Department of Computer Science  
The University of Hong Kong  
Final Year Project  

Interim Report  

A Large-Scale Data Analysis System for MTRC and Coronavirus Analysis  

January 22 2022  

Students:  

Chen Jiaying  3035533123  
Hsu Ya-cheng  3035550250  
Huang Hanting  3035534581  
Yan Junchen  3035533288
Abstract
The coronavirus continues to plague the world, whilst Hong Kong keeps struggling to maintain its prosperity. Pandemic control becomes challenging when it comes to a crowded railway system. Understanding how travel patterns and infections might be related helps to give advice on epidemic controls. The project thus aims at exploring the correlation between railway travellers and COVID cases based on a large amount of data. It will facilitate MTR Corporation and other experts to make adjustments to avoid an outbreak. This is an ongoing project. The objectives of the project are to improve the efficiency, scalability, and security. The progress of the project is smooth as scheduled. Our current outcomes are improvements of user interface and implementation of resilience analysis. We have also evaluated our platform’s performances. The next step involves two more advanced data analysis.
Acknowledgment

First, we would like to show our gratitude to the Department of Computer Science, and Prof. Reynold Cheng for giving us a valuable opportunity to conduct the research, and the enormous support they offer all the time.

We would also like to acknowledge Mr. Shivansh Mittal, who assists our team by giving a detailed review of previous work and continuous suggestions throughout the research.

In addition, we would like to thank the MTR Corporation and Dr. Zhou’s research team for the data and modeling ideas they provide. The immense help is like a cornerstone - without that, the research would not have been possible.

Last, we would like to thank Mr. Cezar Cazan, Ms. Grace Chang and Miss Mable Choi for helping us a lot regarding the interim report writing.
# Table of Contents

Abstract .................................................................................................................. I  
Acknowledgement .................................................................................................... II  
List of Figures .......................................................................................................... IV  
List of Tables ............................................................................................................ IV  
Abbreviations .......................................................................................................... V  
1 Introduction  
1.1 Background ....................................................................................................... 1  
1.2 Existing System ................................................................................................. 1  
1.3 Problem Statement ............................................................................................ 3  
1.4 Project Objectives ............................................................................................. 4  
1.5 Outlines ............................................................................................................. 5  
2 Methodologies  
2.1 Platform Scalability – NoSQL .......................................................................... 6  
2.2 Geospatial Analysis - ESRI ............................................................................... 6  
3 Current Progress  
3.1 UI Improvements ............................................................................................... 8  
3.1.1 Informatic Tips ............................................................................................... 8  
3.1.2 Refined Visualization Layout ........................................................................ 8  
3.1.3 Colour Scale ................................................................................................ 9  
3.1.4 Other Improvements .................................................................................... 10  
3.2 Resilience Data Analysis .................................................................................. 10  
3.2.1 Data Processing ........................................................................................... 10  
3.2.2 Interactive Web Application ......................................................................... 11  
3.2.3 Resilience During Each Surge ..................................................................... 12  
3.2.4 Resilience Group by Week .......................................................................... 13  
3.2.5 Hodrick-Prescott (HP) Filter ....................................................................... 13  
3.2.6 Resilience by Station .................................................................................. 14  
3.3 Performance Analysis ....................................................................................... 15  
3.3.1 Travel Pattern .............................................................................................. 15  
3.3.2 Station Density ............................................................................................. 15  
3.3.3 Passenger Volume ....................................................................................... 16  
3.3.4 Discussion .................................................................................................... 16  
4 Future Plans  
4.1 Advanced Data Analysis ................................................................................... 18  
4.1.1 Essential Riders and Choice Riders .............................................................. 18  
4.1.2 Extreme Riders ........................................................................................... 18  
4.2 Project Schedule ............................................................................................... 19  
5 Conclusions ......................................................................................................... 20  
References ............................................................................................................ 21
List of Figures

