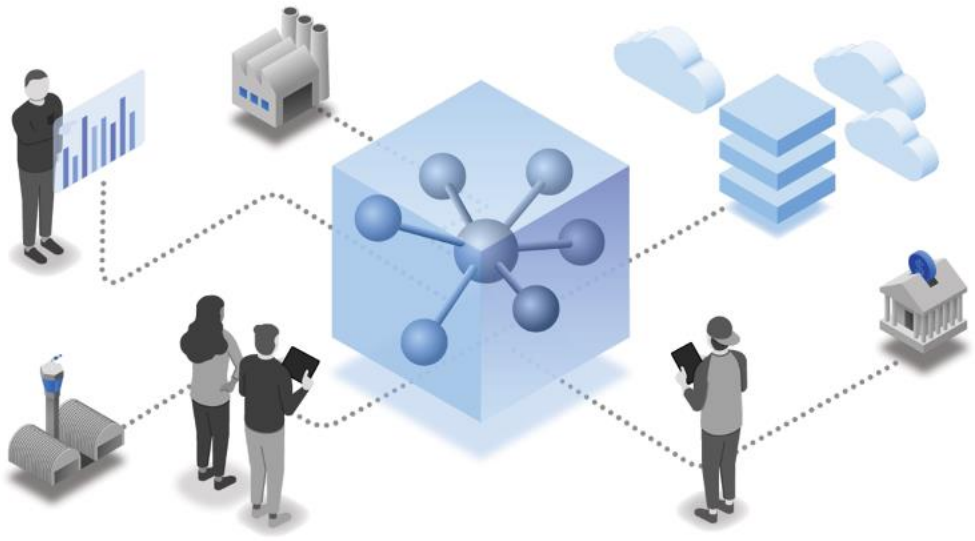
Security on Cloud Pak for Data V3.5

For private cloud and on-premises deployments



By:

Priya Ranjan Sahoo

Security Lead - Cloud Pak for Data

Abhishek Garai

Security Engineer - Cloud Pak for Data

Sandeep Deodikar

Manager Security - Cloud Pak for Data

Sriram Srinivasan

Architect - Cloud Pak for Data, Distinguished Engineer

Table of Contents

Introduction

IBM Cloud Pak® for Data is a fully integrated data and AI platform that modernizes how businesses collect, organize, and analyze data to infuse AI throughout their organizations. Cloud-native by design, the platform unifies market-leading services spanning the entire analytics lifecycle. From data management, Data Ops, governance, business analytics, and automated AI, IBM Cloud Pak for Data helps eliminate the need for costly, and often competing, solutions while providing the information architecture you need to implement AI successfully.

IBM Cloud Pak for Data builds on the streamlined hybrid-cloud foundation of Red Hat® OpenShift® and takes advantage of its underlying resource and infrastructure optimization and management. The solution fully supports multi-cloud environments such as Amazon Web Services (AWS), Azure, Google Cloud, IBM Cloud™, and private cloud deployments.

The IBM Cloud Pak for Data, team is committed to continuously improving product security. In this Security whitepaper, we describe the overall security posture of Cloud Pak for Data.

Cloud Pak for Data follows the IBM Secure Engineering Framework which provides guidance that helps ensure that software is secure by design, in implementation, and in deployment. The framework mandates adhering to the Security and Privacy by Design (SPbD) discipline, which matures the security posture of IBM offerings by introducing security and privacy focus and checkpoints in each offering's lifecycle. The SPbD discipline covers Vulnerability management, Threat Modeling, Penetration Testing, Security QA Testing, Product Security Incident Reporting, and Code Signing. This document will briefly touch upon each of these SPbD discipline areas. This document will also show how IBM Kubernetes Certification & Red Hat Image Certification ensure that the container images that are delivered as part of Cloud Pak for Data meet all the Kubernetes best practices.

Because it’s built on the Red Hat OpenShift container platform, Cloud Pak for Data utilizes all the security features available in the latest OpenShift releases to deliver a secure platform. This document explains how Cloud Pak for Data is utilizing the Red Hat Open shift features and the special security hardening steps that are followed to ensure that the Cloud Pak for Data platform is secure. The document also explains the various security capabilities that are built into the Cloud Pak for Data platform like Authentication and Authorization, User/Role management, Encryption, Audit Logging, and Multi tenancy.

Finally, this document lists the Privacy and Compliance assessments that Cloud Pak for Data product is assessed for, while also providing guidelines on how clients can use that in their preparations for various privacy and compliance assessments, to ensure their own readiness for applicable laws and regulations.

1.0 IBM Secure Engineering Practices

The Cloud Pak for Data offering is developed following a sophisticated and comprehensive Security and Privacy by Design (SPbD) discipline, which is managed by the IBM Hybrid Cloud Business Information Security Office (BISO) Review Board. The objective of the SPbD discipline is to mature the security posture of our offering by introducing security and privacy focus and checkpoints in the offering's lifecycle.

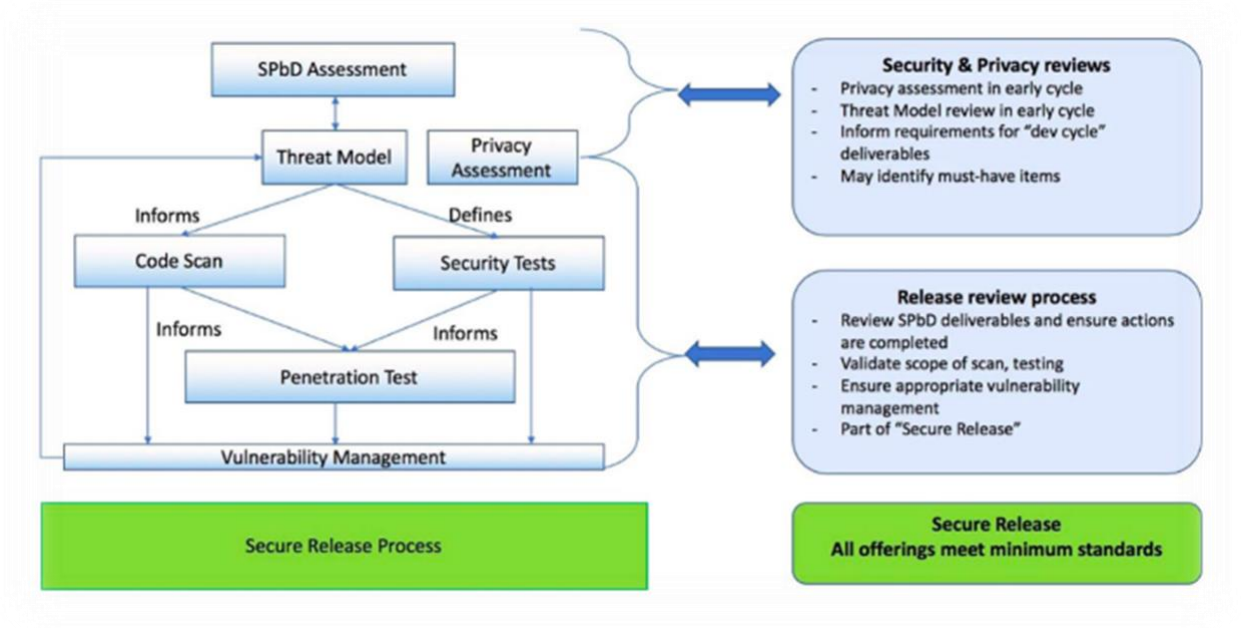


Figure 1.0: Secure Release Process in Security and Privacy by Design discipline

Cloud Pak for Data releases go through a thorough yearly review by the IBM Hybrid Cloud Business Information Security Office (BISO).

For more details, see IBM Security and Privacy by Design.

The following activities are undertaken as mandated by the SPbD process.

1.1 Vulnerability management

Vulnerability management is a proactive approach to find vulnerabilities in systems and applications, so they can be fixed before they have the opportunity to be exploited by attackers. Vulnerability scanning and assessment is done in a continuous cycle for every service on Cloud Pak for Data in the following areas:

- Image vulnerability scan for each container image
- Static source code scan
- Dynamic scan
- Open source scan

The image vulnerability scan checks for vulnerable packages in every new image officially promoted that is using supported operating systems.

