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

ANT401

Ingest data at scale with Amazon OpenSearch Service

Kevin Fallis

Principal Analytics Specialist
Solutions Architect
Amazon Web Services

Yair Green

VP of R&D



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Amazon OpenSearch Service

A managed service that makes it easy to deploy, operate, and scale OpenSearch and legacy Elasticsearch clusters in the AWS Cloud

What is OpenSearch?



Community-driven, open-source search and analytics suite derived from Apache 2.0 licensed Elasticsearch 7.10.2 & Kibana 7.10.2

Consists of a search engine, OpenSearch; a visualization user interface, OpenSearch Dashboards; as well as a series functionality-adding tools and plugins

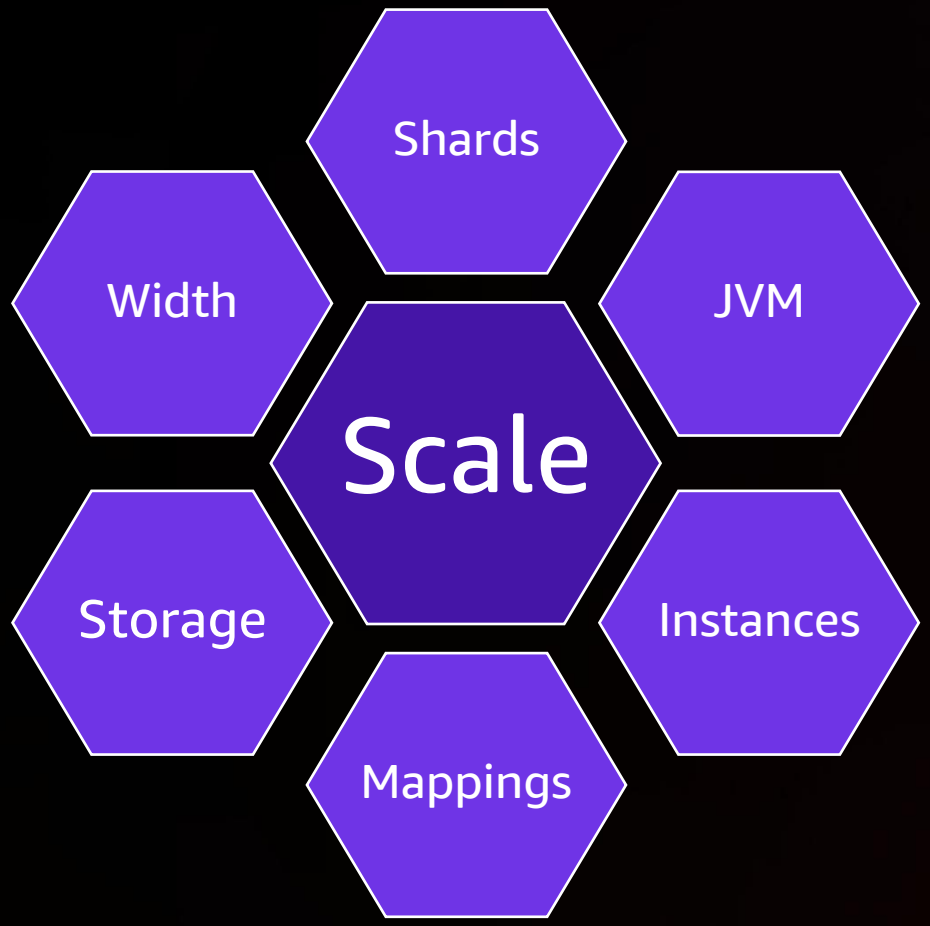
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Designing for scale

