Experimenting Cloud Infrastructure for Tomorrows Big Data Analytics

Amitkumar S. Manekar, G. Pradeepini

Abstract: ***Agile Cloud computing is today's need. Most of the business and enterprises already acquire cloud storage as their mainstream solution for data processing. These mainstream solution are already struggling with limited infrastructure of cloud computing. In day to day business processes large amount of data is generated and these data are migrated on cloud at the end of day. Today's cloud enables data processing, storage and distribution, system are not competent for moving large amounts of data in and out of the cloud. This actually is nothing but challenge for organizations with terabytes of digital content and managing cloud infrastructure for their daily real-time data crunching operation on demand. A high speed transport solution is required for transforming agility of on demand cloud infrastructure and merging big data analytics. Researchers need to developed such kind of mechanism where progressive data will be transferred on cloud with limited infrastructure of cloud computing. Today data migration is still carried out by managing hardware and dumping data on that hardware. No online solution and mechanism which can take care of this progressive data is available. Traditional hardware moving approach i.e. dumping data in hard drive or copy data in to magnetic tape and then transfer this data to cloud enable environment manually have many challenges associated it. In this paper we tried to explore opportunity of migrating big data to cloud infrastructure in optimized way. First part of this research work is discussing traditional data migration techniques and second part is experimenting these traditional data migration techniques on simulation based environment with ORTm algorithm which is modified version of ORT (Optimal Response Time). Here in ORTm we have focused on two parameter basically completion time of the last task and maximum resource utilization.***

Index Terms: Cloud, Data Migration, Migration Algorithm's, ORTm, Big Data Analytics.

I. INTRODUCTION

Big data is a really big term now days. Data has proven its importance from so many decades. All businesses are now a day ultimately running on this data science. Business without data science won't be imagined by any successful business tycoons. The word big data is derived from a complex, unstructured heterogeneous data huge in nature which is very useful for growth of enterprise or organization. Now a days Big data is not restricted to businesses only but also crossed there boundaries to social media, governments and economy etc. [1]. In every literature you will find consensus about the three V's [2] characterizing Big Data: *volume*, *variety* (different types of representations: structured, not-structured, graphs, etc.), and *velocity* (streams of data produced continuously). According to market research Big data or data science has already crossed to $5.8 billion by 2018 for supporting scientific data-intensive research [1][3]. Optimistic peoples and researchers looking double the expenditure for sustainable growth in upcoming year's i.e 2019-2020. To propagate data movement and processing needs, there is a growing trend amongst researchers within Big Data fields to frequently access remote specialized resources and communicate with collaborators using high-speed overlay networks which actually overwhelm of internet bandwidth. Intermediate network of network is connected and with so many electronics devices having end to end connectivity with limited bandwidth [4]. While developing and finding optimize solutions many cases where researchers have random/busty resource demands unfavorable or variable network bandwidth, they are considering to associate local resources with "on-demand" remote resources to form "hybrid clouds," versus just relying on expensive overprovisioning of local resources [5].

Revised Manuscript Received on March 10, 2019.

Mr. Amitkumar Manekar, Research Scholar, Department of CSE, KLEF, Vijaywad, A. P. India.
Dr. G. Pradeepini, Professor, Department of CSE, KLEF, Vijaywad, Andhra Pradesh, India.

Related Work

The cloud computing is a promising new paradigm which enables speedy on-demand availability of server resources (CPU, storage, bandwidth)to users, with minimal management efforts. Recent cloud platforms, as exemplified by Amazon EC2 and S3, Microsoft Azure, Google App Engine, Rackspace [12], etc., organize as hared pool of servers from multiple data centers, and serve their users using virtualization technologies. The elastic on demand nature cloud platform attractive for the execution of various applications, especially computation-intensive ones [13][14]. To fulfill on demand scalable big data transferring of business alighted data and chunk of data migration researchers think there must be a scalable end-to-end network architecture [6] also these mechanisms adapt QoS principal on demand. These developed such flow performance of data-intensive applications [7] are now a days are trending with fundamentally working on that simple WAN architecture developed for transmission of data on low bandwidth. rapid growth of emerging applications like social network analysis, semantic Web analysis and bioinformatics network analysis, a variety of data to be processed [8]. Effective data processing capability is a part of day to day life and it's a challenge for

