A Review on Cloud Based Big Data Analytics

Amitkumar Manekar[a,*] and Dr. G. Pradeepini[b]
[a,*] Corresponding Author: KL University, Computer Engineering, Vijaywada, India.
E-mail: asmanekar24@gmail.com
[b] KL UNIVERSITY/Computer Engineering, Vijaywada, India.

***Abstract*— Today's computing world is facing tsunami and driving without riding on this tsunami towards next generation computing is no choice. So many IT companies decided to grow up with this tsunami like technology. One of these is cloud computing and another is Big data. Currently more than 5 billion mobile users, nearly same facebook, and other social media user generate this tsunami of data. On another side to deliver this services of big data a model called as cloud computing is spreading everywhere as next generations IT Service model. Both technologies continue to evolve. Ultimately, as a cloud, computing development matures, every top mind of organizations will think for development of efficient and agile cloud environment. At the other side, every cloud provider offers the services to the huge number data processing companies that generate data process data and make decision on cloud infrastructure. Ultimately its today's need to think on futures efficient cloud based Big data analytics In this review paper we are focusing on, how we can club Big data and cloud Computing in one frame of development.**

***Keywords*—Big Data, Cloud Computing, Data Management, Distributed Computing.**

I. INTRODUCTION

Big Data is a term refers to Structured, unstructured and Semi structured data that is this data is having variety. Big Data is also referred a term as a data is a huge data set having really huge magnitude [1] i.e. volume (really a huge volume). Big data is that term which arrives before you and your organization has had to deal with before i.e. big data have velocity [4], [7]. This flood of data is generated by connected devices from PCs and smart phones to sensors such as RFID readers and traffic cams, In health care, for instance, clinical data can now come in the form of images (e.g. from X-rays, CT-scan, and ultrasound) and videos [4]. Imaging data collected from one patient alone can easily consume several Gigabytes of storage space. Cloud is term referred as using internet as a backbone for utilization of services on remote servers to store manage and process data rather than local servers or personal computers. Cloud computing basically developed as a heterogeneous service environment for providing computing facility to end users and now a day emerging service as IOT (Internet of Things).

The real big data gets its hidden 'V' besides volume, velocity and variety when that huge/big data get analyzed for discovered patterns [3], derived meaning, indicators for decisions, and ultimately the ability to respond to the world with greater intelligence [16]. A term "Big data analytics", is a set of advanced technologies designed to work with large volumes of heterogeneous data [4]. Resent study shows that cloud become a prominent technology of migrating all applications and services. Researchers thinking for all sectors data should be migrated on cloud for fast decision processing. This ultimately tends to the cloud computing as IAS (Internet as Service). Huge data set will be extracted and made possible decision on basis of knowledge in big data. Extractions required smart scalable analytics services, programming tools, and applications. Big data analytics uses complex data mining algorithms that required efficient high performance processors. Cloud computing infrastructure is able to provide both computational and data processing applications [2], [11]. This paper is organized in chapters, second chapter is for review of cloud computing and big data analytics method and current scenario, third chapter focus on the prominent methods which will be useful for such kind of data analysis, and fourth chapter is analysis and discussing on what are the current challenges in transforming big data analysis in cloud. Finally last we have focus on the future trends and conclude the statement how we can transfer big data analytics in cloud.

Scalable data management is a vision and future for next generation computing. From last decade most of the research has focused on large-scale data management and migration of that data in cloud from traditional enterprises. Cloud computing will provide this data for future decision. Cloud computing infrastructure and operations will have its own set of novel challenges; one of the most research-oriented topics is security [5]. In this article, we will primarily focus on the challenges and opportunities for transforming big data into cloud. To understand the fact we have divided the literature as section II will be focuses on the challenges in front of transformation of big data analytics in cloud and also focuses on the advantages of migration. Section III is about the techniques available in for migration. Section IV is focusing on the analysis and brief discussion about reality of migrating big data into cloud. At last Section V is conclusion for the work migration of big data analytics in cloud computing.

