



Vespa Now: Q2 Product Update

Vespa.ai

June 2026

Agenda

01 Vespa Cloud Updates

02 Search & Query Operators

03 Data Model Improvements

04 Ranking & Scoring

05 Agent Skills & Community

Vespa Cloud Updates

Five improvements to the Vespa Cloud platform

Detailed Metric Dashboards

All internal engineering dashboards now available to customers with built-in explanations.

Index Backup

Automated snapshot backups of indexes for catastrophic recovery without re-indexing.

Fine-Grained Maintenance Controls

Control OS-level maintenance windows using the same pipeline as application deployments.

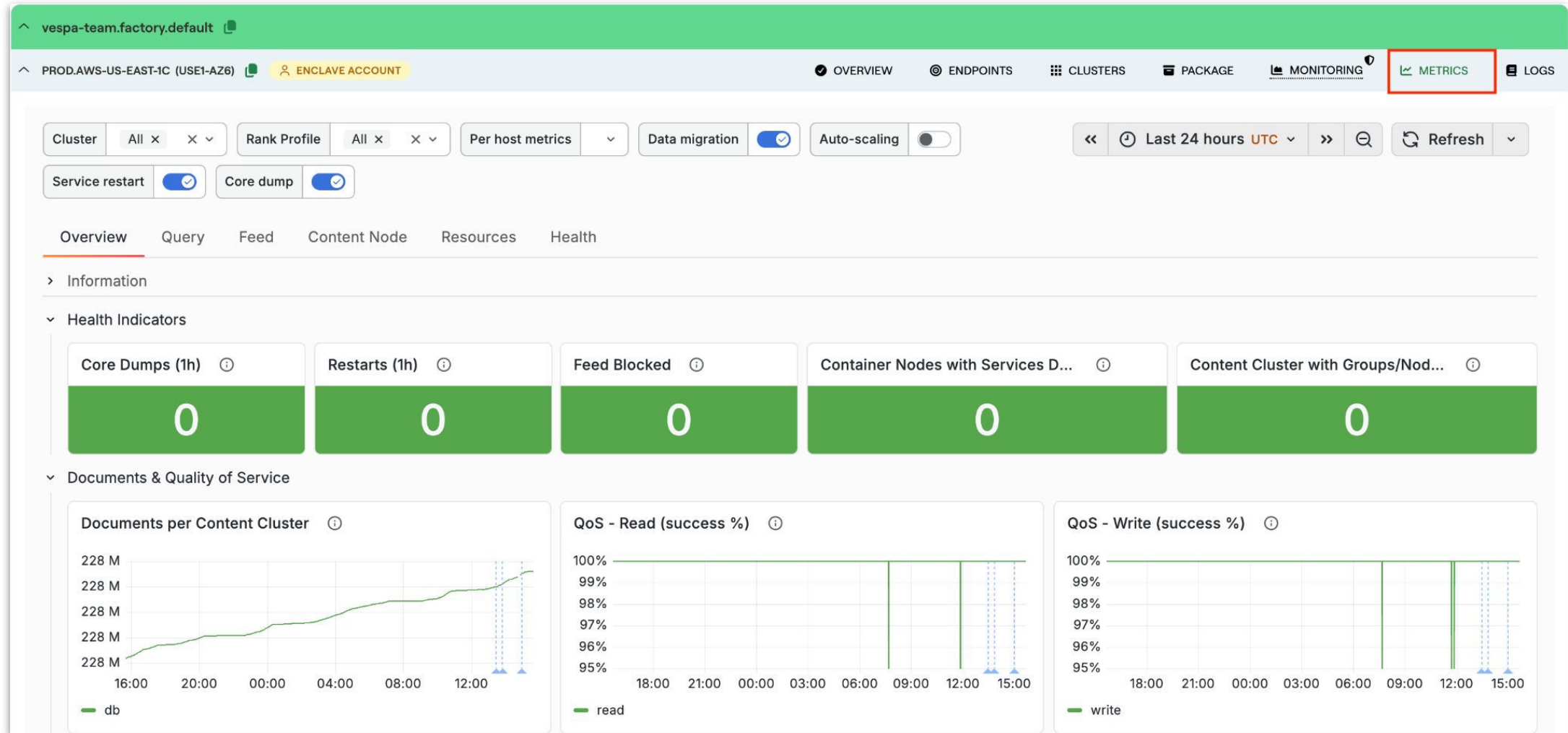
AI Embedding Integrations

Native Voyage AI, OpenAI & Mistral AI embedding APIs for indexing and query time.

