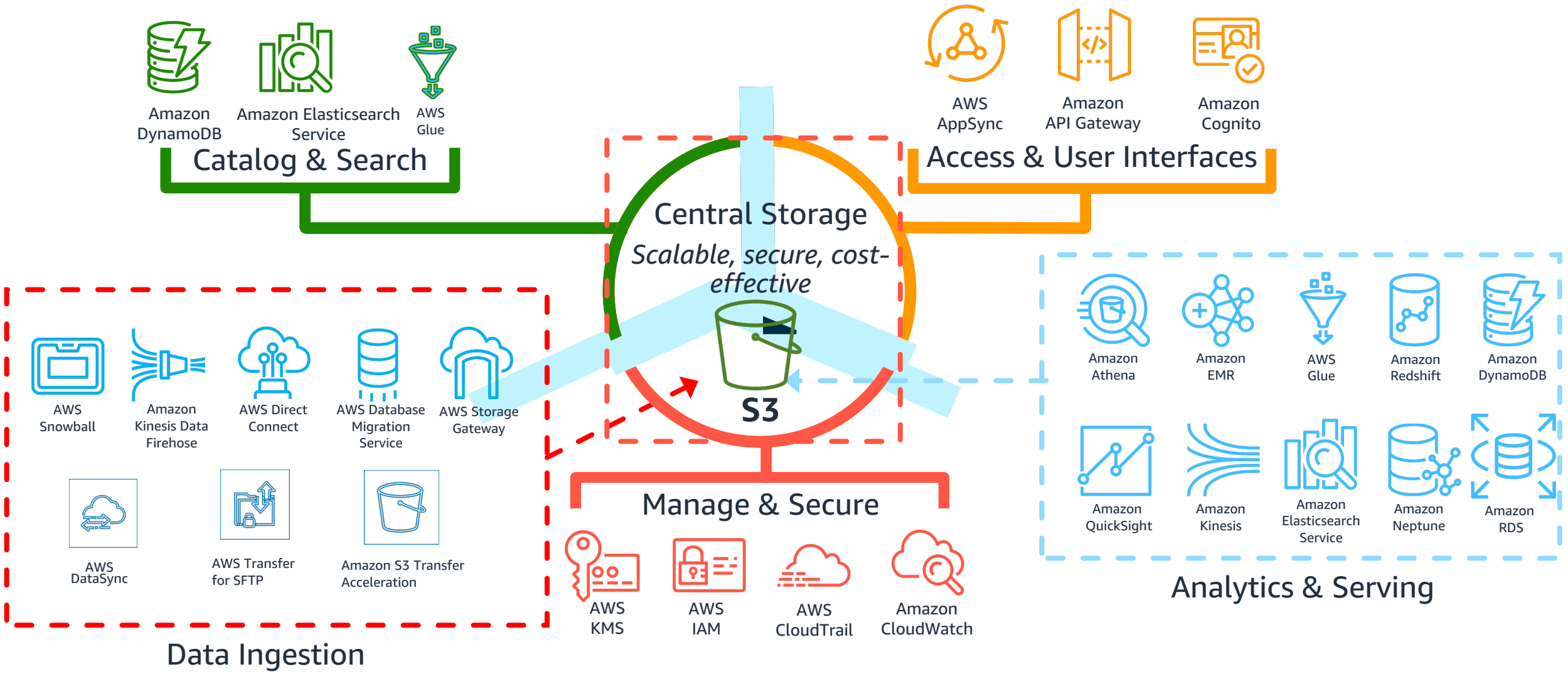
Hydrating the Data Lake

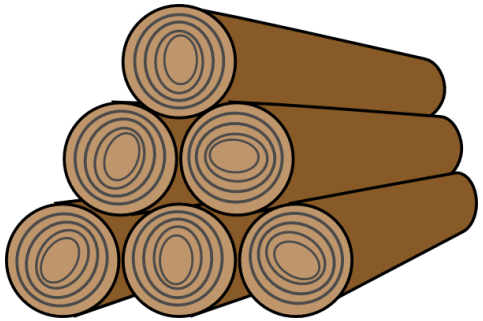
Session's Focus – Working In The Data Lake



