aws partnercast

AWS PartnerCast: Enabling Advanced Search with Amazon OpenSearch and Machine Learning with Amazon SageMaker

Chris Turner & Gitika Vijh

Partner Solutions Architects

turncat@amazon.com

gitikav@amazon.com

OpenSearch Project

OpenSearch is an Apache 2.0-licensed search and analytics suite comprising OpenSearch, OpenSearch Dashboards, and a suite of plugins providing advanced anomaly detection, alerting, observability, and security analytics

>200MM
OpenSearch project downloads since launch

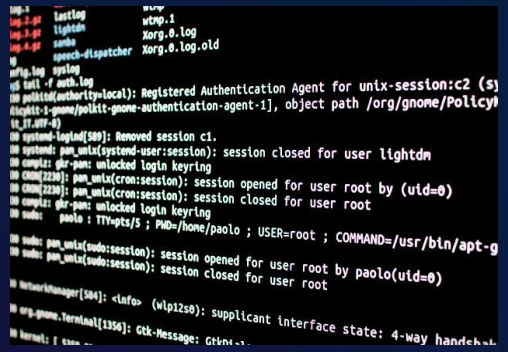
Top 4 search engine
DB-Engines ranking

55+
Partners and growing

10k+ pull request merged
200%+ growth

Multiple service providers
AWS, Oracle, Aiven-Azure, Bonsai-GCP

Evolution of search engines



Text search

Documents

E-commerce

Relevance ranking

Streaming data

High-volume ingest

Near real-time

Distributed storage

Analysis

Time-based visualizations

Nestable statistics

Time series tools

AI/ML techniques

Vector capabilities

Semantic and hybrid search

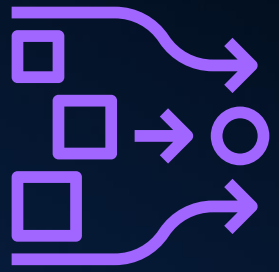
Conversational search

Search Services – a suite of related offerings

OpenSearch and OpenSearch Dashboards

Amazon OpenSearch Service

Amazon OpenSearch Serverless



Amazon OpenSearch Ingestion

Industry examples for real-time search at scale

Find the right product, service, document, or answer quickly--across semi-structured and unstructured data and different facets and attributes.

Retrieve the most relevant search results in large collections in real time, economically and securely

INDUSTRY USE CASES

e-commerce platform: customers find the right product quickly; manage promotions

D**ocument portal:** knowledge base, research articles, investment analyses, health records… Speedy and relevant document search experience.

Recommendation engine (weekly playlist, recipes): Increase user engagement by **delivering personalized recommendations**.

Platform search services: easy to use and snappy search experience with machine learning capabilities.

What’s New with Amazon OpenSearch

NEW

Amazon OpenSearch Serverless

Run OpenSearch on the AWS Cloud without worrying about infrastructure or index and shard strategy

Easy to administer

No sizing, scaling, and tuning of clusters, and no shard and index lifecycle management

Fast

Automatically scale resources to maintain consistently fast data ingestion rates and query response times

