# A Comparative Study on Big Data Applications in Higher Education

**Arba Asha Altaye[1],  Dr. J. Sebastian Nixon[2]**
[1]Lecturer, School of Informatics, Wolaita Sodo University, Ethiopia, arba40asha@gmail.com
[2]Professor, School of Informatics, Wolaita Sodo University, Ethiopia, dr.nixon14@gmail.com

## ABSTRACT

Analysis of Big data is the new technology for extracting hidden information from the large datasets or data deluge, due to its different properties such as velocity, verities and volume. In this paper we discussed about the overview, existing technologies and tools and issues of big data. In the education field it plays a vital role. When the education system becomes digitized it uses large number of data and also its complex. The motivation of this paper is to express the importance of the applications of big data in education sector by analyzing the past and current situation of big data, finding the different applications related to higher education and to motivate educational institutions to implement the Big data applications in their campus and also encourage the researches to find the solutions for the issues in this field.

**Key words**: Big data, analytics, education, Hadoop, Mapreduce.
.

## 1. INTRODUCTION

Big data is an emerging technology that referring to the collections of the huge data sets that cannot be processed using traditional database management systems and existing techniques. Big data start new approaches for data storage, processing models, analysis and visualization of such enormous data size within an accepted time duration that can be reached with exact computational systems [1]. This is characterized into 4Vs: 1. Variety: it belongs to the following: structured, semi-structured,/ unstructured data. 2. Volume: It deals with huge size of data in terms of  petabytes . 3. Velocity: is the rapidness of data generation. 4. Veracity: Refers to the low quality and  inconsistency of data which are detected in the huge sets which are affecting the processing of data [2],[3], [4], [5].

There are considerable opportunities [6] for using big data in higher education. It is important that universities use big data to continue to deliver the very best learning environments for the good of society. There may also be possibilities around using big data to link research to education – both by making better use of latest research practices and outcomes to inform teaching and in enabling research activities to be worked as part of education.Big data is applicable in different organizations. Some of them are listed below: Banking, government, healthcare, manufacturing and education.

Hence the Big data are actually involved to grow the way of industries including the education. In the new era of data the traditional difficulties will be no longer exists, keeping the good methods. The education system will be enriched with new learning methods, making more efficient and targeted.[7].

The Evelyn Rupert [8], has asked 'who owns big data?', noting that number of people, technologies, actions and practices are involved in how data is captured, made and shaped. The technical activities for conducting educational systems are expensive to build such as big data collection, analysis and knowledge production. Specialist technical staff is required to program and maintain them, to design their algorithms, to produce their interfaces. The educational data are treated by the commercial organizations as a potentially lucrative market and 'own' the systems that are now being used to know, see and make sense of education and learning processes. Many of their systems are proprietarily, and are wrapped in IP and patents which makes it impossible for other parties to understand how they are collecting data, what analyses they are conducting, or how robust their big data samples are. Specific commercial and political aims may also be animating the development of educational data analytics platforms, particularly those associated with Silicon Valley where ed-tech funding for data-driven applications is soaring and tech entrepreneurs are frequently developing data-driven educational software and even new institutions.

In this angle, we need to raise important questions such as how educational big data are made, analyzed and circulated within specific fields and institutional contexts that often involve powerful role that possess significant economic capital in the shape of funding and resourcing, cultural capital in terms of the production of new specialist knowledge, and social capital through wider networks of affiliations, partnerships and connections. The question of the ownership of educational big data needs to be located in relation to these forms of capital and the networks where they circulate.

Many organizations can be affording the educational big data. Not all universities, schools and colleges can necessarily afford to purchase a learning analytics or adaptive software platform.

This allows some wealthy institutions being able to benefit from real-time insights into learning practices and processes that such analytics afford, while other institutions will remain restricted.

In this paper, our focus is on the application of big data in education sectors. Educators can make a significant impact on educational systems and they can identify the various activities of the students. . By analyzing big data, they can identify the various activities of the students.

The rest of the survey is structured as follows: Section 2 provides a description of the big data, big data tools and technologies, challenges and opportunities in a big data. Section 3 the research methodology of the survey. Section 4 presents the basic applications of big data in education sector. Section 5 describes educational system management issues with big data .Section 6 summarizes the paper and section 7 contains conclusion and finally references.

