IBM® Cloud Pak for Data — Portworx Reference Architecture

portworx

IBM

1. INTRODUCTION

This document describes the reference architecture for IBM Cloud Pak for Data on RedHat OpenShift with storage managed by Portworx.

IBM Cloud Pak for Data is a modern end-to-end data and analytics platform, providing integrated and flexible workflows for storing, processing data, and helping integrate and unlock the value of customer data.

This reference architecture provides planning, design, and deployment guidelines and considerations for implementing Cloud Pak for Data either in the cloud or on-prem. It is intended for IT professionals, technical architects, sales engineers, and consultants in order to assist them in planning, designing, and implementing the Cloud Pak for Data solution with Portworx as the strategic persistent-storage solution provider.

This document expects you to have knowledge of containers, OpenShift, storage, and big data processing terminology. For more information about Cloud Pak for Data, please see the "Resources" section at the end of the document.

2. CLOUDPAK FOR DATA

Cloud Pak for Data expands and enhances the technology it's built on top of to withstand the demands of your business needs—adding management, security, governance, and analytics features.

With the ever-increasing volume, variety, and velocity of data available to enterprises of all sizes comes the challenge of deriving value from it. That task requires access to multiple data sources, data governance and integration capabilities, flexible and extensible data processing, and an easy data model-inference deployment. Cloud Pak for Data (CP4D) brings the power of artificial intelligence (AI) to enterprises and is an all-in-one data and AI platform that is containerized and deployed on top of OpenShift. Supported by the industry-leading software-defined storage solution from Portworx, Cloud Pak for Data can be built on on-prem systems or public cloud infrastructure to provide a secure environment for data collection, organization, and analysis.

3. PORTWORX ENTERPRISE

Portworx Enterprise is the software-defined container storage platform built from the ground up for OpenShift. By providing scale-out software-defined container storage, data availability, data security, backup, and disaster recovery for OpenShift-based applications running on-prem or across clouds, Portworx has helped dozens of Global 2000 companies—such as Carrefour, Comcast, GE Digital, Lufthansa, T-Mobile, and SAIC—run containerized data services in production.

With Portworx Enterprise, you can

- Run any database or data-rich service on OpenShift, even those that require strict performance, backup & DR, security, and data mobility capabilities.
- Improve application performance and uptime by avoiding the limitations of storage platforms built for VMs—not containers.
- Achieve Zero RPO and < 1 minute RPO Disaster Recovery for mission-critical data services.
- Cut cloud storage costs in half without sacrificing performance.

Portworx Enterprise was named the #1 Kubernetes Storage Platform by GigaOm Research for the breadth of the solution, scale of supported use cases, and list of reference customers.

The Portworx Enterprise platform is made up of the following components that provide everything an enterprise needs to successfully run stateful applications on Cloud Pak for Data and OpenShift. IBM customers can use a certain amount of Portworx for free. See this document for details on what is included in the free offering.

PX-Store — Scalable Persistent Storage for Kubernetes

Built from the ground up for containers, PX-Store provides cloud native storage for applications running in the cloud, on-prem, and in hybrid/multi-cloud environments.

PX-Store includes

- Container-optimized volumes with elastic scaling for no application downtime.
- High Availability across nodes/racks/AZs so you can failover in seconds.
- Multi-writer shared volumes across multiple containers.
- Storage-aware class-of-service (COS) and application aware I/O tuning.
- And much more...

PX-DR — Data Protection for Mission Critical Apps

PX-DR extends the data protection included in PX-Store with Zero RPO Disaster Recovery for data centers in a metropolitan area as well as continuous backups across the WAN for an even greater level of protection.

PX-DR includes

- Multi-site synchronous replication for Zero RPO DR across a metro area.
- Multi-site Asynchronous Replication for DR across a wide area network (WAN).
- The ability to set all DR policies at the container-granular level.

PX-Backup — What Backup for Kubernetes Should Be

PX-Backup allows you to capture entire applications—including data, application configuration, and Kubernetes objects—and move them to any backup location at the click of a button. You can recover entire applications just as easily.

PX-Backup includes

- Continuous backups across global data centers.
- Point-and-click recovery for any Kubernetes app.
- Backup and recovery of cloud volumes from Amazon, Microsoft, and Google.
- The capability to fulfill your compliance and governance responsibilities with a single pane of glass for all your containerized applications.

PX-Autopilot for Capacity Management — Stop Over-provisioning Cloud Storage

PX-Autopilot for Capacity Management allows you to stop over-provisioning storage capacity in the cloud so you can cut your cloud storage bill in half.

PX-Autopilot includes

- The ability to automatically resize individual container volumes or your entire storage pools.
- Rules-based engine that is fully customizable so you can optimize your apps based on performance requirements.
- Integration with Amazon EBS, Google PD, and Azure Block Storage.

PX-Secure — Container Data Security without Compromise

With PX-Secure encryption and access controls, you can move securely at the speed of Kubernetes.

PX-Secure includes

- Cluster-wide encryption.
- Container-granular or storage-class based BYOK encryption.
- Role-based access control for
  - Authorization
  - Authentication
  - Ownership
- Integration with Active Directory and LDAP.

PX-Migrate — Multi-cloud Data Mobility at Your Fingertips

Complete control over your Kubernetes data no matter where it lives.

PX-Migrate includes

- Multi-cloud/multi-cluster application migration.
- Application-consistent snapshots.
- Snapshot-based backup to any cloud.

4. BUSINESS SOLUTION REQUIREMENTS

This section describes the key functional and non-functional requirements for the solution provided.

