DOI 10.20544/HORIZONS.B.03.1.16.P11
UDC 004.65:681.586]:656.053

BIG DATA IN TRAFFIC[1]

Slađana Janković[1], Dušan Mladenović, Snežana Mladenović,
Stefan Zdravković, Ana Uzelac
University of Belgrade
Faculty of Transport and Traffic Engineering, Vojvode Stepe 305
Belgrade, Serbia
s.jankovic@sf.bg.ac.rs

Abstract

It is known that different types of cameras and sensors are the backbone of traffic monitoring today. With the evolution of intelligent devices grows the amount of data that these devices generate. How to collect data is no longer the topical issue of the today. The question is how to store and handle increased amounts of data? Big Data technology gives the answer to this question. The term Big Data refers to an information resource that is characterized by large quantity, high-speed growth and a large variety of data that exceed capabilities of software that are commonly used for storage, processing and data management. It is the amount of data that can be measured in petabytes and the speed of information inflow that is greater than the speed of processing. The main aim of this paper is the familiarization with the Big Data technology, its most popular tools and the advantages in processing large amounts of traffic data.

Keywords–big data; traffic counting; sensors data; Apache Hadoop

INTRODUCTION

The term **Big Data** is often used when talking about the amount of data which exceeds the possibilities of commonly used software designed for storage, processing and data management. We could say that Big Data represents everything which does not fit into Microsoft Office Excel Workbook. It is about the amount of data measured in**petabytes** (1PB = 1015B = 1000TB) and the speed of flow of information that is greater than the speed of processing. An important feature of the data itself in the Big Data concept is the diversity of formats and data sources. The data generated

[1] Original scientific paper

by sensors and other intelligent devices in the field of transport contain all of these features.With the development of intelligent devices theamount of data which they generate grows. For example, one Airbus aircraft during an average flight generates around 1TB of sensory data. Nowadays, the issue ofcollecting data is no longer the main topic for discussion. However, the question is how to store and handle increasing amounts of data? Big Data technologies give the solution to the problem.Therefore, the objective of this paper is to explore the possibilities of using Big Data concept in transportation and traffic engineering.

The first section of the paper outlines the possible examples of application of the Big Data technology in the area of storage and processing of data in traffic and transport. The second section is a brief introduction to the Big Data technology and its most popular tools. The third section describes a working example in terms of storage and processing of data in traffic, where Big Data solution gives better results than traditional software. Finally, the conclusion of this study emphasizes the benefits of Big Data technologies in the field of transport.

BIG DATA IN TRAFFIC - MOTIVATION

Observation of traffic includes the following activities: detecting the presence of vehicles, traffic counting, measuring length, categorization of the vehicles, and many more. In the current play of events, transportation industry cannot meet the rapid growth of data; also, the traditional data processing systems are facing the problem of inefficiency or even sometimes failure [1]. As an important branch of economy, transportation and traffic confront the challenges and promises brought to us by Big Data. Currently, the most widely used data sources are traffic surveillance systems [2]. Big data systems are ideal for monitoring the behaviour of the system, collecting and analysing defects. These systems offer previously unimaginable possibilities for monitoring operation of the system in detail. According to Zeng, the advantages of Big Data concepts are improving the safety of traffic and the efficiency of transportation industry in general [1]. Traffic sensors for data collection are inductive-loop detectors, video image processing systems, pneumatic tubes, global positioning system (GPS), acoustic/ultrasonic sensors, aerial/satellite imaging, and RFID (Radio Frequency IDentification) technology. They can be classified in many ways. Martin et al. proposes that the highly developed detection technologies can be classified into three categories: in-roadway detectors, over-roadway detectors, and off-roadway technologies [3]. Each of these sensors havetheir advantages that provide the real-time information for road users and transportation system operators to make better decisions. Intelligent Transportation Systems (ITS) use big data with the aim to increase energy efficiency, improve traffic safety, reduce air pollution, relieve traffic congestion, and improve homeland security [4]. Every day, companies are investing more and more into this area. Recently, a railway operator invested in an automated system that uses big data to manage the rescheduling of more than 8,000 trains. Likewise, the drivers in Boston can use an app known as "Street Bump" to detect the unmistakable jolt of a pothole. It captures bump data using your GPS location. This data can be used to identify roads in greatest need of repair [5]. Another purpose of using big data analytics is that transportation planners can allocate limited resources in the areas where they can boost the capacity of congested transportation networks [6]. Obviously, there is a need to build an open platform allowing many departments of traffic, companies, and individuals to collaborate and share the same data.

