
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

STG205

Modernize your data archive with Amazon S3

Gayla Beasley (she/her)
Sr. Technical Program Manager
Amazon S3 Glacier
AWS

Andrew Pohl (he/him)
Principal Product Manager
Amazon S3 Glacier
AWS

Kaushik Lohia (he/him)
Technical Program Manager
Efficiency
Stripe

Agenda

Why archive data with AWS?

Amazon S3 archival storage classes

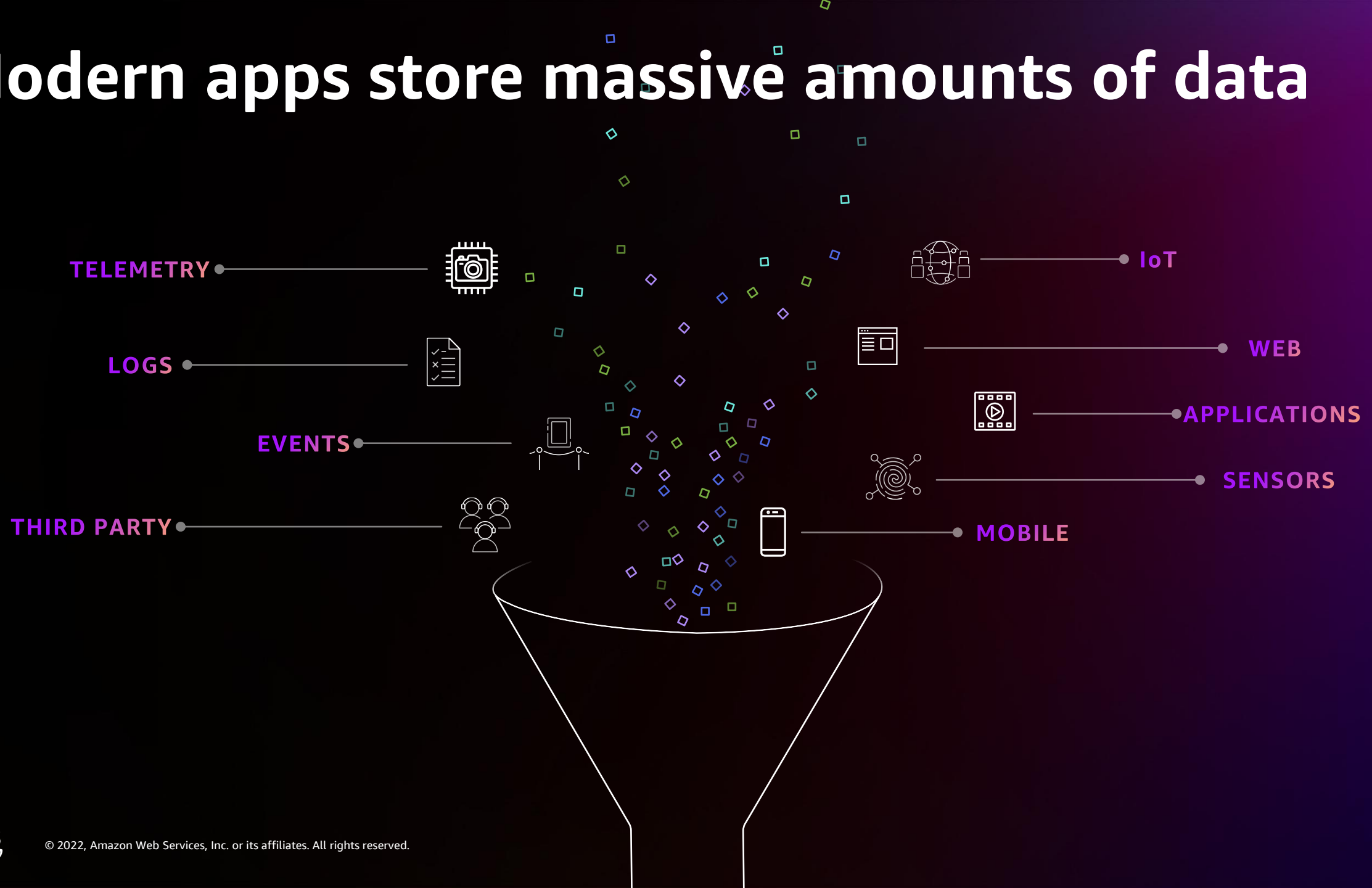
Best practices for retrieving and storing archival data in AWS

Stripe’s journey to archive

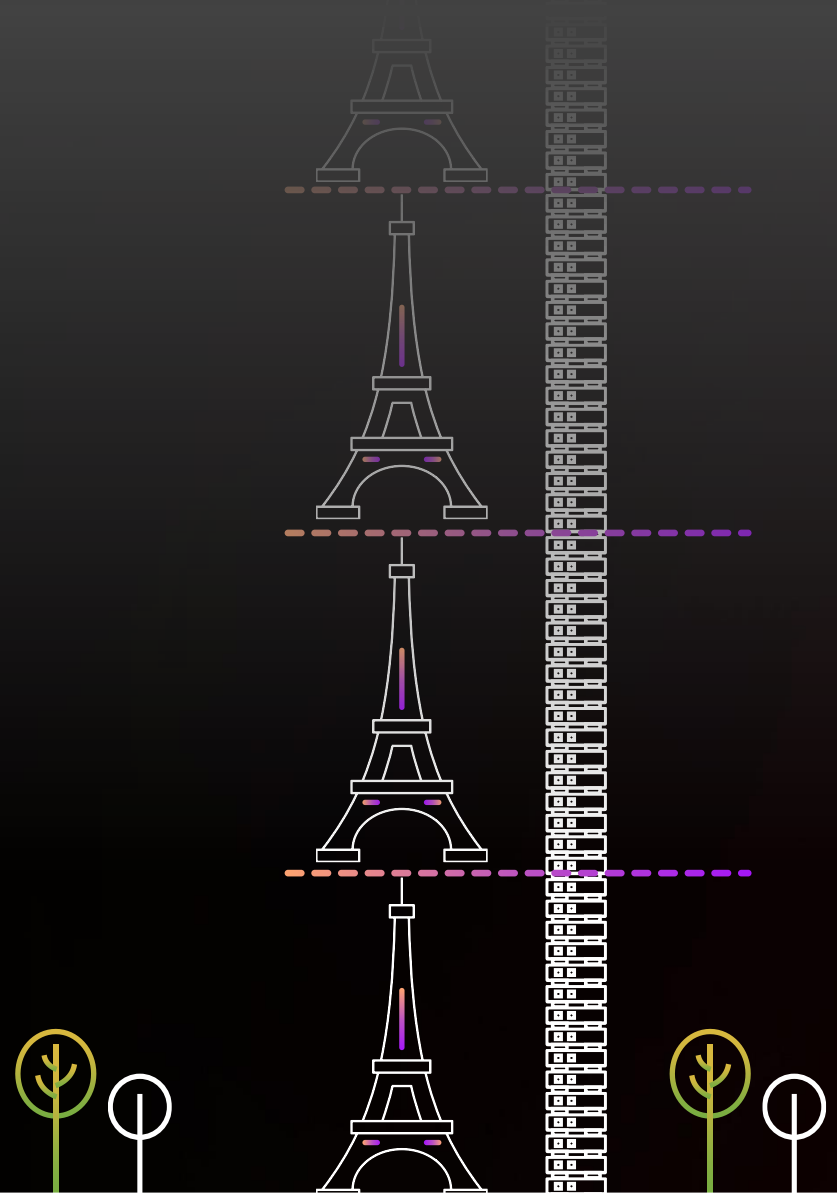
Happy 10th anniversary, Amazon S3 Glacier!