Functional requirements

Modern data and analytics solutions such as Cloud Pak for Data meet the following key functional requirements with an ecosystem that provides a choice of IBM, open source, and third party services:

- Cloud native agility
- Data management capabilities
- Data integration and governance capabilities
- Enterprise catalog to search/use information assets
- Choice of data science tools/technologies to support a diverse data science team
- Integrated cross-persona workflows
- Extensible APIs

Non-functional requirements

- Ease of development/deployment
- Ability to scale up/down services based on usage/seasonal demands
- Multi-tenancy
- Service level isolation to cater to regulatory demands
- Data confidentiality and integrity
- High availability, Disaster recovery

Customers require their big data solution to be fast, easy, and dependable. OpenShift—enhanced further thanks to Portworx—helps CloudPak 4 Data also meet these non-functional requirements in the following ways:

- Fast and easy
  - High Scalability (via being atop OpenShift/Portworx)
  - Easy-to-access data by various user types
  - Multi-tenancy via projects (Kubernetes namespaces)
  - Ease of development and easy management at scale
  - Advanced job management

- Secure and governed
  - Data confidentiality and integrity (possible via Portworx)
  - Efficient response to changing regulations with embedded, sophisticated governance capabilities—including automated discovery and classification of data, masking of sensitive data, data zones, and data lifecycle management
  - Strong authentication and authorization
  - CP4D ActiveDirectory/Kerberos integration support
- Resilient, reliable, and dependable
  - Data protection with snapshots and replication (possible via Portworx)
  - High availability (HA) and business continuity (possible via Portworx)
  - Automated self-healing
  - Insight into software/hardware health and issues

5. PLANNING FOR CP4D DEPLOYMENT AND OPERATIONS

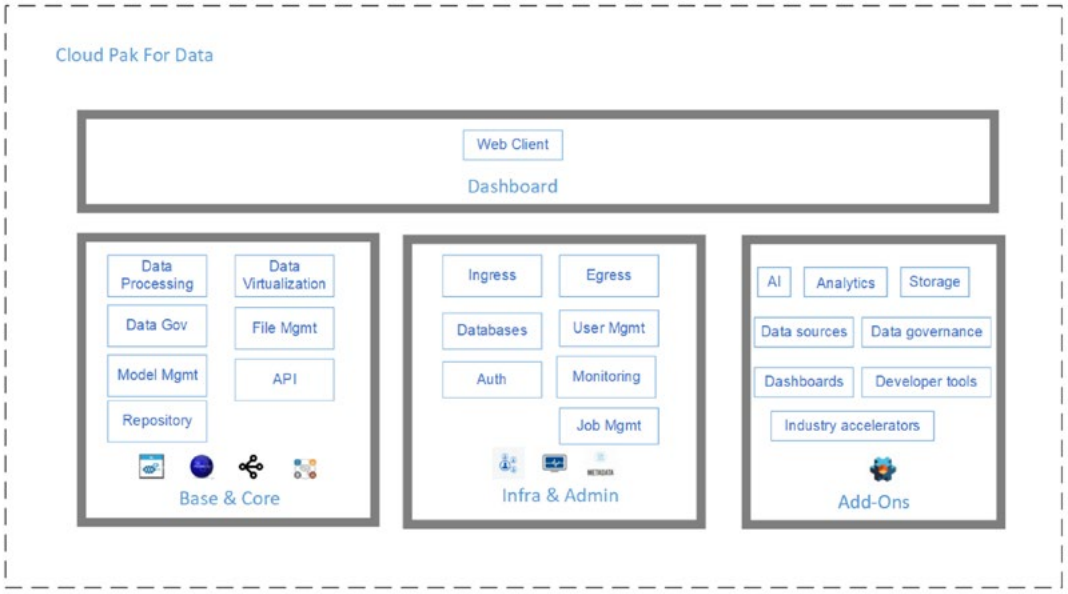
Being able to properly plan for deploying and operating a system as full-featured and advanced as Cloud Pak for Data requires a bit of fundamental understanding of its architecture.

5.1 Architecture

Understanding the various layers of the architecture gives a fuller picture of how they come together to provide a functional solution.

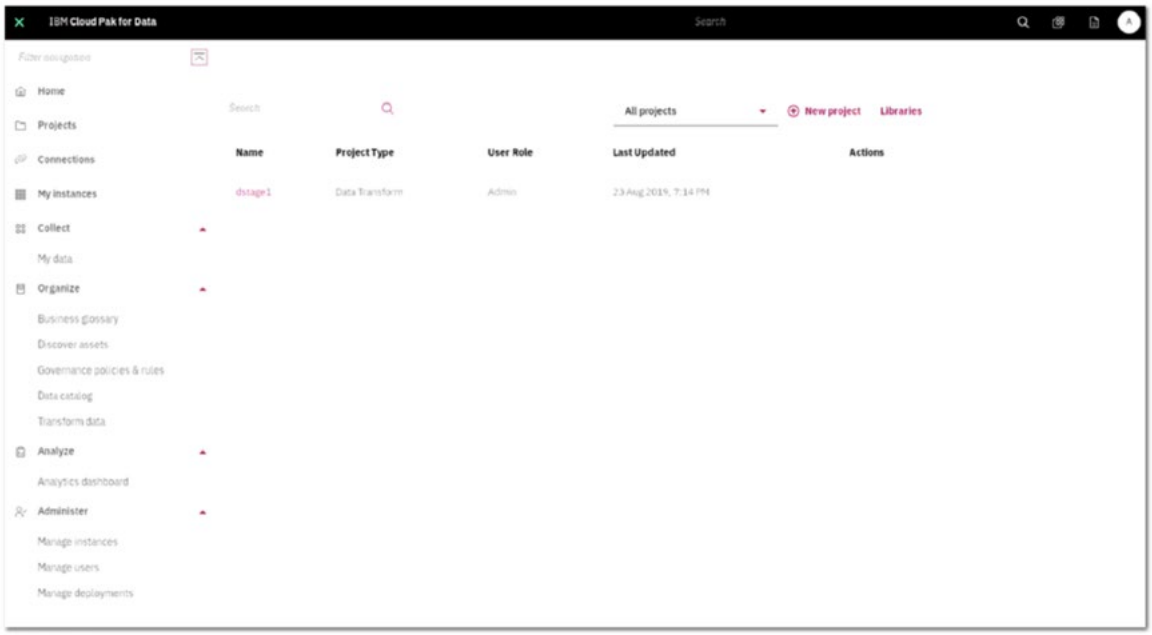
5.1.1 CP4D components and end-user interface

We'll begin with the end closest to the end-user by examining the layout/view of IBM Cloud Pak for Data and its different components, along with what the end-users actually see/interface with.



