OpenSearch v3: A New Era of Search Innovation

Dotan Horovits

OpenSearch Ambassador

CNCF Ambassador

Senior Open Source Advocate, AWS

Aswath Srinivasan

Senior Search Engine Architect

OpenSearch @ AWS

OpenSearch

OpenSearch is the trusted open source platform for AI-powered search, observability, and analytics with built-in security, high performance, and a flexible architecture for modern applications.

OpenSearch
SOFTWARE FOUNDATION

THE
LINUX
FOUNDATION

OpenSearch Software Foundation

aws

IBM

SAP

Uber

ByteDance

Canonical

OpenSource Connections

DigitalOcean

DTEX

Aryn

NetApp Instaclustr

portal26

seacom

aiven

OpenSearch

OpenSearch by the numbers

1.6B+
project downloads

1.25M
monthly page views For opensearch.org

100+
solution providers

3K+
active contributors

400+
active organizations

29
new releases since project launch

140+
GitHub repositories

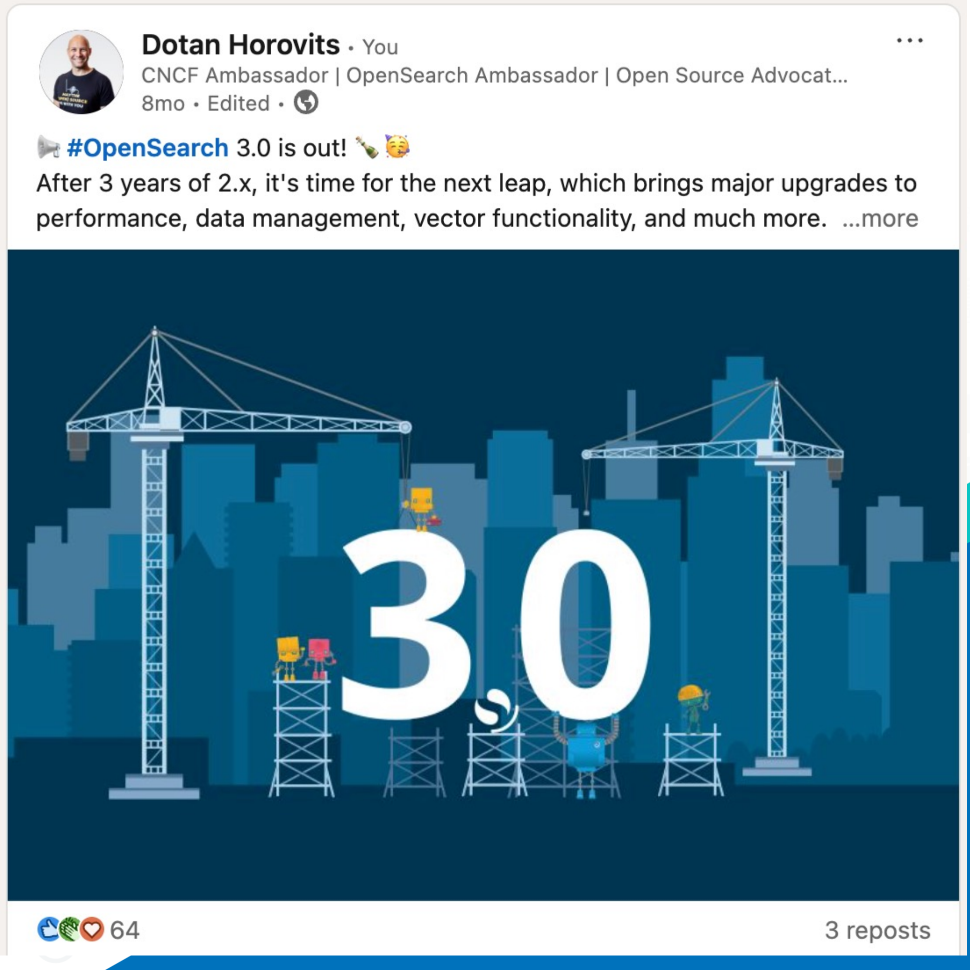
4K+
Slack workspace members

7K+
user forum members

OpenSearch

Next major release: OpenSearch v3!

- 📈 Upgrade to **Apache Lucene v10** and JDK 24
- 📈 Pull-based ingestion
- 📈 Reader-Writer separation
- 📈 Native **MCP** support
- 📈 Expanded **PPL** queries, backed by **Apache Calcite**
- 📈 and much more…



