A Novel Educational Data Mining Approach in Cloud

Khyati Chaudhary

Abstract— **(Cloud computing is a well combination of service oriented characteristics and utility based computing. Growth rate of data in cloud environment is attaining an exponential rate. As a result, there is a need to supervise this gigantic heterogeneous data which can be both unstructured and structured in nature. This data can be managed at varied levels in cloud that is at the end user level, cloud service provider level as well as data center level. Besides, we cannot manage big data with our current methodologies or data mining software tools as big data is a word used to classify the datasets that due to their large size and complexity is not easy to manage. The big data challenge is becoming one of the most exhilarating opportunities for the next years. In the current scenario, we try to present outline of the topic, its current status, controversy and forecast to the future. Cloud computing is the budding technology which is based on pay per use model. It is computing guide where applications, bandwidth, data and IT services are provided via Internet. The purpose of the cloud service providers is to use resource skillful and attain the maximum profit. This leads to task scheduling as a core and thrilling issue in cloud computing. Scheduling is one of the procedures of determining how to compel resources between varieties of feasible tasks. In this paper, we try to elucidate various types of deadline scheduling algorithms that meet the deadline which have been surveyed and analyzed. This work explores the basic features of data mining techniques in cloud computing and secure the data using some proper method. This paper elaborate data mining techniques into cloud computing & making it a hybrid approach.**

Index Terms— **Data Mining, Cloud Computing, Big Data, Scheduling, Image Steganography.**

I. INTRODUCTION

With the progress of IT enabled products in communication and networking system, various organization like education, healthcare, financial institution, customer relationship management are today witnessing a constant and exponential rise of data in every minutes and seconds (IBM, 2014). Cloud comes from the fact that data is located far away beyond a person's reach related to the actual clouds. Cloud computing is a combination of the technologies along with a hardly any of its own unique features. It is often defined as the computing which makes use of computing resources and made resources available to the users as a service over the internet. It is also calculated as the successor of grid computing and promises to fulfill the value visualization of grid computing through its pay per use mechanism. Cloud computing is a metered service in which users can use the computing resources in a pay per use manner that means that they pay only for the part of service they are concerned. Cloud computing is a new generation technology frequently puzzled with utility computing, grid computing and distributed computing. It also promises to offer numerous features like on demand self service, broad network access along with an infinite data storage capability to its users. Cloud computing is a rationally new model next from mainframe to client, server in the early 1980s. In this pattern, details are distracted from the users who no longer have need for expertise over the technology infrastructure "in the cloud" that supports them [1]. Cloud computing illustrate a new supplement, consumption and delivery model for IT services based on the Internet and it classically involve over the Internet provision of dynamically scalable and often virtualized assets. Cloud service providers dynamically allocate servers as per need of a customer. Servers in the cloud are conceptual to the users, may be physical machines or virtual machines. Cloud also includes other resources such as storage area networks (SANs), network equipment, firewall and other security equipments whereas cloud service providers use huge data centers and powerful servers to host web applications and web services. It is Internet-based computing whereby shared resources, software and information are provided to computers. This activity commonly takes the form of web-based tools or applications that users can access and use through a web browser as if it was a program installed locally on their own computer. Cloud computing frequently appears as single points of access for all the consumers computing needs. Commercial offerings are typically expected to meet Quality of Service (QoS) requirements of customers and normally include SLAs (Service Level Agreement). SLA is an agreement between cloud service provider and customer containing various terms and conditions of the cloud service that should be followed by both the parties. Cloud computing is esteemed to many security risks such as privileged user access, data security, authentication, legal issues that need to be measured appropriately. Customers can trust cloud for their computing need if the cloud service provider have the long term feasibility. Due to these challenges, cloud customers thereby need to comprise mechanisms to measure and improve security of their information possessions operating in the cloud. However, detecting illegitimate users could be done if information about the behavior of the impersonate user is taken as a feature outline which is valid only for these user. Early and effective intrusion detection is a major factor in securing a cloud system.

