Pham, Tran-Vu; Tran, Quang-Minh; Truong, Linh; Dam, Khanh-Hoa

Smarter big data analytics for traffic applications in developing countries

Published in:
Big Data-Enabled Internet of Things

DOI:
10.1049/PBPC025E_ch2

Published: 01/01/2019

Please cite the original version:
Smarter Big Data Analytics for Traffic Applications in Developing Countries

Tran Vu Pham and Quang Tran Minh
Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology, VNU-HCM, Ho Chi Minh City, Vietnam
{ptvu, quangtran}@hcmut.edu.vn

Hong-Linh Truong
Department of Computer Science, School of Science, Aalto University, Finland
linh.truong@aalto.fi

Hoa Dam
School of Computing and Information Technology, University of Wollongong, Australia
hoa@uow.edu.au

Abstract

Internet of Things (IoT) generate huge amount of data in real time. Utilization of such data requires appropriate data storage, analytics and computation techniques so that valuable information can be extracted and immediate actions could be taken place. In this paper, we present real time traffic applications, typical IoT-enabled big data applications in smart city context, but for unreliable and fragmented infrastructures in developing countries. We analyze a case study to show challenges that we need to address when implementing big IoT data applications in such infrastructures. Based on that we outline our key design principles and techniques. We present several examples to particularly discuss how we achieve our designs and lesson learned.

1. Introduction

Real time traffic monitoring systems use data generated from IoT devices equipped with global positioning systems (GPS) mounted in vehicles moving on city road networks to extract the current states of the traffic, i.e., traffic flows of roads. The data is then analyzed and traffic information is disseminated to the public on a Web portal and mobile applications. Users can obtain up-to-date information about the traffic and make proper traveling plans to avoid areas having bad traffic conditions. Such systems are well studied and deployed in many cities. These systems often have common characteristics w.r.t. communications, large-scale data storage and high performance computing systems to support traffic analytics. However, most systems are developed and deployed in well infrastructures or in the context of developed countries.

Our experiences in developing such systems are different: we are building such systems in the context of Vietnam, specifically, and in developing countries, generally. In our context, we have faced several different issues w.r.t fragmented ICT infrastructures, weak and unstable networks, highly diverse quality of GPS sensors attached to vehicles, mixed types of vehicles, to name just a few, for a very hybrid multimodal traffics of motorbikes, cars and buses. Even in a not well-connected network whose coverage is limited in Vietnam, currently, our system receives data from around 10,000 vehicles at peak time. The amount of GPS data the system needs to process per day is roughly 4.5 GB, on average 600 data records per second. Due to
the fragmented network and computational infrastructures as well as diverse types of devices with different quality, in our case, we face several challenges that need novel solutions.

1.1 Research Challenges
We will discuss challenges and requirement analysis centered on the following points:

- **Real time processing of data**: even under limited capabilities of networks and computational resources, the data still needs to be processed in a real time manner so that decisions can be made immediately to response to the current traffic conditions. Because of this requirement, the proper solutions need to be balanced between processing time and the accuracy of results.

- **Noises in data**: the system receives huge amount of data, but only a fraction of data is with expected accuracy and currency, due to quality of devices, noises, and network delay/reliability. This problem is severe due to the quality of IoT devices and networks in developing environments. For example, the data may come from many vehicles that are stationary, not participating in the traffic. Late arrival of data due to network problems and low quality of GPS enabled devices are also major sources of noises. The challenge here is how to separate the noise in data in real time before any calculation can take place. Furthermore, this also requires us to devise techniques to improve quality of data not in real-time but not too late for other batch analytics.

- **Heterogeneity of data sources**: given the heterogeneity and diversity of IoT devices, GPS signals are generated from different GPS enabled devices equipped on various means of transportation such as buses, trucks, taxis, mobile users on cars or motorcycles, and so on. Each means of transportation has different traveling behaviors, rate of noises, and quality of data. Therefore, each of these sources of data needs to be processed differently.

