MACHINE LEARNING ALGORITHMS IMPLEMENTED ON HADOOP MAP REDUCE

[1] Nikhil Salvithal, [2] Shriniwas Darshane, [3]Kanchan Chouhan, [4]Dipanjali Chavare

[1]Assistant Professor, [2] Assistant Professor, [3]Research Assistant, [4]Research Assistant
[1234] Computer Science and Engineering,
[1234]SVERI's College of Engineering pandharpur, Pandharpur, India.

Abstract: It isn't easy to analyze huge records. This requires system based structures and technologies for you to procedure. Map -reduce a distributed parallel programming model runs on hadoop surroundings, approaches massive volumes of information. A parallel programming technique can be relevant to the linear regression algorithm and support vector machines algorithm from the system getting to know network to parallelize acceleration on the multicore system for efficient timing efficiency.

***IndexTerms* -** MapReduce; Machine Learning; Machine Learning Algorithms,Data Analytics, Support vector machines, Hadoop, Linear Regression.

I. INTRODUCTION

Machine Learning strategies are fantastically scalable for information analytics. In this paper, an analysis of linear regression and support vector machines algorithms are proposed for enforcing on hadoop for improving the timing performance. The size of informational indexes being gathered and broke down in the business for business intelligence is developing quickly, making conventional warehousing restrictively expensive.

Hadoop is a well-known open-source map-lessen usage which is being utilized as a choice to store and procedure amazingly enormous informational collections on ware equipment. Nonetheless, the guide decrease programming model is low level and requires engineers to compose custom projects which are difficult to keep up and reuse. Large Data examination is rising as a methods factor for accomplishment in various territories, for example, logical, building, research, business, and government exercises.

- Hadoop:

To save statistics, hadoop makes use of its very own dispensed file system, HDFS, which makes information to be had to the hadoop distributed file system (HDFS) is a disbursed report machine designed to run on commodity hardware. It has many similarities with current distributed record systems. But, the variations from other distributed file structures are huge.

Hadoop another arrangement streamlined for in-memory preparing, convey equal handling on immense datasets and effectively bond and between relate with current large information stockrooms in the dispersed capacity environment.

A Hadoop framework keeps up two distinct parts: a MapReduce structure and a circulated document framework. Hadoop is one of the huge information apparatuses that can rapidly process stores of gigantic, muddled unstructured and organized information. Hadoop incorporates a flaw tolerant capacity framework called the Hadoop Dispersed File System. HDFS can store immense measures of in- the arrangement, scale up gradually and endure the disappointment of critical portions of the capacity foundation without losing information.

- HDFS:

HDFS is exceptionally fault-tolerant and is designed to be deployed on low-price hardware. HDFS gives high throughputs get right of entry to application data and is suitable for programs which have massive statistics sets. The time period map-reduce honestly refers to 2 separate and distinct duties that hadoop packages perform.

II. LITERATURE SURVEY

- ➢ Existing System :

Centralized servers: The existing framework contains unified servers for putting away the information, which is not fit for preparing enormous measures of information. Conventional frameworks devour heaps of time while preparing the huge datasets. The existing framework is restrictively costly.

- ➢ Problem Definition:

The centralized servers are used to store and retrieve the data which are not able to handle the huge amount of data and creates a bottleneck while processing multiple files simultaneously.

➢ Proposed System:

We are going to provide solution to address this issue through the use of map-reduce and machine learning algorithms.We have consider hadoop, map reduce and machine learning algorithms for implementation. Also as a responsibility we are going to provide the information to the users for the use in research and education fields.

III. METHODOLOGY

- Modular Design:

The functionality of the proposed application is divided into number of sub modules. The modules to be taken into account are Hadoop map reduce implementation and implementation of machine learning algorithms. These modules while integrated together give the functionality desired out of the application.

- Implementation of Hadoop map reduce:

In this module the data is stored and processed. This is useful to store and retrieve the huge amount of data. We have used hadoop 2.6.5 for the implementation purpose. In the implementation of hadoop we have created three nodes, one master node and two slave nodes. Hadoop is a popular open-source map-reduce implementation which is being used as an alternative to store and process extremely large data sets on commodity hardware. Map reduce divides the task into the key-value pair. It uses intermediate splitting of the data and performs the map on it. After map function reduce is executed.
Reduce is always performed after the map function. The map-reduce programming model is very low level and requires developers to write custom programs which are hard to maintain and reuse. Map Reduce is a framework which is used for making applications that help us with processing of huge volume of data on a large cluster of commodity hardware. Map Reduce will divide the task into small parts and process each part independently by assigning them to different systems. After all the parts are processed and analyzed, the output of each computer is collected in one single location and then an output dataset is prepared for the given problem.

