9th International scientific conference
Technics and Informatics in Education – TIE 2022
16-18 September 2022

Session: *IT Education and Practice*
Original scientific paper
DOI: 10.46793/TIE22.229D

Big Data Analytics Process Implementation on a Educational Data Set Extracted from Online Testing System

Gabrijela Dimić [1*], Ivana Milošević [1] and Ljiljana Pecić [1]
[1] Academy of Technical and Art Applied Studies Belgrade/
Department School of Electrical and Computer Engineering, Belgrade, Serbia
* gdimic@gs.viser.edu.rs

Abstract: *The paper presents an application of the big data analytics process on a data extracted from the educational system for online testing. The procedure of data migration to a non-relational system was implemented. Big data analytics process was realized through query processing in MongoDB and analysis results visualization in Microsoft Power BI. The contribution of the research presented in this paper refers to the importance of applying the big data concept for analyze data set extracted from educational systems in order to discover information and knowledge important for improving the efficiency of the educational process.*

Keywords: *big data; education system; data analytics; MongoDB; Power BI visualization*

1. INTRODUCTION

The rapid expansion of information and communication technologies has significantly influenced the change in the way data is transmitted and stored, and therefore it conditioned the continuous development of various areas of modern life and business. The field of education is one of the areas on which these changes have had a great impact. The most significant result of the strategic process of technology implementation in education is reflected in the development of various systems and platforms that have the ability to collect a large amount of data. The concept of new educational environments for e-learning, based on a three-tier architecture enables flexible design of the learning process [1]. Course management platforms, learning management systems (LMS), open online courses (MOOC), Open-CourseWare (OCW), Open Educational Resources (OER), personal learning environments (PLE) enable the automatic collection of heterogeneous types of data that can be analyzed and used to improve the educational process.

Considering the application of educational platforms and systems, it is evident that huge amounts of data are constantly being generated in different formats from different sources. The volume, variety and speed of data generated daily lead to the big data concept [2], [3].

The emergence of the concept of Big data systems has an increasing influence on various domains of business and social life, including the field of education [4]. The processing and analysis of big educational data enables the discovery of information and knowledge that is useful to teachers, students and the entire education system. In recent years, there has been an increasing need for the application of various methods of analyzing big educational data, with the aim of improving performance in the learning process, discovering and analyzing learning patterns, improving teacher efficiency. The importance of these data analysis is also reflected in the discovery of the reasons that influence the realization of worse outcomes of the learning process. With the removing the weak links from the educational process, the efficiency of the parts of the process that function less well, is significantly improving.

The architecture of Big data technologies enables storing data generated from a wide range of different distributed sources. The information system of some educational institution, as one of the sources, enables storing: student's personal data, data about subjects and achieved grades and all other data related to the study process.

Learning management system provides a database that stores systemic, user information and accumulates large amounts of data from activities such as: access and use of learning materials, testing, performance of various tasks, communication with other participants [5]. Based on predefined input parameters, these systems generate several types of reports. Considering the way of presentation and the amount of data, the process of analyzing the generated reports and

extracting useful information is an extremely difficult and time-consuming task. Therefore, tools that would help these systems to perform the mentioned tasks easier, are necessary. Although some e-learning platforms offer specific tools for creating reports, when there are a large number of students, the process of extracting useful information is complicated. One of the most popular LMS system is Moodle (*Modular Object Dynamic Learning Environment*) [6], which is open source system. It's main advantages are open access and very flexible working method.

Most of the automatically generated educational data is unstructured or poorly structured, and that makes impossible to use traditional tools for their analysis. As existing database management techniques have limitations in the case of using concept of large sets of heterogeneous data, it is necessary to explore new types of technologies, both for storing unstructured data and for their analysis and extraction of useful information and knowledge.

This paper describes a case study of the implementation of big data technologies in the process of analyzing automatically generated educational data. The data analyzed are from the online system designed for entrance exam preparation, and data are stored from 2017 to 2022. The data was migrated from the MySQL relational system to the MongoDB non-relational document-oriented database. Monitoring and visualization of migrated data was performed using the Microsoft Power BI environment for Big data analytics.

