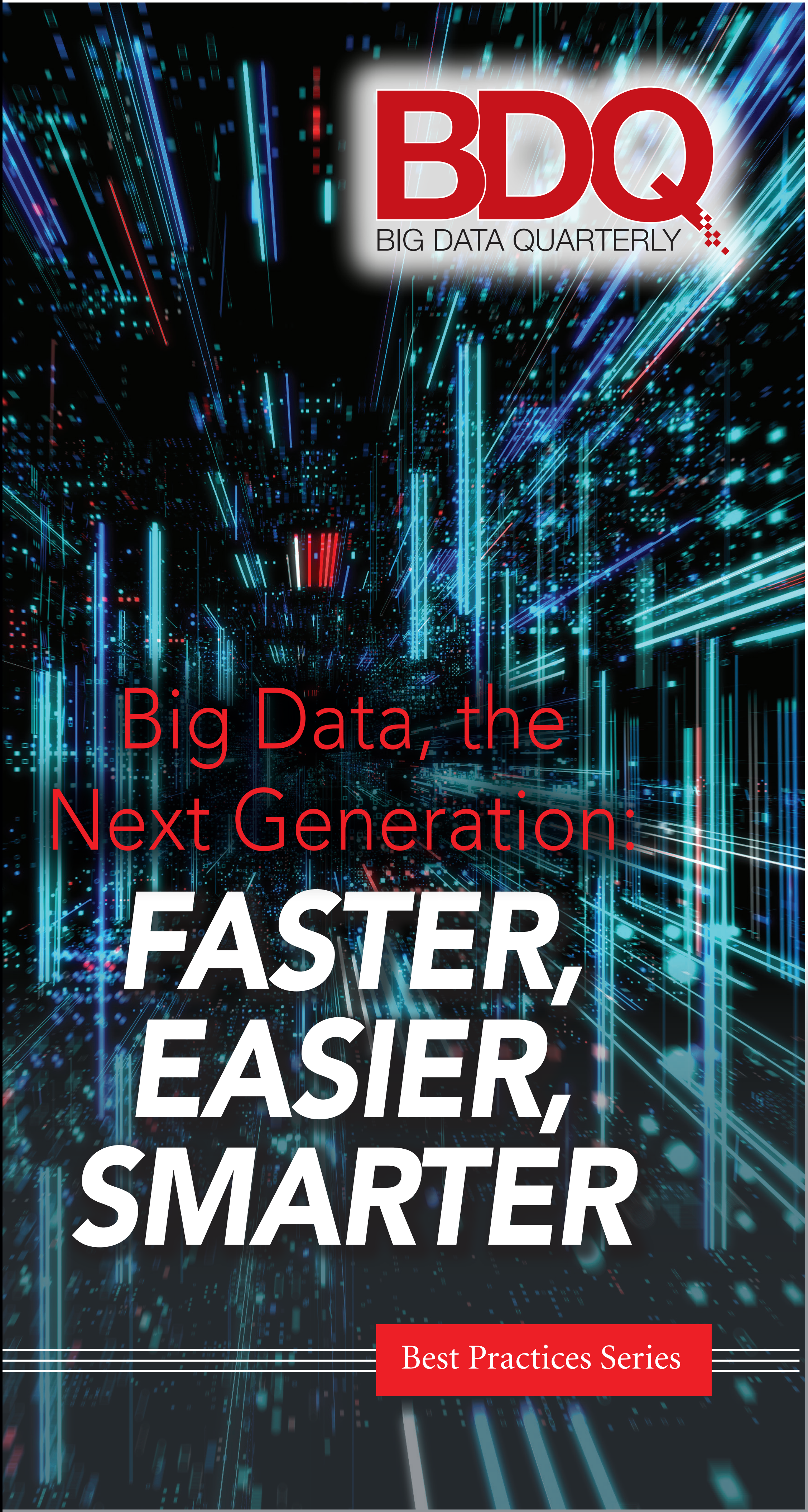


BDQ
BIG DATA QUARTERLY

Big Data, the Next Generation: FASTER, EASIER, SMARTER

Best Practices Series

BDQ
BIG DATA QUARTERLY

THE NEXT GENERATION OF BIG DATA

Best Practices Series

“Big data” in the form we know it, pertaining to volume, variety, and velocity, has been top of mind at enterprises for close to a decade now. Capturing, deploying, and extracting value from it typically required a cadre of data specialists, scientists, analysts, and professionals.

