**Term Paper Group 1 CIS 4560**

Group 1: Alan Liu, Samuel Park, Steven Vallejo, Keith Wang

California State University, Los Angeles

**Abstract:** This document is Group 1’s CIS 4560 term paper. It examines the United States’ spending in the 2017-18 fiscal year. Our project used USAspending.gov’s datasets to help provide a transparent view on government-related spending. Also, to provide a thorough and organized visualization on government spending, in short, this project shows where the American tax payer’s money is going. USAspending.gov is the official source for spending data for the U.S. government. Data is uploaded directly from more than a hundred federal agencies’ financial systems. The first part of the report is the introduction. The second part of the report give the data background information. The third part of the report is the actual data analysis, including queries and geovisualization information.

**1.1 Introduction**

**Background:** Our group decided to examine the United States’ spending during the 2017-18 year. We downloaded .csv files from USAspending.gov for several types of spending including: contracts, awards, and assistance. Awards are money the federal government has promised to pay a recipient. It can and has been awarded to companies, organizations, government entities and even individuals. Contracts are agreements between the federal government and a recipient which will provide the government with some type of goods or service for a fee. Assistance or federal assistance are federal programs, service, or activity that directly aid organizations, individuals, or local governments.

**Goal:** Our goal for this project was to help provide a transparent view on government-related spending. We also want to provide a thorough and organized visualization on government spending.

**Data Source Information:** For our project the data size was ~1.7GB compressed; uncompressed the data was ~12GB. All data was downloaded from USAspending.gov’s data center. In total, seven files were downloaded with information about awards, assistants, and contracts.

**1.2 Data Background Information**

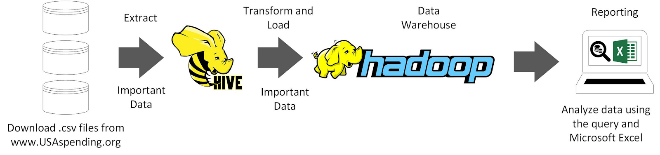
**Hadoop Version:** Hadoop 2.7.1.2.4.2.0-258

**Hive Version:** Hive 1.2.1000.2.4.2.0-258

**Hardware Experimental Specifications:** For this project our Oracle Big Data Compute Edition was 5 nodes. Oracle Compute Units (OCPUs) are defined as the CPU capacity equivalent of one physical core of an Intel Xeon processor with hyper threating enabled. For our project our OCPUs were 10. Memory was 150GB. Storage was 678GB. HDFS capacity 147GB and the CPUS Speed was 2.2 GHz. The Bash command: “lscpu” will show the CPU information, including the speed.

