Data duplication using Amazon Web Services cloud storage

Chapter · December 2020

DOI: 10.1016/B978-0-12-823395-5.00006-9

1 author:

Varaprasad Rao Mangu
CVR College of Engineering
27 PUBLICATIONS **48** CITATIONS

SEE PROFILE

16

Data duplication using Amazon Web Services cloud storage

M. Varaprasad Rao

CMR TECHNICAL CAMPUS, HYDERABAD, INDIA

16.1 Introduction

Data scalability has been a priority for business decision-makers as well as data and research experts since big data became a familiar phrase. This has raised legitimate data quality issues. A few years back, some of those issues were first raised but they have still not been answered adequately. Below are some of the key reasons why the consistency of data remains a concern even now (Web Analytics white paper).

- *Data duplication*: Data duplication is an event where an entity has many copies of the same source of information. Data replication sounds to a layperson as an important problem which any professional data scientist or experienced network administrator should avoid. Sadly, duplication of data is also a very common issue.
- *Inconsistent data formatting*: It can take a lot of time to process if the data formatting is not consistent across the data pools. Only the most robust Hadoop data extraction method will take time to complete the tasks exponentially. If there are many issues with data processing, questions cannot be answered or processed.
- *Incomplete data*: Uncompleted data is yet another important problem that continues to damage organizations. There are countless explanations for why the data of an entity cannot be complete.
- *Data obsolescence*: Another popular issue, which receives almost no attention, is data obsolescence. These data are frequently ignored because it is 100% reliable (or at least at the time it was initially stored). The problem is that the data is no longer meant to be deleted from the system.

16.1.1 Difference between data redundancy and data duplication

The main difference between data redundancy and data inconsistency is that data redundancy happens when the same piece of data is present in multiple database locations, whereas data inconsistency occurs when the same data is present in various formats in multiple tables.

Data Deduplication Approaches. DOI: https://doi.org/10.1016/B978-0-12-823395-5.00006-9

16.1.2 Cloud computing

Cloud computing provides modern computer services on demand, in particular storing data (cloud storage) and processing power, without direct, user-active control. The term describes data centers accessible via the Internet to several customers. There are many vendors to provide cloud services to the end users, such as Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure, IBM Watson, etc. They provide infrastructure as a service (IaaS), platform-as-a-service (PaaS), software-as-a-service (SaaS), network as a service, disaster and recovery as a service, and storage as a service (Wiki/Cloud Computing). Virtualization is the method of running a virtual machine (VM) in a layer which is separated from the hardware itself. The cloud computing services are offered through virtualization.

16.1.3 Data deduplication

Data deduplication includes a method of removing redundant data from a data set in the easiest possible sense. Extra copies of the same data are deleted during the deduplication process, leaving only one copy to be stored. To locate repeated segment samples, data can be analyzed to verify that a single case is indeed a file. Duplicates then are replaced by a pointer pointing toward the storage object. The overall process description of data deduplication is shown in Fig. 16−1.

Consider, for example, one e-mail server containing 100 occurrences of a single 1 MB file attachment, a display of graphics sales to each person on the global distribution team. All 100 display occurrences, requiring 100 MB of storage space, are documented without duplication of data if anyone backups the e-mail inbox. Only a single instance is being stored when data is deduplicated; any subsequent instance is compared simply back to the single copy, which decreases the specifications for space and connectivity to 1 MB maximum.

Deduplication of data, an effective approach to the reduction of records, is gaining increased attention and popularity owing to exponential growth in large-scale storage systems. It strips away redundant information from the server and varies duplicate information cryptographically.

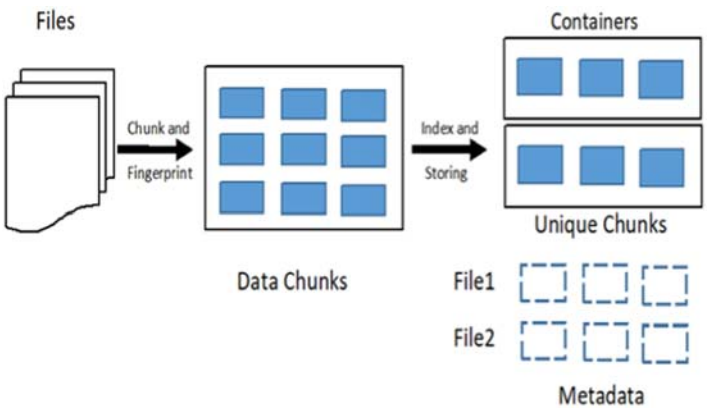

FIGURE 16–1 Processing of deduplication.

An adverse signature as an example of a secure Hash needs huge computation. It is better than conventional compression approaches in wide storage systems.

