How Generative AI is Shaping the Future of Search

MongoDB.

Table of Contents

Introduction

From the beginning of the web with dial-up, technological advances have come in waves of innovation and disruption. Looking back from the early innings of the web with dial-up connections that allowed information access from every household, to the rise of the cloud enabling a whole new era of mobile computing, technological progress has continued to develop at a rapid clip. It wasn't just the velocity of flowing bits that changed, but the form and format too: from desktops to laptops, laptops to phones, phones to tablets, and from mainframes to the cloud. Technology has continued to evolve impacting every industry in the modern economy.

Search technology is no exception. The pages of ten blue links have remained relatively consistent for decades, dominated by a near monopoly of a few large global players. Now with the recent rise of generative AI and Large Language Model-powered search, the status quo looks to be headed towards obsolescence marking a new wave of disruption for the world of search. Today's tech continues to evolve, yet many businesses have been unable to innovate or adopt these new tools as quickly as they are released. Yet some fundamental questions are now at the forefront: How do you provide maximum speed and relevance to ensure the best possible web experience? How do you make sure your customers receive the right information? Where does AI fit in? What actually is vector search and why does it matter?

In this white paper we'll look at where search technology is today, where it's going, and will provide examples to ensure your organization is leading the charge rather than playing catch up. We'll also pose a few critical questions every business should be asking to ensure a future-proofed search strategy.

Let's dive in!

