ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

Cloud Storage problems, benefits and solutions provided by Data De-duplication

Akingbade L.O.
Department of Computer Engineering, Federal Polytechnic Ilaro, Ogun State

Abstract – Storage of data on cloud is a service where data is remotely maintained, managed, and backed up. The service is available to users over a network, which is usually the internet. Instead of storing the data on your computer's hard drive or other local storage device, it allows the user to store data online so that the user can access them from any location via the internet. The cloud provider company makes them available to the user online by keeping the uploaded files on an external server. The internet provides the connection between your computer and the database. Cloud storage reduces hardware and software demands on the user's side, This gives companies using cloud storage services ease and convenience, but can potentially be costly. Users should also be aware that backing up their data is still required when using cloud storage services, because recovering data from cloud storage is much slower than local backup whereby you save it to a remote database. Accessing and storing data on cloud storage has several challenges and security risks, thereby making the cloud information inconsistent and unreliable. *Thus the* ***need for improvement and another approach/dimension into computing arises. This paper identifies Data de-duplication as a solution to the challenges faced by users and providers in cloud storage, some good features and benefits are discussed as boosters for efficient cloud storage.***

Keywords - **Cloud computing, data storage, De-duplication, challenges, solutions.**

I. INTRODUCTION

Cloud computing is an evolving term that describes the development of many existing technologies and approaches to computing into something different. Cloud separates application and information resources from the underlying infrastructure, and the mechanisms used to deliver them. Cloud enhances collaboration, agility, scaling, and availability, and provides the potential for cost reduction through optimized and efficient computing. Organizations are using cloud computing (hereinafter: the cloud) to perform increasingly strategic and mission critical functions. At the same time, companies are facing pressures and challenges to protect information assets belonging to their customers and other sensitive data (McCafferty, 2010). Unsurprisingly security, privacy and availability are among the topmost concerns in their cloud adoption decisions rather than the total cost of ownership (Brodkin 2010). Despite some serious privacy related drawbacks, cloud computing is a lucrative choice to improve productivity in any business environment, where IT is in high demand. To raise the security and privacy of cloud service providers, there need to be more co-operations between world governments so as we can develop a unified global rules and guidance for running a safe cloud computing service. Since users may not retain a local copy of outsourced data, there exist various incentives for cloud service providers (CSP) to behave unfaithfully towards the cloud users regarding the status of their outsourced data. (Wang, 2011).

There are three kinds of cloud storage management to choose from: Public cloud storage, Private cloud storage, and Hybrid cloud storage (Cloud Storage, 2011). The public could is best used for unstructured data, the private could is used by users with more need for customization and more control over their data, the hybrid could is a combination of the private and public cloud for the users that have a need for both types (Cloud Storage, 2011). An effective cloud storage system provides data security and protection, recovery, data lifecycle management, optimized storage solutions and they can be easily provisioned. Any user can customize cloud storage, whether it would be for a single user, small company or large company. The company would need to know exactly what the needs of the business are in order to assess the best use for cloud storage. Cloud storage has multiple benefits and some drawbacks (Booth, 2013).

The purpose of this paper is to provide you with an analysis of the cloud storage and data de-duplication from its begging to what the future holds for this new way to store data. However, a technique called Data de-duplication is introduced into this paper as an efficient method that helps both users and cloud providers to manage and make good use of available space in the cloud for themselves and other users. It also encourages the assurances of cloud data integrity, availability and efficient management.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

