AWS Data, Databases, and Analytics Online Series

How to ingest data seamlessly to build your data lake

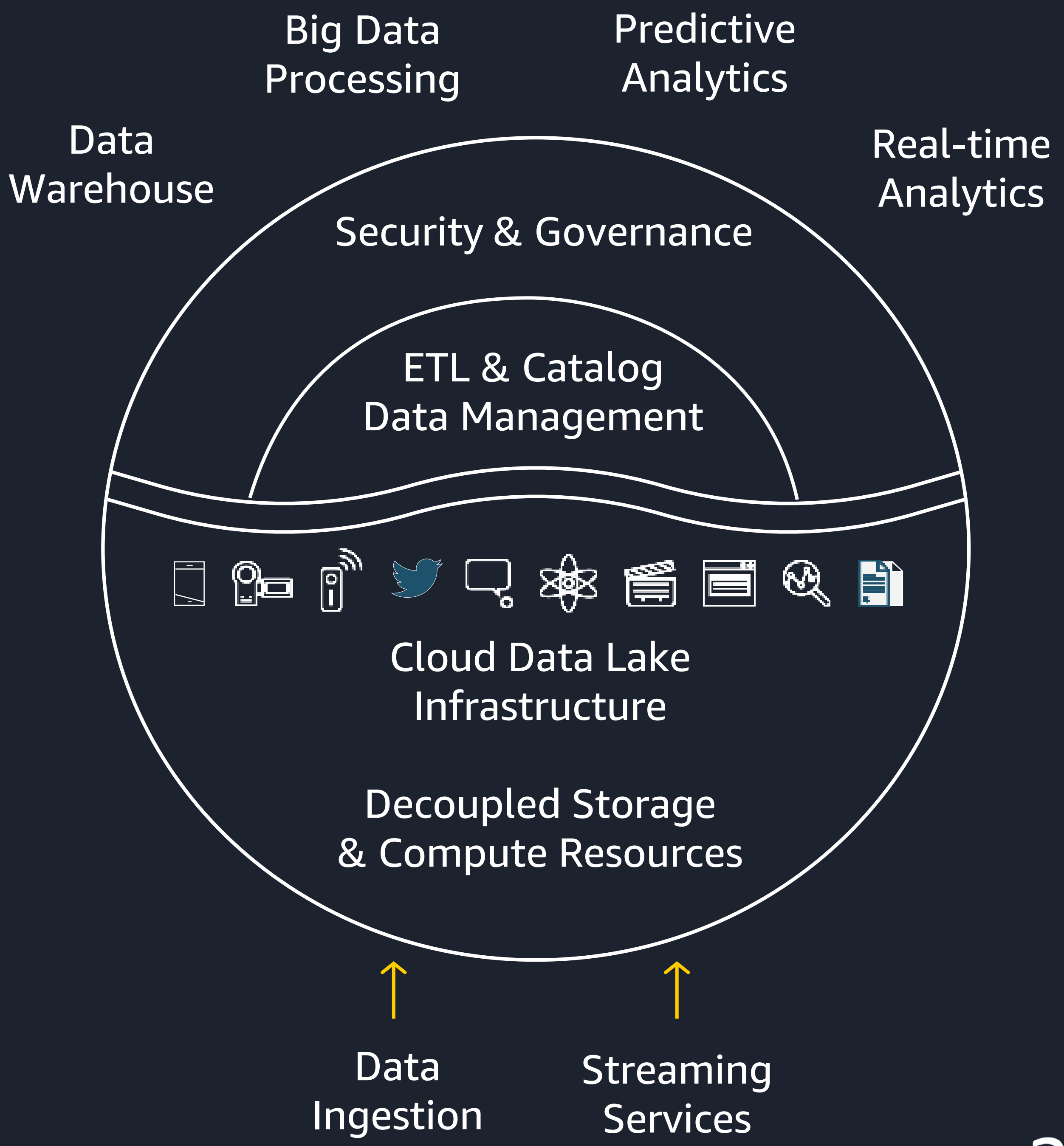
Wali Akbari

Storage Specialist Solutions Architect, AWS

What is a data lake?

A **data lake** is a centralized repository that allows you to ingest, store, and manage structured and unstructured data at unlimited scale.

Then gain insights through analytics and machine learning.



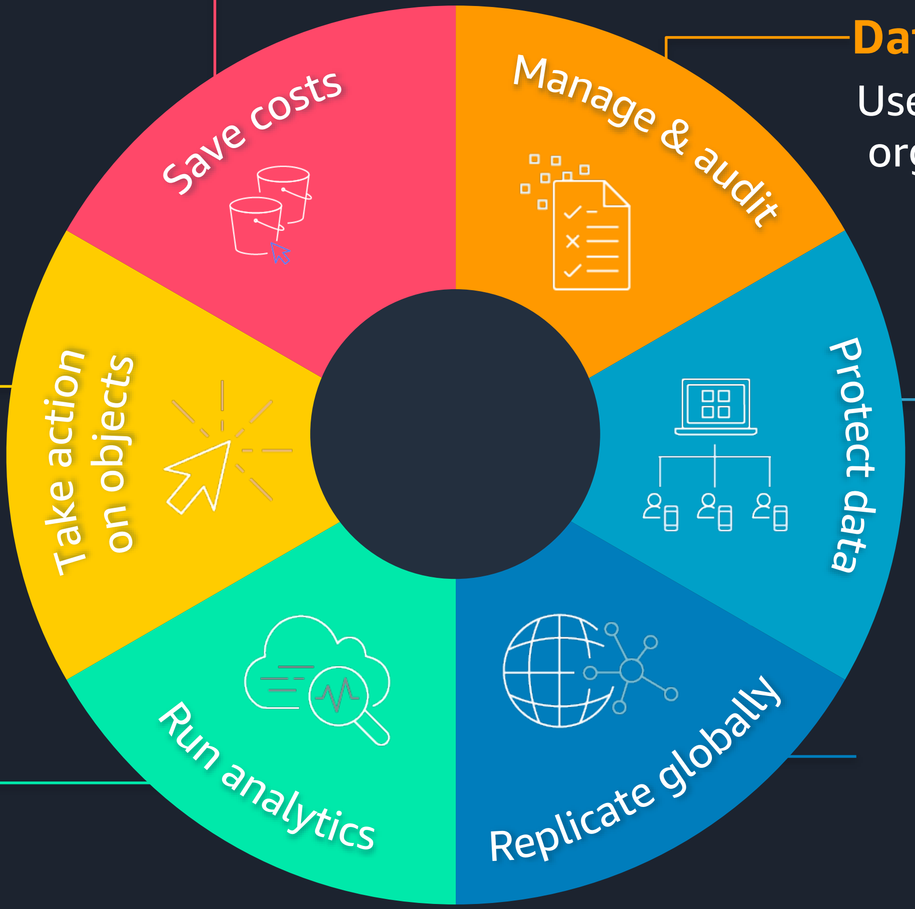
Amazon Simple Storage Service (Amazon S3) - Overview

S3 Batch Operations

Take actions on objects at scale

Analytics & file system integration

S3-integrated analytics applications
AWS Lake Formation to stand up a data lake in days
S3 Select to query data in place
FSx for Lustre for HPC, ML, and media data processing



Data management tools

Use tags, buckets, and prefixes to organize data.

Access management

Configure access to S3 resources.
Block all public access S3 Block Public Access.

Cross-Region Replication

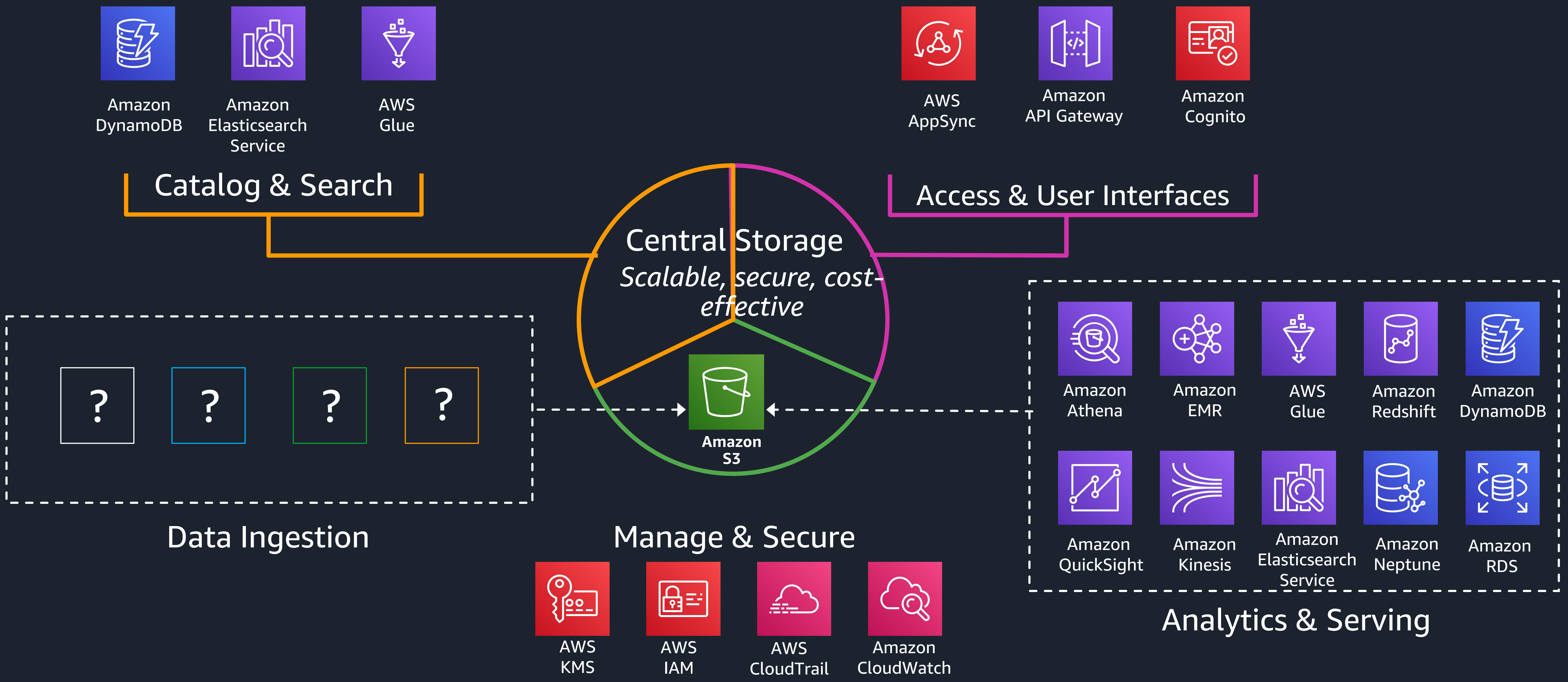
Replicate objects to a different region of your choice.

S3 Storage Classes

- S3 Standard
- S3 Standard-IA
- S3 Intelligent-Tiering
- S3 One Zone-IA
- S3 Glacier
- S3 Glacier Deep Archive

Security by design	Compliance programs	11 9’s of durability	Multi-AZ resiliency	Limitless scalability

Building a data lake on AWS



Amazon DynamoDB

Amazon Elasticsearch Service

AWS Glue

Catalog & Search

AWS AppSync

Amazon API Gateway

Amazon Cognito

Access & User Interfaces

Central Storage

Scalable, secure, cost-effective

Amazon S3

?

?

?

?

