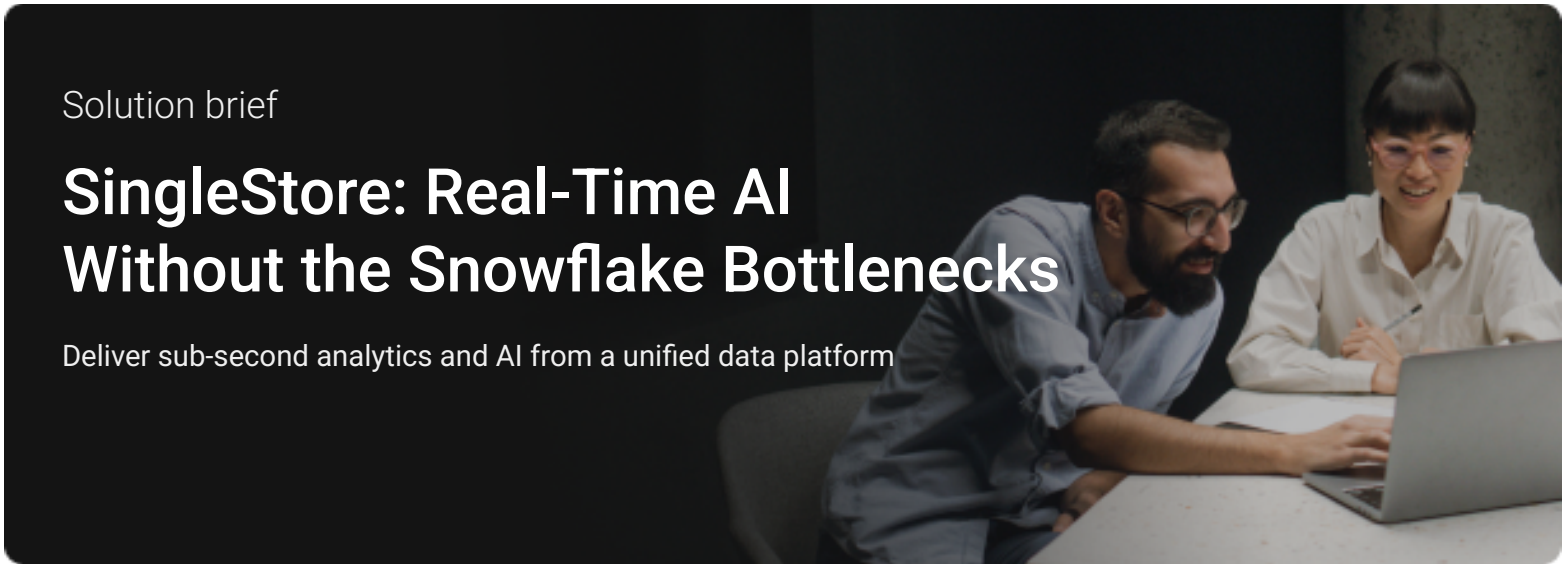

Solution brief

SingleStore: Real-Time AI Without the Snowflake Bottlenecks

Deliver sub-second analytics and AI from a unified data platform

Executive summary

Snowflake transformed cloud analytics, but its architecture was never designed for real-time performance or AI. As enterprises look to power sub-second applications, generative AI and live data experiences, they increasingly face latency, cost and complexity challenges when stretching Snowflake beyond its strengths.

SingleStore offers a millisecond response data engine that complements existing Snowflake environments. It enables technical teams to offload latency-sensitive, always-on workloads — including real-time dashboards, fraud detection and AI inference — while preserving their investment in Snowflake for historical analytics and BI.

This brief outlines where Snowflake's architecture struggles, how SingleStore fits into a modern data stack and the quantifiable benefits of augmenting with a unified real-time platform.

The challenge

Snowflake's architecture is optimized for batch analytics — not real-time speed or operational AI. As technical teams try to stretch it into new territory, they face growing problems:

1. **Latency that breaks sub-minute use cases.** Sub-second dashboards, fraud detection and AI-powered apps suffer. Snowflake's decoupled architecture delays access to fresh data and lacks true transactional support.
2. **AI features that fall short.** While marketed as AI-ready, Snowflake lacks mature vector search, hybrid query capabilities and real-time model serving critical for RAG, recommendations and inference.
3. **Unpredictable and high costs.** Always-on, low-latency applications rack up compute usage. Even mid-sized businesses report Snowflake bills over $30k/month due to opaque pricing and warehouse sprawl.
4. **Stack complexity and engineering overhead.** To compensate, teams bolt on Postgres, Redis, Kafka and Elasticsearch — creating fragile, high-maintenance pipelines and increasing total cost of ownership (TCO).

SingleStore: Real-time AI and analytics on a single engine

SingleStore complements Snowflake by adding a millisecond-level engine purpose-built for SLA driven workloads and AI applications.

1. **Real-time ingestion and querying.** Ingest millions of events per second and serve results in milliseconds — perfect for fraud detection, live dashboards and AI agents.
2. **Built-in AI/ML features.** Serve ML models, power RAG workflows and combine vector, structured + full-text queries in a single call.
3. **Flexible deployment.** Run SingleStore in the cloud (Helios®), your VPC (BYOC) or on-prem. Compatible with Apache Iceberg for smooth integration.

Key benefits

- Sub-second latency for SLA-driven queries
- 50 – 60% lower TCO than multi-database Snowflake + MySQL
- High concurrency to support thousands of users and apps
- Streamlined AI architecture with real-time ingestion + hybrid search
- Predictable costs vs. Snowflake's variable consumption model

How it works/technical differentiators

- **HTAP architecture.** Native OLTP + OLAP on the same data
- **Pipelines.** Real-time ingestion from Kafka, S3 and more
- **SingleStore Kai™ API.** MongoDB-compatible, JSON-native, developer-friendly
- **Advanced vector search.** ANN indexing, hybrid queries, low-latency serving
- **Iceberg support.** Seamless data exchange with Snowflake

Impact

Metric	Outcome
Query performance	Millisecond latency for real-time dashboards vs. seconds+ in Snowflake
TCO reduction	Up to 50% savings by consolidating Snowflake + MySQL
Time to insight	Real-time analytics without ETL delays
Developer efficiency	Fewer pipelines to maintain, faster app delivery

Side-by-side comparison

Feature	SingleStore	Snowflake
Architecture	Unified OLTP+OLAP	Decoupled OLAP
Query latency	Milliseconds	Seconds to minutes
Vector search	Mature, hybrid	Basic similarity
Concurrency	High across workloads	High (only for analytics)
ETL needs	None (real-time ops+analytics)	High (batch movement)
AI use cases	RAG, recs, inference	Model training only
Cost model	Predictable tiers	Variable consumption
TCO (real-time stack)	50 – 60% lower	Expensive + fragmented

Example architecture: coexistence with Snowflake

SingleStore powers real-time ops, AI inference and low-latency dashboards. Snowflake remains the system of record for batch analytics and reporting.

SingleStore as Analytical Ingestion Layer

SingleStore as Analytical Application Layer

Ready to modernize your data architecture?

Discover how SingleStore empowers your team to accelerate AI initiatives, streamline your data architecture and improve performance and cost efficiency — without disrupting your existing Snowflake environment.

- Talk to our team
- Book a demo
- Start a free trial