WHITEPAPER

The Growing Importance of Data Governance with Generative AI

How data governance impacts the success of your GenAI initiatives

Gleecus TECHLABS INC

Forward Thinking Digital Innovation

Table of Content

Executive Summary

The management of data in Generative AI (GenAI) presents unique challenges. GenAI models rely on complex algorithms, and as their complexity grows, explainability diminishes. This opaqueness leads to unique challenges related to data acquisition, usage, retention, and deletion. Key challenges include hidden security risks from sensitive data inadvertently included in training sets, irregular user interfaces that can expose confidential information through natural language prompts, and the inability to forget sensitive information if required by invoking a simple delete function. This underscores the necessity of a robust and specialized data governance framework for GenAI models.

To facilitate a robust data governance, organizations can leverage various tools for data cataloging, data lineage tracking, and data quality maintenance. These tools help automate data discovery, classification, and compliance tracking to simplify GenAI's complexities. Ultimately, a comprehensive approach to data governance will enhance trust, foster ethical AI practices, and strengthen compliance to enable organizations to harness the full potential of GenAI responsibly and effectively.

This white paper elaborates on the unique challenges related to data in a GenAI model and how to tackle them with a regulating Data Governance Framework. Deep dive to learn more about the importance of data governance framework in a GenAI system and the tools and technologies behind maintaining that.

The Unique Challenges Pertaining to Data in LLMs

It is extremely difficult to delete data from LLMs owing to several factors such as the lack of a traditional row delete method, need for complete model retraining, and lack of methods to verify if the data has actually been deleted. As a result, unlearning any information stored within the LLM is difficult. This poses a unique challenge if a user wishes to exercise the Right to Be Forgotten (RTBF) included under the privacy regulations like GDPR and CPRA. This and a few other challenges related to data warrant a special set up of data governance frameworks when dealing with Generative AI (GenAI).

What makes data governance for GenAI challenging?

Following are a few data-related complexities that creep in while training a GenAI model:

Hidden security risks: When a system is being trained on hundreds of terabytes of data, it's easy for some entries containing sensitive information to find their way in. That information will be trained into the AI's neural network. That makes it potentially accessible to users without anyone even recognizing the vulnerability.

Irregular user interfaces: GenAI users don't select from menus. They freely prompt in natural language to a chatbot interface. This flexibility leads to unexpected inputs, such as a user mistakenly inputting private or confidential information. Any records of this input can be a severe compliance threat. The flexibility of prompt-based interactions underpins extra safeguards and audits for UI.

Unexplainability: The algorithms produced by AI training aren't explicitly designed. The mechanics of a GenAI system are not easy to unpack. The opacity of AI systems can make them difficult to trust. Efforts towards explainability help build trust in AI systems.

Expensive testing: Since inputs are flexible, AI system outputs can be chaotic. Imagine the chatbot you trained gives incorrect answers for certain phrasings of a common question. Testing for these failures is prohibitively expensive, so AI systems require consistent monitoring and auditing to stay reliable.

The Role of Data Governance in GenAI Development

Data governance is not just about adhering to regulations; it's about ensuring that data is managed responsibly and efficiently.

Key Components of GenAI Data Governance Strategy

The following are the key components of an effective Data Governance Strategy:

Data Access Control

Specifying who can access specific data and under what circumstances. This involves implementation of the following control strategies:

> **Role-Based Access Control (RBAC):** Implementing RBAC ensures that only authorized personnel can access sensitive information. For example, in a healthcare setting, only doctors and nurses should have access to patient records, while administrative staff may have limited access.

> **Granular Permissions:** As GenAI systems often handle vast amounts of unstructured data, fine-grained access controls are necessary. This means not just controlling who accesses data but also defining what they can do with it (e.g., read, write, train models).

Access controls help organizations comply with regulations such as GDPR and HIPAA by ensuring that sensitive data is only accessible to those with legitimate needs. Automated monitoring tools can track access patterns and flag anomalies for further investigation.

Data Retention

Defining how long data should be retained and when it should be archived or deleted. Retention serves some important purposes in GenAI data management.

