International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 4, April 2015

Secured Multimedia Processing in Big data Using DataCloud and Mapreduce Algorithm

[1]Varnachithra.A, [2]Sahishna Krishna R.G, [3]Neethy C Nair, [4]Navya Anna Moncy, [5]Navin K.S
[1]varna3000@gmail.com,[2]sahishnakrishna@gmail.com, [3]neethynair93@gmail.com,
[4]namzz.moncy92@gmail.com, [5]navinsivendran@gmail.com

Abstract: Abstract— **Big Data is a data analysis methodology enabled by recent advances in technologies and architectures. The scalability issue of big data leads towards cloud computing which offers the promise of big data implementation to small and medium sized businesses. Our paper proposes Big Data processing in cloud through a programming paradigm known as MapReduce. We also propose a novel architecture for the future Internet based on information-centric networking which is called Community Oriented DataClouds where users under common interest are brought together into a community. As a sample application of the mentioned concept, we perform multimedia processing and storage in a private cloud. After data processing, clustering is done and encryption is performed on it. For data encryption and decryption we use blowfish algorithm that supports for all file formats. Integrating all the multimedia processing under a single application and storing it in DataClouds or cloud communities make this a different approach.**

Keywords: **DataCloud, MapReduce**

I. INTRODUCTION

The concept of big data has been endemic within computer science since the earliest days of computing. Big data originally meant the volume of data that could not be processed by traditional database methods and tools. We define "Big Data" as the amount of data just beyond technologies capability to store manage and process efficiently. Today we are thinking in tens to hundreds of terabyte. Thus big data is a moving target. This large amount of data is just beyond our immediate grasp.

The new revolution "The Cloud" was a part of a major technology shift which addressed the issues of handling and processing big data. The Cloud is a shared network of computers through which people and companies store data and run software.

Every individual has had experience with multimedia systems of one type or another. The dictionary definition of multimedia is: including or involving the use of several media of communication, entertainment, or expression. A more technological definition of multimedia, as it applies to communications systems is integration of two or more of the following media for the purpose of

transmission,storage, access, and content creation like text, images, graphics, audio, video etc.

In our proposed system we introduce a sample application involving various multimedia processing on different media files. Under image enhancement we perform gray-scale conversion, image encoding, resizing image and steganography. Under text enhancement we convert word documents to pdf documents and vice-versa. Under audio enhancement noise removal is done. Under video enhancement video compression and quality is improved.

For small scale processing of these data, local database server will be enough. But in case of large scale processing, the large amount of data i.e...Big Data cannot be handled by the local server. This is where "Cloud" plays its part. Among the different types of cloud like public, private and hybrid, private cloud is efficient for secure and personalized data handling.

In cloud data is stored randomly which is more time consuming. A new concept called DataCloud is introduced here. Based on this concept data is clustered according to their domains and stored as cloudlets. Cloudlets are logical spaces in cloud.

II. SYSTEM ARCHITECTURE

The system architecture includes different phases of implementation. The first phase deals with the data processing. In the second phase data clustering and encryption is done. The third phase is the mapping of big data set to DataCloud. The user interaction is explained in the final phase.

Our proposed system is a sample application for storing and retrieval of processed multimedia data.

a) DATA PROCESSING PHASE:

The multimedia data is collected from user and the desired processing is done.

- *Image processing:*

Processing of images is done to bring out its specific features. Image processing is any form of signal processing for which the input is an image, such as a photograph or video frame the output of image processing may be either an image or a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. These techniques are most useful because many satellite images when examined on a color display give inadequate information for image interpretation.
Various processing techniques included are

1. Gray-Scale conversion

In a (8-bit) grayscale image each picture element has an assigned intensity that ranges from 0 to 255. Gray levels represent the interval number of quantization in gray scale image processing. At present, the most commonly used storage method is 8-bit storage. There are 256 gray levels in an 8 bit gray scale image with 0 being black and 255 being white. Another commonly used storage method is 1-bit storage. There are two gray levels, with 0 being black and 1 being white a binary image.
A gray scale image is what people normally call a black and white image, but the name emphasizes that such an image will also include many shades of grey as shown in figure 1. A normal grayscale image has 8 bit color depth = 256 grayscales. A "true color" image has 24 bit color depth = 8 x 8 x 8 bits = 256 x 256 x 256 colors = ~16 million colors.
For images in color spaces such as Y'UV and its relatives, which are used in standard color TV and video systems such as PAL , SECAM and NTSC, a nonlinear luma component(*Y'*) is calculated directly from gamma-compressed primary intensities as a weighted sum, which can be calculated quickly without the gamma expansion and compression used in colorimetric grayscale calculations. In the Y'UV and Y'IQ models used by PAL and NTSC, the rec601 luma (*Y'*) component is computed as

$$Y' = 0.299R' + 0.587G' + 0.114B'$$

where we use the prime to distinguish these gamma-compressed values from the linear R, G, B, and Y discussed above.

