ISSN No. 0976-5697

Volume 5, No. 5, May-June 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Proposing PCSSA-using CBBT for Managing Big Data

Rimmy Chuchra
Department of Computer Science and Engineering
Sri Sai College of Engineering and Technology, Manawala,
Amritsar (India)

R. K. Seth
Dean, Department of Applied Sciences
Sri Sai University, Palampur,
Himachal Pradesh (India)

Abstract: This paper reviews the analysis of big data in cloud computing and data mining and proposes a new architecture design to reduce cost and store data locally with storage node clusters using Cloud Bluetooth Based Technology (CBBT) during sharing of cloud services in which several devices share data information through personal area network with the help of Piconets and Scatternet in order to create multi-user network environment. Any private organization prefers to store their data on personal private cloud storage architecture. Facing the problem of exact storage data location, the utilization of proposed architecture would enable to store data at local level and the proposed architecture is termed as "Private Cloud Storage System Architecture (PCSSA) using Cloud Bluetooth Based Technology (CBBT). The architecture may help to share a cloud service by organization at local level without the interference of cloud service provider by using CBBT and can be accessed more conveniently similar to accessing the data from the hard disk of computer.

Keywords: Big data, cloud computing, data mining, storage, cluster.

I. INTRODUCTION

Big data and cloud computing are two descriptive and emerging technology based trends at present that offers a large number of business opportunities and also posing considerable challenges on the current IT industry and are potential subjects for researchers and scientists around the globe to figure out the problems and their solutions [1, 2]. Big data is not a single technology or initiative but it is a trend across many areas of business. In big data technology, the major concerns are to scale and keep the data at appropriate location; modeling the structured, semi-structured and unstructured data and the integration of collected data from different sources such as spatial and graphical data [25]. The significance of big data lies in its general characteristics such as scalability, elasticity, adhoc resource pooling and low cost set up infrastructure [30]

Big data describes large and complex data sets even they are impractical to manage with their traditional software tools. Online mode, offline mode and real time mode are the three main sources of big data that helps for its generation [23]. The real time big data requires a new breed of data capture and analytics infrastructures that can handle its large volume and high velocity [29]. Any data set must support 5V's; velocity, variety, volume, veracity and variability for maintaining effective and efficient performance in big data affairs [3].

Figure 1: Basic supporting parameters for any Data Set- 5V'S [3].

In order to manage big data, it may be categorized in two ways depending upon its operational and analytical behavior so that big data loads can be easily manageable with the support of 5V's cost effectively while dealing with complex queries without SQL and with SQL depending upon the selection of the category.

The implementation of maintaining and movement of big data may be categorized into three major requirements viz., data modeling services; data migration, transformation and upload service; big data infrastructure configuration and deployment [31]. The Table 1 shows quantity of data handled by some social networking sites in terms of amount of data storage and number of users accessing the data. The estimated quantity of data storage and number of users has been taken from different websites.

Table 1: Estimated Quantity of Big Data World.

Big Data World(BDW)	*Data Storage Amount*	*No. of Users Access BDW*
Google	>100 Peta Bytes	>IT Indexed URL'S.
Facebook	>100 Peta Bytes	>800 Million Users.
YouTube	>=750 Peta Bytes	>20,000 4TB Drivers.
Twitter	>55 B/ Year	>=150 Millions users per day.
Global Text Message	>6.1 TB/Year	>876 users per year.

II. CHALLENGES TOWARDS BIG DATA IN CLOUD COMPUTING AND DATA MINING

In the field of cloud computing and data min ing, Big Data faces following challenges:

a. For improving scalability, there is need to design new architectures that support massive parallel processing architectures for big data companies like Google, yahoo etc [20].
b. Most of the researchers faced problem to deal with multi-network and therefore there is to design new models that support multi-network [20].
c. Designing new algorithms that help to detect hidden patters from big data become too critical.
d. Developing new software's that provides automatic detection of data sets.
e. Proposing new indexing schemes or multi-dimensional index structures to improve the speed [20].

f. Most challenging problem is to draw new parallel computing models besides Map reducing.
g. Proposing a new algorithm for "GARBAGE MINING" that actually helps to mine garbage (that has no value and critical to identify these files) and overcome the limitations of existing data mining techniques [20].
h. Improving data Provenance this integral feature deals with big data and helps to relate evolution history or the origin of that data item was extracted or collected from [20].
i. Google and FaceBook pay for next generation data centre and brings challenges to IT organizations including pressure from corner office to replicate the speed, agility, efficiency etc. [26].

