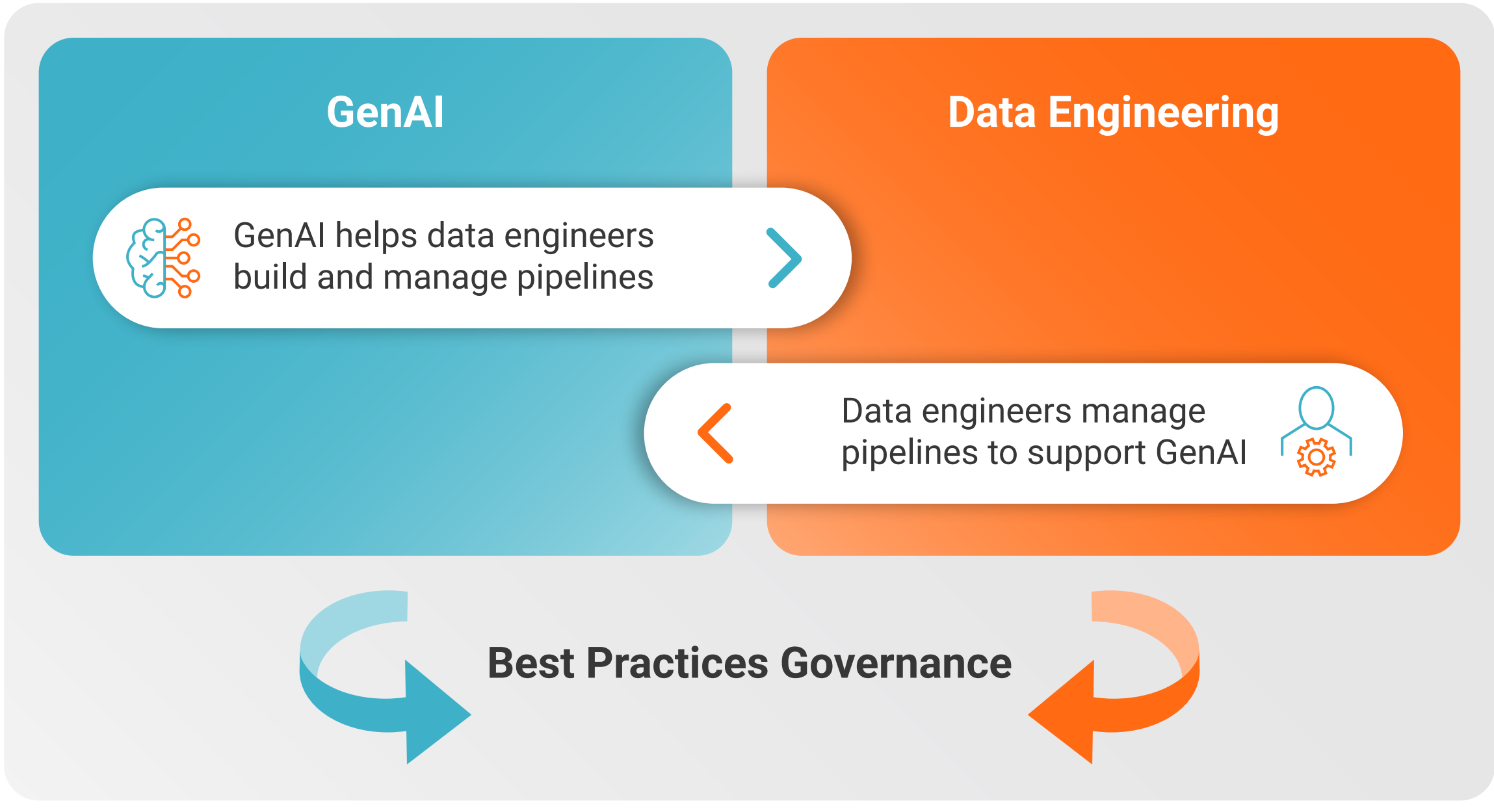


# Achieving Fusion: The Synergy Between GenAI and Data Engineering

Without question, data engineering needs the help of generative artificial intelligence (GenAI) and GenAI needs the help of data engineering. Here’s why: GenAI makes data engineers more productive, while data engineering provides GenAI new levels of innovation.

Uncover the secrets to fostering a dynamic interplay between GenAI and data engineering.

## The Fusion of GenAI and Data Engineering



## How GenAI Helps Data Engineering

### The Problem



Data engineers design, test, deploy, monitor and optimize pipelines that deliver data for analytics. The surge in SaaS applications, mobile apps, IoT sensors, data platforms, analytical tools and business users complicates the management of data ingestion and transformation tasks, making it challenging to seamlessly integrate these varied components.

### The Solution



Language model (LM) platforms and advanced features within pipeline tools can address this challenge. Based on natural-language prompts from data engineers, they generate starter code for data pipelines, suggest ways to debug the code and document the pipelines and related datasets for cataloging. Also, LMs recommend rules for data quality checks and evaluate different architectural approaches for designing pipelines. As a result, LMs save you immense time by automating, accelerating and simplifying the complex and tedious tasks that come with data engineering.

## How Data Engineering Helps GenAI

### The Problem





Companies are starting to build their own GenAI applications that include an LM (or application programming interface (API) that connects to an LM), a conversation user interface (UI) and extra functionality that executes tasks based on LM outputs. To deliver usable outputs, the GenAI applications need usable inputs and pipelines that can transform unstructured data objects into numerical vectors that enrich user prompts and assist LM fine-tuning.


### The Solution



Data engineers can address this challenge by building pipelines that comprise the stages of extraction, transformation and loading (ETL) or ELT. The sequence of extract and load, transform and load again can prepare a company’s domain-specific data for usage by their GenAI applications.

- **Extract and Load**

The pipelines extract relevant text from applications and files, then load it into a landing zone on platforms such as the Databricks Lakehouse or Snowflake Data Cloud. To improve GenAI accuracy, this text should align with master data and meet quality standards.
- **Transform**

The pipelines transform the data to prepare it for LM consumption, followed by converting words to numerical tokens, grouping the tokens into “chunks” and creating vectors that describe the meaning and interrelationships of the chunks.
- **Load**

The pipelines load the embeddings into a vector database (like Pinecone and Weaviate) or vector-capable platforms (like Databricks and MongoDB).

As a next step, data teams can utilize the vectorized data to support GenAI applications in two primary ways:

- 1

Implement retrieval-augmented generation (RAG), which finds relevant content within the vector database and adds it to user prompts so that the LM is more likely to provide high-quality answers.
- 2

Fine-tune the LM by adjusting its parameters to align with the vectorized text.

Both RAG and fine-tuning are instrumental in augmenting the precision of responses generated by GenAI applications — this combination ensures a more accurate reflection of the business context and mitigates the risk of data hallucinations.

Reimagine your data strategy with the combined power of GenAI and data engineering.

LEARN MORE

Source: Eckerson Group, Achieving Fusion: How GenAI and Data Engineering Help One Another, 2024