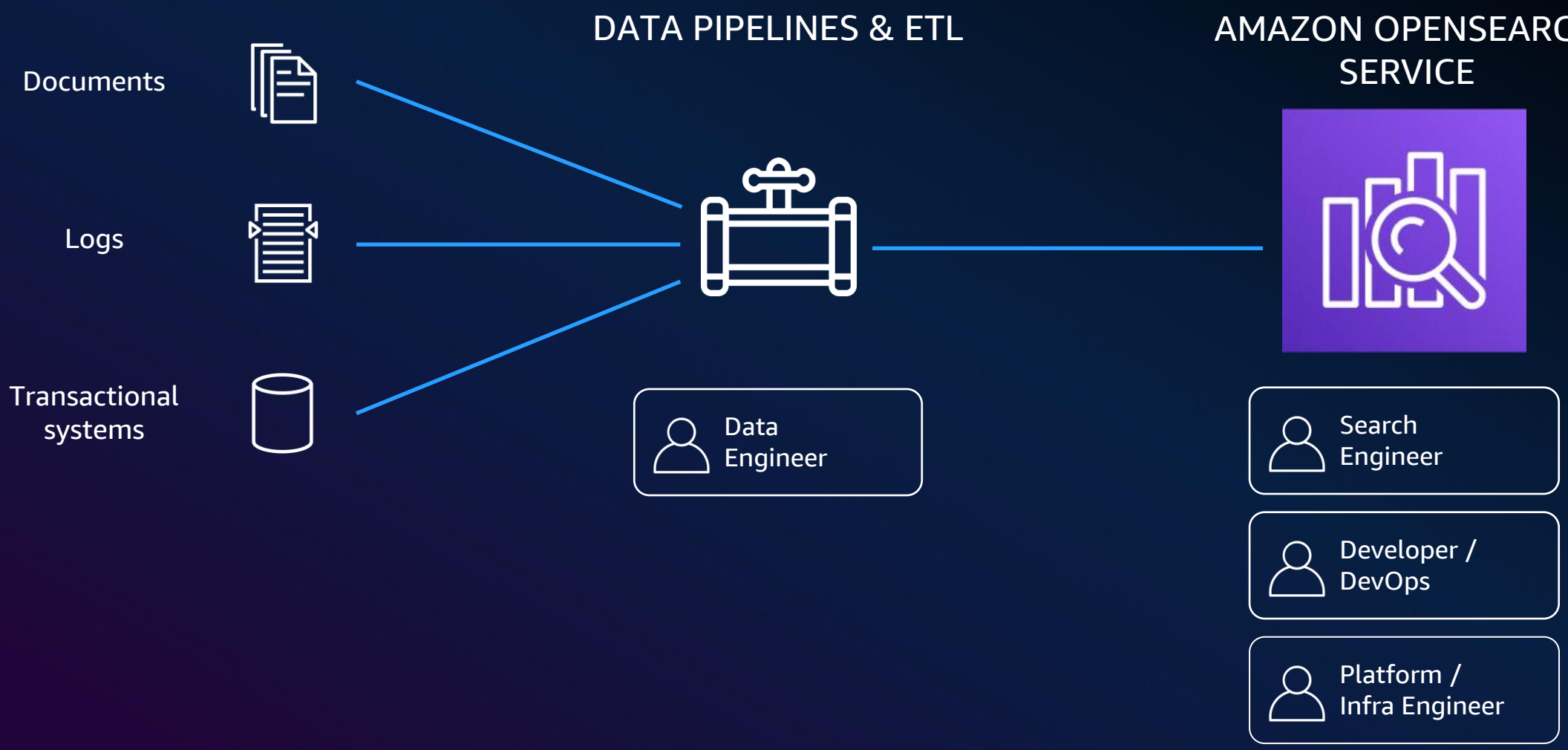
Ecosystem

Get started in seconds using the same OpenSearch clients, pipelines, and APIs

Cost-effective

Pay only for the resources consumed

Amazon OpenSearch Service landscape



NEW

Amazon OpenSearch Ingestion

POWERED BY Data Prepper

Reduce cost

Deduplicate, sample, filter data, and route noisy data to lower cost storage

Enforce data quality

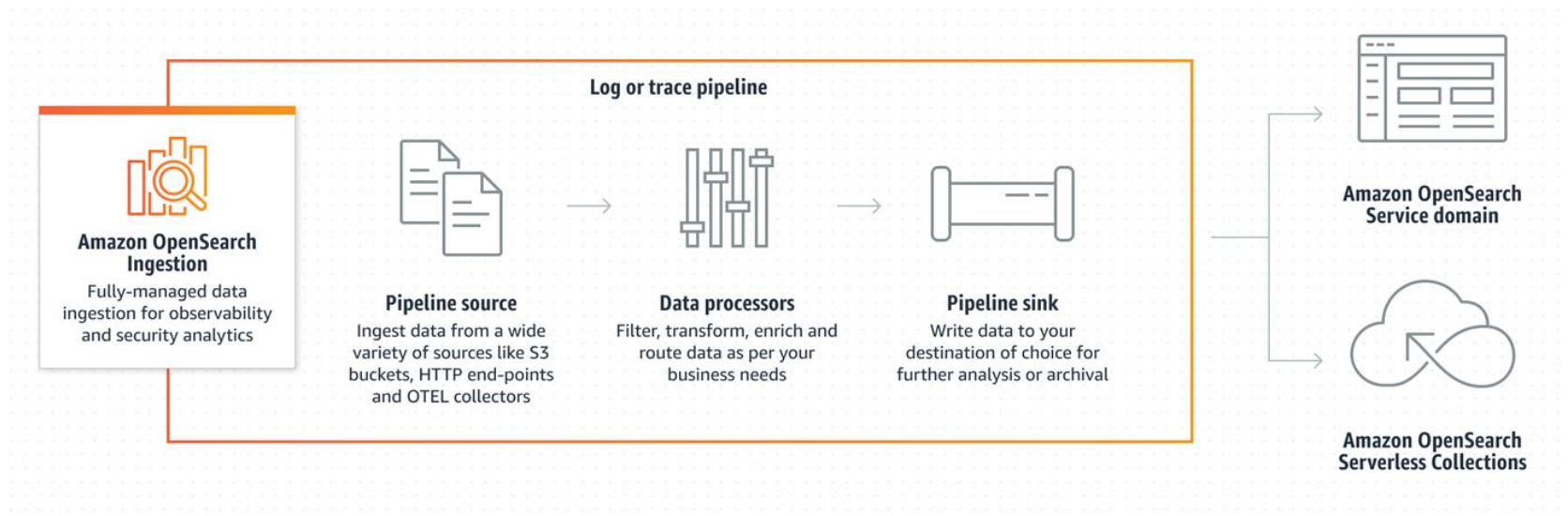
Transform, filter, and enrich data by adopting schemas to accelerate observability and security investigation / resolution times

Protect sensitive data

Redact and obfuscate sensitive information. Route data to maintain compliance with data residency laws.

aws

Serverless Ingestion Pipelines



- Filter, transform, enrich, and route data to managed clusters and Serverless collections
- Serverless implementation provisions, manages, and scales pipelines seamlessly
- Serverless multi-AZ persistent buffering available for push-based sources NEW
- Support for Elasticsearch 7.x and OpenSearch as a source for data migration NEW

Getting data to OpenSearch: tech stack

Collection


Amazon S3


Kinesis Agent


Beats


Fluent Bit


Fluentd


Logstash

Buffering


Amazon Kinesis Data Stream


Amazon Kinesis Data Firehose


ElastiCache


Rabbit MQ


Amazon S3


Kafka

Transformation


AWS Lambda


AWS Glue


Amazon Kinesis Data Firehose

Spark Streaming


Logstash


Flink

Query & Visualization


Amazon OpenSearch Service


OpenSearch Dashboards


Amazon Athena


Kibana


Prometheus


Grafana

Introducing: OpenSearch Ingestion

Collection


Amazon S3

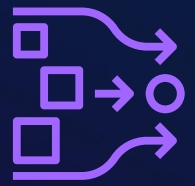

Kinesis Agent


Beats


Fluent Bit


Fluentd


Logstash



OpenSearch Ingestion

Query & Visualization


Amazon OpenSearch Service


OpenSearch Dashboards


Amazon Athena

Kibana


Prometheus


Grafana

OpenSearch OR1 optimized instance family

NEW

ADDRESSING COST AND DURABILITY

80% Indexing Throughput Improvement

High Indexing Throughput

30% Price Performance Improvement

Lower overall cost

11 9s of Durability

Data is indexed to Amazon S3 a durable data store

Automatic Recovery

Automatic recovery from red indices

NEW in the AWS SOLUTION LIBRARY

Migration Assistant for Amazon OpenSearch Service

Simplify and accelerate your migration journey to Amazon OpenSearch Service

Seamless and efficient data migration

Migrate your data from self-managed Elasticsearch/OpenSearch to the Amazon OpenSearch Service managed clusters or Serverless collections without downtime and disruption

Assessment and validation

Run comparative performance and behavioral validation between the source and target environments based on your actual workloads

Monitoring & management capabilities

Monitor and manage the migration process using out-of-the-box dashboards

Open source and customizations

All code is open source enabling you to tailor and customize for your needs

Migration Assistant use cases

- Migrate existing cluster data to Amazon OpenSearch Service
- Migrate live data with traffic capture and replay
- Assess new versions under real workloads prior to upgrading
- Compare performance and behavior between source and target
- Discover optimal hardware and sharding configuration for your workload

