

Group 3, LLC.

Extraction, Transformation, and Load Technical Report

Tech City, USA

**TABLE OF CONTENTS**

1. [Introduction 3](#_1rg4ys32gwdu)
   1. [Summary 3](#_30j0zll)
   2. [Scope 3](#_3znysh7)
   3. [Technologies and resource contributions 3](#_tyjcwt)
2. [ETL Details 4](#_82ggr1k51212)
   1. [Data Import/Extract Sources and Method 4](#_2s8eyo1)
   2. [Data Acquisition 4](#_3rdcrjn)
   3. [Data Transform 4](#_lnxbz9)
   4. [Data Integrity 4](#_1ksv4uv)
   5. [Data Refresh Frequency 4](#_2jxsxqh)
   6. [Data Security 4](#_3j2qqm3)
   7. [Data Loading and Availability 5](#_4i7ojhp)
3. [Data Quality 6](#_rr5pa9omjpzt)

# Introduction

# Summary

This section summarized the final objective of the project, the business problem definition (problem statement) and the expected outcome of ETL.

As home to the future of technology in America, Tech City is using data-driven solutions to better inform their approach to city planning. Incorporating this theme as they plan and goal-set for the City’s budget in the coming years, the Office of the Mayor has partnered with Group 3, LLC. to analyze the spending data for hosts to some of the most successful technology hubs in the country.

# Scope

This section explicitly outlines the disparate data sources that are to be integrated, which components of the overall data science project is in the scope for this initiative and also lists out the components of the data science project that are not in scope here.

Historic operating data from five major US cities were imported from disparate csv files for use in our analysis. The cities’ population data from the US Census was also leveraged to better understand how Tech City might approach their budget on a per capita basis - i.e. to normalize the spending data of the larger cities and account for some the population differences among the sample cities.

Cities considered in our approach:

* Atlanta, GA
* Austin, TX
* Chicago, IL
* New York, NY
* San Francisco, CA

# Technologies and resource contributions

This section lists out the team members and their contributions towards the ETL initiative. Use this section to also outline (or list) the tech stack used to obtain the final outcome.

Team Members:

* Lisa Ashway: Extraction & Transformation
* Asher Collins: Loading
* Amber Reynolds: Quality
* Jonathan Vallecillo: Summarization

Tools:

* Python: NumPy, Pandas, SQLAlchemy
* SQL: PostgreSQL
* QuickDBD

# 2. ETL Details

# Data Import/Extract Sources and Method

The data used to create this database was found primarily on the Data.world website. The cities of Austin, Atlanta, San Francisco, and New York City all uploaded their budget data in CSV form. The city of Chicago uploaded similar budget data on the data.gov website. Budget information for the year 2018 can be found in all datasets, so it was used as the best example of current city spending. Population information was found in excel form from the U.S. Census Bureau. Please see below for links and summaries of the datasets.

* **Atlanta, GA**

Atlanta Ledger and List of Vendors 2016 - 2019

* 1. Source: <https://data.world/brentbrewington/atlanta-open-checkbook>
  2. Type: CSV
* **Austin, TX**

Austin Program Budget Operating Budget Vs Expense Raw Data summary 2018

* 1. Source: <https://data.world/cityofaustin/austin-program-budget-operatin/workspace/project-summary?agentid=cityofaustin&datasetid=austin-program-budget-operatin>
  2. Type: CSV
* **Chicago, IL**

Budget - 2018 Budget Ordinance - Appropriations

1. Source: <https://catalog.data.gov/dataset/budget-2018-budget-ordinance-appropriations>
2. Type: CSV

* **New York City, NY**

City of New York Spending and Budget Tables 2018

* 1. Source: <https://data.world/city-of-ny/ng9f-agux>
  2. Type: CSV
* **San Francisco, CA**

San Francisco Vendor Payments (Purchase Order Summary) 2017 - 2020

* 1. Source: <https://data.world/sanfrancisco/p5r5-fd7g>
  2. Type: CSV
* **Population Data:**