organization for large generated data from their business or operations. Big Data a buzz words is not only more usable for talking about business growth and expansion but more understandable to computer also. For example, modern high-energy physics experiments, such as DZero1, typically generate more than one Terabytes of data per day. The famous social network Website, Facebook, serves 570billion page views per month, stores 3 billion new photos every month, and manages 25 billion pieces of content continues to witness a quick increase. Google's search and ad business, Facebook, Flickr, YouTube, and LinkedIn use a bundle of artificial-intelligence tricks; require parsing vast quantities of data and making decisions instantaneously [8]. On March 29, 2012, American government announced the "Big Data Research and Development Initiative" by keeping view on future market place. Numerously Big data will be a really BIG word and everybody is talking and tried to adopting the big data [9]. Current demanding and expanding computing platforms like grid and cloud computing have all planned to access large amounts of computing power by accumulating resources and offering a single system view. Among all powerful computing environments of today cloud computing is a most powerful and prominent architecture to perform large-scale and complex computing, and has revolutionized the way that computing infrastructure is abstracted and used. Big data and cloud computing are both the fastest-moving technologies identified in Gartner Inc.'s 2012 Hype Cycle for Emerging Technologies [10]. With adaptation of new technology cloud is acquiring stake of big data analytics on cloud infrastructure. Still some new cloud-based technologies should be discover for dealing with big data technology as big data concurrent processing is difficult. New paradigms of big data are emerging with business rapid changing environment. In contrasts Cloud computing is still dealing with basic infrastructure, security and management challenges. Cloud computing is a powerful technology to perform massive-scale and complex computing with delimiting the challenges of dramatically needed and maintained very expensive computing hardware, power, dedicated space and software. Today to expand big data a large computational infrastructure to ensure successful with data processing and analysis is required.

II. PROPOSED METHODOLOGY

Cloud Computing is a highly appreciable ICT revolution of 21st century. With remarkable outcome over a standard infrastructure of computer network and internet cloud environment provide sustainable avenues for migration of future business/operations on cloud based infrastructure. Cloud infrastructure is powerful to perform large scaling and complex computing. The advantages of cloud computing include virtualized resources, parallel processing, security, and data service integration with scalable data storage. Cloud computing can not only minimize the cost and restriction for automation and computerization by individuals and enterprises but can also provide reduced infrastructure maintenance cost, efficient management, and user access [11].In this research work efforts are taken to focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Lastly, open research issues that require substantial research efforts are summarized with an experimental setup by using Cloudanalyst [15] Simulation tool. With resources in a shared pool, illustrates the supply of computing resources from a cloud service provider are combined in a one massive assortment, for serving all users. The frequent provisioning of resources is employed for quickly matching the active resources, once a necessity comes for those resources. This frequent and quick provisioning prevents a scarcity of computing power once the requirement will increase. Virtualization and cloud computing were each developed to maximize the employment of computing resources whereas streamlining processes and increasing efficiencies to cut back the full value of possession[18]. Here primary challenges are transferring this huge and numerous data on cloud in time bound process. Second challenge is maintaining security of this data. Third challenge is maintaining integrity of this data and last but not least is adapting scalable resources in geographically distributed for progressive data. Based on these challenges planning of experiment is done. Here consideration of numerous continuous data random in nature is considered.

A. BLOCK DIAGRAM

Experiments are carried out by using java code on simulation software and results are monitor for changes happen with relative change in inputs of data size. In this section first of all we will discuss about the first challenge while transmisting numerous amount of data on cloud is time i.e the entire process is time bound. Hence cost will be relatively proportion to time required to compute data on existing cloud resources. Here we have to understand that any big Data Analysis required an ecosystem for managing and maintaining data. Figure 1 shows entire pipeline for data analysis.

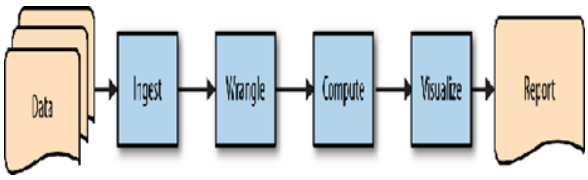

Figure 1: The data science pipeline

Data is captured from various sources and then ingest, wrangle, compute and visualized for preparation of reports. According to management of virtualized resources three cloud service models emerged: infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS). Hadoop like frame work/ technology plays a vital role in improving the quality of Data integrity and availability on single click for delivering right information to right people at right time and reduces its cost and time. Big Data is about

real-time analysis and data driven decision-making process [17]. There are certain stages in data migration as shown in figure 2 which represent block diagram of data migration. First is source and destination mapping, second is extracting data from sources, third is Intermediate staging data, fourth is load into destination table. Big Data migration refers process of Extracting Loading and Transferring large volume of data between commodity computing facilities. This commodity computing facility has their individual limitation for data process. Many companies/applications and institute wants on time data computing facilities. Maintaining and adapting new hardware and software for managing data is tedious and endless for most of the corporate users.

