International Journal of Electronics, Communication & Soft Computing Science and Engineering
ISSN: 2277-9477, Volume 3, Issue 7

Cost Effective Approach for Moving Huge Data to Cloud

Leena B. Dhangar **Prof. Amitkumar Manekar**

Abstract — **Nowadays, Big Data is an emerging topic in the cloud computing arena that is getting a lot of importance. Big data is a collection of data sets which are large and complex and very difficult to process by the traditional tools. Hence the tremendous growth of demands on big data processing imposes a heavy burden on, storage in data centers, which hence includes large expenditure to providers of data centers. So to handle such large amount of data, well-suited resources are required. Thus it becomes necessary to move data to the other place due to some reasons. So In order to move data, systems require measurable amount of time for working out. Therefore, minimization of cost has become an evolving issue for the forthcoming big data era. This paper presents a novel approach for moving huge data to the cloud which minimizes the cost of data migration.**

Key Words — **Big Data, Cloud Computing, Cost Minimization, Data Migration**

I. INTRODUCTION

The continuous increase in the volume and detail of data captured by organization, for instance the growth of social media, Internet of Things (IoT) [3], as well as multimedia, has created an irresistible surge of data in either structured or unstructured format. Data creation is in the works at a tremendous rate [1], referred as big data, and has emerged as a widely renowned inclination. Big data are characterized by three aspects: (a) data are numerous, (b) data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed rapidly. Moreover, Big Data [1] is a term used to describe data sets which have grown so large that traditional storage infrastructures are ineffective at capturing, managing, accessing and retaining them in a tolerable time edge. The thing that separates Big Data from plainly a large archive is the need to process these large data sets.

Big data is a term that refers to data sets or combinations of data sets whose size (volume) [4], complexity (variability), and rate of expansion (velocity) make them difficult to capture, manage, process or analyze by usual technology and tools. Whereas the size used to determine whether a particular data set is considered big data is not firmly defined and continues to transform over era, nearly all analysts and practitioners at present refer to data sets from terabytes to multiple petabytes as big data [1].

Big Data has gained much attention from the academia and the IT business [6]. Information is generated and composed in the digital globe, at a rate that rapidly exceeds the border. In current stages, above 2 billion public are connected to the Internet, and over 5 billion persons possess cellular phone. Near 2020, 50 billion devices are likely to be connected to the Internet**.** At this peak, predicted data creation will be 44 times greater than it was in 2009. At the same time as information is transferred and shared at light speed on optic fiber and wireless network [10], the quantity of facts and the pace of market expansion boost. Nevertheless**,** the hasty expansion rate of such bulky data [8] generates abundant challenges, such as the speedy enlargement of data; transmit speed, diverse data [9], security and cost. Big Data is not a technology, but rather a phenomenon resulting from the vast amount of raw information generated crossways the world, and composed by commercial and government organizations There are numerous sources of big data and the types of data they create differ such as structured data [5], Unstructured data [5] and Semi-Structured data.

Over the past several years there has been a tremendous increase in the amount of data being transferred between Internet users. Escalating usage of streaming multimedia [3] and other Internet based applications has contributed to this surge in data transmission. An additional facade of the augment is due to the expansion of Big Data [18], which refers to data sets that are an order of magnitude larger than the standard file transmitted via the Internet. Big Data can range in size from hundreds of gigabytes to petabytes [11]. Today everything is being stored digitally. Within the past decade, everything from banking transactions to medical history has migrated to digital storage. This change from physical documents to digital files [12] has necessitated the creation of large data sets and consequently the transfer of large amounts of data. There is no sign that the amount of data being stored or transmitted by users is steady or even decreasing. Every year average Internet users are moving more and more data through their Internet connections [12]. Depending on the bandwidth of these connections and the size of the data sets being transmitted, the duration of transfers could potentially be measured in days or even weeks. There exists a need for an efficient transfer technique that can move large amounts of data quickly and easily without impacting other users or applications. Thus Big Data has translated already into the big price because of its high demands on computation and communication resources [14]. It can be predicted that 71% of the worldwide data center hardware spending will come from the big data processing which will be beyond $126.2 billion. Day by day more and more data is created by the internet users through internet attachment [11]. Accordingly considering the bandwidth of these connections and size of these data sets that is being transmitted, duration of data transfer could be measured in terms of days or even weeks.

Therefore it becomes necessary to invent an approach that will minimize the cost [1] of processing of this big data.

