Case Study

Generative AI
Intel® Xeon® Scalable Processors

Intel® Technologies Accelerate IBM watsonx.data Up to 2.7X[1] for Faster Generative AI and Intelligent Decision Making

A second-generation, fit-for-purpose data lakehouse, IBM watsonx.data integrates multiple, optimized query engines for more AI workloads in the enterprise

Executive Summary

Artificial Intelligence (AI) applications are critical to businesses in boosting productivity, innovation, and delivering greater business value. Generative AI (GenAI) offers a range of benefits across industries and applications. However, it's important to consider potential data challenges across data warehouses and data lakes, on-prem or in the cloud, which can slow adoption of AI and increase complexity and expense.

IBM watsonx.data, part of IBM's Enterprise GenAI solutions, empowers business to leverage their data for greater insight and innovation in their business operations. Whether data is on-premises or in the cloud, watsonx.data offers a fast and flexible fit-for-purpose data store.

In addition to leveraging an open data lakehouse architecture and integrating IBM's own intelligent query processing optimizations drawn from decades of delivering high-performance data solutions, IBM and Intel worked together to accelerate watsonx.data on Intel® processors. These optimizations have been proven to deliver higher performance on 4th Gen and 5th Gen Intel® Xeon® Scalable processors. Based on Intel 10TB TPC-DS benchmarking, they take advantage of Intel® Accelerator Engines to boost watsonx.data query performance by up to 2.7X on 5th Gen Intel Xeon Scalable processors and 2.17X on 4th Gen Intel Xeon Scalable processors compared to 2nd Gen Intel Xeon Scalable processors.[1] This level of performance delivers faster time to insight and enables quicker intelligent decision making. Faster queries also enable more work to be done on the hardware, enhancing the value of the solution, while reducing the overall Total Cost of Ownership (TCO).

Data Challenges for AI

Enterprise strategies are maturing, due to integration of AI tools into business operations and analytics. According to Gartner, more than 80 percent of enterprises will have used and adopted GenAI enabled applications by 2026.[2]

But AI feeds on data—lots of data—and AI is only as good as your data. Today, enterprises face real challenges with their data, due to the following:

- Data is in more locations: on-prem, cloud, applications, and silos, which create significant challenges; 82 percent of enterprises are inhibited by data silos.[3]
- Data is in more formats: documents, images, and video. Such variety slows down the creation of quality data; 80 percent of time is spent on data cleaning, integration, and preparation.[4]
- Data quality suffers because it ends up being stale and inconsistent; 82 percent of enterprises say data quality is a barrier on their data integration projects.[3]
- Data curation and cleaning for AI and Business Intelligence may require different tools to act on the same data to achieve desired results.

Data lakehouse solutions have emerged to help address these challenges, but often these 1st-generation lakehouse designs do not meet the various diverse needs of AI and analytic workflows and workloads. For example, they are limited to a single query engine, which complicates supporting some types of workloads. Additionally, some solutions are available on cloud only; they do not support multi-cloud, hybrid-cloud, or on-premises deployments. 1st-generation solutions are limited in their ability to reduce cost and complexity.

IBM watsonx.data – A Fit-For-Purpose Data Store

IBM watsonx.data is a 2nd-generation open data lakehouse, delivering capabilities not part of 1st-generation solutions. watsonx.data is a powerful and flexible data store designed for a wide variety of enterprise Business Intelligence (BI), analytics, and AI workloads. Benefits of watsonx.data include the following:

- **Reduced TCO –** watsonx.data is built on an open source architecture with open source tools and technologies, allowing customers to cost-effectively scale their analytics and AI across all their data, wherever it's located.
- **Flexibility and Fit-for-Purpose Queries –** watsonx.data supports multiple query engines, including Presto and Apache Spark, enabling a broader set of users to tailor data queries to their needs and workloads.
- **Optimized Performance –** watsonx.data is optimized for Intel® architecture and technologies, delivering exceptional query performance and price-performance.

IBM watsonx.data is part of the IBM watsonx platform. The platform is designed to enable enterprises to create and customize predictive machine learning and GenAI models with their data for deep understanding of their business operations and environment. With the platform, companies can store sensitive data in one interface and use it for leveraging AI and ML. IBM watsonx also includes governance capabilities to direct, manage, and monitor the AI activities of your organization.

