International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

Big Data Application: Selection of Sites of INRIX Probe-data Stream for Real-time Performance Monitoring

Sandeep Singh Rawat, Anuj Sharma

Abstract: ***Every year, US transport agencies assess number of people killed and injured on U.S. roads, which witness the enormous increase during the past half-century. With the rapid growth of society, the transportation department is also facing many challenges every day. Challenges include data evaluation, summarize, distributions, capacity, allocation, conception, investigating, modified and security. Nowadays, the Nebraska Department of Roads (NDOR), in collaboration together with the other transport agencies, collects and explores the traffic data at sixty one (61) continuous traffic count sites. The objective of this work is to select the locations based on different criteria of the probe-based data stream compared to fixed sensor data. Based on a significant estimation of the INRIX probe-based data, the study will feature important factors for integrating the probe vehicle stream data into transport related tasks, preparation and supervision actions. This research addresses how big data platform can resolve the questions of traditional traffic management system to identify the traffic anomalies contributory factors, namely miles travelled and confidence score. In future, after collecting real-time TMCs data based on our selected locations, we will assess the reliability and correctness of INRIX probe-based data stream.***

Index Terms: Big data, Inrix, Performance, Probe data

I. INTRODUCTION

This growing interest in big data has perceived around many disciplines globally. Many modern data-intensive application areas which collect and support huge volumes of raw data and need high processing have provided greatly to the development of Big Data Analytics [1]. Today's technology-oriented organizations are collecting and maintaining statistics, which is evaluated in zettabyte magnitudes or even bigger. Also community networking companies have more than active billions of consumers and frequently produce an enormous data. In YouTube alone 78000000000 minutes of data are transferred. According to the National Security Agency, the people those who uses the Internet, transmits nearly 1,622 Pebibytes of data per day [2]. The online digital records has increased many folds in size in just few years [3] and by next year, it's expected to grow up to 35 billion terabytes [4].

Manuscript published on 30 June 2019.
***** Correspondence Author (s)
Sandeep Singh Rawat, Department of Computer Science & Engineering, Anurag Group of Institutions, Hyderabad, India.
Anuj Sharma, Department of Civil, Construction, and Environmental Engineering, Iowa State University; Ames, United States of America.

This outbreak of online figures generates great prospects and transformative possibility for many areas such as retail, hospital, manufacture engineering, public sector administration, educational sectors, and transportation [5]. Almost all technology-driven companies have participated in expanding products using big data tools for supervising, data exploration, models, and their business requirements, building it a dominant area in data science study. Moreover, discovering and extracting a significant pattern from huge data sources for decision making is one of the primary tasks of big data visualization. In addition to mining, assessing massive amounts of data, big data have key challenges for batch analytics, stream analytics and deep learning, including format variation of input data, noisy data, scalability of algorithms, sufficient data storage, fast retrieval of meaningful pattern, etc. Therefore, latest data exploration tools and data administration solutions are necessary when operating with vast data research. For instance, in a current task at the Center for Transportation Research and Education (CTRE) lab at Iowa State University, we observed the high-dimensionality of transportation data, taken from INRIX and researched different practices to deliver the transportation related challenges such as travel time reliability, congestion hour, speed performance, etc. INRIX offers many details like automobile speed, their travel interval, crash data and other facts, which revises and updates these travel section, known as TMCs, at a rate of once every 20 seconds. The final stream data is approximately 9-10 gigabytes per month, or beyond 100 gigabytes annually for the complete Nebraska routes. With the coverage of latest advanced geographical region, the volume of recorded data is projected to grow [6]. This research paper is prepared as follows: Section 2 shows the background and related work in the concerned field and discusses the challenges and opportunities with big data, Section 3 discussed the different criteria for experimental setup, final selected locations and presents the discussion, and Section 5 concludes the paper with future direction.

II. BACKGROUND & RELATED WORK

Today, digital records in all forms and dimensions, is increasing at astonishing rates. It also indicates to a change in thinking in our scientific investigation towards pattern discovery. Technologies for evaluating large-scale information are growing quickly and there is major interest in batch analytic and stream analytic approaches such as Hadoop.

