Enhancing Data Processing And Management For Big-Data Intensive Applications Based On Cloud Computing

R.Rengasamy,M.Chidambaram

Abstract: An important factor that acts as a hindrance to the greater adoption of clouds for scientific computing is data management. The reason behind this is that data-intensive scientific workflow does not possess support for handling data workflow. At present, cloud computing handles data workflow by employing application overlays which map the output of a specific task to another input of specific tasks which may be in pipeline order and this technique was enhanced by MapReduce programming such as Amazon Elastic MapReduce, Hadoop on Azure - HDInsight. To resolve these challenges for managing data, an approach is proposed which can enhance virtual disks. The proposed approach enhances data workflow and management based on the cloud platform.

Keywords: Big data; Data management; Cloud Computing; Map-Reduce; Hadoop; Data Workflow

1. DEMYSTIFYING BIG DATA ANALYTICS

Big data analytics is the process of analyzing a huge amount of data or large volume data sets to interpret useful information and also detect the hidden patterns that help the organization in making out important decisions in the business and industrial sectors. The data could be gathered from several sources that could be in the form of audio, videos and digital images, Internet of Things, social media and many more. The certain organization has been commercially using ETL and RDBMS databases as their advanced analytics process. Certain drawbacks come with the usage of RDBMS and hence it requires a Big Data analytics platform. The main idea behind this platform is to detect invisible correlations and identify hidden patterns, which may offer valuable insights and promote useful and important information. This information could be helpful for the organization is making its important decisions and could lead to the growth of their company. The volume of data entering into the cloud is rapidly increasing; therefore, providers must be able to preserve system availability as well as process a large volume of data. Due to the use of internet, the volume of data is increasing exponentially. The data taken from various resources will be in unstructured format, which is found to be 50 times more than the structured format data. Due to the predefined framework of unstructured data, they cannot be fit into the relational tables.

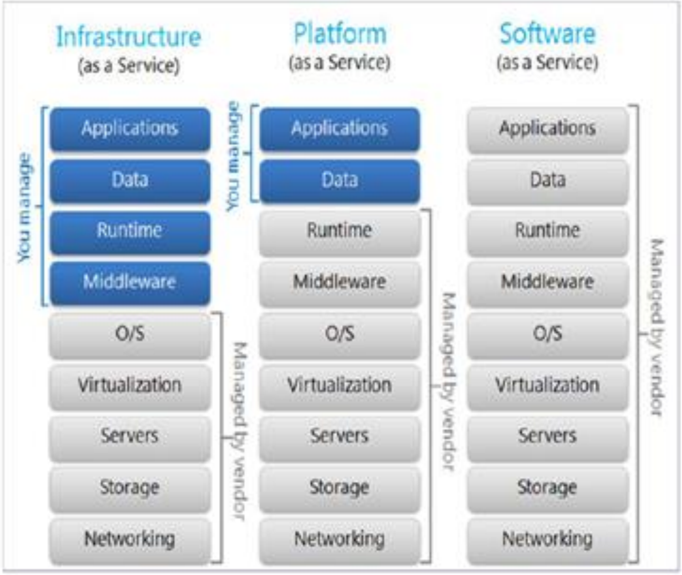

Figure 1: *Cloud Computing Services [1]*

These unstructured data will depend on spatial and temporal information that are considered to be heavy-weight. The techniques involved are very hard for the human understanding when compared to the commercial computer techniques, which is shown in figure 1. In order to solve the difficulty issues, the unstructured data should be converted to their structural format. The main advantage of big data is that the computational cost and storage facility of data will be very low [2]. Relational Databases had been used widely for the storage of complemented data before the advent of big data technology. In this traditional method, unstructured data are neglected. The main concept of big data is to create a whole enterprise architecture that integrates its technique along with the business values [3], thereby taking real-time decisions along with the enhancement in the market values. But, sufficient tools are required to systemize and obtain gain values from an obscured relationship and novice insights [4]. This could increase the production and promotes larger innovation which was provided by the MapReduce process as shown in figure 2. The Philosophy of Big Data can be classified into the generalized articulation of the concepts, theory, and

- *R.Rengasamy, Research Scholar A.V.V.M Sri Pushpam College, Poondi, Thanjavur, Tamilnadu, India .Email:rrsamy74@gmail.com*
- *M.Chidambaram Department of Computer Science Rajah Serfoji Government College, Thanjavur, Tamilnadu, India.*

applications, tools which consist of the overall conduct of big data science and another group is the impact of big data on individuals, society [5].

