Storage Optimization

January 2018

Notices

This document is provided for informational purposes only. It represents AWS's current product offerings and practices as of the date of issue of this document, which are subject to change without notice. Customers are responsible for making their own independent assessment of the information in this document and any use of AWS's products or services, each of which is provided "as is" without warranty of any kind, whether express or implied. This document does not create any warranties, representations, contractual commitments, conditions or assurances from AWS, its affiliates, suppliers or licensors. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Storage Optimization: Choosing the right Amazon S3 and Amazon EBS storage

Storage optimization is the process of choosing the AWS storage services that provide the most appropriate balance of performance, availability, and durability for each of your data storage needs—at the lowest cost. Organizations can tend to think of data storage as an "ancillary service" and neglect the effort of optimizing storage once data is moved to the cloud. Many fail to clean up storage that is no longer being used, leaving these services running for days, weeks, and even months—at significant cost. Up to 7 percent of all cloud spend is wasted on unused storage volumes and old snapshots (copies of storage volumes).[1]

Optimizing storage is an ongoing process

AWS offers a broad and flexible set of data storage options that enable you to easily move between different tiers of storage and change storage types at any time. However, to take advantage of this flexibility and maintain a storage architecture that is both right-sized and right-priced requires an ongoing process. To get the most efficient use of your storage spend, organizations should optimize storage monthly. You can streamline this effort by:

- Establishing an ongoing mechanism for optimizing storage and setting up storage policies.
- Monitoring costs closely using AWS cost and reporting tools, such as Cost Explorer, budgets, and detailed billing reports in the Billing and Cost Management console.
- Enforcing Amazon S3 object tagging and establishing S3 lifecycle policies to continually optimize data storage throughout the data lifecycle.
- Understanding the process of storage optimizing, which is the focus of this whitepaper.

Let's begin by evaluating your storage needs.

Identifying your data storage requirements

The first step in optimizing storage is to understand the performance profile for each of your data sets. AWS storage services are optimized for different storage scenarios—there is no single data storage option that is ideal for all workloads. When evaluating your storage requirements, segment data within each workload by how much you value the data (how durable it is) and how available the data needs to be—and consider data storage options for each separately. To avoid overprovisioning, you can conduct a performance analysis that includes measuring IOPS, throughput, and other variables.

The following are some questions to consider as you evaluate data storage requirements:

- ***How often and how quickly does the data need to be accessed?*** AWS offers storage options and pricing tiers for hot, warm, and cold data.

[1] According to research by RightScale Inc.

- ***Does the data store require high IOPS or throughput?*** AWS provides categories of storage that are optimized for performance and throughput. Understanding IOPS and throughput requirements will help you provision "just enough" and avoid overpaying.
- ***How critical (durable) is the data?*** Critical or regulated data needs to be retained at almost any expense and tends to be stored for long periods of time.
- ***How sensitive is the data?*** Highly sensitive data needs to be protected from accidental and malicious changes, not just data loss or corruption. Durability, cost, and security are equally important to consider.
- ***How large is the data set?*** Knowing the total size of the data set helps in estimating storage capacity and cost.
- ***How transient is the data?*** Transient data is short-lived and typically does not require high durability.[2] Clickstream and Twitter data are good examples.
- ***How much are you prepared to pay to store the data?*** Setting a budget for data storage will inform your decisions about storage options.

Choosing the right AWS storage service

Picking the right AWS storage service for your data means finding the closest match in terms of data availability,[3] durability, and performance.[4] Amazon offers three broad categories of storage services: object, block, and file storage. Each is designed for different storage requirements, giving you flexibility to find what works best for your storage scenarios:

- **Object storage: Amazon Simple Storage Service (Amazon S3)**
  Amazon S3 is highly durable, general-purpose object storage that works well for unstructured data sets such as media content. Amazon S3 provides the highest level of data durability and availability on the AWS platform. There are three tiers of storage: one each for hot, warm, or cold data. In terms of pricing, the colder the data, the cheaper it is to store, and the costlier it is to access when needed (we'll look at a price comparison later in this section). You can easily move data between these storage options to optimize storage costs:

  Amazon S3 Standard
  The best storage option for hot, frequently accessed data, Amazon S3 delivers low latency and high throughput and is ideal for use cases such as cloud applications, dynamic websites, content distribution, gaming, and data analytics.

