Microsoft

Microsoft Partner Project Ready

Implement with Impact

Modern Data Platform with Azure Databricks

<Speaker name or subtitle>

<Date>

Day 3 of 3

Technofocus

Course Plan and Learning Objectives

Day 1

Module 1 - Introduction to Azure Databricks
- Azure Databricks: A Data Intelligent Platform
- Why Azure Databricks
- Decision guide: Azure Databricks vs. Microsoft Fabric

Module 2 - Migration to Azure Databricks
- Microsoft Cloud Adoption Framework for Azure
- Migration strategies
- Data landing zones
- Migration scenarios

Interactive Simulated Lab Experience
- End-to-End Streaming Pipeline with Lakeflow Declarative Pipelines in Azure Databricks

Day 2

Module 3 - Integration with Azure
- Seamless integration with Microsoft Azure services
- Connect to Azure Data Lake Storage (ADLS) Gen2 and Blob Storage
- Leverage Azure Databricks for Azure Cosmos DB Operations
- Secret management with Azure Key Vault
- Connect Azure Databricks to Azure Event Hubs

Module 4 - Integration with Microsoft Fabric and Power BI
- Data Intelligence with Azure Databricks and Microsoft Fabric
- Connect Power BI to Azure Databricks
- Integration with Azure Data Factory
- Mirroring Azure Databricks Unity Catalog

Interactive Simulated Lab Experience
- Setup and use Unity Catalog for Data Management in Azure Databricks
- Real-Time Streaming with Azure Databricks and Azure Event Hubs

Day 3

Module 5 - Integration with Azure AI Foundry
- Azure Databricks connector in Azure AI Foundry
- Mosaic AI and machine learning on Azure Databricks
- Query Generative AI model serving endpoints
- Databricks Assistant, AI/BI Genie and AI Functions on Azure Databricks
- Chat with LLMs and prototype GenAI apps using AI Playground
- Build and optimize agents on your data with Agent Bricks

Module 6 - Security and Governance
- Integrate Azure Databricks with Microsoft Purview
- Integration of Azure Databricks Unity Catalog with Microsoft Purview

Module 7 - Well-architected for Azure Databricks
- Lakehouse implementation: Principles and best practices
- Azure Databricks well-architected framework

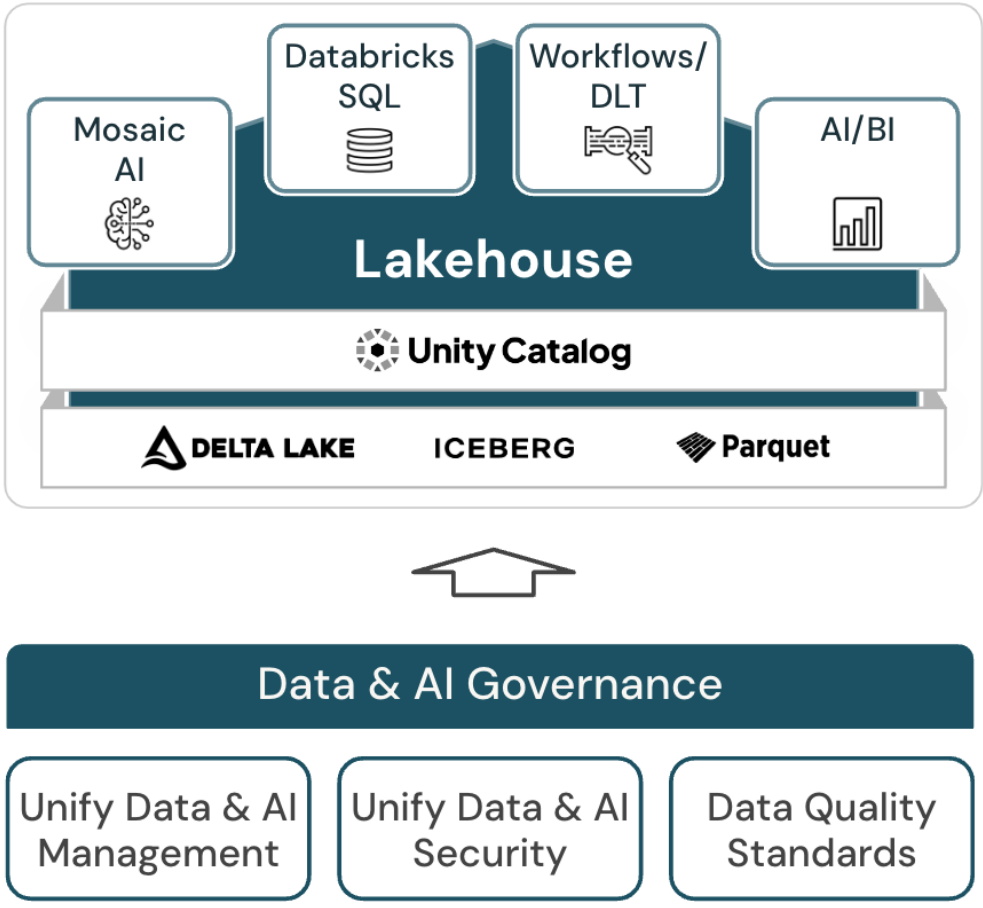
Interactive Simulated Lab Experience
- Responsible AI with Large Language Models using Azure Databricks and Azure OpenAI
- Connect to and manage Azure Databricks in Microsoft Purview

07
Well-architected for Azure Databricks

Lakehouse implementation: Principles and best practices

Principles of Data and AI Governance

The oversight to ensure that data and AI bring value and support your business strategy



Best practices for Data and AI Governance

Unify data and AI management

- Establish a data and AI governance process
- Manage metadata for all data and AI assets in one place
- Track data and AI lineage to drive visibility of the data
- Add consistent descriptions to your metadata
- Allow easy data discovery for data consumers
- Govern AI assets together with data

Unify data and AI security

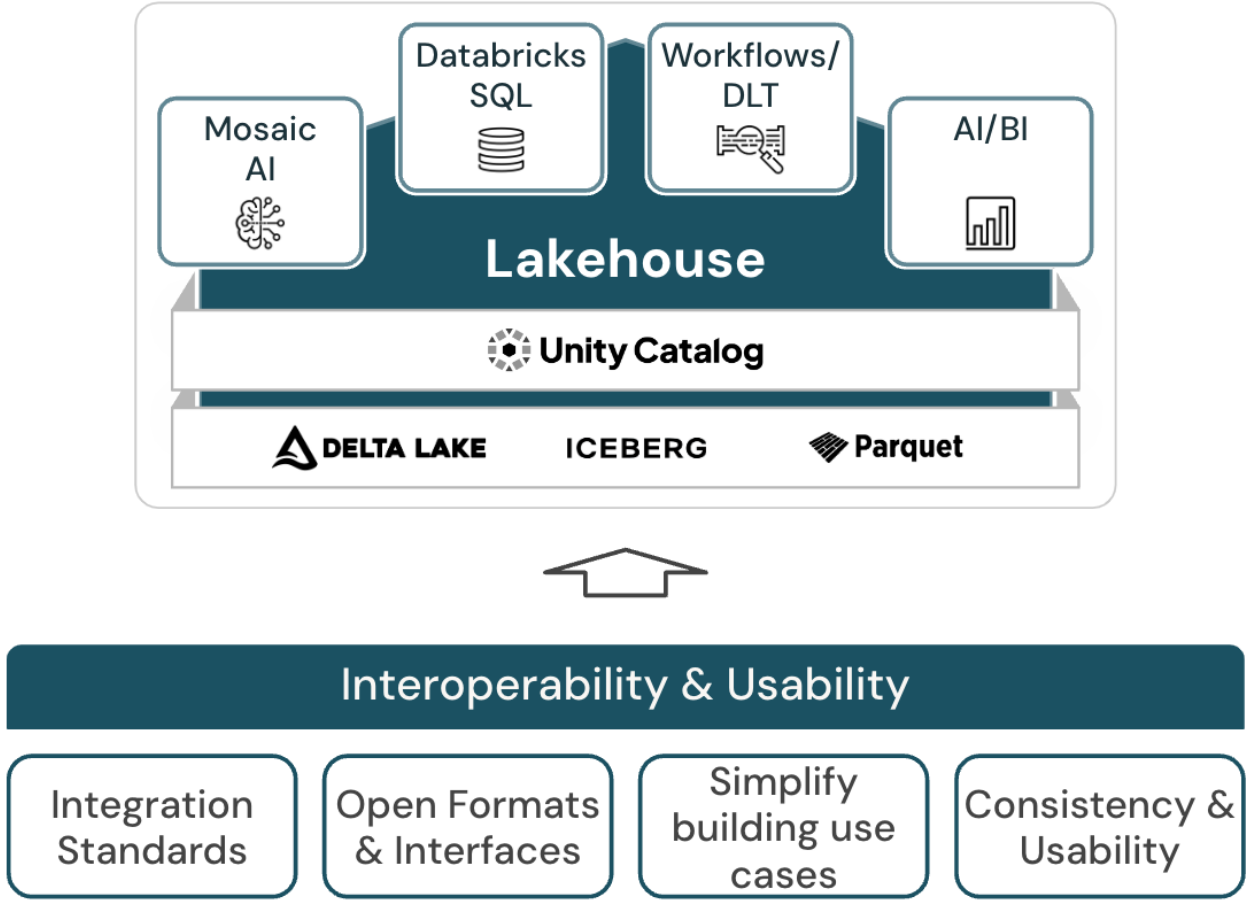
- Centralize access control for all data and AI assets
- Configure audit logging
- Audit data platform events