These people will all be essential for moving forward into the next phase. However, managing and leveraging data is getting a whole lot easier. And this is not coming a moment too soon: Organizations are being inundated with data from the Internet of Things on the outside and artificial intelligence (AI) and machine learning on the inside.

Until recently, data analytics has been delivered through data warehouse and Hadoop-based stores. Typically, both environments have required investments in skills and hardware to support sizable datasets. In recent times, new technology platforms and architectures have emerged that are reshaping—and indeed dismantling—the very concept of big data.

Here are some of the factors leading to the next generation of big data:

> The rise of open source platforms for data analytics—Apache Spark, Kudu, and Impala—provides cost-effective and powerful ways to deliver data and insights in real time to decision makers and systems.

NEXT-GENERATION PLATFORMS HAVE OPEN SOURCE FOUNDATIONS.

The rise of open source platforms for data analytics—Apache Spark, Kudu, and Impala—provides cost-effective and powerful ways to deliver data and insights in real time to decision makers and systems. These platforms aren’t necessarily replacing enterprise data warehouses and business intelligence environments—but are modernizing these solutions.

As AI becomes a more pervasive part of enterprise processes and systems, machine learning means algorithms and programs adjust and are constantly refreshed with an influx of data.

NEXT-GENERATION DATA COMES FROM AND RESIDES IN MANY PLACES ALONG THE IoT CONTINUUM.

"Fog computing," for example, is coming to the fore as a solution of choice for many enterprises seeking more distributed options. Fog computing—as defined by the National Institute of Standards and Technology—"facilitates the deployment of distributed, latency-aware applications and services, and consists of fog nodes (physical or virtual), residing between smart end-devices and centralized cloud services."

NEXT-GENERATION DATA RESIDES IN CLOUDS, DATA CENTERS, BEHIND APIs—AND ALL OF THE ABOVE.

The proliferation of cloud services and APIs has opened up new vistas for big data with almost unlimited capacity. Cloud-based services take much of the burden off enterprise shops in terms of managing data. Data may be provided through data as a service environments, in which any and all data, regardless of where it resides, can be surfaced or virtualized to be available to standardized service layers of applications and systems.

NEXT-GENERATION DATA TAKES ADVANTAGE OF STORAGE ANYWHERE, ANYTIME—WITH ASSURED BACKUP AND RECOVERY.

Storage is no longer confined to disk arrays stacked within data centers; it is available on demand, anytime, anywhere, thanks to cloud computing. There is no longer a ceiling on data capacity within on-premise data centers, no longer a need to employ space-saving techniques such as compression or lifecycle management progresses from disk to tape. In addition, data is resilient and recoverable on-demand in less than a second.

NEXT-GENERATION DATA FOSTERS SELF-LEARNING MACHINES.

As AI becomes a more pervasive part of enterprise processes and systems, machine learning means algorithms and programs adjust and are constantly refreshed with an influx of data. To some degree, humans no longer need to be constantly rewriting or reprogramming systems to meet new demands; the data is steering things in new directions.

NEXT-GENERATION DATA TURNS BIG DATA INTO REALLY BIG DATA.

Big data—based on internal corporate data such as transactions and customer records—has gotten even bigger than previously imaginable, expanding wildly in a geometric sense. Data is now flowing in from sources large and small (as small as miniature cameras) and flowing through enterprises.

NEXT-GENERATION DATA FLOWS THROUGH ENTERPRISES IN REAL TIME.

An emerging generation of tools, platforms, and methodologies has driven down the costs of real-time data analytics. Such data may be locked away in ERP systems, as well as externally within partner systems or the Internet of Things. Cognitive tools and technologies such as AI and machine learning are opening up avenues of discovery that are available across real-time data.

NEXT-GENERATION DATA TAKES MANY FORMS.

When the big data revolution kicked into high gear about a decade ago, it was limited to relational data and NoSQL data. Now, data architectures have opened up to imagery and a variety of content types.