A decade of cold storage and innovation in the cloud

Modern apps store massive amounts of data



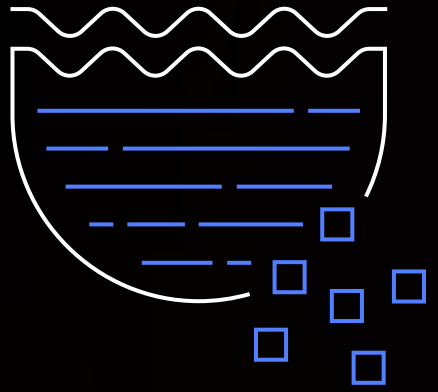
101 ZB
of data created
in 2022



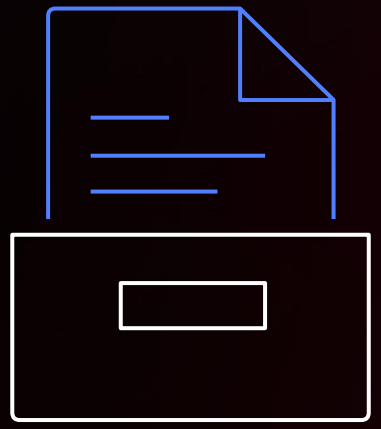
x 18 million

Maximize the value of archival data

Archival use cases on Amazon S3



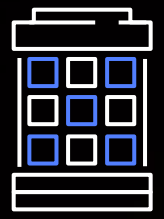
Data retention



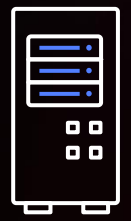
Compliance

Disaster recovery

On-premises archival challenges

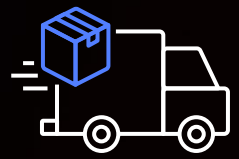


Hosting – space, power, cooling, network



Capital – servers, libraries, tapes

Administration and operations



Off-site storage and transportation

Opportunity cost

Benefits of archiving on AWS

Easier to restore data from archive

Increase value of data

Zero hardware to manage

Lowest cost

Scan for
NASCAR blog

Customers in every industry archive data in AWS

Media & entertainment

Gaming

Healthcare & life sciences

Financial services

Power & utilities

Energy

Manufacturing

Retail

Telecom

Automotive

Education

Government

Amazon S3 storage classes

S3 Glacier Instant Retrieval

What is it?

- For long-lived archive data that requires milliseconds retrieval
- 99.999999999% (11 9s) of durability
- Designed for 99.9% availability

What are the use cases?

- Petabytes of archive data stored for indefinite periods of time
- Only a small percentage of this archive data is accessed each year
- Archive data must be immediately accessible when requested

Amazon S3 Glacier Flexible Retrieval

What is it?

- For long-lived archive data and long-term backup
- 99.999999999% (11 9s) of durability
- Retrievals in 3–5 hours for standard
- Free Bulk retrievals in 5–12 hours

What are the use cases?

- Petabytes of archive data stored for indefinite periods of time
- Data accessed 1–2 times per year

Amazon S3 Glacier Deep Archive

What is it?

- Archiving long-term data that which accessed infrequently
- 99.999999999% (11 9s) of durability
- Retrievals within 12 to 48 hours

What are the use cases?

- Archive data backups that are rarely accessed
- Data that needs to be retained for the long term

Which archive storage class is right for me?

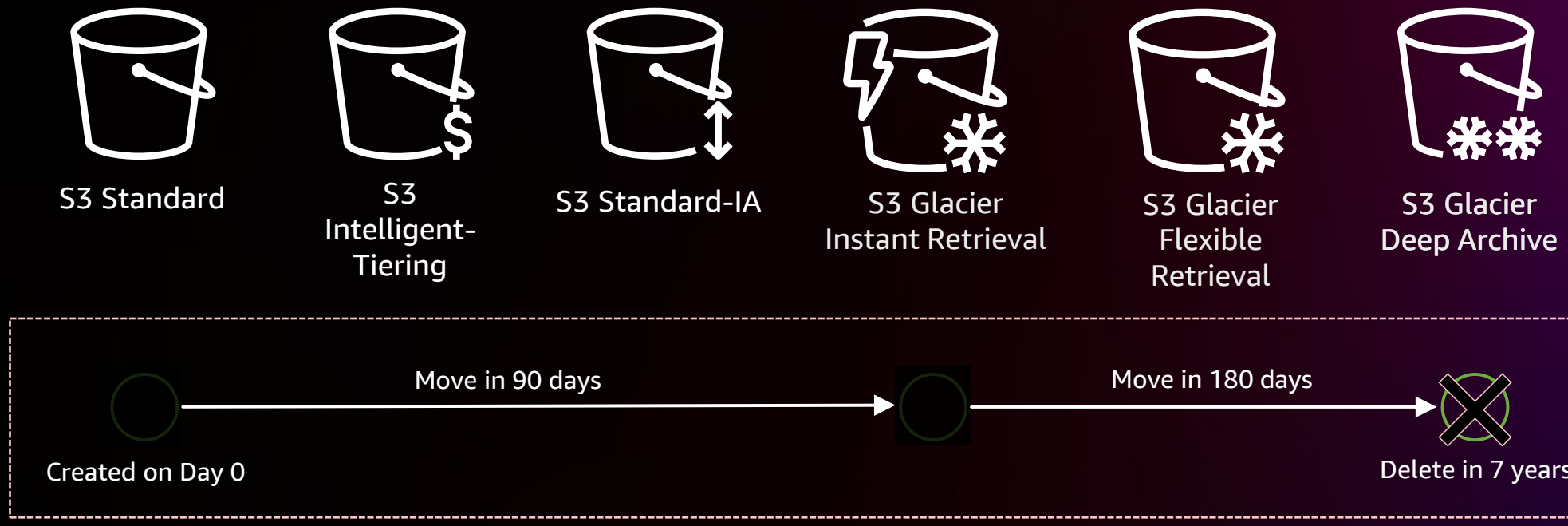
1. Storage cost
2. Retrieval speed
3. Data retention

Choosing between S3 Glacier archive storage

	 S3 Glacier Instant Retrieval	 S3 Glacier Flexible Retrieval	 S3 Glacier Deep Archive
Storage cost	$0.004 per GB-month	$0.0036 per GB-month	$0.00099 per GB-month
Data retrieval	Milliseconds with GET API call	Expedited: 1-5 minutes Standard: 3-5 hours FREE Bulk: 5-12 hours	Standard: Within 12 hours Bulk: Within 48 hours
Minimum object duration	90 Days	90 days	180 days

Tier data to optimize storage costs

Tier data to optimize storage costs



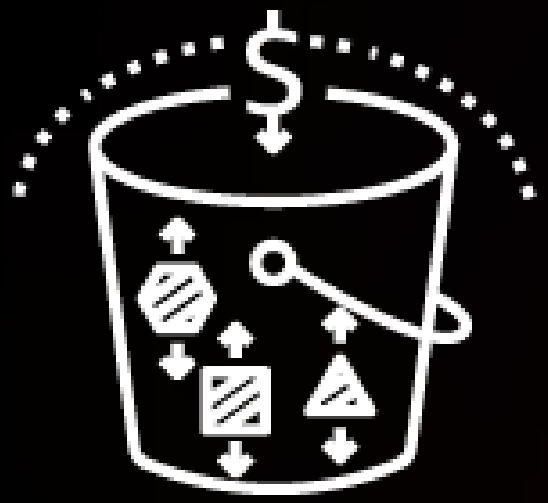
S3 Lifecycle options

Transition and expire data based on . . .

