International Journal for Modern Trends in Science and Technology
Volume: 03, Special Issue No: 01, February 2017
ISSN: 2455-3778
http://www.ijmtst.com

A Survey on Hybrid Cloud with De-Duplication

Ch.Jyosthna Devi[1] | Dasari Lucky Living Ston[2] | Madatala Keerthi[3]

[1]Assistant Professor, Department of CSE, Chalapathi Institute of Engineering and Technology, Guntur, AP, India.
[2,3]UG Students, Department of CSE, Chalapathi Institute of Engineering and Technology, Guntur, AP, India.

To Cite this Article
Ch.Jyosthna Devi, Dasari Lucky Living Ston and Madatala Keerthi, "A Survey on Hybrid Cloud with De-Duplication", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Special Issue 01, 2017, pp. 18-26.

ABSTRACT

A survey on industry trends is been noted where the usage of hybrid cloud architecture can be used which supports, the upcoming industry challenges by providing the efficient way of storing their data in the cloud environment by using the combination of both public and private clouds, So that it provides the facility to store sensitive data in private cloud and less critical data on to the public cloud where huge savings can be made. Since the demand for data storage is increasing day by day and by the industry analysis we can say that digital data is increasing day by day, but the storage of redundant data is excess which results in most of the storage used unnecessary to keep identical copies. So the technology de-duplication is introduced to efficiently utilize the cloud storage system. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data de-duplication. Different from traditional de-duplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new de-duplication constructions supporting authorized duplicate check in a hybrid cloud architecture.

KEYWORDS: *Hybrid cloud; de-duplication; hash numbers; encryption*

I. INTRODUCTION

Despite of cloud computing enormous popularity many companies are being dissatisfied as they have not found what they need in the single cloud environment.

- Private clouds, which companies run internally, are secure and accessible within the local area network.
- Public clouds are operated and run over internet and are less costly, scalable and easy to use.

Until recently, companies could not economically and easily integrate and operate with both the types of architectures as a one cloud system. So nowadays companies are merging both the public and private approaches, creating a single hybrid cloud that will reap the benefits of each stated Panos Tsirigotis, cofounder of hybrid cloud vendor cloud velocity and chief software architect.

Adopting hybrid cloud is very easy for many companies as they will be having in-house cloud and will require only leverage the existing public cloud capabilities noted Professor kantarcioglu.

Business organization data volumes are increasing as companies store and collect huge amount of data for their own use in cloud. According to Business organization strategy group, by industry analysis more organizations prefer to store their data on to cloud. Thus requires organization to have more storage and consume more power and energy for managing and handling the data, more network resources are utilized for transmitting the data and more time is spend on functions such as replication and data backup.

Most of the information that is stored is duplicate data, different sources in the same organizations usually create similar files or duplicate files that already exist by which they can work independently.

If it was possible, IT organizations would only protect the unique data from their backups. Instead of saving everything repeatedly, the ideal scenario is one where only the new or unique content is saved. Data de-duplication provides this basic capability. It offers the ability to discover and remove redundant data from within a dataset. A dataset can span a single application or span an entire organization.

Redundant data elements can be entire files or sub-file data segments within a file. In all cases, the objective of the de-duplication process is to store unique data elements only once, yet be able to reconstitute all content in its original form on demand, with 100 percent reliability at disk speeds.

Data de-duplication is fundamental to improving information protection, streamlining backup operations, reducing backup infrastructure, shortening backup windows, and removing burden from information networks.

II. Literature Survey

Hybrid Cloud is the architecture that provides the Organization to efficiently work on both the private and public cloud architecture in combination by providing the scalability to adopt. Here some of the basic concepts and idea proposed by authors and how best and easy to adopt this environment is explained by Neal Leavitt.

An intelligent workload factoring, service for organization customers which makes the best use of the present public Cloud services including their private owned data canters. It allows the organization to work between the off-premises and the on-premises infrastructure. The efficient core technology that is used for intelligent workload factoring is a fast redundant data element detection algorithm, that helps us factoring all the incoming requests based on the data content and not only on volume of data, Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and Akhilesh Saxena.

