Proc. of The Seventh International Conference on Advances in Computing, Electronics and Communication - ACEC 2018
Copyright © Institute of Research Engineers and Doctors, USA. All rights reserved.
ISBN: 978-1-63248-157-3 doi: 10.15224/978-1-63248-157-3-02

Pragmatic Influence of MapReduce in Big Data using Hadoop: A literature review

Muhammad Kaleem Ullah, Syed Khuram Shahzad

***Abstract*—A term "Big Data" explain innovative technologies and techniques to store, manage, capture, analyze and distribute large size or petabyte data sets with high acceleration and dynamic structures. Big Data is categorized as semi-structured, unstructured, and structured, which results in an inability of traditional data managerial techniques. Data is produced from numerous resources and reached the system at different rates. This immense amount of data is processed in an efficient, and inexpensive manner, a technique of parallelism is practiced. Big Data parameters which include diversity, scale, and complexity required new techniques, architectures, analytics, and algorithms for the purpose of management of data and the knowledge hidden in it. Hadoop is a famous software platform to make data useful for the purpose of analytics to solve problems and structured the Big Data. Distributed processing is enabled by the help of Hadoop for huge datasets across the bunch of servers. It is specially designed to scale from one to thousands of high computing machines, with high fault tolerance degree.**

***Keywords*— Big Data, Hadoop, Distributed Processing**

I. Introduction

A terminology "Big Data" depicts various set of data or arrangement of data sets in which rate of growth, variability and volume are the factors that influence in making them challenging to be processed, captured, examined or managed by chronological tools and techniques, like visual packages, relational database (RDB), and desktop statistics in time to make this helpful. To determine whether a data set is considered Big Data or not, inflexible definitions are given and they change with the passage of time. Currently, most practitioners and experts considered the datasets as Big Data which ranges from 30-50 tera-bytes to several peta-bytes [22]. In figure 1.1 explains the architecture of Big Data consists of multi-layered. It is divided into three levels i.e. are application layer, computing layer, and infrastructure layer.

In fig. 1 colour coding scheme is used to explain the layered architecture of Big Data system. it is divided into three layers, first one is system layer, second is computing layer, third is infrastructure layer.

Muhammad Kaleem Ullah
Superior University Lahore
Pakistan

Syed Khuram Shahzad
Superior University Lahore
Pakistan

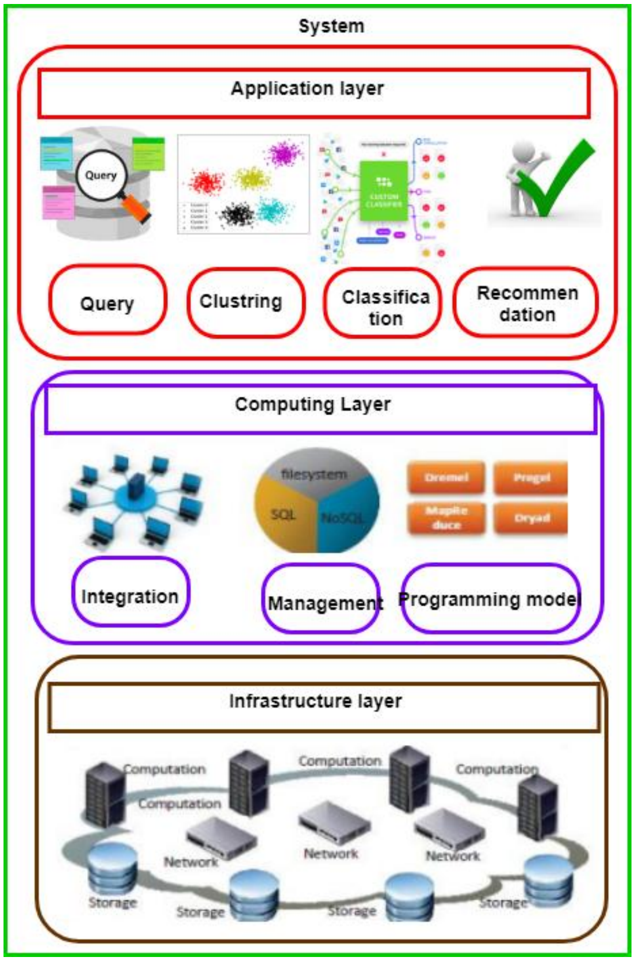


Figure 1- layered Architecture of big data system

A. *Three V's of Big Data*

These 3 V's are the pillars of the Big Data, which include volume, variety, and velocity. Each of these are explained below.

