I.J. Intelligent Systems and Applications, **2017, 4, 61-68**
Published Online April 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2017.04.07

A Survey Work on Optimization Techniques Utilizing Map Reduce Framework in Hadoop Cluster

Bibhudutta Jena
KIIT UNIVERSITY, BHUBANESWAR
E-mail: bibhudutta93@gmail.com

Mahendra Kumar Gourisaria, Siddharth Swarup Rautaray, Manjusha Pandey
KIIT UNIVERSITY, BHUBANESWAR

***Abstract*—**Data is one of the most important and vital aspect of different activities in today's world. Therefore vast amount of data is generated in each and every second. A rapid growth of data in recent time in different domains required an intelligent data analysis tool that would be helpful to satisfy the need to analysis a huge amount of data. Map Reduce framework is basically designed to process large amount of data and to support effective decision making. It consists of two important tasks named as map and reduce. Optimization is the act of achieving the best possible result under given circumstances. The goal of the map reduce optimization is to minimize the execution time and to maximize the performance of the system. This survey paper discusses a comparison between different optimization techniques used in Map Reduce framework and in big data analytics. Various sources of big data generation have been summarized based on various applications of big data.The wide range of application domains for big data analytics is because of its adaptable characteristics like volume, velocity, variety, veracity and value .The mentioned characteristics of big data are because of inclusion of structured, semi structured, unstructured data for which new set of tools like NOSQL, MAPREDUCE, HADOOP etc are required. The presented survey though provides an insight towards the fundamentals of big data analytics but aims towards an analysis of various optimization techniques used in map reduce framework and big data analytics.

***Index Terms*—**MAPREDUCE, Optimization, Big Data, HADOOP, NOSQL, Processing Capabilities.

I. INTRODUCTION

Now- a -days computer has become an indispensable tool in the day-to-day activities. In fact, it is very difficult to get through a working day without it .As the modern civilization gradually more dependent on the age of information which is concentrated in computers, So the Society is becoming more dependent on computers and technology for functioning in every- day life. As the number of users are gradually increasing day by day, so the data generated by them is also enlarging gradually and to handle this large amount of generated data all are facing many difficulties, therefore these data are termed as big data. The presented paper reviews the fundamentals of big data, applications of big data, characteristics of big data, processing capabilities of a system, HADOOP eco system and also gives an idea about various optimization techniques used in MAPREDUCE framework.

A. Sources of Big Data

Big data means large amount of generated and used data, this large volume of data stored as collection of large datasets were not able to be operated using basic computing techniques. It is not simply a data, but it has become a veritable research area and a full fledge subject [1], [2] , which consists of various tools, techniques and frameworks. It embroils the data fabricated by different devices and applications .Some of these applications generating big data are enlisted as follows:

Black Box Data: It is a wedge of airplanes, and rockets, etc. It captures voices of the flight crew, recordings of microphone which is generally consider as big data.

Social Media Data: Social media just like facebook and Twitter contain information and the views posted by the people across the world [3] with day to day increasing in usage of social media by different users and some other sources of big data generation are search engine data and stock exchange data.

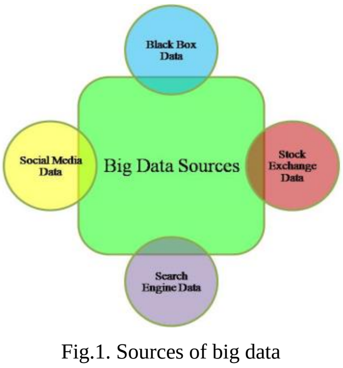

Fig.1. Sources of big data

The above figure [1] represents various big data sources which contribute to the large amount of data being produced. To efficiently analyze these data using various big data tools is a herculean task which could be simplified by inclusion of various optimization techniques in big data analytics.

