Large Scale Intelligent Transportation System Traffic Detector Data Archiving

Tao Qu, Graduate Research Assistant,
Department of Civil and Environment Engineering, University of Wisconsin-Madison
1415 Engineering Drive, Room 1241, Madison, WI, 53706
Phone number: (608)609-3375, E-Mail: tqu3@wisc.edu

Steven T. Parker, Ph.D, IT Program Manager,
Department of Civil and Environmental Engineering, University of Wisconsin-Madison, 2205 Engineering Hall, 1415 Engineering Drive, Madison, WI, USA 53706,
Phone: (608) 265-4921, Fax: (608) 262-5199, Email: sparker@engr.wisc.edu

Yang Cheng, PhD, Research Associate,
Traffic Operations and Safety (TOPS) Laboratory
Department of Civil and Environment Engineering
University of Wisconsin-Madison
1241 Engineering Hall, 1415 Engineering Drive, Madison, WI, USA 53706,
Phone: (608) 262-2524 Email: cheng8@wisc.edu

Dr. Bin Ran, Ph.D., Professor
Department of Civil & Environmental Engineering, University of Wisconsin-Madison,
1415 Engineering Drive, Madison, WI 53706, USA
Phone: 1-608-262-0052 Fax: 1-608-262-5199
Email: bran@wisc.edu
and
School of Transportation, Southeast University
No.2 Si Pai Lou, Nanjing 210096, China

David A. Noyce, Ph,D, P.E., Professor
Director, TOPS Lab
Department of Civil & Environmental Engineering, University of Wisconsin-Madison,
Madison, WI 53706, USA
Phone: 1-608-265-1882 Fax: 1-608-262-5199
Email: noyce@engr.wisc.edu

Submission date: 8/1/2013
Word count: 4,372 + 8 Figure + 4 Tables = 7,372 words

ABSTRACT

Archived traffic data can be used in transportation planning, administration, and research by various entities and agencies. During the past two decades, considerable effort has been dedicated to developing and implementing large-scale traffic data archives. The Wisconsin Traffic Operation and Safety (TOPS) Laboratory at the University of Wisconsin – Madison maintains a statewide traffic detector data archiving and retrieving system, which is developed to enable centralized management of statewide ITS detector and configuration data, optimizing the utilization of massive data on a systematic level, and improving the interactivity and accessibility for integration with other transportation data sources such as lane closure data or incident data. This data archive is currently being enhanced to incorporate higher resolution traffic data by migrating from 5-minute to 1-minute and even 20-second sampling intervals. At the same time, there is a desire to generate aggregated datasets such as hourly, monthly, and annual average values from the raw data. As the traffic data requirements continue to grow, the management of the traffic data archive becomes a complex big data problem. This paper describes a proposed redesign of the TOPS Lab traffic detector archived data management system to improve storage, performance, access, and integration capabilities. Particular detail is given to the data archiving process, including data validation, and support for spatial attributes and GIS data integration.

INTRODUCTION

Traffic data archiving for Intelligent Transportation Systems (ITS) refers to the systematic retention and reuse of operational ITS data, and is required for many application. In the mid-1990s, real-time data from traffic operations was archived and used for purposes beyond traffic control strategies. Archiving these otherwise discarded data provides a rich source of information for evaluating traffic flow characteristics and transportation system performance on a continuing basis *(1)*. To encourage the retention and reuse of ITS-generated data, the archived data user service (ADUS) element of the National ITS Architecture requires that data from ITS systems be collected and archived for historical, secondary and non-real-time uses and that these data be made readily available to users *(2)*. It also provides a general framework for collecting, processing, retaining, and distributing these data. As a result, during the past two decades, many data archiving activities have been conducted by different transportation agencies and considerable effort has been dedicated to developing and implementing large-scale traffic data archives *(3, 4, 5)*. As a supplement or replacement of conventional data sources, the cost-effectiveness of data collection infrastructure is maximized, sampling bias is minimized due to the continuous collection of data, and a better understanding of variability in system performance can be accomplished *(6)*. Once these data are archived, the vast amount of data can be used in transportation planning, administration, and research by various entities and agencies including metropolitan planning organizations (MPOs), state transportation planners, traffic management operators, transit operators, and transportation researchers *(6)*.

The Wisconsin Department of Transportation (WisDOT) has invested heavily over the past two decades in ITS infrastructure. This includes the emergence of a Statewide Traffic Operations Center (STOC) in Milwaukee and the deployment of Advanced Transportation Management Systems (ATMS) software at the STOC and other regional offices. The STOC gathers data from approximately 4,800 traffic detectors that compose the ATMS. A variety of point detection technologies are implemented on the state freeway system including: inductive loop detectors, radar detection, microwave detection, and more recently Bluetooth detection. Generally, these detection systems provide traffic flow data for freeway surveillance and real-time freeway management.

Traffic detector data from the STOC ATMS is archived on a continual basis in the WisTransPortal system, located at the Wisconsin Traffic Operations and Safety (TOPS) Laboratory at the University of Wisconsin-Madison *(7,8)*. The ATMS detector data is available for online query and retrieval through the WisTransPortal VSPOC system *(9)*. VSPOC serves as the primary source of archived Wisconsin ITS traffic detector data and is used by WisDOT, local governments, consulting firms and university researchers for a variety of traffic operations, planning and research purposes.

