Big Data Analytics using Unsupervised Machine Learning

Rahul Khokale[1], NileshSingh V. Thakur[2], Mahendra Makesar[3], Vijay Gadicha[4]

[1,2,3,4]*Nagpur Institute of Technology, Nagpur*

[1]*softrahul@gmail.com,* [2]*thakurnisvis@rediffmail.com,*[3]*msmakesar@nit.edu.in*

[4]*vbgadicha@nit.edu.in*

Abstract

Data has become the most important entity in today's scenario. Enormous data is being generated per second all around the world. Data is increasing with huge volume, high velocity and diversified variety. Hence it is commonly referred as the Big Data. Big data analytics is going to become the significant tool for business intelligence. In this paper, the framework for big data analytics using unsupervised machine learning is proposed. In the paper, representation of subset of stock market data in Hadoop Distributed File System (HDFS) is described. Further data analytics using Artificial Neural Network (ANN) is presented.

Keywords: *Big Data Analytics, Unsupervised machine learning algorithm, ANN*

1. Introduction

Big data is a voluminous collection of enormous datasets. It is difficult to process big data using traditional computing techniques. It consists of many software tools, techniques and frameworks. Big data comprises of huge and complex datasets which can be existed in structured, semi-structured, or unstructured form. These large datasets require large memory for storage. They have to be processed with advanced technology and by using in place computations which means that computation has to be done where the data resides for processing. Big Data possesses a typical 3Vs model, which are velocity, volume, and variety. Velocity refers to real-time speed of data at which it is increasing. A typical example of this would be to perform analytics on a continuous stream of data originating from a social networking site or aggregation of disparate sources of data. Volume refers to the size of the dataset. Big Data volumes are a very large and constantly increasing, its volume can be in terabytes to many petabytes. Variety refers to the various types of the data that can exist, for example, text, audio, still images and video. Big Data usually includes datasets with enormous volumes or sizes. It is very difficult for such systems to process this amount of data within the constraints of stipulated time frame which is required by the business intelligence processes.

1.1 Traditional Big Data Analytics Techniques

Big Data Analytics process is primarily based on statistical data analysis. However, when the data takes enormous volume, increases with tremendous velocity with diversified variety traditional statistical analytical methods are not effective. Hadoop platform is used for big data representation and storage. Various statistical methods can be implemented in R language for data analytics. Hadoop platform is very fast in processing big data as it is based on distributed computing. In this paper, the integrated environment RHIPE which is a blend of R and Hadoop is employed. Fig 1 shows the traditional big data analytics process.

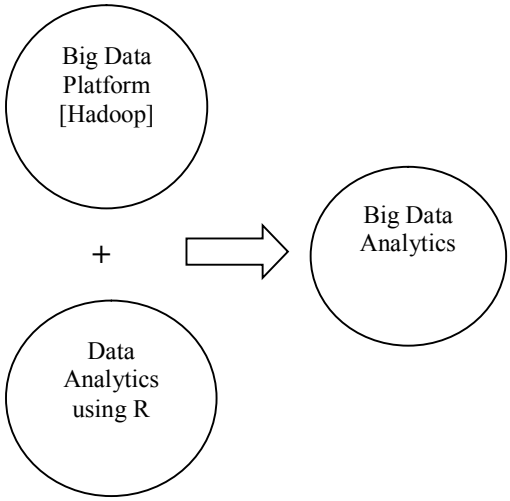

Fig. 1: Traditional Big Data Analytics process

2 Proposed Methodology

In this section, overall methodology related to our work is discussed. The main objective of the paper is the big data analytics using machine learning algorithms. Three types of data viz. Structured, semi-structured and unstructured data is considered. Structured data includes relational databases, semi-structured data includes XML, HTML files and unstructured data comprises of Word files, PDF files, and Text files. Media logs etc. The overall flow of our work is depicted in Figure 2.

i) Acquisition of Stock Exchange Data :

Stock exchange data is a type of big data as it is huge in volume and increasing with fast speed. It holds information about the shares being bought and sold. This information also includes decisions made about shares of the customers of different companies. It is set as an input to our research work, which is showing varying values of frequency of stock market changes.

ii) Representation of Big data in HDFS

Big data is stored effectively and efficiently retrieved by extended form of Google File System (GFS), the Hadoop Distributed File System (HDFS). It provides a distributed file system which is designed in such a way that it can run on commodity hardware. The input dataset are uploaded to the Hadoop directory. These input datasets are used by MapReduce nodes. The Hadoop Distributed File System (HDFS) will divide the input dataset into data splits and store them to Data Nodes in a cluster. Replication of data is taken into consideration for fault tolerance. Formation of bigger servers with heavy

configurations for handling large scale processing is quite expensive. However, as an alternative, many commodity computers can be tied together with single-CPU, as a single functional distributed system. The clustered machines can read the datasets in parallel fashion and provide a much higher throughput. In addition, it is cheaper than one high-end server. This is the motivation behind using Hadoop which runs across clustered and low-cost computers.

