Educational BigData Mining Approach in Cloud: Reviewing the Trend

D. Pratiba
Asst. Prof, Dept. of Computer Science & Engg.
Visvesvaraya Technological University
RVCE, Bangalore, Karnataka India

G. Shobha, Ph. D
Prof . & HOD, Dept. of Computer Science & Engg.
Visvesvaraya Technological University
RVCE, Bangalore, Karnataka, India

ABSTRACT
Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. We present in this issue, a broad overview of the topic, its current status, controversy, and forecast to the future. We introduce four articles, written by influential scientists in the field, covering the most interesting and state-of-the-art topics on Big Data mining

Keywords
component; Big Data, Data Mining

1. INTRODUCTION
With the advancement of IT enabled products in communication and networking system, various organization (e.g. education, healthcare, financial institution, customer relationship management etc) are today witnessing a constant and exponential rise of data in every minutes and seconds (IBM, 2014) [1]. According to Press (2014), performing effective processing on Big Data can elicit various predominant information by rendering the information apparent for increasing the usage rate [2]. An organization can extract various latent transactional attributes for accelerating the business performance by collecting more interesting and accurate information from the big data. According to Devingnes et al [3], precise knowledge discovery from big data will essentially aid an organization to formulate better decision for growth of its business. Big data also permits shaping up better customer segmentation that directly assists to increase the business avenues and hence maximize the growth rate [4]. Finally, effective processing of big data give rise to innovation which is one of the key factors in majority of the business goals [5]. The big data are in the form of audio, video, text, satellite image, medical images and many more originated from different business operation [6]. Storage of big data is never a problem due to ubiquitous services offered by cloud [7][8]. With a massive volume of data being generated, achieving effective storage and processing it becomes a very challenging task with respect to cost and optimization of data. Such preference of big data has raised the attention of research community as solving the key issues of managing big data will eventually lead to innovation and maximization of productivity and thereby a company could visualize it as an opportunity to open up new avenues of business scope with the same resources.

One of the existing applications of the big data is Big Data Analytics [9] which can be defined as a method to evaluate big data to elicit the latent pattern, unknown correlations, and many more interesting facts that could be used to increase the dimension of existing business. One of such significant product is owned by IBM [1] that assists in optimize operation, enhance risk mitigation measures, and create novel business models. However, big data was also termed as ‘noise’ [9]. It is because that it consists of massive redundancies and irrelevant information too for which reason, effective processing is highly essential. With the use of various tool (e.g. data/text/web mining, prediction, analytics etc), knowledge discovery can be achieved that will significant assist in formulating business decisions [10]. However, it is not that easy as it seems like and the domain of big data has evolved very recently. This research proposal will discuss about the problems being identified in using big data and introduce an analytical optimized framework that assists to broaden the scope of business interest. Various social networking applications like Facebook, twitter as well as sales / marketing application etc are generating a massive volume of information along with growth of the dynamic user. The data are produced in the format of text, image, video, audio and many other application oriented format. Although with the existence of cloud, storage doesn’t poses a big issues for big data, however, processing with conventional and existing data mining algorithm seriously posses a biggest threats. Implementation of processing of big data is highly associated with potential execution issues [11][12], where finding the best out of existing tool is always a challenging task [13]. The present paper discusses about the trends of big data with respect to educational institutional and usage over cloud platform. The study also explores various traits and trends of the big data usage with the support of strong evidences in electronic media. This paper discusses about the recent trends and issues of Big data from education sector viewpoint. Section 2 discusses about the evolution of big data followed by discussion on the available tools in Section 3. Section 4 discusses about the topic from educational viewpoint followed by discussion on cloud storage of big data in Section 5. Section 6 discusses about the knowledge discovery process followed by study significance in Section 7. Section 8 discusses about the open issues of the study and finally section 9 concludes the paper by summarizing more issues on the topic.

