International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792

Predictive Analysis and Data Mining Using Hadoop Inspired Map Reduce to Identify Trends and Patterns in Academic Courses

H. R. B. Phani Kumar[1], Bimal Kumar[2], S. Akhilendranath[3]

[1]*M. Tech. Student, Dept. of CSE, Shri Shirdi Sai Institute of Science and Engineering, Ananthapuramu, India*

[2]*Associate Professor, Dept. of CSE, Shri Shirdi Sai Institute of Science and Engineering, Ananthapuramu, India*

[3]*Assistant Professor, Dept. of CSE, Shri Shirdi Sai Institute of Science and Engineering, Ananthapuramu, India*

***Abstract*: Predictive analytics is a kind of analytics that uses both new and historical data to forecast activity, behavior, trends and patterns. It involves applying statistical analysis techniques, analytical queries, mathematical formulas and automated machine learning algorithms to data sets to create predictive models that place a numerical value or score on the likelihood of a particular event happening.**

With the introduction of new technologies and processes from past couple of years, new academic trends introduced into educational system results in huge data which is unregulated and of different forms such as structured, unstructured and semi structured. Existing systems such as RDBMS's fails to process such data from different sources because of its limitations.

And students are facing so many challenges in selecting their course for their industrial training which will help them in their future careers because of huge unregulated data from different sources and it is very hard them to analyze such un even data and take decisions. We can solve this problem by using Big Data analytics.

***Keywords*: Distributed, data mining, educational data mining, Hadoop, MapReduce.**

1. Introduction

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through analysis. Data mining tools allow enterprises to predict future trends.

Data Mining is most popular in modern technology in processing huge data sets and retrieving meaningful data from structured, unstructured and semi structured data. Hadoop is an open-source programming paradigm which performs parallel processing of applications on clusters.

There is huge advantage to Educational sector of following Data Mining Techniques to analyze data input from students, feedbacks, latest academic trends etc which helps in providing quality education and decision-making approach for students to increase their career prospects and right selection of courses for industrial trainings to fulfill the skill gap pertains between primary education and industry hiring students. In [1], Data Mining has great impact in academic systems where education is weighed as primary input for societal progress.

2. Literature review

In [2], Apriori algorithm is implemented and high performance is achieved using Map Reduce Technique of Hadoop framework to collect item sets frequently occurred in dataset. In [3], author has mainly addressed the challenges of using Map Reduce model for computing parallel application of Apriori. According to Author in [4], Big Data Techniques are the necessity in learning environments and present scenario with large amount of unstructured data and introduction of Massive open online courses in Education has stressed upon the need for data mining in Education. In [5], tools of Data Mining like MangoDB, an open-source database and Apache Hadoop are discussed. Data Mining Techniques using these tools help students in choosing their course curriculum. Author in [6] has used Map Reduce Programming paradigm for predicting Student's performance. Author in [7] has performed classification of data using Map Reduce and proposed Data Mining Model for effective data analysis of Higher Education Students. In [8], author has discussed parallel processing of clusters by Map Reduce over large amount of data. MapReduce scales to large array of machines comprising of thousands of machines which solves large computational problems [9]. Authors in [10] have presented review paper on Big Data and Hadoop. The paper has focused on technical challenges and highlighted Map Reduce techniques proposed by different authors.

3. Existing system

The relational database management system (or RDBMS) had been the one solution for all database needs. Oracle, IBM (IBM), and Microsoft (MSFT) are the leading players of RDBMS. RDBMS uses structured query language (or SQL) to define, query, and update the database. However, the volume and velocity of business data has changed dramatically in the last couple of years.

A. Disadvantages with existing system

1) The data size has increased tremendously to the range of petabytes RDBMS finds it challenging to handle such huge data volumes. To address this, RDBMS added more central processing units (or CPUs) or more memory to the database management system to scale up vertically.
2) The majority of the data comes in a semi-structured or unstructured format from social media, audio, video, texts, and emails. However, the second problem related to unstructured data is outside the purview of RDBMS because relational databases just can't categorize unstructured data. They're designed and structured to accommodate structured data such as weblog sensor and financial data.
3) Big data is generated at a very high velocity. RDBMS lacks in high velocity because it's designed for steady data retention rather than rapid growth. Even if RDBMS is used to handle and store "big data" it will turn out to be very expensive.