Establish data quality standards

- Define clear data quality standards
- Use data quality tools for profiling, cleansing, validating, and monitoring data
- Implement and enforce standardized data formats and definitions

Principles of Interoperability and Usability

- The ability of the lakehouse to interact with users and other systems



Best practices for Interoperability and Usability

Define standards for integration

- Use standard and reusable integration patterns for external integration
- Use optimized connectors to ingest data sources into the lakehouse
- Use certified partner tools
- Reduce complexity of data engineering pipelines
- Use infrastructure as code (IaC) for deployments and maintenance

Use open interfaces and open data formats

- Use open data formats
- Enable secure data and AI sharing for all data assets
- Use open standards for your ML lifecycle management

Best practices for Interoperability and Usability

Simplify new use case implementation

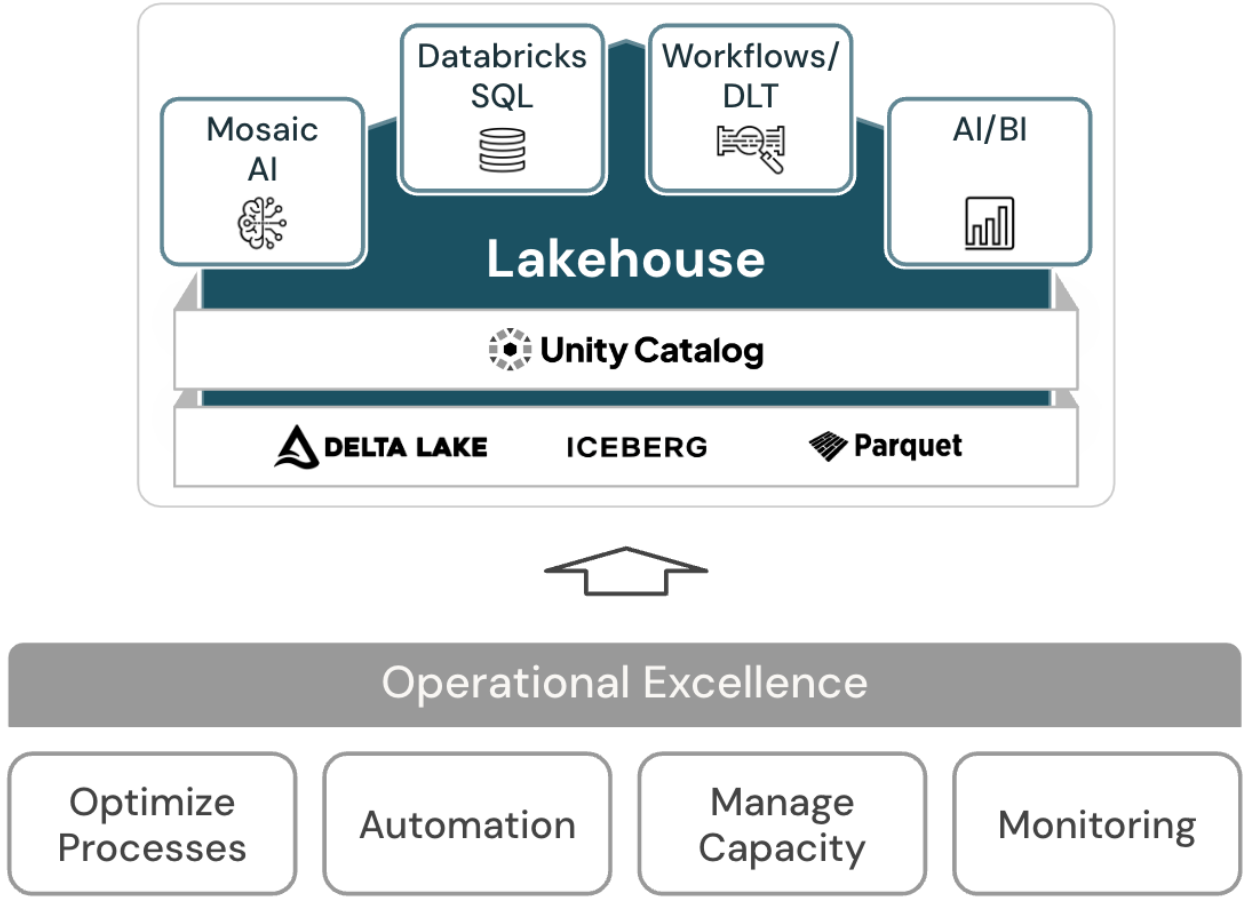
- Provide a self-service experience across the platform
- Use serverless compute
- Use predefined compute templates
- Use AI capabilities to increase productivity

Ensure data consistency and usability

- Offer reusable data-as-products that the business can trust
- Publish data products semantically consistent across the enterprise
- Provide a central catalog for discovery and lineage

Principles of Operational excellence

- Operational excellence addresses the ability to operate the lakehouse efficiently



Best practices for Operational excellence

Optimize build and release processes

- Create a dedicated Lakehouse operations team
- Use Enterprise source code management (SCM)
- Standardize DevOps processes (CI/CD)
- Standardize MLOps processes
- Define environment isolation strategy
- Define catalog strategy for your enterprise

Automate deployments and workloads

- Use infrastructure as code (IaC) for deployments and maintenance
- Standardize compute configurations
- Use Unity Catalog managed tables
- Use automated workflows for jobs
- Use automated and event-driven file ingestion
- Use ETL frameworks for data pipelines
- Follow the deploy-code approach for ML workloads
- Use a model registry to decouple code and model lifecycle
- Automate ML experiment tracking
- Reuse the same infrastructure to manage ML pipelines
- Utilize declarative management for complex data and ML projects

Best practices for Operational excellence

Manage capacity and quotas

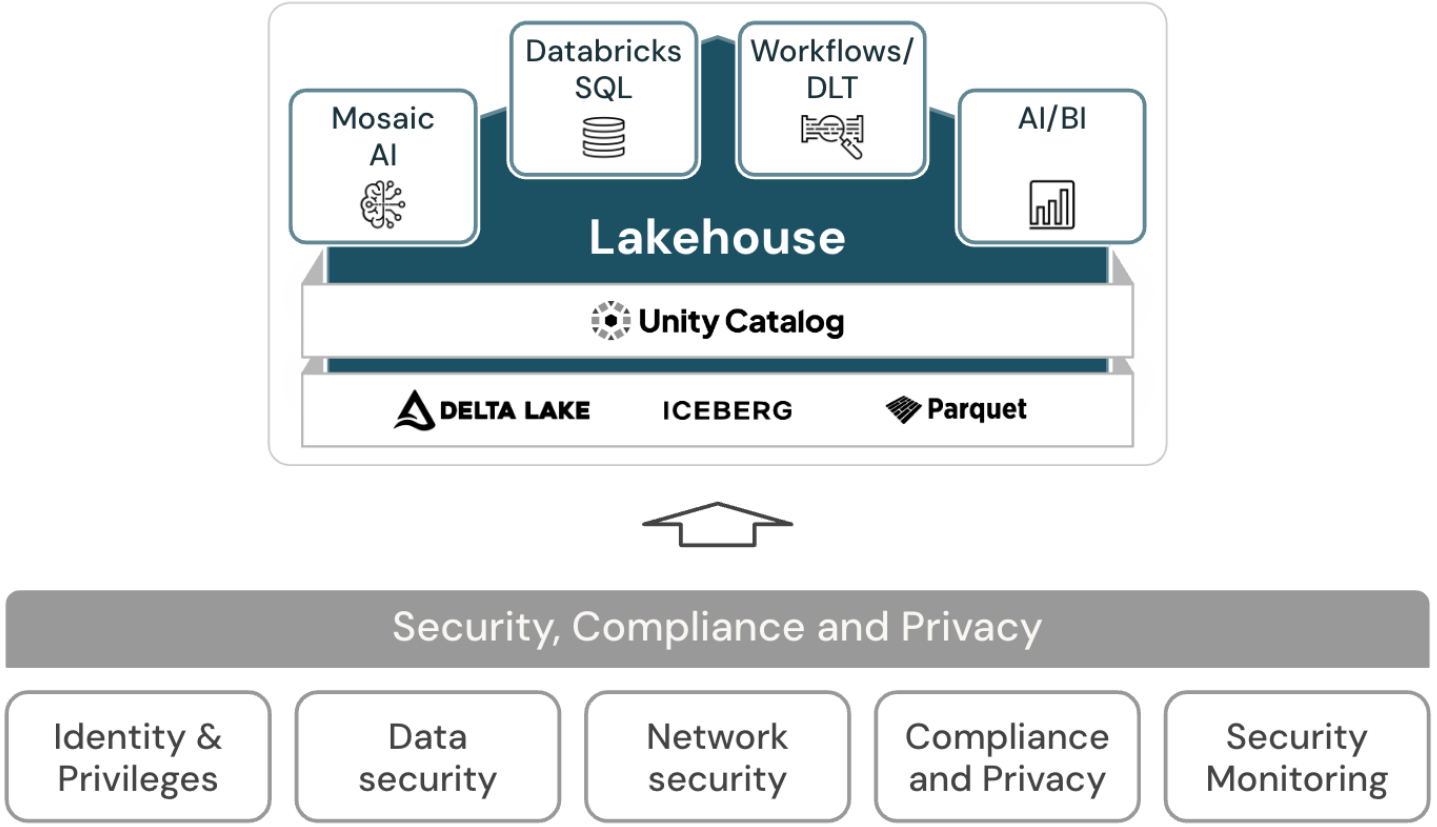
- Manage service limits and quotas
- Invest in capacity planning