Code scans (static, dynamic, and open-source scans) are performed periodically during the release cycle, which include static (SAST), dynamic (DAST), and open-source security testing capabilities.

All of the tools used in the development of Cloud Pak for Data are approved by the IBM CISO (Corporate Information Security Officer).

1.2 Threat modeling

Threat modelling is a cornerstone of the Security and Privacy by Design (SPbD) discipline and Secure Engineering in general. Cloud Pak for Data goes through a detailed threat model review before each release, which helps ensure that potential weaknesses in a solution's design are identified and mitigated during the development cycle.

1.3 Penetration testing

The goal of Penetration testing is to deliver a vulnerability-free offering. Penetration testing is a key defense-in-depth Secure Engineering practice.

Cloud Pak for Data undergoes penetration testing at least once each year. The testing is performed by an independent team that is external to the Cloud Pak for Data product development team. The findings from penetration testing follow the remediation plan defined by the IBM PSIRT (Product Security Incident Reporting Tool) process and are reviewed by IBM Hybrid Cloud Business Information Security Office (BISO).

1.4 Security QA testing

Security QA testing is part of the Security and Privacy by Design discipline. Security QA Testing ensures software systems and applications are free from vulnerabilities, threats, and risks that may cause loss or compromise security**.** Security QA tests are run during the development phase to help ensure that Cloud Pak for Data is secure from both external and internal vulnerabilities and threats.

1.5 Patch management

Cloud Pak for Data releases use a patch management process to provide the fixes for any security vulnerabilities and any critical issues found post release.

1.6 IBM Product Security Incident Response team (PSIRT)

The IBM Product Security Incident Response Team (PSIRT) manages the receipt, investigation and internal coordination of security vulnerability information related to IBM offerings. This process reduces the risk to our customers by ensuring timely identification, analysis, resolution, and reporting of security vulnerabilities. For more details, see IBM Security Vulnerability Management (PSIRT).

1.7 Code signing

Code signing is an integral part of the overall development and build process for Cloud Pak for Data. Code signing is extremely important for building client trust and safely distributing the software because code signing allows clients to:

- Confirm the provenance of a software package.
- Verify the integrity of the software package.

1.8 Container signing

Container signing extends the capabilities of code signing to individual containers and provides:

- Decoupled validations of configuration and functional code.
- Facilitated authenticity of the code for the deployed applications.

All of the container images that are delivered as part of Cloud Pak for Data are signed to ensure the integrity of images.

2.0 IBM Kubernetes certification and Red Hat Image certification

All the containers in Cloud Pak for Data go through IBM Kubernetes certification and use Red Hat Certified Unified Base Images only.

The IBM Kubernetes Certification Framework is a consolidated set of requirements and processes for all IBM Containerized Content being developed to run on Red Hat OpenShift. The IBM Kubernetes Certification Framework builds on the Red Hat Image Certification and Red Hat Operator Certification and ensures that products being delivered as a Cloud Pak or as a standalone product meet the following criteria:

- Meet the requirements for production grade, secure lifecycle management, services integration, and support.
- Are implemented using consistent and opinionated Kubernetes best practices.

The intent of the IBM Kubernetes Certification Framework is to:

- Differentiate IBM Kubernetes content on Red Hat OpenShift.
- Drive consistency for all IBM software to our clients.
- Deliver enterprise capabilities to our clients.

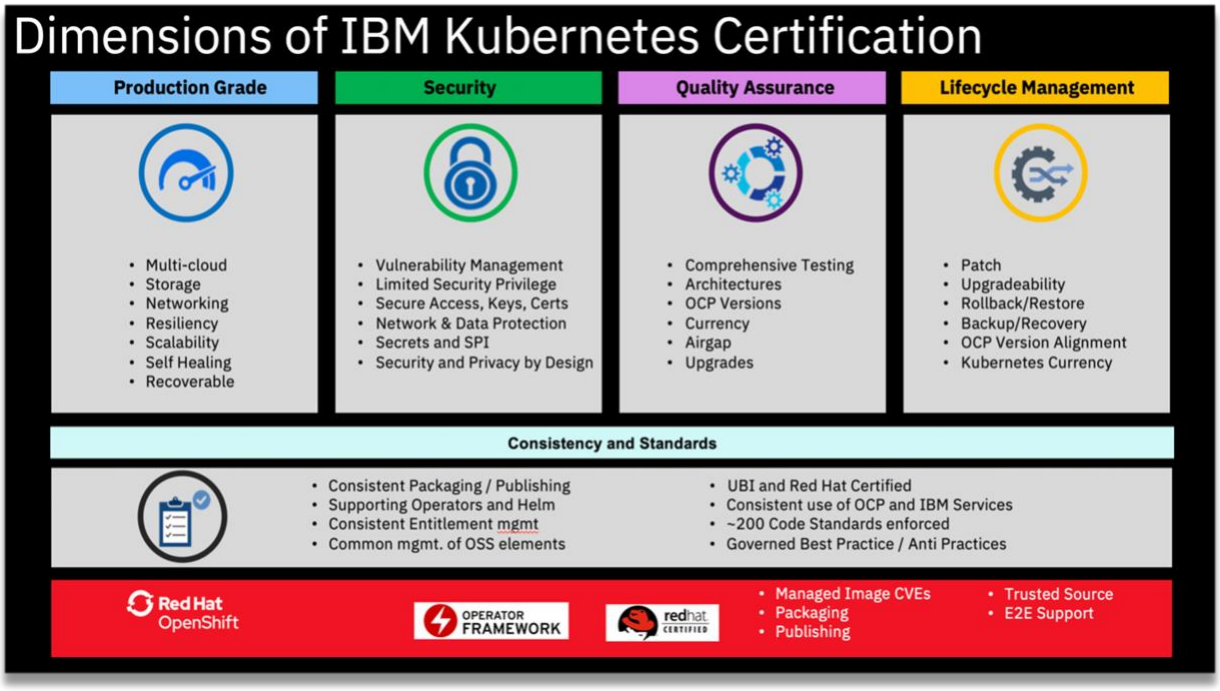


Figure 2.0: Dimensions of IBM Kubernetes Certification

All the containers in Cloud Pak for Data use Certified Red Hat Universal Base images (UBI) only. For more details, see Building, running, and managing containers in the Red Hat documentation.

3.0 Cloud Pak for Data Security features

Security is a paramount need for nearly every enterprise, particularly for organizations in the government, financial services, and healthcare sectors. OpenShift container platform provides a set of security features. These features help protect sensitive customer data with strong encryption controls and improve the oversight of access control across applications and the platform itself. Cloud Pak for Data builds over the security features provided by OpenShift by creating Security Context Constraints, Service Account and Roles in such a manner, that Cloud Pak for Data pods and users have only the least amount of privileges to the OpenShift platform

that is needed for them. Cloud Pak for Data is also security hardened on OpenShift platform and is installed in a secure and transparent manner.

3.1 Security features on OpenShift Container Platform

OpenShift Container Platform (v4.5 and above) enables an improved security posture with the addition of many capabilities that greatly increase the security of the platform.

- OpenShift Container Platform uses Red Hat CoreOS as the immutable host operating system.
- OpenShift Container Platform has stronger platform security with FIPS (Federal Information Processing Standard) compliant encryption (FIPS 140-2 Level 1).
- OpenShift Container Platform has capabilities, such as the Node Tuning Operator, which provides opportunities to further reduce privilege requirements in the Security Context Constraints (SCC).
- OpenShift Container Platform supports encrypting data stored in etcd, which provides additional protection for secrets stored in the etcd database.
- OpenShift Container Platform also provides a Network Bound Disk Encryption (NBDE) feature that can be used to automate remote enablement of LUKS encrypted volumes, making it easier to protect against physical theft of host storage.
- SELinux is mandatorily enabled on OpenShift Container Platform.

3.2 Security Context Constraints

OpenShift provides *Security Context Constraints* (SCC) that control the actions that a pod can perform and what it can access. OpenShift provides a set of predefined SCCs that can be used, modified, or extended by any administrator. Notable amongst them are *privileged* and *restricted*. By default, the execution of any container will be granted the *restricted* SCC and only the capabilities defined by that SCC.

Cloud Pak for Data does not require or recommend the use of the *privileged* SCC in OpenShift because it is too open. However, the *restricted* SCC does not enable the operations of a multi-tenant, multi-user platform like Cloud Pak for Data. Therefore, the Cloud Pak for Data control plane and add-on services carefully identify incremental privileges over *restricted* based on the principle of least privilege.

By default, Cloud Pak for Data creates the following three Security Context Constraints:

cpd-user-scc

The cpd-user-scc is identical to RedHat OpenShift Container Platform's out-of-the-box restricted SCC except for the uid range and fsGroup policy.

- The cpd-user-scc constrains the pods to run only in the uid range: 1000320900 to 1000361000. That is, running as root is prevented.
- OpenShift assigns a unique Multi-Category Security (MCS) label in SELinux to each OpenShift project. This ensure that pods from one namespace *cannot* access files created by pods in another namespace or by host processes with the same uid. For details, see A Guide to OpenShift and UIDs.
- The fsGroup policy is needed to continue to *retain file ownership* & permissions when multiple pods mount the same volume and to work around the Kubernetes problem described in Configure volume permission and ownership change policy for Pods.
- Typical Linux docker capabilities are dropped – [KILL MKNOD SETUID SETGID]. This constraint prevents the use of "privileged container".
- Host network access and Inter Process Communication (IPC) are prevented.
- Access to any host path volumes is prevented.

For reference, see the yaml file: cpd-user-scc.yaml

cpd-zensys-scc

The cpd-zensys-scc is also identical to RedHat OpenShift Container Platform's out-of-the-box restricted SCC except for a single uid use, retention of SETUID/SETGID capabilities, and fsGroup policy.

This SCC is made to set up persistent volume mounts to ensure correct ownership of files, especially in shared volumes. For example, when new users are added to Cloud Pak for Data, each user has a unique uid generated and home directories created whose ownership is then changed to that uid. In addition, permissions are set only for that user. While Cloud Pak for Data uses subPaths in end-user pods (like Jupyter and RStudio) to ensure that only their directories are visible, extra precaution is taken while setting file permissions and ownership for that uid, to ensure that only the required directories are visible.

- The cpd-zensys-scc SCC allows processes to run with the uid 1000321000 only. This SCC is meant for Cloud Pak for Data system pods. Other workloads, especially end-user triggered, will not be associated with this SCC.
- This SCC retains [SETUID SETGID] for the system uid to do chown/chgrp/chmod on mounted volumes, but also drops [KILL MKNOD].
- Other restrictions that cpd-user-scc imposes (including non-root and Multi-Category Security (MCS) label in SELinux) are enforced with cpd-zensys-scc.

For reference, see the yaml file: cpd-zensys-scc.yaml.

cpd-noperm-scc

The cpd-noperm-scc SCC is similar to the out-of-the-box restricted SCC in the OpenShift container platform. The default definition of the out-of-the-box restricted SCC can potentially be modified by customers to alter the priority ordering or make other customizations that could accidentally grant more privileges than appropriate. Hence this custom SCC (similar to restricted SCC) helps mitigate this problem by granting only the very limited amount of privileges. The cpd-norbac-sa (similar to the out-of-the-box default service account) binds to this SCC.

Notes:

1. The SCCs are created only once per cluster. So, even if you have multiple copies of Cloud Pak for Data installed in different namespaces, these SCCs need to be created only once for the entire cluster.

2. Cloud Pak for Data contains multiple add-on services. A few services have exceptions to the rules defined above. These services might define *additional* SCCs, for example to support IPCs. The **cpd adm** command for those additional assembly packages will show the details of what is required for those add-on services.

3.3 Service accounts and roles

Cloud Pak for Data runs in a separate namespace/project on the OpenShift cluster. In the OpenShift project, Cloud Pak for Data creates service accounts and RBAC role bindings for pods to use within that namespace.

- No cluster level access is permitted. All roles impose a restriction to work within that namespace only.
- Two roles are created: **cpd-admin-role** and **cpd-viewer-role**. These roles allow Cloud Pak for Data to ensure that the principle of least privilege can be applied even within the same namespace.

Role	Description
cpd-admin-role	Allows for creation/updates/deletes of Kubernetes deployments, secrets, configmaps, run jobs, and more, but only within that specific namespace.
cpd-viewer-role	Allows for Kubernetes API calls (GETs) but cannot perform any creates or deletes.

- Four service accounts (**cpd-admin-sa**, **cpd-editor-sa**, **cpd-viewer-sa and cpd-norbac-sa**) are created in addition to the default service account, which is created by default in every OpenShift project.

Account	Description
cpd-admin-sa	Bound to the cpd-admin-role and is associated with a system user and the cpd-zensys-scc. This account is used for preparing persistent volume mounts, setting file permissions, and more.
cpd-editor-sa	Bound to the cpd-admin-role also but is associated with the cpd-user-scc, and thus has reduced privileges.
cpd-viewer-sa	Bound to the cpd-viewer-role and the cpd-user-scc. This account can only view Kubernetes artifacts within that namespace.
cpd-norbac-sa	Bound to the cpd-noperm-scc but is not associated with any Kubernetes role thus has no API access. This account is similar to the out-of-the-box default service account. It is used by services or components that should not be granted any RBAC privileges.
default	Automatically created in every OpenShift project, bound to the restricted SCC and is not granted any RBAC privileges; that is, no roles are bound. This default service account will be used for end-user workloads such as notebooks and Python jobs and will not be allowed to perform any kind of actions inside the namespace.

See the following yaml files for reference: cpd-admin-role.yaml and cpd-viewer-role.yaml.

3.4 Security hardening

Security hardening is enforced on Cloud Pak for Data on Red Hat OpenShift. The following security hardening actions are taken:

- Only non-root processes are run in containers. The uid of the processes are in the pre-defined range only. This restriction is enforced by the SCC defined in section 3.2.
- No cluster-admin privileges are allowed. Cluster-admin authority is needed only to set up the project and accounts. Each Cloud Pak for Data instance needs only the privileges within its own RHOS project/namespace.
- Cloud Pak for Data users are typically not granted OpenShift Kubernetes access, and even if they are, it would be only for express purpose of installing or upgrading services inside their assigned OpenShift project.

- Strict use of service accounts with RBAC privileges is enforced, and the least privilege principle is applied. Cloud Pak for Data ensures that any pod that is running end-user code (such as scripts or analytics environments) is not granted any RBAC privileges.
- No host access is allowed. This restriction is enforced by the SCC. There is no access to host paths or networks.
- All pods have restricted resource consumption. Pod resource requests and limits are set for each pod, thereby restricting the consumption. This helps protect against noisy neighbors that cause resource contention.
- Reliability gauges (liveness and readiness probes) are present for each pod to ensure that the pods remain healthy.
- For consumption monitoring, each of the pods on Cloud Pak for Data is annotated with metering annotations to uniquely identify add-on service workloads on the cluster.

Note: Some of the add-on services on Cloud Pak for Data might have exceptions to the security hardening and are being tracked to ensure compliance in future releases.

3.5 Secure and transparent installation process

SSH to OpenShift cluster nodes is not needed to deploy or manage Cloud Pak for Data and its add-on services. The Cloud Pak for Data installation command-line interface is used to deploy and manage the Cloud Pak for Data and its add-on services.

You can install the software on a cluster that is connected to the internet or a cluster that is air-gapped. The following sections explain how the installation process works in each environment.

3.5.1 Installing Cloud Pak for Data on internet-connected clusters

When you run the **cpd** installation command from a client workstation, it downloads Helm charts from a public IBM file server, transfers all images from the entitled IBM Docker registry to your registry server, and deploys the Helm charts to a project in your OpenShift® cluster.

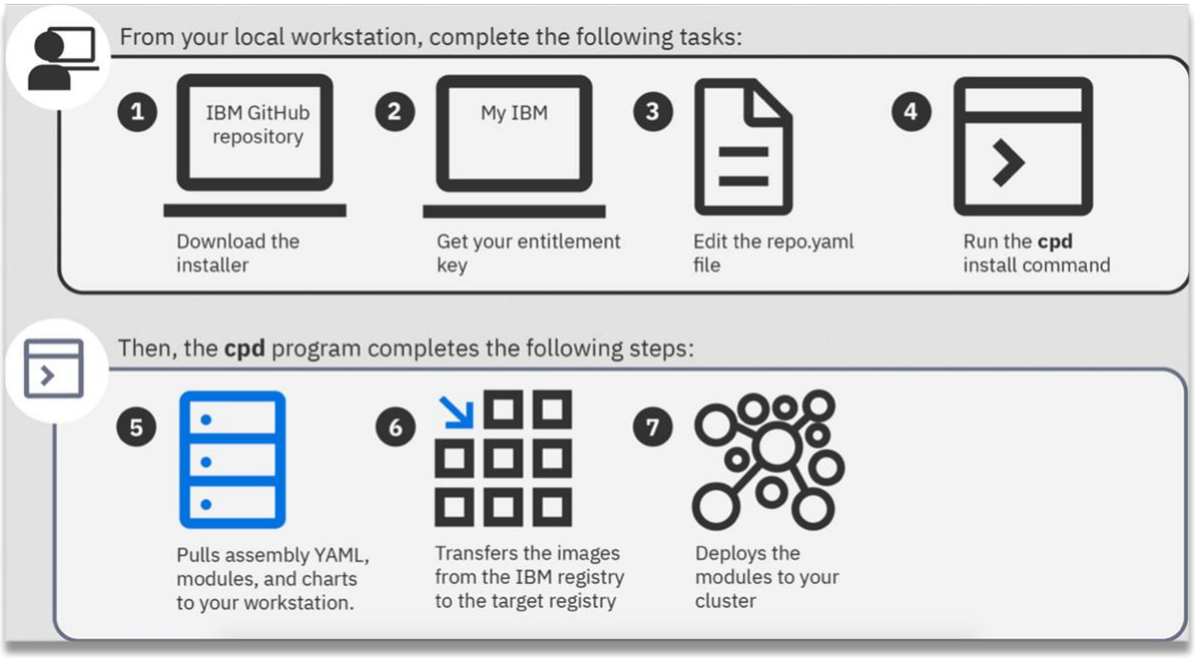


Figure 3.1: Flow of operations when installing Cloud Pak for Data on a cluster connected to the internet

3.5.2 Installing Cloud Pak for Data on air-gapped clusters

When you run the **cpd** download command from a client workstation, it downloads all of the images and Helm charts from a public IBM file server to the client workstation. Next, you move the Cloud Pak for Data command-line interface and the downloaded files to a system that can connect to the registry server and the cluster. From the system that can connect to the cluster, run the **cpd** command to push the images to your registry server. Lastly, you run the **cpd** installation command, which deploys the Helm charts to a project in your OpenShift cluster.

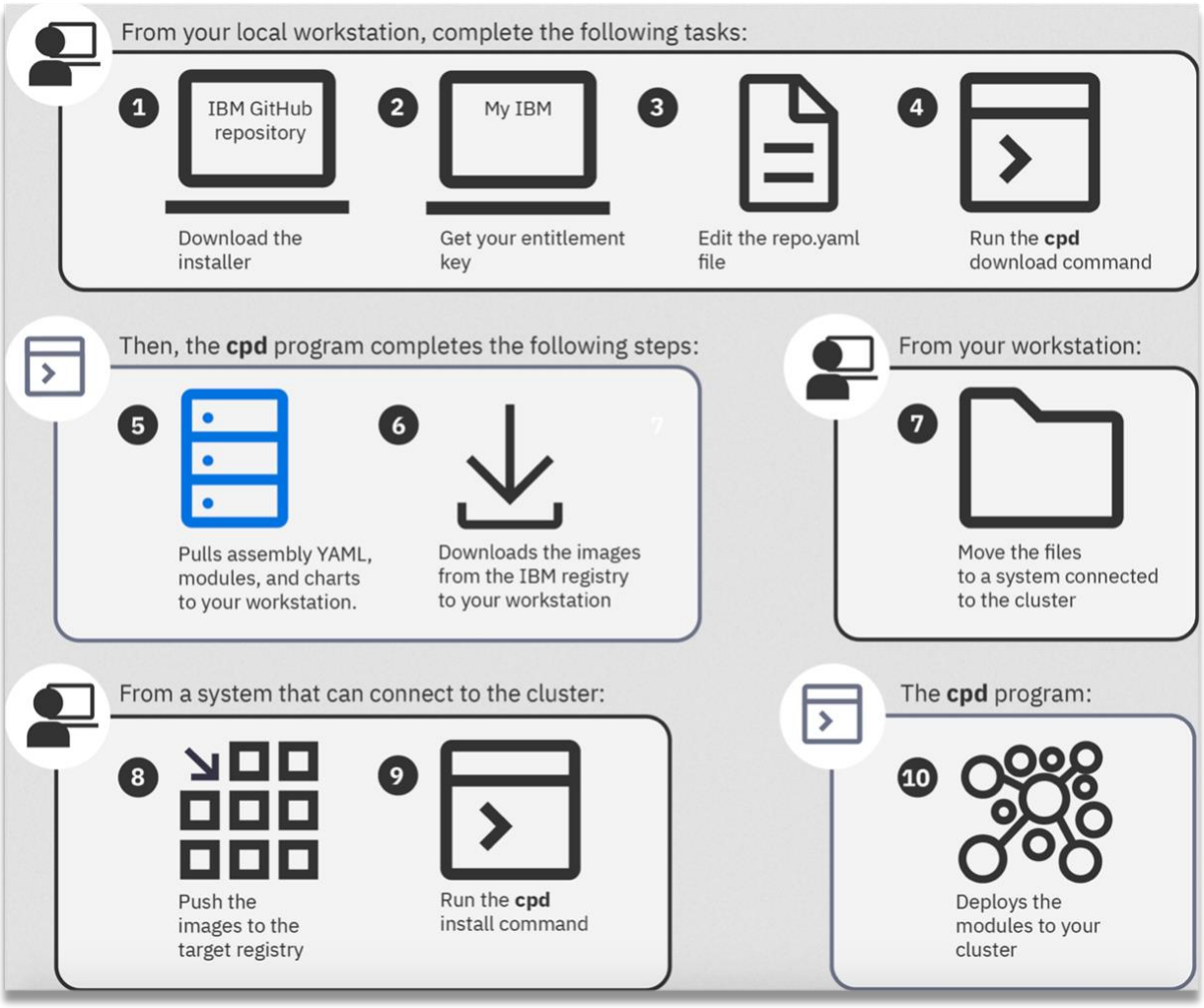


Figure 3.2: Flow of operations when installing Cloud Pak for Data on an air-gapped cluster

3.5.3 Transparent cluster changes

When Cloud Pak for Data in installed on an OpenShift cluster, it makes needed changes to the OpenShift cluster as explained in the previous sections.

To see what changes are needed to be made to the cluster, you can run the appropriate **cpd adm** command for your environment. The **cpd adm** command without the **--apply** option is a dry-run mode and only returns a list of the changes without applying the changes. This enables your security team an opportunity to inspect the SCCs and SAs before applying them to your OpenShift cluster. These changes include the creation of SCCs and SAs and the configuration of all of the necessary resources. You must make the changes to your cluster to ensure that the Cloud Pak for Data control plane can run on your cluster. To apply the changes to the cluster, you can run the **cpd adm** command with the **--apply** option. For more details see Installing IBM Cloud Pak for Data.

4.0 Authentication and authorization

4.1 Managing users

By default, Cloud Pak for Data provides and out-of-the-box user management feature to manage users. The default out-of-the-box user management feature is for getting started quickly and might not be suitable for meeting all the compliance requirements for user management. It is strongly recommended that you use your enterprise-grade LDAP /Active Directory or other supported mechanisms for better security.

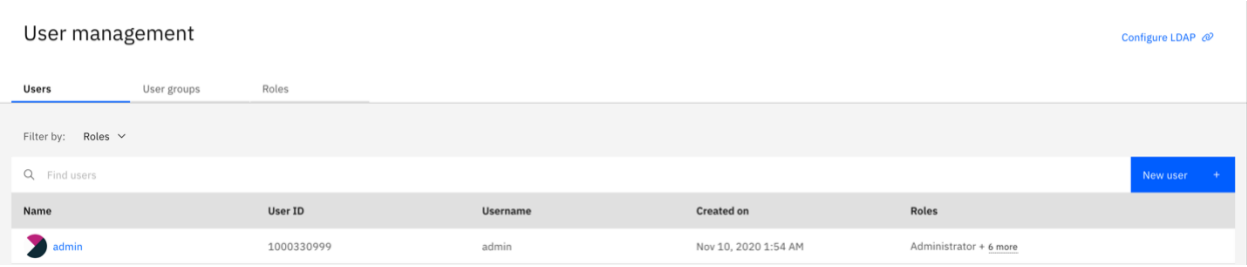


Figure 4.1 Cloud Pak for Data page for managing users

This feature stores the user records in an internal repository database. It stores only the one-way cryptographic hash of the passwords using Secure Hash Algorithm SHA2 and uses unique salts for each user.

The user management features come with a default user (admin). The default admin user is automatically assigned all of the roles, except for Data Engineer. You can edit the admin user to add or remove roles. A Cloud Pak for Data administrator can add new users to Cloud Pak for Data or edit any existing user in this user interface to give users permission to access specific features and optionally give users administrator permissions.

Additionally, for analytics projects, a project administrator can assign Viewer, Editor, or Admin permissions to collaborators.

4.2 Managing user groups

You can create user groups to simplify the process of managing large groups of users with similar access requirements. For example, if you know that multiple users need the same combination of roles, you can add them to a group that is assigned the role combination. If a member of the group leaves the company, you can remove the user from the group.

By default, Cloud Pak for Data includes the **All users** group. As the name suggests, all Cloud Pak for Data users are automatically included in this group. The group is used to give all

platform users access to assets such as the **Connections** catalog. You cannot edit or delete this group.

See Managing user groups for steps to create a new user group.

The user groups feature is not available for some services in the Cloud Pak for Data V3.5 release.

4.3 Configuring LDAP

After you set up Cloud Pak for Data, it is strongly recommended that you use your enterprise-grade LDAP provider for user management. The default user management feature is for getting started quickly and may not be suitable for meeting all the compliance requirements for user management. Cloud Pak for Data provides a user interface for configuring LDAP.

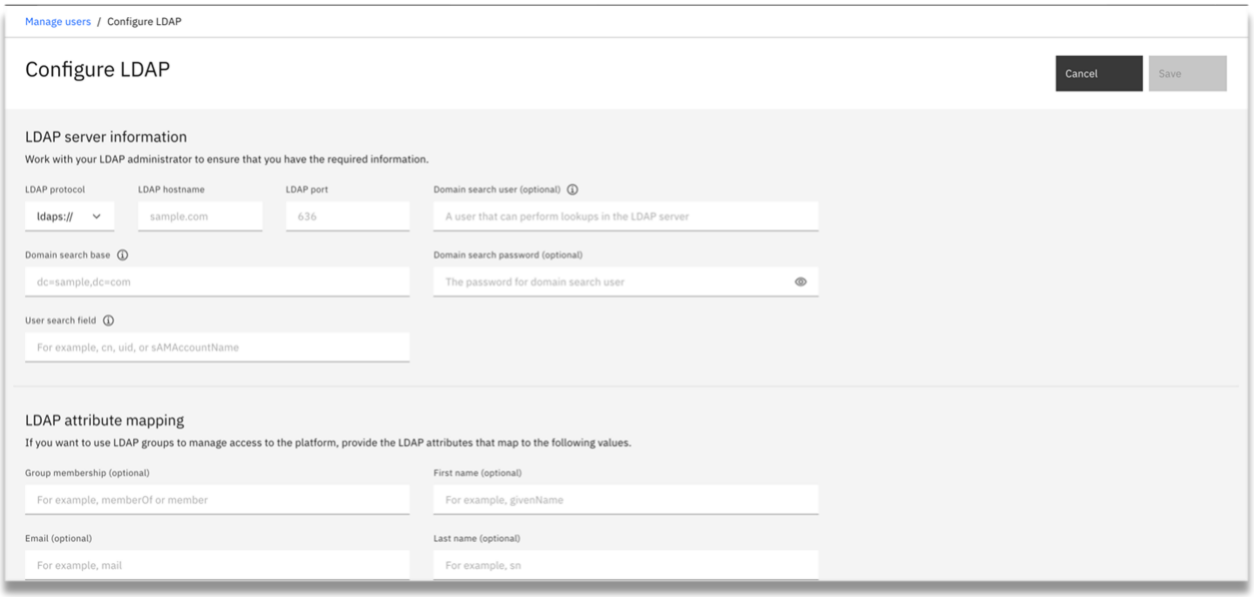


Figure 4.2: Cloud Pak for Data page for configuring LDAP

When LDAP is configured, authentication is delegated to LDAP server for every sign-in to Cloud Pak for Data. Similarly, when SAML Web Single Sign-On (SSO) is setup, sign-in is redirected to SAML instead.

Cloud Pak for Data, as a cache, also maintains a list of users and LDAP groups that were granted access to Cloud Pak for Data (authorized with a specific role). It does this to avoid querying LDAP for every lookup and thereby avoiding any adverse impact to the LDAP server and other

enterprise services. Cloud Pak for Data runs asynchronous jobs which runs at a fixed interval to validate each User's profile and group membership to ensure that this cache is kept up to date with changes done in the LDAP server.

After you grant Cloud Pak for Data administrator privileges to a user in your LDAP server, it is recommended that you disable or remove all users from the internal database repository. See Disabling the default admin user to learn how to disable the default admin user in Cloud Pak for Data.

4.4 Enabling Single Sign-On (SSO) with Security Assertion Markup Language 2.0 (SAML2)

You can also use Security Assertion Markup Language 2.0 (SAML2) for single sign-on (SSO) to the IBM® Cloud Pak for Data web client. If you plan to use SAML2 for single sign-on (SSO), it is strongly recommended that you complete Configuring single sign-on before you add users. If you add users before you configure SSO, you will need to re-add the users with their SAML ID to enable them to use SSO.

Cloud Pak for Data also supports integrating with IBM® Cloud Platform Common Services IAM as a mechanism to work with other Identity providers and to support SSO. This is particularly useful when there are multiple Cloud Paks installed on the same OpenShift cluster. See Integrating with Cloud Platform Common Services for steps to register Cloud Pak for Data with the IBM Cloud Platform Common Services IAM Service.

4.5 Session management

Cloud Pak for Data automatically generates a bearer token when a user signs in, and securely stores information in the user's home directory. This token expires based on the policies defined in the Cloud Pak for Data cluster. When the user signs out, the stored bearer token is cleared.

A Cloud Pak for Data administrator can configure the token expiry time and the web session token refresh period in accordance with your security and compliance requirements. If there is some activity, the web session token refresh period helps to keep the user signed into the Cloud Pak for Data web client by automatically refreshing the token. If a user leaves their session idle in a web browser for the specified length of time as configured in the token expiry time, then the user is automatically logged out of the web client. See Setting the idle session timeout for steps for setting up the token expiry time and the web session token refresh period.

Developers who invoke Cloud Pak for Data APIs from programs that need longer expiry times can use API keys (discussed in the next section) because they don’t expire unless explicitly revoked, instead of the short-term tokens.

On the client, the session information is stored in a secured cookie on the browser. The cookie is https only, not available to JavaScript, and uses SameSite:Lax to avoid cross-domain vulnerabilities. The cookies are cleared from the browser cache when the session is terminated.

4.6 API keys

Cloud Pak for Data v3.5 also supports API keys for authentication. You can automatically authenticate to the Cloud Pak for Data platform or to a specific instance of a service from a script or application. Your API keys are associated with your credentials and are specific to you. Your keys enable you to authenticate without entering your password.

You can generate:

- A platform API key
- An instance API key

Your platform API key enables scripts and applications to access everything that you would typically be able to access when you log in to the Cloud Pak for Data web client. An instance API key enables access only to the specific instance from which is it generated. The instance API key is not available for some services in the Cloud Pak for Data v3.5 release.

See Generating API keys for steps to generate, re-generate or delete a platform API key or an instance API key.

4.7 Managing roles

Cloud Pak for Data comes with predefined roles. The predefined roles that are available depend on the services that are installed on top of Cloud Pak for Data. When you add or approve a user, you must specify the *role* that the user has.

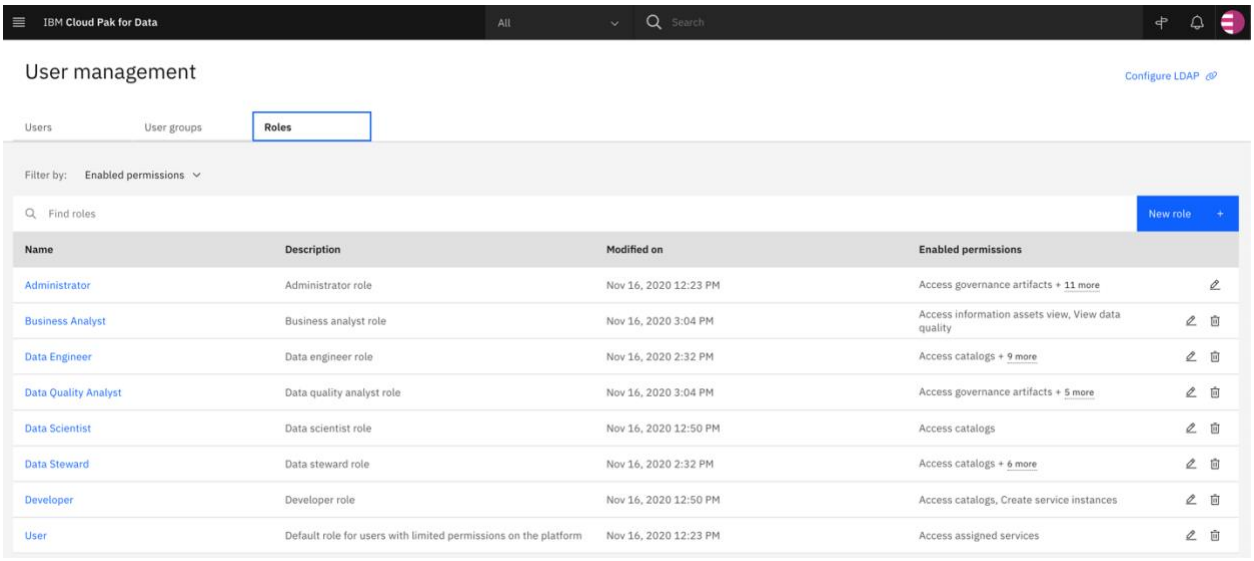


Figure 4.3: Cloud Pak for Data page for managing roles

The following table shows the different predefined roles and their associated permissions in Cloud Pak for Data.

Role (Service that creates the role)	**Permissions**	**Services that contribute permissions**
Administrator (Cloud Pak for Data control plane)	• Administer platform • Create service instances	Cloud Pak for Data control plane
	• Integrate and transform data	DataStage® Edition
	• Analyze data quality • Discover assets • Import metadata • Manage catalogs • Access governance artifacts • Manage governance categories • Manage governance workflows • Manage information assets • Manage data protection rules	Watson™ Knowledge Catalog
Business Analyst (Watson Knowledge Catalog)	• Access catalogs • Access information assets view • View data quality	Watson Knowledge Catalog
Data Engineer (DataStage Edition or Watson Knowledge Catalog)	• Create service instances	Cloud Pak for Data control plane
	• Integrate and transform data	DataStage Edition
	• Access catalogs	Watson Knowledge Catalog

Role (Service that creates the role)	**Permissions**	**Services that contribute permissions**
	• Discover assets • Import metadata • Access governance artifacts • Manage information assets • View data quality	
Data Quality Analyst (Watson Knowledge Catalog)	• Access catalogs • Analyze data quality • Discover assets • Import metadata • Access governance artifacts • Manage information assets	Watson Knowledge Catalog
Data Scientist (Watson Knowledge Catalog)	• Access catalogs	Watson Knowledge Catalog
Data Steward (Watson Knowledge Catalog)	• Access catalogs • Discover assets • Import metadata • Access governance artifacts • Manage information assets • View data quality • Manage data protection rules	Watson Knowledge Catalog
Developer (Watson Knowledge Catalog)	• Create service instances	Cloud Pak for Data control plane
	• Access catalogs	Watson Knowledge Catalog
User (Cloud Pak for Data control plane)	• Access assigned services	Cloud Pak for Data control plane

The default user (admin) is automatically assigned all of the following roles. You can edit this user to remove some of the roles if needed.

- Administrator
- Business Analyst
- Data Engineer
- Data Quality Analyst
- Data Scientist
- Data Steward
- Developer

You can also edit the default roles or create new roles if the default set of permissions doesn't align with your business needs. For more information, see Managing roles.

When you create a new role, you can select from a list of permissions that you want to assign to the role.

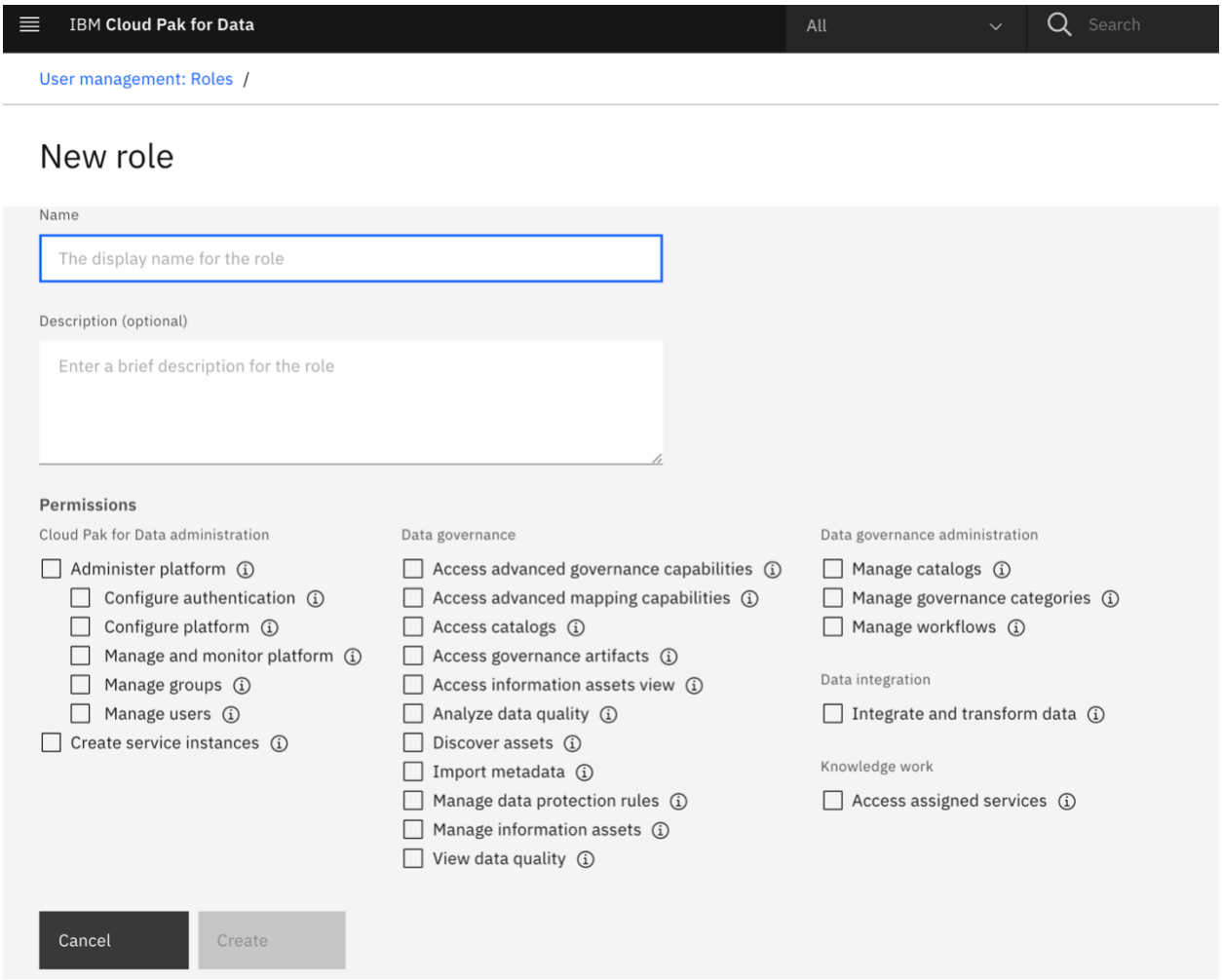


Figure 4.4: Cloud Pak for Data page for creating a new role and assigning permissions

Details about the actions that are associated with each permission are listed Predefined roles and permissions.

5.0 Encryption considerations

Cloud Pak for Data supports protection of data at rest and in motion. It supports FIPS (Federal Information Processing Standard) compliant encryption for all encryption needs.

5.1 Encrypting your storage partition

To ensure that your data in Cloud Pak for Data is stored securely, you can encrypt your storage partition. Cloud Pak for Data supports and is optimized for multiple storage providers on OpenShift:

- IBM Cloud File Storage (ibmc-file-gold-gid storage class)
- Network file system (NFS)
- Portworx
- Red Hat OpenShift Container Storage 4

You can implement encryption at rest for your storage volumes as recommended by the different Storage vendors. See Storage considerations for more details.

If you use Linux Unified Key Setup-on-disk-format (LUKS) to encrypt your storage partition, you must enable LUKS and format the partition with XFS before you install Cloud Pak for Data.

5.2 Using a custom TLS certificate for HTTPS connections

Cloud Pak for Data exposes one HTTPS port as the primary access point for the web client and for API requests. On Red Hat OpenShift, the port is exposed as an OpenShift route. All communications are encrypted using SSL. Only TLS 1.2 highly secure cryptographic ciphers are supported.

The IBM® Cloud Pak for Data installation includes a self-signed TLS certificate that can be used to enable HTTPS connections. By default, this certificate is untrusted by all HTTPS clients. However, you can replace the default certificate with your own TLS certificate.

To use SSL to encrypt communications to and from Cloud Pak for Data, ensure that you use your own SSL certificate and private key (both in PEM format) to enable an HTTPS connection to the Cloud Pak for Data web client. See Using a custom TLS certificate for HTTPS connections.

6.0 Network access requirements

6.1 Required ports

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 Cloud Pak for data cluster.

The primary port is what the Red Hat® OpenShift® router exposes. See Network Access Requirements for details.

When you provision a new service or integration on your Cloud Pak for Data cluster, the services might require connections to be made from outside the cluster. For example, you might require connections when you access databases, or run data virtualization through an ODBC/JDBC connection. If the service or integration requires connections to be made to the cluster, locate the port numbers from each service's Details page and open those network ports.

6.2 Updating the DNS service name

When you install the IBM® Cloud Pak for Data control plane, the installation points to the default Red Hat OpenShift DNS service name. If your OpenShift 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.

7.0 Audit logging

At a high level, audit logging provides accountability, traceability, and regulatory compliance regarding access to and modification of data. Enterprises are often subject to industry requirements for regulatory auditing compliance. Therefore, a holistic auditing solution that works with IBM Cloud Pak for Data requires contributions and coordination of solutions from OpenShift, Guardium, and the IBM Cloud Pak for Data software stack.

Audit logging support within IBM Cloud Pak for Data includes support for generating, collecting, and forwarding CADF (Cloud Auditing Data Federation) compliant audit records for core platform auditable events. You can configure IBM Cloud Pak for Data audit logging to forward audit records to your security information and event management (SIEM) solutions, such as Splunk, LogDNA, or QRadar.

In a multi-tenanted environment, the Audit Logging Service is namespace-scoped and is installed in different namespaces for each instance of IBM Cloud Pak for Data. The audit logging services can be set up to work independently of one another for each tenant. Each tenant's audit records are isolated from other tenant’s records and can be forwarded with different SIEM collector configurations. From a security and compliance perspective, the Audit Logging Service ensures that auditors for each tenant have visibility of records for that tenant only.

Similar to other IBM Cloud Pak for Data core platform services, the Audit Logging Service pod deployment (zen-audit) supports multiple replicas. By default, the zen-audit deployment is configured as a single replica with minimal memory and CPU load. This deployment can be scaled up and out when the need arises.

The audit logging implementation uses Fluentd Output Plugins to forward and export audit records. After you install IBM Cloud Pak for Data, the Audit Logging Service is configured by default to post ingested audit records only to the zen-audit pod stdout log. External SIEM Configuration can be added to *the* zen-audit-config Kubernetes configmap to extend where the Audit Logging Service exports all collected audit records.

For a detailed list of Cloud Pak for Data Audit Events and instructions to configure Cloud Pak for Data to forward audit records to your existing security information and event management (SIEM) solutions, see Exporting IBM Cloud Pak for Data audit records to your security information and event management solution.

8.0 Multitenancy and network security

To make effective use of infrastructure and reduce operational expenses, Cloud Pak for Data can be run in multi-tenanted mode on a single OpenShift cluster, while still maintaining security, compliance, and independent operability.

This is achieved by creating a Red Hat OpenShift Project namespace for each tenant. An instance of Cloud Pak for Data is installed, starting with the control plane, in that namespace which is then assigned to a specific tenant to consume and manage. A selection of individual Cloud Pak for Data services is also installed as needed into that namespace

A tenant is thus offered the benefit of isolation that a dedicated namespace provides. Such as resource quotas, independent configuration of authentication and authorizations mechanisms and audit collections.

The following section describes how best to establish network security for Cloud Pak for Data instances in such shared OpenShift clusters.

8.1 Using the multitenant isolation mode for the OpenShift SDN

During installation of the OpenShift cluster, the openshiftSDNConfig configuration's network isolation mode can be set to 'Multitenant'. See Configuration parameters for the OpenShift SDN network provider.

In this mode, each Project namespace is isolated by default from each other and network traffic is not allowed between two projects.

Cloud Pak for Data also supports the concept of a *tethered project*, a separate namespace from the control plane namespace that can be used to host certain Service instances. However, to provision and manage these service instances, these tethered namespaces would need to be

accessible from the control plane namespace and vice versa for other situations. For details, see Configuring network isolation using OpenShift SDN.

As cluster admin, follow the procedure outlined in this OpenShift documentation reference Configuring network isolation using OpenShift SDN to allow access to and from tethered namespaces.

8.2 Using network policies

If during the installation of the OpenShift cluster, the network isolation mode is set to NetworkPolicy (the default), then traffic between different project namespaces can be defined using custom resources of type NetworkPolicy. See Configuring network policy with OpenShift SDN.

By default, all pods in a project are accessible from other pods and network endpoints. To isolate one or more pods in a project, you can create NetworkPolicy objects in that project to indicate the allowed incoming connections. Project administrators can create and delete NetworkPolicy objects within their own project.

Typically, each tenant would expect isolation inside their dedicated namespace. Therefore, a few network policies need to be created to:

- Deny all traffic.
- Only accept connections from pods within the same namespace.
- Only accept connections from the OpenShift Container Platform
- Only accept ingress traffic at the front door

To make a project deny all traffic by default, define a NetworkPolicy that applies to all pods in that namespace and rejects all incoming traffic into that namespace:

```
kind: NetworkPolicy
apiVersion: networking.k8s.io/v1
metadata:
  name: deny-by-default
spec:
  podSelector:
  ingress: []
```

To make pods accept connections from other pods in the same project, but reject all other connections from pods in other projects, define the following NetworkPolicy:

```
kind: NetworkPolicy
apiVersion: networking.k8s.io/v1
```

```
metadata:
  name: allow-same-namespace
spec:
  podSelector:
  ingress:
  - from:
    - podSelector: {}
```

To make pods in the project accept connections from the OpenShift container platform for example from the OpenShift Ingress Controller and the OpenShift monitoring, define the following Network Policies:

```
kind: NetworkPolicy
apiVersion: networking.k8s.io/v1
metadata:
  name: allow-from-openshift-ingress
spec:
  ingress:
  - from:
    - namespaceSelector:
        matchLabels:
          network.openshift.io/policy-group: ingress
  podSelector: {}
  policyTypes:
  - Ingress
```

```
kind: NetworkPolicy
apiVersion: networking.k8s.io/v1
metadata:
  name: allow-from-openshift-monitoring
spec:
  ingress:
  - from:
    - namespaceSelector:
        matchLabels:
          network.openshift.io/policy-group: monitoring
  podSelector: {}
  policyTypes:
  - Ingress
```

Finally to accept ingress traffic at the front door, define the following NetworkPolicy :

```
kind: NetworkPolicy
apiVersion: networking.k8s.io/v1
metadata:
  name: allow-access-to-front-door
spec:
  podSelector:
    matchLabels:
      component: "ibm-nginx"
  policyTypes:
  - Ingress
  ingress:
  - {}
```

8.3 Enabling access to and from tethered projects

8.3.1 Define a "Tenant Namespace Group"

Since there maybe multiple instances of Cloud Pak for Data, each with its own tethered namespaces, it is important to identify and group each of these namespaces appropriately.

