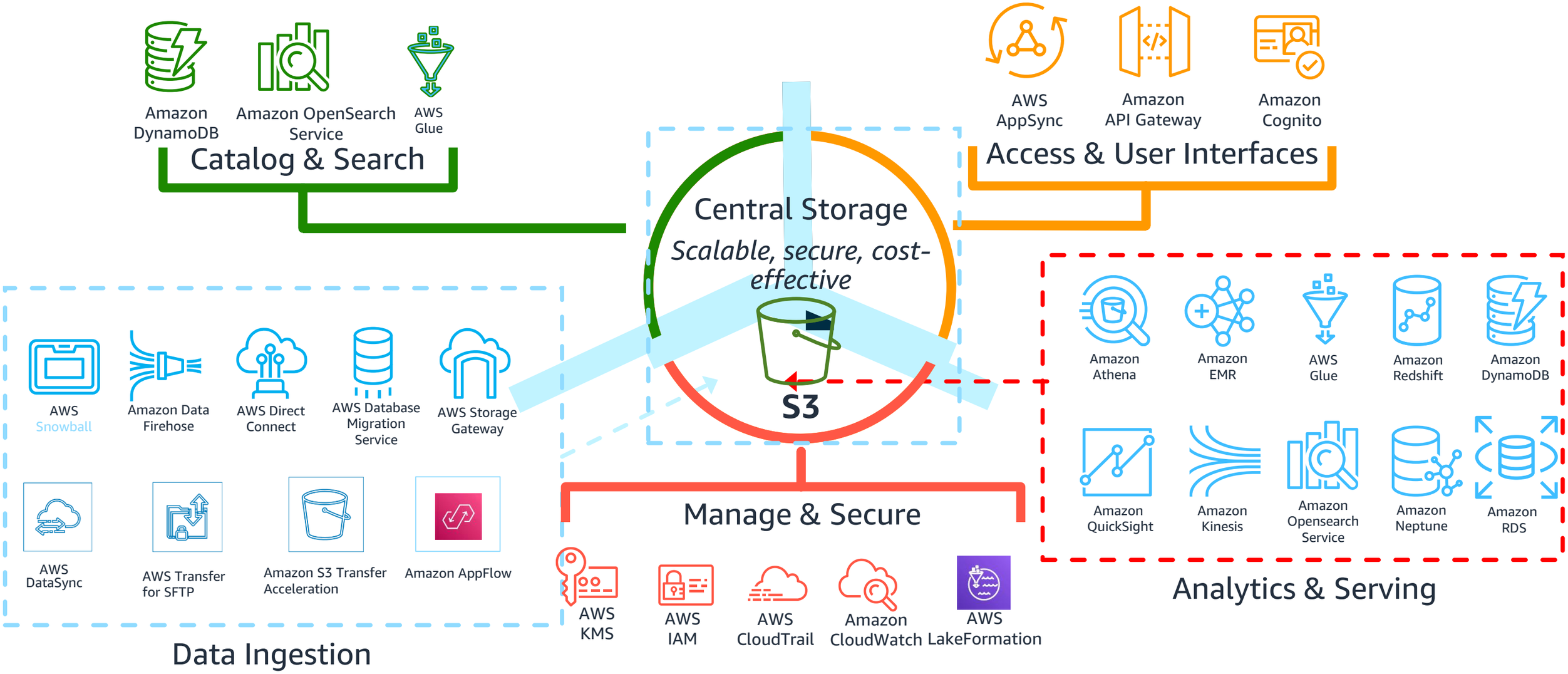
Hydrating the Data Lake

Paige Broderick
Solutions Architect
Amazon Web Services

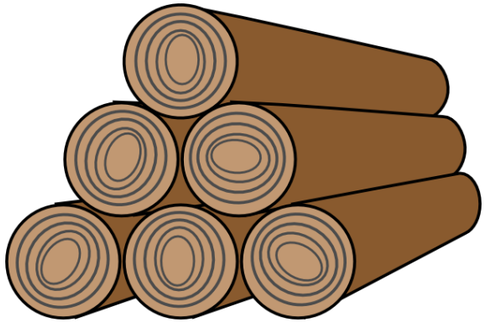
Session's Focus – Working In The Data Lake



Data Sources

Databases

Streams



Logs

Files

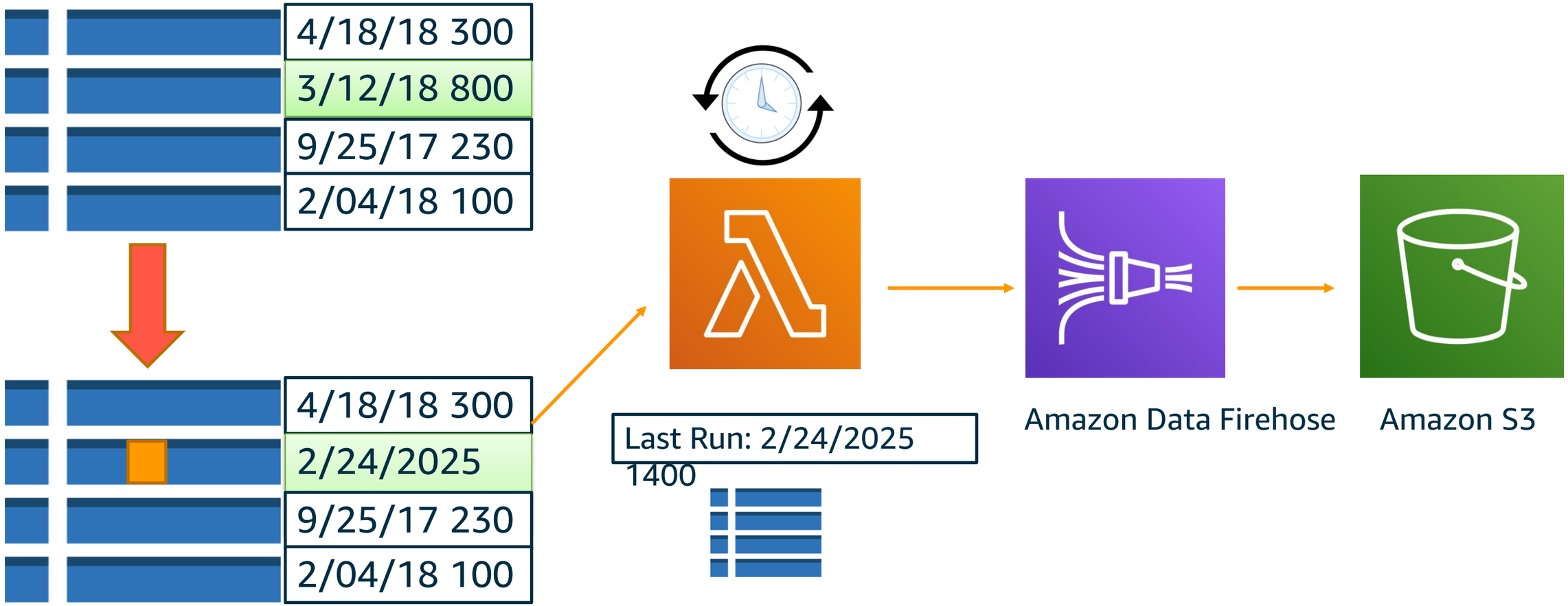
Amazon S3

Change Data Capture

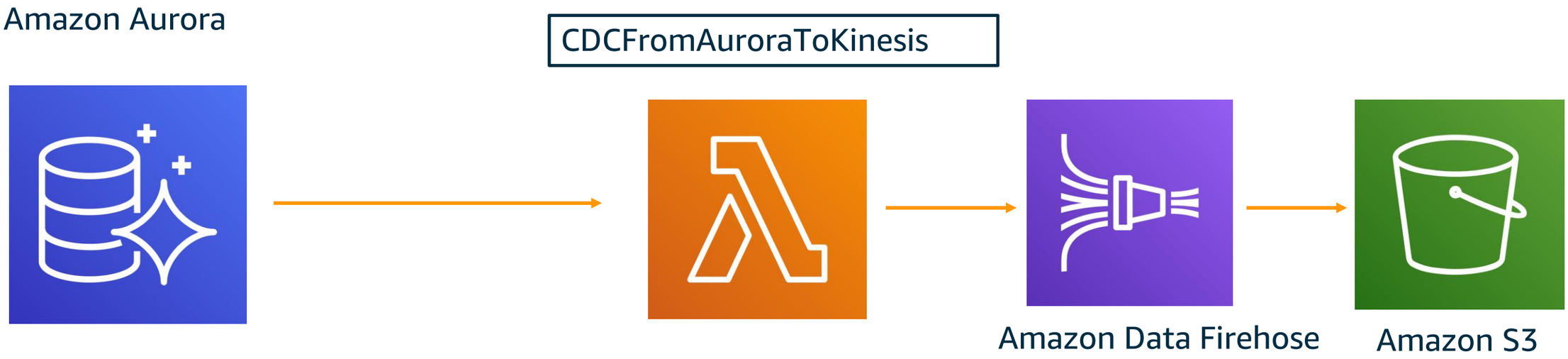
Techniques to Capture Changes

- Timestamp
- Diff Comparison
- Triggers
- Transaction Log

Change Data Capture (CDC)– Timestamp using Lambda



Change Data Capture – Triggers v2




```
CREATE PROCEDURE CDC_TO_FIREHOSE […]
      CALL mysql.lambda_async('arn:aws:lambda:us-east-
      1:XXXXXXXXXXXXX:function:
CREATE TRIGGER TR_Sales_CDC AFTER INSERT ON Sales […]
      CALL CDC_TO_FIREHOSE
```

CDCFromAuroraToKinesis

CDC with DMS Approach

AWS Database Migration Service (AWS DMS)
securely migrate **and/or replicate your databases** *and* data warehouses to AWS

AWS Schema Conversion Tool (AWS SCT)
commercial database and data warehouse schemas to open-source engines **or AWS-native services**, such as Amazon Aurora and Redshift

When to use DMS and SCT?