Since deployment of the VSPOC system in 2007, several significant challenges to the effective use of the WisTransPortal ATMS traffic detector data have been identified. The primary challenges relate to a lack of geo-spatial (GIS) location attributes, ongoing data quality issues, data frequency, and documentation. In addition, the sheer volume of data received from the STOC (over 5 million records per day) presents ongoing data management considerations. This paper will provide a overview of current development efforts by TOPS Lab to upgrade the WisTransPortal ATMS detector data archive and VSPOC system to address these challenges. A specific focus on large scale data management and GIS integration will be discussed, along with implications for improved query capabilities and integration with other WisTransPortal transportation data.

WISCONSIN ITS TRAFFIC DETECTOR DATA

Wisconsin Traffic Data Sources

As in other states, there are several sources of highway traffic detector data in Wisconsin: real-time traffic ITS detector data from the STOC control room ATMS, 511 Traveler Information link speeds and travel times (derived from the ATMS detector data), continuous and short-term traffic counts from the WisDOT Bureau of State Highways (BSHP) planning area, and more recently Bluetooth sensor data. In addition, third party traffic data is available from a variety of sources.

The STOC ATMS monitors real-time traffic volume, speed, and occupancy data from approximately 4,800 freeway detectors in five transportation regions (Table 1). These data are used for real-time management of the transportation system to monitor the traffic state and identify congested area and incidents. At 20-second intervals, each detector records vehicle counts, average speed, and occupancy (i.e., the percentage of the sample period when a vehicle was over the detector). A roadside controller is then polled at that same time interval by the central traffic management software. The ATMS aggregates 20-second data to 1-minute intervals for control room monitoring and traveler information purposes.

As described, traffic detector data from the STOC ATMS in the TOPS Lab WisTransPortal system. The basic objective of the WisTransPortal is to develop capabilities for a statewide ITS data hub to support multiple applications in traffic operations and safety. Those capabilities include integration, management, analysis, and dissemination of real-time and historical ITS operations data through a centralized database and communications infrastructure (7). The data archiving component of the WisTransPortal consists of a collection of automated services that connect to various WisDOT and other ITS data sources and prepare the data for archiving in a common relational database.

TABLE 1 Wisconsin Freeway Detector Distribution

WisDOT Region	Online Date	Controllers	Detectors
SE Region	1996	352	3821
SW Region	2003	107	732
NC Region	2006	48	194
NE Region	2010	23	95
NW Region	2011	1	5

The WisTransPortal traffic detector database is updated every 24 hours with WisDOT ATMS detector data from the previous day. There are three major files are sent from the WisDOT ATMS to TOPS lab for archiving via the WisDOT "ITSNET" fiber network:

- ATMS Detector Configuration Inventory
- ATMS Controller Configuration Inventory
- ATMS One Minute Detector Volume, Speed, And Occupancy Data

The design of the WisTransPortal detector database incorporated elements form the Traffic Management Data Dictionary (TMDD) standard *(10)* whenever possible. Hence, as the guiding principle, the WisTransPortal detector database is structured to support TMDD requirements.

VSPOC Online Tool

The V-SPOC web-interface includes a complete detector database query selection tool, data visualization and export capabilities, data quality reporting, and integration with other ITS data sets in the WisTransPortal database. In order to facilitate the use of this data, the current VSPOC application suite provides mechanisms to organize detector data into spatial (corridors, count locations and controllers) and temporal (time interval) groupings. Besides that, VSPOC was designed as a set of customized modules that take advantage of a common database backend and web-based application platform to provide an integrated system for transportation operations engineering, analysis, research, and reporting functions. Several specialized modules are available for various needs of users and suitable presentation of data including corridor analysis, monthly data retrieval, ramp metering retiming, etc. Figure 1 presents an example of an archive volume plot for sample detector location on I-43 in Milwaukee County.

