S&P Global
Energy

THE TALK

SPGE Enhanced Search: a relevancy engine for Agentic RAG

Hybrid retrieval and business-driven ranking on Amazon OpenSearch Service

From a fragmented content estate to one retrieval layer for humans and agents

IN THIS TALK

- A fragmented estate
- How ranking is built
- Retrieval as a service
- Lessons and roadmap

Tomasz Florin & Artur Barczyński · S&P Global Energy · Haystack Conference

THE ARGUMENT **Retrieval quality — not model choice — decides what a human can find or an agent can answer**

ACT 01 · BUSINESS CHALLENGE

A fragmented estate

Multiple content types, indexed twice, retrieved two different ways

- Fragmented retrieval
- Search requirements
- Choosing an engine

ACT 02 · THE RELEVANCY ENGINE

How ranking is built

Hybrid retrieval first, then business rules applied on top of the score

- Four architectural layers
- Hybrid retrieval
- Business rules as signals
- Freshness decay profiles

ACT 03 · SERVING AI AGENTS

Retrieval as a service

The same index behind copilots, MCP servers and autonomous agents

- Supporting Agentic RAG
- Multi-region architecture
- Index sizing at scale

ACT 04 · WHAT WE LEARNED

Lessons and roadmap

Five lessons from several iterations in production

- Operational lessons
- Future roadmap
- Key takeaways

Two stacks, one index · and a ranking layer the business can actually change

S&P Global Energy

WHERE ENTERPRISE AI PROGRAMS INVEST

- Foundation models
- Prompts
- Orchestration
- Agent frameworks

WHAT ACTUALLY MOVED THE NUMBERS

Retrieval quality, not model choice, decided the outcome

Same foundation models · different retrieval

01 · LEXICAL

Exact still wins

Report identifiers, symbols and company names have to match literally

- BM25 on identifiers
- Vector alone missed them
- Hybrid, not either/or

02 · BUSINESS

Relevance is a decision

The engine cannot know which content the business ranks higher

- Premium over common
- A weight per content type
- Configured, not learned

03 · GOVERNANCE

Trusted, not just similar

Entities and entitlements resolve before anything is ranked

- Ontology-normalised entities
- Entitlements inside the query
- Citations back to source

The gains came from the retrieval layer, not from a bigger model

S&P Global Energy

SECTION 01

Business challenge

the problem, before any of the solution

01

Multiple content types, indexed twice — and retrieved through two stacks that share nothing.

IN THIS SECTION

- Fragmented retrieval
- Search requirements
- Choosing an engine

Two retrieval stacks · partial coverage on one, keyword-only on the other

THE ESTATE

Content estate

Multiple content types — indexed twice, separately.

- Research & analytics reports
- News & market intelligence
- Datasets & dashboards
- PDFs, images & API content

STACK 1 · NEW · ONE CALL, ONE LIST

Own pipeline → OpenSearch → Enhanced Search API

ONE CALL — One ranked list, one response

every content type, merged and ranked once

Keyword + vector, ontology-aware

Live for part of the estate only

CONSUMERS

Served by Enhanced

- Chat AI
- Core platform search

Nothing shared between the two stacks — ingestion · relevance · entitlements · ranking

STACK 2 · LEGACY · ONE CALL PER SECTION

Own pipeline → MarkLogic → Classic Search APIs

CALL 1 — Section 1 — its own ranking

CALL 2 — Section 2 — its own ranking

CALL n — Section n — its own ranking

Own ranking logic per section

Same MarkLogic schema underneath

CONSUMERS

Served by Classic

- Alerts
- Real-time engine
- Content push — 5 systems
- Downloads & preview

Classic: one result list, but a separate API call and its own ranking per section · Enhanced: one API, one list, one source

ONE RETRIEVAL LAYER SERVES

Web apps | APIs | Copilots | AI agents

01 · RELEVANCE

Ranking you can trust

The right documents come back — and come back in the right order

- Improved keyword relevancy
- Results ordered by relevancy
- Business-defined ranking rules
- Search inside document content