Modernize

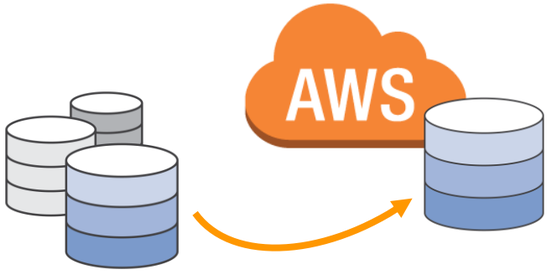
Modernize your database tier –

- Commercial to open-source
- Commercial to Amazon Aurora

Modernize your Data Warehouse –

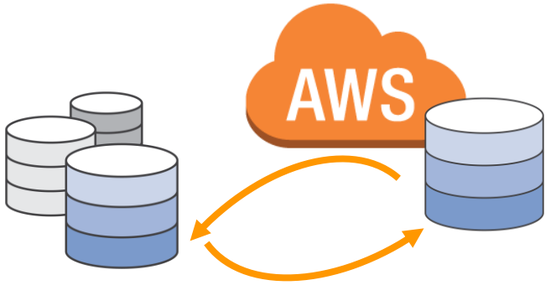
- Commercial to Redshift

Migrate



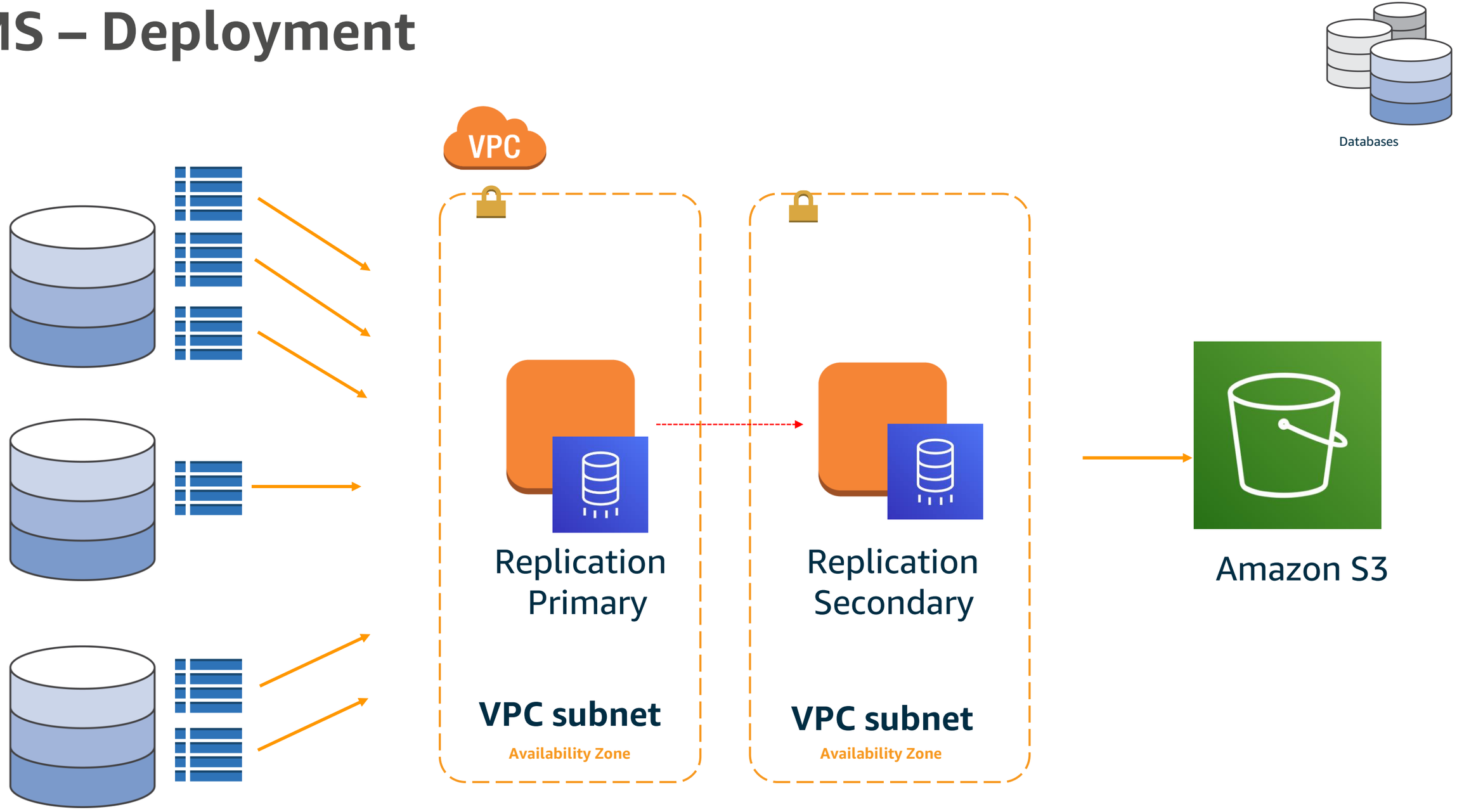
- Migrate business-critical applications
- Migrate from Classic to VPC
- Migrate data warehouse to Redshift
- Upgrade to a minor version

Replicate



- Create cross-regions Read Replicas
- **Run your analytics in the cloud**
- Keep your dev/test and production environment sync

DMS – Deployment



DMS – S3 as a Target

Bulk dump File

```
s3://mybucket/schemaName/tableName
s3://mybucket/hr/employee

/schemaName/tableName/LOAD001.csv
/schemaName/tableName/LOAD002.csv
/schemaName/tableName/LOAD003.csv
...
101,Smith,Bob,4-Jun-14,New York
102,Smith,Bob,8-Oct-15,Los Angeles
103,Smith,Bob,13-Mar-17,Dallas
104,Smith,Bob,13-Mar-17,Dallas
```

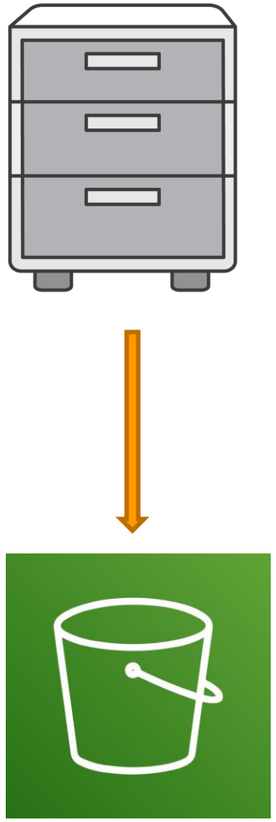
Ongoing CDC Files

```
s3://mybucket/schemaName/tableName

<time-stamp>.csv
<time-stamp>.csv
<time-stamp>.csv
...

I,101,Smith,Bob,4-Jun-14,New York
U,101,Smith,Bob,8-Oct-15,Los Angeles
U,101,Smith,Bob,13-Mar-17,Dallas
D,101,Smith,Bob,13-Mar-17,Dallas
```

Transfer Files to S3



Optimizing Transfers

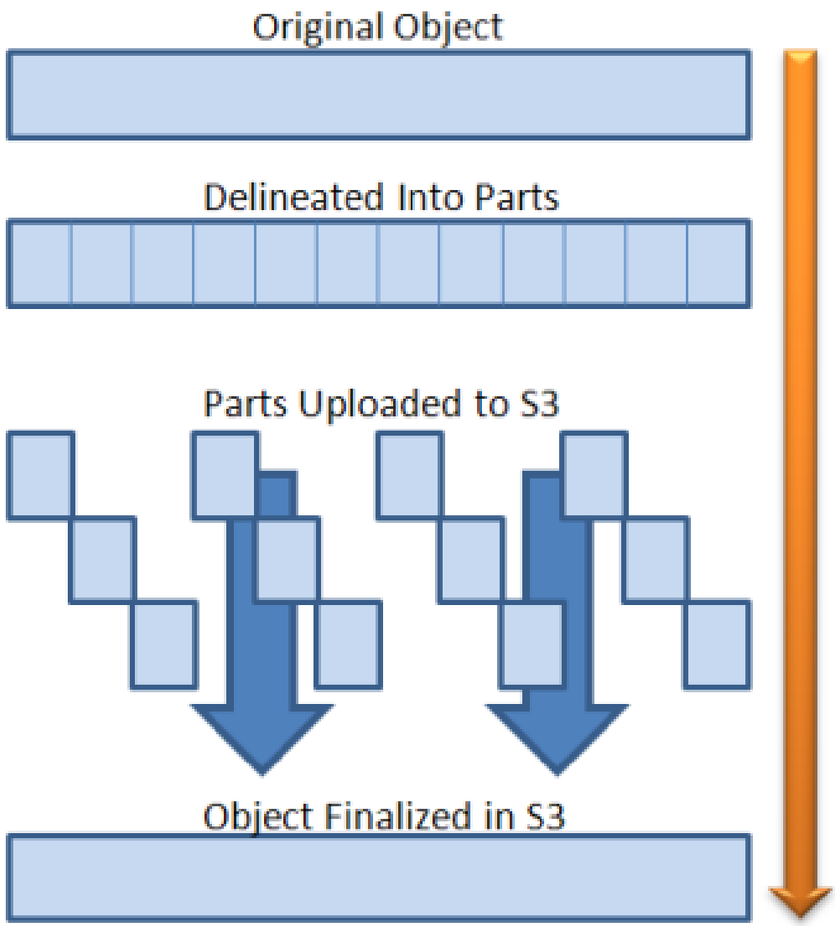
- S3 Multi-Part Upload
- S3 Transfer Acceleration
- AWS Direct Connect