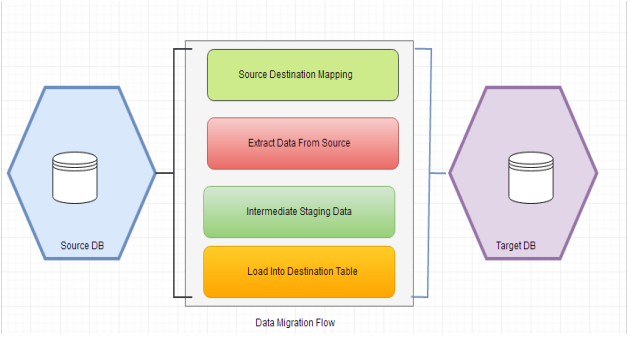

Figure 2: Data Migration Block Diagram

Cloud data center is high end servers and processing units. These data centers are connected by the special WAN links which are mainly fiber optic cables and can allow the data to be transferred at the rate of GBPS (Giga Byte per Seconds). In Cloud Computing if migration will be of an existing application may be moved to the Cloud (Cloud-enabling it) or designed from the beginning to use Cloud technologies (Cloud-native application).

B. ALGORITHM

In experimental setup consideration of all data centers are linked with each other with special data link having giga bytes of speed. These links are very tediously managing traffic and load across all data centers. Data centers are continuously sharing data among each other for synchronizations. Data centers are typically access when user wants to share a data or store a data on some data centers. At peak hours data request traffic is high. This request may divert to different data centers for managing all data center in optimizes in nature. Replication of data is done for backup and security of data. Change in one replica tiger updating among all replica. As shown in figure 3, data center DC1 and data center DC3 wants to share data among each other and wants to synchronize among each other with optimal way.

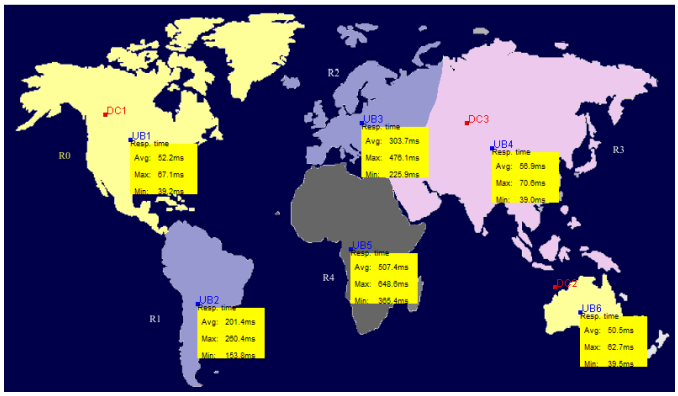

Figure 3: Data Centre Communication

Proposed algorithms start with assumption that each data center is assign with job and each virtual machine are keep busy in execution of individual job. Job distributer has pool of virtual machine bound to it. All jobs are in sorted order in ascending order based on the total execution time i.e JDB (J) ϵ P(VM). Job appears first will be serve to VM available. Unless and until there will be pool of VM is free. If Pool of VM is full then the average time of all VM will be get calculated and available job pool is checked. As is no room to allocated virtual machine is due to all virtual machines are busy in serving the job i.e P (VM) $\neq$ Avail. Job manager waits for job to be finished and keep request in Queue for future processing Q (J). Meanwhile requests are comparing with average completion time Avgas (CT) calculated for each and every VM. if Avgas(CT) in VM_m is less than the Avgas(CT) in VM_n, then the task will be sent to VM_m, else to VM_n

Algorithm ORT$_m$

Input UB(Sorted Req.)

