HUAWEI

Striding Towards the Intelligent World White Paper 2024

Data Storage

Data Is the Key to Unlocking the Digital and Intelligent Future

Building a Fully Connected, Intelligent World

FOREWORD

Humans have been walking the earth for hundreds of thousands of years, but recorded civilization has only existed for a fraction of that time. The key to the development of civilization was the emergence of paper. Paper enabled human experience and knowledge to be recorded, shared widely, and passed down to future generations, and provided a foundation for learning and the evolution and development of ideas. This has contributed to the rapid development of human civilization. It's worth mentioning the forms that this paper could take. In medieval Europe, important words were recorded on parchment, and the production of a book might require hundreds of sheets of sheepskin. This made the benefits of paper prohibitively expensive for ordinary people and hindered the preservation and spread of knowledge.

The invention of digital means of storing and transferring information marked an epochal shift in human history. The word "data" entered the lexicon as a way of referring to information in a digital format. The efficiency with which digital data can be processed is orders of magnitude beyond that of ink on paper, and this has resulted in explosive data growth. Modern data storage media belongs to the same lineage as paper, and data storage is catalyzing the next advances of human civilization, just as paper did in the past. By creating, honing, and innovating data storage, we have become the "papermakers" of the digital world.

There is no AI without sufficient data. As large AI models mature and move towards multi-modality, data is gradually becoming the key to AI development. This is because AI uses computers to simulate human modes of thinking, infer rules from massive amounts of data, and summarize knowledge. We can feed large AI models with the knowledge of different application scenarios, to generate bots that serve as consultants, programmers, and customer support and that have the capacity to learn and evolve independently. The decisive factor of staying ahead in the AI era is data. Specifically, it's how data is generated, stored, and used.

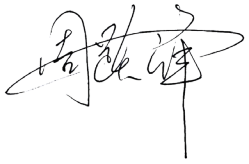
For over 15 years, Huawei has invested heavily in data storage, to produce a premium portfolio of cutting-edge offerings for 26,000 customers in over 150 countries and regions worldwide, for sectors like finance, carriers, government, manufacturing, electric power, energy, healthcare, scientific research & education, and transportation. The *Striding Towards the Intelligent World 2024 White Paper—Data Storage* was

made possible through extensive communication with industry experts, customers, and partners. This paper delves into today's hot topics of digital and intelligent transformation, provides viewpoints about data storage trends and challenges across industries, and offers suggestions for future action. I believe this very meaningful research will bring together more industry forces to drive the data storage industry forward.

Over the past three decades, new technologies and applications have been emerging and generating massive amounts of data. The right data storage, like a cozy "home" for data, lays a solid data foundation that underpins and drives the ongoing growth of new technologies and applications. Huawei Data Storage Product Line is looking forward to working closely with all parties across various industries and making concerted efforts to provide future-proof storage power for new technologies and applications, and create a better future for data storage.

Dr. Peter Zhou

President, Huawei Data Storage Product Line



CONTENTS

Executive Summary

Scaling laws reveal the impact that computing power and the volume of data used to train AI have on AI performance in the current deep learning algorithm framework: the more computing power and effective training data, the better the performance of large AI models. The expanded training resources have led large AI models to evolve from being uni-modal to being multi-modal. As a result, the capabilities and performance of large AI models are going from strength to strength. This has helped push AI beyond central training and into commercialized deployment. AI is already being used in several industries, for things like office assistance, production decision-making, cost reduction and efficiency improvements, management, and the prediction of future trends. AI has also been applied in scenarios requiring higher fault tolerance than before. The commercial application of AI has set off a wave of intelligent transformation and service upgrades across industries. Throughout this process, organizations have come to realize that deepening and accelerating service digitalization to generate large volumes of diverse and valuable data is as important as awakening dormant historical data. Digitalization and intelligentization, which are both bound by data, come together in the form of "digital and intelligent transformation" to enable ever-evolving data infrastructure to meet our ever-increasing service requirements.

Digital and intelligent transformation is expected to become more common. It will play an important role in enabling artificial general intelligence and bringing us closer to a new and intelligent world. This report offers the following insights into current trends and the essential features that data infrastructure needs to have to effectively support digital and intelligent transformation:

1. Large AI models are becoming multi-modal, and the size of computing clusters and data is likely to continue to increase. To streamline system management and unleash the full potential of AI in service scenarios, computing power and storage power need to increase concurrently and the ratio of one to the other needs to remain flexible to adapt to developments in AI.

2. As the size of AI computing clusters increases, interruptions in large-AI-model training are becoming more frequent. This makes more frequent checkpointing and resumable training essential. To achieve this, data access performance must be improved so that checkpoint saving and loading can be done more quickly.

It is also important to note that intelligentization is accelerating digitalization, and as a result, more service data is being generated, which complicates data processing and puts more pressure on data infrastructure.

3. While intelligentization boosts digitalization, enabling the generation of more valuable service data, it also comes with the risk of more frequent ransomware attacks.

4. The increase in the size of AI computing clusters has made the ability to efficiently manage massive amounts of multi-source heterogeneous data a key competitive advantage in AI development. Data map drawing, data ingestion, and data preprocessing should be the main focal points during large-AI-model training.

5. Many industries have experienced difficulties related to infrastructure deployment, large model selection, secondary training, and supervised fine-tuning when deploying AI. Leveraging the capabilities of infrastructure and large-AI-model vendors is the key to accelerating AI deployment.

A new data infrastructure architecture is emerging to meet the needs of large AI models and other new enterprise-level intelligent applications. To optimize data infrastructure for the large-AI-model era, we suggest that enterprises:

1. Use the decoupled storage-compute architecture, which features superb flexibility and independent expansion, to simplify AI computing cluster management and enable on-demand scaling of compute and storage resources. In addition, ensure that data infrastructure is equipped with scale-out, linear performance expansion, and multi-protocol interworking capabilities, as these are essential for success in the digital-intelligent era.

2. Switch to all-flash storage to improve data processing efficiency and meet service requirements in the digital-intelligent era, because it is able to meet the ever-evolving requirements for both digitalization and intelligentization. In addition, leverage emerging data paradigms such as vector retrieval-augmented generation (RAG) and contextual long-term memory storage to simplify data access and enhance computing with storage, thus improving the overall system performance.

3. Build a data resilience system that has both defense and response mechanisms for data-generating digitalization and ever-evolving intelligentization to transition from reactive responses to attacks to proactive and comprehensive protection.

4. Build an AI data lake foundation for AI computing clusters to eliminate data silos and make data visible, manageable, and available.

5. Use training/inference appliances that come with pre-installed infrastructure and tool software to facilitate the deployment of large AI models in various industry scenarios. Moreover, leverage large-AI-model vendors' system integration capabilities to accelerate AI deployment and application.

01

Shifting Gears from Digitalization to Digital and Intelligent Transformation

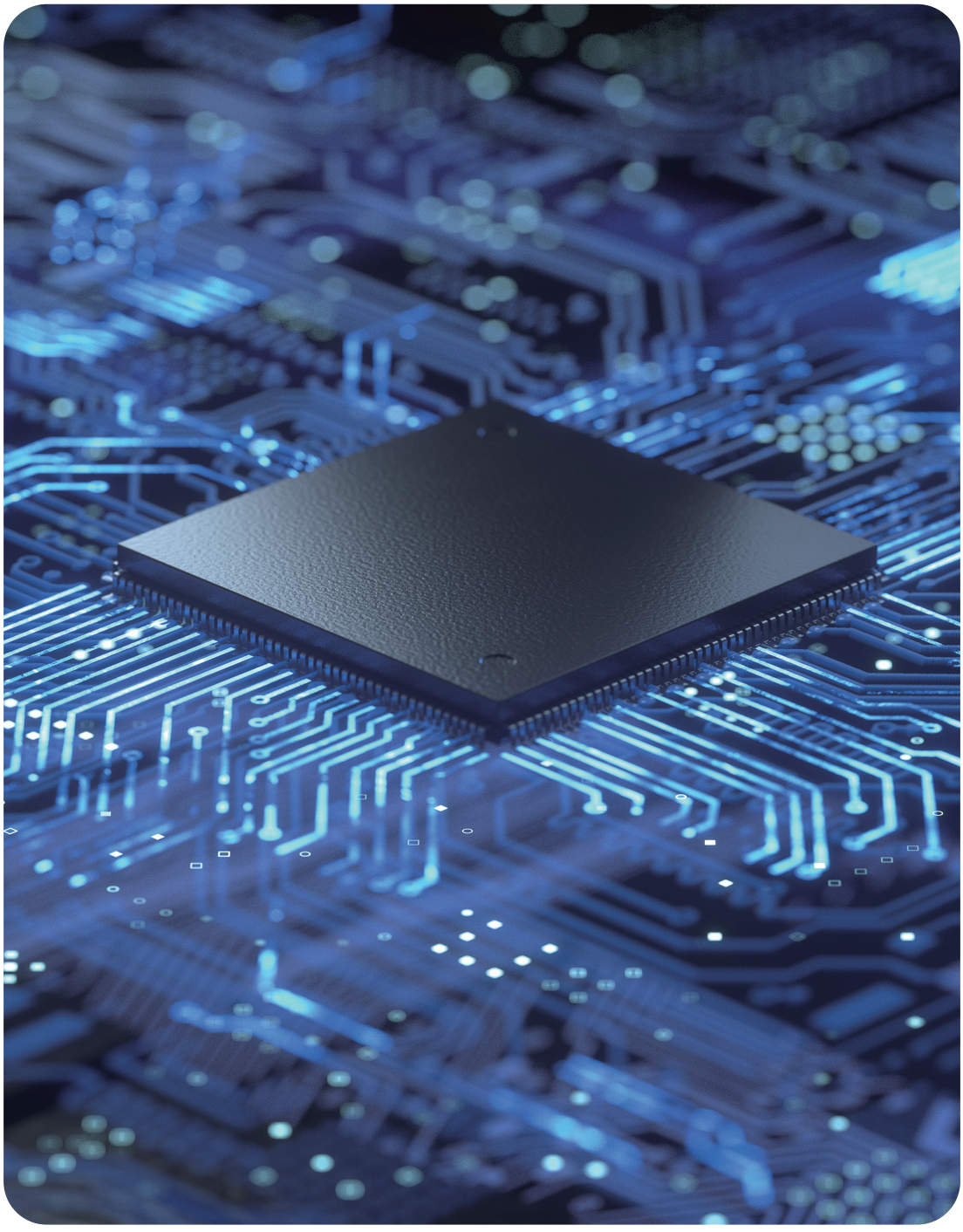
Artificial intelligence was established as a discipline in 1956. For most of the last 70 years of exploration and development it's been an obscure field, but today AI and large AI models are finally booming. Deep learning algorithms and large-scale computing power have made it possible to train AI models on massive amounts of data. Today, as the capabilities and performance of large AI models go from strength to strength, large AI models are gradually shifting from foundation models with central training to more specialized models optimized for use in particular industries, and this has triggered a wave of intelligent transformation and service redesign across industries.

In the era of large AI models, computing power, algorithms, and data constitute the three elements of large model training. Scaling laws reveal the relationship between model performance and factors like computing power and data in the current deep learning algorithm framework: the greater amount of computing power and effective training data, the better AI models you'll get.

Driven by the imperative to scale up, organizations are building larger compute clusters while striving to obtain more data for use in model training. What is emerging from this process is a shift from uni-modal to multi-modal models, and the commercial application of these models in certain consumer-oriented scenarios. For example, the AI-enabled tablets are already on the market, and have been enthusiastically received, thanks to AI functions such as image and text recognition, conference recording transcription, translation, and copywriting.

Unlike consumers, industry users are focused on how large AI models can benefit their business, improve internal operations, and enhance competitiveness. Some have already found fruitful avenues for AI application, such as customer services in call centers, diagnosis and treatment in hospitals, online situated learning, copywriting for advertising, quality inspection in industrial production, and intelligent O&M and autonomous driving for complex networks, and further application scenarios are still being explored. Early experiences have demonstrated that high-quality industry data is indispensable for the practical use of AI in industries. First, it is important for generating industry- and scenario-specific models. For example, you need a certain amount of industry data to perform secondary training and supervised fine-tuning on foundation models to obtain a vertical model oriented to a specific industry. Furthermore, the knowledge bases used to eliminate hallucinations during inference need to be generated from high-quality and up-to-date industry data.

From training foundation models to optimizing models for use in different industries, the quantity and quality of training data determine how far AI can evolve and the extent to which AI applications will be adopted in each industry.



1.1 Finance

The financial industry has led the way in the digital era and pioneered FinTech. The industry is now deeply integrated with large AI models, leveraging vast amounts of data assets from digitalization to gain a first-mover advantage in the digital and intelligent era. Consider banks as an example. AI-driven banks are gradually shifting their focus from office assistance scenarios like intelligent ticket filling and office assistants to more important production scenarios like remote banking and credit risk control assistants. Making the transition from internal office assistance to external service applications requires more powerful fault tolerance, as accurate suggestions and choices must be derived from vast amounts of data. Therefore, efficient collection, quick processing, and high resilience and reliability of mass data have all become new challenges and requirements.

1.1.1 From cost reduction to efficiency improvement: office assistance → service applications

Financial institutions have always been the first to apply emerging IT technologies to service scenarios. Leading firms are already investing in AI, particularly in R&D and the layout of large AI models, in a bid to enhance business operations, product marketing, risk control, and customer services, thereby improving the intelligence of financial services. According to an IDC report, 90% of banks have started exploring the application of AI, making it a key focus for technological banking innovation.

1. In intelligent marketing scenarios, AI technologies analyze vast amounts of user data to deliver personalized financial services that are tailored to customer needs and preferences. This not only improves user experience, but boosts customer loyalty. For example, China's Bank of Communications uses AI to explore customer preferences and enhance customer retention through large AI models. The total transaction volume of wealth management products reaches nearly CNY400 billion by using various wealth management model strategies, 16 times higher than the transactions using traditional methods.

2 In intelligent wealth management scenarios, AI technologies like machine learning and deep learning models are helping investors make more accurate investment decisions. Agricultural Bank of China (Jiangsu Branch) and Industrial and Commercial Bank of China (ICBC) have launched the ChatGPT-like large AI model ChatABC and Ascend AI-based financial foundation model, respectively, to intelligently recommend wealth management products. Meanwhile, Shanghai Pudong Development Bank employs technologies like multi-modal human-machine interaction and knowledge graphs to launch AI "wealth management experts" that suggest suitable products to consumers.

3 In the risk control scenario of credit approval, AI simplifies and optimizes the process from credit decision making to quantitative transactions and financial risk management. One leading bank in the Asia-Pacific region has reduced credit application times from several days to just one minute, with approvals possible in one second.

4 Intelligent customer services are highly effective when applied to financial services. For instance, China Merchants Bank (CMB) uses intelligent customer services to facilitate over 2 million daily online interactions while resolving 99% of user issues. The AI-driven services not only boost efficiency compared with traditional human methods, but offer 24/7 support.

1.1.2 Enhancing multi-source, mass data management and strengthening data resilience compliance

As AI applications become more widespread, new challenges are being posed to financial institutions in terms of data architecture, data resilience, and service continuity, as expanded on below:

1 First up is the management of a large amount of data. The financial industry has now reached the Exabyte (EB) level of data. Take financial intuitions in China as an example. According to a report from Beijing's Financial Information and Technology Institute (FITI) in 2023, financial institutions generally reached the Petabyte (PB) level of data, among which the data volume of large institutions exceeded 100 PB. Furthermore, the average annual growth rate was projected to reach 24.33% over the next five years. In addition, large state-owned banks

have core service systems with storage capacity in the hundreds of PB, while non-core systems, like document images, can reach dozens or even hundreds of PB. Financial institutions must now consider how to ensure highly reliable and efficient access to this vast, diversified service data in order to maximize its value. For example, the institutions will need to invest in high-performance storage devices and optimize storage architectures so as to further integrate AI within the financial sector.

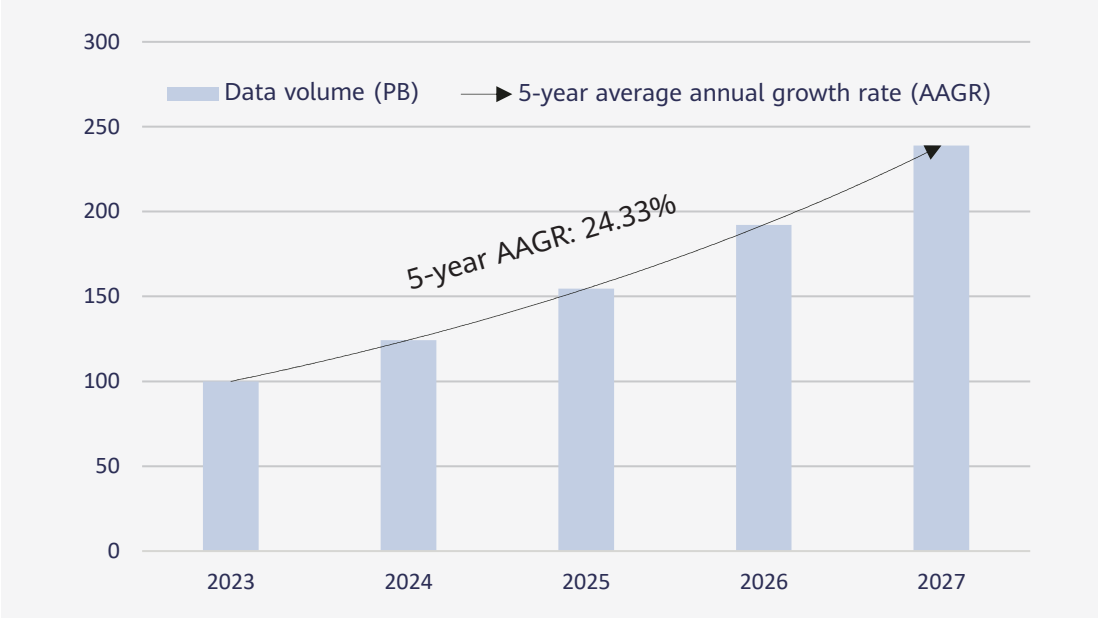


Figure 1: Average annual growth rate of data of a large financial institution

2 Financial services need to handle diverse data types. Over years of accumulation, financial data—including images, videos, audio, and internet logs—has become outdated, complex, and scattered across various service areas and regions. For example, the data formats of core systems of mainframes and minicomputers are incompatible with those of credit card systems on open platforms. This makes it challenging to share user information among credit services, wealth management, and internet services. Integrating such scattered data for AI applications is also a challenge, highlighting the urgent need for a comprehensive data management system. For example, one top bank in China views data as a basic element and strategic resource. When this bank looked to establish a big data resource management system, it faced the following key issues: what data is available, where is the data located, and how can the data be used effectively.

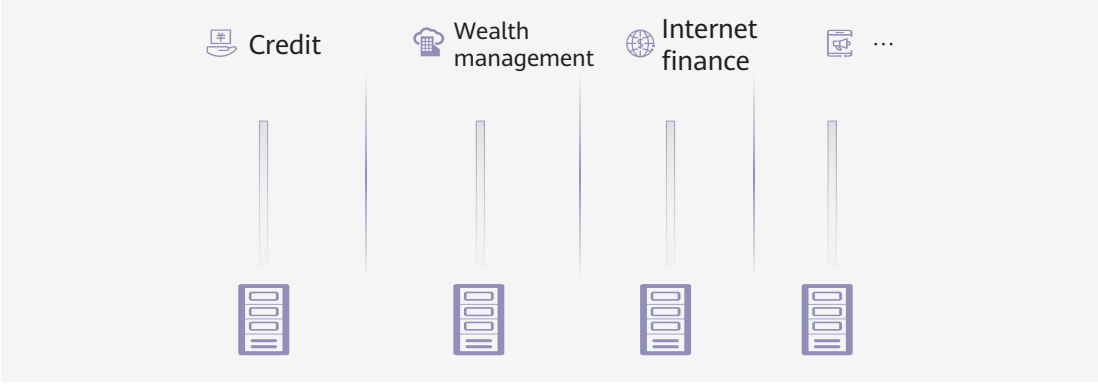


Figure 2: Difficulties in sharing data among different services

3 Data processing in the financial industry must meet the compliance requirements of industry supervision and risk control. Precision marketing that uses AI to provide personalized recommendations and targeted advertisements presents bigger challenges regarding data management and privacy protection, while also raising the bar for financial compliance supervision. At the same time, AI applications increase the risk of data breaches in financial institutions. In May 2024, a well-known US bank suffered a LockBit ransomware attack, compromising the data of approximately one million customers. In June 2024, the National Financial Regulatory Administration of China published on its website that a top bank in China had been fined millions of CNY for inadequate data resilience and disaster recovery management. This demonstrates the importance of methods like disaster recovery-based physical resilience and backup-based logical resilience in today's AI era.