Data Ingestion

Manage & Secure

AWS KMS

AWS IAM

AWS CloudTrail

Amazon CloudWatch

Amazon Athena

Amazon EMR

AWS Glue

Amazon Redshift

Amazon DynamoDB

Amazon QuickSight

Amazon Kinesis

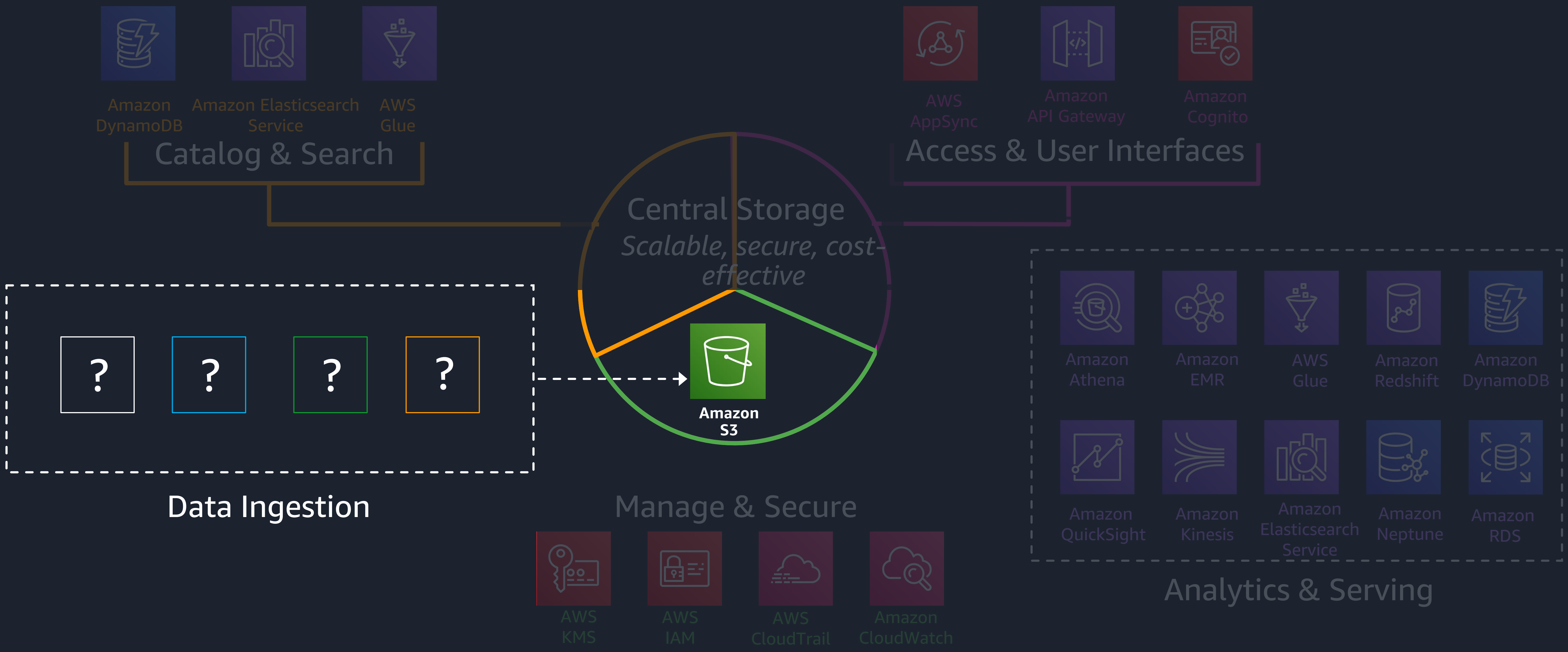
Amazon Elasticsearch Service

Amazon Neptune

Amazon RDS

Analytics & Serving

Building a data lake on AWS – ingesting data



Data ingest challenges

What are some challenges customers face when trying to ingest data

Options

Speed

Time and effort

AWS Storage services portfolio

Block storage

- Amazon EBS

File storage

- Amazon EFS
- Amazon FSx for Windows File Server
- Amazon FSx for Lustre

Object storage

- Amazon S3
- Amazon S3 Glacier

Hybrid

- AWS Storage Gateway

Transport & edge

- AWS DataSync
- AWS Transfer Family
- AWS Snow* Family

Backup

- AWS Backup

aws

AWS Storage services portfolio

Block storage

Amazon EBS

File storage

Amazon EFS

Amazon FSx for Windows File Server

Amazon FSx for Lustre

Object storage

Amazon S3

Amazon S3 Glacier

Hybrid

AWS Storage Gateway

Transport & edge

AWS DataSync

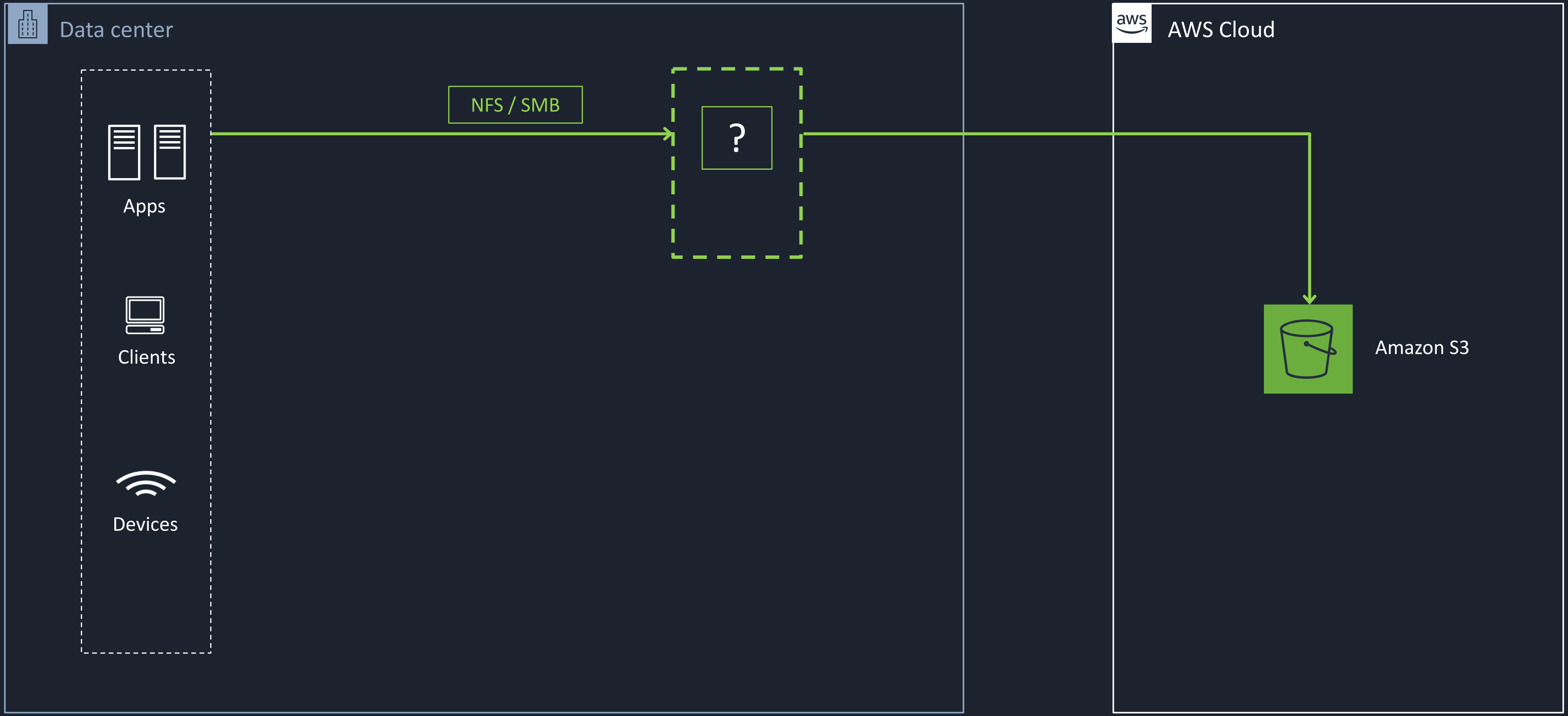
AWS Transfer Family

AWS Snow* Family

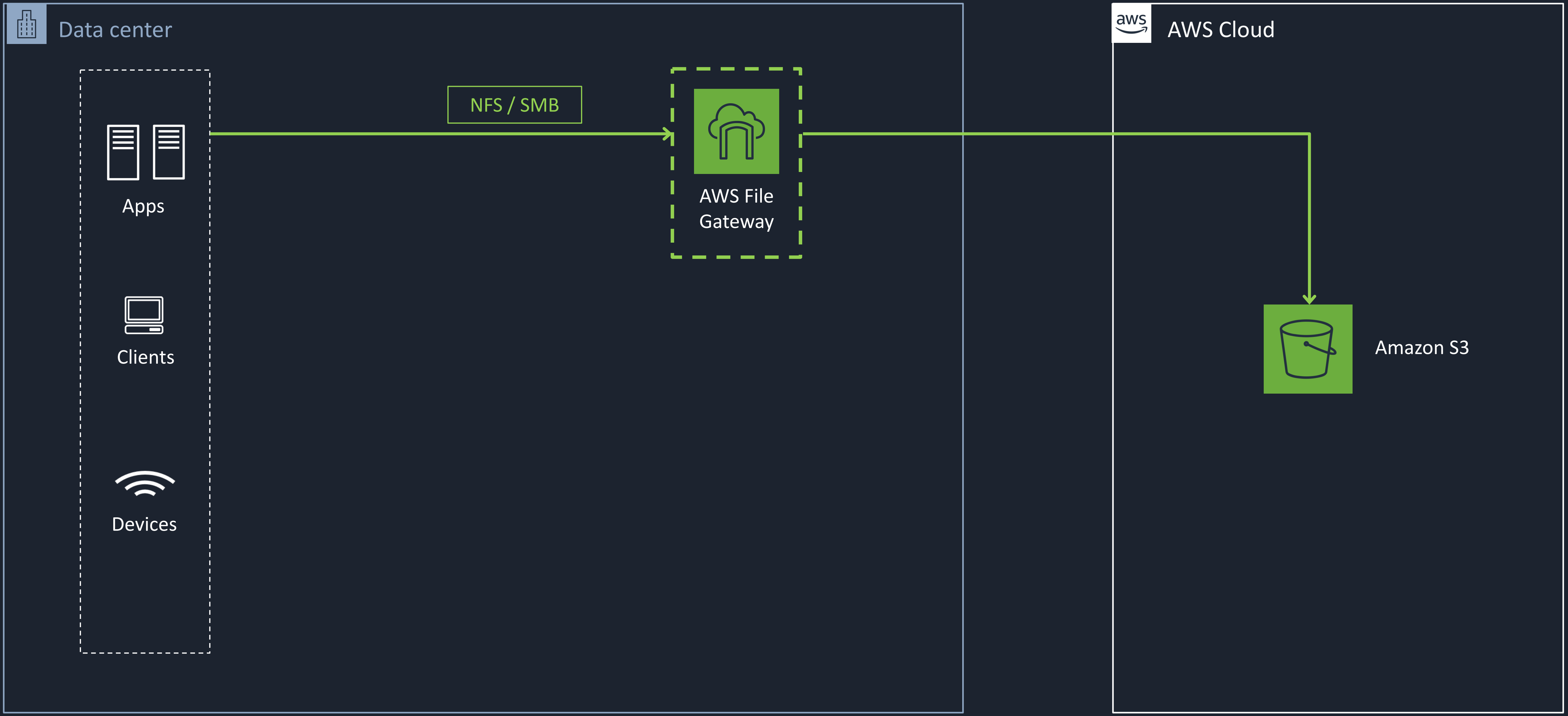
Backup

AWS Backup

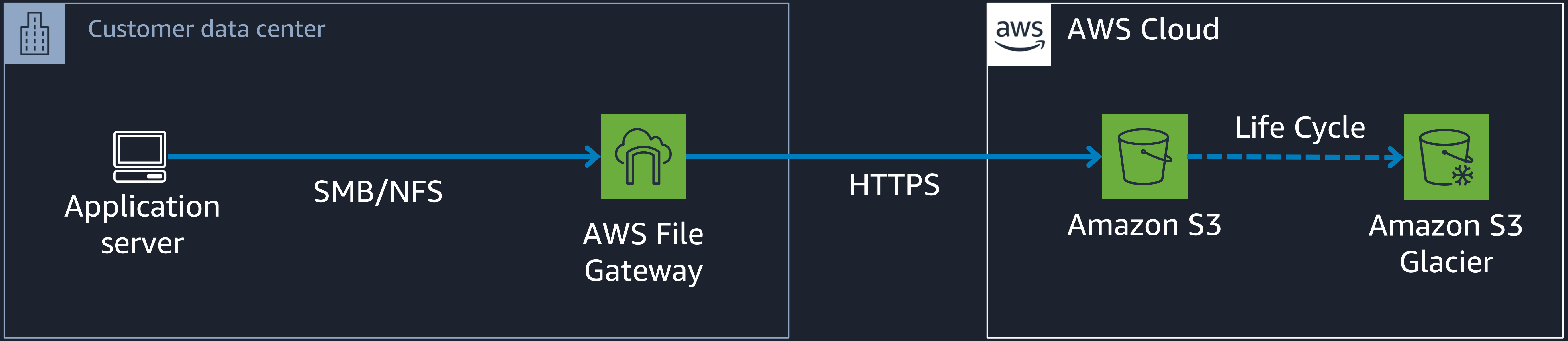
Data Sources – Online ingest



Data Sources – Online ingest



What is AWS File Gateway?



