mariaDB

Datasheet

MariaDB GridGain

Distributed In-Memory Caching and Compute for Real-Time Data Processing and AI Agentic Apps

Extends the capabilities of the MariaDB Enterprise Platform with a native distributed caching and compute layer between your applications and systems of record. Hot data and read-intensive workloads move into a shared-nothing, horizontally scalable memory tier, delivering sub-millisecond response times, ANSI SQL, ACID guarantees and colocated compute. Built for AI/ML and agentic applications

- **Distributed vector search for retrieval-augmented generation (RAG)**
- **Semantic caching and agent memory**
- **Real-time Machine Learning (ML) model execution for decision-making and in-transaction analytics**
- **Fast feature serving**

Architecture

Deploys as a peer-to-peer cluster of server nodes that partition data across cluster RAM, with configurable replicas. The shared-nothing design eliminates single points of failure. Adding nodes automatically rebalances data and scales throughput and capacity linearly, to millions of transactions per second (TPS) and petabytes of data.

Integration and persistence

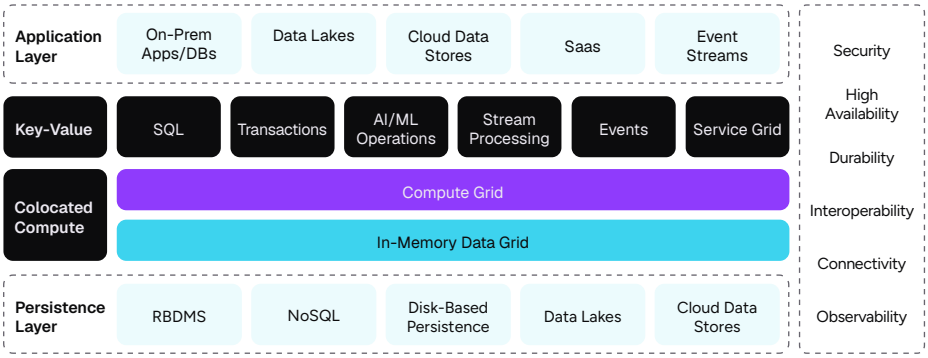
Applications connect through SQL (JDBC/ODBC) or native APIs. Read-through, write-through and write-behind caching keep data consistent with MariaDB or any other system of record. Native persistence stores the full dataset on disk and caches as much as RAM allows, so datasets can exceed memory and each node scales on disk. After a restart, the cluster serves data immediately from its own copy, with no reload from the source.

Colocated compute

Compute runs where the data lives. Colocated processing, continuous queries and streaming ingest eliminate network round trips, so complex calculations run against live data in milliseconds rather than in overnight batch jobs.

Where Customers Deploy It

- **Accelerating application performance**: Keep hot data in memory in front of the database so applications stay fast during traffic spikes.
- **Application modernization**: Speed up an existing application with an in-memory layer and no code rewrite, then migrate off legacy or proprietary databases such as Oracle and DB2 at your own pace.
- **Digital integration hub**: Consolidate and curate data from disparate systems of record into a single, real-time hub that joins across systems and delivers answers in milliseconds.
- **Mainframe offload**: Move high-volume reads and compute off the mainframe to cut MIPS and serve every digital channel, while the core remains the system of record.
- **Real-time decisioning**: Run business rules and ML models against live event streams to drive risk, fraud and pricing decisions at the moment of interaction.
- **Agent memory**: Hold state, checkpoints and cross-session facts for AI agents, so a run resumes after a restart and the next conversation starts informed.



Key Benefits

- **Ultra-low latency** at massive scale, with sub-millisecond data access
- **Distributed in-memory architecture** delivers massively parallel processing and linear scale-out to millions of TPS
- **Colocated compute** that runs business logic where the data resides, reducing network data movement and disk I/O
- **Stream processing and transactional analytics** make decisions within the transaction window
- **Data integrity and consistency** through ACID compliance, read/write-through and disk-based persistence
- **Distributed SQL with full JOIN and aggregations** to answer questions in memory
- **Continuous availability** across racks, data centers and clouds
- **Lower infrastructure spend** by offloading non-core database operations and reducing mainframe MIPS
- **One vendor and one support contract** for the entire data delivery stack

Proven in Production

Trusted by top 10 leaders
in banking, telcos and transportation

300x
faster risk calculations, 3 minutes to 600ms at a global Top 5 bank

99.999%
availability sustained for 5+ years at a large telco solutions provider

200K
transactions per second at a national credit registry

> "We are colocated with the data, so we can take advantage of zero network latency. We are now able to post transactions and do something with them in as little as 20 milliseconds."
>
> — Large U.S. regional bank

RBC · verizon · UBS · HPE · BNP PARIBAS · American Airlines · State Street

Expanding the Capabilities of the MariaDB Enterprise Platform

MariaDB GridGain brings distributed in-memory caching and computing natively into the MariaDB environment — not as a bolt-on, but as a deep integration built for mission-critical data.

