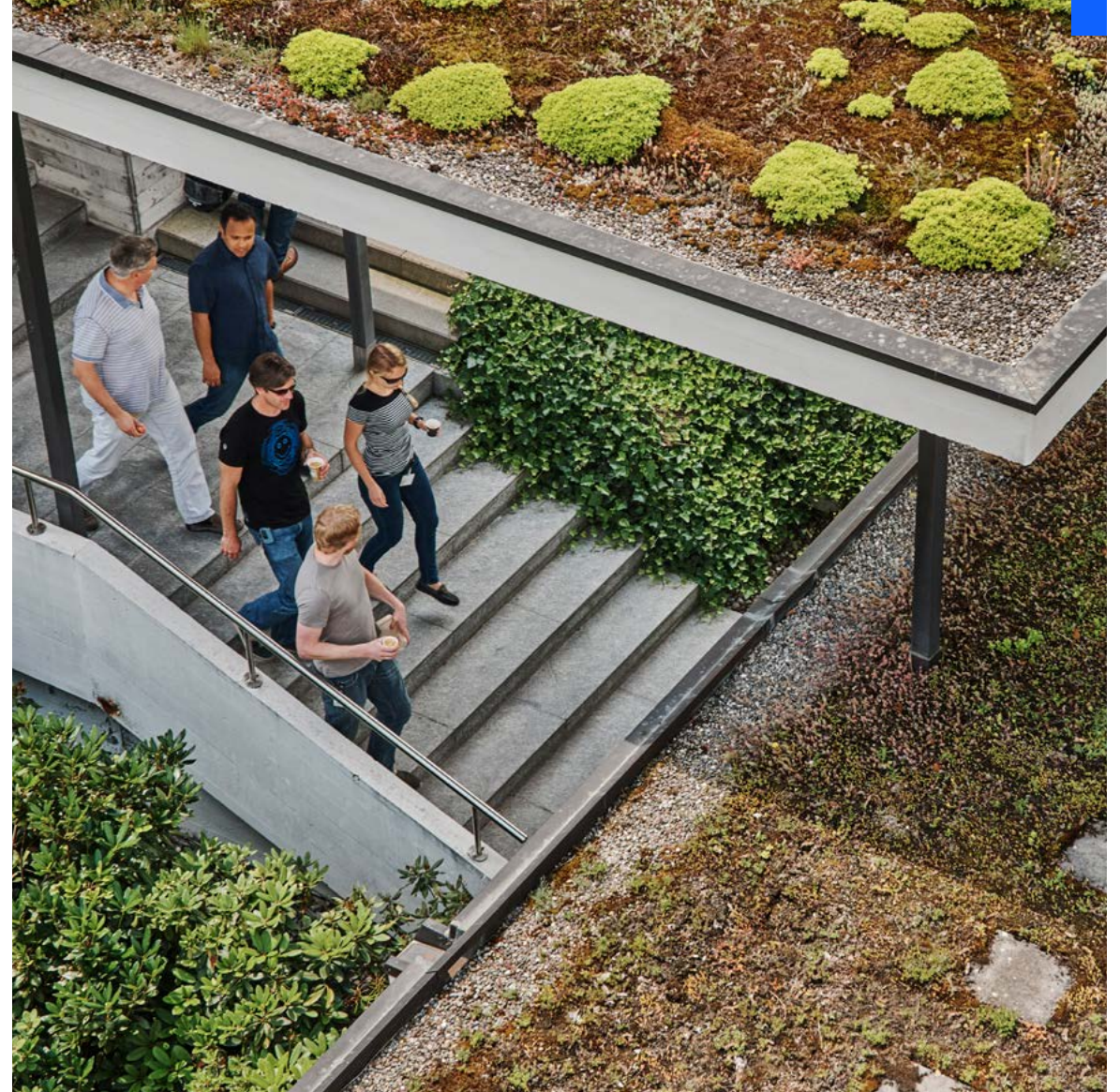
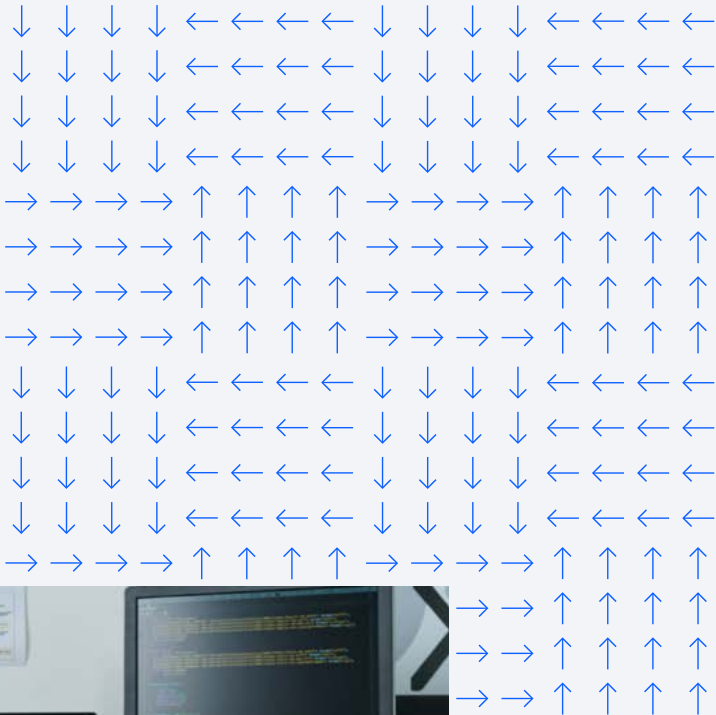




