# A Framework in Big Data Analytics using MapReduce for Education System

[1] Rakesh S Raj, [2]Chandan C S, [3] Monisha D P , [4] Naveena A M , [5] Rajini M R
[1, 2, 3, 4,5] Department of Information Science & Engineering,
Adichunchanagiri Institute of Technology, Chikmagalur,
Karnataka, India.

***Abstract*:-Big data came into existence as the data generation increased exponentially and the traditional data processing tools became incapable to manage those complex datasets. Big data is diversely used in fields like politics, business, weather forecasting, science and research, e-Commerce, healthcare, and so on. Advent of big data analytics has not excluded its impact on the education system. The recent advancements have given raised to the rapid growth in the amount of data stored in educational databases. In this paper, we propose a framework on big data analytics using MapReduce for education system. The analysis of data generated in education sector can enhance the learning process of the student. It helps in discovering student's academic progress, behavior and predicting future performances. The placement analysis can assist the career development of the student. These analysis will help to the overall development of the student and also to achieve enormous productivity of the organization.**

***Keywords* Big data,HDFS, Hadoop, Analytics.**

## 1. INTRODUCTION

Big data consists of huge or complex sets of data. Data grows day by day since they are gathered by mobile devices, microphones, software logs, cameras, various readers and wireless sensor networks. As of 2015, 2.5 Exabyte of data ($2.5*10^{18}$) is created every day. The relational database management systems (RDBMS) and visualization packages are practically not enough to process all the data that is been produced. It needs software which runs on many numbers of servers or systems. Big data differs depending on the capabilities of the users and their software. Big data analytics, computationally uncovers patterns, trends, and associations relating to the behavioral and interactional aspects.

There are various characteristics defining the big data[5].

- **Volume:** Many factors contribute the increase in data volume or size like unstructured data from past years, data streaming from social media, data collected from machines and sensors.
- **Velocity:** Data is streaming at an amazing speed and must be dealt from time to time. Quick response to the data velocity is a major challenge to many organizations.
- **Variety:** Different data formats are present like, structured data, numeric data in traditional databases, unstructured text, documents, video, email, audio, financial transactions etc.
- **Variability:** Data flows are highly consistent in the peak periods along with its other factors like velocity, variety etc.
- **Complexity:** Data now-a-days come from various sources. It is cumbersome to deal the data from different perspectives.

Education is one of the domains behind the success of all other domains (e.g. Medical science, Business). Hence, an effective education system plays an important role in achieving the success of all other domains. Big data in education sector is known as education data mining and learning analytics [1]. Education data mining determines the facts from the data, predominantly unknown knowledge-driven pattern from educational repository so as to emphasize the strengths and weakness of the student by various methods [3]. For an instance, if a student has passed his/her higher school education then we can use prediction methods to find out what will be his/her score in the college entrance exam.

Hadoop, an open-source software framework can be used to store data and run applications on clusters of commodity hardware. The Hadoop framework is basically for reliable, scalable, distributed computing usingsimple programming models. In this paper, a special programming model like MapReduce is implemented for processing the data in an educational repository.

## 2. PROBLEM STATEMENT

### *2.1. Existing System*

In existing system every organization follows manual procedure in which faculty should enter all the details of the students such as attendance, internal marks and the counseling of the student after each internals which is a time taking process and requires lot of paper work and also there is a chance of misplacing of the records. Loss of even a single register/record leads to difficult situation because all the papers are needed to generate the reports.

### *2.2 Proposed system*

This work aims to provide easy way to analyze large amount of data in an educational institution with an Information Exchanging System Using Hadoop[4]. In which information will be stored in the form of small data blocks across various data nodes in the cluster. And whenever a user requests for information, parallel

processing will take place across all the data nodes, and original information will be obtained. This work also involves placement analysis, server log analysis, attendance analysis, and result analysis.

### *2.3 Advantages of proposed system*

It overcomes the traditional limitations of storage and computation since we make use of Hadoop framework with MapReduce. Also helps in analyzing student's performance and behavior.

## 3. SYSTEM DESIGN

### *3.1 Architecture of HDFS*

The architecture of the system, shown in Fig. 1, is based upon the functionality provided by HDFS.

