Brought to you by:

HITACHI
Inspire the Next

Data Fabric

for dummies®
A Wiley Brand

Effectively build and deploy a data fabric

Optimize data access and support compliance

Modernize a data fabric with DataOps

Hitachi Vantara Special Edition

Ed Tittel

About Hitachi Vantara

Hitachi Vantara, a wholly-owned subsidiary of Hitachi, Ltd., guides its customers from what's now to what's next by solving their digital challenges. Working alongside each customer, Hitachi Vantara applies its extensive industrial and digital capabilities to their data and applications to benefit both business and society. More than 80 percent of the Fortune 100 trust Hitachi Vantara to help them develop new revenue streams, unlock competitive advantages, lower costs, enhance customer experiences, and deliver social and environmental value.

Data Fabric

Hitachi Vantara Special Edition

by Ed Tittel

for dummies®
A Wiley Brand

Data Fabric For Dummies®, Hitachi Vantara Special Edition

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

ISBN 978-1-119-79116-4 (pbk); 978-1-119-79117-1 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. For details on how to create a custom *For Dummies* book for your business or organization, contact info@dummies.biz or visit www.wiley.com/go/custompub. For details on licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

Some of the people who helped bring this book to market include the following:

Project Manager: Martin V. Minner

Acquisitions Editor: Ashley Coffey

Senior Managing Editor: Rev Mengle

Business Development Representative: Matt Cox

Production Editor: Tamilmani Varadharaj

Hitachi Vantara Contributors: Madhup Mishra, Anand Sagar Rao Vala, Lothar Schubert

Table of Contents

Introduction

In computer-speak, the term *fabric* covers a range of situations in which strands and types of data form the warp, and a variety of controls and tools the weft. Or maybe it's the other way around? Either way, such a fabric describes how computing, networking, storage, and software components work together to deliver data and services.

What Is a Data Fabric?

A data fabric is a new-ish concept that describes enterprise-wide data management and delivery. Importantly, the data fabric concept is decoupled from any physical implementation and can span infrastructures across multiple clouds, datacenters (core), or even edge systems (such as Internet of Things devices, local machines, or even mobile devices).

Today, the notion of one or more central data repositories that act as hubs for access and storage is giving way to a mesh or data fabric that overlays all data sources. Simply put, a *data fabric* provides each consumer of data with a consistent and coherent view of all the integrated and packaged data they may need, irrespective of location (the user's or the data's). Also, a data fabric is a living, always-evolving collection of capabilities that grows and changes along with whatever organization it serves.

Key capabilities of a modern data fabric include:

- **Intelligent pipeline management:** The means for organizing how data gets onboarded, sanitized, labeled, and incorporated for consumption. Any particular data source in the edge-to-cloud landscape may support one or more (sometimes many) applications. Each step in a pipeline may need to access the same data, which is exactly what an intelligent data fabric can deliver.
- **Policy-based orchestration:** A set of rules and organizing principles whereby data distribution and diversification can come together to make applications function optimally. Diversification lets organizations allocate different types of work where they best fit within an edge-to-cloud

architecture. Distribution permits splitting up workloads across many instances of devices or applications, sometimes to do huge amounts of work in unlikely places.

- **Edge-to-core-to-cloud coverage:** With an edge-to-core-to-cloud data fabric, organizations can distribute data and analytics as needed. But they can consume data in the best locations, too, so it can be found, managed, curated, and sorted efficiently. A data fabric handles the work of assembling and grooming data, so when applications consume it, it's been subjected to a standard set of checks and tools for data quality, stewardship, governance, compliance, and cataloging.
- **Automated semantic enrichment:** When data gets onboarded and incorporated into a data fabric, it gains value to the extent that it is properly labeled and cataloged, with proper integration of multiple, related data sources and views. The idea is to enrich, activate, and govern data assisted by artificial intelligence to drive and improve business outcomes. Some might argue this is the real (and best) reason for implementing a data fabric.
- **Extensible, modular architecture:** A data fabric relies on a well-documented, programmable architecture that is designed to be highly modular. That is, functions are nicely separated and easy to mix and match across storage and compute resources. It's also designed to be easily extensible. Thus, when new data-handling or data-management functions are needed, they are easy to add. Likewise, existing modules are easy to enhance and adapt.

All in all, a data fabric ensures that data from anywhere — from edge, to data center, to multiple clouds, and beyond — becomes easy to find, interpret, and use while remaining safe and secure (and compliant with relevant law and regulations).

Beyond the Book

Between these covers, I've tried to stuff as much information about modern data fabrics as will fit into 48 pages. If you find yourself ready for more such stuffing, please visit HitachiVantara.com, and search on "data fabric." You'll find lots more information, including brochures, white papers, tech briefs, videos, resources, and more. Enjoy!

IN THIS CHAPTER

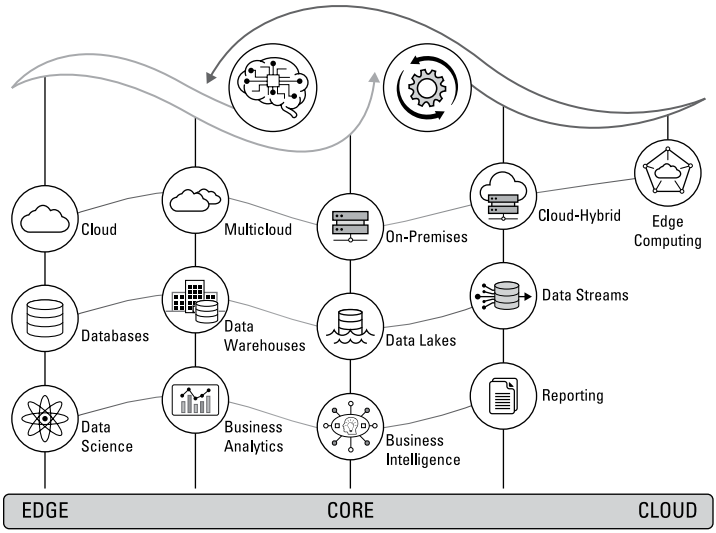
- Uncovering the data fabric's components
- Overcoming implementation challenges
- Grappling with data fabric modernization

Chapter 1
Understanding the Data Fabric

Data is everywhere in modern organizations. It lives in databases, data stores, applications and more. Data may also be found at the network edge in branch or remote offices or industrial operations, in data centers at the network core, or somewhere in-between in multiple clouds. A data fabric brings all these pieces and parts together, whips them into shape, follows best governance practices, and makes sure that data is accessible where and when it's needed. This chapter digs into underlying terms and concepts, and explores how a data fabric works, what it's made of, and what it can do.

Data Fabric Redux

A *data fabric* is best understood as an enterprise-wide data capture, cataloging, management and delivery system that covers an organization's IT infrastructure from the core datacenter to the edge and into the cloud (aka *edge-to-core-to-cloud*). It also brings with it a set of best processes, practices and methodologies for dealing with data. Thus, a data fabric provides data consumers with a consistent, coherent, governed and performance-optimized view of whatever integrated and packaged data they might need, wherever it happens to reside, as shown in Figure 1-1.



Source: Hitachi Vantara

FIGURE 1-1: The data fabric covers all locations and all platforms and formats.

STRUCTURED, UNSTRUCTURED, AND SEMI-STRUCTURED DATA

Data comes in many forms and formats, too. Its internal characteristics permit it to be dropped into one of these three buckets:

- **Structured data** is easy to organize and typically follows a rigid, preset format. Thus, database records are fine examples of such data.
- **Unstructured data** is complex and may include qualitative information that is difficult to reduce into or organize into database records. Case notes and interview transcriptions fall under this broad heading.
- **Semi-structured** data includes characteristics of both structured and unstructured data. Customer interactions, Internet of Things (IoT) data and social media feeds often fall into this category because they include identification, account, time, measurement, and/or transaction data (structured) but may also include descriptions, comments, ratings and review data (unstructured).

A modern data fabric generally incorporates humongous amounts of all three kinds of data beneath its organizational umbrella.