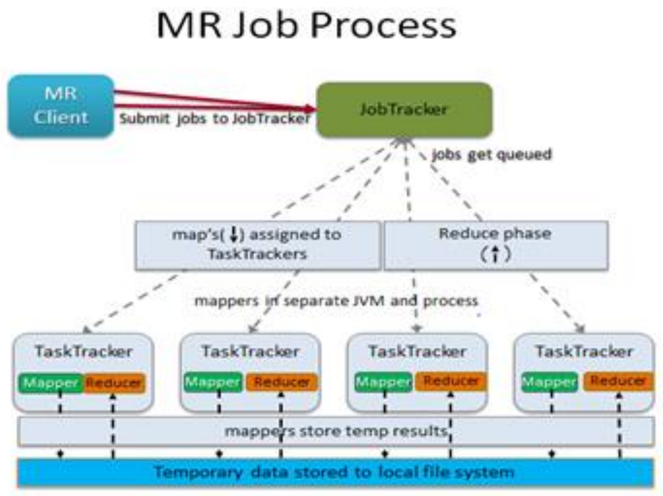
MR Job Process
MR Client
Submit jobs to JobTracker
JobTracker
jobs get queued
map's(↓) assigned to TaskTrackers
Reduce phase (↑)
mappers in separate JVM and process
TaskTracker
Mapper
Reducer
TaskTracker
Mapper
Reducer
TaskTracker
Mapper
Reducer
TaskTracker
Mapper
Reducer
mappers store temp results
Temporary data stored to local file system

***Figure 2**: MapReduce Process*

Cloud computing may be considered as a TCP/IP based high development and integrations of computer technologies for example fast microprocessor, huge memory, high-speed network, and reliable system architecture[6]. The TCP provides cloud computing with a secure base to achieve trusted computing. There is an interesting feature known as the maximum segment size option in TCP available on the Transport layer. This maximum segment size options feature allows the receiver as well as the sender to settle an agreement about the size of the segment which may be large or small. For instance, a small machine that consists of fewer resources can restrict a much bigger machine from sending segments that do not fit and cannot be handled by the small machine because of its larger size. Large segments are more reliable as well as efficient, therefore due to the above-stated reason large segments are most preferred[7]. An indiscriminately small segment will lead to less utilization of bandwidth; the reason for poor utilization of bandwidth is the overhead ratio of the data which will be very low. However, a very large segment will require extremely large IP Datagrams which in-return needs fragmentation and the throughput is decreased if the segment size is above the threshold of fragmentation. Therefore, it is best recommended to select the segment size with large IP Datagram which does not contain fragmentation during the transmission from the sender to receiver [8].

1.1 Microsoft Enterprise Cloud

One of the key products from Microsoft Enterprise Cloud is Nano Server for private clouds and its main objective is to be a lightweight operating system mainly employed as an OS layer for virtualized containers [9]. The term containers refer to faster starting, stateless instances in the OS kernel leading to less usage of resources since the containers do not depend on the OS[10] [11]. This advantage has gained major importance for deployments of faster work-load projects which employs cloud-native application as well as micro services-based components[12]. Nano Server is most compared with Ubuntu Core which is also a lightweight OS. There are many ways in which the Nano Server varies from the Server Core. The Nano Server is primarily built for 64-bit applications as well as tools. Nano Server lacks GUI features, Internet Information Services, Domain Name System as well as local log-on. The admins employ Windows PowerShell, Windows Remote Management along with Windows Management Instrumentation for specific state configuration for the container's settings. Unlike the Server Core, the Nano Server lacks Active Directory Domain Controller, group policy, network interface card teaming, proxy server access and important tool, namely, System Center Configuration Manager and Data Center protection Manager[13]. After the advent of Azure Containers by Microsoft for cloud, container has been the new buzzword for the companies and the main reason is that containers offer greater agility, since they load the right amount of kernel resources and run-time code to execute an application in the container, thus leading to flexibility among the operating environments [14]. Containers are widely employed in a public cloud, for example in Red hat's OpenShift PaaS. The containers are still evolving technology and it still needs major development in the field of security, load balancing as well as trusted connections [15].