Consider this example:

a) The Cloud Pak for Data instance (control plane) in project namespace ‘dev’ is associated with two tethered namespaces – ‘apps-dev’ and ‘db-dev’.

b) Another instance of Cloud Pak for Data in project namespace 'prod' is associated with two other tethered namespaces - 'apps-prod' and 'db-prod'.

To *group* these two sets of namespaces, apply labels to these project namespaces.
For example:

```
oc label namespace dev apps-dev db-dev cpdgroup=dev
oc label namespace prod apps-prod db-prod cpdgroup=prod
```

You can validate that the label has been properly applied by using the **describe** command.

For example:

```
oc describe namespace db-dev
```

To *remove* a namespace from a group, remove its assigned label. For example:

```
oc label namespace db-dev cpdgroup-
```

Note the trailing "-".

8.3.2 Define a NetworkPolicy to allow access only within the group

For example, for the "dev" group:

```
kind: NetworkPolicy
apiVersion: networking.k8s.io/v1
metadata:
  name: dev-group
spec:
  podSelector:
  ingress:
  - from:
      - namespaceSelector:
          matchLabels:
            cpdgroup: dev
```

This allows for all pods from within the same *group* to connect with each other.

However, you can also make the NetworkPolicy even more restrictive.

For example:

```
kind: NetworkPolicy
apiVersion: networking.k8s.io/v1
metadata:
  name: restrict-app
spec:
  podSelector:
    matchLabels:
      app: my-https
  ingress:
  - from:
      - podSelector:
          matchLabels:
            client: my-authorized
        namespaceSelector:
          matchLabels:
            cpdgroup: dev
```

In this case, pods with the label "**app: my-https**" can only be accessed by client pods that are within the same group and that have the "**client: my-authorized"** label**.**

Network Policies can also be defined to apply cumulatively, which means you can combine multiple NetworkPolicy custom resources together to satisfy complex network requirements.

9.0 Privacy and compliance assessments

Cloud Pak for Data provides various features as described in this document and in Security on Cloud Pak for Data that can be used in preparations for various privacy and compliance assessments. This information is not an exhaustive list, due to the many ways that clients can choose and configure features, and the large variety of ways that Cloud Pak for Data can be used in itself and with third-party applications and systems.

Cloud Pak for Data is data agnostic, and so is not specifically aware of the nature of the data that it is handling other than at a technical level (encoding, data type, size). As such, the product can never be aware of the presence (or lack thereof) of personal data. It is up to customers to track whether personal data is present in the data that is being moved by this product.

Clients are responsible for ensuring their own readiness for the laws and regulations that apply. Clients are solely responsible for obtaining advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulations that may affect the clients, business, and any actions the clients might need to take to comply with such laws and regulations.

The products, services, and other capabilities described herein are not suitable for all client situations and might have restricted availability. IBM does not provide legal, accounting, or auditing advice or represent or warrant that its services or products will ensure that clients are ready for any law or regulation.

Cloud Pak for Data platform has completed the following security and privacy compliance assessments, which may or may not apply to all the add-on services on Cloud Pak for Data:

9.1 FISMA

The Federal Information Security Management Act (FISMA) is a United States federal law passed in 2002 that requires federal agencies to develop, document, and implement an information security and protection program.

Cloud Pak for Data version 3.0 is has completed the assessment for FISMA (Federal Information Security Management Act).

9.2 GDPR

The General Data Protection Regulation (GDPR) is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area. GDPR establishes a stronger data protection regulatory framework for processing of personal data of individuals.

Clients are responsible for ensuring their own readiness for the laws and regulations, including the European Union General Data Protection Regulation. See IBM Cloud Pak for Data considerations for GDPR readiness for more help in preparations for GDPR readiness.

9.3 Accessibility

IBM is committed to accessibility. Accessibility features that follow compliance guidelines are included in the content and documentation to benefit users with disabilities. Parts of the user interface of Cloud Pak for Data are accessible, but not entirely. Product documentation for Cloud Pak for Data uses the latest W3C Standard, WAI-ARIA 1.0 to ensure compliance with the United States Access Board Section 508 Standards, and the Web Content Accessibility Guidelines (WCAG) 2.0.

10.0 Additional security best practices

10.1 Network isolation of the RHOS project where Cloud Pak for Data is deployed

As a best practice, it is recommended that you use network isolation to isolate the Red Hat OpenShift project (Kubernetes namespace) where Cloud Pak for Data is deployed. Then, you must ensure that only the appropriate services are accessible outside the namespace or outside the cluster.

For information about network isolation, review the following OpenShift documentation.

- Understand software defined networking options with Red Hat OpenShift
- Using NetworkPolicy objects for custom isolation policies
- Isolating applications in OpenShift projects

10.2 Setting up an elastic load balancer

To filter out unwanted network traffic, such as protecting against Distributed Denial of Service (DDoS) attacks, use an elastic load balancer that accepts only full HTTP connections. For information, see Protecting Against DDos Attacks.

10.3 Disabling the external registry route

For the registry server, you can disable the external route that is used to push images to the registry server when you are not installing Cloud Pak for Data. However, if you leave the route disabled while you install Cloud Pak for Data, the installation documentation won't be accessible.

10.4 IBM and Red Hat's point of view on installing antivirus software on OpenShift Container Platform cluster nodes

We do not encourage our customers to install antivirus on OpenShift Container Platform cluster nodes because it can cause issues with the functioning of the OpenShift Container Platform and Cloud Pak for Data. Before you upload any files that are obtained from untrusted sources to Cloud Pak for Data platform or services, it is recommended that you scan the files externally for any malware or any malicious content.

For customers who want to install additional security software like antivirus on the cluster, it's recommended to use security software that is designed to run in a containerized environment on top of Kubernetes and OpenShift, and that typically uses Daemon sets. Before procuring any such security software, confirm with the security software vendor that the software works and is certified with the specific OpenShift version in use. Legacy host level installation on the cluster nodes might not work well in a Kubernetes and OpenShift environment.

Sometimes, security software could cause various performance issues such as slowdowns or timeouts. In such cases, you need to disable the security software to determine if the issue is related to the use of the security software.

See Is any virus protection software needed for Red Hat Enterprise Linux for how Red Hat Enterprise Linux provides security without antivirus software. OpenShift Engineering also heavily focuses on security features; you can read more about these at Red Hat OpenShift Container Security and at OpenShift Container Platform – Container Security Guide.

References

- IBM Redbooks: Security in Development - The IBM Secure Engineering Framework
  https://www.redbooks.ibm.com/redpapers/pdfs/redp4641.pdf

- Red Hat OpenShift Container Platform Documentation
  https://docs.openshift.com/container-platform/4.5/welcome/index.html

- IBM Security and Privacy by Design (SPbD@IBM)
  https://www.ibm.com/trust/security-spbd

Notices

This information was developed for products and services offered in the U.S.A. IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785 U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:
Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:
INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore this statement might not apply to you. This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
2Z4A/101
11400 Burnet Road
Austin, TX 78758 U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.
Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice and represent goals and objectives only. All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary. This information is for planning purposes only. The information herein is subject to change before the products described become available. This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE
This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.
Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:
© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. _enter the year or years_. All rights reserved.
If you are viewing this information in softcopy form, the photographs and color illustrations might not be displayed.

Trademarks
IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at Copyright and trademark information at ibm.com/legal/copytrade.shtml.
Adobe, Acrobat, PostScript and all Adobe-based trademarks are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, other countries, or both. IT Infrastructure Library is a registered

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 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.