2. EVOLUTION OF BIG DATA
With the constant rise of advancement in the internet technology, ubiquitous computing is also growing exponentially and reaching the end customers day by day. The information is now highly distributed and gain extreme higher degree of mobility due to ubiquitous computing. Hence, it becomes easier for the user to access their resources from multiple computing devices, even on the move. Such communication pattern is also bidirectional, which means that user not only accesses the resources but also share their own resources on the other end [14] [15]. Hence, it can be said that on every seconds, millions of data are being generated and stored in server. This is how the data are growing in every seconds of life, which is now technically called as ‘Big data.’ However, the question arises is how far successful we are

in retrieving some significant information (knowledge discovery) from the big data. As the information carried by big data is highly valuable for business goals and it gives a cut edge prediction capability to the organization by excavating some more secrets about their data which cannot be explored by conventional datamining or data warehousing techniques.

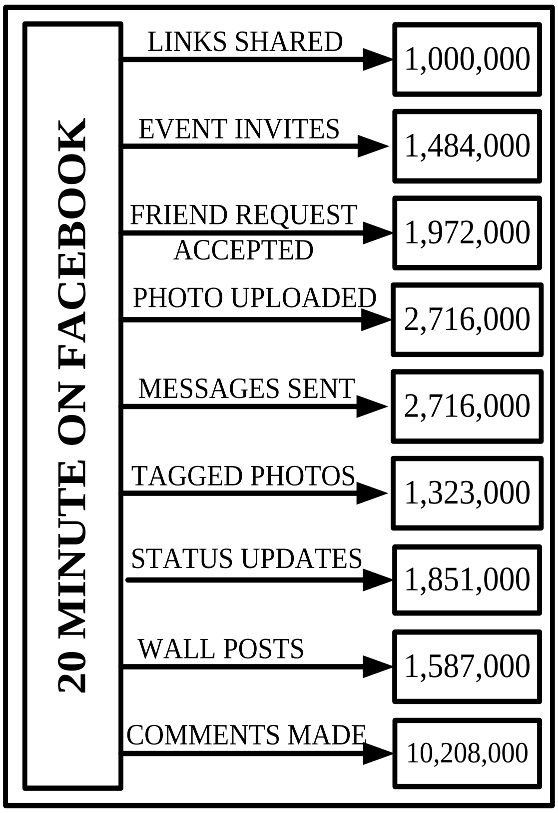

Fig 1: Existing trend of Big Data in Facebook

[Source: http://contemporarycondition.blogspot.in/2012/11/digital-killed-video-candidate.html]

Fig 1 pictorially represents the types of voluminous data generated from the famous social networking site Facebook. From the study conducted in [16], it was found that around i) 1 billion queries are received by Google per day, ii) more than 800 millions of updates are given by users in Facebook, iii) More than 4 billion user views video hosted in YouTube and quarter half of them post their comments on it, and iv) More than 250 Millions of tweets are recorded by twitter in just one day.

Reputed survey oriented organization like Gartner reported the witness of increasing trends of Big Data on the year 2012 [17]. The report has also chalked out that Big Data could be effectively used in maximizing the scope of business, technology, and health sector potentially. It is expected that performing knowledge discovery over Big Data will efficiently open up many avenues of various business scopes keeping with pace of technological modernization [18]. One such effort was also shown by the company Global Pulse who is on constant research data on Big [18]. The company has ongoing research activity for big data being generated from various sources. The same company has published one white paper [18] that optimistically speaks about predominant advantages of Big Data in the viewpoint of real-time awareness, early warning, and real-time feedback. Also, the reports says that with the rise of mobile devices and smart-phones, big data is on constant rise. The next section discusses about the various tools that is frequently heard with Big Data processing.

3. AVAILABLE TOOLS

The work of production and research community has already started exploring about Big Data. Some of them design their own applications/tools to store/process Big data, while some of them still attempt to rely on standards tools. Hence, this section will discuss only such tools that are publically available in processing Big Data as follows,

- **Apache Hadoop**: It is a basically an open source software framework that assists in storing massive volume of data as well as perform processing on the big data. Such framework is also associated by another name MapReduce, which is a programming tool on processing massive volume of data residing in clusters [19]. It is also associated with other framework e.g. Apache ZooKeeper, Scribe, Apache HBase, Apache Pig etc [20].
- Apache S4: It is another product of Apache software foundation that addresses the issues of complex proprietary and batch oriented open source framework [21].
- Storm: It is a software framework for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter [22].