The term —Cloud has many definitions one among them is to provide infrastructure as a service system where the IT infrastructure will be deployed in the particular cloud service provider, data center as virtual machine. The growing popularity of laas will help us to transform the organization present infrastructure into the required hybrid cloud or private cloud. OpenNebula Concept is being used that will provide the features that are not present in any other cloud software, Borja Sotomayor ,Rubén S. Montero and Ignacio M. Llorente, Ian Foster.

Data Deduplication is a technique that is mainly used for reducing the redundant data in the storage system which will unnecessarily use more bandwidth and network. So here some common technique is being defined which finds the hash for the particular file and with that the process of deduplication can be simplified.

In the real world more often we tend to see the data that are two or more in database. The records which are duplicate will share the different keys that will make the duplicates matching task difficult and will result in errors. Errors will usually occur due to lack of standard formats, incomplete information or transcription errors. The through analysis of duplicate record detection literature survey is done in this paper. The duplicate detection algorithm is used which detect s the duplicate records and also some of the metrics are considered that will help us to detect the similar field entry of data that is done. Multiple techniques are presented that will help us to improve the efficiency and the existing tools that is present is being covered, Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios.

De-duplication is the technique that is most effective most widely used but when it is applied across the multiple users the cross-user deduplication tend to have to many serious privacy implications. Simple mechanisms can be used which can enable the cross-user deduplication which will reduce the risks of the data leakage and also some of the security issues are discussed with how exactly to identify the files and to encrypt them while sending is discussed, Danny Harnik, Benny Pinkas, Alexandra Shulman- Peleg.

Data that is being collected from several data sources are being stored in the repositories called data warehouse. During the ETL (Extraction, transformation, loading) or OLTP (On Line Transaction Processing) in the data warehouse we often tend to find the duplicate copies of data in the table. Since the quality of data is very essential to gain the confident of users, more amount of money and time is being spent in obtaining the high quality data. Data cleaning is the process where the dirty data is removed. Here they have discussed

some methods and strategies to remove duplicate data by, Srivatsa maddodi, Girija V. Attigeri, Dr karunakar A.k.

III. HYBRID HIGHLIGHTS

Hybrid cloud can be built using any technology it varies according to different vendors. Key components In many of the situations, implementation of the hybrid cloud has a controller that will keep track of all locations of private and public clouds, IP address, servers and other resources that can run systems efficiently.

Some of the key components include

- Orchestration manager and cloud provisioning for storage, public cloud resources which includes virtual machines and networks, the private and public clouds, which are not necessarily compatible or identical.
- Synchronization element and Data transfer efficiently exchange information between private and public clouds.
- Changing configurations of storage, network and some other resources are being tracked by configuration monitor.

In the Fig 1, the simplest view of hybrid cloud is provided, a single off-premises public cloud and on-premises private cloud is within the Enterprise Datacenter is shown and public cloud establishes the safe connection to store data on to the cloud is indicated by the arrow:

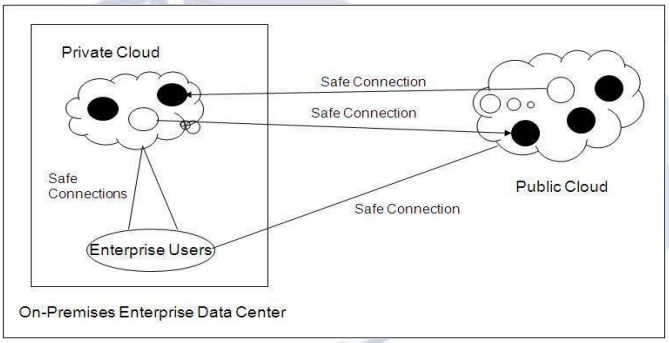

Fig 1.Example of Hybrid cloud Environment

The black circles shows active virtual server images and white circles shows virtual server images which have been migrated by using safe connections. The arrows indicate that the direction of migration. Using safe connections Enterprise users are connected to the clouds, which can be secure HTTP browsers or virtual private networks(VPNs) .A hybrid cloud could also can consist of multiple public or/and private clouds.

A. Integration

One or more private and public clouds integrate to form a hybrid system and it will be more challenging compared to integrating the on premises systems noted Rex Wang. As different clouds usually will have distinct APIs, private, integrating public and legacy systems will often require custom code, said UT Dallas' Kantarcioglu

B. Models

There are mainly two primary hybrid cloud deployment models.

