Microsoft Azure

Maximize Data Value

Preparing your data for AI innovation

Contents

Introduction

AI has the potential to revolutionize industries, automate processes, and uncover insights at a pace that was previously unimaginable. As the excitement around AI continues to build, business leaders realize the importance of ample and diverse data so that AI can effectively identify patterns and make informed decisions. Consequently, data professionals face a pressing challenge—ensuring that their data is primed and prepared to meet the increased demands of AI.

Data leaders understand that the success of AI initiatives hinges on the quality, relevance, timeliness, and accessibility of the underlying data. Poor data quality can easily derail AI initiatives, adding to timelines and increasing costs. As a result, many data leaders are looking for simple, streamlined, and impactful ways to cleanse, integrate, and enrich their data assets to meet the rigorous demands of AI algorithms.

By proactively addressing their data challenges, data professionals can lay the groundwork for successful AI adoption and empower their organizations to realize the full potential of AI investments. This e-book explores how you can reap the maximum benefits from your data by properly preparing it for AI and predictive analytics.

Data complexities and challenges

The impulse to start innovating with AI today is strong. However, it's important to ensure proper implementation from the start. That means addressing data challenges—including privacy, security, regulatory compliance, and potential biases, as well as data quality, an essential element of a solid AI foundation.

Without access to clean, secure, and real-time data, the value of AI output is naturally limited. But attaining data integration and quality is even more complex today because many legacy IT environments were built before the arrival of AI. As a result, data professionals often grapple with complexity as they build and scale their AI models.

This chapter covers two challenges that data pros can—and should—tackle now to prepare data for AI innovation: data movement and data duplication.

Data movement

To support business intelligence initiatives, organizations need clean accurate, and consolidated data, which might be extracted from various sources, transformed for analysis, and loaded into analytical tools or databases for reporting.

Moving data serves several essential functions. It consolidates information from disparate sources or systems, facilitating analysis, reporting, and decision-making. Data movement also supports data-sharing initiatives by allowing teams to disseminate information across departments, teams, or external partners so they can collaborate with their data. Adhering to data localization laws or residency regulations also necessitates data movement, as some data, by law, must be stored in specific geographic locations.

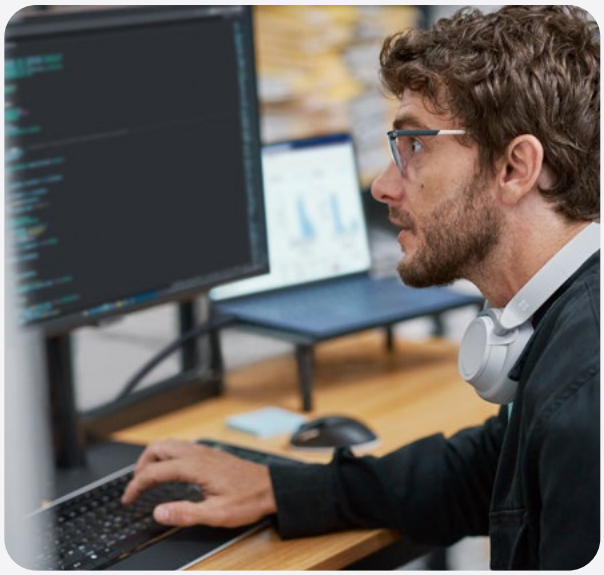
The ability to move data around is fundamental to many operations. However, due to several factors, time spent moving data can reduce the value gained from its use:

- **Latency and delays**
  Moving substantial data volumes introduces latency, delaying real-time applications and the decisions they drive.
- **Network bandwidth constraints**
  Data transfer strains network resources in disparate machine learning systems, especially with high-resolution media and data from sensors.
- **Data consistency**
  Maintaining consistency across replicated or distributed data is complex—and vital for accurate machine predictions.
- **Security and compliance**
  Data transit exposes sensitive information, necessitating encryption and secure protocols, and compliance may restrict cross-border data movement.
- **Costs and resource use**
  Data movement takes up computational resources and drives up costs, making the efficient allocation of resources challenging.
- **Low-performing models**
  Isolated data silos hinder machine learning workflows, slowing down insights and potentially leading to biased models.
- **Increased transformation overhead**
  Data transformation prior to machine learning adds overhead, especially when moving raw data to processing pipelines.

Data duplication

Several factors cause duplicate data. Integrating data from multiple sources or systems can result in duplicate data entries, with the same information stored across different databases.

