GPS Enabled Taxi Probe's Big Data Processing for Traffic Evaluation of Bangkok Using Apache Hadoop Distributed System

Paper Identification number: **AYRF14-060**

Saurav RANJIT[1], Masahiko NAGAI[1, 2], Itti RITTAPORN[3], Fredrik HILDING[3]

[1]Department of Remote Sensing and Geographic Information Systems,
Asian Institute of Technology, Pathum Thani, Thailand
Email: saurav.ranjit@ait.ac.th, nagaim@ait.ac.th

[2]Center for Spatial Information Science,
The University of Tokyo, Japan
E-mail: NAGAIm@iis.u-tokyo.ac.jp

[3]Toyota Tsusho Electronics (Thailand) Co., Ltd
Email: itti@ttet.co.th

Abstract

Probe Taxi have been operated in the Bangkok since the July of 2012 by Toyota Tsusho Electronics (Thailand) Co. Ltd. Approximately 10,000 probe taxi are utilized for the real time traffic information monitoring and it provide the meaningful information of the traffic condition in the region such as travel flow information, best routes etc. GPS devices have been installed in the probe taxies to collect spatial and temporal information every 3 to 5 seconds along with other information. It provides the real time traffic information by calculating the spatial and temporal information of these probe taxies. The spatial information includes the latitude and longitude location of the taxies; on the other hand the temporal information includes the UNIX epoch time. At the same time, the other information such as device ID, speed, direction, taximeter etc are collected from probe taxis. The device ID is the International Mobile Station Equipment Identity also known as IMEI that has unique ID. The main challenge of this study is to handle big data. Approximately 50 millions of data is being collected every day with the file size of 3.5 giga byte. To process this big data, it takes lots of time and resources. Also, extract relevant information from this big data is another challenge along with the filtering out of irrelevant and error data. The objective of this study is to find the suitable method to process the big data and produce the relevant information. Apache Hadoop Distributed System is used to process the big data and Java based programming to perform the operation. The Apache Hadoop software library is a framework for distributed computing of large data across clusters of computers using Programming models. It is designed to scale up from one machine to hundreds of machines, each offering local computation and storage.

Keyword: Big Data, Hadoop Distributed System, Map Reduce, Latitude, Longitude

1. Introduction

Toyota Tsusho Electronics (Thailand) Co., Ltd (TTET) is a one of the member of Toyota Tsusho Corporation belonging to the Toyota Group: one of the largest trading companies in Japan. Located in Bangkok, the company, TTET provides the real traffic information of the Bangkok and some other provinces in Thailand.

Toyota Tsusho Electronics Thailand has installed about 10,000 GPS taxi probe devices in the Taxis and about 200 GPS devices in the Trucks that are operated in the Bangkok in the July of 2012. These GPS devices installed in the Taxies of the Bangkok provided spatial and temporal information every 3 to 5 second along with other necessary information. The real time traffic information is provided by the spatial and temporal information of theses taxis in Bangkok.

The spatial information included the taxis latitude and longitude information and the temporal information includes the epoch time. Along with these the other information includes the device id, speed, direction, taxi meter change, taxi engine state and dilution of precision.The device id is the

International Mobile Station Equipment Identity also known as IMEI. These device ids are unique to each other and ideally no two devices can have same IMEI number. Latitude and Longitude is in the WGS84 datum format.

Big data and Data mining

The amount of data is increasing every day whether the data is structured data from any sensor etc or unstructured data from any mail being sent. Every day new data sets are being piling up in the data warehouse for processing. Big data refers to the collection of data in a large scale such that it becomes so complex and difficult to the handle using traditional data management tool. Hence the big data usually will have the data set with size that is beyond the ability of traditional tool to store, process, analyzed.

The main challenge of the "big data" is the storage of the data, analysis of the data, retrieval of the relevant information and sharing information that has been obtained. As mentioned earlier the big data means the size of the data is huge that is difficult to handle with traditional data management tool. So the first problem associated with the big data would be the storage facilities of such data. In such case a cloud data ware house is needed for the storage of big data. Secondly, since the size of the data is big processing time for such big data would be large as well. Larger the data size, more time it would take to process it. For the real time processing and also processing the historical data it would be difficult to process the big amount of data. To take the measure a distributed system would be need to process and get the relevant information such that information could be retrieved in real time or near real time situation. Only after such process, information from big data can be decimated for meaning full uses.