Some of the significant open source projects for addressing Big data mining issues are as follows:

- **Apache Mahout**: It is one of the scalable machine learning open source software designed exclusively with Hadoop. It has enriched with vast range of machine learning and data mining techniques e.g. clustering, classification, collaborative filtering and frequent pattern mining [23].
- **R**: It is an open source programming language and software environment designed for statistical computing developed by Ross Ihaka and Robert Gentleman [24] and is deployed for analyzing statistically for very large data sets.
- **MOA**: It basically stands for massive online analysis [25] that performs data mining in real time streaming data. It uses the potential features of classification, regression, clustering and frequent item set mining and frequent graph mining. Its advanced version called as SAMOA [26] is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA.
- **Vowpal Wabbit**: It is basically an open source project under the joint initiativeness of Yahoo and Microsoft for developing a scalable and efficient learning algorithm [27].

More specific to Big Graph mining we found the following open source tools:

- **Pegasus**: It is a big graph mining system that is designed using MapReduce and permits determining the patterns and anomalies in massive real-world graphs [28].
- **GraphLab**: It is a high-level graph-parallel system built to computes over dependent records that are recorded as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices [29].
-

4. BD IN EDUCATIONAL SECTOR

The current paper basically intends to discuss about the trends of the big data that is generated exclusively from educational sector. The existing educational system has highly revolutionized from what was there 20 years ago. The trend of higher adoption of technical classes, archival of documents for every session, lecture notes, feedbacks generated by students, instructors, as well as by critics are constantly on the rise to meet up the quality standards of education system for any country. Such big data is emphasized not only from storage viewpoint but also from processing viewpoint. As educational industry has revolutionized along with advancement in technology, so is the growth of big data, which is doubling every year. An authenticated survey report given by McKinsey [30] highlights that educational sector is the next evolving sector (after communication and government) that generates higher quantity of Big data. For better visualization, generation point of big data is highlighted in Table 1.

Table 1 Trend of data generated / stored by sector

	Video	Image	Audio	Text/Numbers
Banking	Medium	Medium	Medium	High
Insurance	Low	Low	Low	High
Securities and investment Services	Low	Low	Low	High
Discrete Manufacturing	Medium	Medium	Low	High
Process Manufacturing	Medium	Medium	Low	High
Retails	Medium	Low	Low	High
Wholesale	Low	Low	Low	High
Professional Services	Medium	Medium	Medium	High
Consumer and Recreational Services	Medium	Low	Medium	Medium
Health care	Low	High	Low	High
Transportation	Medium	Medium	Low	High
Communications and Media[2]	High	Medium	High	High
Utilities	Medium	Medium	Low	High
Construction	Low	Low	Low	Medium
Resource Industries	Medium	Medium	Low	High
Government	High	Medium	High	High
Education	High	Medium	High	Medium

High Medium Low

Source: McKinsey Global Institute analysis, (2011)[30]

With the trend of adopting ICT in the educational establishment, right from enrollment to result declaration is carried out in web based applications. Some of these applications are also supported by the smaller versions of apps in mobile devices for better data mobility feature. Majority of the advance and reputed higher technical educational institution has already adopted various applications that makes the complete administrative work go paperless. This trend is found to be very promising as it saves lots of time against doing unproductive work and stress on more skill development and better communication. Various factors that has been marked responsible for generation of Big Data in educational section are discussed below:

- **Academic Trend**: It has been seen that majority of the educational establishments are now introducing various customized tools for the purpose of various administrative jobs like student's enrollment, fee collection, reporting system etc. Evidence of such trend was seen in literature [31] that discusses about an application called as "Service-Oriented Higher Education Recommendation Personalization Assistant." This application basically assists the students by furnishing recommendation to make their decision further more confirmed. It also assists the students to select their prime classes to be undertaken along with schedule management.

- **Futuristic Technology**: One of the biggest level of uncertainty in the education sector is from the student's side prior taking admission or to take some significant decision for undertaking some specific courses available in the educational institution. Usually, students and parents spends lots of time researching in Google about it and finally ends up in static page that furnishes some outdated or static information. The recent trend in educational sector was seen with customer service or an automated system that interacts with the students and assists them about their uncertainty factor. With the constant rise of mobile apps and various services rendered on 3G enabled devices, the information can be accessed by performing interaction with the customer services or some automated system (IVRS). These types of interactive communication also generated a massive volume of data on the other end and it maximizes potentially [32].

