Managing a Large Volume of Data for Public Engineering Enterprise

Kevin Tantisevi[1] and Jirapon Sunkpho[2]

Abstract

Management of a large volume of data from multiple sources is a challenging task in most public agencies. It requires a good understanding of the underlying data and their structures as well as a good software tool that enables extracting, transforming, and cleansing of data stored in different platforms. Moreover, domain expertise and knowledge are needed in order to reason about these data to enable early detection of business risks and chances. These challenges are unique and differ from system to system and from organization to organization. This paper focuses on these challenges and presents a case study conducted in developing an information management system for a government agency in Thailand to integrate data from various sources and provide business insights for its management team. This paper discusses some technical difficulties experienced in its current information management practice and describes our resolutions to overcome them by using the developed system. Finally, it summarizes key factors that contribute to the success of information management for public enterprises.

Keywords: data integration, data visualization, decision support, information management.

Introduction

Public agencies are typically involved with a large quantity of data. They collect and have access to data as much as, if not more than, the private sector (The Center for Digital Government, 2013). These data come from various sources both internally and externally and are processed within the organizations to serve their business purposes. The U.S. National Institutes of Health (NIH), for example, can access five million pictures of tumors through its cancer research program (FCW, 2012). Once processed, these data are utilized by an executive team to make decisions on long-term strategic plans as well as to monitor and control regular operations.

Information management that involves a large volume of data is a challenging task. The nature of most data today can be described using Gartner's 3Vs as high-volume, high-velocity, and high-variety information that requires new forms of processing to enable enhanced decision-making, insight discovery and process optimization. (Beyer and Laney, 2012). Traditional tools such as relational databases and desktop software for statistics and visualization are no longer adequate. The volume and the velocity of data are increasingly large due to more applications being serviced to residents as well as technological advance in data capture technology. In addition, these data can be in a variety of formats including structured, unstructured and semi-structured formats. One report estimates that merely 15 percent of data generated globally is structured while the remaining 85 percent is unstructured (Tech America Foundation, 2012). If these data are not managed properly, it

[1] Assistant Professor, Department of Civil Engineering, King Mongkut's University of Technology North Bangkok, 1518 Pracharat 1 Road, Bang-Sue, Bangkok, Thailand 10800, Tel: +66-2-5552000 Ext. 8621-9, Fax: +66-2-5874337, E-mail: kevin.t@eng.kmutnb.ac.th.

[2] Assistant Professor, College of Innovation, Thammasat University, 2 Prachan Road, Bangkok, Thailand 10200, Tel: +66-2-6235055, Fax: +66-2-6235060, E-mail: jirapon@tu.ac.th.

can lead to ineffective managerial decision making and missing opportunities for improvement. Hence, there is a need for management of large amount of data stored within organizations to enable effective administration of their business processes.

This paper describes a case study conducted to enhance capabilities of information management systems at the Inter-City Motorway Division, a Thai government agency in charge of development and operation of an inter-city motorway network in Thailand. In this study, a new data consolidation system was implemented to extract data from various sources and integrate them into a centralized database. Based on these integrated data, this system created a unified management dashboard for an executive team to monitor business performances in near real-time. This paper presents an overview of this system and then discusses its application in real-world engineering management tasks. Finally, it suggests lessons learnt during its implementation.

Overview of Previous Information Management for Thailand's Inter-City Motorway Network

In Thailand, an inter-city motorway network is a tolled road network system that connects the Bangkok metropolitan and its surrounding industrial and tourism districts. Comprising two major controlled-access divided highways that totally span 146 kilometers, the inter-city motorway network has been developed, maintained, and managed by the Inter-City Motorway Division, a government sector under the Department of Highways (Tantisevi et al. 2012).

In order to ensure safety, smooth traffic operation, and effective capital budgeting on a motorway network, the Inter-city Motorway Division uses several information systems to collect, process, and analyze data from various sources and places. These systems are composed of different kinds of software applications, databases, and hardware devices, such as tablets, personal computers, closed-circuit television (CCTV) cameras, and global positioning system (GPS) receivers. Figure 1 illustrates a set of information systems used by the Inter-City Motorway Division to support its operations.

