Qlik

WHITEPAPER

DATA QUALITY AND GOVERNANCE

The Six Principles of AI-Ready Data

Establishing a trusted data foundation for AI

Table of Contents

Executive summary	3
Introduction	3
The six principles of AI-ready data	4
The AI Trust score	12
Qlik Talend data foundation for AI	13
Conclusion	14

Executive summary

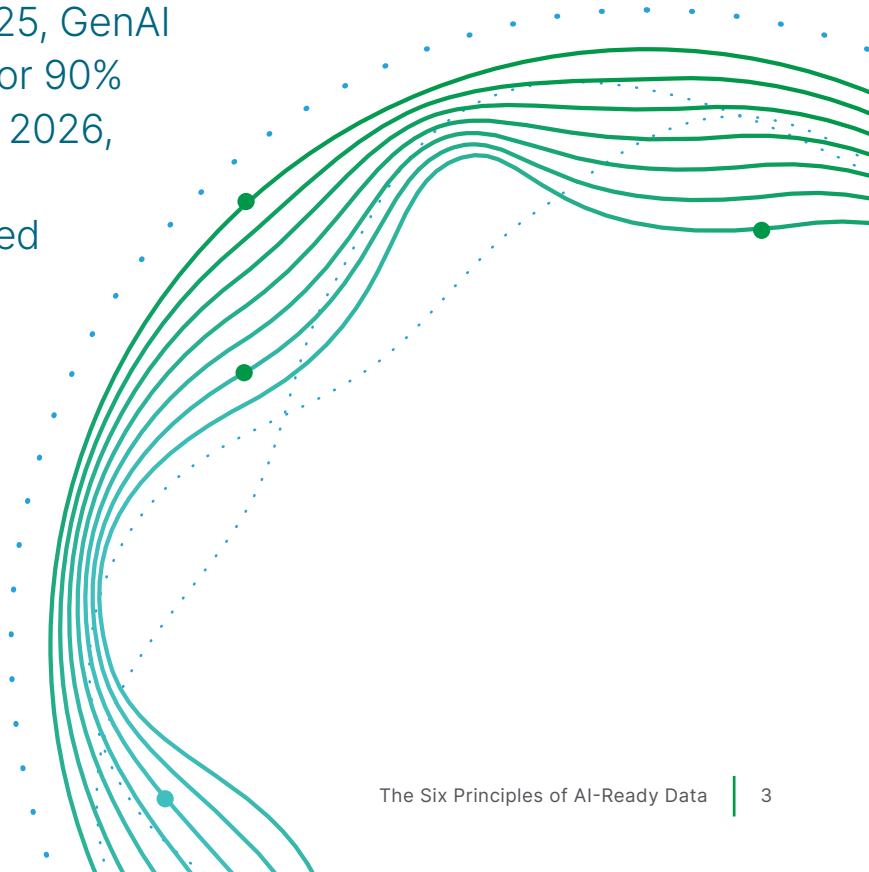
- This document outlines six principles for ensuring that data is ready for use with artificial intelligence (AI).
- These principles are as follows: data has to be diverse, timely, accurate, secure, discoverable, and easily consumable by machines.
- The document also describes the AI Trust Score, which helps assess how well your data adheres to these principles.

Introduction

Artificial intelligence (AI) is expected to greatly improve industries like healthcare, manufacturing, and customer service, leading to higher-quality experiences for customers and employees alike. Indeed, AI technologies like machine learning (ML) have already helped data practitioners produce mathematical predictions, generate insights, and improve decision-making. Furthermore, emerging AI technologies like generative AI (GenAI) can create strikingly realistic content that has the potential to enhance productivity in virtually every aspect of business.

According to Gartner, by 2025, GenAI will be a workforce partner for 90% of global companies, and by 2026, over 80% of enterprises will have deployed GenAI-enabled applications in production.[1]

However, AI can't succeed without good data, and this paper describes six principles for ensuring your data is AI-ready.



[1] "Analysts to Discuss Generative AI Trends and Technologies," Gartner, October, 2023.

The six principles for AI-ready data

It would be foolish to believe that you could just throw data at various AI initiatives and expect magic to happen, but that's what many practitioners do. While this approach might seem to work for the first few AI projects, data scientists increasingly spend more time correcting and preparing the data as projects mature.

Additionally, data used for AI has to be high-quality and precisely prepared for these intelligent applications. This means spending many hours manually cleaning and enhancing the data to ensure accuracy and completeness, and organizing it in a way that machines can easily understand. Also, this data often requires extra information — like definitions and labels — to enrich semantic meaning for automated learning and to help AI perform tasks more effectively.

Therefore, the sooner data can be prepared for downstream AI processes, the greater the benefit. Using prepped, AI-ready data is like giving a chef pre-washed and chopped vegetables instead of a whole sack of groceries — it saves effort and time and helps ensure that the final dish is promptly delivered. The diagram below defines six critical principles for ensuring the "readiness" of data and its suitability for AI use.

The remaining sections of this paper discuss each principle in detail.

