International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 08 | August -2020 ISSN: 2582-3930

Big Data Analytics:Predicting Academic Course Preference Using Hadoop Inspired MapReduce

Kumar Abhirup
Computer Science dept.
Cambridge Institute of Technology
Bengaluru, India

Shivraj Bohara
Computer Science dept.
Cambridge Institute of Technology
Bengaluru, India

Nishant Kumar Singh
Computer Science dept.
Cambridge Institute of Technology
Bengaluru, India

Abhishek Kumar
Computer Science dept.
Cambridge Institute of Technology
Bengaluru, India

Abstract –As technology is becoming advance and evolving exponentially new academic courses were introduced into educational system. As these data are unregulated which become overwhelming for the students to choose and learn technology which are helpful for them in their industrial training and increase their career prospects. So, to solve this problem we designed a system which predict the course preferences for pursuing training for students based on course combination. The first challenge is to store and manage this large amount of unregulated data and second challenge is to convert this unregulated data into structured and meaning information. To conquer the first challenge we are using Hadoop Distributed File System, to store and manage large amount of data. In HDFS the files are stored in redundant fashion across multiple machines which ensure their endurance to failure and parallel processing. And to conquer the second challenge we are using data mining tool to mine frequently occuring course in large amount of dataset. Map Reduce is used to extract useful information which will be helpful in decision making for students. Here, using HDFS, tasks run over Map Reduce and output is obtained after aggregation of results.

Keywords– Hadoop Distributed File System, Data Mining, Map Reduce, frequently occuring course, parallel processing.

I. INTRODUCTION

Data mining is one of the most popular technology in extracting useful and meaningful information from a huge amount of unstructured and distributed data using parallel processing of data. Data mining techniques were used by industries to analyze large amount of data of their customer to understand their behavoiur and purchasing pattern, which help them to take decisions to grow their business. It is also helpful for educational sector to analyze the students feedback, latest academic trends etc which will helps in providing quality education and decision - making approach for students to increase their career prospects.

As there is a huge gap between what the colleges are teaching and the skill requirement of industry, to fulfill this skill gap data mining can be used.

Big Data is the growing field of data mining. Big data deals with a large amount of dataset which is so large and complex that it cannot be handled and processed by any other traditional data processing application software. Big Data gathers all data for storage then processes and analyze large amount of data to reveal patterns, trends, and associations. It deals with unstructured data like Text, MS office file, PDF etc. and structured data may be the relational data. Big Data comes with 3 V's and those are Velocity, Variety and Volume. The first V which is Velocity means as data is increasing at a very fast rate, it is estimated that the value will double in every 2 years. The second V which is Variety which means, like data comes with variety one is structured and other is unstructured , so Big Data can deal with both types of data and last V is volume and it means the amount of data which we can deal with is of very large size of petabytes.

Hadoop is one technique of big data which solves the problem of managing and storing of unstructured and huge amount of data. Hadoop is a collection of open - source software utilities which provide a distributed storage and parallel processing environment. It also provide processing of big data using Map Reduce programming model. Map Reduce job run over Hadoop clusters by splitting the Big Data into small chunks and process the data by running it parallel on distributed cluster.

II. PROPOSED SYSTEM

We developed a system which predict academic course preference according to the trending technology which can help students to decide which course to choose, learn and update their skill according to the industry need. Our system comes with an interface in which user can login through admin credential and do parallel frequent mining on dataset and get the predicted academic course preference as a result in the form of graph.

As shown in Fig:1 we have two buttons present, one is admin screen button and another is user screen button, to do parallel frequent pattern mining which should be done to predict academic course preference we need to go to admin screen.

Fig: 1 - First window screens

And after that as shown in Fig:2 we get a admin login window, by using admin login credential we can login and after successful login we get a new window as shown in Fig:3, by using this window we can do Parallel Frequent Pattern mining by clicking on startParallelFpMining and we get the recommended course as a result and also user can view the academic course preference from most preferable to least preferable course in the form of graph by just clicking ViewGraph button.