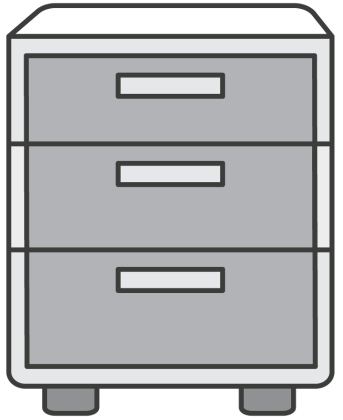
Data Sources

Databases

Streams



Logs



Files

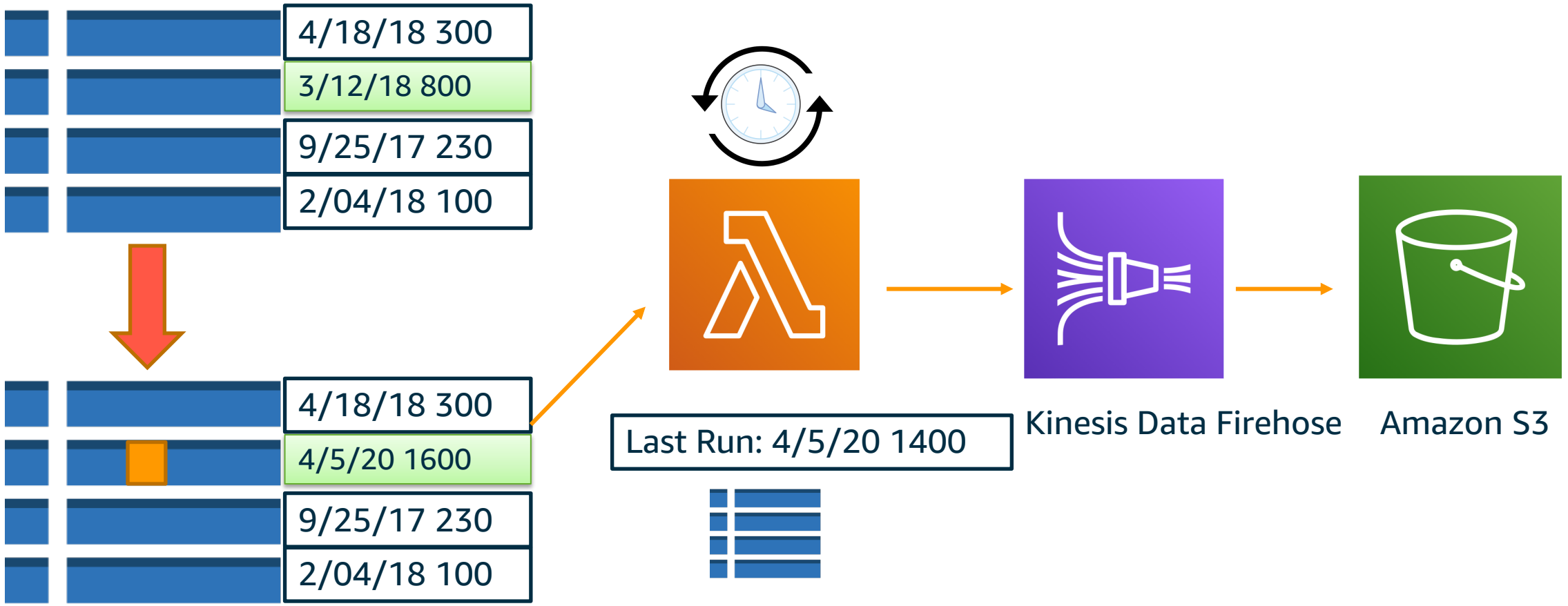
Data Sources - Databases

Change Data Capture

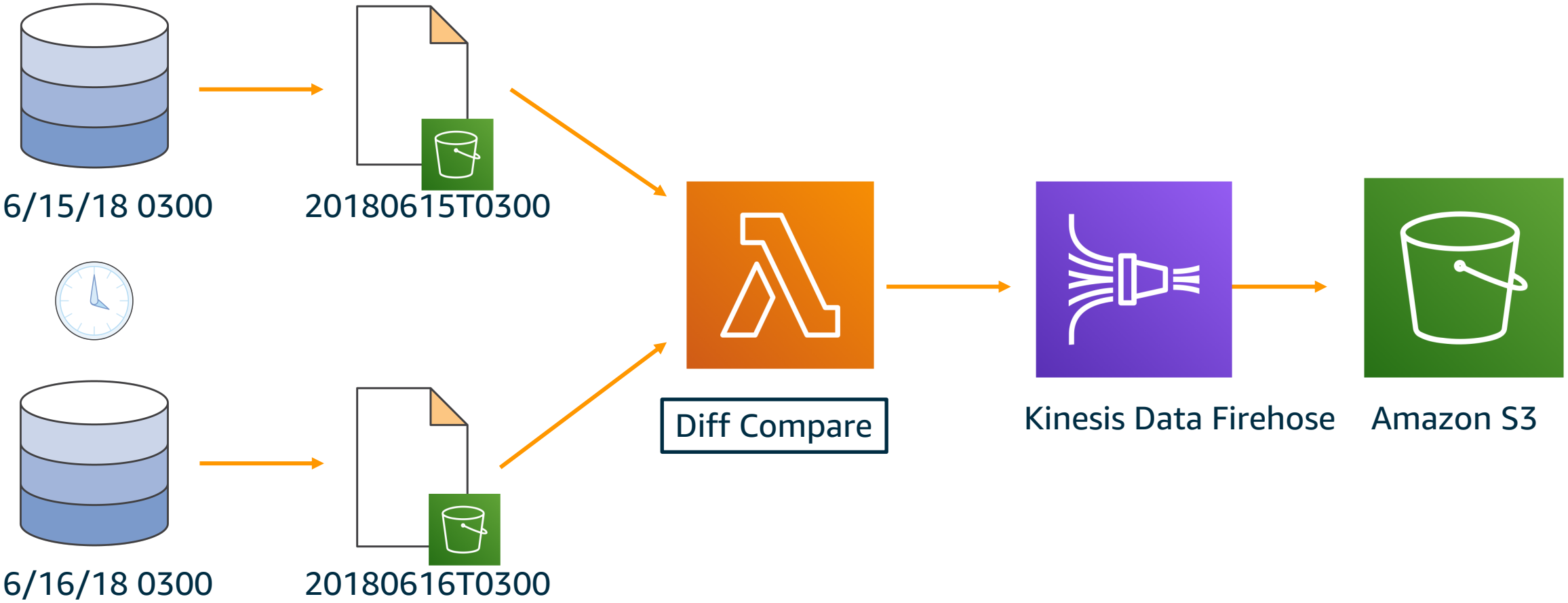
Techniques to Capture Changes

- Timestamp
- Diff Comparison
- Triggers
- Transaction Log

Change Data Capture – Timestamp

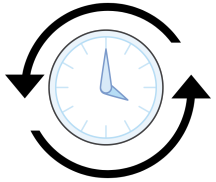


Change Data Capture – Diff Compare



Change Data Capture – Triggers

SELECT
Id: 20982358
Name: Jean-Luc Picard
Rank: Captain
State: Agitated



Write operations to Firehose

Kinesis Data Firehose

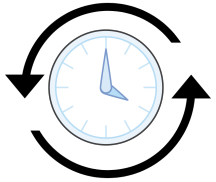
Amazon S3

Table: Roster
Id: 20982358
Operation: Update
Job: **ag8afh8**

SELECT
Table: Roster
Id: 20982358
Operation: Update

Change Data Capture – Triggers

Id: 20982358
Name: Jean-Luc Picard
Rank: Captain
State: Agitated



Delete Job “ag8afh8”

DELETE
Table: Roster
Id: 20982358
Operation: Update
Job: **ag8afh8**

DELETE
Table: Roster
Id: 20982358
Operation: Updat

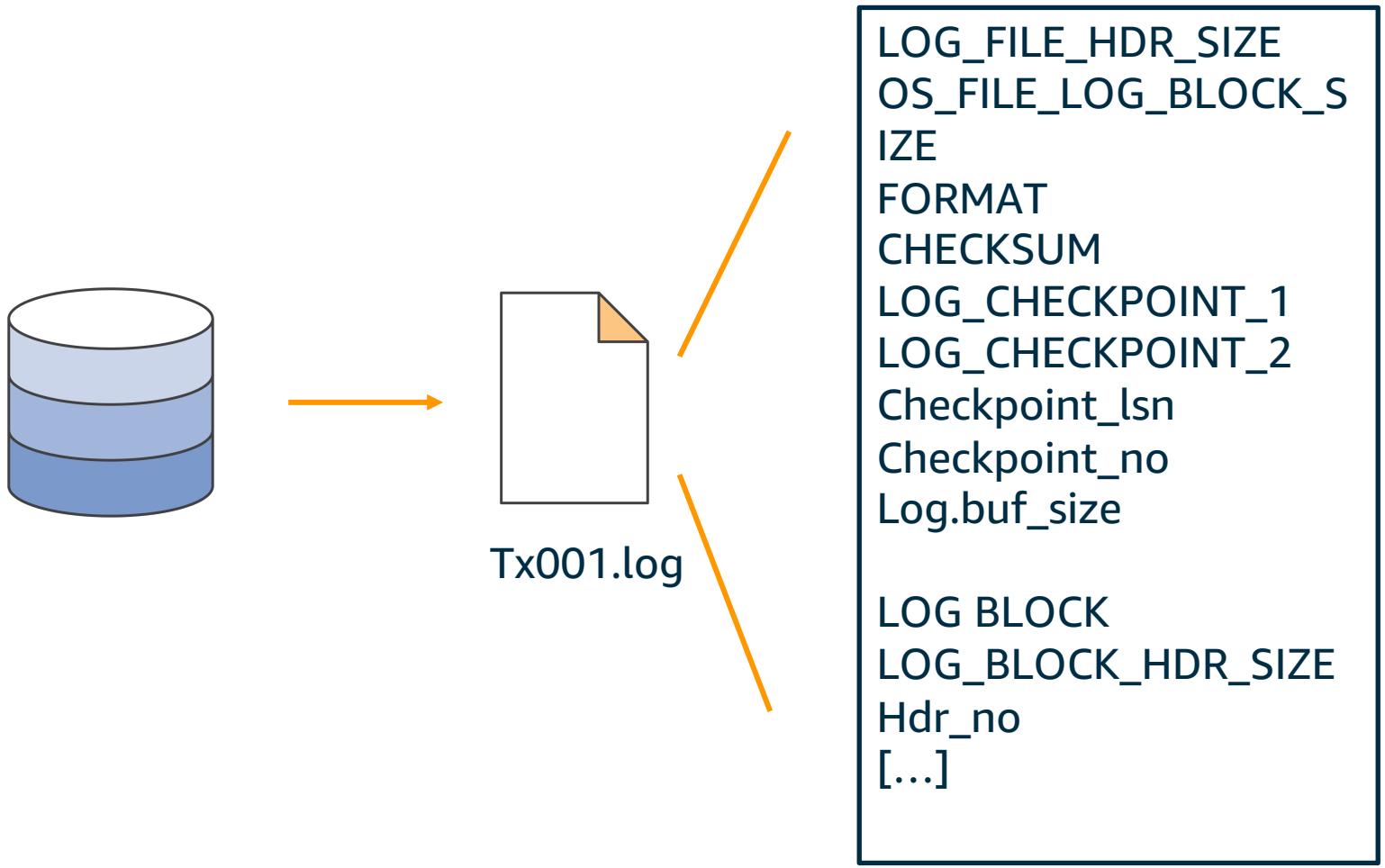
Change Data Capture – Triggers v2

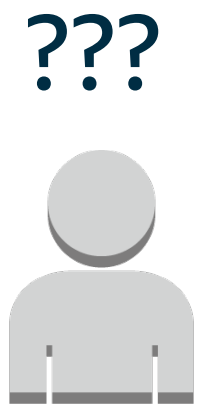

```
CREATE PROCEDURE CDC_TO_FIREHOSE […]
      CALL mysql.lambda_async('arn:aws:lambda:us-east-1:XXXXXXXXXXXX:function:CDCFromAuroraToKinesis

CREATE TRIGGER TR_Sales_CDC AFTER INSERT ON Sales […]
      CALL CDC_TO_FIREHOSE
```