Available Services

- AWS DataSync
- AWS Transfer – SFTP
- AWS Snowball/Snowmobile

Uploading to Amazon S3

- Amazon S3 supports both a single-part upload and a multi-part upload API
- The single-part upload supports objects up to 5 GB in size
- The multi-part upload supports objects **up to 5 TB** in size
- The multi-part upload also enables you to maximize your throughput by using parallel threads
- (!) Cleanup uploaded multi-part chunks in S3



S3 Transfer Acceleration

PUT requests go through the nearest AWS Edge Location

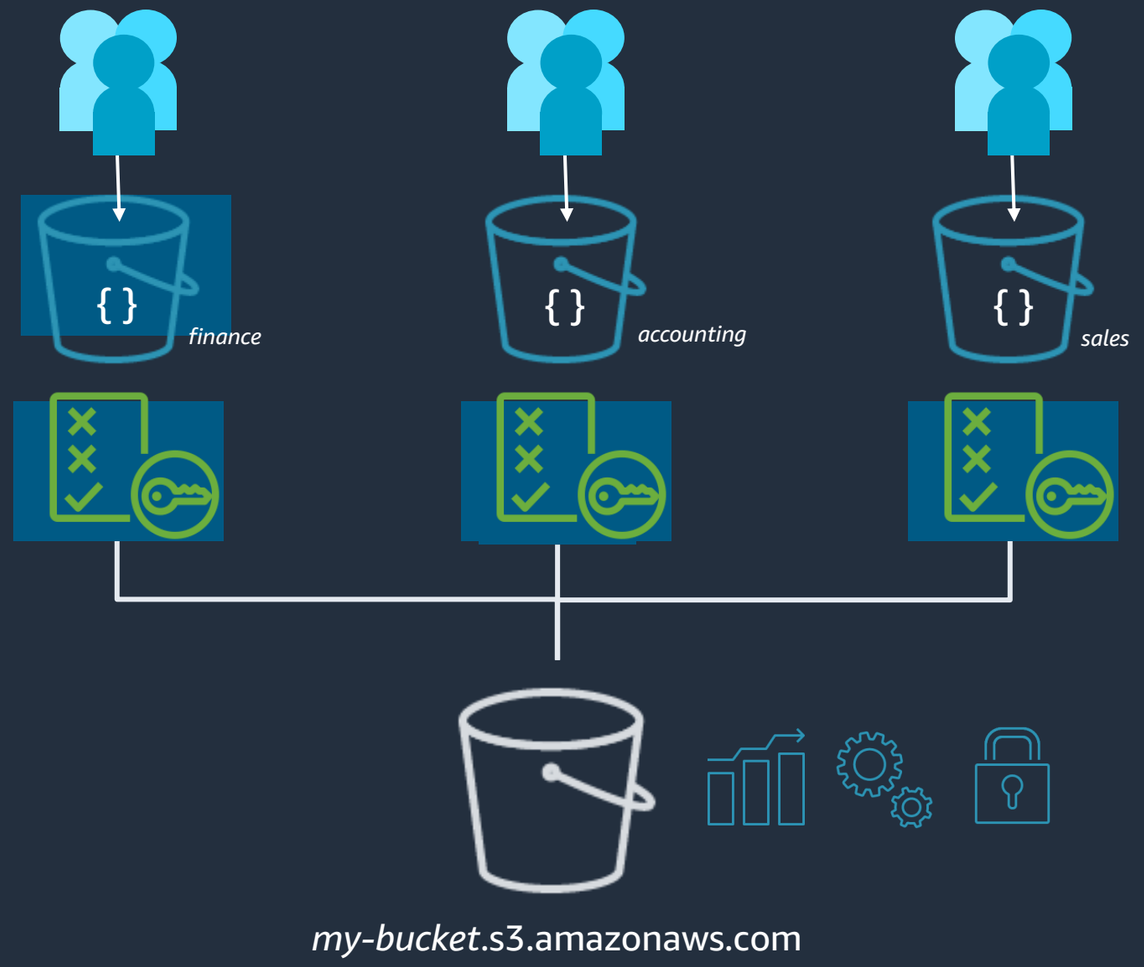
Data transits over the AWS private network rather than Internet

AWS private network optimizes throughput and latency to the AWS Region

Data is not stored in the edge cache

S3 Access Points for shared Data sets

How Access Points Work



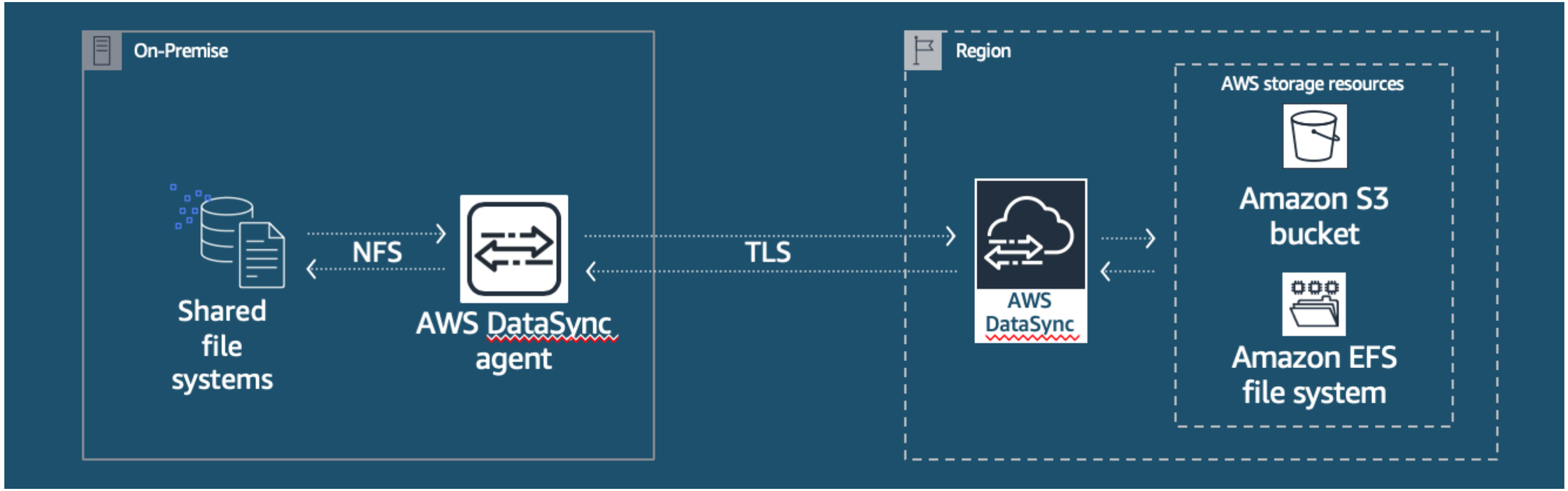
Segment Clients into Distinct Groups
Useful in multi-tenant & data lake environments.

Each group gets their own Access Point
These "bucket names" can be used by clients just as they are today.

Apply a policy to each Access Point
Tailor permissions & access based on who's using the Access Point.

Maintain Centralized Control Over Storage
Many Access Points, but one set of policies for storage management.

