Microsoft

Azure Data & ML strategy

Frederic Gisbert
Cloud Solution Architect
Data Analytics
frgisber@microsoft.com

Narjes Majdoub
Cloud Solution Architect
Data AI
nmajdoub@microsoft.com

Agenda

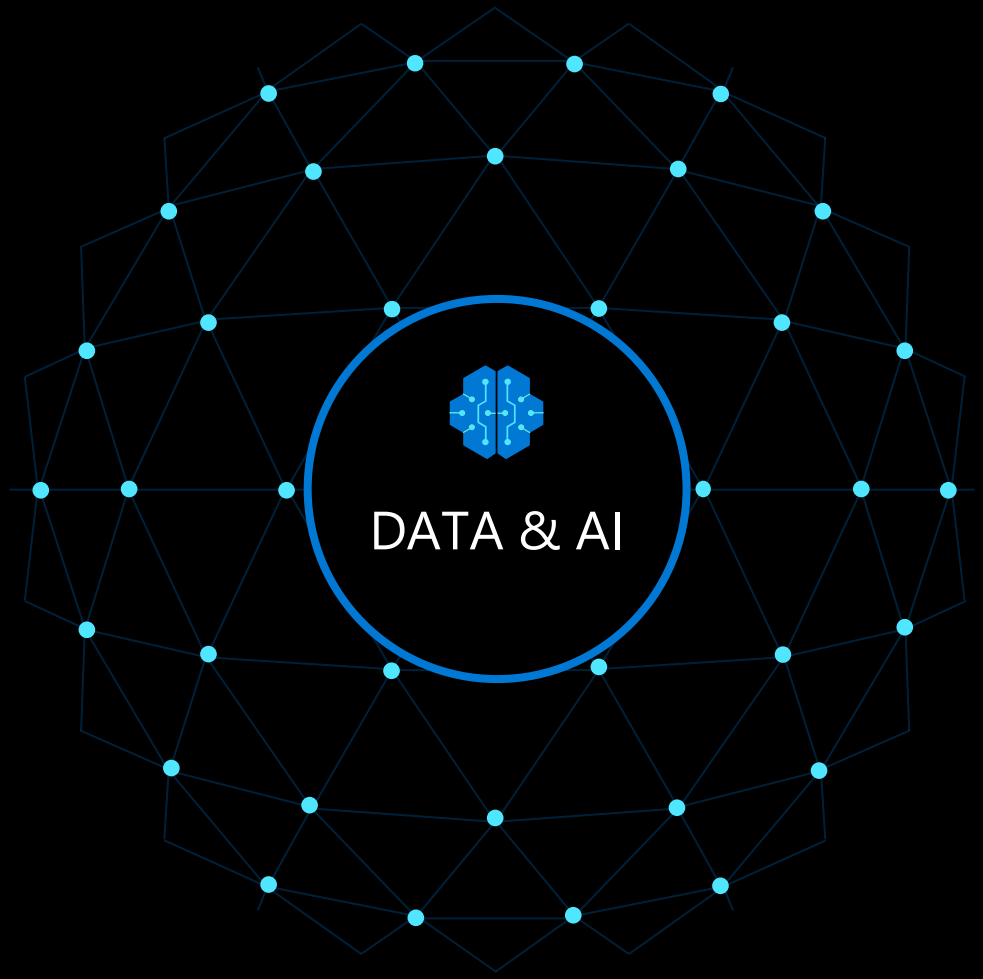
1. Azure data platform overview
2. Azure Synapse Analytics & roadmap
3. Spark in Azure Synapse
4. Azure ML overview & Spark interaction
5. Q &A

Azure data platform overview

"The Data Driven Enterprise"

Engager les clients

Transformer les produits



Optimiser les opérations

Responsabiliser les personnes

Les barrières d'une bonne stratégie de données

Les promesses d'une approche de données moderne

Data-driven business insights

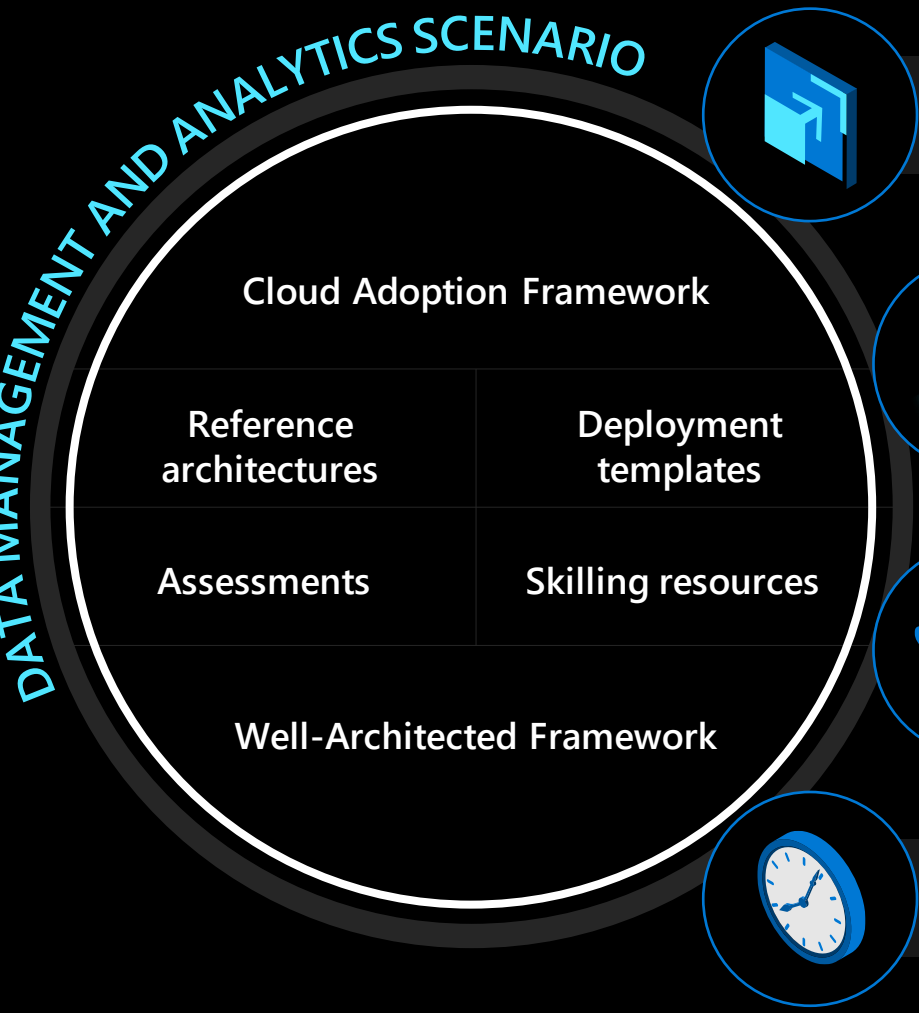
Holistic data management and analytics

Robust governance that enables self-service and flexibility

Secure, compliant, and fully integrated

Well-architected, repeatable, modular patterns

Donner aux métiers les clés de leurs données



Croitre sans augmenter la complexité

Architecture distribuée sous gouvernance de bout en bout

Agnostique en matière de cloud pour prend en charge les approches hybrides et multiclouds

Réduire le temps de déploiement de plusieurs mois à quelques semaines

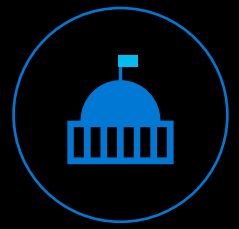
Les obstacles à la valorisation des données

Complexity

- Elaborate architectures that don’t scale.
- More teams engaging with data, not just engineering.

Uncertainty

- Constant evolution of tools and options.
- Lack of clear guidance on where to start and how to bring value.



Governance

- Prioritizing security prevents access to data and stifles innovation.
- Lack of consistent processes and policies create data silos.

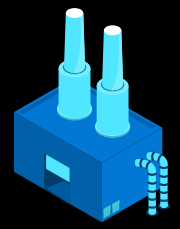
Skilling

- On-prem knowledge and skill sets don’t translate to cloud-based service.
- Need for skilling to adapt to new ways of working.

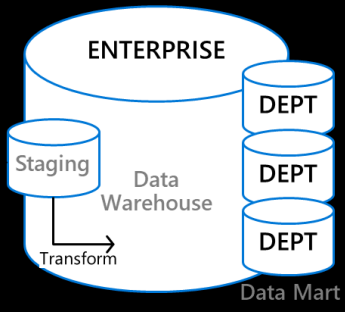
Time to value

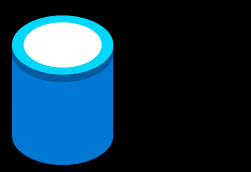
- New solutions can take months to deploy and achieve ROI.
- Inability to provide real-time, trusted data to inform timely business decisions.

Optimisation des plateformes de données

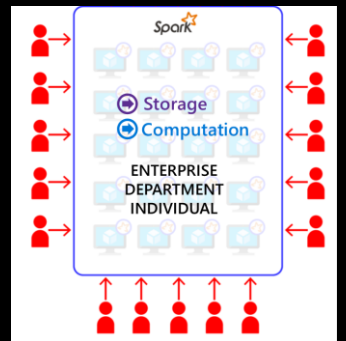


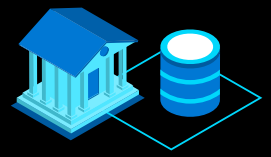
Data Warehouse



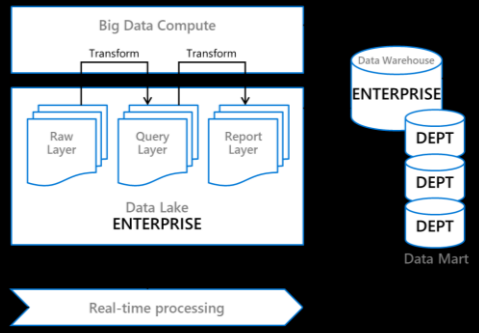


Data Lake

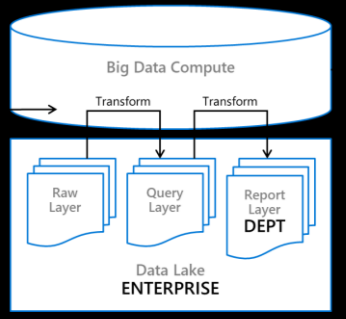




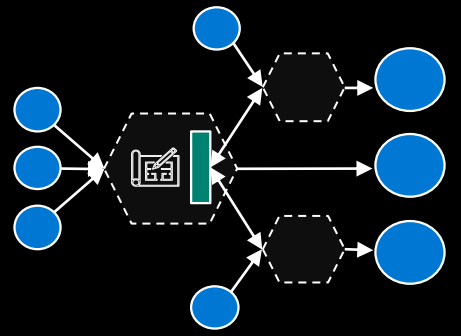
Cloud Data Platform



Data Lakehouse



Data Mesh

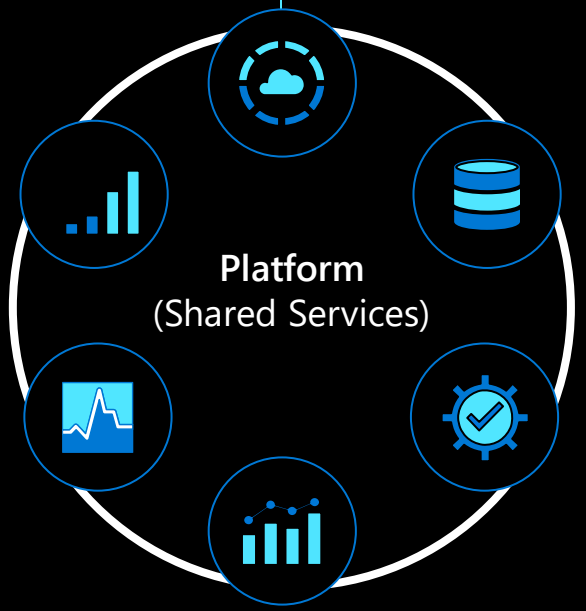


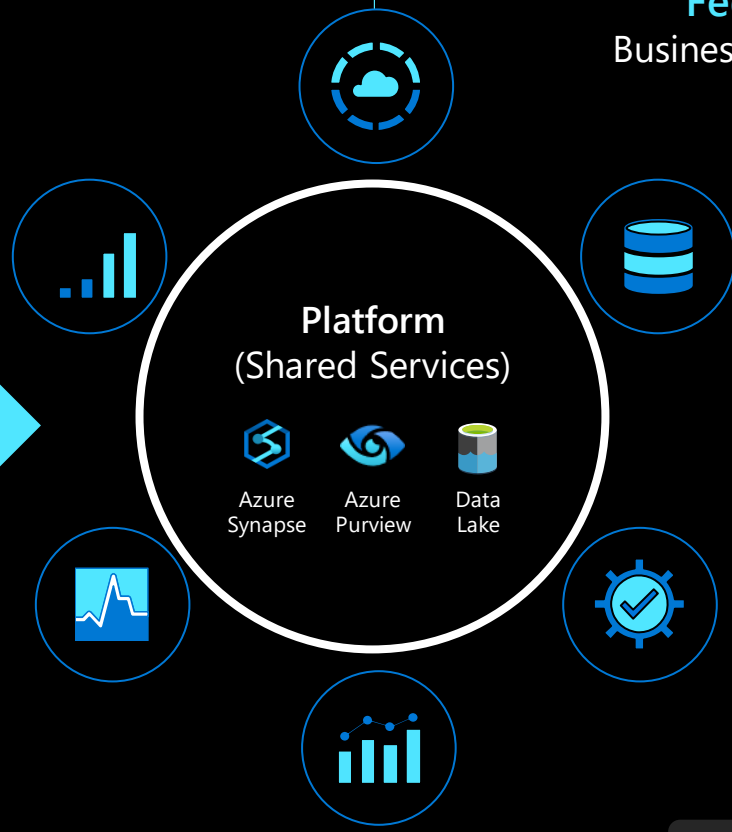
Donner le contrôle aux équipes business

Augmenter l'agilité d'accès aux données pour en tirer un maximum de valeur

Centralized data ownership
IT centrally owns data and all services

- Centralized architectures
- Competing business needs
- Workarounds to enable agility





Federated data ownership
Business owns data-driven projects

- Common governance and management
- Improves quality and security
- Increases efficiency and access

Ressource:
Build an initial strategy

Préparer l'entreprise à la gestion agile des données

L'agilité à grande échelle demande une approche holistique du management de la donnée

- Mettre en œuvre la **gouvernance de données** et **sécurité**.
- Considérer la **donnée comme un produit ("data product")** plutôt qu'un produit dérivé.
- Fournir un **écosystème** "data products" au lieu d'un seul entrepôt de données.
- Créer des domaines de données **"data domains"** pour adresser les secteurs d'activité.
- Donner aux équipes les moyens d'élaborer des solutions **analytiques** qui apportent de la valeur à l'entreprise.
- Moderniser vos **équipes** et vos **opérations**.

