IDC PERSPECTIVE

# Accelerate Innovation and Sustainable Competitive Advantage with a Solid Data Strategy for AI

Ritu Jyoti

## EXECUTIVE SNAPSHOT

## FIGURE 1

### Executive Snapshot: Solid Data Strategy for AI

This IDC Perspective covers the importance of a solid data strategy for AI. It includes the current challenges and provides guidance on putting together a data strategy for AI. It covers data strategy for AI goals and components like data defense versus offense, single source multiple versions, data architecture for diverse data landscapes, intelligent data fabric, feature stores, organizational roles, and data/AI governance.

#### Key Takeaways

- If you have a business, you have data — but data by itself won't let you optimize and improve your business.
- You need to exploit the power of AI to reap significant benefits by seeping AI throughout the value chain of a business. You need a data strategy for AI if you want to turn data into business value.
- There is direct correlation between primary business objectives for AI initiatives and superior business outcomes. As per IDC's *AI StrategiesView 2021 Survey,* early adopters reported 39% improvement in CX and 33% improvement in employee efficiency and accelerated innovation with the roll out of AI solutions in 2020, which is a double-digit surge in improvement of business outcomes compared with 2019.

#### Recommended Actions

- Create workflow of the life cycle of bringing in new data sources into your organization from testing and buying to seamless integration with existing internal data sets and processes. Make sure the process is cross-functional across IT, procurement, legal, compliance, and security.
- Get employees' buy-in and trust for your data strategy for AI with inclusiveness and transparency.
- Select a responsible AI/ML platform with native support for all data types. Embrace an intelligent data fabric that helps automate and enforce universal data and usage policies across hybrid data and cloud ecosystems; automate how data is discovered, cataloged, and enriched for users; and automate how you access, update, and unify data spread across distributed data and cloud landscapes without the need of doing any data movement or replication.

Source: IDC, 2021

November 2021, IDC #US48354321

## SITUATION OVERVIEW

In an enterprise climate where disruption is the norm, businesses live or die by the ability to meet constantly evolving conditions. Those that stay ahead of change — that anticipate it, evolve with it, and even help facilitate it — experience lasting success. Those that fail to adapt don't stay afloat. If you have a business, you have data — but data by itself won't let you optimize and improve your business. Artificial intelligence (AI) is at the heart of digital disruption, and data is foundational and critical to the success of AI Initiatives. While you need to exploit the power of AI to reap significant benefits by seeping AI throughout the value chain of a business, you need a data strategy for AI if you want to turn data into business value.

In the next few years, industry definitions and boundaries will get blurry, and organizations will move from products- to platform-based business models. Organizations will need to redefine their business model, reevaluate their supply chain, and reimagine their customer journey, and AI will be integral to all these efforts.

Even though most big players across industries have "redefined" their business model, chalked out road maps to build/govern their partner ecosystems, and are trying to leverage external/internal data to become aggregators, it is "artificial intelligence" that holds the key to this competitive advantage. Redefinition of business model leads to identification and fulfillment of additional revenue streams from the strengthened partner network. To move ahead of its competitors, a leading market research agency evolved its business model and built a strong partner ecosystem with its key FMCG clients by providing a platform for brand managers to drive product recommendations through AI and deep learning.

Reinventing supply chain implies integrating digital technologies like autonomous vehicles, advanced robotics, and 3D printing to build "supply chains for the future" so that companies can maintain their competitive advantage for the next 5–10 years. A lot of major manufacturing giants all over the world are trying to use autonomous driverless vehicles and drones for transportation of goods to cut down the supply chain costs. Leading American manufacturing conglomerates have leveraged AI for "smart predictive maintenance" to monitor asset health.

Next-gen customers are impatient, well-informed millennials and customers that prefer self-service. Organizations across industries are leveraging AI to reinvent their approach to reconnect with their increasingly demanding customer base by enhanced customer experience management and omni-channel marketing strategy. A leading telco in the United States has pioneered next-gen customer service management by using conversational chatbots and virtual agents to streamline customer interactions. A major telco in Asia identifies and optimizes marketing spend by eyeball tracking and image recognition.