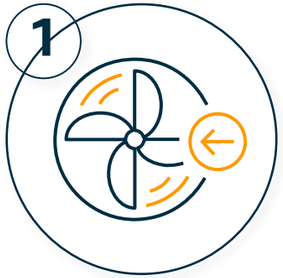
AWS DataSync



Deploy on-premises agent for fast access to local storage → Data transfer over the WAN using purpose-built protocol → Service in AWS writes or reads data from AWS storage services → Managed from AWS Console or Command Line Interface (CLI)

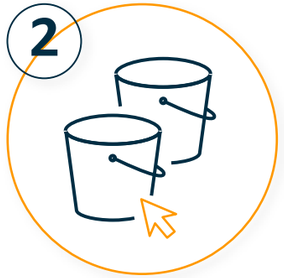
Setting up AWS Transfer

Map your hostname



Associate your hostname with the server endpoint

Select your S3 bucket(s)



Create an IAM role to access the Amazon S3 buckets used for storing data transferred over SFTP

Set up your users

Create and map users to IAM roles to enable them for file operations

Your users can now use your AWS SFTP server endpoint to transfer data

Access Analyzer for S3

Quickly analyze resource policies across your entire AWS organization

Analyzes thousands of policies in seconds for public or cross-account access

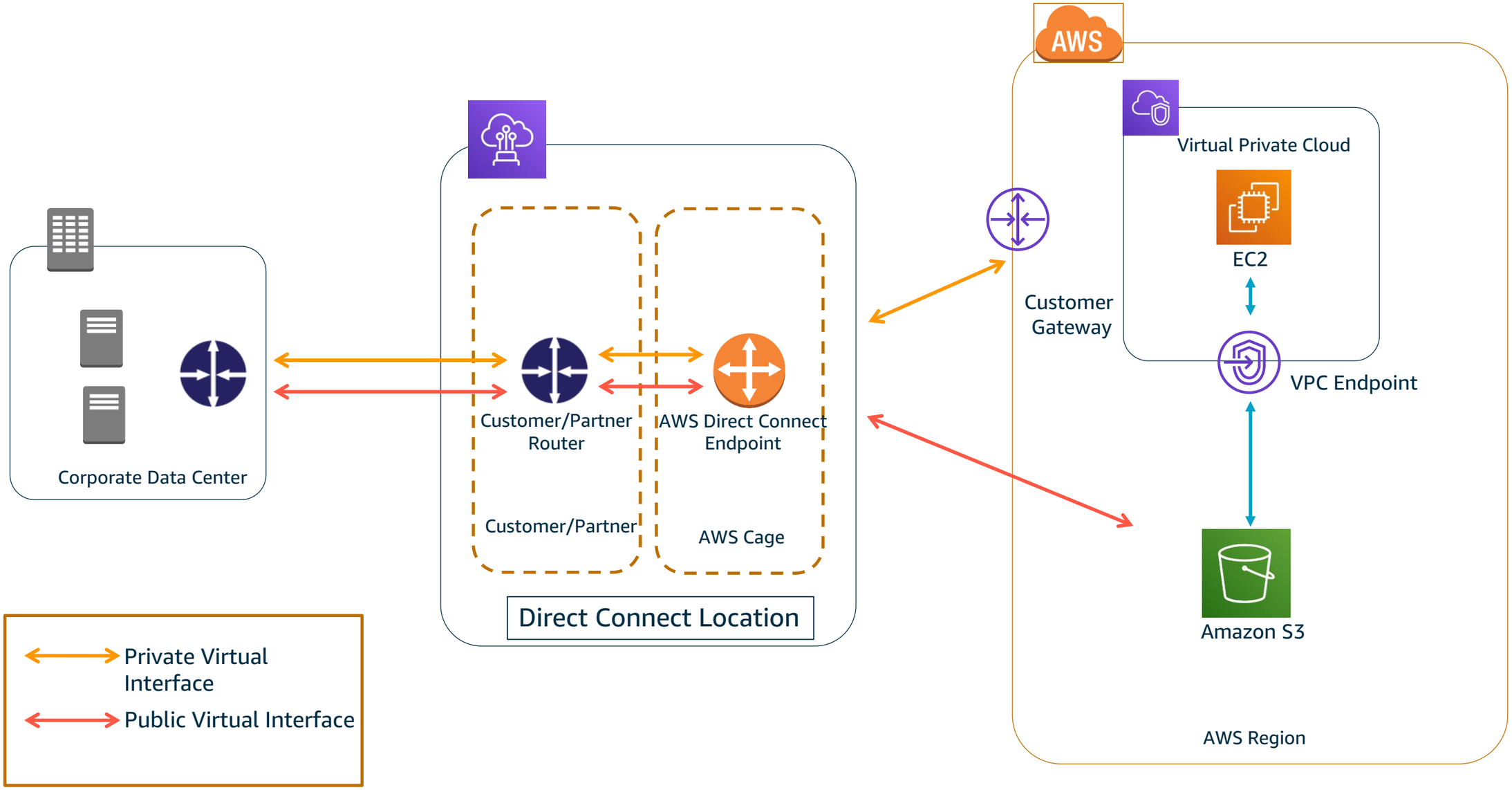
Continuously monitor and analyze permissions

Continuously monitors and automatically analyzes any new or updated resource policy to help you understand potential security implications

Provides the highest levels of security assurance

Uses automated reasoning, a form of mathematical logic & inference, to determine all possible access paths allowed by a resource policy

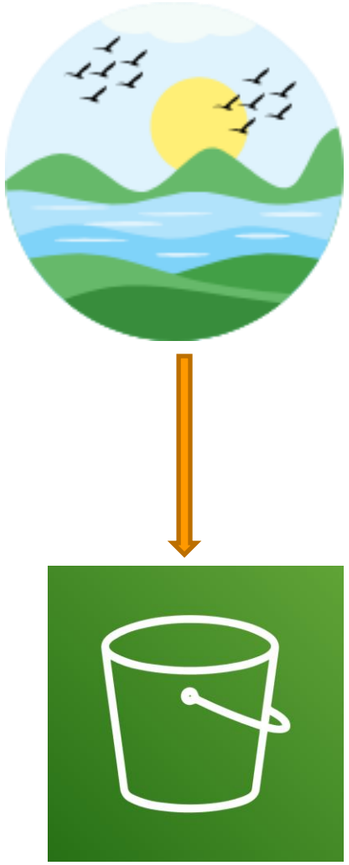
AWS Direct Connect



AWS Snow Family

Use Case	AWS Solution
Cloud Migration, Disaster Recovery	AWS Snowball Edge storage optimized (80 TB HDD, 40 vCPUs, 80GB memory, 1TB SSD) 
Internet of Things (IoT), Remote Locations	AWS Snowball Edge Compute Optimized (104 vCPUs, 416 GB of memory, and 28 TB of dedicated NVMe SSD for compute instances) 
Migrating Exabytes of Data	AWS Snowmobile (100 PB HDD) 
Backpacks on first responders for IoT, vehicular, and drone	AWS Snowcone (8 TB HDD, 4 vCPUs, 4 GB memory, 14 TB SSD) 

Streams



Collecting and Analyzing

- Amazon Kinesis
- Amazon Managed Streaming for Kafka (MSK)
- Example: Clickstream Analytics

Stream Ingestion

Streams

Data from tens of thousands of data sources can be written to a single stream

AWS Toolkits/Libraries

- AWS SDK
- Kinesis Producer Library
- AWS Mobile SDK
- Kinesis Agent

AWS Service Integrations

- AWS IoT
- Amazon CloudWatch Logs
- Amazon CloudWatch Events
- Amazon Database Migration Service (DMS)*

3rd Party Offerings

- LOG4J
- Flume
- Fluentd

* Amazon DMS includes 8 on-premise databases, 1 Azure database, 5 RDS/Aurora database types, and S3

Real-time Streaming on AWS

Streams

Easily collect, process, and analyze video and data streams in real time

Kinesis Data Streams

Collect and store data streams for analytics

Amazon Data Firehose

Load data streams into AWS data stores

Amazon Managed Service for Apache Flink

Analyze data streams with SQL or Java

Amazon Managed Streaming for Kafka

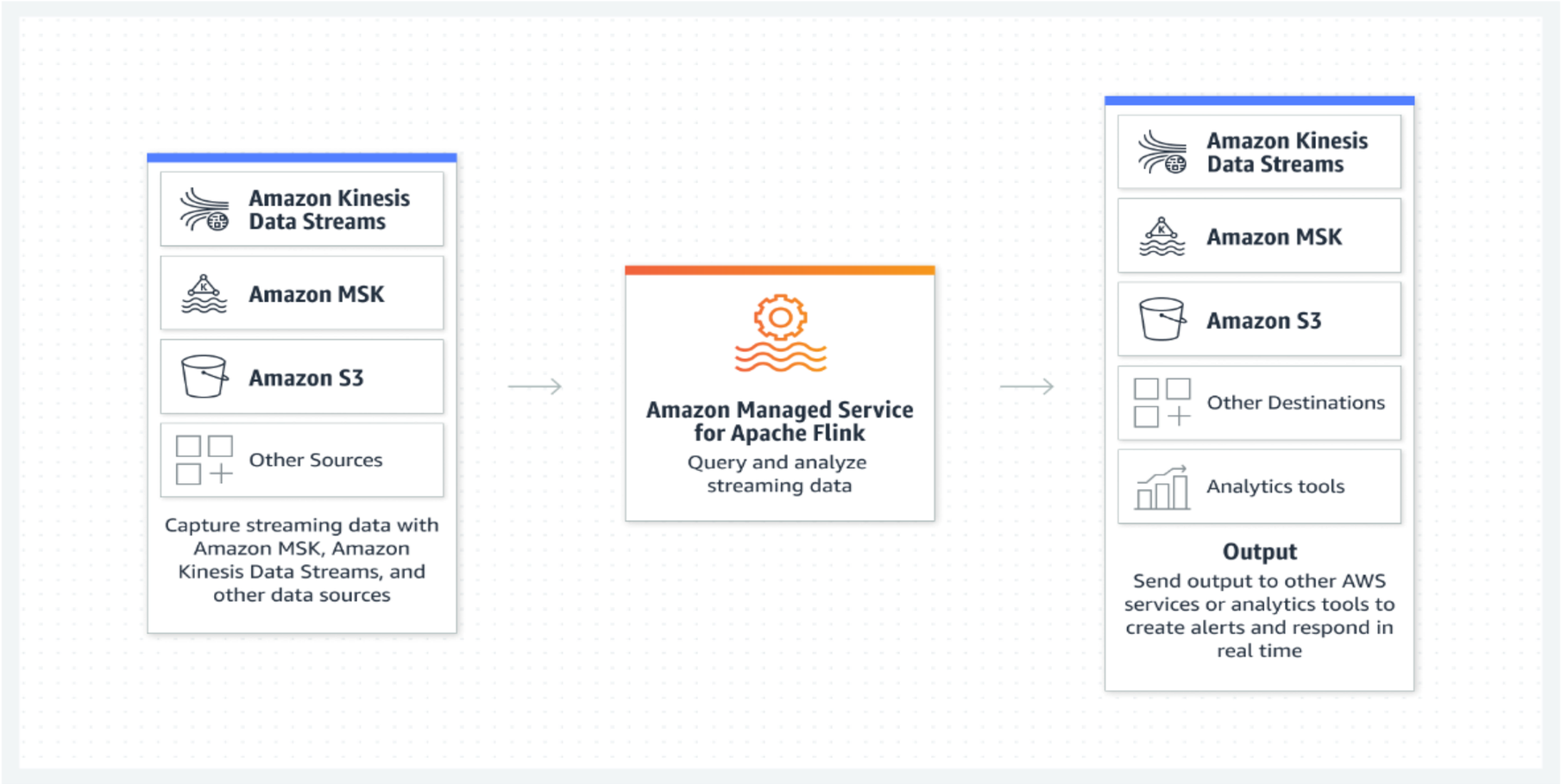
Collect and store data streams for analytics