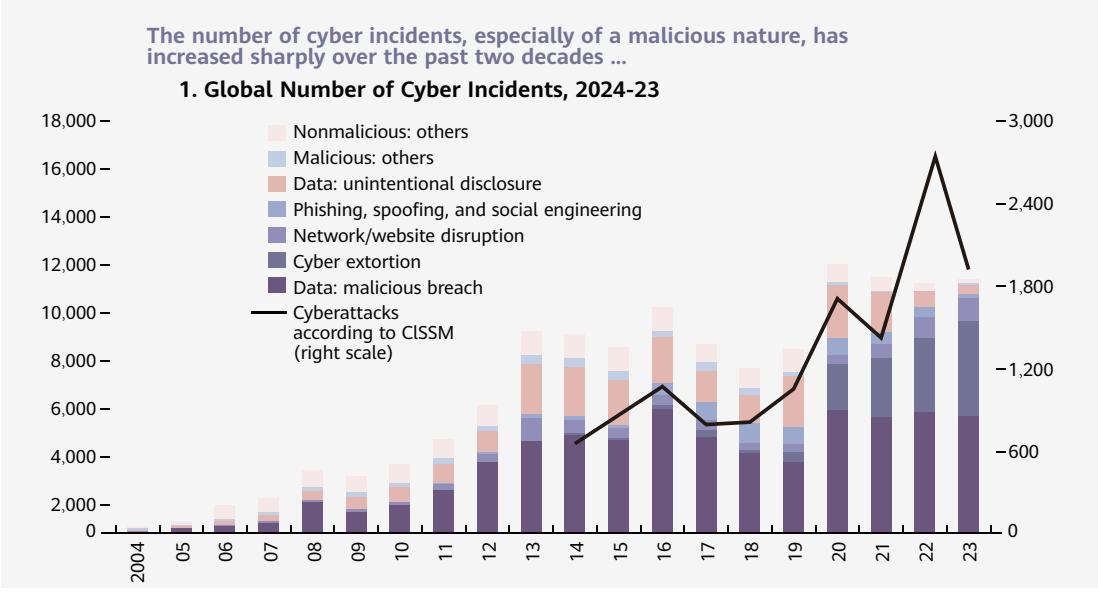


Figure 3: Increasing AI and digital applications increase cyber resilience risks

The Global Financial Stability Report released by the International Monetary Fund (IMF) points out that the rise of AI and digital applications significantly increases cyber resilience risks.

While financial institutions embrace AI technologies to redefine service models and unlock data value, they must also address the challenges AI poses to data management in order to effectively enhance the efficiency and quality of their services.

1.2 Carriers

"Transforming from telco to techco" has become a strategic consensus for most carriers around the world during digital transformation. As GenAI evolves, carriers possess inherent communications advantages in resources, data, and industry expertise, putting them in a position to support AI development and lead in AI application implementation.

1.2.1 From development to application: building up strength for large AI model training and inference to improve internal operation efficiency and enable a wide range of industries externally

So far, global carriers have formed into three waves of AI initiatives. The first wave includes intelligence pioneers like SK Telecom in South Korea and China Mobile, who are developing full-stack AI capabilities across terminals, computing resources, and model applications. The second wave consists of carriers like Singapore's Singtel, Deutsche Telekom, and UAE's e& which have established the Global Telco AI Alliance (GTAA) to create multi-language telecom models for call centers and smart operations. The third wave features pragmatic carriers like Orange and Vodafone who focus on leveraging third-party AI capabilities to enhance efficiency and reduce costs.

In the next two to three years, AI will reshape most carrier applications and services. According to a report by Valuates, the global AI in telecommunication market size is projected to reach US$19.17 billion by 2028, at a compound annual growth rate (CAGR) of 40.6% during 2022-2028. GenAI benefits carriers in two key ways:

1. **Improves service efficiency through the combination of AI applications and carriers' existing services**

AI-powered analysis, policy optimization, and prediction capabilities enhance service systems like network elements (NEs) and networks, improving the intelligent planning, deployment, O&M, and management of telecom networks, and ultimately enabling networks to support L4/L5 autonomous driving. For example, the AI voice robot of telecom company KT in South Korea provides functions like real-time automatic summary, cutting response time to customer requests from 20 seconds to just 5. In addition, China Mobile's anti-fraud system intercepts more than 14 million calls every

month, with 98% accuracy.

2 **Provides external enablement for industry-university-research-user to promote intelligent upgrade**

Carries can use GenAI to directly offer intelligent computing services for large AI model enterprises or educational research institutions, effectively becoming the "shovel sellers" of the AI "gold rush". Additionally, carriers can extend their large AI model capabilities by launching industry AI model applications to industry sectors like public sector, education, and healthcare. For example, China Mobile's Jiutian Large AI Model, collaborating with public sector agencies, has supported the development of a smart public service assistant for Gansu. This assistant creates government knowledge graphs that associate 10 million services, and provides 1 million standard Q&As, delivering efficient digital and intelligent services to 25 million residents in the province.

1.2.2 Activating mass data to facilitate the efficient training and the implementation of large AI models across industries

To grasp the opportunities presented by large AI models, carriers must establish AI-ready infrastructure. But to be AI-ready, carriers must first be data-ready. As AI clusters grow and enter the era of 10,000-level NPUs, the effectiveness of substantial investments will face two key challenges:

Challenge 1: How can carriers' data assets be revitalized, and how can the large AI models better serve their own services?

In the era of digitalization and intelligence, data has emerged as the fifth factor of production—alongside land, labor, capital, and technology—and serves as the core driving force behind the digital economy's evolution. In particular, without sufficient high-quality data, the learning capabilities of these models will significantly diminish. For instance, China Mobile aims to achieve L4 autonomy across its network by 2025, which will require data from over 6 million 4G/5G base stations and 990 million users, along with the national "4+N+31+X" data centers. Currently, the core data volume is 650 PB, with at least 5 PB being generated daily. To achieve the capabilities required for L4 autonomous driving—such as intelligent energy-saving base stations, antenna weight optimization, intelligent complaint management, and network expense auditing—high-value data scattered across provinces, users, and applications must be effectively

organized. This will provide ongoing fuel for large AI models and provide informed strategies on the planning, deployment, O&M, optimization, and operations related to building an L4 network.

Challenge 2: How can carriers reduce AI development and operation costs and expand application at enterprise edge scenarios to accelerate closed-loop AI business?

AI clusters require a substantial amount of investment and consume large amounts of energy. A single round of GPT-3 training consumes the same amount of energy as 300 households consume in one year—which has roughly the same environmental impact as 500 tons of carbon dioxide emissions. One round of Sora training even consumes 1,000 times more power than that. Low AI cluster utilization contributes to these high power and compute infrastructure costs, and ultimately drives up both construction and operation costs. Carriers need to properly plan an intelligent computing foundation that moves away from stacking computing power towards fully unlocking its full potential. Proper configuration of storage cluster performance and high-performance, high-reliability external storage can boost AI cluster utilization.

Edge applications are a key area where GenAI can create profits, especially in the ToB market. Scenarios like self-service medical consultations, industrial manufacturing quality inspection, intelligent financial customer services, and public service assistants are all primed for integrated training/inference solutions that converge private knowledge repositories, GPUs, Retrieval-Augmented Generation (RAG), and scenario-specific large models. Carriers can use one-stop training/inference HCI appliances to quickly develop and launch their AI products to monetize their large AI models. This will bridge the last mile of applying large AI models to real-world tasks. For example, China Mobile's Jiutian HCI Appliance provides out-of-the-box large model services for industry users. It is deployed with a 13.9-billion-parameter large language model and a 1-billion-parameter large vision model. For coal mine customers, this has made production safer and more manageable by enabling multiple AI-powered capabilities like device checks, coal pile-up detection on conveyer belts, foreign object detection, coal volume detection, and personnel violation detection.

1.3 Public Services

AI applications are being rolled out in public service scenarios to improve the management efficiency and risk analysis of public service organizations and enhance user experience. Some of these scenarios include entry and exit management, tax supervision, and public service Q&A assistants.

However, AI adoption in public service governance currently faces a number of risks and challenges such as the difficulty of real-time data sharing, historical data activation, and sensitive data protection.

1.3.1 From services to governance: enhancing efficiency and governance in public services

Oxford Insights' *Government AI Readiness Index 2023* report assessed the AI readiness of national and regional governments around the world to leverage AI for public services. The assessment looked at 42 indicators across 10 dimensions including vision, governance and ethics, and digital capacity. Data is a key factor that drives AI evolution in this case. The total amount of language-related data is eight times that of image data. Data is primarily used for customer service systems, approval systems, and AI-assisted analysis and decision-making. The disparity in data collection, data application, and data resilience is especially evident between high-income countries and low-income countries, highlighting the global digital divide. In the 2023 rankings, the United States ranked first in terms of AI readiness in public services, followed by Singapore and the United Kingdom. China ranked sixteenth. Many countries around the world are seizing on the major opportunities of AI and actively addressing social, economic, technological, and policy challenges related to the implementation of AI in public services. This can be seen in their emphasis on national strategy plans that can bring opportunities for transformation in different sectors of society.

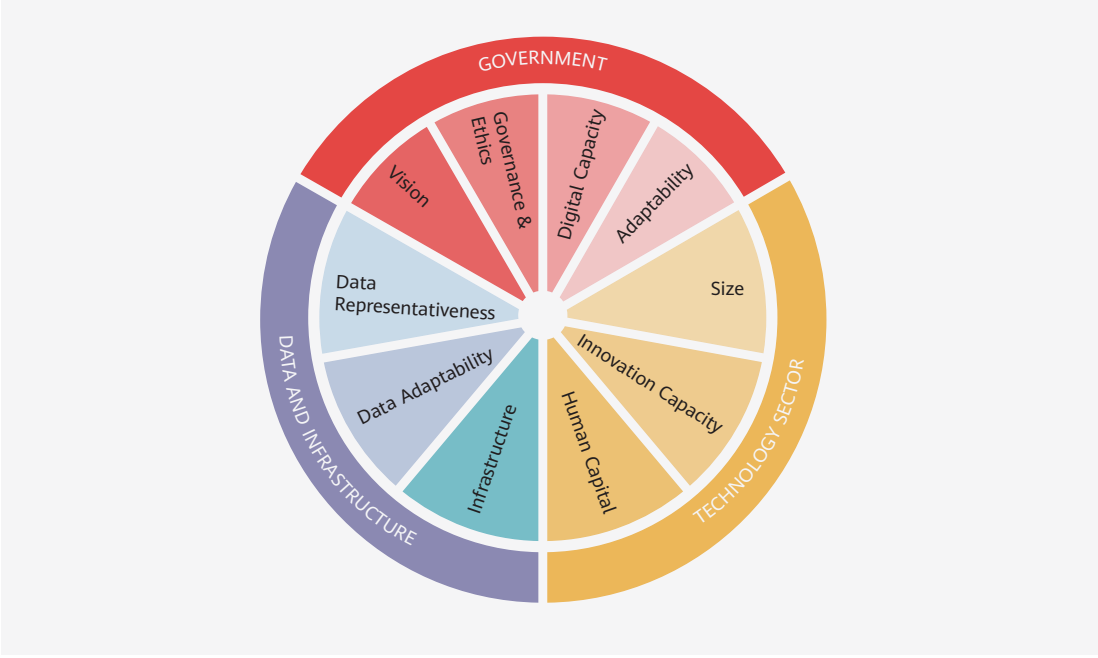


Figure 4: The Pillars of the Government AI Readiness Index

AI is being increasingly adopted in the following public services:

1 Entry and exit management

AI can swiftly process and analyze vast amounts of entry and exit data and enable automatic identity authentication, intelligent risk assessment, real-time data analysis, and migration trend prediction. It can also optimize resource allocation and simplify review processes. For example, AI can speed up passenger identity verification using biometric recognition technologies, reducing both the time required and the likelihood of errors during manual reviews. In addition, it can predict potential resilience threats by analyzing a large amount of entry and exit data, allowing management departments to take proactive mitigation measures. Such intelligent management measures not only improve work efficiency, but also enhance resilience and user experience.

2 Tax system

AI can boost tax management efficiency and accuracy, enabling tax authorities to better access automatic data processing, intelligent data analysis, and AI-powered risk assessment. By leveraging natural language processing to automatically parse tax files and extract key information, AI can also reduce both the time required for and

the likelihood of errors in manual reviews. Furthermore, it can identify potential tax risks by analyzing historical tax data in conjunction with other types of data, helping tax personnel address potential issues earlier. For example, AI can quickly assess the possibility of tax underreporting by comparing real estate companies' transaction data with actual tax filings and incorporating standard cost data for materials such as cement and steel from the construction industry.

3 Public service Q&A assistants

Extensive consulting services are required, as each service department of public organizations frequently has inquiries about policy communication, rule compliance, and specific cases. With natural language processing (NLP) and machine learning technologies, public service Q&A assistants can rapidly understand and answer a wide range of questions around the clock. They are also able to respond through multiple channels, such as government websites, WeChat official account (a platform that allows businesses to publish a variety of content types), and apps to provide policy interpretations, service guidance, and FAQs. This not only reduces the workload of human customer service staff, but also enables more convenient and accurate information access.

1.3.2 Jointly streamlining cross-department data and protecting sensitive data to enhance public services

While the application of AI in public service governance can significantly improve efficiency and service quality, it also creates several new risks and challenges. One such challenge is in the area of real-time data sharing, which demands fast, accurate transmission preventing the formation of data silos. Historical data is also difficult to activate and utilize for a number of reasons, including inconsistent data formats and the sheer volume of most historical data. Additionally, sensitive data protection is very valuable, so encryption technologies and permission management measures must be used to ensure data resilience during transmission and storage.

1 Real-time data sharing

The incorporation of more advanced data sharing and information exchange in China's social credit system is expected to improve trust in the system while improving credit evaluation and oversight for multiple entities, including government departments, enterprises, and individuals. For enterprises, AI can provide their operation status, tax records, and environmental inspection results in real time, detect abnormal behaviors, and issue warnings accordingly. For individuals, AI can assist in developing personalized loan repayment plans based on their personal financials and repayment capabilities. This would help individuals manage debt more effectively, avoid overdue payments, and improve their credit ratings. These AI-powered applications not only improve the efficiency and accuracy of the social credit system, but also promote transparency and fairness. In this regard, data sharing and information exchange are critical. This poses higher requirements on data storage, which shall provide following capabilities to ensure efficient data management.

·Data visibility: Data asset owners and managers should have access to a comprehensive data map that outlines data types, storage locations, and data volumes.

·Data manageability: A mechanism is required to implement policy-based data flow once the data to be aggregated is identified.

·Data availability: Raw data needs to be preprocessed and converted into data that can be identified and directly used by AI.

2 Historical data activation

Public service agencies worldwide are continuously mining value from historical data to improve services. One such example is tax agencies, who are actively leveraging the activation of historical tax data and applying it to AI to significantly enhance the intelligence level of tax management and decision-making in the following aspects.

·Policy formulation: Economic development and tax bases often vary across regions, and so AI can analyze the effect a single policy can have on multiple regions. This enables governments to formulate more targeted regional tax policies.

·Policy effect evaluation: AI can evaluate changes in tax revenue before and after the implementation of a tax policy by analyzing historical tax data for the past five years. These changes can show whether a tax cut boosts economic growth, or increases or decreases tax revenue.

·**Policy outcome prediction:** Through analysis of related historic data, AI can also predict the impact of potential tax incentive policies on investment and development in specific industries over the next few years.

The amount of historical data AI training has proven to need poses stringent read speed requirements on data storage. To support fast training and real-time inference for large AI models, storage systems must provide an ultra-high read speed for fast data access and mass data processing. This requires not only high-performance storage devices like non-volatile memory express (NVMe) SSD storage, but also optimized data management and cache policies to ensure that data can be read and used as quickly as possible.

3 Sensitive data protection

The public service sector involves mass sensitive data. One such example is entry and exit management, which processes typically sensitive data, including personal identity information (such as name, date of birth, and passport number), biometric data (such as fingerprints and iris), travel records (such as entry and exit times, locations, and flight information), and visa information. Although AI technologies make data processing and analysis more efficient, they also introduce new risks for data leaks and abuse. This is particularly concerning during cross-border data transmission, where sensitive information may be exploited by criminals. Public data management systems and technical solutions are needed to combat these risks. Data storage is critical in this regard, as it can act as the last line of defense for data resilience. The must-have aspects it can provide defense are listed as follows:

·**Data encryption:** Encrypts all of the stored data to prevent unauthorized access and data leakage.

·**Access control:** Strictly controls data access to ensure that only authorized personnel can access sensitive data.

·**Data backup:** Periodically backs up data to ensure that data can be quickly restored when it is lost or damaged.

·**Log recording:** Records data access and operation logs for tracing and auditing if resilience events occur.

·**Data isolation:** Isolates sensitive data from other data to reduce the risk of data leaks.

1.4 Manufacturing

In smart manufacturing, AI is being applied to computer-aided design (CAD), demand forecasting, intelligent production scheduling, predictive maintenance, and decision-making support to improve production efficiency and product quality. Meanwhile, many manufacturers are struggling to apply new AI solutions due to challenges including surging data volumes, historical data aggregation, data cleaning, and data labeling during data collection and analysis.

1.4.1 From partial to E2E: boosting efficiency across design, production, operation, and after-sales

The rapid advancement of science and technology has spawned a wider application of AI technologies in the manufacturing industry, ranging from basic after-sales robots to every phase of the production process, greatly improving production efficiency and product quality.

1 AI-assisted CAD

Most manufacturing enterprises widely use CAD during their product design phase. CAD is used for appearance design, component design, mechanical part design, and mold design. Accurate drawings, modeling, and simulation in this phase can speed up production and delivery during the subsequent production phase. In the past, without AI, CAD was conducted by experienced employees, and then the designs had to undergo review and proofing. This process was time-consuming, labor-intensive, and prone to error. AI has changed this process immensely. In the design phase, AI tools can automatically generate CAD system design solutions and new designs can be quickly formed based on previous best practices. The tools also perform multi-phase parallel design, shortening the design period.

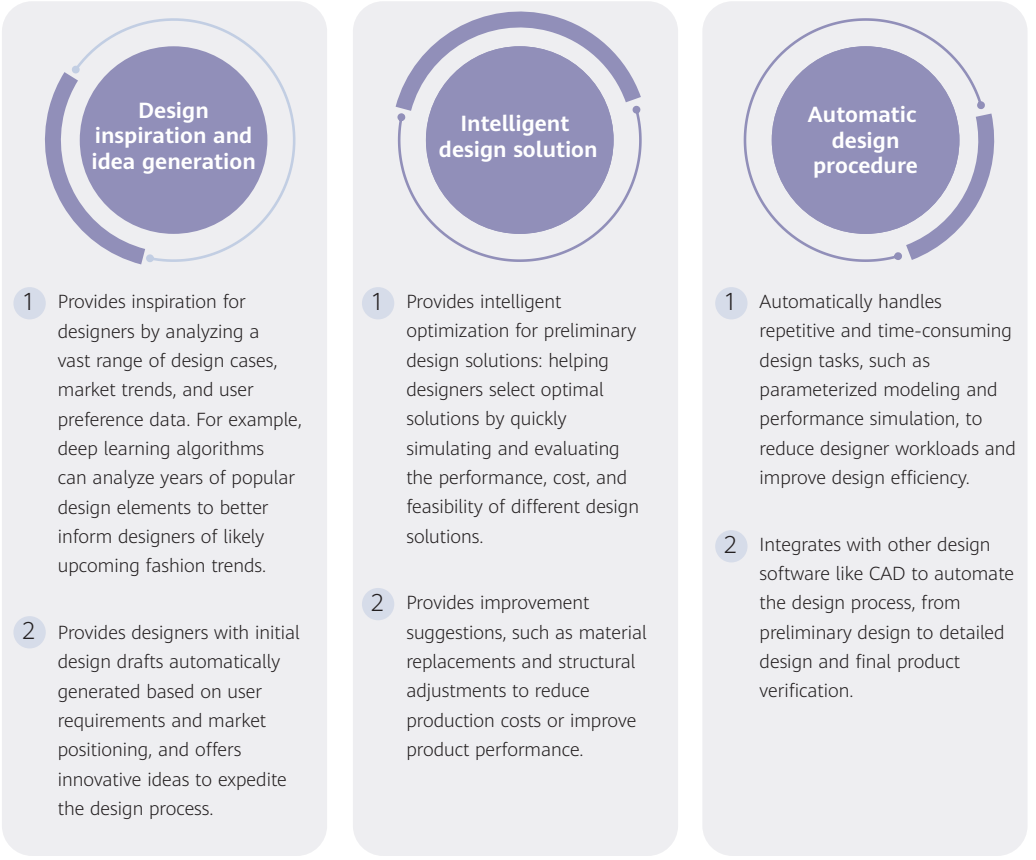


Figure 5: AI-assisted CAD

2 AI-assisted forecasting and smart production scheduling

Throughout the year, manufacturing enterprises have peak and off-peak sales periods, which affect how multiple departments deal with procurement, production, warehousing, and supply. In the past, production schedules relied solely on sales forecasts, but any inaccurate forecasts could significantly impact the entire manufacturing line. In the AI era, AI can help predict product demands across the year by analyzing factors such as historical sales data, supply chains, and market prices, and produce accurate plans and resource allocation that optimize inventory levels, reduce production and logistics costs, and lower production delays and material waste. For example, a leading semiconductor display manufacturer used AI technologies for their automation and intelligent transformation of the production process. By analyzing historical production data and that collected from devices, environments, and products, the manufacturer's efficiency and product quality were significantly improved and production costs were slashed. This intelligent transformation helped the company maintain its leading position in the semiconductor display industry.

3 AI-assisted predictive maintenance during production

Manufacturing equipment will inevitably malfunction or fail, and if this happens during production, such events can cause expensive downtime. Scheduled maintenance, often carried out hourly, can also significantly interrupt production, especially during peak times of order delivery, potentially damaging a company's reputation. Conventional maintenance relies on experienced personnel to monitor devices around the clock. This manual approach was time-consuming and labor-intensive, and it could not be a guarantee of preventing equipment failures. By shifting from reactive maintenance to a proactive approach, the manufacturing industry leverages AI to monitor equipment in real time, predict faults in advance, and maintain devices to reduce downtime and maintenance costs. Furthermore, machine learning algorithms can be used to optimize the production process and adjust product production parameters. AI systems can automatically identify product defects, leading to faster and more accurate product inspection.