- Implementation of Machine Learning algorithm:

Different machine learning algorithms such as SVM, KNN, K-means etc. are used. Machine learning algorithms are collected and implemented on the top of map reduces to increase the processing efficiency. This module intends to find the performance variations in time with an application of supervised and unsupervised machine learning techniques and algorithms and algorithms are productive for data analytics.
To potentially speed up the processing, a unified way of machine learning is applied on MapReduce frame work. A broadly applicable programming model MapReduce is applied on different learning algorithms belonging to machine learning family for all business decisions. By using ML algorithms with Hadoop for better storage distribution will improve the time and processing speed.
The details of design and implementation is discussed in the successive sec-

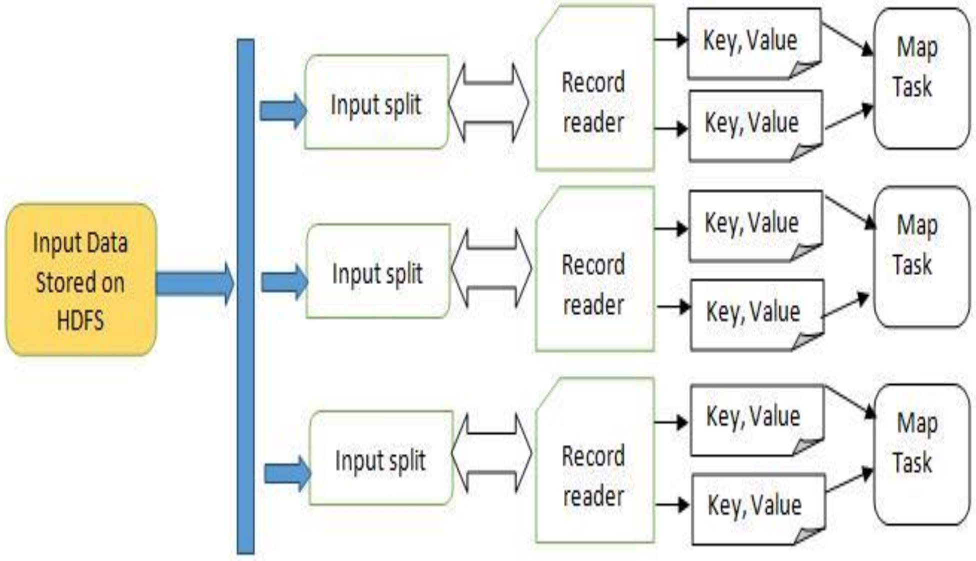
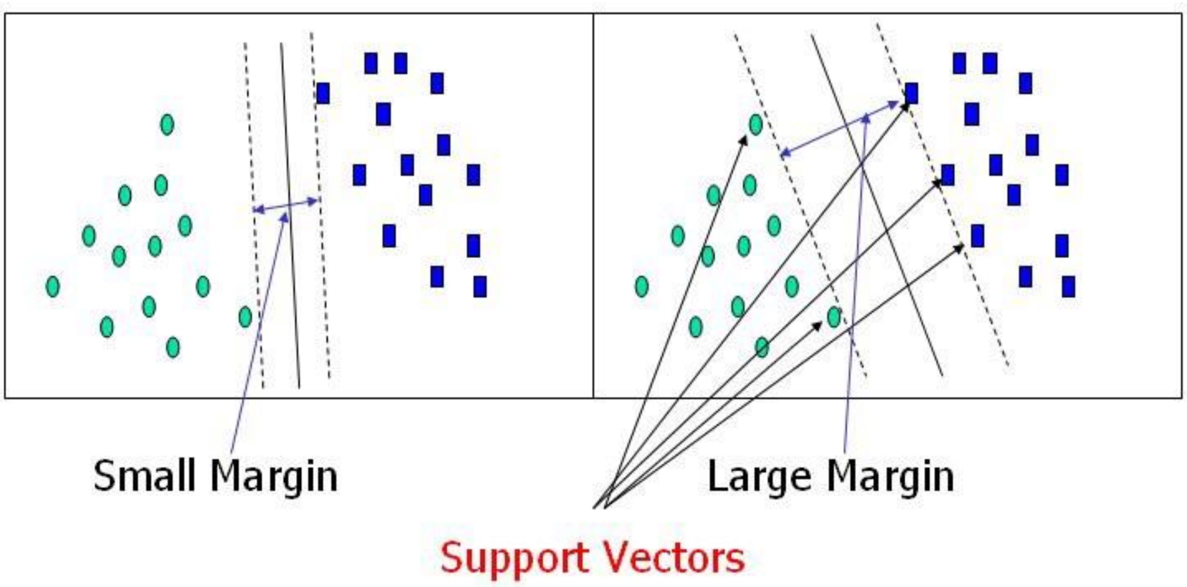