2 CHALLENGES AND THREATS

The evolution of virtualization is going to continue for several years since it has found its application in almost every field of technology and most important field is cloud computing and mobile devices, its evolution has led to the new use-cases for smartphones. As mentioned by Gartner, virtualization in smartphones will increase by 60% in the upcoming years. Since virtualization can only solve the need for higher storage space in smartphones. Most advanced operating systems contain a high level of functionality, thus leading to poor security mess and virtualization solves this challenge, for instance, the virtualization provides better security by partitions[16]. Cybersecurity attacks in big data mainly, advanced persistent threats are increasing nowadays, these Cybersecurity attacks in big data hurt organizations of all types. The recent World Economic Forum considered Cybersecurity attacks in big data as an important challenge and must be addressed in an effective as well as a comprehensive way by international communities and organizations [17]. Generally, cyberattacks can be classified into many groups, few important cyber attacks include hacking national infrastructure, malware attacks and internal misuse of hardware as well as and software malfunction. Therefore, a common definition cannot be derived from Cybersecurity attacks in big data. By integrating these methods, enterprises can tackle Cybersecurity attacks in big data:

- Compulsory monitoring and reporting and conducting an important assessment with threat analysis which can cover people, process, technology and information; and the fundamentals under one place [18].
- Recruiting experts who can respond to such cyberattacks report and find the source of attack by examining the situation of the attack.
- Provide infrastructure and resources from international communities such as distributing response and attack reports between investigators and inform the related stakeholders and refresh important information, controls, and systems.

Detecting and analyzing the cybersecurity incidents is vital since detecting the source is difficult and try to steal confidential data such as intellectual property. Therefore attacks can be non-destructive, unobtrusive. By implementing automated incident response, we can address threats which are evolving and tactics by determining escalating issues or troubleshooting during orchestration process when an incident is obtained from SIEM alert, then it must be recorded automatically and stored in the incident response platform which can analyze and control with security tools such as data Loss Prevention, intrusion detection. Cloud computing systems, as well as log analyzers, should report to experts [19].

3 PROPOSED APPROACH

The latest cloud architecture consists of computational nodes that are not connected to the storage nodes as well as any interaction among the computational nodes and storage nodes results in high latency computational nodes because of data access protocols. Most of these services target storage and does not enable transfers between arbitrary virtual machines precluding any intermediary data storage [20]. Cloud users have to pay for storage and transferring data in/out of these repositories including the cost of leasing the VMs [21]. In recent advancements, cloud providers started to include cloud storage as virtual volumes to the compute nodes, such as in Amazon EBS and Azure Drives. The optimal method is to exploit data locality during storing as well as transferring workflow data as shown in figure 3. One solution was provided by Gfarm as distributed storage solutions, implemented in Eucalyptus which works in the host OS residing in the physical node that can store data in the machine's local storage disks[22].