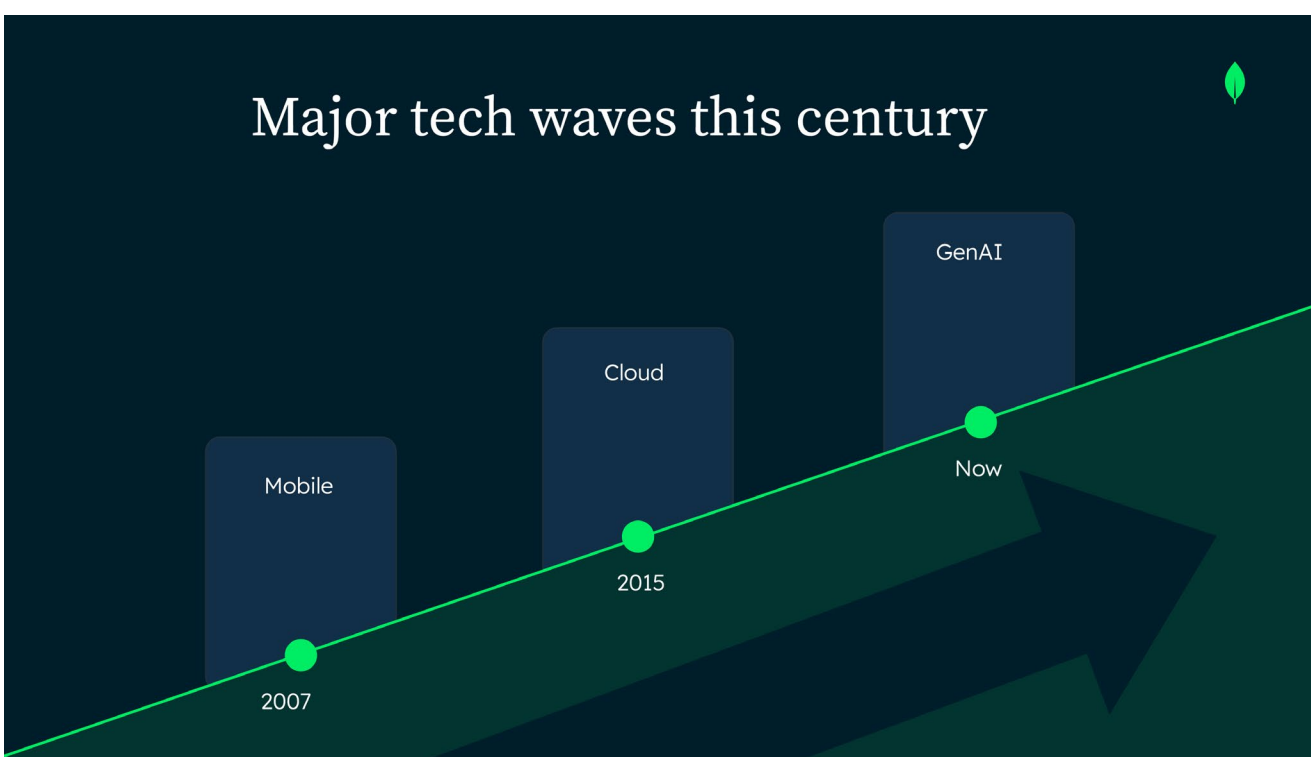

Figure 1. 3 recent major waves of technology

A look at where we've been

The waves of technology mentioned above have all had profound changes not only for society, but specifically for the enterprise. The iPhone, it was proclaimed, would never appeal to businesses, yet the advent of the app store unleashed a wave of enterprise mobility like BYOD (bring your own device), leaving IT teams scrambling to keep up with consumer-centric trends. Similarly, the flexibility of the cloud unlocked new paradigms of enterprise software, giving rise to innovative companies such as Workday, ServiceNow, Salesforce and more, upending incumbent on-premises category leaders in the process.

Search has experienced a similar trajectory, growing from the purview of PhD dissertations to trillion dollar publicly traded companies. The implications for the enterprise have also changed, beginning with rudimentary relational databases using regular expressions (or regex, which was developed back in the 1960's!) to execute crude search queries and match relevant content.

These more utilitarian search methods gave rise to a new breed of "site search" tools, which are a set of technologies and functionality that enables users to search a website's content or product suite with increased speed and relevance. Site search has been especially impactful when it comes to web properties containing a large volume of content to sift through, or for ecommerce companies that constantly need to ensure their product set is current and up-to-date (or risk losing sales!). Popular site search tools including Coveo, Bloomreach, Swiftype, and even Google's Site Search 360 have been some of the key players in years past that placed an emphasis on relevance and helping web users find and discover content. This content might have been buried or hard to find in a given vendor's site. These search tools helped reduce bounce rates, improve user experiences, and ultimately boost conversions.

However, as the move to the cloud continued to become mainstream, these site search-specific vendors became increasingly overtaken by new API-based search specialists that were responsive and adaptive for the shifting technology. Search vendors such as Elasticsearch, Algolia, and Amazon Opensearch were purpose built to take advantage of the always up-to-date cloud architecture to increase relevance, and better understand behavior through user friendly features such as typo tolerance, synonym matching, autocomplete, and more. The advantages of always being on the most up-to-date version with these distributed cloud technologies meant reduced costs, lower maintenance, and a more predictable total cost of ownership.

Future-proofing from the start

At MongoDB we created Atlas Search to give our customers a simple, seamless, and scalable experience for building relevance-based app features that automatically sync with your database. We saw the variety of search use cases our customers had and created a native search experience that solves for limited functionality, performance overhead, high complexity and synchronization issues common in standard database search or third-party bolt-on solutions.

We deliver a fully robust search experience and full functionality including fuzzy search, synonyms, dedicated infrastructure, query analytics, and much more.

This has helped to "future-proof" Atlas Search by tightly integrating all aspects from day one: a search and database experience automatically in sync, with one query language, eliminating the need for extract, transform, and load (ETL). This

also simplifies your tech stack by removing the need to secure, scale, monitor, or back up common with standalone solutions. This unified platform has helped our customers recognize tremendous savings, including getting 10% more time back or decreasing latency times by 70%. We built Atlas Search with a specific focus on a simplified data architecture, driving higher developer productivity, all in one fully managed platform. This architecture and the Atlas platform helped to set the stage for the next generation of search experiences with vector search.

The rise of vector search

Text search remains a critical component of any engaging app or web experience, as ensuring relevant information is accessible and accurate is a top priority for businesses across industries. But text search remains quite linear, as a user searches for some requested information and the business returns relevant results based on the keywords used. Now, with the recent advances in technology and large language models (LLMs), a new search modality has arisen in the form of vector search, which helps solve the challenge of providing relevant results even when the user may not know what they're looking for.

Vector search is a newer form of retrieval that allows you to take any type of media or content, embed it into a vector using machine learning algorithms, then search to find results similar to the target term. The similarity aspect is determined by converting your data into numerical high-dimensional vectors, then calculating the distance between them to determine relevance.