Multicloud data integration for ↔ data leaders



Contents



01

Introduction—a data fabric approach to multicloud integration

02

Why multicloud data integration?

03

The building blocks of multicloud data integration

04

Data fabric—a holistic approach

05

Multicloud data integration success stories

06

Consider these components

07

Create your ideal integrated multicloud data environment

01

Introduction— a data fabric approach to multicloud data integration

Modern enterprises are struggling with increasingly complex data real estates as data becomes more diverse, distributed, and dynamic across hybrid cloud environments. New data sources, applications, and requirements are multiplying. Accessing real-time data in a complex data landscape introduces several challenges that can significantly amplify the difficulties of data integration.

Multicloud data integration addresses this complexity and sprawl across on-premises, multicloud and hybrid cloud environments. When used as part of a data fabric, it can turn a siloed data architecture into one where the right data is delivered to the right place at the right time. A data fabric is a technology architecture approach that enables organizations to unlock insight from their data wherever it resides while keeping true to secure, governed, and performant principles.

Data fabric strengthens compliance with automated data governance and privacy controls while maintaining regulatory compliance no matter where data resides. It also builds the basis for 360-degree customer intelligence and trustworthy AI.

Modern data integration solutions that are part of a data fabric allow you to create flexible, reusable, augmented data pipelines to create and deliver data products across diverse domains and lines of businesses. It enables real-time monitoring, alerting, and analysis of data quality and performance across diverse data sources. Forward-thinking organizations like Highmark Health have already experienced the benefits of such an approach.

Keep reading to get the full story or try it for yourself with our [multicloud data integration trial](#).



Why multicloud data integration?

As organizations adapt to the complexities of the modern data landscape, Chief Data Officers (CDO), Chief Information Officers (CIO), and other data leaders often struggle with the management and integration of this data while ensuring its quality and governance. 76% of data leaders say that data silos hinder cross-departmental exchange, and 74% see this as a competitive disadvantage.¹

In contrast, a sound multicloud data integration strategy can:

- Democratize data access by delivering data where you need it, whether it be on-premises or any cloud, in near-real time, batch, or a virtualized manner
- Provide real-time synchronization of operational and analytic data stores without disrupting or impacting mission-critical data
- Serve the ever-growing population of new data consumers by providing access to data, wherever it resides, in a controlled, regulated manner
- Reduce vendor lock-in and enable greater choice in data environments, allowing enterprises to make selections that meet their unique combination of pricing, performance, security, and compliance requirements
- Ensure optimum health and performance of data at rest and in motion



A sound multicloud data integration strategy can democratize data access by delivering data where and when you need it.

The building blocks of multicloud data integration

The goal of multicloud data integration is to democratize data access across the enterprise, enable continuous availability for mission-critical data, and empower new data consumers, while keeping true to secure, governed, and performant principles. Multi-cloud data integration provides AI with a broader, more diverse dataset, enhancing model accuracy and enabling more robust insights and predictions.

Modern data integration solutions require a data delivery platform that supports different integration and delivery styles while using knowledge graphs, active metadata, and policy enforcements to unify and govern data across a distributed landscape.

By leveraging a flexible architecture that is purpose-built for cloud native workloads, and infusing ML-powered capabilities that allow consumers to unlock insight from their data wherever it resides, modern data integration solutions can provide faster access to high-quality data for downstream analytics, business intelligence (BI), AI, and data intensive application consumers.

Data integration helps combine structured and unstructured data from disparate sources into meaningful and valuable data sets. Integration styles such as bulk/batch integration (ETL or ELT), data replication and streaming integration, change data capture, and data virtualization enable a variety of integration use cases.