Custom Resource Tags

Declare custom tags in deployment.xml for resource tracking and financial monitoring.

Vespa Cloud: Metrics Dashboard



[Read the Blog](#)

Vespa Cloud: AI Embedding Integrations

Voyage AI

- voyage-4 model family
- Use API for docs + local model for queries = zero API calls on query path

[Documentation](#)

OpenAI

- Native API key integration
- Works at document indexing time and query time

[Documentation](#)

Mistral AI

- Mistral embeddings via Vespa Cloud
- Seamlessly integrated into the embedding pipeline

[Documentation](#)

Search & Query Operators

New sample apps

[Bayesian BM25](#) [Hyperencoders](#) [Asia linguistics](#)

Cluster-Size Independent Config

New 'total-' parameter prefix lets you specify values across ALL content nodes. Vespa auto-calculates the right share per node — even during autoscaling.

- `totalTargetHits`
- `total-max-hits`
- `total-keep-rank-count`
- `total-rerank-count`

Near-Matching Aware Ranking

Terms outside the near/onear match window are now excluded from relevance scoring. Only terms appearing within the proximity window contribute to rank scores.

Before: Extra occurrences of term1 outside the window boosted score unfairly

After: Only in-window terms count → more intuitive, accurate relevance

Data Model Improvements

Boolean Array Fields

Pack large amounts of dynamic metadata into memory-efficient bit arrays. Supports `array<bool>` and booleans nested in struct arrays.

```
field my_bits type array<bool>
```

Match Specific Array Elements

Target a specific index in an array during queries. Supports multi-index matching with `elementFilter` in the same query.

```
where my_bits[94085] = true
```

In-Memory Document IDs

From Vespa 8.691, declare document IDs to reside in memory for fast access, similar to attributes, enabling rapid retrieval in queries and visiting.

```
(schema declaration, Vespa 8.691+)
```

Ranking & Scoring Improvements

Search Group Pinning

Consistent pagination across cluster groups

Pin queries to a specific group in multi-group clusters to ensure consistent pagination across result sets. Eliminates relevance drift when paging through results.

firstPhaseMax Rank Feature

Drop low-quality results relative to best-scoring doc

Exposes the highest first-phase rank score among documents on a node during second-phase ranking. Enables relative score thresholding with rank-score-drop-limit.

Geo Distance in Grouping

Group results by distance from any lat/lon point

Group query results by geographic distance bands using `fixedwidth(geo_distance(...))`. Enables location-aware faceted navigation and distance-bucketed result sets.

Agents, Developer Experience & More

Vespa Skills for Agents

A curated collection of skills for coding agents working with Vespa, now open on GitHub.

- Application package management
- Feeding and query operations
- ElasticSearch → Vespa migration

Detect Ignored Write Operations

From Vespa 8.680, the document/v1 API now includes *X-Vespa-Ignored-Operation HTTP* header to flag silently ignored write operations.

Vespa Learn

learn.vespa.ai is a self-paced course that teaches you how to build search, recommendation, and RAG applications with Vespa. You will go from zero to a working e-commerce search engine with hybrid retrieval and machine learning ranking, building it up one piece at a time across six modules.





Vespa.ai Live

Join Vespa's first in-person conference. Featuring speakers from **Walmart, Ravenpack, Etsy** and more.

September 9
Vespa Training

September 10
Conference

Lumiere London



20% off tickets

Q2 Summary

Vespa Cloud

- Detailed metric dashboards
- Index backup & recovery
- Fine-grained maintenance controls
- Voyage AI / OpenAI / Mistral embeddings
- Custom resource tags

Search & Query

- New text() query operator
- Cluster-size independent config
- Near/Onear matching aware ranking
- Search group pinning
- Detect ignored write operations

Data & Ranking

- Boolean array fields
- Match specific array elements
- In-memory document IDs
- firstPhaseMax rank feature
- Geo distance in grouping

Questions?



Vespa.ai