C. Management

Hybrid cloud computing has a technological key called as management. Systems are quickly migrating from single cloud environment to multiple cloud management system. Then later they must manage all types of cloud applications such as platform as a service, infrastructure as a service and software as a service through the whole development and also deployment life cycles.

D. Security

In order to secure hybrid clouds, companies use special techniques such as authentication, access control policies and encryption in both private clouds and public clouds.

These will include the combination of cloud-based security services and managed hosted appliances. Some approaches such as intrusion detection systems and firewalls are always implemented in the hosted environment specifically for use of hybrid cloud architecture, said Jonathan Hogue. ―Since we cannot disclose the sensitive data, companies will requirein SIS implementations. Using sub-file de-duplication, redundant copies of data are detected and are eliminated—even after the duplicated data exist, within separate files. This form of de-duplication discovers the unique data elements within an organization and detects when these elements are used within other files. As a result, sub-file de-duplication eliminates the storage of duplicate data across an organization. Sub-file data de-duplication has tremendous benefits even where files are not identical, but have data elements that are already recognized somewhere in the organization.

Sub-file de-duplication implementation has two forms. **Fixed-length** sub-file de-duplication uses an arbitrary fixed length of data to search for the duplicate data within the files. Although simple in design, fixed-length segments miss many opportunities to discover redundant sub-file data.

(Consider the case where an addition of a person's name is added to a document's title page—the whole content of the document will shift, causing the failure of the de-duplication tool to detect equivalencies). **Variable-length** implementations are usually not locked to any of arbitrary segment length. Variable-length implementations match data segment sizes to the naturally occurring duplication within files, vastly increasing the overall de-duplication ratio (In the example above, variable-length de-duplication will catch all duplicate segments in the document, no matter where the changes occur). keeping limit for the amount of sensitive data which they outsource or they will have to encrypt the sensitive data before outsourcing in public clouds, ―kantarcioglu explained. Encryption based approaches will protect sensitive data when it outsourced to public cloud processing this encrypted data is usually more costly and complex

IV. A Detailed Look at Data De-Duplication

Data de-duplication has many forms. Typically, there is no one best way to implement data de-duplication across an entire an organization. Instead, to maximize the benefits, organizations may deploy more than one de-duplication strategy. It is very essential to understand the backup and backup challenges, when selecting de-duplication as a solution.

Data de-duplication has mainly three forms. Although definitions vary, some forms of data de-duplication, such as **compression**, have been around for decades. Lately, **single-instance storage** has enabled the removal of redundant filesfrom storage environments such as archives. Most recently, we have seen the introduction of **sub-file de-duplication**. These three types of data de-duplication are described below

A. Data Compression

Data compression is a method of reducing the size of files. Data compression works within a file to identify and remove empty space that appears as repetitive patterns. This form of data de-duplication is local to the file and does not take into consideration other files and data segments within those files. Data compression has been available for many years, but being isolated to each particular file, the benefits are limited when comparing data compression to other forms of de-duplication. For example, data compression will not be effective in recognizing and eliminating duplicate files, but will independently compress each of the files.

B. Single-Instance Storage

Removing multiple copies of any file is one form of the de-duplication. Single-instance storage (SIS) environments are able to detect and remove redundant copies of identical files. After a file is stored in a single-instance storage system than, all the other references to same file, will refer to the original, single copy. Single-instance storage systems compare the content of files to determine if the incoming file is identical to an existing file in the storage system. Content-addressed storage is typically equipped with single-instance storage functionality.