Fig 1: Gray scale conversion

2. Steganography

Taking the cover object as image in steganography is known as image steganography as shown in figure 2. Generally, in this technique pixel intensities are used to hide the information. It is the art or practice of concealing a file, message, image, or video within another file, message, image, or video. The advantage of steganography over cryptography alone is that the intended secret message does not attract attention to itself as an object of scrutiny. Plainly visible encrypted messages—no matter how unbreakable—will arouse interest, and may in themselves be incriminating in countries where encryption is illegal Thus, whereas cryptography is the practice of protecting the contents of a message alone, steganography is concerned with concealing the fact that a secret message is being sent, as well as concealing the contents of the message.

Today's steganography systems use multimedia objects like image, audio, video etc as cover media because people often transmit digital images over email or share them through other internet communication application. It is different from protecting the actual content of a message. In simple words it would be like that, hiding information into other information. Steganography means is not to alter the structure of the secret message, but hides it inside a cover-object (carrier object). After hiding process cover object and stego-object (carrying hidden information object) are similar. Due to invisibility or hidden factor it is difficult to recover information without known procedure in steganography. Detecting procedure of steganography known as Steganalysis.

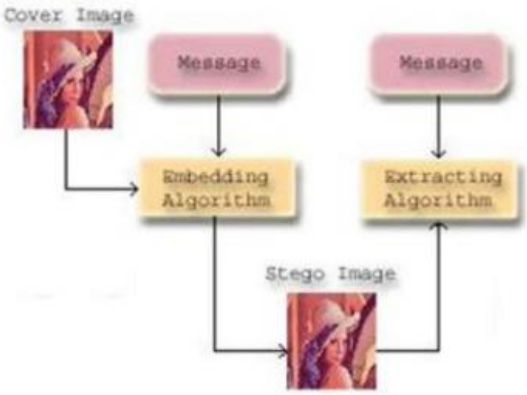

Fig 2: Steganography

3. Image encoding

For providing security image file can be encoded into a string file using base64 encoder. **Base64** is a group of similar binary-to-text encoding schemes that represent binary data in an ASCII string format by translating it into a radix -64 representation. The term *Base64*originates from a specific MIME content transfer encoding.

Base64 encoding schemes are commonly used when there is a need to encode binary data that need to be stored and transferred over media that are designed to deal with textual data. This is to ensure that the data remain intact without modification during transport. Base64 is commonly used in a number of applications, including email via MIME, and storing complex data in XML. The figure 3 shows the above procedure.

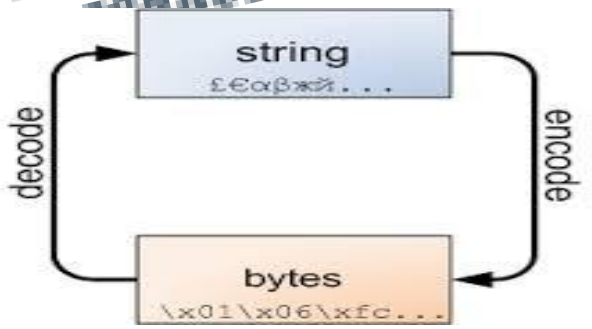

Fig 3: Image encoding

- *Text Processing*

In computing, the term text processing refers to the discipline of mechanizing the creation or manipulation of electronic text. Text usually refers to all the alphanumeric characters specified on the keyboard of the person performing the mechanization, but in general *text* here means the abstraction layer that is one layer above the standard character encoding of the target text. The term processing refers to automated (or mechanized) processing, as opposed to the same manipulation done manually.

- *Video processing*

Video processing uses hardware, software, and combinations of the two for editing the images and sound recorded in video files. Extensive algorithms in the processing software and the peripheral equipment allow the user to perform editing functions using various filters. The desired effects can be produced by editing frame by frame or in larger batches.

