Protection of Data with Data Domain Cloud Tier

[1]T S Vinutha, [2]Sandeep Varma N, [3]Raghuprasad T R

[1]M.Tech Student, [2]Assistant Professor, [3]Senior Manager
Department of Information Science and Engineering
BMS College of Engineering, Bangalore, INDIA

***Abstract*: Data protection plays a vital role in today's IT business. No matter where the data lives, it needs to be protected and as the data protection continues to evolve, the amount and variation of data that need to be protected is also evolving. Today's businesses are faced with explosive data growth, strict retention policies and the pressure to reduce costs. Low-cost, high-capacity object storage in the public, private and hybrid cloud is the way to address these challenges.**
DELL EMC Data Domain system which has deduplication storage systems can solve many of the challenges that companies experience traditional backup, recovery, and replication processes.

Customers are looking for a low-cost storage solution for their long-term retention protection copies. They are considering moving these copies from the DDR to an object storage provided by the cloud service provider, either on premise or off-premise.

Data Domain cloud tier enables simple, cost-effective tiering of data to public, private and hybrid cloud for long term retention. Some of the backup data likely needs to be retained for extended period of time for governance or compliance purposes.

***Index Terms:* Data Protection, Long-term Retention, Cloud Tier**

I. INTRODUCTION

Cloud is just like any other storage device for storing and accessing huge volume of data whenever we want, there are a lot of advantages by using it, one of the extraordinary advantage for cloud user with this is no need of having more computing devices to get or upload information into the cloud, so no need of feeling burden about space utilization and security. Moreover, most of larger and smaller companies are using cloud to make their daily transactions. Because of its advantages and usage by multinational companies it is being considered as future generation information technology.

Long term retention requirements has continued to grow throughout all the industries. Unstructured data, whether its active or archive continues to accumulate at faster rates and it must be kept in readily accessible formats. The data protection tools must be protected from getting overwhelmed by enormous capacity requirements which is generated in today's data centers, and technologies which enables data to be tiered from primary backup to secondary long-term retention are becoming more attractive.
With the ascent of public cloud computing, companies are estimating how they can leverage public cloud resources to become more agile and competitive in market place.

Irrespective of where or how the data are stored, data protection ensures to maintain custody and control over the data assets. Data protection is a key tenet of IT strategy that has withstood over a long period of time.

DELL EMC is the world leader in data protection solutions. In this new era of cloud computing, DELL EMC's software defined data protection solutions will be able to help you gain control of your journey to public cloud, giving you the same level of comfort, control and protection to confidently embark on your journey to any cloud.

II. DATA DOMAIN OVERVIEW

Data Domain is the world's fastest purpose built backup appliance providing top of class streaming deduplication. As such, Data Domain is able to ingest, at full network speeds, from multiple backup sources while providing industry proven storage efficiency. As the product has continued to develop, Data Domain has been structured to take advantage of tiered data storage technologies, such as very large multi-terabyte SATA drive trays, or SSD enabled fast ingest trays. This capability to tier data has been enhanced to utilize cloud technologies to enable more efficient long-term retention.

Data Domain system is designed as a very reliable online system for backups and archive data. As new backups are added to the system, old backups are aged out. Such removals are normally done under the control of backup or archive software based on the configured retention period.

When backup software expires or deletes an old backup from a Data Domain system, the space on the Data Domain system becomes available only after the Data Domain system cleans the data of the expired backups from disk. A good way to manage space on a Data Domain system is to retain as many online backups as possible with some empty space (about 20% of total space available) to comfortably accommodate backups until the next scheduled cleaning, which runs once a week by default.

All Data Domain systems (DDRs) have an active tier - this is the default storage/tier which exists when the file system is created. There are certain models/features that can support additional tiers, for example:

Extended Retention: Allows additional locally attached storage to be added as an 'archive tier' for long term retention of a subset of data

Long Term Retention/Cloud tier: Allows object storage from a supported cloud provider to be added for long term retention of a subset of data in the cloud.

