Analyzing various Issues & Challenges in Big Data BenchMark Datasets

Pooja Singh
Department of Information Technology
NIIST
Bhopal, India
poojasingh.singh375@gmail.com

Asst. Prof. Angad Singh
Department of Information Technology
NIIST
Bhopal, India
angada2007@gmail.com

Abstract— Here in this paper a deep analysis and comparison of various analytics in Big Data is proposed. The paper largely emphases in the various systems and techniques implemented in Big Data for the analysis of Various Services and Data. By analyzing various techniques and comparing their issues and advantages in these techniques a new and efficient system for the study of Big Data is done in future.

Index Terms—Big Data, Cloud Computing, Distributed File System, Benchmark Dataset, Internet Services, Load Balancing.

I. Introduction

Cloud computing presents a new way to supplement the current consumption and delivery model for IT services based on the Internet, by providing for dynamically scalable and often virtualized resources as a service over the Internet. To date, there are a number of notable commercial and individual cloud computing services, including Amazon, Google, Microsoft, Yahoo, and Sales force Clouds can be explained as pools of virtualized resources that can be easily used and accessed. For optimum resource utilization the resources in cloud can be reconfigured dynamically. With the help of strong cloud architectures its mass computing and storage centers organizations and individuals are benefited while utilizing them. A technique Cloud Information Accountability (CIA) framework is based on the notion of information accountability. Unlike privacy protection technologies which are built on the hide-it-or-lose-it perspective, information accountability focuses on keeping the data usage transparent and traceable [2].

Characteristics of high performance scientific applications have been well studied in the high performance computing (HPC) community. As comparable efforts have been prepared in the Big Data area but their exposure is inadequate habitually to applications in the viable area. Data science troubles in research and academia also arrangement with huge numbers but with additional composite applications evaluated to their profitable equivalents. While a complete set of benchmarks is essential to wrap the range of Big Data applications. Massive application domains make us wonder where to start or how to achieve a wide range of coverage.

Figure-1: Application Domains of Big Data

Big data benchmarks are the foundation of those efforts [2]. However, the complexity, diversity, frequently changed workloads—so called workload churns [1] and fast development of big data systems require enormous challenges to big data benchmarking. All Big Data are measured as the advantage of companies, associations and even countries. Extracting the big value from Big Data requires enabling big data systems. After investigating different application domains of Internet services, an important class of big data applications, they give consideration to search engines, which is the most important domain in Internet services in terms of the number of page views and daily visitors. A detailed analysis of search engines workloads and benchmarking methodology has been presented in the paper [3]. An innovative data generation methodology and tool are proposed to produce scalable amounts of big data from a small starting point of authentic data.

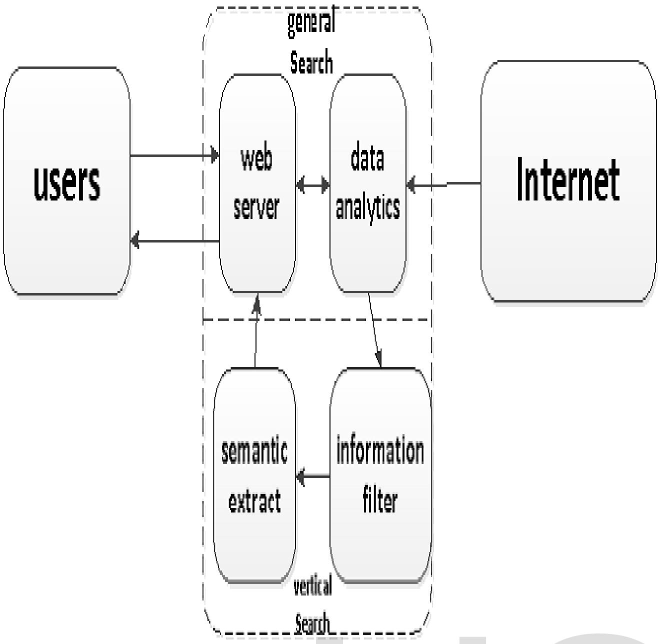

Figure-2: General search and vertical search of Internet Services.

Internet service like search engine, social network, e-commerce. Data are generated faster than ever, the rate of data making will maintain in the approaching years and is expected to increase at an exponential level. These facts evolve the concept of "BigData". The diversity of data and workloads needs comprehensive and continuous efforts on big data benchmarking. Considering the wide utilize of big data schemes for the sake of fairness, big data benchmarks must include diversity of workloads and data sets, which is the prerequisite for evaluating big data systems and structural design. BigData Bench not only wraps extensive application circumstances but also includes diverse and representative data sets. Big Data Benchmarking Requirements:

- A big data benchmark suite candidate must cover not only broad application scenarios, but also diverse and representative real world data sets.
- Big data systems must be handle the four dimensions called "4V" of big data.
- Diverse and representative workloads.
- Covering representative software stacks.
- A big data benchmark suite should keep in pace with the improvements of the underlying systems.
- The benchmarks should be easy to deploy, configure, and run, and the performance data should be easy to obtain.