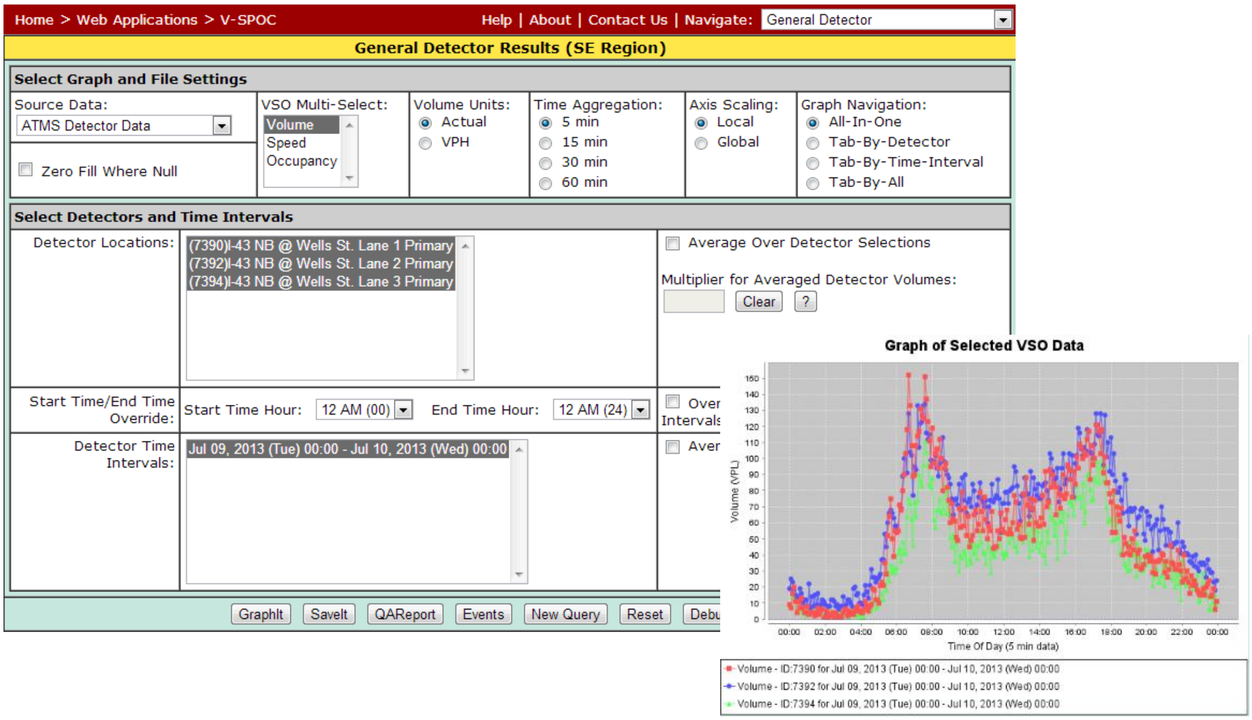

Figure 1 Example of VSPOC modular.

VSPOC ETL

Due to the complexity and distinct set of requirements, it is common to separate the implementation of real-time traffic management system software from the archived data management component. The process of populating the traffic data archive, or moving data from the operation center's database to the traffic data archive, is known as extraction, transformation, and loading (ETL). Incorporating effective ETL process and archiving algorithm for high resolution loop detector data is complex and requires significant attention in design. The current extraction and transformation process used to populate the VSPOC database system is presented in Figure 2.

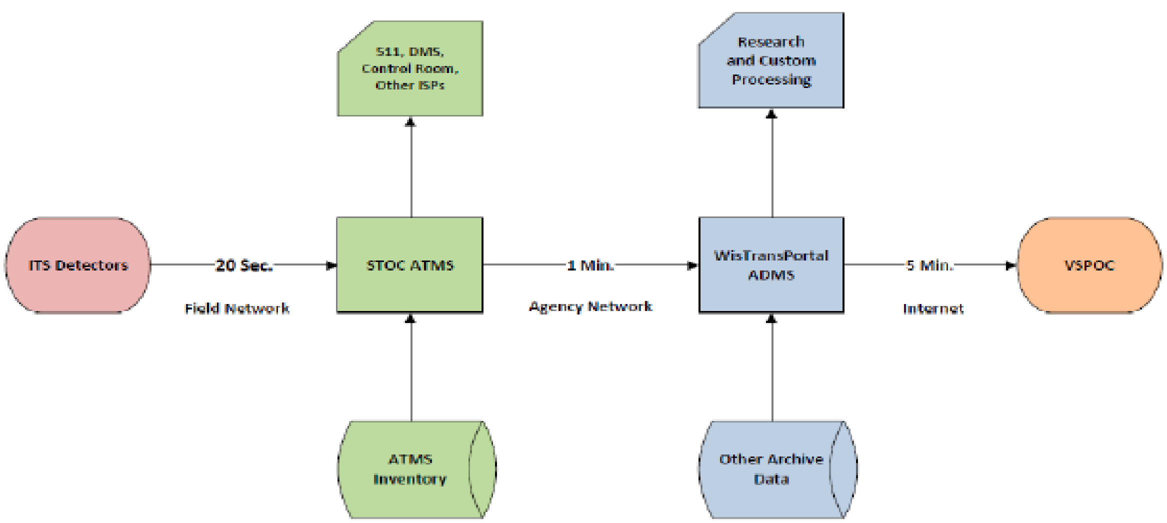

FIGURE 2 WisTransPortal VSPOC detector data architecture.

VSPOC Database Future Trends

The STOC has recently starting providing ATMS detector data to the WisTransPortal in terms of 1-minute intervals. Converting from the previous 5-minute resolution to 1-minute resolution introduces a number of data management challenges. As of 2012, the cumulative record count of traffic detector data in the WisTransPortal had already exceeded four billion. In addition, as the continuous development and investment on ITS devices, the scale of traffic monitoring operations will also grow. As presented in Figure 3, the cumulative record count in 2014 is predicted to be nine billion, which means the amount of two years data are larger than that of past 15 years. The linear trend of data size growth will be five times than before. Consequently, traffic data archiving is becoming a big data challenge which requires a mass data storage solution with high fault tolerance and throughput and sophisticated data management techniques.

