ESG Technical Review

Analyzing the Performance of MapR-DB, a NoSQL Database in the MapR Converged Data Platform

Date: October 2017 **Author:** Mike Leone, Senior Validation Analyst; and Domenic Amato, Associate Validation Analyst

Abstract

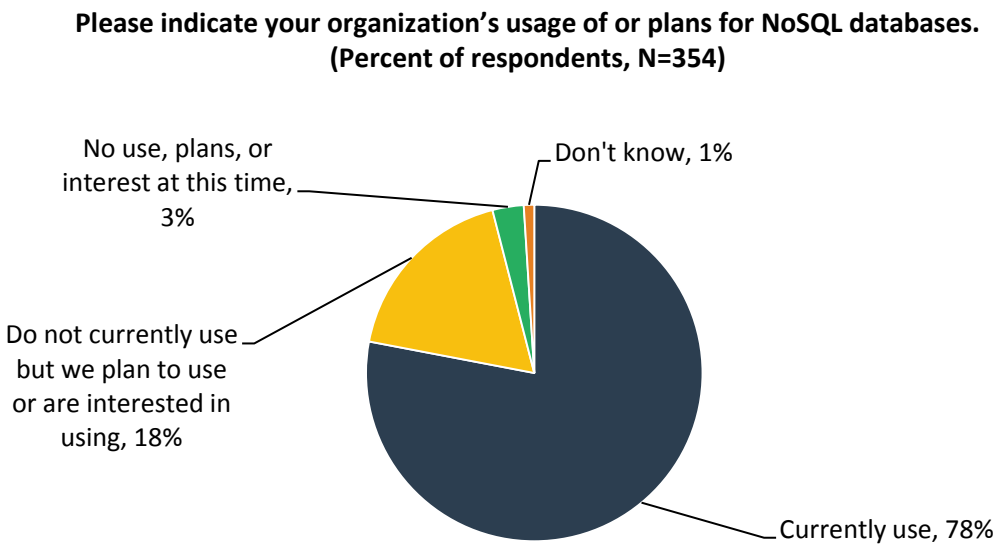
This ESG Technical Review documents and analyzes MapR-DB performance test results. Testing evaluated the performance and scalability of MapR-DB running in the cloud, and we compare results with other leading NoSQL database offerings.

The Challenges

Organizations must be able to scale properly to handle performance needs and expectations as data continues to grow at exponential rates. Traditional relational databases are ideal for handling more structured, predictably sized data sets, but are not equipped to cost-effectively handle the performance and scale that big data requires. As such, organizations are turning to solutions that are more capable of scaling appropriately to handle big data and their underlying workloads, such as data processing and analytics.

NoSQL databases can benefit from modern, distributed architectures to gain the cost-effective scalability that traditional databases are unable to deliver. Because of the performance and simplified structure NoSQL provides, it has become an attractive offering to IT professionals. In fact, ESG research shows that 78% of organizations currently use NoSQL databases, and an additional 18% plan to in the future (see Figure 1).[1]

Figure 1. Use of NoSQL Databases



Source: Enterprise Strategy Group, 2017

While NoSQL databases do offer advantages in performance and cost savings, many of them are still evolving as reliable database options. Some solutions require increased administrative attention and maintenance, resulting in longer downtimes and network bottlenecks, while data inconsistency and data loss during replication, and lack of granular access controls are some of the serious challenges organizations may face while leveraging NoSQL. MapR and its NoSQL database—MapR-DB—are addressing these challenges within MapR's Converged Data Platform.