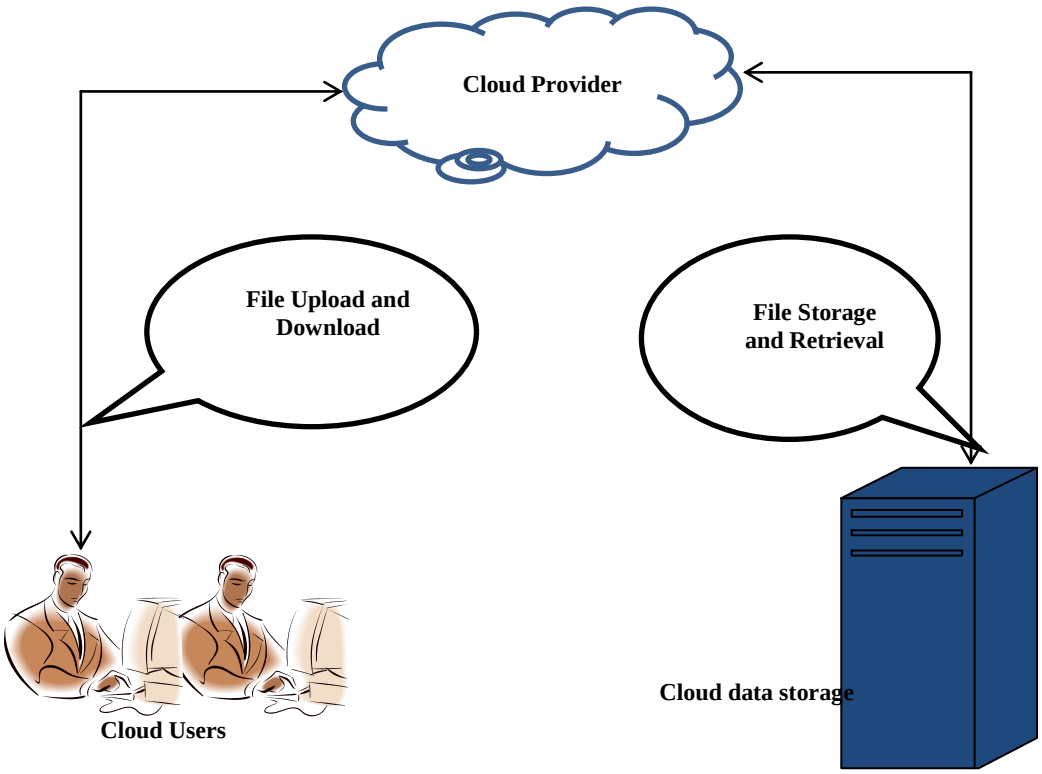

Fig. 1 General Cloud Data Storage Architecture

II. CONCEPT OF CLOUD STORAGE

Cloud storage is achieved through following the concepts of redundancy and repetition. Without these concepts, cloud storage would be very difficult, if not impossible to exist. Redundancy is really the core of cloud storage. Cloud storage at its basic level is just backing up data enough times so that the chance of losing that data becomes nearly irrelevant. (Strickland,2011)Having multiple data servers to store data decreases the chances of losing data. Cloud storage saves or reduces organization's IT expenditure. It also speeds up the development of project. It also raises the need for customizing business process of large organization to take advantage of cloud computing. A single data server store data is good, but ten data servers is a lot better. Data centers are known to house several, even hundreds of data servers. Along with multiple data servers come multiple power supplies. Having all data servers on one power supply would counteract the use of having multiple servers. If one power supply were to power all the servers in a network, and for some reason it went offline, all the servers on that network would go down, rendering the entire network inaccessible. To combat this issue, servers are divided into groups and each is given their own power supply. By doing so, you lessen the chance of all your equipment going offline. Cloud storage system based on distributed system has came into existence like Big Tabel, Casandra, Amazon (David,2013). Having multiples of equipment solves half the problem of cloud storage. The other half is what to do to with all the user data.

The cloud storage concept is a great new technological advancement that will help customers safely and reliably store their data. With multiple data servers and multiple copies of information, companies are able to hedge their bets against the loss of data and server downtime. Companies will no longer need to house their own special equipment for their own data. They can outsource this process to cloud storage businesses and save money and time.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

