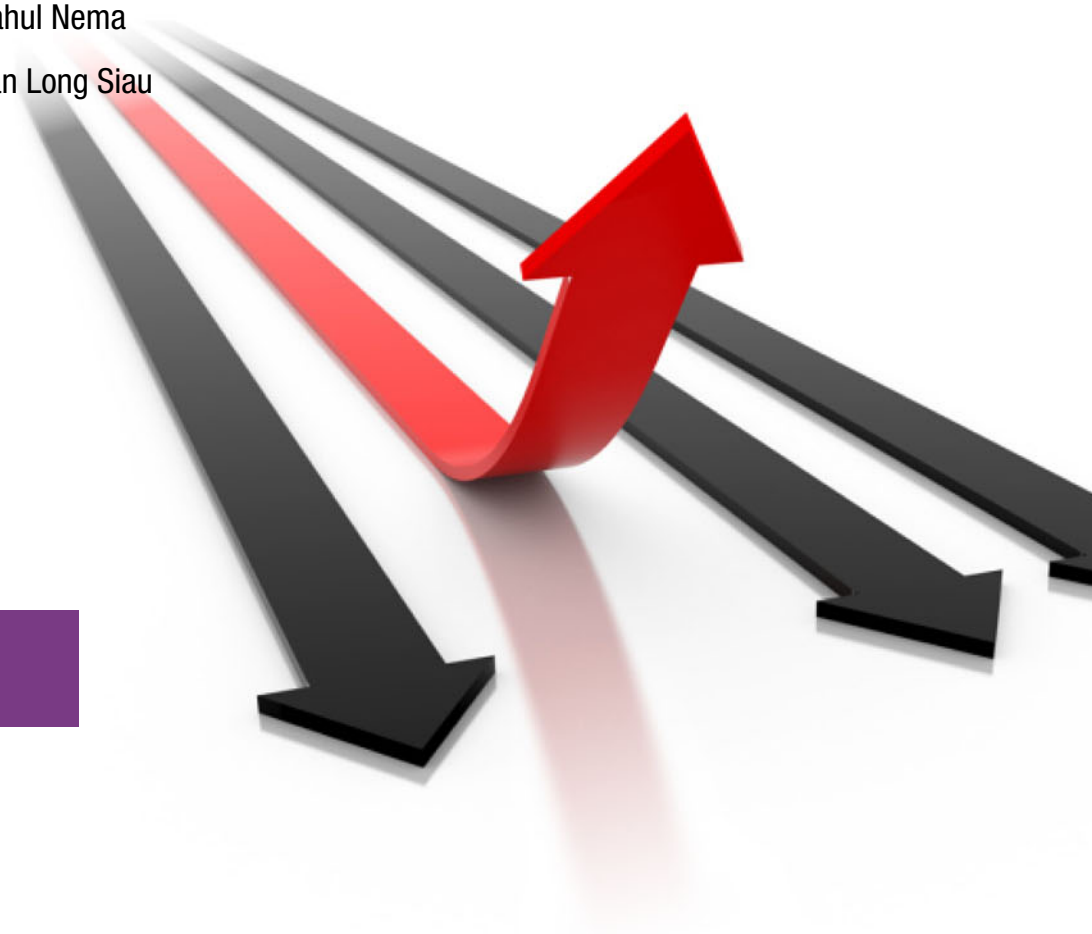
Redbooks
ibm.com/redbooks

IBM Fusion HCI as a Catalyst: Accelerating Growth in IBM watsonx

Karen Medhat
Muhammad Ajmal
Tushar Agrawal
Simon Cambridge
Gayathri Gopalakrishnan
Savitri Hunasheekatti
Suresh Jasrasaria
Hemraj Joshi
Purnanand Kumar
Youssef Largou
Jun Liu
Raydo Matthee
Rahul Nema
Tan Long Siau



Storage

IBM

Redbooks

IBM Redbooks

IBM Fusion HCI as a Catalyst

October 2025

SG24-8600-00

Note: Before using this information and the product it supports, read the information in “Notices” on page ix.

First Edition (October 2025)

© Copyright International Business Machines Corporation 2025. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices ix
Trademarks x

Preface xi
Authors xii
Now you can become a published author, too! xv
Comments welcome xv
Stay connected to IBM Redbooks xv

Chapter 1. Introduction and Overview 1
1.1 The Evolution of Data Platforms 2
1.2 IBM watsonx and the Role of IBM Fusion HCI 2
1.3 Strategic Positioning and Differentiation 3
1.4 IBM Fusion HCI and GPU Integration 3
1.5 Storage Performance and Readiness 4

Chapter 2. Planning and Preparation 5
2.1 Hardware and software prerequisites 6
2.2 Software Prerequisites 7
2.3 Licensing and subscription planning 8
2.4 Network Planning 8
2.5 Security Planning 9
2.6 Storage Planning and Sizing 10
2.7 Deployment Models 11
2.8 Deployment Responsibilities 12
2.8.1 Day 1 Readiness 12
2.8.2 Day 1 Readiness Checklist 13
2.9 Tools and Templates 14
2.9.1 Template Library 14

Chapter 3. Deployment 17
3.1 Component Dependencies and Installation Sequence 18
3.2 Critical Dependencies 18
3.3 Prerequisites Validation 19
3.4 Installation Best Practices 19
3.5 Documentation References 20
3.6 Support and Technical Assistance 20
3.7 IBM Fusion HCI Appliance Deployment 20
3.7.1 Deployment Overview 20
3.7.2 Pre-Deployment Considerations 21
3.7.3 Solution-Specific Deployment Considerations 21
3.7.4 Deployment Process Overview 21
3.7.5 Integration Points and Dependencies 22
3.7.6 Post-Deployment Validation 22
3.7.7 Solution-Specific Validation Procedures 22
3.7.8 Troubleshooting Common Issues 23
3.7.9 Next Steps 23
3.7.10 OpenShift Container Platform Configuration 23
3.7.11 Storage Infrastructure Setup 26

© Copyright IBM Corp. 2025.

3.7.12 Test Environment Overview 27
3.7.13 Storage Performance Validation Tool Installation Guide 27
3.7.14 Fusion Data Foundation Performance Profile 29
3.8 Storage Architecture for AI/ML Workloads 33
3.9 Storage Integration Architecture 33
3.10 AI/ML Performance Optimization Strategies 34
3.11 Storage Performance Validation 35
3.12 Storage Integration with watsonx Platforms 35
3.13 Monitoring and Management 36
3.14 Official Documentation References 36
3.15 Support and Technical Resources 36
3.16 Next Steps 36
3.17 Foundation Platform Deployment 37
3.17.1 Fusion Backup & Restore: 37
3.17.2 Solution Validation and Testing 40

Chapter 4. Management and Operations 41
4.1 IBM Fusion HCI Monitoring 42
4.1.1 Monitoring through the IBM Fusion HCI UI 42
4.1.2 Monitoring with External Tools (Prometheus and Grafana) 42
4.1.3 Troubleshooting and Serviceability 43
4.1.4 Call Home and Support Features 44
4.1.5 IBM Fusion HCI Call Home 44
4.1.6 IBM Fusion HCI Service Page 44
4.1.7 Upgrade and Lifecycle Management 44

Chapter 5. Fusion HCI as the Backbone for IBM watsonx: Technical Foundations of IBM watsonx 45
5.1 IBM watsonx.data: The lakehouse foundation for enterprise-grade generative AI 46
5.2 Background and technical architecture 46
5.3 watsonx.data mechanism for unstructured data processing 47
5.4 Security and governance capabilities: Opening up the AI security management and control chain 48
5.4.1 Unified platform supports multiple AI and data analysis workloads 48
5.5 Typical application scenarios 48
5.6 Atlas Cloud Services Platform 50
5.7 LLM as a Service 51
5.7.1 Model Management and Performance Optimization 51
5.7.2 Security and Multi-Tenant Isolation 51
5.7.3 Data Privacy and Sovereignty 52
5.7.4 Monitoring, Observability, and Compliance 52
5.7.5 LLM Inference Scaling and Cost Control 52
5.7.6 Integration into business processes 53
5.7.7 Data fabric as a service implementation 53
5.7.8 Data integration across heterogeneous sources 53
5.7.9 Metadata and Governance Complexity 54
5.7.10 Data Access Control and Policy Enforcement 54
5.7.11 Latency and Performance Bottlenecks 54
5.7.12 Scalability and Data Volume Management 55
5.7.13 User Enablement and Self-Service Analytics 55
5.8 Introduction: The Dawn of the Conversational Data Stack 58
5.8.1 Defining the Architectural Pillars 58
5.8.2 Data Fabric versus Data Mesh: A Clarification 59
5.8.3 Foundational Capabilities: The New Primitives of Data Interaction 60

5.8.4 Democratizing Access through Natural Language Query (NLQ) 60
5.8.5 Automating Insight with Intelligent Data Catalogs. 61
5.8.6 Ensuring Trust with AI-Driven Data Quality (DQ) 61
5.9 Platform in Practice: An Architectural Blueprint with IBM watsonx 63
5.10 Capabilities:. 63
5.11 watsonx.ai: The enterprise studio for AI builders 63
5.12 Navigating the Labyrinth: Challenges, Risks, and Ethical Imperatives 66
5.13 Technical and Operational Hurdles. 66
5.14 Generative AI-Specific Risks. 66
5.15 Ethical and Governance Mandates 67
5.16 Finally: The Future of Data as a Dialogue. 69

Chapter 6. Applications of IBM watsonx: Use Cases in AI Innovation 71
6.1 IBM watsonx in Action. 73
6.2 Use case: Unlocking Business-Critical Information in Unstructured Enterprise Data . . 73
6.2.1 Business Context and Challenge 73
6.2.2 Solution: IBM watsonx.data and lakehouse architecture 73
6.2.3 Example Use Cases 73
6.3 Case Study: Atlas Cloud Services Accelerates AI Innovation in Morocco 75
6.4 Business Context and Challenge 75
6.5 Industry-specific use cases covered by ACS Platform 76
6.5.1 Use Case 1: AI-Powered Customer Service Automation 76
6.5.2 Use Case 2: Smart HR and Recruitment Automation. 76
6.5.3 Use Case 3: Manufacturing Optimization through Predictive Maintenance 77
6.5.4 Use Case 4: Real-Time Fraud Detection 77
6.5.5 Use Case 5: Application Modernization with Developer AI Assistance. 78
6.5.6 Use Case 6: AI-Driven Supply Chain Intelligence. 78
6.6 Use Case 3: AI Governance in Public Sector with watsonx.governance 79
6.7 Public Sector - Benefits Claim Processing 79
6.8 Benefits of watsonx.gov 79
6.8.1 Resume summarization 79
6.8.2 Hiring recommendation. 80
6.8.3 Credit risk 80
6.8.4 Auto insurance risk 80
6.8.5 Human resources chatbot. 80
6.9 Use case 4: watsonx.governance in Financial Services and Rail. 81
6.10 Achieving desirable outcomes 82
6.10.1 Business Challenges in Rail 83
6.10.2 Solution: watsonx.governance 83
6.10.3 Desired Outcomes 83
6.11 Exercise: Using watsonx.data for Structured Understanding and Vector Retrieval of PDF Invoices 85
6.12 Step 1: Create a project in watsonx.data 85
6.13 Step 2: Create an AWS S3 Connection 86
6.14 Step 3: Build Ingest and Enrich Flow 86
6.15 Step 4: Eight Tasks Automatically Generated by the Flow 86
6.16 Step 5: Generate Data Assets 87
6.17 Step 6: Support for Smart Retrieval and RAG Applications 87
6.17.1 Outcomes and Benefits. 88
6.18 Activating the Intelligent Data Ecosystem: Use Cases for LLM-Enhanced Data Fabric . 88
6.18.1 Industry Spotlight: Real-World Applications and Value Creation 88
6.19 Summary. 92

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

© Copyright IBM Corp. 2025.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at https://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

DB2®
FileNet®
Granite®
HashiCorp®
IBM®
IBM Cloud®
IBM Cloud Pak®
IBM Spectrum®
IBM Watson®
IBM watsonx®
OpenPages®
Orchestrate®
Redbooks®
Redbooks (logo) ®
watsonx®
watsonx Assistant®
watsonx Code Assistant®
watsonx Orchestrate®
watsonx.ai®

Evolution, are trademarks or registered trademarks of Kenexa, an IBM Company.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Microsoft, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Ansible, Ceph, OpenShift, Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM Redbooks publication explores how IBM Fusion HCI serves as a strategic catalyst for accelerating enterprise AI transformation through the IBM watsonx platform. As organizations increasingly recognize artificial intelligence as a competitive differentiator, the underlying infrastructure becomes critical to success. This book demonstrates how the convergence of hyperconverged infrastructure, container orchestration, and AI-optimized platforms creates a foundation for scalable, secure, and high-performance AI deployments.

IBM Fusion HCI represents more than traditional infrastructure; it provides an integrated, software-defined platform that eliminates the complexity typically associated with AI workload deployment. When combined with the watsonx ecosystem, organizations can rapidly transition from data-driven to AI-driven operations while maintaining enterprise-grade governance, security, and performance.

This publication helps you design, deploy, and optimize watsonx solutions on Fusion HCI infrastructure. Whether you're implementing data lakehouse architectures, deploying large language models, or building conversational AI applications, this book provides practical guidance backed by real-world implementations and proven best practices.

The book addresses the complete lifecycle from initial planning through production operations, including comprehensive coverage of storage optimization for AI workloads, GPU integration strategies, and hybrid cloud deployment models. Through detailed use cases spanning financial services, public sector, healthcare, and manufacturing, you'll discover how organizations leverage this integrated platform to unlock insights from unstructured data, implement responsible AI governance, and accelerate innovation.

This book is designed for technical professionals, business leaders, and solution architects who are responsible for implementing enterprise AI initiatives:

- Infrastructure architects and engineers planning AI-ready infrastructure deployments
- Data engineers and AI practitioners implementing lakehouse architectures and machine learning pipelines
- IT leaders and decision makers evaluating infrastructure options for AI transformation
- System administrators deploying and managing watsonx platforms on hybrid cloud infrastructure
- Business stakeholders seeking to understand the technical foundations that enable AI-driven business outcomes

What this book covers

The book provides comprehensive guidance across four key areas:

1. Strategic Foundation: Understand how Fusion HCI and watsonx address the fundamental challenges of enterprise AI, including infrastructure complexity, data fragmentation, and governance requirements.
2. Practical Implementation: Follow detailed deployment procedures, configuration best practices, and performance optimization techniques validated in production environments.

© Copyright IBM Corp. 2025.

3. Technical Architecture: Explore the integration of hyperconverged infrastructure, container orchestration, storage optimization, and AI platforms that enables seamless scaling from proof-of-concept to enterprise-wide deployment.
4. Real-World Applications: Learn from industry-specific use cases that demonstrate measurable business outcomes, including improved decision-making speed, enhanced compliance capabilities, and accelerated innovation cycles.

This book transforms complex technical concepts into actionable guidance, enabling you to successfully implement AI infrastructure that delivers both immediate value and long-term strategic advantage.

Authors

This book was produced by a virtual team working across the world.

Karen Medhat is a Strategic Account Technical Leader and IBM Distinguished Technical Specialist. She is the youngest IBM Certified Thought Leader-Technical Specialist and currently serves as Chair of the IBM UK Technical Consultancy Group. Karen is a leader on the board of IBM Open Innovation Community and a recognized IBM Inventor.

She holds MSc and BSc degrees with honors in Engineering from Cairo University, specializing in AI and wireless sensor networks. Karen serves on the review boards of international conferences and journals in AI and wireless communication, and has authored publications on cloud, IoT, AI, wireless networks, microservices architecture, and blockchain.

Karen has extensive experience in designing application architectures and leading cross-functional teams to deliver successful client outcomes. She frequently mentors IT professionals to help them define career goals, build technical skills, and achieve professional certifications. She has co-created IBM Skills Academy courses and co-created several IBM professional certifications.

Karen is globally recognized with multiple prestigious awards as one of the top women in tech and maintains an eminent professional profile across the industry. She is the IBM EMEA Geo Lead for Business Automation product scouts, bringing client insight directly into IBM's automation products strategy. In her current role, she leads strategic client engagements across regulated industries, aligning IBM's hybrid cloud and AI capabilities with enterprise transformation goals.

Muhammad Ajmal is an Enterprise Architect based in Ireland, currently working with IBM/Red Hat. He brings over 17 years of experience in the IT field and holds an Engineering degree from NED University. His areas of expertise include Hybrid Cloud Architecture, Artificial Intelligence (AI), Infrastructure & Networking, Cybersecurity, IT Governance & Compliance Muhammad has written extensively on open-source technologies, reflecting his passion for innovation, collaboration, and driving enterprise transformation through modern, scalable solutions.

Tushar Agrawal is a Principal Account Technical Leader who creates innovative solutions and drives the adoption of Hybrid Cloud and Artificial Intelligence platforms provided by IBM. Tushar is passionate about integrating AI to augment our day-to-day lives and bring operational efficiency. Tushar focuses on ethical AI principles and collaborates with his customers to develop innovative solutions to address business transformation challenges. Tushar is an experienced enterprise architect for large-scale data processing. Tushar has proven experience designing and successfully launching 50+ complex solutions within budget for large enterprises across multiple industry verticals and segments. Tushar is a research

enthusiast and continuously works on innovation with emerging technologies to co-create Intellectual Property (200+ patent applications) for IBM. Tushar holds a BS in Computer Science & Engineering degree from MNNIT Allahabad. He also holds an MBA degree. He is pursuing a PhD in Software Engineering from North Dakota State University, Fargo, ND.

Simon Cambridge is a Fusion Technical Specialist with IBM in North America. He has over 25 years of experience in the computing industry, working with storage, data and AI solutions. He holds multiple product and industry certifications from IBM, Amazon Web Services (AWS), Microsoft Azure, Nvidia, and Snowflake. Simon's current role focuses on IBM Fusion and Ceph initiatives with Financial Services clients. Simon specializes in storage, data, and AI solutions deployed on OpenShift. He holds a Bachelor of Business degree in Business Information Systems from Massey University in New Zealand.

Gayathri Gopalakrishnan blends two and a half decades of technology leadership with a lifelong passion for learning and storytelling. As an IBM Solution Architect, she has spent over twenty-four years designing systems that power global enterprises—and as an author, she has distilled those lessons into accessible, hands-on guides for the next generation of technologists. Her portfolio includes ten IBM Redbooks and Python Programming Demystified, a practical guide adopted across learning tracks. Honored with the IBM Tech 2023 Award for innovation, Gayathri continues to mentor teams worldwide in AI, cloud, and data-driven design.A postgraduate of the Indian Institute of Science (IISc), she believes technology is most powerful when it serves people - and every solution begins with curiosity.

Savitri Hunasheekatti is a Software Architect at the IBM Software Lab in Bangalore, India. She received her Bachelor of Engineering degree from BEC Engineering College, Bagalkot. With over 20 years of experience in software development, she joined IBM in 2003. In her recent roles, she has contributed to various development projects across IBM POWER® and IBM Z platforms. Her work includes AIX on POWER kernel development and LinuxONE Bare Metal on IBM Z. She is a Plateau holder and actively engages in submitting ideas, participating in hackathons, and staying updated on the latest technologies.

Suresh Jasrasaria is currently a Principal Product Manager at IBM watsonx, where he has played a significant leadership role in bringing innovative data and AI solutions to clients. He has extensive experience in product management, especially in information technology and enterprise storage, having previously held similar roles at IBM in Data Management, Cloud Offerings, and Cloud Object Storage. His technical and marketing expertise includes cross-functional launch and delivery, product positioning, and content creation.

Hemraj Joshi works as an IBM Enterprise Systems Presales Consultant for Gulf Business Machines (GBM) in Kuwait and has over 19 years of industry experience, specializing in enterprise systems, IBM Power, AIX, IBM i, and hybrid cloud environments. His achievements include the IBM Redbooks Silver Author badge and recognition as an IBM Champion for 2025.

Purnanand Kumar is an Advisory Software Engineer at IBM in India. He has over 8 years of experience in software engineering, specializing in storage virtualization, hybrid cloud infrastructure, and containerized solutions. He holds a Bachelor's degree in Computer Engineering from Poornima University. His areas of expertise include hyper-converged infrastructure (HCI), OpenShift, Golang-based backend development, and enterprise storage systems. He has written extensively on IBM storage modernization and disaster recovery architectures, contributing as a co-author to the IBM Redpaper "IBM Fusion HCI: Metro Disaster Recovery Use Case".

Youssef Largou is the Founder and CEO of PowerM and RKube.IO in Morocco. He has more than 20 years of experience in information technology, leading enterprise modernization, hybrid cloud transformations, and AI-driven infrastructure projects. He holds a degree in Computer Science and has built deep expertise in hybrid cloud, IBM Power Systems, quantum computing, software-defined infrastructure, and enterprise AI integration. His areas of expertise include IBM Cloud, LinuxONE, data modernization, and application migration strategies. He has written extensively as an IBM Redbooks and Redpaper author, contributing to works such as "IBM Virtual Machine Recovery Manager for IBM Power Cookbook" and "Practical Migration from x86 to Linux on IBM Z", among others.

Jun Liu serves as the architect for IBM watsonx.data, primarily focusing on the authentication and authorization architecture across various components of the product. His current emphasis is on enhancing data security and governance within the platform. Before working on the watsonx project, Jun Liu was the Technical Architect of Data Virtualization Console. He leads the Data Virtualization Console development and new feature integration team. Before working on the Data Virtualization project, he worked for various projects on Db2® query monitoring and optimization. He has been with IBM since 2005.

Raydo Matthee is a South African technology leader and Solutions Architect at Skunkworks (Pty) Ltd. He has over 15 years of experience in cloud computing, IT security, and enterprise integration. He holds a degree in Information Technology and has built expertise across hybrid cloud, API development, application integration, and secure enterprise architecture. His areas of expertise include IBM Power Virtual Server, cloud security, and high-availability infrastructure solutions. He has written extensively as a co-author of IBM Redbooks publications, including “Power Virtual Server Private Cloud” and “Introduction to Deployable Architectures in IBM Power Virtual Server,” highlighting his contributions to secure and scalable hybrid cloud design.

Rahul Nema is an STSM (Senior Technical Staff Member) and Lead Architect at IBM in India. He has over 18 years of experience in the field of hybrid cloud and enterprise storage solutions. He holds a Bachelor of Technology degree from the National Institute Of Technology Silchar. His areas of expertise include hybrid cloud management, container-native storage, OpenShift integration, cloud automation, and enterprise DevOps. He has written extensively on IBM hybrid cloud offerings and storage modernization, authoring and contributing to multiple IBM Redbooks, including “Multi-Factor Authentication Using IBM Security,” “IBM Cloud Private Application Developer’s Guide,” and technical publications on IBM Spectrum Fusion.

Tan Long Siau is a Storage Software Technical Specialist at IBM ASEAN, bringing over a decade of experience in the storage industry. His expertise spans file and object storage, as well as containerized storage solutions. A seasoned technologist, he possesses knowledge in big data, artificial intelligence (AI), Internet of Things (IoT), and high-performance computing architectures.

Thanks to the following people for their contributions to this project:

Robert Haimowitz and Patrik Hysky
IBM Redbooks, Poughkeepsie Center

Jamie Roszel
IBM RTP

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM® Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on LinkedIn:

  https://www.linkedin.com/groups/2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/subscribe

- Stay current on recent Redbooks publications with RSS Feeds:

  https://www.redbooks.ibm.com/rss.html

1

Introduction and Overview

This IBM Redbooks publication explores how IBM Fusion HCI acts as a strategic enabler for the watsonx® platform. Unlike traditional infrastructure that simply provides resources, IBM Fusion HCI optimizes conditions for AI and data workloads while remaining unchanged itself. It provides the optimal conditions for transformation while simplifying the complex interactions between storage, compute, networking, and AI acceleration.

It provides a detailed roadmap for planning, deploying, and scaling watsonx solutions on Fusion HCI, with a focus on the following areas:

- Prepare infrastructure and select deployment models
- Plan licensing and subscription planning
- Plan network, security, and storage architecture
- Prepare integration with IBM Storage Scale and CAS for real-time Artificial Intelligence (AI)
- Establish role alignment between customer and IBM teams

By integrating with watsonx, IBM Fusion HCI enables organizations to achieve the following outcomes:

- Accelerate time to value for AI and analytics initiatives
- Simplify infrastructure and reduce operational complexity
- Scale seamlessly across hybrid environments
- Maintain consistent governance and security across data and AI assets

This chapter sets the foundation for understanding how IBM Fusion HCI supports the watsonx platform-not just as infrastructure, but as a catalyst for enterprise transformation.

