INNOVATE

DATA EDITION

23 August, 2022

Easy ways to migrate petabytes of data to Amazon S3 using AWS DataSync

Ameen Khan S

Senior Storage Specialist Solutions Architect
Amazon Web Services

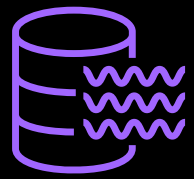
Agenda

- Large-scale data transfer challenges
- Data migration options on AWS
- Why AWS DataSync?
- Customer use case on Hadoop Distributed File Systems (HDFS) data transfer to AWS
- Demo on HDFS data transfer using AWS DataSync
- Key takeaways

Data transfer use cases

Why do customers transfer data to the cloud

Application migration



Data lakes

Sharing Data

Backups

Data management

Large-scale data transfers are challenging

- Building and deploying scripts
- Encrypting and verifying data
- Recovering from errors
- Handling network availability
- Ensuring performance

Data transfer options on AWS

Offline

AWS Snow Family
- AWS Snowball
- AWS Snowball Edge
- AWS Snowmobile

Move terabytes to petabytes of data to AWS using appliances designed for secure, physical transport

Online

AWS Storage Gateway

Sync files with SMB, NFS, iSCSI protocols from on-premise to AWS

AWS Transfer for SFTP

Transfer files in and out of Amazon S3 with SFTP protocol

AWS Data Sync

Sync files from on-premises file storage to an Amazon EFS file system or Amazon S3 bucket or Amazon FSx

Amazon Kinesis Data Firehose

Capture, process, & load streaming data into AWS

Network Optimization

AWS Direct Connect

Establishes private connectivity between AWS and your on-premises resources

Amazon S3 Transfer Acceleration

Makes Internet transfers to Amazon S3 faster

Why AWS DataSync?

Offline

AWS Snow Family
- Snowball
- Snowball Edge
- Snowmobile

Move terabytes to petabytes of data to AWS using appliances designed for secure, physical transport.

Online

AWS Storage Gateway

Sync files with SMB, NFS, iSCSI protocols from on-premise to AWS

AWS Transfer for SFTP

Transfer files In and out of S3 with SFTP protocol

AWS Data Sync

Sync files from on-premises file storage to an Amazon EFS file system or Amazon S3 bucket or Amazon FSx

Amazon Kinesis Data Firehose

Capture, process, & load streaming data into AWS

Network Optimization

AWS Direct Connect

Establishes private connectivity between AWS and your on-premises resources

Amazon S3 Transfer Acceleration

Makes Internet transfers to S3 faster

What is AWS DataSync?

Online data transfer service
that simplifies, automates, and accelerates
copying file and object data to and from AWS storage

Fast data transfer

- Highly optimized, parallel network transfer (up to 100 TB/day)
- Transfers only incremental changes

Easy to use

- Schedule transfers
- Throttle bandwidth
- Filter by file name patterns

Secure and reliable

- End-to-end encryption
- End-to-end data verification
- VPC endpoints with PrivateLink

Fully managed

- Integrates with AWS management and monitoring services
- Direct transfer into all Amazon S3 storage classes

Cost-effective

- $0.0125 / GB transferred
- No minimums

Santos

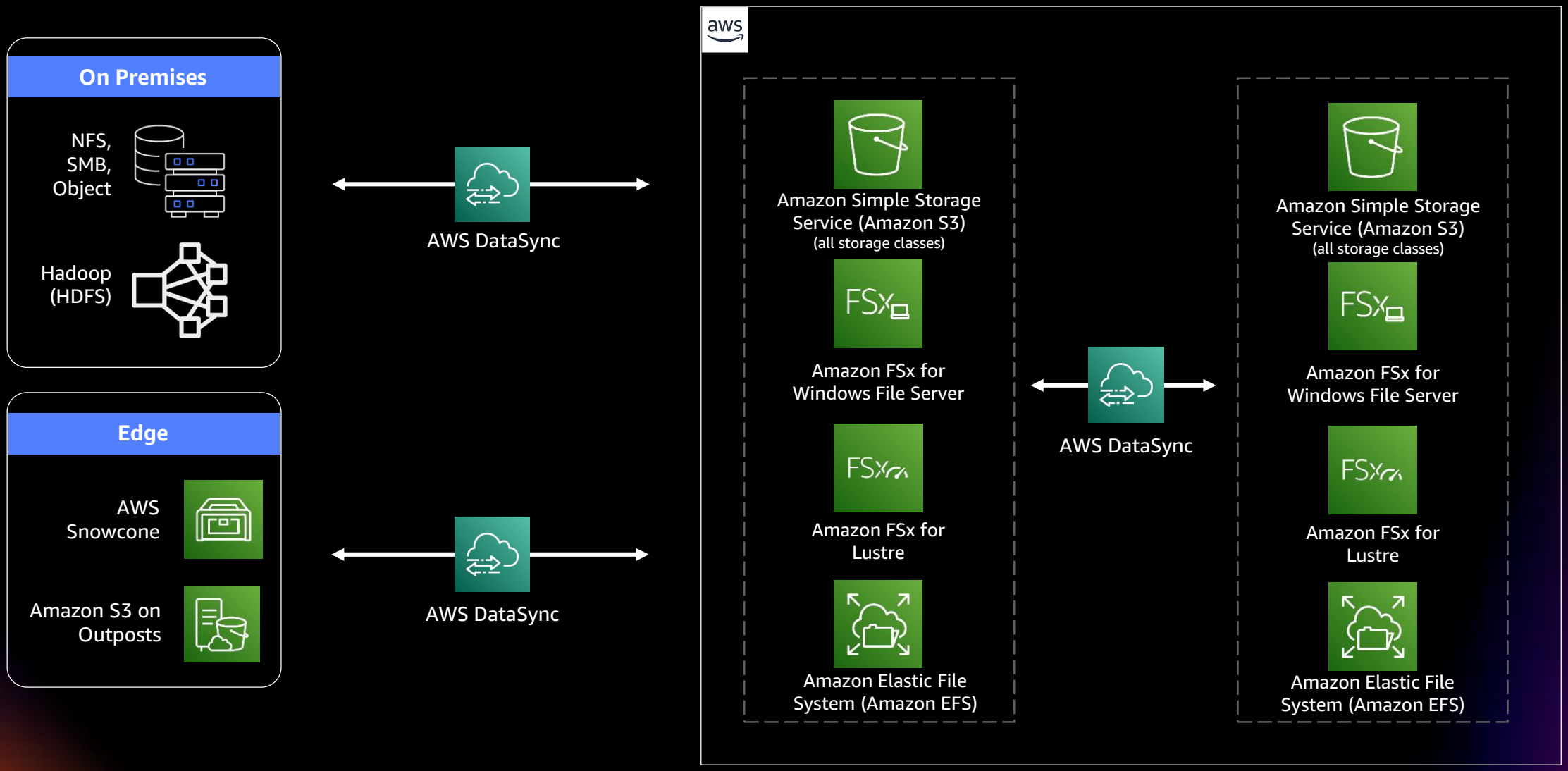
"Prior to learning about AWS DataSync, I had spent over a month creating, testing, and iterating on new scripts to get our backup files into Amazon S3. These scripts needed to be deployed on all of our servers, and were difficult to centrally manage. By transitioning to DataSync, I was able to simplify and automate my backup management. Apart from its simplicity, DataSync also provided additional functionality, such as monitoring and error checking, that was quite valuable. It took less than one hour of my time to set up and get going, and we use DataSync every night to transfer 5 – 50 TB of data, which is roughly 3,000 files. For the task that I needed it for, DataSync was absolutely spot on"

