DEVELOPMENT OF TRAFFIC GEO-APPLICATION BASED ON BIG DATA PROCESSING

Slađana Janković[1],
Snežana Mladenović[1],
Ana Uzelac[1],
Stefan Zdravković[1],
Slaviša Aćimović[1]

[1]Saobraćajni fakultet,
Univerzitet u Beogradu,
Beograd, Srbija

Abstract:
Past is a prelude. Historical data provides signals that indicate what may happen in the future. Analytic tools can be especially useful in helping travel and transportation companies mine and refine data to determine which information is valuable for optimizing business outcomes. In this research we define a methodology for developing traffic geo-applications, based on: 1) relational data model for calculated predefined attributes of the maps, 2) wide column data model for source traffic data store, and 3) schema on read modeling approach for traffic data analysis, ie. calculating of the attributes of the maps. According to the proposed methodology we have developed Traffic Counting geo-application, which enables on-line access, displaying and geo-location of traffic indicators, calculated and stored on Apache Hadoop database. We concluded that our approach is appropriate in scenarios that involve batch Big Data processing and predefined data analysis, but not appropriate for the on-line ad-hoc definition of arbitrary queries.

Keywords:
Hadoop, HDFS, HBase, HiveQL, Shema on Read.

1. INTRODUCTION

Traffic and transportation are now unthinkable without GIS. On the other hand, GIS and Big Data are two parts of a whole [1]. GIS tools search, sift and sieve data from multiple and disparate databases to organize it for better workflows and spatial analysis [2]. They run operations that aggregate terabytes and more spatial information, run analysis, and visualize results as maps. All this occurs in real-time, with multiple data streaming into the existing GIS for better understanding of spatial trends and relationships. The Big Data approach to GIS allows analysis and decision making from huge datasets, by using algorithms, query processing and spatiotemporal data mining [3]. Simply put, this means extracting information from maximum possible sources using established procedures and computational techniques.

For small and medium-sized transport companies and transportation authorities in developing countries GIS tools that can be purchased on the market are too expensive [4]. On the other hand, excellent tailor-made traffic data is the best basis for excellent transportation models. We want

Correspondence:
Slađana Janković

e-mail:
s.jankovic@sf.bg.ac.rs

DOI: 10.15308/Sinteza-2017-458-463

to provide the traffic engineers and authorities with pre-attributed maps tailored to their specific needs, and to hire Big Data technology to calculate and store these attributes.

In the second section of the paper we have tried to answer the question: why Big Data technology in transportation? The third section is dedicated to defining the specific Big Data problem in traffic which we wanted to solve. In the fourth section we describe our solution based on Big Data processing of traffic data and Windows geo-application as graphical user interface (GUI). In the last section, we will discuss about the requirements, possibilities and constraints of our solution.

2. BACKGROUND

Companies focused on logistic management and transportation historically used data warehouses and business intelligence tools to report on and analyze customer behavior, optimize operations, and build advanced routing solutions. As logistics management and transportation networks become larger, more complex and driven by demands for more exacting service levels, the type of data that is managed also becomes more complex [5]. Today, these data sources include:

- Traditional enterprise data from operational systems,
- Traffic & weather data from sensors, monitors and forecast systems,
- Vehicle diagnostics, driving patterns,
- Financial business forecasts,
- Advertising response data,
- Web site browsing pattern data,
- Social media data [6].

Big Data is an approach to generating knowledge in which a number of advanced techniques are applied to the capture, management and analysis of very large and diverse volumes of data. As sensors become more prevalent in transportation vehicles, shipping, and throughout the supply chain, they can provide data enabling greater transparency than has ever been possible [7]. Such data will dwarf today's data warehouses and require Big Data Management Systems for processing and reporting. As a result of the complexity, diversity and stochastic nature of transportation problems, the analytical toolbox required of the transportation analyst must be broad [8].

Where Big Data differs from other technologies is in the sophistication of the analysis it applies [9]. While traditional analysis is often designed around the conditions that allow valid statistical inference about the characteristics of a population based on measurements on a small sample [10], Big Data analysis is built around the possibility of learning about systems by observing them in their entirety [11].

