Big Data for Education in Students' Perspective

G. Vaitheeswaran
Research Scholar,
Department of Computer Science,
St. Joseph's College (Autonomous),
Tiruchirappalli, Tamilnadu, India.

L. Arockiam, Ph.D.
Associate Professor,
Department of Computer Science,
St. Joseph's College (Autonomous),
Tiruchirappalli, Tamilnadu, India.

ABSTRACT
Big Data Analytics is the new technology for extracting hidden information from the large datasets or data deluge, due to its volume, variety, and velocity. This paper presents the overview of big data, its available technologies and tools and discusses the open issues of big data. Big data plays a significant role in education sector. Everything has become digitalized in the educational institutions, which leads to store and process enormous amount of data. Handling the huge amount of data is complex. The main focus of this paper is to propose a new approach to analyze the large streaming data that produced by web server logs of educational institution. The result represents the student's web usage behavior, which supports to make better decisions to improve the student's performance and suggest recommendations for their academic perspectives.

Keywords
Big Data, Big Data Analytics, Web Usage Mining.

1. INTRODUCTION
The Internet Technology makes possible to share and access data from anywhere at any time around the world. Data generating by the users are massive and unpredictable. Everything is digitalized and stored in the repository. The growth of data will never stop. For example, the Yahoo warehouse totaled 170 petabytes (8.5 times of all hard disk drives created in 1995) [1]. The IDC survey reported that, the digital universe will grow by a factor of 10 from 2013 to 2020, i.e., from 4.4 trillion gigabytes to 44 trillion. It more than doubles every two years [2]. This massive growth of data leads to the new technology innovation called Big Data. Big Data is the buzz word, where it can be described with the following characters such as Volume (size of data), Velocity (streaming data), Variety (sources of data such as text, images, videos, audios, etc.), Veracity (uncertainty data) and Value (decision making). According to the leading research team Gartner [3] defined big data as "**Big data** is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." These enormous amounts of data makes possible to analyze and delivers better decision support. Thus big data becomes a hot topic, since technologies such as modern data mining technique, machine learning, and statistical models are help to analyze the available data for the business intelligence. Big data plays major role in all the fields, since processing of data is growing day-by-day. The prominent applications of big data include weather forecasting, social networks, telecommunications, agriculture, e-governance, e-commerce, and education, health care and so on. In the earlier days organization generated data and public consumed the data. Nowadays public generates and consume the data.

Before the emerging of cloud computing and internet of things, organizations have generated structured data and used batch processing technique to place summaries of the data into traditional relational databases. The analysis of such data is exposition and the investigations done on the datasets are on past patterns of business operations. In recent years, new technologies with lower costs have enabled improvements in data capture, data storage and data analysis. Organizations can now capture more data from many more sources and types (blogs, social media feeds, audio and video files). The options to store and process the data have expanded dramatically, and technologies such as MapReduce and in-memory computing provide highly optimized capabilities for different business purposes. The analysis of data can be done in real time or close to real time, acting on full datasets rather than summarized elements. In addition, the number of options to interpret and analyze the data has also increased, with the use of various visualization technologies. All these developments represent the context within which "Big Data" is placed.

Figure 1 [4], represented the big data opportunity on the various sectors. Big data analytics is the process of examining the huge amount of different data types, to discover hidden patterns and measurable outcome for business impact. Big data affects educational sector the same way it does with other sectors. Big data analytics plays major role in educational sector. Big data analytics brings the potential to transform the study of educational data by helping the analysts more efficiently and to deliver more informed conclusions. Educational data includes such as student's enrolment data, fee collection, attendance, e-library, mark statements, lab records, staff profile, etc. This paper discusses about the problems in educational sector and introduces an analytical framework to analyze the student behavior using the large datasets. 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 pose a big issue for big data. The present paper discusses about the trends of big data with respect to educational sector over cloud platform.

This paper discusses about the recent trends and issues of big data in education sector in students' perspective. Section 2 discusses about the evolution of big data and controversies of big data. Section 3 discusses about the role of big data in the educational sector. Section 4 discusses about the proposed approach for analyze the students' behavior. Section 5 discusses about the technologies and tools available for big data. The challenge and issues of big data is discussed in section 6 and finally section 7 concludes the paper by summarizing more issues on the topic.

