POTENTIALS OF ONLINE MEDIA AND LOCATION-BASED BIG DATA FOR URBAN TRANSIT NETWORKS IN DEVELOPING COUNTRIES

Kelsey Lantz*
Ph.D. Student
Glenn Department of Civil Engineering
131 Lowry Hall, Clemson University, SC 29634
Email: klantz@clemson.edu
Tel: +1(724) 825-5951, Fax: +1-(864) 656-2670

Sakib Khan
Ph.D. Student
Glenn Department of Civil Engineering
18G Lowry Hall, Clemson University, SC 29634
Email: sakibk@clemson.edu
Tel: +1(864) 656-3313, Fax: +1-(864) 656-2670

Linh B. Ngo
Research Associate
Big Data Systems Lab, School of Computing
222B, McAdams Hall, Clemson University, SC 29634
Email: lngo@clemson.edu
Tel: +1 (864) 656-5769, Fax: +1-(864) 656-2670

Mashrur Chowdhury
Eugene Douglas Mays Professor of Transportation
Glenn Department of Civil Engineering
216 Lowry Hall, Clemson University, SC 29634
Email: mac@clemson.edu
Tel: +1-(864) 656-3313, Fax: +1-(864) 656-2670

Sarah Donaher
Undergraduate Student
Glenn Department of Civil Engineering
131 Lowry Hall, Clemson University, SC 29634
Email: sdonaher@clemson.edu
Tel: +1(704) 562-2326, Fax: +1-(864) 656-2670

Amy Apon
Professor, Chair of Computer Science Division
School of Computing
221 McAdams Hall, Clemson University, SC 29634
Email: aapon@clemson.edu
Tel: +1 (864) 656-5769, Fax: +1-(864) 656-2670

*Corresponding author
Number of Figures = 4, and Tables = 2; Word count= 8230 (6730 words +250*6)
ABSTRACT

Big data, collected in the form of social media posts and mobile phone location tracking has great potential to inform and manage the planning and operation of transit networks in developing countries. Data is widely available but the challenge, as with developed countries, is figuring out how to best use it. This paper uses a case study method to look at approaches in Nairobi, Kenya; Istanbul, Turkey; and Dhaka, Bangladesh. In Nairobi, Global Positioning System (GPS) location data has been collected to generate the first map of the complex Matatu transit network. In Istanbul, automated fare collection systems are processed to better understand the usage of a Bus Rapid Transit (BRT) system. In Dhaka, researchers are collecting GPS positioning data to manage the city bus networks. Residents of these developing cities are frequent users of on-line media; similar to many other cities in the developing countries. This study revealed that integration of on-line media with location-based data provide a big data scenario, which have potentials for supporting transit operations while posing challenges to manage the data mobility. Of course it is not realistic to apply a one-size-fits-all approach to any problem in the developing world, but together, the case studies show that with the right approach, technical capacity in transitional cities has the potential to grow to support higher-level data processing to make more efficient and more sustainable policy decisions for crucial urban transit networks in developing countries.
INTRODUCTION

Crippling congestion, deteriorating air quality, decreasing road safety and massive growth in energy consumption are symptoms of rapid motorization in developing cities around the world. Transit networks are a critical piece of the mobility puzzle because their effective operation can prevent would-be motorists from taking to the already congested roadways, and in general they move more people with a smaller carbon footprint (1). Leapfrogging beyond traditional traffic sensing techniques, developing countries may capitalize on the technology available today to harness widely available big data to better understand the operation of formal and informal transit networks and subsequently optimize their planning and operation.

Big data is a blanket term that is typically used to describe data that are too complex and too large to be processed using traditional storage and analytic infrastructures. By aggregating prevalent and publicly accessible data sources, big data can provide a wealth of opportunities for improving the planning and operation of the crucial transit networks that cities in developing countries rely on. Growing economies are increasingly taking advantage of data-intensive technologies(2). With the right framework, big data-based knowledge can inform intelligent decision-making and optimize transit networks, thereby saving time, money, energy, and lives through improved safety and reduced travel time. This paper seeks to identify some of the benefits and challenges of using big data to plan and operate urban transit systems and to subsequently develop a replicable architecture for communication within the relevant stakeholders. To do so, we explore case studies in Nairobi, Kenya; Istanbul, Turkey; and Dhaka, Bangladesh to further investigate the practical applications of big data in transit planning and operation.