BigDataBench is for applications from internet services despite the fact that HiBench is related, but anxieties MapReduce [4] approached data analysis and Graph500 is based on graph search to work out supercomputing schemes; additional demonstrated the variety of Big Data use cases. Here they initiate new benchmarks to characterize the application region of general machine learning (GML). Especially, they have selected MDS and clustering classes that are seldom present in the commercially driven benchmarks, yet are salient among Big Data use cases. Our decision is further based on two reasons. First, the implementations of these naturally follow map-pattern [5] with iterations, both are computation and communication intensive, and communications are frequently global evaluated to local neighbor communications establish in HPC applications. These characteristics tend to be comparable in other candidates of the GML category, so they too would advantage from learning MDS and clustering. Subsequent they broadly work with gene sequence and other health data to make available novel means of finding similarities and visualizing them in 3 dimensions (3D) [6] [7] [8] where MDS and clustering are the key components.

As there are many emerging big data applications, here they get an incremental and iterative approach as an alternative of a top-down technique. First of all, they examine the leading application domains of internet services—an essential class of big data applications according to extensively suitable metrics—the number of page views and daily visitors.

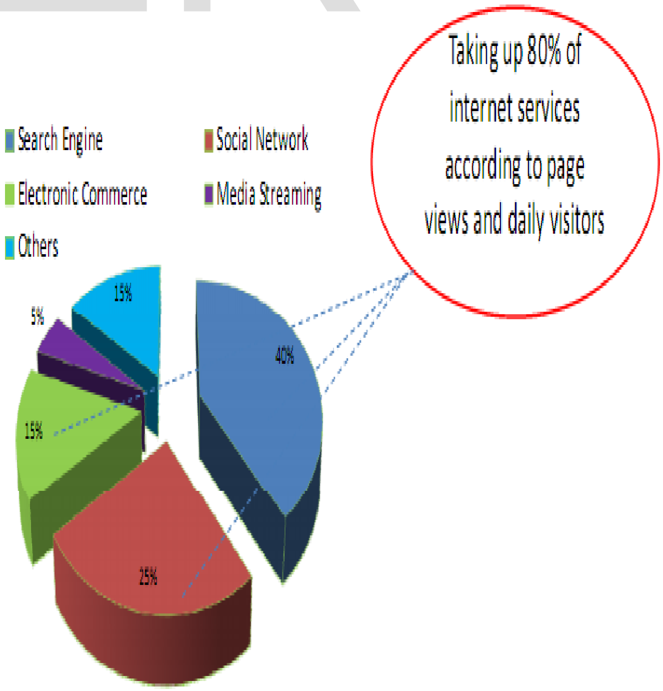

Figure-3: Internet Services Top 20 websites

According to the analysis in the top three application domains are search engines, social networks, and e-commerce, taking up 80% page views of all the internet services in total. And then, we pay attention to typical data sets and big data workloads in the three application domains. We consider data diversity in terms of both data types and data sources, and pay equivalent consideration to structured, semi-structured, and unstructured data. Further, we single out three important data sources in the dominant application domains of internet services, including text data, on which the maximum amount of analytics and queries are performed in search engines [9], graph data (the maximum amount in social networks), and table data (the maximum amount in e-commerce).

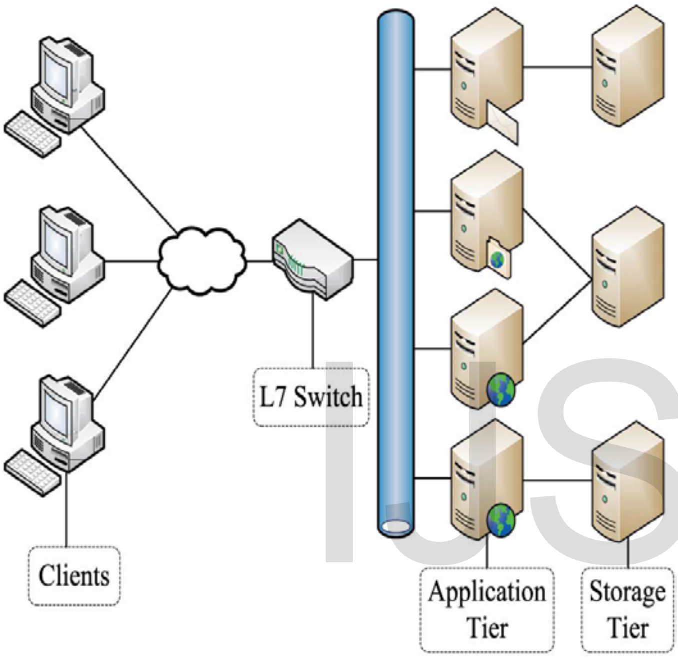

Figure 4. Two Tiered Architecture for Internet Applications

These big data applications are supposed to investigate gigantic amount of data from various data sources from several points of view, uncover new findings, and then deliver totally new values. As big data applications handle extremely huge amount of data compared with conventional applications, there is a high and increasing require for the computational situation, which gather speeds and levels out big data applications. The serious problem here, however, is that the behaviors, or characteristics of big data applications are not clearly defined yet. There is no established model for big data applications right now. Other important data sources, e.g., multimedia data, will be continuously added. With the steady growth of Big Data, the need for a specific benchmark testing the Big Data characteristics of current platforms becomes more important. At the same time, the platforms are becoming more complex as the number of requirements they should address also grows. This makes the creation of an objective Big Data benchmark that covers all relevant characteristics, a complex task.

