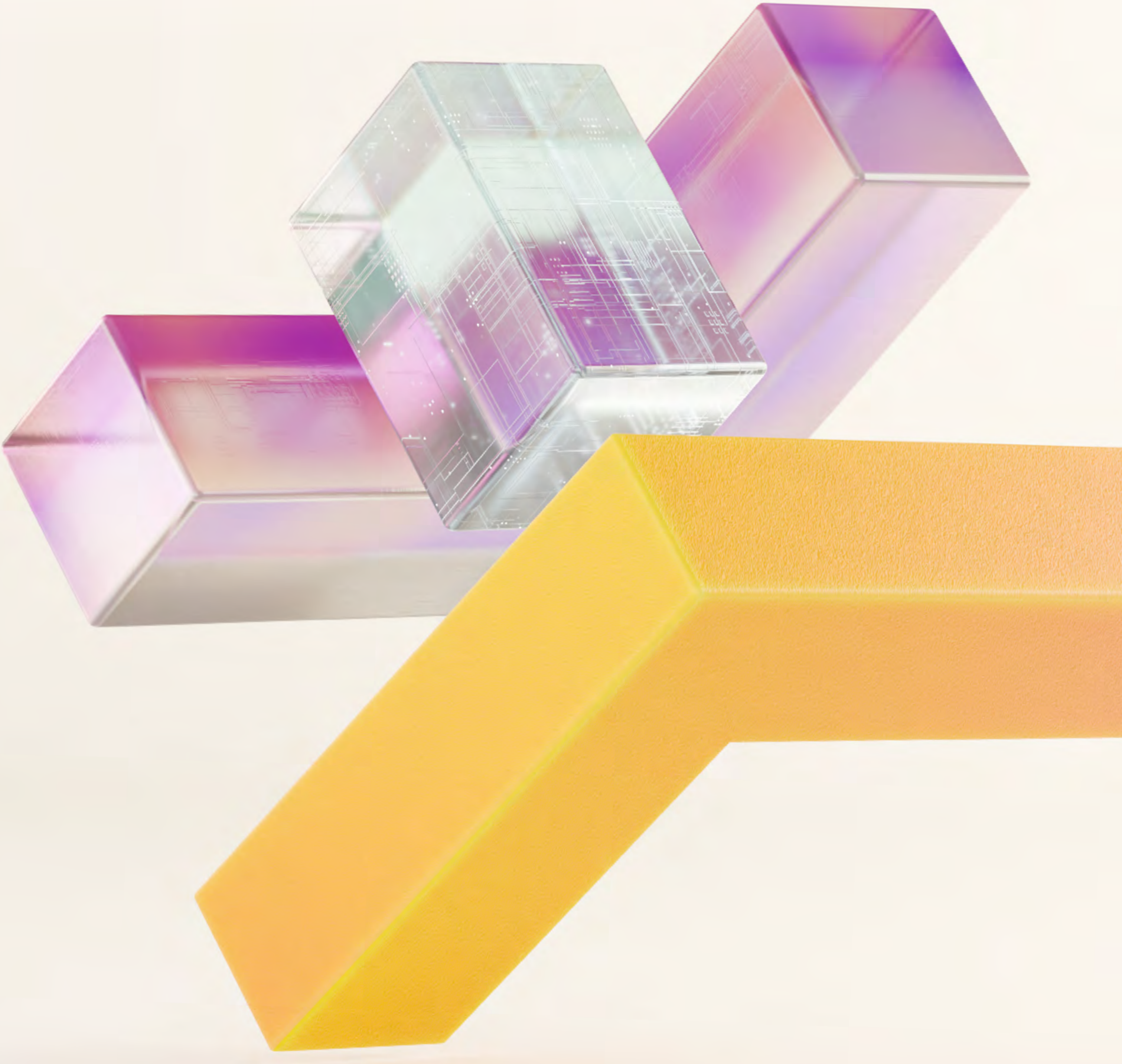
Microsoft

Maximise Data Value

Preparing your data for AI innovation



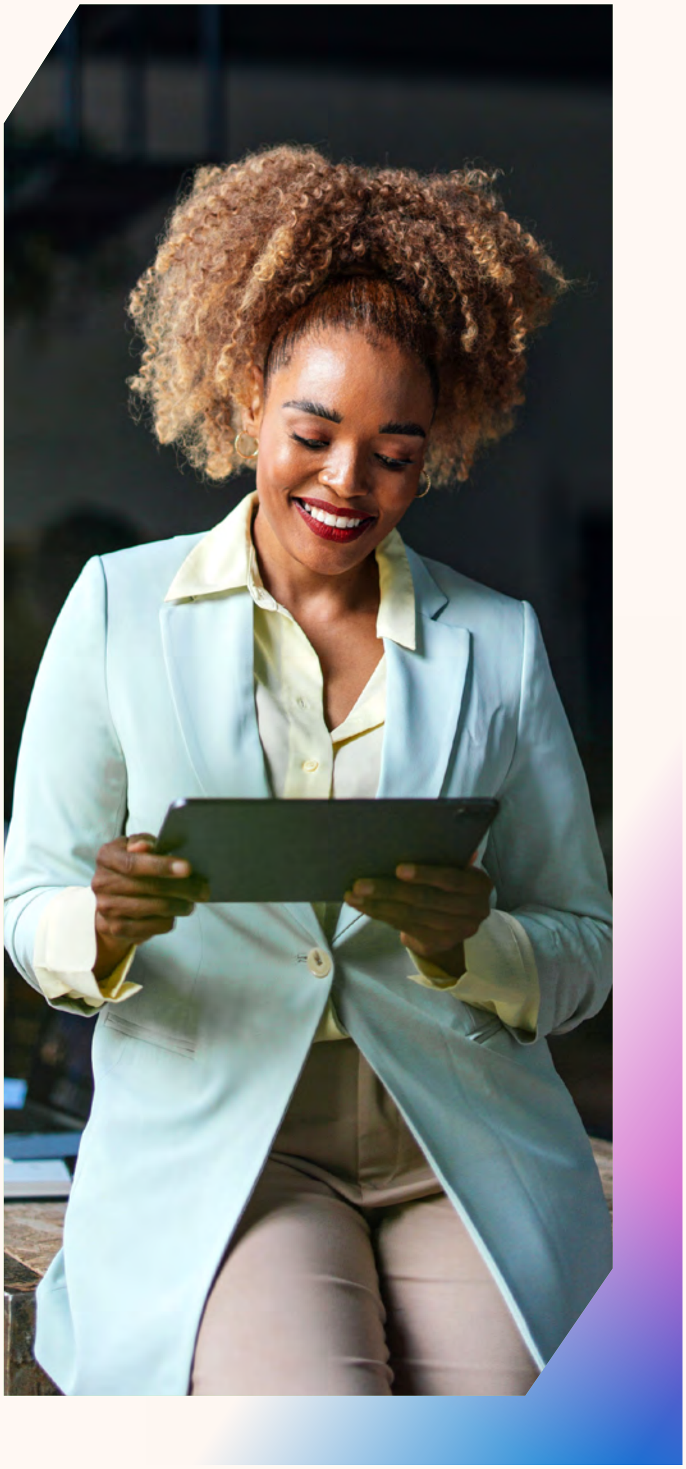
Contents

Introduction

AI has the potential to revolutionise industries, automate processes and uncover insights at a pace that was previously unimaginable. As the excitement around AI continues to build, business leaders realise that without ample and diverse data, AI cannot effectively identify patterns or make informed decisions. Consequently, data professionals face a pressing challenge – ensuring their data is primed and prepared to fuel AI applications effectively.

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 organisations to realise the full potential of AI investments. In this eBook, you'll learn how to reap the maximum benefits from your data by properly preparing it for AI and predictive analytics.



Chapter 1

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. One of the most essential elements of a solid AI foundation is the access to clean, secure, real-time data. Without that access, AI models lack the best information for their purposes – which would naturally reduce the value of their output.

However, attaining such data integration and quality can be a significant challenge. Many IT environments weren't originally designed with AI in mind. This has left data professionals grappling with numerous complexities that make building and scaling their AI models difficult.