Set up monitoring, alerting, and logging

- Establish monitoring processes
- Use native and external tools for platform monitoring

Principles of Security, Compliance, and Privacy

- Protect the Azure Databricks application, customer workloads, and customer data from threats



Best practices for Security, Compliance, and Privacy

Manage identity and access using least privilege

- Leverage multi-factor authentication
- Use SCIM to synchronize users and groups
- Limit the number of admin users
- Enforce segregation of duties between administrative accounts
- Restrict workspace admins
- Manage access according to the principle of least privilege
- Use OAuth or Azure Entra ID token authentication
- Enforce token management
- Restrict cluster creation rights
- Use compute policies
- Use service principals to run administrative tasks and production workloads
- Use compute that supports user isolation
- Store and use secrets securely

Protect data in transit and at rest

- Centralize data governance with Unity Catalog
- Use Azure Managed Identities to access storage
- Plan your data isolation model
- Avoid storing production data in DBFS
- Configure Azure Storage firewalls
- Prevent anonymous read access and apply other protections
- Enable soft deletes and other data protection features
- Backup your Azure Storage data
- Configure customer-managed keys for managed services
- Configure customer-managed keys for storage
- Use Delta Sharing
- Configure a Delta Sharing recipient token lifetime
- Additionally encrypt sensitive data at rest using Advanced Encryption Standard (AES)
- Leverage data exfiltration prevention settings within the workspace
- Use Clean Rooms to collaborate in a privacy-safe environment

Best practices for Security, Compliance, and Privacy

Secure your network and protect endpoints

- Use Secure Cluster Connectivity (No Public IP)
- Deploy Azure Databricks into your own Azure virtual network
- Configure IP access lists
- Use Azure PrivateLink
- Implement network exfiltration protections
- Isolate Azure Databricks workspaces into different networks
- Configure a firewall for serverless compute access
- Restrict access to valuable codebases to only trusted networks
- Use Virtual network encryption

Meet compliance and data privacy requirements

- Restart compute on a regular schedule
- Isolate sensitive workloads into different workspaces
- Assign Unity Catalog securables to specific workspaces
- Implement fine-grained access controls
- Apply tags
- Use lineage
- Use Enhanced Security Monitoring or Compliance Security Profile
- Control and monitor workspace access for Azure Databricks personnel
- Implement and test a Disaster Recovery strategy
- Consider the use of Azure Confidential Compute

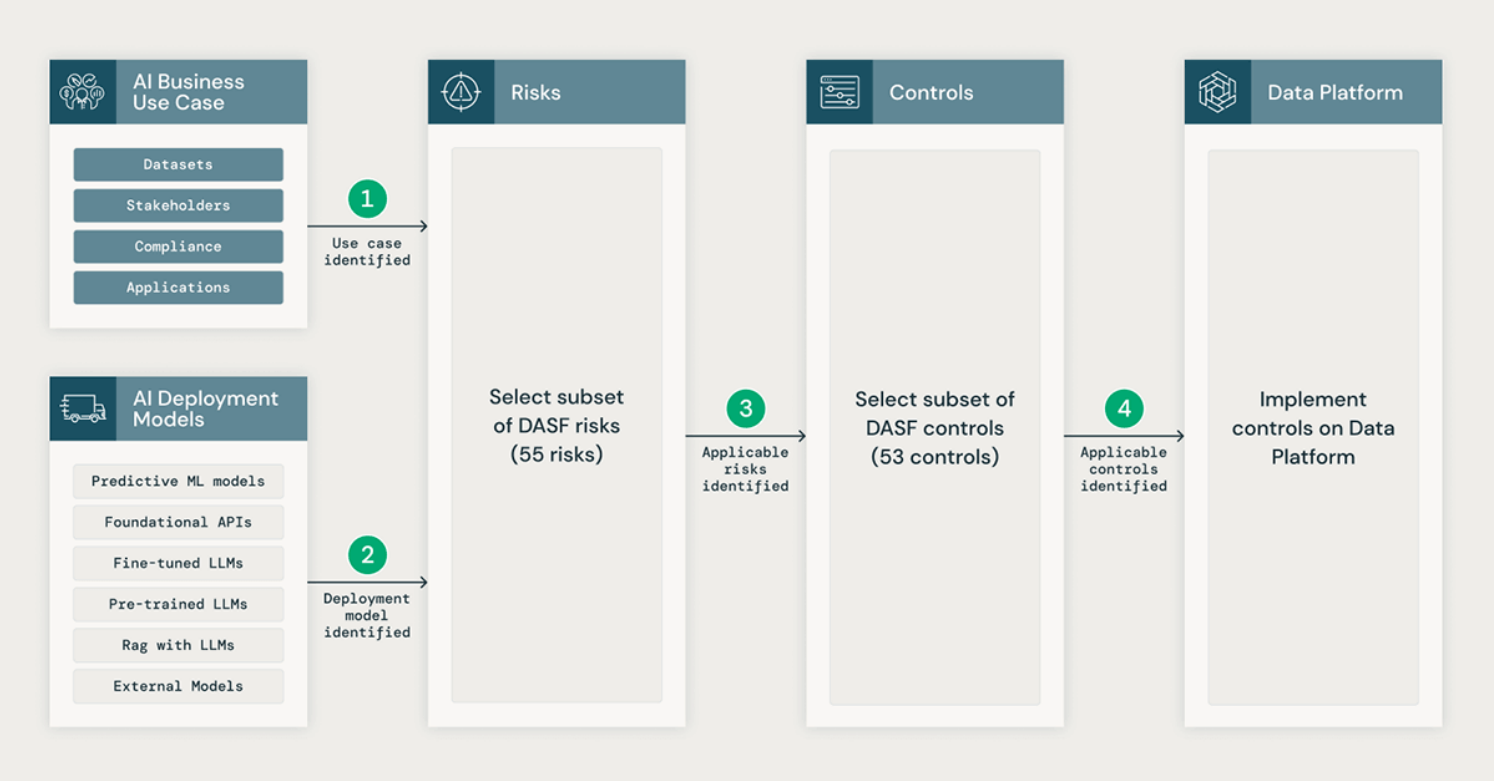
Best practices for Security, Compliance, and Privacy

Monitor system security

- Leverage system tables
- Monitor system activities via Azure logs
- Enable verbose audit logging
- Manage code versions with Git folders
- Restrict usage to trusted code repositories
- Provision infrastructure via infrastructure-as-code
- Manage code via CI/CD
- Control library installation
- Use models and data from only trusted or reputable sources
- Implement DevSecOps processes
- Use lakehouse monitoring
- Use inference tables and AI Guardrails
- Use tagging as part of your cost monitoring and charge-back strategy
- Use budgets to monitor account spending
- Use Azure Policy to create “upper limit” resource controls

Databricks AI Security Framework (DASF)

- Demystify AI and ML
- Provide a defense-in-depth approach to securing AI
- Deliver actionable recommendations

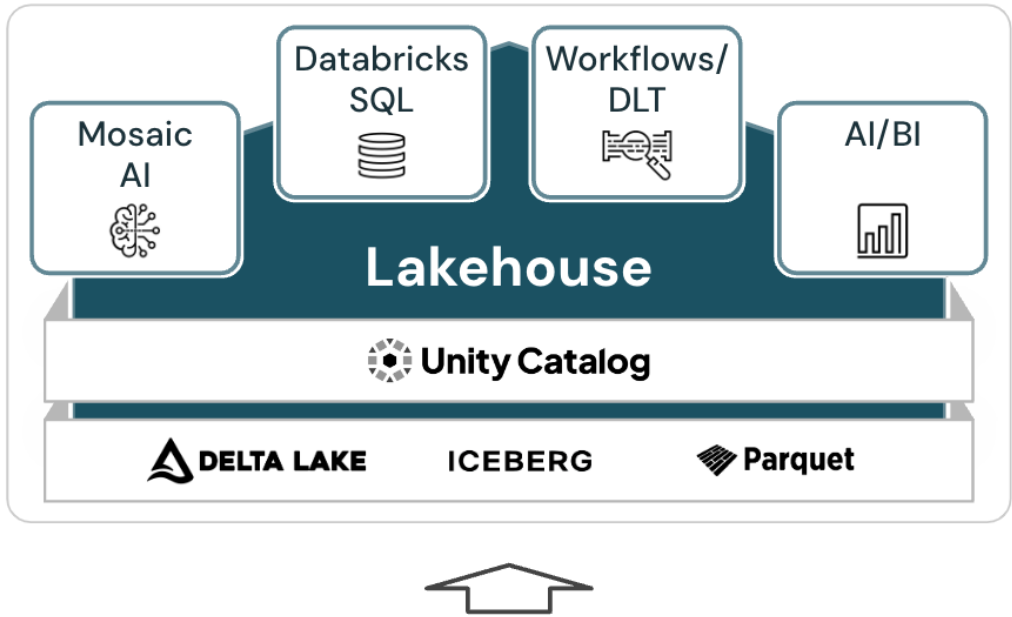


Azure Security Benchmark (ASB)