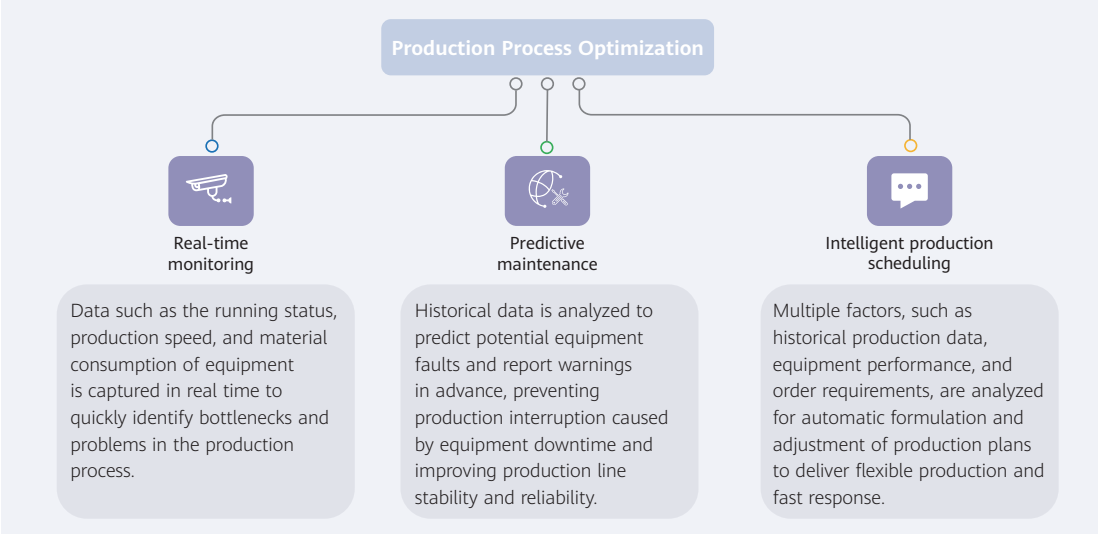


Figure 6: AI-assisted production process optimization

A multinational large-scale digital factory that produces programmable logic controllers (PLCs) overhauled their data infrastructure to integrate digital systems and platforms like the product lifecycle management (PLM), manufacturing execution system (MES), and enterprise resource planning (ERP). The factory used IoT technologies to collect 50 million items of sensor data, equating to around 1 TB that needed to be stored every day. Also, AI technologies enabled different operations like real-time data analytics (analyzing GB-scale machine vision data), production process monitoring, product

quality inspection, and proactive equipment maintenance. This in turn improved transparency and traceability throughout the production process. By utilizing AI across its production line, the factory's time to market was nearly 20% faster, with a 13% boost in efficiency and a significant improvement in product quality.

4 AI-assisted decision-making in operations management

A first-mover advantage can give manufacturing companies a significant edge, allowing them to rapidly gain a greater market share and potentially shape industry trends. The question remains, however, how to utilize market analysis and operations management to enable accurate decision-making. In the past, enterprises would rely on the responses from large-scale user interviews, years of market experience, or cross-departmental discussions to shape their business strategies. Although this could sometimes help win a market share, there was a lack of data support or detailed decision-making processes, and could not be standardized. Nowadays, AI analytics technologies are used to enhance ERP systems, correlating data generated from different phases (product design, production, testing, procurement, warehousing, supply, and sales) and sources (production line, human resources, market trends, and consumption) to support analysis for operational decisions. Such comprehensive end-to-end analysis, grounded in robust data, provides complete information for decision-making. By rapidly analyzing changes in each process and at different times, decision-making can be streamlined and standardized.

5 AI-assisted 24/7 after-sales services

Intelligent chatbots have become ubiquitous in various industries. Providing round-the-clock customer support, chatbots have been widely used in manufacturing, and given the rise of AI, these chatbots are more accurate, professional, and responsive. They can quickly address customer inquiries, resolve issues, and improve customer satisfaction.

1.4.2 Finding value in historical, dormant data and boosting E2E production efficiency

AI is more than just a technological advancement in smart manufacturing; it is a catalyst for a comprehensive digital transformation of the industry. Throughout the smart manufacturing process, from decision making to after-sales services, AI relies heavily on mass data. This includes data for operational analysis, design assistance, demand forecasting, production scheduling, equipment maintenance, product inspection, and intelligent chatbot services. The following are the main challenges of using AI for data analytics:

1 Challenges in data collection and analytics

Real-time data collection and analytics during the production and testing phases can ensure the product yield and long-term normal equipment operations. By leveraging sensors and IoT devices, enterprises can collect equipment data such as temperature, speed, and pressure to enable real-time analytics and monitoring of production status, ensuring stable and reliable manufacturing. In addition to the real-time sensor data of the manufacturing equipment, the product quality inspection data, including images and audio recordings, must be collected for a long time and at a high frequency. To enhance AI analytics accuracy, the data collection intervals will be shortened from hours to minutes or seconds. As the category, type, format, frequency of data collection change, data volumes grow exponentially. The daily data volume has soared from megabytes to gigabytes and even terabytes. A global leader in engineering machinery experienced a surge in the daily data volume, from gigabytes to tens of terabytes, which came from over 560,000 IoT devices. The exponential growth in data collection is fueled by the enhanced AI processing capabilities, which have expanded from megabytes to gigabytes. The company also recognized the huge potential brought by AI to the manufacturing industry. The data surge provided the company with more valuable assets for future innovation, offering the path to stronger share and influence in the market. Specifically, data growth has exploded thousandfold, making the need for high-performance, high-capacity data storage essential.

For large enterprises, finding the way to unleash the value of mass historical data dormant in equipment rooms is also a challenge. Reducing costs and improving efficiency remain a major challenge for the pharmaceutical industry. A leading

pharmaceutical company wanted to boost its product yield and profits without expanding its workforce or production lines at large scale. Despite numerous attempts, they found the solution in their vast trove of historical data. By analyzing vast amounts of historical data aggregated from diverse sources, including production processes and equipment performance, the company identified nine key parameters. Through AI-powered simulations, they optimized these parameters, leading to a 50% increase in medicine productivity and a 3% boost in overall yield. This translated into an annual revenue increase of US$5 million to US$10 million per medicine type. There are many such cases of companies using AI to transform their business. With deep learning algorithms, AI can identify patterns and anomalies from mass volumes of historical data and provide a science-back basis for production, testing, simulation, and decision making. The historical, dormant data is activated and put into use again. Manufacturing enterprises need to aggregate historical data from multiple sources in a fast, simple, and efficient way, without affecting the ongoing production operations. Furthermore, the aggregated data should be efficiently used by the AI system.

2 Challenges in data classification and sorting

(a) Data cleansing: To unlock the full potential of collected data for AI applications, thorough data cleansing is essential. This involves handling missing values, standardizing data formats, and correcting errors to avoid physical and logical errors in collected datasets. An electronic manufacturing group used AI to perform intelligent production scheduling. It found that inference tasks based on raw, unprocessed data consistently yielded inaccurate results, and sometimes led to incomplete or missing output. By leveraging a separate AI industrial data space to interconnect with multiple industrial software systems, the company implemented data aggregation, processing, and cross-verification, ensuring trustworthy, verifiable data and operations. Furthermore, logical data errors were corrected and correct data formats were delivered. This processed data was then used for AI analytics and applications, improving the production scheduling efficiency. To simplify data cleansing and avoid unnecessary work, data must be secure and reliable during the process from collection to storage to avoid non-human-caused data loss and logical errors. Ensuring data accuracy and automatically preventing logical inconsistencies become a problem that manufacturing enterprises must consider. In addition to the challenges in data collection, data cleansing can cause damage or pollution, which requires proactive planning.

(b) Data labeling: Only clean and labeled data can provide relevant context and help the AI model to make accurate predictions during training. Data can have various labels across its lifecycle. For example, production, equipment, operation, and maintenance data all have different uses in the decision making, design, production, scheduling, and after-sales processes. Manual data labeling of different categories, types, and capacities is costly, time-consuming, and human error–prone. Accurate, efficient, and cost-effective data labeling has become a critical challenge for manufacturing companies.

1.5 Electric Power

Electric power is a key infrastructure for national economies, people's livelihood, and enterprise business. The electric power industry continues to face challenges such as power grid scale expansion and load growth. AI-assisted electric power generation management, load prediction of transmission and distribution networks, equipment inspection, and risk identification can effectively improve power supply resilience.

1.5.1 From forecasting to collaboration: accurate electricity supply and demand forecasting facilitates efficient collaboration among power generation, transmission, transformation, and distribution

In the construction of new power systems, the accuracy of electricity supply and demand prediction and the efficient collaboration of power generation, transmission, transformation, and distribution are critical. By leveraging AI technologies, electric power enterprises can accurately predict changes in load and electricity price to better meet supply and demand. This collaboration enhances the overall efficiency of the power system, paving the way for a clean, low-carbon, sustainable, cost-effective, and reliable future of electricity.

1 Electricity generation: AI modeling for better management, less downtime, and better fault identification

Over 90% of the world's top 500 electric power companies rely on intelligent power analysis systems for real-time equipment monitoring, such as thermal or wind turbines and solar panels, helping realize predictive maintenance to reduce unplanned downtime. Enerjisa, a leading Turkish electric power company, leveraged AI analytics to implement real-time monitoring of generator sets and transmission and distribution power lines. This enabled them to reduce equipment downtime by 35%–45% and maintain consistent electricity yield.

In addition, electric power companies install loT sensors in generators and use AI to analyze the sensor information to monitor the motor and component status of generators in real time and identify potential problems in advance. For example, trained on historical wind speed and energy yield data, an anomaly detection model powered by unsupervised learning (an AI algorithm), draws a curve that shows normal operating conditions. By monitoring real-time performance against this curve, the company can identify potential equipment faults before they cause issues, realizing timely maintenance.

2 Electricity supply: Accurate prediction of electricity yield and demand based on AI analytics, solving the problem of renewable energy integration and balancing electricity supply and demand

Historically, estimating electricity generation was straightforward when relying on non-renewable resources like coal and natural gas, but nowadays, predicting the output of renewable energy sources such as solar and wind power presents challenges due to the complex interplay of numerous variables. To complicate it even further, electricity demand cannot be accurately predicted based on historical consumption data, since climate irregularities and lifestyle changes play a significant role. For example, the Australian energy company Red Energy once faced issues with insufficient operating reserve due to low accuracy in their electricity demand forecasting model. This forced them to temporarily purchase electricity at high prices from other power companies, increasing operational costs. After improving their forecasting model with AI, Red Energy achieved a 98% prediction accuracy rate and, through better pre-planning, was able to purchase electricity at lower prices, saving over US$1 million in electricity costs.

3 Electricity consumption: AI-based user analysis identifying abnormal data caused by behaviors such as electricity theft and meter tampering to reduce losses and ensure grid stability

Before the application of AI, power companies were only able to detect electricity thefts during expert inspection or meter replacement, or by conducting random spot checks. These reactive manual approaches led to costly and inefficient detection. Power companies nowadays can use AI to conduct user analysis. Based on existing service rules and any meter tampering history, AI can accurately assess the theft risk of each meter with models analyzing information such as electricity theft behavior patterns

and correlation between electricity consumption and its purposes. The results are then handed over to relevant personnel for further investigation, improving detection accuracy and saving costs. For example, Brazil's second-largest power company not only used AI to identify the risk of organized electricity theft, but also to avoid a monthly loss of several hundred thousand US dollars from theft.

1.5.2 Promoting multi-dimensional and high-frequency data collection and secure data retention for more precise electricity supply and demand forecasting

The power system can monitor and analyze the running status of each process in real time through multi-dimensional and high-frequency data collection. The collected data includes not only the traditional power load and voltage data, but also information regarding the weather, market demand, and device health status. By leveraging this rich data and using AI technologies, power companies can achieve precise control and optimized scheduling across power generation, transmission, transformation, and distribution stages.

1 Using AI to predict power consumption: Increasing the amount and frequency of data collection for more precise prediction results

In the power industry, AI is used to analyze remotely collected electricity data of users to forecast future consumption and prepare for supply and demand. Insufficient data volumes and low collection frequency can lead to deviations in predictions, driving the power industry to continuously increase the data collection frequency of monitors on the user side to meet the requirements of AI analytics models. For example, in IoT metering scenarios, the initial design could predict next month's usage based on weekly or monthly readings for billing. However, while training AI prediction models, it was found that shorter reading intervals could improve prediction accuracy. As a result, data collection frequency was increased to once every few minutes, enabling more efficient predictions and better supply–demand balance. Higher data volumes and collection frequency have increased the demand for greater capacity and performance in data storage devices.

2 Using AI to analyze power equipment: Increasing data collection dimensions and parameters to schedule maintenance in advance and lower downtime

In power generation management, AI is used to analyze electrical component information collected by IoT sensors within generators to identify potential issues in advance and schedule maintenance promptly. In the initial design, parameters such as the aging degree of generator components and the number and types of faulty parts were collected to prepare a replacement parts library. As the volume of data collected by sensors increased, AI training revealed that even weakly correlated data could enhance the prediction accuracy of AI models. Consequently, AI analytics was enhanced to include dimensions like engine operating conditions, device health status, and energy output, improving the capability of hidden issue detection and reducing downtime losses.

3 Electricity protection: Ransomware attacks are a matter of when and not if

The power industry is an integral resource of a state and its people's lives, one that can face significant business disruptions in the event of ransomware attacks. With the increasing digitization of the power sector, it has become a primary target of hacker attacks in recent years. In August this year, cybersecurity company Bitdefender disclosed major vulnerabilities in photovoltaic (PV) plant management platforms operated by Solarman and Deye, which could affect 20% of global PV power generation, involving more than 2 million PV plants in over 190 countries and regions.

New ransomware attacks not only use AI models to generate batches of new ransomware samples, but also feature a longer incubation period and higher covertness, allowing them to easily bypass the common ransomware detection library. For example, a large African power company was attacked by ransomware in recent years and was demanded to pay a ransom of several hundred thousand US dollars.

In the power industry, AI can gather data on the typical operations of essential service production systems and create AI detection and analysis models using data storage devices. These models can identify abnormal behaviors (such as encryption and deletion) on the storage side and detect abnormal storage capacity changes in a short period, helping identify ransomware attacks in the incubation period and reduce attack risks.

1.6 Education and Research

AI-assisted intelligent transformation has led to profound changes in the education and research industry in terms of teaching, research, and management methods. At the same time, construction of IT systems in this industry faces new opportunities and challenges.

1.6.1 From education to exploration: AI amplifies human capabilities by enhancing personalized teaching and scientific research

Intelligentization has been gradually integrated into various education and research scenarios, with intelligent technologies such as large AI models enhancing teaching and research efficiency and quality.

1 AI-assisted personalized/intelligent teaching

Based on students' habits, capabilities, and interests, AI can provide personalized learning plans and resources, increase accuracy of learning evaluation, and assist in intelligent teaching. Typical applications include personalized teaching solutions, intelligent education assistance, integration of diverse teaching resources, virtual classrooms, and real-time learning monitoring, which together provide valuable, real-time feedback for improving teaching quality and efficiency.

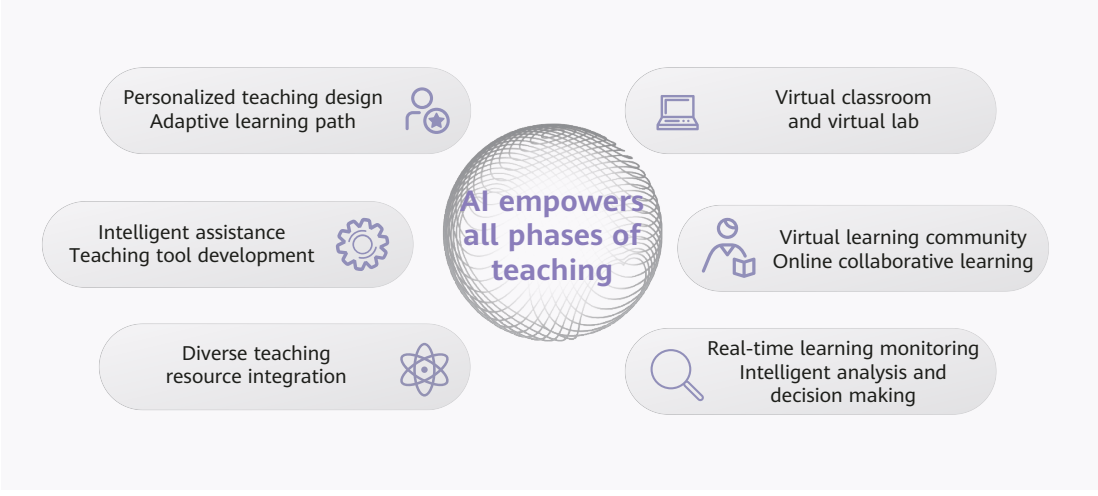


Figure 7: AI empowers all phases of teaching

2 AI-assisted scientific research

Large AI models help researchers quickly filter and analyze massive documents, as well as finding the latest research trends and key concepts through semantic analysis. Concurrently, AI for Science, the emerging scientific research paradigm, is advancing rapidly. It leverages established scientific principles for modeling and explores the patterns within mass data. Harnessing powerful computers, AI for Science is employed to investigate complex scientific problems. For example, in the healthcare industry, AI is used to analyze vast amounts of biomedical data to explore new treatment methods and develop new drugs.

Intelligentization within the education and research industry enhances the efficiency, while also promoting knowledge inheritance and sharing through data aggregation, analysis, and extraction. For example, Shanghai Jiao Tong University has established two scientific research platforms to support both scientific research and teaching services. These platforms interconnect with various AI and HPC platforms and face challenges such as various data access protocols and low data access efficiency. Therefore, a unified storage foundation that supports multi-protocol interworking for various applications is essential for ensuring the efficiency and intelligence of these scientific research platforms.

1.6.2 High-performance, reliable, and resilient data supply underpins AI-driven intelligentization

The ongoing advancement and application of AI in the education and research industry have caused new requirements and challenges in data processing:

1 Real-time analysis of ultra-large and complex datasets

Intelligentization in the education and research industry is characterized by large data volumes, diverse data types, and the promptness of data updates and analysis. For example, personalized teaching leverages intelligent technologies to capture and collect students' facial expressions, actions, and behaviors during class. This data is stored as videos, images, and files, which are then analyzed comprehensively. This has created challenges in data storage capacity and data utilization. The data to be stored is growing exponentially in terms of both volume and complexity, creating pain points

in storage capacity in terms of capacity expansion, equipment room space, and power consumption. The analysis results are used for high-precision predictions of student performance, which in turn inform the intelligent creation of personalized learning plans and teaching suggestions. Promptly processing massive amounts of diverse data under hybrid workloads has become necessary, which requires both high bandwidth for large files, such as videos, and high IOPS for small files, such as AI training data and texts. The difficulty in satisfying both requirements at the same time has led to insufficient data utilization.

2 High requirements on data resilience

Data resilience and privacy protection have become important issues in the AI era, especially when sensitive education information is involved. Scientific research institutions are prone to hacker attacks and ransomware because they have rich financial resources and their scientific research projects usually involve valuable data, some of which even relates to cutting-edge intellectual property rights. At the same time, the network design of education and research institutions is oriented towards data sharing and public access. The multi-device and open access of data across laboratories, offices, and mobile phones has brought huge challenges to data resilience.

3 Efficient data aggregation and mobility

In the education and research industry, data resources from diverse and scattered sources are aggregated and interconnected. Therefore, a more comprehensive data collection and management system is required to implement global data manageability and efficient mobility, ensuring that data from different sources and types is utilized efficiently. Increasing intelligent applications will lead to significantly more requirements for data sharing and collaboration across entities, regions, time points, and domains. However, the siloed construction of existing IT systems can greatly restrict the exploration of data value.

To address these challenges, the education and research industry must establish more efficient, stable, and scalable data infrastructure, which includes efficient data storage solutions, advanced data analysis tools, and strict data management policies. One such example of this is a research team from the Institute of Advanced Agricultural Sciences, Peking University, which is conducting genetics and breeding research for wheat disease

resistance. They do so by utilizing big data and AI technologies to carry out ongoing research on plant genomes, with the goal of significantly enhancing the resistance of major wheat varieties. Crop genome research and analysis involve complex processes and create significant challenges in processing and reading/writing massive volumes of data. First, crop genome research involves mass data generated by genome sequencing, gene expression profiling, and SNP analysis, necessitating a data foundation with superb storage capacity and huge throughput. Second, the entire process of genome sequencing involves continuous reads and writes of fragmented files, which must not be interrupted. Therefore, the storage system that supports sequencing applications must be ultimately stable and reliable to ensure that data is not lost or damaged. Third, electron cryomicroscopy and genome data analysis pose higher requirements on the overall performance and small file processing capability of storage systems. For the research team, one of the most critical issues is storing massive crop genome data while ensuring non-disruptive and high-performance data access.

1.7 Healthcare

Healthcare is a knowledge-intensive industry that stands to benefit most from generative AI. AI is revolutionizing the healthcare industry, including areas such as AI-assisted diagnosis and treatment, drug R&D, and disease prediction. At the same time, data sharing and aggregation, patient privacy protection, and medical data resilience have become challenges for the healthcare industry in the age of AI.

1.7.1 From conventional to AI-assisted healthcare: Diverse AI applications enhance diagnostic efficiency, prevent diseases, and accelerate recovery

The advancement of AI fosters extensive applications in healthcare, leading to significant changes in scenarios like AI-assisted diagnosis and treatment, drug R&D, and disease prediction. The development of AI in the healthcare industry will deeply impact the industry landscape as well as patients' experience.

