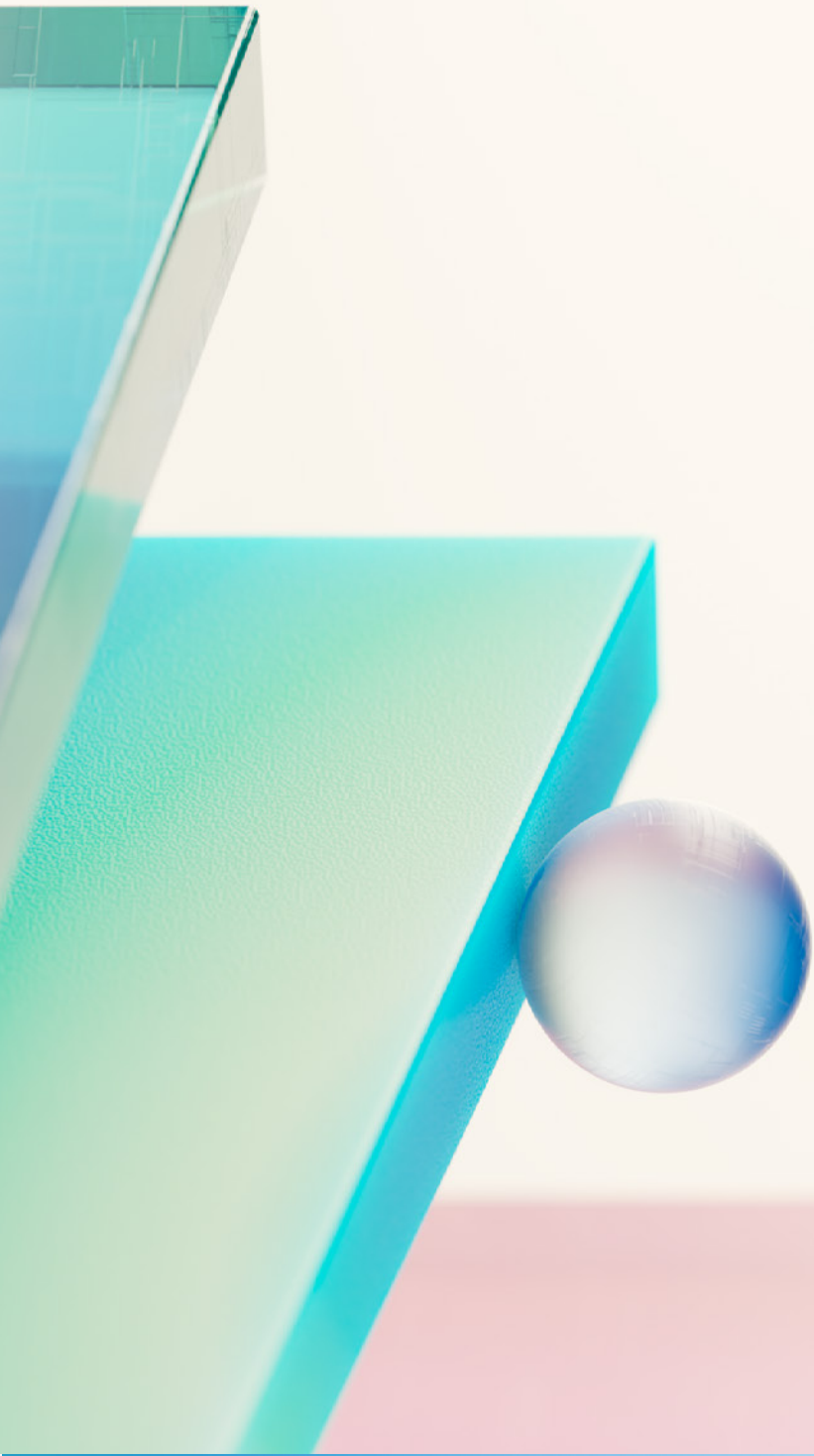
Microsoft

White paper

Modern Analytics, AI, and Governance at Scale

Learn how a strategic framework for data is the foundation for AI innovation

Executive summary



Data has long been the linchpin for making digital transformations possible. Now, after many years into this digital future, we know simply having data isn't enough. Organizations need a system to unlock the value of their data, to power analytics and artificial intelligence (AI) that help sharpen their competitive edge. They also need to adapt their culture, people, and processes so the organization as a whole is maximizing data.

Based on hundreds of engagements and conversations with customers, Microsoft has developed a framework for organizations to adopt a unified analytics and AI ecosystem. Known as Modern Analytics, AI, and Governance at Scale (MA²G), this framework addresses hurdles like siloed data, poor governance, and manual data management. Built on Microsoft Fabric, MA²G helps organizations make sure their data is AI-ready and allows business units to find relevant data assets, without compromising enterprise requirements.