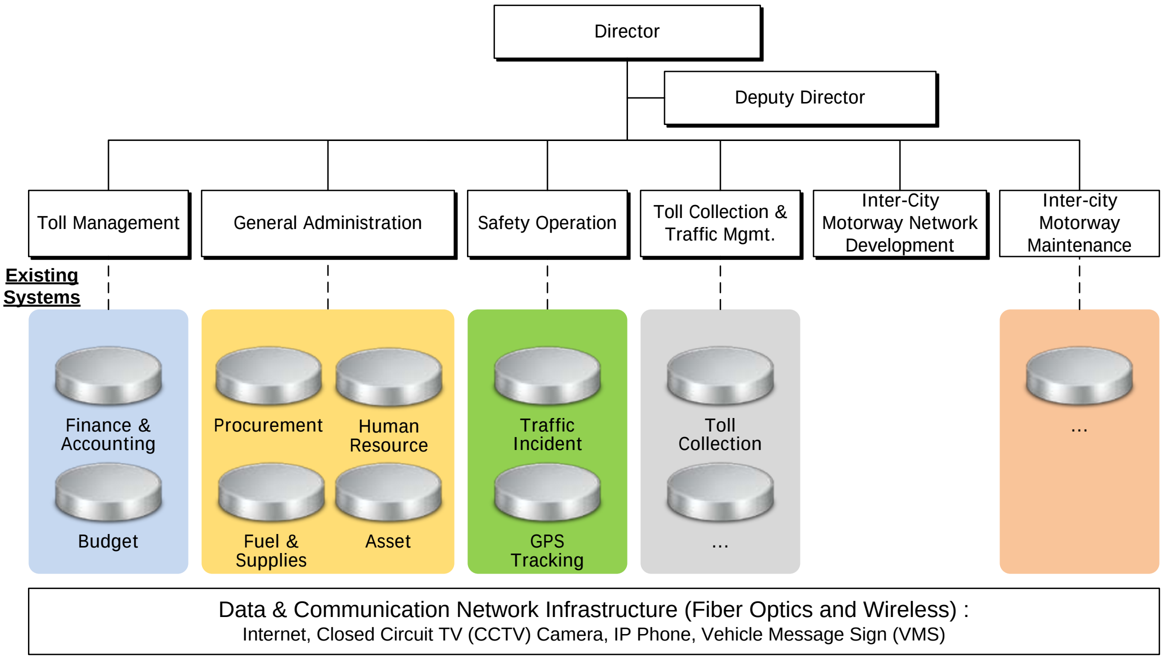

Director
Deputy Director
Toll Management
General Administration
Safety Operation
Toll Collection & Traffic Mgmt.
Inter-City Motorway Network Development
Inter-city Motorway Maintenance
Existing Systems
Finance & Accounting
Budget
Procurement
Human Resource
Fuel & Supplies
Asset
Traffic Incident
GPS Tracking
Toll Collection
...
...
Data & Communication Network Infrastructure (Fiber Optics and Wireless) :
Internet, Closed Circuit TV (CCTV) Camera, IP Phone, Vehicle Message Sign (VMS)

Figure 1. Organizational Structure and Existing Information Systems of the Inter-City Motorway Division

In Figure 1, the existing information systems mainly served two types of organizational functions: general administration and mission-specific tasks. General administration entail back-office works that are typically found in any business and contribute to existence of an organization. Examples of these works include finance, accounting, procurement, human resource, and asset management. Several intranet-based application systems had been used to support these major administration works.

In contrast, organization-specific tasks vary significantly from one business to another. The Inter-City Motorway Division, that takes full responsibility on the development, operation, and maintenance of the inter-city motorways, had specific task forces to manage construction and maintenance projects, collect tolls, and provide roadside assistance services. Therefore, it had a group of information systems to gather data, such as daily traffic volume, roadway incidents, and pavement conditions on the entire roadways. Based on these data, the Inter-City Motorway Division was able to assess its key performance indicators (KPI), such as the quality of services and overall pavement condition index.