Fig. 1: Map reduce Implementation

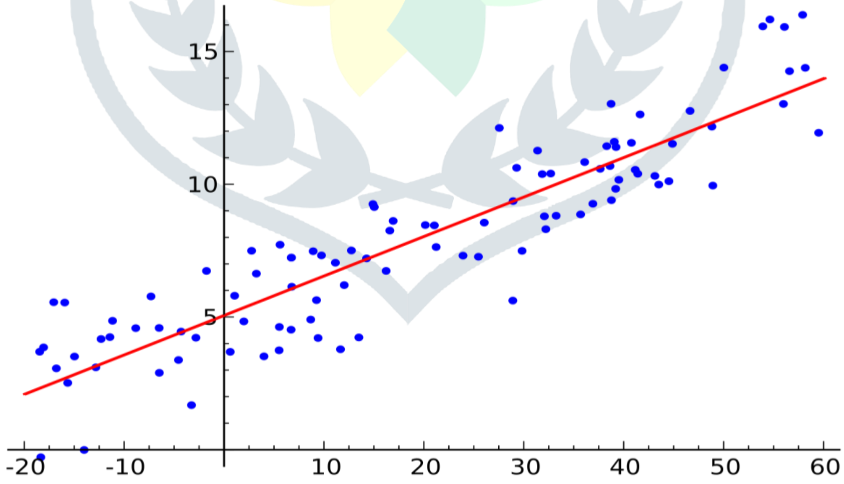
- Support vector machine:

The objective of the support vector machine algorithm is to find a hyper plane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.



- Linear regression :

Method used for the modeling and analysis of numerical data. Regression exploits the connection among or more variables so, that may benefit data approximately one in all them through understanding values of the different. Regression can be used for prediction, estimation, and hypothesis, trying out, and modeling causal relationships.
There may be a linear dating between the unbiased(x) and established(y) variable. The purple line inside the above graph is known as the quality match straight line. Primarily based on the given information points, we strive to devise a line that fashions the factors the first-rate. The road can be modeled primarily based on the linear equation proven underneath. Y = a_0 + a_1 * x ## linear equation



- **Cost Function**

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

IV. MACHINE LEARNING ALGORITHMS:

- K-means :

In centroid based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the K cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. Unsupervised Learning: It models the input into a set of representation produced after the algorithm. This works for finding a new representation of the input data using association learning rules. Association learning rules are procedures that extort system that best explains investigational relationships among variables in data. Grouping of similar data objects in to same cluster and dissimilar objects into divergent groups is referred as clustering.

- **KNN:**

The proposed K-Nearest Neighbor Algorithm is applied for enhancing the unstructured big data analysis. Learning model based on Instance or examples of training data that are considered vital or essential to the model are characteristically built up a database of example data.

They associate new data to the database by utilizing a resemblance measure for finding the ideal match and make a forecast. This is the exact reason why instance based techniques are also called winner-take all methods and memory-based learning.

In order to approximate constant variables, the KNearest Neighbor Algorithm is used. Weighted average of the k nearest neighbors is used by one such algorithm which is weighted by the counter of their distance. KNN Classifier is an instance-based learning algorithm that is based on a distance function for pairs of observations, such as the Euclidean distance and the Manhattan distance.

The basic idea of KNN is very simple. In the classification paradigm, k nearest neighbors of a test sample are retrieved first. Then the similarities between the test sample and the k nearest neighbors are aggregated according to the class of the neighbors, and the test sample is assigned to the most similar class. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct.

A good k can be selected by various heuristic techniques, for example, cross validation. The special case where the class is predicted to be the class of the closest training sample (i.e., when k = 1) is called the nearest neighbor algorithm.

V. CONCLUSION

Executing the map reduces jobs using Hadoop and machine learning will show best results for optimal time efficiency. The actual outcome depends on the final implementation of the model with the data collected from data analytics system.

REFERENCES

I. Alan F. Gates, Building a High Level Dataflow System on top of MapReduce: The Pig Experience In *Proceedings of International Conference on VLDB '09 by ACM in France* 2009.

II. Hadoop: Open-source implementation of MapReduce.

III. Ashish Thusoo, Joy deep Sen Sharma HIVE – A warehousing solution over a Map – Reduce framework *In Proceedings of International Conference on VLDB Endowment in France*,2

IV. Hadoop: Open-source implementation of MapReduce.

V. Hung-chih Yang, Ali Dasdan, and Ruey-Lung Hsiao, Map-Reduce-Merge: simplified relational data processing on large clusters, In Proceedings of ACM SIGMOD Intl Conf. on Management of Data (SIGMOD07), pp. 1029-1040.

VI. Dr. Ananthi Sheshasayee1, J V N Lakshmi2 A Study on Hadoop Architecture for Big Data Analytics, International Journal of Advanced Technology in Engineering and Science.

VII. Dr.Ananthi Sheshasayee1, JVNLakshmi2 Comparison on Machine Learning Algorithm on Map Reduction for Performance Improvemen tin Big Data Indian Journal of Science and Technology.

VIII. Dhyani, B., and Barthwal, A. (2014). Big Data Analytics using Hadoop. International Journal of Computer Applications, 108(12), 15.