IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 28, 2021, accepted March 31, 2021, date of publication April 2, 2021, date of current version April 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3070737

Big Data Platform for Educational Analytics

AMR A. MUNSHI[1], (Senior Member, IEEE), AND AHMAD ALHINDI[2], (Senior Member, IEEE)
[1]Department of Computer Engineering, Umm Al-Qura University, Makkah 21961, Saudi Arabia
[2]Department of Computer Science, Umm Al-Qura University, Makkah 21961, Saudi Arabia

Corresponding author: Amr A. Munshi (aaamunshi@uqu.edu.sa)

This work was supported by the Deanship of Scientific Research at Umm Al-Qura University under Grant 19-COM-1-01-0017.

ABSTRACT Huge amounts of educational data are being produced, and a common challenge that many educational organizations confront, is finding an effective method to harness and analyze this data for continuously delivering enhanced education. Nowadays, the educational data is evolving and has become large in volume, wide in variety and high in velocity. This produced data needs to be handled in an efficient manner to extract value and make informed decisions. For that, this paper confronts such data as a big data challenge and presents a comprehensive platform tailored to perform educational big data analytical applications. Further, present an effective environment for non-data scientists and people in the educational sector to apply their demanding educational big data applications. The implementation stages of the educational big data platform on a cloud computing platform and the organization of educational data in a data lake architecture are highlighted. Furthermore, two analytical applications are performed to test the feasibility of the presented platform in discovering knowledge that potentially promotes the educational institutions.

INDEX TERMS Artificial intelligence in education, education, educational big data, educational data mining.

I. INTRODUCTION

Massive complex educational related data is being produced and with proper management, immense knowledge can be extracted. Over the past two centuries, the world went through a great expansion and enhancement in education quality. The desire to enhance the quality of education is continuous. Recently, artificial intelligence techniques are being utilized to assist in making informed decisions related to improving educational outcomes. Among those artificial intelligence techniques are data mining and knowledge discovery methods [1]–[7].

There are numerous challenges in handling educational data efficiently. In general, these challenges can be categorized into technical and organizational challenges [8]–[11]. The technical challenges can be summed into four main challenges: 1) the capability of the infrastructure at the universities premises to process the produced data, 2) the ability to monitor the effect of the made decisions, 3) the absence of a comprehensive platform tailored for educational organizations for gathering and analyzing the produced data effectively, 4) deploying and using technologies that handle educational data require skilled and talented practitioners. In order to tackle those challenges, the characteristics of educational data must be considered to find appropriate tools that may assist in the gathering and analyzing of the produced data. The educational data is: 1) produced at large volumes, 2) the educational data produced varies in type, for example the produced data could be in a structured or unstructured form, 3) the rate of educational data produced is high in velocity, 4) the produced data needs to be handle in an appropriate manner to extract value. From those aforementioned characteristics, educational data can be considered as a big data challenge. The characteristics of big data is having data that incorporates the 4V's, i.e., volume, variety, velocity and value, which matches the nature of educational data. While handling big data requires costly infrastructure and expertise, there has been much progress using big data analytics in business and industry sectors for enhanced working, effectiveness and informed decision making. However, constructing a big data platform for a specific area can be a complex task due to the lack of rigidity in the produced data. This makes it more difficult to precisely define what the constructed big data platform will achieve. This applies to communication interfaces, communication with other applications and computation on various types of data. Accordingly, constructing

The associate editor coordinating the review of this manuscript and approving it for publication was Francisco J. Garcia-Penalvo.