02 · EXPERIENCE

One result set

Every content section searched, ranked and filtered as one

- Single ranked results list
- Filtering across all sections
- No section-by-section search

03 · DISCOVERABILITY

The main access point

The single front door to everything the platform holds

- Articles & documents
- Dashboards
- Structured data
- Platform functionality

04 · AI & AGENTS

AI-ready retrieval

Transformer models, embeddings and LLMs on the same layer

- Semantic search support
- Relevant source for RAG
- Ingestion as AI-ready data

One platform, one index, one API · not a retrieval stack per application

Keyword search 01

- Exact identifiers & symbols
- Company names & titles
- Lexical BM25 matching

Entitlement filtering 02

- Complex entitlement queries
- Applied inside retrieval
- Not a post-filter

Vector search 03

- Native embedding support
- Hybrid ranking built in
- New vector features fast

Multi-region 04

- AWS multi-region deployment
- Replication & failover
- HA and disaster recovery

Cost & operability 05

- Licensing cost
- Implementation complexity
- Infrastructure & ops burden

Four engines, one benchmark

Lumin	MarkLogic + Semantics	Elasticsearch	OpenSearch
not selected	not selected	not selected	SELECTED

Why OpenSearch cleared all five

- No licence cost
- Elasticsearch-derived API
- AWS is a core contributor
- Managed — no search-infra team

S&P Global Energy

SECTION 02

The relevancy engine

hybrid retrieval, then rules on top of the score

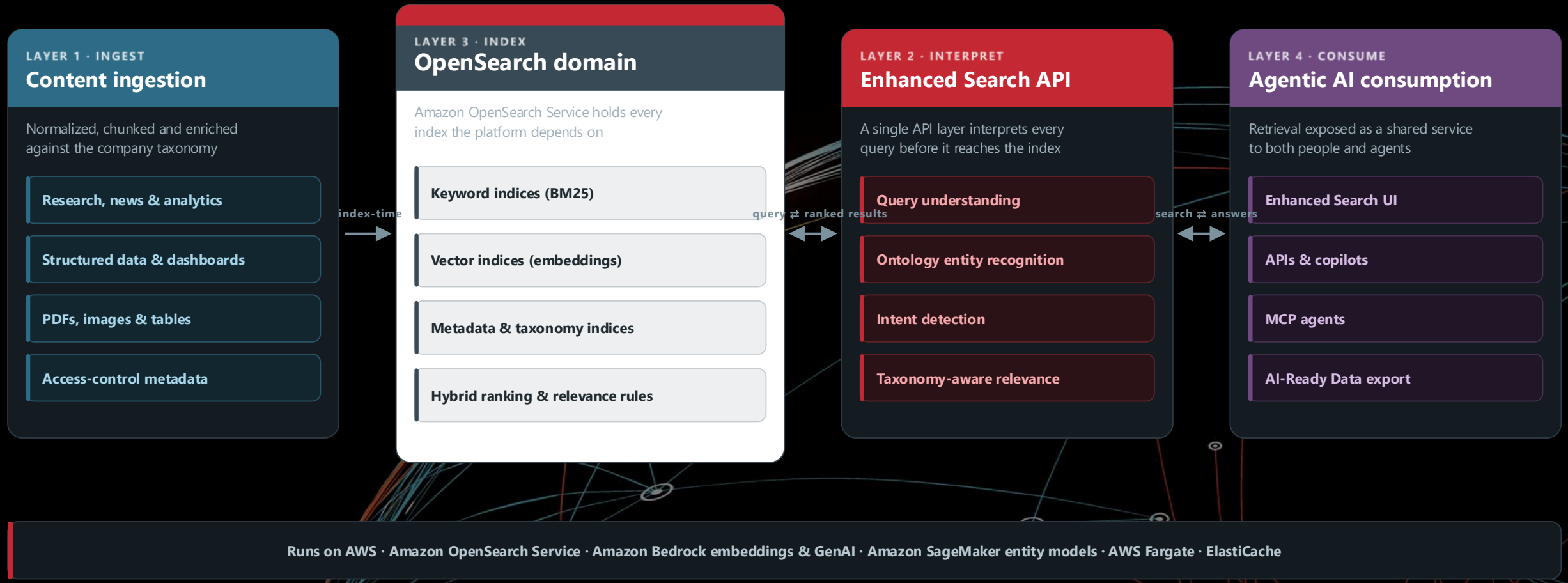
02

Lexical and semantic retrieval run together first — business rules only touch the score after.