CP4D Dashboard

Cloud Pak for Data has a web-based dashboard. From the dashboard, users can take care of their administrator operations, data management, data governance, and analysis in a unified web-based UI. The figure below shows the web client dashboard of Cloud Pak for Data.



CP4D Control Plane

The control plane of Cloud Pak for Data permits modular installation and deployment of services by allowing the user to pick and choose which services to enable. CP4D is provisioned and enabled in terms of units. The base purchase of Cloud Pak for Data includes entitlements for a certain amount of such units, which are spent when selecting which base features should be installed. Additional units for other services can be procured, which would provide additional functionality.

Cloud Pak for Data APIs facilitate programmatic management of users and their access control along with user account management. They can be used to interact with your governance metadata to manage assets, custom asset types, and the association between them. The APIs provide the capability to manage analytics projects and the collaborative use of assets—notebooks, scripts, datasets—allowing users to quickly harvest insight from the data in a repetitive fashion as well as from automated job scheduling. They also automate deployment (from development to production) and help maintain machine learning models, making them accessible through HTTP endpoints on the platform. For more information about APIs, please visit the relevant IBM page.

CP4D infrastructure and administrative components

Cloud Pak for Data makes administration of your enterprise data processing and AI jobs simple and straightforward at any scale. The Infrastructure and Administrative planes enable operators to create/delete users, modify users' profiles, and grant privilege to users. It also can connect to an LDAP server for admission control, using a custom SSL or TLS certificate for HTTPS connections to the web client.

This plane sets the standard for enterprise deployment by delivering granular visibility into and control over every part of the data and AI jobs, which empowers operators to improve performance, enhance quality-of-service, increase compliance, and reduce administrative costs.

Cloud Pak for Data monitors a number of performance and health metrics for services and role instances that are running on your clusters. These metrics are monitored against configurable thresholds and can be used to indicate whether a cluster is functioning as expected. You can view these metrics in the web client, which displays metrics about jobs, pods, services, clusters, and so on. Cloud Pak for Data deploys and integrates several types of database and message systems. For example, Cloudant is a distributed database that is optimized for handling heavy workloads that are typical of large, fast-growing web and mobile apps. Available as an SLA-backed, fully managed cloud and on-prem service, Cloudant elastically scales throughput and storage independently. Another example is using Kafka, which is a distributed commit log service. Kafka functions much like a pub-sub messaging system but with better throughput, built-in partitioning, replication, and fault tolerance. Kafka is a good solution for large scale message processing applications. Influxdb is yet another example—it is a time-series database that can be used to store sensors, logs, and other data over a period of time. Influxdb has seen significant traction and is known for its simplicity and ease of use along with its ability to perform at scale.

Additional CP4D add-ons

Cloud Pak for Data allows customers to extend the functionality with add-ons and integrations. Add-ons are services that are deployed in your Cloud Pak for Data cluster. Integrations are connections to applications that run outside of your Cloud Pak for Data cluster.

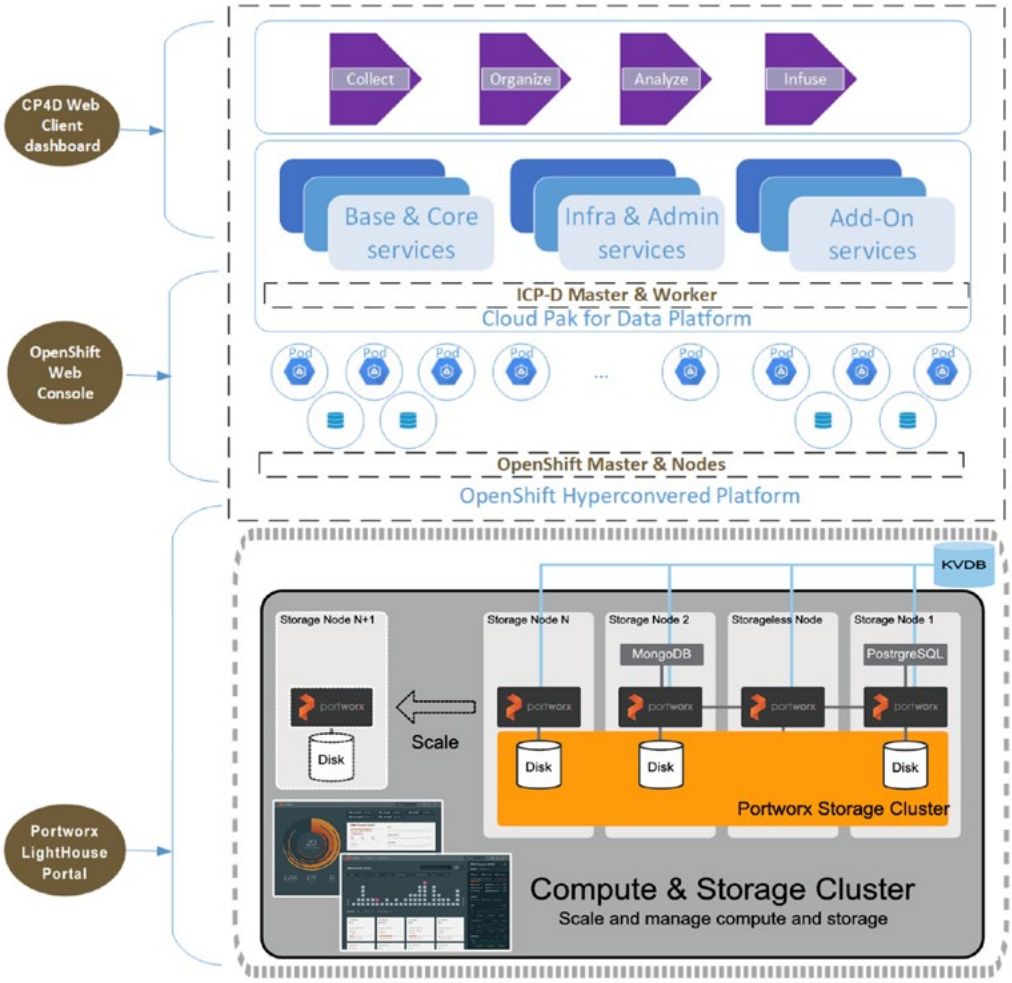
Cloud Pak for Data includes a catalog of add-ons: AI, Analytics, Dashboards, Data governance, Data sources, Developer tools, and Industry accelerators. For a full list of available add-ons see the "Catalog of Data Services" section below or visit the relevant IBM add-ons/services page.

