International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 9 Ver I || Sep 2018 || PP 10-16

A Systematic Review and Agenda for Future Research on Apache Hadoop Framework Tools

Kushal Kanwar[1],Vishal Shrivastav[2]
[1]Department of CSE & IT, Arya College of Engineering & I.T.Jaipur, Rajasthan
[2]Department of CSE & IT, Arya College of Engineering & I.T.Jaipur, Rajasthan
Corresponding Author; Kushal Kanwar

Abstract: *In this electronic life, growing number of organizations are facing the problem of a sudden increase in amount of data. Most of the data would decline under two categories – ordered and unordered. The more important piece of data which is not apprehensive but is hugely admired is the unordered data – likes, clicks, twitter posts, Face book likes, YouTube comments, videos etc. which has all supported development of a new race of database models. Big Data analytics implemented by Apache Hadoop is an open source frameworkthe most information handling engine. Thus; Apache open source Hadoop solves these issues by using multiple frameworks, which provides high speed clustered processing for the analysis of a large set of data smoothly and efficiently on the distributive environment.This paper is an effort to take a look at how Hadoop framework is prospect and designate the enforcement for a large range of evocative use cases, and then analyze, measure, and estimate so that application developers can make apprized choice based upon data size, cluster size, clone aspect, and dividing procedure to meet their performance demands for the unordered data and create it persistent perception to extend alteration.*

Keywords- Big Data, Hadoop, Mapreduce, HDFS.

Date of Submission: 31-08-2018 Date of acceptance: 15-09-2018

I. Introduction

As of late, the issue of huge measure of quick streaming information has picked up a considerable measure of consideration. Enormous Data is characterized as information that is too huge to fit on a solitary server, excessively unstructured, making it impossible to fit into a row and-column database, or too ceaselessly streaming to fit into static information distribution centers [15]. This information is giving enormous chances to reveal new perspectives. Volume, Velocity and Veracity [2], are three fundamental attributes which are utilized to characterize Big Data. Big Data research can be divided broadly into the scheduling of tasks and controlling the rate at which tasks are generating and running.

Apache Hadoop is gathering of open-source programming utilities that encourage utilizing a system of numerous PCs to take care of issues including enormous measures of information and calculation [14]. It gives a product structure to conveyed capacity and handling of huge information utilizing the MapReduce programming model. Every one of the modules in Hadoop are planned with a major suspicion that equipment disappointments are regular events and ought to be naturally dealt with by the system.

Hadoop is a generally received open source instrument which executes the Google's well known calculation show, MapReduce. It is a cluster handling Java based programming model which can process substantial measure of informational indexes in a circulated situation. Hadoop comprise of Hadoop Distributed File System (HDFS) [3], which is an appropriated record framework to store huge measure of information on group, and MapReduce which is a programming model for disseminated handling of information on bunch. MapReduce contains two client characterized capacities, outline lessen which takes after gap and overcome strategy by separating the unpredictable issue recursively.

The center of Apache Hadoop [5] comprises of a capacity part, known as Hadoop Distributed File System (HDFS), and a preparing part which is a MapReduce programming model. Hadoop departs documents into substantial squares and appropriates them crosswise over hubs in a bunch. At that point moves bundled code into hubs to process the information in parallel. This approach exploits information territory, where hubs control the information they approach. This permits the dataset to be prepared quicker and more effectively than it would be in a more ordinary supercomputer design that depends on a parallel record framework where calculation and information are dispersed by means of fast systems administration.

- The base Apache Hadoop framework is composed of the following modules:
- Hadoop Common – contains libraries and utilities needed by other Hadoop modules;

- Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;
- Hadoop YARN – a platform responsible for managing computing resources in clusters and using them for scheduling users' applications;
- Hadoop MapReduce – an implementation of the MapReduce programming model for large-scale data processing [6].
- The Hadoop framework itself is mostly written in the Java programming language, with some native code in C and command line utilities written as shell scripts. Though MapReduce Java code is common, any programming language can be used with "Hadoop Streaming" to implement the "map" and "reduce" parts of the user's program. Other projects in the Hadoop ecosystem expose richer user interfaces.

