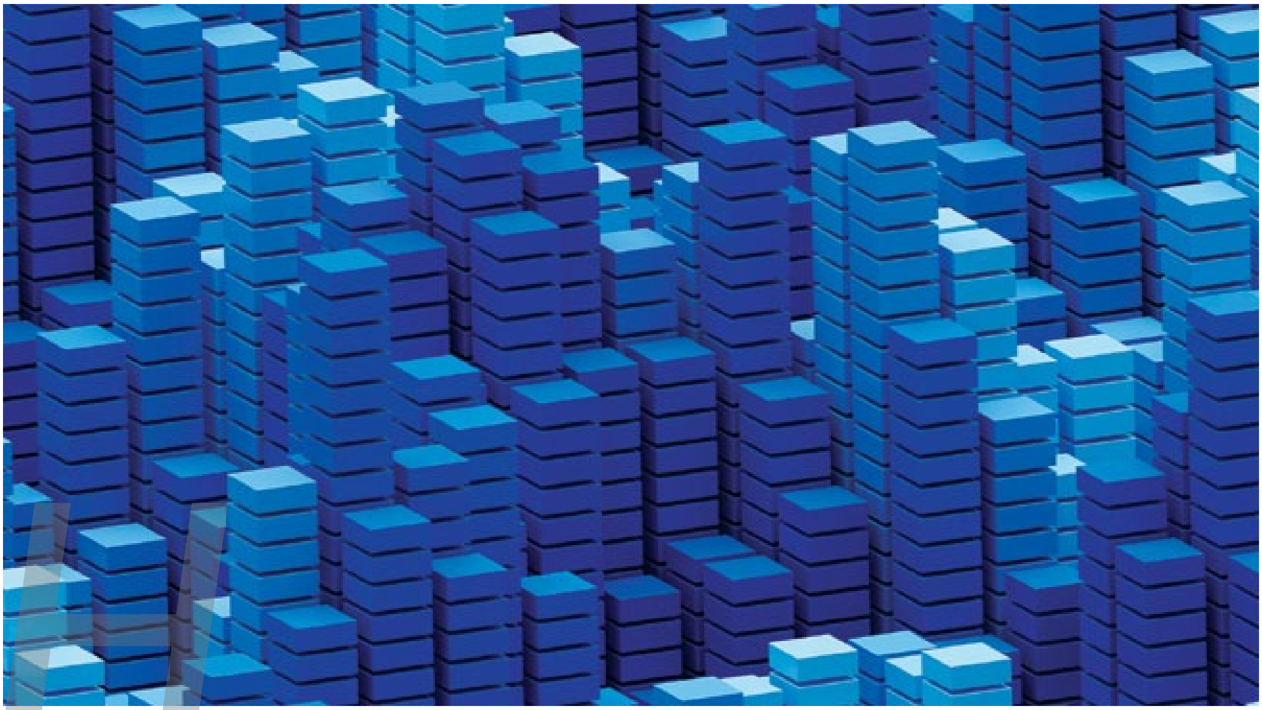
The Data Dividend

"If your data isn't ready for GenAI, your business isn't ready for GenAI"

By Raju Chellam



Here's a corny corporate story: At the annual company picnic, Bob, the IT guy, won the "Employee of the Year" award. Suddenly, everyone wanted to be his friend, even the CEO's dog. "No wonder they say success is relative," Bob laughed. "Because the more the success, the more the relatives."

That paradigm is also true with data – whether authentic or fake – with a twist. The more the data, the more the relationships between authentic data and possibly fake data, with some outcomes, decisions or predictions being at risk. Adding to this mix is the emergence of a new type of data: synthetic data.

Synthetic data is annotated information that computer simulations or algorithms generate as an alternative to real-world data. It is data created in digital worlds rather than collected from or measured in the real world. Why do you need it? Because authenticated data is the new oil in the AI age, but only a handful of companies can afford to have a massive corpus and a continuous stream of authentic data. Others are therefore making their own fuel, one that is both inexpensive and effective.

"It may be artificial, but synthetic data reflects real-world data, mathematically or statistically," notes Nvidia. "Research demonstrates it can be as good or even better for training an AI model than data based on actual objects, events, or people. That is why developers of deep neural networks are increasingly using synthetic data to train their models." A survey in 2019 reported that use of synthetic data "was one of the most promising general techniques on the rise in modern deep learning, especially computer vision," which relies on unstructured data like images and video.

> RESEARCH DEMONSTRATES IT CAN BE AS GOOD OR EVEN BETTER FOR TRAINING AN AI MODEL THAN DATA BASED ON ACTUAL OBJECTS, EVENTS, OR PEOPLE. THAT IS WHY DEVELOPERS OF DEEP NEURAL NETWORKS ARE INCREASINGLY USING SYNTHETIC DATA TO TRAIN THEIR MODELS.

PLACEHOLDER POSITION

Synthetic data can be used as a placeholder for datasets. It is more frequently being used for training of ML (machine learning) models because of its benefit in data privacy, such as in healthcare apps to protect patient data and enhance clinical trials.

"The interest from the healthcare sector stems from the compliance regulations surrounding patient data," IBM reports. "For instance, Health Insurance Portability and Accountability Act (HIPPA) is a US Federal law that protects individuals' information from being discriminated against, which synthetic data helps overcome by creating AI generated data."

About 72% of leading organizations polled in a McKinsey survey said managing data was one of the top challenges that prevented them from scaling AI use cases. The challenge for CDOs (chief data officers) is to focus on changes that can enable GenAI to generate the greatest value for the business.

