CLOUD DATA STORAGE STRATEGY

What should you consider while putting together your strategy for data storage on cloud ?

www.hexaware.com

Highlights

- Cloud storage strategy that outlines what it wishes to achieve in measurable terms
- Parameters to consider while devising a data strategy like Data Security and Performance & Scalability
- File Storage or Block Storage or Object Storage? How to determine the best fit based on the need ?
- What does AWS, Azure and Google have to offer in Cloud Storage Space ?
- Role of Data Mobility in building a multi-cloud storage strategy

Table of Contents

1 Executive Summary

There was a time when enterprises were skeptical about moving data to cloud and would cite security and data mobility as reasons. However, as cloud platforms have matured multifold in terms of data storage, more customers are now looking at cloud for their primary and secondary storage. It has become mandatory in today's world to create a robust Data Storage Strategy to ensure success and uninterrupted business at the end of it all.

Ever-growing data storage volumes resulting in higher cost and complexity have made Data Storage Management a critical component, while adopting cloud. Data storage requirements need to be thoroughly examined and planed for optimum performance and scalability, while maintaining the required security compliance. Organizations are constantly exploring ways to reduce cloud storage CAPEX and OPEX and gain maximum ROI as soon as possible. At the same time, proper planning and execution of strategies play a vital role in successfully achieving the goal, failing to do so can turn out to be a costly mistake.

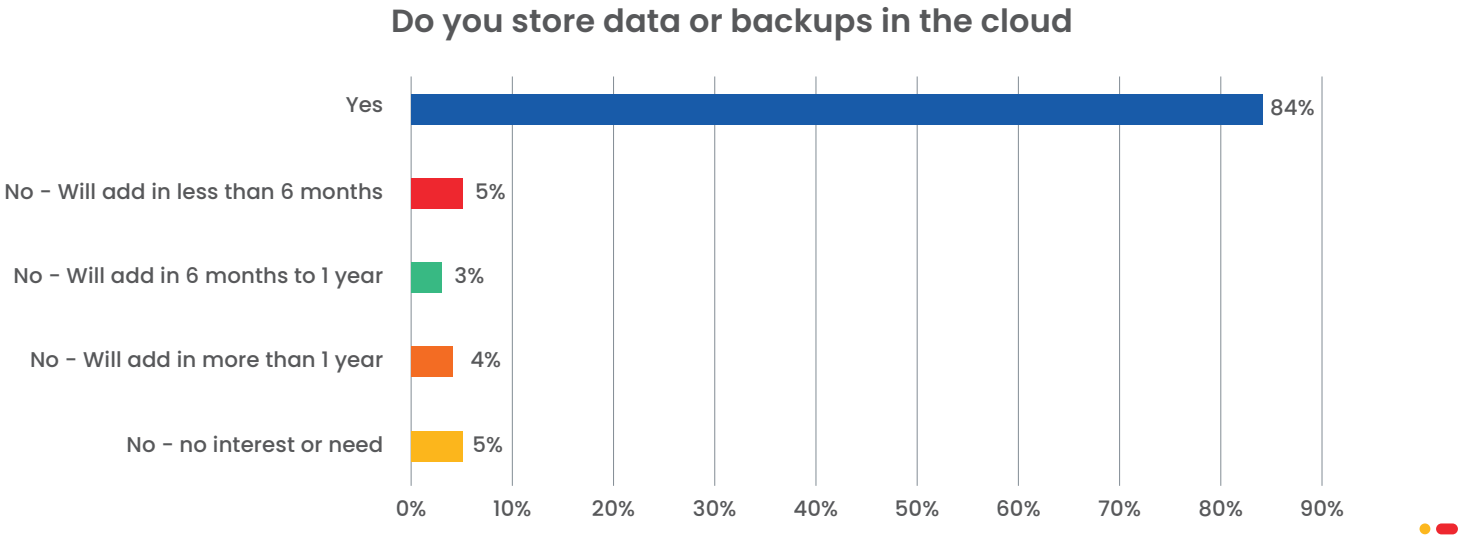
Handling data storage in a multi-cloud environment is one of the biggest challenges customers face today in terms of both data management and cost optimization as higher data transfer rates lead to higher cost incurrence. We will look at the available options of data storage and replication in a multi-cloud setup and how to optimize data mobility to reduce cost.

