© 2019 IJSRST | Volume 6 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology
DOI : 10.32628/IJSRST196292

An Analytical Review on High Performance Cloud Computing in Big Data

Prof. Neha Purohit[1], Sarang A. Mutkure[2], Pranay G. Sawarkar[3]

[1]Assistant Professor, Department of Master of Computer Applications, G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India

[23]PG Scholar, Department of Master of Computer Applications, G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India

ABSTRACT

Cloud computing is the product of the traditional computer technology and network technology development integration. Such as grid computing, distributed computing, parallel computing, utility computing, network storage, virtualization, and load balancing. To perform integrating multiple relative low-cost computing entities into one perfect system with powerful computing ability via the network and with the help of the SaaS, PaaS, IaaS, MSP and other advanced business models distributing this powerful computing ability to the hands of the end user. Cloud computing system mainly uses MapReduce model. The core design idea of MapReduce is to divide and conquer the problem and calculate on data rather than push data to calculate which effectively avoids many communication costs generated during data transmission.

Keywords: Big Data, Cloud computing, High-performance, HPC.

I. INTRODUCTION

With the advent of the digital age, the amount of data generated, stored and shared has been on the rise. From data warehouses, web pages and blogs to audio/video streams, all of these are sources of massive amounts of data. The result of this proliferation is the generation of massive amounts of pervasive and complex data, which needs to create efficiently, stored, shared and analysed to extract useful information [1, 15, 16].

Since innovations in data architecture are on our doorstep, the 'big data' paradigm refers to very large and complex data sets (i.e., petabytes and Exabyte's of data) that traditional data processing systems are inadequate to capture, store and analyse, to seek to glean intelligence from data and translate it into competitive advantage. As a result, Big data needs more computing power and storage provided by cloud computing platforms.

II. BIG DATA MANAGEMENT

The architecture of Big Data must synchronize with the support infrastructure of the organization. To date, all of the data used by organizations are stagnant. Data increasingly sourced from various fields that are disorganized and messy, such as information from machines or sensors and abundant sources of public and private data [2]. Previously, most companies were unable to either capture or store these data, and available tools could not manage the data in a reasonable amount of time. However,

the new Big Data technology improves performance, facilitates innovation in the products and services of business models Big Data technology aims to minimize hardware and processing costs and to verify the value of Big Data before committing significant company resources [3]. Correctly, managed Big Data are accessible, reliable, secure, and manageable. Hence, Big Data applications can apply in various complex scientific disciplines (either single or interdisciplinary), including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry[16]. In the following section, we briefly discuss data management tools and propose a new data life cycle that uses the technologies and terminologies of Big Data.

A. Management Tools

With the evolution of computing technology, large volumes can manage without requiring supercomputers and high cost. Many tools and techniques are available for data management, including Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort [3]. However, companies must develop specialized tools and technologies that can store, access, and analyze large amounts of data in near-real time because Big Data differs from the traditional data and cannot store in a single machine. Furthermore, Big Data lacks the structure of traditional data [4]. For Big Data, some of the most commonly used tools and techniques are Hadoop, MapReduce, and Big Table. These innovations have redefined data management because they efficiently process large amounts of data efficiently, cost-effectively, and promptly. The following section describes Hadoop and MapReduce in further detail, as well as the variousprojects/frameworks that are related to and suitable for the management and analysis of Big Data [5, 6].

B. Hadoop

Hadoop written in Java and is a top-level Apache project that started in 2006. It emphasizes discovery from the perspective of scalability and analysis to realize near-impossible feats. Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could apply in a distributed system. Presently, used on large amounts of data with Hadoop; enterprises can harness data that was previously difficult to manage and analyse. Hadoop used by approximately 63% of organizations to manage a vast number of unstructured logs and events (Sys.con Media, 2011).

C. HDFS

This paradigm applied when the amount of data is too much for a single machine. HDFS is more complicated than other file systems given the complexities and uncertainties of networks [22]. The cluster contains two types of nodes. The first node is a name-node that acts as a master node. The second node type is a data node that acts as a slave node. This type of node comes in multiples. Aside from these two types of nodes, HDFS can also have secondary name-node. HDFS stores files in blocks, the default block size of which is 64 MB. All HDFS files are replicated in multiples to facilitate the parallel processing of large amounts of data.

