RESEARCHERID
THOMSON REUTERS
[INIT: January 2017]
ISSN 2348 – 8034
Impact Factor- 4.022
GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

USAGE OF MACHINE LEARNING AND HADOOP IN SOCIAL MEDIA ANALYTICS

Prof. K. Adisesha[*1] & Prof. Praveen Moses[2]
[*1]HOD, Computer Science Department , Bangalore City College, India
[2]HOD, Computer Science Department , Aditya College, India

ABSTRACT

The current trend in the information technology is data science. Various applications are getting benefit of data science tools and giving birth to new market insights so as to serve the customers in extensive way. In the current paper, we are focusing on the data science context along with machine learning (ML) and Hadoop frame work usage to handle bulk data. The importance of this work is to give the basis of data science and which are all the areas that can be benefited by data science implementation to leverage the capacity of decision making by the companies. Some of the example applications are click stream analysis and sentiment analysis in the platforms like face book, twitter and LinkedIn can be effectively analysed through machine learning usage. The implementation can be done using Python or R language. Hadoop usage is one more dimension in data science projects. The outcome of the paper is to describe the data science importance along with some machine learning algorithms. The other point is usage of R in data science in various applications. The Hadoop integration with R is interesting as R is performing its analysis in RAM which limits the capacity of processing bulk data but with Hadoop we can achieve the usage of bulk data processing with R. In the Machine Learning implementation, we are giving the basics of Mahout also

Keywords- *Social media,Machine learning (ML), Hadoop framework, Decision making, Data processing with R, Mahout*

I. INTRODUCTION

The advent of social media is causing bulk data processing and observing the interests of the users and recommending the users by knowing them is current trend in market. Here the scenarios like sports, health care; Amazon, Netflix and Large Hadron Collider are some of the examples of big data use cases. In sports while a player is entering into ground we can observe his average score, records and strike rate along with the advertisements which he was signed for a product. In health care domain the hospitals analyse the medical data and patient records so as to predict the rate of seeking the re admission into hospitals, similarly to predict the decease which may attack to the patients from their parents. Amazon is a huge market place for companies and retailers. Amazon is having 1.5 billion products and 152 million customer accounts which use big data to store the details. The Netflix is a place where the customers can view various movies which uses 1 petabytes of the storage as to stream the data. One Peta byte of average Mp3-encoded songs would require 2000 years to play. The experiments in the Large Hadron Collider produce about 15 petabytes of data per year, which are distributed over the world wide LHC computing Grid. One petabyte is enough to store the DNA of the entire population of the USA – with cloning it twice. All these scenarios are storing bulk data and required powerful algorithms so as to analyse and produce the strategic decisions and valuable suggestion to companies and customers. The organization of the paper is like section II consists of details about big data and data science basics and their applications. In section III the Machine Learning introduction and frequently used ML algorithms are mentioned. In section IV the usage of R programming to implement Machine Learning algorithms along with the importance of R is described. In section V need for integrating R and Hadoop is described.

II. BIG DATA AND DATA SCIENCE

The current IT industry buzzwords are big data and Data Science and in fact they are driving forces to many scenarios like sports, health care, Amazon and Netflix as above mentioned.

ISSN 2348 – 8034
Impact Factor- 4.022

Big data is referencing the high volume of the data in petabytes or more than that, variety of the data like structured (Schema-oriented), semi-structured (XML) and unstructured data(log files ,audio and video) and velocity of the data refers to frequent changes in the data. There are various challenges are there in big data we are listing some of them

- Data acquisition
- Information search and Analytics
- High volume of data
- High Velocity of processed data
- Information Storage
- Data security and privacy
- High variety of information
- High veracity of information

Big data is a concept where data in the algorithm processing itself becomes a problem and solution to that problem is Hadoop. Hadoop is a frame work of eco system which solves the big data problem with the help of distributed storage and parallel programming model.

