An Overview of Data Storage in Cloud Computing

Isaac Odun-Ayo
Covenant University
Ota, Nigeria
isaac.odun-ayo@covenantuniversity.edu.ng

Olasupo Ajayi
University of Lagos
Lagos, Nigeria
olaajayi@unilag.edu.ng

Boladele Akanle
Covenant University
Ota, Nigeria
bola.akanle@covenantuniversity.edu.ng

Ravin Ahuja
Delhi Technological University
Delhi, India
ravinahujade@gmail.com

***Abstract*—Cloud computing is a functional paradigm that is evolving and making IT utilization easier by the day for consumers. Cloud computing offers standardized applications to users online and in a manner that can be accessed regularly. Such applications can be accessed by as many persons as permitted within an organisation without bothering about the maintenance of such application. The Cloud also provides a channel to design and deploy user applications including its storage space and database without bothering about the underlying operating system. The application can run without consideration for on-premise infrastructure. Also, the Cloud makes massive storage available both for data and databases. Storage of data on the Cloud is one of the core activities in Cloud computing. Storage utilizes infrastructure spread across several geographical locations. Storage on the Cloud makes use of the internet, virtualization, encryption and others technologies to ensure security of data. This paper presents the state of the art from some literature available on Cloud storage. The study was executed by means of review of literature available on Cloud storage. It examines present trends in the area of Cloud storage and provides a guide for future research. The objective of this paper is to answer the question of what the current trend and development in Cloud storage is? The expected result at the end of this review is the identification of trends in Cloud storage, which can beneficial to prospective Cloud researches, users and even providers.**

Keywords-Cloud computing; Cloud Storage; databases; Cloud Infrastructure

I. INTRODUCTION

Cloud computing is defined by [1] as a parallel and distributed computing system consisting of a pool of inter-connected and virtualized computers that are dynamically provisioned and presented a single computing resource to the users based on pre-agreed Service Level Agreements (SLA). It enables users to remotely run their applications as well as store data with the benefit of an on-demand and highly available service; without the burden of local hardware and software management. With Cloud storage, data is stored on multiple third party servers, rather than on the dedicated server used in traditional networked data storage. Third party service providers are entrusted with users' data and for security purposes the exact storage locations of these data are unknown to most people. Cloud computing is positively impacting the IT landscape using the Internet as it enables users pay on a per services usage bases. User concerns are thus shifted from acquisition and maintenance to utilization of facilities made available by Cloud service providers. Cloud computing is about moving services, computation or data for cost and business advantages offsite to an internal or external, location transparent, centralized facilities or contractor [3].

Cloud computing has characteristics that include resource pooling and multi-tenancy [2]. There are three basics service types in Cloud computing: the Software-as-a-Service (SaaS), where applications are made available by Cloud Service Providers (CSPs) over the Internet to the Cloud users; Platform-as-a-Service (PaaS), wherein the CSPs offers the Cloud users platforms for development and deployment of their own applications; and Infrastructure-as-a-Service (IaaS), where the CSPs offers compute, storage, network and other computing resources to the Cloud users. The IaaS users have control over the operating system and applications running on them, while the provider manages the hardware infrastructure. These services are all made available to users anytime and from any location via the web.

Cloud computing also has four modes of deployment, the private Cloud, public Cloud, community Cloud and the hybrid Cloud. The private Cloud is owned and controlled by an individual organization. The facilities could on–promise or off–premise. Private Cloud allow for more secured environment due to internal staff utilization. Public Cloud is owned and managed by major CPSs. These providers own large data centres, sometime spread across different geographical locations. They provide various services that free the customer from expensive infrastructural procurements. Community Clouds belong to several organizations that come together based on shared common interest. The community Cloud may be managed by the community or a third party. Hybrid Cloud is a combination of either private, public or community Cloud. The hybrid Cloud share the same infrastructure but the organizations are unique.

