Microsoft

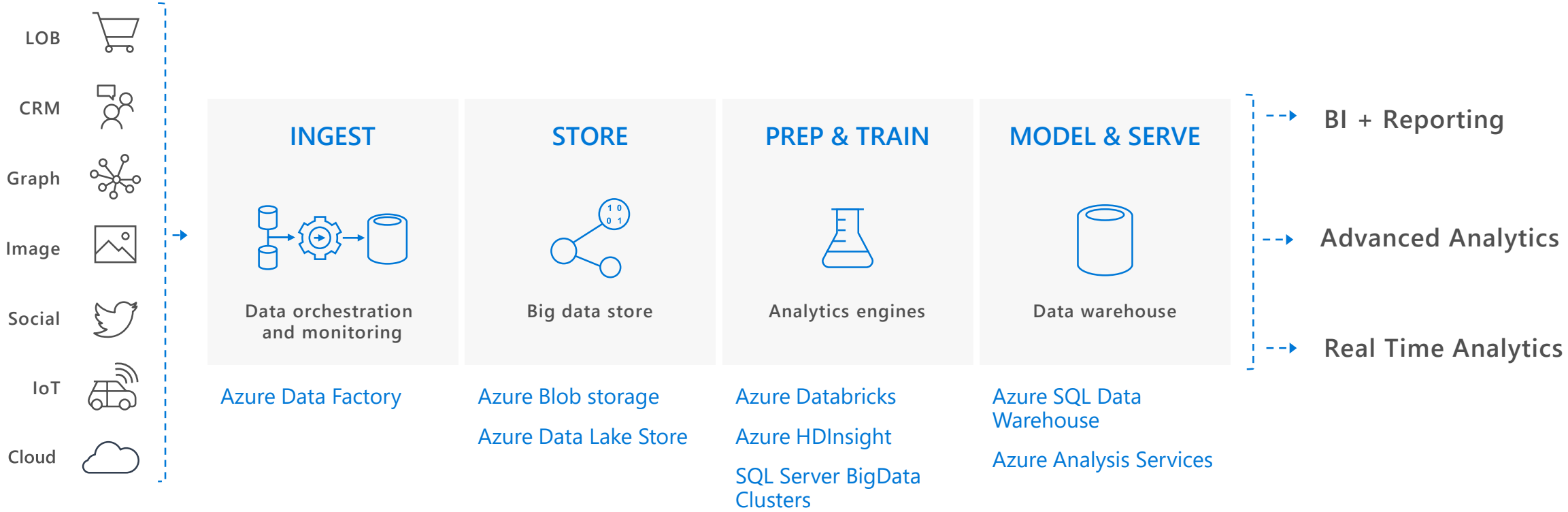
Azure Data Lake Storage

Data will grow to
44 ZB in 2020

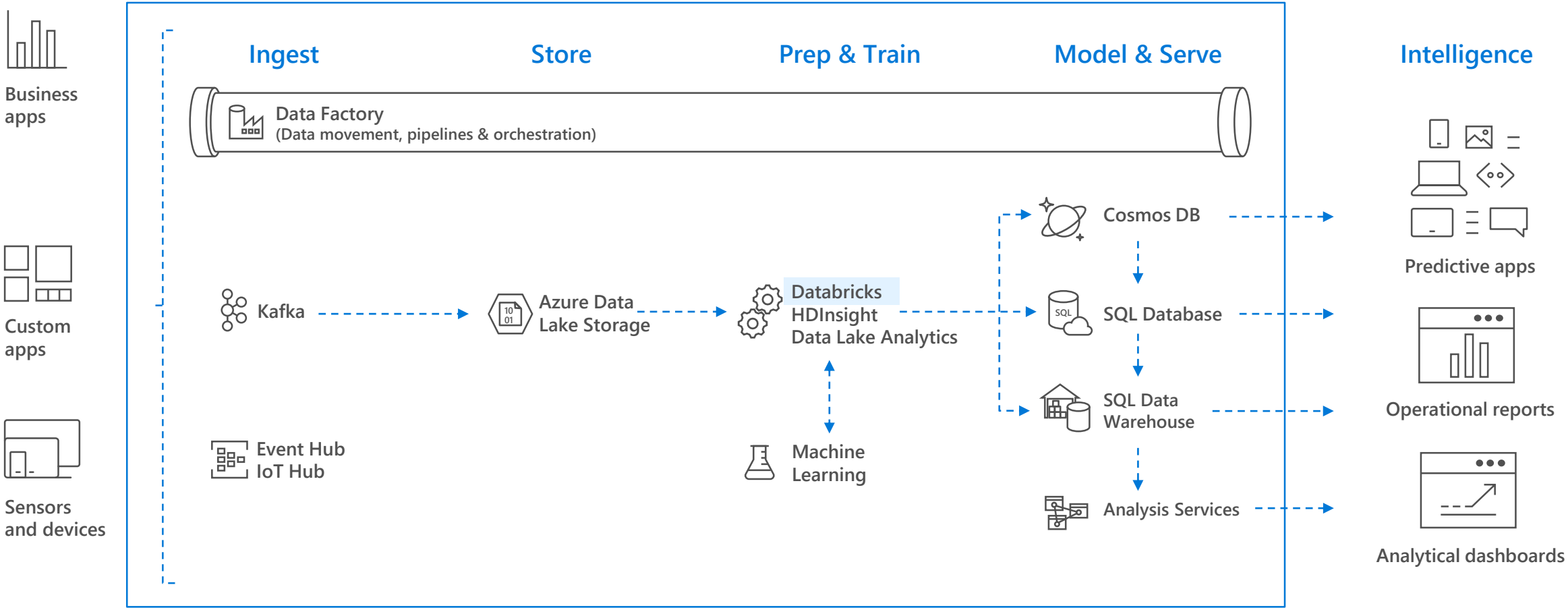
Today, 80% of organizations
adopt cloud-first strategies