I. ***Volume*** depicts the quantity of data. Noticeably, data stored in databases has increased from megabytes to petabytes.

II. ***Variety*** is usually defined as different sources and types of data. Explode of different data is from structured, unstructured, semi-structured, video, XML, audio etc.

III. ***Velocity*** is usually described as data processing speed. For time-dependent actions like capturing

Proc. of The Seventh International Conference on Advances in Computing, Electronics and Communication - ACEC 2018
Copyright © Institute of Research Engineers and Doctors, USA. All rights reserved.
ISBN: 978-1-63248-157-3 doi: 10.15224/978-1-63248-157-3-02

the crime, Big Data necessarily applied because it flows in your business with the purpose of maximizing its value.

B. *Issues in Processing of Big Data*

There are some major issues that we have listed which are usually faced in the processing of Big Data, like completeness and heterogeneity, scale, privacy, and human collaboration.

I. For Incompleteness and Heterogeneity

A considerable heterogeneity is easily permitted when people use up information. In reality, the distinction and richness of native language can give useful depth. While the machine analytic algorithms cannot recognize the distinction except identical data. Consequently, data analysis data should be structured accurately. If the systems can accumulate various items of same structure and size, they can perform utmost capability.

II. Scale

Definitely a word scale brings a knowledge of size to anyone who explores it. From the decades it is keenly noticed that management of a large amount of data and its rapid elevation in its volume is a serious issue. In former times, according to Moore's Law, this issue was reduced by processors acquiring fast speed, which results in providing resources required to manage increased data volume. However, there is a basic shift on-going in these days that volume of data is increasing quickly than resources and the speeds of central processing unit (CPU) are fixed.

III. Timeliness

Large data set for processing, requires a long time for analysis.

IV. Privacy

Data privacy is also a major concern, and it boosts in the situation of Big Data. There are hard rules controlling that what may or may not be performed in electronic medical reports. Especially in the United States, rules for other types of data are not more powerful. Peoples are afraid there for the misuse of their personal data. Maintaining data privacy is a sociological and technical issue.

V. Human Collaboration

Despite the immense advancements made in computing analytics, a human can understand many patterns smoothly while computer systems algorithms can take a hard time. In the current complicated world, mostly it takes many professionals from variant fields to know what is taking place. An analysis system in Big Data takes inputs from different human analysts and mutual investigation results. These numerous professionals can be isolated in time and space when it is extremely costly to gather the whole group collectively in one place. Data system helps their cooperation and admit distributed professionals input.

C. *Solution for processing of Big Data: Hadoop*

Hadoop is a framework for programming, and in distributed computing normally it is used in huge data sets. Google MapReduce is the developer of Hadoop. It is a framework to break down any application into different sections. Apache is the current Hadoop ecosystem and it contains Hadoop distributed file system (HDFS), Hadoop kernel, MapReduce, Zookeeper and Apache Hive. MapReduce and HDFS (Hadoop distributed file system) are described below.

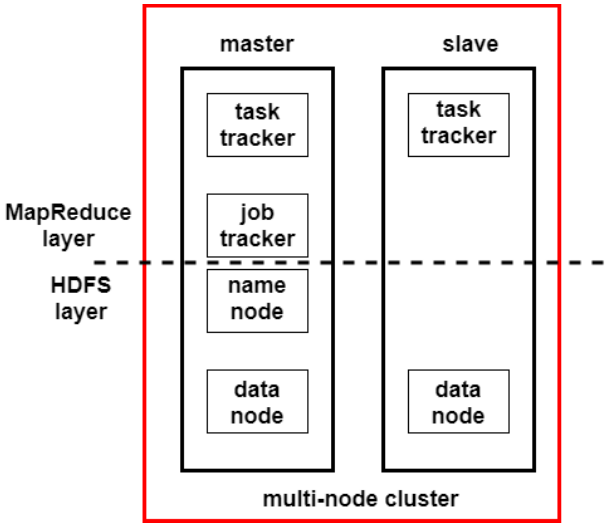


Figure 2 Hadoop Architecture

In fig. 2 Hadoop architecture is explained to give better understanding of the Hadoop software system. It is important to mention that it is divided into two layers one is MapReduce layer, and other is HDFS layer. There are two parts master and slave which include task tracker, job tracker, name node, data node.