2. RELATED WORKS

During the last few years, researches started searching for the implementation of different environments of big data systems and infrastructures, in the process of collecting and analyzing educational data.

In the [7] authors suggesed distributed architecture for processing large educational data. Authors of the paper [8] described an architecture based on the Apache Hadoop cloud environment, intended for the analysis of educational data extracted from the Moodle system.

In the paper [9] authors suggested environment based on map-reduce algorithm for processing data from educational system and HDFS system for storing large amount of data.

Based on results of comparison of three systems, authors in [10] suggested intelligent educational system for large amount of data analysis.

Authors of the paper [11] identified sources of big data in educational environments and they suggest working methods for dealing with those data. In this paper the procedure for projecting appropriate architecture for using the concept big data is also shown.

In the paper [12] authors are discussing the importance of data science, the phenomenom of big data and Learning Analytics in education. A platform whose benefit and importance is reflected in the application of big data techniques in the field of education, has also been considered.

In [13] authors have provided an overview of implemented big data technologies in the field of education. The paper shows the importance of the application of big data analytics methods used in the function of improving the educational process efficiency, the productivity of the higher education institution with the aim of maintaining a competitive advantage on the education market.

3. BACKGROUND

3.1. The theory of Big data

It is difficult to establish a precise and unequivocal definition for the term Big Data, and there is no consensus about the fact from whom and when it was first used. According to some sources, first was John Mashey, one of the experts at The Silicon Graphics, Inc. It is believed that Mashey coined this term during the nineties of the twentieth century [14]. Translated literally, Big Data means "a large amount of data", but this cannot be considered as a definition of this term because it represents a much more complex phenomenon. Economist Francis X. Diebold from the University of Pennsylvania presented scientific article in which he mentions this term in paper's title [15].

Big Data can be defined as a set of technologies used for the process of combining huge amounts of data from various external sources and their analysis, using complex algorithms for making decisions based on knowledge and information that cannot be extracted in a traditional way, using conventional procedures and tools [16].

Big Data enables the management of large sets of different data, at a reasonable speed, in order to be able to analyze them in real time [17]. Also, Big Data implies situations in which existing software tools are not the most suitable for storing and processing targeted huge amounts of data, of different semantic types and structures, in a defined time and with the required accuracy, at a reasonable price [18].

Common for most definitions of this term are three significant aspects of big data that go beyond the capabilities of traditional data processing technologies. These are Volume, Velocity, Variety better known as 3V.

The amount of stored data (Volume), on a global level, is increasing every day at an exponential rate. The growing trend in the amount of available data is accompanied with the need to process it in an adequate manner, in order to reach new

conclusions and make better decisions. Some of the factors that contribute to this growth include: banking transactions, purchases (in retail and online), social networks, smart devices that contain various sensors, scientific projects, medical data (it is estimated that almost 30% of the world's data is some type of medical data, and it is expected that this trend will continue to grow in the coming period).

The challenge of achieving the highest possible speed (Velocity) in Big Data technology has two aspects: the speed with which new data is created and the speed with which existing data is processed and updated. Bearing in mind the importance of this aspect, it is clear that data must be collected, processed and analyzed in real time in order to gain insight into their essence as soon as possible. Since it is crucial for making business decisions to have timely and up-to-date data, it is clear the importance of the fact that it is available at the moment of decision-making. An example of this is sending customized personalized offers to consumers based on their previous purchase history, current location, and the like.

