An Efficient BigDataBench for the Analysis of Internet Services

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— Big data is a growing field that pushes the limits of information collection and analysis. More and more entities are seeking ways to use big data. As the big data industry continues to grow and establish common needs and trends, meaningful benchmarks will be a way to compare different systems and allow engineers to design better solutions and consumers to make informed purchases. Here in this paper a new and efficient algorithm for Internet Things on BigData Bench dataset is proposed. The Algorithm implemented here is applied over various Datasets on different Workloads and the methodology provides higher instructions to be executed per second and generates high SpeedUp ratio as compared to the existing methodology for Internet on Things. The Proposed Methodology also provides Low Instruction Breakdown and the results can't variates for files of less or higher sizes.

Index Terms—BigDataBench, Internet on Things, Hadoop, Cloud Computing, Virtual Machine, Semantic Similarity, Workloads.

I. Introduction

Big data is a growing field that pushes the limits of information collection and analysis. More and more entities are seeking ways to use big data. As the big data industry continues to grow and establish common needs and trends, meaningful benchmarks will be a way to compare different systems and allow engineers to design better solutions and consumers to make informed purchases. Big data is an emerging field for businesses, scientists, and governments around the world. The increase of available data can be mainly attributed to the rapid globalization of the internet and the growth of embedded systems providing real time data. Google estimated that in 2011, 32.77% of the world populations were internet users. That equates to about 2.3 billion users creating data. In March 2012, IBM estimated that 2.5 quintillion bytes of data are created every day [1]. Companies are now offering a range of different solutions to manage and analyze these massive data sets. However, there is not yet an industry standard for benchmarking these systems or comparing their performances.

Big data is one of the fastest growing fields in data processing today. It cuts across numerous industrial and public service areas, from web commerce to traditional retail, from genomics to geospatial, from social networks to interactive media, and from power grid management to traffic control. New solutions appear at a rapid pace, and the need emerges for an objective method to compare their applicability, efficiency, and cost. In other words, there is a growing need for a set of big data performance benchmarks. As the big data industry persists to produce and start widespread requires and developments, significant benchmarks will be a method to evaluate different schemes and permit engineers to plan better explanations and consumers to make knowledgeable acquires. There have been a number of efforts at creating big data benchmarks [2-4]. None of them has increased extensive recognition and large procedure. It continues an indefinable objective to estimate an extensive range of projected big data solutions. The area of big data performance is in a condition where every learning and maintain utilizes a different method. Results from one publication to the subsequently are not equivalent and frequently not even intimately associated, as it was the case for OLTP some twenty years ago and for choice sustain abruptly subsequently.

Figure- 1: Big Data Application Domains.

With the development of computer and network technology, as well as intelligent systems is common used in modern life, big data has become increasingly close to people's daily lives. In 2008, Big Data issue released by "Nature" pointed out the

importance of big data in biology, and it was necessary to build biological big data system to solve complex biological data structure problem [5]. Paper [5] pointed out that the new big data system must be able to tolerate various structures of data and unstructured data, has flexible operability and must ensure data reusability. Furthermore, Big Data plays an important role in the defense of national network digital security, maintaining social stability and promoting sustainable economic and social development [6].

In modern cloud data centers, a large number of tenants are consolidated to share a common computing infrastructure and execute a diverse mix of workloads. Benchmarking and understanding these workloads is a key problem for system designers, programmers and researchers to optimize the performance and energy efficiency of data center systems and to promote the development of data center technology. This work focuses on two classes of popular data center workloads [24]:7 Long-running services. These workloads offer online services such as web search engines and e-commerce sites to end users and the services usually keep running for months and years. The tenants of such workloads are service end users.

– Short-term data analysis jobs. These workloads process input data of many scales using relatively short periods (e.g. in Google and Facebook data centers, a majority (over 90%) of analytic jobs complete within a few minutes [7, 8]). The tenants of such workloads are job submitters.