2 What should be your Data Strategy

Any organization migrating to the cloud should ideally devise a cloud strategy that outlines how the cloud push will achieve the agreed business target and how successful it is. Similarly, data management and storage are major parts of the cloud strategy. This is especially important since businesses are growing their off-premises data storage facilities at a faster rate than ever, and often without much concern for the core elements— the scale, the security, and the robustness. A survey by Unitrends states that participant organizations, who had data storage requirements of about 100TB, adopted cloud for data storage and such need has risen from 11% in 2016 to 31% in 2019 with no real change in the mix of small/medium/large enterprises across the years of survey. Considering this growth rate of all data is stored in the cloud, we must be prepared for an exponential increase in enterprises wanting to adopt cloud storage very soon.

2.1 Devise a Data Strategy

A research survey conducted by Unitrends from 2019 indicates that, overall, 84% of organizations are using the cloud for data storage.



This number will increase in the coming years. As a result, organizations must regard storage as a distinct priority and devise a concrete data strategy. This includes developing an in-depth understanding of the scale of data, type of data, reliability of data, security of data and more. In short, data strategy is a prerequisite rather than a complement to a storage strategy. A data strategy is a vision of how a company will collect, store, manage, share and use data and clearly articulates how data will enable and inspire the business strategy. Basically, it does set the foundation of everything an organization does with its data and therefore, should be specific and actionable. As any strategy it will be a living thing, flexible and adjusting to change in the business strategy or regulation.

In general. a data strategy should,

- Determine what data activities are required to meet the business objectives
- Define the plan and timeline to complete those data activities
- Determine the cultural, architectural, technological and operational changes to be adopted to maximize the value
- Define a roadmap to accomplish the changes required in the company and the timeframe to implement the required data activities
- Analyze the cost implications and define the ROI
- Analyze and define the potential of data monetization

2.2 Data Security

With more and more organizations transferring their mission critical applications and data to the cloud, the already crucial security is becoming even more important. No matter what's the state of data (in rest, motion, or being used) in the cloud, it does not diminish the overall risk of a security incident (data being accessed, modified or stolen). Sometimes the incident can be the result of sheer negligence on the part of the employees, service providers, etc. At other times, cloud data may be compromised because of the vulnerability of the end device that's interacting with it. An example would be employees using weak passwords for their accounts. Applying a security lens to the data strategy and its execution in real time is due. Securing business critical or regulated data must be protected no matter what storage type is being used. The guidance below lists some of the most important activities to secure your data when using cloud storage.

- **Ensure data encryption**
  This is an effective way to ensure sensitive information remains uncompromised in the event of a data breach. Encryption is not only required for data at rest but also for data in motion.

- **Implement strong authentication rules and policies**
  E.g. IAM, Multi-factor Authentication and the requirement to have strong passwords. Always grant least possible access to users.

- **Train employees and create awareness on security breaches**
  Regularly brief and train all employees on the typical and latest approaches that criminals use to gain access. Examples are phishing or the USB stick in the parking lot.

- **Breach drills**
  Conduct breach drills simulating attacks and establish processes in case of an incident. Train employees on the protocols of handling such events and how to avoid them

- **Understand how cloud storage works**
  There is a lot of choice in the storage service market. To decide on the most secure while best fitting and economic services, you need to understand the options and their policies, technologies, applications and controls.

2.3 Performance and Scalability

Today, customer's expectation is about instant access and with the best possible performance. Choosing the right performance parameters is important to measure the KPIs that really matter. Performance parameters for data storage are speed, availability and scalability.

