WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

How Data is Deduplicated using AWS Services

Atharva Sontakke

All India Shri Shivaji Memorial Society -Institute of Information Technology, Pune, Maharastra, India.

I. Introduction

High-performance computing resources, a complex setup environment, and a tonne of storage are requirements for scientific endeavours. With less expensive infrastructure construction costs and more elastic and flexible system extendibility, the emerging cloud computing technology offers scientists an alternative to the traditional grid computing that has been used by scientific projects for decades. It also offers better computing performance and massive storage requirements. Despite these alluring advantages and the rising interest in using research clouds, there are still concerns and obstacles that prevent cloud computing from being widely used in the scientific community. The purpose of this study is to outline the main considerations for using cloud computing in science and to examine how these considerations interact.

As a potential paradigm to serve a wide range of computing needs and application development requirements, cloud computing has drawn considerable attention from both industry and science disciplines. Both private and public cloud services have proven to be capable of offering a scalable set of services that may be used quickly and affordably on an as-needed basis for self-service. By pooling resources over a wide user base, economies of scale are made possible, and a virtual resource can be accessible via a network and elastically modified by the user.

The study of scientific computing focuses on developing mathematical models and numerical methods to solve issues in science and engineering.

Scientific applications are frequently quite complicated and frequently require both a lot of storage space and high-performance computational resources [2, 3]. Users of scientific computing are drawn to cloud computing because it is a less expensive alternative to owner-centric high-performance computing (HPC) and specialised clusters, a platform that is much more dependable than Grids, and a platform that is much more scalable than the biggest commodity clusters [4].

The study of conducting science in the cloud has already begun, as evidenced by early projects like the University of Chicago's Nimbus project[5], the Steinbuch Centre for Computing's Cumulus [6, 7], the NASA Ames Research Center's Nebula Cloud[8,] etc. The viability of cloud computing has been demonstrated in a variety of fields, including high energy physics, computer science, bioinformatics, economics, and educational projects [9,10]. However, there are a number of important problems and difficulties that worry scientific researchers about the widespread adoption of cloud computing.

The findings from the assessment of past research's literature and focus group study were used to develop a two-level hierarchy of consideration elements consisting of four high-level dimensions and ten detailed level influence variables, as shown in TABLE I. Academic journals and actual examples from the disciplines of information technology, scientific computing, and new technology adoption study are searched for throughout the literature review. The focus group consists of one project executive, two senior managers, and two engineers, all of whom are active participants in Taiwan's national science cloud project.

Dimensions	Factors	Description
Relative Advantage(A)	Performance(a1)	Referred to speed of data processing by the cloud system. It includes both virtualized and physical computing resources.
	Elasticity (a2)	Referred to the ability to scale the IT infrastructure dynamically and quickly. It includes scaling out of multiple Virtual Machines, storage resource and computing power.
	Capacity(a3)	Referred to volume of work or data processing capacity that the cloud system can handle. It includes computing and storage capability.
Compatibility(B)	Interoperability(b1)	Referred to the ability of supporting heterogeneous cloud configurations (cloud-to-cloud data migration). It includes applications to utilize multiple distributed heterogeneous resources.
	Integration(b2)	Referred to intra-enterprise IT systems integration (cloud-to- enterprise integration). It includes the connecting with exiting IT systems or legacy systems and services.
	Availability(b3)	Referred to the amount of time the cloud system should be operating. It includes fault tolerance and capacity management.
Complexity(C)	Security (c1)	Referred to the protection of data from theft, corruption, or natural disaster, while allowing the data to remain accessible and productive to its intended users.
	Adaptability(c2)	Referred to ability of satisfying various users' services and application needs.
Environment Condition(D)	Cost(d1)	Referred to the budgets and man power of cloud computing implementation, administrative and maintenance.
	Networking (d2)	Referred to network performance. It includes transaction effectiveness, speed and security of internet.

What is AWS?

Amazon's cloud computing platform, AWS (Amazon Web Services), offers a variety of infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), and packaged software-as-a-service (SaaS) options. An enterprise may be able to benefit from AWS services in terms of resources like computing power, database storage, and content delivery. The internal infrastructure that Amazon.com constructed to conduct its online retail activities served as the foundation for the initial web services that Amazon.com Web Services offered in 2002. It started providing its distinguishing IaaS services in 2006. One of the first businesses to offer a cloud computing platform that scales to meet users' compute, storage, or throughput needs was AWS. For businesses and software developers, AWS provides a wide range of tools and solutions that may be used in data centres in as many as 190 nations. AWS services are available to organisations including governmental bodies, educational institutions, non-profits, and private ones.

AWS is divided into various services, each of which can be customised based on the requirements of the user. Users can view individual server maps and configuration parameters for an AWS service.

The AWS portfolio includes more than 200 services, including ones for application development, database administration, computation, and security. These services fall into the following categories:

- Compute
- Storage
- Databases
- Data Management
- Migration
- Hybrid Cloud
- Networking
- Development Tools
- Management
- Monitoring
- Security
- Governance
- Big Data Management
- Analytics
- Artificial Intelligence (AI)
- Mobile Development
- Messages and Notification

What is deduplication?