III. Video Compression

Video takes up a lot of space. Uncompressed footage from a camcorder takes up about 17MB per second of video. Because it takes up so much space, video must be compressed before it is put on the web. "Compressed" just means that the information is packed into a smaller space. There are two kinds of compression: lossy and lossless. Lossy compression means that the compressed file has less data in it than the original file. In some cases this translates to lower quality files, because information has been "lost," hence the name. Lossless compression is exactly what it sounds like, compression where none of the information is lost. This is not nearly as useful because files often end up being the same size as they were before compression. Compression is a reversible conversion (encoding) of data that contains fewer bits. This allows a more efficient storage and transmission of the data. Software and hardware that can encode and decode are called decoders. Both combined form a codec and should not be confused with the terms data container or compression algorithms.

Bulk of data processing and single data processing can be done.

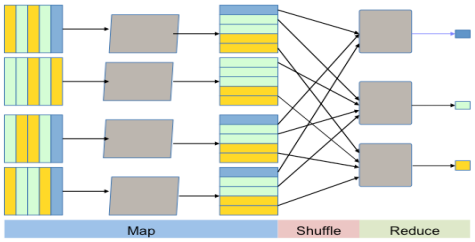

Fig 4: MapReduce framework

IV. MAPREDUCE FRAMEWORK

For clustering and storing of data, we use an efficient algorithm called MapReduce. MapReduce is a programming model used for parallel processing of massive data sets. MapReduce programs solve computational problems by applying, in parallel, a transformation function (called a "mapper") to a (typically large) set of input records; the transformed records are then grouped according to some key value, and the resulting groups are then "reduced" by some aggregation function.

There are three core phases, map, shuffle and reduce.

- the map phase consumes a set of records structured as key-value pairs, and outputs a transformed set of records also structured as key-value pairs
- The shuffle phase gathers all values with the same key into one place, so they can be processed together.
- In the reduce phase, the gathered values from each key are aggregated into a final result.

For e.g. in a college students are categorized based on their departments. Once this categorization is done, further additions are made to this group. A benefit of the MapReduce paradigm is its fault-tolerance. If a mapper fails, then its input split can be sent to another machine to be quickly recomputed. Indeed, horizontal scalability would not be possible without fault-tolerance of some kind, which is an important reason MapReduce has been so successful.

b) CLUSTERING AND ENCRYPTION PHASE

The data is collected from different sources such as PC's, laptops, Smartphone devices, Tablets etc. These data comes from different domains depending on the user. They are processed as per the user's request. Then these large amounts of processed data are to be partitioned according to their domains such as image, text, video and audio. This data is encrypted and are stored.

We can use any encryption algorithms in this context. Here for more easy and secured encryption, we use Blowfish algorithm. Blowfish is a symmetric block cipher that can be effectively used for encryption and safeguarding of data. Blowfish is a variable-length key block cipher. It is suitable for applications where the key does not change often, like a communications link or an automatic file encryptor. It is significantly faster than most encryption algorithms when implemented on 32-bit microprocessors with large data caches.

In this paper, we generate a One Time Password (OTP) for authenticating the user in the user interaction phase which will be explained later. This OTP is taken as the key for the blowfish algorithm. Since it is a symmetric key cryptosystem, decryption is also done with the same key. After encryption the encrypted data is mapped to the logical divisions called cloudlets.

