HONG KONG MONETARY AUTHORITY
香港金融管理局

Practice Guide on Data Excellence

October 2026

Contents

01 Introduction

02 The Need for Data Excellence

03 DataGRID Framework and Evolving Areas of Focus

04 Key Considerations and Suggested Good Practices

05 Looking Ahead

06 Appendix

01 Introduction

1.1 Background

The Hong Kong Monetary Authority ("HKMA") has identified Data Excellence, together with Cyber Resilience, as foundational pillars of its "Fintech Promotion Blueprint".[1] Published under the HKMA's broader "Fintech 2030",[2] the Blueprint provides a tactical framework for fostering responsible innovation and advancing the adoption of more fit-for-purpose technologies. These efforts contribute to the HKMA's broader objective of reinforcing Hong Kong's position as a robust, resilient, and future-ready international fintech hub.

The New Risk Data Strategy is one of four flagship projects under the Fintech Promotion Blueprint. Data has long played a central role in financial services, supporting business, risk management, and regulatory reporting needs. With continued advances in artificial intelligence ("A.I."), data analytics and digital technologies, data is increasingly recognised as a strategic asset that can help Authorized Institutions ("AIs") respond to evolving business needs, enhance operational efficiency, and strengthen risk management.

As organisations face growing demands for greater data volume, velocity, and variety, effective data management is becoming increasingly complex. Given that the quality of underlying data directly affects the trustworthiness of analytics and technology-enabled applications, strengthening data capabilities and governance to enable secure, high-quality foundations for innovation (i.e., achieving Data Excellence) is crucial to supporting responsible and sustainable innovation in the banking sector.

In support of the new strategy, the HKMA has developed this "Practice Guide on Data Excellence" ("Guide"), drawing on industry engagement, supervisory observations and relevant market practices. The Guide sets out the evolving areas of focus for Data Excellence and provides practical reference points to assist AIs in strengthening their data management and governance arrangements.

The Guide is intended to complement, and should be read in conjunction with, applicable HKMA supervisory requirements and guidance. It does not introduce a separate set of prescriptive requirements or a single mandated implementation model. Rather, it is intended to assist AIs in considering how their governance, data quality management and control, and data infrastructure arrangements can support reliable, resilient and responsible use of data, in a manner proportionate to the nature, scale and complexity of their operations and the criticality of relevant data and use cases.

AIs are encouraged to view Data Excellence not merely as a compliance exercise, but as a strategic capability. Robust data governance and management can enhance risk management, bolster institutional resilience, support informed decision-making, and enable the responsible adoption of advanced technologies. By establishing reliable and fit-for-purpose data foundations, AIs can more effectively address their business and risk management needs while supporting responsible innovation.

[1] HKMA. 2026. Hong Kong Monetary Authority - HKMA unveils Fintech Promotion Blueprint: From adoption to advancement. https://www.hkma.gov.hk/eng/news-and-media/press-releases/2026/02/20260203-3/.

[2] HKMA. 2026. Hong Kong Monetary Authority - The HKMA Unveils "Fintech 2030" at the Hong Kong FinTech Week 2025. https://www.hkma.gov.hk/eng/news-and-media/press-releases/2025/11/20251103-3/.

1.2 Purpose

This Guide has been developed as a practical reference to support AIs in pursuing Data Excellence. It highlights key evolving areas of focus and provides illustrative reference points and suggested good practices, rather than prescribing a single standard or implementation approach. AIs are advised to exercise appropriate judgement in determining how these practices may be applied, taking into account their circumstances, including the nature, scale and complexity of their operations, their level of data maturity, and the criticality of the relevant data and use cases.

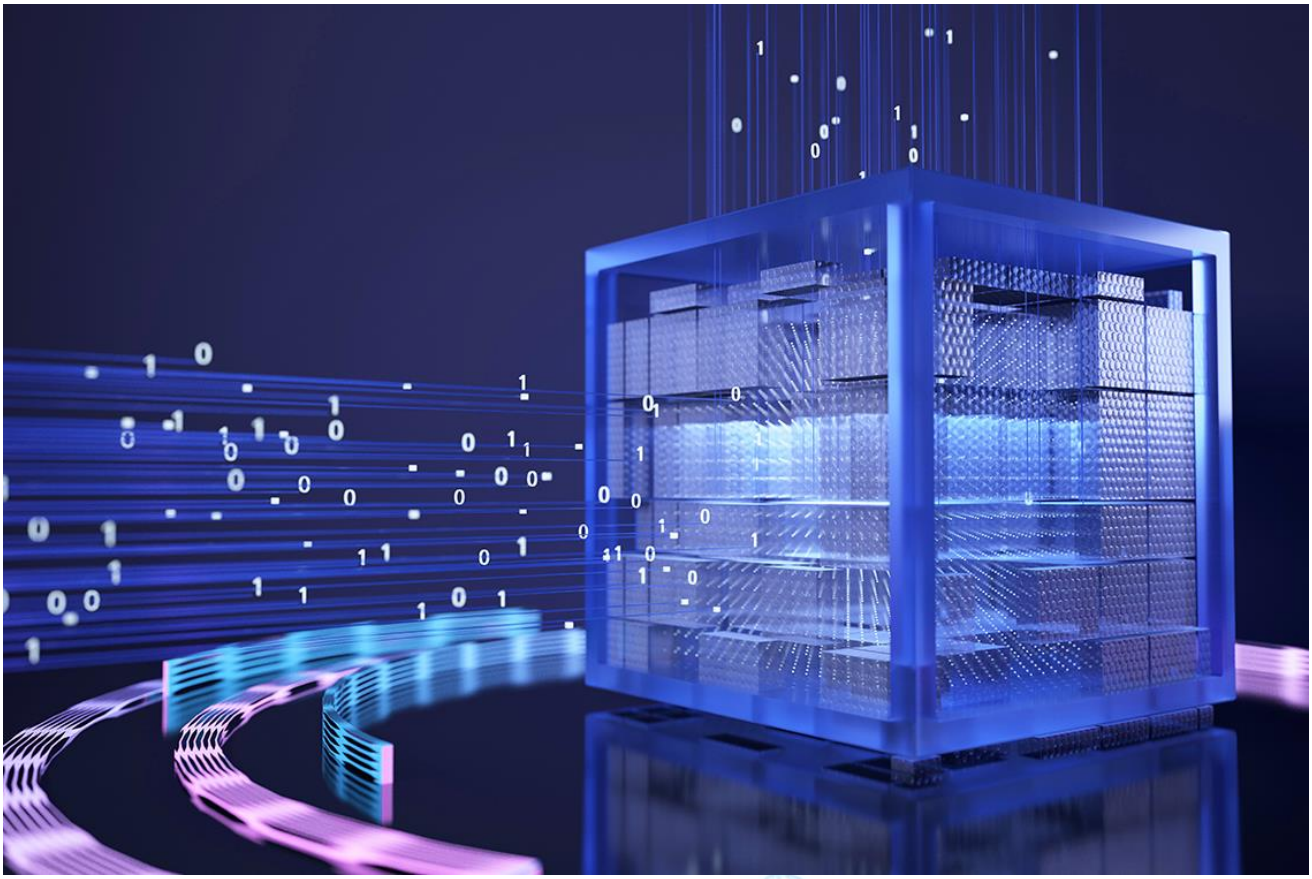
This Guide covers three main topics:

1. **The Need for Data Excellence:** The business and risk management considerations underpinning the need to strengthen data governance and management;

2. **Evolving Areas of Focus:** The key data governance considerations applicable across the DataGRID framework;[3] and

3. **Key Considerations and Suggested Good Practices:** Practical technological, organisational, and governance approaches that may help AIs fulfil key considerations around Data Excellence. The practices are illustrative and include examples from AIs in Hong Kong at varying levels of data maturity. They may support the development of resilient, reliable, and well-governed data foundations but are not intended to prescribe how individual AIs may achieve these outcomes.

The Guide aims to support AIs in making informed decisions when developing, assessing, and/or enhancing their data capabilities. Meaningful progress requires not only investment in technology but also sustained management commitment and effective coordination among business, risk management, technology, and data functions.



[3] HKMA. 2026. Opening Remarks at FiNETech7 on "Data Excellence for Intelligent Risk Management". https://www.hkma.gov.hk/eng/news-and-media/speeches/2026/04/20260402-1/.

02 The Need for Data Excellence

Data Excellence serves as an important foundation for both business operations and effective risk management. As data becomes embedded across decision-making, customer servicing, and regulatory reporting, its quality and surrounding governance increasingly influence how effectively and responsibly AIs operate and respond to emerging opportunities and risks.

This section considers the need for Data Excellence from two complementary perspectives: business and risk management. From a business perspective, well-managed data can support informed decision-making, operational efficiency, and responsible innovation. From a risk management perspective, it can strengthen enterprise-wide risk oversight and maintain operational resilience.

2.1 Business Perspectives

Effective data governance, together with reliable and readily accessible data, can support a range of business outcomes, including the provision of timely and consistent information to management teams, the reduction of manual processing and reconciliation efforts, and a deeper understanding of customers.

AIs may find three areas where these dynamics tend to be visible: (1) strategic and managerial decision-making, (2) operational efficiency and cost management, and (3) the pursuit of new business models, revenue streams, customer experience, and competitiveness.

2.1.1 Strategic and Managerial Decision-making

Reliable, timely and consistently defined data is an important input into sound strategic and management decision-making. Effective data governance helps decision-makers understand the source, meaning, quality, and limitations of the information used, thereby strengthening the dependability and responsiveness of decisions. For example, where business-related performance data is maintained across multiple systems and calculated with inconsistent methodologies, an AI may be unable to develop a timely and/or accurate picture of its current state from an enterprise-level view. These gaps may impact the accuracy of financial projections and undermine senior management's ability to set realistic business expectations.

Disciplined data management and governance can also support a shift from reactive decision-making towards a more forward-looking management approach. AIs that utilise timely data and establish robust data quality controls are better positioned to identify emerging developments and respond promptly. Conversely, fragmented data, reconciliation issues, and poor discoverability can delay strategic decision-making. For instance, where customer-profitability or channel-usage data requires extensive manual consolidation, changes in customer behaviour may be identified late, delaying adjustments to service models and channel strategies.

2.1.2 Operational Efficiency and Cost Management

Beyond decision-making, data management and governance can materially affect operational efficiency and cost. Inconsistent definitions, formats, and unclear ownership may require repeated reconciliation before data can be used. For example, where the same customer or exposure is classified differently across source systems, or where responsibility for resolving data issues is unclear, additional work may be required to produce a consistent record. These activities create recurring operational overhead and, where issues persist across multiple systems, may develop into larger and more costly remediation programmes. Common definitions, clear accountability for data, and embedded data quality controls can reduce rework and facilitate more timely resolution of data issues.

Reporting processes are particularly affected by weaknesses in underlying data. Where reporting inputs are manually assembled from systems using different definitions, AIs may need to perform repeated validation and refinement before reports can be finalised, causing unnecessary delays in reporting cycles. A governed and reusable data foundation can improve comparability across reports and reduce the incremental effort required to respond to new or ad-hoc reporting requirements.

Robust data management and governance are also important enablers of digital transformation. Incomplete or non-machine-readable data can delay system migration and integration, as additional remediation may be required before new platforms and technology tools can operate reliably. Establishing data ownership, quality controls, and lineage early in the transformation lifecycle can reduce subsequent rework and improve delivery predictability.

Taken together, these factors indicate that treating data as a strategic managed asset, rather than an operational by-product, can support more efficient reporting, predictable operating costs, and timely digital transformation. Under-investment in data management and governance may simply defer rather than avoid expenditure.

2.1.3 New Business Models, Revenue Streams, Customer Experience, and Competitiveness

Data capabilities are becoming an increasingly important enabler of service innovation and competitiveness for AIs. The sustainable value of these capabilities, however, depends on customer trust and on whether data is governed and used responsibly. Data governance arrangements should take account of privacy, transparency, consent management, and customer protection from the outset, alongside the pursuit of business opportunities.

High-quality accessible data can support more streamlined and tailored digital journeys. Customer analytics and controlled testing can identify potential improvements in product and service delivery. For example, customer analytics may enable AIs to improve the relevance and usability of digital services by identifying the information, tools or service journeys most useful to customers. Such activities, however, may require clearly defined purposes and data usage boundaries, supported by appropriate customer notices and consent mechanisms where required. Relevant safeguards may include data minimisation, retention controls, and monitoring for unintended customer outcomes.

Improved data governance may also allow AIs to introduce new products and commercial opportunities. For example, alternative data may enhance product attractiveness by enabling AIs to tailor offerings (e.g., credit card deals) according to customer spending habits, potentially opening new customer segments and revenue opportunities. However, its use may also introduce additional risks relating to explainability, fairness, and potential bias. AIs should ensure that data is accurate, appropriately governed, and used in a non-discriminatory manner, with robust protections for privacy and security. Where data is obtained from or shared with third parties, appropriate controls may be required for permitted use cases, access rights, retention periods, and ongoing third-party oversight.

Accordingly, the strategic value of data is not determined solely by the volume of data collected or the number of commercial use cases developed. A sustainable data advantage depends on an AI's ability to demonstrate that data is used responsibly.

2.2 Risk Management Perspectives

Data Excellence supports risk management across two broad areas: (1) risk oversight, exposure management, and regulatory reporting, and (2) data-related technology risk and operational continuity.

Well-managed data, supported by fit-for-purpose data architecture, enables AIs to develop an enterprise-wide view of exposures, identify emerging risks, monitor limits, implement appropriate controls, and respond more effectively to supervisory enquiries. Effective data governance also helps AIs prevent and contain data-related failures, maintain sound operations, and recover from disruptions, thereby strengthening operational continuity.

2.2.1 Risk Oversight, Exposure Management, and Regulatory Reporting

A robust data architecture can support the aggregation of risk data across business lines, legal entities, counterparties, products, jurisdictions, and risk types. This enables management to develop a cross-risk view of the AI's overall risk profile and identify interactions, dependencies, and concentrations that may not be apparent from risk-specific reporting. Where AIs combine information across risk types, they should understand and document the relevant methodologies, assumptions and limitations.

Automated data flows can further strengthen enterprise-wide exposure management by providing timely visibility into emerging risk exposures and supporting the consistent calculation of early-warning indicators. Such indicators may, for example, flag rapid increases in exposures to particular sectors, as well as deteriorating risk factors such as credit quality or liquidity conditions. While automated data flows enable these indicators and thresholds to be calculated and monitored consistently, linking any threshold breaches to clearly defined escalation protocols supports timely and consistent follow-up.

The same data foundation is also fundamental to regulatory reporting, as the reliability of banking returns and risk disclosures depends on the underlying data. Consistent data definitions, traceability, and validation can help reduce reporting errors and repeated manual remediation. Automation, however, does not remove the need for clear ownership, review and accountability.

2.2.2 Data-related Technology Risk Management and Operational Continuity

Beyond keeping data accurate, reliable, and recoverable as it moves through an AI's data architecture, robust data architecture may help mitigate technology risk that could disrupt business operations. As systems and processes become increasingly interconnected, an issue originating in one application (e.g., incorrect data entries or mapping) may propagate to downstream systems and business processes. Fit-for-purpose data governance frameworks reduce the likelihood of such issues arising and constrain their impact when they do, thereby preventing localised data anomalies from escalating into broader disruptions to operational continuity.

Effective risk management also enables AIs to maintain appropriate control of data throughout its lifecycle, strengthening overall operational resilience. When clear prevention, detection, and remediation mechanisms are established, AIs can limit the operational consequences of technology risk and preserve the continuous operations of critical business functions to the greatest extent possible.

03 DataGRID Framework and Evolving Areas of Focus

Recognising the importance of Data Excellence to AIs in business, risk management and regulatory reporting contexts, the HKMA introduced the DataGRID framework under its New Risk Data Strategy during FiNETech7 - *Data Excellence for Intelligent Risk Management*.[4] The framework describes four interconnected attributes of Data Excellence: (1) Granularity, (2) Reliability, (3) Ingenuity, and (4) Discoverability.

For the purposes of this Guide, progress across DataGRID is considered through three evolving areas of focus: (1) governance and accountability, (2) data quality management and control, and (3) data infrastructure. These focus areas are not additional DataGRID attributes or a separate framework. Rather, they provide organisational, procedural, and technological foundations supporting all four attributes.

3.1 The DataGRID Framework

The DataGRID framework provides a reference for considering whether data has the attributes needed to support an AI's business and risk management needs. It considers whether data is sufficiently granular, reliable, and capable of supporting insights from structured and unstructured data,[5,6] and readily discoverable when needed.

