elastic | The Search AI Company

Elasticsearch Capabilities Update

Lovingly Prepared for Docusign

March 2025

Justin Castilla

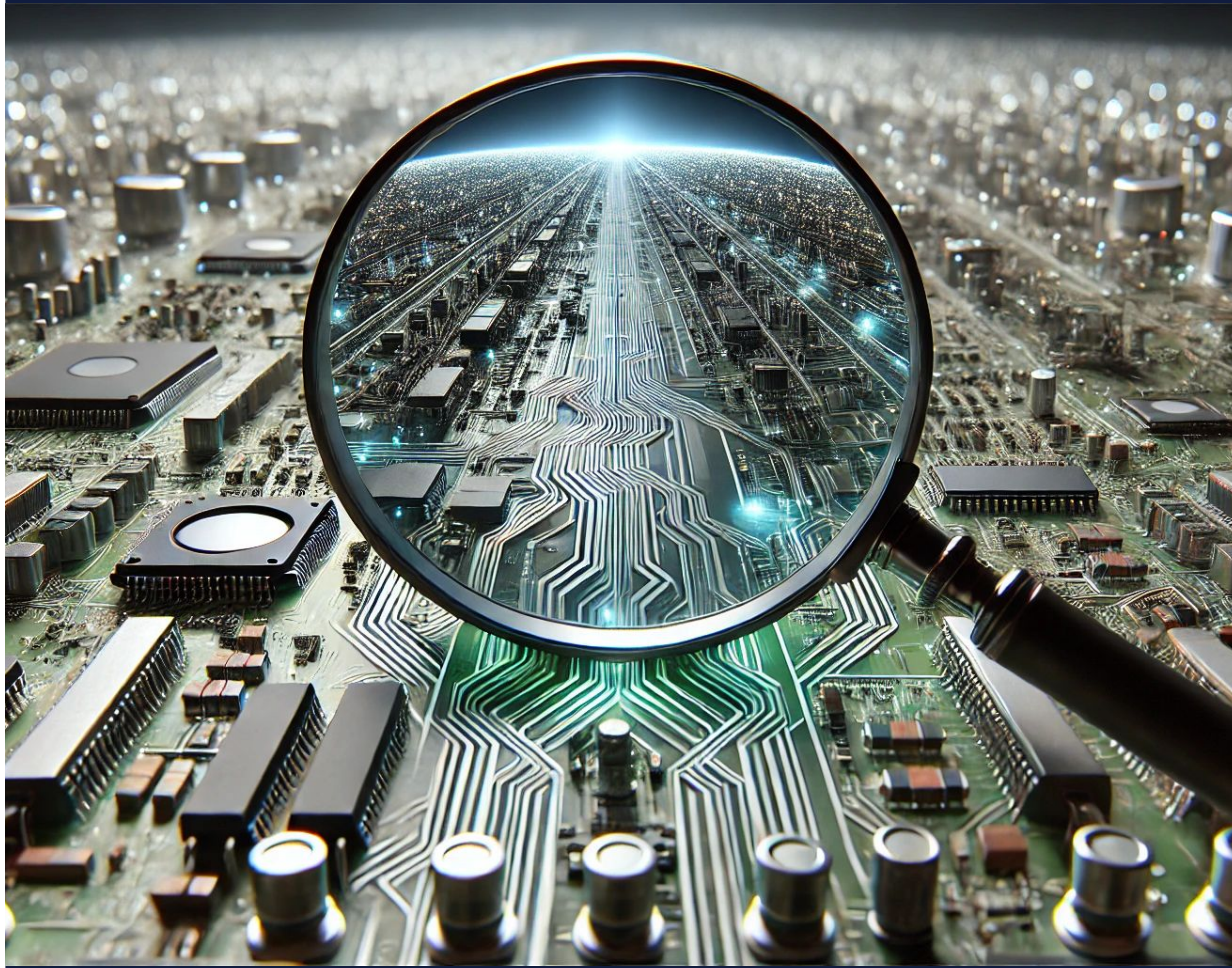
Senior Developer Advocate @ Elastic

Bluesky: @justincastilla.bsky.social

justin.castilla@elastic.co

Talk Agenda:

