

Amazon Backed File System: A Review

S.Malhotra^{1*} and V. Bali²

^{1*}Department of CSE, Panipat Institute of Engineering and Technology, Samalkha, Haryana, India

²Department of CSE, Panipat Institute of Engineering and Technology, Samalkha, Haryana, India

Available online at: www.ijcseonline.org

Received: 18/Mar/2018, Revised: 26/Mar/2018, Accepted: 15/Apr/2018, Published: 30/Apr/2018

Abstract: Cloud Computing is one of the major emerging IT technologies and is highly promising in terms of resources provisioning, unlimited data storage, remote access to applications, easy data backup amenity etc. One of its emerging platforms is AWS S3. It provides storage for the web. It serves developers/users with easy web-scale computing platform. This platform is reliable, scalable, inexpensive and efficient and provides persistent cloud storage. Amazon S3 and amazon.com share common scalable storage root in terms of their infrastructure. The data here is stored in fundamental containers called buckets. A bucket has the capacity to store limitless data. Infinite objects can be stored in a bucket, with each object being able to store a maximum of 5 TB of data. These objects require a unique developer-assigned key to be stored and retrieved by the developers. The data stored this way can be downloaded any instant of time by any user. Permissions to upload or download data to one's own bucket can be granted or denied. The web service always keeps the data secure from unauthorized access using its invulnerable authentication mechanism. Various operations can be executed through the API such as Read an object, delete an object, list keys etc. Amazon S3 associates REST and SOAP API interfaces.

Keywords—Cloud Computing, Cloud Storage, Amazon Web Services, Bucket, Folder, File, Hierarchy, Flat File System.

I. INTRODUCTION

Amazon Simple Storage Service commonly known as Amazon S3, is an Infrastructure as a Service (IaaS) solution which is provided by Amazon Web Services (AWS). Amazon S3 has amplified the IT industry/development activities as it facilitates highly secured and low-latency data storage from the cloud.

The Amazon S3' highly private and scalable interface provides its users the flexibility of data storage or retrieval from any location at any instant of time. It serves the needs of a diverse group of people whether it be IT sector or any other business process, whether it is an individual or a group of folks. It deals with a variety of services that cover the entire range of computing stack, whether it is hardware or the application level. The clients can send requests to create buckets, store and retrieve data in the form of objects, access resources by managing permissions on resources. [1, 2]

The section I in the paper gives the introduction about the Amazon S3 and its services. The second section elaborates the storage infrastructure of Amazon S3 that includes Buckets, Objects, Keys, Regions, Consistency model etc. This section further explains the pros and cons of AWS S3 model. The section three studies the related work done by

various authors in the AWS domain. The next section, the section four extends the Storage Infrastructure to explain Amazon EC2 modeling, its scalability, security and flexibility. The section five describes the proposed methodology based upon the research done in the AWS domain; how one of disadvantages can be overcome to provide the solution with the Amazon S3 flat file system. The section six concludes the research work by providing the conclusion and how the proposed solution will work out. The final section which is the reference section addresses by providing the reference of Journals and authors whose work has been studied in doing the research in the desired domain.

II. STORAGE INFRASTRUCTURE

This infrastructure is built on the core roots of buckets and objects and Amazon S3 API is used to work with these resources. Following include the key feature set of Amazon S3 resources:

Buckets - These are global containers for objects stored on Amazon S3 cloud storage and serve as the organizational structure because they organize the Amazon S3 namespace at the highest level. Each and every object has to be contained in a bucket.

Objects - These are the building block entities of Amazon S3 that have object data and Meta data. The data is stored in the

form of key-value pairs. Hence Key and Version ID are the unique identifications of any object.

Keys - Keys uniquely identify any object inside a bucket. Each object is solely associated to only one key. Hence any object inside Amazon S3 can be mapped or addressed through the combination of a key, bucket name and a version ID (optional).

Regions - It refers to the geographical region where the Amazon S3 buckets that users create, have to be stored. Currently, 14 regions are supported. Objects once stored inside a particular region remain inside that region only until explicitly transferred to any other region.

