TechTarget.com/searchdatamanagement

https://www.techtarget.com/searchdatamanagement/feature/Top-trends-in-big-data-for-2021-and-beyond

Top trends in big data for 2025 and beyond

By Donald Farmer

Global forces, both technical and nontechnical in nature, are reshaping the big data landscape. For example, the continuing economic recalibration in the post-pandemic era has pushed organizations to seek more predictable costs and tangible returns from their data management and analytics investments. Similarly, IT teams are looking for greater flexibility in how they build out data architectures to avoid vendor lock-in and budget overruns.

Meanwhile, the regulatory environment has matured significantly. The patchwork of data privacy laws that characterized the early 2020s has evolved into more standardized -- or at least more pervasive -- legal frameworks. In addition, the increased use of artificial intelligence is pushing organizations to fundamentally rethink how they handle and analyze sensitive information. Privacy is no longer just a compliance checkbox item; it's becoming deeply embedded in the technical fabric of data analytics.

We're also seeing the convergence of several technological breakthroughs. The mainstream adoption of large language models (LLMs) and generative AI tools has democratized access to advanced analytics capabilities, while advances in the development of multimodal AI models and quantum computing are beginning to influence long-term strategic planning.

Collectively, these forces will drive ongoing changes to data strategies and big data environments. Here's a look at seven significant big data trends and their implications for organizations through 2025 and into the future.

1. AI-powered analytics and agentic operations

Over the past two years, AI capabilities have changed big data analytics -- and they continue to do so. As mentioned above, AI brings sophisticated data insights within reach of more users in organizations. Meanwhile, data scientists and other skilled analysts are finding they can use AI to work more efficiently with large data sets.

For example, automated data preparation enables consistent data quality on a very large scale. It detects and corrects data issues, standardizes formats and identifies potential integration points without human intervention. This automation is particularly helpful as organizations struggle to cope with growing data volumes and diverse data sources, from IoT devices to social media feeds and more.

Also, neural network algorithms and LLMs can now identify subtle patterns and correlations in data that would be difficult to detect through traditional analysis methods. Not only is more sophisticated data analysis possible; it also can be directed by business users through natural language processing functionality or chatbots, which act as interfaces to the data or as copilots that help users surface insights.

Looking forward, AI will be increasingly integrated into analytics tools, data management workflows and business operations. AI-powered systems will move toward being able to autonomously monitor data, identify significant patterns and either take actions themselves or alert business stakeholders. These agentic AI experiences promise new levels of efficiency and data-driven automation.

The increased integration of AI and big data comes with challenges, including data governance, AI model management and the need for responsible and ethical AI. Organizations will need to take steps to reduce AI bias, maintain data privacy and ensure that AI-driven analytics applications produce reliable outcomes.

2. Increased focus on privacy-preserving analytics

The use of AI in decision-making processes, often involving data specific to individual customers or patients, is increasing the demand for what's called *privacy-preserving analytics*: technologies and methods that enable data to be analyzed without exposing sensitive or personally identifiable information.

Governance and compliance teams are necessarily concerned about unauthorized access to private information, such as health records, financial data, purchase history and location history. There's a broader issue, too: AI models trained on sensitive data sets might inadvertently amplify biases present in the data. But using decentralized or obfuscated data reduces the risk of privacy exposures and bias amplification, while still enabling effective data analysis.

Differential privacy is one widely used approach. It introduces controlled noise -- in the form of slight changes to data values -- into data sets or query results to obscure individual-level data points while preserving the overall usefulness of the information.

Another popular technique is federated learning, which enables AI and machine learning models to be trained across decentralized data sources without moving the raw data to a central server. Instead, a model learns and then analyzes data in separate processes on local devices or systems, and only the aggregated updates, without sensitive details, are shared with a coordinating server. This method is particularly valuable in industries where data security and privacy are critical for both business and regulatory reasons -- healthcare and financial services, for example.

Expect to see these kinds of techniques increasingly included as options in commercial data and analytics platforms that support big data applications.

3. Cloud repatriation and use of hybrid cloud architectures

The movement of data and applications to the cloud has felt like an unstoppable trend for many years, and the use of public cloud services is still growing strongly. However, a countertrend has emerged in the form of cloud repatriation: Companies are selectively moving certain workloads, including big data ones, back to on-premises data centers or to private clouds instead of public cloud environments.

This doesn't mark a wholesale rejection of cloud computing; rather, it reflects a more mature and nuanced strategy. Cost management is a key driver: Many companies have overrun their cloud budgets, especially for compute-intensive workloads such as AI and machine learning. In these scenarios, the pay-as-you-go model of public cloud services can lead to unanticipated levels of spending, especially if usage rapidly increases. And if there's one thing CFOs dislike more than high costs, it's unpredictability.