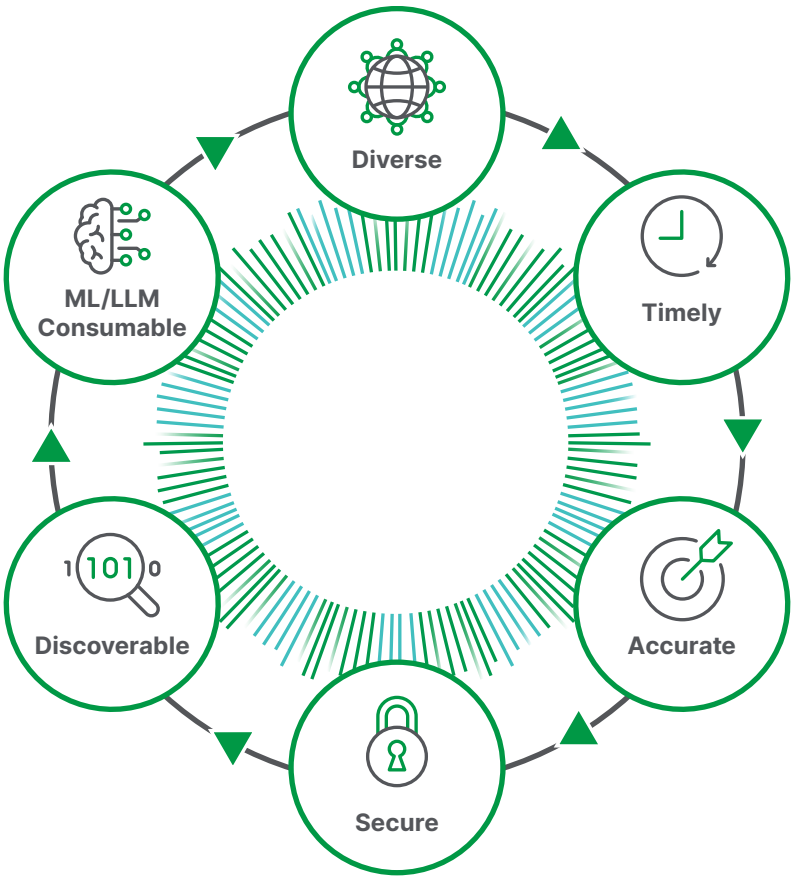


Figure 1. Six Principles of AI-Ready Data

1 Data has to be diverse.

Bias in AI systems, also known as machine learning or algorithm bias, occurs when AI applications produce results reflecting human biases, such as social inequality. This can happen when the algorithm development process includes prejudicial assumptions or, more commonly, when the training data has bias. For example, a credit score algorithm may deny a loan if it consistently uses a narrow band of financial attributes.

Consequently, our first principle focuses on providing a wide variety of data to AI models, which increases data diversity and reduces bias, helping to ensure that AI applications are less likely to make unfair decisions.

Diverse data means you don't build your AI models on narrow and siloed datasets. Instead, you draw from a wide range of data sources spanning different patterns, perspectives, variations, and scenarios relevant to the problem domain. This data could be well-structured and live in the cloud or on-premises. It could also exist on a mainframe, database, SAP system, or software as a service (SaaS) application. Conversely, the source data could be unstructured and live as files or documents on a corporate drive.

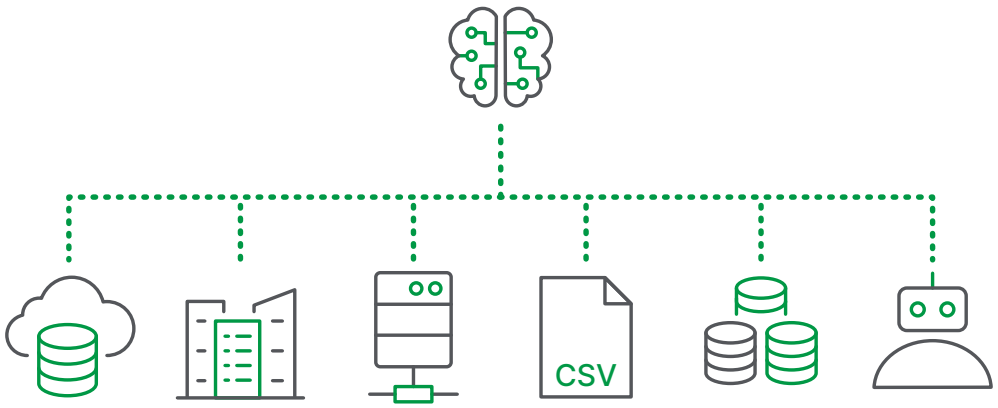


Figure 2. Data Diversity

It's essential to acquire diverse data in various forms for integration into your ML and GenAI applications.

2 Data has to be timely.

While it's true that ML and GenAI applications thrive on diverse data, the freshness of that data is also crucial. Just as a weather forecast based on yesterday's conditions isn't conducive for a trip you plan to take today, AI models trained on outdated information can produce inaccurate or irrelevant results. Moreover, fresh data allows AI models to stay current with trends, adapt to changing circumstances, and deliver the best possible outcomes. Therefore, the second principle of AI-ready data is timeliness.

