Penn State University

Great Valley Campus

Engineering Division

Data Specification for

Election Contributions

Getinet A. Gawo

gag28@psu.edu

Version 1.0

April 9, 2021

**Table of Contents**

[Introduction 3](#_Toc457330103)

[Purpose 3](#_Toc457330104)

[Project Summary 3](#_Toc457330105)

[Requirements Definition 4](#_Toc457330106)

[Considerations 4](#_Toc457330107)

[Document Change Log 4](#_Toc457330108)

[Architecture Design 5](#_Toc457330109)

[Entity Relationship Diagram 5](#_Toc457330110)

[Data Dictionary 5](#_Toc457330111)

[Tables schemas 5](#_Toc457330112)

[Examples of values 5](#_Toc457330113)

[ETL Process 6](#_Toc457330114)

[ETL considerations 6](#_Toc457330115)

[ETL Process Flow with description 6](#_Toc457330116)

[Reporting System 7](#_Toc457330117)

[Conclusions 7](#_Toc457330118)

# Introduction

In American democracy contributions are the most common source that is used to support election campaign. A contribution can be anything that has a value given to support candidates in elections. Monetary contribution is of the support method which is made by cash, check or credit card. The Federal Election Campaign Act puts restrictions on the amount and source of contribution. In addition to this, it sets the threshold that determines whether an individual has qualified as a candidate to receive contributions. The Federal Election Commission (FEC) is independent regulatory agency charged with administering and reinforcing the federal campaign finance law. One of FES’s mission is disclosure of funds raised and spent to influence federal elections.

Since FEC is the agency oversights finance of campaigns, the origin of data used in this study originates from the FEC. The study uses campaign contribution made at individual level, and it does not include those made at a company or organizational level.

# Purpose

The purpose of this study is to understand campaign contributions and identify any pattern or trend that exits in the past data.

# Project Summary

The objective, scope references and related issues for this study are listed below.

1. **Objectives**

* To investigate the trend in election contributions and make predictions.
* To understand the type of contributors and how they evolve over time.
* To understand geographical distribution of contributor with respect to the recipient.

1. **Scope**

Part of study of different systems in the Data Analytics program we want to compare different implementations of systems capable of delivering knowledge to the management. The scope of this project is to analyze the performance of two such systems with different architectures on the same data and same hardware.

The data used in this study are contributions made in January, February, and March of 2020. The scope is limited to the first three months of 2020. The data used in this study comes from contributions made at individual level from Florida, New York, Pennsylvania and Texas.

1. **References**

2020 Presidential Contribution Data by State, ProPublica, October 26, 2020, <https://projects.propublica.org/itemizer/presidential-contributors/2020>

Mission and history, Federal Election Commission, https://www.fec.gov/about/mission-and-history/

1. **Outstanding Issues**

No outstanding point this point.

# Requirements Definition

1. **Goals**

* deploy two systems, load data, and execute simple reports
* articulate the pros and cons of each system
* select better system architecture based on specifics of data to be analyzed

1. **Usability Requirements**

The schema is created using Postgre SQL. Data storage and retrieval to and from the data warehouse use Postgre SQL.

1. **System Security Requirements**

The end use can perform select, insert and update queries on the tables in the schema.

1. **Business Questions**

The data warehouse is used to be a source of data to investigate the trend in election contributions and make predictions, understand the type of contributors and how they evolve over time, understand geographical distribution of contributor with respect to the recipient. Which state contributes the most? How is the trend of the contribution? Which candidate collected the most?

1. **Data Requirements**

Data is provided to us as part of the project. It consists of several Excel files which may be semi-structured. Data is about political contributions in the 2020 election cycle

1. **Design Constraints**

# Considerations

Please add any other considerations, if any, related to the systems in the project such as:

* Postgre database is used.
* DBeaver is used as a user interface to interact with the database.
* [More will be added as we move to the 3rd and 4th part of this project]

# Document Change Log

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Change Date** | **Version** | **CR #** | **Change Description** | **Author and Organization** |
| 02/20/21 | 1.0 |  | Initial creation. | Getinet A. Gawo |
| 03/28/21 | 2.0 |  | Hadoop implementation | Getinet A. Gawo |

# 2. Architecture Design

## 2.1 Relational Data Warehouse

### Design and schema

Inspecting and understanding the data is a key to begin the data warehouse design. The original data contains 28 attributes about the contribution made. Since our business interest is the contribution made in individua basis, it happens that some columns contain same entries. In addition to this, some columns contain the same entry or empty. The attributes (columns) in the data are summarized in the following table. Rows of the same color have the same value in each field.

|  |  |  |
| --- | --- | --- |
|  | **Column Name** | **Description** |
| 1 | filing\_id | File number / Report ID |
| 2 | linenumber | the same as filing\_id |
| 3 | flag\_orgind | contribution made by. In our case all entries are Individual (IND) |
| 4 | org\_name |  |
| 5 | last\_name | last name of contributor |
| 6 | first\_name | first name of contributor |
| 7 | middle\_name | middle name of contributor |
| 8 | prefix | Prefix to the name |
| 9 | suffix | Suffix to the name |
| 10 | address\_one | street address |
| 11 | address\_two | apt |
| 12 | city | City |
| 13 | state | State |
| 14 | zip | ZIP code |
| 15 | employer | Employer |
| 16 | occupation | Occupation |
| 17 | amount | Contribution made. Negative amount shows a refund. |
| 18 | date | The date the contribution was made. |
| 19 | aggregate\_amount | payment made year-to-date |
| 20 | memo\_code | Memo Code |
| 21 | memo\_text | Memo Text |
| 22 | tran\_id | Transaction ID |
| 23 | back\_ref\_tran\_id | The same as transaction Id with 'E" appended |
| 24 | back\_ref\_sched\_name | either the same as filing\_id or empty |
| 25 | prigen | the campaign. For example, primary 2020 |
| 26 | cycle |  |
| 27 | fecid | ID of the recipient |
| 28 | committee\_name | Name of the recipient |

**Declaring the grain:** In this study we are interested in contribution information made towards each candidate by state.

