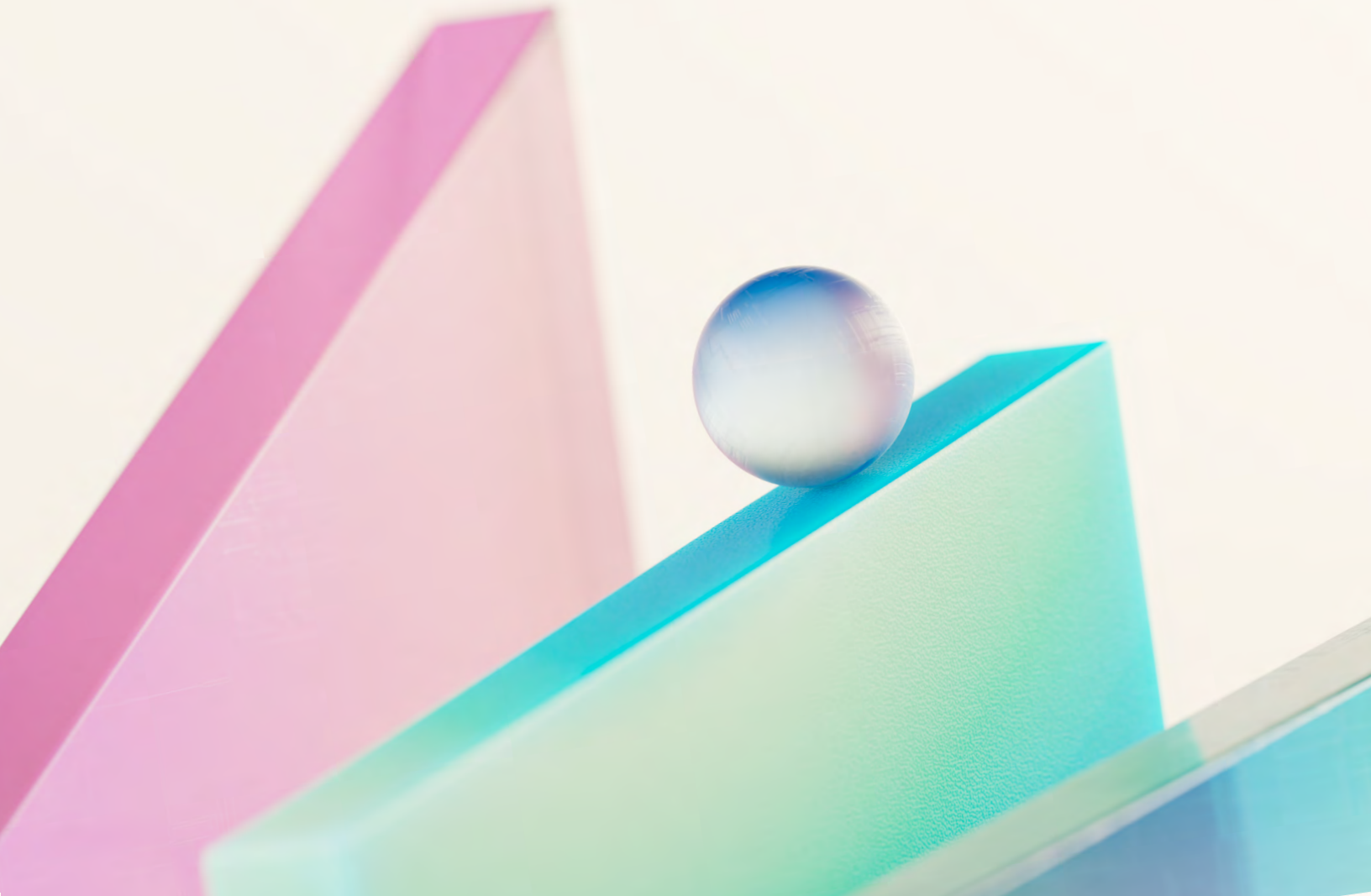
Microsoft

Modern Analytics: A Foundation to Sustained AI Success

Microsoft Fabric eBook series volume 2



Modern Analytics: A Foundation to Sustained AI Success

Data opportunities in the era of AI

The potential of AI is boundless, with profound implications for industries such as healthcare, smart infrastructure, entertainment, retail, banking, logistics, manufacturing and beyond. The significance of data cannot be emphasised enough with the ongoing transition into an era shaped by AI. Data is created and captured everywhere from devices big and small, applications and interactions.