To accomplish the best performance, you need possible measures to reduce latency in your storage systems and how they are accessed, though may not be critical for archival type of storage. Look at implementing storage systems that allow for the least latency; as a generic rule, the closer the data is to the user, the faster it can be accessed. In the later part of this paper we shall analyze the various options available with data storage and their impact on performance like Block (SSD/HDD) Vs Object Vs File. Define norms and definitions of acceptable limits for speed and availability, and devise policies and guidelines to act upon, when these are overrun. Having said that, analyzing below parameters will help choosing the right storage based on your performance and scalability needs,

- Implement a Content Delivery Network (CDN) to keep the data closer to the users geographically, thereby achieving high performance at scale. A CDN allows quick transfer of assets needed for loading static content including HTML pages, javascript files, stylesheets, images, and videos.

- Carefully choose your storage type. Flash memory (SSD) has higher performance than a traditional HDD but can become very expensive at scale. When you are attaching a block storage like in AWS, you have an option to choose the type. Based on the level of performance at scale, go for the right type of drive. Also, choose low cost options for cold storage like S3 in case of AWS.

- Isolate and fix application and infrastructure related bottlenecks that impact your storage performance and scalability. Use the tools that are available with the cloud platform like AWS Cloud Watch, Azure Monitor and Google Stackdriver to diagnose potential bottlenecks within the application stack.

- Optimize caching – a successful cache results in a high hit rate, implements controls such as TTL (time to live) to expire the data cached and cache new data. For high availability, you need to consider in-memory caching through in-memory engines like Redis.

- Determine the network throughput requirements to match the performance and scalability goals. Measure the network's impact on your workload and analyze your workload's networking requirements to understand how the network impacts overall performance and scalability. Also choose appropriately sized dedicated connectivity or VPN for hybrid workloads.

2.4 Plan Ahead

Think cloud storage and systems not just for today, but for many more years to come, with the reliance on cloud only growing multifold. Five years back, cloud storage use cases were more or less restricted to backup and cold storage but today cloud storage has become mainstream. Over the past five years, a range of use cases, such as global content distribution, backup, disaster recovery and data analytics have become reality. It's more about data mobility than just storing it, we will see more about this topic later in this paper. Even concerns like security, privacy and the threat of provider lock-in are much less of an obstacle nowadays. The reason for this is that cloud storage offerings have matured and companies have gained experience and confidence in secondary and primary storage deployments on the cloud. Below is an expert opinion extract from techtarget.com on their 5 years outlook on Data Storage on Cloud.

What features and capabilities do you expect to see added around cloud storage in the next five years ?

- "Enterprise-style data services and service-level guarantees to suit applications and businesses not born in cloud" – Alastair Cooke, Consultant.
- "Intelligent computation and planning engines will decide where both compute processing and its required data will need to live and migrate toward dynamically" – Mike Matchett, Consultant, Small World Big Data
- "More competitive services that don't charge a flat rate for storage used. More public cloud providers will offer built-in backup with their services" – George Crump, Founder and President, Storage Switzerland
- "Simpler, more transparent on-premises-to-cloud and cloud-cloud data migration and ability to reduce the cost of data movement out of the cloud. Standards that enable a unified framework for managing and monitoring hybrid and multi-cloud storage" – Jeff Byrne and Jeff Kato, Senior Analysts, Taneja Group.
- "Parallel file system and Scaling out high performance block storage based on storage class, memory and NVMe" – Marc Staimer, Founder, Dragon Slayer Consulting
- "More compatibility with on-premises to enable different types of hybrid storage environments under consistent workflows and processes. Automated security and protection features become common alongside cloud storage offerings at attractive price points or even free as a value add" – Deepak Mohan, Analyst, IDC

2.5 Operational Considerations

There are also storage-related operational challenges. Storage, unlike compute, degrades over time. Data storage ultimately is connected to a physical appliance, and as it is written and rewritten, its capacity degrades. There can also be physical contaminants that cause data degradation. If you're running a dynamic, containerized application in which clusters detach and reattach from storage resources, every attachment and detachment point is an opportunity for something to go wrong. A data storage strategy isn't something that can be handled once and then ignored— storage resources need to be continuously monitored for cost-effectiveness, performance, and security vulnerabilities.