3.1.1 "G" – Granularity

Granularity refers to data being available at a level of detail and frequency appropriate to its intended use. Different business, risk management, and regulatory reporting activities may require data at different levels, ranging from aggregated portfolio information to individual transaction-level data.

Granularity should be considered in relation to business, risk management and regulatory reporting requirements that the data supports. The objective is not to maximise the volume or detail of data collected, but to ensure that sufficiently detailed and timely data is available to identify exposures, understand trends, and respond to information needs without disproportionate manual processing.

3.1.2 "R" – Reliability

Reliability refers to data being sufficiently accurate, complete, consistent, and auditable for its intended use. Reliable data should retain these characteristics as it moves between source systems and applications.

Reliability supports confidence that different users and functions are working from consistent information. It is particularly important where data informs material business decisions, risk exposure calculations, and regulatory reporting. Where reliability is uncertain, additional verification and reconciliation may be required before the data can be used.

3.1.3 "I" – Ingenuity

Ingenuity refers to an AI's ability to use data to generate meaningful and forward-looking insights. This may involve drawing on structured and unstructured data to identify trends, anomalies, emerging risks, and/or opportunities that may not be apparent through a conventional analysis approach. The growing availability and variety of data,

[4] HKMA. 2026. Opening Remarks at FiNETech7 on "Data Excellence for Intelligent Risk Management". https://www.hkma.gov.hk/eng/news-and-media/speeches/2026/04/20260402-1/.

[5] "Data that has a standardised format for efficient access by software and humans alike." AWS. Accessed 2026. What is structured data?. https://aws.amazon.com/what-is/structured-data/.

[6] "Information not stored in predefined formats, including text, images, audio, and more." Salesforce. Accessed 2026. Unstructured Data Guide. https://www.salesforce.com/data/what-is-unstructured-data/.

spanning both traditional and alternative sources, means AIs are increasingly well-placed to support these advanced analytical use cases. For example, AIs that have more mature data ecosystems may seamlessly integrate, clean, and standardise financial data alongside alternative sources (e.g., transaction patterns, supply chain data, etc.) to help create a more holistic view of a customer's credit risk profile.

Ingenuity is not determined by the sophistication of technology alone. Its value depends on whether the underlying data is fit-for-purpose and whether analytical outputs can be appropriately validated, interpreted and used responsibly with appropriate data privacy and customer protection in mind. When supported by appropriate governance and reliable data, ingenuity can contribute to better-informed business and more effective risk management decisions.

3.1.4 "D" – Discoverability

Discoverability refers to the ability of authorised users to locate and access relevant data in a timely and consistent manner. Data may be available within an organisation but often remains difficult to use when its sources, ownership, definitions, metadata or access routes are unclear.

Effective discoverability supports timely access to information while maintaining appropriate controls over its use. It can reduce delays in responding to business and risk management needs and enable users to spend less time locating appropriate data.

3.2 Evolving Areas of Focus

In developing its approach to Data Excellence, the HKMA has engaged with AIs through industry engagement sessions and ongoing supervisory dialogue to gain practical insights into how AIs at different stages of development are working towards the four attributes of DataGRID.

Through these engagements, the HKMA has identified three evolving areas of focus that support the implementation of DataGRID: (1) governance and accountability, (2) data quality management and control, and (3) data infrastructure. These focus areas aim to serve as organisational, procedural, and technological foundations through which the four attributes of the DataGRID framework can be established and sustained.

The effectiveness of these arrangements may be reviewed and refined over time through relevant governance, monitoring and assurance processes, taking into account the criticality of the data and the complexity of the AI's operating environment.

3.2.1 Governance and Accountability – Organisational Foundation

Governance and accountability establish how decisions concerning data are made, overseen, implemented and monitored across an AI. This focus area considers whether responsibilities and ownership for material data, including critical data elements ("CDEs") used in material business, risk management, regulatory reporting and control processes, are clearly defined and whether appropriate cross-functional oversight is maintained throughout the data lifecycle. This includes, where relevant, establishing appropriate ownership, permitted-use and oversight arrangements for data obtained from external or third-party sources.

Effective governance arrangements establish accountability for defining data requirements, maintaining data quality, determining appropriate access and use, and addressing identified issues when escalated. They also support cross-functional coordination and oversight, particularly where data is shared across different departments for multiple purposes.

The form of governance may vary according to the nature, scale, and complexity of an AI's operations. The governance arrangements, however, should be sufficiently enduring to support DataGRID as an ongoing enterprise capability, rather than purely existing as project-specific or time-limited initiatives.

3.2.2 Data Quality Management and Control – Procedural Foundation

Data quality management and control translate governance requirements into consistent standards, processes, and control measures applied throughout the data lifecycle. This focus area considers whether an AI can maintain the required quality of its data, identify deficiencies, and ensure that material issues are subject to timely escalation for remediation.

Appropriate data controls, monitoring, and review help preserve the data quality as it moves from source systems through processing environments to downstream uses, enabling data issues to be identified and addressed according to their materiality and potential impact.

Data quality management and control are particularly relevant to Reliability, but they also support the other DataGRID attributes. Appropriate standards help preserve Granularity, determine whether data is suitable for analytical use under Ingenuity, and provide the definitions and contextual information needed for Discoverability.

3.2.3 Data Infrastructure – Technological Foundation

Data infrastructure comprises the architecture, systems, and technological capabilities that support data throughout its lifecycle. This focus area considers whether the infrastructure remains fit-for-purpose and can support the data requirements arising from an AI's business and risk management needs.

Effective data infrastructure enables data to be captured, processed, stored, accessed and disposed of across different systems and applications for various use cases. It also supports appropriate interoperability and controlled access where data is distributed across legacy systems, modern platforms, and third-party environments, as well as sufficient operational resilience to maintain, recover or restore the availability and integrity of material data during significant system changes or disruption.

Given that appropriate data infrastructure will vary across AIs, assessing its adequacy calls for a proportionate approach that considers data criticality, system interdependencies, and each AI's existing processes and operating environment. Effective infrastructure planning also responds to evolving business, regulatory reporting and risk management needs *[see Figure 1]*.

Overall, DataGRID and focus areas provide a common reference for assessing an AI's data capabilities. This integrated perspective can help AIs identify whether limitations arise from unclear accountability, weaknesses in data processes and controls, infrastructure constraints, or a combination of these factors.

Figure 1: DataGRID Framework and Evolving Areas of Focus

EVOLVING AREAS OF FOCUS	DATAGRID FRAMEWORK			
	Granularity	**Reliability**	**Ingenuity**	**Discoverability**
Governance & Accountability *(Organisational Foundation)*	• Establishes accountability for determining the level of detail and frequency required	• Assigns ownership of data quality and provides oversight for escalation and resolution	• Provides oversight over how data may be used and the purposes for which analysis is undertaken	• Establishes accountability for data classification, accessibility, and permitted use cases
Data Quality Management & Control *(Procedural Foundation)*	• Applies standards and controls to preserve the required detail and frequency	• Maintains procedures to preserve accuracy and completeness of data	• Helps users assess the suitability of data for analytical and other approved use cases	• Maintains consistent definitions, classifications, and lineage required to locate data correctly
Data Infrastructure *(Technological Foundation)*	• Provides the capacity to capture, process, and store data at the level of detail and frequency required	• Supports consistent, secure, and auditable processing across systems and applications	• Provides the capacity to integrate and process data and make it available for approved analytical uses	• Provides means for authorised users to identify, locate and access relevant data in a timely and controlled manner

Source: HKMA, Quinlan & Associates analysis

04 Key Considerations and Suggested Good Practices

Drawing on industry engagement and supervisory observations, this section examines key considerations associated with each of the four DataGRID attributes and sets out corresponding observed and emerging practices. Each consideration is mapped to both primary and other relevant evolving areas of focus identified in the previous section.

The HKMA recognises that AIs' data capabilities may be at different stages of development. This section is intended to serve as a practical reference to assist AIs in navigating their respective development journeys. It does not prescribe a single implementation approach, and the suggested practices may be considered proportionately to each AI's own circumstances and the materiality of the data and use cases concerned.

4.1 Granularity

To achieve an appropriate level of granularity for a particular use case, organisations need to define the breadth (i.e., the range of data fields and sources required), depth (i.e., the level of detail at which data points need to be captured), and frequency (i.e., how often data needs to be captured and refreshed).

Through its industry engagements, the HKMA has identified three related areas: (1) aligning granularity requirements with new projects and evolving needs, (2) maintaining visibility of current and future data gaps, and (3) ensuring data infrastructure supports granularity requirements. The following subsections set out key considerations for each area, together with corresponding observed and emerging practices *[see Figure 2].*

4.1.1 Aligning Granularity Requirements with New Projects and Evolving Needs

Key Considerations

Primary Focus Area(s): Governance and Accountability
Other Relevant Focus Area(s): Data Quality Management and Control

New projects and evolving use cases often call for data at a different breadth, depth, or frequency from what existing systems provide. Considering these granularity requirements early, before key architectural decisions are made, allows AIs to consider broader data requirements and downstream dependencies as part of the design from the start, which is typically simpler and less costly.

Granularity requirements also benefit from a range of perspectives across business, risk management, compliance, and technology functions. Bringing these perspectives together helps ensure that technical specifications reflect the intended business and risk management outcomes.

With these perspectives in place, AIs are better positioned to weigh the trade-offs involved. Broader coverage, greater depth, and more frequent ingestion may require additional integration, infrastructure, and operational demands. Increased coverage may increase integration and reconciliation requirements while greater depth may increase storage and processing demands. More frequent ingestion may improve timeliness but require additional investment in infrastructure (e.g., source systems, controls, and operational support). A clear view of these trade-offs helps AIs capture the data each use case needs, without incurring unnecessary complexity or cost.

Figure 2: Granularity Summary

Consideration(s)	Suggested Practice(s)	GA[1]	DQ[2]	DI[3]
Aligning Granularity Requirements with New Projects and Evolving Needs	• Embed data governance requirements into architecture design and implementation from the outset • Prioritise foundational data governance capabilities before pursuing complex digital transformation or innovation initiatives, ensuring underlying data assets and governance mechanisms are sufficiently mature • Establish cross-functional processes involving business, risk, compliance, data and technology teams to define and validate granularity requirements before implementation • Document the intended outcomes that determine the required breadth, depth, and frequency, together with the trade-offs (e.g., timeliness vs. source-system capacity)	★	✓	○
Maintaining Visibility of Current and Future Data Gaps	• Establish a proactive, documented assessment process for regular evaluation and avoid reliance on reactive, ad-hoc responses to regulatory or business developments • Assess the current and future states systematically, including data sources, ownership, location, definitions, formats, refresh frequency, data flows, quality limitations and system dependencies, against requirements arising from business and risk management needs • Identify, prioritise and remediate data gaps by comparing current capabilities with future state requirements, including through improvements to existing data, capture arrangements or supporting infrastructure • Embed governance and monitoring mechanisms by escalating material gaps and remediation roadmaps to senior management and tracking progress to closure, with assessment frequency proportionate to data criticality and environmental change • Conduct event-driven reviews where material changes occur, such as new use cases, regulatory reporting requirements, source-system changes, data migrations or system dependencies	★	★	✓
Ensuring Data Infrastructure Supports Granularity Requirements	• Define data requirements, source constraints, and downstream dependencies to select appropriate data capture, extraction, and ingestion arrangements • Match extraction methods to source types and apply validation controls to ensure data is converted and processed reliably • Assess system load, scalability, latency, completeness, recovery needs, and complexity to select between full and incremental extraction • Align ingestion frequency and processing controls with temporal granularity requirements to identify missing, duplicated, aggregated, or incorrectly processed data • Document and periodically review selected arrangements and their rationale to ensure data granularity and frequency remain fit-for-purpose	○	✓	★

Focus Area (Legend)	
Primary Focus Area(s)	★
Other Relevant Focus Area(s)	✓
Less Relevant Focus Area(s)	○

[1]Governance and Accountability, [2]Data Quality Management and Control, [3]Data Infrastructure
Source: HKMA, Quinlan & Associates analysis

Observed Practices

The HKMA has observed that some AIs have embedded data governance processes for granularity requirements at the outset of the design and implementation of their new projects and material changes to existing use cases, rather than applying them only after data has entered the production environment. Embedding these requirements at the outset enables them to assess data sources and ingestion mechanisms against granularity requirements and identify potential downstream impacts before deployment. Doing so reduces the need for subsequent remediation and helps ensure that technical implementation remains aligned with business and risk management needs as the data environment evolves. Where data and systems are already established, the same discipline may be applied at the point of material change (e.g., the introduction of a new use case, a change in regulatory reporting requirements, a source-system migration, or a material amendment to an existing data structure) *[see Industry Example 1 for reference].*

Industry Example 1

Some AIs embed governance into the introduction of new data sources and changes to existing data structures by considering requirements from both business and technical perspectives before production deployment.

From a business perspective, data owners review and categorise new data to establish relevant quality and metadata requirements before the data is introduced into the platform.

From a technical perspective, these AIs use data contracts[7] to codify agreed requirements for data structures and interfaces. Schema and API changes pass through a continuous integration and continuous deployment ("CI/CD") pipeline to validate proposed changes and identify potential downstream impacts before deployment.

As part of implementing these data governance requirements, the HKMA has also observed that some AIs have established cross-functional processes that require collaboration between relevant teams (e.g., business, risk management, compliance, data, and technology functions) to define and validate granularity requirements before technical specifications are finalised. These requirements were documented at an appropriate data domain or use case level, taking into consideration the intended outcomes and material limitations to inform the required dimensions (i.e., breadth, depth, and frequency) *[see Industry Example 2 for reference].*

Industry Example 2

Some AIs implemented cross-functional working groups comprising business and technical representatives to assess whether data was fit for its intended purpose. This arrangement helped them align business requirements with technical implementation and reduce subsequent rework.

Emerging Practices

When incorporating similar good practices from the industry, AIs may also consider prioritising the establishment of foundational data governance capabilities ahead of further innovation initiatives. Doing so may help mitigate downstream architectural risks. Before pursuing strategic digital transformation objectives, AIs can assess whether their underlying data assets and governance mechanisms are sufficiently mature to support complex new use cases safely.

In addition to these processes, AIs may also document key trade-offs (e.g., balancing data timeliness against source-system capacity, data granularity against storage and processing demands, and implementation sophistication against operational complexity and resilience).

4.1.2 Maintaining Visibility of Current and Future Data Gaps

Key Considerations

Primary Focus Area(s): Governance and Accountability, Data Quality Management and Control
Other Relevant Focus Area(s): Data Infrastructure

Beyond individual projects, granularity also depends on how data is governed day to day. Clear governance and accountability for everyday data operations help AIs maintain an up-to-date view of the data they hold and how it is managed (i.e., metadata and data lineage oversight). This gives governance bodies and data users a common basis for judging whether existing data remains sufficient for business and risk management needs.

[7] "Formal agreements between data producers and data consumers that define the quality, structure, semantics and availability of data". IBM. Accessed 2026. What is a data contract?. https://www.ibm.com/think/topics/data-contract.

This view becomes more valuable as needs evolve. When potential gaps (e.g., insufficient granularity or incompatible refresh frequencies) are identified early, AIs can address them in a planned and orderly way, with less disruption to business outcomes.

Observed Practices

The HKMA has observed that some AIs have implemented regular assessment and improvement plans for data granularity to sustain foundational governance over time. These plans are driven by cohesive internal strategies rather than operating as purely reactive, ad-hoc responses to external developments or regulatory pressures. By transitioning from a reactive, project-based methodology towards a proactive, continuous strategy, AIs are better positioned to maintain resilient and scalable data architectures. For instance, a systematic comparison of an AI's current and future states can provide a consistent basis for identifying and prioritising gaps in data granularity *[see Industry Example 3 for reference].*

Industry Example 3

Some AIs established a board-approved data governance framework governing data quality and data security, overseen by a corporate governance committee composed of division heads and board members to ensure executive buy-in and a sustainable rollout.

As part of an enterprise-wide assessment process, a data quality task force monitors data quality on a monthly basis for CDEs identified per key domain. These are reviewed with data owners and stakeholders as a continuous, enterprise-wide function, rather than on a project basis. Furthermore, a second committee, chaired by the Group Chief Operating Officer and comprising senior executives, systematically reviews datasets and sets thresholds, such as tracking and flagging data that is not updated within a set period.