Hadoop is the combination of Hadoop Distributed File System (HDFS) and Map Reduce (MR).
Hadoop is frame work for storing, processing and analysing the big data, allows distributed storage and distributed processing of large data sets across cluster of commodity computers using a simple programming model. Hadoop is apache open source frame work. Hadoop is based on master slave architecture which is the combination of five daemons, Name Node, Data Node, Secondary Name Node, Job Tracker and Task Tracker. The master system consists of Name Node, Job Tracker and Secondary Name node. The slave system consists of Data Node, Task Tracker. The other combination is HDFS (Name Node, Data Node and Secondary Name Node) which deals with storage logic and Map Reduce (Job Tracker and Task Tracker) which deals with Programming model. The storage logic is DFS which is not visible through commands only we can interact with that file system. The Map Reduce is depends on HDFS so as to process the logics.

Various UNIX distributions can be used to launch the Hadoop frame work some of the distributions are RHEL, Cent OS and Ubuntu. For Hadoop frame work implementation there are Apache,AWS, Clod eraHorton Works, MapR are available. All these distributions of Hadoop are serving for the same purpose of implementing HDFS and Map Reduce.

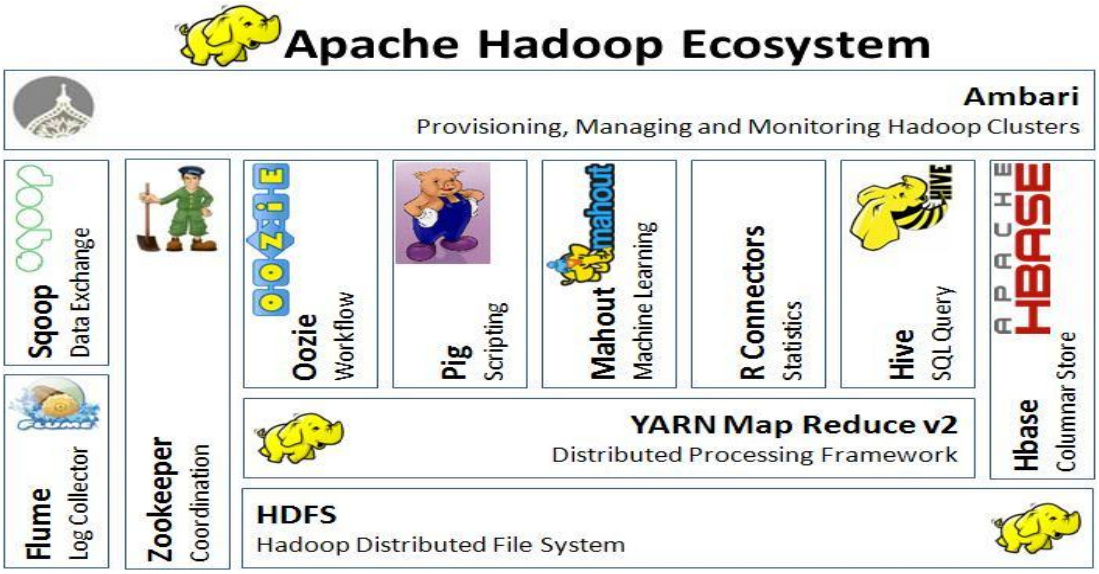

Figure1: Hadoop and Other Platforms in eco system.

ISSN 2348 – 8034
Impact Factor- 4.022

More data usually beats better algorithms, data science is the study of where information comes from ,what it represent and how it can be turned into a valuable resource in the creation of business and IT strategies. Various components are integrated to form the data science. They are

- Statistics
- Domain Expertise
- Data Engineering
- Advance Computing
- Visualization

Data science is the combination of Data architecture, Machine Learning and Analytics. The data scientists main functionality is develop new analytical methods ,respond to and resolve data mining performance issues and business case development, planning and collaboration with various vendors. The primary skills required are R(Python),SQL, Machine Learning, Statistics, Data Visualization and Communication.

