CSP 554 – Big Data Technologies
Project Proposal

Title:

Transit Trends: Analyzing and Visualizing Chicago's CTA Data

TEAM MEMBERS

Gopi Shankar Ravady
gravady@hawk.iit.edu
A20527679

Pranjali Deshmukh
pdeshmukh3@hawk.iit.edu
A20527773

Mihira Gudimetla
mgudimetla@hawk.iit.edu
A20527166

Tejaswini Vishwanath
tviswanath@hawk.iit.edu
A20536544

Yash Amin
yamin@hawk.iit.edu
A20517652

Transit Trends: Analyzing and Visualizing Chicago's CTA Data

Project Statement:

Given: Apply a range of big data tools to explore some interesting data set and derive insights from it. Ingest data, apply transformations, profile the data, summarize it, visualize it.

Abstract:

Public transportation systems like the Chicago Transit Authority (CTA) play a crucial role in urban mobility. Analyzing the CTA dataset provides valuable insights into various aspects of service delivery, rider behavior, and system performance. This project aims to analyze the CTA dataset comprehensively, with a specific focus on crowd analysis, event logs, weather conditions, and other relevant factors. By leveraging data-driven approaches, the project seeks to uncover patterns, trends, and correlations that can inform decision-making processes and improve the overall quality of public transportation services in Chicago.

Introduction:

Public transportation systems play a vital role in urban mobility, and the Chicago Transit Authority (CTA) is no exception. Our analysis will encompass multiple dimensions of CTA ridership, including daily and monthly ridership totals, average weekday boardings, and historical trends. We will explore factors such as day type (weekday, Saturday, Sunday/holiday), route popularity, and station entries to understand peak travel times, popular routes, and areas with high transit demand. Additionally, we'll delve into spatial analysis by examining bus stop locations and their relationship with ridership patterns across different neighborhoods and wards in Chicago. By delving into these aspects of CTA ridership, alongside focusing on crowd analysis, event logs, weather conditions, and other general insights, this project aims to provide a comprehensive understanding of CTA operations and facilitate informed decision-making for optimizing public transportation services in Chicago.

Proposed Methodology:

1. **Data Collection of CTA Dataset and Impacting Factors:** The first step of our methodology involves collecting the CTA dataset, which includes ridership data, service schedules, vehicle locations, fare transactions, and more. Additionally, we will gather datasets related to impacting factors such as event logs, weather conditions, and other relevant data sources.

   Data Source: https://www.kaggle.com/datasets/chicago/chicago-transit-authority-cta-data/versions/8?resource=download

2. **Data Preprocessing/Transformation:** Once the datasets are collected, the next phase will focus on data preprocessing and transformation. We will explore two potential options for this process: Kafka and AWS services. Kafka offers real-time data streaming capabilities, allowing for efficient data preprocessing and transformation. On the other hand, AWS services provide a scalable and cost-effective cloud-based solution for data processing tasks. After evaluating both options, we will finalize the one that best suits our project requirements.

3. **Data Visualization:** After preprocessing and transforming the data, we will proceed with data visualization using Tableau. Tableau offers powerful visualization tools that enable us to explore various insights derived from the CTA dataset and impacting factors. We will create interactive dashboards and visualizations to present key findings related to crowd analysis, event logs, weather conditions, spatial analysis, and other relevant insights. These visualizations will facilitate a better understanding of CTA operations and support informed decision-making for optimizing public transportation services in Chicago.

Our goal is to provide the Chicago Transit Authority and city planners with practical insights through this initiative, enabling them to improve the everyday commuting experience for thousands of locals and tourists alike. We can go through vast amounts of transit data to find hidden patterns and trends by utilizing AWS, PySpark and tableau's powerful features. This will help us identify busy times, well-traveled routes, and places that require better transit services. Our goal is to improve Chicago's public transportation system so that it is more effective, convenient, and available to all. This initiative aims to improve people's lives and shape urban mobility in the future by using data-driven decision-making, not just numbers.

References:

1. Singh, P. K. (2021, December 9). Manage Data with PySpark. Apress eBooks. https://doi.org/10.1007/978-1-4842-7777-5_2
2. Ranganathan, G. "Real time anomaly detection techniques using pyspark frame work." *Journal of Artificial Intelligence* 2, no. 01 (2020): 20-30.
3. https://medium.com/@anitateladevalapalli777/big-data-analytics-project-using-kafka-pyspark-and-tableau-b1c28b7cebad
4. https://www.kaggle.com/datasets/chicago/chicago-transit-authority-cta-data/versions/8?resource=download