AI investment increased by
300% in 2017

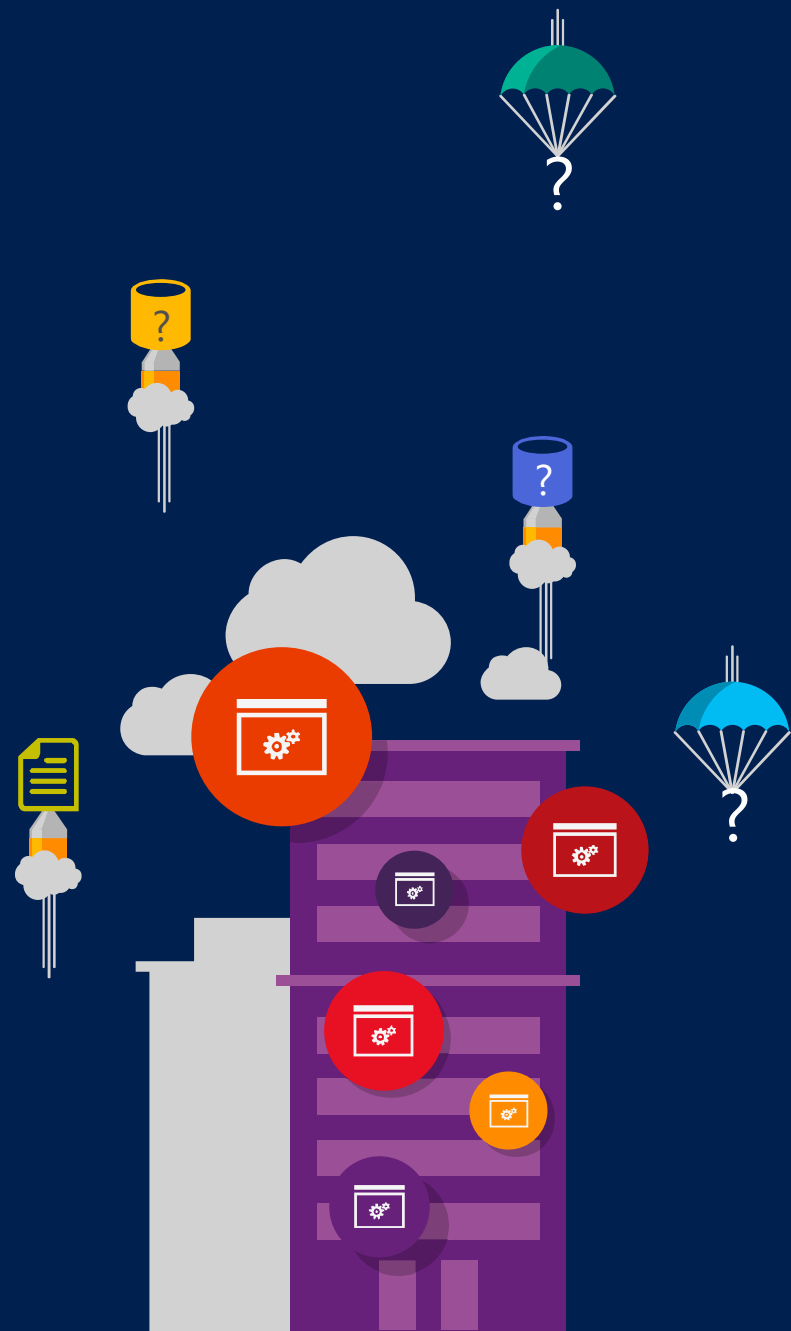
Big data & Data warehouse



BIG DATA & ADVANCED ANALYTICS AT A GLANCE



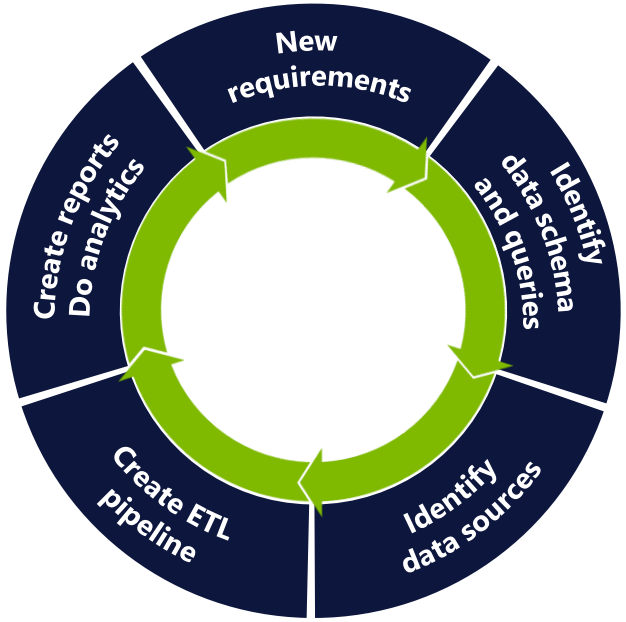
Why data lakes?

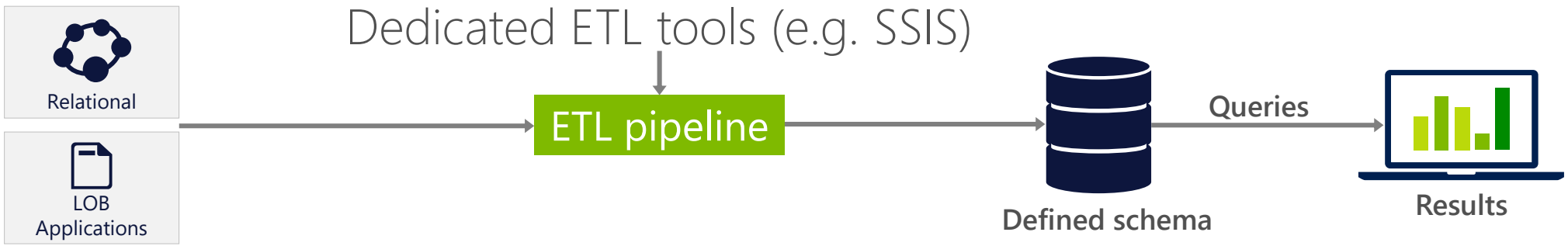


Microsoft

Traditional business analytics process

1. Start with end-user requirements to identify desired reports and analysis
2. Define corresponding database schema and queries
3. Identify the required data sources
4. Create a Extract-Transform-Load (ETL) pipeline to extract required data (curation) and transform it to target schema (*'schema-on-write'*)
5. Create reports. Analyze data

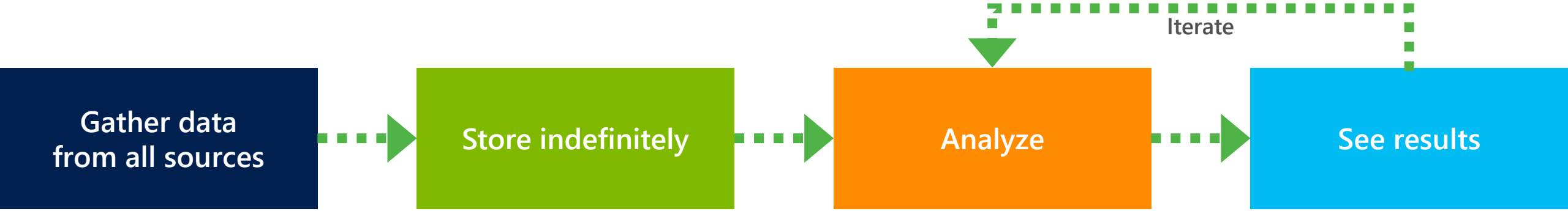




All data not immediately required is discarded or archived

New big data thinking: All data has value

- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time *(schema-on-read)*.
- Apps and users interpret the data as they see fit



Data Lake Store: Technical Requirements

Secure	Must be highly secure to prevent unauthorized access (especially as all data is in one place).
Scalable	Must be highly scalable. When storing all data indefinitely, data volumes can quickly add up
Reliable	Must be highly available and reliable (no permanent loss of data).
Throughput	Must have high throughput for massively parallel processing via frameworks such as Hadoop and Spark
Details	Must be able to store data with all details; aggregation may lead to loss of details.
Native format	Must permit data to be stored in its 'native format' to track lineage & for data provenance.
All sources	Must be able ingest data from a variety of sources-LOB/ERP, Logs, Devices, Social NWs etc.
Multiple analytic frameworks	Must support multiple analytic frameworks—Batch, Real-time, Streaming, ML etc. No one analytic framework can work for all data and all types of analysis.