While file-level de-duplication avoids storing files that are a duplicate of another file, many files that are considered unique by single-instance storage measurement may have a tremendous amount of redundancy within the files or between files. Forexample, it would only take one small element (e.g., a new date inserted into the title slide of a presentation) for single-instance storage to regard two large files as being different and requiring them to be stored without further de-duplication

C. Sub-file De-Duplication

Sub-file de-duplication detects redundant data within and across files as opposed to finding identical files as in SIS implementations. Using sub-file de-duplication, redundant copies of data are detected and are eliminated—even after the duplicated data exist, within separate files. This form of de-duplication discovers the unique data elements within an organization and detects when these elements are used within other files. As a result, sub-file de-duplication eliminates the storage of duplicate data across an organization. Sub-file data de-duplication has tremendous benefits even where files are not identical, but have data elements that are already recognized somewhere in the organization. Sub-file de-duplication implementation has two forms. **Fixed-length** sub-file de-duplication uses an arbitrary fixed length of data to search for the duplicate data within the files. Although simple in design, fixed-length segments miss many opportunities to discover redundant sub-file data. (Consider the case where an addition of a person's name is added to a document's title page—the whole content of the document will shift, causing the failure of the de-duplication tool to detect equivalencies). **Variable-length** implementations are usually not locked to any of arbitrary segment

length. Variable-length implementations match data segment sizes to the naturally occurring duplication within files, vastly increasing the overall de-duplication ratio (In the example above, variable-length de-duplication will catch all duplicate segments in the document, no matter where the changes occur).

So most of the organizations widely use data depulication technology, which is also called as, single-instance storage, intelligent compression, and capacity optimized storage and data reduction.

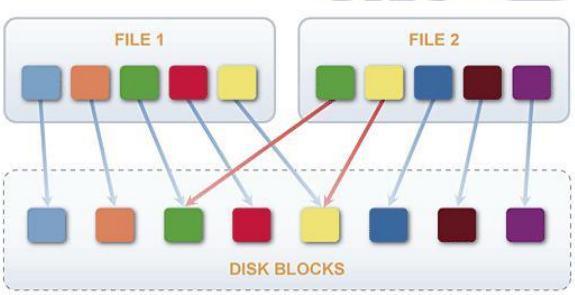

Fig .2 Example of De-duplication

Fig.2 shows, de-duplication finds the redundant data and eliminates all but keeps one copy which creates logical pointers to the file so that users could access the file as and when needed.

Pointers and References.

De-duplication systems removes redundant data that they find then creates a reference or logical pointer to single instance of data that host keeps. There are pointers at places where multiple users store the same single piece of information.

V. DATA DE-DUPLICATIONS TYPES

A. File-level de-duplication

It is commonly known as single-instance storage, file-level data de-duplication compares a file that has to be archived or backup that has already been stored by checking all its attributes against the index.

The index is updated and stored only if the file is unique, if not than only a pointer to the existing file that is stored references. Only the single instance of file is saved in the result and relevant copies are replaced by "stub" which points to the original file.

B. Block-level de-duplication

Block-level data de-duplication operates on the basis of sub-file level. As the name implies, that the file is being broken into segments blocks or chunks that will be examined for previously stored information vs redundancy.

The popular approach to determine redundant data is by assigning identifier to chunk of data, by using hash algorithm for example – it generates a unique ID to that particular block. The particular unique Id will be compared with the central index. In case the ID is already present, then it represents that before only the data is processed and stored before .Therefore only a pointer reference is saved to the previously stored data. If the ID is new and does not exist, then that block is unique. The unique chunk is stored and the unique ID is updated in the Index.

The size of the chunk which needs to be checked varies from vendor to vendor. Some will have fixed block sizes, while some others use variable block sizes likewise few may also change the size of fixed block size for sake of confusing. Block sizes of fixed size may vary from 8KB to 64KB but the main difference with it is the smaller the chunk, than it will be likely to have opportunity to identify it as the duplicate data. If less data is stored than it obviously means greater reductions in the data that is stored. The only major issue by using fixed size blocks is that in case if the file is modified and the de-duplication result uses the same previously inspected result than there will be chance of not identifying the same redundant data segment, as the blocks in the file would be moved or changed, than they will shift downstream from change, by offsetting the rest of comparisons.

Variable block level de-duplication compares varying sizes of data blocks that can reduce the chances of collision, stated Datalinks's Orlandini.

VI. WHEN DEDUPLICATION OCCURS?

A. Inline de-duplication

Inline de-duplication is the most economic and efficient method of de-duplication. It reduces the raw disk space needed in system, since the full, not still de-duplicated data set would never be written to disk. Inline de-duplication reduces time to disaster recovery readiness because the system does not need to wait to utilize the entire data set and before it begins duplication of data at the remote side, it is de-duplicated.