Dotan Horovits • You

CNCF Ambassador | OpenSearch Ambassador | Open Source Advocat…

8mo • Edited •

📢 #OpenSearch 3.0 is out! 🍾 🥳

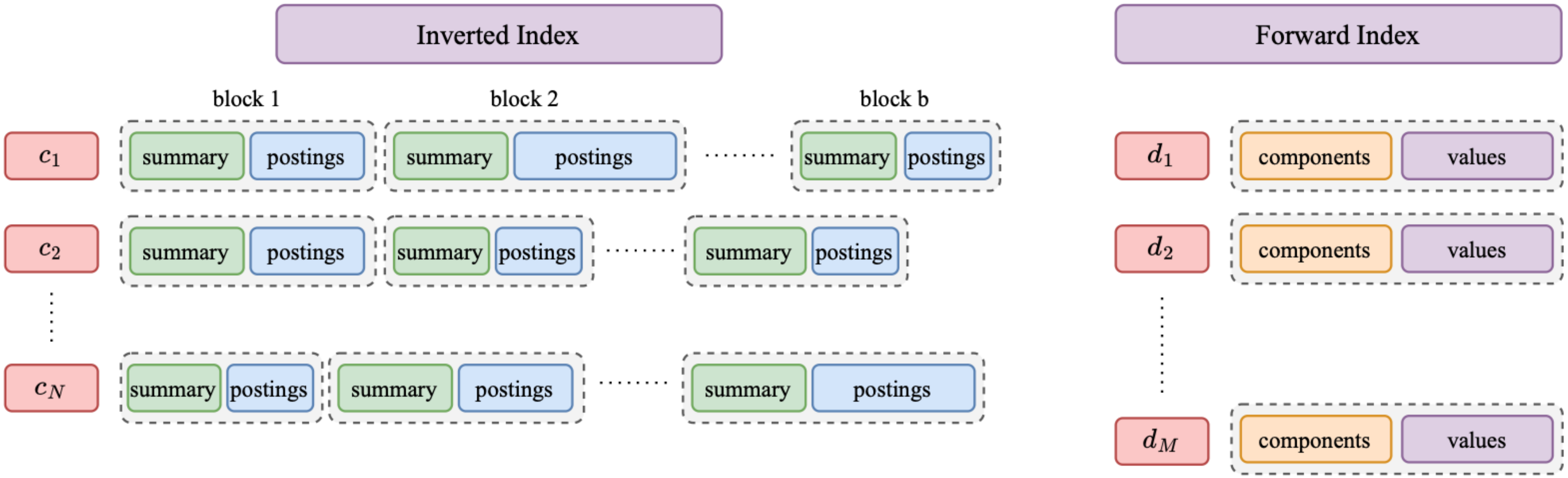
After 3 years of 2.x, it's time for the next leap, which brings major upgrades to performance, data management, vector functionality, and much more. …more