5.1.2 Hardware infrastructure components

This section describes the server hardware infrastructure for the Cloud Pak for Data reference architecture. Although the term hardware is used, it's understood that it can also be virtualized (such as the case with on-prem or in the cloud).

As described in the previous sections, Cloud Pak for Data is deployed on top of OpenShift. In this reference architecture, we deployed Cloud Pak for Data in an OpenShift cluster enabled for enterprise-class storage by Portworx.

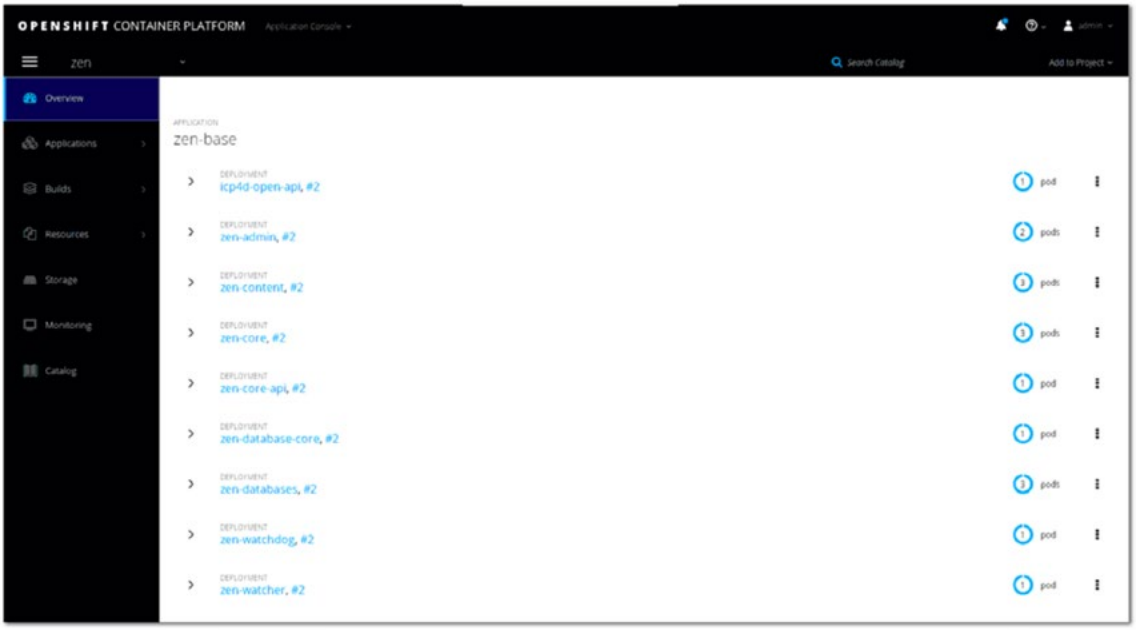
The below figure shows the Cloud Pak for Data deployment architecture with OpenShift/Portworx-powered compute/storage nodes. This deployment consists of one single OpenShift cluster with either on-prem or cloud-based servers and storage.



Each host system is a node that runs the OpenShift components on top of RedHat Enterprise Linux operating system. The Portworx platform provides persistent storage via persistent volume claim (PVC) with any type of attachable storage (anything that Linux can see as a block device). Examples of these are HDDs, SSDs, and NVMe devices.

OpenShift clusters consist of two types of nodes—Master and Worker. A Master node provides the core API and management services for the Kubernetes cluster. A Worker node provides the container run time environment along with compute-resources via OpenShift, storage layer via Portworx, and network services for the Kubernetes cluster.

A Cloud Pak for Data deployment consists typically of an odd number of Master nodes for high availability and three or more Worker nodes. Cloud Pak for Data is deployed as a set of containers that provide a variety of data-focused micro-service applications on the Kubernetes platform, meaning they are implemented on a set of pods on OpenShift/Portworx nodes that make up a cluster. These set of containers/microservices are visible in the following image.



The architectural overview is summarized above; however, for more details, see the relevant IBM CP4D architecture page.

5.2 Infrastructure Requirements

5.2.1 For Cloud Pak for Data and OpenShift

The production deployments requirements as recommended are

- Master—at least three, with 8 x86-64 vCPUs / 32 GB ram
- Worker—at least three, with 16 x86-64 vCPUs / 128 GB (recommended) (64 GB min)

To be able to use the latest version (at the time of this writing) of IBM Cloud Pak for Data version 2.5.x, you will also need to have OpenShift version 3.11 (or higher) set up on all your cluster nodes (OpenShift 4.x will be supported in a future version of Cloud Pak for Data).

You also will need to create an IBM Passport account in order to be able to download the Cloud Pak for Data installation file. In addition, you also need to create a project in OpenShift and create a required security context constraint. Then you need to create a cluster role-binding and bind it to the default service account (more details below).