B. Post-process de-duplication

Post-process de-duplication refers to the type of system where software processes, filters the redundant data from a data set only after it has already been transferred to a data stored location. This is also called as asynchronous de-duplication, and it is usually considered in the situations when managers consider it unfeasible or inefficient to

remove the duplicate data during transfer or before the data is sent to storage location.

C. Client-side de-duplication

Client-side de-duplication is different from all other forms of de-duplication in that duplicate data is first only identified before it has to be sent over the network. This will definitely create burden on the CPU but at the same time reduces the load on the network. Leveraging client side de-duplication gives us lot of advantages, because of the high level of duplicate information in a virtual environment and also the fact that data is sent across a highly congested IP network.

D. Target-based de-duplication

Target de-duplication will remove the redundancies from a backup transmission as and when it passes through an appliance that is present between the source and the target. Unlike source de-duplication, the target de-duplication does not reduce the total amount of data that need to be transferred across a WAN or LAN during the backup, but it reduces the amount of storage space required.

E. Global de-duplication

Global data de-duplication is a procedure of eliminating redundant data when backing up data to more number of de-duplication devices. This situation might require backing up data to more than one target de-duplication system or in the case of source de-duplication it might require backing up to multiple backup nodes which will be themselves be backing up multiple clients.

VII. WHERE DOES DATA DEDUPLICATION OCCUR?

Some technology considerations need to be made when determining optimal de-duplication solutions for using in organizations. Considerations mainly include whether de-duplication must occur at the information backup target or source. Additionally, you should consider the appropriateness of immediate de-duplication or scheduled de-duplication architectures for your backup environment. This section further describes these features.

A. Source-based De-duplication

In the Source based De-duplication the elimination of the redundant data happens at the source. This means that the procedure of data de-duplication is performed at the beginning of the backup processes, before the data is transferred to the backup environment. Source based de-duplication will drastically reduce the huge amount of backup data that would be indeed sent through the networks during the backup process. So there will be substantial reduction in the capacity that is required to store the backup files.

B. Target-based De-Duplication

An alternative to source-based de-duplication is the target-based de-duplication. Target-based de-duplication is performed at the backup storage device. The users need not change their incumbent backup software usually in this type of de-duplication. In Target based de-duplication it is required that all backup files are copied to the backup systems, so in case of target based backup it will not provide us solution that will reduce backup-client-to-target bandwidth requirements.

VIII. HOW DE-DUPLICATION TECHNOLOGY WORKS: STORE ONLY UNIQUE DATA IN A DATA BASE

Data de-duplication compares the data i.e. usually blocks or files and eliminate the redundant data copies that are already present in datasets. It removes the files that are not unique. The process consists of following steps

a. Firstly divide the input data into chunks or blocks.
b. Hash value for each of the block need to be calculated.
c. The values that are got is used to determine whether the blocks of same data is already stored.

Replace the redundant data with the reference or pointers to the block that is already in database.

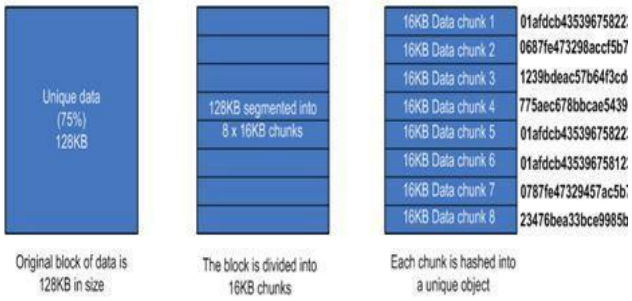

Fig 3. Division of Blocks according to chunks

After the data is chunked, an index is created from the results, and the redundant data can be found and eliminated. Only a one copy of every chunk is stored.

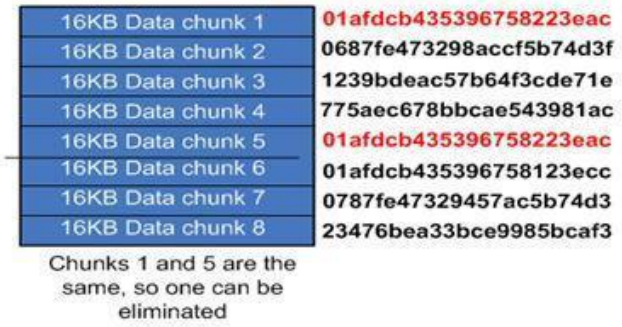

Fig 4. Hash number generation for the data chunks.