D. *The Architecture of HDFS*

Hadoop distributed file system (HDFS) is a storage system of Hadoop. This storage system has the ability to scale and store massive information. Hadoop developed clusters of various machines and then combine their work. The cluster can be made with low-cost computer-systems. In the cluster, if there is the problem with one system, Hadoop remains working with the cluster without disturbing task or dropping data. Hadoop distributed file system (HDFS) handles storage by dividing input data into fragments ("blocks") and these blocks are stored across various servers. Normally, three copies of a file are saved in HDFS by copying them to three separate servers.

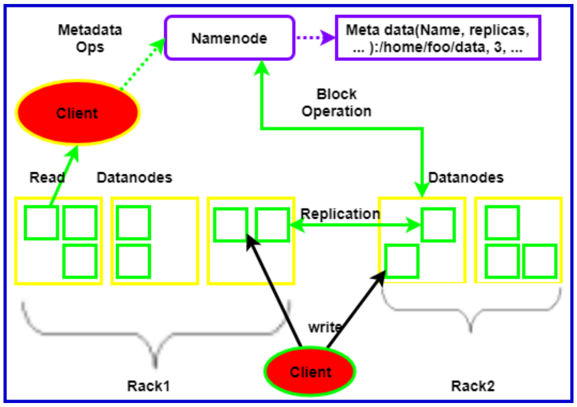


Figure 3 HDFS Architecture

In fig. 3 HDFS architecuter is explained, which consists of rack 1, rack 2, client, and namenode. Each rack contains datanods and able client to read and write into the racks. There are also block operations and metadata operations between racks rack2 – namenode and client – namenode.

E. *Architecture of MapReduce*

MapReduce is the processing backbone of Hadoop. It permits to break the data and problem and execute it in correspondence. According to research viewpoint, it can happen on varied sizes, like a huge data-set could be shortened into the reduced dataset for the analytics purposes. Such type of tasks is written in Java language as MapReduce Jobs. Pig and Hive are other languages, by using the program can be written easily. Their output is placed back to conventional data-warehouses or Hadoop distributed file system (HDFS). Major two functions of MapReduce are described below.

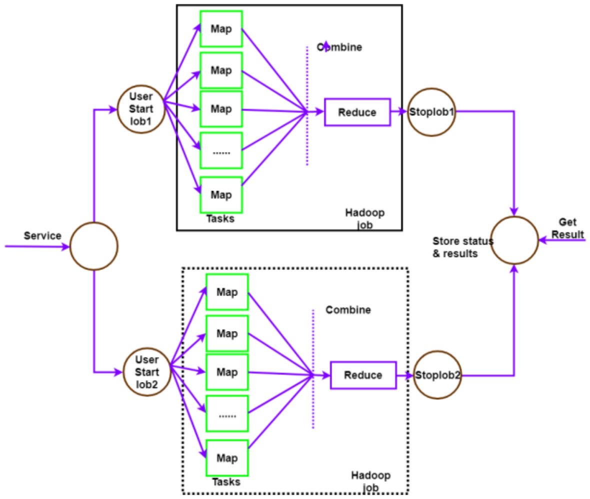


Figure 4 MapReduce Architecture

In fig. 4 there are two major components of the diagrams, which are Map's and Reduce in each Hadoop job. Two users' lobs, service point, two stop lobs, and the results point.

I. ***Map*** This function creates an intermediary set of values by taking input in the form value pairs.

II. ***Reduce*** This function takes the corresponding intermediary key and merges them with associated values.

II. Literature Review

In this section, we have explained the research work done by other researches in the domain of MapReduce using Big Data by selecting research papers from well reputed databases around the globe. We have also limit our search by considering research papers from year 2013 to 2018; to produce a good literature review.

Moreover, our research reviled that few researchers had done a similar work to others in the perspective of domain of MapReduce in Big Data, so we have also combined there researches to produce a good literature.