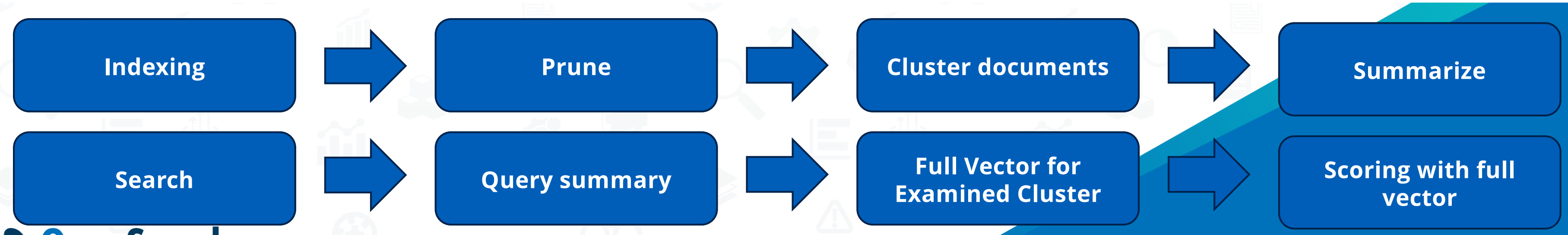
64

3 reposts

OpenSearch

Neural Sparse ANN – SEISMIC





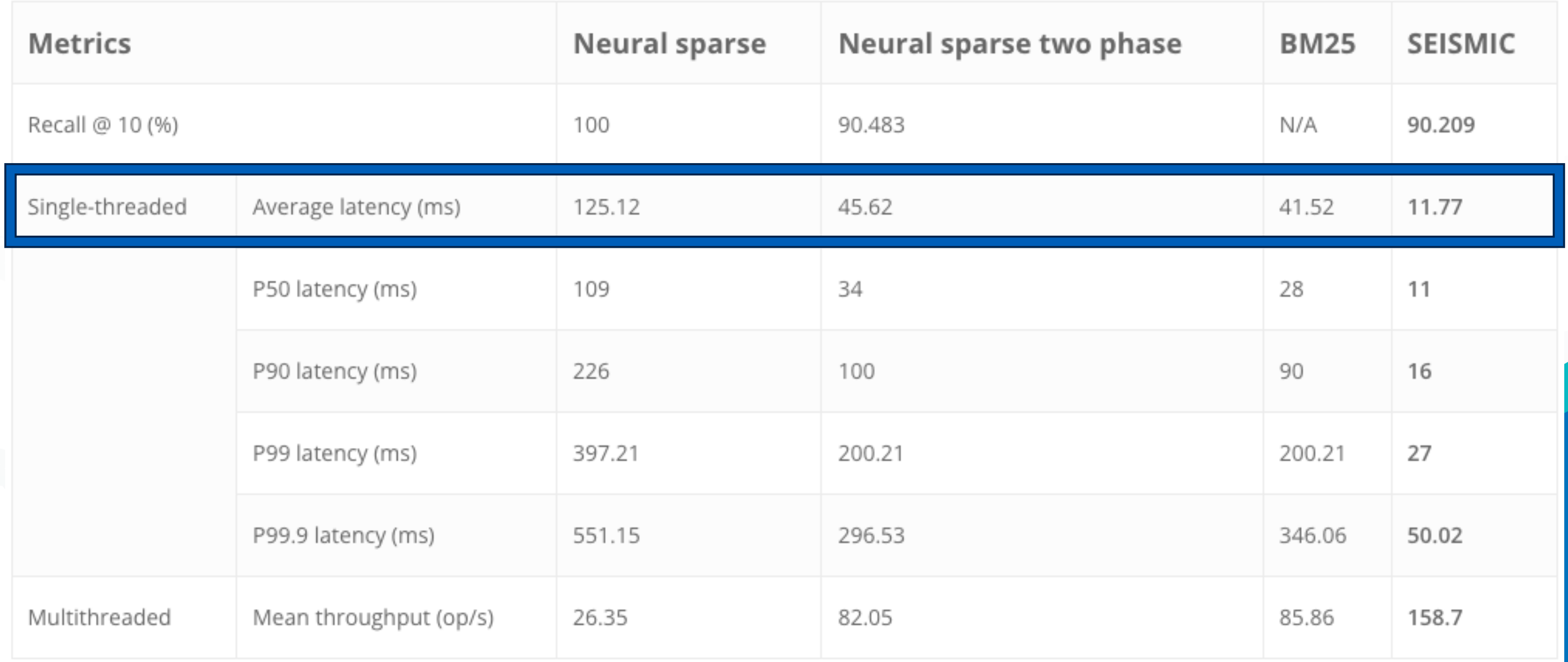
Indexing → Prune → Cluster documents → Summarize

Search → Query summary → Full Vector for Examined Cluster → Scoring with full vector

OpenSearch

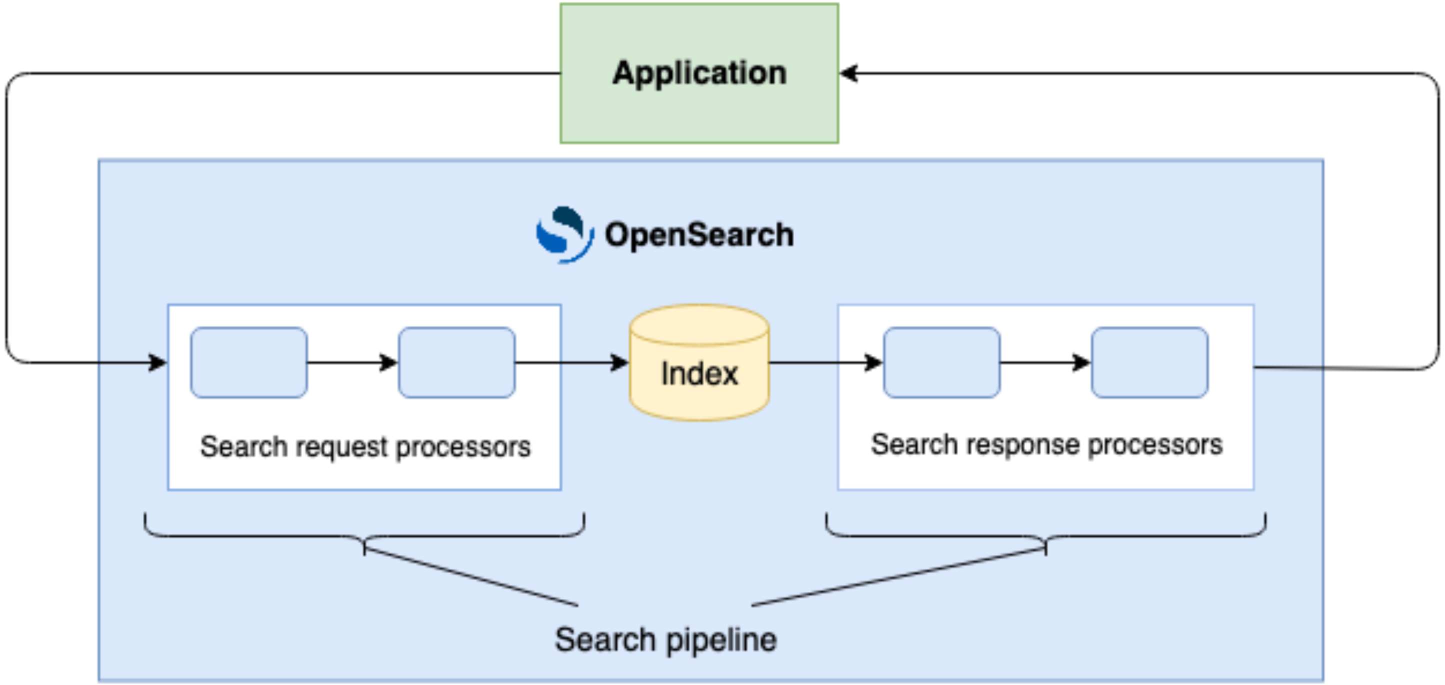
Meet OpenSearch’s SOTA model

Search latency on billion docs (v3-gte, OpenSearch v3.3)



