# Project NAIA:
# Passenger Flow Predictions in Interconnected Multiline Networks

By WONG, Nathan Shih Hao
Advised by: Prof. CHEUNG, Tsz Him James
Supervised by: Hitachi Rail GTS Hong Kong

THE DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

HITACHI

## Introduction


Figure 1: Passenger Flow Analysis (Entry Records) from 22/05/2017 to 29/05/2017
Actual (Green) vs Prediction (Red)

**Project NAIA** is an advanced passenger flow prediction system designed for complex, interconnected transit networks. By leveraging existing infrastructure and data sources—such as **entry/exit logs** and **train schedules**—the system utilizes **machine learning** and **time series forecasting** to effectively model rider behavior while minimizing hardware investments.

By delivering precise origin-destination and station entry predictions, Project NAIA enables transit authorities to implement data-driven strategies that **optimize operational efficiency**, **enhance safety**, and **improve passenger comfort**. These insights allow operators to make informed, proactive adjustments to services, including **optimizing dwell times** for passenger boarding, **managing climate control**, **adjusting train speeds** (speed vs efficiency), and **dynamically directing escalators** based on traffic flow (upward vs downward). Co-op project focuses on the **Kafka to Clickhouse Native Sink** and the **Medium Term Regression Predictions**.

## Objectives

- **Localized Version of the NAIA System:** Develop a system tailored to meet the unique requirements of different regions.
- **Feasibility and Accuracy Validation:** Validate the system to ensure accurate predictions, allowing for actionable insights.
- **Model Optimization:** Explore and optimize various models and techniques within the AI/ML pipeline for enhanced performance.
- **Actionable Strategies:** Identify and formulate actionable strategies derived from data-driven insights that can improve public transport systems.

## Methodology

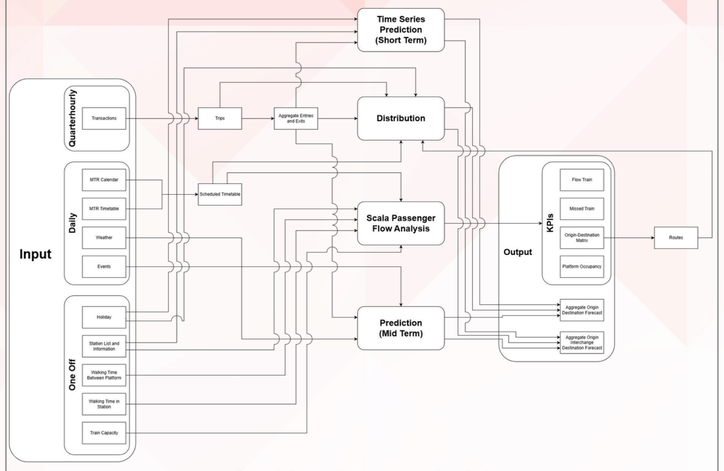


Figure 2: System Architecture and Data Flow

### Development Pipeline

1. **Data Preparation**
- Collect, preprocess, and transform data, incorporating feature selection and one-hot encoding. Data inserted through a Kafka-ClickHouse Native Sink.
2. **Clustering**
- Apply clustering techniques to uncover patterns and group similar origin-destination routes, divided per weekday and weekend.
3. **Mid Term Model Training**
- Train machine learning models on the prepared and clustered data, utilizing an XGBoost model with an 80-20 split.
4. **Mid Term Model Tuning**
- Optimize model parameters to improve performance and accuracy by employing hyperparameter tuning techniques, specifically using Optuna in conjunction with GridSearchCV.
5. **Mid Term Model Evaluation**
- Evaluate the model using daily metrics and origin-destination (OD) measurements, while also benchmarking against data from previous weeks.

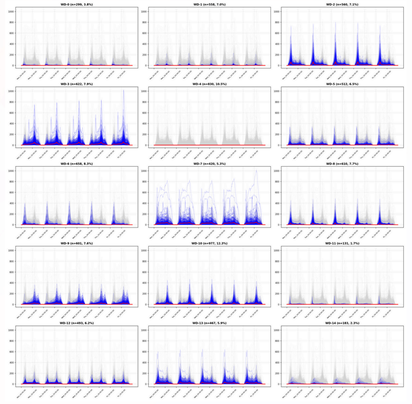

Figure 3: Origin-Destination Route Clustering

## Result

### Result 1 (Normal Monday)

- **1235** RMSE, **908** MAE and **99.9%** R2 value.
- Overall a very good prediction spread out over one day.
- Simple to predict with no external events such as rain, typhoon, public holidays, events. (Normal Monday)
- Two peaks are seen, morning rush and evening rush.
- Represents 90% of all the measured days.

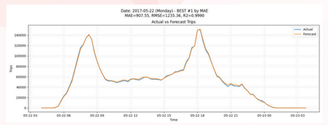

Figure 4: Normal Monday

### Result 2 (Monday, T8 Merbok)

- **25904** RMSE, **12635** MAE and **69.8%** R2 value.
- T8 Merbok called at **5:20pm** on June 12, 2017.
- Happening on a Monday, most commuters are at work.
- Sharp influx of passengers right before normal evening rush show commuters rushing home.
- Represents 1% of all the measured days.

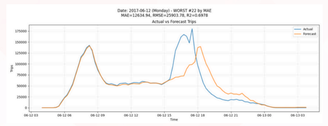

Figure 5: T8 Merbok (Weekday)

### Result 3 (Sunday, T8 Roke)

- **12048** RMSE, **8257** MAE and **69.5%** R2 value.
- T8 Roke called at **9:20am** on July 23, 2017. Subsequently lowered at **1:20pm** on the same day.
- Normal activity resumes at early afternoon, apart from a noticeable decline in ridership among those who are typically returning home.
- Represents 1% of all the measured days.

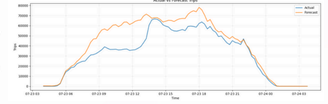

Figure 6: T8 Roke (Weekend)

Figure 7 illustrates the daily distribution metrics for the prediction model. The skewness observed in all metric values indicates that Results 2 and 3 are exceptions rather than typical outcomes.

**Overall, the model performs well in predicting ridership during normal weekdays and weekends; however, it lacks the ability to account for hour-specific events. For instance, a typhoon occurring at 9 AM will impact ridership differently than one at 1 PM. This underscores the need for short-term predictions.**

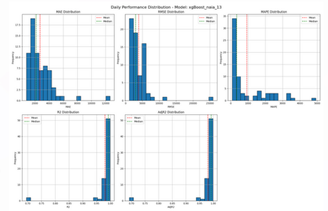

Figure 7: Daily Distribution Metrics

## Conclusion

Throughout Project NAIA, we successfully developed a **medium-term passenger flow prediction model** that achieved **high accuracy metrics** for passenger flow during regular operations. This project showcases the effectiveness of leveraging machine learning and artificial intelligence to uncover both **obvious** and **hidden patterns** within metro systems, demonstrating that **data-driven optimizations** are indeed feasible. However, it is important to note that improvements are still needed, particularly in addressing **hourly fluctuations**, which can be better captured through short-term predictions.

Additionally, the integration of a **ClickHouse-Kafka Native Sink** was also completed during the Co-op Program, although this was not highlighted in this poster.