Fig: 2 - Admin Login Window

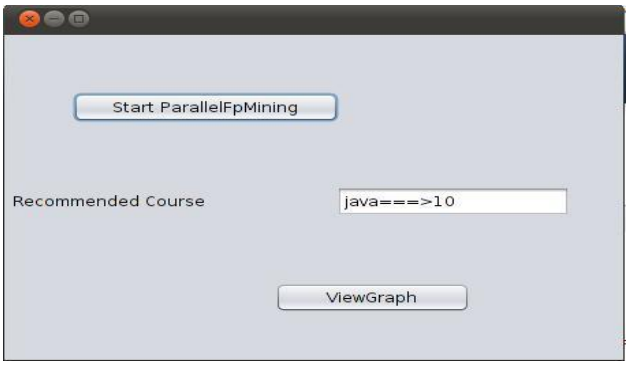


Fig: 3 - Parallel Fp mining window

Parallel Frequent Pattern mining is done on the dataset, the dataset is the collection of course combination entered by the student. As shown in Fig: 1 we also have one more button named user screen, after clicking that we get a new window as shown in Fig: 4 and Fig: 5 in which user can register and then login. After that user can enter the course combination as much they want with space separated. These course combination data helps our system to predict the academic course preference and to create a large dataset resulting in better preferences and help the student to choose most trending and prefered course.

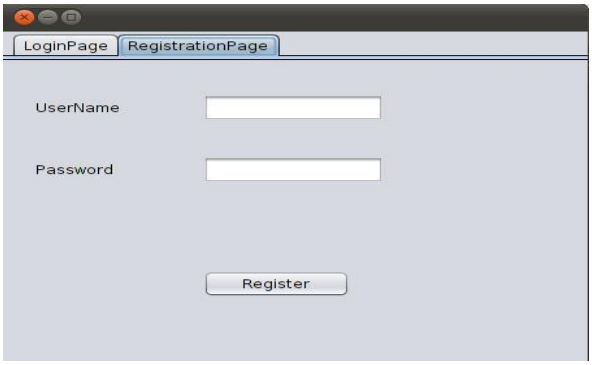


Fig: 4 - Registration Window

Fig: 5 - Login Window

We are using Hadoop environment and it comes with advantages like, we can manage, store and perform parallel processing on large amount of data. The processing time will be reduced due to parallel processing. We had configured Hadoop environment in single node clustering in which one single machine is used and it divides the large amount of data into small chunks and give it the different part of processor present in that single machine to do parallel processing.

Existing System tells about the Apriori scheme for generating the frequent courses. In this Apriori scheme, the data courses should be placed in memory. Problem in existing system is that the candidate courses do not fit into memory for greater depth in the limitation of using Apriori. So, to overcome this problem we are using Apriori algorithm with Frequent Pattern Growth algorithm to do frequent pattern mining. The input dataset taken is implemented through Hadoop, where the clustering for the dataset (course combination) done.

We will mine the frequent course on each cluster in Hadoop and compute the Global Tid and mine the courses to calculate frequent item set. And also we recommend the courses based on the dataset (combination of courses) entered. As the students take the training of courses, based on the courses elected, cluster the courses and recommend the elected course.

III. SYSTEM ARCHITECTURE

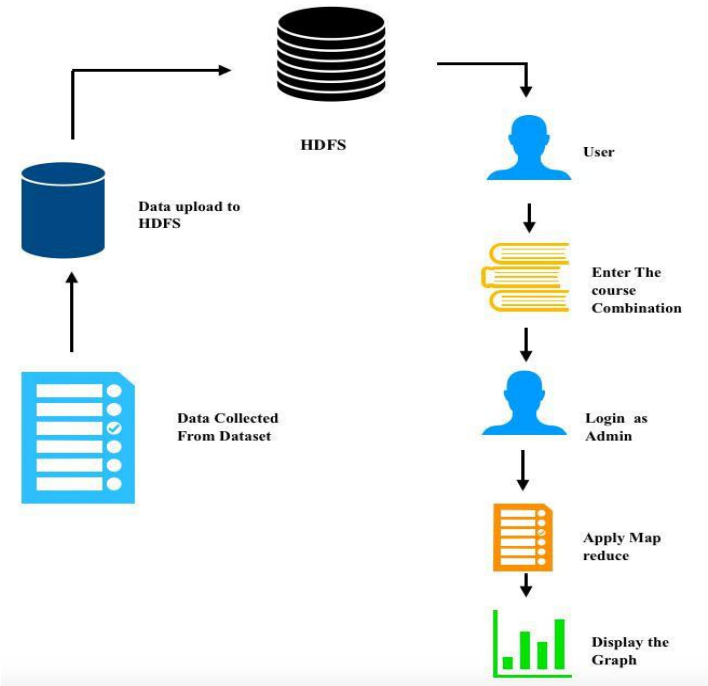


Fig: 6- System Architecture

As shown in Fig: 6, the data is collected from dataset, which is stored in system as a file for frequent pattern mining. The data is then uploaded to HDFS system. HDFS architecture is shown in Fig: 7, HDFS is a block - structured file system in that each and every file is divided into block of fixed sized, these blocks are then stored across cluster of one or more machines. HDFS follows a Master - Slave architecture. The cluster comprises of single NameNode also called Master Node and all other nodes are DataNode also known as Slave Node . The NameNode acts as master and it manages and assign tasks to it's DataNode and also keep track of DataNodes.