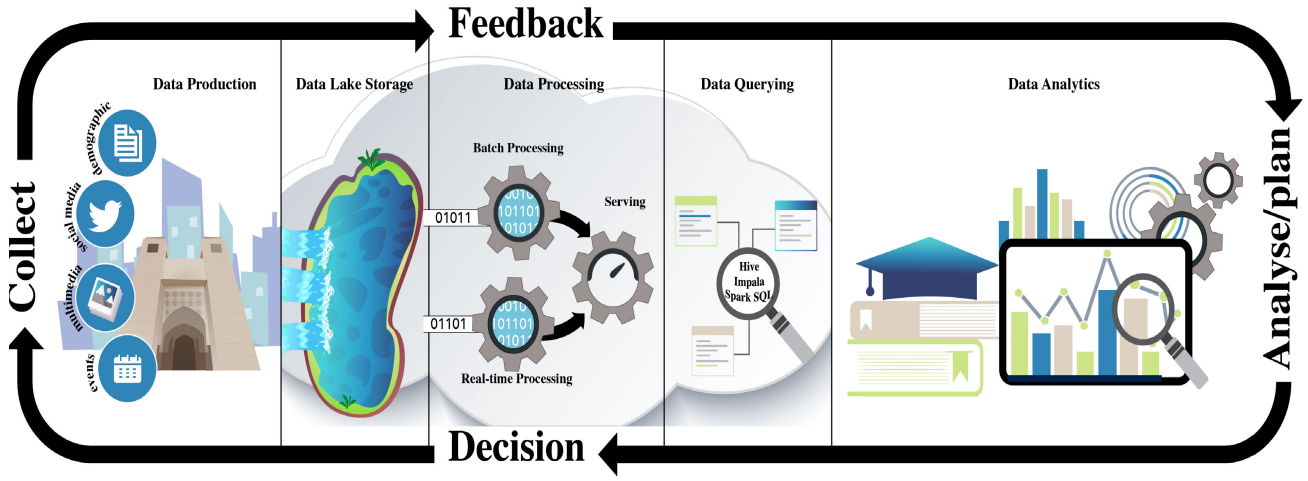

FIGURE 1. The educational big data platform that covers the flow of educational data from data production to data analytics, featuring a feedback loop.

and big data platform for a specific sector is data and application dependant. For example, a big data platform for specific healthcare applications was presented in [12]. Another big data platform to handle the smart grid data for energy optimization was presented in [13], [14]. Also, for fraud detection and prevention in [15]. In education, numerous research has been conducted to improve the quality of education in many aspects. In predicting student academic success [16], predicting the students' final grades [1], and course recommendation [17]. As the volume, variety and velocity of educational data is increasing, there has been growing interest in the education community to utilize this produced big data for improving the educational outcomes in many aspects including enhancing the learning performance of students, enhancing the working effectiveness of instructors and reducing administrative workload. Big data is being considered a revolutionary significance to education, and educational big data is becoming of research interest. Many works highlight the challenges and opportunities of big data analytics for education [18]. In [10] a big data architecture for higher education analytics was presented, in [19], a model for dropout prediction in Edx and massively open online courses (MOOCs) platforms was proposed. Also, a recent work in curriculum reformation for big data in Chinese universities is highlighted [20]. However, a comprehensive platform to handle educational big data that takes into consideration the four aforementioned challenges is of interest. For that, this paper presents an educational big data platform considering the following contributions to field: 1) an infrastructure that is able to handle the educational big data from data production to analytics. 2) includes a feedback loop to monitor the short-term and long-term affects of the decision-making process. 3) A comprehensive platform to perform various educational data analytical applications. 4) an effective environment for non-data scientists and people in the educational sector to apply their demanding educational applications.

The remainder of the paper is structured as follows. Section 2 presents the educational big data platform including the life-cycle of the educational data from data production to data analytics. Section 3, highlights the data lake architecture to store the educational data in a way that maximizes its availability and accessibility for analytical applications. Further, the implementation of the education big data platform on a cloud platform is also presented in Section 3. The application of the educational big data platform on two studies to test the feasibility of the platform is presented in Section 4. Finally, the conclusions are drawn in Section 5.

II. EDUCATIONAL BIG DATA PLATFORM

The educational big data platform can be decomposed into layers. The produced data flows through five subsequent stages: 1) data production, 2) data storing, 3) data processing, 4) data querying, 5) data analytics. Once the data is analyzed and decisions are made with respect to the desired application, the effects of the made decisions are observed through a feedback loop. Fig. 1, presents the educational big data platform and the following subsections highlight the stages of the platform.

A. DATA PRODUCTION

Educational data is being produced from numerous sources at high rates. This includes any data that an educational organization may produce, such as, data generated from administrative processes and systems, student demographic data, coursework, projects, grades, admissions, research, buildings' information, video footage and events. Also, this includes any data that is related to the educational organization, such as, social media data. The data is produced from multiple sources at large volumes and high rates. In addition, the data is produced in various forms i.e., structured, semi-structured and unstructured data. This multiple source and

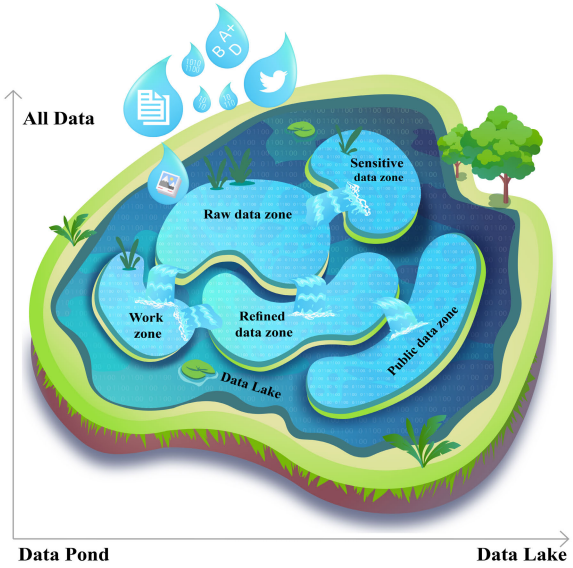

FIGURE 2. Data lake architecture for educational big data.

various form data needs to be stored efficiently in a proper repository.