Variety is caused by the fact that data is collected from heterogeneous sources, in which data can be of different formats and models. The problem that needs to be solved is the extraction of different types of data, in such a way that will enable the connection of their meanings. According to its structure, data can be: unstructured, quasi-structured, semi-structured and structured. Most of the data is unstructured, which complicates the process of processing (because it requires a lot of knowledge and time). Unstructured data basically does not have a predefined model (images, text, video, sound, etc.). Structured data has a clearly defined format and structure and they are most often found in databases or data warehouses. They can be used multiple times and they are suitable for automatic processing. Sources of structured data can be machine-generated (generated by the devices themselves) or human-generated (generated by human-device interaction). Quasi-structured data are textual data that are given in a non-standard format so that they can be formatted, It takes a lot of time, tools and knowledge to do so. Semi-structured data can be defined as structured data that does not fit into the formal structure of the data model. Most often, they do not contain tags that separate semantic elements and do not have a common structure. Unstructured, quasi-structured and semi-structured data can basically be seen as a nominal type of data that requires different and much more complex methods of processing and analysis than in the case of structured data types. In practice, advanced statistical analysis techniques (cluster analysis), artificial intelligence methods (machine learning), etc. are most often applied. Structured data has a clearly defined type, or format, and can be categorized as metric or numerical types whose processing, analysis and interpretation is very precisely defined and relatively simple. However, storing and keeping large amounts of structured data can be problematic. It is important to keep in mind that growth is present in all types of data, but the speed of their generating is evidently different, which can be seen in Figure 1. The growth of structured data has a linear trend, and the growth of unstructured data is exponential function.

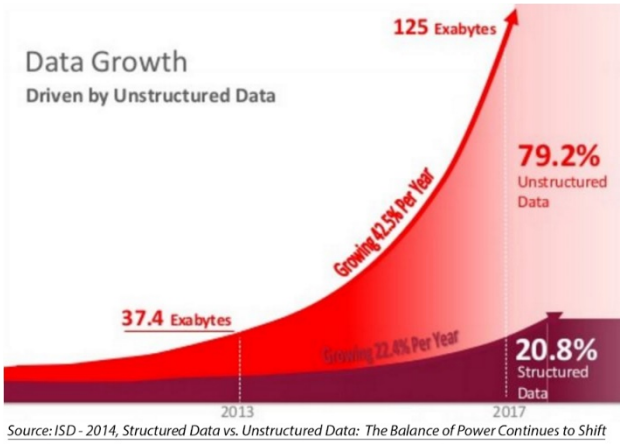

Figure 1. *Display of growth of structured vs. unstructured data [19]*

In recent times, four more Vs are added to the mentioned three important aspects: Veracity, Value, Verification, Visibility.

Veracity and controlling data reliability are a big challenge for Big Data analysts. Data is collected from a wide range of different sources and can be in different formats. Data accuracy is extremely important in the era of continuous and extremely fast generation of huge amounts of data, since huge amounts of data result in a certain amount of inaccurate data, which needs to be extracted in some way. Social media data often contains typos, abbreviations, and colloquial speech. Advanced methods of Big Data analytics, implemented through modern tools, are also adapted to such a structure, relying on the volume of data. Namely, the lack of quality and accuracy is compensated by the amount of available data.

Data collection should not be an end in itself, but the usefulness of the data (value) and the information derived from it is important. Value refers to the ability to convert data into profit. It is believed that, regardless of the activity, Big Data always has the potential to bring profit. An example of how the usefulness of the collected data can be used for making business decisions can be illustrated with the example of sending the nearest available delivery vehicle at the time of ordering (less fuel consumption, less mileage and less time spent).

It is clear that big data cannot be managed using traditional technologies, because in the absence of appropriate software and hardware support, big

data cannot be adequately exploited. The application of artificial intelligence (AI) algorithms and appropriate machine learning methods, which enable automated decision-making, have enabled more efficient and better insight into big data and its better understanding and analysis.

In today's time of growing attempts of various abuses of the weaknesses of computer systems, data verification is an extremely important and actual topic. When clients give permission for storing and processing their data, trust plays a very important role because companies today must do everything in their power to protect data and verify its accuracy. An example of the importance of data verification is reflected in the process of performing online banking transactions, where it is extremely important to check who forwards the request for the transaction, and in order to prevent potential fraud.