Revealing Data Fabric Elements

Numerous interesting responses answer the question: "What kind of things might belong to a data fabric?" Here are some of the most common such elements:

- **Databases and data warehouses:** A *database* is a collection of records stored to match a set of record definitions and relationships called a *schema.* Usually based on specific database engines such as Oracle, MySQL, SQL Server and so on, databases work with their own interfaces to handle queries ad hoc, or with applications that ingest query-based data from databases as they run. A *data warehouse* aggregates structured data from one or more sources (customer orders, medical records, financial or brokerage accounts, and so on) to support ad hoc queries, data mining, business intelligence, and other kinds of analytics.
- **Document repositories:** These provide storage for documents, which generally represent files that include textual and other forms of data (images, diagrams, and so on) in many forms and formats. Documents are a typical type of semi-structured data because they are searchable and usually include regular basic structures.
- **Data lakes and other unstructured data collections:** These store data in its natural (or raw) format as it was generated or collected, usually in the form of amorphous data structures called *object blobs without forcing a schema on it.* Where data warehouses store data in hierarchical files or folders, a data lake uses a flat architecture (no hierarchy) to store data. Hadoop and its massively parallel, highly available, self-healing filesystem (HDFS) is a traditional environment in which data lakes form and operate. In recent years object store-based data lakes are taking over, especially in the cloud.
- **Streaming data:** This is an increasingly common form for semi- and unstructured data. Such data is constantly generated by various sources, such as sensor or device outputs from the Internet of Things (IoT). Data streaming relies on mechanisms to make content immediately accessible and needn't be downloaded as discrete files. Instead, such data streams may be recorded, often in fixed-length circular files (old data is trimmed from the back of the file even as new data is added to the front) or using message queue (such as Apache Kafka or MSMQ).

Facing Data Fabric Trends and Challenges

With all kinds of data scattered around organizations in all kinds of structures, formats and repositories, it should come as no surprise that organizations must face — and overcome — significant challenges when implementing a modern data fabric. In the sections that follow, I examine some industry trends in the data arena, and follow up with an overview of the challenges organizations face.

Trends in data fabrics

The big picture view of data is a big picture indeed. Simply put, organizations are dealing with increasing numbers for:

- **Data volume:** This explodes, as organizations acquire, generate, and collect unimaginable amounts of data. In 2015 the total amount of data stored was around 4.4 zettabytes (1 zettabyte = 1 trillion gigabytes = 10^{21} bytes). In 2020, that number has grown tenfold (44 zettabytes).
- **An ever-increasing variety of data types and formats.** In most organizations, data comes in the form of flat files, document stores, event and process logs, tagged files, graph and relational databases, Hadoop filesystems, and constant data streams from sensors of all kinds.
- **Rapidly expanding streaming data:** The IoT can't help but increase total data volume by orders of magnitude through low-latency streaming. By 2025, IDC says the IoT should create nearly 80 zettabytes of data.
- **Locations and repositories in edge-to-core-to-cloud infrastructures:** TechJury says by the end of 2020, two-thirds of enterprise infrastructures will be cloud-based, plus 82 percent of workloads (`https://techjury.net/blog/cloud-computing-statistics`). Applications are spreading as more organizations turn to data-intensive models and code. Data resides across this entire complex landscape.

Trends in data consumption

With increasing amounts of data to work on, organizations are finding more interesting and innovative things to do with ever-increasing collections of bits and bytes.

Artificial intelligence and machine learning (AI/ML) require special collections of data called "training sets." They're designed to let AI and ML algorithms find meaningful patterns, establish rules, and make data inferences. Once trained, AI/ML is turned loose on real-world data to help organizations find value, improve customer experiences, and innovate more quickly and effectively. As the world changes, algorithms must be constantly re-trained, and new training sets created. However, building upon success with such tools and techniques, companies find their appetites for data and complex analytics increasing dramatically.

Self-service is an emerging trend in organizations where a data fabric is available. Such users (a) know what they're doing, (b) understand the data they're using, and (c) find new and valuable ways to put data to work themselves. Once they truly understand what self-service means, such users find an astonishing number of ways to use it. This democratization of data invariably causes data usage to spike.

Challenges in creating data fabrics

At the same time that sources, uses, and locations for data are exploding, complexities in managing and controlling data are growing rapidly. There's a constant ferment in IT as new technologies, new kinds of data, and new platforms to run and store that data upon keep appearing constantly.

Thanks to cloud and emerging edge-to-core-to-cloud IT infrastructures to integrate and support cloud-based and cloud-linked applications, data is also increasingly distributed across in-house and in-cloud storage facilities and repositories. Most data in various forms is stored across the entire IT landscape.

Work becomes fragmented and incoherent as organizations try to cope with all this complexity. They must understand how to move data, where to store it, how to protect it, and how to make it accessible to authorized users. Sadly, individual practices and processes follow along with specific platforms and applications. This can stymie forming a holistic big-picture view of data. Worse, a fragmented approach makes it nearly impossible to achieve uniformity in data-handling processes, policies and governance.

Traditional "if you build it, they will come" approaches no longer work for data management. Data requirements keep changing constantly, not least owing to fast adoption of ML-based

solutions. A fundamentally different approach to delivery, known as DataOps, is needed. It puts people who understand data and its technologies (data professionals and developers) together with professionals responsible for making IT operate and meet business needs (operations professionals). Thus, DataOps is a conflation of those two worlds, where CI/CD is its mantra. CI = continuous integration, which means all players collaborate to optimize and improve how data is acquired, stored, processed, and used. CD = continuous delivery, which means all players collaborate to keep delivering and deploying new pipelines, analytics, and solutions.

Any data fabric also has an important human side. Applications come with stakeholders, users, and developers, all of whom must understand and buy into that fabric. Otherwise, organizations risk fighting "shadow IT" and other end-arounds that favor "quick and dirty" over consistent and coherent. Chaos is easier to corral when all troops march together.

WARNING

As innovation accelerates, changing and bolstering data management methods becomes increasingly difficult and disruptive. A fragmented, siloed approach is unsustainable.

Grappling with Modernization

A data fabric presents a consolidated data management environment across an organization's edge-to-core-to-cloud infrastructure for all its platforms and applications. Although not a static, single technology stack, conceptually a data fabric provides a consistent view into and a set of controls for managing disparate data and divergent technologies deployed across multiple data centers and edge computing locations, both in multiple clouds and on the organization's premises.

Data modernization is the process of consolidating and aggregating existing fragmented and siloed approaches to managing, storing, and situating data, to be governed and consumed as a single data fabric. That said, any data fabric is too broad and fluid for a single technology or tool to accommodate quickly. Also, it always evolves over time.

But a consistent approach must apply to the organization's data, its data intake and processing facilities, software platforms (in-house and in multiple clouds), and execution environments to cover all the bases.

Modernization also means adapting legacy (and sometimes current) applications to incorporate in and across multiple clouds, both private and public, as well as processing at the network edge. At the same time, organizations must adopt and create new applications and analytics to absorb the deluge of new data so as to transform and improve business operations. Obtaining buy-in from stakeholders, users and developers is much simpler when "what's in it for them" is tangible and valuable. These benefits usually include increased business volume, improved customer satisfaction and relationships, and new business opportunities and value-adds.

TIP

Modernization also involves efforts to address fundamental questions related to how an organization manages its data:

- How must data management change to address a data landscape that includes edge, core, and multiple clouds?
- How can data be made portable so it can reside and be processed where it delivers acceptable performance at the lowest overall cost?
- How can applications be re-engineered or refactored to move data management tasks outside them?
- How can the organization meet an ever-increasing number of governance and compliance requirements systematically and consistently?
- How can applications be built to run where it makes most sense in terms of cost, performance, and availability?
- Last but not least: What changes does it require to culture, processes, and operations metrics?

In today's world, data modernization also incorporates AI and ML into a "smart data fabric" that can self-monitor and self-optimize to help deal with issues inherent to complex, multi-factor situations. At the same time, a smart data fabric also uses automation wherever possible to reduce manual effort and to accelerate and enhance data management processes.

Doing Modernization Right

Given all the requirements that motivate modernization and the questions it can help to answer, you're no doubt curious as to the nature of the effort involved in making modernization happen. Four key forms of activity rule the engineering side of data modernization, all of which any organization will undertake to some degree or another:

- **Rehost:** This keeps data systems and applications more or less intact, and does a "lift-and-shift" to move them from their current runtime environment onto a public cloud such as AWS, Azure, or the Google Cloud Platform (GCP). Call it "old wine, old bottles, new cellar."
- **Refactor:** This takes existing data systems and applications and reworks them to permit them to accommodate and incorporate modern, emerging technologies without rebuilding them completely. This is a case of "old wine in new bottles."
- **Rearchitect:** This recasts old data systems and applications in "digital native" forms and systems to replace existing tools with newer, better, faster equivalent applications. It may be "new wine in new bottles," but we're still talking about wine and bottles.
- **New and different:** This takes old data systems and applications, tosses them out, and builds new ones. It may involve switching to a cloud data lake to replace legacy on-premises databases and data warehouses. It may involve consolidating multiple data puddles into a single, huge cloud data lake. It may bypass old Hadoop-based technologies with new, more flexible (and affordable) cloud-based alternatives. No more wine, no more bottles.

There are many ways to proceed and many options to consider. Most organizations will start with a pilot project or two and see how things go. Then they can learn by working through refactoring and rearchitecting before they tackle something new, completely different — and awesome!

