International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Predicting Academic Course Preference using Hadoop

G. Divya Jyothi, N. Shirisha, P. Subhashini, M. Anusha

Abstract: ***These days new technologies have been introduced by this new academic trends also have been came into existence into the education system. And this leads to huge amounts of data which makes a big challenge for the students to store the preferred course. For this many data mining tools have been invented to convert the unregulated data into structured format to understand the meaningful information. As we know that Hadoop is a distributed file system which is used to hold huge amounts of data this stores the files in a redundant fashion across multiple machines. Due to this it leads to failure and parallel applications do not work. To avoid this problem we are using Mapreduce for decision making of students in order to choose their preferred course for industrial training purpose for their effective learning techniques to increase their knowledge and capability.***

Keywords: Big Data, Mapreduce.

I. INTRODUCTION

The main objective of Data mining is it is used to retrieve meaningful information from huge amount of unstructured and distributed data using parallel processing of data in most of prominent areas in modern technologies.

In that case, if use Hadoop cluster it gives the fastest results without any error problems and also the server doesn't go down. This can be to supply quickest results to the scholars and conjointly to boost the standard of social control choices. For example,if there are so many students not qualifying in one particular subject that can be removed or necessary changes can also be done for future endeavors.There is a way to comprehend the best possible level of quality at intervals the academic activity system is by discovering knowledge from educational info to review the foremost attributes that may have a sway on the scholar's performance.The discovered data may be wont to supply a useful and constructive recommendation to the educational planners in educational activity institutes to reinforce their decision-making method, to improve student's academic performance and cut back failure rate, to higher perceive student's behavior, to help instructors, to enhance teaching and plenty of alternative advantages.

In today's generation huge information, vast amounts of structured and unstructured student information are being created daily. Massive data is difficult to work with and desires massively parallel package running on large amount of computers.

Big data information support an enormous quantity of knowledge and conjointly it takes less time to finish the classification method. Therefore the time quality is low, to keep up accuracy and time quality, the Mapreduce idea is introduced.

Mapreduce could be a programming model that simplifies distributed applications that handle massive information. Increasing digitization of student records means that prognosticative analytics is anticipated to rework teaching and become a key tool in learning additional regarding students. Prognosticative analytics could be a method during which information collected regarding the scholar, generally attending, subjects taken, assessment is employed to know learning patterns, establish talent gaps, predict performance and establish learning opportunities. The effective feature choice technique is needed to research the economical classification algorithmic rule. Handling big data, a novel partition mechanism in MapReduce is also required. This is the main part of process during a hadoop system because it supplies the logic of method. In another words, mapreduce can be a coding system framework that helps in writing applications that processes big information sets victimization distributed and parallel algorithms at intervals Hadoop setting.

Mapreduce program consists of, Map() and Reduce() those are two unit functions. Map performs actions like grouping, filtering and sorting. If it reduce perform aggregates and summarizes the result created by map perform.The output generated by the map perform might be a key value combine(key,value) that acts as a result of the input for reduce perform.This comparatively straightforward plan has widespread applications to business.

Manuscript published on November 30, 2019.
* Correspondence Author
G. Divya Jyothi , Department of Computer Science and Engineering, MLR Institute of Technology, Dundigal, Hyderabad, India. Email: divyag.1605@gmail.com
N. Shirisha *, Department of Computer Science and Engineering, MLR Institute of Technology, Dundigal, Hyderabad, India. Email: grandhishirisha@mlrinstitutions.ac.in

BigData-5V's:

The term "big knowledge" is outlined as data that becomes thus giant that it cannot be processed exploitationstandard strategies. the dimensions of the knowledge which might be thought-about to be huge knowledge could be a perpetually varied issue and newer tools area unit incessantly being developed to handle this huge data. so as to create sense out of this overwhelming quantity of knowledge it's typically softenedexploitation 5 V's: Velocity, Volume, Value, Variety, and Veracity

II. LITERATURE SURVEY

Before Mapreduce, doing this type of calculation would are troublesome. Currently programmers will handle this type issues with an ease. The advanced algorithms have been coded by the knowledge scientists for frameworks so that it becomes easy to use for the programmers. They don't want the department of PhD scientists to develop a whole complex framework. As Mapreduce will work on network which provides a straightforward analysis.

MapReduce is gaining much users as a result of the Apache Hadoop and Spark parallel computing systems. Let the programmers use Mapreduce to run models over huge distributed sets of data and use advanced techniques of math and machine learning techniques so that we can predict the results easily realize patterns, uncover correlations, etc.

A. Existing System

In the existing system we are using Apriori algorithm to collect item sets which are frequently occurred in dataset. And it is also difficult to work on various computational parallel applications using apriori algorithm. We can also use the data mining tools like MangoDB which is an open source database and apache hadoop.