  Amazon S3-Infrequent Access (Amazon S3-IA)
  Use this warm storage option for data that is accessed less frequently, such as long-term backups and disaster recovery. It offers cheaper storage over time, but higher charges to retrieve or transfer data.

  Amazon Glacier
  A cold-storage option, Amazon Glacier is designed for long-term storage of infrequently accessed data, such as end-of-lifecycle, compliance, or regulatory backups. Different

[2] Durability refers to average annual expected data loss.

[3] Availability refers to a storage volume's ability to deliver data upon request.

[4] Performance refers to the number of input/output operations per second (IOPS) or the amount of throughput (measured in megabytes per second) that the storage volume can deliver.

methods of data retrieval are available at various speeds and cost. Retrieval can take from a few minutes to several hours.

- **Block storage: Amazon Elastic Block Store (Amazon EBS)**
  Amazon EBS volumes provide a durable block-storage option for use with Amazon EC2 instances. Use Amazon EBS for data that requires long-term persistence and quick access at guaranteed levels of performance. There are two types of block storage: solid-state-drive (SSD) storage and hard-disk-drive (HDD) storage.

  SSD storage is optimized for transactional workloads where performance is closely tied to input/output operations per second (IOPS). There are two SSD volume options to choose from:

  Amazon EBS General Purpose SSD (gp2)
  The gp2 volumes are designed for general use and offer a balance between cost and performance.

  Amazon EBS Provisioned IOPS SSD (io1)
  The io1 volumes are for latency-sensitive workloads requiring specific minimum-guaranteed IOPS. With io1 volumes, you pay separately for provisioned IOPS, so unless you need high levels of provisioned IOPS, gp2 volumes are a better match at lower cost.

  HDD storage is designed for throughput-intensive workloads such as data warehouses and log processing. There are two types of HDD volumes:

  Amazon EBS Throughput Optimized HDD (st1)
  The st1 volumes are best for frequently accessed throughput-intensive workloads.

  Amazon EBS Cold HDD (sc1)
  The sc1 volumes are designed for less frequently accessed throughput-intensive workloads.

- **File storage: Amazon Elastic File System (Amazon EFS)**
  Amazon EFS provides simple, scalable file storage for use with Amazon EC2 instances. Amazon EFS supports any number of instances at the same time, and its storage capacity can scale from gigabytes to petabytes of data without needing to provision storage. Amazon EFS is designed for workloads and applications such as big data, media-processing workflows, content management, and web serving. Amazon EFS also now supports file synchronization capabilities so that you can efficiently and securely synchronize files from on-premises or cloud file systems to Amazon EFS at speeds of up to 5 times faster than standard Linux copy tools.

Amazon S3 and Amazon EFS allocate storage based on your usage (you pay for use). However, for Amazon EBS volumes, you are charged for provisioned (allocated) storage, whether or not you use it. The key to keeping storage costs low without sacrificing required functionality is to maximize the use of Amazon S3 when possible and use more expensive Amazon EBS volumes with provisioned I/O only when application requirements demand it.

The following table shows comparative pricing for Amazon S3, Amazon EBS, and Amazon EFS:

Amazon S3 Pricing*	***Per Gigabyte-Month***
Amazon S3	$0.023
Amazon S3-IA	$0.0125 (plus $0.01/GB retrieval charge)
Amazon Glacier	$0.004

*Based on US East (N. Virginia) prices.