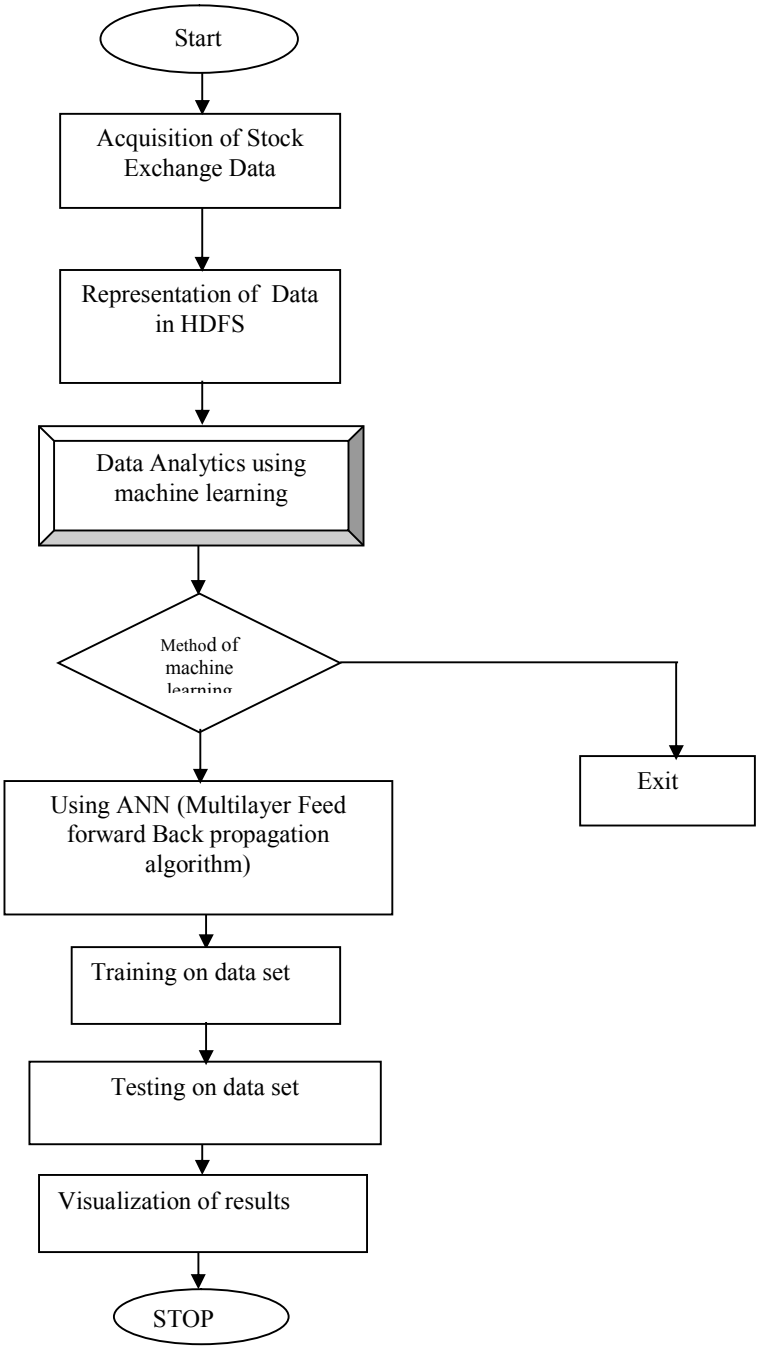

Figure 2: Flow-chart

The code runs across a cluster of computers on Hadoop platform. This process includes the following core tasks that Hadoop performs:

- The input data is divided into directories and files. Files are divided into uniform sized blocks of 128MB.
- Further processing is based on these blocks which are distributed across different cluster nodes. Blocks are replicated for handling hardware failure.

a) **Hadoop Framework** :

In big data, the unsupervised machine learning algorithms can be implemented through MapReduce technique. The data is processed by MapReduce on Hadoop clusters. The data analytics logic is translated to the MapReduce job. The Hadoop Framework is shown in Figure 3.