> **Regulatory Compliance:** Different industries have specific regulations regarding how long certain types of data should be retained. For instance, healthcare organizations must retain patient records for a specified period to comply with legal requirements.

> **Data Lifecycle Management:** Effective data retention strategies help manage the lifecycle of data used in GenAI projects. This includes determining when to delete outdated training datasets or when to archive older models that are no longer in use.

> **Cost Management:** By implementing clear retention policies, organizations can reduce storage costs associated with keeping unnecessary data. This is particularly important in GenAI projects where large datasets are often involved.

Data Security

Implementing measures to protect data from unauthorized access, corruption, or theft. Data security encompasses the following goals:

> **Encryption and Secure Access:** Sensitive data used in GenAI projects must be encrypted both at rest and in transit to prevent unauthorized access. Implementing secure access protocols ensures that only authorized users can interact with the AI models.

> **Data Loss Prevention (DLP):** DLP solutions monitor and protect sensitive information from being shared or accessed improperly. For instance, if an employee attempts to download sensitive training data onto an unsecured device, DLP tools can prevent this action.

> **Incident Response Plans:** Organizations must have robust incident response plans in place to address potential data breaches or security incidents involving GenAI systems. These plans should include steps for containment, investigation, and remediation.

Data Privacy

Safeguarding sensitive data and using it in compliance with privacy regulations. This is very crucial in a GenAI project considering the sensitive information that is dealt with.

> **Anonymization Techniques:** To protect individual privacy, organizations can use anonymization techniques on datasets used for training AI models. This ensures that personal identifiers are removed while still allowing the model to learn from relevant patterns.

> **User Consent Management:** Organizations must ensure that they obtain explicit consent from individuals before using their data for AI training purposes. This is particularly important in sectors like healthcare where patient consent is critical.

> **Transparency and Accountability:** Establishing clear policies regarding how personal data is collected, used, and shared fosters trust among users. Organizations should be transparent about their data practices related to GenAI applications.

Data governance is not just about adhering to regulations; it's about ensuring that data is managed responsibly and efficiently.

It’s inefficient and very expensive to retrain a private LLM to make sure you delete this data if RTBF is invoked, and with a public LLM you won’t have that option. To address these concerns and comply with an RTBF request, you need effective data governance to keep sensitive data out of LLMs.

Data Quality

Artificial intelligence is only as good as the data that fuels it. Poor data quality can lead to misleading or erroneous insights, seriously affecting the outcomes.

Data quality challenges stem from the volume, velocity, and variety of big data, especially since LLMs now tap into the organization's unstructured data sources. Companies looking to develop internal LLMs will need to extend data quality initiatives to include information extracted from documents, collaboration tools, code repositories, and other tools storing enterprise knowledge and intellectual property.