BIG DATA TECHNOLOGY

The term Big Data was created in 2008. The very name "Big Data" clearly indicates that it is about a large amount of data. But how do we know whether, for example, a 10TB relational database is Big Data? The **Big Data dimensions** such as volume, variety and velocity, which were firstly defined by IBM, provide an answer to this question (Fig.1).

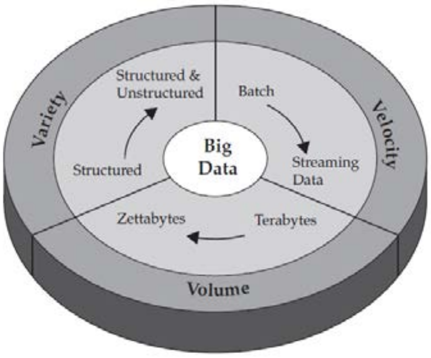


Fig.1.Big Data dimensions

Volume - a high speed of growth of the amount of new data and keeping of the existing data leads to hundreds of terabytes of storage, and even much larger amounts of data. **Variety -** it is no longer enough to keep only the structured data, but also images, information from social networks, logs, sensor information, and so on. **Velocity -** the speed of the new incoming data is big and it is therefore higher than the speed of data processing. If some of the data we manage have these characteristics, then we could say that the system has/isBig Data.As far as the software for storing and processing large amounts of data go, **Apache™ Hadoop®** is currently the only one available.

Hadoop is an open-source framework for storing and processing Big Data in a distributed environment. It contains two modules, one is **MapReduce** and the other one is **Hadoop Distributed File System** (**HDFS**). MapReduce is a parallel programming model for processing large amounts of structured, semi-structured, and unstructured data on large clusters of commodity hardware. HDFS is a part of Hadoop framework, used to store and process the datasets. It provides a fault-tolerant file system to run on commodity hardware. **Apache Hive**is a data warehouse infrastructure tool designed to process structured data in the Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analysing easy. The Hive is built into the “SQL like" query language **Hive Query Language** (HiveQL, HQL) which allows the manipulation of data in the Hadoop in writing queries that are almost identical to standard SQL queries. This tool is easy to integrate within the existing tools such as **Microsoft Office Excel**.

A BIG DATA SOLUTION IN TRAFFIC

We have recognized the **sensory data**, which is widely used in traffic, as a category of data that makes sense to cultivate using the Big Data technologies. The main indicator of the load of a certain section of the road is the **Annual Average Daily Traffic** (**AADT**). For sections where there are **automatic traffic counters**, AADT is calculated on the basis on data generated by the traffic counters on that particular section, throughout the year. We set ourselves the task to calculate the AADT for certain sections of roads and streets in the town of Novi Sad, Serbia and its surroundings, based on data generated by the 14 traffic counters (Table 1).

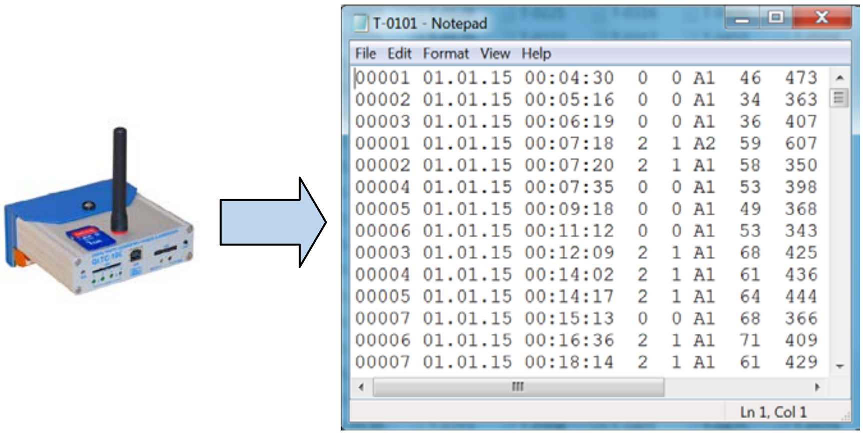
Table 1.Traffic Counters

Traffic Counter Name	Road/Street
Alibegovac	M-21
Bocke	R-107
Bulevar Evrope 3	K. Stankovića - Rumenački put
	Rumenački put - K. Stankovića
Klisa	M-22-1
Paragovo	M-21
Šančevi	R-120
Somborski bulevar (Patrijarha Pavla)	Bulevar Evrope - Vršačka
	Vršačka - Bulevar Evrope