A major component of Cloud computing is storage. Storage could be for an enterprise database or simple storage of data similar to storing information on a local hard drive. In Cloud storage, data is stored in multiple third party services rather than on dedicated servers used in traditional networked data storage [4]. When storing data, the customer "sees" a virtual server, hence it appear that data is stored in a particular place with a specific name, but such a place does not exist in reality. It is just a pseudonym used to reference a virtual space carved out in the Cloud. The users' data could be stored on any computer in any data centre across several geographical locations. The data's actual storage location may differ from time to time as the Cloud dynamically manages available storage locations around the data centres. Although, the data location is virtual, the user sees a static location for his data and can actually manage the storage space as if the user were on a personal computer [4]. A typical Cloud storage architecture includes a master control server and several storage servers as depicted in Fig. 1.

At the most basic level, a Cloud storage system needs just one data server connected to the Internet. A user sends copies of files over the Internet to the data server which then records the information. In the Cloud data storage system, users store their data in the Cloud and no longer have complete control over their data as they would if the data resides on a local computer [5]. Hence, the correctness, security and availability of data being stored on the Cloud server must always be guaranteed. The purpose of this paper therefore is to examine Cloud computing data storage. The paper discusses the Cloud storage architecture and various challenges facing Cloud storage. Thereafter, the events relating to Cloud data storage on the Cloud are highlighted. The rest of the paper is organized as follows: Section 2 examines related work. Section 3 discusses the Cloud data storage architecture. Section 4 highlights developments in the IT industry. Section 5 concludes the paper and suggests future work.

II. RELATED WORKS

In [3], opportunities and challenges of security in Cloud computing are discussed. The paper focused mainly on security concerns relating to the Cloud while also discussing various aspects of data storage. In [6], the authors presented an approach that classified data for security purposes was presented. The approach ensured data privacy and security in storage based on a classification patterns. In [7], the authors noted that the though use of Cloud storage frees users from the resources require for local storage, security of such data becomes a concern. The paper then proposed a storage security mechanism that was shown to be efficient against certain types of attacks but not completely full proof, as it was still vulnerable to other forms of attacks. In [8] another mechanism for security, privacy and access to data stored in the Cloud was presented. Similar to the work of [7], the proposed approach also has its own vulnerabilities. The authors in [9], presented a model similar to the RAID architecture for spreading the storage of data across multiple providers. The model called Redundant Array of Cloud Storage (RACS). The authors simulated the model and concluded that using RACS, users achieve about seven fold cost savings when switching CSPs. In [4], a model for Cloud Storage called High-Availability and Integrity Layer (HAIL) is proposed. The main focus is to ensure the integrity and easy retrieval of user data stored in the Cloud. The HAIL model allows for a secure and efficient Cloud data storage. In [5], the authors proposed a Cloud data storage system with emphasis on supports both for Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP) forms of data processing. The focus was on data freshness and redundancy in storage. In [10], the authors presented an approach for providing an efficient and secure way of sharing data amongst users in a Cloud storage system. The scheme allows secure access to shared data based on different user hierarchies. The authors in [11], presented a paper that discussed Cloud storage reference model and reviewed various challenges of Cloud storage. Like [11], the authors in [12] also presented a review of research works on Cloud storage architecture and related underlining technologies. In [13], the authors focused on guaranteeing the integrity of the user's data on the Cloud. The paper presented a method of performing optimized auditing of data storage security protocols in use in Cloud storage systems. The proposed model safeguards the user data from attacks when stored on Cloud servers. In [14], a survey of data storage security concerns in Cloud computing was presented. The trust of the paper was on identifying various challenges relating to data stored by the CSPs. The authors then presented and discussed possible solutions.

III. CLOUD STORAGE ISSUES

A. *Cloud Storage Architecture*

Cloud storage is composed of thousands of storage devices clustered by network, distributed files system and other storage middleware to provide Cloud storage for users [11]. Generally, storage could be in form of stand-alone arrays, converged infrastructure, hyper converged infrastructure, software-defined storage or public Cloud storage. Storage could also be block, file or object storage. The network infrastructure used in many of these storage systems includes but are not limited to fibre, iSCSI, NFS and SMB. These network infrastructure interconnect storage systems which could be NVMe-based, hybrid arrays, HCI, public Cloud for primary and backup, and storage for containers. The typical structure of Cloud storage includes storage resource pool, distributed file systems, service level agreements and service interfaces among others. A five-layer Cloud storage model comprising of the network and storage infrastructure layer, storage management layer, metadata management layer, storage overlay layer and service interface layer, as described in [7] is shown in Fig. 2; while the Cloud storage architecture [12] is depicted in Fig. 1.; This architecture is as shown in Fig. 1.

