INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 2)

Available online at: www.ijariit.com

Analysis of student academics performance using Hadoop

Diptimayee Baliarsingh
baliarsinghdiptimayee@gmail.com
Bharati Vidyapeeth College of Engineering, Mumbai, Maharashtra

Samiksha Hemant Parab
samikshaprb@gmail.com
Bharati Vidyapeeth College of Engineering, Mumbai, Maharashtra

Vijay N. Patil
vijay.patil.karad@gmail.com
Bharati Vidyapeeth College of Engineering, Mumbai, Maharashtra

ABSTRACT

In recent years the amount of data generated in the educational sector is growing rapidly. In order to gain deeper insights from the available data and extract useful knowledge to support decision making and improve the education service efficient storage management and fast processing analytics is needed.

Academic data of a student helps institute to measure their progress. Students facing severe academic challenges are often recognized too late. Analytics play a critical role in performing a thorough analysis of student and learning data to make an informed decision. Big Data solution enables to analyze a wider variety of data sources and data types which improves the accuracy of predictions. Hadoop platforms provide highly scalable platforms and can store a much greater volume of data at lower cost.

The purpose of the proposed Project is to help in identifying "at risk" students who are not progressing towards graduation early in order to get them back on track. The cause of lack of adequate progression can be identified and addressed. The system proposed will be helpful for educational decision-makers to reduce the failure rate among students. The implementation is done in Hadoop framework. The PAMAE algorithm is implemented for analyzing student's academic data.

Keywords: *Big data analytics, Educational data mining, Hadoop, MapReduce, HDFS, Clustering, Prediction.*

---

1. INTRODUCTION

The amount of data generated in recent years in the educational sector is growing rapidly. It is required for institutes to extract knowledge from huge amount of data collected for better decision making. Data Mining is defined as extracting information or mining knowledge from huge sets of data. Data mining in the field of educational environment is known as Educational Data Mining. Educational Data Mining (EDM) is defined on site http://educationaldatamining.org/ as: "Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in."

A student's performance is measured by how well they perform academically. In the proposed system, academically weak students will be identified and the faculty can decide on necessary steps to be taken to help those students so that by developing proper plans or making a change in their plans they can improve their skills. One of the most important responsibilities of educational institution is to provide quality education. Quality education means that education is produced to students in an efficient manner so that they learn without any problem. For this purpose, quality education includes features like methodology of teaching, continuous evaluation, categorization of the student into similar type, so that students have similar objectives, educational background etc.

In the proposed system Hadoop framework is used. Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. Since student data generated can be in petabytes it is considered as big datasets. In big data, traditional data mining algorithm is translated to MapReduce algorithms to run them on Hadoop clusters. In a distributed computing environment, a Hadoop cluster is a special type of cluster used specifically for storing and analyzing huge amounts of data. Hadoop MapReduce is a software framework which processes vast amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. The MapReduce algorithm contains two important and distinct tasks, namely Map and Reduce. Map phase takes a set of data and converts it into another set of data, where

individual elements are broken down into tuples (key/value pairs). Reduce phase takes the output from Map phase as an input and combines those data tuples into a smaller set of tuples.

2. LITERATURE REVIEW

A number of literature are taken into account.

Midhun Mohan M G & Siju K Augustin (2015) [1] this paper proposes a new approach called Learning Analytics and Predictive analytics to identify academically at-risk students and to predict students learning outcomes in educational institutions. In this paper they take the student data which is clustered using k-means algorithm and prediction of marks is done using multiple linear regression.

Dr. N. Tajunisha, M. Anjali (2015) [2] this paper proposes a system that is useful in identifying weak students who are likely to perform poorly in their studies. This paper identifies that in using MapReduce accuracy of the classification is increased than when it is not. Time complexity of using MapReduce is also reduced. The proposed system which uses MapReduce can be used to process big data too.

Hwanjun Song, Jae-Gil Lee & Wook-Shin Han (2017) [3] this paper proposes an algorithm Parallel k-Medoids Clustering with High Accuracy and Efficiency (PAMAE). PAMAE significantly outperforms other algorithms like GREEDI I n terms of accuracy. In terms of efficiency, PAMAE generates significantly lower clustering error as compared to other algorithms like CLARA-MR.

3. PROPOSED SYSTEM

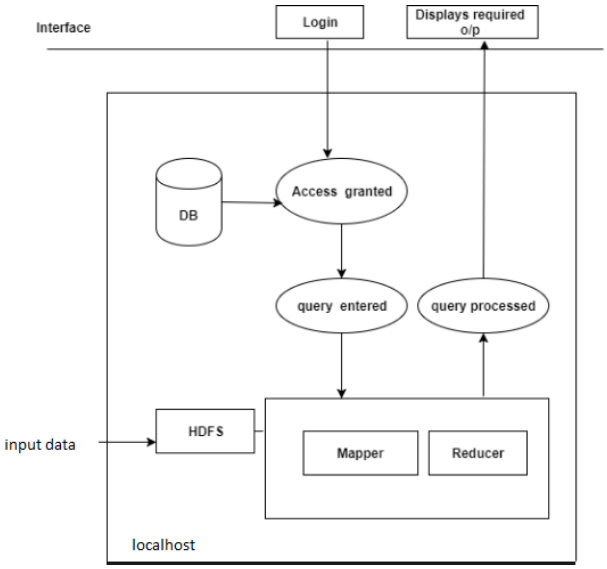

Fig-1: System Architecture

3.1: Hadoop framework

Hadoop is an open-source software framework which is used to store huge amount of data. Hadoop provides background for running applications on clusters for parallel processing on commodity hardware. It provides massive storage for data and enormous processing power. Hadoop is composed of four core components—Hadoop Common, Hadoop Distributed File System (HDFS), MapReduce and YARN. Hadoop framework lowers the risk of huge system failure and unexpected data loss. Hadoop can be easily scaled up from a single server to thousands of hardware machines, each of which offers local computation and storage.