- **Large amount of data**: the size of data causes challenges in processing and storing. A proper storage system needs to be used so that the data can be stored and retrieved for processing at minimal delay.

1.2 Contributions and Paper Structure
In order to address the above challenges, we have developed a system framework for real time traffic applications with the following key solutions:

- A modular and scalable IoT application framework that support various traffic applications in smart city domain,
- A scalable key-value RDMA-based storage system for caching data during processing,
- Real time classification of data and noise reduction using multithreading, and
- Simple but fast analytic techniques for GPS signals.

In this chapter, we will present some examples to illustrate our work. We describe details of the traffic monitoring system for demonstration purpose.

The chapter will include five main parts: Section 2 presents scenarios and requirements. Section 3 describes our analytics framework. We present our big data applications and challenges in Section 4. We discuss related work in Section 5 and conclude the paper in Section 6.

2. Scenario and Requirements

We analyzed the case of traffic management in Ho Chi Minh City (HCMC), the most populous city in Vietnam, to illustrate applications of IoT and Big Data to solve critical problems in urban areas. Currently, there are more than 10 millions of people living in the city. HCMC is pushing toward a smart city program, which aims at utilizing modern
technologies, such as IoT, machine learning, and high performance computing, to improve the living conditions of the city’s residences. The challenges that HCMC is facing, as a mega city in Asia, are city management, transportation (e.g. traffic jams), air pollution, urban flooding, to name just a few. In the scope of this chapter, we focus on transportation issues.

Per regulations, public service vehicles, such as buses, taxis, trucks, and coaches, have to install GPS enabled devices, and frequently send GPS signals in real time to a government office for monitoring their locations and driving behaviors. Typically, the GPS signal consists of ID of the vehicle ID, longitude, latitude, velocity, and timestamp. The GPS data may contain other optional information, such as petrol level, and engine status. The city Transportation Department expects to receive monitoring data from about 67,000 vehicles regularly. Our demonstrating system involves a smaller number of vehicles, about 10,000 at peak time. In addition to the GPS signals from vehicles, we also developed mobile applications for end users (especially motorbikes), enabling them to contribute some limited amount of GPS data to our system.

![Figure 1: Screenshot of Traffic Information System for monitoring traffic conditions of Ho Chi Minh City in real time. The color labels show the traffic flow velocities of the roads.]

The real-time GPS data is valuable for a number of applications in scope of the HCMC smart city program. It can be used for extracting the current state of the traffic in the city to detect traffic jams; one of the most crucial problems and it is the most well known function of our demonstrating applications. A screenshot of the application is shown in Figure 1. In addition, the data can also be used for other types of applications for city residences, such as:

- Predicting arrival times of buses at bus stops.
- Predicting arrival times and routing for taxis.
- Navigating commuters to avoid traffic jams.
- Predicting traffic jams.
- Predicting traveling directions of mobile users and alerting them with real time traffic conditions.

The common characteristic of these applications is that they all need to analyze real time GPS signals to get insights (knowledge) from the data, and then, provide recommendations to the
users in real time. The general processing flow for applications is illustrated in Figure 2. However, for each application, real time GPS data is analyzed differently. For example, for extracting traffic conditions, the analysis is focused on estimating traffic flow velocity and density using velocity information in the signals. For predicting bus arrival times, the analysis module needs to track individual buses’ movement, combines with the traffic flow velocity to make the estimation. Therefore, efficient control the flow of real time GPS data in the system is very important to the applications.

![Figure 2: GPS signal processing flow for traffic applications](image)

