Brought to you by:

IBM

Best Practices in Multicloud Data & AI

for dummies®
A Wiley Brand

Uncover the power of IBM Cloud Pak for Data

Gain value from data wherever it resides

Manage, govern, and secure your data

Judith Hurwitz

Daniel Kirsch

IBM Limited Edition

Best Practices in Multicloud Data & AI

IBM Limited Edition

by Judith Hurwitz and Daniel Kirsch

for dummies®
A Wiley Brand

Best Practices in Multicloud Data & AI For Dummies®, IBM Limited Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

ISBN: 978-1-119-59345-4 (pbk); ISBN: 978-1-119-59362-1 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Contributions by IBM Data & AI

Scott Hebner: Vice President, Marketing

Ijeoma Pelecanos: Program Director, Marketing

Sriram Srinivasan: Distinguished Engineer, Architect

Hemanth Manda: Director, Platform Offerings

Ron Reuben: Program Director, Offering Management

Mukta Singh: Program Director, Offering Management

Brandon MacKenzie: Offering Manager, IBM Cloud Pak for Data

Stacey Redden: Analyst Relations, Hybrid Data Management

Table of Contents

Introduction

Without a cohesive data strategy, your organization will be at a disadvantage in dramatically changing markets. To be successful, you need a data strategy that enables you to access all your core data wherever it resides. This approach requires your organization to have a framework that cleanses and properly governs data. This type of consistency enables business units to bring together their data across silos in order to manage collaboration. This collaborative approach solves many of the most stubborn problems for businesses trying to gain insights and a holistic view of customer data.

What is the relationship between your products and services and key changes in your market? Are there opportunities to sell existing products to a new set of customers in adjacent markets? Can you protect your data so that you both gain the insights you need while protecting the privacy of customers? These are critical questions that you need to answer in order for your business to be protected against the inevitability of change.

The need to understand data isn't new. Businesses are increasingly turning to the hybrid cloud as a pragmatic way to manage all its data. The challenge has always been how you approach the problem. It is unsafe and impractical to think that you can simply move all data to the cloud. In some cases, you want to move the algorithm to where the data resides through data virtualization. You also need to understand the lineage of your data and the rules that dictate appropriate use based on governance and compliance regulations.

About This Book

Best Practices in Multicloud Data & AI For Dummies, IBM Limited Edition, gives you insights into what it means to manage your data in a hybrid cloud environment. This book discusses the issues related to what it means to a business to manage data in a collaborative manner across the business. Your data is only as good as how you can create a data fabric that will help you to disrupt your competition and become an agile business.

In this book, you discover the techniques needed to manage your data across a variety of public and private clouds. You also gain an understanding IBM Cloud Pak for Data.

Foolish Assumptions

The information in this book is useful to many people, but we have to admit that we did make a few assumptions about who we think you are:

- You're already familiar with the power of data and are looking for ways to transform your data into a strategic weapon.
- You're planning a long-term data strategy that will help you manage your data in a safe and predictable way.
- You understand the huge potential value of the data that exists throughout your organization.
- You're a business leader who wants to apply the most important emerging technologies to be as creative and innovative as possible.

Icons Used in This Book

The following icons are used to point out important information throughout the book:

TIP

Tips help identify information that needs special attention.

WARNING

These icons point out content that you should pay attention to. We highlight common pitfalls in taking advantage of machine learning models and algorithms.

REMEMBER

This icon highlights important information that you should remember.

IN THIS CHAPTER

- Preparing to address organizational goals
- Pulling the elements of your data strategy together
- Understanding IBM Cloud Pak for Data
- Preparing for action

Chapter 1 Putting Data in Context with Organizational Goals

We are in the midst of an era where organizations have more data than ever before. Many applications capture information about customers, their purchasing patterns and preferences, and their payment history. At the same time, massive databases focus on specific vertical business applications — for example retail, banking, insurance, and transportation. Combine these data sources with third-party data and packaged on premises and SaaS applications, and now you have a massive data management issue.

REMEMBER

It is not enough to simply manage data. You need to have an overall data strategy that ensures that information can be used to create a sustainable data-driven organization. This chapter explains the requirements to provide business management with an approach to transform and manage disparate data environments. These requirements are explained in context with the IBM Cloud Pak for Data platform.

Understanding Organizational Goals

Business units are often unwilling to share their data with other divisions within their company. These business teams must make sure that the integrity, security, and accuracy of their data is ensured — in the end it is their responsibility. On the other hand, the CEO tasks the CIO, CFO, and the ever increasingly important Chief Data Officer (CDO) with bringing all this data together to be able to understand the business from a holistic perspective. There is now a need and directive to break down the data silos between businesses to gain insights across all of their divisions and organizations. Business leaders want near real-time insight into their Key Performance Indicators (KPIs) — quarterly or monthly "snapshots" — are no longer good enough.

While it may seem straightforward to simply pull data together from different business units, the task is complicated. Data isn't static. Rather, data is fluid and dynamic, ever changing in terms of volume, velocity, and variety. There are both cultural and technical issues that have to be addressed to make real change happen.

Pulling Together the Organizational Threads

Smart CEOs are beginning to demand change. Increasingly, they insist that business units collaborate based on the overall business goals for growth, understanding, and trust. It is not an easy task to change orientation of your company toward a holistic view of corporate data. This transition requires a major cultural change in the way the business understands the value of its data. It is especially complex because many corporate divisions manage complex packaged applications that include specialized data relating to the products or services they offer to customers. Even when business units collaborate they find that the underlying data is constantly changing so that they don't always come to the right conclusions. Is there a single version of the truth? Can we trust the data and the analytics applied to it?

Creating a Culture of Data Sharing

To change culture requires providing each business with a streamlined and self-service approach to sharing data across business silos. Once there is a mandate from management to kick start a data sharing initiative, there needs to be a process for making this happen. There are four imperatives needed to transform the culture.

Focus on business objectives

What do you do once you assemble a cross organizational team? What is the approach that works? One of the biggest mistakes that management makes is to focus primarily on the data or on analytic algorithms rather than on business challenges. The real focus must be on the business objective and the key business challenges. What do you want to learn from your data? Can you find ways to understand how customer expectations are changing? How do you use data to improve decision-making or to help professionals gain meaningful insights?

Create a cross-functional task force

Collaboration needs to begin by creating a task force consisting of representatives of all key business units. Many leading companies have created the role of the CDO to spearhead these task forces and drive the sharing of data. This task force works with management to understand the corporate objectives and should consist of employees who understand the context of data, the governance of data and the business goals of the company. In addition to business leaders and the CDO, include data administrators as part of the team. Finally, in many cases, it’s important to have somebody on the team who understands the required security, privacy, compliance, and regulatory requirements for the specific data that is being analyzed.

Implement a center of excellence based on best practices

While the task force has the benefit of learning to collaborate across business units, the rest of the company typically remains stuck in business as usual. The goal of the task force is to take the newly learned collaboration approach and set the tone across business units. The task force evolves into a center of excellence to instill best practices in the company. Teams from across the business should be trained on these best practices in order to help create a culture of better collaboration. Task force members work across departments to help the company move forward to break down data silos to achieve significant business results.

Identify a key business goal and apply best data practices

Now that your organization understands what you are doing, you need to take incremental steps to achieve your business goals. You need to break your biggest business challenges into manageable and achievable goals and actions.