B. DATA LAKE STORAGE

The produced educational data needs to be stored efficiently in its raw form. This data belongs to an educational institution and is logically related. For that, in this educational big data platform, a data lake architecture is utilized. This allows to store the multiple source and various form produced data at any scale before different analytical studies are performed to uncover insights or find answers for queries in the future. The utilization of a data lake architecture in this educational big data platform is to have all data available in its raw original form. The produced educational data is ingested and organized in the data lake, and called upon for analysis and when needed. However, the ingestion of data from multiple sources may end-up forming a collection of disconnected unorganized data (data puddles) that is unusable (data swamp). In order to avoid the risk of data swamping, in this educational big data platform, the data lake architecture is organized into ''data ponds'' that typically form a data lake. Fig. 2, illustrates the data lake architecture. Based on the data produced for an educational organization, the data lake could be divided into four ponds (zones) taking into account sensitive data:

1) Raw Data Zone: Where data is ingested and kept in its original state.
2) Refined Data Zone: Where the cleaned and processed data is kept.
3) Public Data Zone: Data that is available for public is kept here after cleaning and pre-processing.
4) Work Zone: This zone can be organized into ''data puddles'' for specific users or projects in a flexible manner. This zone is used by data scientists and general people in the educational sector to prepare their data for analytical applications. Once the analytical work is performed at this zone, the data may be moved into the Refined Data Zone for other applications that require the same refined data.
5) Sensitive Data Zone: Where sensitive data is kept. This zone has limited access and can only be accessed by individuals responsible for ensuring that sensitive data does not proliferate into the rest of the data lake.

Within the Raw Data Zone, the data is partitioned into ''data puddles'' based on the data producer for optimal data retrieval purposes. This data lake architecture can service various types of analytical studies from data mining to visualizations and dashboards, and beyond. The aforementioned zones form the data lake architecture utilized in this educational big data platform.

C. DATA PROCESSING

From the previous stage, the data stored in the data lake are organized and ready for consumption. In this educational, big data platform the processing of data is transferred to a Hadoop cluster [21], which consists of master and worker nodes for processing data. Utilizing a Hadoop cluster allows to process various types of data efficiently in their native format, from student records to Twitter feeds. Instead of fixed columns and rows of relational data, the data may have complex structures and a variety of records. Educational analytics require processing of educational data in batch and real-time. For this, a processing architecture that is able to fulfill the processing of big data in batch and real-time is required. The processing architecture of [22] is followed in constructing the data processing layer which allows multi-node massively parallel processing of data.

D. DATA QUERYING

Querying data is the cornerstone of the analytical applications of big data. In the previous subsection, the processing components enabled two main types of big data applications namely, query answering and analytics. As mentioned previously, the educational data may include structured data (student demographic data), semi-structured data (course evaluation and reviews) and unstructured data (campus images or audio streaming). The range of such data makes querying a complicated task, in which the required data is sometimes a result of information retrieval or data analytics (data mining). Based on the requirements of educational organizations, querying components that enable batch and real-time data analytical applications are required. In order to achieve this, in this presented educational big data platform three querying components are considered. Each querying component has a different approach in functioning and executing parallel operations on top of the processing layer. Hive [23] uses MapReduce operations, Impala [24] uses the memory of worker nodes, and Spark SQL [25] uses in-memory computation on top of the Spark processing component. Hence, the presented educational big data platform is able to cover a broad range of query answering and analytical applications.

E. DATA ANALYTICS

The objectives of the data analytics stage is to extract insights and assist in making informed decisions that essentially promote the operation of the educational organization as a whole, including:

- Predicting student performance to provide informed guidance.
- Course recommendations to maximise students' potential and reduce the risk of failures.
- Student major recommendations to maximise enthusiasm and motivation for continued education.
- Extracting features of students that are in risk of dismissal.
- Extracting features of students that maintain high GPA.
- Studying correlations between student success and department facilities.
- Analysing tweets to predict flu outbreaks. student satisfaction or cyber-bullying behaviors.
- Detecting smoke, fire and suspicious actions through image, video or tweets analysis.

The presented educational big data platform features a feedback loop to observe the effects of the made decisions on the educational organization. This could be useful for example, in monitoring the results of students of interest, or monitoring students' GPAs for a specific department after making certain decisions. In general, at this stage, applications that include data mining and knowledge discovery, statistical and table manipulation, and visual analytics could be performed.