Gavin Boyce, Cloud Solution Architect - Santos

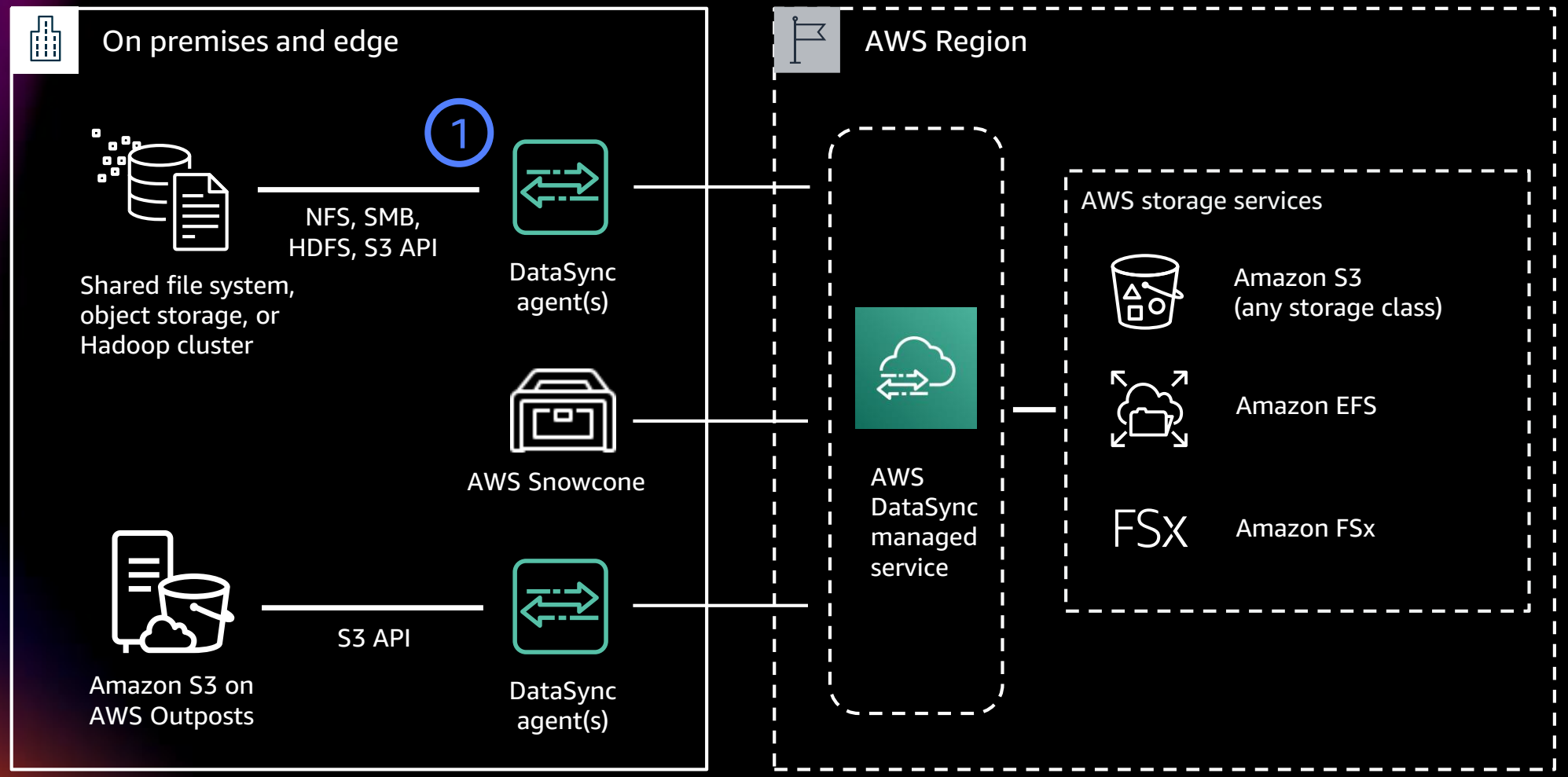
Learn more

aws

What can you do with AWS DataSync?

