# Possibilities of Using Big Data Analytic in Police Work

Djordje Milosevic[1], Dane Subošić[1], Petar Vasiljevic[2], Vojkan Nikolic[1], Branko Markoski[3]
*University of Criminal Investigation and Police Studies, Zemun, Republic of Serbia*
Higher Education Technical School of Professional Studies, Novi Sad, *Republic of Serbia*
*University of Novi Sad , Technical faculty "Mihajlo Pupin" Zrenjanin, ,Republic of Serbia*
djodjolos@gmail.com, dane.subosic@kpu.edu.rs, vojkan.nikolic@kpu.edu.rs,
vasiljevic@vtsns.edu.rs, markoni@uns.ac.rs

**Abstract: The extensiveness of the data available is conditioned by the existence of technical devices whose readings connect to the Internet as both a challenge and a resource. However, in modern conditions, the main problem is not the collection and storage of large amounts of data, but the extraction of useful information from that data. This is also the essence of Big Data analytics. It represents selection and analysis, that is, intelligent decision-making based on a large amount of data collected. Gaining an advantage from its use is a high priority for members of the Ministry of the Interior. This is due to the growth in the use of numerous information systems, reliance on intelligence, eavesdropping and reconnaissance technologies (criminal intelligence, combating cybercrime), as well as the collection of a wealth of public administration data (biometric data, vehicle registration, residence and residence). The rapid processing of data and the development of analysis tools should be in the function of enabling timely and accurate data to be received by field staff (on road traffic safety, crime mapping, etc.). This paper points to the importance of correctly and fully interpreting such data in the limited period of time necessary for efficiency above all in police work. Unsurprisingly, significant financial investment by individual countries into the resources that would gain the ability to accumulate data and process it using Big Data analytics is a contemporary trend.**

**Key words: Big Data, artifact intelligence, MoI of the Republic of Serbia, analytic, criminal intelligence, police work**

## I. INTRODUCTION

Data management and analysis has always been a major challenge for society as a whole, and for the police and other employees of the Ministry of the Interior. The application of modern technologies today leads to the collection of an extremely large number of unstructured data, which poses a great challenge to relational databases. However, data collection and storage alone does not provide input to decision makers without selecting and analyzing such data. Creating analytical departments in the police and other appropriate services, which rely on large database instruments in their work, significantly improves the efficiency and effectiveness of the work of these entities.

Accordingly, this paper deals with a newer way of accessing work with a huge amount of data, which can be referred to as Big data analytics. Therefore, it represents how the data, which by the use of modern technologies will be collected, reduced, analyzed and used. Big data analytics is important because it enables the collection, storage, management and processing of huge amounts of data at high speed, which is inevitable in modern conditions.

## II. BIG DATA

The term "Big data" is a relatively recent date (first used in 1997) and is a system based on a specific information technology. Refers to large-scale structured or unstructured data, that is, to signify large datasets that exceed traditional storage methods by volume.

For example, a Google search yields 275 million Big Data views in 2017, of which 25.8 million are news, 53.3 million are videos, and 760,000 are books; 35.9% of the total findings were issued in the US; to the period from 01.01.2000. to 31.12.2011. 93.2 million data were published while in the period from 01.01.2012. do 10.04.2017. 741 million data entered; in the last 60 minutes compared to the search time 15,100 data were published; etc.

Data that is stored, and characterized as voluminous, is considered to exceed the capacity of commercial storage devices. To generate this type of data, to put it simply, information technology based on hardware, software and objects that send information to the server is used:

Hardware - Servers and clouds of considerable capacity that function as a central storage unit,

Software - tools and applications that connect objects to a server

Objects that send information to the server - e.g. TETRA devices.

The phenomenon of “Big Data”, or “Big Data” analytics, is definitely characterized by: volume, velocity, variety although other characteristics (complexity, probability, sensibility, quality, value, etc.) are also mentioned.

Volume - Many factors contribute to the introduction of data volumes (transaction data stored for years, textual data constantly coming from social networks, etc.). In the past, too much data has created storage problems, but with today's pricing and storage capacity, this is no longer a problem. Still, other problems arise, including determining the importance of certain data in large masses.

Variety - Today, data comes in many different formats. Here we have traditional databases, text files, e-mail, video, audio, financial transaction data, etc. According to some estimates, about 80 percent of the data are non-numerical, but they still need to be included in the analysis and decision-making procedures regarding them.

Velocity - Data processing speeds are two things. The first is the speed of data production and generation, and the second is the speed at which data must be processed to meet certain criteria. Timely response and fast data processing are a major challenge for the largest companies in the world.

