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

PRT266

SPONSORED BY ELASTIC

The latest innovations from Elasticsearch: Drive speed, scale & relevance

Michael Heldebrant, PhD

Distinguished Architect
Elastic

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

Agenda

Relevance innovations powered by vector search

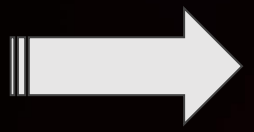
- The power of vector search
- How to implement
- Common use cases
- Personalization
- Ingestion

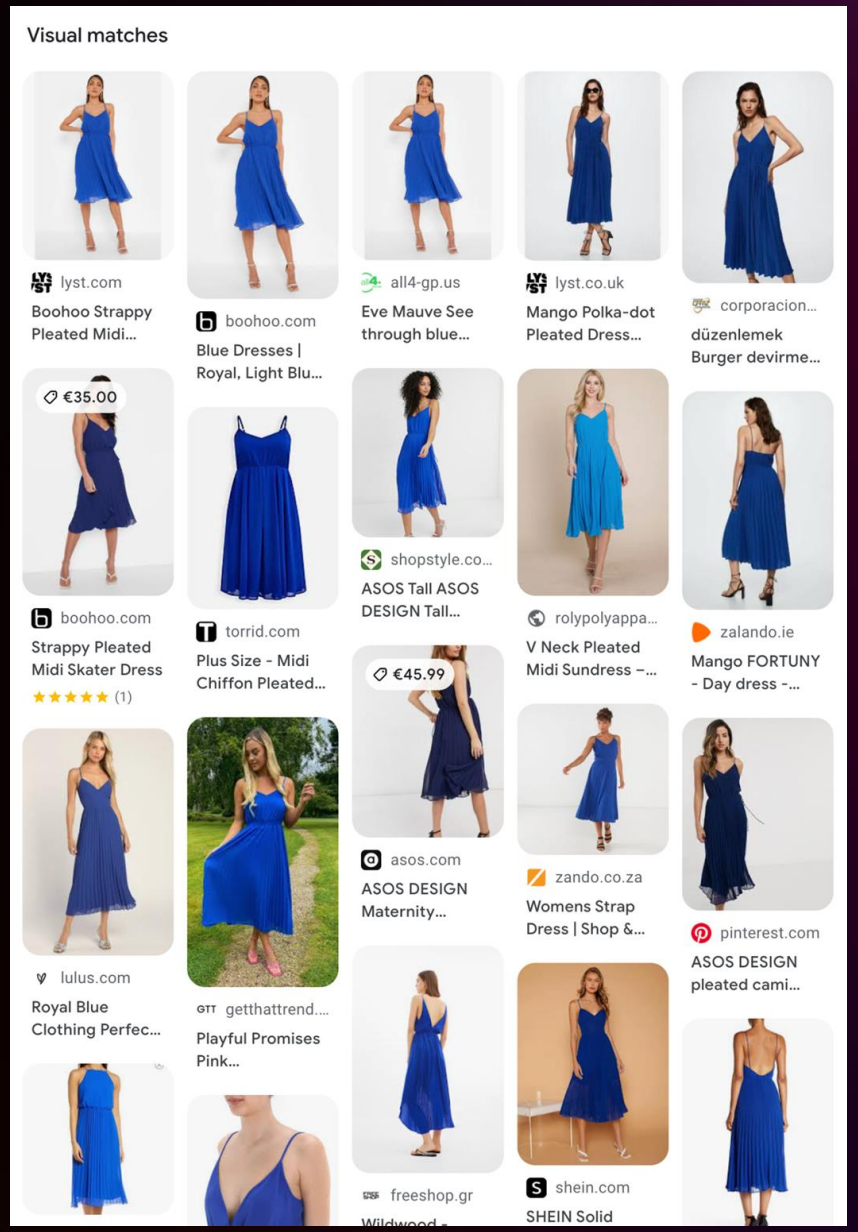
Performance improvements

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

Wanna copy a famous influencer?







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

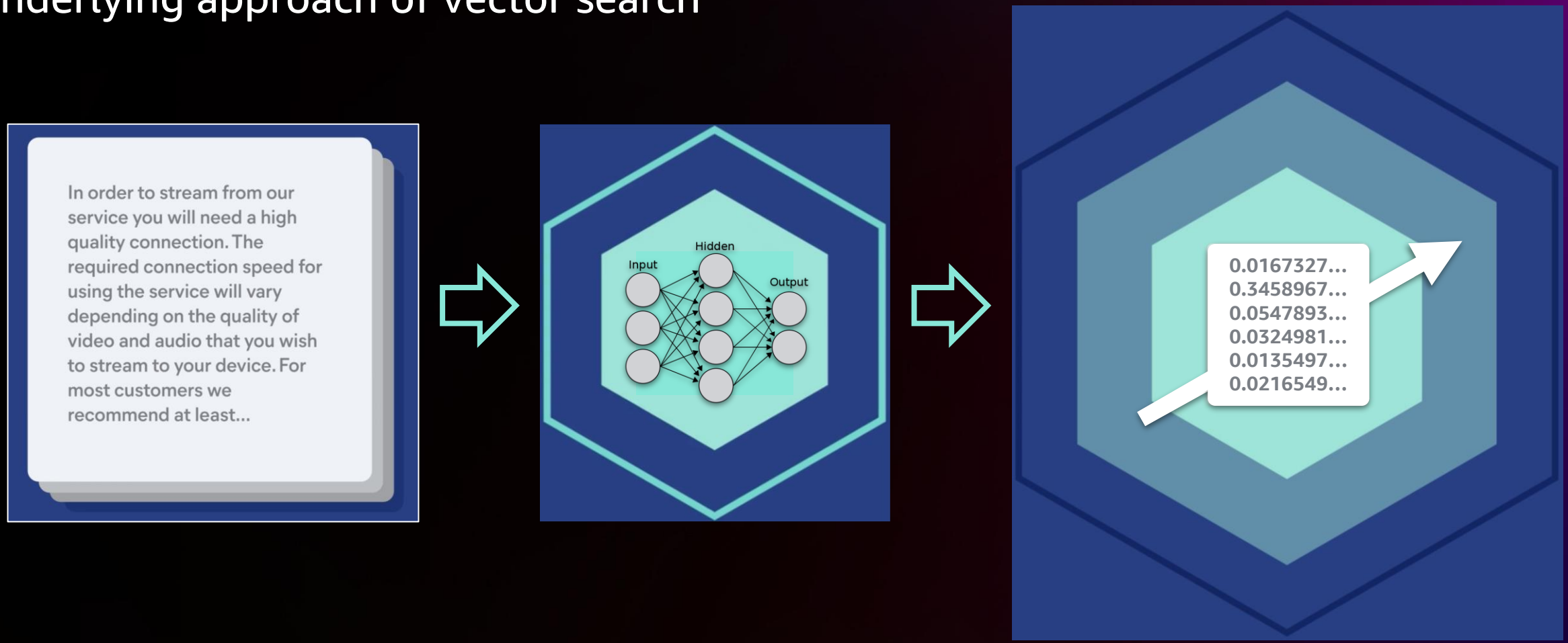
Vector search lets you find what you mean, with higher relevance

- **Semantic search**: understand meaning
- **Multimedia**: not just text, also pictures or sounds
- **Precision**

