Amazon Simple Storage Service (S3): Important Features

Andru Estes
Principal Author
andru-estes

Performing Batch Operations with Amazon S3

S3 Batch Operations

Large-scale

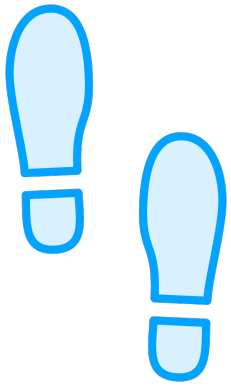
Useful for performing a single operation on a large-scale set of S3 objects at one time

Jobs

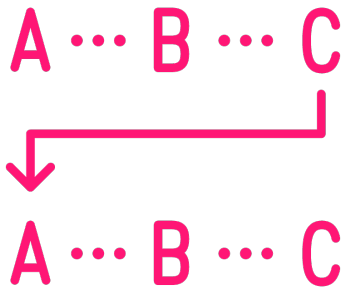
Batch operations are ran via jobs, which use lists of objects, actions to perform, and optional parameters

Image Sources: https://unsplash.com/

What Does S3 Batch Operations Manage?

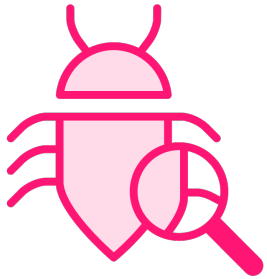


Tracking Job Progress



Retrying Failed Jobs

Notifications of Job Progress



Reporting After Job Completion

S3 Batch Operations Use Cases

Copy or move vast amounts of data between S3 buckets or different storage classes

Add or update tags on a large number of S3 objects

Integrate with other AWS services to transform data formats for analytics

Encrypt thousands or more of unencrypted objects currently living in Amazon S3

Pro Tip: Leverage a combination of S3 Inventory and S3 Select to gather and filter your list of objects.

Filter Objects Using S3 Select and S3 Glacier Select

“After careful consideration, we have made the decision to close new customer access to Amazon S3 Select and Amazon S3 Glacier Select, effective July 25, 2024.”

Citation: https://aws.amazon.com/blogs/storage/

S3 Select

A feature allowing you to use SQL statements to filter the contents of an S3 object and retrieve only the subset of data that you need.

Citation: https://docs.aws.amazon.com/AmazonS3/latest/userguide/selecting-content-from-objects.html

What S3 Select Offers

Works via server-side filtering (*on the Amazon S3 side*)

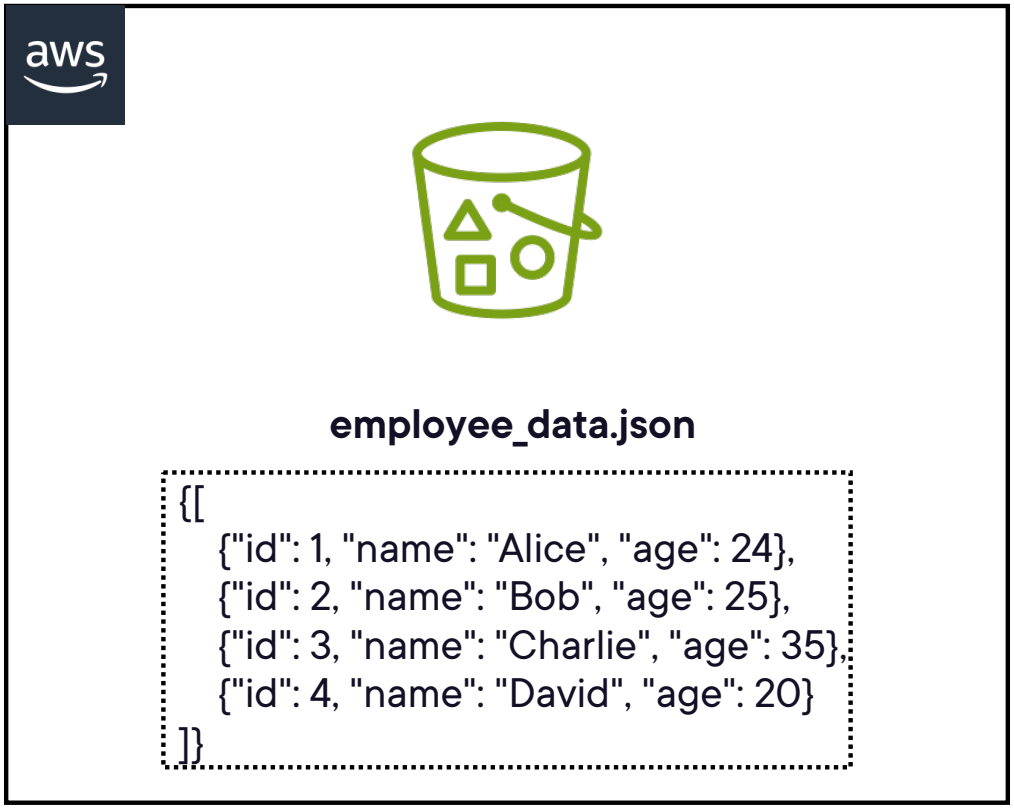
You can only query one object at a time

Supports CSV, JSON, or Apache Parquet format. Also supports GZIP or BZIP2, as long as the compressed objects are CSV or JSON format.

Exam Tip: Reduces the amount of data that S3 transfers, which reduces the cost and latency to retrieve this data!

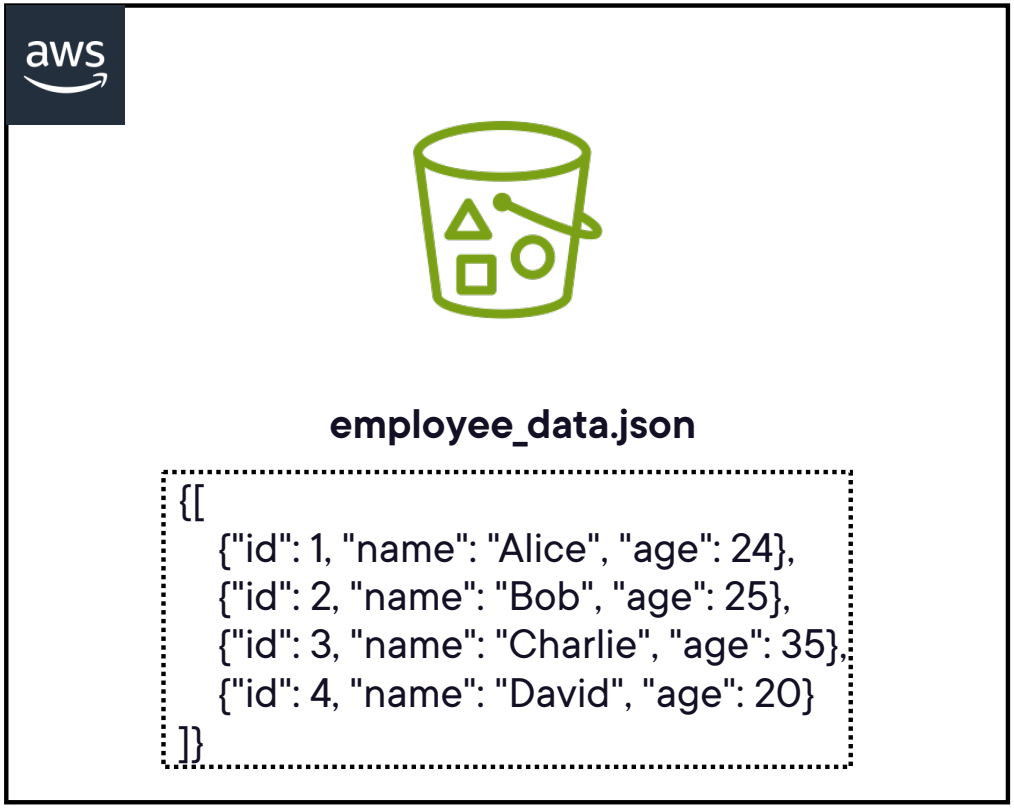
Glacier Select offers the same functionality of S3 Select, but for objects within S3 Glacier.

