Department of Transportation: Data Warehousing and Analytics

Dennis Paolo V. Espiritu, Justin Ethan Raine S. Gorospe, Christian Louies M. Pagaduan, Maria Isabela A. Pilapil, and Marivic S. Tangkeko
IT Department, College of Computer Studies
De La Salle University, 2401 Taft Avenue, 1004 Manila, Philippines
Corresponding Authors: Christian_pagaduan@dlsu.edu.ph and marivic.tangkeko@dlsu.edu.ph

Abstract: The Department of Transportation, which is composed of the rail, air, maritime, and road sectors, uses manual processes in order to produce analytics. This results into the delay of the generation of accurate and timely analytics due to the size of the datasets. The DOTr also cannot maximize the datasets they currently have due to the limited capabilities of their existing processes for their planning and implementation of transportation projects. In order to address these issues, an IT solution with three primary modules were proposed – namely a data submission system, data integration and warehousing application, and a business analytics dashboard. The results of this project include a cleansed and normalized datasets, an executive dashboard, and a meaningful and useful descriptive and prescriptive analytics, such as knowing the busiest months, busiest stations, predicting the peak hours, passenger volume, and train trips. The objective of this study is to extract the data from the various agencies under the rail sector, transform and cleanse the datasets, and load the datasets into analytics. The significance of this study is to provide accurate rail heatmaps, calculate the revenue of the rail lines, and to aid in the decision and policy making for the DOTr's executives.

Key Words: Analytics; transport; rail; executive dashboard; government

1. INTRODUCTION

1.1 Background

Transport planning and management requires and consumes a vast amount of data. These data are used to satisfy the needs of effective transport planning, modelling, evaluation, and policy making (Huang, 2003). Among the data collected from various sources include passenger surveys, traffic monitoring, land use, and socio-economic records. According to Lee-Gosselin and Polak (1997), no single data collection method can provide sufficient information to meet the requirements of the transport models.

The need for such technology in the transport sector is due to the complexity of transportation related data. Problems such as a variety of data types and sources and inconsistency of data models can cause the inconsistency and comprehensiveness of the data.

1.2 Challenges

The primary challenge faced when developing the proposed system is the complexity of the data sets coming from the different agencies under the Department of Transportation. The data sets submitted to the DOTr have inconsistent header rows, multiple sheets inside a large Excel file, uses a dash “-“ instead of 0 for null values, duplicate station names, such as Buendia (MRT) and Buendia (PNR), different fare matrices per rail line, and data that can be as granular as hourly passenger ridership. All train lines, with the exception of the PNR, also have Beep card data, which contains all the transactions of each beep card. A single day alone contains thousands of rows for each rail line.

1.3 Significance and Aim

The project aims to develop a data warehousing and analytics solution using data integration and ETL of the rail sector under the Department of Transportation. There are four main modules in the proposed system, namely the data submission system, the data warehousing and central database, analytics and visualization tools, and the open data platform system.

The significance of this study is to provide immediate, consistent, accurate, and complete analytics that will aid the DOTr in transforming their data into useful information for decision making.

1.4 Scope

This study will cover transport data from the rail sector such as passenger movement, entry and exit data, and revenue in all the MRT and LRT stations excluding air, land and maritime data and will range from 2010 up to the present information. This paper will also cover the employees of the Management Information Service Department of the Department of Transportation excluding the employees of the administrative departments, attached agencies, and other service departments of the DOTr.

2. METHODOLOGY

The researchers used the rapid application development methodology in developing the proposed system. This enables the partner organization, the Department of Transportation, to be involved in the planning, development, and testing stages of the system.

One of the main reasons the proponents’ decided on utilizing the RAD methodology was due to the fact that RAD’S key emphasis is on “fulfilling the business need, while technological or engineering excellence is of lesser importance,“ since the main goal of our proposed system is to process, meet, and fulfill the order specifications, all of which depend on an organized exchange of information between the company. As much as possible, the team wants the partner organization to be involved in the development process to receive constant feedback and meet the correct requirements.

During the interview process, the team was able to gather information about the existing processes of the Department of Transportation, particularly in data gathering from the different sectors. The team was also able to identify the problem areas that the organization has been experiencing using the current system.