IN THIS SECTION

- Four architectural layers
- Hybrid retrieval
- Business rules as signals
- Freshness decay profiles

Retrieval score first, business rules second — never the other way round



S&P Global Energy

1 · UNDERSTAND

Query understanding

Ontology-driven entity normalization maps naming, so precision survives varied phrasing.

- Synonyms & aliases
- Company hierarchies
- Entity recognition
- Intent detection

every query runs both

2a · LEXICAL

Lexical retrieval

Amazon OpenSearch Service BM25 — deterministic, exact matching.

- Report identifiers
- Symbols & companies
- Product names & titles

2b · SEMANTIC

Semantic retrieval

Amazon Bedrock embeddings — queries and content as vectors.

- Concepts & intent
- Natural-language Q&A
- Similar documents

merge

3 · RANK

Hybrid ranking

Both result sets merge through one custom relevance calculation.

- Business importance
- Freshness & decay curve
- Content type (taxonomy)
- Entitlements
- Vector similarity
- Phrase length (vector weight)

One ranked, entitlement-aware result list · served to the Enhanced Search UI, APIs, copilots and MCP agents

SCORE COMPOSITION · ONE PAINLESS SCRIPT PER CLAUSE

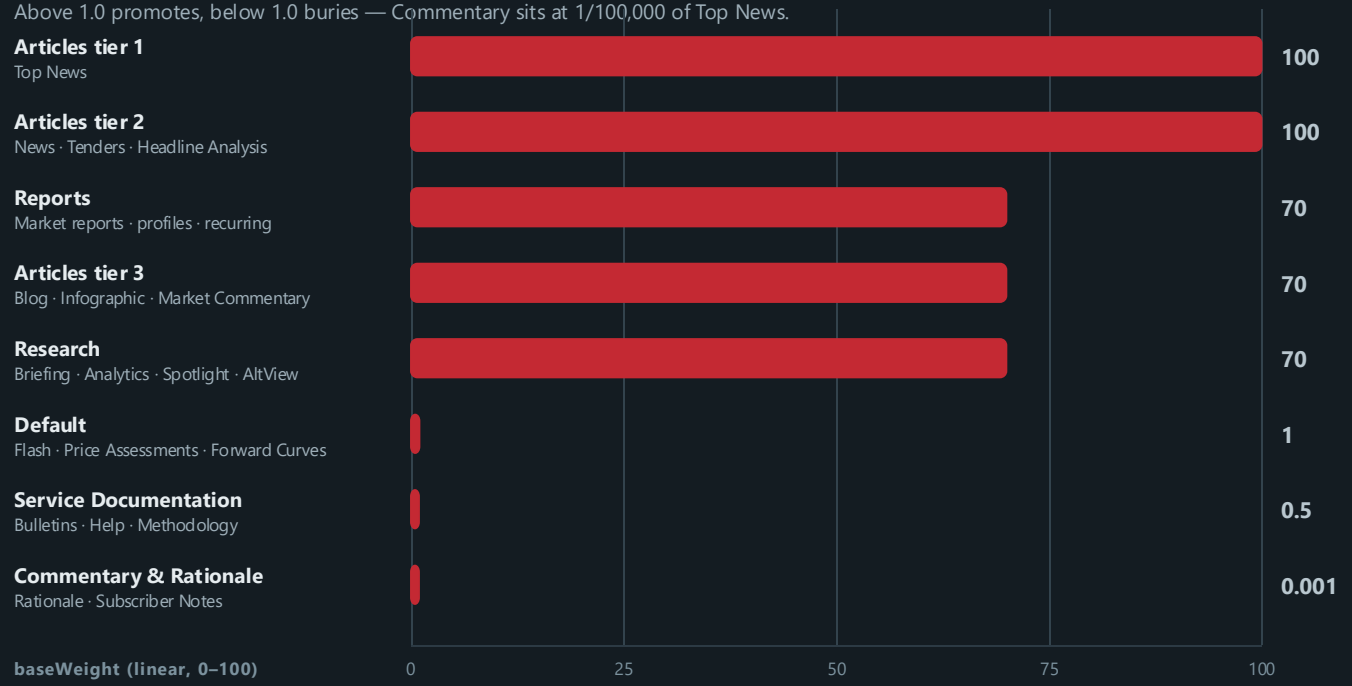
red terms are business rules — configured, not learned