- A virtual or hardware appliance that utilizes a network file system to interface to Amazon S3
- It allows for low latency access to hot data via it's local cache
- Stores file data in its native format as objects in an Amazon S3 bucket
- AWS File Gateway SMB shares can integrate with Microsoft Active Directory

Using AWS File Gateway - Populate your data lake

- Seamless upload of created data to Amazon S3 using a file share
- Ingest existing data into Amazon S3 via a file share

Sharing your data lake contents

- Re-share existing Amazon S3 data via simple file shares

“A big challenge that we face is the integration of the wet labs with the computational aspect of our research.

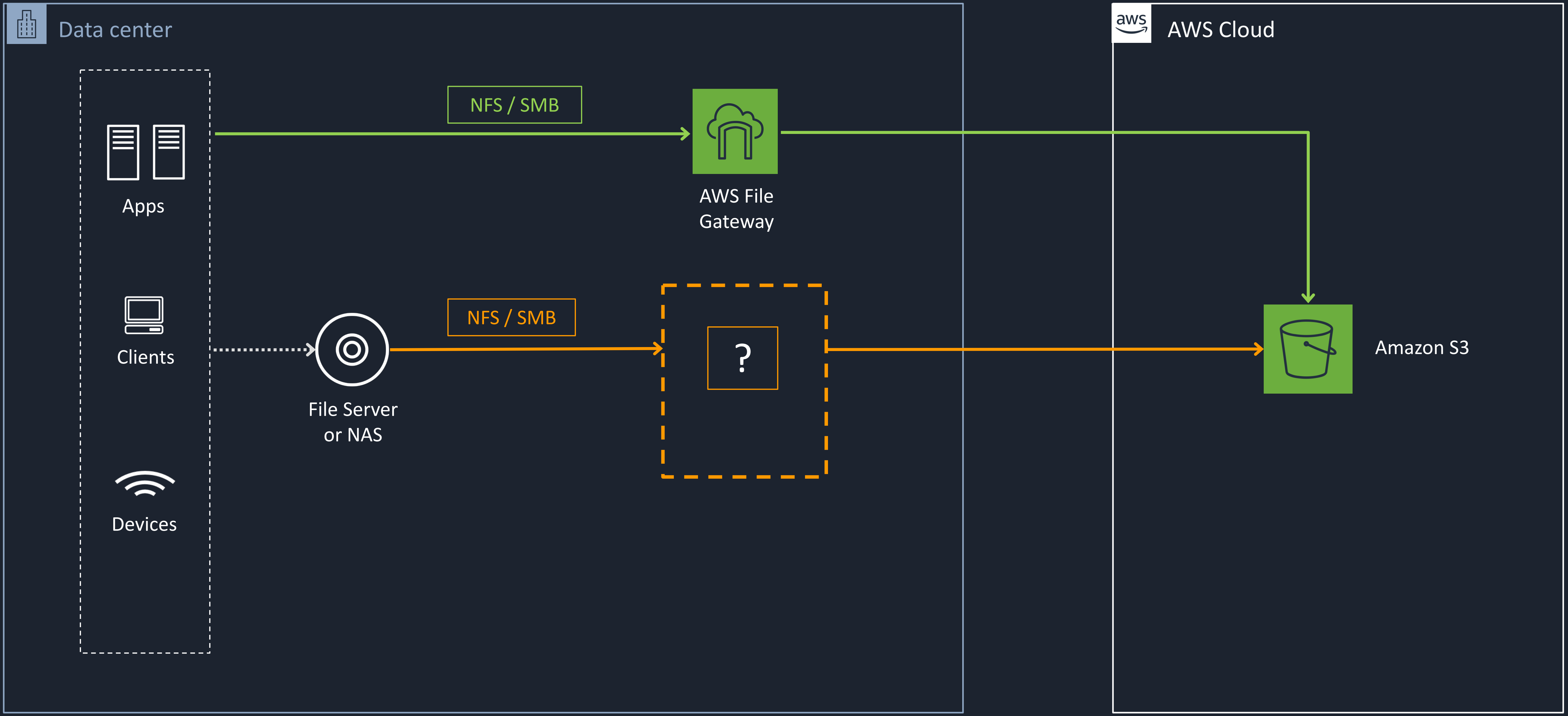
We use Storage Gateway and DataSync to synchronize our on-premise wet labs with our Amazon storage. By having our scientists **immediately** be able to save their files **directly to the cloud**, they can go on with the next experiment without having to wait for the transfer times and they never run out of space."

Lance Smith

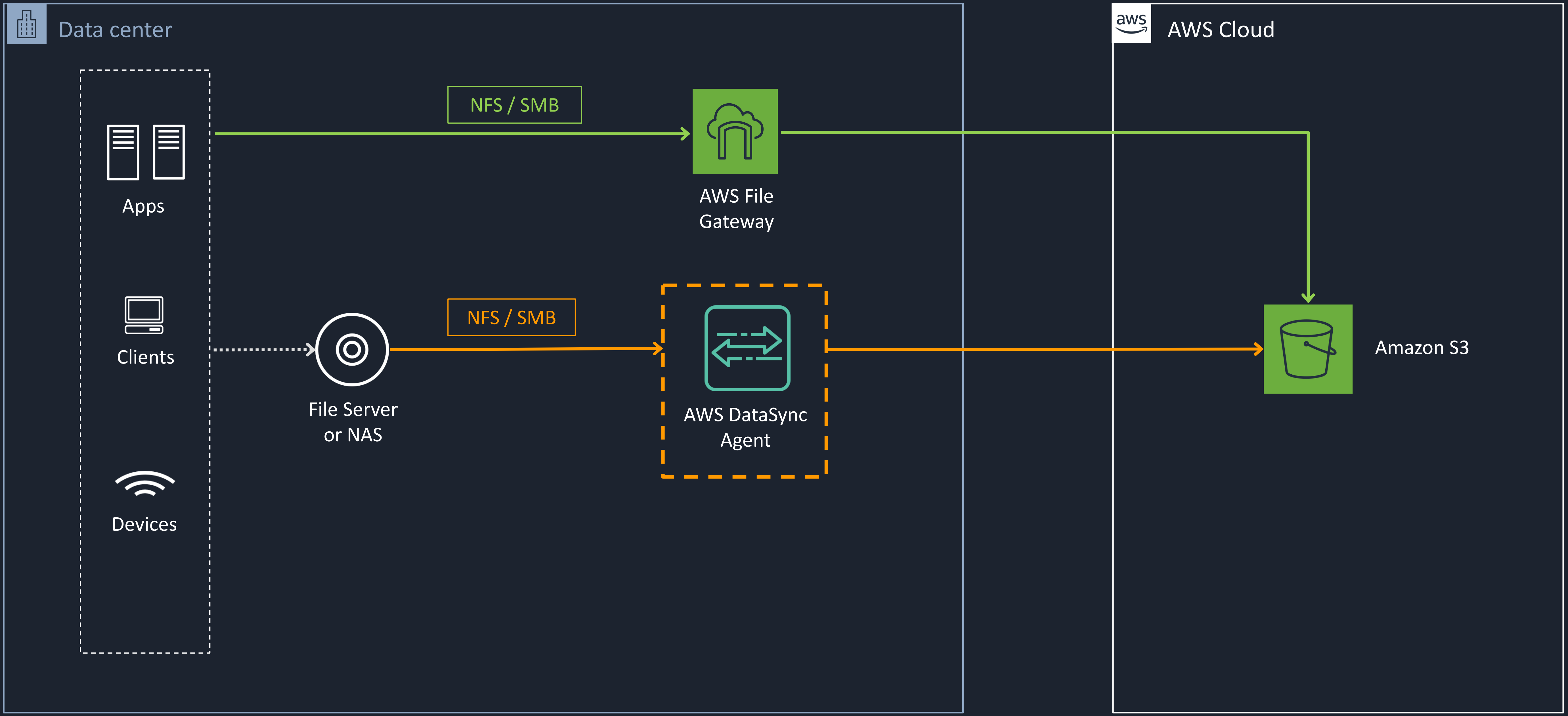
Associate Director - Celgene

https://aws.amazon.com/storagegateway/customers

Data Sources – Online ingest



Data Sources – Online ingest



What is AWS DataSync?

Simplifies, automates, and accelerates your online data transfer

Migrate active application data

Transfer data for timely processing or Archiving

Replicate for data protection and recovery

Transfers up to **10 Gbps** per agent

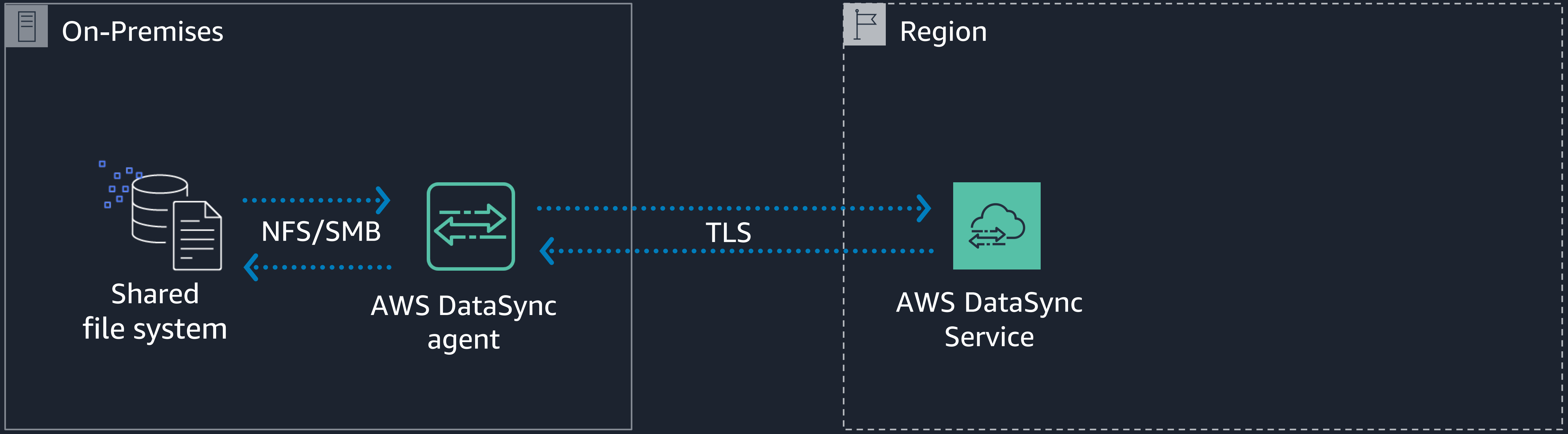
Simple data movement to S3, EFS, FSx for Windows

