I J C T A, 9(2) 2016, pp. 435-444

Big Data with Cloud Virtualization for Effective Resource Handling

K. Venkatachalapathy[1], V. S. Thiyagarajan[2], A. Ayyasamy[3] and K. Ranjani[4]

ABSTRACT

Big data is a challenging task because of its characteristics and presence of data in large amount. Cloud computing deals with handling the remote server with full of integrity, integration of cloud & Big Data is the most challenging task to handle in existing systems. Pattern matching algorithm is used here in order to avoid duplicates during the creation of dynamic dataset. A dynamic dataset is created each and every time when data is uploaded. In this approach a system is defined, which partitions the dataset using balanced partitioning method. This application is developed for medical database and also achieves virtualization in our local machine for effective resource handling. Interfacing unit is included which acts as a bridge between server and virtual clients. Virtual Machines (VMs) can be added or removed as per the request. Aggregation method is used to integrate the partitioned data. The outcome of cloud and BigData is achieved by means of virtual environment scenario.

Keywords: Integrity; Virtualization; partitioning; Big Data; Aggregation.

INTRODUCTION

Cloud computing assembles large networks of virtualized Information and Communication Technology (ICT) services such as hardware resources (such as CPU, storage, and network), software resources (such as databases, application servers, and web servers) and other applications. In industry these services are referred to Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS)[16]. Mainstream ICT powerhouses such as Amazon, HP, and IBM are heavily investing in the provision and support of public cloud infrastructure [15]. Cloud computing spreads rapidly and become a popular infrastructure among all types of organizations. Despite of some initial security concerns and technical issues, an increasing number of organizations have moved their applications and their services in to "The Cloud". These applications range from generic word processing software to online healthcare. The cloud system hits into the processing power of virtualized computers on the back end, thus significantly speeding up the application for the user, which just pays for the used services.

Big data is a famous term used to describe the exponential growth and availability of data in both structured as well as unstructured form. Big data may be significant to business and society as the Internet has become. Big Data is so large that it's difficult to process using traditional database and software techniques [2]. Big Data applications have become a common phenomenon in domain of engineering, science and commerce. Some of the representative applications include disaster management, high energy physics, genomics, automobile simulations, medical imaging, and the like. The "Big Data" problem is defined as the process of collecting and analyzing complex data sets so large that it becomes difficult to analyze and understand manually or by using on-hand data management applications. (e.g., Microsoft Excel). Hadoop can process large volumes of data, assuming the source data files are in a readable and ready state [20].

For example, consider disaster management in case of Big Data application, there is a necessity to analyze "an overflow of online data from multiple sources (feeds from social media and mobile devices)"

[1, 2, 3, 4] Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Chidambaram, Tamilnadu, India, *E-mail: omsumeetha@rediffmail.com*

for understanding and handling real-life events such as flooding, earthquake, etc. Over 20 million twitters posted during Hurricane Sandy (2012) lead to an illustration of the Big Data problem. The statistics provided by the Pear Analytics study expose that almost 44 percent of the Twitter posts are spam and useless, about 6 percent are personal or product promoting, while 3.6 percent are news and 37.6 percent are conversational posts.

During the 2010 Haiti earthquake, text messaging via cell phones and Twitter made headlines as being vital for disaster response, but only some 100,000 messages were actually processed by government agencies due to lack of automated and scalable ICT (cloud) infrastructure. Large-scale, heterogeneous, and uncertain Big Data applications are becoming increasingly common, yet current cloud resource provisioning methods do not scale well and nor do they perform well under highly changeable conditions (data volume, data variety, data arrival rate, etc.). Much research effort have been paid in the fundamental understanding, technologies, and concepts related to autonomic provisioning of cloud resources for Big Data applications, to make cloud-hosted Big Data applications operate more effectively, with reduced financial and environmental costs, reduced under-utilization of resources, and better performance at times of unpredictable workload. Targeting the above mentioned research challenges, this special issue compiles recent advances in autonomic provisioning of Big Data Applications on Clouds.

