Db2 on IBM Cloud Pak for Data

An integrated multicloud data platform
built on Red Hat OpenShift

IBM

Table of contents

Data in context

Data is one of a business's most important resources. However, its value is contingent upon having a comprehensible and complete view of it. Artificial intelligence (AI) technology holds enormous promise in surfacing deeper insights more quickly—accelerating the pace of discovery, broadening the range of data that can be leveraged and automating tasks that previously required human expertise. AI can only be effective if the full range of data is trustworthy, accessible and compatible. And the rise of AI has highlighted weaknesses and limitations that have long existed in data systems—driving the demand for new strategies.

Today you can take your pick of data sources and types: sensor logs, text documents, images, audio recordings, geolocation data, sentiment data and much more. Modern analytics and AI-driven tools can reveal unprecedented insights from enterprise data, but if you're limited to one type of data or even one type of database system, you're missing the complete picture.

This means that although enterprises still have an ongoing need for robust, secure, scalable and high-performance databases, those databases can't exist in isolation. They need to be integrated into a capable data platform.

Consider the typical data flow (Figure 1) between different data roles in an enterprise. Every role has a different focus:

1. Executives set goals and strategy
2. Enterprise architects align the appropriate technology to the business goals and strategy
3. Data engineers build and maintain data pipelines
4. Data scientists build and refine analytic and AI algorithms
5. Developers embed those insights into useful applications

However, all of these roles share a common need for varied, high-quality, trustworthy data. With that data, they can improve decision making, accelerate innovation, improve the customer experience and drive operational efficiency, among other advances.

App developers	Data engineers	Data stewards	Data scientists	Business executives	Enterprise architects
Collect		**Organize**		**Analyze**	
– Data virtualization – Data warehousing – Databases on-demand – Data source ingestion – Distributed processing **Powered by:** Db2 and Db2 Warehouse technologies		– Discovery and search – Data transformation – Data cataloging – Business glossary – Policies, rules and privacy **Powered by:** InfoSphere, DataStage and InfoSphere Information Governance Catalog and Watson Knowledge Catalog		– Data visualization – Machine learning – Model build and deploy – Model management – Dashboards **Powered by:** Watson Studio open source and Cognos	
Multicloud services	– Logging – Monitoring	– Metering – Persistent storage	– Kubernetes – Security	– Identity access management	– Docker registery/ Helm
IBM Cloud Pak for Data					

Figure 1. Example data flow between data roles in an enterprise setting

For enterprises trying to leverage AI technology, the demands on data systems are especially intense. Because AI technology can expand and accelerate traditional business uses for data, AI-driven intelligence can make the difference between purely historical retrospectives and proactive systems that automate insights and make predictions. The opportunity is immense: in 2021, AI augmentation is expected to generate USD 2.9 trillion in business value and recover 6.2 billion hours of worker productivity.[1]

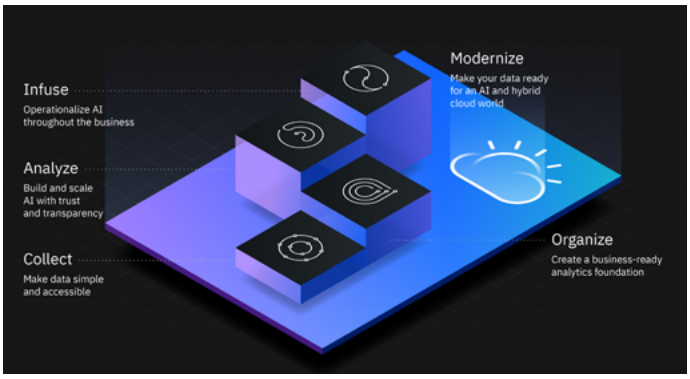


Figure 2. The AI Ladder: five rungs to a successful AI initiative

But effective AI requires top-quality information architecture (IA) underneath. That's why IBM developed the AI Ladder (see Figure 2)—it presents a unique perspective on how organizations can prepare for, and implement, AI technology. Organizations must modernize their data platform; collect, organize, analyze their data; and infuse AI throughout their organization.

