IBM Cloud Pak For Data v3.0

# Introduction

Cloud Pak for Data is a data and analytics platform with built-in governance. It simplifies and unifies how your organization collects, organizes, and analyzes data to accelerate the value of data science and artificial intelligence (AI). This multi-cloud platform delivers a broad range of core data microservices, with the option to add more microservices from a growing services catalog. Cloud Pak for Data offers greater flexibility, security, and control, along with the benefits of the cloud, without having to move your data.

What does IBM Cloud pak for data do?

IBM Cloud Pak for Data is a native cloud solution that simplifies your data ecosystem by letting you connect to your data, govern it, find it, and use it for analysis from a single, unified interface.

IBM Cloud Pak for Data v3.0

Cloud Pak for Data V3. 0 is an integrated data and AI platform that enables organizations to collect, organize, and analyze data.

# Project Summary

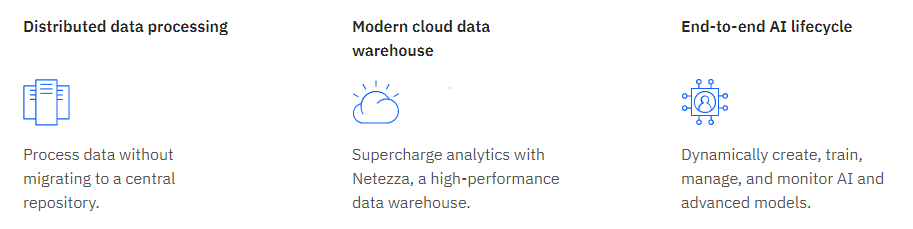
| Website | https://www.ibm.com/in-en/products/cloud-pak-for-data |
| --- | --- |
| Organization/Foundation Name | IBM |
| License | L-DNAA-C22LVZ |
| Open/Proprietary | Proprietary |
| Source Path | https://www.ibm.com/docs/en/cloud-paks/cp-data/3.0.1?topic=overview |
| Brief Description | Connect the right data to the right people at the right time.IBM Cloud Pak® for Data platform helps improve productivity and reduce complexity. Build a data fabric connecting siloed data distributed across a hybrid cloud landscape. |

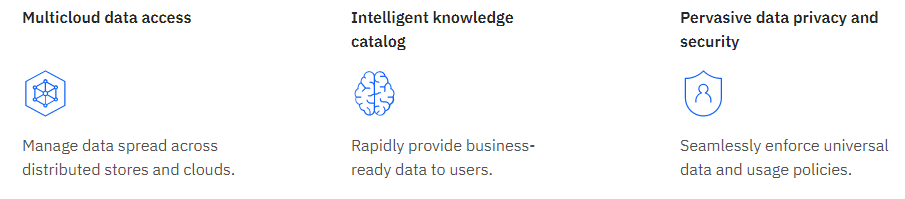
# Project Details

## Key Features

The new add-ons to Cloud Pak for Data enhance preexisting platform capabilities associated with backup and disaster recovery, developer experience, and financial institution client engagement.

* The Portworx third-party add-on provides backup and disaster recovery for Cloud Pak for Data clients, augmenting the existing robust storage capabilities that Portworx provides in Cloud Pak for Data.
* The Personetics third-party add-on is an AI-based personalized engagement platform that enables financial institutions to engage various clients with personalized insights by using proactive personal insights and guidance, self-driving automated money management, and small-to-midsized enterprise (SME) cashflow optimization and guidance.
* The Lightbend Cloudflow is a third-party add-on that helps developers to scalably deploy streaming flows that are written in Spark, Flink, and Akka streams.
* Cloudflow is designed to provide a powerful and cloud-native developer experience for Cloud Pak for Data.





## **Architecture**

### Cloud-native design

* Many companies are embracing cloud concepts because they need reliable, scalable applications.
* Additionally, companies need to modernize their data workloads to use hardware effectively and efficiently.
* By bringing together numerous data and AI services, Cloud Pak for Data enables you to reduce the cost and burden of maintaining multiple applications on disparate hardware.
* It also gives you the ability to assign resources to workloads as needed and reclaim those resources when not in use.