Next, the context and main features of data deduplication are analyzed, then state-of-the-art data deduplication research is summarized and categorized according to core data deduplication workflow. The description and taxonomy of the deduplication will help to define and explain the key design criteria for data-deduction systems. The researchers also address key uses and market developments in the deduction of data and list public sources for deduplication analysis and studies (Wen et al., 2016).

As shown in part by a substantial increase in the average size of content created in 2010 and 2011 from 1.2 to 1.8 ZB, (Economist/Data Deluge) and the volume of information to be generated by 2020 is expected to reach 44 trillion ZB (Gantz & Reinsel, 2020), the volume of digital data in the world is increasing exponentially. As a result of this "content stream," cost-effective control of storage is now one of the very challenging and interesting tasks in the Big Data era in mass data centers. The workload research carried out by the companies of Microsoft (Meyer & Bolosky, 2011; El-Shimi et al., 2012) and EMC (Wallace et al., 2012; Shilane et al., 2012) indicate that about 50% and 85% of information is recurrent, respectively, in main and secondary enterprise applications. As per an International Data Corporation report (DuBois et al., 2011), nearly 80% of companies studies were found to be pursuing data replication strategies in their data centers to decrease data redundancy, thus increasing storage capacity and reducing processing costs.

In addition to minimizing storage capacity by removing duplicate data, the deduplication of data is a major alternative to reduce the data, thus minimizing unnecessary data transmission in bass-bandwidth network environments (Bolosky et al., 2000; Policroniades & Pratt, 2004; Zhu, Li, & Patterson, 2008; Al-Kiswany et al., 2011). A standard data deduplication framework at the chunk side by side divides the incoming data flow into many data fragments (e.g., backing up files, snapshots from databases, virtual systems' images, etc.) each of which is uniquely defined and duplicated through a fingerprints' cryptographically protected hash signature (e.g., SHA-1) (Muthitacharoen et al., 2001; Quinlan & Dorward, 2002). The size of the chunks (Quinlan & Dorward, 2002), such as file blocks, and variable units can be set by content (Muthitacharoen et al., 2001). The doping programs delete redundant data bits and store or upload only one copy to conserve storage space or bandwidth for the program.

Data deduplication is more robust and effective than conventional compression methods for large-scale storage systems. The advantages are, first, deduplication defines and removes redundancy on a chunk or file level, while conventional compression works at the string or byte level, and second, by calculating its cryptographically protected hash-based fingerprint (Wen et al., 2016).

16.2 The workflow of data deduplication

Looking at the state-of-the-art data deduplication work, the key characteristics and distinctive structures of the data are important. These structures split the prevailing technology of

data deduction into eight classes based on the data deduplication process workflow illustrated in Fig. 16−2.

1 Chunking and Hashing

The easiest solution is to reduce the file/data stream to bits of the same size, which is known as the fixed-size chunk (FSC). In FSC, if the insertion or deletion process modifies the data chunk containing the modified part, irrespective of how small, not only does it change the data chunk, but all successive data portions transformation as the limitations of the above said chunks change. This may result in a different duplication FSC-based data deduplication ratio (before modification), otherwise identical chunks may be completely different.

The CDC algorithm, also known as Content-Based Curtail Chunking, was proposed in Lower Bound File Systems (Muthitacharoen et al., 2001) to cover chunk files or data to solve this border-shift issue (Kruus, Ungureanu, & Dubnicki, 2010; Xia et al., 2012) and duplicate identity sources. In particular, CDC uses a file content sliding-window technique and computes the hash value for the window (e.g., Rabin fingerprint). The Rabin (Rabin, 1981) algorithm currently constitutes the most commonly used algorithm to compute the CDC sliding-window hash value for data deduction (Meister et al, 2012; Broder, 1993).

There are various algorithms on chunking such as Two Threshold Two Divisors (Eshghi & Tang, 2005), regression chunking (El-Shimi et al, 2012), Fingerdiff (Bobbarjung, Dubnicki, & Jagannathan, 2006), MAXP (Teodosiu et al., 2006), simple byte (Aggarwal et al., 2010), Gear (Xia et al., 2014), Leap-based CDC (Yu, Zhang, Mao, & Li, 2015), bimodal chunking (Kruus, Ungureanu, & Dubnicki, 2010), subchunk (Romański et al., 2011), frequency based chunking (Lu, Jin, & Du, 2010), and metadata harnessing deduplication (Zhou & Wen, 2013). Various accelerating computational tasks for data deduplication are THCAS (Liu et al., 2009), HPDS (Guo & Efstathopoulos, 2011), P-Dedupe (Paulo & Pereira, 2014), StoreGPU (Al-Kiswany et al., 2008), Shredder (Bhatotia, Rodrigues, & Verma, 2012), and GHOST (Kim, Park, & Park, 2012). Data deduction is a calculation-intensive method with two computational tasks that take time, that depend on multicore-based tasks and GPGPU-based algorithms.

