Educational big data infrastructure: opportunities, design and challenges

I M Putrama¹, G A Pradnyana¹, A A G Y Paramartha¹, I G M Darmawiguna¹, I M A Wirawan¹, I B N Pascima¹, I N S W Wijaya¹ and K Y E Aryanto¹

¹Data Science Research Group (DS-RG)
Faculty of Engineering and Vocational, Universitas Pendidikan Ganesha

Corresponding author: made.putrama @undiksha.ac.id

Abstract. The importance of data to help the management of educational institutions in decision making is recognized by all parties to improve the quality and competitiveness of their alumni in society. However, to process a Big Data, facilities and resource capabilities are needed to convert the data into information that can be easily visualized and consumed by various interested parties. As the inefficiency of the old technology use in helping the businesses' needs, the infrastructure of the big data is important. Big data presents various challenges in infrastructure provision and resource constraints as well as privacy and security concerns. This paper provides an initial development of how to build Big Data infrastructure in an educational institution that is adaptable to existing conditions which provide considerable performance and capabilities to perform information mining to assist in decision making. The scenarios, in this case, provide valuable additional insights for organizations to apply to their environment and can be leveraged to reach critical decision making through data visualization.

1. Introduction

The educational horizon has grown very fast. Students nowadays are easy to find learning materials online, whether in form of text, images, or videos. However, the rapid development of learning sources and other educational supporting data has made the users difficult to extract the needed information [1]. Especially in educational institutions that produce more and more data in various formats every day in the form of lecture materials, student information, admissions, and registration data that are hard to digest for use in decision making only using conventional data warehouse without leveraging more advance technologies [2]. Meanwhile, the data is highly demanded by the institutions to be able to make fact-based policies to improve the quality of services offered.

The Indonesian government has launched a policy for higher education namely "Kampus Merdeka" in early 2020. There are four main points of this policy, including the ease of opening new study programs and opportunities for students to study outside their departments. This implies the importance of making the right decisions in line with the various needs in the academic process that occurs.

One of the higher education's efforts to improve its quality is the optimal use of information technology. The availability of data and information often makes a difference between a university and another. In Industrial Revolution 4.0 era, data played an important role in strategic decision making, especially when entering the era of the data explosion in which universities in various fields were required to handle large amounts of data (big data). In general, big data can be defined as 5Vs, those are: a collection of data that is very large (volume), very fast changing/growing (velocity), comes in various...
forms/formats (variety), and has a certain value (value), provided that it comes from accurate sources (veracity)[3,4]. Big data can offer an invaluable insight and a competitive advantage with supportive technology and organizational resources. [5]. The main thing that distinguishes big data from conventional data sets lies in the management mechanism and its analysis, known as Big Data Analytics.

The use of Big Data Analytics in a university should be an important part of the application of technology in education so that it can fully support the implementation of the policy of Kampus Merdeka. A data-driven university will certainly have advantages, effectiveness, and efficiency in quality improvement efforts. One of its applications is in the aspects of accreditation and curriculum fine-tuning. Big data analytics will be able to predict the target of prospective college students, predict the possibility of early dropped-out students from happening, map student status based on GPA, conduct sentiment analysis as a form of work evaluation in society, mapping lecturers according to research fields, personalized online learning, recommendations for values-based internships, and many more.

For the baseline, this study focused on producing a reference working mechanism (framework) for big data analytics that was appropriate to the higher education field. The question framework related to the development of big data infrastructure, big data processing, and the use of big data in which to adjust the characteristics of data in tertiary institutions to support the policy of Kampus Merdeka. Analysis related to the observation of infrastructure models, mechanisms, technology, and their simulations was carried out at the beginning of the study.

2. State of the Art
The data-driven university is a concept of building a higher education system that in the process applies modern technology to utilize an unlimited amount of data. These data generated from the learning process from year to year for the benefit of real-time decision making, learning analysis, and data mining for the sake of education quality improvement. One of the latest research directions in the field of smart education was the use of Big Data technology. The implementation of this technology have produced various learning innovations such as independent learning media, intelligent tutoring systems, virtual lecturers, robot-based learning, and social learning systems [6]. Another review related to the prospects for modern education based on Big Data was put forward by Li and Zhai [7] who examined adaptive learning systems that can deliver material according to the special needs of learners.

