ISSN 2278-3091
Volume 9, No.1.1, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse6891.12020.pdf
https://doi.org/10.30534/ijatcse/2020/6891.12020

A Distributed Big Data Model for Education Sector

A.Fairuzullah[1], Mohd Helmy Abd Wahab[2], A.Noraziah[3], M. Zarina[4]
[1]Faculty of Computer systems & Software Engineering, Pusat teknologi Maklumat & Komunikasi,
Univerisiti Malaysia Pahang, 26300 Kuantan, Pahang, Malaysia. ;
[2]Department of computer Engineering, Faculty of Electrical and Electronic Engineering,
Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, Malaysia;
[3]Faculty of Computer systems & Software Engineering,IBM Centre of Excellence,
Univerisiti Malaysia Pahang, 26300 Kuantan, Pahang, Malaysia. ;
[4]Faculty of Informatic and Computing,
Universiti Sultan Zainal Abidin,22200, Besut, Terengganu, Malaysia.
*fairuzullah@ump.edu.my

ABSTRACT

Many higher institutions are faced with the problem of managing the ever-increasing volume of data that involves Information Architecture. Most institutions have established long-standing data warehouses and have even deployed several analytics tools, but with the increase in the competition for gifted students increases, while education costs keep the number of potential students low, many institutions devising more ways of analyzing potential students and how to manage students' experience as they are enrolled. Analytics is very important in the thorough analysis of students and their learning data to arrive at better decisions on the courses to be offered in the future. As well as their mix to accommodate both the existing and potential students. With the Big Data systems, Information Technology is positioned to ensure a holistic education system in the aspect of improving the decision-making process compared to the other areas. The predictive analytics and forecasting models used in Big Data guides institutions into making the right investment decisions for a higher institutional impact; data analytics also impact the knowledge domain. This paper proposed a distributed Big Data model for the education sector.

Key words : Distributed, Big Data, Education,

1. INTRODUCTION

The need for cost-effective information processing technologies is ever increasing due to the abundance of several information assets which needs cost-effective measures to be processed. Such innovative information processing techniques must ensure an enhanced decision making, insight, and process optimization, and must be fully utilized in ways that can transform service design to specifically meet the needs of the education in a timely manner. The strategy ought to deliver on the core aspects of Information Communication Technology Strategic Planning [3]. An emerging Big Data (BD) term has opened broad impact, thus, intensely remain to attract various attentions from both scientific experts and the public community in general.

Providing high data availability is very important to enable organizations access to current data where and when they need it [17, 18]. A reliable system is one that can continue to process user requests even when the underlying service is unreliable [19]. Cloud computing provides responsive and scalable resource admission in a utility-like fashion especially for the processing of BD. Cloud is also advantageous in the provision of more fault-tolerant scalable services with better performance [20]. Cloud computing has recently become a label certain types of the data center, or mostly, a group of datacenters [21]. Big Data mainly focus on practicalized service implementation and policy design which encourages a personalized and seamless student-university interaction. As the awareness of the advantages of BD keeps growing, an upsurge in public debate on the trade-off between the benefits and the limitations associated with BD technology is expected, especially in the agencies of the Malaysian Public Education Sector. Several key areas which can be influenced by BD analytics are investigated [4].

Several organizations manage their operations using an integrated database application [16]. Agencies can save time and money if they implement smarter data management techniques which are conscious of the needs of BD analysis. Sourcing data from different organizations or areas benefits several agencies in different ways, especially when there is transparency. Service personalization through BD analytics may add value as it will present a better picture of a university group or an individual student. Human behavior can also be analyzed using BD based on its characteristic granularity. The unification of predictive and problem-solving analytics from several datasets with advanced analytics technologies will enhance their problem-solving capabilities and improve the capability of predictive analytics in offering insights that can encourage decision-making. It is therefore certain that these will provide a basis for several parallel analysis and computations for the extraction of relevant patterns

from educational data. Educational institutions are deploying BD analytics tools to improve their standard and manage their large volume of students' profiles [5].

