To Analyze Power Consumption and Quality of Service using Map Reduce on Hadoop: A Survey

Sandeep Rai
Assistant Professor
Department of CSE
TIT (Excellence) Bhopal

Aishwarya Namdev
M.Tech Scholar
Department of CSE
TIT (Excellence) Bhopal

ABSTRACT
Data is growing at a rate which cannot be handled by the traditional methods of computing. To store and process such data new data analysis and storage techniques have emerged over the last few years. Hadoop is one such parallel processing open source framework which provides distributed storage and processing of Big data. Big Data analytics has emerged as an attractive domain of research these days. For handling big data cloud computing has been used and back end of thetechnology is cluster of resources. Cluster of resources can be formed using a framework like Apache Hadoop. In this paper a survey is performed on big data analysis using Apache Hadoop and other utility tools. For better performance of cloud Quality of service and power consumption should be optimal. So in this survey is discuss resolves around Quality of Service and energy consumption.

Keywords
Big Data, Cloud computing, Quality of service (Qos), Power consumption, Hadoop

1. INTRODUCTION
Big data is the term for any assortment of data sets so large and sophisticated that it becomes robust to process exploitation ancient process applications. Large and complex data sets comprise of variety of structured and unstructured data, often in the quality of too big, too fast or too difficult to be handled by traditional techniques are referred to as "Big Data". There are 4 qualities that under the Big Data, which is Volume, Velocity, Variety, and Veracity – collectively named as the 4Vs. Each of the qualities is described briefly below Volume means the quantity or amount of the data, variety that is the diversity of the data types, velocity means that the speed of the data generation and the speed of which the data need to be processed and veracity means the ability to trust the data to be highly accurate and reliable in times of imperative decision making. Many modern enterprises are now focusing on Big Data, as it is believed to be potentially advantageous in influencing core business processes, providing competitive benefits and induce company revenues and profits. As such many organizations are researching ways to exploit the advantageous features of Big Data. This is done especially in analyzing them to suggest meaningful findings that will lead to better business decisions and to add value to their business. The ability to process and manage large quantity of data in parallel mode is supported by the features of cloud where data analystics concern.

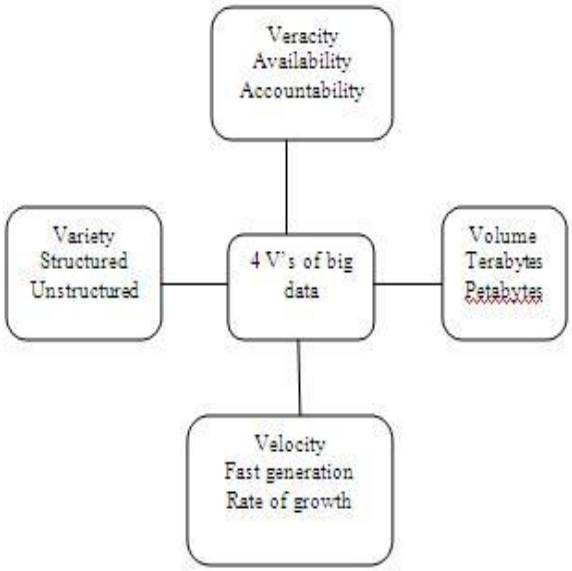

Fig.1 Characteristics of Bigdata

Research areas and its challenges:

There are six major research areas identified as:

Power consumption, Quality of services, Storage and Transport, Accessibility, Inconsistencies, Mobility

2. CLOUD COMPUTING
Clouds are a large pool of easily usable and accessible virtualized resources. These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the infrastructure provider by means of customized Service Level Agreements"Cloud is the essential key to define the cloud computing technology. In rough term, cloud can be equated with use. These interconnected computers can be PCs or network servers. To illustrate, Google may host a cloud that is built on a PC or network servers. This type of Cloud is usually set as private cloud, which limits the use of the cloud only for the company itself. On the other hand, Google owns the cloud that can be accessed by the mass public. This means that any authorized user can access the data and application from his or her computer. The infrastructure and the technology remain invisible to user.

Cloud computing is not to be confused with network computing. Network computing documents or applications are hosted on a single server owned by the company, making

them accessible from any computer on the network. In comparison, cloud computing utilizes multiple servers, multiple networks and also involves multiple companies. Secondly, cloud computing together with its service and storage, can be accessed from anywhere using an internet connection. This is dissimilar to network computing in which the access is limited only within the company's network. Third, cloud computing is also not outsourcing where a company hires an external firm or agency to conduct its computing services. If outsourcing is opted by the company for its data or application, the data however will be set to be limited to the private use of the employees of the company and not to the public via the internet.

3. CLOUD SERVICE MODELS

The National Institute of Standards and Technology (NIST) have classified three service models comprising of Infrastructure, Platform and Software-as-a-service. Each of the services is described briefly below:

Software as a Service (SaaS) – This service allows the access to whole applications to be used over the internet on readily existing cloud infrastructure. Through various platforms, these applications are accessible to many common devices [3]. However, the access may be limited, and secondly, accessing different internet-based SaaS from different browsers, or from different devices such as mobile will result to different presentations of the same application. In this case, the users accessing the applications cannot see or alter the underlying infrastructure but only able to personalize some minor configuration settings.