Keep reading to learn more about MA²G and what it can do for your organization.

When your data is siloed, your organization is siloed

In most cases, organizations struggle to implement an end-to-end analytics and AI ecosystem, not because of the technology, but because they do not plan for, nor address, the role culture, people, and processes play in bringing about digital transformation. In fact, the MIT Technology Review says nearly every time—92%—an organization struggles with data it links back to an issue with the company's data strategy, data governance, and/or data management.[1]

This is because many organizations approach a data challenge with a technology solution—but that only solves a fraction of the problem. Organizations need a wider and deeper focus on solutions that drive cultural changes and align people and processes with technology.

Throughout hundreds of engagements with organizations worldwide to help them become data driven, Microsoft has seen the following top challenges over and over. Here are three common problems associated with culture, people, and processes that impede a unified analytics and AI ecosystem.

Lack of data strategy leads to a siloed ecosystem

After years of implementing different analytics projects in the cloud, organizations continue to build out their ecosystem in a reactive, piecemeal way. Without a well-defined data strategy, solutions become siloed and technical debt increases, which stands in the way of bigger data and analytics innovation. The lack of an analytics foundation for all data also inhibits a thriving ecosystem.

For example, imagine a data warehouse migration that lands data in proprietary data formats. At the same time, Internet of Things (IoT) data is streamed into a data lake store. Data from these two separate projects lands in separate data stores, causing siloed data. To become data driven, the entire organization must be able to build meaningful insights from both sources, regardless of the boundaries between business unit, so they can access data of all attributes and types.

[1] Building a high-performance data and AI organization, MIT Technology Review, April 2021

Poor governance prevents democratized data

Data governance—which includes a clear set of policies, processes, and controls—is critical for organizations to find, manage, and consume data across the business. A lack of data governance and an incomplete understanding of the data can stymie analytics projects. Failure points can run the gamut. Some projects can't access data fast enough, others can't integrate data due to the lack of data relationships, and others require additional data engineering work to be completed before being able to train machine learning models.

All these problems can be avoided if the organizations have enterprise data governance that provides the inventory and context of all data, automated processes to streamline workflows, and policies that automatically manage data access. The goal is to implement robust data governance and data management that enables different analytics projects for different business units.

Manual data management slows time to insights

For data to be consumable, there's a ton of ingestion and data engineering work that must happen. It must be cleaned with approved data quality, integrated to reveal new business insights, and aggregated to become a proper data product. Most organizations tend to build manual data engineering workflows on a case-by-case basis aligned to specific projects.

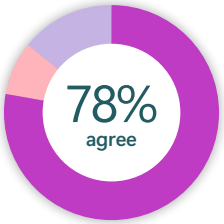
However, this creates fragmented data engineering solutions that become harder to maintain as they grow to include thousands on top of thousands of pipelines. The worst part is, most data engineering tasks are manual. Many organizations use their best people to perform these manual tasks when it should be a process change instead. By implementing proper data management processes with automation, organizations can reassign data engineers to more meaningful work of business data modelling, data aggregations, and calculations.

What will it take to make your business AI-ready?

From chatbots supporting service interactions to generative AI creating sales content, AI is everywhere. And demand for these capabilities is soaring. Just one year after ChatGPT launched, 54% of companies had already implemented generative AI in some area of their business.[2]

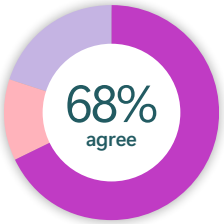
However, a paradox still exists when it comes to AI. While 78% of executives agree that AI is a top business priority,[3] they also recognize that data problems will likely stand in their way of achieving their AI goals. This is because many organizations see AI and analytics as initiatives that can be adopted by investing in technology solutions—but that's just a small piece of the puzzle. Organizations that successfully deploy AI and analytics also change processes, adapt their culture, and support their people to use these new capabilities in effective ways.

AI priorities for executives



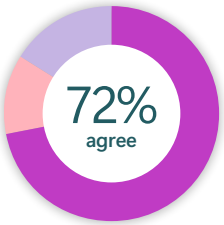
Scaling AI/ML use cases to create business value is a top priority

Disagree 8%
Neutral 14%



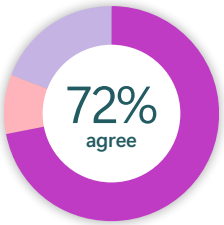
Unifying our data platform for analytics and AI is crucial to our enterprise data strategy

Disagree 12%
Neutral 20%



We favor a multi-cloud approach as a flexible foundation for AI/ML

Disagree 12%
Neutral 16%



Data problems are the most likely factor to jeopardize our AI/ML goals

Disagree 9%
Neutral 19%

[2] 2024 AI Business Predictions, PWC
[3] CIO perspectives on generative AI, MIT Technology Review, July 2023

Modern Analytics, AI, and Governance at Scale

The question every leader is asking themselves right now is: **How can my organization seize the full potential of AI, while safeguarding my business, data, and employees?**

After hundreds of engagements with organizations worldwide, Microsoft has developed a framework to not only adopt AI and analytics, but also use these technologies to their fullest extent. It's known as Modern Analytics, AI, and Governance at Scale (MA^2G) and it outlines the foundational elements organizations need to ensure their data is AI-ready.