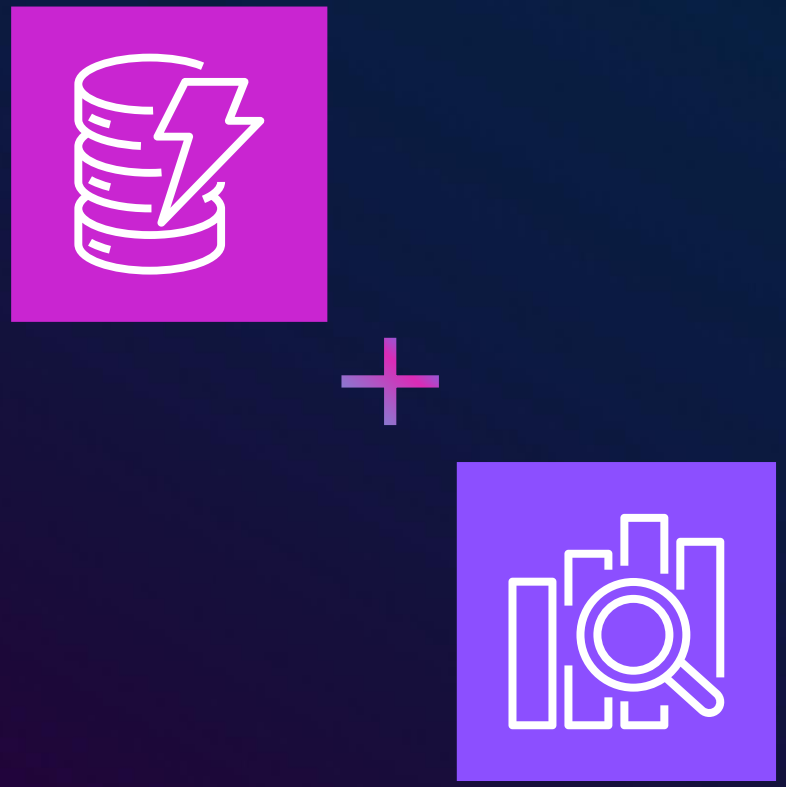
	Platform	Version
Source	Amazon EC2 / EKS	Elasticsearch 7.x (<=7.10), OpenSearch (all versions)
Target	Amazon OpenSearch Service managed clusters and Serverless collections	OpenSearch latest version (2.11) and Serverless

Learn more:

Zero-ETL integration with DynamoDB

NEW

ONE-CLICK DATA SYNCHRONIZATION FROM DYNAMODB



- Select your table and fields right from the AWS console
- Eliminate the need for managing complex ETL
- Quickly build search applications on your data
- Automatically generate embeddings for semantic search