Amazon EBS Pricing*		
Amazon EBS General Purpose SSD (gp2)	$0.10	Per GB-month of provisioned storage
Amazon EBS Provisioned IOPS SSD (io1)	$0.125 $0.065	Per GB-month of provisioned storage, plus Per provisioned IOPS-month
Amazon EBS Throughput Optimized HDD (st1)	$0.045	Per GB-month of provisioned storage
Amazon EBS Cold HDD (sc1) volumes	$0.025	Per GB-month of provisioned storage
Amazon EBS Snapshots to Amazon S3	$0.05	Per GB-month of data stored

*Based on US East (N. Virginia) prices.

Amazon EFS Pricing*	***Per Gigabyte-Month***
Amazon EFS	$0.30

*Based on US East (N. Virginia) prices.

Optimizing Amazon S3 storage

Amazon S3 lets you analyze data access patterns, create inventory lists, and configure lifecycle policies, making it easy to continually optimize S3 storage to keep costs down and ensure optimal performance. You can set up rules to automatically move data objects to cheaper tiers of S3 storage as objects are accessed less frequently—or automatically delete objects after an expiration date. To manage storage data most effectively, use Amazon S3 object tagging to categorize data so that you can specify these tags in your data lifecycle policies.

To determine when to transition data to the right storage class, you can use Amazon S3 Analytics to analyze storage access patterns. Analyze all the objects in a bucket or use an object tag or common prefix to filter objects for analysis. After observing infrequent access patterns of a filtered set of data over a period of time, you can use the information to choose the most appropriate storage classes, improve lifecycle policies, and make predictions around future usage and growth.

Another management tool is Amazon S3 Inventory, which audits and reports on the replication and encryption status of your objects in an S3 bucket on a weekly or monthly basis. This feature provides CSV output files that list objects and their corresponding metadata and lets you configure multiple inventory lists for a single bucket, organized by different S3 metadata tags. You can also query Amazon S3 inventory using standard SQL by using Amazon Athena, Amazon Redshift Spectrum, and other tools, such as Presto, Apache Hive, and Apace Spark.

Amazon S3 also can publish storage, request, and data transfer metrics to Amazon CloudWatch—AWS's resource-monitoring service. Storage metrics are reported daily, are available at one-minute intervals for granular visibility, and can be collected and reported for an entire bucket or a subset of objects (selected via prefix or tags).

With all the information these storage management tools provide, you can create policies to move less-frequently-accessed data S3 data to cheaper storage tiers for considerable savings. For example, moving data from Amazon S3 Standard to Amazon S3-IA can save up to 60 percent (on a per-gigabyte basis) of Amazon S3 pricing. At the end of its lifecycle, when data is accessed on rare occasions, moving it to Amazon Glacier can save up to 80 percent of Amazon S3 pricing.

The following table compares the monthly cost of storing 1 petabyte of content on Amazon S3 Standard versus Amazon S3-IA (the cost includes the content retrieval fee). It demonstrates that if 10 percent of the content is accessed per month, the savings would be 41 percent with Amazon S3-IA. If 50 percent of the content is accessed, the savings would be 24 percent—which is still significant. Even if 100 percent of the content is accessed per month, you would still save 2 percent using Amazon S3-IA.

Comparing 1 Petabyte of Object Storage*

	Content Accessed Per Month	S3 Standard	S3-IA	Savings
1 PB Monthly	10%	$24,117	$14,116	41%
1 PB Monthly	50%	$24,117	$18,350	24%
1 PB Monthly	100%	$24,117	$23,593	2%

*Based on US East prices.

Note: There is no charge for transferring data between Amazon S3 storage options as long as they are within the same AWS region.

To further optimize costs associated to storage and data retrieval, AWS announced the launch of Amazon S3 Select and Amazon Glacier Select (now in preview). Traditionally, data in object storage had to be accessed as whole entities, regardless of the size of the object. Amazon S3 Select now lets you retrieve a subset of data from an object using simple SQL expressions, which means that your applications no longer have to use compute resources to scan and filter the data from an object. Using Amazon S3 Select, you can potentially improve query performance by up to 400 percent and reduce query costs as much as 80 percent. AWS also supports efficient data retrieval with cold-storage service Amazon Glacier so that you do not have to restore an archived object to find the bytes needed for analytics. With both Amazon S3 Select and Amazon Glacier Select, you can lower your costs and uncover more insights from your data, regardless of what storage tier it's in.