III. Data Domain Cloud Tier

DD Cloud Tier is a native feature of DD OS 6.0 (or higher) for moving data from the active tier to low-cost, high-capacity object storage in the public, private, or hybrid cloud for long-term retention. Only unique, deduplicated data is sent from the Data Domain system to the cloud or retrieved from the cloud. This ensures that the data being sent to the cloud occupies as little space as possible. DD Cloud Tier is best suited for long-term storage of infrequently accessed data that is being held for compliance, regulatory, and governance purposes.

Figure1. Overview of Data Domain Cloud Tier

Figure1. Represents the overview of the Data Domain Cloud Tier. DD Cloud Tier is managed using a single Data Domain namespace. There is no separate cloud gateway or virtual appliance required. Data movement is supported by the native Data Domain policy management framework. Conceptually, the cloud storage is treated as an additional storage tier (DD Cloud Tier) attached to the Data Domain system, and data is moved between tiers as needed. File system metadata associated with the data stored in the cloud is maintained in local storage, and it is also mirrored to the cloud. The metadata in the cloud tier shelf of the local storage facilitates operations such as deduplication, cleaning, Fast Copy and replication. This local storage is divided into cloud units for manageability.

The primary value of the Data Domain Cloud Tier is the ability to extend the usable life of the active tier of a Data Domain backup target by moving aged unique data segments on to a secure public, private and hybrid cloud tier. This provides a recoverable long term archival content store accessible by the Data Domain which enforces the retention policies of the business.

IV. Literature Survey

Magnetic tape was one of the earliest methods for storing data. Originally it was a great choice because other methods for long-term storage of data were not really viable[7]. Memory was extremely expensive and required constant power, disks were very large, had very low storage density (and were extremely expensive and not that reliable) and other methods including flexible disks and paper tape were too limiting. Backing up and archiving data to magnetic tape was the obvious choice.

In addition to many different proprietary tape solutions that were available, companies also began to adapt consumer tape technologies such as digital cassettes and helical scan technologies for backup and archive market[8].

Tape started feeling the pressure from competing technologies. For a brief time optical disc libraries looked like a good choice to replace tape, due to reasonable initial capacity points, automation with juke boxes, and the promise of long life[1].

In parallel there were huge advances in hard disk technology. Form factors reduced quickly, density increased dramatically, reliability improved by orders of magnitude and $/MB decreased to levels competitive with tape[3].

Tape has been the core medium for storing the archival data. For years, tape was the only thing which was used for storage of data. Servers were expensive. The disks arrays required for suitable data protection were also too expensive for archive purposes[3].

However, as the storage requirements continue to grow, industry revenue has not grown accordingly. Effectively the industry is challenged to “do more with less”[9].

The cloud technology has changed everything. The economy of scale found in the public cloud has changed the economics of multi-petabyte archives[8]. The shared cost of infrastructure, compute resources, power, cooling is generally much lower in the public cloud compared the operational costs of maintaining a tape based archive. The public cloud brings the promise of “limitless” scale with minimal engineering and technical support requirements.

The public cloud promises far more than lower storage costs. By moving data to the public cloud, collaboration between business units operating in different geographies of the world is simplified[9].

DELL EMC's Elastic cloud storage (ECS) is a multi-purpose cloud storage platform with enterprise-grade capabilities that the public cloud does not provide. ECS is built from the ground up as a software-defined storage solution that can also be deployed on a turn-key commodity server infrastructure. Like the public cloud, ECS can span multiple geographic locations with a single storage infrastructure. ECS can distribute parity data across multiple locations in order to provide enhanced data availability and resiliency while reducing the WAN traffic between data center location[11].

Cloud tiering can be used for long-term data retention in public, private (on-premises) or hybrid clouds such as Dell EMC Elastic Cloud Storage (ECS) and Virtustream Storage Cloud[8].

