International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:9, No:5, 2015

Applications of Big Data in Education

Faisal Kalota

***Abstract*—**Big Data and analytics have gained a huge momentum in recent years. Big Data feeds into the field of Learning Analytics (LA) that may allow academic institutions to better understand the learners' needs and proactively address them. Hence, it is important to have an understanding of Big Data and its applications. The purpose of this descriptive paper is to provide an overview of Big Data, the technologies used in Big Data, and some of the applications of Big Data in education. Additionally, it discusses some of the concerns related to Big Data and current research trends. While Big Data can provide big benefits, it is important that institutions understand their own needs, infrastructure, resources, and limitation before jumping on the Big Data bandwagon.

***Keywords*—**Analytics, Big Data in Education, Hadoop, Learning Analytics.

I. Introduction

DATA driven decision making has been used in businesses and other sectors for a while now. Big Data, which can be utilized in data driven decision making, has become a buzz word in recent years. Other terms that have gained popularity in recent years include analytics such as Business Analytics, Learning Analytics, Risk Analytics, etc. Big Data and Analytics are utilized by different industries for decision making. The objective of this paper is to provide an introduction to Big Data and some of its applications in Education.

II. Big data

Big Data has become a buzz word in recent years and has emerged as an emerging technology. An emerging technology is something that can "Lead to the rapid development of new capabilities" and "Create new opportunities for and challenges to addressing global issues" [1]. For example, sports teams are utilizing data during the game to adjust their game plan; Wearable technology provides data that teams use to monitor stadium experience or player performance [2]. As another example, Big Data analytics allowed Conservation International to monitor the biodiversity of rainforest and help protect the environment [3].

So what exactly is Big Data? Big Data is defined in terms of three Vs; they are Volume, Velocity, and Variety [4]. So it is the sheer volume of data that is being created at an alarming velocity which is in various formats that is collected through various channels such as smart phones, social media, business transactions, people's travel patterns, etc. in digitized format. In every two days the amount of information that is created is equal to what has been created from the dawn of the civilization to 2003, which can amount to approximately five Exabyte [5].

So there may be a question as to what exactly is the difference between Big Data and the ordinary Small Data. There are 10 differences between Big Data and small data as summarized in Table I [6].

A. Big Data, Analytics and Business Intelligence

The purpose of this section is to clarify the relationship between Big Data, Analytics, and Business Intelligence (BI). Big Data is defined as "Dataset with volumes so huge that they are beyond the ability of a typical DBMS [Database Management System] to capture, store, and analyze. The data are often unstructured or semi-structured" [7, p. 607].

Business Intelligence and Analytics complement one another; however, getting BI does not guarantee that you are also going to get Analytics. BI describes the current state of affairs. It is a "retrospective analysis that provides a rearview mirror view on the business reporting on what happened and what is currently happening" [8]. Whereas, analytics or Predictive Analytics is "forward-looking analysis: providing future-looking insights on the business-predicting what is likely to happen" [8].

So BI and Analytics both utilize data, which may be Big Data as well, and allows organizations to understand what is currently going on in the organization and also focus on the what-if type of questions. For example, Universities may use BI to describe the current student demographics in terms of their age, gender, major, GPA, etc. On the other hand they can use analytics or learning analytics to predict future events such as future freshmen class; predicting student success or failure, etc.

Analytics come in different forms such as Business Analytics, Risk Analytics, Learning Analytics, and others. According to the 2014 Horizon Report, the time to adoption for Learning Analytics in Higher Education is one year or less. Research in Learning Analytics requires the analysis of educational data to "deliver personalized learning, enable adaptive pedagogies and practices, and identify learning issues in time for them to be solved" [9, p. 38]. In order for Learning Analytics to work; data is needed; which may come in the form of Big Data.

III. Data Science, Big Data Technologies, and Analytical Methods

This section provides an overview of Data Science, Big Data Technologies, and some of the related Analytical Methods.

Faisal Kalota is a CIS faculty member at the Higher College of Technology at the Dubai Men's Campus in UAE (e-mail: fkalota@hct.ac.ae).