- Bucket
- Prefix
- Object tag
- Object size
- Number of versions

What if my access patterns are unpredictable?

Amazon S3 Intelligent-Tiering

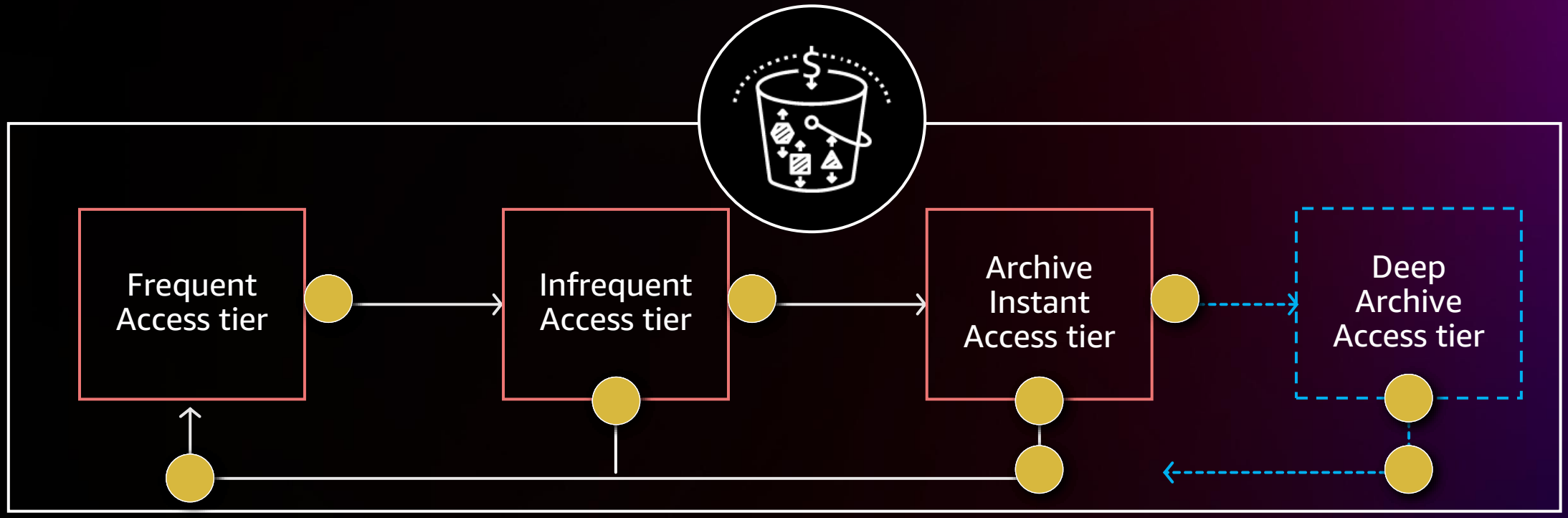


- Automatically moves objects between **three access tiers**
- Enable **Archive and Deep Archive Access tiers** to save even more on archival data
- Optional asynchronous archiving to **realize lowest storage cost in the cloud**
- **No performance impact**, operational overhead, lifecycle fees, or retrieval fees
- Designed for **99.9% availability** and **99.999999999% durability**

$750 million

in savings

Use S3 Intelligent-Tiering by default for data with unknown or changing access patterns



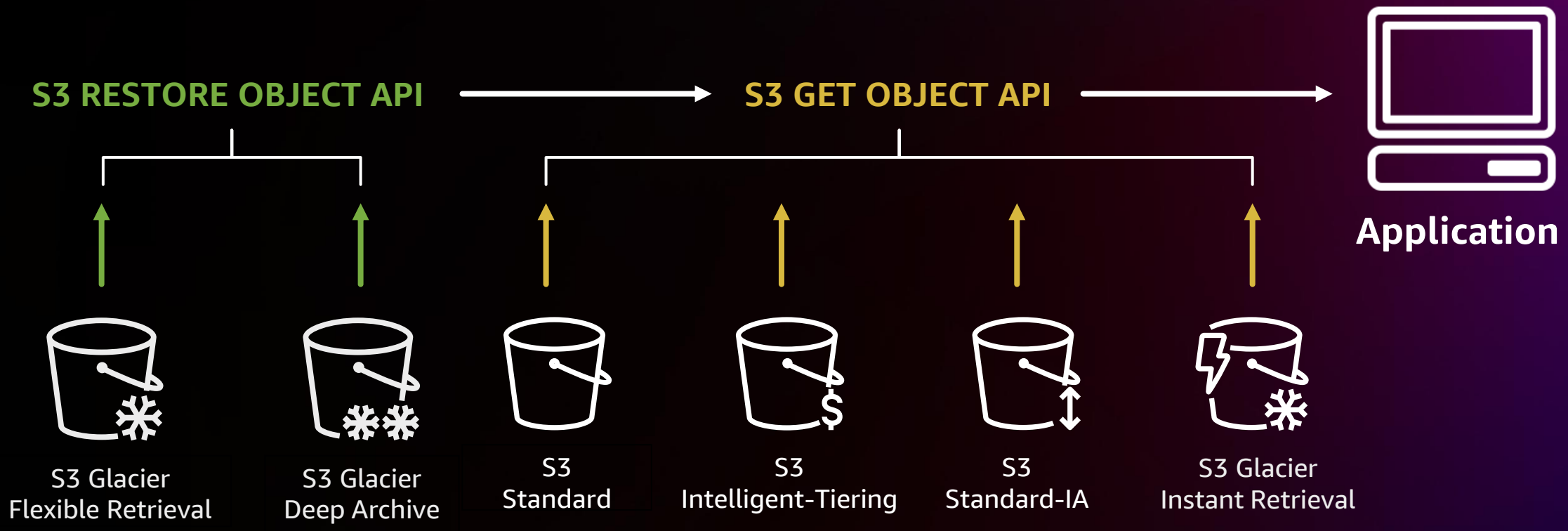
Milliseconds access (automatic)

Minutes to hours (Optional)