II. CHALLENGES AND OPPORTUNITIES

Cloud computing is trends to most efficient and prominent platform for service oriented computing in last two decade. This ultimately transformed cloud computing in revolutionary infrastructure refinement, the most popular infrastructure is a Platform as a service (PaaS) and Software as Service (SaaS) [13], [17]. Finally, with this refinement of paradigm one another paradigm is Infrastructure as Service (IaaS) led down cloud computing to concept where servicers will offer as Elasticity [18], pay-per-user, low affordable investment, low upfront investment, low time to market, and transfer of risks are some of the major enabling features that make cloud computing a ubiquitous paradigm for deploying novel applications which were not economically feasible in a traditional enterprise infrastructure settings [5]. Whereas another trend which leads the next generation computing is progress in business intelligence and analytics of data. Field of big data analytics emergence from these trends where opportunities for BI (Business intelligence) software arise [6]. Big data researchers classify system as one for supporting update heavy applications, and another is for ad-hoc analytics and decision support. Scalable and distributed data management is a visionary paradigm in transforming big data analytics into cloud .initially researchers developed a distributed database [7], [5] and for maintaining the workload parallel database system [8] both were successful. Change in data access is major issue in front of distributed and parallel data, to resolve that a new class of system definitions i.e. Key Value. MapReduce paradigm [9] and its open source implementation platform Hadoop [10] is basically is a solution of this problem of distributed and parallel data bases. The next step is to develop application which works on cloud and manages these big data performed analytics for business intelligence this opens a new possibility. Cloud itself has some features which will support for development of such a complex system which can access data from many sources. These cloud features include scalability, elasticity, fault-tolerance, self-manageability, and ability to run on commodity hardware [5]. Utilization of the features for development of application need a system which can update heavy workloads as internet generates huge amount of data. Here a we focus on discussion on new generation Key-Value data store which has extreme success and adopted by industry i.e. Hadoop MapReduce in next section. Here we can say that their are some challenges in big data migration in cloud namely (i) Scalable Data Management (ii) Data Management for Large Applications (iii) Large Multitenant Databases (iv) security issues for cloud computing, Big data, Map Reduce and Hadoop environment. In next section we are focusing on how these problem get resolve with the current edge technology namely Hadoop and MapReduce [19], [14].

III. TECHNOLOGICAL ASPECT

Complex data analysis and identifying patterns on cloud-based environment is explained with Figure 1. Basically here we have many sources of the data generation. This data is basically massive and progressive. Every data centers have to supply this data in distributed environment and here load balancing is crucial issue. These scalable database management systems have to supply massive data to number of application. Load balancing and security is major issue in big data. From fig we can simply analyzed that these data generation from different sites in massive amount is progressive and has to distribute over cloud environment. Mapping this data to primary sources to their required destination required Hadoop like framework and MapReduce like technology.

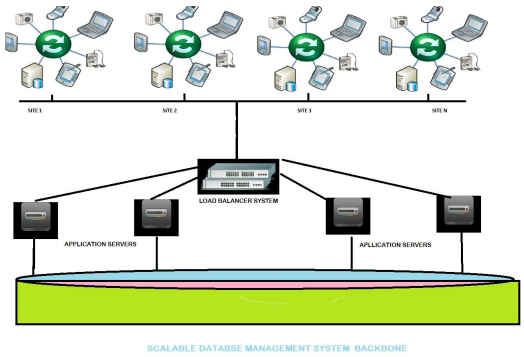

Figure 1. Scalable data sources generating massive data in distributed environment.

Hadoop – Hadoop is open source framework and has two components that are HDFS and MapReduce. HDFS is distributed file system for storing and retrieving data for MapReduce and help to executes jobs for users. Hadoopp form the cluster of data nodes and store data on space utilization of data nodes on cluster [12]. Hadoop runs on heterogeneous environment and may leads to workload problem while data distribution and access. Above the HDFS layer, there is MapReduce engine which consist on operating part for the server. one problem with Hadoop is workload balancing will be handled by transfer of data between racks. Modifying data transfer rate between the racks is another techniques. In figure 2 single node request propagation is shown.