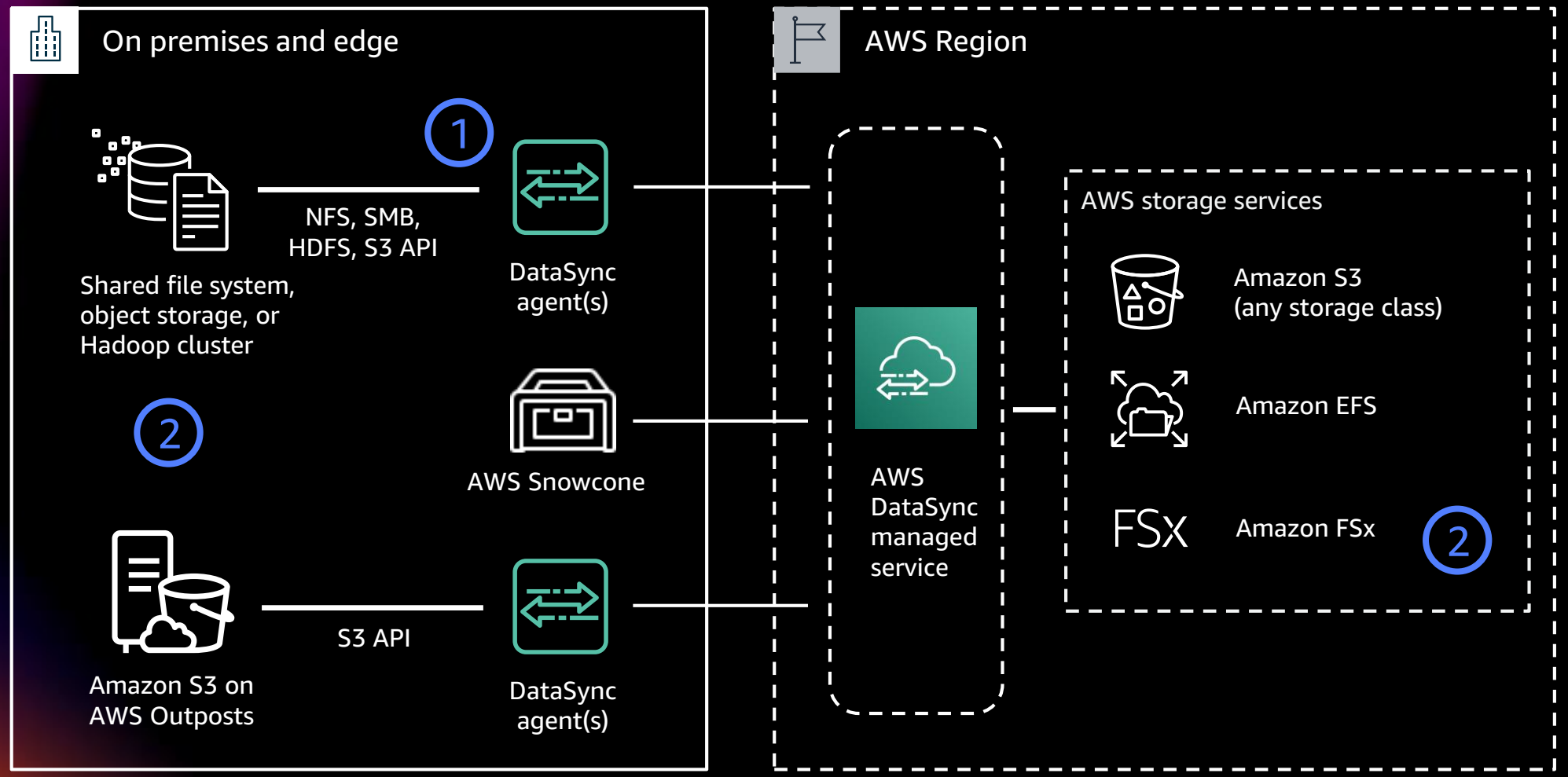


AWS DataSync: How it works



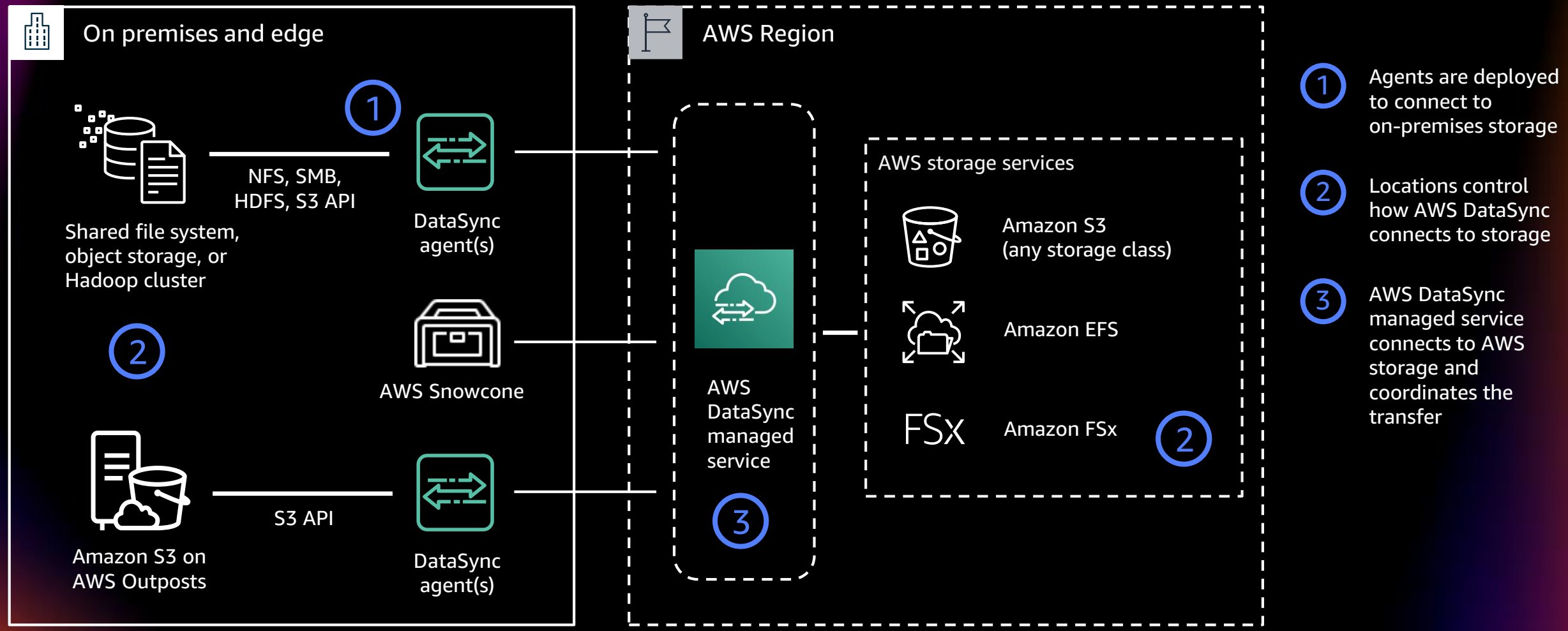
1 Agents are deployed to connect to on-premises storage