**ETL Architecture:** Figure 1.2.1 shows our project’s ETL Architecture.

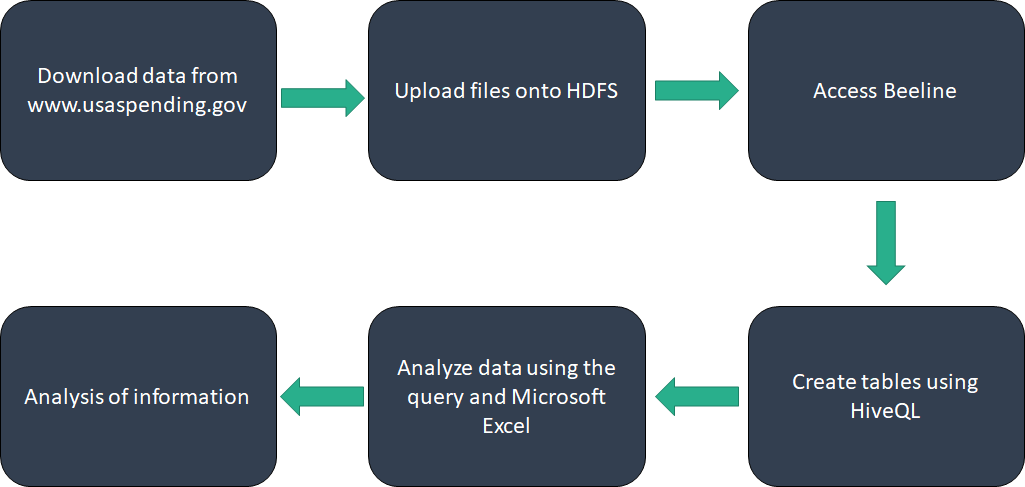
Figure 1.2.1



ETL stands for Extract, Transform, and Load. Extracting the data serves as off-loading data from the data source. Transforming and Load is when the data is integrated and moves the data to the presentation area to be accessed by the end users. First our .csv files were downloaded from USAspending.gov. Then they were uploaded to our Hadoop cluster. After that, the data was extracted by using Hive and then the important data was uploaded on top Hadoop. Finally the data is analyzed using the queries themselves and Microsoft Excel.

**Flow Chart of Data Analysis:** Figure 1.2.2 below shows our flow cart of data analysis.

Figure 1.2.2



The first step in our data analysis flow cart is to download the data from USAspending.gov. The second step is to upload files into HDFS. The third step is to access Beeline. The fourth step is to create tables using HiveQL. The fifth step is to analyze the data using the queries and Microsoft Excel. The last step is to analyze the information that is extracted.

**1.3 Data Analysis**

The following section will demonstrate the queries that were used to extract information from the raw data.

**Table Creation Queries:** Two tables were created before the data could be analyzed.

Figure 1.3.1

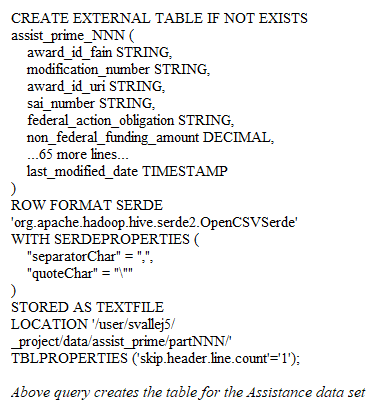
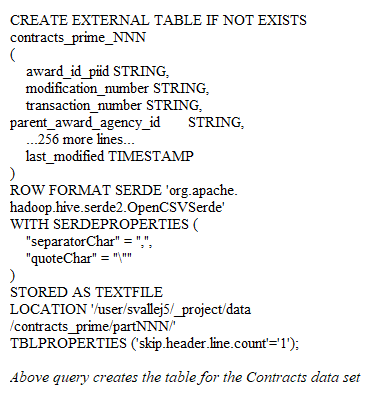


Figure 1.3.2



**Queries for Retrieving Information:** The following are a few of the queries that were used to retrieve information from data.