Figure 1.1 Technology stack ................................................................. 2
Figure 1.2 Existing platform login interface ........................................... 2
Figure 1.3 Existing platform dashboard ................................................ 3
Figure 2.1 Outgoing number of MTR passengers and confirmed cases in April 2020 7
Figure 3.1 Demonstration of informatic tips .......................................... 9
Figure 3.2 Comparison between previous and current visualization page .... 9
Figure 3.3 Colour scale on station density map ..................................... 10
Figure 3.4 Resilience of the MTR system visualization ............................ 11
Figure 3.5 Resilience by station visualization ....................................... 12
Figure 3.6 COVID-19 development history in Hong Kong ....................... 12
Figure 3.7 Resilience of MTR system during First, Second, and Third Surge ... 13
Figure 3.8 Trip amount and Resilience visualization by days in a week ........... 13
Figure 3.9 Daily resilience of MTR system after applying HP filter ............... 14
Figure 3.10 Runtime to render travel pattern ....................................... 15
Figure 3.11 Runtime to render station density ...................................... 16
Figure 3.12 Runtime to render passenger volume ................................. 16

List of Tables

Table 3.1 Five stations with the lowest average rate of recovery .................... 14
Table 3.2 Five stations with the highest average rate of recovery .................. 15
Table 4.1 Definition for different types of Extreme Riders .......................... 18
Table 4.2 Project Schedule ............................................................... 19
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID</td>
<td>Novel Coronavirus Disease</td>
</tr>
<tr>
<td>ESRI</td>
<td>Environmental Systems Research Institute</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>HK</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>HKCHP</td>
<td>Hong Kong Centre for Health Protection</td>
</tr>
<tr>
<td>JWT</td>
<td>JSON Web Token</td>
</tr>
<tr>
<td>MTRC</td>
<td>Mass Transit Railway Corporation</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>UI</td>
<td>User Interface</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Background

After the first COVID-19 case was confirmed in Hong Kong on 23 January 2020 (Cheung, 2020), Hong Kong has been suffering from the outbreak of COVID-19. To control the rapid evolution of the pandemic, social distancing measures were implemented by the Hong Kong government, such as limiting the number of participants in a fitness class or prohibiting group gatherings (Ancheta, 2021). However, it is almost impossible to avoid exposure in the public territory, especially in public transportation, where distancing rules are hard to maintain and people can be easily infected in a confined space with high population density.

Hong Kong residents rely heavily on the Mass Transit Railway (MTR) for transportation. MTR Corporation Limited (2020) disclosed in its annual report that it comprised a 45.3% market share of Hong Kong’s public transportation in 2020, which means that MTR passengers’ data can be deemed to reflect Hong Kong’s population in public transit. Therefore, this project aims to use the MTR passenger dataset to study how the virus impacts passengers’ mobility when different measures are implemented or when the pandemic is in different states, as well as how passengers’ mobility affects the development of COVID-19 in return.

The term big data refers to the usage of large and complex data sets, typically unstructured data, to pinpoint significant characteristics or trends that were not discovered previously (Boyd & Crawford, 2011). Compared to traditional methods of sampling data into smaller volumes, big data involves an important step of extracting value from an extensive dataset. Through this, one can gain the advantage of obtaining additional insights. Consequently, this can lead to more informed future decisions and predictions.

In collaboration with the Hong Kong Mass Transit Railway Corporation (MTRC), this project intends to utilize big data techniques on MTR passenger traffic data from January 2019 to January 2021 to conduct data analysis and generate appropriate visualizations.

1.2 Existing System

In the 2020-2021 academic year, another FYP group also worked on the same project. Within the one-year time, they finished the preliminary data processing and established a completed infrastructure. Therefore, we already have a completed data pipeline, a backend with data models, and an interactive web application. The existing system utilizes Django for backend, React for frontend, and MySQL for the database (see Figure 1.1). REST APIs are set up to provide a stateless, uniform interaction.
between the client and server. SSH channel is used to access the data stored in the MySQL database on the server, hence ensuring the safety of the raw data.