Today, this poses significant hurdles for data teams looking to build AI and machine learning models:



- **Diminished accuracy and reliability**

  Duplicate data introduces inconsistencies and inaccuracies in machine learning models, skewing statistical analyses and causing biased results. Machine learning algorithms learn from data patterns. Duplicate entries can confuse those patterns, compromising predictive accuracy.

- **Increased strain on talent and resources**

  Trying to make sense of data—figuring out what's up to date and where the data originated, among other questions—can put undue strain on data professionals. It also diverts valuable resources away from innovation initiatives.

- **Heightened complexity and processing time**

  Deduplicating data is labor-intensive and diverts resources from essential tasks. AI can automate this, but it demands computational resources and time that could be used for tasks that add value.

- **Increased storage costs and retrieval delays**

  Storing duplicate data increases expenses, particularly in cloud environments. Due to redundant entries, querying, processing, and retrieving data from large data sets becomes sluggish.

- **Higher risk of erroneous output**

  Duplicate data can lead to incorrect conclusions, jeopardizing business strategies. Flawed data undermines the reliability of AI and machine learning models, producing unreliable outcomes.

- **Complex data management**

  Managing duplicate data in distributed systems is intricate, even with AI assistance. Advanced machine learning architectures can enhance deduplication accuracy.

Overcoming data challenges to accelerate AI innovation

By minimizing unnecessary movement and eliminating duplicate records, organizations can ensure that their AI models have what they need to realize the full potential of their data.

This streamlined approach enhances the efficiency of AI algorithms and mitigates the risk of errors, biases, and inconsistencies that might arise from redundant or fragmented data sets. Moreover, optimizing data management practices promotes data governance and compliance, fostering trust and confidence in the AI-driven insights generated from the prepared data. Reducing data movement and duplication is essential to laying the foundation for successful AI initiatives that extract meaningful insights and drive innovation.

Why preparing your data is a critical step to embracing AI

- All artifacts use the same data set without duplication or movement.
- Minimal data movement helps machine learning initiatives yield valuable results.
- Easy discovery and reuse of all data assets by all users drives greater efficiency.
- The efficacy of AI solutions is improved with accurate and reliable data.

Next, you'll explore the essential requirements for preparing your data.

Requirement #1: Assemble the necessary capabilities

To overcome data movement and duplication challenges, organizations need a comprehensive set of unique data capabilities that span data integration, transformation, streaming, querying, visualization, and collaboration.

Think of this requirement as the scene in an action movies when the team of heroes is assembled, and each hero brings their special abilities to the group. These unique data capabilities address data movement and duplication challenges, enabling organizations to derive actionable insights and drive innovation from their data assets.

These essential data capabilities help you build AI models that deliver the most possible value:

1. Data integration

Data integrated from various sources into a single, cohesive platform is easier to access and manage. Combining data involves extracting it from different systems, databases, and applications, transforming it into a consistent format, and then loading it into a single location to create a unified data estate.

Many organizations use data lakes and lakehouses to integrate and unify their data, because cloud-based storage and computing resources often provide a flexible and cost-effective method for enterprise data integration.

Benefits of data integration:

- **Unified view**

  A unified view of your data supports accurate and timely analysis, which can drive meaningful insights. Data integration combines information from multiple sources in a comprehensive dataset, so AI models can make more informed and accurate predictions that lead to better decision making across business functions.

- **Data quality**

  Integrating high-quality data into AI models significantly improves prediction accuracy and reliability. By ensuring that the data used for analysis is consistent, clean, and reliable, organizations can enhance the performance of their AI algorithms and mitigate the risk of errors or biases in decision-making processes.

- **Comprehensive insights**

  Integrated data provides a holistic view of an organization's operations, customer interactions, and market trends. By combining data from diverse sources, organizations can gain comprehensive insights into various aspects of their business, enabling them to more effectively identify patterns, trends, and opportunities for improvement.

2. Data transformation

Data transformation plays a crucial role in converting raw data for analysis, modeling, and decision-making, ensuring that it's accurate, standardized, and suitable for use in different applications.

Benefits of data transformation:

- **Feature engineering**

  Data transformation involves creating relevant features that enhance the predictive power of machine learning models. Organizations can uncover valuable insights and patterns that contribute to more accurate and robust predictions by extracting, selecting, or combining data attributes.

