Learning analytics for higher education: proposal of big data ingestion architecture

Meseret Yihun Amare[1,*], and *Stanislava* Simonova[1]

[1]University of Pardubice, Faculty of Economics and Administration, Institute of System Engineering and Informatics, Studentska 84, 532 10 Pardubice, Czech Republic

Abstract.

Research background: Higher education institutions are generating multiple formats of data from diverse sources across the globe. The data ingestion layer is responsible for collecting data and transform for analysis. Learning analytics plays a vital role in providing decision-making support and selection of suitable timely intervention. The lack of tailored big-data ingestion architectures for academics led to several implementation challenges.
Purpose of the article: The purpose of this article is to propose data ingestion architecture enabled for big data learning analytics.
Methods: The study reviews existing literature to examine big-data ingestion tools and frameworks; and identify big-data ingestion challenges. An optimized framework for the real world learning analytics application was not yet in place at global higher educations. Consequently, the big-data ingestion pipeline is experiencing challenges of inefficient and complex data access, slow processing time, and security issues associated with transferring data to the system. The proposed data ingestion architecture is based on review of recent literature and adapts best international practices, guidelines, and techniques to meet the demand of current big-data ingestion issues.
Findings & value added: This study identifies the current global challenges in implementing learning analytics projects. Review of recent big data ingestion techniques has been done based on defined metrics tuned for learning analytics purposes. The proposed data ingestion framework would increase the effectiveness of collecting, importing, processing and storing of learning data. Besides, the proposed architecture contributes to the construction of full-fledged big-data learning analytics ecosystem of higher educations.

Keywords: *big data architecture; data ingestion; learning analytics; globalization; higher education*

JEL Classification: *D80; I23; M15; O30*

* Corresponding author: yihunm@gmail.com

1 Introduction

In this era of globalization, large amounts of data generated in every area of our lives due to the rapid development of new technologies such as the Internet, social media, Internet of Things (IoT), cloud, smart and mobile devices. As a result, higher education institutions are generating multiple formats of data from diverse sources across the globe. The volume, variety, and velocity of data generated daily lead to the phenomenon of big data with the potential to further improve the values of products and services in different industries [1]. However, the use of analytics in an academic institution is in its infancy [2] in comparison with other business sectors, and the potential for data analytics to impact higher education is growing. There is a growing interest in data analytics at higher education [3] to improve the performance of students; to predict enrolment forecasts; to detect early dropouts and provide targeted interventions to help them remain in the university system, and effectively utilize academic resources [4, 5]. Besides, universities can benefit from big data to enhance the effectiveness of academic faculty and reduce administrative workload [6].

According to [7], learning analytics is the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. The aim of learning analytics(LA) is to evaluate student‘s behaviour in the context of teaching and learning, further to analyse and interpret it to gain new insights, and to provide the stakeholders with new models for improving teaching, learning, effective organization, and decision making [8]. Learning analytics plays a vital role in decision making support and selection of suitable timely intervention.

Big data technologies comprise of architectures and technologies to extract valuable information from large volumes of different data sources. Big data architecture is an overarching system used to ingest and process enormous amounts of data [9].

In recent years, more universities start to use learning analytics to obtain findings on the academic progress of students, predict future behaviours, and recognize potential problems in an early stage [10] using the traditional database analytics, not on big data. In the context of big data in education, some specific big data architectures or framework has is proposed for education [11]. However, there are still limitations in adopting big data analytics architecture for enterprises as current frameworks provide generic architecture for big data analytics [12]. These frameworks do not give a detailed learning analytics process for higher education.

The data ingestion layer is responsible for the collection of data and transforming for analysis. This article aimed to propose data ingestion architecture tailored to big data learning analytics. The paper tries to find out challenges in data ingestion and develop data ingestion architecture for learning analytics based on guidelines and best practices in the area of big data and then are tailored for academic institutions. The contribution of the paper is a new view of the designing ingestion methods that use currently available tools to ingest both batch and stream files. Besides, the paper contributes to the development of full-fledged learning analytics architecture for higher education institutions and future researches in the area of big data analytics. The paper is organized into four sections.

The first section provides the introduction of big data and learning analytics and highlights the current state of the issue in the form of a literature review. The second section describes the methods used in this study. The third section provides the results of the study, briefly describes data ingestion challenges, provides best practices when designing big data ingestion architecture, and introduces the proposed big data ingestion architecture for learning analytics based on identified components. The discussion section describes the main points from the proposed architecture, discusses the scope of the study, and highlights the next direction.