NEXT-GENERATION DATA IS NOT TIED TO ANY VENDOR, NOR DATABASE, FOR THAT MATTER.

In recent years, NoSQL databases have grown in popularity, providing cost-effective and cloud-friendly environments for a variety of data types. In addition, the rise of data lakes also detaches data from particular databases, enabling data to be maintained and available for future applications.

—Joe McKendrick

AEROSPIKE

Maximize the Value of Your Operational Data

Modernize and Transform Your Enterprise Via Real-time Transaction and Analysis Processing

OVERVIEW

It might sound too good to be true: a database system that processes large volumes of operational data in real time while delivering exceptional runtime performance, high availability, and cost efficiency while still keeping your data safe. What if early adopters in banking, telecommunications, and other industries are already harnessing such a database for achieving results that are transforming their businesses in myriad ways? What if published benchmarks demonstrate sub-millisecond response times for high throughput read/write workloads over high data volumes with substantial cost savings compared with traditional alternatives?

This paper introduces key technologies that Aerospike clients are using to modernize their data management infrastructures and realize such impressive (and seemingly impossible) results as:

- Rapid read/write speeds without extensive tuning or a separate data cache
- Substantially smaller footprints than popular alternatives, often leading to 3-year total cost of ownership (TCO) savings of $3-5 million per application
- 24x7 availability, including cross-datacenter replication
- Operational ease during scale-out and maintenance
- Interoperation with popular software offerings, including Apache Hadoop, Spark, and Kafka

Sounds unbelievable, right?

THE TECHNOLOGY IN BRIEF

Aerospike provides a distributed, highly scalable database management system for demanding read/write workloads involving operational data. It was designed to deliver extremely fast—and predictable—response times for accessing data sets that span billions of records in databases of 10s – 100s TB. Other design features address fault tolerance and near 100% uptime even during upgrades and maintenance.

How? By capitalizing on proven architectural approaches—such as distributed computing and parallelism—and developing new technologies to meet business demands that hadn't even surfaced when older systems were originally built. Indeed, Aerospike's patented Hybrid Memory Architecture™ (HMA) drastically reduces traditional I/O and network communication compared with other approaches; it also uses CPU resources considerably more efficiently. The cumulative impact of these features (and others) enables Aerospike to deliver remarkable speed at scale.

APPLICATIONS AND USE CASES

Applications that benefit from Aerospike typically share some or all of these characteristics:

- Service-level agreements (SLAs) that require sub-millisecond database response times
- High throughput for mixed workloads (e.g., 3–5 million operations per second)
- Support for managing billions of business records in databases of 10s–100s TB
- High availability and fault tolerance for mission-critical applications
- High scalability for handling unpredictable increases in data volumes and transactions
- Adaptable infrastructure for managing varying types of data with minimal effort
- Low total cost of ownership (TCO)

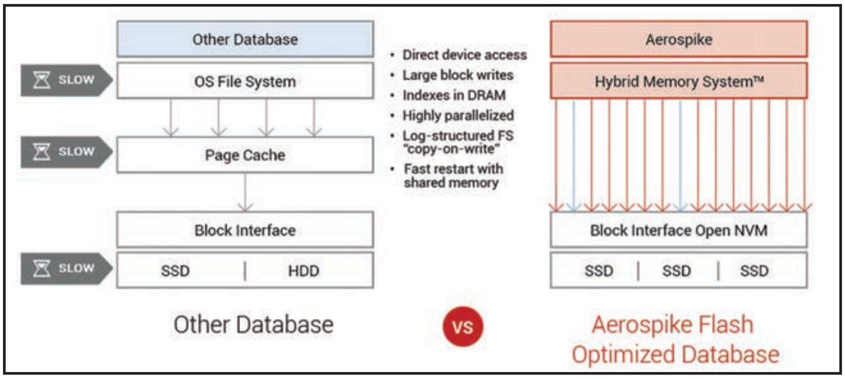
KEY FEATURES AND TECHNOLOGIES

Aerospike is a shared-nothing database system that operates on a cluster of commodity server nodes:

- It's a schema-free, key-value data store.
- Aerospike exploits volatile and non-volatile memory in a distinctive way, providing rapid access to index and user data.
- An intelligent client layer minimizes costly network "hops" needed to access data.
- Immediate record-level consistency and high availability are guiding principles.
- Access management controls and transport encryption protect sensitive data.
- Asynchronous replication across data centers provides disaster recovery.
- Ready-made connectors, a publish/subscribe messaging system, and partner offerings help firms integrate Aerospike into their existing IT infrastructures.