Integrated semantic and hybrid search

SEARCH BY MEANING, NOT BY WORDS

Hybrid search

NEW

Blend neural and lexical search for results better than either technique alone

Fine tuned models

NEW

Fine tune your embedding model with your corpus, even without user behavior data

Sparse vector retrieval

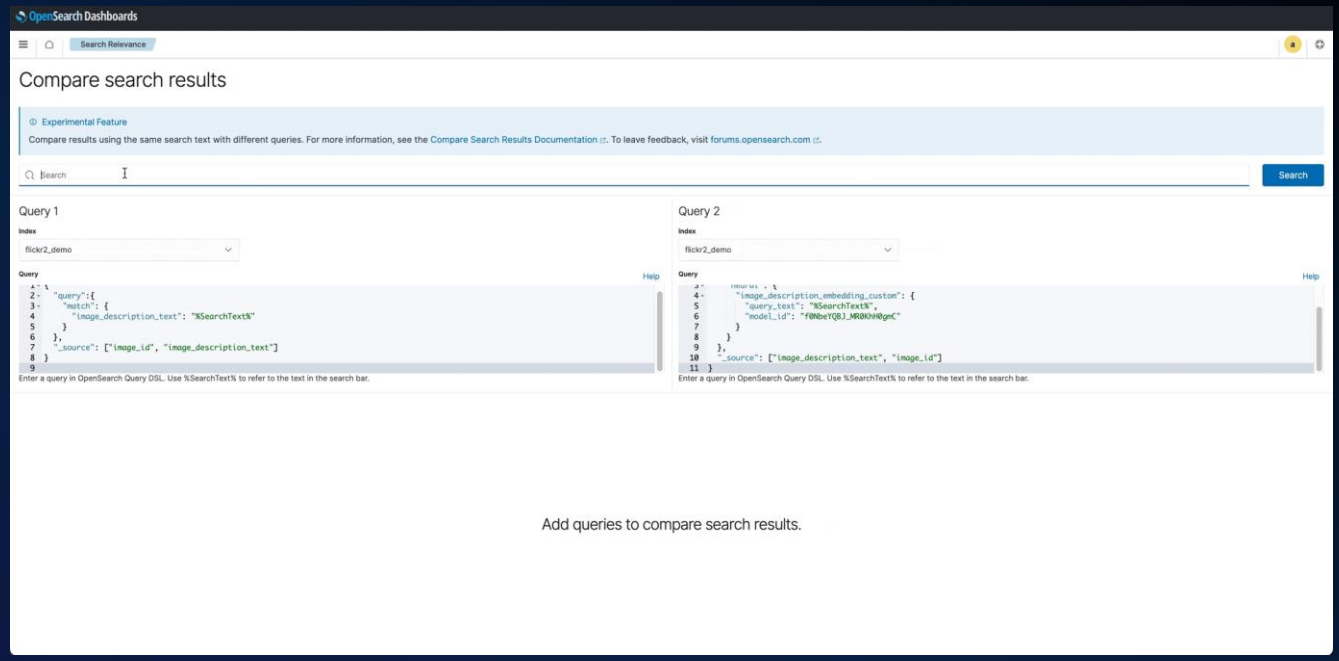
NEW

Semantic understanding of AI models with performance and cost of lexical search

Multimodal search

NEW

Search across images and text without pre-processing or labeling



Connect with ML models to power Neural Search

Integration via AWS Management Console

NEW

SageMaker text embedding models

Amazon Bedrock Titan Text Embedding model

Sparse Encoder model through SageMaker

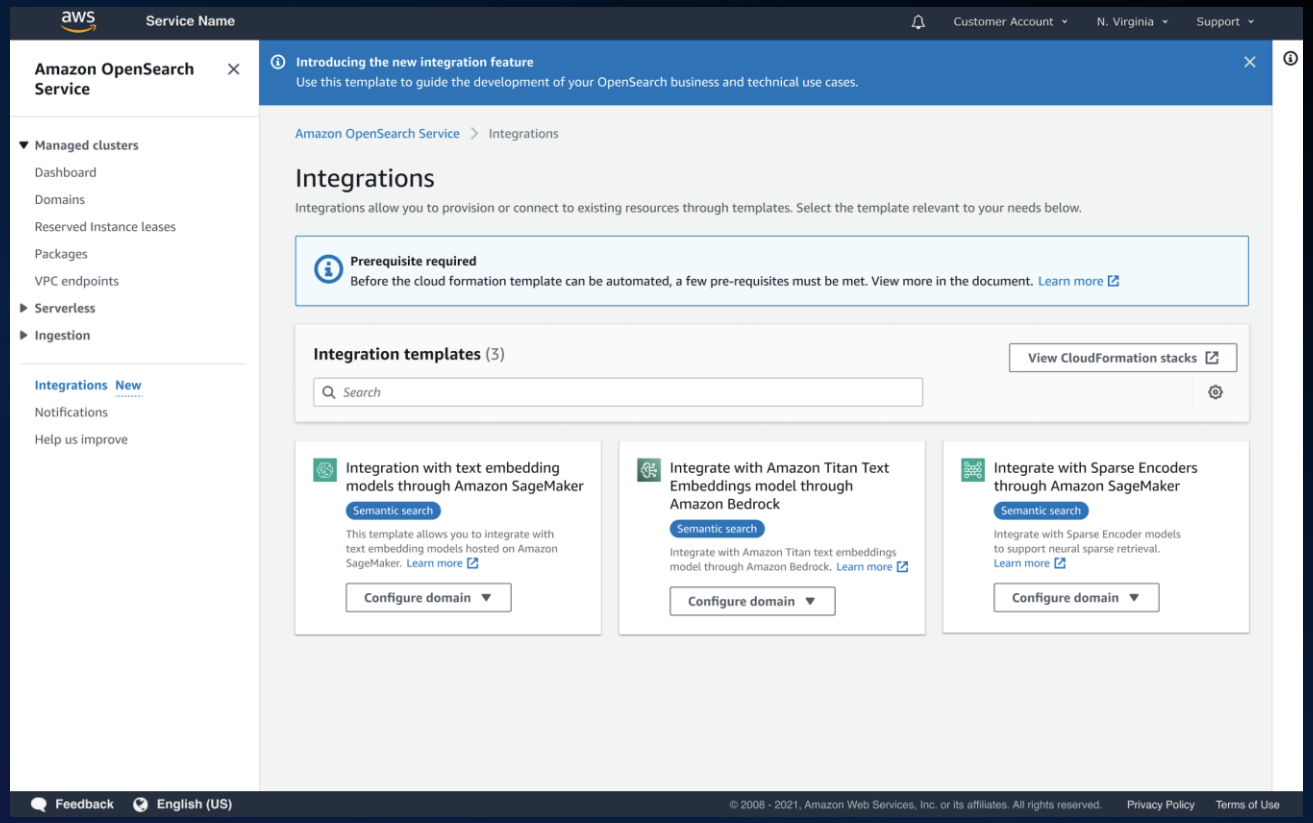
Integration via API

NEW

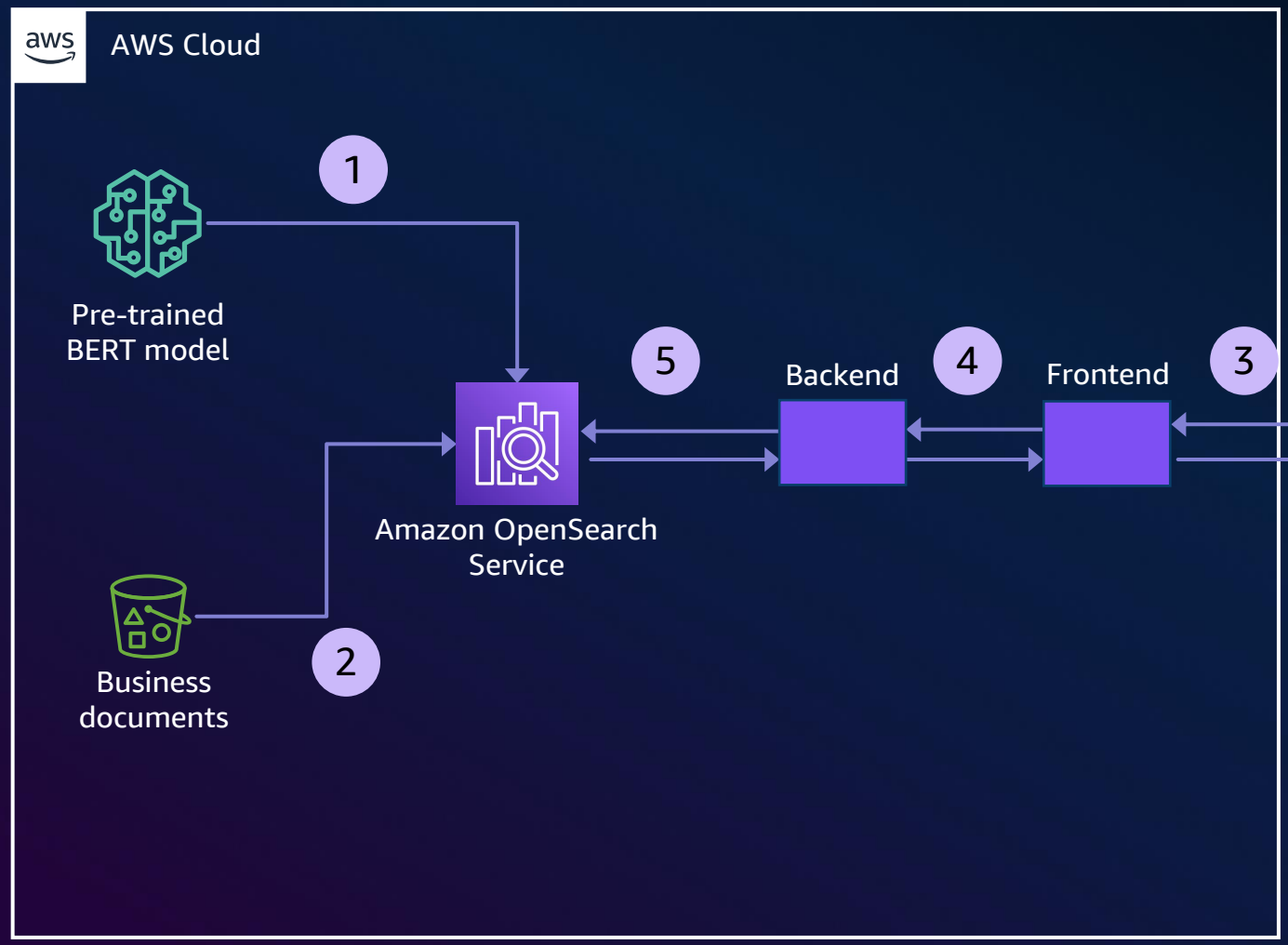
Cohere Embed text embedding model

Amazon Bedrock Titan Multi-modal model

Third-party partner models



Semantic search (neural search plugin)



1. Create a connection to a 3P model hosting service
2. Run neural search pipeline to ingest documents into OpenSearch Service
3. Client submits a search request to API Gateway
4. Amazon API Gateway calls AWS Lambda backend service in Lambda
5. Backend service calls neural search API to get similar documents and return to client