The dataset which was uploaded to HDFS is splitted into small chunks of data by NameNode and assign to different DataNode and also instruct the processing methodology to DataNode.

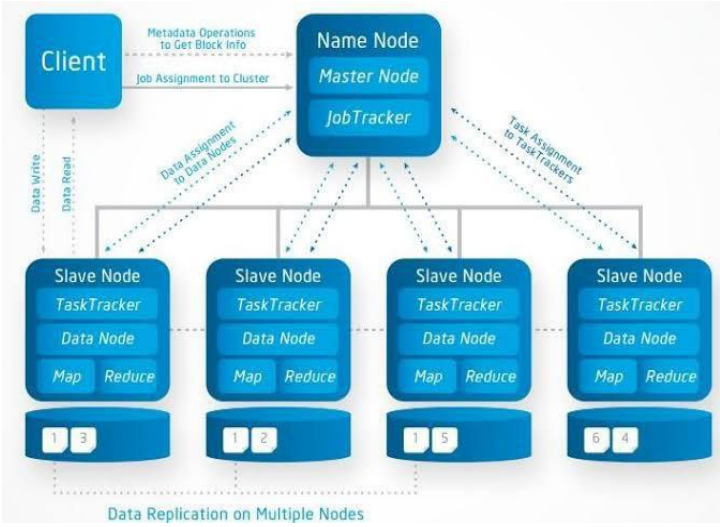


Fig: 7 - HDFS Architecture

Map Reduce is a programming model which we are using for processing huge amount of data. The Map Reduce resides on each and every DataNode for parallel processing. The Map Reduce program works in two phases, one is Map and another is Reduce as shown in Fig: 8. The Map phase splits and do mapping of data and Reduce phase shuffle and reduce the data. To perform Mapping task we are using Apriori algorithm with FP Growth algorithm in Map function to perform frequent pattern mining in our dataset. Reducing task is done once the mapping phase completes and reduce phase is just a aggregation of result of Mapping phase. User can only apply Map Reduce when user login as admin and can view the result in form of graph.

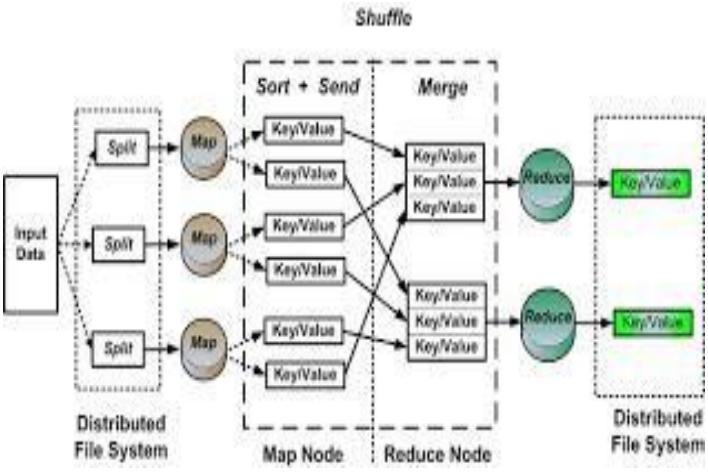


Fig: 8 - Map Reduce Architecture

As use case 1 (shown in Fig: 9) shows a user can register and then login and can enter any number of course combination with space separated, these data are used by the mining algorithm to mine frequent occuring of courses.

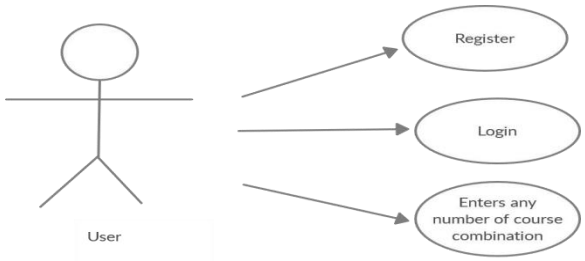


Fig: 9 - Use case 1

And use case 2 (shown in Fig: 10) shows that a user can do mining by applying Map Reduce only when he/she login as admin and can display the result in form of graph.