B. Applications

As is evident by the above retained sources of big data, applications o f big data has also increased a lot. Some of the major applications of big data have been summed up in the following section:

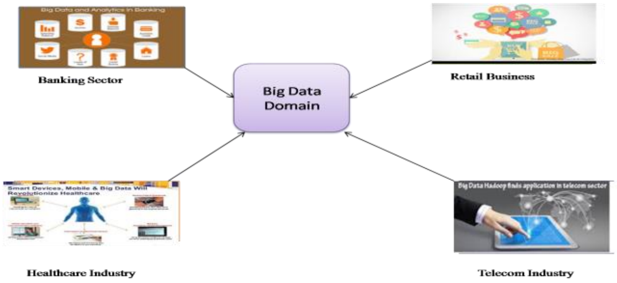

Fig.2.Application of big data

Figure [2] describes the different areas or domains in which big data is currently being used rigorously for different purposes.

Retail Business: The trend of shopping has altered dramatically in recent years as power has transferred to customers.

Banking sector: The Banking industry develops a huge volume of data on a day to day basis.

Healthcare industry: Big data is helpful for making the world a better place and the best example to understand this is investigating the uses of big data in healthcare industry.

Telecom industry: The tremendous growth of smart phones, communications service providers (CSPs) are viewing large extension in the volume of data moving across their networks.

The above mentioned application domains are a subset of the wide range of applications of big data analytics in the current years [4]. Huge amount of data is generated from different application domains and it is very difficult to analyze this vast amount of data. Therefore different optimization techniques and tools are implemented for analyzing the big data and improving the decision making skills in different domains of big data.

C. Big data characteristics

This section describes characteristics of big data which plays a pivotal role for importation of challenges in big data analytics [5].

Volume: Volume indicates to the vast amounts of data generated every second by various applications used for man-machine interaction.

Velocity: Velocity indicates the rate by which data is produced and the rate on which data travels throughout. For example the social media messages are getting viral in few seconds.

Variety: Variety indicates to the various types of data currently use. Now-a-days different types of data are generated such as structured data, unstructured data and semi structured data etc.

Veracity: Veracity indicates to the uncertainty of the data. It also determines the trustworthy of data whether the obtained data is correct or not.

Value: This V generally takes into consider when looking at Big Data i.e Value. It is all well and good having ingress to big data but until the conversation of data into value it is worthless.

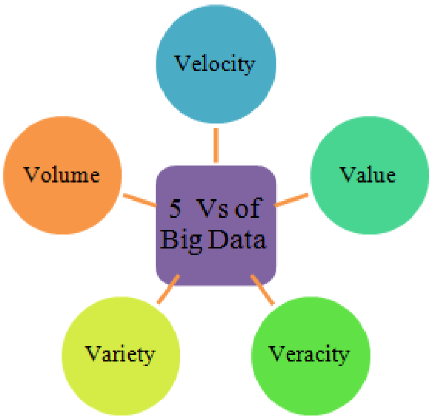

Fig.3.Bigdata characteristics

The above figure [3] describes the different characteristics of big data analytics. These are the 5 basic characteristics in big data analytics. The mentioned characteristics of big data are due to the addition of structured ,semi structured , unstructured data ,However different optimization techniques are used to analyze these data and enhancing the decision making process in different domains of big data.

The reminder of this paper is organized as follows .Section II presents a comparison between SQL VS NOSQL. Section III presents a brief idea about hadoop and its eco system. The motivation for this work is laid out in Section III. Section IV represents different optimization techniques used in map reduce framework and big data. Analysis part is described in Section V. Finally, Section VI concludes this paper and section VII describes the future scope.

II. SQL VS NOSQL & PROCESSING CAPABILITIES

In the era of database technology, databases basically are of two types: SQL and NOSQL .The major difference presents in which way they're built, types of information generally they store, and their storing techniques of the data in to it. Relational databases are normally structured, just like phonebooks that contains phone numbers and addresses.

A. SQL

Structured Query Language (SQL) is a basic computer language for relational database management and data manipulation. SQL is helped to query, insert, update and modify data. It is generally used to interact with the database by using some queries. SQL database follow a fixed table structures and provide the ability to select data from these tables by using structured query language.

B. NOSQL

NOSQL is a term basically refered to define a class of non-relational databases that can scale horizontally to very large data sets but never give ACID guarantees [6],[7]. Nosql data keeps data vary widely in their offerings and have some definite features of its own. The CAP Theorem provided [8] by Eric Brewer in 2000 describes that it is impossible for a distributed environment to be consistent, available, and partition-tolerant at the same time [9].

