International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 6, December 2019, pp. 4970~4978
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i6.pp4970-4978

A proposed architecture of big educational data using hadoop at the University of Kufa

Ahmed Yaseen Mjhool[1], Ahmed Hazim Alhilali[2], Salam Al-augby[3]
[1,2]Information Technology Research and Development Centre, University of Kufa, Iraq
[3]Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Iraq

Article Info

Article history:

Received Feb 26, 2019
Revised May 2, 2019
Accepted Jun 26, 2019

Keywords:

Big data architecture
Distributed systems
Educational data
Hadoop
University of Kufa

ABSTRACT

Nowadays, educational data have been increased rapidly because of the online services provided for both students and staff. University of Kufa (UoK) generates a massive amount of data annually due to the use of e-learning web-based systems, network servers, Windows applications, and Students Information System (SIS). This data is wasted as traditional management software are not capable to analysis it. As a result, the Big Educational Data concept rises to help education sectors by providing new e-learning methods, allowing to meet individual demands and reach the learners' goals, and supporting the students and teacher's interaction. This paper focuses on designing Big Data analysis architecture, based on the Hadoop in the UoK and the same case for other Iraqi universities. The impact of this work, help the students learn, emphasizing the need of academic researchers and data science specialist for learning and practicing Big Data analytics and support the analysis of the e-learning management system and set the first step toward developing data repository and data policy in UoK.

Corresponding Author:

Ahmed Hazim Alhilali,
Information Technology Research and Development Centre,
University of Kufa,
Kufa, Najaf Governorate, Iraq.
Email: ahmed.alhilali@uokufa.edu.iq

1. INTRODUCTION

Recently, Information Technology Systems rated as one of the significant keys to improve and sustain the organizations' businesses. The educational sector has become oriented considerable extent by the technology [1]. Most of the Educational institutions provide for their students the ability to upload and download any required resources that help them to complete and submit assignments through an online platform instead of the traditional methods [1]. Also, they use digital systems to control several services in various campus facilities such as library registration, monitoring students' attendance and so on. However, due to the vast numbers of services that the internet offers for users, IT infrastructure is having problems dealing with this massive data. As a result, the Big Data concept has been used to describe the large quantities of structured or unstructured data, produced by organizations, companies, and institutions [2]. Moreover, the graduates' students are responsible for bringing more data every year. Hence, educational institutions need to use that data as an asset and become driven by it.

Big Data analytics can improve many aspects related to the education process and provide a better understanding of students [3, 4]. Also, new learning methods can be discovered through the adoption of convenient Big Data analytics tools and techniques. Till now, the impact of the Big Data on the educational institutions is limited with no high success rate [2]. As well, the process of applying analytics in the academic environment facing some ethical and financial challenges. However, with proper management of data and an

accredited guideline, education institutes can take control of those challenges. Learning Analytics in an educational institution is based on three models, Behavioral Model, Cognitive Model "which is entirely dependent on the teacher", and Constructivist Model that is dependent on the student to get the knowledge by their own.

This paper proposed the architecture of big educational data under the Hadoop open source framework in the University of Kufa. The aim of the research is to apply the concept of Big Data analytics in order to manage the collected data that cannot be handled with the traditional management methods and enhanced the learning and teaching experiences. Also, to provide researchers, students, and other big data fans a physical powerful tool that supports their projects.

The paper is organized as follows. Section 2 discusses the related works of using big data in the education sector. Section 3 describes the UoK perspective of big educational data. Section 4 explains the Hadoop 2.X framework. The proposed architecture hardware and software components are explained in Section 5. Finally, the conclusion section summarized the key finding of the paper and discussed future works.

2. RELATED WORKS

Currently, the online e-learning tools such as online chats, discussion forums, text messages, digital notes, and various Learning Management Systems (LMS) like Moodle and Blackboard has become more desirable by the users as these tools improve the traditional learning process by making the learning resources available anytime and anywhere through the internet. Using these online services produced a massive amount of data annually. Research by [5] noted that in 2005 the data size was approximately 130 Exabyte and is supposed to reach 40,000 Exabyte according to statistics made by the digital universe. The size of data made applicable in the above scenarios is so tremendous that traditional processing techniques cannot be deployed to process them. Due to the lack of conventional data processing applications, the academic institutions have started researching Big Data technologies to process educational data. Big Data tools can play a vital role in educational institutions because it facilitates the process of storing and retrieving information [2].

