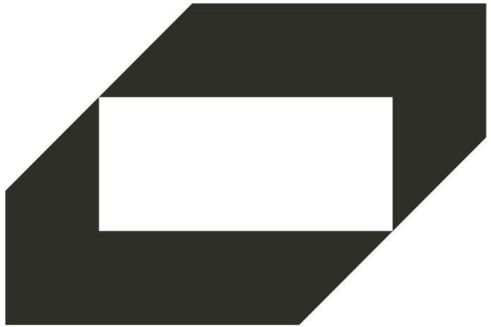
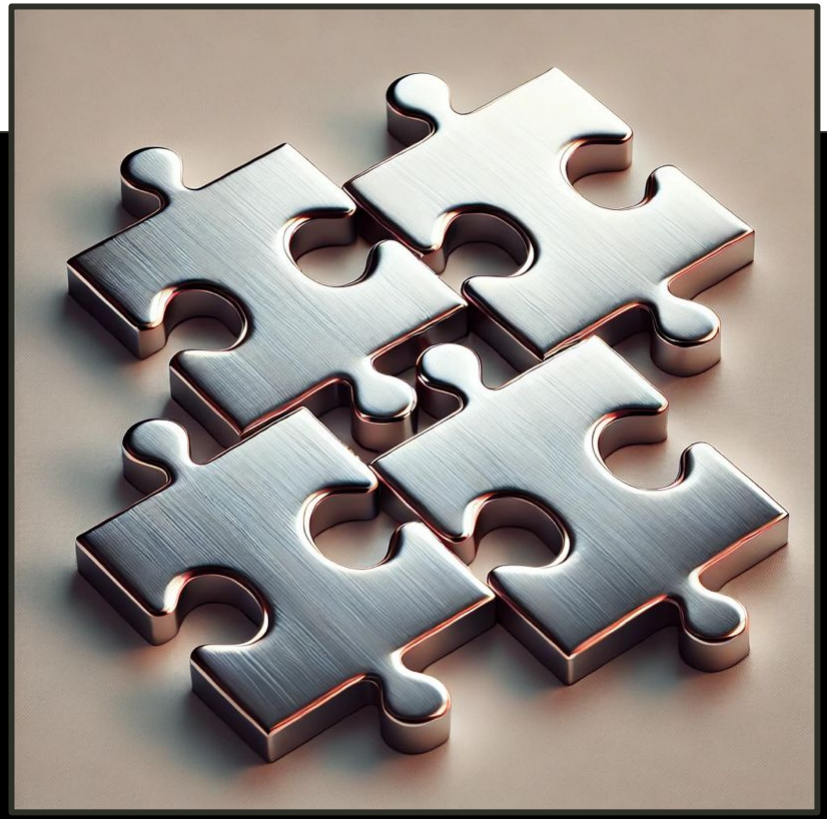


Vespa

Vespa.ai - Application platform for data-driven AI applications



What is Vespa.ai

- Full featured platform for combining AI and data
- Proven over 20+ years at Yahoo and recently worldwide
- Most innovative AI companies like Perplexity.ai, Bigdata.com, Vinted, Spotify are building on Vespa.ai

Why Vespa.ai

- Cost effective at any scale
- Fully managed
- Proven at scale
- Market leading innovation over many years

Getting AI leveraging data out of the lab

Common Barriers

- **Quality:** Achieving consistently high quality requires advanced realtime data processing
- **Operations:** Moving to production requires high availability, scaling data and requests, continuous deployment, upgrades, ...
- **Costs:** Infrastructure and compute resources can skyrocket.

How Vespa Addresses These

- Distributed architecture managing data *and* computations
- Realtime storage, processing and indexing of vectors/tensors, full-text and metadata
- Native data-local tensor and ML model inference
- Managed platform handling all production aspects after source commit