© Copyright IBM Corp. 2025.

1.1 The Evolution of Data Platforms

The evolution of data platforms has been a cornerstone of enterprise innovation. From the early days of database management systems to the rise of data warehouses, data lakes, and now the data lakehouse, each generation has aimed to unlock more value from data. Today's lakehouse architecture blends the scalability of data lakes with the performance and governance of data warehouses-enabling organizations to support both Business Intelligence (BI) and Artificial Intelligence (AI) workloads.

Traditional infrastructure approaches often created silos between storage, compute, and networking layers, resulting in complex deployment cycles, lengthy provisioning times, and operational overhead that slowed innovation. Organizations frequently struggled with:

- Multi-vendor integration complexity
- Inconsistent performance across workloads
- Resource underutilization and scaling inefficiencies
- Extended time-to-value for AI and analytics initiatives

IBM has consistently led this transformation through integrated solutions that eliminate these traditional barriers. The shift toward open data formats, hybrid cloud environments, and AI-native infrastructure reflects a broader need for agility, interoperability, and enterprise-grade governance.

1.2 IBM watsonx and the Role of IBM Fusion HCI

At the heart of the IBM modern data and AI strategy is the watsonx platform, anchored by IBM watsonx®.data, a lakehouse built for enterprise AI and analytics. It provides a unified, open, and hybrid architecture that simplifies data management across structured and unstructured sources. With integrated data fabric capabilities such as governance, quality, and lineage, watsonx.data supports the full lifecycle of data for AI, generative AI, and BI applications.

The watsonx ecosystem includes:

- watsonx.ai®: Provides tools to AI developers and data scientists to build, test, train, and deploy AI models. Tools like AgentOps allow developers to track and analyze all watsonx.ai model interactions automatically. IBM Model Gateway provides a single OpenAI-compatible LLM-as-a-service API, which proxies the routing of requests to LLM providers and currently offers features such as load balancing.
- watsonx.data: The open lakehouse architecture provided by watsonx.data is central to IBM's AI strategy. It allows enterprises to ingest, store, manage, and enrich data at scale. watsonx.data is designed to overcome fragmentation and inefficiencies in AI scalability by providing:
  - separation of storage and compute
  - support for open formats and
  - built-in governance/security tools
- watsonx.governance is an integrated AI lifecycle management to help enterprises ensure compliance, track model accuracy and bias for responsible AI and maintain governance across the business.
- watsonx Orchestrate® acts as an orchestrator for AI agents and automations by enabling composable API-driven integrations to automate workflows required by enterprises for operationalizing AI at scale.

- watsonx BI is a generative AI-powered business intelligence (BI) platform for delivering contextual and personalized insights on complex data using natural language. It intends to democratize business intelligence by simplifying data interactions and delivering actionable insights.

1.3 Strategic Positioning and Differentiation

AI Optimized Stack: IBM Fusion HCI provides an "AI Ready" platform designed to support demands of generative AI, machine learning and data analytics not only in corporate data centers but also at the edge. The infrastructure provides scalability and resiliency by leveraging CPUs and GPUs, high-speed networking (> 25 Gbps) and NVMe storage.

Cost Optimization: IBM Fusion HCI (Hyperconverged Infrastructure) and watsonx are designed to accelerate AI adoption by enterprises by bringing AI closer to the data for improved performance, security and cost-efficiency. IBM Fusion HCI serves as a foundation for AI and Hybrid Cloud by integrating data, hybrid cloud, and governance with a focus on energy efficiency.

Unified Data: IBM Fusion HCI uses watsonx.data as a bridge to unlock valuable insights from unstructured and underutilized data. Enterprises can use these insights for trustworthy and high-impact outcomes in a governed and scalable fashion.

Open Standards: IBM watsonx reduces vendor lock-in through adoption of open standards, API-driven platforms that support "bring your own model" to support a wide range of enterprise AI use cases.

Hybrid Cloud Flexibility: IBM Fusion HCI is built on hybrid architecture that supports on-premises, multicloud and edge deployment making it a strategic choice for dealing with sensitive data and compliance requirements and latency-sensitive applications.

Red Hat OpenShift: IBM Fusion HCI provides a unified container orchestration platform using Red Hat OpenShift. This enables DevOps for multi-cluster management for rapid deployments. The automation enables scheduling of GPU-accelerated workloads, reduces the time-to-market and allows proof-of-concepts to be done in less than a week.

Data Services: IBM Fusion HCI provides backup, restore, disaster recovery and workflow tailored for AI applications. These services ensure that AI workloads have persistent secure data access.

1.4 IBM Fusion HCI and GPU Integration

IBM Fusion HCI provides a robust integration with NVIDIA and AMD GPUs to facilitate compute-intensive workloads - including generative AI, machine learning and data analytics.

Dedicated GPU Server Nodes: IBM Fusion HCI uses optional GPU-accelerated server nodes within rack-mounted architecture. Each GPU server (for example, 9155-G03) can accommodate 1-8 PCIe GPU cards of the same type per server. Supported options include NVIDIA L40S, NVIDIA H100 NVL, and AMD Mi210 GPUs.

Rack Level Scalability: Each Fusion HCI rack supports up to four GPU server nodes, allowing for significant GPU density and processing capability. The GPU server nodes can be added as needed, offering flexibility to scale as workloads grow over time, providing cost-effective scalability.

1.5 Storage Performance and Readiness

Storage performance is a foundational requirement for AI and analytics workloads. IBM Fusion HCI also integrates with IBM Storage Scale and Content-Aware Storage (CAS) to enable real-time, agentic AI solutions. Storage Scale provides high-performance, distributed file systems with NVMe and GPUDirect support, while CAS delivers metadata intelligence and semantic indexing-critical for retrieval augmented generation (RAG) and real-time inference pipelines. IBM Fusion HCI has been validated using IBM's storage performance tools to ensure it exceeds the throughput and latency requirements for IBM Cloud® Paks. Testing across multiple performance profiles (Lean, Balanced, Performance) demonstrated:

- Sequential write throughput exceeding baseline requirements by 11-18x
- Random write throughput exceeding requirements by 3x
- Dynamic tuning using Data Foundation performance profiles for workload optimization

This level of performance ensures that watsonx workloads-especially those involving large-scale model training and inference-can operate without bottlenecks.

2

Planning and Preparation

Planning is essential for a successful deployment. This chapter focuses on ensuring readiness for implementing IBM watsonx.ai, watsonx.data, and watsonx.gov solutions on IBM Fusion Hybrid Cloud Infrastructure (HCI). It is vital to address hardware, software, security, networking, and storage planning early to minimize risks and accelerate value realization.

IBM Fusion HCI brings together compute, storage, and networking in a tightly integrated, AI-optimized infrastructure. When paired with watsonx platforms, it becomes a powerful foundation for scalable, secure, and high-performing AI and data workloads.

This chapter will guide you through:

- Preparing infrastructure and understanding prerequisites
- Choosing appropriate deployment models
- Managing licensing and support
- Planning for network, security, and storage
- Aligning responsibilities between customer and IBM team

© Copyright IBM Corp. 2025.

2.1 Hardware and software prerequisites

A robust and scalable infrastructure is a nonnegotiable prerequisite for the successful implementation of IBM watsonx platforms on Fusion HCI. The goal of this section is to guide infrastructure teams in validating both hardware and software components ahead of deployment.

Here are hardware requirements:

- IBM Fusion HCI Nodes: These purpose-built, certified server nodes form the backbone of the deployment. They're optimized for hybrid workloads and tightly integrated with IBM's AI stack. These nodes provide high-density compute resources and support seamless scaling.
- IBM Fusion HCI node specifications:
  - Minimum 3-node cluster configuration for high availability
  - Enterprise-grade x86_64 processors with virtualization extensions
  - Sufficient memory allocation for Red Hat OpenShift control plane and worker nodes
  - High-performance storage (NVMe recommended) for optimal I/O performance
  - Redundant network interfaces (10 GbE minimum, 25 GbE recommended)
  - Hardware RAID controller support for system drives
- GPU Support: If you're deploying AI/ML workloads, GPU acceleration is critical. IBM Fusion HCI supports AMD MI210, NVIDIA L40S, or NVIDIA H100 NVL GPUs, which are ideal for deep learning model training, large language model (LLM) inference, and AI pipeline acceleration. GPUs ensure the infrastructure can handle complex data science workflows without performance bottlenecks.
- High-Speed Networking: Modern AI and data workloads require high-volume data transfers across nodes and services. Use at least 25 Gbps network interface cards (NICs) to support low-latency east-west (intra-cluster) and north-south (client/server) traffic. Consider redundant NICs for failover.
- Redundant Power and Cooling: Because the environment supports critical workloads, your data center needs dual power sources, UPS systems, and efficient cooling. This reduces the risk of downtime and hardware degradation.
- Network Infrastructure:
  - Managed switches with VLAN support and redundancy capabilities
  - Dedicated management network separate from production traffic
  - High-bandwidth connectivity between nodes (25 GbE or higher for production)
  - Internet connectivity for Container Registry access and updates
- Storage Infrastructure:
  - Dedicated storage network (separate from management and application traffic)
  - Compatible storage arrays or distributed storage capabilities
  - Backup and disaster recovery (DR) infrastructure

Note: Specific hardware models, CPU cores, memory, and storage capacity requirements should be validated with IBM sizing tools and current product documentation.

2.2 Software Prerequisites

- Red Hat OpenShift: This container orchestration layer is required to deploy and manage containerized applications. All IBM watsonx workloads, including those deployed through CP4D, run on Red Hat OpenShift. You should verify version compatibility with IBM Fusion HCI before deployment.
- IBM Cloud Pak® for Data (CP4D): CP4D is the foundation for deploying watsonx.data and watsonx.ai. It unifies data governance, analytics, and AI services. Make sure entitlement and versioning checks are completed in advance.
- IBM Fusion HCI Control Plane: This component manages the Fusion HCI ecosystem, offering tools for deployment automation, monitoring, and scaling. It ensures consistent operations across clusters.
- Validated OS and Firmware Versions: All systems must run operating systems listed in the IBM Compatibility Matrix. You need to validate BIOS, BMC, and firmware versions for compatibility with the IBM HCI stack and supported AI/ML components. Incompatible versions can cause unpredictable behavior or lead to unsupported installation states.
- Base Operating System (OS):
  - Red Hat Enterprise Linux (RHEL) - supported versions as per IBM compatibility matrix
  - Proper subscription and support entitlements
  - System patches and security updates applied
- Container Platform:
  - Red Hat OpenShift Container Platform - version compatibility verified with watsonx components
  - Valid Red Hat OpenShift subscriptions for all planned nodes
  - Container Registry access (internal or external)
- Cloud Pak for Data Foundation: Cloud Pak for Data (CP4D) serves as the foundational data and AI platform that integrates with IBM Fusion HCI to deliver comprehensive data management and analytics capabilities.
- CP4D Integration Considerations:
  - CP4D licensing aligned with watsonx platform requirements
  - Storage class definitions compatible with watsonx.data and watsonx.ai workloads
  - Integration with existing identity management systems
  - Network policy configurations for secure multi-tenant environments
- Additional Software Components:
  - IBM Storage Ceph or compatible storage software
  - Monitoring and logging solutions
  - Backup and disaster recovery software
  - Security and compliance tools

You should conduct infrastructure readiness assessments as a pre-installation milestone to confirm that all prerequisites are met. This prevents rework, accelerates installation, and prepares the environment for long-term scalability. Specific version numbers and compatibility matrices should be verified with current IBM product documentation and support teams.

2.3 Licensing and subscription planning

A successful implementation must account for licenses and support subscriptions early in the project lifecycle. Proper planning in this area helps prevent deployment delays, ensures compliance, and provides ongoing access to technical support and software updates.

Key Considerations:

- watsonx VPC Licensing: Licensing for watsonx platforms is typically based on Virtual Processor Cores (VPCs), which allows for elastic scaling based on compute usage. This flexible model supports dynamic workloads and avoids over-licensing. It is important to forecast expected usage patterns early and align VPC purchases with anticipated growth.
- Red Hat OpenShift Subscriptions: Red Hat OpenShift requires a subscription model that includes access to core packs, technical support, and ongoing updates. Subscription tiers vary depending on deployment size, number of worker nodes, and need for premium support. Accurate node counts and resource planning are essential to avoid under- or over-subscription.
- IBM 9155 Expert Care: This service offering is designed to support the IBM 9155 server family, which includes Fusion HCI nodes. Expert Care provides hardware and software lifecycle support, including next-business-day hardware replacement, proactive firmware and software updates, and direct access to IBM Support. This is critical for enterprises requiring high availability and minimal downtime.

Additional Guidance:

- Collaborate with IBM or Red Hat sales and licensing specialists to validate entitlements and ensure you are selecting the correct licensing model.
- Plan for renewal cycles and budget allocations for multi-year support.

Note: Use the IBM License Metric Tool or similar software asset management tools to monitor usage, maintain audit compliance, and avoid licensing shortfalls or overages. These tools also help consolidate reporting across hybrid deployments involving multiple IBM offerings.

2.4 Network Planning

Network design is critical for availability, performance, and security. An improperly designed network can become a bottleneck or single point of failure, directly impacting application uptime and user experience. In AI and data-intensive environments, where massive volumes of information are processed and moved across systems, a resilient and high-throughput network infrastructure is mandatory.

Key Elements:

- IP address Planning: Allocate sufficient IP address ranges to accommodate various components such as Red Hat OpenShift control and worker nodes, storage systems, application pods, and management consoles. It's important to plan for growth and reserve addresses for future expansion. A good practice is to segment traffic using subnets for better isolation and security.
- Switching Infrastructure: Use enterprise-grade switches that support high availability, redundancy, and VLAN configurations. Redundant paths and dual-homed switches (connected to separate power and uplink paths) help avoid downtime during hardware or

link failures. VLAN tagging enables logical segmentation of traffic-for example, separating management, application, storage, and user traffic.

- SAN/AFM Design: Storage Area Network (SAN) planning must include logical unit number (LUN) mapping, zone policies, and multi-path configurations for failover. For Active File Management (AFM) scenarios, plan for replication paths, source/target mappings, and throughput requirements based on data movement patterns. Using multi-path I/O (MPIO) helps achieve both performance and redundancy.
- Firewall Rules: Pre-define ingress and egress rules for each layer of the environment. This includes allowing secure communication between Red Hat OpenShift components, storage, application services, and external APIs. Also, consider creating a centralized document listing all required ports, protocols, and allowed IP ranges to simplify firewall configuration.

Optional: A sample worksheet or checklist template-covering VLAN assignments, IP subnets, port mapping, and link redundancy-can be included in the appendix or provided in the shared OneDrive project folder to help accelerate design and validation.

2.5 Security Planning

Security needs to be woven into the fabric of the deployment to protect sensitive data, maintain system integrity, and ensure compliance with industry and government regulations. From identity and access controls to encryption and secrets management, every layer of the infrastructure must be designed with security in mind.

Components:

- Role-based access control (RBAC): RBAC enables administrators to define access policies based on roles assigned to users or service accounts. This minimizes the attack surface by ensuring that users have only the permissions they need to perform their tasks. For example, an AI model developer may have access to training environments but be restricted from making configuration changes to the storage backend. Integration with enterprise identity providers like LDAP, Active Directory, or SAML-based SSO platforms is recommended to enforce consistent access policies.
- Encryption: All data must be protected both at rest and in transit. Transport Layer Security (TLS) should be enforced for all communications between system components, APIs, and external interfaces. For data at rest, use FIPS 140-2 compliant encryption algorithms to meet federal and enterprise-grade security standards. Additionally, verify that storage backends such as IBM Storage Ceph are configured to enable native encryption.
- Compliance Alignment: Different industries are subject to different regulatory frameworks such as GDPR (for personal data privacy in the EU), HIPAA (for healthcare data in the US), and others like ISO 27001 and SOC 2. IBM Fusion HCI and watsonx components support implementation of these compliance controls through features like audit logging, access control, encryption, and secure software supply chains. A compliance checklist should be maintained and reviewed regularly as part of governance.
- Secrets Management: Secure storage and access to secrets such as API keys, passwords, certificates, and encryption keys are critical. IBM Fusion HCI can be integrated with external secrets managers such as IBM HashiCorp® Vault, which provides

role-based access, dynamic secrets, and automated expiration. This ensures credentials are not hardcoded or stored insecurely within deployment manifests or scripts.

Security planning must be an ongoing activity-not a one-time task-integrated into every stage of the solution lifecycle: design, deployment, operations, and retirement.

2.6 Storage Planning and Sizing

Storage infrastructure must be scalable, redundant, and high-performance to support the large volumes of data processed in AI and analytics workloads. Improper storage planning can result in latency, reduced performance, or data loss, especially under peak loads or during failure recovery scenarios.

Planning Areas:

- IBM Storage Ceph: Ceph is a highly scalable, software-defined storage solution capable of handling block, file, and object storage within a unified platform. It is ideal for cloud-native and containerized environments. With its ability to scale horizontally and support erasure coding, Ceph delivers high durability and efficiency for both structured and unstructured data. When integrated with Red Hat OpenShift, Ceph supports dynamic provisioning of persistent volumes for AI workloads deployed via watsonx.data and watsonx.ai.
- AFM Configuration: Active File Management (AFM) allows for efficient edge caching and asynchronous replication of data between locations. AFM is useful when extending storage across multiple sites or when ingesting data from edge devices. Configuration includes determining cache policies, data refresh intervals, and consistency models. It also allows organizations to build global data spaces while reducing WAN dependency.
- Sizing Guidelines: Accurately sizing the storage environment is key to ensuring optimal performance and avoiding bottlenecks. Use the following parameters:
  - Daily Ingestion Volume: Estimate the amount of data ingested daily from various sources like sensors, applications, and pipelines. This impacts raw capacity and performance needs.
  - Read/write IOPS Requirements: Determine the expected number of input/output operations per second to ensure the underlying storage system can handle the load, especially for high-concurrency environments like AI model training.
  - Retention Period: Define how long data needs to be retained online before it can be archived or deleted. This influences total capacity and tiering strategies.
  - Growth Rate (TB/month): Forecast future storage needs based on current trends and planned expansions. Over-provisioning slightly ensures buffer capacity for unexpected spikes or data bursts.

It is strongly recommended to use IBM storage sizing and capacity planning templates to translate workload characteristics into infrastructure requirements. These tools help model future needs, compare storage tiers, and evaluate cost-performance tradeoffs for both production and development environments.

2.7 Deployment Models

IBM Fusion HCI offers flexible deployment options based on your infrastructure and compliance needs.

Deployment Type	Characteristics	Best Suited For	Considerations
On-premises	► Fully local deployment with full control ► Complete infrastructure ownership and control ► Data remains within organizational boundaries ► Direct hardware and software management ► Full customization capabilities	► Government, finance ► Organizations with strict data sovereignty requirements ► Environments with consistent, predictable workloads ► Companies with existing data center infrastructure ► Industries with specific regulatory compliance needs ►	► Higher upfront capital investment ► Internal expertise is required for day-to-day operations ► Responsibility for hardware lifecycle management ► Complete control over security and compliance implementation
Hybrid	► Mix of on-prem and public cloud ► Combination of on-premises and cloud resources ► Workload distribution based on requirements ► Data and application portability between environments ► Flexible resource scaling capabilities ►	► AI training at scale ► Organizations with variable workloads ► Companies requiring data locality for some workloads ► Businesses seeking gradual cloud adoption ► Environments requiring disaster recovery capabilities ►	► Network connectivity requirements between environments ► Data synchronization and consistency management ► Security policies spanning multiple environments ► Complexity in management and operations ► ►

Deployment Type	Characteristics	Best Suited For	Considerations
Cloud-based deployment	► Fully on IBM Cloud or partner CSPs ► Leveraging public cloud infrastructure providers ► Reduced infrastructure management overhead ► Elastic scaling based on demand ► Integration with cloud-native services ►	► Fast rollouts, dev/test ► Organizations with highly variable workloads ► Companies seeking to minimize infrastructure investment ► Businesses requiring rapid scaling capabilities ► Teams with limited infrastructure expertise ►	► Data egress costs and bandwidth requirements ► Cloud provider lock-in consideration s ► Compliance with data residency requirements ► Network latency for on-premises data sources ►

Example 2-1 IBM Fusion HCI deployment options

Note: Specific cloud provider integrations and supported deployment patterns should be validated with current IBM partnership and certification status.

2.8 Deployment Responsibilities

A clear understanding of responsibilities between you and IBM ensures smooth deployment and ongoing operations. The following tables show who is responsible for each task before installation begins.

2.8.1 Day 1 Readiness

Task	You	IBM
Rack and Stack Hardware	x	
IP address Allocation	x	
Firmware and BIOS Setup		x
Install Red Hat OpenShift & Fusion HCI		x

Task	You	IBM
Define Security Policies	x	x
Validate Infrastructure	x	x
Data Migration Planning	x	

2.8.2 Day 1 Readiness Checklist

- Infrastructure Readiness:
  - Data center space, power, and cooling capacity verified
  - Network infrastructure installed and tested
  - Internet connectivity established for updates and support
  - Backup and DR infrastructure in place
- Access and Security:
  - Administrative credentials and access procedures defined
  - Identity management system integration planned
  - Firewall rules and network policies configured
  - Security compliance requirements documented
- Organizational Readiness:
  - Project team roles and responsibilities defined
  - Change management processes established
  - Training plans for operational teams created
  - Support escalation procedures documented
- Technical Prerequisites:
  - Hardware compatibility verified
  - Software licensing confirmed
  - Integration requirements documented
  - Monitoring and alerting systems planned

IBM Support Model

- Technical Support:
  - IBM 9155 Expert Care provides comprehensive hardware and software support
  - Single point of contact for integrated solution support
  - Proactive monitoring and maintenance capabilities
  - Access to IBM technical specialists and documentation
- Implementation Support:
  - IBM services team for deployment guidance
  - Technical account management for ongoing relationship
  - Access to IBM knowledge base and community resources
  - Regular health checks and optimization reviews

Note: Specific support offerings and response times should be confirmed with IBM support contracts and service level agreements.

2.9 Tools and Templates

IBM and Red Hat provide a variety of tools to support the planning and deployment process.

Tools Include:

- IBM Capacity Calculator:
  - Forecast infrastructure requirements based on workload characteristics.
  - Model different deployment scenarios and growth patterns.
  - Compare cost-performance tradeoffs across configurations.
  - Generate detailed hardware specifications and recommendations.
- Storage Planning Tools:
  - Estimate Ceph sizing based on workload profiles and performance requirements.
  - Calculate storage efficiency with erasure coding and replication.
  - Plan for data growth and retention requirements.
  - Model backup and DR storage needs.
  - Security Checklist: Align with compliance frameworks.
- Day 1/2 Checklist:
  - Day 1 Deployment Checklist:
    - Pre-deployment validation steps
    - Configuration verification procedures
    - Initial testing and validation scripts
    - Go-live readiness assessment
  - Day 2 Operations Checklist:
    - Ongoing monitoring and maintenance procedures
    - Performance tuning guidelines
    - Backup and recovery validation
    - User onboarding and training procedures
- Network Worksheet: Plan IP ranges, VLANs, DNS, and gateway settings.

2.9.1 Template Library

Configuration Templates:

- Red Hat OpenShift cluster configuration examples
- Storage class definitions for different workload types
- Network policy templates for security segmentation
- Monitoring and alerting configuration samples

Documentation Templates:

- Architecture design documents
- Operational runbooks
- Disaster recovery procedures
- User access management procedures

Access and Availability:

These tools and templates are typically available through:

- IBM Customer Success Portal
- Red Hat Customer Portal
- IBM Business Partner resources

- Project-specific OneDrive or collaboration platforms

Note: Tool availability and access methods should be confirmed with IBM account teams and may require valid support contracts or Business Partner status.

3

Deployment

You install and configure IBM Fusion HCI with watsonx platforms through a multi-layered deployment that integrates infrastructure, platform, and application components. This solution combines IBM hyperconverged infrastructure with Red Hat OpenShift container orchestration and IBM AI and data management platforms to create a unified environment for enterprise AI and analytics workloads.

Understanding how these components interact is essential for a successful deployment. IBM Fusion HCI provides the foundational compute and storage infrastructure. Red Hat OpenShift delivers the container orchestration layer. Cloud Pak for Data serves as the core data and AI platform, although watsonx.ai and watsonx.data extend specialized AI and data lakehouse capabilities.

© Copyright IBM Corp. 2025.

3.1 Component Dependencies and Installation Sequence

Follow a specific installation sequence to ensure proper integration and avoid configuration conflicts. Each component depends on the successful installation and configuration of its predecessors.

