



Technical Research Study



World-Record Performance for Big Data and Analytics

Prowess research draws links from the benchmark strength of Dell™ servers to real-world performance for big-data and analytics workloads.

Executive Summary

Companies from a wide array of industries depend on big data and analytics to acquire and retain customers, better target promotions, identify risks, manage complex supply networks, optimize costs, and improve efficiency. The challenges and needs of organizations in terms of big data and analytics are defined more by variety than broad similarities. However, all big-data and analytics use cases depend on compute performance near where the data resides. And, due to the nature of the data companies use for activities like predictive customer analytics or fraud detection, security is often just as critical.

Due to the performance needs of big data and analytics, the infrastructure to support those workloads can represent a significant investment, which drives the need for rigorous evaluation before purchase. Industry-standard benchmarks can be good for this—and world records can be even better—if they are evaluated in the right way.

Companies need rapid access to their data wherever it's collected, stored, or needed. These benchmarks can provide insight into how quickly data can be collected, processed, and accessed once stored. In order to investigate the relationship between high benchmark performance and potential business value in the real world, Prowess dug deeper into what strong showings in industry benchmarks can mean for businesses deploying world-record servers. Because of its market share and the number of world records Dell Technologies holds across a variety of big-data and analytics scenarios represented by different benchmarks, we specifically looked at Dell™ PowerEdge™ servers.

Among the newest-generation Dell PowerEdge servers, we identified four that have recently set world records in benchmarks specifically designed to provide real-world performance data on big data and analytics workloads:

- **Dell PowerEdge R6515 server (world record, TPC Express Benchmark™ HS [TPCx-HS] 17-node MapReduce and 10-node Apache Spark™ benchmarks)**
- **Dell PowerEdge R7515 server (world record, TPC Express Benchmark™ BB [TPCx-BB] Cloudera® database benchmark and TPC Express Benchmark™ IoT [TPCx-IoT] benchmark)**

Augmenting performance optimizations from Dell Technologies for big-data and analytics workloads, Prowess also found that 3rd Gen AMD EPYC™ processors and Broadcom® network cards can help drive big-data and analytics performance both for benchmarks and real-world applications. In addition, support for PCIe® 4.0 with Broadcom network cards enables dual-port 100 gigabit Ethernet (GbE) network interface controllers (NICs) to accelerate analytics and big-data workloads. And 4th Gen AMD EPYC processor–based platforms can take performance to new levels with even higher core counts and frequencies, larger quantities of higher speed DDR5 memory, and PCIe 5.0 interconnects that can reduce latency.

This study covers the following topics:

- [Industry landscape](#)
- [Prowess research methodology](#)
- [Big-data and analytics benchmarks](#)
- [Behind the performance results](#)



Industry Landscape: Big Data and Analytics

While big data and analytics are well into the maturation phase of the hype cycle, they still play an important role in business operations for organizations in a variety of industries. Whether it's a bank monitoring for fraud, a retailer seeking new customers and striving to retain current ones, a hospital network optimizing costs, or a mid-size manufacturer working to increase efficiency, organizations of all types and sizes rely on analytics to tease out patterns from their data. In a virtuous (or infamous) cycle, analytics often works better the more data is available to analyze—and the more varied that data. This cycle generates a perpetual need for larger, more capable storage, faster networking, and more performant servers to find more value in data—data which must also be kept secure.

Businesses often rely on on-premises servers rather than cloud implementations for a variety of reasons. For big data and analytics, these reasons often center on data gravity and lower latency for analytics. It is often faster and easier to bring compute functions closer to the data rather than bear the cost and time of moving gargantuan amounts of data to centralized compute. Working with data close to where it resides can also reduce the latency for applications. Moreover, regulatory requirements and data-sovereignty laws can also be compelling reasons to keep data on premises, depending on an organization's industry and location. In all cases, speed is a core requirement for businesses when dealing with their data and tapping the analytical value that data contains.

The performance demands of workloads like analytics mean that the data infrastructure must be tuned to meet service-level agreements (SLAs). The interplay of processor, memory size, network bandwidth, and storage subsystems is critical. One prominent tool for comparing server performance for this interplay is benchmark results. Because benchmarks produce numeric results, comparisons between competing systems can feel straightforward.