Your first steps should ensure that you are able to collect all relevant information and have the right level of data governance and data quality. Before you get started, make sure you can answer these questions:

- How many data sources do we have?
- How many of the data sources reside within our own environment versus third-party services?
- Where is our data physically located? Is the data spread across various cloud environments?
- Do we have a variety of data analytics tools that can’t work together easily?
- What are the regulatory requirements for data placement and privacy?
- If we make more data available via self-service, do we have the skills across teams to effectively use the data?

Getting Started on The Journey

To begin achieving the best practices goals, many people may start by integrating data elements into a data lake. This approach can be effective if your organization is heavily using Hadoop as the data management approach. However, in many situations because a data lake ingests raw data, it runs the risk of violating governance and security policies. Many data lake projects have failed to provide value because of governance problems, the inability to have consistent taxonomy, data becoming stale, and corporate collaboration challenges.

Instead of creating another data lake, a more pragmatic approach is an end-to-end data management and analytics platform. Such a platform based on industry standard containers and container management provides maximum flexibility to manage change. Once such a platform is in place, combined with a set of standard Application Programming Interfaces (APIs), it is possible to more consistently and predictably manage and use data in a hybrid and multicloud environment. Key services such as data quality management, security, governance, and connectivity can ensure accurate understanding of highly siloed data. To be successful, you need to have a consistent and predictable way to deal with data — both as it is today and as it is changing in the future. A consistent platform, based on machine learning and artificial intelligence (AI), gives you the freedom to move forward without constraints. This is the design basis of IBM Cloud Pak for Data.

Introducing IBM Cloud Pak for Data

Successful data strategies require establishing a connective tissue across business units so these organizations can collaborate effectively. IBM's approach to creating an end-to-end data and AI platform is called IBM Cloud Pak for Data. IBM Cloud Pak for Data is a pre-integrated Kubernetes-based multicloud platform for data and AI. The platform includes all the components needed to connect, ingest, discover, govern, and analyze data within unified, collaborative workflows. Through a technology called data virtualization, IBM Cloud Pak for Data allows customers to collect, organize, and analyze data wherever it resides — databases, data warehouses, catalogs, cloud stores. Data virtualization eliminates the common task of needing to move data to a centralized location before analytics can be performed (we discuss data virtualization in more detail in Chapter 2). At the same time, the platform gives the business the ability to gain insights to that data, leveraging machine learning models and advanced analytics.

One of the most important characteristics of the platform is the intelligent data catalog. The catalog uses machine learning to discover and understand meta data and automatically map it to business terms. All of the data and analytics services available in

IBM Cloud Pak for Data are containerized. The core design point of this offering is unification, collaboration, automation, and governance. IBM Cloud Pak for Data unifies data and AI services across public and private cloud infrastructures. It also provides a consistent way to enable data engineers, data stewards, data scientists, analytics professionals, and business users to collaborate through workflow so they can unlock the value of their data. Furthermore, automation and governance is infused throughout, enabling self-service analytics.

IBM Cloud Pak for Data has a variety of deployment choices. The platform is designed to run on any cloud and in multicloud environments. IBM Cloud Pak for Data can be used in dedicated, public and private cloud infrastructures, as well as in an on premises data center. For example, IBM Cloud Pak for Data can run on top of IBM Cloud and with IBM's Kubernetes containers. In addition, IBM Cloud Pak for Data can be deployed on Amazon Web Services, Microsoft's Azure cloud services, and Google Cloud Platform, via Red Hat OpenShift Container Platform. IBM has also developed a hyperconverged system that provides pre-integrated optimized hardware for the IBM Cloud Pak for Data platform. This system can be deployed within hours.

Preparing for Action

Successful businesses begin their movement to a data platform by doing upfront preparation. When the entire team ranging from IT staff and technical leaders to business teams understands the business objectives, everyone is better prepared to make strategic decisions. At this stage, the business is ready to take action to set the plan to turn silos of data and applications into a platform for success. There must be a strategic plan that considers everything from the nature of the data, to the accuracy and timeliness of that data. The plan has to be in sync with the security and governance requirements of the business. Even more importantly the approach has to anticipate change. For example, you'll discover new sources of valuable data. Likewise, governance requirements will change as the business moves into new geographies. New competitors will arise, and customer expectations will change. Clearly, data is only a strategic weapon if you have the tools and approaches to put that data to business action.

IN THIS CHAPTER

- **Seeing the value of a Kubernetes-based approach**
- **Examining the IBM Cloud Pak for Data architecture**
- **Preparing for action**

Chapter 2 The Foundational Model for IBM Cloud Pak for Data

Major changes in technology don't happen in a vacuum. Creating a revolution requires innovation and evolution. Data platforms such as IBM Cloud Pak for Data are only possible with the following three major transitions: the maturation of machine learning techniques, the widespread adoption and advancement of open source, and the evolution of commercial cloud-native architecture providing a common platform across hybrid and multicloud. The combination of machine learning models and open source hybrid cloud has led to the development of a flexible open framework.

IBM Cloud Pak for Data's architecture is designed for scalability, extensibility, and hybrid cloud deployments. How can you enable the next generation of advanced analytics and intelligent applications without replacing your existing data platforms? Being successful requires a focus on creating microservices that increase modularity and the ability to leverage services across applications and clouds. In this chapter, you explore the architectural model for IBM Cloud Pak for Data.

The Value of the Kubernetes Foundation

IBM Cloud Pak for Data leverages Kubernetes as the foundation of its architecture. By taking this containerized approach, customers have the flexibility to choose from a variety of deployment options.

What is the value of the Kubernetes foundation for managing data silos? In brief, Kubernetes provides an underpinning to manage your cloud environment including your highly distributed siloed data and analytics workloads. In the next section we will provide an explanation of the containerized services that sit on top of the Kubernetes foundation.

Because Kubernetes is the orchestration layer for containers, it is possible to integrate a variety of underlying platform services that become an integral part of creating a platform that supports and manages a highly distributed data and analytics environment.

Understanding the Foundation of IBM Cloud Pak for Data

The IBM Cloud Pak for Data platform is designed pre-integrated with unified workflows and a consistent user experience so all the elements work together to support a company's data and analytics goals. In many ways, this pre-integrated approach is different than the way many companies have approached data and analytics. Companies have traditionally selected different tools, ranging from data management and governance, to data science and visualization offerings. At this point, it was the responsibility of the IT organization to enable these components to work together. With the emergence of machine learning and artificial intelligence (AI), the challenges of integration become even more complex. The traditional methods of integration can't keep pace with technology innovation and change. Development teams and data scientists now need easy development and to manage entire pipelines, with trust and transparency.

To address customers' end-to-end data, analytics, and AI requirements, the IBM Cloud Pak for Data platform has a number of key components. Each component is intended to support a variety of pre-integrated microservices-based technologies that have been designed to work together. These services are

- **Data collection:** Services to make data collection simple and accessible, regardless of where data lives
- **Data organization:** Services to create a trusted, business-ready analytics foundation and make data ready for AI
- **Data analysis:** Services for scaling insights on demand while adopting AI everywhere with trust and transparency

These services work together within an intuitive user experience, collaborative workflows, and a unified set of underlying cloud-native data and AI services. Collectively, they help automate how organizations turn siloed data into insights, in a manner that can be easily customized to unique data landscapes, on any cloud (public or private).