- Phase 1: Infrastructure Foundation
  - Deploy the IBM Fusion HCI appliance to provide compute, storage, and networking infrastructure.
  - Establish the Red Hat OpenShift Container Platform.
  - Configure integrated storage capabilities.
  - Validate Red Hat OpenShift cluster readiness and node availability.
  - Configure storage classes for persistent workloads.
  - Establish operator lifecycle management.
- Phase 2: Storage and Data Services
  - Configure IBM Storage Ceph for object and block storage.
  - Establish storage performance profiles.
  - Implement Active File Management (AFM) for distributed data access.
  - Deploy Multicloud Object Gateway (MCG) for S3-compatible access.
  - Configure data virtualization layers.
  - Establish data governance frameworks.
- Phase 3: Platform Services
  - Install Cloud Pak for Data core services.
  - Configure identity management and security policies.
  - Establish the foundation for watsonx platform integration.
  - Deploy watsonx.ai for AI model development and deployment.
  - Configure watsonx.data for lakehouse capabilities.
  - Establish platform interconnectivity and resource sharing.
- Phase 4: Validation and Optimization
  - Conduct end-to-end functionality testing.
  - Validate storage performance against workload requirements.
  - Optimize configuration for production workloads.

3.2 Critical Dependencies

- Storage Dependencies:
  - Red Hat OpenShift requires properly configured storage classes before platform deployment.
  - watsonx components require high-performance storage for AI workloads.
  - Data services depend on object storage availability for metadata and artifacts.
- Network Dependencies:
  - All components require consistent network policies and connectivity.

  - Storage replication and data movement require dedicated high-bandwidth connections.
  - Establish security policies before deploying platform components.
- Security Dependencies:
  - Configure identity management integration before enabling user-facing services.
  - Establish certificate management and encryption early in the deployment.
  - Align role-based access control (RBAC) policies across all platform layers.

3.3 Prerequisites Validation

Before you begin the installation sequence, validate completion of the planning requirements outlined in Chapter 2.

- Infrastructure Readiness:
  - Verify hardware specifications against IBM sizing guidelines.
  - Configure network infrastructure according to planning worksheets.
  - Validate storage capacity and performance against workload requirements.
- Software Licensing:
  - Obtain and validate all required licenses.
  - Activate support contracts for all components.
  - Confirm subscription entitlements for multi-year deployments.
- Organizational Readiness:
  - Ensure technical teams are trained on deployment procedures.
  - Establish change management processes.
  - Document and test support escalation procedures.

For detailed validation guidance, see Chapter 2, “Planning and Preparation” on page 5.

3.4 Installation Best Practices

- Sequential Deployment: Always follow the prescribed installation sequence to avoid dependency conflicts. Each phase builds on the previous components and may fail if prerequisites are not properly established.
- Validation at Each Phase: Validate each major component installation before proceeding to the next phase. This minimizes troubleshooting complexity and ensures proper configuration.
- Documentation and Change Tracking: Maintain detailed logs of all configuration changes and customizations. This documentation supports troubleshooting, support interactions, and future maintenance.
- Backup and Recovery Planning: Establish backup procedures for configuration data and critical settings before making significant changes. This enables quick recovery if issues arise during deployment.

3.5 Documentation References

This chapter provides solution-specific guidance and integration considerations. For detailed installation procedures, refer to the official IBM product documentation:

- IBM Fusion HCI Installation Guide
- Red Hat OpenShift Container Platform Installation
- Cloud Pak for Data Installation
- watsonx Platform Installation
- IBM Storage Ceph Configuration

Note: Always verify the latest documentation versions and compatibility matrices before beginning installation.

3.6 Support and Technical Assistance

IBM provides comprehensive support throughout the installation process:

- IBM Expert Care: Proactive support for hardware and software components
- Technical Account Management: Dedicated support for complex deployments
- IBM Services: Professional services for implementation guidance
- Partner Network: Certified partners with specialized expertise

For installation support, contact your IBM technical account manager or open a support case through the IBM Support Portal.

The following sections in this chapter provide detailed guidance for each installation phase, including:

- Solution-specific configuration considerations
- Integration best practices and lessons learned
- Performance optimization techniques
- Troubleshooting common issues

Each section includes references to official documentation while focusing on the unique aspects of deploying these components as an integrated solution.

3.7 IBM Fusion HCI Appliance Deployment

3.7.1 Deployment Overview

IBM Fusion HCI appliance deployment establishes the foundational infrastructure for the watsonx solution stack. This hyperconverged infrastructure provides integrated compute, storage, and networking capabilities optimized for AI and data-intensive workloads. IBM certified technicians or partners typically perform the deployment process to ensure proper hardware configuration and software foundation.

This deployment creates a fully integrated environment where Red Hat OpenShift Container Platform operates natively with IBM storage technologies. This configuration provides the container orchestration layer required for watsonx platforms while maintaining enterprise-grade performance and reliability.

3.7.2 Pre-Deployment Considerations

Hardware Validation: Before deployment begins, you must verify that all hardware components meet the specifications outlined in Chapter 2. Confirm CPU capabilities, memory configurations, storage devices, and network interface specifications. Any hardware discrepancies discovered during deployment can cause delays and may require component replacement.

Network Configuration Planning: Prepare detailed network configuration information, including IP address ranges, VLAN assignments, DNS server configurations, and NTP server details. The deployment process requires this information to configure management interfaces and establish cluster connectivity.

Environmental Requirements: Ensure that data center environmental conditions meet IBM specifications for power, cooling, and physical space. The deployment team validates these conditions before installation begins.

3.7.3 Solution-Specific Deployment Considerations

Storage Configuration for AI Workloads: During deployment, optimize storage configurations for high-throughput, low-latency requirements typical of AI and machine learning workloads. Configure NVMe storage devices for performance and establish storage policies that support both block and object storage access patterns.

Red Hat OpenShift Cluster Sizing: Configure the Red Hat OpenShift cluster to meet the resource requirements of watsonx platforms. Allocate sufficient memory for AI model training, CPU cores for parallel processing, and appropriate storage classes for different workload types.

Network Performance Optimization: Configure network interfaces and bonding policies to support high-bandwidth requirements for data movement between storage, compute, and external data sources. Establish dedicated networks for storage traffic and management operations.

3.7.4 Deployment Process Overview

- Phase 1: Hardware Installation and Configuration: IBM certified technicians perform physical hardware installation, including server rack mounting, cabling, and initial power-on procedures. The deployment includes automated BIOS configuration, RAID setup for system drives, and initial network interface configuration. Stage 1 of the Fusion HCI installation is performed onsite. Key settings such as the NTP server address, DNS IP, Gateway IP and Provisioner node CIDR address are setup with a GUI wizard. IBM works with customer to complete the Fusion HCI installation worksheet. This provides the onsite technician the required inputs to the installation wizard.
- Phase 2: Red Hat OpenShift Platform Installation: In the Fusion HCI stage 2 installation wizard, you can select the Red Hat OpenShift version 4.18.12 or 4.16.38 for Fusion HCI 2.10. Settings such as: Red Hat OpenShift pull secret, IBM entitlement key, proxy settings, image registry settings, Red Hat OpenShift network configuration, MTU and Ingress

certificate will be configured in this stage. At the end of this stage, Fusion HCI will be setup with a three node Red Hat OpenShift cluster and management software.

- Phase 3: IBM Fusion HCI Services Installation: Expand the cluster with the rest of the nodes in HCI before setting up Fusion Data Foundation or Global Data Platform storage. Enabling the Fusion services such as backup and restore, and install essential operators as required.

For detailed deployment procedures, refer to the IBM Fusion HCI Installation Guide available through IBM Documentation.

3.7.5 Integration Points and Dependencies

Storage Integration: The deployment establishes native integration between Red Hat OpenShift and IBM storage technologies. This includes automatic storage class creation and persistent volume provisioning capabilities, which are essential for watsonx platforms that require dynamic storage allocation.

Security Framework: Establish security policies, certificate management, and initial role-based access control (RBAC) configurations. These security foundations must be in place before deploying user-facing platforms and applications.

Monitoring and Management: Establish basic monitoring and management capabilities during deployment. This includes cluster health monitoring, resource utilization tracking, and initial alerting configurations.

3.7.6 Post-Deployment Validation

Cluster Health Verification: Validate that all Red Hat OpenShift nodes are operational and properly joined to the cluster. Verify that control plane components are healthy and that worker nodes are ready to accept workloads.

Storage Validation: Confirm that storage classes are properly configured and that persistent volume provisioning functions correctly. Test both block and file storage capabilities to ensure compatibility with watsonx platform requirements.

Network Connectivity Testing: Verify network connectivity between cluster nodes, external networks, and planned integration points. Test bandwidth performance and latency characteristics critical for AI workloads.

Security Policy Validation: Confirm that security policies are properly applied and that certificate management functions correctly. Test authentication and authorization mechanisms to ensure proper access controls are in place.

3.7.7 Solution-Specific Validation Procedures

Cluster Health Verification: Validate that all Red Hat OpenShift nodes are operational and properly joined to the cluster. Verify that control plane components are healthy and that worker nodes are ready to accept workloads.

Storage Validation: Confirm that storage classes are properly configured and that persistent volume provisioning functions correctly. Test both block and file storage capabilities to ensure compatibility with watsonx platform requirements.

Network Connectivity Testing: Verify network connectivity between cluster nodes, external networks, and planned integration points. Test bandwidth performance and latency characteristics critical for AI workloads.

Security Policy Validation: Confirm that security policies are properly applied and that certificate management functions correctly. Test authentication and authorization mechanisms to ensure proper access controls are in place.

3.7.8 Troubleshooting Common Issues

Node Connectivity Problems: If cluster nodes fail to join properly, verify network configuration, DNS resolution, and firewall rules. Ensure that all required ports are open and that network time synchronization (NTP) is functioning correctly.

Storage Performance Issues: If storage performance does not meet requirements, verify storage device configuration, check for hardware compatibility issues, and validate that storage classes are configured with appropriate performance profiles.

Resource Allocation Conflicts: If resource allocation fails during deployment, verify that hardware specifications meet minimum requirements and that resource allocation policies are properly configured for the intended workload types.

Documentation and Support Resources:

- IBM Fusion HCI Installation Documentation
- Red Hat OpenShift Container Platform Installation Documentation

3.7.9 Next Steps

After completing IBM Fusion HCI appliance deployment, the environment is ready for the next installation phase. The following sections provide guidance for:

- Red Hat OpenShift platform configuration and optimization
- Storage infrastructure setup and performance validation
- Security policy implementation and compliance validation

This infrastructure foundation provides the reliability, performance, and scalability required for enterprise AI and data management workloads, enabling successful watsonx platform deployment.

3.7.10 OpenShift Container Platform Configuration

Configuration Overview

After you deploy the IBM Fusion HCI appliance, you must configure the OpenShift Container Platform to support the unique requirements of watsonx platforms and AI workloads. This configuration phase focuses on optimizing the container orchestration layer for high-performance computing, large-scale data processing, and enterprise-grade security and compliance.

This process transforms the base Red Hat OpenShift cluster into a production-ready platform that supports compute-intensive AI model training, storage-intensive data lakehouse operations, and the security requirements of enterprise and government workloads.

Solution-Specific Configuration Requirements

AI Workload Optimization: Configure Red Hat OpenShift to support the unique characteristics of AI and machine learning workloads. These include long-running training jobs, GPU resource management (if applicable), and dynamic scaling based on computational demands. Configure appropriate resource quotas, limit ranges, and scheduling policies to accommodate the burst nature of AI workloads.

Data Platform Integration: Ensure that the configuration supports the data movement and storage access patterns required by watsonx.data for lakehouse operations. Optimize network policies for high-throughput data transfers and configure storage classes that meet the performance requirements for both structured and unstructured data processing.

Multi-Tenant Security: Enterprise and government implementations require robust multi-tenant configurations. These configurations must provide secure isolation between projects, users, and workloads while maintaining the performance characteristics necessary for AI operations.

Storage Class Configuration and Management

Dynamic Storage Provisioning: Configure storage classes that automatically provision persistent volumes with the appropriate performance characteristics for different workload types. Establish storage classes for high-performance AI workloads, large-scale data storage, and temporary processing volumes.

Performance-Optimized Storage Classes: Create storage classes optimized for the performance profiles supported by the underlying IBM storage infrastructure. These include configurations for high-IOPS database workloads, high-throughput data ingestion, and balanced configurations for general application workloads.

Data Lifecycle Management: Implement storage class configurations that support data lifecycle management policies. These policies should include automatic tiering of data based on access patterns and age, and integration with backup and archival systems.

Operator Lifecycle Management

Essential Operators Installation: Install and configure the operators required for watsonx platform integration. These include storage operators for IBM Storage Ceph, networking operators for advanced network policies, and security operators for compliance and governance.

Operator Update Policies: Establish update policies that balance security and stability with the need for new features. Configure automatic updates for security patches while maintaining control over major version updates that might affect workload compatibility.

Custom Resource Definitions: Configure the custom resource definitions (CRDs) required for watsonx platform integration. Ensure compatibility with existing cluster configurations and avoid conflicts with other platform components.

Network Policy and Security Configuration

Network Segmentation: Implement network policies that segment different workload types while maintaining the connectivity required for data flow and platform integration. Establish dedicated networks for storage traffic, inter-pod communication, and external connectivity.

Security Context Constraints: Configure security context constraints (SCCs) that provide the necessary permissions for watsonx platform components. Maintain security boundaries and

compliance by establishing policies for privileged operations, file system access, and network connectivity.

Service Mesh Integration: If applicable, configure service mesh integration to provide advanced traffic management, security policies, and observability capabilities for complex AI and data platform deployments.

Resource Management and Scheduling

Node Labeling and Taints: Implement node labeling strategies that enable intelligent workload placement based on hardware characteristics, such as GPU availability, memory capacity, and storage performance. Configure taints and tolerations to ensure that specialized workloads are scheduled on appropriate nodes.

Resource Quotas and Limits: Establish resource quotas and limits to prevent contention while supporting the burst capacity requirements of AI workloads. Configure memory limits for large model training and CPU limits for parallel processing.

Pod Disruption Budgets: Configure pod disruption budgets to ensure high availability for critical platform components. These budgets allow for planned maintenance and updates without service interruption.

Monitoring and Observability Configuration

Cluster Monitoring Setup: Configure cluster monitoring to provide visibility into resource utilization, performance metrics, and health status across all platform components. Establish metrics collection for AI workload performance and data processing throughput.

Logging Configuration: Implement centralized logging that captures platform operations and application-level events. Configure log retention policies that balance storage requirements with audit and compliance needs.

Alerting Policies: Establish alerting policies that notify administrators of critical issues while avoiding alert fatigue. Configure alerts for resource exhaustion, performance degradation, and security events.

Validation and Testing Procedures

Cluster Health Verification: Validate that all cluster components function correctly and that resource allocation operates as expected. Confirm that storage classes are properly configured and that persistent volume provisioning works correctly.

Workload Scheduling Validation: Test workload scheduling to ensure that pods are placed on appropriate nodes based on resource requirements and constraints. Verify that resource limits and quotas are enforced correctly.

Network Connectivity Testing: Validate network connectivity between pods, services, and external resources. Test network policies to ensure that security boundaries are properly implemented without blocking legitimate traffic.

Performance Baseline Establishment: Establish performance baselines for the configured cluster. These baselines should include resource utilization patterns, network throughput, and storage performance metrics. Use these baselines to support monitoring and optimization in production environments.

Configuration Best Practices

Namespace Organization: Implement a namespace organization strategy that provides logical separation between platform components and user workloads. Maintain operational efficiency and security boundaries.

Configuration Management: Use configuration management tools and practices to ensure consistency across environments. Facilitate updates and maintenance operations through version-controlled configuration.

Documentation and Change Control: Maintain detailed documentation of all configuration changes. Implement change control processes to ensure that changes are properly tested and approved before implementation.

Troubleshooting Common Configuration Issues

Storage Class Conflicts: If storage provisioning fails, verify that storage classes are properly configured and that the underlying storage infrastructure is accessible from all cluster nodes.

Resource Allocation Failures: If workloads fail to start due to resource constraints, verify that resource quotas and limits are appropriately configured and that nodes have sufficient available resources.

Network Policy Conflicts: If network connectivity issues arise, verify that network policies are not blocking legitimate traffic and that security context constraints are properly configured.

Documentation References

For detailed configuration procedures, refer to the official documentation:

- Red Hat OpenShift Container Platform Configuration Guide: Available through Red Hat Customer Portal
- IBM Storage Operators Configuration: Available through IBM Storage Documentation
- Red Hat OpenShift Security Best Practices: Available through Red Hat Documentation
- Kubernetes Resource Management: Available through Kubernetes Documentation

Next Steps

After you configure the OpenShift Container Platform, the environment is ready for storage infrastructure setup and performance validation. The following sections provide guidance for:

- Storage infrastructure configuration and optimization
- Performance validation and tuning procedures
- Integration with watsonx platform components

This configured Red Hat OpenShift environment provides the foundation for reliable, secure, and high-performance AI and data platform operations. It ensures that subsequent platform deployments can leverage the full capabilities of the underlying infrastructure.

3.7.11 Storage Infrastructure Setup

Assess storage readiness and performance before deploying IBM Cloud Paks. The IBM Software Hub storage performance tool verifies storage performance within a Red Hat OpenShift cluster. Use this tool to validate storage performance before installation and ensure compliance with recommended performance guidelines for each test case.

The tool includes four test cases:

- PVC with ReadWriteOnce (RWO) access mode: Random write 4 KB with 8 threads

- PVC with RWO access mode: Sequential write 1 GB with 2 threads
- PVC with ReadWriteMany (RWX) access mode: Random write 4 KB with 8 threads
- PVC with RWX access mode: Sequential write 1 GB with 2 threads

For more information, see Validation tools for IBM Cloud Paks.

The I/O throughput requirements for IBM Cloud Paks are as follows:

Storage Access Mode	Test	Requirement
PVC with RWO	Random write 4 KB with 8 threads	11 MiB/s or higher
PVC with RWO	Sequential write 1 GB with 2 threads	128 MiB/s or higher
PVC with RWX	Random write with 8 threads	11 MiB/s or higher
PCVC with RWX	Sequential write 1 GB with 2 threads	128 MiB/s or higher

Table 3-1 I/O throughput requirements for IBM Cloud Paks

3.7.12 Test Environment Overview

The storage performance validation tool is an Ansible playbook that collects storage performance metrics from the Red Hat OpenShift cluster. You can also use a separate Storage Validation Tool for IBM Cloud Paks to assess whether a storage provider supports RWO and RWX volumes for Cloud Pak workloads. This tool is available on GitHub.

3.7.13 Storage Performance Validation Tool Installation Guide

The storage validation tool is available on GitHub. The storage validation tool is available on GitHub. Follow this brief installation guide to perform performance testing.

Installation prerequisites:

```
yum install python3.10
ln -s -f /usr/bin/python3.10 /usr/bin/python

python --version
pip --version

pip install ansible
pip install openshift

ansible-galaxy collection install operator_sdk.util
ansible-galaxy collection install kubernetes.core
```

Clone the repository:

```
# clone the storage validation tool
git clone https://github.com/IBM/k8s-storage-perf
```

Log in to Red Hat OpenShift and create a new project and edit the parameters file:

```
# Login to Red Hat OpenShift with the oc command and create a new project
oc new-project storage-perf-testing

cd k8s-storage-perf/
vi params.yml
```

Sample params.yml configuration:

```
# OCP Parameters
ocp_url: https://<required>:6443
ocp_username: <required>  # a cluster admin user
ocp_password: <required>
ocp_token: <required if user/password not available>
ocp_apikey: <required if neither user/password or token not available>

storageClass_ReadWriteOnce: ocs-storagecluster-ceph-rbd
storageClass_ReadWriteMany: ocs-storagecluster-cephfs

run_storage_perf: true

arch: amd64

############################ STORAGE PERFORMANCE PARAMETERS START
######################
# vars file for roles/storage-perf-test
storage_perf_namespace: storage-perf-testing  # openshift namespace/project
where jobs will be executed, it will be created by playbook if not already
existing.
logfolder: '.logs'
imageurl: quay.io/ibm-cp4d-public/xsysbench:1.1

cluster_infrastructure: fusion # optional labels eg ibmcloud, aws, azure,
vmware
cluster_name: storage-performance-cluster # optional labels
storage_type: ocs # optional label eg portworx, ocs, <storage vendor>

# To run the performance jobs on a dedicated compute nodes, set the node label
that meets the criteria.
# The idea is to gather performance data when the jobs are running remotely
from a storage node.
# A cluster administrator can label a node by running this query with
appropriate label key/value:
# oc label node <node name> "<label_key>=<label_value>" --overwrite
dedicated_compute_node:
   label_key: "<optional>"
   label_value: "<optional>"

rwx_storagesize: 10Gi
rwo_storagesize: 10Gi
```

```
file_extra_flags: dsync
#sysbench random read
sysbench_random_read: false
rread_threads: 8            # 1,4,8,16
rread_fileTotalSize: 128m
rread_fileNum: 128
rread_fileBlockSize: 4k     # 4k,8k,16k
#sysbench random write
sysbench_random_write: true
rwrite_threads: 8           # 1,4,8,16
rwrite_fileTotalSize: 4096m
rwrite_fileNum: 4
rwrite_fileBlockSize: 4k    # 4k,8k,16k
#sysbench sequential read
sysbench_sequential_read: false
sread_threads: 2            # 1,2
sread_fileTotalSize: 4096m
sread_fileNum: 4
sread_fileBlockSize: 1g     # 512m,1g
#sysbench sequential write
sysbench_sequential_write: true
swrite_threads: 2           # 1,2
swrite_fileTotalSize: 4096m
swrite_fileNum: 4
swrite_fileBlockSize: 1g    # 512m,1g
############################ STORAGE PERFORMANCE PARAMETERS END
########################
```

Start the performance test:

```
# Disable SSL verification
export K8S_AUTH_VERIFY_SSL=no

#Start the performance test from the root of the repository
ansible-playbook main.yml --extra-vars "@./params.yml" | tee output.log
```

3.7.14 Fusion Data Foundation Performance Profile

Fusion Data Foundation provides an intuitive interface to change the storage performance profile. Log in to the Red Hat OpenShift web console with an admin account. Navigate to **Configure Performance.**

From this interface, you can adjust performance settings to meet workload requirements.

Select the performance profile and click **Save**. It will take a while for the storage engine to reconfigure and restart the pods.

Data Foundation Performance Testing

This section describes the execution of the storage validation tool across three different performance profiles. The results of each test are recorded for comparison.

This testing uses a small seven-node IBM Fusion HCI cluster. Performance results and resource requirements for performance profiles vary depending on the size and configuration of the deployed Fusion HCI system.

Test Environment:

- Fusion HCI v2.9
- Red Hat OpenShift v4.16.39
- Data Foundation v4.16.9

Fusion HCI Hardware:

- 3 x C01 Storage/Compute servers
  - 32 core and 256 GB RAM
  - 2 x 7.68 TB NVMe PCIe Gen4 disks on each server
  - 2 x 960 GB M.2 NVMe OS drives
- 3 x C05 Storage/Compute servers
  - 64 core and 1024 GB RAM
  - 2 x 7.68 TB NVMe PCIe Gen4 disks on each server
  - 2 x 480 GB M.2 NVMe OS drives
- 1 x Compute nodes
  - 32 core and 192 GB RAM
  - 2 x 960 GB SSD OS drives

Figure 3-1 Storage Validation Tool

Sequential Write Performance

Environment	PVC	Test Name	Throughput (MiB/s)
Fusion HCI with DF (Lean)	pvc-sysbench-rwx	seqwr_1g_2	1,372.1
Fusion HCI with DF (Balanced)	pvc-sysbench-rwx	seqwr_1g_2	1,593.8
Fusion HCI with DF (Performance)	pvc-sysbench-rwx	seqwr_1g_2	2,331.2

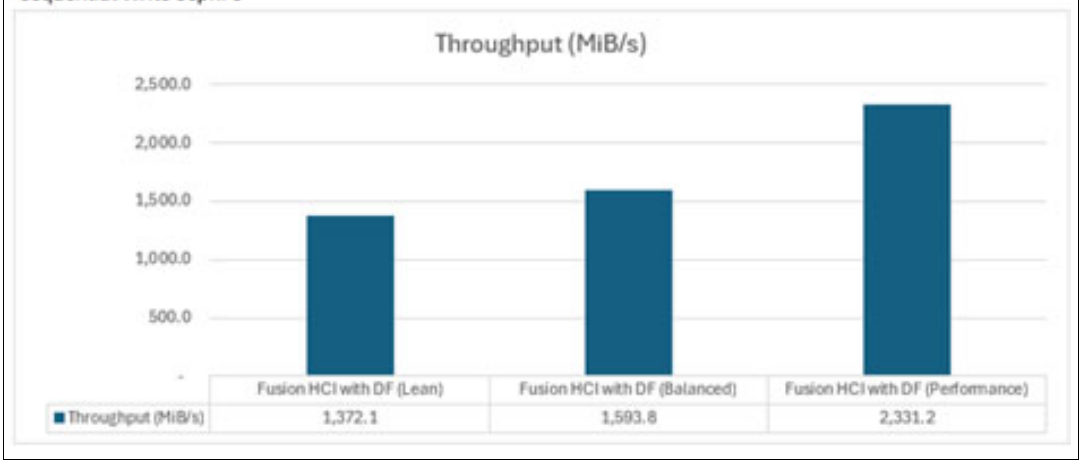


Figure 3-2 Sequential Write CephFS

Sequential Write Block Performance

Environment	PVC	Test Name	Throughput (MiB/s)
Fusion HCI with DF (Lean)	pvc-sysbench-rwo	seqwr_1g_2	1,399.4
Fusion HCI with DF (Balanced)	pvc-sysbench-rwo	seqwr_1g_2	1,777.7
Fusion HCI with DF (Performance)	pvc-sysbench-rwo	seqwr_1g_2	2,006.5

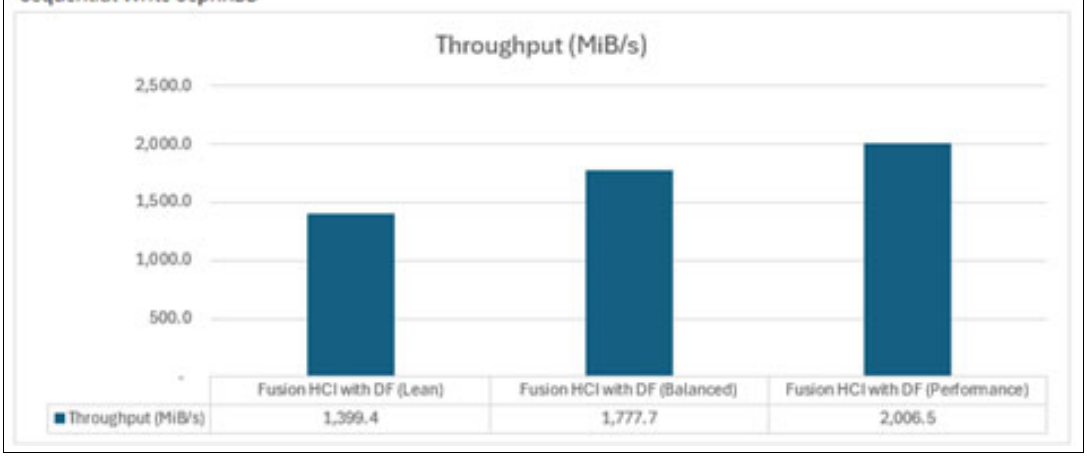


Figure 3-3 Sequential Write CephRBD

Random Write File Performance

Environment	PVC	Test Name	Throughput (MiB/s)
Fusion HCI with DF (Lean)	pvc-sysbench-rwx	rndwr_4k_8	31.0
Fusion HCI with DF (Balanced)	pvc-sysbench-rwx	rndwr_4k_8	34.3
Fusion HCI with DF (Performance)	pvc-sysbench-rwx	rndwr_4k_8	34.8

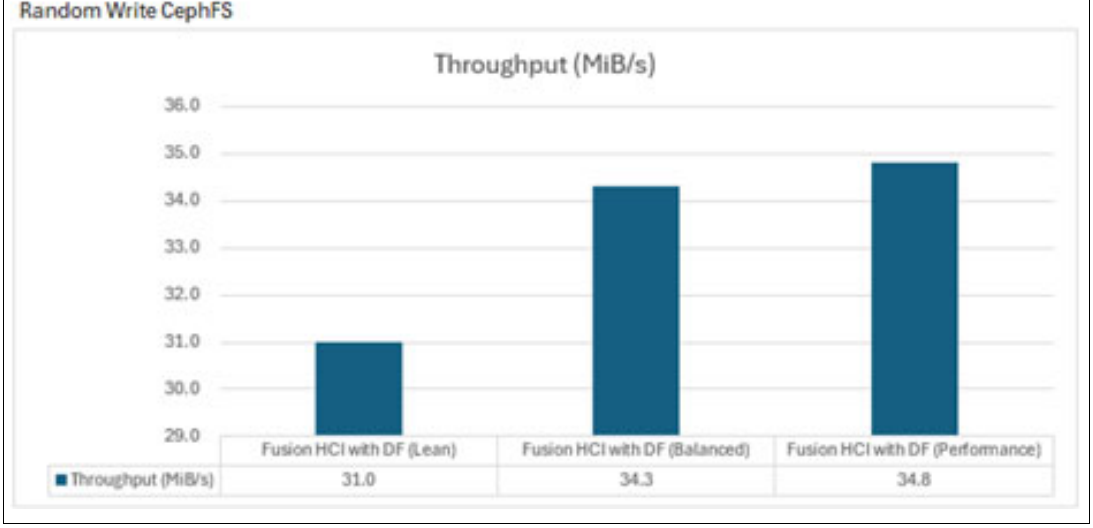


Figure 3-4 Random Write CephFS

Random Write Block Performance

Environment	PVC	Test Name	Throughput (MiB/s)
Fusion HCI with DF (Lean)	pvc-sysbench-rwo	rndwr_4k_8	31.7
Fusion HCI with DF (Balanced)	pvc-sysbench-rwo	rndwr_4k_8	32.2
Fusion HCI with DF (Performance)	pvc-sysbench-rwo	rndwr_4k_8	31.6

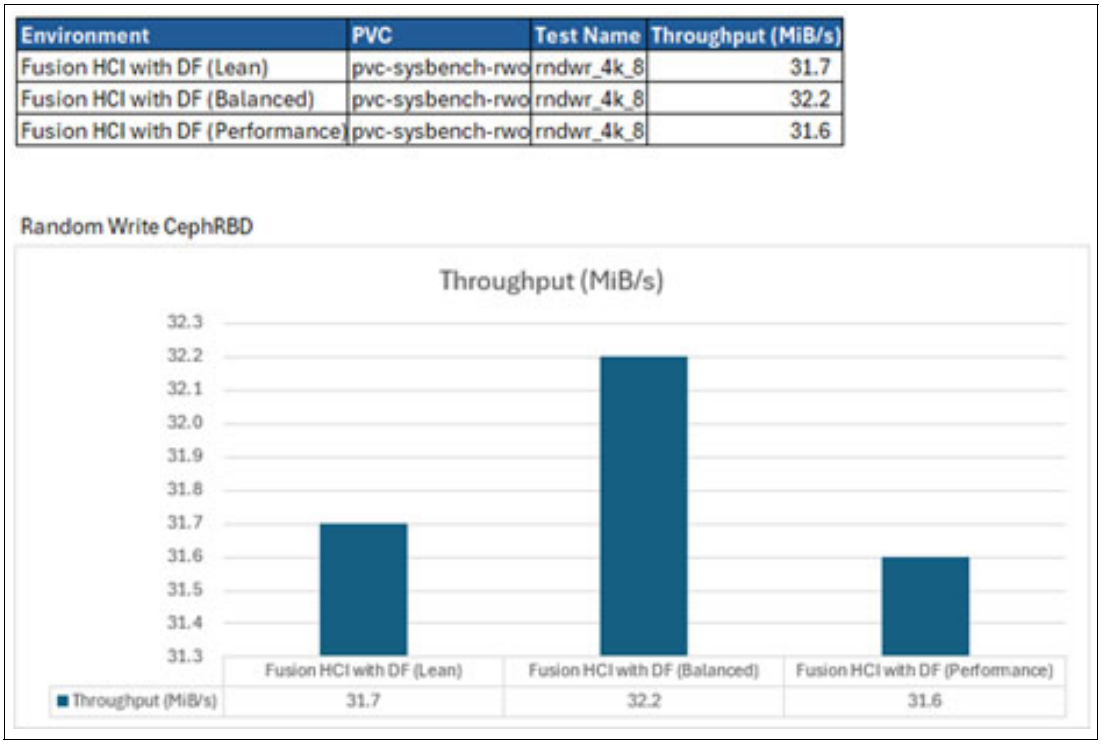


Figure 3-5 Random Write CephRBD

IBM Fusion HCI exceeds the storage performance requirements for IBM Cloud Paks and IBM watsonx products. Sequential write performance for file storage surpasses the throughput requirement of 128 MiB/s by a factor of 11 to 18, depending on the selected Data Foundation performance profile (Lean, Balanced, or Performance Mode). Similarly, sequential write performance for block storage exceeds the 128 MiB/s threshold by a factor of 11 to 16.

In addition, random write operations for both file and block storage exceed the throughput requirement of 11 MiB/s by a factor of three across all three performance profiles.

The default performance profile meets and exceeds the storage requirements for IBM Cloud Paks.

However, if enhanced sequential write performance is required, you can switch to Performance Mode. This profile increases sequential write throughput by approximately 46% for file storage and 13% for block storage.

3.8 Storage Architecture for AI/ML Workloads

The storage infrastructure represents a critical component in the watsonx solution stack, directly impacting the performance and scalability of AI and machine learning operations. AI workloads exhibit unique storage characteristics, including large sequential reads during data ingestion, high throughput writes during model training, and mixed access patterns during inference operations. The storage configuration must accommodate these diverse requirements while maintaining enterprise-grade reliability and performance.

IBM Storage Ceph provides a software-defined storage foundation that scales horizontally and supports the dynamic storage demands of containerized AI workloads. When integrated with IBM Fusion HCI, this storage architecture delivers the performance characteristics required for both structured and unstructured data processing while providing the flexibility to adapt to changing workload requirements.

3.9 Storage Integration Architecture

Unified Storage Platform: The storage infrastructure provides unified block, file, and object storage capabilities through a single platform, eliminating the complexity and performance overhead of managing separate storage systems for different workload types. This unified approach is particularly beneficial for AI workloads that often require multiple storage access patterns within the same application lifecycle.

Container-Native Integration: The storage platform integrates natively with OpenShift Container Platform, providing dynamic storage provisioning, automated lifecycle management, and policy-based storage allocation. This integration enables AI applications to request storage resources programmatically and scale storage capacity based on workload demands.

High-Availability Configuration: Storage infrastructure is configured with redundancy and fault tolerance appropriate for production AI workloads, including data replication policies, automatic failover capabilities, and self-healing mechanisms that ensure continuous data availability during hardware failures or maintenance operations.

3.10 AI/ML Performance Optimization Strategies

Storage Class Optimization: Configure storage classes specifically optimized for different AI/ML workload patterns, including high-throughput classes for data ingestion, low-latency classes for model training, and balanced classes for general-purpose AI operations. Each storage class is tuned with appropriate replication factors, erasure coding policies, and performance profiles.

Data Placement Strategies: Implement data placement policies that optimize storage performance based on access patterns and data characteristics. This includes placing frequently accessed data on high-performance storage tiers and implementing automated tiering policies that move data between performance tiers based on usage patterns.

Caching and Acceleration: Configure caching layers and performance acceleration features that reduce latency for frequently accessed data and improve overall system throughput. This includes client-side caching for read-intensive workloads and write optimization for high-throughput data ingestion operations.

Network Optimization: Optimize storage network configurations to minimize latency and maximize throughput between compute nodes and storage resources. This includes configuring dedicated storage networks, implementing appropriate network bonding strategies, and optimizing network buffer sizes for large data transfers.

Storage Deployment Options in Fusion HCI

There are two internal storage options in Fusion HCI:

- Fusion Data Foundation: Fusion HCI will setup Fusion Data Foundation provider mode. It will become the base storage on the host cluster. With hosted control plane, the virtualized Red Hat OpenShift clusters created in the HCI will be able to consume the storage from the host cluster. The Fusion Data Foundation can be configured with data-at-rest encryption. It supports storing the encryption key in a secret in the cluster (default) or storing the encryption key in an external Key Management Service (KMS).
- Global Data Platform: Fusion HCI will setup Global Data Platform local. Options in this mode includes: Stronger data resiliency and Better storage efficiency. Stronger data resiliency and Better storage efficiency are available when there are more than eleven storage nodes in the HCI. Stronger data resiliency is available when there are less than eleven storage nodes.

The Global Data Platform remote mount feature allows Fusion HCI to remote mount an external IBM Storage Scale cluster. This external storage is presented as a new storage class for application to consume in Fusion HCI.

This option is only available with the following deployment modes:

- Standalone rack with Global Data Platform
- Standalone rack with Fusion Data Foundation
- Expansion multi-rack with Global Data Platform

Global Data Platform with Multi-Cloud Gateway (MCG) is a cloud-native, software object storage gateway. It provides a single S3 compatible interface backed by different storage backends. MCG can connect to external or internal object storage system, internal file systems or container volumes. It also supports features such as caching, tiering, mirroring, deduplication, encryption and compression. MCG is available (by default) in Fusion Data Foundation. If Global Data Platform is chosen, you can enable MCG separately.

3.11 Storage Performance Validation

Importance for AI/ML Success: Storage performance validation is critical for AI/ML implementations because inadequate storage performance directly impacts model training times, data processing throughput, and overall solution effectiveness. The performance validation procedures ensure that the storage infrastructure meets the demanding requirements of AI workloads before deploying production applications.

Validation Against AI Workload Patterns: The performance validation process specifically tests storage performance against the access patterns typical of AI and machine learning workloads, including large sequential reads for data loading, high-concurrency random access for distributed training, and mixed workload patterns that combine read and write operations.

Data Foundation Performance Profiles

Profile Selection for AI Workloads: IBM Fusion Data Foundation provides multiple performance profiles (Lean, Balanced, and Performance) that can be selected based on the specific requirements of AI workloads. The profile selection directly impacts storage performance, resource utilization, and overall system efficiency.

AI Workload Performance Considerations:

- Lean Mode: Suitable for development and testing environments with moderate AI workloads
- Balanced Mode: Appropriate for production environments with mixed AI and general workloads
- Performance Mode: Optimized for high-performance AI operations requiring maximum throughput and minimum latency

The performance profile selection should be based on careful analysis of workload requirements, resource availability, and performance objectives established during the planning phase.

3.12 Storage Integration with watsonx Platforms

watsonx.data Integration: Configure storage policies and access patterns optimized for data lakehouse operations, including support for both structured and unstructured data, metadata management, and query optimization. The storage configuration must support the diverse data formats and access patterns required by watsonx.data implementations.

watsonx.ai Integration: Establish storage configurations that support the model development and deployment lifecycle, including model artifact storage, training data management, and inference data handling. This includes configuring appropriate storage classes for different phases of the AI model lifecycle.

Cross-Platform data sharing: Configure storage policies that enable efficient data sharing between watsonx platforms while maintaining security boundaries and access controls. This includes establishing shared storage namespaces and implementing data governance policies that support collaborative AI development.

3.13 Monitoring and Management

Performance Monitoring: Implement storage performance monitoring that provides visibility into throughput, latency, and resource utilization metrics specific to AI workloads. This monitoring enables proactive identification of performance issues and optimization opportunities.

Capacity Management: Establish capacity management processes that monitor storage utilization and predict capacity needs based on AI workload growth patterns. This includes automated alerting for capacity thresholds and integration with capacity planning processes.

Health and Maintenance: Configure storage health monitoring and maintenance procedures that ensure optimal performance and reliability. This includes automated detection of storage node issues, predictive failure analysis, and maintenance scheduling that minimizes impact on AI workloads.

3.14 Official Documentation References

For detailed storage installation and configuration procedures, see the following. Always verify compatibility matrices and version requirements before beginning storage deployment.

- IBM Storage Ceph Installation Guide: Available through IBM Storage Documentation Portal
- Red Hat Ceph Storage Configuration: Available through Red Hat Customer Portal
- Red Hat OpenShift Data Foundation Deployment Guide: Available through Red Hat Documentation
- IBM Fusion HCI Storage Integration: Available through IBM Fusion HCI Documentation

3.15 Support and Technical Resources

- IBM Storage Technical Support for Ceph-specific issues
- Red Hat Support for Red Hat OpenShift Data Foundation integration
- IBM Fusion HCI Support for integrated storage configurations
- Community resources and knowledge bases for optimization guidance

3.16 Next Steps

With storage infrastructure properly configured and performance validated, the environment is ready for platform services deployment. The following sections provide guidance for:

- Cloud Pak for Data foundation platform deployment
- watsonx platform integration and configuration
- End-to-end solution validation and testing

The optimized storage infrastructure provides the performance foundation required for successful AI and data platform operations, ensuring that subsequent platform deployments can achieve their intended performance and scalability objectives.

3.17 Foundation Platform Deployment

Prior to the installation of IBM Software Hub, Fusion HCI should enable backup & restore feature, operators for GPUs and configure HCI to reserve GPU nodes for AI workloads.

3.17.1 Fusion Backup & Restore:

The Backup & Restore service in Fusion protects your workloads in Red Hat OpenShift. You can restore your data rapidly from local snapshots, storing the backup to any external object storage system that is S3-compliant such as IBM Cloud Object Storage and IBM Storage Ceph. You can also store the backup inside the built-in S3 storage in Fusion HCI. Fusion has a recipe to perform online backup of IBM Software Hub. You can extend this backup feature to become a disaster recovery strategy, leveraging the backup and restore to a different cluster. Refer to the following documentation for details on how to install Backup & Restore:

Installation of Backup & Restore in Fusion

Backup and restore of IBM Software Hub to the same or different cluster

Operators for GPU

The following operators should be installed in Fusion HCI:

- The Node Feature Discovery (NFD) Operator is a prerequisite for the NVIDIA GPU Operator.
- NVIDIA GPU Operator

Reserve Fusion HCI GPU nodes for AI workloads

To prevent Fusion Backup & Restore workloads on GPU nodes, run the **`oc label node/<GPU node name> gpu.isf.ibm.com=""`** command.

Bastion node

The bastion node acts as a secure jump host to Fusion HCI. This can be a virtual machine or a physical server residing outside of Fusion HCI. The bastion node hosts the Red Hat OpenShift CLI (oc) and IBM Software Hub CLI (cpd-cli). For more information, see Installing the IBM Software Hub command-line interface.

Red Hat OpenShift Preparation in Fusion HCI

Update Red Hat OpenShift global pull secret with the following command:

```
cpd-cli manage add-icr-cred-to-global-pull-secret
--entitled_registry_key=${IBM_ENTITLEMENT_KEY}
```

Check that all nodes are updated using the following command:

```
oc get mcp cpd-cli manage
oc get nodes
```

Install prerequisite software

- Install Red Hat OpenShift AI version 2.19.1 with IBM Software Hub version 5.2.1. Many services require Red Hat OpenShift AI such as watsonx.ai, watsonx.data Premium, watsonx Assistant®, and others. For more information, see Installing Red Hat OpenShift AI.

- Install Multi-cloud Gateway (MCG). Fusion HCI includes MCG. This is required if you plan to install watsonx Assistant, watsonx Discovery, Watson Speech services, or watsonx Orchestrate. The default backing store in MCG uses the Fusion Data Foundation Ceph RGW storage class (ocs-storagecluster-ceph-rgw). You can create a new backing store that points to an external enterprise-grade S3 compatible storage such as IBM Storage Ceph and IBM Cloud Object Storage. For more information, see Installing and setting up Multicloud Object Gateway for IBM Software Hub.
- Install Red Hat OpenShift Serverless Knative Eventing. This is required if you plan to install IBM watsonx Assistant or IBM watsonx Orchestrate.

Install shared services

License Service:

```
export PROJECT_LICENSE_SERVICE=cpd-license
export PROJECT_SCHEDULING_SERVICE=cpd-scheduling
export VERSION=5.2.1

cpd-cli manage apply-cluster-components \
--release=${VERSION} \
--license_acceptance=true --licensing_ns=${PROJECT_LICENSE_SERVICE}
```

Scheduling Service:

```
cpd-cli manage apply-scheduler
--release=${VERSION}
--license_acceptance=true
--scheduler_ns=${PROJECT_SCHEDULING_SERVICE}
```

Install an instance of IBM Software Hub:

```
export PROJECT_CPD_INST_OPERATORS=cpd-operators
export PROJECT_CPD_INST_OPERANDS=cpd-watsonx
```

Fusion Data Foundation default storage class:

```
export STG_CLASS_BLOCK=ocs-storagecluster-ceph-rbd
export STG_CLASS_FILE=ocs-storagecluster-cephfs
```

Create projects to host the instance:

```
oc new-project ${PROJECT_CPD_INST_OPERATORS}
oc new-project ${PROJECT_CPD_INST_OPERANDS}
```

Apply required permissions to the projects for an instance:

```
cpd-cli manage authorize-instance-topology
--cpd_operator_ns=${PROJECT_CPD_INST_OPERATORS}
--cpd_instance_ns=${PROJECT_CPD_INST_OPERANDS}
```

Create the instance: cpd-cli manage setup-instance

```
--release=${VERSION}
--license_acceptance=true
--cpd_operator_ns=${PROJECT_CPD_INST_OPERATORS}
--cpd_instance_ns=${PROJECT_CPD_INST_OPERANDS}
--block_storage_class=${STG_CLASS_BLOCK}
--file_storage_class=${STG_CLASS_FILE}
--run_storage_tests=true
```

watsonx Platform Integration

Install watsonx.ai instance

```
export COMPONENTS=watsonx_ai
```

Apply entitlement for watsonx.ai:

```
cpd-cli manage apply-entitlement
--cpd_instance_ns=${PROJECT_CPD_INST_OPERANDS}
--entitlement=watsonx-ai
--production=false
```

Install the services operators

cpd-cli manage apply-olm

```
--release=${VERSION}
--cpd_operator_ns=${PROJECT_CPD_INST_OPERATORS}
--components=${COMPONENTS}
```

Create option file:

```
vi ./cpd-cli-workspace/olm-utils-workspace/work/install-options.yml
########################################################################## #
watsonx.ai parameters
########################################################################
custom_spec:
watsonx_ai: tuning_disabled: false
lite_install: false
```

Start installation

```
cpd-cli manage apply-cr
--release=${VERSION}
--cpd_instance_ns=${PROJECT_CPD_INST_OPERANDS}
--components=${COMPONENTS}
--block_storage_class=${STG_CLASS_BLOCK}
--file_storage_class=${STG_CLASS_FILE}
--param-file=/tmp/work/install-options.yml
--license_acceptance=true
```

Validate the installation

```
cpd-cli manage get-cr-status
--cpd_instance_ns=${PROJECT_CPD_INST_OPERANDS}
--components=watsonx_ai
```

Retrieve the URL and default credentials

```
cpd-cli manage get-cpd-instance-details
--cpd_instance_ns=${PROJECT_CPD_INST_OPERANDS}
--get_admin_initial_credentials=true
```

For more information, see Installing IBM watsonx.ai.

3.17.2 Solution Validation and Testing

Install foundation model with default configuration:

```
oc patch watsonxaiifm watsonxaiifm-cr
--namespace=${PROJECT_CPD_INST_OPERANDS}
--type=merge
--patch='{"spec":{"install_model_list":
["llama-3-2-1b-instruct","granite-3-3-8b-instruct"]}}'
```

For more information, see Foundation models in IBM watsonx.ai.

Check progress of the deployment:

```
oc get watsonxaiifm -n ${PROJECT_CPD_INST_OPERANDS}
```

Verify in Prompt Lab

Open **Prompt Lab** and ask a question

4

Management and Operations

As you increasingly rely on hybrid cloud platforms to run mission-critical workloads, the ability to monitor, troubleshoot, and maintain infrastructure and applications becomes essential. IBM Fusion HCI and IBM Software Hub together provide a powerful, integrated solution for unified observability, proactive diagnostics, and streamlined operations.

IBM Fusion HCI delivers a hyperconverged infrastructure platform optimized for Red Hat OpenShift, with built-in tools to ensure the health, availability, and performance of hardware and system services. IBM Software Hub extends observability to the software layer, offering a centralized environment to monitor and manage IBM solutions deployed on Red Hat OpenShift.

IBM Fusion HCI Monitoring IBM Fusion HCI offers comprehensive observability across hardware and software components through two main methods, designed to suit both visual monitoring and external tool integration.

© Copyright IBM Corp. 2025.