It's critical that you build and deploy low-latency, real-time data pipelines for your AI initiatives to ensure timely data. **Change data capture (CDC)** is often used to deliver timely data from relational database systems, and **stream capture** is used for data originating from IoT devices that require low-latency processing. Once the data is captured, target repositories are updated and **changes continuously applied** in near-real time for the freshest possible data.

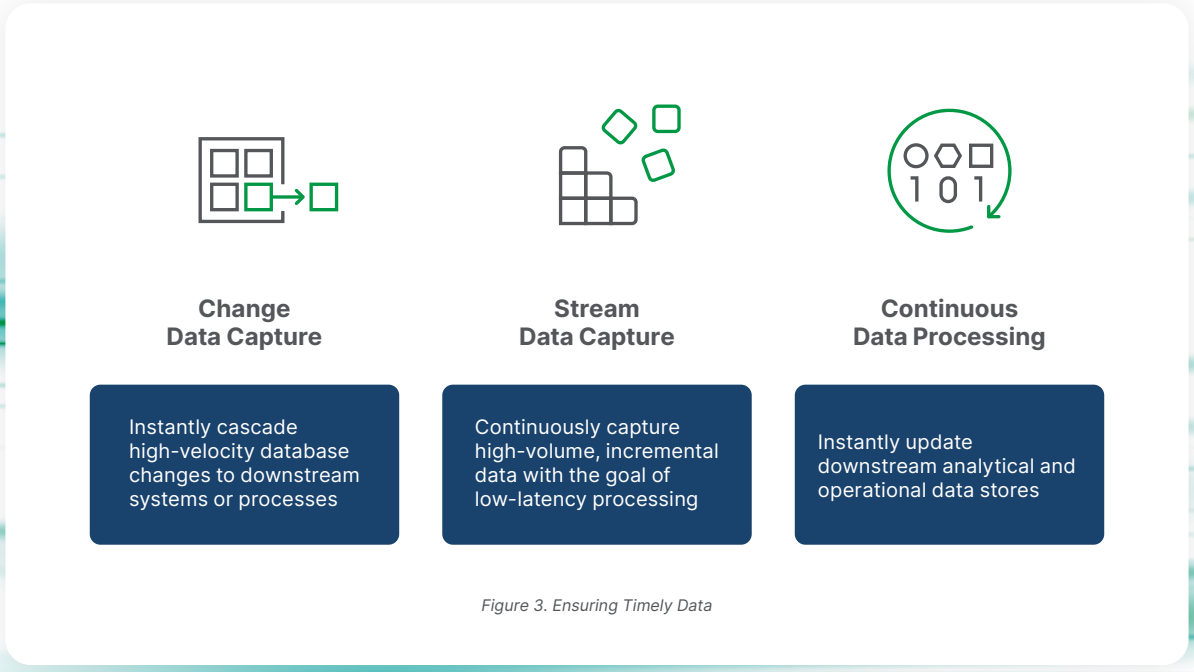


Figure 3. Ensuring Timely Data

Remember, timely data enables more accurate and informed predictions.

3 Data has to be accurate.

The success of any ML or GenAI initiative hinges on one key ingredient: correct data. This is because AI models act like sophisticated sponges that soak up information to learn and perform tasks. If the information is inaccurate, it's like the sponge is soaking up dirty water, leading to biased outputs, nonsensical creations, and, ultimately, a malfunctioning AI system. Therefore, data accuracy is the third principle and a fundamental tenet for building reliable and trustworthy AI applications.

Data accuracy has three aspects. The first is **profiling source data** to understand its characteristics, completeness, distribution, redundancy, and shape. Profiling is also commonly known as exploratory data analysis, or EDA.

The second aspect is **operationalizing remediation strategies** by building, deploying, and continually monitoring the efficacy of data quality rules. Your data stewards may need to be involved here to aid with data deduplication and merging. Alternatively, AI can help automate and accelerate the process through machine-recommended data quality suggestions.

The final aspect is enabling data lineage and impact analysis — with tools for data engineers and scientists that highlight the impact of potential data changes and trace the origin of data to prevent accidental modification of the data used by AI models.