When gaps are identified, data falling below thresholds triggers a documented root-cause investigation and remediation process. This ensures that these AIs continuously monitor critical domains, such as compliance, tax reporting, and customer demographic data.

Emerging Practices

The current state assessment may begin by establishing a clear view of the data already available to the AI and the conditions under which it is managed under established data governance standards. This may involve identifying relevant data sources and designated data owners, where the data resides, how it is defined and formatted, how frequently it is refreshed, and how it is captured and transferred between systems. AIs may also assess known data quality limitations and material dependencies across source systems and downstream processes. These steps provide a baseline for data owners and governance bodies to determine whether existing data and infrastructure can support the AI's data granularity needs.

The future state assessment may then translate intended business and risk management needs into corresponding requirements for data granularity and supporting architecture. Comparing these requirements with the current state can help AIs identify specific gaps and determine how they can be addressed (e.g., through improvements to existing data, changes to capture arrangements, or enhancements to infrastructure).

Identified issues and remediation roadmaps may be formally escalated to senior management and monitored through established governance arrangements and reporting lines until closure. The frequency and depth of each assessment may be calibrated to the criticality of the data and the rate of change in the relevant environment. Additional reviews may also be triggered by ad-hoc developments (e.g., new business cases, changes in regulatory reporting requirements, modifications to source systems, data migrations, and/or material changes in system dependencies) *[see Figure 3].*

4.1.3 Ensuring Data Infrastructure Supports Granularity Requirements

Key Considerations

Primary Focus Area(s): Data Infrastructure
Other Relevant Focus Area(s): Data Quality Management and Control

Another consideration is whether infrastructure can capture and preserve data at the required level of detail, format, and frequency. Many AIs operate a combination of modern and legacy systems, and relevant information may be held in a variety of forms, from structured databases to scanned documents and physical records.

Figure 3: Data Gap Assessment Process

	1	2	3	4
Steps	SYSTEMATIC STATUS QUO DIAGNOSIS	FUTURE STATE SETTING	GAP IDENTIFICATION	ESCALATE AND REMEDIATE
Description	Systematically identify data sources and owners, where data resides, how it is defined, formatted, refreshed and transferred, and known limitations	Translate intended business and risk management needs into corresponding requirements for data granularity and the supporting architecture	Compare these requirements with the current state to identify specific gaps and determine how they can be addressed (data, capture or infrastructure)	Formally escalate issues and remediation roadmaps to senior management; implement corrective actions and monitor progress until full closure
Objective	***Establish a clear view of baseline granularity capabilities across the organisation***	***Define granularity requirements driven by business priorities and risk management needs***	***Identify granularity gaps between baseline capabilities and organisational strategic requirements***	***Escalate identified gaps and address them through structured remediation processes***

Source: HKMA, Quinlan & Associates analysis

How data is ingested also shapes the granularity available downstream. Scheduled batch processing, for example, captures the state of data at set points in time, which suits many use cases, while others may call for transaction- or event-level detail.

When capture and ingestion arrangements reflect both the use case and the characteristics of each source, AIs are able to make full use of the information they already hold.

Observed Practices

The HKMA has observed that some AIs have established a structured process for selecting data capture, extraction, and ingestion arrangements according to the granularity requirements of each material use case. This may include defining the required data fields, level of detail, source format, capture frequency, and retention period, together with relevant source-system constraints and downstream dependencies *[see Industry Example 4 for reference]*.

Industry Example 4

Some AIs determined their ingestion frequency for data elements by balancing cost, business needs, and latency requirements.

These AIs utilise streaming for time-sensitive applications requiring real-time analytics, such as transaction fraud detection, instant AML screening, credit limit assessment on customer loan applications, and Faster Payment System ("FPS") payment events.

For those applications that do not require urgent or real-time data visibility, such as financial and regulatory reporting, these AIs apply batch or daily ingestion, where a daily cadence is sufficient.

Emerging Practices

When applying these structured processes, the appropriate capture or extraction method will depend on the nature of the source, and AIs may need to consider certain implementation nuances when applying different techniques, especially their effect on data completeness and timeliness. For relational databases, AIs may determine whether full or incremental extraction is appropriate, including whether Change Data Capture ("CDC") is supported and suitable for the source system and intended use. For external data, AIs may evaluate the interface options used to obtain the data (e.g., APIs[8] vs. managed file transfers), considering factors such as reliability, security, and consistency. Where information is

[8] "A set of rules and protocols for building and interacting with software applications." SAP. 2026. What is an API (application programming interface)?. https://www.sap.com/hk/resources/what-is-api.

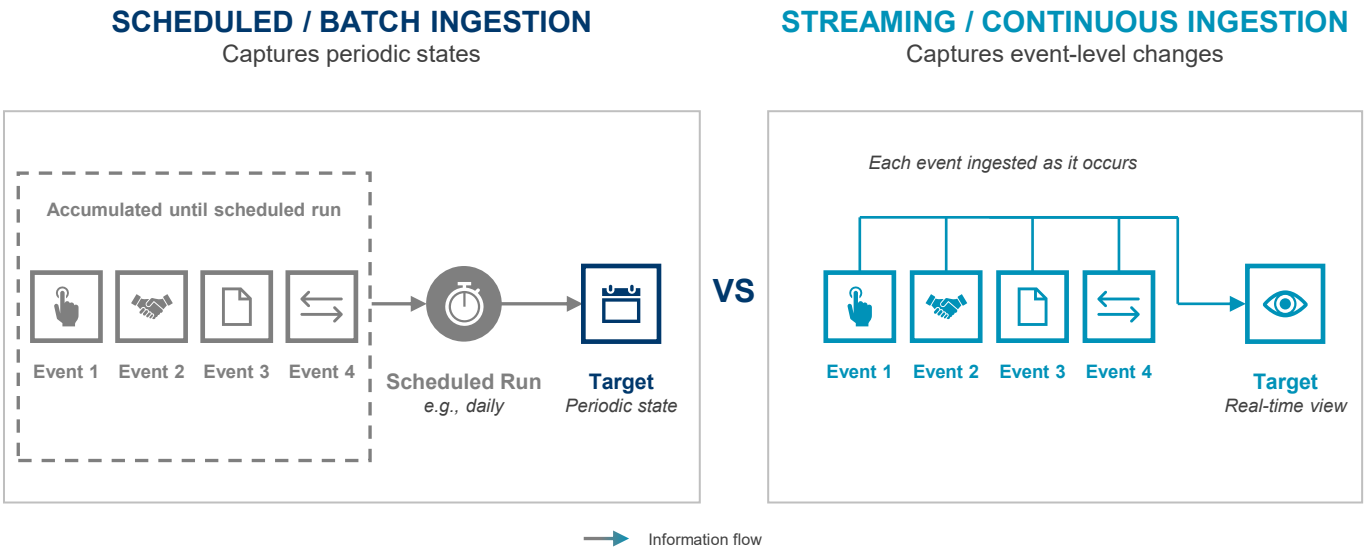
retained in scanned documents or other unstructured formats, AIs may consider the accuracy and limitations of tools (e.g., Optical Character Recognition, etc.) when converting information into a processable form. Appropriate validation can help mitigate risks where extraction errors could affect material downstream uses *[see Section 6.1 in the Technical Appendix for more details].*

When selecting between full and incremental extraction, AIs may consider factors such as the processing load imposed on source systems, data volumes, scalability, latency, completeness, recovery requirements, and implementation complexity. The selected approach can also enable the AI to identify and address incomplete, duplicated, or incorrectly processed data.

Ingestion frequency can similarly reflect the temporal granularity required by the intended use case. Scheduled processing may be appropriate where daily or periodic observations are sufficient. More frequent or continuous ingestion may be needed where time-sensitive data is required. The selected arrangements may include controls to identify omitted records, unintended aggregation, duplicated data, or other processing outcomes that could reduce the required granularity *[see Figure 4].*

AIs may document and periodically review the assessment criteria, selected arrangements, and underlying rationale. This can help AIs determine whether their infrastructure continues to capture and preserve data at the level of detail and frequency required.

Figure 4: Batch vs. Streaming Ingestion



Source: Databricks, HKMA, Quinlan & Associates analysis

4.2 Reliability

Beyond capturing data at the required level of granularity, reliability depends on whether material data remains accurate, complete, and consistently defined, with sufficient traceability from source to output. Through its industry engagements, the HKMA has identified three related areas: (1) establishing governance over consistent data standards and data lineage, (2) strengthening data quality management and controls, and (3) maintaining connected data infrastructure and appropriate data retention. The following subsections set out key considerations for each area, together with corresponding observed and emerging practices *[see Figure 5]*.

Figure 5: Reliability Summary

Consideration(s)	Suggested Practice(s)	GA[1]	DQ[2]	DI[3]
Establishing Governance Over Consistent Data Standards and Data Lineage	• Establish enterprise-level data governance with an appropriate model, senior management participation, cross-functional oversight, and clear roles and responsibilities for data ownership and stewardship • Maintain common definitions, policies, and procedures to apply data standards and manage quality issues consistently across the organisation • Consider adopting automation or A.I.-based tools to reduce manual effort, promote consistent application of standards, and enable earlier identification of data quality issues • Maintain accessible and standardised lineage documentation to explain material data flows, transformations, dependencies, and control points to both technical and business staff • Consider leveraging large language model ("LLM")-based tools integrated with data lineage repositories to let staff query lineage in plain language, with sufficient accuracy checks to validate outputs against authoritative lineage metadata	★	✓	★
Strengthening Data Quality Management and Controls	• Apply proportionate preventive controls at data capture and detective controls at interfaces and before material outputs to prevent or identify errors early • Monitor data volumes, distributions, control results, and other quality indicators to identify unexpected changes and emerging deficiencies • Define quality thresholds, issue-severity criteria, ownership, documentation, escalation, and remediation procedures to manage issues through closure • Designate authoritative data sources and use appropriate master data management capabilities to support consistency, traceability, and aggregate views	✓	★	✓

Figure 5: Reliability Summary (Cont.)

Consideration(s)	Suggested Practice(s)	GA[1]	DQ[2]	DI[3]
Maintaining Connected Data Infrastructure and Appropriate Data Retention	• Establish common definitions, identifiers, matching rules, and authoritative sources for material data to support consistent interpretation across systems • Implement controlled synchronisation mechanisms aligned with business and risk management needs and existing architecture to propagate changes and reduce manual reconciliation • Maintain infrastructure capable of reconstructing the historical data state, processing steps, and specific version used for a material output • Retain historical data, transformation logic, and execution records for periods commensurate with relevant business and risk management requirements • Preserve persistent links between data and processing logic, including model versions and inputs where applicable, to support traceability of material outputs	○	✓	★

Focus Area (Legend)	
Primary Focus Area(s)	★
Other Relevant Focus Area(s)	✓
Less Relevant Focus Area(s)	○

[1]Governance and Accountability, [2]Data Quality Management and Control, [3]Data Infrastructure

Source: HKMA, Quinlan & Associates analysis

4.2.1 Establishing Governance over Consistent Data Standards and Data Lineage

Key Considerations

Primary Focus Area(s): Governance and Accountability, Data Infrastructure

Other Relevant Focus Area(s): Data Quality Management and Control

Reliability starts with consistency. When data of a similar nature is captured using common standards and procedures across the organisation, it is easier to establish a unified view of data.

Data lineage builds on these standards. When traceability is recorded consistently, data can be followed from source to output and understood across organisational silos.

Both of these aspects benefit from enduring governance. Arrangements that operate continuously at the organisational level, supplemented by those that operate throughout the duration of individual projects, may help sustain data standards and data lineage as systems, teams, and business needs change.

Observed Practices

Through its industry engagements, the HKMA has observed that the continuity of data governance arrangements varies across AIs. Some AIs establish governance committees on a project basis, which are set up when a project starts and conclude when it closes. Other AIs set up arrangements that operate continuously at the organisational level. In both cases, AIs may aim to apply data quality and data lineage arrangements consistently across business teams over time.

To establish sustainable data governance, AIs may find it useful to (1) determine an appropriate enterprise-wide governance model for data standards, (2) establish clear roles and outline responsibilities for data ownership and stewardship, and (3) define policies and procedures to ensure data quality.

Similar to addressing data granularity considerations, AIs may need to determine an appropriate enterprise-wide data governance model to ensure data reliability. For example, AIs may need to balance the need for responsiveness, scalability, and cross-functional

communication when selecting between a centralised, decentralised, or federated model to determine where decisions are made and executed *[see Section 6.2 in the Technical Appendix for more details]*.

As part of implementing the data governance model, AIs should also establish suitable roles and responsibilities (e.g., dedicated data stewards or owners) to ensure accountability for data capture, labelling (i.e., data standards), and traceability (i.e., data lineage) for both internal and external data. In addition to data owners, where possible, AIs may consider including senior management in oversight committees to position data governance as a central aspect of business units, while ensuring consistent operational oversight across the organisation *[see Industry Example 5 for reference]*.

Industry Example 5

Some AIs established clear roles and responsibilities supported by comprehensive training programmes to ensure staff competency in fulfilling their duties. These frameworks were designed to maintain segregation of duties, thereby mitigating conflicts of interest and strengthening internal controls.

AIs are also encouraged to utilise internal audit and review functions to evaluate the effectiveness of existing data lineage controls periodically. In particular, AIs may consider formally incorporating regulatory reporting processes and their associated controls into internal audit plans so that controls across the end-to-end data lifecycle are subject to periodic independent review.

To further support implementation, AIs should ensure that policies and procedures covering both data quality and data lineage are sufficiently detailed to allow staff to carry them out. For instance, AIs may consider documenting lineage records in a manner that is accessible and interpretable to business functions, rather than solely to technical staff. This documentation may include maintaining plain-language descriptions of key data transformations alongside technical logs and standardising how lineage documentation is presented across systems.

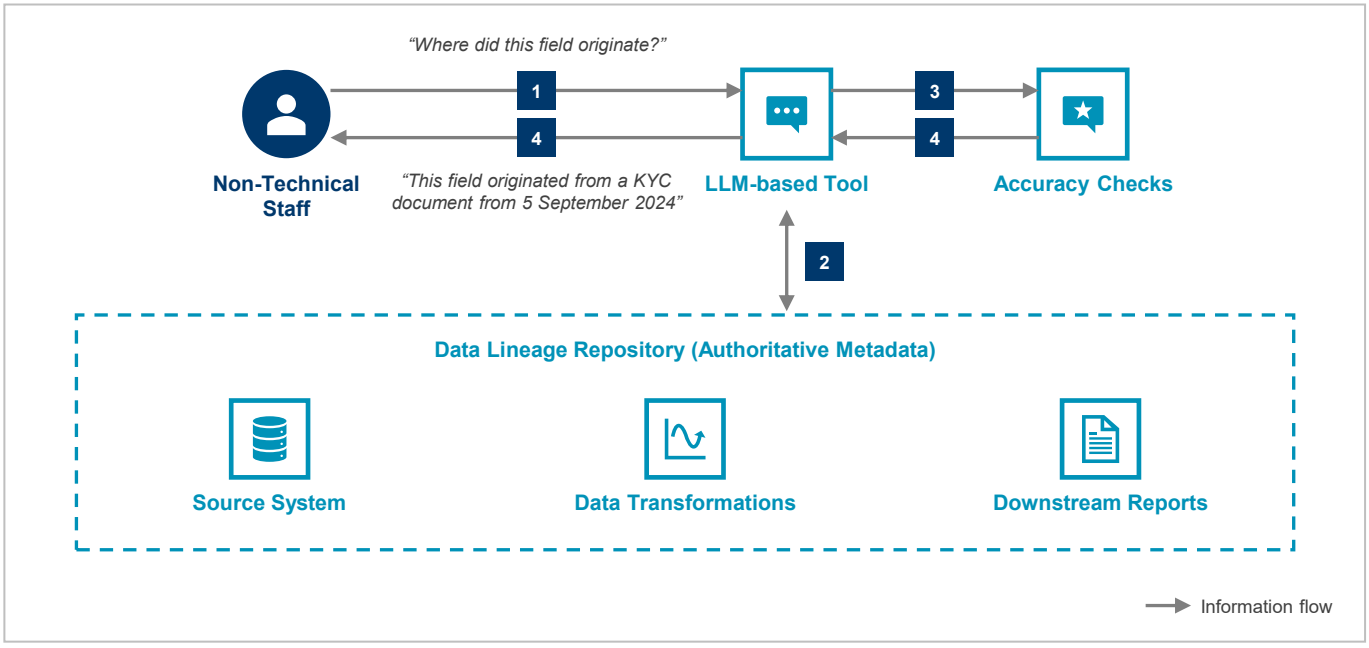
Emerging Practices

Beyond typical data governance arrangements, AIs may identify CDEs and establish data quality Operational-level Agreements ("OLAs") at the governance level. Business units, data owners, and technology staff may collectively define requirements that reflect both business needs and practical implementation considerations. Relevant governance bodies may periodically review whether data quality OLAs remain fit-for-purpose and whether CDEs have approved data quality OLAs and clearly assigned data owners. These reviews may be complemented by automated data quality monitoring tools that continuously assess data against defined OLA thresholds and generate real-time alerts when quality drifts outside acceptable parameters. Doing so allows emerging data quality issues to be identified and addressed as they arise, rather than only being detected at scheduled committee reviews *[see Section 6.3 in the Technical Appendix for more details]*.