Best practices for storing and retrieving archival data in AWS

Over 1 PB
restored every day
from S3 Glacier

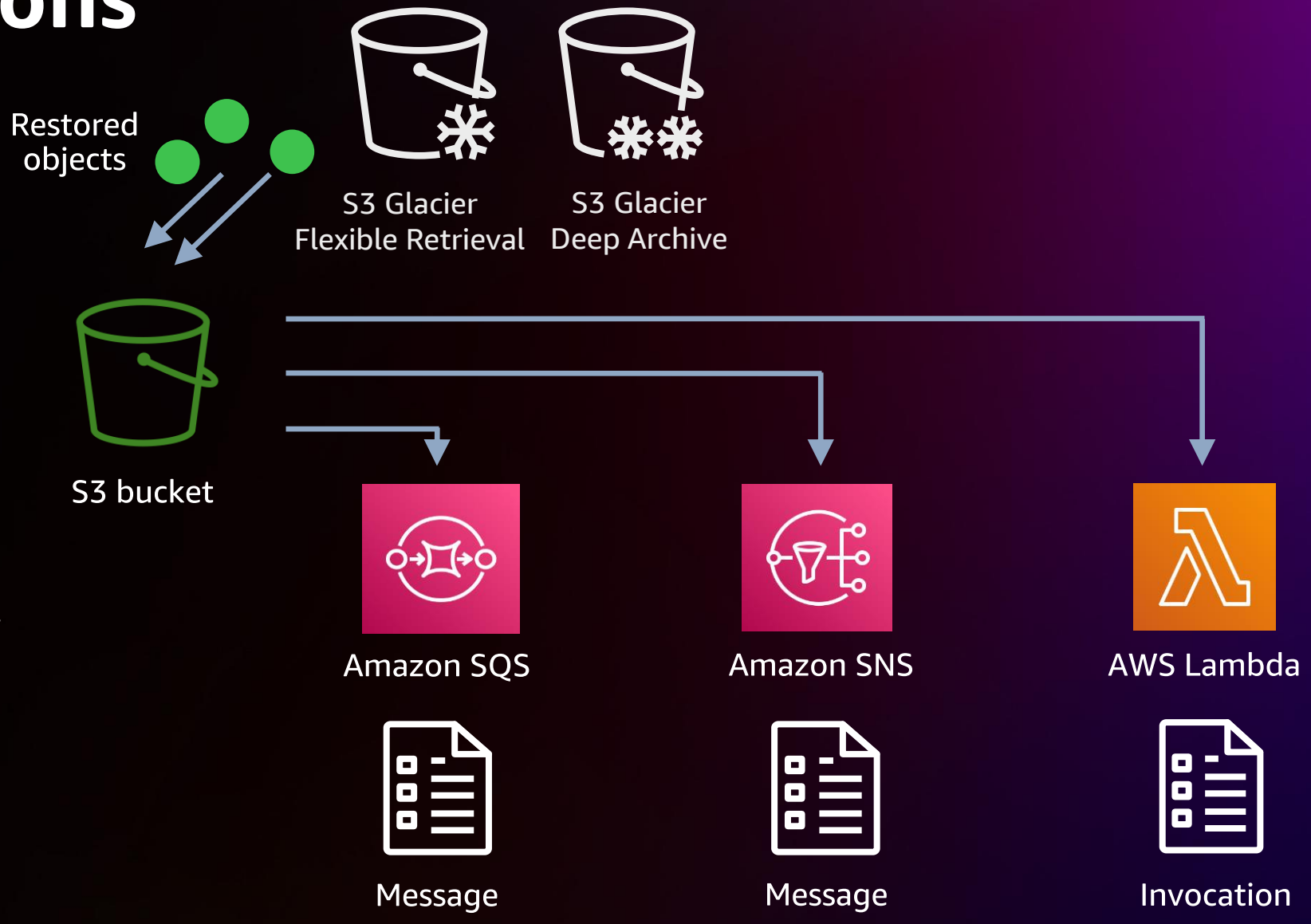
Accessing data from S3



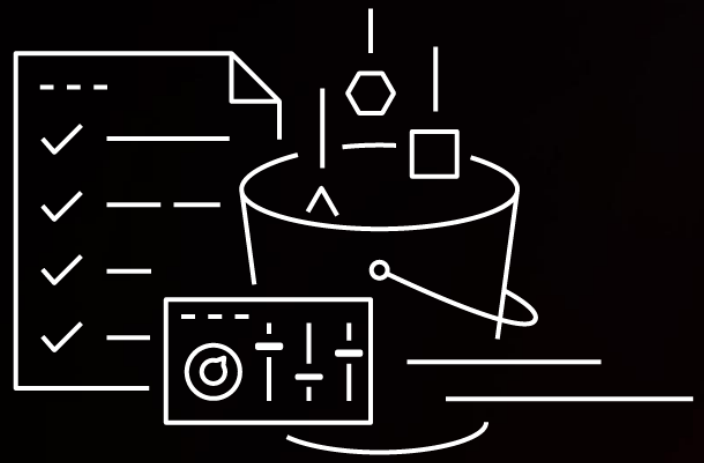
Automate your workflow with Amazon S3 Event Notifications

Reliable & scalable notifications

- At-least-once delivery
- Multiple delivery destinations
- Low latency, no charge
- Notifications on create, delete, restore, and lifecycle actions



Optimize restores with S3 Batch Operations



- Automatic retries
- Management controls
- Notifications
- Auditing

No need to build and maintain an application to call APIs in bulk

Available Now

Amazon S3 Glacier up to 10x restore throughput increase

UP TO 90% FASTER RETRIEVALS FROM S3 GLACIER FLEXIBLE RETRIEVAL, S3 GLACIER DEEP ARCHIVE

Launch blog

Automatically applies to S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive (standard and bulk retrievals) - available at no additional cost

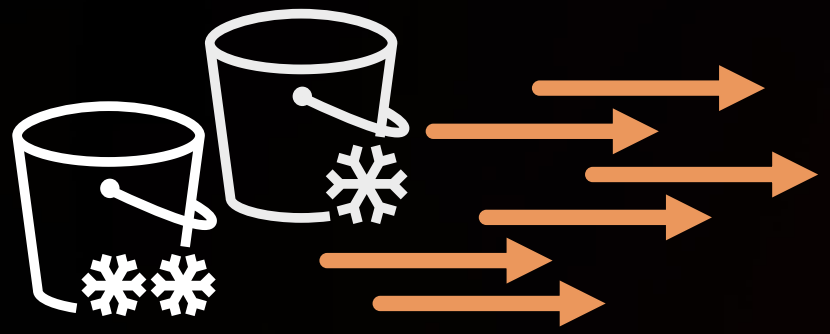
Supports restore requests at a rate of up to 1,000 transactions per second, per account in an AWS Region

Ideal for restoring backups, responding to audit requests, retraining machine learning models, and performing analytics on historical data

Significantly reduces the restore completion time for datasets composed of small objects

aws

Restore object API initiation time at 1,000 TPS



NEW

Use case example

Restore 10 million objects at 10 MB each (~95 TB) to retrain a machine learning model

What is the total restore time from S3 Glacier Flexible Retrieval?

- ~2.8 hours to submit all restore object requests
- Using Standard restore, all objects will typically be restored within 3–5 hours of the last retrieval submission
- Total restore time for this entire workload is typically 5–7 hours