Neural plugin

Load model, then

```
{
  'settings': { 'index.knn': True },
  'mappings': {
    'properties': {
      "plot_text": { "type": "text" },
      "plot_embedding": {
        "type": "knn_vector",
        "dimension": 384
      },
      "year": { "type": "integer" },
      "rank": { "type": "integer" },
      "rating": { "type": "float" },
      "running_time_secs":
        { "type": "integer" },…
```

Map

```
{
  "processors" :
  [
    {
      "text_embedding":
      {
        "model_id": "<model id>",
        "field_map": {
          "plot": "plot_embedding"
        }
      }
    }
  ]
}
```

Pipeline and ingest

```
{
  "query": {
    "bool": {
      "should": [{
        "neural": {
          "plot_embedding": {
            "query_text": "{{query}}",
            "model_id": "<<MODEL_ID>>",
            "k": 10
          }
        }
      }]
    }
  }
}
```

Query

Zero-ETL integration with Amazon S3

PREVIEW

BRINGING POWERFUL OPERATIONAL ANALYTICS TO THE DATA LAKE

Interactive analytics on your data in S3

Queries data where it rests with minimal duplication via zero-ETL

Optionally boost query performance using acceleration

Out of the box visualizations using OpenSearch Dashboards

Improved Operations

Auto-Tune

EventBridge Events

Auto-Tune Events
Software Update Events

In-place updates

In-place updates for cluster-manager

Self-healing

Off-peak hours

Scheduling for software updates

Self-service

Self-service node restart
Self-service process restart

Core Technology

How search engines work - interaction

1

Send data as JSON via REST APIs

2

Data is indexed—all fields searchable, including nested JSON

3

REST APIs, for fielded matching, Boolean expressions, sorting, and analysis

Aggregations

```
GET weblogs/_search
{
  "size": 0,
  "aggs": {
    "Status": {
      "terms": {
        "field": "response.status",
        "size": 10
      }
    }
  }
}
```

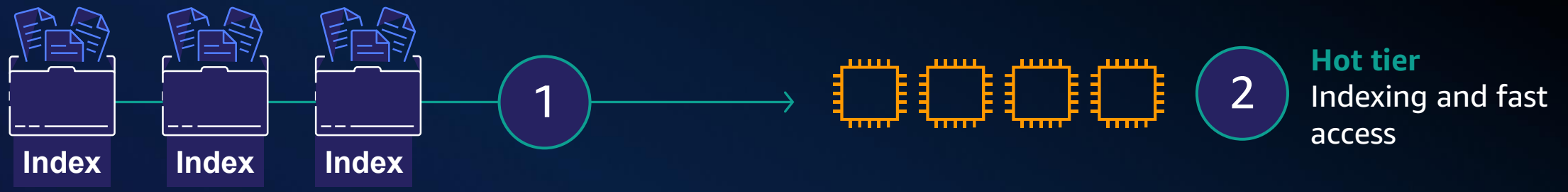
```
...
"aggregations" : {
  "response_codes" : {
    "buckets" : [
      {
        "key" : "200",
        "doc_count" : 12832
      },
      {
        "key" : "404",
        "doc_count" : 801
      },
      {
        "key" : "503",
        "doc_count" : 441
      }
    ]
}}}
```

Data Management

Data lifecycle in Amazon OpenSearch Service

1 Send data to Amazon OpenSearch Service and use Index State Management (ISM) to automate index migrations or deletions

Data lifecycle in Amazon OpenSearch Service

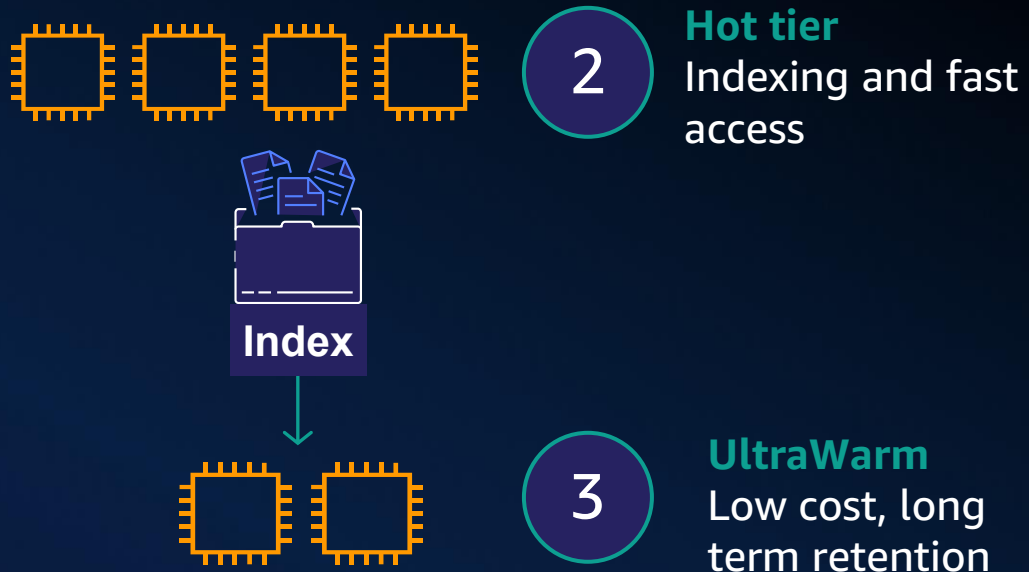


1 Send data to Amazon OpenSearch Service and use Index State Management (ISM) to automate index migrations or deletions

2 Data is indexed and stored in the hot tier

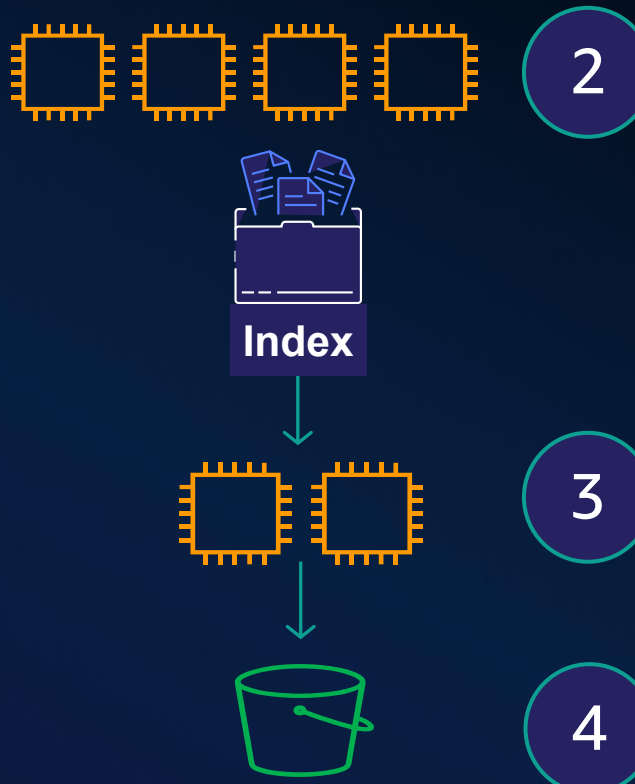
Data lifecycle in Amazon OpenSearch Service