FULL REPORT

To get a copy of the full report, please go to: www.aerospike.com/maximize-operational-data



MariaDB®

The End of Single-Purpose Big Data Platforms

The rise of big data led to a generation of platforms engineered for a single purpose, scaling on commodity hardware.

Apache Hadoop can scale to petabytes of data, but it was engineered for offline analytics. It can't meet the performance requirements of data-driven organizations leveraging ad hoc, interactive queries on near real-time data at scale, and with near real-time latency, to drive faster time to insight. Further, big data is no longer limited to analytical workloads, and because it was engineered for offline analytics, Hadoop can't meet the performance requirements of transactional workloads (regardless of the scale). NoSQL databases can, but they're optimized for point and range queries on small to medium working sets cached in memory, not aggregate queries on most, if not all, data stored on disk. NoSQL databases can't meet the scalability requirements of near real-time analytical workloads in a cost-effective manner—and without schemas and transactions, can't be the system of record for business-critical, mission-critical applications.

The only solution was to deploy a combination of Hadoop, NoSQL databases and relational databases in order to meet *both* performance and scalability requirements, and support *both* transactional and near real-time analytical workloads. It was a complex solution whereby data had to be synchronized between different environments and teams via batch imports, messaging systems or both: Hadoop for offline analytics, NoSQL databases for caching and relational databases for transactions. Further, the lack of standard SQL in Hadoop and NoSQL databases resulted in performance issues with traditional BI and reporting tools.

What if a relational database could scale out *and* support both transactional and near real-time analytical workloads? With MariaDB leading the way, the next generation of big data will be handled by relational databases with scalable, workload-optimized storage engines (row-based for transactional, columnar for analytical), replacing single-purpose big data platforms with hybrid databases capable of scaling both transactional and analytical workloads from a single gigabyte of data to hundreds of terabytes of data—all without sacrificing schemas, transactions or SQL.

MariaDB leverages multiple, purpose-built storage engines to support both transactional and analytical workloads at scale. InnoDB, MariaDB's default general-purpose storage engine, supports transactional workloads up to several terabytes of data. The Spider storage engine extends MariaDB with built-in, transparent sharding to support transactional workloads requiring read, write and storage scalability. And MyRocks, a write- and space-optimized storage engine developed by Facebook and supported by MariaDB, can be used with Spider for unrivaled write scalability and storage efficiency. The final storage engine, MariaDB ColumnStore, extends MariaDB with distributed, columnar data and parallel query processing to support near real-time analytical workloads on hundreds of terabytes of data.

MariaDB is available in two configurations, MariaDB TX and MariaDB AX, both with the world's most advanced database proxy, MariaDB MaxScale. MariaDB TX includes InnoDB, MyRocks and Spider, and is optimized for transactional workloads at any scale. MariaDB AX includes InnoDB and ColumnStore, and it is optimized for scalable, high-performance analytical workloads. It's the same database with the same clients, SQL parser and optimizer, but under the covers, different storage engines support different workloads—and with streaming change-data-capture enabled, the same data too.

In a hybrid transactional/analytical topology, MariaDB TX nodes automatically and continuously stream data to MariaDB AX nodes, enabling analytics on near real-time transactional data without the need for separate databases (often from different vendors), import delays or complex ETL processes. The data is stored in a row-based format in MariaDB TX for high-performance transactional queries, with the same data (or a subset of it) stored in a columnar format in MariaDB AX for high-performance analytical queries.

It's the beginning of the end for single-purpose big data platforms for anything less than a petabyte of structured/semi-structured data. MariaDB is the leading enterprise open source database solution, and the only one delivering the best of both worlds in both dimensions: performance and scalability, transactional and analytical.

MariaDB TX
https://mariadb.com/products/solutions/oltp-database-tx

MariaDB AX
https://mariadb.com/products/solutions/olap-database-ax