Global Positioning System

Global positioning system (GPS) is a space based satellite navigation system which provided positioning and the timing information anywhere in the world. The Global positioning system is governed and maintained by the United State government and is publicly accessible to use it. This system is basically used for military purpose, civilian and commercial purpose. The total of 24 navigation satellite is used for the GPS which are place 6 orbits at the height of about 17700 kilometers from the earth. GPS structure includes three segments i) Space segment ii) Control segment iii) User segment (receiver).

As mentioned earlier GPS can provided the positioning and timing information. Hence the GPS data is the position in term of latitude and longitude and the timing of that. Along with these, GPS can also provide information like speed and direction in case the GPS receiver is placed on the moving platform. The moving platform can be anything from vehicle like cars, buses to UAV or any aircraft. The accuracy of the positioning depends upon the type of the GPS receiver and location where the positioning information has been taken. Highly sophisticated GPS receiver is capable is of giving high accuracy within the meter range whereas the low cost receiver can give the accuracy from 10 meter to 30 meter in range. Location also plays the key factor in the accuracy of the GPS data. In open field area where there is not much interference of the outside world, GPS data is more accurate and precise as compared to the urban area where there are high rise building blocking the GPS receiver from getting the proper signal.

Probe Car Data

Probe car data is the data from the device that is been driven. The data could be position, speed, direction, and timing information from the device that is being driven. These probe car data are essential source of information for the intelligence transportation system as it can provided real time information. The log of these data can be used to evaluate the historical data and get the information from historical data.

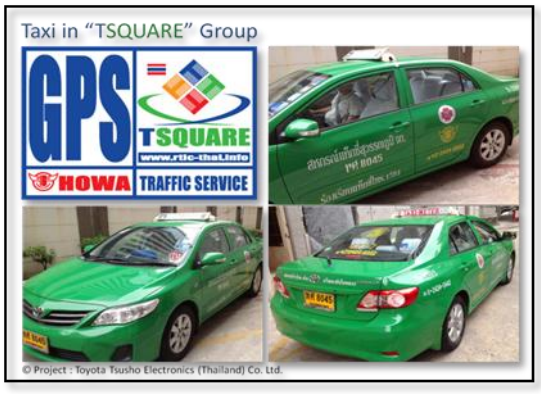


Figure 24 : TSquare Project (Toyota Tsusho Electronic (Thailand) Co., Ltd)

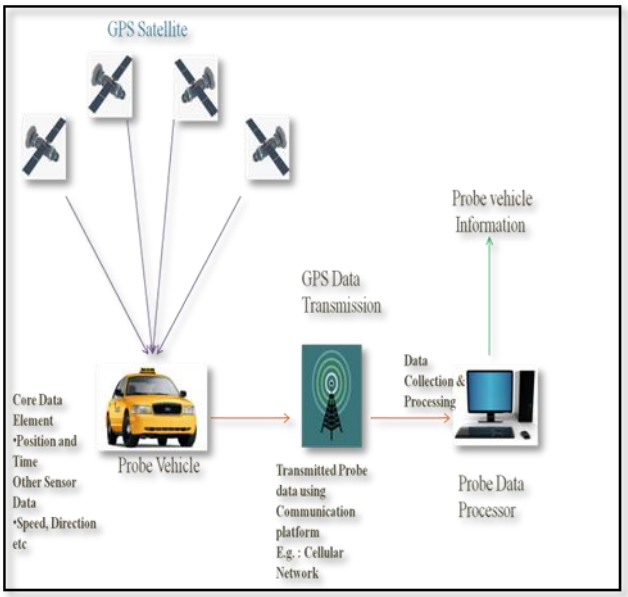


Figure 25 : Prove Vehicle Working

Figure 1 shows project from Toyota Tsusho Electronic (Thailand) Co., Ltd. The name of the project is TSquare which utilizes the probe GPS data to provide traffic information in near real time. Figure 2 show the working of these probe devices.