Change Data Capture – Triggers v2

Change Data Capture – Database Logs





Database Migration Service

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

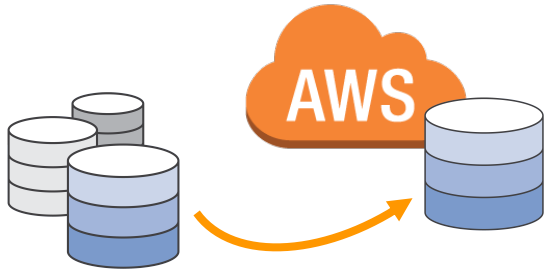
Modernize your database tier –

- Commercial to open-source
- Commercial to Amazon Aurora

Modernize your Data Warehouse –

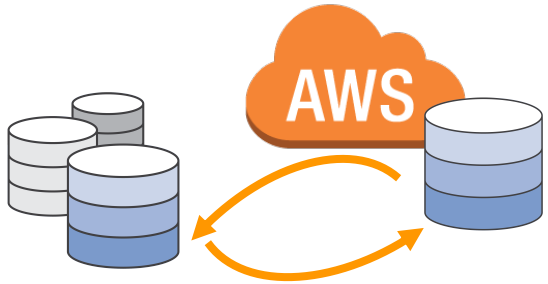
- Commercial to Redshift

Migrate



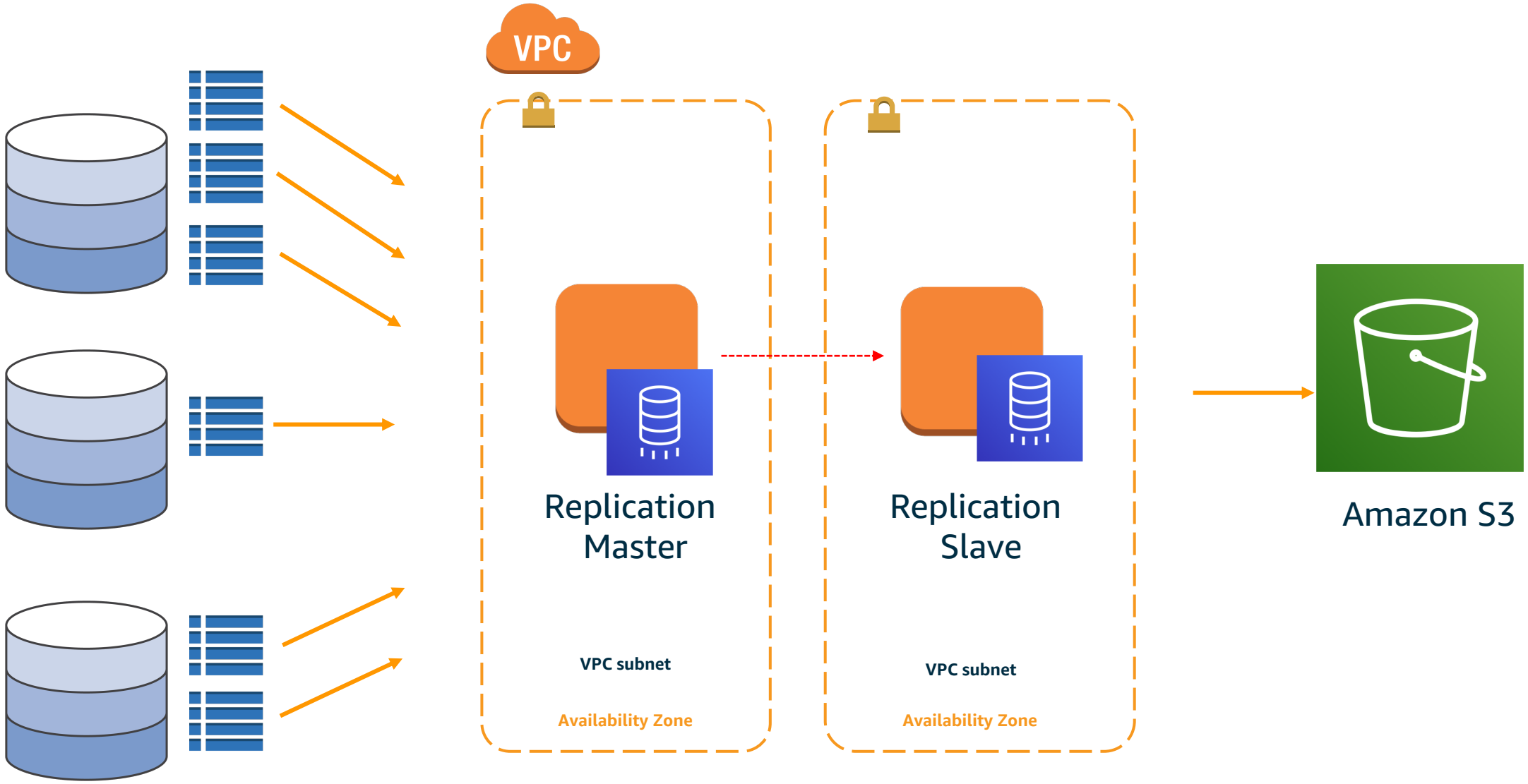
- Migrate business-critical applications
- Migrate from Classic to VPC
- Migrate data warehouse to Redshift
- Upgrade to a minor version