Technical Difficulties in Previous Information Management Practice

Although the information systems shown in Figure 1 sufficed the needs for administrative officers to perform their regular tasks, an executive team needed more abilities to make sense of data integrated from multiple sources in order to make effective decisions. This was hardly achievable by using only these systems due to the lack of appropriate data integration. Each system was originally designed to streamline data-entry and report generation processes within a certain department of the organization. There was no on-line data exchange protocol (e.g., shared database and web-services) implemented between different departmental units. As a result, the only way for other units and the executive team to obtain specific data sets was by printed reports.

While some people are primarily concerned with data security issues and do not want to allow others to directly access their controlled data, we consider such a lack of data integration to be a major impediment to top-level management of the organization. Executive members need to access organizational information in a convenient and timely manner so that they can early identify risks and determine necessary actions to mitigate negative impacts.

In order to integrate data from various sources, some technical challenges need to be addressed. Subsections below discuss those challenges in detail.

Heterogeneity and Data Inconsistency

A major technical challenge in data integration stems from the fact that information systems implemented at the Inter-City Motorway Division were heterogeneous in terms of a variety of database management systems, application frameworks, and operating systems being used. In most government agencies, a competitive-bid program is used to select software companies to develop information systems. While such a bid program ensures that the lowest responsible bidder takes the job, it is difficult to enforce a successful bidder to develop a new information system that is consistent with existing ones. To comply with Regulations of the Office of the Prime Minister on Procurement B.E. 2535 (1992), tender terms and conditions need to be open enough to allow various software companies and products to compete in bidding without prejudice. Therefore, a succesful bidder is likely to select tools and techniques based on its experiences. This results in a complex software environment comprising different applications, databases, and hardware platforms.

The heterogeneity of the existing information systems results in data inconsistency among different platforms. It was found that several information systems stored similar data sets while using different types, formats, and sizes. For example, a finance system stored a vendor's address in one text field while the procurement system used a group of data fields to store street address, district, province and postal code separately.

Moreover, some discrepancies between values of data stored in multiple databases were observed. This problem resulted from the lack of a change data capture (CDC) program implemented in the existing information systems. When data in one database changed, the same data stored in other databases were not updated automatically. Instead, it relied on data-entry officers to manually apply changes in all related databases. Such a manual process is error-prone and takes time for all relevant data to be up-to-date and fully synchronized. There is a need for data profiling and cleansing before all the necessary data are loaded in a unified database.

Concern over Performance Degradation of the Live Systems

Besides data inconsistency, there were major concerns about performance drop in mission-critical systems while data integration is performed. Several information systems are used extensively 24 hour a day to serve critical operations, such as rescue operation and toll collection. These operations cannot be delayed because they are critical to life safety and delays in such operations can escalate traffic congestion on the motorway. Therefore, performance degradation of these systems was considered to be unacceptable by an executive team and field officers. In addition, there are systems that allow users to add and update data only during business office hours. Neither these systems can be interrupted nor they have slow response time while users are accessing them.

Development of Data Consolidation Systems

In order to address the challenges highlighted in the previous section, the Inter-City Motorway Division commenced a data consolidation project. According to its budget plan, this project has two phases. The first phase involves integrating structured data from existing database systems and establishing a business intelligence (BI) platform. The database systems within the scope of the first phase implementation are enumerated in Table 1. The second phase starts after completion of the first phase by consolidating remaining structured data and unstructured data, such as video and images. In the second phase, predictive models will also be derived from collected data to help identify business risks and evaluate possible outcomes. Once the entire project gets completed, it will become a single application platform for executive members to visualize and analyze all necessary organizational information.

Table 1. User Access Behavior of the Existing Database Systems

Existing Systems	Operating Time	
	Days	Hours
Finance & Accounting	Mon – Sat	8:00 – 16:30
Budget	Mon – Sat	8:00 – 16:30
Procurement	Mon – Sat	8:00 – 16:30
Fuel & Supplies Management	Mon – Sat	8:00 – 16:30
Traffic Incident Management	Mon – Sun	0:00 – 24:00
Human Resource Management	Mon – Sat	8:00 – 16:30

This paper put emphasis on development of a data consolidation system in the first phase. In this phase, several tasks were identified and undertaken by a group of IT specialists.

