

# Data fabric architecture delivers instant benefits

Deploy a data fabric with IBM and AWS to integrate and engage all your organization's data for better business outcomes

## **Contents**

- 3** The challenges of data today
- 5** Business benefits of a data fabric
- 6** How IBM delivers a data fabric on AWS
- 7** Data fabric in action on AWS
- 8** IBM Data Fabric on AWS Architecture
- 9** Data fabric is the transformative next step for your enterprise

# The challenges of data today

Data is an integral element of digital transformation for enterprises. But as organizations seek to leverage their data, they encounter challenges resulting from diverse data sources, types, structures, environments and platforms. This multidimensional data predicament is further complicated when organizations adopt hybrid and multi-cloud architectures. For many enterprises today, operational data has largely remained siloed and hidden, leading to an enormous amount of dark data.

## What is a data fabric?

In the past, organizations have attempted to address data access problems either through point-to-point integration or introduction of data hubs. Neither of those are suitable when data is highly distributed and siloed. Point-to-point integrations add exponential cost for any additional end point that needs to be connected, meaning this is a non-scalable approach. Data hubs allow for easier integration of applications and sources but exacerbate the cost and complexity to maintain quality and trust of data within the hub.

The data fabric is an emerging architecture that aims to address the data challenges arising out of a hybrid data landscape. Its fundamental idea is to strike a balance between decentralization and globalization by acting as the virtual connective tissue between data endpoints (see page 4).

Through technologies such as automation and augmentation of integration, federated governance as well as activation of metadata, a data fabric architecture enables dynamic and intelligent data orchestration across a distributed landscape, creating a network of instantly available information to power a business.

A data fabric is agnostic to deployment platforms, data processes, geographical locations and architectural approach. It facilitates the use of data as an enterprise asset. A data fabric ensures your various kinds of data can be successfully combined, accessed, and governed both efficiently and effectively.

## Capabilities and principles of a data fabric

The core of the data fabric architecture is a data management platform that enables the full breadth of integrated data management capabilities including discovery, governance, curation, and orchestration.

However, a data fabric advances and evolves from traditional data management concepts such as DataOps, which only focuses on establishing practices, to increase the level of data operationalization. It is built upon a distributed architecture and advanced technology able to address the needs that arise from extreme diversity and distribution of data assets.

## Data Fabric with IBM and AWS

This partnership combines the #1 leading AI portfolio with the largest global cloud infrastructure, making IBM is the ideal partner for customers' cloud transformation journey on AWS.

AWS and IBM have formed an ecosystem of highly experienced professionals committed to deploying customer solutions on AWS.

IBM is helping customers move existing data fabric workflows to the AWS Cloud, develop cloud-native apps, and optimize their existing AWS Cloud environment.

IBM's expertise in security, enterprise scalability, and open innovation with Red Hat OpenShift helps clients migrate quickly and seamlessly to the AWS Cloud.

## Get ready to learn

This whitepaper explores what a data fabric is, how it differs from previous architectures, what can it achieve for your business, and the role of IBM and AWS in implementation.

A data fabric could be logically divided into four capabilities (or components):

### Knowledge, insights and semantics

- Provides a data marketplace and shopping experience
- Automatically enriches discovered data assets with knowledge and semantics, allowing consumers to find and understand the data