Tekije	M-22-1
Tunel	Irig - Novi Sad
	Novi Sad - Irig
Vojvode Stepe	Rumenačka – S. Jovanovića
	S. Jovanovića - Rumenačka

The aim was to calculate the first attempt by using the conventional software, and then with the help of Big Data tools, and lastly to compare their features and efficiency. To count the traffic at the specified locations, the counters type **QLTC-10C**were used. Data generated by traffic counters are kept in text (.txt) files (Fig. 2). For each vehicle registered by a counter, one record is created in the txt file. Record in the txt file, contains the following information:

- **Index**:a daily number of the vehicles per traffic lane;
- **Date and time**: in the following format: dd.mm.yy hh:mm:ss;
- **Channel**:it may have the values from 0 to 3 depending on the order in which the vehicle encounters a loop;
- **Lane**: it can have values: 0 (vehicle in lane 1) or 1 (the vehicle in lane 2);
- **Vehicle class**: vehicle category;
- **Vehicle speed**:expressed in km/h ;
- **Vehicle length**:length without the correction factor, expressed incm.
-



```
00001 01.01.15 00:04:30  0  0 A1  46  473
00002 01.01.15 00:05:16  0  0 A1  34  363
00003 01.01.15 00:06:19  0  0 A1  36  407
00001 01.01.15 00:07:18  2  1 A2  59  607
00002 01.01.15 00:07:20  2  1 A1  58  350
00004 01.01.15 00:07:35  0  0 A1  53  398
00005 01.01.15 00:09:18  0  0 A1  49  368
00006 01.01.15 00:11:12  0  0 A1  53  343
00003 01.01.15 00:12:09  2  1 A1  68  425
00004 01.01.15 00:14:02  2  1 A1  61  436
00005 01.01.15 00:14:17  2  1 A1  64  444
00007 01.01.15 00:15:13  0  0 A1  68  366
00006 01.01.15 00:16:36  2  1 A1  71  409
00007 01.01.15 00:18:14  2  1 A1  61  429
```

Fig.2.QLTC Traffic Counter and the file which it has generated

Fig.2. displays the file T-0101.txt which stores data generated by a counter on January 1st 2015. The name of the txt file contains a date and the current year when traffic counting was performed. Each counter, during the course of the day, "writes" the data into a text file, so that during one year each counter generates 365/366 files. Txt file size is determined by the volume of traffic in one day at the observed counting place. In our case study, each of the 14 counters had 365 files and each text file kept between 4000 and 14000 records. This led us to the following question: **how to handle more than 50 million records and draw any useful information from such a large number of data?**We first sought the help of one of the most powerful modern tools for spreadsheet and graphic presentation of data **- Microsoft Office Excel 2013**. Firstly, we imported the data generated by one counter during one day. Since each counter has registered fewer than 14,000 recordswithin one day, we found that such txt files can be successfully imported into the Excel worksheets. One only needed to import 365 files and generate 365 Excel worksheets! Off course, the maximum number of worksheets in Excel 2013 workbook is determined by the available memory. Also,in 64-bit environments, the file size is limited by the available memory only. However, it is clear that this strategy did not lead to an efficient solution to the above problem.

Then we changed our strategy and decided to unite the contents of all 365 txt files from one counter into a single text file in order to import this txt file intothe Excel worksheet in the next step. For this purpose we have developed a Windows application in the development environment Microsoft Visual Studio 2015 and Visual Basic. This application created one text file for each counter and enrolled him with all the correct entries from 365 files that he has generated. In this way, we have created 14 large .txt files out of 365 x 14 = **5110 txt files**. The number of entries in each of the 14 txt files is shown in Table 2.

Table 2.The number of entries in the individual txt files

Name of the txt file	The number of records in a txt file
Alibegovac.txt	4 088 971
Bocke.txt	4 363 535
Bulevar Evrope 3 - K Stankovica - Rumenacki put.txt	2 037 270
Bulevar Evrope 3 - Rumenacki put - K Stankovica.txt	1 606 137
Klisa.txt	3 413 757
Paragovo.txt	3 674 736
Sancevi.txt	4 240 577
Somborski bul (Patrijarha Pavla) - Bul Evrope – Vrsacka.txt	4 173 994
Somborski bul (Patrijarha Pavla) - Vrsacka - Bul Evrope.txt	3 758 804
Tekije.txt	5 001 035
Tunel - Irig - Novi Sad.txt	2 107 966
Tunel - Novi Sad – Irig.txt	2 911 248
Vojvode Stepe - Rumenacka - S Jovanovica.txt	4 784 728
Vojvode Stepe - S Jovanovica – Rumenacka.txt	4 504 969
Total	**50 667 727**