Developing countries that lack conventional transportation infrastructure have the unique benefit of being able to capitalize on lessons learned from industrialized countries. The problems that plague an urban transportation network are well known and the solutions do exist(1). Rapid motorization presents a complicated dilemma: rising income growth and the subsequent rise in demand for private vehicles has propelled the transportation sector to be forecasted as the highest source of future greenhouse gas emissions. However, to restrict car use is to restrict the personal freedom of potential drivers and can be expected to be wildly politically unpopular. Ideally, automobile substitutes can be improved in these developing cities. However, this requires significant investments into physical infrastructure solutions such as heavy rail and popular Bus rapid transit (BRT) networks. This could be a challenge in term of the amount of time, effort and money for developing countries. Intelligent transportation systems (ITS) with new software and hardware subsystems must accompany this improvement in physical transit networks in order to face the complex challenges of future urban mobility(3,4). An integrated approach that includes both the hardware of new transit infrastructure and the “software” of system management is critical for urban planners in developing cities (5). Initial transit development policies can be adopted to reduce the impacts on existing physical infrastructures and to reduce the costs to the improvement process.

The planning of any transit development strategy consists of many decisions that are plagued with varying levels of uncertainty. Theoretically, every decision has a certain probability based on some prior information. If we can improve the quality of prior information, this uncertainty will be reduced. This reflects a well-established mathematical theorem of information theory that supports several types of statistical and probabilistic analysis (6). Data available from on-line sources, such as social media and weather services can provide significant data to help the decision making process. However, with the availability with large amount of
data comes the challenge of managing these “Big Data” efficiently and effectively. The objectives of this study are to assess the potential of on-line data sources for supporting public transit operations in developing counties and identify strategies to manage these big data in the context of developing counties.

**Big Data**

For this study, we analyze how certain varieties of big data can be collected and processed to improve the functionality of urban transit systems in developing countries. These data sets include two types of data: location-based data from mobile devices and text-based data from online media sources including Twitter, Facebook, FourSquare, and weather reporting services.

Most individual media posts contain little informational value, but the aggregation of millions of messages over time can reveal important patterns. In absence of other geographic input, texts embedded in social media posts can carry large amounts of information. One of the most significant advantages of using location-based social media data is the valuable ability to track activity purposes, which can lead to more comprehensive planning abilities in the future. Although large amounts of data make sampling error irrelevant, this does not make the sample necessarily representative. While mobile penetration in the developing world is high, Internet access is notably less universal and does therefore not represent an accurate cross section of potential transit ridership. It is worth remembering that Twitter does not evenly represent all people. However, content-based approaches to geo-locating Twitter users is currently being used for public health research and similar methods could be applied to collecting and processing data regarding transit systems. Alternatively, location-based data is a direct report of a physical location, transmitted through a number of sources, including: in-person card payment data, GPS chips in mobile devices, and cell-tower triangulation on mobile devices.

Both formats of data can be configured into an open-source General Transit Feed Specification (GTFS). Public transit schedules and their associated geographic information are defined in this common format. These feeds permit transit agencies to make their data available to developers in order for them to write various applications that utilize that data in an compatible way, making it usable for future uses. By its open source nature, GTFS represents an accessible and wide-ranging platform for future expansion of the initiative to harness big data for the planning and operation of transit networks.

Big data typically refers to data sets with sizes beyond the ability of traditional software tools to capture, curate, manage and process the data within a reasonable amount of time. At a glance, GIS data do not seem to take up a lot of space. Minimally, a GPS record includes latitude, longitude, and a time stamp. In commercial systems, latitude and longitude are typically stored as type double, which occupies 8 bytes for each value. The time stamp can be stored using 4 bytes. This brings the minimal size of a GPS record to 20 bytes. In typical GPS map-matching process, the sampling rates of GPS data on one device can be as high as once every 10 seconds (8640 samples per device per day) or as low as once every 2 minutes (720 samples per device per day). The rates of GPS data collected daily can be calculated as follows:

\[
\text{High-rate bytes per day} = 20 \times 8640 \times \text{number of GPS devices}
\]

\[
\text{Low-rate bytes per day} = 20 \times 720 \times \text{number of GPS devices}
\]

With these equations, we can estimate the amount of data collected daily through GPS devices for Dhaka, Nairobi, and Istanbul shown below in Table 1. The number of GPS-enabled devices
was approximated by multiplying the percent of mobile phone ownership in each city by its transit ridership (14)(15).