### Unified governance and compliance

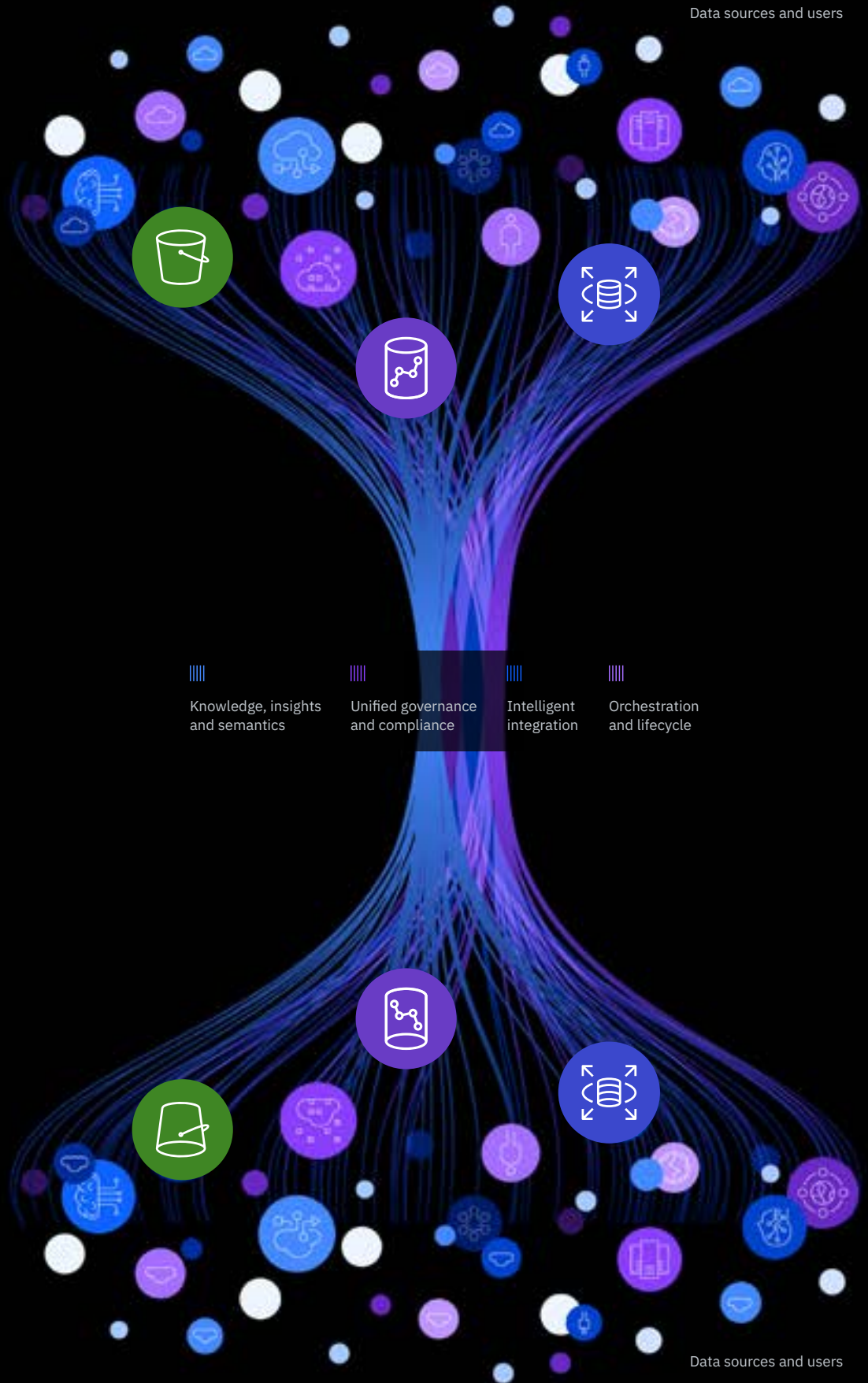
- Allows local management and governance of metadata but supports a global unified view and policy enforcement
- Automatically applies policies on data assets in accordance with global and local rules
- Utilizes advanced capabilities to automate data asset classification and curation
- Automatically establishes queryable access routes for any cataloged assets for increased activation of data

### Intelligent integration

- Accelerates a data engineer's tasks through automated flow and pipeline creation across distributed data sources
- Enables self-service ingestion and data access over any data with local and global deep enforcement of data protection policies
- Automatically determines best fit execution through optimized workload distribution and self-tuning and correction of schema drifts

### Orchestration and lifecycle

- Enables the composition, testing, operation and monitoring of data pipelines
- Infuses AI capabilities in the data lifecycle to automate tasks, self-tune, self-heal and detect source data changes, all of which facilitates automated updates



# Business benefits of a data fabric

Data only delivers business value when it is contextualized and becomes accessible by any user or application in the organization. When implemented correctly, a data fabric helps ensure those values are available throughout the organization in the most efficient and automated way possible. As such, the fabric has three key benefits:

1. Enable self-service data consumption and collaboration
2. Automate governance, protection and security; enabled by active metadata
3. Automate data engineering tasks and augment data integration across hybrid cloud resources

## Enable self-service data consumption and collaboration

By integrating data from multiple sources and analyzing a larger fraction of the enormous amount of data generated daily, organizations gain better insights and respond more quickly to changing business demands. A data fabric rapidly delivers data into the hands of those who need it. Self-service enables the organization as a whole to find appropriate data quicker and spend more time using that data to provide tangible insights.