III. IMPLEMENTATION OF THE EDUCATIONAL BIG DATA PLATFORM

In this section the implementation of the educational big data platform to comply with needs of a university institution is presented. The construction of this platform will serve as a test-bed to develop and validate end-to-end educational related applications with regards to Umm Al-Qura university, Saudi Arabia. The university encompasses seventeen diverse buildings used by a community of over 100,000 students, staff and faculty members, and is one of the largest educational data producer in Saudi Arabia. The following subsections discuss the implementation of the educational big data platform including the data lake on a Google cloud computing platform with respect to the stages presented in Section III.

A. DATA LAKE AND DATA PRODUCTION

At this stage, the data sources and produced data are specified. This includes data such as, present and past student demographic data, projects and research data, buildings' information, laboratory needs, courses data, multimedia data, events and social media related data. This data includes structured, semi-structured and unstructured data that is growing at a vast rate. Thus, the types of data and data sources are specified. This data produced from a variety of data sources needs to be ingested and organized in the data lake cloud architecture. An advantage of utilizing a cloud data lake in this implementation, is that cloud platforms comply in accordance to the Cloud Security Alliance organization that promotes the use of adequate practices for providing secure forms of computing. Also, this allows the related data to be sent from anywhere to the data lake via a network connection. The data is organized with respect to the architecture discussed in Section II-B. Then the data is stored in its raw format in the Raw Data Zone. All data is kept on demand in the Raw Data Zone except for sensitive data. Sensitive data is transferred into the Sensitive Data Zone and then removed from the Raw Data Zone. The sensitive data includes financial, faculty and staff records. It should be noted that data in the Raw Data Zone could also be sensitive, however, only high sensitive data is stored in the Sensitive Data Zone. To minimize the risk of unauthorized disclosure of student personally identifiable information (PII) from education records of Umm Al-Qura university data, deidentification is considered in compliance with the Family Educational Rights and Privacy Act (FERPA) [26]. In this implementation non-private data such as, events, statistical, open data sets are kept in the Public Data Zone for public use. In this educational big data platform, Tableau [27] is used to clean and preprocess data before it is transferred to the other data zones (data ponds). The components that are utilized to perform this cleaning and preprocessing of data are presented later in Subsection III-D.

B. PROCESSING OF EDUCATIONAL BIG DATA

Once the data is stored and organized in the data lake, analytical tools to process and analyze data are possible. In this implementation, Hadoop [21], which is a software framework that stores and processes huge amounts of data is utilized at this stage. It relays on distributed clusters of commodity servers for storing and processing data. This allows to processes huge amounts of data and to perform batch data analytical processing. However, to fulfil the real-time educational data processing, an additional component is required. For this, the setting of [22] is adopted for adding the real-time processing components, Spark [25]. To implement the educational big data platform processing layer, a cloud computing cluster that consists of five nodes is established. The cluster consists of one master node and four worker nodes. The master node is a 16vCPUs 60GB RAM machine and the worker nodes are 8vCPUs 30GB RAM machines, all running Linux operating system. The Hadoop framework and Spark components were setup for storing and processing the educational big data. The Hadoop cluster is primarily used to store and process data, however, in this implementation, the Hadoop and Spark software frameworks are used for processing the educational big data. The reason for not using the Hadoop cluster nodes for storing the entire educational data, is that as the data volume increases the nodes will not be able to scale and additional storage per node is required. Also, utilizing the data lake allows frequent data updates, scalability, reliability and availability among storing in the clustering nodes.

C. EDUCATIONAL BIG DATA QUERYING

At this stage, the data has been stored in the data lake and the cluster nodes are ready to receive workloads to perform on data. Querying the stored data can be accomplished by utilizing querying components. In this implementation, three querying components that enable batch and real-time processing of educational data are utilized. The required data is first extracted from the data lake and stored in the designated Hadoop cluster nodes. This enables scalability and availability of the required data. Each component differs in the way it queries data. Hive and Impala components are utilized for batch processing workloads, whereas, Spark SQl [28] is utilized for low-latency processing workloads. Once the data is processed, it can be removed from the Hadoop cluster and stored back into the data lake. For example, in a data cleaning task, the data is extracted from the Raw Data Zone and stored into the Hadoop cluster, once the data is cleaned, it can be stored into the Refined Data Zone, and removed from the Hadoop cluster.