Pattern matching algorithm is used here in order to avoid duplicates during the creation of dynamic dataset. A dynamic dataset is created each and every time when data is uploaded. There are two operations are performed to handle datasets. 1) Dynamic and 2) Static, which is taken from web [22, 23, 24]. In this approach a system is defined, which partitions the dataset using balanced partitioning method [1, 9] and produces virtualization, which is implemented for effective resource handling. Interfacing unit is included which acts as a bridge between server and virtual clients. VMs can be added or removed as per the request. This process assures QOS for Clients, which is the modification of this proposed implementation. Two Clouds Servers for handling different Jobs in Medical.

This Application is developed for Medical and also achieves Virtualization in our Local Machine. There will be minimum one VM and maximum 4 VMs are assigned per Server. Aggregation method is used to combine the partitioned data. The outcome of cloud and Big Data is achieved by means of virtual environment scenario.

LITERATURE SURVEY

Literature presents various algorithms for effectively handling resource in Big Data with Cloud environment. Here, we review some of the works presented for that. ArtiMohan purkar *et al.*,[1] described Balanced Partition technique which provides better performance with the help of PIG and creates a histogram for the respective partition. Arash Baratloo *et al.*,[4] present a system which allows application programmers to write parallel programs in Java language and allows Java-capable browsers to perform parallel tasks. It comprises a virtual machine model which isolates the program from executable environment. Vijay Thayananthan *et al.*,[6] quantum cryptography provides maximum protection with less complexity that increases the storage capacity and security strength of the big data. In this section, we need to remember the use of symmetric key with a block cipher which is suitable to control the big data security because the design of the block cipher for the big data is very simple. Guy E. Blelloch.,[8] describes that load balancing is important to be noticed because the data element may drop out during the execution of an algorithm. If the element drops out then it also makes other elements remains unbalanced. Craig C. Douglas.,[10] describes a dynamic big-data-driven application system (DBDDAS) toolkit must have in order to provide all of the essential building blocks that are necessary to easily create new DDDAS without re-inventing the building blocks. Luiz Andre Barroso *et al.*,[11] describes different queries run on different processors and, by partitioning the overall index, also lets a single query use multiple processors. to handle this workload, google's

architecture features clusters of more than 15,000 commodity class pcs with fault-tolerant software. this architecture achieves superior performance at a fraction of the cost of a system built from fewer, but more expensive, high-end servers. CISCO Corporation Inc.,[16] describes the Cloud VPN(Virtual Private Network) which is accessible, backward compatible, and based on open standards. Data centers will finally be capable to optimize their resources in the cloud and offer new services: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). Remzi H. Arpaci-Dusseau *et al.*,[17] describes a data-flow programming environment and I/O substrate for clusters of computers. River is designed to provide maximum performance in the common case even in the face of non-uniformities in hardware, software, and workload. Focus Group on Cloud Computing.,[19] describes services in network (L2-L3 connectivity, and VPN and L4-L7 network services) like smart pipes "high-grade networks" for cloud transport service and interconnection of cloud (inter-cloud) in order to assurance a secure and high performance end-to-end Quality of service (QoS) for end users. Jeffrey Dean *et al.*,[21] describes the run-time system which takes care of partitioning the input data, scheduling the execution of program's across a set of machines, handling failures of machine, and managing t required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. Lizhe Wang *et al.*,[12] Commercial and public data centers offer storage, computing and software resources as cloud services, which are enabled by virtualized software/middleware stacks. Private data centers normally build basic infrastructure facilities by combining available software tools and services. They are enabled for resource sharing with grid computing middleware. This software includes cluster management system, resource management and data management system. LiXiong *et al.*,[13] describes that, In order to monitor the stream of data from the data contributors and to assurance differential privacy for the input data streams, a simple yet infeasible approach can be accomplished by issuing an aggregate query for each grouping of the attribute values. Siddaraju *et al.*,[14] describes big data processing can be achieved through a program design paradigm known as Map Reduce. Typical, implementation of the Map Reduce paradigm requires network attached storage and parallel processing. The dataset used for this project is included in the reference [22, 23, 24]. Fig. 1 represents the overall architecture of this project.

