International Journal of Engineering & Technology, 7 (2.7) (2018) 909-912

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper



Studying Cloud as IaaS for Big Data Analytics : Opportunity, Challenges

Amitkumar Manekar[1]*, Dr. Pradeepini Gera[2]

[1]Research Scholar CSE, KLEF, Vijaywad, A.P, India
[2]Professor CSE, KLEF, Vijaywad, A.P, India
**Corresponding author E-mail: asmanekar24@gmail.com*

Abstract

James Watt steam engine revolution was greatest revolution in mankind history in 20th century. In 1776, the first steam engines were installed and working in commercial enterprises. This revolution minimize and make world smaller for human being, now world is connected seamlessly. "Big Data Analytics and Cloud" these two words are second numerous revolutions in 21st century. We are living in an era of information explosion. These two magical terms are nothing but relatively very new and fortunately diverted all market trends to a new era of computation in last decade. As these two emerging technology are their early childhood, many people were confused with its relevancy and applicability. Cloud Computing is Infrastructure based solution for managing data and computational framework. 2016 was a significantly more important year for this volumes data technology or Big Data eco system as large number of enterprises, and organizations are generating data, storing that data and worried about future aspect of that data. In 2017, corporate world take cognizance of their large volumes structured and unstructured data as these enterprises and organizations continuously generating large volumes data. The term big data doesn't just refer to the massive amounts of data existing today, it also refers to the whole ecosystem of Storing or gathering data, Different types of data and analyzing that data. In traditional data ecosystem all leverages are with legacy system. Transforming or migration of these traditional ecosystems to the cloud is full of great challenges and benefits. Cloud computing is an agile and scalable resource access computation paradigm, provides heterogeneous platform seamlessly with infrastructure of internet, exclusively for the trapped and work on pre and post process of big data. Now the challenges are finding opportunity and challenges for managing, migrating and abstracting cloud based big data using cloud infrastructure for future eco system of Big Data Analysis. This paper is basically focused on this issue. We try to reevaluate the facts of existing Cloud Infrastructure as *IaaS* for tomorrow's big data analytics.

Keywords: *Big Data, Cloud, Big Data Analytics, Internet, IaaS, OLM, RFHC, Data Migration.*

1. Introduction

This Big Data is really BIG in nature. 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 day's big data is not concern to business or organization beyond that it expands to every human activity which generates large amount of data such as social media, government, economy etc. [1]. In every literature three V's about Big Data is well explored [2] in some of papers count of V's are more. Actual illustrating Big Data: volume, variety and velocity but this data is valuable also. Variety is associated with diverse natures of representations: structured, not-structured, graphs etc. Velocity associated with streams of data or progressive data and volume is associated with large amount of data. Market estimation related to scientific research is targeted $5.8 billion by 2018 [1][3]. These networks infrastructure are combination of several electronics devices, all these devices are connected with end-to-end circuit provisioning with limited bandwidth mechanism to share data and transmit data [4].While developing and finding optimize solutions many cases where researchers have random/bursty resource demands unfavorable or variable network bandwidth, hybrid cloud is nothing but on demand resource allocation facility form local to remote resource allocation [5]. Cloud computing infrastructure is considered as pool of virtualized resources. These virtualization techniques minimize cost in terms of required infrastructure and other artifact such as hardware, development platforms, and service. These pull of resources are managed in very optimal by implementing elastic in nature by allocation of dynamically assigning and configuring to adjust to a variable load. According to management of policy in elastic virtualization of resources broadly three categories of services are in acceptance namely, infrastructure as a service (***IaaS***), platform as a service (***PaaS***) and software as a service (***SaaS***).

Hadoop like frame work/ technology is a pure open source in nature and provide improved quality of Data integrity and availability on figure tip. Big Data facilitate real-time analysis with future prediction by data driven decision-making process [6]. This migration must be so agile in nature that it efficiently moving the data from one place to another geographically. Traditional hard drive movement is not flexible, nor secure in many ways [7].

Copyright © 2016 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

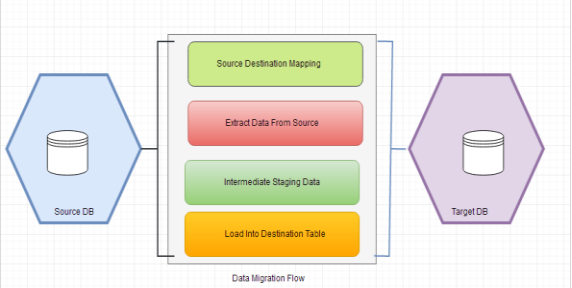


Figure 1: Data Migration Flow

There are certain stages in data migration as shown in figure 1. 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 Analysis required an ecosystem for managing and maintaining data. Figure 2 shows entire pipeline for data analysis.

