Processing Storage Operations trong AWS

Hướng dẫn thực hiện các thao tác lưu trữ hiệu quả trên Amazon S3

Bucket Management

Object Operations

Temporary Access

Batch Processing

Report Generate Date: 2025-10-14

Tổng quan về Processing Storage Operations

Amazon S3

- Dịch vụ lưu trữ Object hàng đầu trong ngành
- Khả năng mở rộng và tính sẵn sàng dữ liệu vượt trội
- Bảo mật và hiệu suất dữ liệu hàng đầu

Vai trò của thao tác lưu trữ

- Quản lý dữ liệu hiệu quả
- Tối ưu hóa, tổ chức dữ liệu
- Cấu hình quyền truy cập vào dữ liệu

Các trường hợp sử dụng phổ biến

Data Lakes

Lưu trữ và quản lý lượng lớn dữ liệu đa dạng

Website Hosting

Hỗ trợ lưu trữ cho các trang web

Ứng dụng di động

Tải lên và truy cập dữ liệu từ thiết bị di động

Backup & Restore

Sao lưu và khôi phục dữ liệu

Archive Storage

Lưu trữ lâu dài với chi phí thấp

Doanh nghiệp

Quản lý dữ liệu doanh nghiệp

IoT Devices

Thu thập và lưu trữ dữ liệu từ thiết bị

Big Data Analytics

Phân tích dữ liệu lớn

Mục tiêu học tập (Learning Objectives)

Thao tác cơ bản với S3

Hiểu và thực hiện các thao tác cơ bản với S3 Bucket và Object

Sử dụng API

Sử dụng các API cơ bản của Amazon S3 để tương tác với dữ liệu

S3 Select

Tối ưu hóa việc truy vấn dữ liệu trong S3 bằng cách sử dụng S3 Select

Presigned URLs

Cấp quyền truy cập tạm thời và an toàn cho các Object thông qua Presigned URLs

Batch Operations

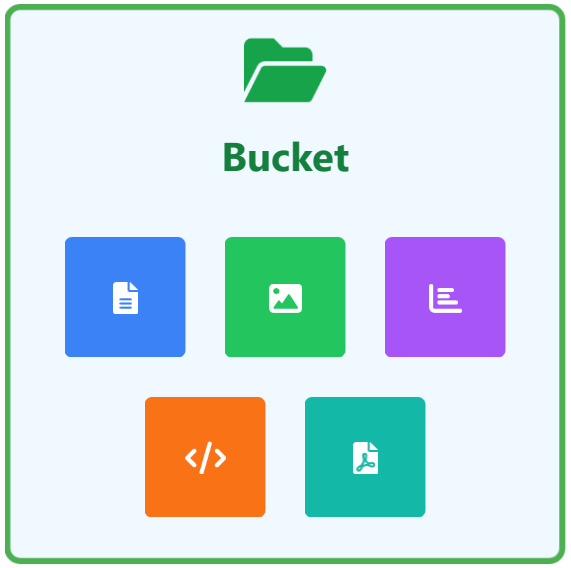
Xử lý và quản lý dữ liệu lớn một cách hiệu quả bằng S3 Batch Operations

Tiến trình học tập

Hoàn thành 0%

Sau khi hoàn thành học phần này, bạn sẽ có thể thực hiện các thao tác lưu trữ trên Amazon S3 một cách hiệu quả và an toàn.

Bucket là gì?



Container cơ bản để lưu trữ các Object trong Amazon S3

Mỗi Object bạn lưu trữ đều phải nằm trong một Bucket

Đặc điểm chính của Bucket

Định danh duy nhất

- Tên Bucket phải là duy nhất trên toàn cầu trong một Region
- Sau khi tạo, không thể thay đổi tên hoặc Region

Tổ chức dữ liệu

- Giúp tổ chức không gian tên (namespace) ở cấp cao nhất
- URL truy cập: https://<bucket-name>.s3.<region>.amazonaws.com/<key>

Quản lý

- Đơn vị để cấu hình quyền truy cập
- Quản lý vòng đời dữ liệu
- Báo cáo chi phí

Một số hiểu biết

> Bucket là đơn vị cơ bản trong Amazon S3, tương tự như thư mục gốc trong hệ thống tệp