2. Literature Review

Barry Storey, Robert Holtom (2003) used the GPS device for the traffic monitoring. The main reason of this GPS device installation was for security system, fleet racking, and satellite navigation. Location, speed, direction etc were being reported by the tracking devices are regular interval of time which the ignition of the vehicle is switched on. The system was primarily designed for real time application but the use if the historical data recorded from these tracking devices has been growing and is being used for the transport and traffic monitoring uses. The number of vehicle used stand around 70,000. Collection information from 70,000 GPS device would generate huge spatial database in the course of time. A system for handling such big data is required as conventional system is not efficient and fast enough to handle big data. Travel time, average speed etc can be determined from the probe vehicle. The probe vehicle is mostly used for the study of travel time. Though these techniques seem very simple, their implementations tend to be quite labor intensive (Taylor, 2000; Wohlson & Haptipkarasulu, 2000). The data collected from GPS will be huge and number of records collected within the short interval of time is very high. GPS device can record 100 gps point in every one second to 1 gps point in one second depending upon the type of gps used. Hence reference, storing and retrieving the GPS data have become very much essential (Joseph OWUSU, Francis AFUKAAR and B. E. K. PRAH(2006).The source of data generated not only by the users and applications but also "machine-generated," and such data is exponentially leading the change in the Big Data space. Dealing with big datasets in the order of terabytes or even peta bytes is a challenging. In Cloud computing environment a popular data processing engine for big data is Hadoop-MapReduce due to ease-of-use, scalability, and failover properties. (Padhy, 2013).

3. Methodology

Data Source

The GPS probe data is being obtained from 10,000 taxi's from Bangkok region that has been installed by Toyota Tsusho Electronics Thailand since the July of 2012. The data is collected 3 to 5 seconds from each device. An average of 50 million records is being collected each day with the average file size of 3.5 giga byte.

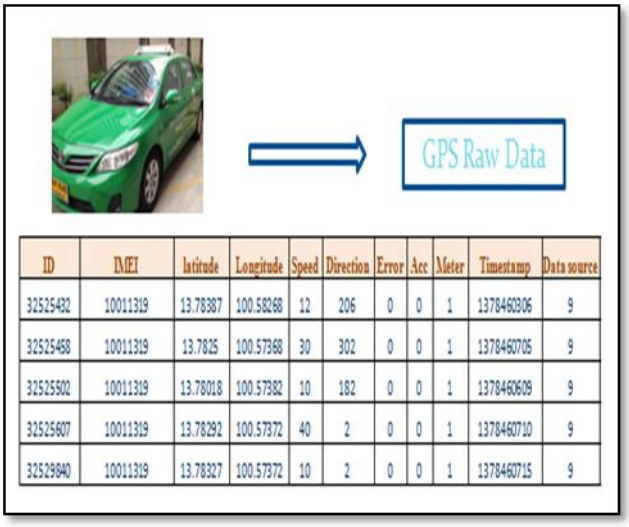
The data parameter are obtained from the taxi probe are as follows.

- IMEI Number: International Mobile Station Equipment Identification (IMEI). The IMEI is the unique identification number given to each mobile device. The IMEI is only used to identify the device.
- Latitude / Longitude : The latitude and longitude of the taxi position.
- Speed / Direction : Speed and direction of the moving taxi
- Error : Error is calculated from geometric dilution of precision (GDOP). GDOP shows the affect of geometry of the satellite on the positioning accuracy. The ideal value of GDOP is 1 which gives the correct positioning. For our case the higher value of GDOP are rejected and not recorded.
- Acceleration : This parameter shows the status of the taxi engine On/Off condition. The value of acceleration is 0 or 1 i.e. on/Off.
- Meter : Meter shows the status of the taxi. The value of meter is 0 or 1. The value 0

shows the taxi has not been used and value 1 show the taxi is being used by the customer.

- Time stamp : Time stamp recorded is the UNIX epoch time stamp. UNIX time is the time system which is described as number of seconds elapsed since 00:00:00 coordinated universal time, 1 January 1970.
- Data source : Data source defines the type of device that is sending the spatial and temporal information.

