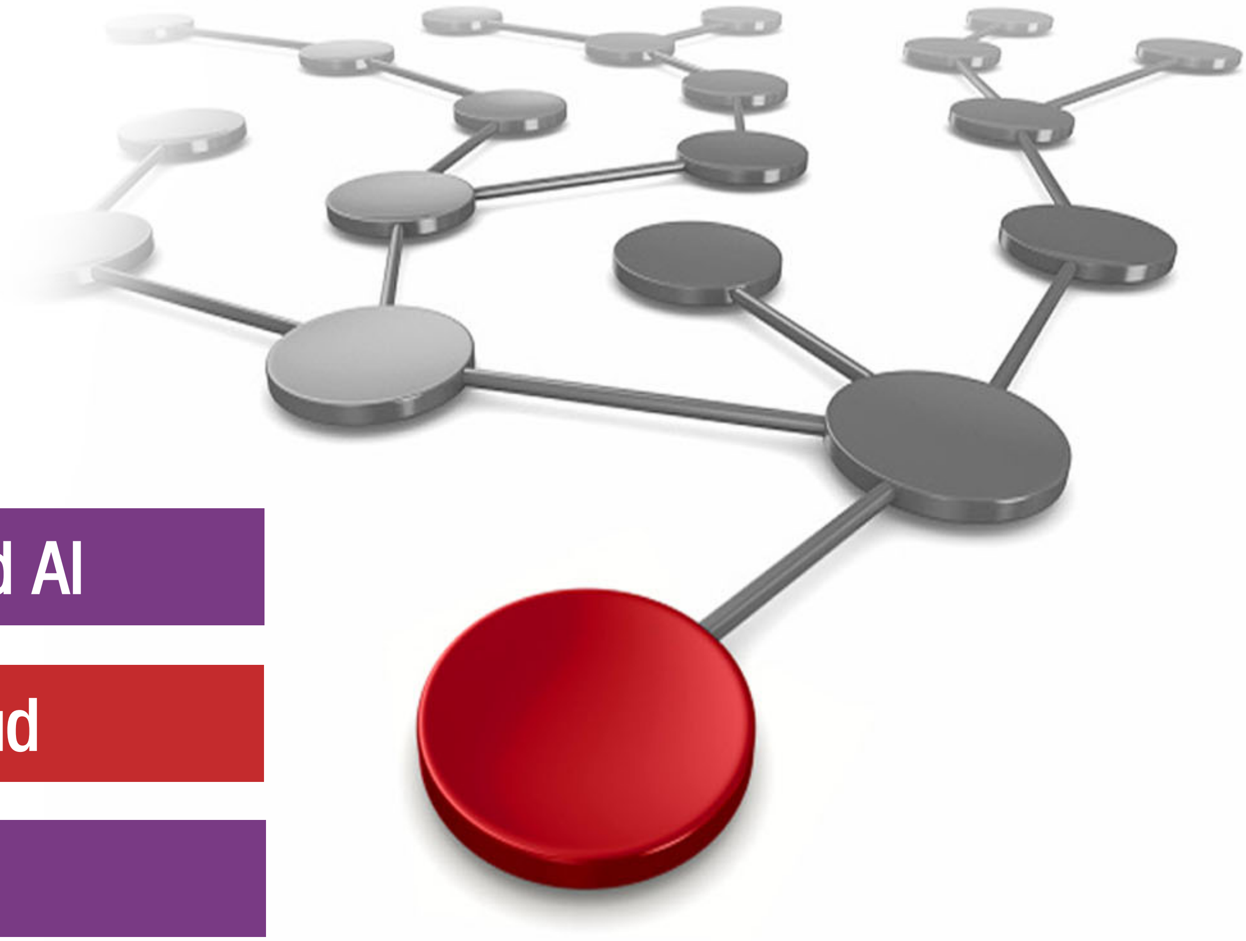
Redbooks
ibm.com/redbooks

Accelerating IBM watsonx.data with IBM Fusion HCI

Larry Coyne
Paulina Acevedo
Eduardo Daniel Ibarra Alvarez
Gabriela Valencia Castillo
Kenneth Hartsoe
Mike Kieran
Alberto Larios
Savitha H N
Khanh Ngo
AshaRani G R
Shyamala Rajagopalan
Ben Randall
Hemalatha B T
Todd Tosseth
Jayson Tsingine
Israel Vizcarra
Zhi Yong Xue



Data and AI

Cloud

Storage

IBM

Redpaper

IBM Redbooks

Accelerating IBM watsonx.data with IBM Fusion HCI

March 2024

REDP-5720-00

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

First Edition (March 2024)

This edition applies to Version 2, Release 7, Modification x of IBM Fusion HCI

Contents

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.

