Brought to you by:

Hitachi Digital Services

Data Reliability Engineering

for dummies®

A Wiley Brand

Gain data advantage via engineering excellence

Boost accuracy & speed with data engineering

Scale data resiliency with data observability

Hitachi Digital Services Edition

Senthil Ramachandran, Ramneek Jaitla, and Madhusudhanan Panchapakesan

About Hitachi Digital Services

Hitachi Digital Services, a wholly owned subsidiary of Hitachi Ltd., is an edge-to-core digital consultancy and technology services provider helping organizations realize the full potential of AI-driven digital transformation. Through a technology-unified operating model for cloud, data, and IoT, Hitachi Digital Services' end-to-end value creation for clients is established through innovation in digital engineering, implementation services, products, and solutions. Built on Hitachi Group's more than 110 years of innovation across industries, Hitachi Digital Services helps to improve people's lives today and build a sustainable world tomorrow.

Data Reliability Engineering

Hitachi Digital Services Edition

by Senthil Ramachandran, Ramneek Jaitla, and Madhusudhanan Panchapakesan

for dummies®
A Wiley Brand

Data Reliability Engineering For Dummies®, Hitachi Digital Services Edition

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

ISBN: 978-1-394-20298-0 (pbk); ISBN: 978-1-394-20299-7 (ebk); ISBN: 978-1-394-28220-3 (ePub). Some blank pages in the print version may not be included in the ePDF version.

Publisher's Acknowledgments

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

Project Manager and Editor: Carrie Burchfield-Leighton

Sr. Managing Editor: Rev Mengle

Acquisitions Editor: Traci Martin

Sr. Client Account Manager: Matt Cox

Table of Contents

Introduction

The world is awash in data. How do you effectively use all this data to inform your business decision making? As more businesses seek to be more data driven, the data doing the driving can often become useless.

Proponents of data reliability believe there's a better way to do it.

About This Book

Welcome to *Data Reliability Engineering For Dummies*, Hitachi Digital Services Edition. This book helps you understand the goals and objectives of data reliability engineering (DRE) and how adopting a culture based on data observability and reliability helps your business.

You discover how to dispose of dated and other kinds of poor-quality data, leaving only the finest, most current, and smartest data in your data warehouse. You make decisions based on the highest quality data you can access.

This book covers several topics, including

- Why you need more reliable data to make better decisions
- The causes of poor data quality and what its costs are
- The relationship between data observability and reliability
- The importance of understanding the lineage of your data, asking yourself questions such as how did this batch of data come to this conclusion
- Identifying problems with your existing data
- The principles of DRE
- The special skills of a data reliability engineer
- How to build an authentic culture around data reliability
- Collaborating to ensure high-quality data
- How Hitachi Application Reliability Centers help you get the most from your data

Icons Used in This Book

Throughout this book, different icons highlight important information. Here's what they mean:

The Tip icon flags useful information or explains a shortcut to help you understand a feature or concept.

Try not to forget the material marked with the Remember icon. It signals an important concept or process that you should keep in mind.

Pay attention here. This icon highlights problems and things you should avoid if you can.

This icon points you to relevant statistics and other news about DRE. You can skip these sections if you aren't the techie type.

Beyond the Book

This book can help you discover more about DRE, but if you want resources beyond what this book offers, Hitachi Digital Services can give you additional insight. Check out the following resources:

- www.hitachids.com
- hitachids.com/service/application-reliability-centers

IN THIS CHAPTER

- Making the move from site reliability to data reliability
- Listing the benefits of data reliability

Chapter 1
Shifting to Data Reliability Engineering

You don't have to be a baseball fan to understand the transformation of business from lucky gut feelings to data-driven decision-making. But it doesn't hurt, either.

Perhaps you've heard the story of *Moneyball*. If you haven't, it's worth two hours of your life to watch the 2011 movie. Better yet, read Michael Lewis's 2003 book.

The short version: For its first 100 years, professional baseball was known to be a game driven by statistics. Young fans tracked player batting averages, home run totals, and earned-run averages on baseball cards. Teams rated and compensated players based on these numbers as much as anything else. When home run king Babe Ruth was asked why his salary was more than the President of the United States, he replied, "I had a better year than he did."

But decades after the era of the Sultan of Swat, as computers started showing up in baseball fans' homes, some of them wondered if there were other, better ways of understanding the game through the statistics they kept. Bill James published his first *Baseball Abstract* in 1977 to identify these new methods, using the same basic statistics Major League Baseball had always kept.

Other people followed, eventually including executives of the baseball operations office of the major league Oakland Athletics (now in Las Vegas).

When the Athletics started winning with players who didn't have a high batting average, but were generating more runs (and thus, more wins) for their teams, other teams noticed. Now everyone embraces analytics when evaluating ballplayers.

What does this have to do with your business? Plenty. You're drowning in a sea of data, but may not be sure if you're tracking the right numbers to make good decisions. What if the data you're using is dated, broken, or just plain wrong? How can you make the right decisions if the data just isn't reliable? Data reliability engineering (DRE) was created to help you.

Moving from Site Reliability to Data Reliability

As with many trends in 21st century technology, DRE owes its existence to Google. About the same time as Michael Lewis published *Moneyball,* in 2003, Google established its site reliability team under the leadership of Ben Treynor. This group's task was resolving outages anywhere in the company's vast domains.

Over the last two decades, site reliability engineering (SRE) has become an engineering discipline itself, with conferences, principles, and certification programs. Site reliability engineers build automated software to optimize application uptime while minimizing *toil* (the kind of work that tends to be manual, repetitive, automatable, tactical, and devoid of enduring value) and reducing downtime. On top of these duties, SREs are known as the firefighters of the engineering world, working to address hidden bugs, laggy applications, and system outages.

Data engineers started looking at the SRE principles and saw some methods that resembled the practices they were using. Eventually, they developed a set of principles for DRE. We cover those in Chapter 2.

The primary goal of DRE is to make high-quality and reliable data available 24/7, from anywhere across the enterprise.

REMEMBER

DRE gives a name to the work of improving data quality, keeping data moving on time, and ensuring that analytics and machine learning products are fed with a healthy set of inputs.

Identifying the Benefits of Data Reliability

While business acknowledges the critical role of data, organizations are still uncertain about making data accessible and reliable. Data often lives in silos, and no one can really guarantee its trustworthiness.

Dr. Thomas Redman runs Data Quality Solutions, which helps companies and people chart a course to data-driven futures. This course focuses on quality, analytics, and organizational capabilities. He notes that knowledge workers waste up to 50 percent of time hunting for data, identifying and correcting errors, and seeking sources to confirm data they don't trust.

Data quality more broadly has been a topic for decades but has gotten markedly more attention recently. You can see the impact of poor data quality when peering at your balance sheets with lower revenue and higher operational costs, both resulting in financial loss. Poor data quality also significantly influences organizational efforts in governance and compliance, leading to rework (toil) and delay.

The increasing complexity of the data stack, the sheer volume, variety, speed, and quantity of data generated and collected, opens the door to more complex issues like schema changes, data drift, downtimes, duplicate data, and other complex issues. The many data storage options, data pipelines, and enterprise applications make data management challenging.

As the amount of data being generated, shared, and stored continues to explode, data-intensive applications like generative artificial intelligence (AI) and large language models (LLMs), not to mention everyday business decision-making, are demanding cleaner, more accurate, and more complete datasets than ever before.

Think about how much of your data is created and consumed by AI right now: Financial planning, managing inventory, helping your customers with product recommendation engines — and now chatbots running on AI!