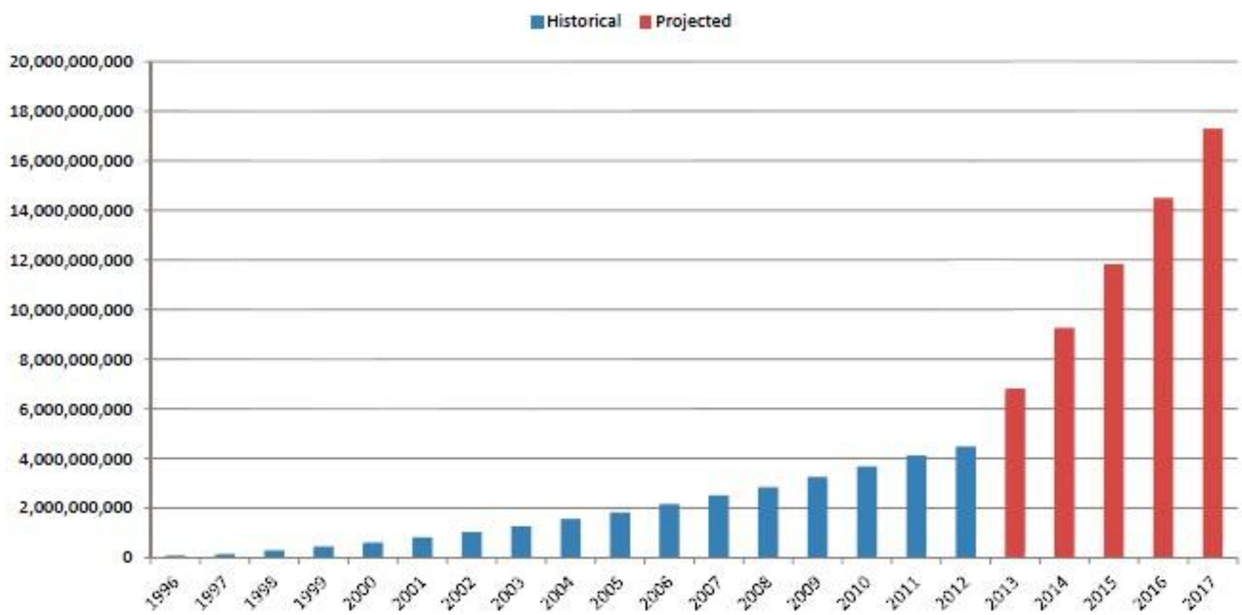

FIGURE 3 VSPOC database cumulative record count.

ARCHIVING PROCESS

The archiving process refers to steps taken for processing and archiving the traffic detector data and system configuration data. To solve the aforementioned big data challenge and support the VSPOC frontend web application, it is necessary to continue improving the effectiveness of the archiving system and data performance in terms of completeness, timeliness, accuracy, and interactivity. Redesigning the archiving process provides a unique opportunity to rethink how to better shape the VSPOC traffic data archive. Hence, the following three objectives are identified:

1. Automation: Increase and refine the automation capabilities of the archiving process
2. Data Validation: Implement front-end data validation steps and report exceptions that occur in archiving process
3. Aggregation: Transform the temporal and spatial grain of raw data to desired temporal and spatial grain of archived data

The proposed archiving process is presented in flowchart in Figure 4. Included discussion is how these objectives are taken into account within the archiving process and how to address identified issues.

