INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Advancements in Engineering, Management, Science & Technology (ICEMST 2021)

Big Data Analytics in MapReduce: Literature Review

Janani S [1], Dr D F X Christopher [2]

[1]*Research Scholar, Department of Computer Science, Rathnavel Subramaniam College of Arts & Science, Coimbatore, Tamilnadu, India.*

[2]*Director, School of Computer Studies, Rathnavel Subramaniam College of Arts & Science, Coimbatore, Tamilnadu, India.*

janusarguru@gmail.com[1]

Abstract

Big Data comprises both structured and unstructured data collected from various sources. For collecting, managing, storing and analyzing the large dataset, an efficient tool is required. Hadoop is an open source framework which processes large dataset and MapReduce in Hadoop is an effective programming model reduces the computation time of large scale database in a distributed architecture. A machine and deep learning algorithm based on MapReduce implemented in huge dataset will reduce processing time. This paper aims to study various MapReduce based model and algorithms to analyze huge data. Also, predicts the way of implementing algorithms in MapReduce to reduce the computing time.

Keywords: MapReduce, Hadoop, Machine Learning Algorithms, Deep Learning, HDFS.

1. Introduction

Data is being generated in all the major sectors include Healthcare, E-Commerce, Social media, Banking, Finance, etc., in the range of peta to Exabyte. Processing this huge dataset in a sequential program increases processing time, whereas, processing big data in a distributed architecture and the MapReduce programming model reduces the processing time with the increase of number of data nodes. An efficient processing can be discovered and automated in collaboration with machine and deep learning in MapReduce framework. Typically, MapReduce technique in Hadoop processes large scale dataset in parallel. Implementing a MapReduce-based machine/deep learning algorithm with increased number of nodes would improve efficiency and reduces processing time. Many authors proposed MapReduce based model, system and approach to analyze big data effectively. In this paper, those approaches were analyzed and compared. [1-5].

1.1 Hadoop

Hadoop is an open source Framework for big data by Apache to store and process big data in a distributed environment. Hadoop's Architecture has two main components:

Distributed File System:

Hadoop Distributed File System is designed to store and process large datasets which runs on commodity hardware. HDFS is similar to other existing distributed file system but the most significant feature is highly fault-tolerant and can be deployed on low cost hardware. In addition, HDFS follows Master/Slave Architecture where metadata is stored in NameNode which acts as a master server and application data is stored in DataNode which acts as a slave server. For reliability, the file content is duplicated on DataNode.

MapReduce:

MapReduce is a programming model to process huge dataset in parallel in a distributed

environment. Mapreduce algorithm proposed by Google to enhance the speed by processing distributed big data in a cloud platform. Major phases involved in MapReduce phases are:

i) **Split**: This phase splits the input into fixed number of pieces to get evaluated in map phase.

ii) **Map** : In the map phase, the data from a data blocks get split and key-value pairs are generated for the data.

iii) **Shuffle**: In this phase, the key-value pairs generated from the map function is passed as an input and clubs together the similar information in it.

iv) **Reduce**: This phase uses the output of shuffle phase and aggregates it, where data reduced into a single output value. Fig.1 depicts the architecture of MapReduce Framework.