Cloud storage monitoring is critical to track multiple metrics simultaneously, monitoring storage resources and processes that are provisioned across virtual machines, services, databases and applications. Required monitoring needs to be configured to track performance metrics and to focus on any type of disruption that may occur on the capacity available for storage. Below are the various native monitoring solutions available across cloud platforms which also cater to cloud storage monitoring.

AWS

AWS CloudWatch provides you with data and actionable insights to monitor your S3 and AWS Storage Gateway and respond to system-wide performance changes, optimize resource utilization, and get a unified view of operational health.

Azure

Azure Storage creates monitoring data by using Azure Monitor, which is a full-stack monitoring service in Azure. Azure Monitor provides a complete set of features to monitor your Azure resources and resources in other clouds and on-premises.

GCP

Using Google Cloud Storage Monitoring you can proactively monitor storage buckets in real time, troubleshoot faults easily, get intelligent alerts and analyze historical data aggregated over a period.

Now that we have a fair idea of the different parameters to consider while you put together a data storage strategy, let's have a look at what data storage types are available on the cloud and what they are called across the different cloud platforms.

3 File Storage or Object Storage or Block Storage

3.1 The Pros and Cons

Whether you're using AWS, Azure, or Google Cloud, the advantages and drawbacks to each type of storage are similar across platforms.

Parameter	File Storage	Block Storage	Object Storage
How it works	Operates like the digital version of a file cabinet, in the same way, you store documents on your personal computer—organized in a logical hierarchy.	Chops data into blocks, and spreads those blocks strategically around multiple environments. When needed, the data blocks are identified by a unique identifier and reassembled.	The data is stored as 'objects' and multiple servers and clients can connect to the same object storage container, using its web address. Object storage can handle detailed metadata and it scales easily.
Cost and Scalability	Cost is low but difficult to scale, can scale potentially to millions of files	Can be expensive but scalable (need to choose Elasticity)	Highly scalable and cost-effective but objects cannot be modified, so ideal for cold storage

Parameter	File Storage	Block Storage	Object Storage
Performance	Performs best fo smaller files	High performing block storages are available	Performs best for big content and higher stream throughput
Analytics	Limited number of set metadata tags	Not ideally used for analytics	Customizable metadata and not limited to number of tags
Geography	Data can be stored across multiple regions	Used only for local storage	Used only for local storage

3.2 Storage Across Cloud Platforms

Below is the naming convention followed for the storage services offered across AWS, Azure and GCP

Parameter	File Storage	Block Storage	Object Storage
AWS	Elastic File System (EFS)	Simple Storage Service (S3)	Elastic Block Storage (EBS)
Azure	Files	Disk Storage	Blob Storage
GCP	File Storage	Block Storage	Object/Blob Storage

