Projet : Cloud Computing

Master Big Data - CMI Big Data

AWS : S3

Réalisé par :

Maroun Gaby -- Lopez Fabien

Guidé par:
M.Khaled Khebbeb

Introduction

With the rapid development of processing and storage technologies and the success of the Internet, IT resources have become cheaper, more powerful and more ubiquitous than ever.

Cloud computing has conquered the market and no one, at least in the near future, will stop it. Amazon, a big company as much as having the wealthiest man in the world as CEO, and of course that didn’t come from nothing, is still one of the leading companies in all the markets it enters and of course she wouldn’t let this huge opportunity or that winning market slip away, so they came in with their "big guns".

In the following article, we will discuss the rise of Amazon Sample Storage, a very important storage platform used today by some large companies and applications such as Netflix because of its object storage, easy to use model and good prices.

We will start by discussing how to use it, continue by citing some of its features and networking, APIs, virtualization and we will finish by talking about it's easy to understand documentation.

1. Business Model :

Amazon S3 pricing can be divided into 4 kinds with the aim of reducing cost and paying only for what you use.

First, Amazon S3 provides us with a pricing grid where the cost depends on the amount of Gb used, the geographic location of our data and the kind of service. For the example below we will take Paris as geographic location of our data :

We have 6 services:

- *S3 standard service* : We can use it on any kind of data but it's usually used for frequently accessed data.
- *S3 intelligent Tiering* : used for data with unknown or changing access patterns (frequent-infrequently).
- S3 Standard Infrequent Access : as the name suggests, allows to store data that will not be used frequently, needs millisecond access and has a long lifespan.
- *S3 One Zone* : used for recreatable infrequently accessed data that needs millisecond access.
- S3 Glacier : can be used for backups or archives that can be retrieved from 1 minute to 12 hours.
- S3 Glacier Deep Archive : used for archives that are accessed more or less twice in a year and, as the previous, can be retrieved within 12 hours.

Second, we can find requests and data retrievals pricing grid. There are the same services as above but here, price depends on the number of requests (generally per piece of 1000 requests). We can find simples requests like Put, Copy, Get, select or some more complex, like lifecycle transition request that can be used for change automatically an object from a service (for example *S3 Standard Service*) to another service (*S3 Standard Infrequent Access*)

Third, we have to pay Amazon S3 for all bandwidth usage :

- From Amazon S3 to the Internet when it's up to 10 Tb/month.
- Data Transfer out from Amazon S3 except if it's to Amazon EC2 and the instance of Amazon S3 and EC2 are in the same region or if it's to Amazon CloudFront.

Amazon S3 propose in addition some transfer acceleration

Finally, Amazon provides one last service : management and replication. With this service, we pay for storage in the selected destination, requests for duplicate the data and bandwidth usage.

2. Installation:

To get started with Amazon Simple Storage Service or S3, you should first set up and log into your AWS account, if you don't already have one, you'll be prompted to create one when you sign up for Amazon S3. You will not be charged for Amazon S3 until you use it.

The second step would be creating a bucket. Every object in Amazon S3 is stored in a bucket. Before you can store data in Amazon S3, you must create and name an S3 bucket.

Buckets are the fundamental containers in Amazon S3 for data storage and their names are globally unique where the namespace is shared by all AWS accounts, this means that after a bucket is created, the name of that bucket cannot be used by another AWS account in any AWS Region until the bucket is deleted.

We should specify the region where we want the bucket to be created and Amazon recommend it to be as close to our geographical please as possible, and that is to optimize latency, minimize costs, or even address regulatory requirements.

By default, we can create as many as 100 buckets and that can be modified to reach a maximum of 1000. Amazon S3 provides APIs for creating and managing buckets. They recommend you to not use the root credentials of your AWS account to make requests such as to create a bucket. Instead, create an IAM user, and grant that user full access.

The third and final step to get started with Amazon S3 would be the first step in your new storage warehouse and that is to start building. Now that we've formed a bucket, we're prepared to add an object to it. An object can be any kind of file: a text file, a photo, a video, and so on. Objects are the fundamental entities stored in Amazon S3. Objects kept in the buckets have an exclusive key value and are recovered using a URL while each object can cover up to 5 TB of data. We can store an infinite quantity of information in a bucket and upload as many objects as we like. We can also download or upload our Data or grant others the permission to do so.

A web-based interface for editing and handling Amazon S3 resources is available via the AWS Management Console. Amazon S3 can be integrated with existing applications through a wide variety of other AWS services.

One of the advantages of using Amazon Simple Storage Service (Amazon S3) is its accessibility to any amount of data, at any time, from anyplace on the web. It is also durable, highly available, and protected.