Scale, performance, reliability

Microsoft

Azure Data Lake Store: no scale limits

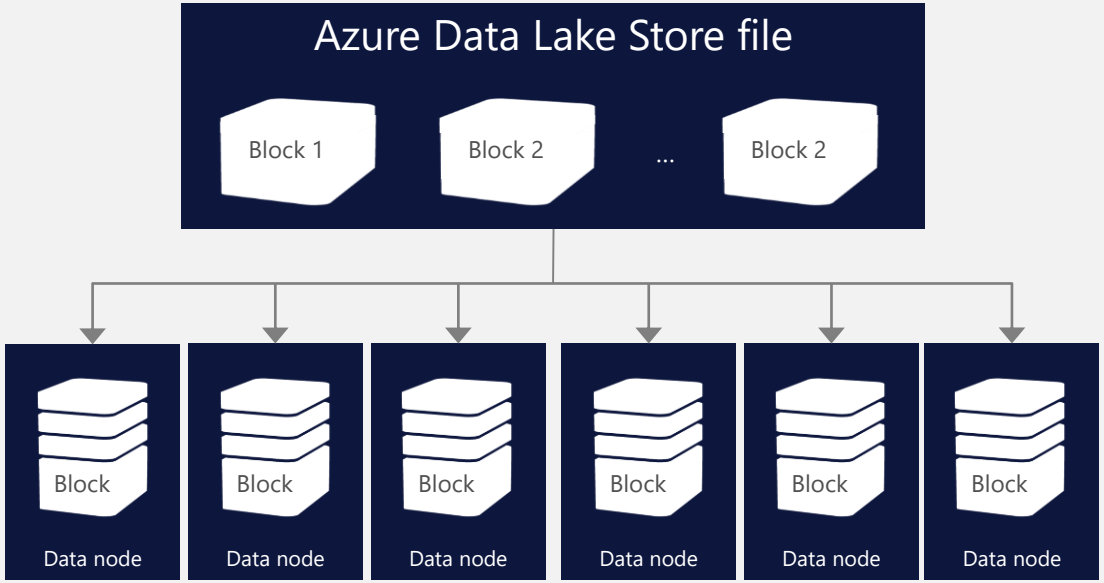
Azure Data Lake Store integrates with Azure Active Directory (AAD) for:

- Amount of data stored
- How long data can be stored
- Number of files
- Size of the individual files
- Ingestion throughput

Seamlessly scales from a few KBs to several PBs

ADL Store Unlimited Scale – How it works

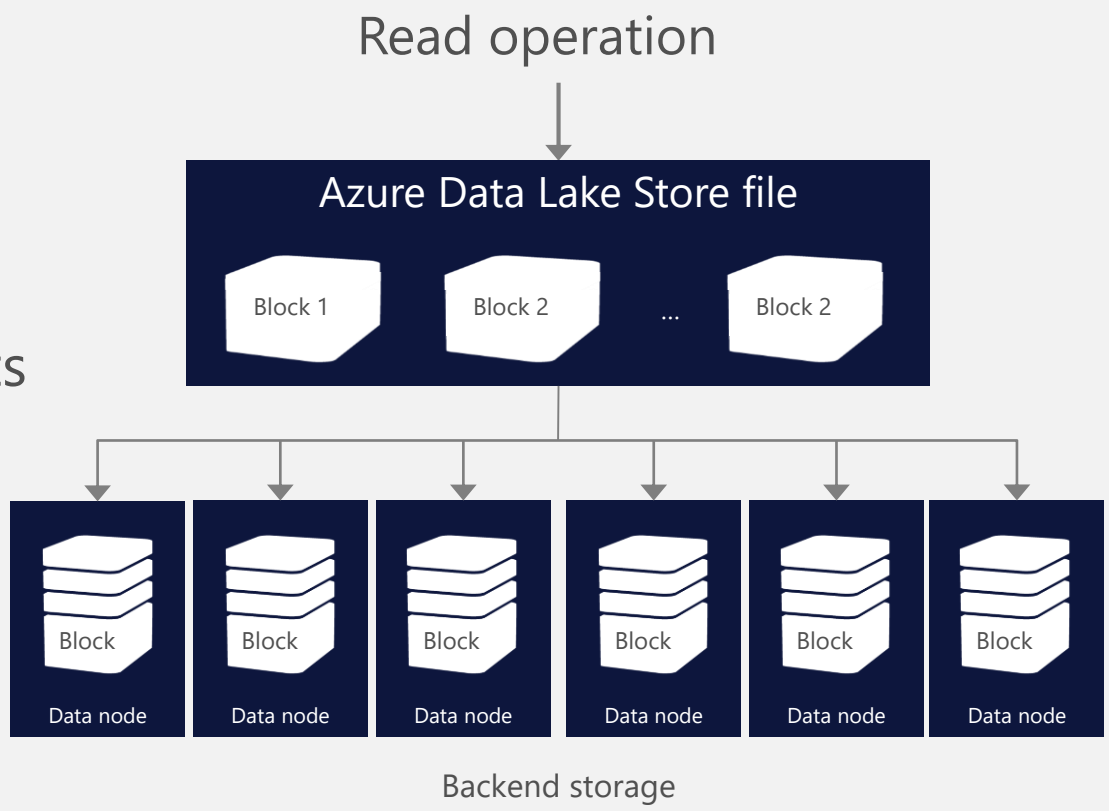
- Each file in ADL Store is sliced into blocks
- Blocks are distributed across multiple data nodes in the backend storage system
- With sufficient number of backend storage data nodes, files of any size can be stored
- Backend storage runs in the Azure cloud which has virtually unlimited resources
- Metadata is stored about each file
  No limit to metadata either.