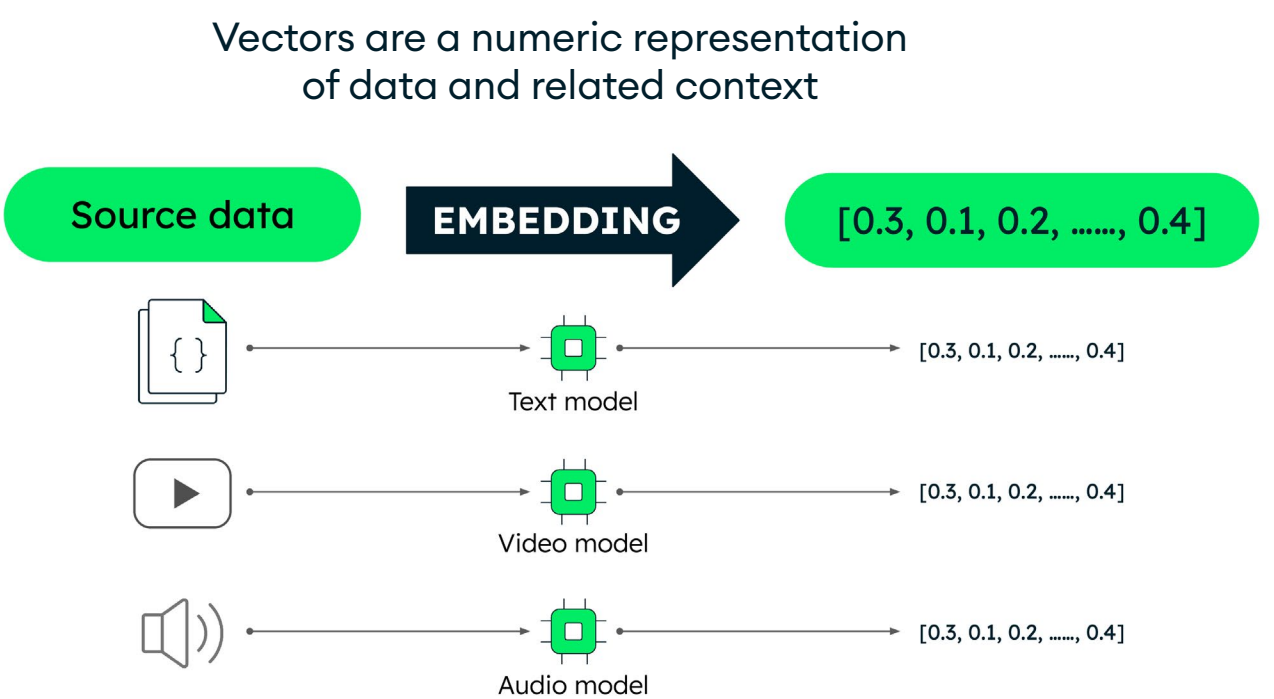

Figure 2. Diagram of vector search

The first key use case for vector search is what's known as semantic search, which helps you find relevant results from unstructured data based on the semantic similarity, or meaning and relationship between items. Rather than looking for specific keywords, the secret sauce with this approach is looking for similarity based on how far away that item is from the target, measured by the distance between vectors.

Semantic search unlocks a whole new world for searching unstructured data – think information buried in audio and video files, PDFs, social media posts, e-books and more. Today nearly 80% of all data is in this unstructured format, with no signs of slowing down in the future, with semantic search helping to make this data discoverable. We've seen customers now enable their users to ask natural language questions and get relevant responses with vector search, providing a much more engaging user experience compared to traditional keyword searches (and seeing a 2x to 3x increase in productivity with the one unified system).