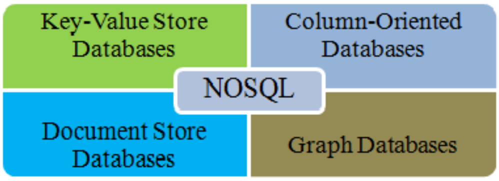

Fig.4. NOSQL Databases Types

C. Comparison between SQL vs NOSQL

Table 1.SQL vs NOSQL

SQL	NOSQL
Schema rigidity is present here	Schema rigidity is not present
Join operation can possible	Join operation can't possible.
Data are stored in a relational model with rows and columns.	Data are stored in either document, key value, graph model.
It follows vertical scaling.	It follows horizontal scaling.
Example:Mysql,db2 etc	Example MongoDB, Cassandra etc

The above table [1] describes some difference between SQL and NOSQL on basis of some characteristics.

D. Processing Capabilities

Processing Capabilities is a feature of a system, model or function that describes its capability to cope and perform under an expanding workload. A system that scales well will be able to maintain or even enlarge its level of performance or efficiency when tested by larger operational demands. Processing Capabilities of the system can also refer to the scalability of the system. Scalability can be defined as the capability of a system or network to control a sprouting amount of work, or its potential to be enhanced in order to lodge that growth. It is of two types:

- Vertical scaling (or scale up/down) signifies to attach properties to a single node in a system, usually including the inclusion of processors or memory to a unique computer. Vertical scaling of trending systems allow us to utilize virtualization technology more efficiently, as it gives more sources for hosting the set application modules to share. For example, Suppose we have 10 TB database in a mid size amazon machine instance .we can easily say that high query rates can exhaust our servers CPU power, can consume all of your RAM and sometimes we will find the working data set is exceeding our storage capacity. So, now this point, we are thinking about adding more CPU cores, more Storage and extra RAM to that instance to enhance the query performance .This is what we called Vertical scaling , means appending extra CPU power and storage resource to a single instance. Major benefit we get from vertical scaling is all of our data is in a single machine, No need to control multiple instance and the major problem we suffer from vertical scaling is the cost efficiency. A powerful machine having immense number of CPU and higher RAM capacity is costlier than a set of small size instances.
- Horizontal scaling separates the data set and distributes the data over multiple servers, or shards. So, you can generate 10 instance each with 1 TB database. Each shard is an independent database, and collectively, the shards make up a single logical database. Horizontally scaling (or scale out/in) signifies to attach extra nodes to a system, such as adding of a new system in to a distributed environment of software domain. The example can be given as scaling out from one Web system to three. Though computer prices have decreased and performance have increased day by day, Therefore high-performance based computing applications like as basaltic analysis have adopted low-cost "commodity" systems for the completion of tasks.

III. HADOOP

HADOOP is an Apache framework evolved in java that grants distributed operating of huge datasets across bunch of computers utilizing general programming models[10],[11]. A HADOOP frame-worked application operates in an environment that gives distributed storage capacity and computation across various clusters. HADOOP is developed to scale up from single server to thousands of systems, each providing local computation and storage [12]. Doug Cutting, Mike Cafarella and team have developed the solution supplied by Google and initiated an Open Source Project named as HADOOP in 2005 and Cutting gave a name to it after his son's toy elephant. Now Apache HADOOP is a registered framework of the Apache Software Foundation.

A. Hadoop Ecosystem

The HADOOP ecosystem refers to the different components of the Apache HADOOP software library [13],[14] as well as to the accessories and tools provided by the Apache Software Foundation for these types of software projects, and to the ways that they work together. Some of the major components of the HADOOP ecosystem has been summed up in the following section.

