Inclusion of Pre-Processing and Time Series Algorithms in Map Reduce Environment using Big Data Analytics

P. Nagaraj, P. Deepalakshmi

Abstract: ***Map Reduce is one of the most effective ways of handling Big Data. Many of existing Data Mining / AI algorithms was developed in Map Reduce to provide effective results. There are many more algorithms including preprocessing algorithms such as Binarization, Normalization etc., Time series algorithms such as Moving average, Sliding Window, Correlation etc., which are not yet implemented in Map Reduce. Although there are not major algorithms they play a vital role in preprocessing and processing chunk data to a meaningful data. In this paper, we proposed a model of including these algorithms in Map Reduce to improve preprocessing outcome of Big Data much faster. The processed data can then be trained by the regression algorithms using Machine learning techniques to preprocess the huge data in a long run automatically.***

Keywords: ***Big Data, Map Reduce, Data Mining, Pre-processing, Time series, Regression.***

I. INTRODUCTION

Big Data is frequently described by 3Vs: the extraordinary Volume of information, the wide Variety of information types and the Velocity at which the information must be prepared. Albeit huge information doesn't compare to a particular volume of information, the searching data is regularly used to depict terabytes, petabytes and even exabytes of information caught after some time. Information may likewise exist in a wide assortment of record types, including organized information, for example, SQL database stores; unstructured information, for example, report documents; or spilling information from sensors. Further, enormous information may include various concurrent information sources, which may not generally be incorporated [1].

Pre-Processing is a technique that forms its information to deliver yield that is utilized as corrected input to another algorithmic phases [2]. The yield is said to be a preprocessed type of the information, which is frequently utilized by some consequent projects like compilers.

Time series is a progression of information focuses recorded (or diagramed) in time request. Most usually, a period arrangement is a grouping taken at progressive similarly separated focuses in time. Accordingly it is an arrangement of discrete-time information. Instances of time arrangement are statures of sea tides, tallies of sunspots, and the day by day closing estimation of the industrial average. Time arrangements are in all respects regularly plotted by means of line diagrams. Time arrangement utilized in insights, signal handling, design acknowledgment, econometrics, Mathematical Finance, Weather estimating, Intelligent Transport and direction anticipating, quake expectation, Electroencephalography, Astronomy, Control Engineering, correspondences building, and to a great extent in any area of connected science and designing which includes transient estimations.

Regression Analysis is a measurable procedure for evaluating the connections among factors. It incorporates numerous strategies for displaying and examining a few factors, when the attention is on the connection between a ward variable and at least one autonomous factor. More explicitly, Regression Analysis causes one see how the normal estimation of the needy variable changes when any of the free factors is differed, while the other autonomous factors are held fixed.

II. RELATED WORK

Our primary focus is on utilizing Map Reduce concept to achieve scalability. Apache Hadoop (YARN) provides an effective implementation of Map Reduce approach. Thus we are preceding the implementation in Apache Hadoop. Machine learning often involves processing of data in large amount i.e., Big Data. Thus while running high-level machine learning programs it is often necessary to increase the effectiveness of running the algorithm [1].

As mentioned earlier the high-level machine learning process involves many levels of pre processing and post processing processes that process, extract or convert the raw data or the output of machine learning program which will be in machine readable format into readable format. These algorithms often run in a sequential manner which increases the time complexity of the overall process even though the high-level machine learning programs are running in scalable manner[2].

Apache Hadoop being a well known platform for processing the information parallely in a distributed environment, numerous Data mining,

Revised Manuscript Received on December 16, 2019.

***** Correspondence Author

P.Nagaraj*, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, India. Email: nagaraj.p@klu.ac.in

P.Deepalakshmi, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, India. Email: deepa.kumar@klu.ac.in

Time Series, Preprocessing algorithms are not executed towards Hadoop [3]. In this work we build up the Data mining, Time Series, Preprocessing algorithms to Hadoop. When these techniques are included Hadoop performance can be improved. These upgrades will quicken the presentation of the various techniques in Hadoop and it will attract-in more similar techniques also be moved to Hadoop platform. It results in a platform where big data can be trained and processed more efficiently rather than applying all techniques separately in a conventional way.