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

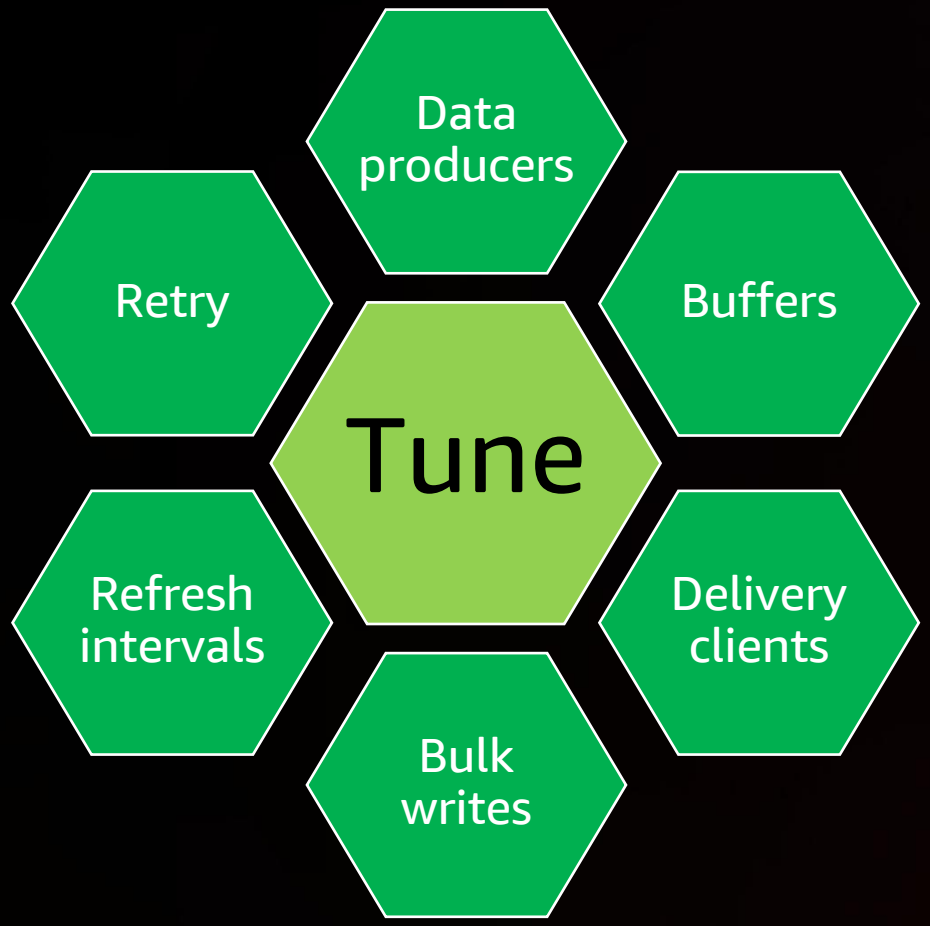
Properties that influence ingestion throughput



- Shards – for high volume, deploy at least a primary shard per node
- JVM – heap is capped at 31.5GB once past 64GB of RAM for most instances; go wider for more
- Instances – prefer R and I series for heavy ingestion
- Mappings – log analytic workloads are mainly aggregations; prefer keywords
- Storage – GP3 for Amazon Elastic Block Store (EBS) provides up to 4x the throughput of GP2
- Width – wider clusters give you more threads, queues, and processing power

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Focus areas for tuning your ingestion pipeline



- Data producers – avoid direct writes at scale and push to a buffer
- Buffers – prefer buffers to mitigate the impact of spikes
- Delivery clients – avoid co-mingling different data delivery solutions; coordinate your clients
- Bulk writes – bulk size influences throughput; test for optimal payload size
- Refresh intervals – set between 30 and 60 seconds
- Retry – when encountering 429 errors, back off exponentially and avoid excessive retry attempts

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data preparation

Transform at the edge

Self-generating _id field adds overhead

OpenSearch ingest pipelines add overhead

Avoid the “swamp” – ingest only what you need

For PII, consider a redaction workflow

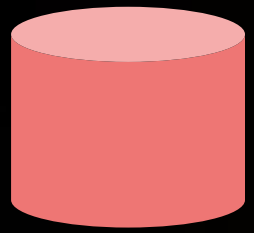
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Four primary ingestion sources

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data sources for Amazon OpenSearch Service

SQL and NoSQL data



Agents via queues and streams



Amazon CloudWatch Logs subscription filters



Amazon S3



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How Syte ingests data at scale with Amazon OpenSearch Service

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

syte

Yair Green

VP R&D @ Syte



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

syte

Syte is the world's first **product discovery platform,** using AI to redefine how people search and discover