Secure and reliable transfers

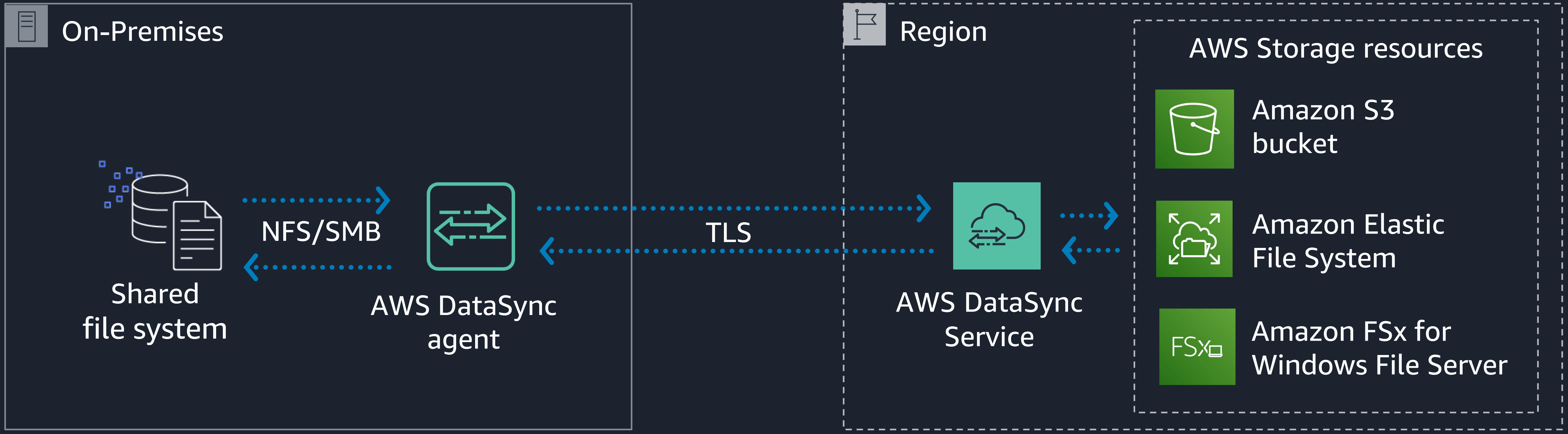
AWS integrated

Pay as you go

How AWS DataSync works



How AWS DataSync works



Highly parallel transfers using optimised network protocol

Fully-managed service scales to send/receive data from agent

Optimised read & write to S3 & EFS using VPC Endpoints

The speed and reliability of *network acceleration* software with the cost-effectiveness of *open source tools*

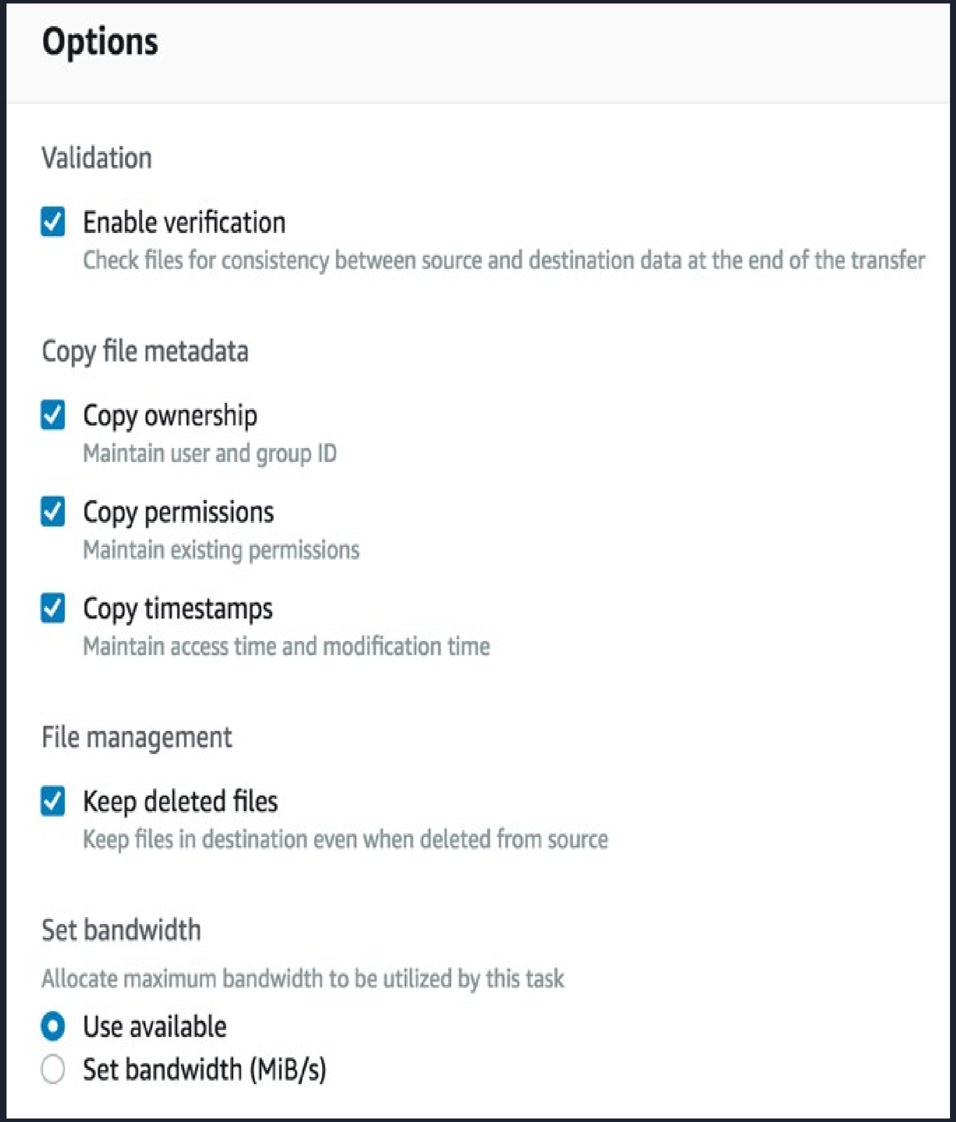
Task options

Invoke via schedule, API, or manually

File-level validation

Copy across file metadata

Throttle bandwidth



Options

Validation

☑ Enable verification
Check files for consistency between source and destination data at the end of the transfer

Copy file metadata

☑ Copy ownership
Maintain user and group ID

☑ Copy permissions
Maintain existing permissions

☑ Copy timestamps
Maintain access time and modification time

File management

☑ Keep deleted files
Keep files in destination even when deleted from source

Set bandwidth

Allocate maximum bandwidth to be utilized by this task

● Use available
○ Set bandwidth (MiB/s)

AWS DataSync usage scenarios

SMB/NFS data transfer → Amazon S3 storage classes

NFS data transfer → Amazon EFS

SMB data transfer → Amazon FSx for Windows File Server

Using AWS DataSync - Populate your data lake

- Simplify & accelerate the transfer of data into Amazon S3
- Utilize for end of event batch or scheduled uploads
- A repeatable data transfer mechanism for different use cases
- Think of its benefits with at-scale data ingest, with simplicity and automation in mind

AUTODESK

"Our petabyte scale data migration journey from on-premises to AWS was accomplished swiftly with minimal effort and was completely self-managed with AWS DataSync. This solution is a game changer!"

Satish Kumar
Infrastructure Engineer

Problem

Wanted to retire multi-petabyte on-premises Data Domain storage system

Data retention policies required data to be retained for many years

Solution

Used Amazon S3 for low cost, pay-as-you-go model as well as versioning support

Used AWS DataSync to seamlessly move data to Amazon S3

Outcome

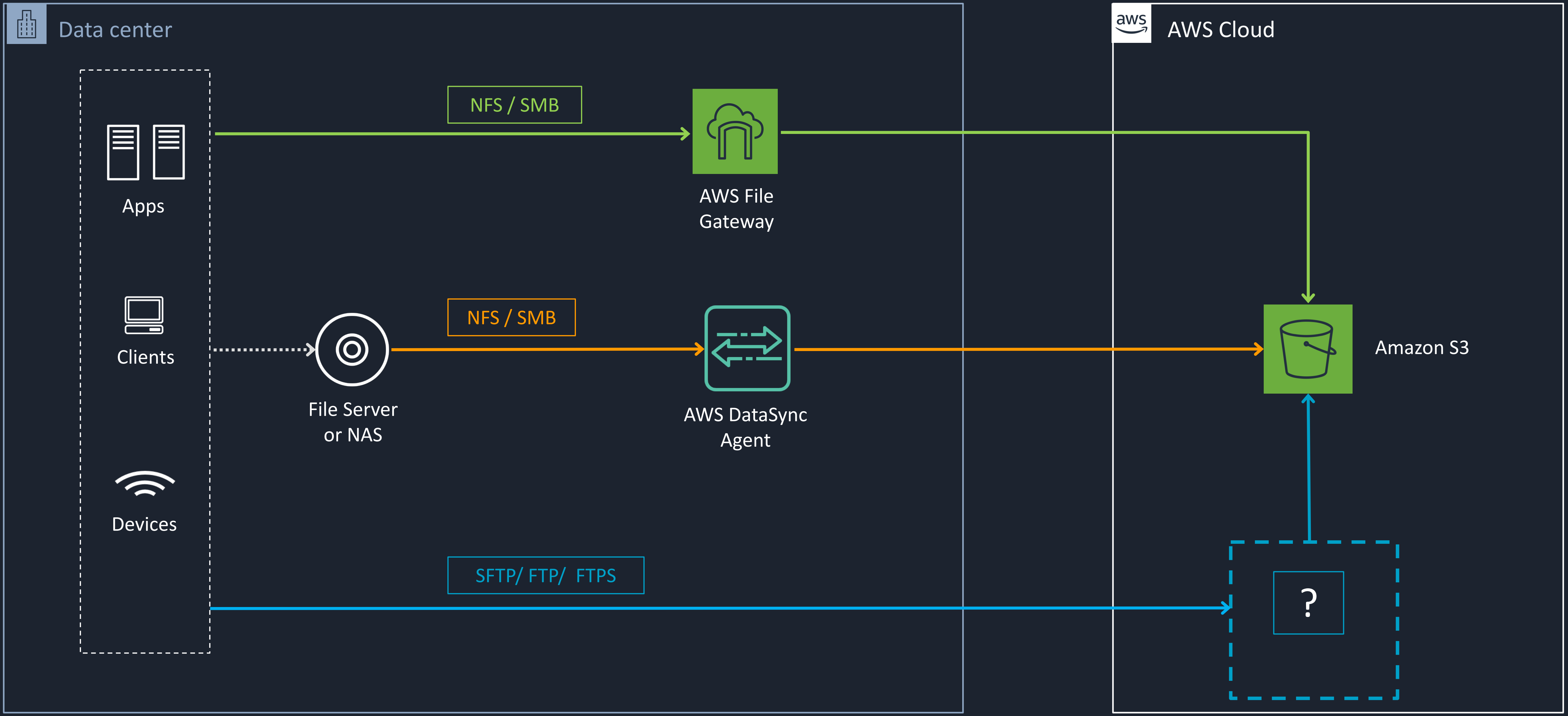
Successfully transferred dataset to Amazon S3 with full byte-for-byte verification