III. PROPOSED SYSTEM

Map Reduce is one of the effective ways of handling chunk data [Big Data]. Majority of the Artificial Intelligence and Data Mining algorithms used in the industry for various purposes are implemented in Map Reduce in order for them to archive scalability.

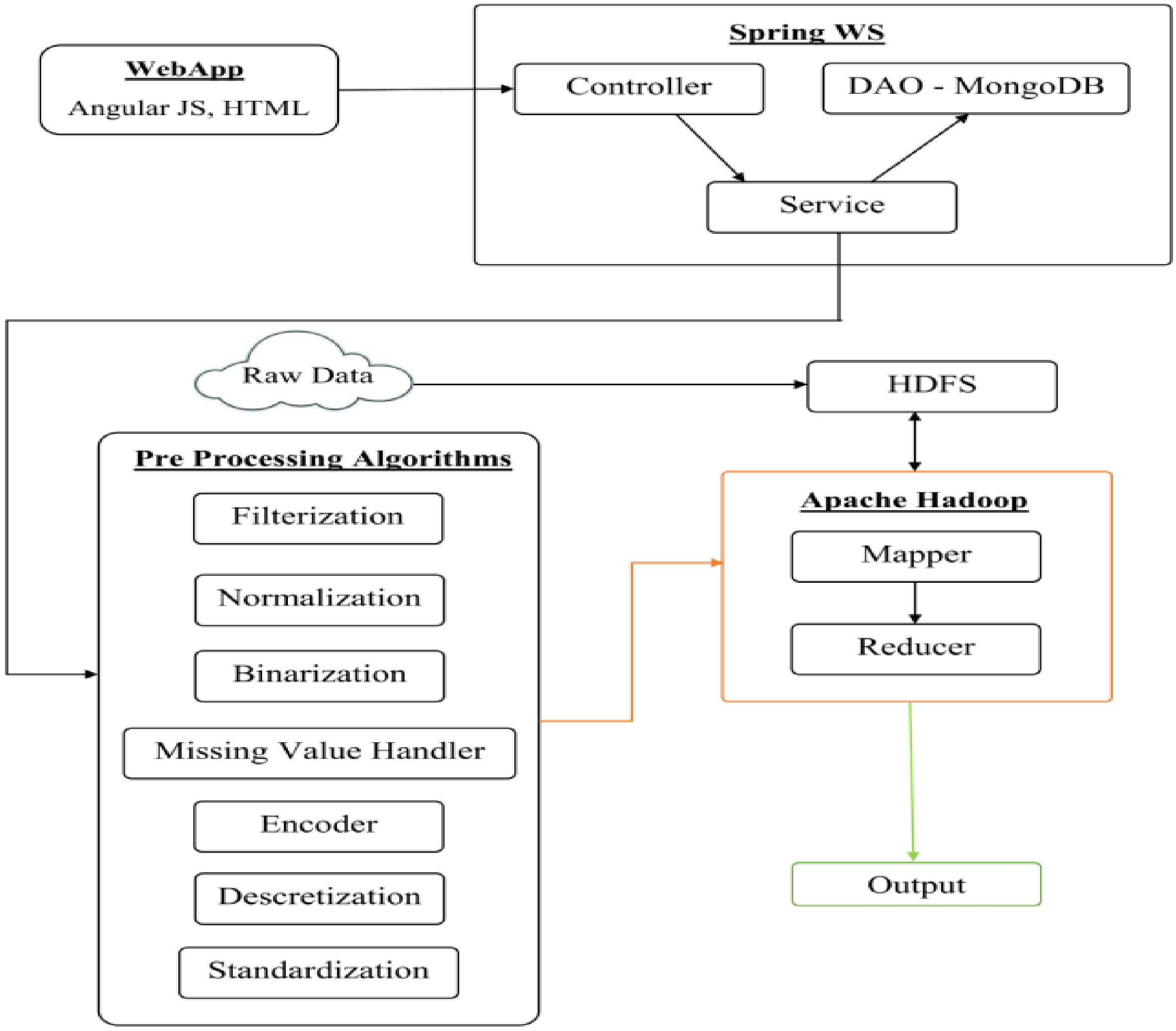

Fig.1 System Architecture

Although time complexity is reduced drastically by running these algorithms in Map Reduce, it also involves multiple smaller level processes that are necessary to convert the raw data into meaningful data, inner process transformation etc.., the proposed system involves implementing these data processing algorithms in Map Reduce.

