redundancy will be enabled with Locally Redundant Storage? and by communicating with sponsors and appropriate contacts with our service providers.

Roles and Responsibilities

All team members will flow through each task together and when tasks can be completed in parallel, tasks will be given to individuals or groups to carry out on their own

Team Lead - Jacob
Permissions

- **Able to publish and execute pipelines in the Azure Synapse Analytics Workspace**
- **Able to create Dedicated SQL Pool and Server for 20231113DE Resource Group**
- **Able to create and manage Azure Databricks Cluster**
- **Manage RBAC and security groups for Data Lake and Dedicated SQL Pool**

Group 1- Infrastructure (synapse / data warehouses) - Sean and Jayson
Permissions-

- Maintain the Synapse environment and ensure the smooth functioning of data warehouses. They will work on optimizing the performance and reliability of the infrastructure to support the overall data processing needs.
- Create DDL scripts to properly integrate the data into the data warehouse with the correct schema and relationships.
-

Group 2- ETL Crew (Databricks / Pipelines) - Danni and Bertrand
Permissions-

- Creation of Apache Spark notebook using Azure Databricks to transform data from bronze to silver to gold layers. Notebooks should be able to be automated and pass in parameter values from Synapse Pipeline.
- Creation of pipeline using Azure Synapse Analytics that links to run notebooks to transform data throughout the medallion architecture and providing correct sinks.
-

Group 3- Analysis / Power BI - Andrew and Paul
Permissions-

- Leverage Power BI to create insightful reports and dashboards based on the transformed data. Their role involves interpreting the data in meaningful ways, providing valuable insights to stakeholders. Within those dashboards provide drillable visualizations and 3-5 KPIs each.
- Query the data as it is transformed and work with other teams to build proper tables containing information that is key in the final analysis and Power BI dashboards.

Schedule

Group ID	Task	Date / Duration	Goal at end of task
1, 2, 3	Write Design Document (Find External Datasets)	1/16 (Full)-1/17(First Half) (1.5 Days)	Finalize a Design Document, Go over Together, Submit
1, 2, 3	Preliminary Data Ingestion and Research	1/17(Second Half) (½ Day)	Pull in Data from USA Spending Document, Start Exploration
1 - Infra 2 - ETL (3 involved for Input)	Bronze -> Silver, Silver -> Gold	1/18 (Full)-1/19 (Full) (2 Days)	Start and Finish Pipelines (In Synapse / Spark Notebooks)
3 - Analysis	Begin Analysis of Silver / Gold	1/19 (Full) (1 Day)	Analyze data as it is created / transformed
1 - Infra 2 - ETL (3 involved for Input)	Pipeline and Application of External Data	1/22 (Full) (1 Day)	Cleanup and Apply External Datasets to Gold Tables
3 - Analysis (1 and 2 involved for input)	Analysis of Final Tables	1/23 (Full) (1 Day)	Queries, Augmentations, and Aggregations of Final Tables
3 - Analysis (1 and 2 involved for input)	Power BI / Slide Creation	1/24 (Full) (1 Day)	Creation of Dashboards, Slides for Presentation
1, 2, 3	Final Cleanup Day	1/25 (Full) (1 Day)	Running through Presentation, Any Errors to be cleaned, etc.

APPENDICES

Links:

- Data Source
- Data Dictionary
- USA Spending Guide

Potential External datasets:

- **USA County Population data**
    - https://www2.census.gov/programs-surveys/popest/datasets/
    - **data dictionary -** https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2020-2022/CO-EST2022-ALLDATA.pdf
- **USA Election Data**
    - https://electionlab.mit.edu/data
- **USA County Poverty, Unemployment, Household Income, Education Data**
    - https://www.ers.usda.gov/data-products/county-level-data-sets/