D. EDUCATIONAL BIG DATA ANALYTICS

The data analytics stage is where useful information and insights are extracted from the educational data that was ingested previously into the data lake. This includes batch and real-time applications that promote the operation of the educational organization. In this implementation, components that allow batch and real-time applications are setup and can be run on top of the querying layer or directly on data stored in the data lake. The components utilized are namely, Radoop, Tableau Desktop, Tableau Prep Builder and Matlab. These components can cover applications including, querying, cleaning, data mining, statistics, and visual analytics. It should be noted that many other analytical components can be used on top of this platform, however, the components illustrated in this implementation are able to perform various educational big data analytical applications, and are sufficient enough for non-data scientists and people in the educational organization to perform their educational applications. The components that run on top of the processing layer connect to the platform through Open Database Connectivity (ODBC) drivers [29]. This connection enables to access the data lake and run queries utilizing the querying layer components. Further, this cloud implementation of the educational big data platform, allows to perform applications remotely through the cluster nodes.

IV. PRACTICAL APPLICATIONS OF THE EDUCATIONAL BIG DATA PLATFORM

In this section, the presented educational big data platform is tested by performing analytical applications on the available data. As mentioned previously, the platform serves as a test-bed to develop and validate end-to-end educational applications for Umm Al-Qura University, and as various types of data become available, the platform is set to handle and perform analytical applications. In the first application, an exploratory data analytical study is presented, were 74,314 student demographic data are ingested into the data lake. Then data mining methods are applied to discover regulations and patterns within the data. In the second application, association rule mining techniques are applied to build a system that recommends the courses a student should enroll-in to potentially achieve a higher grade.

A. MINING STUDENT DEMOGRAPHIC DATA

In this application, it is desired to apply data mining algorithms in order to discover hidden information within students' demographic data. The selected student demographic data includes the student's department, gender, attended high-school, GPA, taken courses, high school average, current academic status, pre-admission assessment test and academic achievement test (pre-admission selected high-school subjects test). For each faculty, the decision tree analysis CART [30] studies the data of students in each faculty and finds new patterns that were not obvious using statistical methods. CART data mining algorithm uses a metric called Gini index for determining the most prominent feature that best divides the data. The Gini index shows the most influential factors of demographic information that divides the students. In this first application, the data concerning the past five years of 4,877 students in the Faculty of Computer and Information Systems is concerned. The CART data mining algorithm is applied on 3,223 students records that have all the aforementioned demographic data (no missing data). The significance of demographic data according to the Gini index are:

1) GPA: 45.68%
2) High School Average: 23.85%
3) Pre-admission Assessment Test: 17.48%
4) Academic Achievement Test: 12.98%

The decision tree is partially shown in Fig. 3. From the resulted decision tree, it can be observed that the GPA of 1.8 is critical in determining the success and failure of students. Students in this faculty having a GPA less than 1.8 are likely to drop-out of the program with a probability of 54%. Further, the probability of drop-out increases up to 90% for students that have a GPA less than 1.15 (Fig. 4), although, 88% of those students achieved relatively high marks in the academic achievement test. From this, it is evident that students who were successful during high school are not immune to drop-out of university programs. This extracted information is useful for academic advisors to make decisions such as giving more attention to those students by scheduling meetings to know the reasons that may lead to their failure. Then further proper decisions may be taken accordingly to reduce the drop-out rate of students. Such students are of interest for further analysis, and decision makers could look into other data related to those students, such as social media data, to understand the reasons behind this significant change in performance. The educational big data platform features a feedback loop to monitor the effect of the decisions made on the attitude of students. Further, the constructed decision tree