Ressource:
Prepare your environment

Data capabilities + Data culture = Unique potential

Les réalités d'aujourd'hui

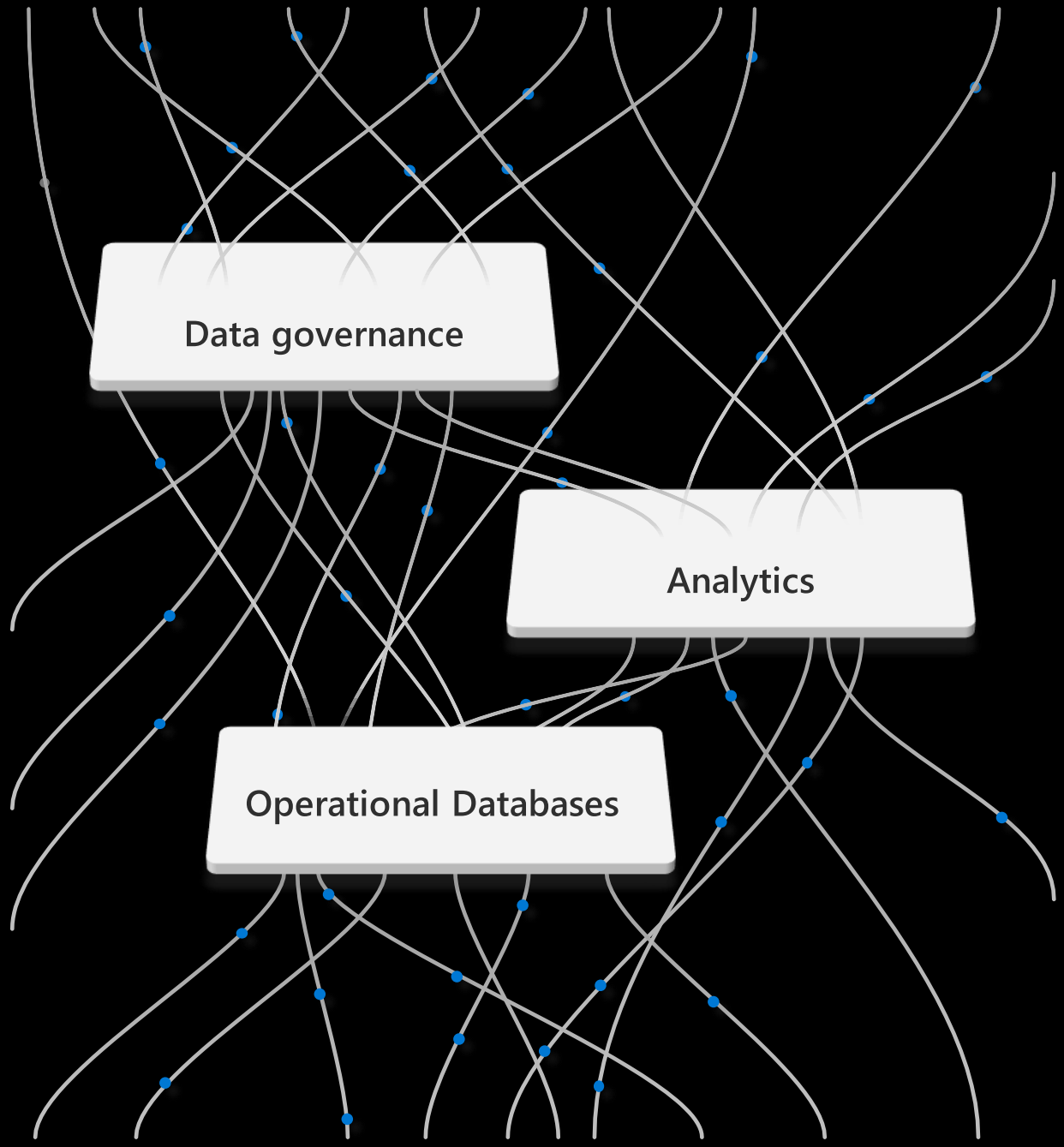
Quelles données ai-je à ma disposition?

Sont-elles dignes de confiance ?

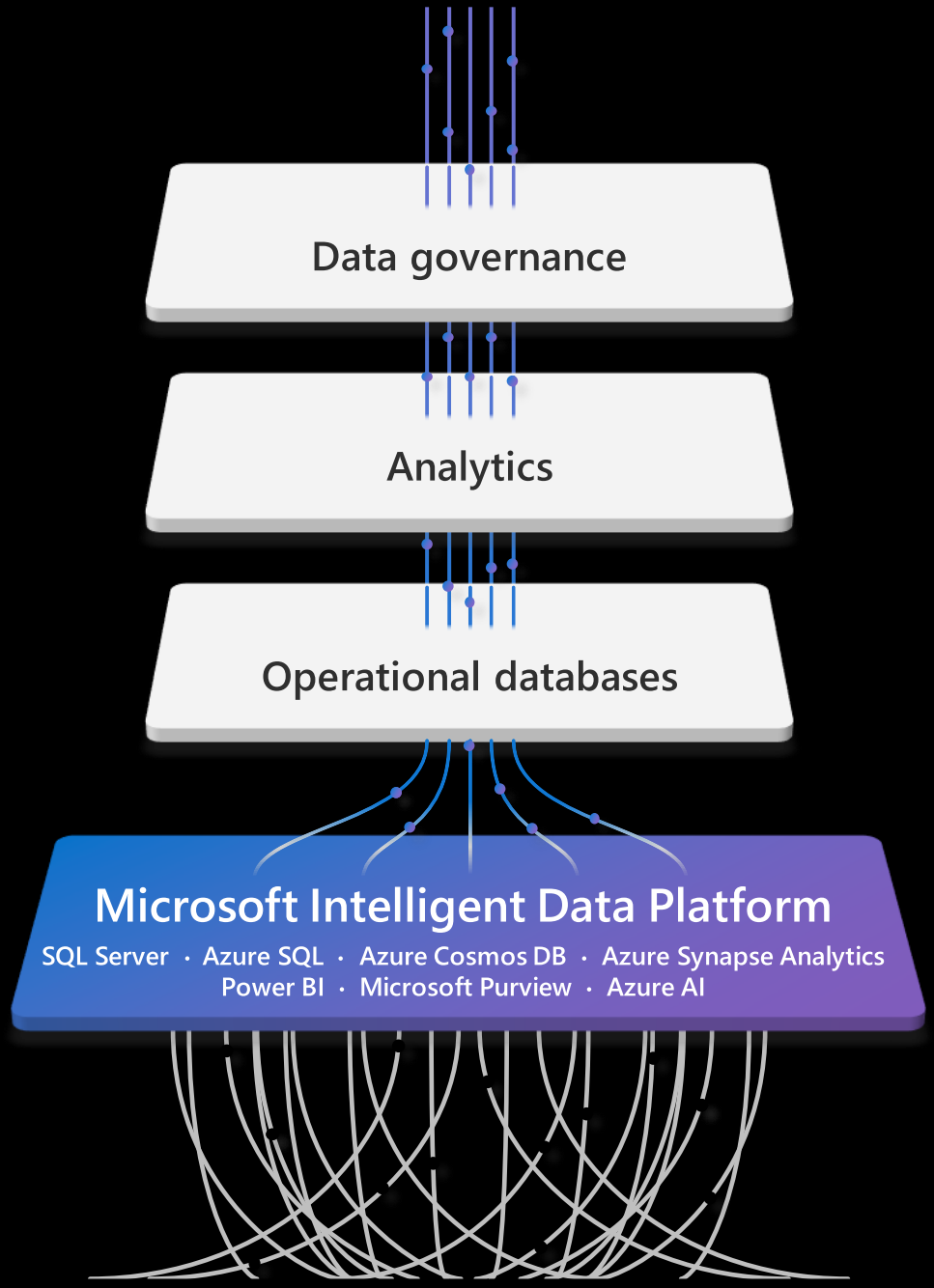
Les personnes peuvent-elles accéder aux données nécessaires pour prendre les bonnes décisions ?

Comment puis-je obtenir de plus rapides informations business des données ?

Quel est mon risque de non-conformité ?



Introduction à la "Microsoft Intelligent Data Platform"



Ce qu'offre la plateforme

Data modernization

Azure is the best destination for all your data

Cloud native applications

Ultra low latency at any scale with Azure Cosmos DB

Analytics and insights

Fastest time to insights

Data science

Responsible, powerful AI on your terms

Governance

AI powered discovery, catalogue, and protection

Common Data Model

Microsoft 365

Microsoft Power Platform

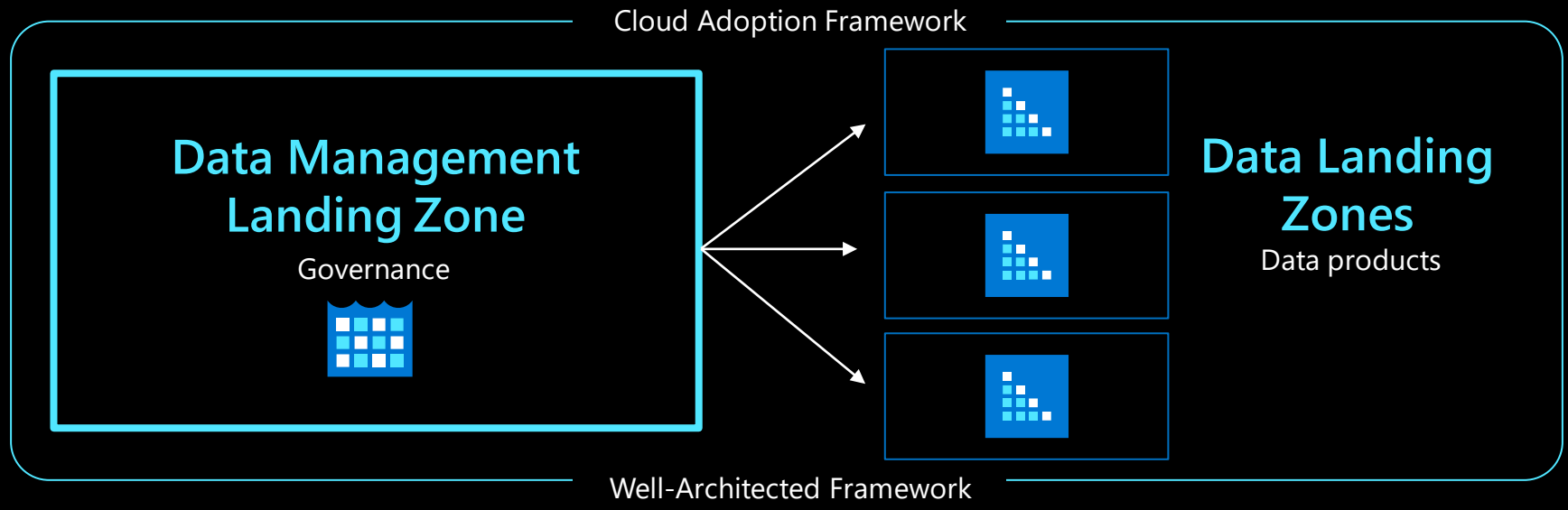
Microsoft Dynamics 365

Microsoft Dataverse

Azure Synapse Analytics & roadmap

Etablir et faire respecter la gouvernance

La "data management landing zone" permet de créer une base solide pour votre plateforme de données.

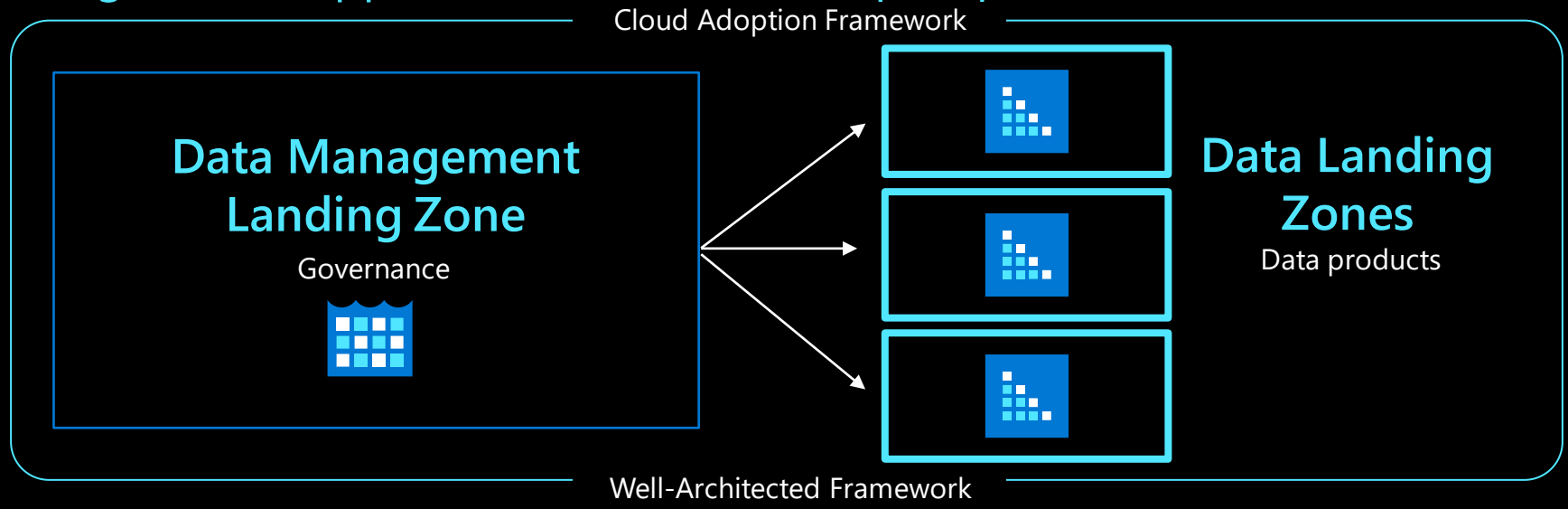


- Responsable de la gouvernance et de la gestion des données de bout en bout pour assurer la cohérence entre les "Data domains"
- Il facilite également la communication pour ingérer les données de l'ensemble de votre patrimoine numérique, y compris les infrastructures multi-cloud et hybrides.
- Héberge le catalogue de données, les systèmes d'audits, de monitoring et les services tiers pour l'automatisation des déploiements
- Définit et met en œuvre des politiques, notamment en matière de sécurité de base, de capacités et de normes.