1. Send data to Amazon OpenSearch Service and use Index State Management (ISM) to automate index migrations or deletions
2. Data is indexed and stored in the hot tier
3. Migrate the index to UltraWarm storage for long-term, low cost storage



Data lifecycle in Amazon OpenSearch Service

1. Send data to Amazon OpenSearch Service and use Index State Management (ISM) to automate index migrations or deletions
2. Data is indexed and stored in the hot tier
3. Migrate the index to UltraWarm storage for long-term, low cost storage
4. Store data in Cold Storage for longer-term, lowest cost storage



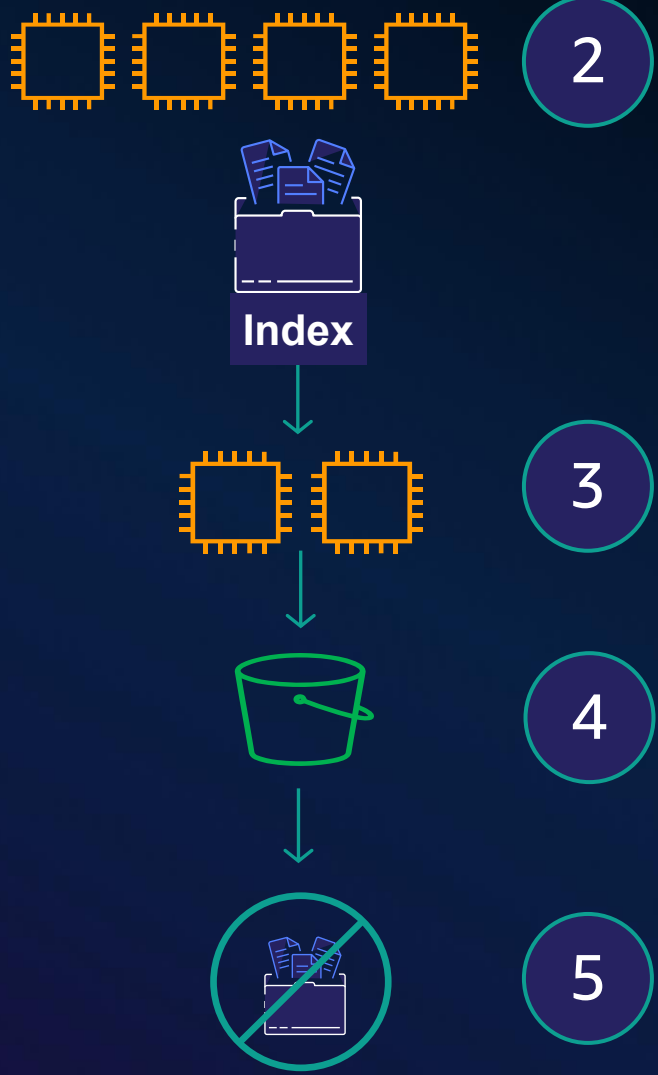
Hot tier
Indexing and fast access

UltraWarm
Low cost, long term retention

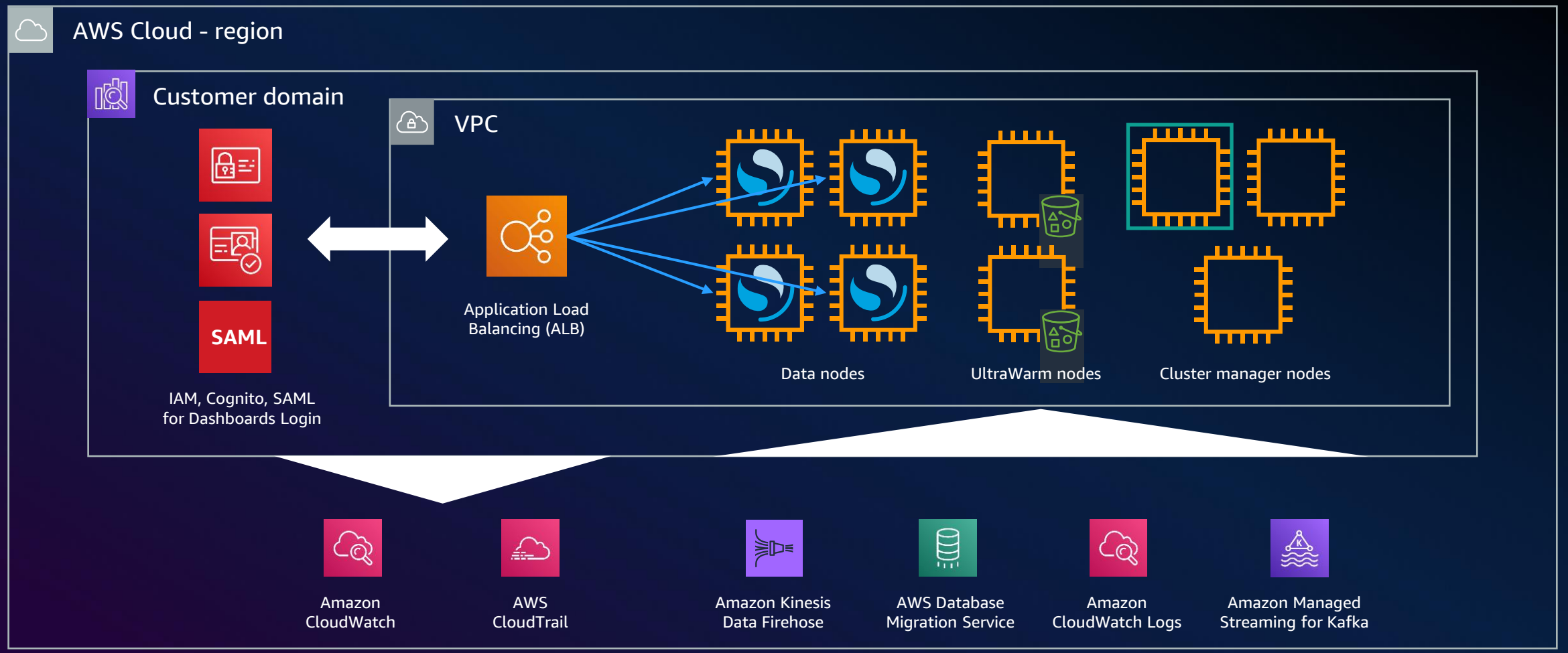
Cold Storage
Lowest cost, longer term retention and on-demand access