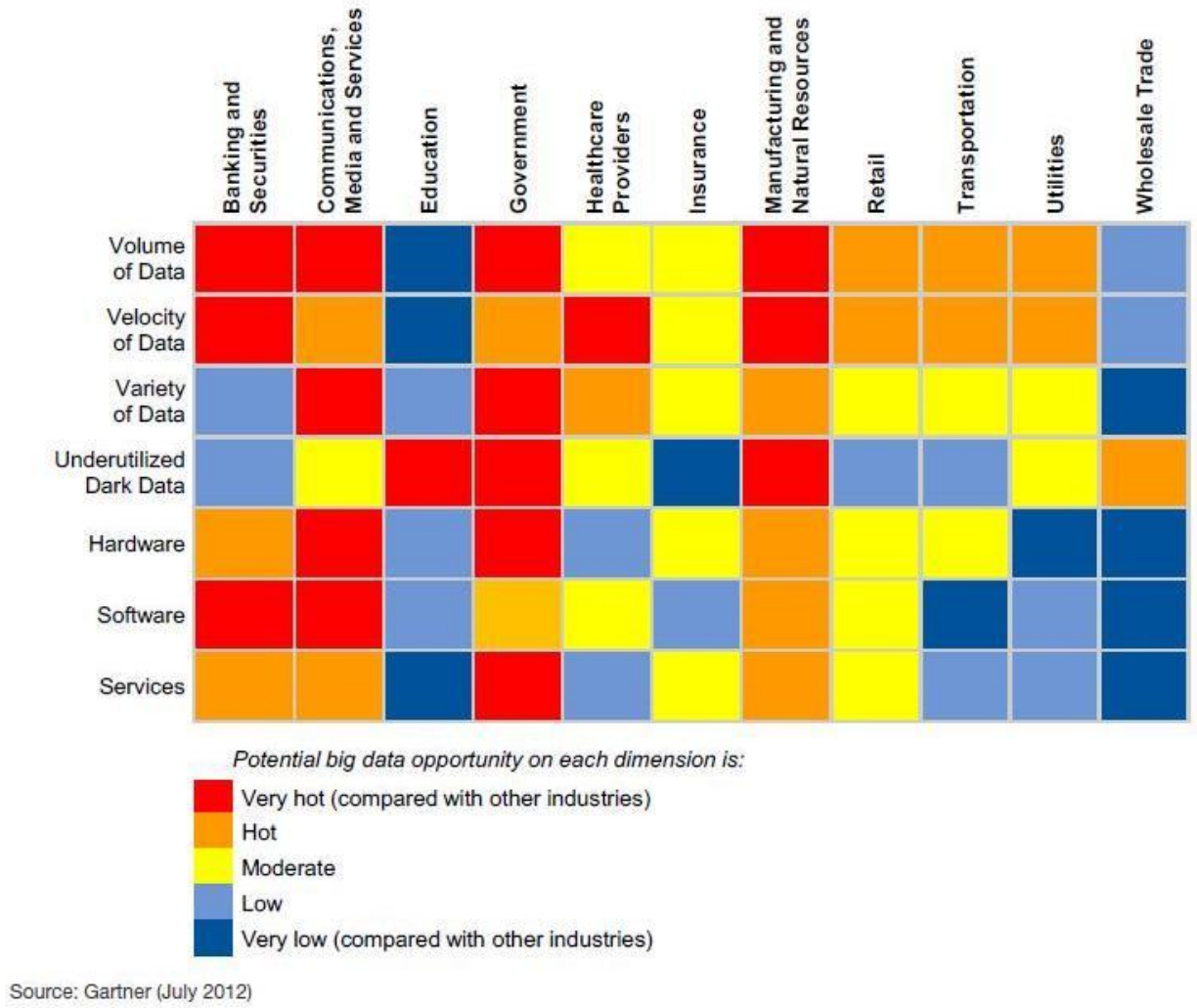

Fig 1: Big Data Opportunity in Various Sectors

2. "BIG DATA" – EVOLUTION AND CONTROVERSY

In olden days, the computers were used for some computational process. This section discusses the gradual development of data management system that happened for the last 5 decades. Since the 1960s, database and information technology has been evolving systematically from primitive file processing to sophisticated and powerful database systems. Later in 1970s, the data management system was introduced. Advance Database Systems (extended relational, object-relational, etc.) was introduced mid-1980s (reference). In late-1980s, the advanced Data Analysis was introduced. In this period the evolution of Data Warehousing and Data Mining has emerged [5]. The web based databases (XML-based) was introduced in 1990s. The steady and incredible progress of computer hardware technology in the past four decades has led to large supplies of powerful and affordable computers, data collection, equipment, and storage media. This technology provides a great boost to the database and information industry, and makes a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis. Later in 2000s the evolution of internet technology Cloud Computing was introduced. This technology leads to produce the massive amount of data. The emerging of social network such as Facebook, LinkedIn, Twitter, Google+, etc. plays major role in generating huge amount of data. The usage of computer has been increased unexpectedly due to the innovation of new technologies. Big data has been the buzz word in public-sectors for just a few years now, but its roots go deep. Figure 2, presents the evolution of big data based on the problems rose on the 3V's (Velocity, Variety, Volume).

As Big Data is a new hot topic, there have been a lot of controversies about it. The problem of handling huge amount of data was discussed in various situation and new technologies were introduced to overcome those situations. The buzz word 'Big Data" came to explore after the evolution of cloud computing, due to the incredible data deluge that drowning the world. There is no need to distinguish Big Data analytics from data analytics, as data will continue growing, and it will never be small again. Bigger data are not always better data. Since most of the data generate by the social networks are not real data. In real time analytics, data may be changing. In that case, volume is not important, than velocity (dynamic change of the data). Big data is not about hadoop and Mapreduce, there are other technologies that are in development progress for analyzing huge amount of data.