Challenge #1: Data movement

Data movement 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.

Data movement is necessary to extract data from operational systems, transform it and load it for analysis. Data movement also supports business intelligence initiatives by extracting data from various sources, transforming it for analysis and loading it into analytical tools or databases for reporting. Adhering to data localisation laws or residency regulations also necessitates data movement, as some data must be stored in specific geographic locations by law.

Being able to move data around is fundamental to many operations. **However, the more time your team spends simply moving data where it needs to go, the less return you stand to gain from its use due to several factors:**

→ **Latency and delays**
Moving substantial data volumes introduces latency, causing real-time applications to suffer delays and negatively impacting prompt decision-making.

→ **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, while compliance may restrict cross-border data movement.

→ **Costs and resource use**
Data movement takes up computational resources and drives up costs, making allocating resources efficiently challenging.

→ **Low-performing models**
Isolated data silos hinder machine learning workflows, slowing down insights and potentially leading to biased models.

→ **Increased transformation overhead**
Pre-machine learning data transformation adds overhead, especially when moving raw data to processing pipelines.

Challenge #2: 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 or files. Human error, system failures, inadequate data governance practices and migration can also lead to the retention of duplicate records.

Data duplication is a common issue as new systems and applications are added to IT ecosystems.

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, but duplicate entries can confuse those patterns and compromise predictive accuracy.

→ **Increased strain on talent and resources**
Trying to make sense of data – figuring out what's up to date, where data originated, etc. – can put undue strain on data professionals. It also diverts valuable resources away from innovation initiatives.

→ **Heightened complexity and processing time**
Deduplicating data is labour-intensive and diverts resources from essential tasks. Although AI-powered deduplication offers automation, it demands computational resources and time that could be allocated to more value-adding tasks.

→ **Increased storage costs and retrieval delays**
Storing duplicate data increases expenses, particularly in cloud environments. Due to redundant entries, retrieving data from large datasets becomes sluggish.

→ **Higher risk of erroneous output**
Duplicate data can lead to incorrect conclusions, jeopardising 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.

Chapter 2

Overcoming data challenges to accelerate AI innovation

By minimising unnecessary movement and eliminating duplicate records, organisations can ensure their AI models have what they need to realise the full potential of their data.

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

Why is preparing your data a critical step to embracing AI?

- → Allow all artefacts to use the same data set without requiring duplication or movement.
- → Minimise data movement to help machine learning initiatives yield the most valuable results.
- → Enable the easy discovery and reuse of all data assets by all users to drive greater efficiency.
- → Improve the efficacy of AI solutions by providing them with accurate and reliable data.

Next, let's explore the essential requirements for properly preparing your data.

Chapter 3

Requirement #1 Assemble the necessary capabilities

In overcoming data movement and duplication challenges, organisations require a comprehensive set of unique data capabilities spanning data integration, transformation, streaming, querying, visualisation and collaboration.

Think of this requirement as the part in action films 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 organisations to derive actionable insights and drive innovation from their data assets.

Here are the essential data capabilities that help you build fit-for-purpose AI models that deliver the most possible value:

1. Data integration

Data integration combines data from various sources to make it readily accessible. It involves extracting data from different systems, databases or applications, transforming it into a consistent format and then loading it into a centralised location for analysis and decision-making purposes.

Many organisations use data lakes and lakehouses to integrate and unify their data. This is 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**
Data integration brings information from multiple sources together to ensure that AI models have access to a comprehensive dataset. This unified dataset helps AI models make more informed and accurate predictions, leading 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, organisations 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 organisation's operations, customer interactions and market trends. By combining data from diverse sources, organisations can gain comprehensive insights into various aspects of their business, enabling them to identify patterns, trends and opportunities for improvement more effectively.

2. Data transformation

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

Benefits of data transformation:

→ **Feature engineering**
Data transformation involves creating relevant features that enhance machine learning models' predictive power. Organisations can uncover valuable insights and patterns that contribute to more accurate and robust predictions by extracting, selecting or combining data attributes.