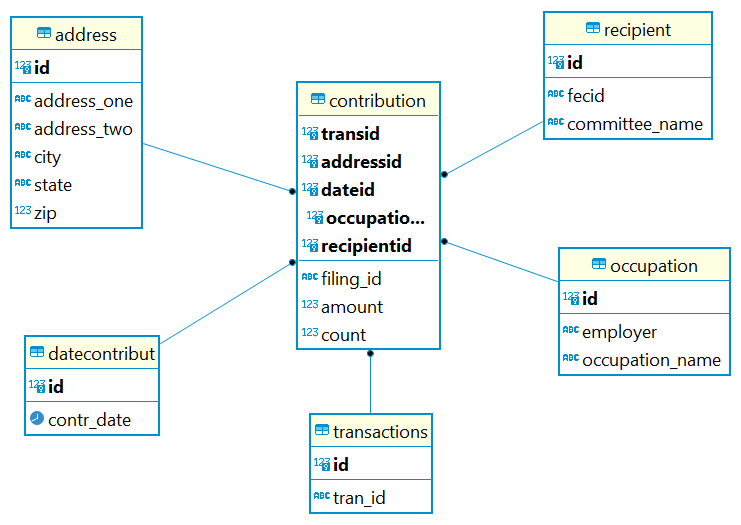
From our business objective point of view, we eliminate one of the repeated columns and columns that do not have importance. Consequently, we keep 14 attributes that has business importance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Column Name** | **Description** | **Null** | **Fact/Dimension** |
| 1 | address\_one | street address | Y | Dimension |
| 2 | address\_two | apt | Y | Dimension |
| 3 | city | City | Y | Dimension |
| 4 | state | State | Y | Dimension |
| 5 | zip | ZIP code | Y | Dimension |
| 6 | employer | Employer | Y | Dimension |
| 7 | occupation | Occupation | Y | Dimension |
| 8 | filing\_id | File number / Report ID | N |  |
| 9 | amount | amount made.  Negative amounts are refunds. | Y | Fact |
| 10 | aggregate\_amount | payment made year-to-date |  | Fact |
| 11 | tran\_id | Transaction ID | N |  |
| 12 | date | date the contribution is made |  | Dimension |
| 13 | fecid | ID of the recipient | N | Dimension |
| 14 | committee\_name | Name of the recipient | N | Dimension |

We create a star schema based on these 14 attributes listed in the above table. Based on this list of attributes we create four dimensions and one fact table. Inspecting the data reveals that a person may make contributions to the same candidate on the same date. The only attributes used to differentiate such two contributions is the transaction ID. For this reason, we will add one dimension that contains transaction Id’s.

These dimensions are Transaction, Date, Recipient, Occupation, Address.

### Tables schemas



The schema will have four dimension and one fact table, and their descriptions are presented as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| ***transaction*** |  |  |  |
| **Description** | The date the payment is made. | | |
| **Attribute** | **Description** | **Type** | Examples of values |
| **Id** | Id of a date | serial | Between 1 and 999999999 |
| **Tran\_id** | Transaction ID | VARCHAR | 5621554 or AE49E5476AC3746F0AF9 |
| **Primary Key** | id | | |
| **Candidate Keys** (if any) | NA | | |
| **Foreign Keys** | NA | | |
| ***dateContribut*** |  |  |  |
| **Description** | The date the payment is made. | | |
| **Attribute** | **Description** | **Type** | Examples of values |
| **Id** | Id of a date | serial | Between 1 and 999999999 |
| **Contr\_date** | Date | Date | 01/01/2020 |
| **Primary Key** | id | | |
| **Candidate Keys** (if any) | NA | | |
| **Foreign Keys** | NA | | |

|  |  |  |  |
| --- | --- | --- | --- |
| ***Recipient*** |  |  |  |
| **Description** | Information of recipient of the contribution. | | |
| **Attribute** | **Description** | **Type** | Examples of values |
| **Id** | Id | serial | Between 1 and 999999999 |
| **fecid** | Federal Election Commission ID | VARCHAR | C00698258 |
| **committee\_name** | Name of the recipient (candidate or committee) | Text | Kamala Harris For The People |
| **Primary Key** | id | | |
| **Candidate Keys** (if any) | **fecid** | | |
| **Foreign Keys** | NA | | |

|  |  |  |  |
| --- | --- | --- | --- |
| ***Occupation*** |  |  |  |
| **Description** | Employment status and employer of contributor. | | |
| **Attribute** | **Description** | **Type** | Examples of values |
| **Id** | Id | serial | Between 1 and 999999999 |
| **employer** | Employer name | VARCHAR | Penn State |
| **Occupation\_name** | Work title | Text | Professor |
| **Primary Key** | id | | |
| **Candidate Keys** (if any) | NA | | |
| **Foreign Keys** | NA | | |

|  |  |  |  |
| --- | --- | --- | --- |
| ***Address*** |  |  |  |
| **Description** | Address of contributor. | | |
| **Attribute** | **Description** | **Type** | Examples of values |
| **Id** | Id | serial | Between 1 and 999999999 |
| **address\_one** | Street name | VARCHAR | 1234 George Ave |
| **address\_two** | Apt or suit numbers | VARCHAR | Apt 102B |
| **city** | Name of the city | VARCHAR | New York |
| **state** | Abbreviation of a state | CHAR (2) | NY |
| **zip** | Zip code | Integer | 10032 |
| **Primary Key** | id | | |
| **Candidate Keys** (if any) | NA | | |
| **Foreign Keys** | NA | | |

|  |  |  |  |
| --- | --- | --- | --- |
| ***Contribution*** |  |  |  |
| **Description** | Address of contributor. | | |
| **Attribute** | **Description** | **Type** | Examples of values |
| **Id** | Id | serial | Between 1 and 999999999 |
| **filing\_id** | File number / Report ID | CHAR (8) | SA17A |
| **amount** | amount made.  Negative amounts are refunds. | FLOAT | 245.60 |
| **count** | Number of contribution | INTEGER |  |
| **transid** | Id of transactions table | inetger |  |
| **addressid** | Id of address table | Integer |  |
| **occupationid** | Id of occupation table | Integer |  |
| **recipientid** | Id of recipient table | Integer |  |
| **dateid** | Id of date table | Integer |  |
| **Primary Key** | **transid, addressed, occupationid, recipientid, dateid** | | |
| **Candidate Keys** (if any) | **tran\_id** | | |
| **Foreign Keys** | **addressid REFERENCES id in address table**  **occupationid id REFERENCES id in occupation table**  **recipientid id REFERENCES id in recipient table**  **dateid id REFERENCES id in datecontribut table**  **transid id REFERENCES id in transactions table** | | |

Based on the above specification and description, all tables have been created in the database. The SQL script used to create these tables is found in appendix. The corresponding entity relationship (ER) diagram of the schema is shown below.