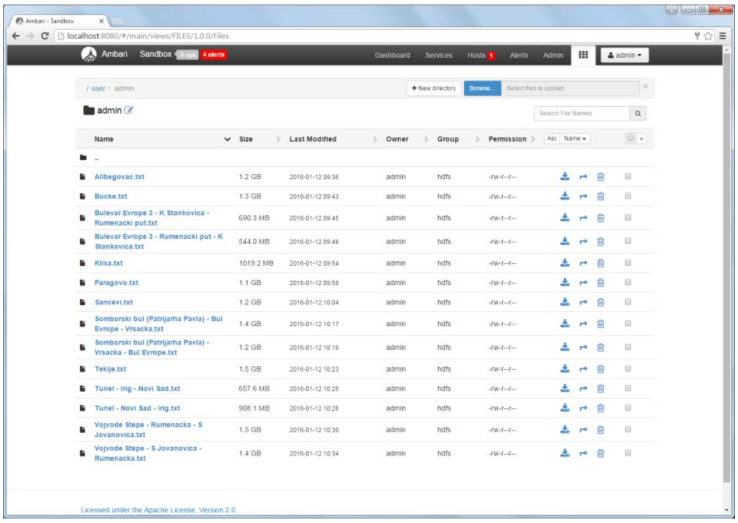
When we tried to import any of the txt files from Table 2 into the Excel worksheet we received a message that the file was too large and cannot be imported into the Excel. Therefore, we found that Excel could not accept all the data which was generated by the counter during one year into a single worksheet. The maximum number of species in the Microsoft Office Excel 2013 worksheet is 1 048 576. However, as it can be seen from Table 2, the number of records generated by the counter during one year ranges from 1,606,137 to 5 001 035. Accordingly, even this strategy did not lead to the solution of our problems.

We sought the help of the latest available tools designed to handle large amounts of data - Big Data tools. We used Apache™ Hadoop®, as it is currently the only available platform. Furthermore, we developed the following **methodology for processing data from traffic counters**:

1. "clean up" txt files of any invalid records,
2. for each counter, consolidate the content of all 365 txt files into a single text file, which would generate 14 large txt file,
3. upload each of the 14 .txt files into the Hadoop Big Data platform HDFS,
4. create a database in the Hadoop platform in which to store data from the 14 uploaded txt files,
5. create useful queries against a previously created database in the Hadoop platform,
6. make available the query results that are stored in a database on Hadoop platform in a conventional environment for analysis and data visualization with Microsoft Office Excel 2013,
7. Analyse the data and graphically display them in Excel.

As it can be noted, we have already carried out the first two steps in the previous attempts to solve the task with the help of Excel. In the next phase,we uploaded 14 large .txt files to the HDFS, using the**Apache Ambari** user interface (UI) (Fig. 3). As it can be seen in Fig.3, the size of every individual txt file amounts up to 1.5 GB!

Using the Apache Hive query language services and HiveQL, we created a database titled **Traffic** at the Hadoop platform (in the same environment - Apache Ambari). Thereafter, for each text file we generated one table in the Traffic database. We “filled” the tables with the data from txt files that are stored on HDFS.