All these factors—from data variety to data users needs to the rise of AI—contribute to a major shift in how databases are deployed and used. Databases are essential to the AI journey—so the benefits of databases such as IBM® Db2®, which offers performance improvements with BLU Acceleration® in-memory computing, and distributed and highly available scalability with pureScale® technology—are still highly relevant to the success of your overall data strategy.

But by integrating these database technologies seamlessly with other parts of the data platform, enterprises can benefit from comprehensive and automated data provisioning while maintaining the performance, security and governance they need. Containerized architectures—specifically those deployed on a cloud-enabled platform such as IBM Cloud Pak® for Data—are key to this transformation.

Using containers for your AI database

The need for a rapid, robust, capable data platform to support AI initiatives is why IBM developed IBM Cloud Pak for Data: a data and AI platform designed to enable governed self-service consumption of data throughout the enterprise. Cloud Pak for Data is a hybrid cloud data and AI platform delivering an information architecture for AI. With Cloud Pak for Data you can unlock the value of all your data on a unified, cloud-native platform and automate how your organization turns data into insights.

By deploying Db2 through Cloud Pak for Data, enterprises can leverage all of the powerful features that make Db2 an industry-leading[2] data platform. As the AI Database, Db2 is built for and powered by AI.

Built for AI

- Exploring data faster with natural language queries
- Popular languages and code examples to build AI applications
- Modeling complex relationships with graph on SQL data
- Analyzing blockchain data natively

Powered by AI

- Data virtualization
- Adaptive workload management and resource optimization
- Machine learning query optimization
- Confidence-based querying

Db2 on Cloud Pak for Data is the containerized version of Db2 Advanced Edition, based on a microservices architecture and optimized for Kubernetes. Through Cloud Pak for Data, Db2 can run natively on Red Hat® OpenShift®, the world's leading container orchestration platform.

By breaking down the Db2 database functions into microservices instead of a monolithic stack, you gain several opportunities:

- **Deploy within minutes;** enable standard deployment and management while retaining flexibility to modify parameters as needed.
- **Gain reliability** due to out-of-the-box enhanced Kubernetes availability and support for high availability/disaster recovery (HA/DR) automated failover.
- **Reduce management burden** with automated updates. Service packs, versions and mods can be deployed with one click.
- **Automate management** by "application group;" administrators can use namespaces to manage access control and provisioning options.
- **Monitor and manage at an application level** thanks to platform and service-level features.
- **Scale microservices independently** to respond to changing needs.

Containerizing your data and AI platform enables you to run Db2 as part of a hybrid cloud environment (combination of cloud and non-cloud platforms) or multicloud environment (clouds from different providers) that uses the appropriate infrastructure for each type of data.

These advantages may account for the recent popularity of containers. Based on the Red Hat Global Customer Tech Outlook 2019, 57% of organizations are already using containers and container usage is also expected to increase by 89% in the next 2 years.[3]

But the advantages of choosing Cloud Pak for Data as your specific version of a containerized architecture go far beyond that. Thanks to thorough integration between Cloud Pak for Data and other elements of the IBM portfolio, you can also easily leverage additional data capabilities, such as machine learning with IBM Watson®; DataOps with Watson Knowledge Catalog and IBM InfoSphere® DataStage®; and integrated data science with Watson Studio and Watson OpenScale™. With your data in Db2 on Cloud Pak for Data, you can more easily access the full scale and expertise of IBM services to design, deploy and manage advanced analytics that help you deliver business value.