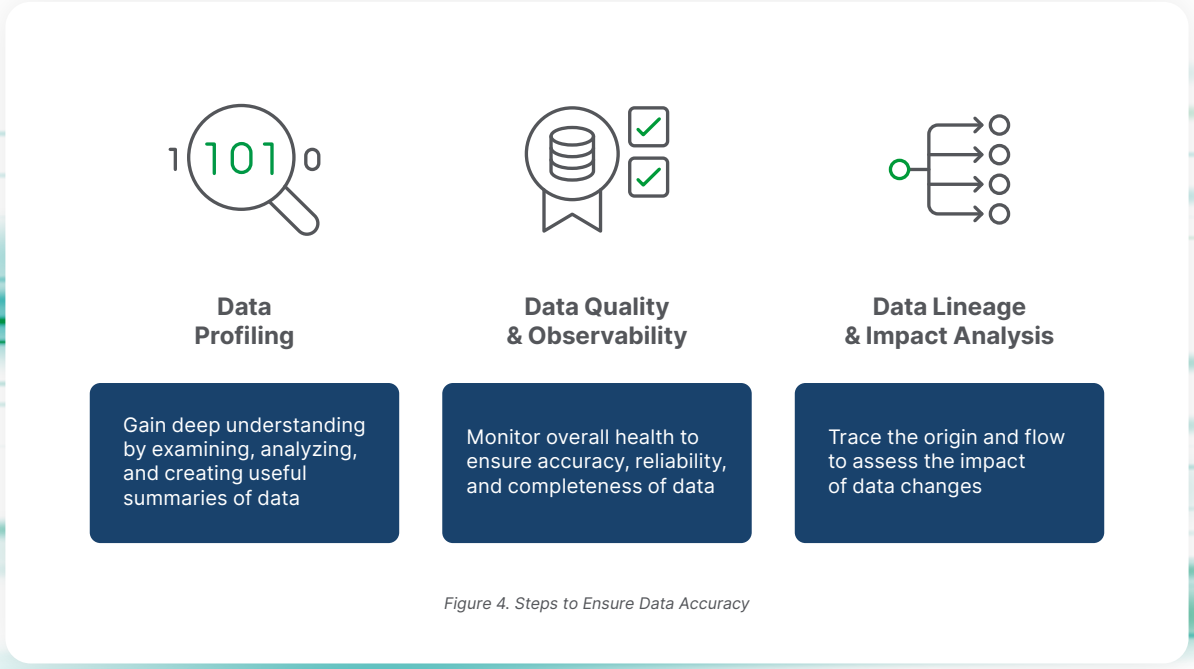


Figure 4. Steps to Ensure Data Accuracy

High-quality, accurate data ensures that models can identify relevant patterns and relationships, leading to more precise decisions, generation, and predictions.

4 Data has to be secure.

AI systems often use sensitive data — including personally identifiable information (PII), financial records, or proprietary business information — and use of this data requires responsibility. Leaving data unsecured in AI applications is like leaving a vault door wide open. Malicious actors could steal sensitive information, manipulate training data to bias outcomes, or even disrupt entire GenAI systems. Securing data is paramount to protecting privacy, maintaining model integrity, and ensuring the responsible development of powerful AI applications. Therefore, data security is the fourth AI-ready principle.

Again, three tactics can help you automate data security at scale, since it's nearly impossible to do it manually. **Data classification** detects, categorizes, and labels data that feeds the next stage. **Data protection** defines policies like masking, tokenization, and encryption to obfuscate the data. Finally, **data security** defines policies that describe access control, i.e., who can access the data. The three concepts work together as follows: first, privacy tiers should be defined and data tagged with a security designation of *sensitive*, *confidential*, or *restricted*. Next, a protection policy should be applied to mask restricted data. Finally, an access control policy should be used to limit access rights.

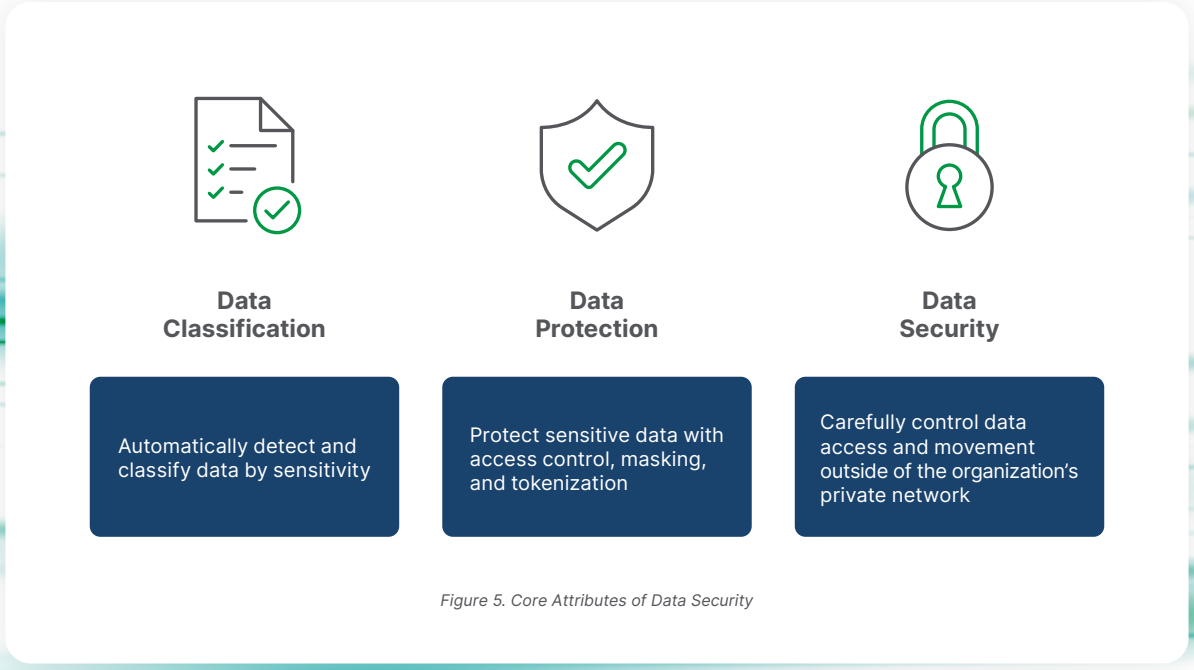


Figure 5. Core Attributes of Data Security

These three tactics protect your data and are crucial for improving the overall trust in your AI system and safeguarding its reputational value.