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

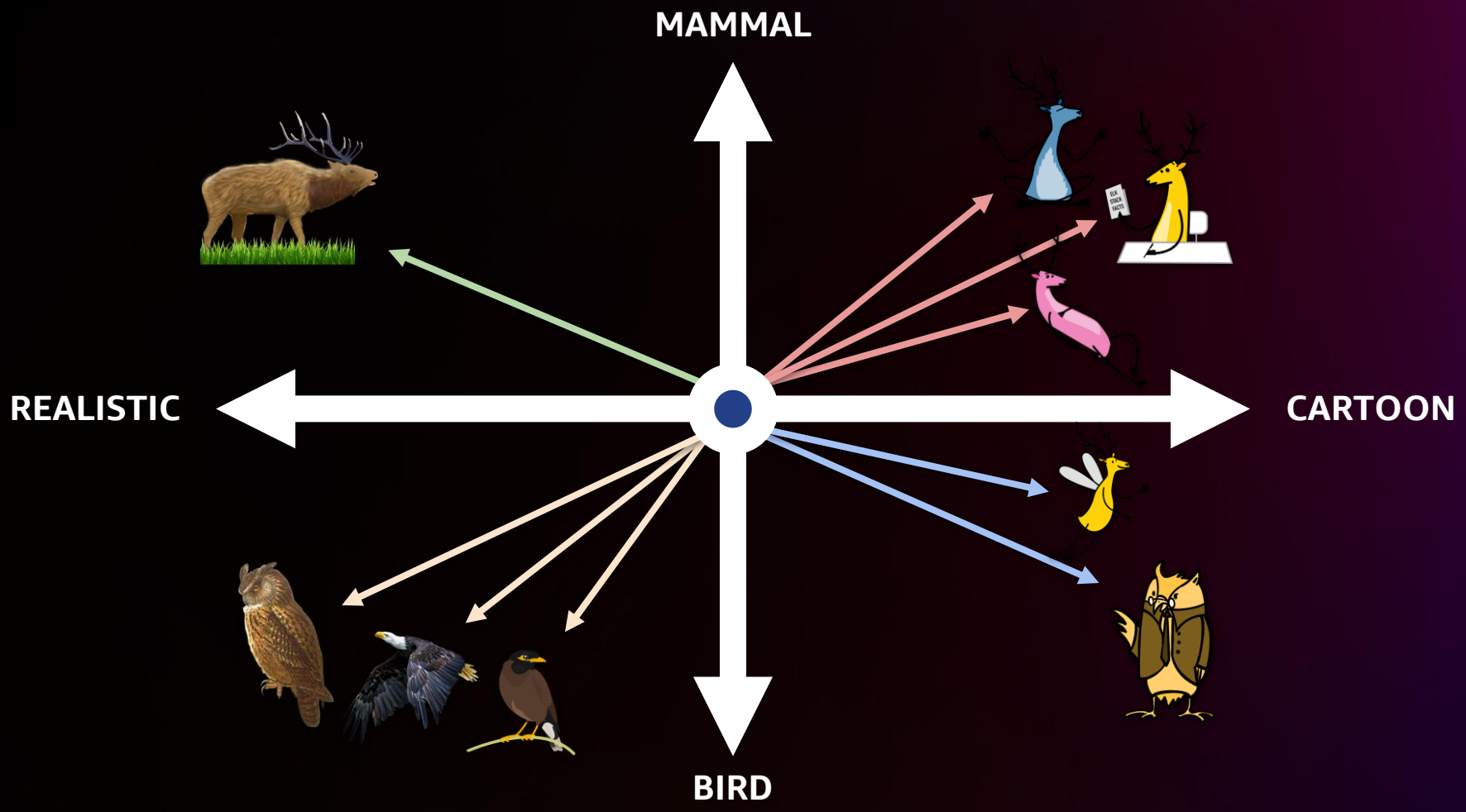
Convert data into vector representation

Underlying approach of vector search



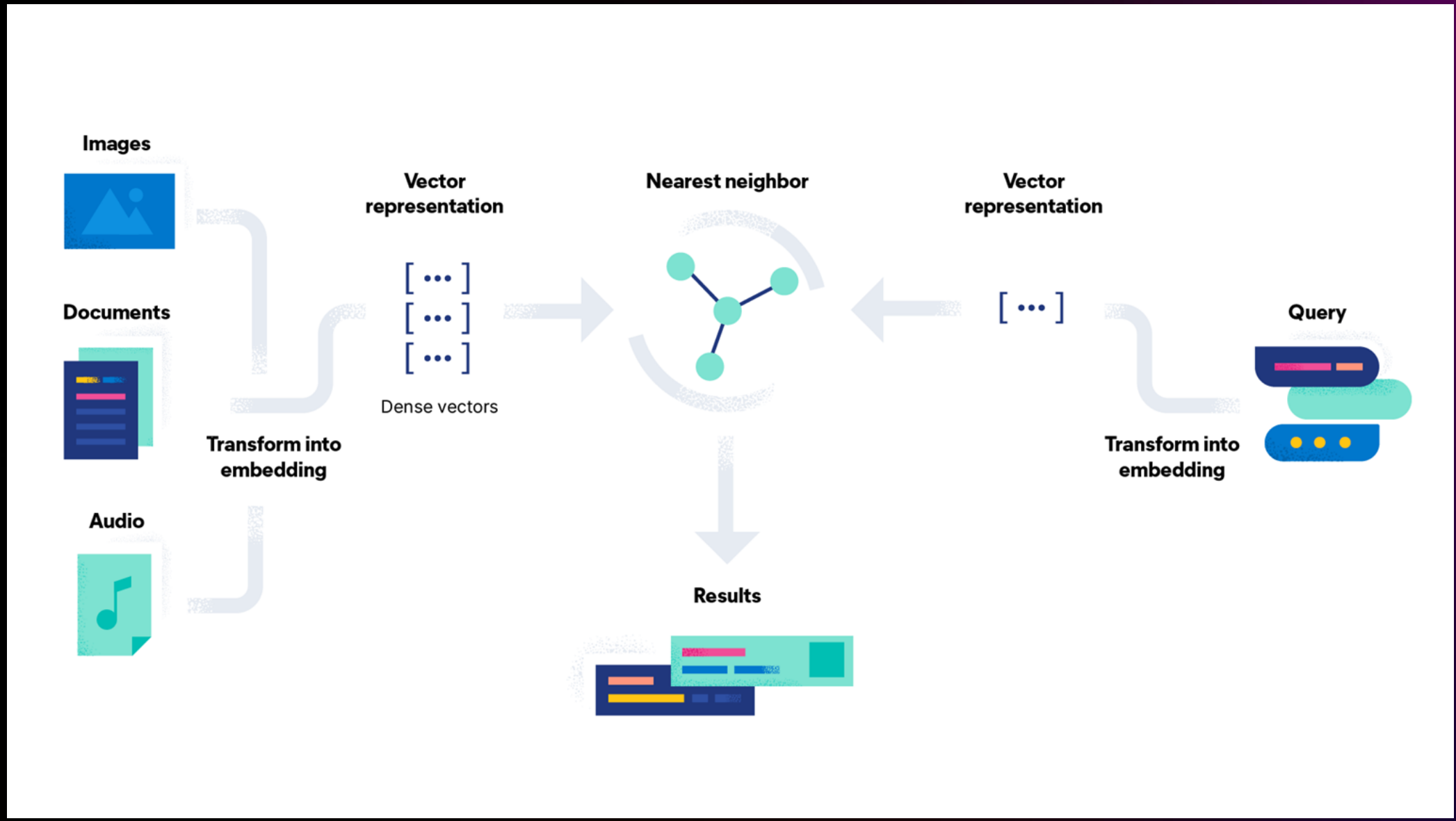
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

In vector space, similar data are nearby



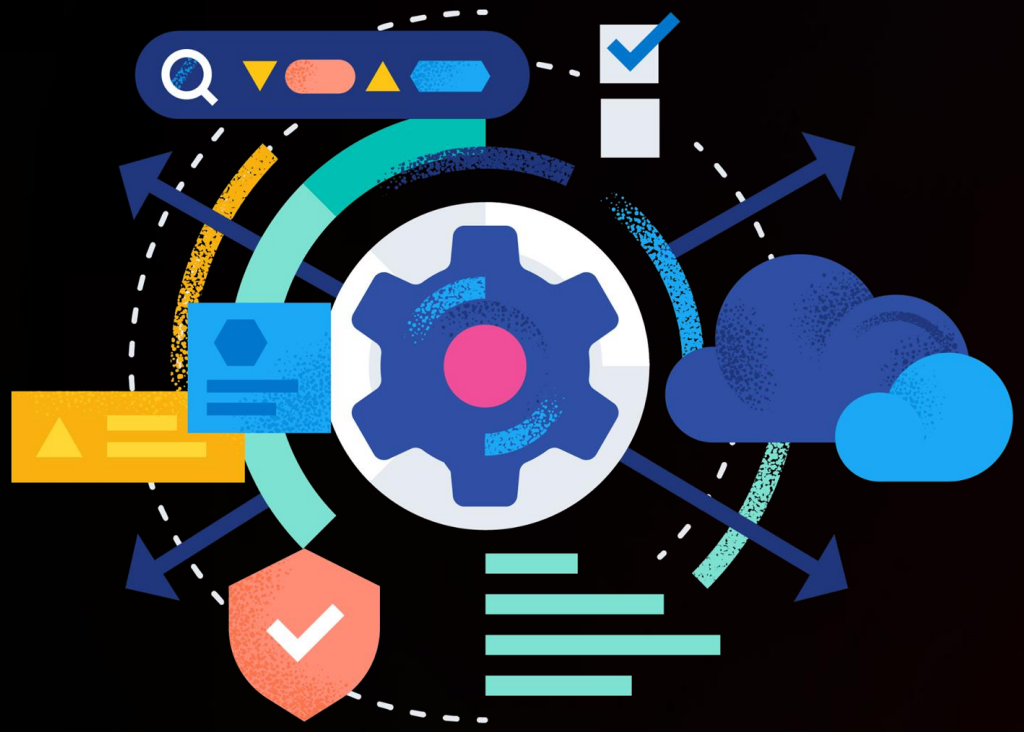
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Vector search returns nearest neighbors of “vectorized” query