5.2.2 For Portworx

Additionally, all worker nodes will need to have Portworx installed and will require at least 1 TB of storage, presented to each node, as one or more raw block devices. Additional requirements are listed on the relevant Portworx prerequisites page.

Per the Portworx Limited Use Rights, you can use the limited free-entitlement version of Portworx for Cloud Pak for Data, which allows the following limits:

- A maximum of 8 storage nodes per Portworx cluster
- A maximum of 5 TB capacity per volume

If you require more nodes in the cluster or volumes larger than 5tb, you should contact sales@portworx.com to obtain a full license to go above these free-tier limits.

In order to ensure that the storage system will have acceptable performance, latency, and throughput of the raw block, devices can be queried using the following:

- Disk latency test - should be comparable/better: **1.8 sec, 286 KB/s**

```
dd if=/dev/zero of=/path-to-installation-directory/testfile bs=512 count=1000
oflag=dsync
```

- Disk throughput test - should be comparable/better: **5.15 sec, 209 MB/s**

```
dd if=/dev/zero of=/path-to-installation-directory/testfile bs=1G count=1
oflag=dsync
```

All of the above-mentioned requirements are overall guidelines; however, greater detail can be found on the relevant IBM CP4D requirements page.

6. DEPLOYMENT/INSTALLATION OVERVIEW

If your cluster will have direct access to the internet, the process is slightly more straightforward. Otherwise, you will be working with what is known as an "air-gapped" cluster, which requires a few special extra steps to manually provide the installation files needed.

To end up with a functional basic CP4D deployment, there are three main stages that need to be implemented. The high-level outline of these stages is as follows (with details in the following section):

1. Provision the server "hardware" infrastructure, either on-prem (possibly virtualized) or in a cloud environment (such as AWS, Google Cloud, Azure, etc.).

    a. Additionally, at this point you would make sure to have an el7-based Linux system (e.g., CentOS 7.x, RedHat Enterprise Linux 7.x, etc.) installed on all the hosts, then set-up OpenShift 3.11.

    b. The OpenShift container runtime interface **must** be configured to be cri-o—and **not** docker. A mixed-mode cluster (where some nodes have different container runtimes) may work for other workloads, but this document will assume cri-o only is used for the worker-nodes that CP4D will run on.

c. As an example of a cloud deployment utilizing the AWS EC2 service, a minimum instance type/count recommended would be three `m4.4xlarge` instance (which meets the 16 core, 64gb requirements in the previous section).

2. Once the OpenShift cluster is running properly, we can then deploy Portworx. (Note: This stage will require the cluster-admin permissions in the OpenShift cluster.)

   a. Included with CloudPak for Data is a bundled Portworx installation that provides the use of the limited-use license (see previous section for details).

   b. Alternatively, if your needs will exceed that which the limited-entitlements license allows, you may instead install the full Portworx Enterprise product, which includes a 30-day trial license (that you can update to a full license at the end of the trial period).

3. Finally, start the deployment of the CP4D control plane:

   a. Obtain the installation files necessary, since before you can install CP4D, you must ensure that the installation files (within the cloudpak4data-ee-v2.5.0.0.tgz tarball) are available on the client system where you will be running `oc` client.

   b. On an air-gapped cluster, the required files must first be manually made available to the cluster before installation.

   c. Finally, the cluster plane can be set up.

4. Complete any post-installation tasks desired.

5. Install and set up any additional data services.

Each stage builds on and requires the previous stage to be performed successfully.

6.1 Stage 1—Infrastructure, operating system, and container orchestration

The steps mentioned in stage 1 are outside the scope of this document, so they will not be discussed in detail. However, there are links in the "References" section at the end of this document where more info can be gathered.

6.2 Stage 2—Deploying and configuring Portworx

You can either use the limited-entitlement Portworx instance that is included with IBM® Cloud Pak for Data (option A), or you can use a trial or existing licensed Portworx Enterprise (option B).

6.2.1 CloudPak-bundled Portworx

If you are using the CP4D entitled Portworx product, the files that you need to set up Portworx are included in the Cloud Pak for Data package on IBM Passport Advantage®.

When you extract the installation package, you will find a tarball named `cpd-portworx.tgz`, and after extracting that, there will be a README and some binaries and scripts.

The README contains the steps needed in more detail, but as an overview of the process:

1. Fetch the container images needed for Portworx install to the local system into the temporary directory `/tmp/cpd-px-images` using

   ```
   mkdir -p /tmp/cpd-px-images; sudo bin/px-images.sh -d /tmp/cpd-px-images download
   ```

   Note: This script needs to be run as the root user on the master node since it needs to use the podman utility to manage the images.

2. Authenticate with the OpenShift cluster using the oc command as the cluster administrator:

   ```
   oc login -u system:admin -n default
   ```

3. Activate the bundled-limited Portworx IBM Cloud Pak for Data license, using

   ```
   bin/px-util initialize [--sshKeyFile ssh-key-file]
   ```

   Note: This step requires being able to ssh as root onto the master/worker nodes; however, some cloud-based VMs (such as on AWS) prevent direct root logins. These restrictions will need to be disabled by editing the /root/.ssh/authorized_keys file.

4. Using the px-images.sh script again, you will now stage the downloaded container images on each host:

   ```
   bin/px-images.sh -e 'ssh -o StrictHostKeyChecking=no -l root' -d /tmp/cpd-px-images load
   ```

   Note: Again, this script needs to be able to log into hosts as the root user since podman is used again.

5. Now the Portworx software installation can begin, using

   ```
   bin/px-install.sh -pp Never install
   ```

   The `Never` argument is used as the PullPolicy since the images were staged onto each node already.

6. With Portworx installed, the various StorageClasses can be created using

```
bin/px-sc.sh
```

In case there are any issues, diagnostics should be collected and sent to Portworx support for troubleshooting. Alternatively, if necessary to re-try the installation, Portworx should only be uninstalled using

```
bin/px-uninstall.sh
```

Note: It is very important that you do not simply delete the various resources—such as Portworx-related StatefulSets/Deployments/etc via `oc/kubectl`—as the installation of Portworx includes systemd services and files outside of the control of what those commands will be able to modify.

