HPE GreenLake

AI's rallying cry: Seize the data — but only the good data

Five ways to optimize your AI strategies around data cleansing, preparation, and management

Enterprises are betting their future on AI, with nearly three-fourths of organizations prioritizing AI over all other digital investments today, according to Accenture.[1] And they're not wrong to go all in on the emerging technology, as many believe AI-driven strategies will speed innovation, boost productivity, and create new, robust revenue streams.

But winning with AI isn't a matter of luck. The spoils will go to companies that can harness the power of their data to get the most out of AI initiatives. Unfortunately, not all data is created equal. So-called dirty data — data that is outdated, incomplete, mislabeled, or insecure — can quickly corrupt the output of the most artfully coded machine learning algorithm.

"Where data quality is bad, you get a bad outcome, no matter how good your machine learning model is," says Chad Smykay, field chief technology officer and distinguished technologist at Hewlett Packard Enterprise. Smykay says the secret to good quality data is no secret at all: Your data needs to be prepared, cleansed, and well organized so data scientists and developers across the enterprise can use it. But producing high volumes of clean data to meet the performance requirements of today's data-intensive AI workloads is both time and labor intensive — huge challenges in a world where data talent is in short supply and speed is of the essence.

How do you overcome these challenges and optimize data quality? Smykay offers five concrete steps to speed your AI deployment through better data operations.

[1] "Reinventing Enterprise Operations," Accenture, May 2023

1.

Establish smart data governance policies

Every company needs a solid data governance foundation that ensures data is well defined and organized while it is being collected — before it is fed into an algorithm or used to train a large language model (LLM). These policies require companies to ask key questions like, "What type of data is this? Where did it come from? Does it apply to my business use case? How old is it? Where should it be stored?"

It's not just a matter of good housekeeping. Effective data governance ensures that the data driving your organization's AI models is not only properly defined but also secure and trustworthy. It also helps to future-proof the organization for privacy laws and other legal regulations likely to emerge in the years to come. These rules, such as the European Union's General Data Protection Regulation (GDPR) today, may further require companies to show their homework on where data originated and how it has been used to train algorithms.

While new software products promise to automate data governance, Smykay says data governance, at least at the outset, should be a people process aligned with technology, managed by a chief technology officer, chief data officer, or chief data and AI officer. The reason? Software most likely won't be able to discern what data will really be useful to achieve a company's AI business needs and goals.

"To get data insights from AI that are going to help you make data-driven decisions, it's critical that the data you're classifying, cleansing, and storing is relevant to your AI business use cases," he says. "That means people who understand how the organization will extract value out of AI need to be the ones that set the rules of the road."

2.

Simplify your data management with a global namespace

When it comes to optimizing your data for successful AI deployments, it's important to work to simplify the management of your global data estate. To achieve simplicity, Smykay recommends companies build a data fabric with a global namespace that manages all their organization's data. Think of it as a single snapshot encompassing every data asset across the distributed, hybrid computing environment, clearly organized with a unified and consistent naming system for the different data sets AI developers may want to use.

"A global namespace is a solution to give developers the easiest way to find data stored in multiple locations, whether it's in multiple public clouds, spread across data centers, or on the edge," says Smykay. "Wherever it's located, they can quickly see and access the data they need."

This 10,000-foot view not only creates a map of all existing data but can also eliminate data waste, says Matt Hausmann, director of marketing for HPE Ezmeral Software. "As much as 30% of an enterprise's data never gets touched," says Hausmann. "A global namespace can lead to more data being used because it becomes visible and accessible." That, in turn, can lead to more enterprise-wide data being tapped to create more accurate AI-powered data insights.

3.

Create a flexible, open, and portable data platform

In today's complex enterprise environment, data is diverse and dispersed among where it originates, where it's stored, and even what shape it takes. Whether structured or unstructured or internally generated or pulled from the internet, today's data is stored in a myriad of formats and must be accessed using multiple application programming interfaces (APIs) that may not speak to one another.

Scattered and scattershot data can pose problems, so consolidation is the best solution, says Smykay. Delivering all a company's diverse, dispersed data assets on a unified data fabric is a critical step toward getting quality data into developers' hands. A single open API data access platform that can support multiple data types gives developers the crucial ingredient they need to succeed: flexibility. "In the analytics and AI space, developers want a data platform that can leverage multiple APIs because it gives them the most flexible way to access and prepare the data they need, no matter what format it's in," Smykay says.

For Hausmann, adopting a unified API data platform is as much about preparing for the future of AI as it is about fueling the models of today. "In the AI space, we're seeing a constant cycle of new breakthroughs, from deep learning and image detection a few years ago to the rise of LLMs last year," he reflects. "A unified data platform with multiformat support for files, objects, streams, tables, and vectors lets you keep evolving your data strategy to meet the future needs of your business."

4.

Build out your data estate's scalability

Data is the fuel that powers AI model training and outputs. Meeting the performance challenges of data-intensive workloads, like the huge volumes of data needed to train generative AI's LLMs, requires massive storage capacity. To accelerate AI data collection, preparation, and deployment, companies need a scalable data management solution — one that spans the hybrid cloud, says Smykay.

For the exabytes of data needed to power today's AI models, hybrid is one of the most affordable and scalable data management strategies, he explains. "Traditional, large hyperscaler solutions in the public cloud have just become too expensive, especially for these types of AI workloads," Smykay says. "Hybrid cloud can meet scalability demands by spreading out assets wherever data is generated or ingested, from an on-premises private cloud to the edge."

Wrangling this data is becoming a huge challenge because the sheer volume of data being generated is accelerating at lightning speed. The amount of data created annually is projected to more than double in size from 2022 levels by 2026, with the amount of enterprise data growing twice as fast as consumer data, according to IDC.[2]

Smykay adds that hybrid cloud's data plane strategy is also flexible and forward looking, best positioned to support whatever innovative AI models arise down the road.

[2] "Global Datasphere, Data Marketplaces, and Data as a Service," IDC, August 2023

5.

Accelerate data pipelines to fuel AI models faster

In terms of value, data comes with an expiration date, says Hausmann. "The value of data typically decays quickly over time. It starts losing value the moment it is created, and depending how you're leveraging it, it could lose most of its value in seconds, hours, days, or weeks," he explains. Companies need to ensure that data is available as quickly as possible to get the most analytical benefits from it, he advises.

But despite this urgent need for speed, 76% of organizations say their current data management processes are unable to keep pace with their business needs.[3] Streaming solutions baked into a unified data platform can help accelerate the process, however. "You've built your AI models, and you've trained them. Now, the challenge is to keep them fed," says Smykay. "The good news is you can keep the data flowing in real time if you change up how it is ingested on the back end with a streaming format."

Faster data ingestion can be invaluable for ensuring data quality, Smykay notes, because it gives developers a shorter lead time for data cleansing and preparation. While automated tools are emerging to help prepare data for AI models, developers and data scientists know that data quality control is too important to outsource to an algorithm. But with the right data solutions in place, companies can position themselves to optimize high-quality data for fast, flexible, and safe AI deployment.

Data growth is not slowing down. In fact, as AI and business analytics continue to evolve, data growth is likely to speed up even further. It's critical to prepare for this reality. When planning and building your organization's data foundation, a data fabric is key. It will allow you to incorporate data spanning your hybrid cloud, support a wide range of data formats, scale with your business, and help your organization build a robust data governance strategy that ensures security and trust.

[3] "Being a Data-First Leader Continues to Matter," Enterprise Strategy Group, October 2023

Learn more at
HPE.com/AI

Visit **HPE GreenLake**

Chat now

Hewlett Packard Enterprise

a50010082ENW