A REVIEW ON HADOOP AND MAPREDUCE TECHNIQUES IN BIGDATA

V.Shanmugapriya [#1] and Dr.D.Maruthanayagam [*2]

[#] *Research Scholar, Periyar University, Salem, Tamilnadu.India.*

[*] *Head cum Assistant Professor, PG and Research Department of Computer Science, Sri Vijay Vidyalaya College of Arts & Science, Dharmapuri,Tamilnadu,India.*

***Abstract*— Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. Big data analytics is the process of examining large amounts of data. Big Data is characterized by the dimensions volume, variety, and velocity, while there are some well-established methods for big data processing such as Hadoop which uses the map-reduce paradigm. In this paper we have discussed the analysis of the Big Data Analytics concepts and some existing techniques and tools, like Hadoop.**

***Index Terms*— Big Data, Hadoop, Mapreduce, BDA,HDFS,Data Mining**

I. INTRODUCTION

Big Data is very familiar term that describes voluminous amount of data that is structural, semi-structural and sub structural data that has potential to be mined for information. Although big data does not refer any specific quantity, then this term is often used when speaking about the pet bytes and Exabyte of data. Big Data Analytics, is the process of examining large data sets that containing a variety of data types i.e., big data to uncover all hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Then analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.

The primary goal of big data analytics is to help companies make more informative business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transactional data, as well as other forms of data that may be untapped by more conventional Business Intelligence(BI) programs. That could include web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey responses, mobile phone call detail records and machine data captured by sensors and connected to the Internet of Things. Big data burst upon the scene in the first decade of the 21st century, and the first organization to embrace it were online and start-up firms. Arguably, firms like Google, LinkedIn, and eBay and Face book were built around big data from the beginning. They did not have to reconcile or integrate big data with more traditional sources of data and the analytics performed upon them, because they didn't have that much of traditional forms. They didn't have to merge big data technologies with their traditional IT infrastructures because these infrastructures didn't exist. Big data could stand alone, big data analytics could be the only focus of analytics, and big data technology architectures could be the only architecture. So big data using Hadoop and

No SQL free software's. *Hadoop and No SQL Free Software:* Hadoop is an open-source distributed file system that is capable of storing and processing large volumes of data in parallel across a grid of commodity servers. Hadoop emanated from companies such as Google and Yahoo, which needed a cost effective way to build search indexes. Engineers at these companies knew that traditional relational databases would be prohibitively expensive and technically unwieldy, so they came up with an alternative that they built themselves. Eventually, they gave it to the Apache Software Foundation so others could benefit from their innovations. Today, many companies are implementing Hadoop software from Apache as well as third-party providers such as Cloud era, Horton works, EMC, and IBM. Developers see Hadoop as a cost effective way to get their arms around large volumes of data. Companies are using Hadoop to process, store and analyse large volumes of Web log data so they can get a better feel for the browsing and shopping behaviour of their customers. Previously, most of the companies outsourced the analysis of their click stream data or simply let it "fall on the floor" since they couldn't process it in a timely and cost effective way.

II. LITERATURE REVIEW

Bhawna Gupta, et al. [1] proposes the use of BDA (Big Data Analytics) for analyzing the enterprise data. The main focus is to gather the unstructured data from all the terminals, processed the data to convert into structured form so that accessing of the data would be easier. BDA describes the simple algorithm for large amount of data without compromising performance. Hadoop is one of the tools which are aimed to improve the performance of data processing. In this approach they are managing the Big Data characteristics of large volumes of enterprise data. If enterprise has an unmet business need for strategic decision making with a high degree of processing, a Revolution Analytics and Hadoop combination offers significant opportunity to gain advantage. Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.

Ulla Gain1, et al. [2] develops BD and symbolizes the aspiration to build platforms and tools to ingest, store and analyze data that can be voluminous, diverse, and possibly fast changing. This strategy is partly descriptive and partly improving. Through launching the term data-milling the Authors try to improve understanding of the phenomenon of BD, as well as, possibilities of data analytics. Launched the term data-milling to represent the searching of the information nuggets from the heterogeneous data. To justify the launched term data-milling, they made the literature review in which they searched the definitions of BDA. Their study shows that BDA is verbosely explained. They used only four statements from 19 to crystallize BDA. The literature review of BDA gave the description of current status of the phenomenon BD. The launched term data-milling improves the understanding of the phenomenon BD, as well as, possibilities of data analytics. There exist large amounts of heterogeneous digital data. This phenomenon is called BD which will be examined. The examination of BD has been launched as BDA.