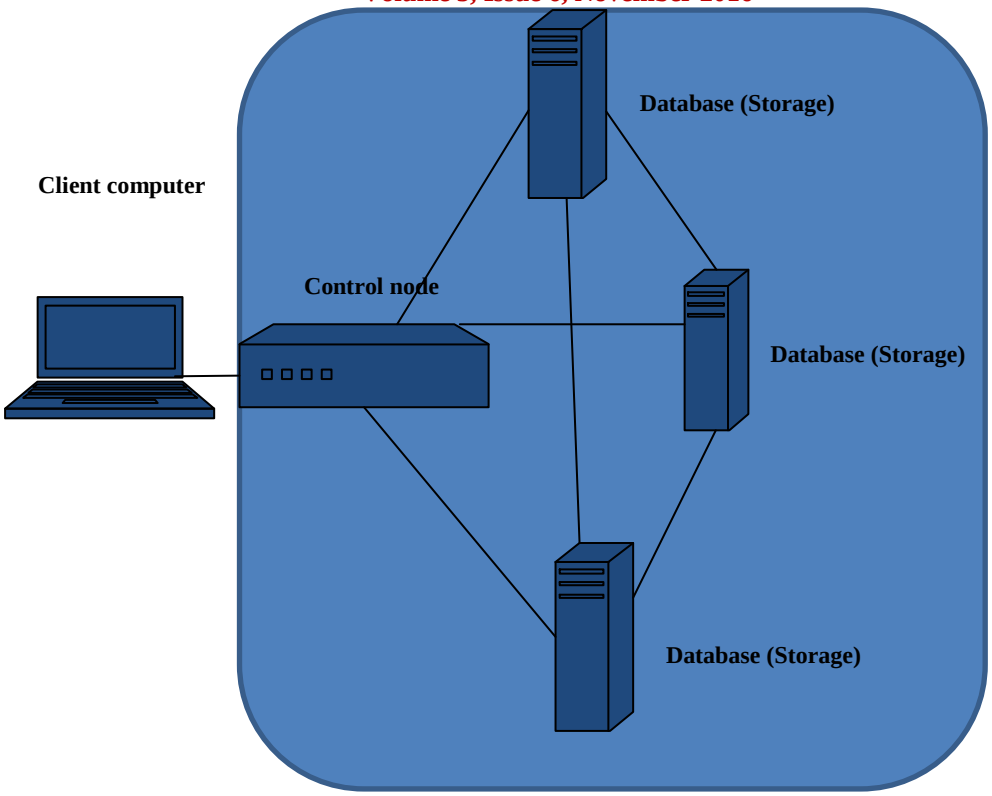

Fig 2. Concept of Cloud

III. CLASSIFICATION OF CLOUD STORAGE

Cloud based system can be easily understood as the composition of infrastructure and service. Here infrastructure mainly consists of hard disk, secondary storage which is virtualized before they are made available to user. There are various services developed by the cloud service provider according to the need of their clients' .Authors have suggested dividing the complete service into three categories (Comsa,2011). They are storage provider who take lot of cloud based storage service provider and provide service to general user. Cloud based service can be deployed in three different model. They are public cloud, private cloud and hybrid cloud(cloud storage,2011). Authors have found out why public cloud has certain limitations for business application. It cannot be said with certainty that public cloud is good for storage purpose. With the increasing information that we humans access , the storage requirements to accumulate that information keeps on increasing , it may be a collection of doc files, movies, software's or some songs, the data requirements keeps on increasing day by day..

Fig 3 Contents of Cloud Storage

Cloud storage can be classified into four types

- **Personal / Mobile Cloud Storage** : This is a type of cloud of type of cloud storage that we use in our daily lifes, We use android and I-phone these days, for our safety we have already synced our phones online so that even if our phone is lost , we can backup it on the new device anytime as per our convenience.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

- **Publicly Available Cloud Storage**: In this type, the user uses a public-ally available cloud which he has either rented or subscribed for a certain period of time. Anyone with access to that cloud with the User's log in credentials could access that data from any part of the world.
- **Privately Available cloud storage**: The Company providing cloud storage services sets up the data center in the user's allocated space. The main reason behind having privately available cloud storage is the requirement of a secure platform and restricted access to data.
- **A hybrid available cloud storage:** In this case the data is available public-ally but some parts of the data is private and hence is restricted to some users only. So this is an On and off model where it can be switched from public to private or private to public anytime as per user's convenience.