MapReduce and Hive, and MapReduce extensions to traditional databases. MapReduce is a core of Apache Hadoop framework; its deployment allows immense scalability across thousands of machines in a cluster and receives considerations for many data processing solutions. MapReduce was initially created to process large-scale batch tasks in a shared nothing environment. The MapReduce has recently drawn plenty of interest for such application that works on large-scale data. MapReduce is a programming structure for handling and analyzing huge volume of datasets related to a variety of real-world problem. The MapReduce concept procures the characteristics of parallel programming that provides simplicity and provides load balancing and fault-tolerance. The key features for MapReduce are the scalability, competence and fault-tolerance as compared to real-time capability, which is the main feature of the traditional managing systems such as DBMSs. Several MapReduce programming frameworks, including Hadoop Streaming [7], Hadoop Pipes [7], PLANET [8] etc. to processed different types of huge information that need unconventional computational power have been developed. The researcher has developed many technologies including Hadoop Online [9], Hadoop indexes (Hadoop++) [10] and column-oriented storage (RCFile) [11], Hybrid method (PCF and DCF) [12] etc. to improve the batch processing jobs using MapReduce. These modern technologies reduce the difference between batch and real-time data handling. However, the choice of choosing the best data handling techniques depends on varieties and source of data, processing time required and then ability to take immediate actions. For instance, Abadi J. Daniel et al. [13] have compared MapReduce paradigm and parallel DBMSs, and the finding display that parallel DBMSs significantly earlier than MapReduce. The Google File System (GFS) [14] that typically uses a MapReduce system and provides an efficient and consistent distributed data storage which is essential for all applications that run on inexpensive commodity hardware. Google's MapReduce technique helps to build distributed applications, which are distributed across multiple nodes in a simpler way with commodity hardware. Coifman, B., and Kim [15] evaluated speeds from INRIX with the parallel loop detector data. Three concerns became noticeable: 1) INRIX speeds incline to delay the loop detector measurements. 2) Most of the occasion he reported speed is similar to the earlier trial, and 3) even though INRIX provides two different types of measures, these measures do not appear to specify the latency or the occurrence of recurring INRIX stated speeds. Researchers have reviewed how INRIX started the original crowd-sourced transportation system with sensor devices in today's automobiles, which offered better description of traffic reports on any location for preparation, exploration and managements of automobile networks. Different investigators described different functionalities to improve transportation handling development which used probe data for CO2 emission reduction. They evidenced to reveal a bottleneck exclusive of the deployment level of the in-vehicle section by manipulating probe-based stream statistics in transport handling exploration. Other researcher suggested a technique for handling probe-based vehicle stream data to assessment of the traffic flow variables like size of the traffic, space mean and degree of traffic level. They used a different technique standard besides using the Kalman filtering technique (KFT) with many sets of assumed traffic data. They recommended the different opportunity to evaluate travel time. Several methods to integrate probe-based stream data into the different transport handling model were utilized to explain the optimization challenge in urban regions and they have validated that to decrease probe-based stream data for huge traffic size with insignificant degradation on the value of transport condition assessment. For moderating carbon dioxide emanations using latest modern transport handling system, we needs several sensors and great equipment and installation charges. Authors have used the probe-based stream data, which are stored by modern tools to analyze successive geographical traffic data, finally shown the possibility of reducing the number of detectors.

A. *Challenges and Opportunities with Big Data*

Leading researchers both from leading universities and companies across the United States have created white paper [17] on challenges and expectations with Big Data. We draw some conclusions from this paper.

- Big Data promises to be one of the most important revolutionize research.
- Their research has suggested that every students' performance evaluation can be generated and then this information may be applied to plan the most valuable methods to students' education.

Exploration of the massive volume of data involves many stages as shown in Figure 1 and every stage encounter the different problems, which are reported in detail in the white paper. The main challenges are as follows:

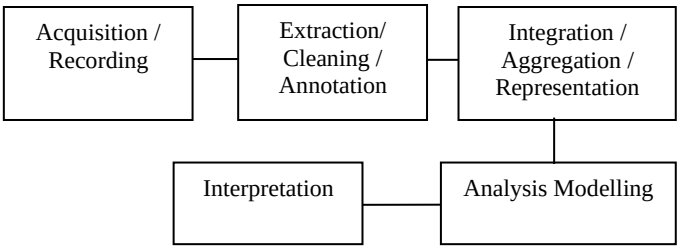

Figure 1. The Big Data Analysis pipeline

- Data acquisition and Recording stage: the major task is to choose data filters, which will discover the valuable data. Another task is to automatically produce the correct metadata to define what data is documented and assessed.
- Information Extraction and Cleaning stage: the major task is to translate the original data in a structures form, which is suitable for analysis.
- Data Integration, Aggregation, and Representation stage: The data are often noisy, diverse and inter-related. The major task is to collect, integrate, and aggregate from all the untrustworthy locations.
- Query Processing, Data Modelling, and Analysis stage: Techniques for querying and mining big data are fundamentally different from the traditional statistical analysis of small datasets.
- Interpretation stage: The main task is to interpret the results obtained from big data analysis. Usually, it involves observing all the hypotheses made and retracting the examination.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