1. Machine Learning

Machine learning is a class of algorithms which is data-driven.ML algorithms grow even more effective at your data scales in volume, velocity and variety, according to Mark Van Rijmenam the more data is processed , the better the algorithm will become. The scenarios like YouTube utilizes recommendation system to bring videos to a user that it believes the user will be interested in. In the biometrics the comparison of finger prints images data based on whorls, arches, and loops.ML categorizes the algorithms into supervised and unsupervised learning. If training data includes both the input and desired results then the algorithms are belongs to supervised learning examples are Naïve Bayes , Support Vector Machine ,RandomForest and Decision Trees. In case of unsupervised learning the model is not provided with the correct results during the training, the examples are K-Means, Fuzzy clustering and Hierarchical Clustering. Clustering is an algorithm which is frequently used in scenarios like telephony companies to identify the frequency of usage and establishing new towers as per the requirements, CISCO new office set up in California by observing about the employees residence which reduce the employees communication minimum, similarly new hospital construction based on the emergency treatment required for accident cases. The purpose of clustering is to identify and study internal structure of the data, summarization and identification of reoccurring patterns.

The process flow of supervised learning can be observed in the following manner.

- Observe the historical data
- Generate Random Sampling
- Derive Training and Test data set
- Apply ML on Training Data set
- Construct the Statistical Model
- Apply the prediction and Validation
- Test Accuracy of Model

The another ML algorithm most commonly used in big data scenarios is Random Forest which is an ensemble classifier made using many decision tree models, ensemble models combine results from different models. The result from an ensemble model is usually better than the result from one of the individual models.

Another important concept in data mining which is used by data scientists to identify the correlation between the various items is association rule mining. The association rule depends on 3 measures to output the conclusion support, confidence and Lift.

ISSN 2348 – 8034
Impact Factor- 4.022

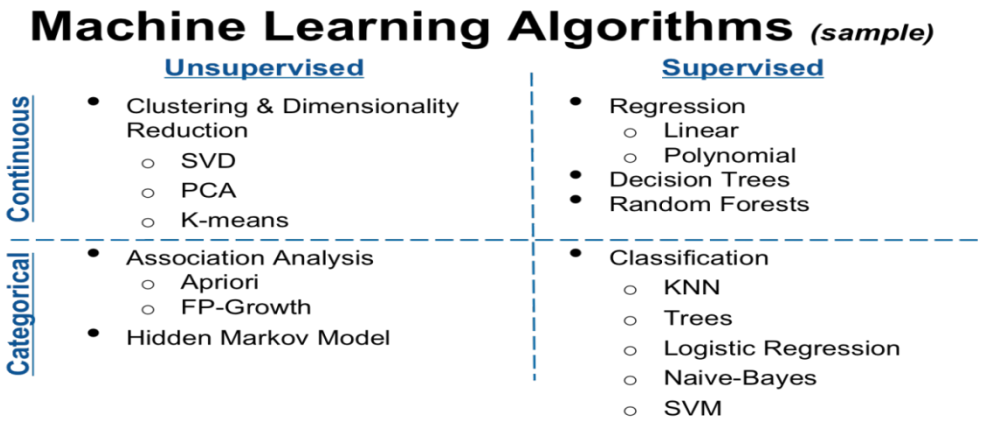

Figure 2: Machine Learning Algorithms

2. R Usage in Machine Learning

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. R has an effective data handling and storage facility, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display. In a typical data analysis R is very much helpful by considering the steps.
Import data, prepare, explore and clean data, Fit a statistical model, evaluate the model, Cross-validate the model, Evaluate model prediction on new data and produce report. R has the following features which are helpful in data processing and visualization

- Clearer, more compact code
- Less debugging
- Object-Oriented and Functional Programming
- Faster execution
- Easy to transform into parallel programming
- Open software available to Linux, windows and Mac

R allows import of the from various sources like CSV, Excel File, Text file, SPSS, SAS, Stata, XML and MATLAB.R supports various data types to handle the data like Numbers, Strings ,Logical, Factors and Data Frames. The statistical functions like min, max, mean, median and quantiles are supported by various plots are supported by R like Strip Charts, Histograms, Box Plots and Scatter plots. The advanced programming constructs like functions are supported by R in two flavours User-defined functions and pre-defined functions.

