Documentation of US Public Transportation Analysis (APTA):
Financial, Operational, Environmental Aspects

Executive Summary:

The public transportation industry has been experiencing drastic external changes in the past few decades. With the heightened emphasis being placed on green energy by society, the rapid emergence of ride-sharing services, as well as the unexpected advent of COVID-19, the U.S. public transportation system is experiencing pressure to successfully adjust. The purpose of this project is to analyze the financial, operational, and environmental factors that should be of particular note, in order to successfully transition the public transportation system into the new world.

Business Case:

The American Public Transportation Association (APTA) is aware of the rapid change in the climate of transportation in recent years. With rising costs, and heightened awareness of green energy and health and safety due to the entrance of COVID-19, APTA is undertaking a grand effort to identify key cost drivers (financially / environmentally) and new opportunities to adjust the public transportation system to be more effective and efficient. By using the provided data source and analyzing the trends in capital and operating expenditures, revenue & funding sources, as well as energy consumption details, we will provide a clearer picture of any current inefficiencies and opportunities that APTA could alleviate / exercise.

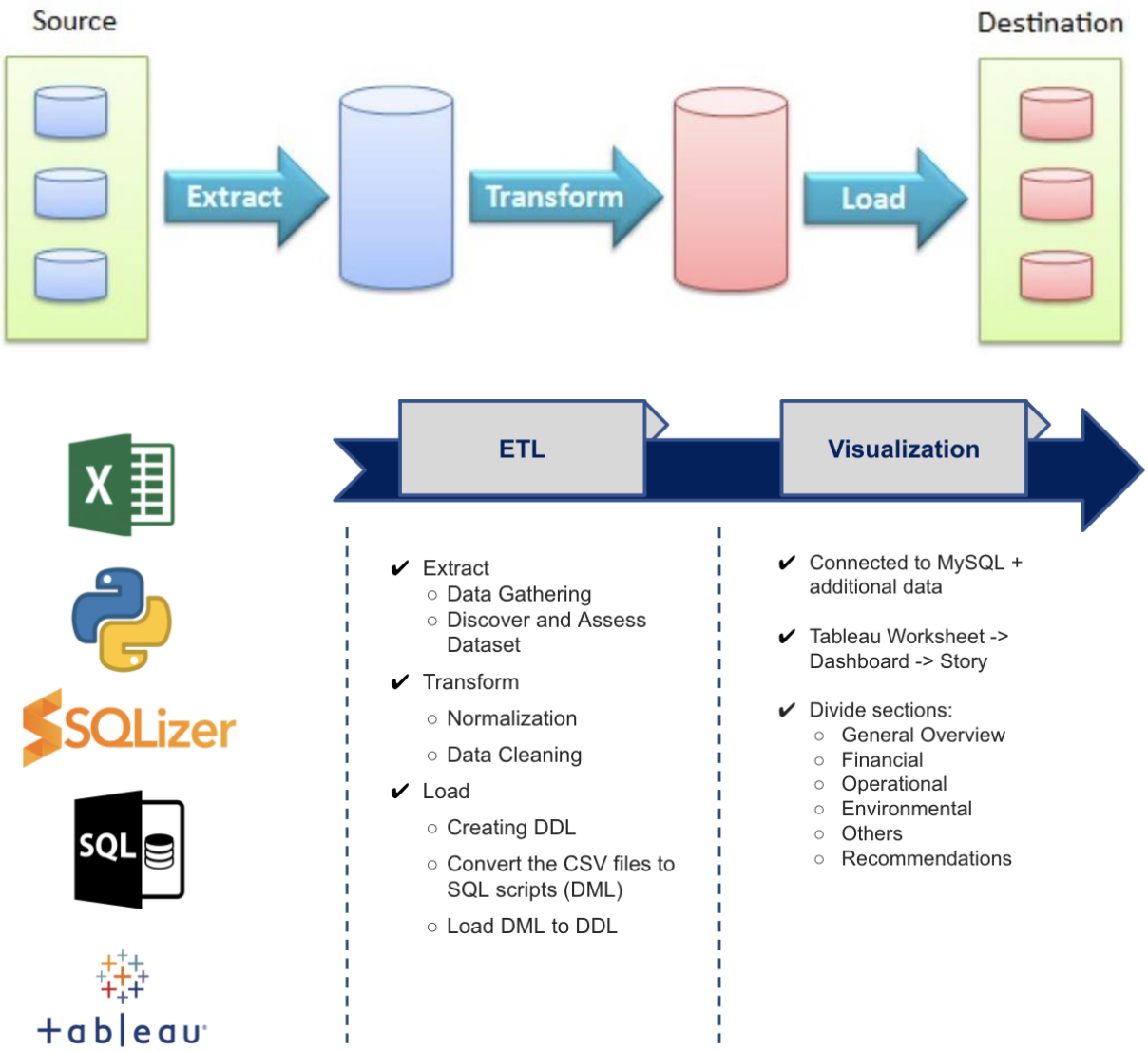
Data Sources:

https://www.apta.com/research-technical-resources/transit-statistics/ntd-data-tables/