Get started with MA^2G

MA^2G was created to give organizations a framework for understanding what it takes to deploy a successful end-to-end analytics and AI platform. It's built on the following three pillars, which describe the people, process, and culture considerations organizations need to keep in mind for AI and analytics to take root.

Enterprise Data Governance includes the set of policies and practices used to discover, describe, and manage data to accelerate responsible data democratization. Data governance ties together the data and analytics stack and automates data operations, such as cataloging, classification, creating lineage, and applying security through policies. Without it, organizations limit their ability to innovate and unlock new insights.

- **Data Management Services**
  - Governance
  - Quality
  - Policy
  - Lineage
  - Classification
  - Catalogue
- **Data Order Service**
  - Rapid Access to Data
  - On Premises or Azure

Data Management Foundation involves the practices and processes that help you create efficiencies with ingesting, storing, protecting, and ultimately serving data to different domains in the organization.

- **Self Service - Automation**
  - Domain Provisioning
  - Workspace Provisioning
  - Data Onboarding
- **Lakehouse and open data formats**
- **Automated Data Operations**
  - Data Virtualization: Shortcuts and Mirroring
  - Ingestion at Scale Solution
  - Data Engineering Acceleration
  - Automation

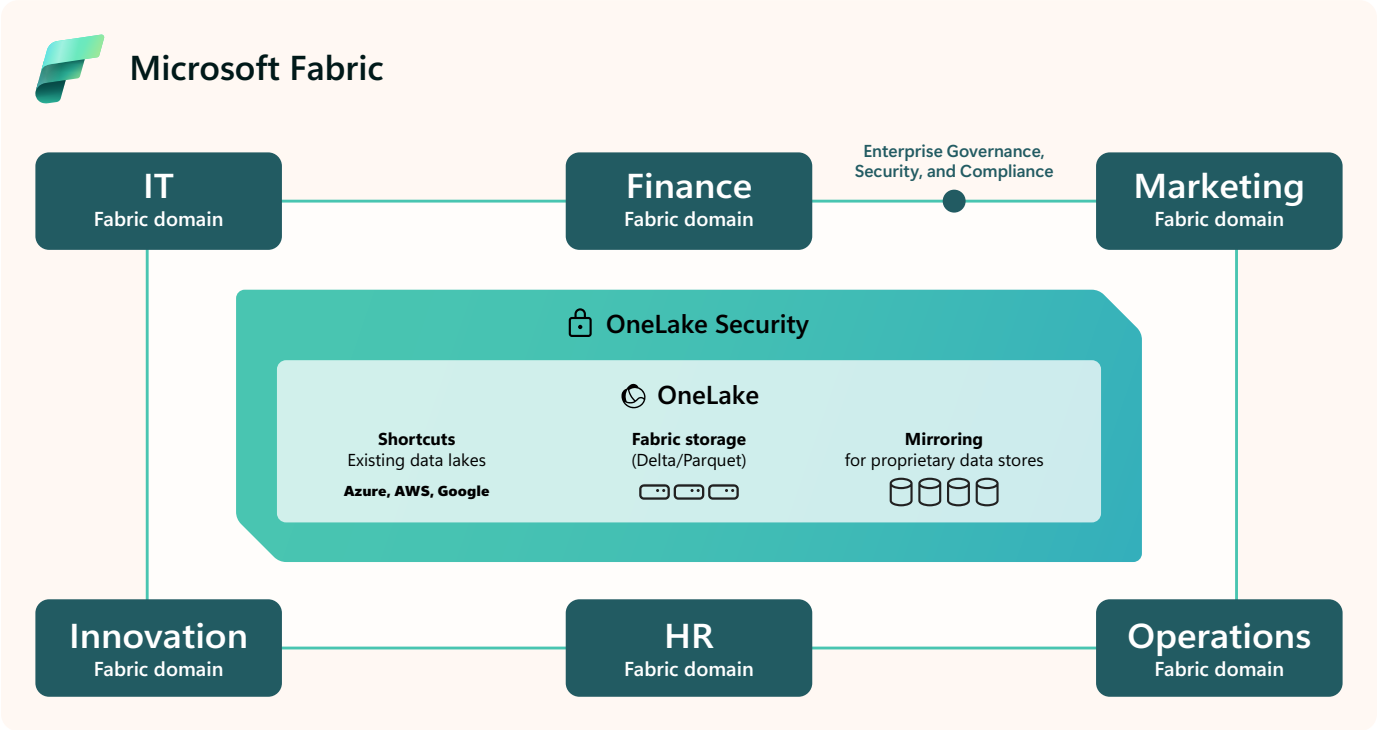
Domains and Data Products describes the environments and services that enable your business units to fully use their data. Allowing departments to self-serve data and analytics enables non-technical users to access, analyze, and build data insights or data products on their own.