WARNING

Unreliable data can lead to wasted time, lost revenue, compliance risk, and erosion of customer trust. Data quality issues crop up in the most mundane situations. When have you seen this happen? You can probably think of instances like the following:

- You get mail from an organization and see your name misspelled in the address window. Do you immediately open the letter to see what they have to tell you? Or does it go straight into the junk mail pile? Your respect for that organization changes.
- Your mailing list has incomplete information, or the information is dated or otherwise obsolete. That new and potentially lucrative account never connects.
- Sales never shares its customer data with Customer Service. Both teams lose an opportunity to create more accurate and complete customer profiles. Your company gets defeated by silos that never connect information.

High data quality ensures that data accurately represents real-world entities and can power reliable insights. Some crucial indicators that signal the necessity of enhancing data quality within your organization include the following:

- Your data platform has recently migrated to the cloud.
- Your data stack is scaling with more data sources, more tables, and more complexity.
- Your data team is growing.
- Your team is spending at least 30 percent of their time firefighting data quality issues.
- Your team has more data consumers than you did a year ago.
- Your company is moving toward enabling self-service business intelligence capabilities.
- Data is a key part of the customer value proposition.

If this sounds like your team, and you're ready to explore this change, dive into the other parts of this book.

IN THIS CHAPTER

- Ruining your data
- Looking at the pitfalls of unreliable data
- Seeing the journey data takes
- Empowering your users

Chapter 2
Understanding Data Reliability

Many companies today rely on data to learn more about customer behavior, improve their own efficiency, and make better-informed business decisions. These enterprises, in today's digital landscape, are gathering data from multiple sources. However, the value of the data is only as good as its quality and reliability.

In most cases, enterprise data is stored in different formats and locations with varying levels of quality, leading to data silos and inconsistencies. The main goal of data reliability engineering (DRE) is to make traceability to the originating source and reliable data available at any time, from anywhere across the enterprise. Data reliability describes the overall consistency and quality of an organization's data.

WARNING

When your business relies on data, both you and your customers need to ensure consistency and accuracy. Undetected errors and data outages can have a severe impact on your business. Incorrect data can lead to poor decisions that affect everything from inventory management and financial planning to product recommendations and support systems.

In this chapter, you discover the many ways data can break and the consequences of using broken, unreliable data.

Breaking Data, Data Pipelines, and Data Infrastructure

In today's era of Big Data, data retrieval can go bad in many ways:

- **Schema changes:** Does the data format conform to the schema? What has changed in the data schema? Who made the changes?
- **Downtimes:** Planned or otherwise, data can become corrupted when the system goes down.
- **Duplicate data:** Same data items appear multiple times and make it cumbersome to identify the right data items.

Keep in mind these examples:

- In June 2023, a wellness chatbot was launched to interact with people seeking 24/7 help. The chatbot was shut down after news reports surfaced that several of its users reported that they'd received advice on losing weight, even when the bot was informed the user had an eating disorder. What seemed to have broken down here is the way the chatbot processed the data it received through its pipeline. The developer claimed there were guardrails put up so chatbot always gave good advice. But clearly something went wrong.
- More seriously, in the fall of 2022, an executive health agency in England found a collection error in their COVID-19 tracking spreadsheet. Over 15,000 positive tests weren't included on the spreadsheet because the spreadsheet hit its data limit. Those patients weren't informed, and the healthcare officials didn't trace their contacts.

Facing the Consequences of Unreliable Data

Data quality issues can result in significant costs, with estimates reaching nearly $12.9 million annually in 2021. This doesn't factor in lost opportunities. Poor data quality can harm your reputation

as a reliable source. If you position your company as data-driven, and if your actions, products, and services are designed to adapt to changing conditions, it's crucial that your data is always reliable.

Grasping the Data Journey

So how do you define data quality? Look for these qualities:

- **Consistency:** You're never comparing apples and oranges because every entry is consistent and standardized.
- **Freshness:** Every entry is current, with no out-of-date material mucking up the analysis.
- **Distinct:** Each data set is unique without repeated or irrelevant entries.
- **Complete:** Data should be representative, giving a clear picture of real-world conditions.
- **Valid:** The data conforms to the formatting required by your business.

The search for quality data begins with rooting out the bad stuff. For that, your data needs observability.

The importance of observability

Data observability is your ability to keep a pulse on your data systems. This happens through tracking, monitoring, and troubleshooting issues to reduce downtime and improve data quality. The ultimate goal is to prevent issues from happening.

Your data observability tools work to make your data pipelines more efficient and useful. A *data pipeline*, shown in Figure 2-1, acts as a conveyor belt that takes data from different places, cleans it up, and organizes it for the purpose of studying valuable business insights. In a typical organization, the data flow architecture involves collecting, storing, and processing data from multiple sources, including internal systems, streaming sources, and external data providers. This processed data is then utilized by business intelligence platforms, user applications, and various business operations to drive insights and actions.

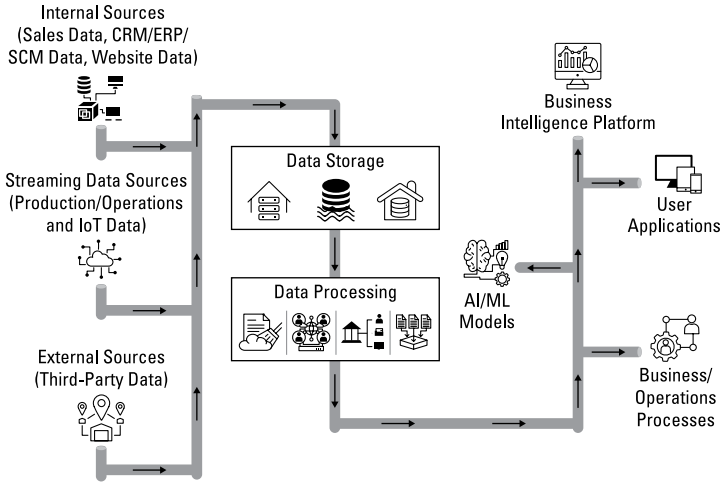


FIGURE 2-1: Data flows from various sources through a data pipeline.

There are many processing steps to prepare enterprise data for analysis. Organizations have a large volume of data from various sources like applications, Internet of Things (IoT) devices, and other digital channels. However, raw data is useless; it must be moved, sorted, filtered, reformatted, and analyzed for business intelligence.

A data pipeline includes various technologies to verify, summarize, and find patterns in data to inform business decisions. Well-organized data pipelines support various big data projects, such as data visualizations, exploratory data analyses, and machine learning tasks.

Looking into the rules of DRE

Many DREs have applied Google's site reliability engineering (SRE) principles to the data world — data and underlying infrastructure. They look something like this:

- **Embrace risk:** The only perfectly reliable data is no data at all. Data pipelines break in unexpected ways; plan for how to manage it effectively.

- **Set standards:** Clarify what your (internal and external) clients and customers can depend on with clear definitions, hard numbers, and cross-team agreements.
- **Reduce toil:** Eliminate busywork! Cut out any repetitive manual tasks from your data platform. You'll get reduced overhead and fewer human errors.
- **Monitor everything:** It's impossible for a data team to understand how their data and infrastructure is behaving without comprehensive, always-on monitoring.
- **Use automation:** Automating manual processes reduces manual mistakes and frees up brainpower and time for tackling higher-order problems. It also reduces toil.
- **Control releases:** When you make changes, two things can happen: things improve, and things break. Having a process for reviewing and releasing data pipeline code helps you ship improvements without causing breakage.
- **Maintain simplicity:** The enemy of reliability is complexity. Minimizing and isolating the complexity in any one pipeline job goes a long way toward keeping it reliable.