### Run on OpenShift

* Cloud Pak for Data runs on top of Red Hat OpenShift, which means that you can run Cloud Pak for Data on :

1. An on-premises, private cloud cluster

2. Any public cloud infrastructure that supports Red Hat OpenShift.

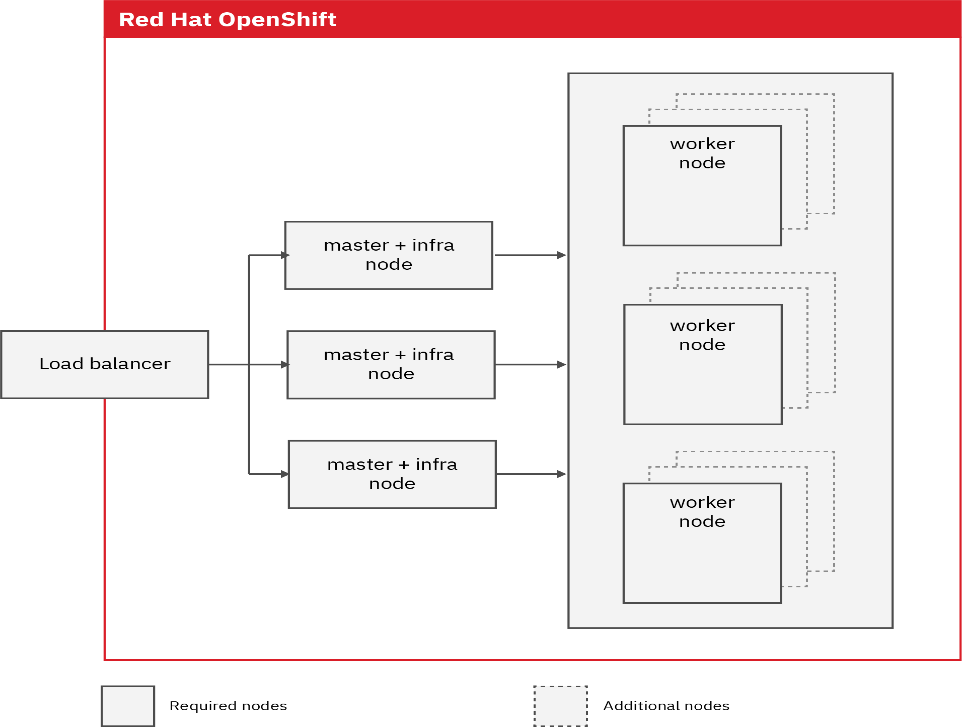
### Storage architecture

* NFS storage
* Portworx storage
* OpenShift Container Storage
* IBM Cloud File Storage

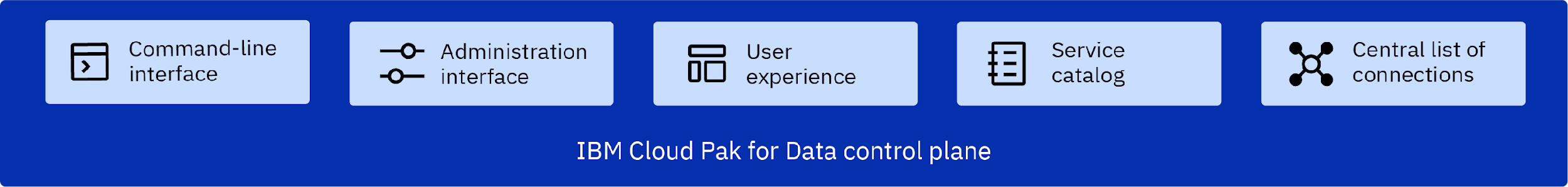
### Cluster architecture

Cloud Pak for Data is deployed on a multi-node cluster. Although you can deploy Cloud Pak for Data on a 3-node cluster for development or proof of concept environments, it is recommended that you deploy your production environment on a larger, highly available cluster with multiple dedicated master and worker nodes.

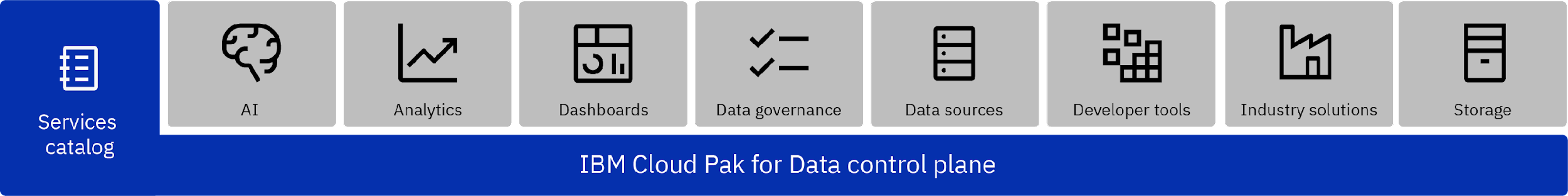
The following diagram illustrates the typical topology of a production-level cluster.