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

Vector search and NLP may be intimidating



Lots of data, labeled

Expertise

Scale processing

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

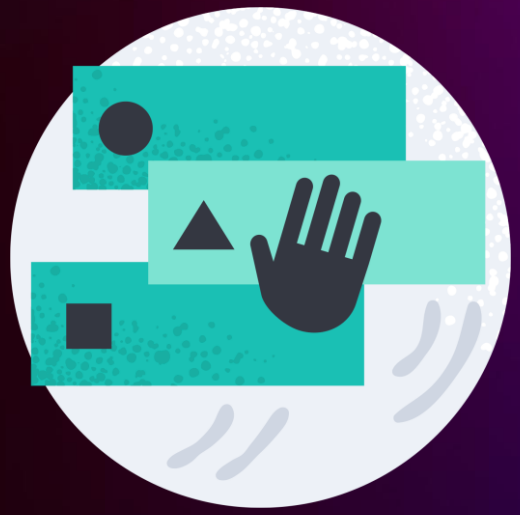
Elastic makes Vector search easy



Just bring the data



Inherent scalability

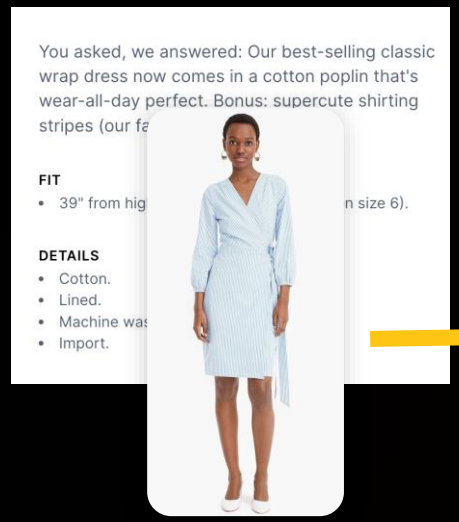


Flexibility

aws

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

Apply vector search in 3 steps



Source data

elastic cloud

2

POST /_doc

Stored in Elasticsearch:

```
{
  "_id":"product-1234",
  "product_name":"Dress",
  "description":"Our best…",
  "Price": 118,
  "color":"blue",
  "fabric":"cotton",
  "embedding":[0.452,0.3242,...]
}
```

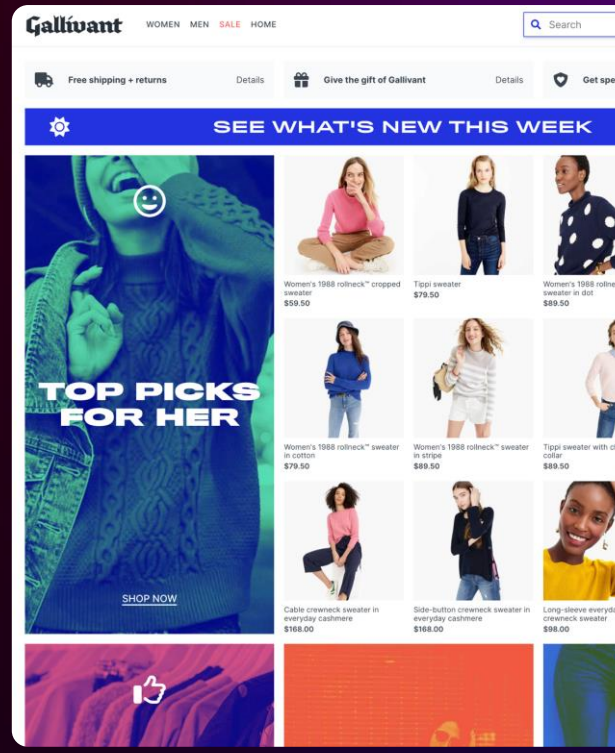
Transformer model

1

3

GET /_search

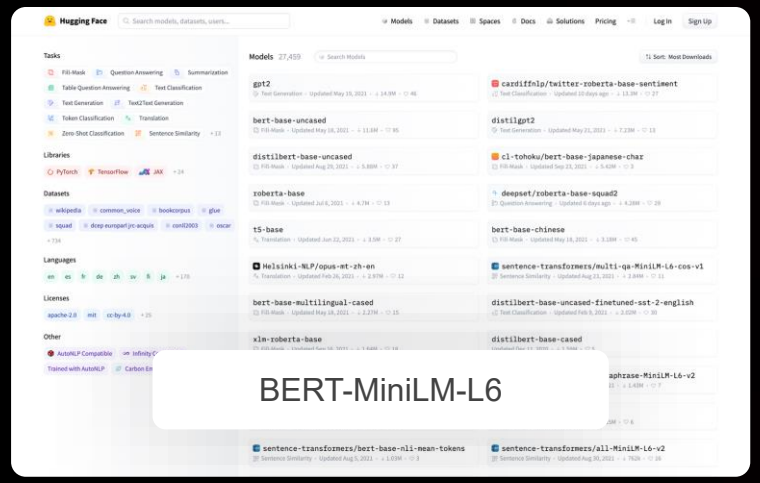
with kNN clause



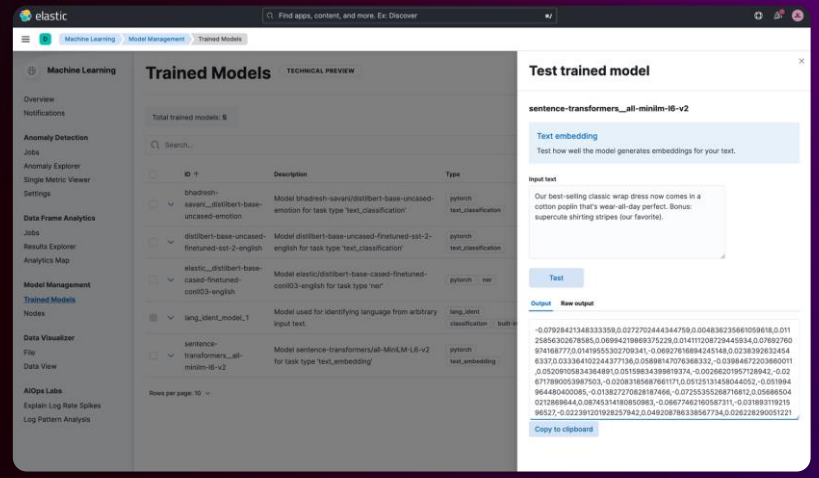
Search-powered application

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

Step 1: Import pre-trained model



$ eland_import_hub_model
--url https://cluster_URL --hub-model-id BERT-MiniLM-L6
--task-type text_embedding --start



PyTorch

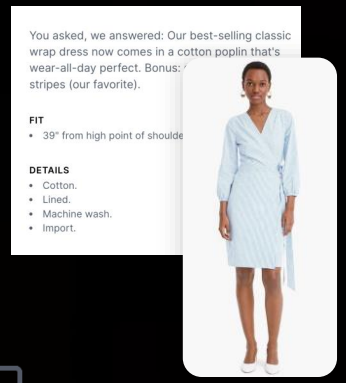
eland

Load the model to the cluster



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

Step 2: Vectorize during data ingestion



Source data

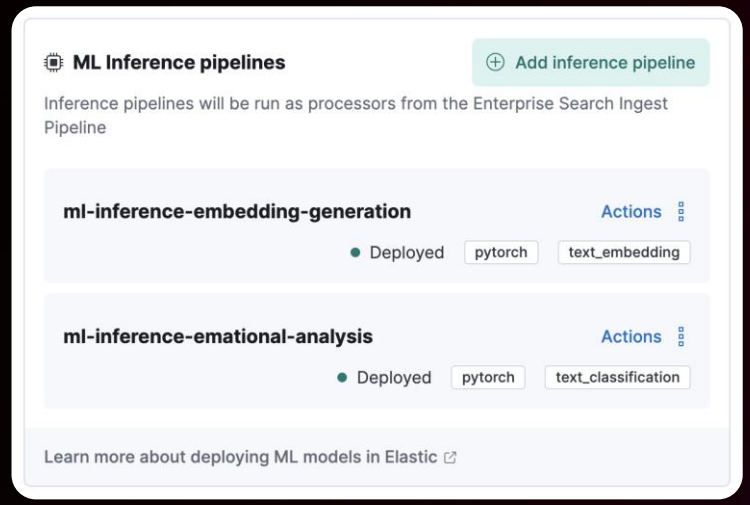
POST /_doc

Standard field indexing for non-vector types

```
{
  "_id":"product-1234",
  "product_name":"Summer Dress",
  "description":"Our best-selling…",
  "Price": 118,
  "color":"blue",
  "fabric":"cotton",
} "desc_embedding":[0.452,0.3242,…]
}
```

POST /_doc

Encoding via inference processor



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

Step 3: Find nearest neighbor, approximately

Query is submitted to the search-powered application:



Generate embedding:

```
POST /_ml/trained_models/my-model/_infer

{
  "docs": {
    "description": "summer clothes"
  }
}
```