TABLE I
DIFFERENCE BETWEEN SMALL DATA AND BIG DATA [6]

Category	Small Data	Big Data
Goals	Usually designed to answer or address a particular goal	It may have an initial goal but it would evolve over the period of time.
Location	Data stored is on one computer or server	Data is stored across multiple servers that may be spread out across geographic locations.
Data Structure	Data is highly structured	Data is not highly structured and may be of different types, and sizes.
Data Preparation	Often the data users prepares their own data.	Data is compiled by different people from different sources. The end users of the data may not be the same people who compiled the data.
Longevity	Data may not be stored permanently, and may be discarded after the initial objective has been met.	Data is stored perpetually. Many extend into the future and past.
Measurements	A single experimental protocol is used to measure the data.	Since data are of different types and formats, the units of measurement vary; and data quality may be hard to verify.
Reproducibility	Projects and the results be can often be reproduced.	Reproducing the same project may not be possible because of the daunting amount of data which is constantly getting updated.
Stakes	Project costs are limited, particularly if there is a failure.	Big Data projects can be expensive, and failed projects may even lead to bankruptcy.
Introspection	Since the data is structured and can be easily organized, individual data points are identified by their row and column location within a spreadsheet or database table	Big Data is unstructured. Therefore the organization of Big Data in some cases may not be structured. Accessing the data, their values, and additional information is achieved using a technique called Introspection
Analysis	Since the data is structured and may be small, their analysis can be easily done.	Big Data analysis are often complicated. They may require supercomputers, parallel processing, and the analysis are done in incremental steps.

A. Data Science Facets and Skills

Data science involves three skills: Math & Statistics Knowledge, Substantive Expertise, and Hacking Skills [10].

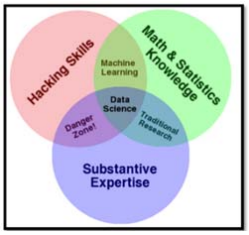

Fig. 1 Data Science Venn Diagram

Fig. 1 represents these skills through a Venn diagram [10]. These skills are valuable on their own; and when combined can provide additional capabilities. Here hacking skills are implied as programming or coding skills. Substantive experience refers to a particular domain such as Health Care, Marketing, etc; and as the name implies having statistical and math skills fall under the final category. Having only two of the related skills may fall under one of the intersecting areas. Danger Zone includes the people who "Know enough to be dangerous" [10]. So they may be good at programming, extracting data, and have good domain knowledge, but may not know a lot about properly interpreting the data. This may be counterproductive in some instances.

B. Big Data Technologies

This section will provide a brief overview of some of the technologies that are used for Big Data and Analytics. Some of the key technology enablers in the Big Data industry are SAS, Microsoft, IBM, and Oracle. However, they are not the only vendors and by no means are the following the only technologies. The sections below are broken down into different sections to discuss different technologies that are utilized for Big Data and Analytics.

1) Storage

Big data is about the three Vs; one of which is volume. The massive amount of data warrants a non-traditional method of data storage. The data must be stored in a distributed environment and must be mirrored as well. This means that not all data is on the same physical device and must be mirrored. From a cost perspective this may not be a viable option for many organization; hence, the significance of cloud storage.

One of the most popular cloud storage vendors is Amazon [11]. Amazon's storage solution is known as Amazon Simple Storage Service (Amazon S3). It is a scalable solution that provides unlimited data storage, at a cost of course. However, the scalability allows the clients to configure options such as storage space, level of redundancy, or the speed at which the data can be retrieved. Since there is no upfront cost related to installation and maintenance, this becomes a lucrative option for many clients to store their data.

2) Cloud Computing

Cloud computing provides computing services through the Internet. These services include Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS). Another type of service that the cloud providers provide is called Data as a Service (DaaS). Like other cloud services, this type of service allows the clients to focus on the data analysis for their use rather than focusing on data collection and data quality. Some of the providers of DaaS include Oracle, EMC, and Factual.com.

a) Hadoop

Hadoop is an open-source framework of application for Big Data. While Hadoop can be installed on any system; often it is used as cloud service. Some of the providers for Hadoop services are Amazon, Cloudera, EMC, Hadapt, Hortonworks, IBM, Informatica, Karmasphere, MapR, Microsoft, and Oracle [12].