- **Normalization**

  Standardizing data through normalization ensures that features are on a consistent scale. This process enhances the model's ability to interpret and generalize patterns from the data, helping to prevent biases toward certain variables and improving the stability and convergence of machine learning algorithms.

- **Dimensionality reduction**

  Dimensionality reduction eliminates redundant or irrelevant features. It also enhances model efficiency and reduces computational complexity, improving machine learning models.

3. Data streaming

Data streaming processes and analyzes data continuously as it's generated or received, in real or near-real time. It's an important part of preparing data for AI because it handles and analyzes large volumes of data quickly, speeding up time to value for insights.

Benefits of data streaming:

- **Timeliness**

  Data streaming enables AI models to process and analyze data in real time, so organizations can react quickly to changing conditions and emerging trends. Rapid response enhances decision-making agility and helps organizations capitalize on time-sensitive opportunities.

- **Event-driven decisions**

  Real-time insights from streaming data facilitate event-driven decisions, because organizations can respond immediately to specific triggers or occurrences. Applications, such as fraud detection, anomaly detection, and stock trading, can detect and respond promptly to critical events, minimizing risks and maximizing opportunities.

- **IoT and sensor data**

  Data streaming is crucial in handling and processing data generated by devices, sensors, and IoT networks across applications and industries. By continuously ingesting and analyzing sensor data streams, organizations can monitor equipment performance, detect anomalies, and optimize operations in real time, improving efficiency, reliability, and safety.

4. Data querying

Using queries or SQL to retrieve specific information from a database or data set, organizations can identify and retrieve the most relevant features, variables, or records required for building and validating AI models. This targeted approach to data retrieval helps to ensure that AI algorithms are trained on high-quality, relevant data, improving the accuracy and effectiveness of the resulting models.

Benefits of data querying:

- **Customized insights**

  By crafting targeted queries, users can extract the most pertinent information from large datasets and uncover valuable insights and trends that inform reporting and decision-making.

- **Ad hoc analysis**

  Data querying facilitates ad hoc analysis, so users can spontaneously perform exploratory data analysis and hypothesis testing. With the flexibility to query data sets as needed, users can explore data relationships, identify patterns, and gain deeper insights into underlying trends quickly, without the constraints of predefined analysis structures.

- **Interactive dashboards**

  Query results are the foundation for interactive visualizations and dashboards that provide intuitive and actionable insights. Users can explore and interact with data visualizations dynamically by populating them with query outputs, gaining deeper understanding and uncovering meaningful insights through interactive data exploration and analysis.

5. Data visualization

Data visualization converts raw data into graphical representations, enhancing understanding and revealing patterns, trends, and relationships that might not be apparent in raw data alone. Visuals, such as charts, graphs, maps, and dashboards, offer an appealing and intuitive way to explore complex datasets, making it easier for stakeholders to interpret and analyze information.

Benefits of data visualization:

- **Insight communication**

  Data visualizations are crucial in conveying complex patterns, trends, and outliers, both visually and intuitively. By representing data through charts, graphs, and diagrams, organizations can effectively communicate insights to stakeholders, facilitating a deeper understanding of data-driven narratives and enabling informed decision-making.

- **Decision support**

  Clear and informative visuals are powerful tools that highlight key information and facilitate data-driven decision-making. Visual representations of data help stakeholders quickly identify meaningful trends, correlations, and anomalies so they can make timely and informed decisions to improve business outcomes.

- **Exploratory analysis**

  Interactive data visualizations help users explore data from different perspectives and angles. Interacting with visualizations, users can manipulate and drill down into data sets dynamically, uncovering hidden patterns, relationships, and insights that might not be obvious through traditional data analysis methods. This exploratory approach to data analysis promotes discovery and fosters a deeper understanding of complex data sets.

6. Data collaboration

Collaborating across data sets, systems, and platforms, data science professionals and teams can build off each other's research and accelerate innovation without complex redundancies. These teams work easily toward common, strategic goals with the help of seamless data-sharing.

Benefits of data collaboration:

- **Cross-functional insights**

  Data collaboration helps to disseminate diverse perspectives, expertise, and domain knowledge. By bringing together individuals from different departments or disciplines, organizations can gain deeper insights into complex problems, driving more informed decision-making and collaborative problem solving.

- **Data governance**

  By establishing clear roles, responsibilities, and processes for managing and sharing data, organizations can maintain data integrity, protect sensitive information, and adhere to regulatory requirements, fostering trust and accountability in data-driven initiatives.