Some organizations that run specialized data workloads and operate under strict regulatory frameworks are also looking at repatriation. For example, financial services firms and healthcare companies are seeking to better manage compliance and data sovereignty requirements with carefully orchestrated hybrid cloud environments that include a mix of cloud and on-premises systems.

4. Data mesh deployments to decentralize data architectures

At its core, data mesh is a data management strategy that's both architectural and organizational in nature. It decentralizes data ownership from corporate IT to individual business domains, such as finance, marketing, HR and operations. Each domain acts as its own data organization, producing and maintaining data products -- ready-to-use data sets, models, dashboards and more -- that are treated as key business assets.

By prioritizing domain-driven design, data mesh enables the teams closest to sets of big data to take control of meeting their particular data preparation and analytics needs. Doing so can reduce the bottlenecks often experienced in centralized data management models.

For a data mesh approach to be successful, the domain teams must have the skills and tools to effectively manage their own data products, even if IT supports them in these efforts. There also must be clear organizational accountability for data management processes. Technically, the success of a data mesh relies on good metadata management and data discoverability. Organizations often deploy data catalogs and self-service analytics tools to make data assets easy to find, understand and use.

There's also a synergy between cloud repatriation and data mesh architectures. While repatriation optimizes workload allocation across hybrid cloud environments, data mesh provides the architectural framework to make that more manageable and effective. Using the two approaches together enables highly flexible environments in which data products can be hosted wherever makes most sense.

For example, a financial services provider could store transaction data on-premises for regulatory compliance purposes while making an anonymized data set available as an analytics data product in the cloud for increased ease of use. In this scenario, a data mesh framework ensures interoperability and accessibility between such environments, without the need to add another data layer.

5. Data lakehouses as the dominant big data platform

Data lakehouses will likely consolidate their position as the dominant architecture for big data analytics in 2025, having proven to be efficient, scalable and cost-effective. A data lakehouse platform combines the flexibility of data lakes for working with raw and often unstructured or semistructured data with the reliability and performance of traditional data warehouses that store consolidated sets of structured data. This integrated approach eliminates the need for separate systems to support data science workloads on the one hand and basic business intelligence reporting on the other.

The single-copy architecture of a data lakehouse reduces data redundancy by avoiding multiple versions of the same data across different platforms. As a result, data engineers can streamline data workflows, and data storage costs can be reduced. The support for diverse data types -- from highly structured relational tables to images and text -- also makes data lakehouses an ideal platform for AI, predictive analytics, real-time data analysis and other advanced analytics applications.

For all these reasons, data lakehouses are likely to be at the heart of enterprise analytics initiatives and big data environments well into the future.

6. The rise of open table formats

A key development as part of the data lakehouse ecosystem is the rise of open table formats, with Apache Iceberg emerging as the most widely used one. Other available options include Delta Lake and Apache Hudi.

Open table formats are designed to manage large-scale tabular data for analytics workloads in a standardized way. They're especially relevant in data lakes or data lakehouses where large data sets need to be stored, queried and updated efficiently.

The table formats provide features such as cross-platform compatibility, transaction support and schema evolution. The latter is particularly important: It's the process of managing and adapting data structures as they change over time, while still maintaining data integrity and backward compatibility.

Open table formats also reduce the risk of vendor lock-in for organizations. By using one, they can avoid finding themselves stuck on a proprietary -- and often expensive -- big data platform because of the difficulty of migrating to a new architecture.

7. Preparations for quantum computing

While quantum computing is still in its early stages, its potential is already shaping the long-term thinking of enterprises in industries that have highly demanding computational needs, such as pharmaceuticals and financial services.

Quantum systems aim to revolutionize complex problem-solving and large-scale simulations by tackling data processing challenges that are beyond the current capabilities of classical computers. Today, organizations experimenting with quantum are looking to model drug and molecular interactions, among other early uses. But the same techniques could be readily applied to training AI models or to complex business scenarios, such as supply chain optimization and financial planning simulations.

Although practical quantum applications remain on the horizon, there's an increasing need to anticipate this new technology playing a role in the future of big data initiatives. That includes upskilling staff in preparation and exploring what hybrid classical-quantum computing approaches would look like. These efforts are likely to become even more pressing in 2025: Expect some R&D breakthroughs that will accelerate the need for quantum readiness in big data environments.

Editor's note: *This article was updated in January 2025 to reflect new trends in big data.*

Donald Farmer is a data strategist with 30-plus years of experience, including as a product team leader at Microsoft and Qlik. He advises global clients on data, analytics, AI and innovation strategy, with expertise spanning from tech giants to startups.

27 Jan 2025