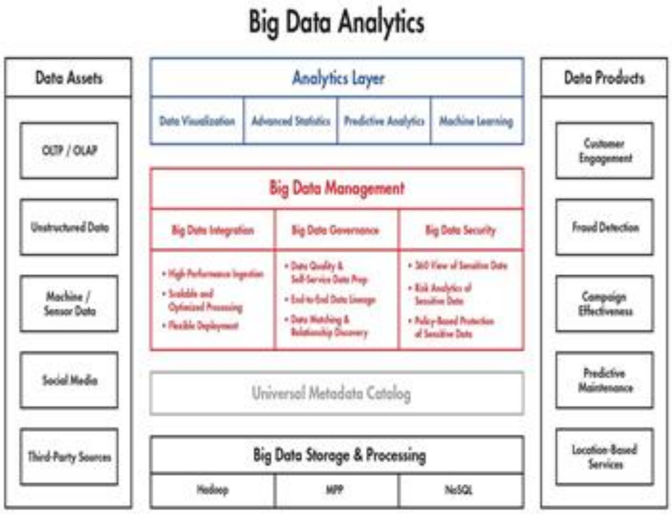

***Figure 3**: Big Data Management [30]*

After the advent of Big data, distributed storage and data management systems had to cope and upgrade their data volumes capacity as well as processing throughput [23]. Cloud storage services are still evolving to match up with big data analytics which can enhance functionality in a reliable and secure way. In-order to support and enhance data-intensive applications for big data applications in cloud computing, at large-scale, it raises the necessity to address the challenges [24]. The proposed techniques will focus on solving the major challenges as shown in table 1.

Challenges	Description	Technique to overcome
Scalability	Regarding the storage infrastructure, big data applications need to leverage a large number of resources effectively such as virtual disks as well as aggregate them elastically so that big data applications can adapt to the growing volume or veracity of the Big Data sets.	Big data can integrate with cloud computing for Big Data applications as long as the services can scale with the computation needs.
Parallel executions	In order to handle Big Data applications partition of computation needs to achieve a high degree of parallelism.	Big data applications can be solved by solving sub-problems. The final results can be aggregated. High-level algorithms can be developed for faster parallel executions such as Fuzzy C-mean for clustering of resources which is well suited for overlapped data set and comparatively effective than the k-means algorithm.
Low-latency and high throughput access to data under heavy concurrency	Big Data processing needs a huge number of a high degree of parallelism which can be employed for data analysis. Determining nodes concurrently access, process and share subsets of the input data needs high-throughput.	A method in which big data application instances accessing the cloud storage system needs to read the input data and write the results, then report their state for monitoring and fault tolerance purposes and to write the log files of their computation.

To resolve these challenges for managing data, an approach is proposed which can enhance virtual disks. The proposed approach enhances the local disk of the virtual machines by changing into a globally-shared data store. Therefore applications can employ local disk of the virtual machine instance which can share input files as well as record the intermediate data[[25][26]. This approach can be further extended into effective data management for the general workflow to enhance data locality for file transfers among the nodes. The observation based on the workflow produces a group of common data access patterns to choose the sufficient transfer protocol, hence increases the transfer rate by a factor of more than 1. Managing a common pool of storage space from a virtual disk in a distributed fashion in an aggregated fashion had been proposed in this paper. This pool helps in handling the data based on the application. Data, when engaged in a striped fashion enhances the load balancing and also scalability [27].This means that data is separated in the form of small chunks and is equally distributed among the local disk [28][29]. Failures could be avoided by replicating every chunk into the multiple local disks. This approach enhances the concurrency of reading and writes access performance since the global I/O workload is equally segregated among the local disks. A larger storage system could be possible by increasing the number of virtual machines, which also enables the data locality by reducing latency and increasing the scalability[30]. A certain set of principles is followed in designing the proposed approach that is selected in such