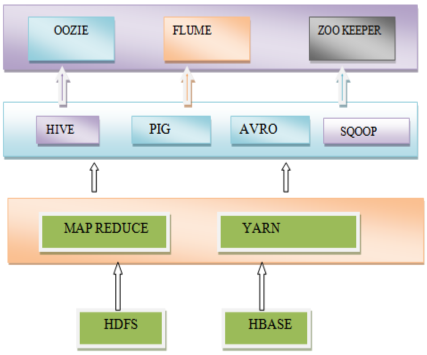

Fig.5. Hadoop Ecosystem

Figure [5] describes some of the components of the HADOOP Ecosystem that are currently using for analysing the big data.

Pig: Apache Pig is generally present in the above of Mapreduce in hadoop ecosystem [15]. Pig is naturally used with HADOOP.

Hive: Hive is a tool is generally responsible for processing the structured data. It is generally lies on the above of Hadoop to analyse the dataset.

HBASE: HBASE is a distributed column-oriented database made on top of the HADOOP file system. It is an open-source project and scalable horizontally.

OOZIE: Apache OOZIE is a scheduler system to run and control HADOOP jobs in a distributed environment. It is generally responsible for combining different complex jobs and run it sequentially [16].

SQOOP: SQOOP is a model designed to transfer data between HADOOP and relational database servers. It generally imports data from relational databases like MYSQL to HADOOP and HDFS platform [17], and vice versa.

YARN (Yet Another Resource Negotiator): YARN is necessary for HADOOP Enterprise to giving resource management and a domain to give frequent operations, and also responsible for data governance tools over HADOOP clusters.

Flume: Apache Flume is a technique ingestion mechanism for aggregating and transmitting huge amounts of running data just like log files, events etc from multiple sources to a centralized data store.

Zoo Keeper: Zoo Keeper is a co-ordination service to manage large set of hosts in distributed manner. Co-ordinating a service in a distributed environment is a complicated process. Zoo Keeper solves this problem by its simple architecture [18].

HDFS (HADOOP Distributed File System): It was evolved by using distributed file system concept. It works on hardware of commodity level. In compare to other distributed systems, HDFS is greatly fault tolerant and configured by using low-commodity hardware. HDFS is used to store plenty amount of data and gives easy access to it. For storing of large amount of data, the files are kept across multiple machines or nodes, which is suitable for the parallel storage and processing. HDFS follows the master-slave architecture where master is the name node and slave is the data node. The name node is the cheap hardware that consists of Linux operating system and a application software This software basically known as name node software. The major responsibilities of name node are: Governs the file system namespace present in HDFS, Responsible for Operating the client's access to files and also responsible for performing the file system operations such as renaming, closing, and opening of files present in HDFS keep track of all the data nodes present in the HADOOP framework. Similarly the data node is a cheap hardware that consists of Linux operating system and a software called as data node software. For each node present in the cluster, there should be a data-node which is responsible to control the data storage of their system. The major responsibilities of data nodes to perform the read write operations according to the client request. Hence some mechanism is present in HDFS which are generally responsible for automatic fault recovery detection

Hadoop Mapreduce Framework: It is an open source framework provided by Apache foundation written in java for distributed processing of large datasets among bunch of computers utilizing simple programming models .It consists of two phases.

Map stage: In map stage the mappers are responsible for extracting the data from different data files and kept it in the HADOOP file system (HDFS)[19]. The input file is moved to the mapper function sequential manner. The mapper operates the data and produces multiple small chunks of data.

Reduce stage: It comprises of both the shuffle stage and the reduce stage and in this phase the reducer operates the data that comes from the mapper and after processing it provides the reduced output in the form of key value pair which will send to the user.

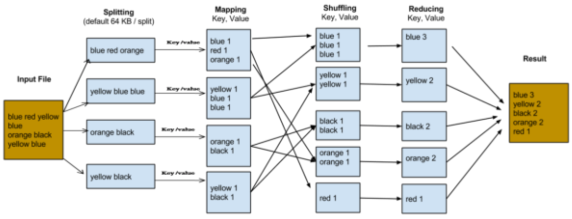

Fig.6. Example of Map Reduce

The figure [6] describes how the data flow occur in the map reduce framework through different stages. It is an example of map reduce word count programme. In the input stage data is extracted from the HDFS, After that the data are splitted in to number of parts and assign to the mappers, Then in the map stage map task will be execute and we will get the output in the form of key value pairs and this output will pass through the combine & shuffle phase and given as input to the reducer. In the reducer reduce task will be executed and we get the required output in the form of key value pair. Different optimization techniques are applied on HADOOP eco system to reduce the execution time for processing the large data sets and to ameliorating the decision making skill in big data domains.