The value of Db2 on Cloud Pak for Data

As shown in Figure 3, Db2 on Cloud Pak for Data combines the Db2 containerized microservices architecture with several additional layers of functionality:

1. **Services layer**
   Get access to over 45 analytics services and templates from IBM and third parties thanks to a containerized microservices version of Db2.
2. **Data virtualization functionality**
   Quickly and easily query multiple data sources without moving your data.
3. **Platform interface layer**
   Speed time-to-value with a single user experience that integrates data management, data governance and analysis for greater efficiency and improved use of resources.
4. **Kubernetes layer**
   Leverage the leading hybrid cloud, enterprise container platform, Red Hat OpenShift, for an innovative and fast deployment strategy.
5. **Infrastructure layer**
   Avoid lock-in and leverage any cloud infrastructures with a multicloud approach.

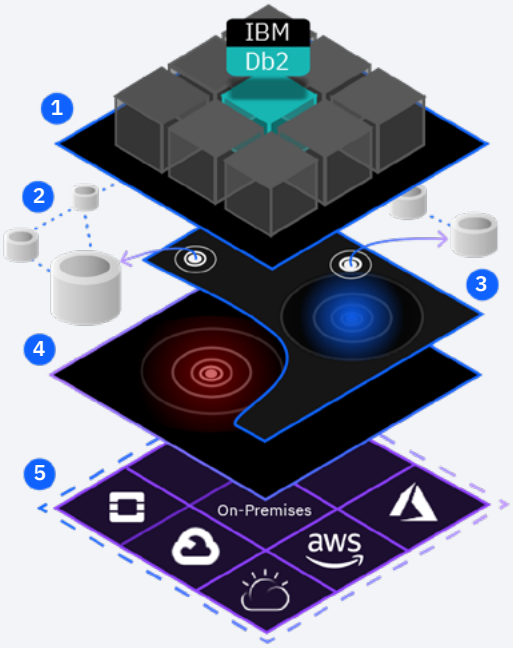


Figure 3. Db2 on Cloud Pak for Data

Three major benefits of deploying Db2 on Cloud Pak for Data

01
Ease of enabling hybrid cloud and multicloud on a single, unified platform

Db2 on Cloud Pak for Data enables enterprises to more easily deploy and manage hybrid cloud and multicloud data platform models. For complex enterprise data platforms, the hybrid cloud model can represent the best of both worlds: control plus efficiency. Enterprises can run sensitive, highly regulated and mission-critical applications and workloads—or workloads with reasonably constant performance and capacity requirements—on private cloud infrastructure. They can run less-sensitive, more-dynamic, or even temporary workloads such as development and test environments for a new application on the public cloud.

With the proper integration and orchestration between the two, enterprises can even leverage both, when needed, for the same workload. For example with “cloud bursting” organizations can leverage additional public cloud capacity to accommodate a spike in demand for a private cloud application.

Multicloud’s inherent flexibility offers benefits including risk mitigation, optimization and ready access to the services you need. Multicloud helps mitigate risk in two ways: by limiting exposure from a single vendor approach and by preventing vendor lock-in. In a multicloud environment, if one provider’s cloud experiences downtime, the outage will affect only one vendor’s service. For example, if your hosted email is down for a few hours, services from other providers, such as a website or software development platform, can still run. And if a vendor is no longer meeting your needs, there is no barrier to choosing a solution from another vendor.

Multicloud also lets users optimize performance by choosing the service that best suits their needs. One service might offer extra functionality or employ a security protocol that makes it easier to meet a user’s compliance requirements.

Another significant multicloud benefit is access to technology. For example, analytics solutions can easily be deployed as cloud services without the up-front capital expense. This also means a user can get the service up and running more quickly, accelerating the time to value.

Similarly, when users have the freedom to choose any provider for any solution, they can access new, innovative technologies more quickly than from a single vendor’s catalog and combine services from multiple providers to create applications that offer unique competitive advantage.