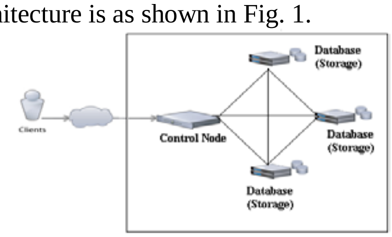

Figure. 1. Typical Cloud storage system architecture [adapted from 12]

Service Interface
Storage Overlay
Metadata Management
Storage Management
Network and Storage Infrastructure

Figure. 2. Cloud Storage Layered Model.

The various layers in the Cloud Storage Model depicted in Fig. 2 are described as follows:

- Network and storage infrastructure: consists of distributed wired and wireless networks interconnecting storage devices.
- Storage management: geographically distributed storage resources are organized by domains and logical entities. In addition, data can be stored by file or blocks in storage media.
- Metadata Management: clusters the global domain data storage metadata information and collaborates different domains for load balancing purposes.
- Storage Overlay: virtualization, service retrieving and redirecting are handles at this layer. A middleware can be used to links distributed data storage devices and then present them as a single and simplified virtual storage network to the users.
- Service Interface: provides clients with a uniform interface to access the Cloud storage system.

B. *Key Isses of Cloud Storage Services*

In Cloud computing, data is stored on multiple third party servers, rather than dedicated servers as used in traditional data centres. The following are issues relating to Cloud storage services as described in [10].

1) Deployment of Cloud Storage

The sale of Cloud storage should be based on application requirements and technology. The common storage networks are integrated by middleware and overlay layer. The geographical location should be selected by the data requirement application. The cost of storage should be optimized based on the deployment mode. Feedbacks from various servers and clients should be collected to adjust the distribution policies and access control.

2) Virtualization and Availability of Cloud Storage

Virtualization is applied to many domains including operating systems, servers, network and storage. Storage virtualization is meant to map logical storage to physical storage in data access procedure. The Cloud virtualization will help to hide storage locations and storage modes from the users. The availability of Cloud involves persistent runtime and recovery.

3) Data Organization

The organization of data in storage could be in database mode, file or block level. The database can be open source or proprietary. The database can only manage some specific data types. The block level is the lowest storage data format, both database and file utilize block level. Block level must be combined with other storage organization mode.

4) Data Migration and Load Balancing

Cloud data migration involves moving data from one storage location to another probably in different locations. The essence is to ensure local balancing in the Cloud storage system. When the storage capacity is used over certain values, the data should be migrated to other Cloud storage units, while keeping the pointers in the old storage position or modify and update the metadata at the same time. Local balancing is meant to keep available storage spaces for latter application in different storage devices on the Cloud. It can improve storage responsibility and availability. Data migration is one of the effective means for load balancing but may lead to bandwidth and I/O processes. Data replication is a type of migration where the original data is preserved. Data replication is a solution to the single point in distributed Cloud storage, which keeps multiple copies of the same content in different storage devices and locations. The ideal Cloud storage system should automatically create needed copies based on the user's access frequency and server workloads.

5) Data Deduplication

Data deduplications deals with storage, backup, recovery and archiving meant to reduce the space occupied in storage by compressing the internal duplication data. Data deduplication is the best way to reduce data volumes, slash storage requirements, minimize data protection cost and risks. In view of the experimental growth expected in data for enterprise and science, there will be need for massive storage, and data deduplication will help to save space and cost.

6) Cloud Storage Security

Cloud storage security involves storage media physical security and data security. Cloud storage and security involves certification, authority, audit and encryption among others. Cloud storage security also affects procedure of storage service, which includes software, hardware, data information, network security and user privacy.

C. *Cloud Storage Security Concerns*

Cloud computing does not provides control to users over the data stored in the Cloud data centres [14]. The Cloud providers have full control and they can perform task such as copying, destroying, modify etc. The lack of control by users, concepts of multi–tenancy and virtualization have high security risks associated with Cloud computing than information stored in traditional data centre. Four security issues are identified with Cloud storage [14].