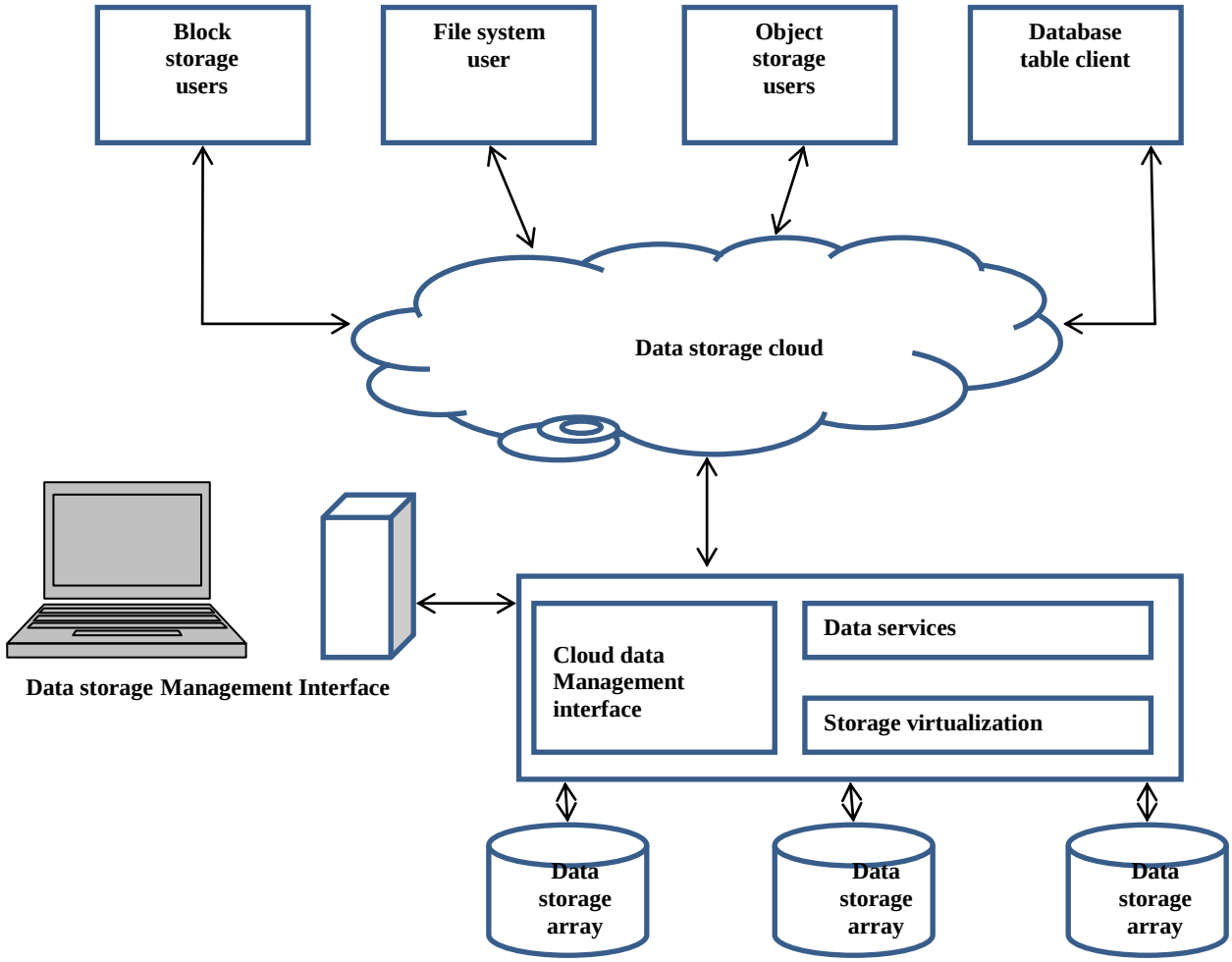

Fig 4 Components of Cloud Storage system

IV. BENEFITS OF CLOUD STORAGE

Cloud storage services are scalable, user friendly, saves bandwidth and accessible across the globe. They can provide zero infrastructure cost implementation of projects. Cloud storage services can be used for diverse purpose like sharing of data, backup, testing of project at reduced cost, database as a service and data.(Li, 2010). The first benefit is instant automatic backup. In short, it functions similar to the Apple's iCloud service. The data is sent to the cloud for storage back up, eliminating the need to save the data in CDs, external hard drives, or servers. The second benefit is security and protection from theft or natural disaster (Comsa, 2011). When these things happened the lost of data is one of the biggest concern for the company and it can potentially bring a company to its knees. For instance, when a company loses data about

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

their customers, such as the customers address, it can be disastrous for the company. The customers address can be use for billing and or shipping, and when a company does not know where to ship or send bills, the company loses revenue. The company can potentially ship order to the wrong address or not sent at all. This type data lost can cost to company customers, time, and money. During natural disaster, a company can be out of power for more than three days. According to article The Cost of Lost Data, (David, 2013) it is estimated that "a company that experiences am computer outage lasting for more than 10 days will never fully recover financially and that 50 percent of companies suffering such a predicament will be out of business within 5 years". This report is alarming for any company, which is why the cloud storage has such a great value to the company. A company can be sure that their data is safe and secure in the cloud, even when they are not physically housed in the same location as the business. It is easier to recover physical property than the precious data. Another benefit mention is the cloud data can be easily accessed whenever or wherever the users are located. One of the greatest benefits to companies when using the cloud is scalability, meaning the ability to expand to meet future needs. With the cloud, the company can increase user without increase their cost for their data storage. As the company grows, users needed to store more and more data, more severs are needed, more electricity needed to keep those servers cool. With the cloud, costs are reduced for maintenance and management of the servers and there is no need to change or add to the infrastructure.

