

Generative AI Technology and Product Update

July 17, 2023

Forward Looking Statements

This presentation and the accompanying oral presentation contain forward-looking statements that involve risks and uncertainties concerning our expectations on the impact of artificial intelligence ("AI") on the technology landscape and the expected strength, performance or benefits of our offerings in the AI space. Actual outcomes and results may differ materially from those contemplated by these forward-looking statements due to uncertainties, risks, and changes in circumstances, including, but not limited to those related to our ability to continue to deliver and improve our offerings in the rapidly changing AI space and our ability to generate revenue from our offerings. Additional risks and uncertainties that could cause actual outcomes and results to differ materially are included in our filings with the Securities and Exchange Commission (the “SEC”), including our Annual Report on Form 10-K for the most recent fiscal year and any subsequent reports filed with the SEC. SEC filings are available on the Investor Relations section of Elastic’s website at ir.elastic.co and the SEC’s website at www.sec.gov. Elastic assumes no obligation to, and does not currently intend to, update any such forward-looking statements, except as required by law.

elastic

Generative AI at Elastic

Product Update



Ken Exner
Chief Product Officer



Matt Riley
GM, Elastic Search



Nick Chow
Principal Product Manager
Elastic Search



Shay Banon
Founder and
Chief Technology Officer

elastic

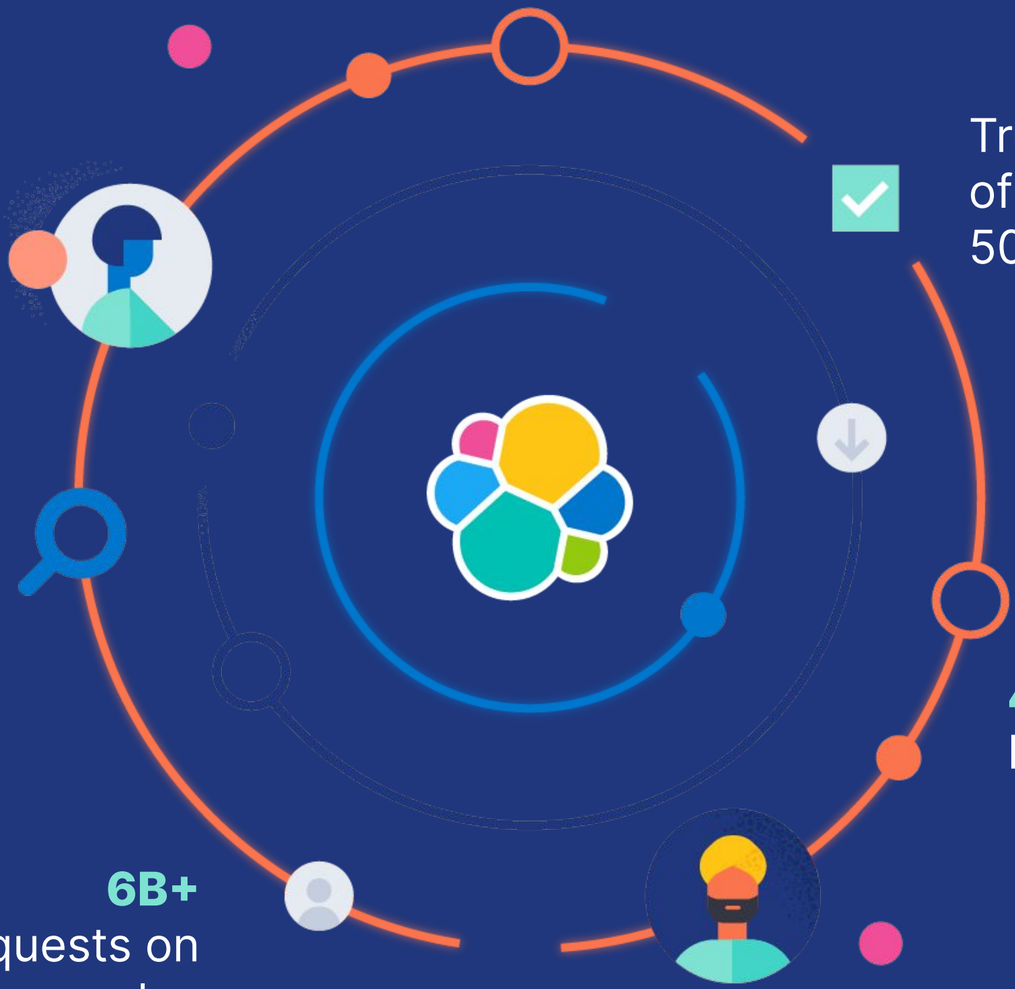
Elasticsearch

The World's Most Popular Search Engine Software

103K+ GitHub Stars

148K+ Commits

Trusted by **50%** of the Fortune 500



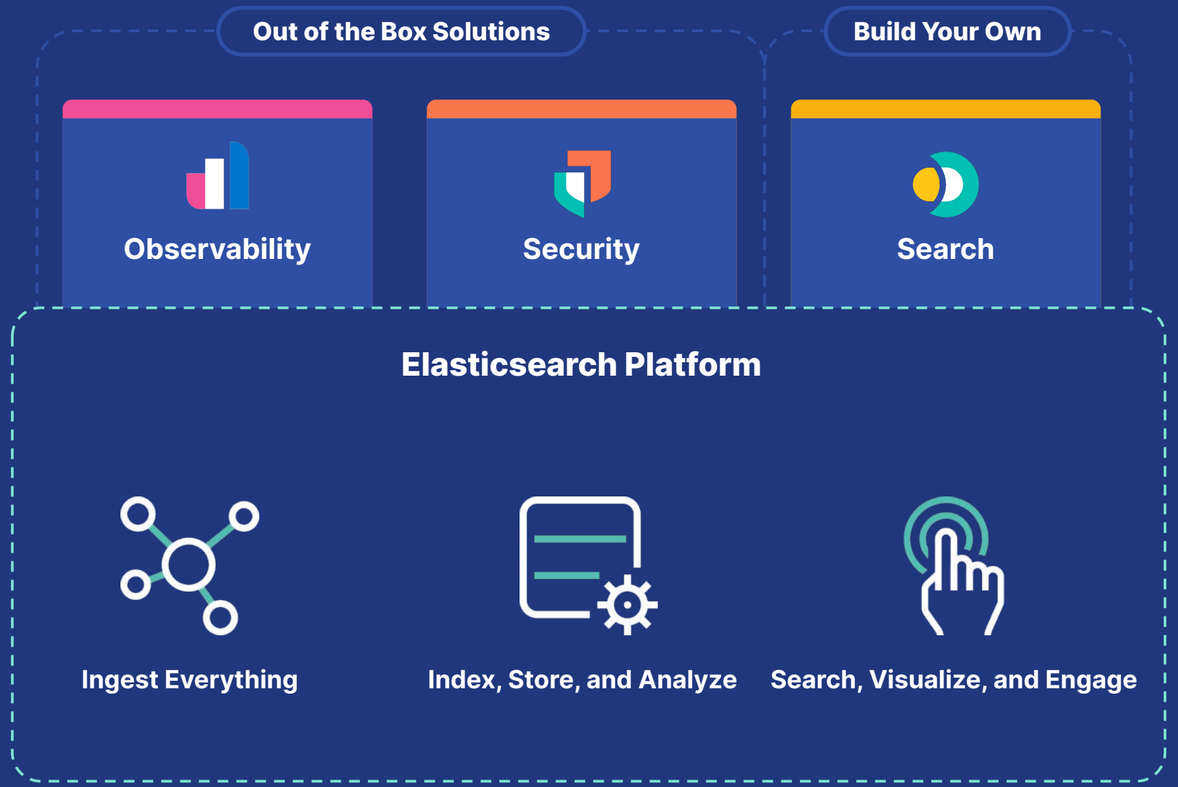
4B+
Downloads

6B+
Search requests on Elastic Cloud everyday

3500+ Unique Authors

elastic

One Data Platform
Out-of-the-box Solutions
And the Freedom to Build Anything



elastic

Generative AI has initiated a dramatic transformation of the technology landscape



Novel Interaction Model

Context-aware chat interfaces, with powerful natural language understanding ability



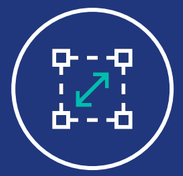
Rising Expectations

Rapid, widespread consumer adoption has transformed consumer expectations



Emerging Ecosystem

A new ecosystem of technologies, startups, and development methodologies



Compelling Event

Enterprises in every industry vertical are rushing to understand and react to changing landscape

elastic

History of AI at Elastic



Foundational technical building blocks within the Elastic Stack

- 2017 — **Support for supervised and unsupervised learning**
- 2019 — Native support for Vector Search
- 2022 — Support for NLP and BYO Transformer models
- 2023 — ELSER - a proprietary zero-shot retrieval model
- 2023 — **ESRE** Elasticsearch Relevance Engine



AI-powered features that enhance our solutions

- 2018 — **Support for forecasting in Observability**
- 2019 — **Support for anomaly detection and AIOps**
- 2020 — **Support for ML-powered detection rules in Security**
- 2021 — ML-powered alerting rules
- 2023 — **Elastic Security Assistant** powered by ESRE

elastic

Elastic provides the bridge between private data and GAI

Sensitive databases
Multi-system / cloud information
Private knowledge bases
Case histories







elastic

Limitations in domain-specific topics can be addressed with various techniques



Fine Tuning

Further training the base LLM with the domain-specific content that is specific to the use case. Expensive to train, expensive to host custom models, and need to be continuous.

Retrieval Augmented Generation

Passing additional, relevant content to the LLM, alongside the original question, as a "context window".
Straightforward, but context window size is limited.

Generative AI

What are the key aspects of the company's 401k policy for an employee in my location and how do I enroll?

original question



- ✓ Takes action and creates new content
- ✗ Generic context
- ✗ Hallucinations – confident tone, erroneous info

elastic

Generative AI + Retrieval Augmented Generation

What are the key aspects of the company's 401k policy for an employee in my location and how do I enroll?

Question as search query

Elasticsearch

Most relevant data to the query

original question

context window



- GAI response based on the most relevant data
- Additional user personalization
- Reduces required compute resources

Domain Specific, Private Content

elastic

Elasticsearch Relevance Engine™

Gives developers a powerful set of **machine learning tools** to **build AI-powered search applications** that **integrate with large language models**

- Vector database
- Ability to host your own transformer model
- Ability to integrate with 3rd party transformer models (OpenAI)
- RRF - hybrid scoring model (vector & textual search)
- Elastic's proprietary ML model
- Integration with 3rd party tooling like LangChain

elastic

Four **key investments** that add up

Vector Database

Native support for fast similarity search in vector space, alongside the existing retrieval models built into Elasticsearch

Bring-your-own Transformer Model

Create vector embeddings from natural language text directly in your Elasticsearch deployment

Native Transformer Model

Elastic's proprietary model, ELSER, improves out-of-the-box ("zero-shot") relevance for any text-based retrieval application

Integrated Relevance Engine

Intuitively integrated into `_search` API and RRF for best-in-class ranking, filtering, document-level permission enforcement

Vector Database

- Fast similarity search in vector space, alongside the existing retrieval models built into Elasticsearch
- Native implementation of the HNSW algorithm for approximate nearest-neighbor search
- Implemented directly in Lucene, the best, most widely-distributed relevance engine in the world.
- Integrated Filtering avoids both pre- and post-filtering limitations

Bring-your-own Transformer Model

- Create vector embeddings from natural language text directly in your Elasticsearch deployment
- Fully integrated workflow, avoids syncing between disparate tools and data sources, and captures full computational workload
- Positioned for interoperability with open source & open model capabilities
- Enables varying levels of sophistication, from teams with data science expertise and those without

Native Transformer Model

Elastic's proprietary model, ELSER, improves out-of-the-box ("zero-shot") relevance for any text-based retrieval application

Brings state-of-the-art relevance to the entire existing Elasticsearch customer base and community

Simple to use and does not require corpus-specific training or data science expertise

Integrated Relevance Engine

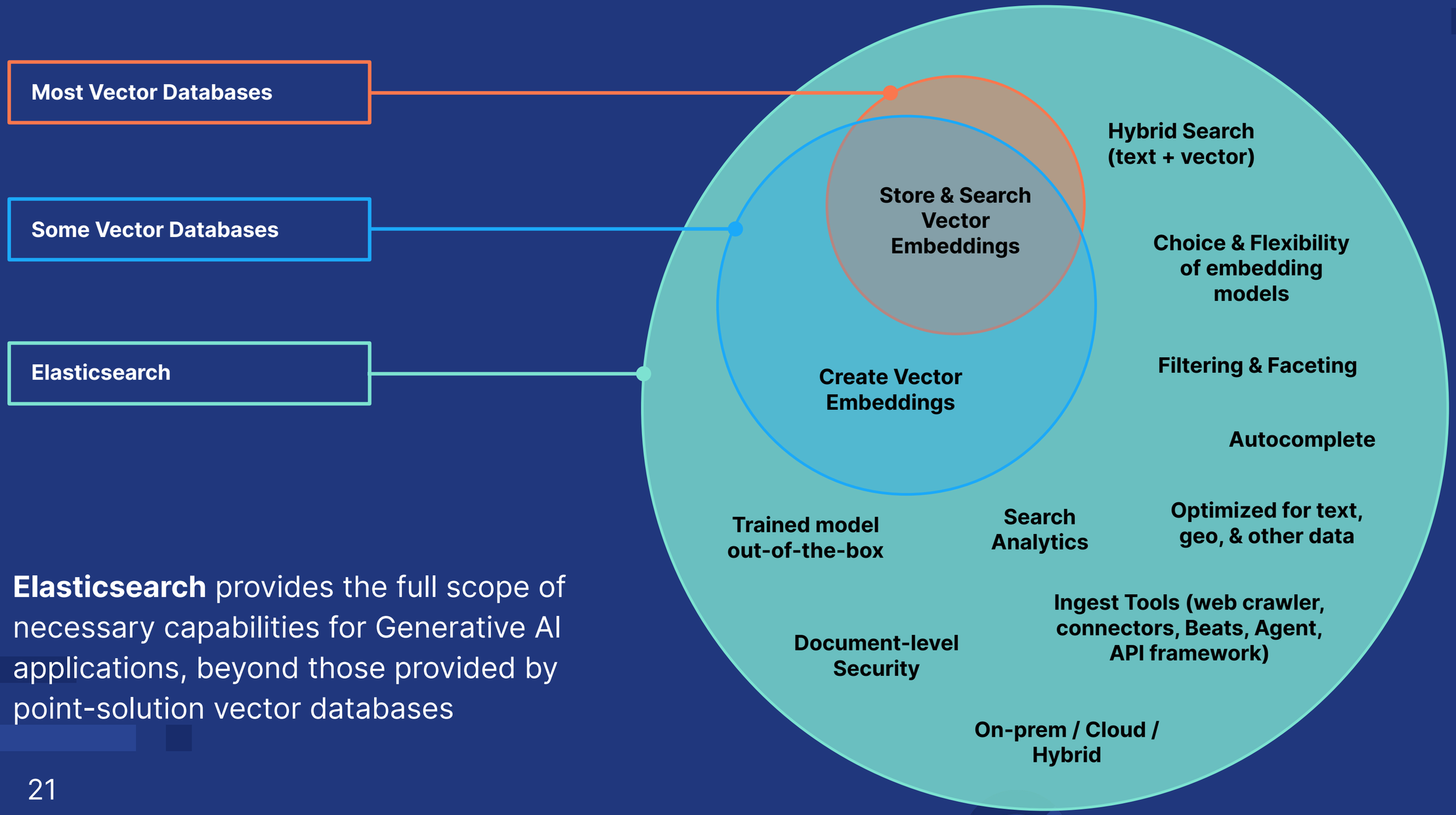
Intuitively integrated into `_search` API and Reciprocal Rank Fusion (RRF) for best-in-class ranking, filtering, document-level permission enforcement

Privacy-first approach w/ native document-level security

Hybrid search capabilities combines the best of text- and vector-based retrieval methodologies

Best-in-class relevance enables retrieval augmented generation optimized for performance, accuracy, and cost

Broad scope of capabilities needed for Generative AI



Elasticsearch provides the full scope of necessary capabilities for Generative AI applications, beyond those provided by point-solution vector databases

elastic

Elasticsearch: production ready search at scale

Deliver at speed

Get up and running in minutes, provide relevant results at the speed

Proven at scale

Wide install base across industry, geo, and use case

Built-in security

OS Hardening, network security controls, encryption, RBAC

Resilient

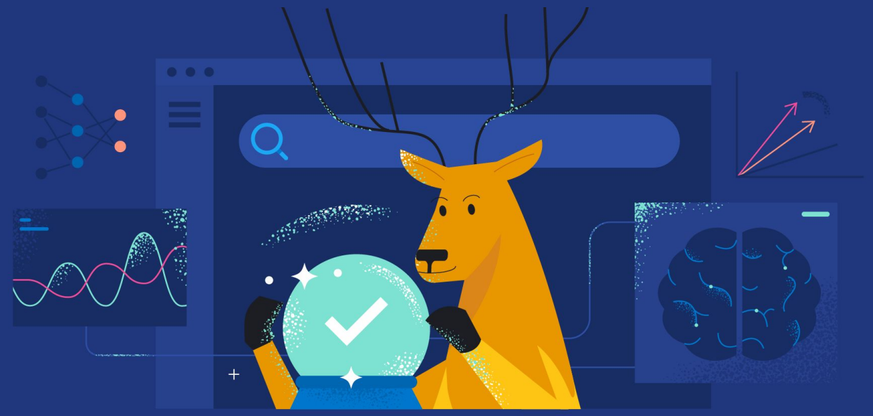
Enterprise ready, backed by support

Flexible deployment

Choose a cloud, hybrid, on-prem, self managed, multi-country

elastic

Generative AI with Elastic



Superior relevance

Bring your own transformer models into Elastic, integrate with third-party models, or use Elastic's Learned Sparse Encoder.

Factual, updated, & contextual data

Use synced data in Elastic for accurate, up-to-date, business-specific results that pass via context window to GAI models.

Privacy and security

Apply Elastic's native support for role-based and attribute-based access control for chat and question answering apps.

Cost effective

Enable precision context windows with Elastic to reduce data footprint and expense.

Generative AI + Retrieval Augmented Generation

What are the key aspects of the company's 401k policy for an employee in my location and how do I enroll?

Question as search query

Elasticsearch

Vector Database | Transformer Models | ELSER | RRF

Most relevant data to the query

original question

context window



- GAI response based on the most relevant data
- Additional user personalization
- Reduces required compute resources



Domain Specific, Private Content

elastic

In Summary

- Elastic provide the bridge between companies' private data and public LLMs
- Retrieval Augmented Generation (RAG) is the best way to pass context for Gen AI
- RAG requires vector search, relevance modeling, transformer integration, inference APIs, and more
- Elastic provides RAG with superior relevance, privacy, security, personalization, cost, scale, and flexibility.