- Elasticsearch
  - BM25
  - Vectorization of Data
    - ELSER (sparse)
    - E5 (dense)
  - Playground
- Automatic Chunking
- BBQ HNSW
- Demo/Workshop



Introduction to Elasticsearch

- What is Elasticsearch?
- Core Components
  - Indices
  - Documents
  - Mappings
- How search works
  - Inverted Index
- Relevance Ranking

Understanding BM25

- What is BM25?
  - term-based scoring
- Strengths
  - keyword precision, ranking by tf-idf
- Limitations
  - poor at semantic similarity
- Why BM25 still matters in hybrid search
  - Fast, less memory usage, lexical precision

Vectorization of Data

- Why Vectorize?
  - Semantic understanding and context
- Text → Embedding → Search
- Requires an embedding model
- Dense vs. Sparse Vectors
- Use cases:
  - Semantic Search
  - Recommendations
  - Hybrid Ranking

ELSER: Elastic Learned Sparse Encoder

- What is ELSER? (Elastic’s sparse vector model)
  - Inverted Index
    - Weighted tokens (e.g. "clause": 0.8, "document": 0.75, "obligations": 0.6)
- Benefits: explainability, Lucene-native
  - You can inspect which tokens matched
  - Individual scoring
  - How each match contributed to scoring
- TL;DR: ELSER is explainable because it outputs token-weight pairs that can be traced through scoring logic — like BM25, but semantically richer.

E5: EmbEddings from bidirEctional Encoder rEpresentations

- What is E5?
  - The more commonly known style of vectors
  - High dimensional count: 384
  - Not human readable
  - Obscured decision process
- When to use dense over sparse
- Works best for: intent matching, semantic similarity
- Multilingual + domain flexibility
  - May convert query and document languages

E5 vs. ELSER cheat sheet

Feature	ELSER (Sparse)	E5 (Dense)
Model Type	Sparse encoder (token-weight vector)	Dense bi-encoder (vector embedding)
Vector Type	Sparse vector (token → weight)	Dense vector (384 floats, fixed size)
Search Backend	Inverted index (Lucene-native)	Approximate kNN (HNSW / BBQ HNSW)
Explainability	✅ Full (token-level scoring visible)	❌ Opaque (semantic similarity score only)
Storage Overhead	Low (sparse vectors, Lucene-friendly)	Moderate to high (dense vectors stored separately)
Latency	Very low (Lucene-optimized scoring)	Low with vector index (HNSW/BBQ HNSW)
Precision	High for lexical or token-semantic matches	High for meaning and intent similarity
Recall	Good within vocabulary scope	Better on rephrased queries / synonyms
Language Support	English-focused (as of today)	Multilingual (via `.multilingual-e5-small`)
Use Cases	Explainable search, hybrid search with BM25	Semantic search, Q&A, recommendation

Playground: Engage your data immediately

- What is it?
- Great for:
  - Testing embeddings (ELSER/E5)
  - Visualizing ranking
  - Trying hybrid queries
  - Quick walkthrough screenshot or live demo references ;)

[Elastic Playground: Using Elastic connectors to chat with your data]

Automatic Chunking in Inference Pipelines

- Why chunk?
  - Model limits
  - Long docs
- Elastic's chunking settings
  - Default is sentence-based
- How to configure + when it kicks in
  - Sentence size
  - Overlap count
- Limitations
  - sentence count thresholds
  - May need to be tuned out of the box