The total of 11 parameters is recorded from 3 to 5 seconds from 10k taxi probe and stored. Data are collected for each day starting from 00:01 am to 12:00 pm. Figure 3 shows the sample GPS data that has been collected.



ID	IMEI	latitude	Longitude	Speed	Direction	Error	Acc	Meter	Timestamp	Data source
32525432	10011319	13.78387	100.58268	12	206	0	0	1	1378460306	9
32525458	10011319	13.7825	100.57968	30	302	0	0	1	1378460705	9
32525502	10011319	13.78018	100.57382	10	182	0	0	1	1378460609	9
32525607	10011319	13.78292	100.57372	40	2	0	0	1	1378460710	9
32529840	10011319	13.78327	100.57372	10	2	0	0	1	1378460715	9

Figure 26 : GPS Probe Vehicle Sample Data Type

Apache Hadoop

The Apache Hadoop is the open source software for distributed computing. Hadoop is a software library that allows the processing if huge data set across clusters of computers using simple programming.

Hadoop Configuration in Distributed Mode
The Hadoop configuration requires minimum of five parameters to be configured. The parameters are as follows.

Hadoop Environment
Hadoop core
Hadoop Distributed File System
Hadoop Map Reduce

Master / Slave configuration

All these parameter need to be configured in each of the cluster depending upon the type of cluster i.e. master or slave.

To process the big gps data, a UNIX based operating system in 5cluster of machine is used. The number of cluster / machine can vary depending upon the data size. The pre-requisite tool includes Java Development Kit (JDK 1.6 or higher), Apache Hadoop version r1.2.1, Secure Shell and Eclipse Integrated Development Environment. The probe taxi data, which is in CSV file format, needs to be stored in the Hadoop Distributed File System (HDFS) for it to be processed and analyzed. HDFS is the primary storage location of Hadoop Distributed System.

The HDFS has two types of node being operated. First is the Name Node which is known as master and the other is the Data Nodes also known as slave or worker. The name node manages the file system of HDFS and regulates access to files within. The data node on the other hand manages the data that are run on them. Each data node is assigned a set of job by the name node to perform. Configuration of Hadoop on machine depends upon the type of machine that is being used i.e. weather the machine is name node or data node. There is only one name node and several data node depending upon the requirement. The software frame work of Hadoop is the Map Reduce.

Map Reduce task is implemented using Java Programming language. Map Reduce framework works on <key, value> pair. Map Reduce views the input as set of <key, value> pair and generates the output as another set of <key, value> pair. In the Map task of Map Reduce, IMEI ID and Unix epoch time is taken as the key while latitude, longitude and taxi meter serves as the value. In the Reduce task of Map Reduce, IMEI is taken as the key while other relevant information is servers as value. The key value pair during both Map and Reduce operation can be changed depending on the requirement and information that needs to be produces from the taxi probe data. Both Map task and Reduce task involves removing of error taxi probe data as well. Filtering out the relevant data from the big data set in the primary task before analysis on the data can be done. Filtering out of data is done both in mapping process and also in reducing process.

Following the filtering of the error data we can extract the relevant information from the big probe data set. Figure 4 shows the flow chart of the overall processing the GPS big data.