**TABLE 1. Bandwidth Required in Each Case Study City**

<table>
<thead>
<tr>
<th>City</th>
<th>Number of devices (GPS-enabled cell phones/GPS-enabled vehicles)</th>
<th>Daily GPS data collected with a high sampling rate (Gigabytes)</th>
<th>Daily GPS data collected with a low sampling rate (Gigabytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nairobi</td>
<td>1,800,000</td>
<td>311.04</td>
<td>25.9</td>
</tr>
<tr>
<td>Istanbul</td>
<td>12,754,334</td>
<td>2203.9</td>
<td>183.6</td>
</tr>
<tr>
<td>Dhaka</td>
<td>6,000,000</td>
<td>1036.8</td>
<td>86.4</td>
</tr>
</tbody>
</table>

In the lowest sampling case, the city of Dhaka will acquire 39 Terabytes (TB) of data every year, and in the highest sampling case, the city of Istanbul will acquire 785 TB of data every year. In term of network bandwidth, we are looking at 8 Mbps (megabits per second) for the lowest case and 208.8 Mbps for the highest case. This network bandwidth does not account for potential bandwidth expansion for video stream traffic camera, which could take up to 600 Kbps (kilobits per second) for each camera (16). Video traffic monitoring could improve the accuracy of traffic management and act as an important component in future ITS initiatives. In another example, city of London’s CCTV system requires a total bandwidth of 1.2 gigabits per second (Gbps) to support up to 500 cameras (17). These estimations give us an idea about the necessary storage capacity and network bandwidth required to support these GPS data streams.

To put this into perspective, the Palmetto Supercomputer, ranked in the top 5 fastest supercomputers for U.S. academic institutions, has only 218 TB for daily storage. The Palmetto Supercomputer has two dedicated T1 lines at 1.54 Mbps per line to connect to the Internet.

For online media data, we include not only communication platforms such as Twitter or Facebook but also data sources that can augment traffic such as weather information, online resources listing public and private events that can impact traffic, and data feeds. These data come in a variety of formats, including and not limited to PDF, CSV, and structured/unstructured XML (18).

Through the above analysis, data coming from GPS-enabled devices and online media sources for traffic management purposes can be classified as big data. We have large volume of data under a variety of formats coming in at a high velocity with a certain degree of veracity due to missing and/or duplicated GPS data or ambiguous/incorrect online data (19). This motivates the need for an Information Technology (IT) infrastructure that can support big data storage and analytics.

Given the high volumes of these data sets collected over time, the capacity of public agencies in developing cities to process the data can be quickly outstripped. Fortunately, there is a global movement to make transit data open for several reasons. Through the Open Government Partnership, governments from more than 55 countries have committed to empower citizens, encourage transparency, condemn corruption, and take advantage of novel technologies in order to strengthen governance with a strong emphasis on open data as a way to achieve this (20). The open source provision of structured datasets by governments will contribute to improving the transparency and accountability of states, enabling new forms of civic engagement, and stimulating economic growth and development (20). For a policy to be effective, planning should be driven by inclusive dialogue, technical expertise and available data. Making more transport
data available enables technology innovation that supports better operations. Additionally, standardizing the format through a structure like GTFS, open source software can be used to develop planning tools. These tools show planners how the city operates, as well as provide a platform to gather public input more efficiently through crowdsourcing. Big, open source transit data collected by telecommunications companies and social media can help to analyze traffic flows and other important urban dynamics.

From an IT point of view, the big data infrastructure for ITS needs to support the four aspects of big data: volume, velocity, variety, and veracity. Volume, velocity, and variety can be supported through storage components and veracity will be handled through processing/analytics components.

**Benefits and Challenges**

The benefits to utilizing big data for the improvement of transit planning and operation are numerous. It delivers cost effective prospects to improve decision making in critical development areas like transportation planning and disaster response. Additionally, traditional household travel surveys in developing countries have proven to be expensive and inefficient relative to the possibilities afforded by big data collection and analysis. Such surveys are not conducted regularly by any local public agencies, so finding a more productive and efficient way to analyze the travel behavior of a complex urban network will save time and effort for city planners.

