2. ETL Details

This section outlines a detailed description of the ETL processes utilized and proposed to achieve the objectives of this initiative.

2.1 Data Import/Extract Sources and Method

For this initiative, our team gathered crime data from five major metropolitan cities across the United States, and their respective weather data. The five cities include Los Angeles, Chicago, Atlanta, Denver, and Boston. The crime data of these cities range over a five year period between 2014 to 2018. Additionally, weather data specific to the above mentioned cities over the same five year period were collected. The table below illustrates the datasets and sources.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Datasets | Sources | Data Statistics | Tools / Modules Used |
| Crime Data | Atlanta | <http://www.atlantapd.org/i-want-to/crime-data-downloads> | 5,458,515  Data Points |  |
|  | Boston | <https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system> | 6,629,480  Data Points |  |
|  | Chicago | <https://www.kaggle.com/chicago/chicago-crime> | 31,261,892  Data Points |  |
|  | Denver | <https://www.kaggle.com/paultimothymooney/denver-crime-data/downloads/denver-crime-data.zip/39> | 9,251,709  Data Points |  |
|  | Los Angeles | <https://www.kaggle.com/cityofLA/los-angeles-crime-arrest-data> | 41,079,624  Data Points |  |
| Weather Data | Cities Above | <https://darksky.net/dev> | 43800 Data Points per City |  |

The weather data was extracted from the XXXXXX API utilizing Python. The team investigated a myriad of different weather data sources, such as Open Weather Map API, however decided on utilizing the XXXXXX API due to cost-benefit calculations, data structure, ease of use and its detailed documentation.

On the other hand, the crime data were available as CSV files from the sources listed above. Fortunately, the process of obtaining the required data for this initiative did not require acquiring permissions or licensing.

2.2 Data Acquisition

To achieve the purposes of this proposal, the team thoroughly selected the five cities based on the sufficient availability of a) duration/years of data, b) locational (latitude/longitude) aspects, c) crime codes/descriptions, and d) dates and times of the crimes. Due to the initiative establishing the observation period from 2014 to 2018, the data is static; however, future endeavors of observing changes in future crime trends or analyses may require updates to the data.

2.3 Data Transformation

Prior to the data transformation process, the team constructed a schema and outline of required data from both the crime and weather data. Once the blueprint of the initiative was created, the data transformation process was performed as follows: a) truncation of dataset to the specified range of years, b) extracting essential columns required for data munging, c) identification and removal of NA/NULL/NaN data points based on their impact to the integrity of the data, d) data calculations and textual format manipulation for the unification of the datasets, and e) organization of columns established in the schema for loading the database. The identification and removal of NA/NULL/NaN data points was a vital part of this process. Three cities (Chicago, Denver, and Boston) contained NaN values for the locational data (latitude/longitude) of the reported crimes. The ratios of the missing data to the total size of the datasets were 1.33%, 0.08%, and 6.2%, respectively. Given the large size of the total size of all crime data, it was safe to assume that the omission of these missing values would not have a substantial impact on our initiative.

2.4 Data Integrity

2.5 Data Refresh Frequency

2.6 Data Security

2.7 Data Loading and Availability