Once the data is divided into chunks, by the results that is obtained index can be created and the redundant data that will be found is removed. Only a single copy of every chunk will be stored. Data de-duplication process can be implemented in the number of different ways. The duplicate data can be eliminated b y simply comparing the files with each other and removing the data that is no longer needed and old.

Hash collisions are potential problem with de-duplication. The piece of data when it receives the hash number, the number is compared with the index of other already existing hash numbers.[8] Some of the algorithms can be used to find the hash numbers for example, SHA, MD[9] than its is encrypted using the AES algorithm and then data is out sourced to the cloud environment.

SHA-1 creates the required cryptographic signatures for security applications. A 160-bit value of SHA-1 creates that isunique for each piece of data.

MD5 -is also designed for cryptographic purposes it creates a 128-bit hash.

Hash collisions usually will occur when two different chunks will produce the same hash value. The chances of this arevery less indeed, but SHA-1 is considered to be most secure of the two algorithms that are above said.

A. Bit-level comparison

The simplest way of comparing the two chunks of data is, by performing, the bit-level by comparing two blocks. The cost that is involved in performing the I/O is required to read and compare them.

B. Custom methods

Here their own hash algorithm is been combined with other methods to identify duplicate data is used by some Vendors.

The hash number after comparing is found to be already in the index, than that piece of data will be considered as the duplicate, and need not be stored again. Otherwise the newly got hash number will be added to the index and the new data is stored. In very rare situations it can be found that for two different chunks of data the same hash number is obtained. Whenever the hash collision will occur, the system will not store the new data because it finds that its hash number is already present in the index. This will give in the false result and will in turn result in data loss. So some of the data vendor will combine hash algorithms to reduce the probability of hash collision. Some vendors will also examine metadata to find and identify data and prevent collisions.

Advantages of De-duplication

De-duplication promises that companies can store many times the data per storage, than before.

- Effectively increased network bandwidth - In case De-duplication takes place at the source end than no copies of data need to be transmitted via the network.
- —Greener‖ environment - Fewer cubic feet of space is required to store the data in both primary and remote locations and less electricity is needed.
- Line-of-business processes continued unimpeded which ensures faster recoveries.
- Buying and maintaining less storage will return us with the faster returns.
- Smaller amount of space is required to keep pointers to the backup data instead of storing the data copy itself.

Cost of storage is less for overall data – As we are storing less.

De-Duplication Downside

De-duplication System are expensive to implement, maintain and purchase said Rob Sims, CEO of Crossroads Systems.

In addition, Companies need to have more data to de-duplicate for saving more money, than they usually did with basic compression technique. Execution of Hashing algorithms and by comparing hashes makes de-duplication to use lot of power for processing and Energy he stated. Most of the companies are using de-duplication in small appliances that can handle up to 100 terabytes data according to Data link's Orlandini, this is not enough for large appliances.

Organizations can maintain performance and increase capacity by using many appliances it can only be used if machines which support clustering,

databases, work with same hash tables said Sims. As many de-duplication system won't offer redundancy, incase large appliance crashes, storage array that is working with it becomes unavailable temporarily to user's Orlandini said.

In addition, Companies need to have more data to de-duplicate for saving more money, than they usually did with basic compression technique. Execution of Hashing algorithms and by comparing hashes makes de-duplication to use lot of power for processing and Energy he stated. Most of the companies are using de-duplication in small appliances that can handle up to 100 terabytes data according to Data link's Orlandini, this is not enough for large appliances.

Organizations can maintain performance and increase capacity by using many appliances it can only be used if machines which support clustering, databases, work with same hash tables said Sims. As many de-duplication system won't offer redundancy, incase large appliance crashes, storage array that is working with it becomes unavailable temporarily to user's Orlandini said.

Security is also threatened by technology because users cannot encrypt data which they have to de-duplicate. Encryption will prevent systems from accurately identifying and reading the stored information that is for de-duplication, stated Curtis Preston, an IT-infrastructure-services vendor.