One of the leading vendors in the development of GIS-T applications is Esri. The most important Esri GIS-T applications are for: Airports and Aviation, Ports and Maritime, Railways, Roads and Highways and for Public Transport [12]. Esri offers the Spatial Framework for Hadoop that allows developers and data scientists to use the Hadoop data processing system for spatial data analysis. The Apache™ Hadoop® software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models [13]. For ArcGIS users Esri offers Geoprocessing Tools for Hadoop. Esri products visualize and analyze big data in a way that reveals patterns, trends, and relationships that reports don't. Even if data exists in many disparate places, Esri technology can pull it all together to help focus decision making.

Metro Transit of St. Louis (MTL) operates the public transportation system for the St. Louis metropolitan region. Hortonworks Data Platform (HDP) helps MTL meet their mission by storing and analyzing IoT data from the city's Smart Buses. HDP® helped the agency cut average cost per mile driven by its buses from $0.92 to $0.43 [14]. It achieved that cost reduction while simultaneously doubling the annual miles driven per bus.

Center for Advanced Aviation System Development (CAASD) provides the US Federal Aviation Administration (FAA) with specialized data analytics, simulation and computer modeling capabilities. "CAASD continuously ingests, stores and analyzes massive amounts of detailed flight data from a variety of sources and enriches it with other data, such as: pilot and air traffic controller voice recordings weather data, terrain maps, air traffic management system data, and aircraft schedules and flight plans. Altogether, CAASD has over a petabyte of data in its Hortonworks Data Platform (HDP™) clusters, deployed both on premises and in the cloud." [15].

3. PROBLEM DEFINITION

The amount of dynamic data generated in transportation industry by various sensors, GPS vehicle location tracking system and other mobile devices each year are developing from PB level to EB [5]. To count the traffic at the specified locations on the state roads in the Republic of Serbia 391 inductive loop detectors were used [16]. These detectors are QLTC-10C automatic traffic counters (ATC). Each counter generated 365/366 text files during one year.

Each file contained about 10,000 records on average, so that the collected data for one year amounted about:

$$391\ [\text{counters}] \cdot 365\ [\text{days}] \cdot 10{,}000\ [\text{vehicles}] = 1{,}427{,}150{,}000\ [\text{records}] \quad (1)$$

The first problem we wanted to solve was to hire Big Data technology to handle such large amounts of data (1) and to calculate the following indicators, from these data:

- Annual Average Daily Traffic (AADT) [vehicles/day],
- Monthly Average [vehicles/day], Daily Traffic (MADT)
- AADT per directions [vehicles/day],
- MADT per directions [vehicles/day],
- AADT by directions and categories [vehicles/day],
- Average speed of vehicles [km/h],
- Standard deviation of the vehicle speed [km/h],
- 85_{th} percentile of vehicle speed [km/h],
- Percentage of vehicles that exceed the speed limit [%],
- Average speeding [km/h], etc [17].

For example, AADT is a measure used primarily in transportation planning and transportation engineering [17]. AADT is a useful and simple measurement of how busy the road is. AADT along with it main characteristics – composition and time distribution (hourly, daily, monthly, yearly) is the basic and key input to the traffic-technical dimensioning of road infrastructure and road facilities. This parameter is used in: capacity analysis, level of service analysis, cost benefit analysis, safety analysis, analyses of pavement construction and for static calculation of road infrastructure objects, traffic forecasting, and others [18]. AADT presents the total volume of vehicle traffic of a highway or road for a year divided by 365 days (2).

$$\text{AADT} = \frac{\text{Total volume of traffic for 1 year [vehicles]}}{365\ \text{days}} \left[\frac{\text{vehicles}}{\text{day}}\right] \quad (2)$$

The second problem we wanted to solve was to enable end-users on-line access, displaying and geolocation of traffic indicators, calculated and stored on Big Data platform. We decided to solve this problem through the development of a Windows geo-application, which will work with a local database, but will also be able to communicate with the database on a Big Data platform. This application should enable GUI to query the Big Data database and display query results in the tables, graphics and on the maps. Also, our application should possess the ability to store the results of the Big Data analysis in the local database. This means, that it is supposed to solve the third problem: the integration of the results of Big Data analysis with the existing data in the local database.