1.1 Data collection at higher education

Data ingestion is the first step in data management systems. As the strategic and modern approach to designing the data pipeline ultimately drives business value, data analysts, managers, and decision-makers need to understand data ingestion and its associated technologies. According to [13], Data Ingestion is "the transportation of data from variety sources and transferring to a storage medium where it can be accessed, used, and analysed by an organization." The destination is typically a data warehouse, data mart, database, or a document store. The data ingestion processes include manual, semi-automatic, and automatic methods [14]. Data ingestion helps to bring various types of data sources from its source into a system where it is easy to be analysed and stored.

Data ingestion is critical and should be emphasized for any big data project, as the volume of data is becoming very large [15]. Data handling is always a challenge and critical activity if the amount of data becomes huge and consists of various formats. As big data systems are designed to process unstructured or semi-structured data, it becomes complex to capture data from different sources. Data ingestion is generally becoming complex in data systems as data sources and processing now includes batch and stream formats which, increase the complexity and management. Besides, the ever-increasing IoT devices are resulting in a large volume of various data sources. Hence, extracting data using traditional data ingestion approaches becomes a challenge [16]. The following section explores the challenges of data ingestion for university big data systems.

1.2 Big data ingestion challenges

The absence of a complete big data architecture framework tailored for higher education institutions that serve as a guideline for an overarching process is one of the existing challenges of implementing big data analytics at higher education [4]. Existing architecture does not give a detailed learning analytics process of big data in higher education. There are no designated specific tools or methodologies for gathering, cleansing using captured data [17].

The traditional standards of data architecture are changing at an ever-increasing rate. Preceding enterprise architectures are undergoing significant technological changes in the face of new trends, including big data, non-relational data stores, IoT, machine learning, and artificial intelligence, data lakes, and many others. The following are the key challenges that can impact data ingestion and pipeline performances [1, 16]:

- Data quality: Data quality is a challenge when we are working with diverse data sources. Inconsistent data formats, data repetition, and missing values would make analysis unreliable. Thus, bringing the data together should be done after the data is appropriately analysed and prepared.
- Slow processes: Writing codes to ingest data and manually creating mappings for extracting, cleaning, and loading data can be cumbersome as data today has grown in volume and has become highly diversified. Therefore, there is a move towards data ingestion automation. The old methods of ingesting data are not quick enough to persevere with the large volume and variety of data sources.
- High complexity: The ever-increasing of new data sources and IoT, academic institutions are confronting challenges to make data integration in order to insights value from their data. The main challenge is the ability to attach to that data source, recognizing and error elimination, and inconsistent data structures.
- Cost: The infrastructure that enables the data acquisition process from the sources and the cost of associated ingestion tools makes the data ingestion task very costly.

- Scaling- The overall performance may decrease if the issue is not addressed correctly during the planning phases of building the architecture.
- Unreliability: Incorrectly ingesting data could result in unpredictable connectivity. This further can disrupt communication and cause loss of data.
- Security: Data security is the biggest challenge that could occur when the data moved from one source to the storage system since data are staged in numerous phases throughout the ingestion process. Consequently, it is challenging to accomplish security standards during the data ingestion process. Data ingestion can compromise compliance and data security regulations, creating further complexity and cost. It requires to introduce advanced security techniques of data encryption and anonymous access to student's sensitive information.

1.3 Data sources for learning analytics

Currently, the big data processing systems are capable of processing huge amounts of a batch, streaming, and real-time. The data ingestion can become a bottleneck or can break a system if not designed according to the requirement [13].

Higher education is producing an enormous volume of data continuously from a variety of sources. Data in the university originates from numerous sources with different formats. These include relational databases, card swipes, student information systems (SIS), company servers, third-party data providers, etc.

According to [18], big data in the academic environment consists of data from courses, modules, experiments learning management systems (LMS), and data coming from the students throughout the education process. In addition, social networks, multimedia, IoT sources, academic records and profiles, demographic characteristics of students, are all sources of data in higher education. Higher education institutions use data drawn from many sources and devices to design and deliver their services, allocate resources, and monitor their performance [19].

2 Methodology

The study reviews existing literature to examine big-data ingestion challenges, adoption guidelines, and evaluating generic architectures. The paper also identified possible big data sources inside academic institutions. This design of data ingestion architecture for learning analytics was done based on the types of devices available for data collection and the different sources of data generated inside academic institutions. For the purposes of this paper, we analyzed studies as a result of which is a big data ingestion architecture tailored for learning analytics has been proposed based on the lambda architecture[20]. The proposed data ingestion architecture was based on a review of recent literature and adapts best practices, guidelines, and techniques to meet the demand of current big-data ingestion issues.