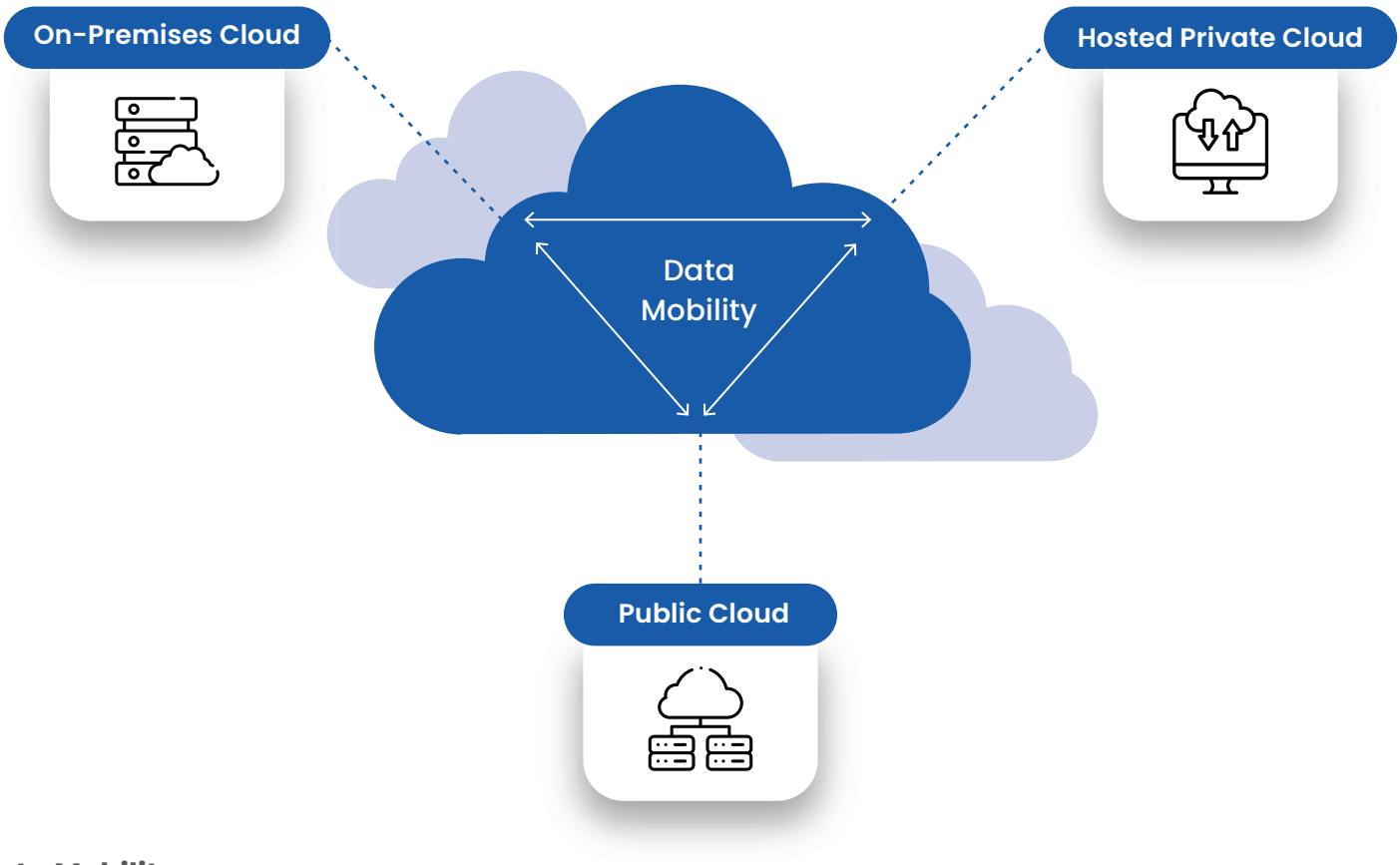
Supplier-integrated services have also started to emerge, with Azure NetApp being the most prominent player in the market. It provides the benefits of NetApp ONTAP technology, integrated directly into Azure using Azure application programming interfaces (APIs) and security capabilities. Supplier-integrated solutions are usually more feature-rich and higher performing than native services. Other platforms have also started to deliver similar services using ONTAP technology.

Also Cloud Platforms today offer High Availability (HA) options across their services including storage. HA means that systems continue to operate even when critical components fail, with built-in resiliency to seamlessly recover from failure.

- Amazon storage services, such as S3, EFS and EBS, provide built-in high availability options. S3 and EFS automatically store data across different Availability Zones, while EBS enables deployment of snapshots to different AZs.
- Azure features Zone-redundant storage (ZRS) which copies your data synchronously across three Azure availability zones in the primary region. For applications requiring high availability, Microsoft recommends using ZRS in the primary region, and replicating to a secondary region.
- Google offers regional Persistent Disk (PD) which is a storage option that provides synchronous replication of data between two zones in a region. Regional persistent disks can be a good building block to use when you implement HA services in Compute Engine.