- **Trend of Academic Lives**: The literature [33] shows that if the learning community assists the students by furnishing their valuable and experienced suggestion against the query or the current issues they (student) are encountering, student satisfaction is raised. Such trends can permissively give cut edge competitiveness among the student community and strengthens their decisive skills too. Various online tutorials and model applications already exists where the login priviledge are given to the enrolled students to experience the online learning benefits. One of such application is called as Persistence Plus that builds upon the academic profile of the student along with their existing trends of academic lives [34]. Such types of tools evidently boost the morale of the students by assisting them to shape up their academic life in better synchronicity with the ongoing courses and activities in their educational establishments. Supported by plugin with existing mobile devices, a student finds it quite comfortable and logical to adopt it for their educational betterment. Practicing of such tools also generates quite a substantial data that are quite hard to be stored in ordinary servers.

- **Performance Monitoring**: Various online tools exists currently that considers the academic or skill sets of the student and assist them to undergo virtual assessment to understand where do they stand in viewpoint of classroom exercise or job interview. Such tools highly assist the students to form a community where the networks of users are constantly on rise and the application is constantly populated with various log files of their performance. Due to free nature of such tools, the adoption among the student community is always high and so is the capturing of voluminous data [35].
- **Virtual Classes**: With the rise of competition in academics, the students are constantly adopting various mechanisms where they could update their skills cost effectively. Such facts give rise to virtual meeting with the tutors/expertise who assist the student by giving their valuable guidance on advances courseware. This is the most ongoing trend and probably the trend will continue in near decade as it is one of the cost effective and time saving tool, which let the student master the skills at the comfort of their location. Various reputed institutions are already adopting cloud based virtual classes to give better skills sets to their students, where various digital contents on virtual classes are shared. Such digital contents are usually PowerPoint presentation and audio files, in few cases, the virtual classes could also be seen offline by enabling the enrolled students to let record the session with authentications. Hence, such types of virtual classes produces enormous data, which could never reside in conventional server and need to take the aid of cloud based storage system.

Hence, it can be seen that with the passage of time, demand of education is increasing day by day and so is quality to be imparted to the students. Such Big Data could find better usage to identify some more cost effective traits in educational sector. Such Big data could potential enhance the educational sector by processing it and by performing various knowledge discovery on it, which is the biggest challenge after all. The biggest question over here is there is a big data generated from educational sector, but what to do about it? How to use this data to increase the scope of existing educational system?

5. STORING EDUCATIONAL BD IN CLOUD

There are various tools (as illustrated in Section III) that uses individual-level data to transform the way higher education is being done today and to provide new data on how it should be done in the future.

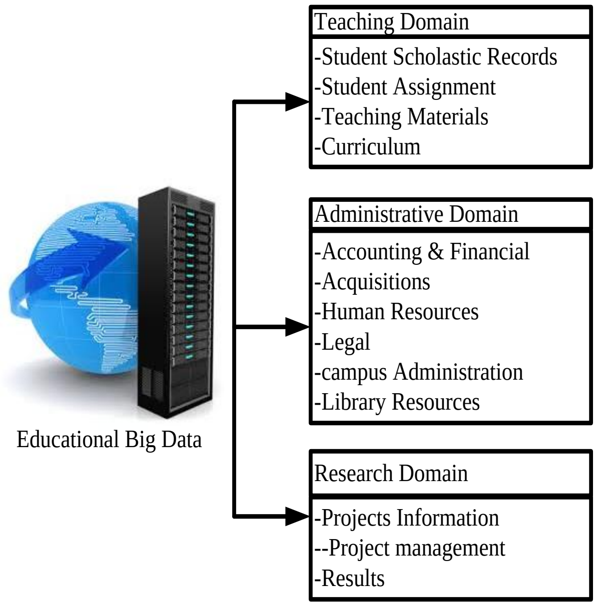

Fig 2: Existing Storage of Educational BD

As discussed earlier, that storage of big data is never a big challenge, but what matters is how much effective information can be retrieved from it, which could substantially ensure return-of-investment (ROI). The existing scenario of evolution and storage of the Big data from educational sector is pictorially exhibited in Fig.2. The above figure shows that the Big data is generated right from teaching domain, administrative domain, as well as research domain when it comes to educational sector. Such Big data couldn't find a place in conventional server for storage and hence adopts cost effective cloud service for it [36]. Following are the prominent existing cloud service provider that caters up the storage requirement of Big Data being generated by educational sector,