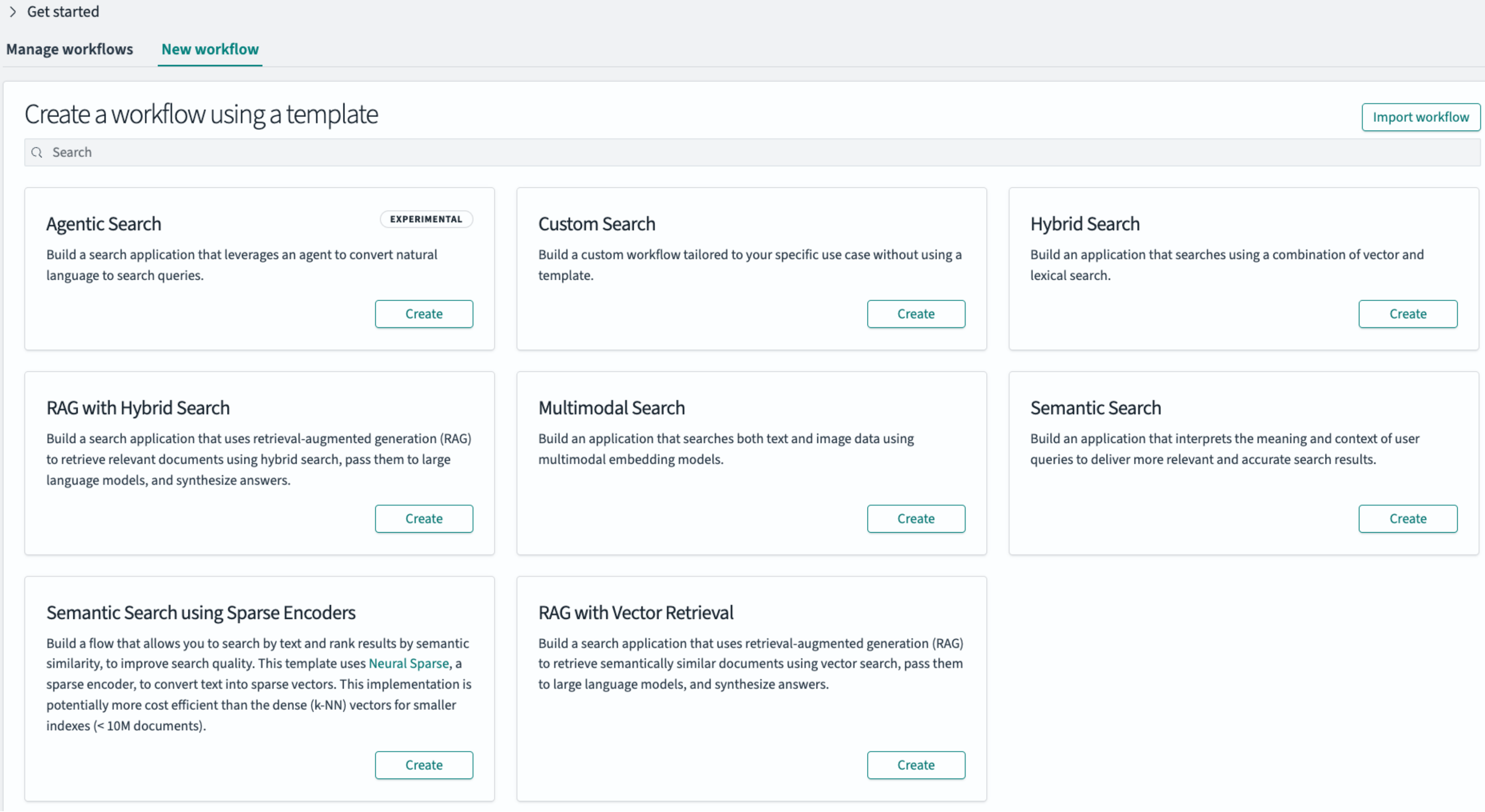
Metrics		Neural sparse	Neural sparse two phase	BM25	SEISMIC
Recall @ 10 (%)		100	90.483	N/A	**90.209**
Single-threaded	Average latency (ms)	125.12	45.62	41.52	**11.77**
	P50 latency (ms)	109	34	28	**11**
	P90 latency (ms)	226	100	90	**16**
	P99 latency (ms)	397.21	200.21	200.21	**27**
	P99.9 latency (ms)	551.15	296.53	346.06	**50.02**
Multithreaded	Mean throughput (op/s)	26.35	82.05	85.86	**158.7**

Sys generated Search Pipeline



OpenSearch

AI Search flow



Create a workflow using a template

Agentic Search

Build a search application that leverages an agent to convert natural language to search queries.