Ressource: [Data management landing zone]

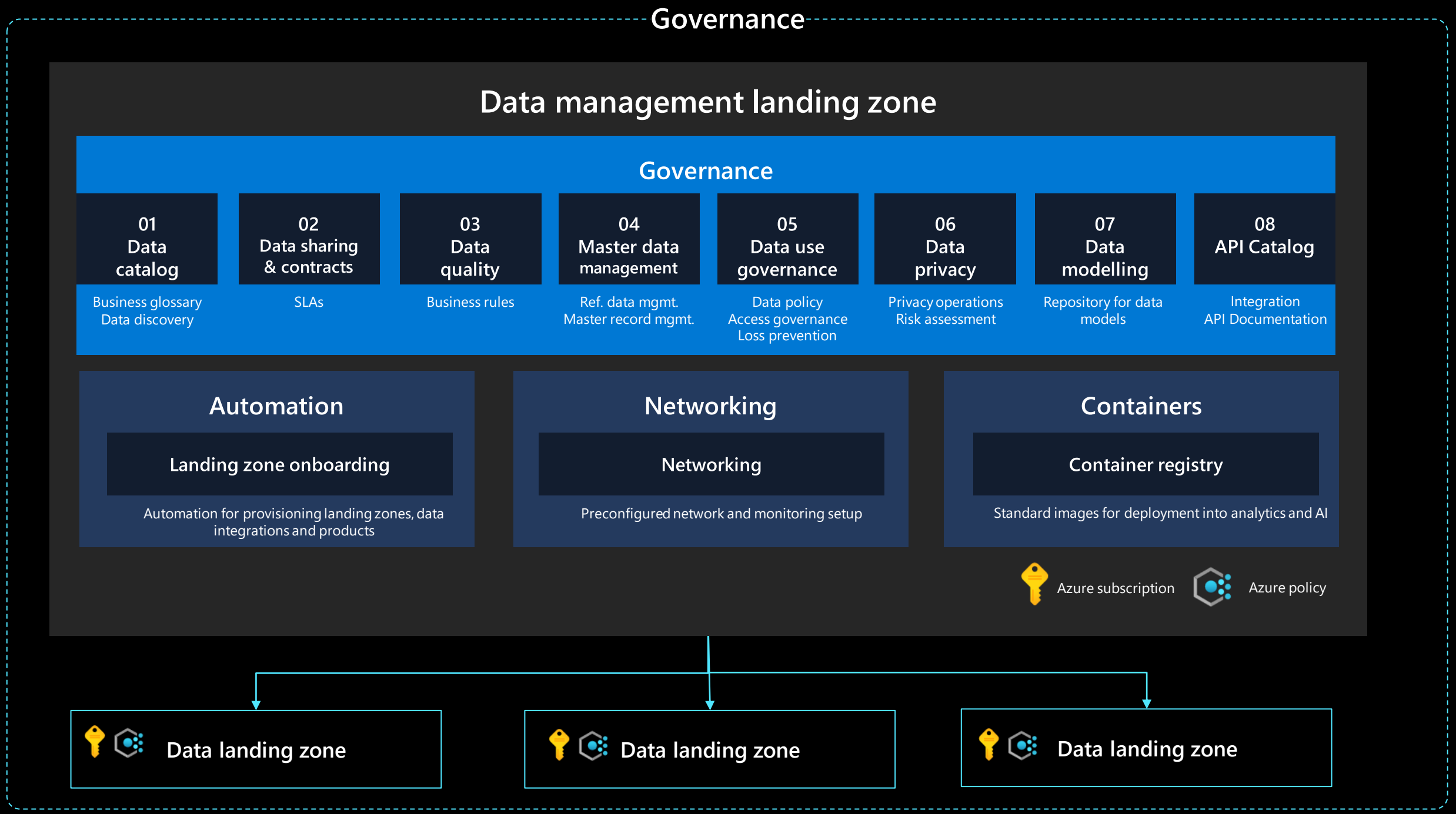
Moteurs d'intégrations de données et d'exposition des "Data products"

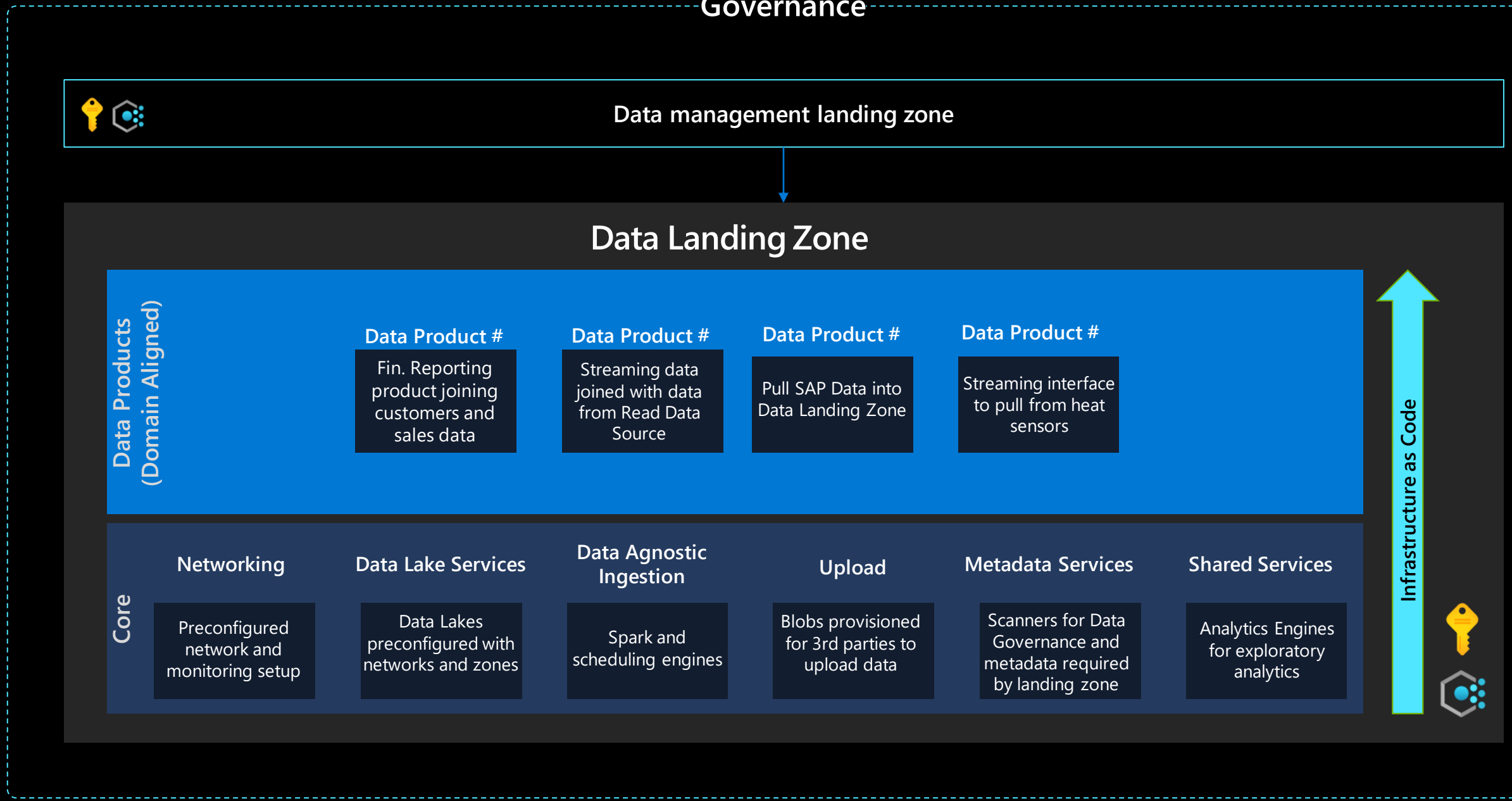
Les "Data landing zones" rapproches les données au plus prêt des métiers.



- Connectée à la "data management landing zone" pour la gouvernance et la compliance
- Construit sur des standards d'infrastructure, tels que les réseaux, le monitoring, les services d'ingestion de données et moteurs de calculs
- Défini par domaine pour faciliter l'agilité
- Chacune des "landing zone" peut héberger plusieurs "Data products", services d'intégration ou analytiques (IA/ML)

Ressource: Overview of the data landing zone


Governance
Data management landing zone
Governance
01 Data catalog
Business glossary
Data discovery
02 Data sharing & contracts
SLAs
03 Data quality
Business rules
04 Master data management
Ref. data mgmt.
Master record mgmt.
05 Data use governance
Data policy
Access governance
Loss prevention
06 Data privacy
Privacy operations
Risk assessment
07 Data modelling
Repository for data models
08 API Catalog
Integration
API Documentation
Automation
Landing zone onboarding
Automation for provisioning landing zones, data integrations and products
Networking
Networking
Preconfigured network and monitoring setup
Containers
Container registry
Standard images for deployment into analytics and AI
Azure subscription
Azure policy
Data landing zone
Data landing zone
Data landing zone

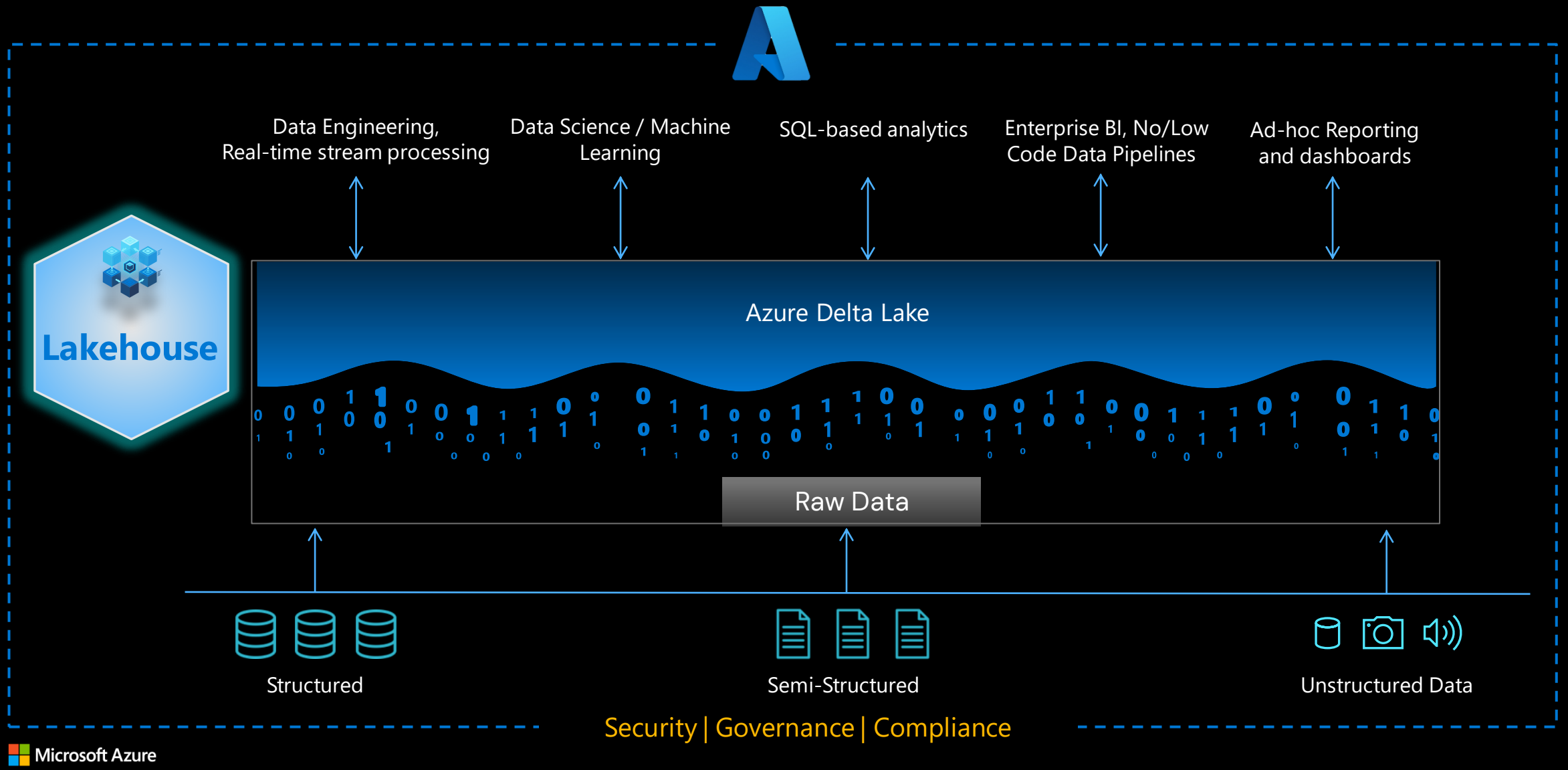

Governance
Data management landing zone
Data Landing Zone
Data Products (Domain Aligned)
Data Product #
Fin. Reporting product joining customers and sales data
Data Product #
Streaming data joined with data from Read Data Source
Data Product #
Pull SAP Data into Data Landing Zone
Data Product #
Streaming interface to pull from heat sensors
Core
Networking
Preconfigured network and monitoring setup
Data Lake Services
Data Lakes preconfigured with networks and zones
Data Agnostic Ingestion
Spark and scheduling engines
Upload
Blobs provisioned for 3rd parties to upload data
Metadata Services
Scanners for Data Governance and metadata required by landing zone
Shared Services
Analytics Engines for exploratory analytics
Infrastructure as Code

Approche moderne de données

Besoin d'expliquer finement l'approche analytique moderne
Besoin d'expliquer les nouvelles avancées des services de données dans le cloud

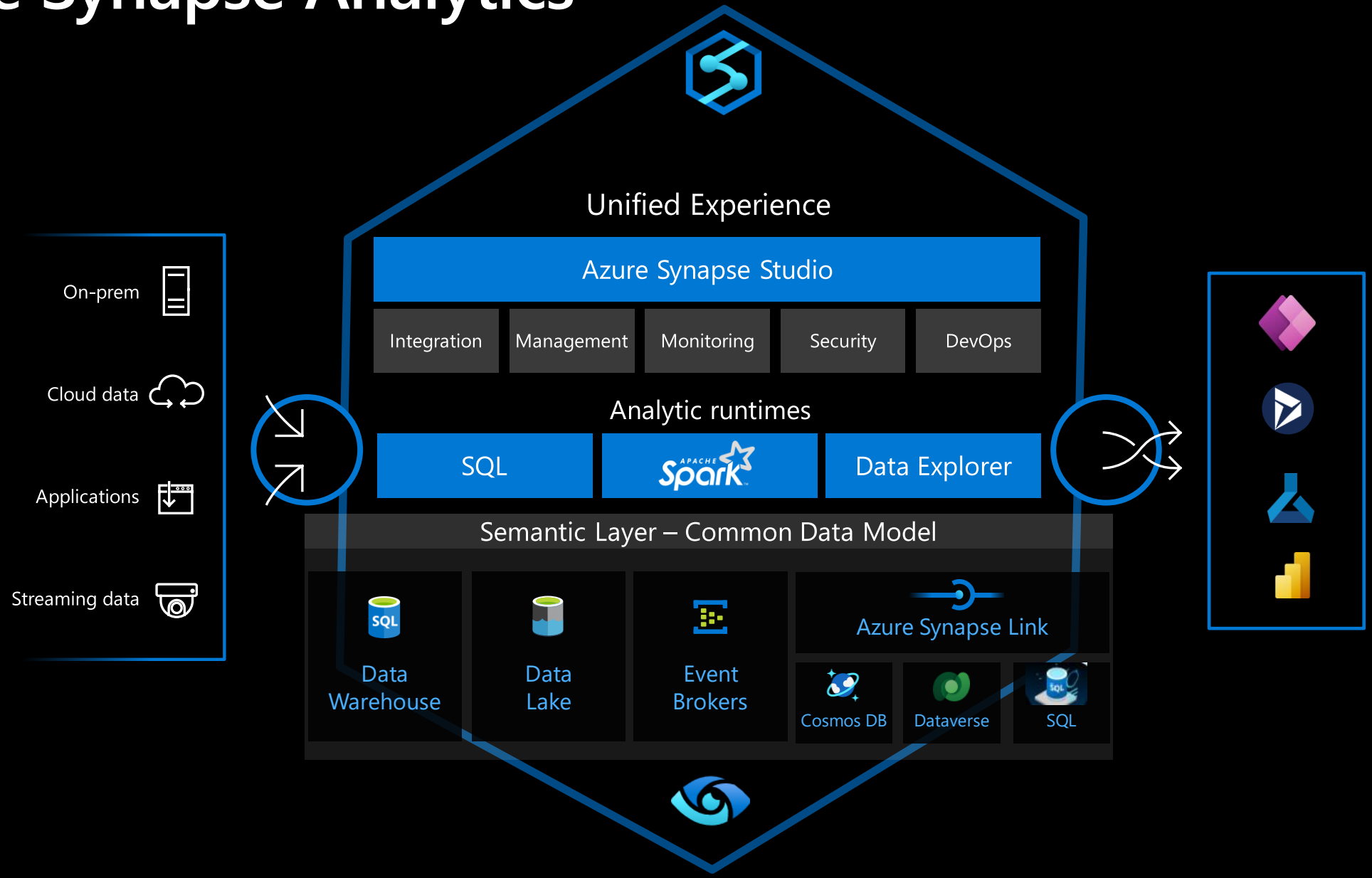
- Détachement de la scalabilité des services, plus de "capacity planning"
- Détachement de la notion de performance
- Consommation "à la demande" de "l'analytique" moderne
- Les limites d'une telle approche
- **Réflexion sur "où" est calculé l'indicateur, modèles hybrides/composites**
- Couts variables pour les métiers (dépendant de l'utilisation), besoin de changer l’approche budgétaire.