Hadoop contains various components or modules such as HDFS (Hadoop Distributed File System). HDFS spreads a

piece of information across multiple nodes. Unlike a traditional database it may contains hundreds of thousands of files that are linked but are distributed across multiple storage entities.

MapRedue is a process for mapping and reducing. Map splits a task into pieces because the hardware may have limited resources. The Reduce process takes the output from the different systems and provides it as a single meaningful output. YARN, which stands for Yet Another Resource Negotiator, is a replacement for MapReduce. It provides additional more robust functionality. PIG is a platform that allows one to write MapReduce Programs. Like any programming language, PIG's language is known as Pig Latin Programming Language. Hive is another module, which allows summarizing queries and data analysis. It also has its own language known as HiveQL.

C.Analytical Methods

Different analytical methods can be applied for Big Data Analysis. These include but are not limited to Basic Statistical Procedures, Visualization, Classifiers, Cluster Analysis, and Association Rule mining. This section provides a brief overview of some of these methods.

1) Basic Statistical Procedures

Tremendous amount of data are generated as a result of various transactions that are made in an online course. These data could also include metadata, which is basically data about the data. The metadata may contain information such as interaction identification, session identification, user identification, page views, session information, task information, and activities [13]. The online generated data could be on different scales such as Nominal, Ordinal, Interval and Ratio. Depending on the goal and data type, one can apply both descriptive and inferential statistical procedures for data analysis. Descriptive statistics can include procedures such as measures of central tendency such as mean, median, mode; dispersion such as range, interquartile range, variance, standard deviation; and shape of distribution such as skewness, kurtosis, and modality [13]. Inferential Statistics, used for Hypothesis testing, could include parametric analyses and nonparametric analyses.

2) Information Visualization

Information Visualization is an area within the field of Human Computer Interaction (HCI). Information Visualization is defined as "computer-generated graphics of complex data that are typically interactive and dynamic" [14, p. 181]. It provides a means of amplifying human cognition, enables them to see patterns, trends, and anomalies in the visualization to gain insight [15]. Visualization maps are often interactive that allows one to drill-in and out to by zooming in and out. In educational environments visualization of user data can provide educators insight into the students' activities in an online environment [16]. For example in MOOCs, that often have thousands of students, a visualization map can provide a first glance at user data. This can provide educators to explore student activities and adjust their instructions. Additionally, information visualizations can also support student by helping them monitor their own performance so that they can take necessary corrective actions [16].

3) Classifiers

Classifiers are a form of data mining technique. A classifier is a "model that predicts that class value from other explanatory attributes" [17, p. 57]. The general idea behind the classifier is a two-step approach. First you must choose a classification method such as Decision Trees, Neural Network or Bayesian Network. Then you select a data set with known class values. The dataset is divided into a training set and test set. The training set develops a classifier based on an algorithm, then the classifier is tested with the test set, however, you would hide the class values. If the classification works accurately on the test set; one can assume that it would work on future data; otherwise, adjustments need to be made.

4) Cluster Analysis

Cluster analysis, often used for explanatory analysis, group cases based on a target variable in such a way that the degree of association between the target variable is maximum if they belong to the same group and vice versa [18]. Clustering can be either hierarchical or non-hierarchical. It provides the benefit of utilizing click-stream data that is collected automatically from the logs such as server logs, LMS logs, etc. However, caution must be exercised in terms of selecting the appropriate clustering algorithm [18].

5) Association Rule Mining

Association rule mining discovers relationship among various attributes in a database. It produces IF-ELSE statements related to the attribute relationships [19]. The goal is to uncover relations among different variables in a database that may appear to be unrelated. For example a supermarket may store data about various customer transactions as listed in Table II. The 1s or 0s under each product represent whether the product was purchased or not. So for Transaction 3, the customer purchased soap and detergent. The intent is to check the relationships among different product purchases. So for example, it may be the case that if a customer purchases Shampoo and Soap, then there is a 75% chance that the customer would also purchase toothpaste. With each transaction you calculate the percentages; for example, if a transaction contains Shampoo, then in 40% of the cases Toothpaste is also part of that transaction.

