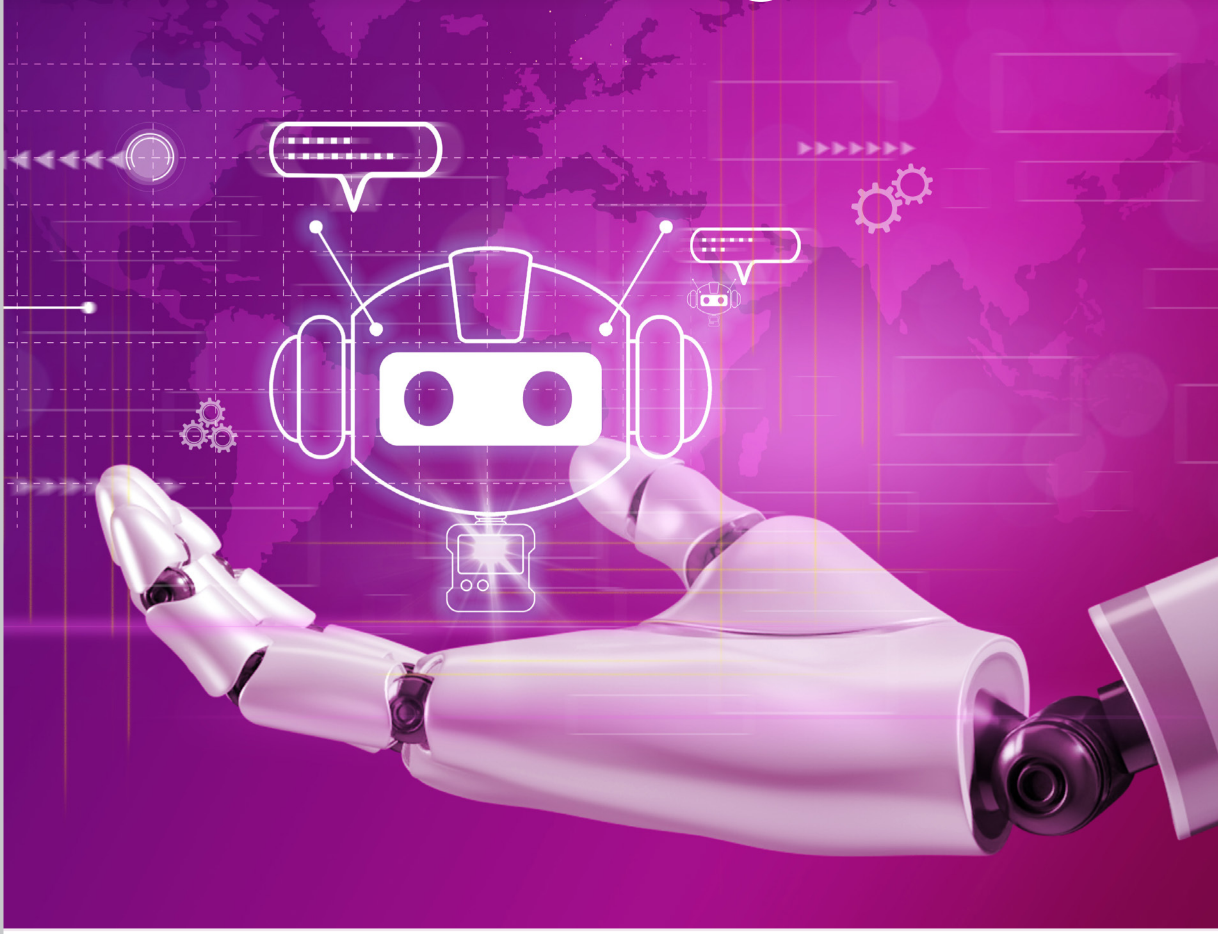