As a result, the inability of relational databases to handle "big data" led to the emergence of new technologies.

4. Proposed system

In proposed system we collect data in various forms and store it using Hadoop Data File System (HDFS). And we process those datasets using Hadoop inspired Map Reduce Framework which does parallel computing

Advantages of Proposed System

1) Huge data can be stored using HDFS with data redundancy and high availability.
2) Hadoop cluster can be horizontally scalable.
3) Hadoop can handle variety of data such as structured semi structured and unstructured data.
4) Hadoop is cost effective and can be managed using commodity software.
5) Response time is very less compared to Traditional RDBMS's system in processing such huge data sets.

5. Hadoop architecture

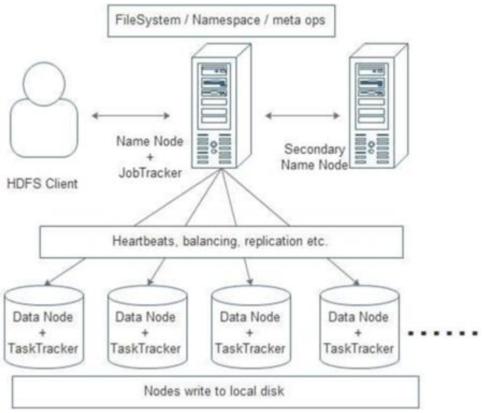


Fig. 1. Hadoop architecture

A. Hadoop Daemons

1) Name node

Name Node works as Master system in Hadoop cluster. Name node stores meta-data i.e. number of blocks, replicas and other details. Meta-data is present in memory in the master. Name Node also assigns tasks to the slave node. As it is the heart of HDFS, so it is good to be deployed on reliable hardware.

2) Data node

Data node works as Slave in Hadoop cluster. In Hadoop HDFS, Data Node is responsible for storing actual data in HDFS. Data Node performs read and write operation as per request for the clients. Data Nodes can also deploy on commodity hardware.

3) Secondary name node

Its main function is to take checkpoints of the file system metadata present on name node. It is not the backup name node. It is a helper to the primary Name Node but it does not replace the primary name node.

4) Resource manager

It is a cluster level component and runs on the Master machine. Hence it manages resources and schedule applications running on the top of YARN. It has two components: Scheduler and Application Manager.

5) Node manager

It is a node level component. Node Manager runs on each slave machine. It continuously communicates with Resource Manager to remain up-to-date

B. Job Tracker and Task Tracker

- The primary function of the job tracker is resource management (managing the task trackers), tracking resource availability and task life cycle management (tracking its progress, fault tolerance etc.)
- The task tracker has a simple function of following the orders of the job tracker and updating the job tracker with its progress status periodically.
- The task tracker is pre-configured with a number of slots indicating the number of tasks it can accept. When the job tracker tries to schedule a task, it looks for an empty slot in the task tracker running on the same server which hosts the data node where the data for that task resides. If not found, it looks for the machine in the same rack. There is no consideration of system load during this allocation.
- HDFS is rack aware in the sense that the name node and the job tracker obtain a list of rack ids corresponding to each of the slave nodes (data nodes) and creates a mapping between the IP address and the rack id. HDFS uses this knowledge to replicate data across different racks so that data is not lost in the event of a complete rack power outage or switch failure.
- Job Performance - Hadoop does speculative execution

where if a machine is slow in the cluster and the map/reduce tasks running on this machine are holding on to the entire map/reduce phase, then it runs redundant jobs on other machines to process the same task, and whichever task gets completed first reports back to the job tracker and results from the same are carried forward into the next phase.

- Fault Tolerance - The task tracker spawns different JVM processes to ensure that process failures do not bring down the task tracker.
- The task tracker keeps sending heartbeat messages to the job tracker to say that it is alive and to keep it updated with the number of empty slots available for running more tasks.
- From version 0.21 of Hadoop, the job tracker does some checkpointing of its work in the filesystem. Whenever, it starts up it checks what was it upto till the last CP and resumes any incomplete jobs. Earlier, if the job tracker went down, all the active job information used to get lost.
- The status and information about the job tracker and the task tracker are exposed vis jetty onto a web interface.