From an architectural perspective, eight core components make up IBM Cloud Pak for Data. These components are

- Ingestion and integration
- Analytical data management and storage
- Data access
- Discovery and exploration
- Actionable insight
- Analytics in-motion
- Information management and governance
- Security and compliance

Figure 2-1 shows the reference architecture for IBM Cloud Pak for Data. You see that information management and governance as well as security stretch across the entire platform. In addition, customers bring their own data from a variety of sources, such as machine and sensor data, images and video, social data, weather data, commercial data sets and system of record data, and can enhance applications built on IBM Cloud Pak for Data with their own intellectual property. This architecture is extensible with add-on microservices, enabling organizations to easily customize and tune the platform to meet their needs. For example, it can be extended with add-on microservices for AI, industry vertical compliance frameworks, or NoSQL databases.

The IBM Cloud Pak for Data platform enables you to enhance your existing analytics-based applications. The insights and machine learning models derived from the platform can be infused into these applications. Some examples of applications that have been enhanced through IBM Cloud Pak for Data include customer insights, planning and analytics, compliance, fraud and security, as well operations applications.

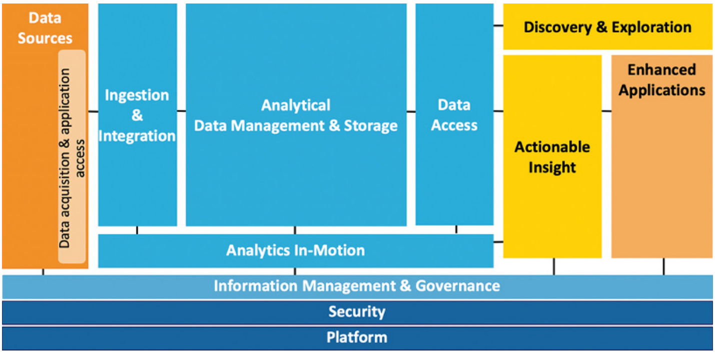

FIGURE 2-1: The IBM Cloud Pak for Data reference architecture.

These foundational components are included in IBM Cloud Pak for Data no matter where it is deployed — in public, private cloud, or your data center. Therefore, for example, if you're deploying the platform on AWS, the entire pre-integrated and containerized platform containing all the components can easily be deployed. The fact that IBM Cloud Pak for Data is an integrated set of services supported by Kubernetes and a comprehensive set of APIs means that all the elements to support data integration are incorporated. At the same time, you can tackle your most immediate analytics needs knowing that all capabilities are in place to support your business needs as they evolve over time.

In addition to the included base components, IBM Cloud Pak for Data can integrate with existing investments as well as add-on services designed to support advancing needs.

Ingestion and integration

REMEMBER

When working with IBM Cloud Pak for Data, the first step is to identify the data sources you're going to include in the platform. You can connect to a wide variety of data, ranging from social data to transactional, application, and third-party commercial data sets. After identifying and acquiring the data, you can ingest and integrate it into the IBM Cloud Pak for Data platform. During this ingestion phase, you verify the quality of the data, cleanse the data, and perform extract, transform, and load (ETL). IBM has integrated a mature set of data discovery, data preparation, staging, and design tools to help streamline the ingestion and integration process. Importantly, ingestion and integration is accomplished in a highly visual, intuitive interface and can easily participate in a broader workflow for a project.

Analytical data management and storage

IBM gives customers flexibility when selecting a data management and storage option for IBM Cloud Pak for Data. In addition to using data virtualization (see Chapter 3 for more details on data virtualization), Db2 Warehouse is part of the included services of IBM Cloud Pak for Data. In addition to the included Db2 Warehouse, customers can choose several add-on databases.

Data access

Connecting business and IT teams with the data they need is critical for any data and analytics platform. As we discuss in Chapter 3, data virtualization is the primary way that developers, data analysts, and data scientists connect with data within the IBM Cloud Pak for Data platform. The data virtualization technology allows analytic queries to be run against data where it resides instead of moving the data to a centralized repository. Data virtualization helps companies maintain high levels of data security, stay compliant with strict governance and regulatory rules, and remove the complexity and cost associated with moving data.

Discovery and exploration

One of the primary challenges that many data analysts and scientists have is finding the correct, relevant data. In many cases, the data that these teams need already exists within the company; however, because of poor labeling, it's difficult to find. The IBM Cloud Pak for Data platform contains enterprise search and cataloging tools to help identify and properly categorize various data sets to ease search.

Refining data consists of cleansing and shaping data. When you cleanse data, you fix or remove data that's incorrect, incomplete, improperly formatted, or duplicated. Shaping data allows you to further prepare your data by filtering, sorting, combining or removing columns, and performing other operations. As you manipulate your data, you build a customized data refinery flow that you can save for automated reuse and modify as use cases evolve.

Actionable insight

Being able to convert data into business insight and predictions is the goal of any analytics platform — collecting large amounts of data is useless unless you are able to gain value. The IBM Cloud Pak for Data platform includes a number of analytic tools that are meant to bring meaning to data. For example, dashboards allow business teams to create interactive visualizations designed for business leadership. Additionally, automation, machine learning, and best practices are built into the business analyst tools so non-data scientists can get actionable insights from data. Alternatively, for data scientists, machine learning libraries and tools can give data scientists all the tools they expect from a data and analytics platform.

The platform is designed to extend with additional analytics tools such as Cognos Analytics and various open source frameworks. In Chapter 4, we discuss the machine learning and AI tools in more detail.

Analytics in motion

In addition to supporting traditional static data that's updated periodically, in some situations, you'll want to gain insights from streaming data that is continuously generated. Common examples of streaming data include Internet of Things (IoT) sensor data, financial data such as credit card transactions and stock tracking applications, and website tracking. For situations that involve streaming data, IBM Cloud Pak for Data includes a service called IBM Streams. This data streaming technology is an established data offering but has been re-architected and containerized to work within the IBM Cloud Pak for Data platform. IBM Streams helps clients ingest, filter, analyze, and correlate massive volumes of continuously generate data.

CONSISTENT DATA CATALOG

A *data catalog* is a metadata repository that's hosted on the platform. The catalog automatically ingests metadata from a variety of data sets across the enterprise. Because the data catalog leverages machine learning, it can discern the meaning of the data without human intervention — a tremendous advancement in speed and quality. For example, by using machine learning, the data catalog can recognize similar columns of data, even if their labels are different. Each new data source that is added is automatically indexed and classified. Having a catalog that can be applied to any data source ensures that analysis will be executed in an accurate manner. One of the powerful capabilities of the catalog is that it enables you to define and enforce governance policies across data sources. Policies can be defined to automatically control access to data resources. To make data understandable for business leaders, the catalog includes a business glossary that links business terms to data assets, policies, and rules. It will automate the process of helping you create 360 views of relevant data. The architecture of the data catalog is designed to provide direct connectors to local data sources. For remote data sources, such as those in the cloud, only metadata is stored in the catalog; the data itself remains stored within the source systems.

The data catalog in IBM Cloud Pak for Data is based on IBM Watson Knowledge Catalog and Watson Knowledge Catalog technology, providing secure enterprise data catalog management that's supported by a data policy framework. This knowledge catalog has AI and machine learning built in at its foundation to help speed the time to analytic results and reduce the chances of human-introduced errors. It connects data and knowledge with the people who need to use it and consists of data policy management to share and control access to assets, catalogs to index and find assets, and projects to work with assets.