TABLE II
SMALL DATABASE WITH TRANSACTIONS

Transaction ID	Shampoo	Soap	Toothpaste	Detergent
1	1	1	1	0
2	1	0	1	0
3	0	1	0	1
4	1	0	0	0
5	0	0	0	1

Association Rule Mining can be applied in online learning environments as well. For example you can have a database with the different type of transactions in an online learning

environment such as logging information, discussion boards, grades, quizzes, etc. Based on these transactions various rules can be formulated. Then for different cases and combinations of attributes the percentages can be calculated.

IV. Applications of Big Data

Big Data and Analytics can be applied to various settings within higher education such as administrative and instructional applications, recruitment, admission processing, financial planning, donor tracking, and student performance monitoring [20]. This section provides a brief overview of some of the applications of Big Data, Analytics and Data Mining.

A. Student Success

With the ever gaining popularity of Massive Open Online Courses (MOOCs), there opens up an opportunity to collect and analyze large amount of data. These data can be used to analyze the relationship among various variables to predict student success or student dropouts. Additionally, most of higher education institutions utilize an Enterprise Resource Planner (ERP) and Learning Management System (LMS). Both of these technologies collect data that can be mined for various purposes.

Analytics can be used to prevent dropout [21] by predicting the factors that may lead to student dropout. For example researchers performed survival analysis to predict student dropout rates on sample of N = 14791 [22]. They found that student performance was a significant predictor of student dropout. Additionally, they found that age was another predictor; with older students more likely to dropout. Military and married students were also less likely to dropout than non-military and unmarried students, respectively [22].

Analytics can spot various outliers that can be utilized for early intervention [21]. For example student attendance and course performance can be easily monitored. Willging & Johnson found that some of the students who dropped out of an online course did not have a set schedule for study time [23]. While the authors cited that their results may not be generalizable due to the small sample size; but it provides a possible variable that should be considered. Data mining techniques can be utilized to see a pattern of students' logging on to the system on a regular basis as well as based on a specific schedule.

There is a correlation between the student satisfaction with their teachers and student success [21]. Data can be mined to understand what makes a great teacher. These data could include, student evaluations, teachers' logging information in the LMS, teachers activities within the LMS, how the teachers facilitates a course using the LMS, etc. These data can be utilized to improve teacher education and training.

Analytics provide academic institution to compare students' performance among their peers, other schools, and district and at the national level [21]. Although, in this case the white paper discusses the implications related to K-12 education, it is equally applicable in higher education as well. For example, data from the same course can be compared to archived data from over the years. This may not have been possible, if it's not done electronically.

With large amount data, data mining becomes possible. For example, researchers applied the techniques of mirroring visualization, sequential pattern mining, and clustering to identify problems in group environments, improving group monitoring, allowing the students to monitor their own progress, and work effectively in groups [24].

Godoy & Amandi recommend an approach to provide students e-learning material based on their profiles, which is based on their interaction with the e-learning system [25]. By using Big Data, such a technique can be applied to the massive data that is generated from online and/or hybrid courses. Based on which student profiles can be built and would create an opportunity to provide customized e-learning material to the students.

B. E-books and Mobile Devices

E-text books are becoming a norm for many students. According to the Horizon Report the time for adoption for Electronic Books was less than a year [26]. This provides an opportunity for additional data mining. Publishers can gather and mine data about the book usage, course content, content presentation, etc. For example, data can be mined to observe student patterns; such as, how much time the students may be spending on a certain page or content.

In one study data mining was used along with Bloom's Taxonomy to understand students' behavior patterns in reading comprehensions tasks [27]. Although, the study was not specified to E-Books, the study can be applicable to E-Books and mobile devices. It is possible to gather students' data related to their reading patterns; and customize instructions based on it. However, this would also involve the applications of machine learning, to adapt to the students' needs.