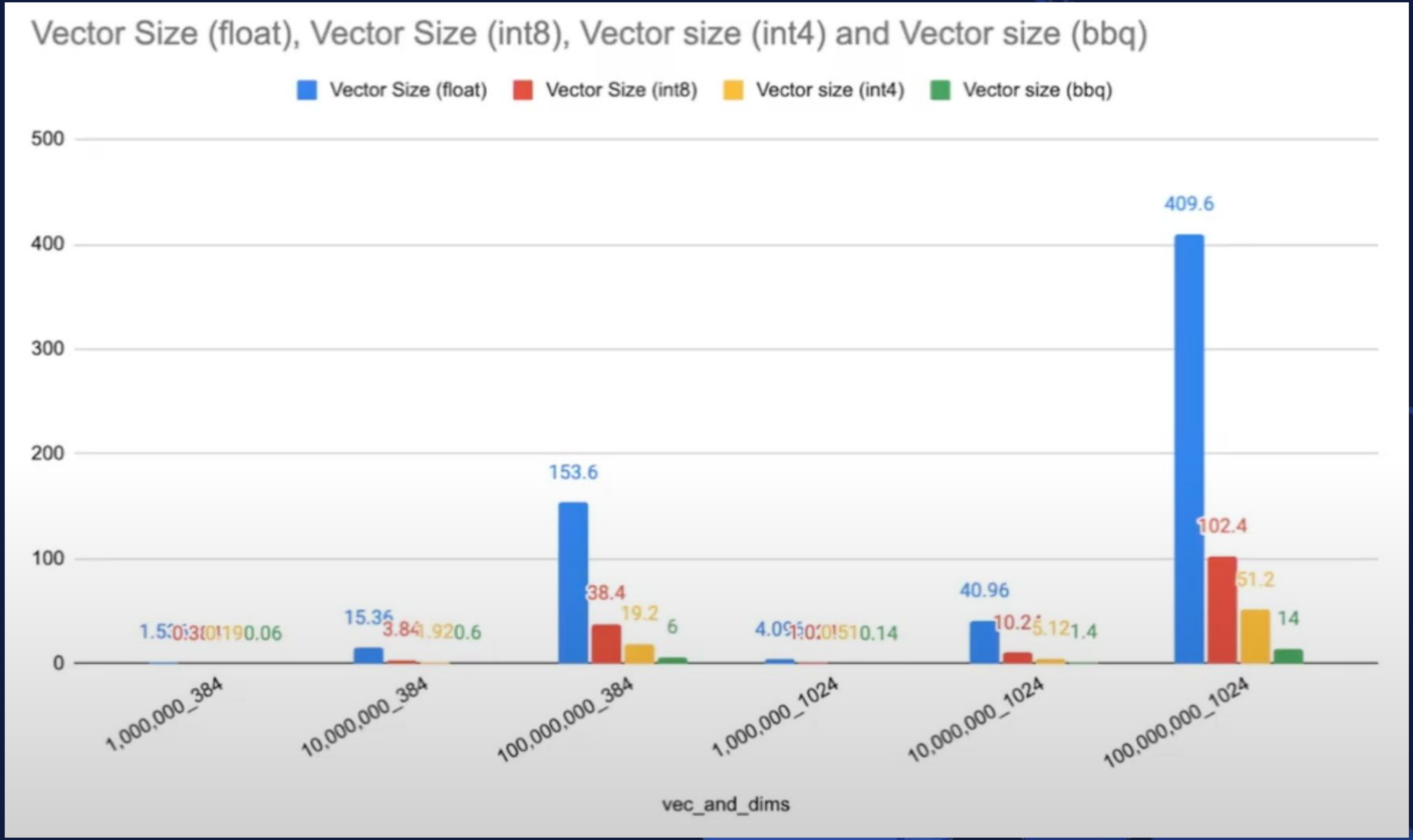
[A recipe for GenAI powered search (RAG) on your PDF treasure]

BBQ HNSW: Better Binary Quantization

- What is BBQ?
- Advantages over classic HNSW
- Smaller memory, faster retrieval
- How to enable it
  - index_options: bbq_hnsw
- Works with dense vectors

A recipe for GenAI powered search (RAG) on your PDF treasure

BBQ HNSW: Better Binary Quantization



Better Binary Quantization (BBQ): From Bytes to BBQ, The Secret to Better Vector Search by Ben Trent

Demo/Workshop

- Manual chunking → embedding → indexing
- Semantic + keyword + hybrid search in action
- Inspect inference output (predicted_value)
- Index settings, mapping, and knn search
- Playground

How to get involved with Elastic in Seattle?

- I am the Pacific Northwest Developer Advocate
- Join the Seattle Elastic Meetup group
- Speak at Elastic meetups in person or virtually!
- Let me know if you'd like another talk or a deeper dive at your own meetup