The Data Fabric Brings Relief

Organizations face many challenges in adopting a data fabric. Working through modernization is serious and non-trivial. Those things said, they should take heart from the assertion that the data fabric brings worthwhile change and numerous benefits to those who buy into the vision. Here are some benefits that can help lighten the load and ease the task:

- Using a data fabric, organizations can find, access, and combine data from all available sources, regardless of data type and location.
- A data fabric works with all data types with ease and facility. Structured, unstructured, and semi-structured are all good.
- The data fabric can provide the agility, speed, scale, and reliability needed for enterprise-grade data systems.
- A data fabric can accommodate multiple instances of all kinds of execution environments, including on-premises data centers, cloud platforms, and edge systems.
- A data fabric can process and provision data at all velocities from streaming data in real-time to scheduled batch jobs (regularly or infrequently, large and small, fast and slow).
- The data fabric provides a consistent, coherent view of, metadata for, and controls over data to meet organizational security, privacy, and compliance needs.
- The data fabric is open ended and extensible, so it can support multiple processing engines, tools, and platforms as technology changes and evolves.
- The data fabric can move data from platform to platform for ready consumption or storage and archiving, no extensive refactoring needed.
- The data fabric can move processing from one execution environment to another without extensive recoding.

In short, a data fabric is a powerful abstraction that covers the whole lifecycle for data — from intake, to enrichment, to application delivery, to storage and archiving, to retirement or deletion — within a single, consistent, policy-driven platform. Aren't you just dying to have one of your very own?

IN THIS CHAPTER

- Understanding DataOps
- Realizing the considerable benefits of DataOps
- Instituting DataOps and improving DataOps practice

Chapter 2
Finding Value in DataOps

There's no doubt that the buzzword *DevOps* — a conflation of the terms "Development" and "Operations" — is all over the IT landscape. It provides the inspiration for a similar and parallel term in the title of this chapter.

Before you can understand *DataOps,* then, you need an explanation of *DevOps.* Simply put, *DevOps* is a set of practices wherein software development (Dev) combines with IT operations (Ops). In particular, DevOps seeks to speed up the development lifecycle and to provide continuous delivery (CD) with continuous integration (CI), driving continuous improvements to deliver more business value faster. This approach builds on Agile software development concepts and approaches. It is also where the notion of CI/CD originates.

For more information on DataOps terms and concepts, visit this page:

```
https://www.hitachivantara.com/en-us/insights/dataops-
   insights/dataops.html
```

Understanding DataOps

DataOps is like DevOps in that it combines two things — data management and IT and business operations — with the DevOps idea of CI/CD included for good measure. In DataOps, the idea is to deliver, in the words of 451 Research (a data-oriented research firm and consultancy) "more agile and automated approaches to data management."

Find this report online at:

```
https://www.ciosummits.com/Online_Assets_Hitachi_Vantara_
   DataOps_Unlocks_Value_of_Data.pdf
```

Because it enables repeatability, speed, and quality, automation is essential to DevOps. Automation is even more important for DataOps because it provides an ability to scale ridiculously. Automation helps in DataOps not only because it saves on human time and effort (as with all automation), but because there's so darn much data that without automation, it might more than mere humans could handle on their own.

Even better, automation works hand-in-glove with artificial intelligence (AI) and machine learning (ML) to groom data for ingestion and to use automated semantic enrichment techniques to improve data quality and value. Thus, automation lets DataOps take organizations where they simply could not go without the speed, power, enhancements, and accuracy it delivers.

Measurement is equally important. To understand the efficiency and effectiveness of data processes, organizations must constantly measure and monitor data quality, usage, movement, access, and more.

In fact, DataOps principles are vital to organizations' data integration and governance efforts. DataOps accelerates the delivery of new data for analytics efforts, improves data quality, and increases trust in data. It even reduces the cost of data management itself.

TIP

One vital key to proper DataOps is a curated catalog of internal and external data. Access to such a catalog supports more and better uses of the data it represents. It also enables the holistic, coherent view of the organization's data that a data fabric is supposed to deliver and ensure. And indeed, the catalog works "behind the scenes" to provide semantic enrichment and data validation.

Realizing the Benefits of DataOps

By putting automation and measurement to work in data operations processes, organizations gain considerable value from that data. Because data can be validated and checked as part of those processes, data made available for analysis and use is more or less guaranteed to be of higher quality.

Where semantic enrichments come into play, that data will also be easier to identify (and hence, also easier to find and use). In fact, semantic enrichment means more usable and useful data that fits extremely well into self-service scenarios where users decide what data they want and what questions to ask of it. As users and teams share data, new meaning gets added, which in turn makes such data more valuable still. Collective enrichment through "tribal knowledge" makes data truly worth treasuring.

Once data is incorporated into a data fabric, it can be situated to give users best access and performance. It doesn't matter if those users are humans building machine learning models or running ad hoc queries, applications consuming data, or analytics or models using data to produce insights or business information. All benefit from faster, more efficient handling.

In addition, DataOps helps make data (and data access) more resilient. Given smart data operations at enterprise scale and using automation (both rule-based and AI-driven), it's easy to position (and re-position) workloads and data to improve performance, or to work around network delays or access issues. This helps businesses leverage their data more fully to ensure successful outcomes.

Understanding Key Tenets of DataOps

Although numerous principles from DevOps carry over into DataOps, these principles are hallmarks of a DataOps approach to data management and data fabric:

- **Collaboration and cooperation:** All players in DataOps work together and form a single understanding of what data and its management means, how it works, and how it supports the data fabric. This includes people from IT and cloud operations departments, data and database professionals from all over, and business stakeholders responsible for applications and services within the organization (and the data and services they consume). Creating a data fabric is as much an exercise in teaching people new ways to think about, analyze, and use data, as it is about choosing and deploying tools and platforms to manage data across its natural lifecycle.

  
  WARNING

  Organizations seeking to create a data fabric ignore the human element only at great peril. Nothing sinks a data fabric initiative faster than unhappy participants who don't understand and buy into the data fabric vision.

- **Heavy automation:** A constant from the operations side, automation ensures that data intake and ingestion produce accurate, properly tagged and labeled, and (where possible) enriched and enhanced data for users, applications, and services to consume. Automation handles initial placement of data and policies for movement, manages active data for performance and other optimizations, directs traffic between online and cold storage (based on activity and usage needs), and much more.
- **CI/CD:** DataOps handles the ongoing influx of data that occurs 24/7/365 through automation and careful monitoring of data fabric resources and data placement. This also permits DataOps to deal with ever-increasing data volumes, as new sources emerge to produce it, and new applications and services appear to consume it. Thus DataOps never stops, always at work to ingest, enhance, deliver, and protect organizational data. In data terms, this is analogous to continuous integration/continuous delivery (CI/CD) in DevOps.

- **Client focus and continuous innovation:** Part of the Agile inspiration for DevOps and DataOps, and a key aspect of its widening uptake and use, is a relentless focus on serving clients who use data in a fabric. Continuous improvement is about doing things better all the time, incrementally, based on objective measurements and subjective priorities and selection criteria. DataOps makes room for both. Continuous innovation speaks to ongoing incorporation and use of emerging tools, technologies, and data sources as part of an ongoing CI/CD process. Data never stops arriving and never stops being ingested and used; DataOps never stops, either.

Building and using a data fabric requires developing a DataOps philosophy, outlook, and practice, along with supporting tools and technologies. The following sections speak directly to this process.

Instituting DataOps

Putting DataOps to work in an organization includes important technologies such as automation, AI/ML, and policy-driven tools and platforms. But it is also very much about working with people. Cultivating DataOps in many organizations often comes as an outgrowth of DevOps. But it requires an understanding of the organization's culture, its policies (be they automated, codified, or implicit in the culture itself), and processes (ditto) to understand what must change to achieve human buy-in and active support from all players.

To a large extent, then, instituting DataOps is also about creating and fostering a culture of comfortable collaboration among key constituencies within an organization. This means getting IT staff, data and database professionals from across the organization, and business stakeholders who control the resources and "own" the data, to interact. At best, each will understand the "other side's" views, priorities, and goals.

Another vital aspect in instituting a workable and sustainable DataOps environment is to create proper, automated workflows. This emerges naturally when comfortable collaboration works among the players. The IT staff naturally wants to make sure the other players get what they need from systems and services. The data and database professionals want to make sure data is properly

tagged and labeled, accessible to the right users (and otherwise a blank slate), and in line with compliance and governance policies and requirements. And stakeholders want to make sure they get the best and most value from their data, as well as the applications and services that consume them. Together, all parties want to succeed by helping each other meet goals and objectives, too. This produces automated workflows that do what needs doing.