Data visibility (Visibility) implies the observability of data, and it is of crucial importance that they are not free accessible to systems and processes, but only to those who have permission to manage them. This dimension additionally emphasizes the importance of data security in information systems.

3.2. Big data concepts

The most important concepts [18] of Big Data domeins are:

- infrastructure,
- storage of large data sets,
- data extraction methods,
- technologies and software for preprocessing, set preparation, analysis and visualization, using systems for business and artificial intelligence, with the aim of making decisions based on significant information extracted from data.

Big Data infrastructure includes data sources, data collection tools and systems, physical data storage medium (servers), data transmission network, data analysis software tools and their working environment, data backup infrastructure, as well as supporting software and results processing.

The basic tasks of Big Data infrastructure include: storage, search, sharing, transfer and analysis of data [20]. Data is stored in databases, found and called from them, and according to the type of data, required security of transactions and types of requested queries, we differentiate:

- relational SQL databases, designed primarily for reliable transactions and ad hoc queries and
- NoSQL databases, which enable: the storage of diverse data, greater speed and flexibility in work. NoSQL databases are non-relational, distributed, open source, horizontally scalable databases. Many of them were developed by companies that tried to find more convenient ways to store or process data for large websites or social networks. There are various solutions, and some of them are: MongoDB, Amazon's Dynamo, Cassandra, CouchDB, Redis, Hbase, BigTable, Hypertable, Voldemort, Riak, ZooKeeper, etc. [21].

3.3. Big data analytics

Big data analytics is a complex process of examining large and diverse data sets, which is carried out with the aim of discovering important information, useful knowledge, hidden patterns, unknown correlations, all in the function of making the best decisions. The three basic phases of Big Data analytics are:

- collecting,
- analysis,
- data visualization.

The data collection phase involves the identification and filtering of relevant data. The collected data is mostly unorganized and inconsistent, contains errors and may be only partially incomplete. The process of collecting raw data from various sources, creating, processing, standardizing, creating an integrated set for analysis is realized within the pre-processing phase and is a complex task. The success of the mentioned phase affects the efficiency of the subsequent application of appropriate methods for detecting legality in the data.

The phase of data analysis involves finding correlations in the data, connecting them in order to discover certain regularities that are not visible at first glance. Data mining methods are used to carry out data analytics processes. Data mining is the process of discovering patterns in large sets that includes machine learning methods, statistics and the use of database systems.

The third phase involves visualization, the method of displaying and analyzing the results of data, in a form suitable for understanding and interpretation. Data are most often displayed in the form of bars, histograms, pie charts, dot charts, Gantt charts, infographics, and the like. The Big Data concept relies on visualization, which is part of Big Data analytics and refers to the graphical, i.e. visual display of the results of processing a large amount of data for easier and more accurate interpretation. It is believed that adequate visualization of big data, in a way that is adapted to the users of that data, will be one of the most important factors in the further development of this technology.

Considering the scale and heterogeneity, the analysis of big data sets is a challenging and complex task. Data integration plays a key role, whereby, differences in data structure and semantics are expressed into a format that can be represented in a computer system.

Considering diversity, variability, occurrence of missing values and redundancy, big data query processing is significantly different from the traditional approach. The existence of redundancy

is reflected in the interconnectedness of big data for solving the problem of missing data, checking conflicting cases and identifying hidden relationships. The lack of connectivity of NoSQL database systems with analytical packages for analysis is a problem associated with big dataa.

4. CASE STUDY

The use of analytics on large data sets extracted from educational environments is a relatively new area of research. The main goal of the conducted research was focused on the analysis and visualization of data generated in the environment of the system for online preparation of the entrance exam in the period from 2017 to 2022.