→ **Normalisation**
Standardising data through normalisation ensures that features are on a consistent scale, preventing biases towards certain variables and improving the stability and convergence of machine learning algorithms. This process enhances the model's ability to interpret and generalise patterns from the data.

→ **Dimensionality reduction**
By eliminating redundant or irrelevant features, dimensionality reduction enhances model efficiency and reduces computational complexity, ultimately improving the performance of machine learning models.

3. Data streaming

Data streaming continuously processes and analyses data as it's generated or received, in real time or near real time. It's an important process in preparing data for AI because it enables organisations to handle and analyse 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 analyse data in real time, allowing organisations to react quickly to changing conditions and emerging trends. This rapid response capability enhances decision-making agility and enables organisations to capitalise on time-sensitive opportunities.

→ **Event-driven decisions**
Real-time insights from streaming data facilitate event-driven decision-making, where organisations can respond immediately to specific triggers or occurrences. This enables applications such as fraud detection, anomaly detection and stock trading to detect and respond promptly to critical events, minimising risks and maximising opportunities.

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

4. Data Querying

Data querying involves retrieving specific information from a database or dataset using queries or SQL. Organisations can identify and retrieve the most relevant features, variables or records required for building and validating AI models by querying datasets. This targeted approach to data retrieval ensures that AI algorithms are trained on high-quality, relevant data, improving the accuracy and effectiveness of the resulting models.

Benefits of data querying:

→ **Customised insights**
Data querying empowers users to extract specific and relevant data subsets tailored to their analysis or reporting needs. By crafting targeted queries, users can focus on extracting the most pertinent information from large datasets, enabling them to uncover valuable insights and trends that inform decision-making processes.

→ **Ad hoc analysis**
Data querying facilitates ad hoc analysis, enabling users to perform exploratory data analysis and hypothesis testing on the fly. With the flexibility to interactively query datasets, users can quickly explore data relationships, identify patterns and gain deeper insights into underlying trends without the constraints of predefined analysis structures.

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

5. Data Visualisation

Data visualisation is the process of presenting data in visual formats such as charts, graphs, maps or dashboards to communicate information. By converting raw data into graphical representations, data visualisation enhances understanding and reveals patterns, trends and relationships that might not be apparent in raw data alone. It offers a visually appealing and intuitive way to explore complex datasets, making it easier for stakeholders to interpret and analyse information.

Benefits of data visualisation:

→ **Insight communication**
Data visualisations are crucial in conveying complex patterns, trends and outliers visually intuitively. By representing data through charts, graphs and diagrams, organisations 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 to make timely and informed decisions that improve business outcomes.

→ **Exploratory analysis**
Interactive data visualisations allow users to explore data from different perspectives and angles. By interacting with visualisations, users can dynamically manipulate and drill down into datasets, uncovering hidden patterns, relationships and insights that may not be obvious through traditional data analysis methods. This exploratory approach to data analysis promotes discovery and fosters a deeper understanding of complex datasets.