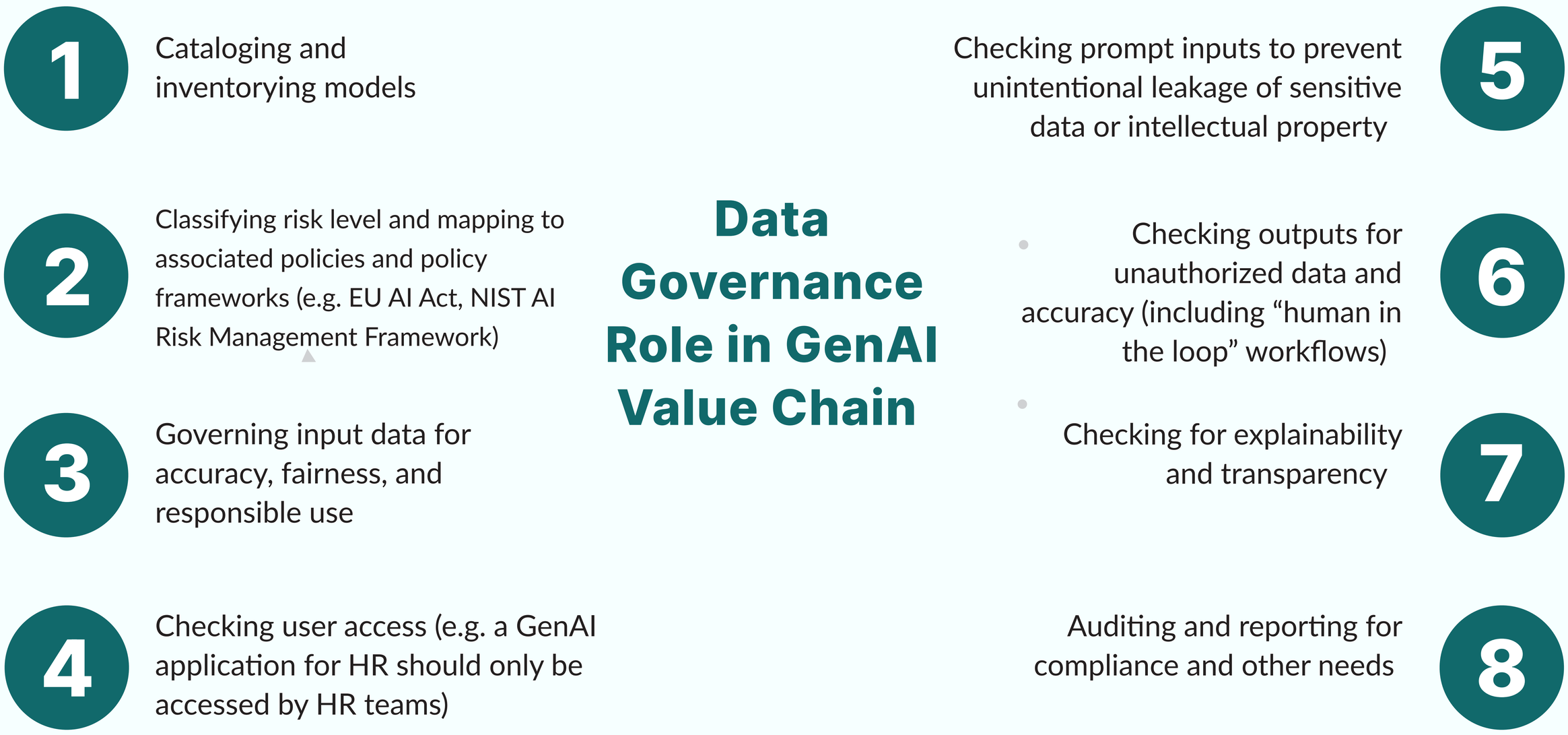
Data quality can be improved using different tools, depending on the business goals and data types.

Data Lineage

Data lineage helps expose the data's lifecycle and answer questions about who, when, where, why, and how data changes. Because AI expands the scope of data and its use cases, understanding data lineage becomes more important to more people in the organization, including people in security and other risk management functions.

Monitoring and Auditing

Regularly monitor and audit the GenAI development process to ensure compliance with data governance policies. This may involve analyzing training data for bias, reviewing model outputs, and conducting security audits.



The Generative AI Data Governance Framework

A foundational framework for GenAI data governance consists of the following:

Data Governance Committee

Establish a cross-functional committee with representatives from IT, data science, legal, and compliance departments. This committee will define data governance policies and oversee their implementation.

Data governance departments oversee data catalogs and communicate data usage policies to help employees tap into centralized data sets and use them for building machine learning models, dashboards, and other analytics tools.

One new, proactive measure data governance leaders should consider is creating prompt libraries where employees can record their prompt use cases and share them across the organizations. This discipline extends the knowledge management practices that many data governance teams already do around maintaining data catalogs and data dictionaries.

Effective data governance involves a multifaceted approach to safeguard sensitive data while allowing organizations to leverage GenAI.