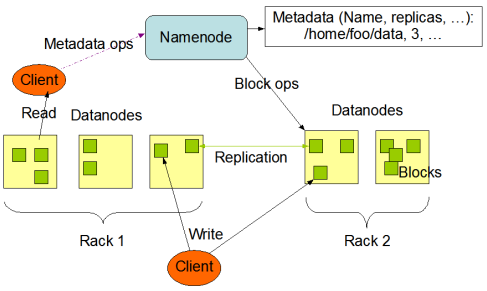


Fig.1 Architecture of HDFS

The Hadoop Distributed File System Architecture consists of two nodes, termed as Name node and Data node .The hadoop clusters consists a single name node and multiple data nodes. The role of name node is to manage the file system by recording and maintaining metadata. This can be controlled by the client's application. Multiple data nodes is use to manage the file storage of devices attached to the cluster. Name node is also known as master and datanode is also known as slaves.

When storing the file into HDFS, it splits the file into one or more blocks and these blocks are stored in a set of data nodes to ensure parallel write or read can be done even on single file. To perform operations like opening, closing and renaming of files and directories in a HDFS client uses a name node. Data nodes are used to serve read and write request from HDFS user. The main functionality of the data nodes is to store and retrieve the blocks when requested by a client application from a name node. In order to manage the updates of the current status, data nodes report the list of blocks that they are storing to the name node periodically.

### *3.1 Architecture of MapReduce*

The architecture of the system, shown in Fig. 2, is based upon the functionality provided by Map and Reducer technology.

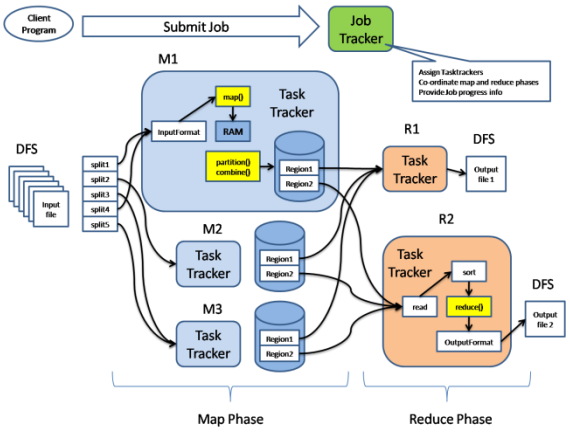


Fig.2 Architecture of MapReduce

Map-Reduce architecture consists of two processing stages, map stage and reduce stage. The intermediate process takes place between these stages which undergoes operations like shuffle and sorting of the mapped data[2].

### *Mapper Phase*

It takes the input as two components called key and value. These key and value are used as pairs i.e. <key-value> pair. During the process stage key is writeable and comparable, but value is only writeable.

### *Reducer Phase*

It takes mapped data as input in the form of shuffled and sorted data. These <key-value>pair data are used to perform required operations to generate desired output.

## 4. METHODOLOGY

### *4.1 General Framework*

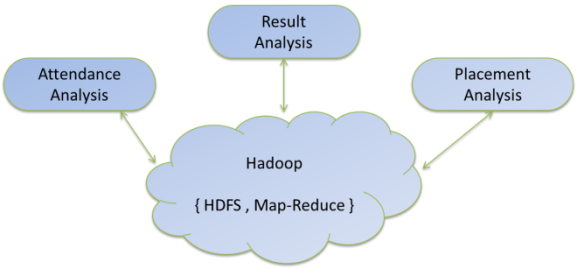


Fig.3 General Framework

Fig 3 shows the generic framework of the proposed system. Each sub-module in the base framework are explained briefly as below,

### *4.1.1 Placement Analysis*

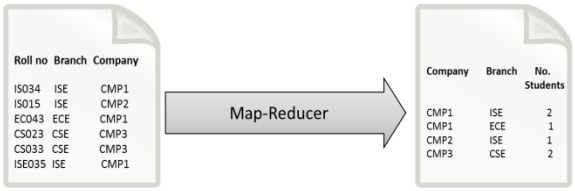


Fig.4 Placement analysis

As shown in above figure Fig 4, the input to the system will be a file havingrecords of all students of BE containing the roll no., branch and company in which he/she is placed. The system will give the report containing the no. of students placed from respective branch in certain companies.