- **Innovation**

  Data collaboration stimulates innovation by creating opportunities for knowledge sharing, idea generation, and experimentation. By encouraging collaboration and open communication among stakeholders, organizations can create a culture of ingenuity with AI, supporting exploration of new concepts, solutions, and approaches to addressing business challenges.

Requirement #2: Modernize your data management architecture

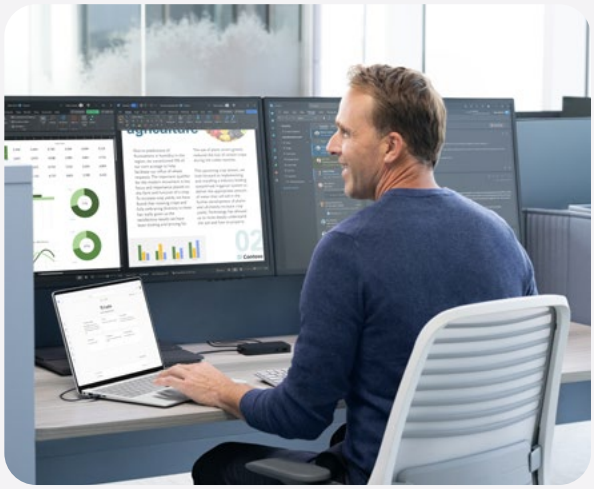
As data volumes grow exponentially, a scalable architecture helps to ensure that organizations can efficiently handle ever-increasing amounts of data, without compromising stability or performance. If you want your AI models to deliver continuous value, scalability and flexibility are critical to accommodating the ever-expanding data requirements of AI and machine learning. For example, an e-commerce retailer has an AI app that provides personalized product recommendations to online shoppers based on their browsing and purchase history. As app engagement increases, the retail company must be able to scale its recommendation engine to accommodate the growing volume of data and diverse customer preferences, which could help the company drive higher sales and customer satisfaction.

The adaptability provided by a modern data management architecture helps to ensure that organizations stay agile and responsive in today's fast-paced business environment, where they must act quickly to proactively seize opportunities and address challenges. When it comes to establishing a scalable and flexible data management architecture, many organizations look to data lakehouses.

What is a lakehouse?

A *data lakehouse* is a modern approach to data management that combines the best aspects of data lakes and data warehouses. Merging a data lake's flexibility and cost effectiveness with a data warehouse's robust data management capabilities, data lakehouses offer a comprehensive solution to the challenges of managing diverse data types.

Unlike traditional data lakes, which often face challenges with data organization and governance, and data warehouses, which can be rigid and expensive to scale, lakehouses offer a unified platform that caters to the evolving needs of modern data analytics. With a lakehouse architecture, organizations can seamlessly ingest, store, and analyze various data sources while maintaining data integrity and governance.



Benefits of lakehouses

A lakehouse presents numerous benefits for teams embarking on AI and machine learning endeavors:

- **Scalable storage and processing**

  Lakehouses are well-suited for managing increasingly large volumes of data.

- **Single source of truth**

  By consolidating data, a lakehouse can minimize data silos and redundancy, helping to ensure consistent information throughout the organization.

- **Optimized for machine learning**

  Lakehouses are designed with indexing protocols optimized for machine learning and data science, enhancing query performance and enabling efficient data exploration for model development.

- **Low query latency**

  Teams can retrieve insights quickly for running complex machine learning algorithms or generating reports.

- **Data freshness**

  Lakehouses maintain up-to-date data freshness through real-time data ingestion and integration with streaming sources, enabling teams to work with the latest information.

- **Data security and governance**

  Lakehouses help organizations to control data access and ensure compliance by managing data security and governance within the platform, safeguarding against data leakage, and maintaining privacy.

OneLake

OneLake is a single, unified, logical data lake that can support your entire organization's data requirements, even across regions. Benefits include:

- **Data storage**

  Using Microsoft Fabric, store lakehouses, warehouses, and other items in OneLake.

- **Scalability**

  Process large volumes of data from various sources to support your entire organization—at scale.

- **Data consistency**

  Reduce latency and improve consistency of data and analysis by using one copy of your data with multiple analytical engines.

- **Governance**

  Manage data governance and compliance within the boundaries of your tenant admin.