Precisely because benchmarks produce clear and seemingly objective results, however, understanding what they measure—and thus what they actually say about server platforms—is crucial. Organizations that ignore the nuance of these benchmarks and blindly chase the top benchmark performers can wind up disappointed when their return on investment (ROI) fails to meet their expectations.



Prowess Research Methodology

In order to investigate the relationship between high benchmark performance and potential business value in the real world, Prowess Consulting dug deeper into what strong showings in industry benchmarks can mean for businesses deploying world-record servers. To simplify our investigation and focus on how individual benchmarks can provide insights into performance in particular facets of big data and analytics, we specifically looked at Dell PowerEdge servers. We did this both because of the large market share for Dell™ servers and because of the number of world records Dell Technologies holds across a variety of big-data and analytics benchmarks.

A single benchmark world record is impressive, but what stands out for big data and analytics workloads is that Dell platforms achieve world records across multiple benchmarks. Each benchmark can be viewed as a piece of the workload puzzle, and the achievement of multiple world records provides good insight into how Dell platforms will operate in real-world environments.

We looked at benchmark results to help determine which platforms offered the best performance for different aspects of big-data and analytics workloads. Our research focused on Dell rack-mount servers because Dell Technologies has the largest market share of servers worldwide (17.2%),¹ and Dell PowerEdge servers are popular workhorse servers built for standard to medium-heavy workload needs. Specifically, for this study, we examined 1U (Dell PowerEdge R6515) and 2U (Dell PowerEdge R7515) rack-mount platforms.

When examining the benchmark results, it is essential to view them through the lens of the most important performance factors. For big data and analytics, these include:

- Performance
- Price/performance

Dell Technologies has optimized its PowerEdge platform, based upon AMD® processors, for big-data and analytics solutions. Big-data and analytics workloads can have specific needs, and PowerEdge servers enable a high degree of flexibility for customers to run their unique workloads according to their specific requirements.

Big-Data and Analytics Benchmarks

Industry-recognized benchmarks can provide insights into common uses of a server platform, and they can help inform customers on whether that platform will meet the needs of the workloads the customer is running. We specifically looked at four benchmarks in order to cover different facets of big-data and analytics workloads:

TPC Express Benchmark™ HS (TPCx-HS)	Measures batch-processing with Apache Hadoop® and in-memory, real-time analytics with Apache Spark™
TPC Express Benchmark™ BB (TPCx-BB)	Measures frequently performed analytical queries for structured data and machine learning (ML) algorithms for semi-structured and unstructured data
TPC Express Benchmark™ IoT (TPCx-IoT)	Measures performance, price-performance, and availability for systems that ingest massive amounts of data from large numbers of devices while running real-time analytics queries

Classic Apache Hadoop® and Apache Spark™: TPC Express Benchmark™ HS (TPCx-HS)

First developed at Google by Jeffery Dean and Sanjay Ghemawat in 2004, MapReduce is a venerable and widely used distributed execution framework that scales well for big-data processing. Apache Hadoop®, created by Doug Cutting and Mike Cafarella at Yahoo in 2005, is a distributed file system in which data processing (such as with MapReduce) can be done directly on the storage nodes, which can help a system scale well for extremely large amounts of data. Apache Spark was designed in 2009 and enables large-scale, in-memory data processing, which gives it greater speed than Apache Hadoop for relatively smaller datasets (although Apache Hadoop can still be less expensive to scale for extremely large bodies of data).

A classic application of Apache Hadoop and Apache Spark is in the retail industry, which uses large amounts of data to create product recommendations for customers. That said, organizations in many other industries (such as finance and healthcare) also have huge datasets and business needs that lend themselves well to both Apache Hadoop and Apache Spark. The TPCx-HS benchmark supports both MapReduce batch-processing using Apache Hadoop and in-memory, real-time analytics using Apache Spark across a broad range of system topologies and implementation methodologies.