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 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
```

Data Sources - Files

Files

Optimizing Transfers

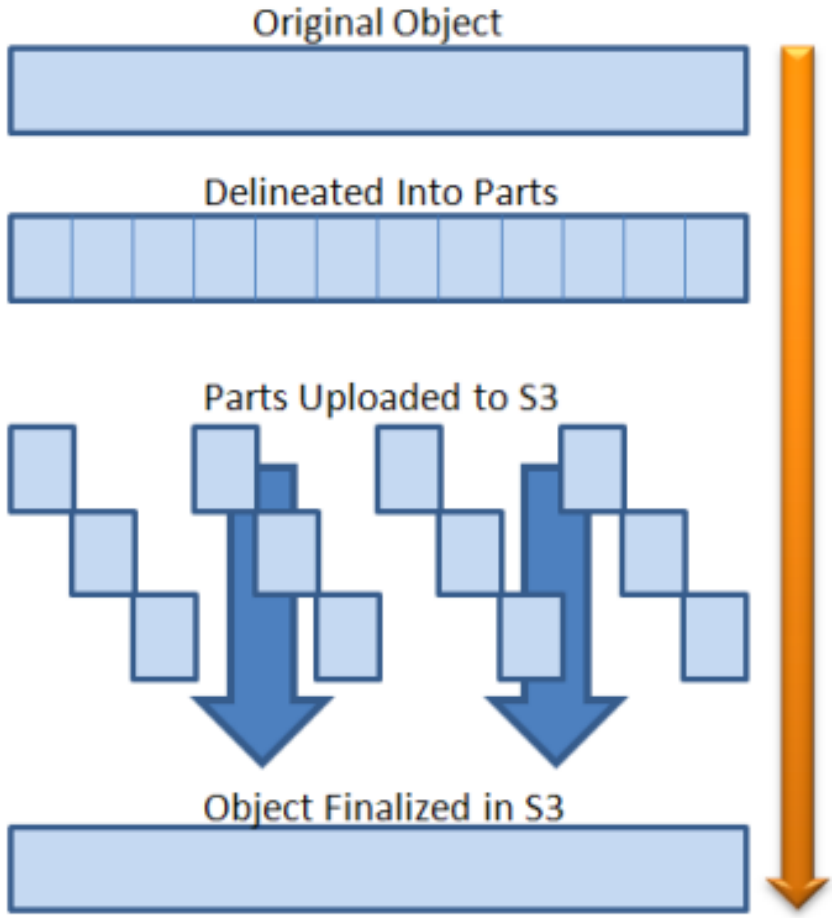
- S3 Multi-Part Upload
- S3 Transfer Acceleration
- AWS Direct Connect

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



S3 Transfer Acceleration

PUT requests go through the nearest AWS Edge Location

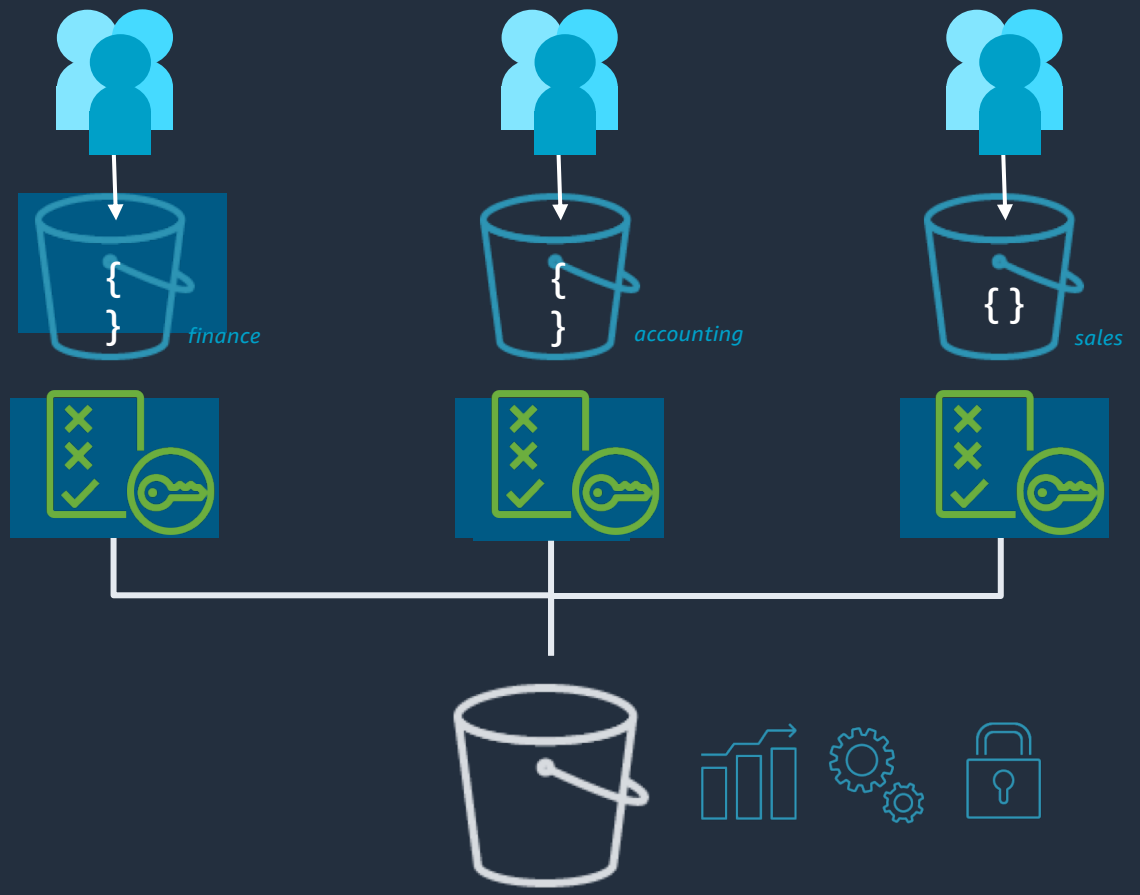
Data transits over the AWS private network rather than Internet

AWS private network optimizes throughput and latency to the AWS Region

Data is not stored in the edge cache

S3 Access Points

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.

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

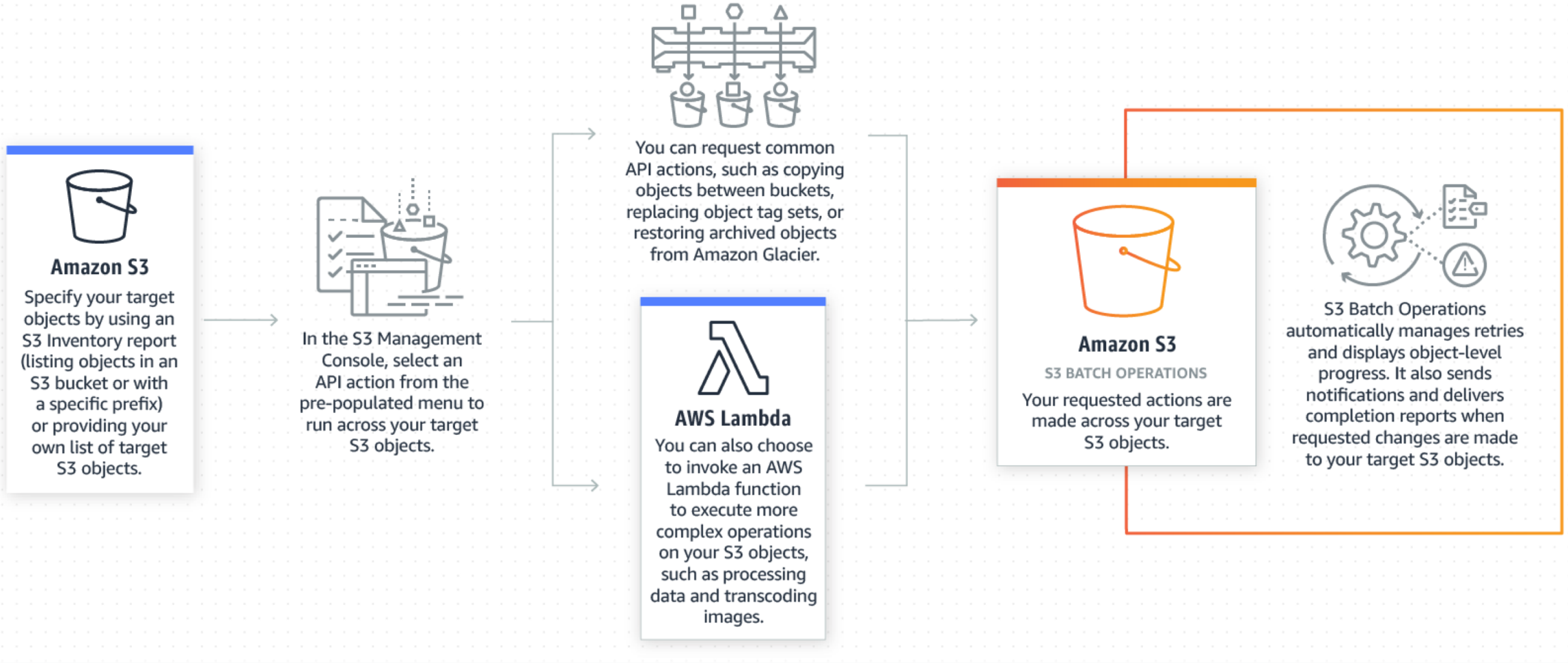
Continuously monitors and automatically analyzes any new or updated resource policy to help you understand potential security implications

Provides the highest levels of security assurance