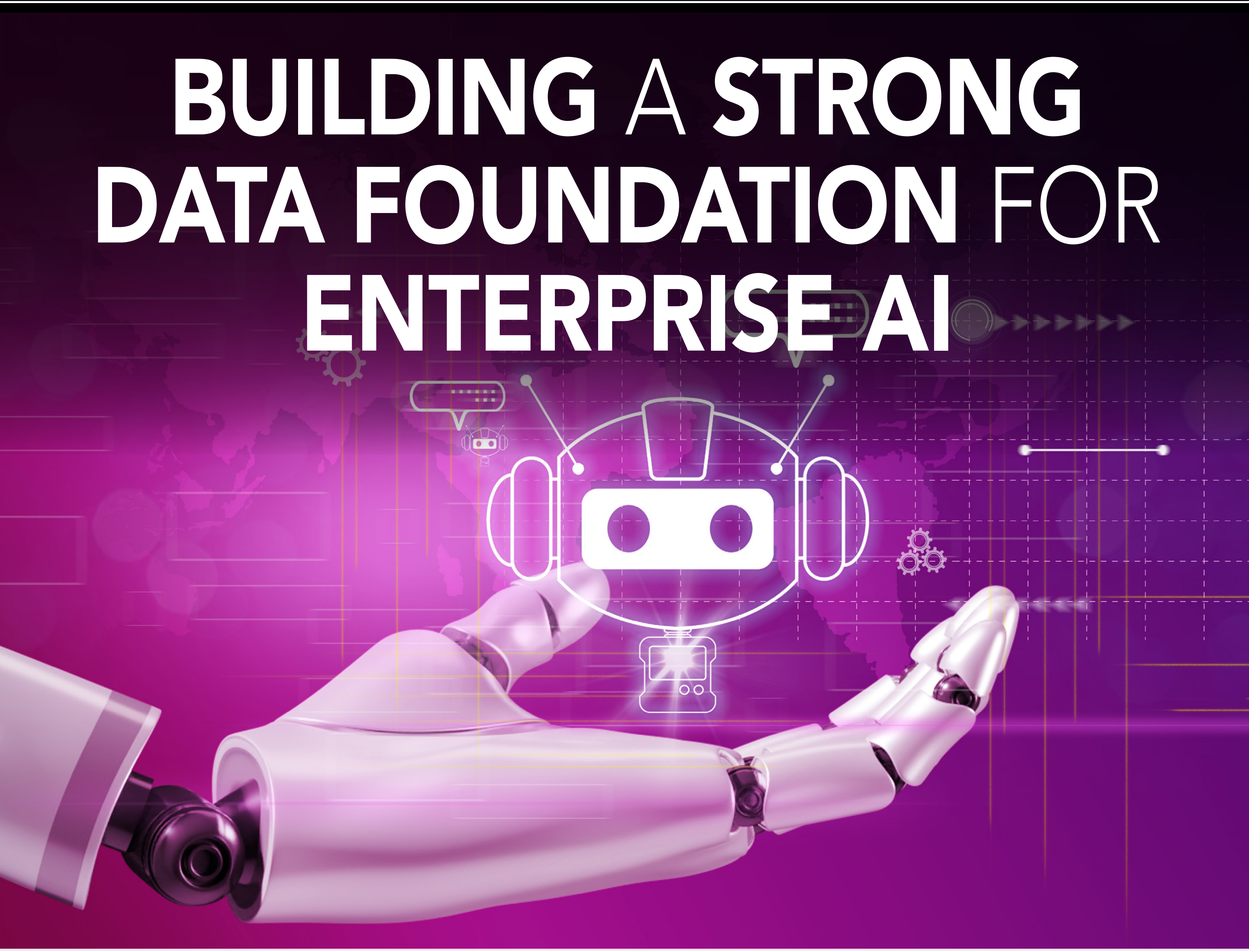
THOUGHT LEADERSHIP SERIES