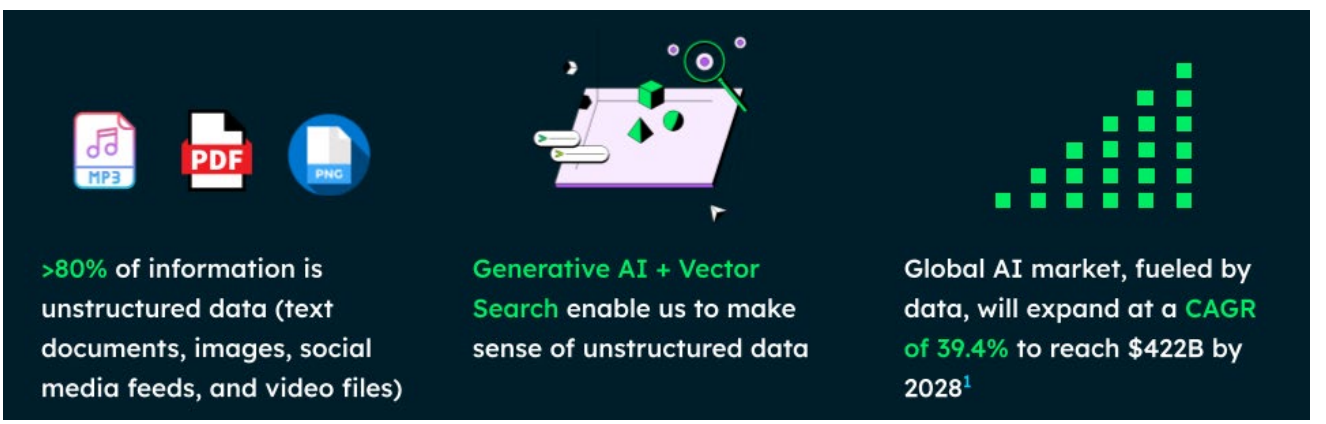

Figure 3. 3 contributing factors in the need for vector search

Enter gen AI and retrieval-augmented generation

The second key use case for vector search is what's known as retrieval-augmented generation or RAG. RAG works by providing the context of proprietary data using vector search to prompt engineer accurate, relevant responses for large language model (LLM) based applications, enabling new generative AI-focused use cases. These vector embeddings represent enterprise data in a way that AI can comprehend, providing context and characteristics that the rows and columns of a relational database could never capture.

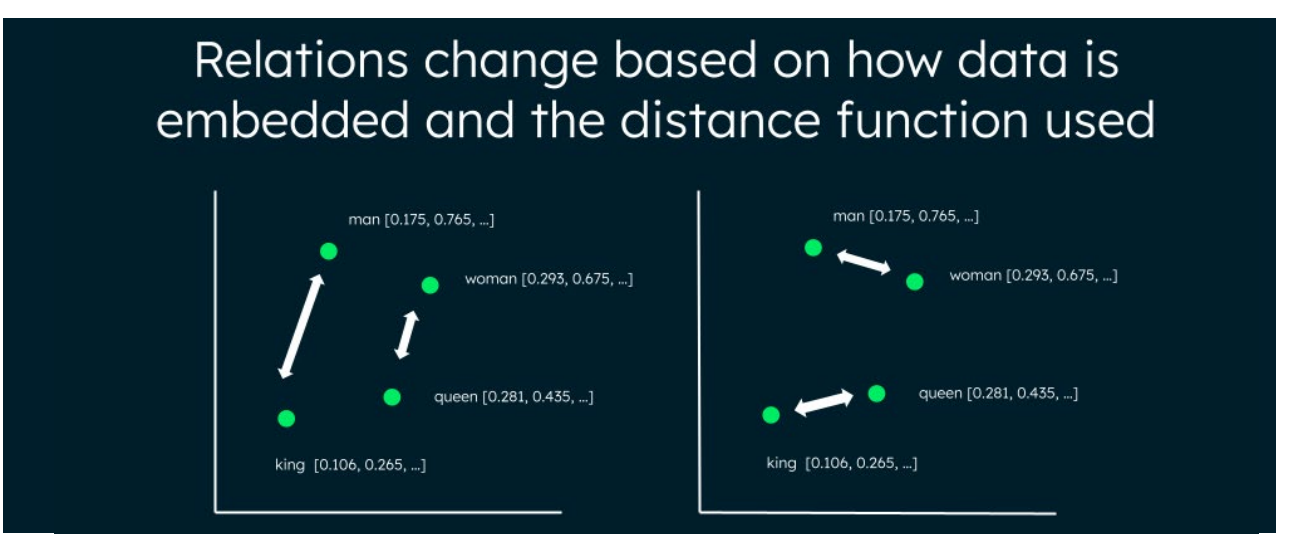

Figure 4. Semantic similarity between similar items

The rise of new AI technologies has created a new paradigm for search (and just about every other business and technology out there), and the stats back it up! According to Bloomberg, the global AI market–fueled by data–will expand at a compound annual growth rate (CAGR) of almost 40% to reach $422 billion by 2028[1]. RAG is the key to unlocking this potential by combining the technological advances in LLMs with business-specific data to better serve customers, increase productivity, and to stay at the cutting edge of technology.

