BusTrust: DE300 Project Proposal

Livia Fingerson and Pedro Lacombe Farina

Introduction

The Chicago Transit Authority (CTA) provides nearly one million rides per day, offering transportation by both bus and train across the city of Chicago and its surrounding areas (Chicago Transit Authority, 2017). With 127 bus routes and 10,588 bus stops, the network of CTA buses plays a critical role in the mobility of large numbers of Chicago residents. Research shows that among CTA riders in 2021, around 70% work in "on-site" jobs, meaning essential industries such as education, retail, and healthcare (Regional Transportation Authority, 2021). Unpredictable or inconsistent bus services can have a disproportionate impact on people who rely the most heavily on public transportation, making it a critical task to develop methods for assessing and improving bus route reliability.

This project aims to address this challenge by using real-time CTA Bus Tracker data to create a data pipeline that produces a CTA Bus Trust Score for each bus route. This score will take several factors into account, including predicted vs actual arrival times, delays, route performance consistency, and the frequency of ghost buses. By combining these metrics into a single score for each route, this project will provide a clear and quantifiable measure of the reliability across routes in the CTA Bus network.

The resulting pipeline will be both scalable and reusable. It will automatically intake new data, preprocess it to detect bus arrivals/delays, and store this information in a structured format. This system will allow CTA stakeholders, including operations teams, city officials, and users, to assess the performance of individual bus routes over time, identify areas for improvement, and work towards a more reliable commuter bus experience in Chicago.

Data sources

The project will rely on three main data sources: the CTA Bus Tracker API, the General Transit Feed Specification (GTFS) Static Feed, and the CTA Bus Routes Daily Ridership. These will be important to understand the observed bus arrival and departure times for each route/stop, the scheduled arrival and departure times for each route and stop, and the average ridership for each route.

The CTA Bus Tracker API is a developer tool available through the Transit Chicago website, which is the official Chicago Transit Authority data platform (Chicago Transit Authority). Data is available through an API after registering for an account on the Transit Chicago website. In this project, this is useful for us to collect information to create the data pipeline, as well as to understand the observed bus arrival and departure times for each route/stop. It will help us identify "ghost buses" and deviations from the official schedule. The data is organized in five endpoints to retrieve real time positions, static metadata, stop locations, and route geometry, available as JSON files updated every 30 seconds. It provides mainly the vehicle id, timestamp, latitude and longitude, speed, route, route direction, trip ID, arrival predictions to the next stop, the list of bus routes, and the stop information (ID, latitude, longitude) per route per direction. Consequently, it would contain between 40 and 60 million rows and consume 10 to 20 GB in JSON files after accumulating data for 14 days (estimated, as it varies depending on the number of bus lines and active vehicles).

Meanwhile, the GTFS Static Feed, published by the CTA and defined on the Google Transit Website, contains ten tables available in .txt files (Google, n.d.). We will use the trips.txt and stop_times.txt files to retrieve route IDs, trip ID, direction ID, scheduled trip ID, scheduled arrival time, scheduled departure time, and stop ID. These will be important to understand the scheduled arrival and departure times for each route and stop, and combined with the CTA Bus Tracker API, it will allow us to quantify delays that later serve as a parameter for the trust score construction. trips.txt contains 49,936 rows and nine columns, and stop_times.txt contains 2,996,069 rows and eight columns.

Finally, the CTA Bus Routes Daily Ridership contains data for the total daily ridership on a per-route basis since 2001, available through the City of Chicago Data Portal (City of Chicago, n.d.). This dataset is especially important for us to calculate the crowdedness level of each route given its capacity, allowing this to be another metric to generate the trust scores. It contains 1,092,474 rows to describe the route, date, day type (weekday, saturday, and sunday/holidays), and number of rides.

Goal Definitions

This project will focus on three main goals:

1. Create a functional database to store CTA bus data;
2. Establish a procedure to identify and input missing bus position and arrival time data;
3. Develop a CTA trust score for each route depending on day type (weekday, Saturday, and Sunday/Holidays).

Currently, the only way to access CTA bus data is through the live API tracker, which can help us see the current position of active buses but does not inform any historical data on the timeliness and reliability of bus routes. As many Chicagoans rely on public transportation for commuting, there exists a need for users to understand any trends. We intend to develop a functional pipeline that extracts data from the CTA Bus Tracker API and stores it in Amazon Simple Storage System (S3), which keeps and retrieves any amount of data from anywhere on the web with a specified frequency. Therefore, we will use the database system technique learned in class. Moreover, as there is a large amount of data, PySpark will allow us to do fast parallel reads, schema handling, joins, and easier data manipulation, related to the distributed computing framework.

Secondly, as we can only periodically collect data and will consider a 14-day data collection process, we will need to identify and input missing bus position and arrival time data based on the cross-comparison between the API and the GTFS Static Feed. This will require the use of the imputation methods discussed, as there will be multiple types of missing data. For example, we expect missing not at random (MNAR) behavior in cases when observations are related to service disruptions, delays, and ghost buses. In addition, the proposed datasets do not inform whether a specific occurrence is a ghost bus or a delay, so we will need to establish a procedure to properly categorize them through time difference thresholds.

Finally, the datasets can provide information on observed and scheduled departure and arrival times but cannot inform any raw performance metrics, such as reliability of a route. Consequently, to develop the CTA trust score, we will need to understand how deviations from the scheduled departure/arrival times, frequency of ghost buses, and crowdedness level impact the bus system reliability. Since there is a big volume of data, parallelized computing will also be necessary, requiring the use of a distributed computing framework.

Overall, the proposed datasets can answer 1. the current position of CTA buses, 2. the schedule for each bus route and direction, 3. the daily ridership for each route by day type. However, they still lack 1. historical data on position and delays, 2. differentiation between the type of delay (ghost bus or standard delays), and 3. performance metrics for each bus route and direction by day type, which are the exact questions this project aims to target.

References:

Chicago Transit Authority. (n.d.). CTA Bus Tracker [API]. https://www.transitchicago.com/developers/bustracker/

Chicago Transit Authority. (n.d.). CTA GTFS (General Transit Feed Specification) [Data set]. https://www.transitchicago.com/developers/gtfs/

Chicago Transit Authority. (2017). CTA Facts at a Glance. CTA. https://www.transitchicago.com/facts/

City of Chicago. (n.d.). CTA Ridership - Bus Routes Daily Totals by Route [Data set]. https://data.cityofchicago.org/Transportation/CTA-Ridership-Bus-Routes-Daily-Totals-by-Route/jyb9-n7fm/about_data

Google. (n.d.). *Google Transit (GTFS) Schedule Reference and Differences. https://developers.google.com/transit/gtfs/reference?csw=1*

Survey shows current riders are more likely to be essential workers or minorities, feel transit is safe.

(2021, April 19). Regional Transportation Authority.

https://www.rtachicago.org/blog/2021/04/19/survey-shows-current-riders-are-more-likely-to-be-essential-workers-or-minorities-feel-transit-is-safe