- **Federation**
  - Autonomous Business Units
  - Domains, Workspaces, Capacities
  - Data Sharing / Collaboration
- **AI Copilots**
  - Empowering Data Practitioners
  - Accelerating Report Creation
  - Enable Data Exploration
  - AI Guided Insights
  - Data Integration

Together, these three solution pillars combine to help organizations achieve MA²G. By first setting business priorities as the guiding North Star, then implementing aspects from each solution pillar, your organization can shift into a whole new paradigm of doing business that empowers everyone to work toward a common goal fueled by data.

Microsoft Fabric powers MA²G

Based on the Microsoft Intelligent Data Platform, the MA²G architecture brings together the best of Azure into a unified SaaS solution powered by Microsoft Fabric and Microsoft Purview, and built on an open and governed data lakehouse. It includes support for Delta Parquet, and emerging open standards like Hudi and Iceberg, to accommodate diverse investments.



A hub for all your data

Most organizations have analytics systems that are a labyrinth of specialized and disconnected services. As businesses aim to adopt modern data capabilities, it's increasingly important to integrate all disparate data into a unified source. OneLake, incorporated into Fabric, unifies all your data in a single, accessible location for comprehensive data management. Plus, it's backed by built-in security and governance, ensuring the protection of all your data.

Simplified data preparation

For AI and machine learning models to be as accurate as possible, they must be built with clean data and in a semi-structured way. Fabric simplifies data preparation and transformation so you can quickly get your data ready for custom AI deployments. Fabric includes analytics experiences such as Data Engineering, Data Factory, and Data Warehouse built into its SaaS-based data platform so that different teams can all find the tools they need and work together.

AI-ready data that's accessible to domains

Once you have a cleaned, robust dataset, you can start building AI and generative AI experiences on top of your data. Fabric allows companies to organize data into domains where data consumers can filter and find content they need for AI or analytics. It also enables federated governance so that each business unit or department can define its own rules and restrictions.

Enterprise Data Governance

Governance is essential to every organization—regardless of the framework or solution implemented—because it lays the bedrock for responsibly democratizing data. Data governance translates your data strategy into data ownership, rules, and policies that improve data discoverability, confidence, security, compliance, and operational efficiencies.

Over the years, data access and use has spread to all corners of companies with individuals making critical business decisions that affect organizations, customers, and shareholders. You might see the finance department visualizing billions of rows of risk data, while analysts in marketing identify customers for a new product. The range of disparate uses, abuses, and copies of data is leading to widespread confusion and risk. Data governance provides the glue that ties together all the data in the analytics stack and ensures the right data is easily accessible by the right people.

Four ways to improve enterprise data governance

Create metadata: Well-governed data includes metadata about its lineage, profile, quality, business context, and classification so it's trusted and useful for all end parties.

Map data assets: With a data map, data consumers can easily and visually inspect all data assets across all domains whether it's physically stored on premises or in the cloud.

Catalog data assets: Data consumers use a data catalog to find all datasets, see if they're complete, and learn the relevant business context associated with the data asset.

Automate governance and security: Instead of manually implementing data security for each data asset—which is prone to errors and highly inefficient—automated governance keeps data secure while still ensuring access of data to the right users.

Within Fabric, you can access a Purview Hub Insights dashboard that shows you:

- **Overview report:** See an overview of distribution, use of endorsement, and sensitivity labeling.
- **Endorsement report:** Drill down and analyze distribution and use of endorsement.
- **Sensitivity report:** Drill down and analyze distribution and use of sensitivity labeling.
- **Inventory report:** Get details about labeled and endorsed items, then narrow down the results through date ranges and filters for workspace, item type, etc.
- **Items page:** Find insights about the distribution of items throughout your organization and endorsement coverage.
- **Sensitivity page:** Discover insights about sensitivity labeling throughout your entire organization.

Democratize data through self-service features

To truly democratize data, organizations need to implement a solution where data is discoverable through a data catalog and domain users can request access without opening a ticket. Microsoft Fabric and Microsoft Purview make this possible. Domain teams (data analytics, data products developers, data owners, etc.) can browse a data catalog in Fabric or Purview to discover new data relevant to their use case.

Fabric displays information about the data, such as metadata with classification, lineage, business terms, related assets, and data owners to help domains make decisions about which sources to explore. Once they select data sources, Fabric makes it easy to grant access to domains if they do not have pre-established permission. By combining Fabric and Purview, you can govern your entire estate and lineage of data. From data source down to Power BI reports, Purview and Fabric work together seamlessly so you can store, analyze, and govern your data without piecing together services from multiple vendors.