1) Data Privacy and Integrity

Cloud computing is vulnerable to treats in the area of data integrity, confidentially, privacy and availability. Due to its simplicity, Cloud users are increasing exponentially and more applications are being hosted in the Cloud. A successful attack on any aspect of data in storage could lead to a breach that can grant unauthorized access to data of all Cloud users. Based on virtualization it is possible for data to be processed by multiple persons due to multi–tenancy. It is also possible for a malicious insider to breach data security during processing. There are is also the unanswered question of what exactly CSPs do with their users' data which they house in their data centres.

2) Data Recoverability and Vulnerability

Due to the elastic nature of the Cloud and other characteristics such as resource pooling and multi-tenancy, data can be breached on the Cloud. The resource allocated to a particular Cloud user may be assigned later to another user later. In terms of memory and storage, a malicious user can employ recovery techniques to obtain data from a previous user.

3) Improper Media Santization

The issue is related to the physical media destruction due to various reasons. There may be a need to change disk or the need to remove data from a disk. In addition, there may be need for termination of service. If the CSP does not sanitize the devices properly, it may be exposed to risks. Also, multi–tenancy contributes to the risk of device sanitization.

4) Data Backup

Data backup is also an issue that must be dealt with carefully. A regular backup is needed by the CSP to ensure the availability and recovery of data in case of intentional and accidental disasters. Moreover, backup needs to be protected against unauthorized access and tampering. There are several security models aimed at guaranteeing using data in storage. SecCloud [11] uses a storage security protocol that not only secures user data uploaded into the Cloud, but also secures computation performed on user data. In [9], a scheme is proposed that allows users to rate the requirement of confidentiality, availability and integrity on a scale of 1–10. The values are used to determine sensitivity rating of user's data and eventual protection. A solution was proposed in [11], based on–demand data correctness verification. The model conducts the verification of Cloud data correctness without explicit knowledge of whole data. It is also possible to encrypt data before outsourcing but there is a lot of overload if such data is to be shared.

5) Data Outage

Many customers wants reliable elastic and highly available storage online. Cloud providers compete on price, guarantees of uptime and availability in the form of SLAs. Cloud providers offer strong protection against component failures, so there is no compelling need to add another fail safe on top of Cloud storage systems. Despite all this, outages occur on the Cloud data centres that leads to loss of data for many Cloud users.

IV. CLOUD STORAGE TRENDS

A. Cloud Storage Projections

The 2016 Computer Weekly IT Priority Survey indicates that Cloud is top priority for IT department while storage and back up for virtualized environment are key issues [15]. According to the survey, it is projected that the total amount being spent on Cloud would overshadow that spent on in-house hardware and software in the not too distant future. From the survey, it was also projected that Cloud compute and storage services would experience a 50 % increase in spend while software and infrastructure would experience 48 % and 33 % increase respectively. It was also reported that about 30 % of correspondents plan to implement a Cloud storage system as a data backup option.

Virtualization has always been a major backbone for Cloud computing and recent surveys report that it would continue to be in the foreseeable future. The 2016 survey [14] projected that Server virtualization still remains a priority for many Cloud users with a 38 % positive feedback from respondent, it is followed by virtualization of data storage with 24 % and virtual desktop environment at 22%. The survey also shows that 23 % of respondent were planning to deploy Cloud virtual servers as backup systems.

B. Cloud Storage Appliances [15]

Cloud storage appliances have evolved to make Cloud a more practical proposition in work and office contexts. They act as translators and accelerators that will allow business systems to access private and public Cloud storage as if it were local storage. Cloud storage is bringing about less hardware to buy and manage, usage–based pricing and easy access from anywhere. However, what works well when storing smart phone photos is not same for enterprise data storage. It is one thing to use a web–based app that backs into Cloud storage, but quite another to use Cloud storage with enterprise applications, even ones that are as apparently as simply file-sharing. That is because most Cloud storage is object based and stateless, accessed via web–friendly APIs, whereas enterprise software is typically file or block–based, although this is changing with the amplification of enterprise. Unlikely legacy applications, in enterprise, web apps are usually designed to cope with the latency and bandwidth issues associated with connections over a wide area network such as Internet. A hardware gateway can help by including local storage as cache or buffer. This is especially useful in common use cases such as Cloud backup and archiving, where local caches can accelerate back up operations and access to online data. Some appliances are discussed in the following sub-sections.