V. Design and Analysis

While many customers are interested in transitioning their data and computation from on-premises to the cloud, this will be a long process as customers gain experience with cloud environments.

Data Domain Cloud Tier is managed by a single Data Domain namespace. There is no separate cloud gateway or virtual appliance required. Data movement is supported by the native Data Domain policy management framework. Figure 2. explains the detailed design of how the data protection is done using the Data Domain Cloud Tier.

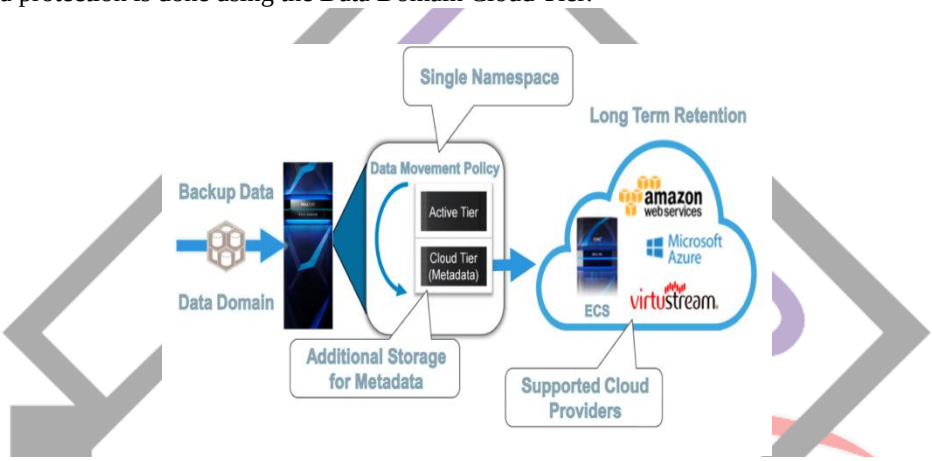

Figure 2. Protection with Data Domain Cloud Tier

With DD OS 6.0, supported cloud storage includes Dell EMC Elastic Cloud Storage, Virtustream, Amazon Web Services, and Microsoft Azure. Additional storage for metadata is required to support the cloud tier. Metadata is used by deduplication, cleaning, and replication operations.

Data Domain Cloud Tier is supported for DDOS 6.0 release and above. Data Domain consists of two tier namely active tier and cloud tier in hardware.

In cloud tier we support two cloud units. cloud tier can have 2 times the maximum capacity of the active tier. For example, if active tier has 1 TIB capacity then cloud tier supports 2 TIB capacity. Each cloud unit can write to a separate supported cloud object store. Each cloud unit can support upto the max active tier usable capacity.

Data is moved to the cloud using the process known as data-movement. Data movement moves the data from active tier to the cloud tier based on data-movement policy.

Every file has a meta-data. Meta-data of the files which is getting moved is stored on active tier and data and metadata is moved into the cloud. There is no multiple data-movement policy set on single mtree. Data moved to particular cloud unit depends on data-movement policy. And only the unique data is sent to the cloud. This is done using the deduplication. Cloud tier has the ability to compress and encrypt data before sending to the cloud.

VI. Architecture of Data Domain Cloud Tier

A file in our active tier is represented by a Merkel tree(MTree) with user data as variable sized chunks. The SHA1 fingerprints of those chunks are grouped together to form the next higher level of the tree. SHA1 fingerprints of chunks are grouped together to form higher level chunks and this continues and represents the entire file. Figure 3. represents the architecture of the cloud tier.

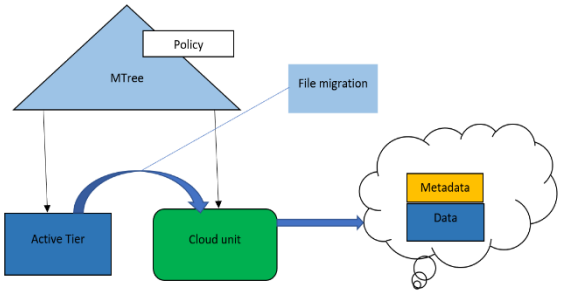

Figure 3. Architecture of DD Cloud Tier

Deduplication happens when different files refer to the same level chunks. As an example, if two files are exactly the same, they would have the same chunk levels. But if two files only partially overlap in content, then some branches of the tree will be identical while other branches will have different fingerprints.