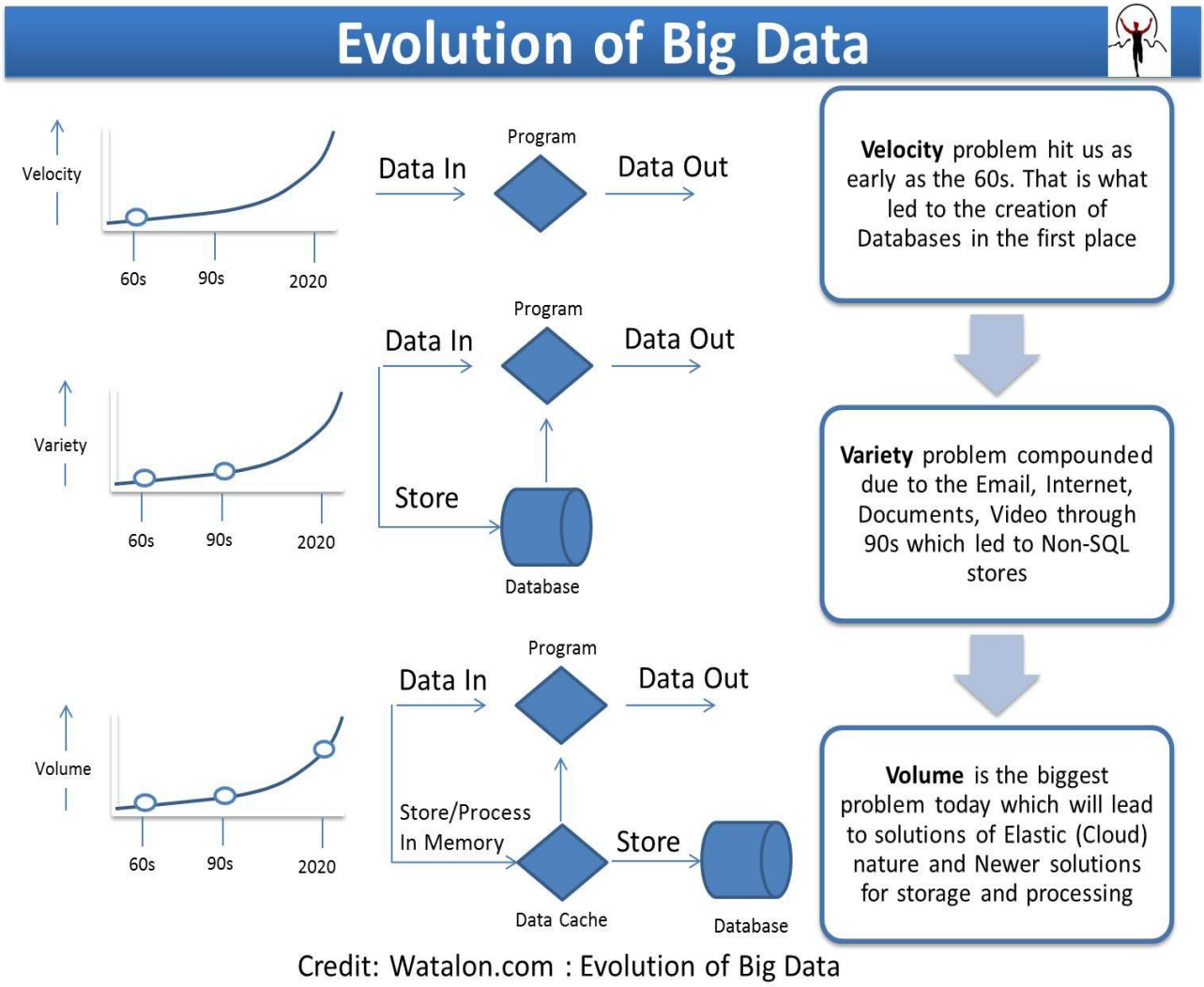

Fig 2: Evolution of Big Data

3. BIG DATA IN EDUCATION

In educational sector, the need and processing of data is growing day by day. Increase in the advancement of the internet technology replaced "chalk and talk" system in the education institutions. The impact of ICT carried out all the data processing such as enrollment to result declaration in web based applications. Most of the modern and esteemed higher technical educational institution has already adopted various applications to make the complete administrative work go paperless. This adoption of ICT is found to be very promising and saves lots of time against doing work and decrease stress on more skill development and better communication. The following factors are leads to the generation of big data in educational institutions [6]:

Academic Trend: Majority of the higher educational institution are now introducing various customized tools for the purpose of various administrative jobs like students admission enrolment, fee collection, teachers and other employees profile maintenance, reporting system etc. Student's regular activities such as internal assessment, attendance records, mark statements, computer laboratory workloads, etc. produces enormous amount of data.

Performance Monitoring: The institutions are supposed to maintain the entire passed out student reports in order to perform statistical report. Applying predictive analytics on the passed out student reports, the institution can predict the upcoming student activities. Various tools exists currently that considers the academic or skill sets of the on going student and assist them to undergo virtual assessment to understand where do they stand in viewpoint of classroom exercise or job interview. The outcome of the predictive analysis provides great impact for the fore coming student to improve their lacking skills to get dream job.

Virtual Classes: [6] With the rise of more competition in academics, the students are regularly 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 advanced courseware. 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 produce enormous data, which could never reside in conventional server and need to take the aid of cloud based storage system.