Vector search differs from traditional keyword search in a couple of other key ways. With vector search, you're searching representations of your data based on a model that transforms it into something AI can comprehend. With traditional keyword search, you're matching on the specific presence of a value or field based on a piece of text. Now with vector search, you're taking a larger corpus of data, transforming it into numerical embeddings, and searching across these embeddings (how you store these vectors is also important). An example would be for the query, "interesting places to visit in manhattan," which for lexical or keyword search could return irrelevant results about visiting Manhattan Beach in California or even bars that serve the classic cocktail, rather than the context and association between keywords.

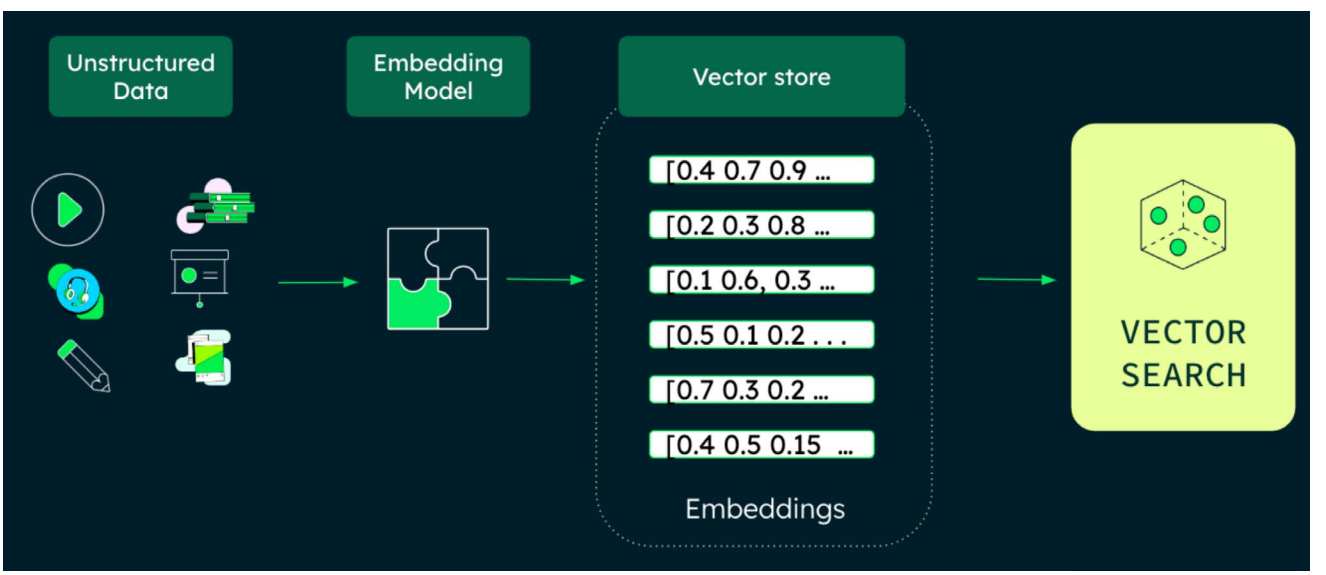

Figure 5. Diagram of vector search

MongoDB's Atlas Vector Search has additional capabilities you won't find in other platforms. First, it's tightly integrated with the Atlas database by default given our unified platform approach. And because MongoDB Atlas is a vector database, organizations can store and search vector embeddings right alongside their operational data, avoiding the need to sync data between your application database and your vector store at both query and write time. It also saves businesses money by eliminating the need for additional technology layers since you can perform semantic search and handle RAG use cases all using the MongoDB Atlas platform. Plus, you don't need to procure a separate bolt-on vector database which leads to fragmented and inefficient developer experiences. When using Atlas Vector Search for gen AI use cases, you can provide additional context to LLM prompts to enhance the output of results, "making them smarter" by augmenting the LLM with relevant contextual information that reduces hallucinations. Lastly, Atlas Vector Search provides a full suite of AI integrations with LLMs and frameworks including LlamaIndex, OpenAI, Hugging Face, LangChain, and more.