In a DataOps environment, organizations also work hard to make the most of their *metadata.* This term means "data about data" and addresses how data is tagged and labeled, relationships between and among data items, compliance and governance requirements for the data, and even patterns and relationships in the data that emerge from AI/ML augmentation. Data about data also indicates where and how it should be housed, when it should be moved (or re-positioned), and when it can be archived or moved to inactive status and storage. All this metadata keeps changing and growing (which is what makes it active) and keeps contributing to improved data accuracy, quality, and availability.

All around the organization implementing DataOps also means adopting the Agile philosophy of continuous integration and continuous delivery (CI/CD). Beyond managing data and its active metadata, this also means automating and improving existing applications and services. It also means seeking out, adopting and incorporating new tools and platforms, applications and services, and data sources as they become available. It's like the old Beatles lyric: "getting better all the time."

REMEMBER

Active metadata is best understood as data about data that is open-ended, flexible, and extensible. Not only does its scope and coverage change over time, it usually becomes richer, more nuanced, more detailed, and more descriptive.

IN THIS CHAPTER

- Onboarding and ingesting data
- Optimizing data for cost and performance
- Putting policy to work in the data fabric

Chapter 3
Data Onboarding, Placement, and Processing

Acquiring, grooming, and enriching, then provisioning and governing data are all important aspects in building and using a data fabric effectively. All these topics provide the focus for the ensuing discussion in this chapter.

Understanding the Basics

When data shows up at your organization's door, you can't let it in unhindered or unchecked. If handled properly, especially within the context of a data fabric, new and incoming data goes through an onboarding process. The ultimate output of the onboarding process is what every organization wants — namely, high-quality data with all tagging, labeling, and metadata that applications, services, and the data fabric need to know what data is made of and everywhere it lives.

Incoming data has a way of showing up in different formats, in drib and drabs, and sometimes even with unhealthy attachments

(or at least, dependencies) on certain applications or services. Only when data is checked and validated can it be consumed in queries, applications, and so on. Other steps follow soon after, as explained next.

Data Onboarding/Ingestion

One important part of the onboarding process involves establishing a connection between the data fabric and the data source. Data has to come from somewhere, right?


TIP

Hooking data sources and a data repository together often involves so-called "cloud or software connectors." These programs hook up with applications, databases, services, and so on, to export their data as input to an onboarding process. Connectors might be put to work anywhere from edge to core to cloud to begin the onboarding process, and to permit consolidation and blending of data from multiple sources. Overall, the idea is to create a holistic and complete view of the data, and thus to produce deeper insights and more accurate analytics.

Organizations can develop and reuse custom connectors, too, as their needs dictate. These include the following:

- **Pre-fabricated, packaged connectors:** These plug into a range of well-known and well-documented data stores to extract their data. They're usually maintained to remain flexible and insulated from changes on the sending side. Thus, onboarding can continue sans difficulty or delay (or extra recoding).
- **Operational technology (OT) and IT sources:** The former represents input from Internet of Things (IoT) devices (sensors, surveillance cameras, and more). The latter represents relational databases, big data stores (on-premises or in the cloud), and enterprise applications. For both kinds, data fabric users can often draw on a rich library of ready-to-run components to prepare and blend incoming data during the onboarding process.
- **SDKs and purchased data:** Specific tools or platforms may include a fee-based set of building blocks to create custom connectors for all kinds of new and legacy applications. These work with the tool's or platform's programming

interfaces (APIs) and software development kits (SDKs) to support custom connectors. Many vendors offer training, developer's guides, consulting services, and examples to enable custom connector development.

Stages: Analyze, enrich, and store

As data is ingested, it is first analyzed to see what it's made of. This is the stage at which duplicates get resolved and eliminated, and when data validation and consolidation occur. Some experts like to call this a *cleaning* or *grooming* process. It results in data that's vetted and checked, labeled and tagged from input sources, labeled and tagged with metadata and other related information, and deduplicated.

But there's more to do with incoming data than cleanup and consolidation. The overall way data moves through an organization involves intake, sanitization, labeling, and enrichment, and finally, incorporation into the data fabric for consumption. This occurs as part of data pipeline management, which uses the metaphor of the paths and means that move data through onboarding, processing, and into use to describe data handling and management across the lifecycle.


TIP

Organizations can add considerable value through semantic enrichment. That is, data gains value to the extent it can be properly labeled and tagged. This may involve using multiple, related data sources and views, and combining them to provide more information than the metadata available from any single source. The impetus is to enrich data with as much information as possible to add to and further clarify what's already there. In the next section, I dig a little deeper into a specific enrichment technique that leans on specialized tools and technologies to handle enrichment based on models, analytics, and related insights.

Once data is onboarded, sanitized, and enriched, it's finally ready to take up full-time residence in the data fabric. That also means it's ready to be used: consumed by services and applications, subject to ad hoc queries, and used in big data models and analytics. A data fabric needs to determine where to house this data, and how best to make it available for consumption while minimizing storage costs and delivering the best possible user experience. Artificial intelligence and machine learning (AI/ML) plays a role in this, too.

Access to AI for identification and tagging

In a modern data fabric, incoming and existing data may be subject to complex analysis and pattern recognition. AI/ML excels at juggling more factors and variables than humans can handle. It can build models and find patterns that humans can't see easily (or at all). This allows data to be more readily identified and tagged using automation, with more and better metadata than would be possible manually, or using a simpler, rules-based approach.

Thus, a data fabric uses what it already knows about data it already has under its purview and control to add insight, identification, and tags to new data coming in. As the volume, types, and relationships among data items the fabric handles grow in number and in kind, it can infer ever more about the universe of data it houses.

REMEMBER

So it is that AI/ML does more than simply add value to new data going through onboarding. It can also add value to existing data already incorporated into the fabric, based on what it's seeing and learning from the new data constantly entering that fabric. This is a surprising and valuable side effect of the constant improvement and constant innovation that DataOps in the fabric runs unceasingly and relentlessly.

Automating avoids wasteful effort

Within the data fabric, dataflow templates make it easier to build and manage data onboarding routines across the many data sources that provide input. Organizations can use templates and automation to create repeatable and reusable workflows. They can also replace dedicated, application- or service-specific (and dependent) pipelines with more standard and flexible general-purpose pipelines.

An automated, template-driven approach helps onboard data quickly and accurately into Hadoop and other platform-specific data stores. Some templates include metadata placeholders set to make sure that ingested data is ready to use as soon as it takes up residence in such a store. And when metadata is included in a template, at runtime the same template can be reused for many different pipelines. This reduces duplication, as well as the overhead for managing thousands of similar pipelines with only minor differences.

TIP

Dataflow-oriented templates may include tools to spot-check data transiting the onboarding process. Such checks often feature analytics like charts, visualizations, and reporting. These are usually accessible during any data preparation step in a pipeline, so the DataOps team can check data during ingestion and quickly resolve quality or accuracy issues.

Optimizing for Cost and Performance

A modern data fabric can also help organizations manage costs while offering improved performance. Storage optimization in the data fabric can help organizations decouple storage and compute usage and costs to improve performance and utilization while reducing storage costs.

For Big Data environments such as Hadoop, Splunk, backup applications, and others, optimization tools show organizations what they're spending on compute and storage costs. These tools can identify opportunities for savings. An example appears in Figure 3-1, and Hitachi Vantara's tool is available at `https://apps.hitachivantara.com/hadoop-tco-calculator/`.

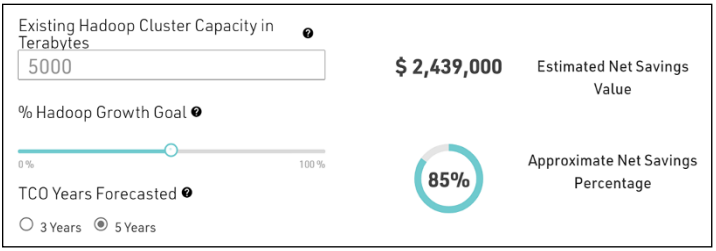


FIGURE 3-1: Various calculators show savings available from using dynamic storage tiering.

TIP

A thorough review of Big Data storage often shows excess costs for inactive or underused data in expensive cloud stores. Certain tools and technologies can actively manage migration of such data between more expensive, live, in-the-cloud storage, and cheaper, near-line, archival storage to provide data when and where it's needed, while saving storage costs.

By separating compute and storage resources, organizations can scale and manage them more efficiently. A well-built data fabric solution can keep track of where resources reside and how they're

used. It can implement a data-tiering approach to match the current type and residence for active use (in the cloud) or inactive and unused (off-line or near-line), as current needs and usage dictate. This approach could cause data to move, for example, from the Hadoop Distributed File System (HDFS) on-premises to an internally owned and managed on-premises object store or a cloud-based S3 object store target without requiring changes to analytics applications.

