Data warehouse Modelling- Independent Expenditures

Document Authorization

Document name	Data warehouse-Independent Expenditures.docx
Author(s)	Vanesa Lopez Garcia, Sabrina Mirtcheva, Daan Pelt, Moritz Steinbrecher, Hsiu-Chi Liu, Andrew Rizk, Pravat Ranjan Pasayat
Department	Group O-2-4, MBD-2018, IE HST
Last modified by	
Last modified date	04-Nov-2018
Authorization type	Baselined
Version	1.0

CONTENTS

1. Introduction

'Independent Expenditures' is a term used in the United States political process to describe political campaign contributions made by an individual or organization, "is not coordinated" candidate (Federal Election Commission, n.d.). The contribution is made towards either the election of a candidate, or the defeat of an opponent.

The Federal Election Commission (FEC) is responsible for monitoring and enforcing campaign finance law in United States federal elections. Although independent expenditures are not considered contributions and have no donation limits, they are still subject to some degree of reporting. Most states require independent expenditure reporting, but there is inconsistency in the types of information the contributor is required to report. Some states mandate reporting on a pre-defined schedule, while other states require reporting when the donation exceeds a specific amount. Further variance in the reporting data comes from whether the contributor is a person or group. The variations in reporting requirements may create inconsistencies in the data, requiring further cleaning in order to gain insights.

We have analyzed the independent expenditures report for 2016 and have made a proposal for how the FEC should design their data warehouse and made some suggestions on how they can make their data 'cleaner' and more useful for stakeholders.

The columns in this report help the reader understand the characteristics of the donor (type, location, occupation, etc.), but also included are some columns which are more relevant for the owner of the data (the FEC). For the purposes of our primary stakeholder, we have suggested to clean the data in what we saw as the most useful for them.

2. Stakeholder Analysis

2.1 Primary Stakeholder: Journalist

We have determined that Journalists are the main stakeholders of this Independent Expenditure dataset. They have the goal to inform the public about the process of political development and stimulate democratic debate on elections. Therefore, they play a key role in providing constituents with fair information about the campaigns of all candidates and their political parties, so that those voting can make informed decisions

The media has the responsibility to inform citizens about any violation of the rights of voters or any fraud committed by the political parties or by the election commission during the electoral process. Journalists' role consists on watching the election process and the election commission to make sure the elections have a legal procedure.

Overall, journalists will be interested in accessing the information contained in the reports offered by the

Federal Election Commission since these reports contain information such as the committee ID, the transaction amount, the transaction date, the candidate being supported, the contributor and the region where the transaction comes from. Furthermore, these reports offer other important fields that complement this information as for instance, the city, the state and zip code, the type of entity that made the contribution and the transaction type so that all the relevant information can be published and disclosed to the public.

2.2 Secondary Stakeholders:

- Political Parties: Political parties usually need information on independent expenditure in order to adapt their political strategies during the whole process of election.

- Analysts: Analysts working independently or in any organization use this information to extract the expenditure pattern from different geographic location and take certain decisions based on the analysis outcome.

With the purpose of analyzing this dataset, the stakeholders chosen to interpret the data are the journalists due to the important role they play during elections on disclosing relevant information to the public, watching the elections process and informing about any violations of the law.

3. Approach to Data Analysis / Cleaning

The Independent Expenditures dataset consists of 22 total fields describing the contribution of an individual, group, or a committee to a candidate in a certain US Federal Election. The columns present detailed information about transactions made to the candidate, for example in terms of amount, date of transaction, contributor, purpose of contribution, etc. This data will be used to create a database model that serves journalists to find and report relevant information about supporters or opposes of candidates. In addition to suspicious and controversial contributions/transactions.

3.1 Attribute and data analysis

- File Number: is a 7-character unique identification number for each file. A few negative values were found in the dataset but will be altered to positive ones. This property is not of use from a stakeholder point of view, so it can be excluded from the model.

- Committee ID: a 9-character variable used to identify a group or an individual making an expenditure. The first character is C which stands for committee and the last 8 characters are unique and constant for each individual or group throughout any expenditure made.