- **Google**: The company Google provides various standard applications based on web-based interfaces and storage system that is operated by user on any client-specific applications (e.g. Firefox). Google already has various applications like Google Drive, Google calendar, Google survey tool etc, which runs on simple browser application and is highly user friendly. Any member from educational sector (students, teachers, administrative) can use it with good security systems almost free of cost. Although majority of the tools are free, but user can pay to increase the storage capacity or to experience some advanced features. Due to simplicity in the interface and user friendly, the technical adoption of Google cloud based product are high on demands from majority of the countries in educational sector [37].
- **Microsoft**: After Google, the next name is Microsoft, who has already flourished into global market by their cloud based educational product called as '*Microsoft Live@edu* [38]. This tool provides the user with a range of services exclusively for the educational sector with efficient storage capabilities. Various such applications are Windows Live Spaces, OneDrive, Windows Live Alerts, etc. Majority of these cloud based storage application are free of cost while user can upgrade to premium version by paying. Although, it is a look-alike of Google products, but there are some of the significant differences in the features in both the products that make the users to select the services based on their utilities and requirements.

- **Amazon**: After Google and Microsoft, the next big name is the most famous Amazon or popularly known as Amazon Web Services [39]. Amazon Web Services has some of the potential features which outperform both Google and Microsoft in terms of storage and processing the data. However, due to cost involved in the product usage, the name ‘Amazon’ comes in third position in our investigational study after Google and Microsoft. One of the potential products of Amazon is Amazon Elastic Computer Cloud which is equipped with MapReduce that can effectively perform the programming of the data being stored in cloud. The products offered by Amazon are also well known for their load balancing properties in peak traffic and generates better analytics that is again highly user friendly. However, the usage factor is restricted to only technical people or the individual who has necessary skills to operate it. It is not meant by common people.

There are some other big names when it comes to cloud service provider apart from Google, Microsoft, and Amazon. Table 2 exhibits the comparative analysis of the supportability of various cloud based services with respect to some parameters e.g. F_1: Better Business intelligence, F_2: Academics and its management, F_3: effective E-Learning, F_4: Enrollment supportability, and F_5: Virtual e-library.

Table 2 Existing Cloud Services for educational data

Solution	Service	F_1	F_2	F_3	F_4	F_5
Microsoft.Live@edu	SaaS-Public			■		
MicrosoftOffice.Live Workspace	SaaS-Public			■		
Microsoft Dynamics CRM Online	SaaS/PaaS-Private/Hybrid		■		■	
CampusEAI Private Cloud	SaaS/IaaS-Private		■	■	■	
Jaspersoft, Rightscale	SaaS/IaaS, Hybrid	■				
Google Docs	SaaS-Public			■		
educationERP.net	SaaS/IaaS-Private		■		■	■
Campus management	SaaS-Private	■	■	■	■	
Coupa e-Procurement	SaaS-Private					

6. KNOWLEDGE DISCOVERY FROM EDUCATIONAL BD IN CLOUD

The amount of big data is increasing exponentially in the area of educational sector with the increasing needs of the online student community as well as various anticipated technological advancement using Cloud computing as discussed in previous section. Hence, there is a need of performing proper utilization of the data that can be used for commercial growth in future. Usually, the elementary information from Big Data can be thought of exploring in its conventional datamining technique as well as data ware-housing techniques. However, the studies has been evolved with the evidences that the conventional datamining algorithms are rightly not applicable in the processing of the Big Data giving rise to impediment towards knowledge discovery process in the educational system [40].