B. Proposed System

MapReduce could be a recent programming model that simplifies distributed applications that handle massive information. For implementing the thought of MapReduce, it's to divide the information and perform the information mining method. Then the collective results created. With the quick emergence of leading edge technologies, ancient information management solutions square measure inadequate to catch up with them. These technological changes square measure golf shot pressure on the adoption of "big information." to understand why massive information is far higher than RDBMS for information analytics we've to understand the benefits of mistreatment massive information for analytics.

III. PREPARE YOUR PAPER BEFORE STYLING

A. Hardware requirement

- Hard Disk of size 1-4 TB
- 8GB RAM
- 64 Bit OS

B. Software requirement

- Eclipse IDE -MARS
- JAVA –JDK-1.7
- WinScp
- Putty
- Hadoop 2.x-hadoop version

IV. SYSTEM DESIGN

A. System Architechture

Mapreduce has became of the foremost often used framework for processing of giant quantity of knowledge hold on in Hadoop cluster. It is used for multiprocessing of giant quantity of knowledge speedily. Firstly, it had been designed by google to produce the correspondence and cut back the fault tolerance of knowledge.

MR processes the info within the type of key price pairs. we are able to select the key price pairs supported our alternative. We need to use the key price pairs for mapreduce as our schema isn't static. after we have static schema we are able to use columns for analysing the info.

Map cut back API can furnish the next choices like process, multiprocessing of giant amounts of knowledge and high accessibility.

The Map cut back work flow can undergoes totally different phases that stores the lead to hdfs with replications at the top. Job Trackers can do the work of checking all the Map cut back jobs that ar acting on the Hadoop Cluster.

The Job huntsman can play an important role in planning jobs and it'll keeps the track of each map and cut back jobs. The task hunter can do the particular map and cut back jobs. Map cut back design principally consists of 2 process stages. 1st one is that the map stage and thus the opposite is cut back stage.

Between these 2 stages there's an extra stage referred to as intermediate stage that will the work of taking the input from the mappers and doing the tasks like shuffle, sort, mix etc.

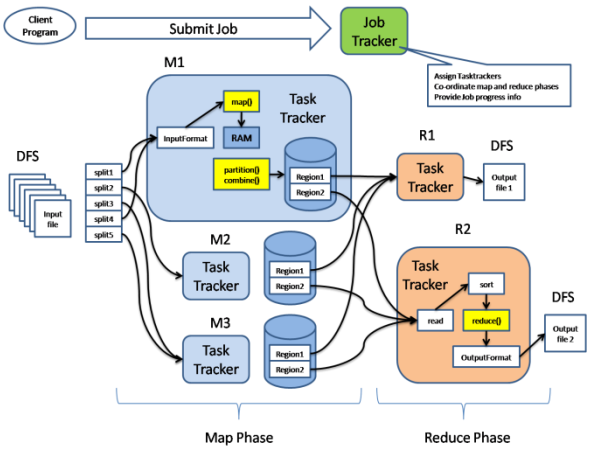


Fig. 1. Mapper Phase

a) Mapper Phase

In this phase, the mapper get the input values from the record reader it sends the key value pairs to the mapper as it can understand only them.

The input received by the mapper will be splited into some other key value pairs so that based on each key they go to the specified reducer which depends on the partitioning constarint.

The out put of this pahse is also key value pairs that is indermediate key value pairs.

b) Intermediate Process

This is the phase which comes in between the map and reduce phases. In this phase many operations are done based on the results required.

In this phase the same key values from different mappers will get into one mapper i.e., here the operations like shuffle, sort etc are done.

It uses the Round Robin algorithm to write the intermediate key values pairs into the local disk so that the condition is obtained correctly.

c) Reducer Phase

This is the second stage of the Map Reduce data flow. In this phase it receives the input from the partitioner and combiner. After performing them the intermediate values are taken as the input for this phase.

The reducers logic will begin with the operations performed by the mapper. It produces the output files like part files which contains the actual output of the analysed data.

Each time when we run a job it shows the number of reducers needed for the our job to jet executed. In the mapreduce configuration file we need to change the permissions so that the partitioning tasks can be performed.

As it is a parallel processing job when one mapper completes it's tasks and sit then the task tracker will assign the other task to it from the mapper which has many tasks to do. This will produce the output which is optimized

By using this we can increase the performance and reduce the processing time.

V. IMPLEMENTATION

Steps: 1

Create a new directory with name Course in your cluster,

EdgeNode@$ HadoopfsmkdirCourse;

Steps: 2

Write your Map Reduce Program in Eclipse,

Course.java

Steps: 3

Create a Jar file,

File.jar

Steps: 4

Copy the jar file, input file to local edgenode using winscp,

File.jar, cc.txt

Steps: 5

Login into your cluster using putty.

