Graduate Certificate Examination SemesterSubject:

Big Data Analytics

APPENDIX A

'iTrans' Case Study

Sample Examination Questions

Appendix A

iTrans– Case Study

A Platform for Smart Transportation
(A Hypothetical Case Study)

Introduction and Background

Smart transport is a vital system within a smart city. Big Data technologies have been widely used in many areas such as designing and planning of smart transportation. Smart transportation will result in operational efficiency, improving end-to-end customer experiences, reducing congestion and increasing flexibility. It can benefit different types of users such as transportation managers, drivers, traffic engineers as well as the urban planners.

iTrans is a platform for smart transportation which is developed by a local start-up company. It provides services for transportation managers to analyze the transportation data and to gain insights into traffic patterns. iTrans has established themselves as a popular people-centric transport solution. Under the open data initiatives of SG government, iTrans uses open data sets from LTA DataMall and NEA.

Current System Architecture

Currently, the target users of iTrans are the transportation managers. They can query the data of interest and efficiently display the traffic flow conditions on major urban streets within the scope of the city. They can also use the platform to analyze the traffic data and find out the hidden traffic patterns in the city.

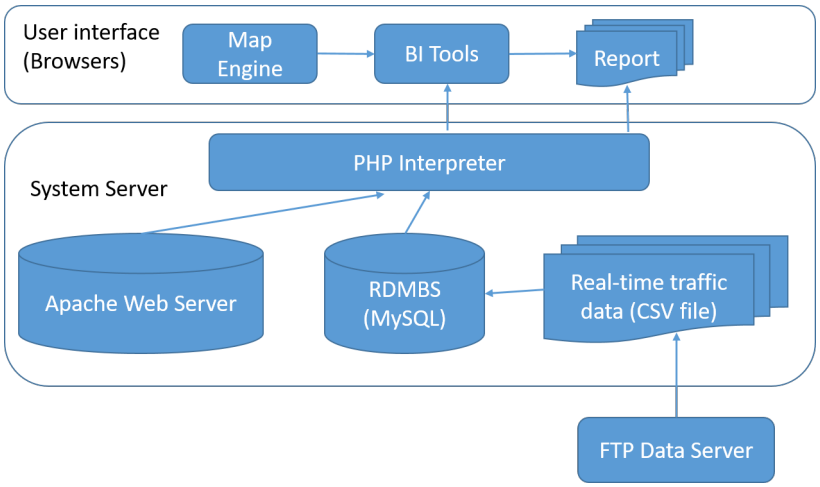

Figure 1 Current Architecture of iTrans

The current architecture layering of the platform is shown in Figure 1. The current iTrans platform consists of three layers: a data layer, a system server layer and User Interface (UI). All data collected for analysis (either batch or near real time) are from iTrans propriety devices.

External and Internal Data Sources

The real-time traffic speed data is provided by our LTA DataMall port. This data is ingested and is stored in the local file system and can be accessed through the File Transfer Protocol (FTP). The data is downloaded every minute and made available for system usage. All data is converted into a Comma-Separated-Value (CSV) file and stored into Relational Database Management System (RDBMS) such as MySQL. The Apache HTTP server was chosen as the web server. PHP is used for setting up a communication bridge connecting these two servers. A user-oriented interface has been constructed. Ajax is used to create interactive web application. Google Maps Application Programming Interface (API) is the core display engine of the system.

The system usage can be described as a transactional query-based dashboard. Users can send various queries via the UI, the Ajax request will be sent through PHP interpreter. The PHP interpreter will in turn query the MySQL databases. Once the query is processed, data in XML format is sent back to the dashboard. Google Maps API visualize this result as a comprehensive map plot as necessary. The real-time flow map mechanism is also incorporated into the current system. XML files are generated in real-time while FTP Downloader downloads the CSV files from the data server. Users' browsers can update the real-time flow map based on the up-to-date XML files.

Challenges due to Growth

iTrans platform is facing challenges due to increasing data volume and variety. There is an increase in the request for embedded devices, such as traffic sensors, cameras, GPS-devices, and smartphones and all these devices are capable of interconnecting and communicating with each other over the Internet. Data from those devices have a variety of forms, i.e., it can be either structured or un-structured. Reports and visualization of current traffic monitoring system are generated using traditional BI tools and execution happens through batch processing. These have to be upgraded to enable dynamic real time information processing for faster decisions and insights generation. The company management now wants to collect data from participating commuter devices such as mobile phones, IoT and Arduino enabled edge devices and ingest into iTrans platform to do further analysis.

Management Directions for Applications of Analytics

The company management has identified the following data characteristics of the current platform:

- The current iTrans platform uses FTP to download data from the LTA DataMall port, the data is in the CSV format;
- All data are stored and processed using RDBMS. End user can use UI to send request through Ajax. Google Maps API is used to display the query results on the map.
- The current iTrans platform only displays traffic flow data using the up-to-date web technologies driven by the database.
- Apart from transportation managers, the company also wants to target other potential users such as road drivers, traffic engineers and urban planners. They have different functional requirements which are summarized as follows:
    - Daily, monthly, yearly transportation patterns for some specific roads
    - Traffic congestion real-time update

- Accident report and update
- Route recommendation
- Crowd mobility analysis

In order to fulfill the above requirements, the company has engaged multiple teams to upgrade the existing system and derive business advantage through Big Data utilization. You are part of the teams as the Big Data Specialist.

