IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

STUDY & ANALYSIS OF EDUCATION USING BIG DATA ANALYTICS AND TECHNIQUES

Aayushi Agrawal*, Devina Khare, Sushrut Singh Sisodiya
Lecturer, Ashokrao Mane Polytechnic Vathar tarf Vadgaon

ABSTRACT

As the data goes on increases, the demand and requirement is also increases. If the past is being considered than we will see first our aim is to collect the data using computers, taps and disk and then we access the data using RDBMS, SQL and ODBC and then we have used some other technologies like OLAP, Data warehouse and data marts for dynamic data delivery. We still are unable to collect and manage this huge amount of data at appropriate level. Still other problems are that uncollected data cannot be analyzed, no quality data is maintained, complexity in terms of storage is more, time required to get information is more. Now, the requirements are more and we required such technologies through which we can achieve the goals in the efficient manner. In this paper, we will discuss the "Big-Data" and the technologies used in it so that we can able to improve the education.

KEYWORDS: Big Data, Education, storage, complexity, analytics.

INTRODUCTION

Data is growing very rapidly day by day and it becomes dynamic in nature so that it becomes more difficult to manage and store the data. If we consider the education system, then the data can be in the form of text, audio and video. The large amount of data will be managed by the colleges and universities, which is hard to store and manage. Due to the reason is that although the colleges have the required amount of information but it is unable to use it in the appropriate manner so that the education systems can work efficiently. Big data in education system is known as Education data mining and learning analytics. As we will see every second human being is accessing the internet via mobile or computers. The huge amount of data will be maintained by the Facebook, Amazon, Google+, Emails. To maintain these data and to access the data required lot of time. The data available here in these websites are in huge amount as known as big data. Big data can be of any form such as structured, unstructured and semi structured. Structured data is the data which is stored in the database in the form of rows and columns. The extraction of information from such data is easy. Structured data is simple in nature but the extraction of information from unstructured data is more complex in nature, because such data can't be stored in database. Over the internet, lot of data is available for the learners. Education system is one of the domains that includes engineering, medical etc. If we make education domain effective so that other domains can be easily succeed. According to IBM, the Big Data is data coming from everywhere i.e. from social media like facebook, transactional data during e-commerce, banking data, government data, transportation data etc. from such sited data can be very in the form of text and numbers, audio, video and images.

Big Data is based on the following basic dimensions:

1. **Volume:** The amount of data which is growing vastly every year. Social media sites are having the large amount of data that are increasing as days passes. In the future, data storage may cross the petabytes also.
2. **Velocity:** Velocity will define the streaming of data. It will tell us at what rate the data is flowing.
3. **Variety:** Data can be in any form. It may be number/text, image, audio and video. All these data is categorized under structured data, unstructured data and semi structured data.

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:** During stream processing, we can generate data in huge amount. As we know that if we consider the education system then we have huge amount of data is present so rapid processing is required to fetch the data frequently. 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. For e.g. internet based application uses streaming processing where processing of applications take time in milliseconds and seconds.
- **Batch processing of data:** In batch processing first we stored the required data in primary memory which then further can be distributed and analyzed. Data is divided into small size of blocks and then we process the

data using parallel processing, 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.