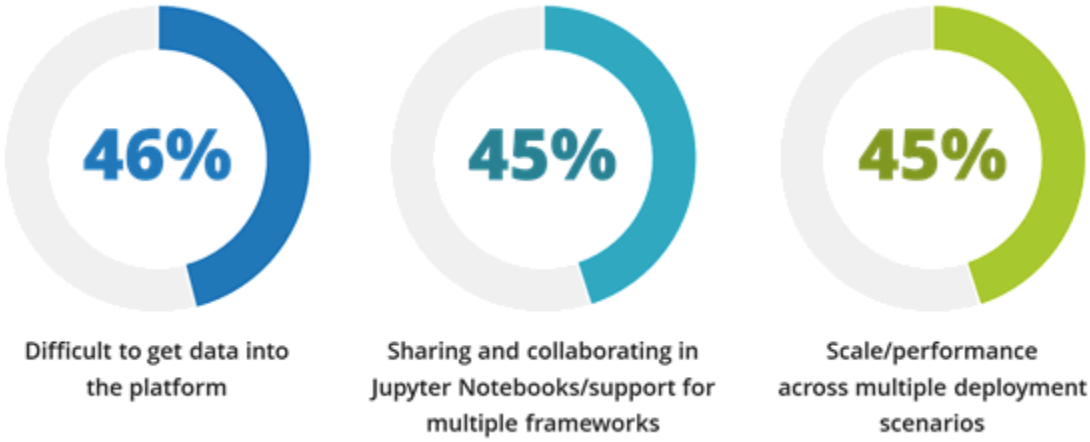
### Why Do You Need a Data Strategy for AI?

As per IDC's *AI Strategies View 2021 Survey,* conducted in April 2021, accelerating innovation, improving operational efficiency, and improving customer experience are identified as the primary business objectives for AI initiatives. There is direct correlation between primary business objectives for AI initiatives and superior business outcomes. Early adopters reported *39% improvement* in customer experience (CX) and *33% improvement* in employee efficiency and accelerated innovation with the rollout of AI solutions in 2020, which is a *double-digit surge in improvement of business outcomes compared with 2019.*

While it is clearly established that AI is the key to innovation and sustainable competitive advantage, a majority of the organizations worldwide are still struggling to move their pilots to production. As per the same study, cost of the solution, lack of skilled personnel (i.e., talent), decision criteria for the solution, lack of adequate volume/quality of data, and lack of operations and trustworthy AI are some of the leading challenges for implementing AI solutions. From a model development perspective, difficulty in getting data into the ML platform is rated to be the top challenge (see Figure 2). Unfortunately, due to the aforementioned challenges, organizations are spending more time on tasks beyond actual data science. For example, as per the same study, organizations spend roughly 21% of the time in data collection/preparation (see Figure 3).