II. Literature Review

Jeffrey Dean Execution of MapReduce keeps running on an expansive group of product machines and is profoundly adaptable [1].an average MapReduce calculation forms numerous terabytes of information on a great many machines. Software engineers and the framework simple to utilize: several MapReduce programs have been actualized and upwards of one thousand MapReduce occupations are executed on Google's bunches each day. Projects written in this utilitarian style are consequently parallelized and executed on a substantial bunch of item machines. The run-time framework deals with the points of interest of apportioning the information, booking the program's execution over an arrangement of machines, taking care of machine disappointments, and dealing with the required between machine Communication. This permits software engineers with no involvement with parallel and conveyed frameworks to effortlessly use the assets of a substantial appropriated framework. Creator proposes Simplified Data Processing on Large Clusters.

S. VikramPhaneendra& E. Madhusudhan Reddy Shown that in past days the information was less and effectively dealt with by RDBMS however as of late it is hard to deal with immense information through RDBMS devices, which is favored as "large information". In this they told that huge information varies from other information in 5 measurements, for example, volume, speed, assortment, esteem and unpredictability. They delineated the hadoop design comprising of name hub, information hub, edge hub, HDFS to deal with enormous information frameworks. Hadoop engineering handle expansive informational indexes, versatile calculation logs administration utilization of enormous information can be discovered in money related, retail industry, medicinal services, versatility, protection [14]. The creators additionally centeredaround the difficulties that need to be looked by endeavors when dealing with huge information: - information security, look investigation, and so forth.

Aditya B. Patel, Manashvi Birla, Ushma Nair (6-8 Dec. 2012) "Addressing Big Data Problem Using Hadoop and Map Reduce" reports the test deal with the big information issues [2]. It portray the ideal arrangements utilizing Hadoop group, Hadoop Distributed File System (HDFS) for capacity and Map Reduce programming structure for parallel handling to process extensive informational collections.

EbinDeni Raj [17] talked about the architectures of Hadoop and constraints in it. These detriments are fathomed in MapReduce 2.0 or YARN. Their examination proposed structure to send a cover over YARN design.

R. Sandeep [18] additionally talked about the confinements of Hadoop as extensive number of duties lie on a solitary procedure. It causes the adaptability issues on the buncheswhere JobTracker needed to screen vast number of TaskTrackers, work accommodation and execution of guide and lessen assignments. The investigation clarified how YARN has developed as re-design of Hadoop to oversee and screen workload, keep up multitenant condition, execute security control.

Tyson CondieUsage of MapReduce keeps running on an expansive bunch of product machines and is very adaptable: a run of the mill MapReduce calculation forms numerous terabytes of information on a large number of machines [3]. Software engineers and the framework simple to utilize: several MapReduce programs have been actualized and upwards of one thousand MapReduce employments are executed on Google's groups each day. Projects written in this utilitarian style are consequently parallelized and executed on an extensive bunch of item machines. The run-time framework deals with the subtle elements of apportioning the information, booking the program's execution over an arrangement of machines, taking care of machine disappointments, and dealing with the required between machine Communication. This permits software engineers with no involvement with parallel and conveyed frameworks to effortlessly use the assets of an extensive circulated framework. Creator proposes Simplified Data Processing on Large Clusters propose a changed MapReduce design in which middle of the road information is pipelined between administrators, while safeguarding the programming interfaces and adaptation to internal failure models of other MapReduce systems. To approve this plan, creator built up the Hadoop Online Prototype (HOP), a pipelining adaptation of Hadoop. Pipelining gives a few imperative focal points to a MapReduce structure, yet additionally raises new plan challenges. To disentangle adaptation to internal failure, the yield of each MapReduce undertaking and occupation is emerged to circle before it is devoured. In this exhibit, we depict an adjusted MapReduce engineering that enables information to be pipelined between administrators. This broadens the MapReduce programming model past bunch handling, and can lessen culmination times and enhance framework use for cluster employments also. We show an adjusted adaptation of the Hadoop MapReduce system that backings online total, which enables clients to see "early returns" from an occupation as it is being figured. Our Hadoop Online Prototype (HOP) likewise bolsters constant questions, which empower MapReduce projects to be composed for applications, for example, occasion checking and stream handling.

Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) “Shared disk big data analytics with Apache Hadoop” Enormous information investigation characterize the examination of huge measure of information to get the helpful data and reveal the concealed examples. Enormous information examination alludes to the Mapreduce Framework which is created by the Google. Apache Hadoop is the open source stage which is utilized with the end goal of usage of Google's Mapreduce Model. In this the execution of SF-CFS is contrasted and the HDFS utilizing the SWIM by the facebook work follows .SWIM contains the workloads of thousands of occupations with complex information landing and calculation designs [4].

III. HADOOP: SOLUTION FOR BIG DATA PROCESSING

Hadoop is a Programming structure used to help the preparing of extensive informational collections in an appropriated processing condition [13]. Hadoop was created by Google's MapReduce that is a product structure where an application separate into different parts. The Current Appache Hadoop environment comprises of the Hadoop Kernel, MapReduce, HDFS and quantities of different parts like Apache Hive, Base and Zookeeper etc.

3.1 Data Storage Layer

HDFS [11], the capacity layer of Hadoop, is a dispersed, adaptable, Java-based document framework skilled at putting away substantial volumes of information with high-throughput access to application information on the network machines, giving high total data transmission over the bunch. At the point when information is pushed to HDFS, it naturally parts up into different squares and stores/imitates the information consequently guaranteeing high accessibility and adaptation to non-critical failure.

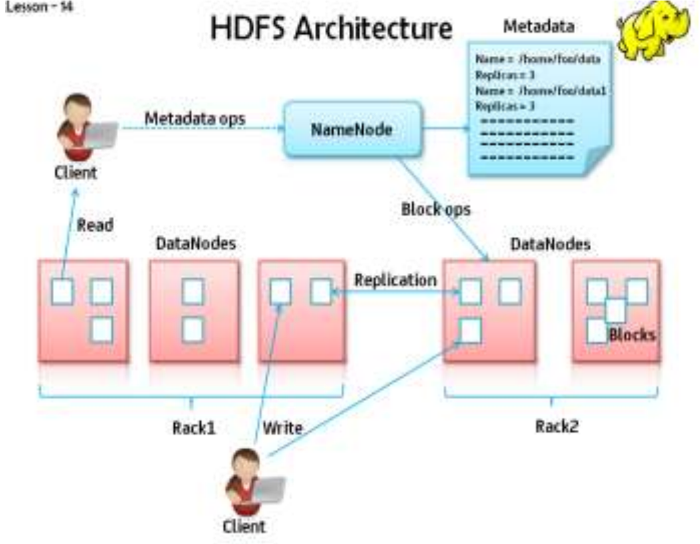

Figure1. HDFS Architecture

HDFS involves 2 vital segments (Fig. 1) – NameNode and DataNode. HDFS works on a Master-Slave engineering model where the NameNode goes about as the ace hub for monitoring the capacity group and the DataNode goes about as a slave hub summing up to the different frameworks inside a Hadoop bunch [8].

HDFS takes a shot at the compose once-read ordinarily approach and this makes it equipped for dealing with such colossal volumes of information with minimum potential outcomes of mistakes caused by replication of information. This replication of information over the group gives adaptation to non-critical failure and flexibility against server disappointment. Information Replication, Data Resilience, and Data Integrity are the three key highlights of HDFS

- NameNode: It goes about as the ace of the framework. It keeps up the name framework i.e., registries and records and deals with the squares which are available on the DataNodes.

- DataNodes: They are the slaves which are sent on each machine and give the real stockpiling and are in charge of serving read and compose demands for the customers. Furthermore, DataNodes speak with each other to co-work and co-ordinate in the record framework tasks.

3.2 Data Processing Layer

Booking, asset administration and bunch administration is planned here. YARN work planning and group asset administration with Map Reduce is situated in this layer.

3.2.1 MapReduce

MapReduce is a product system for disseminated handling of huge informational indexes that fills in as the register layer of Hadoop which process immense measures of information (multi-terabyte informational indexes) in-parallel on huge groups (a great many hubs) of item equipment in a dependable, blame tolerant way[7].

A MapReduce work for the most part parts the info informational index into free lumps which are handled by the guide errands in a totally parallel way. The structure sorts the yields of the maps, which are then contribution to the diminish undertakings.