1. ***Start***
2. *Inputs: Job= J1,J2,.....Jn*
3. *P (Vm1, VM2, VM3…… VMn)*
4. *Try JDB (J) ϵ P(VM)*
5. *If Yes Allocate P(VM) = J(X)and jump to 10*
6. *If No, for the case of P (VM) $\neq$ Avail*
7. *Then JDB (J) ->Q (J)*
8. *Calculate* Avgas(CT) of each VM
9. *If* Avg(CT) in VM_m < Avg(CT) in VM_n
10. *Avail Q(J(X)) to VM_{Less}*
11. *If P (VM) = Avail*
12. *Assign Q (J) -> P(VM) and mapped with J(x)*
13. *JDB (J) ϵ P(VM)*
14. ***End***

Output (Request Serve with optimal response time)

Table 1 represents some symbols used in algorithms and their nomenclature. This algorithm is basically devised for optimal load balancing and segregate the potential of scheduling in cloud computing.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Symbol	Meaning
ORT_m	Optimal Response Time Modified
UB	User Base Request
Job	Individual Job
JDB (J)	Pool of Request
P(VM)	Pool of Virtual Machine
Avgas(CT)	Average Completion Time
VM_n Or VM_m	Virtual Machines
Q (J)	Queue of Job

Table 1 Symbols and their nomenclature

The response time for cloud job scheduler is efficient in this algorithm as compared to other algorithms like RR (Round Robin) and SJF (Shortest Job First) empirically.

C. FLOWCHART

Simulation developed for simulating and calculating response time of ORT_m (optimal response time) algorithms having flow write from inception of request. Below flow diagram figure 4 shows the complete flow of how simulation work. First of all we need to initialize all UB and DC (Data Center). All physical characteristics of Physical machines need to map with each of the data centers. Here configuration of memory like RAM and processing unit are considered. In second step bandwidth is define, bandwidth is depended on the service provider and availability of bandwidth to each data center.

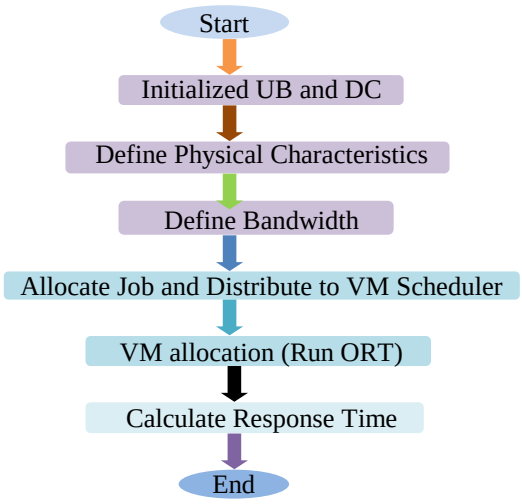

Figure 4: Flow Chart of Simulation

Here simulation work to allocated job initiated by user base which is already initialize in step 1. Now this jobs are mapped to individual data centers and VM scheduler keep track of each job initiate and mapped to VM with available physical infrastructure for optimize response time. If conflict occurs, in such situation where no VM is available then jobs are queued. Role of ORT in existence for fair policy of distribution is applied. Finally response time is calculated and simulation is end.

III. RESULT ANALYSIS

A simulation by using java as a platform is done on Intel i7 with 8 GB or RAM system for the internet users overwhelmed data at peak hours and results are noted. Basically around the globe several internet users are using internet. From the data collected from average statics in end of 2018 internet users per continents are shown in table 2. This data is collected for experimental purpose from and validation is subject to condition of producer of this data [19]. Here for experiments consideration continents based user (millions) is shown in below table 1 [19].

World Regions	Internet Users 30 June 2018	Growth 2000-2018	Internet Users %
Africa	464,923,169	10,199 %	11.0 %
Asia	2,062,197,366	1,704 %	49.0 %
Europe	705,064,923	570 %	16.8 %
Latin America / Caribbean	438,248,446	2,325 %	10.4 %
Middle East	164,037,259	4,894 %	3.9 %
North America	345,660,847	219 %	8.2 %
Oceania / Australia	28,439,277	273 %	0.7 %
WORLD TOTAL	4,208,571,287	1,066 %	100.0 %

Table 2: worldwide internet user's continent wise

Experiments are performed and results are collected basis on algorithms executed by using simulation. While simulations 4 continents are consider and on the basis of 10 user data base is considered from table 1. Four continent have their data centers located which continuously sharing synchronization data among all other data center. Geographically located data center have daytime natural phenomenon to support. When there is day in America at the same time night in Asia. This ultimately does load balancing in natural way for heavy traffic on data centers. Here each data center is having unique strength to deal with average no of users at a time is consider. The physical specification of data center is as follows intel xen- linux based systems having 4 processor and 200GB of RAM are considered.