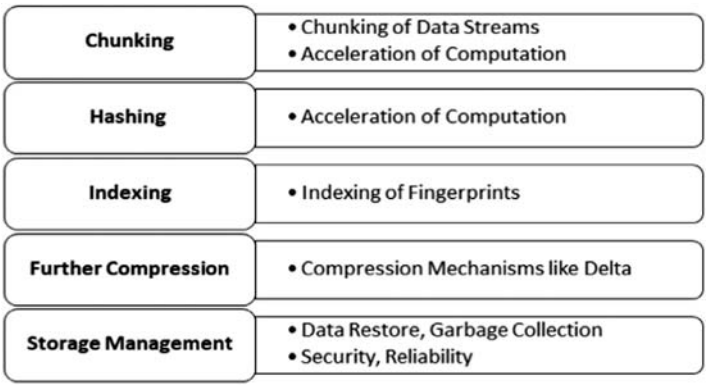

FIGURE 16–2 Data deduplication workflow.

2 Indexing

After the data streams have been chopped and scanned, chunk fingerprints are indexed to identify duplicate and nonduplicate chunks that are important to the deduplication process. The whole chunk fingerprint index can be saved by early deduplication systems for fast duplicate identification.

Depending on the particular method used, fingerprint indexing can be derived accurately or approximately for deduplication. Today, the index lookout process for deduplication is accelerated and the disk bottleneck is eased by four general categories: locally based, similarity dependent, flash-assisted, and cluster-based deduplication approaches.

3 Post-deduplication compression

Apart from the simple compressibility of post-duplication chunks, a high overlap can occur among alike chunks containing only a lesser quantity of bytes. The principal problems of the compression of the delta after deduplication are three long-term stages of identification of resemblances, reading of base parts, and delta encoding.

4 Data restore

The effective restoration and administration of segregated usable space are preferred once duplicate data have been identified and data are saved, and it is called garbage collection (GC). Accordingly, data recovery and garbage collection take two key drawbacks throughout the data duplication framework data storage phase. There are various methods or algorithms used in data restoration in one of the taxonomies, such as primary storage systems, backup storage systems, and cloud storage systems.

5 Garbage collection

In general, GC comprises two essential steps to detect the incorrect parts and to retrieve the storage space. Solutions to the GC can fall into two distinct categories as per the chunk, namely the reference number and the mark-and-sweep.

6 Security and reliability

In particular, Mozy, Dropbox, and Wuala, etc. are the cloud-based storage solutions that address significant security and reliability issues. Because of this security issue, users sharing content parts or files can pose security breaches and issues in a cloud environment (Wen et al., 2016; Guo & Efstathopoulos, 2011). Data deduplication reduces the reliability of the storage systems because the loss of a few critical data chunks can lead to many referenced files being lost (Bhagwat et al., 2006; Li, Lillibridge, & Uysal, 2011).

16.3 Deduplication in Amazon Web Services

16.3.1 Storage on Amazon Web Services

Data storage on AWS is a secure, trustworthy, extensible site for the information. Users can store, view, track, and evaluate the data to minimize costs, improve agility, and speed up innovation in AWS. To construct the base of a cloud IT system, select the web services such as object storage, file storage, and block storage services, backups, and data migration.

The types of storage in AWS are of the following

Amazon S3: Amazon Simple Storage Service (Amazon S3) is an object storing provision that offers high-end scalability, data accessibility and availability, secured storage of data, and fast extraction of data. It is used to store and protect any amount of data for a range of use cases, such as websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics.

Amazon Elastic Block Store: Using Amazon EC2 instances of AWS Cloud, the Elastic Block Store (Amazon EBS) offers a continuous block storage space. All Amazon EBS volumes are optimized within its Availability region, which offers high quality and reliability to protect the user from a product failure. Amazon EBS volumes have reliable and low latency efficiency to the payload function.

Amazon Elastic File System (EFS): The Amazon EFS is optimized for use with AWS Cloud infrastructure and onsite resources by offering powerful, flexible, elastic system files for Linux working loads. This is designed to scale petabytes on demand without interrupting implementation and to dynamically grow and shrink as the customer appends and deletes files, so clients' apps can store them if they need them. The architecture provides highly parallel distributed access to a lot of Amazon EC2 instances, allowing the frameworks with average low latencies to reach significant concentrations of aggregate performance and IOPS.

Amazon FSx for Lustre: Amazon FSx intended for Lustre remains a completely controlled, computationally intensive work file system for high efficiencies, machine learning (ML), and multimedia streaming architectures. The file system is fully controlled. Almost all the applications need a high-performance parallel file system and small latency.

Amazon FSx for Windows File Server: Amazon FSx provides a complete-managed Microsoft Windows native file system that can quickly transfer the Windows program to AWS, which includes file storage. Amazon FSx, based on Windows Server, makes available pooled file storage employing availability, flexibility, and complete support for Server Message Block protocol and NT File System for Windows, Active Directory (AD), and distributed management file system (DFS) systems. Amazon FSx uses solid drives (SSD) storage to provide good quality concerning the degree of throughput, increased IOPS, and stable submillisecond delays to Windows apps and clients.