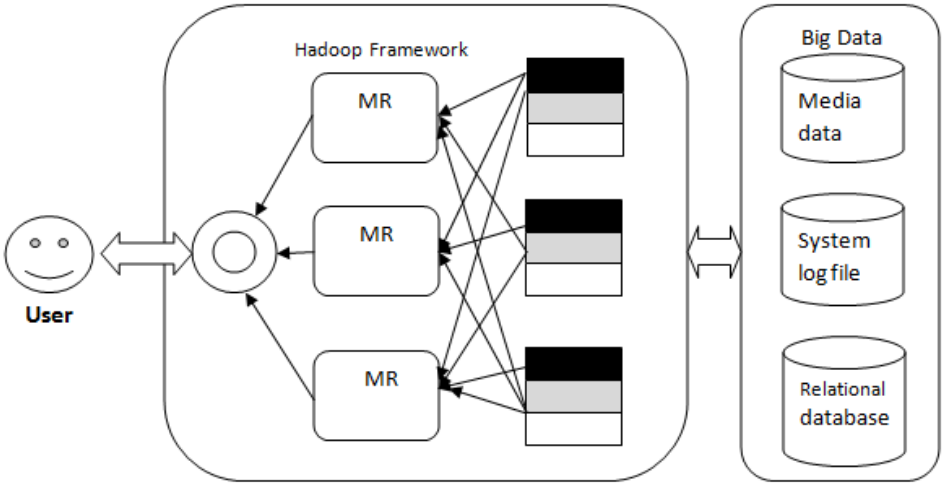

Figure 3 : Hadoop Framework

iii) Data Analytics using machine learning

When data is made available in the prescribed form for data analytics, data analytics operations using unsupervised machine learning such as Artificial Neural Network will be performed for discovering meaningful information from data. This enables to take better decisions towards business intelligence by using data mining techniques. The outcome of predictive analytics is used for decision making in business intelligence.

iv) Artificial Neural Network

We have used Artificial Neural Network (ANN) which is a unsupervised machine learning algorithm for data analytics. The focus is on extracting the hidden information, patterns and trends form the big data. The ANN is employed at each data node in Hadoop MapReduce environment. The vectors of data values [Open, High, Close, AdjClose, and Volume] are applied to the input layer as an input and Daily Returns are predicted as an output.

a) Experimental Setup

To demonstrate the working of big data analytics, we have used stock market data set as an input to our research work, which is showing varying values of frequency of stock market changes. Following table depicts the same.

Yahoo finance data for symbol BP

Date	Open	High	Low	Close	Volume	Adj Close
2013-08-23	41.16	41.54	41.11	41.51	4117400	41.51
2013-08-22	40.82	40.99	40.75	40.91	2808300	40.91
2013-08-21	40.84	40.89	40.51	40.53	4296800	40.53
2013-08-20	41.02	40.90	40.90	4354200	40.90	
2013-08-19	41.29	41.35	41.05	41.10	3633800	41.10

Change frequency calculation for Yahoo Finance data

Change	Frequency
-0.1	20
0.3	2
0.8	1
1.0	22
1.9	12

Rigorous experimentation is carried out on the stock exchange data using R language and the results are shown in Figure 4 and Figure 5 respectively.

v) Visualization of the output

Data visualization is used for displaying the output of data analytics. Artificial Neural Network is simulated using R language. The input vectors such as [High, Low, Close, AdjClose, and Volume] are applied and the predicted values of DailyReturns are shown as output in figure 4.