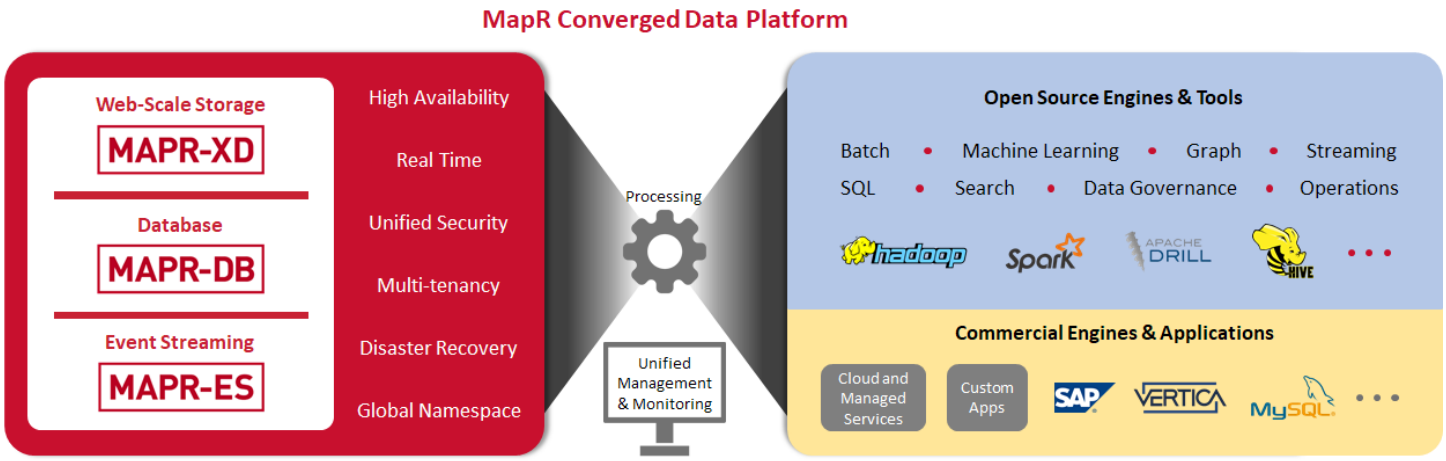
[1] Source: ESG Brief, *Market Disruption: Next-generation Databases*, April 2017.

MapR Converged Data Platform

The MapR Converged Data Platform is a collection of open source engines, tools, and applications that leverage purpose-built MapR platform data to deliver a scalable, reliable, and secure infrastructure for global data-driven applications. The platform integrates Hadoop, Spark, and Apache Drill with real-time database capabilities, global event streaming, and scalable enterprise storage. The distributed MapR architecture offers high levels of availability and reliability with no single points of failure, while working in conjunction with the comprehensive projects in the Hadoop ecosystem to deliver enterprise-grade features and functions on low-cost commodity hardware.

Three core services work together to enable MapR to truly be a converged platform that supports all workloads on a single cluster. MapR-XD serves as the cloud scale data store for the Converged Data Platform, providing distributed, enterprise-grade storage that is high performing, scalable, and reliable. MapR-DB is a global, high-performance distributed NoSQL database for next-generation applications and analytics. Lastly, MapR-ES Event Streaming is a publish-subscribe event streaming system that enables real-time processing of globally produced data in a reliable, scalable fashion.

Figure 2. The MapR Converged Data Platform



Source: Enterprise Strategy Group, 2017

MapR-DB

MapR-DB is an enterprise-grade NoSQL database capable of handling extreme scalability to provide real-time operations and analytics without the extra complexity and cost of a traditional database architecture. Additional benefits include:

- Multi-model flexibility – MapR-DB provides key-value, wide column, and JSON document models, enabling more complex applications that need multiple data models. It can easily handle existing HBase applications without the need for Java virtual machines, and manages a slew of operational data formats allowing for a faster, more flexible development cycle.
- Real-time, event-driven data – MapR-DB harnesses a converged platform approach to NoSQL capable of real-time analytics in addition to other database workloads at enterprise-level size and speed. Platform and filesystem optimizations allow for extreme scalability on thousands of nodes per cluster and trillions of records per table.
- Mission-critical reliability and performance – MapR-DB minimizes downtime and eliminates points of failure using a strong consistency model and replication. It performed consistently during throughput benchmark tests and its unique data structure innovations ensure consistently low latency during various testing scenarios. MapR-DB's optimization is done automatically without the need for additional code.

Performance Analysis

ESG audited and analyzed tests run by MapR to validate the performance capabilities of MapR-DB running in the cloud compared with two other NoSQL databases: Cassandra and HBase. The goal of testing was to first create a common hardware configuration, and then make specific configuration changes to each of the databases to ensure an apples-to-apples comparison.

For hardware, a cloud infrastructure was created in AWS. An eight node EC2 cluster using Amazon's i3.8xlarge instance type was configured in its own virtual private cloud within a single placement group. Eight client nodes (one per physical node) were configured to use enhanced networking adapters (ENAs). Each client ran the Linux 4.11 Linux kernel and CentOS 7.2. Evenly distributed across the cluster was a 2,000,000,000-record database with ten fields of 100 bytes each (1,000 record size) that consumed a total of 2 TB (see Figure 3).