Govern and protect your data with integrated services

Governance is not only crucial to empowering your data consumers, but also gaining their trust while meeting security and compliance requirements. Fabric includes a set of capabilities that help you know, protect, manage, and monitor your organization's sensitive information. It works in tandem with Purview to help govern and manage your entire data estate. Fabric governance and compliance is tightly integrated with Purview, allowing you to create a holistic, up-to-date map of your data landscape with automated data discovery, sensitive data classification, and end-to-end data lineage.

Microsoft Purview provides a unified data governance solution to help manage and govern your data estate.

Learn more >>

Microsoft Purview provides a unified data governance solution

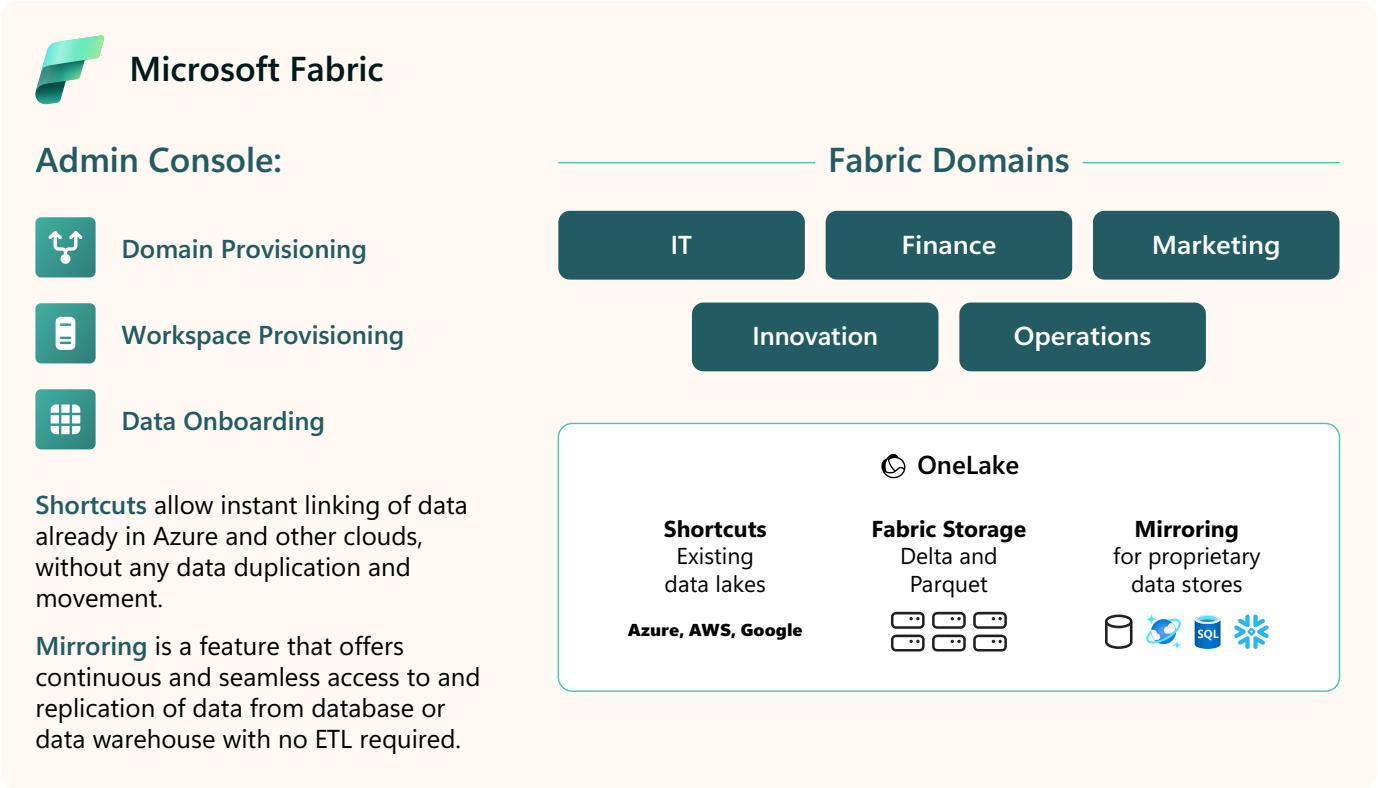
Microsoft Purview allows you to create a holistic, up-to-date map of your data landscape with automated data discovery, sensitive data classification, and end-to-end data lineage. Enable business units to access valuable, trustworthy data management, and take advantage of the following capabilities:

- **Support for multi-cloud data estates:** Automatically scan and catalog all data assets—including machine learning models and Power BI reports—across the organization, whether they're on premises, in Azure, or running on other public clouds.
- **Governance experience:** Develop clear role definitions for administrators, domain creators, data health owners, and data health readers.
- **Business-friendly terminology:** Assign language that follows the data governance experience through data products, domains, quality assessments, and reports.
- **Data scan and search:** Find the data you need across your entire estate and profile data at the source to indicate attributes like min, max, average, and thresholds.
- **Data quality scores:** Generate data quality scores once rules and policies are applied, giving you insights into your data quality relative to your business rules.
- **Metadata analysis:** Capture metadata and data lineage to help personas to decide if data is usable, then use profiling or data quality scans for recommendations.
- **Data health controls:** Ensure your rules and indicators reflect the unique standards of your organization with a set of cloud data management controls.
- **Summarized insights:** Showcase the overall health of your governed data estate with built-in data governance reports.
- **Pre-built integrations:** Extend the value of Purview with integrations for solutions related to master data management and data lineage.

Data Management Foundation

The purpose of data management is to ensure that data is properly collected, stored, processed, analyzed, and used in a secure and efficient manner to support an organization's goals and objectives. Yet, less than one-quarter of organizations have a consistent, global data management strategy in place.

Implementing automation, frameworks, and services can help organizations bolster their data management practices. With a data management foundation that uses Microsoft Fabric and OneLake, you gain both an open and governed data lakehouse for storing data, as well as automated data virtualization that efficiently sends data to domains without overburdening IT teams.



Every workload works with OneCopy and open formats

All compute engines—including Spark, T-SQL, and KQL—automatically store their data in OneLake in a single common format. Once data is stored in the lake, it's accessible to all engines and does not have to be imported or exported. Each compute type has been fully optimized to work with Delta and Parquet as their native format and a shared.

An easier way to onboard data

Fabric helps eliminate data pipelines through capabilities such as shortcuts and mirroring, which bring your data into one platform, without the legwork. In addition, you can also use partner solutions that work with Fabric connectors to move data between stores.

Shortcuts: OneLake shortcuts let you easily onboard data by instantly linking data that already exists in Azure or other clouds through a unified namespace. This eliminates data duplication or movement, reducing latency associated with data copies and staging.

- A shortcut is a symbolic link which points from one data location to another
- Shortcuts make data from a warehouse part of your lakehouse
- You can consolidate data across items or workspaces without changing the data ownership
- Data can be reused multiple times without data duplication
- Existing ADLS Gen2 storage accounts and Amazon S3 buckets can be managed externally to Fabric and Microsoft while still being virtualized into OneLake with shortcuts
- All data is mapped to a unified namespace and can be accessed using the same APIs, including the ADLS Gen2 DFS APIs

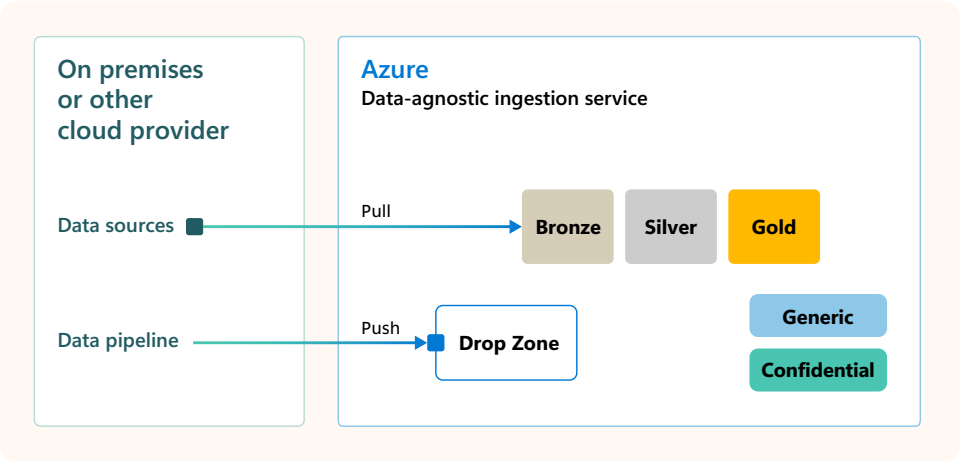
Industry solutions: Fabric includes pre-built, industry-specific solutions that help organizations integrate data from different sources and use rich analytics. Data solutions combine data integration services and, in some cases, machine learning support, so organizations can face industry-specific data challenges. These solutions include retail, healthcare, sustainability, and more.

Mirroring: Fabric offers a mirroring feature that provides continuous and seamless access to—and replication of—data from databases or data warehouses, without ETL. Any database can be accessed and managed via Fabric without having to switch database clients. By just providing connection details, your database is instantly available in Fabric as a Mirrored database.