C. *YARN - Next Generation Hadoop*

In Yarn, the job tracker is split into two different daemons called Resource Manager and Node Manager (node specific). The resource manager only manages the allocation of resources to the different jobs apart from comprising a scheduler which just takes care of the scheduling jobs without worrying about any monitoring or status updates. Different resources such as memory, cpu time, network bandwidth etc. are put into one unit called the Resource Container. There are different App Masters running on different nodes which talk to a number of these resource containers and accordingly update the Node Manager with the monitoring/status details.

D. *Map reduce Framework*

To process any data, the client first submits data and program. Hadoop store data using HDFS and then process the data using MapReduce.

E. *Hadoop Data Storage*

Let's see how Hadoop stores the data

Hadoop Distributed File System: HDFS is the primary storage system of Hadoop. It stores very large files running on a cluster of commodity hardware. HDFS stores data reliably even in the case of machine failure. It also provides high throughput access to the application by accessing in parallel.

The data is broken into small chunks as blocks. Block is the smallest unit of data that the file system store. Hadoop application distributes data blocks across the multiple nodes. Then, each block is replicated as per the replication factor (by default 3). Once all the blocks of the data are stored on data node, the user can process the data.

F. *Hadoop Data Processing*

Let's see how Hadoop process the data.

Hadoop MapReduce is the data processing layer. It is the framework for writing applications that process the vast amount of data stored in the HDFS. MapReduce processes a huge amount of data in parallel by dividing the job into a set of independent tasks (sub-job). In Hadoop, MapReduce works by breaking the processing into phases: Map and Reduce.

Map: It is the first phase of processing. In which we specify all the complex logic/business rules/costly code. The map takes a set of data and converts it into another set of data. It also breaks individual elements into tuples (key-value pairs).

Reduce: It is the second phase of processing. In which we specify light-weight processing like aggregation/summation. The output from the map is the input to Reducer. Then, reducer combines tuples (key-value) based on the key. And then, modifies the value of the key accordingly.

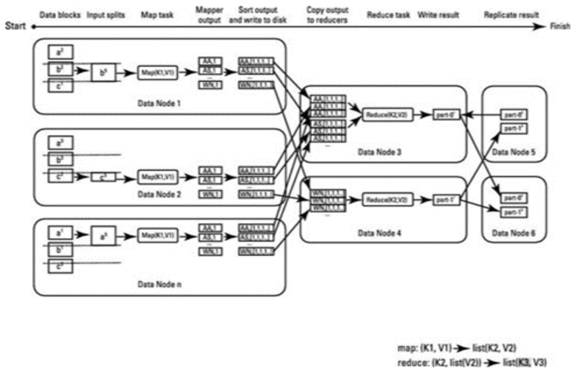


Fig. 2. Hadoop data processing

G. *Data Description*

The Table 1 shows list of courses choose by students for their industrial training. The Pattern of student's choice and latest trends year wise are predicted after processing through Hadoop Mapreduce.

Table 1
Sample data set

Year	Course	Student
2018	Java	S1
2018	C#	S2
2018	Java	S3
2018	SAP	S4
2017	Java	S5
2017	C#	S6
2016	Java	S3
2018	SAP	S4

6. Implementation

- We built a Hadoop cluster with one name node and 25 data nodes configured with Hadoop configuration. And now name node is able to contact all its data nodes.
- The Table 1 data is sample data set. A huge data set of

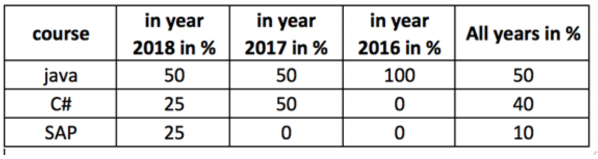
same type is uploaded and saved using HDFS which is distributed across all data nodes with high replication and availability.