This paper canvases the adoption of BD analytics as a part of the "next-generation" frameworks which can meet the demands of higher education institutions. The method and guidance provided by several customer projects and the highlights on the decisions that customers faced architecture planning and implementations Oracle's enterprise architecture approach and framework are articulated in the Open Architecture Development Process (OADP) and the Open Enterprise Architecture Framework (OEAF) [1]. There is an increase in the number of unstructured and machine-generated data such as photos, social media feeds, and videos compared to the structured data. As such, more than 80% of all databases are currently unstructured and there is a need for new technologies which can be deployed to access and manage these datasets [2].

2. TECHNOLOGY AND ANALYTICAL SYSTEMS

Over the years, knowledge management has mainly aimed at developing the capacity to combine information from different sources in order to provide the required insights for an efficient decision-making. The education does not consider just one factor such as student mark or expertise when making decisions. Previously, the collection and storage costs of information restrict the ability of firms to acquire the whole information required to paint a perfect picture. However, such restrictions have been solved by the availability of automated digital information collection systems and the option for cheap information storage. Although there is an abundant data availability currently, however, the relational databases are already exhausting their capabilities in making sense of the available information. University Malaysia Pahang (UMP) is one of the public sector education institution in Malaysia whose mission is "To advance knowledge and educate students in science, technology, engineering and other areas".

The University is committed to generating, disseminating, and preserving knowledge, and to working with others to bring use this knowledge to solve global challenges. UMP is committed to the provision of education which is a combination of rigorous academic commitment and the excitement of new knowledge discovery to its students with the support of a diverse campus community.

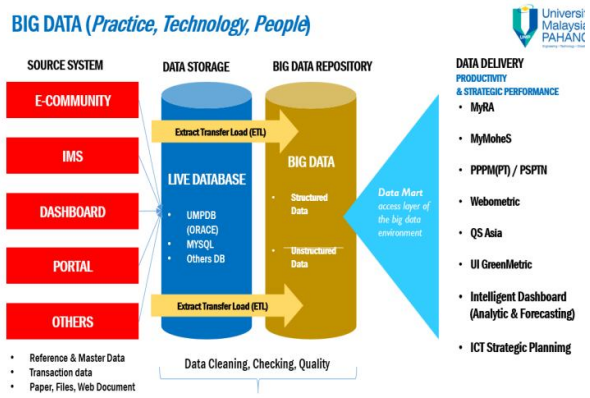

Figure 1: Data Delivery Integrated Education Management System at UMP

Big Data challenges will continue to be difficult with time as the data volume is already huge (almost 10 terabytes) and keeps increasing daily. The rate of data generation is increasing, and this rapid growth is mainly due to the increase in internet connectivity. Furthermore, there is an increase in the types of data generated despite the limited capability of organization to capture and process these data. The current data management methods have failed to handle the large volume of data being generated and there is a need for organizations to devise other means of data management [6].

3. DATA STORAGE: STRUCTURED, SEMI STRUCTURED UNSTRUCTURED DATA

Microsoft developed the Microsoft SQL Server as a relational database management system. This management system is deployed in Structured Data. The server of the database is a software whose major function is to store and retrieve data upon request. The SQL Server 2016, through its SQL Server R Services, extended support for BD analytics and other related analytics applications, making it possible to run analytics applications which are written in the open source R programming language and PolyBase on DBMS. PolyBase is a technology which allows SQL Server users to access and analyze data which are stored in Azure blob storage or Hadoop clusters.

The NoSQL database provides a mechanism for the storage and retrieval of semi-structured and unstructured data. It is modeled in means except for the tabular relations deployed in relational databases. This motivation of this approach includes its design simplicity, finer control over availability, and horizontal scaling. The MongoDB is a cross-platform document-oriented database which belongs to the class of NoSQL databases. The MongoDB favored the JSON-like documents structure with dynamic schemas in place of the conventional table-based relational database structure, making it faster and easier to integrate data in certain types of applications.

Feature of MongoDB

Figure 2 shows the features of MongoDB that suits Big Data for the education sector

Mongo DB supports Map reduce and Aggregation Tools
Java Scripts are used instead of Procedures
Mongo DB is a schema-less Database
Most Importantly Mongo DB supports secondary indexes and geospatial indexes.
Simple to Administer the Mongo DB in cases of failures
Mongo DB designed to provide High Performance
MongoDB stores files of any size without complicating your stack.