Provides prescriptive best practices and recommendations to help improve the security of workloads, data, and services on Azure. Security guidance includes:

- Cloud Adoption Framework
- Azure Well-Architected Framework
- Microsoft Security Best Practices
- Microsoft Cybersecurity Reference Architectures (MCRA)

Principles of Reliability



Reliability

- Design for Failure
- Manage Data Quality
- Scaling
- Recovery Procedures
- Automation
- Monitoring

Best practices for Reliability

Design for failure

- Use a data format that supports ACID transactions
- Use a resilient distributed data engine for all workloads
- Automatically rescue invalid or nonconforming data
- Configure jobs for automatic retries and termination
- Use scalable and production-grade model serving infrastructure
- Use managed services where possible

Manage data quality

- Use a layered storage architecture
- Improve data integrity by reducing data redundancy
- Actively manage schemas
- Use constraints and data expectations
- Take a data-centric approach to machine learning

Best practices for Reliability

Design for autoscaling

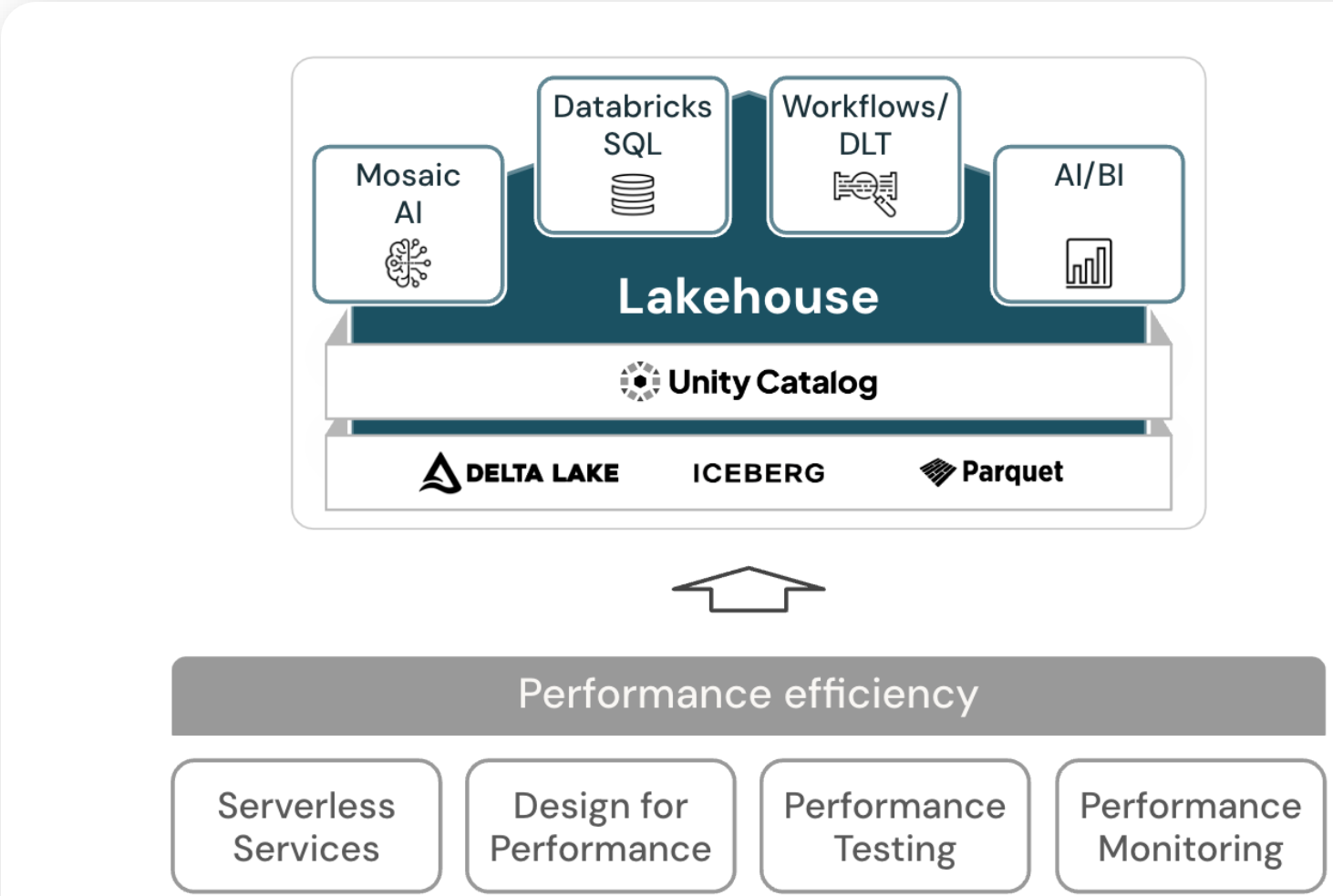
- Enable autoscaling for ETL workloads
- Enable autoscaling for SQL warehouse

Test recovery procedures

- Recover from Structured Streaming query failures
- Recover ETL jobs using data time travel capabilities
- Leverage a job automation framework with built-in recovery
- Configure a disaster recovery pattern

Principles of Performance efficiency

- The ability of a system to adapt to changes in the load



Distributed computing concepts

Vertical scaling

Scale vertically by adding or removing resources from a single machine, typically CPUs, memory, or GPUs

Horizontal scaling

Scale horizontally by adding or removing nodes from a distributed system

Linear scalability

When you add more resources to a system, the relationship between throughput and resources used is linear

Best practices for Performance efficiency

Use serverless architectures

- Use serverless compute
- Use an enterprise grade model serving service

Design workloads for performance

- Understand your data ingestion and access patterns
- Use parallel computation where it is beneficial
- Analyze the whole chain of execution
- Prefer larger clusters
- Use predictive optimization
- Use Unity Catalog managed tables
- Use native Spark operations
- Use native platform engines
- Understand your hardware and workload type
- Use caching
- Use compaction
- Use data skipping
- Avoid over-partitioning
- Optimize join performance
- Run analyze table to collect table statistics

Best practices for Performance efficiency

Run performance testing in the scope of development

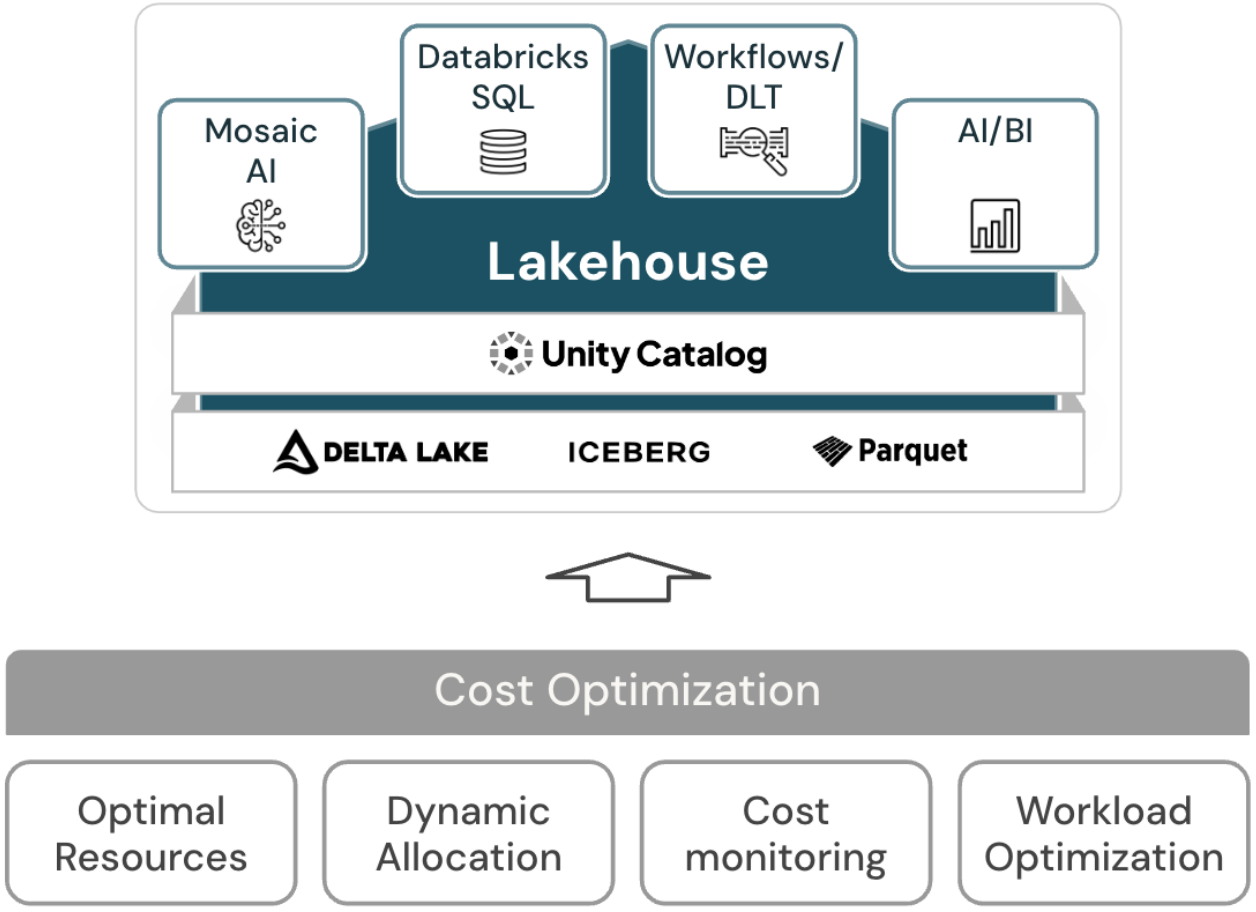
- Test on data representative of production data
- Consider prewarming resources
- Identify bottlenecks