Data Governance Policies

Develop and document data governance policies that address the identified risks. These policies should cover data sourcing, access control, data quality, and human oversight procedures. Some of the common policies defining a data governance framework are the following:

Data Sourcing Policies

Data sourcing policies should align with broader data governance frameworks within the organization. This involves cataloging and inventorying datasets, classifying risk levels, and mapping them to appropriate policies to ensure compliance and responsible use. Organizations must ensure that their data sourcing practices comply with relevant regulations such as GDPR and other privacy laws. Your data sourcing policy should advocate collecting only the essential data needed for your AI models. This approach not only simplifies data management but also reduces the risks associated with unnecessary data accumulation.

Unstructured Data Management

Since unstructured data (like text files) plays a significant role in training large language models (LLMs), your organization needs to develop strong capabilities for discovering, classifying, and managing this type of data. Your policies should focus on cataloging unstructured datasets and linking them with structured data to create a comprehensive view for AI training. Set policies to includes automated checks for accuracy, completeness, and consistency to prevent biases in AI outputs. Create a comprehensive data catalog that includes metadata about the data's origin, sensitivity, and usage context.

Ethical AI Practices

Organizations must establish guidelines that promote fairness, transparency, and accountability in AI usage. This includes regular assessments of AI outputs for bias or ethical concerns and implementing mechanisms for explainability to ensure users understand how decisions are made by AI systems.

Audit Trails and Reporting

Robust audit trails should be maintained to track data access and usage within GenAI applications. This not only supports compliance efforts but also enhances accountability by providing insights into how data is being utilized across different models and applications.

Data Quality Management

The success of GenAI applications heavily relies on the quality of the input data. Organizations should implement policies that ensure data accuracy, completeness, and consistency through processes like data cleansing and validation. This will help mitigate risks associated with biased or inaccurate outputs from AI models.

Data Governance departments are now updating policies to include whether and how to use enterprise data sources in LLMs and open GenAI tools. Developers and data scientists must review these policies and consult with data owners on any questions about using data sets to support GenAI experimentation.

Data Risk Assessment

Conduct a risk assessment to identify potential risks associated with the data used for generative AI training. This assessment should consider factors like bias, fairness, security, and privacy.

Data Governance policies implement the following risk minimization strategies for GenAI-related data management and control:

Data Minimization

Data minimization is the practice of collecting and retaining only the data that is necessary for a specific purpose. By reducing the amount of sensitive personal data in circulation, organizations can significantly reduce the risk of data exposure, easing compliance with data privacy regulations.

Fine-Grained Access Control

Fine-grained access control provides individuals and applications with precisely the level of access needed to perform their tasks, preventing unauthorized access to sensitive data.

Data Anonymization and Pseudonymization

Anonymizing and pseudonymizing data are techniques that can help protect privacy while still leveraging data for AI training. By removing or obfuscating personally identifiable information before it reaches an LLM, organizations can prevent this data from being traced back to individuals, mitigating privacy risks.

Secure Data Sharing Mechanisms

Implementing secure data sharing mechanisms facilitates collaboration and data exchange with trusted third-party services without compromising data privacy. Effective secure data sharing involves the use of encryption, access controls, and encrypted communication channels to share data with trusted partners while maintaining data security.

Data Governance Tools

Utilize data governance tools to automate and streamline data management processes. These tools can include data cataloging systems, access control software, and data lineage tracking tools.

Data governance leaders also review and get involved in these tools:

Data Cataloging Tools

> **Alation Data Catalog:** Alation is recognized for its comprehensive solution that combines AI, machine learning, and automation to facilitate data discovery and governance. Its Behavioral Analysis Engine helps organizations understand data usage patterns, which can inform better decision-making processes in GenAI applications.

> **Informatica Enterprise Data Catalog**: Informatica's solution offers robust metadata intelligence and end-to-end data lineage capabilities. It uses AI-driven automation to discover, classify, and manage data assets across various sources, making it suitable for large-scale Generative AI projects that require comprehensive oversight of data flows.

> **data.world Data Catalog:** data.world stands out for its user-friendly interface and collaborative features. It incorporates GenAI capabilities to facilitate data discovery and management, allowing users to generate natural language descriptions of metadata and convert queries into SQL code easily. This makes it accessible for both technical and non-technical users.