The database and backend are located on the “HinCare” server, which is shared with another project. The web platform is currently deployed on the FYP server and exposes the URL http://fyp20035s1.cs.hku.hk/ for client usage. The login process uses username and password to verify the user identity and returns a JWT authentication token (see Figure 1.2). The platform has three sections, namely query, visualization, and analysis (see Figure 1.3).
The platform can accomplish data filtering and querying for both COVID-19 confirmed cases and the MTR passenger database. Some of the basic data visualization tasks can be realized on the frontend platform, including travel pattern, station density, and passenger volume. The previous team also worked on some contact and behavior-based analysis such as “Someone like you” and “Senser individuals”. These research topics were carried out together with the transportation team, as they helped identify and visualize the indicators for COVID-19 research.

1.3 Problem Statement

Thanks to last year’s effort, this project initiated with a platform that can provide basic functionality. However, some issues arose when putting last year’s product into practical use.

First, as this project is dealing with billions of transaction records of data, the existing database built by the last team could only accommodate one user simultaneously. When one researcher is conducting data analysis utilizing the system, others cannot log in to the system in the meantime, or else the system will break down. Since this project intends to provide data services for researchers interested in the correlation of COVID-19 data and MTR passengers’ data, users may come from varied departments and various institutions, such as transportation departments from universities, industry researchers, public health sectors from governments, etc. It is critical to maintain the system’s scalability so that different users can retrieve the data efficiently even at the same time.

Second, the previous webpage gave users easy access to raw data. It was convenient for researchers to download the desired raw data and conduct analysis outside the
platform. Nevertheless, the confidentiality of the data occupies a predominant position when providing data services. Users should not have access to raw data, especially the access to download the raw data. Therefore, authority management is needed: data should be organized according to their security level and displayed to users holding corresponding authority levels. Furthermore, due to the prohibited access to downloading raw data, the functionality of data visualization and data analysis provided by this platform should be enhanced so that users can achieve their expectations within the same platform even without touching the raw data.

1.4 Project Objectives

Through collaboration with MTRC, this project has 13.5GB of MTR passengers’ data to analyze, ranging from January 2019 to January 2021. Associated with the open-source COVID-19 data retrieved from Hong Kong Centre for Health Protection (HKCHP), this project aims to improve the previous platform and provide a unified platform with highly ensured efficiency, scalability, and security, which can accommodate more advanced data visualization and data analysis for researchers to study the relationship between MTR passengers and COVID-19.

For efficiency, our goal is to shorten the time required for querying and retrieving the data while allowing more users to conduct research on the platform at the same time. If conditions permit, we would like to move our backend to an individual server to ensure the computation power.

As for scalability, we aim to provide a more unified platform with new functionality, so that more data analysis and visualization tasks can be conducted directly on our platform. Specifically, this project will perform mobility analysis and geospatial analysis, which will be carried out using the ArcGIS system provided by ESRI.

The security and safety of the passenger data provided by MTR are of vital importance to our project. In normal circumstances, researchers outside our team should not have direct access to the raw data. While JWT authentication is secure enough for our current demands, we aim to smooth the login process to avoid redirecting. We are also looking at ways to give partial permission to certain users so as to meet their research needs while protecting the privacy of the database.

Our project team will also work closely with researchers from the transportation team in terms of technical support. Our platform should become a useful tool for the researchers to explore different research topics on the relationship between COVID-19 and MTR passengers.
1.5 Outlines
The remainder of this report proceeds as follows. First, methodologies of improving system scalability and presenting geospatial analysis will be elaborated in Chapter 2.1 and Chapter 2.2. Following that, current improvements to perform a more user-friendly UI will be demonstrated in Chapter 3.1, including refinements in informative tips, visualization layouts, and color scale. Chapter 3.2 will introduce the current progress on Resilience Data Analysis, which is to examine the level of recovery in a specified timeframe. Next, in Chapter 3.3, performance analysis regarding the runtime of the current system will be conducted. This report will close with detailed future plans on advanced data analysis and conclusions summarizing this interim report.
2 Methodologies
This chapter proposed two methodologies. The first one is to enhance the platform scalability by migrating the database from MySQL to NoSQL, and the second one is to perform geospatial analysis by utilizing ArcGIS software.