4.1 IBM Fusion HCI Monitoring

IBM Fusion HCI offers comprehensive observability across hardware and software components through two main methods, designed to suit both visual monitoring and external tool integration

4.1.1 Monitoring through the IBM Fusion HCI UI

The built-in web-based UI delivers a real-time graphical view of the physical infrastructure. You can:

- View the health and status of nodes, switches, and service components.
- Inspect individual rack units and hardware elements with color-coded health indicators.
- Zoom, navigate, and scale the rack view for multi-rack environments.
- Quickly identify issues such as hardware degradation or critical failures.

The IBM Fusion HCI interface provides a graphical rack view and a summary of the hardware resource status for the cluster named `isfh14sb12`. It gives an at-a-glance health view of all nodes and switches in the rack, making it easy for you to quickly spot and troubleshoot any degraded or failed components.

Zoom & Navigate

You can zoom in/out to inspect specific parts of the rack. Hover over any component to view its name, health condition (for example, healthy, warning, critical), type (compute node, switch, service node), inventory details, and rack unit (RU) position.

Scalable Rack View

You can add more racks if your environment exceeds 3 racks, which helps with navigation in large deployments.

The Nodes page from the IBM Fusion HCI web interface shows a detailed list of hardware nodes in a cluster, specifically under the **Configured** tab.

Events

The **Events** dashboard provides a detailed, time-stamped list of system events streamed directly from the OpenShift Container Platform.

4.1.2 Monitoring with External Tools (Prometheus and Grafana)

IBM Fusion HCI integrates with Red Hat OpenShift Prometheus instance to export performance metrics from compute, network, and storage layers, as well as system events into the Red Hat OpenShift Event Manager. This data can be consumed by any compatible observability tool (for example, Grafana, ELK) for customized dashboards and alerts. IBM provides ready-to-use Grafana dashboard templates. Logs and CRs are also available for advanced troubleshooting.

Logging and Event Management

IBM Fusion HCI is built to integrate easily with your existing observability stack. You can monitor system health and performance using widely adopted tools like Prometheus, Grafana, and ELK.

Event Integration with Red Hat OpenShift

IBM Fusion pushes system and service events into Red Hat OpenShift's Event Manager. Types of events include hardware (nodes, switches) and data services (backup, restore). These events are automatically visible in any platform already integrated with Red Hat OpenShift's event system.

Metrics Export to Prometheus

IBM Fusion collects performance metrics across storage, network, and compute. These metrics are sent to the Prometheus instance embedded in Red Hat OpenShift. Once in Prometheus, they can be visualized in Grafana or any compatible monitoring system. IBM provides pre-built Grafana dashboards to streamline setup.

Log and Status Monitoring

IBM Fusion HCI offers multiple log sources: hardware and service operations, network activity, storage platform (GDP), and backup and cataloging logs. Status can also be checked via a web-based UI for real-time health monitoring and Custom Resources (CRs) that reflect system state.

IBM Software Hub

IBM Software Hub is a cloud-native platform designed to install, manage, and monitor IBM solutions on Red Hat OpenShift. Key capabilities include standardized installation and upgrade workflows, tools for backup, restore, and resource monitoring, and support for scaling and managing installation health.

The IBM Software Hub Administration Console provides a centralized view to monitor and manage the environment, including system health, alerts, and key administrative tasks.

- The **Platform resource usage** dashboard from the IBM Software Hub interface, providing a real-time summary of system consumption and quotas.
- The **Software Hub Monitoring Dashboard** provides a real-time overview of system health, resource usage, and quota allocations for administrators.
- The **Monitoring** page for watsonx.data dashboard provides a real-time summary of resource usage, system status, and pod-level activity.

For more information, see IBM Software Hub.

4.1.3 Troubleshooting and Serviceability

IBM Fusion consolidates log files and diagnostic tools to streamline resolution. The IBM Support Center is available to assist with both hardware and software problems related to IBM Fusion.

Serviceability Features include:

- Call Home client
- Event manager
- Log collector: These tools automatically gather logs, identify issues, and update support tickets with relevant data.
- Event/Error Code References: A comprehensive list of possible system events and error messages is provided to aid diagnosis.

For more information, see Troubleshooting and serviceability.

4.1.4 Call Home and Support Features

IBM Fusion events are Kubernetes native events that include filter-specific labels.

In the Fusion GUI, navigate to the Events page by selecting the **Events** menu item shown in the left side navigation panel. The **Events** page displays a comprehensive log of system events related to IBM Fusion.

You can filter the event log using the Severity, Category, or Node. Figure 9 shows how to filter the event log using Severity. To tailor the view to your needs, you can customize the columns displayed in the **Event** table. Click the gear icon located in the upper-right corner of the table to select or deselect columns based on your preferences.

4.1.5 IBM Fusion HCI Call Home

The Call Home feature proactively notifies you and IBM Support of hardware malfunctions and critical configuration or environmental conditions. It continuously monitors the health and availability of your storage infrastructure, enabling timely intervention to prevent potential business disruptions.

Click **Settings** -> **Support** in the IBM Fusion GUI left navigation panel. Add details used by IBM to associate support cases with your organization (Company Name, Customer Number, Primary Contact Name, Email Address, Country, Phone Number).

4.1.6 IBM Fusion HCI Service Page

The Services page within the IBM Fusion user interface displays all services that are currently enabled and configured. If certain services were not activated during the initial operator installation, they can be enabled from this page as needed.

4.1.7 Upgrade and Lifecycle Management

The **Upgrades** page from the IBM Fusion HCI user interface provides administrators with an organized view of system components that are eligible for upgrade.

5

Fusion HCI as the Backbone for IBM watsonx: Technical Foundations of IBM watsonx

In the era of data-driven enterprises, the ability to harness artificial intelligence (AI) effectively has become a critical differentiator. IBM watsonx, IBM's next-generation AI and data platform, provides organizations with a comprehensive ecosystem for developing, training, and deploying AI models, including large language models (LLMs) and other generative AI applications. With capabilities spanning AI model development, data management, and governance, watsonx empowers enterprises to operationalize AI securely and at scale.

However, delivering enterprise-grade AI requires more than just software-it demands a robust, high-performance infrastructure. This is where Fusion Hyperconverged Infrastructure (Fusion HCI) becomes relevant. Fusion HCI integrates compute, storage, and networking into a single, software-defined platform, providing the scalability, performance, and reliability necessary for demanding AI workloads. By running watsonx on Fusion HCI, organizations can achieve low-latency data access, efficient GPU acceleration for model training, and simplified deployment of AI services in hybrid or on-premises environments.

This chapter explores the technical foundations of IBM watsonx, highlighting its capabilities and architecture, and examines how Fusion HCI enables enterprises to maximize the potential of AI. The synergy between watsonx and Fusion HCI demonstrates a seamless integration of advanced AI capabilities with enterprise-grade infrastructure, making AI deployment faster, more scalable, and more reliable.

In this chapter, you explore the technical background of Atlas AI platform.

© Copyright IBM Corp. 2025.

5.1 IBM watsonx.data: The lakehouse foundation for enterprise-grade generative AI

IBM IBM watsonx.data platform is IBM's next-generation data platform for big data and AI integration. Built on a lakehouse architecture, it combines open standards with enterprise-grade capabilities to support both structured and unstructured data. This unified approach enables full deployment and operation of generative AI applications across industries.

As digital transformation accelerates, organizations must convert growing volumes of unstructured data-such as contracts, reports, emails, and forms-into AI-consumable knowledge. Traditional data warehouses and lakes often lack the deep processing capabilities required for this transformation, limiting the effectiveness of generative AI in areas like question answering, compliance, and automation.

5.2 Background and technical architecture

IBM IBM watsonx.data platform is built on a modern lakehouse architecture that integrates the following key layers:

- Storage layer: Supports multiple object storage systems (AWS S3, Cloud Object Storage, MinIO) and supports HDFS migration.
- Table format support: Fully supports open table formats such as Apache Iceberg, Delta Lake, and Hudi.
- Metadata and governance layer: Unifies discovery, lineage tracking, classification, data protection rules, and access control for structured and unstructured data using CAMS 2.0 and IBM Knowledge Catalog.
- Query engine layer: Integrates Presto (interactive SQL), Spark (data engineering and AI model training), Milvus (vector retrieval), and AstraDB (read/write NoSQL database).
- AI-enhanced capabilities: Supports document entity extraction, semantic normalization, structured derivation, intelligent query, and retrieval-augmented generation (RAG).

This layered architecture ensures horizontal scalability and deployment flexibility, and it gives generative AI applications strong data support.

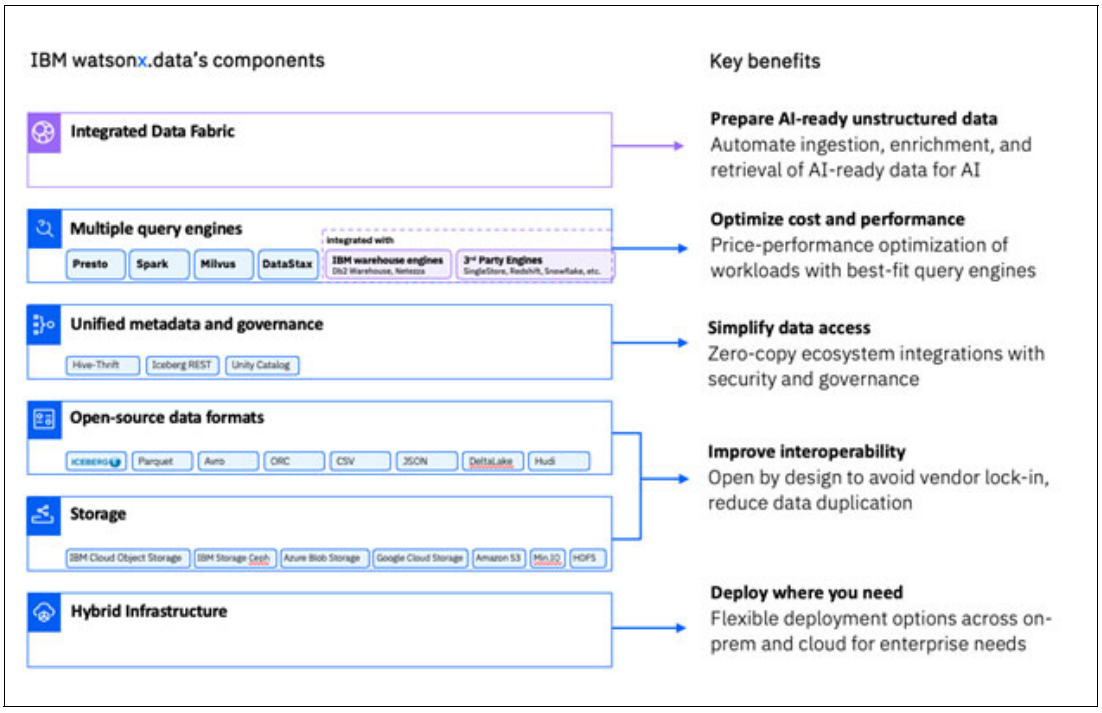


Figure 5-1 IBM watsonx components and benefits of each component

5.3 watsonx.data mechanism for unstructured data processing

Traditional RAG systems mainly rely on text chunking and semantic embedding to convert documents into vectors for LLM retrieval. However, for documents with tables, lists, and key fields-such as invoices, insurance policies, and process forms-this method struggles to extract deep semantic relationships and often leads to discrepancies between generated content and facts.

watsonx.data introduces a dual-channel data derivation mechanism:

- Channel 1 (Semantic Vector): Converts text snippets into vector embeddings and stores them in Milvus to enable semantic retrieval.
- Channel 2 (Structured Representation): Generates structured data in JSON or table format through entity recognition and key-value pair extraction, maps it to the business schema, and stores it in the iceberg table.
- Intelligent Retrieval Service: watsonx's intelligent retriever automatically determines whether to query structured data using SQL, search unstructured content using vectors, or provide hybrid answers based on the natural language query intent.

Tests show that in semi-structured documents-such as purchase orders, application forms, and insurance claim documents-the hybrid retrieval system in watsonx.data improves answer accuracy by more than 40% compared to solutions that use only a vector database.

watsonx.data also provides out-of-the-box, code-free connectors and development kits:

- Connection sources: Supports integration with mainstream document and collaboration platforms such as Box, SharePoint, Google Docs, FileNet®, Slack, Outlook, and others.
- SDK and Operator: Developers can use Python or Scala to extend the entity extraction process and adapt to emerging open source models such as DocTLin and LayoutLM.

- Semantic Schema Mapping: Uses the IKC module in watsonx Data Intelligence to semantically normalize extracted entities and reduce polysemy.

5.4 Security and governance capabilities: Opening up the AI security management and control chain

Unstructured documents often include personalized permission settings across different systems, such as:

- A document is viewable only by project team members
- Contract information becomes available only after signing
- Certain terms are subject to confidentiality agreements

watsonx.data implements the following strategies:

- Automatically inherits the document source's ACL permission model
- Retains and verifies permissions throughout the data consumption path
- Supports fine-grained access control (FGAC) and dynamic data masking

This ensures that only authorized users can access the relevant information, whether it's through RAG Q&A or BI report access.

With built-in data classification and PII detection capabilities, combined with metadata policy management, the system can automatically identify and mark the following types of sensitive content:

- PII such as ID numbers, email addresses, and physical addresses
- Financial and salary data
- Legal and compliance-related terminology

By combining Data Protection Rules, the system can dynamically apply protection mechanisms such as encryption, shielding, and desensitization.

5.4.1 Unified platform supports multiple AI and data analysis workloads

watsonx.data is not limited to improving AI question-answering. It is also an integrated operating foundation that supports data engineering, machine learning model development, business intelligence (BI) analysis, and AI agents.

- BI: Performs user-oriented calculations, aggregations, and tracking through tools such as watsonx BI and Tableau.
- Spark: Handles data engineering and ETL operations on structured data for cleaning and classification.
- watsonx.ai: Builds predictive models to support statistical methods such as sales forecasting, work order allocation, and accounts receivable forecasting.
- AI agents: Automate workflows for document matching, calculation, reconciliation, and optimization.

5.5 Typical application scenarios

- Financial scenarios: Automatically reads invoices and matches purchase orders to budgets.

- Insurance industry: Processes claims forms, identifies claim conditions, and automatically approves and grades them.
- Legal compliance: Analyzes key contract terms, tracks modification history, and supports text comparison.
- Retail and e-commerce: Automatically identifies user comments, classifies after-sales requests, and activates the appropriate agent workflow.

5.6 Atlas Cloud Services Platform

ACS AI Platform offered by Atlasx.ai, presenting a comprehensive full-stack view of how AI is operationalized across industries. The figure below is structured into three key layers:

- Use Cases: The platform supports a diverse range of AI-driven applications, including:
  - Agentic Workflows: Support for autonomous agents that plan and act on behalf of users. This includes integration with Watson Orchestrate.
  - Language-Based Tasks: Use of large language models (LLMs) for summarization, generation, classification, and more-leveraging models such as LLaMA, Mistral, and IBM Granite®.
  - Code Assistant: Helps developers write, review, and modernize code efficiently.
  - Time Series: Supports predictive maintenance, demand forecasting, and sensor data analytics.
  - Geospatial: Processes location-based data-often using vector databases and spatial models-relevant to telecom, utilities, and smart city scenarios.
  - Safety and Confidentiality: Strong commitment to privacy, model alignment, AI explainability, and ethical guardrails-aligned with GDPR and Moroccan law.
  - Powered By: ACS leverages a rich ecosystem of technologies and partners, integrating capabilities from leading providers such as Nvidia and Meta, alongside proprietary innovations from Atlasx.ai. This layer reflects the platform's modular and extensible nature.
  - Foundational AI Pillars: The platform is built on principles that ensure enterprise-grade AI deployment. These pillars define the nonfunctional and strategic architectural principles that underpin ACS offerings:
  - Data Sovereignty: Emphasizes data locality, control, and compliance-central to ACS's positioning as a sovereign platform in Morocco.
  - Multi-Models: Provides flexibility to run and serve various model families, including Meta, Mistral, IBM, and fine-tuned custom models.
  - BYOM (Bring Your Own Model): Allows clients to integrate their own foundation models, whether open source or licensed.
  - Hardware Agnostic: The platform supports a wide range of hardware architectures (x86, ppc64le, s390x, GPU, CPU), ensuring flexibility and cost efficiency.
  - Open and Community-oriented: Demonstrates a strong commitment to open standards and contributions to AI community practices such as fine-tuning and evaluation.
  - Custom Agents: Enables the creation of company-specific AI agents with domain knowledge and task-specific orchestration.

Together, these elements position ACS as a robust, flexible, and sovereign AI platform capable of addressing industry-specific challenges while maintaining alignment with open standards and infrastructure independence.

5.7 LLM as a Service

LLM-as-a-Service provides a scalable, centralized platform that exposes powerful foundation models through secure APIs. These APIs can be consumed by various business units and applications. Instead of building and maintaining isolated LLM stacks for each use case, this approach lets organizations manage AI models like any other strategic enterprise service-with built-in governance, observability, and cost control.

Deployed on sovereign infrastructure using IBM Fusion HCI, Red Hat OpenShift, and watsonx.ai, the service delivers foundation models (such as Granite and LLaMA) in a multi-tenant, high-performance environment aligned with national data policies and enterprise compliance requirements.

5.7.1 Model Management and Performance Optimization

One of the core challenges in LLM-as-a-Service is managing the tradeoffs between model size, performance, and resource consumption. Larger models, such as LLaMA 3 70-B, offer stronger reasoning and understanding capabilities but require significant compute and memory resources-often demanding costly multi-GPU clusters. Deciding whether to serve a general-purpose model, a fine-tuned domain-specific version, or a smaller distilled variant is a strategic decision that affects performance, latency, and cost.

Beyond model selection, lifecycle management adds complexity-from versioning and rollback mechanisms to automated updates. Ensuring that inference endpoints scale under variable load without degrading performance is a complex engineering task.

Solution:

- watsonx.ai: Supports model hosting, fine-tuning, and deployment for Granite models and open LLMs.
- Red Hat OpenShift GPU Operator: Enables dynamic GPU scheduling and multi-GPU inference.
- watsonx Model Studio: Simplifies lifecycle management and prompt tuning.

5.7.2 Security and Multi-Tenant Isolation

LLM services must support sharing across departments or clients while ensuring strict data and usage isolation. Achieving multi-tenancy in a containerized environment like Red Hat OpenShift requires well-structured namespaces, quotas, and role-based access controls. A misconfigured tenant environment can lead to prompt, log, or model endpoint leakage.

Another growing concern is prompt injection attacks, where users craft adversarial inputs to bypass business logic or extract sensitive outputs. Mitigate these risks through input validation, safety filters, and proper sandboxing of interactions.

Solution:

- Red Hat OpenShift: Uses namespaces, RBAC, LimitRanges, and NetworkPolicies to enforce per-tenant isolation.
- Keycloak (integrated): Provides SSO and identity federation with fine-grained access control.
- Istio Service Mesh: Enforces policies, isolates traffic, and secures communication between workloads.

5.7.3 Data Privacy and Sovereignty

LLMs inherently operate on data-whether during pretraining, fine-tuning, or in-context retrieval. Ensuring that data is anonymized, governed, and compliant with local regulations creates a significant operational burden. In countries like Morocco and across the EU, there is growing emphasis on data sovereignty, which requires that all data-and sometimes even AI inference-be processed within national borders.

As a result, infrastructure must be deployed on sovereign hardware, often disconnected from foreign hyperscalers. This adds complexity to hardware procurement, on-premises orchestration, and regulatory compliance.

Solution:

- IBM Fusion HCI: Provides fully on-premises infrastructure for sovereign cloud deployments.
- watsonx.ai on Red Hat OpenShift: Enables model hosting and inferencing within national borders.

5.7.4 Monitoring, Observability, and Compliance

Unlike traditional software services, LLMs are probabilistic-they generate varying outputs and may degrade subtly over time. As a result, robust observability is essential to monitor token usage, latency, throughput, and output quality.

Enterprises must also track bias, toxicity, and fairness, especially in customer-facing applications. Fine-grained logging of who used which model, for which prompt, and when is essential for audits and billing. This requires deep integration between AI services, API gateways, and telemetry systems.

Solution:

- watsonx.governance: Tracks bias, lineage, model drift, and AI usage logs.
- Red Hat OpenShift Monitoring Stack (Prometheus, Grafana): Provides observability for token consumption, latency, and performance.
- Audit logs using Red Hat OpenShift and Keycloak: Enables traceability of identity and access management (IAM) activity.

5.7.5 LLM Inference Scaling and Cost Control

Scaling LLM inference is not the same as scaling web applications. GPUs are expensive and not instantly available. Autoscaling infrastructure to handle variable demand-such as 10x load spikes during a marketing campaign-is especially challenging in on-premises or sovereign environments.

Inference engines like vLLM and DeepSpeed-MII offer performance gains, but tuning them requires deep knowledge of model internals and memory usage. Real-time concurrency management across models and users also demands careful architectural design to avoid bottlenecks or GPU starvation.

Without strong metering and cost allocation, LLM services can quickly become unjustifiable cost centers.

Solution:

- Red Hat OpenShift Horizontal and Vertical Pod Autoscaling: Dynamically scales model services based on demand.
- Red Hat OpenShift Cost Management: Provides per-namespace resource and usage metering for chargeback and cost control.

5.7.6 Integration into business processes

LLM-as-a-Service is only as valuable as its ability to be consumed by the business. Making models accessible to nontechnical users-such as HR, finance, and legal teams-requires integration into tools like Watson Orchestrate or digital worker platforms.

However, abstracting LLM capabilities into business-relevant "skills"-with defined inputs, outputs, and exception handling-is challenging. It requires combining APIs, prompt templates, governance rules, and human-in-the-loop validation in a clean, reusable way.

Solution:

- Watson Orchestrate: Connects LLM capabilities into digital workflows using reusable skills.
- watsonx.ai API: Exposes LLMs as callable REST skills integrated into business applications.

5.7.7 Data fabric as a service implementation

Data Fabric-as-a-Service enables organizations to access all relevant data-across databases, data lakes, files, and APIs-through a unified query and governance layer. With technologies such as watsonx.data, Cloud Pak for Data, and Watson Knowledge Catalog, the platform supports both real-time analytics and long-term data management strategies.

Running on sovereign infrastructure powered by IBM Fusion HCI and Red Hat OpenShift, this service abstracts complexity while enforcing data privacy, lineage, quality, and access control-empowering both IT and business users to make smarter, faster, and more trusted decisions.

5.7.8 Data integration across heterogeneous sources

One of the most fundamental challenges in building a data fabric is unifying data across diverse systems-relational databases (such as Oracle and PostgreSQL), NoSQL stores (such as MongoDB), files (CSV, JSON), data lakes, and APIs.

These sources vary in formats, schemas, update frequency, security models, and protocols. Creating a seamless federated query layer that abstracts these differences is technically demanding. It requires high-performance engines (like Spark), robust connectors, and transformation logic-without replicating data or compromising performance.

Solution:

- watsonx.data: Federates access across structured and unstructured data sources.
- Presto and Spark engines: Integrated for SQL queries and batch data processing.
- Data connectors: Provide native support for Oracle, SAP, DB2®, MinIO, and Cloud Object Storage.

5.7.9 Metadata and Governance Complexity

A true data fabric isn't just a data pipeline-it includes context, lineage, classification, and access policies. Managing metadata at scale across federated sources is a major operational challenge.

You need to ensure that data assets are:

- Discoverable using a catalog
- Tagged with business glossary terms
- Compliant with regulatory rules (such as GDPR and banking secrecy)
- Auditable end-to-end

Setting up and maintaining a system like Watson Knowledge Catalog to ingest, enrich, and govern data consistently is a complex, ongoing task.

Solution:

- Watson Knowledge Catalog: Provides cataloging, lineage tracking, data quality rules, and glossary term management.
- Integration with watsonx.data: Enables metadata-aware queries with governance enforcement.

5.7.10 Data Access Control and Policy Enforcement

In a multi-tenant or enterprise-wide data fabric, not every user should have access to every dataset. Enforcing row-level, column-level, and purpose-based access control across diverse data engines is complex.

Access policies must remain consistent whether data is accessed through notebooks, APIs, BI tools, or orchestration engines. This requires fine-grained integration between the data fabric engine, identity providers (such as Keycloak or Active Directory), and governance tools.

Solution:

- Watson Knowledge Catalog: Applies row- and column-level access policies through its policy engine.
- Keycloak IAM: Enables federated access across Red Hat OpenShift, data APIs, and notebooks.
- Red Hat OpenShift Network policies: Enforces network-level access control between services.