III. EXPERIMENTAL RESULTS & DISCUSSIONS

Present day sophisticated worldwide market, transportation networks support a organization to find in every territory extending the finest promising arrangement of labor, land, tax, and cost — while striving globally. Nowadays most of the transport agencies are trusting mainly on static sensors to gather traffic related reports such as travel time, incidents, traffic speed, etc, to identify or detect which routes are used most, and to either upgrade that road or offer a substitute if there is an extreme volume of traffic. Currently NDOR is providing 61 automatic traffic recorder (ATRs) in several sites. But the total budget of installing and supporting these mounted devices are extremely high in comparison to alternative offered by non-traditional devices. Probe-based stream data is consists of comparatively cheaper practices for collecting traffic related information such as time take to travel from one place to another place and vehicles' speed on Nebraska state. NDOR has now obtained probe-based stream data for augmenting traffic data compilation and evaluating the performance of its management through a third-party vendor INRIX. Presently, Inrix provides mechanisms to store the transport related report on main interstates, non-interstates and urbanized regions. We need to find the location, to assess the consistency and correctness of data stream, which are collected using probe vehicle. Finally, we have selected 16 different locations based on the following criteria:

- Criteria 1 - Based on nearest TMC from ATR mid-point
- Criteria 2 - Based on continuous traffic count data and traffic characteristics on Nebraska Streets and Highways (April 2015) & Automatic Traffic Recorder Data (June 2016).
- Criteria 3 - Winter sections (provided by NDOR)
- Criteria 4 - Level of confidence present in specific areas
- Criteria 5 - Finding the anomaly from CDF distributions

For better judgement we have also studied proportion for heavy truck and Inter quantile range (IQR) for every related TMCs.

A. *Criteria 1 – Based on nearest TMC from ATR mid-point*

Mid points of every ATRs have been calculated to find the nearest distance from each TMCs using R programming.

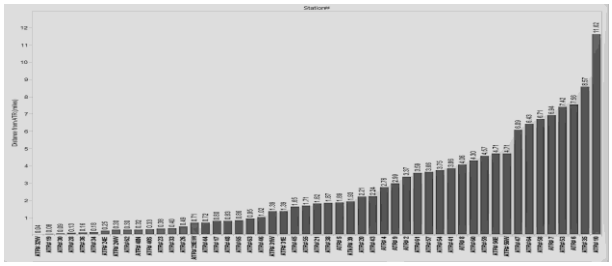

Figure 2. Minimum distance from ATRs station

```
library(argosfilter)
# Reading input file, which consists of latitude and
longitude for ATR and TMCs.
input <- read.csv("TMC.csv", stringsAsFactors = FALSE)
names(input)
# Calculate the distance from ATR to every TMCs
dist <- sapply(2:nrow(p), function(x)
distance(input$start_latitude[x],input$start_latitude[1],inpu
t$start_longitude[x],input$start_longitude[1]) );
#  Map the distance with TMCs
result<- print(paste(input$tmc,dist))
# Output file
write.csv(result,"atr.csv")
# Find the minimum distance from ATR
lapply(dist,min)
min(unlist(dist))
```

Figure 3. Typical code of distance calculation

Figure 2 shows the minimum distance from ATR station and Figure 3 presents the typical code for calculating the distance from ATR to every TMCs.

B. *Criteria 2 – Automatic Traffic Recorder Data*

The Nebraska Department of Roads, in collaboration with the Federal Highway Administration, stored and examines traffic data at 61 locations. Out of these 61 sites, 38 are on rural areas and highways, 8 on low size rural routes, 2 on city roads in Grand Island, 3 on city lanes in Lincoln, 7 on city roads in Omaha, 1 on Interstate in South Sioux town, 1 on a city road in Scottsbluff, and 1 on a city road in Holdrege. A traffic counter senses each vehicle and reports hourly totals automatically. Figure 4 shows the annual average daily traffic for each ATR station.