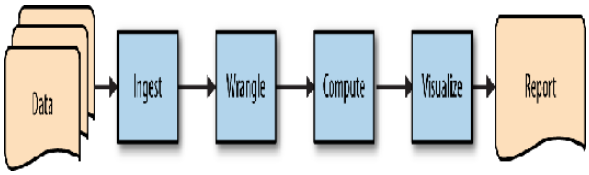


Figure 2: The data science pipeline

In this paper next subsequent chapter depict the issues and challenges faced during migration of big data to cloud infrastructure. The next chapter gives immense flavor of different literature available to finding facts related to challenges and opportunities for migrating big data. The next chapter is about discussion of facts findings and analyzing those facts. Finally a conclusion is drawn by keeping data migration as a operation on cloud based infrastructure and role of cloud infrastructure providing *IaaS*.

2. Literature Survey

Application logs are rich source of information. Logs need to be processed to get the meaningful information out of it. Log mining through hadoop [8] is a challenging task. Processing large amount of data at a single site gives better results. To achieve this, it requires transfer of large continuous progressive data from DC (Data Centre) to another DC. This paper is extension of our previous work [9][10]. Recent cloud platforms, as exemplified by Amazon EC2 and S3, Microsoft Azure, Google App Engine, Rackspace [11][12] etc. are works on virtualization technology as discuss earlier. Cloud elasticity nature in today's cloud platform for on demand resource allocation is mainly computation-intensive ones [13] [14]. Larger data processing by D. Kossmann et. Al. suggets four different architectures by using classic multi-tier database application like partitioning, distribution control and caching architecture [15]. Google File System (GFS) is a chunk-based distributed file system that supports fault-tolerance by data partitioning and replication [16].The current practice is to copy the data into large hard drives for physically transportation to the data center[17][18], or even to move entire machines [19]. This kind of data migration is very unsecure and includes undesirable delay with possible downtown [19]. Now major challenge is dynamic and continuous production of data from different geographical locations like sensor data or astronomical data [20]. Load balancing in cloud based environment is always a research topic with objective of resource allocation. In all sort of computing environment namely distributed, parallel and grid computing researchers also stated many static and dynamic scheduling strategies [21] [22]. Static algorithm's like scheduling algorithm namely ISH [23], MCP [24] and ETF [25] algorithms based on BNP is best for distributed environments with high internet speed and ignorable communication delay while MH [26] and DSL [27] algorithm based on APN take into consideration of the communication delay and execution time so they are suitable for larger distributed environments [21]. A new scheme OLACloud, which is modified for MapReduce structure, to progress the overall performance for running OLA in cloud was suggested by researcher [28][29]. The latest trend was associated with Open source framework and their utilization in transferring big data into cloud. MapReduce, Hive and many more are open source technology is available in market for migration. With this open source research many proprietary milestone was set by different company like Amazon Elastic MapReduce launches all processing nodes in the same EC2 Availability Zone [29][30]. These kind of solutions were very fruitful when data stored in one place. Here main challenges was clearly forget when data in chunk of data centers, inter-data center data moving in the stage of shuffle and reduction is main [30]. In this regard this paper firstly focuses on the geospatial data distribution and accessing overhead of that data in cloud structure. In extension of this research work a new theory will be proposed in future as later stages by our team with comparison of OLM and RFHC algorithm for optimize solution with MapReduce like solution.

3. Data Movement Analysis

All Now a day no companies or institute wants to spare cost on adapting uncertain data storage resources. Maintaining and adapting new hardware and software for managing data is tedious and endless for most of the corporate users. The best solution is managing data and application on Cloud Infrastructure. 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. On the other hand cloud Infrastructure is complex mechanism of different linked hardware and different managing artifact as software. CSPs (Cloud Service Provider's) distribute the data across the geographically distant data centers. Cloud data center is high end servers and processing units. To gain advantages cloud platform, including IaaS and PaaS, all techniques must eliminate the coupling of application to hardware. All data centers are linked with each other by a special link with giga bytes of speed. The links are responsible for managing traffics across all data center. Data centers are access when user request data of user wants to store data, beside of this data centers continuously shared and synchronized data among each other. At peak hours data request traffic is high. This request may divert to different data centers for managing all data center in optimizes way. For backup and security purposes data is stored at multiple locations in the form of replicas. These replicas need to update if data changes in one of the replicas.