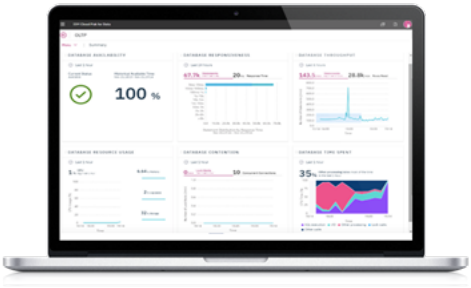
The value of Red Hat in Cloud Pak for Data
The hybrid cloud and multicloud options are enhanced by the advantages of Red Hat OpenShift, upon which Cloud Pak for Data is based. The Red Hat stack, with IaaS, OpenShift and Kubernetes operating together, is particularly beneficial. It allows you to develop secure and scalable Kubernetes applications without being overwhelmed by the complexities of large-scale manual Kubernetes administration. OpenShift provides a comprehensive platform that enables automated operations and provides out-of-the-box support for languages such as Java, Node.js, Ruby and Python. OpenShift also provides supporting services such as monitoring, authentication and authorization and network management. These features that Red Hat adds through OpenShift are not in Kubernetes.

In addition, the included Kubernetes distribution is enterprise-grade, and benefits from hundreds of security, defect and performance fixes in each release. Validated popular storage and networking plug-ins for Kubernetes are also available. And finally, open-source Red Hat tools provide additional functionality options, such as Apache Spark for streaming data, or the popular Python and R languages for machine learning applications. The additional functionality ensures that enterprises leverage essential open source tools necessary to develop, deploy and run applications through the OpenShift platform.

When these varied resources are all part of a single, unified platform in Cloud Pak for Data, they are easier to integrate and manage than they would otherwise be.

02
Ease of data integration and management

Because multi-location deployment options are managed through a single unified platform, Db2 on Cloud Pak for Data can simplify complex management tasks and speed time to value. A single platform that integrates data management, data governance and analysis can offer greater efficiency and improved use of resources. This can improve data access, help maintain data security, and support self-service collaboration among teams by offering unmatched ease of integration between various data resources.



Unified console
A unified data platform needs a unified point of access, so the Db2 console is built on Open Restful APIs. This means all dashboard functionality is available via open APIs.

This single portal allows you to:

- Provision, administer, monitor, manage and optimize the performance of over 100 Db2 databases in any form factor with the interactive, simple to use UI.
- Deliver all data users one coherent, efficient and scalable experience across all data services.
- Collaborate via composable user interface—the Db2 console allows users to compose their own interface by selecting what to display; this selection can be shared with teammates.

Pre-integrated data experiences
With the microservices model on Cloud Pak for Data, standard resource deployments can be provisioned quickly with pretrained abilities tailored for specific industries, such as fraud detection or profile matching. The advantages of these prebuilt experiences include:

- Offers a common look-and-feel with customizable persona-based workflows
- Provides API-like functionality with models residing and running directly in the database, improving performance and minimizing data movement
- Accommodates a well-defined API through the ecosystem, rather than the reliance on collocation and a set file system
- Automates common services, including user management, authentication models, security configurations, provisioning, collaboration and more

IBM ecosystem

With Db2 on Cloud Pak for Data, it is easy to leverage capabilities from the broader IBM and open-source ecosystems. For example, a full suite of AI-enabling IBM solutions is integrated by design, certified to work together, and supported by IBM. They include:

Watson Studio
Build phase of AI: Use open and extensible data science tooling to explore data, prepare data and develop models.

Watson Machine Learning
Run phase of AI: Deploy, retrain and manage models.

Watson OpenScale
Infuse phase of AI: Continuously evaluate outcomes; find KPIs and production metrics.

Watson Knowledge Catalog
Deliver intelligent cataloging, backed by active metadata and policy management.

InfoSphere DataStage
Leverage ETL functionality for hybrid and multicloud environments, including metadata management.

Cognos® Analytics
Use a single business intelligence platform for the entire analytics lifecycle, from discovery to operationalization management.