Amazon S3 Glacier: It is a protected, reliable, and highly inexpensive information processing and enduring standby provider. It is aimed to accomplish durability of 99.9999% and offers complete secured management capacities, which can also help to fulfill only the strictest quality standards. It offers query-in-place features to allow great analytics to be performed effectively on the archive data.

AWS Storage Gateway: This is a blended service that allows AWS cloud storage to be used automatically for on-site applications. The service can be used to backup and archive, recover failure, manage cloud data, and data migration. Clients' frameworks can link to the system using different storage procedures, such as network file system, standard message block (SMB), and iSCSI via a virtual engine or configuration gateway device. This gateway links storage amenities like Amazon EBS, S3, and Glacier, which store AWS files, volumes, and virtual tapes.

16.3.2 Data deduplication through Amazon Web Services

Large databases also contain data redundancy, raising the value of storage. For instance, several consumers can stock several clones or snapshots of the identical file with user file segments. Several repositories are constant over time with software development shareholdings.

By allowing data deduplication for the file system, it can reduce the cloud storage expenses. Data deduplication eliminates or removes redundant data by saving identical data set parts only once. As a background mechanism, the deduplication of data is not significantly affecting the performance of the file system, it is also open to associated clients and stakeholders. Once data deduplication is activated, the background file system is continually and automated scanned and optimized.

The savings that can be made with deductions depend on the nature of the data set, including how many files are replicated. For general purpose file shares, usual savings mean between 50% and 60%. Savings range from 30%−50% in user documents and 70%−80% for software development datasets within shareholdings.

16.3.2.1 Enabling data deduplication

The "*Enable-FsxDedup*" command allows data duplication selection on Windows server as shown in the following:

C:\Users\Admin > Invoke Enable-FsxDedup −ComputerName "Prasad" −ConfigurationName RemoteAdmin -ScriptBlock {Enable-FsxDedup}

Where ComputerName is the name of the computer which is used for data duplication, the ConfigurationName is the name of the server configured, and ScriptBlock is the parameter to set the data duplication is enabled on the selected server.

16.3.2.2 Setting up a schedule for data deduplication

To change the deduplication scheduled plan, the "*New-FSxDedupSchedule*" command is used. Consider the following used in AWS:

C:\Usesrs\Admin > Invoke New-FSxDedupSchedule -ComputerName "CustOptim" -Type Optimization -Days Sat, Sun -Start 10:00 -DurationHours 4

In the above example, the command New-FSxDedupSchedule is used to create a new plan or modifies the existing plan, Name is the name of the plan that is going to be scheduled, Type is a parameter used to optimize the data deduplication, Days is an argument specifying on which days the plan should run for the optimization schedule, Start is a value which says about from which time to start the schedule to be optimized, and the server stops the process for the specified working hours.

16.3.2.3 Retrieving the deduplication configuration

Use get-FsxDedupConfiguration to recover the data deduplication setup for a file.

C:\Usesrs\Admin > Invoke get_FsxDedup −ComputerName "Prasad" −ConfigurationName DupAdmin −ScriptBlcok {Get-FsxDeDeupConfig}

In the above step, it captures the server's configuration with the computer which is used for the deduplication of data and a few lines of the script are used to get the information.

Use the following command to display the volume of storage space saved from running deduplication

C:\Usesrs\Admin > Invoke get-FsxDedupStatus −ComputerName −"Prasad" ConfigurationName DupAdmin − ScriptBlock {Get-FSxDedupStatus} | select OptimizedFilesCount, OptimizedFilesSize, SavedSpace, OptimizedFilesSavingsRate

The following is the result of optimized file space

OptimizedFilesCount	OptimizedFilesSize	SavedSpace	OptimizedFilesSavingsRate
10578	31263649	27199375	87

16.3.2.4 Managing data deduplication

Use Amazon's Command Line Interface for accessing remotely to handle data deduplication on the file system on PowerShell.

The commands users can use for data deduplication are shown in Table 16−1.

16.3.3 Amazon FSx for Windows File Server

Amazon FSx offers Windows Servers, which remain completely maintained and provide backup to a complete file system managed by Windows. The features and performances of Amazon FSx for the Windows File Server are easily lifted and moved to the AWS Cloud.

Table 16–1 Deduplication of data commands and description used in AWS.