- URL của một Object bao gồm tên Bucket, Region và khóa (key) của Object

Các cấu hình chính của Bucket

Permissions

Bucket Policies
Sử dụng ngôn ngữ chính sách dựa trên JSON để định cấu hình quyền truy cập dựa trên tài nguyên

Access Control Lists (ACLs)
Cung cấp quyền đọc và ghi cho các Bucket và Object riêng lẻ

IAM
Quản lý tập trung các quyền kiểm soát tài nguyên AWS

S3 Block Public Access
Ngăn chặn truy cập công khai vào các Bucket và Object S3

Properties

Versioning
Lưu giữ nhiều phiên bản của một Object trong cùng một Bucket

Logging
Ghi lại các yêu cầu được thực hiện đối với Bucket

Static Website Hosting
Cho phép Bucket lưu trữ các trang web tĩnh

Default Encryption
Cấu hình mã hóa mặc định cho các Object được tải lên

Object Management

S3 Lifecycle
Định cấu hình các quy tắc để tự động chuyển đổi Object sang các Storage Class có chi phí thấp hơn hoặc xóa các Object hết hạn

Quy trình vòng đời Object

Các bước tạo một Bucket

1 **Đăng nhập AWS Management Console**

Mở console Amazon S3 tại https://console.aws.amazon.com/s3/

API: HeadBucket

API là một phương thức trong Amazon S3 API để kiểm tra sự tồn tại và quyền truy cập vào bucket.

Mục đích sử dụng

- Kiểm tra xem một bucket có tồn tại hay không
- Xác minh quyền truy cập vào bucket
- Xác nhận tên bucket trước khi thực hiện các thao tác khác

Ví dụ sử dụng AWS CLI

```
aws s3api head-bucket --bucket my-example-bucket
```

HTTP Status Codes

200 OK

Bucket tồn tại và bạn có quyền truy cập

403 Forbidden

Bucket tồn tại, nhưng không có quyền truy cập

404 Not Found

Bucket không tồn tại hoặc không có quyền kiểm tra

Lưu ý quan trọng

API HeadBucket không trả về nội dung, chỉ trả về HTTP status code. Đây là cách hiệu quả để kiểm tra sự tồn tại của bucket mà không cần liệt kê nội dung.

Update Bucket Versioning

Tính năng Versioning

Versioning trong Amazon S3 cho phép lưu giữ nhiều phiên bản của một Object trong cùng một Bucket, giúp bảo vệ dữ liệu khỏi việc bị xóa hoặc ghi đè nhầm.

Cách bật/tắt

- **Bật Versioning:** Trong AWS Management Console, chọn Bucket → Properties → Versioning → Enable
- **Tắt Versioning:** Trong AWS Management Console, chọn Bucket → Properties → Versioning → Suspend

Lợi ích

- Khôi phục các Object bị xóa hoặc ghi đè nhầm
- duy trì lịch sử các phiên bản của dữ liệu
- Trở về phiên bản trước đó của một Object

Cách hoạt động của Versioning

Phiên bản hiện tại (có thể truy cập)

Phiên bản cũ (có thể truy cập)

Phiên bản bị xóa (không thể truy cập)

Quá trình hoạt động

Tải lên Object

Khi bạn tải lên một Object mới, S3 lưu giữ nó như phiên bản đầu tiên

Ghi đè Object

Khi bạn ghi đè lên một Object hiện có, phiên bản cũ được lưu lại

Xóa Object

Khi bạn xóa một Object, nó vẫn được lưu lại dưới dạng phiên bản bị xóa

Các thao tác cơ bản với Object

Các Object là các thực thể cơ bản được lưu trữ trong Amazon S3, bao gồm dữ liệu và metadata mô tả file. Để quản lý hiệu quả các Object này, AWS S3 cung cấp một bộ các thao tác cơ bản:

Upload
Tải một file lên S3 Bucket

List
Liệt kê các Object trong Bucket

Download
Tải Object về máy cục bộ

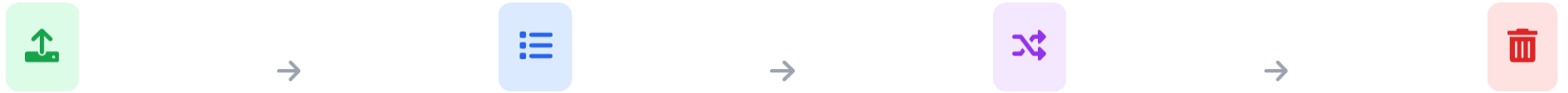
Copy
Tạo bản sao của Object

Move
Di chuyển Object sang vị trí mới

Rename
Đổi tên Object

Delete
Xóa Object khỏi Bucket

Workflow Thao Tác Cơ Bản



Các thao tác này là building blocks cho các ứng dụng lưu trữ S3 phức tạp hơn.