Benefits of data fabric for self-service data consumption:

- Business users have a single point of access to find, understand, shape and consume data throughout the organization.
- A centralized data governance and lineage help users understand what the data means, where it comes from, and how it is related to other assets.
- Extensive and customizable metadata management scales easily and is accessible via APIs.
- Self-service access to trusted and governed data enables line-of-business collaboration with other users.

A Forrester Total Economic Impact study<sup>1</sup> revealed that these capabilities can mean:

USD

5.8M                      459%

in benefits                      ROI

## Automate governance, data protection and security; enabled by active metadata

A distributed active governance layer for all data initiatives reduces compliance and regulatory risks by providing trust and transparency. It enables automatic policy enforcement for any data access, providing a high level of data protection and compliance.

Utilizing AI and machine learning technologies allows data fabric users to increase their level of automation, for example automatically extracting data governance rules based on language and definitions in regulatory documents. This allows organizations to apply industry-specific governance rules in a matter of minutes to help avoid costly fines and ensure ethical use of data wherever it resides.

Benefits of a data fabric for governed virtualization:

- Agility, security, and productivity is increased for data engineers, data scientists, and business analysts.
- Multiple global data sources appear as one database.
- New, industry-leading discovery of personally identifiable information (PII) and critical data elements is possible at massive scale.

These capabilities can mean:

USD

2.4M                      430%

in benefits<sup>1</sup>                      performance improvement<sup>2</sup>

## Automate data engineering tasks and augment data integration

Advanced data engineering means that virtually any data access or delivery process is automated and not requiring any tedious or error-prone coding process. Augmentation of integration utilizes metadata data to optimize the data delivery and access.

Benefits of a data fabric for data engineering and integration:

- Automatically optimized data integration helps accelerate data delivery.
- Automatic workload balancing, and elastic scaling means jobs are ready for any environment and any data volume.
- Resiliency and CI/CD automation are built in.
- The automated process for capturing changes in real time supports delivery of quality data for business processes.
- Machine learning can automate and extend custom data discovery, classification and curation processes, leading to faster time-to-value.
- Continuous analysis can be automatically performed in real time, wherever data lives.

Results from a leading retailer<sup>3</sup> show:

60X                      20X

acceleration in data                      faster customer affinity

delivery time                      analyses

## Provide a 360° view of your customers

Achieve a single, trusted, and comprehensive view of customers by connecting data across all sources in your organization. Minimize risk, increase precision, and improve decision-making speed by centrally managing master data. The single view enables business users within enterprises to extract key insights that facilitate the delivery of a superior customer experience. Also, a comprehensive view helps position products and services to meet customers' needs.

Benefits of a data fabric for a 360° customer view:

- Compile a customized and trusted customer view.
- Enable workflow capabilities to implement governance policies and processes.
- Operationalize master data for deeper analysis and enhanced insights.
- Manage your customer relationships and hierarchies, and get more accurate reporting.

## Support trusted AI

Leverage a data fabric to capture the AI lifecycle to enable greater visibility and automate documentation of the lifecycle. Also, determine which enterprise policies will be enforced during the development and deployment lifecycle.

Benefits of a data fabric for trust in AI:

- Enable the automatic monitoring of models and automatically re-train when applicable.
- Leverage governance rules to automate the control of the model lifecycle.
- Utilize automated monitoring to identify bias, drift, and ensure fairness.
- Make smart recommendations.

# How IBM delivers a data fabric on AWS

IBM Cloud Pak® for Data makes this concept of a data fabric possible. IBM Cloud Pak for Data is an insight platform that simplifies and automates data collection, organization and analysis of data and accelerates the infusion of AI throughout your business.