As an example, the following screen shots are provided to show data redundancy in which data has been duplicated and becomes the part of a problem in big data.

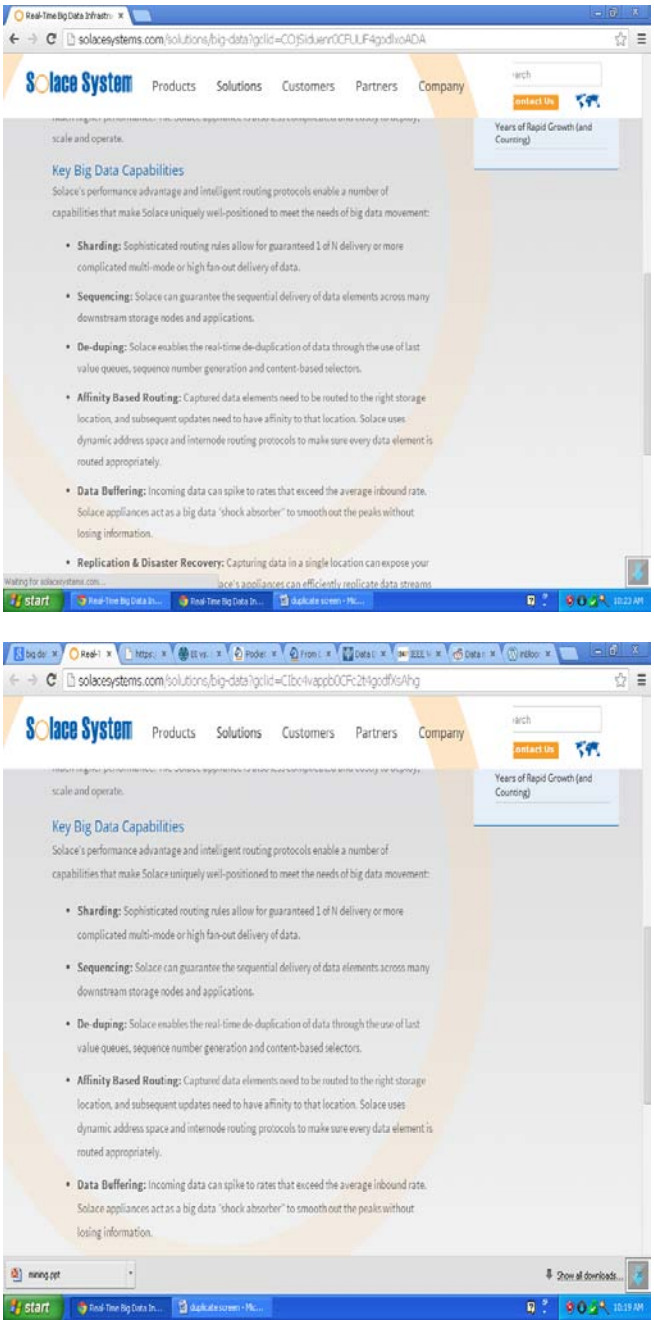

Figure 2: Screenshots Shows Data redundancy at different URL'S.

A. Review of Reducing and Controlling Big Data:

Big data may be reduced by utilizing a variety of scaling algorithms. The alert analytics by data analyst may be provided and "NO SQL Technology" may be used to solve common database management group problems [6]. Reducing the cost of mining operations may helps to improve the efficiency and cost effectiveness [7].