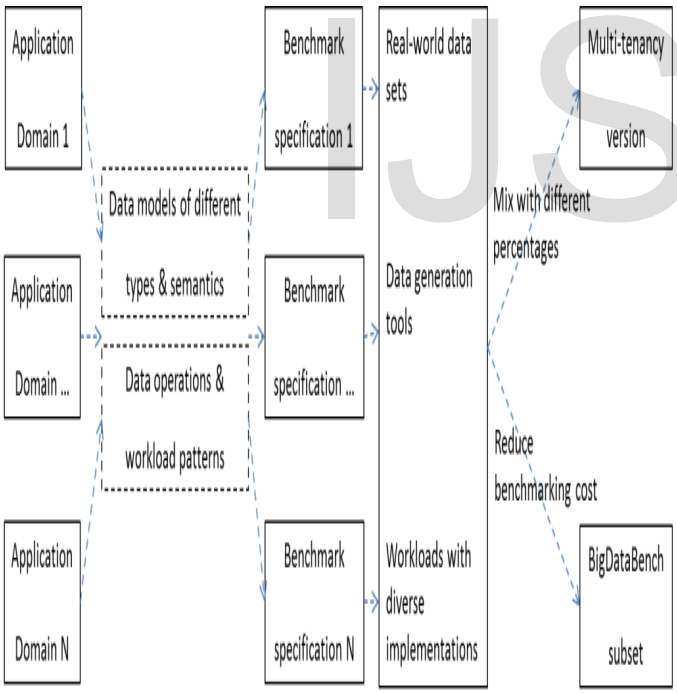

Figure- 2: Big Data Benchmarking Methodology.

II. Literature Survey

In the era of big data it is a phenomenon often appears that useful information is being submerged in a large number of useless information [9]. The data quality of Big Data has two problems: how to manage large-scale data and how to wash it. During the cleaning process, if the cleaning granularity is too small, it is easy to filter out the useful information; if the cleaning granularity is too common, it can't achieve the real cleaning effect.

Leimeister et al. [11] argue that the actors in the Cloud form a business value network moderately than a conventional business significance series. We identify the following actors in a Cloud-centric business value network, Figure 12: IT Vendors develop infrastructure software and operate infrastructure services; Service Providers develop and operate services; Service Aggregators offer new services by combining preexisting services; Service Platform Providers offer an environment for developing Cloud applications; Consulting supports customers with selecting and implementing Cloud services; Customers are the end-users of Cloud services.

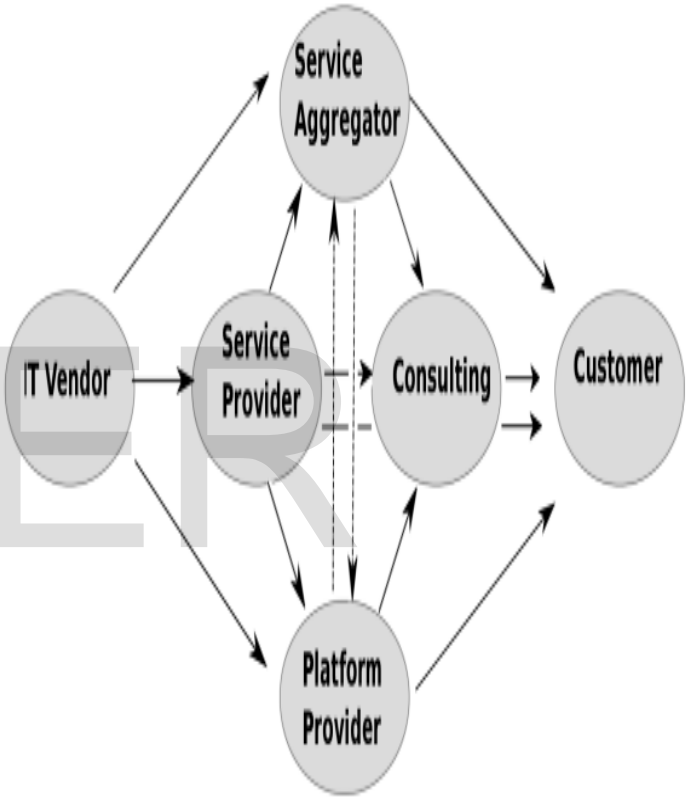

Figure- 3: Cloud Actors and their Value Network [11]

