

# Opportunity and Challenges for Migrating Big Data Analytics in Cloud

Amitkumar S Manekar<sup>1</sup>, Dr. G. Pradeepini<sup>2</sup>

<sup>1</sup>Reserach Scholar KL University, <sup>2</sup>Professor CSE Dept. K L University, Vijaywada A.P, India.

<sup>1</sup>asmanekar24@gmail.com, <sup>2</sup>pradeepini.gera@gmail.com

**Abstract:** Big Data Analytics is a big word now days. As per demanding and more scalable process data generation capabilities, data acquisition and storage become a crucial issue. Cloud storage is a majorly usable platform; the technology will become crucial to executives handling data powered by analytics. Now a day's trend towards "big data-as-a-service" is talked everywhere. On one hand, cloud-based big data analytics exactly tackle in progress issues of scale, speed, and cost. But researchers working to solve security and other real-time problem of big data migration on cloud based platform. This article specially focused on finding possible ways to migrate big data to cloud. Technology which support coherent data migration and possibility of doing big data analytics on cloud platform is demanding in natute for new era of growth. This article also gives information about available technology and techniques for migration of big data in cloud.

**Keywords:** Cloud Computing, Big Data, Data Movement, OLM , RFHC, Migration of Data.

## I. INTRODUCTION

Now a day's cloud is a emerging platform for providing services to corporate world. Many businesses and enterprises already turn to have a cloud based data analytics. Basic problem while adapting cloud based service platform is a migration of data in to cloud. There are various ways and types of services cloud infrastructure offers. With this existing technology some of the challenges are need to resolve on primary role. One major challenge to enterprise adoption of cloud technologies has been the lack of visibility into the security of the platform [10][13]. Cloud service providers are asked to ensure and continuously verify the effectiveness of their security strategies [12]. When choosing a cloud service provider, security should always be a priority. Organizations should ensure that their provider can support their requirements with respect to security and compliance [10]. Companies can mitigate the risks of cloud computing by developing a comprehensive risk assessment prior to choosing a provider and launching a service. Most of the Cloud Backup service providers charge based on data backed up. "Pay as you need" is a service now a days



are more demanding. Data integrity is also a challenge for migration data into cloud. If your data isn't particularly sensitive, then you only have the basic issue of making sure it gets there intact. This is a problem you can mostly address by transferring your data with a tool that checks integrity.

In this work some of the main challenges are discussed to address scientific data analytics. Specially need of identify new scientific workflows from a methodological point of view is focused and discusses thoroughly, and emphasize is given on storage and data management challenges as well as analytics platforms and parallel/distributed paradigms for data intensive science. Furthermore chapters give brief idea about the real challenges and the possible solution for migration of big data into cloud. Article is basically focusing about the related challenges and the available solutions in market.

## 2. RELATED WORK

There are basically three types of migration occurs from commodity hardware to cloud based infrastructure. First one is Enterprise-to-cloud, second is Mobile-to-cloud and last is cloud-to-cloud. In this Enterprise-to-cloud is a common for all and many researchers are currently working on it. Every migration strategy is governed by its own challenges towards cloud migration. Let us discuss each one in details.

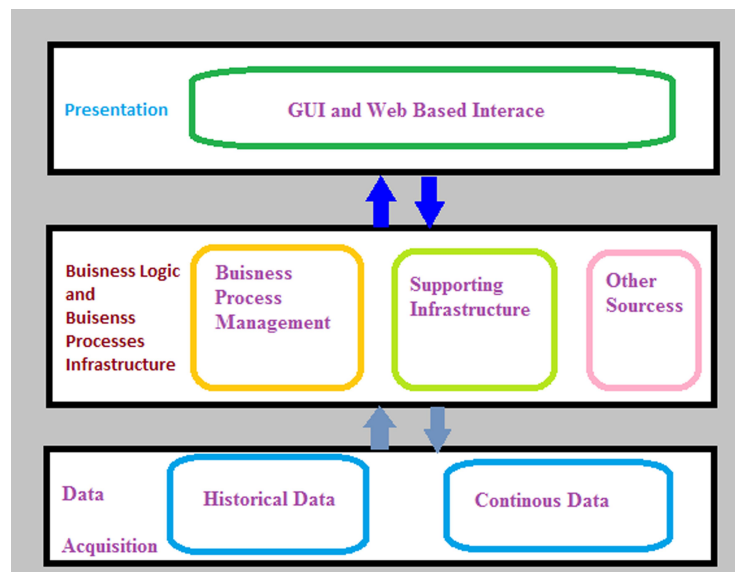
**Enterprise-to-cloud** seems to be where the problems exist. Information is transferred, typically over the open Internet, from servers in the enterprise to public cloud computing providers [14].