Each algorithm has its effect in different range in different datasets. No algorithm is best suited for all datasets. To collect the real time data, a user friendly WebApp can be implemented with AngularJS on the front end, Spring-WS as the middle tier, Mango DB for saving information and Apache Hadoop for Map Reduce[4]. The data from the user is received by the GUI, and the received values are passed to the middle tier. The middle tier will convert the received values to the Hadoop understanding queries and based on the algorithm the processing is done.

Map reduce techniques are used to handle the Big Data to map and reduce the values in distributed network. These algorithms communicate with the cluster and take decision of mapping the data with the pair values. Regression techniques are applied to train the system in an unsupervised manner. Machine learning algorithms like regression logistics are used to train the system in a supervised manner. For these the data sets are partioned into instruction set and execution set. While processing the data list of algorithms are employed as follows:

A. *Pre-processing Algorithms*

1. Missing Value Handler: This Algorithm is used to replace the NULL values for better execution process.

There are 4 kinds of process such as,

a) Filter: Removing the Row containing the NULL values.
b) Mean: Replacing the NULL values by taking Mean for the current column values.
c) Median: Replacing the NULL values by taking Median for the current column values. It is described in equation.(1)

Median= ((array. get ((array_size/2)-2)) + (array. get ((array_size/2)-1)/2)); --- (1)

d) Mode: Replacing the NULL values by taking Mode for the current column values

2. Filterization: This category of algorithm will filter the selective random fields from the cluster of fields. This is used to view the desired field instead of viewing the full fields.
3. Binarization: This category of algorithm will binarize the values. For binarizing the threshold value is assigned, and if the value is lesser than threshold, the value is changed to 0; else the value is changed to 1.
4. Category Encoder: This category of an algorithm will encode the values in the file to the user provided value for security purpose. It is done for both encoding and decoding process.

B. *Time Series Algorithms*

1. Exponential Moving Average: An exponential moving average (EMA) is a type of moving average that is similar to a simple moving average, except that more weight is given to the latest data. In general, to track the price fluctuations faster exponentially weighted moving average method is used.

The graph in Fig.2 shows the exponential moving average time series algorithm. The slope indicates the average moving of faster data samples.

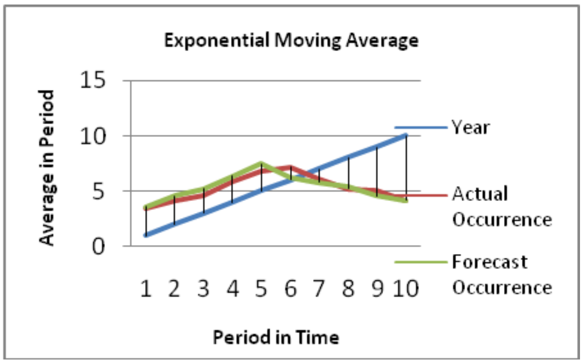

Fig.2 Exponential Moving Average

This focuses on trends that are caused by the most recent data (more weightage). A general equation for smoothing the exponential factor is defined in equation. (2)

$$S_t = \alpha \bullet \text{У}_{t-1} + (1 - c) \bullet S_{t-1} \quad \text{--- (2)}$$

where c is the smoothing factor, and $0 < c < 1$. The smoothed statistic S_t is a simple weighted average of the previous observation У_{t-1} and the previous smoothed statistic $S_{t-1.}$

2. Simple Moving Average (SMA): Simple Arithmetic Moving average can be calculated by adding the actual occurrence of the cost and closing rate of the share for a specific period of time to the number of periods of time in total. It is formed by computing the average price for certain limit. Most moving averages are based on closing prices. General equation for exponential smoothing is given as in equation(3).

$$S_t = (Xt_{-1} + X_{t-2} + \ldots + X_{t-n})/ N \quad \text{--- (3)}$$

where S_t is forecast for the coming period, X_{t-1} Actual occurrence in the past period for up to “n” periods and N is the no of periods to be averaged.