Monitor performance

- Monitor query performance
- Monitor streaming workloads
- Monitor job performance

Principles of Cost optimization

- Manage costs to maximize the value delivered



Best practices for Cost optimization

Choose optimal resources

- Use performance optimized data formats
- Use job compute
- Use SQL warehouse for SQL workloads
- Use up-to-date runtimes for your workloads
- Only use GPUs for the right workloads
- Use serverless services for your workloads
- Use the right instance type
- Choose the most efficient compute size
- Evaluate performance-optimized query engines

Dynamically allocate resources

- Use auto-scaling compute
- Use auto termination
- Use compute policies to control costs

Best practices for Cost optimization

Monitor and control cost

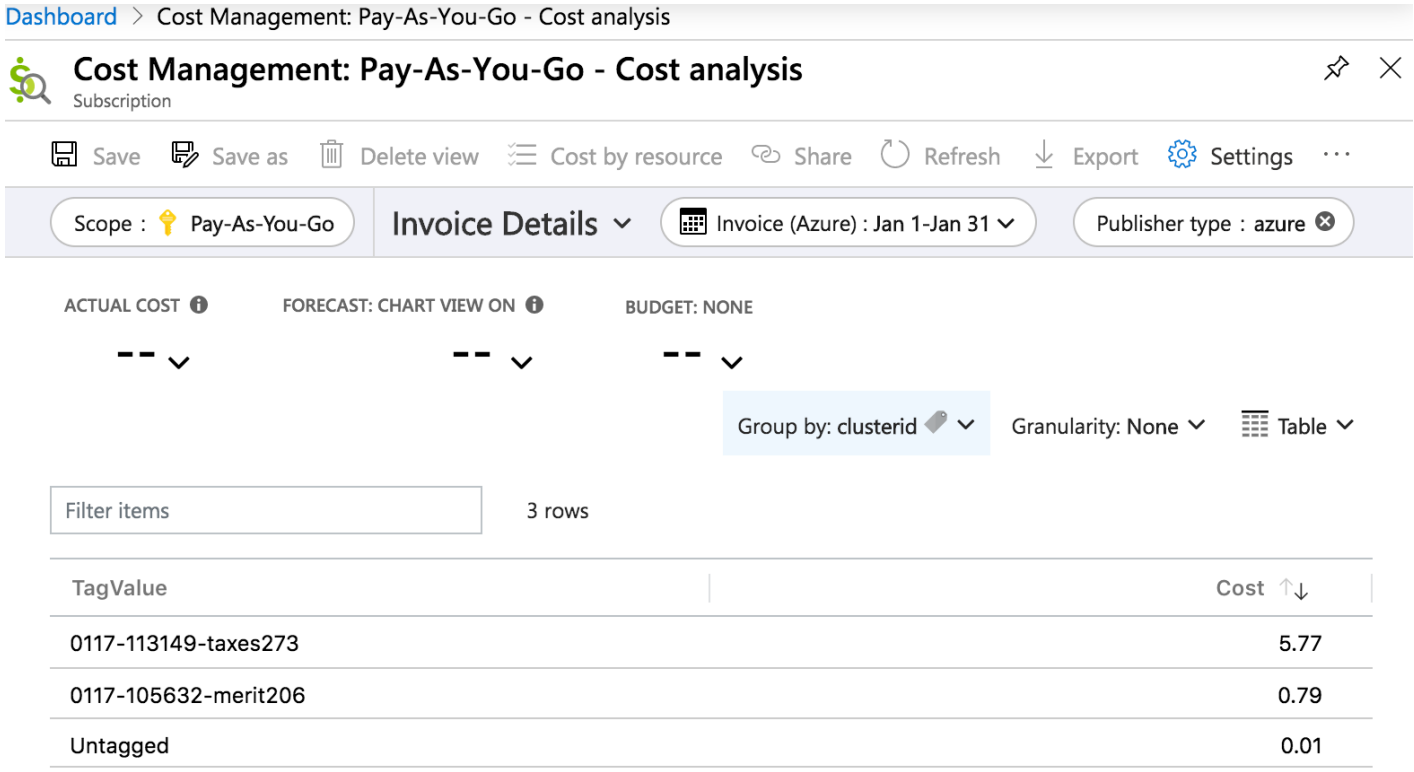
- Setup tagging for cost attribution
- Set up budgets and alerts to enable monitoring of account spending
- Monitor costs to align spending with expectations
- Manage costs to align usage with organizational needs

Design cost-effective workloads

- Balance always-on and triggered streaming
- Balance between on-demand and capacity excess instances

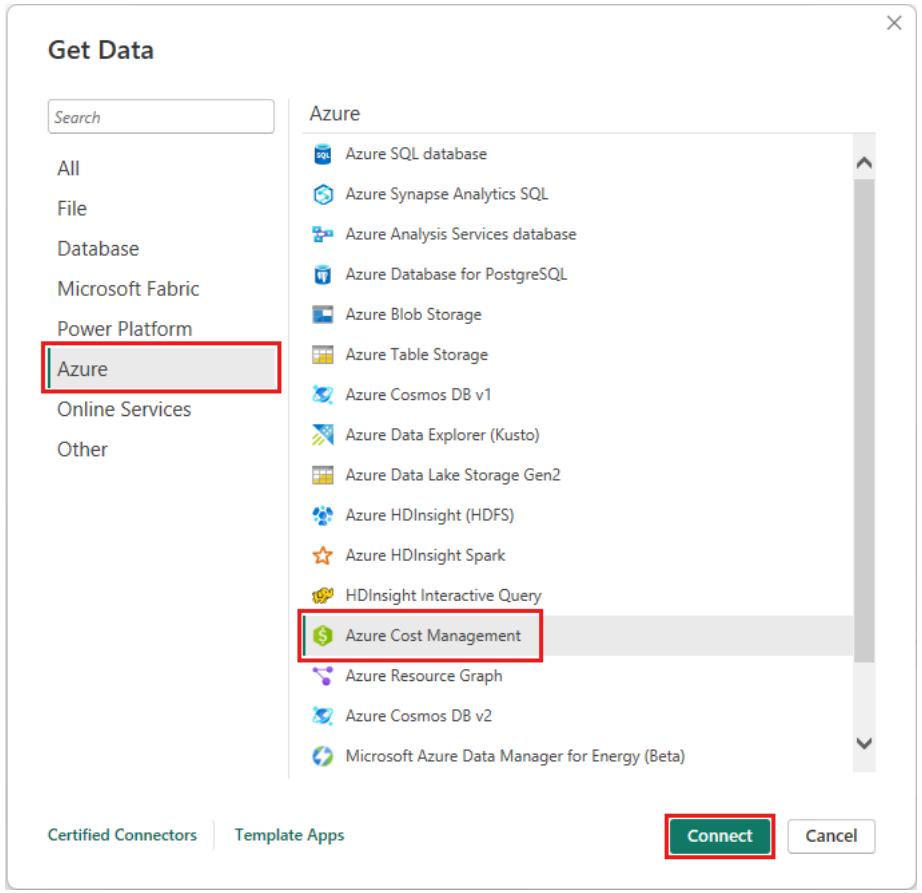
Use tags to attribute and track usage

- **Default tags**: Provide basic metadata like vendor, cluster ID, and creator.
- **Custom tags**: Allow for granular tracking, reporting, and budgeting



Create visuals and reports with the Microsoft Cost Management connector in Power BI Desktop

- Use the Microsoft Cost Management connector for Power BI Desktop to make powerful, customized visualizations and reports that help you better understand your Azure spend
- Currently supports customers with
  - A direct Microsoft Customer Agreement (MCA)
  - An Enterprise Agreement (EA)
  - A Microsoft Partner Agreement



Best practices for users and administrators

Best practices for users

- Delta Lake
- Hyperparameter tuning with Optuna or Ray Tune
- Deep learning in Databricks
- CI/CD

Best practices for administrators

- Introduction to the well-architected data lakehouse
- Cluster configuration
- GDPR and CCPA compliance using Delta Lake
- Identity
- Pools
- Unity Catalog

Best practices for Delta Lake

- Use Unity Catalog managed tables
- Use predictive optimization
- Use liquid clustering
- Remove legacy Delta configurations
- Do not use Spark caching with Delta Lake
- Improve performance for Delta Lake merge
- Reduce the search space for matches
- Control the shuffle partitions for writes
- Enable optimized writes
- Tune file sizes in table

Hyperparameter tuning with Optuna

- An open-source Python library for hyperparameter tuning that can be scaled horizontally across multiple compute resources
- Integrates with MLflow for model and trial tracking and monitoring

```
%pip install optuna
%pip install optuna-integration # Integration with MLflow
```

Python

```python
def objective(trial):
    x = trial.suggest_float("x", -10, 10)
    return (x - 2) ** 2

study = optuna.create_study()
study.optimize(objective, n_trials=100)

best_params = study.best_params
```

Best practices for deep learning on Azure Databricks

Best practices for loading data

- Use Delta Lake tables for data storage
- Use Mosaic Streaming for data loading on PyTorch or Mosaic Composer