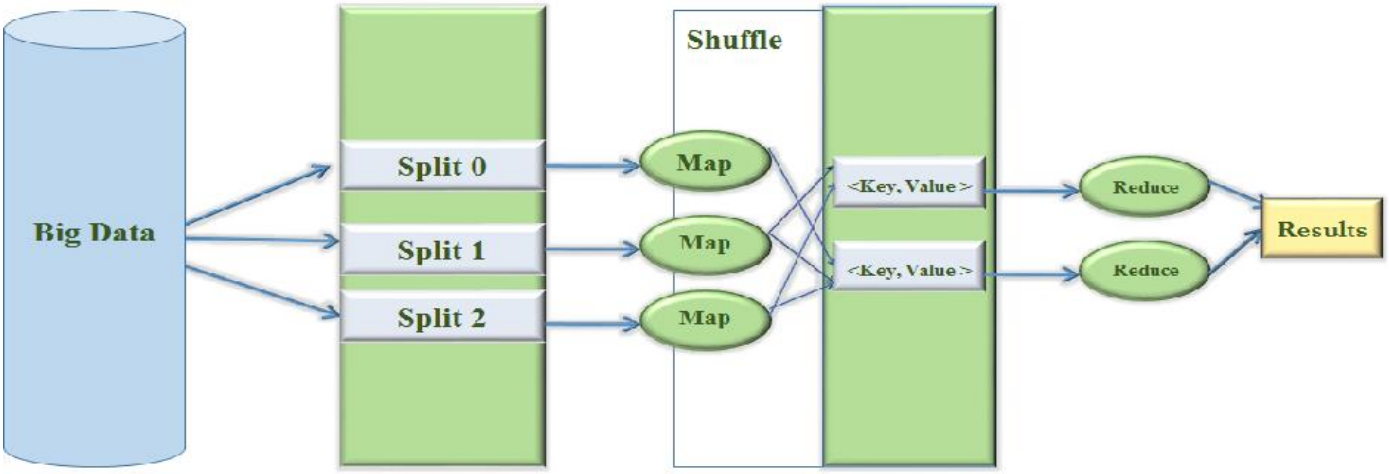

Fig.1. MapReduce Architecture

2. Literature Review

Nasullah Khalid Alham, Maozhen Li , Yang Liu and Suhel Hammoud,[3] proposed an annotation of image automatically using MapReduce based SVM. MRSMO splits the large dataset into smaller subset and this split subset is allocated to a map task. Map function present in the task optimizes the subset in parallel. Output of the map reduce differ in terms of linearity. For linear SVM, partial weight vector from map task penetrates into reduce task to get global weight vector and for non-linearity, the alpha array into reduce task finally gives global alpha array. Anan Banharnsakun[4], recommended a MapReduce incorporated artificial bee colony (MR-ABC) for clustering. This incorporation aims to minimize the sum of squared Euclidean Distance and centroid. The map function retrieves the cluster's centroid from the ABC and it is stored in HDFS. Centroid Value extracted from each bee to calculate the distance value between the centroid values and data record to obtain the minimum distance. Reduce function groups the same key value obtained from map function to determine the average distance and it returned as a fitness value. Daniel Valcarce , Javier Parapar and Alvaro Barreiro[6-8] proposed a MapReduce based recommender system implemented Posterior Probability Clustering algorithm on the basis of matrix factorization followed by Relevance Models. To reduce the complexity, the algorithm is implemented in MapReduce (distributed) framework to obtain the recommender for processing huge dataset. Furthermore, two join strategies, replication and broadcast were involved to make an efficient process. Weizhong Zhao, Huifang Ma and Qing He [9-11], recommended the MapReduce based PKMeans Cluster to analyse huge dataset effectively. Map function assigns the closest center to each sample and reduce function updates the new center and all the samples can be aggregated and determines the total number of samples. Shiva Asadianfam, Mahboubeh Shamsi and Abdolreza Rasouli Kenari[6] proposed a TVD-MRDL algorithm to automate the detection of the violation of drivers using MapReduce technique. Analysis include both structured and unstructured data. The proposed system able to analyze the traffic control center's data and the descriptions predefined by police. To process the image, Deep Learning algorithm named CNN is involved.

Mininath Bendre, Ramchandra Manthalkar, performed a case study to predict the pattern of student behavior of UCAM students by opting an Azure HDInsight big data solution by using its HDFS implementation. The association rules for the events done by the students obtained by implementing the apriori algorithm and further included MapReduce framework. Neha Verma, Dheeraj Malhotra & Jatinder Singh[8], presented a novel approach using association mining for the analysis of market basket to know customer's expectation from retail store. Customer's buying pattern analyzed using the MapReduce based Apriori algorithm implemented using IRM tool. Ms. Vandana Vijay, Dr. Ruchi Nanda2[9], proposed MRC-COVID system to store and process the Covid-19 dataset. In-Memory cache can be implemented on MapReduce to reduce the superfluous operations of disk I/O in runtime. Imparting cache in MapReduce improves performance and reduces the workload of data.

Table.1. A short review of various Machine and Deep learning algorithms implemented in Map Reduce

The comparision of various algorithms incorporated in MapReduce framework is shown in Table 1.

Authors	Algorithm	Data Source	Accuracy	Comparison	Objective
Nasullah Khalid Alham, Maozhen Li , Yang Liu and Suhel Hammoud	MRSMO	Unlabelled image(50 image) Multiclass Classification (5000 images)	93%		To annotate the image automatically
Anan Banharnsakun	MR-ABC	4 Datasets (Iris, CMC, Wine,Vowel)	90%	PKMeans and parallel K-PSO	To minimize the squared Euclidean Distance's sum between the instance of data and cluster's centroid
Daniel Valcarce , Javier Parapar and Álvaro Barreiro	PPC+RM2	Netflix		UB-Pearson, SVD, NMF – Pearson etc.,	Scalable and Distributed based recommender using MapReduce.
Weizhong Zhao, Huifang Ma and Qing He	Parellel K-Means Clustering using MapReduce	Datasets of different sizes			To analyze large dataset effectively
Shiva Asadianfam, Mahboubeh Shamsi and Abdolreza Rasouli Kenari	TVD-MRDL	Images from the traffic controlled center	Efficiency increased by more than 75%	Hadoop in stand-alone mode and Hadoop with more data nodes	To detect driver violations and behavior changes.
Mininath Bendre, Ramchandra Manthalkar	Student Behavior analysis in LMS using Big Data Framework	70 GB of information regarding the behavior of UCAM students			To predict the pattern of student behavior of UCAM students.
Neha Verma, Dheeraj Malhotra & Jatinder Singh	MR based Apriori Algorithm	Retail Transactional Database(generated using IBM tool)	Size up provides better result	Factors: Speed, Size and Scale	To identify the buying pattern of customer
Ms. Vandana Vijay, Dr. Ruchi Nanda2	MRC-COVID	Covid-19 data			To store and process Covid-19 data