Leading organisations harness data to undergo digital transformations and gain a competitive advantage in their respective industries. The current state of AI has amplified these opportunities exponentially. AI's ability to analyse and interpret data on an unprecedented scale allows organisations to derive deeper insights, make more informed decisions and create more personalised experiences for customers.

For instance, an Australian AgTech company, The Yield,[1] uses sensors, data and AI to provide farmers with insightful information about weather, soil and plant conditions. This knowledge helps them make informed decisions, thus promoting sustainable farming techniques. AI is truly revolutionising how organisations understand and utilise data, turning it into a valuable resource that drives innovation, efficiency and business growth.

Commercial AI capabilities today

AI is a vast field with multiple branches, each presenting a unique approach to data interpretation, learning mechanisms and decision-making processes. Understanding these diverse approaches of AI aids in comprehending its complexity and the multitude of ways it's applied. By understanding these various branches, organisations can grasp how they can harness the AI landscape.

- **Machine learning**: Machine learning is an essential subset of AI. It serves as the cornerstone of many AI systems by empowering computers to learn from data and make accurate predictions or deductions. Machine learning operates in a realm where copious amounts of data are generated daily through diverse sources such as text messages, emails, social media posts and numerous sensors embedded within an environment. Data scientists then employ this extensive dataset to train machine learning models capable of making predictions based on discerned patterns.

1 https://news.microsoft.com/en-au/features/how-the-yield-is-using-data-and-ai-to-help-feed-the-world/

Machine learning employs several techniques to analyse data and predict outcomes. These techniques include:

- **Supervised learning**: Handles datasets with labels or a predefined structure, where data serves as a teacher to train the machine, enhancing its predictive abilities.
- **Unsupervised learning**: Processes datasets without any labels or structure. It uncovers hidden patterns and relationships by grouping similar data into clusters.
- **Reinforcement learning**: Involves a computer program that interacts with its environment to learn behaviours, with the learning process based on a system of rewards and penalties.

- **Deep learning**: Deep learning is a more advanced branch of machine learning. It deploys artificial neural networks inspired by the structure of the human brain. These networks pose nested questions, with answers leading to related inquiries. Deep learning necessitates extensive training on large datasets, making it particularly suitable for applications such as image recognition.
- **Generative AI**: Generative AI utilises neural networks to scrutinise data, recognise patterns and generate novel outputs such as text, images or code. Microsoft uses generative AI in various applications, including Azure OpenAI Service, GitHub Copilot and Bing Image Creator.
- **ChatGPT**: ChatGPT, built by OpenAI, is a refined model from the GPT-3.5 series. It employs a transformer-based neural network architecture. Trained on an extensive text dataset and refined through conversational interactions, ChatGPT generates responses by considering the context of the conversation and analysing user inputs. ChatGPT is available in preview in Azure OpenAI Service where developers can integrate custom AI-powered experiences directly into their own applications, all backed by the unique supercomputing and enterprise capabilities of Azure.

Reshape how data is used

AI is predicted to be the ultimate amplifier, augmenting human work by freeing up time for more higher-level activities like solving novel problems. According to the Microsoft Education Team,[2] AI promises significant opportunities in education, enabling personalised learning experiences and democratising access to high-quality educational resources. As the adoption of AI technology increases, it is crucial to ensure its benefits are shared equitably across society, institutions and organisations.

However, to support these AI experiences tailored to a specific organisation, there must be a continuous influx of clean data from well-managed and tightly integrated analytics systems. Unfortunately, due to the fragmented data and AI technology landscape, the analytics systems of many organisations resemble a maze of specialised and disconnected services. To tackle these fragmented data landscapes and disconnected analytics systems, the fusion of data capabilities and AI emerges as a unifying solution. Through AI-enabled data integration and management, organisations have the opportunity to transform a complex landscape of specialised services into an integrated and scalable platform for value creation.