System for online preparation of the entrance exam at the Department of Electrical Engineering and Computer Science, (hereinafter Odsek VIŠER)[22] is based on the principles of three-layer MVC (Model-View-Controller) architecture and integrated within the UPIS section of the official Internet presentation, which is intended to inform future students. Access data and test attempts were stored in a MySQL database in the *test_reception* table, which reached a size of 15.6MiB with 146848 records. Executing queries and generating responses over the *test_reception* table is significantly slowed down, due to its size and number of records.

For the purposes of the described case study, in the data collection phase, the MySQL relational database was migrated to MongoDB [23] document oriented database. MongoDB is an open source NoSQL database written in C++. Unlike relational databases, MongoDB provides weak guarantees of consistency, but that is why it's characteric is better performance in the management and analysis of large data compared to traditional relational models. [24]. Data migration was performed by establishing a connection with the MySQL system using 3T studio GUI for MongoDB. [25]. Data from table *test_prijemni* are migrated into collection *tie_22_analiza*ni, and the relational record format is transformed into a JSON document.

The structure of the JSON document is defined based on the schema of the table *test_prijemni*. The MongoDB system generated a unique identifier for each document*, field _id*, ObjectId of BSON data type. Field *_id* uniquely identifies each collection document, is always 12 bytes long, and contains timestamp information (timestamp), client machine ID, client process ID, 3-byte increment counter. An example of a JSON document is given below.

{"_id"**:**ObjectId**(**"62d0bc4a92d3d4646aa1c668"**),**
"id" ***:*** *NumberInt**(255)**,*
*"broj_pokusaja_matematika"**:** NumberInt**(4)**,*
*"broj_pokusaj_oit"**:** NumberInt**(9)**,*
*"godina"**:** NumberInt**(2017)**,*
*"IP"**:** "93.87.148.30"**,***
*"grad"**:** "Београд"**,***
*"skola"**:** "Електротехничка школа \" Никола Тесла\""*
}

Insight into the contents of the collection was made by quiry *db.tie_22_analiza.find({})* with implemented method find().

By displaying the contents of the collection *tie_22_analiza* it was determined that there were irregularities in the data and missing values for the city and school fields, which were manually filled in by the pupils who were tested. The identified irregularities were eliminated by implementing the data preprocessing process. A method of eliminating documents was implemented in which the occurrence of missing values in important fields (city, school) was determined. Irregularities in the data (typing errors in the name of the school and city, additional empty spaces, etc.) were eliminated by applying data cleaning methods. The data in the fields id, number_of_attempt_mathematics, number_of_attempt_oIT, year, IP were automatically generated after each realized test attempt, so that no errors and irregularities were found in them.

In the analysis phase, the pipeline processing procedure was implemented using queries with the aggregation method. In the first step of the pipeline processing, the $match aggregation operator is included to extract documents based on criteria for a specific year. The extracted documents are passed to the next aggregation step where the grouping operator $group is applied. Grouping of data was done by the IP address of the computer from which it was accessed, and then by city and school. In this step, the cumulative aggregation operator $sum is also defined in order to summarize the access from one IP address. The grouped data was forwarded to the next step, where another grouping was performed with the aim of structuring the data in the documents and summarizing access by school. In the last step of the pipeline processing, the operator $out was applied to save the results by year in separate collections. Figure 2 shows the query used to create document collections by year.

```
db.tie_22_analiza.aggregate([{$match:{godina:2022}},
{$group:{_id:{IP:"$IP",grad:"$grad",skola:"$skola"},
    broj:{"$sum":1}}},
 {$group:{_id:"$_id.grad",
    podaci:{"$push":{skola:"$_id.skola",
                     IP:"$_id.IP",
                     brojPristupa:"$broj"
                     }
            },
          brojPristupa:{$sum:"$broj"}
        }
},
{$out:"analiza_2022"}
])
```

Figure 2. Query to create collections by year

After the analysis, the third phase, visualization, was realized. The visual display of data for each year included the display of the number of accesses by city and school. The analytical software Microsoft Power BI (desktop version) was used for data visualization. Generated collections were exported from MongoDB and imported into the Power BI environment.