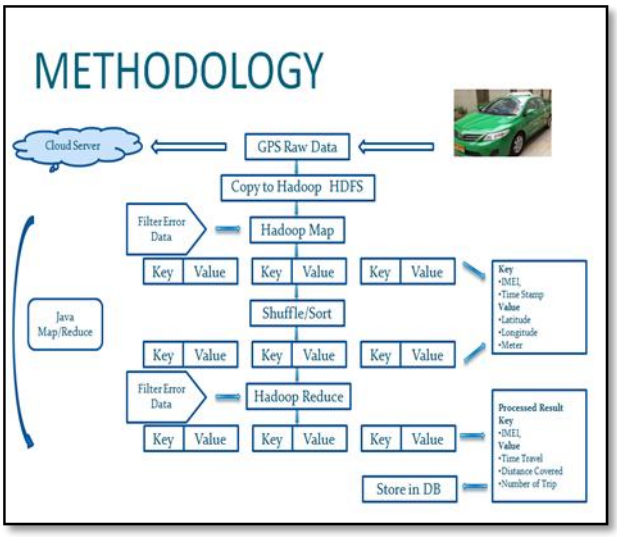


Figure 27 : Data Processing Flow Chart

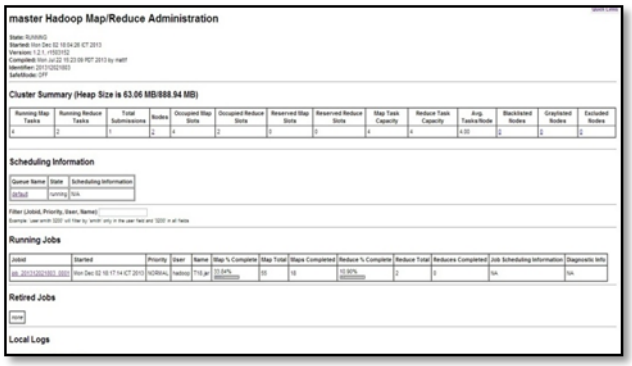


Figure 28 : Apache Hadoop Process Monitor

4. Result

The result obtained from the processing of the big data is shown below. As mentioned earlier the major task before any analysis can be done is the filtering out of noisy data from the big data set. Figure shows the type of error data removed from the big data set. Figure 5 shows the process monitoring of the Hadoop during the map reduce operation. During process both master and slaves cluster can be monitored and check the performance from each cluster. The Table 1, 2 and 3 shows the type of systematic and random error in the data. The error is caused by the device or during transmission of data etc. These errors are device giving same latitude and longitude for all the timestamp, meter status showing 0 all the time or meter status fluctuating. Other error is positioning error in which latitude and longitude are at different location other than Thailand. Figure 6 shows the error in positioning. Figure 7 on the other hand shows the visualization for the GPS probe data around the Bangkok region after removing the noise from the data raw data source.

Removing the error from the big data set allows us to analysis the good data set and extracts relevant information out of it. Table 4 and Table 5 shows the information that can be extracted out from the probe vehicle such as travel time, distance travelled, maximum speed, minimum speed and many more.

Table 7 : Error Latitude and Longitude

ID	IMEI	latitude	Longitude
187032	358688000000159	13.72309	100.56099
187034	358688000000159	13.72309	100.56099
187035	358688000000159	13.72309	100.56099
187036	358688000000159	13.72309	100.56099
187037	358688000000159	13.72309	100.56099

Table 8 : Error Meter Status

Error	Acc	Meter	Timestamp
0	0	0	1378448562
0	0	0	1378448663
0	0	0	1378448668
0	0	0	1378448673
0	0	0	1378448678

Table 9 : Error Meter Status

ID	IMEI	Meter
7186907	10016328	0
7194065	10016328	1
7194720	10016328	0
7197021	10016328	1
7200126	10016328	1