Figure 2: Characteristic MongoDB for Big Data Model For Education Sector

MongoDB: JSON document store

JSON is a format which can be easily read by humans; machines can also efficiently parse it. For instance, a business card in Mongo would look like this:

Query Language

Select _id, name, cell, fax, address from staff_info

{"_id" : ObjectId("4efb731168ee6a18692d86cd"),
"name" : "A.Fairuzullah",
"cell" : "+60199808507",
"fax" : "+6095492199",
"address" : " Lebuhraya Tun Razak,26300,Gambang, Malaysia "
}

Command Mongo Client

To create the database for storing the above profiles, it's needed to give the following commands on the Mongo Client.

db.corporatereocrd.save (name_A);
db. corporaterecord.save (name_B);
db. corporaterecord.save (address);

4. MAP REDUCE AND PARALLEL RDBMS

MapReduce is one of the unstructured data management methods which emerged as a general tool for a specific form of parallel computing. While "Swiss army knife" RDBMS solution works fine for the specific tasks it was designed to handle, MapReduce can work for virtually any form of parallelizable problem [9].

Universality. One of the positives of MapReduce is its universality. It particularly provides a software engineer with an efficient low-level data access, making it possible to handle data that is completely unstructured. Many graph algorithms, for instance, can be implemented easily in MapReduce, but this is not possible with the general purpose databases. This is a well-known problem and the reason for the existence of specialized graph databases like Neo4j. It demands time and software experience in learning how to use MapReduce; good software engineers can learn it within a short period [10, 11]. This is why the learning curve and cost of skillful engineers are not issues for the reputable companies like Google and Facebook.

Customization. The base Hadoop layer in its 10 years of existence has been extended by several frameworks which can run on top of it. Spark (improved the raw Hadoop efficiency) and Apache Storm (for streaming support) are some of the examples of such free frameworks. These efficiencies have sidelined most of the criticism on the inefficiency of Hadoop. After replacing the higher-level Hadoop layers which are inefficient, only the HDFS remained unaltered [12-15]. Google and Facebook are running customized versions of Hadoop/MapReduce with secret details. However, despite the secrecy of these customized versions, there is still news of user data being crunched daily in such systems.

Support of programming language. Any programming language can be used in Hadoop Streaming one. Open source. Apache Hadoop is an open source version of MapReduce which can be easily learned and implemented.

5. DATA ANALYSIS IN EDUCATIONAL EVALUATION

Figure 3 shows the data analytics in education sector. The increase in demand for online learning is driving the need for learning analytics which can capture the interaction pattern of learners with content as well as their discourse around the learning materials. For instance, a significant amount of data such as time spent on a resource or the frequency of posting on a social network can be captured by learning management systems such as Moodle or Desire2Learn. Such data is somehow similar to the website traffic data captured by Google Analytics or Piwik. This data is used by the new generation of tools, like SNAPP for the analysis of social networks, peripheral learners, and degrees of connectivity. Discourse analysis tools, such as those being developed at the Knowledge Media Institute at the Universiti Malaysia Pahang mainstreams.

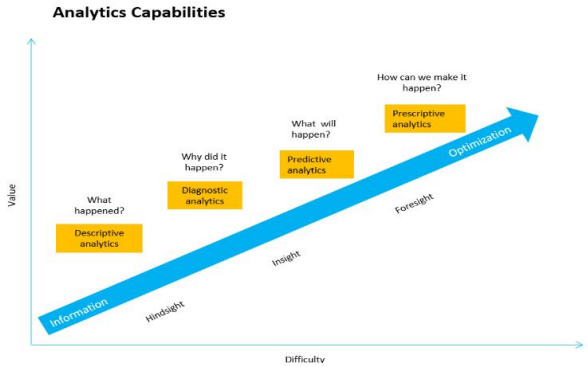

Figure 3: Data Analytics in Education Sector

6. CONCLUSION

The variety and large volumes of real-time data (known as Big Data) provides the educational sector with useful activity insight as they generate huge data volumes in the form of grades, admission, test scores, enrollment numbers, etc. The introduction of online courses by many institutions has increased the amount of available data. Various analytical software and data mining approaches help in the identification of the relevant pedagogic approaches. However, the BD paradigms are currently needed to support the existing data mining methods to increase the efficiency of educational institutions. This paper proposed the use of MongoDB, SQL Server 2017, and MapReduce data storage platforms for the analysis of educational data. Future perspectives can focus on the use of other BD platforms such as Polybase in Microsoft NoSQL and MongoDB.