Bulk Operations: Copy, Sync, Batch

Thao tác hàng loạt

Amazon S3 hỗ trợ các thao tác xử lý hàng loạt dữ liệu, giúp quản lý hiệu quả các tập hợp lớn file.

Lợi ích

- Hiệu quả cho lượng lớn dữ liệu
- Giảm thiểu công sức quản lý thủ công
- Thực hiện đồng thời nhiều thao tác

Copy

Sao chép các file từ một Bucket sang Bucket khác hoặc trong cùng Bucket.

```
aws s3 cp s3://source-bucket/ s3://destination-bucket/ --recursive
```

Sync

Đồng bộ hóa nội dung giữa thư mục cục bộ và Bucket S3, chỉ cập nhật những gì cần thiết.

```
aws s3 sync ./local-folder s3://my-bucket/folder/
```

Batch

Thực hiện hàng loạt các thao tác trên nhiều Objects như Copy, Tagging, Restore.

```
aws s3 batch delete s3://my-bucket --from-file objects.txt
```

Thực hiện thao tác hàng loạt

AWS CLI
Giao diện dòng lệnh cho phép thực hiện các thao tác hàng loạt một cách dễ dàng

Manifest Files
Danh sách các Objects cần xử lý, hỗ trợ cả CSV và JSON

Thao tác PUT (Upload Objects)

Giải thích thao tác PUT

Thao tác chính để tải lên

Phương thức chính để tải một Object lên Amazon S3

Cách cơ bản nhất

Đưa dữ liệu vào một Bucket đơn giản và hiệu quả

Tùy chọn cấu hình

Khi sử dụng PUT, bạn có thể chỉ định:

- Metadata tùy chỉnh
- Storage Class (lớp lưu trữ)
- Tùy chọn mã hóa

Ví dụ và Cách Sử Dụng

Cách sử dụng AWS CLI

AWS CLI

```
aws s3 cp my-file.txt s3://my-bucket/my-folder/my-file.txt
```

Giải thích ví dụ

- my-file.txt: File local cần tải lên
- s3://my-bucket: Bucket đích trên S3
- my-folder/my-file.txt: Đường dẫn lưu trên Bucket

Lưu ý quan trọng

Sau khi tải lên, bạn có thể truy cập Object thông qua URL: https://my-bucket.s3.region.amazonaws.com/my-folder/my-file.txt

Multipart Upload cho file lớn

Giới hạn kích thước

Hỗ trợ các file từ:
100 MB đến 5 TB

Lợi ích

Cải thiện thông lượng
Tải lên các phần song song giúp tăng tốc độ truyền dữ liệu

Phục hồi nhanh chóng
Nếu một phần bị lỗi, chỉ cần tải lại phần đó thay vì toàn bộ file

Tạm dừng và tiếp tục
Có thể tạm dừng và tiếp tục quá trình tải lên

Tải lên trước khi biết kích thước cuối cùng
Bắt đầu tải lên Object ngay khi nó đang được tạo ra

Quy trình 3 bước

1

Initiate Multipart Upload

Khởi tạo quá trình tải lên, S3 trả về một **UploadId** duy nhất

→ S3 trả về: UploadId

2

Upload Parts

Tải từng phần của Object lên S3, mỗi phần có một số thứ tự (**PartNumber**) và **ETag** riêng

→ S3 trả về: ETag cho mỗi phần

3

Complete Multipart Upload

Gửi yêu cầu hoàn tất, S3 sẽ ghép các phần lại thành một Object duy nhất

✓ Kết quả: Object hoàn chỉnh được lưu trữ trong Bucket

GET và HEAD (Retrieving data & Metadata)