An organization can extract a blend of latent transactional attributes for accelerating the business performance by collecting correct information from the big data. Precise knowledge discovery from big data will basically aid an organization to put together better decision for growth of its business. According to Press 2014, performing efficient

Manuscript received May 13, 2015

Khyati Chaudhary, Asst. Prof., Dept. of Computer Sc. & Engg. & U.P.T.U., Lucknow, India

processing on big data can obtain principal information by interpretation of the information actually. Big data also allow shaping up enhanced customer segmentation that openly support to raise the business avenues and consequently maximize the growth rate. Further, effective processing of big data can give rise to innovation which is one of the key factors in mainstream of the business goals. Big data in the form of audio, video, text, satellite image; medical images are originated from unusual business operation. Storage of big data is not at all a problem due to ubiquitous services offered by cloud. With enormous volume of data being generated, achieving efficient storage and processing it becomes a very challenging task with respect to cost and optimization of data. Such leaning of big data has raised the attention of research community as solving the key issues of supervision of big data will eventually lead to innovation and maximization of productivity and therefore a company could predict it as an opportunity to open up new avenues of business scope with the same resources. Cloud computing is an sophisticated way in which IT infrastructure, applications, services are designed, developed and delivered. In this, the value of IT resources can be exhausted on the basis of pay per use model. Numerous computing facility providers like Google, Microsoft, IBM, Yahoo and more deliver data centers in numerous localities around the world to supply cloud computing services. Cloud computing can be defined as a set of computing and communication resources located over distributed datacenters that is shared by many dissimilar users. Some of the cloud computing service models include Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). In Software as a Service (SaaS) model, a pre-made application, along with any required software, hardware, operating system is provided. In Platform as a Service (PaaS) model, an operating system, hardware and network are provided along with the customer installs or develop its own software s well as applications. The IaaS (Infrastructure as a Service) model provides just the hardware and network; the customer installs or develops its own operating systems, software and applications as well [2].

In this time of internet, e-commerce and social activities information as well as data are growing at a exceptional rate. With the fast growth of a variety of internet services and applications, there are usually huge amounts of data; the need for quickly and efficiently manipulating the datasets in a scalable and reliable way is exceptionally high. Data mining applications & techniques are very much useful in the cloud computing model. Cloud computing denotes the new tendency and practice of using a network of remote servers hosted on the Internet to store, manage, and process data comparatively than a local server or a personal computer. We can define data mining in cloud computing is the way of extracting structured information from unstructured or semi-structured web data sources. The data mining in cloud computing allows organizations to centralize the management of software and data storage, with guarantee of efficient, reliable and secure services for their users. The implementation of data mining techniques through cloud computing will let the users to regain meaningful information from virtually integrated data warehouse that reduces the cost of infrastructure and storage. Secure data transformation on internet has been a vision since the coming out of internet. Steganography is one of the solutions to securely transmit data by hiding data in data. For data used to hide data in steganography can be text or image [3]. One of the available applications of the big data is Big Data Analytics which can be defined as a means to estimate big data to obtain the latent pattern, unknown correlations, interesting facts that could be used to increase the dimension of existing business. One of such major product is owned by IBM that assist in optimize operation, improve risk mitigation measures and create novel business models. Cloud computing consists of massive redundancies and irrelevant information because of which effective processing is tremendously essential. With the use of a variety of tool that is data, text, web mining, prediction, analytics, knowledge discovery can be achieved that will considerable assist in formulating business decisions as well. Yet, the domain of big data has evolved, in recent times. This research paper will fall out about the problems being identified in using big data. The ranges of social networking applications like facebook, twitter as well as sales or marketing application are generating a massive volume of information along with growth of the dynamic user. The data are produced in the format of text, image, audio, video and many other application oriented format [4]. Although, the existence of cloud storage does not poses a big issues for big data but processing with conventional and existing data mining algorithm seriously posses a biggest threats. Implementation of processing of big data is greatly linked with potential execution issues where finding the best out of existing tool is always a challenging task. This review paper explore 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 strapping evidence in electronic media. In this paper, the recent trends and issues of big data from education sector viewpoint and discusses about the evolution of big data and make them using available tools.