Firstly, system analysts set up a first-round meeting with an executive team to introduce a data integration concept and get preliminary requirements of the new system. The analysts also interviewed administrative officers to understand their business processes and collected database schema of the existing systems. After having a good understanding of the business requirements and the underlying data, they continued to lay out a system design that includes dashboard design, data specification, and system architecture.

Secondly, system engineers investigated the overall performance of the existing systems and created a data extraction plan. To avoid performance degradation in the mission-critical systems, network data traffic, system input/output (I/O) and CPU load of database servers were monitored for a certain period of time. Table 1 shows a user access behavior of the existing database systems. Figure 2 provides a graphic example of weekly inbound network traffic observed at the Inter-City Motorway head office. Based on the observed behavior of database operations and network traffic, system engineers suggested that data extraction and integration of each system be performed every midnight.

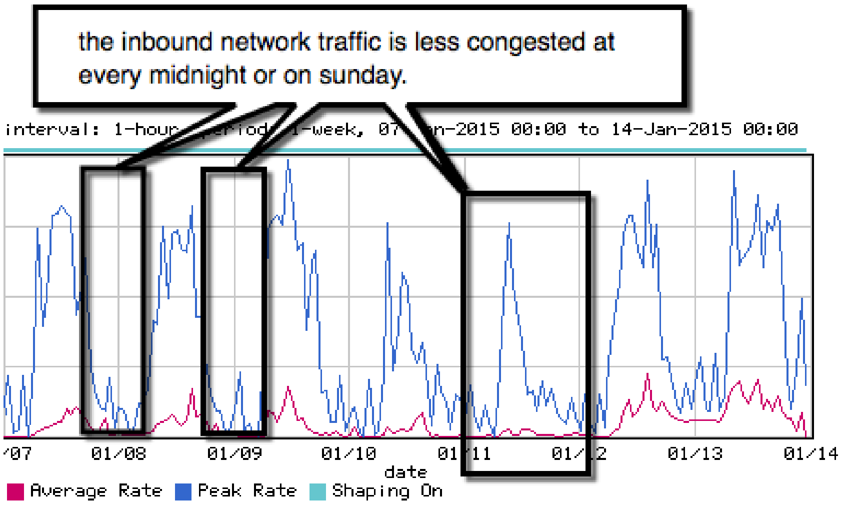

the inbound network traffic is less congested at every midnight or on sunday.

Figure 2. Weekly Inbound Data Network Traffic

After having all necessary information in place, a team of IT specialists continued to develop the data consolidation system. A subsection below describes its system architecture in detail.

Overview of the Data Consolidation System

Figure 3 illustrates an overview of the new data consolidation system. In the figure, the data consolidation system is composed of two main components including database server and business intelligence (BI) server. These two servers were actually built as virtual servers and were physically located on different virtualization hosts. The primary reasons for server virtualization in this project was to better resource allocation and to allow for easier system maintenance that includes system backups, disaster recovery, and future migration. The database server was utilized to collect data from various existing systems. It serves as a central repository of all structured data that can be summarized, analyzed, and finally reported to an executive team. In this project, Microsoft SQL Server was used as a relational database management system (RDBMS).

