Harbr_

ACCELERATING AI ADOPTION THROUGH DATA GOVERNANCE

How a data marketplace enables better data governance and faster AI adoption

White Paper

CONTENTS

ABOUT THE AUTHOR

Anthony Cosgrove

Anthony Cosgrove is the co-founder of Harbr, the data marketplace platform. Previously, Anthony has led global data teams at HSBC.

Introduction

The integration of artificial intelligence (AI) and machine learning (ML) into business strategies is a transformative development, yet it comes with significant challenges related to data governance.

Effective data governance is essential for accelerating AI adoption as it ensures data quality, compliance, and security — all crucial elements for reliable and ethical AI outcomes.

Without appropriate data governance, enterprise implementations of AI introduce risk, experience delays, and ultimately fail. Poor governance undermines the trust and reliability of AI system, leading to:

- inconsistencies in data handling
- increased vulnerability to breaches
- inability to meet regulatory requirements

Additionally, inadequate oversight can result in biased or erroneous AI outputs, damaging the organization's reputation and potentially leading to significant financial and legal repercussions. As businesses increasingly rely on AI to drive decision-making and innovation, the stakes for effective governance have never been higher. Ensuring that AI systems are transparent, accountable, and aligned with ethical standards is not just a technical necessity but a strategic imperative.

This white paper explores two critical aspects of data governance — **governed data access** and **governed data usage** — and how these relate to controlled deployment of AI algorithms and large language models (LLMs).

A robust data marketplace can support these governance objectives. A data marketplace provides a safe and scalable environment for successful AI deployment. Proper governance enabled by a data marketplace will mitigate risks, facilitate smoother AI integration, and ultimately foster innovation and competitive advantage.

Artificial intelligence and machine learning (AI/ML)

An AI or machine learning model processes and analyzes data to recognize patterns, make predictions, and generate insights. These models are trained on large datasets using various algorithms that allow them to learn from the data and improve their performance over time. They can be applied to a wide range of tasks such as classification, regression, clustering, and anomaly detection. By identifying relationships within the data, these models enable businesses to automate decision-making processes, optimize operations, and uncover insights.

Large language models (LLMs)

A large language model (LLM) is a type of AI designed to understand and generate human language. Trained on vast amounts of text data, LLMs can comprehend context, syntax, and semantics. These models perform a variety of language-related tasks, including text generation, translation, summarization, interpreting code, and answering questions.

Data marketplace

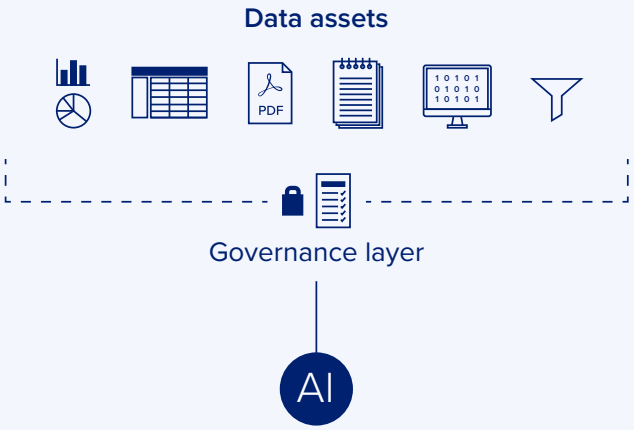
A data marketplace is an online platform designed to facilitate the access, usage, and distribution of data in a secure and governed way. The best private (or enterprise) data marketplaces enable this via user experiences that support the full range of data consumers. A data marketplace leads to quicker access to data, mitigates data governance risks, shortens time to value, and accelerates business outcomes.

The importance of data governance to AI adoption

Governed data access

Governed data access is the cornerstone of secure and compliant AI adoption. It ensures that only authorized users and systems can access specific datasets, thus maintaining the integrity and confidentiality of the data. By implementing role-based access controls, organizations can define precise permissions that align with users' roles and responsibilities. This approach not only protects sensitive information from unauthorized access but also ensures compliance with various regulatory requirements, such as GDPR, HIPAA, and CCPA.