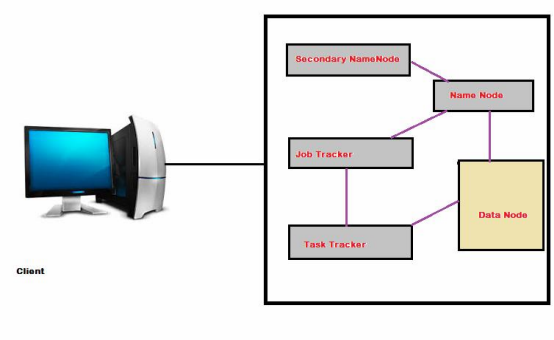

Figure 2. Single Node Representation of Hadoop

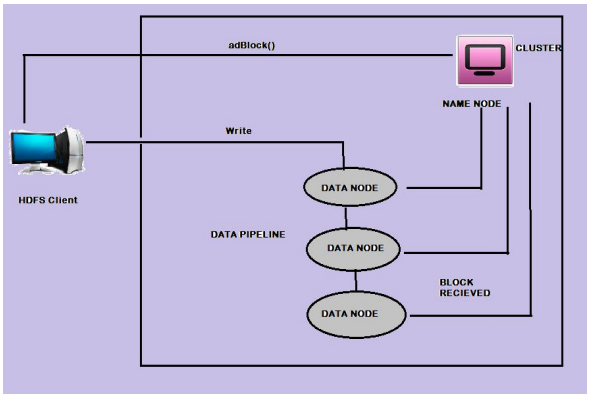

Figure 3. HDFS Representation in Hadoop

MapReduce – MapReduce works with HDFS with functions namely Map and Reduce. Whenever job is assign by the users to Hadoop , immediately input spited in multiple pieces and Map function will be applied to data for generating intermediate result. Intermediate result will be monitor and shuffled for generating final result. Whenever a job tracker gets job from client it will first execute Map task first and then finds processes for every split data [15], [17].

Figure 3 is pictorial representation of HDFS and figure 4 is pictorial representation of MapReduce.

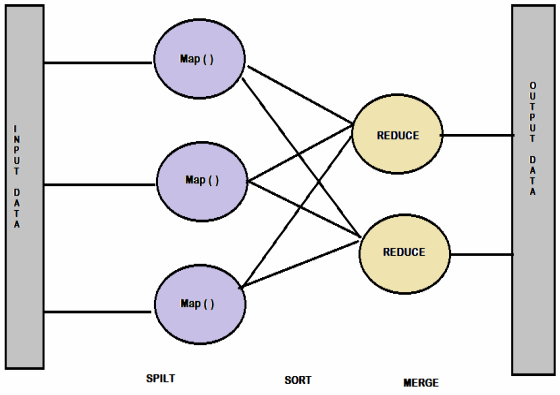

Figure 4. MapReduce in Hadoop

Map/Reduce in distributed is currently using by Google [9]. MapReduce is a functional programming, the main function Map works in collaboration with Reduce. This Mapreduce have advantages reasonable for mapping and aggregation operations. It's Map/reduce through massive data set we can divide into small pieces and distribute that data on different nodes it also manages the load balancing in large data set.

IV. DISCUSSION

In this article we have discussed different issues and challenges of migration big data analytics into cloud environment. We can analyzed that there are majorly few challenges as cloud have its own feature we can use that feature and can build a realistic application for migration of big data analytics. Before development of such application some of the open problem must be address to ensure success of system. Key Value store popularity is advantages result for Hadoop based HDFS and MapReduce Function. We can build a system by using Hadoop and MapReduce technology which will be state of art in scalable data management for heavy workloads. Second designing the next generation cloud application and big data analytics we must have to focus on security. Third one is Improving Database Diversity i.e. providing Data Storage as a Service (DSS).

V. CONCLUSION

A single perfect different system target application which works on distributed heterogeneous environment yet to be developed. For development of such application and cloud as a service provider for data storage and making analytics from diverse source of data we have some challenges. Hadoop MapReduce is likely to be a prominent solution with advancement of some cloud features. Overall security will be the major issue left foe transforming data in cloud and making analytics. With advancement of migration algorithm and Open source techniques next generation computing will be definitely transform to cloud.