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Syte's core technologies

Visual domain

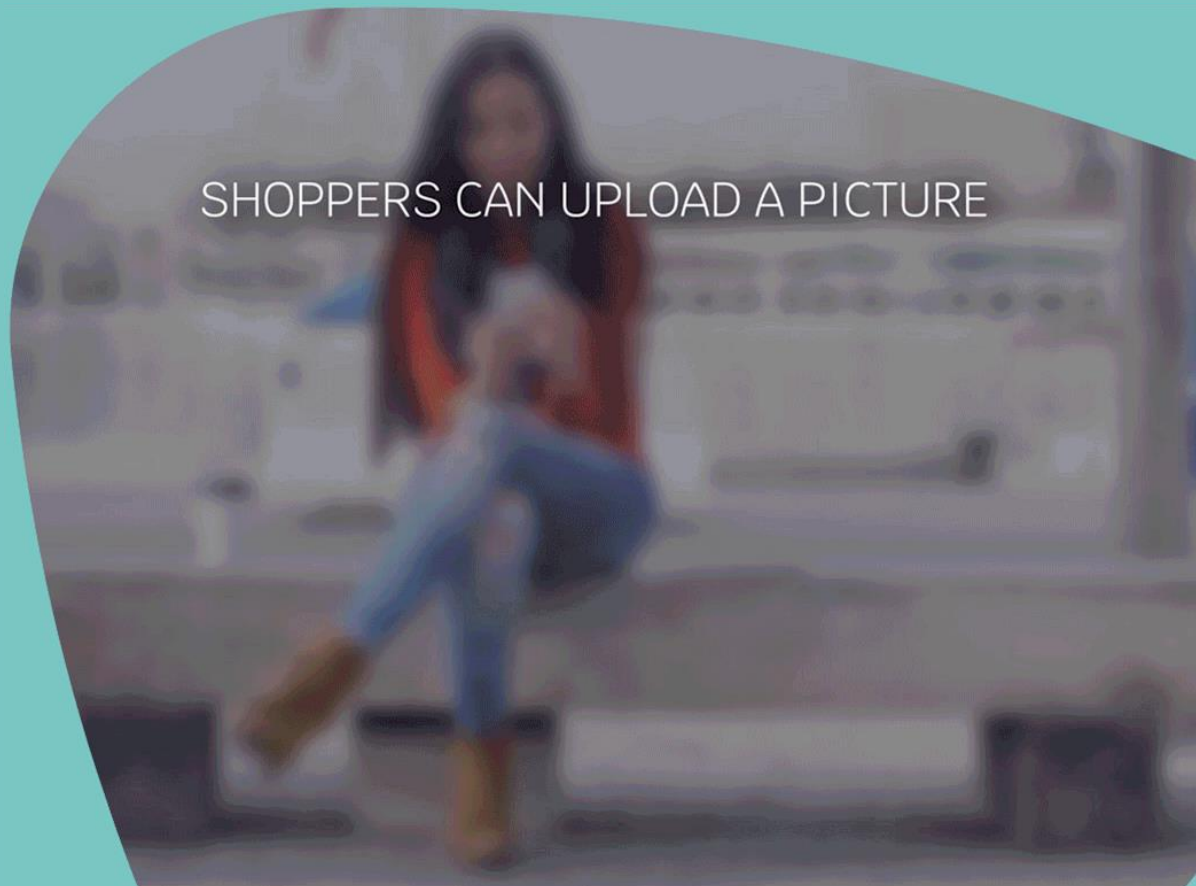
Image-to-text (deep tags)
Visual similarity
Shop the look
Camera search