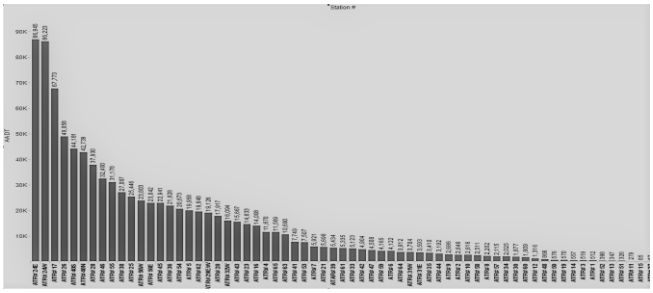

Figure 4. AADT for each ATRs station

C. *Criteria 3 - Winter segments*

We received the list of sites for winter segments from Nebraska Department of Roads (NDOR), refer Table 1. The list comprises the information like start town, end town, start highway, end highway, etc. This information was used to identify the county, route and nearby TMCs and XDs.

Table 1. Winter segments details given by NDOR

District	Start Town	Start RP	Start Hwy	End RP	End Hwy
5	Kimball	20	N-71S	38	L-17B
6	Ogallala	126	US-26/N-61	145	L-51C
4	Elm Creek	257	US-183	279	N-10
1	Utica	366	L-80F	388	N-103
2	Platte River	427	n/a	440	N-50/144th St

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

D. *Criteria 4 - Level of confidence available in particular areas*

Presently INRIX is maintaining 3 different confidence levels (10, 20 and 30) for storing traffic data. Confidence level 10 indicates that it is purely based on the historical data, confidence level 20 indicate that is both real-time data and past data and confidence level 30 indicates that it is purely based on the real-time data. For our evaluation, we have considered only confidence level 30, i.e. real time data.

Figure 5 shows the Nebraska map with average percentage count of traffic data with confidence level 10, 20 and 30 and real time data with confidence level 30 and Figure 6 code was used to calculate the confidence level 30 of real time for each TMCs.

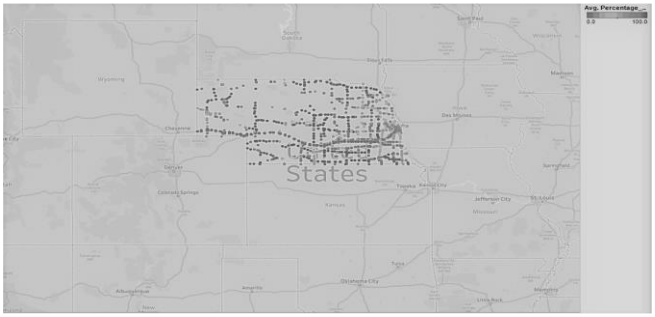

Figure 5. Nebraska map with Average percentage count of traffic data confidence level (10 and 20) and real time data confidence level (30)

```
data        = LOAD 'NebraskaData/2016/{10}' using
PigStorage(',') As (xdseg:int,
c_value:int,segmentclosed:chararray,score:double,speed:
int,avg_speed:int, re_speed:int,
traveltime:double,time:chararray);
reqcol      = FOREACH data GENERATE xdseg,
SUBSTRING(time,0,10) AS (date:chararray),
SUBSTRING(time,11,13) AS (hour:chararray), score;
groupcol    = GROUP reqcol by (xdseg,date,hour);
countcol    = FOREACH groupcol GENERATE group,
COUNT(reqcol.score) AS (counterall:int);
data30      = FILTER reqcol by score == 30;
group30     = GROUP data30 by (xdseg,date,hour);
count30     = FOREACH group30 GENERATE group,
COUNT(data30.score) AS (counter30:int);
joining     =  JOIN countcol by
(group.xdseg,group.date,group.hour), count30 by
(group.xdseg,group.date,group.hour);
out         = FOREACH joining GENERATE
countcol::group.xdseg AS xdseg,countcol::group.date AS
date,countcol::group.hour AS hour, countcol::counterall
AS countall, count30::counter30 AS count30;
uniout      = DISTINCT out;
STORE result INTO 'Sandeep/output/conf_ALL_neb'
USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',');
```