Advanced Tensor Support

Flexible Tensor Model

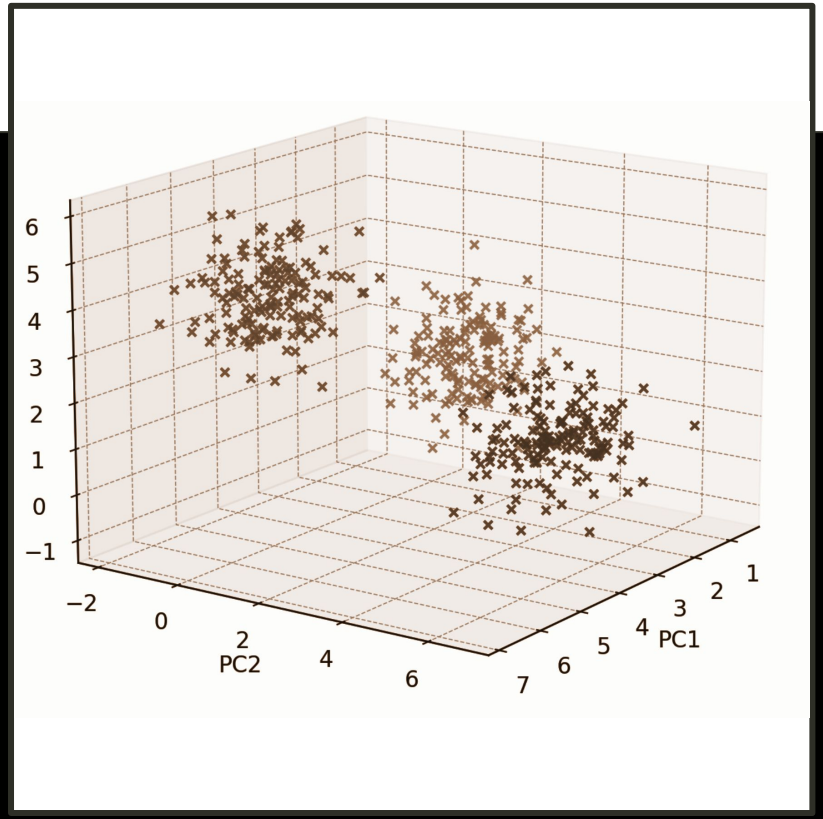
- Efficiently manages both sparse and dense tensor dimensions.
- Unifies scalars, vectors, maps and matrices etc. in one data+computational model.

Core Tensor Operations

- Composable core functions to express any computation over data (map, join, reduce)

Enterprise Impact

- Express any ML model and business logic.
- Minimizes the complexity any time to market for any AI tasks leveraging data.



State-of-the-art ranking/inference

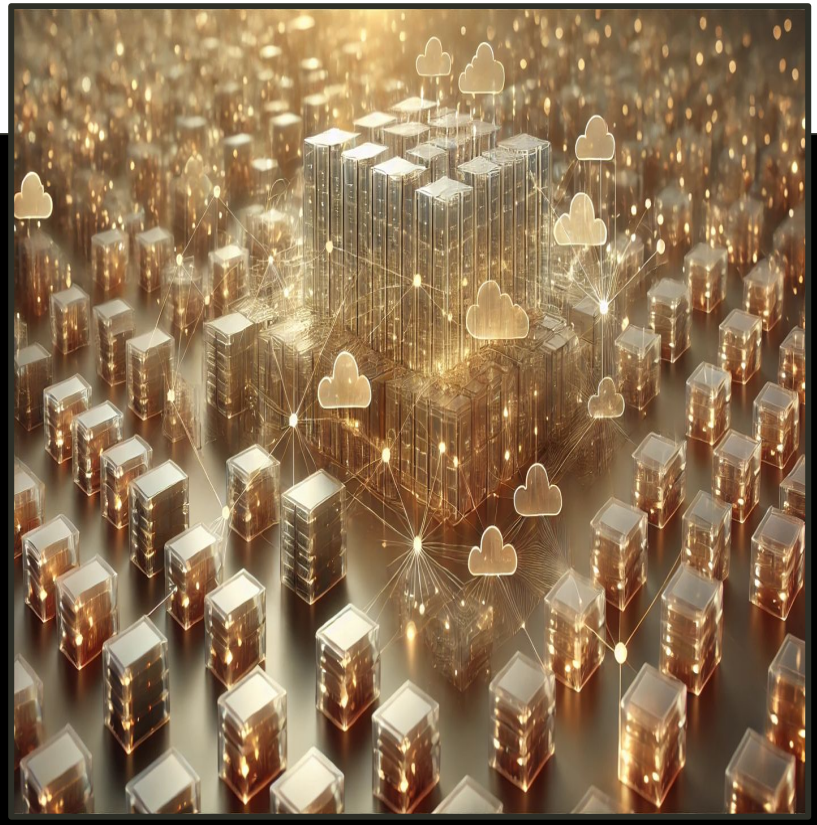
Distributed multi-stage ranking

- Distributed ranking: Avoids data movement.
- Phased ranking: Allocate cpu efficiently.
- Allows any business logic and/or ML model to be applied at each stage.