2.1 Platform Scalability – NoSQL
NoSQL is a form of database with a focus on scaling and fast queries, allowing for frequent application changes. MySQL is another form of database developed with a focus on avoiding data duplication by joining schemas and hence saving storage.

The database of this project will be migrated to a NoSQL database to improve the platform scalability. Since this project must manage billions of transactions of unstructured records, scalability is significantly critical to this big data analysis project to ensure the efficiency and consistency of data querying. The previous team chose MySQL to accommodate the 13.5GB data provided by MTRC, which is the primary choice for general purposes and is widely adopted in the industry. MySQL experts reduce duplicate transactions in the database by creating relational schemas and reducing storage occupation (Schaefer, 2021). However, it appeared to have an extremely slow performance in data querying, especially when dealing with a large dataset. One of this project’s dominating functionality is data query, which also performs in data visualization and analysis. Therefore, it is essential to reduce the time on data queries and improve the efficiency in retrieving data.

NoSQL performs significantly better than MySQL, as data in its databases is stored in a very flexible schema, which is optimal for data querying (Schaefer, 2021). Queries in MySQL, where data is stored in the rigid schema, require time-consuming activity – joining. While in NoSQL, the omission of joining schemas makes queries significantly fast. Although NoSQL requires larger storage as it has duplication issues, storage nowadays is cheap, so NoSQL has no major drawback for this project.

2.2 Geospatial Analysis - ESRI
Geospatial analysis is the most critical part of the visualization sector, as the relationship between MTR passengers and COVID-19 is mostly location-based. This project leverages the ArcGIS system powered by ESRI to perform geospatial analysis, which is more than capable of supporting the projection of our complex and substantial dataset to present geospatial visualization.

We will use the MTR data to investigate the density of passengers in each MTR station and the travel density between stations in the MTR network map.
Popular travel routes will be sketched onto the maps with their endpoints at the starting and ending MTR Station. In addition to this, population density can also be projected to the maps as the navy circles in Figure 2.1, with its size representing the density.

In order to integrate the COVID-19 confirmed cases into the same spatial dimension, we will first retrieve the geocode of the cases’ locations (i.e. their residential buildings) using the geolocation API available in the ArcGIS software. Then, using the Google Maps API, we attempt to match the building to the closest MTR station. As a final step, we will project the density of confirmed COVID-19 cases in the vicinity of each MTR station onto the same map.

There will be a considerable amount of analysis that can be conducted on the projected data. For example, by looking at how the travel patterns of MTR passengers changed over a specific lockdown, how COVID-19 affected the mobility of Hong Kong’s population can be approximated (Zhou, 2021). Furthermore, different districts can be analyzed separately to evaluate how different class of population or how different the purpose of traveling was impacted.
3 Current Progress

3.1 UI Improvements
This part introduces several approaches done on UI to improve the overall user-friendliness of the platform. Three major improvements are informatic tips, refined visualization layout and the colour scale. Other improvements will also be discussed briefly.

3.1.1 Informatic Tips
Informatic tips serve as supplementary information to ensure safe and smooth use of the platform. There are two main purposes. The first is to avoid invalid data input by reminding one of the correct input formats. The platform would spend extra time processing an invalid input, which may add an unnecessary burden on the backend server. Therefore, informing one about expected output will prevent one from lowering the overall platform performances. The second is to let the first-time user be familiar with available functionalities in a shorter time. It will also help current users to master a new functionality when there is an update of the system.

![Figure 3.1 Demonstration of informatic tips](image)

Bootstrap Tooltip component is adopted to realize the functionality. As shown in figure 3.1, essential information will be displayed upon hovering on the component.