Consistency Model - The consistency here refers to read-after-write consistency of data for PUTS of objects in a bucket in any region. Success of PUT request ensures the safety of the data stored. Update operations to a key are atomic. If a put operation is performed on any key followed by a read operation, it might return either the old or the updated data but it will never allow write operation that may result in corrupted data. [1, 3, 4]

There are numerous benefits of Amazon Simple Storage Service, few of which are listed below:

Cost Efficiency - It cuts off the costs in IT industry by eliminating the need of various resources such as hardware, storage and floor space etc. and other operational and maintenance costs thus optimizing the performance as well as capital. Moreover it offers free usage tier to get the users started without any set up cost. This free usage however limits to 5GB. After 5GB, one needs to pay only for the storage which is actually used.

Easy Integration with Third Party - The data stored on Amazon S3 can easily be connected with third party applications via restful APIs and SDK. The application program interface allows the data to be stored in S3 organized uniquely identified buckets. S3 can unite with mobile and web applications both and provide consistently great performance regardless of the application that is in use.

Storage and Distribution- It can easily serve as the basis for CDN (Content Distribution Network) when paired with Amazon Cloud Front. It can then be configured to distribute various virtual instances of data across n number of Amazon Data centres. Hence it is efficient enough in automatic synchronization of high level of data without any point of failures, thus providing primary storage to critical data.

Security - Amazon provides fine-grained access control to the resources stored on S3 bucket. Encryption is supported to ensure the authorized access to data. It uses both client and server-side encryption while downloading and uploading of data respectively. Access logging can also be enabled to track requests for access to bucket.

Static Website Hosting - Developing static websites using HTML / CSS / JavaScript etc makes it affordable and easily configurable with S3. S3 will run it efficiently and faster. It can also be extended to include data driven solution that utilizes Content Management System like Word press/Drupal etc.

Scalability/Elasticity - Amazon S3 ensures high level of Scalability when it comes to storing large amount of data/files in a directory. Its bucket supported structure can store virtually unlimited number of bytes in one bucket which is then automatically managed / organized by Amazon S3 by scaling the redundant data to different servers.

Backup and Recovery-S3 continually implements version controlled backup within buckets. One can maintain scheduled backup, manual backups on internal servers, archive data with enterprise-level archiving tools. It also provides disaster recovery using Cross Region Replication which replicates S3 data across other AWS data centres. Hence S3 has extensive capability to provide data backup and recovery using its super efficient tools, which is the most useful cog across all Amazon S3's benefits. [5]

Performance - Amazon S3 infrastructure is designed in a way to avoid server side latency when clients use multiple applications subsequently/simultaneously. It can provide adequate throughput without any network issues.

Object Tagging - S3 has a smart tagging feature which uses a tagging model across data files and web pages to describe the data in detail which in turn adds meaning to the meta data. This feature can be fully accessed using Amazon APIs. [6]

The cloud computing though being a boon to the society and IT industry, it has downsides too. Some of its disadvantages are listed below:

Security and Confidentiality: The infrastructure and storage of data in cloud is managed by third party called cloud service providers, so the data is very much vulnerable to security risks. The users are always afraid of storing sensitive data and information on cloud. Though the secured passwords and protected accounts ensure the users are authenticated and authorized before accessing the information leading to protection of data but any signs of hacking or security breach may result in loss of business. Hackers can analyze the critical task and hack the accounts to attack the confidential information.

Dependency: The management of the infrastructure in cloud is done by Cloud Service Providers and it becomes very difficult for the It is very difficult for the clients to switch from one Cloud Service Provider to another resulting into dependency on a particular Cloud Service Provider.

Centralized System Failure: Rather than storing data in different physical locations, user's data is stored virtually in

data centres with cloud computing technology that involves the risk of failure of centralized isolated mechanism resulting in the loss of client's data or business. [3]

Breaching of Data: The data stored on cloud can become an attractive target of security/data breach. The breaching of data associated with health information, organization's private information and other intellectual property can be shattering. The impact of damage due to this depends on the sensitivity or level of privacy of the data. The organization may have to suffer fines or other legal actions whenever a security breach occurs because organizations are held responsible for protecting their environments and their own data.