S3 Select Architecture Example

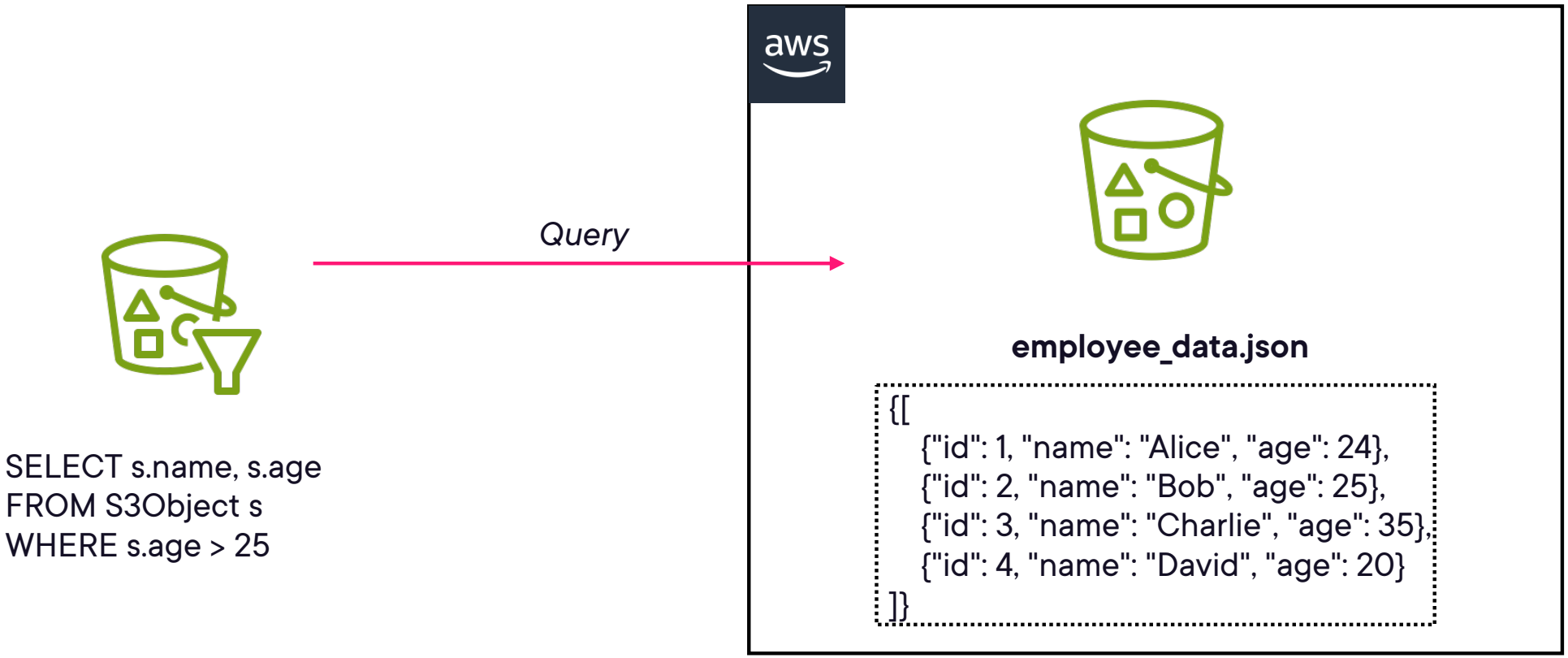


S3 Select Architecture Example

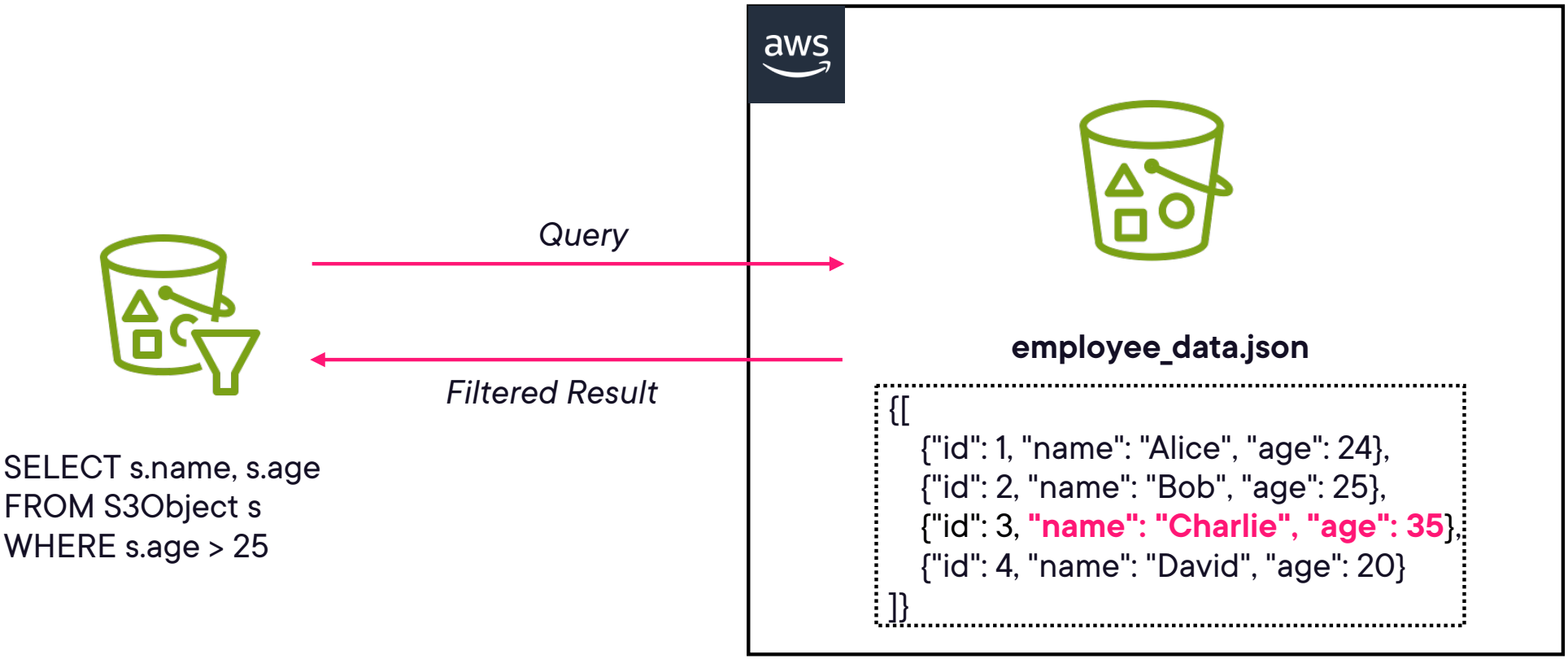
```
SELECT s.name, s.age
FROM S3Object s
WHERE s.age > 25
```