S. Vikram Phaneendra et al. explained that in older days data was in the small amount and easily tackled by the Relational Database Management System (RDBMS) but in recent times it is not easy to handle the large amount of data by RDBMS techniques, which is stated as Big Data. In this author also mentioned that Big Data is different from other data by 5 characteristics such as velocity, volume, variety, complexity, and value. They also illustrated that architecture of Hadoop is consisting of name data node, node, HDFS and edge node to deal with systems of Big Data. Application of Big Data can be found in financial, healthcare, retailing, mobility, and insurance. Authors also discuss the challenges that are faced by systems during handling of Big Data, such as search analysis, and data privacy, etc [1].

Kiran kumara Reddi et al. discussed that Big Data is the blend of unstructured, structured, semi-structured, heterogeneous and homogeneous data. The author also suggested the utilization of NICE MODEL for transmission of large amount of data on the network. The NICE MODEL utilized a forward and store approach by using servers staging. Model is also able to settle the difference between bandwidth and time zone variations. As a future work, it is also mentioned that new algorithm is needed for the transmission of Big Data and to resolve problems related to compression, and security [2].

Jimmy Lin et al. have used Hadoop which is in present time large-scale data analytic tool, but there are also algorithms that are not "nails" in the logic that they aren't agreeable to the MapReduce programming model. The author also focused on easy solutions to find different non-iterative algorithms that solve the same issue. The author suggested gradient, iterative graph, and EM iteration, which is normally implemented as a job of Hadoop with iteration and convergences check. The author also suggested that if you have the hammer, throw entirety away that is not a nail [3].

Wei Fan et al. described that Big Data mining is the ability to extract valuable facts from huge sets of data which was not possible earlier because of variety, velocity, and volume of data. According to the author, some controversies exist about BigData. For BigData processing there are certain tools like Apache S4, storm and Hadoop. PEGASUS and Graph are tools for the analysis of graph in BigData mining. Visualization and Compression are some issues discussed by the author[4].

Albert Bifet et.al discussed that real-time data streaming analysis is the effective and quickest method to get valuable information, letting organization to act suddenly when there is the issue. A large amount of data is generated daily which is called BigData. Some famous tools which are utilized for BigData mining are cascading, storm, apache-hbase, scribe, R, Hadoop etc. According to author our capability to manage the huge amount of data mainly depends on technologies, variant, and tools[5].

Bernice Purcell et.al stated that BigData consists of the large volume of data which is not possible to handle with conventional computer systems. BigData consists of unstructured, semi-structured and structured data. BigData use various methods for data storage including object base warehousing and NAS (network attached storage). Semi-structured and unstructured data is processed by Hadoop. MapReduce is used to find concurrent data which choose data having answers[6].

Sameer Agarwal et.al explained about BlinkDB which is a query engine for executing structured query language (SQL) queries for the huge amount of data in parallel processing. Dynamic sample selection procedure and adaptive optimization framework are the two basic concepts used in BlinkDB. According to dynamic sample selection data is selected on the basis of response time and accuracy of the query [7].

Yingyi Bu et.al introduced a new method termed as “HaLoop”. It is the advanced form of Hadoop. In Hadoop, MapReduce have deficiencies for iterative information. HaLoop facilitates repeated applications to be gathered from current programs of Hadoop without changing and it enhances their performance by the mechanism of inter iteration caching. Author has described implementation, evaluation, and design of HaLoop, which is a recent distributed and parallel system for the support of the analysis of such application that is iterative and large scale[8].

According to Shadi Ibrahim et.al partitioning of data may cause multiple issues such as data transfer and shuffling that may cause unfairness among different data nodes. However, the author has used LEEN and Hadoop techniques to overcome partitioning issues in MapReduce. The experimental analysis reveals LEEN achieve higher efficiency and reduce data shuffle rate as compared to Hadoop that guarantees fair distributions among reduced data notes. The quantities results show LEEN achieves up to 45% performance on various workloads [9].

Kenn Slagter et.al suggested a modified partitioning algorithm which enhances memory consumption as well as load balancing. This is accomplished by partitioner and algorithm for sampling. To measure the results of that suggested algorithm, its operations were comparatively checked with other current partitioning techniques. Workload division is dependent on the algorithm which is used for the data partition. Data sampling is one of the best technique to get rid of data skew problems. Data that is distributed by partitioner mainly relies on the samples and how accurately that samples are examined by partitioning techniques [10].