V. PROBLEMS OF CLOUD BASED STORAGE

Every organization's business process needs to be customized according to cloud storage's needs. The cost for using cloud storage is not clearly specified in SLA.In absence of internet connection cloud based storage service cannot be accessed. Security and privacy are also there. In (Chan,2012) authors have classified the security for different service delivery models of cloud like IaaS,PaaS and SaaS.Such classifications can help us in developing different framework for different service delivery model. In (Strickland, 2011) some of the vulnerabilities of cloud computing are mentioned below:

- **Multitenancy:** Cloud service provider stores the data of multiple users on the same device. This is known as multitenancy. It can be implemented in various ways e.g., creating multiple partition on the disk or using virtualization. It should be the responsibility of the service provider to ensure that data of two user does not mix up with each other. Moreover the data should be permanently deleted and no one should not be able to recover the data from the disk allocated to any user after they have stopped using the service
- **Privacy and Security:** Data stored in local or personal server is always considered safer than data stored on others server. Cloud based service brings dependency on internet. Privacy and availability of data must be guaranteed by cloud service provider. In (Cloud security model, 2011) authors have suggested dividing the data into two category i.e. sensitive and non-sensitive data. Users must encrypt the sensitive data before they upload it on the cloud. Privacy of data also get violated due to illegal data mining by cloud service provider.
- **Performance Unpredictability:** Performance issues related to data retrieval, storage, modification will always be compared before user moves to cloud based service from its local server. Overall performance may also decrease due to travelling of data from a network to another network or due to encryption and decryption of data. Virtualized server and processor sharing give degraded network performance (Brodkin, 2010).They can cause abnormal network delay.
- **Portability and Interoperability:** Cloud based storage faces lock in issues where the user finds it difficult or impossible to move their data from one vendor to another vendor. This task may get more complicated due to different format of data storage, encryption or decryption of data. Deletion of data after it user has moved to different vendor must also be guaranteed.
- **Vulnerability due to mobile cloud computing:** Lots of user access service with mobile phone or smart phone. This raises lot of challenges like network accessibility, network latency due to wireless connection, change in mobile phone, loss of confidentiality due to loss of mobile phone etc. It is easy for hackers to attack vendor's server with virus and malware from mobile system. Along with these threats security issues regarding data confidentiality, availability, and integrity also exist (Aronika, 2012).
- **Data Confidentiality:** Cloud services are virtualized. There are several possible ways to compromise a virtual machine. and compromise the confidentiality of the data stored on them. Attackers can use cache details, system clock to compromise a virtual machine. There are several ways to prevent it like

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

using symmetric key, access control list for easy resource. Symmetric key if shared with cloud service provider would prove it ineffective.(cloud security, 2011)

- **Data Availability and Integrity:** Cloud service provider must guarantee availability and integrity of data stored on their server. There are cryptographic mechanisms for ensuring it. Uptime of service should be well mentioned in SLA.

VI. DATA DEDUPLICATION

In computing, data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage. (Biggar, 2011) .This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. (Chow,2009) Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced.

VII. FEATURES OF DATA DE-DUPLICATION

The Data Deduplication consists of a filter driver that monitors local or remote I/O and a deduplication service that controls the three types of jobs that are available (Optimization, Garbage Collection, and Scrubbing). Inherent in the deduplication architecture is resiliency during hardware failures—with full checksum validation on data and metadata, including redundancy for metadata and the most accessed data chunks.((Subashini, 2012)

Data Deduplication can potentially process all of the data on a selected volume (except a file size less than 32 KB, files in folders that are excluded, or files that have age settings applied). You should carefully determine if a server and attached volumes are suitable candidates for deduplication prior to enabling the feature. We strongly recommend that during deduplication, you regularly back up important data. Data deduplication involves finding and removing duplication within data without compromising its fidelity or integrity.(Network,2011) The goal is to store more data in less space by segmenting files into small variable-sized chunks (32–128 KB), identifying duplicate chunks, and maintaining a single copy of each chunk. Redundant copies of the chunk are replaced by a reference to the single copy. The chunks are compressed and then organized into special container files in the System Volume Information folder.