Data Hijacking: Attackers can falsify the credentials and hijack personal data and information through hijacked credentials that can result in actual user not able to get access to its own data. This can be a devastating threat to user's data. The hijackers can remotely access the sensitive information stored on the cloud. The attackers use various techniques such as scripting bugs and guessed/reused passwords, stealing credentials to gain access to the data.

Compromised identifications and broken authentication: The data stored on cloud is vulnerable to breaching and hijacking because of the poor credentials that people use while storing their data. Organizations many times forget to deactivate the accounts of employees when they have left the organization. Also, people are not concerned about these security threats when the organizations allocate permissions to the employees. Though, strong authentication techniques involve phone-based authentication, One-time password, Unrecognizable strong passwords but people rarely think about these security concerns.

Denial of Service Outbreak: Due to this, the resources and services of a host connected to internet are interrupted or suspended so that the services are many times not available to its intended users. It also takes the security appliances down. Thus cloud providers are better poised.

The data storage infrastructure of AWS has one of the limitations, the way files / folders are stored as flat file system. The users can not store their files and folders in directory structure on cloud via AWS. The proposed methodology hence targets this limitation.

III. RELATED WORK

Several works related to the paper being proposed that present the security of data in cloud computing is as follow:

In 2017, Yibin Li, Keke Gai, Longfei Qiu, Meikang Qiu, Hui Zhao proposed a novel methodology that focused on the problem of the cloud data storage and provided an excellent approach to avoid the cloud operators reach the user's

sensitive data and their approach was. named as Security-Aware Efficient Distributed Storage (SA-EDS) model.

In 2016, Omer Y. Adam, Young Choon Lee, and Albert Y. Zomaya proposed one of the major concerns in Cloud where multiple resources are to be located via Performance Predictability. The optimization model in the paper correlated performance variability that occurred across various instances in cloud. [7]

In 2015, S. Narula described how AWS proves out to be a robust, secure and an excellent model when it comes to the security and storage of data on cloud. AWS being the most trusted provider of cloud computing not only provides the excellent performance and security but is also enriched with many other core operations and services. [8]

In 2014, Monjur Ahmed and Mohammad Ashraf Hossain presented a review on in-depth concepts of cloud computing and cloud data security issues that are inherent within cloud

Infrastructure and cloud computing context. He took into consideration the technical and philosophical factors while dealing with cloud data security issues. The influence of cloud computing includes both, the technical as well as social effects. [9]

In 2012, V. Michael Vrable, S. Stefan and Geoffrey M. Voelker analyzed and prospected solution to network backed file systems on cloud that how LAN based workstations are transparently served by cloud based services and how these continue to provide good performance for enterprise workloads. The paper has proposed the optimization techniques that help in achieving low-cost and good performance along with secured log-structured design. [3]

In 2011, S. Carlin founded that sharing of resources is one the biggest security concern associated with cloud computing. This caters the need to develop and adopt new cloud technologies or techniques that are compatible with cloud architecture and these should replace the older ones. [6]

In 2010, Jesús Hernández Martín, Ioan Raicu focused on IaaS cloud platform after studying the raw performance in terms of I/O. The platform is reliable and easy to use with the increasing number of public cloud platforms and the growth in terms of computing capacity. They have described the tools that can be used for benchmarking which is actually the description of the file storage systems and solutions. [10]

In 2007, Simson L. Garfinkel analyzed the security model of Amazon and based upon which presented the user-report of Amazon of Amazon's computing EC2 and S3 services and concluded that EC2 delivers and continue to provide virtual machines at low cost. This paper deals with the details of

grid computing services between November 2006 and May 2007, which also includes a detailed analysis of the overall system's application program interface. [4]

Cloud storage is driven by the availability of commodity services from Amazon's S3 and other service providers. The flexible feature of cloud storage and services is reminiscent of the motivation for the Plan 9 write-once file systems [11, 12].

