

AWS Storage Optimization

March 2018



Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. 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.

© 2018 Amazon Web Services, Inc. or its affiliates. All rights reserved.

Contents

Introduction	1
Identify Your Data Storage Requirements	1
AWS Storage Services	2
Object storage	2
Block storage	3
File storage	5
Optimizing Amazon S3 Storage	5
Optimizing Amazon EBS Storage	7
Delete Unattached Amazon EBS Volumes.....	8
Resize or Change the EBS Volume Type.....	8
Delete Stale Amazon EBS Snapshots.....	9
Optimizing Storage is an Ongoing Process	9
Conclusion.....	10

Abstract

This is the last in a series of whitepapers designed to support your cloud journey. This paper seeks to empower you to maximize value from your investments, improve forecasting accuracy and cost predictability, create a culture of ownership and cost transparency, and continuously measure your optimization status.

This paper discusses how to choose and optimize AWS storage services to meet your data storage needs and help you save costs.

Introduction

Organizations tend to think of data storage as an ancillary service and do not optimize storage after data is moved to the cloud. Many also fail to clean up unused storage and let these services run for days, weeks, and even months at significant cost. According to [this blog post by RightScale](#), up to 7% of all cloud spend is wasted on unused storage volumes and old snapshots (copies of storage volumes).

AWS offers a broad and flexible set of data storage options that let you move between different tiers of storage and change storage types at any time. This whitepaper discusses how to choose AWS storage services that meet your data storage needs at the lowest cost. It also discusses how to optimize these services to achieve balance between performance, availability, and durability.

Identify Your Data Storage Requirements

To optimize storage, the first step is to understand the performance profile for each of your workloads. You should conduct a performance analysis to measure input/output operations per second (IOPS), throughput, and other variables.

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, consider data storage options for each workload separately.

The following questions can help you segment data within each of your workloads and determine your storage requirements:

- **How often and how quickly do you need to access your data?** AWS offers storage options and pricing tiers for frequently accessed, less frequently accessed, and infrequently accessed data.
- **Does your 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 the right amount of storage and avoid overpaying.

- **How critical (durable) is your data?** Critical or regulated data needs to be retained at almost any expense and tends to be stored for a long time.
- **How sensitive is your 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 your data set?** Knowing the total size of the data set helps in estimating storage capacity and cost.
- **How transient is your data?** Transient data is short-lived and typically does not require high durability. (Note: Durability refers to average annual expected data loss.) Clickstream and Twitter data are good examples of transient data.
- **How much are you prepared to pay to store the data?** Setting a budget for data storage will inform your decisions about storage options.

AWS Storage Services

Choosing the right AWS storage service for your data means finding the closest match in terms of data availability, durability, and performance.

Note: **Availability** refers to a storage volume's ability to deliver data upon request. **Performance** refers to the number of IOPS or the amount of throughput (measured in megabytes per second) that the storage volume can deliver.

Amazon offers three broad categories of storage services: object, block, and file storage. Each offering is designed to meet a different storage requirement, which gives you flexibility to find the solution that works best for your storage scenarios.

Object storage

[Amazon Simple Storage Service](#) (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 Cloud. 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. You can easily move data between these storage options to optimize storage costs:

- **Amazon S3 Standard** – The best storage option for data that you frequently access. 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 Standard - Infrequent Access (Amazon S3 Standard - IA)** – Use this storage option for data that you access 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** – Designed for long-term storage of infrequently accessed data, such as end-of-lifecycle, compliance, or regulatory backups. Different methods of data retrieval are available at various speeds and cost. Retrieval can take from a few minutes to several hours.

The following table shows comparative pricing for Amazon S3.

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

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

Block storage

[Amazon Elastic Block Store](#) (Amazon EBS) volumes provide a durable block-storage option for use with 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 IOPS. There are two SSD volume options to choose from:
 - **EBS Provisioned IOPS SSD (io1)** – Best for latency-sensitive workloads that require 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.
 - **EBS General Purpose SSD (gp2)** – Designed for general use and offer a balance between cost and performance.
- **HDD storage** is designed for throughput-intensive workloads such as data warehouses and log processing. There are two types of HDD volumes:
 - **Throughput Optimized HDD (st1)** – Best for frequently accessed, throughput-intensive workloads.
 - **Cold HDD (sc1)** – Designed for less frequently accessed, throughput-intensive workloads.