AWS DataSync: How it works



1. Agents are deployed to connect to on-premises storage
2. Locations control how AWS DataSync connects to storage

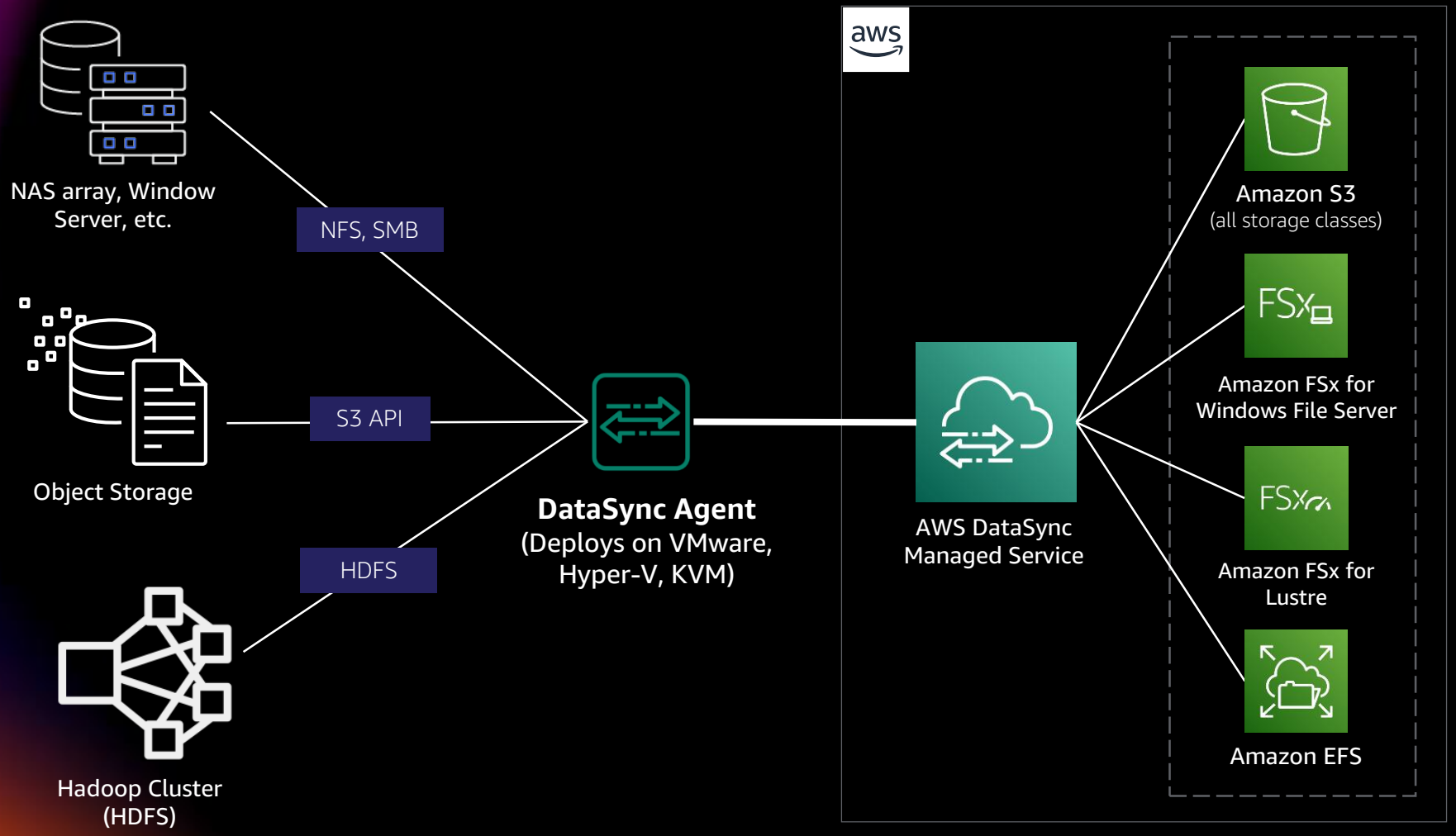
AWS DataSync: How it works



1. Agents are deployed to connect to on-premises storage
2. Locations control how AWS DataSync connects to storage
3. AWS DataSync managed service connects to AWS storage and coordinates the transfer

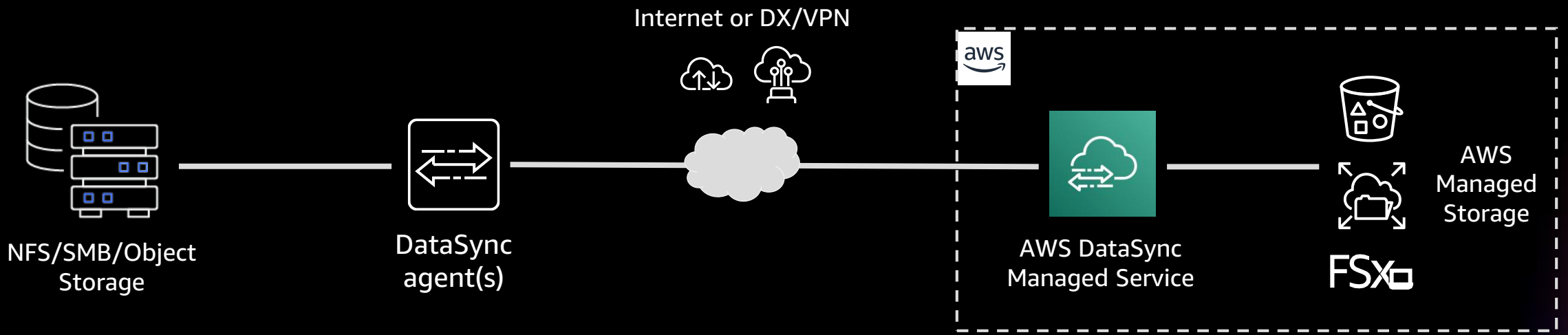
On-premises transfers

Transfer data between on-premises storage systems and AWS Storage services

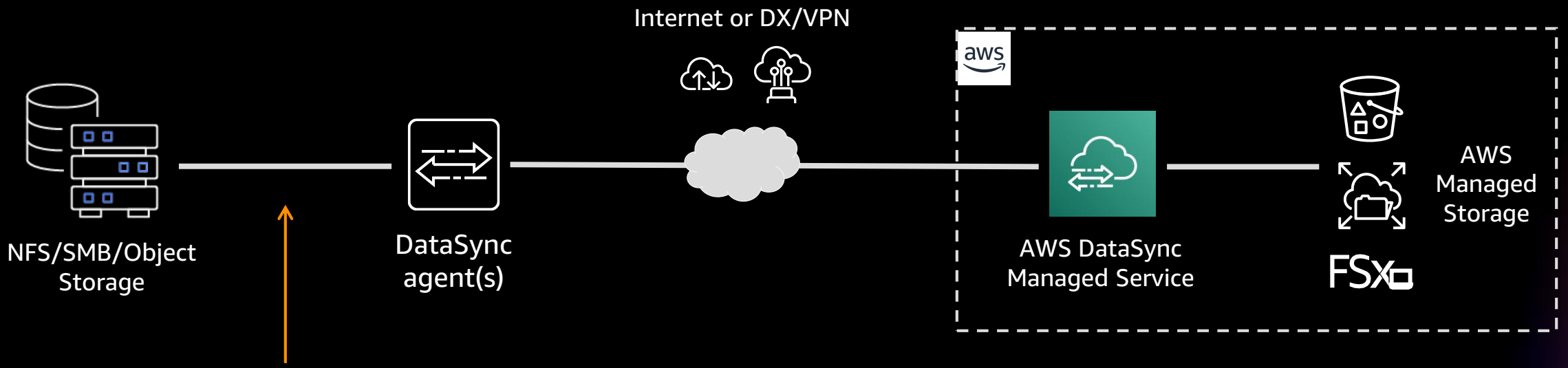


- ✓ Support for a wide variety of on-premises and self-managed storage
- ✓ Copy to and from any supported AWS Storage services
- ✓ Transfer data over the internet or using AWS Direct Connect
- ✓ All traffic between agent and AWS encrypted in flight using TLS 1.2