Data deduplication command	Description
Enable-FSxDedup	Allow file share data deduplication.
Disable-FSxDedup	Deactivates file-sharing data deduplication.
Get-FSxDedupConfiguration	Recovers configuration information for the deduplication of minimum file size and location, compressed setting, alienated categories of files and directories.
Set-FSxDedupConfiguration	Sets deduplication configuration setting for files and its data.
Get-FSxDedupStatus	Restores the status of the deduplication and provides read-only features that classify file system optimization cost-saving and profile, time intervals and finalization information over the last system files tasks.
Get-FSxDedupMetadata	Collects optimum metadata for data deduplication.
Update-FSxDedupStatus	Find the efficiently saved deduplicated data statistics.
Measure-FSxDedupFileMetadata	Tests and finds the available storage capacity which can extract while removing a cluster of directories in the file system. Documents often have pieces in which other files can be exchanged, and the deduplication algorithm determines which pieces are special and therefore are excluded.
Get-FSxDedupSchedule	Returns present schedule status of deduplication.
New-FSxDedupSchedule	Build a data deduplication plan and configure it.
Set-FSxDedupSchedule	Transforms in current data deduplication plans.
Remove-FSxDedupSchedule	Removes a deduplication agenda.
Get-FSxDedupJob	Provide status and data for all deduplicated jobs currently running or queued.
Stop-FSxDedupJob	Cancel a given data deduplication job(s).

Amazon FSx supports a wide variety of business Windows working loads with fully managed Microsoft Windows Server file storage. It also supports the network file system capabilities and the enterprise SMB protocol, which allows admission to the network file storage. This server has a native Windows operating system interoperability and competitive advantage, features, and reliable submillisecond response times, for enterprise application in the AWS cloud.

The code, programs, and tools used today by the development and administrators in Windows will continue to operate unchanged by file storage on Amazon FSx. Apps suitable for Amazon FSx include Windows apps, home directories and content organization, analysis of data, software development, and hypertext storage workloads.

It minimizes the operational complexity of files and storage volumes being set up and delivered as a complete service. Amazon FSx also patches, identifies, and fixes hardware faults, and manages backups on Windows applications. This file system supports creation of file systems, accessing files and folders, sharing of files, storing of data, backing up the data, and storage of volumes. This can be accessed through the DNS name and the performance of the storage is measured in GiB and throughput in MBps.

The server can compute file system instances with SMB protocol to share a specified directory or its subdirectories in Windows. The default file system can be decided by the Operating System at the time of installation, the file share system can be easily created and managed with the other GUI frameworks on Windows.

16.3.3.1 Security and data protection

Amazon FSx provides various safekeeping and passivity levels to guarantee the protection of the user's data. The key that is managed in the AWS Key Management Service is then spontaneously converted for resting (both file and backup) data, it also uses the Kerberos algorithm to encrypt data during transmission.

It also provides access control mechanisms for its file system completely by using Amazon Virtual Private Cloud (VPC) for various security clusters or groups. Amazon also assists various web services to provide API level mechanisms and access policies by IAM, the Ms AD authenticates all the files systems.

16.3.3.2 Availability and durability

This service manages and assures the data on two levels: availability and durability. An availability zone (AZ) is a storage area or place in AWS which assures a highly available service. A single AZ automatically detects and addresses the failures, where multiple AZs maintain duplication or proxy of the existing server configuration within the Region of AWS, to address the issue of any failure of the file server.

16.3.3.3 Managing file systems

Using PowerShell commands or a native interface of Windows GUI can administrate the file server file systems.

16.3.3.4 Price and performance flexibility

The AWS uses a pay-as-you-go method to avail any service. There are separate tariff rates for using storage devices like hard diskettes and SSD with regards to different regions. The user can have the flexibility to choose the type of storage device based on the need of the application.

16.3.3.5 Assumptions

Some assumptions are required here to make use of these services, such as EC2, VM, running in a VMW are environment and its supported types, AppStream, and WorkSpace.

16.3.4 Amazon Elastic File System

Amazon EFS is a service providing a network file system with the autoscale capability of on-site resources. It does not need to stop and pause any applications running on AWS Cloud while creating EFS. This service syncs with scale up or scale down automatically while files can be added or removed by estimating and accommodating the services to the end user without any overhead.

It has two storage classes: (1) standard storage class—a class of storage with limited storage of 5 GB free of cost, where the files can be accessed daily; and (2) infrequent access storage class—an optimized storage class with unlimited storage and paid service, the files are not accessed every day by enabling life cycle of EFS, and the cost is around $0.025/GB-month.

The data may be frequently or infrequently moved into or from the EFS depending on the customer and accessed from a common file system namespace. This may be typically 20% frequently used data and the remaining 80% is infrequently used data, the price is charged for only the files that are actively used and the cost is as low as $0.08/GB-month.

Amazon EFS provides highly scalable and parallel shared access to hundreds of thousands of instances of EC2, allowing the end user to achieve high capacitance utilization levels with consistently low latency levels for the applications.

Amazon EFS is perfect for supporting a wide range of applications related to home files and folders and business applications. For lifting and moving current business systems, customers may use the EFS for the AWS Cloud. Additional use cases include large data analysis, web servers and business intelligence, software and evaluations, news and entertainment workflows, database backups, and storage of containers.