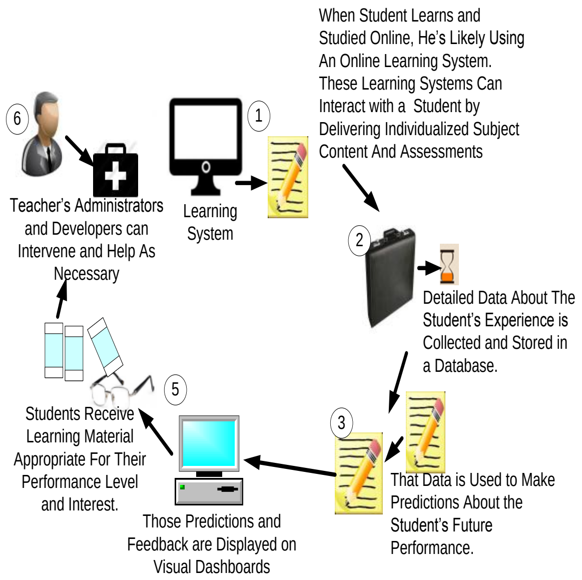

Fig 3: Process of extracting knowledge from big data from cloud

Fig.3 exhibits a possible scenario of e-learning system that gives rise to Big Data. The scenario is considered to understand the possible issues of performing knowledge discovery on it and without performing efficient knowledge discovery; there is no mean to store Big data on cloud. Hence, it is required to understand the potential benefits of knowledge discovery from the Big data that is generated from the educational sector. Some of the significant advantages are as follows:

- **Evaluating the Trends**: one of the first benefits of effective knowledge discovery from Big data is to understand and visualize the trends in the existing educational system and also to predict the same in future [41]. Various hidden patterns could be explored and identified to understand the enhancement scope of existing educational system.

- **Schematic Modelling**: As big data is generated from various individual (student, teachers, & administrative members), so it is important that each of their behaviour and experience be modelled schematically. By performing such activity, the application gets the potential ability of prediction too, that might be sought by every business organization from revenue and customer satisfaction viewpoint. The chances of the educational system enhancement are exponentially high when the policy makers are able to judge their potential investment on retaining Big Data. This information will also play a stepping stone of future direction of educational establishment in terms of enhancing their service quality.

- **Modelling Domain**: An effective extraction of unique information also assists in modelling the domain precisely for creating the essential concepts on a particular topic termed as domain [42]. Such excavation of domain will also draw a unique correlation between various components of the studies. It also assists the students to identify the probable factors that are required to be potentially built over time.

- **User Segmentation**: When various online applications are used by the students, very often it becomes difficult to perform segmentation based on the user (or student) [43]. Therefore, an effective knowledge discovery process from educational big data also ensures understanding the precise segmentation of user. The policy maker can the visualize their potential clients and can design the future products based on their requirements.

For the purpose of knowledge discovery, various cloud based service provider who manages Big data adopts various mechanism as well as technology. For an example MapReduce is frequently used by both Amazon Web Services as well as Google for the purpose of processing massive data. The prime reason behind this is that such technique (like MapReduce) can address massive set of problems from the Big data for the purpose of knowledge discovery. On the other side, various companies like IBM, Microsoft, EMC, and Oracle uses Hadoop frequently for the purpose of storing Big Data and hybridizes the analytics. It was also seen that educational sector usually deploys SaaS for storing and processing Big data. Hence, it can be justified that conventional data mining and data analytics can be enhanced in order to make it compatible with the big data processing.

7. SIGNIFICANCE OF THE STUDY

Today, when in true sense we can say that we are into a global village, where every walks of activities distributed across the globe, is highly integrated and continuously a huge amount of data being generated. The various domains of business like Finance, Social Media, E-Commerce platforms, etc., having their global customer base and it generates mega to petabytes of data, and it can grow exponentially. The situation where the available storage becomes smaller as compare to the size of data, the concept of Big Data came into a picture. Initially, the biggest task across the IT managers had to store such large generated data and further access it seamlessly as needed. In order to achieve the objective of storing unstructured data into distributed file system over a cluster of computers, a benchmark open-source framework named Apache Hadoop is introduced. The programming paradigm MapReduce and setup up the clusters on cloud makes a seamless process of Storage and Access in cost effective way. There are many further efforts are put by different organizations, consortium and researchers to overcome the technical bottleneck aspects to have better storage and retrieval mechanism with the best performance , fault tolerant, scalable. There are evidential facts that every day; our world creates approximately 2.5 quintillion bytes of data. Many business top executives believe that it does not matter how much data you have available, if it is not actionable and how business discovers the real value in large volumes of data is the key to their success.