According to [6], in the successful academic institutions, Big data can be applied to create the completion and outplacement culture, reduce unproductive credits, redesign the instructions delivery, core services that include human resources, academic services, and finance, and optimize the operations and other functions. Furthermore, this article [6] stated that applying data mining concepts on the stored data are bases for future activities associated with higher education organizations. Also, Due to the increasing demands on the big and complex data analysis, many researchers become interested in the field of data mining [5]. However, the McKinsey report shows that the higher education sector represented the power-less link among all the areas of industry in capturing data [6].

Hadoop technique considered as one of the solutions that can overcome problems related to big data in many academic institutions. For instance, in Phoenix University (PU) there is an issue of the massive volume of data such as the data generated by discussion forums, they used Hadoop to deal with that issue. To point out the problems in PU, Hadoop mainly solved them were as follows: firstly, the digestion of large datasets such as (discussion form data, web usage logs). Secondly, data analysis for unstructured data by running a series of scripts [7].

Furthermore, Nottingham Trent University used big data concept to measure four aspects that indicate student engagement which includes swiping ID card into buildings, using the library, and virtual learning environment using electronic submission of assignments. They found that a quarter of the students have low engagement levels depending on the results of applying the measurement [7].

3. PERSPECTIVE BIG EDUCATIONAL DATA IN UoK

In the educational sector, a massive amount of data is generated every day due to the use of the modern learning technologies such as mobile applications, social media, online management system, web resources, LMS student engagements, etc. This section will focus on applications that offered by the UoK for their students, employees, and lecturers. Using the following applications will generate a significant amount of data that cannot be handled with traditional management software in the near future.

3.1. E-learning's data

Most of the e-learning data are generated from the online learning resources and the use of the virtual learning environment. The UoK provides an e-learning platform based on Moodle for their students. Not only that but also UoK launched new distance learning project under Moodle core to serve community learning demands, the project called Kufa Open Online Courses KOOC, it is mostly depending

on streaming short videos to the public in various topics in order to support self-learning and enrich Arabic learning resources on the internet.

According to e-learning server statistics as seen in Figure 1 [8], the number of courses and resources generated on that platform increasing every year due to university view and mission. This virtual environment enables both students and lecturers to exchange information, ask questions, using the discussion board and upload files. Analyzing this data will help the organization's administrator to understand how to enrich the publics' and students' learning experience as well as act upon the unexpected situation and improve it.

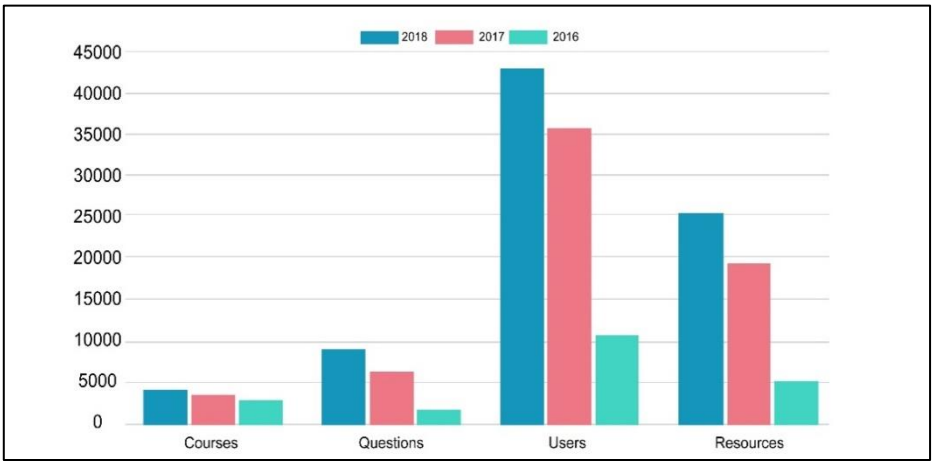

Figure 1. E-learning server statistics [8]

3.2. Network servers' data

The network and internet services serve more than 3500 employee and 65000 students, user interactions with these services will dramatically generate logs, files, reports, etc. on a daily basis. Therefore, a huge amount of data will be ready either quickly managed, processed and stored or deleted due to lack of knowledge of dealing with raw Big Data. In summer 2018, UoK is considered new network system with the more flexible setting to connect and monitor users to the services, therefore, many logs, reports, and files will be generated, and then the more network's files are generated, the deeper knowledge is driven about UoK's network.