The following screenshot from RStudio shows the histogram of GDPs – there are 15 countries having more than 1,000 millions USD GDP; 1 country is in the range of 14,000 – 15,000 millions USD, 1 country is in the range of 7,000 – 8,000 millions USD and 1 country is in the range of 5,000 – 6,000 USD.

ISSN 2348 – 8034
Impact Factor- 4.022

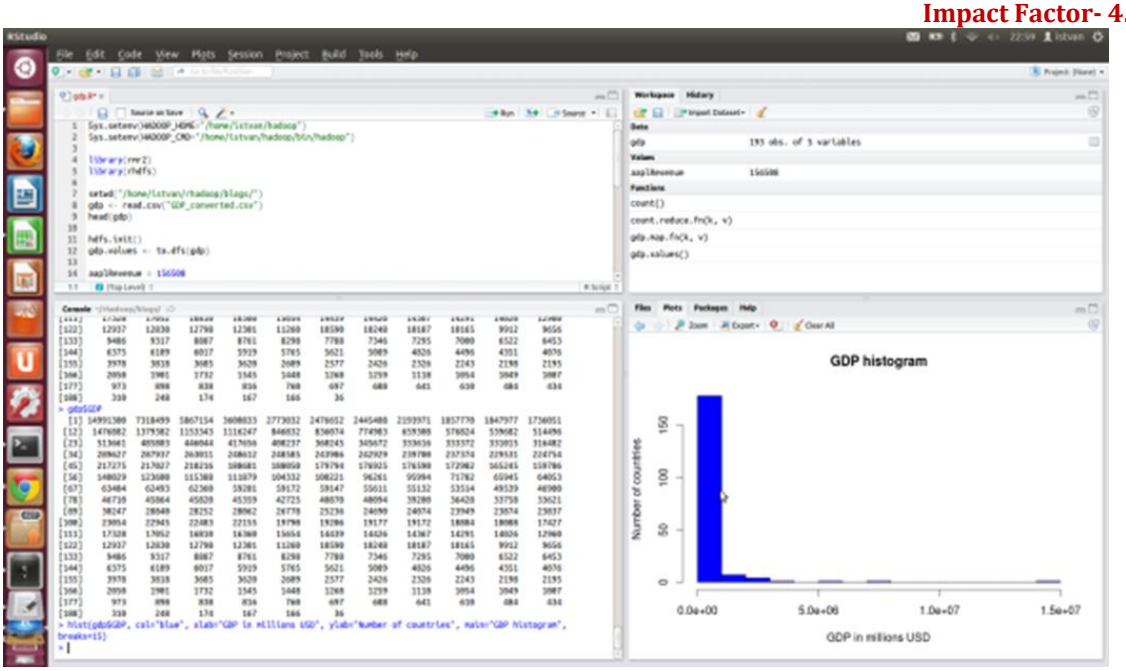

Figure 3: screenshot from RStudio shows the histogram of GDPs.

3. RHadoop

R is best solution for familiarity; Hadoop is best solution for capability. Hadoop comes into picture where the data is large and is out of capacity for the R memory doesn't scale on big data clusters. In R integrated Hadoop, R analyses the data within the node in which it resides, rather than moving it somewhere else to be analysed, which makes R-based data analysis quick.R can leverage Hadoop processing with R and Hadoop Integrated Processing Environment (RHIPE), Using Hadoop Streaming, Using RHadoop Package and ORCH (Used to develop Map Reduce Jobs from R).

Hadoop eco system includes

- HDFS
- Map Reduce
- Hive
- Pig
- Sqoop
- Flume
- HBase
- Oozie
- Zookeeper
- Mahout
- Ambari