Dell Technologies has two world records for **TPCx-HS**, one for a 17-node cluster of Dell PowerEdge R6515 servers using MapReduce and one for a 10-node cluster of PowerEdge R6515 servers using Apache Spark. The 17-node PowerEdge R6515 server achieved MapReduce throughput of 24.6 TB/hr at a price/performance ratio of \$49,795.35/TB/hr.² The 10-node PowerEdge R6515 cluster produced 19.9 TB/hr throughput for Apache Spark at a price/performance ratio of \$36,550.21/TB/hr.²

These servers are powered by 3rd Gen AMD EPYC 75F3 processors, which provide, in addition to processing power, security features for real-world big-data applications such as AMD® Secure Memory Encryption (AMD® SME) and AMD® Secure Encrypted Virtualization (AMD® SEV). Because MapReduce can be used in a variety of applications, strong performance on this benchmark can map to getting real-world results faster for use cases that span many organizations and industry verticals.

Apache Hadoop and Apache Spark for AI and Machine Learning (ML): TPC Express Benchmark™ BB (TPCx-BB)

As originally implemented, MapReduce was extremely Java-focused. Apache Hive™ was developed by Facebook and further enhanced by Netflix, Amazon, and the U.S. Financial Industry Regulatory Authority (FINRA) to provide the abstraction necessary to run SQL queries using MapReduce and thus integrate SQL-based applications with Apache Hadoop.

However, a relative minority of data generated today is structured data suitable for SQL queries. Much more of the data generated from websites, audio sources, video files, and computer logs is semi-structured or unstructured. AI is able to tap the business value latent in that kind of data. For example, companies can monitor what consumers like and don't like by using AI to sift through unstructured social-media comments or integrate and run predictive analytics on data from multiple sources in retail to acquire new customers and retain existing ones.

The TPCx-BB benchmark is designed to address precisely this use case by measuring data throughput for queries in Apache Hadoop-based big-data systems. It measures the performance of both hardware and software components by executing 30 frequently performed analytical queries expressed in SQL for structured data using Apache Hive or Apache Spark and machine learning (ML) algorithms for semi-structured and unstructured data.

Dell Technologies holds a world record for **TPCx-BB** for a cluster of 11 Dell PowerEdge R7515 servers. The cluster achieved throughput of 1,544 GB/min at a price/performance ratio of \$487.85/GB/min.³

In addition to the processing power provided by the 3rd Gen AMD EPYC 7763 processor, our analysis of the benchmark results also indicates that the performance of the Dell servers was boosted by the Broadcom 25 GbE network cards used in the cluster.

While this benchmark sets its context in the retail vertical, strong performance here can map to any organization and use case that needs to feed large amounts of data into AI and ML models, particularly for inferencing.

Internet of Things (IoT): TPC Express Benchmark™ IoT (TPCx-IoT)

The Internet of Things (IoT) represents another large source of unstructured data. What sets IoT apart from other kinds of unstructured data is that IoT devices often generate their data from their environment, such as smart speakers listening to commands or industrial drones collecting agricultural land data. IoT is another part of big data and analytics that could, in theory, touch almost every industry vertical, with healthcare and manufacturing being large current users.

The TPCx-IoT benchmark enables direct comparison of different software and hardware solutions for IoT gateways. Because they are positioned between edge architecture and the back-end data center, gateway systems perform functions such as data aggregation, real-time analytics, and persistent storage. The TPCx-IoT benchmark was specifically designed to provide verifiable performance, price-performance, and availability metrics for commercially available systems that typically ingest massive amounts of data from large numbers of devices, while running real-time analytics queries.

Dell Technologies holds a world record for **TPCx-IoT** for Dell PowerEdge 7625 servers running 3rd Gen AMD EPYC processors. The cluster achieved throughput of 1,617,545,000 records/second at a price/performance ratio of \$329.75/million records/sec.⁴

In addition to the processing power provided by the 3rd Gen AMD EPYC 75F3 processor, our analysis of the benchmark results also indicates that the performance of the Dell servers was boosted by the Broadcom 25 GbE network cards used in the cluster. All of these features helped the Dell servers attain their performance in the TPCx-IoT benchmark, with the Dell Technologies results being especially relevant for organizations that use Cloudera—as the Dell Technologies record-holder did—to store large amounts of data on the edge for analytics.