S3 Select Architecture Example



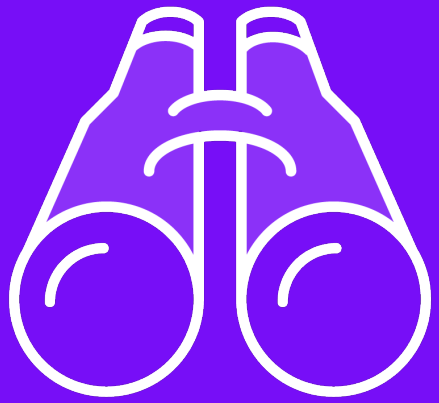
S3 Select Architecture Example



Exam Pro Tip: S3 Select and Glacier Select are best for smaller, simple queries!

Large-scale data retrieval is best done via other AWS services and tools.

Analyzing Data Using S3 Storage Lens



Amazon S3 Storage Lens

AWS cloud-storage analytics feature used to gain organization-wide visibility into object storage and activity

What Does Storage Lens Offer?

Uses metrics to generate summary insights

Aggregates metrics and displays the information in a dashboard

Easily export metrics in CSV or Parquet format to another S3 bucket

Capable of publishing usage and activity metrics to Amazon CloudWatch

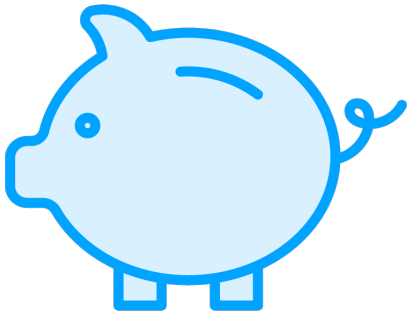
S3 Storage Lens Use Cases

Put best practices into place for data protection within an AWS Organization

Look for ways to become more cost effective with stored objects

Find and report on any data anomalies occurring within your buckets

Storage Lens Metric Category Tiers



Free

Summary, cost optimization, data protection, access management, performance, and event metrics

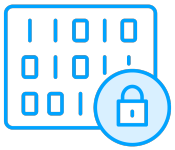
Advanced

Advanced data-protection and cost-optimization metrics, activity metrics, and detailed status-code metrics

Metrics - Part 1

Summary: General insights about your S3 storage, like total storage bytes

Cost-optimization: Help manage and optimize your storage costs



Data-protection: Insights for data-protection, like object encryption

Access-management: Insights for S3 Object Ownership

Metrics - Part 2

Event: Insights for S3 Event Notifications

Performance: Insights for S3 Transfer Acceleration

Activity (Advanced): Details about how storage is requested/uploaded

Status Code (Advanced): Detailed metrics regarding HTTP status codes

Receiving Event Notifications from Amazon S3

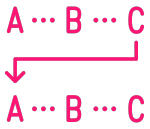
S3 Event Notifications allow you to receive notifications when certain events happen in your S3 bucket.

Curated List of Possible Events

New object created	Object removal	Replication event
Restored objects	Lifecycle events	Object tagging event

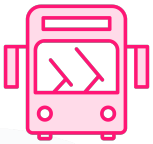
S3 Event Notification Destinations

Amazon SNS Topics. These require topic access policy permissions.



Amazon SQS Queues. These require queue access policy permissions.

AWS Lambda Functions. Require a resource policy allowing invocation.



Amazon EventBridge Buses. Allow for complex filtering and forwarding.

Use Case Examples

Image Processing

Picture uploaded to an S3 bucket triggers a Lambda function to resize it to a thumbnail and upload to a new bucket

Compliance Monitoring

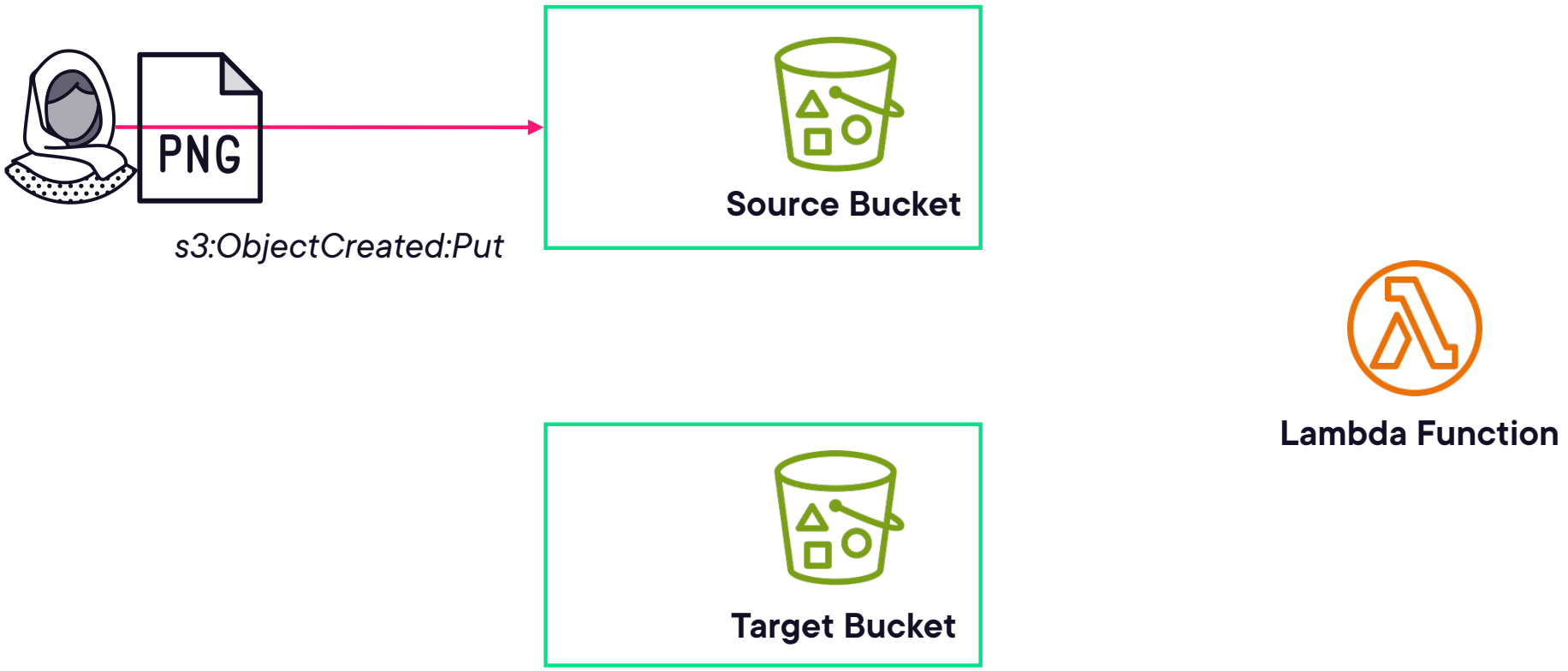
Object uploads trigger workflow to monitor objects for sensitive data and ensure compliance with data protection regulations

Remember that S3 Event Notifications are not real-time!