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®
DS8000®
Enterprise Storage Server®
IBM®
IBM Cloud®
IBM Cloud Pak®
Redbooks®
Redbooks (logo) ®
XIV®

The following terms are trademarks of other companies:

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

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

Preface

Organizations that are expanding from AI pilot projects to full-scale production systems typically need a set of tools for building and deploying foundation models, a container-based application platform, software-defined storage, and hardware on which to run it all. This IBM Redpaper publication describes the IBM® solution for running IBM watsonx.data on premises, with IBM Fusion HCI providing an appliance-based hosting platform, and IBM Storage Ceph providing cloud-scale object storage.

This publication shows how to set up the Storage Acceleration feature, so IBM watsonx.data queries can benefit from a shareable on-premises high-performance cache acceleration. The Storage Acceleration feature is available only on an IBM Fusion HCI.

This paper is targeted toward technical professionals, consultants, technical support staff, IT Architects, and IT specialists who are responsible for delivering data lakehouse solutions optimized for data, analytics, and AI workloads.

Authors

This paper was produced by a team of specialists from around the world working with the IBM Redbooks, Tucson Center.

Larry Coyne is a Project Leader at the IBM International Technical Support Organization, Tucson, Arizona, center. He has over 35 years of IBM experience, with 23 years in IBM storage software management. He holds degrees in Software Engineering from the University of Texas at El Paso and Project Management from George Washington University. His areas of expertise include client relationship management, quality assurance, development management, and support management for IBM storage management software.

Paulina Acevedo is a System Test Architect for the Application and Resiliency Fusion Team. Paulina has been with IBM for more than 17 years and has held several different positions within the Systems organization. She is a certified Project Manager and has been the System Test Project manager for several IBM Storage products.

Eduardo Daniel Ibarra Alvarez is a Systems Test Engineer for the Application and Resiliency Fusion Team. Eduardo has a strong automation and testing background.

Gabriela Valencia Castillo is a System Test Engineer for the Application and Resiliency Fusion Team. Gabriela has extensive experience in functional and non-functional testing, as well as test plan creation, design scenarios, and creation of test cases.

Kenneth Hartsoe is the documentation Content Strategist for IBM Storage Ceph and IBM Fusion Data Foundation, as well as providing content strategy collaboration across multiple solutions within IBM Storage. Kenneth has more than twenty years experience within the storage documentation area, both as a senior technical writer and content strategist, including several years as the storage content strategist at Red Hat.

Mike Kieran is a product marketer currently focusing on Storage for IBM watsonx.data. He has more than a decade of storage marketing experience at Pure Storage, NetApp, and Nimble Storage. Mike is the author of four books on digital color and the recipient of an Emmy for science and technology writing. He is an avid stargazer and astrophotographer.

Alberto Larios is a System Test Engineer for the Application and Resiliency Fusion Team. Alberto collaborates with his team as a systems associate and tests Cloud Pak products on IBM Storage. His expertise lies on the data science field with knowledge on machine learning and modelling.

Savitha H N is a System Test Engineer for the Application and Resiliency Fusion Team. Savitha has been with IBM from past one year and is a systems associate with Cloud Pak. She holds a Devops engineer certification, and specialist in Q/A testing by executing scenarios to Ensures the product is robust and failure scenarios are considered and refactored.

Khanh Ngo is a leader in the IBM Storage CTO office specializing in Data and AI integration with IBM Storage products including the optimization of IBM watsonx.data.

AshaRani G R is a System Test Engineer for the Application and Resiliency Fusion Team. Asha possesses extensive expertise encompassing server hardware, management software, and storage solutions within SAN and DAS domains. She has expertise in bringing up compute nodes for Infrastructure as a Service and performing end-to-end system testing. Additionally, she has a strong background in virtualization technology.

Shyamala Rajagopalan is the senior lead technical content and information architect for IBM Fusion product. She has over 24 years of industry experience, with a significant track record of successfully delivering end-to-end IT documentation solutions to global clients across diverse industries. Her area of contribution includes digital transformation, legacy modernization, integration platforms (middleware, EAI), semiconductor, Cloud, and storage systems. She has performed diverse roles in the information development domain, which includes Information Architect, Technical Writing Manager, Lead Content Developer, Senior technical writer, and Competency Area Mentor.