In another study, Chris Jermaine et.al has proposed an aggression system for large-scale computing. The purpose of the proposed system is to provide a very large scale, data specific and high-performance computing. One basic assumption that was carried out in this work is known as a non-shared environment. The proposed system also give a relationship between two identities known as MapReduce and cloud computing. One major advantage that we have from large-scale computations is related to safe computational resources and accuracy that are being ensured by proposed online large-scale computing[11].

The Jonathan Paul et.al has used variant model related to Bayesian inference to make the model in a ideal state. The proposed approximate solution produced a comparative results with standards estimations. However these estimation techniques solve the distribution problem in a rapid in easiest way related to Big Data 3V views [12].

Kyong-Ha Lee et.al has introduced a significant tool related to data processing that assists the database to understand various technical issues having in MapReduce framework as well as the open source communities.in his work author has also critics porn and corns and characteristics of MapReduce framework . in other contribution is to highlight open issues and challenges related to parallel data analysis in context of MapReduce [13].

Chen He et.al proposed a new method to schedule MapReduce for the enhancement of data locality of MapReduce tasks. Author has merged his technique with Hadoop fair scheduler and FIFO scheduler which is default scheduler of Hadoop. For the evaluation of his techniques, he did comparison with his method as well as without using his method on scheduling algorithm of MapReduce. Comparison findings depicts that his method mostly takes short response time and fast locality rate of data [14].

Vaibhav Pandey et al. has discussed that the MapReduced is a paradigm of computing for that reason this is used in the implementation of Hadoop using BigData. Author also discussed about the neglecting of scheduling algo in Hadoop. So, He discussed that how important the heterogeneity factors in terms of influential factor for performance and over all systems through put. In this survey article Vaibhav also discussed various challenges in the Hadoop designing. To overcome this problem he has discussed several schedulers in the literature. Moreover author also discussed other investigative techniques and environments for Hadoop schedulers [15].

Gabriel M. Alves et al. has obtained a sample of the soil by using tomographic technique which is based on many projections. Author believe that previously big data has shown remarkable potential in terms of data optimization, spotting trends of businesses, decision making in the domain of agriculture. Moreover he has also presented a technique to solve a problem of complexity parallelization by focusing on parallel programming or extreme programming and used MapReduce by considering big data. Gabriel also discussed the future work of his literature by mentioning; parallelization of other algo's for the reduction of time needed in regeneration of images and their 3D analysis [16].

Ashish Kumar Tripathi et al. has proposed a technique consists of three folds, 1. Improving quality of clustering, 2. Improving performances, and 3. A novel methodology. To enhance the clustering quality, author has introduced a Grey Wolf Optimizer technique and validate it by different data sets for clustering issues. Author also mentioned that technique is also used for clustering of huge data sets. Moreover author also mentioned that the proposed technique could be applicable for some realistic clustering applications having large data sets, such as satellite imaging analysis, twitter data analysis, and video analysis [17].

Rifki Sadikin et al. has discussed large data volume and big data solutions in the field of bio informatics for the reduction of computational time in data processing. Author has implemented MapReduce framework for the processing of sequencing of NEXTGen using library of Hadoop-BAM (Binary Alignment Map). Rifki Sadikin also discuss the importance of Hadoop MapReduce, which is it take less computation time and speeds up the large data processing. Moreover also mentioned the future work by stating that more studies are required to depict how it is beneficial for other type of computation in next generation of data processing [18].

Mais Haj Qasem et al. discussed the importance of big data in the domain of matrix multiplication application, he also mentioned that in his research work they have used MapReduce as new method to solve the matrix application problem. He also mentioned that the literature includes the review of techniques used for solving matrix multiplication by considering MapReduce, time complexities, and quantity of mappers for individual techniques. Finally author also mentioned that he has included research papers from year 2010 to 2016 [19].

Pratiyush Guleria et al. has discussed a connection between new emerging technologies implemented in the domain of educational system, larger unstructured data sets which are produced in result of implementation and Data Mining tools used to convert this unregulated information to structured data. Pratiyush also discussed how he used HDFS with MapReduce and results are aggregated to obtain the output [20].