Big-Bench [10] is the modern attempt in the direction of planning big data benchmarks. BigBench focuses on big data offline analytics, thus accepting TPC-DS as the origin and adding up a top novel data types like semi un-structured data, as well as non-relational workloads. Even though BigBench has a entire exposure of data types, its object under test is DBMS and MapReduce methods that declare to give big data explanations, guiding to limited exposure of software stacks. In addition at this time, it is not open-source for simple procedure and acceptance.

Chen et al. [12] found that application outlines from larger-scale MapReduce Clusters organized in Facebook and Cloudera did not fit well-know statistical distributions. In this

case, only real information can return the real system performances and workload features and hence the real world data is desired in big data benchmarks.

Aashish et al. analyze redundancy in the SPEC CPU2006 benchmark suite using micro-architecture metrics. They illustrate suggestion on comparison of the benchmarks and reach your destination at significant subsets; and these subsets are representative of an extensive variety of applications areas without having many benchmarks with comparable features. The research consequence could clearly reduce execution time for system architecture researches [13].

In this paper [14], they present a complete conversation of the BigBench measurement together with the database and the workload. In the development of extending BigBench they have acquired view from leading industry skilled about the significance in addition to entirety of the workload.

III. Proposed Methodology

Consider a cloud Computing Environment with as the number of data centers is the number of Brokers of the cloud is the number of hosts and is the number of physical virtual machines with N number of Cloudlets and Resources .

1. Cloud consumers can submit their requests for the access of resources to the brokers. Each of the requests from the cloudlets is allocated to their respective brokers who can process their requests.
2. Virtual machines can be dynamically started and stopped on a single physical machine according to the incoming requests, hence providing the flexibility of configuring various partitions of resources on the same physical machine to different requirements of service requests. Multiple VMs can concurrently run applications based on different operating system environments on a single physical machine. By dynamically migrating VMs across physical machines, workloads can be consolidated and unused resources can be switched to a low-power mode, turned off or configured to operate at low-performance levels (e.g. using DVFS) in order to save energy.
3. The underlying physical computing servers provide the hardware infrastructure for creating virtualized resources to meet service demands.

Currently, resource allocation in a Cloud data center aims to provide high performance while meeting SLAs, without focusing on allocating VMs to minimize energy consumption. To explore both performance and energy efficiency, three crucial issues must be addressed. First, excessive power cycling of a server could reduce its reliability. Second, turning resources off in a dynamic environment is risky from the QoS perspective. Due to the variability of the workload and aggressive consolidation, some VMs may not obtain required resources under peak load, and fail to meet the desired QoS. Third, ensuring SLAs brings challenges to accurate application performance management in virtualized environments. All these issues requires effective consolidation policies that can minimize energy consumption without compromising the user-specified QoS requirements.

Allocation of Virtual Machines

Here the allocation of virtual machines is based on the entrance of new requests for the provisioning of Virtual Machines and then allocating of virtual machines on hosts and then optimization of the current allocation of virtual machines. The proposed algorithm implemented here uses Bin backing algorithm which is based on Modified Best Fit Decreasing (MBFD) algorithm in which sorting of all VMs in decreasing order of their current CPU utilizations, and allocate each VM to a host that provides the least increase of power consumption due to this allocation. This allows leveraging the heterogeneity of resources by choosing the most power-efficient nodes first.