- **Sub-millisecond latency:** Move hot data from disk to RAM and turn minute-long queries into millisecond responses.
- **Frictionless implementation:** Deploy as a transparent query cache in front of MariaDB with zero application code changes.
- **Automatic data syncing:** For read/write through/behind caching patterns,
  native MariaDB cachestore integration eliminates the need to maintain custom app invalidation logic.
- **Architectural consolidation:** Replace third-party caching layers and fragile batch and ETL pipelines with one unified platform.
- **Geo-distributed performance:** Extend MariaDB's reach with ultra-low latency across regions and data centers.
- **Elastic scalability:** Handle millions of requests by adding nodes, scaling performance as data and concurrent user traffic grow..
- **Single operations view:** Monitor MariaDB databases and GridGain clusters with Enterprise Manager.
- **Unified enterprise support:** One partner, one contract and one expert team for both the database and the cache.

MariaDB GridGain Features

Feature	What You Get
Distributed in-memory architecture	Shared-nothing partitioned in-memory storage delivers sub-millisecond data access and horizontal scalability. NUMA-aware allocation places data in the RAM closest to the processing core.
Persistence	Native durable on-disk storage with fully transactional writes enabled by a write-head log. Store datasets larger than available RAM to scale efficiently without risking memory overflow. Query data whether it is in memory or on disk. After a restart the cluster recovers from its own copy, with no reload from the source.
SQL & transactions	ANSI SQL with DDL and DML, distributed joins, cross-table queries, indexes and aggregations. Fully distributed ACID transactions with user-configurable strict or eventual consistency.
Data access APIs	SQL, key-value (JCache JSR-107), full-text, vector and continuous queries. Native APIs for Java, .NET/C#, C++ and Python, plus JDBC, ODBC and REST.
Compute & analytics	Colocated compute executes logic on the nodes that hold the data, eliminating network data movement and enabling massively parallel processing. Compute grids, service grids and machine learning libraries run in the cluster.
AI/ML	Enable agent memory for durable storage and semantic cache to reduce costs and latency. Feature extraction and ML model execution run in-transaction against live enterprise data. LangChain and LangGraph integrations and native MCP Server.
Vector search	Distributed and tunable vector search for similarity and RAG retrieval.
Streaming & ingest	Confluent-certified connector for Apache Kafka, native Apache Spark integration for RDDs and DataFrames and high-volume data streamers for event-driven data ingestion and processing.
Data lake formats & JSON	SQL COPY imports and exports Apache Parquet, Apache Iceberg and CSV. Dedicated JSON functions accelerate semi-structured and document workloads.
System-of-record integration	Read-through, write-through and write-behind caching with native MariaDB cachestore integration and JDBC dialect adapters for Oracle, SQL Server, MySQL and DB2. Automatic schema import and data loading with no coding required.
High availability & disaster recovery	Configurable backup copies, rack awareness, active-active and active-passive data center replication and zero-downtime rolling upgrades across the cluster.
Backup & recovery	Local, remote, full and incremental snapshots with point-in-time recovery, plus compressed snapshots for persistent clusters.
Security	Authentication and role-based authorization, TLS/SSL transport encryption, transparent data encryption at rest, audit logging and multi-tenancy hardening. Manage user access in-cluster or integrate with external identity systems including LDAP, Active Directory, and OpenID Connect (OIDC).
Observability & management	MariaDB Enterprise Manager provides a unified dashboard across the entire MariaDB Enterprise Platform. MariaDB GridGain Control Center enables in-depth troubleshooting and cluster administration, customizable dashboards, alerting, query tracing and tuning. OpenTelemetry-compatible metrics integrate with existing observability stacks. MariaDB GridGain Kubernetes Operator deploys, scales, upgrades and recovers MariaDB GridGain clusters.
Flexible deployment	Bare metal, VM or Docker; Kubernetes and OpenShift via the certified operator; AWS, Azure and Google Cloud; hybrid, multicloud and inter-cloud topologies.

Start a free trial

Book a demo

© 2026 MariaDB. All rights reserved.

mariadb.com