Lexical BM25 ⊕ Vector × 600k or 60M by phrase length × scoreMultiplier × baseWeight × freshness(category) = final rank

BUSINESS PRIORITY

baseWeight: one multiplier, five orders of magnitude

Each content category carries a business weight applied to every match.
Above 1.0 promotes, below 1.0 buries — Commentary sits at 1/100,000 of Top News.



NAME MATCH CLAUSE

A name match is scored separately

The query fans out into three scored clauses. One of them carries titleWeight on the entity-name field, weighted per category.

- Reports ×1000
- All other categories ×75
- Default (Flash, price data) ×50
- Title & body matches: unweighted

DECAY EXEMPTION

Evergreen vs recency-sensitive

Reports are the only category flagged noDecay — on a name match their age is ignored entirely. Everything else decays.

- Reports only: noDecay = true
- 9 subsections on the evergreen path
- Every other category keeps decay
- Body matches always decay

S&P Global Energy

THE FRESHNESS TERM · ONE FUNCTION FOR ALL 8 CATEGORIES

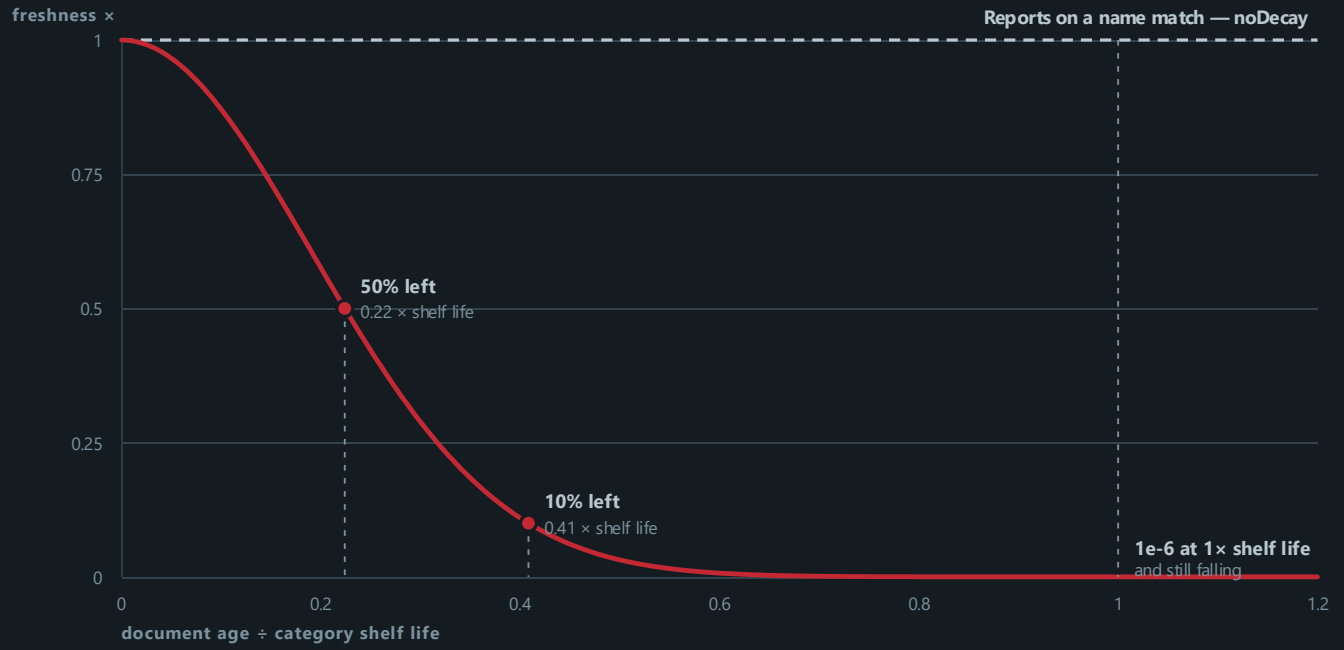
Gaussian · offset 0d · baseline 0 · decay 1e-6 — none of these vary