To further enhance implementation, AIs may consider adopting automation or A.I.-based tools to reduce the manual effort spent by data owners to identify and comply with governance requirements. For instance, an A.I.-enabled policy-lookup tool may allow a data owner to query, in natural language, what standards (e.g., protections or classifications) should be applied to a given dataset, with responses generated from a curated repository of approved policies. Where implemented well, such tools may reduce inconsistent application of standards across business units and shorten the time needed to onboard new datasets into the governance framework.

For data lineage, AIs may consider leveraging LLM-based tools integrated with data lineage repositories to allow staff to query lineage using plain language, such as asking where a particular data field originated from or which downstream reports consume it. Implementing this may make lineage information more accessible to non-technical users and support diagnosis (e.g., identification of data dependencies), without requiring staff to manually navigate complex lineage graphs or interpret system-specific notation. As LLM outputs may be prone to inaccuracy, AIs are encouraged to ensure such tools retrieve answers directly from authoritative lineage metadata and that sufficient accuracy checks are in place to validate the tool's outputs. Such adoption ensures that reviewers moving between teams or platforms are not required to interpret inconsistent formats *[see Figure 6]*.

Figure 6: Plain-language Access to Data Lineage



1	**ASK IN PLAIN LANGUAGE** Staff query lineage in plain language, such as asking where a particular data field originated from or which downstream reports consume it
2	**RETRIEVE FROM AUTHORITATIVE METADATA** The LLM-based tool is integrated with the data lineage repository and retrieves answers directly from authoritative lineage metadata
3	**VALIDATE THE OUTPUT** As LLM outputs may be prone to inaccuracy, sufficient accuracy checks are in place to validate the tool's outputs before they are relied upon
4	**RETURN AN ACCESSIBLE ANSWER** Non-technical users gain lineage insight and support for diagnosis without manually navigating complex lineage graphs or system-specific notation

Source: Snowflake, HKMA, Quinlan & Associates analysis

4.2.2 Strengthening Data Quality Management and Controls

Key Considerations

Primary Focus Area(s): Data Quality Management and Control

Other Relevant Focus Area(s): Governance and Accountability, Data Infrastructure

Data quality is shaped at each stage of the data lifecycle, from the point of capture through every transfer between systems. The earlier an issue is identified, the more contained its impact and the simpler it is to resolve.

Timely identification also supports confidence in management information. When data quality issues are surfaced and escalated proactively, senior management and other data users can rely on data that reflects the underlying business with greater assurance.

Observed Practices

Through its industry engagements, the HKMA has observed that controls may be more effective when applied at relevant stages of the data lifecycle rather than at a single checkpoint *[see Industry Example 6 for reference]*.

Industry Example 6

To mitigate data quality issues arising from data capture, some AIs address data quality at the point of entry through automated validation, preventing low-quality data from entering the system. This approach mitigates the need for fragmented workarounds or ad-hoc exceptions to accommodate data anomalies.

Complementing their data source validation controls, these AIs also implement controls to monitor cross-application data movement. These controls run automatically when data moves from one application to another, preventing upstream errors from propagating downstream.

In addition to controls and monitoring, the HKMA has observed AIs using supporting data management practices to improve consistency and traceability across systems. Master Data Management ("MDM")[9] approaches, for example, may help establish a consistent reference for data elements by defining a single, authoritative "golden record" (e.g., internal client codes, Legal Entity Identifiers ("LEIs")). This may place AIs in a better position to view aggregate exposure (e.g., to a given counterparty, sector, or macroeconomic event), rather than relying on ad-hoc reconciliation across separate technical silos *[see Industry Example 7 for reference]*.

Industry Example 7

Some AIs achieve standardisation via a sanitised golden customer domain schema (i.e., a golden record for all customer data). All external system feeds are transformed back to the standard schema to preserve data lineage and enable common references across applications. This design allows a single view of customer records and facilitates changes to propagate across applications.

Emerging Practices

AIs may consider applying proportionate preventive and detective controls at relevant points throughout the data lifecycle, supported by monitoring and consistent issue-management processes.

Preventive controls at capture and detective controls at system interfaces or before material outputs can help identify deficiencies before they propagate to reporting, risk management, or business decisions.

Ongoing monitoring can complement these controls by identifying unexpected changes in data volumes, distributions, control results or other indicators of potential quality deficiencies. Automated monitoring may support earlier detection, while human judgement and review remain important for exceptions that require contextual assessment.

For example, enhanced monitoring (e.g., via deterministic vs. statistical data controls) may streamline this process by alerting relevant data owners and governance bodies when data quality OLAs are breached, where CDEs do not have approved data quality OLAs or assigned data owners, or where data quality issues not related to CDEs do not have assigned OLAs *[see Section 6.3 in the Technical Appendix for more details]*.

Taken together, these arrangements can support earlier identification, consistent escalation and timely remediation of data quality issues. Their scope, frequency and degree of automation may be calibrated to the criticality of the data and the consequences of control failure, and reviewed as the data environment evolves.

[9] "A foundation for creating a unified, accurate and comprehensive view of business information." Snowflake. Accessed 2026. What Is Master Data Management?. https://www.snowflake.com/en/data-governance/master-data-management/.

4.2.3 Maintaining Connected Data Infrastructure and Appropriate Data Retention

Key Considerations

Primary Focus Area(s): Data Infrastructure
Other Relevant Focus Area(s): Data Quality Management and Control

Reliability also depends on how well systems work together. When an update in one system (e.g., to a client profile or exposure metric) flows consistently to others, data remains aligned across the organisation, and teams can rely on it without manual reconciliation.

Data retention plays a related role. Keeping a record of how data has changed over time, rather than only its latest state, supports auditability and allows data lineage to be maintained across processing steps.

Observed Practices

The HKMA has observed that some AIs have established common definitions, identifiers and matching rules for material data, together with clear designation of authoritative sources by data domain.

Common rules for matters such as date formats, identifier schemes, and counterparty naming can support consistent interpretation across systems and entities. Controlled interfaces and synchronisation processes may be used to propagate relevant changes across systems *[see Industry Example 8 for reference]*.

Industry Example 8

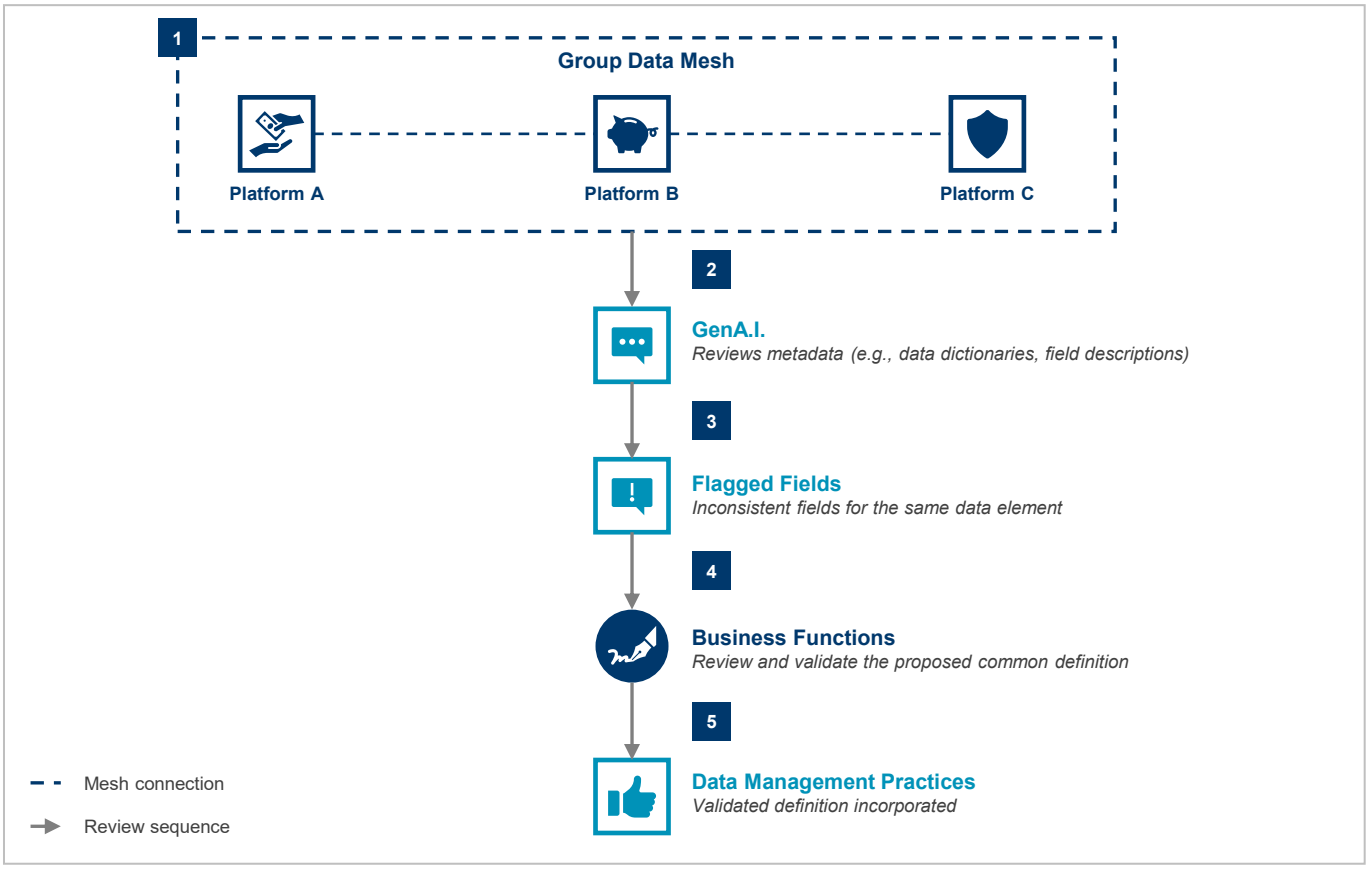
Some AIs highlighted the execution complexity of data governance frameworks. AIs with a local footprint may find governance execution relatively straightforward, while multinational, multi-platform institutions face greater difficulties. With more environments, unifying data definitions and reconciling across systems become increasingly complex.

To address this, some multinational AIs adopted a data mesh at the group level, connecting data platforms rather than duplicating data into a single repository.

They also leveraged generative A.I. ("GenA.I.") as a key enabler, for example, by reading metadata across applications to harmonise definitions.

This, for example, involved GenA.I. reviewing metadata (e.g., data dictionaries, field descriptions, etc.) from different applications and identifying inconsistent fields for the same underlying data element. When discrepancies were identified, the GenA.I. would flag the instance and propose a common definition for review by relevant business functions. The resulting suggestions were then validated by the relevant functions before being incorporated into the AI's data management practices *[see Figure 7]*.

Figure 7: Harmonising Definitions with GenA.I.



1	**CONNECT PLATFORMS** A data mesh at the group level connects data platforms across environments, rather than duplicating data into a single repository
2	**REVIEW METADATA** GenA.I. reviews metadata (e.g., data dictionaries, field descriptions) from different applications across the connected data platforms
3	**FLAG INCONSISTENCIES** GenA.I. identifies inconsistent fields for the same underlying data element and flags each instance where a discrepancy is identified
4	**PROPOSE A COMMON DEFINITION** For each flagged instance, GenA.I. proposes a common definition for review by the relevant business functions
5	**VALIDATE AND INCORPORATE** The relevant functions validate the resulting suggestions before they are incorporated into the AI's data management practices

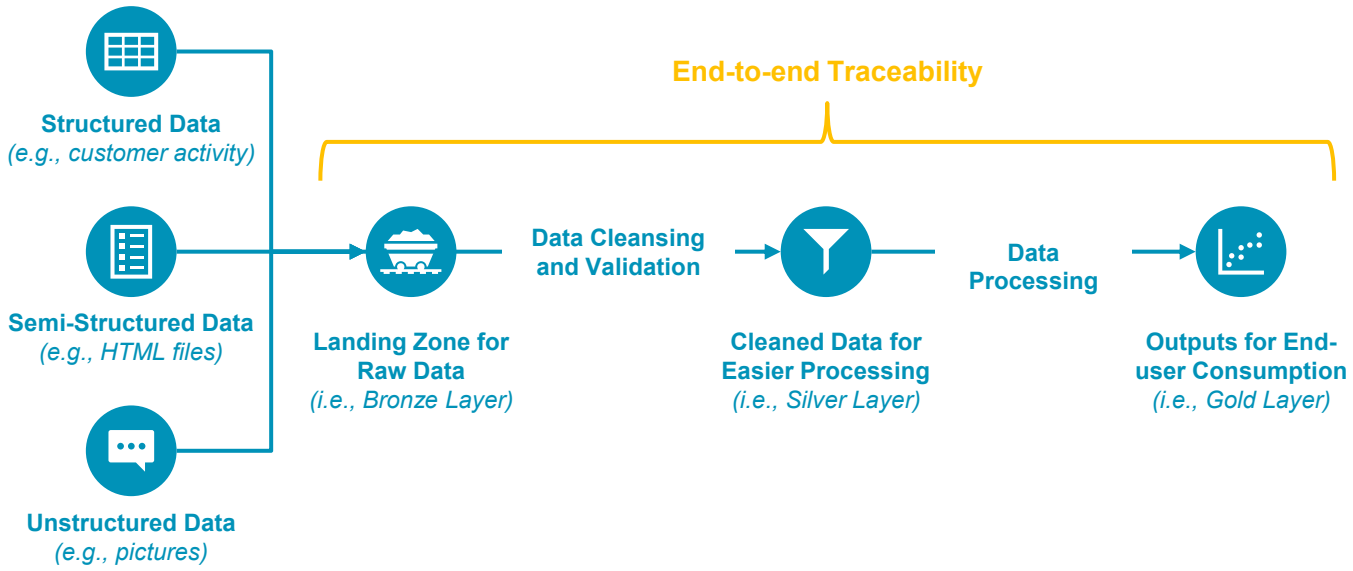
Source: HKMA industry engagement, Quinlan & Associates analysis

AIs may establish clear responsibilities between data owners for maintaining and updating data, as well as appropriate arrangements for reflecting such changes in other systems. For example, AIs may designate one system and its data owner as the authoritative source for a particular data domain, even where relevant updates are subsequently reflected across other systems maintained by other data owners. The appropriate mechanism may be batch-based, event-driven or otherwise aligned with the relevant requirements and architecture *[see Section 6.1 in the Technical Appendix for more details]*.

Emerging Practices

To facilitate auditability across data transformations, AIs may consider establishing data infrastructure capable of reconstructing, for any given point in time: (1) the historical state of a specific data point, (2) the processing steps applied to it, and (3) where applicable, the specific version utilised to produce a corresponding report, model output, or management decision.

Figure 8: Medallion Architecture



Source: Databricks, Snowflake, Google Cloud, Quinlan & Associates analysis

To develop this capability, AIs may consider adopting storage architectures (e.g., data warehouses, data lakes, data lakehouses, data mesh / fabric, etc.) designed to preserve historical snapshots of records beyond the latest version. Maintaining the retrievability of historical data reduces the reliance on manual reconstruction from downstream sources and ensures data can be accessed directly as circumstances require *[see Section 6.4 in the Technical Appendix for more details]*.

AIs may consider retaining historical versions of material data and the associated transformation logic with execution records for periods commensurate with pertinent business and risk management requirements. For example, some AIs implement this through a layered data architecture (e.g., medallion architecture) that separates raw, cleansed, and curated data into distinct layers so that records are not overwritten at intermediate stages *[see Figure 8]*. Combining this with an immutable raw layer, versioned tables, and layer-specific retention policies allows earlier states of the data and the logic applied to them to be retrieved on demand *[see Section 6.5 in the Technical Appendix for more details]*.