6.2.2 Full Portworx Enterprise

The installation of the full-featured Portworx Enterprise on OpenShift is well documented on the relevant Portworx-on-Openshift installation page. Please see that documentation for the detailed steps to follow along.

While setting up this type of Portworx cluster, your configuration must also support dynamic storage provisioning with ReadWriteMany access on the persistent volumes.

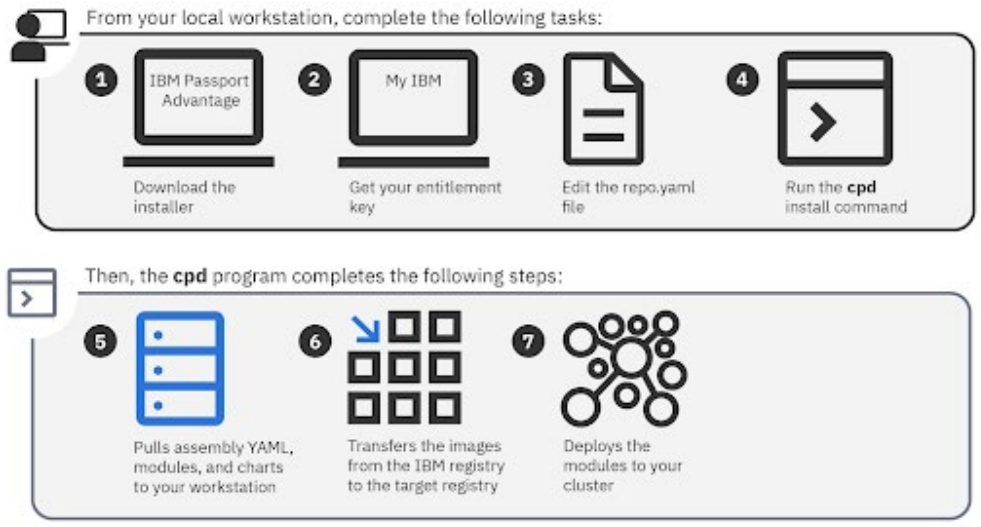
Finally, the cluster-admin also needs to create the storage classes. The full list, along with commands needed, is available on the relevant IBM CP4D specific StorageClasses page.

However, the script from the bundled Portworx tarball discussed earlier in step six can also be used with the full Portworx Enterprise product.

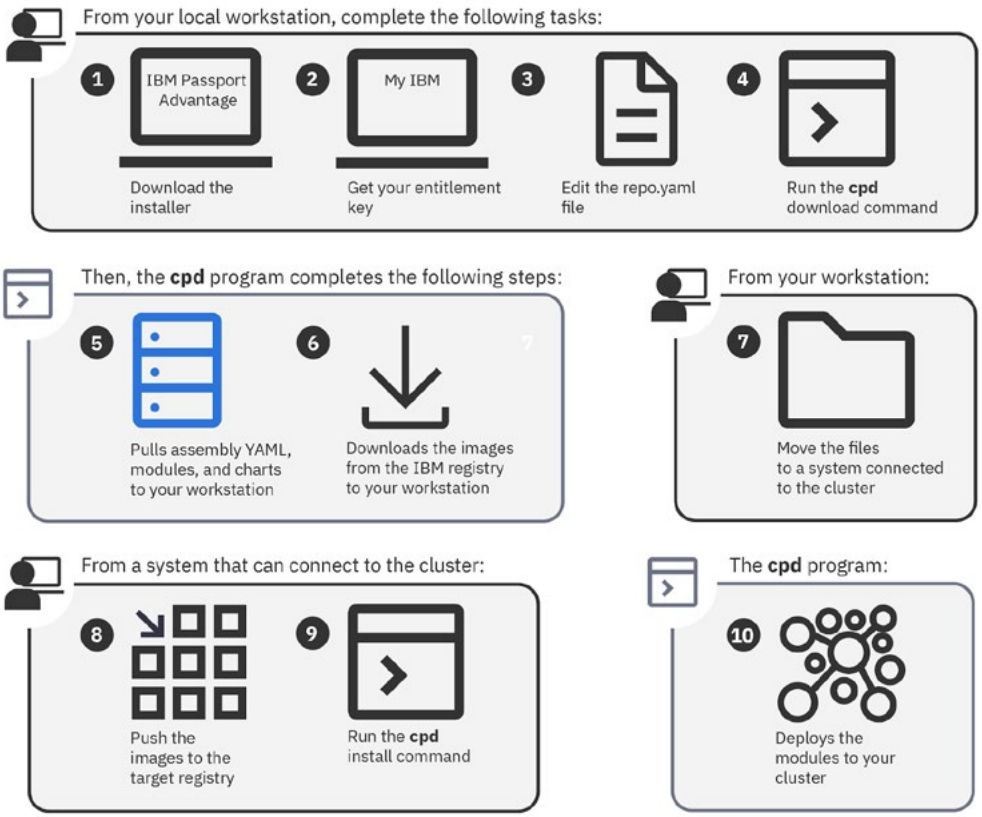
6.3 Stage 3—CP4D base and control plane deployment

After you review the system requirements and other planning information, you can only install IBM Cloud Pak for Data by completing the pre-installation tasks, completing the installation task itself, and then completing the post-installation tasks. When finished, you will have installed the Cloud Pak for Data control plane. Services are installed separately.

For internet-connected OpenShift clusters, the overall process looks like this:



For air-gapped clusters, the process looks slightly different:



Both variations are discussed extensively in detail on the relevant IBM CP4D Install on OpenShift page.