- A full editing experience of the source database is available for the Mirrored database
- Data is replicated into OneLake in Delta format and kept up to date in near real time
- All the Fabric experiences instantly work with the OneLake replica
- Analysts and data scientists can work with real-time data
- The replica protects operational databases from analytical queries

Automated data management

For any sources that don't have mirroring or shortcuts, you can still automatically ingest data into the ecosystem. Automated data services and templates can help improve efficiency around data ingestion, standardization, quality, metadata registration, and access provisioning. These enterprise-level capabilities allow data foundation teams to minimize repetitive, manual work, and they create a foundation in OneLake for domain teams to self-serve data.



Data-agnostic ingestion service

- Pull/push
- Format agnostic
- Data agnostic
- Metadata driven

Data standardization service

- Analytics format conversion
- Versioning
- Merging
- PII handling
- Data quality
- Common Data Model
- Master data unification
- Synchronous processing
- GDPR

Data-agnostic ingestion

Automatically ingest data regardless of its attribute, format, and the domain it belongs to. Organizations can push or pull data from different sources then process it. Metadata-ingestion frameworks or Kafka-based solutions are sample solutions that can be implemented to automate this process.

Data standardization

As your data gets ingested, you can standardize it through processes such as format conversions, versioning, merging, PII handing, and master data management. Use Apache Spark notebooks within Fabric to quickly implement data standardization practices. Additional services related to data quality management address issues such as deduplication, threshold identification, and alignment with master data. Without proper checks on data quality, you run the risk of slowing down time to insights.

Metadata registration and access provisioning

With your data in OneLake, the next step is registering the new data assets in your data catalog so they're instantly discoverable. As a safety net, data governance scheduled scanning of the data hub should register these new assets. Once data is added, another automated service can provision access according to the data classification of data being ingested.

Domains and Data Products

Organizations are shifting their approach from running a center of excellence—where everything is centrally controlled—to using federated domains to provide departments more control and autonomy.

With Microsoft Fabric, every data workload is available in one SaaS experience for all personas in the organization.

Domains and workspaces makes collaboration happen

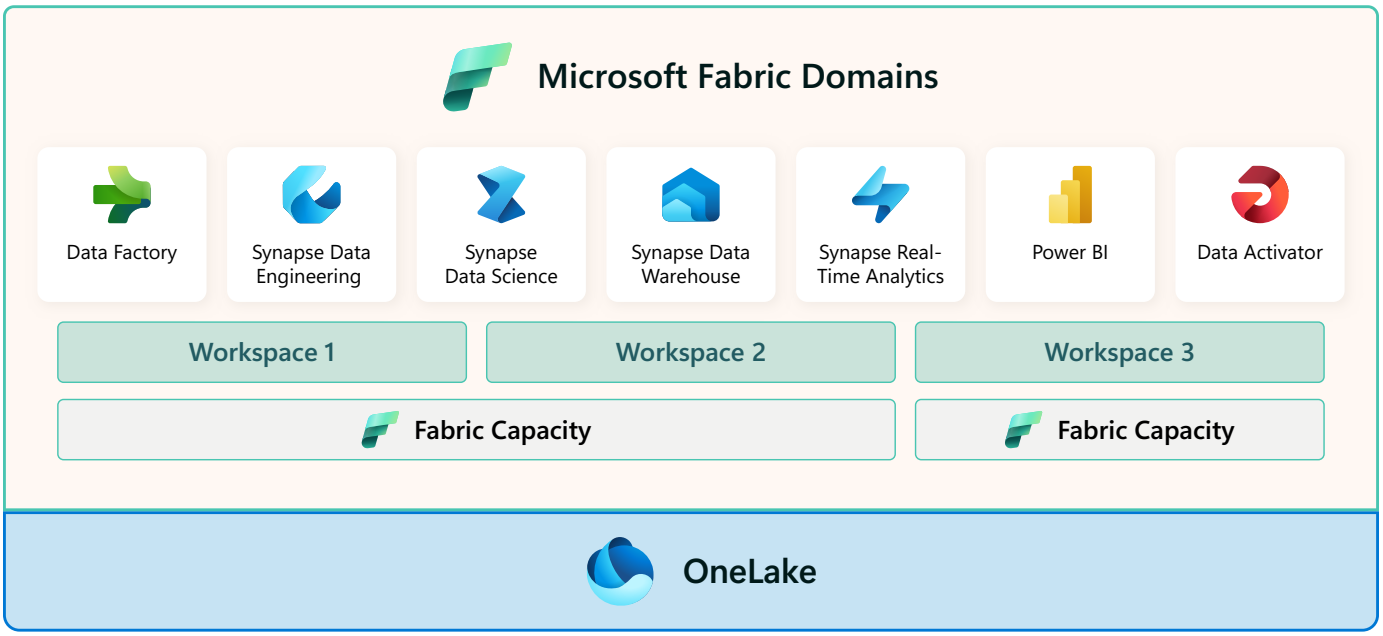
Domain provisioning: Fabric allows you to group data into a domain so that users can find the resources they need that are relevant to their field. For instance, you can create domains by business department such as HR or finance, allowing those teams to manage their data according to their specific regulations, restrictions, and needs.