3.3. Web server's data

The university provides many websites covers its activities, news, and events. Also, each faculty member has its own website where s/he publishes everything related such as events and other directions to his students. With more than 16000 websites for the teaching staff that contains important information such as Curriculum Vitae, contact information, lectures, and published researches. As a result, a huge amount of data related to those websites activates are created.

3.4. Windows applications' data

All administration and management software are developed by the university's programmers and that software basically work in two ways either locally on user computer or on a server that connected to the university intranet. This old-fashion technique prevents data to get bigger because simply there is no hard disk space for it and there is no perspective view strategy on data. Data should be leveraged in the right way in order to get meaningful information out of it and this is what full Big Data system does.

3.5. Students information system's data

The SIS is a local application server used to register and manage students' data during their life study at UoK. SIS's data accumulate all possible students' data such as personal information, images, uploaded files, ID card, official documents, and academic records. Currently, at the time of writing this paper and according to UoK main website, more than 65250 students are registered in the system and this number increasing around 6000 students yearly. Hence, the UoK needs an intelligent solution and to be ready for dealing with Big Data coming up in the near future.

3.6. Datasets repository for researchers

In UoK, data getting bigger every day and there is no clear vision on the data. Academic researcher and scholar must get safe and easy access place for their Big Data sets to do their research work and save back the results, thus, a reliable storage place will be beneficial to other scholars to keep working on the same datasets within deferent research approach. Also, grouping Big Data sets on the same disk space will encourage researchers to propose solutions based on given problems. As a result, our proposed work will be the best place for those who are dealing with Big Data sets.

4. HADOOP 2.X

Hadoop is an open-source framework founded by Apache foundation, and it is used for storing a massive amount of data [4] and running analysis applications across a set of nodes which build on cheaply priced hardware [9]. It offers a massive technique of storage for any type of data, tremendous processing capability, and the competence to handle virtually unlimited concurrent jobs or tasks. The new version of Hadoop 2.x mainly came up with three parts [10]:

a. HDFS is a distributed file system that handles large data sets running on commodity hardware.
b. MapReduce is the heart of Hadoop and a programming paradigm that enables massive scalability across hundreds or thousands of servers in a cluster.
c. YARN is one of the core components and resource management and job scheduling technology in Hadoop. It is responsible for allocating system resources to the various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes.

4.1. Why is hadoop important?

Due to the rapid increase of data size in social media and the internet of things (IoT), Hadoop considers as a solution that can store a vast amount of different type of data and process it quickly. More, a distributed computing model gives Hadoop the ability to process Big Data in a fast manner [11]. As a result, using a distributed model leads to a more powerful process. Furthermore, Hadoop handles node failure in the cluster efficiently by redirecting the data to the other working node and replicate the data x-times automatically according to the system setting. Moreover, flexibility, scalability, and low cost make Hadoop one of the best open-source framework used to analyze Big Data [12].

4.2. Data transportation

This section covers the main software that can push whatever data we have to the Hadoop File System (HDFS) for future processing. In this proposed model, we gave the system administrators multiple ways to handle data between local storage and HDFS such as below [11].

4.2.1. Apache flume

Apache Flume is a service for efficiently collecting, aggregating, and moving large amounts of data to HDFS. It is reliable and distributed framework that has a simple and flexible structure based on streaming data flows. It is robust and faults tolerant with tunable reliability, many failovers, and recovery mechanisms.

4.2.2. Apache sqoop

Apache Sqoop used to move structured data from locally stored SQL databases into HDFS in order to be ingested by Apache HBase for further processing [13]. Thus, all the academic's databases can be linked to Sqoop, which can also be scheduled to ingest data regularly, to get any data updates.

4.3. Saving and processing data

Flume, Sqoop or other frameworks that transport local data, is chopping and storing data on the HDFS. Data in HDFS is scattered over cluster nodes for built-in fault tolerance. HDFS has one master node (name node) and many slave nodes (data nodes) and that nodes reside on commodity computers and each node offers local storage and computation [11]. The name node stores metadata whereas data nodes stores data blocks. Figure 2 [4, 14] shows the basic Hadoop principles for storing and querying Big Data.