Personalization

You may also like
Frequently bought together

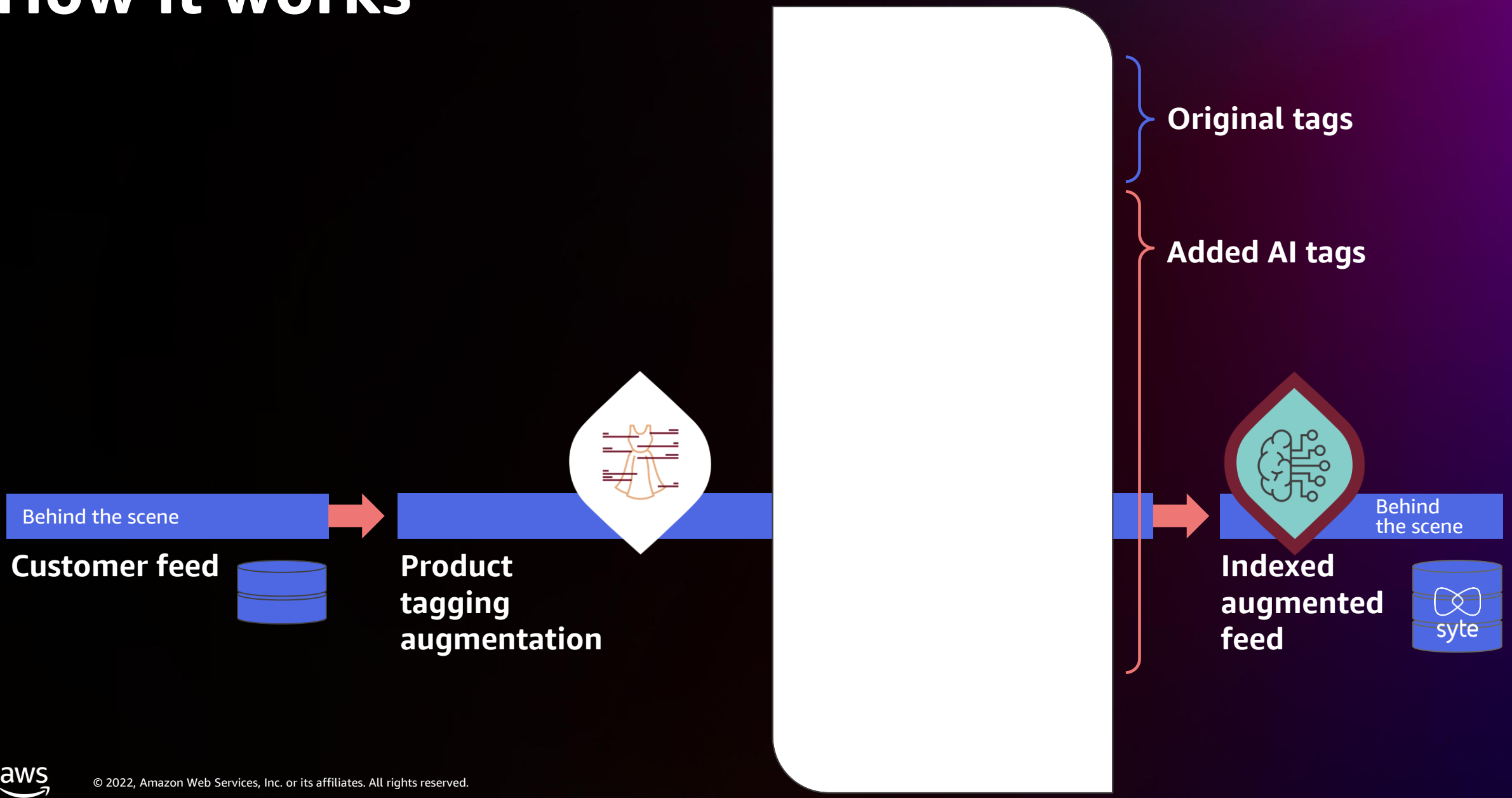
NLP domain

Matching algorithms
Ranking algorithms



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How it works



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Visual flow



1. BB

Dress

2. CBIR

Extract deep tags

Find the most similar items among all the dresses in customer's feed

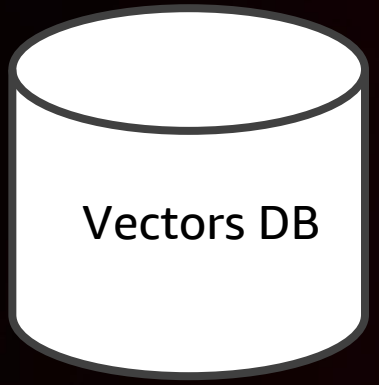
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Find the most similar items within the client's feed

Ingestion – ETL
Client feed processing and indexing



Image vector



Find closest vectors

Search API
Receive image, get vector, and find similar products in client's feed



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Syte ingestion unique value

syte

- Fast onboarding
- Fully automated
- Support scale
- Rich API

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Syte's ingestion flow

Ingestion is a backend system responsible for processing client's data and writing it to the Syte database, which is called indexing process

Catalog

Data is taken **(extracted)** from a source system

Ingestion

Converted **(transformed)** into a format that can be analyzed

Database

Stored **(loaded)** into a database



aws

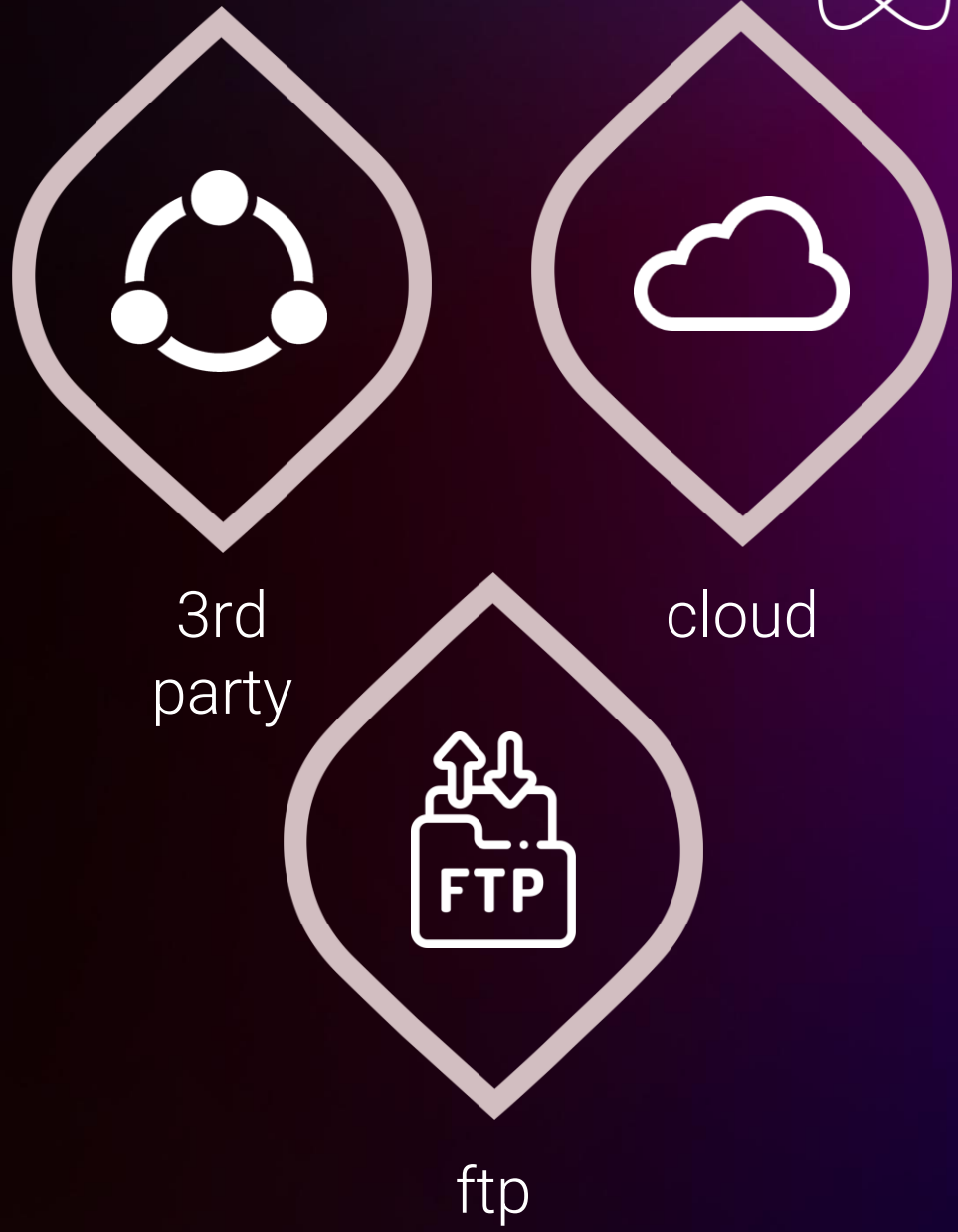
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Steps – Extract