Events can actually take up to a couple of minutes to be delivered.

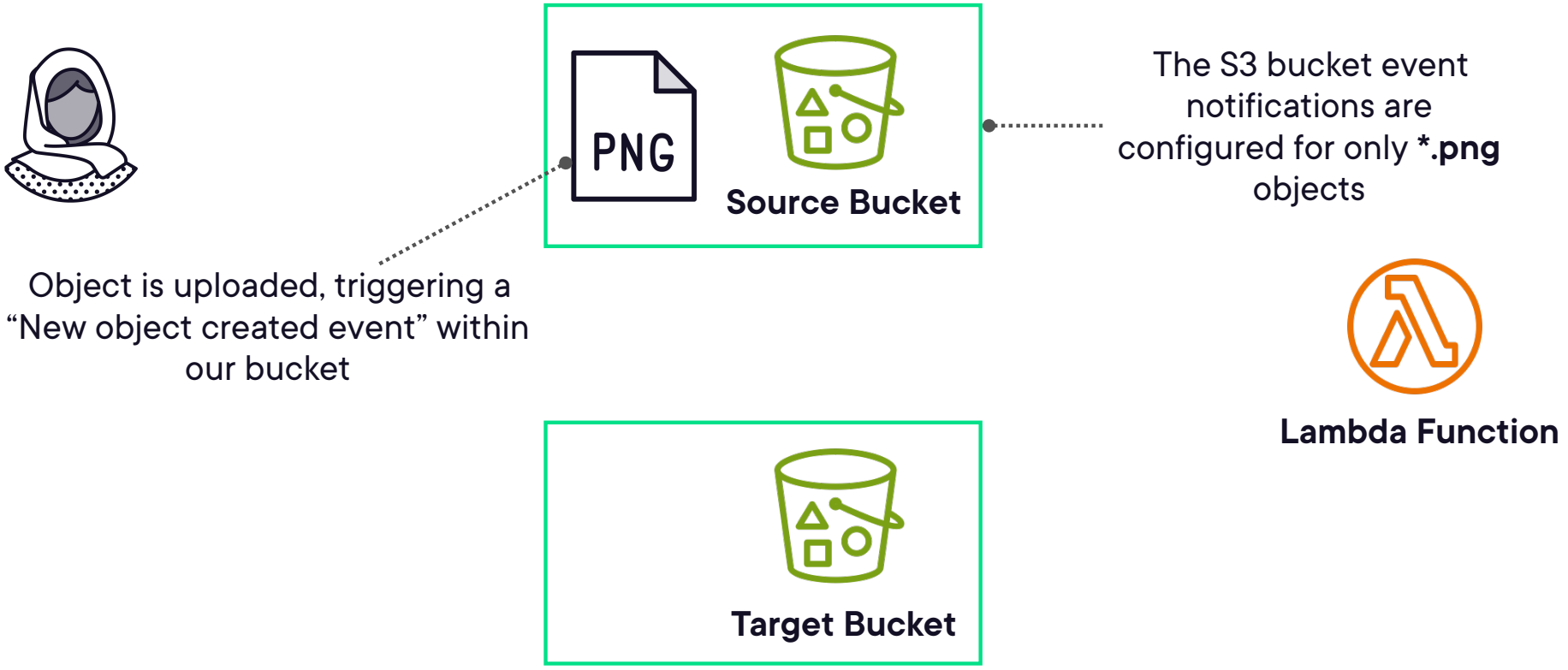
Demo: Trigger a Lambda Function Using Event Notifications

Resizing an image



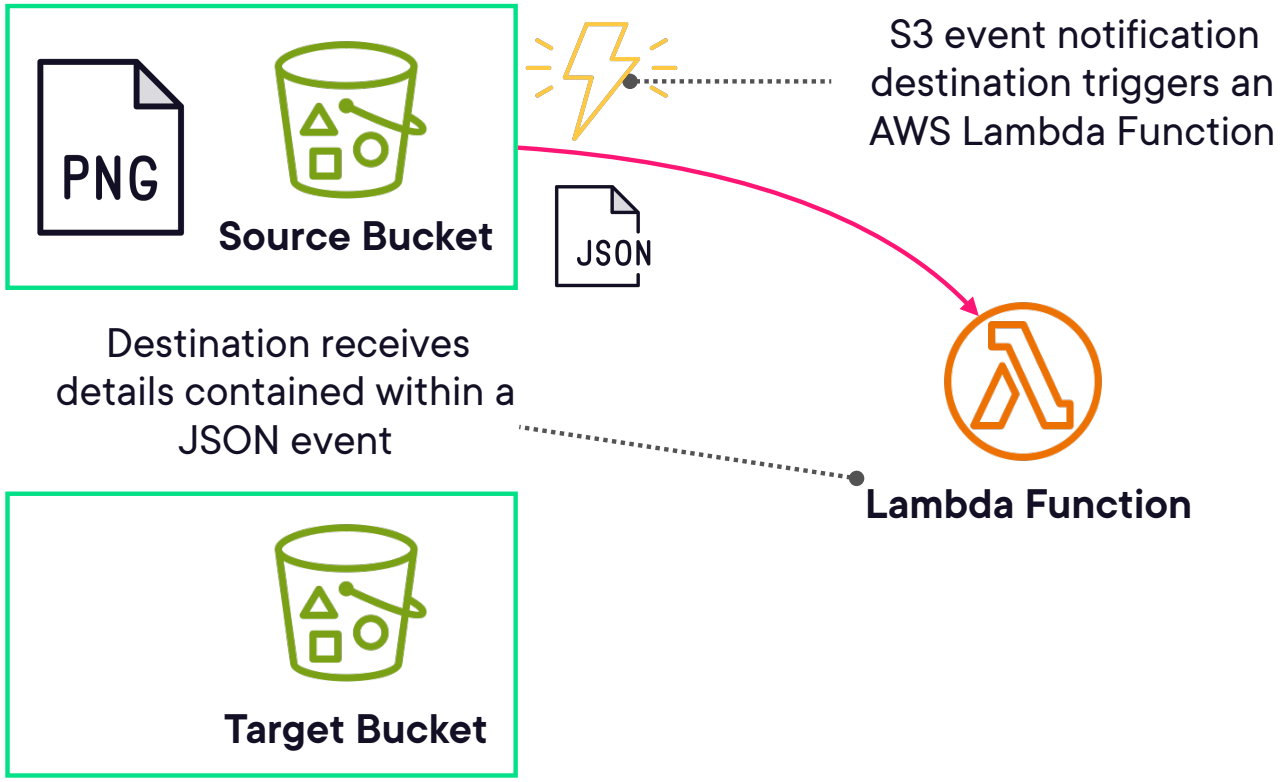
Demo: Trigger a Lambda Function Using Event Notifications