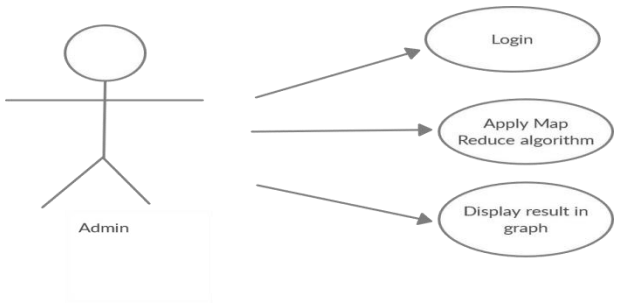
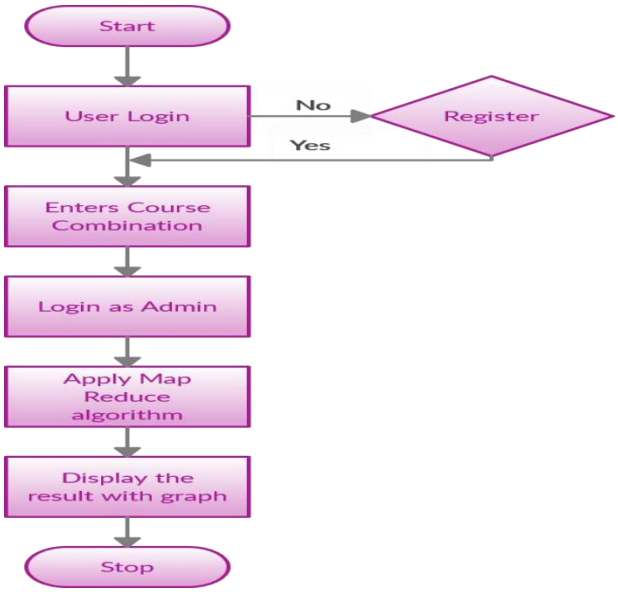


Fig: 10 - Use case 2

The whole flow of our system is shown in Fig: 11 in the form of flow chart.



.

Fig: 11 - Flow Chart

IV. METHODOLOGY

To design the proposed system we had used some methods to create dataset and to process those dataset to get a better result. To create dataset we done online survey and research about the trending course and course combination preference of industry experts, which helped us to design a system which can give a better course preference. Another important part is processing of data, for the better mining of data we used Apriori alogorithm with FP Growth algorithm. And these two algorithm works together in Map function and the aggregation of Map result gives the output which comes under Reduce phase.

Apriori algorithm is used for frequent itemset mining in the larger amount of data. It works in pairwise form, Apriori algorithm mines the large amount of data and find the combination or pair of item which occurs frequently in the itemset. Apriori algorithm is used by many industries to understand the customer purchasing pattern by analyzing their transaction or purchasing details, so they can grow their business by increasing the supply of that particular items. Fig: 12 shows the Apriori algorithm.

F_1 = {frequent items of size 1};
for ($k = 1; F_k$!=ϕ; k++) do begin
  C_{k+1} = apriori-gen(F_k); // New candidates generated from F_k
  for all transactions t in database do begin
    C'_t = subset (C_{k+1},t); // Candidates contained in t
    for all candidate $c \in C'_t$ do
      $c.count$++; // Increment the count of all candidates in C_{k+1} that are contained in t
    end
    F_{k+1} = {$C \in C_{k+1} | c.count \geq$ minimum suport}
    //Candidates in C_{k+1} with minimum support
  end
end
Answer $\cup_k F_k$;

Fig: 12 - Apriori Algorithm

Another algorithm which we are using with Apriori algorithm is FP Growth algorithm. FP Growth algorithm do not work in pairwise way, it is used for finding frequent itemset without candidate generation. FP Growth algorithm represents frequent item in frequent pattern tree. The result of FP Growth algorithm is our final result, which shows the highest to lowest preferred courses which occures frequently and the number of times it occured, which helped us to show result in form of graph.

The FP Growth algorithm, pseudocode with example is shown in Fig: 13.