**Mobile-to-cloud** – mobile cloud is not that big or works in huge aspects as enterprise cloud do. No business data is stored on mobiles devices (smartphones, tablets, netbooks, and so on), and most of the data they depart in the cloud. Data and meta data which is required for an operation is finally transferred to the mobile device. Thus, data-transfer and its performance is usually not an issue [11].

Finally, **cloud-to-cloud** or intracloud (such as within Amazon Web Services) or intercloud transfer of data (such as between AWS and Rackspace). Every time transferring data from one cloud to another cloud is happened through pipe and that's why, data-transfer performance is directly related to the size of the pipe within the cloud service, as well as to performance-enhancing services such as cache services that might be in use[11]. Typically, intracloud data transfer is between tenants, virtual machines, or applications and data stores. The approaches and technology vary greatly, so you should run your own benchmarks to duplicate the scenario you plan to implement[11].

Hadoop distributed file system (HDFS) is fault-tolerant and deployed on low cost hardware. Recently researchers have proposed online algorithms based on cost- minimizing approach [15]. These algorithms at any point of time aggregate and efficiently process the data by selecting the data center [12][15]. Cloud Computing platforms like Amazon EC2 and S3, Microsoft Azure, Google App Engine; have established an forward-thinking way of managing the cloud by the concept of do-it-together using big data and cloud computing as a single platform. These platforms offer the users a shared pool of servers from various data centers and facilities services to all users by virtualization technologies. Data- intensive applications

i.e. big data applications are depending on cloud platforms for execution of their applications as cloud is elastic and resource provisioned[12].MapReduce and Hadoop like Computing framework provides process of large scale data sets. The main problem is with data to be transferred to cloud especially when it is big data. The cost of transferring the data can add up quickly making the data transfer cost an issue. One of the solution given to avoid the huge data transfer cost is from Amazon web service [3] of its new cloud front service.



**Figure 1. Data Migration Layers**

Figure 1. shows Data Migration Layers architecture. This migration is a traditional way of migrating data to data marts / data warehouses. Data generated in business logic is not a traditional data which can now we store on data marts and data warehouse. This traditional architecture is generating Big Volume of data now a day's. Big data offers huge number of prospects such as taming business by analysing the user behaviour's, getting more accurate forecasts in scientific research by analysing large number of previous results etc. As big data provides these opportunities, along with this, there are many challenges while handling big data [1]. Considering the big data, the challenges lie mainly in areas such as data capture, data storage, analysis, representation. To handle these problems, we have many powerful tools like Hadoop [2].Migration includes complete transfer of data from one place to another with minimum cost. Following are the major approaches proposed in the literature.

- I. **Cloudward Bound:** Planning for Beneficial Migration of Enterprise Applications to cloud- Whenever an enterprise application is to be moved to the cloud, both the CSP and application owner has to decide many things such as security policies, access rights, data movement[4].
- II. **PANDORA:** Budget Constrained Bulk Data Transfer- In cloud, there is requirement of transferring bulk data from one data center to another which are geographically at large distance. PANDORA (People And Network moving Data AROund) [5].
- III. **Online Cost Minimization Approach-** The approach, proposed by Linquan Zhang et al. [6] mainly suggests how to decide the best site for the data aggregation. This approach

suggests two algorithms namely Online Lazy Migration (OLM) and Randomized Fixed Horizon Control (RFHC), which minimizes the cost of the data aggregation of geo-distributed applications.

IV. **SSH Model** - An Efficient Transportation Architecture for Big Data Movement- a new data transfer model on packet network which is an extension to the existing per packet forwarding architecture[7].