SYSTEM OVERVIEW

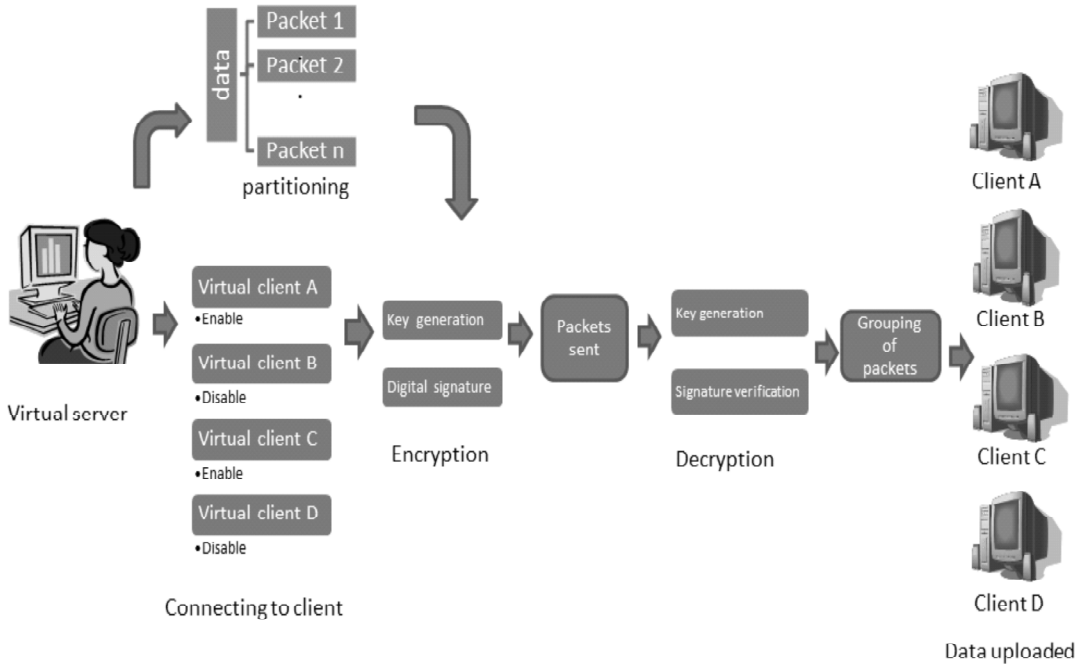

Figure1: Architecture diagram

First, the virtual server chooses the dataset which is needed to be partitioned and send to client. Then, depends upon the buffer value chosen by the user the dataset will be partitioned and created as packets. Buffer value represents number of bytes to create a single packet.

The server then chooses the clients to which the data packets need to be transfer. The server and clients are connected by means of interfacing unit. This interfacing unit connect server to many clients and make transfer of data to more than one client simultaneously. After the connection established, encryption can be achieved by key generation and digital signature. Then packets are transferred. When the packets are received at client end, automatic verification of signature and decryption can be done. Received packets are grouped and viewed by the client in original format.

METHODOLOGY

A. Pattern Matching Algorithm

One of the most actively researched areas of computer science is Pattern matching with numerous papers still being published. This type of category of matching problems includes following:

- String matching exactly: - for example, in a saying handling issue.
- String matching appropriately:- for example, in penmanship identification and optical chaste reorganization.
- Largest common substring
- Retrieval of information and querying
- Spam filtering and plagiarism discovery
- Processing the signal.

Execution of any matching algorithm is judged, as various string similarity measures have been produced that can come close to compare strings and focus a quantitative measure of the level of similarity. The Exact pattern matching algorithm is implemented on the basis of searching and matching contain with other data that is present at that time in the system. Patterns are also helpful for finding the impurity or intruders in the system.

Algorithm

Get values for m and n, the size of text and the pattern

Get values for both the text t1,t2,…tn and the pattern p1,p2,….pn

Set K, the starting location of attempted match to 1

While (k<= (n-m+1)) do

    Set the value of i to1

    Set the value of mismatch to NO

    While both ((i<=m) and (mismatch=NO)) do

        If $P_i \neq T_{k+(i-1)}$ then

            Set mismatch to yes

        Else

            Increment i by 1(move to next character)