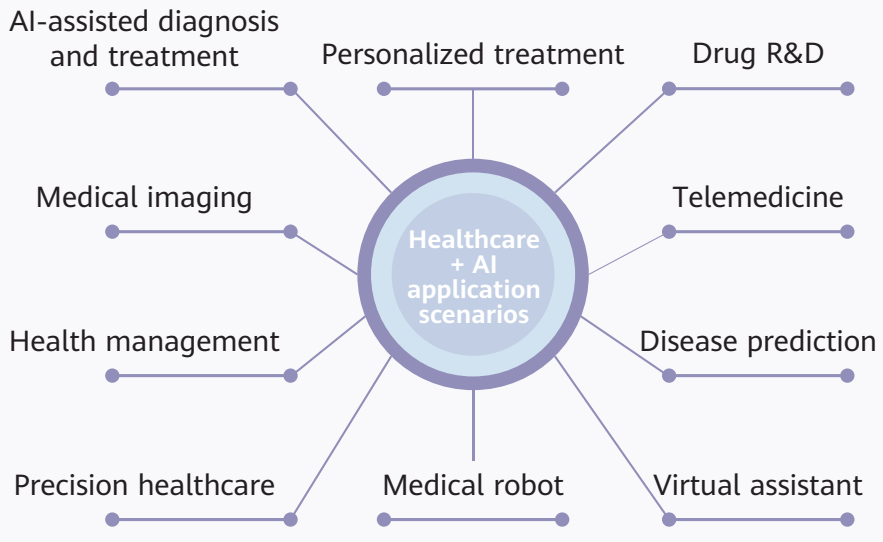


Figure 8: Healthcare + AI application scenarios

1 AI-assisted diagnosis and treatment

The pilot use of AI technology in grassroots healthcare has resulted in the creation of reusable, AI-assisted diagnosis and treatment systems. These application systems offer intelligent triage and other means of AI-assisted diagnosis and treatment to help doctors and improve the intelligence of grassroots healthcare. For example, a device equipped with AI-based segmentation and planning algorithms is applicable to clinical scenarios such as intracranial hematoma aspiration and drainage, as well as intracranial biopsy. In these scenarios, AI is used to find plaques for precise location of bleeding spots in the brain, helping doctors perform surgeries with greater accuracy and safety.

2 Drug R&D

While traditional drug discovery and development are still limited by Eroom's Law, AI technologies are revolutionizing drug R&D through data and algorithm models. Deep learning models enable faster analysis of molecular structures, accelerating the discovery of new drugs and reducing the need for expensive experiments. For example, some research has used AI to successfully identify a drug that works against antibiotic-resistant bacteria. With the help of AI, this drug was discovered within 21 days and verified within 46 days—several years faster than the traditional drug R&D process, at a significantly cheaper cost.

3 Disease prediction

AI and big data model applications provide tools for disease prediction. Disease control departments use these tools to analyze disease-related news and information released by international healthcare organizations, allowing them to more accurately predict trends in disease development and peak periods, and to take appropriate prevention and control measures. Ophthalmologists, for instance, use AI to identify and collect subtle details that are invisible to the naked eye, analyze the retinal changes of patients with specific diseases using big data models, and complete disease detection tasks with clear labels.

1.7.2 Efficient and resilient data sharing protects patient privacy

The application of AI technologies in the healthcare industry has created challenges in data collection, privacy protection, and ransomware protection for hospitals.

1 Difficulties in data collection

Data fuels AI and its quantity and quality determine AI performance. Data sources must be reliable to avoid unreliable data negatively impacting AI training and inference results. To ensure the accuracy of inference results, hospitals must collect training data from reliable sources, such as patients' historical and current medical records.

2 Data privacy risks

The medical field involves a large amount of sensitive data that is associated with patient privacy and value, such as patients' personal identity, health status, diagnosis and treatment records, and genetic information. Data leakage may harm patients, and even threaten social stability, which highlights the importance of data resilience.

At the same time, medical AI application and R&D rely on massive medical data to train algorithms. Greater data volume and diversity can result in more accurate AI analyses and predictions. For that purpose, big data collection, analytics, processing, and cloud storage and sharing are involved, which in turn increase the risk of data leakage.

3 Ransomware attacks

The development of AI technologies enables ransomware attacks to be more targeted, customized, and deceptive. Attackers perform data and behavior mode analyses to effectively select targets and create specific strategies. Additionally, they use AI to adjust their attacks based on victims' responses, adapting them for greater damage. According to *2023 Analysis Report on Ransomware Attacks Targeting Chinese Enterprises*, healthcare has become the most targeted sector. Since 2018, there have been 500 publicly confirmed ransomware attacks against healthcare institutions worldwide, paralyzing nearly 13,000 independent facilities and affecting nearly 49 million patient records. The economic losses caused by these attacks, due to downtime alone, have exceeded US$92 billion. According to third-party statistics, the healthcare industry has had the highest data breach costs for 12 consecutive years, and in 2022, the average data breach cost for healthcare organizations reached US$10.10 million, a sharp increase of 42% compared to 2020.

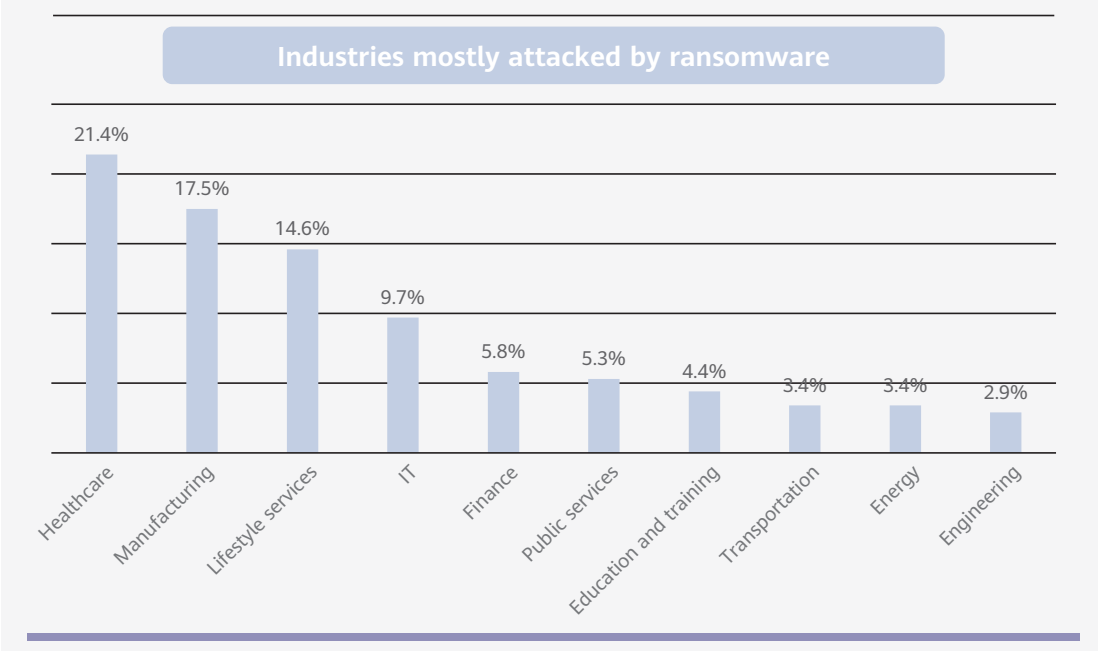


Figure 9: 2023 Analysis Report on Ransomware Attacks Targeting Chinese Enterprises

To effectively address data challenges, the healthcare industry needs professional data storage products that offer advanced technologies, including intrinsic storage resilience, DR and backup, secure and trusted data mobility, and ransomware protection. These technologies are essential for ensuring efficient data storage, robust data resilience, and high data utilization, all of which will help accelerate the intelligentization of the healthcare industry.

1.8 Digital and Intelligent Transformation: Data Is the Key to Success

Digitalization and intelligentization are rapidly transforming industries such as finance, carrier, government, manufacturing, and electric power.

Digitalization converts information from daily life into digital data, enhancing recording, processing, and transmission efficiency. In turn, intelligentization leverages AI computing power for data training and inference, maximizing its value.

Digitalization and intelligentization enhance each other. As they become intertwined, digitalization and intelligentization enter a symbiotic relationship, eventually evolving into digital and intelligent transformation. This process generates insights from data and applies them in order to enhance digitalization across various industries, thus making them more efficient and intelligent. As technology advances, digital and intelligent transformation will have an even greater impact on society, creating an increasingly intelligent world.

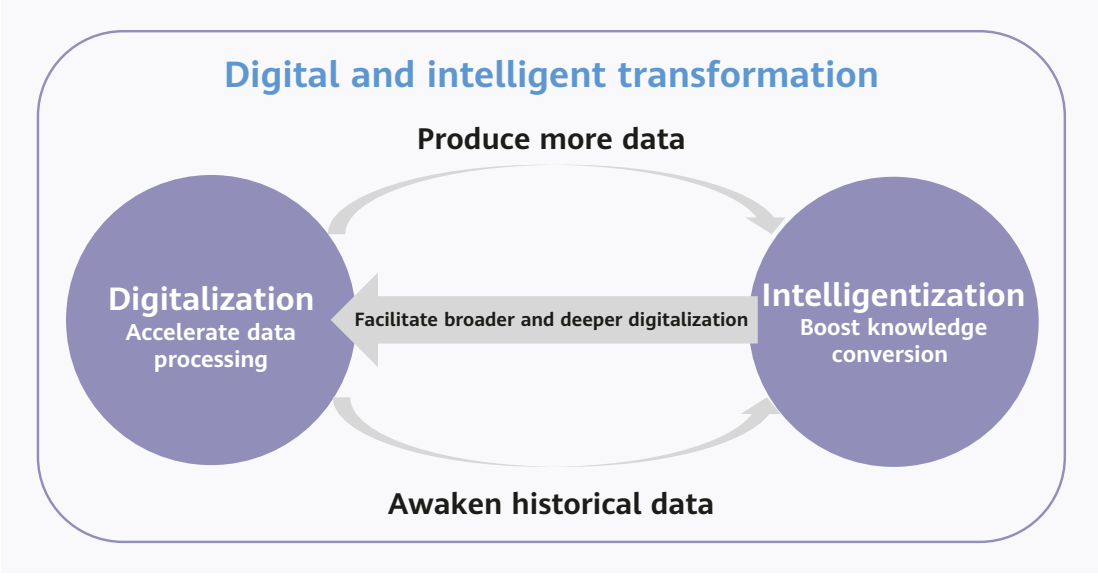


Figure 10: Data powers the shift from digitalization to intelligentization

Digitalization leverages general-purpose computing power, while intelligentization employs AI computing power to process and analyze data, creating value. Ultimately, data serves as the vital link between digitalization and intelligentization, laying the foundation for their convergence and a successful digital and intelligent transformation.

02

Digital and Intelligent Transformation Across Industries Requires High-Quality Data and Efficient Data Processing

Data quantity and quality determine how far AI can go. For example, GPT-1 was trained on 4.8 GB of raw data, while GPT-2 used 40 GB of manually filtered data. Then, GPT-3 took a giant leap to 570 GB of data, filtered from 45 TB of raw data. ChatGPT and GPT-4 were then built on the training data for GPT-3, but with high-quality labels added. GPT-1 to GPT-4 share a similar model architecture, but they differ greatly in model parameter scale, dataset scale, and dataset quality. This explains why their training outcomes are so different. The evolution of GPT has proven many AI scholars' view: AI is data-centric.

A wide range of industries going digital and intelligent currently depend on vast amounts of high-quality data for secondary training and supervised fine-tuning of foundation models, as well as for application inference. In practice, most enterprises provide data for AI computing by waking up historical data, collecting and saving more production data, and manually synthesizing data.

In addition to data quantity and quality, data efficiency is equally important for organizations going digital and intelligent. It helps them review data from multiple dimensions, including data storage, access, energy consumption, and resilience, to ensure that data provisioning is available, fast, and stable.

2.1 Data Awakening: Maximizing the Value of Historical Data

There will be no AI without sufficient data. Data scarcity has become a major obstacle in developing large AI models.

Currently, large AI models are empowering an array of industries while also struggling with a significant challenge due to the lack of massive, high-quality industry datasets. Industry data is rich in domain-specific knowledge, terminology, rules, processes, and logic. This specificity makes it difficult for general-purpose datasets to fully capture. Additionally, industry data is scarce. According to the Beijing Academy of Artificial Intelligence, all known open-source industry text datasets total only about 1.2 TB, falling far short of what most industry models need.

Data plays a pivotal role in the AI field. AI models need a large amount of data for training. The data is used to learn patterns, predict results, and optimize performance. Without sufficient data, the accuracy and effectiveness of a model is limited.

1. Data-driven decision-making: AI systems make decisions based on data. From financial prediction to healthcare diagnosis and beyond, data underpins AI systems' intelligent decision-making.

2. Iteration and improvement: Data helps AI systems consistently iterate and improve. By analyzing user feedback, monitoring performance metrics, and updating data, AI can continuously optimize itself.

3. Personalized experience: Data enables AI to provide a personalized experience for each user. For example, recommendation algorithms can push content based on users' historical behavior and preferences.

Data awakening is a must for transitioning from the digital era to the intelligent era

1 Activating idle service data

A huge amount of data is generated while running services. Some is hot data, which is frequently accessed and may be modified at any time. Other data is infrequently accessed. Although the latter is still stored in primary storage, it is highly unlikely to

be accessed again. For example, once a patient at a hospital recovers from an illness or injury, the large amount of medical image data generated during their treatment may no longer be accessed. This idle data is normally only transferred to other storage devices when the primary storage capacity is used up.

As large AI models grow in size, their demand for training data skyrockets. Incorporating idle service data into training data resources can effectively facilitate model training.

2 Waking up historical archived data

For example, enterprise archives, historical records, and academic documents are currently being extracted and used. Packed with valuable historical information and mass data, these materials are ideal for model training, trend analysis, and service prediction.

(a) Training data diversification: Historical data includes past events, experience, and knowledge. By activating the data, we can obtain more diverse training samples for training machine learning models. This helps improve the accuracy and generalization capability of a model.

(b) Trend analysis: Historical data can be used to predict trends. By analyzing historical data, we can uncover patterns, periodicity, and trends, thus predicting possible future events. This is critical to business decision and planning.

(c) Anomaly detection and fault prediction: Anomaly and fault information contained in historical data can help us build an anomaly detection model for real-time monitoring and early detection of potential problems to prevent losses.

Collecting and managing high-quality training data is a must for continuous AI evolution

Wikipedia currently contains about 420 million words. According to ARK Invest's Big Ideas 2023 report, model training will require a whopping 162 trillion words by 2030. The increasing size and complexity of AI models will undoubtedly lead to a greater appetite for high-quality training data.

In an age of ever-growing computing scale, data will become a major constraint on AI development. As AI models become more complex, the demand for diverse, accurate, and large datasets will continue to grow. The following factors will all be important considerations when managing and waking up historical data:

1 Data source diversity

Collecting data from a variety of sources helps ensure that AI models are trained on diverse and representative samples, thereby reducing bias and improving overall performance. Data infrastructure must be able to store large-scale datasets from diversified sources; quickly read, write, and retrieve data to facilitate model training; protect data from unauthorized access and damage; and ensure data durability and reliability.

2 Data quality

Training data quality is also a critical factor that influences the accuracy and effectiveness of AI models. Prioritizing data cleansing, annotation, and verification can help achieve the highest dataset quality. Data infrastructure powered by technologies such as data cleansing and labeling can also help unlock the full potential of available training data. Data infrastructure must have a high capacity to store high-quality datasets, and, most importantly, implement near-storage computing to build data cleansing, labeling, and verification capabilities on the storage side.

3 Data privacy

As the demand for training data grows, privacy issues will also have to be addressed, and data collection and processing will have to be subject to ethical guidelines, data protection laws, and related regulations. Technologies like privacy computing can help protect personal privacy while still providing useful data for AI training.

2.2 Data Generation and Synthesis: Crafting Data for the Digital-Intelligent Era

Waking up massive amounts of historical data and using the data for AI model training and inference have played a large role in supporting the rapid development of large AI models. However, it has become increasingly clear that, although mass historical data has played an irreplaceable role in the advancement of AI, the data was not created for AI. This means there is still room for improvement in terms of data collection frequency, data format, data diversity, and data retention.

A factory that previously relied on cameras and manual inspections to monitor gasoline tanks can be used as an example. With advances in machine vision models, the factory can now use AI to analyze oil spots on tanks in real time to detect potential leaks before they occur. However, legacy monitoring systems often only retain data for the latest 30 to 90 days, meaning there is very little historical footage of what high-risk oil spots look like, which hampers AI training. Moreover, older cameras often provide high enough video resolution that humans could identify potential issues. By contrast, AI typically needs higher-definition video to predict potential risks more accurately.

Therefore, in this AI-driven age, organizations need to consider how they can fully use existing historical data and improve data generation in existing digital services, thus accelerating their digital and intelligent transformation with improved data quantity and quality.

In addition to generating more high-quality data in existing services, organizations will need to explore data synthesis as an approach for creating data that is difficult to obtain in practice.

2.2.1 Data generation

Typically, the 5F framework can be referenced for the generation and retention of high-quality data for AI applications.

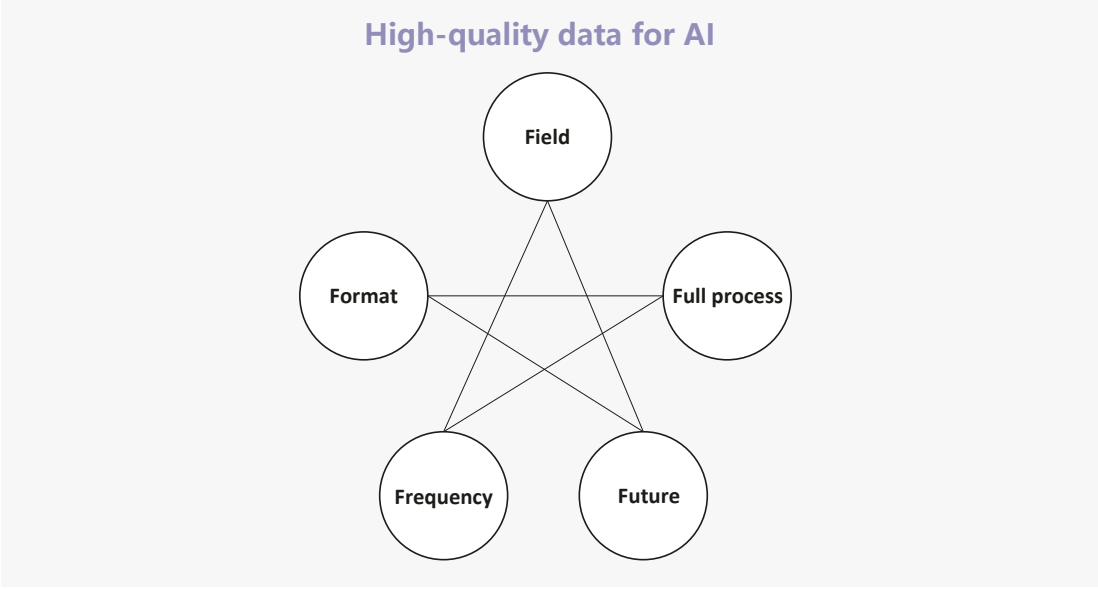


Figure 11: 5F framework for generating and collecting high-quality data

The 5F framework helps organizations generate, collect, and retain more high-quality data for AI by focusing on five dimensions: Field (data generation/collection site), Format (data generation/collection format), Full process (full-process service data), Frequency (data generation/collection frequency), and Future (future-proof data retention).

1 Field: data generation/collection site

Data is generated across a variety of regions and locations, reflecting the diverse scenarios in which it is collected. For example, data can originate from remote outdoor activities (such as oil and gas exploration and oceanographic expeditions), from indoor devices (such as smart meters and home appliances), and from mobile terminals (like smartphones and laptops).

Before the rise of large AI models, data collection and recording were often limited to information that could be immediately processed. Let's take smart meters as an example. Initially, they were designed to facilitate automatic meter reading (AMR), which was intended to replace the traditional manual meter reading process. Later, the implementation of advanced metering infrastructure (AMI) allowed for real-time analysis of power consumption and enhanced the efficiency of power transmission and distribution. Today, many electric power companies are exploring the use of smart meters to gather additional onsite environmental data, such as temperature, humidity, atmospheric pressure, and noise levels. This data enables them to leverage AI for more accurate analysis and predictions, ultimately leading to improved energy efficiency. For

example, in Area A, the average temperature may be about 30°C with a humidity of up to 90%, while in Area B, the average temperature could be about 35°C with a humidity of lower than 5%. Although the data does not directly relate to power supply, the electric power company can use this data to forecast the probability of air conditioner use in different areas. Then, they could anticipate higher usage in Area A than in Area B, allowing them to proactively allocate power resources to the different areas.

2 Format: data generation/collection format

During digitalization, organizations often choose solutions that best suit their current services due to different requirements and budgets. There is also a high probability that AI was not considered in the past. However, as AI is gradually applied to an increasing number of industries, it is emerging as a key consideration that cannot be overlooked during digitalization. In terms of AI, data format is a major dimension.

Data format refers to the way in which information is digitized. For example, an audio clip can be WAV, FLAC, or MP3 format, while an image can be JPG, GIF, or PNG format. In addition to the codec formats mentioned above, the concept of data format also covers definition and resolution.

Organizations need to consider whether their data formats meet current and future AI needs.

3 Full process: full-process service data

Today, AI training is mainly about learning from results, especially correct results. But in the real world, human beings acquire knowledge by learning from both correct and incorrect results, as well as through calculation and deduction processes, such as code programming and drawing design.

The capabilities of AI are constantly evolving as it learns from the data generated by incorrect results and calculation and deduction processes. Yet, such data has neither been saved in a complete way, nor incorporated into our digital process. With the context window being expanded in the AI era, it is believed that learning from such data will be key in order to further develop AI.