## III. Data sources

The development of technologies used to process large amounts of data has contributed to the development of certain areas where such analyzes can be used. For example, great progress can be seen in the field of health or transport. In health care, the number of premature babies can be monitored and, depending on the obtained data, it can be estimated when a certain intervention is needed. In traffic, by analyzing a large amount of data generated by cameras placed on highways, it is possible to predict and regulate congestion and reduce the number of traffic accidents, save fuel, etc.

Data is obtained from a huge number of different sources and comes in different forms. With the rapid development of sensors, smart devices and social networks, data has become more complex, primarily because it now includes not only traditional structured data, but also unstructured data.

Although structured data seems to be well known, in fact, structured data is gaining a new role in light of the Big Data approach. The development of technology enables the emergence of new sources of structured data - often in real time and in large quantities. Such data sources are divided into two categories, computer or machine-generated data and human-generated data:

Computer-generated or machine-generated data - The term machine-generated data usually refers to data produced by a machine without human influence. For example, we classify sensor and web log data in this group

-Sensor data: Examples include radio frequency ID (RFID) tags, GPS data, and more. For example, RFID is rapidly becoming a popular technology. Miniature computer chips are used to track devices remotely. Another example of sensory data sources are smartphones that have sensors such as GPS.

-Web log data: When servers, applications, networks and the like work, they record different information about their activity. The amount of this data can become huge, and this data can be used, for example, to predict security breaches.

Human-generated data - this is data that is provided by people interacting with computers.

Unstructured data is data that does not follow a defined format. Unstructured data is actually the most common data, their number ranges up to 80% of available data. Until recently, however, the technology did not support other ways of working with this data other than storage and manual processing. Unstructured data can be found everywhere, because most people and organizations function on the basis of unstructured data. As with structured data, unstructured data can be machine or human generated. Some examples of machine-generated unstructured data are: satellite images (GoogleEarth), seismic images, etc.

In any case, the main problem today is not the collection of large amounts of data, but the selection and extraction of useful information from them. And it is Big data analytics that enables the obtained data to be understood and to make optimal use of their value.

## IV. BIG DATA ANALYTIC TECHNOLOGY

Technologies that are classified as "Big data" technologies not only support the ability to collect large amounts of data, but also enable their understanding. The main goal of the Ministry of Interior that has access to big data collections should be to use most of the relevant data in its work to make various right decisions. With the development of computer technology, it is now possible to manage huge amounts of data, which previously could only be processed and used with the help of supercomputers, and at a great cost. System prices have dropped and as a result of new techniques for distributed processing are currently in the focus of use. The real breakthrough in Big data technology happened when companies like Yahoo, Google, and Facebook came to realize that they could make money from the large amounts of data their products generated. These companies were faced with the task of finding a way in the form of some new technologies that will allow them to store, access, process and analyze huge amounts of data in real time, so that they can make a lot of money and use the amount of data they have and who participate in their networks. The resulting solutions have led to changes in the data management market. Some of these technologies will be explained further in the paper.

### *A. Spark*

Apache Spark is an open source project developed under the auspices of the Apache Foundation, and is a tool for fast and efficient processing of large amounts of data, and is one of the easiest Big data tools for learning and development. The reason for this lies in the philosophy behind the project, which is based on principles such as a single engine for the development of end-to-end data processing applications, whether batch, streaming or interactive data processing, development using rich and simple APIs, which will support performance optimization, integration with different systems where data is stored, given that moving data between systems is a set of operations, and integration with different components. Spark enables the use of the total RAM memory of a computer cluster to perform complex and demanding tasks when working with data. It's easy to scale - it's very easy to expand the total capacity of a Spark cluster by simply adding RAM or adding a new machine to the cluster.

A mistake that is often made when it comes to Spark's position in the Big Data sphere is that Spark is considered a replacement for Hadoop. Spark may replace some components within Hadoop, such as MapReduce for data processing, but we still need some other Hadoop components, such as HDFS data warehouse. In the previous period, great efforts were made in the development of Spark components used in data processing. The ease of development of Apache Spark applications is reflected in the very rich APIs that are supported in the programming languages Scala, Python, Java and R. API implementations retain some basic concepts from these programming languages that can help in data processing. So if you are developing a machine learning algorithm using the MLlib component in PySpark (Python API for Spark), the calculations will be performed using a NumPy library that is very efficient for tasks like this.

Another use case is streaming data processing. Spark is a very handy solution for tasks like this, and I use it when I need to process, say, Twitter data coming through streaming. Because it integrates easily with various messaging systems, such as Apache Kafka, you can feed the Spark Streaming application with a variety of data from different sources.