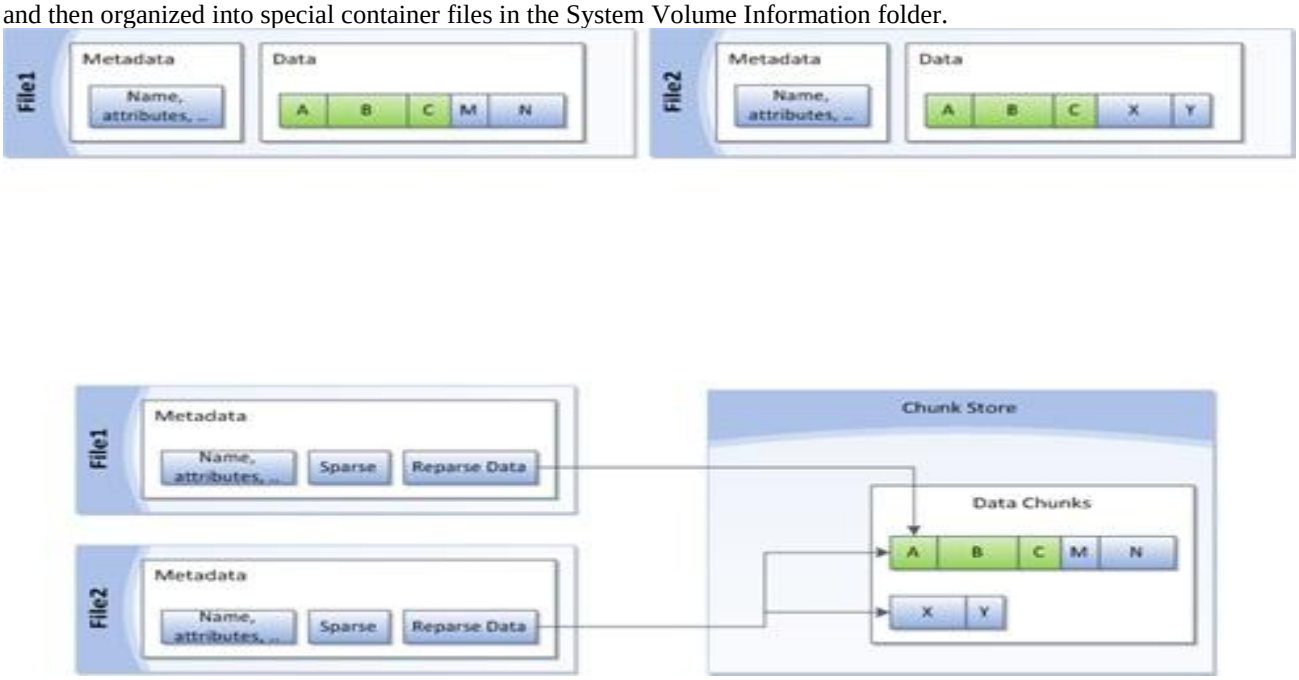

Fig 5. On-disk transformation of files during data deduplication

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

VIII. DATA DEDUPLICATION- CLOUD STORAGE SOLUTIONS OFFERED

To cope with data storage growth in the enterprise, administrators are consolidating servers and making capacity scaling and data optimization key goals. Data deduplication provides practical ways to achieve these goals(Understanding Data deduplication,2011), they include:

- **Capacity optimization.** Data deduplication stores more data in less physical space. It achieves greater storage efficiency than was possible by using features such as Single Instance Storage (SIS) or NTFS compression. Data deduplication uses sub file variable-size chunking and compression, which deliver optimization ratios of 2:1 for general file servers and up to 20:1 for virtualization data.
- **Scale and performance.** Data deduplication is highly scalable, resource efficient, and nonintrusive. It can process up to 50 MB per second in Windows Server 2012 R2, and about 20 MB of data per second in Windows Server 2012. It can run on multiple
- volumes simultaneously without affecting other workloads on the server. Low impact on the server workloads is maintained by throttling the CPU and memory resources that are consumed. If the server gets very busy, deduplication can stop completely. In addition, administrators have the flexibility to run data deduplication jobs at any time, set schedules for when data deduplication should run, and establish file selection policies.
- **Reliability and data integrity.** When data deduplication is applied, the integrity of the data is maintained. Data Deduplication uses checksum, consistency, and identity validation to ensure data integrity. For all metadata and the most frequently referenced data, data deduplication maintains redundancy to ensure that the data is recoverable in the event of data corruption.
- **Bandwidth efficiency with BranchCache.** Through integration with BranchCache, the same optimization techniques are applied to data transferred over the WAN to a branch office. The result is faster file download times and reduced bandwidth consumption.
- **Optimization management with familiar tools.** Data deduplication has optimization functionality built into Server Manager and Windows PowerShell. Default settings can provide savings immediately, or administrators can fine-tune the settings to see more gains.