Backend Storage

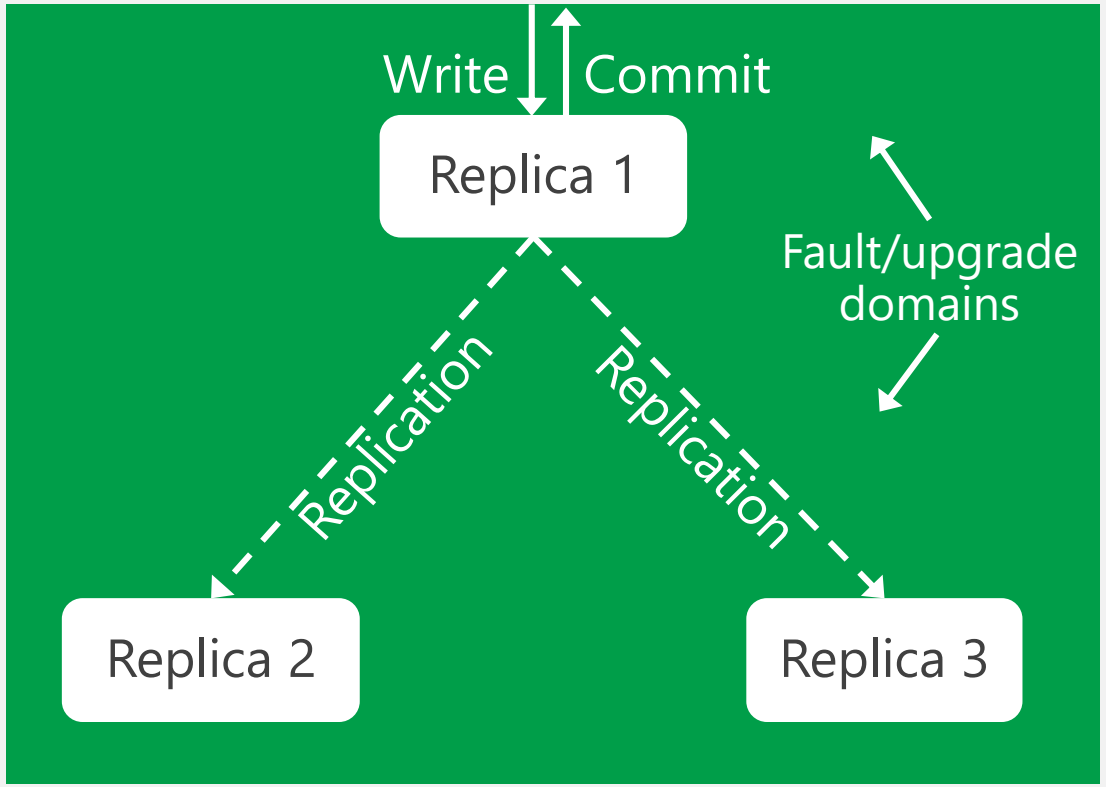
ADL Store offers massive throughput

- Through read parallelism ADL Store provides massive throughput
- Each read operation on a ADL Store file results in multiple read operations executed in parallel against the backend storage data nodes



ADL Store: high availability and reliability

- Azure maintains 3 replicas of each data object per region across three fault and upgrade domains
- Each create or append operation on a replica is replicated to other two
- Writes are committed to application only after all replicas are successfully updated
- Read operations can go against any replica



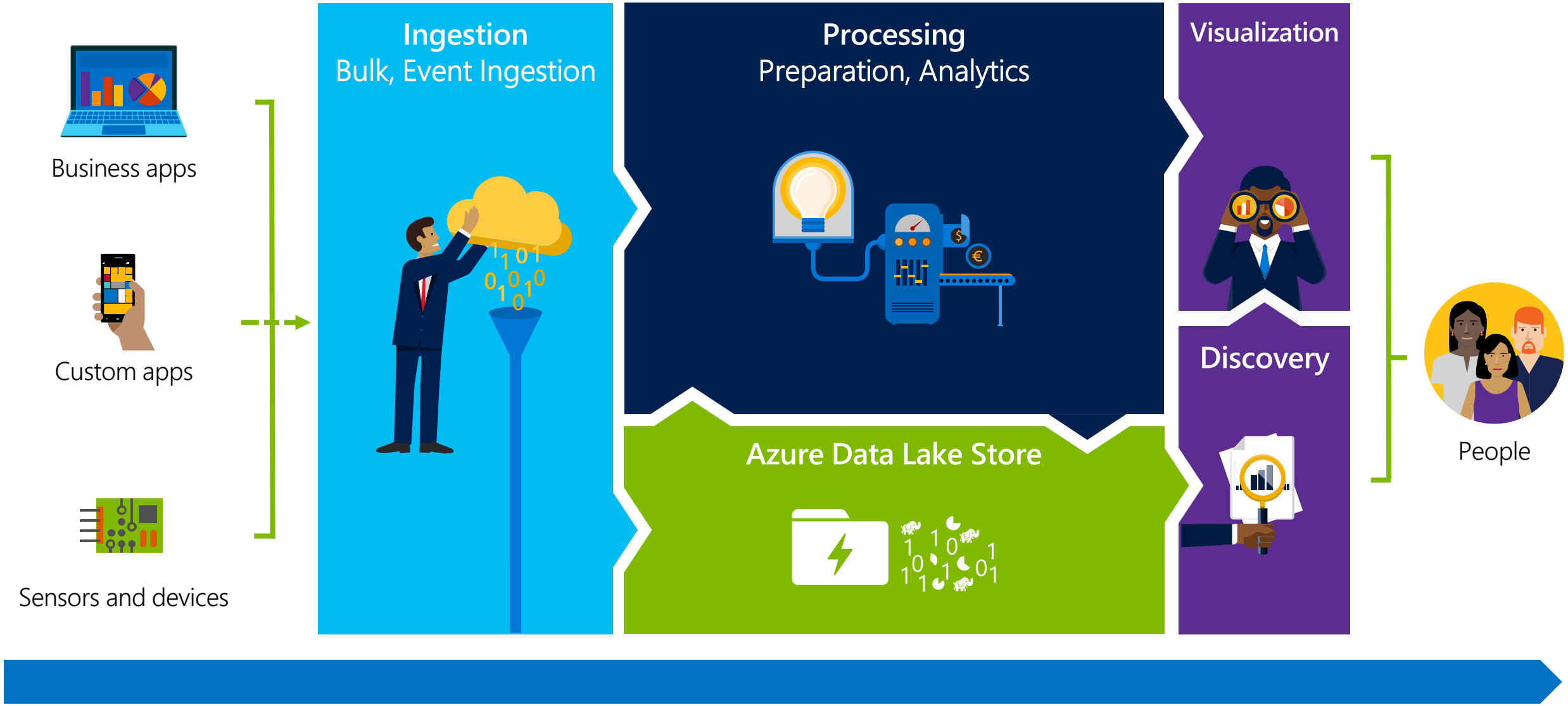
Data is never lost or unavailable even under failures

The building blocks

Ingestion, processing, egress, visualization, and orchestration tools

Microsoft

Big Data Flow



Ingestion tools – Getting started

Data on your desktop

Azure Portal

- Easy to use
- Good for small amount of data
- Analyzing data using Portal

PowerShell

- Upload file and folders
- Control parallelism
- Control format of upload
- Need to use other services