4 Frequency: data generation/collection frequency

In the digital era, there are two ways to collect and retain data during production:

(a) All generated data is recorded, and is matched with the compute and network resources needed to process the data. A typical scenario is the financial industry, such as online transaction.

(b) Generated data is periodically sampled and the sampled data is saved. The sampling period depends on the current processing capability and accuracy of the service system. A typical scenario is scientific research, such as meteorological monitoring stations and scientific breeding.

AI is now bringing about ultra-large computing power, and the problem of data hunger is gradually emerging in the second scenario. That is, AI computing power waits for more high-quality data input. Organizations should consider how to moderately improve data collection frequency and effectively save the high-value data generated in the real world in order to provide more data fuel for AI.

5 Future: future-proof data retention

Before the large AI model era, data was retained for long periods for the purpose of archiving and future reference. In today's environment, it is essential to verify the data retention period in advance, not only to comply with minimum regulatory requirements but also to support the development of AI. Although not all retained data may be utilized immediately, it is prudent to prepare for the future by making moderate advancements.

For example, the customs of one country requires that traveler entry and exit records be kept for five years for querying as needed. However, as large AI models grow mature, five-year-old data may gradually fail to support model training. The customs of the country is therefore now considering extending the data retention period to 10 or even 20 years.

2.2.2 Data synthesis

Data synthesis is a method of generating manual data through computer algorithms or simulations. It simulates statistical features of real-world data, but does not contain or only partly contains real-world data. Data obtained through data synthesis is known as synthetic data, and can be used for various purposes, including data augmentation, data

privacy protection, and model training and testing in the context of data scarcity.

Synthetic data can be generated in the following ways:

1. Statistical distribution–based data synthesis

This data synthesis method first analyzes the statistical distribution of the real datasets, and then determines the statistical distribution rule, such as normal distribution. On this basis, data experts synthesize data from scratch or based on specific initial datasets, and create datasets that are statistically similar to the raw datasets.

2. Machine learning–based data synthesis

This method is essentially the same as the first, with the only difference being that a machine learning model is trained to understand the statistical distribution of the real datasets. In this way, the model is used to generate synthetic data that has the same statistical distribution as the real data.

3. Generative AI–based data synthesis

Generative AI is asked questions or prompts to generate image, text, audio, video, and other types of data. Data synthesis based on generative AI has some similarities to that on machine learning. Generally, the former is used to generate common data such as that related to text and image, and the latter is used to generate statistics required in scientific research scenarios.

4. Randomized algorithm–based data synthesis

Randomized algorithms can generate information like names and home addresses to anonymize the sensitive data found in the raw datasets and protect personal privacy.

Synthetic data has multiple benefits, including the ability to generate endless amounts of data, privacy protection, bias reduction, and data quality improvement. It also allows organizations to use data without violating privacy regulations, while providing a cost-effective way to obtain more data.

However, synthetic data has its limitations. For instance, synthetic data may not fully simulate real data due to its complexity and diversity, and its generation process may require a high level of expertise and skills.

With correct understanding and proper usage, synthetic data is a beneficial supplement to the raw data obtained in the real world. It can address key challenges like data scarcity and privacy protection, allowing it to play a key role in AI research and application development.

2.3 Data Efficiency: Efficient Data Access Enables Efficient Data Processing and Accelerates the Digital and Intelligent Transformation of Industries

In the digital-intelligent era, AI has become the core driving force of social development, and data has become the most important production means. Enterprises' generated and retained data is set to increase exponentially. At the same time, AI applications need mass data ingestion, data loading of hundreds of billions of files, and checkpoint saving and loading during resumable training. Data efficiency is crucial to fully unleashing AI productivity and maximizing value. In the past, data efficiency was improved by optimizing data storage performance, capacity, and reliability. In addition to these three dimensions, the digital-intelligent era will require future-proofing improvements in terms of data paradigm, sustainability, and data fabric to comprehensively improve data efficiency and unleash data productivity.

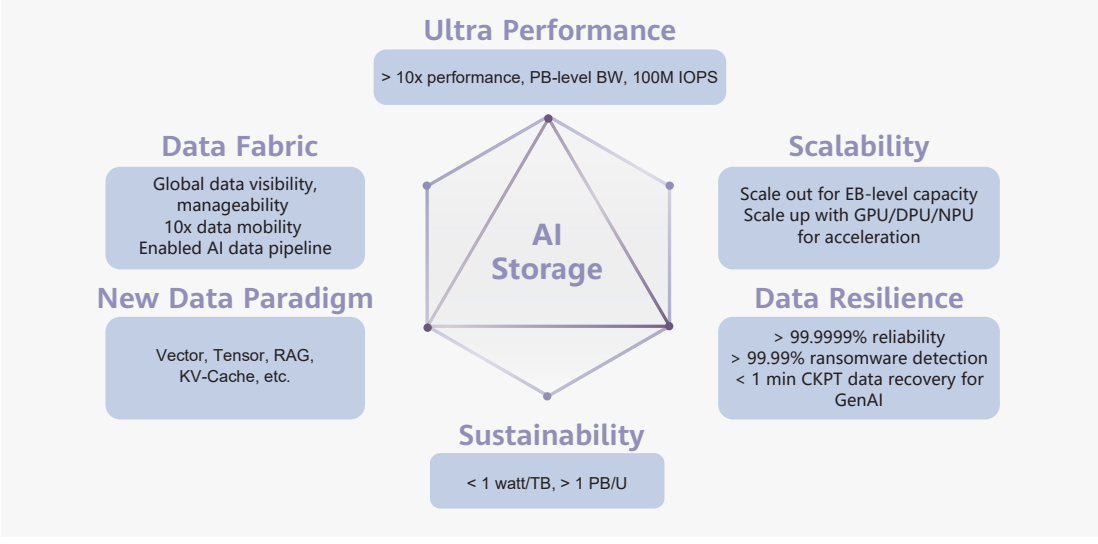


Figure 12: AI storage

1 Ultra performance

Today, large AI model training faces several problems: Data sources are scattered and difficult to ingest; data is frequently migrated; PB-level data preprocessing often takes several days; during training, numerous small files load slowly, resulting in long GPU idle time; and time-consuming recovery from checkpoints causes low GPU utilization. Therefore, higher-performance storage is required to improve the utilization and training

efficiency of the AI training cluster and reduce computing power wait time. Future storage needs to deliver 10-times higher performance than that of conventional storage, PB/s-level bandwidth, and hundreds of millions of IOPS. In this way, mass data loading and checkpoint writing can be completed much more quickly. In addition, storage should support multiple data protocols in order to reduce data replication needs, thus greatly improving the full-process efficiency of generative AI.

2 Scalability

In the AI era, data has become an increasingly critical high-value asset for enterprises, and the data retention rate keeps rising. Enterprises' average data retention period has increased from months to dozens of years, and the data growth rate will increase exponentially. Explosive data growth requires higher storage capacity, and future storage clusters will need to support EB-level capacity scale-out. In addition, each controller enclosure needs to scale up to multiple GPUs, DPUs, or NPUs to support near-storage computing.

3 Data resilience

As data becomes increasingly valuable, higher data resilience becomes indispensable. Comprehensive measures are required to protect data integrity and availability. For one thing, data resilience refers to reliability during production—no data is lost and services are not interrupted. Professional storage devices adopt architecture and technological innovation like node redundancy design to build resilient multi-level reliability mechanisms and achieve 99.9999% availability. This eliminates concerns about the impact of data issues on the running of AI services. For another thing, growing ransomware risks necessitate a comprehensive mechanism, consisting of dynamic detection, proactive defense, and collaborative recovery. In this way, static management will switch to dynamic detection and reactive response to collaborative defense. As a result, the multi-dimensional ransomware protection solution forms a complete defense system that guarantees data resilience.

4 Data fabric

To efficiently use data assets, they must first be visible, manageable, and available. Data fabric indicates that any data can be used anytime and anywhere. Specifically, a unified view of data assets is needed to achieve the global visibility and real-time update of complex data across regions, sites, and vendors. Second, data catalogs become more intelligent. AI and automation technologies are used to automatically label, aggregate, retrieve, and display data. Data is automatically classified and graded by content, compliance, and access frequency. Data fabric also enables automatic data mobility by analyzing hot, warm, and cold data, leading to efficient and economical data storage. In response to the challenge presented by massive amounts of files, data fabric can retrieve specific files from among hundreds of billions within seconds, achieving efficient data retrieval.

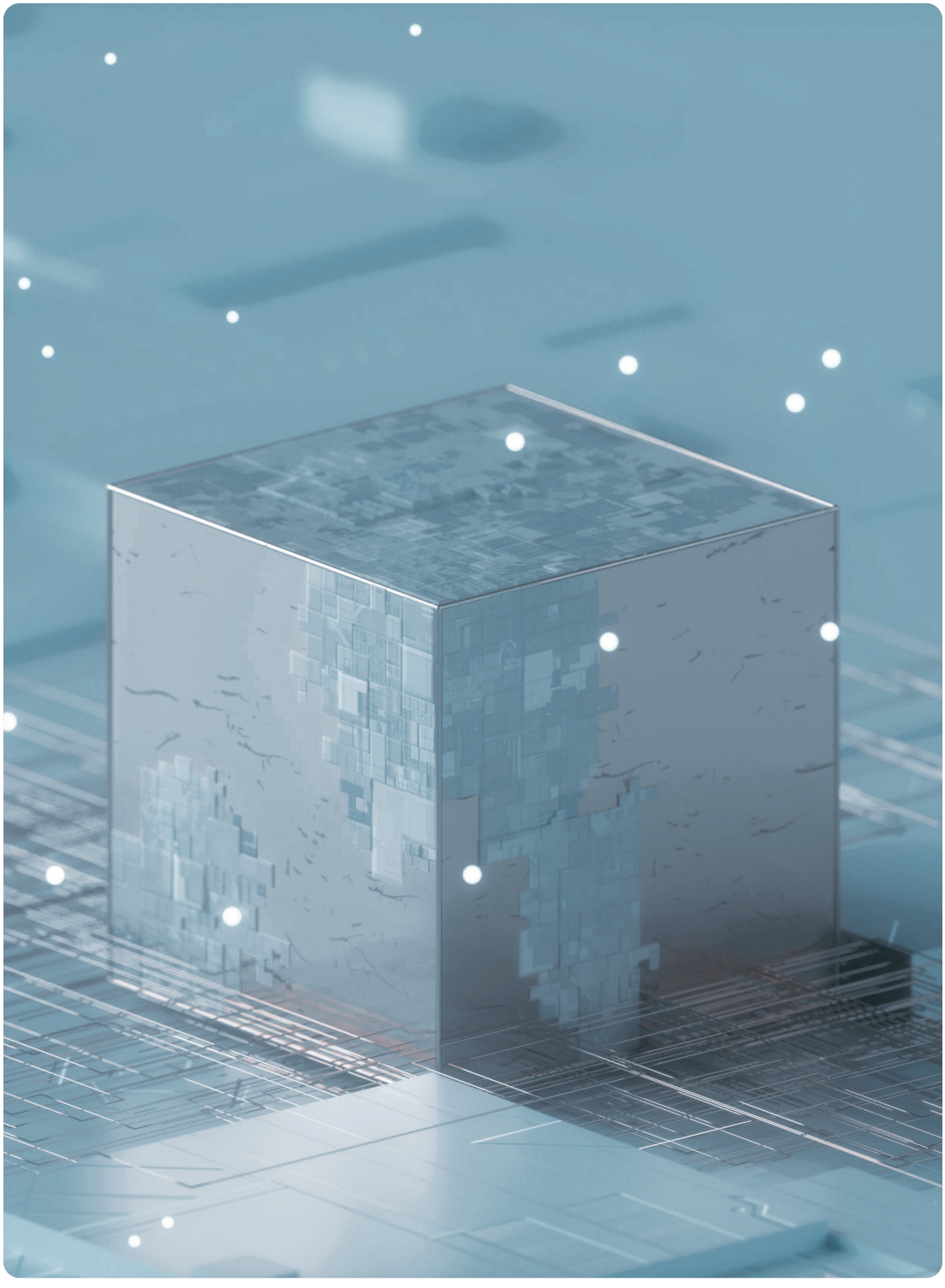
5 New data paradigm

Near-storage computing is used to realize near-data preprocessing, so that storage can take over some data preparation tasks such as data filtering, normalization, transcoding, and augmentation, reducing data migration and improving GPU utilization. At the same time, vectorized storage and retrieval are applied to the latest service data of enterprises, significantly streamlining their access and usage of large AI models. Furthermore, multidimensional tensor data is enabled to support fast data retrieval via an intelligent search engine. Retrieval-augmented generation (RAG) technology works with the embedded knowledge base to eliminate hallucination in large AI models. Multi-layer KV-Cache enables long-term memory storage. Thanks to the enhanced retrieval capability, there are fewer repeated computations, and inference data is shared in the inference cluster. These improvements relieve the computing power load, which means lossless ultra-long sequences with tens of millions of tokens are supported with better inference efficiency and precision.

6 Sustainability

Energy saving and emission reduction are the basis for sustainable social development. By 2026, the power consumption of global data centers is expected to reach 2.3 times that of 2022, which is equivalent to the annual power consumption of Japan. More than half of the power in data centers will be consumed by AI, and storage is the carrier of AI data. Therefore, the increasing data volume will require a more effective energy-saving data storage solution. Optimizing energy consumption per unit of data storage is an inevitable trend in industry development. Innovations in storage media and devices have

brought about outstanding storage energy efficiency (less than 1 watt/TB) and storage density (greater than 1 PB/U). Through media innovation, high-capacity SSDs now provide 10 times more capacity with the same disk size, which can reduce the space and energy consumption of a data center.



03

Data Infrastructure in the Digital-Intelligent Era

3.1 AI-Ready Data Infrastructure Based on the Decoupled Storage-Compute Architecture

AI-ready data infrastructure with a decoupled storage-compute architecture accelerates AI development

As large AI models evolve towards multi-modality, the computing cluster and data scales are continuously expanding, increasing the complexity in system management. Data storage power becomes the key to the continuous and rapid growth of AI.

The decoupled storage-compute architecture effectively simplifies intelligent computing cluster management and facilitates on-demand scaling of compute and storage resources. It offers flexible scaling, linear performance growth, and multi-protocol interworking capabilities to meet the requirements of data infrastructure.

3.1.1 Trends

1 Multi-modal AI training is generating larger volumes of more complex data

The evolution of large AI models from NLP to multi-modality has led to dramatic increases in data volumes and processing complexity. For example, NLP typically involves around 100 billion parameters with simple training data like numbers, texts, images, and audio files. In comparison, multi-modal AI training encompasses trillions or even tens of trillions of parameters, and it added training data like videos and 3D and 4D data files, each up to dozens of gigabytes in size. This fundamentally changes how data is accessed, collected, and organized.

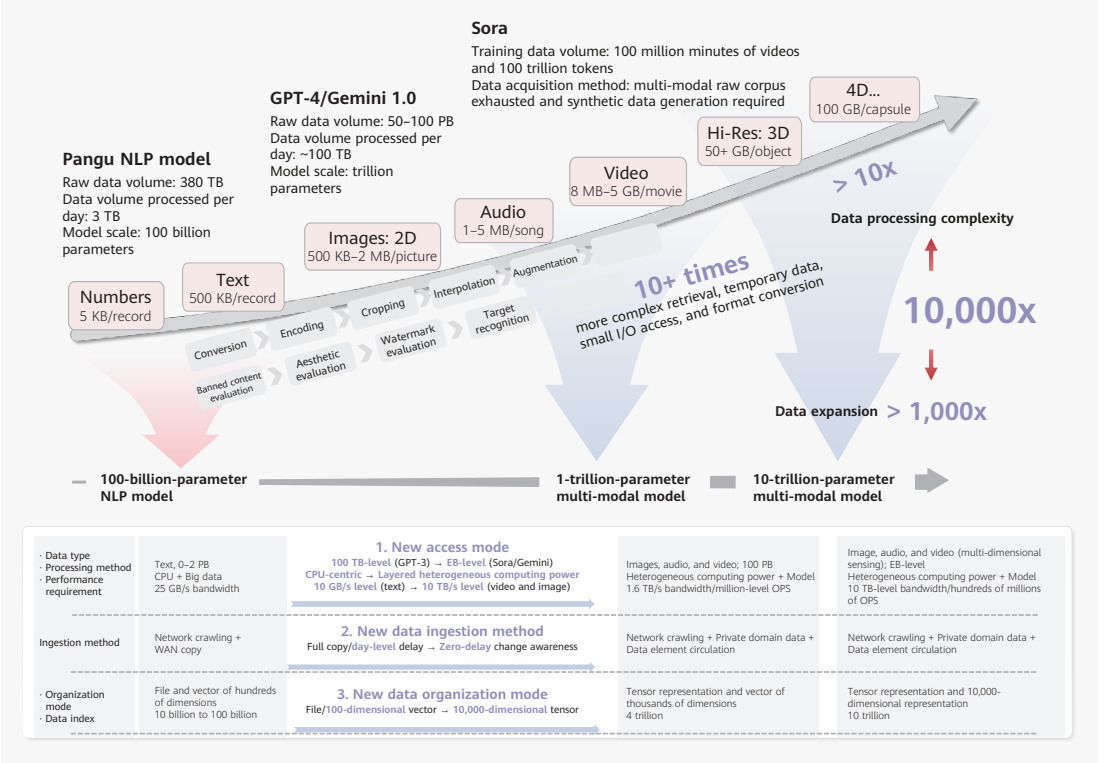


Figure 13: Multi-modal training is generating larger volumes of more complex data

2 AI computing clusters are expanding in scale but declining in computing power utilization

The training and inference process of a large AI model has four phases: data acquisition, data preprocessing, model training, and model inference.

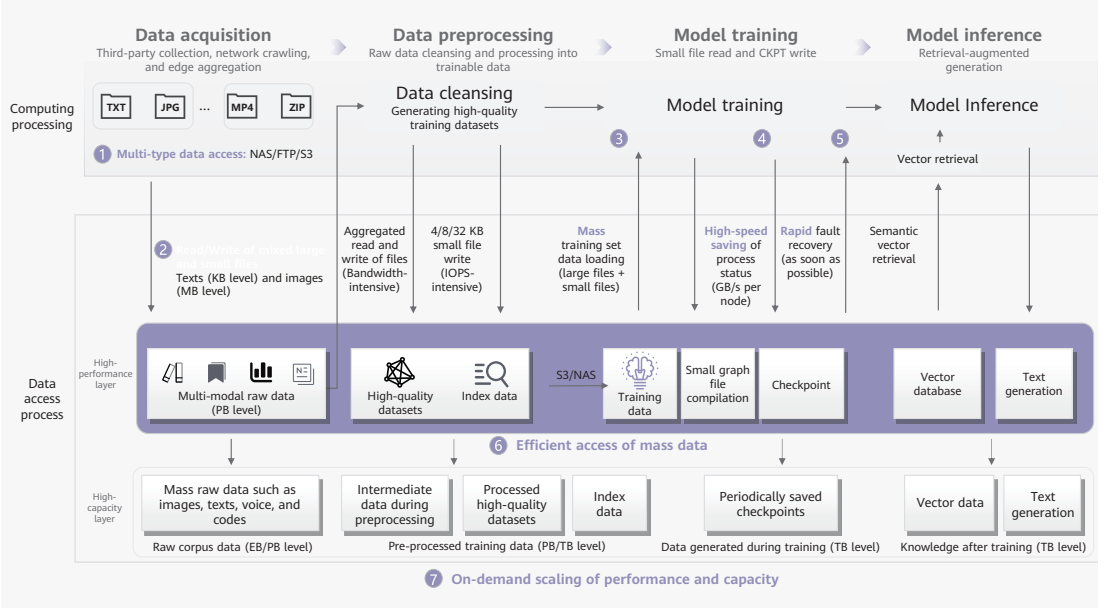


Figure 14: Computing and data access throughout the AI training phases

Phase 1: Data acquisition. Data from various sources is imported to the storage system (usually a data lake). Then analysis software such as Spark is used to collect, filter, cluster, and index the data for subsequent analysis and processing. This phase typically requires exabytes or petabytes of raw corpus data, which is accessed using NAS or S3 protocols. The data access uses a mixed I/O model that involves KB-level files and MB-level images.

Phase 2: Data preprocessing. The data preprocessing software conducts feature extraction, modeling, and vectorization on the cleaned data to form a feature library.

Phase 3: Model training. The AI training cluster executes multiple epochs, adjusts the weights and biases during each epoch to optimize the model quality, and finally creates a model database capable of addressing specific types of problems. In this phase, massive training datasets are loaded to the GPU memory before each training session. During this process, TB-sized checkpoint files are periodically saved to the storage system. In the event of a fault, these checkpoint files must be quickly loaded from the storage system for recovery. Note that this process requires exceptionally high storage performance, as a higher storage performance means better efficiency. When training its Llama 3 large model, Meta used 16,000 GPUs, but the process was hindered by 419 interruptions due to unexpected component faults. An interruption occurred every three hours on average. These interruptions severely impacted the efficiency and stability of the AI training. The service interruption time of a cluster is calculated as follows:

$$=(1-\left(1-\frac{\left(\frac{\text{Backup interval CKPT}}{2}+\text{MTTR}\right)}{\text{MTBF}}\right))*365*24$$

The annual average cluster service interruption time is shown in the chart below:

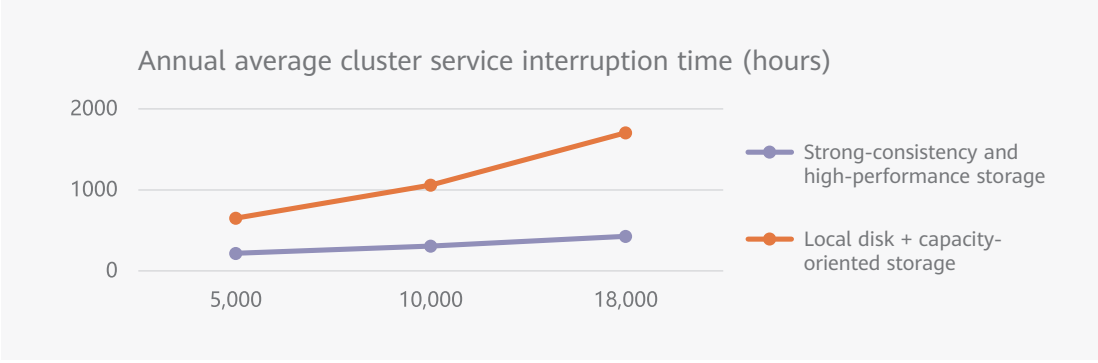


Figure 15: Annual average cluster service interruption time

To minimize the duration of the interruption, it is crucial for the system to promptly read data and resume training after a fault occurs.