ACKNOWLEDGMENTS

We appreciate the Malaysian Ministry of Higher Education for providing the fund for this study under Fundamental Research Grant Scheme RDU140101. We also appreciate the Research Management, Innovation and Commercialization Centre (RMIC, UniSZA) and University Malaysia Pahang for the Internal Grants under RDU170398 and PGRS170304. The authors are also grateful to Universiti Malaysia Pahang for the opportunity to plan and execute this project.

REFERENCES

[1]. Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In ACM SIGMODRecord (Vol. 22, No. 2, pp. 207-216). ACM. https://doi.org/10.1145/170036.170072

[2]. Bakharia, A., Heathcote, E., & Dawson, S. (2009). Social networks adapting pedagogical practice: SNAPP. Same Places, Different Spacesascilite 2009.

[3]. Bhullar, M. S., & Kaur, A. (2012). Use of Data Mining in Education Sector. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 24-26).

[4]. Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An

[5]. issue brief.US Department of Education, Office of Educational Technology, 1-57.

[6]. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[7]. Koedinger, K. R., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An Open Repository and analysis tools for fine-grained, longitudinal learner data. EDM, 157, 166.

[8]. Luján-Mora, S. (2006). A survey of use of weblogs in education. Current developments in technology-assisted education, 1, 260-264.

[9]. Marr, B. (2014). A Talk on Big Data - the 5 Vs Everyone Must Know.

[10]. Moturi, C. A., & Maiyo, S. K. (2012). Use of MapReduce for Data Mining and Data Optimization on a Web Portal. IJCA, 56(7), 39-43. https://doi.org/10.5120/8906-2945

[11]. Nicol, D. J., & Macfarlane Dick, D. (2006). Formative assessment and self regulated learning: a model and seven principles of good feedbackpractice.Studies in higher education, 31(2), 199-218.

[12]. Parry,M. (2012). “Pleased Be eAdvised,” New York Times Education Life, p.25.

[13]. Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

[14]. Wassan, J.T. (2014). Discovering Big Data Modelling for Educational World, Procedia ScienceDirect- Social and Behavioral Sciences 176 (2015) 642 – 649

[15]. Toh, M. Zulfahmi, Noraziah A., Omardin M.A., Zainuddin N., 2016. Managing building checklist plans using BUSCLIS. International Journal of Software Engineering and Computer Systems 1(2), pp. 74-88.

[16]. M. Mat Deris, D. J. Evans, M. Y. Saman, A. Noraziah, (2003). Binary Vote Assignment on Grid for Efficient Access of Replicated Data: Intl. Journal of Computer Mathematics, Taylor and Francis, 80(12): 1489-1498.

[17]. A. Noraziah, M. Mat Deris, NA Ahmed, R. Norhayati, MY Saman, MA Zeyad, (2007). Preserving data consistency through neighbor replication on grid daemon. American Journal of Applied Science 4 (10), 748-755.

[18]. MM Deris, M Rabiei, A. Noraziah, HM Suzuri, (2003). High service reliability for cluster server systems IEEE International Conference on Cluster Computing (Vol. 1, pp. 280-287).

[19]. AAC Fauzi, A. Noraziah, T Herawan, NM Zin, (2012). On cloud computing security issues Asian Conference on Intelligent Information and Database Systems (Vol 7197 pp. 560-569).

[20]. N Khan, A Noraziah, EI Ismail, MM Deris, T Herawan, (2012). Cloud computing: Analysis of various platforms International Journal of E-Entrepreneurship and Innovation (IJEEI) 3 (2). https://doi.org/10.4018/jeei.2012040104