freshness(age) = **0.000001** ^ **(age ÷ scale)²** → **max 1.0, never higher**

ONE SHAPE

The same curve, rescaled per content type

Age is measured from today — or from the upper bound of a date filter, so a historic search re-centres the curve instead of flattening every result.



SHELF LIFE

What the scale means in days

The one parameter that is tuned per content type.

	50% left	10% left	1e-6 left
15 days	3.4	6.1	15
Top News · News · Tenders · Headline Analysis			
30 days	6.7	12.2	30
Blog · Infographic · Market Commentary · Rationale · Subscriber Notes			
60 days	13.4	24.5	60
Briefing · Analytics · Spotlight · AltView			
365 days	81.8	149.0	365
Reports (body match) · Service documentation · Flash · Price Assessments · Forward Curves			
No decay	—	—	—
Reports on a name/title match — age ignored entirely			

"Gentle", "moderate" and "sharp" are not parameters — every category runs the same Gaussian at decay 1e-6. Only the scale changes.

SECTION 03

Serving AI agents

one retrieval layer, opened up to agents

03

The same index sits behind copilots, MCP servers and autonomous agents.

IN THIS SECTION

- Supporting Agentic RAG
- Multi-region architecture
- Index sizing at scale

Same index, same API, same pipeline — agents don't get a separate retrieval stack

THE RETRIEVAL LAYER SERVES

- Chat applications
- MCP servers
- Autonomous agents

query

context & citations

STEPS 1 – 3 · UNDERSTAND

Query understanding

Every query is analysed before it reaches the index

- 1 · Query analysis
- 2 · Ontology entity detection
- 3 · Embedding generation

Entity detection also yields the filters and the action to take

STEPS 4 – 5 · RETRIEVE & RANK

One ranked result set

Lexical and vector retrieval run on every query, then merge into one ranked, entitlement-aware list

- 4 · Hybrid retrieval
- 5 · Ranking

Entitlements are applied inside the query, never as a post-filter. Agents and people share the index and the API — not the ranking.

STEPS 6 – 7 · ANSWER

Answer-ready output

Results come back assembled for the model, with sources

- 6 · Context assembly
- 7 · Document citations

Every claim an agent makes can be traced back to a document

BEHIND THE API

Models & ontology

The same models and taxonomy serve every consumer

- Bedrock embeddings
- SageMaker entity models
- SageMaker intent models
- Company ontology & taxonomy

Filtering stays consistent and governed for every consumer

Shared index, one API, one retrieval pipeline · reusable across every AI application

S&P Global Energy

IN EVERY REGION · 01

Enhanced Search API

An independent API layer per region — stateless, and scaled on demand

- Load balancer across two AZs
- Fargate tasks · AZ1 + AZ2
- Same image, same configuration

Kong routes to whichever region is healthy — clients only see one host

IN EVERY REGION · 02

OpenSearch domain

Its own cluster, not a shared one, spread across two availability zones

- Data nodes in AZ1 + AZ2
- Keyword, vector & metadata indices
- Cross-region replication · active-passive