## FIGURE 2

### Model Development Challenges

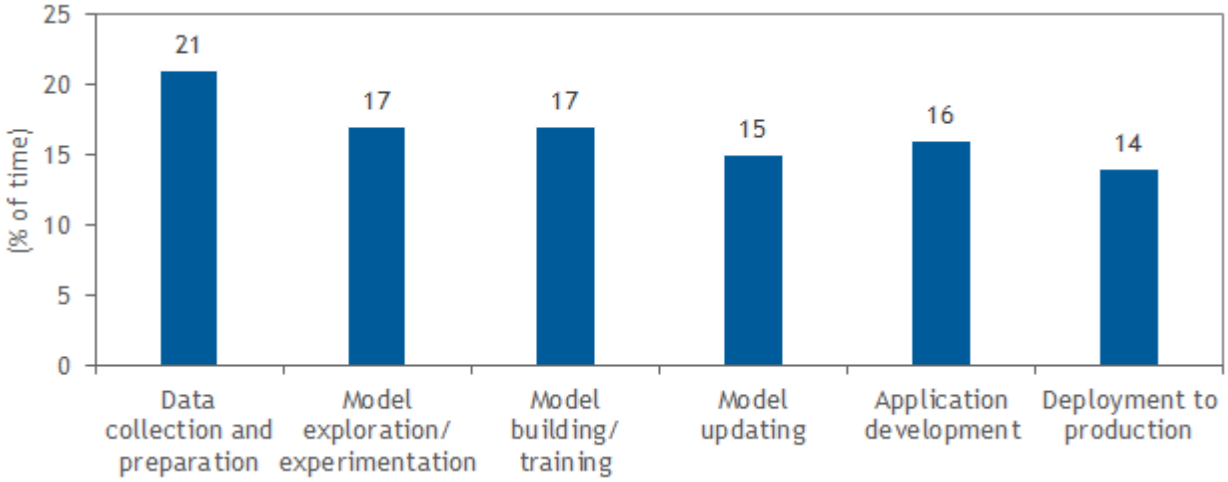



n = 2,000

Source: IDC's *AI StrategiesView 2021 Survey,* CY21

## FIGURE 3

### Time Spent by Workstreams



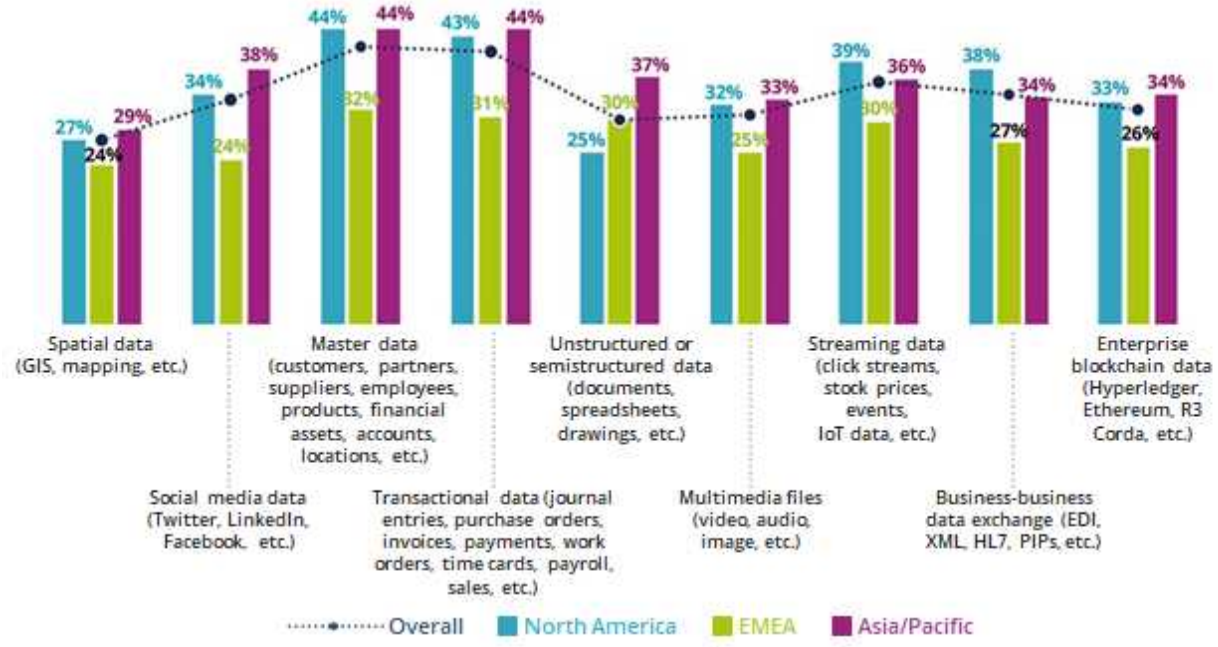

n = 1,366

Source: IDC's *AI StrategiesView 2021 Survey*, CY21

Similarly, while most of the transformative impact of AI can be realized by use of unstructured data, its use is largely untapped (see Figure 4). Data continues to be siloed and distributed (see Figure 5).