4 Building a Multi-Cloud Strategy

In case of applications deployed to an on-premises data center, the likelihood of significant data mobility is low. At the opposite end of the scale, where applications run in any cloud at any time, full flexibility is needed.



Data Mobility:

Implementing data mobility is probably the biggest challenge in deploying and operating in a multi-cloud environment. Data has inertia and takes time to move around. Cloud suppliers charge for egress – data accessed outside of their cloud – so mass migration of data from one cloud platform to another isn't really a practical solution. As a result, the strategies for multi-cloud data mobility tend to fall into one of these three categories: burst on-demand, data replication or abstraction of data.

4.1 Burst On-Demand

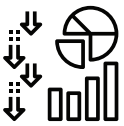
Cloud bursting is all about dynamic deployment of applications that normally run on a private cloud into a public cloud to meet expanding capacity requirements and handle peak demands when private cloud resources are insufficient or not elastic. Cloud bursting can make these private clouds more cost-efficient by eliminating the need to overbuild physical infrastructure to ensure enough capacity to meet fluctuating peaks in demand. Private clouds can be rightsized in terms of compute and storage to accommodate the business's usual demands and the peaks can be handled by a public cloud and a pay-per-use model which makes it cost-efficient.

There are many scenarios where businesses can benefit from cloud bursting. For example, many sectors deal with seasonal spikes that put an extra burden on private clouds. Enterprise data centers may have geographic needs where one location experiences heavy loads and must meet application-specific performance needs unlike the other locations and based on seasonal demands which are region-specific.

Software development projects and analytics are two of the fastest-growing drivers of demand for cloud bursting. For example, organizations in the health and financial sectors are heavy users of cloud bursting and often have fluctuating analytics needs that require faster CPU, applications such as:

Computationally intensive, read-heavy workloads

Batch or burst usage model for longer-running simulations



Risk analysis (credit, hedge funds, portfolio management, actuarial)

Log analysis

The ability to manage demand effectively and seamlessly and potentially bring large volumes of additional CPU power to bear is the only way that many sectors can make this type of data analysis effective in terms of both cost and time.