3 Results

By analysing the available literature it was possible to explore data ingestion challenges, identify university data sources, and develop ingestion mechanisms. This proposal of data ingestion architecture was based on the guidelines and best practices of general big data architectures from works of literature and they are adopted based on guidelines. In order to cope up with the challenges that the data ingestion process is facing, the following best practices were incorporated in the process of developing the big data ingestion architecture:

- Design for performance: In order to benefit from big data at higher education, it's necessary to design an efficient mechanism to ingest available data from different sources and prepare for learning analytics. As recommended by [21], the proposed architecture was designed to process big data by dividing large data sets to allow it to run independently in parallel to satisfy high-performance computing and work efficiently without having to worry about intra-cluster complexities, monitoring of tasks, node failure management.
- Ingestion automation: As the data in the university continues to grow both in volume and complexity, it's wouldn't be possible to curate a huge amount of data using manual techniques. Automating the ingestion process would shorten the time takes for ingestion, increases productivity, and reduces manual efforts. The learning analytics needs several new data sources to be ingested on-demand with minimal user intervention.
- Real-time data ingestion: Some of the university data is time-sensitive and needs to be processed as soon as they are collected. Data is extracted, processed, and stored as soon as it is generated for real-time decision-making.
- Batch ingestion: There are requirements where academic data needs to move at regularly scheduled intervals. This type of ingestion is suitable for repeatable processes.
- The proposed architecture combines the batch and real-time ingestions to utilize batch processing to offer broad views of batch data and real-time processing.

As several tools are emerging, the selection of ingestion tools needs to be done based on the parameters. According to [22], a perfect tool does not exist. Organizations combine different tools for a better solution. Besides, the performance of data ingestion and throughout needs to be verified how fast the system can consume data from the various data source. This helps the user to know information about speed, the number of processed files per minute, the acceptable size for the files for the chosen tool. The following are indicators used to compare ingestion tools:

- Data type: The type of data available for ingestion varies considerably. Thus, the choice of data ingestion depends on the available data types to be used for learning analytics.
- Speed: The result of analytics is based on both real-time and batch data processing. The ingestion processes of these data should be done efficiently to satisfy the needs of students.
- Reliability: Data collection and shipping needs to be reliable by avoiding potential data losses in the process.
- Delivery: To ensure that data are complete after the ingestion process, it should be tested that data from the source matches with the destination.
- Files size: Tools need to be identified based its capability data processing per second as the file size in the university would grow over time.
- Scalability: According to [23], when managing a successful expanding application, the ability to scale becomes a critical need.
- Durability: When the data ingestion becomes out-of-date, it's necessary to make sure that messages are not lost by making sure that the queue is durable.

3.1 Proposed ingestion architecture

A big data architecture enabled for learning analytics framework is proposed as follows. The aim of this ingestion architecture is to attempt to seamlessly integrate the generation and ingestion of cleaned and secured data. It is designed to handle massive quantities of data by taking advantage batch and stream-processing layers of the lambda architecture. Besides, the proposed architecture is designed to utilize technologies that do not necessarily

need a separate batch or streaming layer as it allows attributes of both layers into single layer.