So sánh GET và HEAD

GET Operation

- Tải về toàn bộ nội dung dữ liệu của Object
- Sử dụng khi cần truy cập vào dữ liệu thực tế của file
- Trả về cả metadata và payload (nội dung file)
- Tốn băng thông và thời gian tải toàn bộ dữ liệu

HEAD Operation

- Truy xuất chỉ metadata của Object
- Hữu ích khi chỉ cần kiểm tra sự tồn tại, kích thước, thời gian sửa đổi
- Giết kiệm băng thông và thời gian
- Trả về: Content-Length, Content-Type, ETag, Last-Modified, x-amz-version-id

Trường hợp áp dụng

Khi sử dụng GET

- Người dùng cần xem hoặc tải về file hoàn chỉnh

Khi sử dụng HEAD

- Kiểm tra sự tồn tại của Object trước khi tải

Tối ưu truy vấn với S3 Select

S3 Select là gì?

S3 Select là một tính năng mạnh mẽ cho phép bạn lọc nội dung của một Object ngay trên Amazon S3 bằng cách sử dụng các câu lệnh SQL đơn giản.

Cách hoạt động

- Thay vì tải toàn bộ Object về ứng dụng của bạn và sau đó lọc dữ liệu
- S3 Select thực hiện việc lọc này trên S3
- Chỉ trả về phần dữ liệu bạn cần

Cú pháp SQL

```
SELECT * FROM S3Object [WHERE ...]
```

Lưu lượng dữ liệu

Định dạng hỗ trợ

- CSV
- JSON
- Apache Parquet

Lợi ích

Hiệu năng

Giảm đáng kể lượng dữ liệu truyền qua mạng, tăng tốc độ truy vấn

Chi phí

Giảm chi phí Data Transfer Out vì chỉ dữ liệu được lọc mới được truyền đi

Giảm tải cho ứng dụng

Ứng dụng không cần xử lý lượng lớn dữ liệu không cần thiết

Ví dụ Demo S3 Select

Kịch bản ví dụ

File CSV giao dịch

Bạn có một file CSV **transactions.csv** trong S3 Bucket chứa hàng triệu dòng dữ liệu giao dịch.

Mỗi dòng có các trường: **transaction_id**, **customer_id**, **amount**, **date**.

Truy vấn cần thực hiện

Bạn muốn tìm tất cả các giao dịch của **customer_id = 'C123'** có **amount** lớn hơn 1000.

Lợi ích của S3 Select

- Giảm đáng kể lượng dữ liệu truyền qua mạng
- Giảm chi phí Data Transfer Out
- Ứng dụng không cần xử lý lượng lớn dữ liệu không cần thiết

Cú pháp SQL với S3 Select

```
SELECT s.transaction_id, s.amount, s.date
FROM S3Object s
WHERE s.customer_id = 'C123'
AND s.amount > 1000
```

Ví dụ về kết quả trả về

transaction_id	customer_id	amount	date
TX123456	C123	1500	2025-09-15
TX789012	C123	1250	2025-09-20
TX345678	C123	1800	2025-09-25

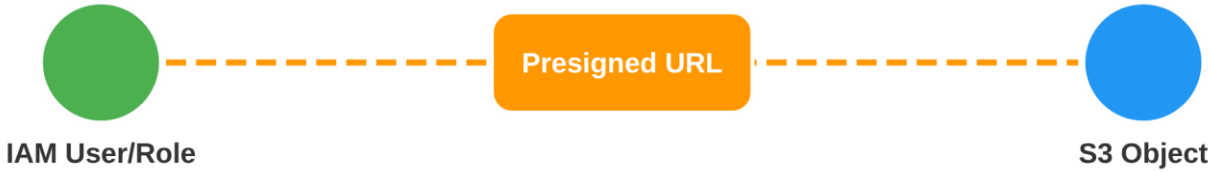
Presigned URLs là gì?

Định nghĩa

> Presigned URL là một URL đặc biệt được tạo ra để cung cấp quyền truy cập tạm thời và có giới hạn thời gian đến một Object cụ thể trong Amazon S3.

