Implementation of NoSQL Database in Data Mining Models for Educational Big Data

Abdus Sattar, Sumon Deb, S. R. Sakib Ahmod, Bulbul Ahmed, Abdullah Al Munaim, Raihana Zannat, Ohidujjaman
[1,2,3,5,7]*Department of Computer Science and Engineering,*
[4]*Senior Software Engineer, DataPath Limited,*
[6]*Department of Software Engineering,*
[1,2,3,5,6,7]*Daffodil International University, Dhaka-1207*

Abstract— Data mining models of education system is limited to relatively small scope of researches and analysis. In the information era, it is the vital issue to retrieve and processing available data in real time. The education system based on relational database has lack capabilities of processing real time data for its large volume and unstructured format. On the other side, NoSQL repositories have strong architecture, it can process big data efficiently in real time by using appropriate model. NoSQL is a non-relational database management system which stands for "Not Only SQL" or "Not SQL", it can store a wide variety of data including document, key-value, columnar and graph formats. Relational database management system uses SQL syntax to store, retrieve and mining necessary data. Instead of NoSQL database system circumscribe a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data. This study proposed an improved model for data mining by NoSQL repositories and transformation of data from unstructured form to structured form using NoSQL technologies. Double view of big data is collected in the database uses data mining technology to retrieve required data. Appropriate process of query a database allows to effective retrieval of imported data for analytical systems. Data is available for analysis instantly after submitted by users. The proposed model is suitable for any institutes or organizations which have a large amount of data.

Index Terms— Big Data, Data Mining, Databases, Educational Data, MongoDB, NoSQL, Polymorphic Data.

1 Introduction

BIG Data is now very common word and it has become popular in scientific community for its grate impact. There is no universal accepted definition for big data. It can be describe by 5V model, which are Velocity, Varity, Volume, Value and Veracity [1]. Velocity is the speed of shifting digital content from data set to data stream. Varity is different kind of natures of data such as texture data, multimedia data etc. Volume describes the size of data, which can be terabyte to petabyte. Value determines by the cost of data and noise renovation of data stream is called veracity. Generally, Relational Database Management System is very popular and common in this time, but it has become unfit for work day by day. The main drawback of relational warehouse is limitation of its storage and data mining capability. On the other side, there is no particular and significant structure of big data. For this, it is a big trouble to handle big data using RDBMS. In this case, NoSQL has various databases such as graph, object, multi-modal etc. Moreover, the main advantage of NoSQL is various type of data storage capability such structured, semi structured and unstructured data. The repository can store real data in its storage like as file, image, attachment etc.In educational world, institutions gather a large amount of structured and non-structured data about students, teachers and other employees as well as course materials. Which can be than used to improve educational systems, as related personnel such as teachers can depends not only their own knowledge but also acquire knowledge from others experiences. Nowadays, it is the big problem to perform data miming in such large storage. There are various kinds of data and the major problem is summarization of these data. In this paper, we proposed a data-mining model using NoSQL, which can perform to generate analytical reports efficiently. In practical world, we need structured data in everywhere. The model has given an accumulated structured data from NoSQL repositories. The model is designed by using MongoDB (NoSQL Database) for its powerful data engine. Thus, the proposed data-mining model can provide better output in short time. It can be fruitful for educational personnel such students, teachers and so on.