To better understand how these principles are applied in practice, take a look at Figure 2-2. It illustrates key components and practices of DRE and visually represents the integration of SRE principles into the data world, focusing on risk management, standards setting, toil reduction, comprehensive monitoring, automation, controlled releases, and maintaining simplicity. Plus, it shows the evolution of data architectures in organizations. The traditional data stack uses structured data and manual, on-premises processes to support analytics and applications. The modern data stack introduces a flexible ecosystem with a lakehouse for diverse data, improving data observability and tool integration. The data first stack, with the Hitachi Unified Data Architecture, focuses on data reliability through a consumption gateway and rule-based data engines. This progression highlights how organizations can improve data monitoring, observability, and reliability for better data-driven decisions.

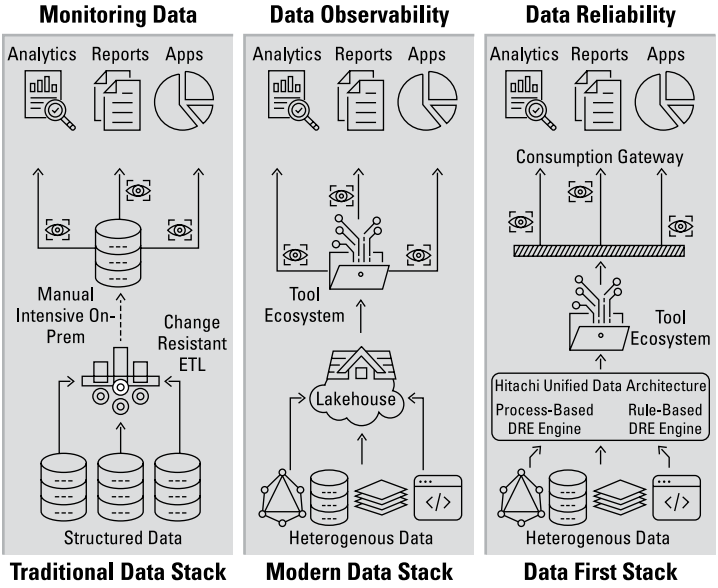


FIGURE 2-2: A comparison of data stacks.

Empowering Users

Data reliability leads to better decisions. Your goal is to empower everyone in your organization to manage data to learn, innovate, and develop new ways of thinking. It starts with building an effective pipeline architecture. Pipelines process data to help you understand your data needs and the data problems to solve.

TIP

Many organizations build their data pipelines with the help of the open source Kafka pipeline framework, originally developed by LinkedIn to handle their messaging load. Several commercial pipeline tools use Kafka as a baseline.

IN THIS CHAPTER

- Gauging data quality
- Configuring data ecosystems
- Discovering self-healing and self-service experiences

Chapter 3
Driving Your Business with Quality Data

Poor data quality can lead to uninformed decision making. To solve your data reliability issues, you need to identify what technical problems you need to overcome. Next comes developing a strategy aimed at delivering better analytical data. In this chapter, you get a basic look at how to do this.

Measuring Data Quality

To measure your data quality, you want to start with determining the size and impact of your existing data reliability problem. This is best understood by measuring data downtime: the periods when you can't make effective, data-driven decisions because the data you're using is inadequate to the task; it's incomplete, dated, missing, or just plain wrong.

IDC reports that just over a quarter (27 percent) of data practitioners fully trust the data with which they routinely work. In its Data Integrity Trends report, Corinium found that 82 percent of respondents believe data quality concerns represent a barrier to their data integration projects. Over 50 percent reported that data quality is very challenging in their organizations, coming in

at number one on the list of data integrity concerns among the survey respondents.


TIP

You can measure data quality in various ways, depending on your domain and business context. Several tools, like DataDog, AppDynamics, and NewRelic, are available to measure data quality and identify quality issues. Software engineers can also monitor the health and performance of their applications by using tools — data teams must also do the same.

Delivering useful information to decision makers

Incomplete data, or data that doesn't account for all available inputs and sources, and other data-related issues give an incomplete picture of the organization. Accurate data can lead to better customer experiences, improved financial performance, and more efficient operations.

Data teams don't need to see everything about their data — they need to find what matters to solve a problem or answer a question. It may impress someone how big the database might be, but if 90 percent of your database is older than the Great Recession, it won't be of much use.

Identifying problems with your data

How do you assess your current data quality? It starts with data validity. Data validity measures a company's data by how usable and applicable it is for business operations. Valid data must follow a specific format and rules; if it doesn't meet these standards, data can't be valid or reliable.

There are various types of validity relevant to your organization. These types include

- Data type checks
- Code checks
- Format checks
- Consistency checks
- Range checks
- Uniqueness checks

Each type of data validity is essential in understanding your company's data performance and management.

To identify data issues, ponder the following questions:

- Is the data up-to-date (fresh)?
- Is the data complete? Do your data tables let you know how healthy your data sources are?
- Are fields within expected ranges?
- Is the null rate higher or lower than it should be?
- Has the schema changed, leading to broken data?
- Do we have data quality metrics?

DETECTING ANOMALIES

Data reliability engineering (DRE) collects and analyses data from multiple sources to detect anomalies. To do this, it uses the following processes:

- **Data collection:** DRE collects data from various sources, Extract, Transform, Load (ETL) logs, and target systems to effectively identify and address data quality issues, maintaining the integrity and consistency of data across systems to ensure a comprehensive view of data quality.
- **Anomaly detection:** By analyzing collected data, DRE detects anomalies related to file availability, data freshness, schema changes, and the availability of key columns.
- **Correlation analysis:** DRE uses data from ETL logs to identify unusual patterns and performance issues, correlating these with data from source systems and target databases.
- **Reconciliation:** DRE performs reconciliation between source data and target systems to ensure that data transformed through ETL processes accurately matches the original input.

Organizations can ensure their data is trustworthy and valuable for decision making, analytics, and operational processes with data quality metrics, such as the ones shown in Figure 3-1.

$$q_i = \sum_{j=1}^{K} q_{ij};\ q_{ij} = 1 \text{ if } v_{ij} \text{ is syntactically accurate}$$

$$\textbf{Accuracy} = \sum_{i=1}^{N} \frac{B(q_i = K)}{N}$$

$$q_i = 0 \text{ if record } i \text{ is not unique}$$

$$\textbf{Uniqueness} = 1 - \sum_{i=1}^{N} \frac{B(q_i = 0)}{N}$$

$$q_i = \sum_{j=1}^{K} q_{ij};\ q_{ij} = 1 \text{ if } v_{ij} \text{ is not NULL}$$

$$\textbf{Completeness} = \sum_{i=1}^{N} \frac{B(q_i = K)}{N}$$

$$\textbf{Timeliness} = \max\{0, 1 - \frac{T}{F}\}$$

$$q_i = \sum_{j=1}^{K} q_{ij};\ q_{ij} = 1 \text{ if } v_{ij} \text{ does not violate constraints}$$

$$\textbf{Consistency} = \sum_{i=1}^{N} \frac{B(q_i = K)}{N}$$

FIGURE 3-1: Data quality metrics are a mathematical foundation for evaluating data quality effectively.

The formulas for metrics calculation include the following:

- Metrics can vary from 0 to 1 *(0 is poor, 1 is excellent)*.
- *B* is a Boolean function returning 1 if the condition is met.
- *vij* is the value at column j and row i.
- For Timeliness metric, *T* is the Age of the data and *F* is the Frequency of the data.

Feel free to add your own questions that are more specific to your situation.