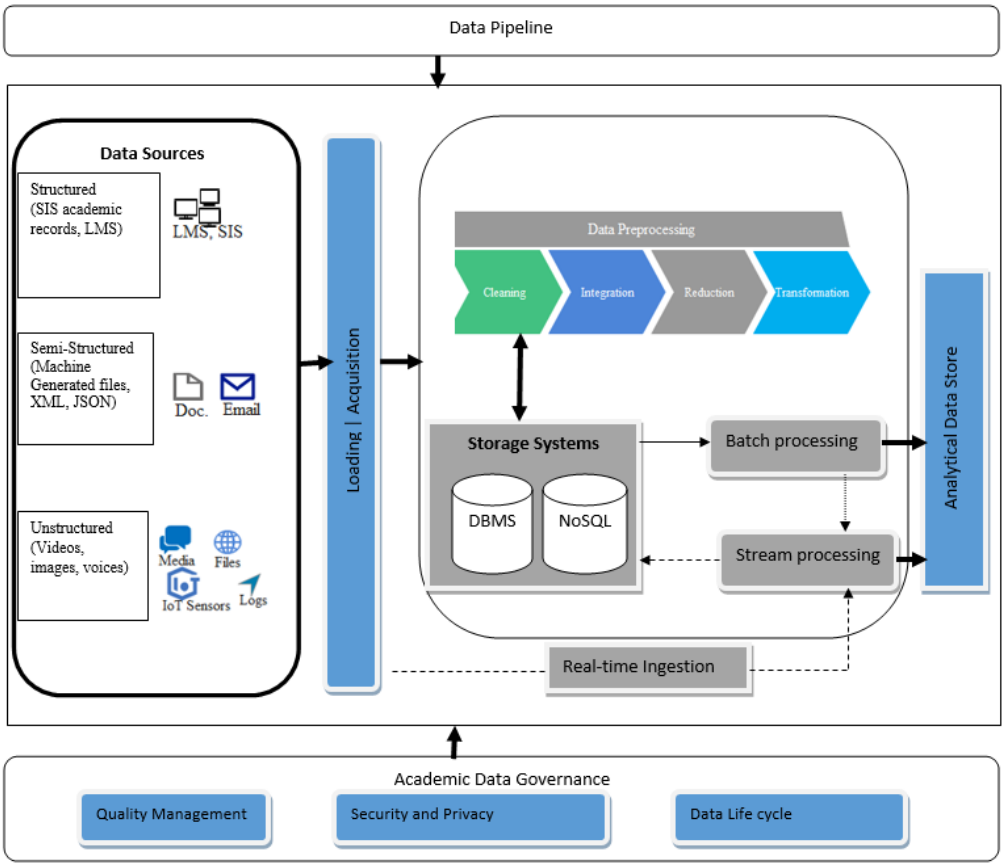

Fig. 1. The proposed big data ingestion architecture for learning analytics.

3.2 Data ingestion components

Big data architecture varies based on a company's infrastructure and needs. The university big data architecture contains the following components:

3.2.1 Data sources

Various data collection devices such as students' cards, social networks, learning management systems (LMS), sensors, and student information systems (SIS) will serve as the data source.

The structured, semi-structured and unstructured data generated from individual students are passed over the data management systems for analysis [4]. Data sources include data from real-time sources such as IoT devices and static files generated through the process of teaching and learning. This component is characterized by accommodating dynamically increasing number of data generating systems and can interact with heterogeneous type of data.

The data sources needed to provide academic decision making are divided into structured data such as traditional academic records; semi-structured data such as logs generated when students participate on discussion forums(XML, JSON, CSV files); and unstructured data such as lecture video, audio, images, PDF files and other documents , files from Social Media platforms and IoT. These data are collected from various locations inside the university premises, and will be stored immediately into appropriate databases, depending on the content format.

3.2.2 Data acquisition

This is the first component from which data from numerous sources begins its next journey. The primary goal of data acquisition is to read data provided in various communication channels, frequencies, sizes, and formats. This layer takes care of categorizing the data for the smooth flow of data into the further layers of the architecture. For the purpose of loading REST APIs or streaming tools can be used.

3.2.3 Pre-processing

This component is capable of cleansing input data, transforming data into an analysis-ready format, and integration-services [24]. The preparation activity is where the transformation portion of the Extract Transfer Load (ETL) cycle is performed, although analytics activity will also likely perform advanced parts of the transformation. Tasks performed by this activity include data validation, cleaning, outlier removal, standardization, reformatting, or encapsulating. Verification or attachment may include optimization of data through manipulations and indexing to optimize the analytics process. This activity aggregates data from different Data Providers, leveraging metadata keys to create an expanded and enhanced dataset.

To make the process easier, data pre-processing is divided into four stages: data cleaning, data integration, data reduction, and data transformation.

- Data cleaning: Data cleaning refers to techniques to clean data by removing outliers, replacing missing values, smoothing noisy data, and correcting inconsistent data.
- Data integration: As the academic data is collected from multiple sources, data integration is a vital part of the process. The following are the most common activities to integrate data: (1). physically bringing the data all to one data store. This usually involves Data Warehousing. (2). Copying data from one location to another using application. It could be synchronous or asynchronous and is event-driven. (3). Virtualization of data using an interface to provide a real-time and unified view of data from multiple sources. The data can be viewed from a single point of access.
- Data reduction: The purpose of data reduction is to have a condensed representation of the data set which is smaller in volume, while maintaining the integrity of original data.
- Data transformation: Transforming the data into form appropriate for Data Modelling is the final step of data pre-processing. Some of the strategies that enable data transformation include: Smoothing, attribute construction, aggregation, normalization, discretization, and concept hierarchy generation for nominal data.