The follower is a live cluster serving queries, not a backup to restore — but failover to it is semi-automated today

IN EVERY REGION · 03

AI services in-region

Embeddings and entity models are called from inside the region

- Bedrock · embeddings & GenAI
- SageMaker · entity & intent models
- Company ontology & taxonomy

Same models and taxonomy either side, so ranking cannot diverge

IN EVERY REGION · 04

State & configuration

Everything the API reads at runtime is regional as well

- ElastiCache · SPOT cache
- Secrets Manager
- S3 config bucket
- CloudWatch logs & metrics

Entitlements, market data and feature flags stay shared platform APIs

WHAT THE DESIGN DELIVERS

- Regional failover
- Disaster recovery
- High availability
- Business continuity
- Consistent entitlements

One stack, deployed twice · the region a query lands in changes nothing about the answer

S&P Global Energy

THE FRONT DOOR

One endpoint

Every consumer reaches the same host — routing is invisible

- Energy Core UI
- Enhanced Search API clients
- Kong · api.platts.com
- Health-based routing

REGION 01 · PRIMARY

us-east-1 — active

A complete stack — API, index, models and cache — serving live traffic.

- Enhanced Search API
- OpenSearch domain
- Bedrock · SageMaker

REGION 02 · SECONDARY

eu-west-1 — also active

The same stack and the same configuration — not a cold standby.

- Enhanced Search API
- OpenSearch domain
- Bedrock · SageMaker

Continuous replication · primary → follower · semi-automated failover

STEADY STATE

Both regions answer

- Traffic served in either region
- Replication keeps both current
- Same entitlements on both sides

REGION LOST

Semi-automated failover

- Routing shifts on health checks
- Promoting the follower is a manual step
- No index to rebuild first

Two active regions, one endpoint · either region can answer every query · entitlements enforced identically in both

THE VECTOR FOOTPRINT · ONE MULTIPLICATION, THREE TERMS

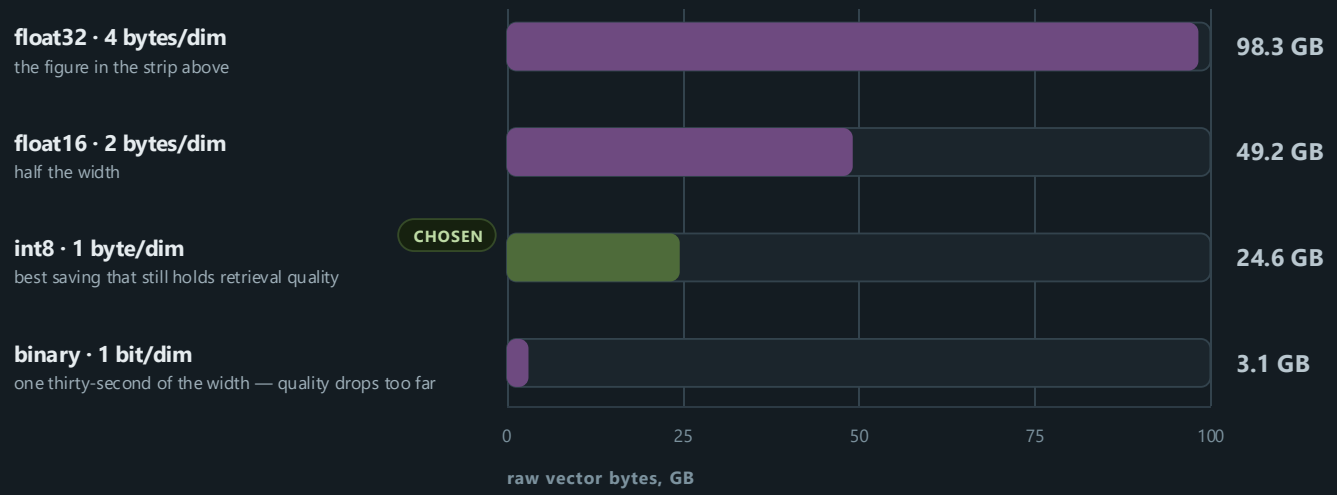
Illustrative · float32 · raw vectors only — not a production figure