Best practices for training deep learning models

- Start with a Single Node cluster
- Use TensorBoard and cluster metrics to monitor the training process
- Optimize performance for deep learning
- Move to distributed training

Best practices for inference

- Online serving
- Batch and streaming inference

Best practices for CI/CD on Azure Databricks

- Version control everything
- Automate testing
- Employ Infrastructure as Code (IaC)
- Isolate environments
- Choose tools that match your cloud ecosystem
- Monitor and automate rollbacks
- Unify asset management

Best practices for Compute configuration

Use compute policies

Compute policies let you create preconfigured compute resources designed for specific purposes, such as personal compute, shared compute, power users, and jobs

Compute sizing considerations

- Total executor cores
- Total executor memory
- Executor local storage

Best practices for Identity

- Configure users, service principals, and groups
- Sync users and groups from Microsoft Entra ID to your Azure Databricks account
- Enable identity federation
- Assign groups workspace permissions

Best practices for Pool

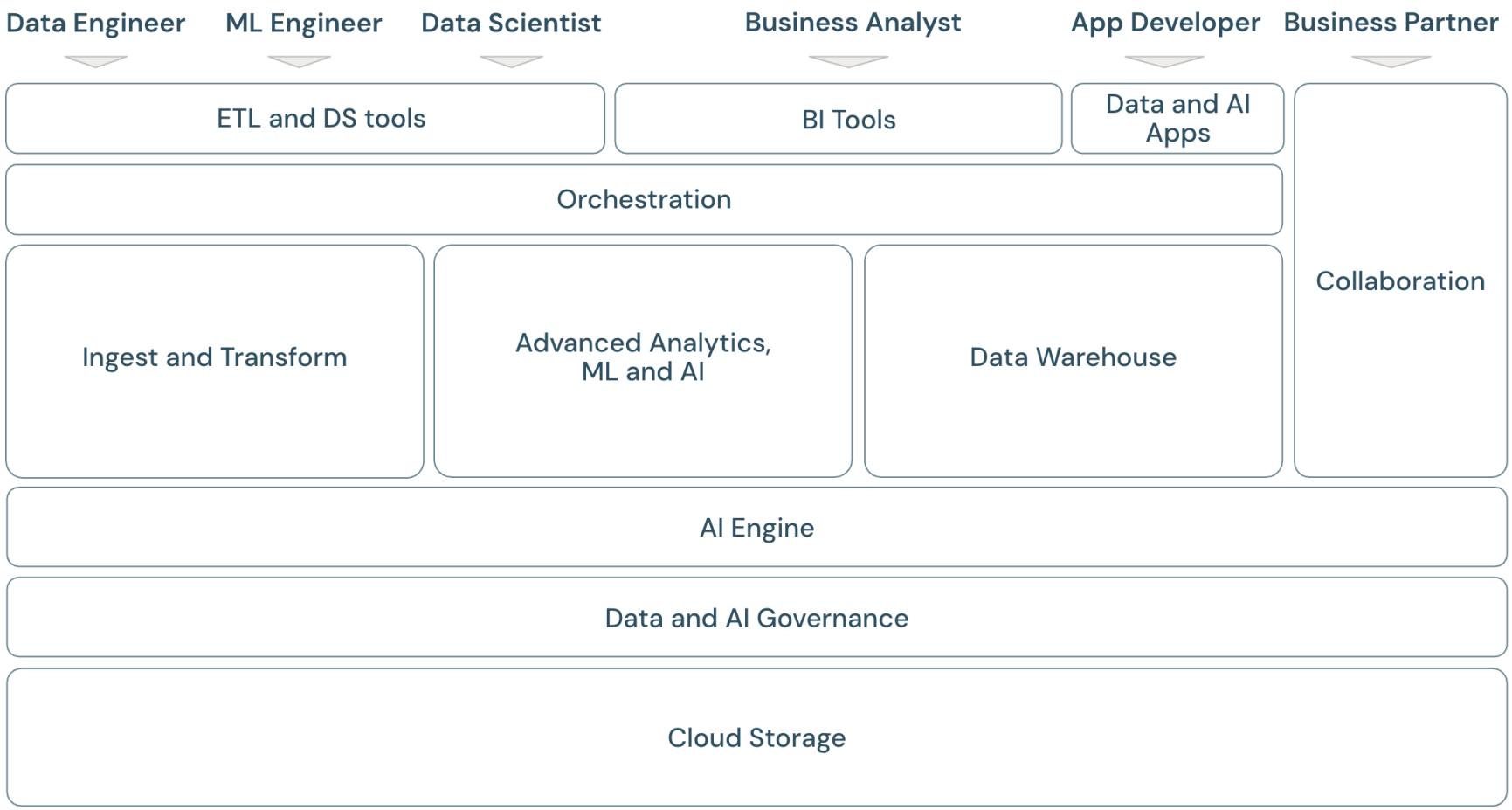
- Create pools based on workloads
- Use spot instance pools
- Tag pools to manage cost and billing
- Configure pools to control cost
- Pre-populate pools

Best practices for Unity Catalog

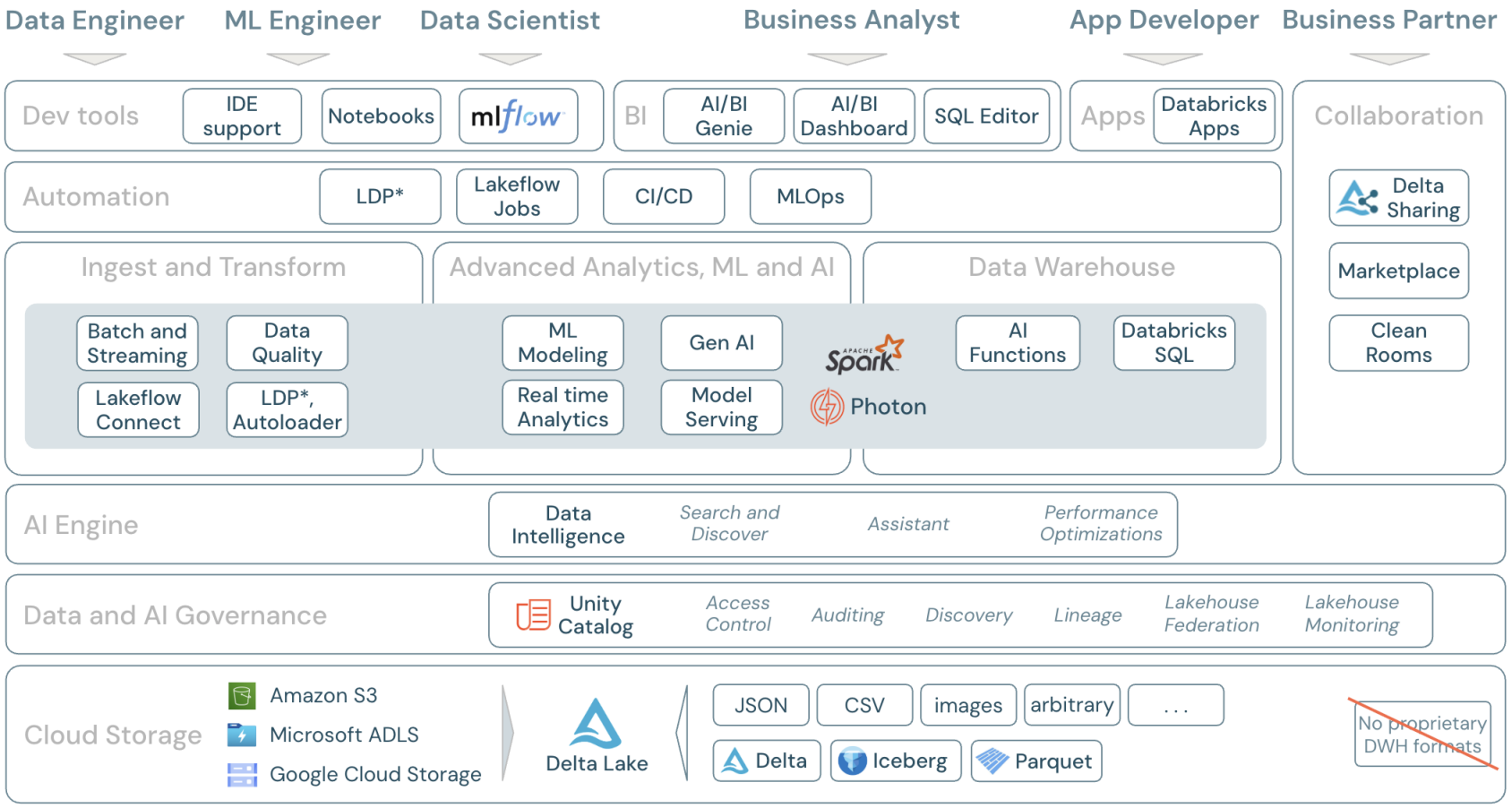
- Plan your data isolation model
- Configure a Unity Catalog metastore
- Organize your data
- Manage external locations, external tables, and external volumes
- Manage compute configurations
- Share data securely using Delta Sharing