6. Data Collaboration

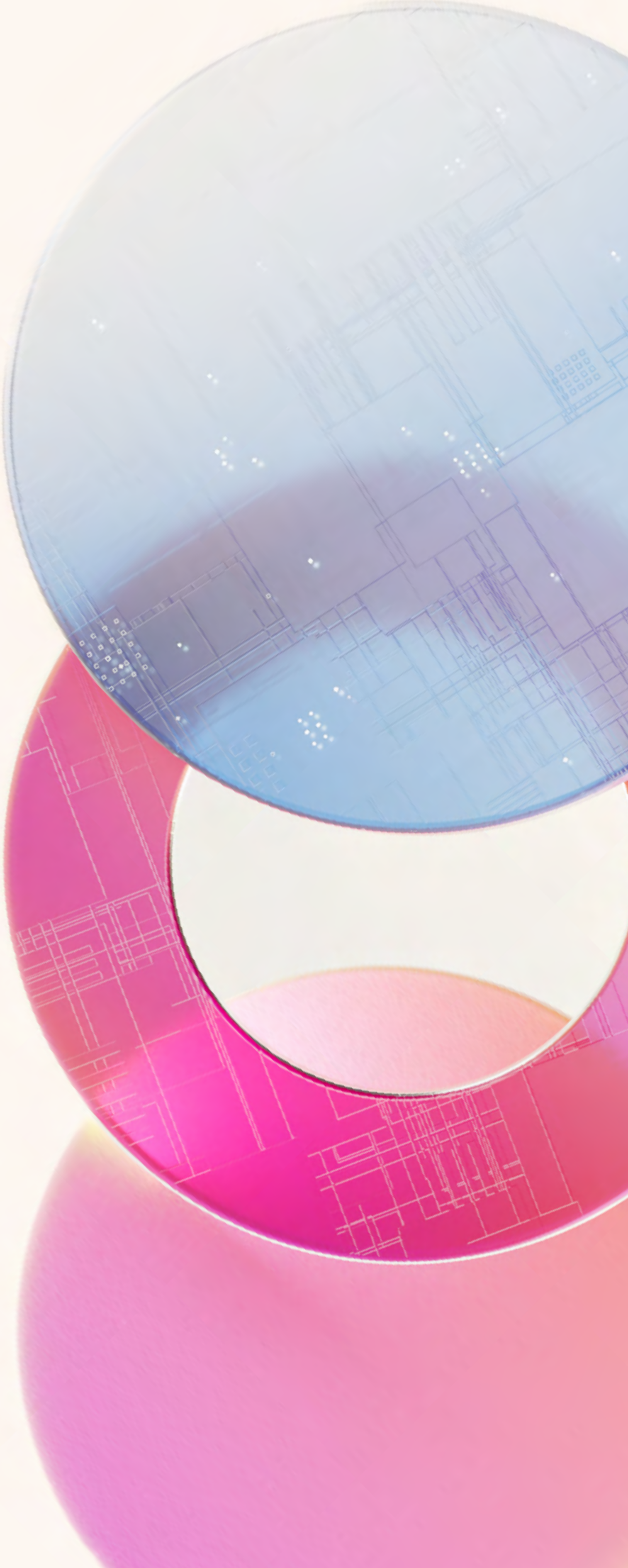
Data science involves professionals working together across systems, data sets and platforms. Collaborating over that data allows scientific teams to build off each other's research and accelerate innovation without complex redundancies. Accordingly, these teams can easily work toward common, strategic goals by enabling seamless data sharing.

Benefits of data collaboration:

→ **Cross-functional insights**
Data collaboration allows for more diverse perspectives, expertise and domain knowledge. By bringing together individuals from different departments or disciplines, organisations can gain fuller 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, organisations 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, organisations can create a culture of ingenuity with AI, allowing for the exploration of new concepts, solutions and approaches to addressing business challenges.



Chapter 4

Requirement #2 Modernise your data management architecture

As data volumes grow exponentially, a scalable architecture ensures organisations efficiently handle increasing amounts of data without compromising stability or performance. If you want to ensure your AI models can continually deliver 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 personalised 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 it drive higher sales and customer satisfaction.

This adaptability provided by a modern data management architecture ensures you stay agile and responsive in today's fast-paced business environment, where organisations must act quickly to seize opportunities and address challenges proactively. When it comes to establishing a scalable and flexible data management architecture, many organisations 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 offers a comprehensive solution to organisations' challenges in managing diverse data types.

Unlike traditional data lakes, which often face challenges with data organisation 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, organisations can seamlessly ingest, store and analyse 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 endeavours.

Let's explore these advantages:

→ **Scalable storage and processing**
Lakehouses are well-suited for managing increasingly large volumes of data.

→ **Single source of truth**
By consolidating data, a lakehouse help minimise data silos and redundancy, ensuring consistent information throughout the organisation.

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

→ **Low query latency**
Teams can quickly retrieve insights 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 helps organisations control data access and ensure compliance by managing data security and governance within the platform, safeguarding against data leakage and maintaining privacy.

Lakehouses offer a fast and reliable path to production, from proof-of-concept to deployment. By providing specialised tools and optimised workflows, lakehouse platforms streamline the development process, allowing teams to iterate on AI and machine learning models rapidly. This accelerated journey to production helps organisations realise the value of their AI initiatives sooner while upholding the highest quality and compliance standards.