II. BACKGROUND

Dr. M. Dakshayini, projected a new scheduling algorithm based on priority and admission control scheme. The priority is assigned to each new admitted queue. Admission of each queue is fixed by calculating acceptable delay and service cost. As the procedure provides the highest preference for very much paid user service requests, overall examining cost for the cloud also increases. On the other hand, Amit Nathani, proposed a algorithm that find number of slots in addition to find one slot while scheduling a deadline sensitive lease. The proposed algorithm applies on two concepts; one is swapping and other is backfilling, while rescheduling already accommodated leases to make space for a newly arrived lease. Results show that by applying swapping and multiple slots concepts, the number of accepted leases increases compared to the various algorithms. Haizea. R. Santhosh focus on providing a solution for online scheduling problem of real-time

tasks using “Infrastructure as a Service” model offered by cloud computing. The real time tasks are scheduled preemptively with the intent of maximizing the total utility and efficiency and also minimize the response time and to improve the efficiency of the tasks. The tasks are migrated to another virtual machine whenever a task misses its deadline. It improves the overall system performance and maximizes the

total utility. Florin Pop, Ciprian Dobre addresses the problem of remote scheduling of periodic tasks with deadline constrains in cloud computing. Starting from traditional addressed scheduling techniques and considering asynchronous mechanism to handle tasks, it analyze the possibility of decoupling event listening from task creation and scheduling, activities that can be placed into a peer-peer relation over a network or to client-server in cloud and consider multiple independent tasks sources that follow with a specific distribution. Nitish Chopra developed a level based scheduling algorithm which executes tasks level wise and it uses the concept of sub-deadline which is helpful in finding best resources on public cloud for cost saving and also completes workflow execution within deadlines. Dr. V. Vaithiyanathan assumes that the incoming tasks can choice their method on the basis of task requirement like minimum execution time or cost and then it is prioritized. The algorithm is named as Task Selection Priority Deadline. The proposed work is experimented and evaluated on cloud-sim toolkit [5]. Outcomes validate the accuracy of the framework and show a momentous improvement over other scheduling methods. Quentin Perret, minimizes the extra-cost implicit from tasks that are executed over a cloud setting by ordering using its laxity and locality. By using CLLF, deadlines are minimized while the total execution time of the job remains in acceptable levels**.** It also manages the locality of task.

A. DIFFERENT TYPES OF CLOUDS

Different types of clouds are:

1. Public Cloud

In a public cloud, the computing resources are made available to the general public. Microsoft, Google and Amazon are some of the public cloud service providers.

2. Private Cloud

Private cloud is a cloud setup in which the cloud infrastructure is developed for a particular organization. Microsoft and HP are some of the private cloud providers.

3. Community Cloud

Community cloud is a cloud setup which is provided for two or more organizations having common interests.

4. Hybrid Cloud

Hybrid cloud is a cloud which is a combination of different types of clouds. Consequently, it can be a combination of public and private or hybrid community.

B. CHARACTERISTICS OF CLOUD

Some of the distinguished characteristics of cloud that makes it better than any other technology available at present are:

1. On Demand Service

Since in cloud computing, computing resources are provided as a service on a utility like basis, as a result the users can demand these services and when needed without any human interference.

2. Agility

Cloud computing provides its users with a very responsive approach. The users can rapidly arrange their applications without worrying about the initial setup.

3. Measured Service

Cloud computing is a measured service in the terms of the computing resources can be provided in a utility basis and the users pay only for the part of service they use.

4. Rapid Elasticity

Cloud computing offers rapid elasticity i.e. the users can allocate and reallocates resources in an elastic manner as per their requirements.

5. Resource pooling

In Cloud Computing, the same resources can be provided to two or more clients as per their requirements. It provides a facility of resource pooling.

6. ***Multi-tenancy***

In cloud, multiple users share the same set of resources in elastic and a scalable manner, by this means employing multiple tenants. Apart from several advantages which cloud computing offers the data storage in cloud is often considered to be insecure because data is held beyond the reach of data owners. But the features offered by cloud are far more worthwhile then its limitations and hence cloud computing is now in the boundary of offering Database as a Service (DaaS) to its end users. Here data base required for hosting information is made available to its end-users as a service. This database as a service is gradually attracting customers but management of cloud data is very different from management of traditional data. Management of data using traditional methods which are expensive is now slowly becoming obsolete in cloud based environments. Therefore in order to cater to the requirements of management of data in cloud many companies have come up with their own solutions for management of data in cloud based environments such as Big Table, Google File system, Map Reduce, Cassandra and Amazons Simple Storage Service(S3).