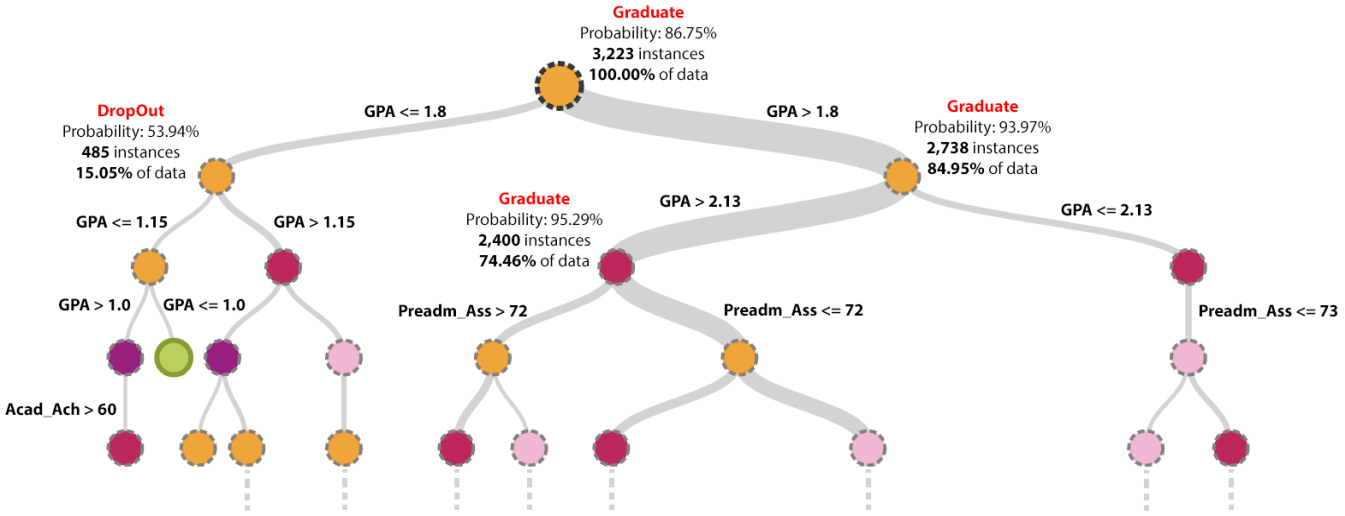

FIGURE 3. Partial view of the resulted decision tree.

TABLE 1. Association rules and corresponding metric values.

Premises	Conclusion	Support	Confidence	Lift
Organizing and Processing Data = A	Human Machine Interaction = A	0.017	0.848	21.6
Intro. To Databases = A	Human Machine Interaction = A	0.010	0.860	21.8
Human Machine Interaction = A, System Analysis and Design = A	Organizing and Processing Data = A	0.010	0.877	43.2
Software Engineering = A	Human Machine Interaction = A	0.010	0.895	22.7
Organizing and Processing Data = A, System Analysis and Design = A	Human Machine Interaction = A	0.010	0.943	24.0

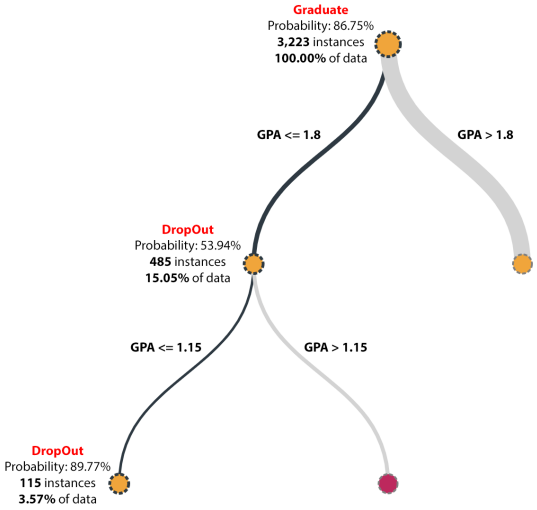

FIGURE 4. Zoomed view of the drop-out branch.

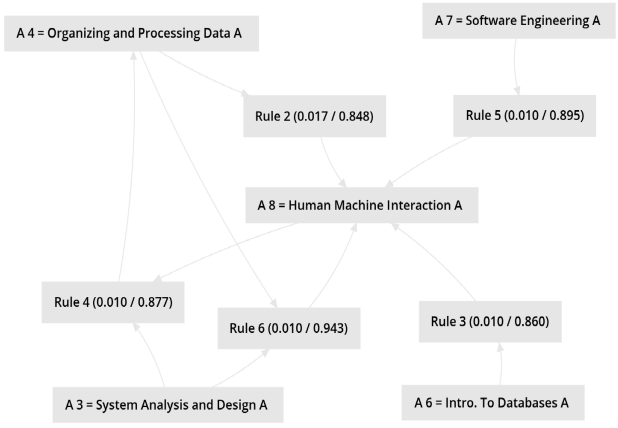

FIGURE 5. Graph representing the resulted top association rules.

can be utilized as a model to predict students that may be at risk of dropping-out. This application can also be applied at the course level, i.e., to extract features of students that are vulnerable to drop-out of a certain course; also, predict and take prior action to reduce drop-out rates.

In the second application to extract further useful knowledge from the data, an association rule mining algorithm namely, FP-growth [31] is applied on five past years for 4,877 students in the Faculty of Computer and Information Systems to find patterns of courses that students are likely to achieve high grades in. This could be useful for students to choose the next course or the elective courses to enroll-in. The associations between courses is evaluated by the following metrics:

1) Support: the number of students that had the courses and achieved high grades as a percentage of the total number of students.
2) Confidence: indicates to how often the extracted rule of those courses occurrence has been found to be true.
3) Lift: is the probability of all those courses occurring together in a rule; greater lift values indicate stronger

associations between sequences of courses with high grades.

The minimum support value to consider that courses are associated was set to 0.01. Table 1, presents the top five resulted association rules with their corresponding support, confidence and lift values. It was observed that many of the extracted association rules had higher lift values which indicates to how adequate the rule is at predicting the result of achieving high grades than just assuming the result in the first place. For example, a student who achieved a grade of "A" in Human Machine Interaction, and System Analysis and Design courses, is likely to achieve an "A" grade in Organization and Processing Data course with a probability of 87%. The association graph of those generated rules is shown in Fig. 5.