BUILDING A STRONG DATA FOUNDATION FOR BUILDING ENTERPRISE AI



database
TRENDS AND APPLICATIONS
2024 | APR

BUILDING A STRONG DATA FOUNDATION FOR ENTERPRISE AI



Poor, uncalculated AI adoption is a daunting reality facing many enterprises aiming to jump onto the AI hype train. Despite AI's unequivocal popularity—and, simultaneously, its rapid pace of innovation—enterprises that leap into the promising pool of AI without proper planning will fail to generate any tangible value.

This "proper planning" is entirely dependent on an organization's approach to data. Only 4% of IT leaders say that their data is AI-ready, according to a report from Gartner, indicating the widening gap between the demand for AI and generative AI adoption and the ability for enterprises to make these projects produce business value. Evidence of this value, as pointed out by AWS' 2024 CDO Insights survey, is low, with only 6% of respondents having had a generative AI application in production deployment—despite 80% of CDOs agreeing that generative AI will transform their organization's business.

Clearly, attempting to embark on AI initiatives without acknowledging the plethora of data-related issues—including data quality, governance, volume, and security—is a recipe for disaster that fails to demonstrate any business value.

THE IMPORTANCE OF DATA QUALITY

Data quality is a massive roadblock for proprietary AI implementation. Low-quality data is a key driver of unsuccessful AI, where training and testing AI based on poor data can lead to biased, inaccurate, or irrelevant results—which will directly impact business decision making, according to a research article from *Applied Sciences*. As the authors of "AI and the democratization of knowledge" suggest, "AI is 'data-hungry'; untrained, it is the ultimate non-specialist."

According to Unisphere Research's "Taking on the Data Quality Challenge in the Age of AI" report, confidence in data quality is slipping, where only 23% of respondents expressed full confidence in their organization's data—which is down by 7% from a survey posing a similar question two years ago. The report further pointed to the fact that AI adoption emphasizes the importance of data quality, with 57% reporting that they find out about data quality issues during the implementation of next-gen data projects.

As the authors of the *Applied Sciences* article suggest, establishing data quality at

> *Low-quality data is a key driver of unsuccessful AI, where training and testing AI based on poor data can lead to biased, inaccurate, or irrelevant results …*

your enterprise is a seemingly Sisyphean task due to the volume of data generated daily and the intrinsic intricacies of data structures. The key to promoting a culture of data quality lies within data monitoring, measurement, and most importantly, governance.

ROBUST DATA GOVERNANCE IS KEY

A central component of a robust data strategy is data governance, which, when in place, positively impacts every working cog of an enterprise's data machine. With over 463 exabytes of data predicted to be generated daily by 2025, data governance is, and will be, fundamental in shaping a strong data foundation to promote AI success. According to Gartner, an agile, adaptive governance approach—as opposed to a traditional, one-size-fits-all solution—will enable enterprises to deliver the value, scale, and speed of governance that is necessitated by digital business.

Successful data governance must encompass several different processes and areas of business, including data stewardship and data privacy and security, according to *Applied Sciences*. By both attributing responsibility and accountability over data quality to data workers, as well as adopting methods of security that protect sensitive information and promote rigid compliance, enterprises will be able to bring greater trust and confidence in data—which is crucial for AI success. Additionally, establishing a data governance framework is conducive to creating a data-driven culture within an enterprise that promotes collaboration and communication between various parts of the business, helping to break down data silos.

Not only does data governance play a significant role in establishing a robust data foundation for AI, but AI and ML initiatives can amplify the efficacy of data governance. AI can be leveraged to enhance monitoring, data preparation, and policy enforcement to ensure that a data governance architecture is as effective as it is seamless to administer and maintain. However, as intersectional as AI and establishing a strong data foundation may appear, putting the necessary safeguards in place that wrangle AI in as-needed is fundamental.

Safeguards and security are essential components of building a strong data foundation to support AI projects. According to the Info-Tech Research Group, privacy is all about personal data, and building an effective data privacy program includes defining how the data is processed, creating notices and capturing consent, and protecting the data itself. This is especially crucial when understanding the ways that threat actors—who advance their attack techniques as quickly as AI technology innovates—may take their aim, including through adversarial attacks, data poisoning, model and data tampering, and more, according to *Applied Sciences*.

As both the privacy landscape and the methods of threat actors continue to rapidly evolve, adhering to the myriad of regulations surrounding AI will be key toward its success—as evidenced by a Statista survey that found that 39% of business leaders cited consumer concerns about privacy and use of data as a primary trust issue regarding AI in a business context. Whether the conversation around data privacy and security focuses on the producer or the consumer, it remains a vital conversation for enterprises to have.

IMPLEMENTING A DATA-CENTRIC APPROACH TO AI

Developing an AI-ready data foundation will require enterprises to focus their efforts on data that serves their AI ambition, according to Gartner. They offer the following key criteria to guide a data-centric approach to AI adoption:

- **Ethically governed**, aligning different artifacts valued by stakeholders around AI principles
- **Secure**, making sure that data isn't being fed into the internet and other enterprises' large language models (LLMs) unless desired
- **Free of bias**, gathering data from diverse sources
- **Enriched** with business rules and tags so the data is ready to be consumed
- **Accurate** through a process of human validation of data

Each of these criteria are dependent and built on one another, Garnter pointed out. This is a consistent theme in imagining the ways that an organization can define a data-centric culture that will fuel its push for AI; recognizing the various interdependencies and the woven nature of a strong data foundation is an implementation imperative.