Custom Search

Build a custom workflow tailored to your specific use case without using a template.

Hybrid Search

Build an application that searches using a combination of vector and lexical search.

RAG with Hybrid Search

Build a search application that uses retrieval-augmented generation (RAG) to retrieve relevant documents using hybrid search, pass them to large language models, and synthesize answers.

Multimodal Search

Build an application that searches both text and image data using multimodal embedding models.

Semantic Search

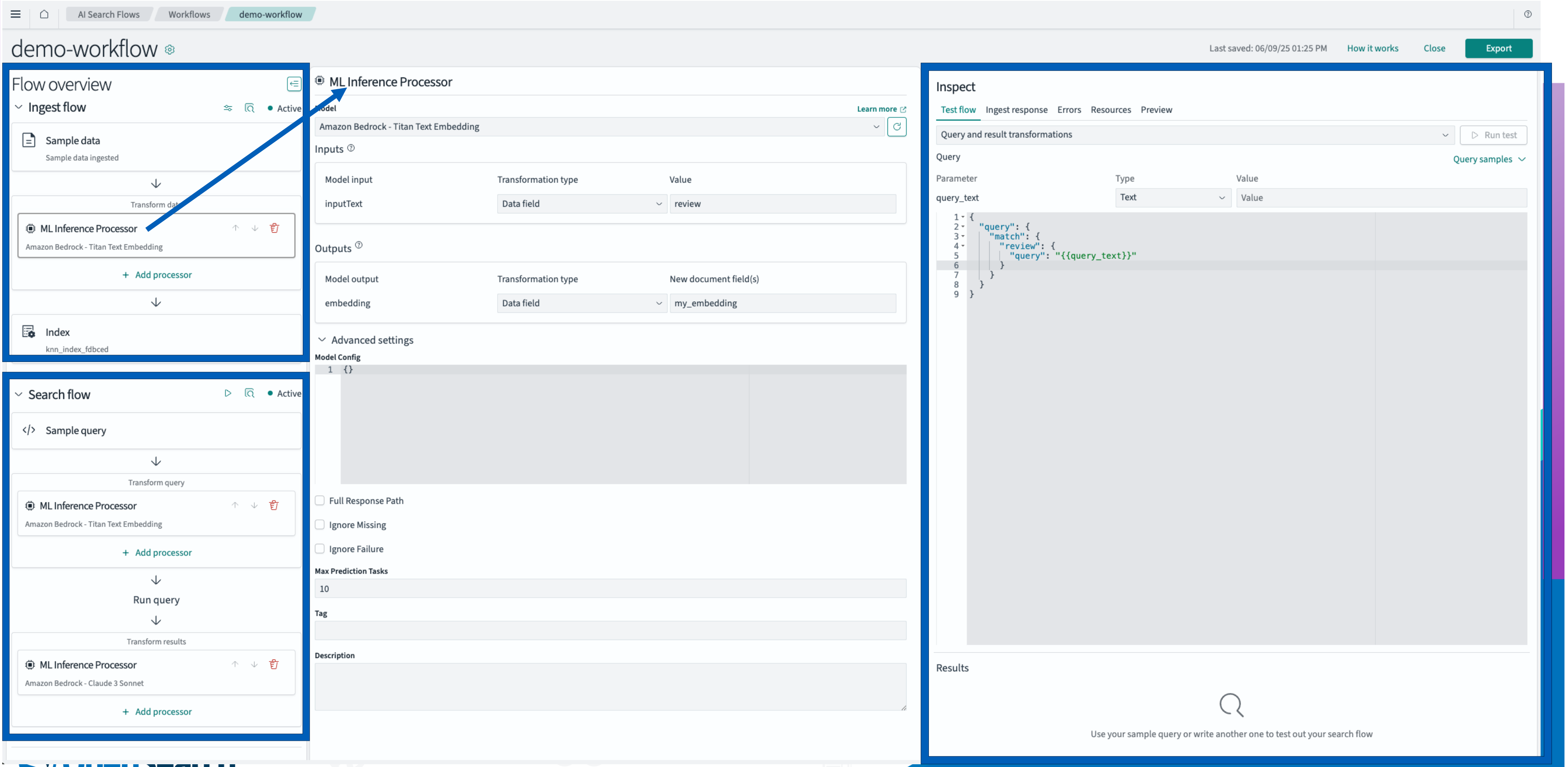
Build an application that interprets the meaning and context of user queries to deliver more relevant and accurate search results.