Beyond preserving historical records, AIs may also consider implementing mechanisms to track data transformations on an ongoing basis. This may involve adopting infrastructure that maintains a persistent linkage between datasets and their corresponding calculation rules, mapping tables, or categorisation logic. Such arrangements help reduce the risk of data and logic drifting out of sync during periodic updates.

Where predictive models or algorithmic routing are deployed, AIs may consider logging model versions, parameters, and the specific input data utilised during each execution. This practice assists in tracing outputs from non-deterministic processes back to their originating conditions, even if the outputs cannot be exactly reproduced.

4.3 Ingenuity

Organisations may hold significant volumes of proprietary data. Their ability to convert such data into meaningful insights depends on governance, data quality management, and infrastructure working together.

Through its industry engagements, the HKMA has identified three related areas: (1) building flexible upstream data pipelines, (2) establishing governance over model selection, and (3) strengthening business cases for model development. The following subsections set out key considerations for each area, together with corresponding observed and emerging practices *[see Figure 9]*.

Figure 9: Ingenuity Summary

Consideration(s)	Suggested Practice(s)	GA[1]	DQ[2]	DI[3]
Building Flexible Upstream Data Pipelines	• Review data processing arrangements periodically and following material changes to confirm continued suitability for evolving requirements • Maintain scalable processing capacity and workload distribution arrangements to reduce bottlenecks as data volumes and complexity increase • Reuse common processing components and business logic across use cases to reduce duplication, simplify maintenance, and promote consistent outcomes • Adopt modular arrangements with defined interfaces and dependencies to enable individual components to be maintained, tested, or replaced with limited disruption	○	✓	★
Establishing Governance over Model Selection	• Define the business problem and intended outcome to establish clear criteria for assessing candidate models • Evaluate candidate models against business-relevant performance, risk, explainability, and operational criteria to select a proportionate approach • Conduct independent validation before deployment to confirm that the selected model performs within approved parameters • Monitor data drift, prediction stability, and business outcomes to trigger timely re-evaluation, retraining, or withdrawal throughout the model lifecycle	★	○	✓
Strengthening Business Cases for Model Development	• Develop reusable enterprise-wide model development and validation capabilities to reduce the cost and complexity of isolated proofs of concept • Implement shared analytics infrastructure, including feature stores and model registries where appropriate, to support reuse, versioning, lineage, and approval • Document build, buy, integration, centralisation, and governance choices to support proportionate investment and model-lifecycle oversight	○	✓	★

Focus Area (Legend)	
Primary Focus Area(s)	★
Other Relevant Focus Area(s)	✓
Less Relevant Focus Area(s)	○

[1]Governance and Accountability, [2]Data Quality Management and Control, [3]Data Infrastructure

Source: HKMA, Quinlan & Associates analysis

4.3.1 Building Flexible Upstream Data Pipelines

Key Considerations

Primary Focus Area(s): Data Infrastructure
Other Relevant Focus Area(s): Data Quality Management and Control

Converting data into insight depends on data pipelines that can keep pace with evolving analytical requirements. Pipelines built around predefined fields and outputs work well where reporting requirements are stable and analytical outputs are known in advance.

As AIs integrate new data sources and explore more detailed analysis, the ability to adapt pipelines without major redevelopment becomes increasingly valuable, allowing new use cases to be supported more quickly and at lower cost.

Observed Practices

The HKMA has observed that some AIs review their data processing arrangements periodically and following material changes and identify whether processing capacity can be scaled, common components can be reused, and individual components can be adapted without necessarily disrupting the wider pipeline.

Data processing arrangements that are suitable for current workloads may become less effective as data volumes, sources, use cases, and system dependencies evolve. As such, AIs may assess whether their processing arrangements can accommodate changes in workload, business scale, and the diversity of data sources. Where processing capacity cannot be adjusted, or workloads re-distributed, increases in data volume or complexity may create unnecessary bottlenecks. Arrangements that scale in step with demand, for example during month-end reporting peaks or the onboarding of a new data source, allow AIs to respond to evolving needs without degrading the timeliness or completeness of their data *[see Industry Example 9 for reference]*.

Industry Example 9

Some AIs implemented data systems architected on a modular basis, segregated by business product or line of business, to enhance operational flexibility and resilience. This design minimises unnecessary dependencies, enabling scalable expansion while reducing the potential impact on interconnected systems.

Emerging Practices

AIs may also consider whether common processing components and business logic can be reused across different use cases rather than developed separately for each application. For example, an AI could utilise the same data validation or cleansing logic across multiple data pipelines (e.g., through a capability registry), rather than embedding the same rules separately within each application. Reuse may reduce duplication and facilitate the consistent and efficient implementation of changes to processing rules *[see Figure 10]*.

Conversely, where similar logic is embedded separately within individual applications, subsequent changes may require multiple implementations, increasing maintenance effort and the risk of inconsistent outcomes *[see Section 6.6 in the Technical Appendix for more details]*.

The adaptability of the wider data pipeline will also depend on the extent to which individual components can be maintained, tested, or replaced independently. For example, AIs could adopt a modular data pipeline (e.g., utilising an ELT approach vs. the traditional ETL approach), which separates key functions (e.g., ingestion, data quality checks, transformation, enrichment, etc.) into distinct components with defined interfaces, allowing an individual component to be updated or replaced without redesigning the entire pipeline. This could also allow new data sources or processing steps to be added without materially changing existing components.

Where components are tightly coupled, changes to one part of the pipeline may require broader testing or redevelopment and may have unintended effects on dependent processes. More modular arrangements, supported by clearly defined interfaces and dependencies, may allow changes to be implemented in a more controlled manner and with less disruption. Without such capabilities, AIs may find it becomes harder to extend or reuse existing components as data needs grow *[see Sections 6.6 and 6.7 in the Technical Appendix for more details]*.

The appropriate degree of flexibility, reuse, and adaptability will vary according to an AI's risk profile, operating model, existing technology environment, and the anticipated scale and complexity of its data processing needs.

Figure 10: Capability Registry

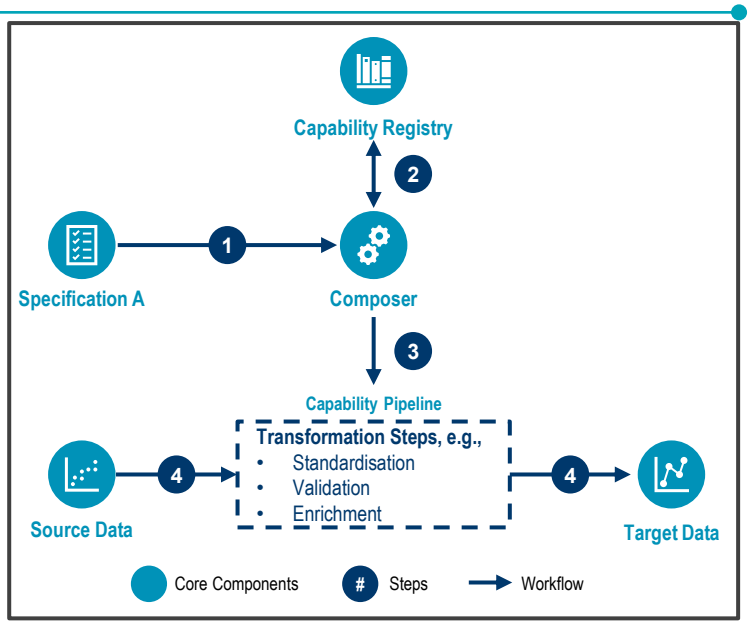
Workflow

1 — Define and Submit Workflow Intent
A user defines what they want the pipeline to do in a structured specification (e.g., standardise counterparty names, validate account IDs, then enrich customer data) and submits it to the composer

2 — Check the Capability Registry
The composer turns the specification into steps and checks the registry to find the approved, reusable processing capabilities corresponding to the requirements. It also checks that the requested capabilities and versions are valid

3 — Compose the Pipeline
The composer takes the specification and the selected capabilities and assembles them into an executable workflow. The composer does not do the data processing itself; it essentially builds the instructions for the processing layer

4 — Execute the Capability Pipeline
The assembled workflow runs the reusable capabilities in the required sequence (e.g., standardisation, validation, enrichment) against the dataset



Source: AWS, Quinlan & Associates analysis

4.3.2 Establishing Governance over Model Selection

Key Considerations

Primary Focus Area(s): Governance and Accountability
Other Relevant Focus Area(s): Data Infrastructure

With data appropriately prepared, the next step is translating it into predictive or prescriptive insights. Extracting insights from large volumes of structured and unstructured data may call for specialised models, such as machine learning ("ML") algorithms.

The appropriate model generally depends on the nature of the underlying data and the intended use case. A consistent basis for evaluating candidate models helps AIs identify those that perform reliably in their specific operational context.

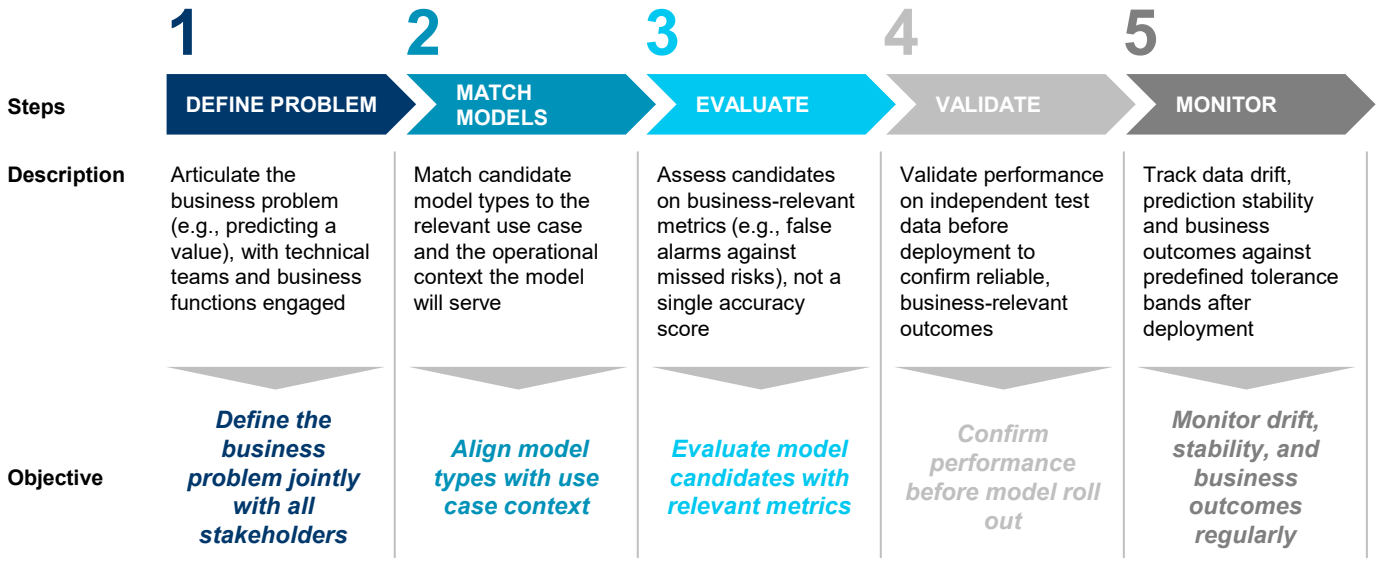
Observed Practices

The HKMA has observed that some AIs have established a structured governance process for identifying an appropriate model, comprising: (1) defining the business problem the model is intended to address, (2) matching candidate model types to the relevant use case, (3) evaluating candidate models against business-relevant performance metrics, (4) validating performance on independent test data before deployment, and (5) monitoring model performance post-deployment.

When adopting such a structured process, AIs may begin model selection with a clear articulation of the underlying business problem, such as flagging potentially suspicious transactions. AIs can ensure that their governance processes outline appropriate engagement between technical teams and business functions to support a shared understanding of the operational context and workflows the model is intended to serve.

In evaluating candidate models, AIs may consider selecting performance metrics that reflect the business objective and risk management needs, rather than defaulting to a single measure of accuracy. For example, models used for fraud detection or risk assessment may warrant metrics that appropriately balance false alarms against missed risks. The objective of independent testing is to identify the model that delivers highly reliable and business-relevant outcomes, rather than the highest score on a single metric *[see Industry Example 10 for reference]*.

Figure 11: Structured Model Evaluation Process



Source: HKMA, Quinlan & Associates analysis

Industry Example 10

Some AIs implemented a model selection strategy that weighs precision against computational efficiency, acknowledging the trade-offs between complex, resource-intensive architectures and simpler, faster alternatives. By evaluating models against applicable selection criteria such as reasoning capabilities and resource requirements, AIs identify models best suited to specific operational scenarios.

Following deployment, AIs may consider maintaining monitoring of model performance over time and defining key metrics to track, such as data drift, prediction stability, and business outcomes. Monitoring results can trigger periodic re-evaluation or retraining when performance falls outside predefined tolerance bands, so that model selection is treated as part of a continuous review cycle rather than a one-off exercise *[see Figure 11]*.

Emerging Practices

For GenA.I. applications, AIs may further consider that model selection can involve assessing not only the underlying model, but also the configuration and approach used to tailor it to the use case. The HKMA's report on GenA.I. "Responsible Innovation with GenA.I. in the Banking Industry – Practical Insights from the GenA.I. Sandbox"[10] provides examples of factors that may be considered, including model capabilities, efficiency, compatibility with the intended use case and the approach used to customise or augment the model *[see Section 6.8 in the Technical Appendix for more details]*.

4.3.3 Strengthening Business Cases for Model Development

Key Considerations

Primary Focus Area(s): Data Infrastructure

Other Relevant Focus Area(s): Data Quality Management and Control

The adoption of specialised models also depends on a sustainable investment case for the supporting data infrastructure. Each model typically requires data pipelines and validation processes, and how these are provided has a significant bearing on the overall cost and complexity of implementation.

Observed Practices

The HKMA has observed that some AIs assess model returns against a broader analytics capability (e.g., as a long-term multi-model asset), rather than a single model in isolation. Doing so may give AIs a fuller view of the value that such infrastructure can deliver over time *[see Industry Example 11 for reference]*.

[10] HKMA. October 2025. "Responsible Innovation with GenA.I. in the Banking Industry – Practical Insights from the GenA.I. Sandbox". https://brdr.hkma.gov.hk/eng/doc-ldg/docId/getPdf/20251031-6-EN/20251031-6-EN.pdf.

Industry Example 11

Some AIs established a centralised development marketplace of pre-approved, reusable functions to streamline solution development. This approach enables users to incorporate standardised components into their solutions, ensuring consistency, reducing duplication of effort, and enhancing overall security by leveraging vetted code assets.

Emerging Practices

AIs may consider further improving the business case for their data infrastructure by adopting a more industrialised approach to model development and validation, transitioning from 'building from scratch' for every project to leveraging standardised, modular building blocks with capabilities such as feature stores (i.e., repositories that consistently serve curated, validated features for both model training and inference) or model registries (i.e., systems that track model versions, lineage, and approval status) *[see Section 6.9 in the Technical Appendix for more details]*. Such an approach may enable AIs to achieve economies of scale in data infrastructure and validation processes and to support the sustained development of analytical capabilities commensurate with their business needs.

When establishing shared feature stores and model registries, AIs may consider implementing granular access controls to govern who can access, modify or deploy particular data features and models. Permissions could be defined by role, data sensitivity, and environment, with access logging and periodic reviews to ensure that the benefits of shared infrastructure do not result in inappropriate access to sensitive data or analytical assets *[see Section 6.11 in the Technical Appendix for more details]*.

4.4 Discoverability

Discoverability is an essential component of Data Excellence for organisations, enabling their end users to identify valuable data assets in a timely manner.

Through its industry engagements, the HKMA has identified four related areas: (1) maintaining visibility into data assets, (2) keeping data inventories and metadata current, (3) establishing consistent entitlement management, and (4) enabling effective data delivery and serving. The following subsections set out key considerations for each area, together with corresponding observed and emerging practices *[see Figure 12]*.