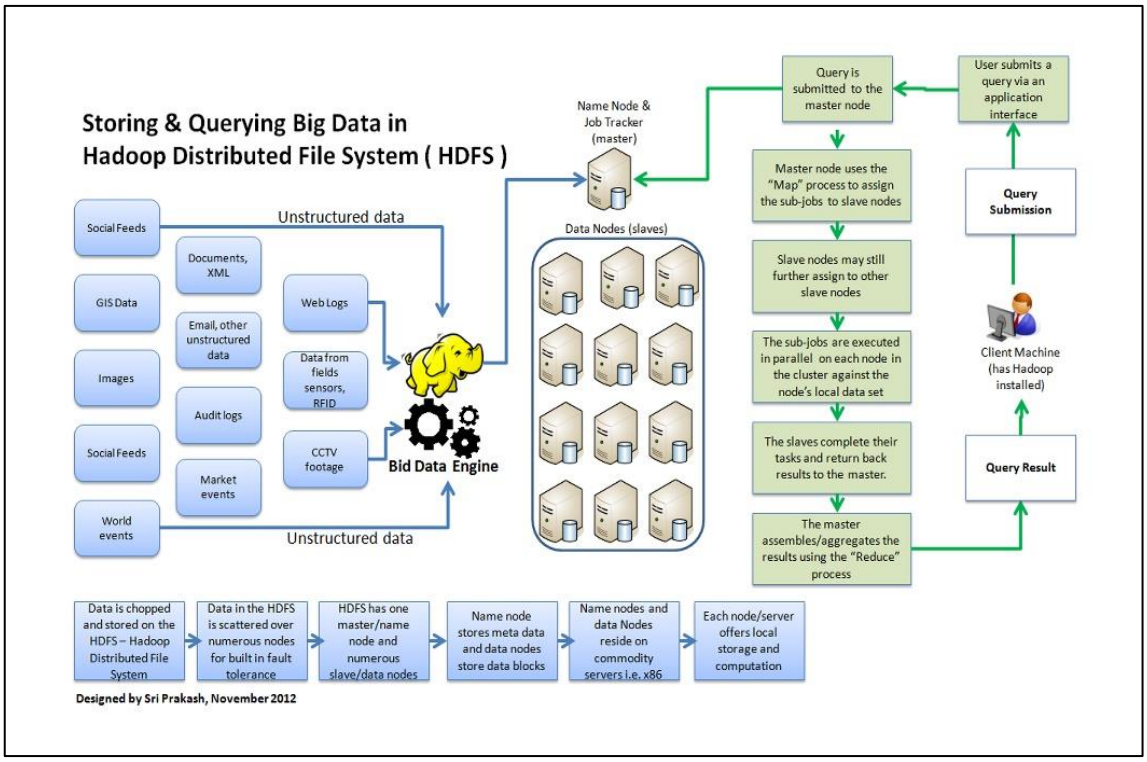

Figure 2. Basic Hadoop principles for storing and querying Big Data [4, 14]

4.3.1. Apache HBase

It is a column-oriented database management system that runs on top of HDFS. It is good for sparse data sets, which are common in many Big Data use cases. HBase is not like a relational database and does not support SQL language. HBase applications are written in Java and support writing applications in Avro, REST, and Thrift. HBase has a set of tables. Each table contains rows and columns and must have a primary key. This key used to access HBase tables [14].

4.3.2. Apache zookeeper

Zookeeper is an open source software that provides centralized infrastructure and services that enable synchronization across a Hadoop cluster. It maintains shared objects required in large cluster environments. In simple words, the zookeeper is developed for coordinating and managing services between cluster for this case zookeeper need to be installed in distributed mode, it is responsible for tracking and recording failures of jobs assigned for a specific task [14].

4.4. Data analytics and visualization

Analyzing data is not enough unless it is shown in an easy way to understand simple graph to the ordinary users and this can be done using cutting-edge technologies such as the tools listed below:

4.4.1. Apache pig

Apache Pig represents a Big Data analysis platform that can express and evaluate data analysis programs due to its high-level-language and infrastructure. The amenable parallelization structure of Pig programs considered essential features that enable them to deal with massive datasets precisely. Currently, Pig's infrastructure layer includes a compiler which generates sequences of Map-Reduce programs and it uses Pig Latin as a scripting language. The Latin scripting language has the following characteristics: programming easily, optimization opportunities and extensibility.