Kinesis Video Streams

Capture and store video streams for analytics

Amazon Managed Service for Apache Flink

Streams



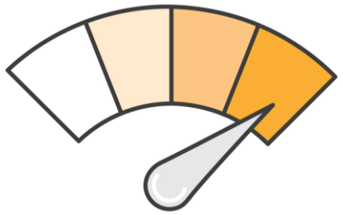
- Interact with streaming data in real-time using SQL or integrated Java applications
- Build fully managed and elastic stream processing applications

KDA Java for sophisticated applications

Streams

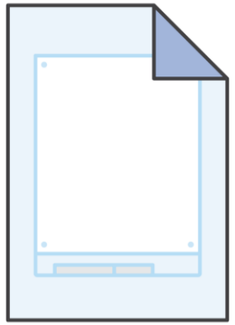
Simple programming

Easy to use and flexible APIs make building apps fast



High performance

In-memory computing provides low latency & high throughput



Stateful Processing

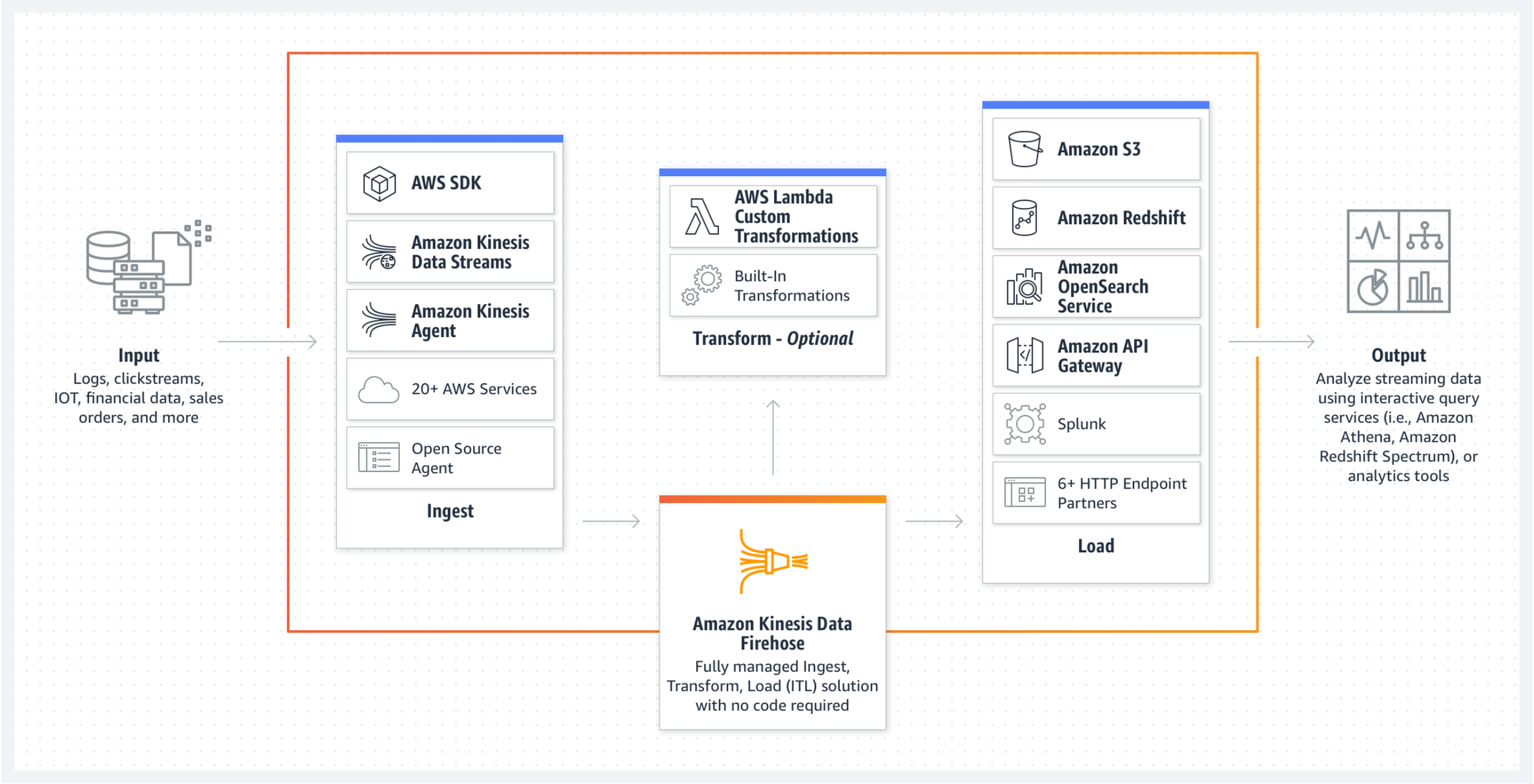
Durable application state saves

Strong data integrity

Exactly-once processing and consistent state

Amazon Data Firehose

Streams



- Zero administration and seamless elasticity
- Direct-to-data store integration
- Serverless continuous data transformations
- Near real-time
- Data format conversion to Parquet/ ORC

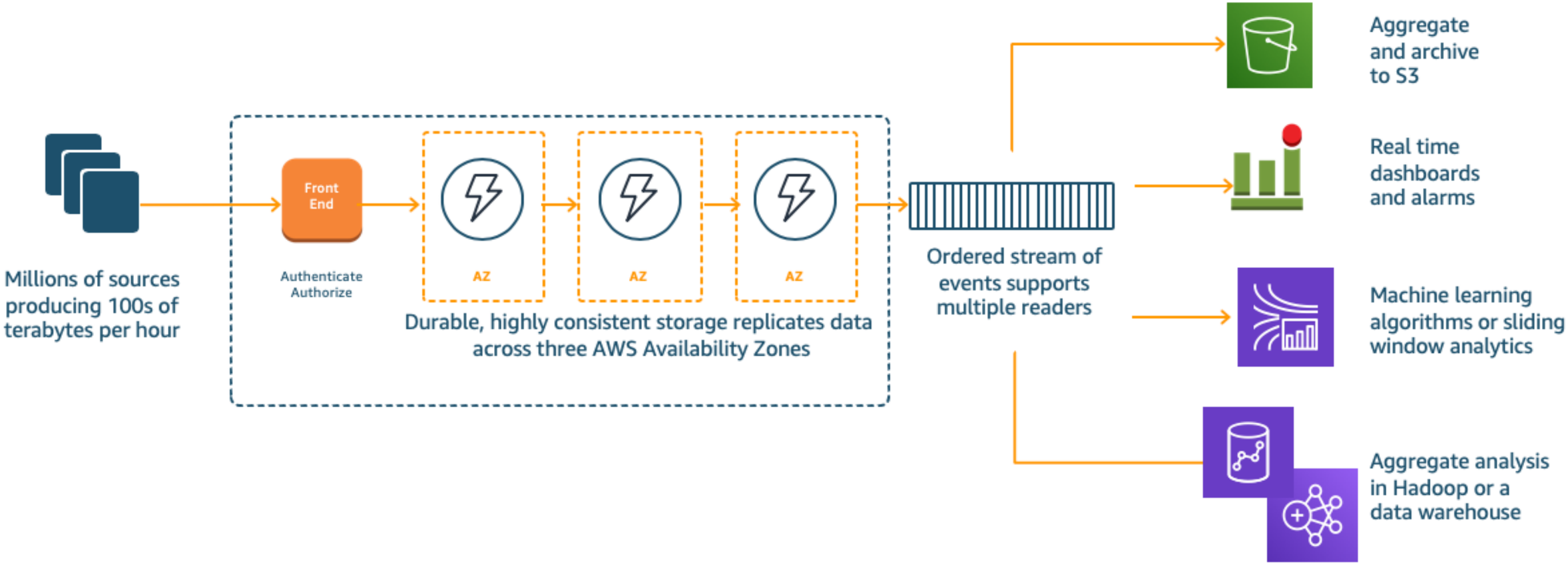
Amazon Kinesis – Streams vs Firehose

Streams

Amazon Kinesis Data Streams is for use cases that require custom processing, per incoming record, with sub-1 second processing latency, and a choice of stream processing frameworks