## 2.2 Hadoop Implementation

First, the data under consideration is loaded on docker container. From docker container, it is loaded on HDFS. The Hadoop implementation starts by creating directories on HDFS to upload the data from four states for three months. We have a total of 12 data pieces that are all uploaded on Hadoop. These data are loaded in HDFS at /user/root/electionData/input. The script used to upload all the data is located in the appendix. It is confirmed that all the data is loaded on HDFS by checking the files in /user/root/electionData/input directory. The output of this is shown in the following image.

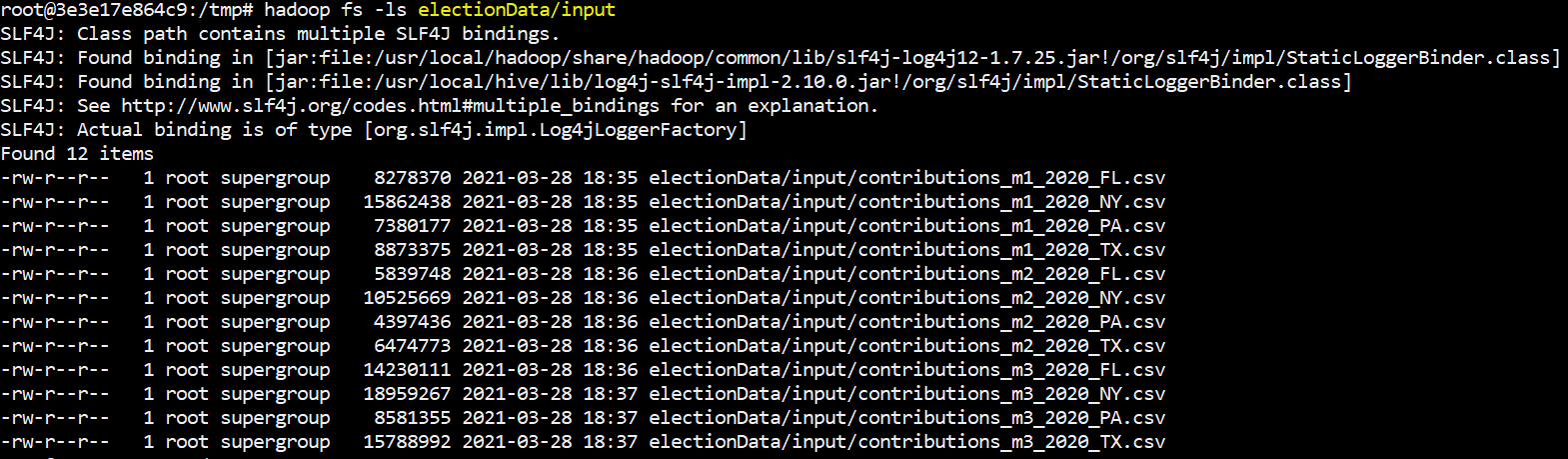


Figure 1: Data loaded on HDFS

## 2.3 Reflective analysis of using a data warehouse vs Hadoop.

The data loading process in Hadoop is very simple compared to the same process in the data warehouse. Hadoop is a very flexible tools to load data. In contrast to data warehousing, there is no schema that we have created to load the loaded. The data was easily loaded by providing the correct script in Hadoop.

3. Data Preparation

## 3.1 Relational Data Warehouse Implementation

## ETL considerations

The contributions reported in the month of January, February and March of 2020 are collected for the state of Florida, New York, Pennsylvania and Texas. There are 12 csv files, 3 for each state, are collected. The data coming from PA for January has date column that has unique date format from the rest of the group. Before it is loading workflow starts, the date column from this data is converted ass shown on the top box in figure 2.

## ETL Process Flow with description

**Loading Dimensions**

There are 12 csv files that contain the required data available to be loaded. There are three files for each state. For convenience, these files are located in the same folder. Loading each file on KNIME is time consuming as we are performing the same process again and again. For example, to load recipient of contributions from all 12 files, we load all the files and perform the aggregation. The same process repeated when loading the date and so on for all dimensions. In order to make the workflow efficient we use a loop. List Files node loads the name of locations csv files. Then Table Row to Variable Start node is used to make the file names variables. File Reader node is used to load the files. The Loop End node lets us loop the loading over the location of each file. In this way, all the 12 files are easily loaded to KNIME.

Once all data are loaded on KNIME, then extraction and aggregation is performed for each dimension. The date column is manipulated in such a way that the format agrees with the table data structure in the database.