By unifying these services, organisations can better deliver a steady flow of clean, well-managed data – a critical foundation for supporting the building of AI applications within specific organisations. This strong coupling between AI and data serves as a catalyst for transformation.

Evolve AI and machine learning into data strategy

The integration of the current generation of AI and machine learning into data strategy is a transformative shift that holds great promise for organisations across diverse industries. These technologies offer a suite of capabilities that revolutionise the way organisations understand and utilise data, thereby reconsidering the opportunities and limitations of their organisational capabilities.

Strategically, organisations can incorporate AI and machine learning across three broad areas – analytics, product enhancement and research and development (R&D). Deep learning, a crucial component of machine learning, makes previously inaccessible data accessible, extracting valuable insights from unstructured data such as video and audio files. This broadens the scope of analytics, offering innovative solutions to complex data science problems and generating new product opportunities.

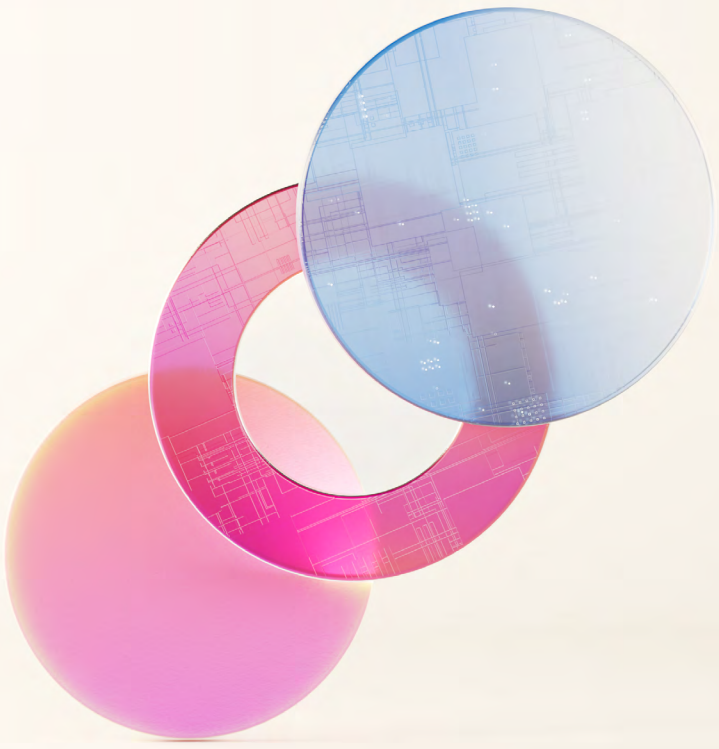
2 https://educationblog.microsoft.com/2023/03/exploring-new-opportunities-with-ai-in-education

In e-commerce, AI and machine learning have significantly improved recommendation systems, making the shopping experience more personalised and efficient. By pushing the boundaries of technical capabilities, these technologies also influence the R&D function, paving the way for groundbreaking discoveries. Therefore, a strategic integration of AI and machine learning into an organisation's data strategy can result in more insightful decision making and potentially disruptive innovation.

Harnessing the potential of AI empowers businesses and organisations to:

- **Balance between analytics, product development and R&D**: Companies must ensure a balanced approach between analytics, product development and R&D. While cost-saving measures may seem attractive initially, investing in new revenue opportunities could provide significant long-term returns.
- **Encourage experimentation in data science projects**: Unlike traditional software engineering projects, data science projects come with inherent uncertainties. It's essential for organisations to foster a culture that encourages experimentation and perceives failure as a path to success, not an end.
- **Maintain a pragmatic outlook**: Amid the hype surrounding AI, it's crucial for businesses to stay grounded. They should critically evaluate claims made about these technologies and understand their fundamental workings. AI, while powerful, is a tool that, when used effectively, can offer unparalleled insights, enhance decision making and drive business growth.

The era of AI brings forth paradigm altering opportunities and challenges in equal measure. By understanding and using these technologies, organisations can unlock unprecedented business value and secure their place in the competitive landscape.