Download the customer's feed

Every client is using a different platform to hold his catalog and the Ingestion system supports many options for downloading





© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Steps – Transform

Make data usable for Syte's needs

For each item ingestion system knows how to:

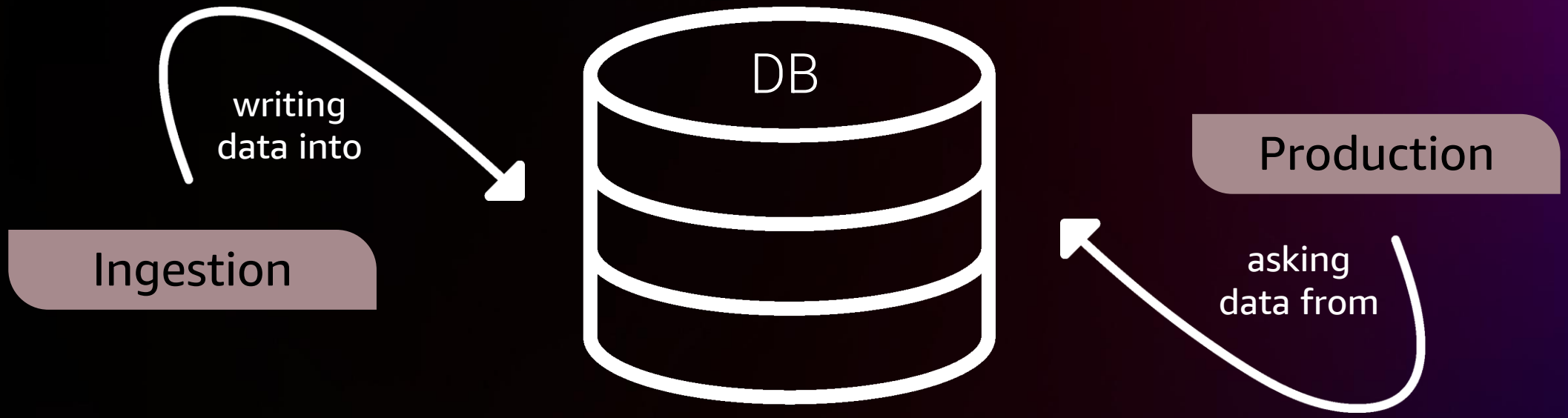
- Map the relevant fields from the catalog to Syte's fields
- Decide which Syte's category the item should be mapped
- Download the image of the item and get it's BBs, deep tags and vectors



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Steps – Load

Writing the relevant output that was created during the process to the DB so when the end-users is searching, we will be able to go and retrieve for them the match results