Jainendra Singh, et al. [3] discusses about Machine Learning (ML) techniques which have found widespread applications and implementations in security issues. Machine Learning algorithms are used in very diverse contexts: 1) to recognize handwritten text, 2) to extract information from images, 3) to build automatic language translation systems, 4) to predict the behavior of customers in an online shop, 5) to find genes that might be related to a particular disease, and so on. This approach focuses on the development of fast and efficient algorithms for real-time processing of data as a main goal to deliver accurate predictions of various kinds. ML techniques can solve the above mentioned applications using a set of generic methods that differ from more traditional statistical techniques. It specifies that the advancement in ML, provides new challenges and solutions to the security problems encountered in applications, technologies and theories.

S. Vikram Phaneendra, et.al. [4] Illustrated that in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as “big data”. In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc.

Sagiroglu, S, et al. [5] describe the big data content, its scope, methods, samples, advantages and challenges of Data. The critical issue about the Big data is the privacy and security. Big data samples describe the review about the atmosphere, biological science and research. Life sciences etc. By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the big data easy to get the information from the complicated data sets.

W. Gao, et al. [6] Complexity, diversity, frequently changing workloads and rapid evolutions of big data systems raise great challenges in big data benchmarking. Most of the big data benchmarking efforts targeted evaluating specific types of applications or system software stacks, and hence they are not qualified much. The bigDataBench not only covers broad application scenarios, but also includes diverse and representative data sets. Compared with other benchmarking suites, BigDataBench has very low operational intensity and the volume of data input has non negligible impact on micro-architecture characteristics.

Yaxiong Zhao, et al. [7] the buzz-word big-data (application) refers to the large-scale distributed applications that work on unprecedentedly large data sets. Google’s Map Reduce framework and Apache’s Hadoop, its open-source implementation, are the defacto software system for big-data applications. An observation regarding these applications is that they generate a large amount of intermediate data, and this abundant information is thrown away after the processing finish. Motivated by this observation, a data-aware cache framework for big-data applications, which is called Dache. In Dache, tasks submit their intermediate results to the cache manager. A task, before initiating its execution, queries the cache manager for potential matched processing results, which could accelerate its execution or even completely saves the execution. A novel cache description scheme and a cache request and reply protocols are designed. Dache is implemented by extending the relevant components of the Hadoop project. Tested experiment results demonstrate that Dache significantly improves the completion time of Map Reduce jobs and saves a significant chunk of CPU execution time.

Vidyasagar S. D, et al. [8] did a survey on Big Data and Hadoop system and found that organizations need to process and handle petabytes of Data sets in efficient and inexpensive manner. According to him if there is any node failure then we can lose some information. Hadoop is an Efficient, reliable, Open Source Apache License. Hadoop is used to deal with large data sets. Author explained its need, uses and application. Now days, Hadoop is playing an important role in Big Data. Vidyasagar S.D concluded that “Hadoop is designed to run on cheap commodity hardware, it automatically handles data replication and node failure, it does the hard work – you can focus on processing data, Cost Saving and efficient and reliable data processing”.

Jian Tan, et al. [9] author talks about the theoretical assumptions, that improves the performance of Hadoop/map reduce and purposed the optimal reduce task assignment schemes that minimize the fetching cost per job and performs the both simulation and real system deployment with experimental evolution. The advantage of this paper is improves the performance of large scale Hadoop clusters. The disadvantage of this paper is environmental factors such as network topologies effect on a reduce task in map reduce clusters.

Jonathan Stuart Ward et.al. [10] did a survey of Big data definition, Anecdotally big data is predominantly associated with two ideas: data storage and data analysis. Despite the sudden Interest in big data, these concepts are far from new and have long lineages. This, therefore, raises the question as to how big data is notably different from conventional data processing techniques. For rudimentary insight as to the answer to this question one need look no further than the term big data. \Big" implies significance, complexity and challenge. Unfortunately the term\big" also invites quantification and therein lies the difficulty in furnishing a definition. The lack of a consistent definition introduces ambiguity and hampers discourse relating to big data. This short paper attempts to collate the various definitions which have gained some degree of traction and to furnish a clear and concise definition of an otherwise ambiguous term.