4.4.2. Apache hive

Apache Hive is a data warehouse platform that facilitates the processes of reading, writing, and managing huge datasets hosted in a distributed storage system. With a command line interface and JDBC driver hive can connect users to its framework.

4.4.3. Apache zeppelin

It is a web-based notebook that enables interactive data analytics. Zeppelin is a new and incubating multi-purposed web-based notebook which brings data ingestion, data exploration, visualization, sharing and collaboration features to Hadoop [15]. It is needed to show graphs and charts of big educational data to the user.

4.5. Information and knowledge

All data analytics from the first step of mining to visualizing aim to show the power of hidden data. This led us to get new knowledge about what we know related to the business model, thus that knowledge would empower decision maker to take a critical decision for improving their business. Automated decision making can happen with the help of machine learning algorithms [11].

4.6. Machine learning

In this field we considered weka software, it is a collection of machine learning algorithms for data mining tasks [13]. Weka support distribution tasks with two packages assign to this purpose, the first one is tied to Hadoop platform called DistributedWekaHadoop2 and the second one is tied to SPARK platform called Distributed WekaSpark.

4.6.1. Apache mahout

It aims to transform Big Data into big information by providing scalable machine learning algorithms' library that can be used on top of Apache Hadoop and the MapReduce model. As well as, Mahout offers data science tools used to explore meaningful patterns in Big Data sets. Equally important, Mahout supports four major data science use cases which include: Collaborative filtering, Clustering, Classification, and Frequent itemset mining.

5. RESEARCH METHODOLOGY

In this section, we will briefly describe the hardware, software components and the technical structure of them. In the proposed system, eighteen accommodate computers were used and those may come up with different hardware specifications (spec) and at least according to the minimum computer spec that can run the basic software most end users operate, which is (Dual core 2.4 GHz Processor, 8 GB RAM, 320 GB 5400 RPM hard drive) [16], however we will consider Table 1 [17] spec as seen below for the proposed work.

Table 1. Cluster requirements

Item	Description
Master	Desktop, laptop, or accommodate server
Nodes	Desktop or laptop
Switch	Netgear ProSafe 24
Network Cabling	CAT6 Ethernet patch cable
RAM	At least 8 GB per node
Hard disk	At least 500 GB per node
Enclosure (shelf or rack)	A 42U 19" rack or cheap IKEA shelve
Socket multiplier	Any type
Power Supply (UPS)	Ex. 10 nodes of Acer X1700 at 220 Watts each amount to 2200/120 = 18.4 Amperes,

As seen in Figure 3, one computer assigned as master (NameNode) that only stores the metadata of HDFS, tree directory of all files in the file system and tracks files across the cluster, accommodate Dell server assigned as a master machine with. Another computer supported as a secondary NameNode as a recovery image to the master NameNode in case it fails. The rest 16 computer treated as slaves (DataNode) to the master in order to run map-reduce jobs in the parallel task. We can deploy DataNode within NameNode's computer just for using more resources, same things with secondary NameNode computer. As a result, we will get a cluster with at least the following spec (18 x Dual core 2.4 GHz Processor, 144 GB RAM, 5.75 TB 5400 RPM hard drive). The operating system of the computers is Linux Centos 6.5 and by using Netgear switch, these computers are given static IP's and connected as a local network to be attached to the intranet network of UoK. Briefly, other hardware requirements for building and installing a Hadoop cluster from commodity components can be found in Table 1. More technical details in this report [17]. All the mentioned components above will be connected as a star network topology as seen in Figure 3.