Behind the Performance Results

Because Dell Technologies has optimized its PowerEdge platform based upon AMD processors, it has achieved numerous world records in benchmarks measuring big-data and analytics performance. In an on-premises implementation, 3rd Gen AMD EPYC processors offer strong performance, performance per watt, and performance per CPU dollar. In the cloud, AMD EPYC systems on chip (SoCs) power high-performance computing (HPC)-optimized infrastructure-as-a-service (IaaS) instances for many cloud service providers (CSPs) including Amazon Web Services® (AWS®), Microsoft® Azure®, Google Cloud Platform™, and others.

4th Gen AMD EPYC processors, built with AMD® Zen 4 microarchitecture, can provide additional performance gains for big data and analytics workloads that can be traced to several platform improvements over the previous-generation platform, including:

- 50 percent increase in core count,⁵ increased thread count, and higher frequencies, which can directly increase processing performance.
- 12 DIMMs/socket (up from 8), which allows organizations to significantly increase available memory. This translates to processing larger datasets faster, particularly for in-memory analytics such as those performed by Apache Spark.
- DDR5 memory support for faster access to data.
- AVX-512 support, which enables 4th Gen AMD EPYC processors to complete more simultaneous calculations in their registers.
- Greater L2 cache, doubled from 512 KiB to 1 MiB per core, which also accelerates operations in memory.
- PCIe Gen 5 support, which enables faster interconnects to move more data with lower latency.
- Enhancements specifically for AI and ML workloads, including support for the bfloat16 numeric type to accelerate the training of AI models and support for INT8 inferencing to increase the performance of already trained models in production.

Overall, 4th Gen AMD EPYC processors operate more efficiently than their predecessors. The Standard Performance Evaluation Corporation's SPEC CPU® 2017 Floating Point Rate results showed a gain in performance of 121 percent in tests run on a system powered by 4th Gen AMD EPYC processors, compared to a system powered by 3rd Gen AMD EPYC processors.⁶ The SPEC CPU 2017 Integer Rates results showed gains of 102 percent.⁷ These processor performance results are reflected in the world-record benchmark results achieved by several of the PowerEdge platforms we examined.

The number of cores in these processors increased by 50 percent, compared to the previous generation, which also boosts performance. At the same time, published specifications from AMD show an increase in maximum default power consumption of only 42 percent, from 280-watt thermal design power (TDP) to 400-watt maximum TDP.⁸ When compared to the SPEC[®] performance results above, these power numbers show the capability for servers built on 4th Gen AMD EPYC processors to provide up to a 55 percent power-performance benefit for businesses running database workloads.⁹

Network cards from Broadcom speed up the flow of data within clusters. Moreover, support for PCIe Gen 5 in both 4th Gen AMD EPYC processors and Broadcom NICs allow the use of dual-port 100 GbE Broadcom network adapters built on the Open Compute Project (OCP) NIC 3.0 form factor. These modern designs reflect a rapid shift in the industry toward 100 GbE adapters built on a more efficient form factor and enabled by PCIe 4.0 and PCIe 5.0. In addition, support for PCIe 4.0 and PCIe 5.0 can provide performance numbers from a single NIC that are on par with dual 100 Gbps NICs. The OCP NIC 3.0 specification enables server manufacturers like Dell Technologies to use more compact designs that can support high-performance adapters with advanced hardware-acceleration capabilities.¹⁰ Advanced adapters such as these from Broadcom can further speed up analytics and batch processing.

AMD[®] Hardware-Based Security

For all of the workloads evaluated in this research study, security considerations are critical. 3rd Gen AMD EPYC™ processors and 4th Gen AMD EPYC processors can provide hardware-based security for big-data and analytics workloads. AMD[®] Secure Memory Encryption (AMD[®] SME) encrypts system memory to protect data in use. AMD[®] Secure Encrypted Virtualization (AMD[®] SEV) protects running VMs so that they are encrypted and isolated from each other and the host-system hypervisor. AMD[®] Secure Encrypted Virtualization-Encrypted State (AMD[®] SEV-ES) encrypts the CPU register contents of stopped VMs to protect the data stored in them. And AMD[®] Secure Boot protects servers during the boot process, providing defenses against rootkits, bootkits, and firmware while servers are most vulnerable.