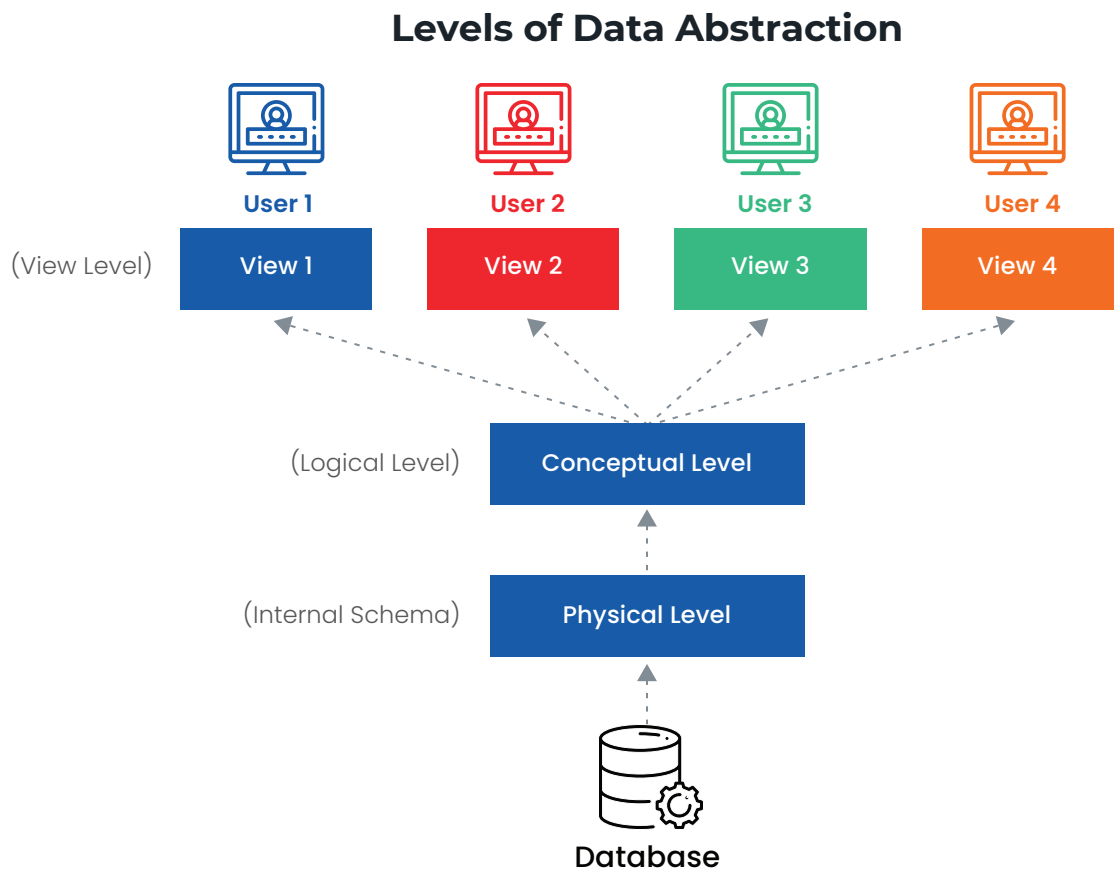
4.2 Replicate Data

Data can be replicated across cloud but replication can be expensive depending upon the number of copies maintained. Keeping copies closely in synchronization to promote scalability, allows companies to work with real time, constantly evolving data. File-based solutions such as Elastifile, CloudConnect or Qumulo QF2 provide the capability to span on-premises or multiple cloud environments, making it easy to expose on-premises data to, for example, analytics services in the public cloud. There are three different types of replication:

- **Transactional Replication –** Replicate a full copy of your database and send continuous updates as your data changes. This makes it easy to keep track of what is altered and if data is lost. Transactional replication doesn't just copy your data changes, it continuously replicates every single change with great accuracy. Normally, this type is used in server-to-server environments.
- **Snapshot Replication -** A snapshot of the database is taken and distributed across servers. Data is sent over exactly as it appears in a specific moment. Typically, snapshot replication will be used when changes to data are sparse. This replication type is great when performing initial synchronization between the publisher and the subscriber but tends to be a bit slower. This is because every snapshot sent is attempting to move multiple data records from one end to the other.
- **Merge Replication –** This type of replication occurs when two or more databases are combined into one singular database. Merge replication allows any data changes to be sent from the publisher (primary server) to one or more subscribers (secondary servers). This replication type is the most complex type because it lets both the publisher and the subscribers make changes to the database. It's typically used in a server-to-client environment.

4.3 Abstract Data

An abstraction will simplify the data by only providing what is relevant and required and ignore unrelated details. So, these solutions separate the presentation of data from the underlying physical storage platform. This provides a single global view of the data, while the physical storage can be across one or many public and private clouds. Examples of this solution are Zenko from Scality or Ctera Enterprise File Services Platform. With full abstraction, data can be physically redirected or replicated between cloud providers to meet the needs of availability, cost and performance. There are three levels of Data Abstraction,



- **Physical –** This is the lowest level of data abstraction. It describes how data is stored in the database. You can get complex data structure details at this level.
- **Logical –** This level comprises the information that is stored in the database in the form of tables. It also stores the relationship among the data entities in relatively simple structures.
- **View –** This is the highest level of abstraction. Only a part of the actual data is viewed by the users. This level exists to ease the accessibility of the data by an individual user. Users view data in the form of rows and columns. Tables and relations are used to store data. Multiple views of the same data may exist. Users can just view the data and interact with the data, storage and implementation details are hidden from them.

Example:

Let's say we are storing customer information in a table. At the physical level, these records can be described as blocks of storage (bytes, gigabytes, terabytes, etc.) in memory. These details are often hidden from the programmers.