3.1.2 Refined Visualization Layout
Besides focusing on the supplementary information, we made efforts to improve the user experience when performing visualization tasks.
Figure 3.2 Comparison between previous and current visualization page

Figure 3.2 provides an example of a modification. Before modification, one should roll down to another page to set or change the data input, as shown in Figure 3.2 (1). To reduce redundant operations, the input table is placed on the right-hand side of the map, as shown in Figure 3.2 (2). However, the modification reduces the display area of the map. To further improve, the input table is further modified to the one in Figure 3.2 (3), which can be folded to maintain the original map size. One more modification on visualization improvement, the colour scale, will be discussed in the next subsection.

3.1.3 Colour Scale
A colour scale is a way to allow clearer visualization. Although data has been filtered and organized before displaying, it is sometimes difficult to distinguish between different display components when having a large display quantity. Adding different colours helps mitigate the issue.
The improvement is demonstrated with an example of a station density map. Figure 3.3 shows a station density map with a colour scale. One can notice that having differences will help separate two overlapping clusters (circles). One can also determine the cluster with the larger station density by selecting clusters with a darker colour.

The colour scale can be further modified. More colours will be needed for displaying larger data quantities. There are two possible ways to improve. One is to estimate the most commonly used data quantity and to set a colour scale accordingly. The other is to adjust the colour scale so that the number of colours is proportional to the number of clusters.

3.1.4 Other Improvements
Besides the major UI improvements, other improvements have been made to facilitate accessibility. Firstly, the unnecessary backend login portal that exists previously has been removed. Currently, one can access the data query function without an extra login procedure. Besides, the wordings are aligned across the platform to avoid the misinterpretation of results.

3.2 Resilience Data Analysis
3.2.1 Data Processing
Resilience is the ability of the system to maintain its level of service or to restore itself to that level of service in a specified timeframe. This is an indicator developed by Dr. Zhou from the transportation department. The resilience ability of the MTR network and each station is measured by the rate of recovery. The rate of recovery can be calculated using the total trip amount of a certain time frame in 2020 divided by the
total trip amount in a corresponding time frame in 2019. A lower rate of recovery shows that the event has a more significant impact on the travel volume. Hence, the higher the rate of recovery, the better the resilience of the system is.

\[
\text{Rate of Recovery} = \frac{\text{TRIP AMOUNT (time frame in 2020)}}{\text{REFERENCED TRIP AMOUNT (corresponding time frame in 2019)}}
\]

To process the data, we use the SQL command to generate CSV files and new tables. We first select all the entries in the 2020 time frame, get the total count and record it in the column TRIP_AMOUNT. The entries for the 2019 time frame are processed in the same way and recorded in the column REFERENCED_TRIP_AMOUNT. The rate of recovery is then calculated by dividing the TRIP_AMOUNT by REFERNECED_TRIP_AMOUNT. We can also group the data by station and study the difference between each station. After data processing, we generate four new tables for the resilience analysis, namely ResilienceTotal_Daily, ResilienceByStation_Daily, ResilienceTotal_Weekly, and ResilienceByStation_Weekly.

### 3.2.2 Interactive Web Application
Currently, we have processed the data from January 2020 to August 2020. Figure 3.4 shows the resilience of the whole MTR network using weekly data. The blue line indicates the referenced trip amount, as in the trip amount in 2019. The green line indicates the trip amount from 2020. Although there are some outliers in the graph, we can observe a rough trend of the rate of recovery. Figure 3.5 shows the rate of recovery of each station. The darker green indicates a larger rate of recovery. It can be observed that the trend of individual stations is roughly the same. However, the rate of recovery differs between stations. Both visualizations are available on our web application. Users can select the start date, end date, and daily/weekly mode of the visualization based on their needs.