Bhuma, Inc.

"One of the unique value propositions for IBM watsonx.data is the flexibility to deploy the lakehouse anywhere—IBM Cloud or customer's private cloud This is particularly valuable for some of our large regulated and security conscious customers.

We are thrilled to have partnered with IBM and deployed watsonx.data solutions for AWS cloud, powered by the new Intel technology, to our financial customers to give the best performance, cost, flexibility—and total governance."

– Srini Gurrapu, Founder & CEO, Bhuma, Inc.

Bhuma is a strategic IBM watsonx ecosystem partner that helps organizations rapidly build real-time data apps, analytics, and GenAI agents purpose built for modern lakehouses such as watsonx and Presto.

Optimized for Analytics and GenAI Performance

Intel and IBM have a long history of deep collaboration to optimize IBM software on Intel processors and accelerators. That relationship continues with optimizations for watsonx.data to deliver the performance and reliability IBM customers expect and trust from IBM.

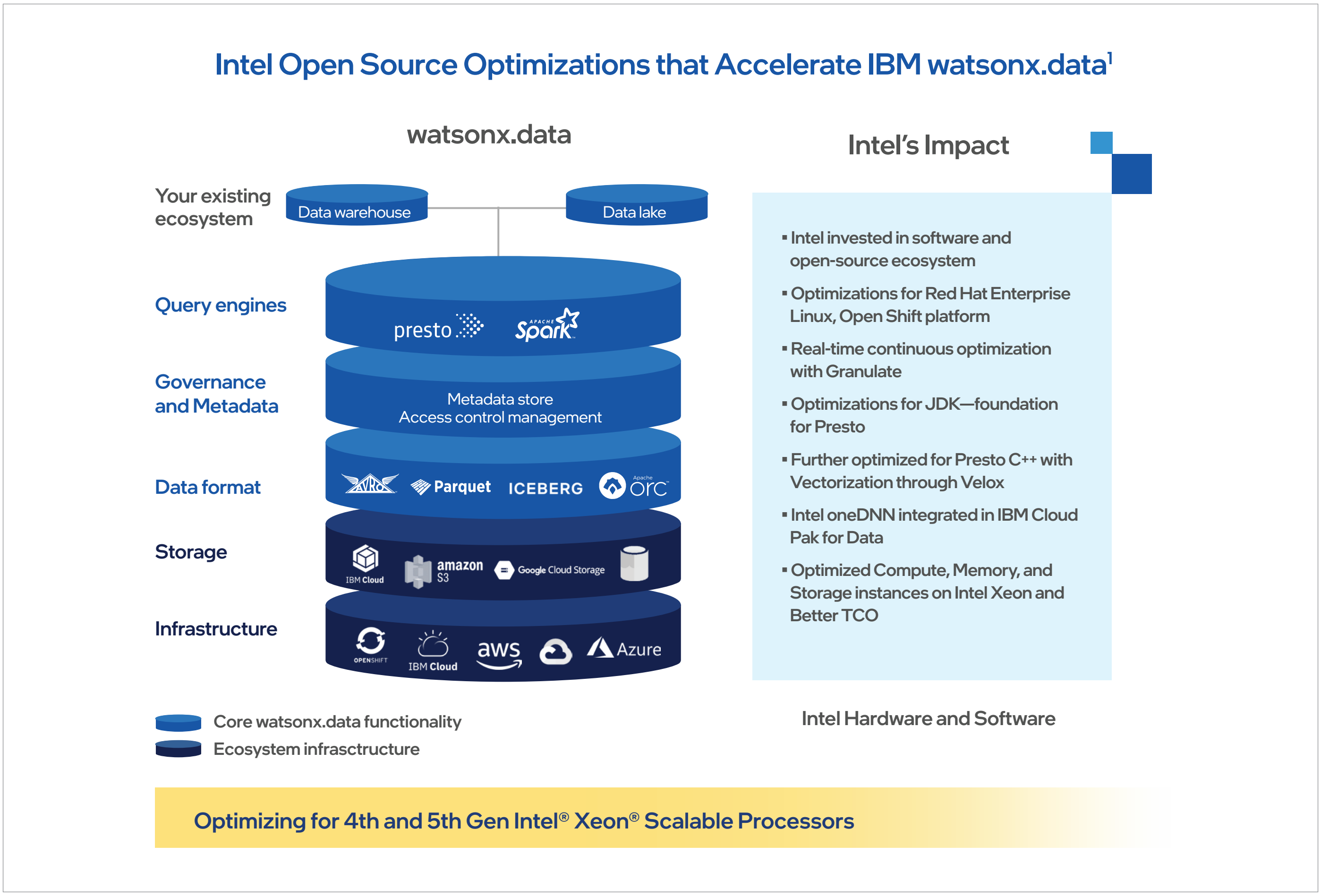
Intel has contributed optimizations for the native Presto query engine's source code. But new Intel optimizations include enhancements on components within Presto.