A similar approach within the data fabric applies to analyzing workloads for placement and execution. A well-built data fabric environment is flexible and accommodating enough so that the right data is managed in the right place at the right time.

Giving Policy a Key Role

The overarching nature of a data fabric makes it ideal for incorporating, accommodating, and enforcing policy and governance on the organization's data. This means that data professionals can incorporate compliance and governance requirements into data fabric policy. Thus, such things are baked into the data management and handling environment so they can neither be ignored nor abandoned. This provides an excellent way for organizations to mitigate substantial legal, financial, and reputation risks they might otherwise incur.

TIP

Policy is a multi-faceted instrument, though, and can do more than risk management. Policy can also establish and enforce data fabric behaviors related to cost management, performance optimization, resource utilization, and more. Thus, using a data fabric improves operational agility because it allows organizations to establish and maintain a set of policies that controls how data is consumed, where it's housed, who has access, and how it gets positioned for immediate use.

Better yet, policy provides a mechanism to let a data fabric operate effectively in an edge-to-core-to-cloud architecture. By dictating how to place workloads based on user location, performance requirements, job priorities, and so on, policy sets terms for how workloads and their data run within an organization's IT infrastructure. And because such policies can be self-adjusting and self-correcting, they can respond to changing conditions and accommodate new sources of data — along with applications and services — to consume them far more easily.

IN THIS CHAPTER

- Understanding the role and vital importance of governance
- Preparing data properly simplifies compliance
- Doing data discovery right
- Handling sensitive data

Chapter 4
Governance, Semantic Enrichment, and Self-Service

A good working understanding of what data represents — and how it must be secured and handled to meet compliance requirements — is essential for organizations with massive data collections. Some of the most important metadata created during semantic enrichment deals with compliance and governance. Finally, the ability for data professionals, business analysts, and other responsible members of the organization to request and create their own uses for its data gives the data fabric an unmatched ability to foster innovation, creativity, and business success. These topics are fodder for this chapter, so pony up to the trough.

Understanding the Basics

Practically speaking, there are two ways that governance and compliance gain great support from a data fabric. To begin with, because data collections are already huge, and growing daily, data fabric helps cope with that increasing data volume and velocity. At the same time, ongoing policy enforcement makes sure

compliance and governance matters are addressed. More importantly, legal and regulatory requirements for privacy, confidentiality, and even "the right to be forgotten" all attach to data: These can pose financial, intellectual property, and reputational risks to organizations that don't follow the rules (and sometimes, jail time for certain corporate officers held responsible).

But data governance goes beyond regulatory compliance requirements. It should include data governance to protect an organization's trade secrets, and to safeguard its brand and reputation. It isn't enough to be able to withstand and pass compliance audits. What's more desirable is to make sure that data is safe and protected, and that access is limited to those with a right and a need to know. Above all, data should be available to those who need it and are authorized to see it, so that manual checks and validations don't hamper genuine productivity or business agility. It's a tricky balance, but one that must be defined, maintained, and constantly adjusted to keep up with changing times, technologies, and data.

Beyond governance, semantic enrichment is primarily about analyzing data to add or update meaning (metadata) to the content in a data fabric. Interestingly, the data fabric can use what it already knows about existing data under its purview and control to enrich incoming data during the onboarding process. Likewise, the fabric can use what it learns from incoming data during that process to enrich existing data as well. Semantic enrichment can even use artificial intelligence (AI) and machine learning (ML) to better identify, tag, and label its data (both incoming and existing) as new insights and patterns emerge from its ongoing and never-ending analysis of its own data holdings.

REMEMBER

It's hard to overstate the value of self-service from the data fabric for any organization. Here, self-service means that users with the right access to the fabric — especially its catalog, analytics tools, and data holdings — can find data, create and run their own analyses, request canned analyses for the data they select, and hook up applications and data of their choosing to make things happen by themselves. No filing job requests with IT and waiting for your turn to come before work gets going! It's a dream come true that opens the door to experimentation and innovation.

Key areas for governance

Laws, rules, and regulations about and for data abound. In the United States, that includes a slew of acronyms such as HIPAA,

PCI, SOX, FCRA, FACTA, COPPA, GLBA, CCPA, and FERPA. In the European Union (EU), the General Data Protection Regulation (GDPR) — which includes a "right to be forgotten" that means companies holding data must find and destroy it at the owner's request — has affected companies around the globe (at least, those wishing to conduct business in the EU).

In broad and general terms, key areas for data governance fall under these headings:

- **Data privacy** is sometimes called *information privacy.* It entails a necessity to preserve and protect personal information collected by any organization from access by any unauthorized third party. Privacy often involves access controls to explicitly manage (and if necessary, audit) access.
- **Data confidentiality** is the practice of keeping private information secret. Thus, confidentiality may involve encryption so that only those with authorized access to decryption keys can view or use the data.
- **Data protection** means safeguarding data from corruption, compromise, or loss.
- **Audit and logging** involve mechanisms to keep track of and record uses of and access to data, including its creation, modification, and deletion. Some governance and compliance regimes require organizations be able to document access and uses. Logs and audit trails are particularly relevant and may be used as evidence to prove or disprove data misuse or malfeasance.

WARNING

Today's compliance and governance regimes require that organizations be prepared for hostile inspection at any time. Thus, tagging data with privacy, confidentiality, protection, and audit/logging metadata is vital. In turn, that means the data fabric can help keep track of such things and report on them as needed. In fact, compliance and governance metadata is essential because all organizations must acquire, manage, and protect private personal data of one kind or another (employees, customers, partners, suppliers, and so on). Beyond audit and compliance concerns, organizations must also do what they can to protect their intellectual property and to preserve their brands and reputation. All of these things require keeping an eye on the data and keeping it safe from unwanted or unauthorized access and disclosure.

Data preparation provides compliance insight

The onboarding and intake process is an ideal time during which policy may be applied to data, including compliance and governance metadata. This means organizations can address privacy, protection, and confidentiality concerns from the get-go. It also means they can produce an audit trail for data items that must be tracked and documented as soon as they enter the data fabric.

Data validation also involves checking accuracy, veracity, and currency of data. Organizations may use third-party data marts or markets to double-check private personal data and to make sure what they see can be confirmed or corrected as needed. This helps ensure data protection by ensuring data integrity.

Data enrichment includes adding metadata from other sources and for purposes such as meeting organizational policy requirements. Because an oversimplification of some such directions might be stated as "follow the law, rules, and regulations," the enrichment process inevitably will include and address metadata to cover governance and compliance matters. From a risk management perspective, one might argue that this kind of data is the most important and valuable of all, because it helps reduce exposure to fines, penalties, legal action, theft, and reputation damage.

REMEMBER

The real nuts and bolts when onboarding and incorporating data items into a data fabric center around data and value identification and tagging. This is when data items acquire type and format descriptions, along with other metadata to address related policies. This may include adding governance, compliance, and security tags (flagging data as private, confidential, or proprietary), applying a security classification (public, secret, top secret, need to know, and so on), and more.

Doing Data Discovery Right

Using data discovery is an important ability for a data fabric. Here's why. Data discovery involves collecting data from across all sources — including databases, data lakes, data stores, file repositories, streaming sources, applications, and more. It also

means consolidating all that data under a single data catalog and platform so that it is easy to find, use, and even move as needs and uses dictate.


REMEMBER

Without data discovery feeding a single master catalog so that one platform can manage and orchestrate an organization's data, there is no data fabric. Data discovery and onboarding are keystones by which a data fabric stands — or falls!

Done right, data discovery involves human expertise and effort as well as automation and computing tools. Humans are best able to deal with setting things up and providing initial sets of training data for AI/ML data discovery when complex analysis is called for. Humans are great at formulating and applying rules for identifying and tagging data when complex analysis is not needed. They are especially good at deciding how policy should be formulated and expressed, and when it should apply to certain types of data.

Automation and programming come into play once templates are established, rules are defined and applied, and policy decisions handled. AI/ML can infer complex relationships that humans may not recognize. Automation can handle routine tagging and labeling in enormous volumes, at great speed, with untiring and predictable accuracy. Programming through export tools and APIs to access incoming data can include rules, enforce policy, identify data, and apply related tags.


TIP

Proper use of data fingerprinting can employ hundreds of features to correctly tag and classify data. Proper use of machine learning means that automated tagging and semantic enrichment improve over time, as what's learned from incoming data and curation efforts boosts overall data quality and value.

Applying Special Rules to Sensitive Data

Onboarding and discovery provide an all-important opportunity to identify and tag sensitive data. During either process, as data items are identified, this should include tagging items as private, confidential, or sensitive. At the same time, other metadata can identify associated compliance, governance, and security requirements.