- **Amendment ID:** a one-character variable (A, N, or T) where 'N' refers to a new report being filed, 'A' refers to an amendment to a previous report, and 'T' refers to a termination report. It is important to note that amendment reports contain transactions recorded in previous reports, therefore a transaction will be recorded again in amendments causing duplication of data during tabulation

- **Report Type:** a 2- or 3-characters' variable that indicates the type of the report filed. Each report covers a certain period of transactions, so codes are given to reports accordingly. Report type column has no missing values and the available data shows no divergence from the known types.

- **Transaction PGI:** a 5-character variable that refers to the type of election (P= Primary, G=General, S=Special, or Other) and the year at which the election is held. 0.8% of the values in this column are missing. A large percentage of the values has election type variable however the year variable is missing. Since our dataset is meant for 2016 elections, PGI values from different years (e.g. 2014, 2015, and 2018) will be excluded from the analysis, and the missing years are assumed to be 2016. Therefore, only the type character will be included to group the dataset by type.

- **Image Number:** a string type variable that consists of either 11 or 18 characters. Based on the FEC website the two formats are as follows:
    - 11-digit image number format YYOORRRFFFF YY- scanning year OO- office (01- House, 02- Senate, 03- FEC Paper, 90-99- FEC Electronic) RRR- reel number FFFF- frame number
    - 18-digit image number format (June 29, 2015) YYYYMMDDSSPPPPPPPP YYYY- scanning year MM- scanning month DD- scanning day SS- source (02- Senate, 03- FEC Paper, 90-99- FEC Electronic) PPPPPPPP- page (reset to zero every year on January 1)

- **Transaction Type:** a string type variable that refers to the type of the transaction being made. There are numerous types of transactions however for the dataset, only 8 types are included (24A, 24C, 24E, 24F, 24K, 24N, 24R, 24Z).

24A	Independent expenditure opposing election of candidate (Represents 19.3% of the data)
24C	Coordinated party expenditure (Represents 0.4% of the data)
24E	Independent expenditure advocating election of candidate (Represents 27.1% of the data)
24F	Communication cost for candidate (only for Form 7 filer) (Represents 0.4% of the data)
24K	Contribution made to nonaffiliated committee (Represents 51.9% of the data)
24N	Communication cost against candidate (only for Form 7 filer) (Represents 0.1% of the data)
24R	Election recount disbursement (Represents 0.1% of the data)
24Z	In-kind contribution made to registered filer (Represents 0% of the data (4 values))

This indicates that most of the dataset is of type 24K (51.9%, to nonaffiliated committee), 24E

(27.1%, advocating election of candidate), or 24A (19.1%, opposing election of candidate). Based on this analysis, the data can be grouped based on these categories.

- **Entity Type:** a 3-character string type variable that indicates the type of the contributor making the expenditure (only valid for electronic filings). Entity type is divided into 7 main groups:
  - CAN = Candidate
  - CCM = Candidate Committee
  - COM = Committee
  - IND = Individual (a person)
  - ORG = Organization (not a committee and not a person)
  - PAC = Political Action Committee
  - PTY = Party Organization

  Using these categories, expenditures can be grouped accordingly. 1.5% (7830 entries) of this column is missing and will be replaced by an unknown entity (UKN).

- **Name:** refers to name of the person contributing to the election, whether it's an individual, a group, or a committee. 0.5% of the column data (2419 values) is missing. Those values will be replaced by a string 'Unknown". This column can be used to aggregate amounts of contributions by a certain individual or a committee.

- **City:** refers to the city for where the contribution is made. This property can be used to categorize how much contributions were made from each city. Some data discrepancies have been noticed in a number of values at which the spelling is incorrect. These values will be corrected to create a match with the corresponding city names.

- **State:** refers to the state for where the contribution is made. This property can be used to categorize contributions by stating 0.5% of state values (2773 values) are missing from the dataset. A decision will be made on how to deal with these values when the data is loaded.

- **Zip Code:** refers to the zip code belonging to area or city of where the candidate is running. The stakeholder can use this information to group transactions for certain areas in a state and/or city. In this column missing, incomplete and error values are detected. These are replaced by "Unknown". The city and state are used to identify the area of the candidate.