Semantic Search using Sparse Encoders

Build a flow that allows you to search by text and rank results by semantic similarity, to improve search quality. This template uses Neural Sparse, a sparse encoder, to convert text into sparse vectors. This implementation is potentially more cost efficient than the dense (k-NN) vectors for smaller indexes (< 10M documents).

RAG with Vector Retrieval

Build a search application that uses retrieval-augmented generation (RAG) to retrieve semantically similar documents using vector search, pass them to large language models, and synthesize answers.



demo-workflow

Flow overview

Ingest flow ● Active

Sample data
Sample data ingested

Transform data

ML Inference Processor
Amazon Bedrock - Titan Text Embedding

+ Add processor

Index
knn_index_fdbced

Search flow ● Active

Sample query

Transform query

ML Inference Processor
Amazon Bedrock - Titan Text Embedding

+ Add processor

Run query

Transform results

ML Inference Processor
Amazon Bedrock - Claude 3 Sonnet

+ Add processor

ML Inference Processor

Model: Amazon Bedrock - Titan Text Embedding

Inputs

Model input	Transformation type	Value
inputText	Data field	review

Outputs

Model output	Transformation type	New document field(s)
embedding	Data field	my_embedding

Advanced settings

Model Config

```
{}
```

□ Full Response Path

□ Ignore Missing

□ Ignore Failure

Max Prediction Tasks: 10

Tag:

Description:

Inspect

Query and result transformations

Query

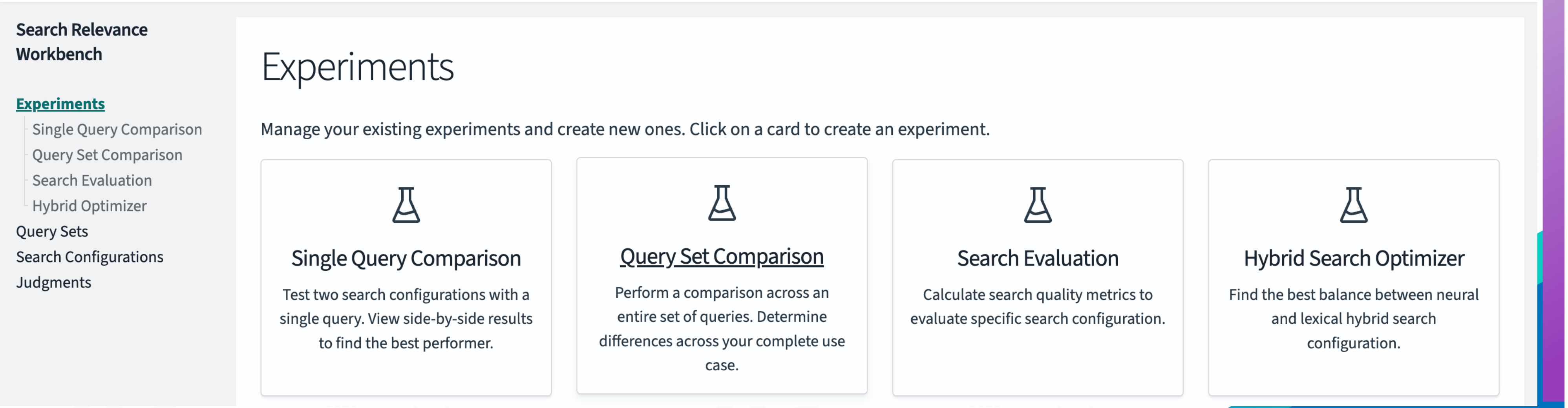
Parameter	Type	Value
query_text	Text	Value

```
{
  "query": {
    "match": {
      "review": {
        "query": "{{query_text}}"
      }
    }
  }
}
```

Results

Use your sample query or write another one to test out your search flow

Search Relevancy Workbench



Search Relevance Workbench

Experiments
- Single Query Comparison
- Query Set Comparison
- Search Evaluation
- Hybrid Optimizer

Query Sets

Search Configurations

Judgments

Experiments

Manage your existing experiments and create new ones. Click on a card to create an experiment.

Single Query Comparison

Test two search configurations with a single query. View side-by-side results to find the best performer.

Query Set Comparison

Perform a comparison across an entire set of queries. Determine differences across your complete use case.

Search Evaluation

Calculate search quality metrics to evaluate specific search configuration.

Hybrid Search Optimizer

Find the best balance between neural and lexical hybrid search configuration.