- ETL and ELT are both data integration processes that move raw data from a source system to a target data repository, such as a data lake or data warehouse. Advanced transformation of the data can either be applied in-flight (ETL), or post-load (ELT). Data sources can be in multiple, different repositories, including legacy systems or systems of record, that are then transferred using bulk or batch integration to a target data location.
- Real-time data integration enables fast, reliable date-based decisions in the era of Generative AI. Streaming data pipelines decreases the staleness of that data produced by traditional data pipelines and enables users to take in structured and unstructured data from various data sources, in real-time, and turn that data into insights.
- Data replication provides the flexibility to replicate data between a variety of heterogeneous sources and targets for dynamic delivery, making it ideal for multisite workload distribution and continuous availability—whether across data centers, from on premises or to the cloud.
- Change data capture captures database changes as they happen, in real-time and delivers them to target databases, message queues, or as part of an ETL solution.
- Data virtualization connects data across data sources and makes it accessible through a single access point—regardless of data location, size, type or format. Data engineers can quickly and easily fulfill ad hoc data integration requests to validate hypotheses or what-if scenarios.
- Data observability provides a comprehensive understanding of data health and performance. With continuous observability and real-time visibility into the data workflows, improve the overall reliability and efficiency of your complex data environment. It helps monitoring data pipelines, detecting anomaly, alerting when data incidents occur and remediating issues.

03

The building blocks of multicloud data integration

Data cataloging can help users easily find and use the right data with a rich and metadata-driven index of cataloged assets. Data catalogs with built-in quality analysis can make inferences and identify anomalies. Some catalogs have built-in refineries that help with data discovery, cleansing and transforming the data with data shaping operations.

Data governance makes the right data easier to find for those who should have access to it, while allowing sensitive data to remain hidden or masked. A governance framework designed to automate the enforcement of data protection policies and backed by automated metadata tagging is ideal for the establishment and enforcement of data governance policies and rules.

Advanced data engineering automates data access and sharing, accelerating data delivery with active metadata. Orchestration can also be used to make DataOps pipelines simpler to build, and run, improving the integration and communication of data flows between data producers and data consumers.

Automatic workload balancing and elastic scaling make jobs ready for any environment and any data volume.

Finally, you'll want an approach that includes solutions to provide self-service access, the ability to query all data, and continuous data availability.



Data fabric—a holistic approach

Gartner predicts that through 2026, manual data integration tasks will be reduced by up to 30% through the adoption of data fabric design patterns that support augmented data integration. In addition, by 2025, data integration tools that do not provide capabilities for multicloud/hybrid data integration through a PaaS model will lose 50% of their market share to those vendors that do.²

As enterprises increasingly adopt hybrid and multi-cloud environments, they require a next-generation data fabric that is not only AI-enabled but also AI-facilitated, offering a more effective and efficient way to manage and integrate data across diverse platforms.

A data fabric is an architectural approach that simplifies data access in an organization to facilitate self-service data consumption. It brings together capabilities like those listed previously as part of a unified architecture, avoiding the cost and complexity of integrating a plethora of point solutions. Instead of a fragmented, group of products that have been stitched together, a data fabric offers a single, holistic solution that is built to work seamlessly.

In addition, a data fabric can address three separate use cases beyond multicloud data integration. These include data governance and privacy, 360-degree views of customers, and MLOps and generative AI. While the three additional use cases are covered in separate ebooks, it is important to remember that the value of a data fabric approach goes beyond the more streamlined and all-encompassing solution for multicloud data integration.



A data fabric offers a single, holistic solution that is built to work seamlessly.

05

Multicloud data integration success stories

Highmark Health ↪

With 720,000 cases annually in the U.S., and a staggering mortality rate between 25–50%, sepsis isn't just life-threatening; it doubles as one of the country's most expensive inpatient conditions, [consuming more than USD 27 billion annually](#).

At Pittsburgh-based [Highmark Health](#) a team realized they could build a model from insurance claims data and easily integrate and deploy insights into the existing clinical services application to identify high risk patients for sepsis. With a data fabric, and operationalizing a multicloud data integration use case, Highmark could handle complex and varied data sets—and even better—unite data scientists, architects and engineers who were collaborating on this first-of-a-kind project.

Highmark's Director of Data Science R&D, teamed with IBM to build a model, then score and identify patients likely to develop sepsis. They were able to eliminate data silos, provide a trusted data source and reduce data preparation by cataloging all attributes in one place. They also integrated insights into the application workflow and were able to monitor them for bias, trust and transparency. Ultimately, the AI development and deployment lifecycle was [reduced from 12 months to six weeks](#).



“I think it was within a couple of days that IBM came back with a deployed model and I was kind of shocked.”