II. LITERATURE SURVEY

Technologies are changing rapidly with lot of competition. In past hardware cost was meaningful, as storage was a big issue for technical development, because of its cost. Software and hardware, both were having same cost at that time. After that software becomes complex in terms of improvement, but effortless to use. Nowadays, with decrement in cost of hardware, the limitation of storage is not a concern. Industry is on the rise; hardware cost is getting lowered so sufficient amount of storage is available without difficulty. Former technology was having particular views on hardware practice, now even 1TB is not a giant agreement for our commodity system. Many social network use Resource Description Framework (RDF) [3]. Facebook's Open Graph [4], Freebase [5] and DBpedia [6] are having structured data. Facebook's Open Graph [4] shows connection of user to its real functioning. Freebase [5] provides structured directories for music. DBpedia [6] provide structural contents from Wikipedia. As per records till 2012, every minute usage of social networking site 'Facebook', having largest number of users, generating share of 684,478 pieces of contents, 'YouTube' users upload 48 hours video, 'Instagram' users share 3,600 new photo and 'Tumbir' sees 27,778 new post published [7]. A Boeing 737 engine generates 10 terabytes of data in every 30 minutes of flight [8]. All these data are information regarding weather conditions, positioning of plane, travelers information and other matter. Thus volume, velocity and complexity of data creation are rising day to day that require tool to handle it and more importantly within time boundary. Conservative database is not adequate for doing all these calculation under the time limit. Here Hadoop fulfill all current requirements. Facebook, Google, LinkedIn, Twitter are establishing their business in Big Data. A lot of companies are still not having Hadoop professionals but they hire those from other companies. In today's world Big Data is moving towards cloud computing. Cloud computing provides essential Infrastructure as CPU, bandwidth, storage spaces at needed time. Organization like Facebook, LinkedIn, Twitter, Microsoft, Azure, Rackspace etc. have moved to cloud and doing Big Data analytic work, like Genome Project [11] that is processing petabytes of data in less amount of time. These technologies use MapReduce, for appropriate functioning. For moving Big Data to cloud, all data is moved and processed at data center [12], as being available at one place, cloud facilities can be easily provided. The data type that increases most rapidly is unstructured data. This type of data is characterized by means of "human information" such as videos, movies, photos, and weather records, log files, and text [11]. According to processor World, unstructured information may account for more than 70% to 80%of all data in organization [14]. These data, which mostly originate from social media, constitute 80% of the data worldwide and account for 90% of Big Data. Currently, 84% of IT managers process unstructured data, and this proportion is projected to drop by 44% in the near future [11]. For much organization, suitable strategy must be developed to manage such data.

Before the cloud era, each application has its own storage server, computing server, application logic as well as individual user interface. Nowadays, an emerging computing Paradigm, called cloud computing provide a shared resource pool, including shared storage resources, shared computing resources, and even shared application logic. All resources are provided on demand and can be charged by pay-as-you-go policy.

The most important issue leftovers how does one move the enormous amounts of data into a cloud, in the very first place? The present observe is to copy the data into large hard drives and dispatch them to the data center [11] [12], or else to move machines completely [13] [14]. Such a shipping method inevitably introduces undesirable delay and possible service downtime, while outputs of the data analysis often needed to be presented to users in the quickest fashion [13]. It is also less secure, given that the hard drives can be infectious with a malicious Program, or lost on the way due to path accident. A more flexible and intelligent data movement strategy is in need, to minimize any potential service downtime. The challenge escalates when we practically consider that data are dynamically and continuously produced, and from different geographical locations, e.g. astronomical data from disparate observatories [15], sensor data from geo-distributed locations 16] and user data from different servers. For dynamic data, an efficient online algorithm is needed, which dynamically guides the transfer of data into the cloud over time; for geo-dispersed data sets with highly cohesion, which cannot be divided into several independent small sets to process in parallel, we have to decide the best data center to aggregate all data onto (e.g., Amazo Elastic MapReduce launches all processing nodes in the same EC2 Availability Zone) [17] given that a MapReduce like Framework is efficient to process data in one data center, but not crossways data centers due to the overhead of enormous amounts of data moving among multiple data centers in the stage of shuffle and reduce.

Many efforts have been made to lower the computation or communication cost of data centers. Data center resizing (DCR) has been proposed to reduce the computation cost by adjusting the number of activated servers via task placement [3]. Based on DCR, some studies have explored the geographical distribution nature of data centers and electricity price heterogeneity to lower the electricity cost [4] [6]. Big data service frameworks, e.g. [7], comprise a distributed file system below, which distributes data chunk and their replica across the data centers for fine-grained load-balancing and high parallel data access performance. To reduce the communication expenditure, a few current study make hard work to get better data locality by placing jobs on the servers where the input data reside to avoid remote data loading [7], [8].Although the above solutions have obtained some positive results, they are far from achieving the cost efficient big data processing because of the following weaknesses. First, data

locality may result in a waste of resources. For example, most computation resource of a server with less popular data may stay at rest. The low resource efficacy further causes more servers to be activated and hence higher operating cost.