There are some challenges in adopting Big Data technology for an organization. The resulting data was generally very large with various formats increasing rapidly every day [8]. Many factors related to IT infrastructure, such as its usability and complexity are taken into consideration. Various studies have been conducted before examining the opportunities and challenges [9,10]. Baig et al [11] examined the advantages, costs, security, and risk aspects of adopting Big Data technology. It was also revealed that currently there are still limitations to suit the theoretical models for adopting Big Data technology. Williamson [12] examined the complexity of higher education infrastructure projects. It was found that there were various hidden architectures related to technology, experts, standards, and practical values that allowed Big Data technology to be adopted in such a way that data infrastructure in higher education is not just a technical program. Data infrastructure is also a practical trigger in the achievement of policy objectives for developing related sectors. His research taken a case study of higher education in the UK, in which the goal of developing higher education based on data infrastructure is one of them so that institutional information can be accessed more broadly, especially for students who have the potential to enter university by paying and as a way to build competitive educational service provision. In addition, it also aimed at accelerating the quality of learning so that UK campuses can be evaluated and ranked more easily. However, his study did not discuss specific technical matters. The characteristics of the higher education infrastructure and the final architecture adopted to implement the Big Data technology used were not covered. His study discussed on how datafication in higher education can be done so that various decisions can be made more easily based on facts in the form of digital data. In making changes to the data-based system of a multinational consulting company, Deloitte, was asked to produce regulations on how the data were collected, shared or exchanged in the higher education environment. In addition, his research also examined the way to standardized data with global quality.
This research is related to various studies on the integrated information system development at the Universitas Pendidikan Ganesha (Undiksha), a state university of education in Indonesia that has been carried out previously [13–17]. Starting from the latest study that has been conducted, this research specifically intended to examine the general architecture of the IT infrastructure that suitable for higher education institutions in Indonesia concerning the Big Data era.

3. System Design
This research generally divided into three stages. The first stage began with preparing a system infrastructure based on big data technology by evaluating the latest technology and designing an effective and efficient infrastructure design. Next, the research continued by performing data preprocessing for the analysis and information extraction. The process was done by aggregating and transforming data from various existing education information systems in structured, semi-structured, and unstructured formats so that they can be analyzed to support future educational needs. The overall processes led to the data availability for conducting data analytics to support the implementation of the policy of Kampus Merdeka. The big picture of all stages carried out in this research as in Figure 1.

![Figure 1. The roadmap of Big Data Infrastructure development for Data-driven University.](image)

The above roadmap is the design and development stage of the Big Data infrastructure that suitable for the implementation of data-driven university. The process were done by performing observation on the models, mechanisms, and the available technologies. To produce a Big Data system infrastructure design that is in line with the data characteristics of the current system, this study initiated a design with the minimum possible specifications but can still accommodate data transformation from the existing system to the Big Data system being built.
4. Results and Discussion

In the early stages, the Big Data platform were tested using Hadoop which was implemented on a standalone computer. The processes were uploading and querying data as well as performance comparisons between processing several types of data through big data techniques and conventional techniques. From the test results, it can be concluded that a significant difference in query response time occurs when Hive accesses more than one table. This happened because Hive required more time to perform table-joins. When joining two or more tables Hive activated the YARN cluster. Based on literature studies, there are several ways to accelerate this, such as optimizing the Hive’s query and increasing the number and capability of cluster nodes. Therefore, the work was continued with testing using several computers (clusters).

![Figure 2. Query Response Time When Viewing and Joining (a) 1 table; (b) 2 tables; (c) 3 tables.](image)

In this study, a cluster was built using 6 computers. The cluster ran on Ubuntu Server 18.04 LTS server operating system. Two computers were configured as Name Node and the rest were configured as Data Node. The cluster was connected to the internet via 8-ports Switch with Gigabits technology capabilities. Each computer in the cluster has specifications as follows.

- Processor: Intel Core i7 3770 CPU 3.40 GHz
- RAM: Kingston 8 GB
- Hard drive: 1 TB
- LAN: 1000-Base-T
- 802.11 Wireless b/g/n PCI-E Mini card
In the proposed architecture, there were total of 6 nodes that were formed into a cluster with configurations: 2 NameNodes and 4 Data Nodes. Both NameNodes were designed using a typical HA Cluster in which one NameNode acted as an active Node Manager and the other one was in Standby condition. By using open software from Apache Ambari, various supporting application requirements for large data processing can be activated during the installation. Some of the initial applications installed were applications related to resource management such as: HDFS, YARN, FLUME, SQOOP, and large data processing applications such as Map Reduce, Hive, and Pig for testing purposes. The Flume and Sqoop applications were two applications that used to initialize data taken from the existing operational systems called E-Ganesh. For the needs of system testing at the end of the study, the sample data were taken from E-Ganesh that sourced from several systems, namely: Lecturer Database, Student Database, and Academic System Database.

The running configured system can be monitored via a Dashboard as shown in Figure 4. The CPU condition, memory, and bandwidth usage of each node in the cluster were easily observed at any time. Adding computers to the cluster or creating new clusters can also be done directly. In addition, Hadoop provides various applications that can be used to process large amounts of data according to the characteristics of the data and the information as needed.
In this architecture, a topology of how the Hadoop architecture was formed clearly explained. This architecture illustrates that the configuration was scalable to use multiple nodes as a compute centre in which additional nodes can be added further as needed. Two nodes that back up each other in an active-passive state as a manager which coordinates workloads when there were jobs to be done. This architecture illustrates more detail of the used parts to receive data in a structured or unstructured format compared to what is shown by Leo [2]. The topology results in this architecture can provide an overview that can be referenced by other educational institutions in implementing the big data architecture they need.