[1]Source: Zion Research - Bloomberg business

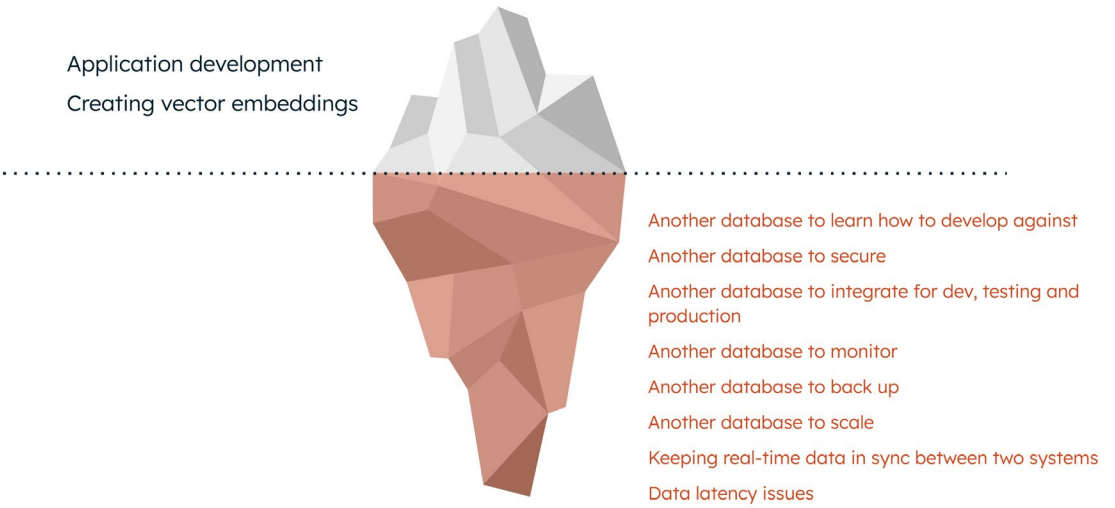

Figure 6. The hidden costs of standalone vector search

Rethinking application search for the enterprise

Utilizing full-text search or vector search provides numerous advantages, allowing businesses to reimagine application search for the enterprise. With the unified MongoDB Atlas platform (providing both full-text search and vector search for semantic and RAG use cases) businesses can remove the possibility of data silos, as Atlas is the only database, vector database, and operational database you need. Unifying all your operational data, metadata, and vector embeddings in Atlas provides your team one system to learn, procure, provision, secure, and scale.