5 Data has to be discoverable.

The principles we've discussed so far have primarily focused on promptly delivering the right data, in the correct format, to the right people, systems, or AI applications. But stockpiling data isn't enough. AI-ready data has to be discoverable and readily accessible within the system. Imagine a library with all the books locked away — the knowledge is there but unusable. Discoverable data unlocks the true potential of ML and GenAI, allowing these workloads to find the information they need to learn, adapt, and produce groundbreaking results. Therefore, discoverability is the fifth principle of AI-ready data.

Unsurprisingly, good metadata practices lie at the center of discoverability. Aside from the technical metadata associated with AI datasets, business metadata and semantic typing must also be defined. **Semantic typing** provides extra meaning for automated systems, while additional business terms deliver extra context to aid human understanding. A best practice is to create a **business glossary** that maps business terms to technical items in the datasets, ensuring a common understanding of concepts. AI-assisted augmentation can also be used to automatically generate documentation and add business semantics from the glossary. Finally, all the metadata is indexed and made searchable via a **data catalog.**

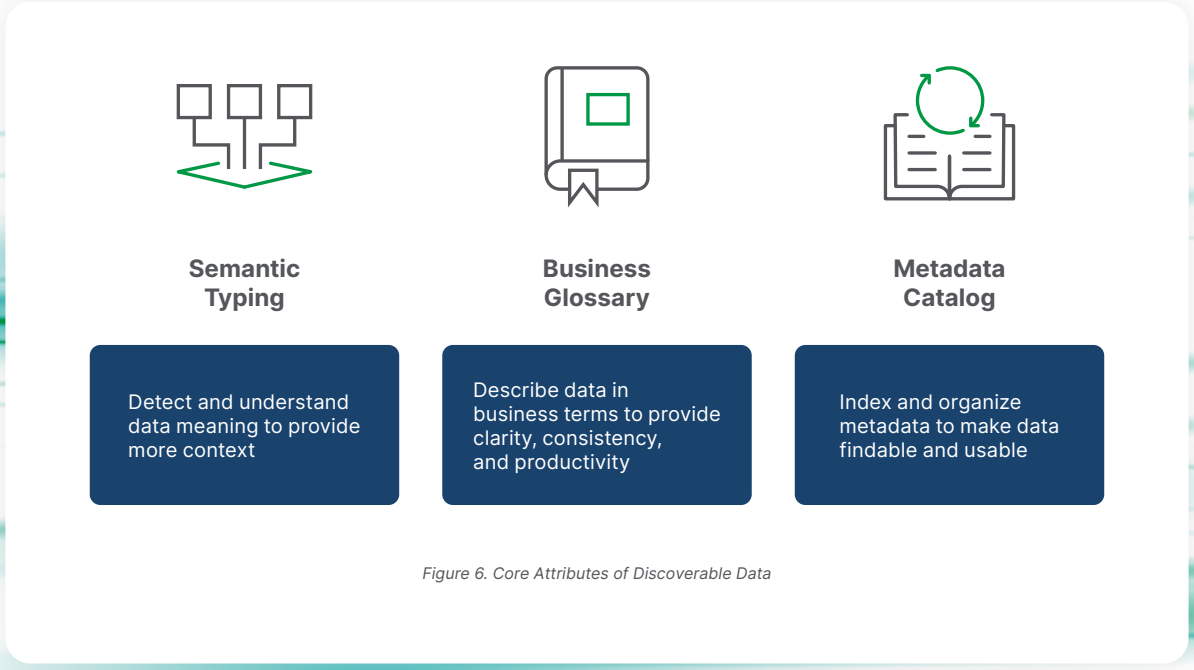


Figure 6. Core Attributes of Discoverable Data

This approach ensures that the data is directly discoverable, applicable, practical, and significant to the AI task at hand.

6 Data has to be easily consumable by MLs or LLMs.

We've already mentioned that ML and GenAI applications are mighty tools, but their potential rests on the ability to readily consume data. Unlike humans, who can decipher handwritten notes or navigate messy spreadsheets, these technologies require information to be represented in specific formats. Imagine feeding a picky eater — if they won't eat what you're serving, they'll go hungry. Similarly, AI initiatives won't be successful if the data is not in the right format for ML experiments or LLM applications. Making data easily consumable helps unlock the potential of these AI systems, allowing them to ingest information smoothly and translate it into intelligent actions for creative outputs. Consequently, making data readily consumable is the final principle of AI-ready data.