IV. Different Mapreduce Optimization Techniques

In this section, the paper discussed about different optimization techniques and methods generally used in map reduce framework and big data. It also gives an idea how different optimization techniques are proposed to determine the number of the Map tasks and the Reduce tasks such that the execution time of the Map Reduce applications can be reduced. The presented survey also proposes methods to gain optimization in Map Reduce by using cross phase technique The basic idea behind the discussed techniques are to consider not only the execution cost of an individual task but also its impact on the subsequent phases performance. It has also provided the approach of isolating Map Reduce clusters in virtual machine, with a continuously adjustable performance based on user-determined spending rates, which can address many of the resource allocation inefficiencies in existing systems. It has also presented the design and configuration of an optimization method by using the multi queries for Map Reduce framework to reinforce the existing multi-queries processing. In the following table the paper has discussed about different optimization techniques and its characteristics.

The below table [2] presents a brief idea about some optimization techniques used in map reduce framework and big data analytics .It also discusses about the characteristics of various optimization techniques and how these techniques are helpful for improving the decision making process in big data domains. It is also described how the optimization approaches are used to maximize the Hadoop framework.

Table 2. Different optimization approaches

YEAR	AUTHOR	OPTIMIZATION TECHNIQUES	CHARACTERISTICS
2016	Luigi Troiano, Alfredo Vaccaro et al [20]	On-line Smart Grids Optimization on Big Data by Case-Based Reasoning	• On-line solution of compelled optimization problems is an necessary requirement for a secure, reliable smart grids operation. • The paper suggests improved computing operators focused on Big Data for smart grids optimization through online. • The basic idea behind this is to resolve a problem of generic constrained optimization by exploiting a module based on the Map Reduce framework, which is an trending computing platform for Big Data management initiated at Google.
2015	Anand Srinivasan, Chung-Horng Lung et al [21] .	Optimizing the performance of Big Data in Mobile Networks.	• The paper proposes an algorithm for cache coherency bandwidth optimization. • In this approach the user data shift is optimized without accommodating the user expectation rate. • The discussed algorithm is compared with trending data transmission techniques
2015	Esma Yildirim, Engin Arslan et al [22].	Big Data Application-Level Optimization Transfers through Pipelining, Parallelism and Concurrency.	• Application-level transfer tuning parameters just like pipelining, parallelism and concurrency are very needful mechanisms for come across data transfer bottlenecks for scientific cloud applications. • Although their optimal values hinge on the environment in which basis the transfers are generally done. By using proper models and algorithms, these parameters can automatically optimized to achieve maximum transfer rate. • The paper proposes the optimization algorithms using these rules and models can provide a gradual increase to the highest throughput on inter-cloud and intra cloud transfers.
2015	Mayank Bhushan, Monica Singh et al [23].	Big Data query optimization by using Locality Sensitive Bloom Filter.	• Due to quick selection of data in network bloom filter has an significant part in searching technique. It operate data in less amount of time and frequently with statistical analysis. • The paper proposes an method to initiate Locality Sensitive Bloom Filter (LSBF) technique in big data and also discusses the uses of LSBF for big data query optimization.
2015	Yunxiang Li and Jiongjun Du [24].	Parameter optimization of the SVM in Big Data domain.	• The paper proposes the technique of sequential optimize parameters focused on the conspiracy plots of cross-validation accuracy. • It always guarantees the optimization of the SVM model on large data sets.
2015	Lancha Liu and Zhu Han [25].	Multi-Block ADMM for Big Data Optimization.	• The paper describes the parallel and distributed optimization algorithms based on alternating direction method of multipliers (ADMM) for resolving optimization problem in smart grid communication networks. • It has introduced the canonical formulation of the large-scale optimization problem and describes the general form of ADMM and focuses on several direct extensions and sophisticated modifications of ADMM to deal with the optimization problem.
2014	Nailah-Al-Madi, Simone A. Ludwigt et al [26].	Map Reduce optimization by cluster Glow-worm Swarm Optimization algorithm	• This paper proposes MRCGSO model which adjusts very well with enlarging data set sizes and scales very near to the optimal speedup.