Azure Databricks well-architected framework

A modern data and AI platform framework

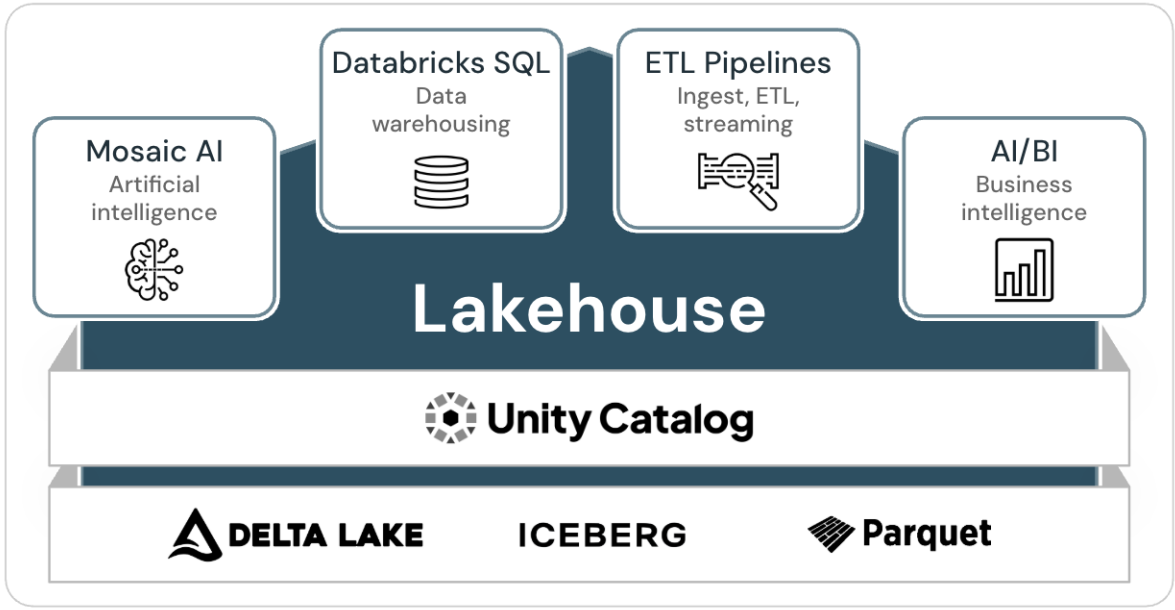


The scope of the Azure Databricks Platform



* Lakeflow Declarative Pipelines

Azure Databricks well-architected framework



Lakehouse specific pillars		*Common pillars*				
Data Governance	Interoperability and Usability	Operational Excellence	Security, Privacy, Compliance	Reliability	Performance Efficiency	Cost Optimization

Interactive Simulated Lab Experience

Interactive Simulated Lab Experience - Advantages

Realistic, interactive simulations eliminate dependencies on trial subscriptions and Azure Passes

Consistent learning experience unaffected by real-time cloud portal updates

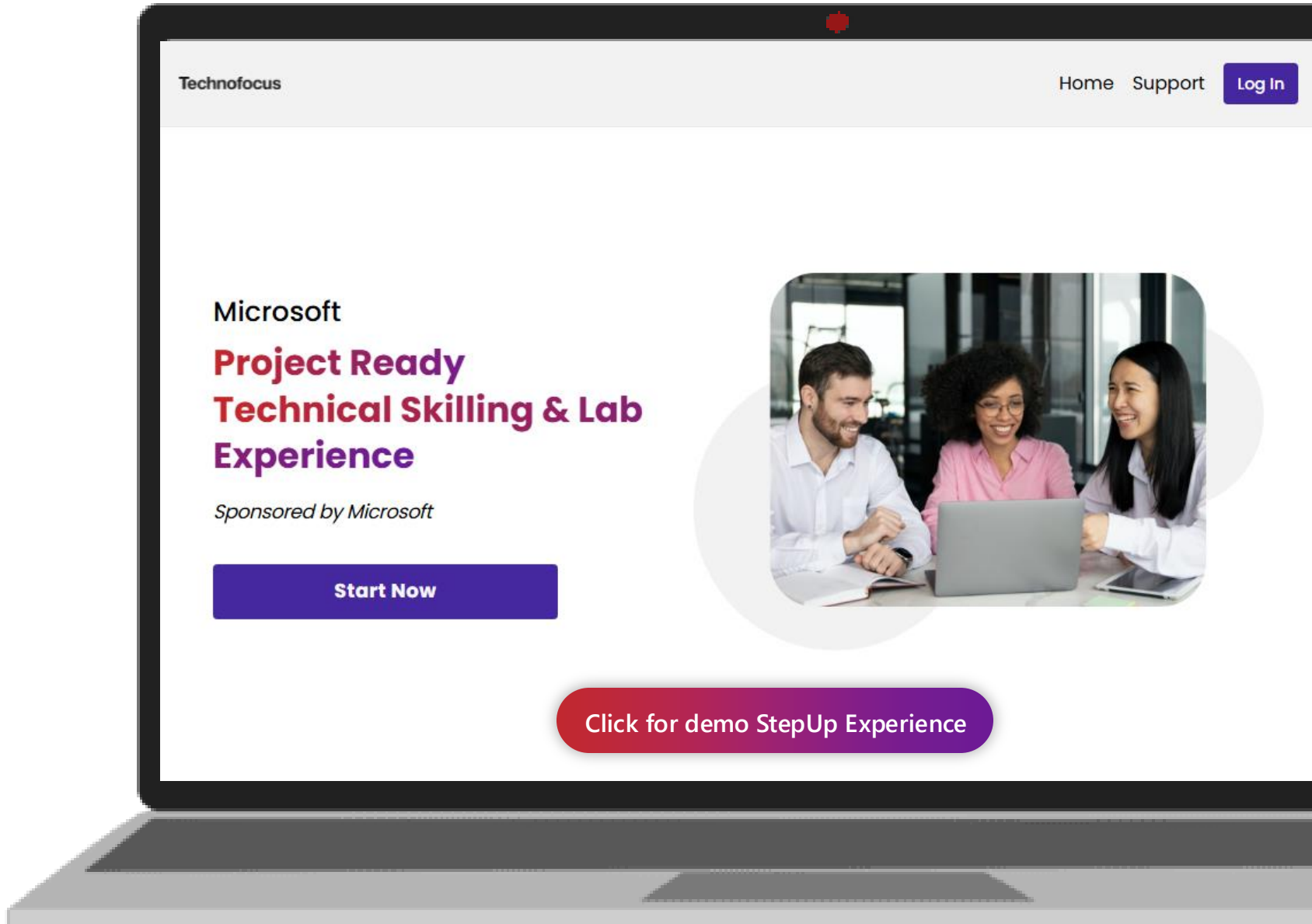
No firewall or security restrictions – available in any standard web browser

Course Availability - FAQs

Do I need to install any software or check system requirements?
Now, the simulated labs are accessible directly from a web browser with no need for software installation or technical setup on the learner's PC.

How is the simulated lab experience integrated into the learning process?
The simulated lab experience is seamlessly accessible from ON24, allowing learners to access simulated labs simultaneously during a live class.

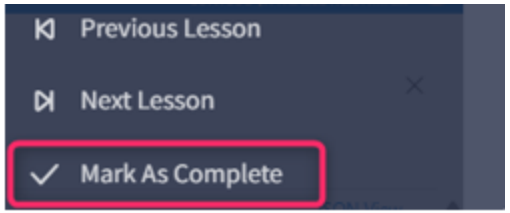
Can I revisit the course and simulated labs after the live session?
Yes, simulated lab access remains available for 30 days from the class start date, ensuring flexibility for learners to revisit and reinforce concepts.



Best Practices- Interactive Simulated Labs

1. Execute the simulated labs in **full-screen mode.**

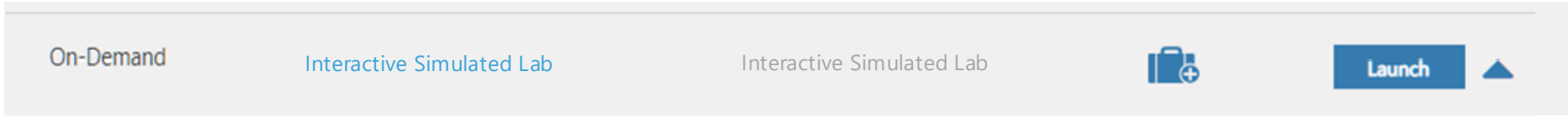
2. After completing the lab, click on the **vertical ellipsis (⋮)** on the title bar, select **Mark as Complete**, then click **CLOSE** to exit.



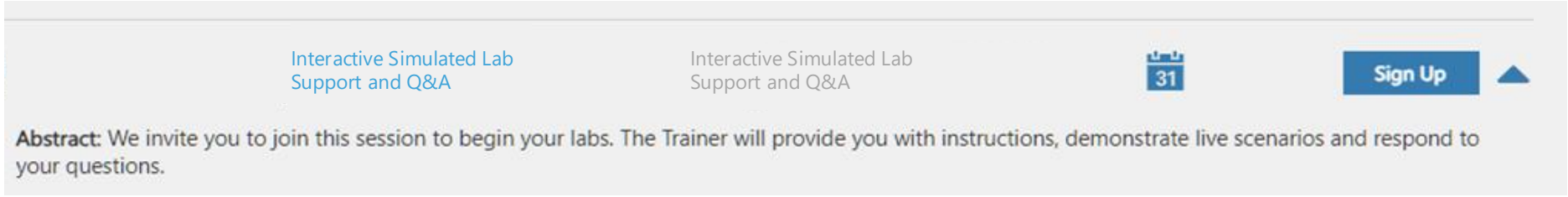
Interactive Simulated Labs

Please note that the interactive simulated labs are only open to a limited number of Microsoft partner participants and are offered on a **first-come-first-served basis**.