City and Town Population Totals

* 1. Source: United States Census Bureau - <https://www.census.gov/data/datasets/time-series/demo/popest/2010s-total-cities-and-towns.html>
  2. Type: XLSX

# Data Acquisition

Since this database outlines the spending of cities in 2018, it is a static database. If further insight is needed for more recent years, the datasets will need to be refreshed. This can be done, so long as the cities do not change their department names or abbreviations, by reloading them into the python files for restructuring using pandas, and then manually uploading them into the SQL database. There are 9 tables in total in the completed dataframe, with tables being no larger than two columns. The rows in these tables are dependent on the number of cities used in the database. No security of pre-requisites exist for extracting this data.

# Data Transform

Each city in this dataset describes their spending in terms of departments. Cities like New York and San Francisco have over 40 departments, while the cities of Atlanta and Austin have 20 to 30. In order to transform the data, each department had to be manually sorted into one of eight categories for city spending. These include: (1) Community Health, (2) Culture & Recreation, (3) General Administration & Finance, (4) General City Responsibilities, (5) Human Welfare & Neighborhood Development,(6) Public Protection, (7) Public Works, Transportation & Commerce, and (8) Other. Additionally, some cities provide a budget spending table on a yearly basis - while others combine their budgets over many years into one table. Therefore, it was important to split the data into groups by fiscal year before analyzing spending. After the tables have been sorted and split, spending was grouped by each of the categories listed above, and summed.

# Data Integrity

The reliability of this database is heavily dependent on the accountability and processes of the city departments who release this data for administration transparency. Additionally, the integrity of the data depends on the manual sorting of department spending into the larger categories that were created. For example, the “Emergency Medical Services” spending in Austin, TX can be sorted into the category of (1) Community Health, or (6) Public Protection. The human error involved in this process does take away some integrity of the final database. Atlanta is also a particular case, because their dataset is based on vendor payouts, and not city spending in its entirety, and therefore does not showcase the entirety of their budget allocation. Despite these issues, we believe that the database shows how each city appropriates funds on the whole, even if every line item is not included.

# Data Refresh Frequency

The datasets are uploaded from cities at varying frequencies, some being quarterly and others yearly - so the frequency at which local data can be updated is yearly at a minimum.

# Data Security

Since this data is all publicly available there is no concern about privacy, encryption, data masking, auditing, backups, etc. If there were things like email, names, addresses, social security numbers we would then need to take these things into consideration and have plans for things like PII, GDPR, CCPA, and a few others depending on the country or state that the consumer whose data is being collected reside.

# Data Loading and Availability

As it stands today the SQL format has been chosen to make the data available for the client. The data was divided into a few different tables to allow growth in case the data for other cities need to be collected and compared. With it being publicly available data this could easily be converted into a webpage making access to the data easier.

# 3. Data Quality

Prior to transforming and loading the data, we reviewed the raw data to ensure it contained the information required to create a properly functioning database. Each table loaded into the database has 2 columns and between 3-5 rows of data, depending on whether all cities contained a specific department. As explained in the transformation section, the original data sets had thousands of lines of data that were parsed down to unique spending categories. When running queries in the SQL database, first run a query to select all from the tables to ensure that only 2 columns are being pulled. In order to ensure the quality of the database, we had one user in our group create and load the database and another user to test it separately.

The ultimate ask of the project was to see what departments/programs do our model cities all have in common? Through the transformation process we found the following commonalities:

* Community Health
* Culture & Recreation
* General Administration & Finance
* General City Responsibilities
* Human Welfare & Neighborhood Development
* Public Protection
* Public Works, Transportation & Commerce
* Other

The other question was what can we expect to budget per citizen for various city services? In order to answer this a query must be run using a selected department or program and dividing the population data from the chosen city. For example if you want to know the budget per citizen for Community Health in Austin, TX for 2018, you can run a query selecting the city of Austin’s budget in Community Health divided by the population of Austin and the expected outcome is roughly 32.71‬ per person annually.