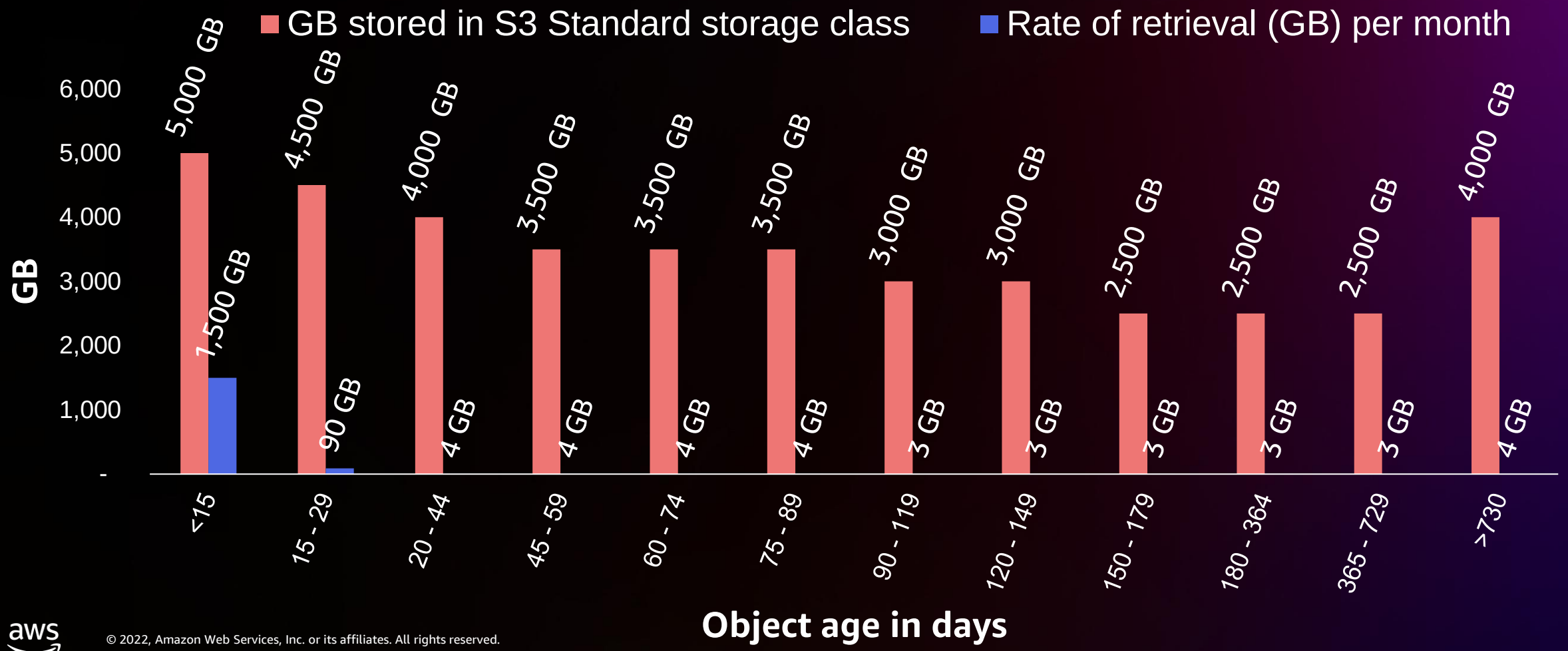
Reducing costs with S3 Glacier storage classes

Lower costs with S3 Glacier storage classes

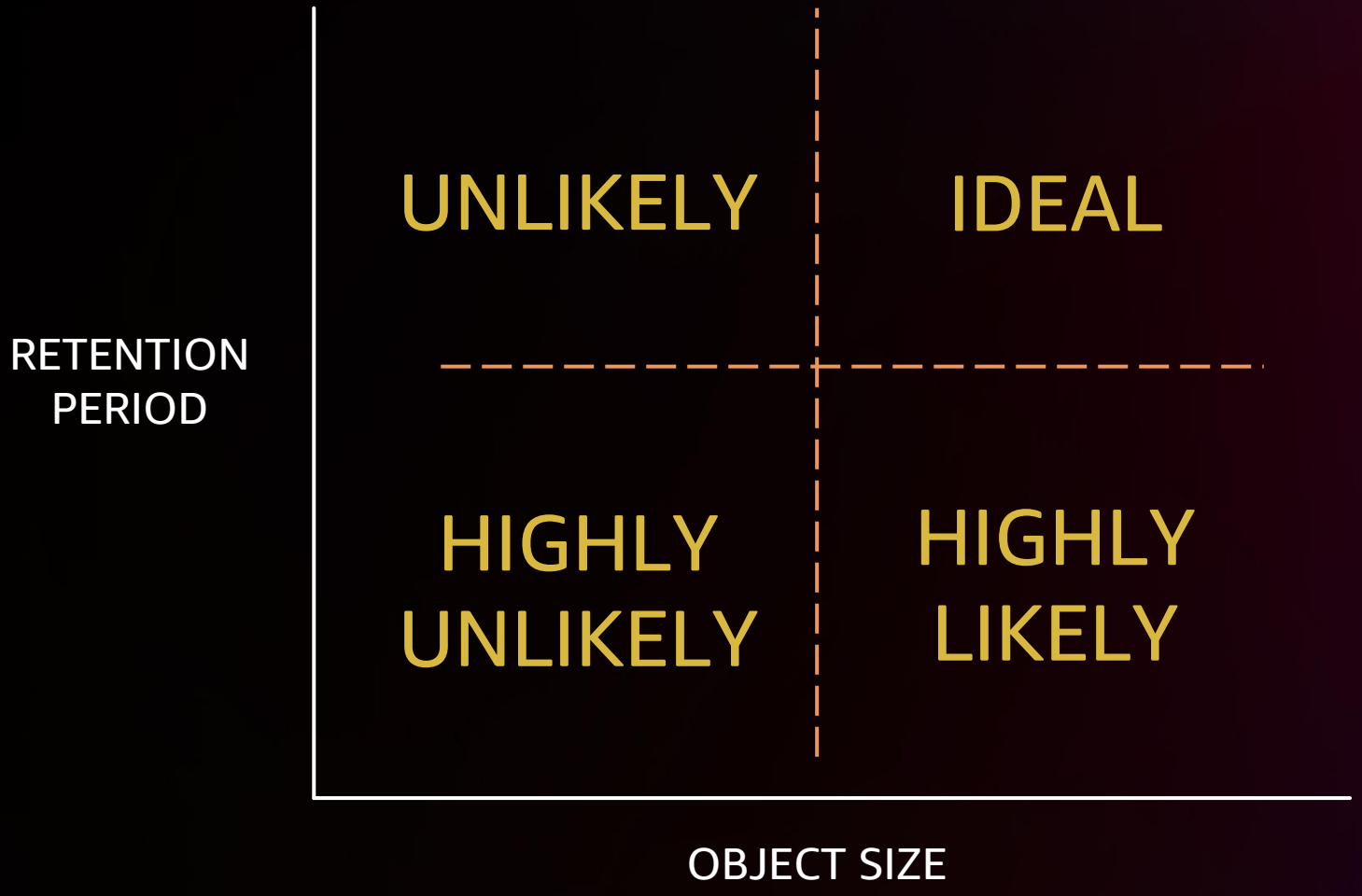
1. Automate savings
2. Retention period
3. Access pattern
4. Object size