Decommissioned on-premises Data Domain

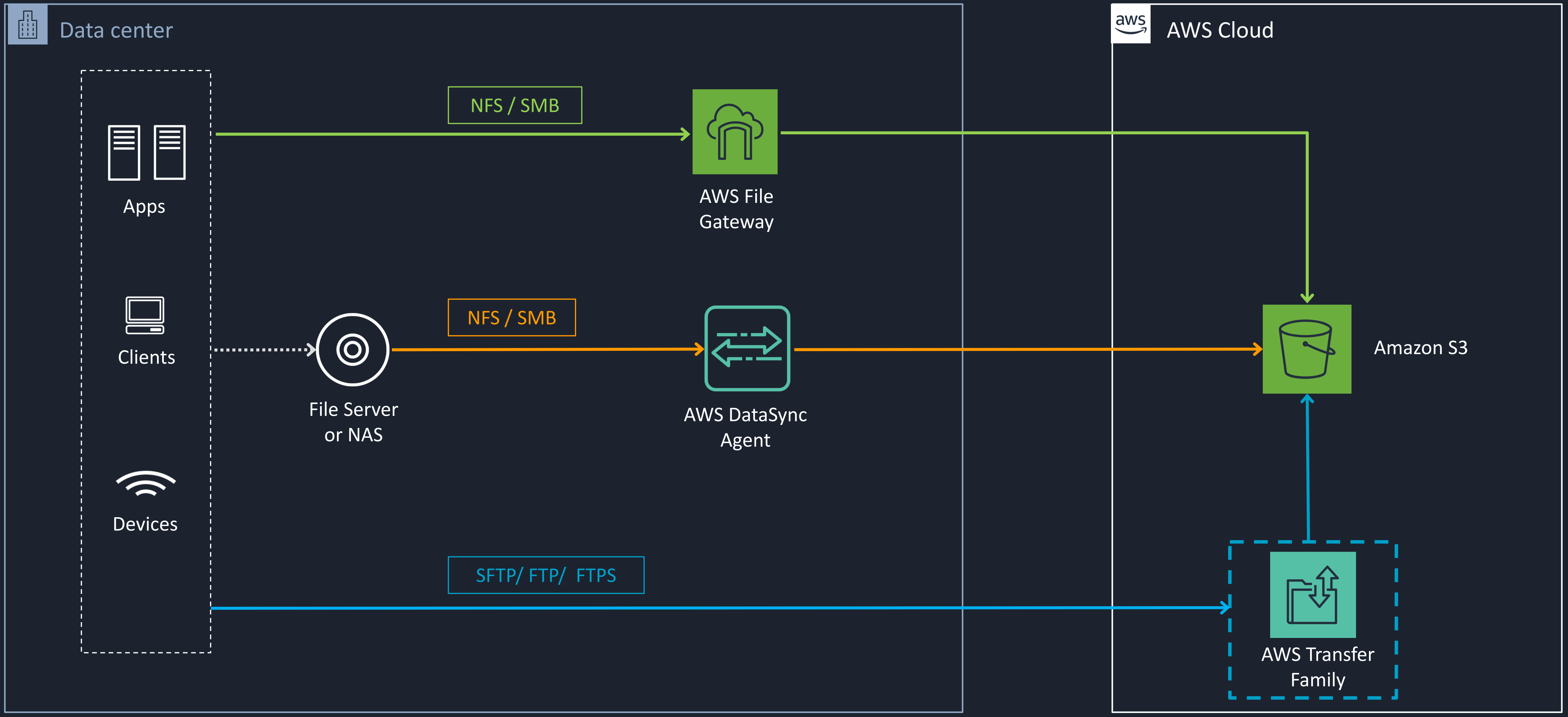
https://aws.amazon.com/datasync/customers

aws

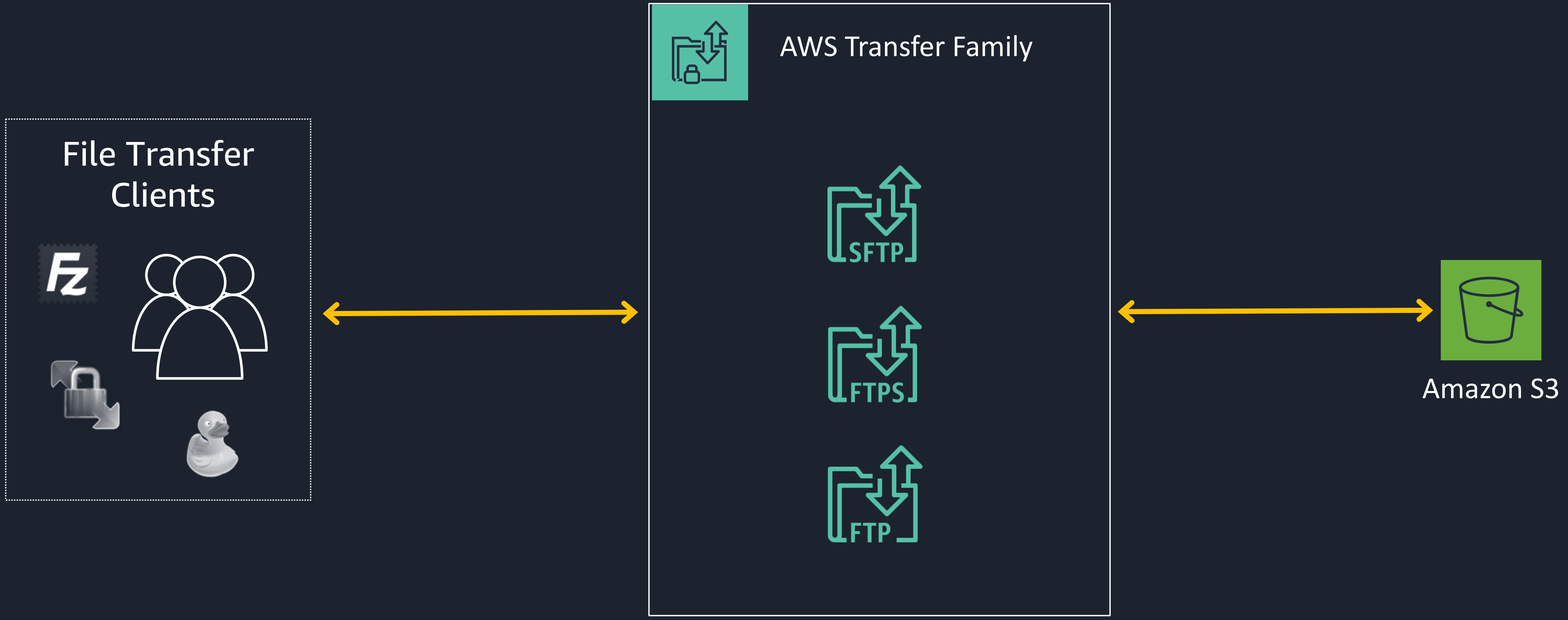
Data Sources – Online ingest



Data Sources – Online ingest



What is the AWS Transfer Family?



AWS Transfer for SFTP

Fully managed SFTP service enabling transfer of data into Amazon S3

Seamless migration of existing SFTP workflows

Fully managed in AWS

Simple

Secure and compliant

Cost-effective

Native integration with AWS services

Simple as 1 – 2 – 3

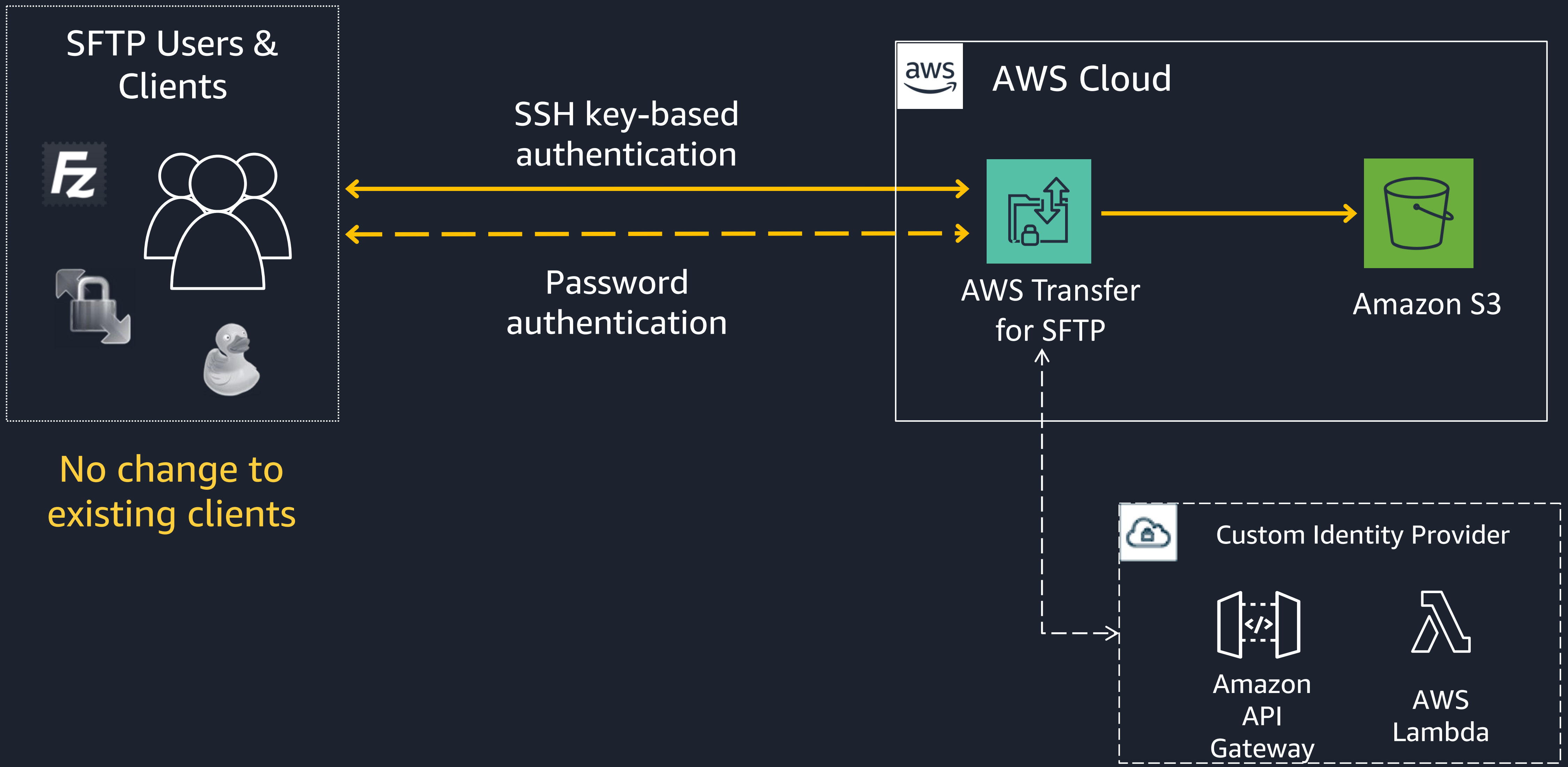
① Deploy an SFTP server endpoint

② Select your target S3 bucket(s)

③ Set up users

How AWS SFTP works

Fully managed, highly available service provides secure access to data in Amazon S3



No change to existing clients

BƎLONG

> “ Our engineers were able to implement a near real-time customer usage analytics framework **within a week**. This timeframe is profound given the Mobile Call Detail Records ingestion comes in varying sizes and frequencies.
>
> AWS SFTP **scales seamlessly** and makes the files available as S3 Objects, which is just perfect for our needs.
>
> This enabled **event driven ingestion** of data into our **Data Lake**.”