The graph in Fig.3 shows the simple moving average of time series algorithm. The slope indicates the average time series analysis of repeated data samples.

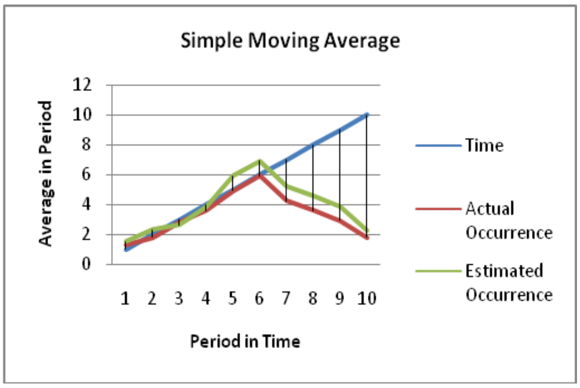

Fig.3 Simple Moving Average

C. *Regression Techniques*

1. Linear Regression: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe data. Explanatory variable and dependant variables are the two variables to be considered. In other words it is a mathematical technique for finding the straight line that best-fits the values of the linear function, plotted on a scatter graph as data points. If a “best-fit" line is found, it can be used as a basis for estimating the future values of the function by extending it while maintaining its slope.
2. Logistic Regression: Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. A dichotomous variable is used to observe the result which can possibly accept only two values.

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.,). It can be described in equation (4).

Logit (p) =1/ (1+ ($b_0X_0+b_1X_1+b_2X_2+b_3X_3+\ldots+b_k X_k$)) -- (4)

where p is the probability of presence of the characteristic of interest.

The graph in Fig.4 shows the linear and non-linear representation of the performance of linear regression and logistic regression over a period of time t. Here X axis indicates the threshold of mapping and Y axis indicates time.

Linear regression models are used to show or predict the relationship between two variables or factors. We are using a straight line method to get the objective for pre-processing. We can use the same equation to predict the points at different values of x which result in a straight line. The factor that is being predicted is called the dependent variable.

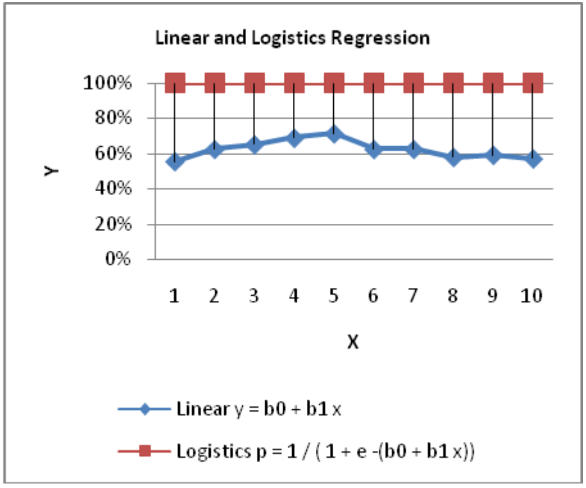

Fig.4 *Linear and Logistic Regression*

Logistic regression models are used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Regression analysis will furnish with a condition for a diagram so we can make forecasts about our information.

IV. CONCLUSION AND FUTURE WORK

Although time complexity is reduced drastically by running time these algorithms in Map Reduce, it also involves multiple smaller level processes that are necessary to convert the raw data into meaningful data, inner process transformation etc., Existing systems uses Conventional implementation of these algorithms that takes up time since it involves Big Data. This system involves implementing these data processing algorithms in Map Reduce which results in better performance in both training as well as in predicting the data pattern. Numerous stages for enormous scale preparing have attempted to face the dangerous of Big Data in last year's. These stages attempt to bring nearer the bring advances to the standard client by concealing the specialized nuances got from appropriated situations. We utilized the Hadoop stage to expand the adaptability as far as use.