    End of the loop

    If mismatch=NO then

Print the message there is a match in position

Print the value of k

Increment k by 1

End of the loop

Stop we are finish

(B) Partition Algorithm

Hadoop is most popular for best performing the distributed computing, but it's simple partitioning methodology does not preserve correlation between data chunks. So there is need for partitioning Framework like FARQ in which partitioning helps for balancing data chunks into corresponding partitions. These partitions hold data for increasing processing speed. According to large data record field partitioning algorithm is splitting and analyzing that particular record. Also, it is assigned a record transfer data from large tables to small tables. For bursting query performance, the partitioning algorithm plays a vital role. The sampling of data is necessary for the analysis of big data because the data is present in large amount. Sampling has various methods, one of the most famous method is stratified sampling in which sampling takes place among independent groups and select only one sample for improvement and reduction of errors. This project is based on the idea of stratified sampling partitioning algorithm. This algorithm divides space values into different groups and subdivides groups into different portions according to server spaceavailable for particular partition. Partition algorithm is expressed for data set DS as Partition (Ds) = (G, pn) = (Vi, random [1, Vrange]) Where pn is number of a partition in group G, random function is a random number in [1; Vrange] , and Vi is a Group Identifier (GI) for the group G. For initial condition GI is equal to < 0;0; 0 > then length of group is [0; 1]. For GI is equal to <x; 0; 0 > then length of group is [2x ; 2x + 1]. For GI is equal to < x; y ; 0 >then length of group is [2x +y ; 2x +y +1]. For GI is equal to < x; y ; z > then length of group is [2x +y +z ; 2x +y +z + 1].

Algorithm steps

- **Input: Record(R), Vector Set VS**
- **Output: Partition identifier PI**
- Record has to parse into different column families.
- Compute Group Identifier (Gi) with value ranges as stated above. Get partition vector Vp from VS with Gi and set
- Vpi=<Gi ; Vrange> Set the target for Partition identifier,
- PI =<Gid; random [1; V pi * Vrange] ; Build sample in partitioning PI;
- count P I count P I + 1;
- sum PI sum PI + N;
- sample PI sum x; y ;z ;range =count PI ;
- RI Hash (PI; counter PI);
- Send Record to partition PI; return PI;

We use mean value of aggregation that generates samples, given as Sample = S UM =count, where S U M - sum of value from aggregation, and count- number of records in current partition. P I sent to partition is generated by input record R.

(C) Aggregation

The aggregation query is nothing but the aggregate functions used in the query like SQL, oracle, MySql and Sybase. There is Online Aggregate (OLA) that is used for improving the interactive behavior of database.

For effective operations on database, batch mode is performing a key role. The traditional way is that user query and waits till database come to an end of processing entire query. On contradict to OLA, the user gets estimated results side by side as query is fired. In 1997, ArtiMohan purkar[1] includes that Hellerstein proposed the OLA for group-by aggregation queries for just one table. Here aggregation is used to grouping the packets, when it reaches the client side.

(D) Interfacing Unit

It acts as a connecting bridge between virtual server and virtual clients in order to transfer data between them. By using this interface unit, it can able to connect and send data to more than 3 clients from server. The server and client can be connected by giving their corresponding IP address, so that it gets enabled and delivering excellent end to end performance across wide-area [5]. Port numbers for VMs are assigned in coding itself. It also indicates whether the VMs are enabled or not to send the data.

Virtualization

Virtualization is reducing the need for physical hardware systems; saves cost and provide incremental scalability of hardware resources [3]. Virtualization requires more bandwidth, processing capacity and storage space compare to traditional server or desktop, if the physical hardware is going to host multiple running virtual machines on it.

Virtual Machine

A Virtual Machine (VM) is an operating system or application environment that is installed on software which duplicates dedicated hardware. The end user has the same experience on a virtual machine as they would have on dedicated hardware. Specialized software called a hypervisor imitates the server's CPU or PC client, hard disk, memory, network and other hardware resources completely, enabling virtual machines to share the resources. The hypervisor can imitate multiple virtual hardware platforms which are isolated from each other, allowing virtual machines to run on different platforms such as Linux and Windows server operating systems on the same underlying physical host. VMs use hardware more efficiently, which lowers the quantities of hardware and related maintenance costs. Also it reduces power and cooling demand. They also easily managed because virtual hardware does not fail. Administrators can take advantage of virtual environments to simplify backups, disaster recovery, new deployments and basic system administration tasks. VMs can easily move, copied and reassigned between host servers to adjust hardware resource utilization. Because VMs on a physical host can consume unequal resource quantities (one may hog the available physical storage while another stores little), IT professionals must balance VMs with available resources. In this project, VMs are created by using java coding.