Simulation is performed by implementing optimized response time (ORTm) as mentioned in section 3 as proposed algorithms with improved strategy of queuing mechanism for on time heavy load or user request. Optimized performance is captured with this algorithm. Below is table 2 showing the response time for different user based. Here we have considered cost of each VM is 0.1$ per hr. memory cost 0.05 $/Sec. Storage cost 0.1 $/Sec. Data transfer cost 0.1 $/GB. All data centers are having same parameter of physical specification.

User base	Avg (ms)	Min (ms)	Max (ms)
AMERICA	49.99	42.39	60.14
ASIA	50.23	39.14	62.63
AUSTRALIA	50.62	43.38	57.88
EUROPE	50.09	38.63	61.13
Overall response time	50.17	38.63	62.63
Data Center processing time	0.49	0.00	0.89

Table 3: computing request per user base region wise.

Table 3 shows execution time for each user base w.r.t region in average minimum and maximum in millisecond after execution of ORT_m.

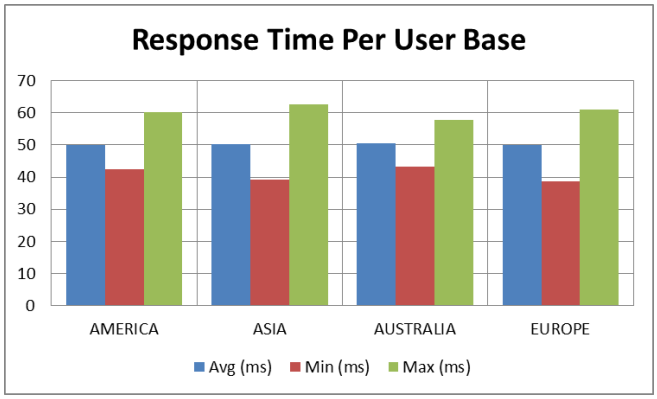

Figure 5 : Response time calculated by modified ORT algorithms.

Figure 5 shows that overall response time of each of the continent with relative user base considered in table 3. Here consideration of optimal response time with modified ORT algorithms is implemented. Overall data is calculated and cost is measure for users base.

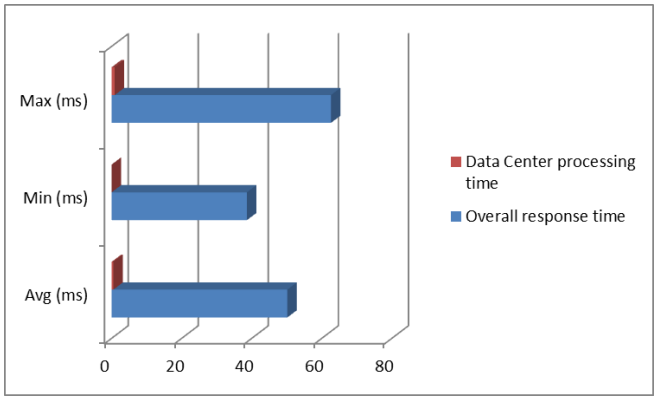

Figure 6 : Optimal Processing time for Data center with response time.

Overall response time is equal for all data center. With the help of simulation we have calculate overall cost for VM is 0.9 $ and total data transfer cost is 1.1 $ and overall cost is 2.1 $. Fig. 6 show optimal processing time for data centers with minimum response time. This shows cost of processing transaction on each data centers are relatively proportion to time processing. As a part of analysis, here an ORTm is always beneficial as compared to RR imperially. In RR algorithms unnecessary data processing of one data centers propagate delay in processing numbers of job based on result here analysis is drawn as (i) if service initiate and data centers are closed then the users improves the quality of service (average completion time in this case) (ii) to meet most efficient response time a well-defined necessary volume is required in the data centers to meet the peak demand. (iii) If the current response time is increasing and is greater than the best response time for the data center plus some pre-defined threshold the ORTm gives fair result in cost optimization.

IV. CONCLUSION

In result we have discovered some fact that overall cost for current scenario in which we are processing data in optimized way is not sufficient. Several different approached for optimizing cost can be implemented on data center site. Ultimately different cloud services have different cost with respective service and facility concern. Service quality can be further improved by the application of load balancing at the application level across data centers (by different service brokerage policies) and also at virtual machine level within data centers. But the levels of improvement achieved depend largely on the load balancing algorithms employed.