It is a local service data management system for high quality and durability within and across multiple AZs. Amazon EC2 can use AZ, regional, and VPC file systems, while on-site servers can access the file system through either the AWS Direct Connect or the AWS VPN.

16.3.4.1 Benefits

- POSIX-compliant shared file storage
- Scalable performance
- Dynamic elasticity

- Fully managed
- Cost-effective
- Security and compliance

16.3.5 Amazon FSx for Lustre

The Amazon File Systems for Lustre renders launching and running of most common high-performance file systems with simple and low cost. It is used in many applications, such as video analytics, HPC, and ML algorithms, etc.

This is an open-source system built to store the applications quickly in order to maintain the computing speed. As the system is being used to access very large data processing with less cost, lesser transfer delay operates on hundreds of thousands of IOPS, and hundreds to thousand GBps of efficiency or throughput.

Amazon FSx now allows the worker to use Lustre file systems as a completely managed solution for any workflow that needs processing speed. It removes the conventional complexity with which Lustre cloud servers can be set up and managed, so that it will have a high-performance file system in minutes. It also provides multiple options for cost optimization.

It integrates with the S3 storage file system to process the large data values, and it projects S3 objects like files and directories to write changed data returns to the S3 bucket.

16.4 How to deduplicate

The data deduplication is performed in seven steps in AWS and is described as follows (https://aws.amazon.com/articles/).

Step 1: Create a File System

- Go to the AWS Console
- Choose Create File System
- Configure network access section, choose default VPC
- Select the AZs. Chose all default setting like subnets, IP address, and Security group
- Add the name of the file system, and optional tags
- Select Lifecycle Management
- Keep Bursting and General Purpose for performance and throughput
- Enable encryption
- Set File system policy and create a file system

Step 2: Send the configuration file to an EC2

- Create a key pair
- Launch EC2 Instance from types of Amazon Machine Images
- Configure Network and subnet details
- Add storage, tags

- Assign security group, protocol, port range, and accessing IP number
- Launch the instance

Step 3: Write Data for File Share

- Go to a text editor.
- Write the text "He is a Good person" in the editor.
- Save the contents to a shared location
- Find the text file that would like to share by File Explorer

Step 4: Backup File System

- Open the Amazon FSx console
- Choose to Create a backup
- Choose Backups

Step 5: Transfer Files Using DataSync

- Activate the environment by downloading and deploying the agent
- Configure the locations of source and destination
- Create a task and deploy that task
- Choose and executes the task to move contents of files from the source to the destination

Step 6: Clean Up Resources

- Connect EC2 instance
- Unmount file system
- Delete the file system
- Terminate the EC2 instance

Step 7: Amazon FSx File System Status

The data deduplication is performed dynamically in AWS by creating and managing file systems. Fig. 16–3 shows deduplication using various web services offered in AWS. Here in the process of deduplication, the data can be input from various sources outside of AWS like mobile, wearables, big data, etc. and on-premises backups and recovery systems. A VPC can

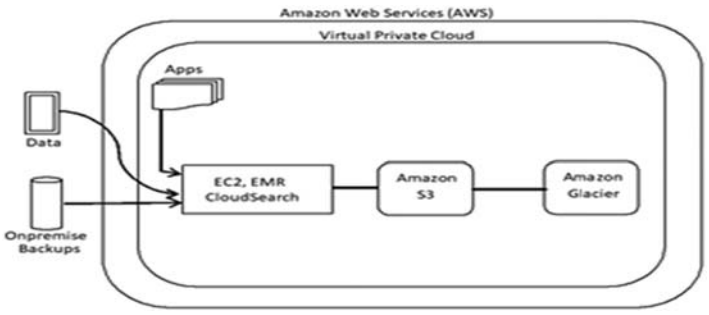

FIGURE 16–3 Data deduplication using file server system.

be created in which the file system is used to create backup data, as well as deduplication of data using Simple Storage Service (S3) in the form of streams. Once set up, the deduplication process automatically executes the deduplication process with the specified period. AWS provides a forum for the development of fault-tolerant software applications. The AWS framework allows users to build load-balancing systems with limited administrative interaction and investment capital.

16.5 Integrate and deduplicate datasets using AWS Lake Formation FindMatches

AWS Lake Formation FindMatches is a new ML transform that enables matching of records across different datasets as well as identification and removal of duplicate records, with little to no human intervention. FindMatches is part of Lake Formation, a new AWS service that helps to build a secure data lake in a few simple steps.

To use FindMatches, the customer does not have to write code or know how ML works. Here data does not have to include a unique identifier, nor must fields match exactly.