Amazon S3 is similar to a storehouse for Internet data. Amazon S3 offers access to reliable, fast, and cheap data storage infrastructure. With that in mind, it’s been designed to make

web-scale computing easier by allowing you to stock and recover any capacity of data, at any time, from within Amazon EC2 or anywhere on the internet

Amazon S3 also keeps multiple redundant duplicates of your data and allows simultaneous read or write access to these data objects by many distinct users or application threads. You can use these redundant data to recover rapidly and reliably from any kind of failures.

All IaaS (infrastructure as a service) AWS management, administration and access functions in the AWS Management Console are obtainable in the AWS API and CLI.

The AWS Command Line Interface (AWS CLI) is an open source tool that permits you to interact with AWS services using commands in your command-line shell. With a small change in the settings, the AWS CLI allows you to start running commands that implement functionalities equivalent to the browser-based AWS Management ones. It can be used on Linux shells, Windows command line or by running commands on Amazon Elastic Compute Cloud (Amazon EC2). You can also use what you learn to develop programs in other languages (Java, Python, …) by using the AWS SDKs.

3. Authentication:

Authentication to amazon S3 is done using an email address that serves as a username and password. This one can be improved with the addition of a multi-factor verification system. This involves adding an additional layer of security using an MFA Device.

However, we can need to access this data using a machine. In this case, Amazon S3 offers us ID keys. Within the limit of two per account. We can see below what this style of key looks like:

In addition, it is possible for a “root” user to create permanent end users (IAM user). They are assigned to roles and permissions for the various files on Amazon S3. A role is a kind of policy that can apply to a group of users. A single user can have “infinite” roles but can only use one at a time.

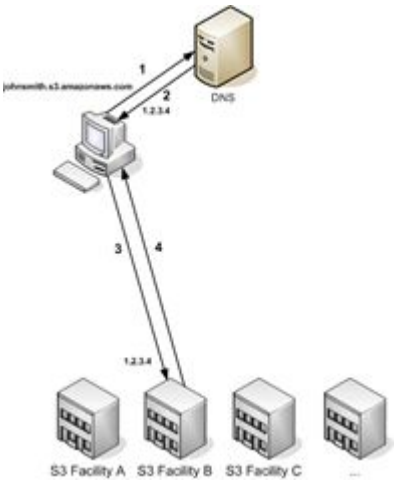
In addition, the root user can create temporary end users with a temporary token with which the end user can be identified for a short period (between a few minutes and several hours). When a token is out of date, the end user can request a new one if they have the rights. This token can be used by the root user, for example, when they prefer to avoid creating a permanent end user.

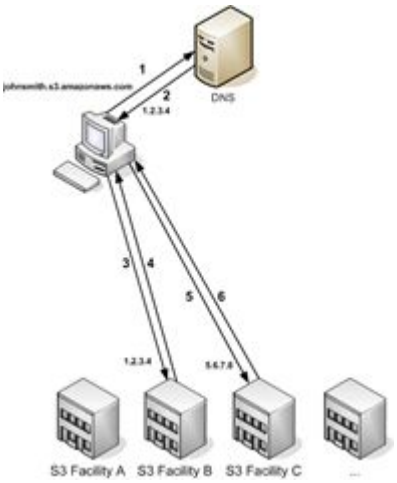
4. Network management :

S3 is a highly scalable public cloud storage service that uses objects instead of blocks or files and it is accessed using web-based protocols that use standard HTTP(S) (Amazon S3 website endpoints do not support HTTPS) and a REST-based application programming interface (API). It will use HTTP requests, recover their URI component, adapt it to S3 object name and with getObject() it gets content (using one of available S3 SDKs, for example AWS python SDK) Support for SOAP over HTTP is deprecated, but it is still accessible over HTTPS. However, new Amazon S3 features will not be maintained for SOAP

Amazon S3 uses the Domain Name System (DNS) to forward requests to services that can process them. This system works excellently, but temporary routing errors might rise. If a request reaches the wrong place, Amazon S3 responds with a temporary answer that tells the requester to return the message to the right place. If a request is incorrectly designed, Amazon S3 uses permanent redirects to offer steps on how to execute the request properly. To use this feature, you must have an application that can handle Amazon S3 redirect responses.

One of the design necessities of Amazon S3 is extremely high availability. One of the ways to meet this condition is by modifying the IP addresses associated with the Amazon S3 endpoint in DNS as required. These changes are automatically reflected in short-living users, but not in some long-lasting users. Long-lasted users will need to make different actions to re-resolve the Amazon S3 endpoint occasionally to benefit from these changes.