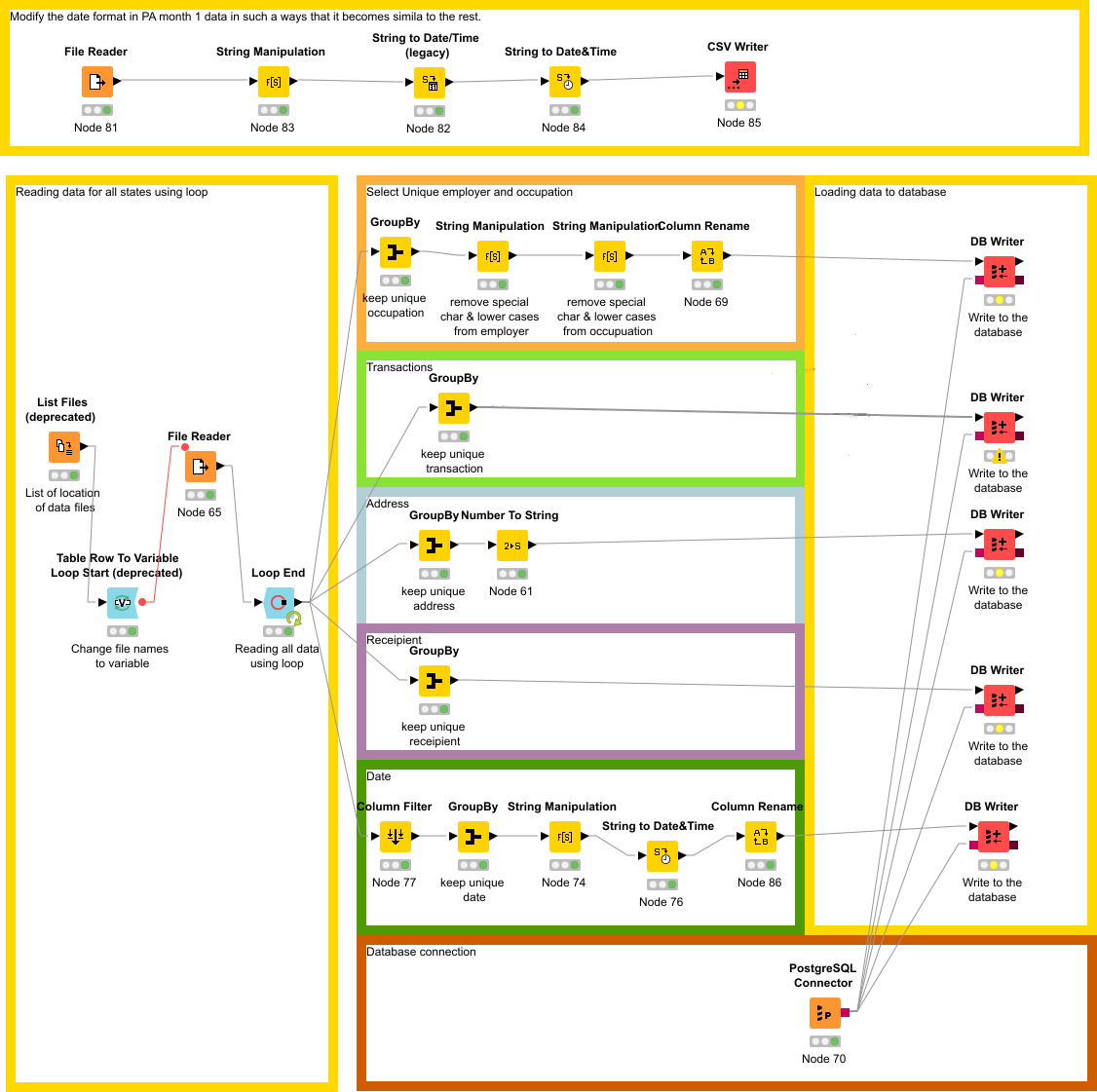


Figure : Workflow to load dimensions.

**Loading the fact table – Contribution**

Following the dimensions, the fact table is created based on its definition given in section 2.1.

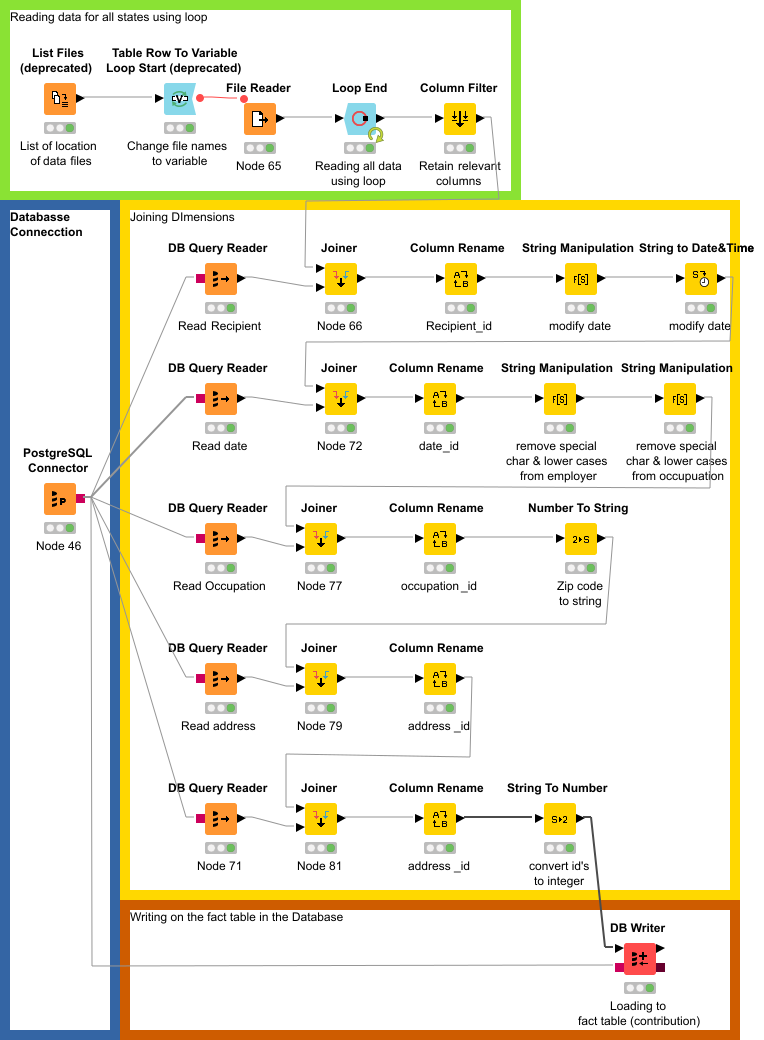


Figure : Workflow to load the fact table.