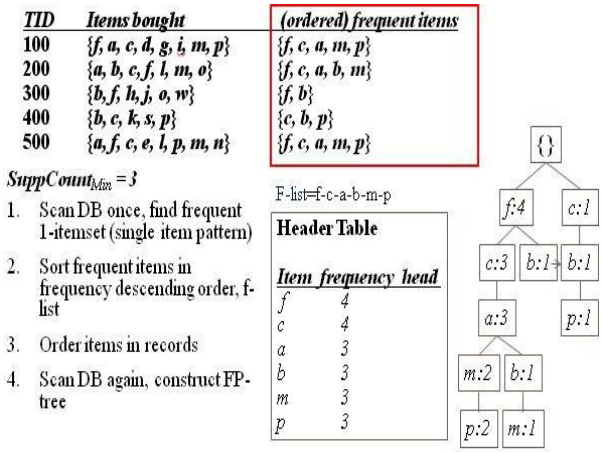


Fig: 13 - FP Growth algorithm

V. RESULT

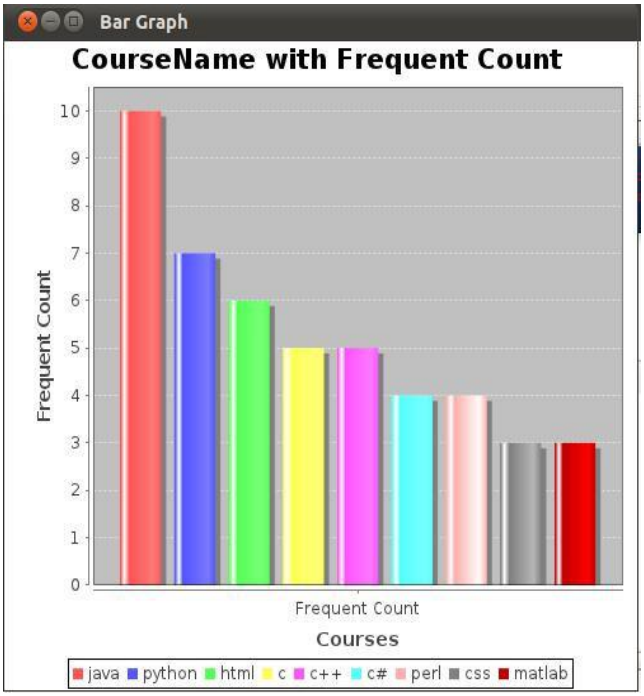


Fig: 14 - Preferred Courses

As shown in Fig: 14 the final result is displayed in form of graph. The graph shows the preferred courses ranging from most preferred to least preferred. When user login as admin, user can do parallel frequent mining and get result.

CONCLUSION

This system provides a academic course preference which help student to choose courses according to the trend in industry to increase their career prospects. Hadoop environment is used to reduce the probabality of system failure and to achieve parallel processing of large amount of data to reduce the execution time. Map Reduce programming model is used to do processing of large amount of data in a parallel fashion, as it can run on different DataNode parallelly. Hadoop environment also provide data security for our system. Apriori algorithm with FP Growth algorithm is used to mine greater depth in data and frequently occuring of itemset, these two algorithm is used in Map function. The input dataset taken is implemented through Hadoop where the clustering for the dataset done and we mine the frequent course on each clusters in Hadoop. The result is displayed in graph format so it will be easy to understand and help in making decision based on preferred course by the system.

ACKNOWLEDGEMENT

It is a great pleasure to thank our project guide Dr. ShashiKumar D.R., HOD of Dept. Computer Science and Engineering, for his guidance and continuous support throughout the course of doing this project.

REFERENCES

[1] Prajwal M G, Anand Shankar: “Big Data Analytics: Predicting Academic Course Preference Using Hadoop Inspired Map Reduce”. School of Computing And Information Technology, Reva University, Bengaluru, (2019)

[2] Xin YueYang, Zhen Liu, Yan Fu, “Map Reduce as a Programming Model for Association Rules Algorithm on Hadoop”, Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on, pp. 99-102. IEEE, 2010.

[3] Jongwook Woo, "Apriori-Map/Reduce Algorithm." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[4]B.Manjulatha,AmbicaVenna,K.Soumya,"Implementa -tion of Hadoop Operations for Big Data Processing in Educational Institutions",International Journal of Innovative Research in Computer and Communication Engineering,ISSN(Online) : 2320-9801,Vol. 4, Issue 4, April 2016.

[5]Shankar M.Patil,Praveen Kumar, “Data Mining Model for Effective Data Analysis of Higher Education Students Using MapReduce”,IJERMT,ISSN:2278-9359,Vol.6,Issue 4,April 2017.