The fig. 2 hubs and the capacity hubs are the same, that is, the MapReduce structure and the Hadoop Distributed File System are running on a similar arrangement of hubs. This setup enables the system to successfully plan errand on the hubs where information is as of now present, bringing about high total data transmission over the bunch. Part assignment on the slaves, checking them and re-executing the fizzled undertaking. The slaves execute the errands as coordinated by the ace [7].

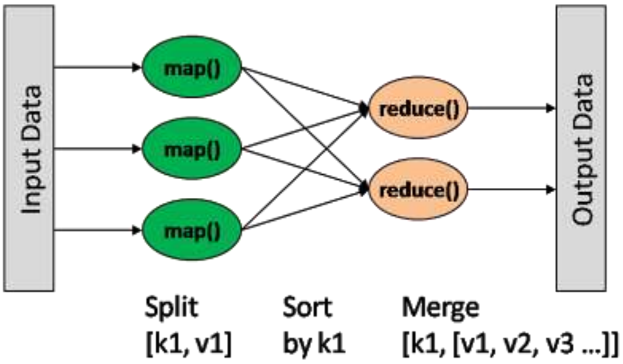

Figure2. MapReduce Architecture

Despite the fact that the Hadoop system is actualized in Java, MapReduce applications require not be composed in Java. Hadoop Streaming is a utility which enables clients to make and run employments with any executable (e.g. shell utilities) as the mapper and additionally the reducer. Hadoop Pipes is a SWIG-good C++ API to execute MapReduce applications.

3.2.2YARN

YARN (Yet another Resource Negotiator) shapes a vital piece of Hadoop 2.0.YARN is awesome empowering agent for dynamic asset usage on Hadoop structure as clients can run different Hadoop applications without bothering about expanding workloads. The incorporation of YARN in hadoop likewise implies versatility gave to the information handling applications.

YARN is a center hadoop benefit that backings two noteworthy administrations:
- Global asset administration (ResourceManager)
- Per-application administration (ApplicationMaster)

3.2.2.1 Resource Manager:

Asset Manager, in YARN design (Fig.3), is preeminent expert that controls every one of the choices identified with asset administration and portion. It has a Scheduler Application Programming Interface (API) that arranges and calendars assets. In any case, The Scheduler API doesn't screen or track the status of utilizations.

The principle motivation behind presenting Resource Manager in YARN is to streamline the usage of assets all the time by dealing with every one of the limitations, which include limit ensures, decency in distribution of assets and so forth. Hence, YARN Resource Manager is in charge of the considerable number of undertakings. Asset Manager plays out the entirety of its errands in incorporation with NodeManager and Application Manager [13].

3.2.2.2 Application Manager:

Each occurrence of an application running inside YARN is overseen by an Application Manager, which is in charge of the arrangement of assets with the Resource Manager. Application Manager additionally monitors accessibility and utilization of holder assets, and gives adaptation to non-critical failure to assets. As needs be, it is in charge of consulting for suitable asset compartments from the Scheduler, observing of their status, and checking the advance.

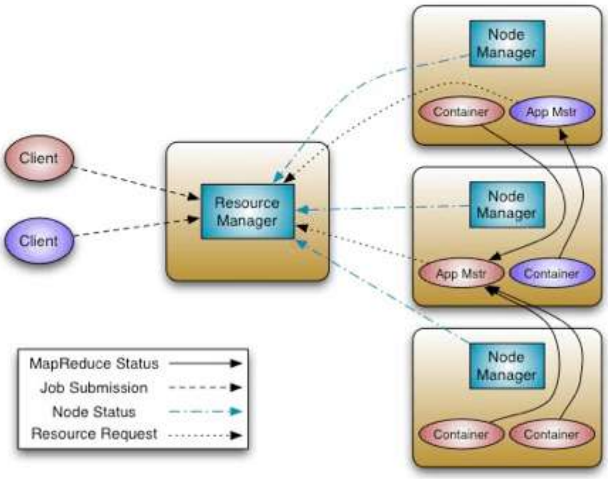

Figure3. YARN Architecture

3.2.2.3 Node Manager:

NodeManager is the per-machine slave, which is in charge of propelling the applications' holders, checking their asset use, and announcing the status of the asset utilization to the Resource Manager. NodeManager deals with every hub inside YARN bunch. It gives per-hub benefits inside the group. These administrations run from dealing with a compartment to observing assets and following the strength of its hub.
YARN benefits incorporate effective asset use, exceptionally versatility, past Java, novel programming models and administrations and readiness.