III. EVOLUTION OF BIG DATA

With the constant rise of 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 excessive higher degree of mobility due to ubiquitous computing. Therefore, it becomes easier for the user to access their resources from multiple computing devices. Such communication pattern is also bidirectional. Hence it can be said that on every seconds, millions of data are being generated and stored in server. This is the way the data are growing in every seconds of life, which is now technically called as "Big data." As the information carried out by big data is highly valuable for business goals and it gives a cut edge prediction capability to the organization by excavating more secrets about their data which cannot be explored by conventional data mining or any data warehousing techniques.

A. NEED OF CLOUD COMPUTING AND STEGANOGRAPHY

Cloud computing is the term used as a metaphor for the Internet and cloud computing is a kind of distributed computing prototype where different services such as servers, storage and applications collectively known as configurable computing resources that are rapidly equipped and released with minimal management efforts. The cloud computing model allows access to information and computer resources from anywhere. With cloud environment, data can be

replicated in different locations so it provides disaster recovery servers and data storage at low cost.

However, cloud is best solution to match uncertain scaling demands. Organizations can store plenty amount of data even more than what they can store on private computer systems or servers. The scope of the research work is to extract the useful information from large amount of data and store at cloud in secure fashion and then make predictions required by the organization. But that prediction which is generated as a result of mining should be secure from interception [7]. In this way, steganography is the best option for sending information secretly because it hides the survival of secret message and provides more security. The security module which is used as image steganography. As images are the most popular because of their frequency on the Internet, so the prime focus is to increase capacity to provide better security during transmission. In this work, Edge detection based approach for image steganography has been reviewed. Canny edge detection method produces single pixel thick, continuous edges even in noise conditions that are the ability to detect true weak edges. Because editing in edge areas cannot be detected well by human eye but editing in smooth areas can be detected easily. Small size companies can reduce their capital and operational expenditure for their increasing computing needs as no longer do IT professional need to worry about keeping software up to date.

IV. DATA MINING

With the beginning of 21st century, it has been a recognized fact that we are in an information technology driven society where knowledge has proved to be a precious asset to any individual, organization or government. Data mining provides that boost to enable organizations to turn their huge amounts of data into expensive information or knowledge that not only could meet customer demands but also keep the future scenario in check. Data mining is defined as a "type of database analysis that attempts to discover useful patterns or relationships in a group of data."

Data mining tracks these patterns and relationships using data analysis tools and techniques to build models. Organizations that wish to endure to these changes should be proactive and have to anticipate what their current and prospective customer desire. Association rule mining addresses the bonding of togetherness or connection of objects also known as association rule. It finds rules connected with frequently co-occurring items. It is mostly used for market basket analysis, cross-sell, root cause analysis, in-store placement, and defect analysis. Association rule mining is usually divided up into two parts: Firstly, minimum support is applied to find all frequent item sets in a database. Secondly, these frequent item sets and the minimum confidence constraint are used to form rules. There are a group of different association rule algorithms. Apriori algorithm is the simplest and best-known for association rules mining. It is a significant algorithm for mining frequent item sets for boolean association rules. The business environment faced by all organizations has changed a lot with customers becoming more demanding in terms of their needs and in terms of products as well as services that they require.

V. AVAILABLE TOOLS

As the work of production and research community has already started exploring about big data. Some tools that is publically available in processing big data as follows,

(i) ***Apache Hadoop****:* It is a basically an open source software framework that assist in storing massive volume of data as well as perform processing on the big data. Such frameworks are associated by another name "MapReduce", which is a programming tool on processing massive volume of data residing in clusters.

(ii) ***Apache S4:*** It is the product of Apache software foundation that addresses the issues of complex proprietary and batch oriented open source frameworks.

(iii) ***Storm:*** It is a software framework for streaming data-intensive distributed applications which is similar to S4 and developed by Nathan Marz at Twitter. Some of the significant open source projects for addressing big data mining issues are as follows:

(i) ***Apache Mahout****:* It has enriched with vast range of machine learning and data mining techniques that is clustering, classification, collaborative filtering and frequent pattern mining. It is one of the scalable machine learning open source software designed exclusively with Hadoop.