The biggest challenge in its implementation is the availability of data that is owned by each university. Even if it is available it is not a well-documented data. Meanwhile, big data infrastructure will be effective when the owned data is diverse and large in number as well as processing needs that require high-intensity performance. One of the most dominant uses in the future is to support decision-making carried out by various agencies that require valid data. Further, geographical conditions and network facilities owned by an area, especially in areas that have not been reached by a network with adequate speed, are also the challenge in its implementation. These challenges require further study.

Although the academic data used in this research was taken from Undiksha, these data had similar characteristics to data from Higher Educational institutions. Thus, the architectural design of the Big Data system generated in this research can be implemented in various Higher Educational institutions, especially in Indonesia.

5. Conclusion
In the research, an infrastructure design based on big data technology has been developed to perform large-scale processing for structured, semi-structured, and unstructured data, which is more scalable than the data warehouse systems running at Undiksha which is still limited to structured data management. The future work of this research is to prepare a data transformation mechanism from the current system and perform data preprocessing so that the data analytics process can be carried out in the future. In addition, the proposed infrastructure needs to be studied further to improve compliance with the criteria including available data characteristics such as volume, velocity, veracity, and so on.
Acknowledgment
This work is supported by the Department of Research Institutions and Community Service, Universitas Pendidikan Ganesha (grant number: 914/UN48.16/LT/2020).

References
[1] Wassan J T 2015 Discovering Big Data Modelling for Educational World Procedia - Social and Behavioral Sciences 176 642–9
[2] Santoso L W and Yusof 2017 Data Warehouse with Big Data Technology for Higher Education Procedia Computer Science 124 93–9
[3] Toba H 2015 Big Data : Menuju Evolusi Era Informasi Research Gate 1–6
[4] Yu J H and Zhou Z M 2019 Components and development in Big Data system: A survey Journal of Electronic Science and Technology 17 51–72
[5] Morabito V 2015 Big data analytics vol 21
[6] Shi W, Liu X, Gong X, Niu X, Wang X, Jing S, Lu H, Zhang N and Luo J 2019 Review on Development of Smart Education Procedia - IEEE International Conference on Service Operations and Logistics, and Informatics 2019, SOLI 2019 157–62
[7] Li Y and Zhai X 2018 Review and Prospect of Modern Education using Big Data Procedia Computer Science 129 341–7
[8] Tyagi A K, Priya R and Rajeswari A 2015 Mining Big Data to Predicting Future Journal of Engineering Research and Applications www.ijera.com ISSN5 2248–962214
[9] Matsubula F and Mnkandla E 2017 A big data architecture for learning analytics in higher education 2017 IEEE AFRICON: Science, Technology and Innovation for Africa, AFRICON 2017 pp 951–6
[10] Ternier S, Scheffel M and Drachsler H 2018 Towards a cloud-based big data infrastructure for higher education institutions Lecture Notes in Educational Technology pp 177–94
[11] Baig M I, Shuib L and Yadegaridehkordi E 2019 Big data adoption: State of the art and research challenges Information Processing and Management 56
[12] Williamson B 2018 The hidden architecture of higher education: building a big data infrastructure for the ‘smarter university’ International Journal of Educational Technology in Higher Education 15 1–26
[13] Aryanto K Y E, Dermawan K T and Putrama I M 2019 Fundamental Architectures of Datacenter ECR System Prototype by Applying Principle of Messaging Services - IOPscience Journal of Physics: Conference Series
[14] Aryanto K Y E, Dermawan K T and Putrama I M 2020 Management of Integrated System Prototype By Applying Principle of SCRUM International Journal of Innovation and Learning
[15] Permana A A J and Pradnyana G A 2019 Recommendation Systems for internship place using artificial intelligence based on competence - IOPscience Journal of Physics: Conference Series
[16] Pradnyana G A and Permana A A J 2018 SISTEM PEMBAGIAN KELAS KULIAH MAHASISWA DENGAN METODE K-MEANS DAN K-NEAREST NEIGHBORS UNTUK MENINGKATKAN KUALITAS PEMBELAJARAN JUTI: Jurnal Ilmiah Teknologi Informasi 16 59
[17] Putrama I M, Dermawan K T, Dantes G R and Aryanto K Y E 2017 Architectural evaluation of data center system using architecture tradeoff analysis method (ATAM): A case study Proceedings - 2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017 (Institute of Electrical and Electronics Engineers Inc.)