Conclusion:

MapReduce is an effective framework to process large data set in parallel. Machine learning and deep learning algorithm implementation in MapReduce results in better performance. In this paper, the various algorithms, systems and models proposed by authors related to the big data analytics in MapReduce model were overviewed and the efficiency of the proposed algorithms were discussed. There is a dearth in focusing feature engineering in this implementation. In future, an effective algorithm to process feature engineered large dataset to be implemented in MapReduce will be proposed.

References

Journals

[1]. Fatih Gürcan, "Major Research Topics in Big Data: A Literature Analysis from 2013 to 2017 Using Probabilistic Topic Models", 978-1-5386-6878-8/18 ©2018 IEEE.

[2]. Next-Generation Big DataAnalytics: State of the Art, Challenges, and Future Research Topics, DOI 10.1109/TII.2017.2650204, IEEE Transactions on Industrial Informatics

[3]. Alham NK, Li M, Liu Y and Hammoud S, "A MapReduce-based distributed SVM algorithm for automatic image annotation", Comput Math Appl 62(7):2801–2811, 2011.

[4]. Banharnsakun A, "A MapReduce-based artificial bee colony for large-scale data clustering", Pattern Recogn Lett 93:78–84, 2017.

[5]. Cantabella M, Martínez-Espana R, Ayuso B, Yanez JA and Munoz A, Analysis of student behavior in learning management systems through a big data framework. Futur Gener Comput Syst 90:262–272, 2019.

[6]. Shiva Asadianfam, Mahboubeh Shamsi and Abdolreza Rasouli Kenari, "TVD-MRDL: traffic violation detection system using MapReduce-based deep learning for large-scale data", Multimedia Tools and Applications, https://doi.org/10.1007/s11042-020-09714-8, September 2020.

[7]. Daniel Valcarce , Javier Parapar and Alvaro Barreiro, "A MapReduce implementation of posterior probability clustering and relevance models for recommendation", Engineering Applications of Artificial Intelligence, https://doi.org/10.1016/j.engappai.2018.08.006, September 2018.

[8]. Neha Verma, Dheeraj Malhotra & Jatinder Singh, "Big data analytics for retail industry using MapReduce-Apriori framework", Journal of Management Analytics, 2020 https://doi.org/10.1080/23270012.2020.1728403

[9]. Ms. Vandana Vijay, Dr. Ruchi Nanda2, "MRC-COVID (Map Reduce with Cache) System for Big data Analytics",International Journal of All Research Education and Scientific Methods (IJARESM), ISSN: 2455-6211 Volume 8, Issue 12, December-2020, Impact Factor: 7.429.

Symposium

[10]. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler "The Hadoop Distributed File System", 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Sunnyvale, California USA.

Conference

[11]. Weizhong Zhao, Huifang Ma and Qing He, "Parallel k-means clustering based on MapReduce," in IEEE International Conference on Cloud Computing, pp. 674–679: Springer.

Biography

S.Janani has completed her Master degree in Computer Applications from kalasalingam University (MCA, 2015) and M.Phil from Dr.SNS Rajalakshmi college of Arts and Science, 2018. . She Qualified the UGC – NET Examination for eligibility for lecturer ship held on December, 2018. Currently, she is working as an Assistant Professor at KPR College of Arts, Science and Research. She has academic experience of 5 years Her specialization area is Big Data. Published journals and presented papers in an international conference.

Dr.D.F.X.Christopher is serving as a Director (Administration) & Associate Professor in School of Computer Studies at R.V.S (Rathnavel Subramaniam) College of Arts and Science. His primary expertise in Agile Software Engineering and secondary in Data

Structures & Algorithms and Networking, and his research focuses on determining software requirements stability based on complexity point measurement adopting multi-criteria fuzzy approach augmenting quality models. He associated with Academics since 1998, currently holding an experience of 21 years in Teaching and 17 years in Research. In 2014 he earned Ph.D. in Bharathiar University, where he had received an M.Phil in 2002. He produced 62 M.Phils' and currently guiding 8 Ph.D scholars' in the area of Software Engineering and Networking. He is serving as an Academic Expert for the board of studies for various institutions in India. He is a senior editorial member for the journals viz. International Association of Engineers (IAENG), Institute of Research Engineers and Doctors (IRED), Universal Association of Computer and Electronics Engineers (UACEE) and International Association of Computer Science and Information Technology (IACSIT).He severed as a Principal IT consultant for various startup software companies in and around Coimbatore.