The presented educational big data platform is capable of performing further applications, and as the university makes addition types of data available, further analytical applications that potentially promote the educational outcomes can be performed.

V. CONCLUSION

This paper presented a comprehensive platform for educational big data analytics. The objective of the platform was to handle complex educational big data while considering four main contributions to field: 1) an infrastructure that is able to handle the educational big data from data production to analytics. 2) presenting a comprehensive platform to perform various educational data analytical applications. 3) including a feedback loop to monitor the short-term and long-term effects of the decision-making process. 4) introducing an effective environment for non-data scientists and people in the educational organizations to apply their demanding educational applications. The presented platform was implemented on a cloud computing platform and is consistent to the Lambda architecture design and principals, to allow batch and real-time processing of data. Further, the educational big data platform utilized a data lake repository, and the ingested data was organized to avoid data swamping. An implementation of the educational big data platform to comply with the needs of a university organization was presented. The construction of this platform serves as a test-bed to develop and validate end-to-end educational related applications with regards to Umm Al-Qura University, Saudi Arabia, and for educational institutions in general. Furthermore, two analytical applications were presented to test the platform. In the first application, an exploratory data analytical study was presented to discover regulations and patterns within the data. Then a model was built to predict students that may drop-out of the university programs. In the second application, association rule mining techniques were applied to build a system that recommends courses a student may enroll-in, to potentially achieve a higher grade. The decision-makers at the organization may take corrective actions to enhance the educational outputs. The presented educational big data platform features a feed-back loop to monitor the effects of the decisions taken. The impact of this platform surpass the presented applications and as various types of data become available, the platform is capable of performing other analytical applications.

REFERENCES

[1] T. Devasia, V. T P, and V. Hegde, "Prediction of students performance using educational data mining," in *Proc. Int. Conf. Data Mining Adv. Comput. (SAPIENCE)*, Mar. 2016, pp. 91–95.

[2] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Predicting grades," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 959–972, Feb. 2016.

[3] E. Kurilovas, "Advanced machine learning approaches to personalise learning: Learning analytics and decision making," *Behaviour Inf. Technol.*, vol. 38, no. 4, pp. 410–421, Apr. 2019.

[4] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 3, May 2020, Art. no. e1355.

[5] U. K. Saba, S. U. Islam, H. Ijaz, J. J. P. C. Rodrigues, A. Gani, and K. Munir, "Planning fog networks for time-critical IoT requests," *Comput. Commun.*, vol. 172, pp. 75–83, Apr. 2021.

[6] G.-E. Zaharia, T.-A.-I. Şoşea, R.-I. Ciobanu, and C. Dobre, "Machine learning-based traffic offloading in fog networks," *Simul. Model. Pract. Theory*, vol. 101, May 2020, Art. no. 102045. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569190X19301765

[7] A. Shakarami, M. Ghobaei-Arani, and A. Shahidinejad, "A survey on the computation offloading approaches in mobile edge computing: A machine learning-based perspective," *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107496. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128620311634

[8] C. Dede, A. Ho, and P. Atan, "Big data analysis in higher education: Promises and pitfalls," *Educ. Rev.*, vol. 51, no. 5, pp. 23–34, 2016.

[9] N. Manokaran, V. Varathan, and S. Deepak, *Cloud-Based Big Data Analytics in Smart Educational System*. Hershey, PA, USA: IGI Global, 2017.

[10] F. Matsebula and E. Mnkandla, "A big data architecture for learning analytics in higher education," in *Proc. IEEE AFRICON*, Sep. 2017, pp. 951–956.

[11] K. L.-M. Ang, F. L. Ge, and K. P. Seng, "Big educational data & analytics: Survey, architecture and challenges," *IEEE Access*, vol. 8, pp. 116392–116414, 2020.

[12] H. Khazaei, C. McGregor, M. Eklund, K. El-Khatib, and A. Thommandram, "Toward a big data healthcare analytics system: A mathematical modeling perspective," in *Proc. IEEE World Congr. Services*, Jun. 2014, pp. 208–215.

[13] A. A. Munshi and Y. A.-R.-I. Mohamed, "Data lake lambda architecture for smart grids big data analytics," *IEEE Access*, vol. 6, pp. 40463–40471, 2018.

[14] A. A. Munshi and Y. A.-R.-I. Mohamed, "Big data framework for analytics in smart grids," *Electr. Power Syst. Res.*, vol. 151, pp. 369–380, Oct. 2017.

