BIG DATA IN CLOUD CHALLENGES AND SOLUTION

M.Amudha[1], T.Ambika [2], P.Sangeetha[3], B.S.Sangeetha[4].

[1] *Student Research Scholar, Department of Computer Science, Shri Sakthikailassh Women's College, Tamil Nadu, India Country*
[2] *Student Research Scholar, Department of Computer Science, Shri Sakthikailassh Women's College, Tamil Nadu, India Country*
[3] *Student Research Scholar, Department of Computer Science, Shri Sakthikailassh Women's College, Tamil Nadu, India Country*
[4] *Head of the Department of Computer Science, Shri Sakthikailassh Women's College,Tamil Nadu, India Country*

Abstract

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications. So a hadoop cluster is used in Big Data for a data analysis methodology enabled by recent advances in technologies and architecture. A Hadoop cluster is essentially a computational cluster that distributes the data analysis workload across multiple cluster nodes that work to process the data in parallel. However, big data entails a huge commitment of hardware and processing resources, making adoption costs of big data technology prohibitive to small and medium sized businesses. Cloud computing offers the promise of big data implementation to small and medium sized businesses. Despite the fact that cloud computing offers huge opportunities to the IT industry, the development of cloud computing technology is currently has several issues. This study presents an idea for introducing the cloud and big data are infiltrating company practices, and in some cases, they're doing it together. The intersection of the two brings many challenges and opportunities for companies that choose to process and store big data in the cloud. The aim of this idea is to provide a better understanding of the design challenges of cloud computing and identify important research directions in this big data increasingly important area. We'll describe a series of studies by which we and other researchers have assessed the effectiveness of these techniques in practical situations. Finally, in this study we will show how this idea could be implemented in a practical and useful way in industry.

Key Words: Big data, Cloud computing, Challenges in Big Data, Hadoop cluster,.

1. INTRODUCTION

Big Data is a data analysis methodology enabled by a new generation of technologies and architecture which support high-velocity data capture, storage, and analysis (Villars, Olson, & Eastwood, 2011). Data sources extend beyond the traditional corporate database to include email, mobile device output, sensor-generated data, and social media output (Villars, Olofson, & Eastwood, 2011). Data are no longer restricted to structured database records but include unstructured data – data having no standard formatting (Coronel, Morris, & Rob, 2013). Big Data requires huge amounts of storage space. While the price of storage continued to decline, the resources needed to leverage big data can still pose financial difficulties for small to medium sized businesses. A typical big data storage and analysis Infrastructure will be based on clustered network-attached storage (NAS) (White, 2011).

Data storage using cloud computing is a viable option for small to medium sized businesses considering the use of Big Data analytic techniques. Cloud computing is on-demand network access to computing resources which are often provided by an outside entity and require little management effort by the business (IOS Press, 2011). A number of architectures and deployment models exist for cloud computing, and these architectures and models are able to be used with other technologies and design approaches (IOS Press, 2011). Owners of small to medium sized businesses who are unable to afford adoption of clustered NAS technology can consider a number of cloud computing models to meet their big data needs. Small to medium sized business owners need to consider the correct cloud computing in order to remain both competitive and profitable. This idea is based on concept of natural cloud [2]. Different weather is based on different cloud on the sky. So, we can use different cloud template as based for designing a cloud computing [8, 9, 10] systems [12] with different character [14, 17]. First of all, this idea back to concept of various types of clouds on the sky. As shown in Figure 1, we have different characters of the cloud and each character has different style and makes different weather. It means different types of a cloud are useful for different atmospheres and geographies. As first step, we'll consider this idea which is based on natural cloud then we'll back to cloud template architecture.

1.1 Big Data and the Cloud

The term big data is derived from the fact that the datasets are so large that typical database systems are not able to store and analyze the datasets (Manyika et al., 2011). The datasets are large because the data is no longer traditional structured data, but data from many new sources, including e-mail, social media, and Internet-accessible sensors (Manyika et al., 2011).