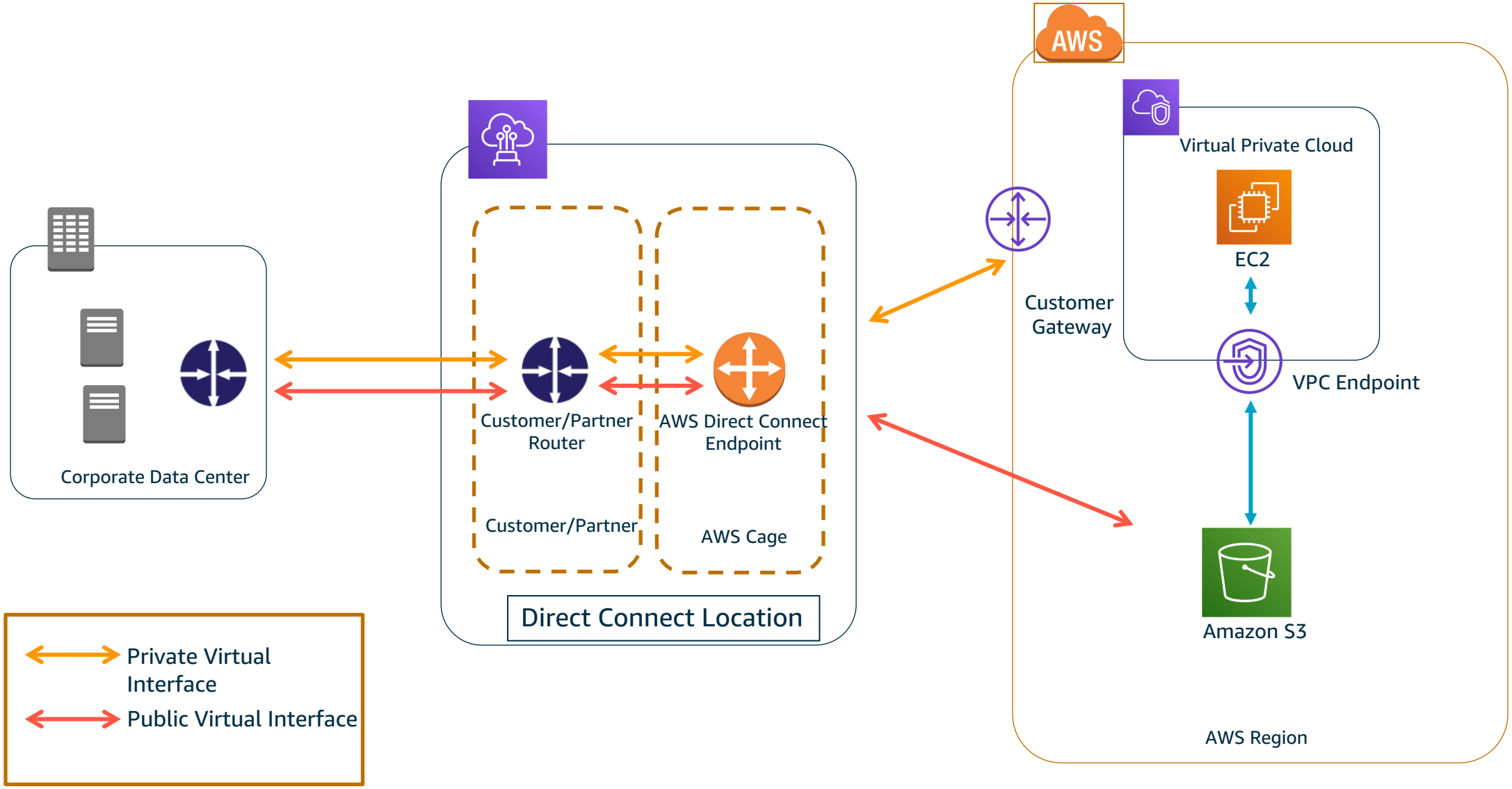
Uses automated reasoning, a form of mathematical logic & inference, to determine all possible access paths allowed by a resource policy

S3 Batch Operations

Manage billions of objects at scale, change object properties, perform storage management tasks



AWS Direct Connect



AWS DataSync

Online transfer service that simplifies, automates, and accelerates moving data between on-premises storage and AWS

Fast data transfer

Easy to use

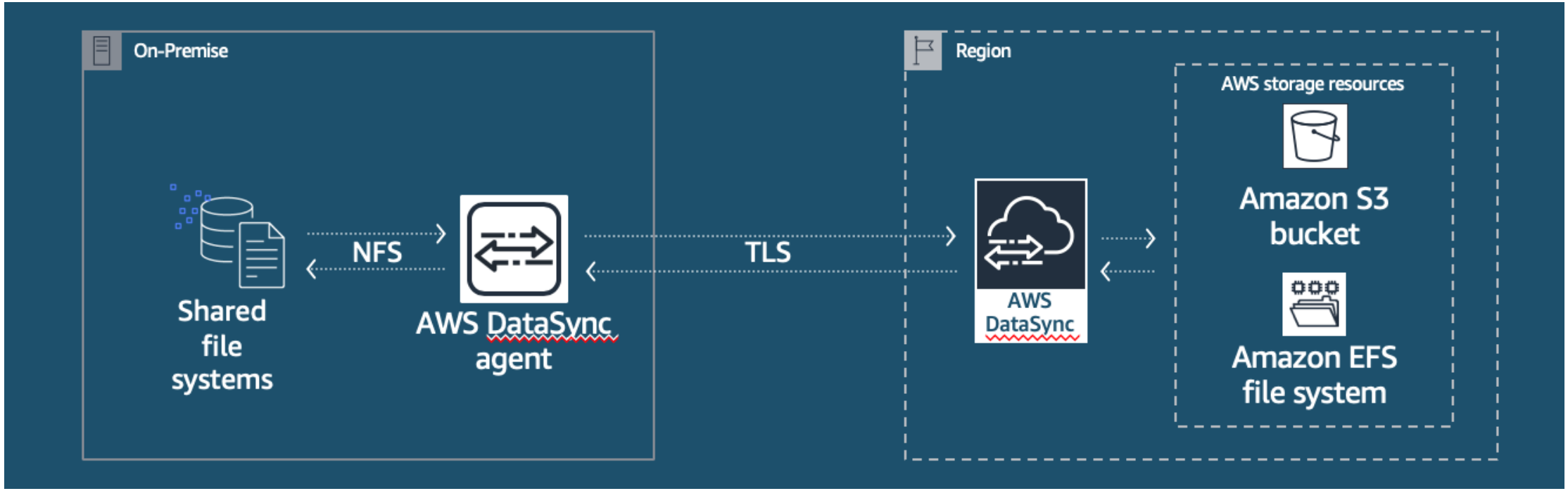
Secure and reliable

Cloud integrated

Cost-effective

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

How AWS DataSync works



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)

AWS Transfer for SFTP

Fully managed SFTP service for Amazon S3

Seamless migration of existing workflows

Fully managed in AWS

Secure and Compliant

Native integration with AWS services

Cost-effective

Simple to use

Setting up AWS Transfer

Map your hostname

Associate your hostname with the server endpoint