ADL Tools for Visual Studio

- Integrated experience
- Drag-and-drop
- Programmatic Analytics

CLI

- Linux, Mac
- Most features of PowerShell

Data located in other stores

Azure Data Factory

- Copy Wizard for intuitive one-time copy from multiple sources

AdlCopy

- Copy data easily from Azure Storage at least cost

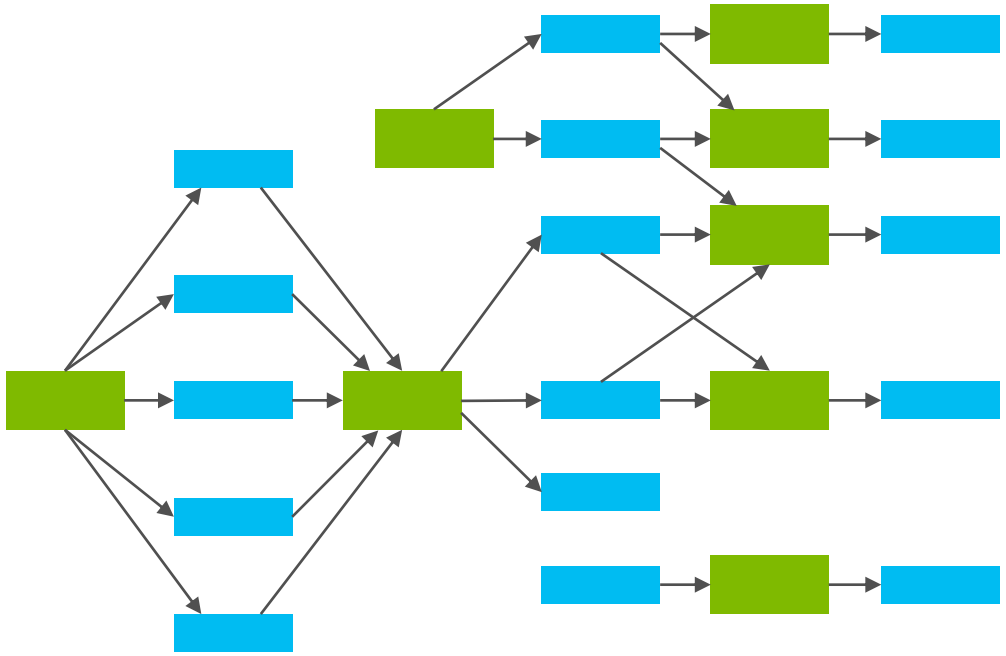
OSS tools on HDI

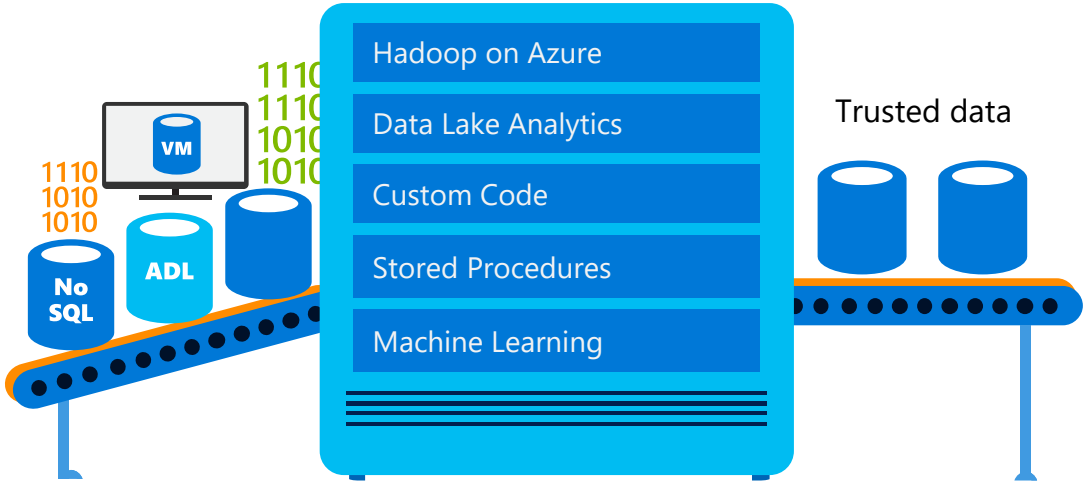
- Distcp, Sqoop
- If analyzing data using HDInsight

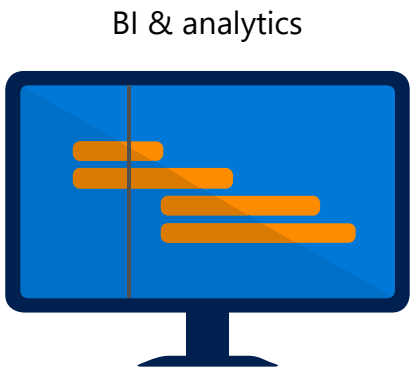
Azure Data Factory

Compose, orchestrate & monitor data services at scale

- Fully managed service
- Any data on-premises or in the cloud
- Single pane of glass management
- Global service infrastructure
- Cost Effective