Access patterns with S3 storage class analysis

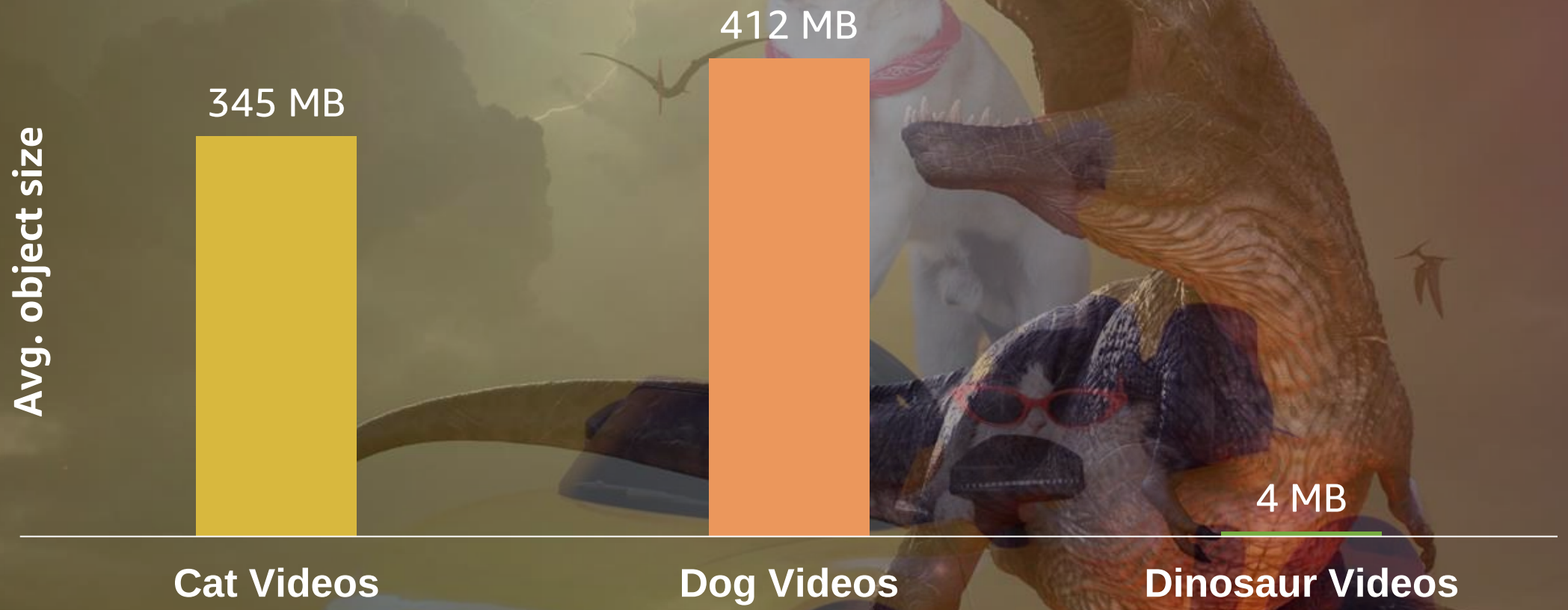
HOW MUCH OF MY S3 STANDARD STORAGE IS ACCESSED ON AVERAGE?



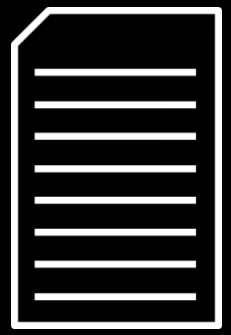
Object size economics – likelihood of saving money with S3 Glacier storage classes



Analyzing object size with S3 Storage Lens



Object size with inventory reports and Amazon Athena



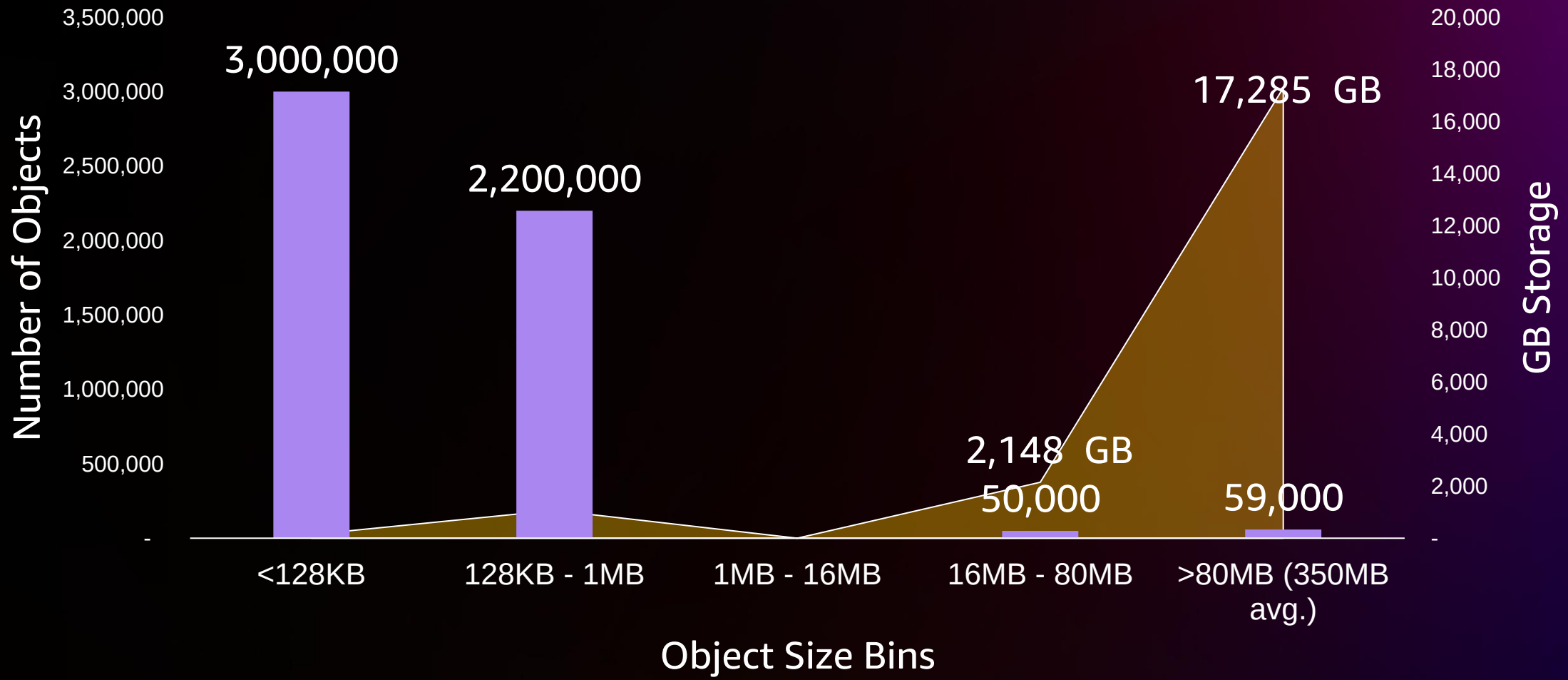
List objects less than 128 KB

```
SELECT key,size
FROM <athena table name>
WHERE dt = '<inventory date>' AND size < 131072;
```

Count number of objects less than 128 KB

```
SELECT COUNT(*) as numbobjects_less_than_128KB
FROM <athena table name>
WHERE dt = '<inventory date>’
AND size < 131072;
```

Object size with inventory reports and Athena



Saving on archival costs with the S3 Lifecycle object size filter

~57% objects <128 KB
<1% of data

3 Million small objects <128 KB

2.2M **medium** at ~500KB

109K large objects >16MB, most data at avg 350MB

Archiving data to the S3 Glacier storage classes using Lifecycle

Pro tip:
Use an object size filter of at least 128 KB to immediately save on storage spend


