Development of a Statewide Transportation Data Warehousing and Mining System under the Louisiana Transportation Information System (LATIS) Program

Introduction

Considerable amounts of data are collected by metropolitan and state authorities that are either not used or accessed and analyzed with difficulty. This situation has been exacerbated by the recent increase in information from Intelligent Transportation Systems and other data collection programs. To address this situation, the Louisiana Transportation Research Center (LTRC) initiated a study investigating the design of a data warehousing/data mining system that would be applicable to metropolitan areas in the state, or as a statewide model.

Objective

The objectives of this study are to assess whether introducing a data warehousing/data mining system in Louisiana would be feasible and beneficial. The study sets out to identify the features of the most suitable system for the state as well as to outline options available in its implementation.

Scope

The scope of this project was limited to the design of a data warehousing and mining system and did not include implementation of any aspect of the design other than to verify its feasibility. The design was set within the context of Baton Rouge, but the system is generic in its application and could be applied to any metropolitan area or as a statewide system.

Research Approach

The methodology employed in designing the data warehouse/data mining system included visiting many of the agencies that collect data and collecting sample data from them. Advisory and stakeholder committees were formed to identify potential applications of the data. The applications were traced back to the data resulting in some being dropped or modified to suit the data and its quality. From that juncture, the data was tracked forward again to the applications modeling the transformations necessary to get the data into the format required.

The study also included an extensive examination of several Data Management Systems in other states. This provided information on how they function and the benefits they

gain by having access to real-time and archive data. The report documents the results of lengthy interviews and surveys with key Intelligent Transportation System (ITS) data users and managers to identify implementation challenges and lessons learned from practitioners experienced in Data Management System deployment or use.

Conclusions

A data warehouse performs three main functions: (1) extraction/transformation/loading (ETL) of source data, (2) storage of the data, and (3) client access capabilities. ETL consists of acquiring, cleansing, formatting, merging, and purging of source data. Much of the time spent on this activity involves data quality checking. Data storage has to account for different types of data, and it is common to separate static data, e.g., sensor location, from dynamic data, e.g., lane occupancy. Data retrieval involves issues such as marketing, authorization, data transfer, and security.

Recommendations

- It is both feasible and desirable that a data warehouse/data mining system be introduced in the state.
- The warehouse should have a permanent staff consisting of at least a technician, database analyst, and computer programmer. A transportation engineer needs to be either a full-time or part-time staff member.
- Start with the simplest applications with visible, short-term payoffs. New applications should be added singly in an incremental manner.
- Keep in mind that the warehouse is not a static system but grows application by application; avoid trying to be all things to all users at the outset.
- Data mining should be considered an application to be added after the base applications are operational.
- Maintain the archival, read-only nature of the data; this means resist the temptation to make it an operational database with disparate users supplying updates spontaneously.
- The warehouse should have uncontested ownership of the data it collects.
- The initial implementation should ultimately extend past the main storage phase and include the configuration of necessary tools, such as a Geographic Information System (GIS), on client workstations.