Figure 12: Discoverability Summary

Consideration(s)	Suggested Practice(s)	GA[1]	DQ[2]	DI[3]
Maintaining Visibility into Data Assets	• Identify and maintain a map / catalogue of unstructured and semi-structured data assets to improve enterprise-wide visibility of their location, format, ownership, and accessibility • Assess and record data sensitivity, business value, and lifecycle status to support prioritisation and proportionate governance • Establish metadata standards to promote consistent interpretation, interoperability, and compatibility with relevant tools and data-sharing arrangements • Define validation, storage, access, and protection requirements for extracted information to maintain reliability, traceability, and appropriate handling	★	✓	★
Keeping Data Inventories and Metadata Current	• Combine automated metadata capture with accountable owner and steward validation to keep classifications, definitions, permitted uses, and catalogue records up to date • Present lineage, known limitations, quality indicators, and access conditions to help users assess relevance without implying universal accuracy or accessibility	○	✓	★
Establishing Consistent Entitlement Management	• Establish an enterprise-wide access model with standard roles, clear ownership, consistent approval criteria, and controlled exceptions to promote consistent entitlement decisions • Conduct periodic and event-driven entitlement reviews calibrated to data sensitivity and privilege to identify excessive or outdated access • Revoke unnecessary or expired access promptly to reduce the risk of inappropriate data access or use • Document material access decisions, exceptions, duration, and outcomes to maintain a complete and traceable record • Apply proportionate context-aware or granular controls where standard roles do not adequately address complex access requirements	★	○	✓

Figure 12: Discoverability Summary (Cont.)

Consideration(s)	Suggested Practice(s)	GA[1]	DQ[2]	DI[3]
Enabling Effective Data Delivery and Serving	• Define service requirements for material data uses to establish appropriate freshness, volume, availability, security, traceability, and recovery expectations • Select delivery mechanisms suited to the architecture and use case to provide controlled and resilient data retrieval • Implement capacity, performance, error-handling, timeout, dependency, escalation, and retry controls to maintain service resilience without amplifying failures • Provide authorised users with governed self-service capabilities, certified data, common definitions, controlled workspaces, and clear support routes to reduce reliance on intermediaries • Provide training on data meaning, limitations, access conditions, and permitted uses to support appropriate and confident use of self-service tools	✔	○	★

Focus Area (Legend)	
Primary Focus Area(s)	★
Other Relevant Focus Area(s)	✔
Less Relevant Focus Area(s)	○

[1]Governance and Accountability, [2]Data Quality Management and Control, [3]Data Infrastructure

Source: HKMA, Quinlan & Associates analysis

4.4.1 Maintaining Visibility into Data Assets

Key Considerations

Primary Focus Area(s): Governance and Accountability, Data Infrastructure

Other Relevant Focus Area(s): Data Quality Management and Control

Unstructured and semi-structured data can accumulate across systems, business units, and archival repositories over time. A clear view of these assets helps ensure that they are known to the business units best placed to use them, and that redundant or obsolete data can be identified for disposal.

Visibility is also the starting point for using and governing data effectively. When the location, format, sensitivity, and ownership of data assets are known, AIs can apply appropriate access controls and data quality standards, and realise the risk management and commercial value of the data they already hold.

Observed Practices

The HKMA has observed that AIs adopt a range of approaches to building and maintaining a map of their unstructured and semi-structured data assets within their data infrastructure. The appropriate approach will depend on the AI's existing technology landscape and operating model, varying according to its circumstances *[see Industry Example 12 for reference]*.

Industry Example 12

Some AIs utilise a data catalogue as the foundational bridge between data discoverability and advanced analytics. By using metadata management to map and link structured, semi-structured, and unstructured data, these AIs enable search, indexing, and downstream A.I. capabilities. This integrated access and cross-referencing between structured datasets and unstructured content allows AIs to create a unified data layer that supports consistent analysis and A.I. applications.

Emerging Practices

AIs may consider three related elements that are relevant to mapping their data assets: (1) how information about data assets (e.g., location, format, and ownership) is consolidated or made accessible across the organisation, (2) how data assets are identified and kept up to date, so that the inventory remains accurate as systems and business activities change, and (3) how the sensitivity and business value of the data identified is assessed and recorded *[see Section 6.10 in the Technical Appendix for more details]*.

Each of these involves trade-offs around governance consistency, coverage, accuracy, cost, and the ability to keep pace with changes in source systems. Assessment of these trade-offs may consider the nature, scale, and complexity of the AI's data holdings.

Appropriate metadata standards can support the usability and interoperability of the data map. Their design may take into account the AI's internal business context and the extent to which compatibility with external tools or data-sharing arrangements is relevant.

Once information is extracted from unstructured or semi-structured data, AIs may then define clear rules for how it is validated, stored, governed, and made available for downstream analysis. Such workflows are particularly important where the underlying data contains sensitive information (e.g., customer communications, identity documents, or contracts). AIs may consider establishing appropriate processes, such as validation checks and access controls, to ensure that extracted data remains reliable, traceable, and handled in a manner consistent with applicable data protection requirements.

A well-maintained data map underpins access governance and delivery arrangements. By clarifying ownership, it allows consistent controls to be applied to unstructured and semi-structured assets. Over time, this foundation can also enable broader integration of unstructured data into advanced use cases.

4.4.2 Keeping Data Inventories and Metadata Current

Key Considerations

Primary Focus Area(s): Data Infrastructure
Other Relevant Focus Area(s): Data Quality Management and Control

Beyond knowing which data assets exist, staff need timely and accurate information about them. Current metadata allows data users to assess the relevance of data quickly and retrieve it when needed, without lengthy clarification or approval steps.

Keeping metadata current is easier when it is captured consistently, with common naming conventions applied across business units and systems. This allows data catalogues to deliver their full value and reduces repeated clarification and reconciliation work.

Observed Practices

The HKMA has observed that some AIs have assigned clear ownership to data catalogues / data dictionaries to ensure that the metadata stays current. By assigning responsibility to specific teams or individuals, these institutions help ensure that metadata remains accurate, comprehensive, and up to date. This structured approach not only supports consistency across different business units but also enhances transparency, making it easier for stakeholders to locate, interpret, and rely on critical data assets *[see Industry Example 13 for reference]*.

Industry Example 13

Some AIs implement data dictionaries to enhance organisational transparency, providing users with clear visibility into available data assets and their definitions. These dictionaries establish explicit data ownership and integrate embedded request and approval workflows, ensuring that both requesters and approvers have a clear and consistent understanding of the specific data being accessed, thereby facilitating informed and accountable decision-making.

Emerging Practices

AIs may also look to automated tools to assist in capturing and updating metadata, while accountable data owners and stewards remain responsible for confirming data classifications and permitted uses. Appropriate maintenance and review arrangements may also help ensure that catalogue information remains complete and current as data assets and requirements change *[see Section 6.10 in the Technical Appendix for more details]*.

Automated data catalogues and lineage tools can enhance data relevance by automatically capturing and maintaining metadata on data assets (e.g., definitions, ownership, etc.), while tracking how data moves and changes across systems. This may help AIs maintain a more current view of their data environment, supporting auditability and confidence in data integrity. For example, a data catalogue could act as a live context layer, automatically identifying and documenting a newly created customer data table (e.g., owner, key fields, sensitivity classification), and capturing updates as they occur. Lineage tools could trace transformations from the source system to downstream reports without manual intervention, embedding the most current information directly into the metadata. Combined with more intelligent search (e.g., via LLMs), this could allow AIs to treat their data catalogues as an active part of data consumption, surfacing errors and flagging issues in real time, rather than relying on manual review and metadata updates to catch problems after they propagate downstream.

Updated metadata in the catalogue may not, in itself, be regarded as assurance that the underlying data is accurate, suitable for a particular purpose, or accessible to all users. AIs may therefore consider presenting relevant information (e.g., data lineage, known limitations, permitted uses, etc.) to support effective data discovery while maintaining data quality and controls.

4.4.3 Establishing Consistent Entitlement Management

Key Considerations

Primary Focus Area(s): Governance and Accountability
Other Relevant Focus Area(s): Data Infrastructure

Timely access to data for approved activities depends on access arrangements that remain effective as users, responsibilities, and projects change. Clear governance policies help AIs balance the protection of sensitive data with the operational flexibility that different business units need, so that access is neither broader nor more restrictive than required.

Consistency across the organisation is equally important. When access is granted and recorded in the same way across business units and platforms, AIs can see who approved an entitlement, for what purpose, for how long, and whether it was subsequently reviewed or revoked. This applies whether data is distributed across systems or consolidated into a central repository or "golden record".

Observed Practices

The HKMA has observed that some AIs have strengthened access governance by establishing standardised entitlement lifecycle processes, defining standard roles, conducting periodic reviews, and outlining clear conditions for access revocation.

AIs may consider establishing an enterprise-wide access model built around standardised access roles (e.g., role-based access controls), providing a consistent basis for granting access across business units and systems. Responsibilities for administering the model may be clearly assigned, with accountable data owners determining whether access to their data is appropriate, and relevant functions (e.g., business, risk management, etc.) administering permissions and monitoring compliance. In some cases, the HKMA has observed some AIs automating the entitlement process to mitigate operational risk and manual workloads *[see Industry Example 14 for reference].*

Industry Example 14

Some AIs leverage Single Sign-On ("SSO") integrated with Active Directory, synchronised with centralised HR employment records, to automate user access provisioning and de-provisioning. This integration reduces reliance on manual interventions, thereby minimising operational risks while ensuring that access rights are updated in a timely and effective manner.

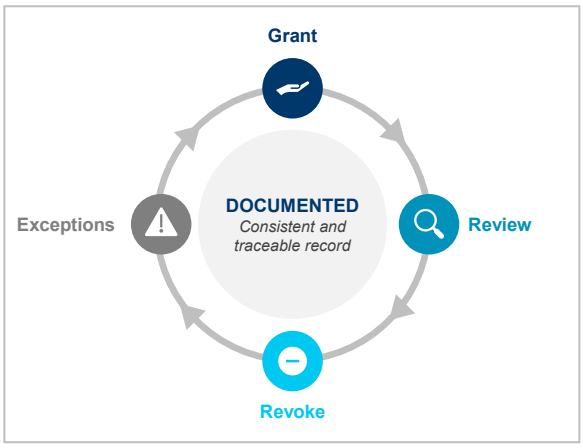
Emerging Practices

The access lifecycle may incorporate both periodic and event-driven reviews. Periodic reviews may be calibrated according to the sensitivity of the data and privilege of the access concerned, while event-driven reviews may be triggered by relevant changes to a user's circumstances.

Access that is no longer required may be revoked promptly, including following staff separation or the expiry of a time-bound approval. Changes to data access are recommended to be documented in a consistent and traceable manner *[see Figure 13].*

Depending on the AI's risk profile and the complexity of its access environment, standard role-based models may not fully address decentralised or complex data holdings, such as unstructured data with dispersed ownership. In these cases, AIs could consider non-traditional mechanisms like attribute-based or just-in-time access controls, which grant

Figure 13: Access Entitlement Lifecycle



Stage	Suggested Practice
Grant	Standardised access roles (e.g., role-based access controls) give a consistent basis for granting access, with accountable data owners determining whether access to their data is appropriate
Review	Periodic reviews calibrated to the sensitivity of the data and privilege of access, plus event-driven reviews triggered by relevant changes to a user's circumstances
Revoke	Access that is no longer required is revoked promptly, including following staff separation or the expiry of a time-bound approval
Exceptions	A controlled exception process (e.g., approvals, purpose-based, time-bound access); recurring exceptions are incorporated into standard roles / attributes or policies

Beyond roles: attribute-based or just-in-time access controls grant permissions dynamically for decentralised or complex data holdings

Source: HKMA, Quinlan & Associates analysis

permissions dynamically. Adopting such models can refine how access is granted and reviewed *[see Section 6.11 in the Technical Appendix for more details]*.

4.4.4 Enabling Effective Data Delivery and Serving

Key Considerations

Primary Focus Area(s): Data Infrastructure
Other Relevant Focus Area(s): Governance and Accountability

The value of discoverable data is realised when it can be delivered to end users at the speed they require. Data infrastructure that can accommodate peaks in request volumes and payload sizes,[11] and that supports efficient querying of the underlying data, allows data to reach users reliably when it is needed.

Delivery also depends on the user experience. When end users have visibility of available data and intuitive tools for self-service, they can obtain the data they need directly and in a timely manner.

Observed Practices

The HKMA has observed that some AIs have defined service requirements for material data uses, including required freshness, volumes, peak demand, availability, security, traceability, and recovery. The delivery mechanism may involve APIs, governed query services, batch interfaces, event-based feeds, data virtualisation, or other arrangements appropriate to the architecture and use case *[see Industry Example 15 for reference]*.

Industry Example 15

Some AIs improve data discoverability by establishing a group-wide data catalogue for defining and categorising data. This catalogue clearly defines all data to ensure that business and technical metadata are unified across the group. Any changes to definitions require sign-off from a group design authority. This allows users across the group to understand the available data within the organisation, enhancing data discoverability.

Moreover, these AIs strengthen data literacy within the organisation to further enhance data discoverability. For example, they conduct user training on BI tools.

Emerging Practices

Relevant controls may include capacity and performance monitoring, error handling, timeout thresholds, dependency mapping, procedures for escalation, and retry logic.[12] Retry arrangements may be designed carefully so that repeated requests do not amplify an existing capacity or service failure. The objective is not to introduce a particular connecting layer, but to provide controlled and resilient retrieval appropriate to the underlying requirements.

[11] "Data that is sent in a request or received in response." HubSpot. 2024. *What is Payload?*. (https://blog.hubspot.com/website/what-is-payload#what-is-a-payload-in-api).

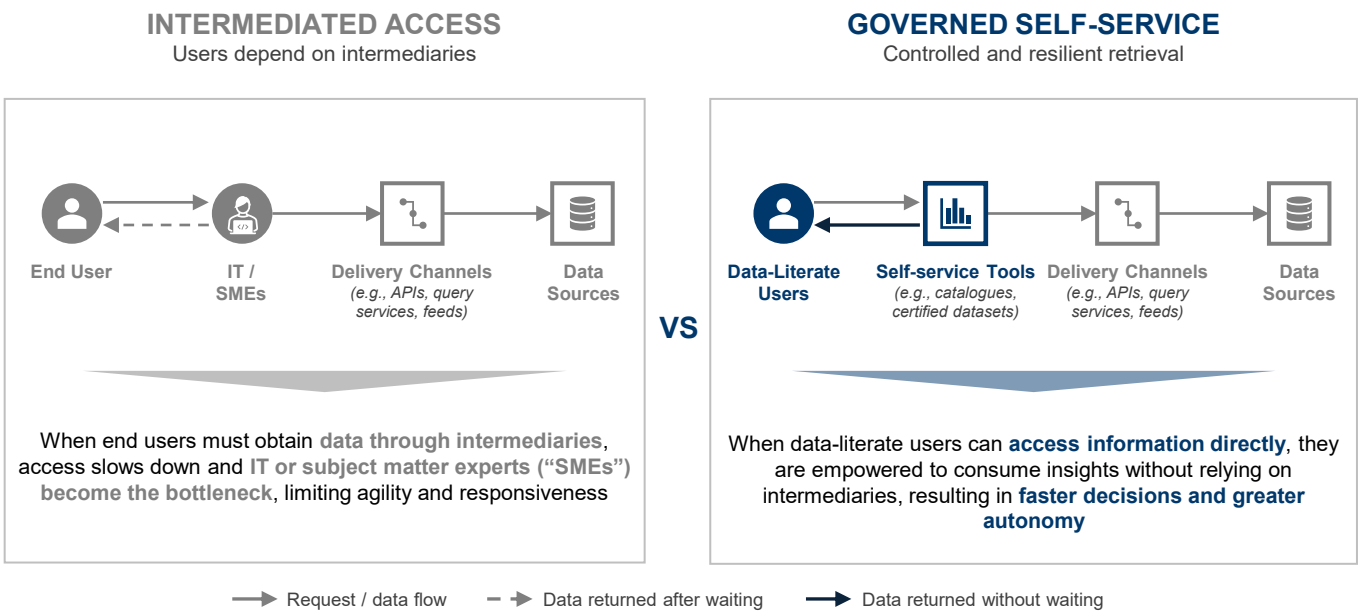
[12] "An application [that] handles transient failures when it tries to connect to a service or network resource, by transparently retrying a failed operation." Microsoft. Accessed 2026. Retry pattern. https://learn.microsoft.com/en-us/azure/architecture/patterns/retry.

AIs may also consider providing self-service discovery and analytical capabilities appropriate to the needs and skills of authorised users. These may include searchable catalogues, certified datasets, common business definitions, governed semantic layers, controlled workspaces, reusable queries and dashboards, and clear routes for requesting access or specialist support.