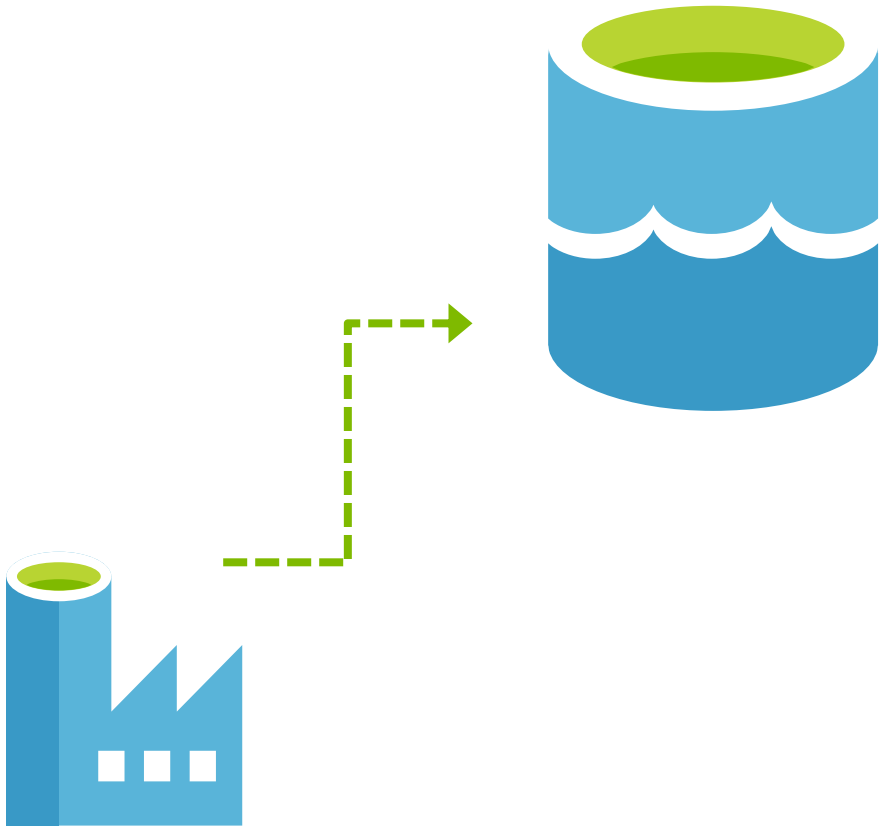






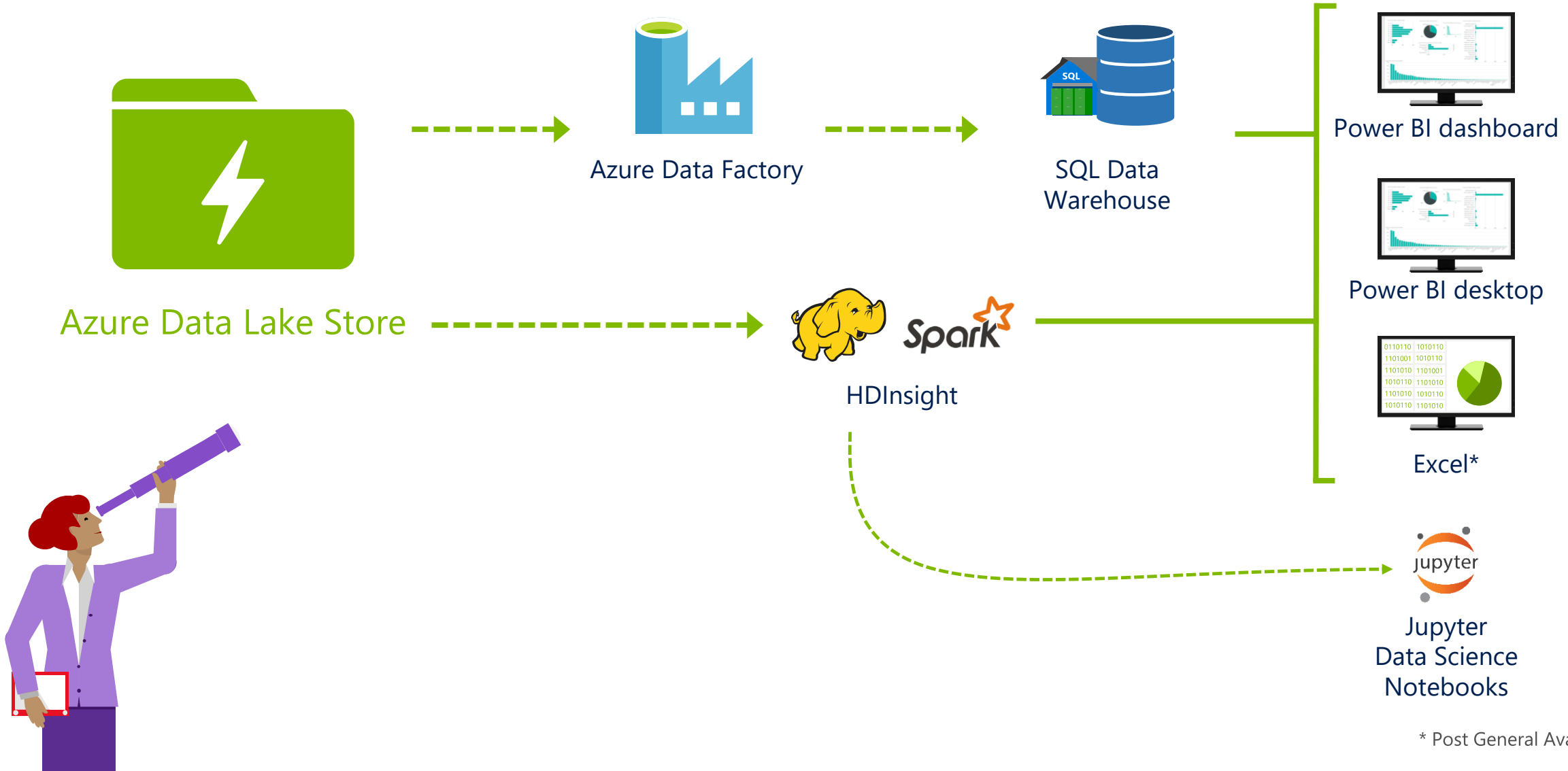
Azure Data Factory

Connects ADL Store out-of-the-box to all your stores



Category	Data store	Supported as source	Supported as sink
Azure	**Azure Data Lake Store**	●	●
	Azure Blob storage	●	●
	Azure SQL Database	●	●
	Azure SQL Data Warehouse	●	●
	Azure Table storage	●	●
	Azure DocumentDB	●	●
Databases	SQL Server*	●	●
	Oracle*	●	●
	MySQL*	●	
	DB2*	●	
	Teradata*	●	
File	HDFS*	●	
	Others	●	

Visualizing data

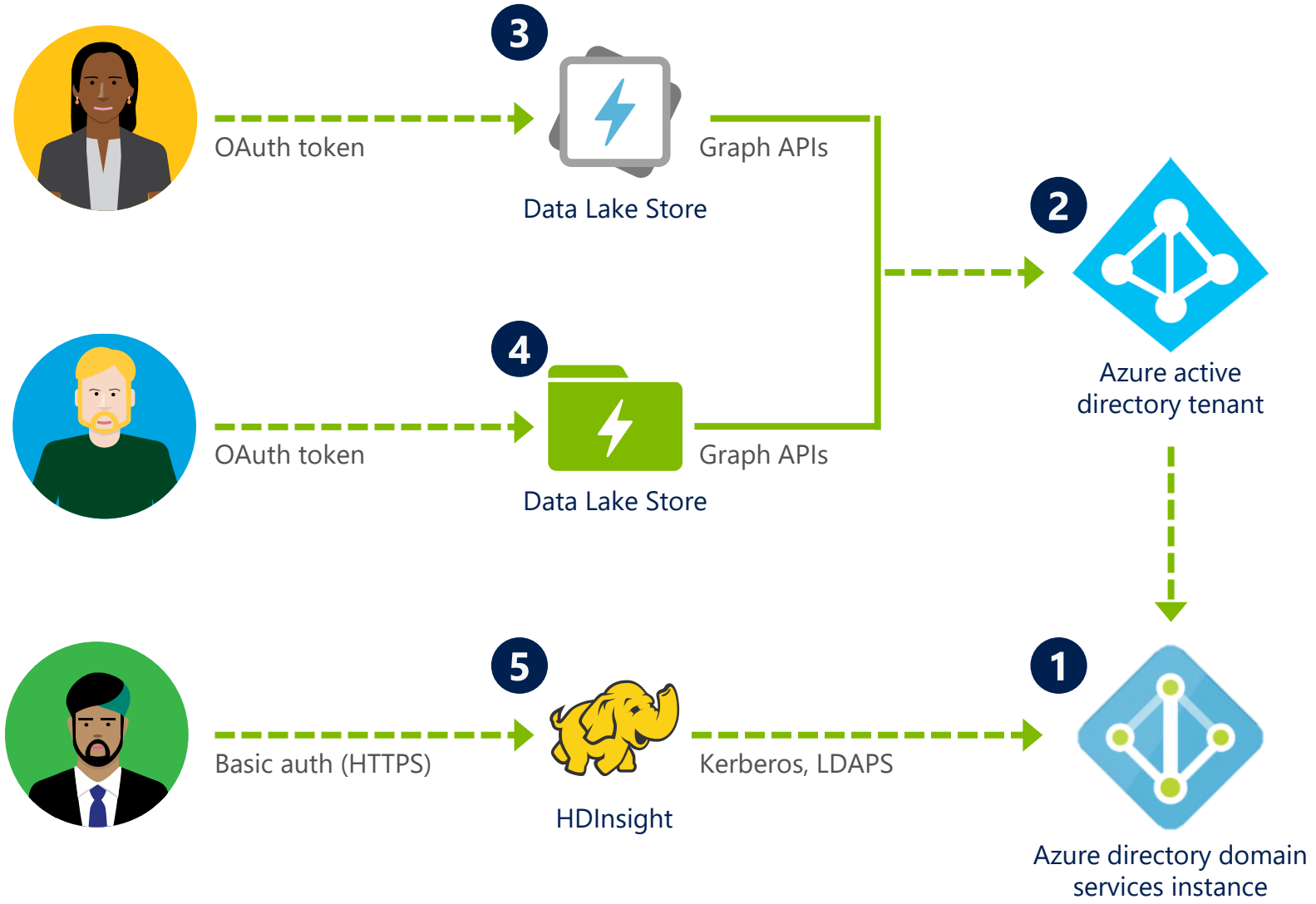


* Post General Availability

ADL Store Security: AAD integration

- Multi-factor authentication based on OAuth2.0
- Integration with on-premises AD for federated authentication
- Role-based access control
- Privileged account management
- Application usage monitoring and rich auditing
- Security monitoring and alerting
- Fine-grained ACLs for AD identities