IV. AWS STORAGE METHODOLOGY

Cloud computing extends the storage and scalability of data on a very large scale which is a boon especially to IT enterprise when compared to traditional solutions where IT services are subjected to physical and logical controls. [13]

Cloud computing basically has two major components —the front end and the back end. The front end refers to the client, the devices and the applications whereas the back end refers to the cloud itself which is the data warehouse. The whole cloud monitors the client demands. [13]

The paper focuses entirely on cloud data storage which is backed up by Amazon Web Services. Amazon EFS which stands for Amazon Elastic File System is a scalable, easy to use and simple file storage system to be used with Amazon EC2 instances on AWS cloud. It provides a simple and user friendly interface for the users to create and configure file systems.[14] Moreover, the storage capacity that it provides is elastic; can grow or shrunk automatically depending upon the user's storage requirements, users can add or remove files depending upon the storage needs.

The Amazon EFS file system when mounted on Amazon EC2 instances provides a file system interface and access semantics that allow its users to easily integrate EFS with existing tools and applications. Many instances of Amazon EC2 can simultaneously access EFS file system and hence provides a common data source for users' applications to run on multiple EC2 instances. [10]

Amazon backed up file system is highly available, durable and provides excellent performance to serve a wide variety of users that includes ERP, content management systems, web applications, database backups and servers, big data and analytics and other developer tools.[9,15]

V. PROPOSED PRINCIPLES AND METHODOLOGIES

The paper proposes an efficient and resourceful technique to eradicate the storage structure limitation of AWS. This flexible distributed scheme provides explicit dynamic sustenance to ensure the data is stored in a directory structure on cloud instead of flat file system.

The proposed paper deals with storage of user's data on cloud in a flexible way such that the files and the folders can be stored in a hierarchical manner and this hierarchy can extend to a great level and user can move to any node or any folder in the hierarchy by clicking the desired node in the structure.

Each user is first authenticated before he/she can access the data stored on the cloud to ensure the security of data. [1, 3, 10, 11, 12]. There would be one common storage environment for all users.

VI. CONCLUSION

The paper hence reviews various cloud storage tools and algorithms. The data stored on the cloud via AWS has to be secured, readily available and durable. Hence the promise of AWS to ensure that the cloud storage and computation will be seamlessly outsourced from individual to massive data centres distributed on the world. The limitation of AWS of storing files as flat file system has also been addressed and storage of this data on cloud in the form of directory structure can be the solution to eradicate the problem so that users can make the maximum benefits of AWS storage.

V. REFERENCES

- [1] R. Uhlig et. al., Microsoft Cloud Technology, IEEE Internet Computing, May 2005, Vol. 38, No. 5.
- [2] Shivalal Mewada, Umesh Kumar Singh and Pradeep Sharma, "Security Enhancement in Cloud Computing (CC)", ISROSET-International Journal of Scientific Research in Computer Science and Engineering, Vol.-01, Issue-01, pp (31- 37), Jan -Feb 2013.
- [3] A. Bessani, M. Correia, B. Quaresma, F. Andr'e, and P. Sousa. DepSky: Dependable and Secure Storage in a Cloud-of-Clouds. In EuroSys 2011, Apr. 2011.
- [4] Parneet Kaur and Sachin Majithia, "Various Aspects for Data Migration in Cloud Computing and Related Reviews", International Journal of Computer Sciences and Engineering, Volume-02, Issue-07, Page No (83-85), Jul -2014, E-ISSN: 2347-2693
- [5] N. Zhu, J. Chen, and T.-C. Chiueh. TBBT: Scalable and Accurate Trace Replay for File Server Evaluation. In Proceedings of the 4th USENIX Conference on File and Storage Technologies (FAST), Dec. 2005.
- [6] A. Muthitacharoen, R. Morris, T. M. Gil, and B. Chen. Ivy: A Read/Write Peer-to-Peer File System. In Proceedings of the 5th Conference on Symposium on Operating Systems Design and Implementation (OSDI), Dec. 2002.
- [7] Omer Y. Adam, Young Choon Lee, and Albert Y. Zomaya, "Constructing Performance-Predictable Clusters with Performance-Varying Resources of Clouds", 2016.
- [8] S. Narula. "CLOUD COMPUTING SECURITY:AMAZON WEB SERVICE". 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015.
- [9]J. Howard, M. Kazar, S. Nichols, D. Nichols, M. Satyanarayanan, R. Sidebotham, and M. West. Scale and Performance in a Distributed