Deploying self-service analytics tools and intuitive dashboards can also reduce reliance on IT intermediaries, particularly in larger AIs where the volume of data-producing systems may exceed the capacity of data-literate business users. The objective is not simply to add a front-end layer, but to enable non-technical users to formulate and run queries with a reasonable degree of confidence. This increase in accessibility, in turn, allows IT and data teams to focus on higher-value activities rather than routine report generation.

AIs may provide appropriate technical training that enhances users' data literacy,[13] including their understanding of the meaning, limitations, and permitted uses of relevant data, rather than focusing solely on how to operate a particular business intelligence ("BI") tool *[see Figure 14]*.

Figure 14: From Intermediated to Governed Self-service Delivery



Source: HKMA, Quinlan & Associates analysis

[13] "The ability to explore, understand, and communicate with data in a meaningful way." Tableau. Accessed 2026. Data literacy explained: Definition, importance, examples, and more. https://www.tableau.com/data-insights/data-literacy/what-is.

05 Looking Ahead

Looking ahead, as the risk landscape continues to evolve, Data Excellence will remain a strategic foundation for the next phase of banking transformation. Data goes beyond a compliance requirement and should be recognised as an asset. By embedding robust governance, quality controls, and modern infrastructure, AIs can unlock data as a secure, scalable, and responsible asset. This shift will empower smarter risk management, sharper decision making and responsible innovation.

Data capabilities across the DataGRID framework: granularity, reliability, ingenuity and discoverability, can support the reliable and fit-for-purpose use of data for advanced analytics, more agile risk management and reporting, and the continued development of granular data reporting. They are also fundamental to the New Risk Data Strategy under the Fintech Promotion Blueprint, which seeks to strengthen AIs' data management capabilities. Only with high-quality data can AIs unlock the full potential of advanced A.I. and DLT applications. These capabilities can also enable AIs to move beyond layering technologies over existing processes. They offer AIs the opportunities to leverage advanced technologies to reshape customer engagement, product and service delivery, and participation in evolving financial-market ecosystems. Data Excellence should therefore form an integral part of each institution's strategic review of its business model amid digital transformation.

As a cross-cutting enabler of responsible and sustainable innovation, Data Excellence contributes to Fintech 2030's vision of reinforcing Hong Kong's position as a robust, resilient, and future-ready international fintech hub. To translate this vision into reality, the pace and priorities of each AI's journey should remain proportionate to its strategy and risk profile.

06 Appendix

6.1 Data Capture and Ingestion

The appropriate data capture mechanism will vary by source type. AIs may select the method best suited to each source, such as Change Data Capture ("CDC") or snapshotting for relational databases, Application Programming Interface ("API") polling for external feeds, and Optical Character Recognition ("OCR") or Natural Language Processing ("NLP") tools for extracting structured data from unstructured content (e.g., scanned physical documents).

For data capture, AIs may decide between two approaches: (1) snapshotting and (2) CDC.

Snapshotting extracts the full or partial state of a source system at each run. This method is often considered conceptually simple to implement, but places higher load on source systems and scales poorly with data volume, since each run re-reads parts of or the entire dataset regardless of the underlying changes.

CDC, by contrast, captures incremental changes (e.g., inserts, updates, and deletions), typically via transaction or redo logs. This can reduce the load on source systems and support lower-latency data propagation. However, CDC is generally more complex to implement, depends on the compatibility and configuration of the source system (e.g., access to write-ahead logs or equivalent records), and typically requires an initial snapshot for historical loading or backfilling.

For data ingestion, AIs may also determine how frequently captured data needs to be ingested.

Batch processing transfers data through scheduled jobs, such as "ETL" or "ELT" processes conducted at fixed intervals. It is generally simpler to build and operate and may require less infrastructure but inherently introduces latency. It is therefore more suitable for use cases that can tolerate delayed data availability.

Event-driven and streaming architectures, supporting near-real-time or continuous data availability, propagate data continuously as it is generated. These require greater engineering investment and operational maturity, including capabilities to manage message ordering, fault tolerance, backpressure, and delivery guarantees (e.g., at-least-once or exactly-once processing).

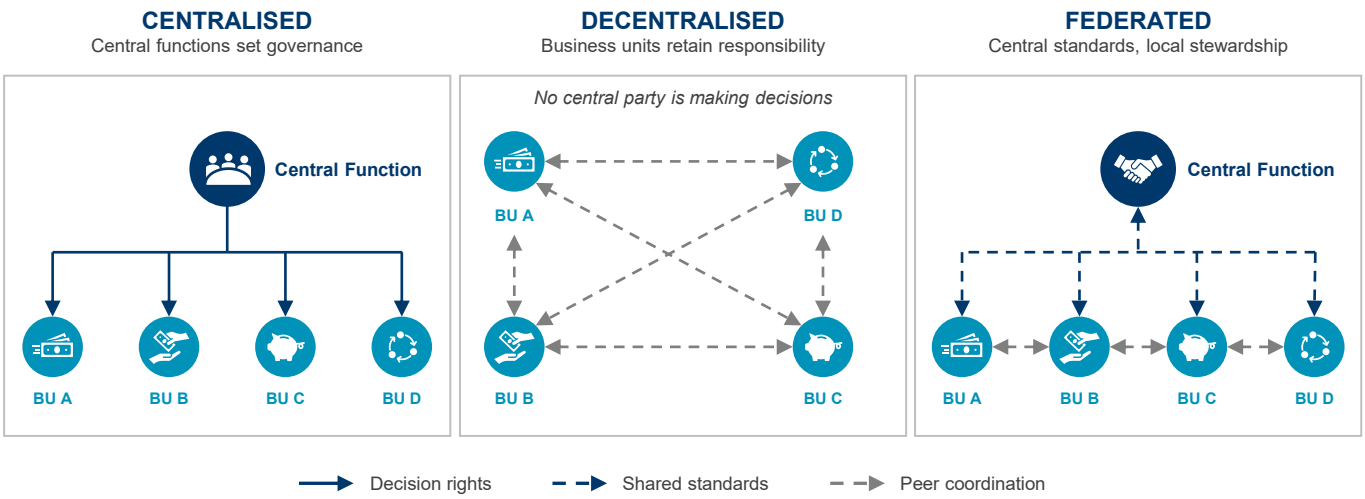
AIs may calibrate both the data capture mechanism and ingestion frequency to the requirements of each use case, rather than applying a single standard across the organisation, reflecting the differing needs of relevant functions, including business, risk management, compliance, and operations.

6.2 Data Governance Operating Models

Data governance operating models define where decision rights sit, who owns data, and how governance standards are applied across the enterprise. AIs may consider three common types of models, depending on their organisational realities: centralised, decentralised and federated models. Each involves different trade-offs between control, responsiveness, consistency and scalability.

Centralised models retain centralised functions, which set governance policies, standards and controls, with business units operating within this common framework. This arrangement provides consistency and clear

Figure 15: Data Governance Operating Models



Source: IBM, Snowflake, Databricks, Microsoft, HKMA, Quinlan & Associates analysis

accountability across the entire enterprise, but can create bottlenecks and reduce responsiveness to domain-specific needs, as central teams lack business-specific contexts. AIs may find this model the most appropriate where standardisation and control are paramount.

Decentralised models utilise individual business units or data domains to retain responsibility for governance and decision-making. Without a central party to make decisions, this arrangement enables faster decisions and stronger alignment with local business needs, but increases the risk of inconsistent definitions and fragmented controls without strong cross-functional communication. AIs may find this model the most effective where business units have sufficient data maturity and operate with significant autonomy, such as subsidiaries across multiple geographies.

Federated models balance centralisation and decentralisation. Enterprise-wide governance (e.g., policies, standards and guardrails) is defined centrally, while business units retain responsibility for implementation and day-to-day stewardship. This balances enterprise consistency with local responsiveness and can scale effectively across complex organisations. However, AIs may note that this model requires clear decision rights and strong coordination to monitor business-unit adherence.

The appropriate model for AIs will depend on organisational structure, regulatory and risk requirements, data sensitivity, business-unit autonomy and data maturity. In practice, organisations may also apply different models to different data domains rather than adopting a single model enterprise-wide *[see Figure 15]*.

6.3 Data Quality Management and Controls

AIs may consider strengthening data issue identification by automating pipeline-wide data quality monitoring and controls, especially with deterministic validation and statistical techniques. These automated validation controls reduce the burden on operational staff, who would otherwise need to identify and correct errors manually at a later stage, often with less surrounding context.

At the point of capture, deterministic validation, which involves mandatory field requirements, ensures real-time checks at the point of entry. This approach may cover syntax, schema, and format, preventing common input errors before they are recorded. While these rules can be relatively straightforward to implement, they may occasionally prove brittle or inflexible, as a data point could meet formatting requirements while remaining substantively unlikely in certain contexts.

As such, AIs may also consider statistical techniques as an alternative or supplement to deterministic validation. Such checks are designed to flag or reject entries that fall outside parameters defined by statistical analysis (e.g., z-score outlier detection, moving-average control charts). However, statistical approaches also require periodic recalibration as data distributions change over time. Consequently, institutions might choose to reserve these more resource-intensive methods for high-risk data domains rather than applying them uniformly across all systems.

AIs may consider integrating both approaches for a more comprehensive framework for data quality controls. As deterministic and statistical checks tend to have distinct failure modes, they may be better suited when done together. For instance, AIs may apply rules-based checks during initial ingestion, followed by statistical evaluations to identify subtle anomalies within formally valid data.

Beyond point-of-capture controls, AIs may also consider ongoing data quality monitoring to periodically re-assess data already in the pipeline, which may help detect degradation that emerges after initial capture. The same deterministic and statistical techniques discussed above for point-of-capture controls could also be applied for monitoring purposes. Deterministic rules may be run periodically across existing data to identify errors (e.g., inaccurate historical customer data), while statistical techniques could be applied to detect changes in data distributions (e.g., trends in missing records) that may not be apparent at the point of entry alone.

6.4 Fit-for-purpose Data Infrastructure

Evaluating modern data architecture requires AIs to align their architectural choices with specific operational objectives rather than pursuing a single universal standard. In short, AIs may consider (1) data warehouses, (2) data lakes, (3) data lakehouses, and (4) data mesh / fabric.

Data warehouses generally focus on structured and relational data to support predictable query activity. These systems usually rely on schema-on-write principles, which can deliver high query performance for user access. However, a notable trade-off can include rigidity when handling unstructured information or rapidly evolving data schemas. Uncoordinated mart proliferation may also recreate data duplication issues across business units, requiring careful governance to maintain consistency. When designing a data warehouse, AIs may prioritise mapping out each stage of the data warehouse (including initial schema design for ingestion, integration with other databases, and end presentation to users) to help ensure high-performance querying.

Data lakes tend to support diverse data types through a schema-on-read approach, prioritising raw data retention and exploratory workloads. This design often accommodates machine learning initiatives where data scientists need access to unprocessed files. A key advantage of data lakes is their capacity to store vast volumes of varied data cost-effectively. Conversely, these environments frequently require careful management of file formats alongside strict folder conventions to avoid devolving into disorganised data swamps. Query performance on raw data may also lag behind structured alternatives, such as data warehouses, without substantial indexing efforts. AIs adopting a data lake may consider setting up a multi-zone storage strategy with consistent folder-naming conventions and partitioning standards to help prevent the environment from becoming disorganised.

Data lakehouses attempt to bridge these approaches by combining cost-effective object storage with transactional layers. This allows platforms to support both user queries (e.g., business intelligence reporting) and advanced analytical workloads. However, data lakehouse platforms introduce transactional table formats, which may require careful evaluation of user support, access, and version control. Implementing multi-layered medallion patterns to document transformation logic clearly is often necessary to help maintain system reliability. To implement a data lakehouse design, AIs may incorporate open-source transactional table formats paired with a medallion architecture to incrementally clean and structure data.

Data mesh and data fabric paradigms tend to shift the focus towards organisational scalability and automated connectivity. These frameworks often treat data as a distinct product or seek to abstract physical storage locations through virtualised integration layers. These approaches can be used in conjunction with data warehouses, data lakes, and data lakehouses. An advantage of these paradigms is improved organisational alignment and flexible access to distributed data sources. However, decentralised ownership or virtualised integration layers may introduce complexity in tracing data provenance and maintaining consistent security policies. Robust metadata synchronisation is, therefore, often essential to manage operational overhead effectively. An AI adopting a data mesh may map out self-contained business domains with dedicated data product owners, whereas adopting a data fabric usually focuses on building a metadata-driven virtualisation layer to help connect distributed sources.

6.5 Data Lineage: The Medallion Approach

To retain greater visibility across their data transformation, AIs may consider structuring storage across distinct refinement states, from raw ingestion through to curated outputs. As such, AIs may adopt a tiered data design, such as that of a medallion architecture, which separates and stores data in its various states in three main layers (i.e., Bronze, Silver, and Gold) during data transformation.

More specifically, the Bronze layer stores all captured data, preserving an accurate history of ingested records. Batch or streaming inputs from manual, structured, or unstructured sources would be stored in this landing zone for raw data.[14] The Silver layer applies data cleansing and harmonisation procedures (e.g., deduplication rules, enforcing data types, etc.) for easier processing. All processed inputs are stored for greater traceability. The last layer, i.e., the Gold layer, builds on top of the Silver layer, creating and storing aggregated, clean outputs for downstream users and applications. Preserving intermediate steps in this manner may make it more practical for AIs to trace data back to its point of origin or even pivot calculation methodologies, as all the underlying data is captured.

When adopting the medallion architecture, AIs need to consider two main factors: (1) storage and table format and (2) layer separation. In terms of storage and table format, AIs evaluating modern analytical platforms generally choose between two dominant options, though hybrid and other types of implementations exist in parallel. The two main options are (1) native warehouse layering used with data warehouses and (2) open table formats used with data lakehouses.

Native warehouse layering implements the three layers as logical schemas under a data warehouse storage architecture. This approach may be more straightforward to govern (e.g., via SQL-native access controls), as security policies and user permissions can be managed centrally within a single database engine. However, this technique may offer less granularity on file-level versioning or time travel due to the limitations of a data warehouse.

Open table formats, on the other hand, store the files of each layer alongside discrete metadata layers that map into structured tables. As logs track all file-level modifications, this approach generally provides more reliable data processing guarantees (e.g., ACID transactions), greater flexibility to update and control data structures (e.g., schema enforcement and evolution), and increased access to historical data states (e.g., time travel or versioning) than native warehouse layering. In particular, this combination may enable point-in-time reconstruction of a given layer, which could be useful for demonstrating what data looked like at a particular decision point rather than only its current state. However, this approach tends to depend on the maturity of compatible query engines and tooling, and it may introduce additional operational overhead due to managing file compaction and metadata growth as data volumes increase.

While native warehousing and open table formats represent distinct foundational choices, they are no longer strictly mutually exclusive. Many AIs are adopting converged architectures. For example, an AI may utilise open table formats for its raw and cleansed data layers to maximise cost-efficiency and historical tracking, while leveraging a native data warehouse engine for the highly aggregated serving layer to achieve sub-second query performance. Furthermore, major commercial data warehouses increasingly support seamless integration with open table formats, allowing AIs to decouple their storage and compute strategies.

In addition to storage and table format, AIs may need to consider layer separation, namely, how the boundaries between Bronze, Silver, and Gold are enforced. IT staff may determine, as a matter of architectural principle rather than product selection, whether to separate the relevant data layers (1) physically, through separate storage locations or buckets; (2) logically, through shared storage with distinct schemas and permissions; or (3) through a hybrid of physical and logical separation. This decision can provide the basis for subsequent architectural and security configurations.

Physical separation isolates data layers using distinct storage buckets, isolated networks, or separate cloud tenancies. This configuration might provide robust security boundaries and could limit blast radii during potential breaches, as access to the Bronze layer does not inherently grant network pathways to the Gold layer. The downside of this method could involve increased data movement latency and higher administrative burdens, as engineers might need to replicate governance policies across multiple disparate environments.

[14] "Unprocessed information that might come from the cloud, on-premises resources, or edge-computing devices." AWS. Accessed 2026. What is a Data Cloud?. https://aws.amazon.com/what-is/data-cloud/.

Logical separation maintains all three layers within shared storage environments but enforces boundaries through distinct database schemas, namespaces, and role-based access permissions. This architecture could reduce infrastructure complexity by potentially eliminating the need to physically move data between isolated environments. Conversely, a single misconfigured access policy under this approach may inadvertently expose unmasked Bronze data to users authorised to access only aggregated Gold analytics.