Prathyusha Rani Merla et al. has analyzed the data of YouTube using Hadoop MapReduce. He mentioned that he has used Amazon web service for the Hadoop multi-node clustering. Prathyusha also discussed how data is obtained from API and stored into Hadoop Distributed File System (HDFS). Finally he discuss how data processing is done by the help of MapReduce. Author also discuss the future work by mentioning that the data must be transformed to decisions which has great influence in the real world. To show real world importance he also stated that it can be helpful for businesses to extract useful information from the unstructured data [21].

A. *Parts of Hadoop*

MapReduce and HDFS are two major parts of Hadoop, each has a master and slave which further has data node and task traker, only job tracker and name node are included in master part. These master and slave collectively become part of multi-node cluster.

To provide better understanding of the MapReduce in Big Data using Hadoop we have made a table of Components of Hadoop which is in practice nowadays.

In the Table 1 we did a comparative analysis among different components of Hadoop that are in practice nowadays. These key components are: HBase, MongoDB, Hive, Cassandra, and Drizzle. This comparative analysis among mentioned components is carried out based on the concepts of concurrency, replication, consistency, and durability.

TABLE I. COMPONETNS OF HADOOP

Name of Concepts	**Table Column Head**					
	MongoDB	***HBase***	***Redis***	***Drizzle***	***Hive***	***Cassandra***
Database model	Document store	Wide column store	Key-value store	Relational DBMS	Wide column store	Wide column store
Concurrency	Yes	Yes	Yes	Yes	Yes	yes
Language	C++	Java	C	C++	Java	Java

Name of Concepts	**Table Column Head**					
	MongoDB	***HBase***	***Redis***	***Drizzle***	***Hive***	***Cassandra***
Consistency	Immediate consistency	Immediate consistency	-	-	Immediate consistency	Eventual consistency, immediate consistency
Replication Method	Master-slave replication	Selected replication factor	Master-slave replication	Master-Master replication, Master-Slave Replication	Selected replication factor	Selected replication factor.
Durability	Yes	Yes	Yes	Yes	Yes	Yes

III. *Conclusion*

This paper describes the concept and significance of Big Data along with 3V's (volume, velocity, and variety). We have also highlighted processing problems on Big Data as well as the technical issues related to Big Data. In the literature, we have identified challenges related to data scalability, heterogeneity, error handling, poor structure, provenance, privacy, and visualization; which are the key factor that we face at every stage of Big Data. We did this literature review on pragmatic influence of MapReduce in the domain of Big Data using Hadoop that is an open source software used for Big Data processing. Furthermore we are aimed to do a systematic literature review in the domain of Big Data as we have realized this is the most influential and hot domain of research in computer sciences nowadays.

References

[1] Phaneendra, S. and Reddy, E. (2013). Big Data- solutions for RDBMS problems- A survey. International Journal of Advanced Research in Computer and Communication Engineering, [online] 2(9). Available at: https://www.ijarcce.com/upload/2013/september/71-o-V_I_K_R_A_M_-big_data_solution_for_rdbms_problems.pdf [Accessed 24 Jun. 2018].

[2] Kiran kumara Reddi and Dnvsl Indira "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355} [Accessed 24 Jun. 2018]

[3] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013). [Accessed 24 Jun. 2018]

[4] Umasri.M.L, Shyamalagowri.D ,Suresh Kumar.S "Mining Big Data:- Current status and forecast to the future" Volume 4, Issue 1, January 2014 ISSN: 2277 128X [Accessed 24 Jun. 2018]

[5] Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) 15–20 DEC 2012[Accessed 24 Jun. 2018]

[6] Bernice Purcell "The emergence of "Big Data" technology and analytics" Journal of Technology Research 2013. [Accessed 24 Jun. 2018]

[7] Sameer Agarwal†, Barzan MozafariX, Aurojit Panda†, Henry Milner†, Samuel MaddenX, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data" Copyright © 2013ì ACM 978-1-4503-1994 2/13/04[Accessed 24 Jun. 2018]

[8] Yingyi Bu _ Bill Howe _ Magdalena Balazinska _ Michael D. Ernst "The HaLoop Approach to Large-Scale Iterative Data Analysis" VLDB 2010 paper "HaLoop: Efficient Iterative Data Processing on Large Clusters. [Accessed 24 Jun. 2018]

[9] Ibrahim, S., Jin, H., Lu, L., He, B., Antoniu, G. and Wu, S. (2013). Handling partitioning skew in MapReduce using LEEN. Peer-to-Peer Networking and Applications, [online] 6(4), pp.409-424. Available at: https://link.springer.com/article/10.1007/s12083-013-0213-7 [Accessed 23 Jun. 2018].