III. PROCESSING OF BIG DATA IN CLOUD COMPUTING

Cloud computing as an essential application environment for big data has attracted tremendous attention from the research community. Remarkable

progress of Big data networking has also reported in this area. In this section, we introduce Big data research issues and solutions related to Cloud Computing [18, 20]. Individually, we are interested in the following topics: opportunities and challenges of Big data networking in Cloud Computing, cloud resource management of big data, and performance optimization of big data in Cloud Computing.

Data is the central element of communication and collaboration on the Internet and all the applications that built on this platform. The immense popularity of data-intensive applications like Facebook, LinkedIn, Twitter, Amazon, eBay, and Google+ contributes to increasing requirement of storage and processing of data in the cloud environment [20].

Therefore, they require high-performance processors to do the job. The cloud provides an excellent platform for big data storage, processing, and analysis, addressing two of the primary requirements of big data analytics, high storage, and high-performance computing [21].

The cloud-computing environment offers development, installation, and implementation of software and data applications 'as a service.' Three multi-layered infrastructures namely, platform as a service (PaaS), software as a service (SaaS), and infrastructure as a service (IaaS) exist. Infrastructure-as-a-service is a model that provides computing and storage resources as a service. On the other hand, in case of PaaS and SaaS, the cloud services provide software platform or software itself as a service to its clients [19].

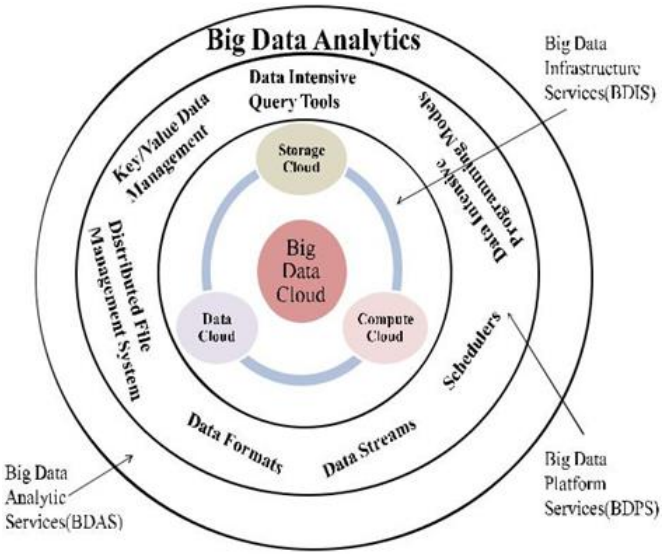

Figure 1. Big Data Cloud Computing

Since innovations in data architecture are on our doorstep, the 'big data' paradigm refers to very large and complex data sets (i.e., petabytes and exabytes of data) that traditional data processing systems are inadequate to capture, store and analyze, to seek to glean intelligence from data and translate it into competitive advantage [16]. As a result, Big data needs more computing power and storage provided by cloud computing platforms. In this context, cloud providers, such as IBM, Google, Amazon, and Microsoft, provide network-accessible storage priced by the gigabyte-month and computing cycles priced by the CPU-hour.

Although big data is still in the preliminary stages, comprehensive surveys exist in the literature [1, 9–11, 20]. This survey article aims at providing a holistic perspective on big data and big data-as-a-service (BDaaS) concepts to the research community active on big data-related themes, including a critical revision of the current state-of-the-art techniques, definition, and extensive researches issues. Following this introductory section, Sect. 2 presents related work approaches in the literature, including the architecture and possible impact areas. Section 3 demonstrates the business value and long-term

benefits of adopting big data-as-a-service business [18].

Another significant challenge is the delivery of Big data capabilities through the cloud. The adoption of Big data-as-a-service (BDaaS) business models enables the efficient storage and management of massive datasets and data processing from an outside provider, as well as the exploitation of a full range of analytics capabilities (i.e., data and predictive analytics or business intelligence provided as service-based applications in the cloud).