Chapter 5

Requirement #3 Unify your data solutions

How Microsoft Fabric and Azure Databricks work together

Another key requirement for successful AI-powered analytics is having 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 empower 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 lakehouse in Azure, enabling the processing of all data types so users can deploy, share and maintain enterprise-grade data and AI solutions at scale.

The secret to these two platforms' compatibility is OneLake, the logical data lake that enables seamless data integration. By minimising data movement and duplication, the lake-based architecture reduces the time and costs required to build and scale high-performing AI models, letting enterprises speed up their data-driven initiatives and enlighten their decision-making processes.

Unifying its data delivered value for Dener Motorsport in the form of fast insight and time savings.

"Before we used Microsoft Fabric and real-time analytics, it was probably 30 minutes before the engineers knew something was wrong with a car, could get the data, analyse it and provide a solution. Today, that process is done in minutes. Since we have a gap of no more than an hour between practice and qualifying, those time savings make a huge difference."

Dener Pires, CEO, Dener Motorsport

Read the story ›

Using Azure Databricks contributed to a 20% reduction in operational costs and enhanced user satisfaction with its AI-enhanced mobile platform.

"With Azure Machine Learning and Azure Databricks, we observed a 20% increase in daily users reading recommended articles and a 15% increase in viewership of recommended video highlights. I don't believe we could ever have gained a 15% increase with the rule-based model, even with more rules."

Adolfo Almedia, Senior AI Engineer, MultiChoice

Read the story ›

Chapter 6

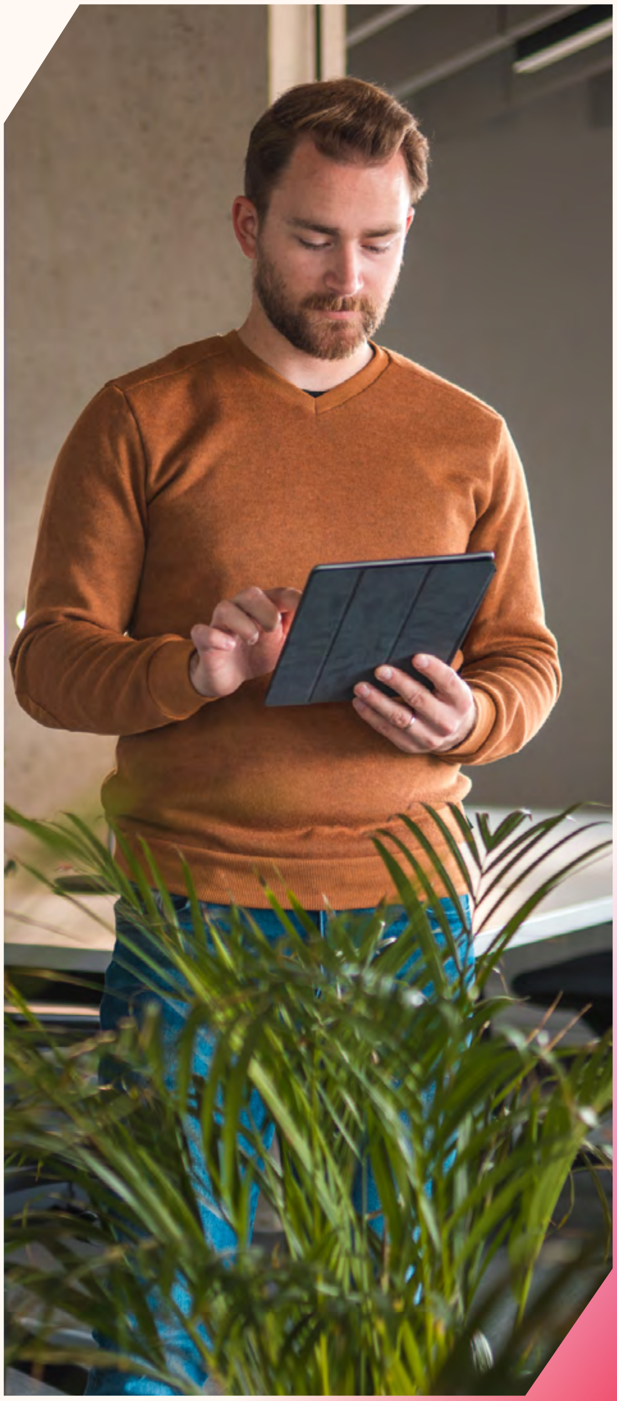
Azure Databricks + Microsoft Fabric: Better together