4. DEVELOPMENT OF TRAFFIC COUNTING GEO- APPLICATION

We decided to carry out the calculation of the above indicators over such a large amount of data on Apache Hadoop Big Data platform. Our solution was applied through a case study of the analysis of traffic data for ten locations on the state roads and streets in the town of Novi Sad, Serbia which the traffic counters generated during the 2015. The solution was implemented through the following phases:

1. Data which was generated by ten QLTC-10C ATC were collected in text files on file server.

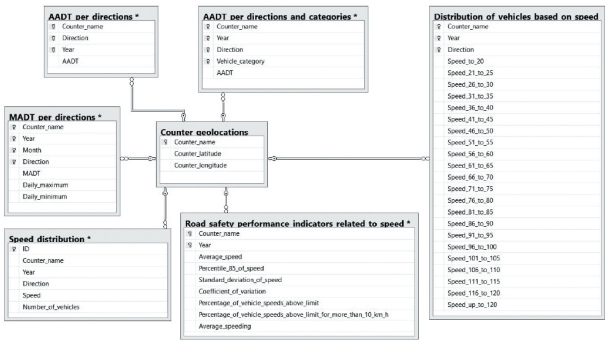

Fig. 1. Relational model of traffic data

2. Based on the structure of data contained in files generated by traffic counters and based on the requirements of traffic engineers, in terms of the traffic data analysis and required traffic indicators, a relational data model was designed. Based on the data model shown in Fig. 1. Microsoft SQL Server 2012 database Traffic Indicators was designed and created. This database was designed for local storage of the results of batch data processing on the Big Data platform, as a one-time operation, for each year. Each of the tables in the database contains indexed fields.
3. For the storage and processing of traffic data the Apache Hadoop platform and Apache HBase was chosen. Apache HBase is a column-oriented Database Management System that runs on top

of Hadoop Distributed File System (HDFS). We have developed a non-relational wide-column data model, taking into account the structure of the source files and designed relational data model.

4. Using the Apache Ambari user interface, on the Hortonworks Sandbox - a single-node Hadoop cluster, with the help of HiveQL query language, based on developed non-relational data model, Apache HBase Traffic Analysis database was created. This database is NoSQL database. NoSQL is a current approach for large and distributed data management and database design.
5. Another application was designed to "clean up" the text files of any invalid records generated by traffic counters. For each counter, this application has consolidated the content of all 365 .txt files into a single large text file. During this operation the structure of source files is not changed. Instead, files are only normalised. Then, we uploaded each of the ten large .txt files into the HDFS. Using HiveQL query language we "filled" the HBase tables with the data from the .txt files.
6. We carried out numerous HiveQL queries on the Hadoop Traffic Analysis database resulting in useful information on traffic volumes, traffic structure, vehicle speeds, etc. HiveQL has the powerful technique known as Create Table As Select (CTAS). This type of HiveQL queries allow as to quickly derive Hive tables from other tables in order to build powerful schemas for Big Data analysis. This data modelling approach is known as schema on read. The example of CTAS query shown in Fig. 2. creates a new table in Hadoop Traffic Analysis database.

```
CREATE TABLE AADT_per_directions_and_categories AS
SELECT Counter_name, SUBSTRING(TRIM(Date), 7, 4) As Year,
Direction, Vehicle_category, FLOOR(COUNT(*)/365) As AADT
FROM All_counters
GROUP BY Counter_name, Year, Direction, Vehicle_category;
```

Fig. 2. One example of Create Table As Select queries in Hadoop database.

The table AADT_per_directions_and_categories has a same structure as the namesake table in relational data model on Fig. 1. This will allow the integration of data calculated on the Hadoop platform and data in existing SQL database. Queries results were traffic volume indicators and traffic safety indicators, for each counting place: AADT; AADT by directions; MADT by directions; AADT by directions and vehicle categories; average speed of vehicles; standard deviation of the vehicle speed; coefficient of variation of vehicle speed; 85th percentile of vehicle speed; percentage of vehicles that exceed the speed limit; average speeding, etc.