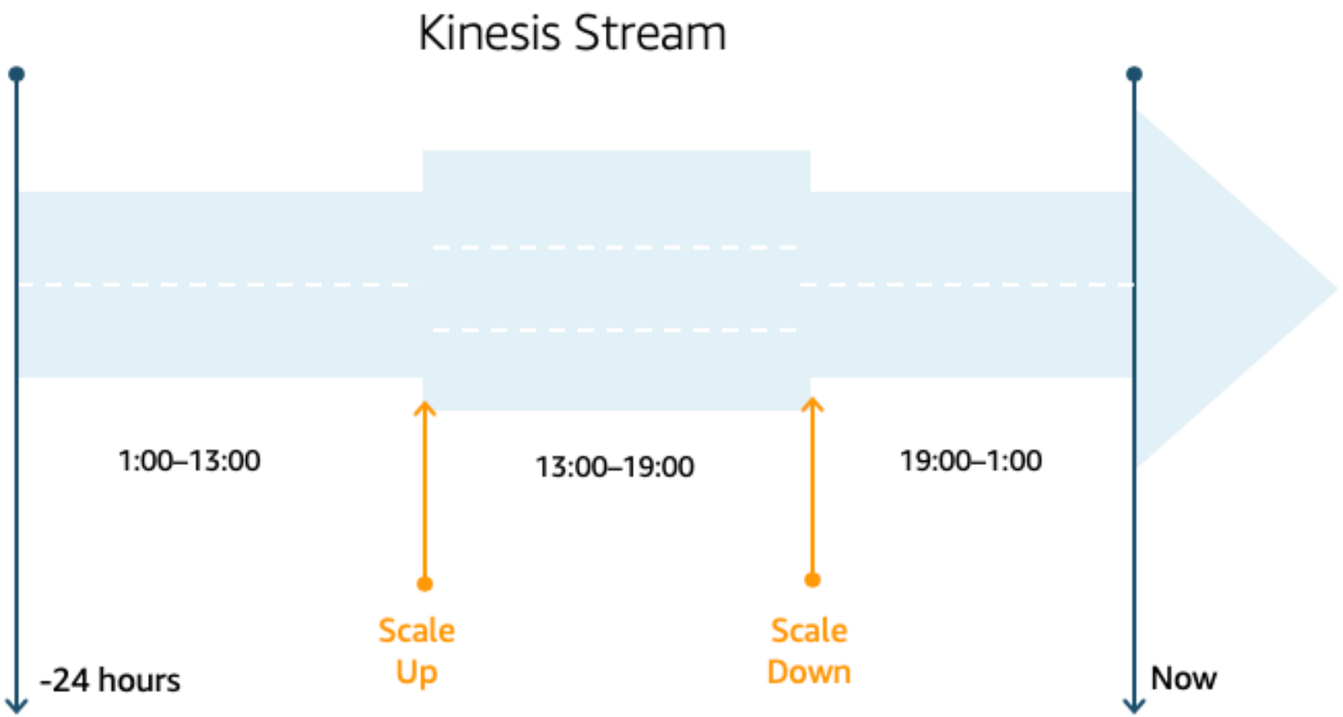
Amazon Data Firehose is for use cases that require zero administration, ability to use existing analytics tools based on Amazon S3, Amazon Redshift, and Amazon ES, and supports zero buffering

Kinesis Data Streams – How it works



Kinesis Data Streams – How it works cont.

Streams



- Data streams are made of Shards
- Each Shard ingests data up to 1MB/sec, and up to 1000 TPS
- Each Shard emits up to 2 MB/sec
- All data is stored for 24 hours – 7 days
- Scale Kinesis data streams by splitting or merging Shards
- Replay data inside of 24 hours – 7 days window

***Note**: You can raise data retention period to up to 7 days by enabling extended data retention or up to 365 by enabling long-term data retention using the console, the CLI or the API call.*

Amazon Kinesis Data Streams On-Demand

Streams

Simple to use
Simplify streaming data processing by eliminating capacity management

Flexible scaling
Automatically scale capacity in response to changing data volumes

Automated high availability
Provide built-in availability and fault tolerance by default

Lower your costs
Pay per gigabyte of data written, read, and stored

Amazon Managed Streaming for Kafka (MSK)

Streams

- Fully compatible with Apache Kafka v3.x (and some other 2.x versions)
- AWS Management Console and AWS API for provisioning
- Clusters are setup automatically
- Provision Apache Kafka brokers and storage
- Create and tear down clusters on-demand
- Cruise Control to more easily scale, partition management, and balance I/O

Amazon MSK Serverless

Streams

Easily run Apache Kafka clusters without rightsizing cluster capacity

Instantly scale I/O without worrying about scaling capacity up and down or reassigning partitions

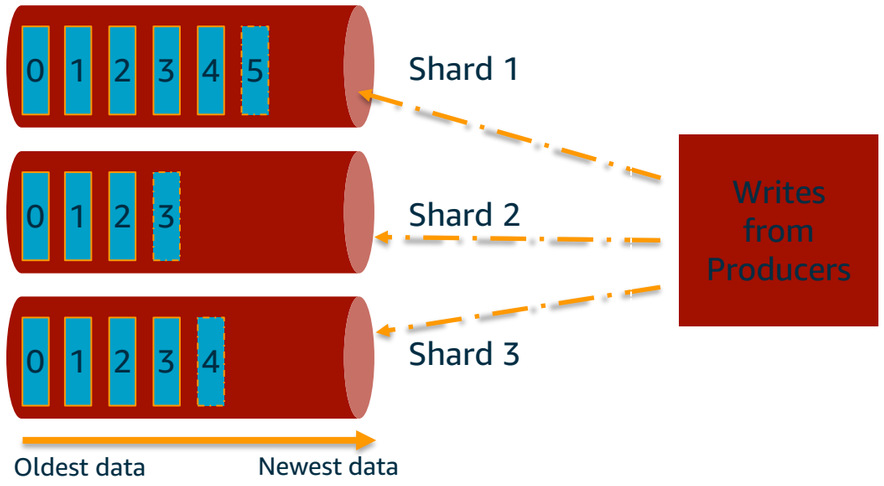
Pay for the data volume you stream and retain