In addition to the benefits afforded by big open source data for transit planning and operation, several challenges plague the process. Common to all big data endeavors, security and privacy concerns and interoperability challenges may limit the scale of projects at this time. Encrypting data, though useful for addressing privacy concerns, will also increase the size of bandwidth needed to support the networks. Completing these projects in the developing world means that we also have to overcome low technological infrastructure as well as economic and human resource scarcity. It is particularly important to note that each city in the developing world has its own individual assets and weaknesses, and that it is impractical to apply a one-size-fits-all solution to a diverse array of transit networks.

**METHODS**

This paper evaluates the usage of big data collection and processing for transit networks in the three urban areas: Nairobi, Istanbul, and Dhaka. Nairobi’s network is informal, unregulated and fluid; data is being collected regarding the routes and stops in order to understand the system and improve it. Istanbul enjoys a fairly functional BRT network, and data collected from the automated fare collection service is helping planners to optimize the network. Dhaka has an overcrowded bus system that lacks a map or clearly marked stops, and researchers are using GPS data to better understand the workings of the system. Analysis of existing systems in these cities revealed that with adequate technical capacity, information and communication technology can greatly improve the sensing and processing of data regarding transit network operations in developing cities. There are many opportunities to utilize this information, including formalizing bus stops to increase the safety of riders, regulating and standardizing fares, and promoting open source data reserves to encourage the development of mobile phone applications to assist existing and new transit ridership. We then generate potential architectures for communication technology, and data storage and processing.
Table 2 shows organizational and data details for the three cities that were the focus of case studies of applications of big data for transit planning and operation. Nairobi, Istanbul and Dhaka are all large, rapidly growing metropolitan regions in developing countries that are making an effort to harness big data to optimize their transit networks. For this paper, we identify approaches to collecting and processing data. We develop architecture for existing and future data flows within each city and we calculate the required bandwidth capacity for the data flows.

<table>
<thead>
<tr>
<th>Transit System</th>
<th>Agency</th>
<th>Sample Data Types Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informal Matatu Network</td>
<td>University of Nairobi</td>
<td>GPS locations from mobile phones</td>
</tr>
<tr>
<td>BRT</td>
<td>Metropolitan Municipality of Istanbul</td>
<td>User fares</td>
</tr>
<tr>
<td>Informal Private Bus Network</td>
<td>Urban Launchpad</td>
<td>GPS locations and rider surveys</td>
</tr>
</tbody>
</table>

Nairobi

It is often argued that Africa’s greatest challenge is its infrastructure. Growth has outpaced capacity of infrastructure, especially with respect to urban mobility. Large African cities like Nairobi are in desperate need of evidence-based policy advising, where big data can play an important role.

A popular response to heavy congestion problems in cities like Nairobi has been significant investment in extensive public transit systems, including both underground and overland metro systems. Substantially fewer resources have been allocated for improving and expanding public bus networks, which are often poorly maintained and quite crowded. In many cases, privately owned minibus operators have met the need for affordable public transit by serving areas where standard bus routes are insufficient for the respective travel demand (22).

Researchers with the University of Nairobi and their collaborators are working to standardize and make transit data available for the network of Matatus, the informal city bus system. Using open source transit software in conjunction with existing digital mapping efforts, an inclusive framework for acquiring and mapping transportation data will emerge (23). The data includes 130 routes and stops and has already spawned the development of two useful apps. These apps can now help to gather other useful data on accidents and congestion via crowdsourcing.

Industry partners are also getting on board with the open source big data movement in Nairobi. IBM has pioneered a solution that enables Nairobi commuters to utilize their mobile devices to advise traveling routes subject to current estimates of traffic congestion.

Using special algorithms to read and transmit images from CCTV cameras located around the metropolitan region of Nairobi, both drivers and public transit passengers can use their mobile devices to stay updated on current road conditions and receive recommendations for alternate routes. Only 36 cameras are currently operational, so IBM researchers have used network analytics to augment the data, enabling the system to forecast otherwise undetected patterns of mobility. The system works on both basic mobile phones with an SMS-based system and on smart phones through an app where users can view a city map that displays several
routing options around existing traffic jams. Researchers at IBM are developing the capabilities to include various other types of data including public safety, road conditions and scheduled construction in order to produce a more nuanced perspective on urban mobility.
FIGURE 1 A Transit management architecture for Nairobi – Produced using Turbo Architecture
Figure 1 shows a transit management architecture for Nairobi developed by the authors based on the data available to them. The attached architecture for Nairobi is based on the Matatu transit system. It does not enjoy a ready framework of big data system like Istanbul has, but some of the required subsystems are already there. IBM has one of its global laboratories there, which collects data from the CCTV cameras installed in the local roads. This IBM center can act as the central data management center. Also the transit map of the Matatu route is already developed. Proper integration between this transit system with the data center (IBM lab) will help to have a complete framework of collecting transit data. The mobile network and online media of Nairobi can fill the gap of information flow regarding incident occurrence or transit vehicle status. Nairobi Traffic Command Center is suggested to act as the public transit management center. Using the data provided by the traffic command center, vehicles will be updated of route condition, fare schedule, and other factors.