## 2. BIG DATA
In this section we discussed the concepts of big data.

### 2.1 Big data definition

Big data is identified as a huge amount of data collected through time and is difficult to handle using traditional database management tools. Social media activities, business activities, photos, videos, sensors, e-mails, text files and application logs are sources of big data [9].

Big data may be described by velocity, volume and variety characteristics.[10] To extract useful information from big data, one needs a great processing power, analytical capabilities and skills big data generally ranges from several TB (Terabytes) to several PB (Petabytes) and even EB (Exabyte)[11][7]**.**

Big data often relates to unstructured and semi-structured data[12] that appears in all kinds of places, from various sources and formats; for example web content, Twitter posts, Face book photos and free form comments. Their outputs (website log files, social media sentiment analysis, video streams) can be analyzed to provide a range of insights that previously were not possible. For example, previously 10 years took to decoding the human genome to process but now within a week it can be done. Currently, a key concern is that big data projects are highly dependent on skilled developers and computer programmers and in order to avoid limited diffusion and to democratize use, user-friendly developer tools are needed.

### Big data Technologies and Tools

Many technologies and tools are used to store, manage and analyze big data. Technologies and tools that support big data techniques include, but are not limited to [13].

Hadoop is an open source software framework for processing huge amounts of data on a distributed system for a given problem. It is de- signed to store, manage, and analyze hundreds of Terabytes and even petabytes of data. It is considered as one of the big data analysis plat- forms supporting Windows, Linux, and OS X operating systems. It is comfortable to work with multiple data sources, large scale processing and large volumes of data, such as transaction data, social media data and weather location-based data. Hadoop components include:

> - **Hadoop Distributed File System (HDFS):** It is developed to cover large clusters of commodity hardware based on java-based file system that supports scalable, credible and redundant data storage.

> - **MapReduce engine:** This is a frame work for distributed processing of large data sets on computer clusters. It includes the following tasks: 1. monitoring, 2. re- executing any failed tasks and 3. Task scheduling.

> - **NoSQL database (Not Only SQL):** is a latest technology of databases with new characteristics; being non-relational, distributed, horizontally scalable, schema-free, supports easy replication, supports simple API and keeps numerous quantity of data. There are different types of databases that fit into this; such as document stores and key-value stores. This database type focus on the storage and retrieval of huge volumes of structured, unstructured and semi-structured data. The class of NoSQL DBMS' includes HBase that is the non-relational data store for Hadoop and MongoDB that was de- signed to support homogenous databases.

> - **R:** is an open source programming language which supports statistical computing and graphics. It has an increasing importance as a tool for computational statistics, visualization and data science.

### 2.2 Big Data Analytics

It is a technique to identify the patterns and trends from a large volume of data to extract its value and correlations [14][15]. According to [16], the steps involved in data retrieval for big data are as follows: 1) Data acquisition, 2) Data Extraction, 3) Data Collation, 4) Data structuring, 5) Data visualization, and 6) Data Interpretation.

### Challenges in Big data
Big data is more complex and analyze the large datasets is more difficult. Combine both big data and analytics is a baffling for the researchers and organizations. The following are the main challenges [17].

**Analytics Architecture**: The batch processing techniques were used for analyze the historic data using the existing analytic architecture. Nathan Marz introduced lambda Architecture for analyzes the real-time / streaming data.[18].

**Time Evolving Data:** Data may flow in and out, with dynamic change, so analyzing the streaming data is major problem since technologies are immature.

**Compression**: There are 2 major techniques: Using compression technique, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling technique, we are losing information, but the gains in space may be in orders of scale.[19]

**Visualization**: One of the main advantages of big data Analytics is to visualize the results. As the data is enormous, it is very difficult to find easy to use visualizations. New technologies and approaches to put in the picture and demonstrate stories will be needed. For example the photographs, infographics and essays [20].