![Figure 3.4 Resilience of the MTR system visualization](image-url)
3.2.3 Resilience During Each Surge
For COVID-19 in Hong Kong, there have been three surges from January 2020 to August 2020 (Figure 3.6).

i) First surge: 23 Jan 2020 - 17 Feb 2020
ii) Second surge: 16 Mar 2020 - 11 Apr 2020
iii) Third surge: 1 July 2020 - 31 Aug 2020.

Therefore, we looked into each surge to study how the MTR system performed during the above three periods (Figure 3.7). During the first surge, when HK residents were first made aware of the coronavirus, the number of trips decreased significantly. The rate of recovery drops from around 90% to approximately 50% and stays at around 50% for an extended period of time. For both second and third surges, there is a slight drop in the rate of recovery from around 60% to 40% during the surge. But the rate of recovery returned to around 60% near the end of the surge. This observation suggests that the outbreak of COVID-19 made people reduce their travel amount, especially...
when it was first discovered. The following two surges also have an impact on traveling amount, but not as severe as the first surge.

Figure 3.7 Resilience of MTR system during First, Second, and Third Surge

3.2.4 Resilience Group by Week
During the plotting of daily data, we noticed that the trip amount usually has a spike on Fridays and a drop on Saturdays and Sundays. The pattern is aligned with our intuition because there are much fewer commuters during weekends. To further study this pattern, we did another analysis by grouping the data by days of the week. Figure 3.8 (1) shows that the cycle exists both before and after COVID happens. Figure 3.8 (2) shows the pattern after dividing the trip amount by its referenced amount. The same pattern that the rate of recovery is lower on weekends than on weekdays still exists, meaning the reduction of the trip amount is more significant on the weekends. This suggests that more travelers are avoiding unnecessary trips during weekends.

Figure 3.8 Trip amount and Resilience visualization by days in a week

3.2.5 Hodrick-Prescott (HP) Filter
HP filter is commonly applied during the analysis to remove short-term fluctuations associated with the cycle (Will Kenton, 2021). Removal of these short-term fluctuations reveals long-term trends. Since the daily resilience data have the weekly cycle mentioned in section 3.2.4, we can apply the HP filter to smooth the data. Figure 3.9 shows the result of applying the HP filter to the daily resilience data. After smoothing the line, we can clearly observe three prominent downward trends, around
Jan 15, Mar 11, and Jul 1, respectively. The observation aligns well with the three major surges from Jan 2020 to Aug 2020.

Figure 3.9 Daily resilience of MTR system after applying HP filter

3.2.6 Resilience by Station
Besides studying the resilience of the MTR system, we can also explore the resilience data by the station. We take the average of the trip amount and the referenced trip amount of all stations. Different stations were affected by COVID-19 somewhat differently. The five stations in Table 3.1 took the strongest hit. We can provide some reasonable explanation after a little research. Trip amounts from Lo Wu and Lok Ma Chau decreased a lot because of border control. Disneyland Resort, Racecourse, and Ocean Park lost most of their passengers because of the recession in tourism. Meanwhile, the five stations in Table 3.2 were not significantly affected by COVID-19. This might be because the residents who live near these stations are very dependent on MTR for traveling.

Table 3.1 Five stations with the lowest average rate of recovery

<table>
<thead>
<tr>
<th>Station</th>
<th>Trip Amount</th>
<th>Referenced Trip Amount</th>
<th>Rate of Recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lo Wu</td>
<td>7794.31</td>
<td>79841.26</td>
<td>9.76%</td>
</tr>
<tr>
<td>Lok Ma Chau</td>
<td>22072.45</td>
<td>211490.15</td>
<td>10.44%</td>
</tr>
<tr>
<td>Disneyland Resort</td>
<td>6358.28</td>
<td>27209.51</td>
<td>23.37%</td>
</tr>
<tr>
<td>Racecourse</td>
<td>2699.67</td>
<td>6769.81</td>
<td>39.88%</td>
</tr>
<tr>
<td>Ocean Park</td>
<td>12811.21</td>
<td>31599.13</td>
<td>40.54%</td>
</tr>
</tbody>
</table>
Table 3.2 Five stations with the highest average rate of recovery