### *B. Hadoop*

Search engine innovators such as Yahoo! and Google have been tasked with finding a way to extract meaning and value from the vast amount of data their systems collect, ie. to understand at the same time what information they collect, as well as how to incorporate that information into their business and improve their business. Hadoop allows companies to easily manage large amounts of data. Hadoop lets big problems be broken down into smaller ones so analysis can be done quickly and cheaply. By breaking down these big problems into smaller parts that can then be processed in parallel, and after the processing is completed, this information is collected and grouped in order to issue the final results. Hadoop is a software framework derived from MapReduce and BigTable systems. Hadoop allows

MapReduce-based applications to run on large clusters of regular hardware. It is designed to parallelize data processing using nodes to increase computation speed and reduce response. Hadoop consists of two main components, a highly scalable distributed file system that supports the amount of data measured in petabytes, while the other component is the MapReduce system.

### C. Mapreduce

MapReduce is a solution introduced by Google as a way to efficiently execute a set of functions over vast amounts of data in a serial way. The map component distributes a programming problem or task to a large number of systems and manages task scheduling in a way that balances load and manages error recovery. After the distributed processing is completed, another function called "reduce" is called, which joins all the elements back together to provide the result. One example of MapReduce usage could be the task of determining how many pages of a book are written in each of some 50 different languages. MapReduce is a programming model for processing large data sets using parallel, distributed algorithms in a cluster.

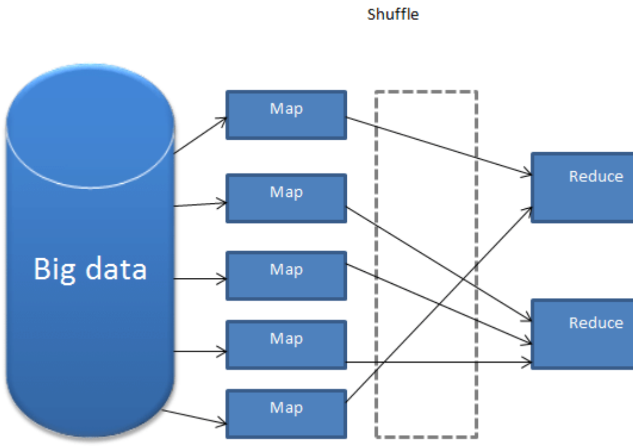

Figure 1. Mapreduce workflow

## V. BIG DATA ANALYTIC IN POLICE WORK

It has already been pointed out that Big Data analytics is about analyzing, forecasting, and intelligent decision making based on the large amount of data collected. Gaining an advantage based on its use is a high priority for both the police and other employees in the Ministry of Interior. This is due to the growth in the use of numerous information systems, reliance on intelligence, eavesdropping and reconnaissance technologies (criminal intelligence, combating cybercrime), as well as the collection of a wealth of public administration data (biometric data, vehicle registration, residence and residence).

The mentioned growth of the use of such funds burdens the activities of employees in MoI because a large amount of data for processing is accumulated. The rapid processing of data and the development of analysis tools should be designed to enable field staff to receive timely and accurate information (eg on road safety, crime mapping, etc.).

A particular problem is undoubtedly the (un) availability of experts capable of analyzing large amounts of data and then correctly interpreting different types of complex data structures based on their skills and work experience. This actually reflects the importance of the application of algorithms and Big Data computerized instruments, the application of which creates a sufficient amount of information of satisfactory quality.

Thus, during just one day of a simple complexity mission (eg in traffic control), the drone delivers to the control center 10 terabytes of data, of which only 5% is subject to analysis while the rest is stored. The lack of conditions for analyzing the remaining 95% of the data minimizes the overall quality of the decisions made at the tactical decision level.

High-quality video connectivity, high-resolution photo uploading, text content transfer such as coordinates, sensor readings and more make up such a high transmission capacity. In doing so, Big data enables the drone to collect video-audio data during the flight and to provide algorithmic "readout" through the video, which identifies traffic problems.

Admission of new members of the police force is an important element of the quality of police work. Big Data analytics tools (such as Google Trends, Google AdWords, and Google Correlate) provide layered insights into changes in the younger population's view of policing over time; bring the advertising of suitable jobs closer to those who search for possible jobs in that sector on the Internet (according to the way they are searched, it is possible to predict what the candidate searched on the Internet for several months before applying for a job), and the like. This improves the process of advertising the Ministry of the Interior in the media space, attracts candidates for admission and anticipates the intentions of those interested in admission to the service.

It is rightly considered that big data analytics enables more efficient police action by reducing the time of searching voluminous data, and at the same time increasing the processing speed of collected unstructured data, thus detecting hidden patterns and recognizing anomalies or similar information that can serve better and faster decision-making suppression of criminal activities.

The collection of an extremely large amount of data and the need to select and analyze those with intelligence and security implications, concludes that there is a need to improve existing practices to control, analyze and track the activities of individuals or groups through Big Data analytics creation of special Big Data departments within the analytical segment of intelligence and security agencies, not only in large and specialized services, but also in all others. Namely, in all countries there is a need for greater control of the telecommunication traffic of citizens in the country and abroad, in the sense that the data is adequately manipulated and quality intelligence is created. Therefore, it is not surprising that among the announced new jobs for the highly qualified workforce of the German BND service (Bundesnachrichtendienst), about 67% are jobs for IT specialists and cyber-infrastructure specialists.