IX. CONCLUSION

The cloud has helped change IT's role in the business. Cloud computing based storage when properly implemented will be the best form of cloud based services, it helps provide a new architecture to address the storage, management and analysis of fast growing machine-generated data, despite its critics and drawbacks it seems that Cloud Computing is here to stay. Having considered the risks and benefits of cloud storage, the solutions and pros offered by Data deduplication will definitely encourage more and more users and cloud providers to benefit a lot from cloud computing. Data storage helps in providing advanced scalability, manageability and the potential to collapse compute and storage together on the same processing nodes, cloud storage will guide in new levels of efficiency and economics into enterprise data centers. Data deduplication help storage system by providing IT administrators with the capability to proactively manage their cloud space and cloud environment, all with minimal effort and at a low cost. Future research will include a study regarding the level of adoption and the implementation of Data deduplication for Cloud data management.

REFERENCES

[1] Aronika.R and Paul Rajan.(2012) Evolution of Cloud Storage as Cloud Computing Infrastructure Service. IOSRJCE.1 (1), pp-38-45.

[2] Balakrishnan.S, and Karthikeyan.S (2011) Introducing Effective Third Party Auditing for Data Storage Security in Cloud. Proc. of IJCST Vol. 2, Issue 2.

[3] Booth .G and Soknacki,A. (2013).Cloud Security: Attacks and Current Defenses.8th annual Symposium on Information And Assurance.

[4] Brodkin, J. (2010). Problems with SaaS security. Network World, 27(18), 1-27.

[5] Chan, D. & Zhao, H., (2012). Data Security and Privacy Protection Issues in Cloud, Computing. Conference on Computer Science and Electronics Engineering. IEEE, pp. 647-651.

[6] Chow, R. (2009). Controlling Data in the Cloud: Outsourcing Computation without Outsourcing Control. In Proceedings of the 2009 ACM workshop on Cloud computing security. pp. 85-90.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

[7] Cloud Security Model. (2011). Retrieved from Cloud Computing Best Practices Networkhttp://BestPracticeCommunity.com Cloud Storage. (2011). Retrieved from Search Storage: http://searchcloudstorage.techtarget.com/definition/cloud-storage.

[8] Comsa, G. (2011). An Intro to Cloud Storage for Business - A few reasons cloud storage is a must-have tool for all businesses today. Cloud Computing Journal, online.

[9] David M. and Smith, P. (2013). The Cost of Lost Data - The importance of investing in that "ounce of prevention". Los Angeles: Graziadio Business Review.

[10] IBM Software (2011) Effective storage management and data protection for cloud computing.

[11] Li,.Z. and Xiao Zhang (2010)Study on cloud storage system based on distributed storage System. ICCIS .

[12] McCafferty, D. (2010). Cloudy Skies: Public versus Private Option Still up In the Air.Baseline, 103, 28-33.

[13] Networks, T. (2011). "What You Should Consider Before Using Cloud Storage". Retrieved from thrive networks.com: http://thrivenetworks.com/blog/2011/08/25/whatyou-should-consider-before-using- cloud-storage.

[14] Strickland, J. (2011). "How Cloud Storage Work". Retrieved from HowStaffWoks.com: http://computer.howstuffworks.com/cloud- computing/cloudstorage3.htm.

[15] Subashini, S. and Kavitha, V (2012) A survey on security issues in service delivery models of cloud computing. Journal of Network and Computer Applications, 34(1), p.1-11. 2012.

[16] Understanding Data Deduplication (2013) Druva, 2009. Retrieved 2013-2-13.

[17] Wang, G. (2010). The impact of virtualization on network performance of Amazon. Data Center. In Proceedings - IEEE INFOCOM.