C. Finance and Budgets

Market Entry Strategies are used in businesses to evaluate how a business can enter different markets sectors. They generally develop various business models and analyze the results to assess the feasibility to enter the market. Likewise, academic institutions can collect data and utilize analytics to enter new markets. Such models may also help them plan their future operations more reliably.

V. Concerns and Considerations

There are some concerns that should be considered for implementing Big Data and Analytics. They are related to ethical considerations, capturing data electronically, and lack of experts in the area of big data and analytics [20].

With Big Data a large amount of data is captured, which may come from different sources that are part of the institutions' databases [20]. This creates a privacy issue for the students, faculty, and staff. It's important for institutions to take necessary measures to address the ethical and privacy issues related to data collection.

Big Data and analytics work best if the data is available in

electronic format [20]. The LMS and ERP can compile electronic data that can be mined. However, if an instructor is delivering the course in a traditional face-to-face format, then it would become difficult to translate the manual data into electronic data. This requires institutions to consider their Big Data strategies carefully, because it comes with an upfront cost and ongoing operational cost.

Data Science is a combination of three main knowledge areas: Hacking skills (programming skills), Statistical knowledge, and domain knowledge of the particular field [10]. There is also a lack of experts in the areas of Big Data [20]. Simply having knowledge of one or two of the knowledge areas may not be sufficient to utilize the full benefits of big data. Therefore, academia and industry should form a bigger partnership to provide learning opportunities that would produce data scientist.

VI. Research Trends

Big data can lead to big benefits. It is not only meant for big organizations [28] and even small organizations can benefit from them. Likewise, academic institutions can also benefit from Big Data. However, institutions should understand that not all data is created equally [28] and institutions should understand their data and context. Big Data and learning analytics have taken center stage in recent years. Institutions can apply data mining techniques and analytics to gain an understanding on different topics such as, administrative and instructional applications, recruitment, admission processing, financial planning, donor tracking, and student performance monitoring [20]. There is ongoing research being conducted related to Big Data and Student performance [29]-[31]. Additionally, there are conferences specifically geared towards Learning Analytics such as the Learning Analytics and Knowledge (LAK) conference.

VII. Conclusion

The purpose of this paper was to provide an overview of Big Data, some of the technologies used in Big Data, and its application in academia. With the tremendous amount of data that is generated from the Internet, Learning Management Systems and other academic applications, it becomes important for academic institution to at least explore how it can benefit their institution.

References

[1] M. Treder, "The definition of emerging technologies," 2010. (Online). Available: http://ieet.org/index.php/IEET/more/treder20101206.

[2] T. Olavsrud, "10 ways big data is changing the business of sports," 2014. (Online). Available: http://www.cio.com/article/2687035/big-data/164523-10-Ways-Big-Data-Is-Changing-the-Business-of-Sports.html.

[3] T. Olavsrud, "How big data is helping to save the planet," 2014. (Online). Available: http://www.cio.com/article/2683133/big-data/how-big-data-is-helping-to-save-the-planet.html.

[4] E. Dumbill, "What is big data? An introduction to the big data landscape," *O'Reilly Radar: Insight, Analysis, and Research about Emerging Technologies*, 2012. (Online). Available: http://radar.oreilly.com/2012/01/what-is-big-data.html.

[5] E. Schmidt, "Every 2 days we create as much information as we did up to 2003," 2010. (Online). Available: http://techcrunch.com/2010/08/04/Schmidt-data/.

[6] J. J. Berman, *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*, Waltham, MA: Elsevier, 2013.

[7] K. Laudon and J. Laudon, *Management Information System: Managing the Digital Firm*, 13th ed. UK: Pearson, 2014.

[8] B. Schmarzo, "Business analytics: moving from descriptive to predictive analytics," 2014..

[9] NMC, "The horizon report: 2014 higher education edition," 2014.

[10] D. Conway, "The data science venn diagram," 2010. (Online). Available: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

[11] B. Poulson, "Techniques and concepts of big data (Online Training Course)," 2014. (Online). Available: http://www.lynda.com/sdk/Hadoop-tutorials/Techniques-Concepts-Big-Data/158656-2.html.