Figure 3. Test Bed

Source: Enterprise Strategy Group, 2017

A Cassandra 3.11 cluster running on a dedicated cluster was used for testing. Prior to running tests, a full compaction was completed after the data load phase. Even though this operation required over eight hours to complete, this time was not included in the benchmark numbers. This is an important consideration, as Cassandra performance is impacted while the compactions are running, which is typically done after business hours. To improve the consistency of the database, both read and write consistency were set to two, while replication was set to three. For storage, four disks of AWS ephemeral (instance) storage were set to RAID 0 using the XFS filesystem. These settings reflect widely used best practices.

For HBase, testing was completed on a cluster running a commercial Hadoop platform, using HBase version 1.2 with short-circuit reads enabled. The default write-ahead log (WAL) settings were used and HDFS replication was set to three. Similarly to Cassandra, four disks of AWS ephemeral storage were set to RAID 0 using the XFS filesystem.

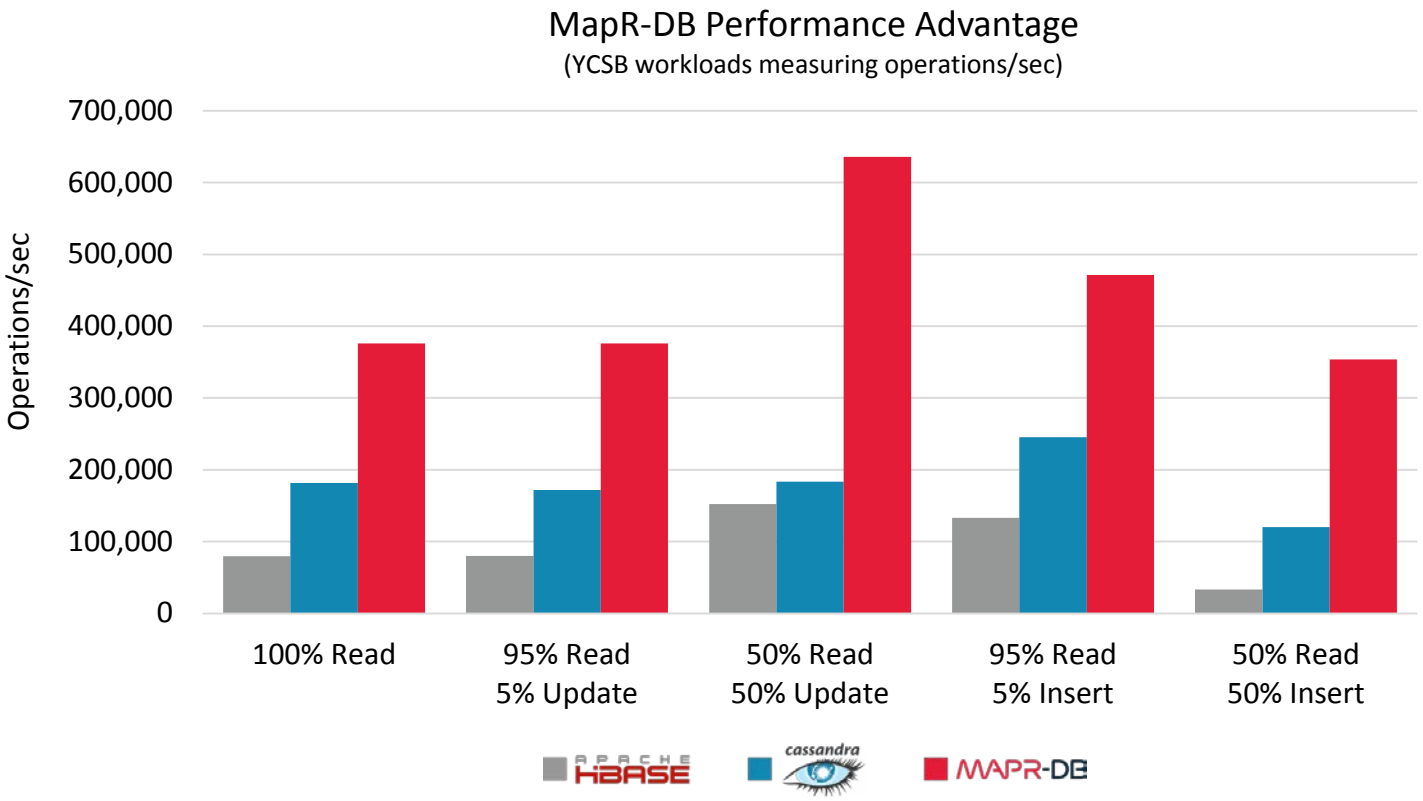
Finally, MapR version 5.2 was tested with four MapR-FS instances and four storage pools across four volumes of AWS ephemeral disks, running on the MapR Converged Data Platform.