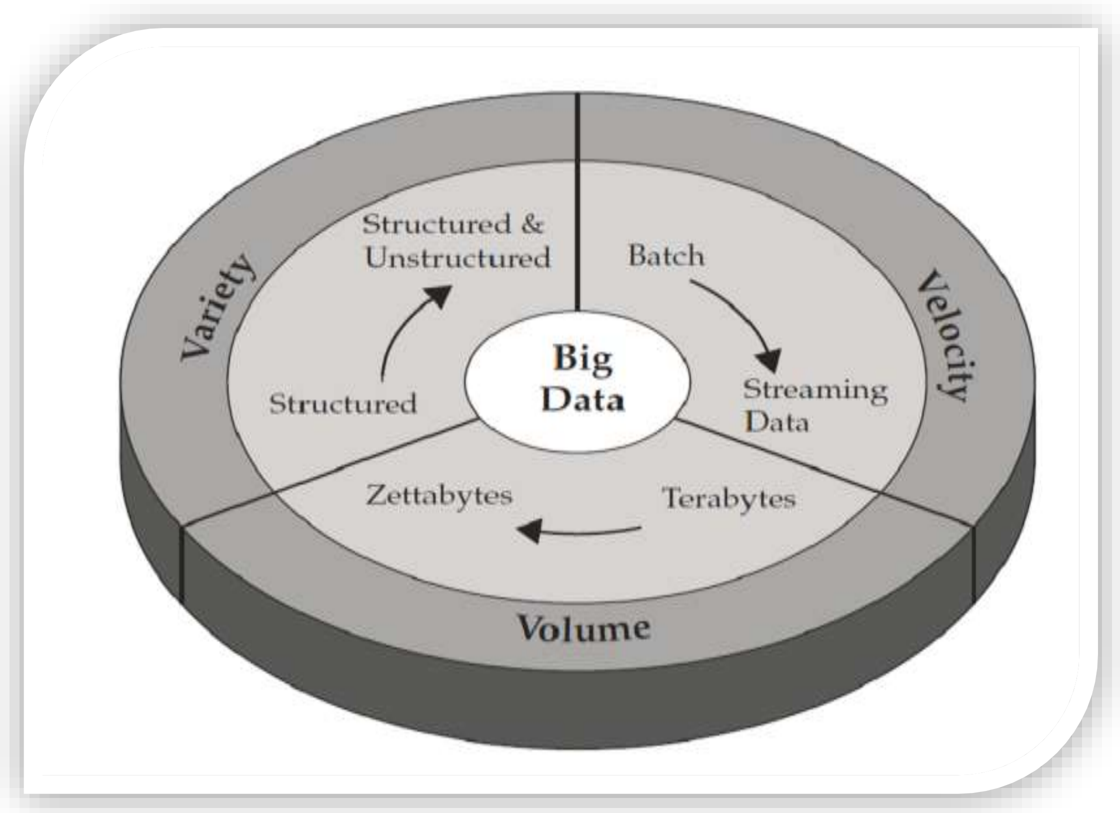

Figure 1: Dimensions of Big Data (Volume, Velocity and Variety)

CHALLENGES IN EDUCATION SYSTEM

The use of big data and education learning analytics are exiting to learn but in future it has many challenges. Some of the challenges over the technical side is to integrate the data comes from different sources, on different platform and different vendors. We required maintaining the data interoperatability, so that data can be marked in consistent ways across systems, as well as software solutions to pull those data together. The universities and their staffs are required to develop the new level of knowledge in the field of data management and analysis. Institutions and education systems will have to find the way through these anxieties to develop procedures that allow them to access and process personal data for educational purposes, while ensuring that they collect data in sensitive ways, protect those data and ensure that they are used only by authorized users for agreed purposes.

Apart from it, there are many challenges to maintain the big data:

- **Storage:** Storage is one of the major issues while working on data. As we know that the amount of data generated by the internet is in huge size as compared with the education data. But in future the data will be in larger size and to store such an large data older tools like RDMS are unable to store and process on such big data. In today's day the hard disk is in the range of terabytes but in future it can be extended. To overcome this challenge, databases that don't use traditional: SQL based queries are used. Compression technology is used to compress the data at rest and in memory.
- **Analysis:** Data analysis is one of the crucial problems, because we have data available in variety. Some of the data in education system is maintained in the form of hard copy. The analysis of data required more time and resources to process on such huge data. To overcome this problem, leveled out architectures are used to

process the data in a distributed manner. Data is clustered into the number of process and distributed over the network and then process by individual to generate the final result.

- **Reporting:** Traditional reports involve display of statistical data in the form of numbers. When large amount of data are involved, traditional reports become difficult to interpret by human beings. In those cases the reports need to be represented in a form that can be easily recognized by looking into them. The Big Data technologies overcome these challenges using various techniques.