With data fabric on AWS, you can connect to AWS data, workloads, and build, deploy and manage AI at scale. IBM Cloud Pak for Data is the enabler for a business digital transformation.

IBM Cloud Pak for Data leverages the following capabilities to deliver the business-ready data a data fabric demands. All these capabilities play a part in supporting the data fabric architecture.

### 1. Metadata-based knowledge core

Facilitates the discovery of data sources and catalogs; enriches data assets; and performs analysis to extract insight for more automation using AI. The knowledge core is used to power the marketplace with semantic search.

### 2. Self-service data marketplace

The next-generation data catalog that helps data consumers, such as business analysts, retrieve data from across the data landscape of the enterprise.

### 3. Smart integration

Enables data consumption by extracting, virtualizing, transforming, and streaming data. It is integrated with the knowledge core to automate data integration, and has the intelligence to decide which integration approach is best suited based on workloads and data policies. It can also be used for data preparation as part of data engineering workloads or to create data products. Finally, it provides the ability to publish updates to data products.

### 4. Governance

Deliver governed data fabric with virtualization across all data sources. Catalogs and curates metadata, defines data policies for privacy, curates data, captures data lineage, and performs other tasks related to security and compliance. This layer understands the data format (for example, structured vs. unstructured) and data significance (for example, public vs. protected) and applies the correct policies to each bit of data and each prospective user. Instead of applying standards and rules to data manually, this integrated platform capability means they can be applied at the organizational level and propagate throughout the various data resources as needed. Analytic models in different tools can talk to one another; data policy enforcement at the granular level can be largely automated.

### 5. Unified development and operations

Enables a unified lifecycle to configure and run all the aspects of the data platform in production.

# Data fabric in action on AWS

The best way to understand the value of a data fabric is to see the business benefits that actual organizations have achieved.

You can get IBM Cloud Pak through the AWS Marketplace and get data fabric, governance and privacy capabilities on AWS up and running in less than 4 hours.

## Illumina

### Business Challenge

Illumina, a global leader in genomics, creates leading-edge applied genomics technology for its clients. Genomics is highly dependent on data, but industry trends and competitive information are not readily available. They needed to be able to prepare short-term and long-term sales forecasts and capacity plans that quickly adapt to their expanding markets.

### Solution

Seeking to use predictive analytics with artificial intelligence, they worked with the IBM Data Science and AI Elite team to identify their sales forecasting priorities and develop a proof of concept (POC) that included IBM Watson Studio on IBM Cloud Pak for Data. The company then worked with IBM Consulting Services and the Advanced Analytics team to develop AI/ML models that forecast global territories, improving accuracy, efficiency, and granularity across Illumina’s product portfolio. The solution provided data-driven insights into customer trends and a second data point for vetting revenue projections, demand planning, and annual budget planning.

### With a data fabric, this organization is able to:

- Improve shipment forecast accuracy by 19.7% using advanced analytics.
- Fine tune customer segmentation and discover data insights to guide their future data strategy.
- Test environment deployed in less than 3 days.

Data consumers / Business processes		
<b>Unified lifecycle</b> <ul style="list-style-type: none"> <li>– Design</li> <li>– Compose</li> <li>– Build</li> <li>– Test</li> <li>– Orchestrate</li> <li>– Deploy</li> <li>– Operate</li> <li>– Monitor</li> </ul>	<b>Data self-service</b> Exploration, collaboration, data marketplace, provisioning	<b>Unified data governance, security, compliance, financial governance</b> <ul style="list-style-type: none"> <li>– Data policies</li> <li>– Data lineage</li> <li>– Data quality</li> <li>– Data privacy</li> <li>– Cost optimization</li> </ul>
	<b>Smart integration</b> Data ingestion, transformation, preparation, virtualization, real-time data	
	<b>Augmented knowledge</b> Discovery, catalog, semantics, ontologies, graph, metadata, master data, reference data, event-driven data	

Figure 1. The capabilities of the IBM Cloud Pak for Data platform— all of which support AI development and the data fabric

## A Manufacturing Company in Asia Pacific

### Business Challenge

One of the significant challenges for the client’s IoT solution is hybrid data management. There is a considerable volume of data to be handled – it includes both structured data and semi-structured data. The structured data is mainly about the elevators, escalators, and sensors on them. It contains essential information about devices, the relationship between devices, maintenance plans, etc. And for the semi-structured data, it’s mainly time-series data generated by sensors.