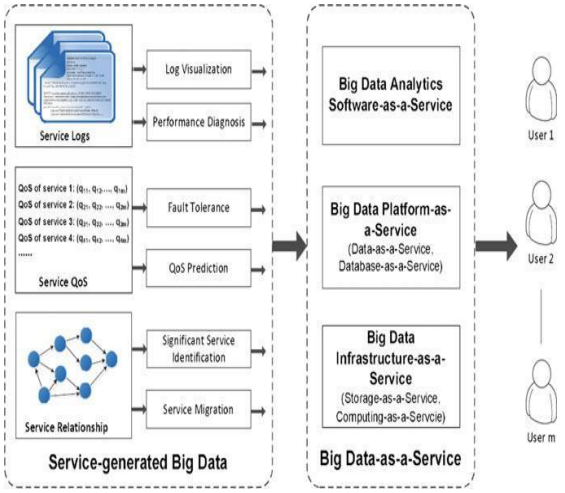

Figure 2. Service-Generated Big Data and Big Data-As-A-Service

IV. PERFORMANCE OPTIMIZATION

Performance optimization is yet another classic and essential topic in cloud computing because appropriate optimization techniques will provide better application experiences with comparable or even less system resource consumption, compared to non-optimized cases.

A dataflow-based performance analysis tool for big data cloud, i.e., Hitune, presented in [Dai11]. Hitune shown to be useful in assisting users doing Hadoop performance analysis and system parameter tuning. Limitations of existing approaches, such as Hadoop logs and metrics compared and discussed. A few interesting case studies on Big data processing in cloud computing environment depicted [17, 18] . Efforts of the Fujitsu laboratory based on the data store and complex event processing, as well as workflow description in distributed data processing.

A recent online cost-minimization algorithm depicted in [Zhang10]. The proposed work specifically focused on real-time cost minimizations for uploading massive and dynamic data onto the cloud. The two online algorithms have achieved competitive cost reduction ratios [19]. However, the proposed methods evaluated on a limited scale. The proposed algorithms need further evaluated at more extensive and more competitive scales, e.g., data streaming applications with larger topologies.

V. PERFORMANCE OPTIMIZATION

In scientific applications, data commonly represented by a multi-dimensional array-based data model. For instance, the widely used Community Earth System Model (CESM) software package consists of four separate modules simultaneously simulating the earth atmosphere, ocean, land surface, and sea-ice, and each module uses the multi-dimensional arrays data model [9]. A typical example is a 3-dimensional temperature data with longitude, latitude, and time dimensions. It is often needed to compute the moving average, median, lowest and highest temperature with specified conditions such as areas and periods. Such computed results will further correlated with the computed results from other parameters, such as the humidity and wind velocity, to predict weather conditions [10].

The current way of conducting such processing is to read the required data (e.g., a sub-array of the affected area) from storage servers to compute nodes, perform computations on desired data with specified conditions, such as those data shown in a shaded area, and then write the output back to storage [7,8]. For CESM, an experimental test shows that the data access and movement time for the calculation of the moving average, median, lowest and highest degrees can occupy 88.2%, 95.4%, 96.6%, and 96.6% of the total execution time on a cluster, where 128GB of data retrieved to 272 nodes for processing.

CESM has data retrieval and processing phases and computing and simulation phases, as many scientific Big data applications do. The basic idea of the new decoupled HPC system architecture is to change the conventional architecture to handle these two phases differently on different nodes. Such architecture decouples nodes into compute nodes and data processing nodes [14, 15]. These nodes are mapped with computation-intensive operations and data-intensive operations respectively. Computation-intensive operations executed on massive compute nodes. Data-intensive operations executed on dedicated data processing nodes. In other words, the decoupled architecture reshapes the current pattern of retrieve - compute - store cycles into retrieving (generate) - reduce - compute - reduce - store cycles as shown in Figure 1, where the reduce phases are designed to conduct offloaded data-intensive operations and reduce data size before moving data across the network. This retrieval, reduce, compute, and store phases can be pipelined to overlap the I/O, communication, and computation times. From one point of view, the decoupled architecture is an enhanced framework of MapReduce [10], where one node with its local storage does not conduct the reduction, but a set of (data) nodes and the global storage so that that parallel computing features can maintain. From another point of view, the data nodes are the data-access accelerators, to speed up data accesses and reduce data size before sending data across the network.

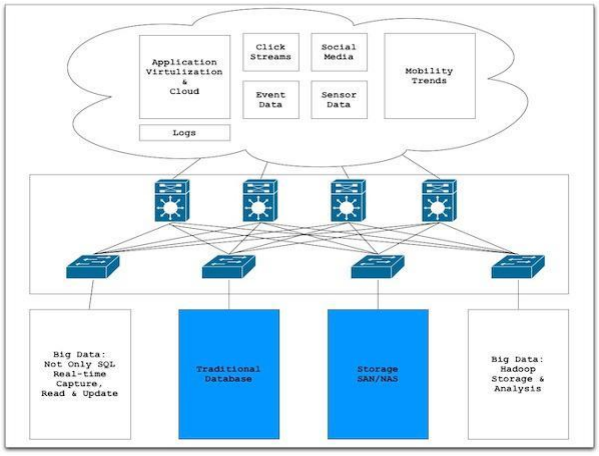

Figure 3. General Framework of Big Data Networking.