OpenSearch

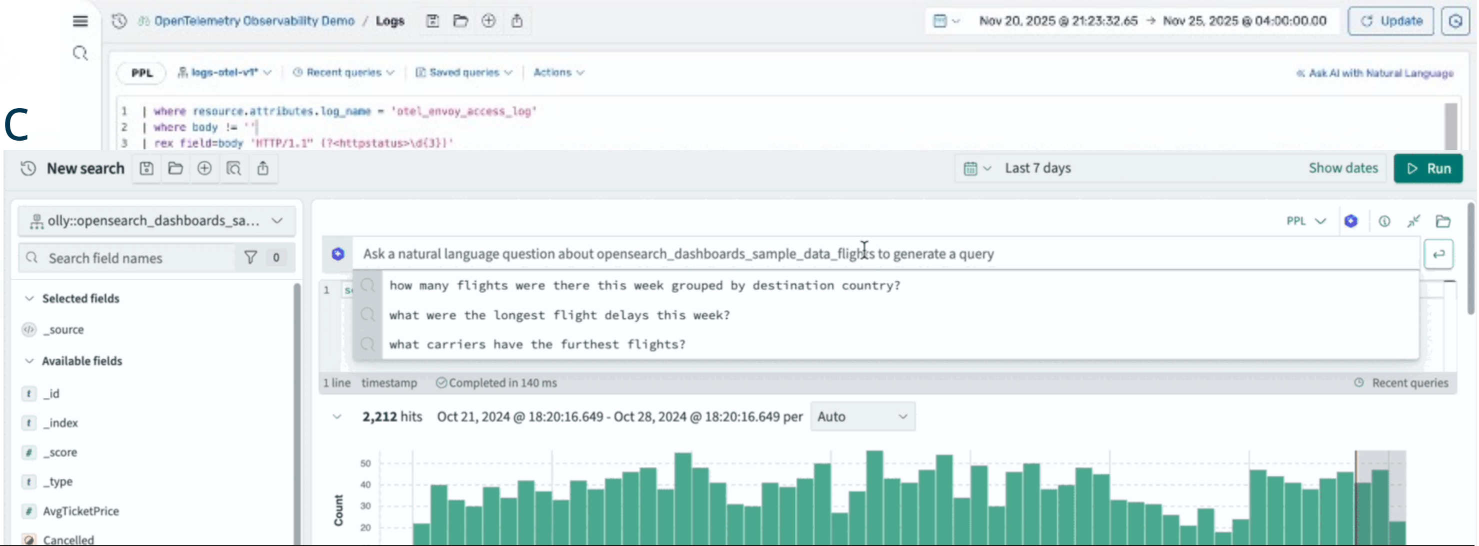
OpenSearch

Agentic Search

Demo

- PPL – Build Viz via Natural Language
- BQ – ADC & RR for improving recall
- Batch Inference Semantic Highlighting
- Late interaction
- Persistent Agentic Memory
- Radial Search,

RRF, zscore, MMR



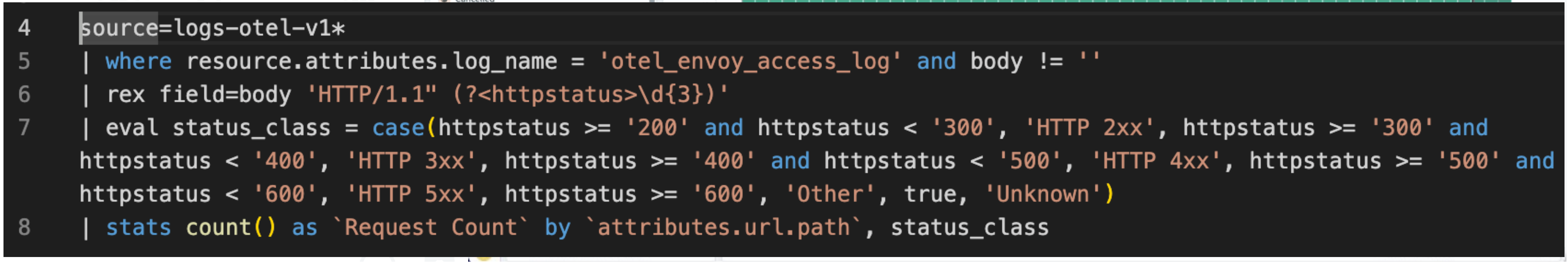
```
| where resource.attributes.log_name = 'otel_envoy_access_log'
| where body != ''
| rex field=body 'HTTP/1.1" (?<httpstatus>\d{3})'
```

Ask a natural language question about opensearch_dashboards_sample_data_flights to generate a query

how many flights were there this week grouped by destination country?

what were the longest flight delays this week?

what carriers have the furthest flights?