Furthermore, governed data access contributes to maintaining high data quality. By controlling who can add, modify, or delete data, organizations can prevent data corruption and ensure that only verified and accurate data is used in AI training and analysis. Comprehensive audit trails provide visibility into data access activities, facilitating easy audits and compliance checks, thereby building trust and accountability.



Governed access to data assets: *The governance layer controls which data assets an AI or LLM has access to. Data assets can include tables, data science notebooks, visualizations, PDFs, and more.*

Governed data usage

While a system to govern data access will control *who* can view the data, governed data usage dictates *how* the data can be used. This distinction is crucial for ethical considerations as well as for driving consistent data practices across an organization. By defining and enforcing precise usage policies, organizations can ensure data is used in line with legal, ethical, and regulatory standards and requirements. This prevents misuse and ensures that AI models are developed and used responsibly without creating risk.

Governed data usage involves setting terms and conditions, often formalized through data contracts or subscription plans. These contracts and subscriptions specify how data can be used, shared, and processed, providing clarity and accountability. They can also be used by tools and systems to enforce and track data usage, ensuring compliance or identifying potential risks and issues so that corrective action can be taken.

Governed data usage also supports consistency across the organization. When all departments adhere to the same data usage policies, these become easier to understand, and the risk of data misuse is minimized. Consistency is also important when developing and using AI models, as it ensures that models are trained on data that meets the organization's standards for quality and intended use and the models are used in the appropriate way for a given use case. Due to the nature of some AI models, such as LLMs, consistent data and models may not be enough to drive consistent outcomes — there's also a need for consistent prompts.

Control over AI models

Controlling which AI model can be used with a given dataset is another critical aspect of data governance. This control ensures that AI models are trained on appropriate data, enhancing their accuracy and effectiveness. By whitelisting specific algorithms, or specifying which algorithm can be used for which use case, organizations can prevent the development of unsuitable or biased models and the use of an inappropriate model for a given use case, thereby eliminating poor, unreliable outcomes driven by the use of AI.

This control also helps mitigate the risk of bias in AI models. By carefully selecting the datasets used for training, organizations can ensure that their AI systems make fair and unbiased decisions. Additionally, controlling AI algorithms in this way and auditing their usage supports regulatory compliance and may provide some level of future-proofing against future laws and regulation. Many industries already have strict regulations regarding the use of AI, and this is an area that is likely to continue developing. Therefore, organizations best suited to take advantage of AI should have in place strict data governance rules over access, usage, and models.

Using a data marketplace to govern AI

A robust data marketplace enables governed data access, governed data usage, and control over which AI algorithms can be used. As such, they provide a safe and scalable way to enable usage of AI and LLMs. A good data marketplace should offer comprehensive integration capabilities connecting to existing systems, databases, and catalogs when creating data products. They should also integrate with a wide range of third party tools for consumers to access and use the data contained in data products. This interoperability, coupled with tight governance over access and use, enables data marketplaces to be the point of specifying, enabling and monitoring how data can be used with AI and LLMs.