## FIGURE 4

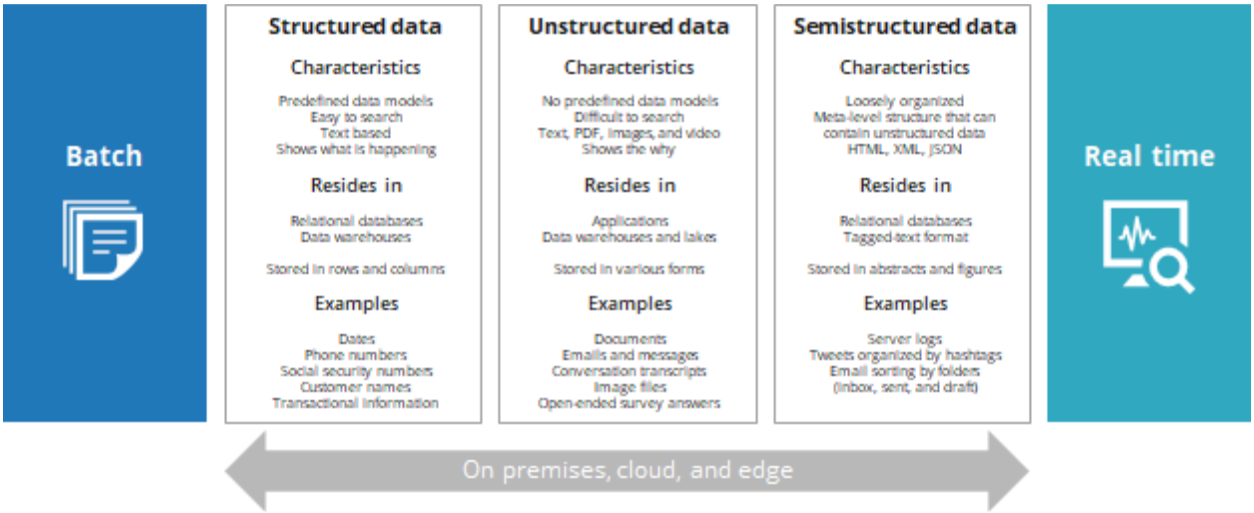
### Types of Data Used for AI/ML Solutions




Source: IDC's *AI StrategiesView 2021 Survey*, April 2021

## FIGURE 5

### Siloed and Distributed Data




Source: IDC, 2021

Hence, the transformative power of artificial intelligence needs to start with a solid data strategy. The data strategy should encompass a foundational data architecture that can address the complexity of today's diverse data landscapes. Let us start with the goals and components of a data strategy for AI.

## Data Strategy for AI: Goals and Components

While most companies recognize that their data is a strategic asset and foundational for AI initiatives, many are not taking full advantage of it to get ahead. The goal should be to move from using descriptive analytics ("what happened?") to prescriptive analytics ("how can we make it happen?"). Essentially from information and insight to foresight.

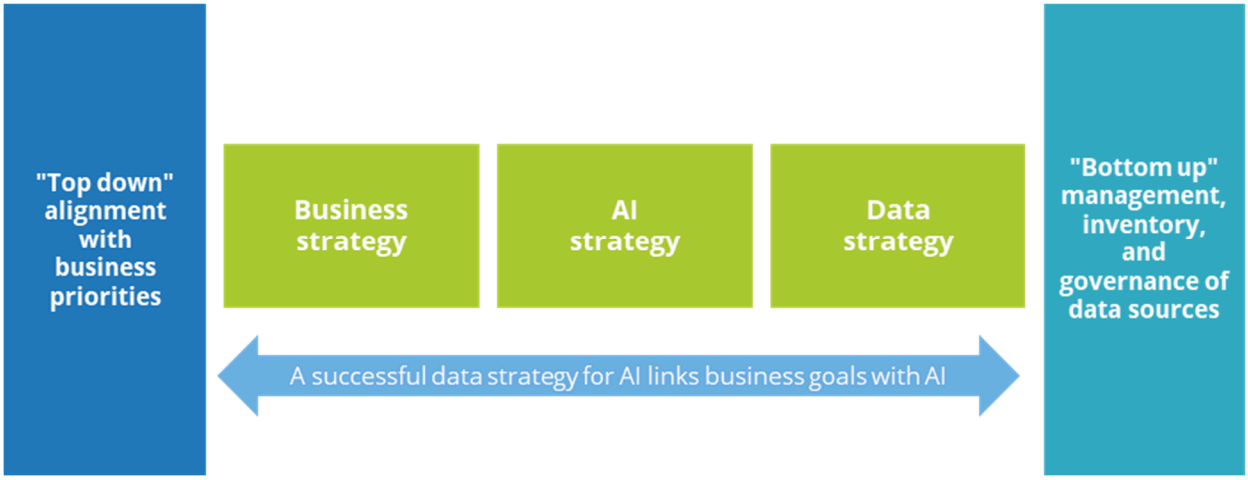
Data strategy for AI refers to the tools, processes, and rules that define how to manage, analyze, and act upon business data. A data strategy for AI helps you make informed decisions based on your data. It also helps you keep your data safe and compliant. In short, a business without a data strategy for AI is poorly positioned to operate efficiently and profitably or to grow successfully. Your data strategy for AI should reinforce and advance your overall business strategy, which refers to the processes you use to operate and improve your business. The first step of defining the business requirements is to identify a champion, all stakeholders, and SMEs in the organization. The champion of the data strategy for AI is the executive leader who will rally support for the investment. Stakeholders and other SMEs will represent specific departments or functions within the company.