Designing Your Data Ecosystem

The key to data reliability is to create an observability framework. One of the best ways to ensure observability into your data is by defining its sources, or *lineage.*

When you're combining analytical data from multiple sources to help you decide how to move forward, know what those sources are and whether they're the best quality sources you can find.

When data breaks, the first question is always "where?" Data lineage provides the answer by telling you which upstream sources and downstream processors were affected, as well as which teams generate the data and who accesses it. Good lineage also collects information about the data (also referred to as *metadata*) that speaks to governance, business, and technical guidelines associated with specific data tables, serving as a single source of truth for all consumers.

Lineage is an important piece of the best observability tools. The information you get from them should give you great confidence that you're using the right information to decide critical questions.

A data lineage chart such as the one in Figure 3-2 is a natural visualization to show how data flows from its source to its destination.

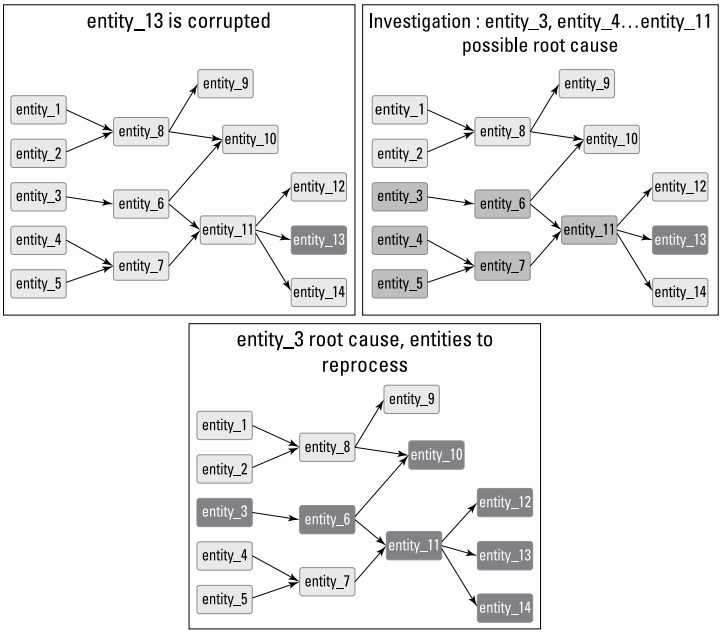


FIGURE 3-2: An example of a data lineage chart.

WARNING

Beware, though. Some data lineage tools are better at generating great-looking graphs and charts than at identifying your problems.

Using data fabric and data mesh

Modern architectures like data fabric and data mesh are quite popular these days, and they complement data reliability in different ways. At the highest level, a data mesh is an organizational paradigm and a data fabric is a layer of technology.

What is a data fabric?

By harnessing the magic of metadata through artificial intelligence (AI) and machine learning (ML), a data fabric can generate previously undreamed-of connections between disparate data sources. A data fabric can mean different things, depending on who you're talking to, but as a starting point, a data fabric

- Integrates and connects all your organization's data intelligently and efficiently by abstracting underlying complexity
- Minimizes disruption by enabling a highly adaptable data management strategy with augmented data integration and management
- Doesn't care what deployment platforms, data processing methods, data delivery methods, locations, and architectural approaches you use because it's independent
- Can manage large volumes of data — structured, unstructured, or semi-structured

What is a data mesh?

A *data mesh* is a data platform architecture that embraces the thought that data is everywhere in the enterprise. It leverages a domain-oriented, self-serve design and aims to democratize data to help data consumers discover, understand, trust, and use data and data products to drive data-driven decisions and initiatives.

Data mesh focuses on organizational change and enables domain teams to own the delivery of data products, understanding that the domain teams are closer to their data so they understand their

data better. Data teams view data mesh as a prime opportunity to transition from monolithic data platforms to data microservices (business contextual services) architecture.

Core to data mesh is the concept of breaking apart the monolithic architecture and monolithic kind of custodianship or ownership of the data around domains in an organization. Data warehouses and data lakes can still exist in the mesh architecture, but they become just another node in the mesh.

Supporters of data mesh advocate distributed, domain-based ownership and custodianship of data and building data products that are self-described and atomic, more easily managed and delivered at the domain level. This makes data observability (and reliability) essential to data mesh. To effectively manage the pipelines, the owners need as much observability into the pipelines as possible.

These data products are sharable with other domains and interoperable with other data products that form the data mesh. A data mesh manages data as a distributed network of self-describing data products.

As data sharing and more distributed architectures like the data fabric or data mesh come to prominence, processes and workflows that treat data like a developing, cross-disciplinary entity will become industry standards.

Fitting in DataOps

While data reliability engineering (DRE) was born from site reliability engineering (SRE), you'll almost hear as much about DataOps as SRE.

DataOps believes businesses need to treat data as a valuable asset and must manage and process it efficiently. It emphasizes the importance of collaboration between different teams, such as data engineers, data scientists, and business analysts, to ensure that everyone has access to the right data at the right time.

DataOps also encourages a culture of continuous improvement and innovation, as teams work together to identify and address

bottlenecks and inefficiencies in their data pipelines and processes. You can probably see the connection between these two practices.

TECHNICAL STUFF

The Data Science Council of North America calls data observability an "outcome" of the DataOps movement. It states that you can have the most advanced automation and algorithms to monitor your metadata, but those will only benefit from organizational adoption. However, anyone can adopt DataOps as an organization, but DataOps should be a well-documented philosophy that doesn't impact output without the technology to support it.

REMEMBER

Hitachi Digital Services provides a suite of DRE services as part of Hitachi Application Reliability Center (HARC). Grounded in proven DataOps methodologies and practices, HARC delivers a robust, prebuilt solution for data reliability that effectively addresses a wide range of data challenges. This offering encompasses reliability architecture, design, and implementation, optimizing data reliability, observability, and governance through various methods.

The key features are HARC include the following:

- High visibility into data quality, trustworthiness, and age, aiding users in deploying data in innovative ways
- A self-healing, predictive maintenance approach to manage pipelines, preventing disruptive incidents
- A shift-left strategy, moving data back to source applications to enhance data intelligence and governance at the source

For more information, visit `hitachids.com/service/application-reliability-centers`.

Creating Reliable Data Systematically

When starting on the technical journey toward data reliability, you also may hear about self-healing architecture and self-service data. What do these mean and where do they fit? In some ways, these describe goals for your data reliability journey.

Building a self-healing data architecture

WARNING

Because data pipelines are a lot of small processes that work together to ensure your data loads into the target database, one or more of these processes could fail, resulting in the following issues:

- **Poor data quality:** Inconsistent or poor-quality processed data
- **Technical issues:** Network failures, system failures, and bugs in the pipeline
- **Human error:** Incorrect adjustments to the data pipeline, unauthorized changes, and general mismanagement of the pipeline
- **Changes in data:** Business requirements changed at the source — for example, a new column added to the table or changes to data types or structures — causing issues with ingestion, transformation, and loading
- **Scalability:** Increased volume of data, leading to data pipeline failures
- **Lack of maintenance and monitoring:** Ineffectively maintaining and monitoring the pipelines leading to eventual failures

Self-healing systems empower data engineers to start reconciling incidents through automatic repairs. This cuts down the time to get data back on track for analysis. This type of architecture responds to evolving conditions, rapidly identifying any breakdowns and allowing for instant reboots or adjustments to components. As a result, you get enhanced system availability and manage operational expenses effectively.

The goal of a data reliability engineering project is to eliminate as many points of failure as possible with a self-healing system.