Second, the links in networks vary on the transmission rates and costs according to their unique features [9], e.g., the distances and physical optical fiber facilities between data center. Conversely, the presented routing approaches among data centers fail to exploit the link diversity of data center networks. Due to the storage and computation capacity constraint, not all tasks can be placed on the same server, on which their equivalent data reside. It is inescapable that certain data must be downloaded from a remote server. In this case, routing approach matters on the transmission cost. As indicated by Jin et al. [10], the transmission cost, e.g., energy, nearly proportional to the number of network connection used. The more connection used, the higher cost will be incurred. As a result, it is essential to lower the number of links used while satisfying all the transmission requirements.

Third, the Quality-of-Service (QoS) [24] of big data tasks has not been considered in existing work. The QoS [24] of any cloud computing tasks is first determined by where they are placed and how many computation resources are allocated. Moreover, the transmission rate is another prominent factor since big data tasks are data-centric and the computation task cannot proceed until the corresponding data are accessible. Existing studies, e.g., [3], on universal cloud computing tasks mainly focus on the computation capacity constraints, while ignoring the constraints of transmission rate. To conquer above weaknesses, the cost minimization [21] problem for big data processing via joint optimization of task assignment, data placement, and routing in geo distributed data centers was taken into consideration. Large-scale data centers have been deployed all over the world providing services to hundreds of thousands of users. According to [11], a data center may consist of large numbers of servers and consume megawatts of power. Millions of dollars on electricity cost have posed a heavy burden on the operating cost to data center providers. Therefore, reducing the electricity cost has received significant attention from both academia and industry [5], [11] [13]. Among the mechanisms that have been proposed so far for data center energy management, the techniques that attract lots of attention are task placement and DCR. DCR and task placement are usually jointly considered to match the computing requirement. Liu et al. [4] reexamine the same problem by taking network delay into concern. Fan et al. [12] study power provisioning strategies on how much computing equipment can be safely and efficiently hosted within a given power budget. Rao et al. [3] investigate how to reduce electricity cost by routing user requests to geo-distributed data centers with accordingly updated sizes that match the requests. Recently, Gao et al. [14] propose the optimal workload control and balancing by taking account of latency, energy utilization and electricity price. Liu et al. [15] reduce electricity cost and environmental impact using a holistic approach of workload balancing that integrates dynamic pricing.

To tackle the challenges of effectively managing big data, many proposals have been proposed to improve the storage and computation process [25]. The key issue in big data management is reliable and effective data placement. To accomplish this purpose, Sathiamoorthy et al. [16] present a novel family of erasure codes that are efficiently repairable and offer higher reliability compared to Reed-Solomon codes. They also analytically show that their codes are optimal on an identified tradeoff between locality and minimum distance. Yazd et al. [8] make use of flexibility in the data block placement policy to increase energy efficiency in data centers and propose a scheduling algorithm, which takes into account energy efficiency in addition to fairness and data locality properties. Hu et al. [17] propose a mechanism allowing linked open data to take advantage of existing large-scale data stores to meet the requirements on distributed and parallel data processing. Shachnai et al. [21] investigate how to determine a placement of file copies on the servers and the amount of load capacity assigned to each file copy so as to minimize the communication cost while ensuring the user understanding. Agarwal et al. [22] recommend an automated data placement mechanism Volley for geo-distributed cloud services with the consideration of WAN bandwidth cost, data center capacity limits, data inter-dependencies, etc. Cloud services make use of Volley by submitting logs of data center requests. Volley analyzes the logs using an iterative optimization algorithm based on data access pattern and client location, and outputs migration recommendations back to the cloud service. Cidon et al. [23] discover MinCopysets, a data replication placement scheme that decouples data distribution and replication to improve the data durability properties in distributed data centers. Recently, Jin et al. [10] propose a joint optimization scheme that simultaneously optimizes virtual machine (VM) placement and network flow routing to maximize energy savings. Existing work on data center cost optimization, big data management or data placement mainly focuses on one or two factors. To deal with big data processing in geo-distributed data centers, it is essential to jointly consider data placement, task assignment and data flow routing in a systematical manner.

CONCLUSION

Previous technologies developed an optimization model for migrating enterprise IT applications on a hybrid cloud. They focus on workflow migration and application performance optimization by deciding modules to be moved to the cloud. And also they focus on static scenarios with fixed amount of bulk data transfer instead of considering dynamically generated data. This paper presents an online approach for moving big data to the cloud which minimizes the cost of data transfer.