REFERENCES

[1] A. Abouzeid, K. B. Pawlikowski, D. J. Abadi, A. Rasin, and A. Silberschatz. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. *PVLDB*, 2(1):922–933, 2009.

[2] D. Agrawal, S. Das, and A. E. Abbadi. Big data and cloud computing: New wine or just new bottles? *PVLDB*, 3(2):1647–1648.

[3] D. Agrawal, A. El Abbadi, S. Antony, and S. Das. Data Management Challenges in Cloud Computing Infrastructures. In *DNIS*, pages 1–10, 2010.

[4] White paper on "Solution Brief Big Data in the Cloud: Converging Technologies , How to Create Competitive Advantage Using Cloud-Based Big Data Analytics.

[5] Divyakant Agrawal ,Sudipto Das and Amr El Abbadi "Big Data and Cloud Computing: Current State and Future Opportunities", EDBT 2011 , March 22–24, 2011, Uppsala, Sweden ACM 978-1-4503-0528-0/11/0003.

[6] Hsinchun Chen, Roger H. L. Chiang et. al. "Business intelligence and analytics :from big data to big impact" MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012 pp-1165-1189

[7] J. B. Rothnie Jr., P. A. Bernstein, S. Fox, N. Goodman, M. Hammer,T. A. Landers, C. L. Reeve, D. W. Shipman, and E. Wong.Introduction to a System for Distributed Databases (SDD-1) ACM Trans. Database Syst. , 5(1):1–17, 1980

[8] D. J. Dewitt, S. Ghandeharizadeh, D. A. Schneider, A. Bricker, H. I. Hsiao, and R. Rasmussen. The Gamma Database Machine Project. IEEE Trans. on Knowl. and Data Eng., 2(1):44–62, 1990.

[9] Hadoop Distributed File System: Architecture and Design http://hadoop.apache.org/common/docs/r0.18.2

[10] Venkata Narasimha Inukollu et. al. " Security Issues Associated With Big Data Incloud Computing", International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014 pp- 45-56.

[11] Tharam Dillon et. al. "Cloud Computing: Issues and Challenges", 2010 24th IEEE International Conference on Advanced Information Networking and Applications, 2010 IEEE DOI 10.1109/AINA.2010.187, pp- 27-33.

[12] AiLing Duan et. al. "Research and Practice of Distributed Parallel Search Algorithm on Hadoop_MapReduce", 2012 International Conference on Control Engineering and Communication Technology, 2012 IEEE DOI 10.1109/ICCECT.2012.131, pp-105-108.

[13] Amrit Pal, Pinki Agrawal et. al. "A Performance Analysis of MapReduce Task with Large Number of Files Datasetin Big Data Using Hadoop", 2014 Fourth International Conference on Communication Systems and Network Technologies, 2014 IEEE DOI 10.1109/CSNT.2014.124, pp- 587-591.

[14] Xiaofei Hou, Ashwin Kumar T K et. al. "Dynamic Workload Balancing for Hadoop MapReduce", 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, 2014 IEEE DOI 10.1109/BDCloud.2014.103 pp- 56-62.

[15] A reason to take Big Data seriously - http://miipharos.com/big-data-analysis-ibeacon/

[16] Big Data Analytics - Academia.edu- *w.academia.edu.*

[17] Cloud Computing Portability and Interoperability : Cloud Portability and Interoperability http://www.opengroup.org/cloud/cloud/cloud_iop/cloud_port.htm

[18] R.Saranya, V.P.MuthuKumar "Security issues associatedwith big data in cloud computing ",International Journal of Multidisciplinary Research and Development, Volume :2, Issue :4, April 2015 ISSN: 2349-4182 pp- 580-585

[19] Dominique A. Heger "Big Data Analytics Where to go from Here- http://www.datanubes.com/mediac/BigDataFutureDHT.pdf.