Data lifecycle in Amazon OpenSearch Service

1. Send data to Amazon OpenSearch Service and use Index State Management (ISM) to automate index migrations or deletions
2. Data is indexed and stored in the hot tier
3. Migrate the index to UltraWarm storage for long-term, low cost storage
4. Store data in Cold Storage for longer-term, lowest cost storage
5. Delete the index at end-of-life



Managed service deployment Architecture



Using OpenSearch for Search

SEARCH IS THE FOUNDATION FOR ALL USE CASES

OpenSearch is an information retrieval system

Getting the most relevant results for the requestor

Text search, faceting, geospatial, auto-complete, fuzzy matching

Make modern AI/ML technologies available to all OpenSearch users

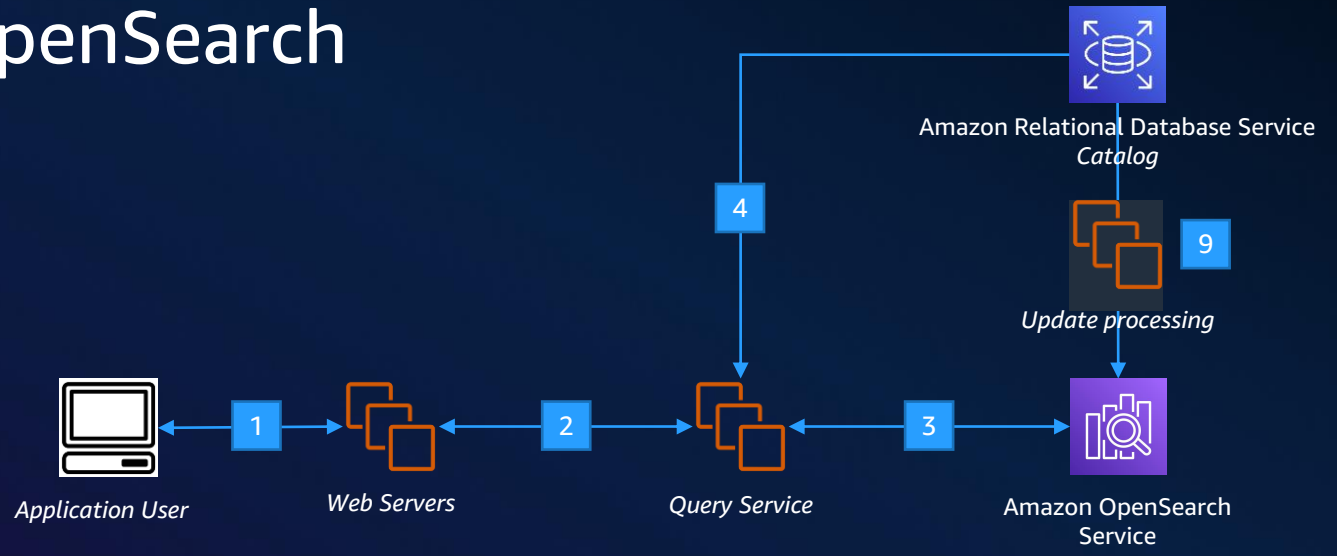
Search basic

Use OpenSearch in tandem with a durable store

1-3: Send queries, receive responses

4: Retrieve source records from durable store

5: Send changes to OpenSearch



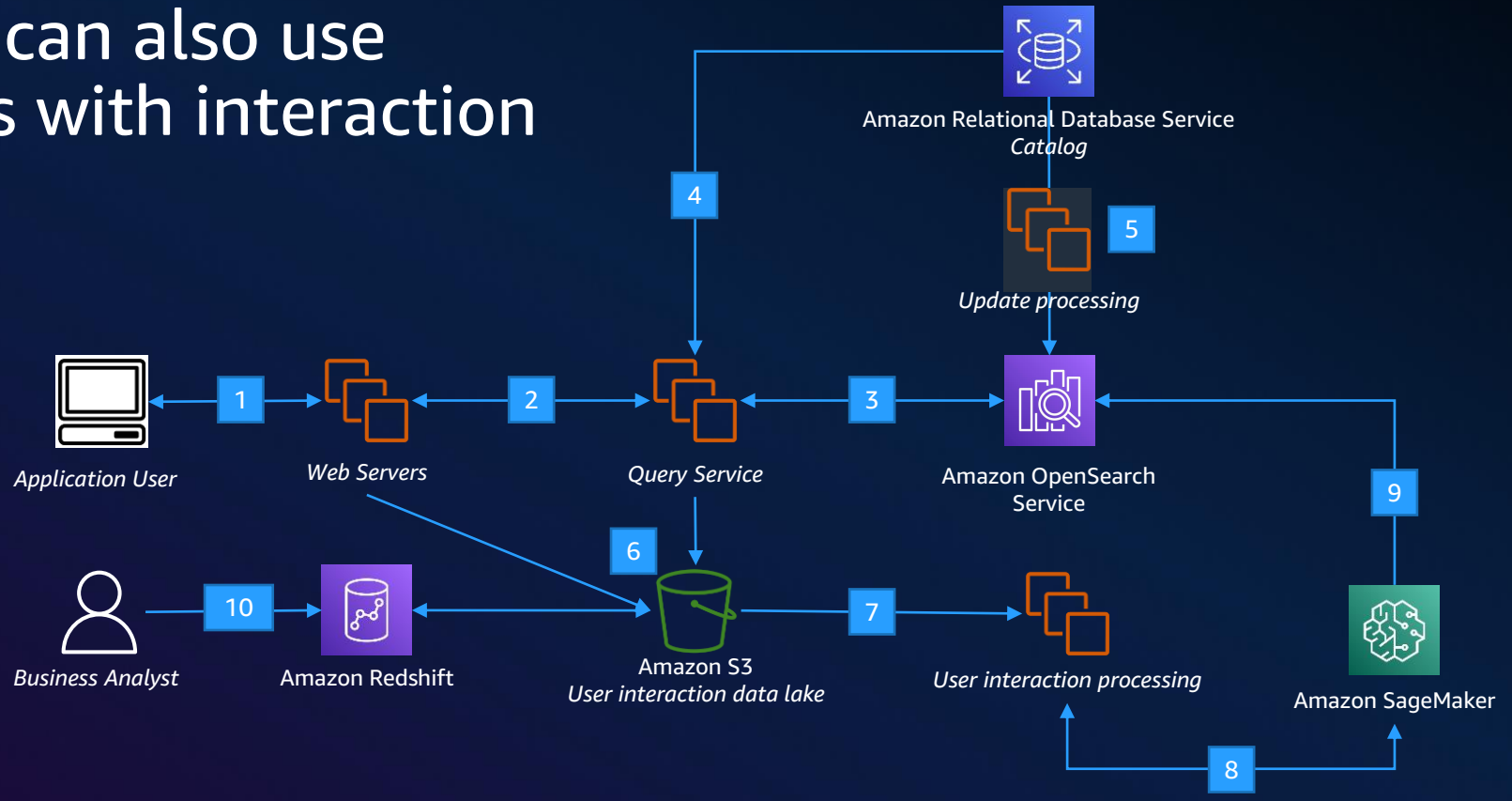
Search advanced

Gather and use user behavior to improve search relevance

6: Send interaction data to S3 (queries, results, clicks, purchases)