> **Google Cloud Data Catalog (Dataproc Metastore):** This fully managed solution supports both cloud and on-premises data sources, enabling users to search using natural language queries. It integrates seamlessly with other Google Cloud services, making it a strong choice for organizations leveraging cloud infrastructure for their GenAI applications.

Data meshes for delegating the management of the data to those creating it.

Data Mesh

Data Mesh enables domains to customize governance policies while adhering to global standards. Each domain is accountable for the quality of its data products, fostering transparency and trust. This ensures that only high-quality data is used in GenAI models. By establishing clear protocols for data sharing and access control, a Data Mesh enhances interoperability between different data domains.

Vector Databases

Vector databases play a vital role in managing the scalability and complexity inherent in GenAI and large language models (LLMs). Optimized for high-dimensional data, these databases facilitate rapid access to relevant information through advanced indexing techniques like Approximate Nearest Neighbor (ANN) algorithms.

As the complexity of GenAI models increases and data volumes expand, vector databases can seamlessly scale horizontally across multiple nodes or clusters, ensuring efficient performance even under heavy loads. They are specifically designed to handle diverse data types, including images and natural language, which helps preserve the richness and granularity essential for effective model training and output generation.

Additionally, vector databases enable sophisticated semantic search capabilities that surpass traditional keyword-based methods, allowing Generative AI models to retrieve contextually relevant data more effectively. This enhanced retrieval process significantly improves the quality of the outputs generated by these models, making vector databases indispensable in the landscape of Generative AI.

Real-time monitoring tools

Dynatrace is well-known for its AI-powered monitoring and analytics capabilities, making it a top choice for complex IT environments. It provides real-time insights, automated root cause analysis, and predictive analytics, which are essential for maintaining system performance and ensuring effective governance in GenAI applications.

Datadog offers a comprehensive platform for monitoring cloud applications, providing deep insights into both infrastructure and application performance. Its capabilities include real-time monitoring, alerting, and integration with various data sources, making it suitable for managing the complexities associated with GenAI systems.

New Relic is recognized for its robust observability features that leverage AI-driven insights. It provides comprehensive monitoring solutions for applications and infrastructure, helping organizations maintain visibility and control over their Generative AI models and the systems they interact with.

Data Lineage Tracking Tools

Collibra is a comprehensive data governance platform that features automated lineage mapping and maintenance. It provides visibility into data flows across systems, making it a popular choice for organizations looking to improve their data management practices.

MANTA offers end-to-end lineage tracking and impact analysis, integrating with various data platforms. Its focus on automated data mapping makes it a favored tool among organizations aiming to streamline their data governance efforts.

OpenLineage is an open-source initiative that provides a standard for metadata and data lineage collection. It is gaining popularity for its ability to track data across multiple processing systems, making it suitable for complex environments.

Tokern is noted for its specialized approach to cloud data warehouses and its ability to provide column-level lineage from various databases. Its integration capabilities with open-source data catalogs make it a strong contender in the market.

Data Quality Tools

> *The global data quality tools market size was valued at over USD 4 billion in 2022 and is expected to grow 17.7% annually. In fact, higher growth is expected now that many companies are experimenting with AI tools and LLMs.*

Traditional data quality tools can deduplicate, normalize data fields, validate data against business rules, detect anomalies, and compute quality metrics. Following are some tools leveraged to ensure data quality for GenAI model training.

Master data management tools (MDM)

Helps organizations connect multiple data sources and create a source of truth around business entities such as customers and products.