## 3.2 Hadoop Implementation

## 3.3 Reflective analysis of data preparation in relational data warehouse vs Hadoop.

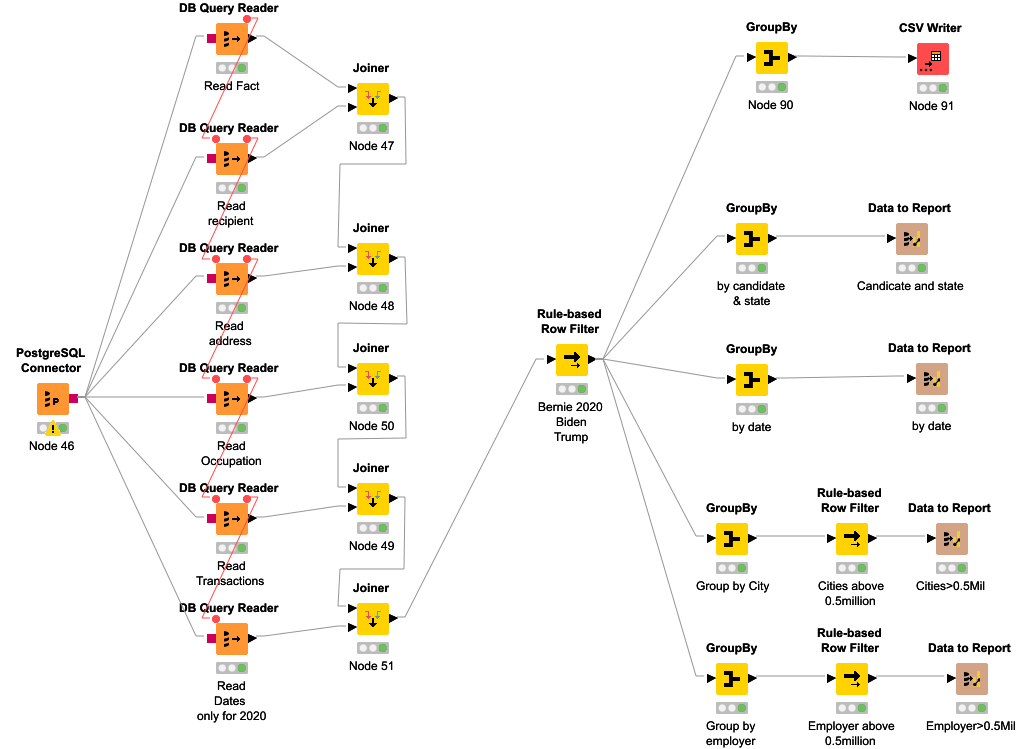
To be completed with project number 2

4. Reporting System

Provide screenshot and sample results with discussion on potential knowledge that was elicited.

## 4.1 Relational Data Warehouse Implementation

The data that is persisted in the database is imported to KNIME using the following workflow. For reporting purpose, contribution between January – March 2020 that are made towards Bernie 2020, Biden For President and Donald J.Trump For President, Inc. are presented in the reporting.



**Summary**

Contribution toward the three candidates is summarized in figure 4. Contributions made in Florida and Texas states favor republican candidate, while Pennsylvania and New York to democratic party. For each candidate, contribution made from New York is the highest, and that from PA is the least. Texas and Florida made almost the same number of contributions towards Donald Trump.

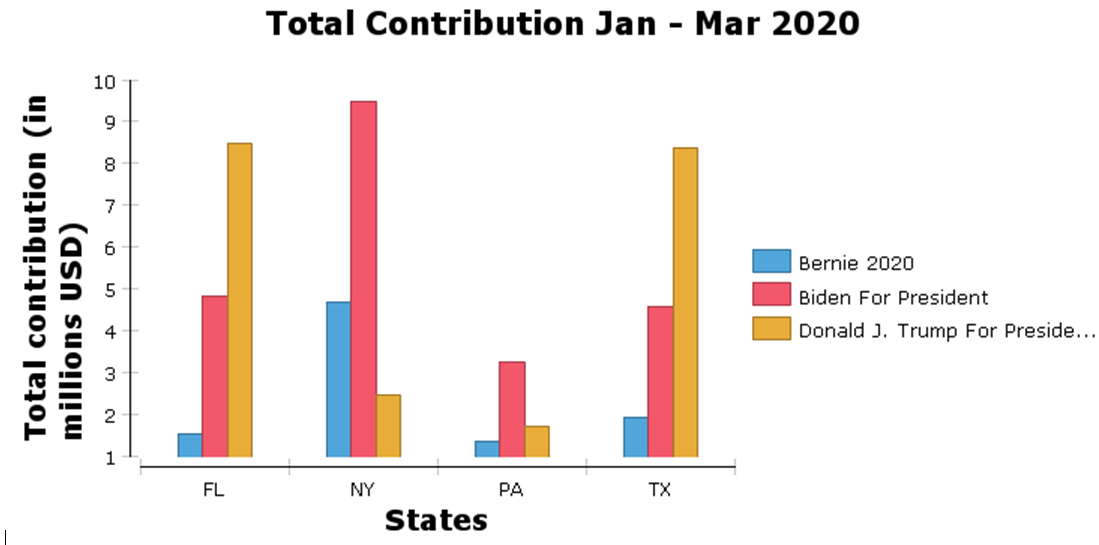


Figure : Contribution by State

Figure 5 and 6 shows the trend of contributions made between January and March of 2020. Spikes in contribution are observed at the end and beginning of a month. This is shown in figure 4 near end of January and beginning of March. Moreover, form the four states the highest contribution was made at the beginning of March. Contribution from PA is still the least and that from NY remains the highest.