Conclusion

Benchmark results in general (and world-record results in particular) are about more than bragging rights for server manufacturers. Interpreted correctly, best-in-industry results in benchmarks can offer insights as to how servers could perform in real-world use cases to quickly return results from companies' valuable data. Because of the market share of Dell Technologies and the number of world records the company holds, its PowerEdge servers provide a natural opportunity to examine how benchmark results can map to performance benefits for organizations in production. While no mapping of benchmark performance (world record or otherwise) is 1:1, our investigation shows that the performance of Dell PowerEdge servers across four benchmarks indicates strong big-data and analytics performance across several use cases in a variety of industries. Viewed in aggregate, the world records held by Dell Technologies across different benchmarks demonstrate how it developed platforms that take advantage of the individual components' strengths in order to deliver real value for a variety of workloads to its customers.

Appendix A: Benchmark Performance Links

- TPCx-HS top performance results: www.tpc.org/tpcx-hs/results/tpcxhs_perf_results5.asp?version=2
- TPCx-BB top performance results: www.tpc.org/tpcx-bb/results/tpcxbb_perf_results5.asp
- TPCx-IoT V2 top performance results: www.tpc.org/tpcx-iot/results/tpcxiot_perf_results5.asp?version=2

Appendix B: Dell Technologies System-Specification Links

- Dell PowerEdge server specification sheets: www.dell.com/en-us/dt/servers/poweredge-rack-servers.htm

¹ History-Computer. "The 10 Largest Server Companies In The World, And What They Do." September 2022.

<https://history-computer.com/largest-server-companies-in-the-world-and-what-they-do/>.

² TPC. "TPCx-HS Top Performance Results." Accessed October 20, 2022. www.tpc.org/tpcx-hs/results/tpcxhs_perf_results5.asp?version=2.

³ TPC. "TPCx-BB Top Performance Results." Accessed October 20, 2022. www.tpc.org/tpcx-bb/results/tpcxbb_perf_results5.asp.

⁴ TPC. "TPCx-IoT V2 Top Performance Results." Accessed October 20, 2022. www.tpc.org/tpcx-iot/results/tpcxiot_perf_results5.asp?version=2.

⁵ Tom's Hardware. "Zen 4 Madness: AMD EPYC Genoa With 96 Cores, 12-Channel DDR5 Memory, and AVX-512."

www.tomshardware.com/news/zen4-madness-amd-epyc-genoa-with-96-cores-12-channel-ddr5-memory-and-avx-512.

⁶ Up to 121 percent higher SPEC® Floating Point performance comparing top-bin 4th Gen AMD EPYC™ processors with top-bin 3rd Gen AMD EPYC processors based on SPEC Floating Point rate score of 1,410 achieved on a Dell™ PowerEdge™ R7625 server powered by AMD EPYC 9654 processors, compared to a score of 636 achieved on a Dell PowerEdge R7525 server powered by AMD EPYC 7763 processors. Scores accessed as of November 10, 2022. See Standard Performance Evaluation Corporation benchmark results. <http://spec.org/benchmarks.html>.

⁷ Up to 102 percent higher SPEC® Integer Rate performance comparing top-bin 4th Gen AMD EPYC™ processors with top-bin 3rd Gen AMD EPYC processors based on SPEC Integer rate score of 1,660 achieved on a Dell™ PowerEdge™ R7625 server powered by AMD EPYC 9654 processors, compared to a score of 821 achieved on a Dell PowerEdge R7525 server powered by AMD EPYC 7763 processors. Scores accessed as of November 10, 2022. See Standard Performance Evaluation Corporation benchmark results. <http://spec.org/benchmarks.html>.

⁸ AMD. AMD EPYC 7003 Series processors specifications webpage. www.amd.com/en/processors/epyc-7003-series.

⁹ 55 percent CPU performance per watt improvement calculated using the SPEC® Floating Point score of 1,410 achieved on a Dell™ PowerEdge™ R7625 server powered by AMD EPYC™ 9654 processors with a processor cTDP of 400 watts, compared to a score of 636 achieved on a Dell PowerEdge R7525 server powered by AMD EPYC 7763 processors with a processor cTDP of 280 watts.

¹⁰ Broadcom. NetXtreme E-Series OCP NIC 3.0 Ethernet Adapters Product Brief. 2021. <https://docs.broadcom.com/doc/12395120>.