Graphic interpretation of the data was performed using grouped bar and grouped bar histograms.

Figure 3 shows the visualization of the number of accesses by city.

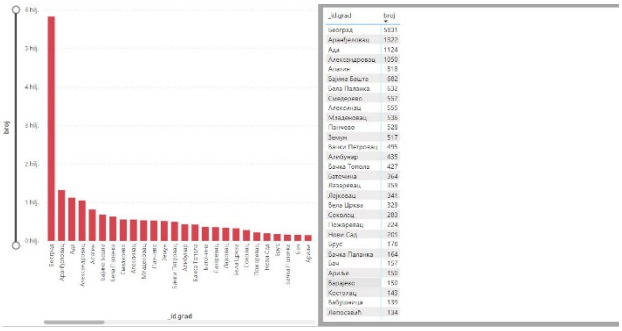

Figure 3. *Number of accesses per cities for the year 2022.*

Figure 4 shows the visualization of accesses by city and school.

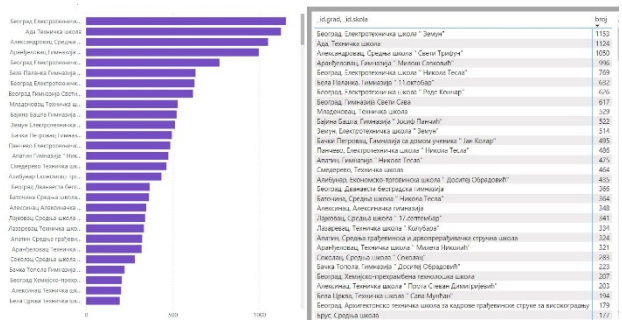

Figure 4. *Number of accesses per cities and per schools for the year 2022.*

Both graphical interpretations are for year 2022.

4. 1. Results and discussion

In this paper, we have presented a study that investigateed educational system for online education using the big data analytics.The system is primarily designed for the higher education students, in order to improve their interest in studying at the Department VIŠER of The Academy of Technical and Art Studies. The volume and structure of the stored data exceeded the capabilities of the relational model, which made the process of extracting data for analysis more complicated. It has been made migration and transformation of data from relation to non-relation database. Queries and aggregation framework implemented in MongoDB document – oriented database management system were used for processing large – scale data. Visualization of the processed data was realized using the Microsoft Power BI tool.

Figure 3 shows that the highest number of accesses was recorded by the students of Belgrade secondary schools. A slightly smaller number of accesses, but not insignificant, was recorded from the students from: Aranđelovac, Ada, Aleksandrovac. On Figure 4, first rows in table legend indicate that in addition to the expected interest of the students from technical schools, the Department VIŠER is also a choice for the students from other secondary schools.

The related studies mentioned in the second section were made to discover architecture and environment for processing a large amount of educational data in different ways, by using different methods. Proposed system for analyzing factors affecting students performance and dropout by Alblawi and Alhamed [7] is based on implementation Apache Spark analytics engine and HBase column-oriented non-relational database management system. Machova, Komarkova and Lnenicka [8] described methodology for processing Moodle log files, which store information about all users, no metter what role is the user's role (the instructor, course editor, student) and their mutual interactions. They proposed model in the cloud with the use of the Apache Hadoop cluster. Swathi et al. [9] proposed framework for data analysis and student's performance in education system using map reduce algorithm for processing large–scale data stored in Hadoop Distributed File System (HDFS). Based on comperative investigation of Education Big Data Analysis System architectures, Chen et al. [10] proposed framework for analysis and educational data model to achieve better performace in education process. Michalik et al. [11] have been focused on determining and identifying data sources from education enviroment. Authors have proposed methodology for work with large–scale education data, based on environment arhitecture from the technical perspective. Klašnja-Milićević et al. [12] have developed framework for processing and analyzing education data, based on Apache Hadoop architecture which contain: Map – Reduce programming model, Hive non-relational database management system, Pig client – side application in Java and machine learning algorithms as Mahout library.