8. OPEN ISSUES

This section present brief discussion of various open issues that needs to be overcome in the area of Big Data utilization exclusively in Educational sector. The educational sector is changing day by day giving rise to generation of various data charecteristics [44][45]. Handling Big data is just a rise of new era in current work of ubiquitous computing where the researchers need to deal with certain issues that remains unaddressed in the past attempt as following,

- **Distributed Datamining**: The conventional data mining techniques cannot be directly applied in Big data as it has enormous computational challenges, which is yet to be seen for mitigating. Number of standard research attempt in this direction was not found much.
- **Time Series Analysis**: Various applications like stock price prediction as well as meteorological predictions are based on last few years of data collection, where still the error rate persists. Hence, massive data which are generated over a large period of time is highly difficult to be made and no evidence is found in literature about it.
- **Analytics Architecture**: A very big research gap is found when it was attempted to explore a technique that creates a bridge between heuristic data and real-time data simultaneously. Although an attempt is found in literature [46], where Lamda architecture is found to have mitigation measure against this issue, but the technical adoption of this architecture and benchmarking is not witnessed in any literature.
- **Compression**: Although storage is never a big issue in Big data, however, it should be lightly taken as using storage services in cloud cost money. Significant work of Feldman et al. [47] has attempted to address compression of big data, however, the software reliability of it is still questionable.
- **Statistical Analysis**: It is said that the value of big data increases if statistical analysis can be performed on it. However, performing statistical analysis is still an open issue till date on Big data [48].
- **Visualization**: One of the most challenging task in processing big data is to create an user interface to visualize the significant information from the Big data [47]. Till date, evolutions of such user interfaces are very rare and very few.

9. CONCLUSION

There are various problems in processing big data which is strongly associated with the effectiveness of knowledge discovery process, which are briefly discussed below:

- The existing processing of Big Data doesn't emphasis on Metadata [49]. Metadata is responsible for validating data transformation with accuracy. Hence, ignoring metadata will influence processing of Big Data.
- The conventional analytics (data warehousing) are not applicable in Big data as Big data has massive volume of data which requires first processing with precise outcome [49].
- Another bigger problem identified is the non-applicability of conventional data mining algorithms [50].

Hence, due to all above issues, the knowledge discovery process cannot be ensured which will lead to degraded nature of processing the data of no use for the organization. Hence, even after investing a heavy amount of money towards storage of big data, the organization need to think about what best information they can get from the big data. Our future work will be in direction of designing a framework that could address the issues discussed in this manuscript.

10. REFERENCES

[1] IBM, Data growth and standards. Retreived from: http://www.ibm.com/developerworks/xml/library/x-datagrowth/index.html?ca=drs- [Accessed 13th March 2014].

[2] G. Press, G, 'A Very Short History Of Big Data, An Article of Forbes', Retreived from http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/[Accessed 13th March 2014]

[3] M.D. Devignes, M. Smail, E. Bresso, A. Coulet, C. Raïssi, A. Napoli, , "Knowledge discovery from biological Big Data : scalability issues", International Journal of Metadata, Semantics and Ontologies, vol. 5, Iss.3, pp.184-193, 2010

[4] B. Schmarzo, B, Big Data: Understanding How Data Powers Big Business, John Wiley & Sons, Business & Economics, pp.240 pages, 2013

[5] K. Davis, K, Ethics of Big Data: Balancing Risk and Innovation, O-Reilly Media, Computers, pp.82 pages, 2012

[6] N. Veeranjaneyulu, M.N., Bhat, A. Raghunath, “Approaches for Managing and Analyzing Unstructured Data”, International Journal on Computer Science and Engineering, Vol. 6 No. 01, 2014

[7] C.F. McCaul, B.W. Scotney, G.P. Parr, S.I. McClean, “A Cloud based End-To-End Big Data System”, PGNet, ISBN: 978-1-902560-27-4, 2013

[8] M. Peters, J. Buffington, and M. Keane, ‘Cloud Storage: the next Frontier for tape’, White paper of Enterprise Strategy Group, 2013