[12] D. Henschen, "12 Hadoop vendors to watch in 2012," *InformationWeek Connecting the Business Technology Community*, 2012. (Online). Available: http://www.informationweek.com/big-data/software-platforms/12-hadoop-vendors-to-watch-in-2012/d/d-id/1102410?page_number=13.

[13] J. Sheard, "Basics of statistical analysis of interactions data from web-based learning environments," in Handbook of educational data mining, C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker, Eds. Boca Raton, FL: CRC Press, 2011, pp. 27–42.

[14] Y. Rogers, H. Sharp, and J. Preece. *Interaction Design: Beyond Human-Computer Interaction*, 3rd ed. John Wiley & Sons Ltd, 2011.

[15] K. Card, D. Mackinley, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann, 1999.

[16] K. Silius and Kailanto, "Visualizations of user data in a social media enhanced web- based environment in higher education," *Int. J. Emerg. Technol. Learn.*, vol. 8, no. 2, pp. 13–19, 2013.

[17] W. Hamalainen and M. Vinni, "Classifiers for educational data mining," in *Handbook of educational data mining*, C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker, Eds. Boca Raton, FL: CRC Press, 2011.

[18] P. D. Antonenko, S. Toy, and D. Niederhauser, "Using cluster analysis for data mining in educational technology research," *Educ. Technology Research and Develop*, vol 60, pp. 383-398, 2012.

[19] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD Conference*, 1993.

[20] A. Picciano, "The evolution of big data and learning analytics in american higher education," *J. Asynchronous Learn. Networks*, vol. 16, no. 3, pp. 9–20, 2012.

[21] IBM, "Analytics for achievement: Understand success and boost performance in primary and secondary education," Somers, NY, 2014.

[22] D. Niemi and E. Gitin, "Using big data to predict student dropouts: Technology affordances for research," in *Proceedings from the International Association for Development of the Information Society (IADIS) International Conference on Cognition and Exploratory Learning in Digital Age*, 2012.

[23] P. Willging and S. Johnson, "Factors that influence students' decision to dropout of online courses," *J. Asynchronous Learn. Networks*, vol. 13, no. 3, pp. 115–127, 2009.

[24] J. Kay, I. Koprinska, and K. Yacef, "Educational data mining to support group work in software development projects," in *Handbook of educational data mining*, C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker, Eds. Boca Raton, FL: CRC Press, 2011, pp. 173 – 185.

[25] D. Godoy and A. Amandi, "Link recommendation in e-learning systems based on content-based student profiles," in *Handbook of educational data mining*, C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker, Eds. Boca Raton, FL: CRC Press, 2011, pp. 273 – 286.

[26] NMC, "The horizon report: 2011 edition," 2011.

[27] T. Peckham and G. McCalla, "Mining student behavior patterns in reading comprehension tasks," in Proceedings of the 5th International Conference on Educational Data Mining, 2012.

[28] M. Korolov, "10 big myths about big data. Network World," 2014. (Online). Available: http://www.networkworld.com/article/2173703/software/10-big-myths-about-big-data.html.

[29] V. Protonotarios, G. Stoitsis, K. Kastrantas, and S. Sanchez-Alonso, "Using multilingual analytics to explore the usage of a learning portal in developing countries," *J. Asynchronous Learn. Networks*, vol. 17, no. 2, pp. 101–117, 2013.

[30] K. Xiangsheng, "Big data x-learning resources integration and processing in cloud environment," *Int. J. Emerg. Technol. Learn.*, vol. 9, no. 5, pp. 22–26, 2014.
[31] J. Yoo and M.-H. Cho, "Mining concept maps to understand university students' learning," in *Proceedings of the 5th International Conference on Educational Data Mining*, 2012.

Faisal Kalota. B.S. Computer Science, Northeastern Illinois University, Chicago, IL, USA, 1999. M.S. Computer Science, Northeastern Illinois University, Chicago, IL, USA, 2003. Doctoral Instructional Technology, Northern Illinois University, DeKalb, IL, USA, 2010.

His professional background includes Embedded Systems, IT Strategy & Planning, Project Management, Instructional Technology, and Teaching.