Platform as a Service (PaaS) – This service allows deployment, management and alteration of certain range of applications on readily existing cloud infrastructure. This can be achieved by utilizing "programming languages, libraries, services, and tools supported by the provider. Similar to SaaS, users cannot see, manipulate or alter the underlying infrastructure, but are able to configure some minor settings of the applications. In addition, the settings in PaaS enable more user control of the applications and to some extent, the application environment.

Infrastructure as a Service (IaaS) – This service requires the infrastructure including storage, networking capacity and related resources to enable the user to run their own software or platforms. The consumer however, does not manage the infrastructure. Instead, the users are able to modify the amount that they requested, in addition to having some control over the storage managements, operating systems, and their deployed applications. The consumer also has limited control of some networking components e.g. the firewall

4. QUALITY OF SERVICE

Even though the cloud has greatly simplified the capacity provisioning process, it poses several novel challenges in the area of Quality-of-Service (QoS) management. QoS denotes the levels of performance, reliability, and availability offered by an application and by the platform or infrastructure that hosts it. QoS is fundamental for cloud users, who expect providers to deliver the advertised quality characteristics, and for cloud providers, who need to find the right tradeoffs between QoS levels and operational costs. However, finding optimal tradeoff is a difficult decision problem, often exacerbated by the presence of service level agreements (SLAs) specifying QoS targets and economical While QoS properties have received constant attention well before the advent of cloud computing, performance heterogeneity and resource isolation mechanisms of cloud platforms have significantly complicated QoS analysis, prediction, and assurance. This is prompting several researchers to investigate automated QoS[5] management methods that can leverage the high programmability of hardware and software resources in the cloud.

This paper aims at supporting these efforts by providing a survey of the state of the art of QoS modeling approaches applicable to cloud computing and by describing their initial application to cloud resource management. Introduce quantitative models that may facilitate Hadoop operators and developers in distinguishing opportunities to enhance energy potency.

5. RELATED WORK

There has remained a big quantity of previous work in energy efficient computing systems. Comprehensively reviewing the prevailing literature closely associated with work.The new architecture referred as Advanced Control Distributed Process Architecture (ACDPA) by connecting the abstraction property of Software Defined Networking (SDN) and also the distributed process power Hadoop [11]. Software Defined Networking (SDN) is an access where abstraction is used to simplify the network in two layers: 1) Used for controlling the traffic. 2) Used for forwarding the traffic. SDN use openflow protocol that is an open protocol for controlling and design the switches in the network. The system offers abstraction for the control of network traffic and also provide fast processing of big data i.e. network traffic. The control is done using SDN controller opendaylight [16] and the processing is done use Hadoop. The system accepts huge amount of data from SDN data plane through wire shark and offer it to Hadoop for process. The result from Hadoop is feedback to the SDN controller that controls the Quality of Service (QoS).

An proposed an architecture [12] that provides better security to the cloud and fulfill Quality of Service (QoS) requirements, using Software Defined Networking (SDN) and Hadoop.

Kerberos is also used to enhance the security & provide authentication and Single Sign On (SSO). Their contribution includes: Set the QoS requirements for group of users of the cloud using the concept of SDN. Using SDN the segregation of the flows thus controlling the QoS can be accomplished.

A measurement-driven methodology [13] to evaluate the impact of replication on Database-as-a-Service environments. The technique builds upon an analytical model represents the database cluster configuration combined with an environment model to shows the transient replication stages. The main aims are exploit fluid modeling procedure to approximate response time percentiles for replicated relational DBaaS platforms, Procedure evaluates the performance of a relational database cluster hosted on DBaaS platforms, methodology estimate under variable workloads and dynamic cluster-reconfigurations and evaluated the concussion of replication overhead and time on performance stability and its effect on Database cluster performance.An introduce the cluster content caching structure in Cloud Radio Access Networks (C-RANs) that takes full advantage of centralized signal process and distributed caching [14].The structures proposed enhance the Quality of Service (QoS) and minimize the power consumption in real-time services. In particular, redundant traffic on the backhaul may be reduced because the cluster content cache provides a part of needed content object for Remote Radio Heads (RRHs) connected to a standard edge cloud. The tractable expressions are derived for each energy efficiency and effective capability performance that shows

that the proposed structure will improve Quality of Service guarantees with a low power cost of local storage. The joint architecture of RRU allocation and RRH association has been studied to any improves the performance of cluster content caching. They defines energy consumption for high throughput workloads of several of virtual machines running the hadoop system by an Open Nebula Cloud [20]. The main target of this work is to focus how the power consumption can be related to the number of Virtual Machines (VMs) & the associated workload generated on a physical server. It is monitored and understood & sub-sequent exposed to the user. In this, the approach considers two types of workloads: Virtual machines deployed on the server and Data analysis algorithm executed on the virtual machine. An introduction how power usage varies over time and as the number of machines increases from individual racks to cluster of up to 5000 servers. Their key searches and contributions are Power capping using dynamic power management can enable additional machines to be hosted but is more useful as a safety mechanism to prevent overload situations. Determined time intervals when large groups of machines are operating close to peak power levels suggesting that power management and power gaps techniques can be more easily exploited at the datacenter-level.CPU voltage/frequency scaling, a technique targeted at energy management, has the potential to be moderately effective at reducing peak power consumption once large groups of machines are considered.