L'approche Lakehouse, multi moteur analytique



Microsoft Azure

Azure Synapse Analytics

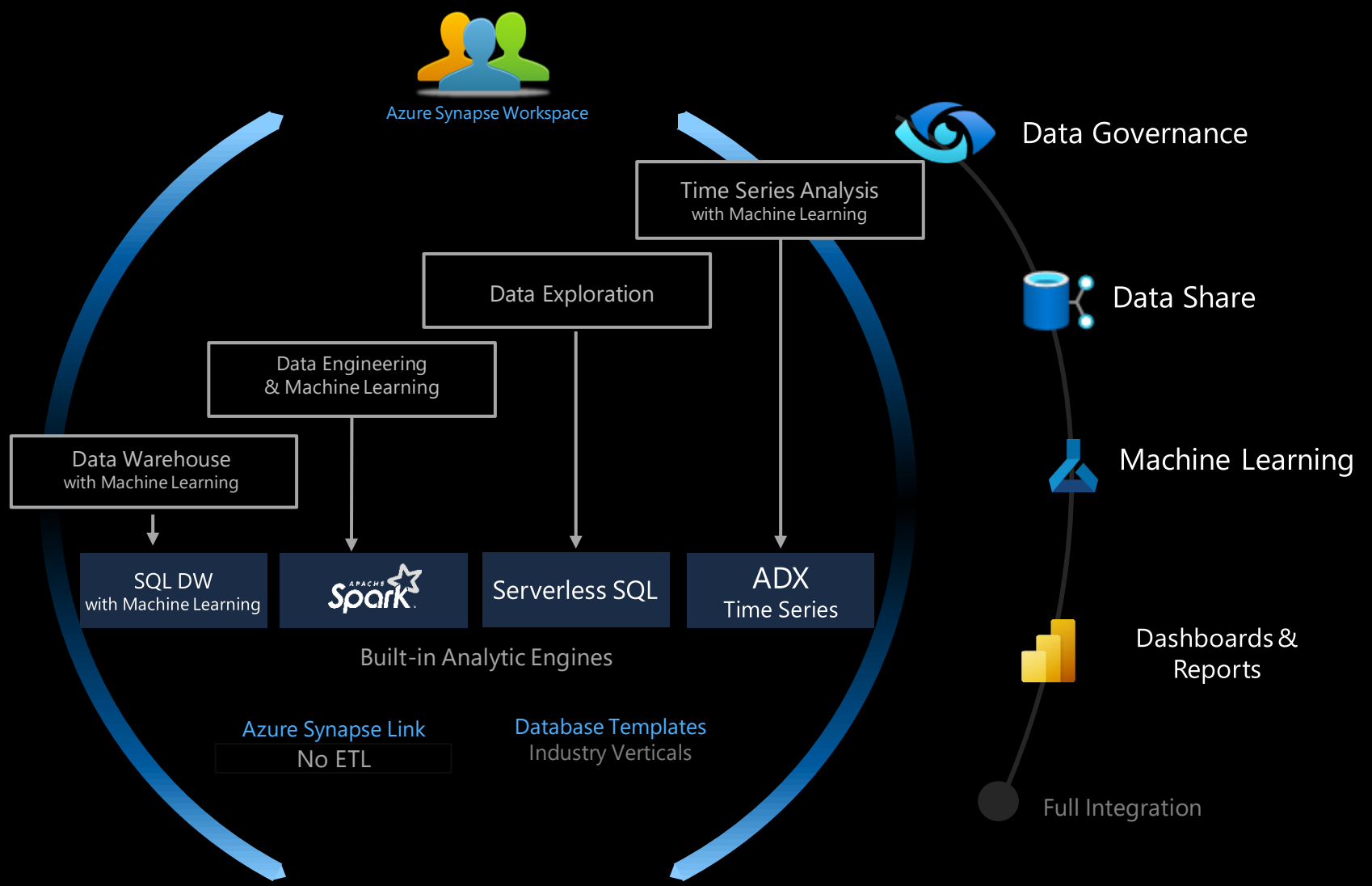


Azure Synapse Analytics

Unified, Secured, and Integrated Ecosystem

Azure Synapse Analytics

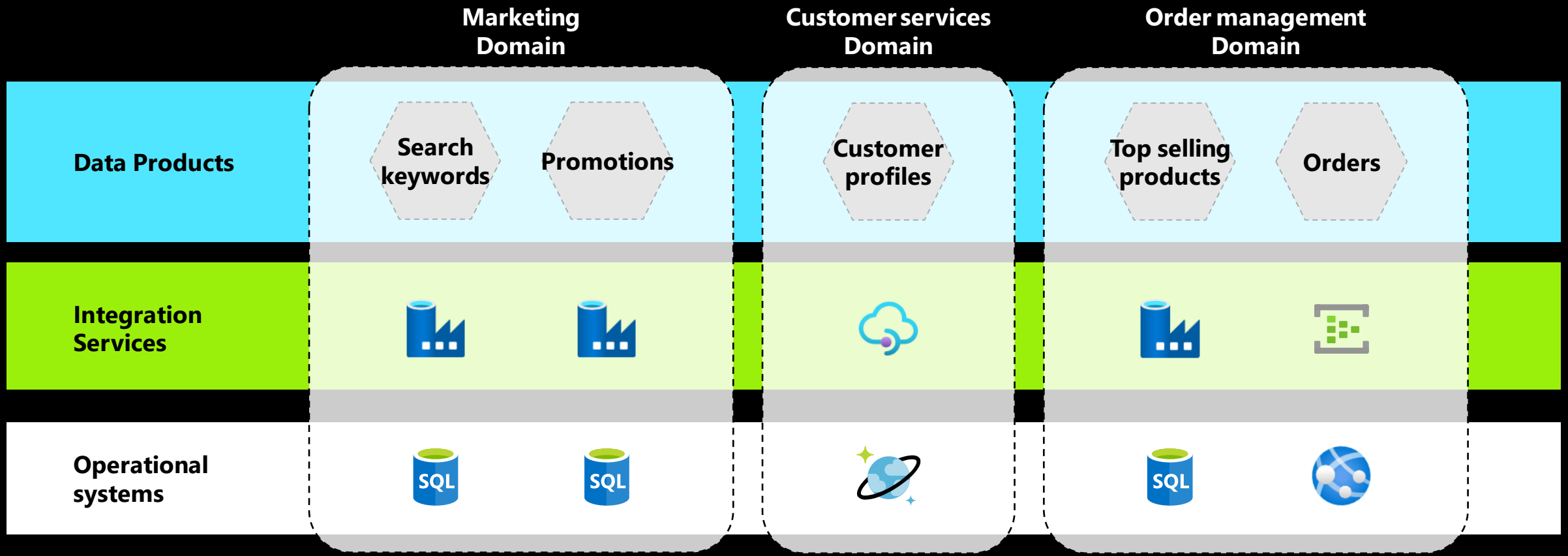
Unified, Secured, and Integrated Ecosystem



Polyglot Data Store: On-premise Cloud data Operational Data Streaming data

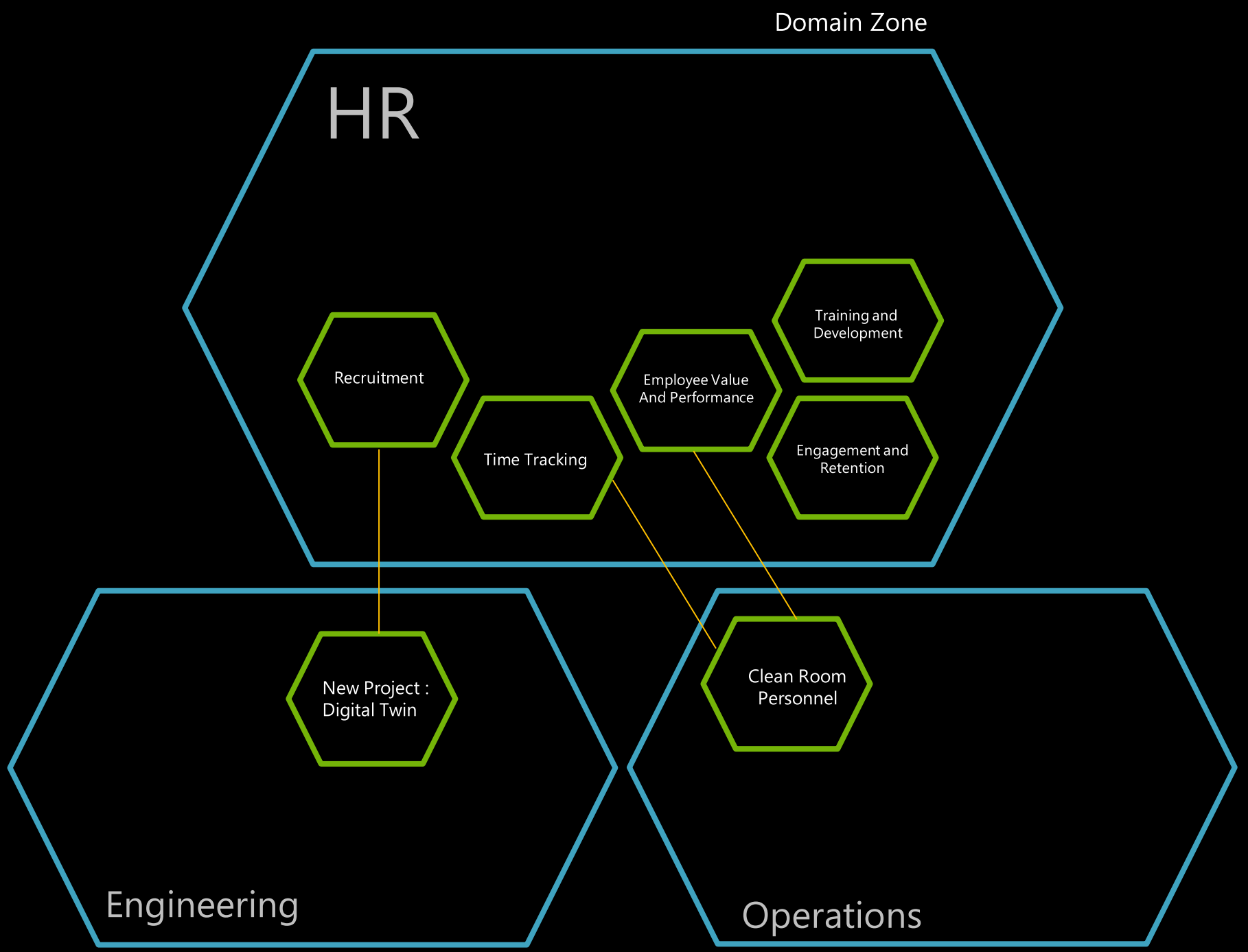
Qu'est ce qu'un "Data domain" ?

- Un domaine est simplement un ensemble de personnes généralement organisées autour d'un objectif business commun.
- Ce domaine va créer des "Data product" pour d'autres domaines ou utilisateurs, indépendamment d'autres contraintes / d'autres domaines.
- Le domaine doit s'assurer que les données sont accessibles, utilisables, disponibles et respectent des critères de qualité prédéfinis.
- Le domaine est responsable de l'évolution des "Data product" dépendant des retours des utilisateurs. Il est aussi de sa responsabilité de gérer le cycle de vie d'un "Data product".