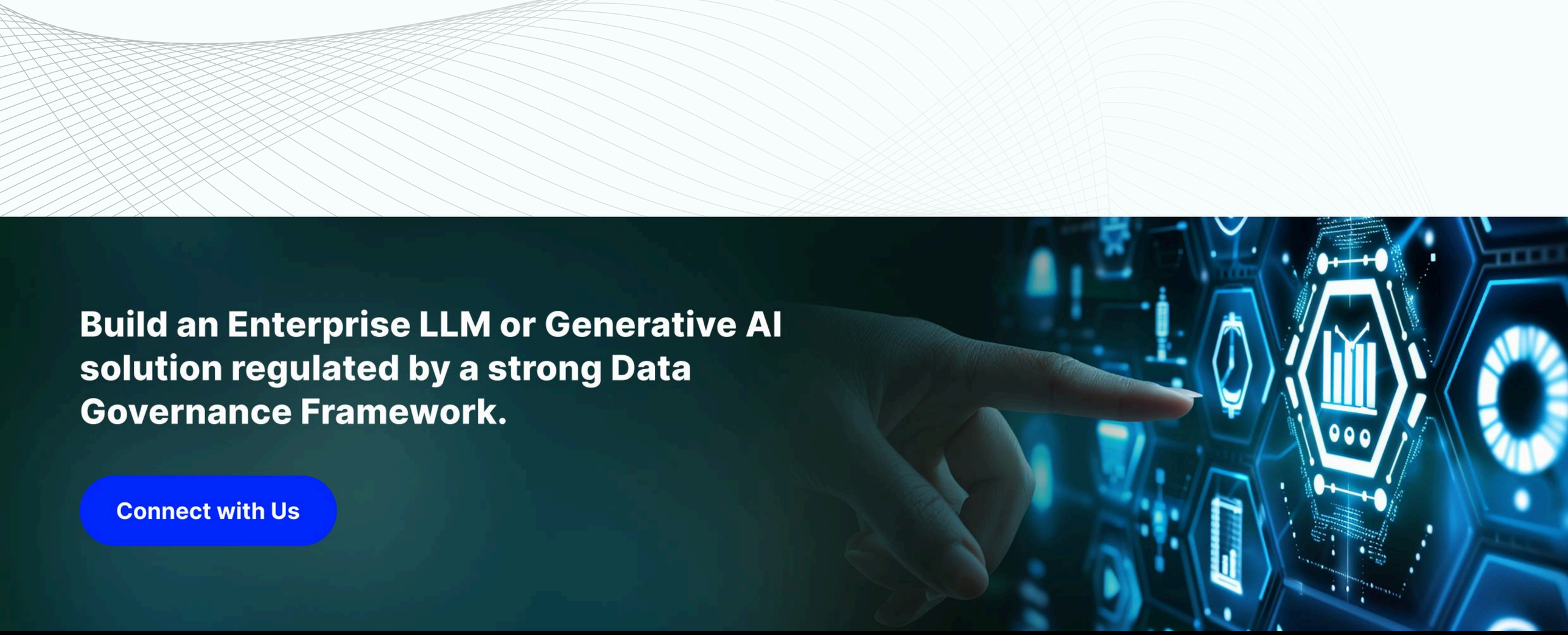
Customer data platforms

(CDP) are specialized tools for centralizing customer information and enabling marketing, sales, customer service, and other customer interactions.

About Us

Gleecus TechLabs Inc. is one of the fastest growing IT innovation partners for startups, SMBs, and enterprises that help clients envision, build, and run more innovative and efficient businesses. We envision the data engineering and governance strategy for enterprises exploring enterprise-grade Generative AI solutions to modernize their operations and boost customer satisfaction with cutting-edge solutions.

Our team builds, trains, and maintains GenAI models and integrates with existing enterprise workflow to open new horizons of knowledge and innovation for organizations. Our expertise in framing and handling data governance for large-scale real-time AI/ML solutions makes us the go-to-guy for enterprises looking to introduce and scale GenAI and Enterprise LLMs for business excellence.



About Gleecus TechLabs Inc.

Gleecus TechLabs Inc.(ISO 9001:2015 and ISO 20000-1:2018) is a Forward Thinking Digital Innovation Company that empowers businesses to achieve their Digital Transformation goals. We partner with leading Enterprises, SMES and Startups to realize their true digital potential leveraging our deep focus and expertise on Cloud, Data, AI/GenAI, Automation, Product Engineering & Cyber.

Forward Thinking Digital Innovation

Email: hello@gleecus.com

Phone: +1 347 947 2022

www.gleecus.com