Data science projects in Microsoft Fabric

Data science projects can be complex, particularly when dealing with large volumes of data. However, with the right structure and tools, these projects can be managed effectively and efficiently. Microsoft Fabric provides an extensive suite of resources to facilitate the lifecycle of a data science project.

The lifecycle of a data science project can be broken down into several key stages:

- **Understand the business**: A critical initial step, this involves defining the problem that the data science project is trying to solve and aligning it with the overall business objectives. The aim here is to ensure that the project's goals are clear, measurable and directly linked to the organisation's strategic objectives.

- **Acquire and understand the data**: This step involves gathering relevant data and analysing it to extract valuable insights. This stage can be challenging due to the sheer volume of data and the need to identify and extract the most relevant information. However, Fabric provides tools to aid in data ingestion and exploration, making this process easier.

- **Build the model**: The modelling stage involves using the collected and understood data to build predictive models. This can be done using various methodologies, such as machine learning or AI. Fabric offers a range of tools and features to facilitate this process, including the use of MLflow for experiment tracking and model registration.

- **Deploy the model**: Once the model is built, it needs to be deployed. This involves integrating the model into the organisation's existing systems and processes, ensuring that it functions correctly and delivers the expected results. Fabric provides tools to support this process, including the PREDICT function for generating batch predictions.

- **Manage the project**: Finally, managing the project involves monitoring its progress, making necessary adjustments and ensuring that it stays on track to meet its objectives. Fabric provides a range of resources to support project management, including a visual representation of the data science process and the ability to track experiments and manage models.

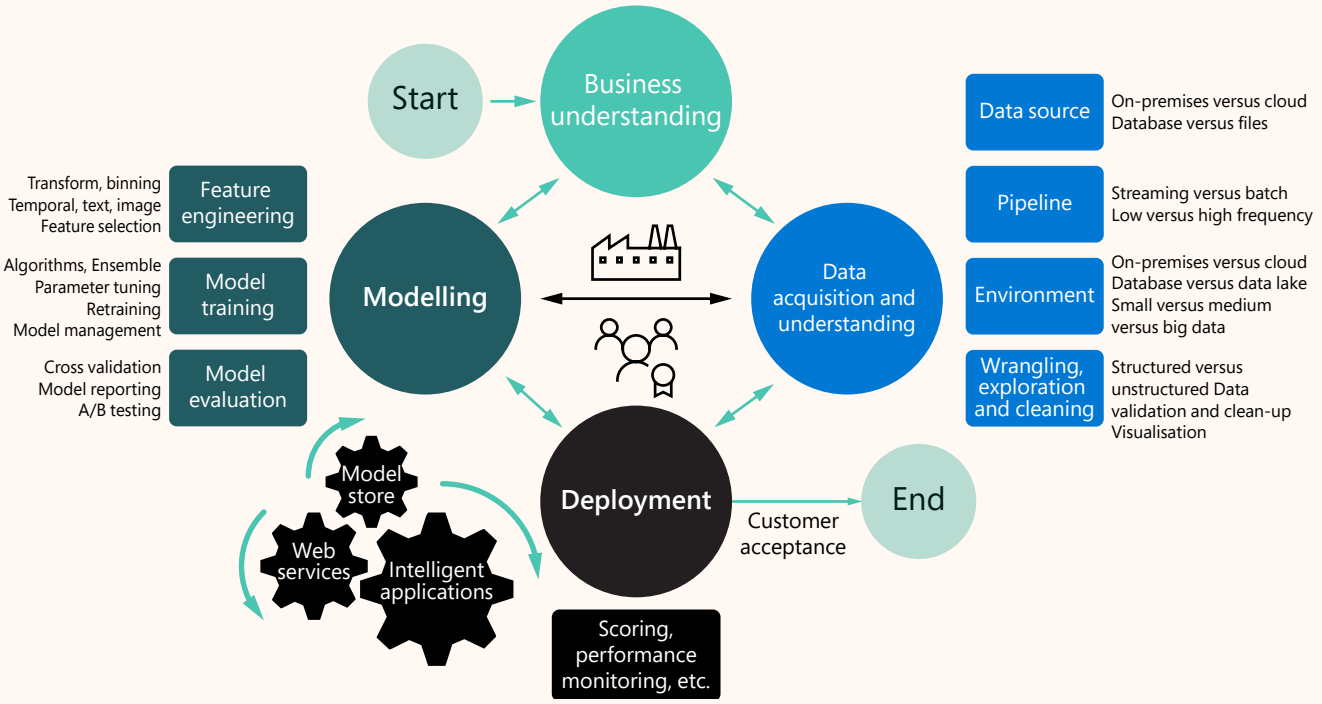


Figure 1: Data science lifecycle

Each of these stages comes with their own distinct considerations, but with the right approach and toolset, they can be effectively addressed.