V. Analysis

The Presented survey paper has analyzed the task assignment strategy, where all the map tasks can be assigned to the Task tracker containing the data input fragments for the tasks. Taking into account the regression based model can increase the performance and reduce the average response time of the map reduce applications effectively.

Table 3. Comparison between optimization techniques

TYPES	PERFORMANCE	ACCURACY	COMPLEXITY	COST
On-line Smart Grids Optimization on Big Data by Case-Based Reasoning.	High	High	Moderate	Low
Optimizing the performance of Big Data in Mobile Networks.	Moderate	Moderate	Low	High
Big Data Application-Level Optimization Transfers through Pipelining, Parallelism and Concurrency.	High	High	High	Moderate
Big Data query optimization by using Locality Sensitive Bloom Filter.	High	High	Low	Low
Parameter optimization of the SVM in Big Data domain.	Moderate	Low	Moderate	Low
Multi-Block ADMM for Big Data Optimization.	High	Low	Moderate	High
Map Reduce optimization by cluster Glow-worm Swarm Optimization algorithm	High	High	High	High
Convex Optimization for Big Data	High	Moderate	Low	Low

Table 4. Applications of optimization approaches

TYPES	APPLICATIONS
On-line Smart Grids Optimization on Big Data by Case-Based Reasoning	This optimization technique can be useful in banking sector, retail sector to optimize big data.
Optimizing the performance of Big Data in Mobile Networks.	This optimization technique is useful for optimizing the big data in telecom industry.
Big Data Application-Level Optimization Transfers through Pipelining, Parallelism and Concurrency.	This optimization technique can be helpful in insurance sector and banking sector for fraud detection by optimizing the big data.
Big Data Application-Level Optimization Transfers through Pipelining, Parallelism and Concurrency.	This optimization technique can be helpful in insurance sector and banking sector for fraud detection by optimizing the big data.
Parameter optimization of the SVM in Big Data domain.	This optimization technique is supportable for healthcare industry to optimize the big data.
Big Data optimization in Bioinformatics with Swarm Algorithms.	This optimization technique is widely used in bio medical industry for optimizing the big data.

According to the above table [3], table[4] On-line smart grids optimization technique provides high performance, gives accurate result but it follows a complex architecture [27].The optimization technique used in big data for enhancing the performance of mobile networks follows a simple architecture but unable to provides the accurate result [28]. However the application -level optimization of big data provides high performance , gives accurate result and follows a simple architecture[29].The Big Data query optimization by using Locality Sensitive Bloom Filter offers high performance , high accuracy and also follows a simple architecture [30] .On the other hand the optimization technique used for SVM in Big Data offers moderate performance ,low accuracy and follows a moderate level architecture [31] .While Multi block ADMM optimization technique gives high performance , low accuracy and follows moderate level architecture [32]. The Map Reduce optimization by cluster Glow-worm Swarm Optimization algorithm provides high performance, high accuracy, and follows a complex architecture.

VI. Conclusion

The survey paper discusses about various sources of big data causation and has been outlined based on multiple applications of big data. It has presented the adaptable characteristics of big data and also given some idea about various tools like NOSQL, MAPREDUCE, HADOOP which are required to process the big data. Although the presented survey provides an overview towards the basic fundamentals of big data analytics but focuses towards multiple optimization techniques used in MAPREDUCE framework and big data analytics .It proposes different optimization techniques to minimize the execution time of the map reduce applications. However the proposed survey has discussed the regression based optimization method for MAPREDUCE applications and given an idea about the scalable design and implementation of a clustering algorithm in MAPREDUCE.