- **Collaboration**

  Create as many workspaces as you need, with distributed ownership and access policies.

Every Microsoft Fabric tenant receives one instance of OneLake, automatically provisioned, with no extra resources to set up or manage.

Requirement #3: Unify your data with Microsoft Fabric or Azure Databricks

Successful AI-powered analytics require a unified environment for your analytics tools and solutions. The field of data and AI has grown enormously, with teams adding new products and capabilities to their tech stacks over the years. By combining different analytics capabilities, such as data processing, real-time analytics, and business intelligence, under one digital roof, Microsoft Fabric and Azure Databricks help users to derive deeper insights, drive innovation, and extract maximum value from their data resources.

- **Microsoft Fabric** is a unified, end-to-end analytics platform that combines essential data and analytics tools, such as data engineering, data science, real-time analytics, and business intelligence.
- **Azure Databricks** offers an open data lakehouse in Azure that can process all data types, so users can deploy, share, and maintain enterprise-grade data and AI solutions at scale.

"We help users unlock the value of their data by asking questions specific to their business to make better data-driven decisions."

—Srihari Kumar,
Chief Product Officer,
Insights and Data, JLL

Read JLL transforms real estate with Azure AI Services

"Microsoft Fabric quickly emerged as a key component, seamlessly integrating various elements to create a comprehensive analytics solution for us and for our clients."

—Rajiv Phougat,
Principal and Chief Technology Officer for Data and Analytics, KPMG US

Read KPMG transforms its global data analytics with Microsoft Fabric

Drive better business decisions with visualization

Democratized business data

Today's low-code business intelligence tools eliminate historic obstacles to getting business value out of your data. No longer do you need a skilled developer, a data scientist, or an analyst to derive value from the volumes of data at your fingertips. Self-service analytics and generative AI solutions deliver fast, permission-based access to data analysis and AI insights that help your users create more value.

Faster insights

With in-app and in-context insights, delivered in real or near-real time, everyone in your organization can be equipped to make quick decisions in response to change, respond to issues as they occur, solve small problems before they become big ones, and identify innovative opportunities as they emerge. Applications with built-in triggers can be more responsive, too, reacting to positive or negative events with immediate and decisive action.

Data visualization

Communicating trends and insights visually is key to understanding data. And you no longer need a background in graphic design or database design. Using tools like Power BI and Microsoft Fabric, you can transform data trends and insights into visual presentations that make sense to your audience, inviting more engagement and deeper understanding that can lead to business value.

Seamless governance

With all your data managed from a single platform, data security and control are streamlined. Although there's just one data lake, ownership, data certification, access, and compliance can be distributed. Trusted data, trusted governance.

A data foundation for business value

To easily drive business value broadly across your organization, you need a solid data foundation, including:

- A unified, lake-based data estate, intuitively organized, that minimizes errors and inconsistencies and streamlines analytics.
- Generative AI solutions, ready to unlock insights from your data using performance-optimized models.
- Robust tools for self-service analytics that facilitate insights in real time, accelerating new opportunities and innovation.
- Layered security and governance to help you provide data access where and when it's needed, delivering immediate value without compromising data protection.

With Microsoft Fabric and Azure Databricks, you get end-to-end data and analytics platforms to help you unify data, transform it, and uncover insights that can improve your business outcomes.

Customer stories

→ **Chanel** combines exceptional creation with cutting-edge technologies powered by Microsoft AI & Fabric to continually elevate client experience

Read the story

→ **AT&T** migration to Azure Databricks catalyzes technical staff, advances business goals

Read the story

Start unifying your data for AI

There's no time like the present to start reaping the business benefits of real-time data insights, based on the volumes of data you already collect, across your organization.

With an end-to-end data and analytics platform, you start by unifying your siloed data—on-premises and in the cloud—into a single source of truth. Transform it to create a clean, consistent data source, ready for custom, generative AI solutions with performance-optimized models and self-service analysis. Add security and compliance, opening up data across your organization based on permissions and governance.

Before long, users across departments, from data science to human resources to marketing, will be using self-service tools and scalable models to discover and communicate insights that could transform your business.

Take the next steps

To get to know Microsoft data solutions, download these e-books:

Getting Started with Azure Databricks

Microsoft Fabric: Unified Analytics for Data-Driven Innovation

To explore the tools, practices, and policies that Microsoft has created to uphold responsible AI principles, reference:

Empowering responsible AI practices | Microsoft AI