Algorithm: Modified Best Fit Decreasing (MBFD)
Input: HostList & VmList
Output: Allocation of VM's

1. First of all sort the list of virtual machine lists in decreasing order of their Utilization.
2. For each of the Virtual machine repeat
3. manpower ← MAX
4. allocatedHost ← NULL
5. for each of the host in HostList do
6. if host has enough resource for VM then
7. power ← estimatePower(host,VM)
8. if power < manpower then
9. allocatedHost ← host
10. manpower ← Power
11. if allocatedHost ≠ NULL then
12. allocated VM to allocatedHost
13. return allocation

- **Data content similarity (SimC)**

It is the Cosine similarity between the term frequency vectors of d1 and d2:

$$SimC(d1,d2) = \frac{V_{d1} * V_{d2}}{\left\|V_{d_1}\right\| * \left\|V_{d_2}\right\|} \quad (1)$$

WhereVd is the frequency vector of the terms inside data unit d, ||Vd|| is the length of Vd, and the numerator is the inner product of two vectors.

- **Number of Common Neighbors**

It is defined as the total number of nodes that are connected directly in relationship with node x and y for unweighted network,

$$CN(x,y) = \varphi(x) \cap \varphi(y) \quad (2)$$

Where, φ is the set of neighbors of node x.

$\varphi$is the set of neighbors of node y.

To calculate link prediction between nodes for unweightednetwork common neighbors can be calculated as,

$$CN(x,y) = \sum_{z \in \varphi(x) \cap \varphi(y)} w(x,z) + w(y,z) \quad (3)$$

- **Jaccard Coefficient**

It is defined as the highest proportion of common neighbors to the total number of neighbors in the network. The Jaccard Coefficient can also defined for weighted as well for unweighted network.

For unweighted network,

$$JC(x,y) = \frac{\varphi(x) \cap \varphi(y)}{\varphi(x) \cup \varphi(y)} \quad (4)$$

For weighted network,

$$JC(x,y) = \sum_{z \in \varphi(x) \cap \varphi(y)} \frac{w(x,z) + w(y,z)}{\sum_{a \in \varphi(x)} w(a,x) + \sum_{b \in \varphi(y)} w(b,y)}$$

Predict the most valuable words from the text documents having most similarity between words.

(4.5)

IV. Result Analysis

The Table shown below is the analysis and comparison of various Workloads on Existing and propose work. The proposed Methodology implemented here provides better L3 Cache Configuration.

	L3 cache MPKI of different configurations in big data workloads	
Workloads	**Existing Work**	**Proposed Work**
WordCount	2.1	1.7
Scan	3	2.3
Sort	2	1.2
Read	1.4	0.8
PageRank	1.45	0.85
Index	1.4	0.8

Table 1: L3 Cache MPKI of different configurations in big data workloads

The Table shown below is the analysis and comparison of various Workloads on Existing and propose work. The proposed Methodology implemented here provides better MPIS.

	MIPS of different workloads	
Workloads	**Existing Work**	**Proposed Work**
WordCount	16000	21000
Scan	1000	3000
Sort	6000	7400
Read	1000	3500
PageRank	3000	7000
Index	1000	2200

Table 2: MIPS of different Workloads

The Table shown below is the analysis and comparison of various Workloads on Existing and propose work. The proposed Methodology implemented here provides better Speedup.

	Speedup of Different Workloads	
Workloads	**Existing Work**	**Proposed Work**
WordCount	1	2.4
Scan	1	3
Sort	1	1.7
Read	1	3
PageRank	1	3.1
Index	1	2.5

Table 3: Speedup of different Workloads

The Figure shown below is the analysis and comparison of various Workloads on Existing and propose work. The proposed Methodology implemented here provides better L3 Cache Configuration.