AWS DataSync network path for on-premises transfers

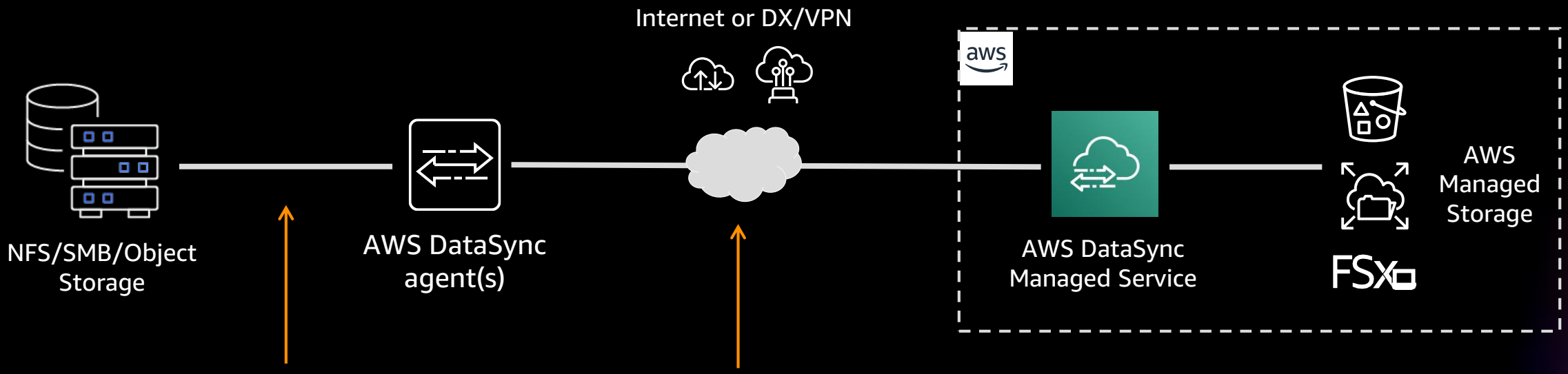


AWS DataSync network path for on-premises transfers



This needs to be fast and low latency. Install the agents as close to the customer-managed storage as possible

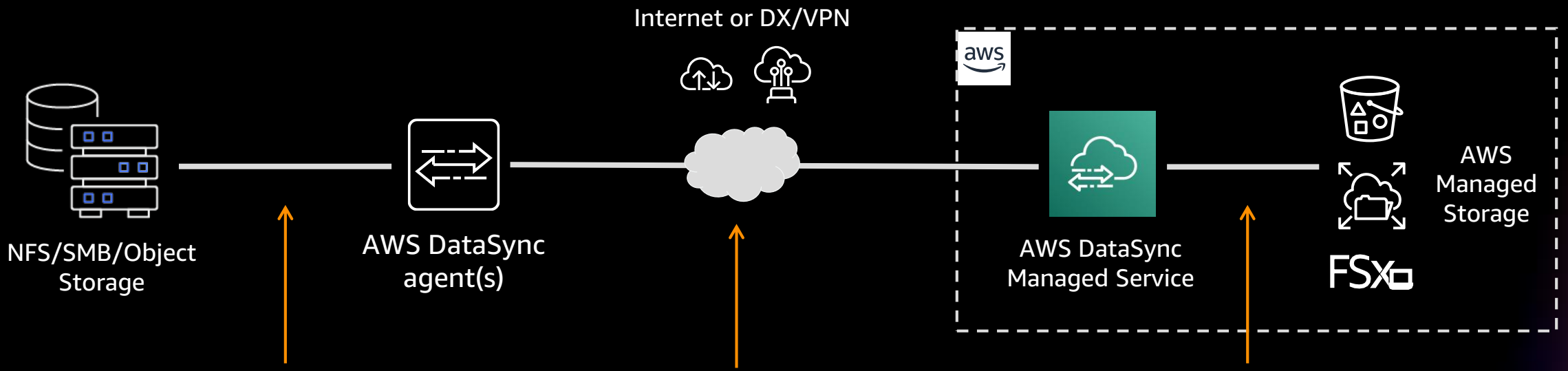
AWS DataSync network path for on-premises transfers



This needs to be fast and low latency. Install the agents as close to the customer-managed storage as possible

AWS DataSync is optimized for traversing WAN links. All data is encrypted in flight using TLS

AWS DataSync network path for on-premises transfers

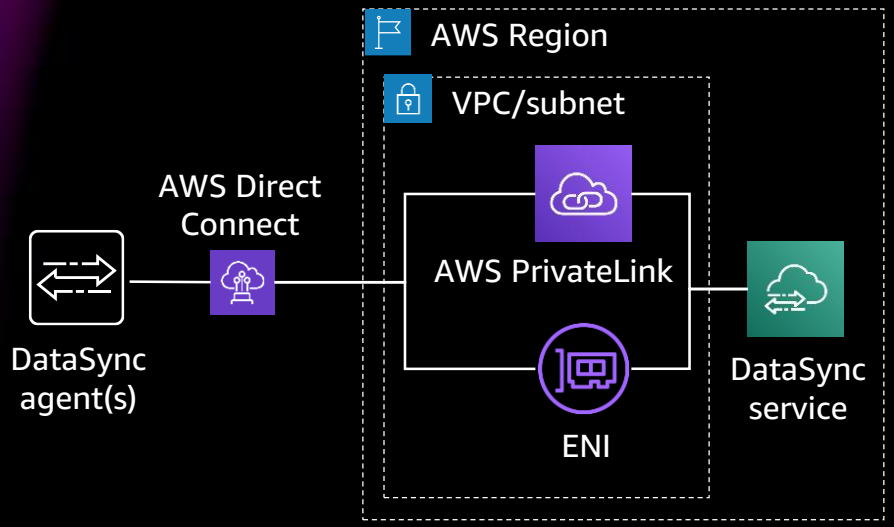


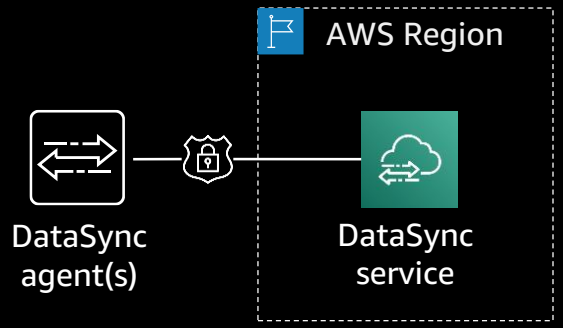
This needs to be fast and low latency. Install the agents as close to the customer-managed storage as possible