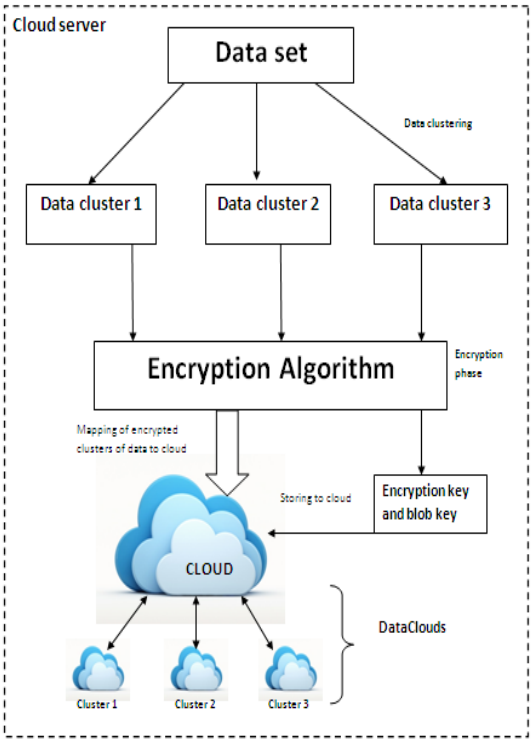

Fig 5: System architecture

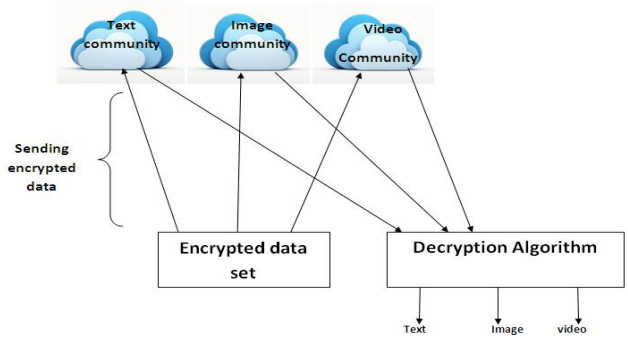

Fig 6: Storage and retrieval

c) MAPPING OF BIG DATA TO V. DATACLOUDS

Using MapReduce framework, we map the processed and encrypted data to the DataCloud according to their domains. The images are mapped to the image cloud and text data are mapped to text cloud, likewise for video and audio. For lower end user applications we can just map the data to the DataClouds with respect to their domains.

For Big data storage, we consider DataClouds as communities. The Data cloud is evolved from two divisions; the physical and the logical space. The logical space captures user relationships and interests in data content. Here, we connect the users who share common interests in one data centric service, which will divide the users into different groups, called *logical communities*. The physical space manages physical network connections among users. A user can access the data from the community if he is authorized by simply requesting the community manager as < Service Type | Hierarchical Community Name (HCN) |Domain Name >. Since communities are established, data dissemination could be efficiently achieved among them without duplicate transmissions, which significantly reduces the communication overhead. Also when divided into groups of common interests, we can perform easy search and fast access which in turn reduces network traffic and increases time efficiency.

Network Entities

DataClouds consists of several kinds of new network entities, which are described as follows. The figure 7 explains the DataCloud architecture.

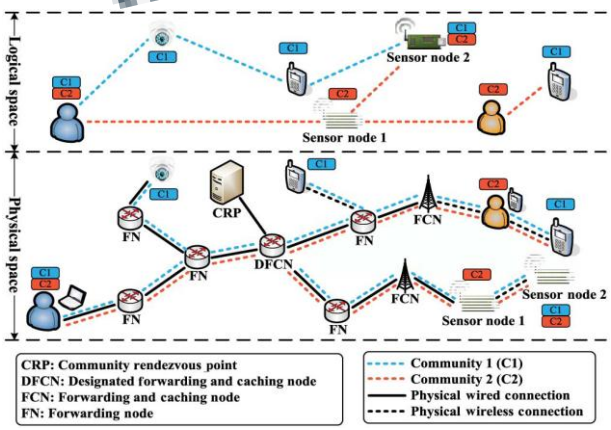

Fig 7: DataCloud architecture

1) *CRPs*: In DataClouds, there is one CRP in each administrative domain, which records the CIDs of all the communities in this domain. The main functionalities of CRPs include searching for desired communities based on users' interests and assisting them to join the communities in physical space.

2) *FNs*: FNs work as intermediate nodes to forward data and control messages. The routing function at FNs is achieved with two modules, called the *Forwarding Information Base (FIB)* and the *Request Pending Table (RPT)*. FI maintains the CIDs of different communities and the corresponding incoming and outgoing interfaces for data forwarding. The RPT records the routes created between users and the CRP for the transmissions of control messages when the users initiate new communities or join existing communities as shown in fig 3.

3) *FCNs*: FCNs have all the functionalities of FNs. In addition, FCNs are equipped with memories to cache/store data items in order to facilitate future data retrieval.

4) *Designated forwarding and caching nodes (DFCNs)*: When a new community is initiated, the CRP will designate one FCN to it, which serves as the root of the communication structure of the community in physical space, i.e., every member in this community must be connected to this FCN so that it can share data with or retrieve data from others.

Community Management

In DataClouds, communities are created and maintained through the following operations.

1) *Community initialization*: When a user has data to disseminate but there are no communities existing for her, the user can create a new community for the users with the same interest to share the data. To do so, the user needs to register the new community at the local CRP, and the CRP will assign a DFCN to the community. After that, a route is established between the user and the DFCN, which forms the original communication structure of the community.

2) *Community joining*: After querying the CRP, a user can discover and join the communities in the network that meet their interests in order to publish or receive the desired data. Similar to community initialization, the user has to find a route and connect to the DFCNs of the corresponding communities. In this way, the community, or the cloudlet, is expanding.