Name	Size	Last Modified	Owner	Group	Permission
..					
Alibegovac.txt	1.2 GB	2016-01-12 09:36	admin	hdfs	-rw-r--r--
Bocke.txt	1.3 GB	2016-01-12 09:43	admin	hdfs	-rw-r--r--
Bulevar Evrope 3 - K Stankovica - Rumenacki put.txt	690.3 MB	2016-01-12 09:45	admin	hdfs	-rw-r--r--
Bulevar Evrope 3 - Rumenacki put - K Stankovica.txt	544.0 MB	2016-01-12 09:48	admin	hdfs	-rw-r--r--
Klisa.txt	1019.2 MB	2016-01-12 09:54	admin	hdfs	-rw-r--r--
Paragovo.txt	1.1 GB	2016-01-12 09:59	admin	hdfs	-rw-r--r--
Sancevi.txt	1.2 GB	2016-01-12 10:04	admin	hdfs	-rw-r--r--
Somborski bul (Patrijarha Pavla) - Bul Evrope - Vrsacka.txt	1.4 GB	2016-01-12 10:17	admin	hdfs	-rw-r--r--
Somborski bul (Patrijarha Pavla) - Vrsacka - Bul Evrope.txt	1.2 GB	2016-01-12 10:19	admin	hdfs	-rw-r--r--
Tekije.txt	1.5 GB	2016-01-12 10:23	admin	hdfs	-rw-r--r--
Tunel - Irig - Novi Sad.txt	657.6 MB	2016-01-12 10:25	admin	hdfs	-rw-r--r--
Tunel - Novi Sad - Irig.txt	908.1 MB	2016-01-12 10:26	admin	hdfs	-rw-r--r--
Vojvode Stepe - Rumenacka - S Jovanovica.txt	1.5 GB	2016-01-12 10:30	admin	hdfs	-rw-r--r--
Vojvode Stepe - S Jovanovica - Rumenacka.txt	1.4 GB	2016-01-12 10:34	admin	hdfs	-rw-r--r--

Licensed under the Apache License, Version 2.0.

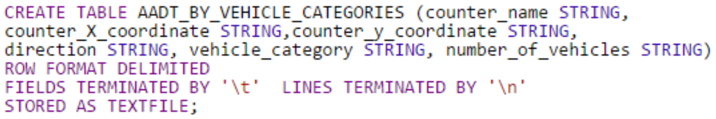
Fig.3. Uploaded TXT files to HDFS using the Apache Ambari UI

Figure 4 shows an example of HiveQL query with whose help we were able to "pour” all the data from the file Alibegovac.txt into the Alibegovac table.

```
LOAD DATA INPATH '/user/admin/Alibegovac.txt' OVERWRITE INTO TABLE alibegovac;
```

Fig.4.An example of the HiveQL query used for "pouring” data from txt file into the Traffic database table