2 Literature Review

In ICT era, educational data field is essential to work on analysis. Data mining process that produces knowledge from dataset, as system needs and disclose hidden patterns. To improve education field, many researcher tried to establish better system by the virtue of data mining technology. There are huge research work on data mining technique with structured and unstructured data. Many researchers work with data mining for relational database. Till now, it is not popular to practice with NoSQL. Marcin Mazurek proposed a model titled to combine both RDBMS and NoSQL technologies in an analytical system [2]. A B M Moniruzzaman and Dr. Syed Akhter Hossain has given a discussion to provide classification, evaluation and characteristics of NoSQL databases for Big Data Analysis [3]. Aryan Bansel, Horacio Gonz´alez–V´elez, Adriana E. Chis tried to provide data standardization and classification stages with efficient mapping and they discuss different between cloud-based and NoSQL data stores [4]. Benymol Jose and Sajimon Abraham tried to solve the problem of inserting unstructured data. They discussed flexibility of NoSQL to various sorts of data, structured data, semi structured data or unstructured data [5]. F.N. Chernilin, M.M. Rovnyagin, A.V. Guminskaia, O.V. Myltsyn, V.M. Kinash, A.V. Kuzmin and A.P. Orlov have given disscussion to accelerate NoSQL-system and amplify their functionality [6]. The mentioned papers are helpful for my research work. In these papers, they worked with Relational Database Management System (RDBMS) and NoSQL. Moreover, they tried to provide acknowledgement and depth information of structured and unstructured database to analysis Big Data. In this paper, we proposed a model for data mining with NoSQL repositories efficiently. Applying the model, a better output could be achieved for proper analysis of educational big data. It will be more useful for educational data analysis.

3 Research Methodology

3.1 Research Instrumentation

Data mining techniques and concepts are uses in various perspectives. Educational System, Customer Relationship Management, Healthcare, Banking, Fraud Detection, Market Analysis, Real Estate, Criminal Investigation, Manufacturing, Engineering, Education are the most common Data Mining applications. Educational data mining is now one of the most demanding issue. There are so many techniques are used for educational data mining such as Decision Tree, Naïve Bayse, Regression, Nural Networks, K-Nearest Neighbors and so on. These techniques provide us various knowledge to discover such as association rule, classification, regression and clustering. These techniques can be useful us to retrieve data using NoSQL. As research input, system need collected data from various university and some other educational institutions. To develop this system, the experimental model is designed by Python programming language. The collected data is stored in JSON format. Dataset is produced by using MongoDB. Then the model produce expected output by implementing the data mining technique on the generated dataset using "Matplotlib" library of Python. Here are also used "Numpy", "Seabon" and "Bokeh" library of python in implementing time. The models are assesses by calculating accuracy through sklern preprocessing.

3.2 Data Collection Procedure

Firstly, it is mandatory to identify the Big Data and determine the origins from which come. Suppose that we dealing with a varsity, which is very similar to any other universities. After collecting data of one university, it need try to collect data from other educational institutions. Main differences start to emerge at the level of elementary and secondary schools. Then this data will be separated on various groups based on its origins, as shown in figure 1. The sources of collected data can be different wayssuch as shown in bellow:

- Documents in offline form
- Server's Logs
- Public portal
- University information
- Data from social network system

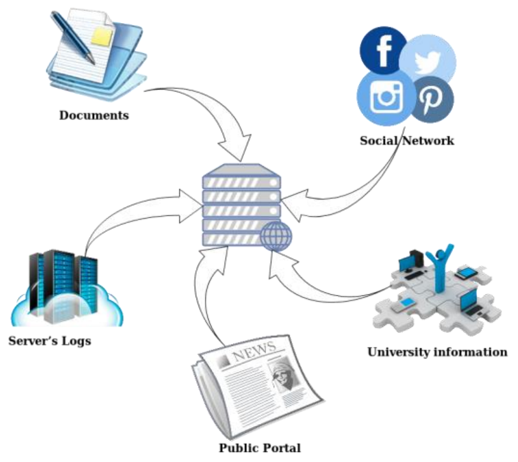


Fig. 1. Big data collection from various sources

3.3 Usage Scenario and Design Principle

The system targets various groups. This system will be useful to researchers, teachers, students and system of education. This study discussed various use-case scenarios for the system:

1. Education Sector
Actors: Teachers, Students, Stuff.
Input: Students Information, Teacher's Information, Stu dents attendance results, others education information.
Output: Big data analytical report, structure dataset, aggre gated statistics for educational system, visualized pattern.
Tools: Desktop application.

2. Data distribution from large amount of both structured data and unstructured data.
Actors: Database administrator.
Input: Stored data from various repositories.
Output: Extracted dataset of structured and unstructured data, algorithms.
Tools: NoSQL repositories such as MongoDB, RDBMS.