3.3 Data Access Layer

The Management layer where the request is sent to the Data Processing Layer. The Hadoop ecosystem [20] includes other tools to address particular needs. Hive is a SQL dialect and Pig is a dataflow language for that hide the tedium of creating MapReduce jobs behind higher-level abstractions more appropriate for user goals. HBase is a column-oriented database management system that runs on top of HDFS. Avro provides data serialization and data exchange services for Hadoop. Sqoop is a combination of SQL and hadoop. Zookeeper is used for federating services and Oozie is a scheduling system. In the absence of an ecosystem [12], developers have to implement separate sets of technologies to create Big Data solutions.

IV. VERSIONS OF HDAOOP

4.1 Hadoop 1

Hadoop 1.x supports just MapReduce (MR) preparing model. It does not bolster non-MR apparatuses.MR does both preparing and bunch asset administration. It has restricted scaling of hubs to 3000 hubs for each group. 1.x Works on ideas of spaces – openings can run either a Map errand or a Reduce undertaking as it were. A single Namenode to deal with the whole namespace.1.x has Single-Point-of-Failure (SPOF) – as a result of single Namenode-and if there should be an occurrence of Namenode disappointment, needs manual intercession to overcome.MR API is good with Hadoop 1.x. A program written in Hadoop1 executes in Hadoop1.x with no extra records.1.x has an impediment to fill in as a stage for occasion handling, gushing and constant activities.

Limitations of MRv1

Different Hadoop MapReduce businesses particularly assignments related with the science information, for example, genetic information, manage the game-plans similarities, super-groupings and sub sequences in

DNA. Such endeavours, if all else fails, require different MapReduce Jobs to get to relative information conventionally. For a DNA gathering arranging undertaking, if an n-nucleotide long movement exists in a particular DataNode, by then any superstring can be found in an indistinguishable DataNodes from it were [10]. Neighbourhood Hadoop has flexibility between assignments because fundamentally we can see that no condition between occupations. So in neighbourhood Hadoop same information dealt with such an expansive number of times. When we will process the vague information, again...again, it will cause the capability lessening and it is the weakness of the nearby Hadoop and decreases execution of the bundle.

4.2 Hadoop 2

Hadoop 2.x allows to work in MR as well as other distributed computing models like Spark, Hama, Giraph, Message Passing Interface) MPI &HBase coprocessors. YARN (Yet Another Resource Negotiator) does cluster resource management and processing is done using different processing models.2.x has better scalability up to 10000 nodes per cluster.It Works on concepts of containers and can run generic tasks. Multiple Namenode servers manage multiple namespace.2.x has feature to overcome SPOF with a standby Namenode and in case of Namenode failure, it is configured for automatic recovery.MR API requires additional files for a program written in Hadoop1x to execute in Hadoop2x.it Can serve as a platform for a wide variety of data analytics-possible to run event processing, streaming and real time operations [9].Apache Hadoop YARN system or MapReduce v2 propels past the clump preparing model of MapReduce to help close constant, forward and in reverse chainable, ceaseless handling model for Big Data [19]. YARN parts the two fundamental obligations of JobTracker i.e. asset administrator and employment planning for discrete daemons. Asset Manager and Node Manager oversee applications in the new form. A Per application ApplicationMaster is dependable to arrange assets from Resource Manager. It works with Node Manager to execute and screen holders and their asset utilization

4.3 Hadoop 3

Java rendition 8 is the base requirement as the majority of the reliance library record utilized is from java8. HDFS help for deletion encoding. (Eradication coding is a method for solidly putting away data).YARN course of events benefit v.2 (improved versatility and dependability). Here many new UNIX shell API, alongside old Bug Fixed. It Supports more than 2 NameNode. Map decrease turned out to be speedier, especially at delineate gatherer and rearrange occupations by 30% when contrasted with hadoop2.x. Hadoop now underpins combination with Microsoft Azure Data Lake as a contrasting option to Hadoop-perfect filesystem. New usefulness intra-DataNode adjusting is included, which is conjured through the hdfs plate balancer CLI. New strategies for arranging daemon stack sizes. Strikingly, auto-tuning is presently conceivable in view of the memory size of the host, and the HADOOP_HEAPSIZE variable has been belittled.