Take the reading and writing of a checkpoint as an example. During the training process, each GPU writes a checkpoint shard, and the shards of all GPUs are combined into a complete checkpoint. An error in any shard will render the training epoch invalid.

In the following figure, each training node generates multiple shards at T0 to form a complete checkpoint 0 (CKPT 0). If these shards are stored on the local disks of the servers, the nodes asynchronously replicate the shards to the external storage.

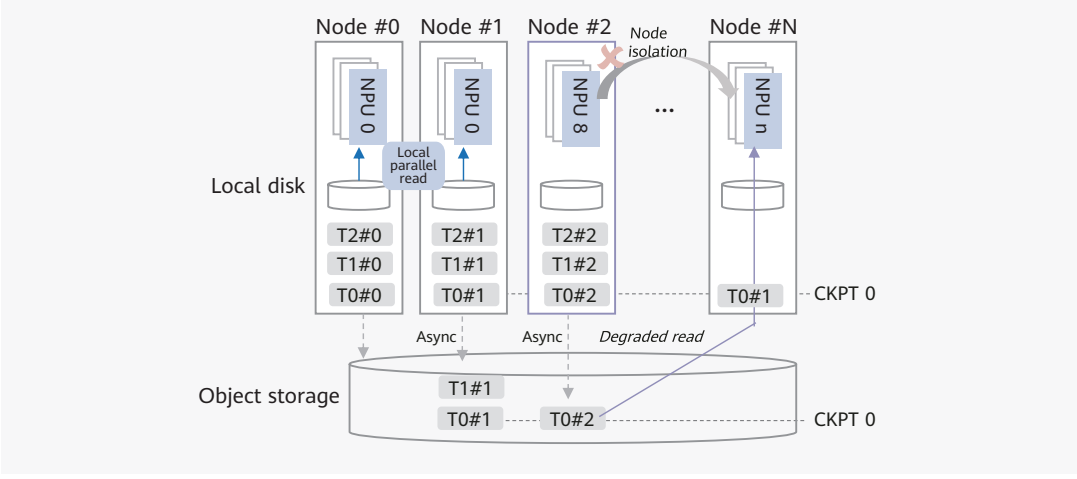


Figure 16: An invalid training epoch due to node faults

If node 2 is faulty, the training job removes the faulty node and switches to a new node N. However, because data is not shared between the local disks of the servers, the training job must load data from the external storage.

Since checkpoint shards are asynchronously replicated to the external storage, the training job can only load the shards that were replicated a few epochs ago. This renders the most recent epochs invalid. In addition, poor performance of the external object storage will prolong the data loading time, and this in turn will delay the training job. That means a fault at one node slows down the recovery efficiency of the entire cluster.

Phase 4: Model inference. To improve the accuracy of large model inference for user queries, enterprises often fine-tune their large models with private domain knowledge and leverage Retrieval-Augmented Generation (RAG) to enhance the readiness of responses.

3 Hallucination is common in AI inference

AI hallucinations may be caused by various factors:

(a) Insufficient data and poor data quality for the training of foundation models. Using inaccurate or incorrect data during the training of foundation models will cause AI hallucinations. The incorrect data may come from errors in data collection and processing or inherent errors within historical data. Inaccurate data directly undermines the judgment and prediction capabilities of the model, resulting in unreliable results. Moreover, the biases of training data in different groups, categories, or scenarios will amplify the issue in the results, which further compromises the fairness and universality of the model. For example, an object recognition model primarily trained on data from light-colored objects may struggle to recognize dark-colored objects. Some data may also become outdated over time. Models trained on outdated data will not be able to adapt to the latest application scenarios and changing requirements. If there is not enough training data, the model will have limited generalization capabilities, making it difficult to adapt to new scenarios and unfamiliar data. This type of model generally performs well on the training set, but underperforms in actual applications or on test sets. Large datasets should be diverse enough to ensure that models have sufficiently extensive knowledge and processing capabilities to be used in a wide range of scenarios. Datasets lacking diversity will restrict the application scenarios and hinder the performance of the models.

(b) Insufficient industry data and poor data quality for the secondary training and fine-tuning of foundation models in specific industries. When the secondary training of a foundation model is conducted on limited industry data, the model is prone to overfitting, resulting in poor performance in adapting to new and unfamiliar data. This is a common issue in machine learning and deep learning, especially in complex model structures. In addition, small-scale datasets may fail to include all of the important industry scenarios, and this will mean that the models are not adequately representative, leading to inaccurate predictions in actual applications. The distribution of data for a specific industry may have an obvious long-tail effect, meaning that the majority of data is concentrated in a few categories, while other categories may lack data. As a result, models may perform well in common certain categories but struggle in others. Poor quality industry data, often due to noise, errors, inconsistent labels, and missing key information, can also impair a model's judgment accuracy and adversely affect its performance in actual applications.

(c) Lack of industry consensus, basic knowledge, and real-time information in inference. A foundation model that lacks an understanding of industry consensus and basic knowledge may fail to address core industry issues during inference, leading to superficial analysis results. This type of model may not be able to provide effective support for decision-making due to the lack of necessary industry knowledge and may instead compromise the accuracy and reliability of decisions. To accurately identify and learn industry-specific patterns and rules, models must be trained on a substantial amount of professional knowledge. Furthermore, models need real-time industry information, such as price fluctuations in financial markets and inventory information in supply chain management, to maintain their timeliness and effectively predict future trends. Without access to real-time information, model outputs may become outdated and fail to keep up with industry changes.

3.1.2 Suggestions

1 Use the decoupled storage-compute architecture to enable independent deployment and on-demand evolution of computing and storage power

A decoupled storage-compute architecture that separates computing power from storage power is particularly important for the deployment of large AI models. This architecture effectively improves resource utilization and provides powerful support for model training and inference. It enables independent and on-demand scaling of compute and storage resources, reducing the required level of investment and preventing resource waste. Furthermore, this architecture changes the current situation of merely accumulating computing power during large AI model training. It allows for the use of high-performance and highly reliable external storage systems to optimize the entire AI training process, thereby reducing interruptions in training jobs and enhancing the availability of computing power. To ensure load balancing within a cluster, users can increase compute resources to process larger data volumes during peak hours without worrying about storage bottlenecks. For data-intensive tasks, users can separately enhance storage performance to increase the overall processing speed. Users can select the most cost-effective combination of resources based on trends in resource prices and service requirements to effectively control costs.

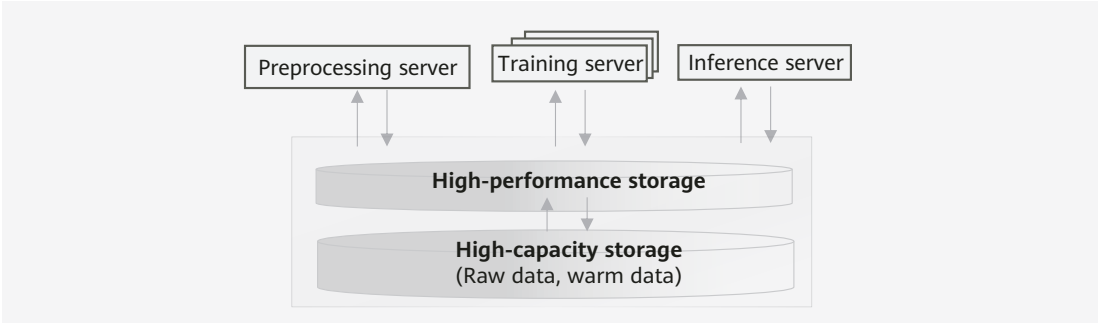


Figure 17: Decoupled storage-compute architecture for on-demand evolution

AI is advancing alongside computing power, algorithms, and data. The decoupled storage-compute architecture allows independent upgrades of compute and storage resources. Users can employ the latest processors or algorithms to boost computing performance, or apply new storage technologies to increase the data read speed without impacting any other part of the system. AI models and algorithms are evolving rapidly. The decoupled storage-compute architecture can easily adapt to these evolutions. For example, when a new AI model requires additional compute resources, users can add GPUs or TPUs without worrying about storage bottlenecks. The decoupling of compute and storage resources simplifies the integration of the latest technological advancements, such as a new neural network architecture or an optimized algorithm, only with an upgrade at the relevant compute or storage layer. This architecture also supports multi-tenancy environments, enabling different users to share compute and storage resources while maintaining isolation and privacy. Independent data storage is more focused on ensuring data resilience and backup while minimizing the risk of data errors and loss. A professional AI storage system featuring high performance (bandwidth and IOPS), flexible scaling, and robust reliability is essential for improving cluster availability.

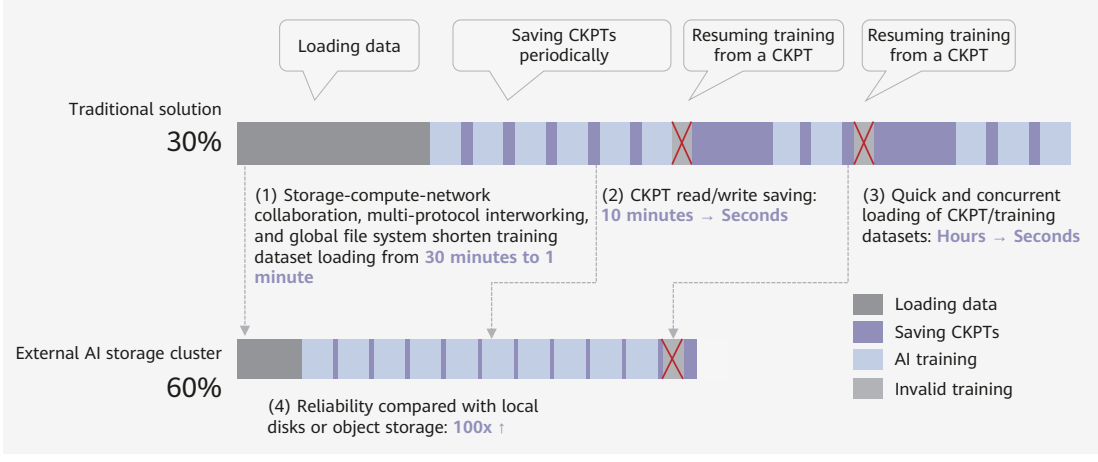


Figure 18: Reliable and professional AI storage improves cluster availability

2 The data infrastructure should have scale-out capabilities, enabling performance to increase proportionally to capacity

Large AI models have evolved from handling a single data type, such as texts, to processing multiple types of data, including texts, images, audio files, and videos. This shift towards multi-modality and even full-modality significantly increases the size of training datasets from TB to PB or even EB level. Meanwhile, the number of parameters in large AI models has surged from hundreds of billions to tens of trillions. This means that the demand for computing and storage resources has also increased. Storage systems must adapt to these changes, offering EB-level capacity and performance scaling in line with capacity expansion. As the model complexity increases, data access and preprocessing also become more complex. Storage systems must not only enable the high-speed storage and access for larger volumes of data but also support complex data processing tasks. Therefore, the systems must be able to scale out to additional GPUs, DPUs, and NPUs to accelerate I/O processing. The AI storage system should be designed with both high-performance and high-capacity layers and present a unified namespace so that data can be stored at different layers based on predefined policies set during initial write tasks. Additionally, data can be automatically migrated between the layers based on policies related to access frequency and time, which optimizes overall performance and capacity utilization. To meet the storage and access requirements throughout the entire AI process, the AI storage system must cover all phases—from data acquisition and preprocessing to model training and deployment. This simplifies data transfer and reduces the time and resources required for data migration. An ideal storage architecture should feature a fully symmetric design, eliminating the need for independent metadata service nodes. As the number of storage nodes increases, users can scale the system's total bandwidth and metadata access capabilities proportionally to ensure high performance during AI training.

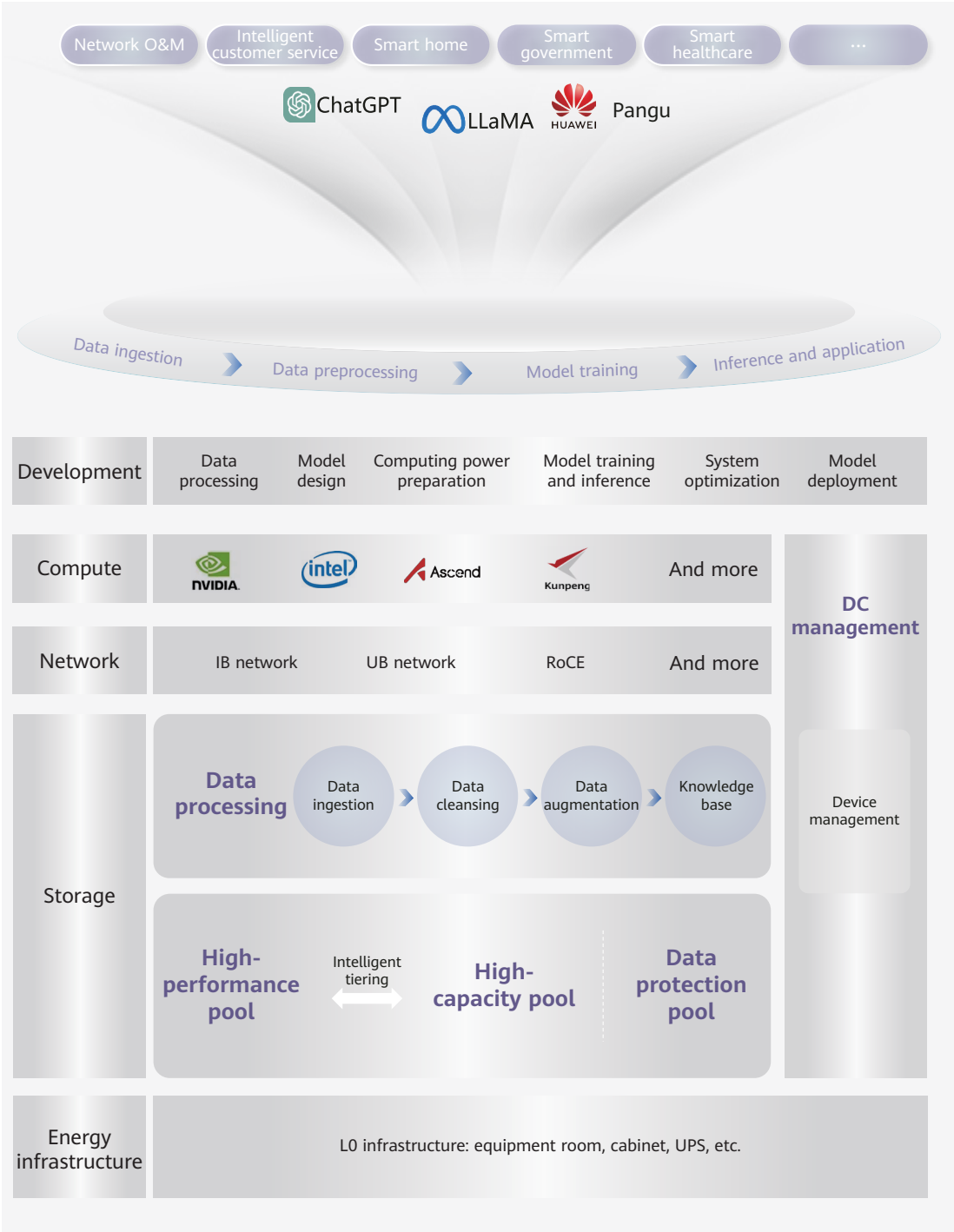


Figure 19: Huawei AI Data Lake Solution

3 The data infrastructure should support multi-protocol interworking

The data preprocessing phase of AI includes four key steps—data cleansing, integration, conversion, and mitigation. However, these steps often take up a lot of time and resources. The data preparation process must not only handle large volumes of data but also ensure accuracy and consistency. Given the diversity and complexity of data sources, various issues may arise during processing, such as data loss, inconsistency, and redundancy. These issues require careful handling and verification, making the data preparation phase one of the most time-consuming parts of an AI project. For example, preprocessing PB-level data can take several months.

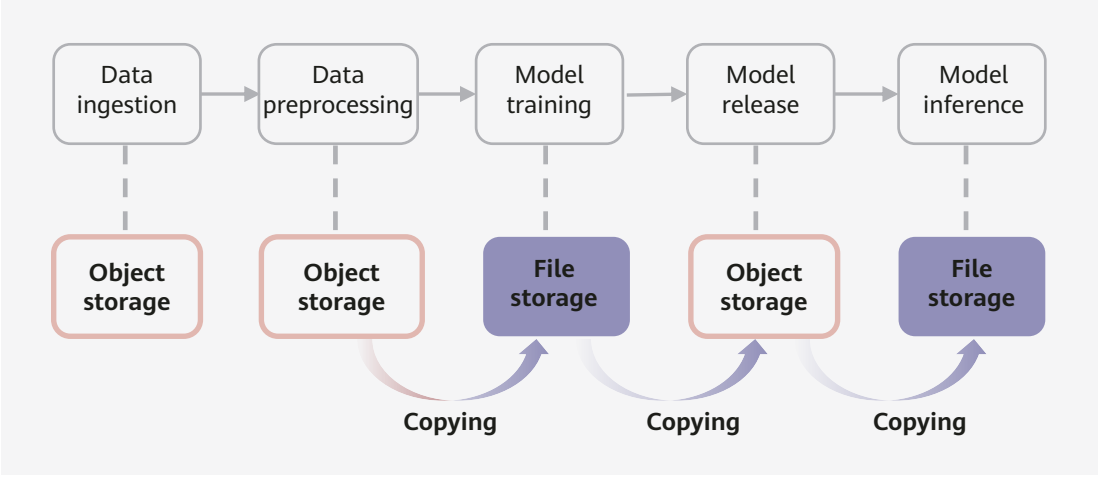


Figure 20: Multiple data copies due to varying protocols

As shown in Figure 20, due to varying data protocols, data must be copied between storage devices multiple times. Training preparation involves copying hundreds of millions of files, which can take anywhere from days to weeks.

For example, training Huawei's Celia voice assistant, which is powered by the Pangu model, involves processing 2 PB of raw data. During the data cleansing step, this data expands to over 30 PB to meet upstream service requirements, making this step last several months. Different protocols may be used throughout the AI tool chain. Excellent AI storage should support multiple protocols such as NAS, HDFS, and object and enable lossless semantics for each protocol, ensuring seamless ecosystem compatibility with native protocols. In addition, throughout each phase of AI training, zero data copy and zero format conversion are required. A global file system with multi-protocol interworking is essential for enhancing data preparation efficiency and preventing unnecessary data copying between data centers and devices. This also eliminates the

need for data copying throughout data processing, training, and inference, accelerating the deployment and parallel processing of big data and AI platforms while reducing wait times and performance losses. Moreover, storage systems must also support high-performance processing of dynamic hybrid loads. During data import, preprocessing, and model training, the storage system should be able to handle simultaneous read and write operations for both large and small files while maintaining high performance, particularly during the intensive write operations needed for checkpoint generation (see Figure 21).

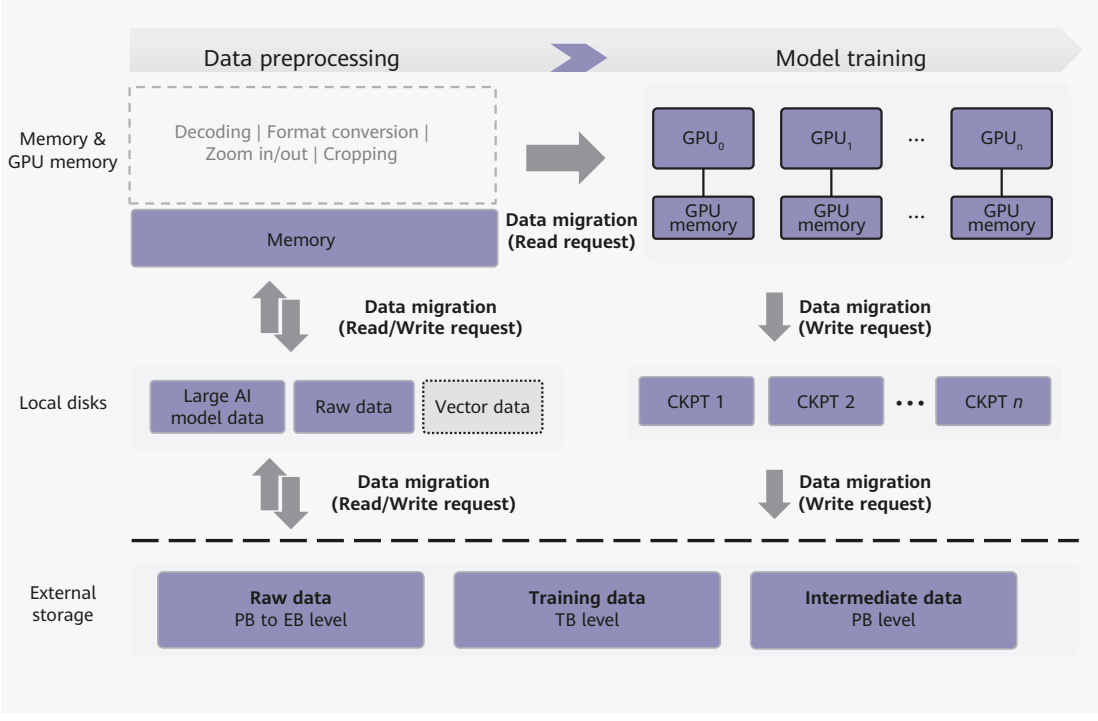


Figure 21: Impacts of PB-level data migration on overall efficiency during AI data preprocessing

3.2 Efficient Data Processing with All-Flash Storage

All-flash storage boosts data processing efficiency and unlocks data value

As large AI model computing clusters grow, the issue of idle computing power waiting for data becomes more pronounced. To optimize the utilization of computing power, data access efficiency must be accelerated. Additionally, the quick digital transformation driven by intelligent upgrades generates more service data, increasing the complexity and pressure on digital infrastructure.