Evaluated the advantages of building systems that are power efficient across the activity range.

6. PROBLEM IDENTIFICATION

There is a variability observed in the power consumed by multiple runs of the similar workload. With the increase in execution time, the variability gets reduced. Thus, there can be limitation of using the power related metrics in case of short running jobs even when there are private clouds present.

In previous techniques the response time, balance utilization of resources of a host and load distribution on all the hosts is only based on the VM scheduling techniques, which further refined to upgraded technique with large data processing This scenario can be controlled in the future work.

7. CONCLUSION

In this paper a survey is performed on big data analytics using cloud based on Apache Hadoop. The objective of survey is comparison of power efficient techniques. Here quality of service techniques which provides low power consumption are studied. And survey paper gives an idea to know about the issues and challenges of big data in cloud environment. As the big data refers the large volume of data and cloud computing can offer the better scalability for the large data. And also describes the techniques which are useful for analyzing the large data, which is generated from difference sources. In order to handle big data use the standard tools like Map Reduce and Hadoop. The future research scope hints an idea for better management of big data in cloud environment.

8. REFERENCES

[1] D. P. Acharjya, Kauser Ahmed P, Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, School of Computing Science and Engineering, IT University Vellore, India, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016

[2] J.Dean, S.Ghemawat,"MapReduce: simplified data processing on large clusters," Communications of the ACM, 51(1):107–113, January 2008.

[3] NIST Definition of Cloud Computing V15, csrc.nist.gov/groups/SNS/cloudcomputing/ cloud-def-V15.doc

[4] Annual energy review 2008. Energy Information Administration, U.S. Department of Energy, June 2011.

[5] "Green Grid data center efficiency metrics: PUE and DCIE," White Paper, The Green Grid, December 2012.

[6] SPECpower ssj2008. Standard Performance Evaluation Corporation, http://www.spec.org/power ssj2008/, April 2013.

[7] Google, " Data center efficiency measurements," The Google Blog, 2014. [8]C. Belady, "In the data center, power and cooling costs more than the IT equipment it supports," Electronics Cooling Magazine, 13(1):24–27, February 2015.

[8] J. Koomey, "Worldwide electricity used in data centers," Environmental Research Letters, 3(3), September 2015.

[9] "International energy annual 2006," Energy Information Administration, U.S. Department of Energy, June 2016.

[10] A. Desai, Nagegowda K S, "Advanced Control Distributed Processing Architecture (ACDPA) using SDN and Hadoop for identifying the flow characteristics and setting the quality of service(QoS) in the network,"2016 IEEE International Advance Computing Conference (IACC), Bangalore, 2015, pp. 784-788.

[11] A. Desai, Nagegowda K S, Ninikrishna T, "Secure and QoS aware architecture for cloud using software defined networks and Hadoop," 2017 International Conference on Computing and Network Communications (CoCoNet), Trivandrum, 2015, pp. 369-373.

[12] R. Osman, J. F. Pérez, G. Casale, "Quantifying the Impact of Replication on the Quality-of-Service in Cloud Databases," 2016 IEEE International Conference on Software Quality, Reliability and Security (QRS), Vienna, 2017, pp. 286-297.

[13] Z. Zhao, M. Peng, Z. Ding, W. Wang, H. V. Poor, "Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 5, pp. 1207-1221, May 2017.

[14] S. Ghemawat, H. Gobioff, S.-T. Leung, "The Google file system," SIGOPS Oper. Syst. Rev., 37(5):29–43, 2003.

[15] J. Ammer, J. Rabacy, "The energy-per-useful-bit metric for evaluating and optimizing sensor network physical layers," In Sensor and Ad Hoc communications and Networks, 2006.

[16] SECON '06. 2006 3rd Annual IEEE Communications Society on, vol. 2, pages 695–700, Sept. 2017. Energy star enterprise server specification, United States April 2017.

[17] L. A. Barroso, U. H¨olzle, "The case for energy-proportional computing," Computer, 40(12):33–37, 2017.

[18] K. Lim, P. Ranganathan, J. Chang, C. Patel, T. Mudge, S. Reinhardt, "Understanding anddesigning new server

architectures for emergingwarehouse-computing environments," In ISC A '08:Proceedings of the 35th International Sym posium on Computer Architecture, pages 315– 326, W ashington, DC, USA, 2008. IEEE Computer Society.

[19] Gridmix. HADOOP-HOME/src/benchmarks/gridmix in all recent Hadoop distributions.

[20] Javier Conejero, Omer Rana, Peter Burnap, Jeffrey Morgan, Blanca Caminero, Carmen Carrion, "Analyzing Hadoop power consumption and impact on application QoS," Elsevier, pg. no. 213-223, March 2017