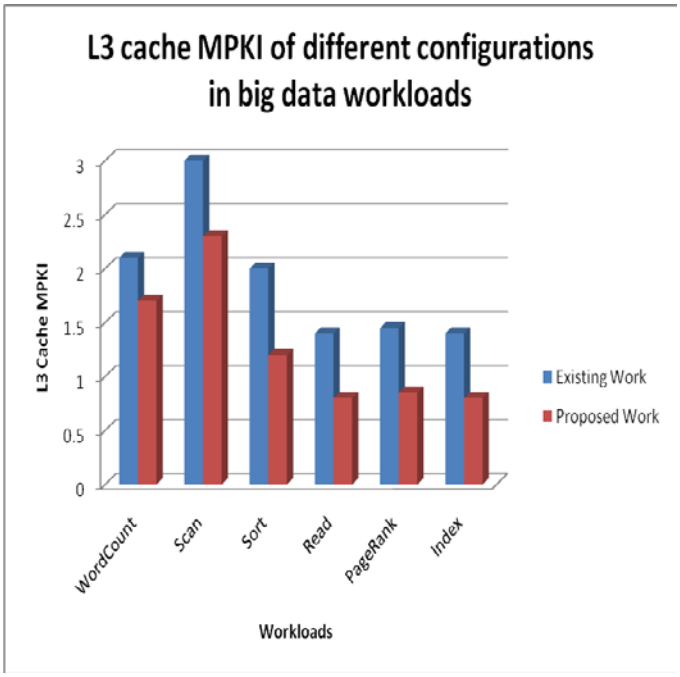

Figure 4: L3 Cache MPKI of Different Configurations

The Figure shown below is the analysis and comparison of various Workloads on Existing and proposes work. The proposed Methodology implemented here provides better MIPs.

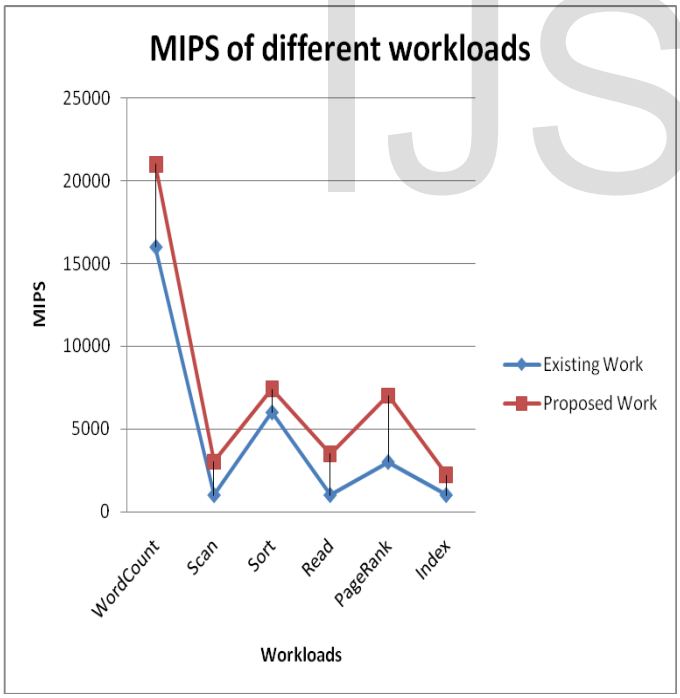

Figure 5: MIPS of different workloads

The Figure shown below is the analysis and comparison of various Workloads on Existing and propose work. The proposed Methodology implemented here provides better Speedup.