DNS routing request steps

1. The user makes a DNS request to get an object stored on Amazon S3.
2. He receives one or more IP addresses for facilities that can process the request. In this example, the IP address is for Facility B.
3. The client makes a request to Amazon S3 Facility B.
4. Facility B returns a copy of the object to the client.

Temporary request redirection steps

1. The user makes a DNS request to get an object stored on Amazon S3.
2. The user receives one or more IP addresses for facilities that can process the request.
3. The client makes a request to Amazon S3 Facility B.
4. Facility B returns a redirect indicating the object is available from Location C.
5. The client resends the request to Facility C.
6. Facility C returns a replica of the object.

As mentioned before, Amazon S3 is a REST service. You can forward requests to Amazon S3 using the REST API or the AWS SDK wrapper libraries that fold the original Amazon S3 REST API, making your programming responsibilities simpler. Every contact with Amazon S3 is either authenticated or anonymous. And by authentication, they mean a process of validating the identity of the requester trying to use an AWS service. To be Authenticated, a request must contain a signature value that confirms the request source. The signature value is, in part, made from the requester's AWS entree keys (access key ID and secret access key).

Amazon Simple Storage Service is able to access S3 buckets with the Internet Protocol version 6 (IPv6), additionally to the IPv4 protocol. Amazon S3 dual-stack endpoints can hold requests to S3 buckets over IPv6 and IPv4. Accessing Amazon S3 over IPv6 is free of charge.

This architecture diagram explains how the project is implemented. The PNG picture is loaded from Amazon S3 bucket maven.s3auth.com, which is not readable anonymously.

5. Virtualization management

The Amazon S3 service does not allow advanced virtualization management. Indeed, this service allows in terms of virtualization to manage :

- Virtual compartment hosting.
- export and duplicate virtual machines from their image, if these images are on an Amazon S3 bucket and are moved to an EC2 instance.

Virtual compartment hosting :

Virtual hosting consists in distributing the load of several websites on a single server. apparent host (bucket name). Since bucket's names are unique, this will not be a collision problem. The latter allows us a total control over the usage of resources and regions. A host name could be like this:

https://bucket-name.s3.Region.amazonaws.com/key name

where :

- Bucket-name represents the name of the bucket containing the resources,
- Region represents the different regions of Amazon S3 Data centers (see this link https://docs.aws.amazon.com/fr_fr/general/latest/gr/s3.html)
- Key name is the name of the resource.

Exporting and replicating virtual machines from their image:

This part is a bit on the edge of Amazon S3. Indeed, Amazon S3 and Amazon EC2 are intimately linked. Thus Amazon S3 allows, thanks to the command line interface AWS, to import or export images of virtual machines so as to run EC2 instances on them (without additional cost).

The fact of being able to import and export images of existing virtual machines thus makes it possible to make backups or duplicates of the latter. In addition, AWS Server Migration Service (AWS SMS) allows you to select a server and an MV image and duplicate it in order to move it very easily. It even makes it possible to regularly program duplicates in order to keep versions of the server.

6. API / Access :

As mentioned before, you can forward requests to Amazon S3 using the REST API which is provided for creating and handling bucket messages which in their turn can be authenticated or anonymous. Authenticated access requires authorizations that AWS can use to validate your requests. When making REST API calls right from your code, you generate a signature using valid identifications and involve the signature in your request.

Making REST API calls directly from your code can be heavy. It necessitates you to add the needed code to compute a valid signature to confirm your requests. Amazon S3 recommend the following alternatives instead:

Use the AWS SDKs to forward your requests. Using this, means you don't have to write the needed code to compute the signature since the SDK clients verify your requests by using access keys that you provide. Unless you have a better option not to, they believe that we should always use the AWS SDKs and AWS CLI to make Amazon S3 API calls.

Even if we have the valid credentials to authenticate our requests, we cannot create or modify Amazon S3 resources unless we have been granted the permission (permission to create a bucket or get an object for example). Root credentials can grant us all the needed permissions although it's not recommended. IAM users are preferred instead.

A rise in the competing services based on the S3 API, that uses the standard programming interface, has been given by the broad adoption of Amazon S3 and its tools. Although, they have different technologies and support different business models. A cloud storage standard give the opportunity to the providers to make their services supporting communication between all kind of clients and that comes back to them with the following benefits:

1. Inspiring small companies to this market by putting a set of rules and a level of working fields that may make the competition fairer.

2. Turning the focus of the vendors and developers to improving their own products as a replacement for concentrating on compatibility.