Ben Randall is the User Experience Architect for IBM Fusion and works on product development and design. He has worked in the enterprise storage industry for 21 years, focusing on technologies such as disaster recovery, backup and restore, container native storage, software defined storage, high performance computing, and SAN monitoring.

Hemalatha B T is a System Test lead for Application and Resiliency Fusion Team. Hema has been with IBM for over 15 years and was associated with Power System Performance before moving to the functional/system test area working for products like IBM Cloud Pak® for Systems, IBM Fusion. An enthusiastic quality assurance person who thrives to ensure high quality and perfectly functioning systems are delivered to customers.

Todd Tosseth is a Software Engineer for IBM in Tucson, Arizona. Joining IBM in 2001, he has worked as a test and development engineer on several IBM storage products, such as IBM DS8000®, IBM Storage Scale, and IBM Enterprise Storage Server®. He is working on IBM Cloud Pak as a system test engineer, with an emphasis on Cloud Pak storage integration.

Jayson Tsingine is an Advisory Software Engineer in the IBM Systems Storage Group based in Tucson, Arizona. Jayson has worked in numerous test roles since he joined IBM in 2003, providing test support for storage products, including IBM XIV®, DS8000, FlashSystems, Hydra, Spectrum Scale, Spectrum NAS, and Cloud Pak Solutions. He holds a BS degree in Computer Science from the University of Arizona.

Israel Vizcarra is a test specialist that works for the Cloud Pak Storage Test Team. He graduated as a Mechatronics Engineer and he has 8 years of experience in the Quality Assurance area for the storage organization. Israel has actively participated in different roles from Functional Verification, System level, and Automation Testing.

Zhi Yong Xue is an Architect of Fusion Storage in China. He has 15 years of experience in software design and development as a developer and architect at IBM. He holds a Bachelor's degree in Exploration Technology and Engineering from the University of Petroleum. His areas of expertise include Storage and Cloud computing.

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

Patrik Hysky
IBM Redbooks®, IBM Infrastructure

Pramod Thekkepat Achutha, Tara Astigarraga, Scott Colbeck, Karli Collins, Marcel Hergaarden, Ted Hoover, Lisa Huston, Kedar Karmarkar, Moushumi Kalita, Joshua Kim, Ana Lilia Sanchez Llamas, Ajay Lunawat, Boda Devi Manikanta, Kevin Shen, Bill Stoddard, Chris Tan, Thiha Than, Garth Tschetter, Hans Uhlig
IBM Infrastructure

David Wohlford
IBM CHQ, Marketing

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 papers to be as helpful as possible. Send us your comments about this paper 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

Solution overview

Organizations that are expanding from AI pilot projects to full-scale production systems typically need the following components:

- A set of tools for building and deploying foundation models
- A container-based application platform
- Software-defined storage
- Hardware on which to run it

This publication describes the IBM solution for running IBM watsonx.data on premises, with IBM Fusion HCI providing an appliance-based hosting platform, and IBM Storage Ceph providing cloud-scale object storage. This publication shows how to set up the Storage Acceleration feature, which is only available on IBM Fusion HCI, so IBM watsonx.data queries can benefit from a shareable on-premises, high-performance cache acceleration.

This paper is targeted toward technical professionals including consultants, technical support staff, IT Architects, and IT specialists who are responsible for delivering optimized for data, analytics, and AI workloads.

This chapter includes an overview covering the background of data lakes and how the IBM solution of IBM watsonx.data, IBM Storage Ceph, and IBM Fusion HCI accelerated infrastructure works to improve on-premises performance and improves cost efficiency. The architecture of the solution and components are also described.

1.1 Overview

This section describes the evolution of data lakes, the emergence of data lakehouses, and IBM watsonx.data lakehouse, IBM Storage Ceph, and the IBM Fusion HCI accelerated infrastructure solution.

From data warehouse to data lake

During the past 20 years, large organizations have changed the way they aggregate data for analytics and business intelligence (BI) purposes. The original approach was to build a single monolithic database, or data warehouse, and then analyze specific subsets of the data through an extract, transform, load (ETL) process based on queries by using structured query language (SQL).

Data warehouses are often used for repeatable reporting and analysis workloads such as monthly sales reports, tracking of sales per region, and website traffic. But building and maintaining a data warehouse is a costly, time-consuming process, and data warehouses work only with structured data.

Moving data warehouses to the cloud doesn't solve the problem. Sometimes, it makes them even more expensive, and they're still not well suited to machine learning or AI applications.