We compared our results in terms of Pre-processing algorithms and time series algorithm to get the predicted output. In resultant graph we used regression techniques for change point analysis.

REFERENCES

1. Nandimath, J., Banerjee, E., Patil, A., Kakade, P., Vaidya, S., & Chaturvedi, D. (2013, August). Big data analysis using Apache Hadoop. In 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI) (pp. 700-703). IEEE.
2. Manikandan, S. G., & Ravi, S. (2014, October). Big data analysis using Apache Hadoop. In 2014 International Conference on IT Convergence and Security (ICITCS) (pp. 1-4). IEEE.
3. Afzali, M., Singh, N., & Kumar, S. (2016, March). Hadoop-MapReduce: A platform for mining large datasets. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1856-1860). IEEE.
4. Kumar, P., & Rathore, D. V. S. (2014). Efficient capabilities of processing of big data using hadoop map reduce. International Journal of Advanced Research in Computer and Communication Engineering, 3(6), 4421-4425.
5. Nandakumar, A. N., & Yambem, N. (2014). A survey on data mining algorithms on apache hadoop platform. International Journal of Emerging Technology and Advanced Engineering, 4(1), 563-565.
6. Shafer, J., Rixner, S., & Cox, A. L. (2010, March). The hadoop distributed filesystem: Balancing portability and performance. In 2010 IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS) (pp. 122-133). IEEE.
7. Bendre, M., & Manthalkar, R. (2019). Time series decomposition and predictive analytics using MapReduce framework. *Expert Systems with Applications*, *116*, 108-120.
8. Schörgenhumer, A., Kahlhofer, M., Chalupar, P., Grünbacher, P., & Mössenböck, H. (2019, January). A Framework for Preprocessing Multivariate, Topology-Aware Time Series and Event Data in a Multi-System Environment. In *2019 IEEE 19th International Symposium on High Assurance Systems Engineering (HASE)* (pp. 115-122). IEEE.
9. Zhu, S., Qiu, X., Yin, Y., Fang, M., Liu, X., Zhao, X., & Shi, Y. (2019). Two-step-hybrid model based on data preprocessing and intelligent optimization algorithms (CS and GWO) for NO2 and SO2 forecasting. *Atmospheric Pollution Research*.
10. Smith, V., Portillo-Quintero, C., Sanchez-Azofeifa, A., & Hernandez-Stefanoni, J. L. (2019). Assessing the accuracy of detected breaks in Landsat time series as predictors of small scale deforestation in tropical dry forests of Mexico and Costa Rica. *Remote sensing of environment*, *221*, 707-721.
11. Friedman, J. (2018). *An algorithm for finding best matches in logarithmic expected time* (No. SLAC-PUB-1549). SLAC National Accelerator Lab., Menlo Park, CA (United States).
12. Bullock, E. L., Woodcock, C. E., & Holden, C. E. (2019). Improved change monitoring using an ensemble of time series algorithms. *Remote Sensing of Environment*, 111165.
13. Ramot, M., & Gonzalez-Castillo, J. (2019). A framework for offline evaluation and optimization of real-time algorithms for use in neurofeedback, demonstrated on an instantaneous proxy for correlations. *NeuroImage*, *188*, 322-334.
14. Liu, C., Zhang, Q., Luo, H., Qi, S., Tao, S., Xu, H., & Yao, Y. (2019). An efficient approach to capture continuous impervious surface dynamics using spatial-temporal rules and dense Landsat time series stacks. *Remote Sensing of Environment*, *229*, 114-132.

AUTHORS PROFILE

Nagaraj. P is currently working as an Assistant Professor in the department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu. He is pursuing his PhD in the area of Health Care Recommender System using Data Analytics.

Deepalakshmi. P is currently working as a Professor in Department of Computer Science and Engineering at Kalasalingam Academy of Research and Education (KARE), Virudhunagar, Tamilnadu, India. She is also serving as Dean, School of Computing. Her research interest includes Optimization Techniques, Distributed Computing, Network Security, Data Analytics, machine Learning Techniques. She also takes care of KARE ACM student chapter as faculty mentor.