II. Conclusion

The Paper mainly focuses on the survey of all the existing techniques that are implemented for analyzing the characteristics of Internet or Benchmark Dataset. Hence by analyzing the various Benchmark Dataset on various Datasets a new and efficient technique is implemented in future.

References

[1] L. A. Barroso and U. H¨olzle. The datacenter as a computer: An introduction to the design of warehouse-scale machines. Synthesis Lectures on Computer Architecture, 4(1):1–108, 2009.

[2] W. Gao, Y. Zhu, Z. Jia, C. Luo, L. Wang, J. Zhan, Y. He, S. Gong, X. Li, S. Zhang, and B. Qiu. Bigdatabench: a big data benchmark suite from web search engines. The Third Workshop on Architectures and Systems for Big Data (ASBD 2013), in conjunction with ISCA 2013.

[3] W. Gao, Y. Zhu, Z. Jia, C. Luo, L. Wang, Z. Li, J. Zhan, Y. Qi, Y. He, S. Gong, et al., \Bigdatabench: a big data benchmark suite from web search engines," arXiv preprint arXiv:1307.0320, 2013.

[4] Jeff Dean and Sanjay Ghemawat, *MapReduce: Simplified Data Processing on Large Clusters.* Sixth Symposium on Operating Systems Design and Implementation, 2004: p. 137-150.

[5] Jaliya Ekanayake, Thilina Gunarathne, Judy Qiu, Geoffrey Fox, Scott Beason, Jong Youl Choi, Yang Ruan, Seung-Hee Bae, and Hui Li, Applicability of DryadLINQ to Scientific Applications. 2010.

[6] Geoffrey L. House Yang Ruan, Saliya Ekanayake, Ursel Schütte, James D. Bever, Haixu Tang, Geoffrey Fox. Integration of Clustering and Multidimensional Scaling to Determine Phylogenetic Trees as Spherical Phylograms Visualized in 3 Dimensions. in C4Bio 2014 of IEEE/ACM CCGrid 2014.

[7] Larissa Stanberry, Roger Higdon, Winston Haynes, Natali Kolker, William Broomall, Saliya Ekanayake, Adam Hughes, Yang Ruan, Judy Qiu, Eugene Kolker, and Geoffrey Fox. Visualizing the protein sequence universe. in Proceedings of the 3rd international workshop on Emerging computational methods for the life sciences. 2012, ACM.

[8] Yang Ruan, Saliya Ekanayake, Mina Rho, Haixu Tang, Seung-Hee Bae, Judy Qiu, and Geoffrey Fox. DACIDR: deterministic annealed clustering with interpolative dimension reduction using a large collection of 16S rRNA sequences. In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. 2012.

[9] H. Xi, J. Zhan, Z. Jia, X. Hong, L. Wang, L. Zhang, N. Sun, and G. Lu. Characterization of real workloads of web search engines. In Workload Characterization (IISWC), International Symposium on, volume 11, pages 15–25. IEEE, 2011.

[10] Lei Wang, Jianfeng Zhan, Chunjie Luo, Yuqing Zhu, Qiang Yang, Yongqiang He, Wanling Gao, Zhen Jia," BigDataBench: a Big Data Benchmark Suite from Internet Services", High Performance Computer Architecture (HPCA), 2014 IEEE 20th International Symposium, 2014.

[11] Shengsheng Huang, Jie Huang, Jinquan Dai, Tao Xie, Bo Huang," The HiBench Benchmark Suite : Characterization of the MapReduce-Based Data Analysis", IEEE 2010.

[12] Jianfeng Zhan, Lixin Zhang, Ninghui Sun, Lei Wang, Zhen Jia, and Chunjie Luo," High Volume Computing: Identifying and Characterizing Throughput Oriented Workloads in Data Centers", 2013.

[13] Lei Wang, Jianfeng Zhan, Chunjie Luo, Yuqing Zhu, Qiang Yang, Yongqiang He, Wanling Gao, Zhen Jia," BigDataBench : a big Data Benchmark Suite from Web Search Engines", 2013.

[14] Zhen Jia, Lei Wang, Jianfeng Zhan, Lixin Zhang, and Chunjie Luo," Characterizing Data Analysis Workloads in Data Centers", 2013.

[15] Zijian Ming, Chunjie Luo, Wanling Gao, Rui Han, Qiang Yang, Lei Wang, and Jianfeng Zhan," BDGS: A Scalable Big Data Generator Suite in Big Data Benchmarking", 2014.

[16] Ahmad Ghazal, Tilmann Rabl, Minqing Hu,Francois Raab, Meikel Poess, Alain Crolotte, Hans-Arno Jacobsen," BigBench: Towards an Industry Standard Benchmark for Big Data Analytics", ACM 2013,

[17] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica," Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing", 2011.