Making Data Consumable for Machine Learning

Data transformation is the unsung hero of consumable data for ML. While algorithms like linear regression grab the spotlight, the quality and shape of the data they're trained on are just as critical. Moreover, the effort invested in cleaning, organizing, and making data consumable by ML models reaps significant rewards. Prepared data empowers models to learn effectively, leading to accurate predictions, reliable outputs, and, ultimately, the success of the entire ML project.

However, the training data formats depend highly on the underlying ML infrastructure. Traditional ML systems are disk-based, and much of the data scientist workflow focuses on establishing best practices and manual coding procedures for handling large volumes of files. More recently, lakehouse-based ML systems have used a database-like feature store, and the data scientist workflow has transitioned to SQL as a first-class language. As a result, well-formed, high-quality, tabular data structures are the most consumable and convenient data format for ML systems.

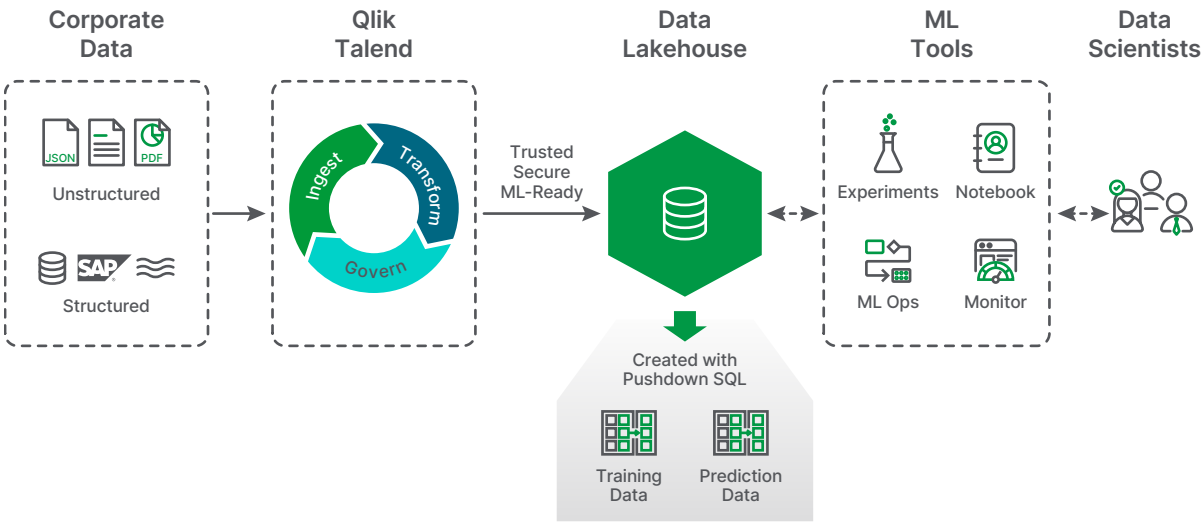


Figure 7. Making Data Consumable for ML

Making Data Consumable for Generative AI

Large language models (LLMs) like OpenAI's GPT-4, Anthropic's Claude, and Google AI's LaMDA and Gemini have been pre-trained on masses of text data and lie at the heart of GenAI. OpenAI's GPT-3 model was estimated to be trained with approximately 45 TB of data, exceeding 300 billion tokens. Despite this wealth of inputs, LLMs can't answer specific questions about your business, because they don't have access to your company's data. The solution is to augment these models with your own information, resulting in more correct, relevant, and trustworthy AI applications.

The method for integrating your corporate data into an LLM-based application is called retrieval-augmented generation, or RAG. The technique generally uses text information derived from unstructured, file-based sources such as presentations, mail archives, text documents, PDFs, transcripts, etc. The text is then split into manageable chunks and converted into a numerical representation used by the LLM in a process known as embedding. These embeddings are then stored in a vector database like Chroma, Pinecone, and Weviate. Interestingly, many traditional database vendors — such as PostgreSQL, Redis, and SingleStoreDB — also support vectors. Moreover, cloud platforms like Databricks, Snowflake, and Google BigQuery have recently added support for vectors, too.