The partner organization provided the raw data sets that were used in the existing system, such as granular data on passenger traffic, train information, airport information, incident data, and revenue data. These datasets can be as granular as hourly data to monthly data.

The team also consulted with the partner organization on the recommended technologies. Since this is a government agency, the partner organization shall procure the licenses needed to launch this system into production.

With intensive research, interviews, and consultation with the partner organization, the proponents have identified the following problem areas of the existing system.

1. The file management system is not immediately accessible
2. The submitted datasets are inconsistent and inaccurate
3. The reports delivered to the department executives are incomplete.

Due to the problems mentioned, the Department of Transportation encounters difficulty in planning and implementing transportation projects.

The researchers have studied existing systems that can be used to integrate and implement to address these problems.

2.1. Framework

This capstone projects aims to address the problem areas in the Department of Transportation's existing system and processes. There are four main modules in the proposed system, namely the data submission system, the data warehousing and central database, analytics and visualization tools, and the open data platform system.

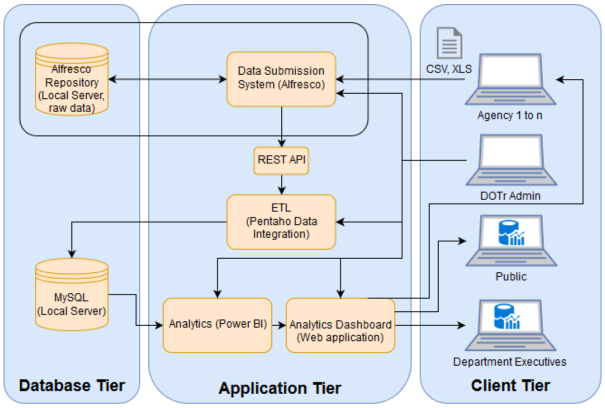
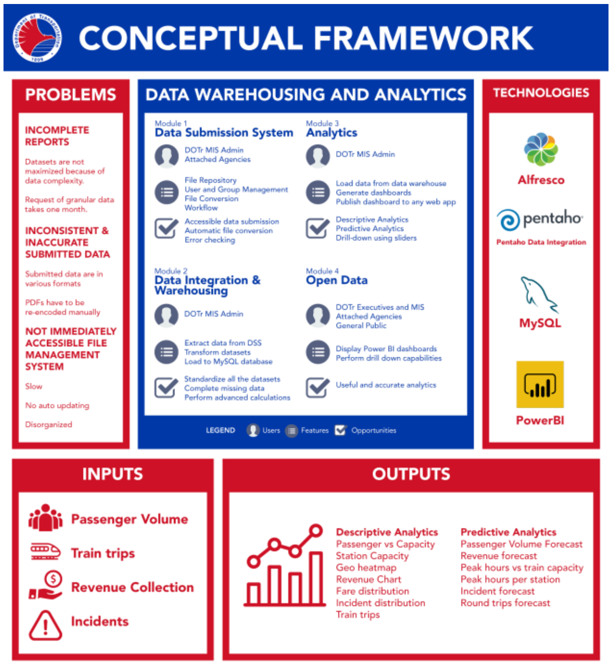


Fig. 1. Solutions Architecture



CONCEPTUAL FRAMEWORK

PROBLEMS

INCOMPLETE REPORTS

Datasets are not maximized because of data complexity.

Request of granular data takes one month.

INCONSISTENT & INACCURATE SUBMITTED DATA

Submitted data are in various formats

PDFs have to be re-encoded manually

NOT IMMEDIATELY ACCESSIBLE FILE MANAGEMENT SYSTEM

Slow

No auto updating

Disorganized

DATA WAREHOUSING AND ANALYTICS

Module 1
Data Submission System
DOTr MIS Admin
Attached Agencies
File Repository
User and Group Management
File Conversion
Workflow
Accessible data submission
Automatic file conversion
Error checking

Module 3
Analytics
DOTr MIS Admin
Load data from data warehouse
Generate dashboards
Publish dashboard to any web app
Descriptive Analytics
Predictive Analytics
Drill-down using sliders

Module 2
Data Integration & Warehousing
DOTr MIS Admin
Extract data from DSS
Transform datasets
Load to MySQL database
Standardize all the datasets
Complete missing data
Perform advanced calculations

Module 4
Open Data
DOTr Executives and MIS
Attached Agencies
General Public
Display Power BI dashboards
Perform drill down capabilities
Useful and accurate analytics