These limitations led to the concept of the data lake, which is a centralized repository that can store massive volumes of data in its original form so that it's consolidated, integrated, secure, and accessible. Data lakes are designed to accommodate all types of data from many different sources:

- Structured data, such as database tables and Excel sheets
- Semi-structured data, such as herbages and XML files
- Unstructured data, such as images, video, audio, and social media posts

Because data lakes are massively scalable and can handle all types of data, they are ideal for real-time analytics, predictive analytics, and machine learning or AI. They are also typically less costly than data warehouses.

Data lakehouse architecture

The data lakehouse is an emerging architecture that offers the flexibility of a data lake with the performance and structure of a data warehouse. Lakehouse solutions typically provide a high-performance query engine over low-cost object storage along with a metadata governance layer. Data lakehouses are based around open-standard object storage and enable multiple analytics and AI workloads to operate simultaneously on top of the data lake without requiring that the data be duplicated and converted.

A key benefit of data lakehouses is that they address the needs of both traditional data warehouse analysts who curate and publish data for business intelligence and reporting purposes; and of data scientists and engineers who run more complex data analysis and processing workloads.

IBM watsonx.data, shown in Figure 1-1, is built on an open lakehouse architecture, supported by querying, governance, and open data formats for accessing and sharing data.

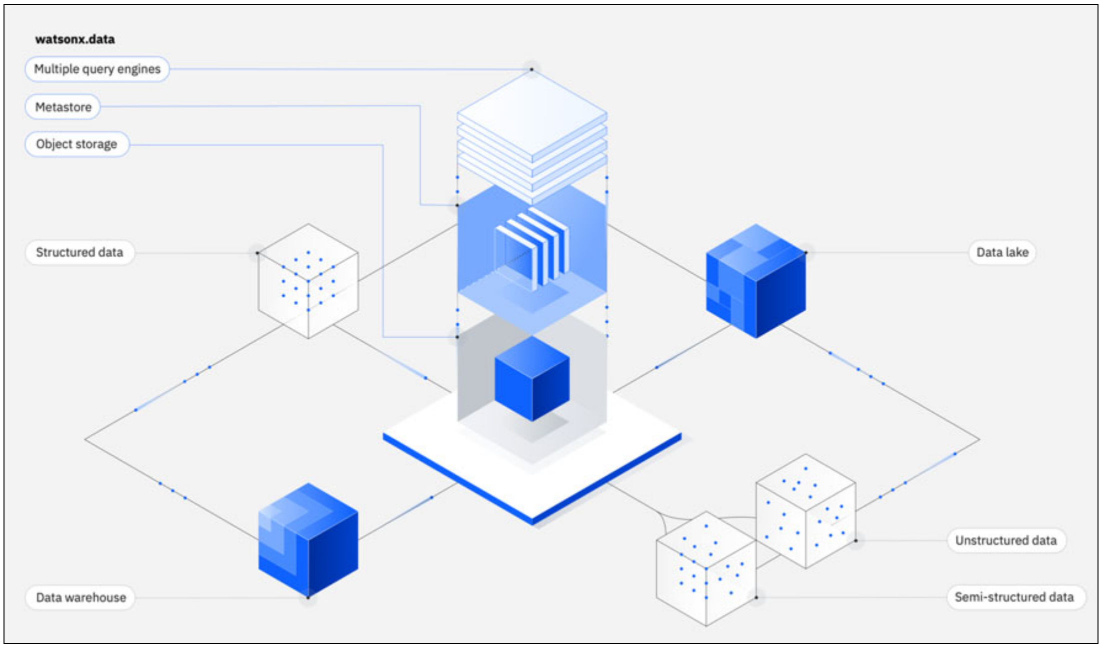


Figure 1-1 IBM watsonx.data provides an ideal platform for building and scaling AI applications

IBM watsonx.data, IBM Storage Ceph, and the IBM Fusion HCI accelerated infrastructure solution

Administrators of today's modern data lakehouses are required to think about storage optimizations as a top priority and a two-tiered approach. The first tier is an on-premises high-performance acceleration layer, which provides superior storage bandwidth with a cost-effective caching approach for the hybrid cloud object storage. The second tier is the low-cost persistent storage for your on-premises storage needs. With the combination of IBM Fusion HCI as your first tier solution and IBM Storage Ceph as your second tier solution, an organization can improve query performance with Storage Acceleration, significant cost advantage, and superior data management capabilities. IBM watsonx.data can take advantage of both of these tiers when using the IBM Fusion HCI and IBM Storage Ceph.