As an outline, the steps in this stage look like this:

1. A new account for administering CloudPak is created and permissioned:

```
sudo htpasswd -b /etc/origin/master/htpasswd cpadmin mycpadmpassword && oc
adm policy add-cluster-role-to-user cluster-admin cpadmin
```

2. We setup a session to run the oc commands using

```
oc login -u cpadmin -p mycpadmpassword
```

3. Before we can prepare and install CP4D, we have to make sure our entitlement key is populated into the `repo.yaml` file that was created when the installation package was extracted.

   This will be available from the entitlements page once you've created a profile on the My IBM service, and it is usable only after the CP4D product has been procured. You will need to place that key value into the `repo.yaml` (at end of the *apikey:* line).

   Without this step in place, step six can fail.

4. The control plane is prepped to be installed using

```
chmod +x bin/cpd-* && bin/cpd-linux adm --repo repo.yaml -a lite -n cpd
--apply
```

   With the above parameters, we will work with the lite assembly (basic control plane but no additional services included yet) into the **cpd** namespace.

5. The control plane services (installed in the next step) need storage configured to utilize Portworx by setting the StorageClass the services should use and be specified in an override file. Please see the relevant IBM page on this topic.

6. The control plane can now be finished being loaded up:

```
sudo bin/cpd-linux -s repo.yaml -a lite --verbose -o cp-override.yaml
--target-registry-password $(oc whoami -t) --target-registry-username
cpadmin --insecure-skip-tls-verify --transfer-image-to docker-registry.
default.svc:5000/cpd -n cpd -c portworx-shared-gp
```

   Note: This command uses sudo because it must have root-level access, as it will download via podman all necessary containers for the control plane and load them into the OpenShift internal docker registry (so they can be deployed within the cluster).

The control plane should now be installed and accessible from the web at the address output at the end of the installation.

6.4 Stage 4—Post-install

After you install Cloud Pak for Data, the following tasks wcan be completed.

If your cluster is configured to use a custom name for the DNS service, a project administrator or cluster administrator must update the DNS service name to prevent performance problems because when you install the IBM Cloud Pak for Data control plane, the installation points to the default Red Hat OpenShift DNS service name. Details of this step can be found on the relevant IBM post-installation page.

To ensure secure transmission of network traffic to and from the Cloud Pak for Data cluster, you need to configure the communication ports used by the network. Details of this step can be found on the relevant IBM post-installation page.

Finally, after you install Cloud Pak for Data, you can configure the web client. Details of this step can be found on the relevant IBM post-installation page. This may also involve the use of a custom TLS certificate for HTTPS access.

Afterwards you may want to enable email notifications, enable Single-Sign on via SAML, setup federation of user-access via LDAP, or otherwise manually manage users.

6.5 Stage 5—Additional CloudPak for Data service installations

You may also at this point decide to set up additional data services/features available for CP4D, which are listed in the next section.

The exact process for their installation is detailed on the relevant IBM CloudPak for Data service installation page (full list in next section). For examples of what these pages look like, see the Watson Knowledge Catalog installation page or the Watson Studio installation page.

As an overview of the installation step, the process is very similar to the installation of the control plane—using the final two commands that use cpd-linux tool detailed in the control-plane installation above, except that instead of specifying the parameter *lite* as the assembly to work with, we will specify the service abbreviation itself. So for example, installing Watson Studio, the assembly parameter becomes *wsl*, and for Watson Knowledge Catalog, it is *wkc*.

The exact *assembly* parameter value (name) can be found on the relevant service's install page. Step five (control panel section) will need to be performed to create the right override file.

Additionally, if specified, you should pay attention to any outlined cluster preparation steps on the relevant service's stage prior to the installation page.

6.5.1 Catalog of Data Services

Cloud Pak for Data includes an enterprise-class data-services catalog that helps to create a cohesive information architecture.

Service Type	From	Data Service Offering	Incl. w/ CP4D	Doc links		Description
AI	IBM	Watson™ Assistant	no *	install	usage	Build conversational interfaces into any app, device, or channel.
AI	IBM	Watson Assistant for Voice Interaction	no *	install	usage	A bundle of services that deliver a Watson-based voice automation system.
AI	IBM	Watson Discovery	no *	install	usage	Find answers and uncover insights in your complex business content.
AI	IBM	Watson Knowledge Studio	no †	install	usage	Teach Watson the language of your domain.
AI	IBM	Watson™ Language Translator	no *	install	usage	Translate text into different languages.
AI	IBM	Watson Machine Learning	yes	install	usage	Deploy machine-learning models into production at scale.
AI	IBM	Watson Machine Learning Accelerator	no *	install	usage	An end-to-end, deep learning platform for data scientists.
AI	IBM	Watson Natural Language Understanding	no †	install	usage	Take your understanding of unstructured data to a new level by extracting entities, keywords, sentiments, and more.
AI	IBM	Watson OpenScale	yes	install	usage	Infuse your AI with trust and transparency. Understand how your AI models make decisions to detect and mitigate bias.
AI	IBM	Watson Studio	yes	install	usage	Unleash the power of your data. Build custom models and infuse your business with AI and machine learning.
AI	IBM	Watson Speech to Text	no †	install	usage	Quickly convert audio and voice into written text.
AI	IBM	Watson Text to Speech	no †	install	usage	Convert written text into natural-sounding speech.
Analytics	IBM	Analytics Engine Powered by Apache Spark	yes	install	usage	Automatically spin up lightweight, dedicated Apache Spark clusters to run a wide range of workloads.
Analytics	non-IBM	Analytics Zoo for Apache Spark	free	install	usage	An analytics and AI platform that unifies TensorFlow, Keras, and BigDL for distributed Apache Spark environments.