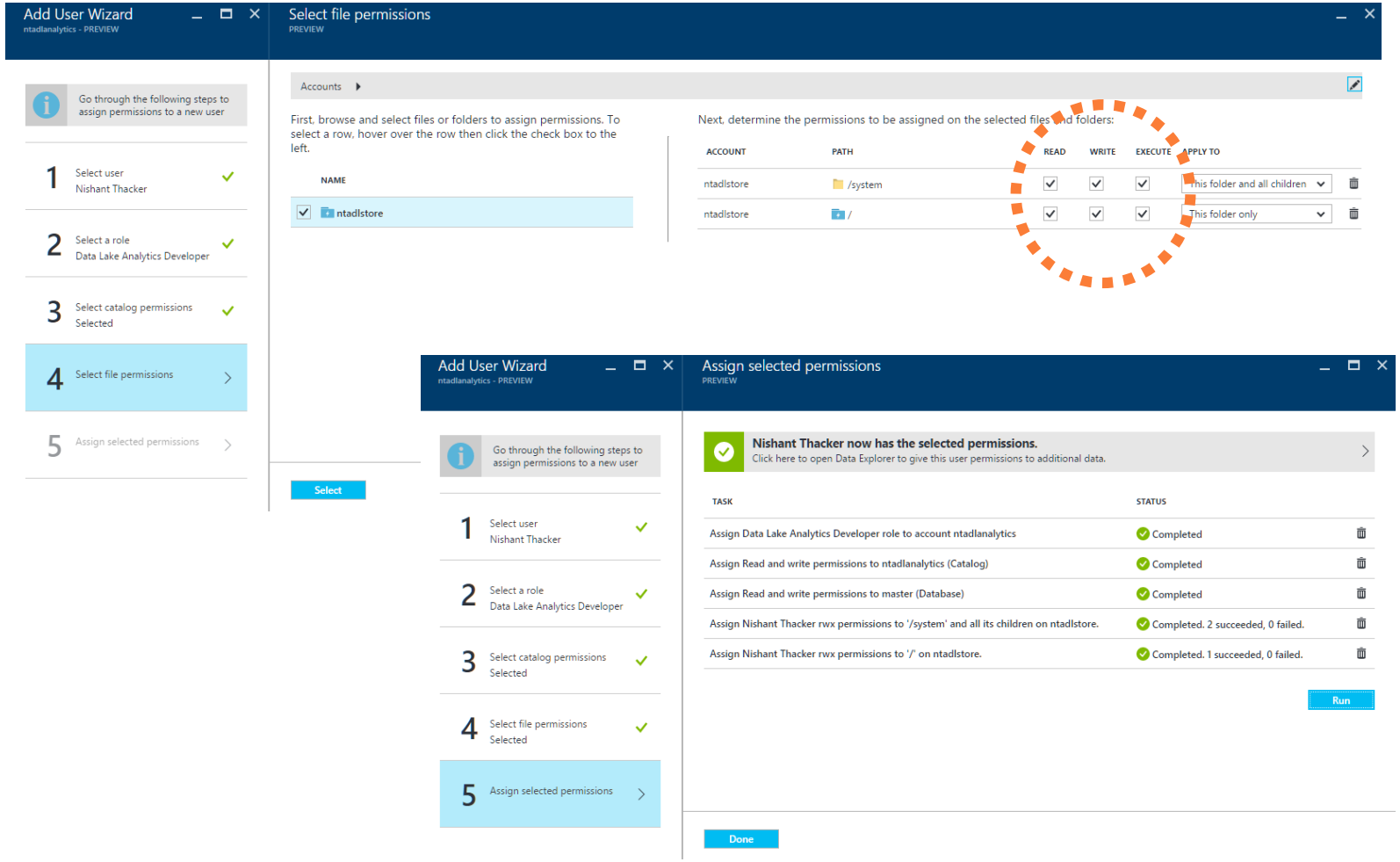
Leveraging Azure Active Directory



1. Create ADDS instance in separate VNET
2. Add users to AAD Tenant
3. Add users to ADLA RBAC roles
4. Add users to ADLS RBAC roles & file system ACLs
5. Join HDInsight cluster to ADDS instance

ADL Store security: Role-based access

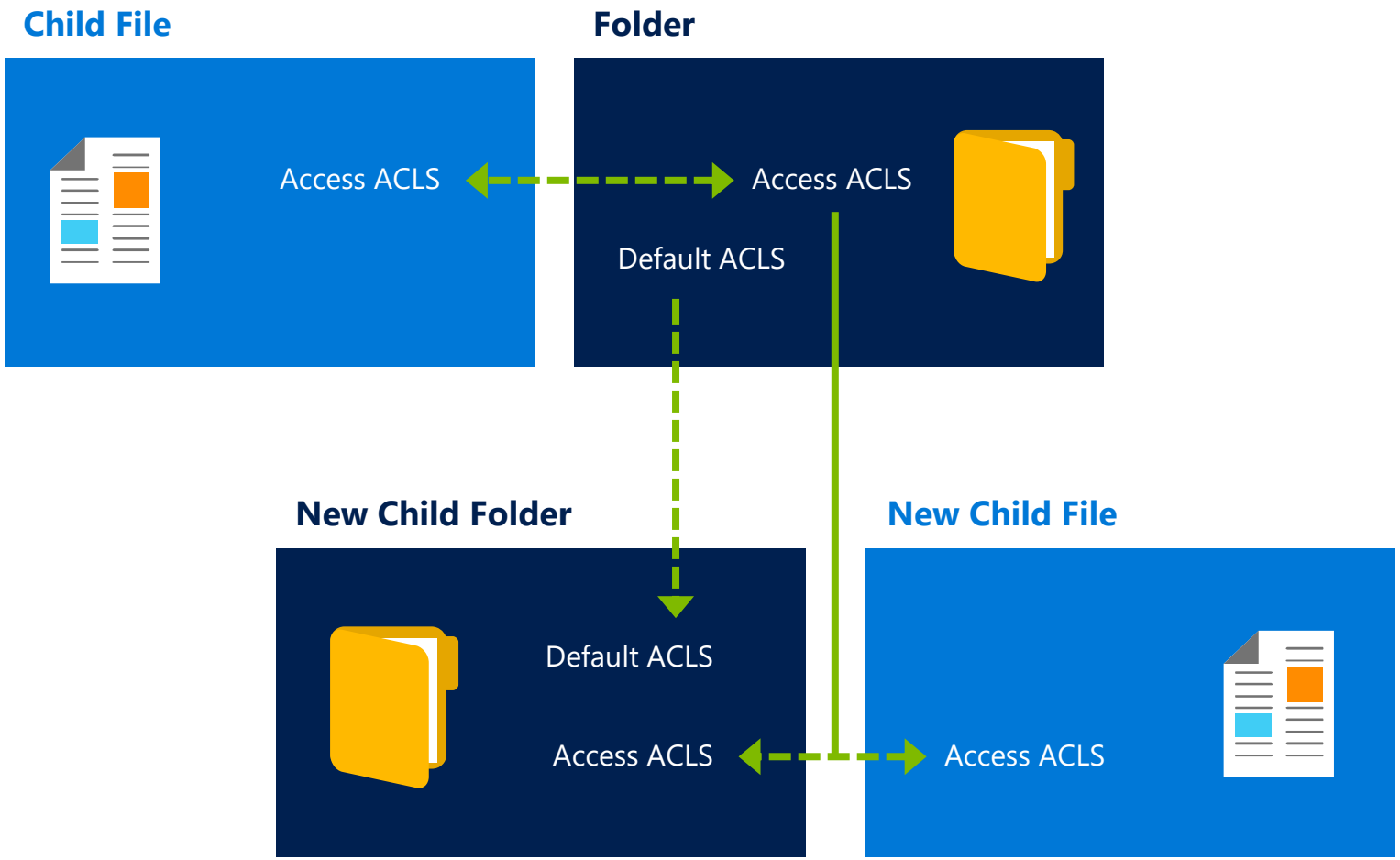
- Each file and directory is associated with an owner and a group
- Files or directories have separate permissions (read(r), write(w), execute(x)) for owners, members of the group, and for all other users
- Fine-grained access control lists (ACLs) rules can be specified for specific named users or named groups