4. BDA APPROACH FOR PREDICTING THE STUDENTS' BEHAVIOR

The roles of big data in educational sector are discussed efficiently in the previous topic. Big data analytics (BDA) is the process of examining the huge amount of different data types, to discover hidden patterns and measurable outcome for business impact. Hence, applying the big data analytics in the education sector will provide valuable insights to improve the student knowledge and their academic functioning. Big data analytics will play a major role in the education sector. The existing model analyzed the student behavior based on their academic performance [7]. Nowadays students like to spend more time before the internet. Since internet technology makes possible to get all the resources in and around the world at any time. They download e-content such as e-books, magazines, articles, video and audio lectures, etc. The proposed framework model is for examining the student behavior based on surfing the internet. Figure 4, represents the overall process of the proposed framework.

4.1. Process of the proposed model

The proposed framework model consists of three phases. They are discussed below:

4.1.1. Phase – 1: Authentication process:

The institution provides separate login for each student. Each student must have to login with their user name and password to browse the internet. This phase will help to monitor the student login activity and prevent the intruders to access the internet. According to the academic level (Undergraduate and Postgraduate and Research), each student will categorize by some privileges. This also allows the student to browse, based on the privilege granted to them.

4.1.2. Phase – II: Maintaining Web Server Logs:

Maintaining log files produces huge amount of data. The web log contains both the offline and online (streaming) data. This situation leads to the "Big Data". Handling the enormous amount of data is complex process. Log files in web servers maintain different types of information, discusses below. Table 1 [8], presents the contents of the web logs. These contents are the raw data for analyzing the web logs.