- **Employer:** this column states the employer of the contributor. In the total dataset we find only 10 values (0.0%) for this attribute. The values do not add value compared to the total dataset, therefore this column is disregarded from an analysis point of view.

- **Occupation:** this column states the occupation of the contributor. In the total dataset we find only 10

values (0.0%) for this attribute. The values do not add value compared to the total dataset, therefore this column is disregarded from an analysis point of view.

- Transaction Date: this column represents the date of when the transaction was made. Transaction Data is important for identifying 24h- and 48h Reports and is considered an import variable for analysis. In the dataset 1,998 missing values (0.0%) were detected, in additions to some errors. Errors in the far past (1975) and in the future (2017-2019, or 2025). These values are replaced by "Unknown".

- Transaction Amount: this column represents the amount of a transaction made by a contributor. This column is one of the main variables for analysis to see how much is contributed. Negative values are detected and amounts equal to zero (1538 values). Since negative amounts aren't logical because transactions are contributions, we assume these are errors and should be positive. These amounts are adjusted to positive amounts. Amounts equal to zero are removed since contributions of zero are irrelevant.

- Other ID: this column represents a unique ID numbers given by the FEC to a candidate, entity or committee that is contributing. Names that are linked to one Other ID may be different. In the column no missing values or errors are detected.

- Candidate ID: Candidate ID is the unique number given by the FEC to a candidate. The candidate will keep his or her Candidate ID across election cycles as long as the candidate is running for the same office. This attribute is relevant to stakeholders to see who is contributing to this candidate, from where and how much. No errors or missing values are found within this column.

- Transactions ID: this column has unique identifiers for a specific committee and report. This ID number is only valid for electronic filings, so they need to appear in an FEC electronic file. Multiple transaction can be filed with the same Transaction ID when they belong to the same report and include amendments. In this column no missing values and/or errors were identified.

- Memo Code: this column if amounts of transactions need to be included in the itemization total. This is indicated by an X. Transaction with an X are not included to the itemized independent expenditures. Blank values in this column are included to the total of itemized independent expenditures. No errors have been found in this column.

- Memo Text: this column shows descriptions of activities. Memo text is available on itemized amounts on schedule A and B. These transactions are included in the itemization total.

- Sub ID: the sub ID is the unique row ID given by the FEC when compiling this dataset. This ID number

is linked to the row rather than a specific transaction. This ID number is not relevant for our primary stakeholder and therefore can be deleted from the dataset.

3.2 General Assumption:

Remove rows: Since the amount of a transaction is important for analysis, rows cannot be deleted. Only duplicate rows and Transaction Amounts equal to zero can be deleted, in addition to duplicate transactions with similar Transaction IDs.

4. Data Warehouse Approach

Before designing a data warehouse, it is a crucial step to assess the most appropriate model. This assessment depends on the business context, stakeholders' needs and understanding of the underlying dataset.

For the selection of the data warehouse approach, we followed below steps.

1. Identified the underlying business entities and their respective relationships, as described in section 5.1.
2. Validated logical entities and relations with potential data warehouse approaches (Star, Snowflake and Data Vault schema modelling approaches.)
3. Identified the modelling approach that best fits our dataset and then proceeded with designing the model.

Star schema:

The star schema serves well for mapping the multidimensional data into a relational database and makes the information stored in critical dimension tables easy to access. Querying the data stored in a star scheme does not require complex queries and only few computing powers.

Snowflake schema:

The snowflake schema on the other hand favors data normalization. Thereby, it helps to reduce data redundancy, fosters integrity of data and ultimately improves the performance of the database. Likewise, less storage is needed as redundancy is eliminated. As a result, a clearer picture of how the attributes relate to each other and the business hierarchy is created. Furthermore, a snowflake model offers more flexibility to advance the data model by including additional dimensions, hence is more scalable and easier to maintain.

Data vault:

The Data Vault is a detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business. It is a hybrid approach encompassing the best of breed

between 3rd normal form (3NF) and star schema. The design is flexible, scalable, consistent, and adaptable to the needs of the enterprise. It is a data model that is architected specifically to meet the needs of today's enterprise data warehouses.

Data vault at the same time has below disadvantages.