1) On-Premise Gateway

In this model, an appliance (physical or virtual) sits on the premises and is connected on one side to the internal LAN and the other to the Cloud. It might take Cloud storage and present it to servers such as iSCSI block LUNs or as CIFS file – server volumes. These devices can also include local storage tier for certain data for performance reasons.

2) Cloud Controllers

As well as gateway capabilities, these devices aim to provide services similar to those offered by the traditional enterprise storage arrays, except that the data is stored in the Cloud. They add features such as data deduplication, compression and encryption and Cloud–based clones and snapshots.

3) Cloud Integrated Storage

These provide a higher degree of integration between Cloud and local storage. In this model, data is dynamically moved to the most appropriate tier based on policy. Hybrid Cloud storage arrays are now being developed that are deployed in-house but have built-in Cloud integration capabilities that enable them add and utilize storage tier located within the Cloud.

4) Cloud Resident Gateways

These are similar to Cloud integrated storage but resident in the Cloud as a virtual appliance, these serve applications that have been migrated to the Cloud. For example, Avere's CloudFusion gateway takes the different tier of Cloud storage available to it like Amazon EC2 RAM, solid state disk or bulk S3 storage and builds them into a virtual tiered network–attached storage filer. Some common examples of Cloud–resident gateway are as follows:

a. Amazon's AWS storage gateway: This sends only changed data as a means of saving bandwidth and allows primary data to stay on-premise.
b. Microsoft StorSimple: a hybrid local storage device with Cloud connectivity. It is designed to work as primary on–premises storage, while using Azure for Cloud – based archiving, backup and data recovery.
c. Barracuda Backup: This acts primarily as an on–premise backup system but includesdata deduplicating on to the Cloud.
d. Nasuni filers. This blends local disk and flash storage with Cloud storage, creating a Cloud – integrated unified storage system able to **s**erve block and file workloads.

C. IBM Cloud Object Storage.

Until recently, enterprise did not have too many options in terms of deploying high–performance object storage solution across both the Cloud and the on–premise data centres. A 2016 executive brief sponsored by IBM title "Which Cloud storage service delivers the performance you need? Comparing IBM Cloud object storage and Amazon S3", reported the introduction of IBM Cloud Object Storage [17]. The solution enabled users take control of their Cloud and provided options to choose the optimum between a plethora of various deployment models, costs options, and performance for each workload. A concise comparison of some of the features provided by the IBM system versus those of Amazon's Simple Storage Service (S3) is shown on Table 1

1) Throughput Results

When the environment were configured comparably, IBM Cloud object storage dedicated service delivered 1.9x higher "read" throughput and 3.3x higher "write" throughput than S3. The higher performance was attributed to a single–tenant architecture of IBM which maximizes the server resources available to the workload. In addition, the authors also reported that the write–optimized IBM Cloud object storage dedicated delivered 1.7x faster "read" and 9.9x faster "write" performance than S3.

2) Latency

In this test, the authors reported that both IBM and AWS systems were configured with similar settings and tested comparably using similar indexing configurations, applying a constant request rate of 420 requests per second. Latency results measured at the 95^{th} percentile indicating that IBM Cloud object storage dedicated service delivered a "read" latency which was an average of five times lower than Amazon's S3; while the "write" latencies were on an average 6.5 times lower than AWS's S3.

V. CONCLUSION

Cloud computing provides compute, storage and application services among others to users over the Internet. The resources made available to users by CSPs has reduced the need for expenditures on infrastructure. The Cloud is used for numerous activities but prominent among them are computation and storage. This paper focused on Cloud storage. A review of Cloud storage systems, architecture, models and challenges was done. A comparison of some of the storage features offered by two popular Cloud Storage Service Providers IBM and Amazon was also done. In conclusion it is important to note that despite certain Cloud challenges particularly in terms of security and privacy, Cloud storage is still being adopted at a tremendous rate; and research works are still on-going in a bid to further push the boundaries of Cloud storage adaptation