The GPS data from different sources are very different in terms of quality, quantity and other characteristics. In terms of quality, devices from different manufacturers have different in quality, and hence, leading to the variety of the data accuracy. The quality of the data is also affected by the quality of communication networks used. Currently, most GPS devices on vehicles use GPRS services from mobile networks. GPS signals from mobile devices, via mostly on at 3G networks, often give best accuracy. However, mobile devices often have problem with batteries. Therefore, their online time is very intermittent. Another issue is that the quality of mobile networks is changing very different, even within a small area, causing many issues for data delivery. In terms of quantity, the frequency that signals are sent to the servers are supposed to be a signal per 15 second for each device. However, in reality, the frequency for each device may be different due to network condition, quality of devices, and physical setting (thus we allow, for mobile devices, the frequency to be set by the users). The characteristic of GPS data is also very different. As buses run mostly on major roads of the city at regular intervals, they reflect well the traffic conditions of these roads. However, if we rely only on bus data, we lack of information for smaller roads. The coverage of taxis and trucks are larger, but their traveling pattern is irregular. Coaches move mostly in outer city roads for destinations at long distances, and therefore, data from coaches is not very useful for inner city applications. Mobile (motorbike) users can be anywhere, such as on roads, in houses, offices, or coffee bars. Hence, deciding whether a mobile user is on roads or not is also a challenging issue. Most of the time the vehicles are in idle states, but still sending real time signals to servers. To know whether a vehicle is stuck in a traffic jam or in idle state is also a hard problem.

In summary, the above-mentioned issues make the development of applications related to traffic very challenging. Many constraints and techniques need to be developed for specific developing countries that are very different from developed environments.

3. Analytics System Framework for Traffic Applications

3.1 Design objectives
To address issues described in Section 2, we have designed a system framework for utilizing GPS data provided by different sources from the city that tackles the following requirements:

- Able to process large volume of GPS data in real time to give timely responses to end users.
- Efficiently deal with the heterogeneity of GPS data from various sources, with differences in quality, quantity and characteristic.
- Allow different analytic modules, both real time and offline, to be plugged-in the system to support different kinds of applications.
In the following, we will discuss the framework and our unique design features.

### 3.2 Framework Overview

Figure 3 depicts the overall system framework. Generally, the system takes as inputs different GPS data sources by GPS Data Logger. The Raw Data is stored in database for offline analytics and archiving. A copy of Raw Data is also kept in Memory Storage for Real time Analytics. Because different sources of GPS data have different characteristics, and each application requires to process the data differently, the Raw Data is classified and fed in different Analytics modules by Data Router. As the volume of data is large, real time analytics must be done in parallel. The Raw Data can also be processed by Offline Analytics modules to build data models for different purposes (e.g. prediction traffic conditions, and user movement). The Data Model extracted by Offline Analytics modules can be utilized to improve the quality of Real time Analytics. When all the analytics are done with the data, the Decision Maker finalizes the results, and stores them in databases for later use by Web Applications or Mobile Applications. For quick responses to user requests, copies of the results are also cached in Memory Storage.

![Figure 3: Analytics system framework for GPS data](image)

### 3.3 GPS Data Providers

The GPS Data comes to the system from many different sources, such as cars, buses, taxis, trucks, coaches and mobile devices. Except for mobiles, GPS signals from other sources do not come directly to our servers. Instead, the signals from devices on cars, buses, etc. are collected by their device providers, named as GPS Data Providers, via GPRS/3G networks. The providers then send the collected data to our RabbitMQ server in real time via the Internet. The data received by RabbitMQ Server is forwarded to GPS Data Logger for archiving and further processing. Because the data has to be processed at multiple steps before coming to our server, the delays of data packages when arriving at the system from different providers may be varied.

For mobile devices, as we build our own applications for collecting GPS data from mobile users, the GPS signals are sent directly to Data Logger using network connections built atop the UDP protocol. Even though the UDP protocol is not reliable, the use of UDP helps the system maintenance become easier. As connections of mobile devices to the Internet are intermittent, with UDP we do not need to continuously monitor and reset up the connections when there are network interruptions.
3.4 Offline Analytics
The framework is designed to allow the system to work with different offline analytics modules. Recall that our infrastructures are fragmented: this offline analytics is not just due to different types of postmortem or predictive traffic analytics but also due to the need to prioritize and balance computation demands. Depending on the kinds of application, different offline analytics methods can be applied. The results of offline analytics processes, usually data models, are used to improve the quality of real time analytics and final decisions. For example, for traffic notification applications, machine learning techniques can be used to extract users traveling patterns. The patterns can then be used to predict the user movements on roads in real time so that the right notifications about traffic conditions can be sent to the users. For bus arrival time prediction, machine learning techniques can also be applied to raw bus data from the past to extract bus behaviors. The knowledge is used by real time analytics modules to improve the quality of predictions.