(E) Smart Frame Processing's

In cloud computing application domains of large-scale, suggest a secure cloud computing based framework for big data management in smart grids, which is called "Smart-Frame". The main idea of our framework is to build a hierarchical structure of cloud computing cores to provide different types of services for information management and big data analysis. Privacy is ever-growing concern in our society and is becoming a fundamental aspect to take into account when one wants to use, publish and analyze data involving human personal sensitive information [7]. So, in addition to this structural framework, they present a security solution based on identity-based encryption, signature and proxy re-encryption to address critical security issues of the proposed framework. In this project, for encryption, digital signature and decryption RSA with SHA1 is used as jar file. First, keys are generated for the data which is to be sent. Then, digital signatures are created for each and every partition. Once the packet reached at client side, automatic decryption takes place.

(F) Data Integrity Addressing

It takes a lot of understanding to get data in the correct shape so that you can use visualization as part of data analysis. For example, if the data comes from social media, there is needed to know who the user is in general sense. One solution to this challenge is to have the appropriate domain expertise in place. Make sure the people examining the data have a deep understanding of where the data arrives from, what viewers will be consuming the data and how they will interpret the information.

Of course, not all of the information would be of value if it was tracked, but certainly some of it would be valuable. The growth of useful information should be an inspiration for the implementation of Big Data technologies and practice. The complexity rises in managing, securing, storing and extracting values [18]. Even if the data can find and analyze quickly and put it in the suitable context for the audience who will be consuming the information, the value of data for decision-making purposes will be endangered if the data is not accurate or in time. This is a challenge with any data analysis, but when considering the volumes of information involved in big data projects, it becomes even more pronounced. Again, data visualization will only prove to be a valuable tool if the data quality or integrity is assured.

DATASET

For this project implementation, static dataset are taken from UCI Repository, which are freely available in internet. This paper contains 2 types of operation. 1) Dynamic, which is created each and every time when medical data is uploaded. 2) Static, which contains minimum of 4 data sets. These data sets contain information about people who are affected by disease such as thyroid, diabetes, blood pressure, lung cancer, liver disorder and heart disease. This paper uses dataset which contains above 1 lakh of records and size of 18,711 kb.

PERFORMANCE EVALUATION

The performance of the proposed scheme is defined by means of security as well as successful data uploading with big data with virtual cloud scheme. Extensive security and performance analysis shows that the proposed schemes are provably secure and highly efficient. We believe all these advantages of the proposed schemes will shed light on economies of scale for resource maintenance. To understand the practicality of the integration of Big Data with Virtual Cloud and preventive measures, we analyze its security strength, evaluate its running time overhead via tested experiments.

Analysis of running times

This system proposes the design and implementation of a practical big data scheme for data maintenance in remote storage with cloud mechanisms. We augment the implementation of the virtual scheme and construct preventive measures via secured cryptography schemes, a code that allows virtual clients to remotely verify the integrity of random secret keys of long-term archival under a server setting. This aims to achieve several design features. First, it preserves fault tolerance and repair traffic saving as in big data with cloud mechanisms. Second, it assumes only the thin cloud interface, i.e., the server only need to support the standard transactional functionalities. Third, it exports several tunable parameters that allow clients to trade performance for security.We Implement and evaluate its overhead over the existing big data implementation schemes through extensive tested experiments in a storage environment. We evaluate the running times of different basic operations, including upload, check, download, and repair, for different parameter choices of our scheme. In addition, we evaluate the monetary cost overhead of big data should it be deployed in commercial servers. Our work demonstrates the feasibility of enabling attack protection, fault tolerance, and efficient data transactions for remote storage. Fig. 2, indicates the partitioning of dataset or resource set whose partitioning speed is bits per second.