All-flash storage improves data processing efficiency and meets service requirements in the digital-intelligent era, because it is able to meet the ever-evolving requirements for both digitalization and intelligentization. In addition, emerging data paradigms such as RAG and contextual long-term memory storage simplify data access and enhance computing with storage, thus improving the overall system performance.

3.2.1 Trends

1 Complex preprocessing of massive amounts of multi-source heterogeneous data requires comprehensive data governance

In today's digital environment, data can come from a host of sources, be it social media, IoT devices, online transactions, or sensor networks. As data redundancy and complexity grow, enterprises need to remove invalid and noisy data from vast, disordered datasets to identify valid features and retain valuable information. This requires more powerful and intelligent data processing technologies to store, analyze, and govern data.

Take autonomous driving training as an example. The vehicles require devices to record and be tested in extreme road and weather conditions so as to enhance behavior prediction and decision-making in different scenarios. One company, Waymo, has 100,000 miles of driving data stored in its public datasets comprising data from 1,000 test vehicles. A single test vehicle generates over 20 terabytes of data per day, and there

is a need for faster processes from test and verification to large-scale commercial use in a market with increasingly fierce competition. The automotive industry requires fast data reading to more efficiently extract high-quality algorithms (transfer learning, small sample learning, and self-supervised learning) from vast datasets and improve model adaptability.

This is also the case in healthcare, specifically, medical image analysis, such as CT scans, MRI, and X-rays. A whole-body CT scan generates thousands of images, resulting in gigabyte-scale data volumes, though not all this data is referenced in order to make a diagnosis. A tiny tumor or abnormal tissue may show up on a small area over several images, while the rest is considered healthy tissue but irrelevant information, meaning key lesion identification and screening must be more efficient.

Typically, traditional data storage systems focus on data access, management, and backup recovery, and prioritize relational databases and file systems to ensure data durability and accessibility. Today, however, enterprises have gradually shifted to comprehensive data governance, underlining the need for lifecycle management spanning integration, cleaning, labeling, protection, compliance, and value mining. This requires data to be integrated across systems, platforms, and devices into a unified environment, to realize unified visibility and analysis, cross-department data collaboration, and cross-region data sharing. In retail, integrated data (such as sales data, customer feedback, and inventory information) across online channels and offline stores provides a full-stack, omni-channel customer information view, which helps optimize inventory management and marketing strategies.

2 Large-scale computing power requires fast data access from data storage systems

Every generation of deep learning model contains more neural network layers and parameters than the previous ones, leading to high data dimensions and magnitude. Traditional data processing methods are inadequate for modern AI model training. Conventional relational databases and storage use indexes and relational models to process high-dimensional data (such as embedded vectors) and complex queries, but at a very low efficiency. Retrieving data from a million records takes between 1 and 5 seconds, while the vector database designed for high-dimensional data takes only tens of milliseconds.

Vector data enables fast similarity calculation and k-nearest neighbor (k-NN) search from millions of data points, which is essential for image retrieval, text matching, or other operations that process a large amount of data. In marketing and recommendation systems for e-commerce, for example, the similarity and relationship between a user and an item are calculated using feature vectors, which are then leveraged to provide personalized recommendations.

3 Real-time data processing is an essential requirement

AI technologies have now become the heartbeat of many industries, like finance (transactions), autonomous driving, and smart manufacturing, which all need incredibly fast processing of data. Consequently, beyond handling periodic static data like quarterly reports using traditional data analysis features, modern storage systems must be able to process and analyze dynamic data flows within milliseconds to inform decisions and provide differentiated advantages. For example, the Nasdaq stock exchange processes market data (stock prices, transaction volume, order information) from all over the world, handling millions of data packets and orders every second to execute real-time transaction decisions. Following the emergence of stream processing frameworks, such as Apache Flink and Kafka Streams, data storage must be able to integrate diverse data formats and quickly respond to data read and write requests. This will fuel real-time analysis and training and enable dynamic decision-making for AI systems.

3.2.2 Suggestions

1 Build a comprehensive data governance infrastructure

There are two major benefits of data storage's transformation to comprehensive data governance: quick ingestion and aggregation of mass heterogeneous data from multiple sources, and efficient extraction of training data from massive amounts of data using a data preprocessing tool chain.

Comprehensive governance can be divided into three layers. First, the device management layer, manages all data storage devices in a data center through unified O&M. The second, data management layer, uses the global file system to integrate data scattered across data centers into a unified data map to provide visualized management and scheduling. Last is the data filtering layer. Here, raw data is filtered, or preprocessed, and forms high-quality datasets so as to be efficiently processed by multiple analysis platforms, including AI ones.

2 Leverage all-flash storage and innovative semantics to provide data efficiently to computing systems

All-flash storage offers very fast data reads and writes that enable high IOPS and low latency, which are the bedrock of modern data centers that must provide ultra-fast real-time data processing and analysis.

With high performance, large capacity, and low power consumption of flash storage, both the centralized architecture (for relational databases) and scale-out architecture (for mass unstructured data) provide the perfect duo in terms of performance and capacity in a limited space. This meets the high-speed data access requirements for large-scale computing systems and unleashes its power.

Innovative data access semantics (memory, vector, etc.) shorten the path between computing systems and data for faster access to resources.

3 Unify data infrastructure platforms to implement efficient data transfer

Data lifecycle management ensures efficient and reliable handling across all stages, from generation and storage to processing, archiving, and destruction. Multi-protocol convergence and interworking allow data to be efficiently transferred across different storage and computing environments migration-free, which saves both time and resources, and also ensures data resilience and integrity during transmission. This further improves the data processing efficiency.

3.3 Intrinsic Resilience of Storage: A Critical Requirement

Storage is the foundation of data resilience and must start strong

While intelligentization boosts digitalization, enabling the generation of more valuable service data, it also comes with the risk of more frequent ransomware attacks.

Against this backdrop, we need to build a data resilience system that has both defense and response mechanisms for data-generating digitalization and ever-evolving intelligentization. By leveraging the intrinsic resilience of storage, this system transitions from reactive responses to attacks to proactive and comprehensive protection.

3.3.1 Trends

1 Growing data volume and limited backup window call for powerful backup systems

The boom of large AI models such as ChatGPT and Pangu has created a demand for data mining capabilities to find hidden patterns and insights in huge volumes of structured and unstructured data. The associations, trends, and rules revealed can then be used to train large AI models and inform decisions. The push for data value mining encourages users to collect frequently used and high-dimensional data, leading to an exponential increase in data volume and a significant boost in its value.

The explosive growth of data poses new challenges to data backup. Take short-term data retention for example. Backup storage needs to back up more high-value data in the same time window. This is possible but requires enhanced backup media, like all-flash systems, and next-gen architecture, such as those using deduplication and compression algorithms to store more data, or a data appliance that features data passthrough. For long-term data retention, many AI models invoke historical data for repeated training, where the same data is copied multiple times and used in different environments. In doing so, backup and archive media must go beyond simple data

retention and automatic tiering of warm and cold data but instead provide rapid transitions between backup and archive data. In this landscape, current products adopt an integrated backup and archive architecture for storing data with both short- and long-term retention needs, with the automatic tiering function enabling rapid recovery of long-term retention data.

2 A comprehensive data protection strategy is imperative as AI makes ransomware attacks easier to launch

The rise of generative AI has significantly advanced the automation of traditional data protection, but at a cost. Ransomware variants have evolved so much that it's easier than ever for cybercriminals to launch attacks. Research shows that with the emergence of tools such as WormGPT and FraudGPT, generative AI has increased the number of phishing email attacks by 135%. This is supported by the latest market research reports, which indicate the wide scope of generative AI and cloud corresponds to a sharp increase in bad bots, which account for 73% of the total Internet traffic. In one case, a Japanese man with no professional IT knowledge developed ransomware capable of encrypting computers and demanding ransoms, simply by using generative AI.

Indeed, generative AI has optimized the attack method, which has made it increasingly difficult to identify. Consider how automated botnet attacks enable attackers to quickly and accurately scan for vulnerabilities or launch attacks on networks, greatly increasing the impact and effectiveness of cyberattacks. In November 2023, a hacker organization in China used ChatGPT to optimize ransomware programs, scan vulnerabilities, obtain licenses by penetrating data protection systems, and implant ransomware. This specific attack paralyzed all servers of the target company and was followed by a ransom demand.

3.3.2 Suggestions

1 Use all-flash backup storage to enhance backup efficiency

All-flash solutions are essential to improving backup efficiency: All-flash media for faster backup and recovery within the original time window; deduplication and compression algorithms to maximize storage capacity and back up and restore more data; 3-in-1 architecture (backup software, server, and storage) with data passthrough feature to boost reliability and avoid link disconnections common in traditional server stacking

setups. In environments with both long- and short-term data retention needs, an integrated backup and archive architecture enables data tiering without compromising performance and seamless transitions between backup and archive data.

2 Build a multi-layer ransomware protection system that combines both defense and response mechanisms to transition from reactive to proactive protection

The combination of storage and network infrastructure provides multi-layer detection and end-to-end protection against ransomware attacks. The network-storage collaborative design helps effectively prevent (pre-attack), accurately identify (in-attack), and quickly recover (post-attack) from attacks, ensuring that ransomware protection is no longer reactive but proactive. Threats are detected and intercepted quickly before they can strike, protecting data from unauthorized encryption and theft. The collaborative design enables the storage devices to be used for fast, secure data recovery, building a comprehensive ransomware protection system.

3.4 AI Data Lakes Enable Visible, Manageable, and Available Data

Building an AI data lake foundation to eliminate data silos and make data visible, manageable, and available

The increase in the size of AI computing clusters has made the ability to manage massive amounts of multi-source heterogeneous data a major challenge. Data map drawing, data ingestion, data preprocessing, hierarchical management of mass data, and data resilience should be the main focal points during large-AI-model training.

Building an AI data lake foundation for digital and intelligent transformation and eliminating data silos through data fabric can enable sufficient storage, free mobility, and full utilization of mass multi-source heterogeneous data.

3.4.1 Trends

1 Data is becoming the differentiating factor that determines AI competitiveness

The industry has reached a consensus that there can be no AI without sufficient data. The quantity and quality of data determine how far AI can evolve. According to the *2023 Global Trends in AI Report*, the orderly and effective management of data assets is the top challenge in building AI infrastructure, and that is followed closely by data resilience and compute performance challenges. In the future, 20% of the potential and effectiveness of large AI models will be determined by algorithms and 80% will be determined by data. Model rankings from DataLearner show that Meta's Llama 3 model, which has 70 billion parameters and uses 15 trillion tokens, scored 82 points, while their Llama 2 model, which has 70 billion parameters and uses two trillion tokens, only scored 68.9 points. Enterprises need to monitor the accumulation of core data assets such as industry data and daily operations data. Sufficient and high-quality data will help enterprises significantly improve the effectiveness of AI training and inference.

Finance	Credit assessment	Investment consulting	Personalized recommendation	Risk assessment	Programming assistant	Compliance management
Healthcare	Self-service consultation	Electronic medical record	Drug-effect evaluation	Healthcare assistant	Genome sequencing	Epidemic warning
Government and public service	Digital twin city	Critical incident warning	Intelligent report generation	Intelligent meeting assistant	Government affairs assistant	Government hotline
Internet	Creative collaboration	AI-assisted search	Online translation	Marketing copywriting	Online education/training	Online Q&A
Manufacturing	Industrial quality inspection	Production resource planning	Industrial robot	Predictive maintenance	Knowledge graph	Supply chain management
Electric power	Fault diagnosis	Line inspection	Distribution network operation optimization	Scheduling calculus	Electricity consumption forecast	Statistics report
Oil & Gas	Fault identification	Reservoir layer prediction	Reservoir sweet spot identification	Smart construction site	Chemical refining	Intelligent review
Education	Subject dialogue assistant	Consolidation learning assistant	Exercise recommen-dation assistant	Chinese writing assistant	English writing assistant	Math calculation assistant
Transportation	Transportation planning	Accident prevention	Congestion alleviation	Freight supervision	Hub management	Parking violation handling
Carrier	Intelligent customer service	Telecom fraud prevention	Expense inspection	Network planning & optimization	Virtual secretary	XR call

Figure 22: Service scenarios of large AI models in various domains

2 Managing data assets is an essential part of implementing AI practices

Data quality is one of the core issues of data asset management. More than 80% of the entire AI workflow is focused on preparing high-quality data. Most enterprises use data from multiple sources. The quality of data varies between sources, making it difficult to quickly prepare large volumes of data for AI model training. Warehousing key data assets and list-based management are important data asset management strategies for enterprises working to implement AI practices.

High-quality Q&A pairs can significantly improve the fine-tuning effect on large AI models during model training. Manually generating Q&A pairs tends to be inefficient and can produce inconsistent quality. Self-QA and Self-Instruct technologies are used in the industry to automatically generate a corpus of high-quality Q&A pairs.

RAG is key to improving the inference accuracy of large AI models. Enterprises need to vectorize data assets and store them in the vector database for efficient information retrieval and generation with RAG.

3 More and more industries are starting to use large AI models for inference

With the increase in the number of parameters and the context length of large AI models, the capacity of the vector retrieval library increases from tens of millions to billions. As a result, the retrieval latency increases and precision decreases, and index reconstruction takes several weeks, thereby hindering the commercial use of model inference. The context length determines the memory and inference capabilities of

large AI models. Long sequence inference can enrich semantics and help generate more coherent and accurate content, and ultra-long sequences have been widely used for large AI model inference. However, many challenges, such as insufficient inference computing power and slow inference response, have come with the use of long sequences. Therefore, the lossless long sequence has become a main focus. Since the single-server inference mode cannot meet service requirements, cluster-based inference must be used. Additionally, the multi-layer KV-Cache technology is built by leveraging independent external storage with strong consistency to provide the long-term KV-Cache for the inference cluster. This simulates the human brain's thinking process. The use of queries to eliminate repeated computing and the sharing of inference process and result data help reduce the pressure on inference computing power. The efficiency and cost of model inference have become the determinants of competitiveness.

4 Enterprises that effectively use AI are gaining a competitive edge

The use of large AI models is evolving from general applications such as knowledge Q&A, and text-to-image and text-to-video conversion, to comprehensive applications that feature large AI models + Copilot assistance + Agent autonomous decision-making. The enterprises that can skillfully evaluate the capabilities of large AI models and find ways of optimizing and using them are becoming more competitive.

Take the financial sector as an example. AI model technologies are helping banks perform precise customer profiling so that they can provide better personalized recommendations and customized services. Human-machine interactions are streamlining processes for intelligent customer services and enabling intelligent bank branches, which greatly improves user experience.

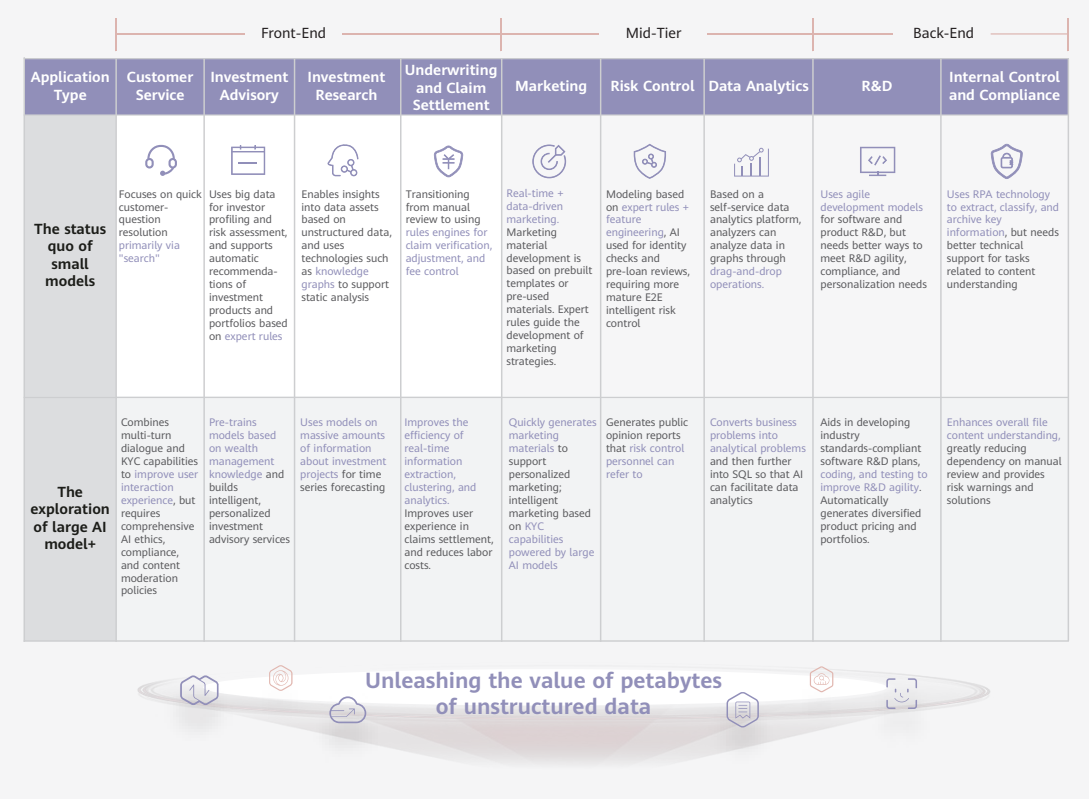


Figure 23: Unleashing the value of petabytes of unstructured data

In the healthcare industry, large AI models are already being used to improve patients' pre-hospital experience through intelligent registration and triage. Image-assisted diagnosis and treatment, pathological diagnosis, and precision medicine can reduce doctors' workloads and improve the efficiency and quality of diagnosis. After diagnosis, AI can help patients with health management through functions like knowledge Q&A, thus driving the shift from reactive treatment to proactive prevention.

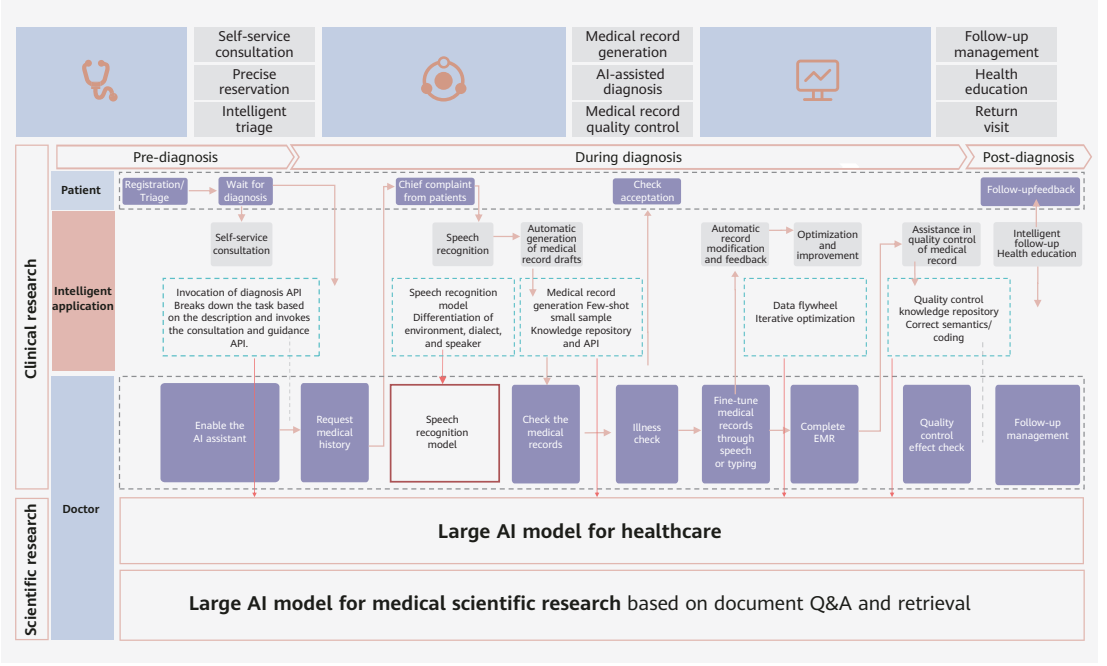


Figure 24: Workflow of the large AI model for healthcare

3.4.2 Suggestions

1 Build a unified AI data lake to make data assets visible, manageable, and available

More industry and enterprise knowledge is required to iterate and upgrade large AI models. Currently, huge amounts of enterprise data assets are scattered across branches and production sites. This multimodal data may come from service systems in different regions, business or ecosystem partners, or different vendors' public or private clouds. This may result in data silos which restrict the development of large AI models.