Resizing an image



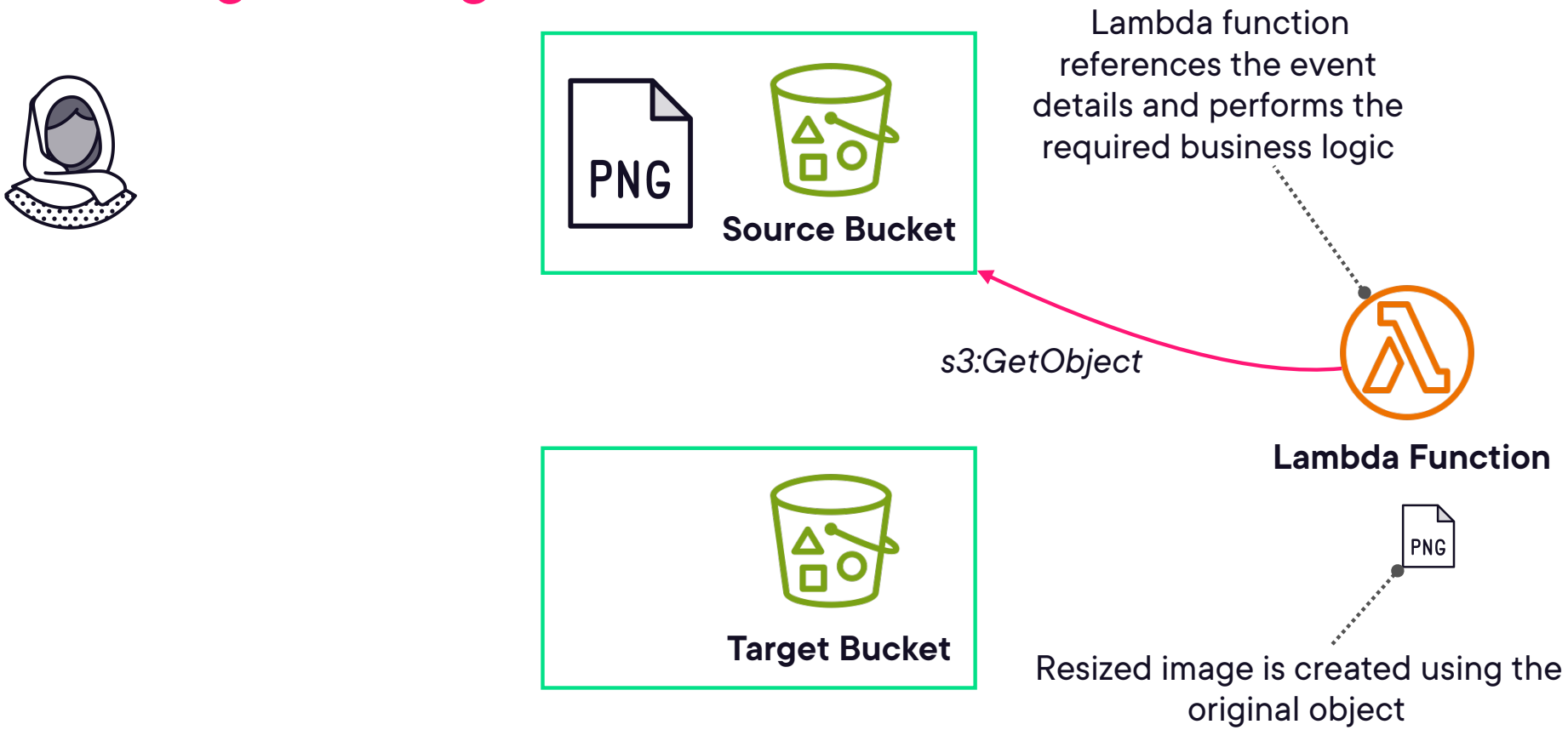
Demo: Trigger a Lambda Function Using Event Notifications

Resizing an image



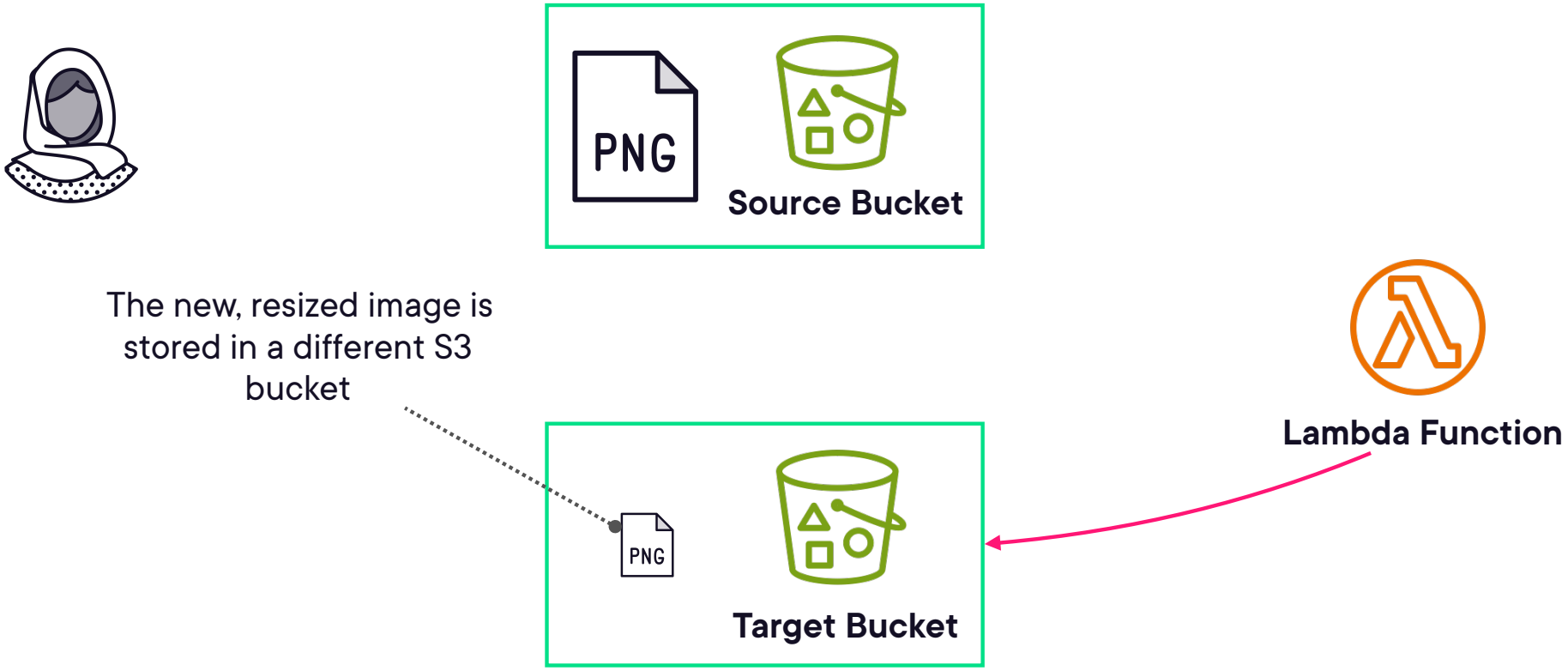
Demo: Trigger a Lambda Function Using Event Notifications