[9] M. Minelli, M. Chambers, A. Dhiraj, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, John Wiley & Sons, Business & Economics, pp. 224 pages, 2012

[10] J.K. Pal, “Usefulness and applications of data mining in extracting information from different perspective”, Annals of Library and Information Studies, Vol.58, pp.7-16, 2011

[11] D. Boyd and K. Crawford. Critical Questions for Big Data. Information, Communication and Society, 15(5):662{679, 2012

[12] J. Lin. MapReduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not a Nail! CoRR, abs/1209.2191, 2012

[13] N. Taleb. Antifragile: How to Live in a World We Don't Understand. Penguin Books, Limited, 2012

[14] A. Petland. Reinventing society in the wake of big data. Edge.org, http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data, 2012

[15] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001

[16] U. Fayyad. Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. http://big-data-mining.org/keynotes/#fayyad, 2012

[17] Gartner, http://www.gartner.com/it-glossary/bigdata

[18] UN Global Pulse, http://www.unglobalpulse.org

[19] Apache Hadoop, http://hadoop.apache.org

[20] P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Companies,Incorporated, 2011. SIGKDD

[21] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed Stream Computing Platform. In ICDM Workshops, pages 170{177, 2010.

[22] Storm, http://storm-project.net.

[23] Apache Mahout, http://mahout.apache.org.

[24] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0

[25] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa.cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010

[26] SAMOA, http://samoa-project.net, 2013

[27] J. Langford. Vowpal Wabbit, http://hunch.net/~vw/,2011

[28] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS:Mining Billion-Scale Graphs in the Cloud. 2012

[29] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In Conference on Uncertainty in Arti_cial Intelligence (UAI), Catalina Island, California, July 2010

[30] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung Byers, Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011

[31] N. Pirani, “New Software Personalizes College Experience,” OrangeCounty Register, September 29, 2010, <http://www.ocregister.com/news/sherpa-268815-college-students.html>.

[32] http://www.ivrsdevelopment.com/ivrs_education.htm

[33] V. Tinto, “Learning Communities: Building Gateways to Student Success,” National Teaching & Learning Forum, vol. 7, no.4(May1998),<http://www.ntlf.com/html/lib/suppmat/74tinto.htm>.

[34] http://persistenceplusnetwork.com/

[35] https://www.pa.nesinc.com/TestView.aspx?f=HTML_FRAG/PA001_TestPage.html

[36] G. Fox, Big Data in the Cloud: Research and Education, PPAM, 2013

[37] http://www.google.com/enterprise/apps/education/products.html

[38] http://www.microsoft.com/education/en-cy/solutions/Pages/live-edu.aspx

[39] http://aws.amazon.com/education/

[40] C. Romero, S. Ventura, Educational Data Mining: A Review of the State-of-the-Art, Transactions on Systems, Man, and Cybernetics, IEEE Transactions On Systems, Man, And Cybernetics, 2010

[41] R. Sallam, M. Beyer, N. Heudecker, Key trends in Big Data technologies, An article from The Connected Business, 2013

[42] R. Schutt, Big Data Domain Surfing, An Article from Introduction to data Science, Columbia University, 2012. Retreived from http://columbiadatascience.com/2012/09/11/big-data-domain-surfing-part-1/

[43] O. Hasan, B. Habegger, L. Brunie, N. Bennani, E. Damiani, A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case, IEEE International Congress on Big Data, 2013

[44] C. Parker. Unexpected challenges in large scale machine learning. In Proceedings of the 1st International Work-shop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and

Applications, BigMine '12, pages 1{6, New York, NY, USA, 2012. ACM

[45] V. Gopalkrishnan, D. Steier, H. Lewis, and J. Guszcza. Big data, big business: bridging the gap. In Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Big-Mine '12, pages 7{11, New York, NY, USA, 2012. ACM

[46] N. Marz and J. Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications, 2013

[47] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In SODA, 2013

[48] B. Efron. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010

[49] Y. Dandawate, “Big Data: Challenge and Opportunities”, Business Innovations through technology, vol.11, No.1, 2013

[50] W. Fan, & A. Bifet, “Mining Big Data: Current Status, and Forecast to the Future”, ACM-SIGKDD Explorations, Vol.14, Iss.2, pp.1-5, 2012