V. Conclusion

A review to enormous information challenges is given and different openings and uses of enormous information has been talked about. The paper describes Hadoop which is an open source programming utilized for handling of Big Data. This paper depicts the Hadoop Framework with 3 versions and its segments HDFS and MapReduce. The Hadoop Distributed File Framework (HDFS) is a disseminated record framework intended to run on item equipment. Hadoop assumes an imperative part in Big Data. This paper will help to improve understanding of Hadoop tools by comparing their performance using all 3 versions of Hadoop.

Acknowledgements

The authors would like to thank Dr. akhilPandey and Vishal Shrivastav for their thoughtful suggestion & timely advice. The authors are very grateful to all the writers & publishers of all the books & journal papers that have taken as references while conducting this research. The authors would like to thank their friends, juniors for their cooperation & motivation to complete the journal. Finally, the authors would like to extend their thanks to their parents whose continuous inspiration, sacrifice & support encouraged them to complete the journal.

References

[1]. Jeffrey Dean and Sanjay Ghemawat "**MapReduce: Simplified Data Processing on Large Clusters**" **OSDI 2010**

[2]. Aditya B. Patel, Manashvi Birla, Ushma Nair,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"

[3]. 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 978-1-3503-0032-2/10/06.

[4]. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"

[5]. Apache Hadoop Project, http://hadoop.apache.org/, 2013.

[6]. Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" in International Journal of Advance Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013.

[7]. Yu Li; Wenming Qiu; Awada, U. ; Keqiu Li,,(Dec 2012)," Big Data Processing in Cloud Computing Environments"

[8]. http://searchcloudcomputing.techtarget.com/ definition/Hadoop

[9]. K. Bakshi, "Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012

[10]. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013.

[11]. Bowen Tian, Yun Tian, Yijie Sun, Trevor Hurt, Brandon Huebert, Waymon Ho, Yuting Zhang, Danqi Chen, "A secure data allocation solution for heterogeneous Hadoop systems: SecHDFS", *Performance Computing and Communications Conference (IPCCC) 2016 IEEE 35th International*, pp. 1-8, 2016.

[12]. Longhua Guo, Mianxiong Dong, Kaoru Ota, Qiang Li, Tianpeng Ye, Jun Wu, Jianhua Li, "A Secure Mechanism for Big Data Collection in Large Scale Internet of Vehicle", *Internet of Things Journal IEEE*, vol. 3, no. 2, pp. 601-610, 2017.

[13]. Ralf Lammel, "Google's MapReduce programming model - Revisited", *Science of Computer Programming*, vol. 70, no. 2008.

[14]. 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).

[15]. T. Davenport, "Big Data at work: Dispelling the myths, uncovering the opportunities," in Harvard Business Review Press (2014).

[16]. H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient Scheduling Focusing on the Duality of MPL Representatives," Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.

[17]. E. D. Raj, Nivesh J.P., M. Ninnala and L.D. Dhinesh Babu, "A scalable cloud computing deployment framework for efficient MapReduce operations using Apache YARN," in Infonnation Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE (2014).

[18]. R. S. Raj and G. P. Raju, "An approach for optimization of resource management in Hadoop," in Computer and Communications Technologies (ICCCT), 2014 International Conference on. IEEE (2014).

[19]. E.S. Chan, D. Gawlick, A Ghoneimy, Z. H. Liu, "Situation aware computing for Big Data," Big Data (Big Data), in 2014 IEEE International Conference on IEEE (2014).

[20]. Deepika P, Anantha Raman G R," A Study of Hadoop-Related Tools and Techniques", IJARCSSE, Volume 5, Issue 9, September 2015.

Kushal Kanwar1" A Systematic Review and Agenda for Future Research on Apache Hadoop Framework Tools"'International Journal of Engineering Science Invention(IJESI), vol. 7, no. 9, 2018, pp. 10-16