**Poor data quality** can significantly affect the effectiveness of big data projects. As analysis on big data grows, so too will the need for validation, standardization, enrichment and resolution of data.[21]

### 2.3 Opportunities in Big Data

Big data analytics will open many opportunities for the educational institution. It will useful for the learners to administrators. These opportunities include[22]: Cross collaboration and comparisons among the institutions would become more comfortable, improved knowledge flow and learning success across the organization would be achieved, learning effectiveness would be improved through the self-measurement of learners and educators, Cost reduction through managing financial performance could be possible and the learning and academic risk and complexity could be reduced.

### 3 RESEARCH METHODOLOGY

The main aim of this study is to report on the applications of big data in education sector by examining the literature, identifying different applications, presenting open research questions and compiling a comprehensive reference list to assist researchers.

In order to achieve this objective, a comprehensive review of the literature was performed. The reviewed literature included journal articles, conference papers, and edited volumes. Relevant literature was identified by querying scholarly databases for the terms "application of big data in education". Returned results were downloaded and read. We have tried to review many papers and select relevant articles for our survey study. Totally 32 papers were deeply reviewed.

### 4 Basic Applications of Big Data in Education Sector

#### 4.1 Student Performance Prediction
Predictive Analytics enables prediction of student's behavior by analyzing various activities performed by the student while interacting with the Learning Management System. Using data mining techniques we can predict the student's performance based on their activities then take measure to improve them[23].

The following Figure 1 shows that how to evaluate or predict the student performance by using C5.0 algorithm. In data mining techniques, the many classification methods are used to predict the student's performance. In that C5.0 is one of the best decision tree classification algorithms. It can handle continues and categorical values. It can handle numeric attributes. Attribute selection is the basic step to construct a decision tree. Entropy, information gain, and gain ratio is used for attribute selection process, using attribute selection C5.0 algorithm select which attribute will be selected to become a root node of the decision tree and so on. Comparing with ID3 and C4.5, C5.0 is the highest speed and pre-pruning was performed.
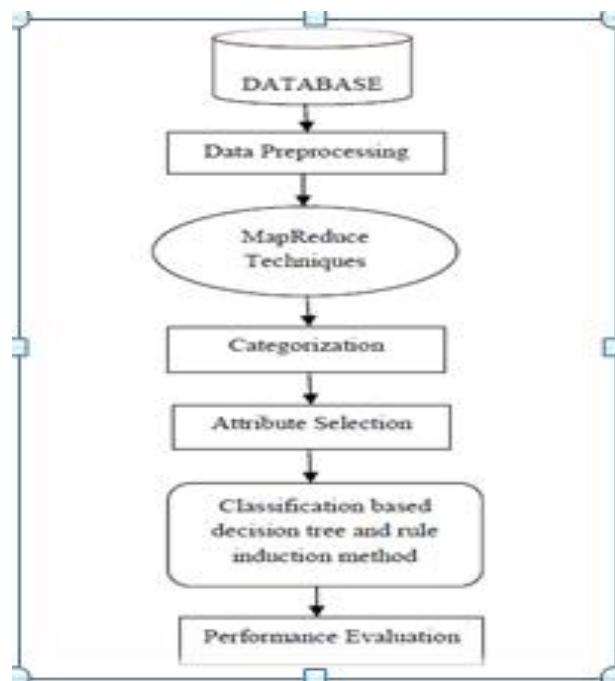


**Figure :**. Student performance prediction

#### 4.2 Behavior Detection
Joseph Grafsgaard [24] described about the student's facial expressions after the class by the following activities: engagement, frustration and learning outcomes of students using gesture detection and posture tracking algorithms.

In addition to the above, Seong Jae Lee[25], described a framework that learns individual policies from the movement of the players in the game and builds a cognitive model. He stated that this type of modeling will help in understanding learning processes of the user who interacts with the system and in adapting the learning environment to the user.

#### 4.3 Attrition Risk Prediction
Big data techniques are used to predict the risk of attrition associated with students. Some of the students are dropped out of courses their activities are monitored and engagement score

is predicted. The predicted score was found to be dependent on the attrition rates.

Lalitha Agnohotri [26], proposed a new model named "Student At Risk" that is capable of calculating risk ratings for new joiners. She used that the historical data to model the student's behavior and the model was used to calculate the attrition risks of new joiners. The ratings can be used to identify the students at risk and take needed measures to retain those students in advance.