The value of the knowledge catalog is that it provides a consistent way to manage policies across all your analytical models across your organization. It serves as a single source of truth for data engineers, data stewards, data scientists, and business analysts. Active policy management helps your organization protect and govern data, so it's ready for self-service analytics and fuels AI at scale.

Information management, governance, and security

A mandate of any complete end-to-end data platform is that it must have robust data management, governance, and security capabilities. In addition, the platform needs to allow full visibility into the lineage of data.

TIP

To help make information management and governance easier for clients, IBM Cloud Pak for Data includes a business glossary, policies and rules, and a data discovery service. These tools stretch across the entire platform. For example, if you put certain rules around customer data when it is ingested, the integrity of those rules will remain intact throughout the platform. Likewise, security features that are included with IBM Cloud Pak for Data protect data, analytics, and machine learning models throughout their entire life cycle. Some of the security capabilities include role based access control (RBAC), continuous monitoring, and encryption. In Chapter 6, we discuss the security capabilities in more detail.

Preparing Data for AI

One of the design points for IBM Cloud Pak for Data is to provide a platform to enable an enterprise to build, run, and manage AI models and create applications that scale for enterprise workloads. AI and machine learning have become an imperative to help organizations compete in an increasingly complex business climate. AI and machine learning provide new capabilities to uncover hidden insights and patterns, predict and shape future outcomes, and enable people to perform higher value work. The best way to prepare is to be able to take advantage of open source tools and capabilities that are the source of incredible innovation. This is precisely why IBM Cloud Pak for Data incorporates popular open source frameworks such as Python, R, and Scala.

Organizations have long valued the simplicity and elasticity of public cloud environments. These clouds deliver time-to-value and enterprise-grade scalability that can lower total cost of ownership and enable the agility they desire. However, when using a public cloud, organizations often need to move data, which leads to loss of control and lock-in. Furthermore, many organizations need to collect, organize, and analyze data behind the firewall on private clouds, customized by country, yet be able to connect to data on various public clouds.

EXPLAINING IBM CLOUD PAK FOR DATA SYSTEM

Many organizations want the flexibility and scalability of IBM Cloud Pak for Data without having to do their own implementation and management. Therefore, IBM recently announced a new deployment option called IBM Cloud Pak for Data System. This is a hyperconverged integrated hardware and software system that is preconfigured. It combines high-end hardware with integrated data and AI software, delivering the agility and scalability of the public cloud but behind the firewall. The environment can be implemented in a few hours rather than months because of the pre-integration.

IBM Cloud Pak for Data System is an all-in-one hyperconverged system for organizations that want to quickly and cost effectively stand up new private cloud instances of IBM Cloud Pak for Data that is

- Pre-integrated with all the necessary systems and software components
- Deployed as a complete private cloud in less than four hours, with no assembly required
- Provides dynamic scalability of compute, storage, and software with on-demand plug-and-play configuration

IN THIS CHAPTER

- **Making sense of disparate data**
- **Analyzing data without moving it**
- **Understanding how data virtualization works**

Chapter 3 Managing Data across Silos

Organizations are drowning in data. What frustrates management is that it's unable to bring the right data elements together across business silos in order to make well-informed decisions. Many business leaders don't even know what other useful data may exist or what information they may need. Management needs to enable its teams to identify and gain meaningful insights from this highly distributed data at the right speed and at the right level of security and compliance.

This chapter delves into the pragmatic ways that businesses can leverage their data without spending countless hours and massive amounts of money trying to move the data. Instead, they can use technology to bring together their data in a safe and compliant manner. You also explore techniques such as consistent data cataloging and data virtualization that can streamline the process of making data work for the business.

Getting Clarity from Data Silos

The process of bringing together data from different sources can be complex because new data sources are constantly becoming available. You have to understand the nuances of each data source. These data sources may be composed of uncountable data stores based on different data types. What is the underlying metadata? How do you translate cryptic data elements to match their business context?

As data analysts rush to query data across silos, there are a number of unanticipated risks. How do you know if the account number in one data source aligns with the way accounts are numbered in another area of the business? How can you be sure that as you combine customer data you are not exposing private information? You need to be able to

- Keep track of the meaning of each data asset in your catalog.
- Understand the rules required to manage each data source as well as automation rules, data classes, and business labels.
- Take advantage of advanced data science and machine learning techniques to bring sophistication and consistency to your data analytics process.

Machine learning and artificial intelligence (AI) tools and their supporting platforms can transform the way organizations can access data and understand nuances of distributed data in a way that results in being able to better predict outcomes. By infusing machine learning into data management, a business can more easily understand its data. For example, customer data with words like "street" and "lane," are likely a customer's address; while a string of nine numbers may likely be a social security number.

REMEMBER

You must make sense of all your distributed data sources in a cohesive and predictable way so the end result will be trusted, business-ready data. At the same time, consider the right way to access data from silos at speed while retaining privacy controls and security.

Maybe you need to move all your key data into a single data lake. A data lake is a powerful mechanism to store and manage highly diverse data sets with common metadata and management. The data lake is designed to take advantage of inexpensive storage and can manage a massive amount of both structured and unstructured data. Data lakes can be useful, for certain use cases — for example, in the area of the Internet of Things (IoT) — sensor and other semi-structured data can be put directly into a data lake. By using the data lake, teams don't need to add structure to the data in order to place it into a traditional database.

WARNING

However, many companies have struggled to gain value from their data lake initiatives. While the data lake may contain massive amounts of data, assuring that the data in the lake is relevant, accurate, timely, and secure are all major challenges, especially given the fluidity of today's data landscapes. In addition, understanding the data's lineage and whether it can be relied on is also a concern.

Controlling and Managing Highly Distributed Data

An important development in distributed data management is the IBM Cloud Pak for Data's service called Data Virtualization. The Data Virtualization service is based on a peer-to-peer architecture that utilizes the processing power of every data source while accessing the data stored at each local source. This approach avoids latency by enabling data to be accessed in real time. It also enhances governance as erroneous data issues are eliminated because ETL (extract, transform, load) and duplicate data storage are not needed. As such, processing times are greatly accelerated bringing real-time insights to applications or analysts more quickly and dependably than existing methods.

The platform is designed so data can be collected, organized, and analyzed wherever it resides, overcoming the accessibility challenges of data silos. Aside from improving access to distributed data, the IBM Cloud Pak for Data platform provides services ensure predictability, manageability, security, and compliance across highly distributed data.

An enterprise data catalog can also help codify operational and privacy regulations in order to enable businesses to operate in a consistent and predictable manner. Managing the governance of disparate data sources can be a challenge if it is not executed in a consistent manner. What are the process and policy rules that have to be applied to all data sources? What are the relationships between data elements that need to be considered? A data catalog provides a consistent and predictable way to manage privacy and process rules while mapping data assets to common business terms. The catalog automates lineage of data in order to help you locate and retrieve information about data objects. The catalog assists in determining the meaning, characteristics, and usage of data. The goal is to help improve productivity within the context of meeting regulatory, privacy, and compliance requirements.

REMEMBER

The ability for business leaders to translate data elements into business terms is essential for making the catalog an important resource for the business to understand their data.