H.Herodotou, *et al*. [11] provides a technique to implement self tuning in Big Data Analytic systems. Hadoop's performance out of the box leaves much to be desired, leading to suboptimal use of resource, time and money. This paper introduces Starfish, a self tuning system for big data analytics. Starfish builds on Hadoop while adapting to user needs and system workloads to provide good performance automatically, without the need for users to understand and manipulate the many tuning knobs in Hadoop. Explores the MADDER properties (i.e Magnetism, Agility, Depth, Data-lifecyle-awareness, Elasticity, Robustness). The behavior of a map reduce job is controlled by settings of more than 190 configuration parameters. If the user does not specify the settings, then default values are used. Good settings for these parameters depend on job, data, and cluster characteristics. Starfish's Just In Time Optimizer addresses unique optimization problems to automatically select efficient execution techniques for map reduce jobs.

Chris Jermaine et.al. [12] Proposes a Online Aggregation for Large-Scale Computing. Given the potential for OLA to be newly relevant, and given the current interest on very large-scale, data-oriented computing, in this paper we consider the problem of providing OLA in a shared-nothing environment. While we concentrate on implementing OLA on top of a MapReduce engine, many of author's most basic project contributions are not specific to MapReduce, and should apply broadly. Consider how online aggregation can be built into a MapReduce system for large-scale data processing. Given the MapReduce paradigm's close relationship with cloud computing (in that one might expect a large fraction of MapReduce jobs to be run in the cloud), online aggregation is a very attractive technology. Since large-scale cloud computations are typically pay-as-you-go, a user can monitor the accuracy obtained in an online fashion, and then save money by killing the computation early once sufficient accuracy has been obtained.

Jeffrey Dean et.al. [13] Implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on thousands of machines. Programmers and the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google's clusters every day. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine Communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. Author proposes Simplified Data Processing on Large Clusters.

Chen He Ying Lu David Swanson, et.al [14] develops a new MapReduce scheduling technique to enhance map task's data locality. He has integrated this technique into Hadoop default FIFO scheduler and Hadoop fair scheduler. To evaluate his technique, he compares not only MapReduce scheduling algorithms with and without his technique but also with an existing data locality enhancement technique (i.e., the delay algorithm developed by Facebook). Experimental results show that his technique often leads to the highest data locality rate and the lowest response time for map tasks. Furthermore, unlike the delay algorithm, it does not require an intricate parameter tuning process.

Tyson Condie, et.al. [15] propose a modified MapReduce architecture in which intermediate data is pipelined between operators, while preserving the programming interfaces and fault tolerance models of other MapReduce frameworks. To validate this design, author developed the Hadoop Online Prototype (HOP), a pipelining version of Hadoop. Pipelining provides several important advantages to a MapReduce framework, but also raises new design challenges. To simplify fault tolerance, the output of each MapReduce task and job is materialized to disk before it is consumed. In this demonstration, we describe a modified MapReduce architecture that allows data to be pipelined between operators. This extends the MapReduce programming model beyond batch processing, and can reduce completion times and improve system utilization for batch jobs as well. We demonstrate a modified version of the Hadoop MapReduce framework that supports online aggregation, which allows users to see "early returns" from a job as it is being computed. Our Hadoop Online Prototype (HOP) also supports continuous queries, which enable MapReduce programs to be written for applications such as event monitoring and stream processing.

III. CONCLUSION

New techniques of big data such as **Hadoop** and **MapReduce** create alternatives to traditional data warehousing. Traditional Hadoop with combination of new technologies explores a new scope of study in various fields of science and technologies. This paper explains the big data concept and its importance in business. The technologies used for big data processing mainly Hadoop and Map Reduce. Management tools for big data are explained and also the big data techniques are discussed.