1. A design that doesn't have a balance between Business and IT needs. Hugely IT influenced design.
2. Extensive set of rules and recommendations to follow, causing qualities issues later due to non-adherence.
3. Does not support easy data access. It introduces many tables joins.
4. Two to three-fold explosion of the number of tables. This increases semantic complexity of the database

Selected schema design – Snowflake and the motivation behind

As contributions serves well as a logical center of the model, it was chosen as the fact table, complemented by other entities which provide information such as who contributed, to whose campaign, when and where the contribution was reported. These entities relate to a certain contribution, which makes a central star schema most appropriate. Likewise, the high cardinality of the recipient, date and report dimensions, respectively the one-to-many relationships between the fact- and dimension-tables are suitable for a Star approach.

However, since not all dimensions are denormalized it, the result is a multidimensional Snowflake model. The contributor dimension, assumingly the main subject of analysis requires to be normalized into additional profession and geography dimensions to reduce redundancy of the data, whilst the logical hierarchy behind the dimensions remains through referential integrity.

One main priority for the data warehouse model selection is the flexibility of the model to be enhanced by complementary data. The FEC gathers an array of data-bases storing information about participators in independent expenditures, that can be beneficial for analysis purposes. Additional dimensions to the model can become important not only for analysis, but also for potential regulatory requirements for information.

The advantage of storage reduction comes into play considering the data-base's long-term orientation. Efficiency of the data warehouse is a main priority. The reduced data redundancy of the Snowflake-schema also limits the overall efforts for the FEC to maintain the data warehouse. Ultimately, which model suits the data warehouse best depends on the type of analysis stakeholders aim to perform with it.
As described, journalists may perform very diverse use cases for data analysis, which requires reduced complexity of queries and number of joins. Although there is no operational retrieval of data, the focus is on the ease using the data warehouse.

5. Data Warehouse Design

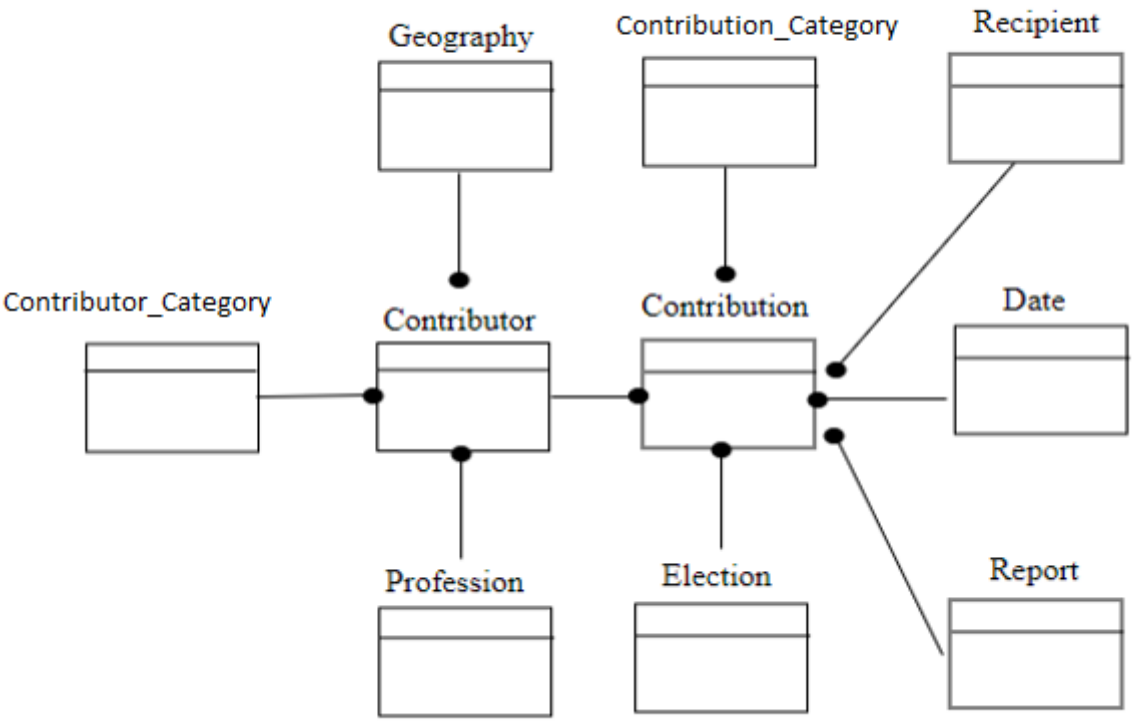
5.1 Conceptual data modelling:

During this phase, we identified highest level relationships between different entities.