The transformation engine is capable of moving, cleaning, splitting, translating, merging, sorting, and validating data. For example, structured data such as that typically resided in students record might be extracted from student's information systems (SIS) and subsequently converted into a specific standard data format, sorted by the specified and then the record validated against data quality rules.

3.2.4 Real-time ingestion

Since the university data includes real-time sources, the architecture includes a way to capture and store real-time messages for stream processing. This is can be a simple data store where incoming messages are dropped into a folder for processing or a message ingestion store to act as a buffer for messages to support scale-out processing, reliable delivery, and other message queuing semantics.

3.2.5 Stream processing

After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis. The processed stream data is then written to an output sink. The real-time data can be queried using suitable analytical tools in later stage. The stream computing should support high performance real-time or near real time processing. Since real-time processing involves no grouping at all, data is sourced, manipulated, and loaded as soon as it's created or recognized by the data ingestion layer.

3.2.6 Batch processing

Since the university data sets are large, it's necessary to process data files using long-running batch jobs to filter, aggregate, and prepare the data for analysis. Usually these jobs involve reading source files, processing them, and writing the output to new files. The data ingestion layer collects and groups source data and sends it to the destination system periodically. Groups may be processed based on any logical ordering, the activation of certain conditions, or a simple schedule.

3.2.7 Storage system

The data storage system consists of big database management system with all capabilities like buffering and real-time query optimization. This phase is also responsible for data pre-processing and data-cleaning. In other words, the main functionality of the data storage and management system is to process and convert raw data into a form that can be very efficiently processed by the analytics engine [4]. The data storage principles are based on compliance regulations, data governance policies and access controls. The data storage methods can be implemented and completed in batch processes or in real time. This component is receives data from the various data sources and stores it in the most appropriate manner. Batch processing data is generally stored in a distributed file storage systems that are capable of storing high volume data in different formats. Other structured data can be stored using RDBMS while unstructured data would be stored in NoSQL databases.

3.2.8 Data governance

Data quality and security is the pillar of the proposed learning analytics ingestion architecture. This component consists of data quality management, data life-cycle management, and data security and privacy management that emphasize how to harness data in the organization [25].

The data quality management can be viewed as the processes, governance, policies, standards, and tools for managing data. Data is quality is regularly monitored for

completeness, accuracy, and availability to support the learning analytics process and decision making.

The life-cycle management is includes archiving, data warehouse maintenance, testing, disposal and removal of data throughout the entire data analytics process.

The security and privacy management of big learning data includes all the platforms for providing institution-wise activities of data discovery, configuration assessment, monitoring, auditing, and protection. It is essential to implement high-level data control mechanisms to preserve students' personal data privacy. In addition, it's necessary to adopt organizational policies, standards, and compliance requirements in line with the existing data privacy regulations.

3.2.9 Data pipeline

In order to continuously gain insights from academic big data, the ingestion and pre-processing, storing, loading to an analytical data store, and generating insights needs to be automated in a data pipeline.

4. Discussion

The proposed architecture is technology agnostic and it can use available tools for collecting batch and stream files for analytics based on the requirements and institutions affordability. The architecture is designed in such a way to accommodate all types of data generated in academic institutions.

The architecture presented in this paper has the capability to gather, pre-process, clean various types of higher education data for learning analytics. Improvement over efficiency can be achieved batch and real-time data are integrated. The data governance system is included in the architecture to ensure the security and privacy of data is maintained; preserve data quality and manage the data life-cycle throughout the ingestion system. The entire ingestion process is connected to a data pipeline to automate the processes from the source to analytical data stores.

The paper focused on designing the ingestion architecture of learning analytics system and further stages of the architecture were not included. It's recommended to introduce a testing mechanism to verify that the data adequately extracted from multiple sources are correctly loaded to the system.

5. Conclusion

Properly designed data ingestion mechanism can help academic institutions utilize data to support decision-making and improving teaching learning performance. It reduces the complexity of bringing data from multiple sources together and allows analytics on various data types and schema across the university. The proposed big data ingestion architecture for learning analytics in academic institutions has been proposed by identifying big data sources in the academic environment, identify main components of data ingestion, analysing existing challenges, following best practices and guidelines, and tailoring for academic institutions.

The architecture comprises of a data pipeline for automation of ingestion process, data ingestion techniques and storage for analytical data that could enhance the data ingestion process by making it seamless, secure, and guaranteed data quality for decision making. The paper can contribute to further studies and construction of full-fledged big data learning analytics architecture for higher educations.