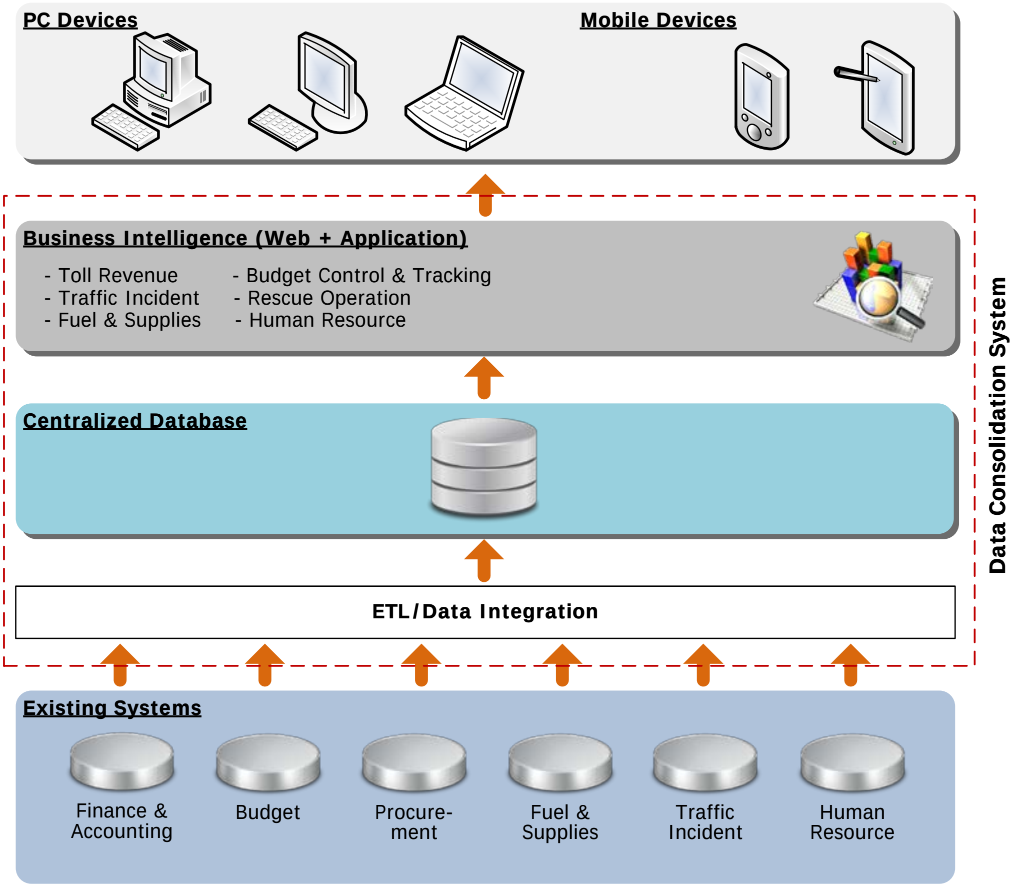
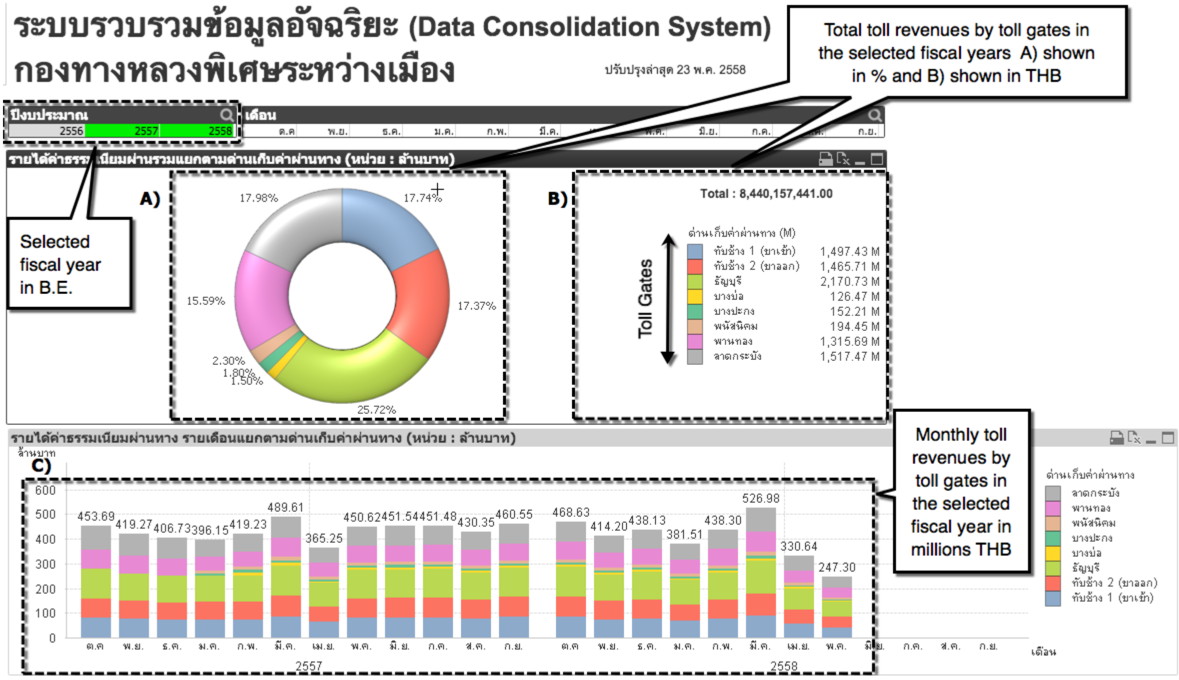