<u>High-level entities:</u>

1. Contribution (primary entity – expenditure detail)
2. Date (Secondary entity- date when contribution is made)
3. Election (Secondary entity – election in which contribution is made)
4. Report (Secondary entity – Report in which contribution is registered)
5. Contribution_Category (Secondary entity – Type of contribution made)
6. Recipient (Secondary entity – Person or group that received the donation)
7. Contributor (Secondary entity – person or group that contributed)
    a. Profession (Tertiary entity – Professional details of the contributor)
    b. Geography (Tertiary entity – Geographic location details of the contributor)
    c. Contributor_Category (Tertiary entity – Type of contributor)

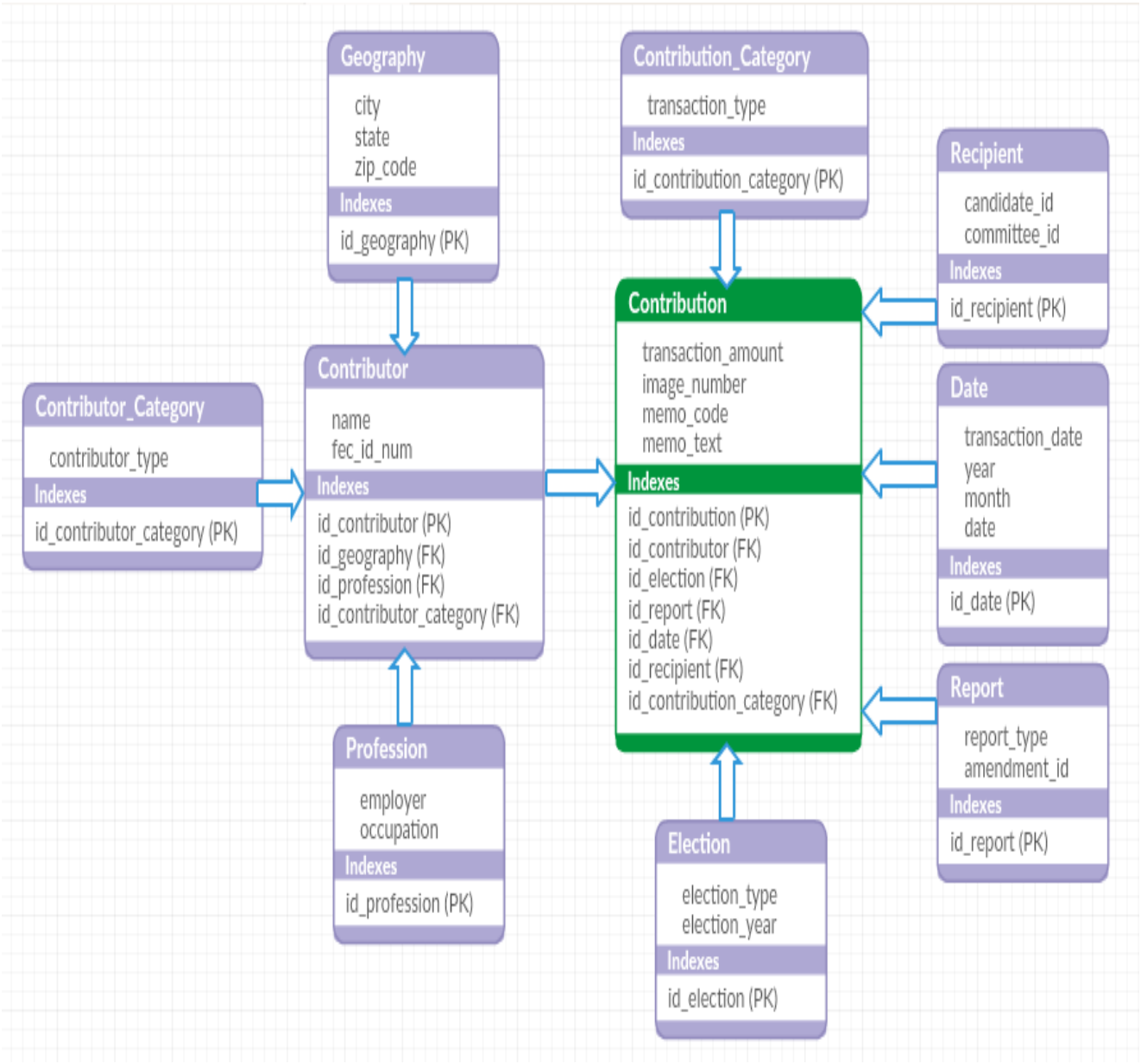
<u>Conceptual data model diagram:</u>



5.2 Logical data modelling:

During this phase, we identified specific entities, attributes and relationships involved in business function. This will serve as the basis for the creation of the physical data model.

Logical data model diagram:



5.3 Physical data modelling:

During this phase of data modelling, we implemented the logical data model in ´MySQL work bench´. We transformed the logical data model into proper selected tool specifics, adding types etc.

Physical data model diagram:

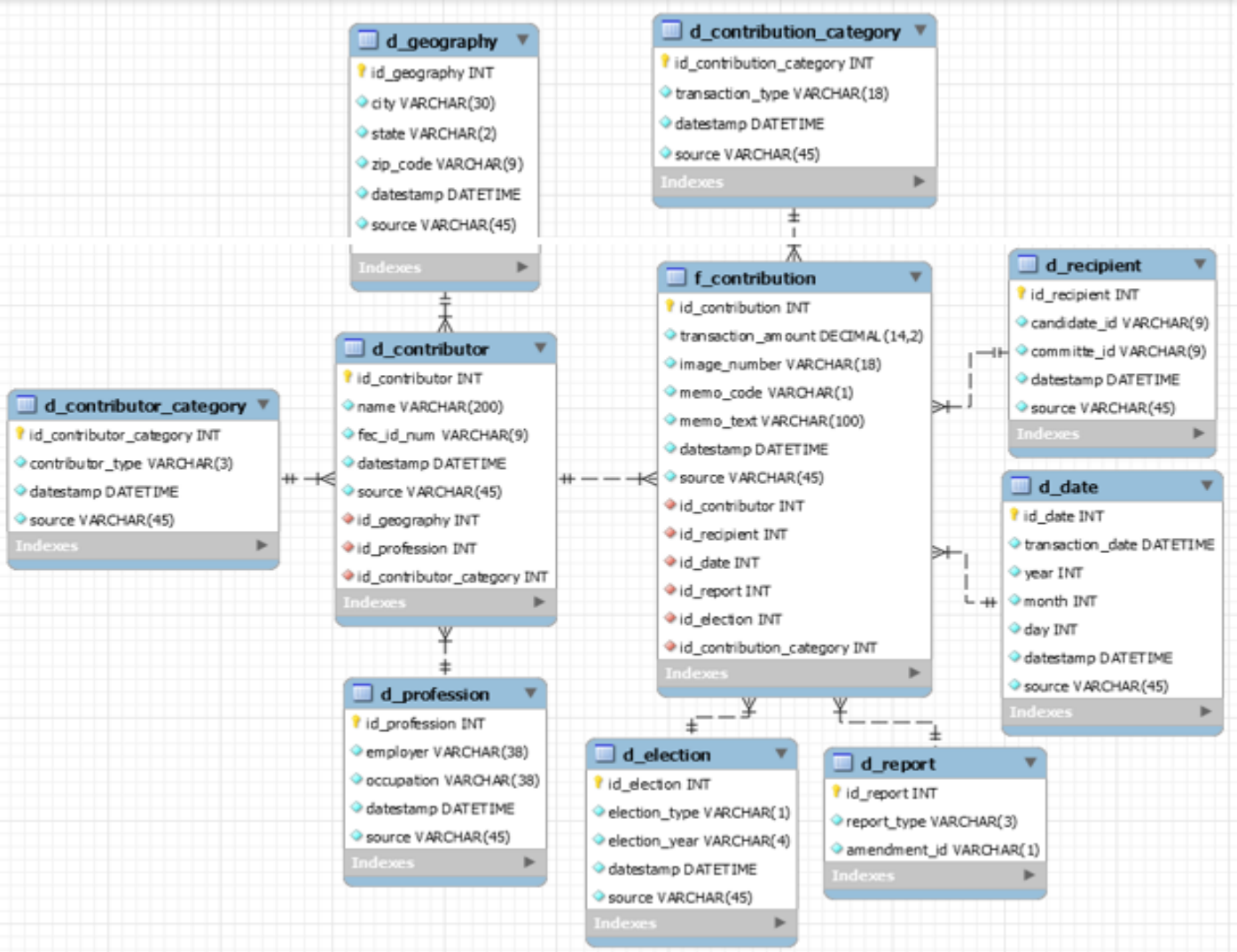


Table Description:

* Convention followed:

- Fact table has a prefix ‘f_’
- Dimension tables have a prefix ‘d_’

1. **f_contribution:** The fact table contains all the attributes related to monetary contribution. The reason of choosing this as a fact table is that contribution is the key of the whole dataset, in other words, this is the most important data for all kinds of stakeholders.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_contribution	INT	✓		✓	✓	✓
transaction_amount	DECIMAL(14,2)			✓		

Image_number	VARCHAR(18)			✓		
memo_code	VARCHAR(1)			✓		
memo_text	VARCHAR(100)			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		
id_contributor	INT		✓	✓		
id_recipient	INT		✓	✓		
id_date	INT		✓	✓		
id_report	INT		✓	✓		
id_election	INT		✓	✓		
id_contribution_category	INT		✓	✓		

2. **d_contributor:** This dimension table contains information of contributors, including name, professional background and detailed address, which will identify each contributor uniquely.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_contributor	INT	✓		✓	✓	✓
name	VARCHAR(200)			✓		
fec_id_num	VARCHAR(9)			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		
id_geography	INT		✓	✓		
id_profession	INT		✓	✓		
id_contributor_category	INT		✓	✓		

3. **d_geography:** This dimension table contains the details of contributors' address.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_geography	INT	✓		✓	✓	✓
city	VARCHAR(30)			✓		
state	VARCHAR(2)			✓		
zip_code	VARCHAR(9)			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		

4. **d_contributor_category:** This dimension table contains the category information of the contributors.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_contributor_category	INT	✓		✓	✓	✓
contributor_type	VARCHAR(3)			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		