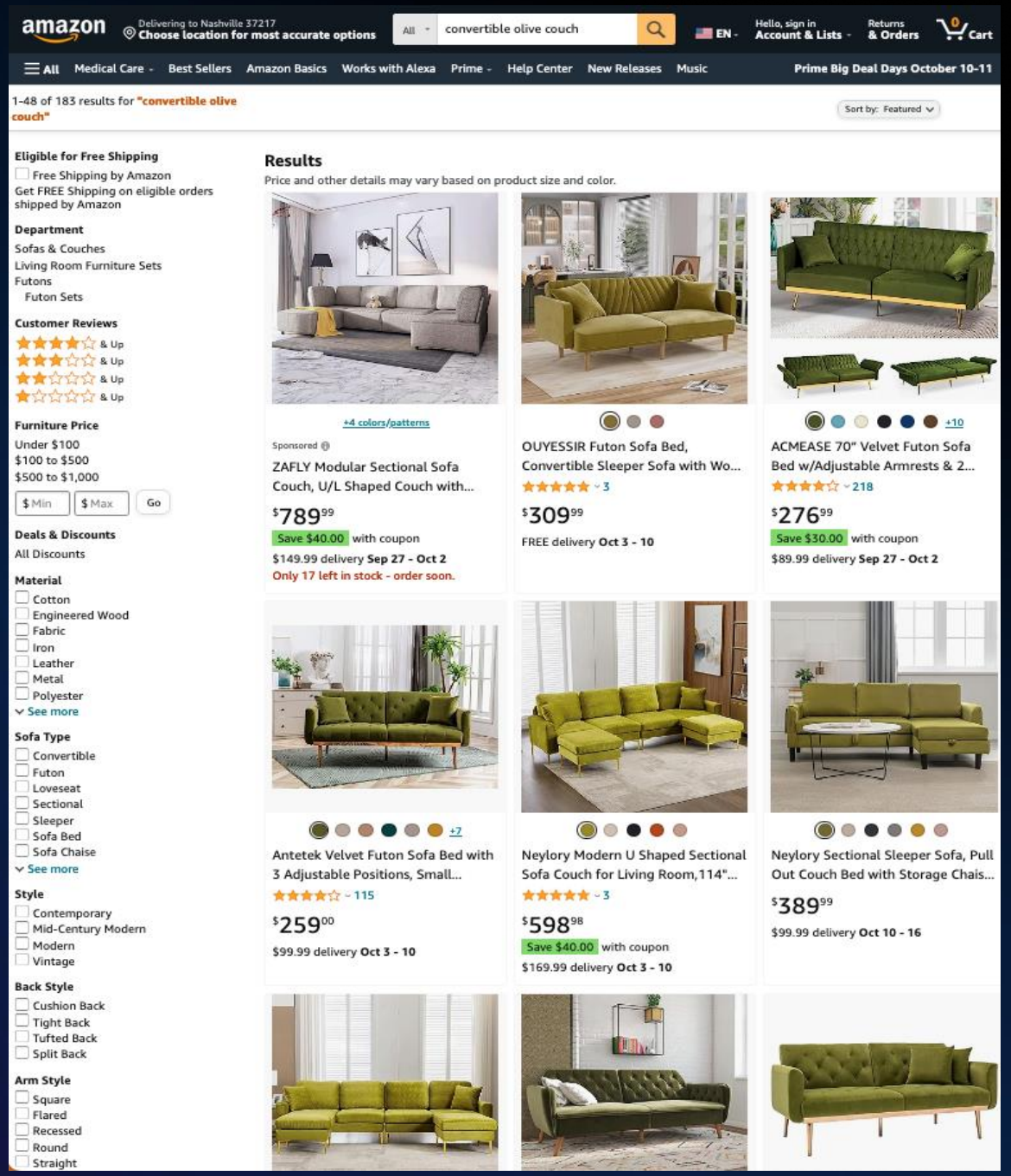
7-9: Build a model and send to OpenSearch

10: Business analysts can also use BI and other tools with interaction data in S3



Text Search

- Search engines match terms to objects in a catalog
- Match free text or structured data
- Relevance determines sort order
- Facets enable users to drill in to the search results to narrow by attribute value



OpenSearch ML Connector

STREAMLINING ML INTEGRATIONS

Simplify & operationalize vector hydration

Build native integrations with embedding services

Leverage vector search to power your generative AI applications

Secure and manage access to the ML models

Leverage the distributed design of OpenSearch to build a **stable, scalable** vector database

OpenSearch ANN supported algorithms

	NMSLIB-HNSW	FAISS-HNSW	FAISS-IVF	Lucene-HNSW
Max dimension	16,000	16,000	16,000	1024
Filter	Post-filter	Post-filter	Post-filter	Filter while search
Training required	No	No	Yes	No
Distance formula	l2, innerproduct, cosinesimil, l1, linf	l2, innerproduct	l2, innerproduct	l2, cosinesimil
Vector volume	Tens of billions	Tens of billions	Tens of billions	< Ten million
Indexing latency	Low	Low	Lowest	Low
Query latency & quality	Low latency & high quality	Low latency & high quality	Low latency & low quality	High latency & high quality
Vector compression	Flat	Flat Product Quantization	Flat Product Quantization	Flat
Memory consumption	High	High Low with PQ	Medium Low with PQ	High

OpenSearch Service vector search public content

OpenSearch Service's vector DB capabilities

OpenSearch Serverless vector engine

Semantic workshop

Vector search video

Multi-modal

Semantic benchmarks

Scaling for vectors

Building chatbots

aws partnercast

Thank you!

Please join us again for another PartnerCast session

https://aws.amazon.com/partners/training/partnercast/

Demo