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Syte's ingestion tech stack

syte

Programming languages

python

C++

ERLANG

AWS services



Amazon OpenSearch Service



Amazon S3



Amazon EFS



Amazon EC2



Amazon EKS



Amazon MSK

Databases

AEROSPIKE

mongoDB

redis

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

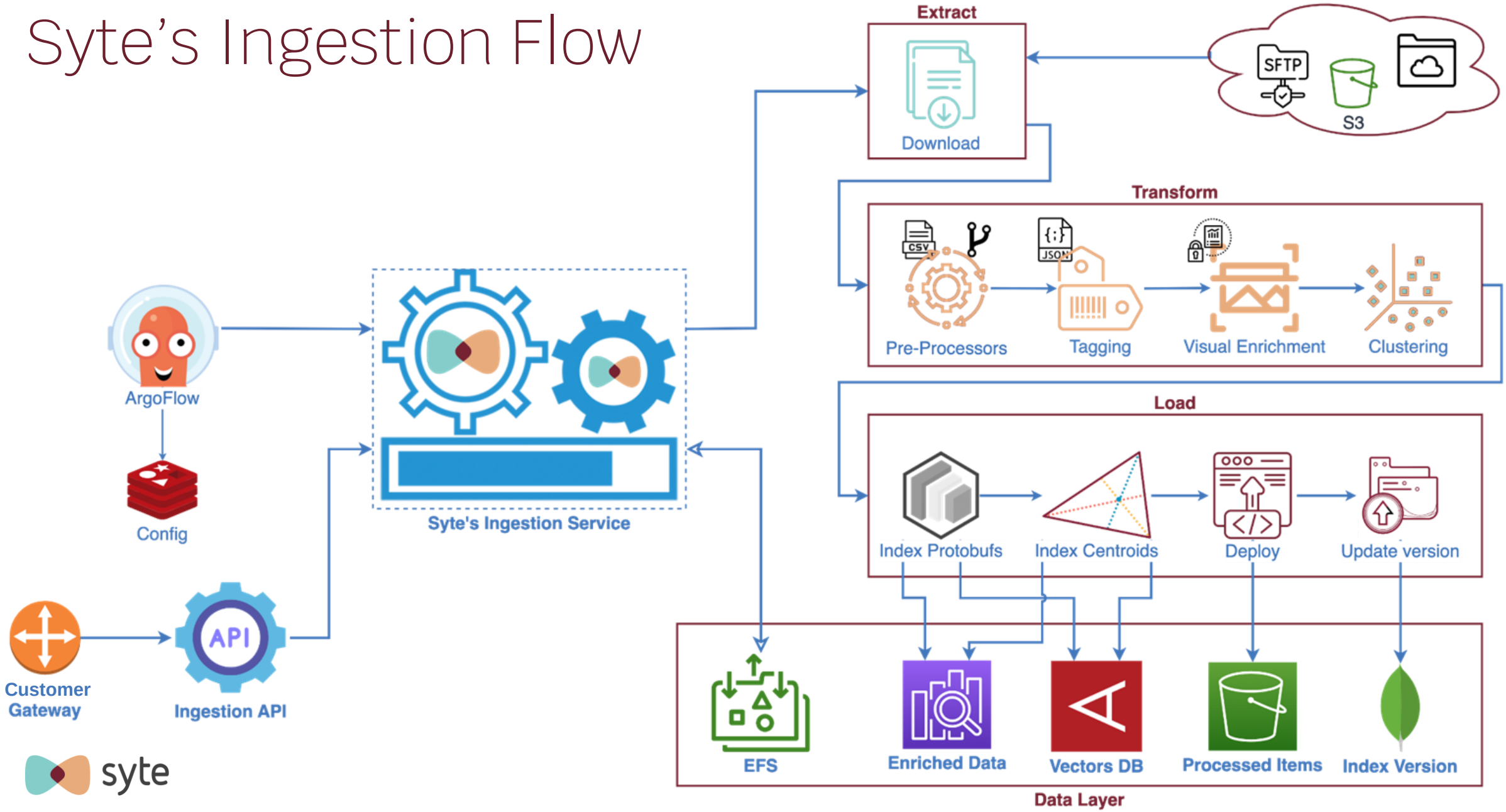
Syte's ingestion automation

The single-tenant ingestion flow is fully automated and orchestrated using Argo Workflows



