Introduction to IBM Cloud Paks

IBM Cloud Pak offerings are an integrated set of AI-infused software solutions for hybrid cloud that help you fully implement intelligent workflows in your business to accelerate digital transformation. Cloud Paks pull together a lot of the capabilities that IBM had in the past as on-premises license software, and are designed to support variety of use cases:

* **WebSphere Hybrid Edition** (formerly Cloud Pak for Application) - helps developers to build, deploy and run applications in microservices framework, it includes everything you need to modernize your app portfolio
* **Cloud Pak for Business Automation** - automates business processes and streamline operations, it helps to digitize and automate more types of work at scale
* **Cloud Pak for Data** - accelerates the shift toward artificial intelligence (AI) usage, it simplifies and automates how you collect, organize, analyze, and infuse data across your business
* **Cloud Pak for Integration** - helps integration specialists, it connects systems and applications for better decisions
* **Cloud Pak for Watson AIOps** - designed for proactive multicloud management, it helps to securely automate end-to-end lifecycle management
* **Cloud Pak for Security** - helps you find and respond to threats and risk and enables you to assess hidden threats, make more informed risk-based decisions, and respond to incidents faster

**Cloud Paks Architecture**

The container architecture enables all Cloud Paks to run more efficiently because multiple software elements are isolated from each other and, unlike virtual machines, containers share the operating system kernel with the underlying host so system calls are made directly.

As the answer for the complexity of multiple containers management, Cloud Paks use **Red Hat OpenShift**, which is built on the open-source system - Kubernetes.

**General concepts of containers and Kubernetes will be covered in a later course of this Learning Path.**

The IBM Cloud Paks environment has changed over the past two years and continues to evolve as IBM has been enabling the software to run in containers. In most cases, IBM has rewritten the capabilities to be container-based cloud-native application.

## Deployment options - interactive version

Cloud Pak for Data is a true cloud agnostic platform that can be deployed anywhere depending on client’s business needs and requirements. The various deployment options of Cloud Pak for Data provide a great flexibility to the clients who can decide how they want to manage their infrastructure, starting from on-premises environments that are fully owned and managed by the client, ending at Software as a Service option where Cloud Pak for Data environment is in IBM charge.

The graphic below illustrates several different options to deploy and manage Cloud Pak for Data environment.

**IBM Cloud Pak for Data System**

If you do not have an existing on-premises cluster and you want to get up and running quickly, consider IBM Cloud Pak for Data System. IBM Cloud Pak for Data System is an all-in-one cloud-native data and AI platform in a box that provides a preconfigured, governed, and secure environment to collect, organize, and analyze data.

* Red Hat OpenShift: Red Hat OpenShift is already deployed on your IBM Cloud Pak for Data System.
* Cluster architecture: The system lets you flexibly expand or reduce storage and compute with plug-and-play nodes.
* Services: After you install Cloud Pak for Data, you can decide which services you want to install.

**On-Premises**

If you want to ensure that your environment is running securely behind your firewall, or you have an existing on-premises Red Hat OpenShift Container Platform cluster, you can deploy Cloud Pak for Data in your own private cloud.

* Red Hat OpenShift: You must deploy an instance of Red Hat OpenShift Container Platform on your cluster.
* Cluster architecture: Cloud Pak for Data is deployed on a multi-node cluster.
* Services: After you install Cloud Pak for Data, you can decide which services you want to install.
* **Infrastructure as a Service**
* If you already use any of the supported cloud solutions: IBM Cloud, Amazon Web Services (AWS), Microsoft Azure or Google Cloud, and you do not want to set up and manage your own hardware, you can deploy Cloud Pak for Data in that particular infrastructure.

**Managed OpenShift**

Cloud Pak for Data includes entitlement to the Red Hat OpenShift Container Platform. You can deploy Cloud Pak for Data on the Red Hat OpenShift environment that is managed by IBM which is also a recommended solution or use self-managed OpenShift.

* Cluster architecture: At a minimum, your cluster must contain 3 master nodes and 3 worker nodes. You can deploy additional worker nodes to support your workload.
* Services: After you install Cloud Pak for Data, you can decide which services you want to install.

**Software as a Service**

* IBM Cloud Pak for Data as a Service is a “pay-as-you go” subscription model for a starter set of IBM Cloud Pak for Data. It might be right for you if you already use IBM Cloud to run business-critical applications and you do not want to set up and manage your own deployment of Cloud Pak for Data.

Cloud Pak for Data as a cloud-native solution with a single user interface (UI)

* Cloud Pak for Data is data and AI **cloud-native** software which provides a **single platform** solution where different services can be deployed depending on your business needs.
* Find out what the practical reason is for mentioning these technical features at the beginning of this course.
* Cloud Pak for Data helps to modernize the way of working with the data to operationalize AI for faster business insights and increased **return on investment** (ROI). The platform’s infrastructure is designed to do two things.

**Consolidate software:**

* From The Cloud Pak for Data platform you can access services that span the **entire analytics lifecycle**. It means that within the tool you can: collect, analyze and organize data, and infuse the Machine Learning models into applications. IBM Cloud Pak for Data gives the power to “lift and shift” data and AI workloads to its platform, and eliminate the expensive need to integrate competitive solutions.

**Container Platform**

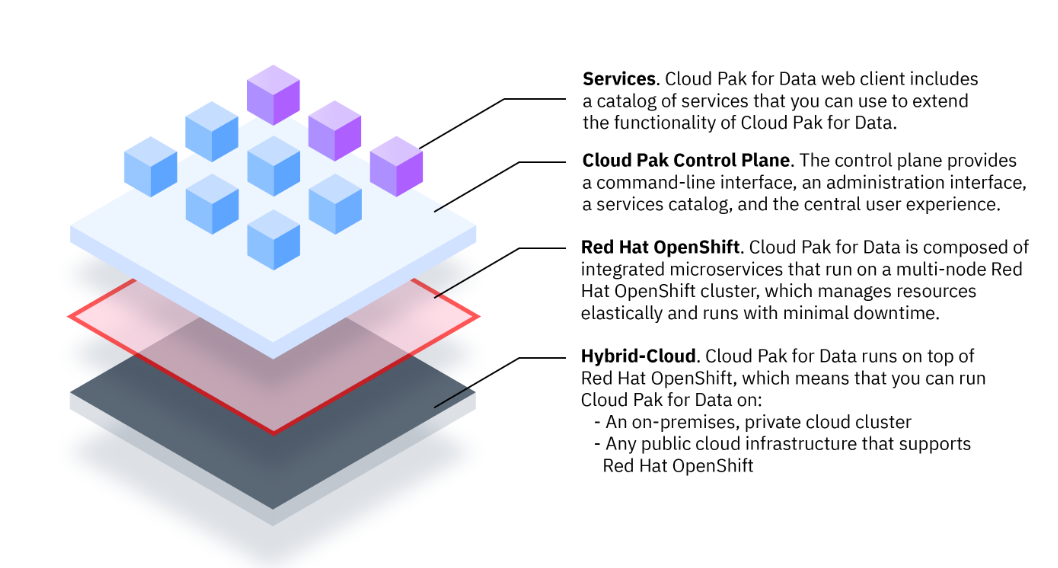
* For your information architecture, Cloud Pak for Data brings Red Hat OpenShift’s container management and orchestration which gives flexibility to run on top of **any private, public cloud, or hybrid cloud** infrastructure based on a microservices architecture. That helps reduce administration expenses and simplifies provisioning.

## Core technical features of Cloud Pak for Data

Let's examine the illustration which appeared on the video with more details. You can also skip the picture and read the text below it. It gives a general overview of the Cloud Pak for Data architecture. Pay attention to two features – unified platform where all services lie, and hybrid-cloud layer which is connected with the platform via Red Hat OpenShift.

The top layer of Cloud Pak for Data consists of services. The catalog of services like Watson Knowledge Catalog or Data Refinery can be found in Cloud Pak for Data web client. Services lie on the unified platform called Control Plane. Thanks to the control plane user can access administration interface, previously mentioned services catalog, a command-line interface and UI in general. Thanks to Red Hat OpenShift Cloud Pak for Data can run on:

* an on-premises, private cloud cluster
* any public cloud infrastructure that supports Red Hat OpenShift

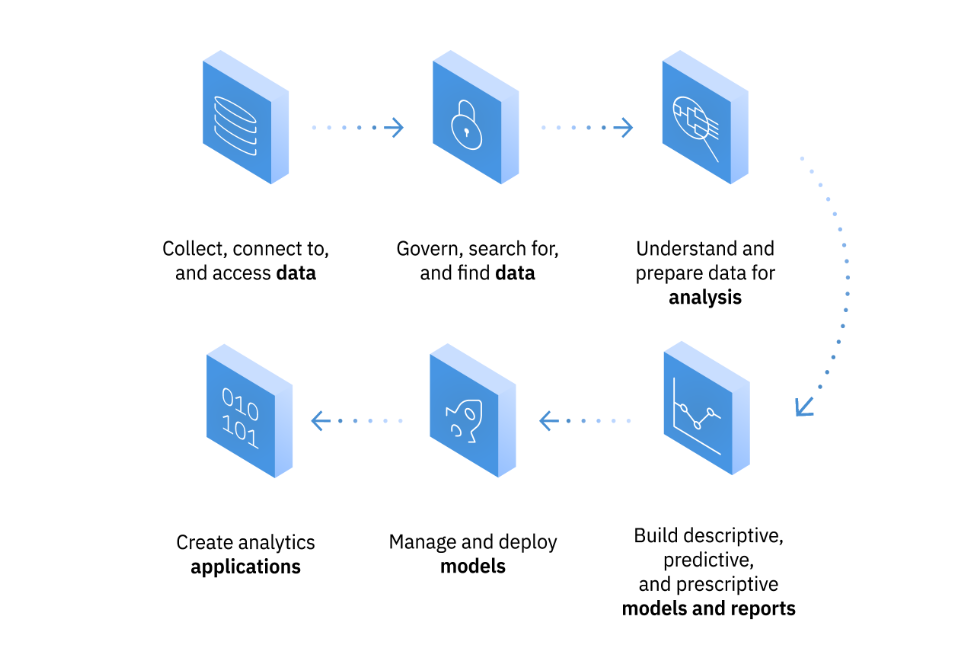


Core functional features of Cloud Pak for Data

To be successful, your enterprise must leverage the power of **artificial intelligence** (AI). With Cloud Pak for Data you can scale AI throughout your organization and deliver solutions to the business questions faster.

IBM defines AI as "**the application of machine learning to build systems that simulate human thought processes**". Cloud Pak for Data provides a suite of services that support you in your journey to AI.

Let us take a look at the common steps in a typical data and analytics workflow:



Those can be summarized in a simpler way: **Collect**, **Organize**, **Analyze**, and **Infuse**. All these steps can be done in Cloud Pak for Data and are **functional features** of Cloud Pak for Data.

**Collect**

First things first. You want to make it easy to consume and access your source data. Cloud Pak for Data helps you connect to your data, no matter where it lives. Cloud Pak for Data includes a**Platform Connections**page that lists connections that can be used by multiple services. Some Cloud Pak for Data services support additional data sources that you can connect to from the service.

With **Watson Query**, you can query data across many systems without having to copy and replicate data. It also can simplify your analytics and make them more up to date and accurate because you are querying the latest data at its source.

**Organize**

After you have collected your data, you will want to create a trusted analytics foundation on the data, with business meaning. The **Watson Knowledge Catalog** service helps you organize your data through data classification and governance. With this service, you can develop an information architecture that is on-point and ready to keep up with the scale of your data.

**Analyze**

Once you have organized your data, you are ready to scale business insight with artificial intelligence everywhere. Cloud Pak for Data includes numerous **analytics services** that can help you generate scalable insight on demand.

For example, with Cloud Pak for Data you can use:

* **Analytics Dashboards** to create stunning dashboards to quickly visualize data
* **Streams** to build solutions that drive real-time decisions by combining streaming and stored data with analytics
* **SPSS Modeler** to create flows to prepare and blend data, build and manage models, and visualize the results
* **Notebooks** to run your own code that process data and immediately show the results of your computation

**Infuse**

Eventually you will be able to operationalize artificial intelligence with trust and transparency throughout your enterprise.

With Cloud Pak for Data, you can make AI a part of your standard operating procedure. For example, you can build smarter applications with premium **Watson services** and deploy machine learning models into production at scale with **Watson Machine Learning**. You can also deploy trusted AI across your business with **Watson OpenScale.**

A data science project begins with business goals being set, typically by the business executives and data science leaders within the organization. From there, the team works to understand the use case that have been prioritized, and follows a development and prototyping process that ultimately leads to the deployment of models into production.

Here is an example of the process flow that an organization may follow to create and manage a data science project.

The company which exists on the market for 3 years provides an innovative product to the customers: a subscription for chocolate bars with different delivery options: chocolate is delivered every two weeks, one month or every two months.

At some point they observe increased amount of clients cancelling their subscription in winter, when the company expects a higher sales rate. They decide to investigate.

The process starts with formulating a business problem: why the chocolate bars subscription sales rate dropped this winter. The team trying to understand the business problem asks the question to the business analyst “How do we retain the customer?”

Busines analyst will look for the data which potentially could answer this problem. He directs his way to the data steward to check the current cancellation rate. Data steward in accordance to private regulations, curates data already available in the Watson Knowledge Catalog ensuring it does not include any personal information.

Business analyst confirmed that available information do not solve the issue. It appears that there are no potential candidates for this particular business problem, and cured information are irrelevant in this case so for a deeper understanding he asks data scientist for more insights. Data scientist creates a project in Cloud Pak for Data, uses historical data, and thanks to other services available on a platform builds a machine learning model. As a conclusion: the customers that are with the company for more than 2 years and order mostly a chocolate with a mint flavor have a 75% probability to cancel the subscription in winter. Finally, to apply the model to current customers, a developer writes an application which offers mint chocolate to customers with subscriptions longer than 2 years, with a discount.

**Roles and permissions:**

As the typical data science workflow requires variety of competences, there are different personas involved in the process, where each individual has different relationship with data. For this, Cloud Pak for Data includes predefined roles that are available depend on the services that are installed on the platform. Find out what are predefined roles and how each of them fits into the project’s landscape.

You and other people in your company login to the Cloud Pak for Data as a**user.** A Cloud Pak for Data user must be associated with **one or more** Cloud Pak for Data **roles.**

There are already existing, **predefined user roles** like data steward or data engineer. In case if needed - these can be customized.

**Role** consists of one or more permissions. A permission is one or more actions or tasks that can be performed on the cluster. An example of permission is*Administer platform,* users with this permission can manage access to the web client, configure connection to an LDAP server etc.

You can edit the **predefined user roles**to add or remove permissions/create new roles with the available permissions.

## Example of user collaboration:

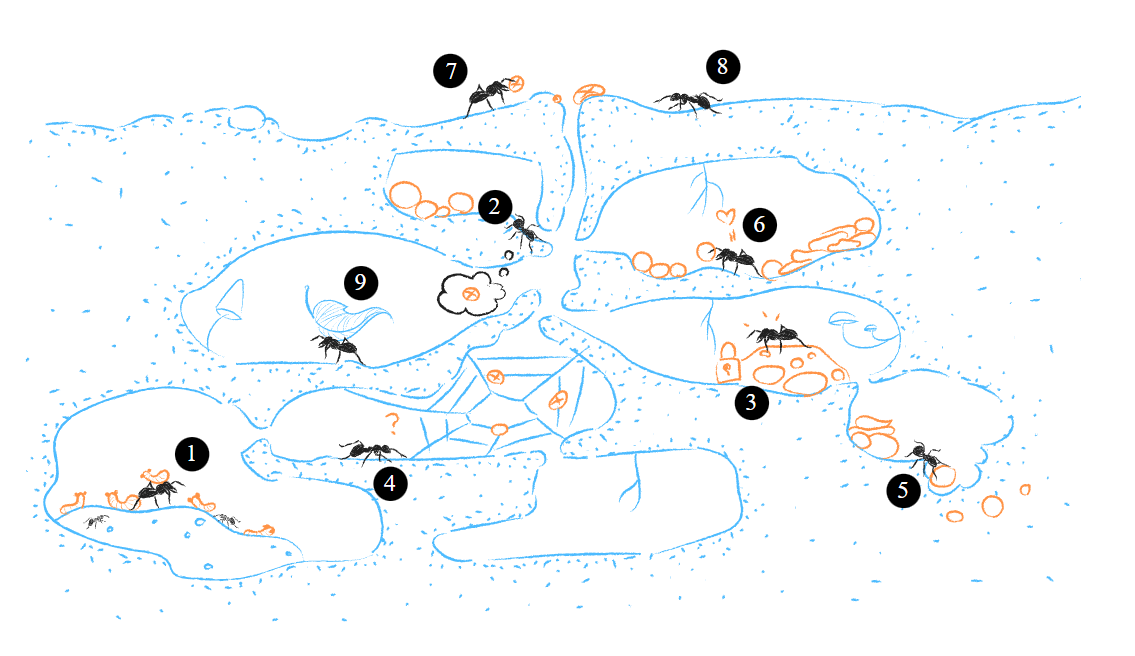
The diagram below demonstrates collaboration between data scientist, data steward and data engineer. If you recall the company which delivers chocolate in a subscription model, the business analyst had to reach out to the data scientist who used historical data in order to come up with more insights. During this process, the data scientist may work with two other personas: the data steward and the data engineer. The first one governs data and ensures regulatory compliance, then approves the data for use. The second one architects data pipelines and ensures operability to deliver data which is missing.

In the first step, the data scientist requests a data set. In the second step, the request has to be approved by the data steward . After getting the approval, the data engineer can fulfill data request in the third step. The data scientist builds and publishes the model as the fourth step. In the final, fifth step, the data steward approves the model which is deployed and captured in the enterprise catalog.



**Ant's tale - text version:**

1. **Administrator.** In a typical nest there are nurse-ants which take care of injured fellows as well as ant's larvae. The nurse ant as Administrator assigns us a role of Business Analyst.
2. **Business Analyst**. We got the first question from the ant's-business - "What are the best chunks of dirt to build our nest?" (which in our case is the metaphor of "bring us the data which will solve our business problem").
3. **Data Steward**. We address this question to the Data Steward who protects all chunks already available for a build and in case we do not find any satisfying piece we can direct the issue to the Data Scientist.
4. **Data Scientist.** Data Scientist Ant has the ability to untangle relations between interesting pieces and by building Machine Learning models find out what solves our issue.
5. **Data Engineer.** For this the Ant can ask Data Engineer to dig and extract more balls of dirt from the ground.
6. **Data Quality Analyst**. In the process Data Quality Analyst can be engaged to avoid using chunks of bad quality, on the picture you can see that the Data Quality Ant picks only perfectly round pieces of dirt.
7. **Developer.** Finally proper pieces are identified and used by Developer to build the anthill.
8. **User.** Apart from specific predefined user roles it is possible to assign a regular user role. The User Ant can access assigned services.
9. **Leaf**. Main organ of photosynthesis and transpiration.



**Business scenarios including services:**

You can treat Cloud Pak for Data like a canvas where different services can be placed and connected in order solve particular business problem. Find out what are exemplary scenarios that can be fulfilled within the platform and which services are needed for each case.

Previously in this course you found out what the typical steps are taken in a data science project. Now you will get a view on examples of how you can answer common business questions using Cloud Pak for Data and which services and extensions are needed in each case.

Example 1: AI for customer care

For this business case you will need Watson Assistant and optionally Watson Discovery. You can automate customer care to increase customer satisfaction while decreasing time to resolution and call volume. With Watson Assistant, you can provide AI-powered, automated assistance to your customers or employees through web or mobile applications and over the phone. You can also enable your customer service agents to handle customer inquiries more efficiently and effectively.

You can supplement your solution with Watson Discovery to unlock insights from complex business content, such as manuals and contracts.

Example 2: Data Ops

You can enable your workforce to access critical data without exposing sensitive data. With Watson Knowledge Catalog you are able to create and enforce policies to govern your business data and understand the quality and lineage of your data. Data Refinery helps you to prepare and refine your data.

**Learn about the subsidiary that implements Cloud Pak for Data:**

Amsel Corporation owns also a subsidiary company called Amsel Real Estate Corp, which is a credit lender under pressure since the ROI on loans is flattening because a high percentage of clients are defaulting on paying back their loan. Amsel Real Estate Corp needs to better identify clients with a high credit risk so that the ROI on loans can be improved.

The company relies on its data science team for credit risk scoring. The team’s lead discusses with leadership and management on how the credit risk model can be improved. Finally, they come to the following conclusions:

They must have an ability to use more data sources at the same time and the data needs to be provided in a timely manner so that the data scientists do not have to wait and lose many hours of productivity.

Using few data sources might involve transferring huge amounts of data from the source to the destination over the network, and thus slowing down network traffic at the expense of production.

  Additional data sources might include Personal Identifiable Information (PII) data which needs to be hidden from the team. Therefore, the new data sources need to be governed so that they align with the company's data protection policies.

The data science team needs to improve the credit risk model by using recently developed modeling techniques that met the legal and business requirements, e.g. fairness and explainability, which will generate trust in the solution.

After many months of exploring and investigating, in cooperation with the management team and taking into consideration the statements listed above, the data science team decides to implement a new product, IBM Cloud Pak for Data. The new software is going to help them to unlock the value of their data and create an information architecture for AI. Cloud Pak for Data is used to collect, organize, analyze, and infuse AI into the data with a scalable Red hat OpenShift Platform.

In all three learning paths designed for Cloud Pak for Data administrator you will be introduced to Emily, who has been working for Amsel Real Estate Corp for over three years as one of the system administrators. Emily has a great experience in installing and setting up Linux systems and servers. She is interested in new technologies, e.g. cloud computing and containerization.

The most of the company’s infrastructure is already containerized so the system administrators have already a good technical background for setting up, managing, and monitoring the Cloud Pak for Data platform. However, switching to Cloud Pak for Data is still a big technology step for Amsel Real Estate Corp and a great challenge for the whole IT department, including Emily, who is going to be the main Cloud Pak for Data administrator.

Because Cloud Pak for Data is based on Red Hat OpenShift Container platform, Emily as a Cloud Pak for Data administrator is going to work very closely with Maria, the Red Hat OpenShift administrator, who primarily focuses on administering and supporting the whole Red Hat OpenShift Container Platform ecosystem for the organization. Maria will also be the one who transfers to Emily all the knowledge that she needs to have about installing, managing, and monitoring a cluster on Red hat OpenShift Container Platform. Next Emily will install the Cloud Pak for Data on the Red Hat OpenShift cluster and with Maria’s help she will learn how she can manage and monitor the cluster.

Another person that Emily is going to work with very closely is Anne, the Cloud Pak for Data solution architect. Anne is an expert when it is about designing the platform architecture, planning, and designing cloud solutions, monitoring, and operationalizing AI. She will recommend industry best practices and configuration settings, that Emily will use when building her Cloud Pak for Data environment.