Data Tools:

MySQL, SQL, Excel, SQLizer, Python, Tableau

Data Preparation and Cleaning:



1. Extract

a. Data Gathering
Dataset used in this final project comes from the official website of The American Public Transportation Association (APTA). We proceeded with the National Transit Database (NTD) produced by the Federal Transit Administration that covered all public transportation agencies in the US from 2016-2020.

b. Discover and Assess Data
This step aims to understand the data, find potential data crossing the same storyline, and consider the next step that needs to be assessed in the particular context of financial, operational, and environmental factors before doing data transformation.

2. Transform

a. Normalization
In this step, we split columns in the CSV file that were not required and merged the required ones, then normalized the tables into the third normal form.

b. Data Cleaning
We applied some cleaning activities such as missing value checking, tackling outliers and irregular data, eliminating duplicated values, and fixing structural errors including incorrect naming or capitalization. All these processes were done to the CSV datasets using python. data completeness, data format, data accuracy

Summary of data cleaning process:

Tools	**Python (Jupyter Notebook)**	**Excel/CSV**
Activity & Results	**data.head():** to show the upper 5 data from the dataset as a data example. **data.info():** to show the name of columns, number of non-null data, and data type. In this process, we can know which column has a different number of rows, how many rows are null, and whether the data type is already in accordance with what we expected or not. After this step, we fix the data in excel to make the process easier and faster. **data.describe():** to show descriptive statistics on the dataset, make sure that they look reasonable to be proceed.	**Agency Table** - Added in 4 cities that were left blank o FRS Transportation, Inc o Puerto Rico Maritime Transport Authority o SeaStreak, LLC o Bay State Cruise Company **Capex** - Removed $ signs from values in 2016, 2018, 2019 **Funding** - Removed $ signs from FundingAmount values in 2016, 2018, 2019 **Headcount** - Rounded HeadcountAmount to the nearest whole number **HourlyWages** - Replaced (blank) with the value 0 in the HourlyWagesAmount column
		LaborHours - None **Opex** - Replaced OpexAmount (blank) with 0 - Remove dollar signs whenever possible in the OpexAmount column in 2018, 2019

		FuelEnergy - There were vehicles of Mode SR with value – for VehicleCount. Replaced these values with 0

3. Load

After cleaning the CSV files, we loaded them into a relational database and ready to be visualized.

We converted the XLSX/CSV files into SQL DML scripts using SQLizer for easier data upload as we experienced difficulty in directly loading the CSV files into MySQL database (technical challenge especially with macOS).

Other than the original database from MySQL (focusing on the data from 3 states), we also connected additional excel files to Tableau to cover a little bit of the full state's data to complete the analysis, due to time constraints and capacity issues.

Database Platform Considerations:

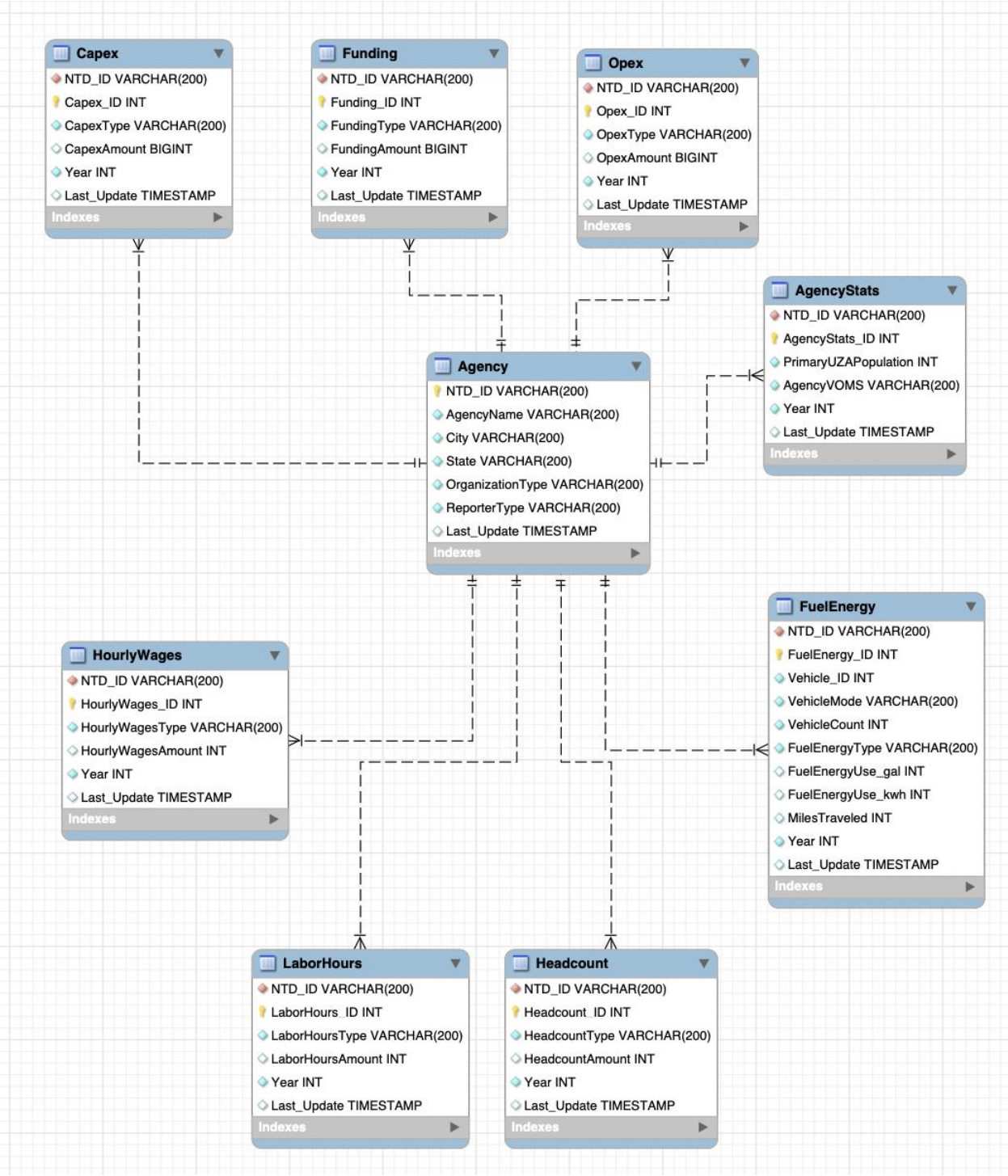
MySQL:

- High familiarity and simplicity.
- Sufficient to accommodate the required data loading, retrieval, and analysis.
- Based on external references, MySQL has good usability across platforms/operating systems and is renowned for being secure while freely available as an open-source.

Data Model:

- The datasets are normalized to remove redundancy, partial functional dependency, and transitive dependency.
- For example, transitive dependency exists in Opex Amount -> Opex Type -> NTD_ID (Agency). We split these attributes into separate tables and we assign surrogate keys (as primary keys) to each new table.
- This process results in a new normalized relational database called 'APTA' with multiple main tables for analysis:
  - Agency
  - AgencyStats
  - Capex
  - Opex
  - Funding
  - Labor Hours
  - Headcount
  - Hourly Wages
  - Fuel & Energy
- The tables are related/connected by NTD_ID.
- The EER diagram is enclosed.

EER Diagram - APTA



Data Analysis (Insights, Reports, Dashboards)

The full analysis can be accessed in our Tableau Story workbook (.twbx).

Conclusion and key recommendations regarding US Public Transportation:

- Shift to sustainable green energy, with better efficiency, e.g. Illinois shift from diesel to bio-diesel.
- Change the structure of expenditure: maintain the level of operational labor, try to improve on high-skilled labors that support the transformation prior to the start of infrastructure funding.
- Given the low current demands as a result of the global pandemic, begin to punctually restructure transportation modes in preparation for post-pandemic demands.

Overall Project Lessons Learned and Recommendation

- **LessonsLearned**
  - Even when the tools are compatible with all operating systems (OS), unique/specific problems may still arise for each OS
  - Difficulty in importing data directly from CSV to MySQL, with regards to capacity and file conversion format
  - Source-to-target mapping requires significant time for completion
  - Due to time constraints, we need to adjust the approach outside the normal process to incorporate more required data
- **Recommendation**
  - Ensuring the usability across the operating systems is crucial before moving forward with the tools and methods
  - For data input to MySQL, better to convert the source CSV files to SQL scripts
  - Allocate sufficient time for table transformation (source-to-target mapping) during ETL
  - It is possible to use an additional approach/workaround to add more information

References

- https://www.apta.com/research-technical-resources/transit-statistics/ntd-data-tables/
- https://www.transit.dot.gov/ntd/ntd-data
- https://www.apta.com/research-technical-resources/research-reports/