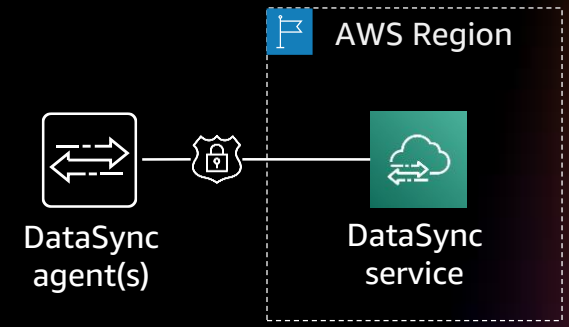
DataSync is optimized for traversing WAN links. All data is encrypted in flight using TLS

This is fast, internal networking automatically configured by DataSync. Customer only manages security groups for accessing Amazon EFS or Amazon FSx

AWS DataSync network endpoint types







VPC endpoints

Data remains within VPC

Connect over AWS Direct Connect to private VPC/subnet

Public endpoints

Internet-facing service endpoints

Connect over the internet or AWS Direct Connect with an internet gateway

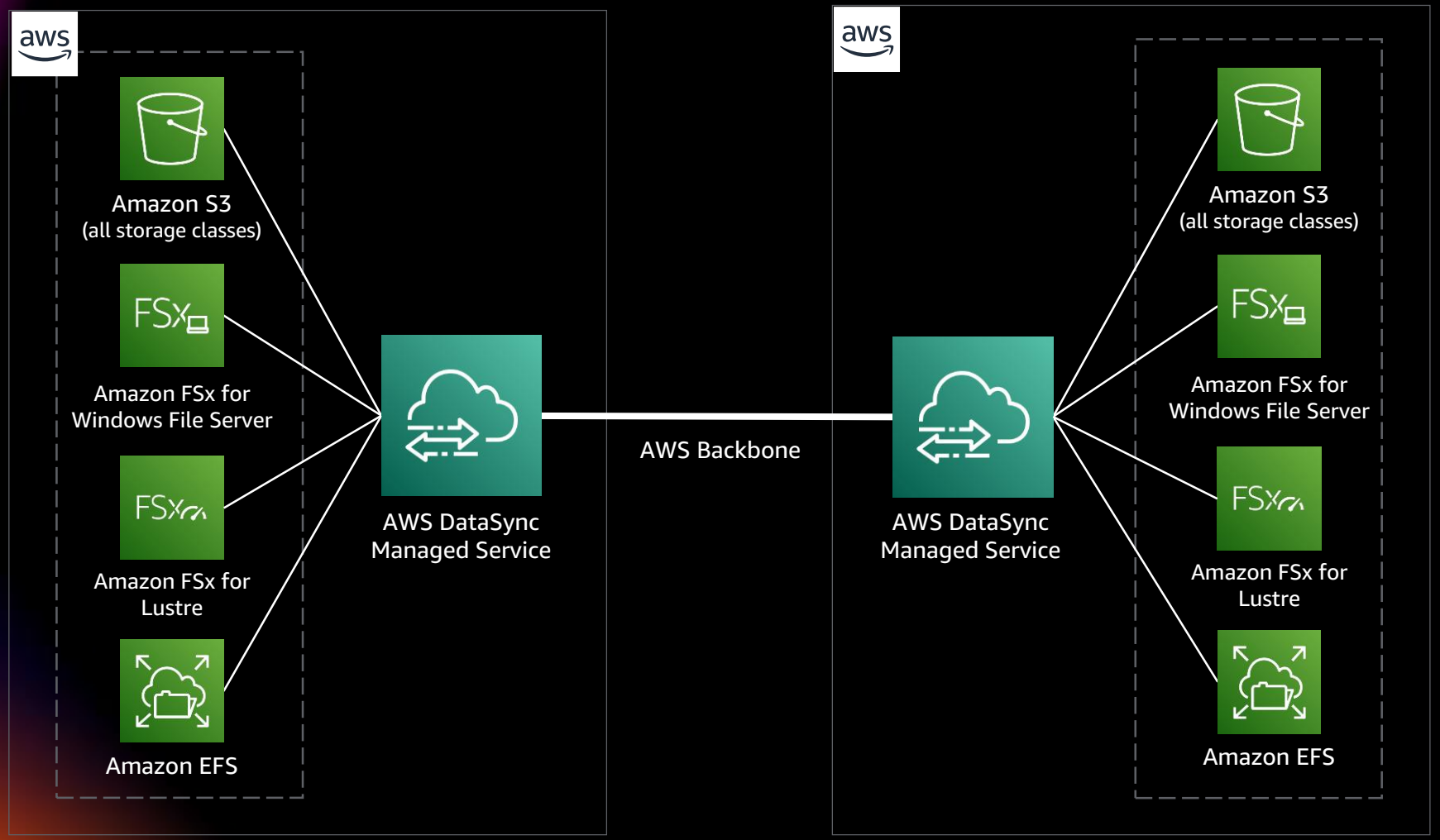
FIPS endpoints

FIPS-compliant endpoints

Connect over the internet or AWS Direct Connect with an internet gateway

Data transfers within AWS

Transfer data between AWS Storage services quickly, easily, and securely



- ✓ No infrastructure to deploy or manage
- ✓ Copy data between any supported AWS Storage services
- ✓ Copy data within the same region or across regions
- ✓ All traffic stays within the AWS network
- ✓ All traffic encrypted in flight using TLS 1.2

Customer use case

Petabytes of HDFS data transfer to Amazon S3

Business requirements

- Customer wants to transfer > 6 PB of data in < 4 months
- Current source of data is in HDFS (Historical & Incremental)
- Destination storage on AWS is Amazon S3
- Customer industry vertical is telecom hence security is a priority