The characteristics of big data present data storage and data analysis challenges to businesses. A typical model for in-house storage of big data is clustered Network-Attached Storage (Sliwa, 2011). The configuration would begin with a network-attached storage (NAS) pod consisting of several computers attached to a computer used as the (NAS) device. Several NAS pods would be attached to each other through the computer used as the NAS device. Clustered NAS storage is an expensive prospect for a small to medium size business. A cloud services provider can furnish the necessary storage space for substantially lower costs.

1.2 Structure of Hadoop Cluster

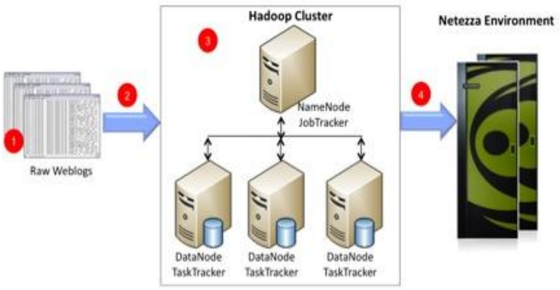

Fig.1 Hadoop Cluster

i.Hadoop cluster:
A hadoop cluster is a special type of cluster that is specially designed for storing and analyzing huge amounts of unstructured data.Ahadoop cluster is essentially a computational cluster that distributes the data analysis workload across multiple cluster nodes that workto process the data in parallel..

ii. Benefits of hadoop cluster:
The primary benefit to using Hadoop clusters is that they are ideally suited to analyzing big data. Big data tends to be widely distributed and largely unstructured. The reason why Hadoop is well suited to this type of data is because Hadoop works by breaking the data into pieces and assigning each "piece" to a specific cluster node for analysis. The data does not have to be uniform because each piece of data is being handled by a separate process on a separate cluster node.

2. Significant challenges in Big Data

Various types of challenges had different approach and each of them has different purpose:

2.1. A comprehensive approach to using big data:

Most companies collect gobs of data but they don’t have comprehensive approaches for centralizing the information. According to a recentsurvey by LogLogic, 59% of the more than 200 security officers who responded say they are either using disparate systems for gathering data, not managing log data, or they use antiquated spreadsheets. The right analytics tools can definitely help to streamline and make sense of all this data, but a well-conceived strategy for collating data sources from different silos is still necessary.

2.2. Effective ways of turning “big data” into “big insights”:

No matter how you slice it, data is just that – data. In and of itself, data doesn’t necessarily provide decision makers with the kind of insights they need to do their jobs effectively or to take the next best actions based on discoveries about customer trends or other revelations about market conditions. This is where the right analytics tools are needed to help data scientists and business leaders make sense of the volumes of data that are pouring into their organizations. This includes the use of data visualization tools that can be used to help put data into context

2.3. Big data skills are in short supply:

There’s already a shortage of data scientists in the market. This includes a scarcity of people who know how to work well with large volumes of data and big data sets. Companies need the right mix of people to help make sense of the data streams that are coming into their organizations. This includes skills for applying predictive analytics to big data, a skill set that even most data scientists lack.

2.4. Getting the right information into the hands of decision makers:

Companies should use analytics “to avoid getting buried under the humongous amount of information they generate through various outlets,” according to a recent ZDNet Asia interview with XMG analyst Jacky Garrido. It’s true – too many companies lack coherent approaches to utilizing the gushers of customer and business data that are flowing into their organizations. As Garrido notes, as data is gathered, it needs to be mapped out. Moreover, critical data needs to be separated from insignificant or unnecessary data (e.g. inconsequential comments made by customers on Facebook or Twitter).