TECHNIQUES AND TOOLS USED IN BIG DATA

- **Techniques:** The challenges faced in processing Big Data technologies are overcome by using various techniques. The most popular techniques used in educational data mining are listed below.
  - **Regression –** Regression is used in predicting values of a dependant variable by estimating the relationship among variables using statistical analysis.
  - **Nearest Neighbor –** In this technique the values are predicted based on the predicted values of the records that are nearest to the record the needs to be predicted.
  - **Clustering –** Clustering involves grouping of records that are similar by identifying the distance between them in an n-dimensional space where n is the number of variables.
  - **Classification –** Classification is the identification of the category/class to which a value belongs to, on the basis of previously categorized values.
- **Tools:** Several Open source tools exist which help in taming Big Data [9] Some of the top tools are listed below.
  - MongoDB is a cross platform document oriented database management system. It uses JSON like documents instead of table based architecture.
  - Hadoop is a framework that allows distributed processing of big datasets across clusters of networked computers using simple programming models.
  - MapReduce is a programming model and framework used by Hadoop. It enables processing huge amount of data in parallel on large clusters of compute nodes.
  - Orange is a python based tool for processing and mining big data. It has an easy to use interface with drag & drop functionalities with variety of add-ons.
  - Weka is a java based tool for processing large amount of data. It has a vast selection of algorithms that can be used in mining data.

Education data mining and learning analytics are the basic approaches used in education system. As we know that to maintain the big data over the internet is the crucial task. To extract the relevant information from the existing data is becoming a challenge for researchers. We have lots of data available in education system that includes the personal information of the students, their academic records, personal data of the faculties their academic records, sports activity records, other curricular activities records, workshop, expert lecture, guest lectures will be conducted by the department wise. To maintain these data and to retrieve the relevant information from these data is a challenge. We can use semantic web mining in education system because it provides the rich representation of information and knowledge. 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. We required learning analytics to make the system more efficient and effective. Analysis of learners and learning is required. For faculties as well as students analysis is much more important factor. Because as we know that the technologies is growing rapidly and to make efficient teaching faculties must have the knowledge of latest trends and technologies so that they can guide to the students in the better way. Universities/ colleges are required to develop such a system through which in the entire session they can analyze the effective teaching of the faculties. The workshop and expert lectures for the faculties needs to be arranged by the department on the regular interval on teaching methodology. For the learners it is very important that colleges will analyze the students on the basis of their academic and extra activities performances. Every students need to be analyzed so that we can effectively improve the quality of the students. 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) Generation of Data

2) Possession of Data
3) Data Repository
4) Analysis of Data

In the first phase of big data we will generate the data from different sources. The data may also be generated via autonomous sources. Suppose if we are applying the big data on the engineering education system then we have to generate all the data relevant to this. This simply means that generation of data is domain specific. 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.

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).
6. Zailani Abdullah, Tutut Herawan, Noraziah Ahmad and Mustafa Mat Deris, “Mining significant association rules from educational data using critical relative support approach”, Procedia-Social and Behavioral Sciences, Vol. 28, pp. 97-101, 2011.
7. Brijesh Kumar Baradwaj and Saurabh Pal, “Mining educational data to analyze students' performance”, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, pp. 63-69, 2011.
8. Umesh Kumar Pandey and Saurabh Pal, “A Data mining view on class room teaching language”, International Journal of Computer Science Issues, Vol. 8, No. 4, pp. 277-282, 2011.
9. Toon Calders and Mykola Pechenizkiy, “Introduction to the special section on educational data mining”, ACM SIGKDD Explorations Newsletter, Vol. 13, No. 2, pp. 3-6, 2012.
10. R.B Bhise, S.S Thorat and A.K Supekar, “Importance of data mining in higher education system”, IOSR Journal Of Humanities And Social Science, Vol. 6, No. 6, pp.18-21, 2013.