## 4.4 Skill Estimation

It is used to estimate the skills of the students so that the education system can be altered to suit the student's skills. Skills were calculated based on the interaction of the student with the system or in the message boards or discussion forums.

Paulo Blikstein [27], used a tool called "NetLogo". He logs the mouse inputs of students into the lab machines through the software and with help of the data logged finds the error rates and progress rates of the students.

Beheshti [28], assessed the skills of learners using synthetic and real data sets then compared the differences between the details of the skills then analyzed . In that the real data provides more accurate results.

## 5 EDUCATIONAL SYSTEM MANAGEMENT ISSUES WITH BIG DATA

Big data is a knowledge system which changes the objects of knowledge and social theory in many fields while also having the potential to transform management decision-making theory[29]. Currently major issue in higher education institutions is handling a massive information system concerned with student learning-teaching process and result consistency and delivering mechanism to be operated in secure way by using web service. The other matter related here is ensuring the quality of education system and promoting disciplined professionals that can be computed with industrial standard. So that big data in higher education covers database systems that store large quantities of longitudinal data on students' right down to very specific transactions and activities on learning and teaching.

In general, big data Analytics could be applied to examine student entry on a course assessment, discussion board entries, blog entries which could generate thousands of transactions per student per course.

It can impact the educational system practice, from enhancing students experience to improved academic programming, to more effective evidence-based decision making, and to strategic response to changing global trends. Big data promises to turn complex, often unstructured data into actionable information.

According to Hilbert [1], big data delivers a cost-effective prospect to improve decision making. However, the

implementation of big data is a challenge in the case of technology, cost and designing fittest architecture to cover overall necessary issues. There is a tremendous cost associated with development of algorithms, data collection and storing of data to mine data. Since most of the institutional systems data are not interoperable, the aggregation classroom, online data and administrative data make additional challenges.[30]

By using the administrative and operational data we can evaluate the performance, issues and future progress of the institution [31]. Siemen [30], indicated that it is now well positioned to start addressing some of the key challenges currently facing higher education.

To minimize the challenge in case of management, divide the targeted processes in different category or group and assign directing body for each category. As in the figure 2 below visualized, the conceptual framework for security camera based problem identifying technique is in three categories as classroom area category, administrative office area category and students' dormitory and surrounding field category. In case of classroom, the data to be handled include instructor regular class attendance and course delivery, students' action during lecturer class examinations period. This enables the better controlling mechanism to evaluate education quality at start point.
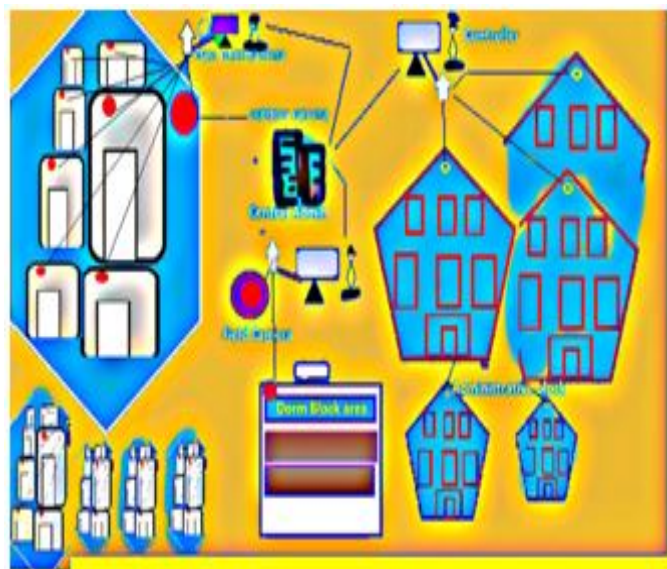


**Figure 1:** Conceptual framework of big data security camera in higher education administrative technique

Here what we observe from the figure 2 above is general about security camera based administrative control mechanism. As visualized above, there are three conceptual category of camera poisons and separate administrators for each three categories and central administrators for future data store and decision provider. Around the classroom block the main activity is live during normal lecturer class and particularly during examination time to handle cheating situations even if the instructor is unwilling to act punctually and then the malpractice picked up and provided to the central administrator

and that the process managed correctly with video image evidence. It may not need to be online every day for every hour. It on when needed and off during no lecturer class week. Among the technologies Mobile DVR Systems GP-550NDF is fit for this area. This is because mobile DVR systems are capable of supporting mirror recording and high speed video compression. Furthermore, equipped with fixed 4 mm lens, these systems are suitable for recording in both day and night. The Figures 3 & 4 are some example of Security Cameras.