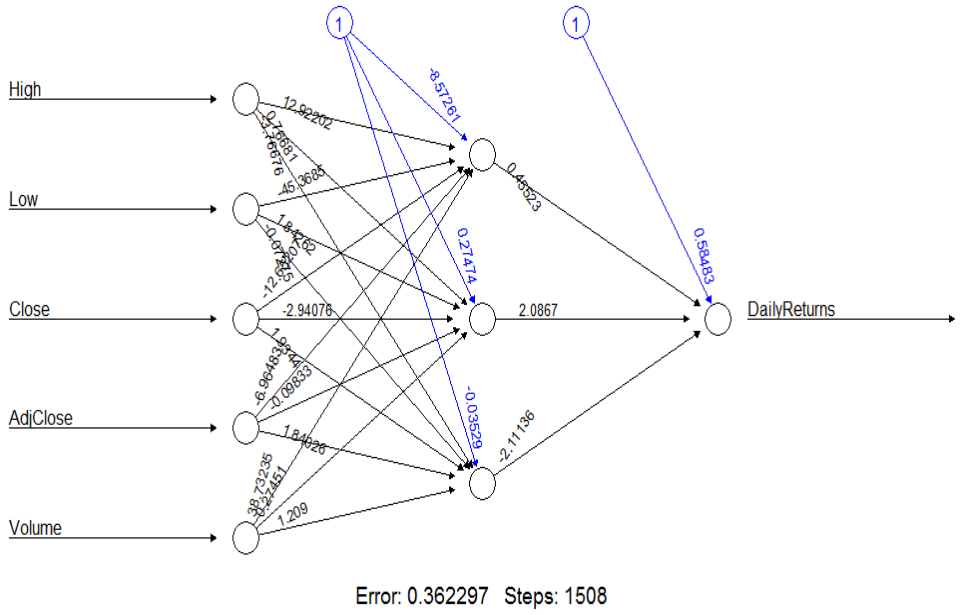

Figure 4 : A multilayer feed-forward neural network

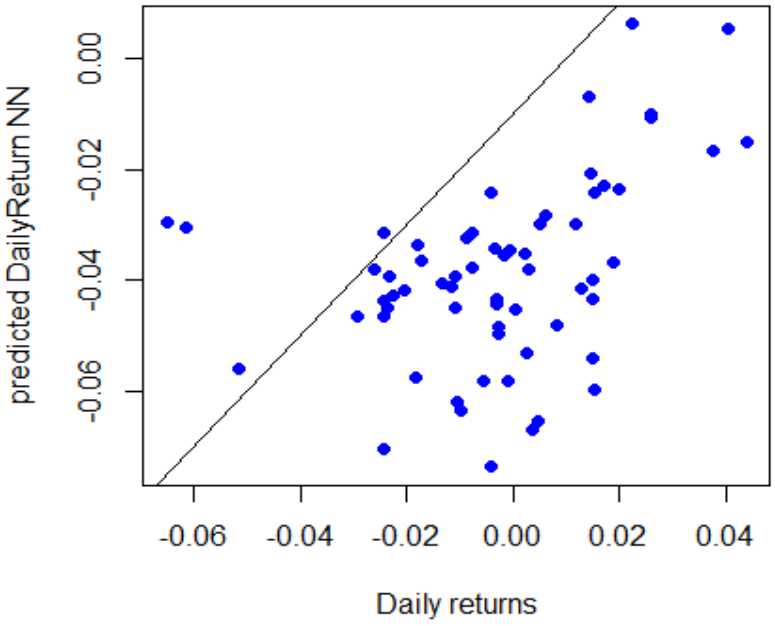

Figure 5 : Actual Daily Returns Vs predicted DailyReturns

As shown in Figure 5, the comparison of actual Daily Returns Vs predicted DailyReturns is shown in Figure 5.

Conclusion:

In this paper, we have presented the technique for big data analytics using unsupervised machine learning algorithm. In the first part, representation of big data in Hadoop Distributed File System (HDFS) is discussed. It is the requirement of data analytics that the data is required to be stored into HDFS. As the data is massive in size and volume Hadoop framework with Map Reduce is required for efficient and effective data processing and analytics. Artificial Neural Network (ANN) is used for data analytics as the data analytics process requires some kind of intelligence. This technique can be suitable for various types of big data analytics.

References

[1] Alexandra L'heureux , Katarina Grolinger, Hany F. Elyamany and Miriam A. M. Capretz , "Machine Learning With Big Data: Challenges and Approaches", IEEE Access March 2017,Vol. 5

[2] Althaf Rahaman.Sk, Sai Rajesh.K, .Girija Rani K, "Challenging tools on Research Issues in Big Data Analytics", International Journal of Engineering Development and Research, 2018, Volume 6, Issue 1

[3] Gianluca Bontempi, Yann-Ael Le Borgne, "Predictive modeling in a big data distributed setting: a scalable bias correction approach", IEEE International Congress on Big Data, 2016

[4] Mr. Shrikant Rangrao Kadam , Vijaykumar Patil**, "** Review on Big Data Security in Hadoop", International Research Journal of Engineering and Technology, 2017, Volume: 04 Issue: 01

[5] Tilwani Mashook, Patel Malay, Pooja Mehta, "Security and Privacy-A Big Concern in Big Data A Case Study on Tracking and Monitoring System", IJIRST 2017

[6] R.Kalaivani**, "**Security Perspectives on Deployment of Big Data using Cloud: A Survey", International Journal of Advanced Networking & Applications, (2017) Special Issue, Volume: 08, Issue: 05 Pages: 5-9

[7] Dr. Venkatesh Naganathan, " Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends", International Research Journal of Engineering and Technology. May-2018, Volume: 05 Issue: 05

[8] Junfei Qiu, Qihui Wu, Guoru Ding* , Yuhua Xu and Shuo Feng, "A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing 2016, Springer Open Access

[9] Nada Elgendy and Ahmed Elragal, "Big Data Analytics: A Literature Review Paper", Springer International Publishing Switzerland 2014. LNAI 8557, pp. 214–227

[10] Rahul Khokale, Ashwini Bonde, "Big Data Analytics Framework for Business Intelligence: A Review", International Journal of Research and Science Publication, Volume 01 Issue 01, Page No. 36 to 39