REFERENCES

[1] Lin Gu, Deze Zeng, Peng Li, and song guo, “Cost Minimization for Big Data Processing in Geo-Distributed Data Centers” pp. 2168-

6750, 2013.

[2] Hong Xu, Chen Feng, "Temperature Aware Workload Management In Geo-distributed Datacenters," in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) ACM, pp. 33–36, 2013.

[3] S. A. Yazd, S. Venkatesan, and N. Mittal, "Boosting energy Efficiency With mirrored data block replication policy and energy Scheduler," SIGOPS Oper. Syst. Rev., vol. 47, no. 2, pp. 33–40, 2013.

[4] H. Jin, T. Cheocherngngarn, D. Levy, A. Smith, D. Pan, J. Liu, and N. Pissinou, "Joint Host-Network Optimization for Energy-Efficient Data Center Networking," in Proceedings of the 27th International Symposium on Parallel Distributed Processing (IPDPS), pp. 623–634, 2013.

[5] L Y. Wu, C. Wu, B. Li, L. Zhang, Z. Li, and F. Lau, "Scaling Social Media Applications into Geo-Distributed Clouds," in *Proc. IEEE INFOCOM*, Mar. 2012.

[6] M. Lin, Z. Liu, A. Wireman, and L. Andrew, "Online Algorithms For Geographical Load Balancing," in *Proc. IEEE IGCC*, 2012.

[7] T. Lu and M. Chen, "Simple and Effective Dynamic Provisioning for Power-Proportional Data Centers," in *Proc. IEEE CISS*, Mar. 2012.

[8] Mathew R. Sitaraman and P. Shenoy, "Energy-aware Load Balancing In Content Delivery Networks" Proc. IEEE INFOCOM, Mar 2012.

[9] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, And C. Hyser, "Renewable and Cooling Aware Workload Management for Sustainable Data Centers," in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) ACM, pp. 175–186, 2012.

[10] R. Kaushik and K. Nahrstedt, "T*: A data-centric cooling energy costs reduction approach for Big Data analytics cloud," in 2012 International Conference for High Performance Computing, Networking, Storage and Analysis (SC), pp. 1–11, 2012.

[11] M. Cardosa, C. Wang, A. Nangia, A. Chandra, and J. Weissman, "Exploring MapReduce Efficiency with Highly-Distributed Data", in *Proc. ACM MapReduce*, 2011.

[12] B. Cho and I. Gupta, "Budget-Constrained Bulk Data Transfer Via Internet and Shipping Networks," in *Proc. ACM ICAC*, 2011.

[13] M. Lin, A. Wierman, L. L. Andrew, and E. Thereska, "Dynamic Rightsizing for Power-proportional Data Centers," in *Proc. IEEE INFOCOM*, April 2011.

[14] X. Cheng and J. Liu, "Load-Balanced Migration of Social Media to Content Clouds," in *Proc. ACM NOSSDAV*, June 2011

[15] Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Greening Geographical Load Balancing," in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, pp. 233–244, 2011.

[16] B. Cho and I. Gupta, "New Algorithms for Planning Bulk Transfer Via Internet and Shipping Networks," in *Proc. IEEE ICDCS*, 2010.

[17] D. Logothetis, C. Olston, B. Reed, K. C. Webb, and K. Yocum, "Stateful Bulk Processing for Incremental Analytics," in *Proc. ACM SoCC*, 2010.

[18] 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. IEEE AINA*, 2010.

[19] 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.

[20] M. Hajjat, X. Sun, Y. E. Sung, D. Maltz, and S. Rao, "Cloudward Bound: Planning for Beneficial Migration of Enterprise Applications to the Cloud," in *Proc. ACM SIGCOMM*, August 2010.

[21] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing Electricity Cost: Optimization of Distributed Internet Data Centers in Multi Electricity- Market Environment," in Proceedings of the 29th International Conference on Computer Communications (INFOCOM). IEEE, pp.1–9, 2010.

[22] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan, "Volley: Automated Data Placement for Geo-Distributed Cloud Services," in the 7th USENIX Symposium on Networked Systems Design and Implementation (NSDI), pp. 17–32, 2010.

[23] 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.

[24] J. Cohen, B. Dolan, M. Dunlap, J. M. Heller stein, and C. Welton, "Mad Skills: new analysis practices for big data," Proc. VLDB Endow.vol. 2, No. 2, pp. 1481–1492, 2009.

[25] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on Large clusters," Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.