5.7.11 Latency and Performance Bottlenecks

Federated queries can become slow, expensive, and error-prone-especially when pushed to source systems not built for large-scale analytics. This makes query planning and optimization a top priority. You need smart caching, adaptive pushdown strategies, materialized views (such as Iceberg tables), and possibly query rewriting. Without these optimizations, the data fabric becomes a bottleneck instead of an enabler.

Solution:

- Iceberg Table Format: Enables query pushdown, partition pruning, and versioned data access.

- Spark Caching and Materialized Views: Precomputes frequently used queries to improve performance.
- Query Planner in watsonx.data: Optimizes execution paths for federated queries.

5.7.12 Scalability and Data Volume Management

As data volumes grow, storage, indexing, and compute resources must scale accordingly, especially when supporting a lakehouse with petabytes of structured and unstructured data.

Object storage engines like IBM Spectrum® Fusion must be configured for high concurrency and throughput. Iceberg or Delta tables must be intelligently partitioned, compacted, and versioned to avoid query slowdowns and data duplication.

In multi-tenant environments, resource isolation and cost tracking are essential-yet most open source data engines don't provide these capabilities natively.

Solution:

- IBM Spectrum Fusion: High-speed object storage with native support for Iceberg and S3.
- watsonx.data + MinIO: Provides a compatible object layer with data locality and replication.
- Red Hat OpenShift Storage Operators: Enable scale-out persistent volumes on demand.

5.7.13 User Enablement and Self-Service Analytics

Building the backend is only half the job. A successful data fabric must empower business users with tools to search, preview, query, and combine data-without writing code.

However, delivering true self-service is challenging:

- Interfaces must be intuitive yet powerful.
- Access policies must be enforced seamlessly.
- Training and change management are essential to help users transition from Excel or siloed reporting tools.

Solution:

- Watson Studio: Offers notebooks, drag-and-drop tools, and visual analytics for a range of user skill levels.
- WKC Search & Browse: Provides a business-friendly interface to discover, preview, and request access to data.

The following outlines a comprehensive AI and data platform architecture designed to deliver secure, scalable, and sovereign capabilities. Organized into six core pillars, each section details the underlying technology stack, key functional capabilities, and the business value it provides. Together, these pillars enable organizations to harness advanced AI, streamline data management, ensure compliance, and support national digital transformation initiatives.

- Infrastructure Foundation
  - Technology Stack: IBM Fusion HCI, Red Hat OpenShift, Spectrum Fusion
  - Key Capabilities:
    - Hyperconverged compute + storage

        - GPU-native Red Hat OpenShift
        - Unified container and VM hosting
- Business Value:
  - Sovereign, secure platform
  - Scalable and resilient
  - Cost-efficient on-prem AI infrastructure
- LLM-as-a-Service (LLMaaS)
  - Technology Stack: watsonx.ai, Watson Orchestrate, Istio, Keycloak
  - Key Capabilities:
    - Foundation model hosting (Granite, Llama)
    - Inference APIs
    - Multi-tenant isolation
    - Workflow integration
  - Business Value:
    - Instant AI access across departments
    - API monetization potential
    - No vendor lock-in
- Data Fabric-as-a-Service
  - Technology Stack: watsonx.data, Watson Knowledge Catalog, Presto, Spark, Iceberg
  - Key Capabilities:
    - Unified access to multi-format data
    - lakehouse architecture
    - Metadata and data lineage management
  - Business Value:
    - Democratized analytics
    - Zero-copy data strategy
    - Lower ETL and storage overhead
- Observability and Governance
  - Technology Stack: watsonx.governance, Red Hat OpenShift Monitoring, Keycloak, Prometheus/Grafana
  - Key Capabilities:
    - Model performance and bias monitoring
    - Query metering and SLA alerts
    - Centralized IAM and SSO
  - Business Value:
    - Compliance with AI and data policies
    - Usage accountability
    - Secure, auditable operations
- Consumption/API Layer

- Technology Stack: REST/gRPC APIs, Watson Orchestrate UI, Notebooks, BI tools
  - Key Capabilities:
    - Easy integration into business tools
    - RAG-ready architecture
    - API/skill-based orchestration
  - Business Value:
    - Fast time-to-value for users
    - Supports both technical and business personas
- Sovereignty & localization
  - Technology Stack: On-prem deployment, Moroccan infrastructure, local tuning pipeline
  - Key Capabilities:
    - Local data storage
    - Custom LLMs for Moroccan dialects
    - Compliance with national regulations
  - Business Value:
    - Trusted national AI platform
    - Supports Digital Morocco 2030
    - Data residency ensured

5.8 Introduction: The Dawn of the Conversational Data Stack

A paradox of abundance and inaccessibility defines the modern enterprise data landscape. While organizations collect more data than ever before, this information often fragments across a complex web of on-premises databases, multi-cloud platforms, data lakes, and SaaS applications. This distribution creates data silos, security risks, and significant bottlenecks that hinder timely, data-driven decision-making. In response to this challenge, two powerful technologies are creating a new architectural paradigm: Data Fabric and large language models (LLMs). This fusion promises to transform the enterprise data stack from a passive system of record into an active, conversational, and intelligent ecosystem.

5.8.1 Defining the Architectural Pillars

To understand the transformative potential of this convergence, you must first define its core architectural pillars.

Data Fabric: A data fabric is an architectural approach, not a singular product, that facilitates the end-to-end integration of diverse data pipelines and cloud environments through intelligent and automated systems. Its fundamental principle is to decouple data access from data location. Rather than physically consolidating all data via traditional Extract, Transform, Load (ETL) processes, a data fabric leaves data where it makes the most sense-in its source systems-and centralizes only the governance and metadata management layers. It primarily achieves this through two key technologies:

1. Data Virtualization: Creates an abstraction layer that connects to different sources and integrates only the necessary metadata, providing a unified, real-time view of all data regardless of where it resides.
2. Augmented Data Catalog: Serves as a centralized, intelligent inventory of all data assets. It uses metadata to help you discover, understand, and trust the data available to you. The ultimate objective of a data fabric is to achieve enterprise-wide data democratization and enable self-service analytics.
3. Data Fabric as a Service (DFaaS): Given that implementing a data fabric can be organizationally and technically challenging, requiring the assembly of numerous disparate components, many cloud providers and specialized vendors now offer it as a managed service. DFaaS lowers the barrier to entry by abstracting away the complexity of managing the underlying infrastructure, allowing you to consume data fabric capabilities in a scalable, hybrid, or multi-cloud context. This service-based model shifts the burden of component integration and maintenance to the provider, accelerating your time-to-value.
4. Large language models (LLMs): LLMs are a class of artificial intelligence models that train on vast quantities of text data. Through deep learning techniques, they develop a sophisticated ability to understand, generate, and manipulate human language to perform a wide array of tasks, including text summarization, translation, and conversational question-answering. Within the enterprise, their most crucial capability is the interpretation of unstructured data and the understanding of user intent expressed in natural language.

The symbiosis of these technologies creates a powerful new system. The data fabric provides the governed, integrated, and trustworthy data backbone-the what. The LLM acts as the intelligent, conversational interface and processing engine-the how. Together, they address the central challenge of modern data management by simplifying access to a complex and distributed data environment.

5.8.2 Data Fabric versus Data Mesh: A Clarification

The discourse on modern data architecture often involves two prominent, and sometimes confused, paradigms: data fabric and data mesh. Although both aim to solve the problems of data silos, their philosophies and approaches differ significantly.

- Data Fabric emphasizes a technology-driven, unified approach. It focuses on creating a seamless data layer through intelligent integration and automation, with central management of governance and metadata, even though the data itself remains distributed.
- Data Mesh promotes a decentralized, sociotechnical strategy. It advocates for domain-oriented data ownership, treating data as a product that cross-functional teams closest to the data manage and own.

This represents a classic conflict between centralized intelligence (fabric) and distributed ownership (mesh). However, the infusion of LLMs into the data fabric architecture may offer a resolution to this tension. An LLM-enhanced data fabric can deliver the primary user-facing benefit of a data mesh-empowering domain experts to interact with data using their own terminology and context-though retaining the core technical benefit of a data fabric: robust, centralized governance and security. As a business analyst, you can ask a question in plain English, and the LLM interface provides the answer; in the background, the data fabric's virtualization and governance layers securely resolve the query against the appropriate, controlled data sources. This synthesis offers the best of both worlds, potentially positioning the LLM-enhanced data fabric as the dominant architectural pattern in the era of generative AI. Market analysis from firms like Gartner bolsters this view, as they have questioned the long-term viability of the data mesh concept, suggesting it may become "obsolete before plateau".

The concepts of Data Fabric and Data Mesh represent two distinct approaches to modern data architecture. While both aim to maximize the value of distributed data, they differ in philosophy, governance, ownership, and focus. The following comparison highlights their core characteristics side by side.

- Philosophy
  - Data Fabric (Centralized/Automated): A unified, technology-driven architecture that creates a seamless layer over distributed data.
  - Data Mesh (Decentralized/Sociotechnical): A decentralized, sociotechnical approach focused on organizational structure and domain ownership.
- Data Ownership
  - Data Fabric: Data remains in source systems, but governance and metadata are centrally managed.
  - Data Mesh: Data is treated as a “product” owned and managed by specific business domain teams.
- Governance Model
  - Data Fabric: Centralized governance, security, and access policies enforced across the entire fabric.
  - Data Mesh: Federated computational governance with distributed autonomy for domain data teams.
- Key Enabler
  - Data Fabric: Intelligent automation, data virtualization, active metadata, and AI/ML.

- Data Mesh: Domain-oriented design, data-as-a-product thinking, and self-serve infrastructure platforms.

- Primary Focus
  - Data Fabric: Technology-centric — automating data integration and delivery to provide a unified view.
  - Data Mesh: People-and-process-centric — empowering domain teams and promoting data accountability.

5.8.3 Foundational Capabilities: The New Primitives of Data Interaction

The integration of LLMs into a Data Fabric as a Service (DFaaS) architecture is not merely theoretical; it enables a set of concrete, transformative capabilities. These foundational use cases represent a new set of primitives for how you interact with data across an enterprise, moving from rigid, technical processes to fluid, intuitive dialogs.

5.8.4 Democratizing Access through Natural Language Query (NLQ)

Problem: Historically, accessing data for analysis has been the domain of technical experts. When you want to ask questions of your data, you have had to rely on data engineers or analysts to write complex SQL queries or build static dashboards. This dependency creates significant bottlenecks, slowing down decision-making and preventing widespread data exploration.

Solution: The combination of LLMs and a data fabric directly addresses this challenge through text-to-SQL generation. This capability allows an LLM to act as a universal translator, converting user questions posed in everyday language into structured database queries that the system can run. The data fabric architecture is the critical enabler that makes this translation accurate and secure. The process works as follows:

1. You ask a question like, "What were our total sales for the top 5 products in the Northeast region last quarter?"
2. The LLM receives this natural language input. To translate it correctly, it needs context about the available data.
3. The data fabric provides this context by giving the LLM secure access to the database schema (table names, column definitions, and their relationships).
4. A semantic layer within the fabric maps technical database terms to common business concepts (for example, mapping a column named cust_rev_q1 to the business term "Quarter 1 Customer Revenue"), which helps the LLM correctly interpret your intent.
5. Crucially, the fabric's data virtualization layer presents a single, consistent SQL interface to the LLM, even if the underlying product, sales, and customer data reside in completely different systems (for example, a cloud data warehouse, an on-premises ERP system, and a SaaS CRM).
6. The LLM uses this context to generate a precise SQL query, which the fabric's query engine then runs. The system returns the results to you, often in a summarized or visualized format.

Value Proposition: This capability profoundly democratizes data access. It empowers you-from sales managers to marketing executives-to conduct self-service analytics, test hypotheses, and gain insights in real time without technical intermediaries. This accelerates the pace of data-driven decision-making across the entire organization. Leading platforms

like IBM's watsonx.data highlight this as a core feature, offering a "conversational interface to easily find, augment and visualize data... no SQL required".

5.8.5 Automating Insight with Intelligent Data Catalogs

Problem: An augmented data catalog is a cornerstone of a data fabric, serving as the "shop window" for your organization's data assets. However, missing, outdated, or poorly written metadata often severely limits the utility of these catalogs. Manually creating and maintaining high-quality descriptions for thousands of tables and columns is an unscalable task for your data stewardship teams, rendering the catalog difficult to search and its assets untrustworthy.

Solution: LLMs offer a powerful solution by automating the creation and enrichment of metadata at scale. This generative AI-powered metadata curation transforms the data catalog from a static inventory into a dynamic knowledge base. Key techniques include:

- Automated Description Generation: LLMs can analyze the structure of a data asset (its schema, column names, data types) and even sample data to automatically generate clear, human-readable descriptions for tables and columns.16 For instance, an LLM can infer that a column cryptically named ytd_dist_amt should be described as “year to date distribution amount.”
- Retrieval-Augmented Generation (RAG): To ensure the generated descriptions are not just generic but are grounded in the specific context of your enterprise, the system employs the RAG technique. Before generating a description, the LLM retrieves relevant information from a trusted corporate knowledge base, such as business glossaries, technical design documents, or examples of other well-documented data assets. This grounds the LLM's output, dramatically improving its accuracy and relevance while preventing "hallucinations.”
- Automated Tagging and Classification: Beyond descriptions, LLMs can automatically scan data and classify it, for example, by identifying columns that contain Personally Identifiable Information (PII) or by applying business-relevant tags. This enhances both the searchability of the catalog and the enforcement of governance policies.

Value Proposition: An AI-enriched catalog makes data discovery dramatically more efficient and intuitive. It improves the overall usability and trustworthiness of your data assets, significantly reduces the manual workload on your data stewards, and fosters a shared, consistent understanding of data across your enterprise.

5.8.6 Ensuring Trust with AI-Driven Data Quality (DQ)

Problem: The adage "garbage in, garbage out" has never been more relevant than in the age of AI. LLMs and other machine learning models are exquisitely sensitive to the quality of their training data; feeding them inaccurate, inconsistent, or biased data will inevitably lead to unreliable models, flawed insights, and untrustworthy outcomes. Traditional data quality tools, which are often based on rigid, manually defined rules, struggle to validate complex and unstructured data types like free-text documents, images, or audio.

Solution: LLMs can function as a sophisticated data quality engine, performing context-aware checks that were previously impossible to automate. This moves data quality beyond simple validation to a deeper, semantic level of integrity assessment.

- Validating Unstructured Data: LLMs excel at understanding the content of unstructured data. You can prompt them to analyze free-text fields to check for specific quality attributes. For example, an LLM can read through thousands of customer reviews to flag

those containing toxic language, or it can verify that a product's technical description includes all legally required safety warnings.

- Semantic Anomaly Detection: LLMs can identify outliers based on meaning, not just statistical deviation. You might use a multimodal LLM to scan a dataset of product images and flag a picture of a cat in a catalog of dog food, a semantic error that traditional tools would miss.
- Automated DQ Rule Generation: In a powerful application, you can give an LLM a sample of high-quality data and ask it to infer the deterministic rules (often as SQL queries) that define its quality. It can then generate an entire suite of data quality tests automatically, which you can apply to the full dataset.

Value Proposition: AI-driven DQ is fundamental to building the trusted data foundation that you require for enterprise-grade AI. It ensures that the data flowing through your fabric is reliable, consistent, and fit for purpose, which is essential for training accurate models and building your trust in AI-driven insights. This capability is a cornerstone of achieving the "enterprise data readiness" you need for generative AI, a key objective for platforms like IBM watsonx.

These three foundational capabilities create a powerful, self-reinforcing cycle. Natural Language Query lowers the barrier to data interaction, driving you and more users to the data and increasing demand. This increased demand heightens the need for an effective data catalog; if you cannot find the right data to ask questions about, the value of NLQ is lost. AI-powered cataloging meets this need by making assets easily discoverable and understandable for you. As you and more users and applications rely on this data, its quality becomes paramount to maintain trust. AI-driven data quality ensures the data you discover and query is reliable, which in turn builds your confidence and encourages even greater use of the system via NLQ. This feedback loop of demand, discovery, quality, and trust is the engine that drives true data democratization.

However, this automation introduces a new imperative. Although LLMs can generate queries, metadata, and quality rules, they are also prone to errors, such as factual hallucinations, subtle biases, or security vulnerabilities. Simply accepting AI-generated outputs without oversight is a significant risk, which security frameworks like the OWASP Top 10 for LLMs identify as "insecure output handling". The solution is to embed a "human-in-the-loop" protocol, where human experts treat AI outputs as suggestions for review and approval. This signals a fundamental evolution in the role of the data professional: a shift away from the manual labor of building pipelines and writing queries, and toward the higher-value work of being a curator, validator, and overseer of an AI-driven data ecosystem.

5.9 Platform in Practice: An Architectural Blueprint with IBM watsonx

To move from abstract architectural concepts to concrete implementation, you can examine how a major technology provider packages and delivers an LLM-enhanced DFaaS platform. The IBM watsonx platform, comprising three core components, serves as a practical blueprint for how providers implement these capabilities in a commercial offering.

The Foundation - watsonx.data: The Hybrid, Open Data lakehouse

Watsonx.data serves as the foundational data store and access layer of the fabric, embodying the principle of accessing all data, wherever it resides, through a single point of entry. It is the component that builds the governed, AI-ready data backbone.

5.10 Capabilities:

Hybrid and Open Architecture: Hybrid and Open Architecture: You can run Watsonx.data anywhere-on-premises, on IBM Cloud, or on other public clouds like AWS. It uses open data formats (for example, Apache Iceberg) and open source engines, which helps you avoid vendor lock-in and share a single copy of data across different analytics workloads without costly duplication. This hybrid nature is a strategic acknowledgment that enterprise data is and will remain distributed.

Multi-Engine Query Federation: A key feature is its multi-engine architecture, which allows you to pair the right workload with the most cost-effective and performant query engine, such as Presto for interactive queries or Spark for large-scale data processing.15 This directly supports the data fabric goal of optimizing workloads and can reduce data warehouse costs significantly.

AI-Ready Data Preparation: Watsonx.data includes integrated capabilities for creating vectorized embeddings from data. This is a crucial preprocessing step for enabling Retrieval-Augmented Generation (RAG) use cases, directly preparing the data for the intelligence layer.

5.11 watsonx.ai: The enterprise studio for AI builders

The Enterprise Studio for AI Builders watsonx.ai provides the comprehensive suite of tools, models, and computing resources you need to build, tune, and deploy generative AI applications on top of the trusted data that watsonx.data provides.38 It is the platform's intelligence engine.

Capabilities:

Foundation model Library: Watsonx.ai offers you access to a curated library of foundation models trained by IBM, as well as popular open source models. It also provides the flexibility for you to upload and deploy your own custom models, ensuring you can use the right tool for your specific use case.

Prompt Engineering and Tuning: The platform includes tools like the "Prompt Lab" where you can experiment with prompt engineering, create reusable prompt templates, and evaluate different model and prompt combinations. Crucially, it allows you to fine-tune these models on

your enterprise's own proprietary data, sourced securely from the watsonx.data fabric, to improve accuracy and align the model's responses with specific business contexts.

Seamless Integration: Developers design the studio for tight integration with watsonx.data, ensuring that you have seamless access to the high-quality, governed data required for all stages of the AI lifecycle, from training and tuning to production inferencing with RAG.

The Framework - watsonx.governance: The Trust and Compliance Layer

Watsonx.governance addresses the critical need for risk management, transparency, and compliance for all AI models operating within the fabric. It acts as the trust and safety framework for the entire platform.

Capabilities:

End-to-End AI Monitoring: It provides you with tools to monitor both traditional machine learning models and new generative AI assets throughout their lifecycle. It evaluates models against specified thresholds for performance metrics like fairness, accuracy, and drift. For LLMs specifically, you can configure monitors to detect risks such as the presence of hateful language or the inadvertent leakage of PII in model outputs.

AI Factsheets for Transparency: This feature automatically captures and centralizes critical metadata for each AI model into a comprehensive "factsheet." This creates a full audit trail, documenting who built the model, what data it was trained on, its performance evaluations, and its stage in the lifecycle. This transparency is vital for explaining model behavior and meeting regulatory compliance goals.

Integrated Risk Management: Watsonx.governance integrates with enterprise risk management tools like IBM OpenPages® to develop and enforce governance workflows, assess the risk level of AI use cases, and ensure that all AI development and deployment aligns with internal policies and external regulations.

The tripartite structure of the watsonx platform-Data, AI, and Governance-is a direct commercial implementation of the virtuous cycle of data democratization. Watsonx.data provides the unified, high-quality data foundation (enabling discovery and ensuring quality). Watsonx.ai consumes this data to power intelligent applications that drive user demand. Finally, watsonx.governance wraps around the entire process to manage risks and ensure responsible deployment, thereby building the trust necessary for widespread adoption. This integrated stack demonstrates that a mature DFaaS offering must be more than a loose collection of services; its value lies in the seamless flow of data, models, and metadata between the data, AI, and governance layers.

The LLM-Fabric architecture is built on three core components, each serving a distinct purpose within the ecosystem. Together, they provide the data foundation, AI intelligence, and governance framework needed to deliver trusted, scalable, and high-performing large language model solutions. Below is an overview of each component and its key capabilities.

- watsonx.data — The hybrid, open data lakehouse; the data foundation of the fabric
  - Core Function: Serves as the unified data layer for structured and unstructured sources.
  - Key Capabilities:
    - Unified Access: Single point of entry to all data via virtualization and connectors.
    - Multi-Engine Optimization: Matches workloads with fit-for-purpose engines (for example, Presto, Spark) to balance cost and performance.

        - AI-Ready: Integrated vector embedding capabilities to prepare data for retrieval-augmented generation (RAG).

- watsonx.ai — The enterprise studio for AI builders; the intelligence engine of the fabric
  - Core Function: Provides tools and models for building, tuning, and deploying AI applications.
  - Key Capabilities:
    - Foundation models: Access curated IBM and open source models, or bring your own.
    - Prompt Engineering & Tuning: Experiment with prompts and fine-tune models using proprietary fabric data.
    - Conversational AI: Build natural language query (NLQ) and other generative AI applications.
- watsonx.governance — The trust and compliance layer; the governance framework for the fabric
  - Core Function: Ensures AI models operate within ethical, regulatory, and quality standards.
  - Key Capabilities:
    - End-to-End Monitoring: Tracks fairness, drift, accuracy, and LLM-specific risks such as PII leakage.
    - AI Factsheets: Automatically documents the model lifecycle and metadata for transparency and auditing.
    - Risk Management: Integrates with risk platforms to enforce policies and maintain regulatory compliance.

5.12 Navigating the Labyrinth: Challenges, Risks, and Ethical Imperatives

Although the fusion of LLMs and data fabric presents a compelling vision for the future of enterprise data, implementing it presents significant challenges. Successfully navigating this new landscape requires a clear-eyed understanding of the technical hurdles, AI-specific risks, and profound ethical responsibilities involved.

5.13 Technical and Operational Hurdles

The Data Quality Prerequisite: The principle of "garbage in, garbage out" is dangerously amplified in the context of LLMs. The performance, accuracy, and safety of the entire enhanced fabric architecture hinge on the quality of the underlying data. LLMs trained or grounded on inconsistent, erroneous, or biased data will inevitably produce unreliable and potentially harmful outputs. The sheer scale of modern datasets, which can contain trillions of tokens, makes comprehensive manual assessment of data quality nearly impossible, forcing reliance on automated filtering techniques that may not catch subtle issues.

Architectural Complexity: Even when delivered "as-a-Service," a data fabric remains a complex system of systems. Integrating the fabric with a multitude of legacy platforms, managing intricate data dependencies, and ensuring consistent performance across a distributed, hybrid-cloud environment are substantial technical challenges that require specialized expertise.

Scalability and Performance: LLM training and inference are notoriously compute-intensive processes. The network architecture that supports them differs fundamentally from traditional data center designs. LLM training, which often involves parallel processing across hundreds or thousands of GPUs, generates massive volumes of "east-west" traffic (server-to-server communication). This requires an ultra-fast, lossless network fabric capable of handling synchronous, bursty data flows, a significant departure from networks optimized for "north-south" (user-to-server) traffic.

5.14 Generative AI-Specific Risks

The Open Worldwide Application Security Project (OWASP) has identified a top 10 list of vulnerabilities specific to LLM applications, many of which are directly relevant to the data fabric context.

Large language models (LLMs) introduce a range of risks that must be addressed to ensure trustworthy, compliant, and secure AI deployment. Within the LLM-Fabric, these risks are categorized by type, with specific mitigation strategies embedded into the architecture. The following outlines the primary risk categories, their descriptions, affected use cases, and how the fabric mitigates them.