5. **d_profession:** This dimension table contains the professional background of contributor.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_profession	INT	✓	✓	✓	✓	✓
employer	VARCHAR(38)			✓		
occupation	VARCHAR(38)			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		

6. **d_recipient:** This dimension table contains recipient information.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_recipient	INT	✓	✓	✓	✓	✓
candidate_id	VARCHAR(9)			✓		
committe_id	VARCHAR(9)			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		

7. **d_contribution_category:** This dimension table contains the additional detailed information of the contribution.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_contribution_category	INT	✓	✓	✓	✓	✓
Transaction_type	VARCHAR(3)			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		

8. **d_date:** This dimension table contains date information of the contribution made.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_date	INT	✓	✓	✓	✓	✓
transaction_date	DATETIME			✓		
year	INT			✓		
month	INT			✓		
day	INT			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		

9. **d_report:** This dimension table contains the information related to payment reports.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_report	INT	✓	✓	✓	✓	✓
report_type	VARCHAR(3)			✓		
amendment_id	VARCHAR(1)			✓		

10. **d_election:** This dimension table contains the information related to the election for which contributions are made.

Column Name	Datatype	Primary Key	Foreign Key	Not Null	Unique	Auto Increment
id_election	INT	✓	✓	✓	✓	✓
election_type	VARCHAR(1)			✓		
election_year	VARCHAR(4)			✓		
datestamp	DATETIME			✓		
source	VARCHAR(45)			✓		

6. References:

1. https://public.enigma.com/datasets/independent-expenditures-2016/dc15823c-6e6d-4e7f-bf7d-50939f56730b
2. https://www.fec.gov/data/