- File System. ACM Transactions on Computer Systems (TOCS), 6(1):51–81, Feb. 1988.
- [10] Jesús Hernández Martín, Ioan Raicu. Performance evaluation of AWS; Exploring storage alternatives in Amazon Web Services. http://datasys.cs.iit.edu/reports/2012_AWS-storage-benchmarks-TR.pdf.
- [11] R. Kotla, L. Alvisi, and M. Dahlin. SafeStore: A Durable and Practical Storage System. In Proceedings of the 2007 USENIX Annual Technical Conference, June 2007.
- [12] J. Li, M. Krohn, D. Mazières, and D. Shasha. Secure Untrusted Data Repository (SUNDR). In Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementation (OSDI), Dec. 2004.
- [13] P. Mahajan, S. Setty, S. Lee, A. Clement, L. Alvisi, M. Dahlin, and M. Walfish. Depot: Cloud Storage with Minimal Trust. In Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation (OSDI), Oct. 2010.
- [14] Meiko Jensen, JorgSchwenk et al. “2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015.
- [15] Yibin Li, Keke Gai, Longfei Qiu, Meikang Qiu, Hui Zhao, “Intelligent cryptography approach for secure distributed big data storage in cloud computing”, 2017.
- [16] Simson L. Garfinkel. An Evaluation of Amazon’s Grid Computing Services: EC2, S3 and SQS. <https://dash.harvard.edu/bitstream/handle/1/24829568/tr-08-07.pdf?sequence=1>
- [17] R. Pike, D. Presotto, S. Dorward, B. Flandrena, K. Thompson, H. Trickey, and P. Winterbottom. Plan 9 From Bell Labs. USENIX Computing Systems, 8(3):221–254, Summer 1995.
- [18] S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In Proceedings of the 1st USENIX Conference on File and Storage Technologies (FAST), 2002.
- [19] P. Shanthi Bala, “Intensification of Educational Cloud Computing and Crisis of Data Security in Public Clouds”, IJCSE Vol. 02, No. 03, 2010, 741-745.
- [21] M. Rosenblum and J. K. Ousterhout. The Design and Implementation of a Log-Structured File System. ACM Transactions on Computer Systems (TOCS), 10(1):26–52, 1992.
- [22] C. Ruemmler and J. Wilkes. A trace-driven analysis of disk working set sizes. Technical Report HPL-OSR-93-23, HP Labs, Apr. 1993.
- [23] R. Sandberg, D. Goldberg, S. Kleirnan, D. Walsh, and B. Lyon. Design and Implementation of the Sun Network Filesystem. In Proceedings of the Summer USENIX Technical Conference, pages 119–130, 1985.
- [24] J. Satran, K. Meth, C. Sapuntzakis, M. Chadalapaka, and E. Zeidner. Internet Small Computer Systems Interface (iSCSI), Apr. 2004. RFC 3720, <http://tools.ietf.org/html/rfc3720>.
- [25] M. Vrable, S. Savage, and G. M. Voelker. Cumulus: Filesystem Backup to the Cloud. In Proceedings of the 7th USENIX Conference on File and Storage Technologies (FAST), Feb. 2009.

Authors Profile

Ms. Shilpa Malhotra pursued Bachelor of Technology from Asia Pacific Institute of Information and Technology, Staffordshire University in 2014. She is currently pursuing Master of Technology from Panipat Institute of Engineering and Technology, Kurukshetra University. She is currently working as Software Quality Analyst in an IT firm. She has 2 years of work experience in the same domain and one year of teaching experience.



Dr. Vikram Baliis currently working as Head of the Department at Panipat Institute of Engineering and Technology, Kurukshetra University. His qualifications include Ph.D., ME, M.Tech, B.Tech in Computer Science and Engineering. His main domain area is Cloud Computing. He has working knowledge in various other domains such as Networking, Digital Image Processing etc.