3) *Community leaving*: As the interests of users might

change over time, they can choose not to share data with other users in certain communities anymore by leaving it. When a user decides to leave a community, the user needs to inform the DFCN so that the route between the user and the community can be removed from the communication structure. In this way, the community, or the cloudlet, is shrinking.

4) *Community deletion*: When there are no users in a community, the DFCN will inform the CRP to delete the corresponding record for this community. The DFCN is freed from that community and can be reassigned to new communities in the future. Thus, the community or cloudlet disappears.

Thus the big data set is anonymized and mapped to the cloud. The blob keys, values and encrypted keys are stored along with the metadata in the database server. By the help of an internet service provider (ISP), the user can access the data from cloud. In the cloud the anonymized data are grouped as different cloudlets as DataClouds based on their domains. For example, the big dataset of an insurance company can be grouped as different communities like healthcare, education etc. Thus big data can be efficiently mapped to cloud.

d) USER INTERACTION PHASE

In the user interaction phase, if a user needs to access a data, he query for the data to the Internet Service Provider. The request is forwarded to the cloud database administrator. The administrator checks whether the user is valid or not. If the user is valid then data access will be granted else access denied. When the user login through the interface, an OTP is generated a send to his valid email id. The user can select his desired type of multimedia and can process it. This processed data is encrypted by using the OTP as key which was generated at the time of his login. Then the encrypted data is stored in the cloud. Using the same key we can decrypt the data and download it for use.

CONCLUSION

As a sample application we see multimedia processing (that include image,text,video and audio) in private cloud .In image enhancement we introduced grey scale conversion, image resizing, image encoding and stegonography.For text we performed text to pdf conversion and vice versa .We performed compression techniques on video and audio. In order to map the processed data to cloud, we used map reduce algorithm where we see mapping, reducing and shuffling phases. We get a blob key as a result of this algorithm. Using this blob key the processed data is retrieved from the cloud .For small data set we can divide the private cloud into different clusters and store processed data .For large scale dataset, our proposed concept- data cloud is used where the processed data within the same community are brought together via communication structure, and the community structure is intelligently maintained so that data items can be efficiently disseminated or retrieved. A backup of the original data is stored in the private cloud for future references. Our proposed system solves the issue of scalability, in a local database server. Also the concept of data cloud facilitates the fast retrieval of big data application.

REFERENCES

[1] Xuyun Zhang, Wanchun Dou, Jian Pei, Surya Nepal, Chi Yang, Chang Liu, and Jinjun Chen, Member, IEEE *" Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud"* IEEE TRANSACTIONS ON COMPUTERS, TC-2013-12-0869.

[2] Hao Yue, Student Member, IEEE, Linke Guo, Member, IEEE, Ruidong Li, Member, IEEE,Hitoshi Asaeda, Senior Member, IEEE, and Yuguang Fang, Fellow, IEEE "*Data Clouds: Enabling Community-Based Data-Centric Services Over the Internet of Things*" IEEE INTERNET OF THINGS JOURNAL, VOL. 1, NO. 5, OCTOBER 2014.

[3] Sangeeta Bansal, Dr. Ajay Rana Department of Computer Science & Engineering Amity University, Noida (U.P.) India "*Transitioning from Relational Databases to Big Data*" Volume 4, Issue 1, January 2014 International Journal of Advanced Research in Computer Science and Software Engineering.

[4] http://www.datahouse.com/assets/files/Cloud-1.0.pdf "*INTRODUCTION THE CLOUD The Evolution of The Cloud*".

[5] Samira Daneshyar and Ahmed Patel School of Computer Science, Centre of Software Technology and Management (SOFTEM), Faculty of Information Science and Technology, "*EVALUATION OF DATA PROCESSING USING MAPREDUCE FRAMEWORK IN CLOUD AND STANDALONE COMPUTING*" International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.6, November 2012

[6] T. Morkel , J.H.P. Eloff , M.S. Olivier "*AN OVERVIEW OF IMAGE STEGANOGRAPHY*" Information and Computer Security Architecture (ICSA) Research Group Department of Computer Science.

[7] R. ANANDHI, K. CHITRA "*A Challenge in Improving the Consistency of Transactions in Cloud Databases – Scalability*"

International Journal of Computer Applications (0975 – 8887) Volume 52– No.2, August 2012.

[8] Kalpana Parsi, M.Laharika Department of Computer Applications, Sreenidhi Institute of Science and Technology," *A Comparative Study of Different Deployment Models in a Cloud*" Volume 3, Issue 5, May 2013 International Journal of Advanced Research in Computer Science and Software Engineering.