24 million chunks × 1,024 dimensions × 4 bytes per float = 98.3 GB of raw vectors

ARITHMETIC, NOT A MEASUREMENT

What the encoding costs, before any tuning

The same 24 million chunks at four vector encodings. This is the byte-width identity — the ceiling on what quantization can buy, not a measured saving.



Compression and selective indexing act on the other two terms — this axis only moves bytes per dimension.

THREE LEVERS

One per term of the equation

The footprint has three terms — each lever attacks a different one.

Selective indexing

fewer chunks

Only content that genuinely needs vector retrieval gets an embedding at all.

Quantization

fewer bytes per dimension

Narrower encodings buy exactly what the chart on the left shows — no more.

Compression

quantized on the way in

Full-precision vectors stay on disk for distribution; OpenSearch only ever holds the quantized int8 copy.

AND VECTORS ARE NOT ALL OF IT

- Chunked document text
- Document metadata
- Access-control metadata
- Generated summaries

Retrieval for AI is a storage problem before it is a ranking problem · raw text is the smallest thing the index holds

S&P Global Energy

Lessons 1-3 are the same three pillars from the opening slide, learned the hard way

LESSON 1 · BUSINESS PILLAR

Relevance over models

Improvements in retrieval quality produced larger gains than upgrading the underlying large language models.

LESSON 2 · LEXICAL PILLAR

Hybrid beats pure vector

Exact matches remain critical in enterprise search, so lexical and semantic retrieval have to run together.

LESSON 3 · GOVERNANCE PILLAR

Entitlements belong inside

Filtering after retrieval adds unnecessary cost and risk. Entitlements must be part of the query itself.

LESSON 4 · PLATFORM

Share the infrastructure

A unified retrieval platform for search and AI teams reduces duplication and accelerates AI adoption.

LESSON 5 · PLATFORM

Explainability is essential

Engineers, product owners and users all need to understand why a given piece of content was returned.

LESSON 6 · PLATFORM

Data leaves the platform too

The API isn't the only output: the same pipeline also publishes an AI-Ready Data schema customers use to build their own RAG.

The next phase moves relevance tuning from hand-written rules to learned, observable signals

01 · QUERY

Understand the query

What the words mean, not just what they say

- Query intent detection
- Taxonomy-driven phrase & entity recognition

02 · USER

Know the user

Who is asking, and what they usually need

- User profile detection & profile-based boosting
- Usage-signal relevancy boosting

03 · AGENT

Retrieval shaped for agents

Multi-step, tool-using retrieval, not a single query

- Advanced reranking & multi-hop workflows
- Agent-driven retrieval strategies
- Document-aware MCP tools

04 · THE POINT

Relevance as a product

Tuning becomes observable, measurable and iterative

- Relevance admin app & automated optimization
- Retrieval observability & AI-guided ranking experiments

Roadmap: retire primary → follower replication for full active-active across regions

Retrieval is the foundation agents stand on — investment here compounds across every AI application

S&P Global Energy

WHAT THE THREE ACTS DELIVERED

Lexical still wins

BM25 + hybrid, not either/or

Relevance is a decision

Business weights, configured not learned

Trusted, not just similar

Entitlements resolved before ranking

One platform, every agent

No duplicated retrieval stacks

THE VERDICT, RETURNED

Retrieval quality is still the highest-leverage investment as AI initiatives keep expanding

Argued at the open — proven across the platform since

Enhanced Search API: one retrieval layer for enterprise search and agentic AI

S&P Global
Energy

THE TALK, CLOSED

Thank you

Questions?

Tomasz Florin & Artur Barczyński · S&P Global Energy · Haystack Conference

S&P Global
Energy