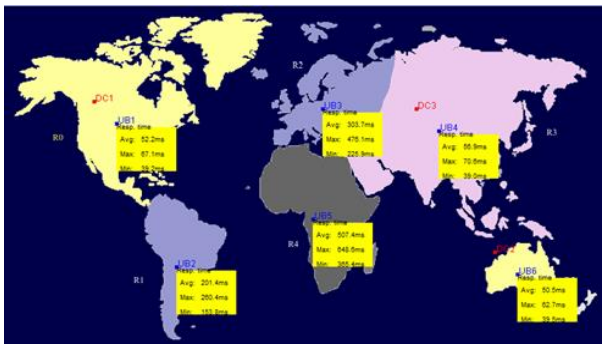


Figure 3: Data Centre Communication

As shown in figure 3, all data centers (DC1 to DC3) are communicate and synchronized with each other. In figure 3, world is divided into 6 different regions called as user base and we have

Considered 3 data centers from one CSP's. These entire data center are synchronized and communicate to each other in optimal way. Considering this existing traffic in the network, available bandwidth needs to be utilized efficiently. The processes such as log data analysis or the sensor data analysis need to be done in such a way that overall bandwidth cost can be reduced. We have identified some of the Core architectural principles of clouds to serve as IaaS for migration and tomorrow's big data analysis. Principal and their mechanism are explained in following table number 1 and 2. Currently there is a tremendous requirement of framework which can solve scalable and parallel processing of progressive data in commodity hardware. Table 1 is a summary of required parameter for seamless data transfer.

Table 1: Core architectural principles of clouds to serve as IaaS for big data analysis

Principals	Definition
On-demand self-service	In this cloud resources automatically bound with less interaction with service provider in physical and virtual way.
BNA(Broad network access)	In this a standard mechanism is facilitated to access all resources over network in diverse platform fashion.
MT(Multi-tenancy)	In this with multiple tenants allocation for all resources guaranteed but they are isolated from each other.
RP(Resource Pooling)	In this location-independent fashion is use to serve multiple multi-tenant customers in a simple way with on demand dynamic allocation and reallocation.
RES(Rapid Elasticity and Scalability)	In this on demand rapid, elastic, and auto scaled out up, out, and down mechanism is used for allocation
ME(Measured Service)	In this Service by virtue of use is targeted by monitoring, controlled and scale up.

Table 2 is a comparison based on variants of approaches in three types of migration namely IaaS, SaaS and PaaS. In table 1 the observation and fundamental analysis on some parameter are considered and counted in Table2. Finally there is a scope for the findings of migration strategies and an enhancement in VM Migrations, Online data migration strategies can be planed. Resultant shows that there is huge scope in IaaS as comparative to other migrations like PaaS and SaaS. IaaS is relative complex and further will be studied for proving the fact finding in this work in near future. Many companies and solution provider like Azure, Google and Rackspace are providing some template based solutions. Some of open source frameworks also support all these three Migrations. Still there is a scope of enhancement for managed lifecycle based framework by considering parameters in table 2. Security is a missing thread in all.

Table 2: Comparative Study of Three Types of Services

PRINCIPALS	MIGRATION to IaaS	MIGRATION to PaaS	MIGRATION to SaaS
Migration workload	Less	Moderate	Less
Migration complexity	Easy	Moderate	Easy
Adaption	No need	Compatibility Issue with Application	No need
On-demand self-service	Provided with Refactoring	Moderate Level	Provided
Broad network access	Present	Present	Present
Multi-tenancy	Yes	NO	NO
Resource pooling	Strongly Required	NO	NO
Rapid elasticity and scalability	Manage By strong Resource allocation policy	Less	NO
Measured service	Storage	Computing	Analytics

4. Conclusion

In these work findings about most advanced research in migration of Cloud as Infrastructure for future big data analysis has been performed till date but security issue is not notice. Architectural and legal issues related to this effort were identified by means of a motivating scenario for finding Cloud Iaas for futurist Big Data Analysis. Facts investigate cover different aspects of data migration to and from the Cloud and include some important aspects as a service from table 1 and 2. We have noticed not only the list of the migration scenarios, but also one of the patterns is claimed to be complete. There are future aspects as Cloud as Iaas for Big Data analytics strongly. The extension of this work will be simulated presented in near future as research work.

Acknowledgement

First Author is research scholar in KLEF under noble guidance of Dr. G. Pradeepini at CSE department. This work is a comparative study and analysis for the research work title "Big Data Analysis in Cloud, a Cost minimizing Approach" in KLEF, CSE Department. Subsequent part of this research work and result of simulation will be published in near future with new migration strategies and comparative study based on analysis shown in this paper with SaaS and IaaS.