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Syte's Ingestion Flow



Syte's ingestion in numbers



>800
feeds in production

>350,000,000
processed items every day

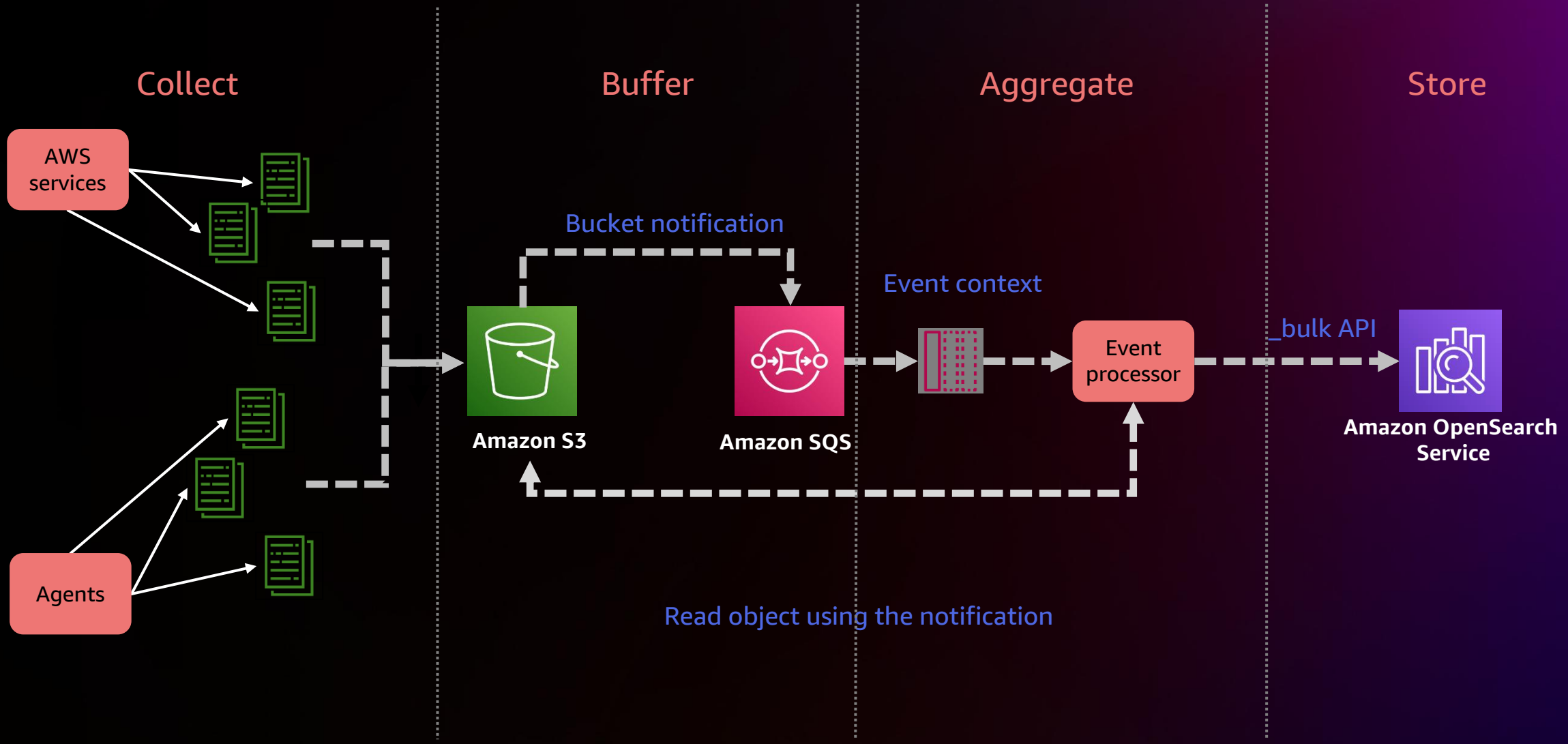
>2,500
ingestion cycles every day

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Popular ingestion solutions

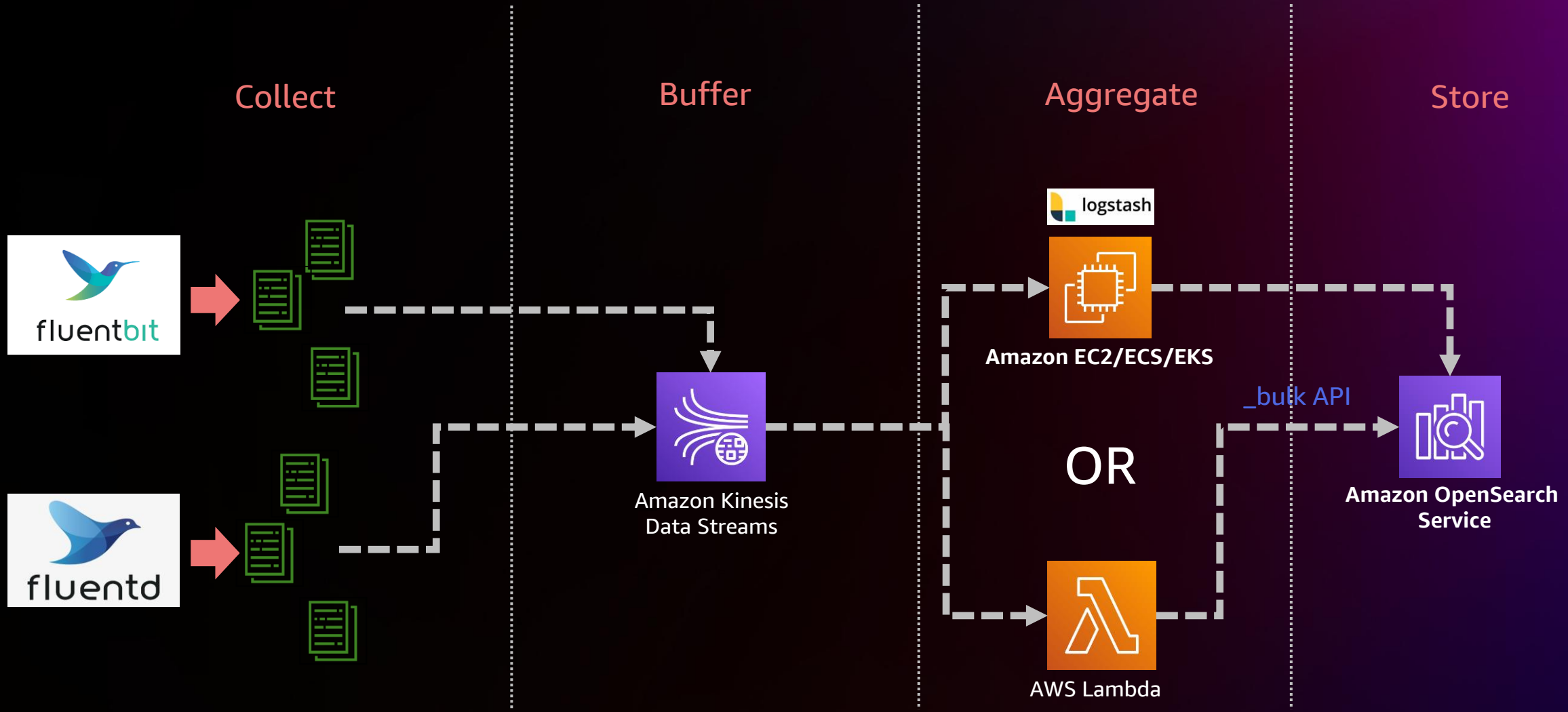
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Generic pattern for Amazon S3 sourced data