Dynamic, multi-purpose ranking

- Incorporate real-time signals.
- Integrates efficiently with metadata like e.g user behavior signals.
- Multiple rank profiles over the same data for different use cases, bucket tests etc.

Dynamic Scalability and Data Management



Low Latency

- Millisecond retrieval even with massive datasets.
- Ideal for consumer-facing internet scale and internal enterprise apps.

Vector Optimization

- Vector size reduction techniques like MRL and BQL.
- Reduce memory requirements while maintaining performance.

Late Interaction Models

- MaxSim Hamming distance close to the data layer.
- Flexibly trade accuracy against cost.

AI platform simplification

Unified Architecture

- Combines data processing, indexing, search, ranking and ML models into one platform.
- Easy to extend functionality through chained component architecture.
- Ensures scalability, and eliminates the need for stitching together multiple tools.

CI/CD

- Automated deployments and continuous upgrades.
- Reduces operational overhead and speeds innovation.

Reduced Complexity

- Less time spent managing infrastructure.
- More time for delivering business value.

Emerging Technologies Driving AI

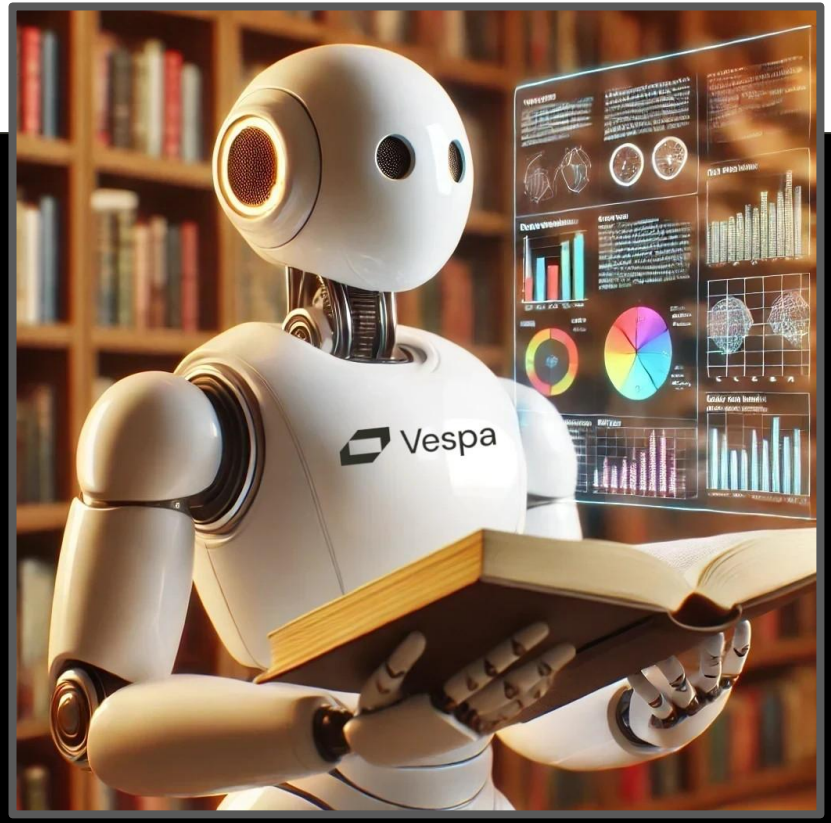
Vision-Language Models

- Vespa's tensor framework can manage complex embeddings at scale.
- Enables multimodal search and richer AI insights.

Future-Proof Architecture

- Vespa's tensor computations make any inference over data easily expressible.
- Positions organizations to leverage cutting-edge research.

Document digitisation at scale with Vespa.ai and colPali



What we will show:

- Moving away from OCR to VLM gives something which was not possible before
- Understanding visuals and text during search
- Being able to find relevant information from mixed media

Benefits of using Vespa.ai

- Query at any scale with minimal latency
- High accuracy not sacrificing speed or cost
- Delivered as fully managed platform