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*, TechnicalReport, 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 designpattern 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 networkmanagement for big data processing in the cloud, Proceedings of USENIX HotCloud, 2013.
[6] Dasari Madhavi, B.V.Ramana"De-Identified Personal Health Care System Using Hadoop", International Journal of Electrical and Computer Engineering ,Vol. 5, No. 6, December 2015, pp. 1492~1499
[7] Linquan Zhang, Chuan Wu" Moving Big Data to The Cloud", 2013 Proceedings IEEE INFOCOM, ISSN -978-1-4673-5946-7/13 pp-405 to 409.
[8] B. Kotiyal, A. Kumar, B. Pant, R. H. Goudar, Big Data Mining of Log File through Hadoop, International Conference on Human Computer Interactions (ICHCI),2013.
[9] Prasad Teli, Manoj V. Thomas, K. Chandrasekaran, An Efficient Approach for Cost Optimization of the Movement of Big Data, Open Journal of Big Data, vol.1, no.1, pp. 4-15, 2015.
[10] C. Wilson, H. Ballani, T. Karagiannis, A. Rowstron, Better never than late: Meeting deadlines in datacenter networks.
[11] Bing Wu, Deirdre Lawless, Jesus Bisbal, Jane Grimson. 1997. Lagacy System Migration: A Legacy Data Migration Engine. 17th International Database Conference.
[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 Applicationsin Cloud Computing Environment," in Proc. of IEEE AINA, 2010.
[15] D. Kossmann, T. Kraska, and S. Loesing, "An evaluationof alternative architectures for transaction processing in thecloud," in *Proceedings of the 2010 international conferenceon Management of data*. ACM, 2010, pp. 579–590.
[16] S. Ghemawat, H. Gobioff, and S. Leung, "The google filesystem," in *ACM SIGOPS Operating Systems Review*, vol. 37,no. 5. ACM, 2003, pp. 29–43.
[17] WSImport/Export, http://aws.amazon.com/importexport/.
[18] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational Solutions to Large-scale Data Management and Analysis," Nat Rev Genet, vol. 11, no. 9, pp. 647–657, 09 2010.
[19] *Moving an Elephant: Large Scale Hadoop Data Migration at Facebook,http://www.facebook.com/notes/paul-yang/moving-an-elephant-largescalehadoop-data-migration-at-facebook/10150246275318920.*
[20] R. J. Brunner, S. G. Djorgovski, T. A. Prince, and A. S. Szalay,"Handbook of Massive Data Sets," J. Abello, P. M. Pardalos, andM. G. C. Resende, Eds. Norwell, MA, USA: Kluwer AcademicPublishers, 2002, ch. Massive Datasets in Astronomy, pp. 931–979.
[21] JianhuaGu, Jinhua Hu et. al. "A New Resource Scheduling Strategy Based onGenetic Algorithm in Cloud ComputingEnvironment", Journal of Computers, doi:10.4304/jcp.7.1.42-52 ,VOL. 7, NO. 1, JANUARY 2012 pp-42-52.
[22] Wei Wang, "A reliable dynamic scheduling algorithm based onBayes trust model," Computer Science, 2007.
[23] Rewinin H E, Lewis T G, Ali H H, "Task Scheduling in paralleland Distributed System Englewood Cliffs," New Jersey: PrenticeHall, 1994, pp. 401-403.
[24] Wu M, Gajski D, Hypertool, "A programming aid for messagepassing system," IEEE Trans Parallel DistribSyst, 1990, pp.330-343.
[25] Hwang J J, Chow Y C, Anger F D, "Scheduling precedencegraphics in systems with inter-processor communication times,"SIAM J Comput, 1989, pp. 244-257.
[26] Rewinin H E, Lewis T G, "Scheduling parallel programs ontoarbitrary target machines," J Parallel DistribComput, 1990, pp.138-53.
[27] Sih G C, Lee E A, "A compile-time scheduling heuristic for-Interconnection-constraint heterogeneous processor architectures,"IEEE Trans Parallel DistribSyst, 1993, pp. 175-187.
[28] Yu-Xiang Wang, Jun-Zhou Luo, Ai-Bo Song, FangDong, "Partition-Based Online Aggregation with SharedSampling in the Cloud", Journal of Computer Scienceand Technology, November 2013, Volume 28, Issue 6,pp 989-1011.
[29] Amazon Elastic MapReduce, http://aws.amazon.com/elasticmapreduce/.
[30] M. Cardosa, C. Wang, A. Nangia, A. Chandra, and J. Weissman, "Exploringmapreduce efficiency with highly-distributed data," in Proc. OfMapReduce 2011.