Select your S3 bucket(s)

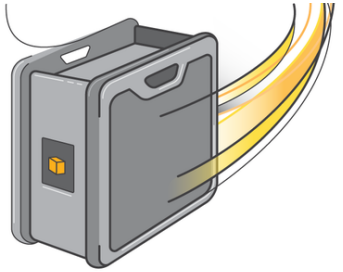
Create an IAM role to access the Amazon S3 buckets used for storing data transferred over SFTP

Set up your users

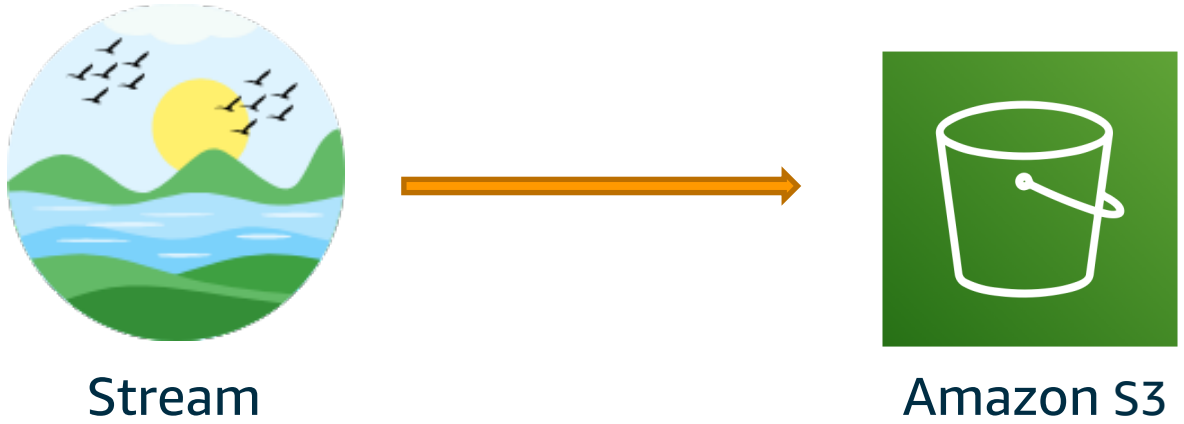
Create and map users to IAM roles to enable them for file operations

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

AWS Snowball/Snowmobile

Use Case	AWS Solution
Cloud Migration, Disaster Recovery	AWS Snowball 
Internet of Things (IoT), Remote Locations	AWS Snowball Edge
Migrating Exabytes of Data	AWS Snowmobile 

Data Sources - Streams

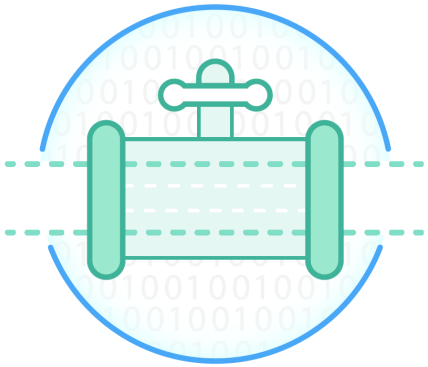


Streams

Collecting and Analyzing

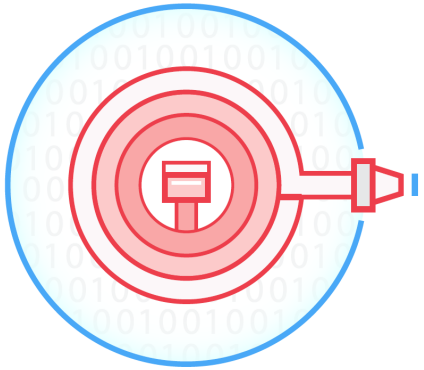
- Amazon Kinesis
- Amazon Managed Streaming for Kafka (MSK)
- Example: Clickstream Analytics

Amazon Kinesis - Stream Processing on AWS



Streams

- Capture streaming data for downstream processing
- Allow multiple processors to read streams at their own rate



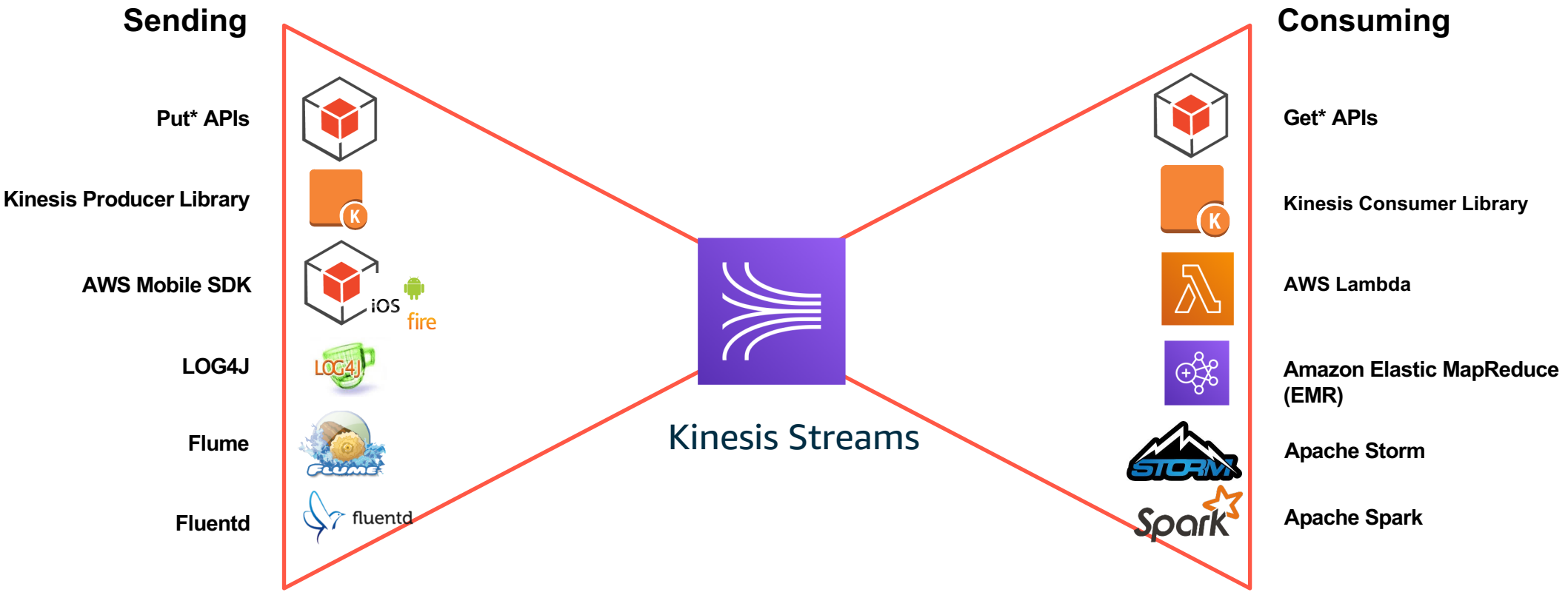
Firehose