Resizing an image

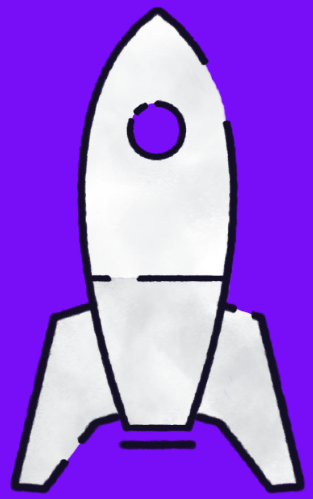


Demo: Trigger a Lambda Function Using Event Notifications

Resizing an image



Faster Content Transfer with S3 Transfer Acceleration



Amazon S3 Transfer Acceleration

Bucket feature that speeds up data uploads and downloads to and from Amazon S3 buckets

Optimizes transfer speeds from across the world by using customer-facing, globally distributed edge locations.

Reasons for Using Transfer Acceleration

Want to upload all objects to a central bucket from customers spread around the globe

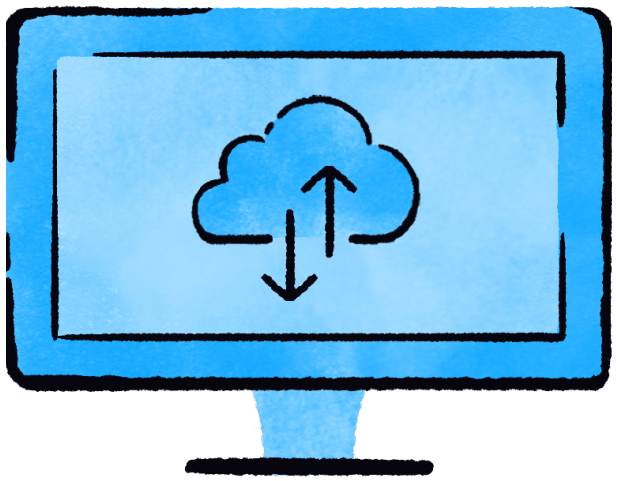
Need to transfer terabytes of data on a regular schedule from around the world

You are not able to use all of your available bandwidth over the internet

It is good to know that using Transfer Acceleration might not always result in a performance gain!

Offloading Costs Using S3 Requester Pays

Requester Pays



The typical, default behavior for Amazon S3 is that the bucket owner pays the costs associated with:

- Storage volume of objects
- Data transfer out of S3 for requests

Requester Pays is a feature that allows you to offload the costs of requests (*transfer and storage*) to the originating caller

The requester cannot be anonymous! They must be authenticated via AWS IAM.

You configure the entire bucket to be Requester Pays!

Requesters must either include x-amz-request-payer header...

Or they can add
the “RequestPayer” parameter
in their REST request.

Serving Websites from S3 with Website Endpoints

Static Websites on Amazon S3

You can use S3 to host static websites, such as .html sites

Client-side scripts are okay, but **no server-side processing** is allowed

Dynamic websites, such as those that require database connections, cannot be hosted on S3!

Need to allow public read access for these to work!

Image Source: https://unsplash.com/

Amazon S3 scales automatically to handle demand for your static websites!

S3 Website Endpoint Examples

When you configure your bucket as a static website, the website is available at the AWS Region-specific website endpoint of the bucket.

The endpoint syntax varies based on the Region your bucket is in.

Dash Example

http://pluralsight.s3-website-us-east-1.amazonaws.com

S3 Website Endpoint Examples

When you configure your bucket as a static website, the website is available at the AWS Region-specific website endpoint of the bucket.

The endpoint syntax varies based on the Region your bucket is in.

Dash Example

http://pluralsight.s3-website-us-east-1.amazonaws.com

Bucket name

S3 Website Endpoint Examples

When you configure your bucket as a static website, the website is available at the AWS Region-specific website endpoint of the bucket.

The endpoint syntax varies based on the Region your bucket is in.

Dash Example

http://pluralsight.s3-website-us-east-1.amazonaws.com

Region with a dash (-) in front

S3 Website Endpoint Examples

When you configure your bucket as a static website, the website is available at the AWS Region-specific website endpoint of the bucket.

The endpoint syntax varies based on the Region your bucket is in.

Dot Example

http://pluralsight.s3-website.us-east-1.amazonaws.com

S3 Website Endpoint Examples

When you configure your bucket as a static website, the website is available at the AWS Region-specific website endpoint of the bucket.

The endpoint syntax varies based on the Region your bucket is in.

