Empower AI infrastructure with IBM Fusion HCI

Unleash agility with IBM Fusion: The easiest way to deploy OpenShift applications and watsonx

Highlights

The proven, fast track to on-premises watsonx deployment

Supercharging AI: IBM watsonx on fusion delivers unparalleled performance and efficiency

Secure AI innovation: IBM watsonx on Fusion HCI delivers on-Premises power with ironclad data control

IBM watsonx deployed on IBM Fusion HCI (Hyperconverged Infrastructure) offers a powerful and efficient solution for organizations seeking to leverage advanced AI capabilities. As the easiest way to deploy and manage a self-service on-premises cloud for watsonx, Fusion significantly accelerates the deployment process, reducing it from weeks to days. This engineered, integrated OpenShift system not only simplifies application operations through automated day-2 monitoring, maintenance, and resiliency services but also provides unparalleled flexibility in scaling. Organizations can easily add GPU servers for watsonx.ai inferencing, prompt tuning, and training, while also having the option to start a data lakehouse on Fusion and grow with Ceph-ready nodes.

The system's high performance is particularly noteworthy, outperforming Databricks at 60% of the cost. When combined with watsonx, it accelerates queries by an impressive 7x to 90x – a feat proven by the IBM CIO office use case. This remarkable speed improvement, coupled with the ability to grow Fusion HCI with CPUs and storage as needed, ensures that organizations can adapt to changing demands efficiently. The combination of ease of deployment, scalability, and superior performance makes IBM watsonx on Fusion HCI an ideal choice for enterprises looking to harness the full potential of AI while maintaining control over their infrastructure.

90x

IBM Fusion HCI System with local caching capabilities and IBM watsonx.data accelerates remote S3 queries by 90X[1]

60%

IBM watsonx.data with IBM Fusion HCI is able to deliver equivalent performance to Databrick’s Photon engine, at less than 60% of the cost[2]

Secure AI Innovation: IBM watsonx on Fusion HCI Delivers On-Premises Power with Ironclad Data Control

IBM watsonx on IBM Fusion HCI offers a robust solution for organizations seeking to harness AI power while maintaining strict data control. This on-premises deployment addresses critical concerns related to data security, regulatory compliance, and data sovereignty. By keeping data within the organization's infrastructure, companies can more easily meet stringent security requirements and industry-specific regulations. The integration enables easier on-premises AI deployments without compromising on data governance or cybersecurity measures. This solution also provides flexibility in data management, offering accelerated on-premises deployment and querying through IBM Storage Fusion HCI integration. For organizations dealing with sensitive information or operating in highly regulated industries, IBM watsonx on Fusion HCI balances innovation with compliance and security needs, allowing complete data control while accessing cutting-edge AI technologies.

The proven, fast track to on-premises watsonx deployment

watsonx running on IBM Fusion offers a proven, low-risk approach for organizations looking to implement artificial intelligence on-premises. As a reference architecture for watsonx deployments, Fusion provides a turnkey solution that significantly accelerates the adoption process, enabling rapid setup of Proof of Concept (POC) projects and full-scale implementations. This streamlined approach is particularly beneficial for businesses that have already invested in watsonx or those requiring swift deployment while maintaining strict data sovereignty and security standards. By leveraging IBM Fusion, companies can confidently integrate watsonx into their existing infrastructure, maximizing their AI investment while retaining full control over their data and processes. This makes it an ideal choice for industries with sensitive data or regulatory constraints, offering a fast track to harnessing the power of watsonx in a secure, on-premises environment.

Figure 1. IBM Fusion HCI

Supercharging AI: IBM watsonx on fusion delivers unparalleled performance and efficiency

IBM watsonx deployed on IBM Fusion offers significant performance advantages, making it an attractive solution for organizations seeking high-performance AI capabilities. The integration of watsonx with IBM Fusion HCI provides superior price-performance ratios, particularly for data-intensive AI workloads. One key feature is the storage acceleration capability, which enables watsonx.data queries to benefit from a shareable on-premises high-performance cache acceleration, enhancing overall query performance. The architecture of IBM watsonx.data, combined with IBM Storage Ceph and IBM Fusion HCI, works synergistically to improve on-premises performance while simultaneously enhancing cost efficiency. Furthermore, IBM Fusion HCI with watsonx supports GPU-accelerated applications through its integration with NVIDIA H100 GPUs, allowing for even more powerful AI processing capabilities. These performance enhancements not only accelerate AI model training and inference but also streamline complex data management tasks, enabling organizations to derive insights faster and more efficiently from their AI initiatives.

To learn more about IBM watsonx.data on IBM Fusion, contact your IBM representative or IBM Business Partner, or visit IBM Fusion.

1. IBM Fusion HCI System with local caching capabilities and IBM watsonx.data accelerates remote S3 queries by 90X *vs not using IBM Fusion System caching capabilities with watsonx.data*
2. Based on IBM internal testing of Presto C++ 0.286 on a hyper-converged infrastructure setup with 1 master + 75 worker nodes, 1009 vCPUs, 18 TB memory, 344.8 TB of filesystem storage, distributed RAID and 50 GB network compared to public Databricks 100TB TPC-DS Query benchmarks published in 2021 with 1 master + 256 worker nodes, 2112 vCPUs, 16.1 TB Memory, 528.2 TB of total storage and 10 GB Network. Pricing calculations are based on IBM watsonx.data pricing as of 7 May 2024 and Databricks published pricing for Photon as of 7 May 2024. Results are based on testing conditions and pricing as of the dates shown. Actual costs and performance can vary depending on individual client configurations and conditions. Results are derived from the Databricks SQL 8.3 benchmark and as such is not comparable to published Databricks SQL 8.3 benchmark results, as results do not comply with the Databricks SQL 8.3 benchmark specification.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
September. 2024