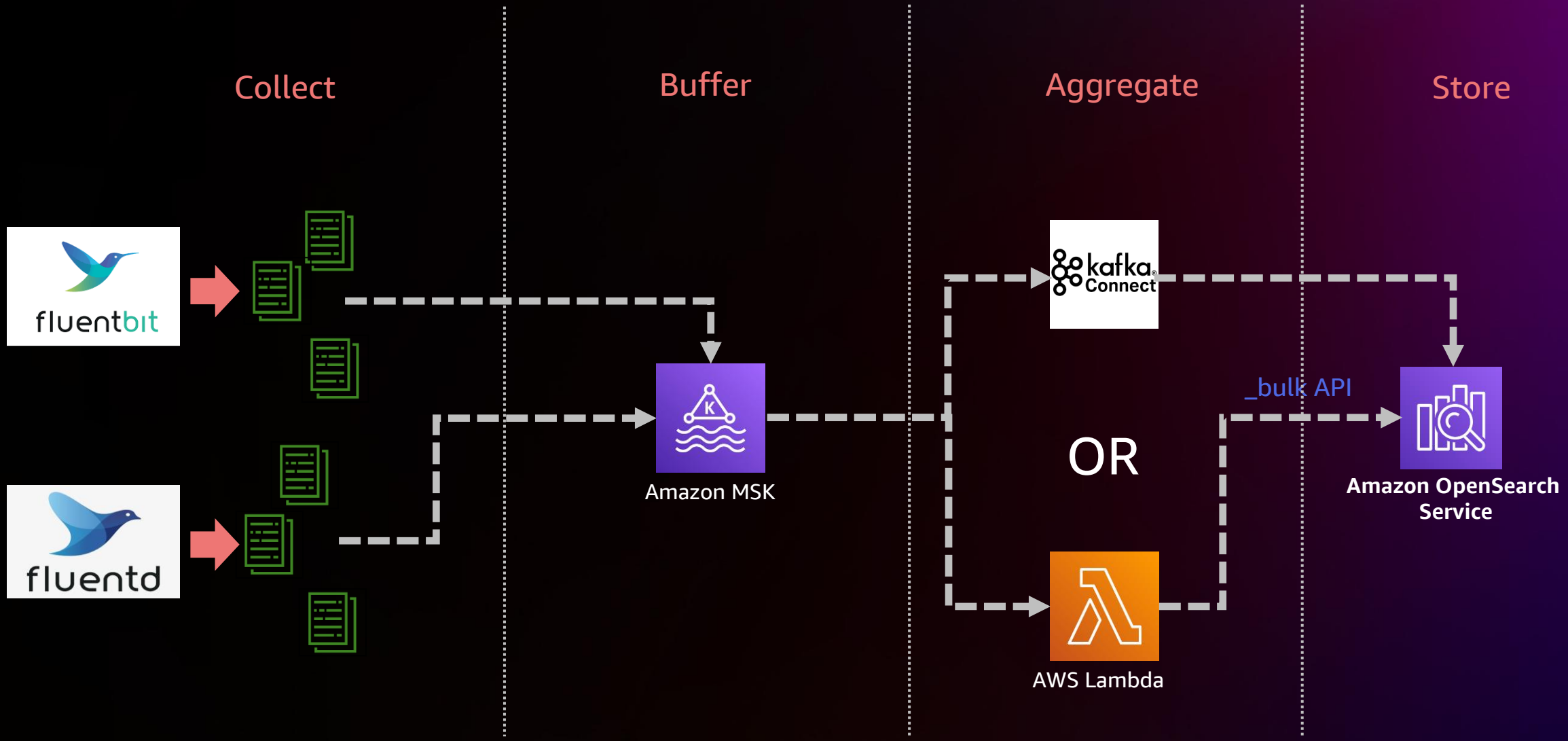
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Kinesis Data Stream sourced data



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon MSK sourced data



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Kevin Fallis
kffallis@amazon.com

Yair Green
yair.g@syte-vc.com



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

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.