We have end to end verification that verifies the data received from clients is correctly written to disk. The user data and internal structures are read from disk and verified against their checksums soon after they are stored on disk platter. The verification is done by various filesystem layers.

As files are migrated to cloud tier, the respective chunks go through a deduplication process relative to chunks already transferred to the cloud tier. After deduplication, the unique chunks are written to the containers, and these containers are written to object storage in the cloud. Because it is slower to access data in the cloud, we store various types of metadata on local storage.

VII. Conclusion

The task of backing up data and retaining data for long term retention is not easy as it used to be. Organizations have been catapulted into a new and constantly changing array of choices. All are seemingly viable, making it difficult to determine the best solution as part of business continuity strategy. The public cloud presents its own set of challenges including how to avoid vendor lock-in and how best to meet the compliance and governance concerns. With Data Domain cloud tier, customer can move their data to the cloud confidently for long term retention purposes and recall their data whenever needed.

References

[1] Data Domain overview [online]. Available: https://www.emc.com/collateral/white-papers/h11534-why-datadomain?isKorealPage=false&domainUrlForCanonical=https%3A%2Fwww.emc.com
[2] Data Domain Invulnerability Architecture [online] Available: https://www.emc.com/collateral/software/White-papers/h7219-data-domaindatainvularchwp.pdf?isKoreaPage=false&domainUrlForCanonical=https%3A%2Fwww.emc.com
[3] Best way to archive data to cloud Available: https://storageswiss.com/2017/12/05/best-way-to-archive-data-to-the-cloud/
[4] Data Domain Product Overview Available: www.enterprisestorageforum.com/.../dell-emc-data-domain-product-overview/
[5] Data Domain Backup Storage Available: https://www.dellemc.com/en-us/data-protection/data-domain-backup-storage.html
[6] DELL EMC ECS Monetizing Archives Available: https://dellemcevents.com/uploads/Dell-EMC-ECS-Monetizing-ME-Content-Archives.pdf
[7] DELL EMC ECS Monetizing Archives Available: https://dellemcevents.com/uploads/Dell-EMC-ECS-Monetizing-ME-Content-Archives.pdf
[8] Data Domain Cloud Tier Available: https://www.emc.com/collateral/brochure/h15266-emc-data-domain-cloud-tier-ecs.pdf
[9] Cloud Computing [online] Available: https://www.synopsys.com/software-integrity/resources/knowledge-database/cloud-computing.html
[10] Aws cloud services [online] Available: https://s3.amazonaws.com/academia.edu.documents/
[11] Amazon Elastic Compute Cloud (Amazon EC2), http://aws.amazon.com/ec2/, 2009
[12] Dr. Ramalingam Sugumar, and K. Raja, "Enhanced Data Security Methodology for Cloud Computing Environment", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2018.
[13] Vivek Paul, Supriya Pandita, and Prof. Meera Randiva, "Cloud Computing Review", International Research Journal of Engineering and Technology, 2018.
[14] Muhammad Jawad, Muhammad Bilal Quresh, and Usman Khan, " A robust optimization Technique for Energy Cost Minimization of Cloud Data Centers", IEEE Transactions on Cloud Computing, 2018.
[15] Shyam Patidar, Dheeraj Rane, and Pritesh Jain, "A survey paper on Cloud Computing", Second International Conference on Advanced Computing & Communication Technologies, 2012.

[16] DP Mishra, P Amitha, KS Reddy, N Jami and MD Prasad, "Encrypted Data Management with Deduplication in Cloud Computing", European Journal of Advances in Engineering and Technology, 2018.