3.2: MapReduce

MapReduce is a software framework which provides ease of writing software applications whose function is to process a huge amount of data on a number of clusters of DataNodes. The data can be structured or unstructured form. The term MapReduce is actually a combination of two phases-Map phase and Reduce phase. The Map phase is designed for sorting and filtering the input data and Reduce phase which is designed for summing up the data. The MapReduce framework operates exclusively on <key, value> pairs. Each map task reads the input as a set of (key, value) pairs and produces a transformed set of (key, value) pairs as the output. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (key, value) pairs to reduce task, which groups them into final results.

3.3: HDFS

HDFS stands for Hadoop Distributed File System, it provides data storage across Hadoop clusters. The data stored in HDFS can be accessed across all the nodes in a Hadoop cluster. It connects together the file systems on many local nodes to create a single file system. The datasets need to be processed and uploaded to Hadoop Distributed File System (HDFS). It can be used by various nodes with Mappers and Reducers in Hadoop clusters. HDFS uses two types of nodes-NameNode and DataNode. NameNode acts as the master node. It is responsible to manage the file system metadata, mapping of blocks to DataNodes. DataNodes act as slave nodes. Datanode store the actual data. The DataNodes takes care of reading and writes operation with the file system.

Parallel k-Medoids Clustering with High Accuracy and Efficiency (PAMAE) algorithm

Phase I: PAMAE starts by simultaneously performing random sampling with replacement. That is, it creates m random samples whose size is n. Then, a k-medoids algorithm A is performed in parallel against each sample. We note that any existing algorithm of finding medoids globally can be a candidate for A. Among them sets of medoids, each of which was obtained from each sample, PAMAE searches for the set of medoids that minimizes the clustering error. This set of medoids, denoted by {θˆ1, . . . , θˆk}, is regarded as the best set of seeds and is fed to Phase II.

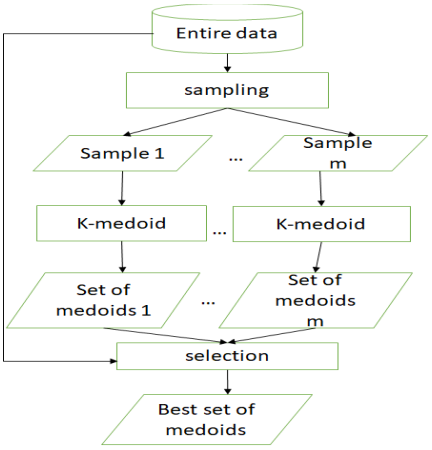

Fig-2.1: Phase I of PAMAE algorithm

Phase II: Using the seeds derived in Phase I, PAMAE simultaneously partitions the entire data set just like the assignment step of the k-means algorithm. That is, it assigns each of non-medoid objects to the closest medoid. Then, the algorithm updates the medoid of each partition (i.e., cluster) by choosing the most central object, similar to the update step of the k-means algorithm. Last, the entire set of data objects is partitioned into clusters if needed.

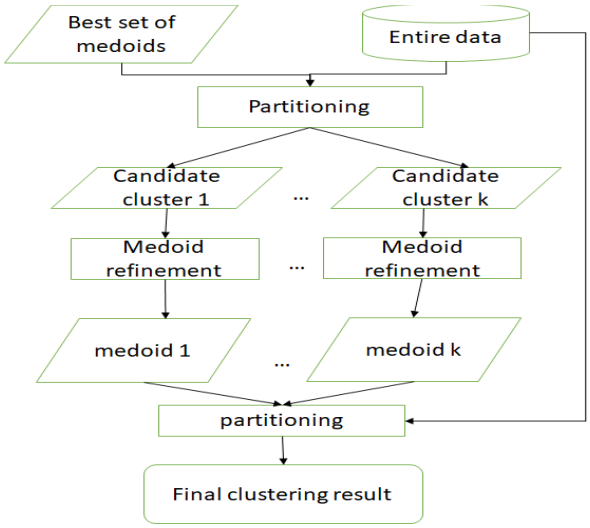

Fig-2.2: Phase II of PAMAE algorithm

4. CONCLUSION

The user can identify academically at-risk students in educational institutes. This study will help to the students and the teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reducing the failure rate and taking appropriate action for the next semester examination. Hadoop MapReduce framework provides parallel distributed processing and reliable data storage for large volumes of data files.

5. ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our guide Prof Vijay Patil, our Project Co-coordinator Prof S. M. Mhatre, HOD Prof. S. M. Patil, and our Principal Dr. M. Z. Shaikh for their guidance. We also thank Bharati Vidyapeeth College of Engineering for providing us an opportunity to embark on this project.

6. REFERENCES

[1] Midhun Mohan M G & Siju K Augustin, "A bigdata approach for classification and prediction of student result using MapReduce," 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS).

[2] Dr. N. Tajunisha, M. Anjali "Predicting student performance using MapReduce", International Journal of Engineering and Computer Science ISSN:2319-7242 Volume 4 Issue 1 January 2015

[3] Hwanjun Song, Jae-Gil Lee & Wook-Shin Han (2017), "PAMAE: Parallel *k*-Medoids Clustering with High Accuracy and Efficiency", KDD 2017 Research Paper.