a way to tolerate the contradicting constraints of cloud providers and progress in performing with high computational time. The proposed approach had been typically designed for the system with the concurrency-optimized PaaS-level cloud storage that adapts the following design principle in exploiting data locality. The system will depend on the local disk of the virtual machine so that the input files could be shared by saving the intermediate or output files. Through this technique, it is not required for the cloud middleware or the application to be changed. Moreover, any additional storage is not required by the system due to the availability of the virtual disks. Microsoft Azure cloud platform is used in this paper for implementing the following approach. There are three loosely-coupled components in the proposed approach architecture. The specific one is the Initiator component that plays the part in deploying, setting and launching the data management system more transparently. Additionally, it also plays a role in customizing the scientific environment, which is normally required by every application. This could be readily implemented and organized for any form of cloud API by exposing a generic stub. Prior knowledge is required regarding the infrastructure for most of the storage solutions. All the required information could be obtained by cloud middleware interaction and thereby required information could be obtained. The roles are assigned and the instances are differentiated by the set of parameters with the help of the initiator running inside each node of the cloud. Based on the user policies such as the storage entities, numbers, etc, the system could be deployed with the separate nodes. However, the elasticity of the system will be supported by the Initiator component and the computing platform could scale itself up and down in the system by enabling new nodes and thereby leaving the deactivated ones. The Local Storage Agent plays the role in combining the virtual disks into shared uniform storage that is expelled to applications. This is also non-special as it does not rely on any specified storage solution. As backend storage, any form of distributed file system could be used that should be capable of adapting the cloud infrastructure. This should not change the cloud middleware and thereby the storage services could be managed with the storage agents by composing the entities of the solution. The layers are represented by the client API that helps in viewing and accessing the storage by the applications. A set of primitives helps in accessing the application by BLOBs with the entire concurrency thereby supporting data manipulation with transparency though all can have access with the same BLOB. This is typically similar to the public cloud such as Amazon S3, Azure BOBs as one could obtain the data (READ) from the system, modify them within the specified range inside the BLOB (WRITE) and also include additional data to the BLOB (APPEND). When these operations are performed, all the interactions among the storage entities and the call could be hidden with the help of client API. VM local storage is mainly composed of virtual block-based storage devices; unlike the attached storage volumes it promotes access to the physical storage of the compute nodes that use the remote cloud storage in linking the virtual machines. VM takes the virtual disk as a device that could be accessed and formatted if they are known to be physical devices. But it is said to be an ephemeral storage option as they could only hold data till the VMs lifetime. It is cleared if the VMs are terminated. This does not support for the long-time storage option as this would signify leasing the computation nodes for long periods.