3.5 Data Router and Real time Analytics
Real time Analytics is the core of the systems. As any IoT system, the incoming data stream needs to be constantly processed to extract immediate insights so that timely decisions can be made to respond to changes in the real environments. In the transportation domain, real time GPS data are useful to many different applications. For this reason, the framework is designed in a way that different real time analytics modules can be used the real time GPS data stream. Depending on the need of applications, the Data Router classifies the data packages in the real time data stream and forks main data streams into substreams and feeds them to the right analytics modules.

The Data Router splits the main data stream into substreams for different purposes:
- For applications that require to process large amount of data, such as monitoring traffic conditions (i.e. traffic flow velocity and density), it is not possible for one machine to process the data of the whole city, the Data Router splits the main data stream into smaller ones so that they can be analyzed in parallel to speed up the processing.
- Each data source may also have different characteristics, as analyzed in Section 2, and require different analysis techniques. In this case, the data with similar characteristics is grouped together in substreams for further processing.
- Each application may require a different set of input data. For example, for bus arrival time prediction, GPS signals from buses may only be needed. The Data Router may split the data stream depending on the need of applications.

3.6 Decision Maker
Based on the intermediate results of Real time and Offline Analytics, the Decision Maker concludes the results, which are stored for use by end users applications. For different types of application, different decision mechanisms can be applied. For example, to estimate the velocity of traffic flow of a particular road segment, the decision can be simply an average calculation of the all velocity attributes of carried by the GPS signals having coordinates falling on the road segment in a particular period of time. However, to predict a user trajectory on roads in the city, a much more complex decision model needs to be used.

3.7 Mobile and Web Applications
Mobile and Web Applications provide interfaces for users to interact with the system. We use RESTful services so that application front ends can be easily integrated with backend systems. The system can also be extensible with services oriented architectures.
4. Big Data Applications and Challenges

4.1 In-Memory Storage

One of the challenging problems when developing traffic applications is handling real time GPS data. The data comes from many different sources and needs to be processed in real time. Traditional relational databases are not suitable. The available NoSQL databases such as MongoDB\(^1\) are usable, but their performance, especially for read operations, does not meet the need of the applications. The reason is that MongoDB is key-value operation orientated. It does not take into account time and spatio characteristics of the data. For real time traffic applications, a database needs to be able to handle large amount of real time data, particularly:

- Able to store and scale with constantly increasing amount of data.
- Support many read operations with low latency, as many different applications need to read data from the data stream in real time.

![In-Memory Storage System](image)

In order to reduce latency of read operations, the storage system is separated into two layers: in-memory (cache) and persistent storages on disks, as shown in Figure 4. Initially, we have tried Memcached\(^2\) for in-memory storage. Its performance did not meet our requirements. Memcached does not scale well when the amount of data increases. By exploiting the time-spatial characteristics of GPS data, we have designed a new in-memory storage system for traffic applications. Design of the system has been improved over time and published in [1] [2] [3] [4]. Because GPS data also shares common time-spatial characteristics with other classes of IoT datasets, we believe that our in-memory system can be used efficiently for other classes of IoT applications.

---

1. https://www.mongodb.com
Figure 5 shows the geographical distribution of GPS data on an overlay of HCMC map in four continuous days. The distribution is almost static. This is explainable because the traveling patterns of city residences are repeated day by day. Based on this distribution, we have designed a scalable mechanism for distributing GPS data into data nodes. Any new indexing and querying mechanisms were also developed for the storage system. Experiment results showed that this new indexing and querying mechanisms could improve the system performance significantly in comparison with traditional methods [3].