The control over big data is another factor for enhancing the efficiency and may be achieved by designing new mechanisms and policies for data movement along with controlling unstructured and semi-structures data in addition to improving the capacity of cloud storage architecture. The major concerns have been figured out comprehensively about reducing and controlling big data and are described as follows:

a. For controlling big data, there is need to design new data centers that helps to consume less energy by its components (e.g. computing, storage, networking) and also helps to save bandwidth and become more energy efficient.
b. By designing more scalable methods for combinatorial data mining on case control data by including maximum number of entries [9].
c. For controlling big data, the important factors are to handle the processing demands that are posted by the big data. To handle such type of processing demands most supportive technologies are to be used like NO SQL databases, Hadoop & map reduce etc. [10].
d. Use distributed computing rather than centralized computing so that the chances of single hub of failure decrease [11].
e. Focus on much deeper analysis and better utilization of computing.
f. With the help of SKYROCKET, a large amount of data generated by the internet can be easily controlled and that required advanced analysis also [12].
g. Use HADOOP to manage the clusters of computers that actually use potential privacy "gotchas" in cloud computing [13].
h. Use multi-step analysis instead of single step of analysis for taking most accurate decision and easily find out separate clusters. It helps to increase the query optimization. The main objective is to enable to specify a large number of potential analyses that basically helps to select carefully parameterized choices. It facilitates to reduce or minimize human burden in the iterative process of data exploration [14].
i. Use streaming of operational intelligence from logs [15].
j. Helps to enhance network security in indirect manner [16].
k. Try to work with cloud resources that handle structured and un-structured data rather than work with existing data warehousing technology which is too expensive. The extra benefit in this case is to provide scalability at low storage cost associated with building data hub on the cloud [17].
l. Replace physical server with virtual servers on a single physical machine and these virtual servers may be dynamically created and destroyed like MANET (Mobile Adhoc Network).

m. Need to focus on the relation between performance, power and energy efficient parameters and try with different versions of Hadoop [18].
n. Use HDFS (Hadoop Distributed File System) rather than NTFS (New Technology File System) using proper set of rules. By using different protocols that may help to increase in performance without increasing the power consumption [18].
o. Need to design a new layer for MAC (Media access control) when providing networking [19].
p. By improving the storage processing large datasets, there is need to create a control communication granularity [19].
q. Use of robots for transfer of data on the cloud rather than humans which would also helps to enhance the daily operation capability for making accurate decisions [21].

III. OVERVIEW OF ROAD MAP OF PROPOSED CONCEPT

In order to meet challenges such as storage, capture, sharing, search, transfer, analysis and virtualization [24] and also to improve the effectiveness and lower the cost of existing application, the aim is to enhance productivity and making the data availability faster by allowing users to access systems and files remotely via cloud or otherwise [28].

The clustering such as millions of data sets that are dominated by particular collective operations which are common in MPI (Multi-Programming Interfaces) [5] along with parallel and elastic computing are among the common problems in the big data. The problem of path may be solved by using various tools like machine learning, Bayesian, clustering etc [4]. For solving the problems related to big data, the following steps has been proposed by the authors to enable the proposed procedure utilizing the Private Cloud Storage System Architecture (PCSSA) using Cloud Bluetooth Based Technology (CBBT). Fig. 3 shows the overview of road map of research concept in order to develop architecture for solving the big data problem.

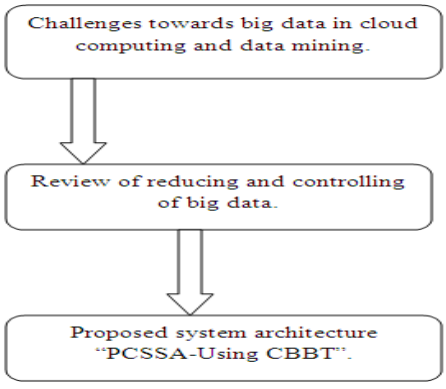

Figure 3: Overview of research enabling to develop system architecture.

IV. PROPOSED SYSTEM ARCHITECTURE "PCSSA USING CBBT"

The proposed architecture design using "PCSSA using CBBT" is shown in Fig. 4.