R components with Hadoop eco system are

- rhdfs
- rhbase
- rmr2
- plymr

ISSN 2348 – 8034
Impact Factor- 4.022

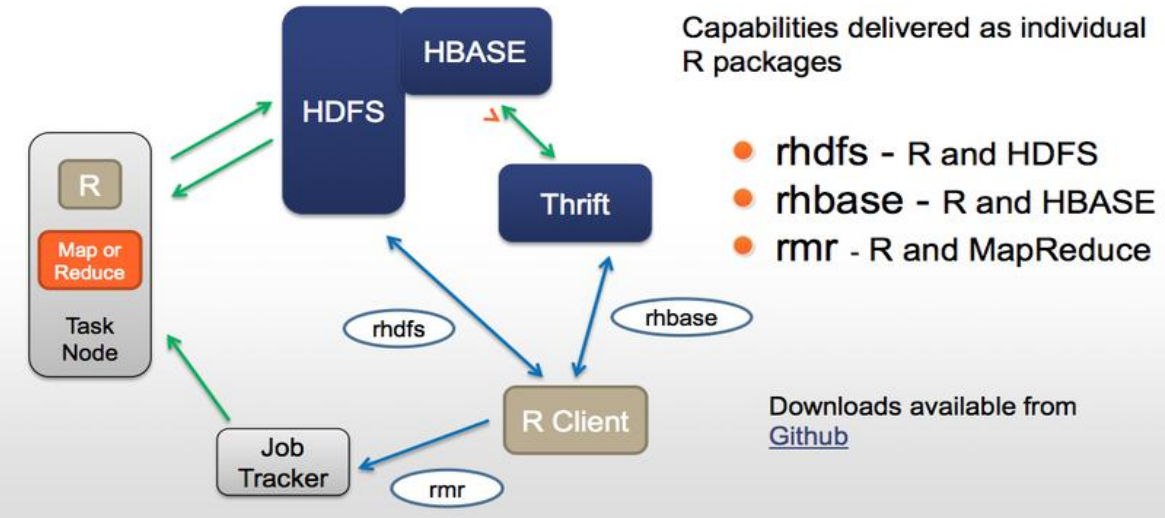

Figure 4:R and Hadoop Integration abstract View.

III. DISCUSSION

This paper describes the initial points of Hadoop and R along with ML algorithms. Hadoop is mainly for storing bulk data whereas R is for data analysis. The ML algorithms will learn from user given data and ML helps in recommendation systems, classification and predictions. R is having rich set of functionalities to process ML. The problem with Hadoop is data analysis is not up to the mark but the frame work is having the capability of parallel programming and distributed storage based on commodity hard ware. R is good at data analysis but not capable of storing bulk data so the combination of R and Hadoop gives the capability of storing the bulk data and analysis of data.

IV. CONCLUSION

In the paper we have described the initial points on usage of Hadoop and R along with ML algorithms in Social Media Analytics. Hadoop is mainly for storing bulk data whereas R is for data analysis. The ML algorithms will learn from user given data and ML helps in recommendation systems, classification and predictions.We trust the article will helps the beginners who want to know basics of Hadoop and data science, ML and R, as an extension of this paper we are trying to get another paper so as to explain some advance aspects of ML and other tools of Hadoop and additional programming aspects of R.

REFERENCES

1. *S. Lohr, "The age of big data," N. Y. Times, vol. 11, 2012.*
2. *S. Madden, "From Databases to Big Data.," IEEE Internet Comput., vol. 16, no. 3, 2012.*
3. *P. Zikopoulos, C. Eaton, and others, Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.*
4. *A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, "Big data," Manag. Revolut. Harv. Bus Rev, vol. 90, no. 10, pp. 61–67, 2012.*
5. *R. Appuswamy, C. Gkantsidis, D. Narayanan, O. Hodson, and A. Rowstron, "Scale-up vs Scale-out for Hadoop: Time to rethink?," in Proceedings of the 4th annual Symposium on Cloud Computing, 2013, p. 20.*
6. *A. S. Tanenbaum and M. Van Steen, Distributed systems. Prentice-Hall, 2007.*
7. *C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Inf. Sci., vol. 275, pp. 314–347, 2014.*

8. *I. Mashal, O. Alsaryrah, and T.-Y. Chung, "Performance evaluation of recommendation algorithms on Internet of Things services," Phys. Stat. Mech. Its Appl., vol. 451, pp. 646–656, 2016.*
9. *www.google.com*
10. *www.cloudera.com*