(ii) ***R****:* It is an open source programming language and software environment designed for statistical computing developed by Ross Ihaka and Robert Gentleman and is deploy for analyzing statistically very large data sets.

(iii) ***MOA****:* It stands for massive online analysis that performs data mining in real time streaming data. It uses the potential features such as classification, regression, clustering and frequent item set mining and frequent graph mining. It hass advanced version called as SAMOA is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA.

(iv) ***Vowpal Wabbit****:* It is basically an open source project under the joint imitativeness of Yahoo and Microsoft for developing a scalable and efficient learning algorithm.

Some other specific open source tools to Big Graph mining are:

(a) ***Pegasus****:* It is a big graph mining system that is being designed using MapReduce and permits determining the patterns and anomalies in massive real-world graphs.

(b) ***GraphLab****:* It is a high-level graph-parallel system built to compute over dependent records that are recorded as vertices in a large distributed data-graph. Some of the algorithms in GraphLab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices [8].

VI. BIG DATA IN EDUCATIONAL SECTOR

This work basically intends to discuss about the trend of the big data that is generated from educational sector. 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. 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. The 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 have been marked responsible for generation of big data in educational section are as follows:

(a) Academic Trend: The 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 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. It also assists the students to select their prime classes to be undertaken along with schedule management.

(b) Futuristic Technology: One of the biggest levels 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. 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.

(c)Trend of Academic Lives: Various online tutorials and model applications already exist where the login privileged 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. Such types of tools markedly boost the morale of the students by helping them to shape up their academic life in better synchronicity with the ongoing courses and activities in their educational establishments. Practicing of such tools also generates quite a substantial data that are hard to be stored in ordinary servers. It has been supported by plugin with existing mobile devices; a student finds it quite comfortable and logical to adopt it for their educational betterment.

(d) 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.

(e) Virtual Classes: As 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.

VII. STORAGE OF EDUCATIONAL BIG DATA IN CLOUD

There are various tools as discussed above which 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. Storage of big data is never a big challenge which could substantially ensure Return OF Investment (ROI). Big data is generated from teaching domain, administrative domain as well as research domain when it comes to educational sector. Following are the prominent existing cloud service provider that caters up the storage requirement of big data being generated by educational sector:

(i) ***Google****:* Google provides various standard applications based on web-based interfaces and storage system that is operated by user on any client-specific applications that is, Firefox. Google already has various applications like Google Drive, Google calendar, Google survey tool etc. which runs on browser application and is highly user friendly. Any member from educational sector for example, students, teachers, administrative can use it with good security systems almost free of cost. Even though 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, technical adoption of Google cloud based product are high on demands from majority of the countries in educational sector.

(ii) Microsoft*:* Microsoft who has cloud based educational product called as Microsoft Live@edu. This tool provides the user with a range of services absolutely for the educational sector with efficient storage capabilities. Some of the 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.

(iii) Amazon*:* Amazon web services have some of the latent features which can surpass both Google and Microsoft in terms of storage and processing the data. One of the potential products of Amazon is Amazon Elastic Computer Cloud which is well equipped with MapReduce that can effectively

perform the programming of the data being stored in cloud. Yet, the usage factor is restricted to only technical people or the individual who has necessary skills to operate it.

VIII. KNOWLEDGE DISCOVERY FROM EDUCATIONAL BIG DATA IN CLOUD

The amount of big data is growing exponentially in the area of educational sector with the rising needs of the online student community as well as various probable technological advancement using cloud computing. Consequently, there is a need of performing proper deployment of the data that can be used for commercial growth in future. Generally, the elementary information from big data can be thought of exploring in its conventional data mining technique as well as data ware-housing techniques. Scenario is measured to recognize the possible issues of performing knowledge discovery on it; there is no mean to store big data on cloud [9]. Thus, it is required to know the possible benefits of knowledge discovery from the Big data that is generated from the educational sector. Some of the major advantages are as follows:

(i) Evaluating the Trends: One of the profit 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. A variety of hidden patterns could be explored and well-known to understand the improvement scope of existing educational system.