Enhancement to the Platform

The company management wants to achieve the following:

- Collect data from multiple data sources with variety of formats which may or may not be structured.
- Store and process large amount of data with sizes varying from terabytes to petabytes.
- A technical architecture that can support spatiotemporal, graph, statistical, prediction and fusion queries. The process can be batch, stream and graph processing.
- Process real-time data including video streams.

The existing data collection layer was designed on an ad-hoc basis. Though there are existing tools such as file read libraries, and batch scripts, data manipulation is not automated in iTrans platform. The current data architecture includes ingestion, query processing and visualization layers as shown in Figure 2 below.

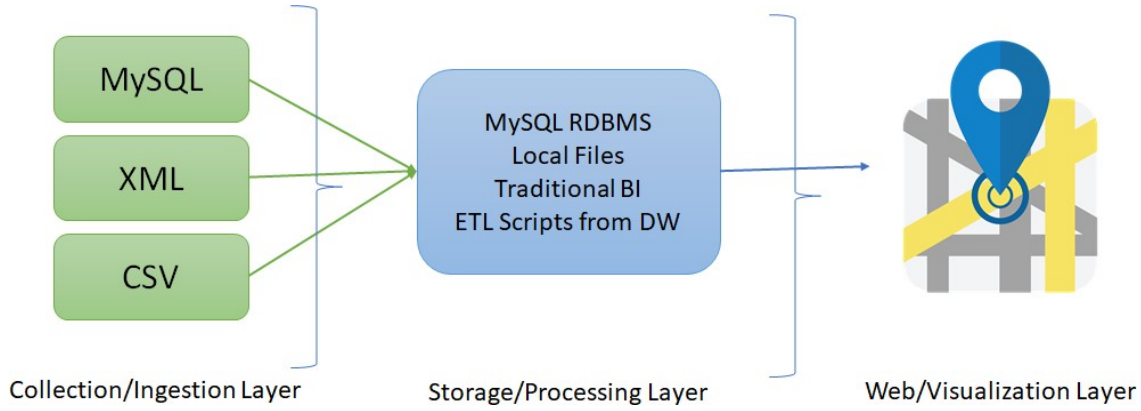

Figure 1 Data Architecture

The data collection/ingestion layer ingests data from limited data sources (CSV, XML and RDBMS only), performs data preprocessing for visualization. However, the architecture has issues with elaborate batch processing. Also, the architecture does not support machine learning, stream processing and graph processing outcomes. The current architecture only supports ad-hoc queries to display the traffic flow conditions on major urban streets within the scope of the city. Due to the time resource limitations, the company management decides to focus on the following enhancements and requirements on the current iTrans platform. The enhanced platform has major improvements on the traffic congestion monitoring, crowd mobility analysis and a new application for "Intelligent Accident Identification". Some key issues and information regarding these three use cases are provided below:

- Traffic congestion monitoring: This project currently collects traffic flow data from LTA DataMall. The system can also collect data from preferred routes and predict

expected travel time from the same data mall. As a new initiative, the platform now is expected to track sensors from more than 5,000 partner logistics vehicles and the real-time location data of buses to facilitate better transport planning to meet user demands. Thus, iTrans wishes to incorporate planning and prediction features, especially in CBD areas, to monitor additional activities such as actual travel time, route choices and queue length at different timings. Monitoring the system with this increased surge of signals from IoT devices is a challenge. Additional data fields to be handled are:

  - Average Traffic movement/density: current traffic movement on the roads from LTA DataMall as a stream of structured data.
  - Fleet location, individual speed and fuel status via fleet's mobile devices and cellular network communication.
  - Geospatial information: whole island geospatial information
  - Current climate information from NEA data sharing platform

- Crowd mobility analysis: Call Detail Record (CDR) is the information captured by the telecom companies when clients make call, send SMS and connect to the internet. It can be used to analysis people crowd mobility behaviours in a city. It would benefit the public transportation by identifying the people density. Traffic managers can map new routes to decongest crowded roads and reduce travel time.
- Accidents identification: CCTV cameras are installed island wide with permission from LTA. Live video streams from these cameras can be used to identify the characteristics of urban road traffic accidents. Images collected through CCTV cameras at intersection of roads can be sent to the enhanced platform to understand and analyse the road traffic accidents on a real time basis.

Closing Remarks

In addition to the current capabilities of the existing platform, the proposed new platform will integrate and enable more enhanced capabilities. It will have the ability to process and handle increased amount of structured and unstructured data from a larger variety of sources (sensors, mobile devices, cameras) that are aggregated by the platform. The platform will clean, transform and use the data to make predictions and analysis on specific outcomes. The enhanced platform will incorporate prediction and reasoning feature as it stores the data and uses it for data modelling via statistical and machine learning algorithms leveraging different types of traffic data. The enhanced platform will provide recommendation routes for drivers, up-to-date traffic status for transportation managers, and help the urban planners predict the future traffic volume identifying the bottlenecks for better road planning by using the results of current traffic data analysis.