**Figure 2:** Security Camera

The other category of surveillance system is around administrative block to secure activities at office area. There is also their own controller as admin and directs necessary information to central administrator. The filed camera established for campus security at all that built in sensitive area around entrance, sport area, TV room, dorm block, etc. IR Camera is best suit for this area as this series is a high resolution, weather-proof, day & night IR security camera from GPES that delivers superior image quality with an IR effective range from 15M to 120M. The camera utilizes 420TVL -650 TVL to deliver crystal clear images and its sensitivity of 0.00Lux allows it to produce clear images even in complete darkness. In addition to its weather-proof housing, this series also comes with a wide range of functions, such as AGC, BLC, and Auto White Balance, OCR, Face recognition etc. Both the models, one with IR LED board and other with LED array, are available.

The other issues seek for big data management in educational system is managing official website information provided as comment. Updating web information frequently is also requirement for quality education to provide updated course based materials and academic calendar others also e-library information and student statistics and over all academic progress keep updated and that can be easily prepared for report purpose.[32]



**Figure 3:** Field camera

## 6 SUMMARY

The following table shows the short summary of the paper.

**Table 1:** Summary Table

| Applications | Descriptions |
|---|---|
| Performance Prediction | Student's performance can be predicted by analyzing student's interaction in a learning environment with other students and teachers |
| Attrition risk Detection | By analyzing the student's behavior, risk of students dropping out from courses can be detected and measures can be implemented in the beginning of the course to retain students. |
| Student skill estimation | Estimation of the skills acquired by the student |
| Behavior Detection | Detection of student behaviors in community based activities or games which help in developing a student model |
| **Challenges** | **Description** |
| Time Evolving Data | Data may flow in and out, with dynamic change, so analyzing the streaming data is major problem since technologies are immature. |
| Poor data quality | Poor data quality can significantly impact the effectiveness of big data projects. As analysis on big data grows, so too will the need for validation, standardization, enrichment and resolution of data |
| Analytics Architecture | The existing analytic architecture is used for analyze the historic data using batch processing techniques. Nathan Marz introduced lambda Architecture for analyzes the streaming/real-time data. |
| Training practitioners | To develop better understanding of the practitioners about the system will be time consuming. It would be difficult for the learners and educators to present information in an accessible and informative way and therefore would be inflexible to collaborate with the system. |
| **Opportunities** | **Description** |
| collaboration | Cross collaboration and comparisons among the institutions would become more comfortable. |
| knowledge flow | Improved knowledge flow and learning success across the organization would be achieved. |
| Cost reduction | Cost reduction through managing financial performance could be possible |
| Learning effectiveness | Learning effectiveness would be improved through the self-measurement of learners and educators. |

The following figure 5 shows that the number of articles reviewed by their respective publication year. Our article review was between 2011 to 2017. From reviewed articles 10 articles are published in 2014. Totally 32 articles were reviewed.
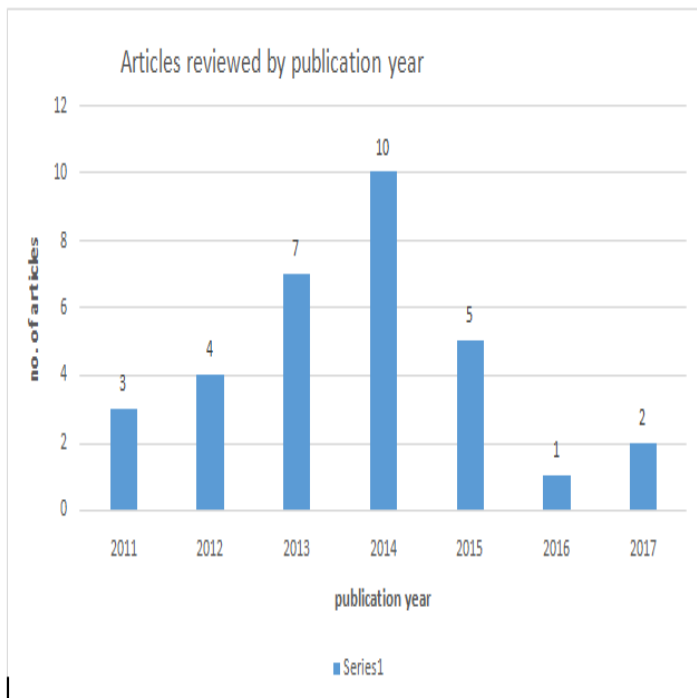
**Figure 4:** Articles Reviewed by publication year