A hybrid approach attempts to balance these factors by physically isolating the raw Bronze layer while logically separating the refined Silver and Gold layers within a unified analytical environment. This design might offer strict protection for sensitive, unprocessed source data while potentially enabling faster query performance and simplified governance for downstream business intelligence workloads. Selecting amongst these separation strategies will require AIs to weigh their specific security tolerances against their operational resource constraints.

6.6 Managing Inflexible Data Pipelines

To support adaptable data pipeline architectures, AIs may consider how they structure their data processing across three key dimensions: (1) the distribution of processing, (2) the integration of business logic, and (3) the modularity of the pipeline itself. These decisions can influence how easily data systems are maintained, scaled, and adapted as future business requirements evolve.

AIs may first evaluate the distribution of processing, specifically, whether it can be (1) centralised or (2) distributed, depending on their anticipated workloads and operating models. Centralised processing often consolidates transformation logic within a shared platform. As this approach unifies processing logic, it might be suitable when multiple business functions rely on shared datasets or common processing needs.

Alternatively, distributed processing allocates computing tasks across multiple systems or domains. This approach might improve scalability and resilience as processing demands increase over time. In scenarios where different business units have distinct processing requirements or operate at varying scales, distributed processing could allow for greater flexibility without necessarily creating centralised bottlenecks. Since these approaches often serve different institutional needs, some organisations may find value in combining elements of both to support consistency while accommodating diverse operational requirements.

AIs may also assess the extent to which business logic may be (1) embedded within processing pipelines or (2) separated from the underlying infrastructure. This consideration could influence how easily processing rules might be updated as future business requirements change. Embedding business rules and transformation logic directly within workflows often simplifies initial implementation, as it can reduce the number of moving parts during early development phases. Where processing requirements appear stable and closely aligned with a specific application, an embedded approach may be appropriate.

Conversely, separating business logic from the infrastructure could provide greater flexibility in environments where business rules or data validation requirements are expected to evolve. This separation might allow logic to be updated somewhat independently of the underlying platform. It could also support the reuse of processing components across multiple use cases, potentially reducing the need to redesign pipelines completely when requirements shift.

Finally, AIs could examine whether pipelines might be better implemented as (1) monolithic workflows or (2) modular processing components. This choice is likely to affect how easily pipelines can be maintained and adapted to future architectural shifts. AIs may also consider maintaining a capability registry, which provides a centralised record of available processing components, their functions, and how they can be used or reused across pipelines.

Monolithic pipelines tend to consolidate processing activities within a single continuous workflow. This unified approach can simplify initial implementation, as all operations exist within one shared codebase. This design could be applicable when processing requirements remain relatively straightforward, and the pipeline supports a limited number of data sources, transformations, or downstream applications. It may also reduce the coordination required between components in less complex environments.

Modular pipelines, by contrast, separate processing into independent components that can often be deployed separately. This strategy may be suitable where processing requirements show increasing complexity or

Figure 16: ETL vs. ELT

Dimension	Extract → Transform → Load	Extract → Load → Transform
Applicable Scenarios	Legacy databases and complex multi-source analytics	Advanced analytics, such as real-time processing and machine learning
Process Sequence	Transform before load	Load raw data first, transform after
Data Residence	Separate processing server	Target data warehouse
Supporting Data Format	Structured data *(Cleaned, transformed)*	All types, both structured and unstructured *(Native format re-queried as needed)*
Speed / Latency	Slower *(An extra step before load)*	Faster *(Near-real-time via parallel processing)*
Data Auditability	Unfavourable *(Raw data may not be retained)*	Favourable *(Historical audit trail retained)*

Source: AWS, Google, Microsoft, Quinlan & Associates analysis

where new data sources are expected to be integrated gradually. Adopting modular pipelines could make it easier to update individual processing functions without necessarily affecting the entire workflow. A capability registry could further support this approach by providing a centralised view of available processing components and their intended uses.

In some cases, raw data may also not be retained in an accessible form. Where this occurs, AIs may find it difficult to reprocess historical data using updated processing techniques (e.g., cleansing, enrichment, or modelling), as the original data has already been transformed. While these limitations may be manageable for smaller datasets and narrower use cases, they can become harder to sustain as scale increases (e.g., in terms of volumes, source diversity, and analytical complexity).

6.7 ETL vs. ELT

AIs may adopt either an "Extract, Transform, Load" ("ETL") approach or an "Extract, Load, Transform" ("ELT") approach (or a combination of both), depending on their business and risk management needs *[see Figure 16].*

Under an ETL approach, data is cleaned and transformed before being loaded into the target environment. This can lower storage and bandwidth requirements and simplify downstream processing. However, it may remove granularity required by future models or use cases, potentially necessitating subsequent re-engineering of the data capture pipeline.

Under an ELT approach, data is captured at the finest practicable level of granularity and stored close to its native form, with transformation and aggregation performed downstream as required. This preserves flexibility for future or unanticipated use cases but may increase storage and processing costs, depending on the nature of the data.

6.8 Technical Considerations for GenA.I. Model Selection

The HKMA has identified a number of criteria that AIs may consider when evaluating GenA.I. models. These may include:

- **Performance and resource requirements:** More complex models may offer higher precision but can require greater training or fine-tuning time and more computationally demanding inference. AIs may therefore consider the trade-off between model performance, model size, inference speed and resource requirements based on the needs of the use case.

- **Open or closed source:** Open models may allow greater customisation through fine-tuning on proprietary data, while closed models accessed through APIs may offer greater ease of use and lower operational overhead, but with greater restrictions on customisation and less transparency over model updates or architecture. Closed models may also require robust third-party data governance.

- **Licence compatibility:** AIs may consider whether the model's terms and conditions are consistent with the institution's policies and guidelines, including whether commercial use is permitted.

- **Domain specificity:** AIs may consider whether the application requires domain-specific capabilities. For example, a model specialised for code generation may be more appropriate for coding-related applications.

- **Reasoning and multilingual or multimodal capabilities:** AIs may consider whether the application requires reasoning, multilingual or multimodal capabilities, and whether the model's reasoning capabilities are appropriate for the requirements of the application, including considerations such as audit trails and output explainability.

- **Context window:** Where use cases involve lengthy inputs, AIs may consider whether the model supports a sufficiently large context window.

- **Model size and efficiency:** AIs may consider whether inference speed is consistent with the requirements of the use case, particularly where users expect near real-time responses, and balance model size against efficiency.

- **Benchmarks:** Public benchmarks can provide a starting point for comparing LLM capabilities. However, their relevance may depend on the specific use case, and AIs may therefore consider whether benchmark results are representative of the banking task or capability being assessed.

- **Technical compatibility:** AIs may consider compatibility with relevant development frameworks and whether the model supports efficiency-focused fine-tuning approaches, such as Low-Rank Adaptation, where required.

- **Hardware requirements:** AIs may consider whether the model's GPU memory and other infrastructure requirements are compatible with the institution's infrastructure, both for day-to-day operations and ongoing fine-tuning where applicable.

- **Community and ecosystem support:** AIs may consider the availability of documentation and development examples, as well as whether the model is regularly maintained and updated.

6.9 Feature Store and Model Registry Design

This section outlines practical options for feature stores, model registries, and the broader centralisation question, so that an institution can match its technical approach to its own data maturity, latency requirements, and governance structure.

A feature store architecture typically falls into one of three patterns: (1) batch, (2) online, or (3) dual online / offline, each suited to different operational demands. A batch feature store computes features on a schedule, often through nightly ETL jobs, and persists them in a data warehouse or data lake. This pattern suits models that do not require real-time inputs, such as monthly credit scoring, and carries a relatively low engineering overhead. It is generally not usable for latency-sensitive use cases like real-time fraud detection. An online feature store computes and serves features with low latency. This pattern is necessary for transaction-time fraud scoring and similar use cases, but it introduces architectural complexity around stream processing, cache invalidation, and consistency between online and offline values. A dual online / offline architecture attempts to address that complexity by maintaining a single feature definition that serves both a batch store for training and an online low-latency store for inference. Utilising this architecture can reduce training-serving skew, though it tends to require greater platform investment and may only be justified once multiple teams begin sharing features.

The decision to build or buy a feature store turns on a trade-off between speed to value and control. A vendor-managed or cloud-native feature store service typically offers faster time to value and a lighter internal engineering burden, but it may introduce vendor lock-in and raise data residency or regulatory considerations for AIs. An open-source, self-hosted option provides greater control and avoids licensing fees, yet demands internal engineering capacity to maintain, scale, and secure the platform. An in-house built feature store allows maximum customisation to an AI's internal data architecture

and control framework, but it usually entails the highest upfront cost, the longest build time, and an ongoing maintenance commitment.

Model registry architecture also presents a parallel set of choices, centred on integration depth and governance role. A standalone model registry tracks model versions, lineage, and approval status independently of any single training platform. This approach is flexible and tool-agnostic, though it requires separate integration with deployment infrastructure. A platform-integrated registry is tightly coupled with a cloud provider's training and deployment pipeline. Integration effort is generally lower, but the AI may become tied to that provider's ecosystem. Some AIs configure the registry to act as a governance gate, so that a model cannot progress to production without passing through defined approval stages. This configuration creates a built-in control point that can mirror model risk governance requirements, adding process rigour. However, it requires the registry to be wired into deployment tooling rather than used only as a passive catalogue.

Both feature stores and model registries raise a centralisation question that applies across these components. A fully centralised, enterprise-wide store or registry can maximise reuse and consistency and is often easier to govern and audit. It may, however, become a bottleneck if a central team gatekeeps every addition, and it may not fit business units with highly distinct data domains. A federated or domain-based arrangement, with separate stores per business line supported by shared standards and a common catalogue layer, offers more autonomy and faster iteration for individual teams. The trade-off is that consistent quality and security controls become harder to enforce, and there is a risk of duplicated effort across domains.

AIs will generally benefit from requiring teams to make an explicit, documented choice on each of these dimensions rather than mandating a single architecture. The dimensions include feature store topology, build-versus-buy strategy, model registry design, and the centralisation model. A documented choice allows the rationale to be justified against the institution's own data maturity, latency needs, and governance structure, which can help secure board-level support and align investment with actual risk and reuse patterns.

6.10 Preparing Unstructured and Semi-Structured Data for Analytics

AIs may consider several factors when building a map of their unstructured and semi-structured data assets. The choices span (1) system architecture, (2) discovery methods, and (3) classification techniques. Each option carries a distinct set of trade-offs that may be evaluated against the AI's existing technology landscape and operating model.

In terms of system architecture, AIs may consider whether to adopt (1) a centralised catalogue, (2) a federated approach, or (3) a data virtualisation layer. A centralised catalogue ingests metadata from across the organisation into a single repository, which can simplify governance and create a unified audit trail, though it may require significant upfront integration work and can struggle to keep pace with rapidly changing source systems. A federated model, sometimes described as a data mesh approach, leaves metadata indexed in place under the stewardship of individual business domains, which can scale more easily across siloed organisations and keeps ownership close to subject matter experts, but it demands robust federated governance to maintain consistent metadata quality. A data virtualisation layer avoids data duplication by querying source systems directly, which may reduce storage costs and provide real-time access; however, query performance can be a concern and not all unstructured data types, such as certain proprietary document formats or streaming media, virtualise effectively.

Discovery mechanisms, either (1) automated or (2) manual, also shape the catalogue's coverage and reliability. Automated crawlers and connectors can scan repositories to extract technical metadata (e.g., format, size, and location) with minimal manual effort; however, doing so requires significant input and governance to ensure its accuracy. Manual registration relies on data owners to self-declare data assets, which can improve the richness of business context but may result in incomplete coverage if owners are not consistently engaged. A hybrid approach, where automated scanning surfaces candidate assets and business or data owners then confirm ownership, sensitivity, and permitted use, often balances efficiency with accountability.

Once data assets are identified, classification and tagging frameworks, whether rule-based, machine learning-based, or with human-in-the-loop, also help determine the data's sensitivity and analytical value. Rule-based pattern matching, such as regular expressions for personally identifiable information, offers transparency and predictability but can miss context-dependent sensitivities. On the other hand, machine learning-based auto-classification can surface content-based sensitivity or topic tags at scale, though model outputs may require periodic validation to guard against drift or bias. For datasets that are particularly sensitive or ambiguous, human-in-the-loop validation remains an appropriate safeguard, allowing subject matter experts to review and override automated classifications.

In particular, selecting metadata standards involves a trade-off between interoperability and internal relevance. Adopting open standards (e.g., Dublin Core) can facilitate tool compatibility and future data sharing, whereas proprietary taxonomies may reflect internal business use cases more accurately but risk locking the AI into specific vendor ecosystems or complicating external data exchanges. To balance these considerations, AIs may consider applying internal extensions to an open base standard.

6.11 Access Roles

To manage more complex access scenarios, AIs may explore attribute-based access control ("ABAC")[15] to complement role-based access control ("RBAC")[16] models. The former can provide greater flexibility by evaluating contextual factors (e.g., role, location, device security posture, time of access, purpose of use, sensitivity level, etc.) at the time of each request. This makes access decisions more dynamic and risk-sensitive, rather than relying solely on permissions granted earlier.

As a complementary good practice, AIs may also consider implementing both row-level and column-level RBAC to refine data-level permissions, where row-based restrictions limit access to specific records based on user attributes, and column-based restrictions mask or deny access to particular fields within those records.

For particularly sensitive systems and datasets, AIs may consider implementing just-in-time ("JIT") access models[17] that grant time-bound permissions for specific tasks, with access automatically expiring once the approved task or access window has closed.

[15] "An access approach in which access is mediated based on attributes associated with subjects (requesters), such as location, time of creation, access rights, etc. Access to an object is authorised or denied depending upon whether the required (e.g., policy-defined) correlation can be made between the attributes of that object and of the requesting subject." National Institute of Standards and Technology (U.S. Department of Commerce). Accessed 2026. Glossary. https://csrc.nist.gov/glossary/term/attribute_based_access_control.

[16] "Access control based on user roles (i.e., a collection of access authorisations a user receives based on an explicit or implicit assumption of a given role). Role permissions may be inherited through a role hierarchy and typically reflect the permissions needed to perform defined functions within an organisation. A given role may apply to a single individual or to several individuals." National Institute of Standards and Technology (U.S. Department of Commerce). Accessed 2026. Glossary. https://csrc.nist.gov/glossary/term/role_based_access_control.

[17] "An access control approach that grants time-limited, task-specific privileged permissions to a human or non-human identity only when needed, and revokes those privileges immediately after the work is done." Palo Alto Networks. Accessed 2026. What Is Just-in-Time (JIT) Access?. https://www.paloaltonetworks.com/cyberpedia/what-is-just-in-time-access-jit.

6.12 List of Abbreviations

- **A.I.:** Artificial intelligence
- **ABAC:** Attribute-based Access Control
- **ACID:** Atomicity, Consistency, Isolation, Durability
- **AI:** Authorized Institution
- **AML:** Anti-money Laundering
- **APIs:** Application Programming Interfaces
- **BI:** Business Intelligence
- **CDC:** Change Data Capture
- **CDEs:** Critical Data Elements
- **CI/CD:** Continuous Integration and Continuous Deployment
- **DLT:** Distributed Ledger Technologies
- **ELT:** Extract, Load, Transform
- **ETL:** Extract, Transform, Load
- **FPS:** Faster Payment System
- **GenA.I.:** Generative Artificial Intelligence
- **GPU:** Graphics Processing Unit
- **HKMA:** Hong Kong Monetary Authority
- **HR:** Human Resources
- **IT:** Information Technology
- **JIT:** Just-in-time
- **KYC:** Know-your-customer
- **LEIs:** Legal Entity Identifiers
- **LLM:** Large Language Model
- **MDM:** Master Data Management
- **ML:** Machine Learning
- **NLP:** Natural Language Processing
- **OCR:** Optical Character Recognition
- **OLA:** Operational-level Agreement
- **RBAC:** Role-based Access Control
- **SMEs:** Subject Matter Experts
- **SQL:** Structured Query Language
- **SSO:** Single Sign-On

6.13 Acknowledgement

The HKMA would like to express its appreciation to KPMG China and Quinlan & Associates for sharing their expertise, practices and experience in Data Excellence, which have contributed to the development of this Guide.

The HKMA also gratefully acknowledges the valuable insights, practical perspectives and constructive feedback provided by AIs through the HKMA's ongoing engagement and consultative sessions. These contributions have helped ensure that this Guide reflects the evolving data-related priorities, key considerations and practices of the banking sector.