Transformer model

PyTorch

Issue kNN query using the _search endpoint

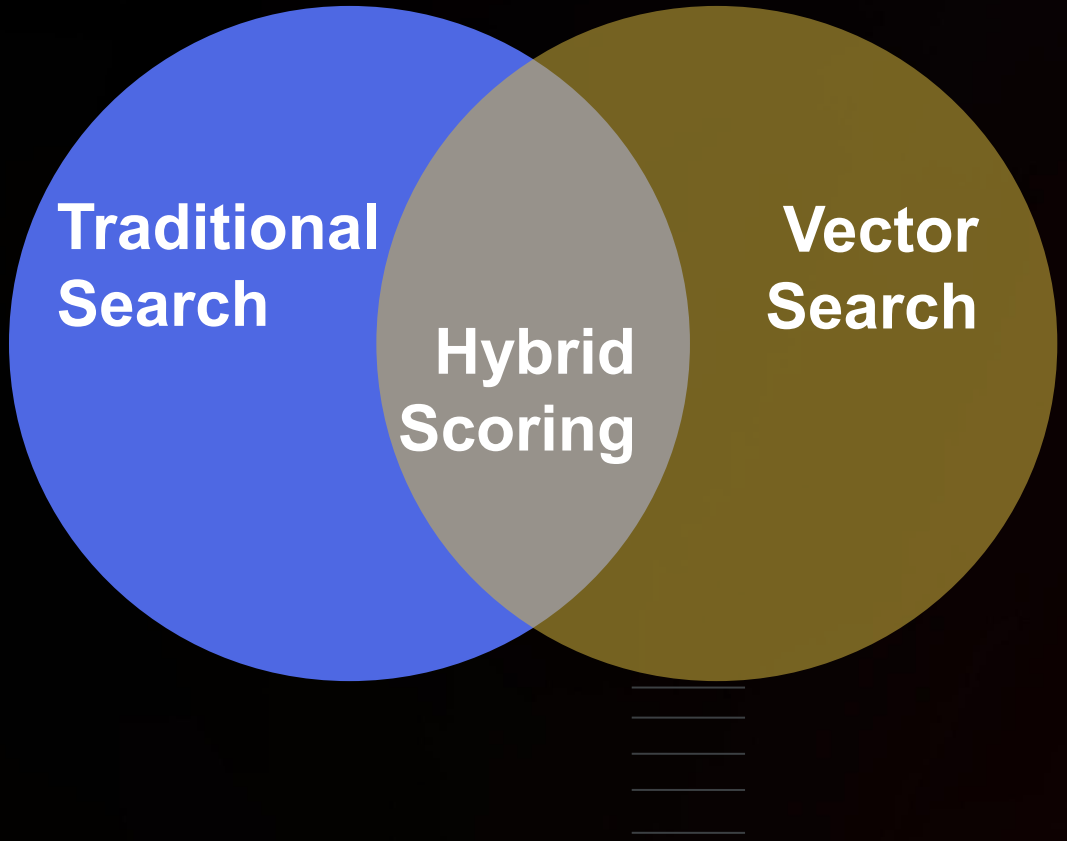
```
GET product-catalog/_search

{
  "knn": {
    "field": "desc_embbeding",
    "query_vector": [0.123, 0.244,...],
    "k": 5,
    "num_candidates": 50,
    "boost": 0.1,
    "filter": {
      "term": {
        "department": "women"
      }
    }
  }
}
```

aws

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

Hybrid scoring gets you the best of both worlds




```
GET product-catalog/_search
{
  "query": {
    "match": {
      "description": {
        "query": "summer clothes",
        "boost": 0.9
      }
    }
  },
  "knn": {
    "field": "desc_embbeding",
    "query_vector": [0.123, 0.244,...],
    "k": 5,
    "num_candidates": 50,
    "boost": 0.1,
    "filter": {
      "term": {
        "department": "women"
      }
    }
  }
}
```

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

How will you apply vector search?

Product similarity search

"Do you sell black v-neck shirts that look like this?"



Answer technical support

"What are the troubleshooting steps for ___?"

Query medical knowledge

"Is lithium used to treat bipolar disorder?"

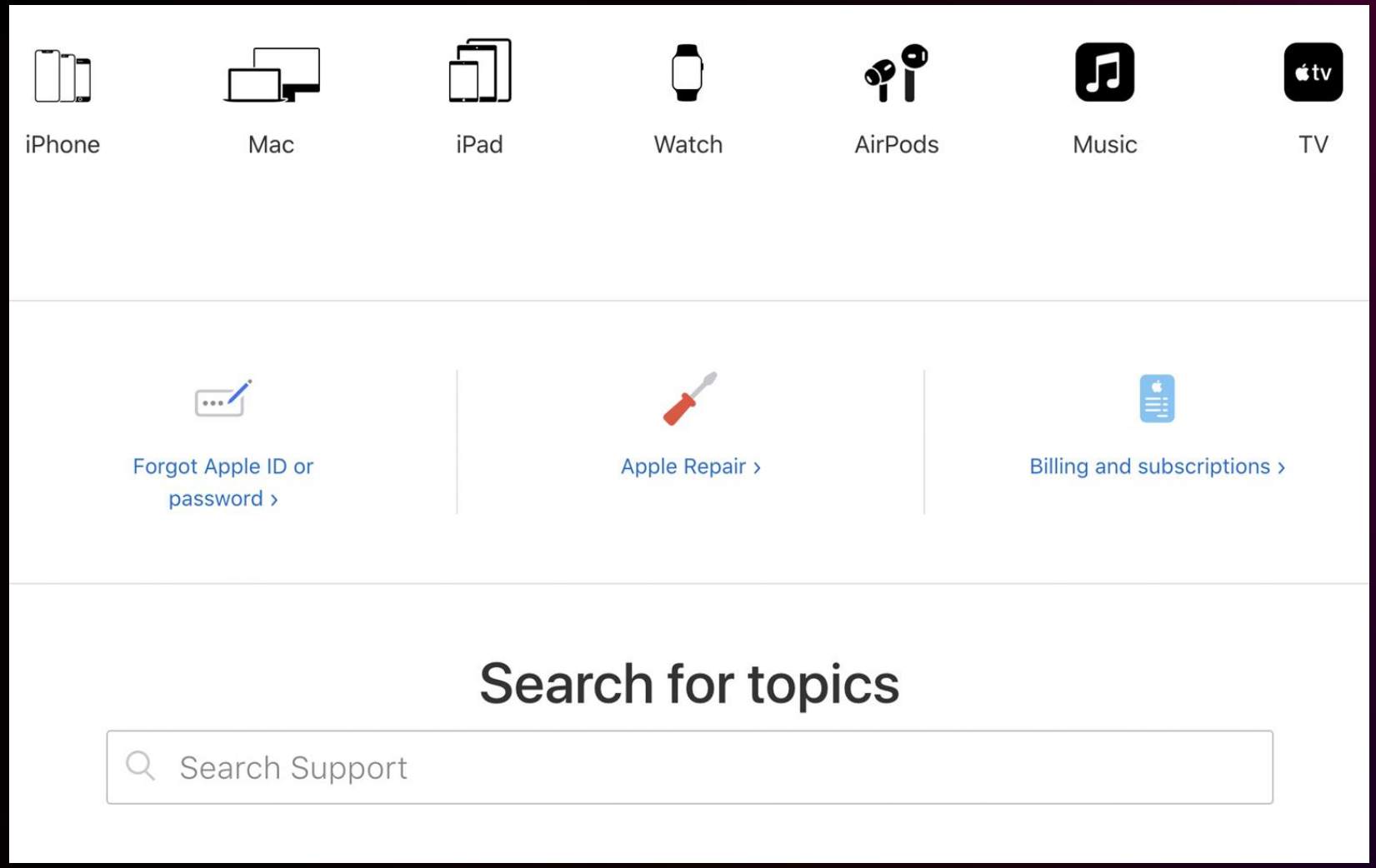
React to user sentiments

Identify poor customer interactions before they lead to escalations

aws

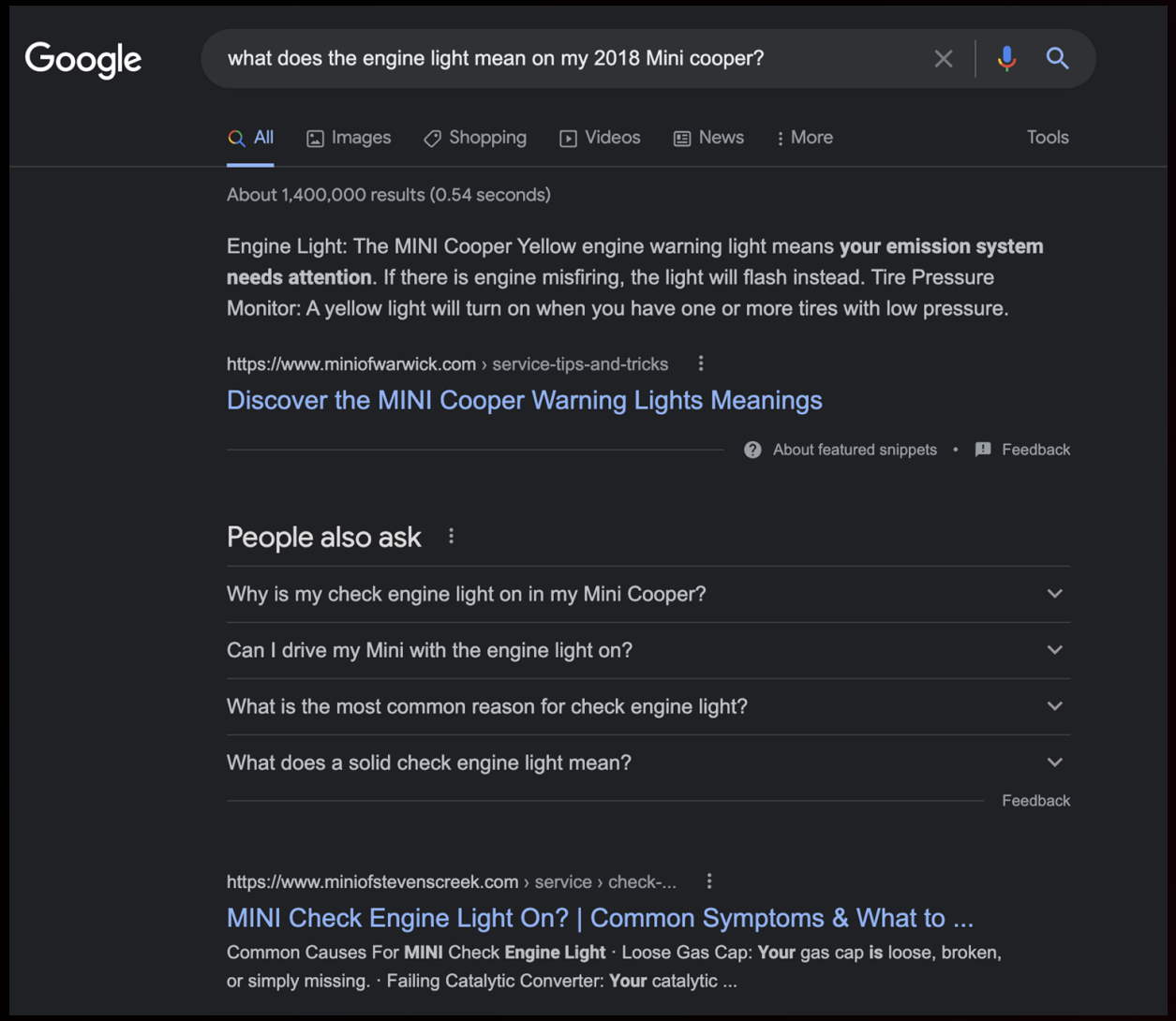
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