FindMatches helps with the following ways:

- Match customers: link and integrate customer records across different datasets, even where fields do not match exactly (for example, due to different name spellings, address differences, and missing or inaccurate data).
- Match products: match products across different vendor catalogs and SKUs. This can be done even when records do not share a common structure.
- Prevent fraud: identify potentially fraudulent accounts compared to previously known bad actors.
- Match other data: match addresses, movies, parts lists, etc. In general, if a human being could look at database rows and determine that they were a match, there is a good chance of matched data through FindMatches app.

16.6 Additional services and benefits

16.6.1 StorReduce

- StorReduce is a partner of AWS.
- StorReduce will minimize or delete the use of CAPEX to deduplicate hardware otherwise.
- When tape data is transferred to the cloud, the StorReduce system is easily integrated into the Amazon S3 API. Those data can therefore easily be accessed by all existing AWS services such as Amazon CloudSearch and Amazon EMR. This setup is difficult with deduplication offers on-site.
- With the customer data growing, StorReduce can scale rapidly to meet its needs without purchasing new hardware.

16.6.2 Commvault

- It provides a robust business intelligence framework to transfer, control, and use data across files, programs, databases, hypervisors, and the cloud with Commvault support for AWS.
- This connects substantial data security, backup, retrieval, strategic planning, and eDiscovery functionality on one unified platform—all fully integrated with AWS services.

16.7 Comparison of Cloud backup services with AWS, GCP, Azure

AWS provides various services on-demand for public and private customer entities. It allows customers to choose the services from IaaS such as computational power (EC2), data storage (S3), etc. on a pay-per-usage basis, platform-as-a-service (PaaS), and software-as-a-service (SaaS). It also promotes enterprise solutions in scalable services, and the sophisticated automated backup service is also enabled for all the AWS cloud services for individual and enterprise users.

More than 50 cloud computing services are delivered by GCP. The infrastructure-as-a-service (IaaS), PaaS, SaaS, GCP storage, and data and ML services are supported. GCP is extremely customizable and offers open-source support. The strength of GCP is very easy to integrate with third-party cloud service providers with a low-cost solution.

The cloud infrastructure software Microsoft Azure provides business-level applications for IaaS, SaaS, and PaaS. For the complete production cycle, Azure provides customers with features such as the popular VM solution for scalable and quick computing power. Microsoft Azure strengths are the fully integrated backup and recovery system for all Azure tools that are simple and user-friendly.

16.8 Key terms and definitions

Policy: Policy is the systematic method of instructions and acceptable results for decision-making. It is a statement of intent and it is applied as protocol or procedure.

Reliable: A trustworthy service informs the customer if delivery fails, while a "nontrustworthy" service fails to notify the customer.

Service-oriented architecture (SOA): The architecture of the SOA is a web service layout application framework where certain applications offer other modules through a communication protocol like SOAP, normally through the network. Service orientation standards are irrelevant to any manufacturer, company, or technology.

Virtual machine: A VM is a simulation of a certain computer device in the computation process. Digital machines run similarly to a real or presumed computer's computer architecture and functions, and they can be complemented by specialist hardware, software, or a

combination of the two. Digital machines are divided into two major groups, based on their use and the degree to which each actual computer corresponds.

Web service: A web application is a way in which two smart devices interact across a web. It is a software feature that is delivered at a web address with the service that is also useful in a machine framework.

References

Aggarwal, B., et al. (April 2010). Endre: An end-system Redundancy Elimination service for enterprises. In *Proceeding of 7th USENIX Conference on Network System Design Implementation*, 14−28.

Al-Kiswany, S., et al. (June 2008). Storegpu: Exploiting graphics processing units to accelerate distributed storage systems. In *Proceedings of 17th International Symposium on High-Performance Distributed Computing*, 165−174.

Al-Kiswany, S., et al. (June 2011). VMFlock: Virtual machine co-migration for the cloud. In *Proceedings of 20th International Symposium on High-Performance Distributed Computing*, 159−170.

Bhagwat, D., et al. (September 2006). Providing high reliability in a minimum redundancy archival storage system. In *Proceedings 14th IEEE International Symposium on Modelling, Analysis, Simulation Computing Telecommunication Systems*, 413−421.

Bhatotia, P., Rodrigues, R. & Verma, A. (2012). Shredder: GPU-accelerated incremental storage and computation. In *Proceedings of 10th USENIX Conference File Storage Technologies*, 1−15.

Bobbarjung, D., Dubnicki, C. & Jagannathan, S. (May 2006). Fingerdiff: Improved duplicate elimination in storage systems. In *Proceedings Mass Storage Systems Technology*, 1−5.

Bolosky W. J., et al. (August 2000). Single instance storage in Windows 2000. In *Proceedings of the 4th USENIX Windows Systems Symposium*, 13−24.

Broder A. (1993). Some applications of Rabin's fingerprinting method. In *Sequences II: Methods Communication in Security Computing Science*, 1−10.

Cloud Computing <https://en.wikipedia.org/wiki/Cloud/computing/> Accessed June 2020.