While 93% of CDOs agree that a robust data strategy is crucial for getting value out of generative AI, 57% have not made any changes to their data strategy yet, according to the AWS survey. Taking the necessary steps to build a strong data foundation, as opposed to blindly diving into AI implementation, is worth the challenges and costs it may impose. After all, jumping onto the AI bandwagon *without* putting in the work to ensure its success will be more costly for both the enterprise and its workers in the long run. ■

—Sydney Blanchard

Building Trust with Users of Generative AI Capabilities

Pythian
love your data

A new technology is only as valuable as the measure of how widely adopted it becomes across an organization. Countless examples of exciting technology were never viewed as successful due to low adoption, often driven by a lack of trust in the product or its ability to add a level of value overcoming the cost, loss of privacy, or added strain of usage. User trust is the single most important dynamic to ensuring success of new capabilities and technologies.

With the interest and early experiences with Generative AI (GenAI), there is a hyper awareness of the accuracy produced by large language models (LLMs). Users are aware of the potential for hallucinations and are hyper-conscious of the output, regularly reviewing results for accuracy. Each time a user successfully uses a GenAI tool and receives a correct answer, their trust builds. Conversely, when a user receives an incorrect answer there is a loss of trust, maybe not absolute, but measurable and non-recoverable if not corrected. GenAI tools must be deployed with this in mind, ensuring user feedback is easy to provide, rapidly acted upon, and automated in testing to prevent reoccurrence.

> "By properly educating users, aligning processes to tools, designing rich UIs, and monitoring for data drift you can build and maintain a foundation of trust with the users and the GenAI capabilities you are deploying in your own organization."

LLMs are the race car of the technology world. They consume data faster than any previous analytical tool and require near constant investment in tuning and optimization to maximize their performance. While this enhances their ability to identify and leverage patterns and relationships in data, it makes it extremely difficult to debug, analyze, and understand when hallucinations occur. The speed at which we see GenAI technologies being deployed is both a blessing and a curse. The user community is becoming educated and aware at hyper rates of speed but also keenly aware of incidents across the community that affect their own trust in the technology.

To support the inevitable debugging of output and retraining of LLMs and supporting analytical models, organizations must work to ensure that discrete steps within well-known business processes are enhanced 1:1 with the use of GenAI tools. Implementing a GenAI solution that automates multiple steps of a process at the same time can lead to confusion about where bad data, process, or misunderstanding is entering in and contributing to incorrect output. By breaking down corporate processes to discrete steps and applying GenAI to smaller steps early and then growing, you minimize the chances of errors manifesting for users.

By tying GenAI to discrete steps within existing business processes, we can bookend each implementation with a human for reviewing output, determining accuracy, executing, and submitting examples of incorrect or inaccurate data. We gain the value of GenAI for acceleration and efficiency, while also working to build trust with the teams that know the business process best.

Users can only correct data they can see, they can only identify gaps when the output is clear and described. When building the user interface (UI) for GenAI tools, it is impactful to embed results into existing tools and workflows to reduce the impact of "swivel-chair" between applications. The UI should also include the ability to describe how the results manifested, report errors, and unmask obfuscated data when appropriate to fully understand the results and impact on downstream process steps.

Even after your organization has built an effective system to identify and correct GenAI results due to poor quality or incomplete data, you must contend with the ever-changing data landscape present in every organization. The constant shifting of data transformations, upstream technologies, application integration, and user input leads to drift in all analytical results. Organizations must establish processes early to monitor and correct for drift of results over time and accept the organizational dynamic of a never-ending list of user reports that must be addressed and corrected.

Users have to trust the results, must have paths to report errors, and a closed look to know that corrective action was taken. Systems will make incorrect recommendations, even with the best tuning, our data will change and evolve, which will lead to new and undiscovered problems. By properly educating users, aligning processes to tools, designing rich UIs, and monitoring for data drift you can build and maintain a foundation of trust with the users and the GenAI capabilities you are deploying in your own organization.

If you'd like to learn more about GenAI and Pythian's capabilities to support your organization in preparing for innovative technologies, download our "Capitalize on the Art of the Possible" eBook here. ■

PYTHIAN
www.pythian.com

Hallucination-Proof your GenAI Business Apps

melissa

Core Data Quality, Unbiased Training Datasets, and Semantic Inference for Accurate GenAI

AI-ENABLED DATA QUALITY, "GOLD STANDARD" REFERENCE DATASETS AND SEMANTIC SUPERVISION

Generative Artificial Intelligence (GenAI) capabilities, for years limited to R&D and high-profile projects, can now be usefully applied within many businesses.