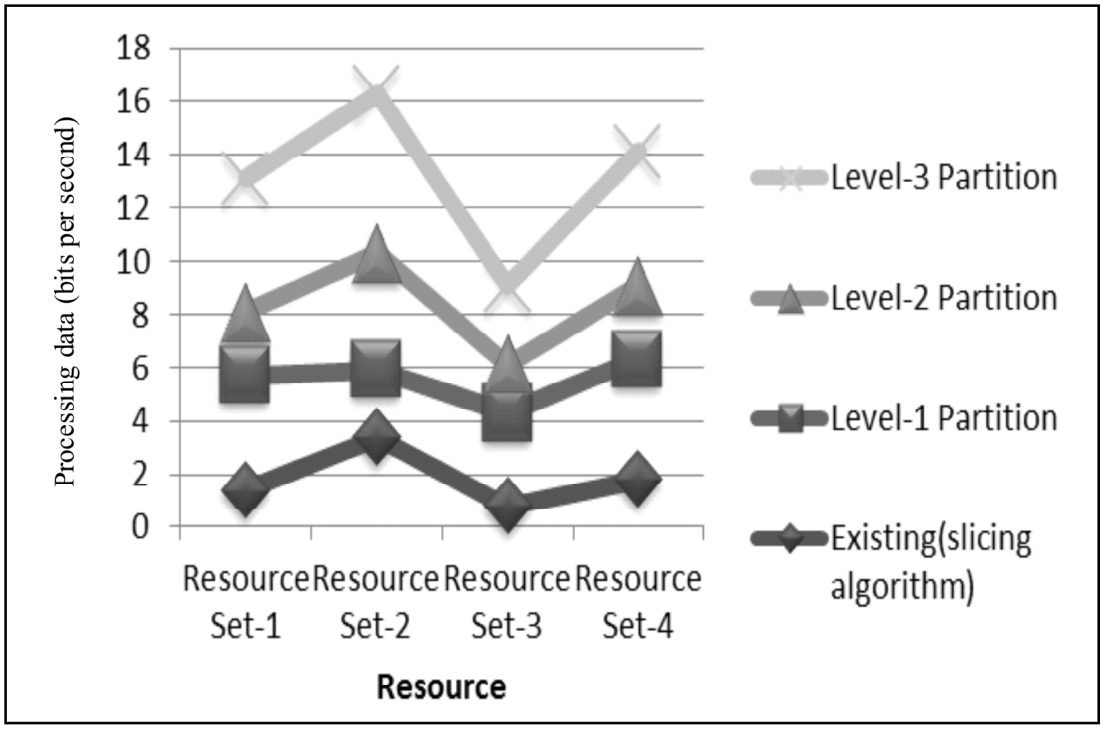

Figure 2: Comparison graph (Partitioning algorithm Vs. Slicing algorithm)

Fig. 3, represents the time taken to upload the dataset in client. In the case of bigdata, many problems arise when we upload the data as whole like bottleneck problem, network traffic, loss of data, etc., In order to avoid this problem, we partition the dataset.