Figure 6. Typical code of calculating real time data

E. *Criteria 5 - Finding the anomaly from CDF distributions*

In probability concept and statistics, the cumulative distribution function (CDF) of a real-valued random variable X, assessed at X, is the likelihood that X will take a value smaller than or equal to X. For continuous allocation, it provides the neighborhood below the probability density function from minus infinity to X. Cumulative distribution functions are also utilized to state the distribution of multivariate random variables [16]. Cumulative distribution function determines the cumulative likelihood of a failure hitting earlier than a specified period. The Equation (1) for the cumulative distribution function is provided by:

$$F(x) = P(X <= x) = \int_{0,-\infty}^{x} f(s)\, ds \qquad (1)$$

Here the lower limit is specified as zero or negative infinity. The value of the lower limit changes from distribution to distribution. For example, the normal or Gaussian distribution has a lower limit of negative infinity, while the Weibull distribution has a lower limit of zero. Note that the value of the cdf continually approaches 1 as time reaches infinity. This is because the region beneath the curve of the probability distribution function is always equivalent to 1, and the cumulative distribution function is basically calculating the region under the probability distribution function curve from zero to the point of interest.

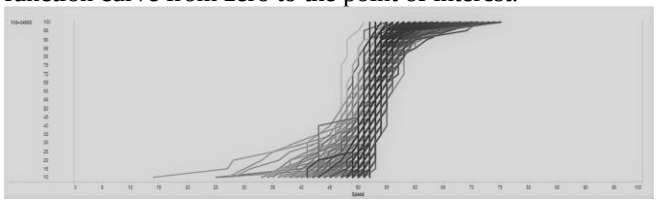

Figure 7. Reliability curve for TMC 118+04560

Figure 8. Typical code of calculating travel time/mile CDF quantile day wise

Figure 7 illustrate the reliability curve for one TMC and Figure 8 shows the code of calculating travel-time per mile CDF quantile day wise.

```
define Quantile datafu.pig.stats.StreamingQuantile('0.10',
'0.15', '0.20', '0.25', '0.30',  '0.35', '0.40',  '0.45', '0.50', '0.55',
'0.60', '0.65', '0.70',  '0.75', '0.80', '0.85', '0.90', '0.95', '1.00');
data        = LOAD 'Nebraska_Inrix/{2014,2015,2016}' using
PigStorage(',') As (code:chararray, time:chararray,
speed:double, avg_speed:double, re_speed:double,
traveltime:double, conf:double, cvalue:double);
data        = FILTER data BY conf == 30.0 AND traveltime >
0.0;
tmc_data = LOAD
'Nebraska_Inrix/others/others_2016/TMC_Identification.csv'
using PigStorage(',') AS (tmc_segement:chararray,
road:chararray, direction:chararray, intersection:chararray,
state:chararray, county:chararray, zip:double,
start_lat:double, start_long:double, end_lat:double,
end_long:double, miles:double, road_order:int);
joining     = JOIN data BY code, tmc_data BY tmc_segement
USING 'REPLICATED';
reqdcol     = FOREACH joining GENERATE data::code AS
code, data::traveltime AS traveltime,
SUBSTRING(data::time,0,10) AS
(date:chararray),(float)(data::traveltime/tmc_data::miles) AS
tt;
reqdcol1    = FOREACH reqdcol GENERATE code,
traveltime, date, tt;
date_code_group = GROUP reqdcol1 by (code, date);
all_quantile  = FOREACH date_code_group GENERATE
group.code AS code, group.date AS date,
Quantile(reqdcol1.tt) AS myquantile;
STORE all_quantile INTO 'Sandeep/output/travelpermile/I80'
USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',');
```


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

F. *Criteria 6 - Other Techniques - Interquartile Range*

The interquartile range is frequently applied to discover anomalies in data. These anomalies or outliers are observations that fall below Q1 - 1.5(IQR) or above Q3 + 1.5(IQR). The interquartile range (IQR), also identified as the mid-spread, or H-spread, is helped to explain the difference between the upper and lower quartiles, as shown in Equation (2). Quartiles allocate a rank-ordered dataset into four equal segments. The values that divide each segment are called the first, second, and third quartiles; and they are represented by Q1, Q2, and Q3, respectively. Quantile('0.25','0.5','0.75') yields the 25th, median, and the 75th percentiles.