LEGEND Users Features ✓ Opportunities

TECHNOLOGIES

Alfresco

pentaho
Pentaho Data Integration

MySQL

PowerBI

INPUTS

Passenger Volume
Train trips
Revenue Collection
Incidents

OUTPUTS

Descriptive Analytics
Passenger vs Capacity
Station Capacity
Geo heatmap
Revenue Chart
Fare distribution
Incident distribution
Train trips

Predictive Analytics
Passenger Volume Forecast
Revenue forecast
Peak hours vs train capacity
Peak hours per station
Incident forecast
Round trips forecast

Fig. 2. Conceptual Framework

3. RESULTS AND DISCUSSION

This study will cover transport data from the rail sector such as passenger movement, entry and exit data, and revenue in all the MRT and LRT stations excluding air, land and maritime data and will range from 2010 up to the present information. This paper will also cover the employees of the Management Information Service Department of the Department of Transportation excluding the employees of the administrative departments, attached agencies, and other service departments of the DOTr.

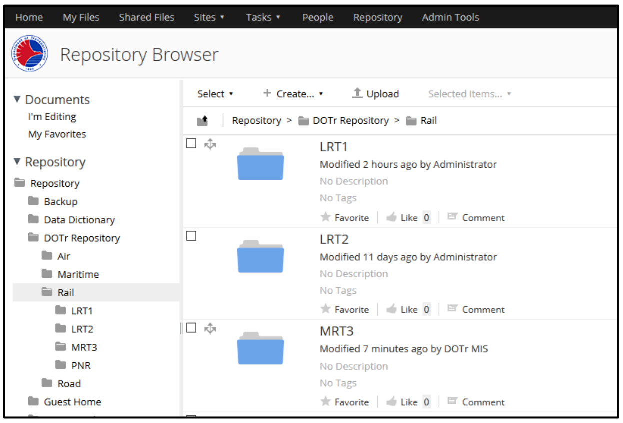
3.1. Modules

The researchers have developed a data warehousing and analytics solution for the Department of Transportation. The system has the following modules:

1. Data Submission module, powered by Alfresco

   Alfresco Enterprise Content Manager is a business application that is used to digitize processes, manage content, and securely govern information (About Alfresco, n.d.). It will be used as the main Document Submission System (DSS) of the proposed system. Alfresco allows users to be created that can access certain sites, each having its own file repository and access restrictions. The system shall have separate sites for the LRT, MRT, and PNR.

This module enables the different agencies under the Department of Transportation to send datasets and replaces the current method of sending emails. This enables the DOTr to check the data and ask for a revision without using the email. This module will also serve as the primary repository for the raw files of the received data, such as CSV and XLSX files.



Home My Files Shared Files Sites ▾ Tasks ▾ People Repository Admin Tools

Repository Browser

▾ Documents
I'm Editing
My Favorites
▾ Repository
Repository
Backup
Data Dictionary
DOTr Repository
Air
Maritime
Rail
LRT1
LRT2
MRT3
PNR
Road
Guest Home

Select ▾ + Create... ▾ Upload Selected Items... ▾

Repository > DOTr Repository > Rail

☐ LRT1
Modified 2 hours ago by Administrator
No Description
No Tags
Favorite | Like 0 | Comment

☐ LRT2
Modified 11 days ago by Administrator
No Description
No Tags
Favorite | Like 0 | Comment

☐ MRT3
Modified 7 minutes ago by DOTr MIS
No Description
No Tags
Favorite | Like 0 | Comment

Fig 3. Data Submission module

2. Data Integration module and warehousing module, powered by Pentaho Data Integration and MySQL

The Pentaho Data Integration application simplifies the process of ETL (extraction, transformation, and loading) with the use of a graphical interface. It allows multiple file input types and can interact with database managers to integrate data (Pentaho Data Integration - Accelerate Data Pipeline, n.d.). The files received from the DSS are cleaned through ETL using the Data Integration tool. Kettle scripts are run on each file to ensure that there are no null values and incorrect data formats. All the data from the datasets will be extracted via REST API from the data submission module and will be checked for errors and duplicates.