Analyzing Data Wherever it Resides

Gaining insights from massive data sets across their companies — and outside the companies in various clouds — has long been a major challenge for many enterprises. One of the key challenges is to bring the right data together and to analyze the results in a way that is both efficient and cost effective. When organizations are dealing with data that is internal, they may have relied on data warehouses, data marts, and data lakes. However, organizations increasingly need to analyze large amounts of highly distributed data from their own data sources as well as third-party data sources, including highly dynamic cloud data. In reality, moving data to one centralized environment is just not realistic. Besides the cost, complexity, and time involved in moving data, there are also serious security and governance concerns. Also, in some situations, corporate governance requires that mission critical data be stored behind the firewall. In addition, if you're operating across countries, local regulations may not permit movement of certain data assets across boundaries.

Seeing the Value of Data Virtualization

How does an organization leave data where it resides and at the same time gain consistent insights from the data? Data virtualization is an important emerging approach that can successfully manage data access and manipulate and query data without having to move the data into a single repository or warehouse. In essence, data virtualization is a peer-to-peer architecture where queries are broken down and sent closer to the data sets. After all the sub-queries are processed, results are combined along the way, eliminating the application entry point/service node as the bottleneck. Data virtualization simplifies data analytics and keeps information up to date and accurate because you're querying the latest data at its source. Because of the common user interface across all of the IBM Cloud Pak for Data services, the data is auto-discovered and the metadata is automatically available to the user.

IBM Cloud Pak for Data brings six core advancements to the world of data virtualization, making it a more practical approach than ever before, all within a cloud-native environment:

- **Access current data:** Get always-current analytics across distributed data sources. Experience a single virtualized data repository where your SQL applications can connect and your analytics get access to always-current distributed data.
- **Improved performance:** Automatically self-organizes your data nodes into a collaborative network for computational efficiency, leveraging networked devices for polynomial processing gains.
- **Secure to the core:** Data isn't cached in the cloud or on other devices. Credentials for your private databases are encrypted and stored at the local device. Data is privately managed on that device.
- **Flexibility:** Support for multiple application query languages (SQL, stored procedure languages, R, and Python) and data sources such as Cloudera/Hadoop, Db2, Db2 Event Store, Informix, Oracle, PostgreSQL, Microsoft SQL Server, Cloud Object Store, MongoDB, and Teradata.
- **Simplicity:** A single, intuitive console with an interactive interface to query data, manage users, and visualize data-node constellations.
- **Self-optimization:** The virtualization is self-optimized and automated through machine learning and adaptive algorithms that continuously tune its performance.

The Elements of a Data Virtualization Service

The data virtualization service is the data consumption layer needed to bring elements together to support the abstraction of data. There are two reasons why you would want to manage data through virtualization. Firstly, the data you want to analyze may be so sensitive or regulated that moving it is simply not an option. For example, it isn't realistic for a healthcare organization to move all of its patient records to the cloud without considerable amounts of time and money spent on removing any trace of personably identifiable information. This is further complicated when businesses have to deal with data protection requirements across different geographies. Another example is an insurance company's underwriting data. While there might not be local and federal regulations prohibiting the movement of underwriting data, this type of data is the lifeblood of any insurance company and represents the insurance company's ability to successfully compete and offer new, innovative offerings. Therefore, the company would never want this type of intellectual property exposed to competitors.

In addition to security concerns, the data might be so complex and large that it doesn't make sense to move to another environment. For example, a retailer's transactional system that contains inventory data across locations, customer buying history, and warehousing information is large, constantly updated and extensive. Trying to move all of this data to a centralized repository while maintaining the data integrity and timeliness will be a major challenge.

In all these cases, data virtualization allows an organization to gain a comprehensive understanding of its data without moving the source data. This supports the ability of data scientists to build machine learning models using a broader data set to apply new levels of analytic insights.

How IBM's Data Virtualization works

IBM's Data Virtualization service is designed as a clustering technology that creates a virtual data set. A query is issued by the application/data consumer layer (see the next section) against the system as if it is a single data source. A Data Virtualization service node (or Coordinator) receives the request and distributes the work to data nodes (for example, databases, flat files, and so on). Data nodes self-organize into a constellation whereby they can collaborate with a small number of peers. Peer devices collaborate to perform the analytics on their data and return the result through the constellation mesh. The coordinator receives mostly finalized results from just a fraction of devices, completes and finalizes the query result, and returns it to the application.

An application or the data consumption layer

An application or the consumption layer uses the Data Virtualization service to bring together a single view of data across silos. This eliminates the need for ETL and duplicate data storage, and processing times are greatly accelerated. This brings insightful real-time results to decision-making applications or analysts more quickly and dependably than existing methods.

The role of the service node

A query is issued by the application against the system as if it is a single data source. A service node (or coordinator) receives the request and distributes the work out to data source nodes (for example data assets that are part of organizational silos). The head node coordinates the whole data virtualization process.

The role of the data source node

A data source node works with its peer data source node(s) to address the data request coming down from the query sent by the service node. These data source nodes are part of organizational data silos.

The constellation

Data source nodes self-organize into a constellation whereby they can collaborate with a small number of peers. Peer data source devices collaborate to perform the analytics on their data and return the result through the constellation.

To enable all the data to work as one unified virtual data object to the application or data consumer layer, the Data Virtualization service builds this peer-to-peer mesh. This service gives the application/data consumer layer the ability to connect to data sources globally to query and govern as if the data was within one database. In this way, rather than moving data to where the query is being performed, the data virtualization service pushes the query to the data source and encrypts the resulting data set used for the query.

A mesh network of the constellation dynamically connects data nodes in the most efficient route possible to increase efficiency and speed. One of the benefits of a mesh network is that it's dynamically self-organizing and self-configuring. This technique is helpful for better performance, allowing higher parallelism.

Because of the architecture of IBM Cloud Pak for Data, the Data Virtualization service works in conjunction with the data catalog so data can be easily identified based on context and meaning. In addition, the data virtualization works hand in hand with defined governance policies so it ensures the proper usage of information. Users will have to have the right level of access before they're able to use the data as part of their data query process. This is essential to ensure security and to appropriately govern the use of data.

The Entire Process of Data Virtualization

In many data marts, data lakes, or warehouses, it is common to have multiple copies of the same data source that are no longer relevant or accurate. These data challenges can have serious problems when business leaders are making decisions based on out of data information.

TIP

One of the benefits of data virtualization is that you can avoid the complexities of the traditional ETL process because the data stays in place. In this way, you have real-time access to data at its source, which eliminates the data consistency challenge with constantly changing data. The data virtualization process ensures that any changes to the data source are kept in sync and are easily integrated with governance solutions. Its modular architecture allows for fast deployment.

IN THIS CHAPTER

- » Infusing machine learning and AI into the data platform
- » Looking at IBM Cloud Pak for Data analytics services

Chapter 4

Building, Running, and Managing AI Models

Creating a unified view of data across your organization is no easy task. You need to bring together data that's designed to operate with different databases, different technologies, and different structures, and all with different purposes. In this chapter, we discuss the requirement to deploy and manage machine learning models and artificial intelligence (AI) as part of an overall data platform. At the same time, we describe the capabilities that are available within the IBM Cloud Pak for Data platform, designed to make the use of machine learning and AI easier for businesses.

Leveraging Machine Learning within a Data Platform

Creating a consistent and predictable way to manage data across an enterprise is much more complicated than it may seem. You need to discover all sources and ensure that there are consistent definitions and rules across these sources. You also want to make sure that requirements for compliance and security are followed and are consistent. If you have only a couple of data sources it is relatively easy to manually review and compare sources. However, in reality, most organizations have hundreds of internal and external data sources that they rely on to perform in-depth advanced analytics.