In the related studies mentioned in the second section, a big educational data processing framework, based solely on the Apache Hadoop architecture, was proposed, but it did not include any visualization process. The authors had analyzed the data extracted from the educational process of teaching and learning, but they had not analyzed data using variables, like city and secondary schools of potential prospective students. In this paper, a study was conducted with

the intention of measuring these data using the proposed framework, based on the implementation of the MongoDB database management system integrated with the Microsoft Power BI visualization tool.

The Big Educational Data analytics framework presented in this paper includes:

- Data set extraction tool and techniques,
- aggregation enviroment for querying data,
- management tool for visualization of results.

Figure 5 provides a indication of all phases proposed Big Educational Data analytics framework.

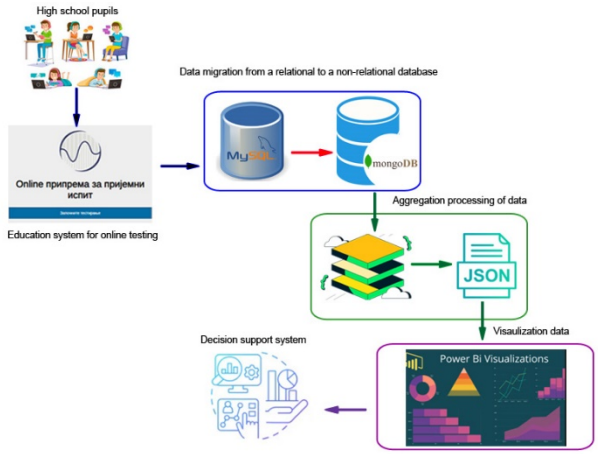

Figure 5. *Model of Big Educational Data analytics framework.*

We have developed proposed model according to the essesential aspects for the implementation of big data analytics process in the domen of education: extracting data from data source, programming environment, and presentation results to the end user. In this paper, we have described the big data concept, infrastructure, techniques, and analytical process, that can be used for achieveing better performance of educational systems.

We would like to highlight the importance of the concept of big data and the implementation of analytical processes in education, as for the improvement of the teaching and learning process, but also for profiling basic demographic data about potential future students. This research also points to the development of a framework that can support data sets created by integrating multiple different educational sources, in addition to developing the key presentation and performance techniques for the collection and processing of large amounts of data.

The limitation and obstacle to the conducted research is the lack of appropriate technical infrastructure for the realization of the theoretical concept of big data analytics process. The expansion of the proposed framework can be realized by horizontal scaling, i.e. by adding additional clusters for storing and processing large - scale data. A distributed environment would enable a more flexible and faster integration of several different data sources, and thus the possibility of monitoring basic data about: students, the learning process, achieved and unrealized results, from their enrollment to the end of their studies.

5. CONCLUSION

Big data analytics undoubtedly occupies a significant role in the future of education.

The paper presents research focused on the application of the process of big data analytics to data extracted from the educational system for online testing. Due to the large amount of data stored in the MySQL database of the system, the procedure of data migration to a non-relational system was implemented.

The process of big data analytics was realized through three phases. Data sets were generated by year and the necessary groupings were performed, in order to prepare for the graphical interpretation of the data. By identifying the unique IP addresses of the computers, the visual access involved displaying the number of accesses by cities and schools.

The results of the analysis provided a better insight into the interest of secondary school students in studying at the Department VIŠER helped in making the right decisions, in order to improve interest where the number of accesses was lower.

In the further continuation of the research, it is planned to apply the big data concept to data extracted from several educational systems (learning management system, information system), with the aim of achieving better efficiency of the educational process.

REFERENCES

[1] Horton, W., Horton, K. (2013). *E-learning tools and technologies*. *A consumer's guide for trainers, teachers, educators, and instructional designers*. Wiley Publishing Inc.