VII. Future Scope

Big Data analytics is still in the initial stage of development, since existing Big Data optimization frameworks and tools are very restricted to solve the actual Big Data problems completely, Though, further scientific investments [33].From both governments and enterprises end should be contributed into this scientific paradigm to develop some new optimization techniques for big data and by observing different optimization techniques we want to propose an aggregator assisted hadoop framework to enlarge the performance of the hadoop framework by minimizing the execution time.

References

[1] HUI JIANG1, KUN WANG1, YIHUI WANG, MIN GAO, YAN ZHANG, "Energy Big Data: A Survey", Digital Object Identifier 10.1109/ACCESS.2016.2580581. Tim Mattson," HPBC 2015 Keynote Speaker - Big Data: What happens when data actually gets big", Parallel and Distributed Processing Symposium Workshop (IPDPSW), 2015 IEEE International.

[2] Carson K. Leung, Hao Zhang, "Management of Distributed Big Data for Social Networks", 2016 16th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing.

[3] Setareh Rafatirah , Avesta Sasan , Houman Homayoun," System and architecture level characterization of big data applications on big and little core server architectures " IEEE, 2015.

[4] Mohand-Sa ̈ıd Hacid, Rafiqul Haque, " Blinked Data: Concepts, Characteristics, and Challenge " , 2014 IEEE World Congress on Services.

[5] Nicolas Sicard, B ́en ́edicte Laurent, Michel Sala, Laurent Bonnet, " REDUCE, YOU SAY: What NoSQL can do for Data Aggregation and BI in Large Repositories " 2011 22nd International Workshop on Database and Expert Systems Applications.

[6] Karamjit Kaur and Rinkle Rani," Modelling and Querying Data in NoSQL Databases ", 2013 IEEE International Conference on Big Data.

[7] Richard K. Lomotey and Ralph Deters, "Terms Mining in Document-Based NoSQL: Response to Unstructured Data " , 2014 IEEE International Congress on Big Data.

[8] Eva Kurekov á, "Measurement Process Capability – Trends and Approaches", MEASUREMENT SCIENCE REVIEW, Volume 1, Number 1, 2009.

[9] Mehul Nalin Vora, " Hadoop-HBase for Large-Scale Data", 2011 International Conference on Computer Science and Network Technology.

[10] Apache Haddop HDFS homepage http://hadoop.apache.org/hdfs.

[11] Tom White, "Hadoop: The Definitive Guide", 1st edition, O'Reilly Media, June 2009, ISBN 9780596521974.

[12] Nagesh HR, Guru Prasad “High Performance Computation of Big Data: Performance Optimization Approach towards a Parallel Frequent Item Set Mining Algorithm for Transaction Data based on Hadoop MapReduce Framework” International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.1, pp.75-84, 2017. DOI: 10.5815/ijisa.2017.01.08.

[13] Siddharth S Rautaray, and Manjusha Pandey, “Single and Multiple Hand Gesture Recognition Systems: A Comparative Analysis”, I.J. Intelligent Systems and Applications, 6 (11), 57-65, 2014.

[14] Apache PIG Homepage - http://pig.apache.org/

[15] Apache Hive Homepage - http://hive.apache.org

[16] Apache Sqoop Homepage - http://sqoop.apache.org/

[17] Apache Zoo Keeper Homepage - http://zookeeper.apache.org/

[18] Jeffrey Dean and Sanjay Ghemawat , " Map Reduce: Simplified Data Processing on Large Clusters " , IEEE Micro, 23(2):2228, April 2005.

[19] Troiano, Luigi, Alfredo Vaccaro, and Maria Carmela Vitelli. "On-line smart grids optimization by case-based reasoning on big data", 2016 IEEE Workshop on Environmental Energy and Structural Monitoring Systems (EESMS), 2016.

[20] Ramaprasath, Abhinandan, Anand Srinivasan, and Chung-Horng Lung. "Performance optimization of big data in mobile networks", 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), 2015.

[21] Esma Yildirim, Engin Arslan, Jangyoung Kim, Tevfik Kosar. "Application-Level Optimization of Big Data Transfers through Pipelining, Parallelism and Concurrency", IEEE Transactions on Cloud Computing, 2016.