Dot Example

http://pluralsight.s3-website.us-east-1.amazonaws.com

Region with a dot (.) in front

Demo: Host a Website in Amazon S3

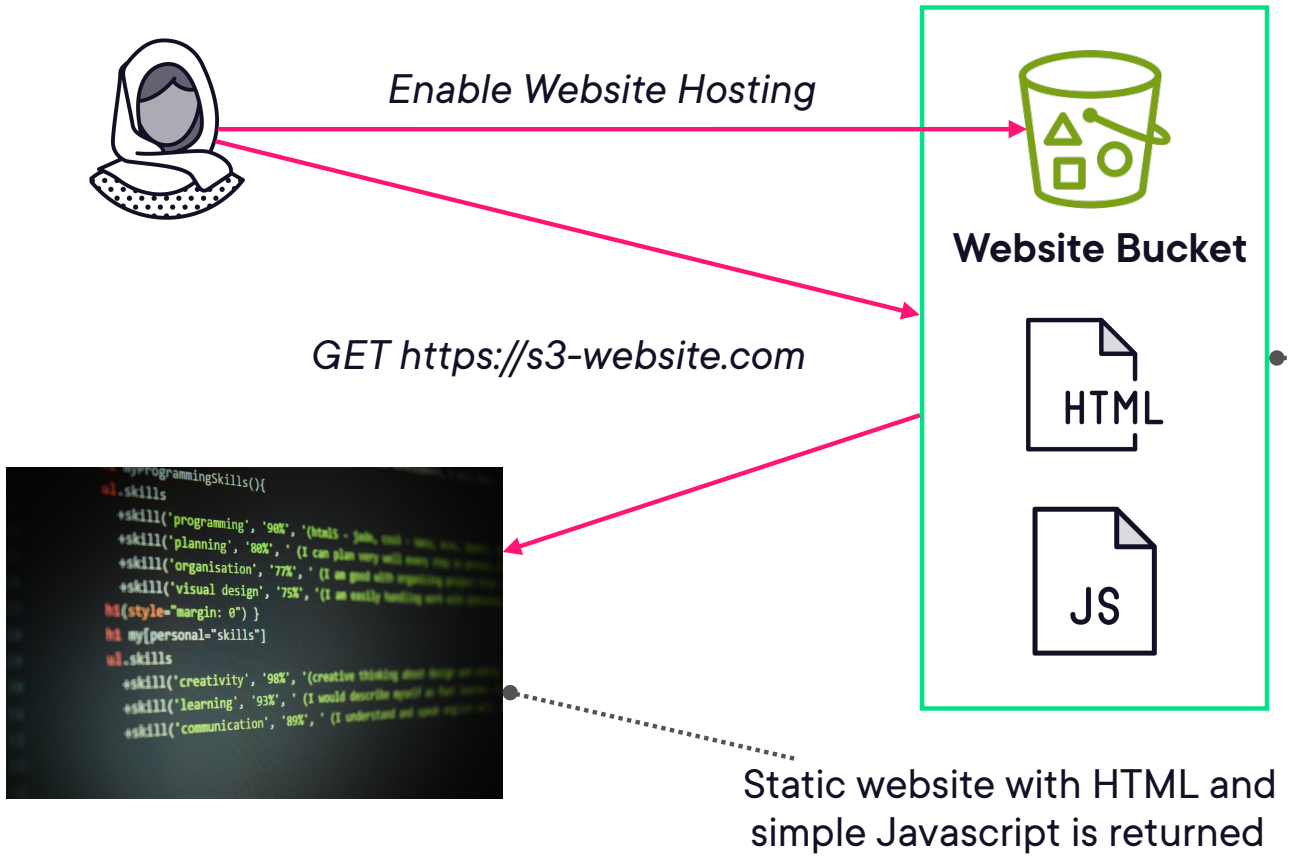


Image Source: http://unsplash.com

Optimizing S3 Performance

Amazon S3 Performance

S3 has extremely low latency, allowing you to get the first byte out of S3 within 100-200 milliseconds

3,500 PUT/COPY/ POST/DELETE HTTP requests per second, per prefix

5,500 GET/HEAD HTTP requests per second, per prefix

Exploring S3 Prefixes

Maximize your S3 performance by leveraging different prefixes!

mybucketname/folder1/subfolder1/myfile.jpg

mybucketname/folder2/subfolder1/myfile.jpg

mybucketname/folder3/myfile.jpg

mybucketname/folder4/subfolder4/myfile.jpg

Exploring S3 Prefixes

Maximize your S3 performance by leveraging different prefixes!

mybucketname**/folder1/subfolder1**/myfile.jpg

mybucketname**/folder2/subfolder1**/myfile.jpg

mybucketname**/folder3**/myfile.jpg

mybucketname**/folder4/subfolder4**/myfile.jpg

Each highlighted portion is a prefix.

So we have four prefixes:
- /folder1/subfolder1
- /folder2/subfolder1
- /folder3
- /folder4/subfolder4

Prefix Performance

You can get better performance by spreading your reads across different prefixes

If we used the 4 prefixes in the last example, you could achieve 22,000 GET requests per second (5,500 * 4)

S3 Byte-Range Fetches

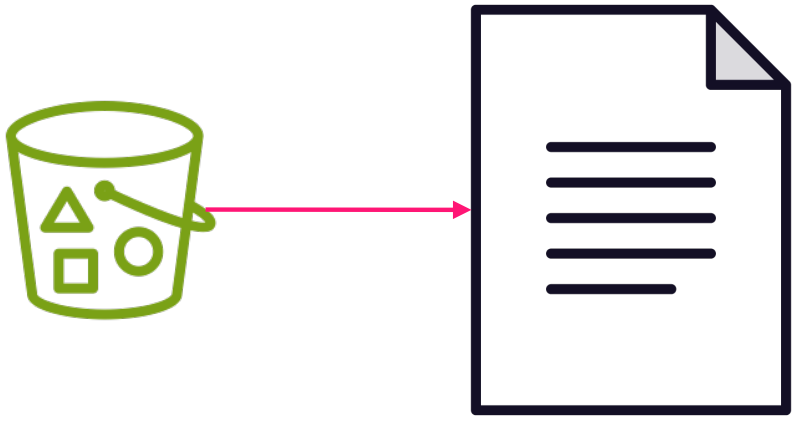
Parallelize downloads by specifying byte ranges (*specific portions of an object*).
If there's a failure in the download, it's only for the specific byte range.
This offers better resiliency for download failures and higher aggregate throughput!

S3 Byte-Range Fetches

Parallelize downloads by specifying byte ranges (*specific portions of an object*).
If there's a failure in the download, it's only for the specific byte range.
This offers better resiliency for download failures and higher aggregate throughput!