Presto's native worker component is built on a Java Virtual Machine layer. Presto C++ is a native implementation of that layer that takes advantage of Intel® Advanced Vector Extensions 512 (Intel® AVX-512). Integrating optimized Presto C++ significantly accelerates queries in Presto when run on Intel processors, including the 4th Gen Intel Xeon Scalable processors and 5th Gen Intel Xeon Scalable processors.

Additionally, IBM leveraged their experience with designing and optimizing database query engines for their products, such as Db2, to integrate an intelligent query optimizer in watsonx.data. The query optimizer dynamically creates a plan for each query to execute the fastest search of the data, accelerating results from large datasets.

A Data Lakehouse Built on Open Source

Using an open source architecture (Figure 1), watsonx.data offers the flexibility afforded by open source tools and easily scales to accommodate expanding data. Intel contributes to many open source projects, such as Presto, Presto C++, OpenJDK, and Apache Spark, enabling highly performant software. Intel's optimizations take advantage of the many Intel® technologies and Intel Accelerator Engines built into Intel Xeon Scalable processors and Intel® Xeon® CPU Max Series processors.



Intel Open Source Optimizations that Accelerate IBM watsonx.data[1]

watsonx.data
Your existing ecosystem
Data warehouse
Data lake
Query engines
presto
Apache Spark
Governance and Metadata
Metadata store
Access control management
Data format
AVRO
Parquet
ICEBERG
Apache orc
Storage
IBM Cloud
amazon S3
Google Cloud Storage
Infrastructure
OPENSHIFT
IBM Cloud
aws
Azure
Core watsonx.data functionality
Ecosystem infrastructure
Intel's Impact
▪ Intel invested in software and open-source ecosystem
▪ Optimizations for Red Hat Enterprise Linux, Open Shift platform
▪ Real-time continuous optimization with Granulate
▪ Optimizations for JDK—foundation for Presto
▪ Further optimized for Presto C++ with Vectorization through Velox
▪ Intel oneDNN integrated in IBM Cloud Pak for Data
▪ Optimized Compute, Memory, and Storage instances on Intel Xeon and Better TCO
Intel Hardware and Software
Optimizing for 4th and 5th Gen Intel® Xeon® Scalable Processors

Figure 1. watsonx.data is built on an open source architecture with Intel optimizations built in.

Multiple Query Engines for the Right Data in the Right Project

watsonx.data supports multiple analytics engines, including open source Apache Spark and Presto. Presto is an open-source, distributed SQL query engine, designed for analytics against data of any size. Presto is fast, reliable, and efficient at scale. Apache Spark is an open source, unified analytics engine for large-scale data processing and machine learning pipelines. With Spark, clients can run data engineering, data wrangling, data science, and machine learning workloads with data parallelism and fault tolerance. Both Presto and Spark are optimized for Intel architecture, taking advantage of Intel Accelerator Engines, such as Intel AVX-512.

Additionally, watsonx.data supports external analytics and query engines that are compatible with the Apache Iceberg open table format. IBM's Db2 and Netezza are being enhanced to read from and write to the Iceberg format so they can then participate in the lakehouse architecture and share the same data.