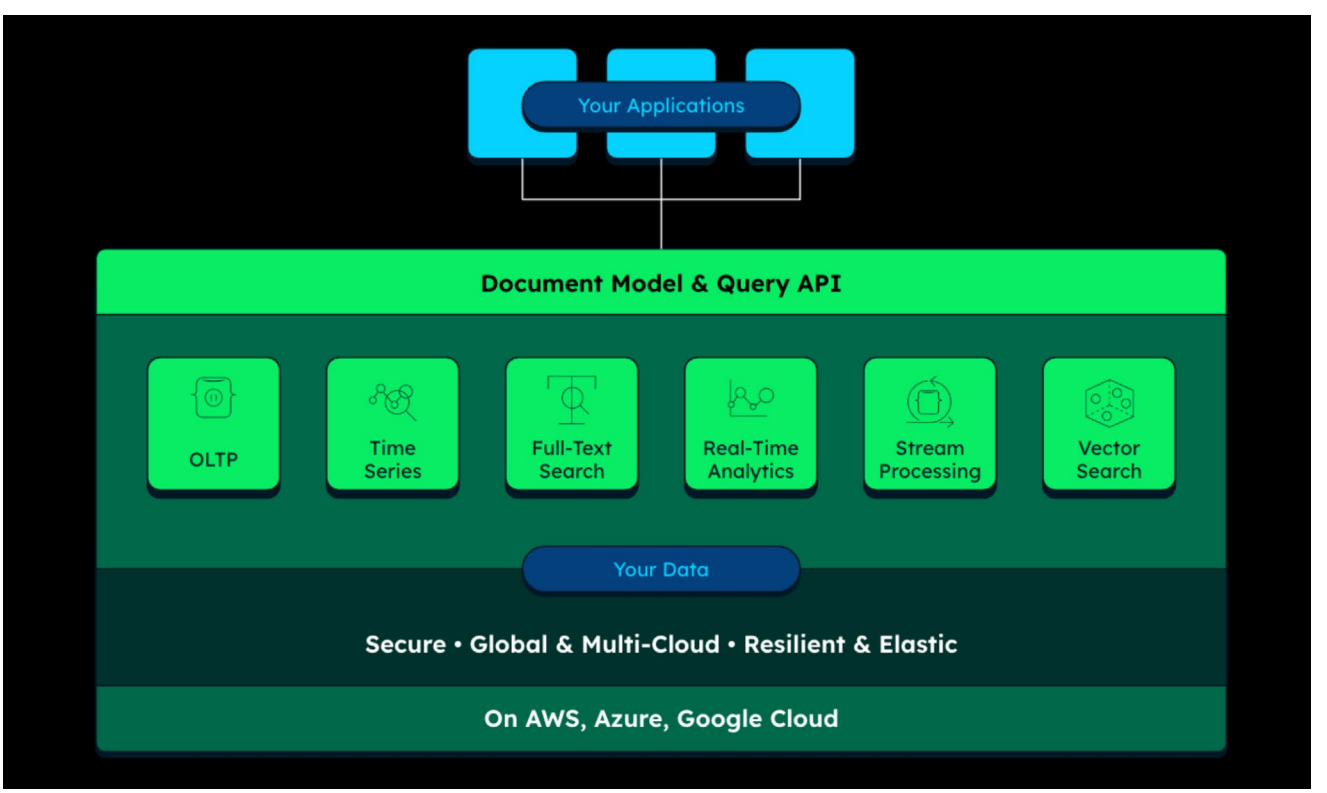

Figure 7. The unified Atlas platform

The additional benefit of the unified Atlas platform is the ability to utilize hybrid search, bringing together the power of full-text and vector search functionality with different types of data, including text, graph, and geospatial. Hybrid search offers the best of both worlds, combining the accuracy of full-text search with semantic and RAG use cases, to ensure the highest relevance and best experience so users find exactly what they're looking for. This hybrid approach is something your developers (and end customers) will love. Plus the additional benefits of avoiding the synchronization tax and potential for data duplication while delivering higher availability, security, and scalability – all with multi-cloud deployability. Hybrid search unifies the best of both search worlds, complete with all the benefits of the unified Atlas platform.

The other critical consideration is infrastructure, and the ability to scale your search and database needs independently, eliminating the risk of possible resource contention and resulting downtime. With Atlas Search Nodes for both search and vector search, you get dedicated compute resources allowing you to scale search and database workloads independent of one another leading to superior performance and higher availability. And best of all it's available on your cloud of choice, whether you're building on AWS, Google Cloud, or Microsoft Azure.

With the introduction of Atlas Search Nodes, we've taken the next step in providing builders with ultimate control, giving them the ability to remain flexible by scaling search workloads without the need to over-provision the database. By isolating your search and database workloads while keeping the search cluster data synchronized with operational data, Atlas Search and Atlas Vector Search eliminate the need to run a separate ETL tool, which takes time and effort to set up and is yet another fail point for your scaling app. In fact, we've seen a 40% to 60% decrease in query time for many complex queries, while eliminating the chances of any resource contention or downtime.