4 EVALUATION

The evaluation section presents results that discover the goal of the proposed approach which mainly concentrates on enhancing data management as well as resource allocation. The main focus is the throughput which is determined when a client carries out a group of operations on a dataset and its size will gradually increase. This scenario is suitable for the applications which consume more time when the process manipulates its data set independently of other processes.

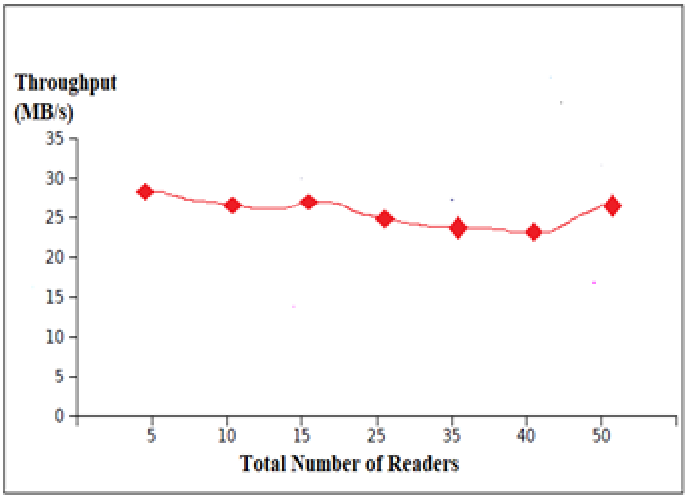

***Figure 4:** Read Throughput Performance of Proposed Approach*

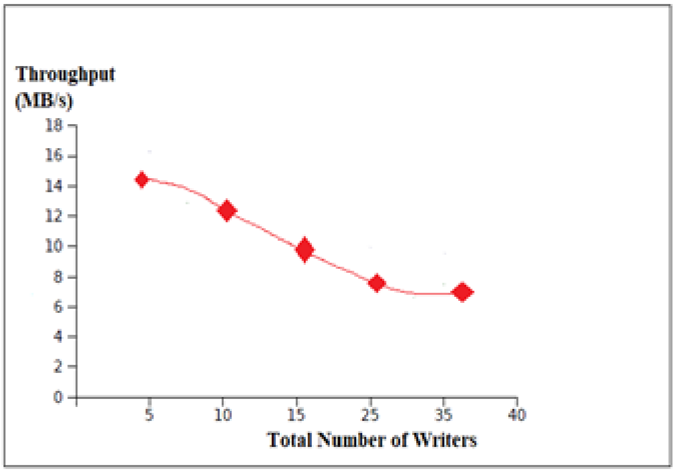

***Figure 5**: Writers Throughput of the proposed approach*

To examine the proposed approach, the Grid mix benchmark was employed, Grid mix benchmark is provided by Hadoop distribution. This includes combine, sort, and select. The whole experiment is done in a Hadoop cluster consisting of 2-way

64-bit 2.8GHz Intel Xeon machines that has 4GB of RAM and runs a 2.6.17 Linux kernel and connected by employing the Gigabit Ethernet network. The execution of data workload is performed proposed approach and its throughput are recorded and compared. A-Brain scientific application is employed for this purpose, as shown in figures 4 and 5, the experiment concentrates on the aggregated throughput under high concurrency for readers and writers of data. When data is fixed the size the input set increases from 30MB and it is evident that the data workload is processed up to 30 % faster. These results can be deployed into write-intensive maps as well as read-intensive reducers that can enhance data management by decreasing the overall execution time.

5 CONCLUSION

Google published a paper on the Map Reduce architecture that enables the processing of a huge amount of data with the parallel processing method in the year 2004. The queries based on the analysis are split up in the map step process and gets distributed across the parallel nodes and also proceed simultaneously. Reduce step process delivers and groups the query data. Map Reduce process is the most popular process in the Big data analytics and Hadoop also relies on the Map-Reduce architecture. These applications require a high-performance storage system that can allow VMs to acquire shared data concurrently. But most reference commercial clouds offer object stores for example in S3, as well as Azure Blobs can be accessed through high-latency REST (HTTP) interfaces. In such cases, the applications must be modified such a way data can be adapted for actual access technique.

REFERENCES:

[1]. Cloud Service : https://rajivramachandran.wordpress.com/2012/06/19/cloud-service-models-iaas-vs-paas-vs-saas/

[2]. Big Data Architecture https://en.wikipedia.org/wiki/Big_data#Architecture

[3]. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system." In: arXiv preprint arXiv:1603.02754 (2016).

[4]. K. Liu, H. Jin, J. Chen, X. Liu, D. Yuan, Y. Yang, A Compromised-Time-Cost Scheduling Algorithm in SwinDeWC for Instance-Intensive Cost-Constrained Workflows on a Cloud Computing Platform, Int. J. High Perform. Comput.Appl.

[5]. K. Kaur, A. Chhabra, G. Singh, Heuristics based genetic algorithm for scheduling static tasks in homogeneous parallel system, Int. J. of Comp. Sci. and Sec.

[6]. R.Rengasamy and M.Chidambaram, "A Novel Predictive Resource Allocation Framework for Cloud Computing", In Proceedings of International Conference on Advanced Computing and Communications Systems (ICACCS), 2019.

[7]. R.Rengasamy and M.Chidambaram "Challenges and Oppurtunities of Resource Allocation Frameworks for Big data Tools in Cloud Computing", International Journal of Computer Sciences and Engineering, Vol.6, Issue 12. Dec-2018, e-ISSN-2347-2693.

[8]. P. Sempolinski, D. Thain, A Comparison and Critique of Eucalyptus, OpenNebula and Nimbus, in: Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science, CLOUDCOM '10, IEEE Computer Society, Washington, DC, USA, ISBN 978-0-7695-4302-4, 417–426, doi: 10.1109/CloudCom.2010.42, URL http://dx.doi.org/10.1109/CloudCom.2010.42, 2010.population, Future Generation Computer Systems 27 (8) (2011) 1035–1046, ISSN 0167-739X, doi:10.1016/j.future.2011.04.011, URL http://dx.doi.org/10.1016/j.future.2011.04.011.

[9]. B. Palanisamy, A. Singh, L. Liu, and B. Jain. Purlieus: locality-aware resource allocation for MapReduce in a cloud. In Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2011.