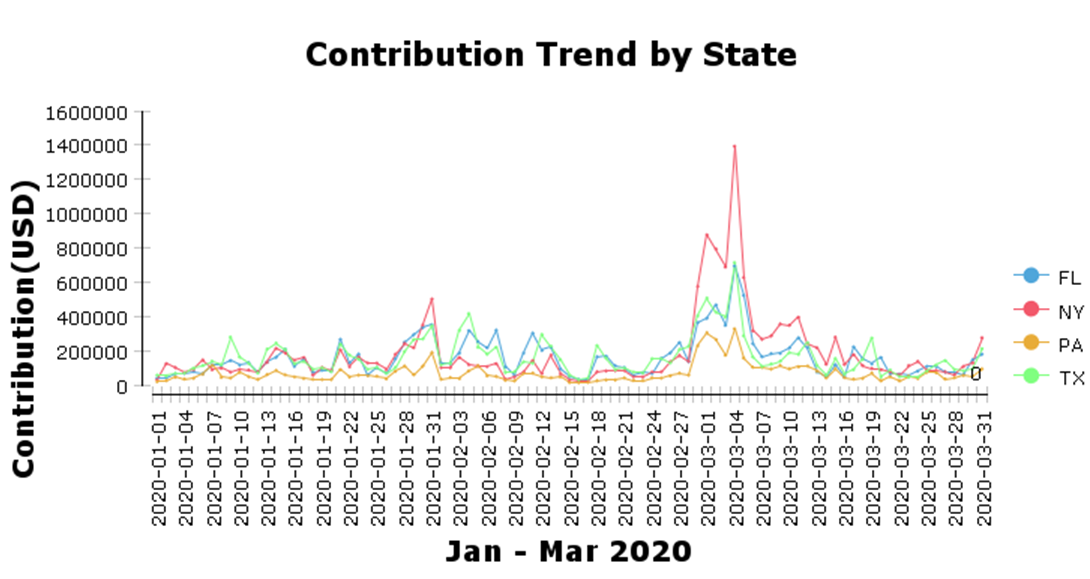


Figure : Contribution Trend by State

Contribution for Biden slowly started, but it spikes in the first two weeks of March. For the remaining days of the months the contributions remain the close for the three candidates.

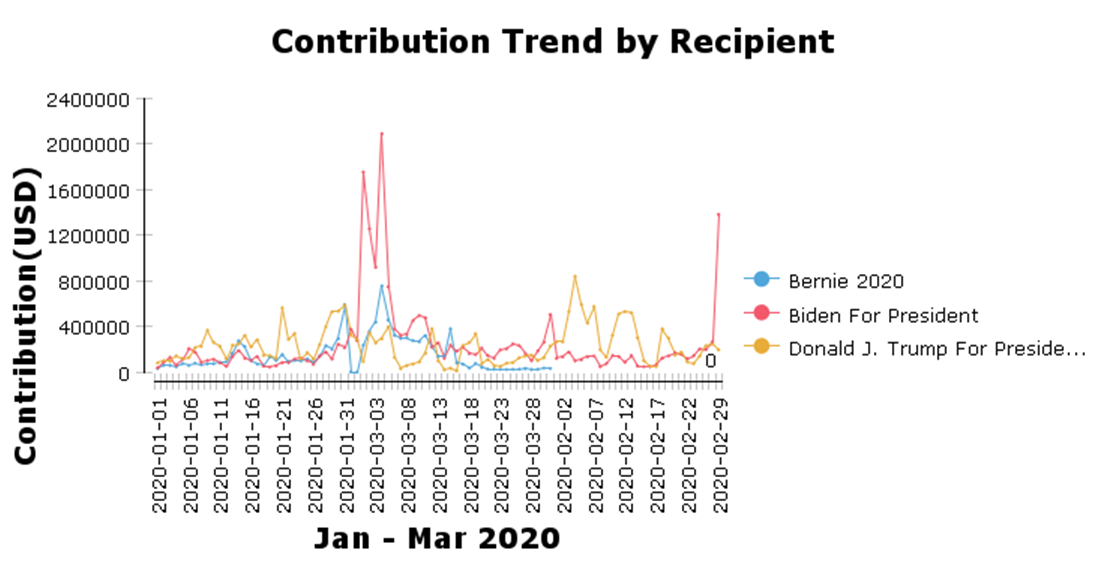


Figure : Contribution Trend by candidate.

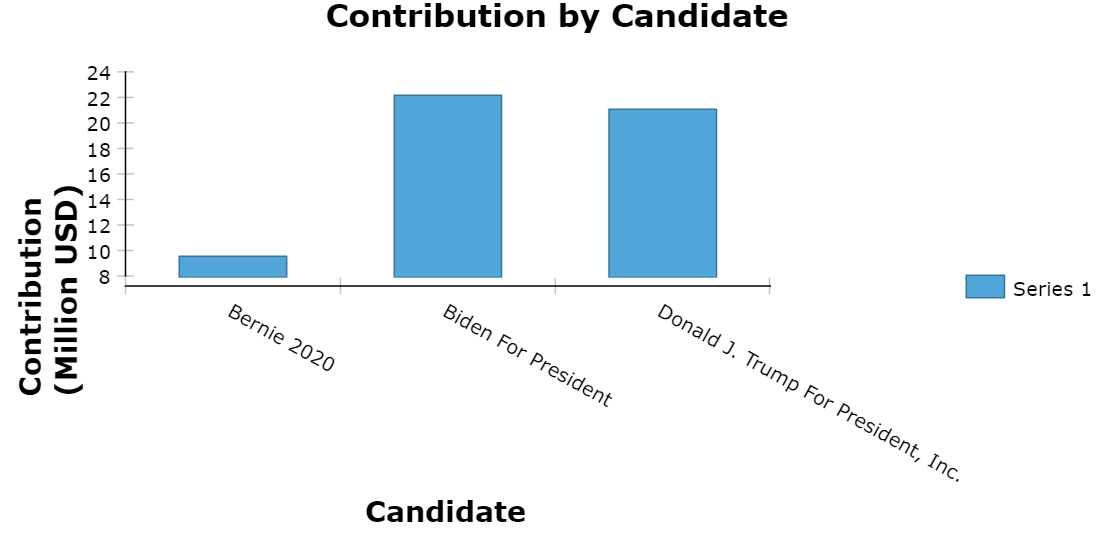


Figure : Contribution by candidate

We studied the top cities, contributed $500,000 are presented in figure 9. New York city is the top contributor, which contributed around 6.5 million dollars the two democratic candidates. Except New York City and Brooklyn, other cities collected most of the contributions toward one candidate. The top contributing cities are from TX (Austin, Dallas, Houston). Austin, Dallas and Houston cities lean towards democratic party. From figure 7, the contribution collected by Biden and Trump are almost equal. But this equality is not reflected on figure 8. Almost half of Biden’s collection sourced from the top cities. However, Trump’s collection is not from the top cities. This suggests that cities are strong holds of democratic party.