3. Data Analysis.
Actors: Data scientist.
Input: Dataset of structured and unstructured data.
Output: Data analytical reports, visualization patterns, algo rithms.
Tools: Statistical packages of programmatic language like Python or R.

4. Predictive Modeling.
Actors: Users, Researchers.
Input: Relational repositories, Anomaly detected data.
Output: Graphical Report, Prediction.
Tools: Data mining tools, visualization package of Python.

To achieve the goal of our proposed system, the above usage scenario has been determined and it will be useful for any institutions.

3.4 System Architecture

There are two types of dataset in our proposed system, which are structured and unstructured dataset. The reason is large volume and variety of dataset. There are many institutions, which have various kinds of data in their own format.

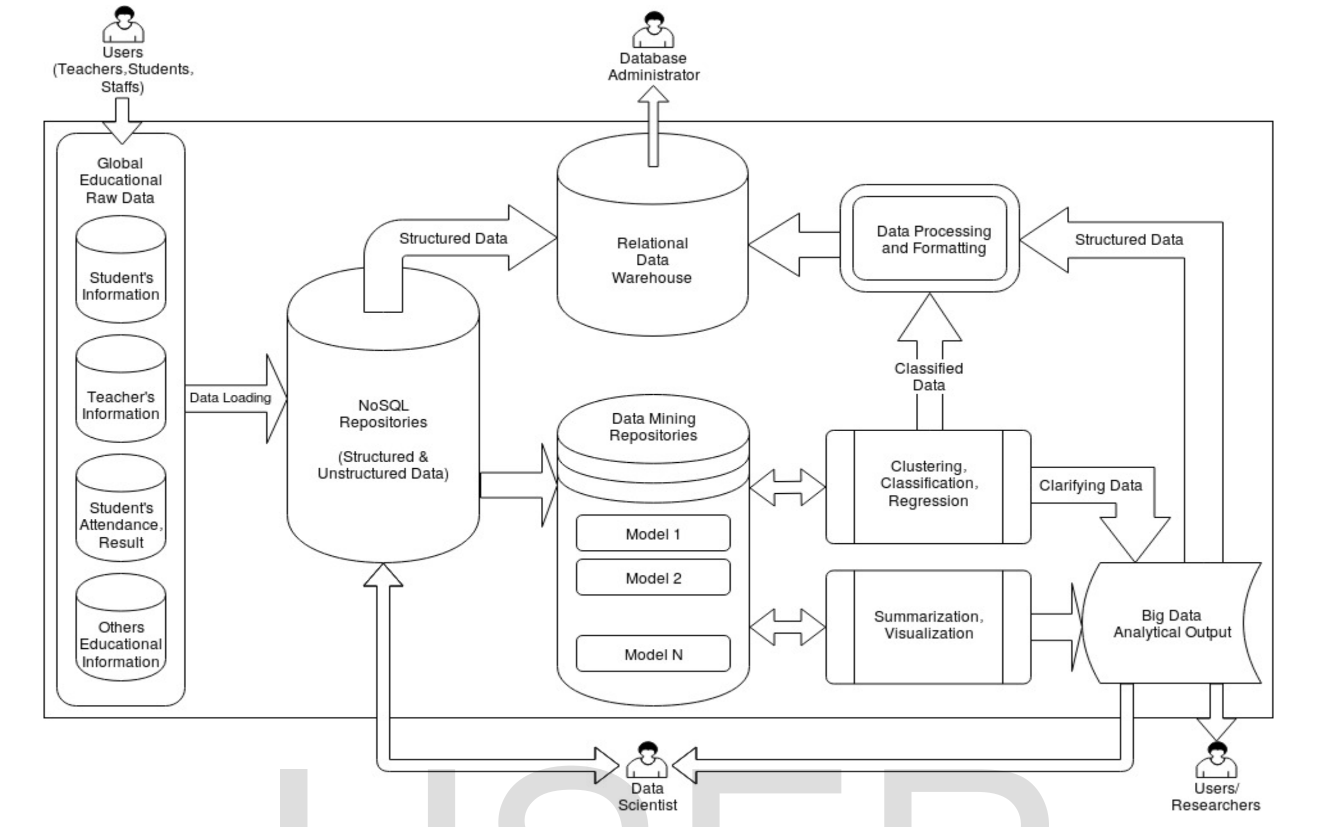


Fig. 2. System Architecture

These have become an unstructured data when all data are combined in a single database system. In this paper, our prime concern is to retrieve those data from large volume of data efficiently. To do this, data are stored in NoSQL repositories, then classify and cluster these dataset using data mining technology, as shown in Figure 2. The architecture focuses our proposed model, by which, users can retrieve and mining data efficiently and they get all analytical reports in the system. The architecture is performing with proper algorithm, which has given in next section. The working procedure has shown below:

Data Sources

Data stored by different actors such as teachers, students, and stuffs. They collect data from various sources such as universities, schools; educational institutions and so on about student's information, teacher's information, student's attendance, results etc. They are mainly collected as data. Collected data can be structured or unstructured form.

NoSQL Repositories

Global educational raw data is loaded in NoSQL repositories like as MongoDB. Which is able to store both kinds of structured data and unstructured data in Database. In NoSQL database, data are stored in JSON format. For this, user can able to store any kinds of data as JSON format. The data can be document, key-value, columnar or graph. The NoSQL database can also provide user with relational data.

Data Mining Repositories

After retrieved data are stored in data mining repositories, which is then process by data mining model. Data are manipulated using data mining processes are include clustering, classification, summarization and visualization.

Relational Data Warehouse

It is only able to gather structured data from repositories. The model passes unstructured data to Hadoop for transform into structured data by perfuming some operations. It is the final output of our system, which provide analytical output to users in various form that depends on the user's requirements; such as graphical, tabular, SQL and so on. The generated output, which is structured form, can also be store in relational data warehouse.

3.5 Proposed Model Algorithm

After find and identify a data source it begins processing data with them. The algorithm of thee proposed system is shown in figure 3. This figure shows that, first collect data from various sources, then load data into NoSQL database. The collected data can be in both form, which are structured or unstructured format. After loading data into NoSQL database, data processed by data mining techniques. There are various processing techniques such as clustering, classification, regression, Summarization etc.