TABLE 1: COMPARATIVE PERFORMANCE OF IBM CLOUD OBJECT STORAGE VERSUS AMAZON S3

Feature	IBM Cloud Object Storage	Amazon S3
Single/multi-tenancy options	▪ Multi-tenant ▪ Single-tenant (Dedicated Service)	▪ Multi-tenant only
Deployment options	▪ On premise (appliance or licensed software) – managed by enterprise or IBM ▪ IBM Cloud ▪ Unified hybrid deployments	▪ Cloud only
Customization and control	▪ With Dedicated Service, dynamic control over performance, on a workload basis ▪ Visibility and reporting	▪ No customized control or workload visibility ▪ Standard reporting
API support	▪ OpenStack Swift ▪ S3 Compatible API ▪ Simple object API ▪ NSF/SMB	▪ S3 only

REFERENCES

[1] R. Buyya, C. Yeo, S. Venugopal, J. Broberg & I. Brandic, 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems. Journal of Future Generation Computer Science. vol. 25 no. 6 pp. 599-616

[2] P. Mell, T. Grance, The NIST Definition of Cloud Computing, NIST Special Publication 800-145, 2011

[3] M. Ali, S. Khan, A. Vasilakos, Security in Cloud Computing: Opportunities and Challenges, Information Sciences 305 (2015) 357-383.

[4] K. Bowers, A. Juels, A. Oprea, HAIL: A High-Availability and Integrity Layer for Cloud Storage, CCS'09, November 9–13, 2009, ACM 978-1-60558-352-5/09/11.

[5] Y. Cao, C. Chen, F. Guo, D. Jiang, Y. Lin, B. Ooi, H.Vo, S. Wu, Q. Xu, ES2: A Cloud Data Storage System for Supporting Both OLTP and OLAP, Accessed on 24 May 2017

[6] R. Shaikh, M. Sasikumar, Data Classification for Achieving Security in Cloud Computing, Procedia Computer Science 45 (2015) 493 – 498.

[7] C. Wang, Q. Wang, K. Ren, N. Cao, W. Lou, Towards Secure and Dependable Storage Services in Cloud Computing, IEEE Transactions on Cloud Computing Date of Publication: April-June 2012 Volume: 5 , Issue: 2.

[8] A. Kumar, B. Lee, H. Lee, A. Kumari, Secure Storage and Access of Data in Cloud Computing, ICTC 2012, 9 78-1-4673-4828.

[9] H. Abu-Libdeh, L. Princehouse, H. Weatherspoon, RACS: A Case for Cloud Storage Diversity, SoCC'10, June 10–11, 2010, ACM 978-1-4503-0036-0/10/06.

[10] Q. Liu, G. Wang, J. Wu, Efficient Sharing of Secure Cloud Storage Services, Access on 24 May 2017

[11] J. Wu, L. Ping, X. Ge,Y. Wang, J. Fu, Cloud Storage as the Infrastructure of Cloud Computing, 2010 International Conference on Intelligent Computing and Cognitive Informatics. DOI 10.1109/ICICCI.2010.119.

[12] W. Zeng, Y. Zhao, K. Ou, W. Song, Research on Cloud Storage Architecture and Key Technologies, ICIS 2009, November 24-26, 2009, ACM 978-1-60558-710-3/09/11.

[13] A. Singh, S. Pasupuleti, Optimized Public Auditing and Data Dynamics for Data Storage Security in Cloud Computing, 6th International Conference on Advances in Computing & Communications, ICACC 2016, 6-8 September 2016, Cochin, India. Procedia Computer Science 93 (2016) 751 – 759

[14] N. Vurukonda, B. Rao, A Study on Data Storage Security Issues in Cloud Computing, 2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016) , Procedia Computer Science 92 (2016) 128 – 135

[15] A. Adshead Cloud, compliance and data protection top storage priorities for 2016, TechTarget, ComputerWeekly Publication

[16] B. Betts Cloud storage appliances: what are they and who provides them, 2016, TechTarget, ComputerWeekly Publication

[17] L. Stadtmueller, Which Cloud Storage Service Delivers the Performance You Need? Comparing IBM Cloud Object Storage and Amazon S3, An Executive Brief Sponsored by IBM, 2016 Stratecast.