7. In the Microsoft Visual Studio 2015, based on relational data model from Fig. 1., a Windows Forms application – Traffic Counting was developed. This geo-application will be used to interact with SQL Server database Traffic Indicators.

Traffic Counting geo-application has the following features:

a) An intuitive GUI that allows traffic engineers to define the query parameters and start executing the queries against the database Traffic Analysis on the Apache Hadoop platform was created. Access to the Hadoop database from the Windows Forms application was enabled with the help of Hortonworks Hive ODBC Driver;

b) The results of the query of the Hadoop database, with the help of Hortonworks Hive ODBC Driver, were stored in the local SQL Server database;

c) Furthermore, a user interface for graphical and tabular display of query results was generated. Finally, for geo- location of query results in the Traffic Counting application we utilized Bing Maps and OpenStreetMaps.

In Fig. 3. is shown one window from Traffic Counting application, that displays the results of CTAS query shown in Fig. 2. - AADT by directions and vehicle categories.

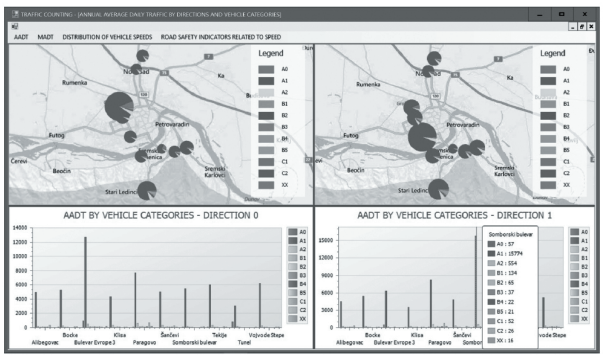

Fig. 3. TRAFFIC COUNTING geo-application – Annual Average Daily Traffic by directions and vehicle categories

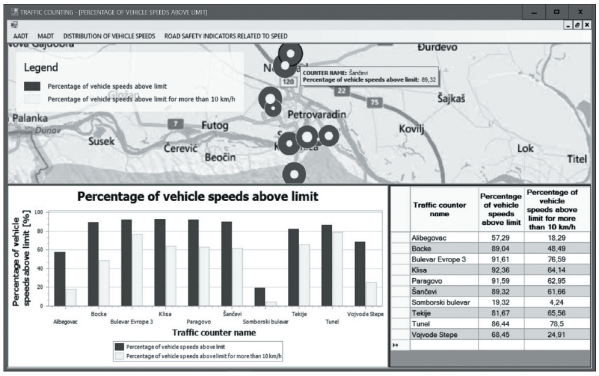

Fig. 4. TRAFFIC COUNTING geo-application – Percentage of vehicle speeds above limit

On the graphs and on the maps in Fig. 3., for example, can be seen that the intensity of traffic flow is very uneven per direction, for counting places Bulevar Europe 3 and Somborski Bulevar. In Fig. 4. is shown one window from Traffic Counting application, that displays percentage of vehicle speeds above limit, and percentage of vehicle speeds above limit for more then 10 kilometers per hour, for each counting place. On the graph, on the table and on the map in Fig. 4., for example, can be seen that the percentage of vehicle speeds above limit for counting place Šančevi is 89.32 percent, and that the percentage of vehicle speeds above limit for more then 10 kilometers per hour, is even 61.66 percent. Contrary to that, at a counting place Alibegovac 57.29 percent of vehicles exceeding the speed limit, while only 18.29 percent of vehicles exceeding the speed limit by more than 10 kilometers per hour. V.

5. CONCLUSION

The proposed methodology for development of traffic geo- application over the local SQL Server database and over the Hadoop database is appropriate in scenarios that involve batch Big Data processing and predefined data analysis. This model is not appropriate for on-line ad-hoc definition of arbitrary queries against Hadoop database. The proposed integration model of local data and data on Big Data platform implies schema on read modeling approach.