Steps: 6

Copy the input file from local to hadoop cluster.

EdgeNode@$ hadoopfs–copyFromLocal cc.txt cc.txt

Steps: 7

Run the Mapreduce Program

EdgeNode@$ hadoopfs jar File.jar Data.txt output

Steps: 8

Result can be seen through command interface as,

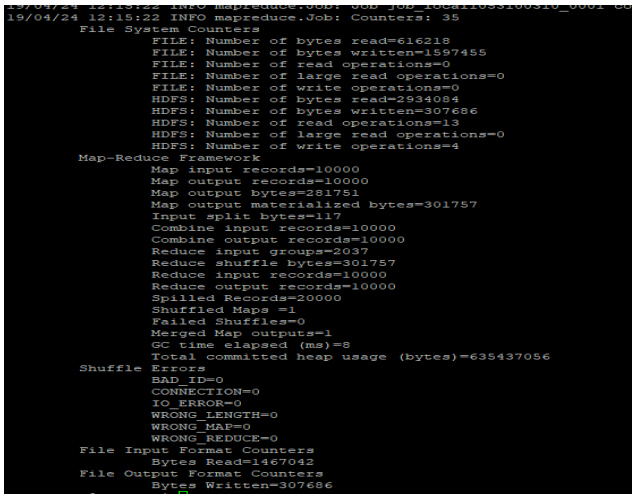
EdgeNode@$ hadoopfs -cat output/part-*

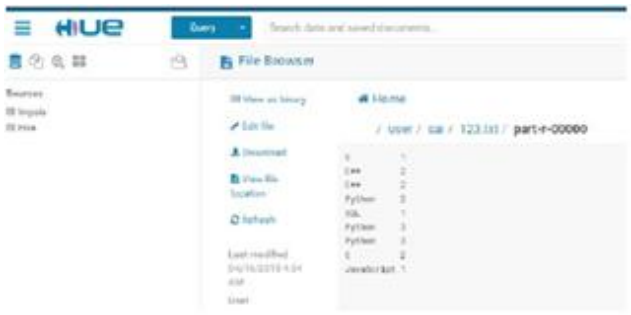
Or

You can use GUI for results

By using the address of the cluster : 172.16.103.68:8888

VI. RESULTS





VII. CONCLUSION

Storing the data and processing the data is the hectic task now-a-days in various businesses. Asdata is increasing tremendously we do not have appropriate ways to store and process this data. Traditional methods are not capable of handling this data. In this scenario, Hadoop comes into picture.HDFS solves the problem of storing and MapReduce can be used to process data. In our project we are using MapReduce framework to analyze student data. There are many ways to store and process the data, but these days' storage and handling of data has become a major issue, instead of using traditional databases we are moving towards a new Solution for dealing with large amount of data that is HDFS in Hadoop.

The main objective of our project is to analyze student data in order to provide information for the education system to maintain records and help in making faster and useful decisions for the benefit of the students. We have taken a data set of students which contains student roll number and credits in a csv format. Next steps are as follows. We obtain results from moving the data from local to Hadoop and execute the MapReduce program by converting it in to a jar file. From these results we can analyze students who are graduated and who are not graduated and credits obtained.
This makes education system know the number of graduates from their institution and their by take necessary steps to increase their graduate percentage.

VIII. FUTURE ENHANCEMENT

Our project can be used as basis for development of student website were results access is far faster when compared to traditional formats. This method is highly reliable as it can deal with huge volumes of data. We are going to include more courses for better result generation.

REFERENCES

1. Sai Prasad K., Amarnath Reddy E.," *A new document representation approach for gender prediction using author profiles*",In Advances in Intelligent Systems and Computing,2019
2. Raghunadha reddy T., Gopi Chand M., Hemanath K.," *Location prediction of anonymous text using author profiling technique*",In International Journal of Civil Engineering and Technology,2017.
3. Madhuravani B." *Notification of data congestion intimation [NDCI] for IEEE 802.11 adhoc network with power save mode*", Smart Innovation, Systems and Technologies,2018.
4. https://www.guru99.com/create-your-first-hadoop-programs.html
5. http://a5academics.com/tutorials/83-hadoop/840-map-reduce-architecture.
6. https://readwrite.com/2013/0557/23/hadoop-what-it-is-and-how-it-work/
7. https://data-fair.training/blogs/hadoop-partitioner-tutorials/

AUTHORS PROFILE

G Divya Jyothi, M.Tech., from JNTUH University, having five yeare of teaching experience.

N. Shirisha, M.Tech., (Phd.), doing research in big data security, ISTE.,completed B.Tech.,M.Tech., from jntuh university. Having nine years of teaching experience.