Such requirements cover a broad range of actions and tags, but will include some or all of the following (and more):

- Enable auditing and or logging of access and use
- Flag data as personally identifiable information (PII, a well-understood data category for privacy regimes)
- Associate with one or more named governance or compliance regimes (HIPAA, PCI-DSS, and so on)
- Classify data within a security classification scheme
- Turn on encryption and/or pseudonymization

As a matter of policy and typical operation, the tools and platforms used within any data fabric will apply and enforce access controls on the catalog, data items, and tools under their purview and control. This must be considered as an equally important aspect of managing and controlling access to data, especially sensitive data.

How a Data Fabric Improves Governance

The most important way in which a data fabric improves governance is by baking related rules and requirements right into the data items it manages. Having an enterprise-wide data catalog that provides end-to-end visibility to data also makes it more accessible to authorized parties, and easier to govern. But by making sure that policy applies to all of its data items, the data fabric sets the stage to address governance and compliance matters and concerns. The next frontier for organizations to cross is in defining and applying automated policies and rules to address compliance and governance matters.


TIP

Properly implemented, a data fabric embeds the rules and requirements for compliance and governance in all of its actions and activities. Then, as the data fabric discovers, onboards, uses, stores, and ultimately, deletes, and destroys old or unwanted data items, applicable policies at every stage make sure that privacy, confidentiality, and protection are part of each phase and action that occurs.

IN THIS CHAPTER

- Empowering innovation with self-service
- Embedding analytics into business processes
- Placing and archiving data intelligently
- Realizing true multicloud flexibility

Chapter 5
Premier Data Fabric Use Cases

Data fabric is of no use in and of itself. Its value is exclusively created through use of data to drive business outcomes. Hence, use cases bring needs and outcomes into sharper focus. By definition, a *use case* depicts or describes a specific situation in which a product or service can be put to work. The use cases in this chapter showcase some of the important and valuable things a data fabric can do for an organization. Each organization must work through and document its own important use cases to design and build a data fabric architecture.

Gaining Incredible Value from Self-Service

Within a more traditional organization, departments and business functions work more or less independently. In terms of designing, creating, and using systems and applications, this translates into a situation that keeps departments and functions at arm's length. For example, the process to obtain a new analytical tool for marketing staff to analyze market potential for proposed products is very time-consuming with numerous process steps between various business and IT teams.

What's missing from these interactions is the amount of time it takes to get from one step to the next. In a traditional organization, any step in the sequence may take days to weeks. The entire process can easily take several months. Alas, once the application is deployed, it may not be what the user wanted, or requirements may have changed in the interim.

This approach reflects a slow, deliberate and outmoded approach to development and delivery of software and tools, especially for analysis. It also explains the impetus so often found — in sales and marketing units especially, where time is of the essence — to hire a consultant instead, pay them big bucks to do it *now*, and work around, instead of with, the IT department. That likelihood increases with the length of the IT backlog and how long it takes to get their time and attention. In a nutshell, this explains the phenomenon known as "shadow IT" — other departments taking on and taking over what IT usually does, outside IT's normal procurement channels, processes, procedures, and policies. This doesn't happen for no reason.

Self-service to the rescue!

With a data fabric, analysts in marketing (and other departments not named "IT") can take advantage of self-service tools to create analyses — including one-offs and recurring items — for themselves.

Instead of waiting for IT, data analysts and professionals in other departments turn to the data fabric's catalog to explore data items of interest. Using efficient metadata-based search capabilities, they can quickly and easily assemble a collection of data items they wish to explore, investigate, and analyze.

Next, sales or marketing analysts can turn to a set of visual tools to design data pipelines and processes. These include a variety of analytical items to operate on their self-selected datasets and various graphs, charts, and other formats to display results. They can tweak and tune the data, tools, and visualizations until they get things "just right." Presto! Self-service at work, and another satisfied user gets things done as quickly as they can, with no need for IT involvement.

Building and managing models

Some of the most interesting action in today's modern infrastructure comes from the insights that artificial intelligence (AI) and machine learning (ML) bring to light (together, AI/ML). There's work involved in using these technologies properly, so here's a short detour through the data science lifecycle as it applies to ML.

ML starts with a model of how things work, and uses learning algorithms to analyze data handed to it. Such learning cycles back into the model to improve accuracy and representational ability. Thus, ML uses step-by-step instructions for handling data (like normal programs). However, its algorithms compare results obtained from using various methods and making specific choices. Over time, repeated choices of methods and options that produce the best results cause the model to evolve and improve, and its results to do likewise.

ML starts out with carefully selected sets of data called *training data* before it gets turned loose on real, live input. Analysts from a targeted department (sales, marketing, and so on) work with data professionals to create training datasets to start the ML process. Then, they examine early results to make sure models produce sensible and usable results. As you might expect, a certain amount of tweaking and fiddling is required to get training datasets going (and staying) in the right direction. Once training is far enough along, ML may then take real data as input. Here again, the analysts and data professionals make sure that the model produces interesting and meaningful results, so further twiddling is inevitable.


TIP

Because real, live data tends to change with time, ML models must be tweaked constantly to keep them current and accurate. ML models are subject to data drift (where data values and clusters change over time) and schema drift (where data definitions and interpretations change over time). Constantly accounting for drifts keeps overall ML model accuracy high.

A data fabric helps analysts and data professionals throughout the data science lifecycle for ML. It helps them extract and manage "features," and possibly capture them in a feature store within the data fabric. It provides data and metadata used to set up initial training sets, and additional intelligence used to adjust them and the ML model as it goes into production.

In practice, AI/ML-derived insights informs a data fabric how to better organize and optimize its own holdings and metadata. As it turns out, such metadata may be used to drive rule-based automation as well as human comprehension. Thus, these same insights can also guide humans around the fabric for their own purposes (usually to create business value and improve business outcomes).

Embedding Analytics into Business Processes

Robotic process automation (RPA), is an approach that empowers users with tools to create digital robots that automate business processes. By adding analytics into existing software that implements or handles business processes, organizations can take advantage of the insights and information such analytics provide to increase efficiencies, eliminate errors, improve time-to-completion, and boost the quality and engagement of work for their employees and contractors.

RPA provides a set of methods and tools that individual users can employ in cookie-cutter fashion to stamp out lots of instances of specific analysis and reporting. This applies equally to frequent tasks that IT professionals automate to handle provisioning networks, servers, and user PCs. But it also works with workflows like procure-to-pay in purchasing operations, and for typical line-of-business applications in accounting such as accounts receivable (AR), accounts payable (AP), and even purchase order (PO) handling.

Tools available through RPA facilities include optical character recognition to translate images into meaningful text for analysis and data intake, mathematical tools to check transaction accuracy and reconcile accounts, statistical tools to detect anomalous or out-of-bounds transactions, and more. Analysts work with RPA to create automated tasks to update business process software using simple visual programming tools, and applications already at their disposal, without requiring input or servicing from the IT department. And of course, RPA capabilities work across the entire data fabric and explain how self-service works both easily and well.

Placing and Archiving Data Intelligently

With the onset of the General Data Protection Regulation (GDPR) on May 25, 2018, companies in (and doing business with citizens of) the EU found themselves required to provide all data they keep on customers or users on demand. They also found themselves subject to "the right to be forgotten." That is, under the GDPR any user or customer can request at any time that their accounts be closed, and all data stored about them deleted. Further, this law requires that information be provided — or forgotten — no more than 30 days after receipt of any such request.

A major European financial services provider found itself initially unable to comply with these provisions of the GDPR. Why? Because they had so many different and separate data stores, databases, document archives and so on, they could neither be sure that they'd found everything related to a particular customer account, nor be sure that all such information was in fact forgotten upon request.

Building a data fabric proved to be a workable remedy for their problem. By discovering all their data holdings, and putting everything in a global catalog, this company was able to find any and all data holdings for each and every customer. Thus, they could produce that information on demand, or forget it if asked to do so. They further combined a lot of their scattered storage archives into a single object store, effectively reducing the cost of petabytes of regulated data.

Another scenario for data retention is purely archival, and comes from legal departments everywhere. Under a court order, or as part of legal proceedings, certain data holdings and related logs and audit trails from the organization may need to be produced. These might relate to pending audits, regulatory or compliance actions, shareholder actions, or intellectual property and other lawsuits.


TIP

Rather than maintaining all such data in readily available, high-performance storage (in the cloud, or on-premises), an organization can simply take a snapshot of all of data it needs to satisfy legal hold and electronic discovery (aka e-discovery) requirements. Then, it can archive the snapshot in a near-line or offline data store. The production data fabric can change and evolve, as

use and circumstances dictate, without having to hang onto the contents of that snapshot. This offers legal protection, considerable convenience, and even potential cost savings (that increase over time).

Attaining Multicloud Flexibility

By far, the biggest value and advantage that a data fabric delivers to an organization is its profound platform- and location-agnostic flexibility. This reduces at a high level to "Any cloud, any storage, any analysis." What does this mean?