[22] Mayank Bhushan , Monica Singh , Sumit K Yadav ," Big Data query optimization by using Locality Sensitive Bloom Filter ",IJCT, 2015.

[23] Liu, Yunxiang, and Jiongjun Du. "Parameter Optimization of the SVM for Big Data", 2015 8th International Symposium on Computational Intelligence and Design (ISCID), 2015.

[24] Lanchao Liu and Zhu Han , " Multi-Block ADMM for Big Data Optimization in Smart Grid " , IEEE, 2015.

[25] Al-Madi, Nailah, Ibrahim Aljarah, and Simone A. Ludwig. "Parallel glowworm swarm optimization clustering algorithm based on MapReduce", 2014 IEEE Symposium on Swarm Intelligence, 2014.

[26] A. Ramaprasath, K. Hariharan, A. Srinivasan, “Cache Coherency Algorithm to Optimize Bandwidth in Mobile Networks”, Springer Verlag, Lecture Notes in Electrical Engineering, Networks and Communications, Chapter 24, Volume 284, 2014, pp 297-305.

[27] Ziv J., Lempel A., “A Universal Algorithm for Sequential Data Compression,” IEEE Transactions on Information Theory, Vol. 23, No. 3, pp. 337-343.

[28] E. Yildirim, J. Kim, and T. Kosar, “Optimizing the sample size for a cloud-hosted data scheduling service,” in Proc. 2nd Int. Workshop Cloud Computing. Sci. Appl., 2012.

[29] Mayank Bhushan & Sumit Yadav, “Cost based Model for Big Data Processing with Hadoop Architecture,” volume 14 Issue 2, Year 2014.

[30] Gunjan Varshney1, D. S. Chauhan2, M. P. Dave,” Evaluation of Power Quality Issues in Grid Connected PV Systems”, International Journal of Electrical and Computer Engineering (IJECE), Vol. 6, No. 4, August 2016, pp. 1412~1420.

[31] N.E. Ayat, M. Cheriet, C.Y. Suen, “Automatic model selection for the optimization of SVM kernels,” Artificial intelligence in medicine, vol. 38, no.10, pp. 1733-1745, 2005.

[32] Hamid Bagheri, Abdusalam Abdullah Shaltooki., " Big Data: Challenges, Opportunities and Cloud Based Solutions", International Journal of Electrical and Computer Engineering (IJECE), 2015.

Authors' Profiles

Bibhudutta Jena, is a CSI Accredited Student. Currently pursuing M. Tech (Computer Science and Engineering) at the School of Computer Engineering, KIIT University, Bhubaneswar. His areas of interest Data Analytics ,Data mining etc . He can be reached at bibhudutta93@gmail.com.

Mahendra Kumar Gourisaria is Professor at the School of Computer Engineering, KIIT University, Bhubaneswar. He has more than a decade of teaching and research experience. He has published numbers of Research Papers in peer-reviewed International Journals and conferences. His areas of interest include data mining and Cloud Computing. He can be reached at mkgourisaria2010@gmail.com.

Siddharth Swarup Rautaray, Ph.D (Computer Science), Member of IEEE is Professor at the School of Computer Engineering, KIIT University, Bhubaneswar. He has more than a decade of teaching and research experience. Dr. Rautray has published numbers of Research Papers in peer-reviewed International Journals and conferences. His areas of interest is Image Processing/DA/Human Computer Interaction. He can be reached at siddharthfcs@kiit.ac.in

Manjusha Pandey, Ph.D (Computer Science), Member of IEEE is Professor at the School of Computer Engineering, KIIT University, Bhubaneswar. She has more than a decade of teaching and research experience. Dr. Pandey has published numbers of Research Papers in peer-reviewed International Journals and conferences. Her areas of interest is WSN, Data analytics etc. She can be reached at manjushafcs@kiit.ac.in

How to cite this paper: Bibhudutta Jena, Mahendra Kumar Gourisaria, Siddharth Swarup Rautaray, Manjusha Pandey,"A Survey Work on Optimization Techniques Utilizing Map Reduce Framework in Hadoop Cluster", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.4, pp.61-68, 2017. DOI: 10.5815/ijisa.2017.04.07