NLP #1: No more browsing manuals/FAQs



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

Query with question-answering model




```
POST _ml/trained_models/deepset__minilm-uncased-squad2/_infer
{
  "docs":[{ "text_field": "My name is Peter and I live in London"}],
  "inference_config": {
    "question_answering": {
      "question": "Where do I live?"
    }
  }
}
```

```
{
   "inference_results": [
       {
           "predicted_value": "London",
           "start_offset": 31,
           "end_offset": 37,
           "prediction_probability": 0.9948918266325819
       }
   ]
}
```

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

NLP #2: Prevent customer service escalations

Possibly one of the worst service experiences at [redacted] We checked in at 5pm and were seated for about 20 minutes and no waiter. Not busy at all. We saw a table being seated 10 minutes after us that was served right away. We had to go back to host desk to ask for our waiters name and host was flustered at best.

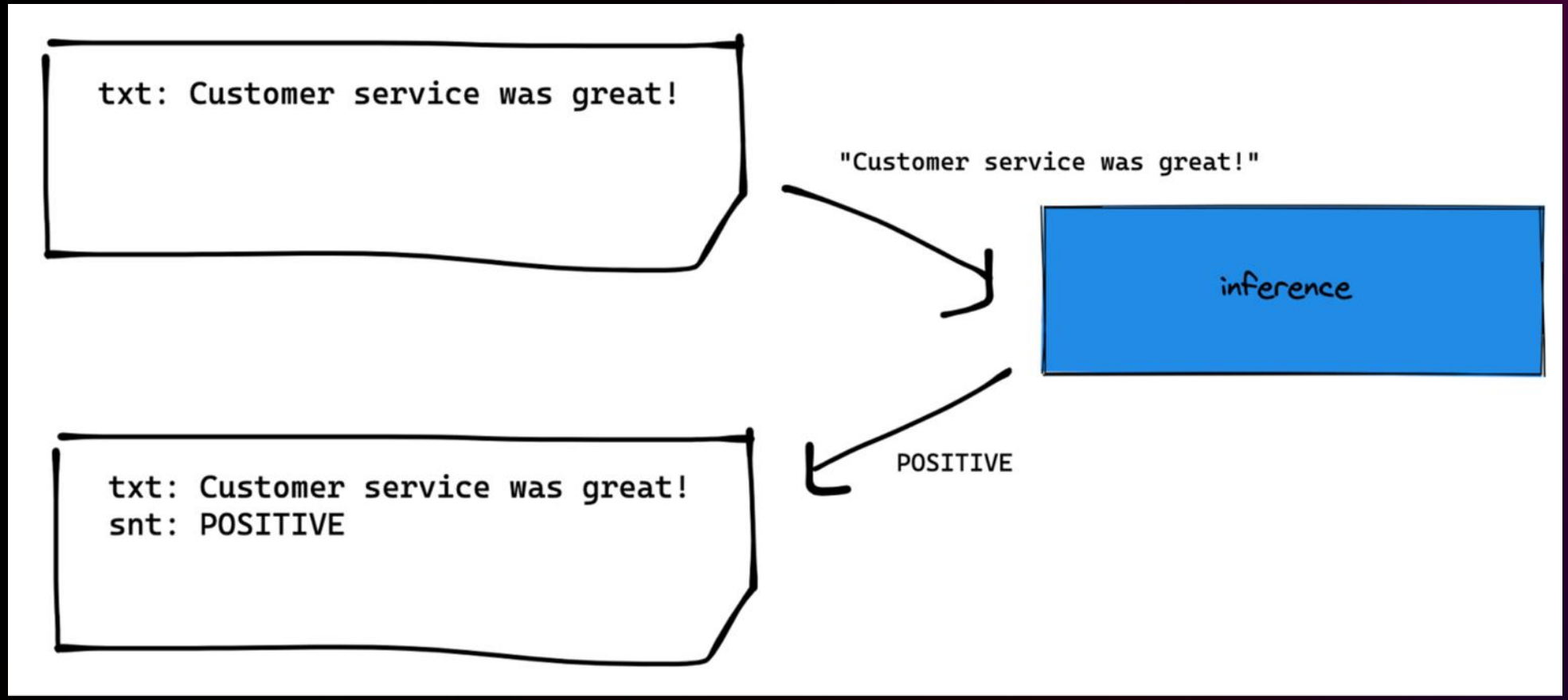
I received a text from Open Table saying I missed my reservation almost an hour after we checked in. On my way out I asked them to check their system and note that they would have a table ready for me shortly. I said no just check m showed them my receipt.



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

Sentiment models classify feedback

Monitor your brand in social media, get notified after feedback!



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

Personalization drives digital commerce outcomes

Online shoppers today

88%

are more likely to continue shopping on websites that offer a personalized experience

84%

report personalization already influences their shopping decisions

68%

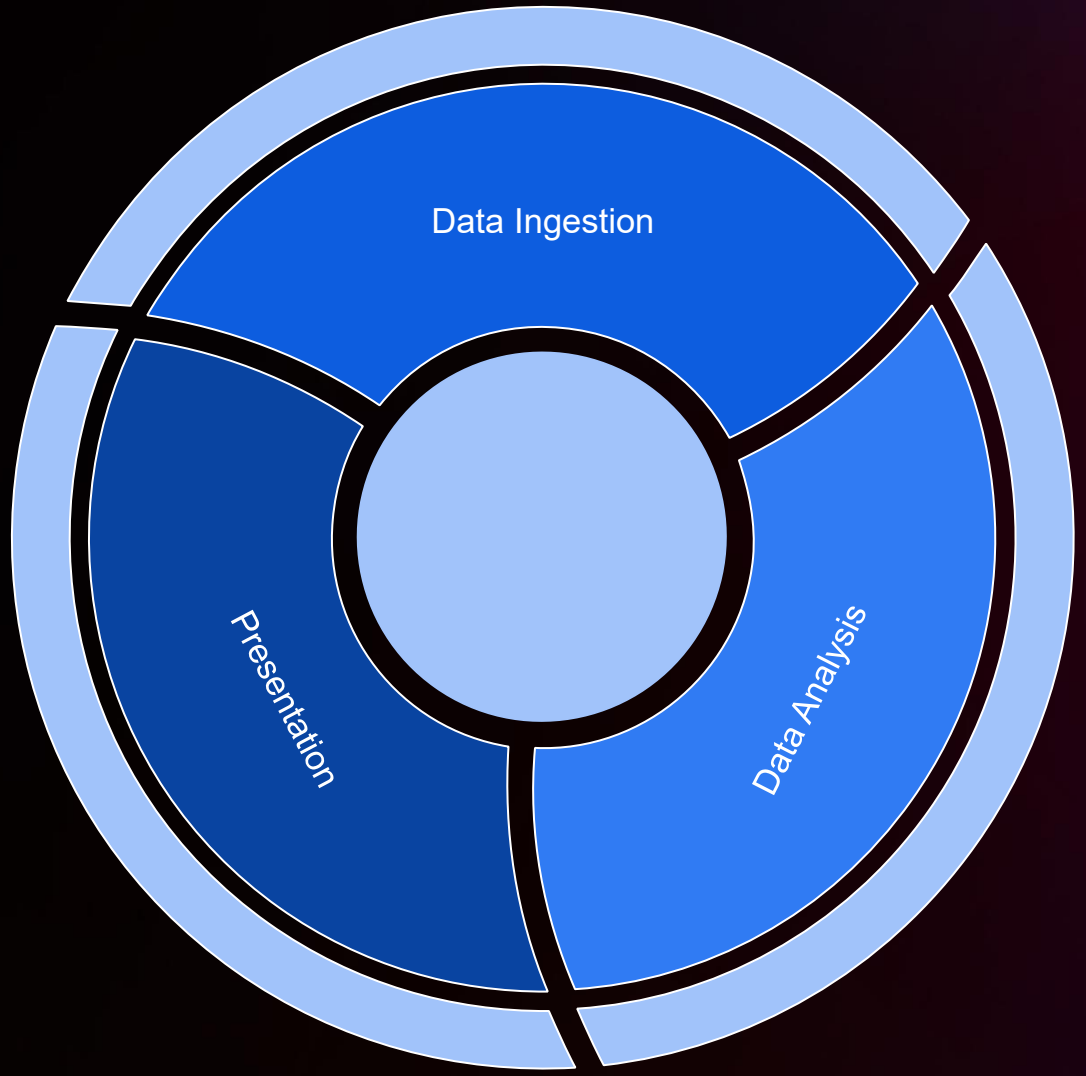
have purchased items they did not intend to initially, due to personalized recommendations

