Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analytics in Education Using Big Data

Sachin Sharma, Diksha Sharma, Pankaj Vaidya
Department of CSE, Shoolini University,
Solan, India

Abstract—Big Data is an emerging technique from past year. Tremendous amount of data which can't be managed easily by our traditional software applications such as (RDBMS, OODBMS).In this area a lot of domain comes under such as business, transport, biology and education etc. In this paper we are presenting about Big Data in education analytics, methods and technology which are helpful for improving education system in both perspective such as learners and leaning both.

Keywords—Big Data; OODBMS; RDBMS; hadoop; EDM; learning analytics, data abundance.

I. INTRODUCTION

Big Data consist of data which is growing rapidly and which is dynamic in nature so it becomes difficult to manage such tremendous amount of data. Every day data growing rate is quite high. Data comes from all the social media sites such as facebook, Flickr, Google+, emails, video, audio etc contributes a lot to Big Data. Some of the data is Structured, Un-structured and semi-structured. Extracting information from rows and column i.e. structured data is simple but extraction from unstructured form of data is complex. A lot of data is available over the net for the learners. Education is one of the domain behind the success of all other domains (e.g. Medical sciences, Business).So by making education system effective we can achieve the success of all other domains.

Big Data in education also known as Education data mining and Learning analytics. There are a lot of case studies for the betterment of education system. As an instance could be the data analysis of 3,747 students in Massachusetts in 3 school districts. The outcome of these analyses should be student's readiness for the college or career[2].

II. WHAT ACTUALY BIG DATA IS?

A. Definition

Big Data is defined by the IBM as follows: "Data coming from everywhere; posts on social media sites, transactional data during e-commerce, videos, texts etc" [3].

Another definition of Big Data is "exponential growth of data which is highly dynamic in nature i.e. structured, un-structured and semi-structured".

B. Big Data Dimensions

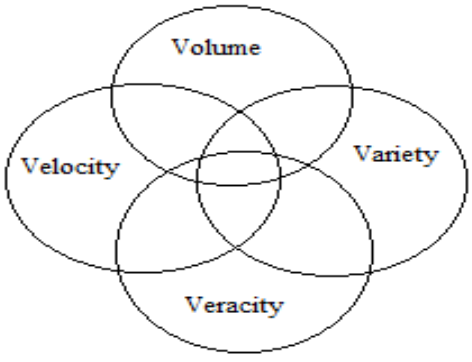

Fig. 1 Dimenssions of big data

Big Data is based on four basic dimensions:

1) *Volume:* The amount of data which is growing vastly every year. Social media sites are greatest contributor on increasing the volume of data. Now we have data in petabytes but in near future it will become zettabyte.
2) *Velocity:* The speed of increasing data or we can say the rate of data flow from various resources.
3) *Variety:* Data which is coming from various resources having different format such as structured, un-structured and semi- structured(log files, server logs, videos, emails etc) .
4) *Veracity:* Data that we have from various resources must be accurate. We cannot make decisions on the basis of inaccurate data.

All of the dimensions shown in the fig.1 are interrelated with each other i.e. volume depends upon the variety, velocity as well as veracity of the data. In other way volume is increased by variety of structured, unstructured data and also the speed of flowing data i.e. in two days near about 8 petabyte of data is generated through the social media sites. So all these data generated by the user contributes to the volume of Big Data and data we get from different-different sources, so accuracy of the data is can't be determined easily but this type of raw data is also contributes for increasing the size of data.

Several companies working with big data also rely on other dimensions:

5) *Complexity:* Tremendous amount of data we have in various formats such as structured, un-structured and semi-structured. So extracting information from social media sites (facebook, Flickr, Twitter, Blogs) is quite complex.