## CONCLUSION

It is indisputable that the extensiveness of the available data is conditioned by the existence of technical devices whose readings are connected to the Internet as both a challenge and a resource. For example, video analytics are of increasing importance, which implies the generation of information based on video content (eg eavesdropping and surveillance, etc.). However, in modern conditions, the main problem is not the collection and storage of large amounts of data, but the extraction of useful information from that data. Today's technologies provide an opportunity to understand such data and reap the value of it. This is the essence of Big Data analytics.

This paper points to the importance of correctly and fully interpreting such data in the limited period of time necessary for efficiency above all in police work. Unsurprisingly, significant financial investment by individual countries into the resources that would gain the ability to accumulate data and process it using Big Data analytics is a contemporary trend. Thus, for example, the effectiveness of Big Data analytics for policing has already been validated using a similar methodology to identify unregistered vehicles in some Chinese cities (Shanghai).

Within the framework of criminal intelligence, the fight against cybercrime, the application of special investigative techniques (raster search), crime mapping and other modern forms of police work, it concludes the need to improve existing practices to control, analyze and monitor the activities of individuals or groups through Big Data analytics. The problem of collecting extremely large amounts of data and selecting and analyzing them also points to the need to create special Big Data departments within the analytical segment of police and security services, not only in large and economically advanced countries but also in other countries. Therefore, it is not surprising that among the new job openings for the highly qualified workforce of the German BND (Bundesnachrichtendienst), as many as two thirds are jobs for IT specialists and cyber infrastructure specialists.

## REFERENCES

[1] G. Šimić, A. Jevremović, I. Franc, M. Veinović.: *Baze podataka*, Univerzitet Singidunum, Beograd, 2013. Godina.

[2] Ž. Milojević, Lj. Dulović., *Velike baze podataka – Big Data i Primena u vojno-bezbednosnom sistemu*, Vojno Delo DOI: 10.5937/vojno delo1803236M

[3] Judith Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman - Big Data for Dummies, 2013.

[4] Bhatele A, Yeom JS, Jain N, Kuhlman CJ, Livnat Y, Bisset, KR, Kale LV, Marathe MV: Massively parallel simulations of spread of infectious diseases over realistic social networks. In: Proceedings - 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2017. p. 689–694 (2017).Srinath Perera, Thilina Gunarathne - Hadoop MapReduce Cookbook, 2013.

[5] Kesarev S, Severiukhina O, Bochenina K. Parallel simulation of community-wide information spreading in online social networks. Sci: Commun Comput Inf; 2018

[6] Severiukhina O, Kesarev S, Petrov M, Bochenina K. Parallel forecasting of community-wide information spread with assimilation of social network data. Procedia Comput Sci. 2018;136:228–35. https://doi.org/10.1016/j.procs.2018.08.260.

[7] Alam M, Khan M. Parallel algorithms for generating random networks with given degree sequences. Int J Parallel Program. 2017;45:109–27. https://doi.org/10.1007/s10766-015-0389-y.

[8] Quan Y, Jia Y, Zhou B, Han W, Li S. Repost prediction incorporating time-sensitive mutual influence in social networks. J Comput Sci. 2018;28:217–27.

[9] Li M, Wang X, Gao K, Zhang S. A survey on information diffusion in online social networks: models and methods. Information. 2017;8:118. https://doi.org/10.3390/info8040118.

[10] Mei S, Zarrabi N, Lees M, Sloot PMA. Complex agent networks: an emerging approach for modeling complex systems. Appl Soft Comput J. 2015;37:311–21. https://doi.org/10.1016/j.asoc.2015.08.010.

[11] Vega-Oliveros DA, Berton L, Vazquez F, Rodrigues FA. The impact of social curiosity on information spreading on networks. 2017. https://doi.org/10.1145/3110025.3110039

[12] Zhang ZL, Luo XG, García S, Tang JF, Herrera F. Exploring the effectiveness of dynamic ensemble selection in the one-versus-one scheme. Knowl-Based Syst. 2017;1(125):53–63

[13] Jallad KA, Aljnidi M, Desouki MS. Big data analysis and distributed deep learning for next-generation intrusion detection system optimization. J Big Data. 2019;6:88.

[14] Benqdara S. Anomaly intrusion detection based on a hybrid classification algorithm (GSVM). Int J Comp Appl. 2019;181(36):0975–8887.

[15] Chakir EM, Moughit M, Khamlichi YI. An effective intrusion detection model based on SVM with feature selection and parameters optimization. J Theor Appl Inf Technol. 2018;96(12):3873–85.