In order to reduce read latency, we developed a Key-Value Links (KVL) data model, by utilizing Remote Direct Memory Access (RDMA) for our in-memory storage. From design perspective, KVL model is an extension of the common Key-Value (KV) data model. Each pair of KV is added with an additional link, similar to pointer, to a related data item. To fetch the related data item, the link is chased via RDMA Read operation. Because RDMA Reads can bypass kernel operations, they can help to reduce the latency significantly. This in-memory storage system architecture is described in detail in [4].
4.2 Filtering Unusable Data for Real time Analytics

GPS data sources continuously generate many signals, but not all the signals are usable for real time applications. There are many reasons that make the data is unusable:

- Inaccuracy of the data signals, due to quality of devices
- Noises in data
- Errors in map data (coordinates of roads shown on map different from their real physical locations)
- Errors in timing: unsynchronized clocks (timestamp of signal is earlier/later than the actual time), network delays/interruptions (late arrival of data packets)
- Data from stationary vehicles (not participating in traffic).

Therefore, the received data needs to be cleaned or corrected before use. This step is necessary for all traffic applications. In other words, we need to know which signals are from vehicles that are participating in the traffic to further process. The rest of the signals will be filtered out. By analyzing the data manually, we discovered the following information:

- About 35% of the data could not be mapped onto any road segments. This could be because of noises in data, inaccurate GPS signals, or errors in map data.
- About 8% of the data arrives at our server having packet timestamps either earlier or later than expected. For real time processing, we need to have data arrive within a time window, around the current time. However, due to network delays, network interruptions, unsynchronized clocks, or some other reasons, the timestamps (recorded by the devices) of the signals are outside of the current time window, hence, not usable.
- About 38 data signals have zero speed values. These data signals come from stationary vehicles, not participating in traffic. This because the GPS devices keep generating signals and send them to the servers 24/7, independently from the operations of the vehicles.

When filtering out the above data signals, only about 17% of the raw data is used for further processing in real time. Bringing more data into the calculation will significantly improve the quality of the results. However, correcting errors in data to improve its utilization is a challenging problem.

We have taken a couple of methods to improve the utilization of the data:

- For signals from buses, as buses always travels on fixed routes, we can decides the buses’ routes based on offline analytics, then, errors caused by incorrect map data or inaccurate GPS signals can be reduced by shifting the coordinates of the signals to the road segments belonging to expected routes.
- For signals from devices physically attached to vehicles, as the vehicles also traveling on roads, we can extend the margins of road segments to include more signals to the calculation when mapping the signals to road segments.
- It is a little more complicated for signals from mobile devices, as the connections of mobile devices to the Internet are intermittent, and the devices can be anywhere (e.g. on roads, coffee bars, offices, etc.). To deal with intermittent Internet connection, when the connection is dropped, the device will temporarily store in the device’s buffer. The buffered data will be uploaded to the server when the device is back online. Some of the data may be unsynchronized, hence not be used for real time calculation, but still valuable for offline analytics. To detect whether a device participates in traffic or not, we have built an offline model based on neural networks to learn the moving characteristics of mobile users. The model can then be applied to detect the traveling status of mobile users.
The above actions could help to improve slightly the utilization of the data, but the challenge we are still facing is how to integrate the models resulted from offline analytics to real time calculations. If too much calculation and complicated models are required for real time processing, it will cause more delay in delivering final results. We need to consider the tradeoff between accuracy of computation and delivering time.

4.3 Traffic Monitoring and Prediction
Providing real time traffic information is currently the main functions of our demonstrating system. It is currently running live at http://traffic.hcmut.edu.vn (as only a demonstrating application, the real time GPS data fed in the system is varied from time to time). Figure 1 shows a screenshot of the system. The overall system architecture is depicted in Figure 4, which is a specific implementation of the analytics system framework described in Section 3.

The main objective of the system is to extract the current traffic conditions (in terms of traffic flow velocity) of city road segments and show the results in the city map. In order to get the traffic flow velocity of a particular road segment (limited by two continuous intersections), we gather all the GPS signals having the coordinates fallen in the segment in the observation time window (currently, it is period of 15 minutes from the current time back). The flow velocity is the average of all the speed values carried by the GPS signals. As the number of road segments in the city is quite large, the calculations are done in parallel.