REMEMBER

To help identify the key components of your self-healing infrastructure, keep in mind the following four categories and their questions:

- **Current state:** How do you monitor or stream the environment's current state?

- **Baseline:** How do you compare your current state to a certain set of standards (such as risk, vulnerability, or compliance)?
- **Remedy:** If the current state doesn't match the baseline standards, what remediates the drift or vulnerable state?
- **Automation:** How do you automate the remediation process? It wouldn't be self-healing if there weren't an automated remediation, now would it?

Considering self-serve experiences

There's an ugly debate going on in IT circles about self-service data access. Democratizing data can be a great thing, but not always.

Data reliability engineers fall on the side of self-service because the people making the decisions have different data needs and can differ on what makes up high-quality data beyond the basics. Self-service is important because line-of-business professionals and analysts can access and work with data and data visualization directly. They're also supported by (not dependent on) IT and data professionals to carry out their work.

Self-service programs remove technical boundaries and can empower people to use their own subject matter expertise — after all, they know the problems they're tackling best, and they know what data they need — to generate insights and execute their work.

IN THIS CHAPTER

- Identifying data engineering skills
- Finding out the best practices for data reliability

Chapter 4
Molding Data Reliability Engineers

Data reliability engineers work to create standards, process, alignment, and tooling to keep data applications — like dashboards and ML (machine learning) models — reliable, without slowing down the organization's ability to handle more data. Their work also evolves their data pipelines.

Data reliability engineers are crucial for smooth data operations, ensuring quality, movement, and reliability for analytics and machine learning. When managers and executives hear *data reliability engineering* (DRE), they often think they can't afford to hire another engineer. They then move on to some other practice that they think will cost less and hope to bring at least as much benefit to the company as DRE.

Now certainly there are an increasing number of people who have developed the DRE skill set, and hiring a specialist in data reliability is rarely a bad idea. Yet, if you have a team of specialists managing your company's data, it's likely that at least one of them can cultivate their skills and talent in data reliability. Alternatively, you could build a team of data engineers who can collectively lead the data reliability function.

This chapter helps you identify the people with those skills. We can't promise that your new reliability specialist won't want a salary bump, though.

Identifying Data Engineering Skills to Build On

Data reliability engineers don't just put out fires. They put the guardrails in place to prevent them. They enable agility for analytics engineers and data scientists, keeping them moving quickly knowing that safety guards are in place to prevent changes to the data model from affecting production.

Data teams are always trying to balance speed with reliability. The data reliability engineer achieves that balance. DREs handle databases, data pipelines, deployments, and system availability in collaboration with various teams, and they combine elements from data engineering, software development, and systems administration.


TIP

If you're a data engineer, you can help your organization recognize the importance of the work that's data reliability engineering. You get the freedom to prioritize this work appropriately, plus you can make sure it doesn't get sidelined instead of other projects.

Supporting data pipelines

Data reliability really starts by fixing and maintaining the state of your data pipelines. Data pipelines require regular maintenance and monitoring to ensure they function correctly. Not maintaining and monitoring the pipelines effectively can lead to failures.

Data reliability engineers should be comfortable in data engineering and scripting languages. Testing, observability, and transformation tools are useful here, too, but a data engineer should never run a pipeline with untested data.

Any experience with self-healing pipelines (see Chapter 3) is a clear bonus as well.

Modern data architecture

Familiarity with data mesh and data fabric helps your company rethink how to organize data. Find more information about these architectures in Chapter 3.

Data governance concepts

Managing service-level agreement (SLAs), service-level objective (SLOs), and service-level indicator (SLIs) contracts with your clients (internal and external) measures reliability and performance. You need these governance measures to maintain high-quality data systems.

Pipeline orchestration and monitoring tools

Orchestration is composing or building complex structures from a single responsible block, element, or component. The goal of pipeline orchestration is to streamline and optimize the execution of frequent, repeatable processes and to help data teams manage complex tasks and workflows. Anytime a process is repeatable, and you can automate related tasks, use orchestration to save time, increase efficiency, and eliminate redundancies.

Meanwhile, you need data monitoring tools to ensure high-quality data in your pipelines. Weed out outdated material, corrupted data, and other items that hamper your decision making.

Figure 4-1 illustrates how pipeline orchestration and monitoring tools interact to manage and maintain data quality throughout the entire process. This visual representation emphasizes the importance of both orchestration and monitoring in ensuring efficient and reliable data pipeline operations.

Cloud platforms

Cloud-based data management enables efficient storage, extraction, transformation, archival, and security measures for your valuable information. A well-managed integrated cloud solution helps businesses erase data silos and provides a single source of truth for each data point.

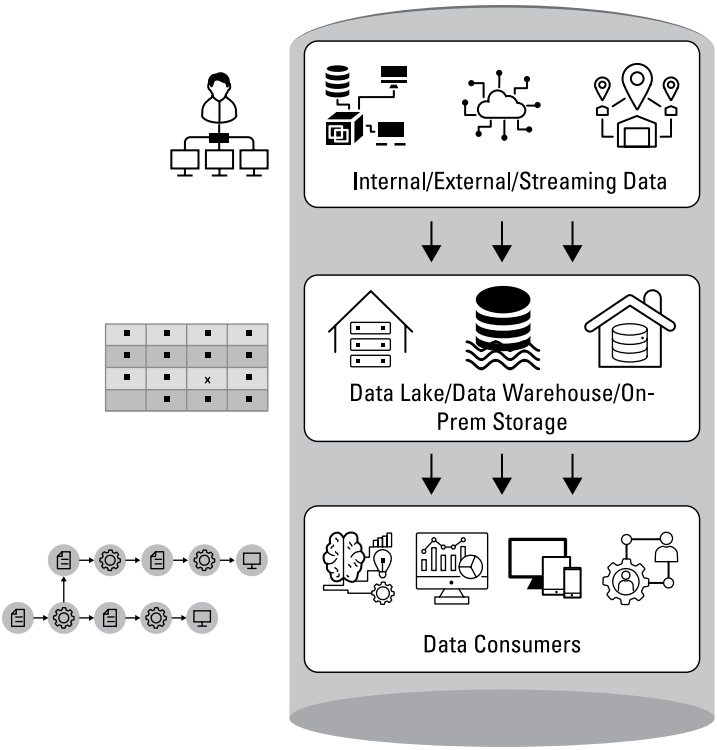


FIGURE 4-1: An overview of data flow from sources to consumers.

Real-time analytics provided by cloud systems assist your decision-making process. Add in complicated machine learning systems and external application programming interfaces (APIs), and you have a recipe that demands consistent, up-to-date, error-free, and deduplicated data.

Some analysts believe that cloud-based data can require more vigilance to maintain data quality standards. To keep data reliable, your DRE should be aware of the following potential issues:

- Risk of data loss when moving data within the cloud or even between the cloud and infrastructure
- Inaccurate timestamps and similar trivial issues
- If one tool has been upgraded and your other tools aren't
- Risk of restructuring data in a new format that your cloud-based tools can't process

Infrastructure provisioning tools

Tools like Kubernetes and Terraform help manage and provision infrastructure and capacity for processing of huge data volumes. Knowing how to work with them is critical for ensuring scalability and reliability.

IP networking

You need networking knowledge for debugging issues that appear during data transmission or access over a network.

Data platform design and scalability

Data platforms run from simple databases, to data warehouses data lakes up to Big Data processing systems and technologies are foundational to DRE. Your DRE should be familiar with the design and deployment of these database management tools. After deployed, they should be able to normalize the data and optimize queries.