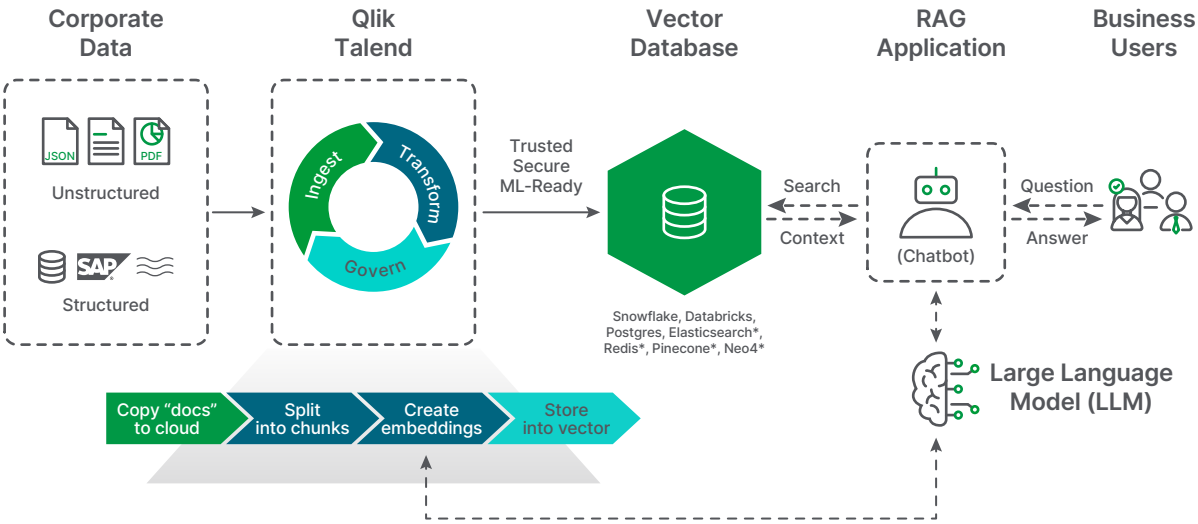


Figure 8. Making Data Consumable for GenAI

Whether your source data is structured or unstructured, Qlik's approach ensures that quality data is readily consumable for your GenAI, RAG, or LLM-based applications.

The AI Trust Score

Having defined the six core principles of data readiness and suitability, the questions remain: can the principles be codified and easily translated for everyday use? And how can the readiness for AI be quickly discerned? One possibility is to use Qlik's AI Trust Score as a global and understandable readiness indicator.

The AI Trust Score assigns a separate dimension for each principle and then aggregates each value to create a composite score, a quick, reliable shortcut to assessing your data's AI readiness. Additionally, because enterprise data continually changes, the trust score is regularly checked and frequently recalibrated to track data readiness trends.

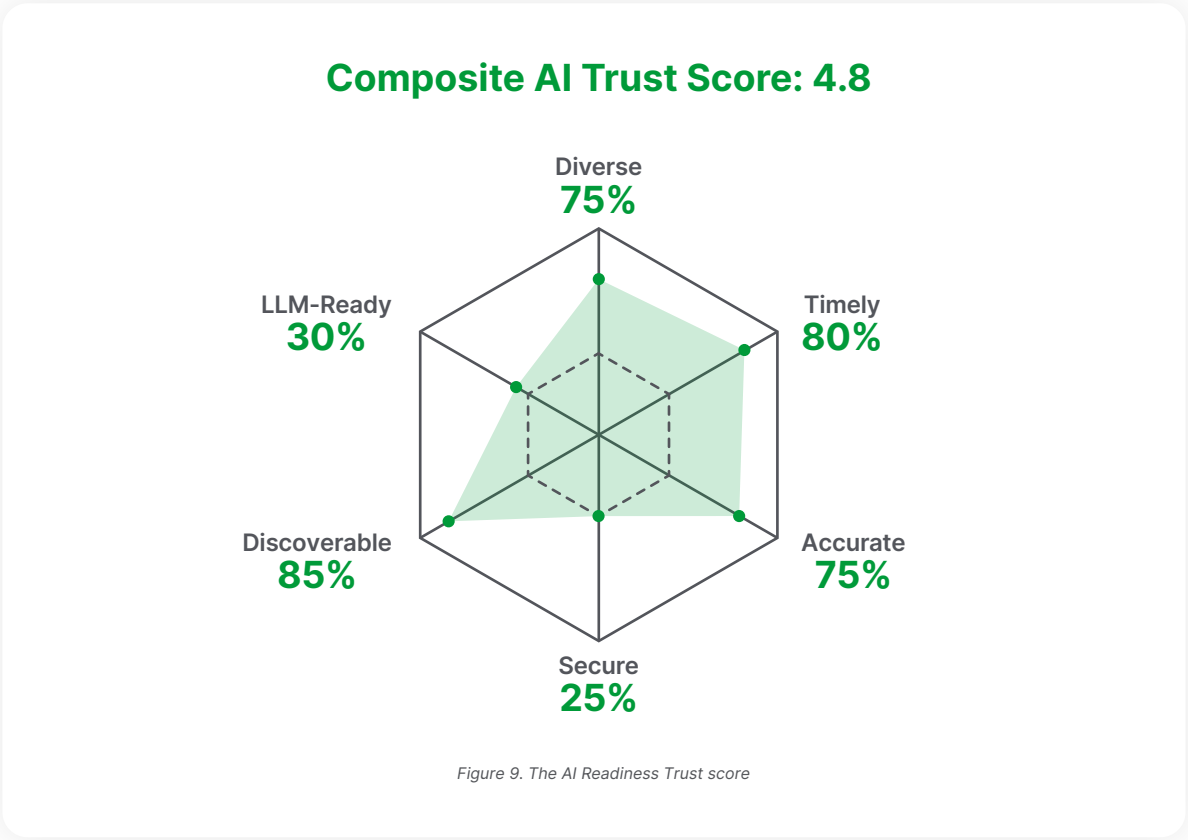


Figure 9. The AI Readiness Trust score

Your AI Trust Score aggregates multiple metrics into a single, easy-to-understand readiness Score.