$$IQR = Q3 - Q1 \quad (2)$$

```
define Quantile
datafu.pig.stats.StreamingQuantile('0.25','0.50','0.75');
data = LOAD 'Nebraska_Inrix/{2014,2015}' using
PigStorage(',') As (code:chararray, time:chararray,
speed:double, avg_speed:double, re_speed:double,
traveltime:double, conf:double, cvalue:double);
data30 = FILTER data BY conf == 30.0 and speed > 0.0;
selfeature = FOREACH data30 generate code,speed;
group30 = GROUP selfeature by (code);
myquantile = FOREACH group30 GENERATE
selfeature.code AS code, Quantile(selfeature.speed) AS
quantile;
iqr = FOREACH myquantile GENERATE
code,(quantile.quantile_0_75 - quantile.quantile_0_25) AS
iqr;
uniqiqr = DISTINCT iqr;
STORE uniqiqr INTO
'Sandeep/output/myquantilebyCode1415' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',');
```

Figure 9. Typical code of calculating inter quantile range for each TMCs

Figure 9 demonstrates the typical code used to calculate inter quantile range for each TMCs.Input probe-based stream data taken from INRIX were analyzed using Hadoop technology and then we have employed visualization tools like Tableau and R

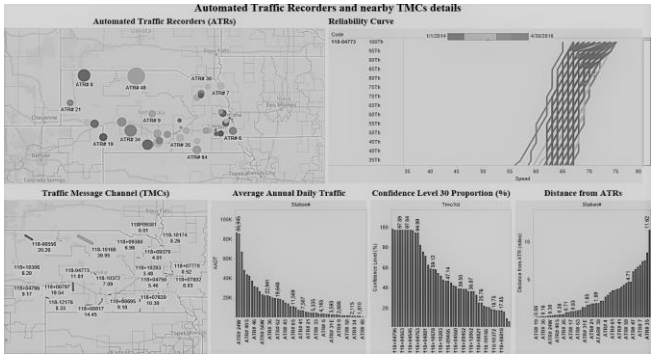

Figure 10. Dashboard view of 16 selected locations

programming to recommend best 16 locations to check the consistency and correctness of data stream as compared to fixed sensor data. Figure 10 shows the consolidated observation of all the selected automatic traffic recorder sites in the state of Nebraska with their consistency curve, adjacent traffic message controls, their respective average annual daily traffic, real time data with confidence level (30) and least distance from ATRs mid-point.

G. *Discussions*

Latest sophisticated technologies in future can not only offer user feedback and plenty opportunities but also going to endure to impact and enrich revolution inside the different transportation agencies. We propose the few suggestions to transportation industry reflecting the practice of stream data to enhance traffic handling related decision-making administration:

- To combine probe-based stream data with other major data sources, state DOTs and transportation agencies should associate the probe-based streams to the linear referencing approach, as many state DOTs and agencies specify road sections based on a linear referencing approach.
- TMCs segment lengths are fluctuates critically, from 0.5 miles to around 8 miles. These industries have to observe whether these TMCs receiving data stream are enough for planned functionalities.
- This study was focused mainly on freeways consist of both interstates and non-interstates, upcoming job must be concentrated on a greater investigation of metropolitan areas and this can be achieve by installing additional fixed sensors on both expressways and metropolitan regions.

IV. CONCLUSION & FUTURE WORK

Different techniques were used to identify 16 locations to assess the reliability and correctness of INRIX data versus static mounted sensor device data. For better evaluation we have also taken the number for heavy truck and inter quantile range (IQR) for all selected TMCs. In future, we will be collecting the real-time TMCs and XDs data based on our selected locations and then estimate the reliability and correctness of INRIX data. The precision of the data stream will be assessed with different measures such as roadway type, average vehicle speed or volume, the percentage of truck traffic, road segment length, and probe penetration.

Many directions for future enhancements are open. Among them, we can mention:

- To assess the reliability and correctness of a probe-data stream using different metrics like Absolute Average Speed Error (AASE) & Average Speed Bias (ASB), etc. and
- use of machine learning techniques like classification and clustering to find the anomalies.

ACKNOWLEDGMENT

This work was supported by the Nebraska Department of Roads, Lincoln, NE under the project titled “Evaluation of Opportunities and Challenges of Using INRIX data for Real Time Performance Monitoring and Historical Trend Assessment”. The authors are thankful to the anonymous referees for their useful comments.