Istanbul
Residents of Istanbul, Turkey rely on a centrally operated BRT system. Data collected from automated fare collection systems is being processed daily. Istanbul’s bus lines use a smart card to enables speedier payment and the collection of user data. The card is swiped by users when they get on a bus and at their destination when they collect their refund, if they do not use the entire line. The card can track a rider’s origin station, destination, frequency of rides, and some personal characters of the rider (including their status as a student, adult, or senior). Though this creates a constant incoming stream of transportation data, the information is not yet achieving its highest potential.

There are still issues halting the progress of analyzing this data in order to improve Istanbul’s transportation. The amount of data is overwhelming and, as to date, no efficient way of sorting through and analyzing it has been implemented. Furthermore, the data is not collected for the purpose of using it to analyze transportation, so some additional data of demographic, land use, and spatial nature, for example, will likely be needed in order to create relevant, actionable conclusions.

Istanbul’s Bus Rapid Transit (BRT) system, which travels between the Asian and European parts of the city, also uses the Istanbulkart, or smart payment card. This bus system is environmentally friendly and greatly helps relieve congestion. However, 22% of commuters in Istanbul still drive their own cars to work. Since the BRT network operates in primarily separate lanes, it maintains an appeal in its own right as traffic congestion increases in regular travel lanes. If the bus system was streamlined and made more appealing to the public, traffic congestion could continue to be significantly reduced. Big data could be used to accomplish this. For example, the analysis of big data led to the conclusion that a high volume of passengers was consistently overwhelming the station of Mecidiyekoy. In response, a new station was built, greatly lessening the strain on the Mecidiyekoy station. The greatest roadblock on the way to utilizing the BRT’s data collection system is the fact that data is only collected for those travelers who collect the refund on their Istanbulkarts at their destination station—are mere 36% of travelers.

Another project, called Insights in Motion, utilizes IBM analytics software to study how to make transportation more “convenient and comfortable.” The study is currently being conducted in two cities: Dubuque, Iowa and Istanbul, Turkey. This project looks at “cities in motion” and aims to decrypt the direction and motivation behind the movement of a city—basically, how and why people move around. This new information will be used to develop new infrastructure while retrofitting old infrastructure to its full potential. This project analyzes data
gathered from transportation systems, census records, and data collected through cellular device.

In Istanbul, this project hopes to have big impacts. It aims “to reduce operating expenses by
40%, meet 37% more demand, reduce average commuter time by 60% and reduce per-traveler
combustion emissions by 40%.”

The Istanbul Metropolitan Municipality briefly opened up all their transport data from
November 1-3, 2013. This was to allow for a citywide competition in which teams used this data
to code and design “smart city applications.” It focuses on three domains: smart mobility, smart
tourism and smart participation.

Istanbul could be coming into a new era of efficient transportation if projects like Insights
in Motion continue to be implemented. These projects use big data to create efficiency, allowing
us to know where and how to allocate resources, when certain areas of infrastructure are most
likely to be overwhelmed, and where people travel.

Many transportation analysts and experts have begun to realize that supplying open and
thorough data is the best and possible only way to begin developing smarter, sustainable cities.
Though Turkey still does not have an open data policy, events like the previous mobile
application competition will hopefully encourage the Istanbul Metropolitan Municipality to open
up their transit data and allow individuals and companies to begin analyzing the data in order to
create a more sustainable Istanbul.