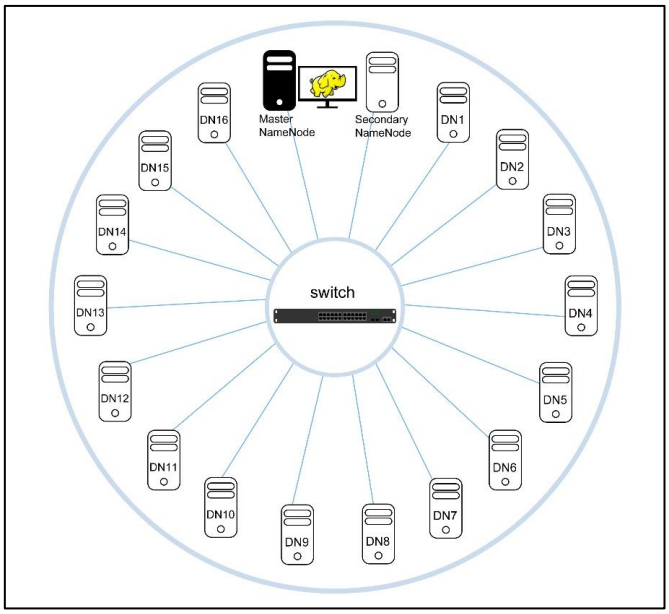

Figure 3. The proposed structure

The proposed architecture is divided into four major layers as seen in Figure 4: (1) Local Data Layer, (2) Transportation Data Layer, (3) Processing and Storing Data, (4) Data Analytics and Visualization Layer. The first layer is the main data generator which includes many systems, applications and devices like servers, computers, digital libraries, websites, etc. The second layer consists of many highly reliable mechanism frameworks that responsible to move data easily and safely on top of the distributed system, thus, selecting a specific framework depends on speed and type of data thrown from machines. Layer 2 has two methods for transportation data, first one is called semi-automated method that allows users and systems to upload data into local shared spooling directory on a regular basis to let Flume and Sqoop getting data into HDFS; second one is called fully-automated method which is customizing Flume or Sqoop with a special script to be connected with local machines. The third layer has the main cores of Hadoop (MapReduce and HDFS) and other software that manages tasks and jobs; for in-memory and real-time processing, Apache Spark is the best choice for that matter. The last fourth layer, it has all application that can clean, sort, analyze and visualize data in order to develop a smart model with the ability to take a decision for solving a certain problem under Mahout Framework.

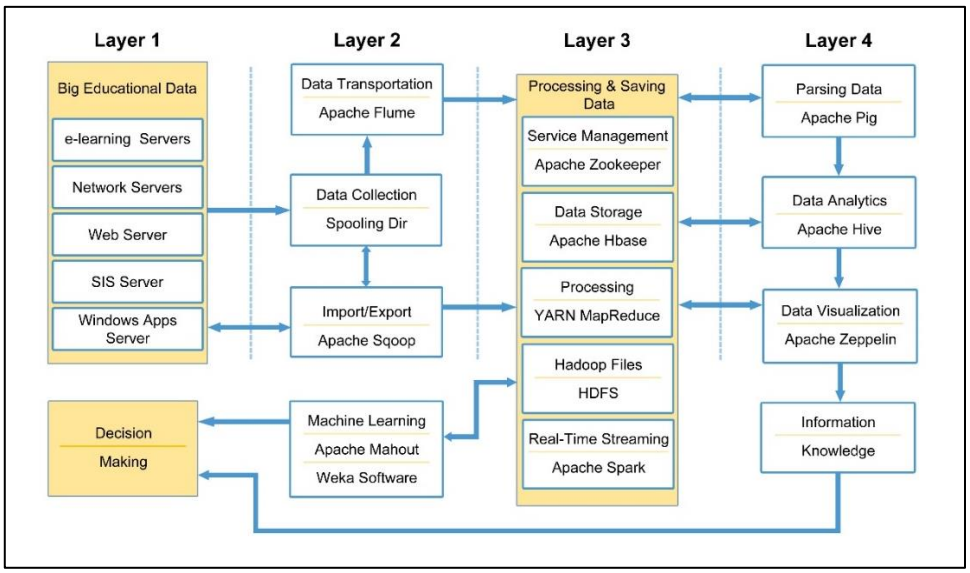

Figure 4. Proposed architecture

6. CONCLUSION

The proposed Big Data architecture is obsolete, and a new data model that integrates Hadoop to the existing systems is structured on Information Technology Systems' world of UoK. Therefore, the design is offered on a private network, and it is designed to be used as a Big Data Lab for students and researchers and as a solution for growing Big Educational Data in UoK. We recommend considering Hadoop open-source software and its environment software on a private network and managing by UoK's engineers, and not using public cloud solutions that offered by big software companies like Google or Amazon, because the last one is expensive especially for a university in a developing country like Iraq. Thus, the evolution of open-source products for managing and analyzing educational data will allow Iraqi universities to benefit from this new trend that empowers today's education. In our future research, we intend to implement the proposed design in multiple node Hadoop clusters and evaluate its performance working with structured data from our university LMS and unstructured data from Social media.