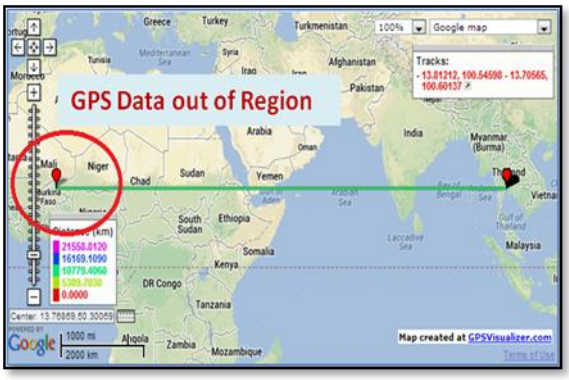


Figure 29 : Error Latitude and Longitude

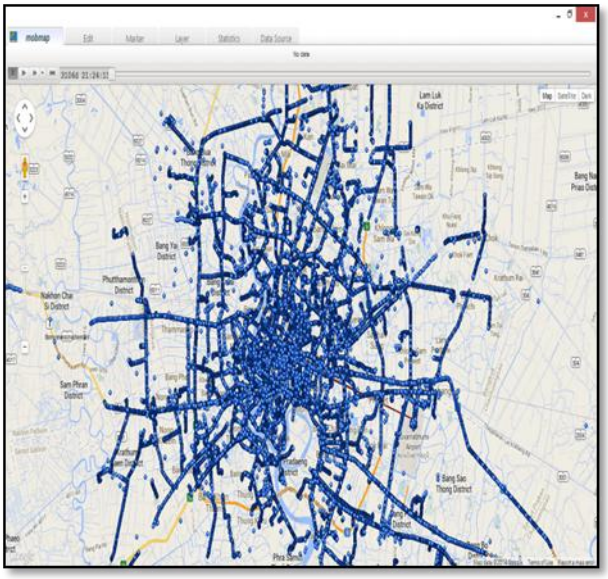


Figure 30 : Visualization of the GPS probe data

Information Extracted from the Probe data processing.

Table 10 : Information Extracted from GPS Probe Vehicle data

IMEI	Number of Records	Distance Travel	Time Travel
10000023	555	297486.4771	572.4166667
10008908	867	450682.1861	1430.683333
10011265	2527	156671.9971	771.75
10011317	2387	241907.382	1405.216667
10011328	6834	557990.9337	1418.9

Table 11 : Information Extracted from GPS Probe Vehicle data

Maximum Speed	Minimum Speed	Center of Gravity	
144	2	13.6564	100.2093
128	2	13.5934	99.22405
108	2	13.3801	97.50961
122	2	13.26782	98.18546
148	2	13.45985	98.44365

Conclusion/Discussion

With the use of Apache Hadoop Distribution System, the processing time of the big data has been reduced sharply. For the data set of 50 millions GPS records, the processing time is reduced to about 10 minutes with 2 clusters which would have taken long time if processing is done conventionally. This is processing and extracting the information from the data of 3.5 giga byte is only about 10 minutes. Filtering of error data was carried out using the programming logic during the Map Reduce operation and helps reduce the error for further processing.

Reference

[1] Padhy, R. P. (2013). Big Data Processing with Hadoop-MapReduce in Cloud Systems (pp. 16~27), Bangalore, Karnataka, India. International Journal of Cloud Computing and Services Science (IJ-CLOSER).

[2] Owusu, J., Afukaar, F., Prah, B.E.K (2006). Towards Improving Road Traffic Data Collection: The Use of GPS/GIS, Accra, Ghana, March 8-11, , Promoting Land Administration and Good Governance 5th FIG Regional Conference

[3] Storey, Barry., & and Holtom, Robert (2003). The use of historic GPS data in transport andtraffic monitoring, TEC.11.03/p376-379 IT IS

[4] Tong , D., Merry, C.J., Coifman, Benjamin (2005). Traffic Information Deriving Using GPS Probe Vehicle Data Integrated with GIS , Center for Urban and Regional Analysis and Department of Geography The Ohio State University 1036 Derby Hall 154 North Oval Mall Columbus, Ohio 43210, USA

[5] B.L.Malleswari.,I.V.MuraliKrishna., K.Lalkishore.,M.Seetha.,Nagaratna, P. Hegde (2009). THE ROLE OF KALMAN FILTER IN THE MODELLING OF GPS ERRORS

[6] Samet, Hanan (2009) . Sorting Spatial Data By Spatial Occupancy, Institute for Advanced Computer Studies Computer Science Department, University of Maryland eoSpatial Visual Analytics: Geographical ormation Processing and Visual Analytics for Environmental Security, pp. 31?43. Springer Business Science Media, Berlin, 2009.

[7] Brakatsoulas, Sotiris., Pfoser, Dieter., Salas, Randall., Wenk, Carola (2005), On Map-Matching Vehicle Tracking Data , Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005

[8] Aslam, Javed., Lim Sejoon., Pan Xinghao (2012). City-Scale Traffic Estimation from a Roving Sensor Network, SenSys'12, November 6–9, 2012, Toronto, ON, Canada

[9] Francis, Deja Hepziba., Madria, Sanjay., Sabharwal, Chaman (2007) . A scalable constraint-based Q-hash indexing for moving objects, Information Sciences 178 (2008) 1442–1460, Rolla, MO 65401, United States.