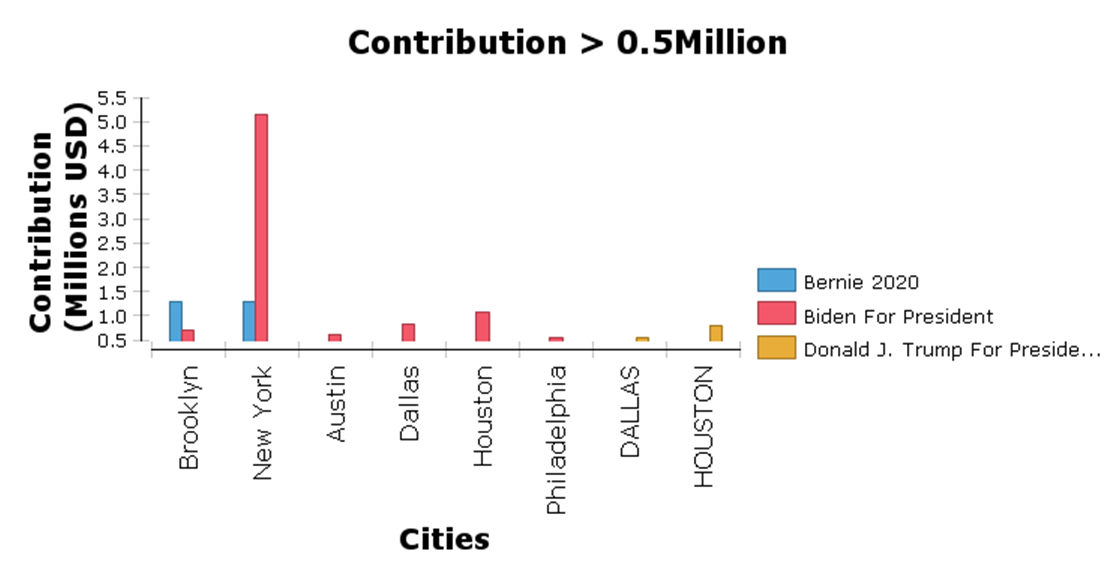


Figure : Top cities

We can also see contributors from the perspective of their employers. Figure 9 suggests that retirees have made the huge contribution to Trump, and we can infer that retirees lean towards republican party. Most of contribution made to democratic party came from those who are unemployed.

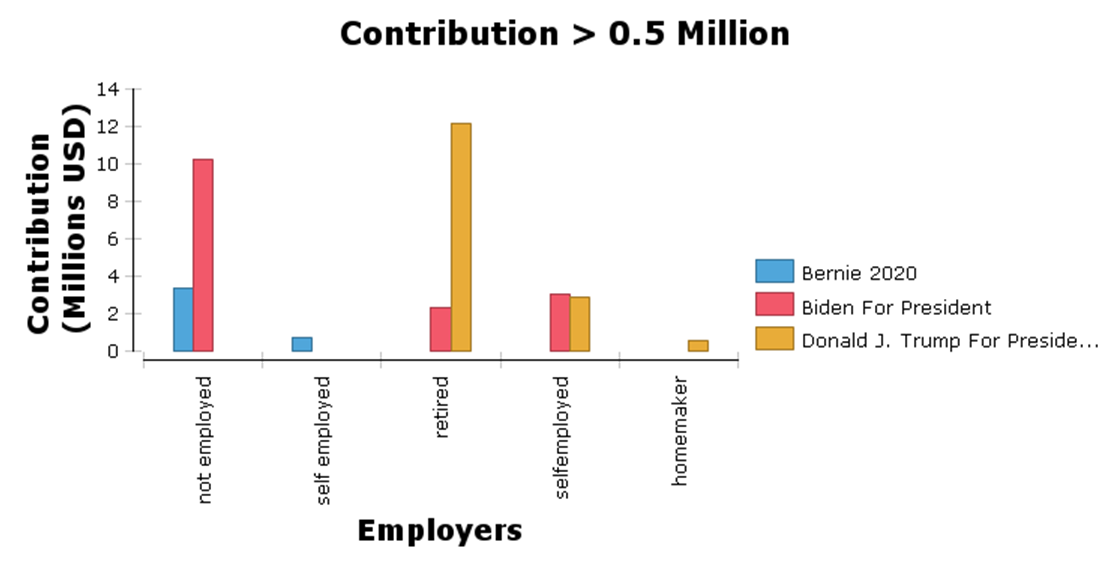
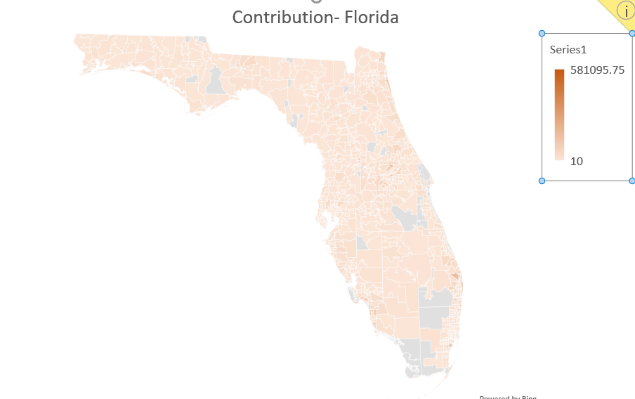
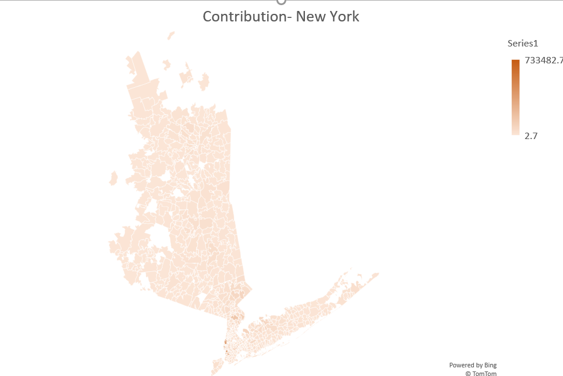
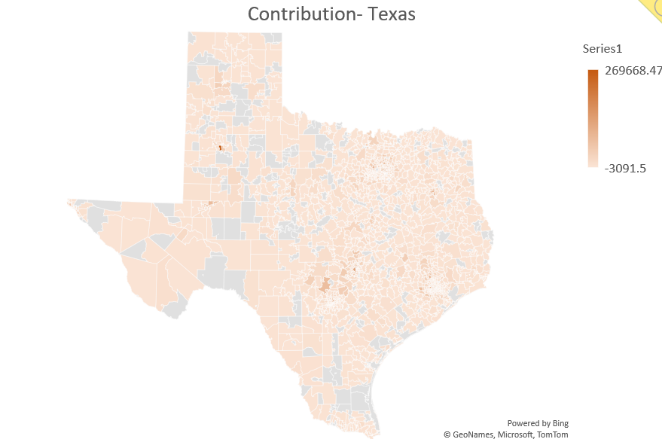
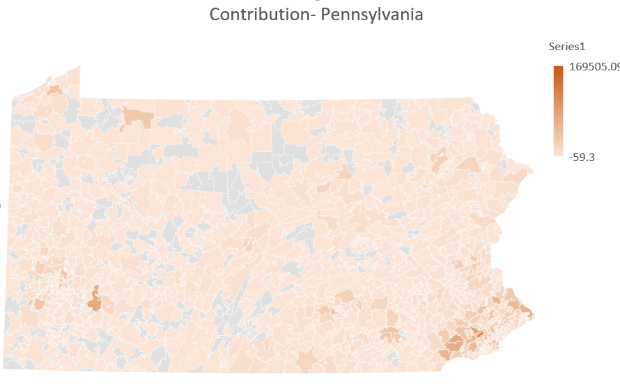