Lambros Kallianiotis
Engineering Principal - Belong

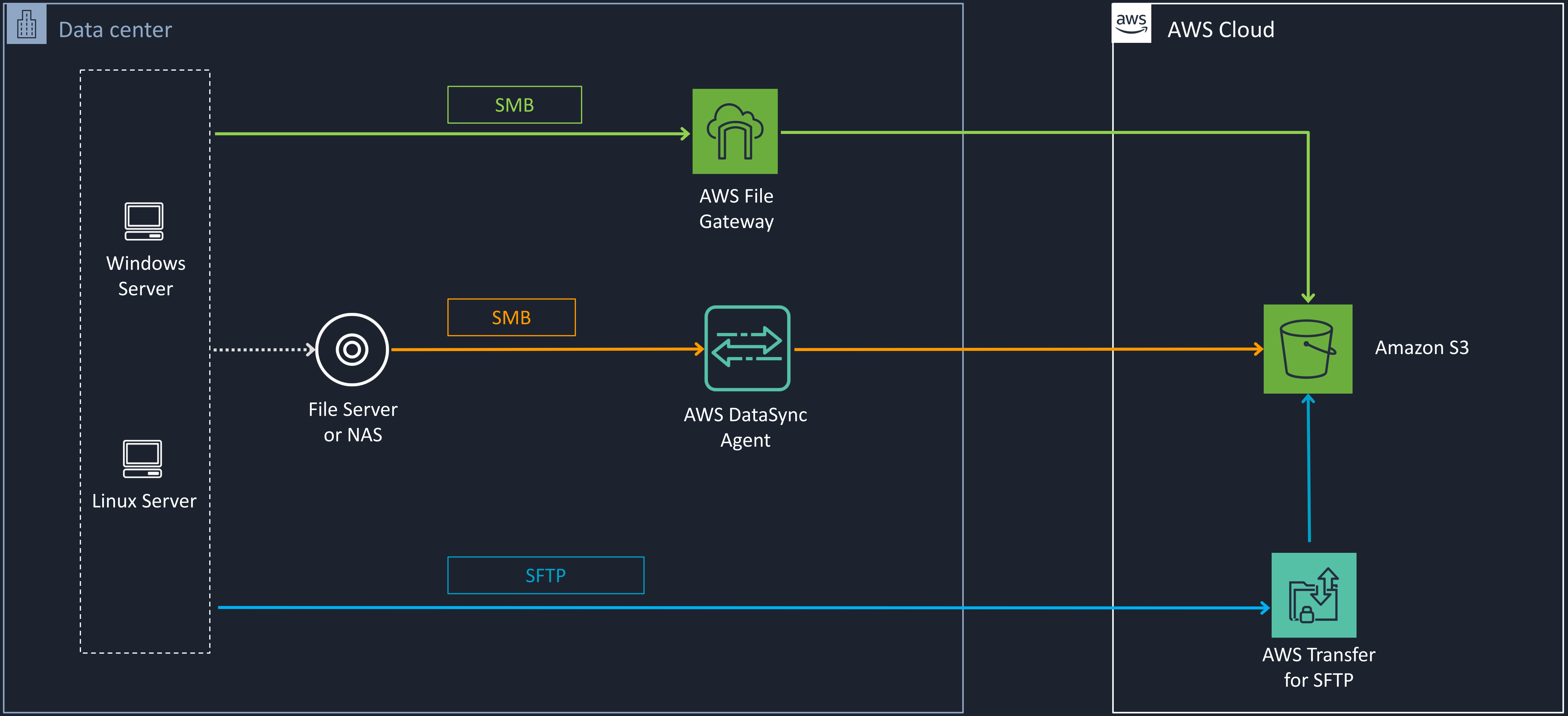
https://aws.amazon.com/aws-transfer-family/customers

Using AWS Transfer for SFTP - Populate your data lake

- Seamless upload of data to Amazon S3 using SFTP clients
- Think of the service’s simplicity at scale
- Utilize for end of event data uploads