This article was supported by the grant No: SGS_2020_18 of the Student Grant Competition.

References

1. Ip, R., Ang, L., Seng, K., Broster, C., Pratley, E. (2020). Big educational data & analytics: Survey, architecture and challenges. *IEEE access*, 8, 116392-116414.
2. Kwon, Y. O. (2013). Data analytics in education: Current and future directions. *Journal of Intelligence and Information Systems*, 19(2), 87–99.
3. Daniel, B. (2014). Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, 46(5), 1-17.
4. Matsebula, F., Mnkandla, E. (2017). A big data architecture for learning analytics in higher education. In D. R. Cornish (Ed.), *2017 IEEE AFRICON Conference* (pp. 951-956). Cape Town: IEEE.
5. Shacklock, X. (2016). *From bricks to clicks: The potential of data and analytics in higher education*. London: Higher Education Commission.
6. Jha, S., Jha, M., O'Brien, L. (2018). A step towards big data architecture for higher education analytics. In *5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)* (pp.178-183). Nadi: IEEE.
7. Viberg, O., Hatakka, M., Balter, O., Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98-110.
8. Siemens, G., Long, P. (2011). Penetrating the fog: analytics in learning and education. *Educause Review,* 46(5), 30–40.
9. Alley, G. (2019, March 4). *What is big data architecture? Big data zone*. Retrieved from : https://dzone.com/articles/what-is-big-data-architecture
10. Leitner, P., Khalil, M., Ebner, M. (2017*). Learning analytics in higher education - A literature review*. Switzerland: Springer International Publishing AG.
11. Ang, K. L, Ge, F. L., Seng, K. P. (2020). Big educational data & analytics: Survey, architecture and challenges. *IEEE Access*, 8(1), 116392-116414.
12. Gong, Y., Janssen, M. (2020). Roles and capabilities of enterprise architecture in big data analytics technology adoption and implementation. *Journal of theoretical and applied electronic commerce research*, 16(1), 37-51.
13. Moorman, C. (2020, September 15). *Data ingestion: the first step to a sound data strategy.* Stitchdata. Retrieved from : https://www.stitchdata.com/resources/data-ingestion/
14. Lee, J., Wei, T., Mukhiya, S. K. (2018). *Hands-on big data modeling. Effective database design techniques for data architects and business intelligence professionals*. Birmingham: Packt Publishing Ltd.
15. Singh, C., Kumar, M. (2019). *Mastering Hadoop 3: Big data processing at scale to unlock unique business insights.* Birmingham: Packt Publishing Ltd.
16. Shekhar, C. (2019, August 20). *Big data ingestion: Parameters, challenges, and best practices*. Datapine. Retrieved from : https://www.datapine.com/
17. Hadwer, A., Gillis, D., Rezania, D. (2019). Big data analytics for higher education in the cloud era. In *2019 IEEE 4th International Conference on Big Data Analytics* (pp. 203-207). Suzhou: IEEE.
18. Banica, L., Radulescu, M. (2015). Using big data in the academic environment. *Procedia of Economics and Finance*, 33, 277-286.

19. Williamson, B. (2018). The hidden architecture of higher education: building a big data infrastructure for the smarter university. *International Journal of Education technologies for higher education*, 15(12), 1-26.
20. Demertzis, K., Iliadis, L., Anezakis, V. D. (2019). A machine hearing framework for real-time streaming analytics using lambda architecture. In J. Macintyre et al. (Eds.), *International Conference on Engineering Applications of Neural Networks* (pp. 246-261). Springer, Cham.
21. Amare, M. Y., Simonova, S. (2019). Overview of big data challenges, opportunities and its applications in the context of public administration organizations. In K. S. Soliman (Ed.), *Proceedings of the 34rd International Business Information Management Association Conference 2019 (IBIMA)* (pp. 12200 – 12209), Madrid: IBIMA Publishing.
22. Matacuta, A., Popa, C. (2018). Big data analytics: Analysis of features and performance of big data ingestion tools. *Informatica Economica*, 22(2), 25-34.
23. Isaacson, C. (2014). *Understanding big data scalability: Big data scalability series, Part I.* New Jersey: Pearson Education.
24. Osman, A. M. S. (2019). A novel big data analytics framework for smart cities. *Future Generation Computer Systems,* 91, 620-633.
25. Wang, Y., Kung, L., Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126, 3-13.