B. Data integrity

By breaking data into blocks, de-duplication removes the boundaries that separate all data groups. This creates problems for organizations that comply with government related regulations that require companies to separately keep types of different financial records, he elaborated. Since data is broken into blocks, reassembled and de -duplicated he noted, —Lawyers will be lining up with security & integrity questions when a company need to prove that data produced is the actually stored data.

Technology related industries such as pharmaceuticals, telecommunications and financial services have already adopted de-duplication, explained Scott Gidley, chief officer with a data-management, Data flux and integration vendor. However technology can consume lot of resources for processing, energy, and also costly is not suited to all end users.

Nevertheless de-duplication will definitely become common feature same like compression in coming five years if, it will be less costly, predicted David Russell, vice president for strategies and storage technologies with market research firm Gartner Inc. He stated, —It is no longer technology that is emerging but it is something that is found to be in early mainstream stage. The economics are more compelling to ignore.

How Does Encryption Affect Data De-Duplication

De-duplication works by removing the redundant blocks, files or data and encryption turns data into the data stream which is random by its nature. So if you encrypt data first that is random, it is impossible to de-duplicate it. So the data must be de-duplicated first and later encrypted.

Applications

Hybrid clouds are mainly built to suit any of the IT environment or architecture, whether it might be any enterprise wide IT network or any department. Public data which is stored can be analysed from statistical analyses which is done by social media, government entities can be used to enhance and analyse their own corporate data stand which is internal to gain the most form of perusing hybrid cloud Benefits. But analysis of big data and high performance computing that is involved between clouds is challenging.

IX. CONCLUSION

Using Hybrid cloud architecture for the IT Industries provides lot of benefits with the use of both public and private clouds and adopting de-duplication to store data in the cloud will provide us better storage benefits at lower costs. Hybrid cloud use is big data processing. A company, for example, could use hybrid cloud storage to retain its accumulated business, sales, test and other data, and then run analytical queries in the public cloud, which can scale to support demanding distributed computing tasks.

REFERENCES

[1] Neal Leavitt, "Hybrid Clouds Move to the Forefront. Published by the IEEE Computer Society, MAY 2013.

[2] Danny Harnik, Benny Pinkas, Alexandra Shulman-Peleg "Side Channels in Cloud Services Deduplication in Cloud Storage. COPUBLISHED BY THE IEEE COMPUTER AND RELIABILITY SOCIETIES, NOVEMBER/DECEMBER 2010.

[3] http://searchcloudcomputing.techtarget.com/tutorial/Hybrid-cloud-computing-explained

[4] https://education.emc.com/academicalliance/documents/EAA_Content/Exercises/An%20EMC%

[5] David Geer, "Reducing the Storage Burden via Data Deduplication.computer.org , December 2008.
[6] https://www.daniweb.com/images/attachments/0/WP_Deduplication_US_Letter_090702.pdf
[7] http://en.wikipedia.org/wiki/Data_deduplication#Deduplication_overview
[8] http://www.computerworld.com/article/2474479/data-center/data-deduplication-in-the-cloud-explained--part-one.html
[9] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", IEEE Transactions on Parallel and Distributed Systems, Volume: PP, Issue: 99, Date of Publication :18.April.2014
[10] Yang Zhang, Yongwei Wu and Guangwen Yang, Droplet: a Distributed Solution of Data De-duplication, ACM/IEEE 13th International Conference on Grid Computing,2012
[11] http://www.computerweekly.com/report/Data-deduplication-technology-review
[12] Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and Akhilesh Saxena, Intelligent Workload Factoring for A Hybrid Cloud Computing Model , Published by the IEEE Computer Society ,2009
[13] Borja Sotomayor, Rubén S. Montero and Ignacio M. Llorente, Ian Foster, Virtual Infrastructure Management in Private and Hybrid Clouds, Published by the IEEE Computer Society, 200.
[14] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, Duplicate Record Detection: A Survey, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 1, JANUARY 2007
[15] Srivatsa maddodi, Girija V. Attigeri, Dr karunakar A.k, Data Deduplication Techniques and Analysis. Third International Conference on Emerging Trends in Engineering and Technology IEEE, 2010