On the other hand if the peak volume is allocated throughout, there will be a substantial amount of the time where that capacity is not fully utilized. This is not economical. Here new solution governing VM migration in heuristic way should be the need. Prior analysis of data computation should be recognized. As a future of this research work comparison of all available algorithms like FCFS dynamic scheduling algorithms and RR will be evaluated with modified ORTm. In conclusion a strong point for future research is required to discuss that whenever user data base varies on peak time then a ultimate automated increase in the VM count by creating more VMs is necessary. At the same time when peak user data base reduce in idle time the reduce the VM count by releasing VMs mechanism need to implement for better result of ORTm.

REFERENCES

1. Shui Yu, Xiaodong Lin et. al. “Networking with Big Data” CRC Press Taylor and Francis Group ISBN 978-1-4822-6350-3, 2016.
2. D. Laney, 3D Data Management: Controlling Data Volume, Velocity & Variety, Technical Report, META-Group, 2001.
3. A. Robbins, Network modernization is key to leveraging big data, http://www.federaltimes.com
4. E. Dart, L. Rotman, B. Tierney, M. Hester, J. Zurawski, The science DMZ: A network design pattern for data-intensive science, Proceedings of IEEE/ACM Supercomputing, 2013.
5. A. Das, C. Lumezanu, Y. Zhang, V. Singh, G. Jiang, C. Yu, Transparent and flexible network management for big data processing in the cloud, Proceedings of USENIX Hot Cloud, 2013.
6. X. Yi, F. Liu, J. Liu, H. Jin, Building a network highway for Big Data: Architecture and challenges, IEEE Network Magazine, 28(4), 5 – 13, 2014.
7. L. Borovick, R. L. Villars, The critical role of the network in Big Data applications, Cisco White paper, 2012.
8. Changqing Ji, Yu Li et.al. “Big Data Processing in Cloud Computing Environments” 2012 International Symposium on Pervasive Systems, Algorithms and Networks, ISSN -1087-4089/12 IEEE DOI 10.1109/I-SPAN.2012.9 pp- 17-24
9. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-dea.
10. http://www.gartner.com
11. L. Chih-Wei, H. Chih-Ming, C. Chih-Hung, Y. Chao-Tung, And Improvement to Data Service in Cloud Computing with Content Sensitive Transaction Analysis and Adaptation, Computer Software and Applications Conference Workshops(COMPSACW),2013IEEE 37th Annual, 2013,pp.463–468
12. Rackspace, http://www.rackspace.com/.

13. M. Armbrust, A. Fox, R. Grifth, A. D. Joseph, R. Katz, A. Konwinski,G. Lee, D. P. A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," EECS, University of California, Berkeley, Tech. Rep., 2009.
14. S. Pandey, L. Wu, S. Guru, and R. Buyya, "A Particle Swarm Optimization(PSO)-based Heuristic for Scheduling Workflow Applications in Cloud Computing Environment," in Proc. of IEEE AINA, 2010.
15. Bhathiya Wickremasinghe ; Rodrigo N. Calheiros ; Rajkumar Buyya , "CloudAnalyst: A CloudSim-Based Visual Modeller for Analysing Cloud Computing Environments and Applications", 20-23 April 2010, DOI 10.1109/AINA.2010.32, pp-
16. S. Ghemawat, H. Gobioff, and S. Leung, "The google file system," in *ACM SIGOPS Operating Systems Review*, vol. 37,no. 5. ACM, 2003, pp. 29–43.
17. AWS Import/Export, http://aws.amazon.com/importexport/.
18. Cloud computing demystifying saas, paas and iaas. www.cloudtweaks.com/2010/05/cloud-computing-demystifying-saas-paas-and-iaas/, 2010.
19. https://www.internetworldstats.com/stats.htm

AUTHORS PROFILE

First Author Mr. Amitkuamr S. Manekar research scholar in KLEF formally known as KL university under noble supervision of Dr. G. Pradeeepini , Professor in CSE department of KLEF, Vijaywada , A. P. India. This research work is a part of research topic "Moving big data analytics in cloud".

Second Author Dr. Pradeepini Gera is Professor in CSE department of KLEF , Vijaywada , A. P. India. She has vast experience in teaching and guiding research in the field of Data mining, Big Data and Cloud Computing. She have many international Scopus , Web of Science and IEEE paper on her name.