**Bipartisan 9 Research & Analytics, LLC**

Sam Tsai

Vinh Phan

Cornelia Harris

**Extraction, Transformation,**

**and Load Technical Report**

TABLE OF CONTENTS

1.                Introduction 3 1.1 Summary                                   3

1.2              Scope                                                                             3

1.3              Technologies and resource contributions                   3

1.4              Definitions, Acronyms and Abbreviations                  3

  2.                ETL Details                                                                    4

  2.1              Data Import/Extract Sources and Method                4

2.2              Data Acquisition                                                         4

  2.3              Data Transform                                                          4

  2.4              Data Integrity                                                              5

  2.5              Data Refresh Frequency                                           5

  2.6              Data Security                                                               5

  2.7              Data Loading and Availability                                 5

  3.                Data Quality                                                                6

1. **INTRODUCTION**

The Bipartisan 9 Research & Analytics, LLC is a communications infrastructure and media research company that provides ETL services and marketing insights to organizations seeking to strategize political campaigns.

***1.1 Summary***

Republican Senator Max Beiner is running for office in 2020 and his campaign strategist has decided to use YouTube to reach their target market.   A proposal was submitted to commission the Bipartisan 9 to provide information on which political campaigns were most trending on YouTube in the United States. Beiner’s campaign strategist hopes to view the messages trending and use this information to help understand the public’s response to political campaigns via YouTube.

***1.2 Scope***

The Bipartisan 9 Research & Analytics, LLC has a reputation for building solid data models leading to a streamlined ETL (Extract-Transfer-Load) process, delivering valuable insights to their clients by collecting organic data and/or utilizing reliable data sources in their analysis.

The Bipartisan 9 Research & Analytics, LLC have decided to utilize datasets from Kaggle, a cloud-based, public data platform to satisfy the requirements outlined in the proposal.

***1.3 Technologies and resource contributions***

The Bipartisan 9 Research & Analytics, LLC organization is led by its founding members who play an integral role in the ETL process.

Sam Tsai - Senior ETL Developer

Vinh Phan - ETL Developer

Cornelia Harris - Data and Reporting Analyst

Tech stack used in ETL process: Python and PostgreSQL

***1.4 Definitions, Acronyms and Abbreviations***

·   ETL - Extraction of data from a source, Transformation of data into a usable format, and Loading of data into a database.

·  Groupby - a function that groups rows that have the same values into summary rows.

·   Trending stats - an upwards or downwards shift in a data set over time.

·   Schema - is a structural representation of the entire database.

·   Database-is a collection of facts or information about the considered object.

·   API key - a code that gets passed in by computer applications. The program or application then calls the API or application programming interface to identify its user, developer or calling program to a website.

·   Postgres - a free and open-source relational database management system.

·   Political Bias - is a bias or perceived bias involving the slanting and altering of information to make a political position or political candidates seem more attractive.

·   Youtube\_id - The ID that YouTube uses to uniquely identify the channel.

·   snippet - The snippet object contains basic details about the channel, such as its title, description, and thumbnail images.

· bias - prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.

**2. ETL DETAILS**

After a brainstorming session between the team at Bipartisan 9, the first step was to decide the best data sets to use in the ETL process. An extensive search on various dataset websites: Data World, Reddit, Google dataset search engine, and Kaggle proved the most pertinent datasets to satisfy the requirements of this proposal was found on Kaggle. Our goal was to find if there was a political bias slant to trending videos on YouTube in the US during the year 2017.

***2.1 Data Import/Extract Sources and Method***

**Trending YouTube Video Statistics dataset:** (Daily Statistics for Trending YouTube Videos) This dataset includes data on approximately 6 months (2017-11-14 to 2018-06-14) of trending YouTube videos in specified countries. This data was acquired using an API key for YouTube Data API.  The dataset includes 10 country codes which are two-letter country abbreviations according to ISO 3166-1.  The data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. The data also includes a category\_id field, which varies between regions. This information was provided in both .csv and .JSON formats. File size: 514.24 MB

**YouTube Political Bias:** (YouTube Dataset with Generated Features for Political Bias Detection) This dataset provides basic information, in JSON format, regarding 421 YouTube channels with an average of 10 videos per channel. Each channel contains information for corresponding media outlets generated from [https://mediabiasfactcheck.com](https://mediabiasfactcheck.com/). Each video contains information downloaded from YouTube API. This dataset doesn’t represent all videos in a given channel (CNN has more than 140K videos), however, it represents the actual video count for videos for this particular channel in the dataset as well as with their corresponding YouTube Ids. File size: 2.62 GB

***2.2 Data Acquisition***

Per the details outlined in the proposal, the client would like information during the year 2017. After extensive research on datasets, the research team found two datasets that will be used in the ETL process:  **Trending YouTube Video Statistics dataset:** (Daily Statistics for Trending YouTube Videos) and **YouTube Political Bias:** (YouTube Dataset with Generated Features for Political Bias Detection).