### Modular platform



### Integrated data and AI services



## System Requirements

### **Minimum hardware requirements**

The following *minimum* recommendations are for demos and proof-of-concept with Cloud Pak for Data.

| Node role | Hardware | Number of servers | Available vCPU | Memory |
| --- | --- | --- | --- | --- |
| Master + infra | x86-64 | 1 master and 1 infrastructure on the same node | 8 vCPU | 32 GB RAM |
| Worker/compute | x86-64 | 2 worker/compute nodes for NFS; 3 worker/compute nodes for Portworx | 16 vCPU | 64 GB RAM |

Minimum requirements for Red Hat OpenShift Container Platform Version 4.5.4 or later fixes:

| Node role | Hardware | Number of servers | Available vCPU | Memory |
| --- | --- | --- | --- | --- |
| Master + infra | x86-64, ppc64le | 3 master (for high availability) and 3 infrastructure on the same 3 nodes | 4 vCPU | 16 GB RAM |
| Worker/compute | x86-64, ppc64le | 3 worker/compute nodes | 16 vCPU | 64 GB RAM |

### Storage requirements

An additional 200 GB of free space in the root file system on all of the nodes. Cloud Pak for Data with all services installed can use up to 700 GB of storage space, leaving up to 300 GB of storage space available for user data. You can add additional capacity depending on your user data volume requirements.

### Disk requirements

To prepare your storage disks, ensure that you have good I/O performance, and prepare the disks for encryption.

**I/O performance :** To ensure that the storage partition has good disk I/O performance, run the disk latency test and the disk throughput test:

**Disk latency test:** The value must be better or comparable to: 4096000 bytes (4.1 MB, 3.9 MiB) copied, 1.5625 s, 2.5 MB/s

**Disk throughput test:** The value must be better or comparable to: 1073741824 bytes (1.1 GB) copied, 5.14444 s, 209 MB/s

The OpenShift cluster should include:

* Ensure that you have the appropriate container runtime for your environment:

| **Container runtime** | **OpenShift version** | **Storage type** | **Storage requirement** | **Notes** |
| --- | --- | --- | --- | --- |
| CRI-O version 1.13 or later (recommended) | 4.5, 3.11 | NFS, IBM Cloud File Storage, Portworx | 200 GB | Thread info |
| Docker with the overlay2 driver | 3.11 | NFS | 200 GB | Overlay info |

## Current usage

### Automated AI lifecycle

Use case: Data science teams need integrated systems to manage assets across the AI lifecycle. And, Chief Data Officers want to ensure that AI models and the data that is associated with them are properly governed. With Cloud Pak for Data, you can manage the end-to-end AI lifecycle from a single platform. When you install the services that support the automated AI lifecycle, you can:

* Govern your data
* Prepare your data
* Build models
* Deploy and manage your models at scale
* Govern your AI models
* Detect and mitigate bias and drift in your AI models
* Understand how your AI models make predictions

### Data Modernization

When you install the services that support the data modernization use case, you can virtualize data to enable self-service access to data in real time and at scale without creating redundant data. You can also:

* Prevent your data from becoming dark data by using discovery and cataloging
* Turn raw data into trusted data through data preparation and governance
* Enforce data privacy policies
* Understand the lineage and context for your data

### Data Ops

Enable your workforce to access critical data without exposing sensitive data. Easily create and enforce policies to govern your business data. Cloud Pak for Data includes services that enable you to:

* Prepare and refine your data
* Understand the quality and lineage of your data
* Govern your data in accordance with relevant regulations

### AI for financial operations

Deliver more agile and reliable plans and forecasts to drive better business performance. With Planning Analytics, you can automate and integrate planning across your organization for:

* Financial planning and analysis
* Workforce planning
* Sales forecasting
* Supply chain planning

### AI for customer care

Provide exceptional support. Automate customer care to increase customer satisfaction while decreasing time to resolution and call volume. With Watson Assistant, you can provide AI-powered, automated assistance to your customers or employees through web or mobile applications and over the phone. You can also enable your customer service agents to handle customer inquiries more efficiently and effectively.

### Self-service analytics

### Data is your most valuable asset. Turn your data into a competitive edge with a data warehouse and business intelligence (BI). Use your data to build reports and dashboards that help you derive business insights. Cloud Pak for Data includes the services you need or persist, visualize, and analyze your data at scale.

## Alternatives to IBM cloud pak for Data

* Google Cloud BigQuery
* [MATLAB](https://www.g2.com/products/matlab/reviews)
* G2 Deals
* SAP Analytics cloud
* Tableau
* Posit
* RapidMiner
* [Amazon EC2](https://www.g2.com/products/amazon-ec2/reviews)
* [Qlik Sen](https://www.g2.com/products/qlik-sense/reviews)se

## Reference/ Acknowledgement

<https://www.ibm.com/products/cloud-pak-for-data> : The Link was used to refer the information for the Key Features section and Introduction.

<https://www.ibm.com/docs/en/cloud-paks/cp-data/3.0.1?topic=planning-architecture> : The Link was used to refer the information for the Architecture section.

<https://www.ibm.com/docs/en/cloud-paks/cp-data/3.0.1?topic=planning-system-requirements> : The link was used to refer to the information for the System Requirements.

<https://www.ibm.com/docs/en/cloud-paks/cp-data/3.0.1?topic=use-cases> : The link was used to refer the information for the Current Usage section.