REMEMBER

The only way to accurately and systematically provide consistency across data silos is to leverage machine learning algorithms within your data management processes. This process requires determining and discovering metadata, business rules, governance, and security. Machine learning models are ideal for helping to manage all of these operations within a data platform. Machine learning algorithms can help to automate data cleansing, tagging unlabeled information, understanding the metadata within different data sources and more.

As a unified data management platform, IBM Cloud Pak for Data has infused machine learning algorithms into the fabric of the platform — it's everywhere. The platform automatically crawls through enterprise data sources to discover metadata and map a business vocabulary to the underlying data and analytics assets.

TIP

The key advantages of integrating machine learning into a data platform include the following:

- Makes data analysis easier for all users
- Reduces errors and automates routine tasks
- Provides a mapping between metadata, the data catalog, and the business glossary
- Discovers underutilized and embedded data
- Allows data scientists to focus on analysis rather than data preparation
- Automates the design and generation of AI models

Analytics Services within IBM Cloud Pak for Data

IBM Cloud Pak for Data has built in services to manage your data platform, including data virtualization, data catalog, business glossary, platform security, data policies, and more. In addition, a number of services are available to support analytics, machine learning, and AI use cases. In this section, we explain these integrated services.

Built-in analytics services

IBM offers several tools that are built into IBM Cloud Pak for Data. They're intended to help customers create a foundation for their advanced analytics. They are

- **Cognos Dashboards:** Cognos Dashboards enable business analysts to incorporate drag-and-drop visualizations into their data environments. In this way, end-users can explore data within the IBM Cloud Pak for Data environment through a visual interface.
- **Watson Studio and Watson Machine Learning:** IBM Cloud Pak for Data includes Watson Studio and Watson Machine Learning, which are open-source-based services for data scientists and data engineers. IBM Cloud Pak for Data supports all the most popular open source data science frameworks and tools, such as Jupyter, Zeppelin, RStudio, TensorFlow, Keras, and more.

Service offerings

Numerous offerings are seamlessly integrated with IBM Cloud Pak for Data, thereby enabling you to easily extend your AI capabilities as your needs and sophistication grows. These offerings include

- **Watson Studio Premium:** In addition to the open source tooling in IBM Cloud Pak for Data, an add-on for Watson Studio provides access to SPSS Modeler, Decision Optimization, Hadoop Execution, AutoAI, and many other advanced data science capabilities.
- **Watson OpenScale:** Watson OpenScale provides insights into how AI models are making decisions, such as model explainability. This helps determine the accuracy of the model and how it arrived at its recommendations. OpenScale can also detect bias in the model through simulation. The simulation mimics the model in order to determine outcomes.

- **Watson Assistant:** Watson Assistant is a service designed to add a conversational interface into any application, device, or channel.
- **Watson Discovery:** Watson Discovery is designed to rapidly ingest normalize and enrich index and search your structured and unstructured content. The benefit of Watson Discovery is that it makes it easier to scale in order to support large amounts of data needed for the selected algorithm.
- **Watson Compare & Comply:** This service uses AI to help understand contracts and documents.
- **Watson Speech to Text and Text to Speech:** This engine combines information about grammar and language structure with knowledge of composition of audio signal to transcribe customer care calls, meetings, conference calls, and so on.
- **Watson Knowledge Catalog Premium:** Watson Knowledge Catalog is a governance catalog designed to help analytics professionals find, curate, and share data and analytical assets. The catalog enables users to better understand the relationships between data and analytics assets, and allows for self-service while enforcing policies and rules.
- **Cognos Analytics:** Cognos Analytics is a business intelligence service for interactive visualization, dashboards, storytelling, and reporting.

Bringing the Parts Together

TIP

One of the benefits of IBM Cloud Pak for Data is the ability to leverage open source software, IBM's proprietary analytics software, and non-IBM software on a unified data and AI platform. The driving force behind the creation of IBM Cloud Pak for Data is to support teams through unification and simplification across data, tools, and users.

IN THIS CHAPTER

- **Ensuring trust and transparency of data**
- **Maintaining data privacy and compliance**
- **Understanding how IBM Cloud Pak for Data changes the approach to data quality**
- **Improving data quality with IBM Cloud Pak for Data**

Chapter 5 The Imperative for Data Quality

With the advancements in advanced analytics, it's not surprising that organizations are increasingly relying on their data to predict the future of their business, make better decisions, and enable higher value work. The power of machine learning models is helping analysts and data scientists to gain insights that were never available in the past, and apply them to improve business outcomes. Therefore, the accuracy of this data becomes a business imperative. When you remove the original context of how the data is used, there could be problems when that data is used in a different form and with a different use.

Trust and transparency has become an outright imperative. In addition to emerging privacy and regulatory issues, more organizations demand trust and transparency, as they become increasingly data driven. It's the old adage of "garbage in, garbage out." Organizations need to put greater emphasis on three key principles that feed all downstream usage of data:

- Know your data to ensure how it can be used.
- Trust your data, knowing its business-ready.
- Use your data with transparency and lineage.

In this chapter, we focus on the need to know, trust, and use data across silos. To be successful, you need to understand metadata, business rules, and the overall meaning of data. This requires organizing your data sources, governing them appropriately, and making them usable by many different users within your business.

Needing Consistent Data

WARNING

What's holding businesses back from gaining the maximum benefit from their data? One of the most problematic issues is that different parts of a business may define and label data differently. One division may define an individual person as the "customer" while other divisions may consider an individual's company the "customer." Without a consistent way of defining and describing data (a taxonomy), a business can't achieve a single view of the customer. In addition, you may have a problem if there are errors in the data or if there is conflicting data. Someone may have inaccurately entered data that has gone unnoticed for years. When that data was rarely used, no one paid attention to that data set. Errors that were once considered insignificant can potentially cause serious harm as more and more data is brought into the decision process. Ironically, even when all the data is consistent, there can be so much data from so many sources that it's difficult for analysts to understand the meaning of the information.

If you can't trust the accuracy of your data and have transparency on its lineage and usage, your organization will be at a competitive disadvantage. For example, if you lack consistent metadata, it will be almost impossible to analyze data across silos — rather than having data scientists focus on actionable insight, they will be bogged down with the complexity of preparing data for analysis. One of the biggest problems suffered by data managers is the inability to separate data from the underlying application and processes. For example, CRM application data is often tagged in specific ways that are proprietary to that application. Generalizing those data identifiers can be a challenging task. The lack of flexibility and consistency will lead to failure.

Requirements for Data Quality

How can you move to a plan that ensures that your data is ready to be used in multiple situations to solve a variety of business problems? You can't simply assume that your data is accurate and secure. You need to have a plan and an approach that prepares you for the world beyond the current uses of the data. It is increasingly more complex for organizations to be able to ensure quality for their massive amounts of internal and external data. How do you know if your data can be trusted? There are four criteria needed to support your organization:

- **Integration:** Data must be prepared in a way that makes it possible to link data elements together to support a wide variety of use cases.
- **Accuracy:** Each data element has to be consistent and correct. Something as simple as misspellings or abbreviations can cause problems. Ensuring a single version of the truth is an absolute imperative.
- **Availability:** Data has to be prepared in advance so it can be available to support a variety of situations, increasingly in self-service scenarios. Data must be accessible no matter where it resides in the most efficient and compliant manner.
- **Timeliness:** Data changes more rapidly every day as it grows in volume, variety, and velocity. Businesses need to trust that the data they're using to make decisions is up to date.