- Buffer records in a stream into a single output for more efficient storage
- Automatic flushing of buffer to S3, ElasticSearch, Redshift, or Splunk

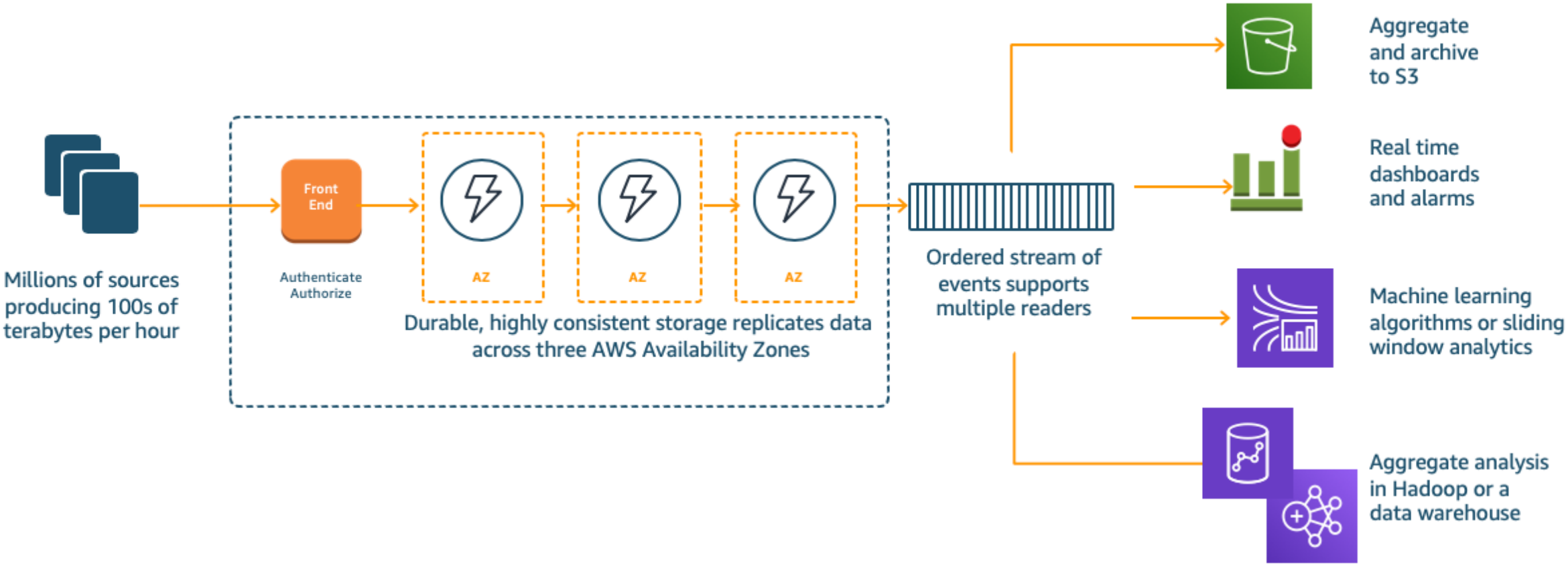
Analytics

- Create time windows over streams and perform aggregate operations using SQL
- Join together multiple streams and output to new streams

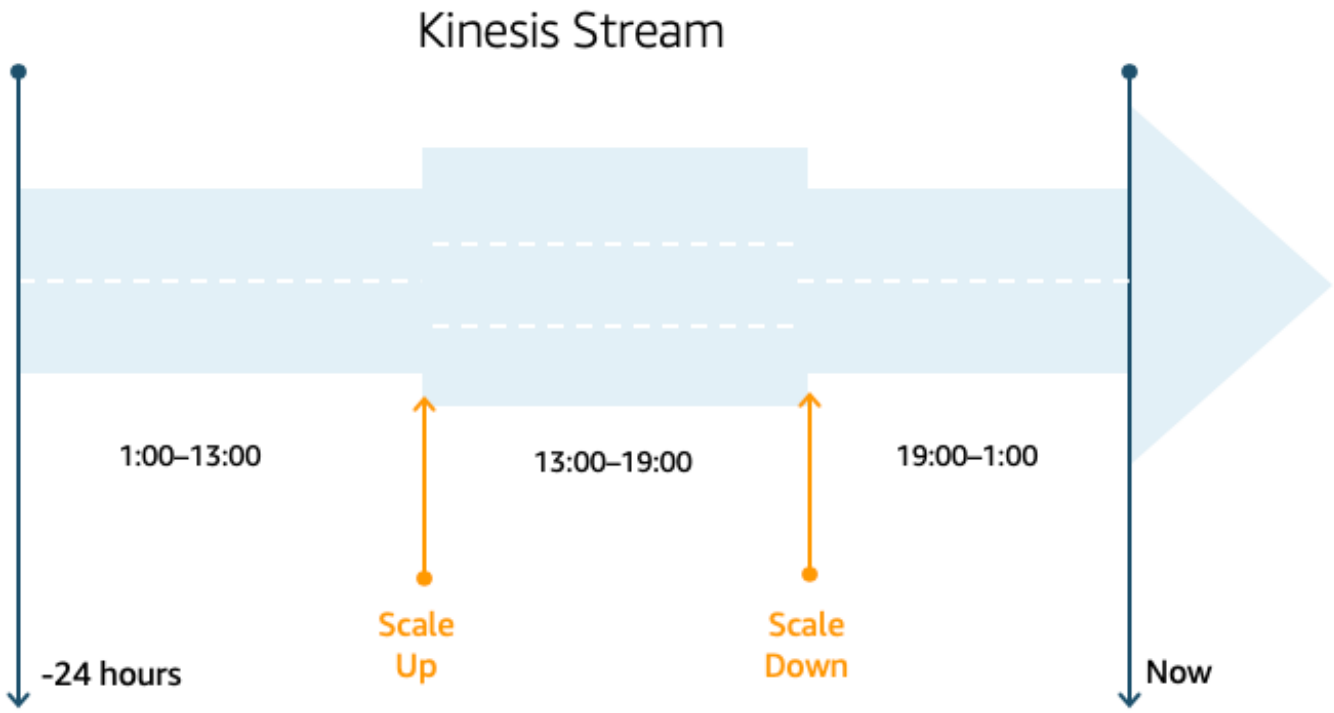
Kinesis – Lots of Integrations



Kinesis – How it works



Kinesis – How it works



- 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

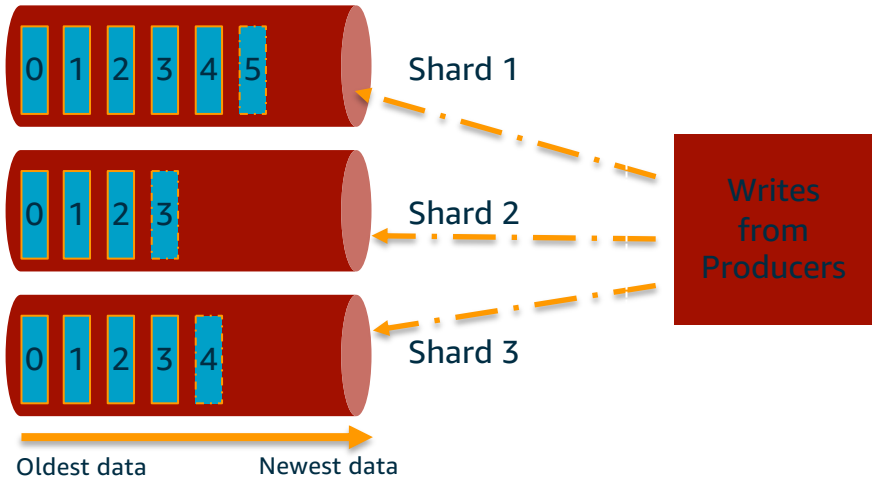
Amazon Managed Streaming for Kafka (MSK)

- Fully compatible with Apache Kafka v1.1.1
- 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

Comparing Amazon Kinesis Data Streams to MSK

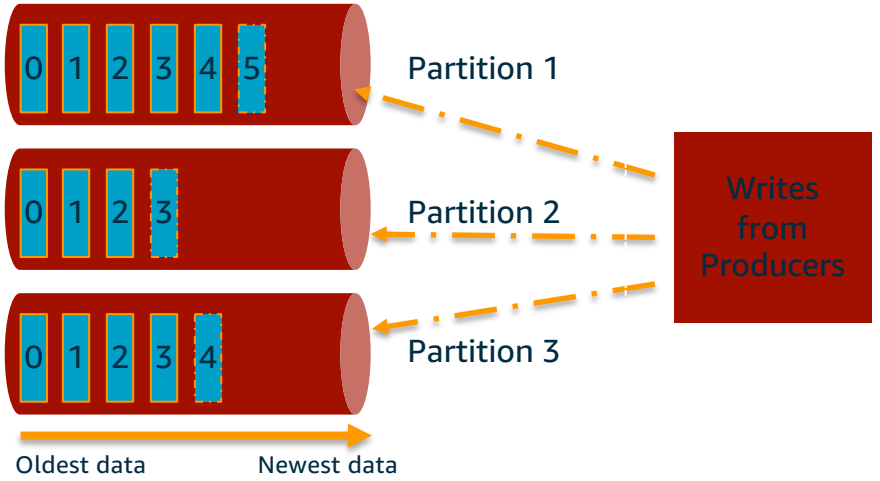
Amazon Kinesis Data Streams

Stream with 3 shards



Amazon MSK

Topic with 3 partitions



Comparing Amazon Kinesis Data Streams to MSK

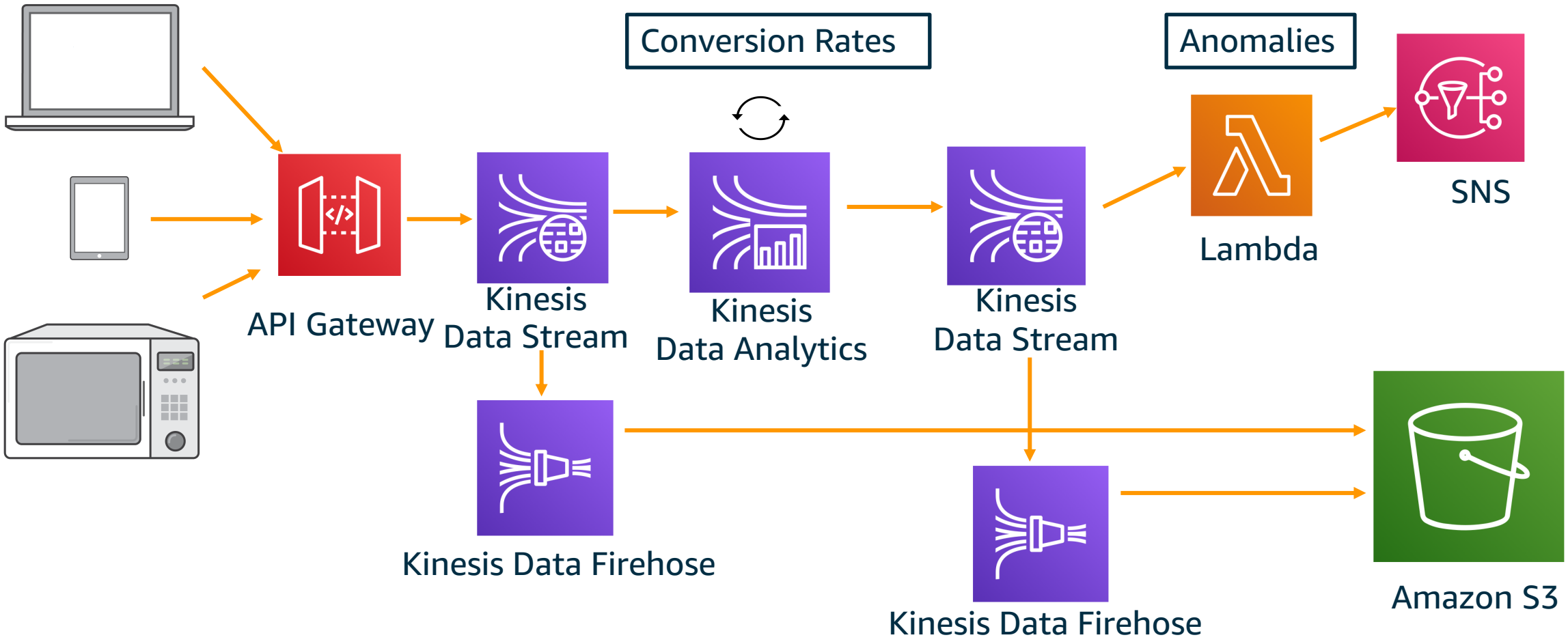
Amazon Kinesis Data Streams