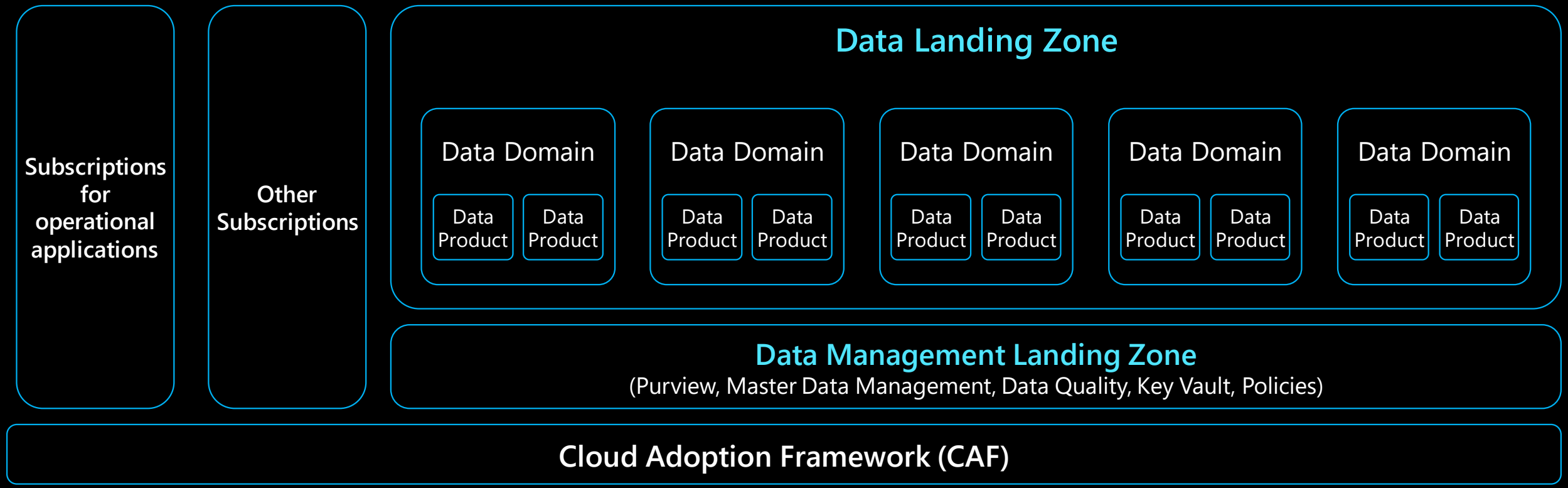
Domain Zones

Data Products



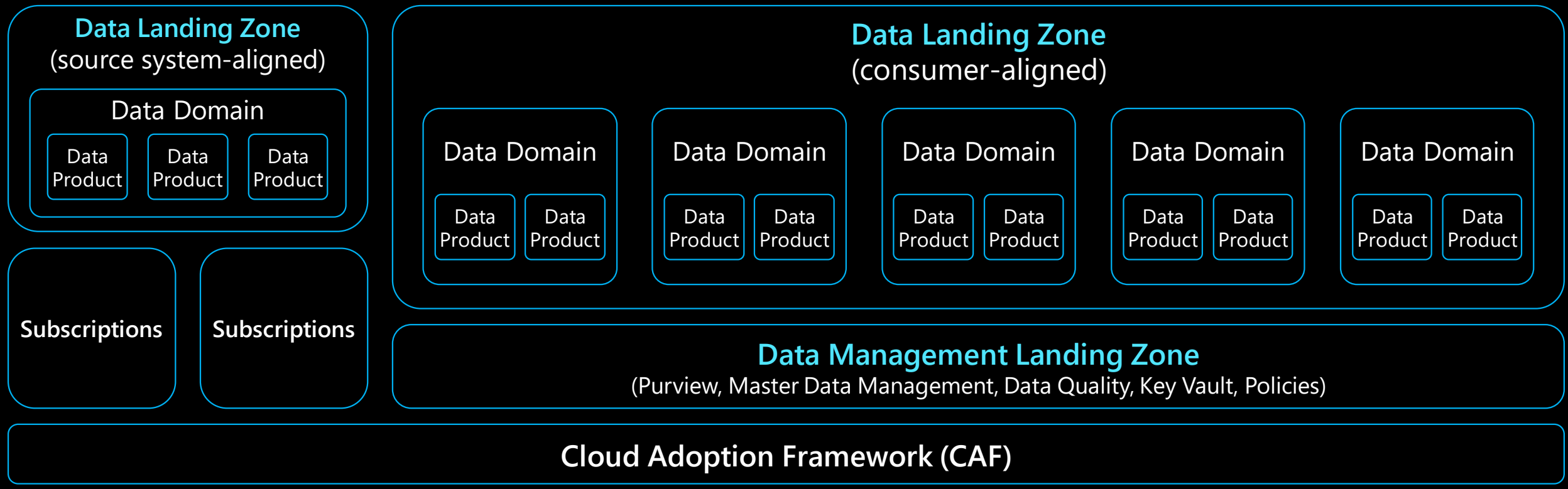
Data Mesh Pattern I

Single landing zone



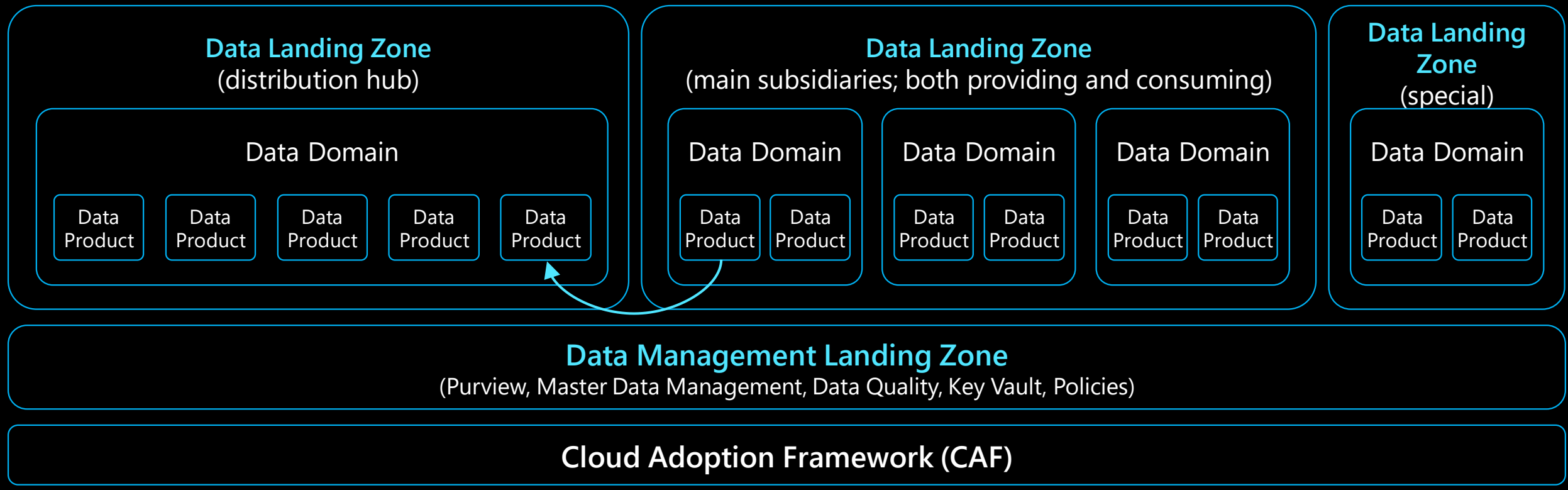
Data Mesh Pattern II

Source system- and consumer-aligned landing zones



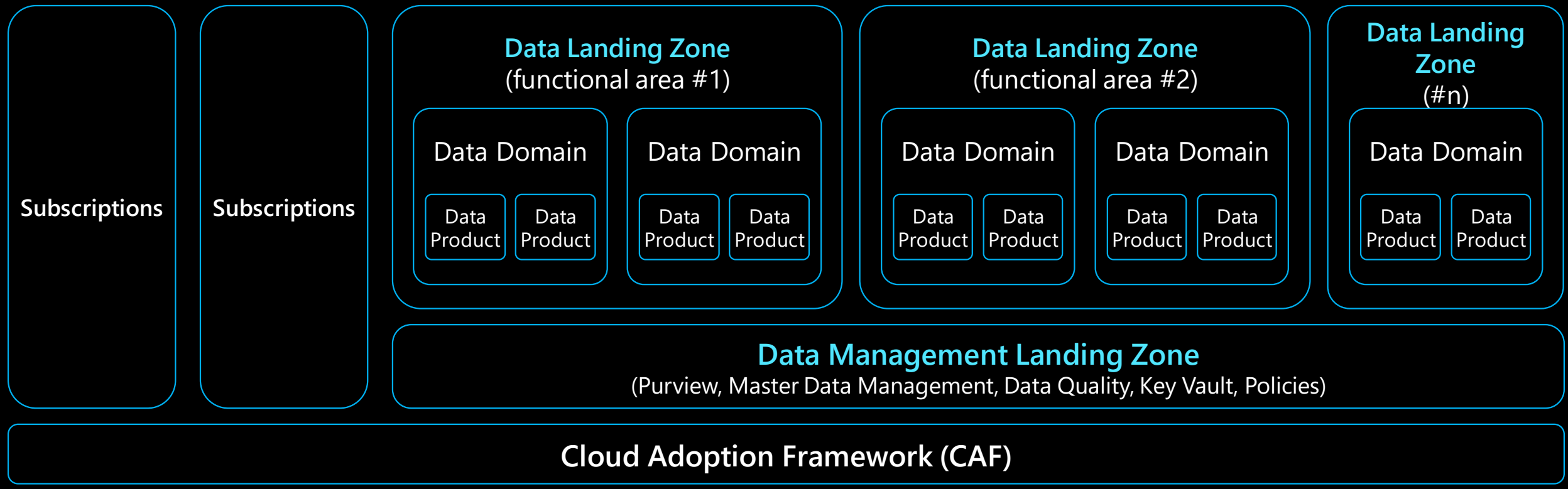
Data Mesh Pattern III

Hub-, generic- and special data landing zones



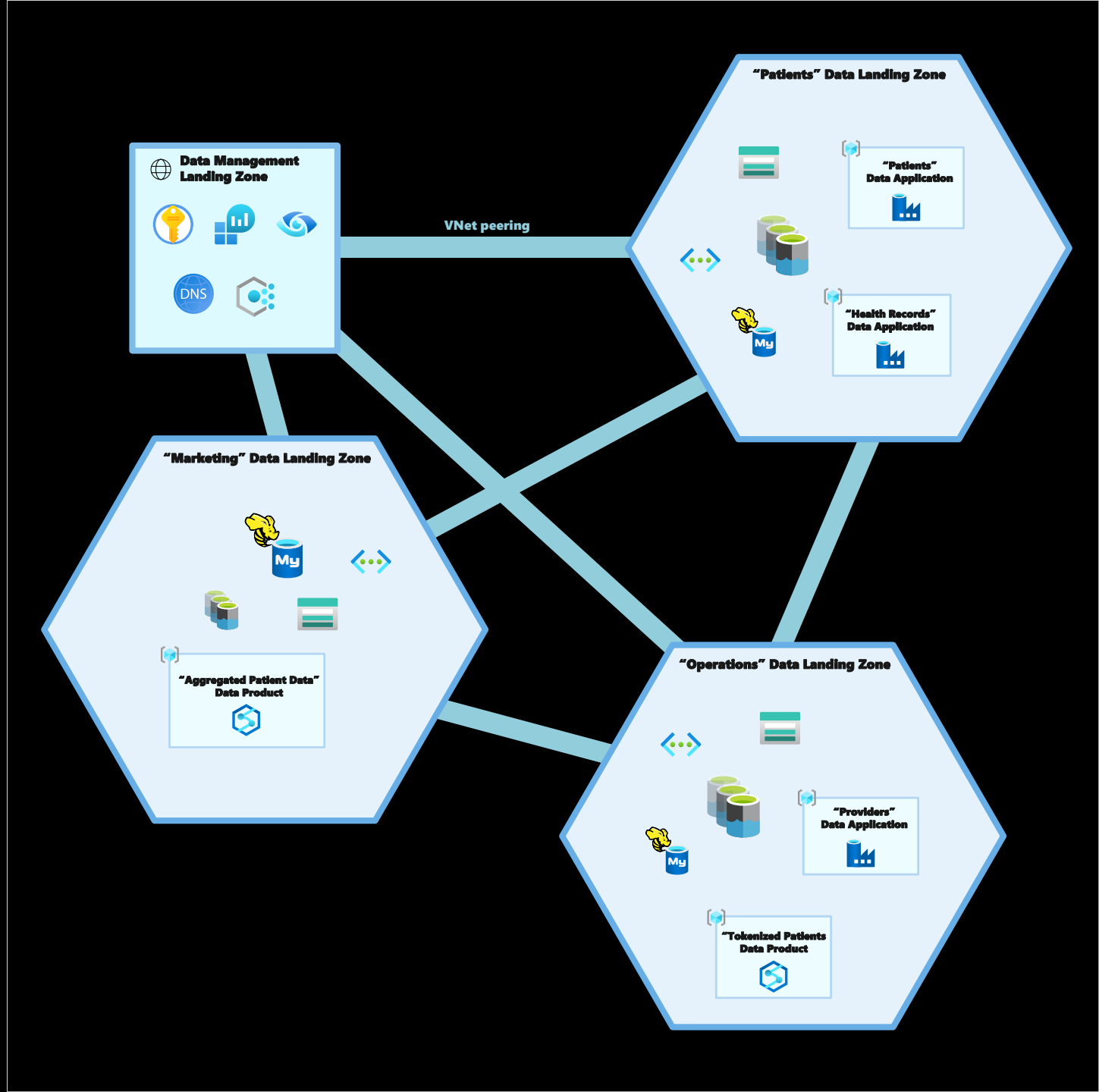
Data Mesh Pattern IV

Functional and regionally aligned data landing zones



Exemple de domaines

(Pattern IV)



Spark in Azure Synapse

Data Engineering with Spark

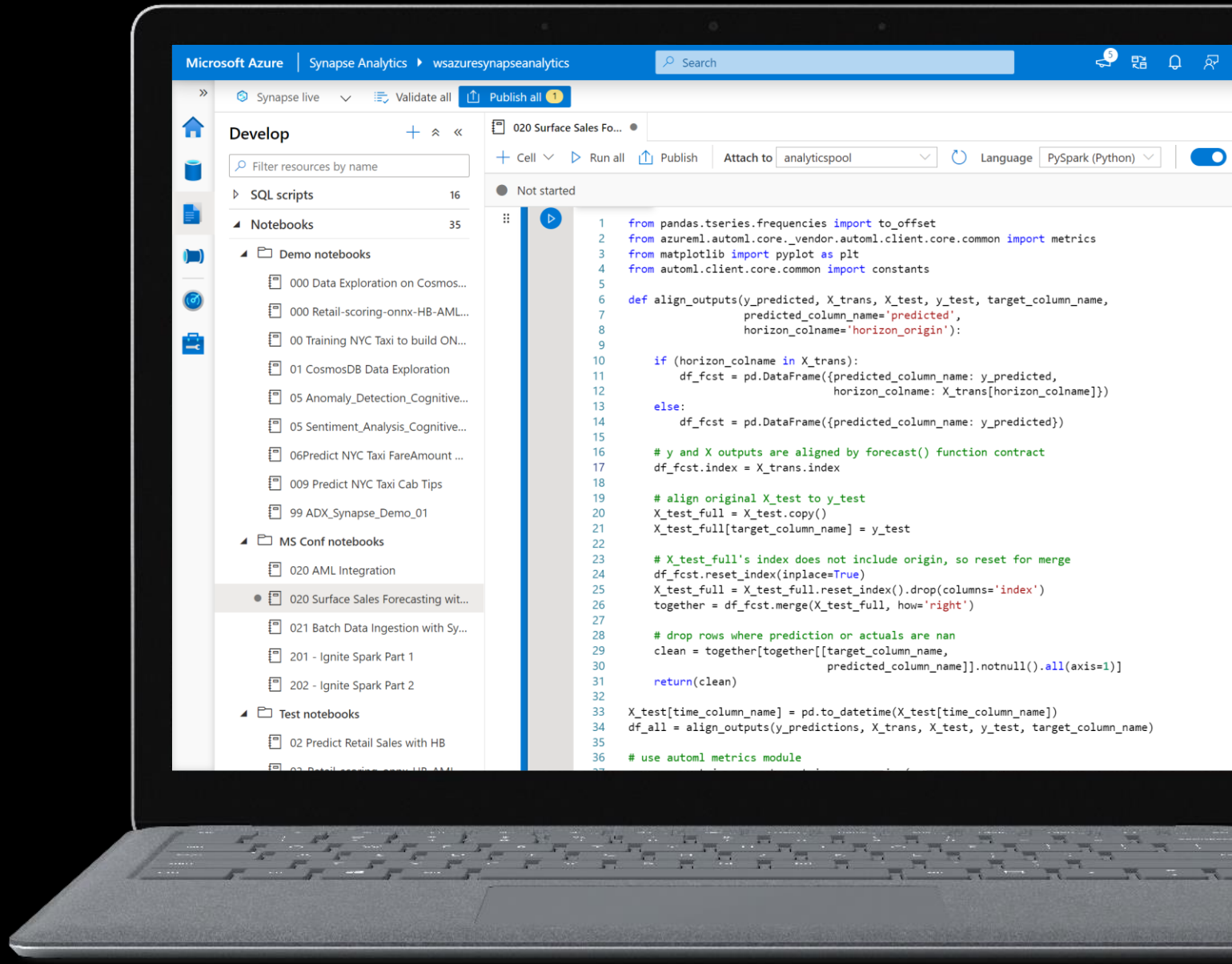
Code-first Data Engineering

PySpark, Scala, SQL and C# languages supported

Author multiple languages in a single notebook

Analyze & transform data from the data warehouse, data lake, and real-time operational data from one place

Synapse Spark uses Spark 3.2 runtime, which includes Delta Lake 1.0



What is Apache Spark?