Establishing Best Practices

To ensure data quality throughout the entire data stack, data reliability engineers should follow best practices, including

- **Monitoring:** Setting up systems to detect when data is inaccurate or unreliable
- **Alerting:** Letting parties know when you find an issue
- **Testing:** Using automated tests to ensure that data is accurate and reliable

  Writing tests for things like uniqueness and not_null allows organizations to verify their assumptions about source data. It is also common for organizations to ensure that data is in the correct format for their team to work with and that the data meets their business needs.

  Before setting your tests, clearly understand the data, what to expect from it, and what "bad data" looks like.
- **Documenting:** Keeping documentation up to date so that others can understand how the data is being used and maintained

REMEMBER

- **Data cleaning (or cleansing):** Preparing data for analysis by removing poor-quality data from a data set

  The data cleaning process is becoming more distributed, with both the data engineering team and data producers responsible for ensuring clean data. Everyone in the company has a key role in maintaining data integrity, so educate them about the importance of data cleaning.

- **Data enrichment:** Merging and adding either first- or third-party data to data sets you're already working with

  By enriching data, organizations can add more value to their data sets, which in the end makes data more useful and reliable.

You'll find an abundance of tools have emerged in the space to automate this process for companies. These tools check data sets for format, consistency, completeness, freshness, and uniqueness.

TIP

Hitachi's DRE suite of services, powered by Hitachi Application Reliability Centers, deploys an R3 approach: Reveal, Resolve, and Regulate. This approach, shown in Figure 4-2, improves the strength of your data infrastructure.

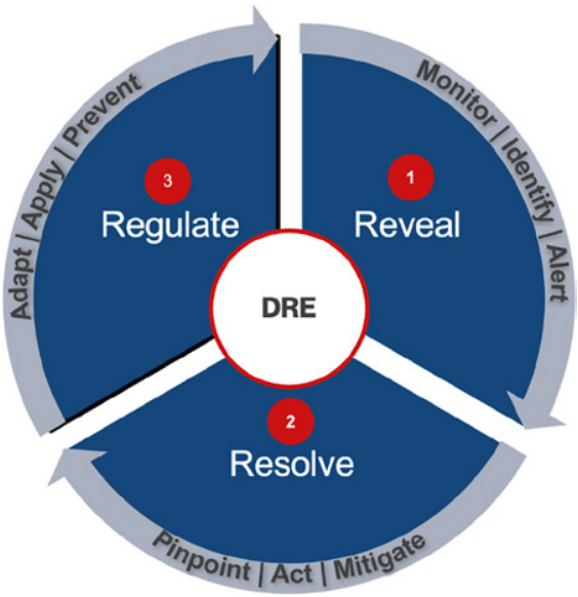


FIGURE 4-2: Hitachi's R3 approach to DRE.

In the DRE R3 methodology, you want to ask yourself some questions relating to each section:

- **Reveal:** What's the current state of our enterprise's data?
- **Resolve:** How does our company detect and resolve issues with or in the data pipeline?
- **Regulate:** Where is our enterprise at risk from a compliance perspective?

Answering and addressing these questions involve many different areas, but generally include a proactive approach with a minimal amount of manual work. Taking this approach will incorporate predictive maintenance of data pipelines, data observability, fault-tolerant architecture, and real-time monitoring to ensure efficient and reliable data management.

Head to `hitachids.com/service/application-reliability-centers` to find out more about Hitachi's suite of services.

IN THIS CHAPTER

- Enhancing your business culture
- Investing in reliable data management
- Fostering data reliability

Chapter 5
Building a Data Reliability Culture

Transforming a business into a high-quality data-driven one isn't just about the technical fixes I discuss in this book. You also need to shift the culture.

Data culture is a company's efforts to engage with all employees and unite the organization with a shared sense of purpose. Data culture is essential for improving your company's workflow and core operations and helps solve data-related issues to guarantee an efficient workflow.

Trusted data is the foundation of good business decisions. Because everyone wants to use trusted data, everyone in the company has to understand their role in building trust in that data. This notion means aligning the culture of your business around generating and maintaining high-quality data. A company requires a healthy culture around the usage and management of its data to guarantee that data is trustworthy, reusable, and supports business operations.

A good data quality culture shows that all your stakeholders clearly understand what data quality means to their business decisions. In this chapter, you get the knowledge of how to make this work.

Creating a Culture of Reliability

A culture centered on data quality is one where stakeholders proactively contribute to improve the quality of data. You start valuing the data and trust the date for better decision making and for indications of the next best actions.

Your stakeholders can also leverage data to drive better business outcomes. Data quality culture means that *every* individual in your organization shares the same mindset to meet data quality goals.

Building your company's data culture helps streamline business operations while creating a unified team of employees with a clear vision to grow and expand your business. Because a healthy data culture is essential for your company, you can benefit from seeking data observability platforms to manage your company's data processes. A comprehensive data culture in modern data stack systems is crucial for today's data-driven business world.

A healthy data culture in your organization provides extra support to you and your team and helps boost the effectiveness of your daily tasks. Data culture and data observability are essential to reduce time spent on menial tasks and focus on other organizational projects and decisions.

But how do you get there? Take a look at a few steps:

1. **Generate top-to-bottom energy for data quality improvement.**

   This step often starts at the top of the company, where the decisions get made. Executive support is critical, but anyone with influence and clout, especially among the data engineers, data scientists, analysts, or the DataOps team, can start the conversation. As momentum builds, people will want to join in.

   Make sure you have people in your company who act as data stewards. They provide business context and serve as points of contact for others in your organization who may have questions or need more information. These folks are extensions of your team in locations around the world. They help manage communication, break down organizational barriers, and drive governance across different sites.

2. **Break down all information silos.**

   Enterprises often have dozens or even hundreds of applications that hold data used by different people and different teams. Figure 5-1 shows how this works (or doesn't, as in this case). You have two teams managing social media with different analytical tools. Three other teams manage customer relationships, supply chains, and resources. One of those teams is "Master Data Management," but doesn't appear to have any connection to the other five silos! Nobody is sharing the original insights these applications generate, leading possibly to a variety of responses to the insights they perceive.

   By eliminating silos, organizations can use data as an asset, transforming their businesses for the better and outperforming their strategic goals. Perhaps the Master Data Management team will earn their title too.

   A culture change begins by making your teams aware of silos and inviting solutions to break them. It is also essential to recognize and appreciate that data quality is not a one-time activity or one team's responsibility.

   Documenting data quality practices can help your teams understand why they're critical. Everyone can see where things go wrong and how they can prevent problems. Tools with collaborative workflows can assist you in addressing data quality issues across the organization.

3. **Establish a robust data quality program.**

   Define the SMART goals for your program. Each of these goals are

   - **Specific:** Not "we will change."
   - **Measurable:** Define exactly what success means.
   - **Achievable:** Complete within the time frame you select.
   - **Relevant:** Make sure it's first relevant within *your* organization.
   - **Time-bound:** When is the due date?

   Invite everyone to ask themselves where they fit in — and give them resources to identify their niche (such as monitoring, cleaning, or just identifying aging data). A predictive, self-service data quality solution can power active monitoring at scale, even for diverse sources and frequent migrations across storage systems.