Curren Katz

Director of Data Science R&D
Highmark Health

Consider these components

Your organization needs a modern data integration stack that ensures the data is accessible, consistent, trusted and ready for AI. Effective data integration is a cornerstone of a next-generation data fabric, providing the foundation for AI-powered data management and analytics. IBM Data Fabric is open and acts as a trusted data foundation for all your operations, analytics and AI implementations.

The capabilities of Data Fabric allow you to ingest, explore, prepare, manage, govern and serve petabyte-scale data for business-ready AI. The capabilities are data intelligence and data integration.

Data intelligence is designed to secure your data landscape with inbuilt governance, catalog, data sharing and AI capabilities.

Data integration provides modular data integration and management capabilities to provide flexible and scalable solution for every use case under the data fabric landscape. It is designed to create, manage and monitor data pipelines that ensure accurate and real-time data is accessible from every corner of your data fabric.

The three core data integration capabilities that ensure you can deliver trusted, reliable and fit-to-purpose data at every corner of your organization are as follows.

06

Consider these components



IBM DataStage

IBM® DataStage® is an industry-leading data integration tool that helps you design, develop and run jobs that move and transform data. At its core, the DataStage tool supports extract, transform and load (ETL) and extract, load and transform (ELT) patterns. Get flexible and near real-time data integration, both on-prem and on the cloud, with a scalable ETL platform and workload balancing that can execute faster.

[Learn more about IBM DataStage →](#)



IBM StreamSets

StreamSets is designed to create smart pipelines to integrate low-latency data for Generative AI. With StreamSets, build, run, manage and monitor streaming data pipelines at scale, across environments. You can transform the data in flight as it is collected. It improves real-time decision making and reduces the data latency risks associated with complex dataflows across an organization.

[Learn more about IBM StreamSets →](#)



IBM Databand

Databand is a proactive, continuous data observability tool that detects data incidents earlier, resolves them faster and delivers more trustworthy data to the business. Databand is designed to work with any data orchestration, data integration or workflow automation tools you might be using, bringing all the monitoring data under one roof. Data teams can maintain optimal performance and meet ever-evolving service level agreements (SLAs) for their data pipelines.

[Learn more about IBM Databand →](#)

Create your ideal integrated multicloud data environment

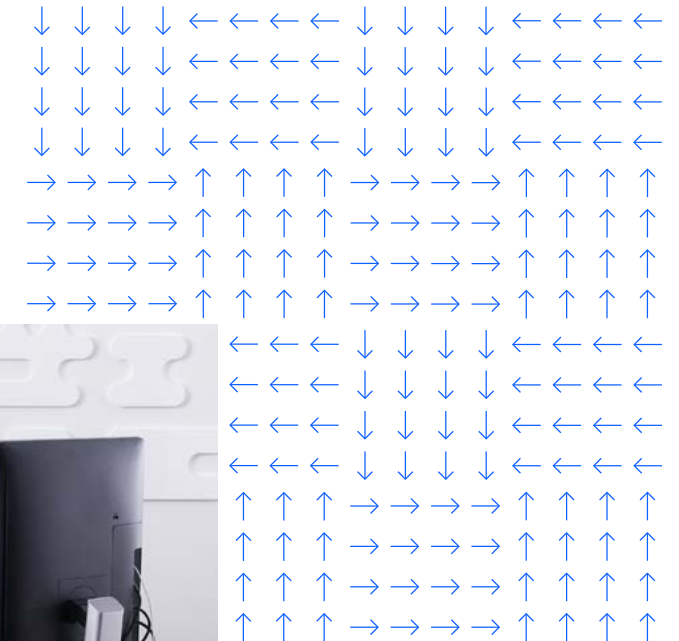
If you're excited to finally tame the data sprawl and provide simpler, regulated data access for all types of data users, we encourage you to take advantage of a few resources. Foremost is the free multicloud [data integration trial](#), which will provide hands on experience with a data fabric designed for this use case. Second, review the information on our [multicloud data integration website](#) to learn even more about the benefits. And, finally, reach out to one of our experts either by [scheduling a time online](#), talking with your IBM representative or reaching out to one of our business partners.

Check out the other three data fabric use case ebooks:

[Data Governance and Privacy](#)

[Customer 360](#)

[Trustworthy AI and MLOps](#)





© Copyright IBM Corporation 2024

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
October 2024

IBM, the IBM logo, ibm.com, IBM Cloud Pak, IBM DataStage, and IBM Watson are trademarks of International Business Machines Corp., 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 www.ibm.com/legal/copytrade.shtml.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

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.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

01 2023 Industry study: Data success factor, Xphere by XPLM, 2023.

02 Magic Quadrant for Data Integration Tools, Gartner, 2023.