Figure 2 shows the proposed transit management architecture for Istanbul developed by
the authors based on the data available to them. Among the selected study areas, Istanbul already
has an active framework of big data management to manage the data collected from
Istanbulkart (traveler card) of the local BRT system. In Istanbul, traffic sensors and cameras are
installed on the highways. Data collected from the sensors and cameras can help the Istanbul
Department of Fire Brigade in case of incidents or fire hazard. In addition to that, massive
quantity of media and wireless network generated data are expected to be collected by the data
center, which will archive every possible data for future planning purposes.
FIGURE 2 Transit management architecture for Istanbul - Produced using Turbo Architecture
Dhaka

As the most densely populated city in Bangladesh and one of the fastest growing cities in the
world, high levels of congestion with heterogeneous traffic frequently plague the city of
Dhaka. Dhaka is poised at a critical moment in its development where it must choose whether to
allow rampant motorization to choke its already congested streets or to determine the best way to
improve public transit to enable a more seamless integration of an already heterogeneous traffic
situation (27). A common policy reaction to such motorization is to increase roadway capacity in
order to decrease congestion. However, it is well known that this strategy tends to encourage
long-term motorization patterns through induced demand. In other words, expanding capacity
attracts new trips by improving driving conditions, which causes the city and its transport
network to become more and more auto-centric. An alternative to rampant infrastructure growth
to accommodate the endlessly growing fleet of private vehicles is to invest in technology that
will improve a more sustainable transit network in the long term (5). With the right architecture,
Dhaka can accomplish this arduous but necessary task. It should be noted that the Asian
Development Bank approved a loan for a Dhaka BRT system in 2012, which has not yet
materialized (28). We analyze the urban transit system as it exists today and plan for a future that
could accommodate a centrally planned BRT corridor.

Researchers are exploring how smart phone technology can improve bus and minibus
public transportation network through GIS tracking. The bus network is independently owned
and operated and largely unregulated, leaving many questions as to the true effectiveness of the
service. This situation is expected to improve in near future due to integrated GPS system with
the proposed bus rapid transit system (BRTS) in the Dhaka city. With the installed GPS
technology, real time data regarding the BRTS arrival or departure time, and incident location,
can be transferred instantly to a transit management center.

Figure 3 shows a future transit management architecture for Dhaka developed by the
authors based on the data available to them. In Dhaka, current transit system consists of transit
vehicles from multiple private and public bus companies on every route. Existing traffic control
center is inactive due to lack of coordination between the transit vehicles, control center and
traveler. Also, road excavation due to drainage system management and underground cable
connection is a very common scene in Dhaka, which often blocks the bus stops and hinder the
normal embarking or alighting of passengers. Also Bangladesh fire service and Civil Defense
acts as an emergency management center in case of any incident occurs like fatal accident or fire
hazard in public buses. In the architecture developed in this study, these subsystems are included
as they are inevitable for smooth transit system flow. As shown in Figure 3, Bangladesh Road
Transport Authority (BRTA) can play the role of the data management center with its current
manpower and infrastructure. As suggested, BRTA will receive information from traffic control
room, City Corporation, fire service and wireless network data. Transit vehicles will be assigned
by the central traffic control center. Traveler from Bangladesh already uses SPASS, an
automated fare collection system. They can get information from a cellular network. Data will
flow from media and cellular network to BRTA. BRTA already maintains the log of transit
vehicle fitness information, thus from the archive, transit vehicle compliance to maintenance and
safety requirements can be checked. Also from the database of vehicle registration number,
detailed information regarding the vehicle can also be derived.
FIGURE 3 Transit management architecture for Dhaka - Produced using Turbo Architecture
FUTURE POTENTIALS

In order for big data to work effectively for transit planning and operation, technical capacity must be developed and maintained in the developing world. An overview for a vision for the ITS information infrastructure is shown in Figure 4.

FIGURE 4 Multiscale/multiresolution Design of the ITS Information Infrastructure and the available options at each storage level

This infrastructure includes a flexible hierarchical distributed system that can incorporate design and configuration tradeoffs at several levels. For example, tradeoffs to be considered are when and where the data will be stored, processed, and cached in different layers of the data infrastructure decision support system to optimize metrics such as response time, cost, and reliability. Each layer of this infrastructure represents a subset of the computing infrastructure that must be designed to support the characteristics of the size and velocity of the data streams.