INDUS
Scan for
Indus OS blog

Stripe's journey to Amazon S3 savings

Kaushik Lohia (he/him)
Technical Program Manager, Efficiency

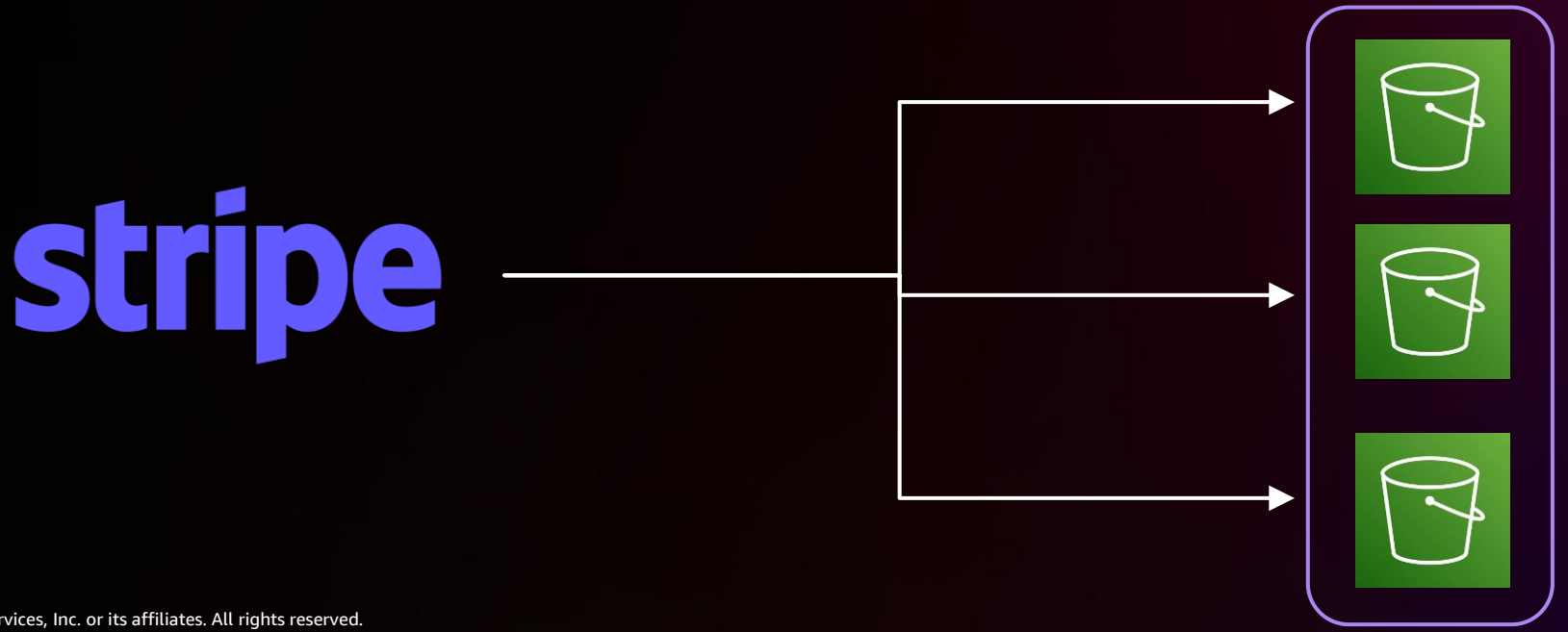
Agenda

- Stripe’s Amazon S3 journey
- More savings via S3 Intelligent-Tiering Archive Access tiers
- Results and retrospective

Stripe's S3 journey

Millions of companies use Stripe

Stripe handles more than **500 million API requests a day**



Stripe's S3 journey

Stripe <> AWS

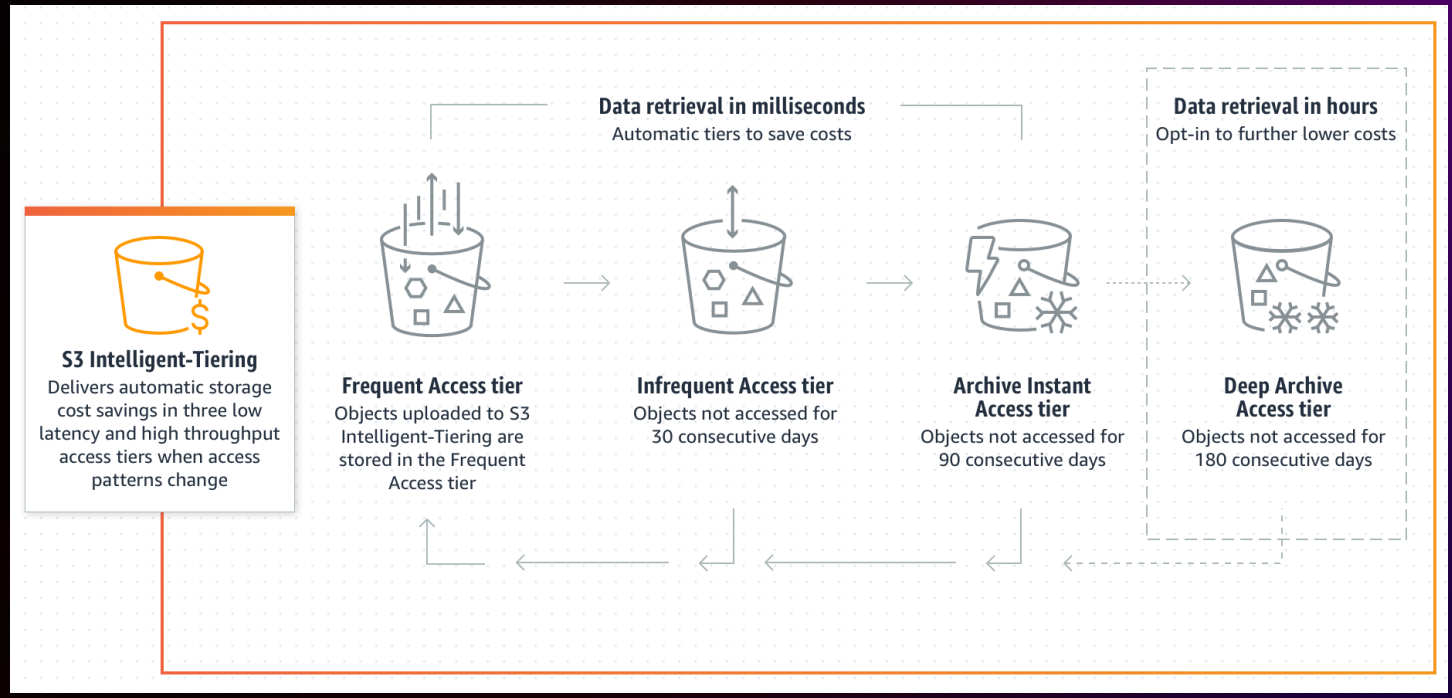
2018: S3 Intelligent-Tiering – optimize costs and understand data

2021: S3 Intelligent-Tiering – Archive Instant Access

2022: S3 Intelligent-Tiering – Deep Archive Access

More savings via Deep Archive Access

- Starting to take advantage of archive access tiers for greater savings
- Constraints:
  - **Data candidacy**
    - User experience
    - Compliance
  - **Developer productivity**



