CHECKLIST

Leveraging a Data Fabric for Generative AI

By Fern Halper, Ph.D.

denodo

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

Leveraging a Data Fabric for Generative AI

By Fern Halper, Ph.D.

The introduction of generative AI has changed the landscape of AI. Generative AI—a subset of artificial intelligence that involves systems designed to generate outputs such as images, music, text, or other forms of media based on its training data—has become top of mind for many organizations. In a recent TDWI survey, for instance, generative AI was among the top four priorities for analytics in 2024, ranking higher than machine learning.

Use cases for generative AI are starting to move past chatbots and employee onboarding systems that might utilize some company-specific information toward implementations that *require* company data about customers or products. For example, a chatbot might be enhanced to include customer loyalty information to provide modified responses based on whom the chatbot is conversing with. Another

To leverage a logical data fabric for generative AI:

1. Understand your business objectives and use cases
2. Consider new technologies and processes
3. Determine the data architecture
4. Identify the right data for your use case
5. Incorporate business semantics and enhance metadata
6. Ensure data quality, security, and governance
7. Monitor and manage generative AI applications

use case might involve sending targeted marketing messages developed by generative AI to specific customer segments. Still another use case might make use of real-time delivery information.

A popular technique that uses company data in a generative AI system is retrieval-augmented generation (RAG). RAG can provide data in context to foundation models such as large language models (LLMs) without needing extensive retraining that can be computationally intensive. RAG combines specific information with prompts to enhance the relevance of the model's output.

To effectively use an approach such as RAG, it is important to have a solid data foundation. Yet organizations still struggle with data silos. In TDWI research we see that most respondents utilize some sort of hybrid data management environment. This might consist of on-premises and cloud-based platforms or multicloud platforms. Legacy systems are also still part of the landscape.

The hybrid nature of the data environment further complicates building generative AI models with company data. How do you utilize data relevant to the question being asked when it is distributed among systems, stored in various formats, and requires different access methods? How do you ensure that data provided to a generative AI model is accurate and timely? Can you provide the proper access controls to ensure that the person asking the question has access rights to the data and that you are not disclosing data they should not see? How will your company explain the inputs to and outputs from models such as LLMs? In many cases, a logical architectural approach can help.

This TDWI Checklist Report explores how to implement generative AI applications using a logical architectural approach such as a data fabric together with techniques such as RAG. Adopting RAG may seem straightforward for a single use-case prototype covering unstructured data (e.g., documents) or dynamic real-time data from a single source, but the challenge lies in scaling it to meet the diverse needs of a variety of teams and use cases (such as customer service, data analysis, and decision support). For instance, a customer service chatbot utilizing RAG must access multiple data sources—such as an ERP system, support ticketing system, CRM system, and internal API endpoints—to provide comprehensive and accurate responses to customer inquiries. This will require careful planning and strategic implementation.

1 Understand your business objectives and use cases

The first step in any successful AI implementation is to determine the business objectives and use cases for the technology. Business issues should be driving the deployment of generative AI; it shouldn't be the other way around.

Different business problems require different AI approaches. Although generative AI is an exciting technology, it may not be the best AI technology for every use case. The choice between generative AI and other types of AI, such as predictive AI, should be driven by the specific business objectives and needs of the project. Predictive analytics, for instance, is highly effective in use cases such as retention analysis, cross-selling, forecasting, and fraud detection. It

leverages statistical and machine learning techniques to predict future outcomes based on historical data. In contrast, generative AI excels at creating new content or data that mimics a given data set. It can be used for generating marketing emails, summarizing text, or utilizing dynamic data in a chatbot. Understanding these distinct capabilities and aligning them with business goals ensures that the chosen AI technology effectively supports strategic objectives. Of course, some use cases will utilize both traditional and generative AI.

By pinpointing and clearly defining the use case, your organization can tailor its generative AI models and methodologies to fit the specific needs of your problem, which enhances the effectiveness and efficiency of the AI solution. Addressing a specific, relevant problem or opportunity within your organization increases the likelihood that the AI solution will have a meaningful impact on business outcomes, driving value and supporting strategic goals. It also supports practical issues such as resource allocation.