Enterprises need to build a unified data lake foundation to make data assets visible, manageable, and available across regions. First, they need to create a unified data asset view that enables global visibility and real-time updates of complex data across regions, sites, and vendors. They also need to enable intelligent data cataloging, which includes automatic data labeling, aggregation, retrieval, and display, and allows for automatic data classification and grading by content, compliance level, and access frequency. To truly make data available, they need to achieve storage-compute-network collaboration to enable efficient access to and processing of ingested data. Large AI models need to be continuously fed data in order to effectively serve enterprises' service systems. This requires unified data scheduling across organizations, regions, and applications.

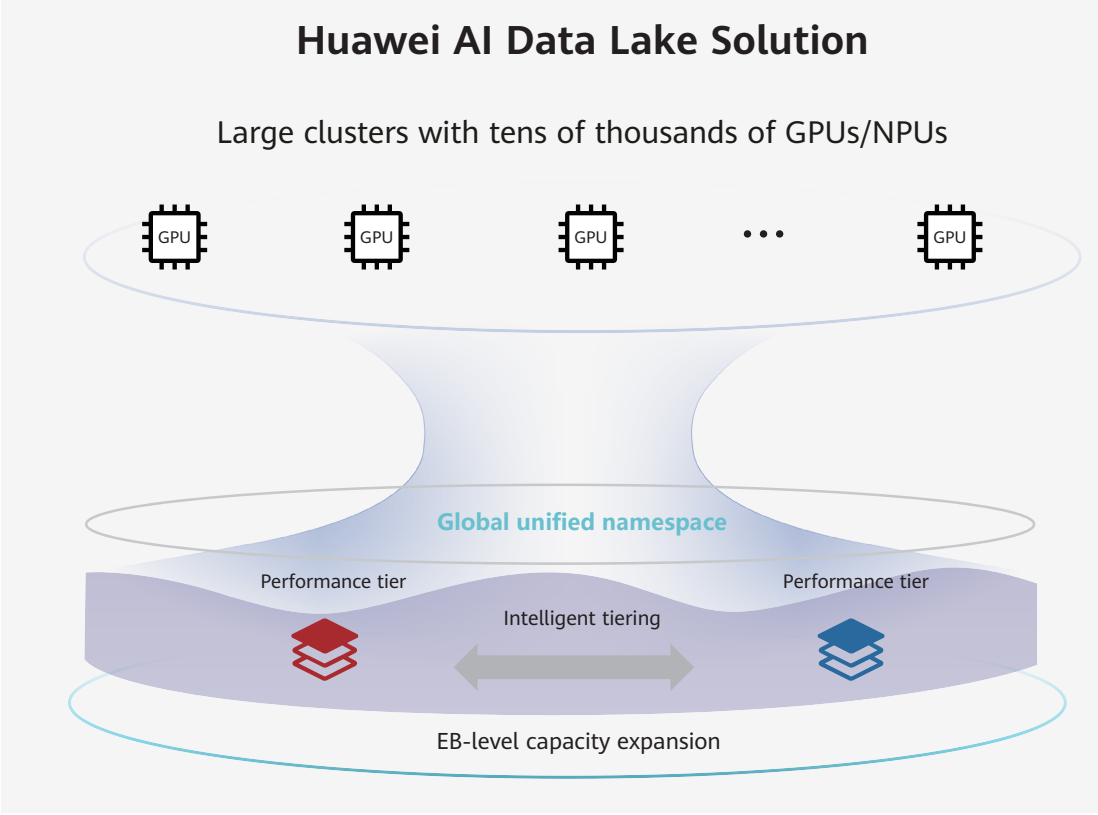


Figure 25: Huawei AI Data Lake Solution

2 Choose professional AI storage for model training to improve computing power utilization and maximize the returns on AI investment

For large AI models, scaling laws continue to prove effective. The number of model parameters is growing from hundreds of billions to trillions, the scale of clusters is expanding from thousands of GPUs to tens of thousands, and training datasets are growing from terabytes to exabytes. This increases the technical complexity of large AI models and the volume of data for processing, which means that more frequent retraining and tuning of larger-scale models are required. Inadequate AI infrastructure can lead to additional costs for enterprises undergoing intelligent transformations. NVIDIA is working with storage vendors on high-performance AI training clusters powered by a standard file system and the Share Everything storage architecture. In its technical proposal for a next-generation intelligent computing center, Oak Ridge National Laboratory argues that only AI-optimized storage can deliver the performance and reliability that large AI models need to process exabytes of data.

Enterprises need to carefully plan out their intelligent computing foundations and select

dedicated AI storage for AI workload optimization to improve cluster utilization. In this way, they can move away from simply stacking computing power and start to fully unlock its potential. They also need to optimize storage cluster performance, and select high-performance and reliable external AI storage. This can improve cluster utilization by over 10% and reduce unnecessary investment due to idle computing power.

3 Adopt technologies such as RAG and long sequence to improve the performance and accuracy of model inference

As enterprise knowledge and data change frequently, periodic training of large AI models cannot ensure timeliness and accuracy. The RAG technology is widely used for the AI-based reconstruction of enterprise applications when the benefit of the application outweighs the reconstruction costs. When generating a result, the large AI model retrieves related knowledge from the database and generates an answer with references, thereby improving the credibility of the inference result. In the inference phase, multi-turn dialogue and context processing depend on the model's memory capacity. The three-layer cache mechanism that covers xPU, DRAM, and external storage SSD can extend the memory period of the model from hours to years to improve the inference accuracy. In addition, historical results can be queried instead of repeating the inference to reduce computing power consumption.

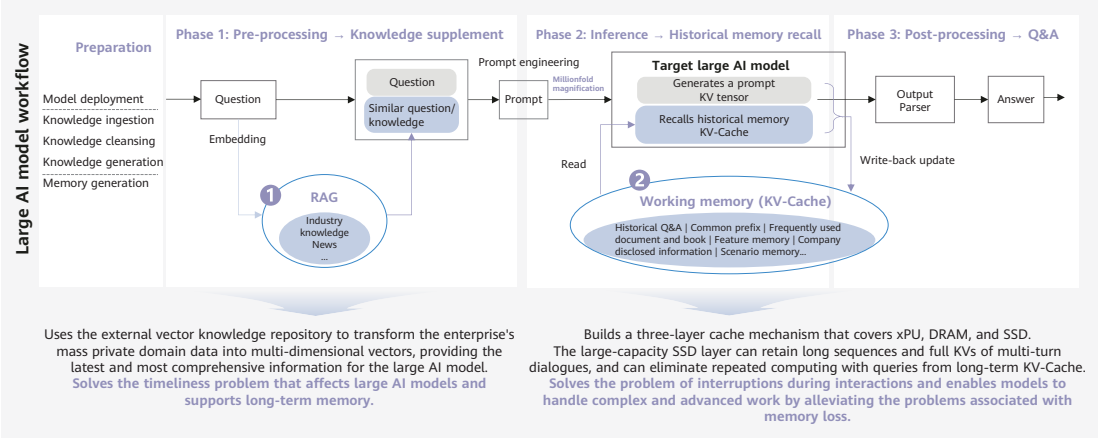


Figure 26: Technical architecture of RAG and KV-Cache

4 Strengthen classified and hierarchical data protection by means of disaster recovery, backup, and ransomware protection

Large AI models are trained based on mass data that contains sensitive information such as personal data and enterprises' private domain production data. A wide range of data resilience risks emerge as large AI model technologies are developed. Sample data poisoning attacks can cause models to produce misleading results which can significantly affect decision-making. The theft of model files can render substantial investments worthless. Ransomware attacks that encrypt training data can interrupt large AI model training and jeopardize production resilience.

Enterprises need to focus on the classified and hierarchical management of data assets, determine data owners and users, ensure data compliance, and build a comprehensive resilience solution that covers management, applications, networks, and storage. As the final carrier of data, storage can provide a full suite of intrinsic resilience solutions, including storage hardware and software system resilience, data DR and backup, ransomware protection, and resilience management to form the last line of defense for data.

5 Implement an AI-talent cultivation mechanism and organize practical activities on large AI models

The use of large AI models is evolving from general applications such as knowledge Q&A, and text-to-image and text-to-video conversion, to more complex applications that feature large AI models + Copilot assistance + Agent autonomous decision-making. Large AI models are shifting from being in peripheral support roles in enterprises to playing key roles in production and in improving operational efficiency.

Enterprises should thoroughly evaluate their readiness for generative AI applications from the perspectives of top-level design, organizational structure, talent, and teams. For top-level design, enterprises should develop methods for evaluating and tracking the use of open-source large AI models, data, and training models. Additionally, they should learn from the current best practices for AI infrastructure. For organizational structure, enterprises should establish dedicated teams for data resilience, privacy, and ethics. For cultivation of talent and teams, they should create a talent development system to train more professionals with a thorough understanding of and practical experience in large AI models, and particularly in storage for large AI models.

3.5 Training/Inference Appliances for Accelerating the Deployment of Large AI Models Across Industries

Training/Inference appliances streamline the deployment of large AI models, driving digital and intelligent transformation across a vast range of industries

The AI boom is in full swing, and various industries are striving to integrate AI into their applications but encounter challenges in infrastructure deployment, model selection, secondary training, and fine-tuning. The training/inference appliances come pre-integrated with infrastructure and tools, designed in collaboration with large-AI-model vendors to enable fast AI deployment and accelerate digital and intelligent transformation across diverse industries.

3.5.1 Trends

1 Data quality may not be consistent and preparation can take a long time

The process of transforming extensive raw enterprise data into usable datasets is complex and time-consuming. One of the primary challenges is collecting sufficient amounts of representative and high-quality data, as this often requires integrating data from multiple sources. Additionally, the process of cleaning data, which involves removing noise, errors, and duplicates, is labor-intensive and requires significant amounts of time and effort. Accurate data labeling is another critical challenge. This step is essential for model training and requires professional involvement, further extending the preparation time. Moreover, data collected from different departments can differ in quality during the preparation phase and this can negatively impact the performance of large AI models.

2 Hardware selection is difficult, delivery takes a long time, and O&M is costly

Large AI model applications require specialized hardware, including computing, storage, and network devices. However, selecting hardware is challenging due to the variety of options and complex performance parameters involved. Additionally, the process

of assembling, commissioning, testing, and running hardware is complex. Monitoring, maintenance, and upgrades further complicate things and drive up operational costs.

3 Large AI models can suffer from serious hallucination problems which can affect inference accuracy

In complex scenarios, large AI models can produce distorted outputs, leading to what is known as model hallucination. This issue not only reduces the accuracy of the models, but can also have serious consequences when incorrect information is used in critical decision-making. In academic research and knowledge dissemination, such inaccuracies may mislead readers and researchers, potentially resulting in ethical and legal risks.

4 Data resilience is not guaranteed, and core data assets such as models are prone to leakage

High-value industry data is one of the core assets of enterprises. As a result, there are high requirements on data resilience. For model vendors, industry-specific models are core components that enable enterprise applications, so they need to ensure resilience and reliability while avoiding the risks of model leakage. Key resilience challenges include:

(a) Data privacy: Training data may contain sensitive information, such as personal identity information and financial data.

(b) Model resilience: Models are vulnerable to attacks, such as parameter tampering or malicious code injection, which can compromise outputs and reliability.

(c) Adversarial attacks: Attackers might use adversarial examples to deceive models, leading to incorrect outputs.

(d) Model explainability: The black-box nature of large AI models complicates output interpretation, potentially undermining model trust and reliability.

(e) Model sharing: Sensitive information, such as model parameters and training data, can be exposed during model sharing.

(f) Model deployment: During deployment, models face resilience threats like network attacks and malware injection, which can compromise their resilience and reliability.

5 It may take a long time to see a strong return on investment (ROI)

As large AI models are still in their exploration phase, the expected ROI on hardware and software may fall short, and this can result in operational costs that exceed the budget. In tasks like large-scale data processing, graphics rendering, and deep learning, low GPU utilization can hamper efficiency and delay task completion. These inefficiencies impact the speed of innovation and development for enterprises and research institutes, and affect product launches and the publication of research results.

3.5.2 Suggestions

1 Pre-integrate a data preprocessing tool chain to quickly generate high-quality training datasets

High-quality data is the cornerstone for AI to achieve accurate inference. Vendors of professional AI storage typically provide data preparation tool chain components with dozens of high-performance AI operators to automatically clean data (such as parsing, filtering, deduplication, and replacement) in various formats. These operators help enterprises swiftly transform raw data into valuable datasets.

2 Deploy full-stack, pre-integrated training/inference appliances for large AI model applications across industries

The training/inference hyper-converged appliance (HCI appliance) is an out-of-the-box solution. Pre-integrated and pre-tuned, it provides all the necessary compute, storage, and network components without the need for complex model selection, assembly, and setup. This significantly reduces time and labor costs. Many vendors, including Huawei, offer pre-configured appliances with GPU/NPU servers, networks, and professional storage. In addition, full-stack device management software is pre-installed to manage and maintain key hardware and software components, including compute, storage, network, and container platforms. This reduces the routine O&M workload for IT teams, allowing them to focus on AI services rather than infrastructure maintenance.

Many of these appliances provide high-performance confidential execution environments and confidentiality protection measures to safeguard data and models. With built-in data protection and ransomware defense mechanisms, the core data of enterprise users is protected against leaks and damage.

Moreover, most appliances support horizontal scaling, enabling enterprises to combine multiple appliances into a larger, integrated training and inference platform. This scalability allows for flexible deployment of large AI model applications, spreading investment over multiple cycles and mitigating ROI risks.

3 Use RAG knowledge base to reduce hallucinations for accurate inference

Training/Inference appliances are equipped with a knowledge base embedded with high-quality datasets. The built-in RAG tool can retrieve relevant information from the base, guiding the inference process to the correct context and effectively reducing hallucinations during user interactions. With real-time updates to the knowledge base, the model's responses remain up-to-date. Additionally, a built-in evaluation component continuously assesses and tracks inference accuracy, ensuring reliable model performance.

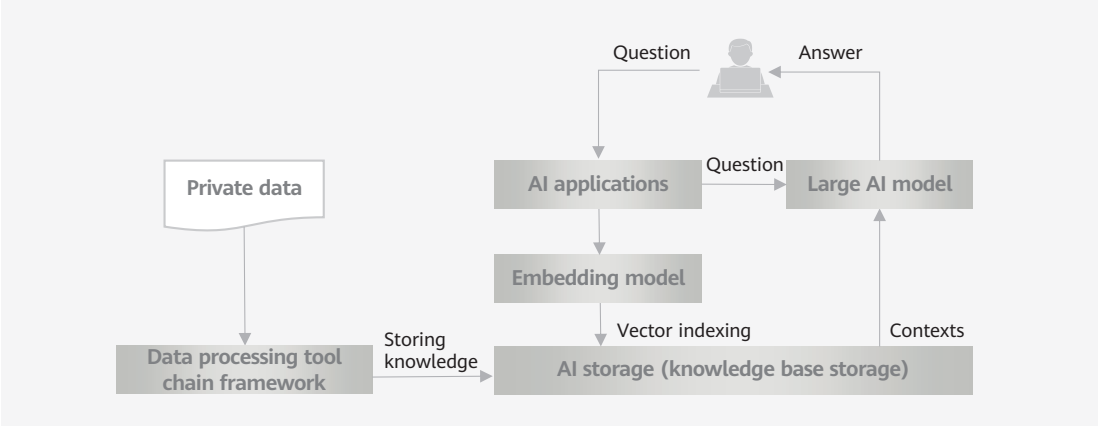


Figure 27: Working process of the RAG knowledge base

4 References

1	Government AI Readiness Index 2023 https://oxfordinsights.com/ai-readiness/ai-readiness-index/
2	Risks and Challenges of Integration of Artificial Intelligence in the Public Sector Governance https://www.secrss.com/articles/50686
3	Thoughts on Introducing Artificial Intelligence in Tax Risk Analytics https://mt.ucass.edu.cn/info/1035/1904.htm
4	2023 Analysis Report on Ransomware Attacks Targeting Chinese Enterprises https://www.qianxin.com/news/detail?news_id=10812
5	2023 White Paper on Artificial Intelligence Development in Telecommunications https://www.iotku.com/news/88956963267805184O.html
6	White Paper on the New Generation of Artificial Intelligence Infrastructure https://www.sensecore.cn/whitepaper.pdf
7	Research on Typical Application Scenarios and Standards System of Artificial Intelligence in Intelligent Manufacturing https://www.engineering.org.cn/ch/10.15302/J-SSCAE-2018.04.018
8	2024 China AI+Manufacturing Industry Research Report | 36Kr Research Institute https://36kr.com/p/2853458940627588
9	Generative AI in Manufacturing: Application Paradigms, Trends, and Challenges https://www.tisi.org/27034
10	How Can Generative AI Optimize the Production Process for Manufacturing Enterprises? https://m.huxiu.com/article/3082157.html
11	Special Report on AI-enabled Manufacturing: Focusing on Nine Key Segments https://www.sohu.com/a/673958291_121124366
12	What is Synthetic Data? https://aws.amazon.com/what-is/synthetic-data/?nc1=h_ls
13	Status Quo, Prospects, and Challenges of Synthetic Data Technologies https://www.secrss.com/articles/55976
14	Data-centric Artificial Intelligence: A Survey https://arxiv.org/pdf/2303.10158
15	AI Development Prospects in Banking: From Customer Service to Risk Management https://finance.sina.cn/2023-02-24/detail-imyhuqpt1751967.d.html

16	Financial Technology Competition in the Banking Industry Intensifies, and the Large AI Model Has Become a New Competition Point https://xhrbcj.com/newsDetail?id=d89bc8524614dbf5aa38dd0e92b16757&type=2
17	Agricultural Bank of China Unveils ChatABC: An Independent Financial Large AI Model Application https://www.cebnet.com.cn/20230331/102868614.html
18	ICBC Launches the First Foundation Model for Financial Industry: Wide Applications Powered by Ascend AI https://ai.qianjia.com/html/2023-04/04_400384.html
19	Application of Large AI Models in Banking: A Case Study of Shanghai Pudong Development Bank https://new.qq.com/rain/a/20240131A09C0400#:~:text
20	Financial Storage Infrastructure Development Research Report https://www.yunduijie.com/businessreport/view/3956.html
21	The Future Is Coming: Global Rankings and Prospects of Cities Leading in AI Innovation and Convergence https://www.baogaozhan.com/57705.html
22	Global Financial Stability Report https://www.imf.org/en/Publications/GFSR/Issues/2024/04/16/global-financial-stability-report-april-2024?cid=bl-com-SM2024-GFSREA2024001
23	2023 Global Trends in AI Report https://www.weka.io/wp-content/uploads/files/resources/2023/08/2023-Global-Trends-AI-Report.pdf
24	Comprehensive Evaluation Comparison Table for Large Language Models https://www.datalearner.com/ai-models/llm-evaluation
25	OLCF-6 Request for Proposals https://www.olcf.ornl.gov/draft-olcf-6-technical-requirements/
26	Shanghai Jiao Tong University Supercomputing Platform User Manual https://docs.hpc.sjtu.edu.cn/
27	The Peking University Institute of Advanced Agricultural Sciences Builds a Genomic Analysis Platform: How Seeds Drive the World https://e.huawei.com/cn/case-studies/solutions/storage/institute-of-advanced-agricultural-sciences-peking-university
28	How AI is Creating Explosive Demand for Training Data https://www.unite.ai/how-ai-is-creating-explosive-demand-for-training-data/

29	Global Data Center Energy Consumption to Double by 2026 https://www.stdaily.com/index/kejixinwen/202401/a3bb743e34134d159c5e7f1e50069ce7.shtml
30	AI Is Driving the Future of Renewable Energy Development https://www.sas.com/zh_tw/insights/articles/analytics/ai-renewable-energy.html
31	60 Hurts per Second – How We Got Access to Enough Solar Power to Run the United States https://www.bitdefender.com/blog/labs/60-hurts-per-second-how-we-got-access-to-enough-solar-power-to-run-the-united-states/
32	Scam email cyber attacks increase after rise of ChatGPT https://technologymagazine.com/articles/scam-email-cyber-attacks-increase-after-rise-of-chatgpt
33	Bad Bots Account for 73% of Internet Traffic: Analysis https://www.securityweek.com/bad-bots-account-for-73-of-internet-traffic-analysis/
34	Growing Frequency of Hacker Attacks Targeting Citizens' Personal Information http://www.chinapeace.gov.cn/chinapeace/c100047/2024-01/05/content_12705164.shtml

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P. R. China
Tel: +86-755-28780808
www.huawei.com

Editorial Consultant:
Peter Zhou
Managing Editors:
Xiao Degang, Pang Xin, Yang Bailiang, Fang Weifeng, Fan Jie, Wang Zhen, Shen Shen, Sun Rui, Xue Han, Huang Ting, Luo Yi, Qiu Donghua
Editors-in-Chief:
Gong Tao, Qiu Fangjia
Section Editors:
Hua Xian, Han Mao, Pang Liangshuo, Chen Yang, Gao Tan, Miao Yonggang, Jiang Huahu, Chen Zhenhua, Feng Zhen, Li Wenxiu, Liu Weiqi, Chen Hui, Lan Guoping, Chen Lin, Yu Leqing, Zeng Fan, Yuan Yanlong, Cao Xiaohui, Fan Jue
Translators:
Wang Jing, Yu Shanshan, Lu Shasha, Liu Jiawen, Tan Chuan, Mei Wuzhi, Yin Yingying, Guan Hanwen, Luo Qianrui, Xiao Yue, Wu Mengni, Chen Xin, Li Jingwen, Gong Qinglu, Daniel Mark Curran, Leo Stephen Gallagher, Kyle Melieste, Gavin Wills, CHRISTOVA EKATERINA ROUMENOVA, YOUNG MEGAN MARY, Jimmy Ding
(The names above are listed in no particular order.)