Overcome risks associated with data science projects

Navigating the complex landscape of a data science project is often fraught with various risks and challenges:

- One of the most significant hurdles is ensuring data quality. When data is inaccurate or of poor quality, it can skew results, leading to ineffective and potentially misleading outcomes.
- Data security and privacy are other critical aspects to consider. With stringent regulations such as the General Data Protection Regulation (GDPR), data must be handled responsibly to avoid legal repercussions.
- The scarcity of skilled data science professionals presents a challenge. The interdisciplinary nature of data science, which encompasses statistics, machine learning and programming, makes finding individuals with the requisite skill set challenging.

- Integrating the insights and solutions derived from data science projects into existing business processes can be a complex task and may face resistance from various stakeholders.

Best practices for managing a data science project

Effective management of a data science project calls for the adoption of several best practices:

- Start with clearly defined objectives that provide a roadmap for the project.
- Ensure the data used is of high quality and relevant to the problem at hand.
- Assemble a skilled team and the use of the right tools. Microsoft Fabric, with its capabilities to simplify data ingestion, exploration, modelling, deployment and project management, proves invaluable in this regard.
- Conduct regular project reviews to help identify potential issues early on, allowing for timely course correction.
- Keep stakeholders engaged and informed throughout the project to ensure their buy-in and support.
- Prioritise data security and compliance with data protection regulations to safeguard the project from potential legal issues.

The lifecycle of a data science project can be complex and challenging. However, with the right tools and resources, such as those provided by Fabric, data scientists can ensure that their projects deliver valuable insights and contribute to their organisation's strategic objectives.

Architecture of a data science project in Microsoft Fabric

Data science projects require a well-thought-out architecture to effectively manage and execute them. By utilising Microsoft Fabric, a data science project can be structured in a way that ensures a seamless flow of processes, from data ingestion to visualisation and real-time analytics.

Fabric offers a comprehensive framework for implementing data science projects. This architecture is built to handle end-to-end data science scenarios that include:

- Data ingestion from various sources
- Data exploration and visualisation
- Data cleansing
- Model training and evaluation
- Batch scoring
- Visualising prediction results

The Fabric architecture is designed to work with different components of a data science scenario, such as data sources, models, storage and analytics. Fabric makes it easy to connect to various data services, cloud platforms and on-premises data sources for data ingestion. The data science experience on Fabric supports data cleansing, transformation, exploration and featurisation using built-in experiences on Spark and Python-based tools.

Understand the different components of a data science scenario

The architecture of a data science project with Microsoft Fabric involves several components.

- **Data sources**: Fabric allows quick and easy connection to Azure data services, other cloud platforms and on-premises data sources for data ingestion. Data scientists can ingest data into Microsoft Fabric notebooks from various sources, including a built-in lakehouse, data warehouses, Power BI datasets and Apache Spark and Python-supported custom data sources.

- **Data exploration and preparation**: Fabric supports data cleansing, transformation, exploration and featurisation using built-in experiences on Spark and Python-based tools like Data Wrangler and the SemPy library.
- **Models and experiments**: Fabric allows users to train, evaluate and score machine learning models using built-in experiments and model items. It seamlessly integrates with MLflow for experiment tracking and model registration and deployment.
- **Storage**: Fabric uses Delta Lake as the standard storage layer, enabling all Fabric engines to interact with the same dataset stored in a lakehouse. It supports both structured and unstructured data in file-based storage and tabular format.
- **Analytics and insights**: Data from a lakehouse can be consumed by Power BI for reporting and visualisation. It can also be visualised in Fabric notebooks using Spark or Python native visualisation libraries.