VI. CHALLENGES

While the rise of big data yields enormous opportunities for individuals, organizations and the society, it also raises significant privacy and ethical issues [16, 17]. These issues are factors that may lead to situations in which the underlying analytic models and infrastructures are likely to impact privacy negatively from both a legal and an ethical perspective and hence represent possible obstacles for the big data's potential to be fully realized.

Big data analytics essentially requires very high computing capabilities to drive data into meaningful insights. High-performance data analytics, HPDA, seeks to widen the HPC and Big data analytics application domains by augmenting with other related technologies [17,19]. However, varied and complex requirements of big data analytics pose many challenges at micro as well as macro level. At the micro level, there are unusual and specific issues about statistical modelling of big data. At the macro

level, big data analytics challenged by the complexities of working computational prototypes.

- **Data storage and management:** Since big data are dependent on extensive storage capacity and data volumes grow exponentially, the current data management systems cannot satisfy the needs of big data due to limited storage capacity. Also, the existing algorithms are not able to store data efficiently because of the heterogeneity of big data.

- **Data Transmission and Curation:** Since network bandwidth capacity is the major drawback in the cloud, data transmission is a challenge to overcome, especially when the volume of at enormous huge. For managing large-scale and structured datasets, data warehouses and data marts useful good approaches. Data warehouses are relational database systems that enable the data storage, analysis, and reporting, while the data marts are based on data warehouses and facilitate the analysis of them. In this context, NoSQL databases introduced as a potential technology for large and distributed data management and database design. The significant advantage of NoSQL databases is the schema-free orientation, which enables the quick modification of the structure of data and avoids rewriting the tables.

- **Data processing and analysis:** Query response time is a significant issue in big data, more time needed when traversing data in a database and performing real-time analytics. A flexible and reconfigured grid along with the big data pre-processing enhancement and consolidation of application and data-parallelization schemes more efficient active approaches for extracting more meaningful knowledge from the given data sets.

- **Data privacy and security:** Since the host of data or other critical operations performed by third party services or infrastructures, and security issues witnessed concerning big data storage and processing. The current technologies used in data security are mainly static data-oriented, although big data entails the dynamic change of current and additional data or variations in attributes. Privacy-preserving data mining without exposing sensitive personal information is another challenging field to investigate.

VII. CONCLUSION

The information-driven economy relies on the actionable insights extracted from data analytics. The era of data revolution has paved a way to the need of convergence of paradigms like High-Performance Computing and Big Data Analytics. The amalgamation of these paradigms is a herculean task involving various aspects of data management and computing efficiency. HPC with Big data has given rise to the evolution of the data storage technologies and computing models. The transformation of traditional analytical paradigms to cater to the requirement of the intense data applications and High-Performance Computing is the need of the hour.

The convergence of the paradigms "High-Performance Computing" and "Big Data Analytics" can lead to a sustainable solution for the data-driven applications. The continuous flow of "real" data which is the predominant type of data seen in data-intense applications needs to be handled by a different architectural platform termed as "Real Time Analytical Framework." The computational requirements of these newer models are different

from the traditional models, and hence the evolution of the models becomes the critical challenge of High-Performance Data Analytics.

VIII. REFERENCES

[1]. Big Data: The next frontier for innovation, competition and productivity. James Maniyka, Executive summary ,McKinsey Global Institute ,May 2011, http://www.mckinsey.com/mgi/publication/big. data/MGI _big_data_exec_summary.pdf>.