Testing was completed using the Yahoo Cloud Serving Benchmark (YCSB) version 0.12. YCSB is an open source framework for evaluating the performance of key-value and cloud-based serving datastores. This includes NoSQL databases like MapR-DB, Cassandra, and HBase. The platform consists of two components: a client and the core workloads. The client is the extensible workload generator, while the core workloads consist of a set of scenarios that emulate real-world workloads to be executed by the client. YCSB provides sample code for many open source projects, and the platform is extensible, enabling users to create new interface layers to benchmark any data-serving system. Five workloads were run for two hours each across the three similarly configured platforms:

1. 100% read
2. 95% read – 5% update
3. 50% read – 50% update
4. 95% read – 5% insert
5. 50% read – 50% insert

The results of testing with a focus on operations/sec are shown in Figure 4 and Table 1.

Figure 4. Performance Analysis – Operations/sec



Source: Enterprise Strategy Group, 2017

Table 1. Performance Analysis – Operations/sec/cluster-wide

Database	YCSB Workloads				
	100% Read	95% Read 5% Update	50% Read 50% Update	95% Read 5% Insert	50% Read 50% Insert
HBase	79,387	80,033	152,075	133,282	33,303
Cassandra	181,570	171,751	183,493	245,370	119,967
MapR-DB	375,830	375,865	635,809	471,498	353,648