The datasets will also be transformed and standardized into a data model. These data will then be be loaded into the MySQL data warehouse for analytics. The MySQL server was deployed locally for the development of this system, however it will later on be deployed on the DICT servers during final implementation.

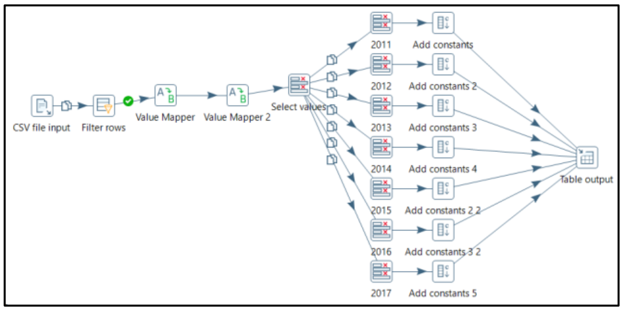


Fig 4. Data Integration Module

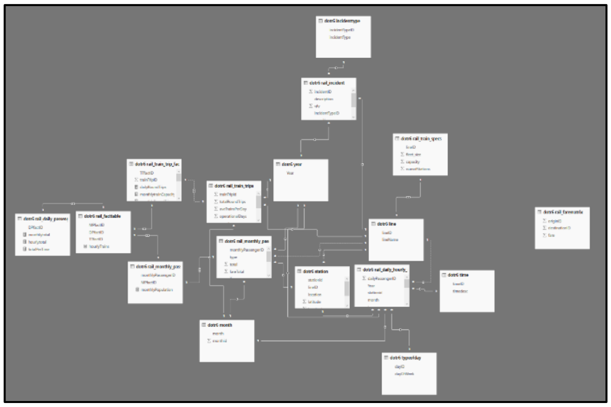


Fig 5. Data model

3. Business Analytics module, powered by Power BI.

Microsoft Power BI is a powerful tool used in data analytics, which is capable of handling large amounts of data to be loaded and visualized. Aside from a user interface with an easier learning curve designed for non-technical users. Microsoft Power BI has more drill down capabilities as well as allowing user to directly publish there charts online or onto a different platform through accessing the web. This module is an interactive executive dashboard that is used to visualize the datasets that were stored in the data warehouse.

Visualizations include geomaps, bar graphs, line graphs, and tables to indicate the descriptive analytics of the data. Credentials to this module will be given to the department executives for viewing. This module is primarily for visualization of data with a few predictive analytics such as peak hours at any given time of the day or week. Aggregated results from the multiple agencies under a single sector are also included in this module, such as LRT1, LRT2, MRT3, and PNR aggregated results in one page due to similarities in passenger volume and peak hours.

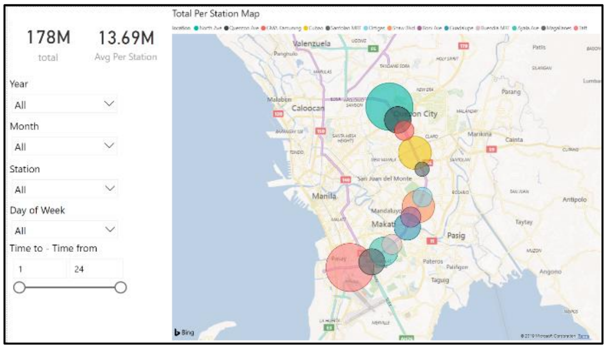


Fig. 6.1. MRT Dashboard with Heatmap

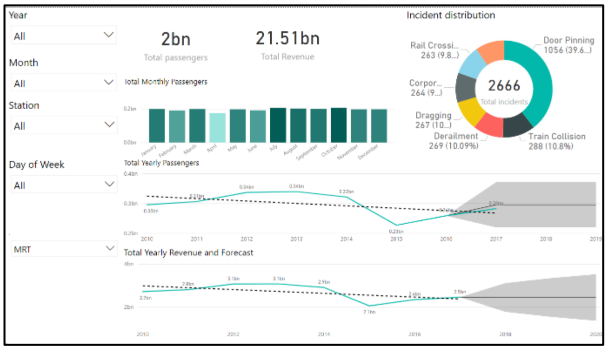


Fig 6.2. MRT Dashboard with Passenger, revenue, and incident analytics

4. Open Data module, powered by DKAN via data.gov.ph.

Using the website under the data.gov.ph, the reports and analytics generated from the transformed data can be accessed by the general public. For information security purposes, only public data (for example, ridership volume and station traffic) will be displayed in the open data portal. This module is backed by DKAN open data platform.