Figure 3. System Architecture of the Developed Data Consolidation System

The business intelligence (BI) server was used to create a management dashboard for an executive team. It loads data from the database server into its memory and presents them in a graphic and intuitive way. In this project, a commercial BI software system, Qlikview (QlikTech International AB, 2011), was chosen because it required relatively less knowledge about software programming, when compared with other traditional BI tools. Another reason for selecting Qlikview was that we wanted to develop the first prototype and presented it to an executive team as soon as possible. It had been difficult for the executive team to explain what they really wanted to visualize until they saw a working system with some actual data. The prototype was used to help the executive team generate new ideas and provide suggestions for improving the system. Since Qlikview mainly targets analysts and business users who want to quickly and interactively build data analytics, it could fit these requirements.

Applications of the Data Consolidation System

After successully implemented, the data consolidation system has shown potentials in many areas. Most importantly, it provides a graphic and user-interactive environment for users to analyze organizational data and explore association among different data sets. For instance, Figure 4 shows a screen capture of a management dashboard for visualizing toll revenue data. The dashboard comprises a set of drill-down charts that enable an executive team to examine toll revenues from different angles and at multiple levels of detail. The executive team can select a specific period of time to view the monthly and total revenues (Figure 4A and 4C). Also, he or she can click on these charts to see toll revenues collected by specific toll gates. This visual presentation is useful for the team to estimate seasonal losses of revenue during the long holidays, such as New Year and Songkran Festivals (Thai traditional New Year), when motorists are allowed to use the motorway network for free. In addition, the toll data of each toll gate presented in the dashboard can be exported and utilized in combination with traffic counts collected from an electronic traffic counting system for internal auditing purposes.



ระบบรวบรวมข้อมูลอัจฉริยะ (Data Consolidation System)
กองทางหลวงพิเศษระหว่างเมือง

ปรับปรุงล่าสุด 23 พ.ค. 2558

Total toll revenues by toll gates in the selected fiscal years A) shown in % and B) shown in THB

Selected fiscal year in B.E.

Monthly toll revenues by toll gates in the selected fiscal year in millions THB

รายได้ค่าธรรมเนียมผ่านรวมแยกตามด่านเก็บค่าผ่านทาง (หน่วย : ล้านบาท)