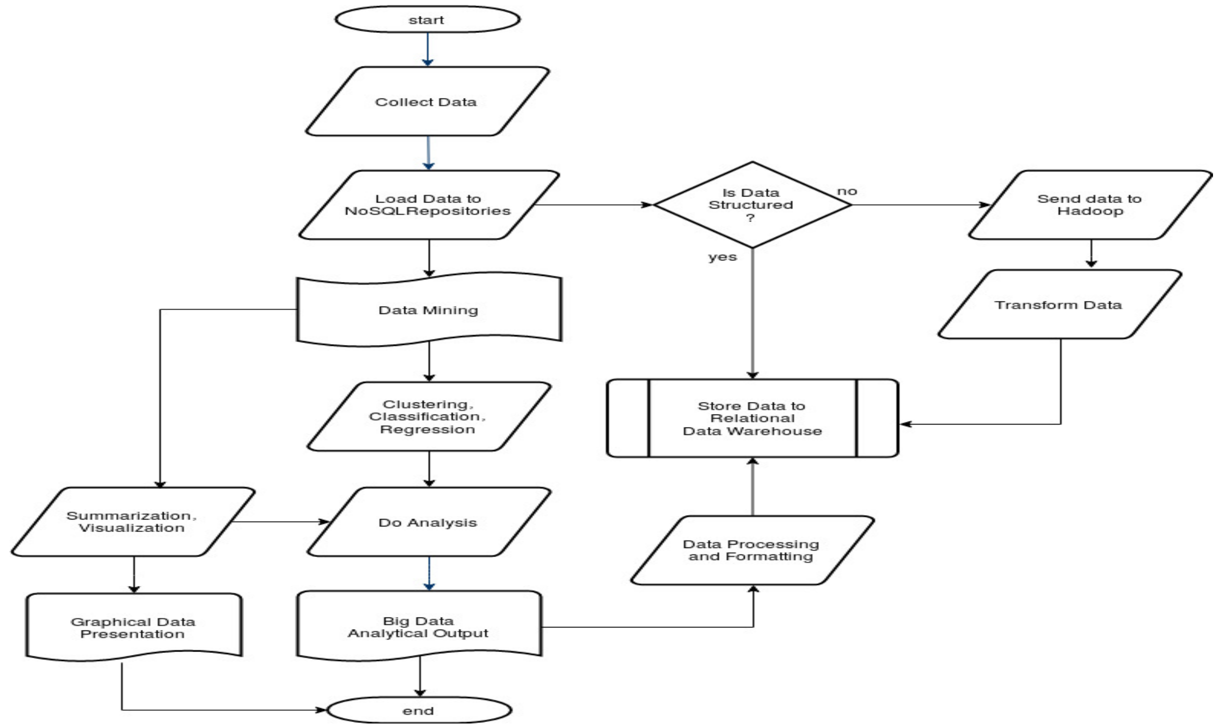

start
Collect Data
Load Data to NoSQLRepositories
Is Data Structured ?
no
yes
Send data to Hadoop
Transform Data
Data Mining
Clustering, Classification, Regression
Store Data to Relational Data Warehouse
Summarization, Visualization
Do Analysis
Data Processing and Formatting
Graphical Data Presentation
Big Data Analytical Output
end

Fig. 3. Flow Chart of Proposed Model

After processing data, processed data can be provide to data scientists or data researchers for analysis. Users can also receive processed data in graphical form. After analysis data, system provides big data analytical output. The output can be also stored in data warehouse by processing and formatting data. The flowchart is described as follows:

Step 1: Collect input data in structured or unstructured form and produce the data into JSON format.

Step 2: Load data in NoSQL repositories. NoSQL database such as MongoDB is preferable.

Step 3: If loaded data has already in structured form then data can be stored directly in relational data warehouse; Otherwise data is sent to Hadoop for transforming into structured form.

Step 4: For analysis data of NoSQL repositories, data would be mining from NoSQL database.

Step 5: Retrieval data can be clustering, classification or re gression for analysis.

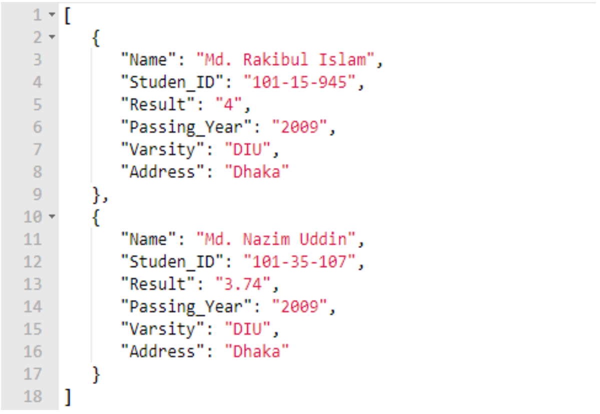
Step 6: Retrieval data can be also summarization for visualized and graphical report.

Step 7: After classification structured data can be also stored in relational data warehouse.

Step 8: Through proper analyzing of data using clustering, classification, regression and summarization, data analyst can get expected outcome from stored data of NoSQL database.

4 Experimental Result And Discussion

Generally there are different kinds of data in real life and all over data produced in unstructured form. To evaluate performance it needs collected data in JSON format. For performing the model, there have collected data from various universities. These data have some similarities. All those data are converted in JSON format as shown in figure 4.