Table 1: Web Server Log Format

Variable	Contents
User name	Name or IP address of the students.
Path Traversed	Identifies the navigation path taken by the user with in the web site.
Time & Date	Identifies the time & date of the user login.
Time stamp	Time spent by the student in each web page.
Page last visited	The page that was visited by the student before leaves the website.
Bytes	Stored the information of downloads and number of copying activity.
User Agent	This stores information about the browser from where the student sends the request to the web server.
Request type	Methods (GET, POST) used for information transfer is noted.

4.1.3. Phase – III: Analyzing the Web Server Logs:

Student access the cloud services, e-content and local database through the internet and the logs are maintained in the web server. There are two main approaches to analyze the huge amount of web log data, they are:

- Batch processing: This technique is used to analyze the historic or offline data.
- Real-time processing: This technique is used to analyze the online/streaming data. During the days before examination, the access of internet seems higher than the normal days. This access produces the streaming web log data.

The web logs contain noisy data. Using the web usage mining technique [9], the raw web logs are cleaned and user segmentation will be identified. The cleaned data may further convert into the required data format. For example, MOA (Massive Online Analysis) provides various techniques and algorithms to analyze the streaming data [10]. Using the student web usage logs and implementing the stream mining techniques, the log data may categorize by most viewed sites, most listened lectures and accessed e-books. Through this we could recommend or suggest useful information and websites by using the recommendation analysis. We can also compare the student performance based on their internet usage and the marks obtained in the academic examination. This result can improve the student academic career. Web usage mining consists of three main steps [11]:

- Preprocessing
- Pattern Discovery
- Pattern Analysis

Figure 3, represents the process of the web usage mining to extract the hidden information from the web logs.