Schema on read is the revolutionary concept that we don't have to know what we're going to do with our data before we store it. When we access data, when we query it, then we determine the structure we want to use. There is always a time cost to imposing a schema on data. In schema on write strategies, that time cost is paid in the data loading stage. In schema on read strategies, that time cost is paid when we query data.

Flexibility and reuse of raw/atomic data make the schema on read approach appropriate in scenarios of sharing data of public interest. In further research we will try to take advantage of this approach in sharing information of public interest in the field of transport, between various stakeholders.

ACKNOWLEDGMENT

This work has been partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under No. 036012. The data generated by automatic traffic counters provided by the company MHM - Project from Novi Sad.

REFERENCES

[1] Michael F. Goodchild, "GIS in the Era of Big Data", Cybergeo: European Journal of Geography [Online], Cybergeo anniversary, Online since 25 April 2016, connection on 21 March 2017. URL: http://cybergeo.revues.org/27647

[2] M. Loidl, G. Wallentin, R. Cyganski, A. Graser, J. Scholz and E. Haslauer, "GIS and Transport Modeling-Strengthening the Spatial Perspective", International Journal of Geo-Information, vol. 5(6), pp. 1-23, 2016.

[3] S. Shekhar, "Transportation Data Mining: Vision & Challenges", Transportation Research Board Meeting 182, January 23rd, 2011.

[4] P. Russom, Integrating Hadoop into Business Intelligence and Data Warehousing, The Data Warehousing Institute, Renton, WA, 2013.

[5] G. Zeng, "Application of Big Data in Intelligent Traffic System", IOSR Journal of Computer Engineering, vol. 17(1), 2015, pp. 1-4.

[6] J.P. Rodrigue, The Geography of Transport Systems. Routledge, New York, 2017.

[7] 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, vol. 58(B), pp. 380-394, 2015.

[8] S.P. Washington, M.G. Karlaftis, F. Mannering, Statistical and Econometric Methods for Transportation Data Analysis, CRC Press, Taylor & Francis Group, Boca Raton, FL, 2010.

[9] Transportation Research Board, Statistical Methods in Highway Safety Analysis, TRB Executive Committee, 2001. P.

[10] Sridhar, N. Dharmaji, "A comparative study on how big data is scaling business intelligence and analytics", International Journal of Enhanced Research In Science Technology & Engineering, vol. 2(8), pp. 87–96, 2013.

[11] P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, Geographic Information Systems and Science, Wiley, Hoboken, NJ, 2015.

[12] Esri.com, "GIS for Transportation", 2016. [Online]. Available: http://www.esri.com/industries/transportation. [Accessed: 25-March- 2017].

[13] T. White, Hadoop: The Definitive Guide. O'Reilly Media, Inc. USA, 2015.

[14] Hortonworks.com, "Metro Transit", 2016. [Online]. Available:https://hortonworks.com/customers/metro-transit-of-st-louis/. [Accessed: 23-March-2017].
[15] Hortonworks.com, "MITRE", 2016. [Online]. Available: https://hortonworks.com/customers/mitre/.[Accessed: 23-March- 2017].

[16] K. Lipovac, M. Vujanić, T. Ivanišević, M. Rosić, "Effects of Application of Automatic Traffic Counters in Control of Exceeding Speed Limits on State Roads of Republic of Serbia", Proceedings of the 10th Road Safety in Local Community International Conference, April 22-25. 2015, Kragujevac, Serbia, pp. 131-140.

[17] S. Janković, S. Zdravković, S. Mladenović, D. Mladenović and A. Uzelac, "The Use of Big Data Technology in the Analysis of Speed on Roads in the Republic of Serbia", Proceedings of the Third International Conference on Traffic and Transport Engineering - ICTTE Belgrade 2016, November 24-25. 2016, Belgrade, Serbia, pp. 219-226.

[18] S. Janković, D. Mladenović, S. Mladenović, S. Zdravković and A. Uzelac, "Big Data in Traffic", Proceedings of the First International Conference Transport for Today's Society – TTS 2016, May 19-21. 2016, Bitola, Macedonia, pp. 28-37.