Total : 8,440,157,441.00

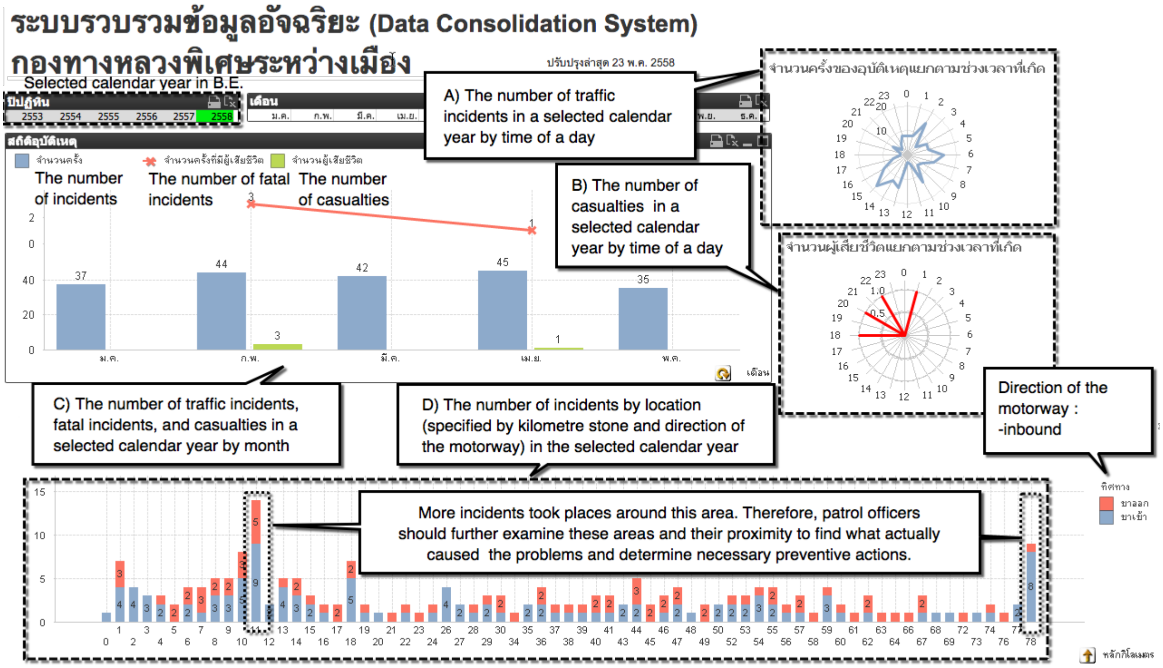
ด่านเก็บค่าผ่านทาง (M)	
ทับช้าง 1 (ขาเข้า)	1,497.43 M
ทับช้าง 2 (ขาออก)	1,465.71 M
ธัญบุรี	2,170.73 M
บางบ่อ	126.47 M
บางปะกง	152.21 M
พนัสนิคม	194.45 M
พานทอง	1,315.69 M
ลาดกระบัง	1,517.47 M

รายได้ค่าธรรมเนียมผ่านทาง รายเดือนแยกตามด่านเก็บค่าผ่านทาง (หน่วย : ล้านบาท)

Figure 4. Screen Capture of a Dashboard for Visualizing Toll Revenues

Figure 5 depicts another view of the dashboard used for monitoring traffic incident statistics. Road accidents are one of the major causes of deaths in Thailand and attribute to approximately 13,500 road traffic deaths (the rate of 38 deaths per 100,000 population) per year (WHO, 2013). Therefore, it is necessary that the executive team pay attention to these statistics. The dashboard shown in Figure 5 provides the executive team with a brief summary of the number of traffic incidents, injuries, and fatalities on the motorway. It also visualizes where and when traffic incidents occur most. Based on such information, the executive team can identify problematic locations (e.g., 11 kilometers from the start of the motorway) to further assess road hazard conditions. If the corresponding assessment result shows a high accident risk level, the executive team will determine road-safety actions to mitigate such risks.

With different pieces of information being compiled together and presented in a form that is easy to understand, the data consolidation system enables the executive team to have access to all critical data. Another key benefit is its capability to expand in the future. Its centralized data repository has laid a solid foundation for the organization to make continuous improvement to serve its business needs. As data sets were cleansed and integrated, more software systems can be easily developed by utilizing existing data sets. Domain-specific knowledge can also be applied to assist in decision-making.



ระบบรวบรวมข้อมูลอัจฉริยะ (Data Consolidation System)
กองทางหลวงพิเศษระหว่างเมือง
ปรับปรุงล่าสุด 23 พ.ค. 2558
Selected calendar year in B.E.
ปีปฏิทิน
2553
2554
2555
2556
2557
2558
เดือน
ม.ค.
ก.พ.
มี.ค.
เม.ย.
พ.ย.
ธ.ค.
สถิติอุบัติเหตุ
จำนวนครั้ง
The number of incidents
จำนวนครั้งที่มีผู้เสียชีวิต
The number of fatal incidents
จำนวนผู้เสียชีวิต
The number of casualties
37
44
42
45
35
3
1
ม.ค.
ก.พ.
มี.ค.
เม.ย.
พ.ค.
เดือน
จำนวนครั้งของอุบัติเหตุแยกตามช่วงเวลาที่เกิด
จำนวนผู้เสียชีวิตแยกตามช่วงเวลาที่เกิด
ทิศทาง
ขาออก
ขาเข้า
หลักกิโลเมตร

A) The number of traffic incidents in a selected calendar year by time of a day

B) The number of casualties in a selected calendar year by time of a day

C) The number of traffic incidents, fatal incidents, and casualties in a selected calendar year by month

D) The number of incidents by location (specified by kilometre stone and direction of the motorway) in the selected calendar year

Direction of the motorway : -inbound

More incidents took places around this area. Therefore, patrol officers should further examine these areas and their proximity to find what actually caused the problems and determine necessary preventive actions.