The following table shows comparative pricing for Amazon EBS.

Amazon EBS Pricing*	Per Gigabyte-Month
General Purpose SSD (gp2)	\$0.10 per GB-month of provisioned storage
Provisioned IOPS SSD (io1)	\$0.125 per GB-month of provisioned storage, plus \$0.065 per provisioned IOPS-month
Throughput Optimized HDD (st1)	\$0.045 per GB-month of provisioned storage
Cold HDD (sc1)	\$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.	

File storage

[Amazon Elastic File System](#) (Amazon EFS) provides simple, scalable file storage for use with EC2 instances. Amazon EFS supports any number of instances at the same time. 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 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 and you pay for what you use. However, for 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 EBS volumes with provisioned I/O only when application requirements demand it.

The following table shows pricing for Amazon EFS.

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

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

Optimize Amazon S3 Storage

Amazon S3 lets you analyze data access patterns, create inventory lists, and configure lifecycle policies. You can set up rules to automatically move data objects to cheaper S3 storage tiers as objects are accessed less frequently or to automatically delete objects after an expiration date. To manage storage data most effectively, you can use tagging to categorize your S3 objects and filter on these tags in your data lifecycle policies.

To determine when to transition data to another storage class, you can use [Amazon S3 analytics storage class analysis](#) 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. If you observe infrequent access patterns of a filtered data set over time, you can use the information to choose a more appropriate storage class, 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 S3 objects 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 Apache Spark.

Amazon S3 can also publish storage, request, and data transfer metrics to [Amazon CloudWatch](#). 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, by moving data from Amazon S3 Standard to Amazon S3 Standard-IA, you can save up to 60% (on a per-gigabyte basis) of Amazon S3 pricing. By moving data that is at the end of its lifecycle and accessed on rare occasions to Amazon Glacier, you can save up to 80% of Amazon S3 pricing.

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

Comparing 1 Petabyte of Object Storage*

	Content Accessed Per Month	S3 Standard	S3 Standard - 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. 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% and reduce query costs as much as 80%. AWS also supports efficient data retrieval with 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.

Optimize Amazon EBS Storage

With Amazon EBS, 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 Amazon EBS, monitor volumes periodically to identify ones that are unattached or appear to be underutilized or overutilized, and adjust provisioning to match actual usage.

AWS offers tools that can help you optimize block storage. [Amazon CloudWatch](#) automatically collects a range of data points for EBS volumes and lets you set alarms on volume behavior. [AWS Trusted Advisor](#) is another way for you to analyze 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.

Delete Unattached Amazon EBS Volumes

An easy way to reduce wasted spend is to find and delete unattached volumes. However, when EC2 instances are stopped or terminated, attached 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**, which indicates that they are not attached to an 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. 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.

Resize or Change the EBS 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 EBS volumes to determine if throughput is low. If you have a current-generation EBS volume attached to a current-generation 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 tips can help you optimize your EBS volumes:

- For General Purpose 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% above maximum IOPS utilization.
- You can save by reducing provisioned IOPS or by switching from a Provisioned IOPS SSD io1 volume type to a General Purpose SSD gp2 volume type.
- 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.

Delete 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 that are over 30 days old and delete them to reduce storage costs. Deleting a snapshot has no effect on the volume. You can use the AWS Management Console or AWS Command Line Interface (CLI) for this purpose or third-party tools such as Skeddly.

Storage Optimization is an Ongoing Process

Maintaining a storage architecture that is both right-sized and right-priced is an ongoing process. To get the most efficient use of your storage spend, you should optimize storage on a monthly basis. 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.

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, you want to implement Amazon S3 lifecycle policies to automatically move data to cheaper storage tiers as data is accessed less frequently. For Amazon EBS block stores, monitor your storage usage and resize underutilized (or overutilized) volumes. You also want to delete 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.