Source: Product Over Price: The Critical Role Personalization Plays in Converting Online Searches into Sales, commissioned by Elastic and conducted by Wakefield Research

aws

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

Personalization is a *_data_* problem



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

Elastic capabilities enable all stages of a personalization implementation

Data Ingestion

- Ongoing usage metrics with Search Analytics
- Prebuilt integrations with external data sources
- Custom integrations with use case specific data

Data Analysis

- Visual analytics with Kibana dashboards
- Broad and flexible set of analytics capabilities
- Speed, scale, relevance

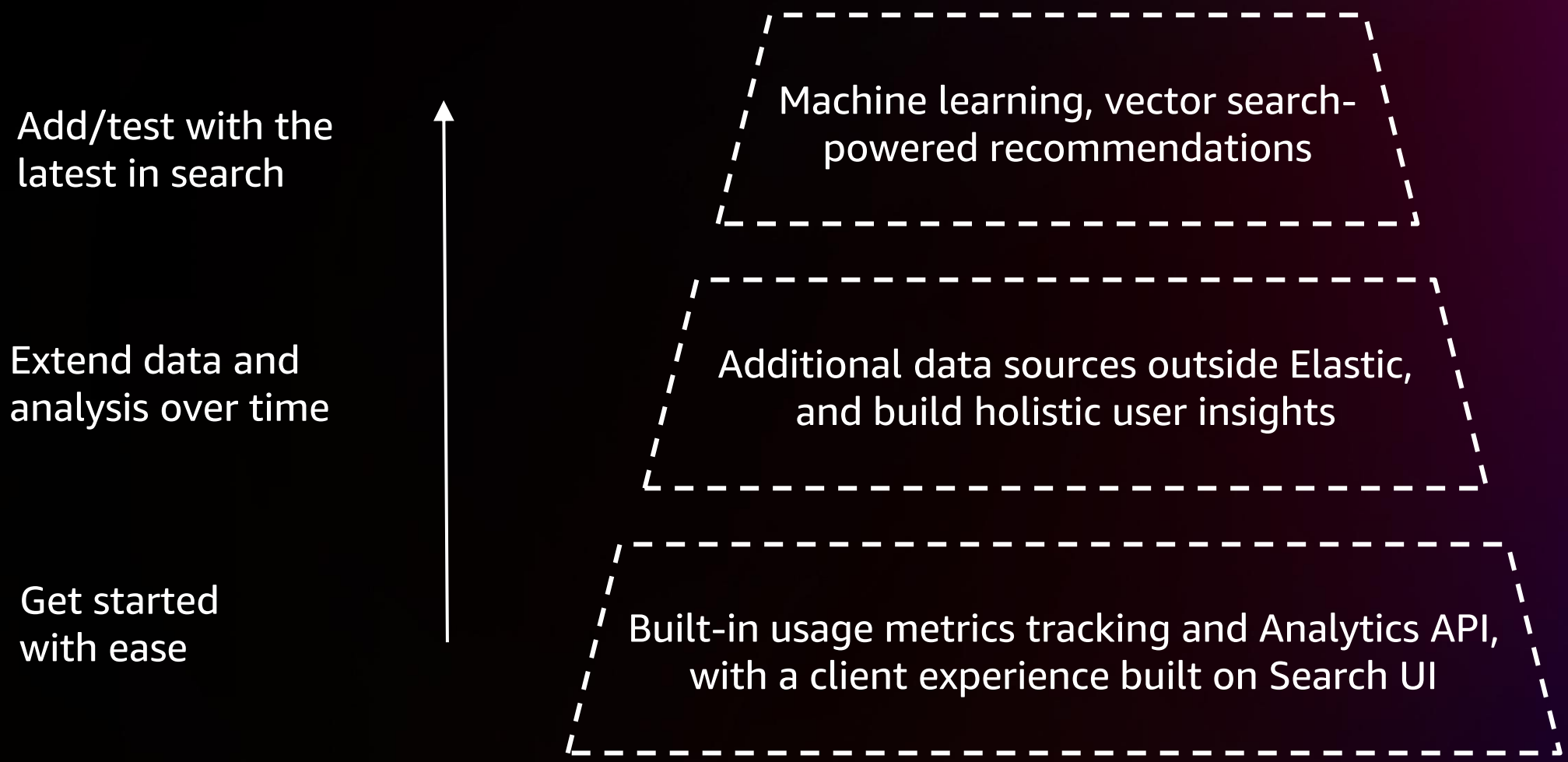
Presentation

- Built in tooling to manage search responses
- APIs to refine responses at runtime
- Accelerate development with Search UI

aws

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

Personalization journey with Elastic

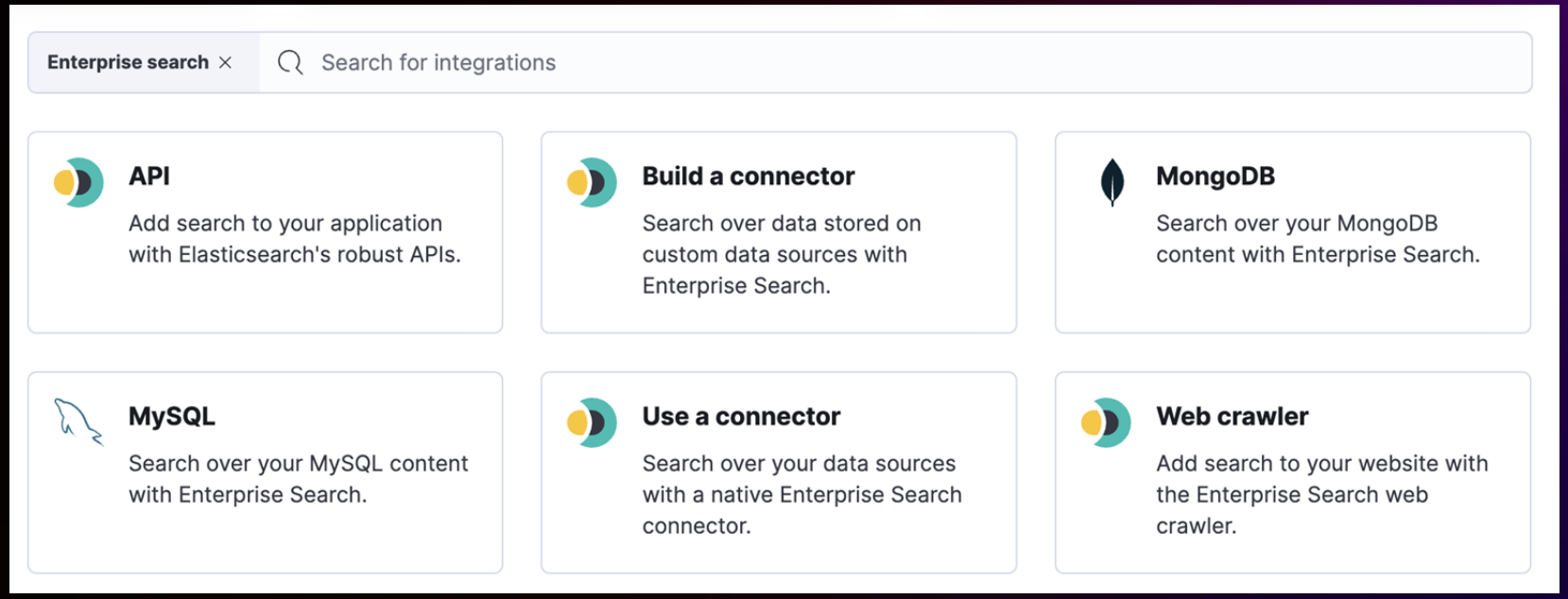


aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Crawl or connect searchable data

Accelerate building search experiences with ingestion flexibility and scalability

- **Native web crawler** with PDF extraction and simple authenticated crawls
- Native connector clients for **MongoDB, Gitlab,** and **MySQL**
- Open code connector clients in **Ruby** and **Python**
- New connectors – **py framework**