Figure 5. Screen Capture of a Dashboard for Visualizing Traffic Incidents

Lessons Learnt from Implementing a Data Consolidation System

After successfully implementing the system for managing large organizational information, we have identified three key success factors as follows: understanding the underlying data, choosing the right tool, and applying domain-specific knowledge to serve business needs. The first important issue to the success of managing large data set for public organization is to understand the underlying data that is to be managed. By understanding, we mean ones need to know what data are being kept, who is responsible for managing those data, how they are stored, how often they are updated, and most importantly, how a particular data item relates to others. The association among data sets in the organization is also very important information when trying to integrate different piece of information together and to be able to visualize new insight that was not able to visualize before.

Another key point in managing a large data set is selecting the right tool to manage the underlying data. There are several software tools both commercially and open-sourced that are applicable for managing large data sets. However, each software tool has its own strengths and weaknesses. By choosing the right tool, the resources required to implement the system is also minimized. However, if available commercially or open-sourced tools cannot address the requirements, an organization can choose to develop the right tool to customize its own needs. A public organization may explore the applicability of a particular software tool by piloting a small project and then evaluating if it meets performance, budget, and security requirements.

Once having a good understanding of the existing data and selecting the right tool, one needs to apply domain-specific knowledge to make sense of the data and better decision-making. Graphic presentation of data sets may allow executive members to easily monitor business performances. However, it is also necessary for them to derive analytics from the existing data sets. This is the optimization part where domain expertise and a good understanding of business processes come into play to make use of the selected tool at its potential.

Conclusions

This paper presents a case study conducted to manage a large volume of data at a Thai government agency. In this study, a new data consolidation system was implemented to extract data from various sources and integrate them into a centralized database. This paper describes an overview of the system and then discusses key success factors learnt from our implementation. These factors include 1) understanding the underlying data, 2) choosing the right tools, and 3) applying domain-specific knowledge to continuously improve the organization. Finally, as a public organization collects more and more of citizen and public data, it is necessary to provide public the access to the data. Giving public access will put tremendous pressure on the organization on ensuring both transparency and security of their data. However, the result is rewarding when people start to utilize these data to create innovation and make good use of them to give positive impacts to the society.

References

Beyer, M. and Laney, D., 2012. *The importance of 'Big Data': a definition*. Available from: http://www.gartner.com/DisplayDocument?ref=clientfriendlyUrl&id=2057415.

FCW, 2012. *Research report: big data.* Available from: http://fcw.com/BigDataResearch.

Office of the Prime Minister, 1992. *Regulations of the office of the prime minister on procurement B.E. 2535.* Bangkok: Office of the Prime Minister [in Thai].

QlikTech International AB, 2011. *QlikView tutorial: version 11 for Microsoft Windows®.* 1st Edition. Lund, Sweden.

Tantisevi, K. Sunkpho, J. and Surabal, S., 2012. A GPS-based traffic incident management system: a case study of Thailand's inter-city motorways. *In: Proceedings of the 14th international conference on computing in civil and building engineering, 27-29 June 2012, Moscow, Russia.*

Tech America Foundation, 2012. *Demystifying big data: a practical guide to transforming the business of government*. Available from: www.techamericafoundation.org/official-report-of-the-techamerica-foundations-big-data-commission

The Center for Digital Government, 2013. *Big data big promise*. Available from: http://images.erepublic.com/documents/CDG13_SPQ1_V.pdf.

WHO, 2013. *Global status report on road safety 2013: supporting a decade of action*. Luxembourg: World Health Organization.