Challenges

- Tight timeline as the source HDFS system license is due for renewal
- Capacity issues with current system
- Incremental data to sync
- Offline transfer
- Data transfer at scale

Technical requirements

- 10Gb/s Dx Network bandwidth between on-prem to AWS
- Transfer mechanism to be on-demand and at scale
- Run parallel transfers and optimal bandwidth utilization

Understand your network bandwidth

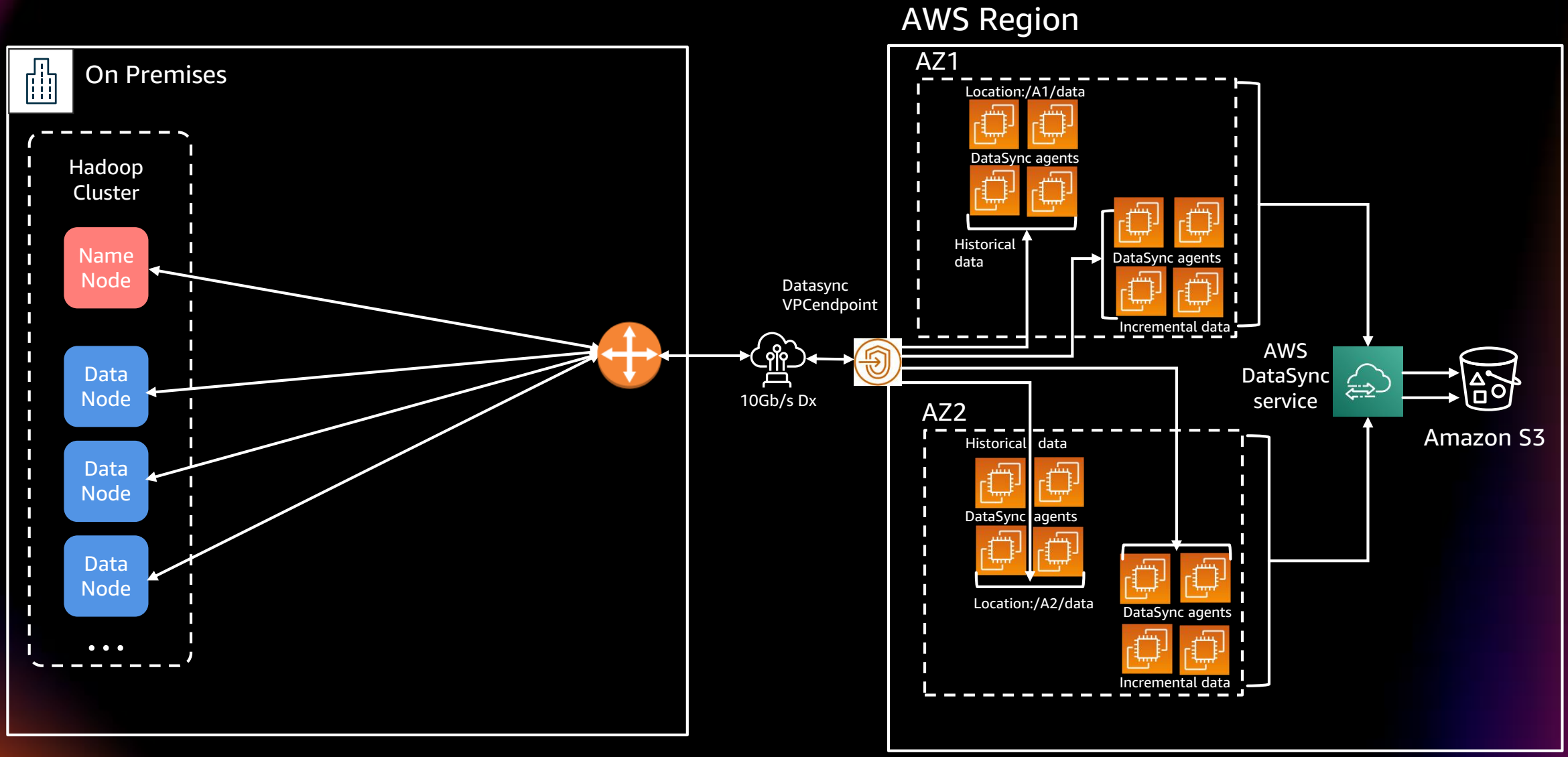
	100 Mbps	1 Gbps	10 Gbps
1 TB	30 hours	3 hours	18 minutes
10 TB	12 days	30 hours	3 hours
100 TB	124 days	12 days	30 hours
1 PB	3 years	124 days	12 days
10 PB	34 years	3 years	124 days

- Plan for **available** bandwidth
- Task bandwidth can be throttled in MiB/s
- If running multiple tasks, aggregate throttling across tasks

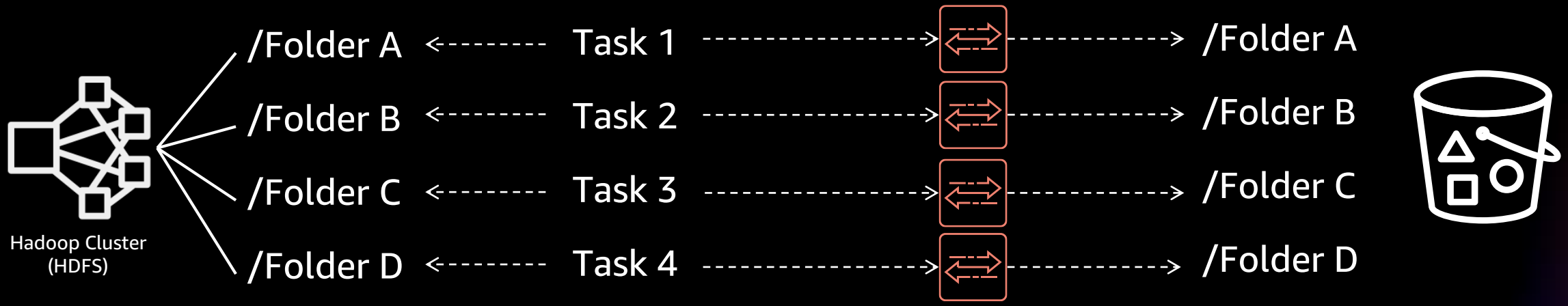
How do we solve with AWS DataSync?