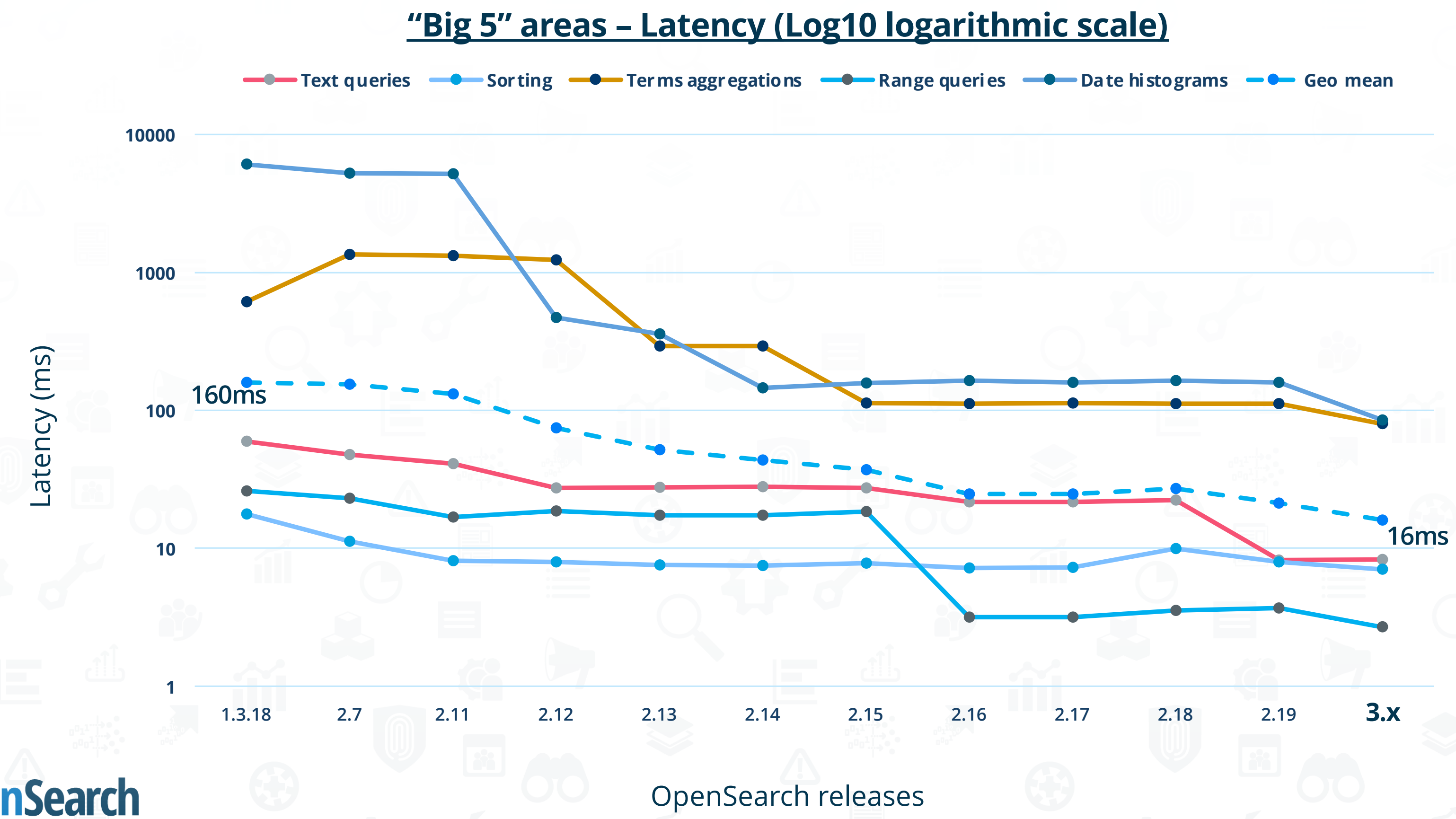


```
source=logs-otel-v1*
| where resource.attributes.log_name = 'otel_envoy_access_log' and body != ''
| rex field=body 'HTTP/1.1" (?<httpstatus>\d{3})'
| eval status_class = case(httpstatus >= '200' and httpstatus < '300', 'HTTP 2xx', httpstatus >= '300' and
httpstatus < '400', 'HTTP 3xx', httpstatus >= '400' and httpstatus < '500', 'HTTP 4xx', httpstatus >= '500' and
httpstatus < '600', 'HTTP 5xx', httpstatus >= '600', 'Other', true, 'Unknown')
| stats count() as `Request Count` by `attributes.url.path`, status_class
```

Performance improvments

OpenSearch v3

Performance improvements



OpenSearch

Performance improvements

- **GPU acceleration: 9.3x** indexing speed, reducing costs by **3.75x**
- **Lucene on FAISS:** nearly **doubles QPS** at 32x quantization
- **gRPC/protobuf transport: ~50% reduction** in client-side processing time, and **~20% higher throughput** for vector search workloads
- **Apache Arrow support**: eliminate serialization overhead
- **Pull based ingestion** for Streaming service like Apache Kafka
- **Reader-Writer separation:** decouple indexing and search workloads for predictable performance
- **Derived Source:** 2x storage savings
- **Star-tree index** for complex aggregation over large set of data

OpenSearch 3.5 — coming in February

- Skip list support for aggregations
  - github.com/opensearch-project/OpenSearch/issues/19384
- New Application Performance Monitoring experience
  - github.com/opensearch-project/dashboards-observability/issues/2545
- Built-in agent observability
  - Experimental in 3.5, looking for feedback
  - github.com/opensearch-project/dashboards-traces/blob/main/GETTING_STARTED.md

OpenSearch

Find us at

Github

Events

Blog

Slack

Forum

User groups

OpenSearch

Thank you!

Dotan Horovits

Aswath Srinivasan