Data Extraction Process:

Both datasets are dynamic and would need to be updated if new date parameters are required. The datasets were collected using YouTube API. An API key would be needed to access the data if new search date parameters are requested or if the data from the specified timeframe requires an update.  Once a valid API Key has been obtained, it should be stored inside a text file named api\_key.txt. It can also be targeted with the --key\_path parameter. Once the data is collected, both files either .JSON or .csv files can be downloaded and read using Python programming and the Pandas library.

***2.3 Data Transformation***

Political Bias Dataset:

Of the two datasets, the most transformation took place in the Political Bias dataset. This dataset, provided in .JSON format, was downloaded and initially read into Python via Jupyter Lab. A view of the first 5 rows of data showed that the data was very nested, therefore a process of unpacking the data into rows for visualization was necessary. The columns deemed relevant from this dataset and for the scope of this project were: 'snippet', 'videos\_information', 'bias', 'youtube\_id'.  The first step in the transformation process was to unpack the ‘videos\_information’ column so that each row is a unique id mapped to a ‘youtube\_id’ column. Using the .iloc function we were able to view the video id’s and the video counts per row. The lambda function was used to unpack each record and the groupby function was used to append each record to the video id column. A new dataframe was created containing both the ‘youtube\_id’ and ‘videos\_information’ columns. Another dataframe was created with the ‘snippet and ‘bias’ columns. These tables were saved as test21.csv and test1join.csv respectively to be imported to Postgres database for further manipulation. The Jupyter notebook file used to clean the data has been provided as ETL.ipynb.

Trending Statistics Dataset:

This dataset was provided in a .csv format and was downloaded and read into Python via Jupyter Lab. The first 5 rows of data were viewed to see the columns in the dataset.  The ‘tags’ and ‘description’ columns were not needed for the requirements of this proposal, therefore, they were dropped from the dataset and the column titled ‘views’ was changed to ‘yt\_views’.  This cleaned data was then saved and exported as ‘USvideosREV.csv’ for further manipulation in Postgres. The Jupyter notebook file used to clean the data has been provided as ETL2.ipynb.

***2.4 Data Integrity***

Both datasets used are considered reliable because it uses information pulled directly from the YouTubes website using YouTube’s API.  The Trending Statistics dataset is a daily record of the top trending YouTube videos, however, for the purposes of this request, this data does not need to be updated.

***2.5 Data Refresh Frequency***

This data does not need to be updated per the requirement of this proposal.

***2.6 Data Security***

This data does not have any security requirements and does not have any federally mandated considerations.

***2.7 Data Loading and Availability***

After the transformation of the data, the new tables were saved as .csv files to be uploaded to a Postgres database. Prior to the upload, the schema was created with the pertinent columns and datatypes to prepare for the joining of the three tables using Postgres. The schema has been included as a resource to this proposal.

Adjustments to the schemas were required to import the tables properly. Regarding the US Video table, the ‘trending date’ object was originally valued as a ‘timestamp’ object (integer) but needed to be converted to ‘VAR CHAR’ to be read in as a string object. We also had to remove the ‘description’ column from the US Video data because it used the character (” \n”) which prompted an error. The parameters of the Escape Character had to be changed to quotation marks (“) from apostrophes (‘) when importing the files to avoid an import error.

Once the schema was properly structured and tables were loaded into the database, a query tool was created and the two tables from the Political Bias dataset were combined using the “join” function. This resulting table was exported as political\_bias\_joined.csv. Then, this table was joined with the US Videos table on the video\_id column and the results were saved as political\_bias\_us\_joined.csv. The query tool was saved as us\_query.sql.

**3. DATA QUALITY**

The results of the three tables combined into one showed 14 matches out of a total of 24,374 videos. In other words, out of 24,374 videos from various political outlets, only 14 videos made it to YouTube trending. These 14 matches all came from the Vox or CNN YouTube channels which are both labeled as left-wing biased. This observation shows that very few political videos manage to trend on YouTube. All the trending channels were considered left wing resulting in a possible left-wing bias.

The outcome of the ETL process is considered a success because not only did it provide insight as to which political videos trend the most on YouTube, it also showed that political videos do not trend a lot in comparison to other video content on YouTube. With the result from the ETL, the client may consider other media outlets to target with their campaign efforts.

The Bipartisan 9 Research and Analytics, LLC has provided the client with political\_bias\_us\_joined.csv to view the results.  The client will need to use any program that can open .csv files to load it.