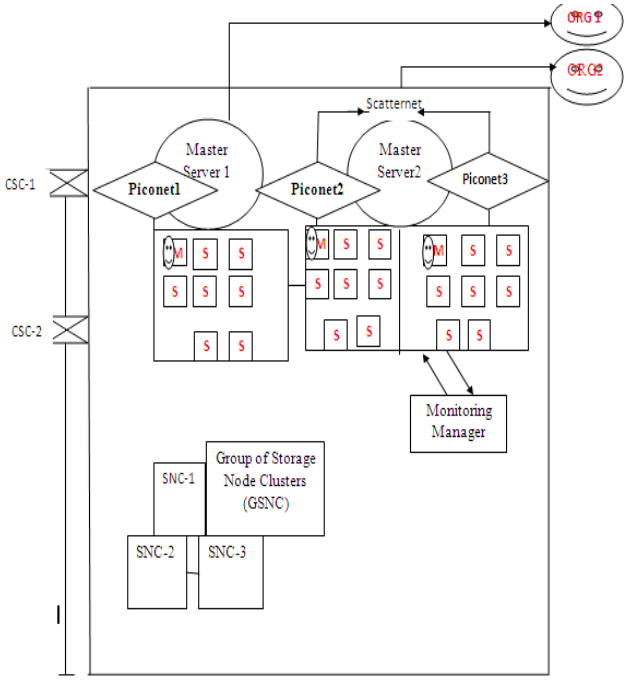

Figure 4: Proposed Architecture: "PCSSA-Using CBBT"

The description of the configuration of the architecture design shows that there are two master servers named as MS-1 and MS-2 comprising "Piconet" and "Scatternet". Further, Piconet consists of eight devices that include one master device and seven slave devices. A maximum of seven slave devices could be attached with one piconet. Scatternet comprises Piconets. As an example one Scatternet would consist of two mater devices and fourteen slave devices. CSC is abbreviated for Cloud Service Client and is anode that helps to provide network connectivity. GSNC is A Group of Storage Node Clusters and helps to store data on nodes in the form of clusters. SNC (Storage Node Cluster) is a place where data is actually stored. All the tasks performed under the system are supervised by Monitoring Managers.
Cloud storage system aims to provide users with transparent, efficient and local disk like storage space where users can save their data and use local storage space [22]. "PCSSA-Using CBBT" is a new proposed cloud storage system architecture that is described as Private Cloud Storage System Architecture using Cloud Bluetooth Based Technology. The objective of this proposed architecture is to reduce the data access time, increase performance, to reduce cost as well as power.
Fig. 5 shows the architectural workflow communication in which two organizations are taken as an example to make the concept of architecture more descriptive.

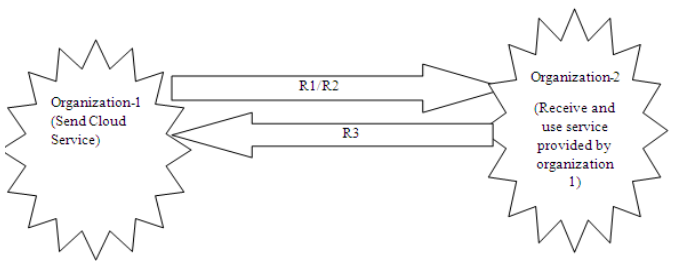

Figure 5: Communication flow for accessing cloud service used by private organizations without interference of cloud service provider.

Table 2 describes the abbreviations used in Fig 5 for depicting the communication flow between the two organizations without the interference of cloud service provider.

Table 2: Abbreviations used in proposed Architecture.

R1	Request for the use of already purchased cloud service from organization 2.
R2	Request for the data access stored in cloud system storage architecture from organization 2.
R3	Reply.

Cloud data stored in a number of data centers can be made available by the cloud service providers. There are number of cloud storage system architectures available in particular data centre. The design goal of any data center is to connect users to applications using three core infrastructure components that meet specific requirements such as compute, storage and networking [27].

A. Working:

When organization 1 sends request to organization 2 for accessing already purchased cloud service and if during that time slot, the server-1 is free, and then request will be handled by its master during receiving request. It would first figure out the type of request and operations to be applied on it and after that forwards the requests to its own slaves to process requested operation.

Piconet has maximum of seven slaves and suppose in case user wants to add more than seven devices under a single piconet, then request pending status will be enabled and then it will automatically shifted to server 2 that consists of Scatternet comprising two Piconets and would follow the same procedure as discussed earlier. In case all the devices of both the servers are busy and cannot process the requests, then the user must have to wait until green signal comes to step ahead.