- We developed a mapper module which generates intermediate datasets. These mappers are launched on all data nodes where and all input file got distributed. All these mappers run parallelly on data nodes and generate intermediate outputs. These Intermediate outputs are key value pairs for eg.<key,value>
- All intermediate outputs go through the shuffling phase and shuffled as per its key and all keys are aggregated and generate another set of intermediate output which is also in the form of key value pair. But the value is nothing but aggregated output of that particular key for eg.<key,[val1,val2,…]>
- We developed a reducer module which does predictive analysis and apply appropriate mathematical transformations on aggregated data and identify tends and patterns. Later these trends and patterns get transformed into some statistical outputs and get saved as final output report which is called reduced output.
- We developed a driver module to track and log all details of job once submitted to Hadoop cluster. This will track the job and log all outcomes including all errors. Also Driver module sets appropriate mapper and reducer classes for inputs uploaded. And configure output paths for results generated.
- The Table 2, shows the trends of students opted academic courses for their industrial training for past years.

7. Results and discussion

The input data sets for past years for course selection in various branches and streams were collected and saved on HDFS. We will feed this data to Hadoop Map reduce Framework for processing them. The data sets get spitted into various small data sets and distributed to various data nodes and mappers are launched to each small dataset. All mappers run parallelly and feed their intermediate output to reducers after shuffling. Reducers further reduce the intermediate output and finally generate final output which is saved in the system in the format shown in Table 2.

Table 2
Final results



course	in year 2018 in %	in year 2017 in %	in year 2016 in %	All years in %
java	50	50	100	50
C#	25	50	0	40
SAP	25	0	0	10

8. Conclusion

The Map Reduce approach is used for running jobs over HDFS. Using Map Reduce, the application can be scaled to run over multiple machines in a cluster and for that Hadoop cluster is used. The Map Reduce Framework consists of Map and Reduce Functions with single Resource Manager which acts as a master and one Node manager which acts as slave per cluster node. The input dataset is fed into the mapper and after passing through shuffle phase, reducer displays the output after aggregating the tuples obtained from mapper and are in the form of <key, value> pair. The dataset shows that the students have opted for multiple course combinations for industrial trainings and the data becomes unstructured as well confusing for students to opt for course for trainings. The results in this paper shows that large volume of course combinations in the form of input dataset after passed through mapper function in Map Reduce Framework which runs the job in parallel on a single node cluster using HDFS, it converts the data into individual tuples and the meaningful data obtained from Reducer function classifies the data of course combinations opted more by students and strengthens the decision-making of students as well institutions to prefer demanding course for Industrial trainings. Apart from it, the predicted result from Hadoop programming framework which is the emerging field of Data Mining also helps Management to stress over these courses in their curriculum to improve student skills and increases employment chances for them.

References

[1] Sonali Agarwal, G. N. Pandey, M. D. Tiwari, “Data Mining in Education: Data Classification and Decision Tree Approach”, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.

[2] 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.

[3] Xin Yue Yang, Zhen Liu, Yan Fu, “MapReduce 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.

[4] Katrina Sin,Loganathan Muthu,"Application of Big Data in Education Data Mining and Learning Analytics – A Literature Review" ,ICTACT Journal on Soft Computing, Vol. 5, Issue 4,July 2015.

[5] B. Manjulatha, Ambica Venna, K. Soumya,"Implementation of Hadoop Operations for Big Data Processing in Educational Institutions", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2016.

[6] N. Tajunisha, M. Anjali, "Predicting Student Performance Using MapReduce”, IJECS, Vol.4, Issue 1, Jan 2015, pp. 9971-9976.

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

[8] Madhavi Vaidya,"Parallel Processing of cluster by Map Reduce", IJDPS, Vol.3, No.1,2012.

[9] Jeffrey Dean, Sanjay Ghemawat,"MapReduce: Simplified Data Processing on Large Clusters",Google, Inc, OSDI 2010.

[10] Harshawardhan S. Bhosale, Devendra P. Gadekar,"A Review Paper on Big Data and Hadoop", IJSRP, Vol. 4, Issue 10, Oct 2014.