Direct integration

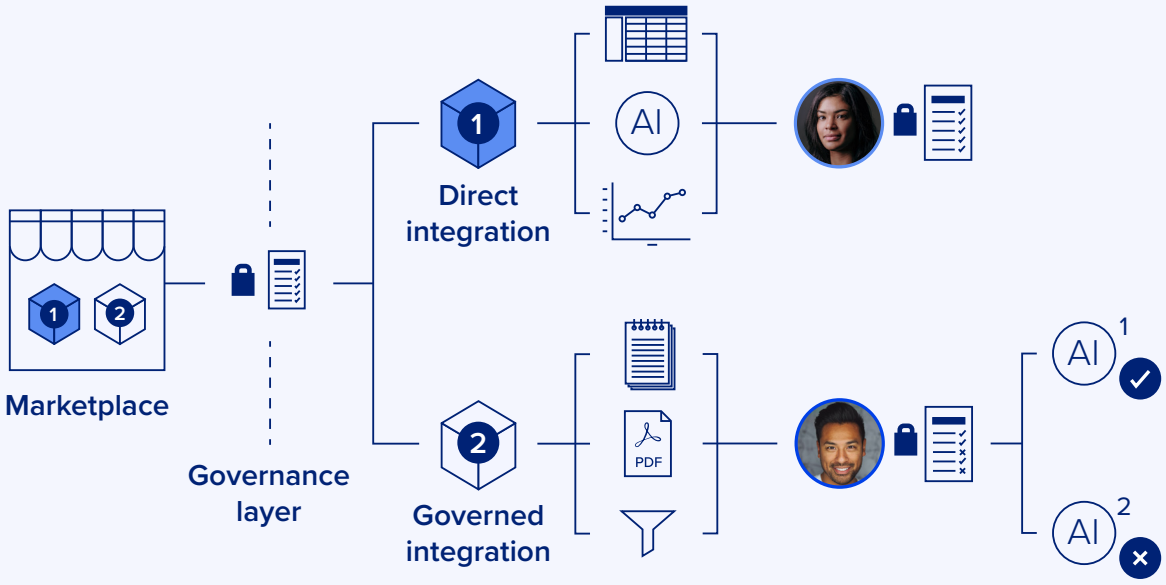
One approach is to integrate AI models directly into data products, enhancing their value by combining them with diverse data assets such as files, tables, and notebooks. This integration allows for the creation of comprehensive data products that not only include raw and processed data but also advanced analytical capabilities. For instance, a data product might feature an AI model designed to predict customer behavior, supported by extensive datasets stored in tables, supplementary documents in files, and interactive notebooks that provide detailed analysis and visualization. This holistic approach enables users to access and use data and AI insights in a unified environment, driving more informed decision-making and enabling advanced data-driven solutions.

Governed integration

An alternative approach is to treat AI models and LLMs as generic capabilities available via the marketplace and governed by the subscription-based permissions. This method allows organizations to flexibly apply AI tools to various data products while maintaining stringent control over their use.

Subscription-based permissions specify whether certain datasets can interact with particular AI models, ensuring compliance and adhering to ethical standards. For instance, a data owner might want to allow a language model to be used for sentiment analysis of customer feedback data, while restricting any access to sensitive financial records.

This approach provides a scalable, secure framework, enabling the strategic use of AI tools across the enterprise while safeguarding data integrity and compliance.



Direct vs. governed integration: *The data products are accessed in a data marketplace. Data product 1 features an AI model that is directly integrated into the data product. Data product 2 follows the governed integration model, where the AI models are not part of the data product itself, but rather a capability available via the data marketplace. In both cases, the end users are granted permission based on their role, use case, and type of subscription.*

Benefits

Regardless of how it's configured, the ability of a data marketplace to support governed data access and usage, alongside control over AI algorithms provides numerous benefits.

1. It enhances security and compliance by ensuring that data is accessed and used appropriately. This is crucial for building trust with customers and stakeholders and for avoiding regulatory issues.
2. It improves the quality of AI models. By controlling access to high-quality, relevant data and enforcing strict usage policies, organizations can develop more accurate and reliable AI models, leading to better decision-making and business outcomes.
3. It supports ethical and responsible AI development. Governance features ensure that AI models are developed and used ethically in line with specific use cases, reducing the risk of bias and ensuring alignment with organizational values and standards.
4. The scalability to help organizations grow their AI capabilities without compromising on data governance. This flexibility is essential for keeping pace with the rapidly evolving AI landscape and maintaining a competitive edge.

Conclusion

The scalable, sustainable, and secure adoption of AI requires strong data governance. Accelerating it, requires even more stringent measures - robust governance over data access, data usage, and the interactions between data and AI models or LLMs. A data marketplace that provides these capabilities can effectively broker the interaction between data and models, creating a secure and scalable environment to test, deploy and manage AI. By adopting a marketplace with these capabilities, organizations can build a trusted foundation for their AI initiatives, rapidly enhance their AI capabilities, and ultimately drive better business outcomes.

TAKE THE NEXT STEP

To continue on your data governance and AI journey, the best thing to do is speak to people who are on the same path. Harbr works with some of the world's biggest and most innovative data teams to deploy data marketplaces at scale — and governance is at the heart of what we do. If you'd like a free consultation to help you get to the next level, book in a chat with us today. We'd be delighted to help.

Learn more at harbrdata.com

Get in touch →