3.Cloud Big Data Challenges

Fig.2 Cloud Big Data using Hadoop

Some important cloud big data challenges are follows

3.1.Vertical scaling :

Vertical scaling achieves elasticity by adding additional instances with each of them serving a part of the demand. Software like Hadoop is specifically designed as distributed systems to take advantage of vertical scaling. They process small independent tasks in massive parallel scale. Distributed systems can also serve as data stores like NoSQL databases, e.g. Cassandra or HBase, or filesystems like Hadoop's HDFS. Alternatives like Storm provide coordinated stream data processes in near real-time through a cluster of machines with complex workflows.

3.2.The interchangeability :

The interchangeability of the resources together with distributed software design absorbs failure and equivalently scaling of virtual computing instances unperturbed. Spiking or bursting demands can be accommodated just as well as personalities or continued growth.

3.3.Unlimited Resoureces:

Renting practically unlimited resources for short periods allows one-off or periodical projects at a modest expense. Data mining and web crawling are great examples. It is conceivable to crawl huge web sites with millions of pages in days or hours for a few hundred dollars or less. Inexpensive tiny virtual instances with minimal CPU resources are ideal for this purpose since the majority of crawling the web is spent waiting for IO resources. Instantiating thousands of these machines to achieve millions of requests per day is easy and often costs less than a fraction of a cent per instance hour.

3.4.Mining operations :

Mining operations should be mindful of the resources of the web sites or application interfaces they mine, respect their terms, and not impede their service. A poorly planned data mining operation is equivalent to a denial of service attack. Lastly, cloud computing is naturally a good fit for storing and processing the big data accumulated form such operations.

FUTURE WORK

This is a new idea and it's requiring more study as we listed below:

1-Compare and contrast the benefits of operating Big Data infrastructure in the cloud vs on-prem data centers.

2-Cover how cloud helps companies to derive faster time to value from big data.

3-Talk about how agility and flexibility of the cloud benefits big data infrastructure and completely changes the model

4-Cover how new advances in cloud security and compliance and progressively changing perceptions around those topics in the enterprise and causing more and more enterprises to consider the cloud option for their big data infrastructure.

.

3. CONCLUSIONS

Cloud computing enables small to medium sized business to implement big data technology with a reduced commitment of company resources by means of using hadoop clustering. The processing capabilities of the big data model could provide new insights to the business pertaining to performance improvement, decision making support, and innovation in business models, products, and
services. Benefits of implementing big data technology through cloud computing are cost savings in hardware and processing. With the emergence of cloud computing system as a computer science paradigm in which computing is done exclusively on resources leased only when needed from big data centers, scientists are faced with a new platform option. However, the initial target often cloud computing system does not match the characteristics of the scientific computing workloads, also often scientists are require customize their cloud based on requirement. In this paper we introduced an idea of using hadoop cluster for splitting the large data that available in big data and overcoming to cloud computing issues. Our main finding is that the cloud computing systems are requiring a revolution such as using cloud template for different purpose on a cloud.

ACKNOWLEDGEMENT

We thank all sponsors in the footnote on the first page for funding this ongoing research project and all volunteers for their involving this research project. We would also like to thank the anonymous referees for their constructive and valuable comments.

REFERENCES

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G…Zaharia, M. (2010, April). A view of cloud computing. Communications of the ACM, 53(4), 50-58. DOI: 10.1145/1721654.1721672.

[2] Rouse, M. (2010b, August). Infrastructure as a Service. Retrieved from http://searchcloudcomputing.techtarget.com/definition /Infrastructure-as-a -Service-IaaS

[3] Cisco. (2009). Infrastructure as a Service: Accelerating time to profitable new revenue streams. Retrieved from http://www.cisco.com/en/US/solutions/collateral/ns341/ns991/ns995 /IaaS_BDM_WP.pdf

[4] Salesforce.com. (2012). The end of software: Building and running applications in the cloud. Retrieved from http://www.salesforce.com/paas/

[5] Géczy, P., Izumi, N., &Hasida, K. (2012). Cloudsourcing: Managing cloud adoption. Global Journal of Business Research, 6(2), 57-70.