### Solution

Data collection and data analytics capabilities are handling the hybrid data in huge volumes – combined as data stores when implementing the solution. This solution leveraged IBM Cloud Pak for Data data management capabilities, including Watson Studio, and data stores, including IBM DB2 and MongoDB.

### With a data fabric, this organization is able to:

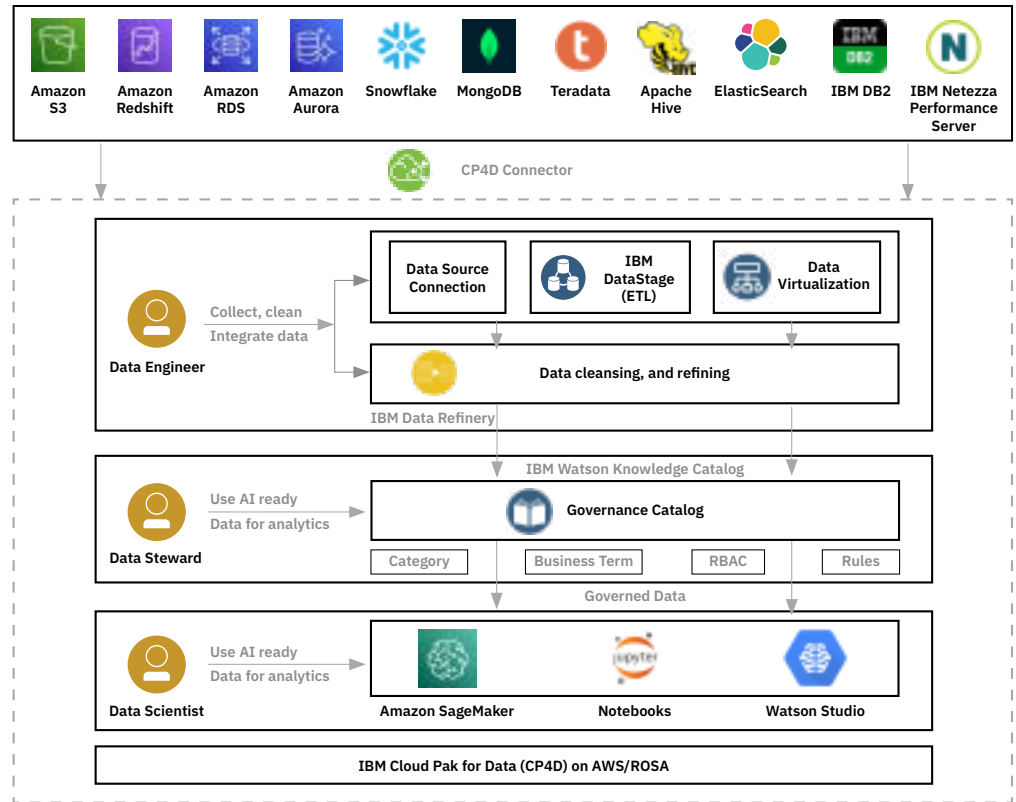
- Manage 10,000+ elevators and escalators
- Connect various equipment (elevators, escalators, doors) and develop capabilities for internal usage (monitoring and maintenance).
- Provide additional pathways for new service development, such as equipment management for building owners and mobile applications for end-users.

# IBM Cloud Pak for Data on AWS Architecture

IBM® Cloud Pak for Data lets you connect to your data on AWS. It supports a variety of data sources such as Amazon S3, Amazon Redshift, Amazon RDS, Amazon Aurora, Snowflake, MongoDB, Teradata, Apache Hive, IBM DB2, Netezza performance servers and more.