- Factual Integrity
  - Specific Risk: Hallucination
  - Description: LLM generates plausible but factually incorrect or nonsensical information.
  - Primary Use Cases Affected: NLQ, Automated Reporting, Research
  - Mitigation Strategy within the Fabric:

        - Retrieval-Augmented Generation (RAG): Ground LLM responses in a trusted, curated knowledge base provided by the data fabric.
- Security
  - Specific Risk: Prompt Injection
    - Description: Malicious users craft inputs to bypass safety filters or run unintended commands (for example, generating malicious SQL).
    - Primary Use Cases Affected: NLQ, Code Generation
    - Mitigation Strategy: Input sanitization and human-in-the-loop validation; treat LLM-generated code or queries as suggestions requiring human approval.
  - Specific Risk: Sensitive Information Disclosure
    - Description: LLM inadvertently leaks sensitive data (PII, trade secrets) from its training set or prompt context.
    - Primary Use Cases Affected: All use cases, especially in Healthcare and Finance.
    - Mitigation Strategy: Robust governance with strict access controls; use AI monitors (for example, watsonx.governance) to detect PII in outputs.
- Bias & Fairness
  - Specific Risk: Algorithmic Bias
    - Description: LLM inherits and amplifies societal biases from its training data, leading to discriminatory outcomes.
    - Primary Use Cases Affected: Credit Scoring, Hiring Recommendations, Medical Diagnosis
    - Mitigation Strategy: Fine-tune models on balanced, representative enterprise data from the fabric; continuously audit outputs for fairness.
  - Specific Risk: Training Data Poisoning
    - Description: Malicious actors contaminate the training data to compromise model integrity or introduce vulnerabilities/biases.
    - Primary Use Cases Affected: All use cases relying on third-party models.
    - Mitigation Strategy: Use private/trusted models; fine-tune only on carefully vetted enterprise data managed by the fabric.
- Compliance
  - Specific Risk: Copyright & Plagiarism
    - Description: LLM generates content that infringes on copyrights or plagiarizes protected material.
    - Primary Use Cases Affected: Content Generation, Report Summarization
    - Mitigation Strategy: Document provenance using the fabric's data lineage and metadata tools (for example, AI Factsheets) to trace data and content origins.

5.15 Ethical and Governance Mandates

Beyond technical risks, the deployment of LLM-enhanced data fabrics carries profound ethical responsibilities.

Algorithmic Bias: As noted, LLMs can absorb and scale societal biases present in their vast training data. This can lead to deeply inequitable and discriminatory outcomes in critical

applications like loan approvals, medical diagnoses, or recruitment screening. Mitigating this requires a proactive commitment to curating diverse and representative datasets, implementing fairness-aware algorithms, and continuously auditing model behavior.

Data Privacy and Copyright: The methods that train large foundation models, which often involve scraping immense swathes of the public internet, raise serious legal and ethical questions about user consent, data privacy, and copyright infringement. Enterprises leveraging these models must be acutely aware of the provenance of their AI tools and the potential for legal and reputational liability.

Accountability and Transparency: The "black box" nature of deep learning models makes it extremely difficult to explain why an LLM produced a specific output. This lack of interpretability creates a major barrier to adoption in regulated industries, as it complicates efforts to assign accountability when an AI system makes a harmful error. This necessitates the adoption of governance tools like AI Factsheets, which create a transparent audit trail for a model's entire lifecycle.

Data Justice: A truly ethical approach must go beyond simply mitigating bias and consider the principles of "data justice." This framework examines the power imbalances inherent in the AI ecosystem: Who funds and builds these models? Whose data is being used, often without their knowledge or consent? And who benefits from their deployment? These questions are critical to ensuring that this powerful technology does not further entrench existing inequalities or harm marginalized communities.

This analysis reveals a key consideration: the "governance paradox." The very LLM capabilities that democratize data access and unlock immense value, such as NLQ, simultaneously create an explosion in the governance and security challenge. As the number of users who can query data expands from a few hundred data professionals to tens of thousands of business users, the "attack surface" for data governance grows exponentially. Each new user and every natural language query represents a potential vector for error, misuse, or a security breach. This paradox means that the technology making the fabric more powerful also makes it more vulnerable. Therefore, the governance layer itself must become intelligent and automated. Manual governance processes cannot possibly scale to meet this challenge. An organization needs AI to effectively govern AI, which is precisely the value proposition of AI-powered governance solutions.

5.16 Finally: The Future of Data as a Dialogue

The convergence of large language models and Data Fabric architecture represents not an incremental improvement, but a fundamental paradigm shift in how you interact with your data. This synthesis transforms the corporate data stack from a passive, rigid system of record into an active, intelligent, and conversational ecosystem. The analysis has shown that this fusion democratizes access through Natural Language Query, automates insight through intelligent data catalogs, and ensures trust through AI-driven data quality checks. These capabilities are unlocking tangible value in data-intensive industries like finance, healthcare, and manufacturing, solving long-standing problems related to regulatory compliance, personalized services, and operational efficiency.

However, a new class of challenges accompanies this powerful new paradigm. The success of an LLM-enhanced fabric depends on solving the immense challenge of data quality at scale. It introduces novel security vulnerabilities and amplifies ethical concerns around bias, privacy, and accountability that demand robust, AI-powered governance frameworks. The path forward requires a strategic, human-in-the-loop approach where you evolve from builders of infrastructure to curators and validators of an intelligent system.

The trajectory of this technology is clear. The future sees an even deeper integration of generative AI within the data fabric itself, automating not just data access but also core data operations like pipeline generation, workload optimization, and system self-healing. With market projections showing a massive expansion from approximately $3.5 billion in 2025 to over $17 billion by 2032, investment in and adoption of these integrated platforms will accelerate dramatically.

Ultimately, the vision for the LLM-enhanced data fabric is to make interacting with an organization's entire, complex data landscape as simple and intuitive as having a conversation with a trusted, omniscient, and expert colleague. This architecture provides the blueprint for building that future, transforming data management from a series of technical procedures into a continuous, intelligent dialogue that drives business at the speed of insight.

6

Applications of IBM watsonx: Use Cases in AI Innovation

In today's fast-paced digital landscape, organizations across industries are facing increasing need for innovation, adapt, and transform. As the volumes of data continue to grow, and the complexity of business challenges increases, the need for intelligent, efficient, and scalable solutions has never been more pressing.

IBM IBM watsonx platform is a comprehensive AI and data management platform that empowers organizations to unlock the full potential of their data, drive innovation, and achieve significant business success. Several organizations adopted watsonx with its advanced technology and comprehensive capabilities, to overcome complex challenges, improve efficiency, and create new opportunities for growth.

This chapter explores how IBM IBM watsonx platform is transforming industries through different use cases that are addressed by watsonx. We examine how organizations leverage watsonx to unlock insights from unstructured data, enhance AI governance, and accelerate innovation in diverse sectors such as finance, government, and cloud services.

In this chapter, we delve into use case of IBM watsonx in action. We explore how this powerful platform helps organizations in different sectors to transform their businesses, overcome challenges, and achieve remarkable success.

We examine how a leading enterprise used watsonx to unlock business-critical information in unstructured enterprise data, gaining actionable insights and improving decision-making. We will look at how Atlas Cloud Services leveraged watsonx to accelerate AI innovation in Morocco, driving growth and competitiveness in the region.

We showcase use cases that highlights the capabilities of watsonx in the government sector, demonstrating how this platform can be used to drive public sector innovation, improve citizen services, and cfreate a more sustainable future.

We explore how watsonx.governance addresses critical AI governance challenges in two highly regulated sectors: financial services and the rail industry.

© Copyright IBM Corp. 2025.

Through these case studies, we aim to provide a comprehensive understanding of IBM watsonx, its capabilities, and its potential to modernize operations, transform businesses and organizations across industries. Whether you are a business leader, a technologist, or a policy-maker, this chapter provides you with valuable insights and inspiration to unlock the power of AI and data in your own organization.

At the end of the chapter, you get a hand-on exercise to try the platform on unstructured data.

6.1 IBM watsonx in Action

In this section, we delve into the use case where IBM watsonx can be used and case study of IBM watsonx in action. We explore how this powerful platform can help organizations in different sectors to transform their businesses, overcome challenges, and achieve remarkable success.

6.2 Use case: Unlocking Business-Critical Information in Unstructured Enterprise Data

6.2.1 Business Context and Challenge

As digital transformation accelerates, enterprises are generating vast volumes of unstructured data-contracts, reports, emails, forms, and more. This data often contains essential business knowledge, yet remains underutilized due to its complexity and lack of structure. Traditional data warehouses and lakes struggle to process this data effectively, limiting the potential of generative AI in areas such as:

- Automated question answering
- Regulatory compliance
- Operational decision-making

Organizations face a critical gap: they have the data, but not the means to extract value from it at scale.

6.2.2 Solution: IBM watsonx.data and lakehouse architecture

IBM watsonx.data addresses this challenge by combining the scalability of data lakes with the performance of data warehouses-creating a lakehouse architecture optimized for AI workloads. This unified platform enables seamless access to both structured and unstructured data, making it AI-ready.

Key capabilities include:

- Dual-Channel Retrieval Mechanism: Combines semantic search powered by LLMs with metadata-based retrieval, improving accuracy for semi-structured documents by over 40%.
- Augmented Data Catalogs: Automatically tag and classify documents using AI, enabling faster discovery and governance.
- Natural Language Interfaces: Allow users to query complex documents conversationally, without needing technical expertise.

6.2.3 Example Use Cases

- Legal & Compliance: Extract clauses from contracts, identify regulatory risks, and ensure auditability.
- Customer Service: Summarize call transcripts, detect unresolved issues, and improve response quality.
- Operations: Search across maintenance logs and incident reports to accelerate root cause analysis.

Outcomes

By deploying IBM watsonx.data within a governed data fabric, organizations can:

- Unlock Hidden Value: Extract insights from unstructured data without manual reformatting.
- Boost Accuracy: Improve retrieval precision for semi-structured documents by 40%+.
- Accelerate Decisions: Enable business users to self-serve insights using natural language.
- Ensure Compliance: Maintain traceability and governance across all data interactions.

6.3 Case Study: Atlas Cloud Services Accelerates AI Innovation in Morocco

6.4 Business Context and Challenge

In a bold move to drive national digital transformation, Atlas Cloud Services (ACS) partnered with IBM and Power Maroc to build a sovereign AI and cloud platform aligned with the Digital Morocco 2030 strategy. This case study showcases how the alliance leveraged the full IBM watsonx stack-including watsonx.ai, watsonx.data, and watsonx.governance-alongside IBM Fusion HCI and Red Hat OpenShift to deliver a secure, high-performance, and locally governed AI infrastructure.

The platform empowers Moroccan enterprises across sectors-finance, public services, manufacturing, and telecom-with capabilities such as LLM-as-a-Service, data fabric-as-a-service, and AI-powered automation. By ensuring data sovereignty, multi-model flexibility, and ethical AI governance, ACS is not only enabling innovation within Morocco but also setting a blueprint for responsible AI adoption across North Africa.

A key challenge in this context is ensuring platform sovereignty-storing and processing data within national borders while delivering world-class AI capabilities. This requires a carefully architected technology stack that combines high performance with compliance and localization.

Technical Solution Architecture

IBM IBM watsonx.data platform is becoming the core platform for intelligent enterprise transformation. By enabling deep mining, unified governance, and intelligent retrieval of unstructured data, the platform improves the question-answering quality of existing generative AI agents and expands the boundaries of data asset utilization.

With ongoing enhancements to the GPU-accelerated Velox engine, SaaS model integration, and compatibility with the Power architecture, IBM watsonx.data platform is positioned to become the unified data and AI platform for the next generation of intelligent enterprises-driving organizations from data-driven to AI-driven.

The strategic collaboration between Atlas Cloud Services (ACS), IBM, and Power Maroc (PowerM) aims to accelerate AI-driven digital transformation across Morocco and the broader North African region. The initiative supports the goals of the "Digital Morocco 2030" strategy by enabling businesses to adopt AI-powered solutions through a sovereign, hybrid, and multi-cloud platform tailored to the local market's needs.

A key pillar of this initiative is platform sovereignty, which ensures that ACS hosts all data and AI workloads within Moroccan borders, under Moroccan governance, and in alignment with national data protection laws. ACS's sovereign cloud is built on IBM infrastructure and software technologies, including IBM Power10 servers, IBM Fusion HCI, Red Hat OpenShift, watsonx.ai, watsonx.data, Cloud Pak for Data, Watson Orchestrate, and Watson Code Assistant. This stack ensures a high-performance, secure, and resilient environment for AI operations.

The technology stack integrates:

- IBM watsonx.ai for building and deploying foundation models and enterprise AI applications.
- watsonx.data as an open, governed lakehouse for AI workloads.
- Cloud Pak for Data for end-to-end data lifecycle and AI lifecycle management.
- watsonx Code Assistant® to accelerate application modernization and software development.
- Watson Orchestrate for business process automation through AI agents.
- rKube.IO, PowerM's modernization and DevSecOps hub, to orchestrate AI, security, and transformation pipelines.
- IBM Power10 systems for AI and data-intensive computing with enhanced inferencing capability.
- IBM Fusion HCI for hyper-converged infrastructure supporting private and hybrid cloud models.
- Red Hat OpenShift as the container orchestration platform standardizing AI workloads and scaling across clouds.

This integrated stack enables Moroccan organizations to modernize securely while preserving national data sovereignty and enabling local control over AI governance.

6.5 Industry-specific use cases covered by ACS Platform

6.5.1 Use Case 1: AI-Powered Customer Service Automation

Industry: Banking, Telecom, Utilities

Problem:

- High customer service volume
- Inconsistent and slow response times
- Difficulty scaling support operations

Solution:

- Deploy IBM Watson® Orchestrate to automate service ticketing, customer inquiries, and backend workflows
- Use watsonx.ai for NLP-based understanding of customer queries in Arabic, French, and English
- Integrate chatbots into call centers and digital platforms (web/app)
- Use rKube.IO to connect AI assistants to legacy CRM systems

Outcomes:

- Faster resolution times
- 24/7 multilingual support
- Reduced agent workload and training time
- Enhanced customer satisfaction and retention

6.5.2 Use Case 2: Smart HR and Recruitment Automation

Industry: Public Sector, Corporations, Education

Problem:

- Manual recruitment cycles
- High attrition and unstructured HR data

Solution:

- Use AI models in watsonx.ai to evaluate CVs, identify top talent, and score candidates
- Integrate with internal HRMS via rKube.IO
- Use Watson Orchestrate for onboarding workflows (account creation, policy training)
- Use watsonx.data to ensure secure access to employment history and performance records

Outcomes:

- Reduced time-to-hire
- Improved talent retention
- Transparent, bias-mitigated selection process
- Data compliance with Moroccan labor laws

6.5.3 Use Case 3: Manufacturing Optimization through Predictive Maintenance

Industry: Phosphate, Automotive, Textile, Agritech

Problem:

- Reactive maintenance culture
- Equipment failures affecting productivity

Solution:

- Stream sensor data to Cloud Pak for Data
- Deploy predictive models on Red Hat OpenShift running on IBM Power10
- Use watsonx.data dashboards for performance monitoring
- Trigger AI alerts and automate technician dispatch with the right parts

Outcomes:

- Reduced downtime
- Optimized maintenance costs
- Real-time insights into plant operations

6.5.4 Use Case 4: Real-Time Fraud Detection

Industry: Finance, Insurance, FinTech

Problem:

- Complex, evolving fraud tactics
- Siloed data sources across departments

Solution:

- Train LLMs on historical transaction data using watsonx.ai
- Implement anomaly detection models via watsonx.data
- Enable unified data access across compliance, risk, and operations
- Deploy within Atlas Cloud for full data residency and sovereignty

Outcomes:

- Fraud detection latency under seconds
- Adaptive risk-scoring system
- 360° view of transactions across platforms
- Compliance with GDPR and Moroccan banking regulations

6.5.5 Use Case 5: Application Modernization with Developer AI Assistance

Industry: Cross-sector (GovTech, Retail, Logistics)

Problem:

- Legacy codebases and outdated applications (Java, COBOL)
- Low developer efficiency

Solution:

- Use watsonx Code Assistant to refactor and modernize legacy code
- Migrate workloads to Red Hat OpenShift using rKube.IO modernization blueprints
- Provide code suggestions and auto-documentation within CI/CD pipelines
- Run compute-intensive builds on IBM Power10 with Fusion HCI

Outcomes:

- Reduced time-to-modernization
- Enhanced developer productivity
- Future-proof, microservice-ready architecture

6.5.6 Use Case 6: AI-Driven Supply Chain Intelligence

Industry: Logistics, Retail, Manufacturing

Problem:

- Delays in sourcing and procurement
- Lack of inventory visibility

Solution:

- Use AI to forecast demand, delivery delays, and supplier risk
- Integrate with ERP systems using rKube.IO connectors
- Deliver visualizations and alerts through dashboards on Cloud Pak for Data

Outcomes:

- Inventory optimization
- Significant reduction in stockouts
- Supplier diversification and proactive sourcing

Strategic Benefits to Morocco

- Boost in sovereign AI capabilities hosted within ACS
- Local data processing and storage on Moroccan soil
- Training and employment in AI and cloud computing through UM6P partnerships
- Enhanced economic competitiveness through digital innovation
- Scalable AI foundation for regional expansion into North and West Africa
- Alignment with national sovereignty and data governance priorities

6.6 Use Case 3: AI Governance in Public Sector with watsonx.governance

IBM watsonx.gov provides an end-to-end platform for governing, managing and monitoring AI activities. It ensures regulatory compliance, transparency, risk management and responsible AI deployment for both traditional machine learning models and generative AI (gen AI) models.

6.7 Public Sector - Benefits Claim Processing

A government agency implemented an automated system to process and approve welfare benefit claims (for example pensions, unemployment, subsidies). The system uses machine learning (ML) and natural language processing (NLP) models for document verification, eligibility evaluation and fraud detection. It is important to:

- Ensure compliance with data privacy laws and ethical AI standards
- Monitor for bias, fairness and drift in the AI model output to prevent unjustified approvals or rejection of claims
- Provide transparent and explainable decisions and comply with audits
- Streamline approval, deployment and monitoring of the AI models and risk management

6.8 Benefits of watsonx.gov

- Provides documented decisions to help agency staff and citizens understand "why" a claim was approved or denied
- Provides tools and workflows that reduce regulatory risk and loss of public confidence with privacy, fairness and audit trail
- Provides operational efficiency through centralized AI model lifecycle management
- Provides scalable governance by supporting AI expansion across departments and ensuring high standards

6.8.1 Resume summarization

In this use case, a human resources department would like to use a large language model (LLM) to create brief summaries of job applicant resumes to help deal with a large volume of applications. Summarization is one of the most common use cases for LLMs. However, this use case can pose some risks:

- Resumes traditionally contain personal data, some of which may be considered sensitive
- Hallucinations and other incorrect summarizations might negatively affect the perception of a candidate's viability
- Any attempt to introduce AI into hiring decisions will always be met with extra scrutiny

6.8.2 Hiring recommendation

In this use case, a human resources department would like to use a traditional machine learning (ML) model to assist in hiring decisions. Candidates are scored in a variety of categories, and the model attempts to predict which ones will be successful employees based on those scores. This is a binary classification model. This model has the following risks associated with it:

- Models of this nature are at risk for unfair bias based on biased training data
- Decisions made by the model can impact hiring decisions, and are subject to higher scrutiny
- Auditors are likely to request that any decisions the model made be explained
- Any attempt to introduce AI into hiring decisions will always be met with extra scrutiny

6.8.3 Credit risk

In this use case, the corporate finance department would like to evaluate applications for credit cards using a traditional machine learning (ML) model. This model has the following risks associated with it:

- Models of this nature are at risk for unfair bias based on biased training data
- Decisions made by the model can impact financial decisions, and are subject to higher scrutiny
- Auditors are likely to request that any decisions the model made be explained
- Financial data used by the model is considered highly confidential

6.8.4 Auto insurance risk

In this use case, an insurance company would like to evaluate the riskiness of auto insurance policies based on a number of factors, including age and gender of the driver, make and model of the vehicle, previous insurance claims, and proximity of the driver's home address to locations where traffic accidents frequently occur. This model has the following risks associated with it:

- Models of this nature are at risk for unfair bias based on biased training data
- Decisions that reject policies as too risky will likely be subject to audits for explainability
- Personal information such as home addresses will be used in the model
- Indirect bias may exist when judging distance to areas with frequent traffic accidents, if those areas are near neighborhoods with high minority populations

6.8.5 Human resources chatbot

In this use case, the company's human resources department would like to use retrieval-augmented generation (RAG) to answer questions regarding employee benefits, compensation and health insurance. This model has the following risks associated with it:

- The model will accept direct input from users, so data poisoning attacks are possible
- The model will need to guard against any hateful, aggressive, and profane (HAP) speech
- The model must be monitored to ensure it does not unwittingly leak personal identifiable information (PII) for other users

- Hallucinations are possible, so the model must be able to provide citations for relevant information

6.9 Use case 4: watsonx.governance in Financial Services and Rail

To address complex AI governance challenges, both financial institutions and rail operators are leveraging watsonx.governance to ensure responsible, transparent, and compliant AI adoption. The solution enables centralized model lifecycle management, automated monitoring, policy enforcement, and explainability-tailored to each sector's regulatory and operational needs.

Outcomes include improved risk management, regulatory readiness, operational efficiency, and scalable, ethical AI practices across diverse environments.

Business Challenges in Financial Services

Financial institutions are increasingly adopting AI and machine learning to enhance decision-making, automate operations, and improve customer experiences. However, this rapid expansion introduces several critical challenges:

- Fragmented AI Ecosystems: Models are deployed across multiple cloud platforms and embedded in third-party applications, making centralized oversight difficult.
- Limited Governance Capabilities: Existing tools often lack the depth to manage model lifecycle, enforce governance policies, and monitor performance across environments.
- Regulatory Pressure and Uncertainty: Institutions must prepare for evolving AI regulations that demand transparency, accountability, and ethical use of AI in financial decision-making.
- Operational and Reputational Risk: Without robust governance, models may be biased, noncompliant, or poorly documented-leading to financial loss and reputational damage.
- Difficulty Operationalizing Responsible AI Principles: While many institutions have defined ethical AI frameworks, they struggle to embed these principles into day-to-day model development and deployment.

Solution: watsonx.governance

To address these challenges, financial institutions are turning to Watsonx.governance, a purpose-built AI governance solution designed to support responsible, scalable, and compliant AI adoption.

Key Capabilities of watsonx.governance:

- Model Lifecycle Management:Enables end-to-end tracking of AI models from development to retirement, ensuring visibility and control at every stage.
- Automated Documentation and Monitoring: Generates audit-ready documentation and continuously monitors model performance, drift, and compliance metrics.
- Policy Enforcement and Risk Controls: Applies governance policies consistently across environments, helping institutions meet internal standards and external regulatory requirements.
- Explainability and Transparency: Provides tools for model explainability, bias detection, and lineage tracking to support ethical and defensible AI use.

- Integration with Responsible AI Frameworks: Aligns with organizational principles for ethical AI, enabling institutions to operationalize values such as fairness, accountability, and transparency.

6.10 Achieving desirable outcomes

By implementing watsonx.governance, financial institutions aim to achieve:

- Enterprise-Wide Risk Management: Centralized oversight of AI model risk across business units and technical environments.
- Regulatory Readiness: Confidence in meeting current and future regulatory requirements related to AI governance.
- Operational Efficiency and Trust: Streamlined governance processes that reduce manual effort and increase stakeholder trust in AI systems.
- Ethical and Responsible AI Adoption: Embedding ethical principles into AI workflows to ensure models are fair, transparent, and aligned with institutional values.
- Scalable Governance Frameworks: A flexible foundation that supports growth in AI adoption without compromising control or compliance.