### *4.1.2 Result Analysis*

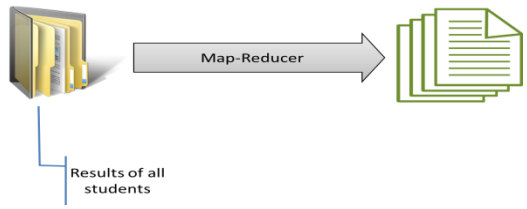


Fig.5 Result analysis

As shown in above figure Fig 5, the input to the system will be a file having result of each student of BE containing the roll no., branch and each subject marks. The system will give the report containing the no. of students and result of the particular subject.

### *4.1.3 Attendance Analysis*

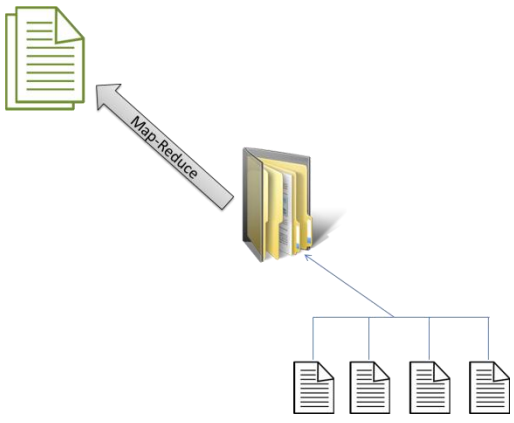


Fig.6 Attendance analysis

As shown in above figure Fig 5, the input to the system will be a attendance of each and every subject. Based on the attendance of each subject average attendance is calculated.
From the average attendance no. of students in the defaulter list is calculated.

### *4.2 Implementation*

4.2.1 Mapper Algorithm

**Algorithm** MAP (key, value)

**Input**: Filename.csv, the offset *key*, the sample *Value*

**Output**: <*key*', *value*'>pair, where the *key*' is the index of the closest center point and *value*' is a string comprise of sample information

1. Read every entry of student data from input file
2. Construct the sample *instance* from *value*;
3. Check if the data is valid, if not skip that line
4. Compute the grade of student
5. Represent grade as key and percentage of student as its value
6. Output < *key*, *value*>pair;
7. End.

The input to the system will be a file having result of each student of BE containing the roll no., branch and each subject marks this algorithm produce the grade and percentage of its value as <grade, percentage>.

4.2.2 Reducer Algorithm

**Algorithm** REDUCE (*key*, *Value*)

**Input**: *key* is the index of the cluster; *Value is* the list of the partial sums from different host

**Output**: < *key*, *value*>pair, where the *key*' is the index of the Cluster, *value*' is a string representing the new center

1. Read each instance from the intermediate file, which contains mapped data in sorted manner
2. Initialize a counter NUM = 0 to record the number of enters in the same cluster;
3. Initialize a counter TOTAL to record the sum of values in the same cluster;
3. If the new key is same as old then
4. Increment the counter NUM and add the value to counter TOTAL
5. If key did not match then print recorded the values in key-value pairs
6. Reset the counter and record current values
7. Divide the TOTAL by *NUM* to get the average.
8. Output average

Reducer will sort the Mapped data, and it will produce the output as a number of students, particular grade and average percentage of all students.

## 5. CONCLUSION

Using predictive analytics on the data that is collected give educational institute insights in future student outcomes. These predictions can be used to change a program if it predicts bad results on a particular program or even run a scenario analysis on a program before it is stared. Universities and colleges will become more efficient in developing a program that will increase results thereby minimizing trial and error. This work aims to cope with changing requirement and making easy and efficient storing and retrieving of both structured and unstructured datasets. 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.

## REFERENCES

[1] S Rajeswari, R Lawrance "Classification model To Predict the Learners Academic Performance using Big Data" ,ICCTIDE,2016

[2] Maedeh Afzali, Nishanth Singh, Suresh Kumar "Hadoop MapReduce:A Platform for Mining Large Datasets",in International Conference on Computing for sustainable global development(INDIAcom),2016.

[3] G. Vaitheeswaran, and L. Arockiam,:Big Data for Education in Student's Perspective",IJCA and ICACCTHPA-2014

[4] Sachin Sharma, Diksha Sharma, PankajVaidya (2014): "Analytics In Education Using Big Data",IJARCSSE,2014.

[5] D Laney.3-D data Management:Controlling Data Volume,Velocity and Variety. META Group Research Note,February 2012