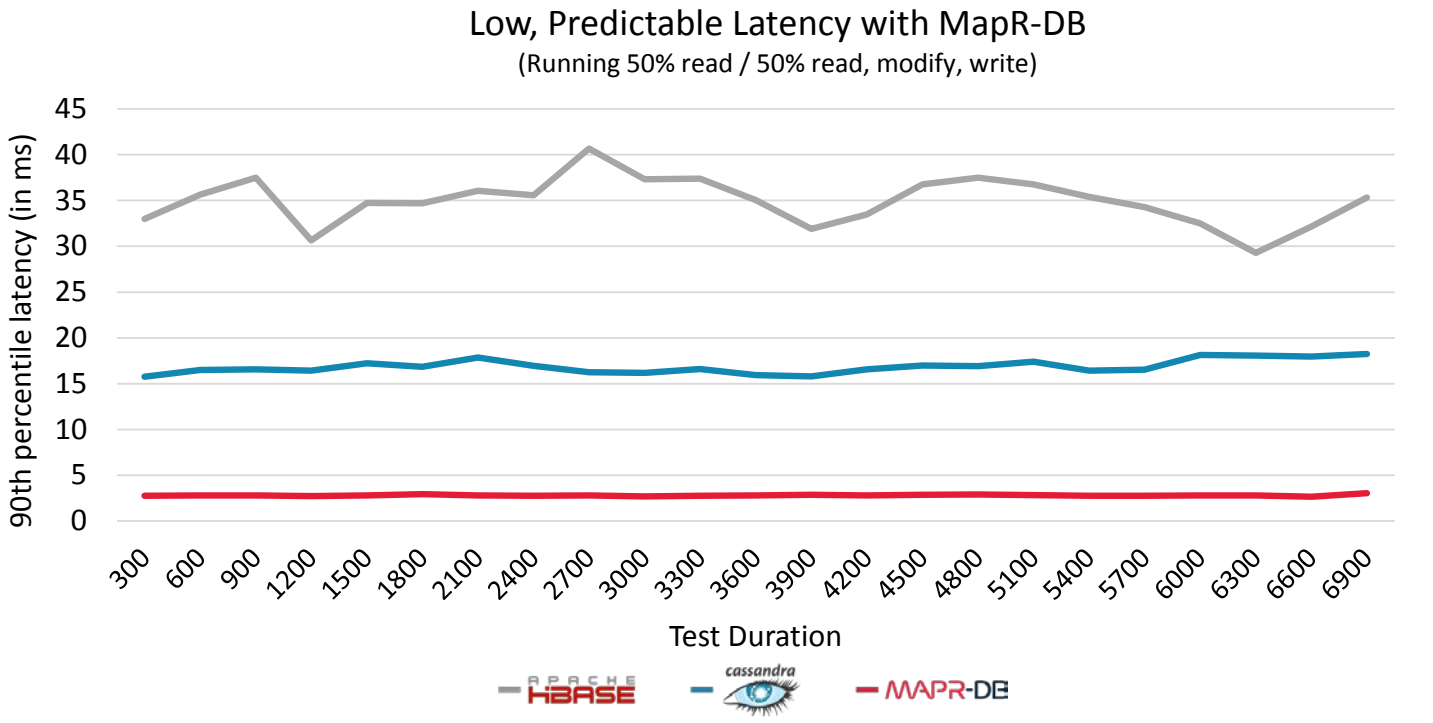
Source: Enterprise Strategy Group, 2017

What the Numbers Mean

- Three databases were similarly configured and database performance was evaluated after running five different preconfigured workloads with YCSB.
- MapR-DB yielded an average performance improvement of 2.5x more operations/sec than Cassandra and 5.5x more than HBase.
- Benefits with MapR-DB were recognized across all the mixed workloads, with the largest on the 50% read/50% insert test. MapR-DB outperformed HBase by over 10x.
- Though Cassandra yielded higher performance than HBase in all tests, MapR-DB went a step further, dwarfing Cassandra by as much as 3.5x.

The next performance metric measured was latency, as this is a true indication of responsiveness and can be directly tied to the end-user experience. The goal of this phase of testing was to yield low, predictable latency throughout the duration of the test. A workload of 50% read and 50% read/modify/write was run for two hours (immediately after load, for all three databases) and ESG focused on the ability to deliver a premium class of service by measuring 90th percentile latency. This means that 90% of completed operations fall below the latency measurement at a given period. Excluding test ramp-up, ESG analyzed latency every five minutes, and the results are shown in Figure 5.

Figure 5. Performance Analysis – Latency



Source: Enterprise Strategy Group, 2017

With a goal of delivering both low and predictable latency, MapR-DB delivered. Latency never exceeded 3ms over the course of the two-hour run, with a minimum of 2.7 ms. These are the results organizations want when trying to deliver a positive end-user experience. Though Cassandra could deliver predictability, the average latency was 75% higher than MapR-DB. HBase did not meet either of the goals—not only was latency significantly higher, but also there was essentially no predictability. Average latency was 6.4x higher with HBase when compared with MapR-DB.

Why This Matters

When leveraging NoSQL databases to meet the performance requirements of demanding operational analytic workloads, organizations want more than just speed—they also want consistency. This enables a predictable end-user experience that meets the needs of a dynamic business that continues to crave a real-time response. In the end, organizations gain insights faster and therefore make better decisions to positively impact the business.

ESG validated MapR-DB's ability to deliver faster, more consistent performance in an AWS cloud environment compared with open source offerings HBase and Cassandra. In analyzing test results from five core workload tests run on three similarly configured deployments, MapR-DB yielded an average performance improvement of 2.5x more operations/sec than Cassandra and 5.5x more than HBase. When looking at latency for one of those workloads, MapR-DB proved to not only be the fastest, but also the most predictable, while Cassandra yielded higher latency and HBase proved unreliable in delivering fast or consistent latency.

The Bigger Truth

As organizations continue to prioritize internal initiatives related to being more data driven, faster time to value and faster time to insight are key to the evaluation process. With the growing number of tools and components within the data management and analytic ecosystem, constant questions arise: Build or buy? Use open source or proprietary software? Run on-premises or in the cloud? Regardless of an organization's preferences in response to those questions, performance is top of mind when making a final decision. The ability to constantly handle growing data sets, while continuing to the meet ever-increasing demands of the business and end-users is all but essential to making an appearance at the top of evaluation or POC lists. For operational intelligence specifically, frameworks such as Spark coupled with a NoSQL database serve as an ideal platform to handle those requirements, but not all solutions are created equal. A balance between speed, consistency/predictability, scalability, flexibility, and cost are all factors when homing in on the ideal solution.

The MapR Converged Data Platform serves as an ideal foundation that ties together a collection of open source engines, tools, and applications with the purpose-built MapR platform to meet the dynamic needs of a transformative, data-driven business. MapR-DB, a key pillar in the MapR Converged Data Platform, serves as the enterprise-grade NoSQL database delivering high performance, reliability, and scalability for applications and analytic workloads. Being seamlessly integrated into MapR's converged platform, MapR-DB delivers the scalability, performance, and reliability organizations require to support the real-time needs of the business.

ESG validated the high levels of predictable performance MapR-DB can yield in a cloud environment. Performance tests were run on MapR-DB, HBase, and Cassandra with a focus on operations/sec and latency. Not only did MapR-DB yield an average gain of 2.5x more operations/sec than Cassandra and 5.5x more than HBase, but it also delivered the lowest, most consistent latency.

When looking for a solution to meet the performance needs of a data-driven business that operates in real time, ESG suggests evaluating the MapR Converged Data Platform with MapR-DB—a robust, enterprise-grade NoSQL database.