Couchbase

TCO BRIEF

Achieve High-Performance AI Applications and Lower TCO With Couchbase

Executive Summary

Modern enterprises are navigating a critical inflection point where the demand for high-performance applications and the transformative power of artificial intelligence (AI) converge. To remain competitive, organizations need a database platform that not only manages mission-critical workloads with speed and reliability but also serves as a flexible, cost-effective data foundation for advanced AI capabilities. Couchbase is revolutionizing this landscape with its performance-driven and economically efficient database platform.

Couchbase delivers an AI-ready unified database platform with high throughput and low latency, distributed scalability, and cost-efficiency. As organizations confront challenges like escalating infrastructure costs, data processing bottlenecks, and the complexities of integrating and managing siloed solutions, Couchbase offers a unique path forward. We will demonstrate how Couchbase significantly reduce total cost of ownership (TCO), while enhancing developer agility, and unlocking next-generation performance for both traditional applications and sophisticated AI workloads.

Introduction

The pace of digital transformation is accelerating, driven by the exponential growth of data and the burgeoning adoption of AI. Businesses across all sectors are under pressure to build smarter and more responsive applications. This requires a data infrastructure that can support large-scale, real-time AI workloads without compromising on performance, reliability, or affordability. This is precisely where Couchbase excels, offering a modern database platform that combines high-performance and flexible workload support with cost-efficient solutions designed for the demands of today’s IT environments.

With its powerful Capella Database-as-a-Service (DBaaS), Couchbase provides the essential tools for developing, deploying, scaling, and evolving sophisticated AI-powered applications. From healthcare and finance to retail and telecommunications, organizations are using Couchbase to power critical applications. Couchbase delivers the speed, efficiency, and scalability necessary to turn data into a true competitive advantage.

The Modern Data Dilemma

As enterprises embrace AI, they encounter a new set of data-related challenges that legacy databases and fragmented data architectures are ill-equipped to handle. These challenges create significant barriers to innovation and growth.

- **The rising cost of high performance:** AI and agentic workloads are resource-intensive, often requiring specialized hardware and massive-scale infrastructure.
- **Slow data processing in mission-critical applications:** End users expect instantaneous responses, whether they are accessing a profile, completing a transaction, or interacting with an AI-powered assistant. Delays caused by slow database queries or inefficient data retrieval directly impact user experience, customer satisfaction, and revenue. Adding AI interactions adds complexity and increases overall response latency.
- **Difficulties in scaling AI workloads:** AI development is a dynamic field. A rigid data infrastructure can stifle innovation by making it difficult to work within your AI ecosystem, manage diverse data types (like vectors for similarity search), and scale workloads. Most AI development to date has driven POCs, but not led to enterprise deployments. Many organizations find themselves locked into specific vendors or platforms, hindering their ability to adapt.
- **Constraints in leveraging LLMs for real-time applications:** The power of large language models (LLMs) is undeniable, but integrating them into real-time applications presents significant challenges. High API costs, network latency, and redundant queries can make LLM-powered features prohibitively expensive and slow.

Couchbase Delivers Performance and Cost-Efficiency by Design

Couchbase addresses these challenges with a unified platform built on a multi-model, memory-first architecture. This approach redefines the relationship between speed, scale, and cost.

HIGH-PERFORMANCE CACHING ARCHITECTURE

Central to Couchbase's success is its integrated in-memory caching layer. This is not an add-on or a separate tier; it is a native component of the database engine. This design ensures that the most frequently accessed data is served directly from RAM, enabling real-time data access at sub-millisecond speeds. By embedding caching as a core function, Couchbase eliminates the complexity and latency associated with bolting on third-party caching systems like Redis, while guaranteeing ultra-low latency for key-value operations.

- **Real-world impact:** LinkedIn, one of the world's largest professional networks, leverages Couchbase for the seamless login experience of its nearly one billion members. The platform relies on rapid key-value lookups from Couchbase's caching technology to provide instant profile access, handling massive concurrent traffic without performance degradation.

SCALABLE INFRASTRUCTURE FOR MISSION-CRITICAL APPLICATIONS

Couchbase was designed from the ground up for distributed environments. Its architecture allows for effortless horizontal scaling across commodity nodes, providing virtually unlimited flexibility for organizations as their data volume and user base grow. Its active-active replication capabilities ensure high availability and disaster recovery, making it the platform of choice for industries like finance and e-commerce that demand always-on reliability.

- **Flexible deployment models:** Organizations are not locked into a single deployment strategy. Couchbase can be deployed across multicloud environments, in on-premises data centers, or at the edge, offering tailored scalability that meets business requirements without sacrificing performance.

Performance and cost are often seen as a trade-off. Couchbase breaks this paradigm. Couchbase consistently outperforms competitors, supporting similar workloads with fewer and smaller resources. This architectural efficiency directly translates to lower operational costs and reduced energy consumption that help offset the additional costs of running AI models.

- **Example:** Rakuten, a global technology leader, reduced its infrastructure costs by 60% after migrating to Couchbase. The company achieved the scalability it needed to support its massive user base but with substantially lower hardware and operational overhead compared to its legacy system.

Enabling Advanced AI Workloads With Couchbase

Couchbase is more than just a high-performance database; it is a comprehensive platform engineered to support the entire lifecycle of AI application development.