REFERENCES

[1] V. Vatsala, *et al*., "A Review of Big Data Analytics in Sector of Higher Education," *Int. J. Eng. Res. Appl.*, vol. 7, pp. 25-32, 2017.
[2] B. Logica and R. Magdalena, "Using Big Data in the Academic Environment," *Procedia Econ. Financ.*, vol. 33, pp. 277-286, 2015.
[3] B. Manjulatha, *et al*., "Implementation of Hadoop Operations for Big Data Processing in Educational Institutions," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 4, 2016.
[4] P. S. G. A. Sri and M. Anusha, "Big data survey," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 4, 2018.
[5] H. Fatlawi, *et al*., "An efficient hybrid model for reliable classification of high dimensional data using k-means clustering and bagging ensemble classifier," *J. Theor. Appl. Inf. Technol.*, vol. 96, pp. 8379-8398, 2018.
[6] J. Murumba and E. Micheni, "Big Data Analytics in Higher Education: A Review," *Int. J. Eng. Sci.*, vol. 6, pp. 14-21, 2017.
[7] N. Baker, "3 Universities That Are Using Big Data - QS," 2017. Available: https://www.qs.com/3-universities-that-are-using-big-data/?fbclid=IwAR2eH_8StopgXHIlnGn5IQvnwIKmDQVM_xTWZc5oq6FDp4_vwNkUzG4xZOQ.
[8] Information Technology Research and Development Centre University of Kufa, "E-learning server statistics," 2018. Available: http://elearning.uokufa.edu.iq/?page_id=100.
[9] J. Parsola, *et al*., "Post Event Investigation of Multi-stream Video Data Utilizing Hadoop Cluster," *Int. J. Electr. Comput. Eng*., vol. 8, pp. 5089, 2018.
[10] Apache Hadoop, "Hadoop – Apache Hadoop 2.9.2," 2018. Available: https://hadoop.apache.org/docs/r2.9.2/.
[11] W. Tom, "Hadoop: The Definitive Guide," *Fourth Edition The Definitive Guide Storage and Analysis at Internet Scale*, 2015.
[12] Insights SAS, "What is Hadoop?" Available: https://www.sas.com/en_us/insights/big-data/hadoop.html.
[13] Waikato University, "Machine Learning at Waikato University." Available: https://www.cs.waikato.ac.nz/ml/index.html.
[14] K. Sin and L. Muthu, "Application of Big Data in Education Data Mining and Learning Analytics – a Literature Review," *ICTACT J. Soft Comput.*, vol. 5, pp. 1035-1049, 2015.
[15] Hortonworks Inc., "Apache Zeppelin." Available: https://hortonworks.com/apache/zeppelin/.
[16] Clark University, "Our Recommended Computer Specifications." Available: https://www2.clarku.edu/offices/its/purchasing/recommendations.cfm.
[17] J. L. Leidner and G. Berosik, "Building and Installing a Hadoop/MapReduce Cluster from Commodity Components," *LOGIN*, vol. 35, 2010.

BIOGRAPHIES OF AUTHORS

Ahmed Mjhool got his BSc. Degree in computer engineering form University of Technology/Iraq in 2009 and master's degree in computer engineering form Florida Institute of Technology/ USA in 2017. During his master study, Mjhool had an industrial experience with big companies such as GE and Alstom at Florida site. After that, he joined IT research center as a researcher and college of engineering as a lecturer at the University of Kufa. His research topics interest are Big Data Analytics, Machine Learning Algorithms, Apache Hadoop, Spark, Complex Embedded Systems, Internet of Things, and Wireless Sensor Networks.

Ahmed Hazim Alhilali received the BSc. Degree in computer science from Imam Ja'afar Al-Sadiq University in Computer Science in 2009 and the Master degree in Information Technology from the University of Technology Sydney, Sydney, Australia, in 2016. My project research focuses on the cloud computing pricing models to provide a comparative study that helps the educational institutions to choose the suitable one depending on a group of factors. Other research interests are Healthcare Monitoring Systems, Internet of Things (IoT), Attendance Monitoring Systems.

Salam Al-augby received his BSc. Degree in Electronic and Electrical Engineering from MEC in 1997 and the Master degree in Computer Science from the University of Technology, Iraq in 2005. He got his PhD degree in IT in Management from University of Szczecin, Szczecin, Poland in 2015. The area of interests are Data Mining, Text Mining, Behavioral Finance, Sentiment Analysis, IT in Management, Big Data analysis, Social media analysis, and Natural Language Processing.