Finally, watsonx.data can be deployed as containerized software on-premises and is available as SaaS on IBM Cloud and AWS, offering multiple deployment options according to enterprise needs.

Up to 2.7X Higher Analytics Throughput on Intel Technologies

Using two different analytics workloads—a Decision Support workload and IBM Big Data Insights (BDI)—Intel engineers benchmarked query throughput-per-node of IBM watsonx.data. Throughput was measured across three different hardware configurations to evaluate the gen-to-gen performance across three generations of Intel Xeon Scalable processors. These included 2nd Gen Intel Xeon Scalable processor (baseline), 4th Gen Intel Xeon Scalable processor, and 5th Gen Intel Xeon Salable processor. Configurations are described in the footnotes below.[1]

Benchmarks (Figure 2) reveal for Decision Support, that 5th Gen Intel Xeon Scalable processors deliver up to 2.7X higher throughput compared to 2nd Gen Intel Xeon Scalable processors. 4th Gen Intel Xeon Scalable processors provide up to 2.17X higher throughput.[1] For IBM Big Data Insights, the 5th Gen Intel Xeon Scalable processor delivered up to 2.41X higher throughput, while 4th Gen Intel Xeon Scalable processor achieved 2.03X higher throughput.[1]

Additionally, the benchmarks also assessed the value of software optimizations to Presto C++ and impact on throughput of the query optimizer in watsonx.data (Figure 3). The open source Presto analytics engine provided a baseline measurement. Against this baseline, optimized Presto C++ with query optimizer runs queries over 4X faster on the two latest generations of Intel Xeon Scalable processors.[1] The following charts show the results.

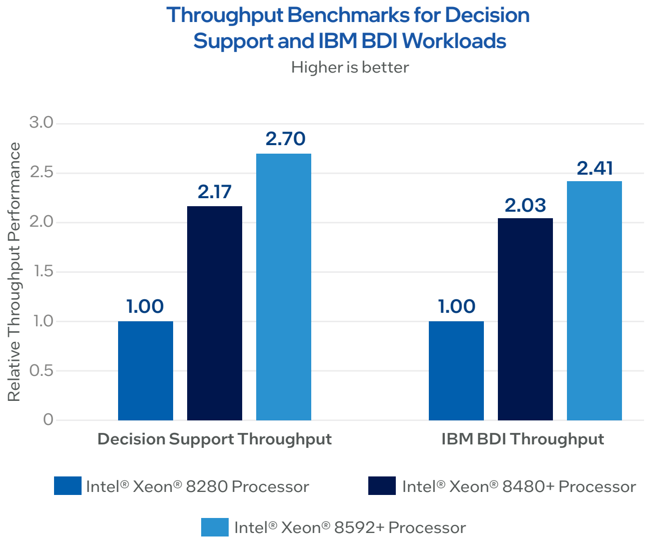


Figure 2. Decision support and IBM BDI throughput benchmark.[1]

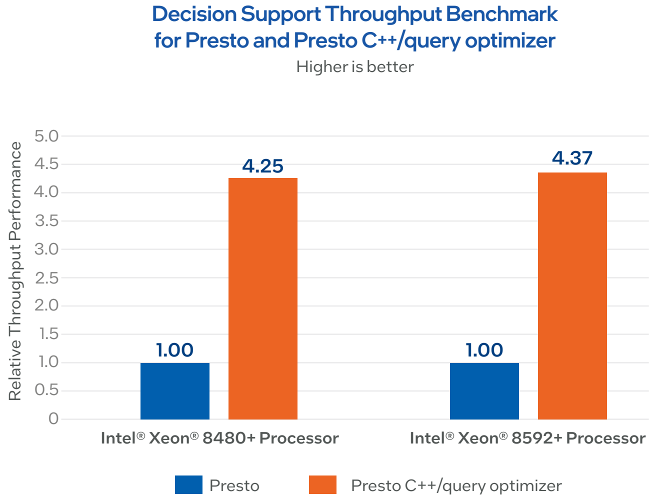


Figure 3. Software optimizations benchark.[1]

Conclusion

As enterprises increasingly leverage the power and benefits of GenAI to gain insight and enhance operations, the speed of queries and analysis from their data becomes paramount. IBM watsonx.data is designed to not only accelerate analytics on massive data lakehouse architectures, but also to deliver greater flexibility to use fit-for-purpose analytics engines for the right application.