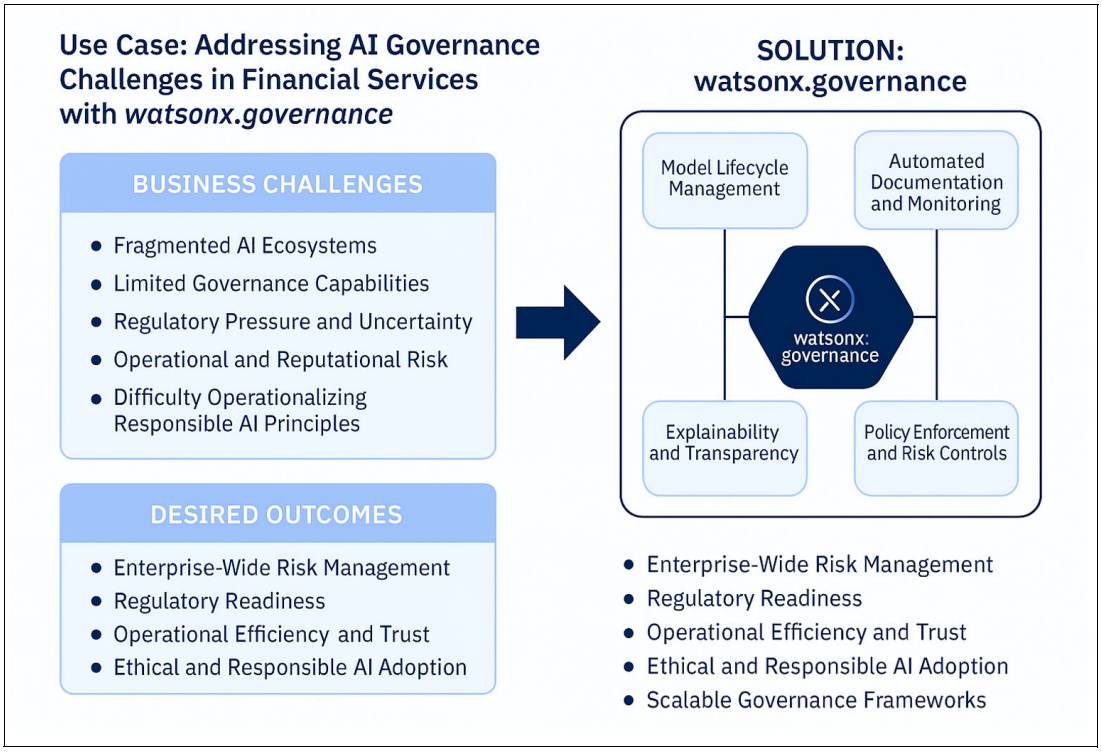


Figure 6-1 watsonx.governance use case in financial services

6.10.1 Business Challenges in Rail

Rail operators are increasingly adopting AI and machine learning to optimize operations, improve safety, and enhance passenger experience. However, this transformation introduces several governance challenges:

- Safety-Critical AI Applications: AI models are used in predictive maintenance, scheduling, and safety monitoring-areas where errors can have serious consequences.
- Distributed Data and Model Environments: Models are deployed across edge devices, control centers, and cloud platforms, making governance complex and fragmented.
- Regulatory and Compliance Pressures: Rail operators must comply with transportation safety regulations, data protection laws, and emerging AI-specific legislation.
- Lack of Transparency and Accountability: Operational teams often lack visibility into how AI models make decisions, especially in automated systems like signaling or fault detection.
- Difficulty Operationalizing Ethical AI: While many rail organizations have adopted principles around safety, fairness, and transparency, embedding these into technical workflows remains a challenge.

6.10.2 Solution: watsonx.governance

Watsonx.governance provides a comprehensive framework to address these challenges by enabling rail organizations to:

- Manage Model Lifecycle Across Environments?Track and govern models deployed in control systems, cloud platforms, and edge devices.
- Automate Documentation and Monitoring?Ensure models are continuously monitored for drift, performance degradation, and compliance violations.
- Enforce Safety and Regulatory Policies?Apply governance controls aligned with transportation safety standards and data protection regulations.
- Enable Explainability and Auditability?Provide clear insights into model decisions, supporting accountability in safety-critical applications.
- Operationalize Ethical AI Principles?Embed fairness, transparency, and safety into model development and deployment workflows.

6.10.3 Desired Outcomes

By implementing watsonx.governance, rail operators aim to:

- Ensure Safe and Reliable AI Operations: Minimize risk in safety-critical systems through robust governance and monitoring.
- Achieve Regulatory Compliance: Meet current and future regulatory requirements with confidence and transparency.
- Improve Operational Efficiency: Streamline governance processes and reduce manual oversight through automation.
- Build Trust in AI Systems: Enhance stakeholder confidence in AI-driven decisions across operations and passenger services.

- Scale AI Responsibly: Support innovation while maintaining control and accountability across the AI lifecycle.

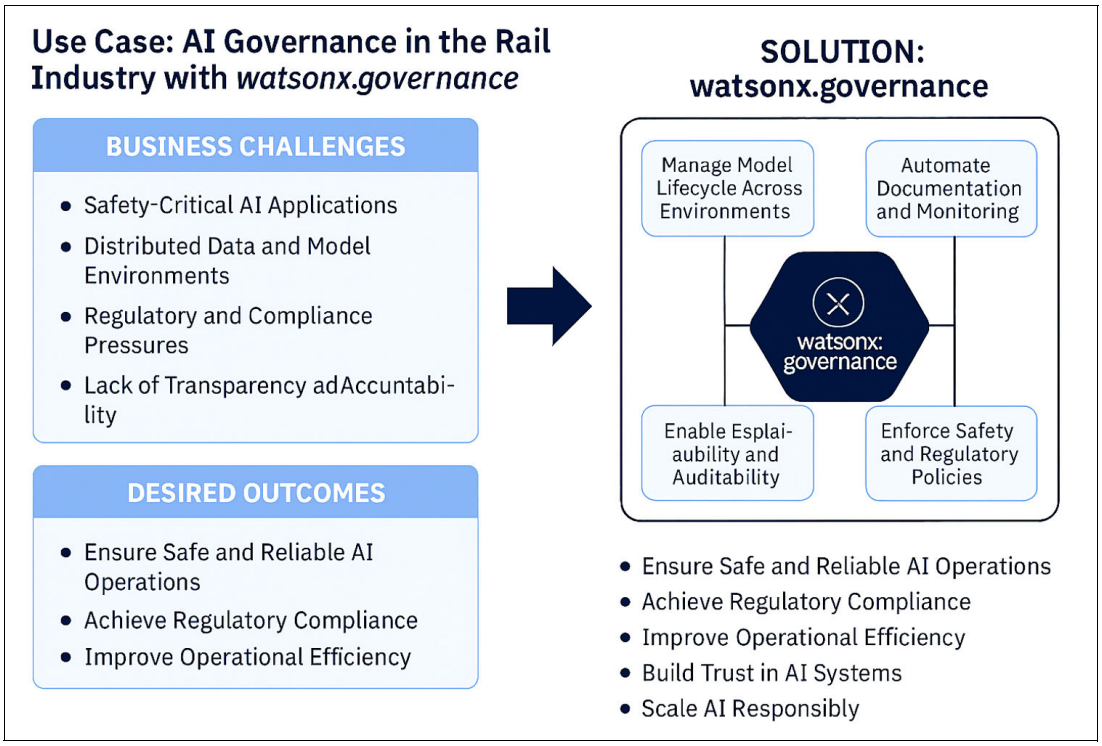


Figure 6-2 watsonx.governance use case in Rail Industry

6.11 Exercise: Using watsonx.data for Structured Understanding and Vector Retrieval of PDF Invoices

This hands-on exercise demonstrates how watsonx.data transforms unstructured PDF invoices stored in AWS S3 into structured data and semantic vectors to support Retrieval-Augmented Generation (RAG) applications.

Summarization of the workflow steps:

1. Create a project in watsonx.data
2. Use your IBMid credentials to login to your watsonx.data instance and create a new project
3. Connect to AWS S3
4. Securely connect to an S3 bucket containing PDF invoices.
5. Ingest and Enrich Flow
6. Ingest PDF files and enrich them by:
   a. Storing semantic vectors in Milvus.
   b. Storing structured data in Iceberg tables (queryable using Presto).
7. Automated Tasks (8 Total):
   a. Read files, extract entities, chunk documents, generate vectors, classify data, match terminology, and export to Milvus and Iceberg.
8. Generated Data Assets:
   a. Document Set, Entity Table, ACL Table, Enrichment Metadata, Smart Retrieval and RAG Support
9. Enables semantic + SQL search:
   a. Example: "What invoices exceed $10,000 in the past quarter?" triggers:
10. Vector search in Milvus.
11. SQL query in Presto.
12. Results are fed into an LLM for accurate, context-rich answers.

6.12 Step 1: Create a project in watsonx.data

- Log in to watsonx.data
- Go to your watsonx.data instance or your internal Cloud Pak for Data environment
- Use your IBMid or enterprise credentials.
- Navigate to the Projects Section
- From the main dashboard, click on "Projects" in the navigation menu.
- Select "Create a project".
- Choose Project Type
- Select "Analytics project" or "Data engineering project" depending on your use case.

- For unstructured data processing and enrichment (like the PDF invoice case), choose "Analytics project".
- Configure Project Details
- Name your project (for example, Invoice Processing with watsonx.data).
- Optionally add a description and tags.
- Choose a storage location (Cloud Object Storage or an existing bucket).
- Add Collaborators (Optional)
- You can invite team members to collaborate on the project.
- Create the Project
- Click **Create** to initialize the project workspace.

6.13 Step 2: Create an AWS S3 Connection

- In your project, first create an AWS S3 connection.
- You provide the bucket name and IAM certificate, and the system accesses the bucket through a secure connection.
- The bucket stores many invoice files in PDF format, organized by year or customer number.

6.14 Step 3: Build Ingest and Enrich Flow

Create an Ingest and Enrich Unstructured Data flow in your project to perform document ingestion and semantic enhancement.

- Source: Select the AWS S3 connection you just created. Specify the file paths and set the file format to PDF.
- Target: Configure output for two objectives:
- Store the semantic vectors of document chunks in the Milvus vector database.
- Store structured business data in an Iceberg table under the Presto engine for querying and analysis.

When you save the flow, the system automatically generates an executable flowchart that shows the data transfer logic.

6.15 Step 4: Eight Tasks Automatically Generated by the Flow

After you save and submit the flow, the system automatically performs the following tasks:

- Read the source file: Reads PDF files sequentially from AWS S3 and streams them to a local buffer for processing.
- Structural analysis and entity extraction: Uses an intelligent document parsing model to analyze the structure and layout of each PDF and extract key-value pairs (for example,

invoice number, date, amount, supplier). Simultaneously reads and stores the access control list (ACL) of the source file for later output to the permission control table.

- Chunking: Splits each document into multiple chunks based on logical structures such as paragraphs, tables, and invoice entries.
- Vector generation: Uses embedding models such as BERT or watsonx embedding to encode each chunk into a high-dimensional vector.
- Entity identification and data classification: Matches extracted entities to data classes (for example, date, amount, ID) to support downstream data governance.
- Terminology matching and grading: Matches entities to business terminology definitions in the enterprise dictionary (for example, "amount payable," "invoicing date"). Grades sensitive data to support audit and compliance control.
- Export vector data to Milvus: Writes the vectors and associated metadata (chunk ID, document ID, source) to the Milvus vector database.
- Output structured entities to Iceberg table: Writes all extracted entities, document metadata, and ACLs to a Presto-queryable Iceberg table, generating structured entity tables, ACL tables, and document indexes.

6.16 Step 5: Generate Data Assets

After the flow runs successfully, you can view the following generated objects in Data Assets:

- Document Set: A processed PDF document and its version history.
- Entity Table: Structured field data extracted from invoices.
- ACL Table: A record of access rights for each document.
- Enrichment Metadata: Includes classification labels, terminology matches, vector chunk mappings, and more.

These assets support enterprise applications such as regulatory analysis, reconciliation verification, and audit traceability.

6.17 Step 6: Support for Smart Retrieval and RAG Applications

With the integrated Retrieval Service, you can enable users to perform semantic searches using natural language.

A user might enter: "What invoices exceed $10,000 in the past quarter?"

The search service converts the query into two forms:

- Vector query: Sent to Milvus to locate the most relevant document chunks.
- SQL query: Sent to Presto to filter structured entity tables accurately.

The system feeds the search results back to the LLM (such as a watsonx.ai foundation model) as part of the prompt, significantly improving fact-based responses.

This solution demonstrates how watsonx.data transforms massive volumes of unstructured documents into searchable, embeddable, and manageable data assets that support intelligent Q&A, compliance audits, and business insights:

- Automates end-to-end data ingestion, governance, and analytics

- Builds RAG workflows that support both vector search and SQL queries
- Enhances enterprise semantic understanding, data security, and generative AI capabilities

6.17.1 Outcomes and Benefits

This exercise demonstrates how watsonx.data transforms unstructured documents into searchable, embeddable, and manageable data assets that support:

- Intelligent Q&A with improved accuracy
- Compliance audits with proper governance
- Business insights from previously inaccessible data

By enabling deep mining, unified governance, and intelligent retrieval of unstructured data, the platform improves question-answering quality of generative AI agents and expands the boundaries of data asset utilization.

6.18 Activating the Intelligent Data Ecosystem: Use Cases for LLM-Enhanced Data Fabric

6.18.1 Industry Spotlight: Real-World Applications and Value Creation

The foundational capabilities of an LLM-enhanced data fabric are not just abstract technical primitives; they are powerful tools that you apply to solve concrete, high-value business problems across a range of industries. By unifying disparate data and providing an intelligent interface, this architecture unlocks new efficiencies and insights in sectors with complex data challenges.

Financial Services: Fortifying Governance and Insight

Massive data volumes, a complex and ever-changing regulatory landscape, and the constant threat of sophisticated fraud define the financial services industry. An LLM-enhanced data fabric provides a secure, governed, and intelligent platform uniquely suited to addressing these challenges.

The following use cases illustrate how an LLM-enhanced data fabric can address complex business challenges by combining governed, unified data access with intelligent language model capabilities. Each example outlines the business problem, the solution enabled by the fabric, and the key value proposition delivered.

Automated Regulatory Intelligence

- Business Problem: Manually tracking, interpreting, and implementing ever-changing regulations (for example, Basel IV, CSRD) is slow, costly, and prone to compliance errors.
- LLM-Enhanced Fabric Solution: Intelligent agents within the fabric monitor regulatory sources. LLMs interpret new rules, compare them to internal policies, and generate impact assessments. The fabric's lineage ensures traceability.
- Key Value Proposition: Reduced compliance risk and cost; faster adaptation to regulatory change; auditable proof of compliance.

Advanced Fraud Detection

- Business Problem: Fraudsters use complex, multi-channel schemes that evade traditional, rule-based detection systems.

- LLM-Enhanced Fabric Solution: The fabric unifies transaction data, customer communications, and unstructured notes in real-time. LLMs analyze this holistic view to detect subtle anomalies and interpret text for hidden risks.
- Key Value Proposition: Higher fraud detection rates; prevention of financial loss; improved security for customers.

Agentic Investment Research

- Business Problem: Analysts spend excessive time gathering and summarizing data from financial reports, news feeds, and earnings calls to assess risk and opportunity.
- LLM-Enhanced Fabric Solution: The fabric provides governed access to internal and external data. An LLM agent can be tasked to research a company, automatically summarizing reports, analyzing news sentiment, and generating a detailed risk profile.
- Key Value Proposition: Accelerated research cycles; deeper, more comprehensive insights; more informed and agile investment decisions.

Cross-Process Narrative Governance

- Business Problem: Maintaining consistent tone, reasoning, and justification across numerous reports (capital plans, ESG disclosures, board communications) is a major challenge.
- LLM-Enhanced Fabric Solution: An “LLM-as-a-Judge” mechanism compares narratives across documents stored in the fabric, ensuring alignment with policy and historical precedent, and flagging inconsistencies for SME review.
- Key Value Proposition: Enhanced disclosure consistency and quality; reduced manual review effort; stronger corporate governance.

Intelligent Scenario Validation

- Business Problem: Aligning new stress test scenarios (for example, for ICAAP) with historical regulatory positions and internal assumptions is a complex, manual process.
- LLM-Enhanced Fabric Solution: LLM agents compare proposed scenarios against a fabric-accessible knowledge base of past submissions and models, identifying gaps and recommending revisions to ensure alignment.
- Key Value Proposition: More robust and defensible stress testing; improved alignment with regulatory expectations; faster scenario development.

Life Sciences: Accelerating Personalized Medicine and Research

Healthcare data is notoriously siloed across Electronic Health Records (EHRs), laboratory systems, medical imaging archives, and patient-worn devices. This fragmentation is a major obstacle to delivering personalized care, accelerating clinical research, and managing operations efficiently. A data fabric creates a unified, longitudinal view of this data, which LLMs can then analyze to uncover profound insights.

The following healthcare-focused use cases demonstrate how an LLM-enhanced data fabric can unify fragmented data, enable advanced analytics, and deliver actionable intelligence to clinicians, researchers, and healthcare administrators. Each example includes the business problem, the LLM-fabric-enabled solution, its key value proposition, and supporting snippet counts.

Personalized Patient 360° View

- Business Problem: Clinicians lack a complete, holistic view of a patient’s history, which is fragmented across multiple systems, hindering personalized treatment.

- LLM-Enhanced Fabric Solution: The fabric virtualizes and connects records from EHRs, labs, and patient apps. An LLM summarizes the complete patient journey into a concise narrative for the clinician at the point of care.
- Key Value Proposition: Improved diagnostic accuracy; more effective, personalized care plans; better patient outcomes.

Clinical Trial Acceleration

- Business Problem: Integrating and analyzing diverse data sources (for example, trial protocols, patient data, adverse event reports) is a major bottleneck in drug development.
- LLM-Enhanced Fabric Solution: The fabric provides a unified data layer for trial data. LLMs help identify eligible participants faster, analyze unstructured notes, and monitor for adverse events in real-time.
- Key Value Proposition: Accelerated time-to-market for new therapies; reduced trial costs; more efficient trial management.

Hospital Operations Optimization

- Business Problem: Inefficiencies in patient flow, staff scheduling, and supply chain management lead to increased operational costs and staff burnout.
- LLM-Enhanced Fabric Solution: The fabric connects disparate operational systems (admissions, billing, inventory). LLMs analyze this unified data to predict admission rates, optimize schedules, and identify bottlenecks.
- Key Value Proposition: Reduced operational costs; improved resource allocation; enhanced patient experience.

Enhanced Insurance Fraud Detection

- Business Problem: Detecting fraudulent claims requires correlating information across claims data, medical records, and payment histories, which are often in separate systems
- LLM-Enhanced Fabric Solution: The fabric connects these disparate datasets. AI models powered by LLMs analyze the unified data to spot unusual patterns, such as billing for procedures inconsistent with a patient's clinical history.
- Key Value Proposition: Reduced financial losses from fraud; lower insurance premiums; protection of the healthcare system.

Medical Literature Summarization

- Business Problem: Healthcare professionals struggle to keep up with the vast and rapidly growing body of medical literature.
- LLM-Enhanced Fabric Solution: LLMs can process and summarize large volumes of medical literature, clinical records, and scientific papers, providing clinicians with quick access to the latest evidence-based findings.
- Key Value Proposition: Faster access to critical information; improved evidence-based decision-making; streamlined workflows.

Manufacturing: Building the Sentient Factory Floor

The manufacturing sector faces a unique challenge in bridging the divide between Information Technology (IT) systems (like ERP and CRM) and Operational Technology (OT) systems (like sensors, SCADA, and engineering diagrams). This IT/OT gap prevents a holistic view of the factory floor, hampering efforts in predictive maintenance, quality control, and supply chain management. An industrial data fabric, enhanced with domain-specific LLMs, contextualizes and unifies this data.

The following industrial use cases demonstrate how an LLM-enhanced data fabric can unify operational technology (OT) and information technology (IT) data, enabling predictive insights, real-time decision-making, and improved operational efficiency. Each example outlines the business problem, the solution, and the key value proposition.

Predictive Maintenance

- Business Problem: Unforeseen equipment failures result in costly production downtime and reactive, inefficient maintenance cycles.
- LLM-Enhanced Fabric Solution: An industrial data fabric contextualizes real-time OT sensor data with IT maintenance logs and engineering manuals. LLMs analyze this rich data to predict failures before they happen.
- Key Value Proposition: Reduced unplanned downtime; lower maintenance costs; increased asset lifespan and operational efficiency.

Automated Quality Control

- Business Problem: Identifying product defects and quality deviations often relies on manual inspection, which is slow and error-prone.
- LLM-Enhanced Fabric Solution: The fabric provides real-time production data to LLMs, which can recognize patterns and anomalies in product specifications, automatically flagging deviations from quality standards.
- Key Value Proposition: Improved product quality; reduced waste and rework; proactive quality assurance.

Supply Chain Demand Forecasting

- Business Problem: Lack of real-time visibility across the supply chain leads to inaccurate demand forecasts, resulting in inventory stockouts or overages.
- LLM-Enhanced Fabric Solution: The fabric integrates data from suppliers, logistics, and sales. LLMs analyze this data alongside external market trends to generate more accurate and dynamic demand forecasts.
- Key Value Proposition: More resilient and efficient supply chain; optimized inventory levels; reduced carrying costs.

Worker Safety Enhancement

- Business Problem: Complex industrial environments pose significant safety risks to workers.
- LLM-Enhanced Fabric Solution: By analyzing real-time sensor data and historical incident reports within the fabric, an LLM can identify potential hazards and issue real-time alerts to workers or power immersive safety training simulations.
- Key Value Proposition: Reduced workplace accidents; improved worker well-being; stronger safety culture.

IT/OT Data Contextualization

- Business Problem: Technicians on the factory floor lack easy access to the full context of an asset, which is split across OT data streams and IT documents.
- LLM-Enhanced Fabric Solution: The fabric creates a “digital twin” of assets. A technician can use an LLM-powered conversational interface to ask, “Show me the maintenance history and common failure points for this pump.”
- Key Value Proposition: Faster root cause analysis; empowered frontline workers; bridging of the IT/OT data divide.

6.19 Summary

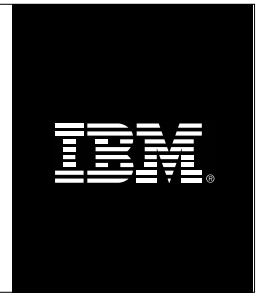
Across these diverse industries, a common thread emerges: the LLM's ability to process and contextualize vast amounts of unstructured data (regulatory text, clinician's notes, engineering manuals) and integrate it with the structured data managed by the fabric unlocks the primary value. This fusion of quantitative and qualitative information allows organizations to analyze their entire data estate holistically for the first time.

However, these applications critically depend on domain-specific grounding. A generic, internet-trained LLM cannot provide reliable regulatory advice or medical diagnoses. The data fabric's most crucial role, therefore, is to serve as the "AI immune system" for the enterprise. It provides the trusted, curated, and domain-specific knowledge base needed for techniques like RAG, ensuring that the LLM reasons over the company's single source of truth rather than the noisy, unreliable public internet. This positions the data fabric not merely as a data access layer, but as an essential risk mitigation platform for deploying generative AI safely and effectively in the enterprise.

IBM Fusion HCI as a Catalyst: Accelerating Growth in

SG24-8600-00

ISBN

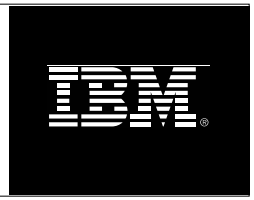


(1.5” spine)
1.5”<-> 1.998”
789 <->1051 pages

IBM Fusion HCI as a Catalyst: Accelerating Growth in

SG24-8600-00

ISBN



(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

IBM Fusion HCI as a Catalyst

SG24-8600-00

ISBN

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

IBM Fusion HCI as a Catalyst

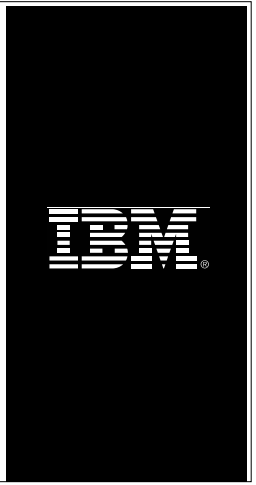
(0.2”spine)
0.17”<->0.473”
90<->249 pages

(0.1”spine)
0.1”<->0.169”
53<->89 pages

IBM Fusion HCI as a Catalyst: Accelerating

SG24-8600-00

ISBN

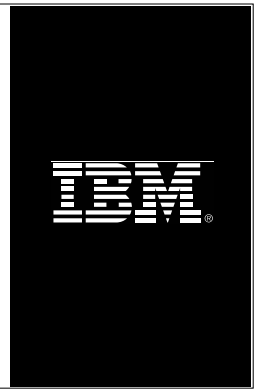


(2.5” spine)
2.5”<->nnn.n”
1315<-> nnnn pages

IBM Fusion HCI as a Catalyst: Accelerating Growth in watsonX.ai, watsonX.data, and

SG24-8600-00

ISBN



(2.0” spine)
2.0” <-> 2.498”
1052 <-> 1314 pages

SG24-8600-00

ISBN 0738462306

Printed in U.S.A.

Get connected