### 3. CHALLENGES

Based on the above comparison and analysis, we identified the following challenges that could be research topics in the future. 1) Holistic methodology 2) Scalability in IaaS. Beside of this There are many future important challenges in Big Data management and analytics that -arise from the nature of data: large, diverse, and evolving

**Analytics Architecture**- Optimal architecture of an analytics system should be to deal with historic data and with real-time data at the same time

**Statistical Significance**-Important to have significant statistical results, and not be fooled by randomness

**Distributed Mining**- Distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods[8].

With this some sort of Security and Challenges is also present

**Network level**-These challenges are deals with network protocols and network security, such as distributed nodes, distributed data, Internode communication.

**Authentication Level**- These challenges are categorized under user authentication level deals with encryption/decryption techniques, authentication methods such as administrative rights for nodes authentication of applications and nodes, and logging.

**Data Level**- These challenges are particularly data integrity and availability such as data protection and distributed data.

**Generic Types**- These are categorized under general level are traditional security tools, and use of different technologies [8].

Besides of this cloud computing problem major chunk of unanswered question for completing second steps for migration of big data is answering questions like What is the data involved, how big is it and where is it located? What about continuous and infrequent data ? etc....

There are some challenges for mitigating big data analytics over cloud computing environment.

**Big Data Storage** - A Big Data cloud must incorporate a highly scalable, efficient, low-cost data storage platform. Chief uses of the Big Data cloud storage platform are for archiving, governance and replication, as well as for discovering, acquiring, aggregating and governing multi-structured content. Typically, it would support these uses through integration with a high capacity storage area network architecture [9]

**Big-data processing** - A Big Data cloud should support massively parallel execution of advanced data processing, manipulation, analysis and access functions [9].

**Big-data development** - A Big Data cloud should support Big Data application development which involves modelling, mining, exploration and analysis of deep data sets [9].

#### 4. ANALYSIS AND IMPROVEMENT STRATEGIES

Transforming or migrating Big Data in the cloud and for efficient secure data transfer of corporate world, it must first identify Big Data business applications all over cloud approaches have an advantage that other approaches be deficient in providing. The realistic scenario between the Big Data and cloud paradigms is so broad that you could validly claim you're doing cloud-based Big Data with an existing on premise Hadoop, NoSQL, or enterprise data warehousing environment [11]. Keep in mind that cloud computing includes "private cloud" deployments in addition to or in lieu of public cloud environments Table 1. representing some of the required and measure characteristics for migration of big data into cloud infrastructure.

**Table 1: Required and measure characteristics for migration of big data into cloud infrastructure**

Measures	Required Characteristics
<i>High performance</i>	maximize throughput using techniques such as pipelining and parallel data streams
<i>Reliable.</i>	At every stage of data transfer, periodically checks transfer performance, recovers from errors by retrying transfers, and notifies users of various events (such as errors and success)
<i>Secure</i>	Implementation of security approaches with respect to user authentication and authorization, securely manages the storage and transmission of credentials to endpoints for authentication, and supports optional data encryption.
<i>Third-party transfer</i>	Third-party transfers between two remote endpoints. That is, rather than maintain a persistent connection to an endpoint, users can start a transfer and then let Centralized system manage it for the duration of the transfer
<i>High availability</i>	distributed, replicated, and redundant hosting model need to deploy for high availability of data
<i>Accessible</i>	In software-as a-service (SaaS) provider, users can access it capabilities without installing client software locally
<i>On-demand self-service</i>	Enabling a Big Data cloud service customer to self-provision cloud services--both physical and virtual--automatically and with minimal interaction involving the cloud service provider.
<i>Rapid elasticity and scalability</i>	This enables Big Data cloud resources to be rapidly, elastically, and automatically scaled out up, out, and down on demand.

From above table we can simply predict the technological aspects available in current methods. There are some solutions like OLM , RFHC which provide faster transformation in cloud. These methods actually an online methods and that's why required higher bandwidth. An online lazy migration (OLM) algorithm and a randomized fixed horizon control (RFHC) algorithm have change the scenario of migration data in to cloud infrastructure. One of the practical analysis OLM algorithm achieves a worst-case competitive ratio of 2.55 without the need of any future information and regardless of the system scale, under the typical settings in

real-world scenarios [16]. Improvement strategies are likely to be govern the fundamental issues like CPU, storage, bandwidth, topologies and data platform.