Further benefits from the open-source ecosystem include:

- Authoring tools—Jupyter notebooks, R Studio, H2OAI
- Machine learning, optimization—Spark, Anaconda
- Deep learning runtimes—Torch, Caffe, Keras, TensorFlow
- Coding and visual modeling options—Graph, Python, R and more
- Model lifecycle management—automatic versioning, release updates and SLAs
- Simplified lifecycle management—rolling upgrades via Helm

Security and reliability
Security and reliability are fundamental to Cloud Pak for Data. The containerized architecture helps isolate any failures so that, for instance, a console-level failure does not impact the overall system. And data security is improved by the ability to centralize data governance policies, apply them to all data in the enterprise and use the data without moving it or granting wider access. Cloud Pak for Data is designed with security in mind and equipped with many security characteristics, including access control, authentication and authorization; data protection; and security logging.

Db2 supports secure engineering development practices: threat modelling, risk assessment, static and dynamic code analysis, penetration testing, container scanning and more. It is built to support security operations best practices such as audit log consolidation and analysis, user access management, and security incident management.

All of these features make data governance and compliance initiatives much easier. For example, compliance controls can be defined by outside agencies, and Db2 includes system security plans for maintaining compliance security postures. Db2 on Cloud Pak for Data is Federal Information Security Management Act (FISMA) High ready with system security plans, spanning 350 controls.

It's because of these advantages—deployment model flexibility, easy integration, and unified management—that Db2 on Cloud Pak for Data can support the ultimate goal of enterprise data: quick and easy access to insights in your data, wherever it resides, without needing to move that data first.

03
Query across multiple data sources without moving data

With Db2 on Cloud Pak for Data, you can achieve a trusted view of all your data and comprehensive insight without moving data around, without duplicating it and exposing it to security risks.

These two sets of features help you achieve a singular view of the data:

Data governance and self-service analytics
Data in Db2 is automatically integrated with governance capabilities for auto discovery, cataloging and search. These automatic processes are subject to data security policies and access rules, and they enable self-service data discovery by the widest possible range of data users. Because the data is pre-cleaned and pre-validated, and because access controls can be centrally administered, those looking to use data can do so quickly with the knowledge that the data is of high quality and they can't access anything they aren't supposed to.

Ecosystem and virtualization
The Db2 family includes capabilities optimized for OLTP (Db2 Database), OLAP (Db2 Warehouse) Hadoop/big data (BigSQL) and streaming "Fast Data" (Event Store). Db2 also benefits from additional formal partnerships with service providers such as MongoDB, Postgres, and more, also available in Cloud Pak for Data.

Data virtualization deserves special mention, because this pervasive technology is integral to the overall goal of collating insights from different data without needing to move all data to a single repository, which may be impossible (due to different types of data) or merely very difficult (due to cost, latency and security risks). Virtualization helps overcome the roadblocks inherent to data silos.

Without virtualization, the tedious task of searching through, moving and governing data resources is left to highly paid and skilled data teams. This work could take up 80% of a data scientist's time just a few years ago,[4] and despite advances is still seen as burdensome.[5] Data transfers also take time and can incur surprising fees. Although the IBM Cloud® does not charge for data transfers, many other providers do.

By contrast, governed data virtualization (Figure 4) provides the ability to view, access, manipulate and analyze data without the need to know or understand its physical format or location, and without having to move or copy it. Data virtualization connects data silos as if they were a single data set. The process performs analytics queries on the servers where data lives, and then returns the results to the original application.

With Db2 on Cloud Pak for Data, you can avoid the silo (data subset), governance, security, and duplication problems associated with moving data. You also free up your data scientists to focus on strategic initiatives and model building, rather than manual data cleansing and manipulation.

Learn more
Download the paper or view the video for more about virtualization and eliminating data siloes with Cloud Pak for Data.