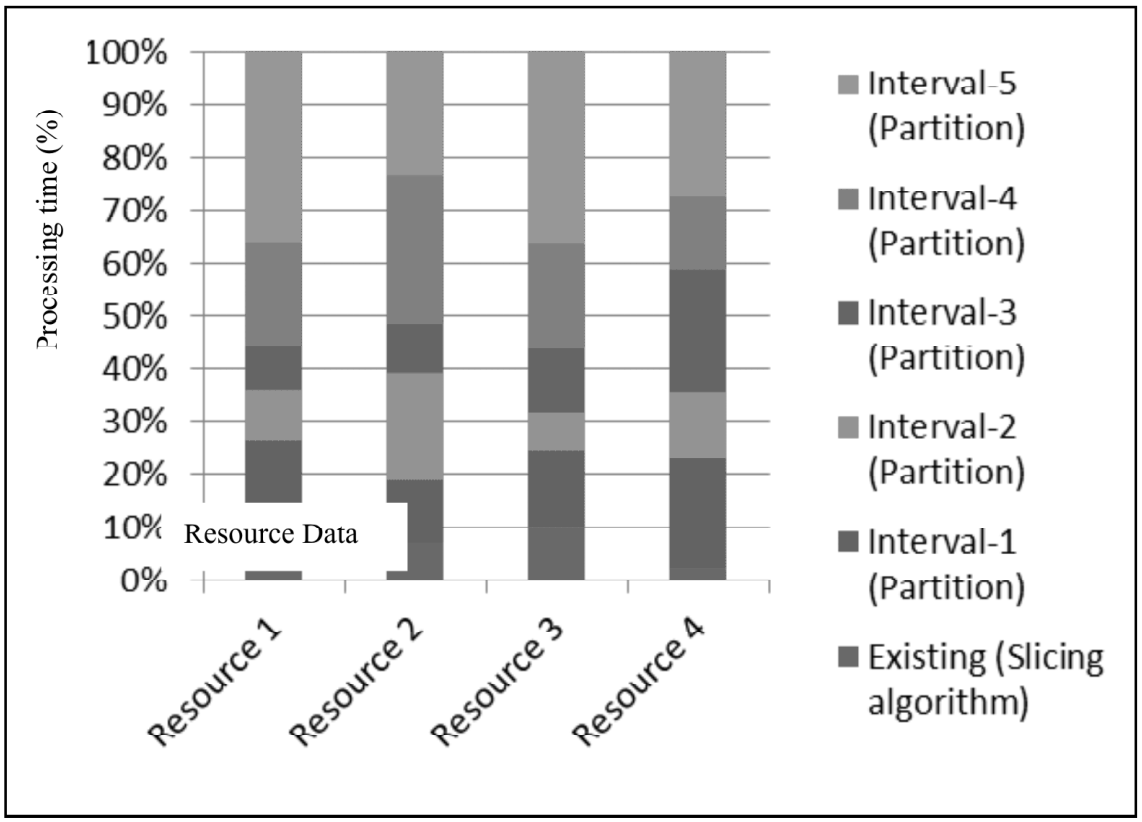

Figure 3: Processing time (Partitioning algorithm Vs. Slicing algorithm)

Here balanced partitioning algorithm is used which can be able to partition the static as well as dynamic dataset and more efficient then existing algorithms like slicing. Partitioning refers to splitting the dataset corresponding to buffer value and bytes value. These segmented data's form the packets. Partition increases as increased in byte value and decreases as increased in buffer value. Partitioning can be calculated using formula

Partition = (byte value / buffer value).

In Fig. 2, level 1 reaches the maximum of 4.5 partitions at resource set 4, level 2 reaches the maximum of 4.4 partitions at resource set 2 and level 3 reaches the maximum of 6 partitions at resource set 2. This algorithm gives better performance than the existing algorithm.

After partitioning the data is uploaded into corresponding client. Depends upon the packet size and network traffic the time varies during simultaneous upload of dataset in different client. In Fig 3, five time intervals are considered. Partition 1 takes the maximum time of 28% to upload the resource set1. Partition 2 takes the maximum time of 40% to upload the resource set2. Partition 3 takes the maximum time of 58% to upload the resource set4. Partition 4 takes the maximum time of 78% to upload the resource set2. Partition 5 takes the maximum time of 100% to upload. This shows the efficiency of uploading the entire dataset as packets in 100% time. The security mechanisms like key generation and digital signature are also included in this work.

CONCLUSION

Big data is nothing but structured as well as unstructured, uncertain, real-time data that is present in a massive amount. Handling such data's and integrate with cloud is quite difficult. In this system, balanced partitioning technique, virtualization and aggregations are implemented in order to achieve the integration. In cloud computing application domains of large-scale, propose a safe cloud computing based framework for big data management in smart grids, which is called "Smart-Frame-". The main idea of our framework is to build a hierarchical structure of cloud computing centers to provide different types of computing services for information management and big data analysis.

ACKNOWLEDGEMENT

I wish to express my sincere thanks and deep sense of gratitude to Prof. V. Srinivasan, M.E., Professor and Head, Department of Computer Science and Engineering and Director of AIC, for having me the opportunity to undertake this project. I would like to convey my heartiest thanks to my project guide Dr. K. Venkatachalapathy, MCA, Ph.D., Professor, Department of Computer Science and Engineering, for all his help and support. He with his extreme patience has guided me in situation of need for which I am extremely grateful.

REFERENCES

[1] ArtiMohanpurkar and Prasadkumar Kale, "Big Data Analysis using Partition Technique", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6, issue 3, pp. 2871-2875, 2015.