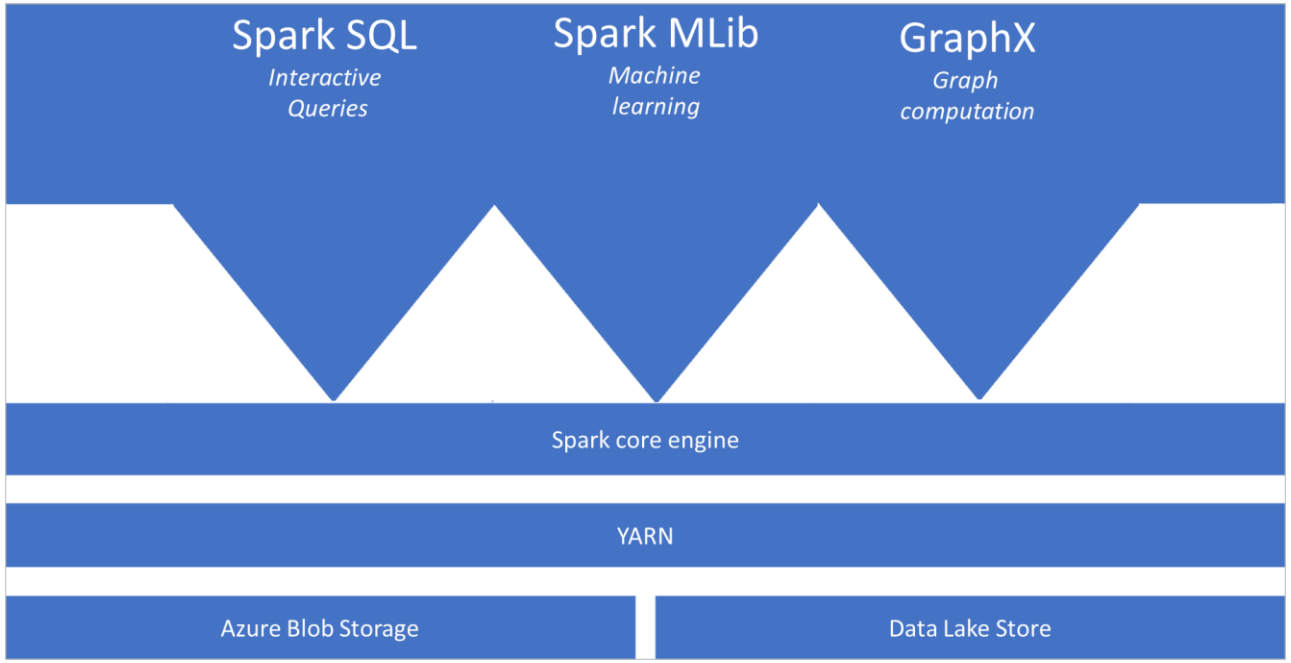
Parallel processing framework

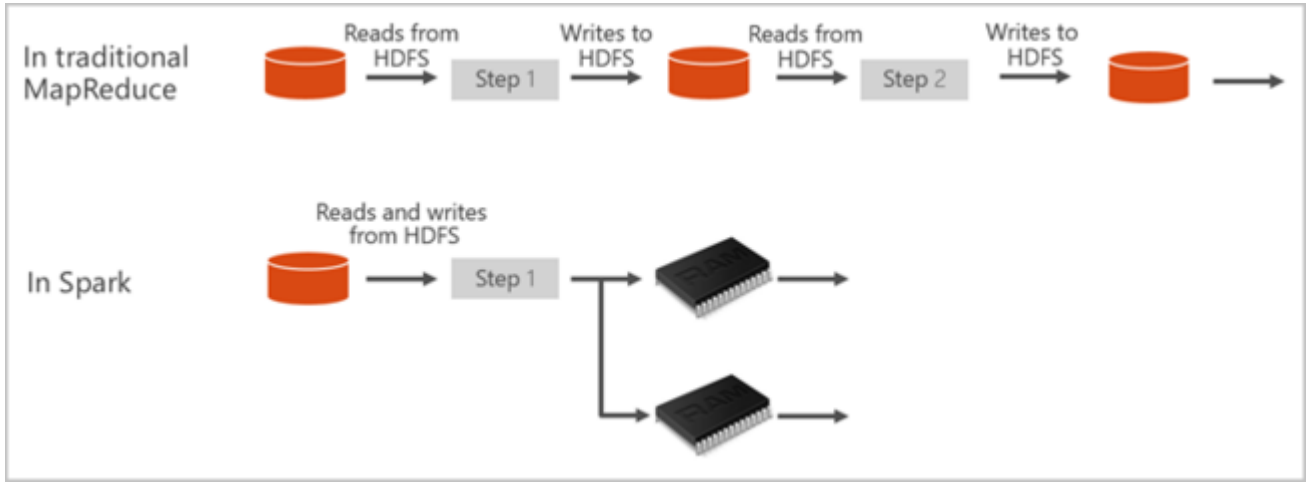
In-memory processing engine to boost the performance of big-data analytic applications

Much faster than disk-based applications

Integrates with multiple programming languages

Supports many workloads: Data Engineering, SQL, ML, Graphs., etc.





General Availability

Q2 2022

Spark 3.2

Enables developers can leverage the latest innovations in the Spark ecosystem

Pandas (Koalas) integration

A highly popular and flexible library with broad industry adoption

Adaptive Query Execution (AQE) enabled by default

Significant improvements in query performance out-of-the-box

Small Query execution improvements

Small queries run faster due to reduced initialization overhead

Motivation for Spark Pool in Synapse

Speed and efficiency	• Spark instances start in approximately 2 minutes for fewer than 60 nodes and approximately 5 minutes for more than 60 nodes. • The instance shuts down, by default, 5 minutes after the last job executed unless it is kept alive by a notebook connection.
Ease of creation	Quickly create a new Spark pool in Azure Synapse using: • Azure portal • Azure PowerShell • Synapse Analytics .NET SDK.
Ease of use	Synapse Analytics includes a custom notebook derived from Nteract. You can use these notebooks for interactive data processing and visualization.

Motivation for Spark Pool in Synapse

REST APIs	Spark in Azure Synapse Analytics includes Apache Livy, a REST API-based Spark job server to remotely submit and monitor jobs.
Support for Azure Data Lake Storage Generation 2	Spark pools in Azure Synapse can use: • Azure Data Lake Storage Generation 2 • BLOB storage.
Integration with third-party IDEs	Azure Synapse provides an IDE plugin for JetBrains' IntelliJ IDEA that is useful to create and submit applications to a Spark pool.

Motivation for Spark Pool in Synapse

Pre-loaded Anaconda libraries	• Spark pools in Azure Synapse come with Anaconda libraries pre-installed. • Providing close to 200 libraries for machine learning, data analysis, visualization, etc.
Scalability	• Can have Auto-Scale enabled, so that pools scale by adding or removing nodes as needed. • Spark pools can be shut down with no loss of data since all the data is stored in Azure Storage or Data Lake Storage.

Spark Pool - Configuration

- **Nodes**
  - Apache Spark pool instance consists of one head node and two or more worker nodes with a minimum of three nodes in a Spark instance.
  - The head node runs additional management services such as Livy, Yarn Resource Manager, Zookeeper, and the Spark driver.
  - All nodes run services such as Node Agent and Yarn Node Manager.
  - All worker nodes run the Spark Executor service.

Spark Pool - Configuration

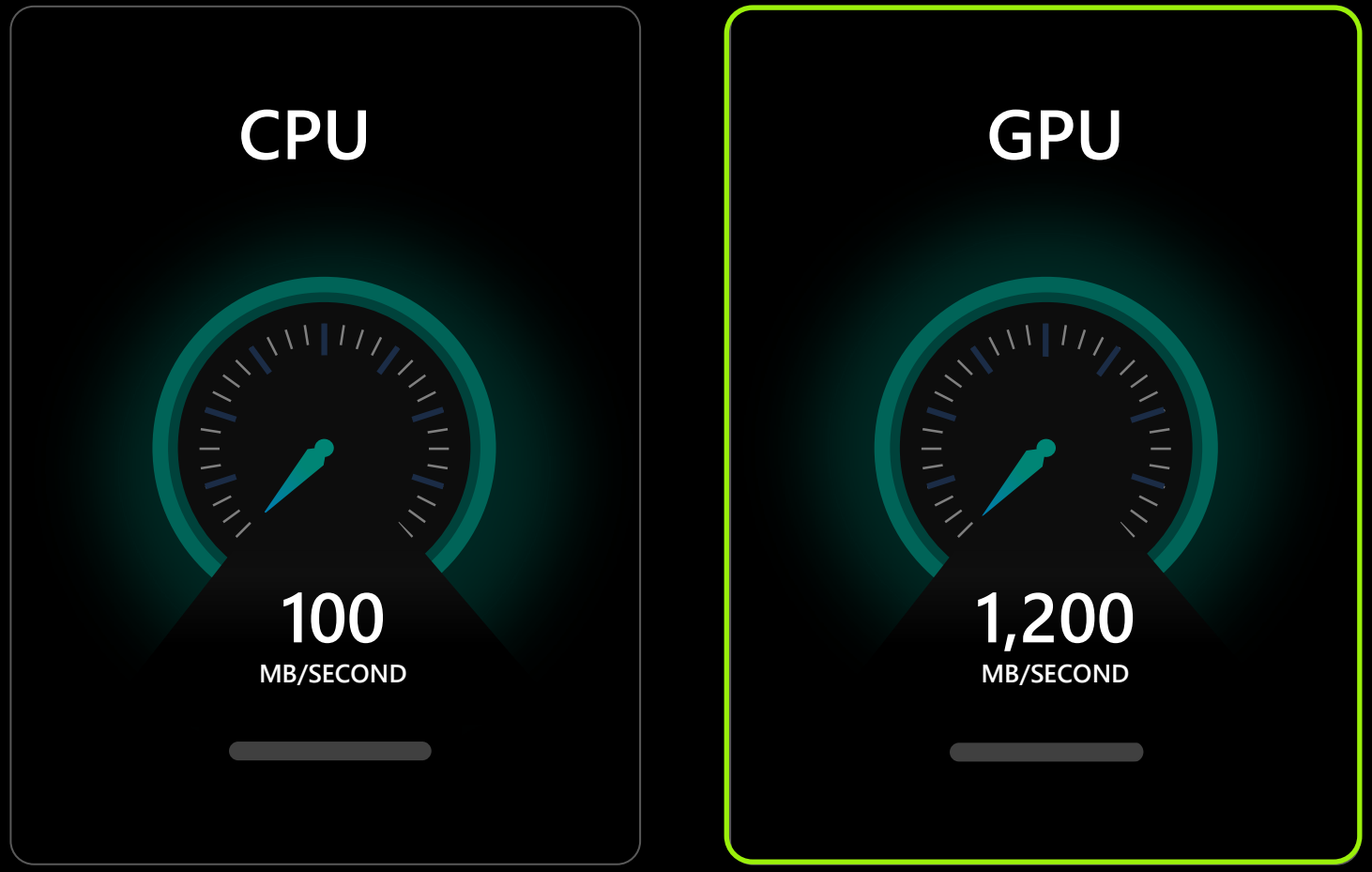
- **Node Sizes**
  - Node sizes can be altered after pool creation, instance may need to be restarted.

Size	vCore	Memory
Small	4	32 GB
Medium	8	64 GB
Large	16	128 GB
XLarge	32	256 GB
XXLarge	64	432 GB
XXX Large (Isolated Compute)	80	504 GB

Public Preview

GPU Accelerated Workloads

Accelerates data transformation and reduces ML model training time by dramatically increasing throughput vs. traditional CPU



Pools accélérés par GPU - Azure Synapse Analytics | Microsoft Learn

Generally Available

https://aka.ms/spark

SynapseML

A machine learning library that's

Simple

Quickly create, train, and use distributed machine learning tools in only a few lines of code.

Scalable

Scale ML workloads to hundreds of machines on your Apache Spark cluster.

Multilingual

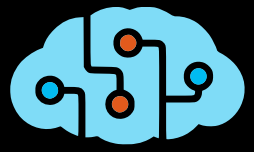
Use SynapseML from any Spark compatible language including Python, Scala, R, Java, .NET and C#.

Open

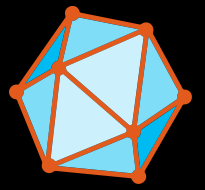
SynapseML is Open Source and can be installed and used on any Spark 3 infrastructure including your local machine, Databricks, Synapse Analytics, and others.

Synapse ML = Apache Spark ML +

Gradient Boosting



Cognitive Services



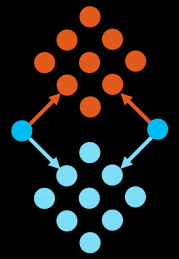
Deep learning

Reinforcement learning

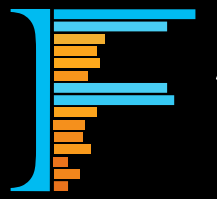
Responsible AI

Language Modeling

Search engine creation



Content retrieval



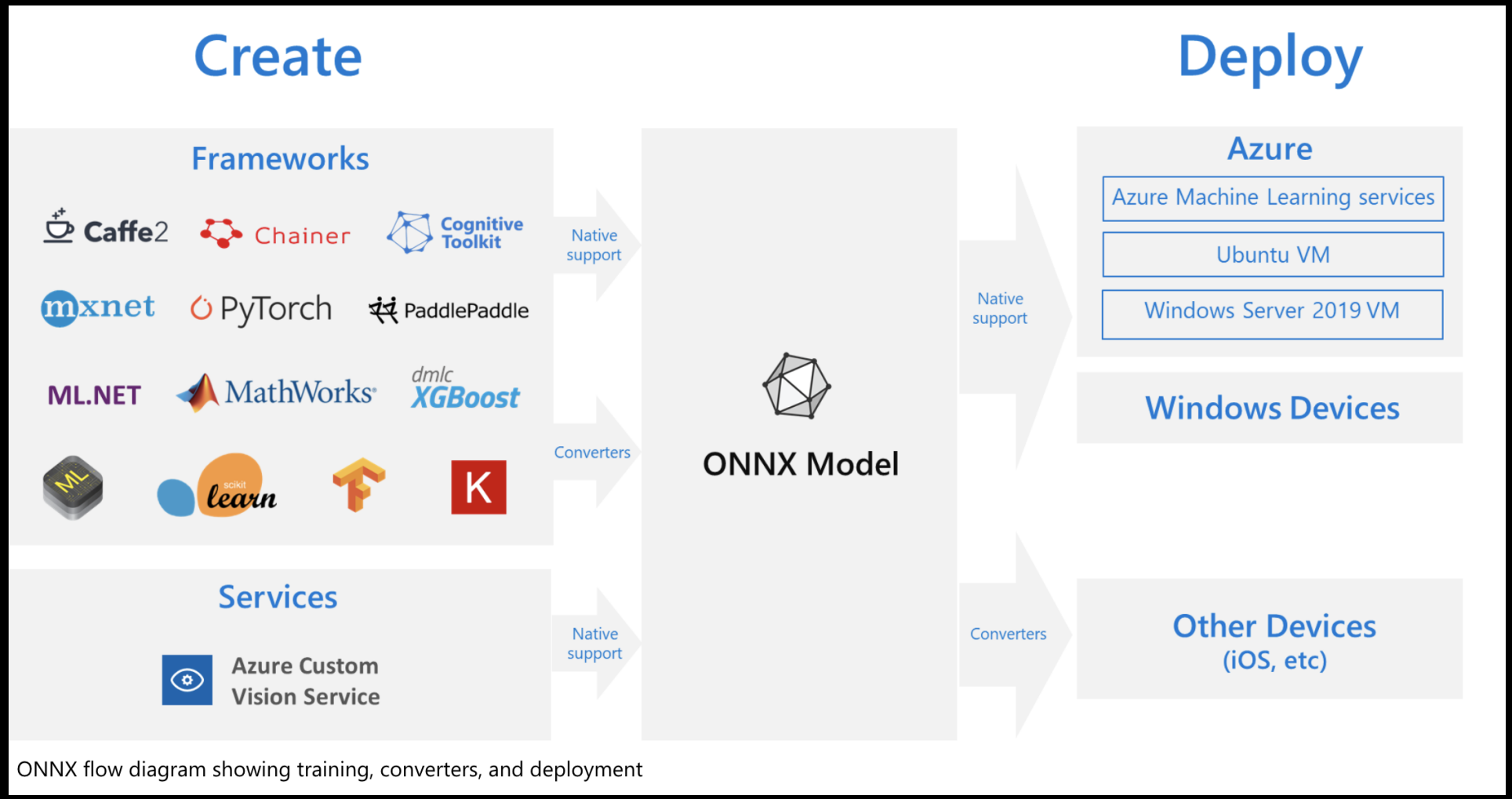
Anomaly detection

Cybersecurity

Explainable Models

Image processing

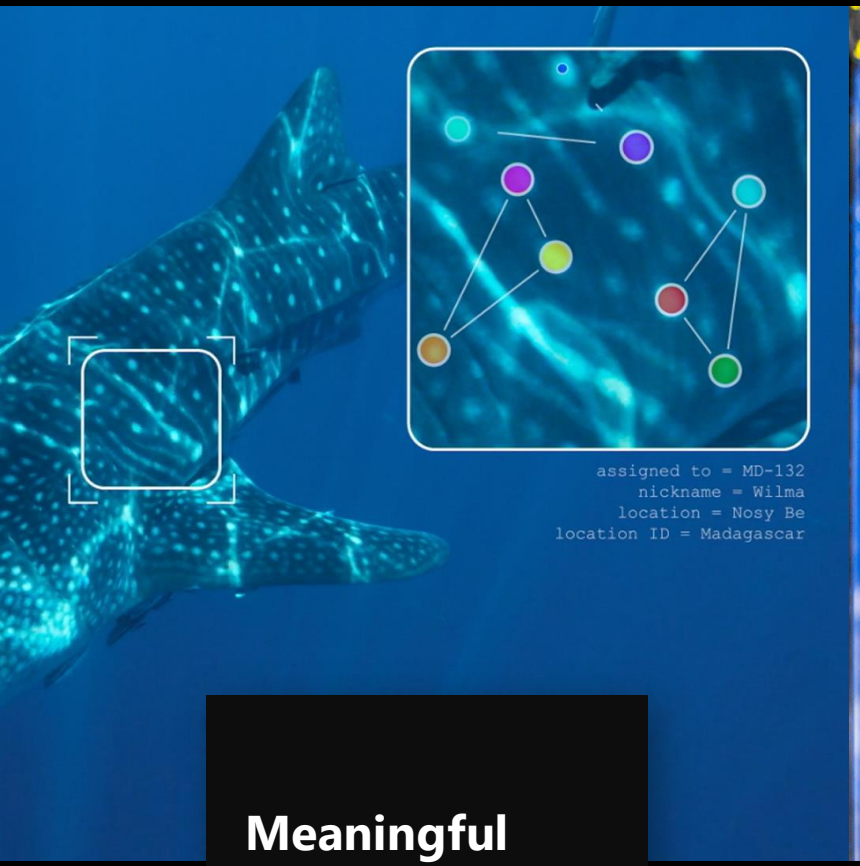
What is ONNX?