BUILD AI APPLICATIONS WITH UNMATCHED FLEXIBILITY

Modern AI applications require a platform that can process and index vast volumes of unstructured and semi-structured data. Couchbase's native support for JSON and the powerful SQL++ query language provide a familiar and flexible interface for developers. This allows businesses to seamlessly integrate AI logic, process vector embeddings for similarity searches, automate complex workflows, and adopt a wide range of cutting-edge AI techniques in a single database platform. There is no need for multiple, siloed technologies. Features like vector search are built directly into Couchbase, allowing for efficient similarity searches that power recommendation engines, semantic search, and retrieval-augmented generation (RAG). Organizations can co-locate AI models in the Couchbase environment using NVIDIA cutting-edge NIM platform, which improves security and lowers latency.

SEAMLESS SUPPORT FOR YOUR LLM STRATEGY

Couchbase empowers organizations to adopt the LLM strategy that best suits their security, cost, and performance needs – whether it's leveraging publicly available models (like those from OpenAI or Google) or deploying proprietary, self-hosted ones.

CACHING FOR AI WORKLOAD OPTIMIZATION

The cost of interacting with LLMs, measured in tokens and API calls, can quickly become prohibitive. Couchbase's caching capabilities are uniquely suited to optimize these interactions. By caching the results of LLM queries and prompts, applications can drastically reduce redundant calls to the LLM service. This not only lowers operational costs but also significantly reduces latency, leading to a more responsive user experience. Semantic caching takes this a step further by identifying and consolidating queries with similar meanings, further minimizing token usage and API expenses.

CUSTOMER SUCCESS: FICO, REAL-TIME FRAUD DETECTION AT GLOBAL SCALE

Challenge: FICO, a leader in predictive analytics, is responsible for protecting over 65% of the world's credit card transactions from fraud. This requires a data platform capable of delivering data for machine learning inference in real time, at a massive scale.

Solution: FICO adopted Couchbase for its high-speed caching and data retrieval with the need to deliver data in under 10 milliseconds. These features are critical for their machine learning-based fraud prediction models, which must analyze transaction patterns and customer profiles in real time.

Outcome: With Couchbase, FICO achieves the sub-millisecond query times necessary for effective real-time fraud detection. The platform's ability to handle workloads with instant data retrieval has significantly improved the performance and efficiency of their AI-driven fraud detection models.

Couchbase's Technical Edge

Couchbase's competitive advantage stems from its unified platform and superior, measurable performance.

- **Unified platform advantages:** Unlike fragmented data strategies that require separate databases for operational, analytical, caching, and search workloads, Couchbase unifies these functions into a single, scalable solution. This consolidation simplifies the data architecture, reduces administrative overhead, and lowers TCO.
- **Performance metrics that make a difference:** Internal testing and independent benchmarks consistently show that Couchbase outperforms competitors with fewer nodes and higher scalability.

CUSTOMER SUCCESS: JINMU, REAL-TIME DATA FOR AI ASSISTANT

- **Challenge:** Jinmu, an IT data consulting firm, needed a document-based NoSQL database for a client's AI Assistant project. The solution required a consistent, highly available key-value store with vector search capabilities, SQL syntax support, and the ability to process time-series data to support an application for virtual meetings.

- **Solution:** Jinmu chose Couchbase to power the client's AI Assistant. The assistant captures and organizes real-time conversations, summarizes discussions, and provides notifications. Couchbase's vector search allows the assistant to quickly retrieve relevant context from past exchanges and supports retrieval-augmented generation (RAG) for accurate, timely responses.
- **Outcome:** With Couchbase, the AI Assistant efficiently manages 70 million documents and handles 8,000 operations per second with 10ms response times. The database supports query acquisition under high concurrency, significantly improving the application's performance and stability for external users.

Measurable Business Outcomes

Adopting Couchbase translates directly to tangible business benefits:

- **Reduced ownership costs:** New customers report saving on average more than 30% on their data architecture after switching to Couchbase
- **Enhanced scalability and agility:** The platform easily adapts to changing business needs without requiring costly hardware overhauls or architectural redesigns.
- **Accelerated AI adoption:** Native vector search, flexible JSON support, and seamless LLM integration enable development teams to build and deploy AI-ready workflows faster.
- **Improved speed and user experience:** Sub-millisecond response times for critical operations drive better end-user experiences and higher customer satisfaction.
- **Simplified infrastructure:** A unified platform minimizes technical complexity, freeing up valuable administrative and development resources to focus on innovation.

Conclusion

Organizations aiming to build next-generation applications and leverage cutting-edge AI must prioritize data platforms that deliver elite performance, operational simplicity, data flexibility, and cost-efficiency. A fragmented, multiple database approach is no longer sustainable in an AI-driven world. With its unified, memory-first architecture, Couchbase empowers businesses to run mission-critical workloads at real-time speeds, scale without limits, and innovate with AI – all while reducing total cost of ownership.

By consolidating operational, vector search, real-time analytics, and AI data services onto a single platform, Couchbase removes the friction and complexity that hinder digital transformation. Businesses can deliver superior customer experiences, accelerate time to market for new features, and build a resilient foundation for future growth.

We invite your team to explore Couchbase's AI-ready platform today. Start with a free trial of Couchbase Capella or learn more by exploring our extensive library of customer success resources and technical documentation.

Couchbase

Modern customer experiences need a flexible database platform that can power applications spanning from cloud to edge and everything in between. Couchbase's mission is to simplify how developers and architects develop, deploy and consume modern applications wherever they are. We have reimagined the database with our fast, flexible and affordable cloud database platform Capella, allowing organizations to quickly build applications that deliver premium experiences to their customers—all with best-in-class price performance. More than 30% of the Fortune 100 trust Couchbase to power their modern applications.

For more information, visit www.couchbase.com and follow us on Twitter.