However, if data quality (DQ) and semantic fundamentals aren't well managed, GenAI initiatives can return clearly inaccurate gibberish. Even worse, GenAI applications can return "hallucinations"—factually inconsistent and/or subtly biased results that may seem accurate and useful, but that can lead to chaos if applied in real-world business applications. Even clean business data can serve as a source of hallucinations if training data is biased or incomplete or if the application supervision is biased.

In order to avoid GenAI hallucinations in your business, ground your AI projects in clean, meta-data rich business data; in curated and meta-data rich reference data for training; and in semantically informed results supervision.

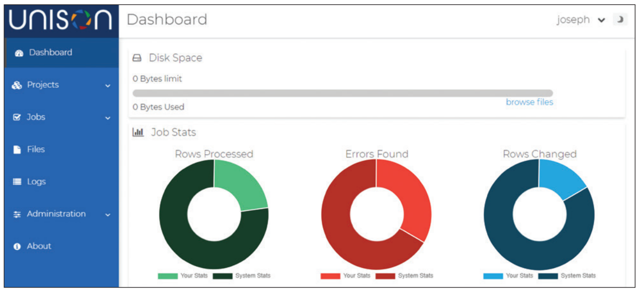


Figure 1: Apply low-code, AI-enabled tools for active DQ, to prepare your business data for accurate GenAI outputs, and to compare output to expected results (Ref: Unison, Melissa)

START WITH ACTIVE BUSINESS DATA QUALITY (DQ)

Effective AI requires clean data. Dirty business data leads to failed AI efforts. This has become a truism but is still often overlooked in practice! For effective outcomes, every GenAI project should start with active DQ and core DQ management, including profiling, deduplication, cleansing, classification, and enrichment[1].

Vendors like Melissa Data Corporation (Melissa) are integrating AI-enabled applications such as Chatbots and semantically informed DQ rules generators within unified DQ project environments. This makes it possible to generate and apply active rules to cleanse, integrate, harmonize, and enrich data referenced by your GenAI application.

IDENTIFY AND APPLY CURATED, MULTI-SOURCE, META-DATA RICH REFERENCE DATASETS

If reference datasets and training are biased, AI applications will inevitably betray their users. Large Language Models (LLM), like the "Gemini" GenAI application, have offered recent examples of hallucinations resulting in negative business outcomes[2].

"Gold Standard" reference datasets and empirically grounded supervision are critical to ensuring accuracy from your GenAI applications (hallucination-proofing). For example, Melissa curates and integrates multiple sources to create the highest quality demographic, customer, firmographic, and geographic reference data resources.

In addition to serving as training datasets, gold standard reference datasets can support downstream reasoning and active AI result supervision by comparing outcomes with expected reference data patterns and content—to flag or correct potential errors for application of new business rules and/or AI supervision to ensure accurate outcomes.

SUPERVISE AND REFINE RESULTS WITH AI ENABLED, SEMANTICALLY AWARE RULES GENERATION

Hallucination-proofing can be enabled by combining complementary AI technologies. For example, machine reasoning can auto-generate and apply DQ rules informed by expert "semantics" such as those provided by Financial Industry Business Ontology (FIBO) or National Center for Biomedical Ontologies (NCBO). These can be applied to assess and correct GenAI outcomes.

Computers can apply machine reasoning to infer DQ rules based on logic and reference data. Upstream, inference grounded in accurate reference data can be useful for data cleansing and enrichment. Downstream, ontology-informed, reference-based reasoning can supervise and flag or auto-correct GenAI outputs. If the system has reference data and semantic reasoning capabilities, it can automatically generate and apply DQ rules.

For a medical example, an expert ontology combined with curated medical reference data such as the Unified Medical Language System (UMLS) can automatically "know," for example, that if a specific drug was prescribed as a patch, then the dose will be greater than 15 milligrams (rather than a low dose pill). If the GenAI application hallucinates that a patch is dosed below 15mg, machine reasoning can detect and correct that result—potentially saving lives.

CONCLUSIONS

To hallucination-proof your GenAI application, build from tools and resources that support empirical accuracy! Ground your GenAI projects with clean, well-managed business data, gold standard reference data, and active AI-enabled DQ and result supervision.

Melissa offers unmatched quality for reference datasets and AI-enabled unified "low-code" DQ and project management tools. Ask Melissa about our AI-enabled tools and resources to address your DQ needs. ■

REFERENCES:

1. How to Prepare Your Data for GenAI. Kennedy, Russ: Forbes, Forbes Technology Council: January 16, 2024.
2. Google can't guarantee its Gemina AI App won't be Biased. Mearan, Lucas: Computerworld. February 29, 2024