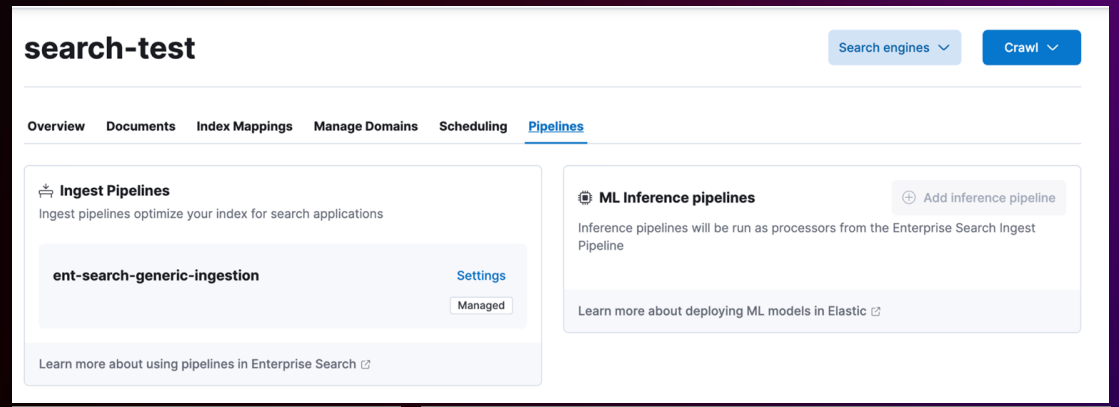


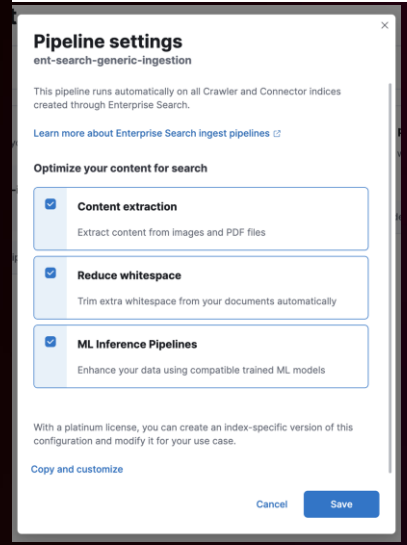
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

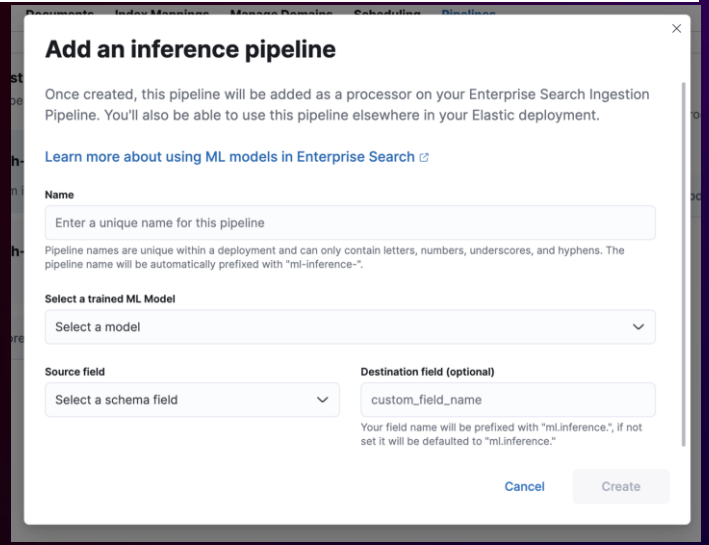
Pipeline management for ML models

Choose what to ingest, how to transform ingested data, and ML models needed

- Managed, custom, and inference pipelines
- Tune relevance at ingest time







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

Topics we don’t have time to cover

- Named Entity Recognition
- Zero-shot classification
- Applying NLP and vector search to Observability or Security