First, it means that the data fabric is tied to no specific cloud platform. Thus, workloads can run anywhere they might be needed, on clouds public or private, without requiring additional programming. A primary goal of the data fabric, after all, is to abstract the data and related applications and services from the platforms they run on.

Second, it means that data in the data fabric is tied to no specific data store nor even to any specific type of data storage. In fact, a data fabric may use tiered storage behind the scenes to match live cloud storage consumption with live data currently in use in the cloud (or likely to be used again soon, based on observed usage patterns and frequencies). In a more general sense, a data fabric can deploy and use storage based on specific criteria related to cost, performance, or compliance requirements instead of mere happenstance. This helps control costs while providing at least an acceptable experience for users — usually better than that.

Third, it means that data in the data fabric can accommodate any kind of analysis or analytics that the organization might want to run against its data. A data fabric offers an extensible, open-ended set of functions and pipelines for analytics and business processes. It also supports creation and maintenance of AI- and ML-based models and analyses. But because it is extensible and open-ended, a data fabric can also incorporate and accommodate new tools and technologies as they emerge.

All of these capabilities combine so that organizations can attain multicloud flexibility.

IN THIS CHAPTER

- Driving to fabric modernization
- Climbing the modernization curve
- Staying agile through constant activity
- Delivering the data fabric in many forms

Chapter 6
Modernizing the Data Fabric

Because most organizations already own and use data assets galore — including databases, data stores, data lakes, document repositories, streaming data, and more — adopting and deploying a data fabric means drawing all those assets under its umbrella.

REMEMBER

The data fabric is a consolidated data management environment that extends across an organization's edge-to-core-to-cloud infrastructure for all platforms and applications.

This invariably involves a modernization process. The idea is to take existing fragmented and siloed approaches to managing, storing and situating data into the data fabric's single, consistent, policy-driven, self-service environment, supported by DevOps and DataOps principles.

Modernization is a vital job for organizations seeking to deploy a data fabric. It involves nothing less than a review of all data assets relevant to a business case, and all applications and services that consume and produce data, with an eye to opening the data fabric's umbrella to cover everything.

What's Driving Fabric Modernization?

Given that business goals, needs, and objectives ultimately drive adoption and use of technology, it's important to understand what's behind the ongoing mad rush to institute data fabrics in today's marketplace. To that end, the following factors are at work to force modernization, whether organizations are ready and willing, or otherwise:

- Today's market conditions — especially financial crises arising from the pandemic — demand that any technology investments provide a faster return on investment. This puts added pressure on existing investments at the same time it encourages cautious and critical consideration of new one. Where the data fabric is concerned, this favors use cases at the high end of the volume and variety curve where bigger payoffs prevail.
- Given a fractured data landscape found currently in most organizations, it's important to understand that no one platform or solution cures all ills or addresses all problems. That is, one size does *not* fit all data assets nor analysis, storage, or governance and compliance needs. That's why extending a data fabric's umbrella over all existing assets relevant to a business case is where things must start, with a firm resolve to keep extending its coverage as new tools, technologies, and data sources enter the future picture.
- Amid all the time and effort involved in finding, deploying, and maintaining a data fabric, the real burning question has to be: Where and how to get it right? Since talent is scarce, more often than not this means partnering with a solution or service providers with the right expertise, tools, and technologies.


TIP

Organizations need to pick their data fabric partners carefully. Make sure they have the tools and technologies, and also the experience and expertise to help you acquire, deploy, customize, and maintain all the elements needed to extend a modern data fabric across your organization.

Climbing Modernization's Curve

When it comes to doing the work of modernization, organizations need to carefully examine their current data assets and holdings and decide how to get them underneath the data fabric. As I explain in the following sections, any number of strategies can serve that end. Figure 6-1 introduces these strategies with an illustrative diagram.

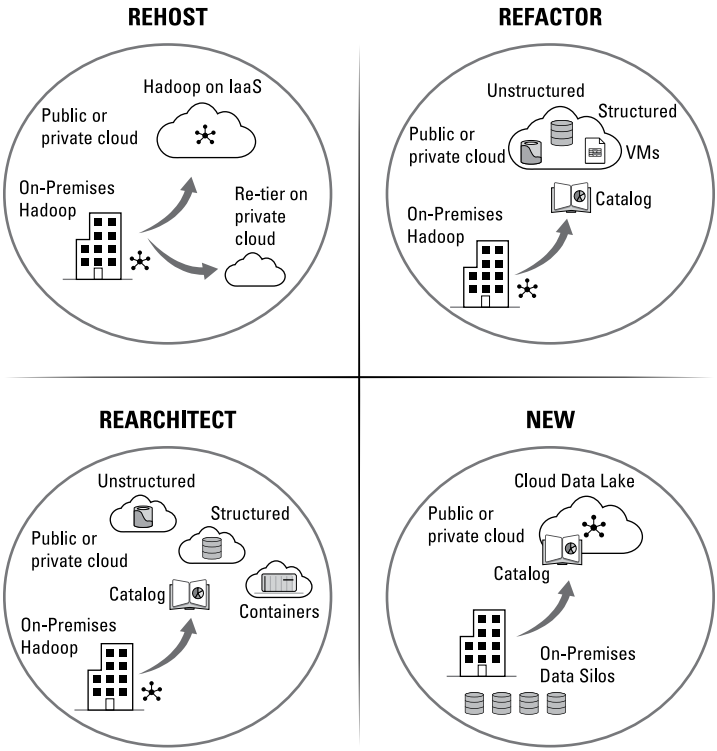


FIGURE 6-1: Four strategies to modernize assets and applications in a data fabric.

WARNING

Some of the upcoming strategies may serve as stopgaps in the short run, while others put the data fabric to more nuanced and powerful use. Those latter strategies speak most directly to business drivers (faster ROI, cohesive platform and POV, and best/most effective use of technology) described in the preceding section.

TIP

In the following section titles, I use the abbreviation *A&A* to mean "Assets and Applications."

Re-host A&A

Re-hosting appears in the upper-right corner of Figure 6-1. It involves what pundits call "lift and shift migration" of existing systems. That means taking systems from their current runtime environments and moving them into the cloud with little or no change to code and data.

On the plus side, rehosting is fast, involves little implementation cost, and is easy to do. It usually involves no disruption of ongoing activity. But its minuses are massive and increase over time. Because applications and data don't change, they don't incorporate advanced, active metadata. This poses issues— most important: no policy enforcement for security, compliance, or governance, no catalog, no self-serve.

There's more: Lift and shift transplants what worked in the datacenter or at the edge and does the same in the cloud. This can incur unnecessary, higher pay-as-you-go costs doing things the old way. Over the long haul, this gets expensive!

Refactor A&A

Refactoring is a well-known discipline in computer science. It involves restricting existing computer code without changing its external behavior. Refactoring seeks to improve application design, structure, and implementation without changing its functionality. This lets organizations switch to new and emerging technologies without changing the way applications work or altering what they do.

On the plus side, refactoring usually delivers improved code readability and maintainability, reduced complexity, and a cleaner more modern object model that improves extensibility.

Refactoring involves replacing an existing system with a new system and incurs disruption as users cut over from old system to new (both systems often run in parallel for a while, to make sure the new system is ready to take over for good). On the minus side, refactoring involves more time, effort, and cost.

Rearchitect A&A

Rearchitecting involves changing functionality and behavior in addition to everything that refactoring entails. Essentially, re-architecting means designing and building new "cloud-native" systems to replace existing ones. Along the way, it makes fullest possible use of data fabric tools and capabilities.

On the plus side, rearchitecting enables easy, straightforward migration of applications across the organization (edge-to-core-to-cloud) and uses everything the data fabric can do. On the minus side — you guessed it — rearchitecting takes more time and effort, and costs more than refactoring. The degree of disruption for rearchitecting is about the same as that for refactoring because both involve a changeover from an old, existing system to a new one.

Build New A&A

Only two classes of application make sense for a "build new" scenario: existing applications already scheduled for retirement that need replacement, and brand-new stuff that nobody has built yet, but for which a clearly felt (and stated) need exists. This is what happens to new things that come along once the data fabric is available.


TIP

The time, effort and cost of building a new application depends on the application, but it will take some of all of those things to make it happen. The disruption that new systems cause also varies, but usually relates to learning new ways to work and get things done. These can be profound hills to climb, but with profound pots of gold at their summits.

Staying Agile with CI/CD

An up-to-date approach to data fabrics requires implementing a DataOps practice within the organization to work its magic. Among other things, this means following an Agile methodology known as CI/CD, which stands for "continuous integration, continuous delivery." This means that data professionals and analysts, business stakeholders, and IT people work together to make sure they understand the organization's data, the applications

and services that produce and consume that data, and the business value that these things provide.