REFERENCES

1. National Research Council Frontiers in Massive Data Analysis. The National Academies Press, Washington, DC, 2013. https://bigdatawg.nist.gov/pdf/FrontiersInMassiveDataAnalysisPrepub.pdf
2. The National Security Agency: Missions, Authorities, Oversight and Partnerships [Online]. https://fas.org/irp/nsa/nsa-story.pdf
3. J. Gantz and D. Reinsel, “Extracting Value from Chaos”, Hopkinton, MA, USA: EMC, June 2011.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

4. J. Gantz and D. Reinsel, “The Digital Universe Decade—Are You Ready”, Hopkinton, MA, USA: EMC, May 2010.
5. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute [Online], May 2011. https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf
6. http://www.inrix.com
7. https://hadoop.apache.org
8. Biswanath Panda et al., “PLANET: massively parallel learning of tree ensembles with MapReduce”, in Proc. of the VLDB Endowment, v.2 n.2, August 2009 [doi>10.14778/1687553.1687569]
9. Jens Dittrich et al., “Hadoop++: making a yellow elephant run like a cheetah (without it even noticing)”, in Proc. of the VLDB Endowment, v.3 n.1-2, September 2010.
10. Tyson Condie et al., “MapReduce online”, in Proc. of the 7th USENIX conference on Networked systems design and implementation, San Jose, California, pp.21-21, April 28-30, 2010.
11. He, Yongqiang et al., “RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems”, in Proc. International Conference on Data Engineering, 2011. 1199-1208. 10.1109/ICDE.2011.5767933.
12. Ye Zhou, Amir Esmailpour, "Improvements in Big Data Hadoop Several hybrid efficiency methods", in Proc. ASEE 2014 Zone I Conference, April 3-5, 2014.
13. D. J. Abadi et al., “Column-stores vs. row-stores: how different are they really?”, in Proc. SIGMOD Conference, pp. 967–980, 2008.
14. Sanjay Ghemawat et al., “The Google file system”, in Proc. ACM symposium on Operating systems principles, NY, USA, pp. 19-22, October 2003. [doi>10.1145/945445.945450]
15. Coifman, B., and Kim, Seoungbum, “Assessing the performance of Speed Info Sensor”, Report: The Ohio Department of Transportation, Office of State wide Planning and Research, 2013.
16. https://en.wikipedia.org/wiki/Cumulative_distribution_function
17. Q. Shi, M. Abdel-Aty, “Big Data Applications in Real-time Traffic Operation and Safety Monitoring and Improvement on Urban Expressways”, Transportation Research Part C: Emerging Technologies, Vol. 58, pp. 380-394, 2015.

AUTHORS PROFILE

Dr. Sandeep Singh Rawat obtained his Bachelor of Engineering in Computer Science from National Institute of Technology, Surat (formerly REC, Surat), his Masters from Indian Institute of Technology, Roorkee and doctorate degree from University College of Engineering, Osmania University, Hyderabad. Dr. Sandeep was a Research Scholar and Visiting Professor at Iowa State University, USA from 2016 to 2018. Dr. Sandeep is an experienced researcher and consultant with proven leadership abilities in areas of Knowledge Extraction, Information System & Machine Learning, Database, Information Retrieval, data warehousing, cloud computing and Groups Decision Support Systems. He has published/presented high-quality research papers in International, National Journals and proceedings of International Conferences. He has organized many workshops and conferences.

Dr. Sharma is an associate professor in the Department of Civil, Construction, and Environmental Engineering at Iowa State University. Dr. Sharma’s research has been recognized by numerous funding agencies, including the National Science Foundation, Federal Highway Administration, National Institute of Health, several state departments of transportation, and multiple city public works departments. Dr. Sharma has been granted awards of more than $10 million, which contribute to fund his research. He has served as principal investigator/co-principal investigator for over 32 research projects over 9 years.His research uses big data-driven discoveries to help make better short-term (usually automated control) and long-term (policy) decisions. Dr. Sharma instrumented and is currently leading the REACTOR (REaltime AnalytiCs of TranspORtation data) Laboratory. The lab is able to ingest multiple streams of real-time data to assist in driving transportation policy decisions. Research efforts at the lab are focused on data ingestion, real-time analytics, batch processing, visualization/front end development, and archiving of numerous data streams