"If your data isn't ready for GenAI, your business isn't ready for GenAI," McKinsey says. "GenAI could add the equivalent of US$2.6 trillion to US$4.4 trillion in annual economic benefit across 63 use cases. Pull the thread on each of these cases, and it will lead back to data. Your data and its underlying foundations are the determining factors to what's possible with GenAI."

In determining a data strategy for GenAI, CDOs might consider adapting a quote from US President John F. Kennedy: "Ask not what your business can do for GenAI; ask what GenAI can do for your business." Focus on value is a long-standing principle, but CDOs must particularly rely on it to counterbalance the pressure to "do something" with GenAI, McKinsey advises.

SYNTHETIC SENSE

To be clear, there are three key types of synthetic data:

- **Fully Synthetic:** No real data is used with this technique. The computer program may use real-world data characteristics to narrow down and estimate realistic parameters. The data generator will identify the density function of features in the real data and estimate parameters. The data is then randomly generated and provides a strong privacy protection.
- **Partially Synthetic:** This technique replaces only a portion of selected sensitive features with synthetic values and keeps some real data or existing unstructured data. This technique is useful when data scientists are trying to fill in the gaps in original data and is done to preserve privacy in the newly generated data.
- **Hybrid:** A combination of real and synthetic data that takes random records from a real dataset and pairs it with close synthetic records. "This technique has advantages from both fully and partially synthetic data," IBM says. "While it can provide good privacy preservation, the drawback is the longer processing time and more memory."

What's driving organizations to use synthetic data in their test and production environments? A slew of regulations covering data privacy, and soon, AI. About 18% of enterprises are integrating synthetic data to address privacy regulations and facilitate secure data exchange in insurance services.

"Due to regulations surrounding AI, 40% of AI algorithms utilized by insurers in the policyholder value chain will utilize synthetic data to guarantee fairness within the system and comply with regulations by 2027," reports IDC. "This integration of AI spans from underwriting to marketing and claims handling. However, concerns about privacy and bias will require insurers to develop guidelines that align with evolving regulations like the EU AI Act to ensure compliance, address biases, and enhance transparency."

TESTING TIMES

Since synthetic datasets maintain statistical properties that closely resemble the original data, they can produce precise training and testing data that is crucial for model development. For example, training computer vision models often requires a large and diverse set of labeled data to build highly accurate models. Obtaining and using real data for

> LEADERS CAN USE TECHNIQUES SUCH AS DIFFERENTIAL PRIVACY TO ENSURE ANY SYNTHETIC DATA GENERATED FROM REAL DATA IS AT VERY LOW RISK OF DEANONYMIZATION.

this purpose can be challenging, especially when it involves PII (personally identifiable information).

"Two common use cases that require PII data are ID verification and ADAS (automated driver assistance systems), which monitor movements and actions in the driver's area," says Gartner senior director analyst Alys Woodward. "In these situations, synthetic data can be useful for generating a range of facial expressions, skin color and texture, as well as additional objects like hats, masks, and sunglasses. ADAS also requires AI to be trained for low-light conditions, such as driving in the dark."

Why does it matter? Because efforts to manually anonymize and deidentify datasets – or remove information that links a data record to a specific individual – are time-consuming, labor-intensive, and prone to errors. This can delay projects and lengthen the iteration cycle time for development of ML models. Synthetic data can overcome many of these pitfalls by providing faster, cheaper, and easier access to data that is similar to the original source, suitable for use, and protects privacy.

"Moreover, if manually anonymized data is combined with other publicly available data sources, there's a risk it could inadvertently reveal information that could lead to data reidentification, thus breaching data privacy," Woodward says. "Leaders can use techniques such as differential privacy to ensure any synthetic data generated from real data is at very low risk of deanonymization."

BROAD BASE

The bottom line: How can you optimize the value of data with the capabilities offered by GenAI? Build specific capabilities into the data architecture to support the broadest set of use cases. That's necessary because the scope of value has gotten much bigger because of GenAI's ability to work with unstructured data such as chats, videos, and code.

"This represents a significant shift because data organizations have traditionally had capabilities to work with only structured data, such as data in tables," McKinsey says. "Capturing this value doesn't require a rebuild of the data architecture, but the CDO will need to focus on two clear priorities."

The first is to fix the data architecture's foundations. While this might sound like old news, the cracks in the system a business could get away with before will become big problems with GenAI. Many of the advantages of GenAI will simply not be possible without a strong data foundation.

To determine the elements of the data architecture on which to focus, the CDO could start identifying the fixes that provide the greatest benefit for the widest range of use cases, such as data-handling protocols for PII, since any customer-specific GenAI use case will need that capability.

The second is to determine which upgrades to the data architecture are needed to fulfill requirements of high-value use cases. The key issue: how to cost effectively manage and scale the data and information integrations that power GenAI use cases.

"If they are not properly managed, there is a significant risk of overstressing the system with massive data compute activities, or of teams doing one-off integrations, which could increase complexity and technical debt," McKinsey says. "These issues can be further complicated by the business's cloud profile, which means CDOs must work closely with IT leadership to determine compute, networking, and service use costs."

Since we started this column with a corny corporate story, let's end with a slightly sobering saga: At a "family day" corporate party, Jim, the usually reserved accountant, had one too many pegs. By the end of the night, he was dancing on tables and confessing secrets. The next day, his colleagues had a field day mocking him. "No wonder alcohol is considered an effective solvent," Jim sighed. "Because it effectively dissolves marriages, families, and careers." DQ

Raju Chellam is a former Editor of Dataquest and is currently based in Singapore, where he is the Editor-in-Chief of the AI Ethics & Governance Body of Knowledge, and Chair of Cloud & Data Standards.

maildqindia@cybermedia.co.in