Step 1 : Click on **Launch** in the event curriculum to launch your interactive simulated labs



Step 2 : Join the **Live session for Interactive Simulated Lab Support and Q&A Session**

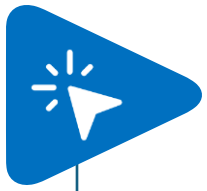


Interactive Simulated Lab Experience

Lab 4 - Responsible AI with Large Language Models using Azure Databricks and Azure OpenAI

Lab 5 - Connect to and manage Azure Databricks in Microsoft Purview

Troubleshooting Simulated Labs connectivity



Launch test lab:
https://stepup.technofocus.ai/learn/testintcours1

Optimized for resolution:
1440 x 900

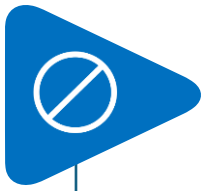
Use a supported operating system:

- Windows 10 or later
- macOS 10.12 or later

- Also supported on Tablets and Mobile devices with large screen.

Use a supported browser

- Chrome (**preferred**)
- Microsoft Edge



Ensure connection is not blocked by your company VPN/Firewall rules

URLs to be whitelisted:
*.navattic.com
labs.technofocus.ai

LevelUp: Accelerating Partner Success with On-Demand Skilling

On-Demand Partner Enablement

Access comprehensive training resources anytime to accelerate partner onboarding and skill development.

Be Project-Ready

Equip teams with the knowledge and tools needed to confidently deliver on real-world projects from day one.

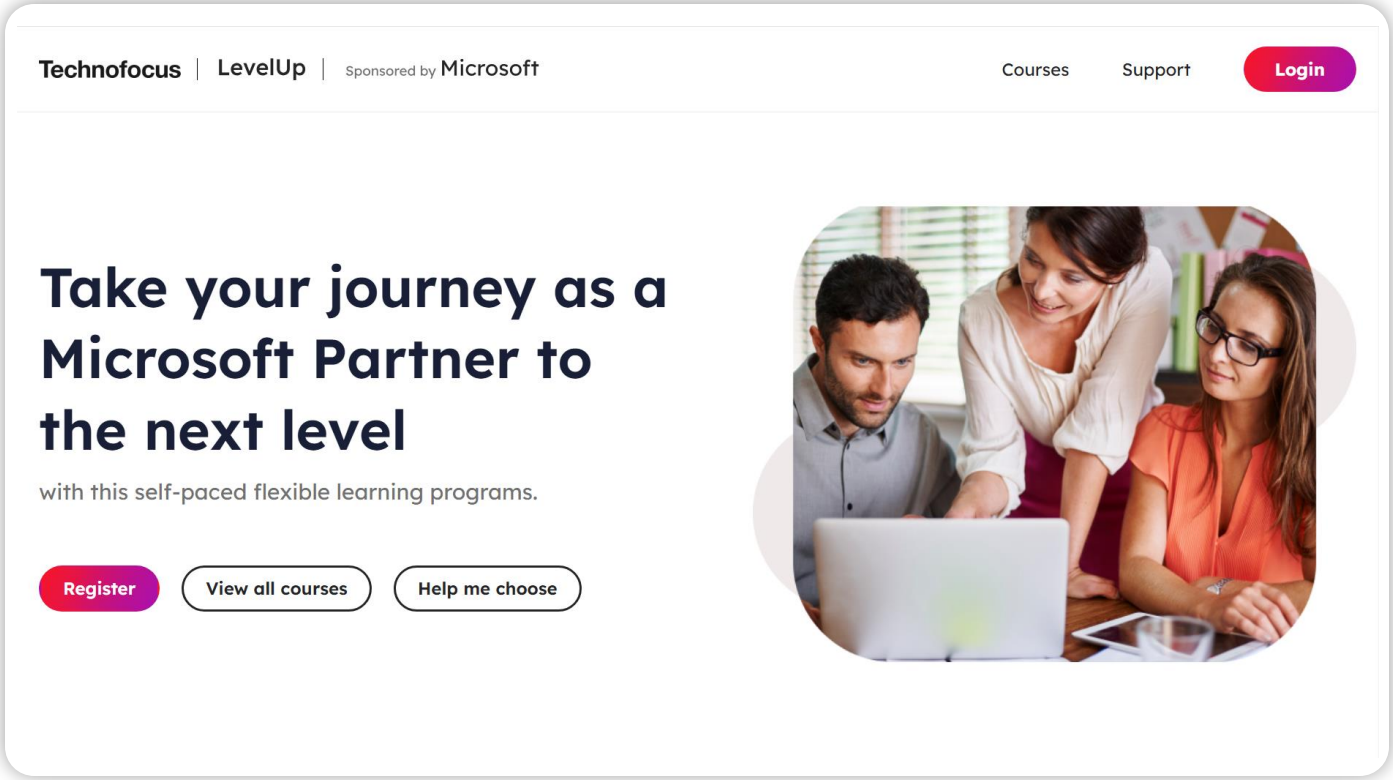
Integrated Hands-On Labs

Reinforce learning through immersive, practical lab environments that simulate real business scenarios.

End-to-End Enablement,Sales, Presales & Technical

Deliver role-specific training that empowers sales, presales,and technical teams with the skills to drive results.

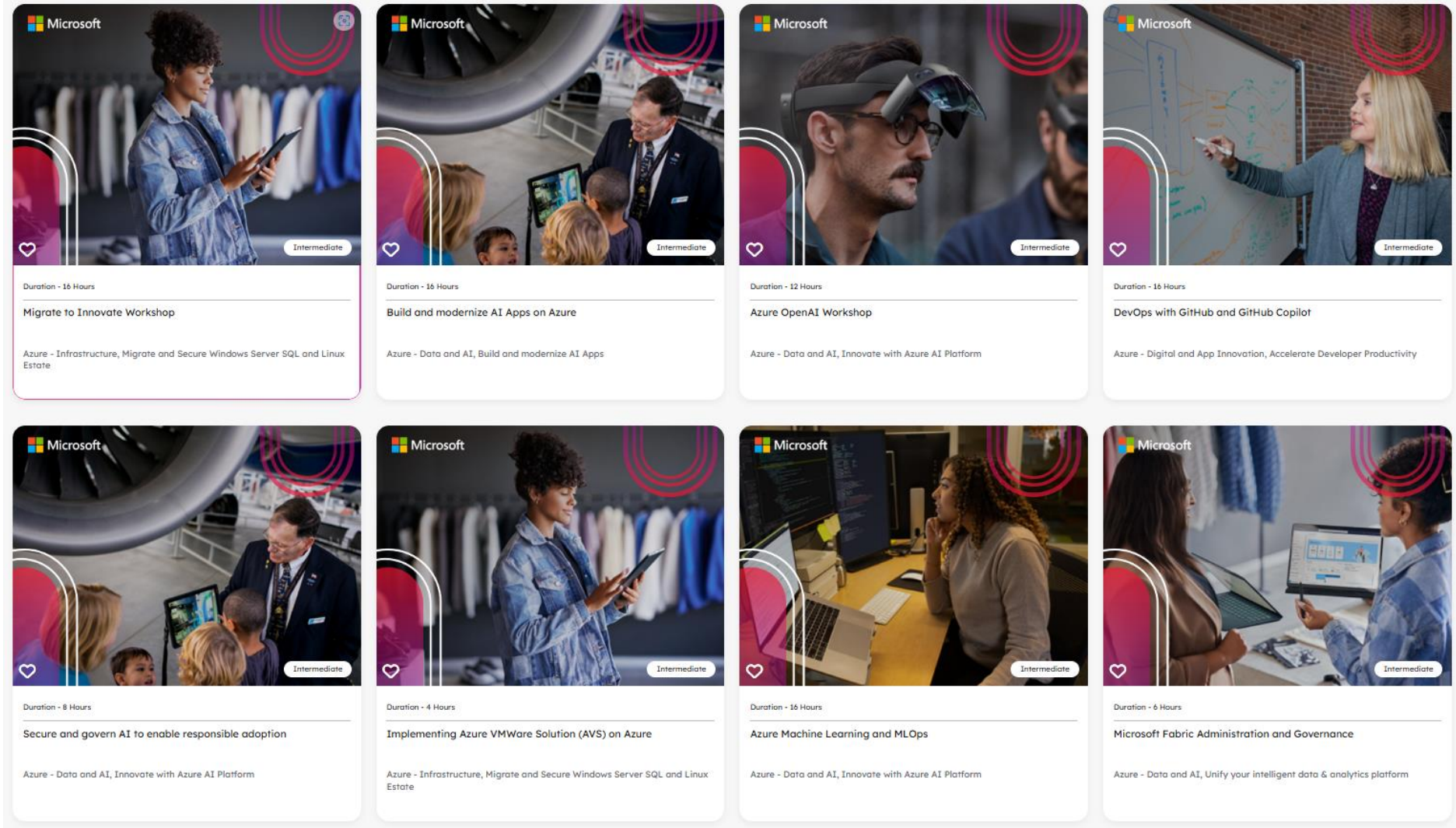
Sign Up today
skillupwithlevelup.com



18K+	55+	84	10+
Enrollments	AI courses	Courses with Labs	Years of successful training

Azure courses on Microsoft LevelUp

ENROLL TODAY



Microsoft

Technofocus

Thank You!