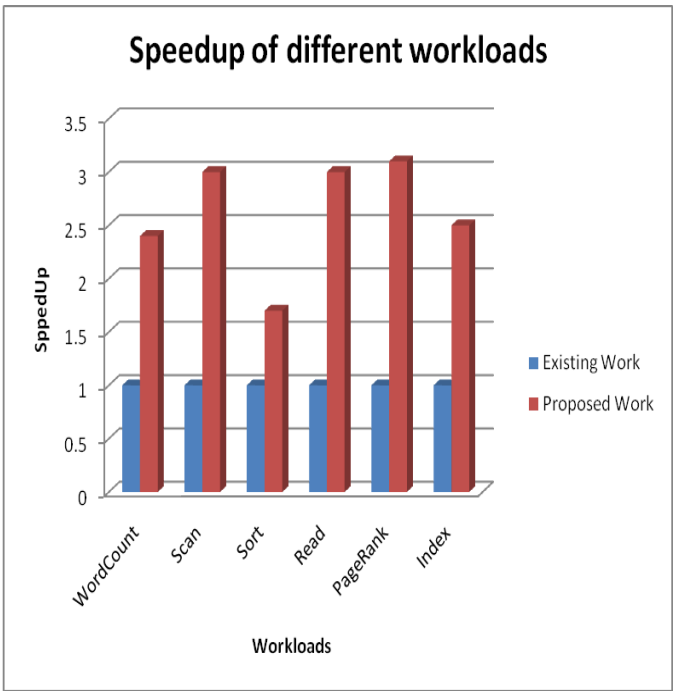

Figure 6: Speedup of different Workloads

V. Conclusion

The proposed methodology implemented here for the Internet Services on BigData Bench Dataset using Similarity Metrics is proposed. The Algorithm implemented here is applied over various Datasets on different Workloads and the methodology provides higher instructions to be executed per second and generates high SpeedUp ratio as compared to the existing methodology for Internet on Things. The Proposed Methodology also provides Low Instruction Breakdown and the results can't variates for files of less or higher sizes.

References

[1] Singh, S.; Singh, N., "Big Data analytics," Communication, Information & Computing Technology (ICCICT), 2012 International Conference on , vol., no., pp.1,4, 19-20 Oct. 2012.
[2] B. Cooper et al. Benchmarking cloud serving systems with ycsb. In SOCC 2010
[3] Z. Fadika, E. Dede, M. Govindaraju, and L. Ramakrishnan.Benchmarking mapreduce implementations for application usage scenarios. In GRID 2011
[4] M. Ferdman et al. clearing the clouds, a study of emerging scale-out workloads on modern hardware.In ASPLOS 2012.
[5] John Boyle. Biology must develop its own big-data systems.Nature. 2008, 499(7): 7.
[6] Wang Yuan-Zhuo, Jin Xiao-Long, Chen Xue-Qi.Network Big Data: Present and Future [J].Chinese Journal of Computer. 2013, 36(6):1125-1138.
[7] Reiss, C., Tumanov, A., Ganger, G.R., Katz, R.H., Kozuch, and M.A.: Heterogeneity, dynamicity of clouds at

scale: Google trace analysis. In: Proceedings of the Third ACM Symposium on Cloud Computing, pp. 7. ACM, 2012
[8] Chen, Y., Alspaugh, S., Katz, R.: Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads. VLDB 5(12), 1802–1813, 2012
[9] Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony, and David R. Easterling. Climate Data Challenges in the 21st Century [J]. Science, 2011, 331(6018):700-702
[10] A. Ghazal, M. Hu, T. Rabl, F. Raab, M. Poess, A. Crolotte, and H.-A.Jacobsen.Bigbench: Towards an industry standard benchmark for big data analytics. In SIGMOD 2013.
[11] Leimeister, S., Böhm, M., Riedl, C., Krcmar, H.: The business perspective of cloud computing: Actors, roles and value networks. In Alexander, P.M., Turpin, M., van Deventer, J.P., eds.: ECIS. 2010.
[12] Y. Chen et al, "We Don't Know Enough to make a Big Data Benchmark suite". Workshop on Big Data Benchmarking. 2012
[13] A. Phansalkar, A. Joshi, and L. K. John, "Analysis of redundancy and application balance in the SPEC CPU2006 benchmark suite". 2007. In Proceedings of the 34thannual international symposium on Computer architecture (ISCA '07). ACM, New York, NY, USA, 412-423.
[14] ChaitanyaBaru, MilindBhandarkar, "Discussion of BigBench: A Proposed Industry Standard Performance Benchmark for Big Data" 2015.