(ii) Schematic Modelling: As big data is generated from various individual that is student, teachers & administrative members, so it is significant that each of their behaviour and experience be modelled schematically. This information will also play a stepping stone of future direction for educational establishment in terms of enhancing their service quality. By performing such activity, the application gets the latent ability of prediction that might be required by every business organization from revenue and customer satisfaction viewpoint.

(iii)Modelling Domain: An extraction of unique information assists in modelling the domain specifically for creating the necessary concepts on a particular topic termed as domain. 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.

(iv)User Segmentation: When various online applications are used by the students, oftenly it becomes difficult to perform segmentation based on the user. Therefore, an efficient knowledge discovery process from educational big data also ensures understanding the precise segmentation of user. For this purpose of knowledge discovery, a variety of cloud based service provider who manages big data adopts various mechanism as well as technology [10]. For example, MapReduce is frequently used by both Amazon Web Services as well as Google for the purpose of processing massive data. The major reason behind this is that such technique like MapReduce can address massive set of problems from the big data for the principle 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. Thus, it can be defensible that conventional data mining and data analytics can be improved in order to make it compatible with the big data processing.

IX. CONSEQUENCES OF THE STUDY

In current scenario, we can say that we are into a global community where every walk of activities distributed across the globe is integrated and constantly a huge amount of data being generated. Various domains of business like finance, social Media, e-commerce platforms etc., having their global customer base and it generate mega to peta bytes of data and it can grow exponentially. The situation where the accessible storage becomes smaller as compare to the size of data, the concept of big data came into a depiction. Initially, the biggest task across the IT managers had to store such large generated data and further access it flawlessly 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 as well as setup up the clusters on cloud makes a faultless process of storage and access in cost effective way. There are many efforts that 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 details that every day, our world creates approximately 2.5 quintillion bytes of data. Many business top executives suppose that it does not matter how much data is available, if it is not actionable and how business discovers the real value in large volumes of data is the key to their success [11].

X. SOME OPEN ISSUES

Brief discussion of a variety of open issues that needs to be overcome in the area of big data utilization exclusively in educational sector. Educational sector is changing day by day giving rise to generation of different data characteristics. Handling big data is just a rise of new era in existing work of ubiquitous computing where the researchers need to deal with definite issues that remains unaddressed in the past effort are as follows:

(a) Distributed Data mining: Conventional data mining techniques cannot be directly applied in big data as it has huge computational challenges which are yet to be seen for justifying. Number of research attempt in this direction was not found much.

(b) Time Series Analysis: A range of 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 great period of time is exceedingly difficult to be made.

(c)Analytics Architecture: A very big research gap has been found when it was attempted to discover a method that creates a bridge between heuristic data and real-time data at the same time.

(d) Compression: While storage is never a big issue in big data however, it should be lightly taken as using storage services in cloud cost money.

(e) Statistical Analysis: Value of big data increases if statistical analysis can be performed on it. Still performing statistical analysis is still an open issue till date on big data.

(f) Visualization: One of the most challenging tasks in processing big data is to create user interface to imagine the

significant information from the big data. Evolution of such user interfaces are very rare and very few.

CONCLUSION

There is a diversity of problems in processing big data which is strongly connected with the effectiveness of knowledge discovery process which are briefly discussed as follows:

(i) The existing processing of big data does not emphasis on metadata. Metadata is responsible for validating data transformation with accuracy. For this reason, ignoring metadata will influence processing of big data.

(ii) Conventional data warehousing are not applicable in big data as big data has enormous volume of data which requires first processing with precise outcome.

(iii) Another bigger problem identified is the non-applicability of conventional data mining algorithms.

Therefore, due to all the 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 intense amount of money towards storage of big data.

Scheduling is the fundamental matter in the management of application execution in cloud environment. In cloud computing environment, number of different resources is delivered as a service in the method of virtual machines and these machines are scheduled by scheduling algorithm. Typically they all are work on to minimize the execution time, reduces cost and meet the deadline. Hence, we need to develop scheduling algorithm for cost optimization.

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] 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

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

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

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

[8] 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

[9] 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

[10] N. Pirani, "New Software Personalizes College Experience," Orange County Register, September 29, 2010, <http://www.ocregister.com/news/sherpa-268815-college-students.html>.

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