DuBois L., et al. (March 2011). Key considerations as deduplication evolves into primary storage. In *White Paper 223310*.

El-Shimi A., et al. (2012). Primary data deduplication-large scale study and system design. In *Proceeding of Conference USENIX*, 1−12.

Eshghi, K., & Tang, H. K. (2005). *A framework for analysing and improving content-based chunking algorithms*. Palo Alto, CA: Hewlett Packard Lab., Tech. Rep. HPL-2005-30(R.1).

Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, the biggest growth in the far east. In *IDC Analyse the Future*.

Guo F., & Efstathopoulos, P. (June 2011). Building a high-performance deduplication system. In *Proceedings USENIX Conference USENIX Annual Technologies Conference*, 1−14.

Kim C., et al. (2012). GHOST: GPGPU-offloaded high-performance storage I/O deduplication for the primary storage system. In *Proceeding International Workshop Programming Models Application Multicores Manycores*, 17−26.

Kruus E., Ungureanu, C. & Dubnicki, C. (February 2010). Bimodal content defined chunking for backup streams. In *Proceedings of 7th USENIX Conference on File Storage Technologies*, 1−14.

Li, X., Lillibridge, M., & Uysal, M. (2011). Reliability analysis of deduplicated and erasure-coded storage. *ACM SIGMETRICS Performance Evaluation Revision*, *38*(3), 4−9.

Liu, C., et al. (December 2009). A novel optimization method to improve de-duplication storage system performance. In *Proceedings of the 15th International Conference on Parallel Distributing Systems*, 228−235.

Lu, G., Y. Jin, & Du, D. H. (August 2010). Frequency-based chunking for data de-duplication. In *Proceedings of IEEE International Symposium Model Analysis Simulation Computing Telecommunication Systems*, 287−296.

Meister, D., et al. (June 2012). A study on data deduplication in HPC storage systems. In *Proceeding of International Conference on High-Performance Computing Network Storage Analysis*, 1−11.

Meyer, D. & W. Bolosky (February 2011) A study of practical deduplication. In *Proceedings USENIX Conference on File Storage Technologies, San Jose, CA, USA*, 229−241.

Muthitacharoen, A., et al. (October 2001) A low-bandwidth network file system. In *Proceedings of ACM Symposium on Operating Systems Principles*, 1−14.

Paulo, J., & Pereira, J. (2014). A survey and classification of storage deduplication systems. *ACM Computing Survey*, *47*(1), 11.

Policroniades, C., & Pratt, I. (June 2004). Alternatives for detecting redundancy in storage systems data. In *Proceedings of USENIX Annual Technologies Conference General Track*, 73−86.

Quinlan, S., & Dorward, S. (January 2002). Venti: A new approach to archival storage. In *Proceedings of USENIX Conference on File Storage Technologies*, 1−13.

Rabin, M. O. (1981). Fingerprinting by random polynomials. In Centre for research in computing technology. Aiken Computing Lab.

Romański, B., et al. (May 2011). Anchor-driven sub chunk deduplication. In *Proceedings of 4th Annual Intelligent System Storage Conference*, 1−13.

Shilane, P., et al. (February 2012). WAN optimized replication of backup datasets using stream-informed delta compression. In *Proceeding of 10th USENIX Conference on File Storage Technologies*, 49−64.

Teodosiu, D., et al. (2006). Optimizing file replication over limited bandwidth networks using remote differential compression. *Microsoft Research TR-2006-157*.

The data deluge (February 2010). *The Economist*, <http://www.economist.com/node/15579717> Accessed May 2020.

Wallace, G., et al. (February 2012). Characteristics of backup workloads in production systems. In *Proceedings of 10th USENIX Conference on File Storage Technologies*, 33−48.

Wen, X., et al. (September 2016). A comprehensive study of the past, present, and future of data deduplication. In *Proceedings of the IEEE*, Available from: https://doi.org/10.1109/JPROC.2016.2571298.

Xia, W., et al. (June. 2012). P-dedupe: Exploiting parallelism in data deduplication system. In *Proceedings of 7th International Conference on Network Architectures and Storage*, 338−347.

Xia, W., et al. (2014). Ddelta: A deduplication inspired fast delta compression approach. *Performance Evaluation*, *79*, 258−272.

Yu, C., Zhang, C., Mao, Y., & Li, F. (June 2015). "Leap-based content defined chunking—Theory and implementation. In *Proceedings of 31st Symposium Mass Storage Systems Technology*, 1−12.

Zhou, B., & Wen, J. (October 2013). Hysteresis re-chunking based metadata harnessing deduplication of disk images. In *Proceedings of 42nd International Conference on Parallel Process.*, 389−398.

Zhu, B., Li, K., & Patterson, R. H. (February 2008). Avoiding the disk bottleneck in the data domain deduplication file system. In *Proceedings of 6th USENIX Conference on File Storage Technologies*, vol. 8, 1−14.