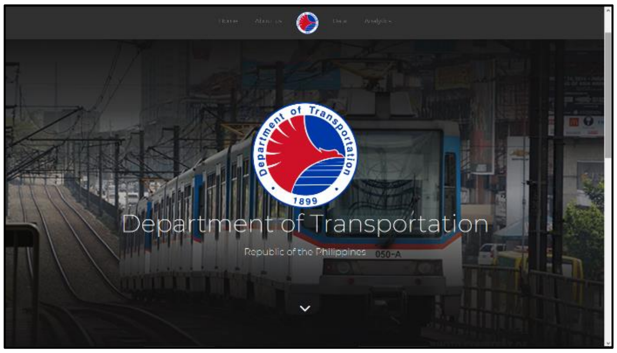

Department of Transportation
1899
Department of Transportation
Republic of the Philippines

Fig 7.1. DOTr Open Data landing page

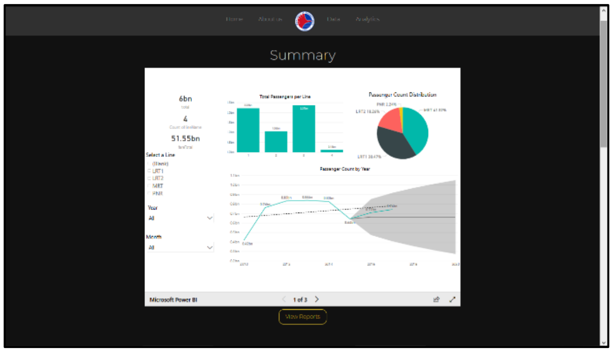


Fig 7.2. DOTr Open Data for rail sector

The expected results of this system are the different types of analytics produced and the useful information that can be obtained from the analytics for decision making and policy making.

Through these modules, the file management system has been made accessible to both agencies and the department executives, the submitted datasets are now consistent and accurate due to the intensive ETL processes, and the reports are now complete when delivered to the department executives.

4. CONCLUSION

With the use of this proposed system, improved decisions, such as when to send out more trains or deploy more security personnel, how many trains must be procured to meet the increasing demand of passengers, and policy-making, such as extending the operating hours of trains, can now be done through the information that came from analytics. These datasets can now be understood by department executives and stakeholders through visualization. Important analytics include financial data, passenger volume, train trips, and incident data can be used by both the department executives and agencies to accurately plan and implement transportation projects.

5. ACKNOWLEDGMENTS

The proponents would like to express their gratitude to the following people for making this proposed system possible:

To the Department of Transportation, specifically Dir. Paul Visaya, Mr. Varsolo Sunio, Ms. Ruth Montes, Mr. Edgar Fernando for giving us the opportunity to assist them in developing this system.

To Mr. Johannes Paulus Acuña for assisting us in developing the system and recommending us to the Department of Transportation

To Mr. Oliver Malabanan for helping us in writing the project in methods and Ms. Marivic S. Tangkeko for the guidance and knowledge in making this project possible.

6. REFERENCES

About DOTR. (2018). Retrieved October 4, 2018, from GOVPH: http://dotr.gov.ph/2014-09-02-05-01-41.html

Acuña, Johannes Paulus (2018). Personal Communication.

Huang, Z. (2003). Data Integration for Urban Transport Planning. Retrieved from https://webapps.itc.utwente.nl/librarywww/Papers_2003/phd_theses/zhengdong_huang.pdf

Kawabata, Y., Aoki, H. (2009). Metro Manila Strategic Mass Rail Transit Development. Japan International Cooperation Agency. Retrieved from https://www2.jica.go.jp/en/evaluation/pdf/2008_PH-P171_4.pdf

Luke, R. (2017). The failure of transport megaprojects: lessons from developed and developing countries. Pan-Pacific Conference XXXIV: Designing New Business Models in Developing Economies. Retrieved from https://www.researchgate.net/publication/320267809_The_failure_of_transport_megaprojects_lessons_from_developed_and_developing_countries

Sunio, V. (2018). Personal communication.

Visaya, P. (2018). Personal communication.