Using IBM Cloud Pak for Data to Prepare for Data Quality

Data quality can't be managed in isolation in a data platform. In an era where machine learning is becoming the focal point of advanced analytics, data quality has to be integrated into the overall approach to data management. Data quality services are deeply integrated into the IBM Cloud Pak for Data platform. IBM is leveraging its mature portfolio of data quality offerings to support the ability to break down data silos in a consistent and predictable manner.

How is IBM Cloud Pak for Data different? Traditionally, IBM has offered a variety of separate offerings that address everything from data cleansing to data integration to data quality and ingestion tools. Now all these data quality offerings have been modularized and containerized as part of the IBM Cloud Pak for Data platform. These capabilities are supported by standardized APIs, workflows, and user experiences. These service APIs can be used to onboard data into the data catalog. Each service can connect to other data quality services through the APIs. The workflows and user experience make this all accessible.

TIP

One of the benefits of this platform approach is that the process of discovering data and its attributes can be automated. This means that when data from a source is identified, it's matched against the metadata from the data catalog. As customers identify key use cases, they're able to add services on top of the IBM Cloud Pak for Data platform. For example, customers can take advantage of IBM's Data Refinery offering. The refinery (which is part of Watson Studio) is a tool that automates the process of transforming raw data into usable information that can be integrated into the data catalog. IBM Watson Knowledge Catalog is another example. The Watson Knowledge Catalog is a unified data catalog that can help your data users quickly find, curate, categorize, and share data, analytical models, and their relationships with other members of your organization.

Organizing Data

In order to establish a consistent and predictable way to manage all your data across silos, you have to start with an organizational structure. Because of the complexity of the data environment, there are a number of elements that have to be considered and integrated together to achieve success.

Discovery and search

If you don't know what data you have and where it resides, nothing is going to be achieved. Therefore, it is imperative to have the ability to discover what data elements are contained within your various data sources. Discovery can't simply be the process of finding data; there needs to be a way to understand the context for data elements, including what they mean and how they relate to data constructs in other data sources. After you're able to search for data and identify data, you are in a better position to begin to catalog that data so it can be used properly to make business decisions.

Data transformation

One of the biggest challenges of integrating data silos is the requirement to map data that exists in different forms so they can easily communicate and integrate. It is therefore necessary to provide a set of tools that can understand and then transform the data into a common format. For example, some data is stored in highly structured traditional relational databases while other data sources are unstructured text. Transforming this data into a consistent format will be the first step in breaking down data silos.

Creating a data catalog

A well-designed data catalog becomes a single trusted source that enforces policy and governance of the highly distributed data. One of the benefits of the data catalog is that it can pull metadata from existing data sources. The catalog then classifies data through the use of machine learning models. The classification capability benefits customers by automatically understanding what the data means in context with the data source and its defined use. The data catalog works in concert with the data virtualization capabilities detailed in Chapter 3.

IBM's Data Catalog is a unified environment that is intended to help users find the data sources they need and then curate and categorize the data so it can be shared. After creation, the catalog can be used to create analytical models and codify the relationships between data sources in a consistent and predictable way.

Business glossary

A business glossary works hand in hand with the data catalog in order to help translate complex metadata into terminology and definitions that are understandable by business professionals. That is, business-ready. The business glossary makes it possible to create and categorize business terms while assigning ownership for who is allowed to manage the information.

It’s All About Trusting the Data

In the era of the hybrid and multicloud environment, it is imperative to be able to manage data across silos. It isn’t enough to simply integrate data between sources. Success means bringing together data in a trustworthy manner. Context matters in creating a corpus of data that can be used to analyze data across divisions to help the business make future decisions. This requires having governance embedded within the platform so data quality, as well as access rules and restrictions, aren’t an afterthought.

IN THIS CHAPTER

- » **Securing and governing data through its life cycle**
- » **Looking at IBM's data catalog**
- » **Complying with IBM Cloud Pak for Data**
- » **Securing your data while performing analytics**
- » **Approaching governance holistically**

Chapter 6 Protecting Personal Data for Security and Compliance

The goal and aspiration for many organizations is to democratize data so it can be used across the business to support analytics for a variety of users. How are companies democratizing data and making it available to everyone within the organization while also assuring that it is secure? Are you able to protect sensitive data while performing in-depth analysis? If you anonymize data are you still able to glean actionable insight from the incomplete data? The best approach is to create an environment where data is managed based on ubiquitous and embedded security and compliance. This chapter discusses the type of services and approaches that protect the company's use of data across silos and the protection of customer, partner, and internal data.

Securing and Governing Data through the Entire Analytics Life Cycle

Businesses are confronted with a variety of security and compliance regulations. It is critical that data be protected from the point of ingestion, through storage, analysis, and model building. In addition, emerging data governance and compliance rules mandate that companies must be able to quickly locate specific types of data about customers and employees. For example, requirements such as the EU GDPR (European Union General Data Protection Regulations) and the California Consumer Privacy Act (CCPA) are designed to give consumers greater control over data about them. Regulations like these are increasingly common and require businesses to ensure personal information is protected.

Coping with new industry regulations such as GDPR and CCPA is a constant battle for the enterprise. To make it easier to understand the impact of a new regulation to existing assets, the regulatory accelerator service in IBM Cloud Pak for Data incorporates machine learning techniques to help co-relate regulatory terms to business terms within the enterprise business glossary. It also provides the ability to map enterprise data assets to regulatory terms and enables policy definition and enforcement across the enterprise to ensure privacy.

WARNING

In addition to financial and legal penalties, failure to comply with best practices and regulations puts customer trust in jeopardy. Consumers don't want to conduct business with companies that don't protect personal data. Likewise, businesses won't partner with other businesses that don't take data governance, compliance, and security seriously.

Adhering to data regulations is complex. First, you need to understand the types of data that you are gathering. What types of industry and regulatory provisions apply to those data types? In addition to regulations, do you have internal best practices in place to protect certain data types? Is the personal data you are using in your analysis actually improving the analysis of data?

There is no single step that can properly govern and secure data. Instead, you need to layer a number of technical and corporate best practices to protect data. For example, encrypting data on its own won't be sufficient to satisfy compliance requirements. To properly understand how your organization is approaching compliance, you need to understand the meaning of different data fields stored across disparate sources. You have to know what business terms mean and if they are relevant to compliance. Date of birth, for example, might be stored in many different formats across your business's data sources. Furthermore, other types of data, like personal preferences that are stored in a customer database, must be identified and stored.

The Role of IBM's Data Catalog

IBM's Data Catalog provides a consistent and predictable way to manage data for security and compliance. This web-based tool enables you to explore and understand information. The catalog would also contain a taxonomy of business terms based on personal data, such as first name, last name, address, date of birth, and industry specific terms. For more details on data catalogs see Chapter 2.

Within IBM Cloud Pak for Data, you can store essential sets of terms in order to classify all your data across information sources. With the classification capabilities within a data catalog, you are able to identify different types of information that must be highlighted and protected. Furthermore, the machine learning technology that's built into a data catalog helps organizations more quickly identify sensitive information to which governance rules may be applied.