## 5. CONCLUSION AND FUTURE WORK

Cloud services can be considered for storing big amount of files by the corporate world. As the cloud provides virtual centralization of applications, storage, computing and can be accessed by web applications. For migrating large dataset there are some challenges with existing infrastructure. The cloud enables to efficiently manage the large set of images with high bandwidth transfer. It can scale the capacity as and when needed. It can manage the authentication encryption and security protocols. By using fast and secure protocol sessions and virtual on demand servers it is possible to transfer the data set to the cloud by utilizing the maximum bandwidth. In future we are trying to find best possible way to migrate data in cloud with minimum standards. Our future research work is finalizing Hadoop like benchmarks for migration data from commodity hardware and comparing existing methods like OLM and RFHC etc.

## ACKNOWLEDGMENT

This Research work is a collaborative work under department of Computer Science and Engineering for identifying and optimizing data migration effective techniques in cloud based environment.

## REFERENCES

- [1] C. L. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314-347, 2014.
- [2] Hadoop, "Welcome to apache hadoop," <http://hadoop.apache.org/>, accessed May 27, 2015.
- [3] [aws.amazon.com/importexport/productdetails](http://aws.amazon.com/importexport/productdetails).
- [4] M. Hajjat, X. Sun, Y. W. E. Sung, D. Maltz, S. Rao, K. Sripanidkulchai, and M. Tawarmalani, "Cloudward bound: Planning for beneficial migration of enterprise applications to the cloud," *SIGCOMM10*, 2010.
- [5] B. Cho and I. Gupta, "Budget-constrained bulk data transfer via internet and shipping networks," *ICAC11*, 2011.
- [6] L. Zhang, C. Wu, Z. Li, C. Guo, M. Chen, and F. C. M. Lau, "Moving big data to the cloud: An online cost-minimizing approach," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 2710 - 2721, 2013.
- [7] W. Hu, W. Sun, Y. Jin, W. Guo, and S. Xiao, "An efficient transportation architecture for big data movement," *ICICS*, 2013.

[8] Dr.Jangala. SasiKiran et.al. “Recent Issues and Challenges on Big Data in Cloud Computing”, IJCST Vol. 6 ISSN : 0976-8491 (Online) ISSN : 2229-4333 (Print), ISSue 2, AprIl - June 2015pp- 98-102.

[9] James Kobielus, IBM; Bob Marcus, ET Strategies A report on “ Deploying Big Data Analytics Applications to the Cloud: Roadmap for Success” by Cloud Standard Customer Council May 2014 pp-1-21.

[10] Rick Blaisdell white paper on “Security Concerns when Migrating Apps into the Clouds”, Jun 13, 2013 reference <https://www.rickscloud.com/security-concerns-when-migrating-apps-into-the-clouds/>

[11] David Linthicum Article on <http://www.infoworld.com/article/2619818/cloud-computing/all-you-need-to-know-about-cloud-data-transfers.html>

[12] Sreekanth R,Dr.Gondkar RR et al “MOVING BIG DATA MEDICAL IMAGE EHR TO THE CLOUD: AN APPROACH TO DESIGN FAST PERFORMANCE TRANSPORT SOFTWARE USING FASP” IJRETS, ICICSIT 2015 pp-1-6.

[13] VanT.K.Tran et. al. “Size Estimation of Cloud Migration Projects with Cloud Migration Point (CMP)”, 2011 International Symposium on Empirical Software Engineering and Measurement,

[14] <http://cloudnewsdaily.com/public-cloud>.

[15] Ravi Prakash “DEVELOPING A RECOMMENDER SYSTEM AND WEB 2.0 TOOLS TO ENHANCE A BLENDED LEARNING MODEL ”, International Journal of Research in Engineering, Technology and Science, Volume VI, Special Issue, July 2016, ISSN 2454-1915 pp- 1-7.

[16] Linquan Zhang et. al. “Moving Big Data to The Cloud”, Proceedings IEEE INFOCOM, 2013 DOI- 978-1-4673-5946-7/13 pp- 405-410.