[2] Anitha S Pillai and Jisha Jose Panackal, "Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets", Procedia Computer Science, Vol. 50, pp. 347-352, 2015.

[3] Andrea C. ArpaciDusseau, Remzi H. ArpaciDusseau, David E. Culler, Joseph M. Hellerstein and David A. Patterson, "High-Performance Sorting on Networks of Workstations", ACM 0-89791 -911 -419710005, pp.243-254, 1997.

[4] ArashBaratloo, Mehmet Karaul, ZviKedem and PeterWyckoff, "charlotte: metacomputing on the web",Appeared in the 9th International Conference on Parallel and Distributed Computing Systems (PDCS), 1996.

[5] Andrea C. ArpaciDusseau, Remzi H. ArpaciDusseau and MironLivny, "Explicit Control in a Batch-aware Distributed File System", USENIX Association, 2004.

[6] Vijay Thayananthan and AiiadAlbeshri, "Big data security issues based on quantum cryptography and privacy with authentication for mobile data center", Procedia Computer Science, Vol. 50, pp. 149-156, 2015.

[7] Anna Monreale, Salvatore Rinzivillo, Francesca Pratesi, FoscaGiannotti and Dino Pedreschi, "Privacy-by-design in big data analytics and social mining", EPJ Data Science, 2014.

[8] Guy E. Blelloch, "Scans As Primitive Parallel Operation" Carnegie Mellon University, November 1989.

[9] ChingHsien Hsu, Daiqiang Zhang, KennSlagter and Yeh-Ching Chung, "An improved partitioning mechanism for optimizing massive data analysis using Map Reduce", Springer, pp. 540-555, April 2013.

[10] Craig C. Douglas, "An Open Framework for Dynamic Big-Data-Driven Application Systems (DBDDAS) Development" ICCS 2014, 14th International Conference on Computational Science, Volume 29, Pages 1246–1255, 2014.

[11] Luiz Andre Barroso, Jeffrey Dean and UrsHolzle, "web search for a planet: the google cluster architecture" IEEE Computer Society 0272-1732/03, 2003.

[12] Lizhe Wang and Rajiv Ranjan, "Processing Distributed Internet of Things Data in Clouds", IEEE Cloud Computing. pp. 2325-6095, February 2015.

[13] Li Xiong, VaidySunderam, Liyue Fan, Slawomir Goryczka and LaylaPournajaf, "PREDICT: Privacy and Security Enhancing Dynamic Information Collection and Monitoring", International Conference on Computational Science, ICCS 2013.

[14] Siddaraju, Sowmya C L,Rashmi K and Rahul M, "Efficient Analysis of Big Data Using Map Reduce Framework", International Journal of Recent Development in Engineering and Technology, Vol. 2, Issue 6, June 2014.

[15] Satoshi Tsuchiya, Yoshinori Sakamoto, Yuichi Tsuchimoto and Vivian Lee "Big Data Processing in Cloud Environments", FUJITSU Sci. Tech. J., Vol. 48, No. 2, pp. 159-168, April 2012.

[16] CISCO Corporation Inc., "Tomorrow's Internet. Today", DDM11CS3243, June 2011.

[17] Remzi H. Arpaci-Dusseau, Eric Anderson, Noah Treuhaft, David E. Culler, Joseph M. Hellerstein, David Patterson and Kathy Yelick, "cluster i/o with river: making the fast case common", Computer Science Division, University of California, Berkeley, DRAFT.

[18] JhonGantz and David Reinselet., "The digital Universe in 2020: Big Data, Bigger Digital Shadows, and Bigger Growth in the Far East", EMC Corporation, February 2013.

[19] Focus Group on Cloud Computing., "Cloud computing benefits from telecommunication and ICT perspectives", part. 7, February 2012.

[20] Fred Zimmerman., "Bringing Big Data into the Enterprise", Enterprise Executive Magazine, June 2013.

[21] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplifed Data Processing on Large Clusters" Google Inc, OSDI 2004.

[22] Thyroid dataset is taken from: *https://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/*

[23] Heart disease dataset is taken from: *https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/*

[24] Other dataset link: *https://archive.ics.uci.edu/ml/datasets.html*