Next is to define the strategic goals and tie department activities to organization goals. It's natural for goals to exist at the company and department level, but the stated goals for both levels should sync up. These objectives are most effectively gathered through an interview process that starts at the executive level and continues down to department's leaders. Through this process, you will discover what leaders are trying to measure, what they are trying to improve, questions they want answered, and, ultimately, the KPIs to answer those questions.

By starting with gathering and documenting the business requirements, you can overcome the first roadblock to many AI projects: knowledge of what the business is trying to accomplish (see Figure 6).

### FIGURE 6

**Building an Enterprise Data Strategy for AI**




Source: IDC, 2021

## Goals

To craft an effective data strategy for AI, you need to work toward several goals:

- **Innovation:** Any successful business creates new value or efficiency through innovation. Innovation should be a central goal as you create and implement data strategy for AI.
- **Addressing the needs of users:** Your data strategy for AI must support and empower your users — anyone within your organization who helps power the business.
- **Addressing risk and regulations:** An effective data strategy for AI must address your business' data security risks and compliance requirements, which can vary widely between different types of industries.

And, because your IT department will be responsible for implementing and overseeing the tools and technologies that power your data strategy for AI, consider IT resources and capabilities when setting your goals. Also, ensure that you have the goal to keep the costs of IT resources down.

## Components

When it comes to the components of an effective data strategy for AI, having a CDO and a data management function is a start. However, neither can be fully effective in the absence of a coherent strategy for organizing, governing, integrating, and deploying an organization's information assets to seamlessly feed into AI proof of value. Given the high business impact of an AI initiative, ensuring smart data readiness for AI is the responsibility of all C-suite executives, starting with the CEO. The following are the critical components of an effective data strategy for AI:

- **Data for AI — defense versus offense:** Data for AI defense and offense are differentiated by distinct business objectives and the activities designed to address them. Defense is about minimizing downside risk. Activities include ensuring compliance with regulations (such as rules governing data privacy and the integrity of financial reports), using machine learning (ML) to detect and limit fraud, and building systems to prevent theft. Data for AI offense focuses on supporting business objectives such as increasing revenue, profitability, and customer satisfaction. It typically includes activities that generate customer insights or integrate disparate customer and market data to support managerial decision making.

  Offensive activities tend to be most relevant for customer-focused business functions such as sales and marketing and are often more real time than is defensive work, with its concentration on legal, financial, compliance, and IT concerns. Every company needs both offense and defense to succeed, but getting the balance right is tricky. Putting equal emphasis on the two is optimal for some companies. But for many others it's wiser to favor one or the other. Some company or environmental factors may influence the direction of data strategy: strong regulation in an industry (e.g., financial services or healthcare) would move the organization toward defense and strong competition for customers would shift it toward offense. The challenge for CDOs and the rest of the C-suite is to establish the appropriate trade-offs between defense and offense and to ensure the best balance in support of the company's overall data strategy for AI. A company's position on the offense-defense spectrum is rarely static. A company's industry; competitive, regulatory environment; and AI strategy will inform its data strategy for AI. For example, hospitals operate in highly regulated environments where data quality and protection are paramount. They emphasize on defense versus offense. Banks are heavily regulated and require strong data defense. But they also operate in dynamic markets and so typically devote equal attention to offense. Retailers are less regulated, work

with limited sensitive personal data, and must react rapidly to competition and market changes. They typically emphasize offense over defense.