Figure 1.3.3

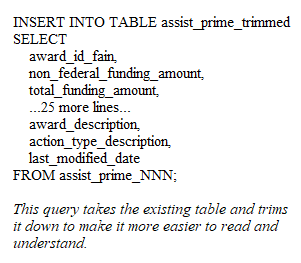


Figure 1.3.4

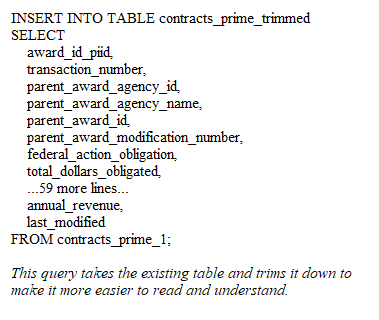


Figure 1.3.5

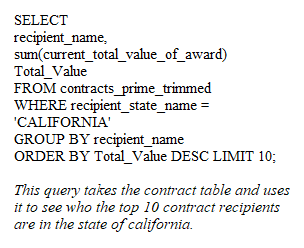


Figure 1.3.6

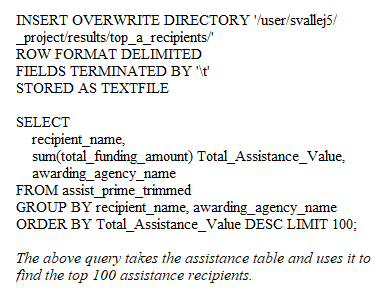


Figure 1.3.7

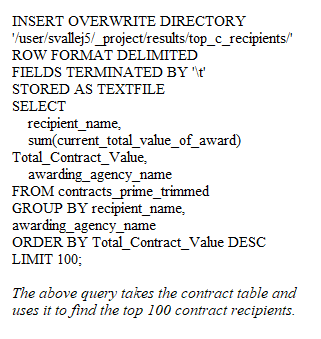


Figure 1.3.8

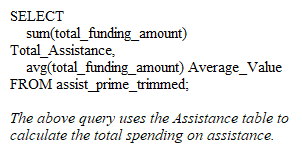
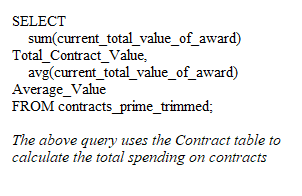


Figure 1.3.9



**Geovisualization:** This is short for geographic visual. This refers to a set of tools and techniques supporting the analysis of geospatial data by using interactive visualization. We used it for two, assistance data and contract data.

Figure 1.3.10

INSERT OVERWRITE DIRECTORY '/user/svallej5/\_project/results/assis\_trimmed/'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

SELECT

action\_date, total\_funding\_amount,

recipient\_name,

recipient\_address\_line\_1,

recipient\_city\_name, recipient\_state\_name,

awarding\_agency\_name,

assistance\_type\_description, award\_description

FROM assist\_prime\_trimmed;

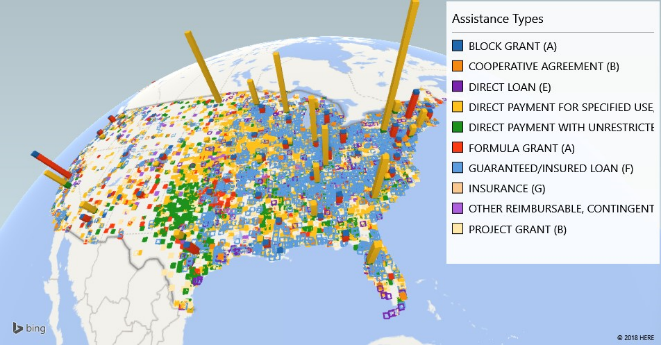


Figure 1.3.11

INSERT OVERWRITE DIRECTORY '/user/svallej5/\_project/results/contr\_trimmed/'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

SELECT

action\_date, current\_total\_value\_of\_award,

recipient\_name,

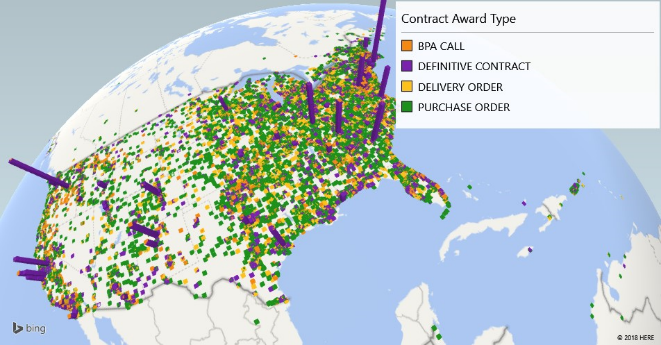
recipient\_address\_line\_1,

recipient\_city\_name, recipient\_state\_name,

awarding\_agency\_name,

award\_type, award\_description

FROM contracts\_prime\_trimmed;



**1.4 Conclusion**

**Conclusion:** Our project used USAspending.gov’s datasets to help provide a transparent view on government-related spending. Also, to provide a thorough and organized visualization on government spending, in short, this project shows where the American tax payer’s money is going. USAspending.gov is the official source for spending data for the U.S. government. Data is uploaded directly from more than a hundred federal agencies’ financial systems. The first part of the report is the introduction. The second part of the report give the data background information. The third part of the report is the actual data analysis, including queries and geovisualization information.