More savings via Deep Archive Access

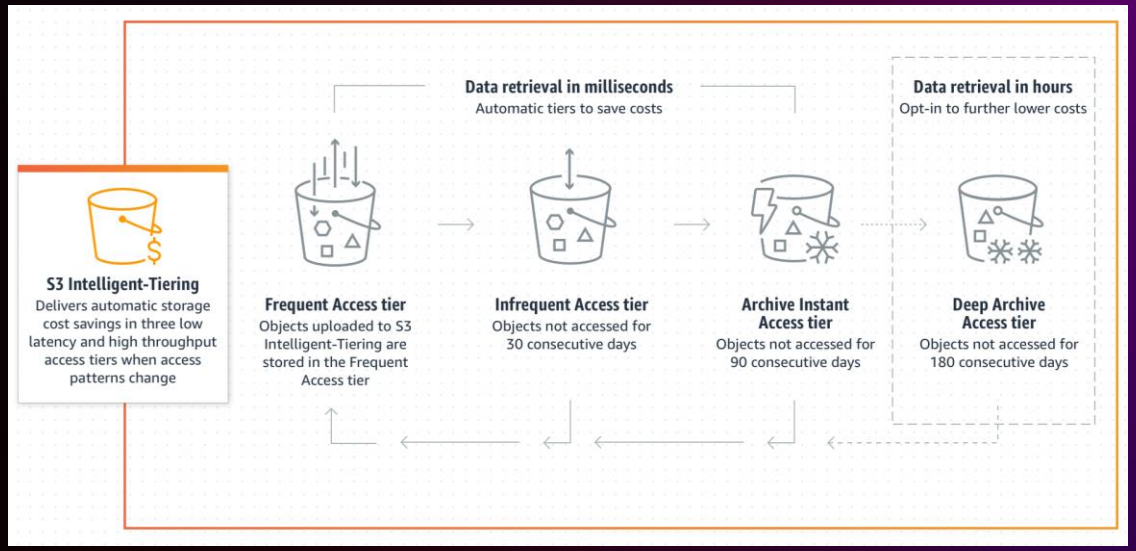
DATA CANDIDACY

- What data can we move across petabytes of data?
- Establish prefix growth rates – with **S3 Storage Lens** and **Usage Reports**!
- Build artifacts and latency data

More savings via Deep Archive Access

DATA CANDIDACY

- Historical latency profiles
- Production software artifacts for replaying transaction data
- SLA: Hours-days, not minutes

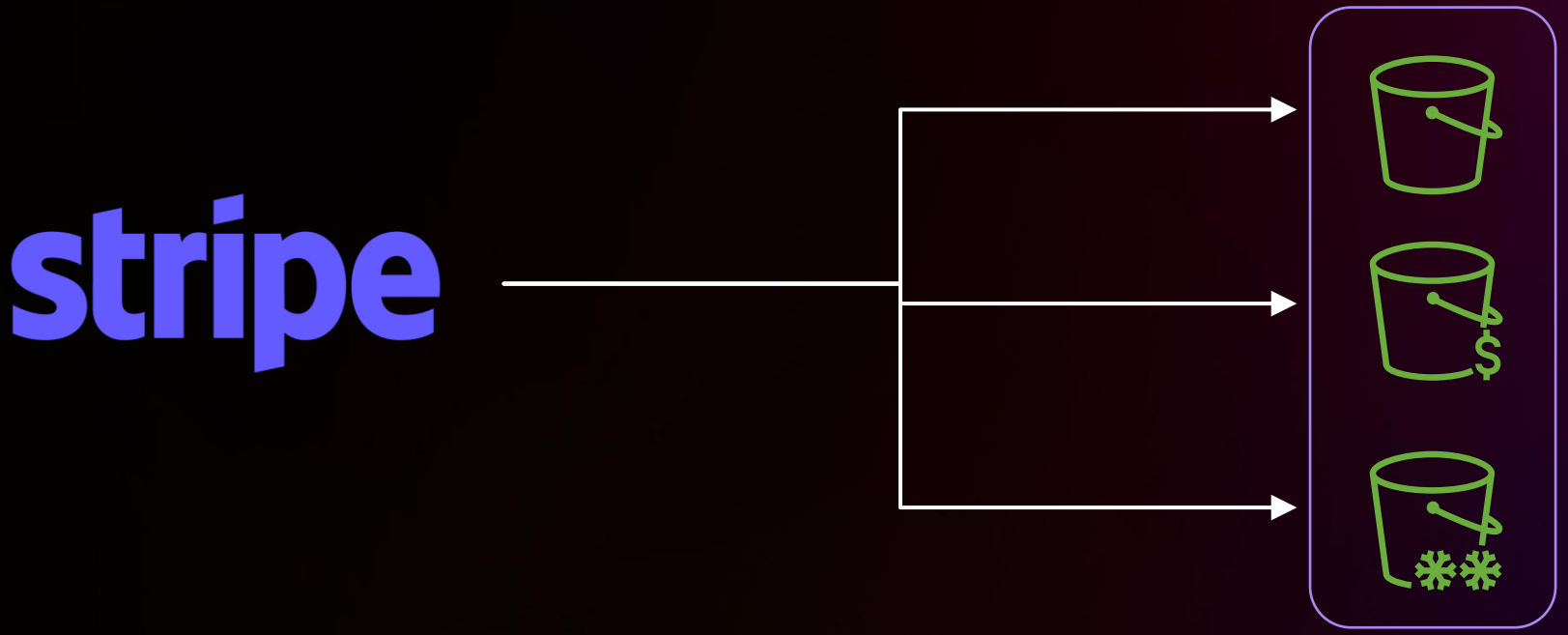


Not accessed in >6 months + Rare use = Deep Archive

More savings via Deep Archive Access

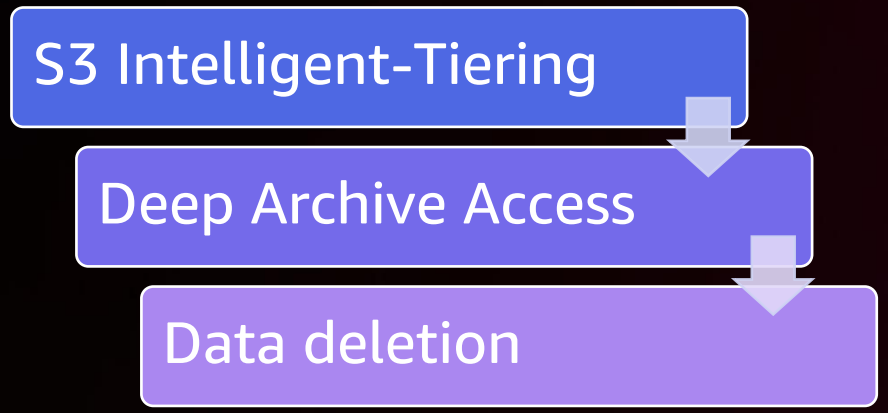
DEVELOPER PRODUCTIVITY

- Avoid re-inventing the wheel for restoring
- Keep it simple – prefix or list of objects
- Maintain trust



Results and retrospective

- 10% of Stripe's data now in Deep Archive Access!
- Large audit restore in 2022 – 1-line Ruby command to execute on time
- Using S3 Intelligent-Tiering and Deep Archive Access to understand data access



Retrospective

Data storage is cheap, until it isn't

Revisit storage assumptions

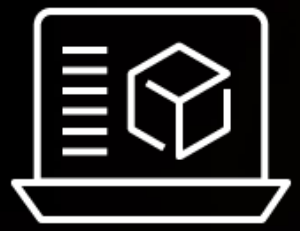
S3 Intelligent-Tiering and Deep Archive Access established a way for Stripe to meet compliance needs ***without holding onto unnecessary data***

Tutorial:

Getting started using the Amazon S3 Glacier storage classes

Continue your AWS Storage learning

Build a learning plan



Set your AWS Storage Learning Plans via **AWS Skill Builder**

Increase your knowledge

Use our **Ramp-Up Guides** to build your storage knowledge

Earn AWS Storage badges

Demonstrate your knowledge by achieving **digital badges**

aws.training/storage

Thank you!

Gayla Beasley
gybsl@amazon.com

Andrew Pohl
andrepoh@amazon.com

Kaushik Lohia
LinkedIn: kaushik1111

Please complete the session survey in the **mobile app**