Cách hoạt động

- Người dùng sở hữu Presigned URL không cần thông tin xác thực AWS
- URL được ký điện tử (cryptographically signed) bằng thông tin xác thực của người tạo
- URL có thời gian hết hạn (mặc định 15 phút, tối đa 7 ngày)
- Chỉ cho phép các hành động được chỉ định (ví dụ: chỉ GET hoặc chỉ PUT)



Hết hạn sau 15 phút (tùy chỉnh được)

Sử dụng phổ biến

Chia sẻ file an toàn

Cho phép người dùng không có tài khoản AWS tải xuống các báo cáo, hóa đơn

Tải lên từ client

Ứng dụng di động hoặc web tải ảnh, video lên S3 mà không cần nhúng thông tin xác thực

Truy cập tạm thời cho đối tác

Cung cấp quyền truy cập có giới hạn cho các đối tác bên ngoài

Presigned URLs khác với việc cấp quyền truy cập công khai: chúng có thời gian hết hạn và có thể bị hủy bỏ.

Cách hoạt động và trường hợp sử dụng Presigned URLs

Cách hoạt động

IAM principal (user/role) có quyền `s3:GetObject`

Thực tiễn tốt nhất về bảo mật (Best Practices)

Bảo vệ Presigned URLs

Presigned URLs cung cấp quyền truy cập tạm thời đến Object S3, do đó việc bảo vệ chúng là cực kỳ quan trọng.

Bất kỳ ai có URL này đều có thể truy cập tài nguyên được cấp quyền!

Ví dụ rủi ro

Nếu URL bị lộ, ai đó có thể tải xuống hoặc tải lên dữ liệu (tùy thuộc vào quyền được cấp) cho đến khi URL hết hạn.

Giới hạn thời gian hết hạn

Sử dụng thời gian hết hạn ngắn nhất có thể (principle of least privilege). Điều này giảm thiểu rủi ro nếu URL bị lộ.

Chỉ cấp quyền cần thiết

Khi tạo Presigned URL, chỉ cấp các quyền tối thiểu mà người nhận cần (ví dụ: chỉ `GET` nếu mục đích là tải xuống).

Bảo vệ Presigned URL

Coi Presigned URL như một thông tin nhạy cảm. Tránh lưu trữ hoặc truyền tải URL qua các kênh không an toàn.

Giới hạn quyền của người tạo

Đảm bảo IAM user/role tạo Presigned URL chỉ có quyền truy cập vào các Object mà họ được phép chia sẻ.

Sử dụng phiên tạm thời

Khi tạo Presigned URL, nên sử dụng các phiên tạm thời (temporary sessions) thay vì thông tin xác thực dài hạn để tăng cường bảo mật.

S3 Batch Operations

Giới thiệu

S3 Batch Operations là một giải pháp được quản lý của Amazon S3, cho phép thực hiện các thao tác lưu trữ trên hàng triệu hoặc hàng tỷ Objects một cách hiệu quả và tự động chỉ với một yêu cầu duy nhất.

Cách hoạt động

Tạo Job

Tạo một Job mới trong S3 Batch Operations

Manifest

Cung cấp danh sách các Objects cần xử lý

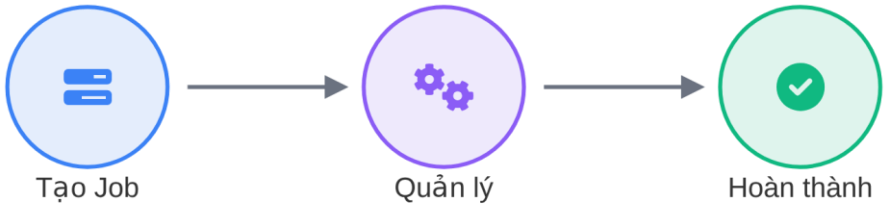
Operation

Chỉ định hành động muốn thực hiện

Quản lý

S3 Batch Operations quản lý thực thi, thử lại, theo dõi tiến độ

Quy trình



Lợi ích

Hiệu quả

Xử lý khối lượng lớn dữ liệu mà không cần phát triển mã tùy chỉnh phức tạp

Tự động hóa

Giảm thiểu công sức quản lý thủ công

Các loại thao tác được hỗ trợ trong S3 Batch Operations

Copy objects

Sao chép Objects từ Bucket này sang Bucket khác, hoặc trong cùng một Bucket.

Hoạt động sao chép hàng loạt giữa các Bucket

Replace tags

Thay thế hoặc thêm các Tag mới cho Objects.