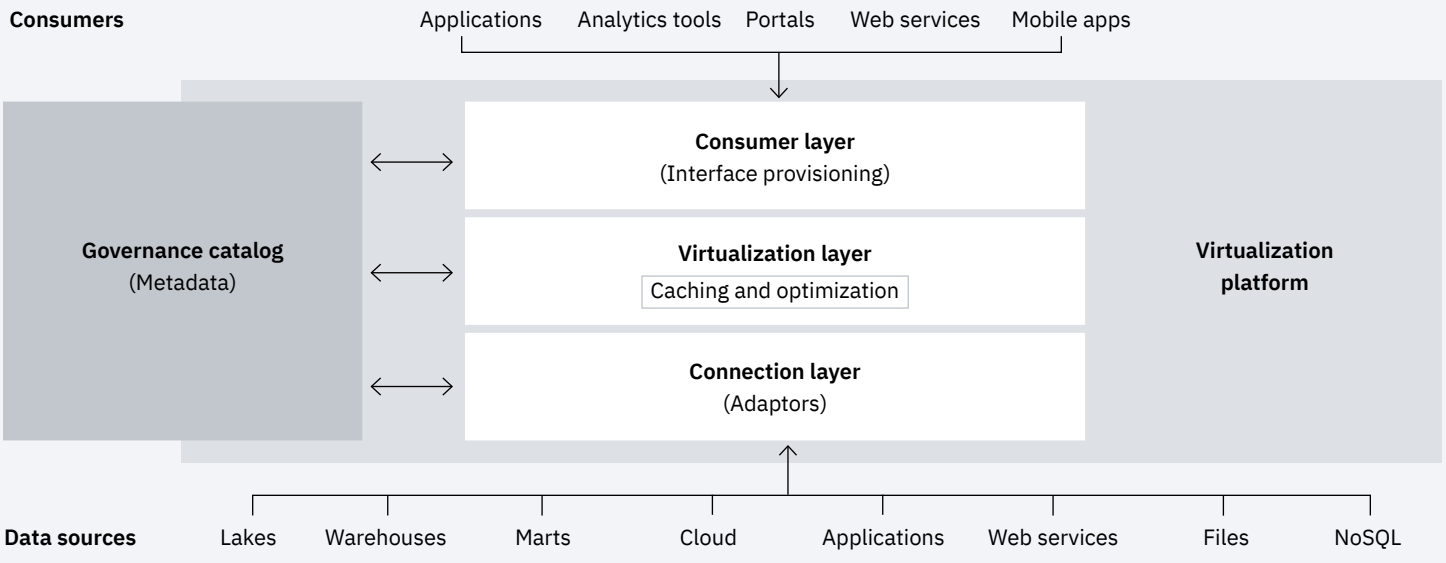


Figure 4. Example of how the virtualization layer enables connection of any data to any consumer

Next steps

When deployed via IBM Cloud Pak for Data, Db2 is more than a robust data repository in which you can collect data. It becomes part of a microservices-based data platform that also helps you organize and analyze your data, infusing AI capabilities throughout your enterprise.

Db2 on IBM Cloud Pak for Data

1. Built with AI capabilities, and built for AI projects
2. Data virtualization unlocks insights without requiring data movement
3. Pre-governed self-service data access through a single touchpoint

Db2 on Cloud Pak for Data offers a unique combination of containerized architecture, Red Hat infrastructure, data connectivity and a broader IBM capabilities ecosystem, making it a compelling choice for enterprises that want to prepare their data foundations for the opportunities ahead.

To get started try Cloud Pak for Data for free. Schedule a free one-on-one consultation with a Hybrid Data Management expert.

IBM Corporation
New Orchard Road, Armonk, NY 10504
Produced in the United States of America
June 2020

1 “Reshaping Business with Artificial Intelligence” by Sam Ransbotham, David Kiron, Philipp Gerbert, and Martin Reeves. MIT Sloan Management Review, September 2017. www.sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/

2 “The Forrester Wave™: Data Management For Analytics, Q1 2020” by Noel Yuhanna. Forrester, February 2020

3 www.redhat.com/en/blog/red-hat-global-customer-tech-outlook-2019-automation-cloud-security-lead-funding-priorities

4 “Cleaning Big Data” by Gil Press, Forbes, March 2016. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says

5 “What do Data Scientists Really Do?” by Hugo Bowne-Anderson, Harvard Business Review, August 2018 https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists