



SUCCESS STORY

UNITED STATES

SOFTWARE & TECHNOLOGY

ELASTICSEARCH

Contextual AI delivers next-generation enterprise AI powered by Elastic

When it comes to [enterprise AI](#), [RAG \(Retrieval Augmented Generation\)](#) is a cornerstone technique creating massive impact for enterprises. The technology enables organizations to incorporate proprietary data with [large language models \(LLM\)](#) to deliver unique, context-dependent AI experiences for employees and customers. End users can interact directly in a familiar chat experience or embedded directly into enterprise workflows and apps, solving a range of tasks,

including enterprise knowledge search, compliance and risk analysis, and customer support.

However, traditional techniques using frozen RAG components are still prone to hallucinating even after fine-tuning and other accuracy boosts. This is where [Contextual AI](#), supported by Elastic, takes AI projects stuck in POC and makes them production ready.



Achieves RAG accuracy of 90%+

Agents built with Contextual AI achieve 90%+ accuracy, powered by Elastic's effective retrieval.



Scales to handle massive volumes of content

With Elastic, Contextual AI's platform operates across millions of diverse documents, managing repositories with 22 million chunks.



Boosts efficiency

With Elastic, Contextual AI takes advantage of its multi search API enabling multiple search types, including semantic and keyword search, using a single API call.

Contextual AI provides a unified context layer that effectively feeds enterprise data and relevant business context into LLMs, centered around advanced RAG and context engineering techniques. The platform includes models, components, and workflows that have been specifically designed for accuracy in complex knowledge tasks, including document understanding, retrieval, reranking, query planning, grounded generation, attribution, and tool use. Elastic's vector and semantic search technology plays a critical role, enabling the Contextual AI platform to deliver its solution at scale to enterprise customers.

Select Agent Template

Speed up agent creation with pre-configured templates. Choose from our curated templates for common use cases like customer support, financial analysis, and company Q&A, or save your own custom configurations for future use.

Default



Company Policy Q&A

Customer Support

Financial Analyst

Legal Researcher

Root Cause Analysis Copilot

Taking RAG to the next level

Mathew Hogan, Tech Lead, Contextual AI, identifies three areas where Elastic enables the company to push the boundaries of context engineering and RAG. First, most RAG solutions combine various open-source tools and functionalities to try and improve accuracy and minimize hallucinations, with typical accuracy results in the 65% to 75% range.

Contextual AI's advanced RAG system is built on a unified platform underpinned by Elastic, allowing customers to quickly optimize and efficiently fine-tune their agents to achieve the 90%+ accuracy required for high-value and complex production use cases.

The second key advantage is scalability. Contextual AI's agents operate effectively across millions of documents. Its largest individual data repository contains approximately 14 million chunks, sourced from more than 60,000 documents. To manage this, the company uses a custom document understanding pipeline that chunks and embeds the data, then indexes it into its [Elasticsearch cluster](#). Typically, the incoming data is [unstructured](#) and multimodal, comprising PDFs, HTML files, and documents that contain images, tables, schematics, diagrams, and charts. [Elastic's vector database](#) handles it all.

[Hybrid search](#) is the third reason Contextual AI chose Elastic as the foundation for its vector database and search technology. Elastic supports both BM25 keyword algorithms and embedding-based search methodologies. With BM25, Elastic provides a range of useful features, including customizations, synonym support, and stemming—allowing for the consideration of various word forms.



“Elastic’s comprehensive support for BM25, combined with its vector search capabilities within the same database, means we can conduct both types of searches simultaneously without the complexity of managing separate services.

Junaid Saiyed

Head of Engineering, Product & Design, Contextual AI

Based on my analysis of Apple's Form 10-Q filing for Q1 2023, I can provide a comprehensive overview of the company's business performance during this period:

Apple's total net sales decreased by 5% or \$6.8 billion during the first quarter of 2023 compared to the same quarter in 2022, primarily attributed to the weakness in foreign currencies relative to the U.S. dollar. 2

Let me break this down into key performance areas:

- **Products Segment:** The company's Products gross margin decreased during the first quarter of 2023 compared to the same quarter in 2022, primarily due to weakness in foreign currencies relative to the U.S. dollar and lower Products volume. 1
- **Services Performance:** Services net sales increased during the first quarter of 2023 compared to the same quarter in 2022, driven primarily by higher net sales from cloud services, the App Store, and music. 1

Wearables, Home and Accessories
Wearables, Home and Accessories net sales decreased during the first quarter of 2023 compared to the same quarter in 2022 due primarily to lower net sales of AirPods, partially offset by higher net sales of Watch.

Services
Services net sales increased during the first quarter of 2023 compared to the same quarter in 2022 due primarily to higher net sales from cloud services, the App Store® and music.

Gross Margin
Products and Services gross margin and gross margin percentage for the three months ended December 31, 2022 and December 25, 2021 were as follows...

Research, deployment, and flexibility

The Contextual AI research team also uses Elastic's vector database to test and compare embedding models using [semantic search](#) through vector similarity queries. By indexing data as embeddings and running [approximate nearest neighbor](#) searches, they can evaluate how well different models capture meaning and context in tasks such as FAQ retrieval, product search, or customer support. This setup allows for scalable, real-world testing and supports hybrid search to reflect more complex use cases that combine semantic understanding with traditional filters.

Gurnoor Singh Khurana, Member of Technical Staff, Contextual AI, stresses the value of consistency between research and production: "A significant advantage for us is that our platform team also uses Elasticsearch as their deployment solution. This ensures alignment between the research and platform environments."

This means that if a particular approach works well during the research phase, Contextual AI can be confident it will scale seamlessly into production, avoiding complications that can arise from using disparate database technologies.

Another strength of Elastic is its multi-cloud flexibility. Even though Contextual AI primarily operates on Google Cloud, it can easily deploy to customers' preferred AWS or Azure regions. Elastic's self-hosting capabilities are also crucial for clients who need on-premises or VPC solutions due to strict data compliance regulations and cloud migration policies.

Hogan also highlights Elastic's comprehensive API coverage. From saving snapshots and inspecting indices to running complex searches, the team can handle everything through well-documented APIs without building custom tools. Of particular interest is the multi search API, which supports multiple types of searches, including hybrid search, in a single API call. This streamlines their workflow and makes working with Elasticsearch both efficient and developer friendly.



The versatility of Elasticsearch is a significant asset. It provides us with sales flexibility and the agility to rapidly accommodate novel deployment requirements from our customers.

Junaid Saiyed

Head of Engineering, Product & Design, Contextual AI

Start your free trial

See for yourself how your business can benefit from Elastic in the Cloud, with a free 14 day trial.

[Get started](#)