Cập nhậtMetadata và Tagging cho nhiều Objects cùng lúc

Restore objects

Khôi phục Objects đã được lưu trữ trong các lớp lưu trữ S3 Glacier.

Tự động hóa quá trình khôi phục từ lưu trữ dài hạn

Invoke Lambda

Kích hoạt một hàm Lambda trên mỗi Object để thực hiện các tác vụ tùy chỉnh.

Thực hiện các xử lý tùy chỉnh như chuyển đổi định dạng

Change Object Lock

Áp dụng hoặc thay đổi cài đặt Object Lock để bảo vệ Objects khỏi bị xóa hoặc ghi đè.

Bảo vệ dữ liệu quan trọng khỏi việc bị vô tình xóa

Update ACLs

Thay đổi Access Control Lists (ACLs) cho các Objects.

Cập nhật quyền truy cập cho nhiều Objects cùng lúc

S3 Batch Operations hỗ trợ thực hiện các thao tác này trên hàng triệu Objects một cách hiệu quả và tự động

Xử lý tập kết quả lớn (Large Result Sets)

Cơ chế phân trang (Pagination)

- Amazon S3 trả về kết quả theo từng trang (paginated)
- Mỗi phản hồi `ListObjectsV2`

Ví dụ: Batch Copy hàng ngàn Objects

Kịch bản

- Sao chép 100,000 file log từ Bucket source-logs

Tóm tắt nội dung chính

Quản lý Buckets và Objects

- Tạo và cấu hình Bucket với thuộc tính như Permissions, Versioning
- Thực hiện thao tác cơ bản: Upload, List, Download, Copy, Move, Rename, Delete
- Quản lý vòng đời của Objects thông qua Lifecycle

Thao tác API cơ bản

- Upload Objects sử dụng PUT và Multipart Upload
- Download và HEAD Objects
- API HeadBucket và HeadObject để kiểm tra sự tồn tại và metadata

Tối ưu hóa với S3 Select

- Truy vấn và lọc dữ liệu trực tiếp trên Object S3
- Sử dụng cú pháp SQL để thực hiện các câu lệnh SELECT
- Giảm chi phí và tăng hiệu suất so với tải về client

Cấp quyền truy cập tạm thời

- Presigned URLs để chia sẻ quyền truy cập Objects một cách an toàn
- Thời gian hết hạn có thể cấu hình (từ 15 phút đến 7 ngày)
- Thực tiễn bảo mật: giới hạn thời gian, chỉ cấp quyền cần thiết

Xử lý dữ liệu lớn

- S3 Batch Operations để thực hiện các tác vụ hàng loạt
- Hỗ trợ nhiều loại thao tác: Copy, Tagging, Restore, Invoke Lambda
- Quản lý tự động: thử lại, theo dõi tiến độ, báo cáo kết quả

Xử lý tập kết quả lớn

- Phân trang kết quả trả về từ API
- Sử dụng NextContinuationToken để duyệt qua tất cả các trang
- Thực hiện vòng lặp các yêu cầu để xử lý toàn bộ tập kết quả

Câu hỏi ôn tập

1 Multipart Upload

Khi nào bạn nên sử dụng Multipart Upload thay vì thao tác PUT thông thường để tải Object lên S3?

Hint: Xem xét kích thước file và nhu cầu về hiệu suất

2 Presigned URLs

Sự khác biệt chính giữa Presigned URL và việc cấp quyền truy cập công khai cho một Object là gì, và trường hợp sử dụng nào phù hợp cho mỗi phương pháp?

Hint: Xem xét về bảo mật và thời gian truy cập

3 S3 Select

Lợi ích của việc sử dụng S3 Select so với việc tải toàn bộ file về client để xử lý là gì, đặc biệt về hiệu năng và chi phí?

Hint: Xem xét về lượng dữ liệu truyền và xử lý

4 S3 Batch Operations

Hãy mô tả một kịch bản thực tế mà S3 Batch Operations có thể giúp giải quyết một cách hiệu quả.

Hint: Xem xét về xử lý hàng loạt Objects

Lưu ý

Các câu hỏi này nhằm kiểm tra kiến thức quan trọng về các thao tác lưu trữ trong Amazon S3. Hãy dành thời gian để suy nghĩ về từng câu trước khi xem đáp án.