Compute systems running 4th Gen Intel Xeon Scalable processors and 5th Gen Intel Xeon Scalable processors accelerate IBM watsonx.data queries and analytics over 2X faster than servers built on previous 2nd Gen Intel Xeon Scalable processor-based systems. Performance enhancements to watsonx.data comes not only from improvements in the latest generation of Intel Xeon Scalable processors, but also due to the following:

- Intel software optimizations in Presto C++
- IBM's intelligent query optimizer in watsonx.data

The level of throughput improvements from the newest hardware and optimized software mean that enterprises can do more with less. IT departments can build compute solutions or cloud instance clusters with fewer nodes to accomplish more analysis and achieve new insights using IBM watsonx.

Get started with scaling your AI workloads using all your data at IBM cloud and **watsonx.data**.

[1] **Source:** Intel benchmarking; see testing results in this document. Your costs and results may vary. Performance varies by use, configuration and other factors. Learn more at the Performance Index site.

Configurations:
4th Gen Intel Xeon Scalable processor: 3 nodes of 4th Gen Intel Xeon Processor (Sapphire Rapids) Node Configuration: 1-node, 2x Intel® Xeon® Platinum 8490H, 60 cores, HT On, Turbo On, Total Memory 2048GB (32x64GB DDR5 5600 MT/s [4800 MT/s]), BIOS 3B05.TEL4P1, microcode 0x2b0004b1, 2x Ethernet Controller 10-Gigabit X540-AT2 , 100GB E810, 1x 1.7T INTEL SSDSC2KG019T8, 4x 3.5T INTEL SSDPF2KX038TZ, Red Hat Enterprise Linux 9.2 (Plow), 5.14.0-284.30.1.el9_2.x86_64. Pre-release version of watsonx.data 2.0 (build 80). Optimizer Plus only applied to queries with performance benefit. Test by Intel as of 04/15/24.

5th Gen Intel Xeon Scalable processor: 3 nodes of 5th Gen Intel Xeon Processor (Emerald Rapids) Node Configuration: 1-node, 2x INTEL® XEON® PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 2048GB (32x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 3B05.TEL4P1, microcode 0x21000161, 2x Ethernet Controller 10-Gigabit X540-AT2, 100GB E810, 1x 1.7T INTEL SSDSC2KG019T8, 4x 3.5T INTEL SSDPF2KX038TZ, Red Hat Enterprise Linux 9.2 (Plow), 5.14.0-284.30.1.el9_2.x86_64. Pre-release version of watsonx.data 2.0 (build 80). Optimizer Plus only applied to queries with performance benefit. Test by Intel as of 04/15/24.

2nd Gen Intel Xeon Scalable processor: 4 nodes of 2nd Gen Intel Xeon Processor (Cascade Lake) Node Configuration: 1-node, 2x INTEL® XEON® PLATINUM 8280, 28 cores, HT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR54 2933 MT/s [2666 MT/s]), BIOS SE5C620.86B.02.01.0017.110620230543 microcode 0x5003604, 2x Ethernet Controller 100GB E810, 1x 1.7T INTEL SSDSC2KG019T8, 4x 3.5T INTEL SSDPF2KX038TZ, Red Hat Enterprise Linux 9.2 (Plow), 5.14.0-284.30.1.el9_2.x86_64. Pre-release version of watsonx.data 2.0 (build 80). Optimizer Plus only applied to queries with performance benefit. Test by Intel as of 05/06/24.

[2] Gartner: https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026

[3] Goldberg, Michael: "Are Data Silos Killing Your Business?" Dun & Bradstreet, May 2018 - https://www.dnb.com/perspectives/marketing-sales/data-management-strategies-avoiding-data-management-silos.html

[4] Data Integrity Trends: Chief Data Officer Perspectives in 2021, Corinium - https://www.precisely.com/app/uploads/2021/06/Data-Integrity-Trends-2021-Corinium-Intelligence.pdf

0624/PB/HBD/PDF
Please Recycle