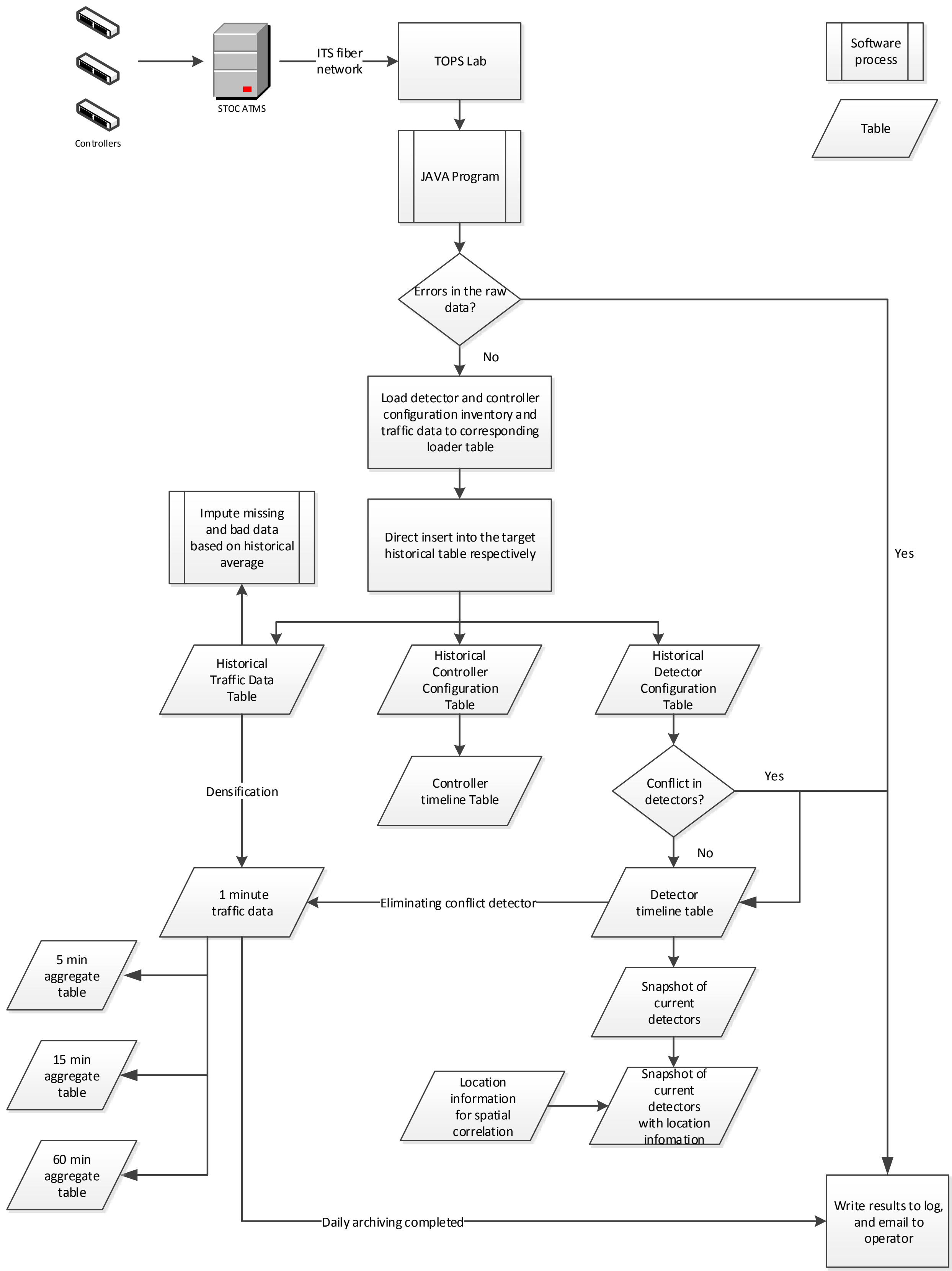

FIGURE 4 Proposed VSPOC archiving process

Automation

A Java program is developed to implement the archiving process including retrieving source data from the STOC, importing the raw data into a temporary loader table, and processing the loader data to the target database. Once deployed, it can be used for centralized management. This process is completely automated – it runs as a scheduled "cron" job on a separate virtual server assigned for data acquisition and writes a log of its activity. In the current VSPOC system, only the detector data import process is automated. The process for updating the detector and controller configuration data is a manual procedure. Moreover, the automated detector data archiving component does not incorporate a full loader process and is therefore limited in its post-processing and data validation capabilities, as described below.

Data Validation

Structural errors in the detector data files arise from system faults in the ATMS itself, differences in how detector data is represented in the ATMS compared to the WisTransPortal, and in the packaging and data transmission process. The Java archiving program handles structural errors in the input data to support high quality of data acquisition. In each step, specific constraints have been set for data validation. These constraints are used to automate the process of identifying and loading the qualifying data into the archive system. Rules are implemented to handle conflicts, where possible, in order to capture as much data as possible without compromising the archive. Moreover, all exceptions are caught and reported by email. Checking and analyzing the reported errors enables developer to resolve them at first time, which aims at dynamic data maintenance and continuous improvement on VSPOC. Once in the archive system, data remains online and accessible.

Aggregation

Data aggregation can be used to reduce storage requirements on large datasets and to improve the performance of analytic processing over vast amounts of fine grain data. Whereas the temporal and spatial granularity at the traffic operations center is generally limited to point detections and travel time routes with respect to 1-minute time intervals, it is desirable for the VSPOC archive to supports a range of aggregation levels. The frontend of VSOPC suite is a web user interface for data query, data visualization, data exporting, quality reporting, and corridor analysis. By using physical tables to store traffic data at desired aggregation levels, the underlying database system can better support the various needs of VSPOC users such as providing monthly or annual averages of daily traffic patterns. In addition, data aggregation can smooth out irregular data values and eliminate noise that may report occasionally because of detector or communication failures. Pre-processing aggregate data and storing it physically in the database improves performance but increases data management requirements. A description of typical temporal and spatial granularity requirements in an operation center and a traffic archive with Wisconsin's practice is provided in Table 2.

TABLE 2 Grain Description

Aggregation Level	Operation Center's Database	Traffic Data Archive
Temporal	Intervals between two successive polling operations of same detector by a controller. STOC: 20 seconds Travel times and link speeds for control room and traveler information: 1 minutes	Intervals between two successive database records for the same detector location. VSPOC: 1-minute, 5-miunte, 15-minute, 60-minutes, daily, monthly.
Spatial	Geographic area served by a detector. STOC: a transverse section of a lane.	Transverse section of a lane, detector location, roadway segments such as link or corridor. VSPOC: lane, link, corridor.

DATABASE DESIGN

The new database design targets the VSPOC backend to address the identified issues in current archiving process and disadvantages of underlying database, as well as targets data product in terms of necessary improvement on data quality and data coverage. In particular, the following four objectives are the primary concerns during the redesign of database: 1) Establish historical archives of system configuration data; 2) Accomplish densification of traffic data of current online detectors; 3) Synchronize internal ID and external ID for each detector (internal ID refers to an individual detector, external ID points to a physical detector location); and 4) Enhance GIS attributes of detector. The proposed design makes sure the archived data are accurate, sufficient, and easily accessible through a high performance database system. The database environment for VSPOC is the WisTransPortal Oracle Database 11g Enterprise Edition, which can improve database performance and reliability.

Table Design for Configuration Data

The table design for configuration data primarily address the challenge of management on historical configuration data and synchronization and transformation between internal ID and external ID. Figure 5 shows the data flow of system configuration data and relationship among each functional table. The archiving process leverages this design to automate data processing on system configuration data and to generate current snapshot of online detector.