Data deduplication is a procedure that gets rid of extra copies of data and drastically reduces the amount of storage space needed.

Deduplication can be implemented as a background process to remove duplicates after the data has been stored to disc or as an inline procedure to remove duplicates while the data is being written into the storage system.

To maximise savings, deduplication, a zero data loss technology, is conducted both as an inline procedure and as a background process. In order to avoid interfering with client operations, it is executed opportunistically as an inline process. To maximise savings, it is done thoroughly in the background. Deduplication is enabled by default, and the system executes it without any user input on all aggregates and volumes.

Because deduplication operations take place in a specialised efficiency domain that is distinct from the client read/write domain, there is very little performance overhead. No matter what programme is used or how the data is accessed (NAS or SAN), it operates in the background.

II. Related Work

Amazon Web Services (AWS) works with institutions and research labs all around the world to process complicated workloads for researchers. AWS offers scalable, affordable, and secure database, storage, and computing capabilities to quicken the pace of science. Researchers can progress research utilising game-changing technologies like artificial intelligence (AI), machine learning (ML), and quantum with the help of Amazon Web Services (AWS). Researchers collaborate with people all across the world, sharing their findings. In order to speed up innovation, AWS also gives researchers access to open datasets, funding, and training.

In partnership with top universities as well as domestic and foreign research organisations, AWS promotes innovative research. The National Science Foundation, the National Aeronautics and Space Administration, the European Space Agency, and the National Institutes of Health are a few examples. AWS also engages in computer, biomedical, engineering, quantum information, and space scientific research and development.

The companies that offer cloud storage assume full responsibility for data protection and also offer dependable data access. In addition to this, employing a cloud storage service has a lot of other advantages [2]. There are numerous cloud storage companies, including Apple iCloud, Amazon Drive, Google Drive, Dropbox, Box, and Microsoft Onedrive. Even though the cloud offers users storage space as a service, redundant copies of data prevent it from being used effectively. Such redundant copies will unnecessarily take up storage space. This will result in a less effective use of storage capacity. To promote the better utilisation of storage capacity, numerous storage optimization techniques are used. They are data deduplication, data compression, thin provisioning, snapshots, and clones.

TABLE I. DATA GROWTH RATE

Year	Data volume (ZB)
2006	0.16
2007	0.28
2008	0.48
2009	0.8
2010	1
2011	1.8
2015	8
2020	40

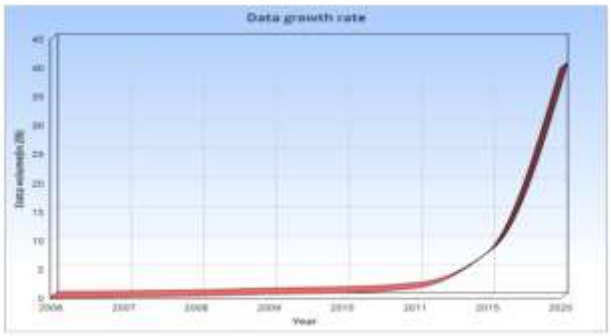

Fig. 1. Growth Rate of Data.

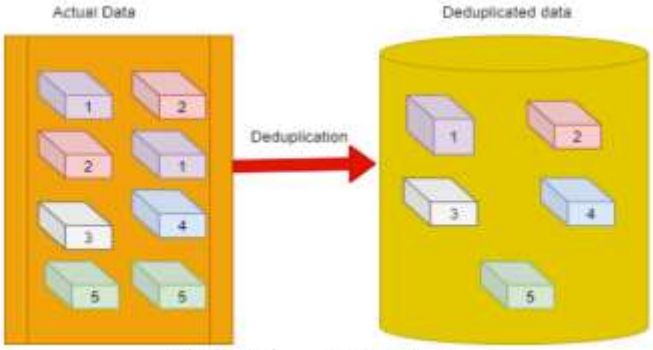

Fig. 2. Process of Deduplication.

III. Research methodology and results

According to CNBC, the COVID-19 pandemic is one reason why providing contemporary, affordable virtual healthcare options, like telemedicine, with the cloud is a rapidly increasing field of healthcare.

Interactions between patients and doctors can now take place virtually and in real time thanks to telehealth technology. The ability to receive care without needing to visit a doctor's office is a major feature that provides advantages for people in remote or hard-to-reach places and patients with mobility challenges. These solutions can be helpful for both patients and care providers. These approaches can also lessen the physical traffic in hospitals, which is crucial during the pandemic.

Customers from all across the world discuss how developing on Amazon Web Services (AWS) enables them to scale, innovate, and operate at scale to enhance the experience for patients and healthcare professionals.

In collaboration with the whole education community, including students, teachers, administrators, and researchers, Amazon Web Services (AWS) is on a mission to hasten the digital revolution of education.

With the help of AWS, more than 14,000 educational institutions of all sizes—from elementary and secondary schools to colleges and universities—are able to fulfil their essential missions and achieve their most important institutional priorities.