Big Data analysis can be obtained through the stream processing of data and batch processing of data on the basis of processing time requirements:

- *Stream processing of data:* Tremendous amount of data is generated during the stream processing in the form of stream so rapidly processing of data is required. There is an assumption that we do mostly that fresh data is taken during stream processing. A stream generates huge amount of data which can't be stored in any memory place. That's why we need faster processing of data. In this approximation of result values we achieve. For e.g. online application uses streaming processing where processing of applications takes time in milliseconds and seconds.
- *Batch processing of data*: In batch processing firstly data is stored in memory which can be distributed and then analyzed. Data is divided in to small size chunks and then processing of data is obtained through parallel computing i.e. where simultaneous execution of the data happens and intermediate results are obtained. After that all the intermediate results are combined together i.e. integrate to obtain the final results. For instance MapReduce is used for doing batch processing. Batch processing is widely used for getting the result in real time.

III. DATA ABUNDANCE

Data comes from everywhere available over the internet, on the mobile devices and computers. Each click, every post on the Facebook, every tweet and online reading of a document can leave a digital trail. For students there is MOOCS(Massive Open Online Courses) are available and also LMSs is there which contains all the student record such as their assignments, attendance records, test records etc which will provide their complete assessment. By finding each and every learners performance organization authority can do intervene by providing help to those learners who are at the risk of dropping out. It helps the learner for increasing their performance. Day by day data is explored over the internet. Research related data is also there which helps the learner to get information from that source. It helps the learner to get ideas from the existing research and development technologies and to build or generate new ideas which explore the particular domain.

Social media sites like Facebook, Google, Flickr, Twitter is the one of the most popular source of exploring data. Within two days amount of data is increased in petabytes through the posting blog, uploading videos and pictures etc. In an organization such as in universities and colleges they have their historical records of their students related to their marks, their personal information is also available somewhere. So this detail also contributes to exploring the data. All the data sources which have shown in fig. 2 are the main reason of data abundance at the university level.

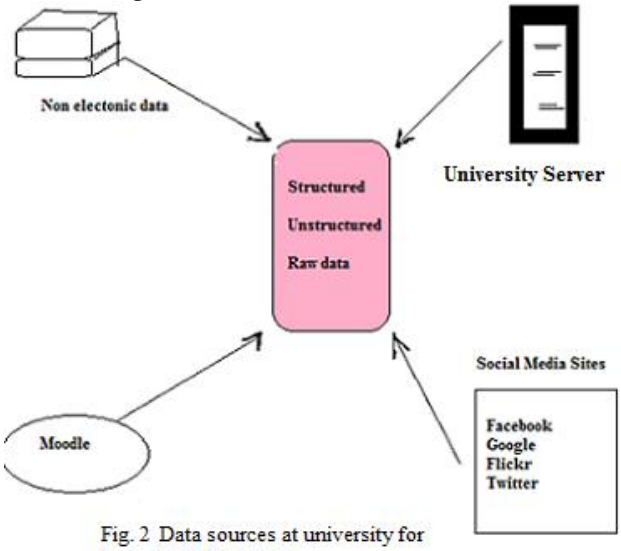

Fig. 2 Data sources at university for learners and learning

The fig. 2 shows all the data sources available for the learners as well as for the learning. So it contributes to exploring data in education using Big Data. We are discussing in this paper about the one particular domain of Big Data which is education. Data abundance is due to the autonomous sources of data available for the learners and learning are as:

1) Non electronic form of data contains learner records which can be historical which are maintained in the form of hard copy of an individual files.

2) University server contains a lot of log files, server logs and cookies etc which provide detail of net surfing of individual learners. So all those details or records are helpful for finding student or we can say learners interest are growing in which direction i.e. either in the studies or they are wasting their precious time by doing chat, watching movies etc whole day. As a consequence of this learners performance surely affected or degraded.
3) Moodle is used by the most of university and colleges which contain all the records of the learners their attendance, test marks and assignment records by which an organization easily find out the individual performance of learner and also can do intervene. So that student will know about their performance and also they can compare their performances with other batch mates. For e.g. In Shoolini University we are using euniv moodle for analyzing learner's performances.
4) Social media sites are there such as Facebook, Google, Flickr, Twitter etc are the sites which contributes a lot for increasing data.