Continuous integration also translates into self-monitoring and self-measurement for the data fabric so that it keeps getting better at managing data through its entire lifecycle (intake, enrichment, delivery and use, storage, and ultimately, retirement and deletion/destruction). Continuous delivery means the data fabric and its data holdings are continuously subject to change through extensions and enhancements of existing tools and technologies. At the same time new tools, technologies and data sources appear and are integrated and accommodated within the data fabric.

REMEMBER

At any given time, the data fabric will be its best, and do its best, for its users and their organization. Over time, what's best must keep getting better — or it isn't working right.

Delivering the Data Fabric

A data fabric can be delivered in a variety of forms, through various technology and service offerings:

- **On-premises:** The data fabric runs on equipment in a data center somewhere. This usually applies only for limited-use scenarios or pilot testing because the cloud is ubiquitous now.
- **Private cloud:** The data fabric runs as a service through a private cloud from an application or service provider. This will be the starting point for many production deployments.
- **Public cloud:** The data fabric runs as a service through any number of public clouds. This is of increasing interest as it allows to engage in one or more such clouds, where all should interoperate freely.
- **Hybrid cloud:** The data fabric runs on-premises, at the edge, and in the cloud as a truly hybrid offering.

TIP

Hitachi Vantara provides services, tools, and expertise to design, pilot, deploy, and operate data fabrics. Visit `https://hitachivantara.com` and search on "data fabric" (or add specific modifiers, if you like). You'll find a wealth of material on this subject there, including whitepapers, videos, data sheets, and more.

IN THIS CHAPTER

- Modernizing the fabric from edge-to-core-to-cloud
- Extracting benefits via cataloging, enrichment, and more
- Driving global insights throughout the business
- Delivering on the promise of data fabric

Chapter 7
Ten Fabulous Fabric Facts

Here are ten worthwhile facts about data fabrics to peruse and ponder:

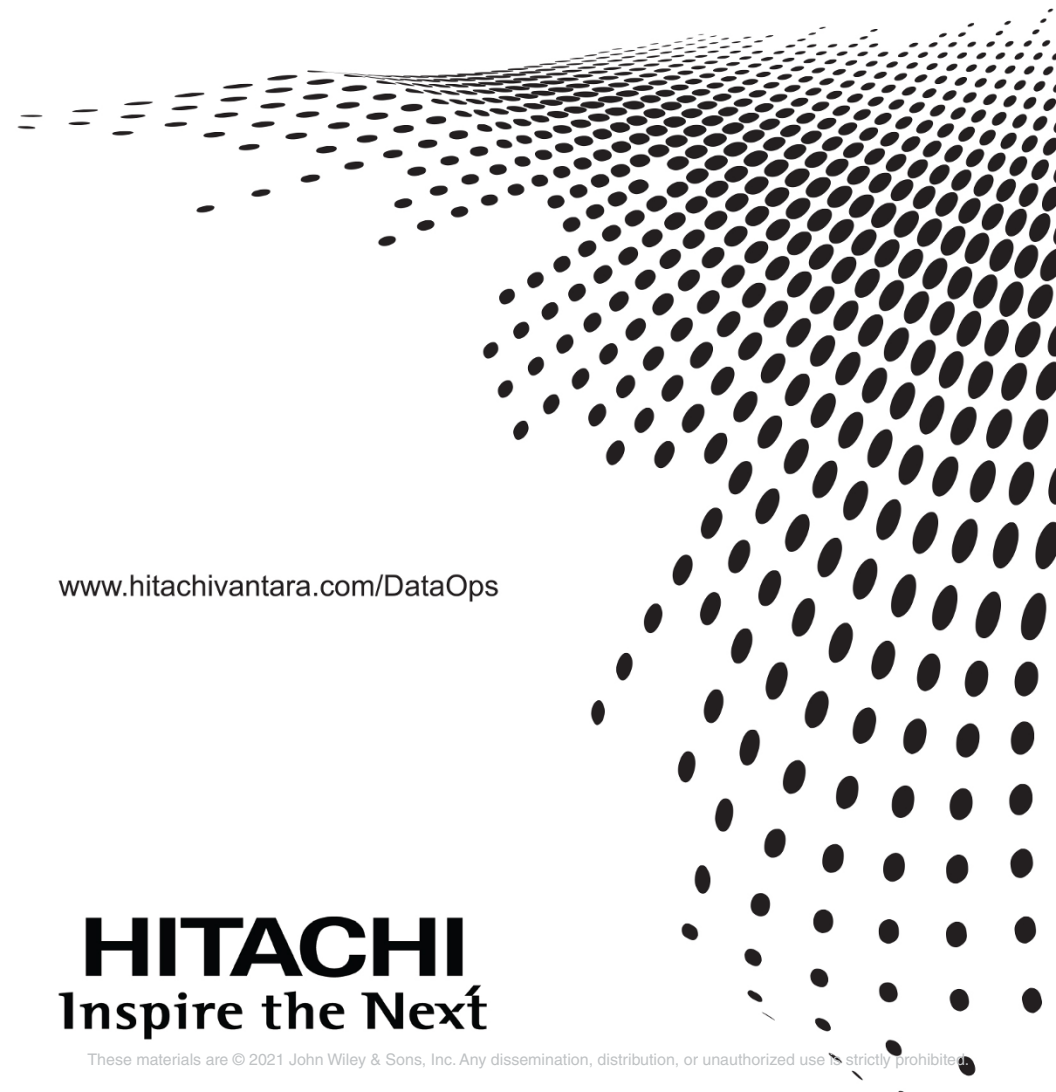
- **Using intelligent data tiering for cloud cost savings:** With only active, needed data in the cloud, and tiering other data in cheaper near- or off-line S3 storage, organizations can save big overall on cloud costs.
- **Making edge-to-core-to-cloud coverage doable**: A proper data fabric seamlessly integrates data from remote edge facilities, core data center operations, and public cloud to meet your unique business demands. A proper data fabric solution makes this into a real game changer.
- **Making the data catalog key:** A proper data catalog accelerates data discovery and metadata tagging to secure sensitive data, infer hidden relationships, and improve data self-service and insights.
- **Using semantic enrichment to increase data value:** Semantic enrichment adds key business specific metadata and grooms data collections for faster, more meaningful

analyses. The net-net is a big boost in insights and inferences.

- **Delivering the goods on privacy and compliance:** Proper identification and labeling of data puts organizations ahead on compliance and governance: They know just what they must protect, audit, log, and report on.
- **Enduring benefits from DataOps:** Integrating the people who know the data — in business and what it represents — with the people in IT who build the applications that ingest, handle, and analyze that data creates better applications and insights faster and more cheaply. What more could you want?
- **Making fabric automation work:** Automation is key to all modern IT operations. Effective use of data fabric automation makes data more accessible and available. It ensures optimal end-user experiences interacting with data, and better insights when using AI-based models.
- **Driving data insights, improving the data fabric:** A modern data fabric supports data-driven applications where they'll do the most good. It also fosters increasing data value through semantic enrichment, optimized storage, and enhanced analytic access. Continuous improvement also means more innovation and better services.
- **Consolidating management to drive global insight:** A modern data fabric offers a single, coherent view of data repositories and resources wherever they reside. By eliminating duplication, offering clean, enriched data sources, and optimizing access, organizations benefit from more and better insights.
- **Making dual use of AI/ML:** AI technologies not only provide ever-improving models for data analysis and insights, they also drive the operation and behavior of the data fabric itself for improved reliability and performance.
- **Adopting client obsession and CI:** CI stands for "continuous integration." With a modern data fabric, organizations can zero in precisely and directly on their clients to understand and serve them better. Bringing CI to the mix means that levels of service and improvement just keep getting better and better. Ultimately that's what makes a data fabric worthwhile.

Build your data fabric
with Hitachi Vantara

Engage the experts and lay the foundation for AI-driven business with a modern and modular data infrastructure engineered for DataOps.



www.hitachivantara.com/DataOps

HITACHI
Inspire the Next

Deliver an end-to-end data management solution

A *data fabric* provides each consumer of data with a consistent and coherent view of all the integrated and packaged data they may need, irrespective of location. It is a living, evolving collection of capabilities that grows and changes along with the organization it serves. This book explains the ins and outs of data fabrics, and how they fit best into a modern enterprise.

Inside...

- Explore the components of a data fabric
- Implement key DataOps concepts
- Optimize data for cost and performance
- Make the most of data governance
- Support compliance, insight, and action
- Achieve data modernization
- Explore data fabric use cases

HITACHI
Inspire the Next

Ed Tittel is an author, trainer, and consultant with more than 100 technology books to his credit.

ISBN: 978-1-119-79116-4
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.