Workspace provisioning: You can create a workspace to collaborate with teammates in your domain and create collections of items such as lakehouses, warehouses, and reports.

Microsoft Fabric experiences promote org-wide collaboration

Because it's built on a single, unified SaaS platform, Fabric brings together Power BI, Azure Synapse, and Azure Data Factory so data teams can collaborate in a single workspace, on the same copy of data. Fabric provides each domain with core experiences designed to work together seamlessly.

Each experience is tailored to a specific persona and a specific task, allowing different domains to find the tools they need to create their own data products.



Data Factory offers a modern data integration experience to ingest, prepare, and transform data from a rich set of data sources. Data Factory brings Fast Copy capabilities to both dataflows and data pipelines so you can move data between your lakehouse and data warehouse in Fabric at blazing speed.

Synapse Data Engineering provides a world class Spark platform with great authoring experiences, enabling data engineers to perform large scale data transformation and democratize data through the lakehouse. The Spark integration with Data Factory also enables notebooks and Spark jobs to be scheduled and orchestrated.

Synapse Data Science allows you to build, deploy, and operationalize machine learning models within your Fabric experience. It integrates with Azure Machine Learning to provide built-in experiment tracking and model registry.

Synapse Data Warehouse provides industry-leading SQL performance and scale. It fully separates compute from storage, allowing independent scaling of both the components. Additionally, it natively stores data in the open Delta Lake format.

Synapse Real-Time Analytics gives you a way to focus and scale up your analytics solution while democratizing data for both citizen data scientists and advanced data engineers. As a fully managed big data analytics platform, Real-Time Analytics utilizes a query language and engine so you can search structured, semi-structured, and unstructured data.

Power BI provides business owners the ability to access all their data in Fabric quickly and intuitively to make better decisions with data. This experience allows organizations to turn unrelated data sources into coherent, visually immersive, and interactive insights.

Data Activator monitors data in Power BI reports and automatically takes actions when certain patterns or conditions are detected. This allows you to build a digital nervous system that acts across all your data, at scale and in a timely manner.

Copilots reduce the heavy lifting

Microsoft Fabric includes several copilots that act as interactive aides, lightening the load on engineers, scientists, and analysts so they can expedite the journey from raw data to meaningful insights. From data preparation to report building, Copilot and other generative AI features offer new ways to analyze data, generate code, and create visualizations in Fabric and Power BI.

Copilot for Data Science and Data Engineering provides intelligent code completion, automates routine tasks, and supplies industry-standard code templates to facilitate tasks like data enrichment and the creation of analytical models. Copilot offers contextual code suggestions and prompts that adapt to specific tasks, helping you code more effectively and with greater ease.

Copilot for Data Factory supports both citizen and professional data wranglers in streamlining their workflow. It provides intelligent code to transform data, as well as code explanations to help you understand complex tasks.

Copilot for Power BI allows you to create Power BI reports automatically. You can generate summaries of existing reports or ask for suggestions on which reports to create based on your data. Prompts like “Create a page to examine next month’s forecast” yield visualizations that help you spot trends and patterns quickly.

LLM capabilities power your generative AI applications

As a platform for analytics and AI, Microsoft Fabric is well suited to support the use of large language models (LLMs) for the creation of generative AI applications. Fabric and SynapseML offer unique LLM capabilities so you can build solutions that can handle question-and-answer tasks or document summaries, for example.

Extract insights from unstructured data: Use Fabric to tap into information stored in unstructured documents like PDFs. You can load PDF documents into a Spark DataFrame, read the documents using the Azure AI Document Intelligence in Azure AI Services, and use SynapseML to split the documents into chunks.

Integrate Azure OpenAI: Apply LLMs at scale by integrating Azure OpenAI Service and SynapseML. Azure OpenAI can be used to solve natural language tasks by prompting the completion API. Through SynapseML, you can use Apache Spark distributed computing framework to easily process millions of prompts.

Generate embeddings: Connect Azure OpenAI Service and use SynapseML to generate embeddings in a distributed manner that allows you to efficiently process large volumes of data. You can also store the embeddings in a vector store using Azure AI Search and search the vector store to answer users' questions.

Get ready for the era of AI

As organizations forge ahead in the era of AI, clean data from well-managed and highly integrated analytics systems is critical. Frameworks like MA²G make sure your systems support governance, data management, and domains, allowing organization to create customized AI and analytics experiences. Through Microsoft Fabric, all the data and analytics tools you need are available in one, end-to-end platform.

Interested in learning more about MA²G and Fabric?

Reach out to an Azure sales specialist or your Microsoft sales representative for best practices on analytics, help getting started with Fabric, and more. Or ask about visiting a local Executive Briefing Center or Microsoft Training Center.