Assuring Data Security

Attacks on corporate, customer, and employee data are nearly constant and have become highly complex and targeted. Traditional methods of protecting sensitive data are no longer sufficient. For example, firewalls alone can't stop intruders. In addition, you must consider the fact that many breaches are perpetuated by insiders who are either well-meaning employees who make a mistake, or employees or contractors that have ill-intent.

Within the IBM Cloud Pak for Data, there are a number of capabilities to help secure data from both outsiders and insiders. Within the platform's governance services, it is possible to either mask data such as a credit card number or scramble numbers so that the data becomes a random set of numbers that cannot be interpreted. Combining encryption with advanced masking techniques will help ensure your data is secure and compliant.

In addition, IBM Cloud Pak for Data offers privacy enforcement through the supported integration with Guardium, IBM's enterprise data security offering. Sensitive data found during data discovery can be automatically protected by IBM Guardium.

REMEMBER

One of the most important ways that IBM Cloud Pak for Data helps maintain data security is through the data virtualization technology. Data virtualization allows customers to perform analytics and build machine learning-based applications without needing to move data to an analytics environment. By keeping the data where it resides, the existing security framework around your data doesn't need to be disrupted. Additionally, because your data can remain on premises, data virtualization helps you avoid objections to moving data to the cloud. See Chapter 3 for more details on data virtualization.

Ensuring a Holistic Approach to Governance

One of the benefits of IBM Cloud Pak for Data is that it's designed as a set of foundational services based on a container and multicloud architecture. The core foundation is that all services can work together to create a modular platform. Governance baked into the core is a key guiding principle of IBM Cloud Pak for Data. Leveraging IBM's expertise and experience in the Unified Governance and Integration space, governance-oriented services made available on the platform are a combination of workforce efficiency enhancing technologies that enable best practices such as governed self-service and regulatory oriented services that help the enterprise stay compliant with industry regulations. Thanks to data profiling, discovery, metadata extraction, data stewardship, data quality assessments, data classification, and a data transformation service that all feed a metadata catalog, enterprises can truly make governance not only about regulatory compliance but also an enabler of better business outcomes.

IN THIS CHAPTER

- Creating business objectives
- Understanding the need to use an open and standards-based platform
- Encouraging collaboration
- Enabling hybrid multicloud support
- Making sure you have security and governance
- Embedding machine learning into a data platform

Chapter 7 Six Best Practices for Managing Data

Beginning your journey to turn data into a strategic tool for growth can be complicated. There are often many opportunities to take advantage of your data in new ways in order to impact business change. The good news is that you can bring together data sources in boundless ways that would never have been possible in the past. The bad news is that you have to be able to select a starting point that allows you to have early success so you can begin your journey with the support of senior management. How can you start your transformation journey with quick wins? This chapter focuses on six best practices for getting started on your data journey.

Establish Your Objectives

When you are building your data and artificial intelligence (AI) strategy, begin by establishing a set of short-term and long-term objectives. While it would be great to upend your company overnight, you must recognize that you can't do everything at once. To get organizational buy-in, you need to be able to demonstrate quick and impactful "wins." Therefore, set short-term goals that can be quickly achieved as you set the longer-term strategy for transformation.

REMEMBER

The following are top mandates for any data and analytics strategy:

- **Access data anywhere.** Define a strategy that enables access across all your data sources. Your goal should be enabling access to all your data no matter where it resides. Providing a consistent method of access will prepare you for the future.
- **Focus on the quality of your data.** Just because you have massive amount of data, you are not guaranteed success. You have to be sure that the data you are using is clean, well defined, accurate, and timely.
- **Ensure security and compliance.** As soon as you begin to use your data to understand customer expectations and future requirements, you have to guarantee that sensitive information is protected. Your customers and partners expect you to protect their privacy. You need to make sure that you're following compliance guidelines.
- **Your AI models need to be well understood.** As your company implements advanced AI and machine learning models, it is imperative that you are able to understand what is happening behind the scenes. You will begin to make critical business decisions based on your data and models your team creates. You need to explain where your conclusions came from. You therefore need to be able to explain what's behind the models.

Implement an Open Architecture

Before you can begin to implement a meaningful data strategy, you need to think about your technology foundation. The most pragmatic approach is to create an open information architecture. This open approach allows your teams to flexibly leverage innovative and proprietary services that will differentiate your data and AI strategy from your competitors. Take advantage of open source standards, so you aren't locked into a single vendor. When hiring new employees, open source communities have large networks of developers who will be familiar with your analytics technology stack.

Establish a Collaborative Environment

REMEMBER

Because you have set up an open information architecture, you have the ability to encourage collaboration. Business teams need to be able to work with their technical counterparts in order to establish a meaningful data strategy. In addition, technical teams in different locations can collaborate on a common platform. Successful organizations are creating centers of excellence that include stakeholders from across the business and IT in order to provide best practices.

Enable Hybrid Multicloud Support

Most businesses rely on a hybrid computing environment that includes on premises resources, and a variety of cloud technologies, including Software as a Service (SaaS) applications, as well as cloud infrastructure. Business are simply too complex to rely on a single computing platform — different business units and use cases require different technologies. In addition, advances in the sophistication and the cost effectiveness of cloud computing has changed the dynamics of data and AI. To prepare for the future, you must develop a strategy to support multiple public and private clouds.

Infuse Security and Governance at the Core

Without a well-orchestrated and properly executed security and governance strategy, your company's livelihood will be at risk. In addition to regulatory and legal expenses, failure to protect customer data will result in distrust from customers and partners. A data driven organization must have a plan for managing and analyzing data securely. Different data types will have different requirements.

For example, from a security and financial perspective, it may make the most sense to leave some data where it currently resides. In other situations, you may want to aggregate data into a data lake or an analytics staging ground. Likewise, in some scenarios, distributing data and analytics to the edge for speed and agility is the best approach. Governance requirements should be followed across your entire data and AI environment to protect your business' reputation and adhere to overall regulations, while encouraging self-service analytics.

Embed Machine Learning Throughout

A practical way to manage massive amounts of data across your business is to leverage tools that have machine learning built into them. These machine learning-based tools help to automate processes such as metadata management and governance to speed time to results, reduce mistakes, and allow analysts to focus on business results rather than data preparation. Incorporating machine learning into your data and analytics platform makes it possible to keep data rules and policies current as the business changes.

Accelerate your successful journey to AI

To be successful with your journey to AI, you need a platform that's capable of managing, governing, and securing your data throughout the analytics process. This platform must also be able to analyze data no matter where it resides. In addition, the platform should support the open source technologies, containerization, and the latest machine learning libraries. In this book, we feature IBM Cloud Pak for Data and tell you how it can help your business accelerate its journey to AI.

Inside…

- Understanding IBM Cloud Pak™ for Data
- Uncover the platform's architecture
- Analyze all your data
- How to secure and govern data
- Using a data platform across clouds
- Understanding the business imperative

IBM

Judith Hurwitz, President of Hurwitz & Associates, is a consultant, thought leader, and coauthor of 10 books, including *Augmented Intelligence*, *Cognitive Computing*, and *Big Data Analytics For Dummies*. **Daniel Kirsch,** Managing Director of Hurwitz & Associates, is a researcher, author, and consultant in AI, cloud, and security.

ISBN: 978-1-119-59345-4
Part #: 95027395USEN-00
Not For Resale

9 781119 593454

for dummies®
A Wiley Brand