[15] B. K. Jha, G. G. Sivasankari, and K. R. Venugopal, "Fraud detection and prevention by using big data analytics," in *Proc. 4th Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Mar. 2020, pp. 267–274.

[16] S. Qu, K. Li, S. Zhang, and Y. Wang, "Predicting achievement of students in smart campus," *IEEE Access*, vol. 6, pp. 60264–60273, 2018.

[17] L. Huang, C.-D. Wang, H.-Y. Chao, J.-H. Lai, and P. S. Yu, "A score prediction approach for optional course recommendation via Cross-User-Domain collaborative filtering," *IEEE Access*, vol. 7, pp. 19550–19563, 2019.

[18] L. Cen, D. Ruta, and J. Ng, "Big education: Opportunities for big data analytics," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 502–506.

[19] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Dropout prediction in edx MOOCs," in *Proc. IEEE 2nd Int. Conf. Multimedia Big Data (BigMM)*, Apr. 2016, pp. 440–443.

[20] X. Li, X. Fan, X. Qu, G. Sun, C. Yang, B. Zuo, and Z. Liao, "Curriculum reform in big data education at applied technical colleges and universities in China," *IEEE Access*, vol. 7, pp. 125511–125521, 2019.

[21] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia, B. Reed, and E. Baldeschwieler, "Apache Hadoop YARN: Yet another resource negotiator," in *Proc. 4th Annu. Symp. Cloud Comput.*, Oct. 2013, pp. 1–16, doi: 10.1145/2523616.2523633.

[22] N. Marz and J. Warren. (2015). *Big Data: Principles and Best Practices of Scalable Real-time Data Systems. Manning*. [Online]. Available: https://books.google.com.sa/books?id=HW-kMQEACAAJ
[23] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: A warehousing solution over a map-reduce framework," *Proc. VLDB Endowment*, vol. 2, no. 2, pp. 1626–1629, Aug. 2009.
[24] J. Li, "Design of real-time data analysis system based on impala," in *Proc. IEEE Workshop Adv. Res. Technol. Ind. Appl. (WARTIA)*, Sep. 2014, pp. 934–936.
[25] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia, *Spark SQL: Relational Data Processing in Spark*. New York, NY, USA: Association for Computing Machinery, 2015.
[26] Protecting Student Privacy. (Oct. 2012). *Data De-Identification: An Overview of Basic Terms*. [Online]. Available: https://studentprivacy.ed.gov/resources/data-de-identification-overview-bas%ic-terms
[27] Tableau Software. *Tableau*. Accessed: May 26, 2020. [Online]. Available: https://onlinehelp.tableau.com/current/pro/desktop/en-us/help.htm#save_sav%ework_packagedworkbooks.html
[28] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia, "Spark SQL: Relational data processing in spark," in *Proc. 2015 ACM SIGMOD Int. Conf. Manage. Data*. New York, NY, USA: Association for Computing Machinery, 2015, pp. 1383–1394.
[29] (2016). *Cloudera ODBC Driver for Impala*. Cloudera, Palo Alto, CA, USA. Accessed: Jan. 2, 2020. [Online]. Available: http://www.cloudera.com/documentation/other/connectors/impala-odbc/latest/Cloudera-ODBC-Driver-for-Impala-Install-Guide.pdf
[30] W.-Y. Loh, "Classification and regression trees," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
[31] C. Borgelt, "An implementation of the FP-growth algorithm," in *Proc. 1st Int. Workshop Open Source Data Mining Frequent Pattern Mining Implement. (OSDM)*, 2005, pp. 1–5.

AMR A. MUNSHI (Senior Member, IEEE) received the B.Sc. degree in computer engineering from Umm Al-Qura University, Makkah, Saudi Arabia, in 2008, and the M.Sc. and Ph.D. degrees in computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2014 and 2019, respectively.

He is currently an Assistant Professor with the Department of Computer Engineering, Umm Al-Qura University. His research interests include artificial intelligence, data mining, smart grids, and big data analytics. He is a member of the Golden Key International Honor Society. He also serves as an Editor for the *Alberta Academic Review* journal.

AHMAD ALHINDI (Senior Member, IEEE) received the B.Sc. degree in computer science from Umm Al-Qura University (UQU), Makkah, Saudi Arabia, in 2006, and the M.Sc. degree in computer science and the Ph.D. degree in computing and electronic systems from the University of Essex, Colchester, U.K., in 2010 and 2015, respectively. He is currently an Associate Professor of artificial intelligence (AI) with the Department of Computer Science and a Researcher of CIADA, UQU. He is also involved in AI algorithms, focusing particularly on machine learning and optimization with a willingness to implement them in the context of decision making and solving combinatorial problems in real-world projects. His current research interests include evolutionary multi-objective optimization and machine learning techniques.