Qlik Talend data foundation for AI

The need for high-quality, real-time data that drives more thoughtful decisions, operational efficiency, and business innovation has never been greater. That's why successful organizations seek market-leading data integration and quality solutions from Qlik Talend to efficiently deliver trusted data to warehouses, lakes, and other enterprise data platforms. Our comprehensive, best-in-class offerings use automated pipelines, intelligent transformations, and reliable Datasets quality to provide the agility data professionals crave with the governance and compliance organizations expect.

So, whether you're creating warehouses or lakes for insightful analytics, modernizing operational data infrastructures for business efficiency, or using multi-cloud data for artificial intelligence initiatives, Qlik Talend can show you the way.

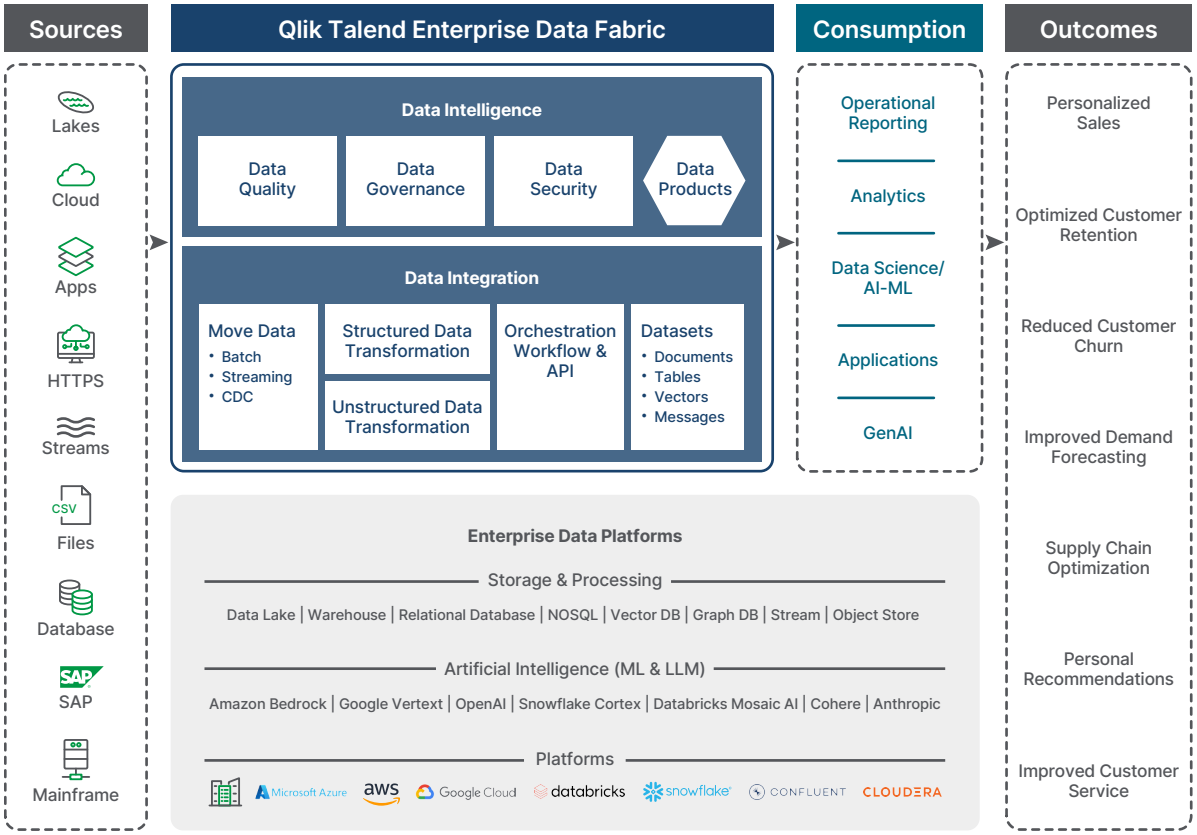


Figure 10. Qlik Talend Enterprise Data Fabric for AI & Analytics

Conclusion

Despite machine learning's transformative power and generative AI's explosive growth potential, data readiness is still the cornerstone of any successful AI implementation. This paper described six key principles for establishing a robust and trusted data foundation that combine to help your organization unlock AI's true potential.

Want to unlock your organization's true AI potential?

Start Here

About Qlik

Qlik transforms complex data landscapes into actionable insights, driving strategic business outcomes. Serving over 40,000 global customers, our portfolio leverages advanced, enterprise-grade AI/ML and pervasive data quality. We excel in data integration and governance, offering comprehensive solutions that work with diverse data sources. Intuitive and real-time analytics from Qlik uncover hidden patterns, empowering teams to address complex challenges and seize new opportunities. Our AI/ML tools, both practical and scalable, lead to better decisions, faster. As strategic partners, our platform-agnostic technology and expertise make our customers more competitive.

qlik.com

24-DQ-0001-01-WHM