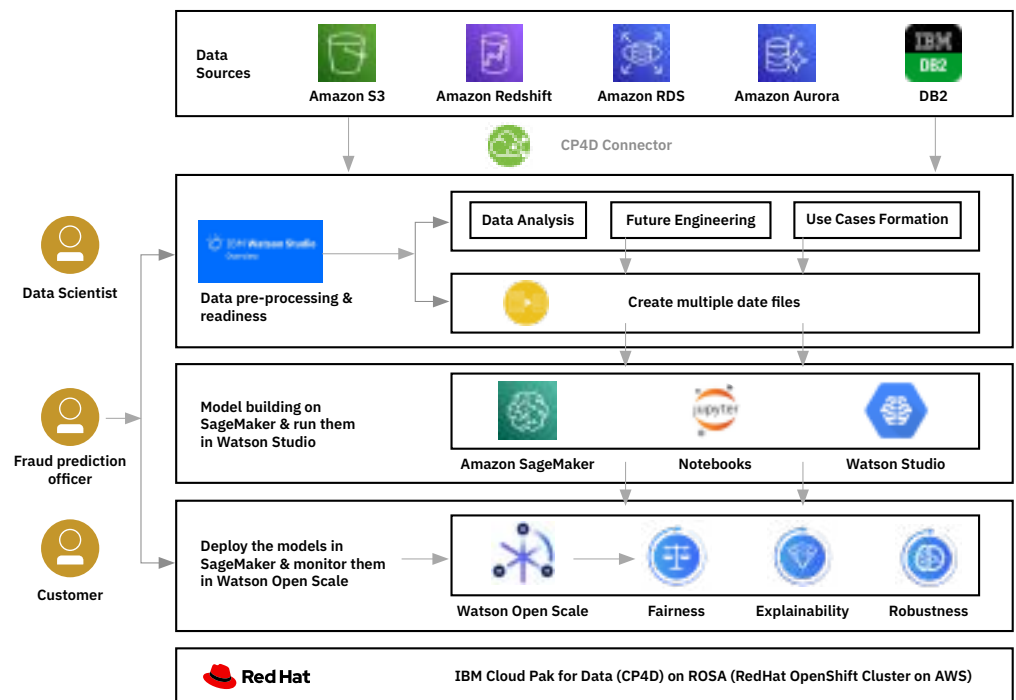
## Use Case: Data Governance

Deliver governed data fabric with virtualization across all data sources.



## Use Case: Trustworthy AI

Deliver both prescriptive and predictive capabilities for model explainability, to detect model bias, fairness, decay and drift.





# Data fabric is the transformative next step for your enterprise



With a data fabric built on IBM Cloud Pak for Data technology, you can hyper-automate data discovery, data governance, and data consumption on AWS.

Employ a data fabric to enable faster time-to-value for business users, higher productivity for data engineering and operations, and greater governance and compliance fidelity.

[To get started, find IBM Data and AI on the AWS Marketplace](#)

© Copyright IBM Corporation 2022  
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

IBM Corporation  
New Orchard Road  
Armonk, NY 10504

Produced in the United States of America  
May 2021

IBM, the IBM logo, and IBM Cloud Pak are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on [ibm.com/trademark](http://ibm.com/trademark).

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. It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs. 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.

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed, misappropriated or misused or can result in damage to or misuse of your systems, including for use in attacks on others. No IT system or product should be considered completely secure and no single product, service or security measure can be completely effective in preventing improper use or access. IBM systems, products and services are designed to be part of a lawful, comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective. IBM DOES NOT WARRANT THAT ANY SYSTEMS, PRODUCTS OR SERVICES ARE IMMUNE FROM, OR WILL MAKE YOUR ENTERPRISE IMMUNE FROM, THE MALICIOUS OR ILLEGAL CONDUCT OF ANY PARTY.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

- 1 A Forrester New Technology: Projected Total Economic Impact™ Study Commissioned By IBM, February 2020
- 2 Performance measurements were gathered within a controlled test environment at IBM Silicon Valley Labs using IBM data virtualization against various 100TB data sources. The measurements taken in May 2019 and performance gains are compared to IBM federation.
- 3 Based on experience by an IBM customer.