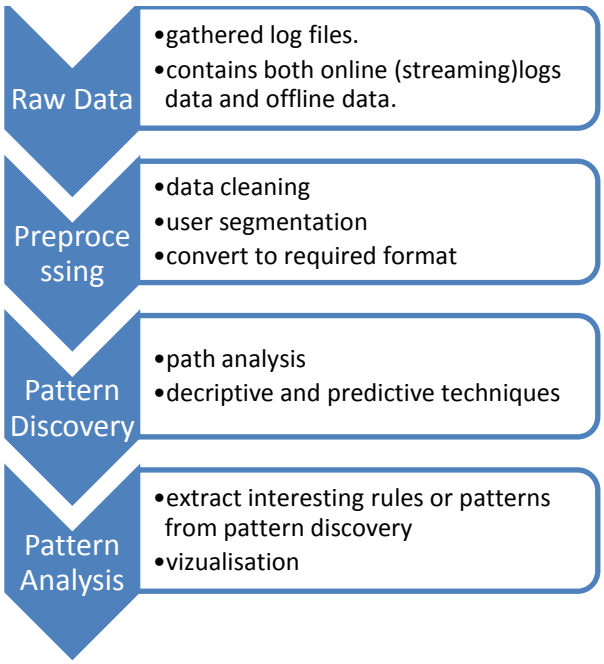

Fig 3: One High Level Web Usage Mining

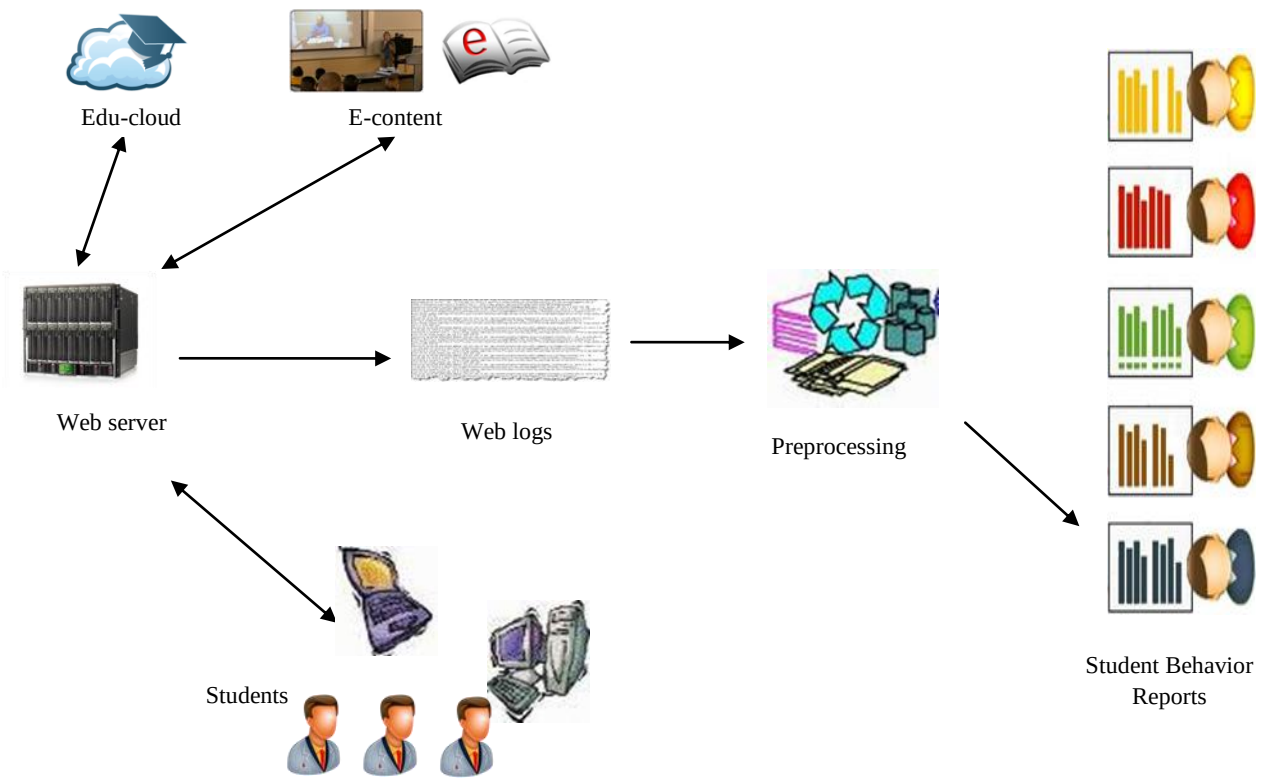

Fig 4: Approach to Analyze the Student Behavior

5. TECHNOLOGIES AND TOOLS

As we enter the "petabyte age", the existing data analysis approaches exhibits their own limits. Data analysis is the process of examining available data in order to extract useful information. Decision makers commonly use this information to drive their choices. The quality of the information extracted by this process greatly benefits from the availability of extensive datasets. The big data analytics plays major role in analyzing the huge amount of datasets. According to Giga Spaces [12], nearly 80 per cent people in IT are either using or planning to use dedicated big data tools to manage massive amounts of data in their organization. Some of them design their own applications or tools to store and retrieve the huge amount of data. Some of the practicing technologies and tools are used for big data analytics are listed below:

5.1. Existing Technologies

- Column-Oriented Databases.
- Schema-less Databases or NoSQL Databases.
- Data fusion and integration.
- Data mining techniques.
- Statistical Learning
- Genetic Algorithms.
- Machine Learning.
- Natural Language Processing.
- Time series analysis.
- Visualization.
- MapReduce.

5.2. Open Source Tools Available for Big Data [13]

Apache Hadoop is an open source software framework for data-intensive distributed applications. Hadoop distributes the enormous of data in pieces over a series of nodes running on commodity hardware by using MapReduce. Now it is the most popular technologies for storing the structured (traditional database), unstructured (images, videos, audio, etc.) and semi-structured (XML) data that comprise Big Data. Hadoop is available under the Apache License 2.0. It is associated with other framework like Hbase, Hive, Pig, Zookeeper, Sqoop, Yarn, etc.

R is an open source statistical software environment and programming language designed for statistical computing and visualization. R has been commercialized by the company, Revolution Analytics, which is pursuing a services and support model encouraged by Red Hat's support for Linux. It is available under the GNU General Public License.

Scribe is a server developed by Facebook and released in 2008. It is proposed for aggregate log streaming data in real time from a large volume of servers. Facebook designed it to meet its own scaling challenges, and it now uses Scribe to handle tens of billions of messages a day. Scribe is available under the Apache License 2.0.