- Build an architecture which can scale during the transfer
- Data security and integrity at all levels of online data transfer
- Optimal utilization of tasks and locations

Architecture



Tasks & locations



Partition large data sources by copying from different folders

Use multiple agents to run tasks in parallel, to fully utilize bandwidth

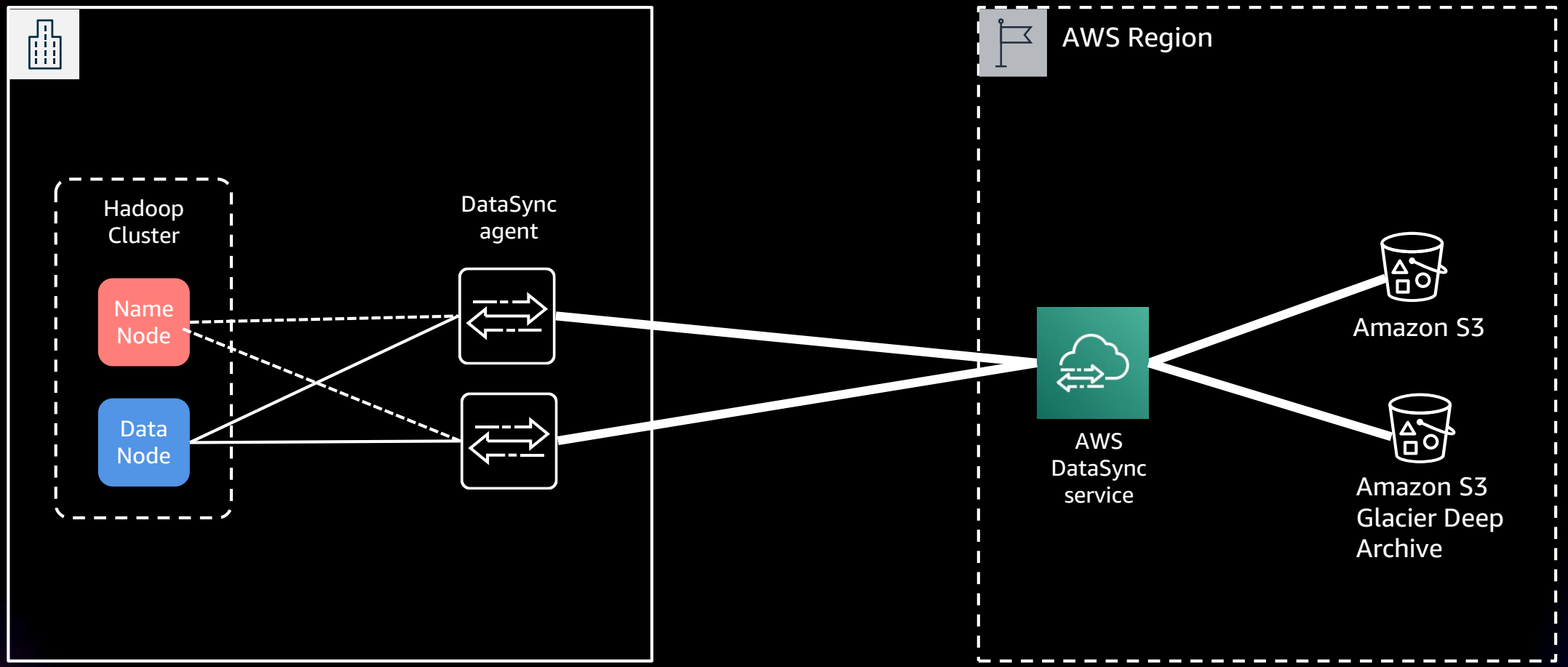
Use different folders/prefixes per task when copying to the same destination.

Best practices

1. Know your end state and the time line for migration
2. Understand your network requirements to make a conscious decision online vs offline
3. Plan the network requirement in advance with allowed port list, firewall configuration etc
4. Divide the data transfer tasks & location based criticality of data
5. Include parallel tasks and pattern per location is key

Demo – Data transfer on using AWS DataSync

Demo



Demo takeaways

1. Migrate HDFS file data from Hadoop cluster to Amazon S3 standard and Amazon S3 Glacier Deep Archive

2. Create task and location creation options

3. Review the transferred files in Amazon S3 and the performance metrics

Recap / Key takeaways

1. Large-scale data transfer challenges
2. Data migration options on AWS
3. Why AWS DataSync?
4. HDFS data transfer to AWS customer use cases
5. Demo - Data transfer on using AWS DataSync
6. Next steps

Other resources

Blog

https://aws.amazon.com/blogs/storage/using-aws-datasync-to-move-data-from-hadoop-to-amazon-s3/

Github Link

https://github.com/aws-samples/aws-datasync-migration-workshop

Visit the AWS Data resource hub

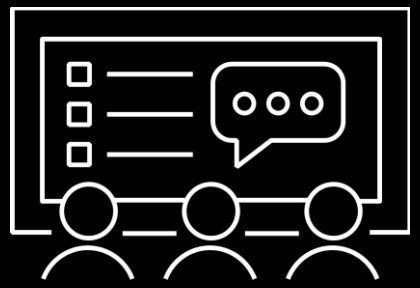
A modern data strategy can help you manage, act on, and react to your data so you can make better decisions, respond faster, and uncover new opportunities. Dive deeper with these resources today.

- Harness data to reinvent your organization
- In unpredictable times, a data strategy is key
- Make data a strategic asset
- Rewiring your culture to be data-driven
- Put your data to work with a modern analytics approach
- … and more!

https://tinyurl.com/data-hub-aws

Visit resource hub

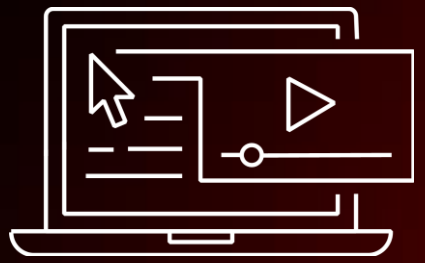
AWS Training and Certification for Data and Analytics



AWS Data & Analytics FREE Training Resources

Discover how to harness data, one of the world's most valuable resources, and innovate at scale.

https://bit.ly/3Ntlhy7



AWS Data Analytics Learning Plan

This learning plan expose you to the fastest way to get answers from all your data to all your users. It can also help prepare you for the AWS Certified Data Analytics - Specialty certification exam.

https://bit.ly/3wBVjD1

AWS Certified Data Analytics - Specialty

Earning AWS Certified Data Analytics – Specialty validates expertise in using AWS data lakes and analytics services.

https://go.aws/3lwF0RR

Thank you for attending AWS Innovate – Data Edition

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-apj-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

Thank you!