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

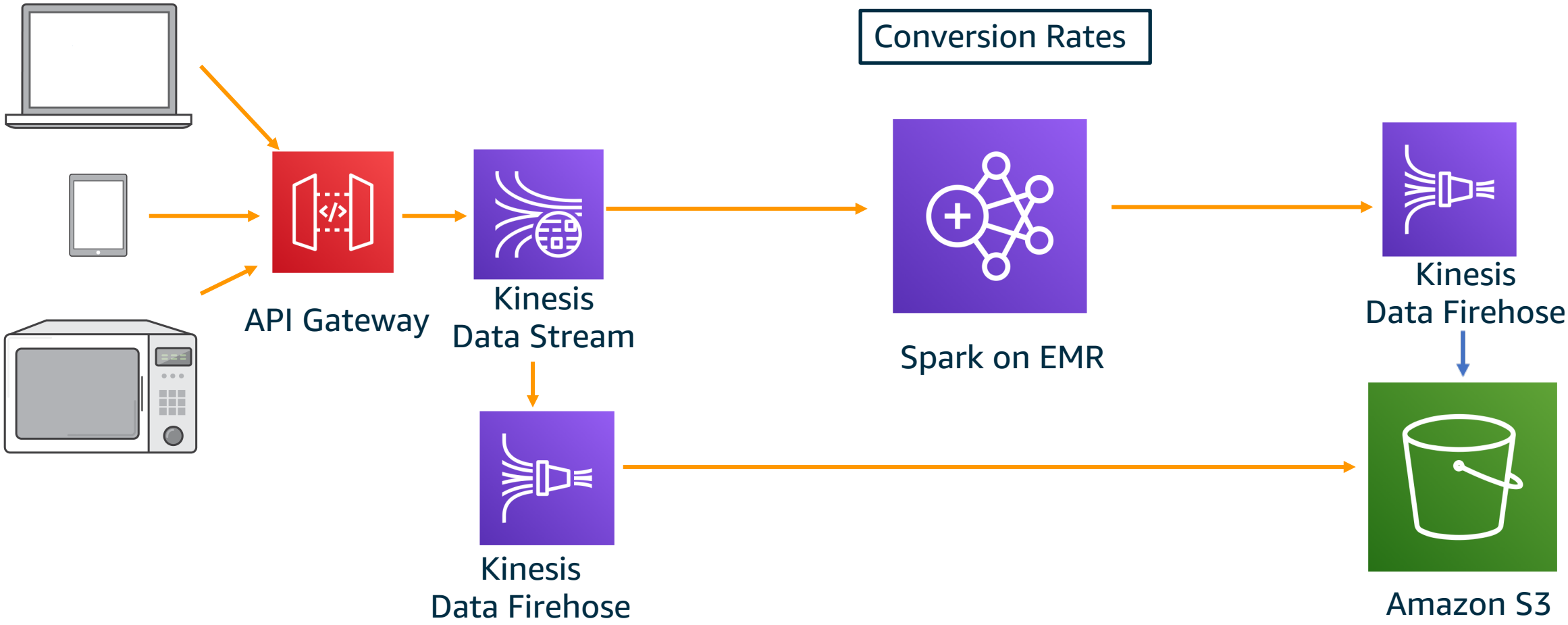
Amazon MSK

- 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

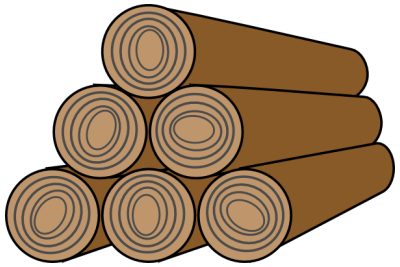


Clickstream with Spark on EMR



Data Sources - Logs

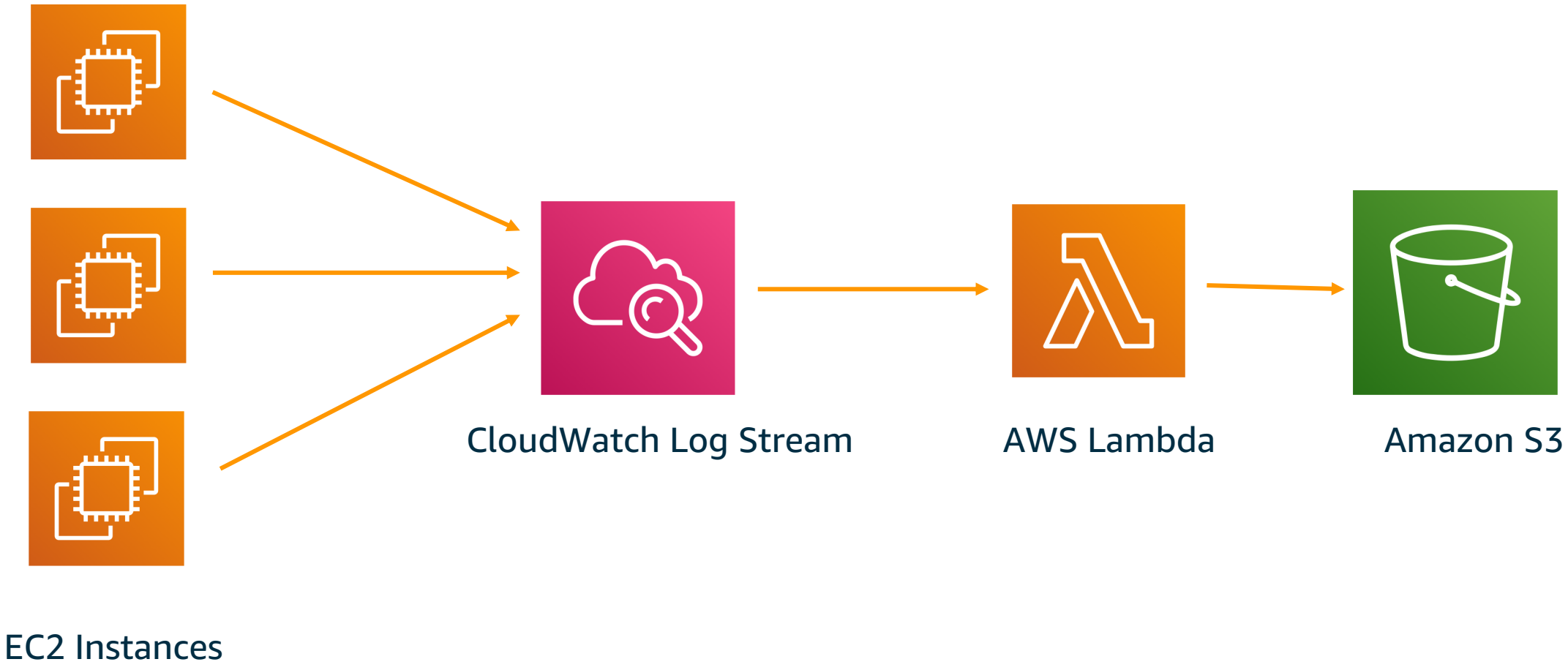
Logs



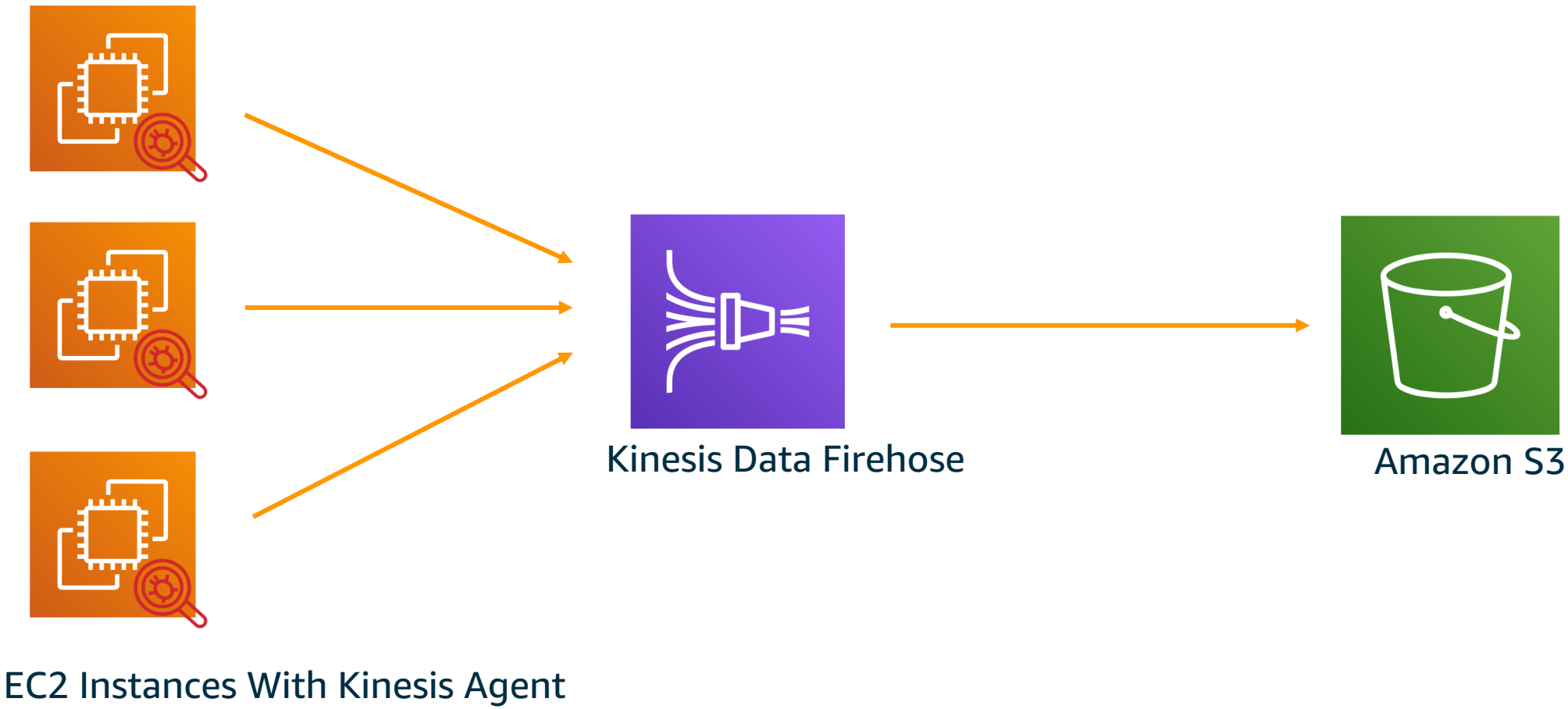
Collecting and Analyzing

- Amazon CloudWatch
- Amazon Kinesis
- Other Options

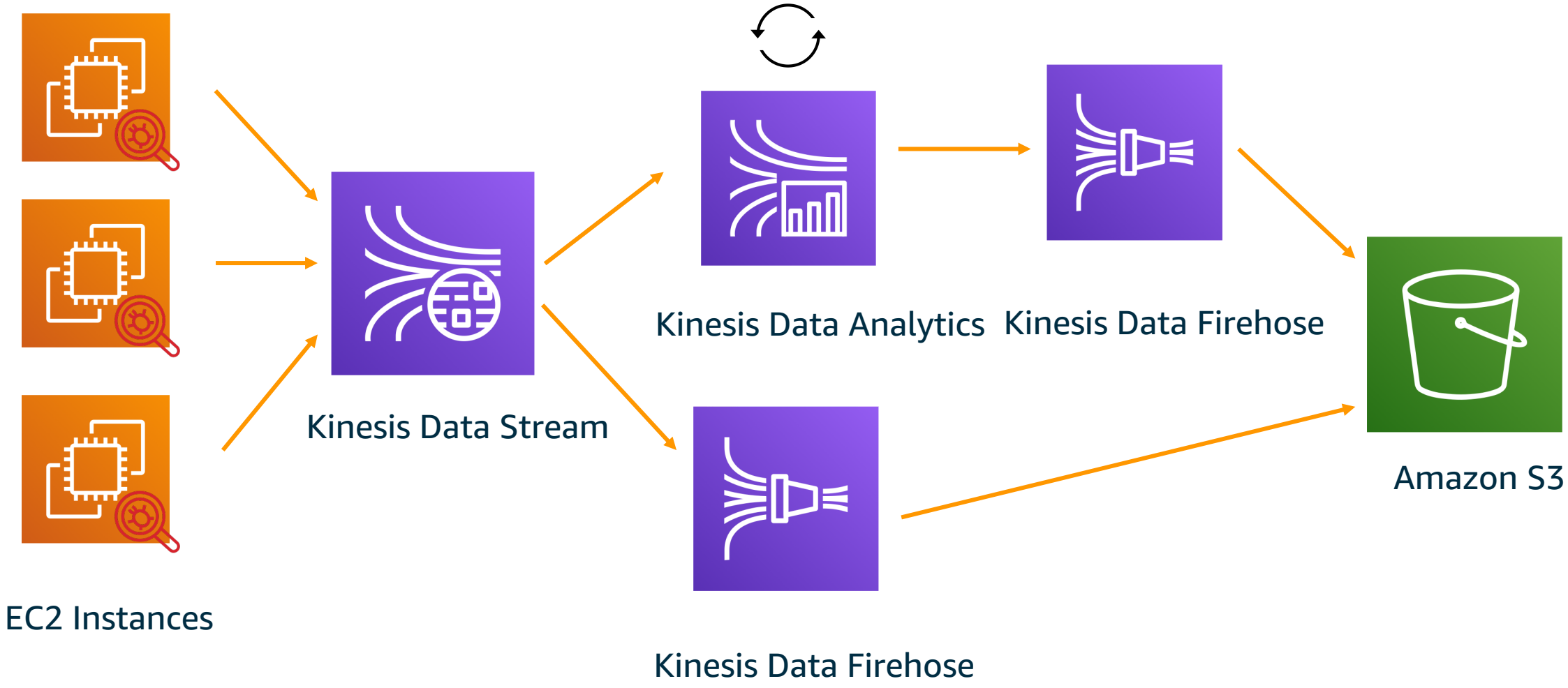
Logs – CloudWatch Agent



Logs – Kinesis Agent



Logs – Kinesis Agent (with Analytics)



Logs – Other Options

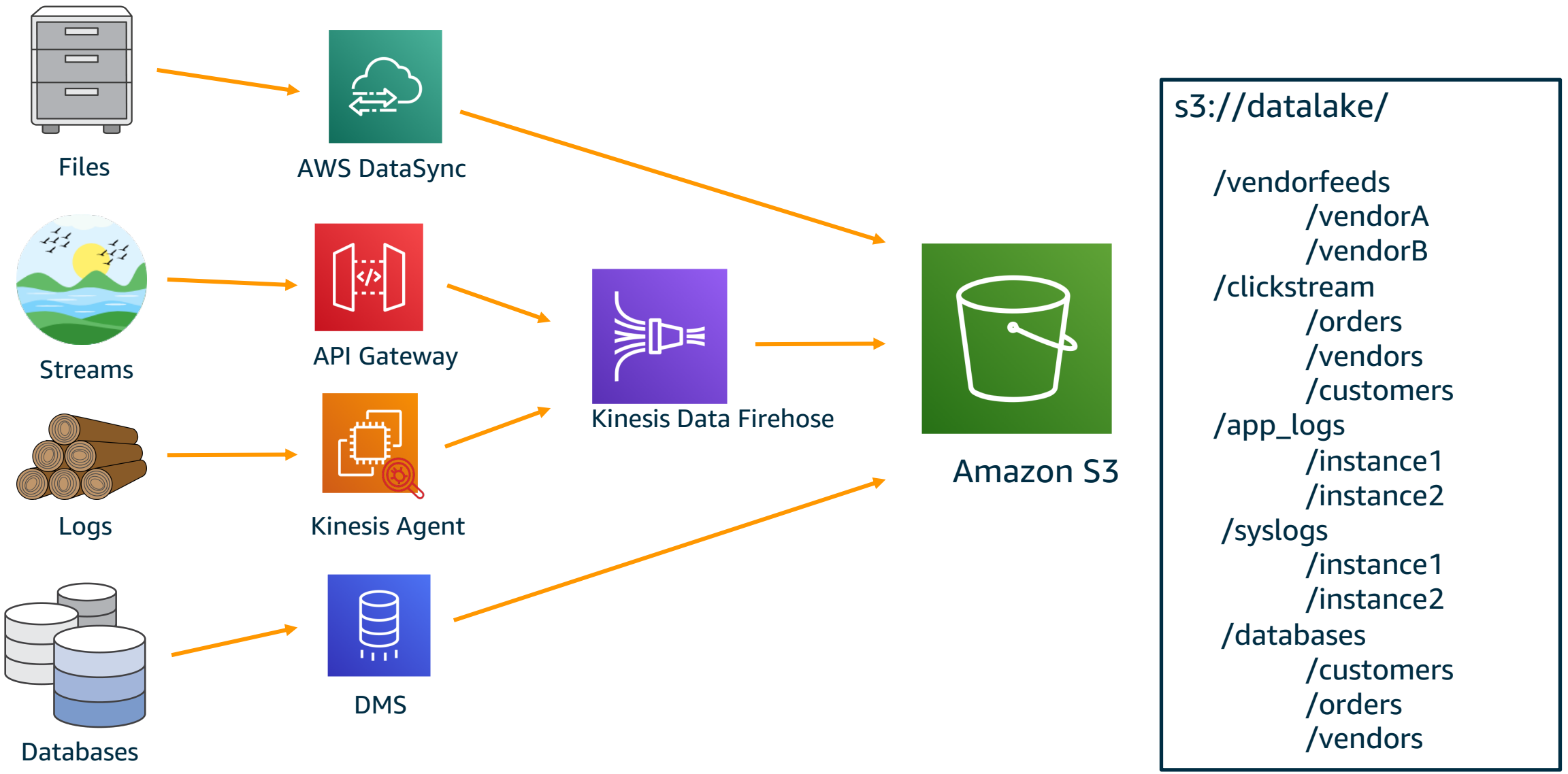
Use OSS agents, like Flume or Fluentd

- 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



Options for data transfer

AWS
Direct Connect

Amazon
Kinesis Data
Firehose

Amazon Kinesis
Data Streams

Amazon Kinesis
Video Streams

Amazon S3
Transfer
Acceleration

AWS
Storage
Gateway

AWS
Snowball

AWS
Snowball Edge

AWS
Snowmobile

AWS
DataSync

AWS
Transfer
for SFTP

Thank you!