Figure : Top employers

The distribution of collected contributions in each state is shown from figure 10. In New York state, the distribution of contributions is almost from all zip codes. In Pennsylvania, most contributions were made from eastern section of the states. In Florida, Texas and Pennsylvania, there are pockets of the states from which no contribution has been made. These pockets could be potential target in future to campaign for donations.

Figure 10: Contribution made by zip code.



## 4.2 Hadoop Implementation

To be completed with project number 2

## 4.3 Reflective analysis of result in relational data warehouse vs Hadoop.

To be completed with project number 2

Conclusions

**Project 1:** This project was very helpful in understanding concepts of database warehousing and ETL. Schema and tables creation were accomplished seamlessly. The ETL part of the project was time consuming. There were 12 data sources to be merged, processed, and loaded. Once the data is processed on KNIME, persisting in the database was very time consuming. This is due to the computing capability of my machine and the size of the data. For example, one of the dimensions, took 3 hours to upload the data on the database. In general, in this project he loading part was very slow.

Overall conclusions of the project. In project 2, add a reflective analysis of the advantages and disadvantages of the two implementations.

APPENDIX

**SQL script used to create the schema.**

-- election.address definition

-- Drop table

-- DROP TABLE address;

CREATE TABLE address (

id serial NOT NULL,

address\_one varchar NULL,

address\_two varchar NULL,

city varchar NULL,

state bpchar(2) NULL,

zip int4 NULL,

CONSTRAINT address\_pk PRIMARY KEY (id)

);

-- election.address definition

-- Drop table

-- DROP TABLE address;

CREATE INDEX address\_zip\_idx ON election.address USING btree (zip);

CREATE TABLE election.transactions(

id serial NOT NULL,

tran\_id varchar NOT null,

CONSTRAINT transactions\_pk PRIMARY KEY(id);

);

CREATE INDEX transaction\_idx ON election.transactions USING btree (tran\_id);

-- election.datecontribut definition

-- Drop table

-- DROP TABLE datecontribut;

CREATE TABLE datecontribut (

id serial NOT NULL,

contr\_date date NULL,

CONSTRAINT datecontribut\_pk PRIMARY KEY (id)

);

CREATE INDEX datecontribut\_contr\_date\_idx ON election.datecontribut USING btree (contr\_date);

-- election.occupation definition

-- Drop table

-- DROP TABLE occupation;

CREATE TABLE occupation (

id serial NOT NULL,

employer varchar NULL,

occupation\_name varchar NULL,

CONSTRAINT occupation\_pk PRIMARY KEY (id)

);

CREATE INDEX occupation\_occupation\_name\_idx ON election.occupation USING btree (occupation\_name);

-- election.recipient definition

-- Drop table

-- DROP TABLE recipient;

CREATE TABLE recipient (

id serial NOT NULL,

fecid varchar NOT NULL,

committee\_name text NOT NULL,

CONSTRAINT recipient\_pk PRIMARY KEY (id)

);

CREATE INDEX recipient\_committee\_name\_idx ON election.recipient USING btree (committee\_name);

-- election.contribution definition

-- Drop table

-- DROP TABLE contribution;

CREATE TABLE contribution (

filing\_id bpchar(8) NOT NULL,

amount float8 NULL,

count int4 NULL,

transid varchar NOT NULL,

addressid int4 NOT NULL,

dateid int4 NOT NULL,

occupationid int4 NOT NULL,

recipientid int4 NOT NULL,

CONSTRAINT contribution\_pk PRIMARY KEY (trans\_id),

CONSTRAINT contribution\_address\_fk FOREIGN KEY (addressid) REFERENCES election.address(id),

CONSTRAINT contribution\_datecontribut\_fk FOREIGN KEY (dateid) REFERENCES election.datecontribut(id),

CONSTRAINT contribution\_occupation\_fk FOREIGN KEY (occupationid) REFERENCES election.occupation(id),

CONSTRAINT contribution\_recipient\_fk FOREIGN KEY (recipientid) REFERENCES election.recipient(id)

);

CREATE INDEX contribution\_addressid\_idx ON election.contribution USING btree (addressid);

CREATE INDEX contribution\_dateid\_idx ON election.contribution USING btree (dateid);

CREATE INDEX contribution\_occupationid\_idx ON election.contribution USING btree (occupationid);

CREATE INDEX contribution\_recipientid\_idx ON election.contribution USING btree (recipientid);

CREATE INDEX contribution\_transid\_idx ON election.transactions USING btree (transid);

**Loading data to HDFS**

# Loading election data on HDFS.

hadoop fs -put /tmp/election\_data/contributions\_m1\_2020\_FL.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m1\_2020\_NY.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m1\_2020\_PA.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m1\_2020\_TX.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m2\_2020\_FL.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m2\_2020\_NY.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m2\_2020\_PA.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m2\_2020\_TX.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m3\_2020\_FL.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m3\_2020\_NY.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m3\_2020\_PA.csv /user/root/electionData/input

hadoop fs -put /tmp/election\_data/contributions\_m3\_2020\_TX.csv /user/root/electionData/input