Analytics	IBM	Cognos® Analytics	no *	install	usage	Self-service analytics, infused with AI and machine learning, enable you to create stunning visualizations and share your findings through dashboards and reports.
Analytics	non-IBM	Datameer	no *	getting started		Unlock the value in your raw data. Bring data from across your enterprise into a single, unified view so that everyone can blend, prepare, and explore the data to uncover hidden answers.
Analytics	IBM	Data Refinery	yes	install		Simplify the process of preparing large amounts of raw data for analysis.
Analytics	IBM	Decision Optimization	no ‡	install	usage	Evaluate millions of possibilities to find the best solution to any given problem.
Analytics	IBM	Execution Engine for Apache Hadoop	no ‡	install	usage	Explore data or build and deploy models on your Apache Hadoop cluster.
Analytics	non-IBM	Figure Eight	no *	install	usage	Transform text, images, audio, and videos into annotated training data to fuel your machine learning initiatives.
Analytics	non-IBM	Intel Distribution of Python	free	install	usage	Popular Python libraries, including analytics and machine learning libraries, accelerated for Intel architecture.
Analytics	non-IBM	Intel Deep Learning Reference Stack - PyTorch	free	install	usage	A Deep Learning Reference Stack for PyTorch that is optimized for Intel architecture. Requires AVX-512.
Analytics	non-IBM	Intel Deep Learning Reference Stack - TensorFlow	free	install	usage	A Deep Learning Reference Stack for TensorFlow that is optimized for Intel architecture. Requires AVX-512.
Analytics	non-IBM	Operational Analytics for ERP	no *	install	n/a	Visualize and analyze operational SAP ERP data in Cognos Analytics.
Analytics	IBM	SPSS® Modeler	no ‡	install	usage	Create flows to prepare and blend data, build and manage models, and visualize the results.
Analytics	IBM	Streams	yes	install	usage	Build solutions that drive real-time business decisions by combining streaming and stored data with analytics.
Analytics	non-IBM	WAND Foundation Taxonomies	no *	req for info		WAND Industry and Business Function Taxonomies provide relevant terminology to jump start your business glossary.

Dashboard	IBM	Analytics Dashboards	yes	install	usage	Identify patterns in your data with sophisticated visualizations. No coding needed.
Data Governance	IBM	DataStage® Edition	no *	install	usage	Effortlessly deliver data at the right time to the right place with integration, transformation, and delivery of data in batch and real time.
Data Governance	IBM	Regulatory Accelerator	yes	install	usage	Streamline the process of complying with regulations.
Data Governance	non-IBM	Senzing	no *	install	n/a	Real-time AI for entity resolution that scales with your data. Exploit the power of your data with minimal data preparation and transformation. Discover the people and places at play in your data.
Data Governance	IBM	Watson Knowledge Catalog	yes	install	usage	Find the right data fast. Discover relevant, curated data assets using intelligent recommendations and user reviews.
Data Sources	non-IBM	CockroachDB	no *	install	usage	A SQL database with the scalability of a NoSQL database and the reliability of a transactional RBDMS.
Data Sources	IBM	Data Virtualization	yes	install	usage	Query many data sources as one.
Data Sources	IBM	IBM Db2® Advanced Edition	no *	install	usage	Relational database that delivers advanced data management and analytics capabilities for transactional workloads.
Data Sources	IBM	IBM Db2 Event Store	yes	install	usage	Data store designed to rapidly ingest and analyze streamed data for event-driven applications.
Data Sources	IBM	IBM Db2 Warehouse	yes	install	usage	Data warehouse designed for high-performance, in-database analytics. Runs on a single node for cost-efficiency or on multiple nodes for improved performance.
Data Sources	IBM	Db2 for z/OS® Connector	yes	install	usage	Create databases in Db2 for z/OS and work directly with the data from Cloud Pak for Data.
Data Sources	non-IBM	MongoDB	no *	install	usage	Scalable, open source NoSQL database.
Data Sources	non-IBM	PostgreSQL	yes	install	usage	Open source object-relational database designed for developers.
Developer Tools	IBM	Jupyter Notebooks with Python 3.6 for GPU	yes	install	usage	An optional development environment for Watson Studio that enables you to create Jupyter Notebooks that use GPU-accelerated Python 3.6 libraries.

Developer Tools	IBM	Jupyter Notebooks with R 3.6	yes	install	usage	An optional development environment for Watson Studio that enables you to create Jupyter Notebooks that use R 3.6 libraries.
Developer Tools	non-IBM	Knowis Solution Suite for Banking	no *	install	usage	Design, implement, and deliver microservices that transform digital business opportunities into cloud-native solutions.
Developer Tools	non-IBM	Lightbend Platform	no *	install	usage	Develop and deploy Reactive Microservices, real-time streaming pipelines, and machine learning pipelines.
Developer Tools	IBM	Open Source Management	yes	install	usage	Make it easy for developers and data scientists to find and access approved open source packages.
Developer Tools	IBM	RStudio Server with R 3.6	yes	install	usage	Optional development environment for working with R.
Industry Solutions	IBM	Financial Crimes Insight®	no *	install	usage	Simplify the process of detecting and mitigating financial crimes with AI and regulatory expertise.
Industry Solutions	non-IBM	Prolifics Customer Prospecting Accelerator	no *	install	usage	Unlock reliable sales leads and increase customer acquisition.

Footnotes:

“no *” - not included with Cloud Pak for Data but available/priced separately
“no †” - included with Watson API Kit package, otherwise available/priced separately
“no ‡” - included with Watson Studio Premium package, otherwise available/priced separately

7. RESOURCES

For more information, see the following resources.

Red Hat OpenShift

- Product page: https://www.redhat.com/en/technologies/cloud-computing/openshift
- Version 3.11 manual: https://docs.openshift.com/container-platform/3.11/welcome/index.html

IBM Cloud Pak for Data

- Product page: https://www.ibm.com/analytics/cloud-pak-for-data
- Manual guide: https://www.ibm.com/com.ibm.icpdata.doc/zen/overview/overview.html

IBM DB2 warehouse

- Product page: https://www.ibm.com/products/db2-warehouse

portworx

Portworx, Inc.
4940 El Camino Real, Suite 200
Los Altos, CA 94022
Tel: 650-241-3222 | info@portworx.com | www.portworx.com