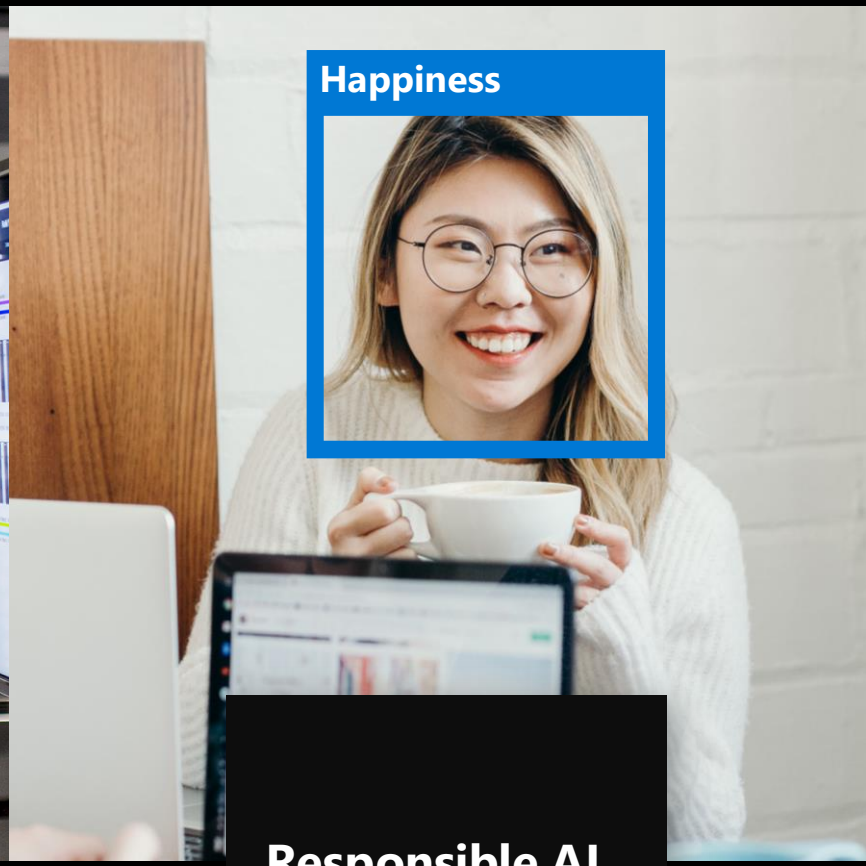


ONNX flow diagram showing training, converters, and deployment

Azure ML overview & Spark interaction

Microsoft's AI Approach





Meaningful Innovation

Empowering People

Responsible AI

Democratizing AI for every employee

More Personas Building AI Applications

Azure Machine Learning

Azure Machine Learning

Seamless studio experience

Notebooks | Automated ML | Designer

Responsible ML tools

Comprehensive MLOps

Reproducibility | Automation | Deployment | Re-training

Governance

Unified management across clouds and on-premises

Azure		Edge & Hybrid	
Serverless Compute	Managed Kubernetes	Azure Arc-enabled Kubernetes	Edge/IoT Devices

Structured Data

Unstructured Data

All other ML scenarios; NLP, Vision, IoT, etc.

Business Apps

Models: AKS
Batch predictions: SQL DB

Analytics & Govern

Prepare Data | Build & Train | Deploy | Manage & Monitor

Azure Machine Learning Service

Make data scientists to be more productive
Enable your organization to manage the ML lifecycle through MLOps

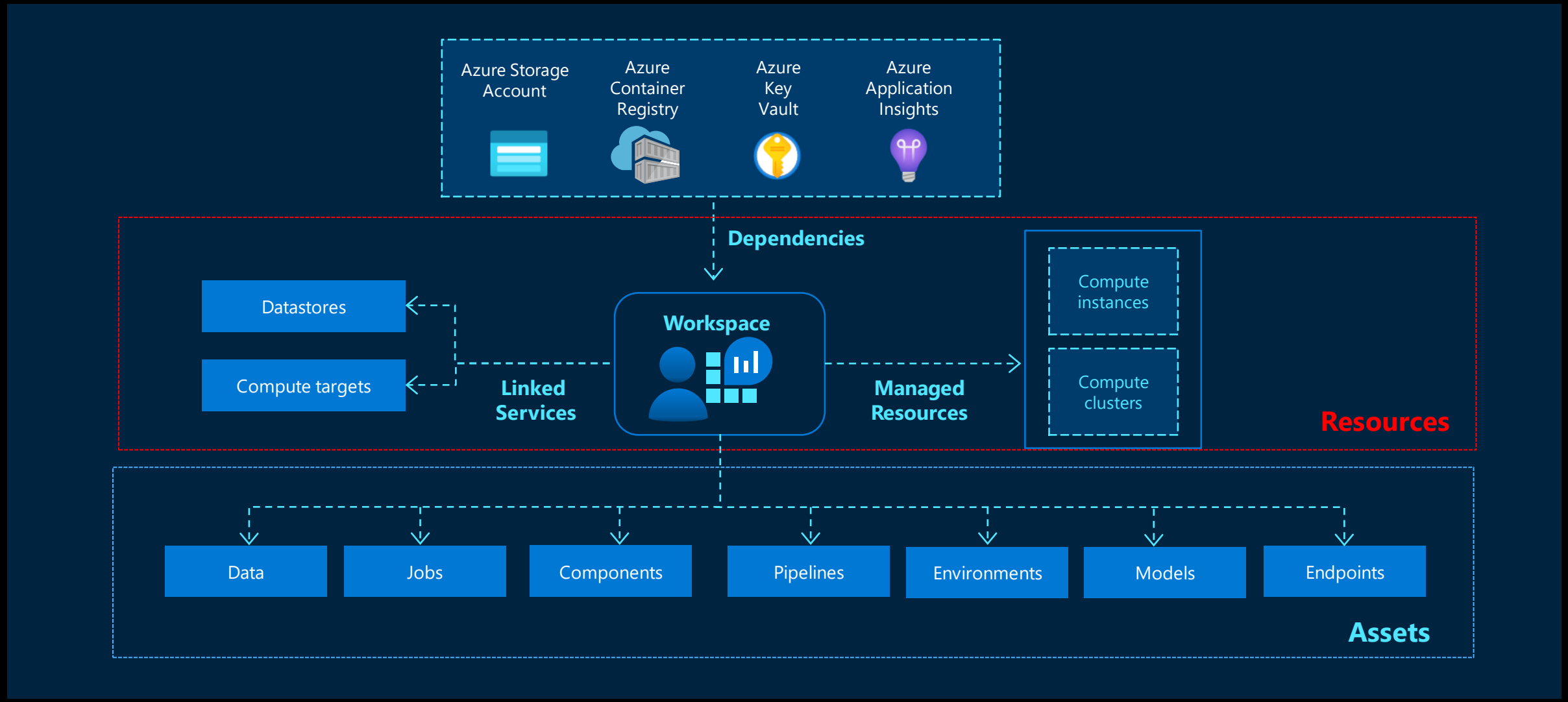
Azure Cloud Services + Python SDK + Cross-Platform CLI

CLI v2 supports python, R, Java, Julia or C#

That enables you to:

- ✓ Prepare Data
- ✓ Build Models
- ✓ Train Models
- ✓ Manage Models
- ✓ Track Experiments
- ✓ Deploy Models

Key Elements of Azure Machine Learning



Azure ML service

Key Artifacts

Instantiation of an image either in Web service (ACI, AKS or FPGA) or via an IoT Module (Docker container).

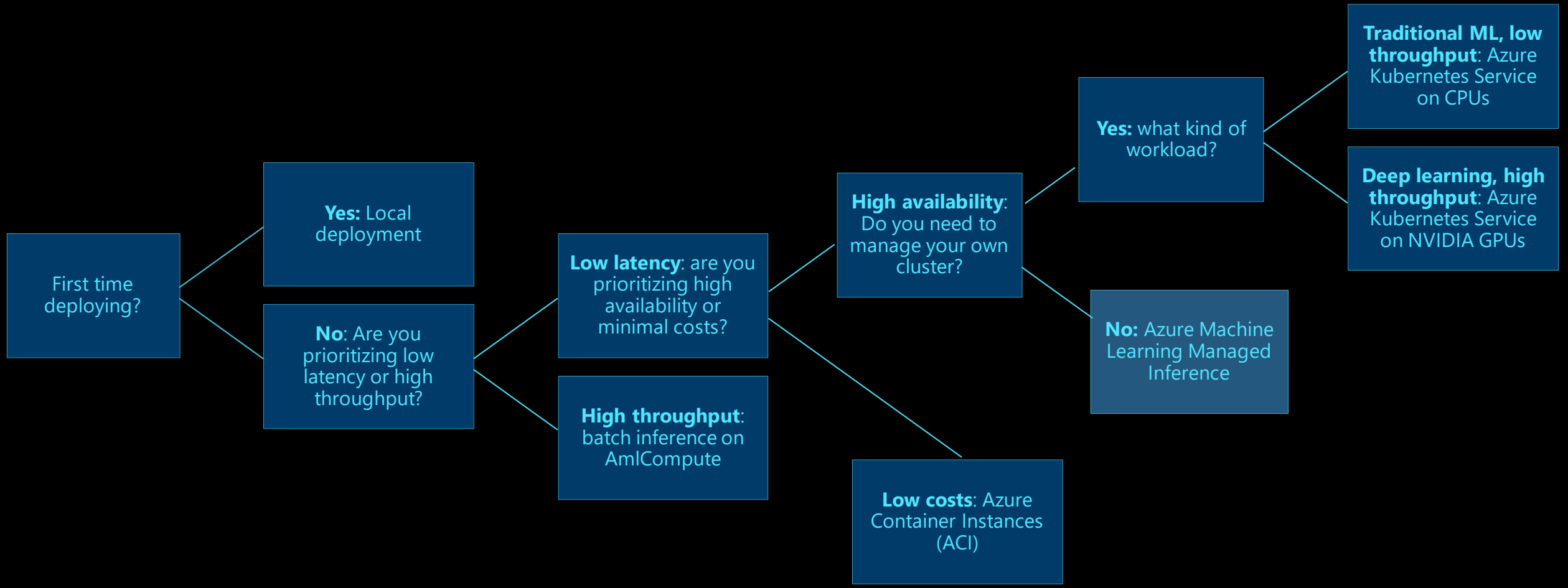
Workspace

- Models
- Models registry
- Experiments
- Pipelines
- Compute Instance
- Compute Ta
- Images
- Images registry
- Deployment / Endpoints
- Dataset / Datastores

We need to specify:

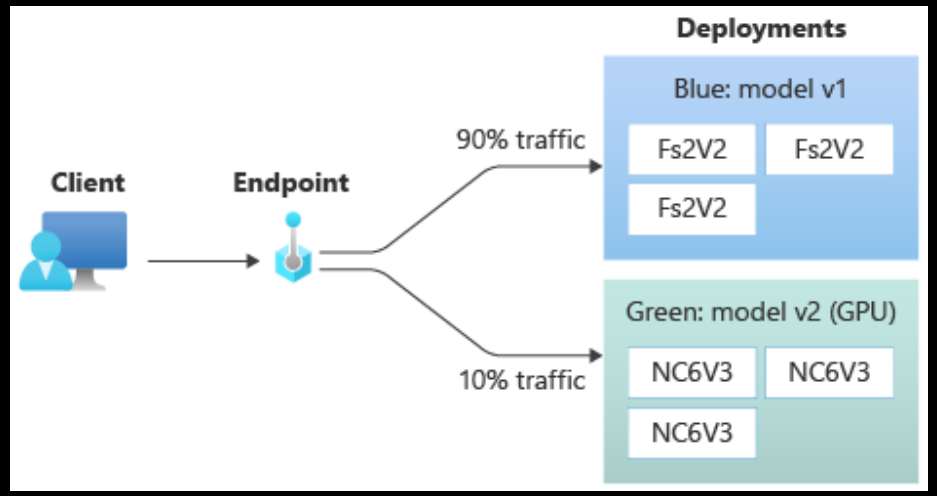
- Model assets
- Scoring script
- Environment
- Compute size + scale settings

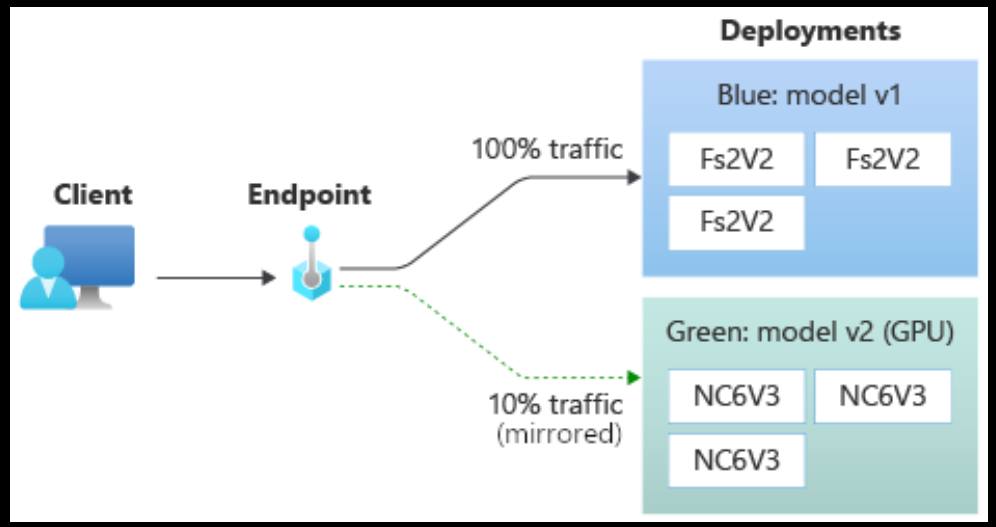
Choosing an inferencing target



Managed Endpoints - GA

•Azure Machine Learning **managed endpoints**, now generally available, help developers and data scientists more easily deploy large-scale machine learning models for both real-time and batch inferencing.