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

Performance enhancements

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

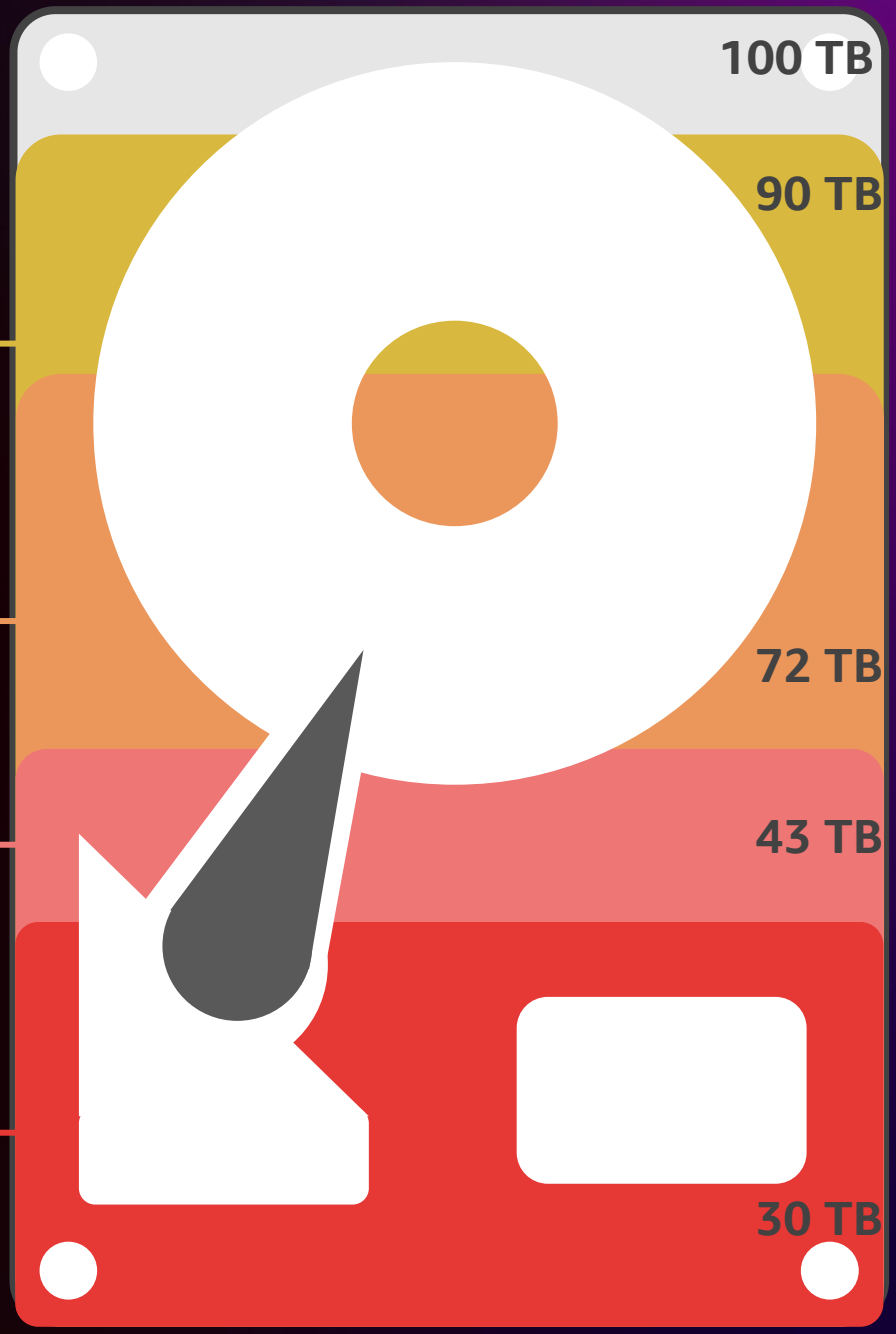
Storage savings

7.14 match only text
Save up to 10%

8.1 doc-value-only fields
Save up to 20%

8.4 synthetic _source
Save up to 40%

8.5 routing and sorting by TSID
Save up to 30%



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

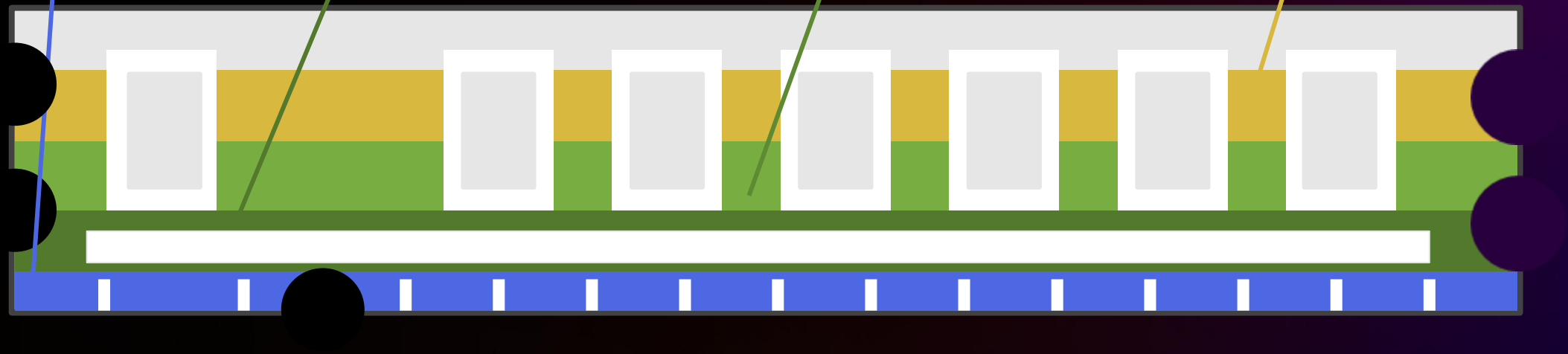
Memory savings

7.16 92% Memory Savings on idle shards

7.14 Avoid global ordinals in composite aggregation

8.3 New sharding 3,000 indices per GB of heap

8.3 New sharding guidance 1 kb per field



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

Performance

7.14 geotile grid aggregation over geo_shape 15% faster

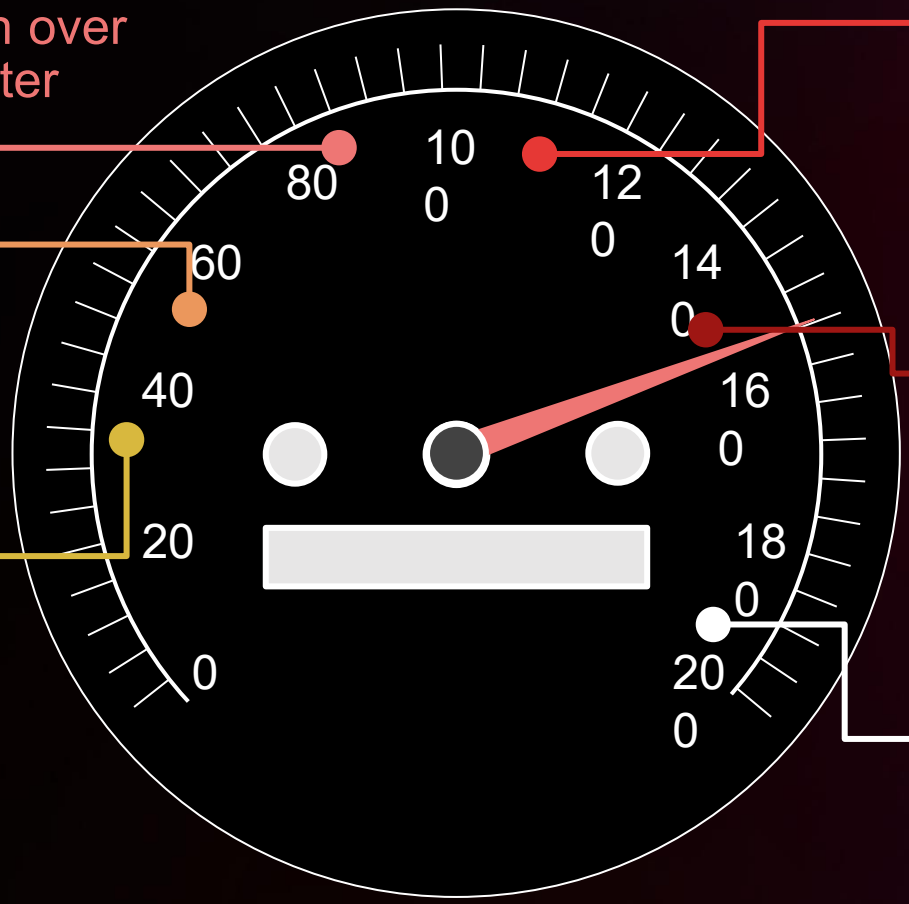
7.16 search_after on sorted long fields 20% faster

7.13 terms agg 20%faster

8.2 range queries 20% faster

7.11 date histograms 8x faster

8.2 random sampler aggregation 80x faster on count aggregation



aws

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

Search 1PB in minutes stored on S3



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

Intelligently store and search everything

	Accessed Frequently	Accessed Less Frequently	Accessed Less Frequently	Accessed Intermittently	Accessed Rarely
Data Usage					
Performance	milliseconds - seconds	seconds - 10s of seconds	seconds - 10s of seconds	10 seconds - minutes	minutes - 10s of minutes

Serverless Elasticsearch

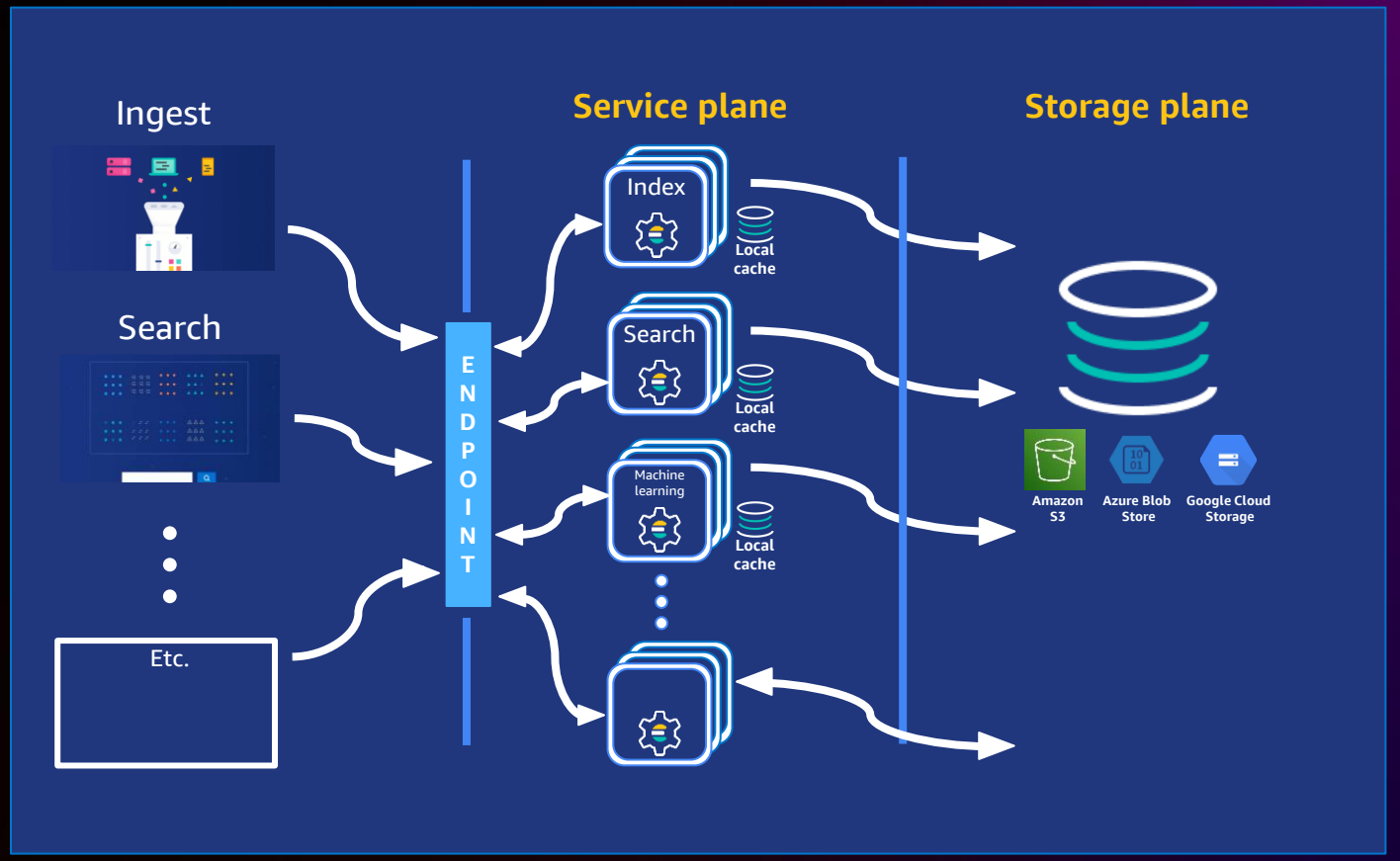
Next generation managed service

Scale compute & storage independently

Consume only what you need

Versionless

Autoscaling – index and search



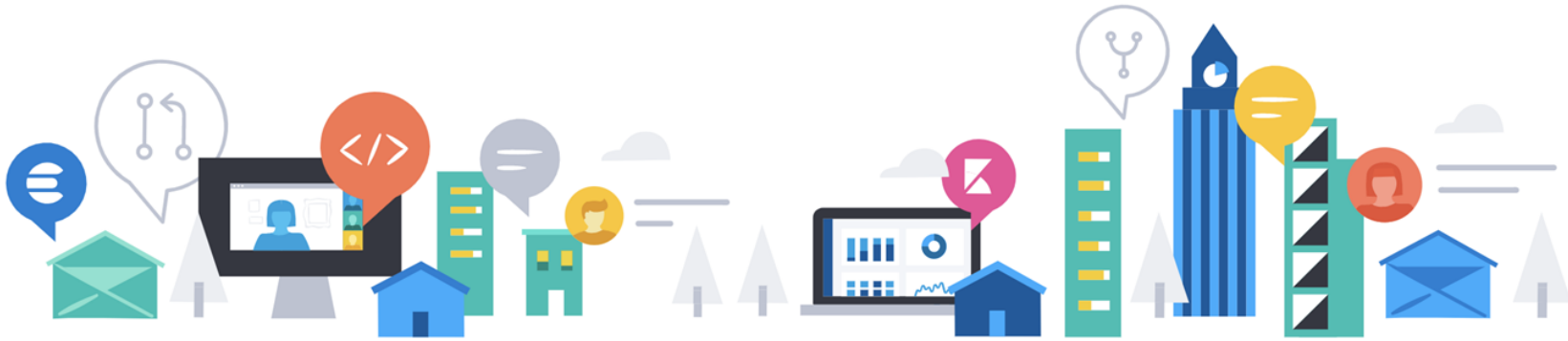
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Community resources

Elastic Contributor Program

Elastic User Groups

Elastic for Students and Educators



discuss.elastic.co

Community YouTube Channel

Elastic Stack Community Slack Workspace

aws

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

Thank you!



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

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