Finally, having a clearly defined use case allows your organization to set specific, measurable goals and KPIs (key performance indicators) for the AI project. This makes it easier to evaluate the success of the initiative and make data-driven decisions about future AI investments and developments. At TDWI, we've seen that those organizations that have clearly defined metrics are more likely to build success with analytics and AI. It is a virtuous circle; showing success helps build more success.

2 Consider new technologies and processes

It will be important to determine what your tech stack looks like and how your organization can use it to support your identified business goals and use cases. This includes new components that are required for utilizing generative AI against company data. These components are both technology and process related.

On the technology front, some of the new components include:

- **Foundation models.** Foundation models are large-scale, pre-trained machine learning models that serve as a base or foundation for a wide range of downstream tasks. These models are trained on vast amounts of data (typically internet-scale) and can be fine-tuned or adapted to specific applications with additional data. There are numerous kinds of foundation models available, some better suited to specific tasks than others. For instance, encoder models might work best for sentence classification tasks; decoder models such as GPT for text generation; encoder and decoder models for translation. Typically, organizations will experiment with off-the-shelf LLMs that have been trained to generate text as a starting point for generative AI. These include GPT, BERT, Gemini, Claude, Llama, and others.
- **Vector embeddings and vector databases.** Many generative AI applications require input in a way that a computer can understand. This is often in a vector format (a numerical format that requires data to be transformed) and the

model input is called a vector embedding. Vector databases are designed to store, retrieve, and manage high-dimensional vector representations of data efficiently. This means that organizations may need to first transform their data into a vector embedding and then store it. Some cloud platforms include vector databases as part of their offerings.

- **Data access.** RAG was mentioned earlier as a framework for helping LLMs access real-time data in context. It can be useful for applications that require precise and contextually accurate information. RAG works by first retrieving data from an external source system that has been vectorized, then taking this data together with the initial prompt (e.g., how many units a customer bought) and responding in natural language. RAG provides contextual information to an LLM or other kind of foundation model. It can be fairly simple to build an application using retrieval-augmented generation if there is one data source, but applications may require multiple siloed data sources. Development frameworks such as Langchain (a Python library) are one way to construct a RAG application.
- **Development frameworks.** Developers need frameworks to provide structure and templates for building applications. There are numerous frameworks that developers might use to build generative AI applications. A popular one is Langchain, which provides the structure to connect company data to LLMs and makes it easier to build RAG models. Langchain provides a comprehensive framework for building complex applications using LLMs. Langchain can help create and manage workflows that involve multiple steps and components (e.g., RAG, which is a specific technique). Other frameworks for building applications include LlamaIndex and LangDock.

Of course, it will be important to train developers and others how to use and interact with these new generative AI technologies. This will include training on these tools and frameworks as well as establishing support mechanisms for troubleshooting and maximizing the benefits of the technology.

3 Determine the data architecture

Your organization won't succeed in deploying generative AI against its own data unless it has a solid data foundation in place. At TDWI, we see that many organizations have a hybrid architecture deployed. They typically utilize both on-premises and cloud platforms or have multicloud environments. There may be legacy, as well as ERP and CRM, systems. Organizations struggle with data silos and bringing together different data types for AI. They also struggle with pipelines. That means that it may be difficult to overcome latency issues in traditional ETL/ELT approaches if the application requires recent transactions. For instance, a customer might have performed a transaction, but it may take time to be part of the generative AI response.

Silos can make techniques such as RAG difficult if the application must query different systems to obtain data, such as order data from a CRM system and customer trouble ticket information from another system. To overcome these silos, organizations are taking several architectural approaches to data integration. One is physical, the other logical.

In the physical approach, the goal is to centralize all the company's data into a cloud data platform. This might include a data lakehouse that combines a data lake and a data warehouse that provides warehouse

data structures and data management functions on low-cost platforms, such as cloud object stores. Although centralizing data seems like a good idea, it is often difficult, if not impossible, to centralize all data in one physical location. Data is rarely all in one source in a company.

The logical approach utilizes a data fabric, which brings together disparate data in an intelligent fashion. The data fabric maps and connects relevant application data stores with metadata to describe data assets and their relationships. In a data fabric, there can be siloed data; on top of this, however, is an abstraction layer (a semantic layer or a virtualization layer) that contains the metadata. This layer integrates distributed data sources to present a unified view of data, making it appear as though all data comes from a well-integrated source, regardless of its actual location. This can be very helpful, for example, when structured data is in an on-premises data warehouse, unstructured data is in a cloud data lake, and data is also needed from an ERP system.