## 7 CONCLUSION

The study reported on the applications of big data in education sector by examining the literature, identifying different applications, describing challenges and opportunities of big data and compiling a comprehensive reference list to assist researchers. To achieve comprehensive review of the literature was performed. The reviewed literature included journal articles, conference papers, and edited volumes. Overall we reviewed 32 papers.

Rather than other business institutions, higher education institutions struggling with the issue of big data handling such as academic progress, administrative progress, student's information management and employ management. Similar to business, government and healthcare institutions, Higher education institutions share some of the same reasons for adopting analytics, especially in the areas of financial efficiency; expanding local and global impact; addressing new funding models during a changing economic climate and responding to the demands for greater accountability. It is fact that in the institutions data are growing, but most of them are scattered across desktops and come in various formats, making it difficult to retrieve or consolidate. To utilize these data in a proper manner, the following are the key requirement for implementation of big data in higher education: 1. Ability to analyze diverse information sets, 2. Origin and consolidating data, 3. Managing and governing the data. Central predicting system to manage efficiently the issue of big data in the higher educational institution is concern point for future. We hope in the future all the educational institutes will implement the big data applications and make use of it.

## REFERENCES

[1] R. Kirkpatrick, "Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, 'Business Intelligence And Analytics: From Big Data To Big Impact, Big Data Analytics An Oracle White Paper', MIS Quarterly vol. 36 no. 4, pp. 1165-1188/December 2012.," *Big Data*, vol. 1, no. 1, pp. 3–4, 2013.
https://doi.org/10.2307/41703503

[2] "Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, 'Business Intelligence And Analytics: From Big Data To Big Impact, Big Data Analytics An Oracle White Paper', MIS Quarterly vol. 36 no. 4, pp. 1165-1188/December 2012."
https://doi.org/10.2307/41703503

[3] "Youssef, M., Gamal Attiya, and El-Sayed Ayman. 'New Framework For Improving Big Data Analysis Using Mobile Agent.'"

[4] "Kale Suvarna Vilas, 'Big Data Mining', Journal of CSMR, CSMR (International Journal of Computer Science and Management Research eETECME), October 2013."

[5] "Rong Liu, Qicheng Li, Feng Li, Lijun Mei, Juhnyoung Lee, Big Data Architecture for IT Incident Management, 2014 IEEE."
https://doi.org/10.1109/SOLI.2014.6960762

[6] S. Anirban, "Big Data Analytics in the Education Sector : Needs , Opportunities and Challenges," vol. 3, no. 1, pp. 1425–1428, 2014.

[7] A. S. Drigas and P. Leliopoulos, "The Use of Big Data in Education," vol. 11, no. 5, pp. 58–63, 2014.

[8] E. Ruppert, "Rethinking empirical social sciences," *Dialogues Hum. Geogr.*, vol. 3, no. 3, pp. 268–273, 2013.
https://doi.org/10.1177/2043820613514321

[9] K. ElDahshan and H. Mancy, "HPC based Modeling, Analyzing and Forecasting of a Century of Climate Big Data," *Int. Congr. Glob. Sci. ...*, no. December, pp. 39–50, 2015.

[10] S. R. Azzam and Y. Ramadani, "1 ) Big data elements A . General B . Education Sector B . Cloud Computing Internet of Things ( loT ) D . Business Intelligence," pp. 3–8, 2016.