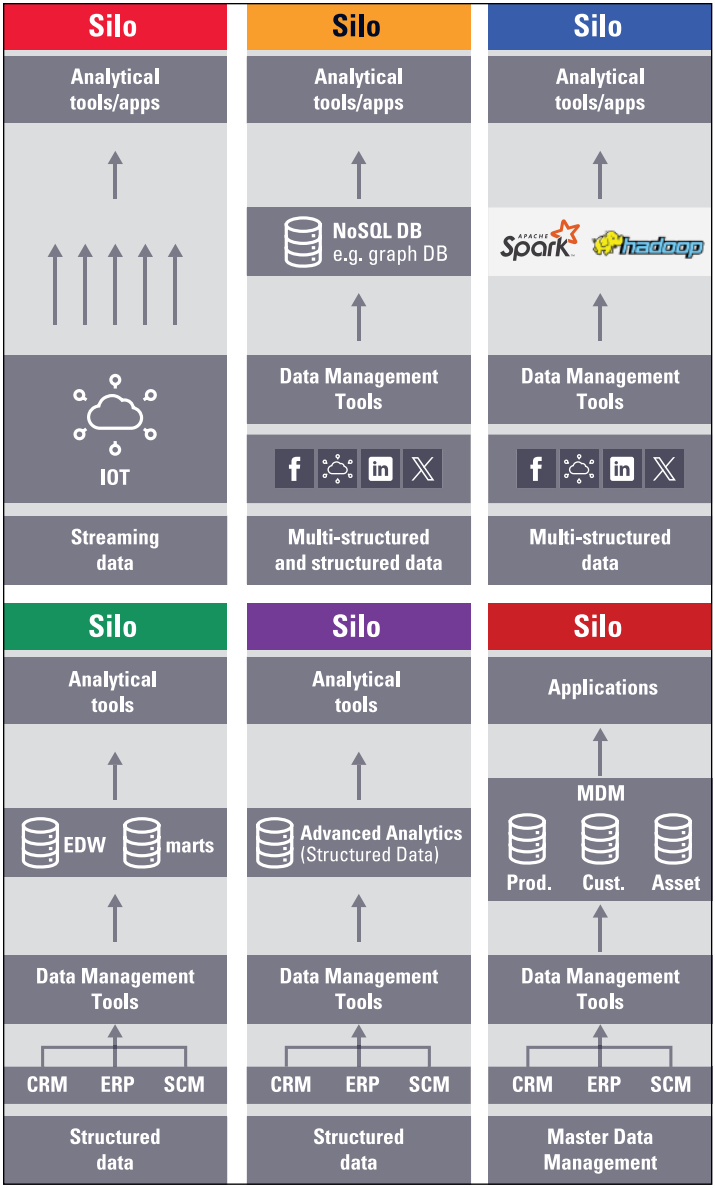


FIGURE 5-1: Siloed data that isn't shared across the business is ineffective data.

TIP

Flip back to Chapter 3 for more about managing a self-service data program.

4. **Enable self-service data quality.**

   When you show confidence in your business users, they'll respond to the program. They can quickly identify issues, assess their business impact, set priorities, and assign tasks to the right person for action.

5. **Measure, communicate, and improve continuously.**

   In short, once you're on this path, keep going. Don't forget to communicate your successes. With regular updates to all, you can always build the momentum for creating the data culture.

A company requires a healthy culture around the usage and management of its data to guarantee that data is trustworthy, reusable, and supportive of business operations. When you've built a data quality culture, you'll see

- Greater trust in business decisions
- Better and faster access to the relevant data needed to make decisions
- Compliant data use and storage
- Streamlined quality and governance processes
- A shared commitment to maintaining high-quality data

DATA CULTURE AND DATAOPS

The returns for those who fostered a data culture were made clear during the COVID-19 pandemic. Brick-and-mortar businesses had to speed up moves to digital delivery of customer service and engagement. That shift required massive data collection from alternative sources. Organizations that prepared for the data integration found opportunities. Those that didn't fell behind.

The improved efficiency with which organizations address data privacy regulation and compliance is another example of effective DataOps. Today, regulations, including the EU's General Data Protection Regulation (GDPR) laws and California's Consumer

(continued)

(continued)

Privacy Act (CCPA), require companies to deliver data to any individual who asks to see what the organization is holding. Fulfilling such requests takes time and money. Good data management practices at data-driven organizations with reliable DataOps manage, find, and deliver data much more efficiently.

When done correctly, DataOps is a catalyst for creating a data culture. The result removes the barriers between people and data technologies, creating a positive feedback loop that strengthens the entire organization.

Investing in Data Management

Investing in data management is the essential first step toward establishing a data culture. While data management is complex, it enables the seeds of the data culture to take root. The self-service business intelligence tools that are the hallmark of a data-driven organization are only as good as the data they integrate with. IT must address that need.

This means dealing with some significant hurdles. Integrating data while maintaining security is complex and tricky. Increasing regulations and changing expectations regarding personal data have set the bar very high. When an organization integrates disparate data sources, it needs to be mindful of the results. Two data sources, neither of which are sensitive on their own, can easily yield unintended inferences.

Bringing data from legacy systems will be challenging. Organizations have sizable investments in their IT infrastructure. A universal rip-and-replace approach is likely to be unacceptable, even where a clearly better solution exists. At the same time, today's data is everywhere, spread across applications and multiple clouds, and located on-premises and off. All these resources will need to work together.

Quality of data will also need to be addressed. An organization's "data past" needs to be confronted. You may expose a history of underinvestment in data architecture and management, but you won't replicate it.

Confronting these challenges today, head-on, will enable the self-service data culture of tomorrow. A good DataOps program reduces the friction that hindered self-service users in the past. It also eliminates headaches for IT department staff, who were the object of the resulting frustration. A good DataOps program is all but invisible to the end-users of the data.

REMEMBER

The data reliability life cycle is an organization-wide approach to continuously and actively improve data health and eliminating data quality issues by applying best practices of DevOps to data pipelines. If it sounds like a lot to handle, that's because it is. But it's worth it. Investments in addressing these issues creates a self-service data model in which decision makers will engage. Those users will be the first generation of an emergent data culture.

IN THIS CHAPTER

- Supporting data pipelines
- Understanding data architecture
- Grasping data governance
- Knowing pipeline orchestration and monitoring tools
- Digging into cloud computing

Chapter 6
Ten Content Areas and Tools for Data Reliability Engineers

Data reliability engineers need to know a lot of things. In this chapter, we break down ten sample areas (okay, you caught us; there are only nine). These content concepts and tools represent what data reliability engineers need in each area and why.

Cloud Infrastructure

If you're working in cloud computing, make sure you have experience with cloud services such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform and their data services (AWS S3 and Google BigQuery).

REMEMBER

Having experience with cloud services is crucial because these platforms offer strong, scalable, and secure environments for storing, processing, and managing data. They also provide tools and services to improve data reliability, such as automated backups, disaster recovery options, and advanced monitoring capabilities.

Data Platform Design and Scalability

When it comes to storing information in databases, data reliability engineers should be familiar with the following data stores:

- **Relational databases:** MySQL, PostgreSQL, SQL Server, Oracle
- **NoSQL databases:** MongoDB and Cassandra
- **Cloud Datawarehouse:** Snowflake, Google BigQuery, Amazon Redshift
- **Data lakes:** Azure Data Lake Services, S3, Databricks, Big Query Cloudera/Hadoop

TIP

Making yourself familiar with these data stores helps for following reasons:

- **Versatility in data management:** Different data stores are optimized for different use cases. Relational databases are suitable for complex queries and transactions, while NoSQL databases are better for handling large volumes of unstructured data.
- **Scalability and performance:** Each data store has unique strengths in scalability and performance. Knowing how to leverage these capabilities helps engineers design systems that efficiently handle growth and high demand.
- **Data integration:** Modern data ecosystems often involve integration across multiple data sources. Being familiar with various data stores allows engineers to seamlessly integrate data, ensuring consistency and reliability across the platform.
- **Cost efficiency:** Cloud data warehouses and data lakes offer scalable and cost-effective solutions for storing and processing large datasets. Understanding these options helps engineers choose the most cost-effective solution without compromising performance.
- **Data reliability and availability:** Knowledge of these data stores helps engineers implement robust backup, recovery, and disaster recovery strategies, ensuring high data availability and reliability critical for business operations.
- **Optimized data processing:** Different data processing needs may require different storage solutions. Familiarity with various data stores enables engineers to optimize data processing workflows, improving system efficiency.