The logical data architecture can help with generative AI RAG applications that require data from different systems (this is described in more detail in number 5). The fabric helps to ensure security across the systems and provide the metadata to choose the right system from which to get the data and run the query.

4 Identify the right data for your use case

Generative AI that makes use of data such as customer data is more complex to implement than simply pulling information from a product documentation manual. It may require querying multiple kinds of data. For instance, if a prompt is a question about billing by region for a certain kind of customer, it will need to determine which system(s) to query, where to run the query, and what kind of SQL to use. In a data fabric approach, high quality metadata populated in the vector database via the embedding process might be used.

The LLM can leverage a vector database the same way that end users can leverage a data catalog to identify the best data to use. Business semantics presents file names in terms understood and commonly used by the end users. It also provides comprehensive descriptions of views and individual fields. All this data is available to allow users to find and use the best data for their specific needs. Similarly, an LLM can utilize the metadata to understand the structure, context, and relevance of different data sets. That means that by utilizing a fabric approach, the right data can be used for an application, not simply the data that is easiest to access.

On the organizational front, it will be important to work with domain owners to identify the best data for the application. At TDWI, we see more organizations moving to a data product model. Data products are derivative assets created from data. These products can run the gamut, from enriched data sets provided to a customer to a dashboard that provides the output of machine learning models to external partners—or apps that use derived data for a specific industry. Data products promote the documentation of data assets. These data assets, often managed by domain owners, can be detailed and fed into a vector database. This allows LLMs to automatically identify the best information to respond to prompts. Therefore, collaborating with domain owners and ensuring thorough documentation of data assets can be very beneficial in determining the best data available.

5 Incorporate business semantics and enhance metadata

If your organization is building a generative AI application using RAG and data from different sources, the data fabric approach can be helpful because it can provide access to multiple data repositories in real time and supports complex queries that involve many JOINS and source-optimized SQL. Of course, that requires business semantics and enhanced metadata as mentioned earlier.

The process might work something like this, using a framework such as Langchain to "chain" together different components and create a RAG application.

- Data sources are integrated via a data fabric approach. This approach utilizes a semantic layer that contains business and other metadata.
- Table schemas are extracted and vector embeddings (using models of your choice) are created using data from the fabric.
- A question is posed via a prompt. Langchain is used to gather the table schema and other context-relevant information.
- This context-relevant information is passed to the LLM that generates the SQL that is then executed against the appropriate data source in the data fabric.
- Once the results are received, the original prompt is packaged/augmented with the context retrieved from the data fabric and sent back to the LLM.
- The LLM then sends the response, in natural language, back to the user.

This kind of process works because business semantics are integrated into the logical data abstraction layer to provide business-friendly metadata terms that add context for the LLM. Enhancing metadata and incorporating business semantics improves data usability and relevance, leading to more accurate and contextually appropriate responses from the LLM, thereby enhancing decision-making.

6 Ensure data quality, security, and governance

Of course, it will be important to make sure that the data used in any generative AI application maintains its integrity, is of high quality, and is relevant and timely. Data quality is critical for the effectiveness, reliability, and ethical integrity of generative AI systems. Ensuring high data quality helps you build models that are accurate, unbiased, consistent, and performant, ultimately leading to better outcomes and user experiences. Likewise, organizations will need to develop broader data governance policies to ensure that data is compliant, used ethically, protected, safe, and performant.

Although data governance is always at the top of the list of priorities for data management in TDWI surveys, aspects of data governance for AI, such as bias detection, explainability, and responsibility, often do not rank as high. However, as organizations move forward with generative AI, they will have to build a plan to govern new

data types for AI, such as unstructured text data or image data, as well as govern the AI models themselves (see number 7).

Key factors for data governance and data quality management in generative AI include:

- **Ensure data quality and explainability.** In TDWI surveys, less than 50% of respondents say they are satisfied with the quality of their data. This may be because they are struggling to address new data types. Organizations will need to put tools and processes in place to ensure data quality to build trust in AI systems or users may not adopt them. At TDWI, we see that automated tools are becoming more popular to address data profiling and data quality. They are also being used to generate lineage—to track the origin of data, allowing stakeholders to trace it back to the original sources. This helps in understanding where the data came from and ensures its authenticity and reliability.