[2] Ip, RH., Ang, LM., Seng, KP., Broster, JC., Pratley, JE. (2018, August). *Big data and machine learning for crop protection*. Computers and Electronics in Agriculture. 151, 376-83. doi: 10.1016/j.compag.2018.06.008.

[3] Seng, KP., Ang, LM., Ooi, CS.(2016, Jul). *A combined rule-based & machine learning audio-visual emotion recognition approach.* IEEE Transactions on Affective Computing, 9(1), 3-13.doi:10.1109/TAFFC.2016.2588488.

[4] James, M. (2011, May). *Big Data: The next frontier for innovation, competition and productivity.* McKinsey Global Institute. https://www.mckinsey.com/business-

functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation, (accessed, 2022, Jun).

[5] Mostow, J., Beck, J. (2006, Jun). *Some useful tactics to modify, map and mine data from intelligent tutors.* Natural Language Engineering, 12(2), 195-208.

[6] Rice, W.H. (2006). *Moodle e-learning course development. A complete guide to succeful learning using Moodle*. Packt publishing.

[7] Alblawi, A.S., Alhamed, A.A. (2017, Nov 16). Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, NLP and analytics. In2017 IEEE conference on big data and analytics (ICBDA), 124-129.

[8] Machova, R., Komarkova, J., Lnenicka, M. (2016, Oct 10). *Processing of big educational data in the cloud using apache Hadoop*. 2016 International Conference on Information Society (i-Society), 46 – 49.

[9] Swathi, R., et al. (2017, Jun 15). *Systematic approach on big data analytics in education systems.* 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), *420-423.*

[10] Chen, Jinhua, et al. (2017, Mar 10). *Research on architecture of education big data analysis system.* 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), 601-605.

[11] Michalik, P., Štofa, J., Zolotova, I., (2014). *Concept definition for Big Data architecture in the education system.* In2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI), 331-334.

[12] Klašnja-Milićević, A., Ivanović, M., Budimac, Z., (2017, Nov). *Data science in education: Big data and learning analytics.* Computer Applications in Engineering Education, 25(6), 1066-78.

[13] Muhammad, R.N., Tasmin, R., Aziati, A.N., (2020, Apr). *Sustainable competitive advantage of big data analytics in higher education sector: An* Overview. InJournal of Physics: Conference Series, 1529(4), 042100.

[14] *Diebold, F.X.,* (2012). *On the Origin (s) and Development of the Term'Big Data.*

[15] Diebold, F.X., (2003, Feb). *Big data dynamic factor models for macroeconomic measurement and forecasting.* InAdvances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society, 115-122.

[16] EDPS (2015). *Meeting the challenges of big data. European Data Protection Supervisor (EDPS).* Retrieved April 14, 2017.

[17] Erl, T., Khattak, W., Buhler, P., (2015). *Big Data Fundamentals Concepts, Drivers and* Techniques. Prentice Hall.

[18] Yu, S., Guo, S.,(2016). *Big Data Concepts, Theories, and Applications*. Springer.

[19] https://www.ravenpack.com/blog/quant-investing-big-data-analytics, (accessed, 2022, July).

[20] Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A., & De Laat, C. (2012, December). *Addressing big data challenges for scientific data infrastructure.* In4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, 614-617.

[21] Moniruzzaman, A. B. M., & Hossain, S. A. (2013). *Nosql database: New era of databases for big data analytics-classification, characteristics and comparison.* arXiv preprint arXiv:1307.0191.

[22] Prokin, D., Čoko, D., Dimić, G., Đenić, S., Savić, A., Bogojević, B., (2016). *Online aplikacija za testiranje učenika za pripremu prijemnog ispita.* Proc. INFOTEH, Jahorina, Bosnia and Herzegovina, 15, 710-713.

[23] Shannon, B., Eoin, B., Kristina, C., (2109, December). *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage.* O'Reilly Media; 3rd edition.

[24] Li, Y., & Manoharan, S. (2013, August). *A performance comparison of SQL and NoSQL databases.* In 2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 15-19.

[25] *https://studio3t.com/*