IV. CONCEPT DEFINITION

Big Data in education is also known as EDM (Education Data Mining) and LA (Learning Analytics).Huge amount of data is available over the internet for the learners/researchers/students related to their courses, their researches but extracting relevant information is big challenge. Data sources available for the learners such as classroom learning, social media sites i.e. (Google, videos, emails, facebook, twitter) and PSLC data shop which is India's leading repository of software tools. In education data mining the need of extracting the relevant information from huge amount of data is a big challenge. So we can use the concept of semantic web mining here. Semantic web mining is used to present the rich representation of information or knowledge. So in education data mining extracting the information by semantically understand the meaning of content is required for real time retrieval of the information from huge amount of data.

Learning Analytics is required in education to make it effective. Analysis of learners and learning is required. For learners perspective analysis can be done by their assessment on the basis of their on-task and off-task behavior. E.g. LMSs is the best way for student assessment. For the Learning perspective analysis can be done by make changes to the existing teaching methods, assessment criteria and by altering the contents. Intelligent curriculum is required to introduce in the organization for the learners i.e. introducing the subject related to the current new technologies so that the students show their keen interest to know about the particular technology. By knowing the latest development in the technology they become aware about what actually happen in that field? It will provide them new ideas for development which is helpful for doing their research as well as in developing some advance technologies. Basically our motive for doing analytics in education using Big Data is to find or we can say explore new insights both for the learners and learning.

Big Data contains the structured as well as unstructured data. Structured data can easily used for decision making but unstructured data is need to cleansing, preprocessing and then transformation of data needed. After transforming data by using OLAP techniques it can be used for decision making. Resulted data which we get by applying all these techniques are in the form of histogram, pie charts and bar charts. Analytics in education is required for making decisions on the basis of learner's individual performances.

Big Data has following phases:

1) Data generation
2) Data possession
3) Data storage
4) Data analysis

In first phase which is data generating phase data is generated through the scientific research, business, and social media sites. Data is generated via autonomous sources. Data generation is basically domain-specific i.e. related to particular domain either in medical sciences, business or education.

Data possession phase in which data is acquired through the distributed and longitudinal sources. Acquired information from distributed sources may be either structured, unstructured or in the form of raw data. So pre-processing, transformation of data is required using any of techniques such as OLAP, ROLAP, MOLAP.

After pre-processing and transforming data, now data is stored in the hardware infrastructure or in the data management system such as RDBMS, OODBMS etc.

When data storage is done then after that analysis of data is required through the methods used for analyzing data which is prediction, correlation mining, and pattern discovery.

A. Methods

Methods are used for education data mining and learning analytics are as:

1) Prediction
2) Structure discovery
3) Relationship mining
4) Discovery with models

In the first method prediction is basically used to predict the future or sometimes used to make inference about present. For example if a student has passed his/her higher school education then we can use prediction method to find out what will be his/her score in the college entrance exam. Prediction can be categorized in to classification, regression.

- In classification we can predict through categorical i.e. in the form of either correct or wrong. For e.g. school records, test data, survey data, Form filling etc.
- Regression is used for binary classification i.e. in the form of 0 and 1.

The second method is about structure discovery i.e. to find out the pattern in which data exist. I.e. in the form of histogram, bar charts and pie charts. Structure discovery is classified in to clustering, factor analysis etc.

- Clustering is used for building clusters at which similar information is put in a cluster while the dissimilar information is put in the other clusters. That means number of clusters are there and each cluster contain the information which is relevant to the particular one.
- Factor analysis is used for the analyzing the factors in the available dataset, variables or we can say how variables and datasets are grouped together on the basis of which common factor?

In third method which is about relationship mining to find out the relationship between the variables in a data set. For e.g. Association rule mining and correlation mining.