<table>
<thead>
<tr>
<th>Station</th>
<th>Trip Amount</th>
<th>Referenced Trip Amount</th>
<th>Rate of Recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nam Cheong</td>
<td>48525.17</td>
<td>52005.90</td>
<td>93.30%</td>
</tr>
<tr>
<td>Tsuen Wan West</td>
<td>66285.51</td>
<td>78210.76</td>
<td>84.75%</td>
</tr>
<tr>
<td>LOHAS Park</td>
<td>40133.19</td>
<td>47616.97</td>
<td>84.28%</td>
</tr>
<tr>
<td>Wong Chuk Hang</td>
<td>31178.79</td>
<td>39376.48</td>
<td>79.18%</td>
</tr>
<tr>
<td>Kam Sheung Road</td>
<td>32425.84</td>
<td>41117.56</td>
<td>78.86%</td>
</tr>
</tbody>
</table>

3.3 Performance Analysis
The runtime was computed via chrome developer tools, where the time taken to send requests, to wait, and to download content were recorded automatically by the system.

3.3.1 Travel Pattern
This metric was precomputed in the “travel pattern by day” table. With google developer tools, the runtime for 1 month, 3 months, 5 months, and 7 months with two starting dates were recorded and presented in Figure x.x respectively. A watershed can be spotted at 2 months, where a timespan shorter than 2 months would complete rendering in 10 seconds, but that longer than 2 months all consistently complete rendering at about 2 minutes.

![Travel Pattern Runtime](image)

Figure 3.10 Runtime to render travel pattern

3.3.2 Station Density
This metric was precomputed in the “station density by day” table. With google developer tools, the runtime for 1 month, 3 months, 5 months, and 7 months with two starting dates were recorded and presented in Figure x.x respectively. These tasks with various timespan were all completed in within 1 second.
3.3.3 Passenger Volume
This metric was precomputed in the “passenger volume by day” table. With google developer tools, the runtime for 1 month, 3 months, 5 months, and 7 months with two starting dates were recorded and presented in Figure x.x respectively. Tasks were all completed within 1 second, presenting an increasing trend as timespan lengthened.

3.3.4 Discussion
Even though the three metrics were precomputed in tables, aggregation was still needed to sum up the number from each day. The long runtime taken for travel pattern likely resulted from the complexity of its grouping requirement. For travel pattern, the
precomputed table needed to be further group by two variables, namely entry station and exit station; for station density, grouping by station should be enough; for passenger volume, card type was the grouping variable. There are 96 MTR stations currently, and considering the size of our trip records, grouping requirement of $O(n^2)$ complexity would likely be a heavy burden on server. The potential reason for such system lag may be due to the volume of data and its computation can’t fit into buffer, whose capacity was equivalent to the memory needed for computation of the 2-month timespan. Those with timespan shorter than 2 month can retrieve data directly from the buffer, and that longer would have to retrieve hard disk, resulting in a consistent increased runtime.
4 Future Plans

4.1 Advanced Data Analysis
Apart from resilience, there are some other indicators that we can investigate. We have done some preliminary analysis on the following two indicators. With the identification of essential riders, choice riders, and different types of extreme riders, we can perform further analysis and visualize the Spatio-temporal distribution of these riders. This analysis will help differentiate the impact of COVID-19 has on different types of riders. Also, identifying essential stations can help decide the focus of epidemic control and prevention.

4.1.1 Essential Riders and Choice Riders
The first one is essential and choice riders. Essential riders are those who stick to their original traveling patterns. Choice riders are those who instantly respond to the outbreak incidents of COVID-19 and change their travel behavior.
To ensure our data’s accuracy, we want to exclude those occasional travelers. So we first select a week before the COVID outbreak and choose those riders who have traveled more than three days in a week, which we considered frequent MTR users.
For the second step, we compute the travel frequency and trip length of those frequent MTR users, both before and after the COVID outbreak. After that, we can compute the difference between the two travel frequency numbers. Those with a huge difference are considered choice riders as they have means other than MTR for travel. On the other hand, those with lesser differences are considered essential riders because they stick to MTR even after COVID. Furthermore, we can also compute the difference between before and after COVID-19 for each station’s income and outgoing trips to identify the essential stations.