At the core of future analytics and AI applications lies an open and governed lakehouse foundation, encompassing a perfect blend of data lakes and data warehouses. Microsoft Fabric and Azure Databricks offer the best of both worlds – a completely unified data experience.

Below are some of the components of these solutions that make them effective as an end-to-end analytics stack:

→ OneLake: **Unify data in a centralised repository.**
Before OneLake, organisations would establish multiple lakes for different business units, even though it involved extra effort and more resources. OneLake, on the other hand, provides a single place for organisational data, enabling different teams to work from the same data set. It's also compatible with Azure Databricks, which means data engineers, data scientists and analytics can access the same data from their Azure Databricks ecosystem.

→ OneLake shortcuts: **Connect data silos without moving or copying data.**
Shortcuts in OneLake act as symbolic links to other storage locations. They unify data across domains, clouds and accounts by creating a single virtual data lake for the entire enterprise. Shortcuts help eliminate edge copies of data and reduce process latency associated with data copies. Copilot helps create reports, generate summaries and create compelling visuals that tell the story of your data simply and clearly.



→ **Comprehensive analytics experiences: Enable data collaboration between data teams.**
Fabric fosters collaboration between data teams by integrating technologies like Azure Data Factory, Azure Synapse Analytics and Power BI into a single unified platform. Each team works within the same environment, sharing data and insights seamlessly.

→ **Machine learning in Fabric (MLflow): Create machine learning models for predictive analytics.**
MLflow offers a standard packaging format for models, making them usable across downstream tools. It supports multiple model versions so data scientists can track and compare iterations and helps simplify model management without requiring training or fine-tuning.

→ **Microsoft Copilot in Fabric and Power BI: Transform and analyse data, generate insights and create visualisations and reports.**
By analysing your data and generating insights, Copilot helps turn your data into actionable information and deliver it to team members in the way that makes the most sense for their unique roles. ∂Copilot in PowerBI can help you visualize your data using natural language. And, with Copilot in Fabric, each workloads' capabilities are enhanced with AI.

→ **Azure OpenAI: Apply large language models at scale.**
Run advanced AI models on your enterprise data without extensive training or fine-tuning. Azure OpenAI offers a range of analytical use cases, including predictive analytics and forecasting. Users can build generative AI models that analyse historical data to identify patterns and improve efficiency in supply chain management, inventory forecasting and demand planning.

Customer stories

Belfius uses Azure Machine Learning, Azure Synapse Analytics and Azure Databricks to increase its agility in building reliable machine learning models.

Read the story ›

Milliman increases data access and querying using the OneLake component of Microsoft Fabric, Power BI and Azure Data Factory.

Read the story ›

Get started with unifying your data for AI

Fabric and Azure Databricks create an end-to-end analytics stack that help you unify and transform your data, letting you surface fast, accurate insights. By unifying your data within a single environment, you can better ensure the security and compliance of that data while also enabling the easy adoption of Azure capabilities for your innovation initiatives. Supported by a unified data platform, your AI projects will be more likely to succeed and deliver the desired business outcome.

By using Fabric and Azure Databricks together, you can build and scale cost-effective, quick, performance-optimised AI and machine learning models without data movement or duplication. With the time saved from not having to move and prepare data, your teams can instead focus on building secure and reliable models that deliver the highest value results.

Take the next steps

Contact sales ›

Experience the next generation of AI-driven data analytics.

Try Fabric for free ›

Find out how cloud-scale analytics helps realise more value from your data.

Learn more ›

Get expert assistance for building with AI.

Explore Azure Innovate and applications ›

This document is provided 'as-is'. Information and views expressed in this document, including URL and other internet website references, may change without notice. You bear the risk of using it. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.