[10] Kenn Slagter • Ching-Hsien Hsu "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: 11 April 2013, © Springer Science+Business Media New York 2013. [Accessed 24 Jun. 2018]

[11] Pansare, N., Borkar, V., Jermaine, C. and Condie, T. (2018). Online Aggregation for Large MapReduce Jobs. Proceedings of the VLDB Endowment, [online] 4(11), pp.1135-1145. Available at: https://www.researchgate.net/publication/220538853_Online_Aggregation_for_Large_MapReduce_Jobs [Accessed 23 Jun. 2018].

[12] Jonathan Paul Olmsted "Scaling at Scale: Ideal Point Estimation with 'Big-Data" Princeton Institute for Computational Science and Engineering 2014. [Accessed 24 Jun. 2018]

[13] Lee, K., Lee, Y., Choi, H., Chung, Y. and Moon, B. (2012). Parallel data processing with MapReduce. ACM SIGMOD Record, 40(4), p.11. [Accessed 24 Jun. 2018]

[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 [Accessed 24 Jun. 2018]

[15] Pandey, V. and Saini, P. (2018). How Heterogeneity Affects the Design of Hadoop MapReduce Schedulers: A State-of-the-Art Survey and Challenges. Big Data, 6(2), pp.72-95. [Accessed 30 Jun. 2018]

[16] Alves, G. and Cruvinel, P. (2018). Big Data infrastructure for agricultural tomographic images reconstruction. 12th International Conference on Semantic Computing. [Accessed 30 Jun. 2018]

[17] Tripathi, A., Sharma, K. and Bala, M. (2018). A Novel Clustering Method Using Enhanced Grey Wolf Optimizer and MapReduce. Big Data Research. [Accessed 30 Jun. 2018]

[18] Sadikin, R., Arisal, A., Omar, R. and Mazni, N. (2017). Processing Next Generation Sequencing Data in Map- Reduce Framework using Hadoop-BAM in a Computer Cluster. 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE). [Accessed 30 Jun. 2018]

[19] Qasem, M., Sarhan, A., Qaddoura, R. and Mahafzah, B. (2017). Matrix Multiplication of Big Data Using MapReduce: A Review. IT-DREPS Conference, Amman, Jordan Dec 6-8, 2017. [Accessed 30 Jun. 2018]

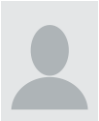
[20] Guleria, P. and Sood, M. (2017). Big Data Analytics:Predicting Academic Course Preference Using Hadoop Inspired MapReduce. 2017 Fourth International Conference on Image Information Processing (ICIIP). [Accessed 30 Jun. 2018]

[21] Merla, P. and Liang, Y. (2017). Data Analysis using Hadoop MapReduce Environment. 2017 IEEE International Conference on Big Data (BIGDATA). [Accessed 30 Jun. 2018]

[22] Gordon, K. (2013). What is Big Data?. ITNOW, 55(3), pp.12-13. [Accessed 30 Jun. 2018]

About Author (s):

He was born at Hasilpur in 1992, a small city of south Punjab, Pakistan. He did his B.Sc computer systems engineering from the university college of engineering and technology, I.U.B. He is doing MS software engineering from the superior university Lahore. Moreover he is interested to excel in the field of research & development with research interest in Big Data, health care, software engineering and data mining.



He was born and raised in Multan. Which is an industrial city of south Punjab, Pakistan. He did his B.Sc computer science from the Baha-ud-Din Zakariya University Multan, Pakistan. He did his MS computer science from Quaid-i-Azam University Islamabad, Pakistan. He did his PhD from Knowledge Technologies Institute, Graz University of Technology, Graz, Austria. Moreover his research interests are Knowledge Based Systems, Content Modeling and Visualization, Ontology & Semantics Information System, Semantic Web, Linked Open Data, HCI and Usability Engineering.