4.1.2 Extreme Riders
Extreme riders are another indicator we plan to study, which are travelers who commute too early, too late, or last a long time. There are four different types of extreme riders: namely early birds, night owls, tireless itinerants, and recurring itinerants. Table 4.1 gives the definition of each type. We identify those travelers because they are often located on the edge of a city, and employees need to spend more commuting time or start earlier. We can analyze if COVID impacts their travel pattern differently.

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early birds (EBs)</td>
<td>First trip &lt;= 6AM</td>
</tr>
<tr>
<td>Night owls (NOs)</td>
<td>Last trip &gt; 10PM</td>
</tr>
<tr>
<td>Tireless itinerants (TIs)</td>
<td>Length of one trip &gt; 1.5 hour</td>
</tr>
<tr>
<td>Recurring Itinerants (RIs)</td>
<td>&gt;= 30 trips on weekdays of a week</td>
</tr>
</tbody>
</table>
4.2 Project Schedule

Our further plan will adhere to the project schedule in Table 4.2 but is subject to change based on our implantation progress.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Event</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>2021 Sep</td>
<td>• Detailed project plan</td>
<td>Completed</td>
</tr>
<tr>
<td></td>
<td>• Project web page</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Project research</td>
<td></td>
</tr>
<tr>
<td>2021 Oct - Nov</td>
<td>• Understand codebase</td>
<td>Completed</td>
</tr>
<tr>
<td></td>
<td>• Database migration</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Authentication integration</td>
<td></td>
</tr>
<tr>
<td>2021 Dec</td>
<td>• Update the latest dataset</td>
<td>Completed</td>
</tr>
<tr>
<td></td>
<td>• Improve platform performance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Improving the user interface</td>
<td></td>
</tr>
<tr>
<td>2022 Jan</td>
<td>• Finish preliminary implementation</td>
<td>In progress</td>
</tr>
<tr>
<td></td>
<td>• First presentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Detailed interim report</td>
<td></td>
</tr>
<tr>
<td>2022 Feb</td>
<td>• Essential and Choice riders</td>
<td>Pending</td>
</tr>
<tr>
<td></td>
<td>• Database migration demo</td>
<td></td>
</tr>
<tr>
<td>2022 Mar</td>
<td>• Extreme riders</td>
<td>Pending</td>
</tr>
<tr>
<td></td>
<td>• Integration of rules of mobility patterns</td>
<td></td>
</tr>
<tr>
<td>2022 Apr</td>
<td>• Finalized tested implementation</td>
<td>Pending</td>
</tr>
<tr>
<td></td>
<td>• Final report</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Final presentation</td>
<td></td>
</tr>
<tr>
<td>2022 May</td>
<td>• Project exhibition</td>
<td>Pending</td>
</tr>
</tbody>
</table>
5 Conclusions
As the COVID-19 pandemic is still ongoing as a major public health burden for the world, this project aims to facilitate mobility research taking into account the epidemic development by building a scalable and secure one-stop platform for data query and ArcGIS visualizations. The team has developed additional interface functionalities to not only add some interactive design such as an explanation box and a progress bar but also streamline the login procedure. To increase the scalability of the backend, the team is working on replacing the current relational database with a new NoSQL database. Such improvements from both the front-end and backend can help build the current platform to be more robust and scalable, which is of core importance with the increasing number of concurrent operations on the platform. The team hopes the platform can facilitate the use of mobility data by researchers to look into ways to contain the spreading of the coronavirus, and most importantly, infer knowledge to keep the world prepared in case of future outbreaks.
References