  Often data fabric solutions will provide these features as part of the solution or tightly integrate (via a technical integration) with solutions that do so.

- **Comply with security and privacy regulations.** Generative AI models must also comply with security and privacy regulations. This will include the ability to classify sensitive information so it can be handled properly in generative AI models. Additionally, only those who are authorized to use certain data should be able to do so for generative AI. That means that at a minimum, it is important to implement robust access controls. These can be role-based access control (RBAC) or attribute-based access control (ABAC). Tags should be assigned to sensitive data at the view, row, or column level. As part of the data fabric, it should be possible to set up security rules that apply to multiple users and view them all at once to make managing these rules easier. Traditional security mechanisms such as robust authentication, encryption, and masking will continue to be important for data both at rest and in transit.

  In the data fabric, the logical access layer establishes a central location for defining security and governance policies globally. Centralized security and governance ensure data compliance and protection, reducing the risk of data breaches and maintaining customer trust.

- **Mitigate risks.** It will be important to identify and safeguard against the potential reputational, financial, and legal risks of using generative AI and LLMs. For example, generative AI application output should be safe from hallucinations (where LLMs output wrong or irrelevant information). Data leakage (where private information is leaked) should be prevented. Work is being done to develop techniques to help detect risks such as a BLEU score (which measures the quality of the text output by a generative AI model), perplexity score (measures the ability of a language model to forecast upcoming words), and F1 score (measures accuracy). Human-in-the-loop processes are also being used to oversee models. It is still early days, however, as additional technologies are being developed. This is an area worth an enterprise's careful attention.

7 Monitor and manage generative AI applications

No matter how your company is using AI technologies, putting models into production comes with concerns that organizations need to address, especially when dealing with company data.

As is the case with traditional AI models, generative AI models and applications will need to be governed to ensure that the models put into production are trusted and that they retain their integrity. This includes putting operations (Ops) personnel in place to ensure that the models are versioned and documented. The Ops team may also track the model to ensure that it doesn't drift or become stale. In the case of generative AI, the team should also ensure that the model doesn't start to produce hallucinations and that high-quality data is fed to the model. Details about the data sets used should be part of the metadata provided by the data fabric. Some organizations are starting to utilize analytics catalogs that store information about models.

When using generative AI with your company's data, your concern must extend beyond the AI model itself. It will also be important to continuously monitor the performance and effectiveness of the RAG implementation. For instance, it will be important to know if upstream data being fed to the model changes or if a vector database fails or SQL can't be run.

The approach can be updated or refined as necessary to adapt to evolving enterprise needs and ensure sustained success.

Concluding thoughts

The integration of the data fabric with generative AI applications using techniques such as retrieval-augmented generation (RAG) represents an advancement for organizations looking to harness their proprietary data. The data fabric's capabilities to manage, index, and provide context to changing data, simplify access to distributed data, and ensure efficient query execution enable generative AI applications to access and utilize the most up-to-date corporate information.

There is a good deal of potential for generative AI applications to transform business operations. However, realizing this potential requires a careful and strategic approach to data management. Techniques such as RAG, supported by a comprehensive data fabric, can be important for ensuring that AI applications can effectively leverage company data.

About our sponsor

Denodo is a leader in data management. The award-winning Denodo Platform is a leading logical data management platform for delivering data in the language of business, at the speed of business, for all data-related initiatives across the organization. Realizing more than 400% ROI and millions of dollars in benefits, Denodo's customers across enterprises in 30+ industries all over the world have received payback in less than six months.

For more information, visit denodo.com.

About the author

Fern Halper, Ph.D., is vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, machine learning, AI, cognitive computing, and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. She has taught at both Colgate University and Bentley University. Her Ph.D. is from Texas A&M University.

You can reach her by email (fhalper@tdwi.org), on X/Twitter (x.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).

About TDWI Checklist Reports

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

About TDWI Research

TDWI Research provides industry-leading research and advice for data and analytics professionals worldwide. TDWI Research focuses on modern data management, analytics, and data science approaches and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of business and technical challenges surrounding the deployment and use of data and analytics. TDWI Research offers in-depth research reports, commentary, assessments, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

Transforming Data With Intelligence™

A Division of 1105 Media
6300 Canoga Avenue, Suite 1150
Woodland Hills, CA 91367

E info@tdwi.org

tdwi.org