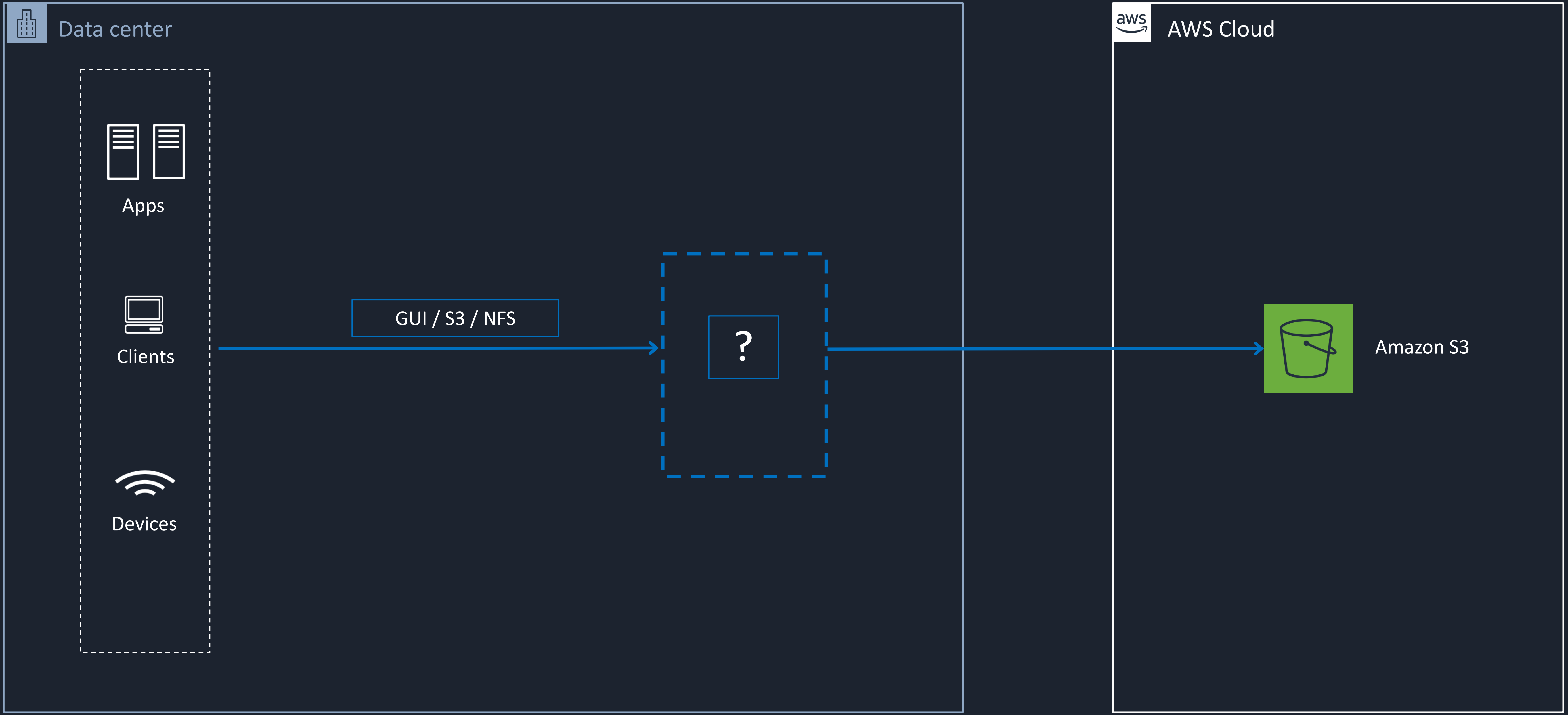
Demo

Demo Setup – Online Ingest

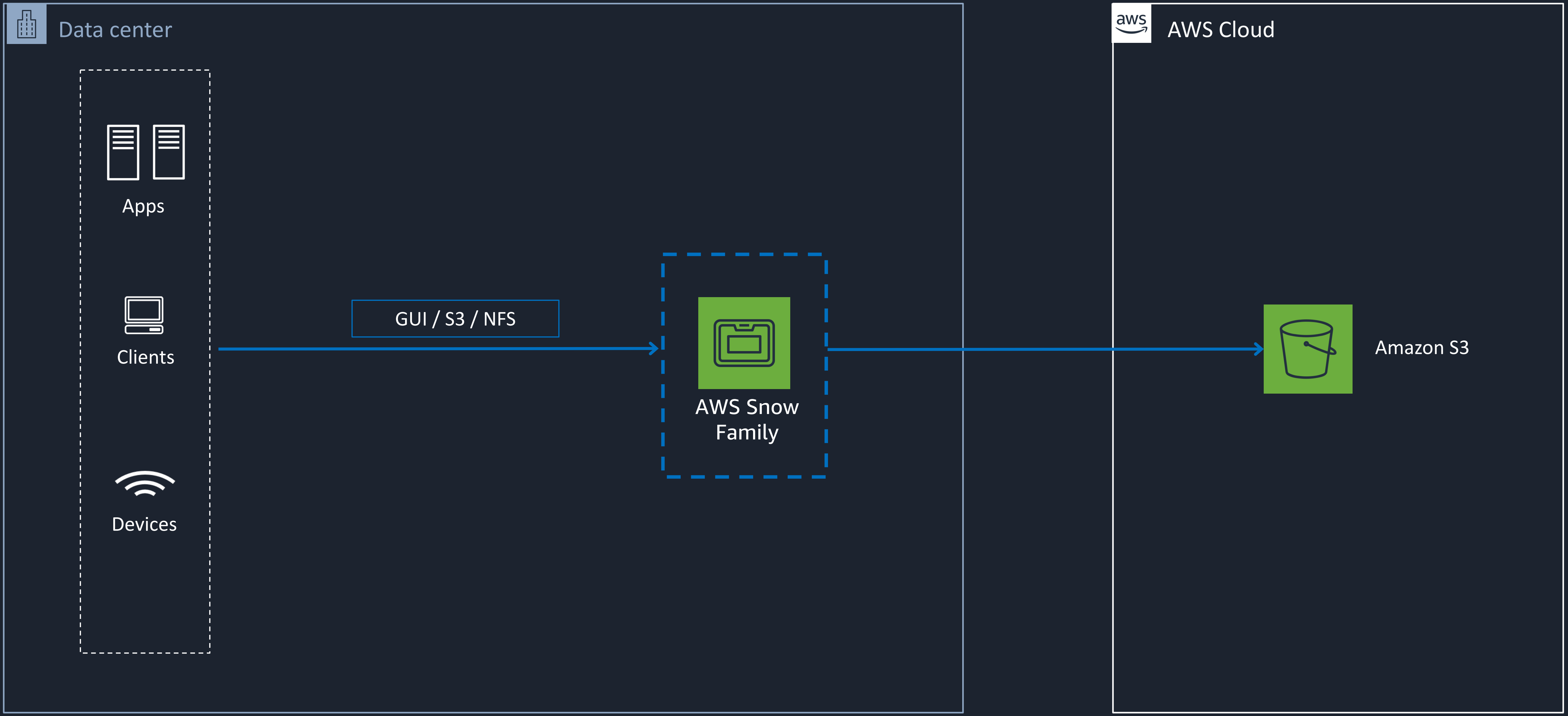


Offline data ingest methods

Data Sources – Offline ingest



Data Sources – Offline ingest



AWS Snow Family Portfolio

AWS Snowball Edge

Data transfer & edge compute

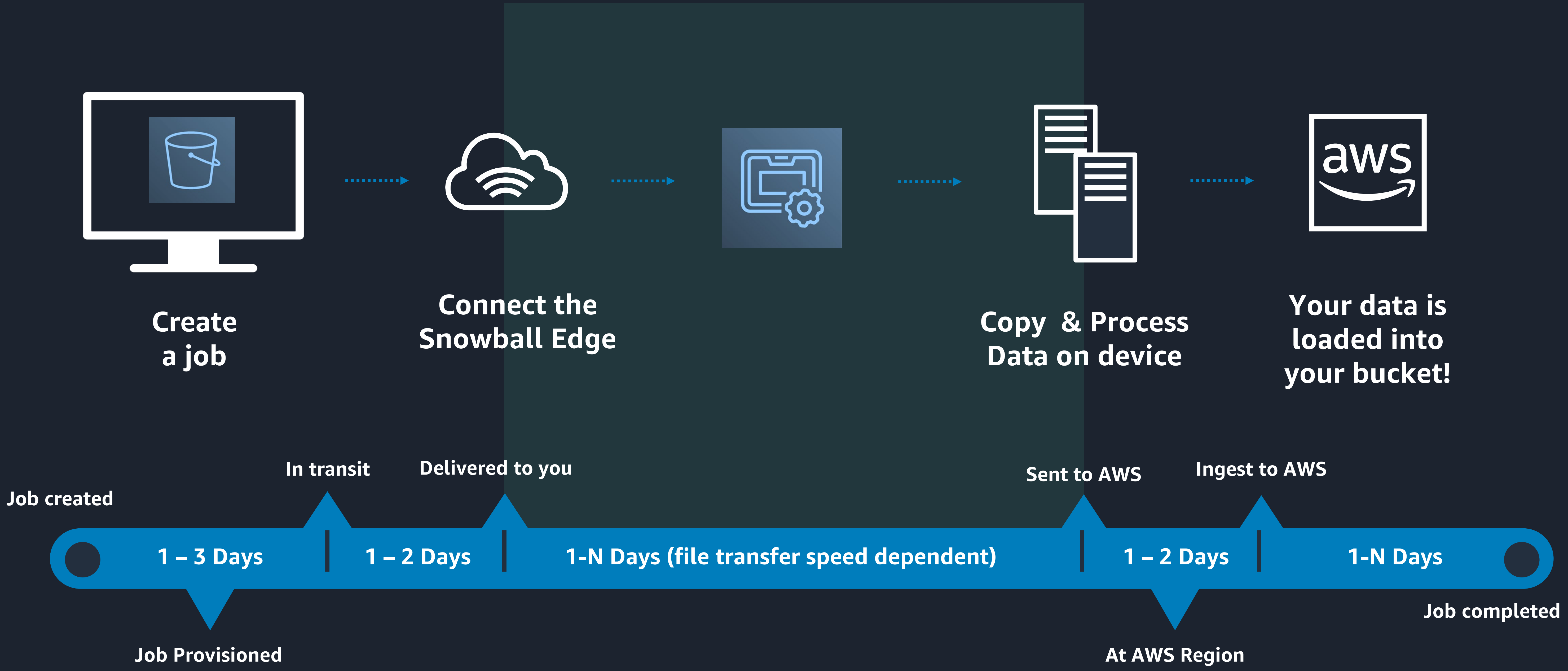
- 42/100TB storage capacity (S3)
- 10/25/40GE networking
- Data encryption end-to-end
- Rugged 8.5 G impact case
- Chain of Custody, Tamper Detection
- Rain and dust resistant
- EC2/AMI support for edge computing
- NFSv4 Server

AWS Snowmobile

20+ PB data transfer

- Exabyte-scale storage in a 45ft container (90PB s3/Glacier/EBS)
- 10/25/40GE networking
- Data encryption end-to-end
- S3/Glacier Data import
- Dedicated security personnel
- GPS tracking, alarm monitoring, 24/7 surveillance, and optional additional security

AWS Snowball Edge import workflow



Create a job

Connect the Snowball Edge

Copy & Process Data on device

Your data is loaded into your bucket!

Job created

Job Provisioned

In transit

Delivered to you

Sent to AWS

At AWS Region

Ingest to AWS

Job completed

1 – 3 Days

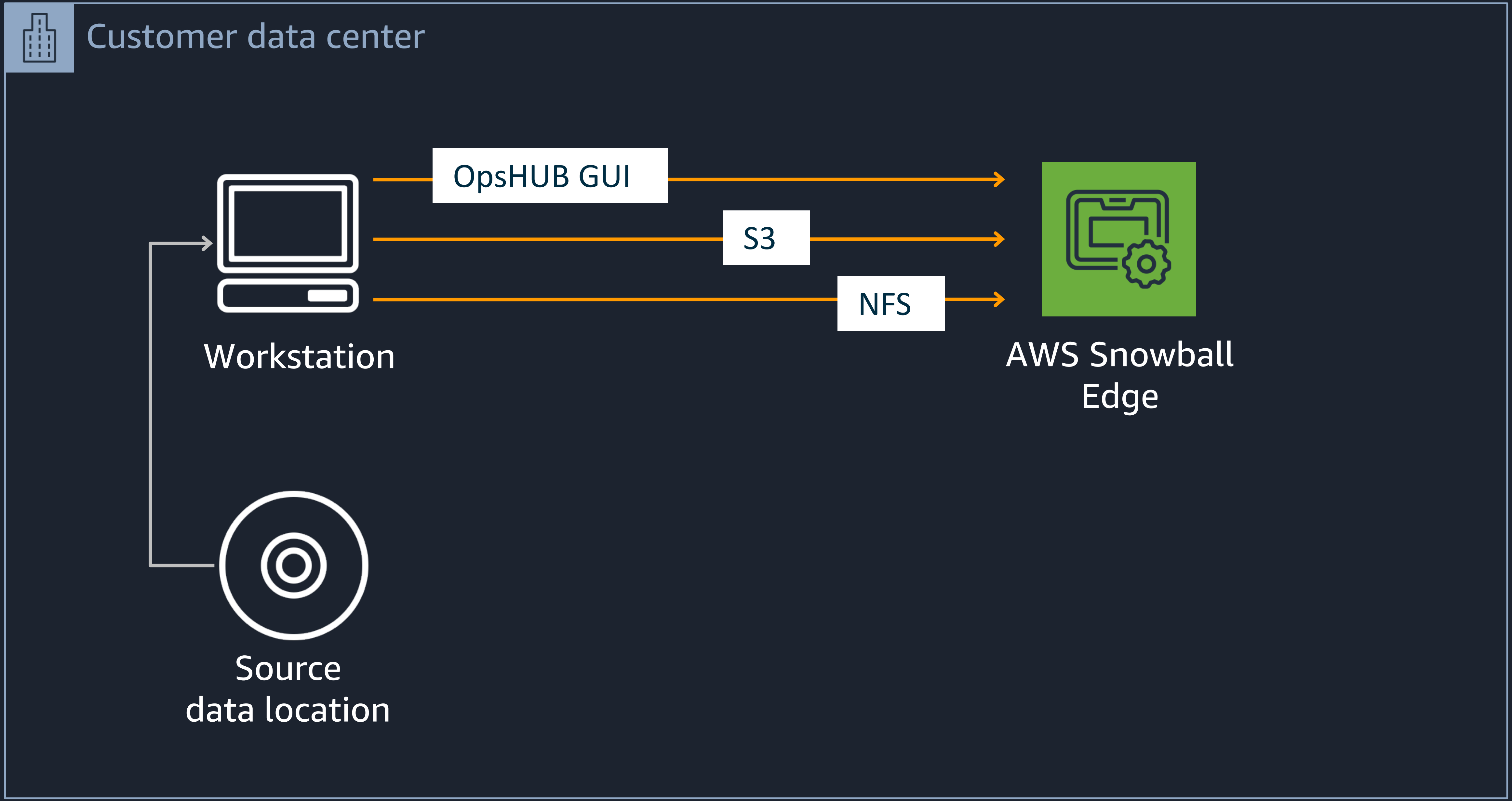
1 – 2 Days

1-N Days (file transfer speed dependent)