Comparing Amazon Kinesis Data Streams to MSK

Streams

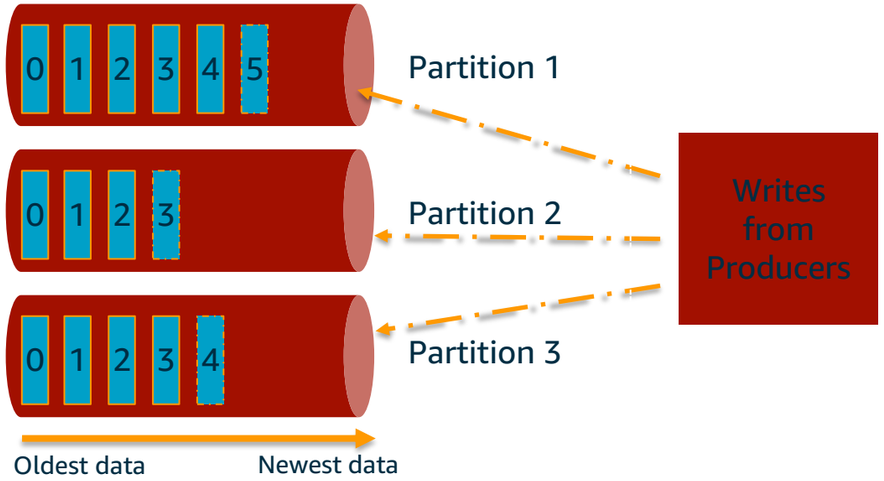
Amazon Kinesis Data Streams

Stream with 3 shards



Amazon MSK

Topic with 3 partitions



Comparing Amazon Kinesis Data Streams to MSK

Streams

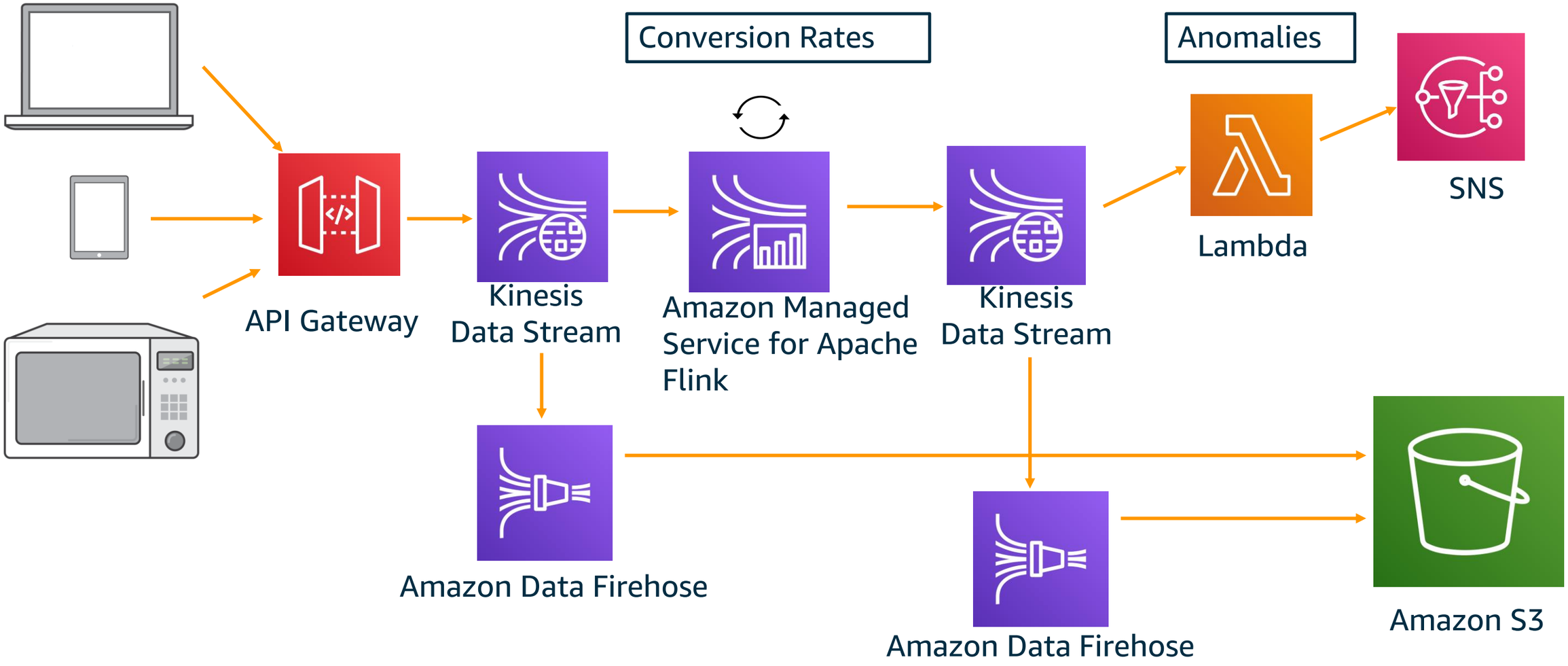
Amazon Kinesis Data Streams

- AWS API experience
- Throughput provisioning model
- Seamless scaling
- Typically lower costs
- Deep AWS integrations

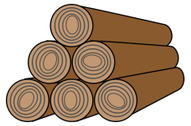
Amazon MSK

- Open-source compatibility
- Strong third-party tooling
- Cluster provisioning model
- Apache Kafka scaling isn’t seamless to clients
- Raw performance

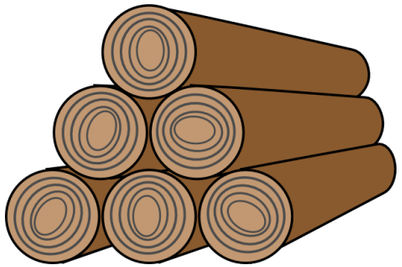
Clickstream with Real-Time Analytics



Logs



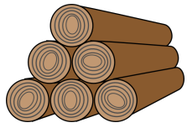
Logs



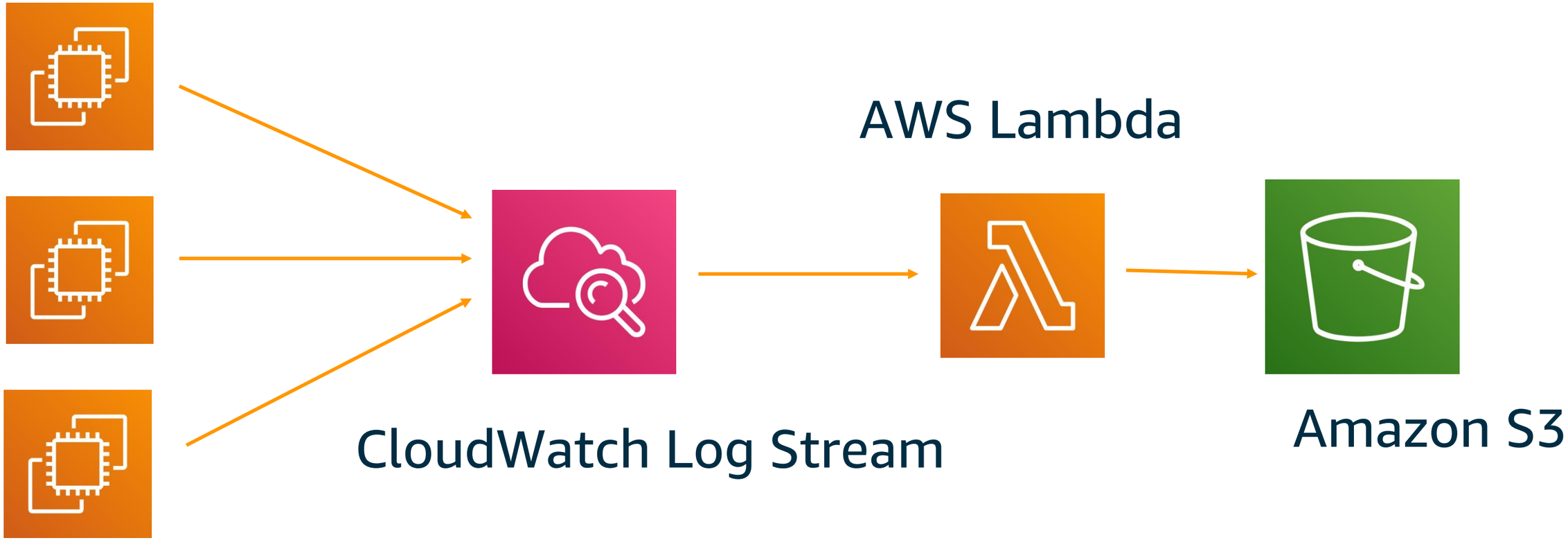
Collecting and Analyzing

- Amazon CloudWatch
- Amazon Kinesis
- Other Options

Logs – CloudWatch Agent

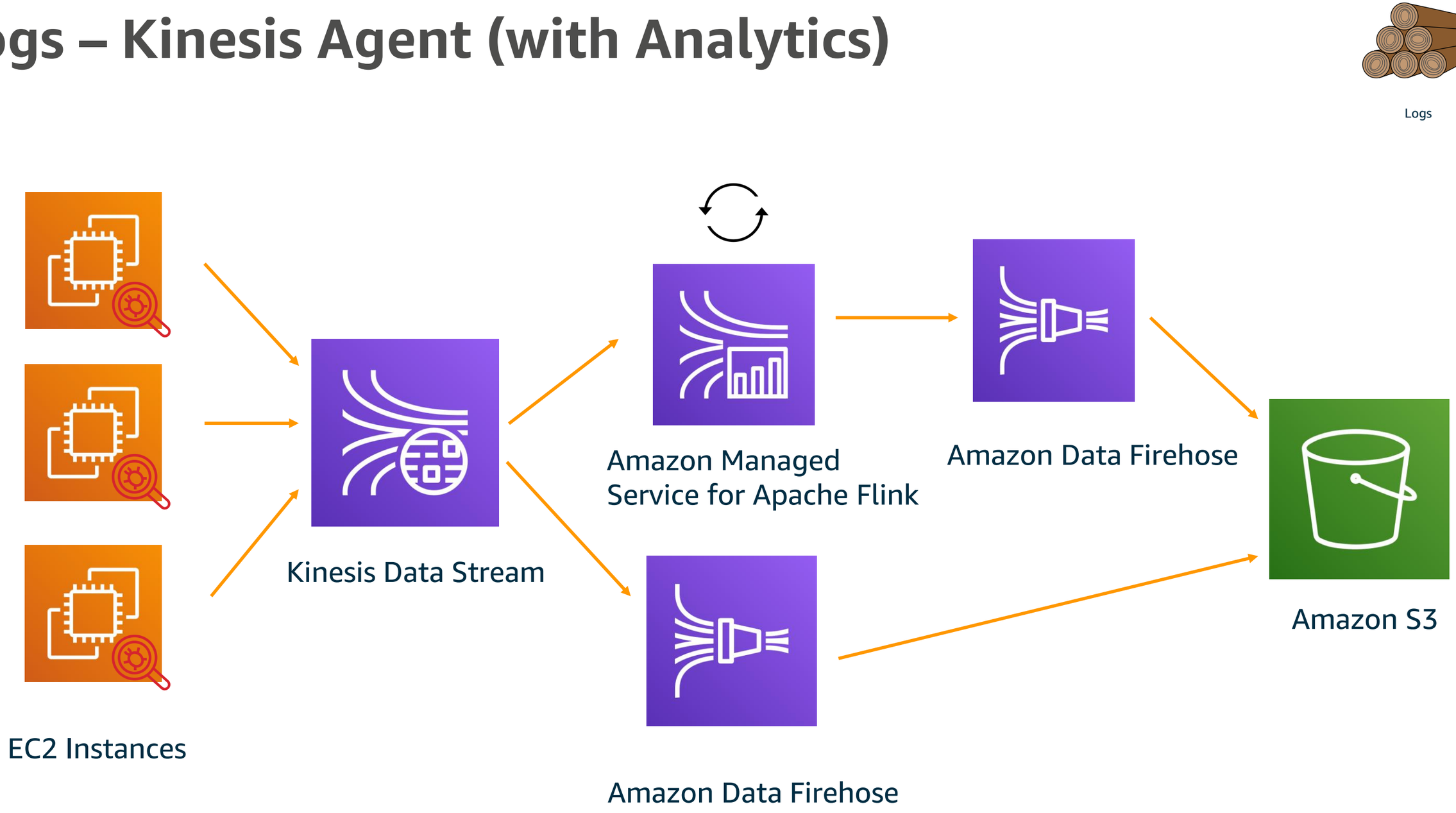


Logs

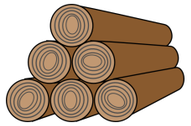


EC2 Instances

Logs – Kinesis Agent (with Analytics)