Amazon Web Services (AWS) is used by public sector healthcare organisations all over the world to analyse their data, enhance patient care in the communities they serve, and provide teams with the resources they need to assess, diagnose, and treat patients in real time. By utilising AWS, healthcare businesses may lessen the time and effort needed to run their current clinical and non-clinical workloads and give patients access to new analytics and machine learning tools that satisfy local and international security and privacy regulations.In future research, we hope to examine the various deduplication techniques in relation to different types of digital data.

IV. Conclusion

AWS is made to enable suppliers, ISVs, and application providers to swiftly and securely host your apps, whether they are SaaS-based or not. To access AWS's application hosting platform, use the AWS Management Console or well-documented web services APIs.

You can choose the web application platform, programming language, operating system, database, and other services you require with AWS. You get a virtual environment through AWS that you may fill with the programmes and services your application needs. This facilitates the migration of current applications while keeping open the possibility of creating new solutions.

There are no upfront costs or long-term obligations; you simply pay for the processing power, storage, and other resources that you really utilise. Visit the AWS Economics Center for further details on comparing the expenses of other hosting options with those of AWS. We now know why the majority of programmers pick AWS as their cloud computing provider.

With the use of deduplication algorithms, cloud storage can be used effectively. Deduplication strategies come in many different forms, and each one has pros and cons that can be effectively used in the right circumstances. In this essay, we examined the advantages of employing the deduplication approach for effective cloud storage space use and talked about several deduplication techniques in various dimensions. In-line with variable-sized chunk level deduplication, among other techniques, will help us to efficiently utilise cloud storage space because it does not require additional storage space, as it does with the post-process method, because it does not store duplicate data, and because variable-sized chunk level deduplication looks for whether or not even a portion of the data is duplicate rather than the entire data set is duplicate. The primary goal of this research project is to contrast the various deduplication methods that are currently available. Applying the appropriate deduplication techniques in each case will help the various groups of cloud storage providers.

V. Acknowledgment

I thank my seminar guide Mrs.Punashree Patil of the information technology faculty, AISSMS-IOIT for providing their valuable opinions and knowledge for this study.

VI. References

[1].”Key Consideration Factors of Adopting Cloud Computing for Science”

Chia-Lee Yang*1,2, Bang-Ning Hwang *3 and Benjamin J.C. Yuan1

1 Institute of Management of Technology, National Chiao Tung University, Hsinchu, Taiwan

2 National Center for High-Performance Computing, Hsinchu, Taiwan

3 Graduate Institute of Business Administration, National Yunlin University of Science and Technology, Yunlin, Taiwan

[2].”Leveraging Platform Basic Services in Cloud Application Platforms for the

Development of Cloud Applications”

Fotis Gonidis, Iraklis Paraskakis

southeast European Research Centre

International Faculty of the University of Sheffield,

City College

Thessaloniki, Greece

Anthony J. H. Simons

Department of Computer Science

The University of Sheffield

Sheffield, UK

a. j.simons@sheffield.ac.uk

[3]. “On-Demand Automated Fast Deployment and Coordinated Cloud Services”

Hsi-En Yu, Yi-Lun Pan, Chang-Hsing Wu, Hui-Shan Chen, Chi-Ming Chen and Kuo-Yang Cheng

National Center for High-performance Computing

Taiwan, R.O.C. {yun, serenapan, hsing, chwhs, jonchen, kycheng @nchc.narl.org.tw }

[4].”A Comprehensive Study of Different Types of Deduplication Technique in Various Dimensions”

G.Sujatha1, Dr.Jeberson Retna Raj2

Department of Computer Science and Engineering

School of Computing, Sathyabama Institute of Science and Technology, Chennai, India1, 2

Department of Networking and Communications, School of Computing

SRM Institute of Science and Technology, Chennai, India1

[5].Video on “Why most coders use AWS? Link: https://youtu.be/BSGcQi2WNPg

[6]. Article link services provided by AWS: https://aws.amazon.com/blogs/publicsector/delivering-modern-accessible-virtual-healthcare-solutions-cloud/

[7].Video article on “Data Deduplication using AWS” Link: https://youtu.be/PrESGquZBG0

[3]. “On-Demand Automated Fast Deployment and Coordinated Cloud Services” Hsi-En Yu, Yi-Lun Pan, Chang-Hsing Wu, Hui-Shan Chen, Chi-Ming Chen and Kuo-Yang Cheng National Center for High-performance Computing Taiwan, R.O.C. {yun, serenapan, hsing, chwhs, jonchen, kycheng @nchc.narl.org.tw }

[4].”A Comprehensive Study of Different Types of Deduplication Technique in Various Dimensions”

G.Sujatha1, Dr.Jeberson Retna Raj2

Department of Computer Science and Engineering

School of Computing, Sathyabama Institute of Science and Technology, Chennai, India1, 2

Department of Networking and Communications, School of Computing

SRM Institute of Science and Technology, Chennai, India1

[5].Video on “Why most coders use AWS? Link: https://youtu.be/BSGcQi2WNPg

[6]. Article link services provided by AWS: https://aws.amazon.com/blogs/publicsector/delivering-modern-accessible-virtual-healthcare-solutions-cloud/

[7].Video article on “Data Deduplication using AWS” Link: https://youtu.be/PrESGquZBG0