- **Single source, multiple versions:** A company's data architecture for AI describes how data is collected, stored, transformed, distributed, and consumed. It includes the rules governing structured formats, such as databases and file systems, and the systems for connecting data with the business processes that consume it. Many organizations have attempted to create highly centralized, control-oriented approaches to data and information architectures. These top-down approaches are often not well suited to supporting a broad data strategy for AI. Although they are effective for standardizing enterprise data, they can inhibit flexibility, making it harder to customize data or transform it into insights/foresights that can be applied strategically. A more flexible and realistic approach to data architecture for AI involves both a single source of truth (SSOT) and multiple versions of the truth (MVOTs). The SSOT works at the data level; MVOTs support the management of information. The SSOT is a logical, often virtual and cloud-based repository that contains one authoritative copy of all crucial data, such as customer, supplier, and product details. It must have robust data provenance and governance controls to ensure that the data can be relied on in defensive and offensive activities, and it must use a common language — not one that is specific to a particular business unit or function. Thus, for example, revenue is reported, customers are defined, and products are classified in a single, unchanging, and agreed-upon way within the SSOT. Not having an SSOT can lead to chaos. An SSOT is the source from which multiple versions of the truth are developed. MVOTs result from the business-specific transformation of data into information — data imbued with "relevance and purpose." Thus, as various groups within units or functions transform, label, and report data, they create distinct, controlled versions of the truth that, when queried, yield consistent, customized responses according to the groups' predetermined requirements.

  Consider how a supplier might classify its clients Bayer and Apple according to industry. At the SSOT level, these companies belong, respectively, to chemicals/pharmaceuticals and consumer electronics, and all data about the supplier's relationship with them, such as commercial transactions and market information, would be mapped accordingly. In the absence of MVOTs, the same would be true for all organizational purposes. But such broad industry classifications may be of little use to sales, for example, where a more practical version of the truth would classify Apple as a mobile phone or a laptop company, depending on which division sales was interacting with. Similarly, Bayer might be more usefully classified as a drug or a pesticide company for the purposes of competitive analysis. In short, multiple versions of the truth, derived from a common SSOT, support superior decision making.
- **Data architecture for AI:** A first step in defining your data architecture for AI is determining what data sets exist among business units across the company. Enterprise data catalogs are useful tools for this purpose. If you don't have a data catalog, review data sources with your team and the users who work with the data. You need a data pipeline to ingest raw data from disparate sources and replicate it to a destination for use by data scientists/engineers. Data identification, ingestion, storage, and analysis are all parts of a data architecture for AI. Considerations like whether you proceed with extract, transform, and load (ETL) or extract, load, and transform (ELT) will be fundamental. As per our research, to support proofs of value for AI, more and more organizations are proceeding with ELT. The ability to document and implement your data architecture for AI is essential for a consistent, predictable data strategy for AI. It also makes it easier to scale your data operations as your needs change. It's unlikely that all data will be available within the organization and that it already exists in a place that's accessible. So you need to work backward to find the source. Data is an issue in most AI projects. Synthetic data can help change this situation. Synthetically generated data can help

companies and researchers build data repositories needed to train and even pretrain machine learning models. Make sure you incorporate the same sets of guardrails for synthetic data as for the real data.

The next fundamental question is whether you should federate or virtualize data. While data virtualization provides a single point of access to data that hides its distributed or heterogenous details, data federation is a virtual database that provides a common data model and access point for distributed data and heterogenous data sources. We expect data virtualization to be the norm for AI as data has gravity, and subject to use cases, data needs to be available where AI training and inferencing occur. We are seeing adoption of intelligent data fabric that include automated data governance and protection, self-service data consumption, and automated data integration. Intelligent data fabric uses AI to automate complex data management tasks and universally discover, integrate, catalog, secure, and govern data across multiple environments. Users will be able to benefit from the intelligent unification of diverse data types and architectures — like data lakes, data catalogs, data warehouses, and other data integration platforms — into one common data foundation without the need to copy or move information. Users across different personas and roles will also be able to search, understand, and consume data throughout the organization through a single point of access.

Today, AI-powered high-performance universal query engines simplify the data landscape and empower users to easily query data across hybrid cloud, multicloud, and multivendor environments. AI-powered automated cataloging helps overcome the challenges faced by managing a complex hybrid and multicloud enterprise data landscape and helps ensure that data consumers can easily find and access the right data at the right time regardless of location. Last, AI-powered automated privacy intelligently automates the identification, monitoring and, subsequently, enforcement of policies on sensitive data across the organization.

A "feature store" is a platform on which companies can store, process, and maintain the features that make up an AI model, along with the description of what the feature is, what it does, and how it was built. As a refresher, a feature refers to the concept of a relevant data signal, not the data itself. Feature stores have become the key piece of data infrastructure for machine learning platforms. Feature stores emerged several years ago with big tech companies and advanced business users of AI, such as Uber Technologies Inc., among the first to create them. The concept is being embraced by more companies as they accelerate AI adoption. So make sure you include a feature store in your data strategy for AI early in the project life cycle, even if the initial ML organization is small.