The advantage of the current method is that it is very simple, with minimal amount of calculations in real time, but gives acceptable results. However, to further improve the accuracy of the results, a number of issues need to be addressed:

- What is the best time window for a particular segment? If the time window is small, there will be less GPS signals fallen in the segment, then the estimated results will not be reliable. However, if the time window is large, it will not reflect well the current state of the traffic. Furthermore, GPS signals are not distributed equally on city road segments. Using the same time window may be good for some segments but not appropriate for other segments.
- How GPS data from heterogeneous sources, with different level of accuracy, contribute to the final estimation? Currently, because the number of GPS signals fallen within a segment is relative small, a simple averaging is acceptable. However, when there is a mix of signals from different sources, a more appropriate method may be required.
- How historical data can help to improve the accuracy of real time calculation? Furthermore, how the current and past data can help to predict the traffic in near future? This issue leads to traffic prediction problem.
- How to inform city residences the current state of city traffic? Broadcasting traffic information to all city commuters is not a good option, because people tend to travel in a particular area of the city interested in some particular types of information. Receiving too much information that is not relevant to user context will result negative impression. To send proper information to users, we need to learn user habit (traveling patterns and interests). This issue needs to be addressed.

Predicting future events in traffic (e.g. arrival times and routing for taxis, traffic jams, or traveling directions) is highly challenging due to the inherent uncertainty, temporal and spatial dependencies, and the dynamic nature of traffic. The analytics component of our framework therefore plays an important part in learning from potentially large datasets and automatically forming a deep understand of traffic patterns and dependencies. Such an understanding along with relevant memory of experience will be helpful in making predictions and relevant interventions.
A number of challenges exist in building an effective traffic model. First, we need to capture various dimensions of dependencies, including spatial and temporal dependencies (e.g. time and location of vehicles). Second, handling long-term dependencies: future traffic events may depend on previous events and interventions. For example, a car accident occurred on a road may result in a traffic jam on that road a few minutes later. Third, accurate traffic prediction requires a sufficiently rich representation of traffic events. The challenge here is how features representing a traffic event can be automatically learned from GPS data. Fourth, we need to model the episodic and irregular timing in traffic. Traffic can be seen as a series of episodes and time between traffic events are largely random. We thus need to address the episodic and irregular timing of events of interests when building traffic prediction models. Finally, modeling confounding interactions between traffic events and interventions. For example, directing traffic from a congested road to other roads (as an intervention) may result in congestions elsewhere.

The recent advances of deep learning techniques in machine learning offer a powerful solution to address the above challenges. Deep learning techniques are capable of dealing with dimensionality issues and modeling deep complex non-linear dependencies with distributed and hierarchical feature representation. One of the most widely used deep learning models is Long Short-Term Memory (LSTM), a special kind of recurrent neural network that is highly effective in learning long-term dependencies in sequential data such as time series in traffic data. Currently, we focus on developing an end-to-end traffic prediction model, which does not require manual feature engineering; i.e. features are automatically learned from GPS data. Our model will be based on LSTM, and thus will be capable of reading historical traffic data, memorizing a long history of experience, inferring the current state of traffic, predicting future events and risks, and finally recommending actionable interventions.

We have implemented a deep learning model for predicting traffic conditions of road segments based on LSTM. The inputs to the model are traffic conditions of the predicted segments from the past. Experimental results showed that the current model is good for extracting regular traffic patterns from past data and using the patterns for the prediction. However, accuracy of the prediction is limited when there are ad hoc traffic incidents. The model needs to be revised in order to improve the quality of prediction.

4.4 Trip Planning in City Bus Networks
With large amount of GPS data received from city buses, we also use the data to improve bus services. We have developed solutions to help people plan their trips when traveling with city buses. The two challenges we need to address: (i) predicting arrival times of buses on a particular bus route, and (ii) planning optimal traveling routes from a user current location to desired destination. These problems are even more challenging because of current the conditions of HCMC city bus network. The city are suffereing from regular traffic jams, therefore, it is almost impossible to have an on time bus timetable. The arrival time of the current bus is relative to the last one on the same route. Hence, when planning for a trip, the current state of the bus network, including locations of buses and traffic conditions, needs to be taken into account.