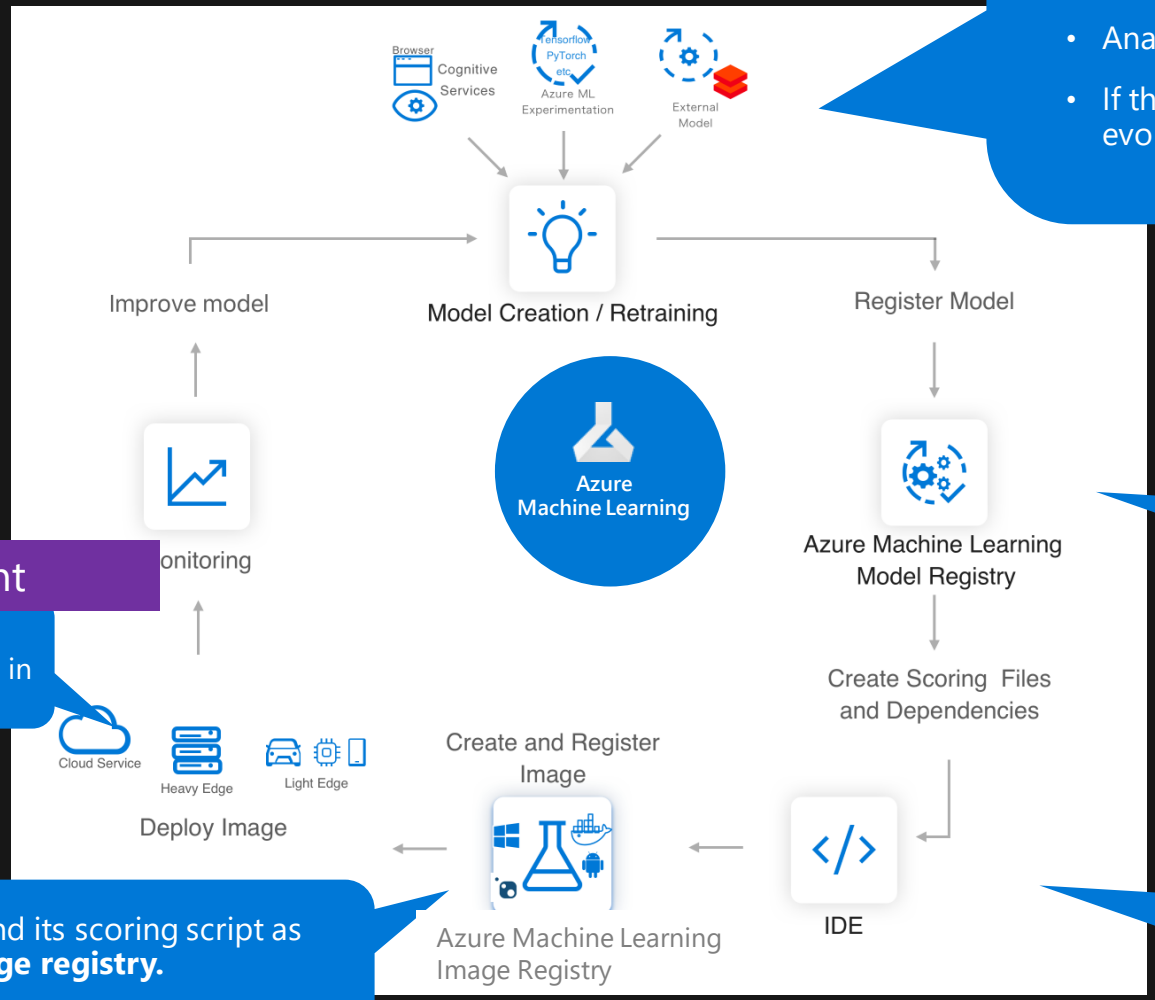




https://docs.microsoft.com/en-us/azure/machine-learning/concept-endpoints

Azure ML service

AI/ML Lifecycle



- Development of a training script in Python.
- Creation and configuration of a **compute target**.
- Run the training script on the compute target.
- During model training, the compute target stores the different executions in a **datastore**. These recordings are grouped together in an **experiment**.
- Analysis of the metrics concerning the executions of an experiment.
- If the metrics do not meet the expected results, return to the first step and evolve the **training script**.

Model creation and training

When a run meets expectations, it is registered as a **model** in the **model registry**.

Model registration

Development of a **scoring script**.

Scoring

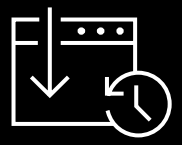
Saving the model and its scoring script as an **image** in an **image registry.**

Image creation

Model deployment

Finally, we deploy the image as a **web service** in Azure.

MLOps: Supporting Technologies



Infrastructure as Code

- **Azure Resource Manager** Templates
- **Azure ML Python SDK & CLI**
- Azure SDK's

CI/CD

- **Azure DevOps Pipelines**
- **Azure ML Pipelines**
- Azure Repos / GitHub
- Azure Boards

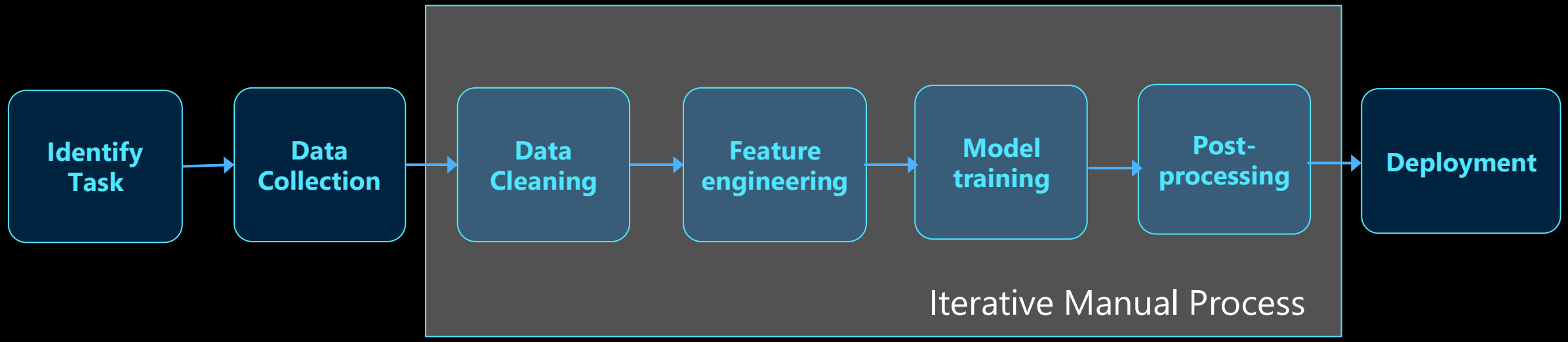
Testing / Release / Monitoring

- **Azure DevOps** for automated testing
- **R** - Runit and testthat
- **Python** - PyUnit, pytest, nose, …
- **Azure ML & MLFlow Tracking**
- **Azure Data Prep SDK** (analyse/profile)
- **Azure ML Model Management** (Instrumentation, Telemetry)
- **Azure ML – Data Drift Detection**
- **Azure Monitor** for app telemetry

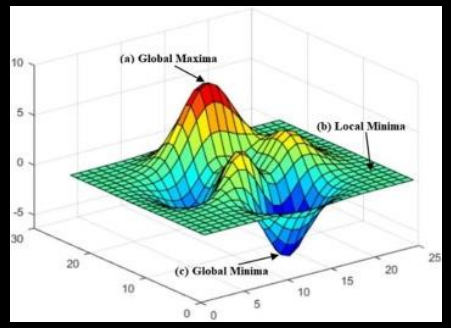
Azure DevOps + Azure ML

Fast, easy to use, well-controlled MLops lifecycle

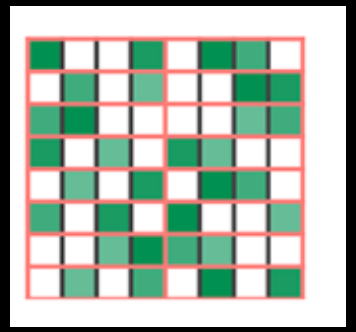
Challenges in Designing ML pipelines



Complex search space



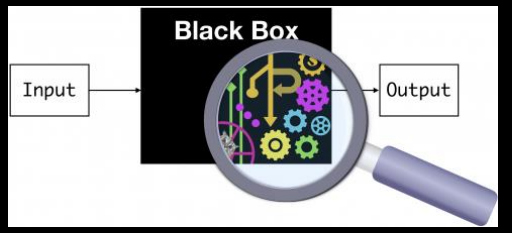
Sparsity of good configurations



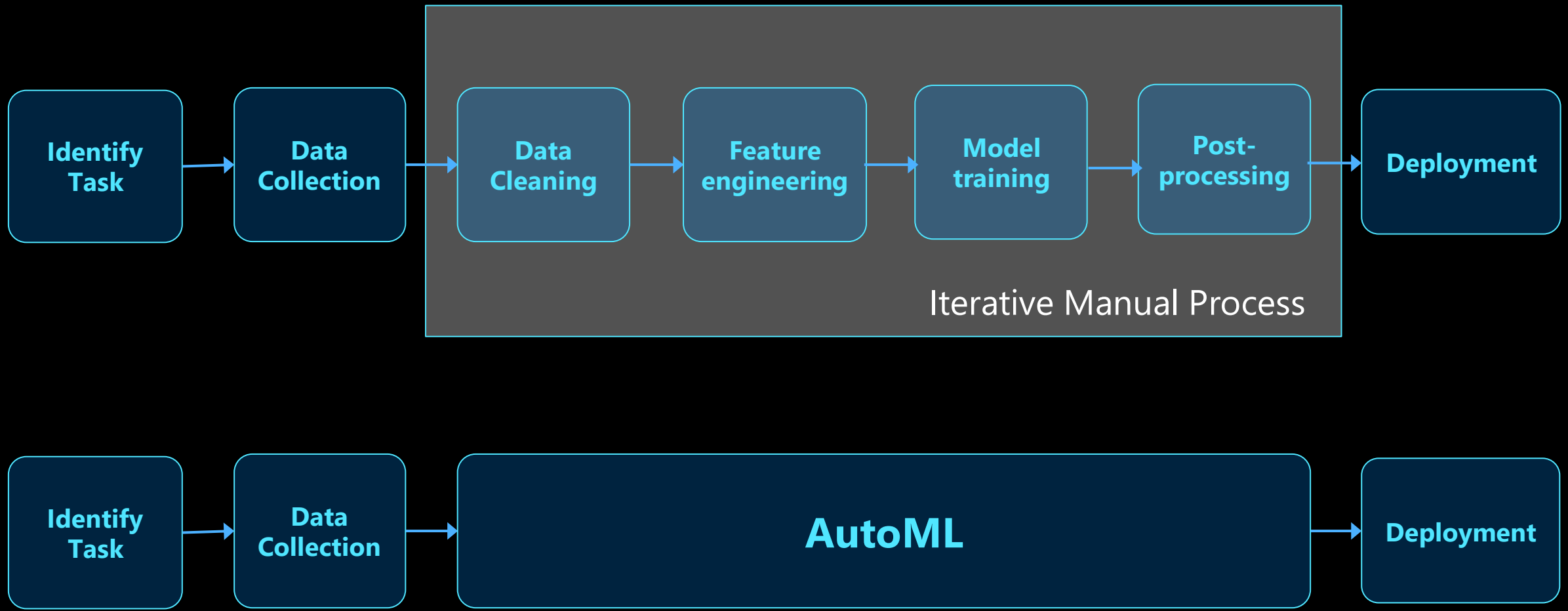
Expensive evaluations

Noise on observations

Black-Box Problem



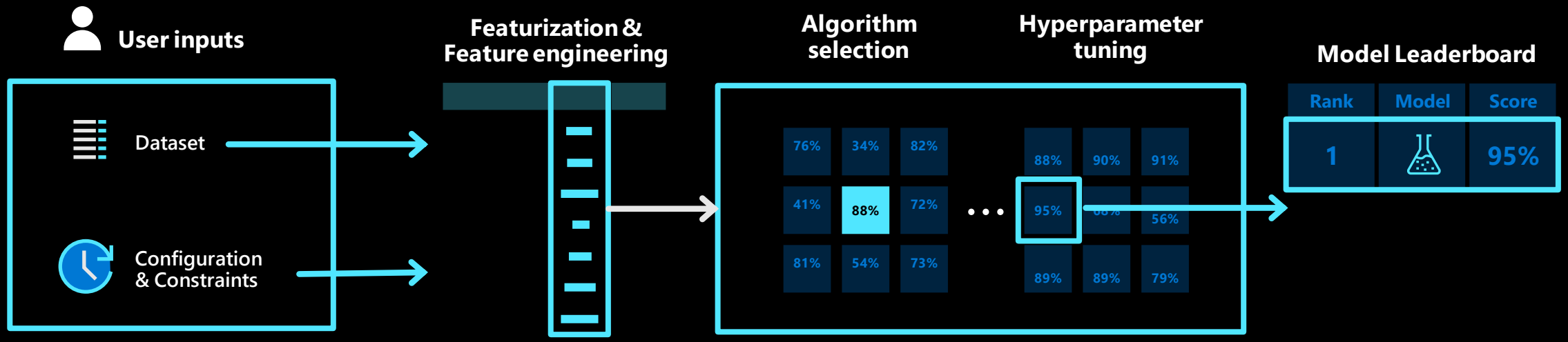
From Manual to Automated ML



What is Automated ML?

Automated machine learning (automated ML) automates feature engineering, algorithm and hyperparameter selection to find the 'best model' for your data.

Loop until reaching **exit criteria**



AML Studio - Automated ML

Data

☑

Data cleaning support

Automated ML currently supports automated data cleaning

Feature

☑

Feature engineering

Most time consuming part when done manually can now be done within minutes.

Algorithm

☑

Pick and play

Testing many different algorithms at once.

Tuning

☑

What to leave out

Hyperparameter tuning: what to include what to leave out

Ranking

☑

Ranking

Having an overview of the best performing models based on accuracy & speed.

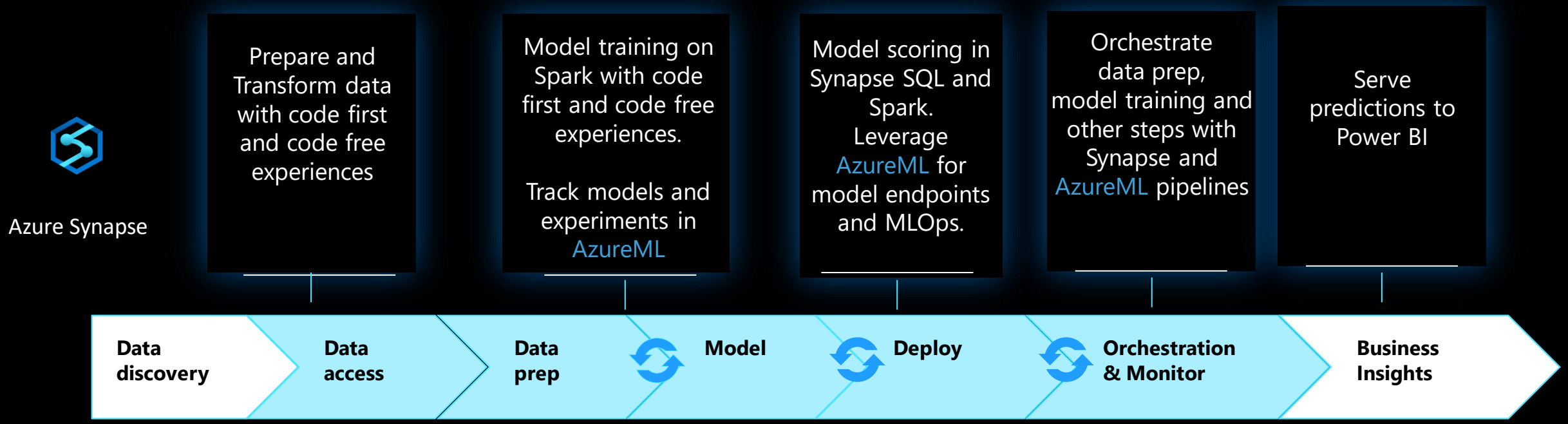
Explaining

☑

Justification

Being able to explain what created an outcome and what features had the most significant impact

Synapse & Azure ML: Supporting the full Data & AI lifecycle



Data wrangling with Apache Spark pools (preview) - Azure Machine Learning | Microsoft Learn

MachineLearningNotebooks/how-to-use-azureml/azure-synapse at master · Azure/MachineLearningNotebooks (github.com)

Q&A