At the lowest level, streaming process, document store technologies such as MongoDB or CouchDB can handle incoming data stream with high velocity. These database technologies can be implemented directly upon the standard Linux file system on the local hard drives of the computing infrastructures for this level. In a performance evaluation study, MongoDB are found to be able to support write volume of up to 768Mb/s (29). With the document store acting as a buffer, data will be transferred to an extensible record store such as HBase on the next layer, Extraction-Transform-Load (ETL) process, and possibly imported into a MySQL database at the streaming process level to support real-time data analytics. Built on top of the Hadoop Distributed File System (HDFS) (30) and based on Google’s Big Table technology (31), HBase not only can provide massive tabular storage with billions of rows and hundreds of thousands of columns for unstructured data but also can be dynamically extended and are extremely resilient against hardware failures. Both document store and extensible record store can support unstructured data under a variety of different formats (32). After the ETL process, the data then
can be integrated into a common massive-scale data warehouse, from which long term analytic and strategic planning jobs can be run. Once again, a storage mechanism such as HBase would be useful at this layer, as HBase can scale to petabytes of data and can easily be distributed across geographically different locations.

One of the challenges in collecting transportation data, particularly at developing countries, is the lack of data standards. As a result, all the sample technologies described in Figure 4 are chosen such that they can support the ingestion of data in any format/structure, as they do not rely on the traditional restrictions of tabular data in standard RMDBS. However, we still need to integrate these unstructured data into a common format usable for analyst. This challenge, together with the issues of missing and inaccurate data, are often handled through machine learning techniques (33,34). Built on top of HDFS, HBase can take advantage of Mahout, a scalable machine learning and data mining library. For other types of data analytics, HDFS enables a wide variety of programming tools and libraries. One such tool is MapReduce (35), a batch-oriented programming library that can support hundreds of thousands of concurrent computing processes. Another tool is Spark (36), an in-memory analytic tool to support very fast data processing.

There exists previous work that utilizes Hadoop HDFS/MapReduce in different aspects of traffic management such as estimating vehicle carbon emissions (37), providing dynamic route guidance (38), and predicting traffic congestion (39). With the previously described capability of the Hadoop-based storage infrastructure and the related supporting components, this provides a theoretical validation to the feasibility and potential of our proposed framework.

Given the proper framework for collecting, processing, and storing large quantities of data relevant to the planning and operation of transit networks, developing cities will be better equipped for sustainable and cost-effective decision making in the future. Planner will be able to use more realistic models to determine a course of action.

The last twenty years have seen the birth of Intelligent Transportation Systems (ITS), which has typically focused on applying communication and information technology to more traditional transportation functions like toll collection and incident detection. The increasing diversity and prevalence of mobile devices may change ITS at its core as they introduce new information sources into transportation planning and operations, including mobile sensor networks and real-time information dissemination. Mobile phones are a medium for both creating and receiving valuable urban information; this technology is poised to revolutionize the way we exchange and process data in complex metropolitan areas in the developing world. Real-time connectivity for enabling data exchange between vehicles and infrastructure(40), which is supposed to be a major transportation innovations in the future, will provide another opportunity to improve transit services with buses communicating with infrastructure that includes traffic signals for receiving priority and for enhancing other transit related services.

CONCLUSIONS

In order to be equipped to make smart decisions about the future of critical formal and informal urban transit networks in developing countries, we need access the data to contextualize the problems and to model future scenarios based on real information. Access to such data, including social media aggregations and GIS data, can be difficult in areas with low technical capacity and processing power. It is our argument based on this study, however, that it is worth overcoming such technical obstacles in order to leapfrog beyond a traditional transit-planning model to a more accurate, cost-effective, and sustainable process involving big data for transit planning and
operation. Furthermore, traffic management policies can be derived from big data in order to assist with future improvements of the physical transit infrastructure.

This study revealed that integration of online media with location-based data provide a big data scenario, which have potentials for supporting transit operations while posing challenges to manage the data mobility. It is not feasible to apply a one-size-fits-all approach to any problem in the developing world, but together, these cases show that with the right approach, technical capacity in transitional cities can grow to support higher-level data processing to make more efficient and more sustainable policy decisions for crucial urban transit networks. For cities at a crossroads between committing to an auto-centric future and improving alternative modes of transportation in our energy insecure future, it is important to consider the short-and long-term impacts of smarter transit planning and operation fueled by more reliable data sources.
ACKNOWLEDGEMENT

This material is based upon work supported by the Department of Education GAANN Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Education.

REFERENCES


11. LocationManager [Internet]. Android Developer; Available from: http://developer.android.com/reference/android/location/LocationManager.html