[10]. J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," ACM Commun., vol. 51, Jan. 2008, pp. 107-113.

[11]. M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, and I. Stoica. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In Proceedings of the 5th European Conference on Computer systems (EuroSys), 2010.

[12]. Udendhran R, "A Hybrid Approach to Enhance Data Security in Cloud Storage", ICC '17 Proceedings of the Second International Conference on Internet of things and Cloud Computing at Cambridge University, United Kingdom — March 22 - 23, 2017, ISBN:978-1-4503-4774-7 https://dl.acm.org/citation.cfm?doid=3018896.3025138.

[13]. Suresh, A., Udendhran, R., Balamurgan, M. et al. "A Novel Internet of Things Framework Integrated with Real Time Monitoring for Intelligent Healthcare Environment "Springer-Journal of Medical System (2019) 43: 165. https://doi.org/10.1007/s10916-019-1302-9.

[14]. Udendhran R., Balamurgan M. (2020) An Effective Hybridized Classifier Integrated with Homomorphic Encryption to Enhance Big Data Security. In: Haldorai A., Ramu A., Mohanram S., Onn C. (eds) EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing. EAI/Springer Innovations in Communication and Computing. Springer, Cham

[15]. Suresh, A., Udendhran, R. & Balamurgan, M. " Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers " Springer - Journal of Soft Computing (2019). https://doi.org/10.1007/s00500-019-04066-4.

[16]. D. Kreuter, "Where server virtualization was born", Virtual Strategy Magazine, July 2004

[17]. Hive Performance Benchmarks. https://issues.apache.org/jira/browse/HIVE-396 .

[18]. S. L. Faraz Ahmad and T. V. Mithuna Thottethodi. MapReduce with communication overlap (marco). http://docs.lib.purdue.edu/cgi/viewcontentcgi?article=1412&context=ecetr, 2007.

[19]. Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing, STOC '00, pages 163–170, New York, NY, USA, 2000. ACM

[20]. Lun Li, David Alderson, John C Doyle, and Walter Willinger. Towards a theory of scale-free graphs:

Definition, properties, and implications. Internet Mathematics, 2(4):431–523, 2005. .

[21]. Tim Mather, Subra Kumaraswamy, and Shahed Latif. Cloud Security and Pri- vacy: An Enterprise Perspective on Risks and Compliance. O'Reilly Media, Inc., 2009. .

[22]. Xiaoqiao Meng, Vasileios Pappas, and Li Zhang. Improving the scalability of data center networks with traffic-aware virtual machine placement. In INFOCOM, 2010 Proceedings IEEE, pages 1–9, 2010

[23]. Issawi, S. F., Halees, A. A., & Radi, M. (2015). An efficient adaptive load-balancing algorithm for cloud computing under bursty workloads. Engineering, Technology, & Applied Science Research, 5(3), 795-800.

[24]. Jena, S. R., & Ahmad, Z. (2013). Response time minimization of different load balancing algorithms in cloud computing environment. International Journal of Computer Applications, 69(17), 22-27.

[25]. LaCurts, K. L. (2014, June). Application workload prediction and placement in cloud computing systems (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Cambridge Massachusetts.

[26]. Lee, R., & Jeng, B. (2011). Load-balancing tactics in cloud. In Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge CyberC Discovery, pp. 447-454.

[27]. Mahmood, Z. (2011). Cloud computing: characteristics and deployment approaches. In the 11th IEEE International Conference on Computer and Information Technology, pp. 121-126.

[28]. Mathur, S., Larji, A. A., & Goyal, A. (2017). Static load balancing using SA Max-Min algorithm. International Journal for Research in Applied Science & Engineering Technology, 5(4), 1886-1893.

[29]. Nema, R., & Edwin, S. T. (2016). A new efficient balancing algorithm for a cloud computing environment. International Journal of Latest Research in Engineering and Technology, 2(2), 69-75.

[30]. What is the effective way to handle Big Data? https://www.zarantech.com/blog/effective-way-handle-big-data/