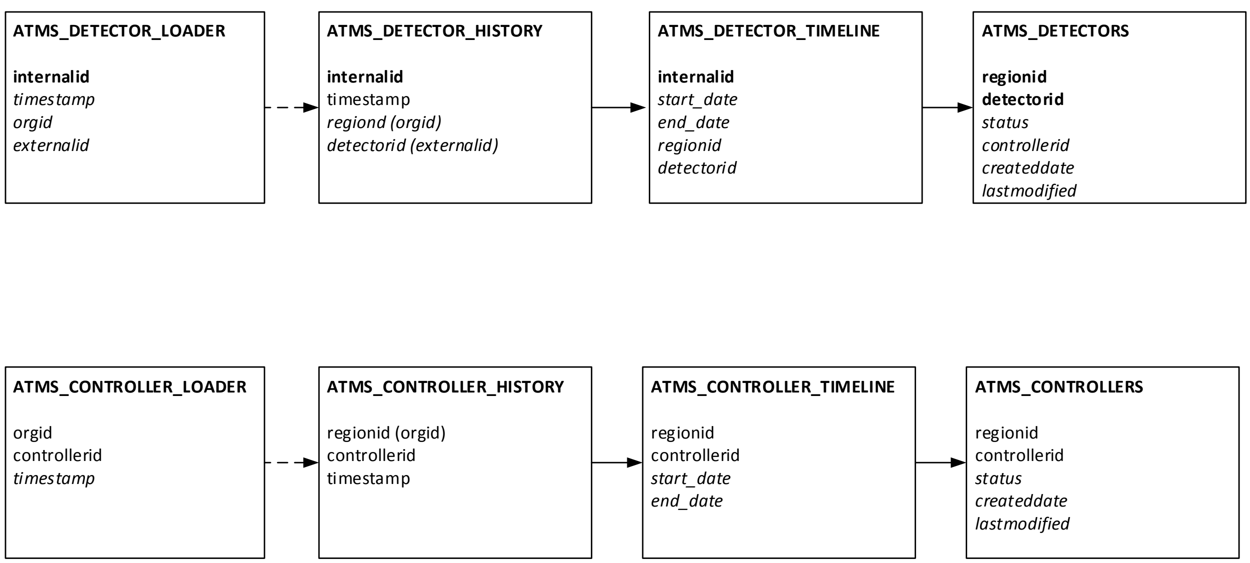

FIGURE 5 Database design and data flow for detector (controller) configuration inventory.

Staging Loader table

Two intermediate staging databases (DETECTOR_LOADER and CONTROLLER_LOADER) serve as loaders to import detector system configuration data. The primary purpose of staging loader table is to accommodate the significant real-time process requirement when populating vast amount of data, providing maximum availability and stability during archiving process. As shown in the flowchart of archiving process, each night, the Java archiving program retrieves the system configuration data and then loads the data to specific loader table for the previous 24-hour period. As a buffer, the staging loader tables can process and validate the received data before it goes to archive, and prevent duplicate recorder in historical table. It ensures the further process operates on clean, correct and useful data. After inserting the data with timestamp into historical table, the loader tables will be emptied to recycle databases.

Historical table

The VSPOC database system accumulates new configuration information but still needs to retain older information to preserve point in time snapshots of the ATMS system configuration. Two historical (DETECTOR_HISTORY and CONTROLLER_HISTORY) tables are then designed to keep historical information of system configuration to date. It provides an efficient way to update system configuration data on regular basis, and allows original data to be preserved to ensure complete, reliable integrity for the life of archived data. In addition, it prevents the potential risk from system crash or failure of Java archiving program. Backup and recovery runs faster based on historical table, disaster recovery therefore is less costly. Moreover, the historical tables enables further development upon raw data, facilitating diagnosis algorithm for detector health and data availability. After the data is inserted into historical table, qualified raw data is archived exactly as it is received in real time. The entire current day's data are then populated to the VSPOC database backbend's historical table in a nightly batch after the end of day.

Timeline table

The detector and controller timeline table (DETECTOR_TIMELINE and CONTROLLER_TIMELINE) present a timeline for each individual detector and controller. The need of timeline table stems from the detector management strategies. For each detector, it either can be identified by INTERNALID which is unique lifetime identifier for an individual detector, or the pair of REGIONID and EXTERNALID, a unique identifier for an individual detector within a given WisDOT transportation region (REGIONID) field detector data collection system. For example, if a detector is broken, a new detector with a different INTERNALID will be installed to replace the previous one. However, the EXTERNALID for this detector location remains the same. Figure 6 concludes three types of detector ID conflicts.

By grouping data by detector internal ID then ordering by timestamp, two new attributes are created, which are Start_Date and End_Date indicating online and offline date for a certain detector or controller (identified by INTELNALID or CONTROLLERID). By matching the online and offline date of each detector or controller to their corresponding EXTERNALID, traffic data can be precisely retrieved by locating the online detector for specific time and location. Table 3 is an example of how timeline table solve conflict detector issue.

Three types of conflict detector issue:

1. Overlap
   The new detector becomes online before the current one become offline. There is an overlapping period.
2. Gap
   The new detector becomes online after the current one become offline. There is a gap between the offline date of the current detector and the online date of the new detector.
3. Return
   The current detector become offline, after a period of time, it returns back and becomes online.

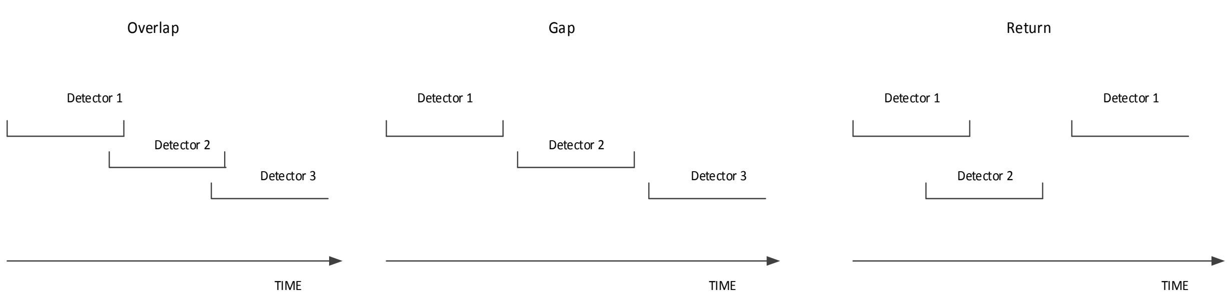

FIGURE 6 Illustration of three types of detector ID conflict

TABLE 3 Example of How Timeline Table Solves Conflicting Detectors

INTID	REGIONID	EXTID	FIRST REPORTED	START_DATE	END_DATE
1001	1	100	01-01-2012	01-01-2012	03-03-2012
1002	1	100	03-01-2012	03-04-2012	06-06-2012
1003	1	100	02-06-2012	06-07-2012	

The design of the timeline table can also support the systematic retention of data generated from detectors over time. If any unexpected problems found in historical data, it can allow user easy to access the data and locate the problem.