- At the logical level, these records can be described as fields and attributes along with their data types, their relationship among each other can be logically implemented. The programmers generally work at this level because they are aware of such things about database systems.
- At view level, users just interact with the system with the help of GUI and enter the details at the screen, they are not aware of how the data is being stored and what data is being stored; such details are hidden from them.

The key to making these solutions work are principles such as automation – being able to run storage services from command line interfaces or APIs. Solution implementation also needs to consider incremental changes in data and should be able to automatically move only the changed data between locations, as this reduces the impact of egress charges.

5 Conclusion

Managing data is not a 'one-size-fits-all' proposition. Every organization faces unique challenges in terms of data storage and availability. Any plan to implement a comprehensive data storage and management system must consider the various parameters that impact the performance, scalability, security and cost. Organizations need to have a clear idea of the business value their data represents and understand how to leverage storage solutions that allow them to maximize that value.

In today's world, organizations have a wide variety of options when it comes to managing data and they should take advantage of the unique characteristics of each. A multi-tier approach that utilizes different storage technologies is essential for optimizing performance. Frequently used data, for instance, may be stored on faster, more easily accessible devices like solid-state drives (SSDs), while back-up data may be kept on more cheaper options available for cold data. Matching enterprise data storage technology to business needs can greatly improve efficiency and ensure that companies are able to maximize the value of their data.

Always thinking a little further ahead is the key to success for all organizations. This is incredibly important for enterprise data storage because managing a data solution can carry significant ongoing costs. In many cases, companies focus too much on the up-front costs of implementing a data management solution. Those data strategy costs represent only a small portion of the total cost of ownership. They must also think about how their storage needs might scale in the future and what degree of flexibility they may need as the business grows.

Cloud-based storage technology has come a long way over the last decade. While storing critical data in a cloud environment may have given some organizations pause in years past, now there are many cloud vendors offering storage solutions that are far more secure and versatile than on-premises solutions. Whether it is a public cloud environment or a customized private cloud, organizations should evaluate their cloud options closely and ensure to create a good cloud storage strategy with all the discussed parameters and achieve their business goals within a planned budget.

References

- The growing adoption of cloud storage – 2019 survey findings
- The future of cloud storage: Experts look 5 years out

Author

Ganesh Kamalanathan

Principal Consultant

Hexaware

About Hexaware

Hexaware is the fastest growing next-generation provider of IT, BPO and Consulting services. Our focus lies on taking a leadership position in helping our clients attain customer intimacy as their competitive advantage. Our digital offerings have helped our clients achieve operational excellence and customer delight. We are now on a journey of metamorphosing the experiences of our customer's customers by leveraging our industry-leading delivery and execution model, built around the strategy - 'Automate Everything™, Cloudify Everything™, Transform Customer Experiences™.' Hexaware services customers in over two dozen languages, from every major time zone and every major regulatory zone. Our goal is to be the first IT services company in the world to have a 50% digital workforce.

Learn more about Hexaware at **www.hexaware.com**

NA Headquarters
Metro 101, Suite 600,101
Wood Avenue South, Iselin,
New Jersey - 08830
Tel: +001-609-409-6950
Fax: +001-609-409-6910

India Headquarters
152, Sector – 3
Millennium Business Park
'A' Block, TTC Industrial Area
Mahape, Navi Mumbai – 400 710
Tel : +91-22-67919595
Fax : +91-22-67919500

EU Headquarters
Level 19, 40 Bank Street,
Canary Wharf,
London - E14 5NR
Tel: +44-020-77154100
Fax: +44-020-77154101

APAC Headquarters
#09-01, One Finlayson
Green, 1 Finlayson Green,
Singapore - 049246
Tel : +65-63253020
Fax : +65-6222728

Australia Headquarters
Level 3, 80 Mount St
North Sydney NSW 2060
Australia
Tel : +61 2 9089 8959
Fax : +61 2 9089 8989