3. Give opportunities to increase size and swap solutions at any time easily.

4. It's flexible to the increase in the demands and can be easily changed.

The S3 API has become so popular as a way to store objects. Consequently, many applications have been designed to natively support the Amazon S3 API which includes applications that write data to Amazon S3 and Amazon S3-compatible object stores.

7. Application Domain :

Amazon S3 is designed to perform well on the storage of large amounts of data of all types. Its bucket system makes it possible to store a theoretically infinite amount of data.

In addition, Amazon S3 has a large number of services such as:

- Amazon S3 glacier that allows you to archive data or store long-term backups at a very low price.
- the AWS SMS service which makes it possible to make backups with a regular time interval.
- S3 Intelligent-tiering which offers a service to check the use of user data in order to minimize costs by moving, if necessary, certain misused data from one storage system to another more suitable.

In addition, Amazon S3 guarantees us what it calls the 11 9, that is to say, it gives us a durability of 99.999999999% (there are 11 times the number 9, hence the name) which equivalent, for example by storing 10 million objects with Amazon S3 to lose 1 unique object every 10,000 years

These services thus guarantee a safe use of our data: there is very little risk of losing its data because of the hardware (replication) but in order to convince companies, Amazon S3 must also guarantee data security which is divided in two groups:

- physical security resolved by infrastructure and network inspections
- security of hosts and endpoints resolved via data encryption applications (Thales E-security, ...) as well as a functionality allowing the owner of an account to authorize and manage a large number of users on their data with AWS Identity Services.

8. Visibility (notoriety) :

AWS in general, is leading the cloud market since some time now and its biggest competitors are Google and Microsoft Azure. And that reflects also on the Amazon S3 market place which makes it the leader in its domain, that means storage. One of the reasons that are making S3 lead is the uniqueness of its features, the object storage that is making life easier for a big number of users and the ability to do analytics easily on the data.

According to stackshare, Amazon Simple Storage Service is leading the race way ahead of Azure storage and Google Cloud Storage and is preferred by users for its reliable and user friendly platform.

Some billion dollars companies use Amazon S3 to store their data, take for example:

Netflix, they deliver billions of hours of content to their clients around the world and they also use S3 as the data lake for their Big Data analytics solution. Netflix also invented a tool, S3mper, to address the Amazon S3 limitations of eventual consistency. S3mper stores the filesystem metadata: filenames, directory structure, and permissions in Amazon DynamoDB.

Airbnb keeps its backup data and files on Amazon S3, with over 10 petabytes of user pictures and as a born-in-the-cloud solution, they try to find new ways for the data analysis.

Photo hosting service SmugMug is another big company that has been using Amazon S3 since 2006, and they believe that they have $1 million in storage costs.

FINRA uses Amazon S3 to ingest and store data for over 75 billion market events daily and AWS Lambda functions to format and validate the data against more than 200 rules.

GE uses Amazon S3 to store and protect a petabyte of critical medical imaging data for its GE Health Cloud service, which connects hundreds of thousands of imaging machines and other medical devices

While not forgetting other big companies such as Spotify, Pinterest… or Amazon itself.

9. Documentation :

Through the research for this report, we were helped by the detailed documentation of the Amazon S3 where you can find the steps to get started with Amazon S3 easily while defining S3 buckets and objects.

We can also find details for developers about coding and different programming languages SDKs and how to use them in Amazon S3.

We can also find an API description detailing the signature and what is S3 API made of while helping the user set it up.

We can even find a S3 console guide that helps us navigate through the console and understand the meaning of each button.

Moreover, a very helpful youtube channel give us useful advisor for the understanding of this cloud technologie (https://www.youtube.com/channel/UCT-nPlVzJI-ccQXlxjSvJmw)

One of the problems with this documentation is that one can find translation errors between the English and French versions which sometimes mean the exact opposite.

Another problem is the regular appearance of the Amazon ordering site which saturates Google results with books on the subject sought

Conclusion :

In conclusion, Amazon S3 is a powerful service in its field, that is to say the storage of large amounts of data. And because of the multitude of additional services it provides, it is a tool of choice for any business, small or large.

In addition to its storage and computing power and its ancillary services, Amazon S3 stands out for its price and its ease of use. Its deserved notoriety no longer needs to be redone. We understand why companies like Netflix who need adequate storage space as well as a very frequent access request have chosen to use it.

The use of Cloud technologies such as virtualization or services responding to Cloud problems, for example data replication, show that this technology is indeed part of Cloud technology.

In this article, we have examined the Amazon S3, covering its essential concepts, its business model, its installation, its authentication steps, its network management, as well as its value and importance in the market. As the development of cloud computing technology is so fast, we hope that our work will provide a good description on the subject and help beginners to understand the process and the way to get started easily.