1 – 2 Days

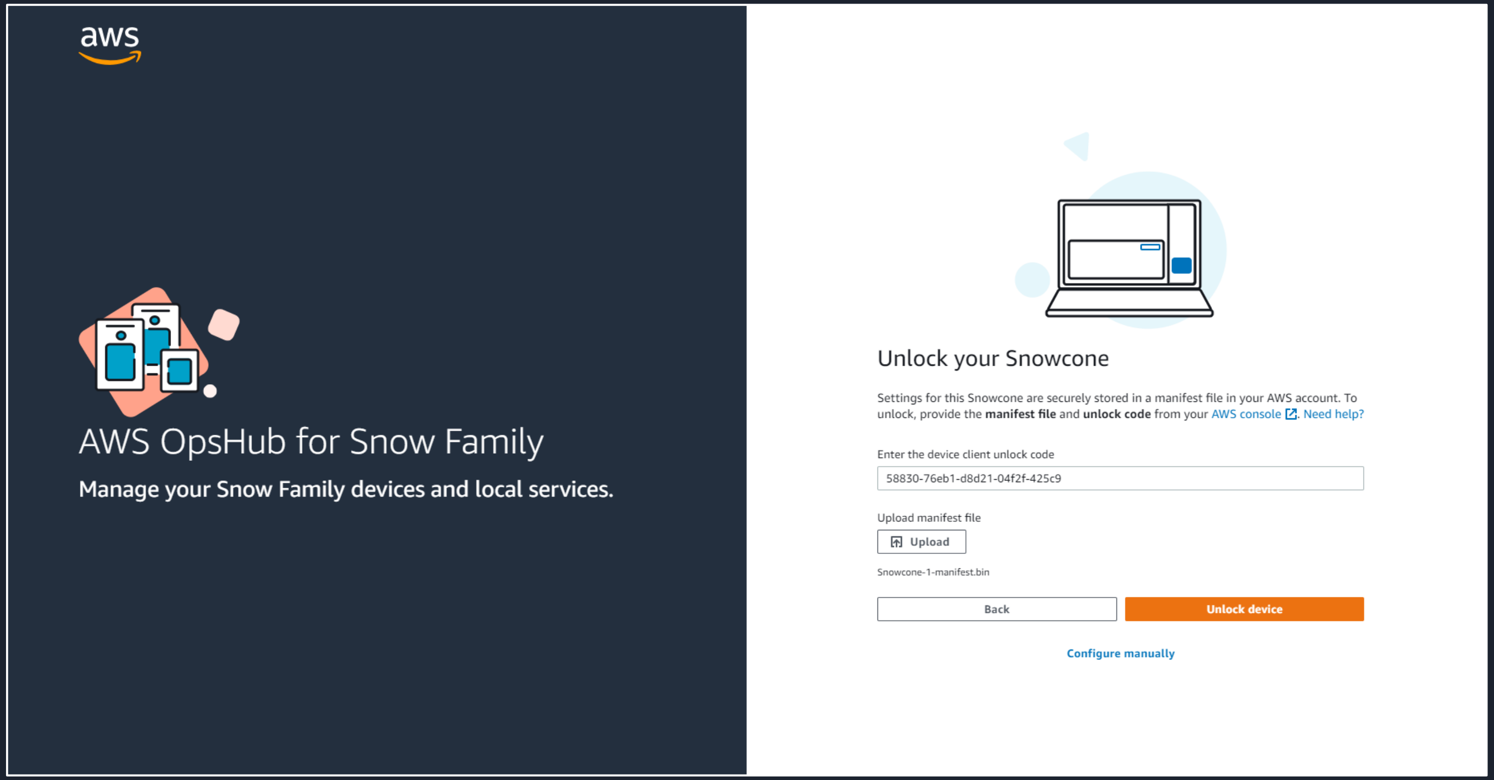
1-N Days

Ingesting data into AWS Snowball Edge



AWS OpsHub for Snow Family

GUI for customers to easily manage Snow devices



AWS OpsHub for Snow Family

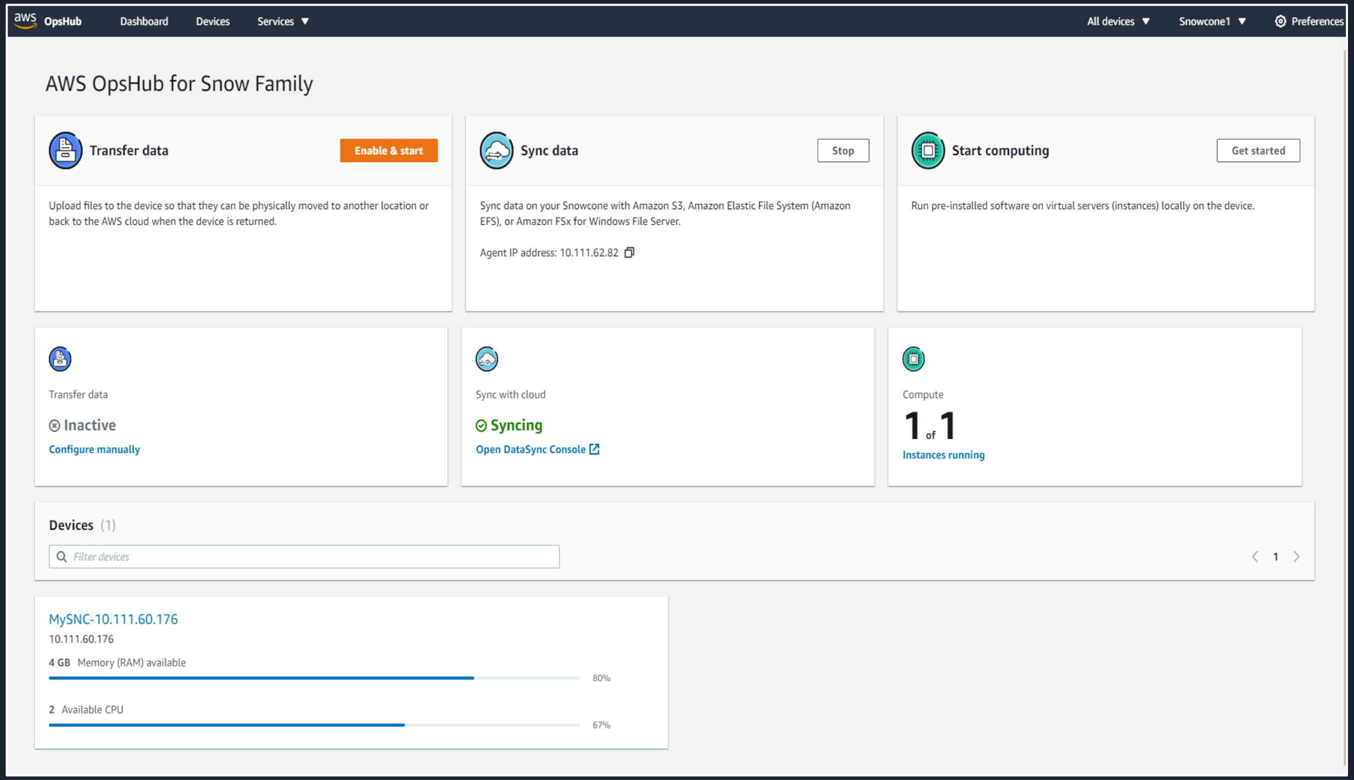
Manage your Snow Family devices and local services.

Unlock your Snowcone

Settings for this Snowcone are securely stored in a manifest file in your AWS account. To unlock, provide the **manifest file** and **unlock code** from your AWS console. Need help?

Enter the device client unlock code: 58830-76eb1-d8d21-04f2f-425c9

Upload manifest file: Snowcone-1-manifest.bin



AWS OpsHub for Snow Family

Transfer data

Upload files to the device so that they can be physically moved to another location or back to the AWS cloud when the device is returned.

Sync data

Sync data on your Snowcone with Amazon S3, Amazon Elastic File System (Amazon EFS), or Amazon FSx for Windows File Server.

Agent IP address: 10.111.62.82

Start computing

Run pre-installed software on virtual servers (instances) locally on the device.

Transfer data: Inactive

Sync with cloud: Syncing

Compute: 1 of 1 Instances running

Devices (1)

MySNC-10.111.60.176
10.111.60.176

4 GB Memory (RAM) available 80%

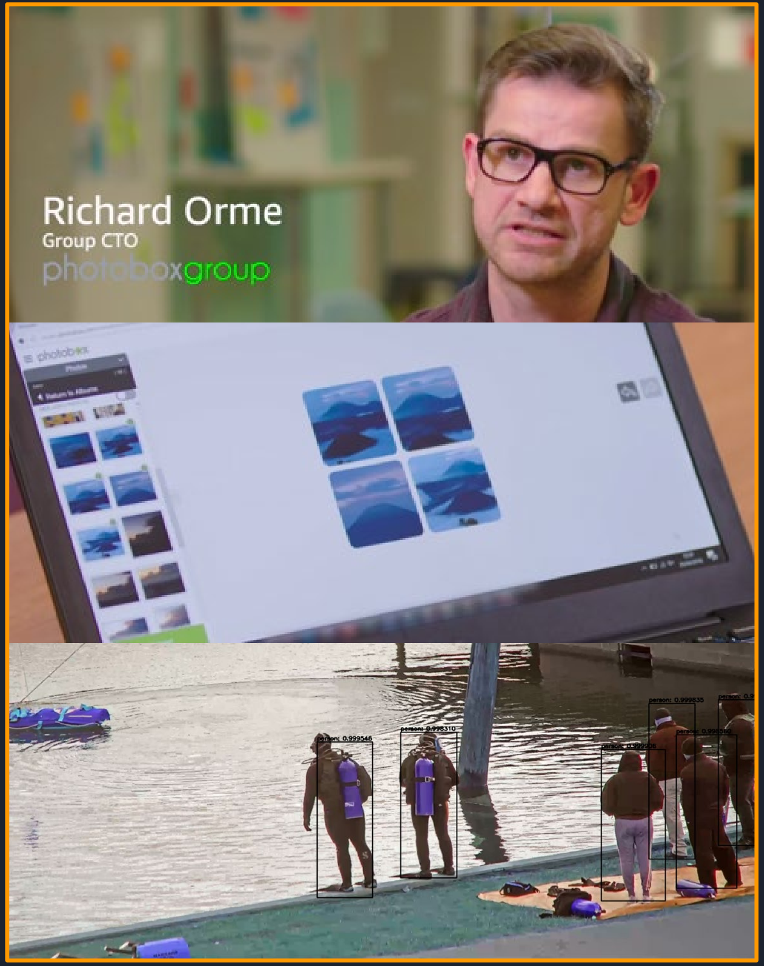
2 Available CPU 67%

Using AWS Snowball Edge - Populating your data lake

- Bulk data transfers - data is not required immediately
- Bulk data transfers - limited network bandwidth
- Perform data transformation at edge using AWS Snowball Edge compute to pre-process data before ingest into Amazon S3 data lake

Data transfer at scale with AWS Snowball Edge

PHOTOBOX



Europe’s #1 online photo service

Migrated to AWS from two data centers

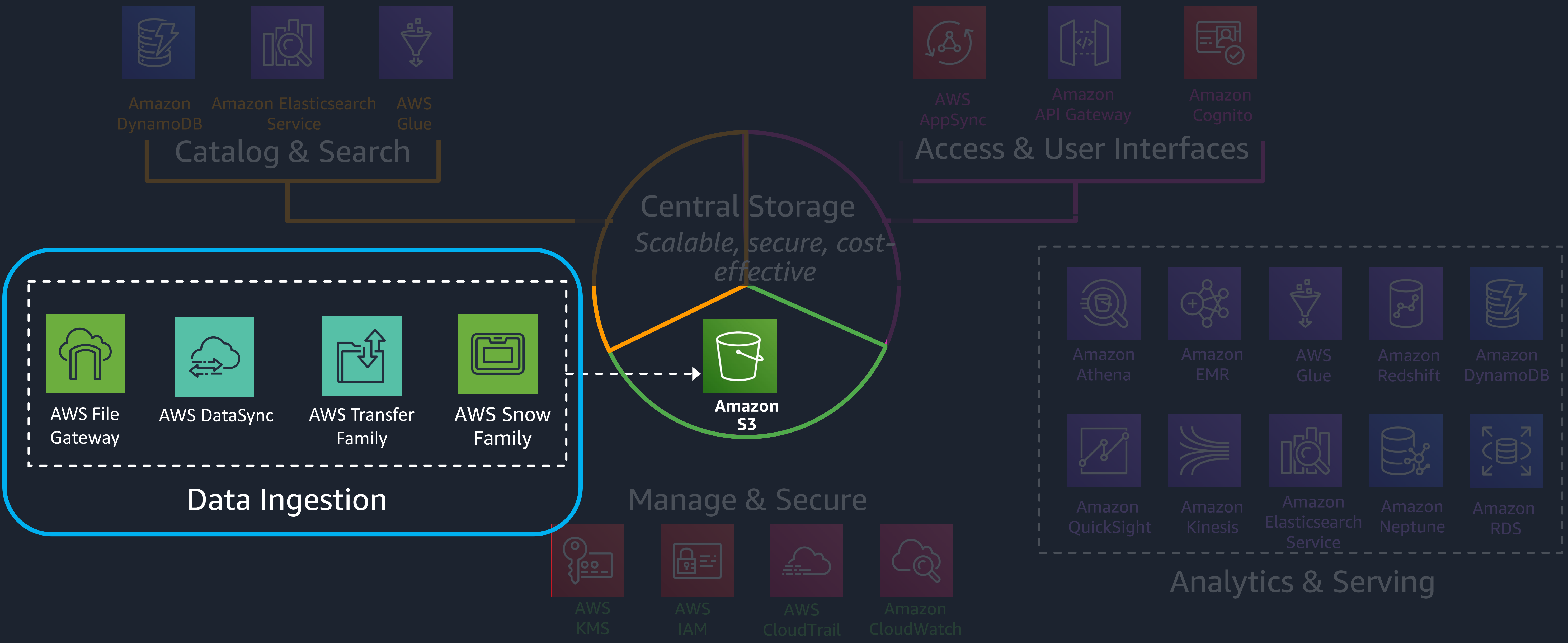
- Used AWS Snowball Edge to move 10 PB (5.7 billion) of photos from Dell EMC Isilon and IBM Cleversafe to Amazon S3
- Needed Amazon S3 data durability; higher than colos and other clouds

Shifted investments and focus to innovation and product development for customers, away from IT infrastructure

> “We’ve reduced costs. We’ve improved our customer experience. Generally, we’ve made our website faster. And that’s because AWS manages that infrastructure in a way we could never do internally.”
>
> **— Chris Astall, Group Director of Architecture**

https://aws.amazon.com/solutions/case-studies/photobox/

Recap – Ingesting data into your data lake on AWS



Get up and running with these resources

Get hands-on experience with the AWS online data migration workshop

https://github.com/aws-samples/aws-online-data-migration-workshop

AWS DataSync : https://aws.amazon.com/datasync

AWS Snow Family : https://aws.amazon.com/snow

AWS Storage Gateway : https://aws.amazon.com/storagegateway

AWS Transfer Family : https://aws.amazon.com/aws-transfer-family

AWS Training and Certification

Training for the Whole Team

Explore tailored Data or Database learning paths for customers and partners

Flexibility to Learn Your Way

Build cloud skills with free digital Data training courses such as “The elements of Data Science”, or dive deep with classroom training

Validate Skills with AWS Certification

Demonstrate expertise with a Data industry-recognized credential (Data analytics and Database Specialty AWS Certifications)

aws.amazon.com/training/

Visit the Data, Databases, and Analytics Resource Hub for more resources

Dive deeper with these newly created whitepapers and e-books to help you uncover new insights and value from your data

- An introduction to cloud databases
- Enter the purpose-built database era
- Harness the power of data
- Creating a modern analytics architecture
- The data-driven enterprise
- … and more!

https://tinyurl.com/aws-data-databases-analytics

Visit resource hub »

Thank you for attending
AWS Data, Databases, and Analytics Online Series

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

- aws-apac-marketing@amazon.com
- twitter.com/AWSCloud
- facebook.com/AmazonWebServices
- youtube.com/user/AmazonWebServices
- slideshare.net/AmazonWebServices
- twitch.tv/aws

Thank you!