V. CONCLUSION

Due to exponential increase in growth of data, the problem of managing Big Data has been reviewed and analyzed. A variety of methodologies utilizing new electronic based technologies is being used to meet the challenges of reducing and controlling big data. Being a critical issue for the researchers, a more convenient way for managing big data involving a new architecture "PCSSA-Using CBBT" has been proposed that may help to deal with the problems associated with big data more efficiently. The proposed architectural design would resolve the problem of users' exact locations of data storage. The technique may help to share cloud service among organizations without the interference of cloud service provider.

VI. REFERENCES

[1]. Borkar V,Carey MJ,Li C.Inside "Big Data Management":ogres,onions or parfaits?proceedings of the 15th International Conference on Extending Database Technology(EDBT'12),2012:3-14.

[2]. Chaudhuri S.what Next? A half-Dozen Data Management Research Goals for Big Data and cloud the cloud.Proceedings of the 31st Symposium on principles of Database systems (PODS' 12), 2012:1-4.

[3]. Ec.europa.eu/digital-agenda/futurium/en/content/big-data.

[4]. https://www.google.co.in/#q=what+is+the+common+problem+of+big+data+in+cloud+computing+and+data+mining&start=20

[5]. https://www.google.co.in/#q=what+is+the+common+problems+big+data+problems+in+cloud+computing+and+data+mining&start=0

[6]. https://www.google.co.in/#q=what+is+the+common+problem+of+big+data+in+cloud+computing+and+data+mining&start=7.

[7]. www.bigdatamining.com.

[8]. https://www.google.co.in/#q=what+are+the+common+big+data+problems+m+cloud+computing+and+data+mining&start=20.

[9]. https://www.google.co.in/#q=how+to+control+big+data+problem+in+data+mining+and+cloud+computing.

[10]. searchbusinessanalytics.techtarget.com/definition/big-data-analytics.

[11]. https://www.google.co.in/#q=how+to+control+big+data+problem+in+data+mining+cloud+computing=10.

[12]. https://www.google.co.in/#q=how+to+control+big+data+problem+in+data+mining+and+cloud+computing&start=20.

[13]. https://www.google.co.in/#q=how+to+control+big+data+problem+in+data+mining+and+cloud+computing&start=60.

[14]. pages.cs.wisc.edu/~raghu.research.html.

[15]. https://www.google.co.in/#q=how+to+control+big+data+problem+in+data+mining+and+cloud+computing&start=80.

[16]. www.windowssecurity.Com/articles-tutorials/misc_tutorials/misc_network_Security/big-data-security-perespective-part1.html.

[17]. www.aditi.Com/Solutions-on-cloud/big-data/big-data-cloud-computing/.

[18]. https://www.cs.ualberta.Ca/news-events/event-Calender/2014/big-data-and-cloud-computing-seminar.

[19]. https://www.google.co.in/#q=how+we+control+the+problem+of+big+data+in+cloud+computing+and+data+mining&start=100.

[20]. www2.cs.slu.edu/~dche/Publications/from%20Big%20Data%20To%20Big%20Data%20Mining.pdf.

[21]. www.isn.ethz.ch/Digital-Library/Articles/Detail/?Ing=en&id=173004.

[22]. file:///C:/Documents%20and%20Settings/Administrator/My%20Documents/Downloads/cloud%20storae%20architctur%20paper%202013.pdf.

[23]. www.mongodb.com/learn/big-data.

[24]. en.wikipedia.org/wiki/Big-data.

[25]. www.calit2.uci.edu/uploads/Media/Text/BigAndCloudy-Carey.Pdf.

[26]. utsa.edu/today/2013/11/Clouddata.html.

[27]. Wikibon: org/blog/the-future-of-data-center/.

[28]. www.federaltimes.Com/article/2030709/DEPARTMENTS/307090011/Cloud-seen-answer-big-data-needs

[29]. solacesystems.Com/Solutions/bigdata?gclid=COjSiduenrOCFUUF4godLXoADA.

[30]. blog.Sqlauthority.Com/2013/10/15/big-data-role-of-cloud-computing-in-big-data-day-11-of-21/.

[31]. www.theserverlabs.Com (big-data-hpc.html.