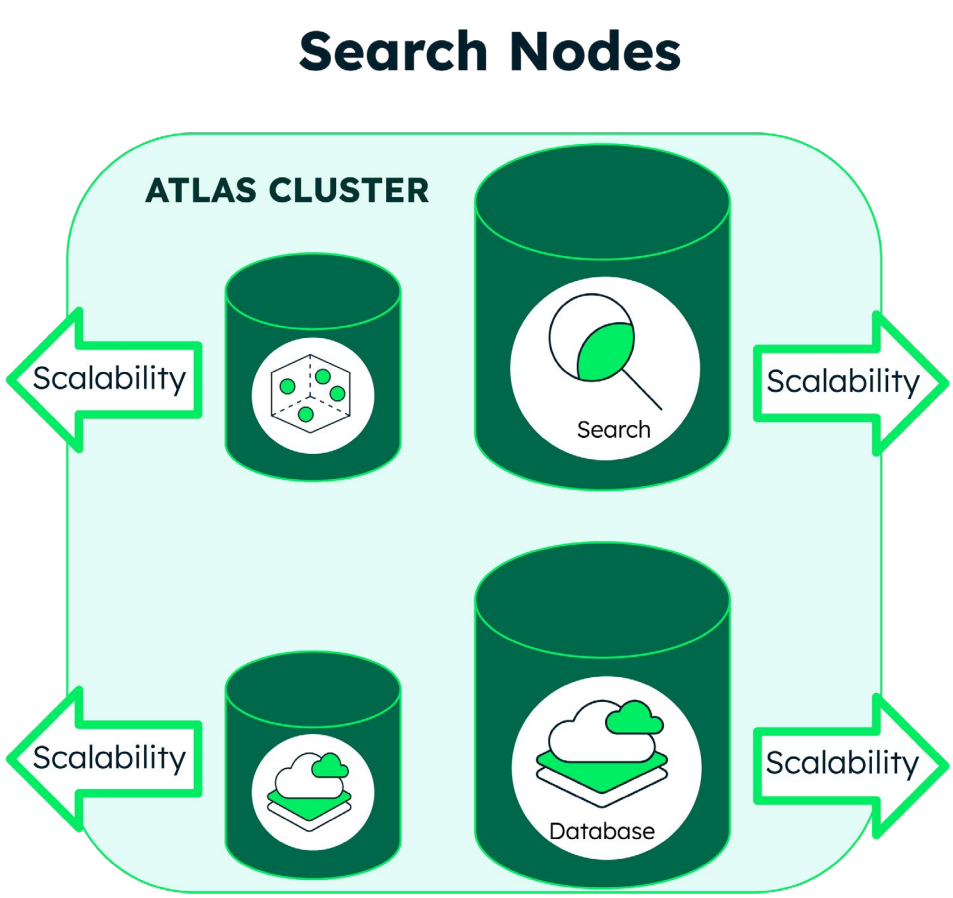

Figure 8. Search Nodes dedicated architecture

The four critical questions to future-proofing your search strategy

Now that you know a little bit more about the evolution of search, including where it's headed, we wanted to provide a checklist of questions to share with your team to ensure you're able to future proof your own corporate strategy. We also offer some MongoDB resources and guidance around future-proofing.

This is not meant as an exhaustive list but rather a catalyst to start the discussion with your team around what's important to your business and how you can use search to increase your competitive advantage, no matter what industry you operate in.

The four critical questions for for future-proofing your search strategy are:

1. **What is your entry point, and how does your developer team get started?**
   - Is there a free experience to try and test without any commitment?
   - Is there a local developer experience, in addition to cloud functionality?
   - Are there programmatic methods to get started?
2. **Where should my data live?**
   - With MongoDB Atlas, we combine operational data with your metadata, as well as vector embeddings for those taking advantage of vector search. Search indexes and vector indexes live in the same platform as your operational data, avoiding data duplication and the chances your data gets stale or out of date.
3. **How can I ensure access to the best search frameworks and integrations?**
   - Atlas search capabilities are built with the established search industry standard of Apache Lucene.
   - Avoid having to build with proprietary search technology and being "locked in."
   - Atlas provides access to a robust set of partners and frameworks, including LangChain, LlamaIndex, OpenAI, Hugging Face and more.
4. **How do I future-proof my tech stack?**
   - Ensure you are building and scaling with one single platform and one unified solution.
   - Ensure your search architecture is optimized for maximum control, flexibility, and improved outcomes with workload isolation (known as Search Nodes in Atlas).
   - Ability to monitor your search analytics and fine turn your results based on search data.
   - Ability to leverage advanced functionality, including hybrid search to unify full-text and vector search.

Learn More

For more information on MongoDB Atlas Search, visit our website or sign up for Atlas Search today.

Author:
Elliott Gluck