[6] Oracle. (2012). Oracle platform as a service. Retrieved from http://www.oracle.com/ us/technologies/cloud/oracle-platform-as-a-service-408171.html

[7] Jackson, K. L. (2012). Platform-as-a-service: The game changer. Retrieved from http://www.forbes.com/sites/kevinjackson/2012/01/25/platform-as-a-service-the-gamechanger.

[8] Cole, B. (2012). Looking at business size, budget when choosing between SaaS and hosted ERP. E-guide: Evaluating SaaS vs. on premise for ERP systems. Retrieved from http://docs.media.bitpipe.com/io_10x/io_104515/item_548729/SAP_sManERP_IO%23104515_EGuide_061212.pdf

[9] Carraro, G., & Chong, F. (2006, October). Software as a service: An enterprise perspective. Retrieved from http://msdn.microsoft.com/en-us/library/aa905332.aspx #enterprisertw_topic3.

[10] Santosh Kumar, R. H. Goudar, "Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications: A Survey", International Journal of Future Computer and Communication, Vol. 1, No. 4, December 2012.

[11] www.jason.org/digital_library/201.aspx accessible on March 6, 2013.

[12] T. Dillon, C. Wu, E. Chang, "Cloud Computing: Issues and Challenges," 2010 24th IEEE International Conference on Advanced Information Networking and Applications(AINA), pp. 27-33, DOI=20-23 April 2010.

[13] J. F. Yang, Z. B. Chen, "Cloud Computing Research and Security Issues," 2010 IEEE International Conference on Computational Intelligence and Software Engineering (CiSE), Wuhan pp. 1-3, DOI=10-12 Dec. 2010.

[14] S. Zhang, S. F. Zhang, X. B. Chen, X. Z. Huo, "Cloud Computing Research and Development Trend," In Proceedings of the 2010 Second International Conference on Future Networks (ICFN '10). IEEE Computer Society, Washington, DC, USA, pp. 93-97, DOI=10.1109/ICFN.2010.58, 2010

[15] B. Grobauer, T. Walloschek, E. Stöcker, "Understanding Cloud Computing Vulnerabilities," 2011 IEEE Security and Privacy, pp. 50-57, DOI= March/April 2011.

[16] W. A. Jansen, "Cloud Hooks: Security and Privacy Issues in Cloud Computing, "Proceedings of the 44th Hawaii International Conference on System Sciences, 2011.

[17] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. "A view of cloud computing", Communications of the ACM,53(4), 50-58, 2010. [9] Mell, P., & Grance, T.,"The NIST definition of Cloud computing (draft).", NIST special publication, 800(145), 7, 2011.

[18] Foster, I., Zhao, Y., Raicu, I., & Lu, S. "Cloud computing and grid computing 360-degree compared", In Grid Computing Environments Workshop, 2008. GCE'08 (pp. 1-10). IEEE, 2008.

[19] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility", Future Generation computer systems, 25(6), 599-616, 2009.

[20] Youseff, L., Butrico, M., & Da Silva, D., "Toward a unified ontology of cloud computing", In Grid Computing Environments , 2008. GCE'08 (pp. 1-10). IEEE, 2008.

Author profile:

M.Amutha received her M.Sc Degree in Information Technology from Srithar University,Rajasthan, India. Her interest of area for Research Work is in Cloud Computing and Networking.

T.Ambika received her M.Sc Degree in Information Technology from Anna University,Chennai, India. Her interest of area for Research Work is in Cloud Computing and Networking.

P.Sangeetha received her M.Sc Degree in Information Technology from Anna University,Chennai, India. Her interest of area for Research Work is in Cloud Computing and Networking.

Guided by B.S.Sangeetha M.C.A., M.Phil, B.Ed., Head of the Department of Computer Science in Shri Sakthikailassh women's college. Her interest of area for Research Work is in networking.