- Organizational roles: A data strategy for AI should include attention to organizational roles by documenting who does what with the data, to facilitate collaboration and avoid duplication. Four main types of users typically implement and enforce data strategy for AI:
  - Data engineers, who oversee the data pipeline and are responsible for building an efficient, reliable data architecture
  - Data scientists, who work with data that the pipeline delivers
  - Data analysts, who specialize in analyzing and interpreting data
  - Business managers, who help manage data operations and review data reports

  When coordinating roles, consider everyone in the organization who uses data in any way, even if working with data is not a primary part of their job responsibilities. For example, an account manager who records customer information has a role to play in data collection and a sales manager may need insights to help plan the next marketing campaign. Your data

strategy for AI should document the roles of each team member or group. In addition, when a business maintains multiple data sets, its data strategy for AI should specify who "owns" which data, meaning who is responsible for storing, safeguarding, and interpreting the different data sets.

- **Data/AI governance:** The foundation for effective data management for AI is data and AI governance, which establish the processes and responsibilities that ensure the quality, security, fairness, explainability, lineage, and transparency of the data used across an organization. For example, data governance for AI might specify that a manager must archive data in an offline location if it's no longer in daily use. Or a data/AI governance policy may require data encryption and differential noise to bolster security and privacy.

  You should update data governance policies as your business needs change. A data governance program will ensure that:

  - Calculations used across the enterprise are determined based on input from across the enterprise.
  - The right people have access to the right data.
  - Data lineage (where did the data originate and how was it transformed since that origination) is defined.

AI governance takes leadership, sometimes helping navigate through difficult conversations.

## ADVICE FOR THE TECHNOLOGY BUYER

Data strategy for AI will vary greatly depending on the size, nature, and complexity of your business and AI strategy. To accelerate innovation and time to value and enjoy sustainable competitive advantage, technology buyers are advised to:

- Build the talent pool of industry domain and technical expertise.
- Get employees' buy-in and trust for your data strategy with inclusiveness and transparency.
- Create a workflow of the life cycle of bringing in new data sources into your organization from testing and buying to seamless integration with existing internal data sets and processes.
- Make sure the process is cross-functional across IT, procurement, legal, compliance, and security.
- Select a responsible AI/ML platform with support for all data types.
- Embrace an intelligent data fabric that helps:
  - Automate and enforce universal data and usage policies across hybrid data and cloud ecosystems.
  - Automate how data is discovered, cataloged, and enriched for users.
  - Automate how you access, update, and unify data spread across distributed data and cloud landscapes without the need of doing any data movement or replication.

## LEARN MORE

### Related Research

- *IDC FutureScape: Worldwide Artificial Intelligence and Automation 2022 Predictions* (IDC #US48298421, October 2021)

- *Market Analysis Perspective: Worldwide Artificial Intelligence Software, 2021* (IDC #US48243221, September 2021)
- *Manage AI/ML Business Risks and Thrive with Trustworthy AI* (IDC #US48235521, September 2021)
- *AI StrategiesView 2021 Premium: Banner Tables* (IDC #US47638621, April 2021)
- *Feature Stores: Critical for Scaling ML Initiatives and Accelerating Both Top-Line and Bottom-Line Impact* (IDC #US47223320, January 2021)

## Synopsis

This IDC Perspective covers the importance of a solid data strategy for AI. It includes the current challenges and provides guidance on putting together a foundational data strategy for AI. It covers data strategy for AI goals and components like data defense versus offense, single source multiple versions, data architecture for diverse data landscapes, intelligent data fabric, feature stores, organizational roles, and data/AI governance.

"Artificial intelligence (AI) is changing the rules of the game for almost every industry. AI applications are fueled by data to function and provide outputs. The success of an AI model is highly dependent on the relevance and accuracy of the data that is fed into it. Hence creating an appropriate data strategy is a prerequisite for building and deploying a successful AI model." – Ritu Jyoti, group vice president, AI and Automation research at IDC. "Without a data strategy for AI, an organization's efforts will be greater than necessary, risks will be magnified, and chances of success will be reduced."

## About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

## Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com

---