- Association rule mining is used for finding the relation between the variables. For e.g. Market basket analysis is a technique which is used for predicting what customers can most probably buy if he/she bought milk and bread. By this prediction market can analyze the demand of their customers.
- Correlation mining is used for finding the relationship between different variables in the dataset. Suppose if we have 50 variables in a dataset then to find how they are correlated with each other and on the basis of which features.

Discovery with model is very popular method for analysis. In this method pre-existing model (which is developed by the prediction methods and clustering) use that model and applied to the data that the person want to evaluate.

B. Technology

In Big Data processing of data by the existing methods is too complex and also time consuming. So there is a need of the new technologies are required. Parallel computing is very popular for processing of huge amount of data at real time. In parallel computing simultaneous processing of multiple tasks happen at a time.There are some of the technologies which are used in the big data:

1) Hadoop: It is an apache open source software project which is based on distributed processing of large data sets across clusters of computers. Hadoop has Characteristics such as flexibility, scalability, cost efficiency, fault tolerance.
2) Hadoop Distributed File System (HDFS): It provides storage of large distributed files across distributed servers.
3) MapReduce: Originally owner of MapReduce is Google but now it is incorporated by the apache. Two functions are performed by the MapReduce. First is Map () and another is Reduce ().
   Map () is used for filtering and sorting of data. For e.g. list of the students by their last name.
   Reduce () is used for summarizing the data.
4) Apache hive: It is based on data warehouse infrastructure which is built in top of Hadoop. It is used for summarizing data, query and making analysis.
5) NoSQL: It provides less constrained database system then SQL. It provides mechanism for storing and retrieving data which is less constrained [1][5].

C. Importance of EDM and LA for higher education

There is a lot of importance of education data mining (EDM) and learning analytics (LA) for Big Data in education:

1) Evidential data which is obtain through the methods such as prediction, clustering and classification helps to make decisions at the organizational level.
2) Learning Management Systems (LMSs) such as MOODLE, Desire2learn which contains all the records of the learners helps in generating their assessments. By finding learners individual performance academic authority can intervene and assist them to improve their performances [4].
3) Organization system can introduce innovative skills in teaching to make academic organization adroit.
4) Through the prediction method we can give suggestion to the higher authority of the academic organization. For e.g. (Why students are not taking admission in their university and colleges and in which way their university and colleges lack behind compared to others university?) So by making these analyses at the academic level they can do changes to their infrastructure and also in teaching techniques. Analysis helps in improving or we can say increasing efficiency and productivity of the organization.

V. CONCLUSION

Analytics and Big Data have immense influence to predict the future of education. Nowadays there is a growing need of analysis techniques and technology in the entire domain such as in government, business etc. In education domain big data and analytics will help to improve learners and learning skills and to achieve immense productivity and efficiency of the organization. It helps in making decisions. Big Data in education and analytics helps us to achieve successful future of education for the learners.

REFERENCES

[1] D. Jayathilake et al: "A study into the capabilities of NOSQL databases in handling a highly heterogeneous tree," in Information and Autonomation for sustainability (ICIAFS), pp. 106-111, Beijing 2012.

[2] M.O.Z. San Pedro, R.S.J.D.Baker, A.J.Bowers, N.T.Heffernan(2013): "Case study on Predicting college enrollment from student interaction with an Intelligent Tutoring System in Middle School." Proceeding of the 6th International Conference on Education Data Mining, 177-184.

[3] Official Webpage Of IBM Company: http://www-01.ibm.com/software/data/bigdata/

[4] P. Campbell.John, B.DeBlois.Peter, and G.Oblinger. Diana,"Academic Analytics: A New Tool for a New Era," EDUCAUSE Review, vol.42, no.4(July/August2007),pp.4057,http://www.educause.edu/library/erm0742.

[5] R.D.Schneider, Hadoop for Dummies, John Wiley, Mississauga (2012).