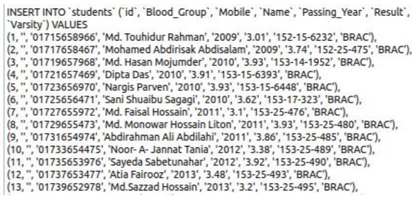
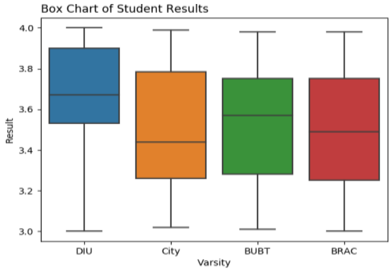
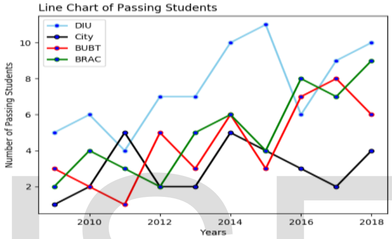
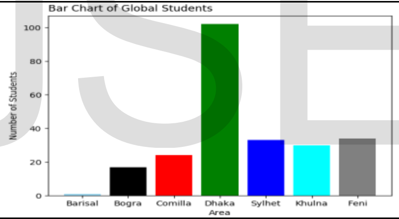
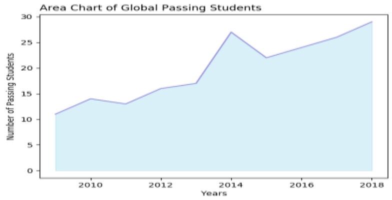
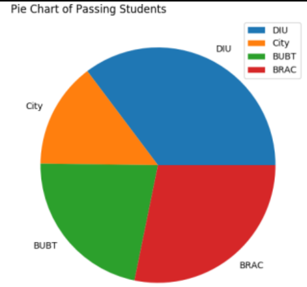


```
[
    {
        "Name": "Md. Rakibul Islam",
        "Studen_ID": "101-15-945",
        "Result": "4",
        "Passing_Year": "2009",
        "Varsity": "DIU",
        "Address": "Dhaka"
    },
    {
        "Name": "Md. Nazim Uddin",
        "Studen_ID": "101-35-107",
        "Result": "3.74",
        "Passing_Year": "2009",
        "Varsity": "DIU",
        "Address": "Dhaka"
    }
]
```

Fig. 4. Collection of Data in JSON Format

To test the system, it is most important to check the outcome. The experimental outcome and explanation shown in table-1.

TABLE 1
EXPERIMENTAL RESULTS AND DISCUSSIONS

Input Data	Expected Output	Experimental Output	Explanation
Unstructured data in JSON format.	Structured data in SQL format.	 INSERT INTO `students` (`id`, `Blood_Group`, `Mobile`, `Name`, `Passing_Year`, `Result`, `Varsity`) VALUES (1, '', '01715658966', 'Md. Touhidur Rahman', '2009', '3.01', '152-15-6232', 'BRAC'), (2, '', '01717658467', 'Mohamed Abdirisak Abdisalam', '2009', '3.74', '152-25-475', 'BRAC'), (3, '', '01719657968', 'Md. Hasan Mojumder', '2010', '3.93', '153-14-1952', 'BRAC'), (4, '', '01721657469', 'Dipta Das', '2010', '3.91', '153-15-6393', 'BRAC'), (5, '', '01723656970', 'Nargis Parven', '2010', '3.93', '153-15-6448', 'BRAC'), (6, '', '01725656471', 'Sani Shuaibu Sagagi', '2010', '3.62', '153-17-323', 'BRAC'), (7, '', '01727655972', 'Md. Faisal Hossain', '2011', '3.1', '153-25-476', 'BRAC'), (8, '', '01729655473', 'Md. Monowar Hossain Liton', '2011', '3.93', '153-25-480', 'BRAC'), (9, '', '01731654974', 'Abdirahman Ali Abdilahi', '2011', '3.86', '153-25-485', 'BRAC'), (10, '', '01733654475', 'Noor- A- Jannat Tania', '2012', '3.38', '153-25-489', 'BRAC'), (11, '', '01735653976', 'Sayeda Sabetunahar', '2012', '3.92', '153-25-490', 'BRAC'), (12, '', '01737653477', 'Atia Fairooz', '2013', '3.48', '153-25-493', 'BRAC'), (13, '', '01739652978', 'Md.Sazzad Hossain', '2013', '3.2', '153-25-495', 'BRAC'),	Relational data warehouse such as MySQL or SQL server can gather only structured data. So it is essential to transform unstructured data into structured form. In this model, Hadoop technology is used for data transformation process.
Student results of different universities.	Result comparison between different universities.		Data analyst need to analyze educational data for comparing or rating different educational institutions based on student results. To produce this analytical report, our system at first classify data from NoSQL database and then summarize the classified data. The produced report can be graphical, tabular or textual form.
Student information with passing year of different universities.	Year based number of passing students for individual university.		To generate the analytical report of the total number of passing students based on year, the system at first classify passing year and then accumulate and cluster the data. Educational personnel can be generated these kind of report by using this model.
Student information with their residential area/location.	Area based number of students, globally comparison of these data.		For educational survey, it is most important to know student information of various cities. The system can generate report based on area by classifying area and university of all over student information.
Global Student information with passing year.	Ratio analysis of yearly grow up of number of passing students.		Sometimes we need to know, how many students are passed from all or specific number of universities in a particular year. To produce the report, system summarize all student information based on passing year.
All passing student information of different universities.	Ratio analysis of total number of passing students and comparison between different universities.		To compare different universities, at first data should be regretted for removing redundant data and then classify and summarize data based on particular university. The generated report can be shown in graphical form for ratio analysis of total number of students of different universities.