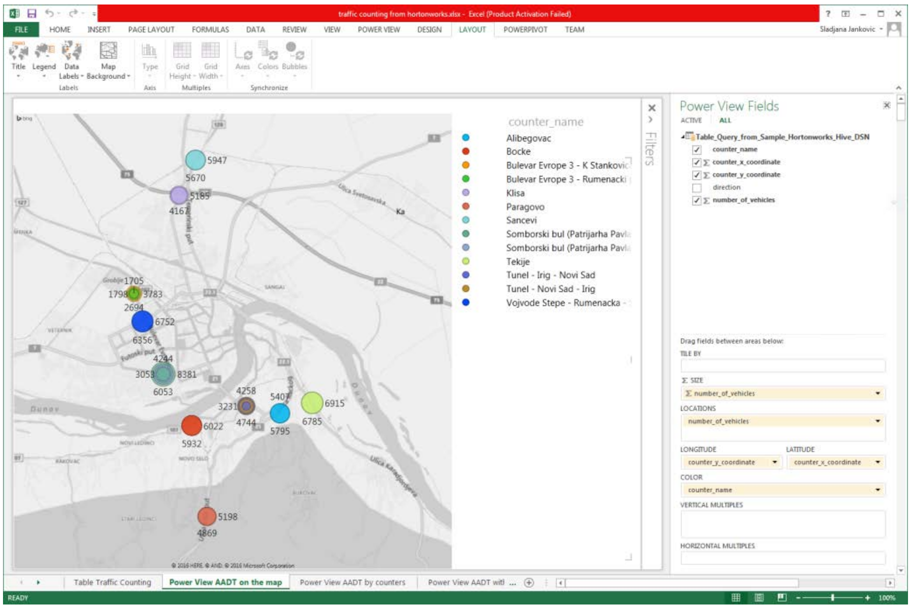
After that, we created a table named **all_counters**in the Traffic database in which we were able to move all the data from all 14 tables (of the database) with the help of HiveQL. Furthermore, we carried out numerous HiveQL queries/tests on the “all counters” table resulting in useful information on traffic volumes, structure and vehicle speeds, etc. Some of these queries/testswere: AADT per directionsfor each meter; Monthly Average Daily Traffic (MADT) for each month of the year according to directions, for each meter; AADT by directions and vehicle categories for each counter, etc. Finally,we placed the query results in the new tables orviewsof Trafficdatabase. Figure 5 shows an example of query used for creating tables intended for the admission of query results that were calculated by AADT by categories of vehicles and in the direction of their movement.



```
CREATE TABLE AADT_BY_VEHICLE_CATEGORIES (counter_name STRING,
counter_X_coordinate STRING,counter_y_coordinate STRING,
direction STRING, vehicle_category STRING, number_of_vehicles STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'  LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```

Fig.5.Example of HiveQL query that creates one of the tables in the database

In order to allow the visualization and geo-location of query results, we made available all the tables and views from Traffic database in Microsoft Office Excel 2013. This was done with the help of **Microsoft Query Wizard**. Analysis and graphical presentation of the data were made possible with the help of Excel's add-ons,**Microsoft Power View** and **Microsoft Power Map**. Figure 6 shows us one of the power view worksheets that graphically display the data from HadoopTraffic database. With the help of GPS coordinates counter which are stored in the Traffic database, the AADT size for each counter, for one of two directions, is shown in the**Microsoft Bing Map**. In this way, besides the graphic presentation of data in terms of traffic volumes, the geo-location was also conducted.



traffic counting from hortonworks.xlsx - Excel (Product Activation Failed)

counter_name
- ● Alibegovac
- ● Bocke
- ● Bulevar Evrope 3 - K Stankovic
- ● Bulevar Evrope 3 - Rumenacki
- ● Klisa
- ● Paragovo
- ● Sancevi
- ● Somborski bul (Patrijarha Pavl
- ● Somborski bul (Patrijarha Pavl
- ● Tekije
- ● Tunel - Irig - Novi Sad
- ● Tunel - Novi Sad - Irig
- ● Vojvode Stepe - Rumenacka -

Power View Fields

Table_Query_from_Sample_Hortonworks_Hive_DSN
- ☑ counter_name
- ☑ Σ counter_x_coordinate
- ☑ Σ counter_y_coordinate
- □ direction
- ☑ Σ number_of_vehicles

Drag fields between areas below:

TILE BY:

Σ SIZE: Σ number_of_vehicles

LOCATIONS: number_of_vehicles

LONGITUDE: counter_y_coordinate

LATITUDE: counter_x_coordinate

COLOR: counter_name

VERTICAL MULTIPLES:

HORIZONTAL MULTIPLES:

Fig.6.Example of visualization and geo-location of data contained in the Traffic database in Excel

CONCLUSION

Solving this simple task - calculating AADT for certain sections of roads and streets in the Republic of Serbia, pointed to weaknesses and limitations of modern tools for tabular and graphical presentation of data - Microsoft Office Excel 2013. In order to become familiar with Big Data technologies in this field, we have developed Big Data solutions that enable the storage, processing and graphical visualization of data generated by traffic counters on roads and streets in the Republic of Serbia. We used the Apache Hadoop Big Data platform to store and process all the data which could not have been performed using the Excel. Then we displayed the results of data processing, which we made with the help of Big Data technology, using the tabular and graphic display of Excel.

In this case, we have been convinced in the power and efficiency of Big Data technologies, especially when it comes to the conduction of "SQL-like" queries over massive amounts of data. In addition, we have been assured in the ability to connect Excel worksheets with the Big Data sources, as well as an enormous capability of Microsoft Office Excel 2013 tools in the field of visualization and geo-location of data. Due to the fact that traffic generates huge amounts of data, which is often necessaryto graphically display and geo-locate, the best solution is imposed by Big Data technologies in cooperation with the Excel.

ACKNOWLEDGMENT

This work has been partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under No. 036012.

REFERENCES

[1] G. Zeng, “Application of Big Data in Intelligent Traffic System“, IOSR Journal of Computer Engineering, 2015, 17(1), pp. 01-04.

[2] Q. Shi and 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, 2015, 58(B), pp. 380-394.

[3] P.T. Martin et al., 2003, “Detector Technology Evaluation“. Mountain-Plains Consortium, Available: http://www.mountain-plains.org/pubs/pdf/M

[4] PC03-154.pdf.

[5] D. Ni, “Traffic Sensing Technologies - Characteristics, Experimental Methods, and Numerical Techniques“ in Traffic Flow Theory, Ist ed. Waltham, Massachusetts, 2015, ch. I, sec. I, pp. 14-29.

[6] G. Leopold., 2014, September 02, "Big Data Gets Green Light for Traffic Management", Datanami. Available: http://www.datanami.com/2014/09/02/

[7] big-data-gets-green-light-traffic-management/.

[8] G. Leopold., 2014, July 02, "Big Data Helps Drive Transportation Planning", Datanami. Available: http://www.datanami.com/2014/07/02/big-data-helps-drive-transportation-planning/.