Logs - Other Options



Logs

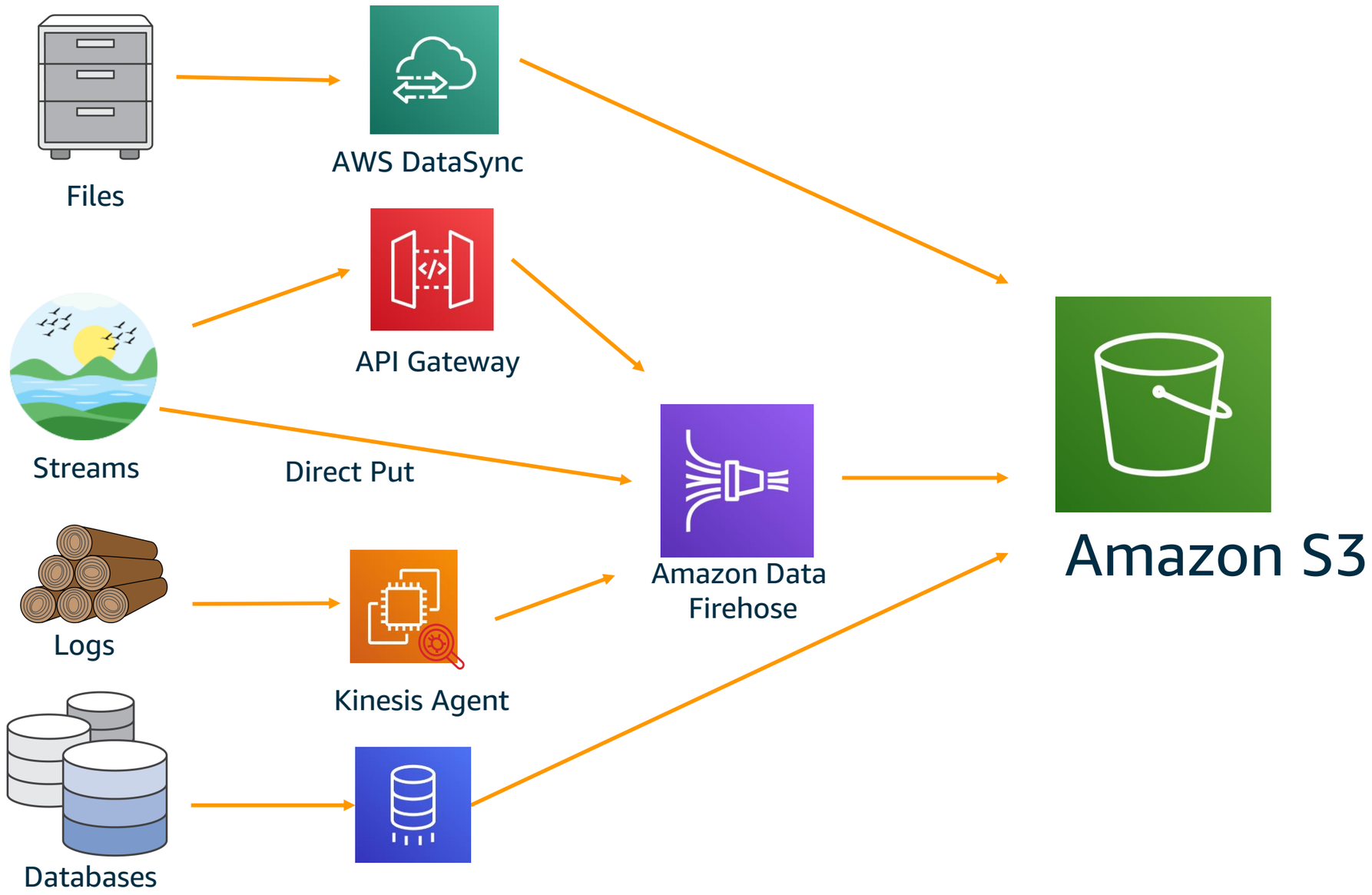
Use OSS agents, like Flume or Fluentd

- Pre-batch PUTS for better efficiency
- See https://github.com/awslabs/aws-fluent-plugin-kinesis

Make a tweak to your existing logging

- log4j appender option
- See https://github.com/awslabs/kinesis-log4j-appender

Summary – Ingestion into S3 (Data Lake)

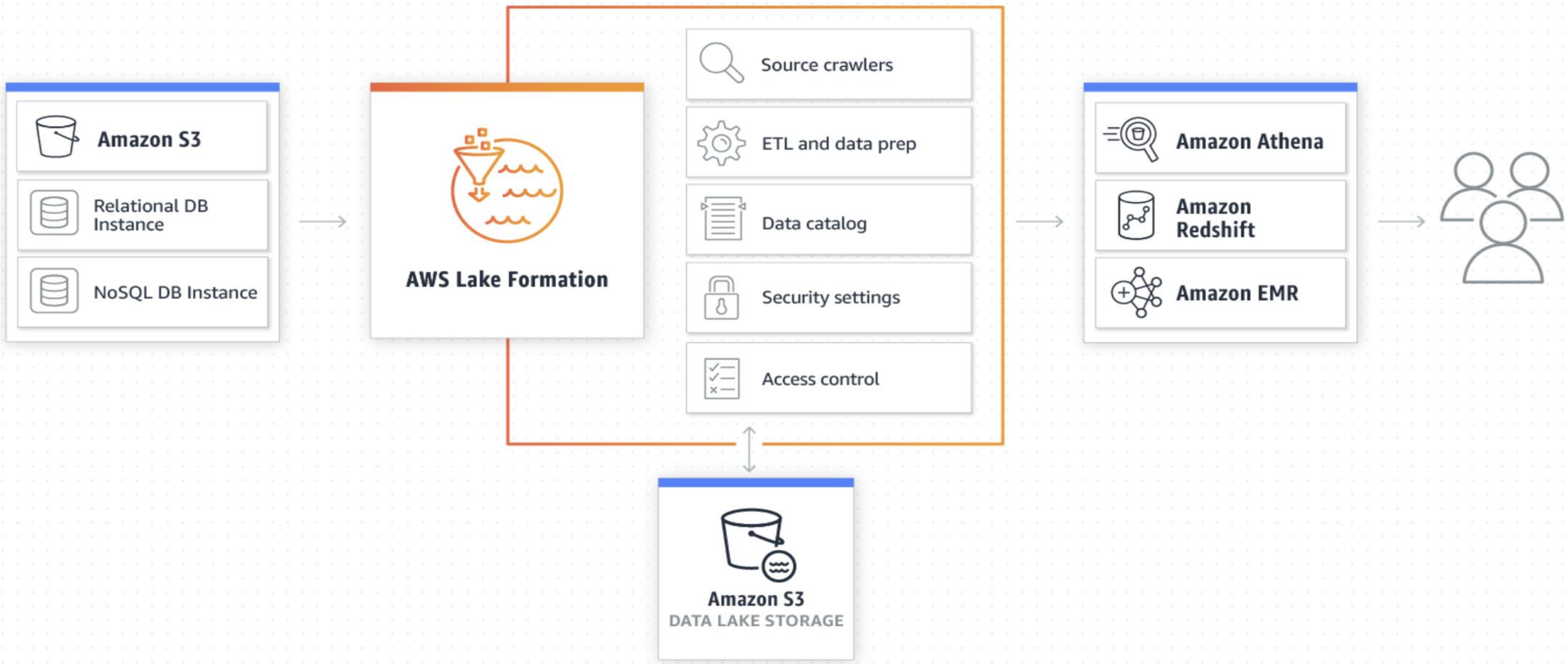


s3://datalake/

/vendorfeeds
/vendorA
/vendorB
/clickstream
/orders
/vendors
/customers
/app_logs
/instance1
/instance2
/syslogs
/instance1
/instance2
/databases
/customers
/orders
/vendors

Ingesting data using AWS Lake Formation and AWS Glue

AWS Lake Formation



With blueprints

You

1. Point us to the source
2. Tell us the location to load to in your data lake
3. Specify how often you want to load the data

Blueprints

1. Discover the source table(s) schema
2. Automatically convert to the target data format
3. Automatically partition the data based on the partitioning schema
4. Keep track of data that was already processed
5. You can customize any of the above

Data Ingestion with Glue

Options for data transfer

AWS Direct Connect

Amazon Data Firehose

Amazon Kinesis Data Streams

Amazon Kinesis Video Streams

Amazon S3 Transfer Acceleration

AWS Storage Gateway

Amazon Appflow

AWS Snowcone

AWS Snowball Edge

AWS Snowmobile

AWS DataSync

AWS Transfer for SFTP

Thank You!

Paige Broderick
jbropaig@amazon.com