Real-world examples and case studies

Microsoft Fabric is used in various industries to streamline data science projects. For instance, in the sustainability and energy sectors, Fabric aids renewable energy providers in combining real-time data from wind turbines or solar panels with customer data, thereby predicting future power demands. It enables providers to enhance the efficacy of energy sources based on usage trends and to identify new business models through AI-driven analysis.

In the financial services industry, it provides a platform for combining customer portfolio and market data, strengthening risk prevention through scalable compute and analytics power. In the public sector, Fabric enables governments to combine research and data across the public health sector, using machine learning and AI to identify health risks and trends. It also helps governments to anticipate customer demand on a local and global scale, ensuring the right product gets to the right customer at the right time.

The architecture of a data science project using Fabric provides a robust framework for managing complex data science scenarios, helping organisations unlock the potential of their data and drive business insights.

Augment Copilot in Microsoft Fabric

Choosing the right technology for data science can be a complex puzzle with an overwhelming mix of tools. Microsoft Fabric simplifies this process by acting as a master conductor, orchestrating various data tools to create a harmonious symphony of data analytics. Here's how Fabric streamlines data science projects:

- **Unified platform**: By integrating Azure Data Factory, Azure Synapse Analytics and Power BI, Fabric creates a single platform that optimises data utilisation for professionals.
- **OneLake**: OneLake, a key innovation, acts as a vast library for data storage, simplifying data sharing and collaboration across the organisation.
- **Direct Lake mode**: A groundbreaking feature that allows for high-speed data access directly from OneLake, enhancing performance and efficiency.
- **Open data formats**: Fabric adopts open data formats like Delta Lake and Parquet as its native storage formats, alleviating supplier lock-in concerns and reducing data replication needs.
- **Collaboration**: Fabric facilitates teamwork with Git integration for Power BI datasets and reports, enabling easy tracking of changes, merging updates and reverting to previous versions.
- **Unified capacity model**: This model simplifies the purchasing and management of resources by providing a single pool of compute power for all Fabric workloads, thereby reducing costs and complexities.

Copilot in Microsoft Fabric offers advanced generative AI features allowing data professionals to interact with the platform in highly productive ways. For example, users can simply describe the visuals and insights they're looking for, and Copilot will analyse and pull the right data into a report. Additionally, it enables fast tailoring of reports, generation and editing of Power BI programmatic code DAX, creation of narrative summaries, and ability to ask questions about data, all in conversational language. Copilot in Fabric supercharges productivity and value generation of end users, even for those without a technical background.

Find more value in your data

Data fuels AI's ability to learn, adapt and deliver precise results. Microsoft Fabric is an innovative solution that recognises the underlying foundation on which AI constructs its understanding and decision-making capabilities. Fabric transforms the raw power of data into a force driving AI's evolution, seamlessly integrating modern data structures and data science into a shared SaaS foundation. By providing a unified data lake and centralised administration, Fabric directly addresses the prevailing challenges in data management and analytics.

Moreover, Fabric plays a crucial role in data science projects. Its robust architecture supports every stage of analytics, from data ingestion and processing to visualisation and real-time analytics. Data scientists can manage complex projects effectively and efficiently with Fabric, deriving valuable insights that drive strategic decision making.

The integration of Copilot in Microsoft Fabric makes the platform even more user-friendly. It allows users to generate code, build machine learning models and create data pipelines using simple, everyday language. This significantly reduces the learning curve for adopting new technology, making advanced data science accessible to a wide range of users.

Microsoft Fabric stands as a pillar for the AI-driven future of data management and analytics. It is revolutionising businesses across various sectors, making advanced data science accessible to all and significantly contributing to the sustained success of AI.

Next steps

- Start transforming your data with the Microsoft Fabric free trial
- Read this blog to learn more about Synapse Data Science in Microsoft Fabric
- Learn how to create end-to-end solutions in Microsoft Fabric
- Discover more about the Microsoft Fabric integrated analytics platform. Read Vol 1, Microsoft Fabric: The Essential Guide for Decision Makers
- Read Vol 3, Manage and Govern Your Data with Microsoft Fabric and Microsoft Purview, to learn more about the governance aspects of Microsoft Fabric