By familiarizing with these data stores, engineers can design and implement scalable, reliable, and efficient data platforms that meet the diverse needs of modern enterprises.

Modern Data Architecture

If you're designing your data ecosystem, you need the following:

- **Data mesh:** Focuses on organizational change and enables domain teams to own the delivery of data products, understanding that the domain teams are closer to their data so they understand the data better
- **Data fabric:** Focuses on an intelligent data management architecture that unifies data from various sources, enhances data governance and security, and leverages metadata through artificial intelligence (AI) and machine learning (ML)

Flip back to Chapter 3 for more information.

Data Pipelines Tools

If you're working with data pipelines, you need the following skills and tools:

- **Proficiency in languages commonly used in data engineering, such as Python, Java, or Scala:** Essential for building, managing, and optimizing data pipelines
- **Data observability tools, such as Bigeye and Acceldata:** Helps ensure data quality and reliability by monitoring data flows, detecting anomalies, and providing visibility into data health
- **Data testing tools, such as Great Expectations and dbt tests:** Crucial for validating data quality and ensuring that data meets specified criteria before it is used in analytics and reporting
- **Skills in scripting for automation and data manipulation tasks:** Vital for automating repetitive tasks, transforming data, and integrating various data pipeline components, which increase efficiency and reduce the likelihood of human error, particularly in automation and data manipulation

- **Transformation tools, such as dbt:** Enables engineers to transform data within the data warehouse by using SQL-based transformations, which helps standardize data transformation processes, making them more reliable and maintainable
- **Reverse-ETL, such as Census and Hightouch:** Facilitates data movement from data warehouses to operational systems to ensure that actionable insights derived from data analytics are effectively utilized across the organization


REMEMBER

Understanding and mastering these skills and tools is crucial for data engineers because they ensure the creation and maintenance of robust, efficient, and reliable data pipelines. These capabilities enable engineers to handle the complexities of modern data ecosystems, ensure data quality, and deliver actionable insights to stakeholders.

Pipeline Orchestration Tools

If you're using pipeline orchestration, you need tools such as the following:

- Azure Data Factory
- AWS Glue
- Airflow
- Prefect
- Dagster


REMEMBER

By mastering these pipeline orchestration tools, data reliability engineers can build and maintain efficient, reliable, and scalable data workflows that meet the demands of modern data-driven enterprises.

Data Governance Knowledge and Tools

When it comes to regulating your data, you need the following:

- An understanding of encryption, access controls, and secure data transmission techniques

- An understanding of SRE fundamentals, including service-level agreements (SLAs), service-level objectives (SLOs), and service-level indicators (SLIs)
- Discovery and governance tools for managing data ownership and documentation, including quality metrics, data catalog, and lineage

With this knowledge, you can:

- Ensure data security and protect sensitive data through robust encryption, access controls, and secure transmission methods.
- Meet compliance requirements and adhere to regulatory standards and industry best practices for data security and governance.
- Optimize data reliability and use SRE principles to maintain high performance and reliability of data services.
- Enhance data management and leverage discovery and governance tools to manage data ownership, quality, and documentation effectively.

Monitoring Tools

If you're using certain monitoring tools, you need the following skills:

- Deploying monitoring and alerting systems such as Prometheus, Grafana, New Relic, or Datadog
- Scripting for automation and data manipulation tasks

REMEMBER

With these skills, you can ensure robust and efficient monitoring of your systems, enable proactive issue resolution, and maintain high performance and reliability.

Infrastructure Provisioning

In the area of infrastructure provisioning, you need the following:

- Familiarity with containerization tools such as Docker and orchestration platforms such as Kubernetes and Terraform
- Experience with CI/CD tools such as Jenkins or GitLab CI
- Analytics and ML tools used by their counterparts in analytics
- Experience with big data platforms such as Hadoop and Spark
- Knowledge of data streaming platforms like Kafka

REMEMBER

With these skills, professionals can ensure that the infrastructure is robust, scalable, and capable of supporting modern applications and data workflows efficiently and reliably.

Manage Services with the Hitachi Application Reliability Center

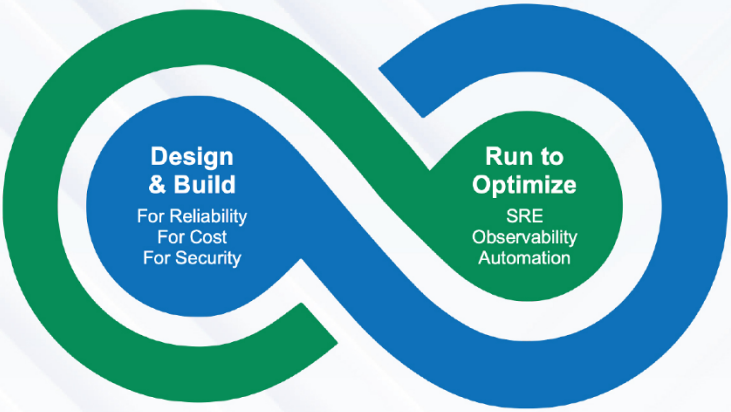
Hitachi Application Reliability Centers (HARC) combine cloud experts and industry-leading methodologies to help you chart your cloud journey. It ensures your application portfolios are always focused on what really matters — your business outcomes.

HARC offers comprehensive services to optimize resilience and cost for always-on business. Design, build, run, and operate workloads across private, public, hybrid, and multicloud environments.

Start your journey at `hitachids.com/service/application-reliability-centers`.

HITACHI
Inspire the Next

Hitachi Application Reliability Centers



Design, build, run, and operate your cloud workloads to run your resilient, cost-optimized business. Leverage best-in-class industry experts to deploy cutting-edge cloud modernization processes and technologies.

Engineering-led, Site Reliability Engineering-based cloud centers of excellence for all your cloud workloads.

HitachiDS.COM/AppReliability →

Hitachi Digital Services, a wholly owned subsidiary of Hitachi Ltd., is an edge-to-core digital consultancy and technology services provider helping organizations realize the full potential of AI-driven digital transformation. Through a technology-unified operating model for cloud, data, and IoT, Hitachi Digital Services' end-to-end value creation for clients is established through innovation in digital engineering, implementation services, products, and solutions. Built on Hitachi Group's more than 110 years of innovation across industries, Hitachi Digital Services helps to improve people's lives today and build a sustainable world tomorrow.

Data reliability engineering: SRE for data

This book is a comprehensive resource on data reliability engineering (DRE) — a practical approach similar to site reliability engineering (SRE) that equips businesses with the confidence to make informed decisions. Gain insights into how DRE ensures data integrity amid increasing volume and complexity, utilizing tools like data lineage and observability frameworks. Moreover, discover the advantages of self-healing technologies and self-service data management and fostering a data-driven culture.

Inside...

- DRE — the SRE for data
- Managing data quality and downtime
- Journey toward data reliability
- A proactive approach to data quality
- Advancing data management
- Essential areas and tools for DRE

Hitachi Digital Services

Go to Dummies.com™
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-394-20298-0
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.