Other tables and views

The database system also involves additional tables (views) in the database to maximum flexibility and availability. Two important tables are defined as:

ATMS_DETECTORS: Current snapshot for detector configuration information by STOC. A status flag indicate whether the detector is current or historical. The value is either “C” for current or “H” for historic.

VSPOC_DETETCORS: Additional detector configuration information including location information entered through VSPOC configuration interface.

Table Design for Traffic Data

The database design for traffic data ensures quality and availability of traffic detector data. As the same logic of configuration data processing, a loader table (ATMSREPORTS_LOADER) is used to filter bad record in the raw data, while ATMSREPORTS_1MIN serves as historical table for traffic detector data. Figure 7 shows the data flow of traffic detector data, procedures for handling conflicting detectors, and densification.

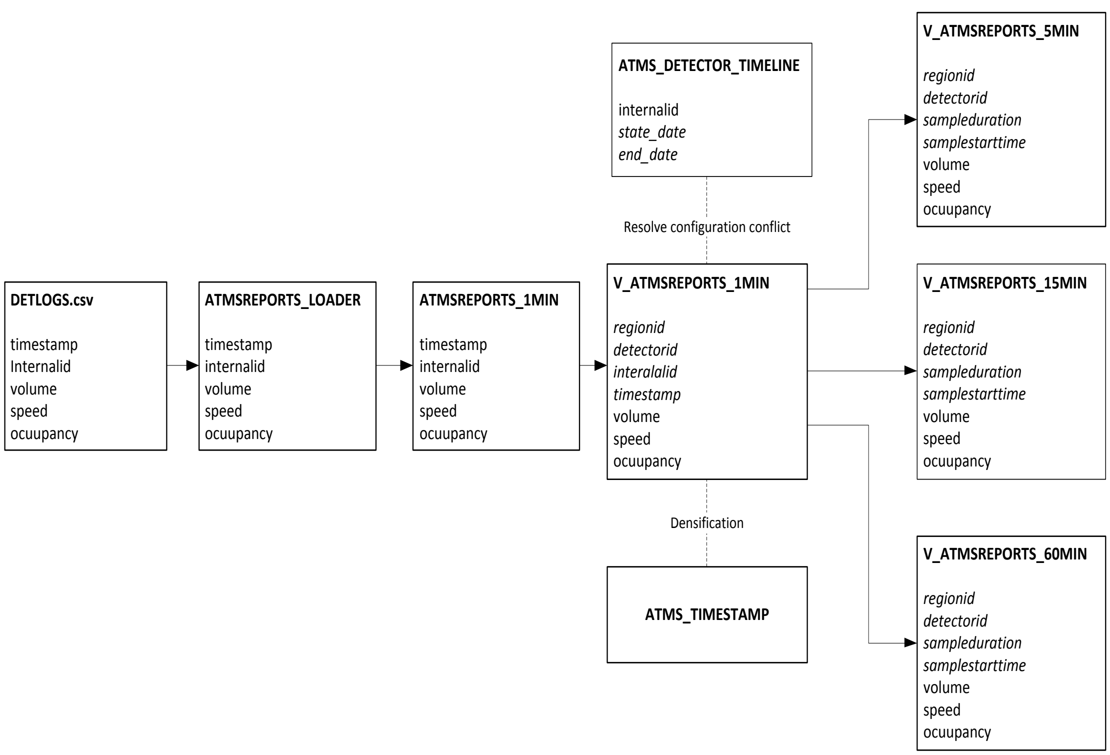

FIGURE 7 Database design and data flow for traffic detector data.

Densification

It is common for the detector logs to exhibit short-term and long-term break in data availability in terms of discontinued time series. In order to provide accurate reporting and analysis functions through VSPOC, it is desirable to work with a dataset that has a uniform sample space. To generate a complete set of records for all time periods, a pre-defined template table consisting of 1-minute timestamps was created to join with the source detector data. To screen out offline detectors, the raw data join the DETECTOR_TIMELINE table first. Then by joining the timestamp table, a database view (ATMSREPORTS_1MIN) is generated which provides a dense presentation of traffic data of current online detector by filling in NULL values at timestamps where archived data is missing. All detector reports table must be uniform in its representation of time, by doing this, data densification is pushed down to the database level where it most efficiently solved.

GIS Attributes

As described, the WisTransPortal design incorporates two functions: ITS data archiving and ITS data dissemination. The current data archiving capabilities are primarily focused on integrating and correlating six existing WisDOT and regional data sources: traffic detector data, lane and ramp closure data, traffic incident data, historical crash reports, freeway traffic video, and road weather condition. Each subsystem is designed according to high level architecture to be a component of a highly-integrated, highly-interactive traffic management database system.