5 CONCLUSION

The proposed model architecture obtains from evaluation real life implementation of educational database. This article covered implementation of data repositories based on relational database. The research is creating a large amount of data in educational repositories that is not being utilize perfectly. It is possible to get advantage from the large educational dataset by using data mining processes which serves as stronger tools for characterization, investigation, analysis and prediction. Nowadays, many people are using data mining techniques in various field but it is limited used in educational perspectives.

Moreover, it is more limited for NoSQL repositories. In this proposed model, NoSQL database is used which provides an efficient way to retrieve data from database that contains a large volume of data**.** However users can analyze data from such database is able to retrieve data in a short time to use that information for analysis, characterize, investigation and preparing valuable reports. This model can be helpful for any organizations especially for educational institutions which play with large volume of data. This model will provide advantages to the both students and teachers also and they can support their institutions by quality assurance of education system. It is more effective to improve our education system. Applying data mining techniques with NoSQL repositories on educational data which discloses some significant region on education system where analysis with data mining has acquired benefits such as identification performance of different student, comparison between different teachers, students and educational institutions, identify students satisfaction for a specific area, student evaluation, students course registration planning, analysis the enrolment head count and so on. The article primarily investigates, analyzes and compares student performance with NoSQL database for achieving efficient data mining from various data sources. The study has developed a significant system which can be apply for analysis of student activities. The study of this research makes easy to analysis student performances, global educational environment evaluation and so on. It is more desirable to improve education system.

REFERENCES

[1] P. C. ZIKOPOULOS, C. EATON, D. deROOS, T. DEUTSCH, AND G. LAPIS, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data," Published by McGraw-Hill Companies, 2012.

[2] Mazurek, Marcin., "Applying NoSQL Databases for Operationalizing Clinical Data Mining Models", Springer International Publishing, 2014, pp. 527--536.

[3] Moniruzzaman, A B M & Hossain, Syed., "NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison", Int J Database Theor Appl. 6, 2013

[4] Bansel, Aryan & Gonzalez-Velez, Horacio & Chis, Adriana. "Cloud-Based NoSQL Data Migration", 2016 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP).

[5] B. Jose and S. Abraham, "Exploring the merits of nosql: A study based on mongodb," International Conference on Networks & Advances in Computational Technologies (NetACT), Thiruvanthapuram, 2017, pp. 266-271

[6] M. M. Rovnyagin et al., "Modeling NoSQL Systems in Many-nodes Hybrid Environments," IEEE 11th International Conference on Application of Information and Communication Technologies (AICT), Moscow, Russia, 2017, pp. 1-4.

[7] Arnaud Castelltort, Anne Laurent, "Extracting fuzzy summaries from nosql graph databases" FQAS: Flexible Query Answering Systems, Oct 2015, Cracow, Poland. pp.189-200

[8] Bifet, Albert., "Mining big data in real time", Informatica (Slovenia), 2013, pp. 15-20.