Optimizing Amazon EBS storage

For Amazon EBS storage, it's important to keep in mind that you are paying for provisioned capacity and performance—even if the volume is unattached or has very low write activity. To optimize storage performance and costs for this type of storage, monitor volumes periodically to identify ones that are unattached or appear to be underutilized or overutilized, and adjust provisioning to match actual utilization.

AWS tools that can assist with block storage optimization include Amazon CloudWatch, which automatically collects a range of data points for EBS volumes and lets you set alarms on volume behavior. AWS Trusted Advisor is another option for analyzing your infrastructure to identify unattached, underutilized, and overutilized EBS volumes. Third-party tools, such as Cloudability, can also provide insight into performance of EBS volumes.

Deleting unattached Amazon EBS volumes

Identifying unattached volumes and deleting them is an easy way to reduce wasted spend. When Amazon EC2 instances are stopped or terminated, attached Amazon EBS volumes are not automatically deleted and will continue to accrue charges since they are still operating. To find unattached EBS volumes, look for volumes that are “available,” indicating that they are not attached to an Amazon EC2 instance. You can also look at network throughput and IOPS to see whether there has been any volume activity over the previous two weeks, or look up the last time the Amazon EBS volume was attached. If the volume is in a nonproduction environment, hasn’t been used in weeks, or hasn’t been attached in a month, there is a good chance you can delete it.

Before deleting a volume, store an Amazon EBS snapshot (a backup copy of an EBS volume) so that the volume can be quickly restored later if needed. You can automate the process of deleting unattached volumes by using AWS Lambda functions with Amazon CloudWatch.

Resizing or changing volume type

Another way to optimize storage costs is to identify volumes that are underutilized and downsize them or change the volume type. Monitor the read-write access of Amazon EBS volumes to determine if throughput is low. If you have a current-generation Amazon EBS volume attached to a current-generation Amazon EC2 instance type, you can use the Elastic Volumes feature to change the size or volume type, or (for an SSD io1 volume) adjust IOPS performance—without detaching the volume.

The following are tips to keep in mind when optimizing EBS volumes:

- For SSD gp2 volumes, you’ll want to optimize for capacity so that you’re paying only for what you use.
- With provisioned IOPS SSD io1 volumes, pay close attention to IOPS utilization rather than throughput, since you pay for IOPS directly. Provision 10–20 percent above maximum IOPS utilization.
- You can save by reducing provisioned IOPS or by switching from a provisioned IOPS volume type to SSD gp2.
- If the volume is 500 gigabytes or larger, consider converting to a Cold HDD sc1 volume to save on your storage rate.
- You can always return a volume to its original settings if needed.

Deleting stale Amazon EBS snapshots

If you have a backup policy that takes EBS volume snapshots daily or weekly, you will quickly accumulate snapshots. Check for “stale” snapshots over 30 days old and delete them to reduce storage costs. Deleting a snapshot has no effect on the volume. You can use AWS Console or AWS Command Line Interface (CLI) for this purpose, or third-party tools such as Skeddly.

Conclusion

Storage optimization is the ongoing process of evaluating changes in data storage usage and needs, and choosing the most cost effective and appropriate AWS storage option. For object stores, this means implementing Amazon S3 lifecycle policies to automatically move data to cheaper storage tiers as data is accessed less frequently. For Amazon EBS block stores, optimization focuses on monitoring storage usage and resizing underutilized (or overutilized) volumes. It's also about deleting unattached volumes and stale Amazon EBS snapshots so that you're not paying for unused resources. You can streamline the process of storage optimization by setting up a monthly schedule for this task and taking advantage of the powerful tools by AWS and third-party vendors to monitor storage costs and evaluate volume usage.