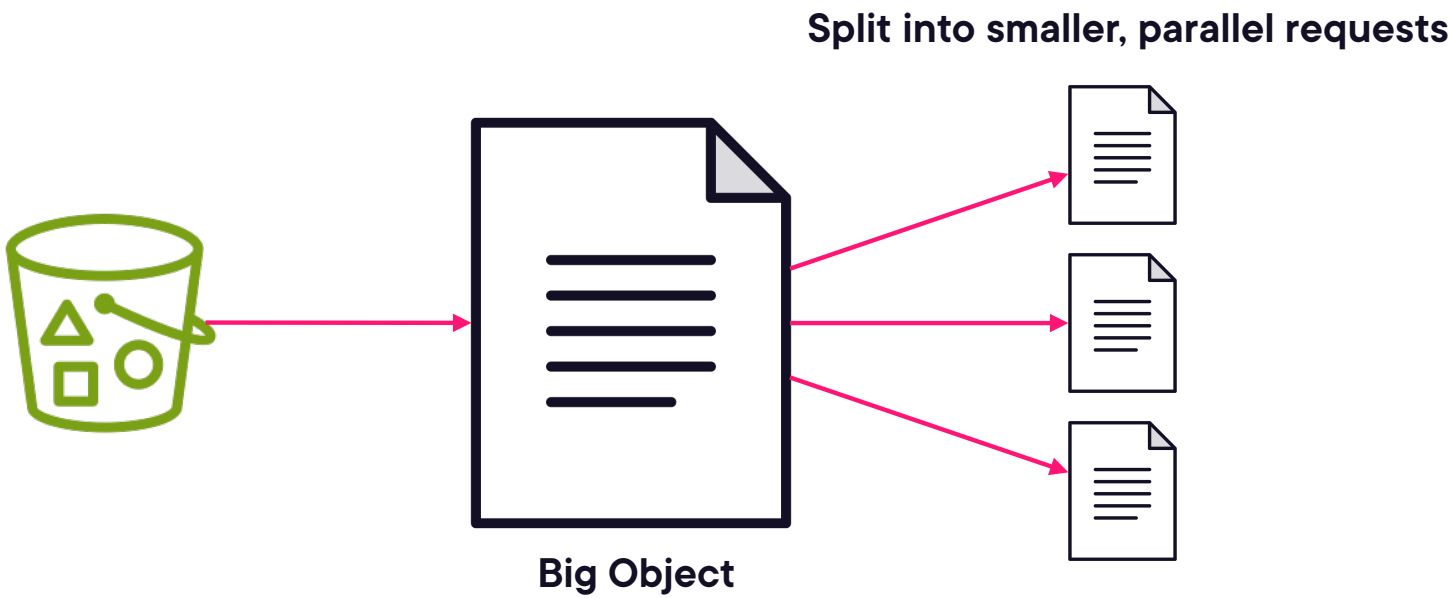
S3 Byte-Range Fetches

Parallelize downloads by specifying byte ranges (*specific portions of an object*).
If there's a failure in the download, it's only for the specific byte range.
This offers better resiliency for download failures and higher aggregate throughput!



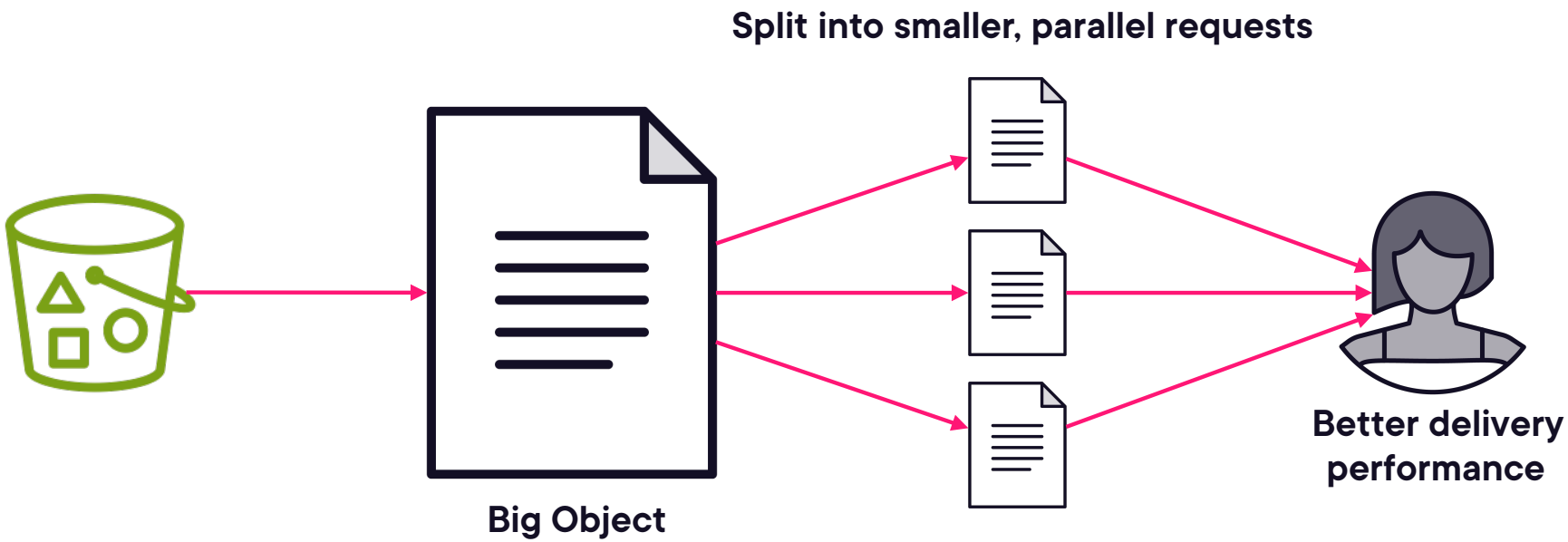
S3 Byte-Range Fetches

Parallelize downloads by specifying byte ranges (*specific portions of an object*).
If there's a failure in the download, it's only for the specific byte range.
This offers better resiliency for download failures and higher aggregate throughput!



S3 Byte-Range Fetches

Parallelize downloads by specifying byte ranges (*specific portions of an object*).
If there's a failure in the download, it's only for the specific byte range.
This offers better resiliency for download failures and higher aggregate throughput!



Pro Tip: If you use multipart upload to PUT an object into S3, then typically it is good to GET them in similar part sizes.

This is one potential scenario for using Byte-Range Fetches.

Module Summary and Exam Tips

S3 Batch Operations can be leveraged to perform operations on a large-scale set of S3 objects at once.

S3 Select and Glacier Select

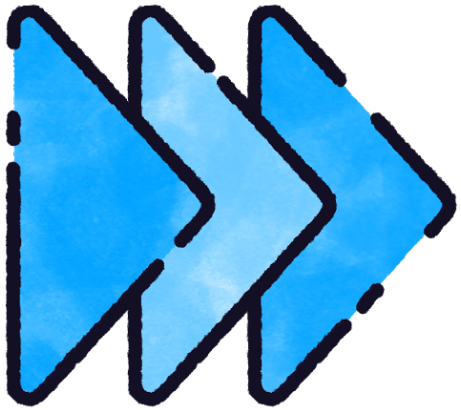
Use SQL statements to perform server-side filtering of S3 objects

Supports CSV, JSON, or Apache Parquet, as well as GZIP and BZIP2 compression

Best used for smaller, simple selections to reduce data retrieval costs and latency

Use S3 Storage Lens to gain organizational-wide insights about your S3 buckets and objects, such as hotspots and data anomalies.

S3 Event Notifications



Trigger workflows based on several events that can occur within your buckets

Good events to remember:

- s3:ObjectCreated:*
- s3:ObjectRemoved:*
- s3:Replication:*
- s3:ObjectTagging:*

Remember the destination types:

- AWS Lambda Functions
- Amazon SQS Queues
- Amazon SNS Topics
- Amazon EventBridge

Amazon S3 Transfer Acceleration can be useful if you need to speed up data requests to and from Amazon S3 buckets around the globe.

Amazon S3 Websites

Static Websites

Easily host static websites within your Amazon S3 buckets, but make sure your bucket policy allows public access

Client-side Scripts

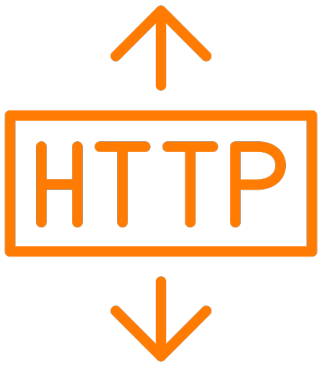
Client-side processing is fine, like JavaScript functions, but there is no server-side processing capability

Amazon S3 can scale automatically to handle variable static website demands!

Amazon S3 Performance

Prefix Performance

Plan your S3 prefixes appropriately so that you can maximize HTTP requests



Byte-Range Fetches

Byte-Range fetches allow for improved download performance and resiliency against failures