ElasticSearch is a distributed, powerful open source search server. It's a scalable solution that supports near real-time search and multitenancy without a special configuration. It has been adopted by a number of companies, including Mozilla and StumbleUpon. It is available under the Apache License 2.0.

Apache Cassandra is an open source distributed database management system developed by Facebook to power its Inbox Search feature. Facebook abandoned Cassandra in favor of HBase in 2010, but Cassandra is still used by a

number of companies, including Netflix, which uses Cassandra as the back-end database for its streaming services. Cassandra is available under the Apache License 2.0.

MongoDB is another popular open source NoSQL data store. It stores structured data in JSON-like documents with dynamic schemas called BSON (for Binary JSON). MongoDB has been adopted by a number of large enterprises, including MTV Networks, craigslist, Disney Interactive Media Group, The New York Times and Etsy. It is available under the GNU Affero General Public License, with language drivers available under an Apache License. The company 10gen offers commercial MongoDB licenses.

Apache CouchDB is still another open source NoSQL database. It uses JSON to store data and JavaScript as its query language. MapReduce and HTTP are used for an API. The BBC uses CouchDB for its dynamic content platforms, while Credit Suisse's commodities department uses it to store configuration details for its Python market data framework. CouchDB is available under the Apache License 2.0.

Apache Storm is a free and open source distributed real-time computation system. Storm makes it at ease to reliably process limitless streams of data, doing for real-time processing. Storm is simple and it can be used with any programming language.

6. 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 major challenges for researchers and companies to handle big data [14],

Analytics Architecture: The existing analytic architecture is used for analyze the historic data using batch processing techniques. Nathan Marz [15], introduced lambda Architecture for analyzes the streaming/real-time data.

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: Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two 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 [16].

Visualization: A main task of BDA 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 in the beautiful book "The Human Face of Big Data"[17].

The other open issues in research perspectives:

- Big Data Science and Foundations.
- Big Data Infrastructure.
- Big Data Management.
- Big Data Search and Mining.
- Big Data Security & Privacy.
- Big Data Applications.

The other major challenges [18] includes not only the issues of scale, but also heterogeneity, lack of new data structure, error-handling, security and privacy, timeliness, and visualization, at all levels of the study from data acquisition to result interpretation. These technological challenges are common across diversity of application domains, and it is not cost-effective to concentrate on the context of one domain alone.

7. CONCLUSION

This paper discussed the impact of big data and cloud computing in the education sector. On the other hand, many technical challenges are discussed in this paper. This paper has proposed a framework to improve the academic performance of the student by analyzing the web logs. In future, our work will focus on collaborating cloud services and big data analytics in the educational sector, the collaborative model will be cost-effective and supports better decision making in the educational sector.

8. REFERENCES

[1] https://www.ida.gov.sg/~/media/Files/InfocommLandscape/Technology/TechnologyRoadmap/BigData.pdf, as on 08-5-2014

[2] IDC. the 2011 Digital Universe Study: Extracting Value from Chaos, http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf, as on 08-05-2014.

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

[4] Infocomm Development Authority of Singapore, "Big Data", 30, Nov 2012.

[5] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", chapter - 1

[6] D. Pratiba, G. Shobha, "Educational BigData Mining Approach in Cloud: Reviewing the Trend", IJCA *(0975 – 8887)*, Volume 92 – No.13, April 2014.

[7] R. Sallam, M. Beyer, N. Heudecker, "Key Trends in Big Data Technologies, An Article from the Connected Buisness", 2013.

[8] http://httpd.apache.org/docs/current/logs.html, internet source as on 07/05/2014.

[9] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai, "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & Its Applications", IJNSA, Vol.3, No.1, January 2011.

[10] 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.

[11] L.K. Joshila Grace, V. Maheswari, and Dhinaharan Nagamalai (Jan 2011)"Web Log Data Analysis and Mining" in Proc CCSIT-2011, Springer CCIS, Vol 133, pp 459-469

[12] White Paper, "Big Data Survey", Gigaspaces, 2012.

[13] http://www.cio.com/slideshow/detail/51062, internet source on 07/05/2014.

[14] 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.

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

[16] 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.

[17] R. Smolan and J. Erwitt, "The Human Face of Big Data", Sterling Publishing Company Incorporated, 2012.

[18] A community white paper developed by leading researchers across the United States "Challenges and Opportunities with Big Data", 2012.