In order to deal with the above issues, we model the bus network as a stochastic schedule-based transit network. A number of solutions that utilize both real time and offline analytics using GPS signals from the city have been developed:

- Planning traveling path on city bus network with least expected travel time and minimum number of transfers: the solution is based on correct-labeling algorithm and Bellman’s principle of optimality. Experimental results show that running time of the algorithm is suitable for real time applications [5].
- Finding reliable path with earliest arrival time: due to stochastic nature of the transit network, the success of a planned route, involved a number transfers, depends on the
Finding alternative traveling paths on bus network: the commuters may have personal preferences on bus routes. Therefore, we developed a path finding algorithm that returns multiple similar cost paths so that they can choose a path to travel based on their preferences [7].

We have developed, experimented and published technical details of these solutions. However, due to lack of human resources, we have not yet implemented the solutions in our live demonstrating system.

5. Related Work

Recent work related to this chapter from two closely related domains: Smart Traffic (or Intelligent Transportation Systems - ITS) and, broadly, Smart City. There have been many systems for traffic management as well as the use of big data techniques for such management systems is known [8] [9] [10]. In this chapter, we focus more on specific characteristics and techniques that are designed for solving issues related to developing countries.

In Smart Traffic domain, one of the earliest proposals for ITS system architecture for developing countries could be from the World Bank [11]. However, this proposal only focuses the functional and service aspects of ITS; such features are common in various architectures. Technical solutions, such as how data should be stored and processed in fragmented environments, are not mentioned. Also in Smart Traffic domain, there are other suggested system architectures for building analytics applications [12] [13] [14]. In general, we have seen the widely use of big data storage, streaming processing and services in the cloud for ITS but there is a lack of discussions w.r.t how to deal the lack of resources, unreliable networks and bad quality of data.

In Smart City, different system architectures have also been introduced [15] [16] [17]. Similarly proposed architectures for traffic applications, they are very generic [15], very specific to parallel data processing [16], or utilization of network technologies [17], whereas, in our work, we aims at support different analytics methods and applications.

In ubiquitous computing and sensor networks, we have seen many works discussed techniques to deal with data quality [18] [19] [20] [21] and smart city network connectivity issues [22]. However, such works have not been integrated into the big data pipelines, which bring real time data from Things to the analytics algorithms through various phases. In our work, we particularly focus on developing suitable solutions that are well integrated with the big data pipelines, and deployed and tested with traffic in developing countries.

6. Conclusions

In this chapter, we have introduced a scalable, IoT based system framework for traffic applications in HCMC, Vietnam. This system utilizes the GPS data generated by buses, taxis, trucks, coaches, etc. and mobile devices for real time traffic application such as traffic monitoring, traffic predictions, bus trip planning. The key advantages of the framework are: (i) supporting a variety of real time applications that utilize the same set of real time data, (ii) allowing the integration of real time and offline analytics to improve the quality of results, and (iii) providing mechanism for splitting/duplicating real time data stream into substreams for further analysis by specific applications and/or speeding up the calculation by parallel processing. The chapter also reports a number of demonstrating applications we have developed, and discusses their current status and how big data analytics techniques could be employed to improve quality of results. As learned from developing these applications, with
real time analytics system, there should always be a combination of real time and offline analytics, and for particular applications, the tradeoff between the cost of real time computation and quality of results need to be considered.

References


[11] Toshiyuki Yokota and Richard J. Weiland, "ITS System Architectures For Developing Countries," Transport and Urban Development Department,


[21] Barnaghi, Payam; Bermudez-Edo, Maria; Tönjes, Ralf, "Challenges for Quality of Data in Smart Cities," *J. Data and Information Quality*, vol. 6, no. 2-3, June 2015.