REFERENCES

[1] Bhawna Gupta , Dr. KiranJyoti ,”Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3867-38702014

[2] Ulla Gain1 ,VirpiHotti ,”Big Data Analytics for Professionals, Data-milling for Laypeople “,International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 4, Number 1 (2014), pp. 33- 402013

[3] Jainendra Singh , “Big Data Analytic and Mining with Machine Learning Algorithm”, World Journal of Computer Application and Technology 1(2): 51 -57, DOI: 10.13189/wjcat.2013.010205,2014

[4] S.Vikram Phaneendra & E.Madhusudhan Reddy “Big Data- solutions for RDBMS problems- A survey” In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[5] Sagiroglu, S.; Sinanc, D., (20-24 May 2013), Big Data: A Review

[6] W. Gao, Y. Zhu1, Z. Jia, C. Luo, L. Wang, Z. Li, J. Zhan, Y. Qi, Y. He, S. Gong, Xiaona Li,S. Zhang, and B. Qiu. BigDataBench: a Big Data Benchmark Suite from Web Search Engines. in The Third Workshop on Architectures and Systems for Big Data(ASBD 2013) in conjunction with The 40th International Symposium on Computer Architecture, May 2013.

[7] Yaxiong Zhao, Jie Wu. Dache: A Data Aware Caching for Big-Data Applications Using The MapReduce Framework. Proc. 32nd IEEE Conference on Computer Communications, INFOCOM 2013, IEEE Press, Apr. 2013, pp. 35-39.

[8] Vidyasagar S. D, A Study on “Role of Hadoop in Information Technology era”, GRA - GLOBAL RESEARCH ANALYSIS, Volume : 2 | Issue : 2 | Feb 2013 • ISSN No 2277 – 8160.

[9] Jian Tan; Shicong Meng; Xiaoqiao Meng; Li ZhangINFOCOM, “Improving ReduceTask data locality for sequential MapReduce” 2013 Proceedings IEEE ,1627 – 1635

[10] Jonathan Stuart Ward and Adam Barker “Undefined By Data: A Survey of Big Data Definitions” Stamford, CT: Gartner, 2012.

[11] H.Herodotou,H.Lim,G.Luo,N.Borisov,L.Dong,F.B.Cetin,and S. Babu. Starfish: A Selftuning System for Big Data Analytics. In CIDR, pages 261–272, 2011.

[12] Niketan Pansare1, Vinayak Borkar2, Chris Jermaine1, Tyson Condie “Online Aggregation for Large MapReduce Jobs” August 29September 3, 2011, Seattle, WA Copyright 2011 VLDB Endowment, ACM

[13] Jeffrey Dean and Sanjay Ghemawat “MapReduce: Simplified Data Processing on Large Clusters” OSDI 2010

[14] Chen He Ying Lu David Swanson “Matchmaking: A New MapReduce Scheduling” in 10th IEEE International Conference on Computer and Information Technology (CIT’10), pp. 2736–2743, 2010

[15] Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein “Online Aggregation and Continuous Query support in MapReduce” SIGMOD’10, June 6–11, 2010, Indianapolis, Indiana, USA. Copyright 2010 ACM 978-1-4503-0032-2/10/06.

V.Shanmugapriya received her M.Phil Degree from Periyar University, Salem in the year 2007. She has received her M.C.A Degree from Madurai Kamaraj University, Madurai in the year 2002. She is working as Assistant Professor, Department of Computer Science, PGP College of Arts & Science, Namakkal, Tamilnadu, India. She is pursuing her Ph.D Degree at Periyar University. Salem, Tamilnadu, India. Her areas of interest include Big Data and Data Mining.

Dr.D.Maruthanayagam received his Ph.D Degree from Manonmanium Sundaranar University, Tirunelveli in the year 2014. He has received his M.Phil, Degree from Bharathidasan University, Trichy in the year 2005. He has received his M.C.A Degree from Madras University, Chennai in the year 2000. He is working as Head cum Assistant Professor, PG and Research Department of Computer Science, Sri Vijay Vidyalaya College of Arts & Science, Dharmapuri, Tamilnadu, India. He has 14 years of experience in academic field. He has published 1 book, 19 International Journal papers and 27 papers in National and International Conferences. His areas of interest include Grid Computing, Cloud Computing and Mobile Computing.