[11] M. C. B. et al. J. Manyika, "Big data : The next frontier for innovation , competition , and productivity," no. May, 2011.

[12] D. Rajeshwari, "State of the Art of Big Data Analytics : A Survey," vol. 120, no. 22, pp. 39–46, 2015.
https://doi.org/10.5120/21395-4456

[13] "S. Tae Bae J. Kim, G. Wang. A Survey of Big Data Technologies and HowSemantic Comput- ing Can Help. International Journal of Semantic Computing, 8(1):99–117, 2014." .

[14] "Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 42-47). IEEE."
https://doi.org/10.1109/CTS.2013.6567202

[15]    "Toshniwal, R., Dastidar, K. G., & Nath, A. (2015). Big data security issues and challenges. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(2), 15-20."

[16]    "Villars, R.L. et al. (2011) Big Data: What It is and Why You Should Care. IDC."

[17]    W. Fan and A. Bifet, "Mining Big Data : Current Status , and Forecast to the Future," *ACM SIGKDD Explor. Newsl.*, vol. 14, no. 2, pp. 1–5, 2013.

[18]    N. Marz and J. Warren, "Big Data, Principles and best practices of scalable real-time data systems," *Big Data - Princ. best Pract. scalable real-time data Syst.*, vol. 37, pp. 1–23, 2015.

[19]    D. Feldman, M. Schmidt, and C. Sohler, *Turning Big data into tiny data : Constant-size coresets for k -means , PCA and projective clustering*. 2013. https://doi.org/10.1137/1.9781611973105.103

[20]    T. M. Scholz, *Big Data in Organizations and the Role of Human Resource Management: A Complex Systems Theory-Based Conceptualization*. 2017.

[21]    O. Stevenson, O. Marsch, Maurovich-Horvat, and Lajos, "Big Data and Education: What's the Big Idea?," *Big Data Educ. What's Big Idea?*, no. 3, p. 33, 2014.

[22]    B. Daniel, "Big Data and analytics in higher education: Opportunities and challenges," pp. 1–17, 2014. https://doi.org/10.1007/978-3-319-06520-5_1

[23]    K. Sin and L. Muthu, "APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS – A LITERATURE REVIEW," vol. 6956, no. July, pp. 1035–1049, 2015.

[24]    J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Predicting Learning and Affect from Multimodal Data Streams in Task-Oriented Tutorial Dialogue," *Proc. 7th Int. Conf. Educ. Data Min.*, no. Edm, pp. 122–129, 2014.

[25]    S. J. Lee, Y. Liu, and Z. Popovi, "Learning Individual Behavior in an Educational Game : A Data-Driven Approach," *Proc. 7th Int. Conf. Educ. Data Min.*, no. Edm, pp. 114–121, 2014.

[26]    L. Agnihotri and A. Ott, "Building a Student At-Risk Model : An End-to-End Perspective," *Proc. 7th Int. Conf. Educ. Data Min.*, pp. 209–212, 2014.

[27]    P. Blikstein, "Using learning analytics to assess students' behavior in open-ended programming tasks," *Proc. 1st Int. Conf. Learn. Anal. Knowl. - LAK '11*, no. May, p. 110, 2011.

[28]    B. Beheshti and M. C. Desmarais, "Predictive performance of prevailing approaches to skills assessment techniques : Insights from real vs . synthetic data sets," *Proc. 7th Int. Conf. Educ. Data Min.*, pp. 409–410, 2014.

[29]     danah boyd and K. Crawford, *Critical Questions for Big Data*, vol. 15, no. 5. 2012.
https://doi.org/10.1080/1369118X.2012.678878

[30]    B. K. Daniel and R. Butson, "Technology Enhanced Analytics (Tea) in Higher Education," *Int. Conf. Educ. Technol.*, no. Oblinger 2012, pp. 89–96, 2013.

[31]    A. G. 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.
https://doi.org/10.24059/olj.v16i3.267

[32]    "https://www.indiamart.com/gp-eco-solutions/security-servillance-systems.html (accessed on 06/20/2017).