[2]. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf [Accessed on 2nd January 2015]

[3]. Beyond the hype: Big data concepts, methods, and analytics, Amir Gandomi , Murtaza Haide, International Journal of Information Management 35 (2015) 137–144

[4]. https://analyticsacademy.withgoogle.com/course01/asses/pdf/DigitalAnalyticsFundamentals-Lesson2.1TheimportanceofdigitalanalyticsText.pdf [Acesses on 27th December 2014]

[5]. Big Data Meets High Performance Computing Intel® Enterprise Edition for Lustre* software and Hadoop combine to bring big data analytics to high performance computing configurations.http://www.intel.com/content/dam/www/public/us/en/do cuments/white-papers/big-data-meets-high-performance-computing-white-paper.pdf

[6]. Agrawal, D., Das, S., El Abbadi, A.: Big data and cloud computing: current state and futureopportunities. In: Proceedings of the 14th International Conference on Extending Database Technology (EDBT/ICDT'11), pp. 530–533 (2011)Amazon Web Services, Inc.: Elastic Compute Cloud (EC2). http://aws.amazon.com/ec2 (2015). Accessed 18 Oct 2015

[7]. Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A.S., Buyya, R.: Big data computing and clouds: trends and future directions. J. Parallel Distrib. Comput. 79–80, 3–15 (2015)

[8]. Batalla, J.M., Kantor, M., Mavromoustakis, C.X., Skourletopoulos, G., Mastorakis, G.: A novel methodology for efficient throughput evaluation in virtualized routers. In: Proceedings of the IEEE International Conference on Communications (ICC 2015)—Communications Software, Services andMultimedia Applications Symposium (CSSMA), London, UK, pp. 6899–6905 (2015)

[9]. Zheng, Z., Zhu, J., Lyu, M.R.: Service-generated big data and big data-as-a-service: an overview. Proceedings of the 2013 IEEE International Congress on Big Data (BigData Congress), pp. 403– Santa Clara, California (2013)

[10]. Zhang, Linquan, et al. "Moving Big Data to The Cloud: An Online Cost Minimizing Approach." IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 31.12 (2013): 1.http://i.cs.hku.hk/~fcmlau/papers/info13-lq-m.pdf

[11]. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of "big data" on cloud computing: review and open research issues. Inf. Syst. 47, 98–115 (2015)

[12]. IBM Corporation: IBM big data & analytics hub: the four V's of big data. http://www.com/infographic/four-vs-big-data (2014). Accessed 18 Oct 2015

[13]. IBM Corporation: IBM social media analytics software as a service. http://www-03.ibm.com/software/products/en/social-media-analytics-saas (2015a). Accessed 18 October 2015.

[14]. Ranjith, J. M. Balajee and C. Kumar," Trust computation methods in mobile ADHOC

network using glomosim: A Review" International Journal of Scientific Research and Modern Education, Vol. I, Issue I, pp. 777-780, Nov.2016.

[15]. Janarthanan Y, Balajee J.M, and Srinivasa Raghava S. "Content based video retrieval and analysis using image processing: A review."International Journal of Pharmacy and Technology 8, no.4 (2016): 5042-5048.

[16]. Jeyakumar, Balajee, MA Saleem Durai, and Daphne "Case Studies in Amalgamation of Deep Learning and Big Data." In HCI Challenges and Privacy Preservation in Big Data Security, pp. 159-174. IGI Global, 2018.

[17]. Kamalakannan, S. "G., Balajee, J., Srinivasa ,"Superior content-based video retrieval system according to query image"." International Journal of Applied Engineering Research 10, no. 3 (2015): 7951-7957.

[18]. Priya, V., Subha, S., & Balamurugan, B. (2017). Analysis of performance measures to handle medical E-commerce shopping cart abandonment in Informatics in Medicine Unlocked.

[19]. D Lakshmi narayanan. J and Balajee. J," A study of behavior on information system in a university campus by analysis of people mobility" International journal of research in computer application & management, Vol. 6, Issue 7, pp. 29-31, Jul.2016.

[20]. Ranjith, D., J. Balajee, and C. Kumar. "In premises of cloud computing and models." International Journal of Pharmacy and Technology 8, no. 3 (2016)

[21]. Sethumadahavi R Balajee J "Big Data Deep Learning in Healthcare for Electronic Health Records," International Scientific Research Organization Journal, vol. 2, Issue 2, pp. 31–35, Jul. 2017.

[22]. Ushapreethi P, Balajee Jeyakumar and BalaKrishnan P, Action Recongnition in Video Survillance Using Hipi and Map Reducing Model, International Journal of Mechanical Engineering and Technology 8(11), 2017,pp. 368–375.

Cite this article as :

Prof. Neha Purohit, Sarang A. Mutkure, Pranay G. Sawarkar, "An Analytical Review on High Performance Cloud Computing in Big Data", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 6 Issue 2, pp. 441-448, March-April 2019.
Journal URL : http://ijsrst.com/IJSRST196292