As mentioned, the VSPOC_DETECTOR table contains location information for detectors. Besides using a latitude and longitude pair to represent a detector location, it can be modeled by a single control point implemented by a system of fixed landmarks. A landmark is defined as a physical, identifiable point on a highway, such as an intersection, milepost, bridge, or a virtual point, such as a county boundary line over a highway. In VSPOC, landmark and corresponding offset, which comprise a direction and a distance, can be used to provide highly accurate location for detectors on Wisconsin State Trunk Network (STN), the WisDOT GIS-based linear referencing system for state and federal highway in Wisconsin. These processes enable the new VSPOC system to incorporate data from internal WisDOT data systems and applications with minimal user interaction. The integration of data is the key to next generation data archived and management system as a single data source correlated with other related data sources. In addition, TOPS lab is on progress to visualize traffic detectors location by mapping all the detectors on STN linear referencing system. Figure 6 shows the mapped detector location on STN linear referencing system.

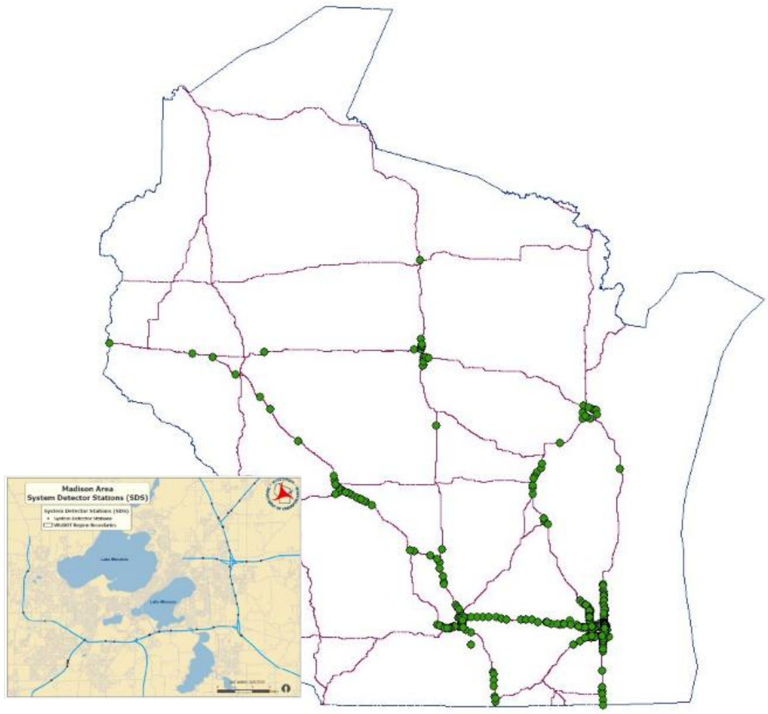

FIGURE 8 Mapped detector on stn linear referencing system.

CONCLUSION

This study describes experience and implementation on large-scale ITS traffic detector data archiving. The identified issues and potential solutions are universal and therefore other regions around the country can be benefit from this experience. Towards a next generation of traffic data archive, the VSPOC system will continue dedicating on data availability and accessibility for research and decision support. Future work of this study may include following directions. First, a comprehensive algorithm is needed to optimize system performance. Second, by identifying attributes of data quality that are relevant to ITS data archiving, data quality assessment and statistical techniques are desirable to improve data quality. Third, continuous improvements are needed on GIS network, data modeling, and high level system integration to leverage other data sources such as lane closure data or incident data to support decision making.

ACKNOWLEDGMENT

The WisTransPortal system and VSPOC traffic detector data archive were developed through sponsorship of the Wisconsin Department of Transportation Bureau of Traffic Operations (BTO). Ongoing data archiving and system enhancement activities with respect to the ITS traffic detector data are conducted by TOPS Lab in coordination with the BTO Statewide Traffic Operations Center.

REFERENCES

1. Turner, S. M. *Guidelines for Developing ITS Data Archiving Systems.* Report 2127-3. Texas Transportation Institute, Texas A&M University
2. *System, 2001.Archived Data User Service (ADUS): An Addendum to the ITS Program Plan.* ADUS Program. U.S. Department of Transportation, Sept. 1998.
3. Brian L. Smith and Simona Babiceanu. Investigation of Extraction, Transformation, and Loading Techniques for Traffic Data WarehousesFederal Highway Administration. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1879, TRB, National Research Council, Washington, D.C., 2004, pp. 9–16.
4. Robert L. Bertini, Steve Hansen, Andrew Byrd, and Thareth Yin. Experience Implementing a User Service for Archived Intelligent Transportation Systems Data. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1917, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 90–99.
5. Dilruba Ozmen-Ertekin, Kaan Ozbay. Dynamic Data Maintenance for Quality Data, Quality Research. OECD. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
6. *ITS Data Archiving: Five-Year Program Description.* ADUS Program. U.S. Department of Transportation, March 2000.
7. Parker, Steven. and Tao, Yang .WisTransPortal: A Wisconsin Traffic Operations Data Hub. *Applications of Advanced Technology in Transportation* (2006): pp. 611-616.
8. Wisconsin DOT, TOPS Laboratory UW-Madison. *V-SPOC Traffic Detector Database Tools*. http://transportal.cee.wisc.edu/applications/vspoc.html.
9. TOPS, Wisconsin Traffic Operations and Safety Laboratory *The Wistransportal Project* http://transportal.cee.wisc.edu/.
10. Institute for Traffic Engineers (2004). *Traffic Management Data Dictionary (TMDD), Version 2.1.*