Granular control of file and folder access

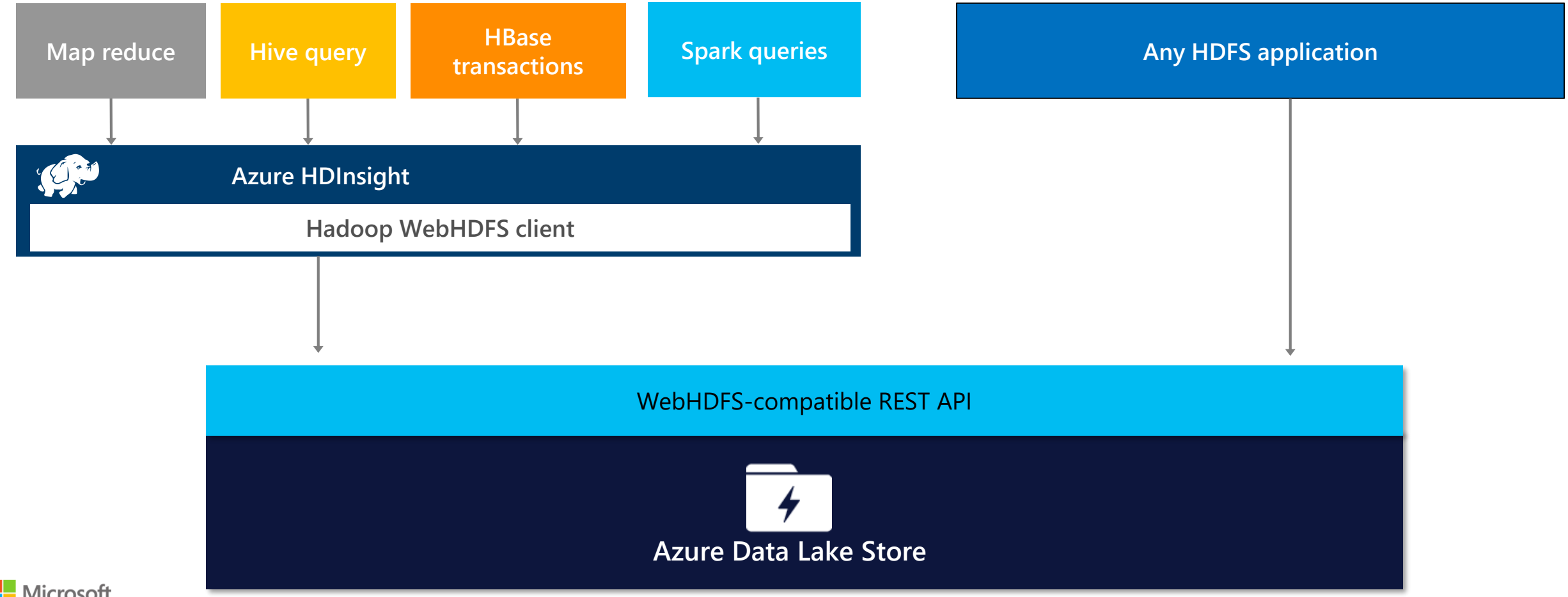
POSIX-Style ACLs with full compatibility with HDFS/WebHDFS

- Generate default ACLs for files and folders
- Customize for fine-tuned control
- Access ACLs control how a user can access to the file or folder
- Default ACLs used to construct the Access ACL of new children
- Default ACLs copied to the Default ACL of new child folders

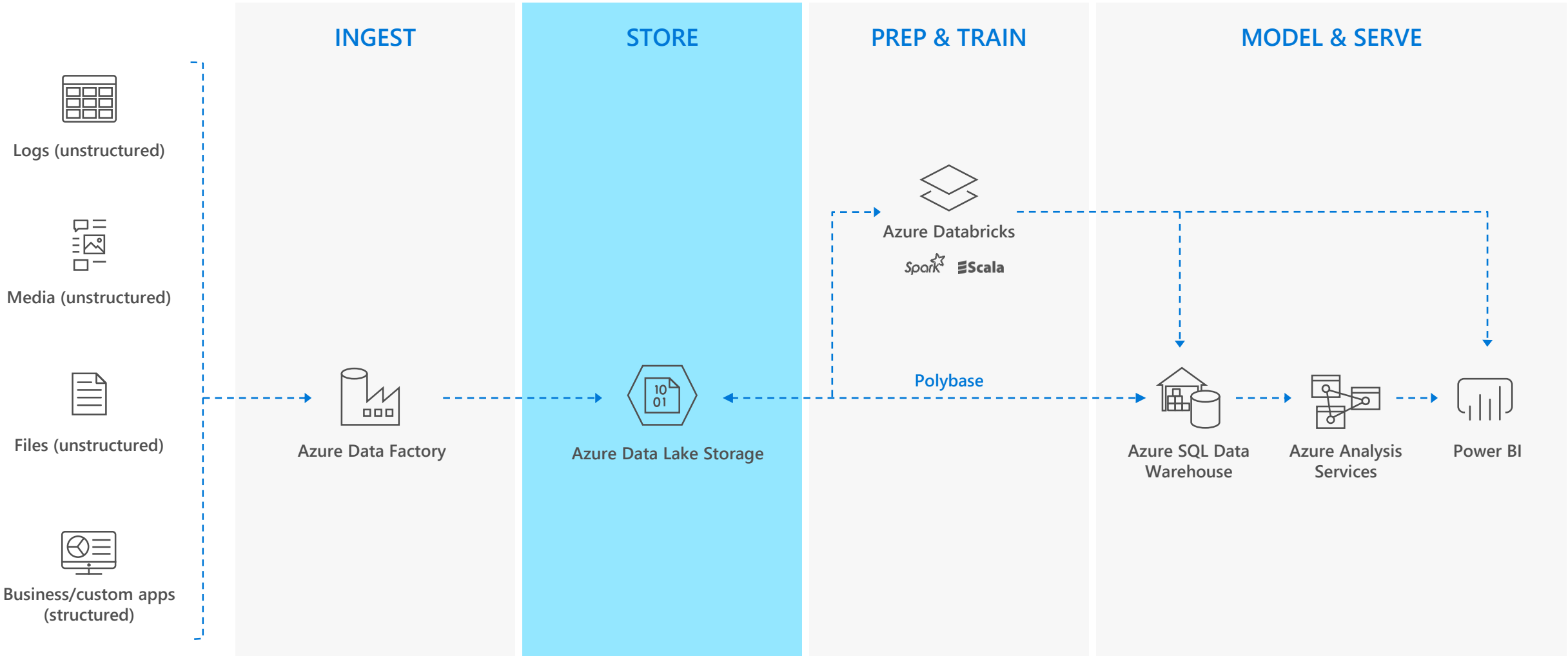


ADL Store is HDFS-compatible

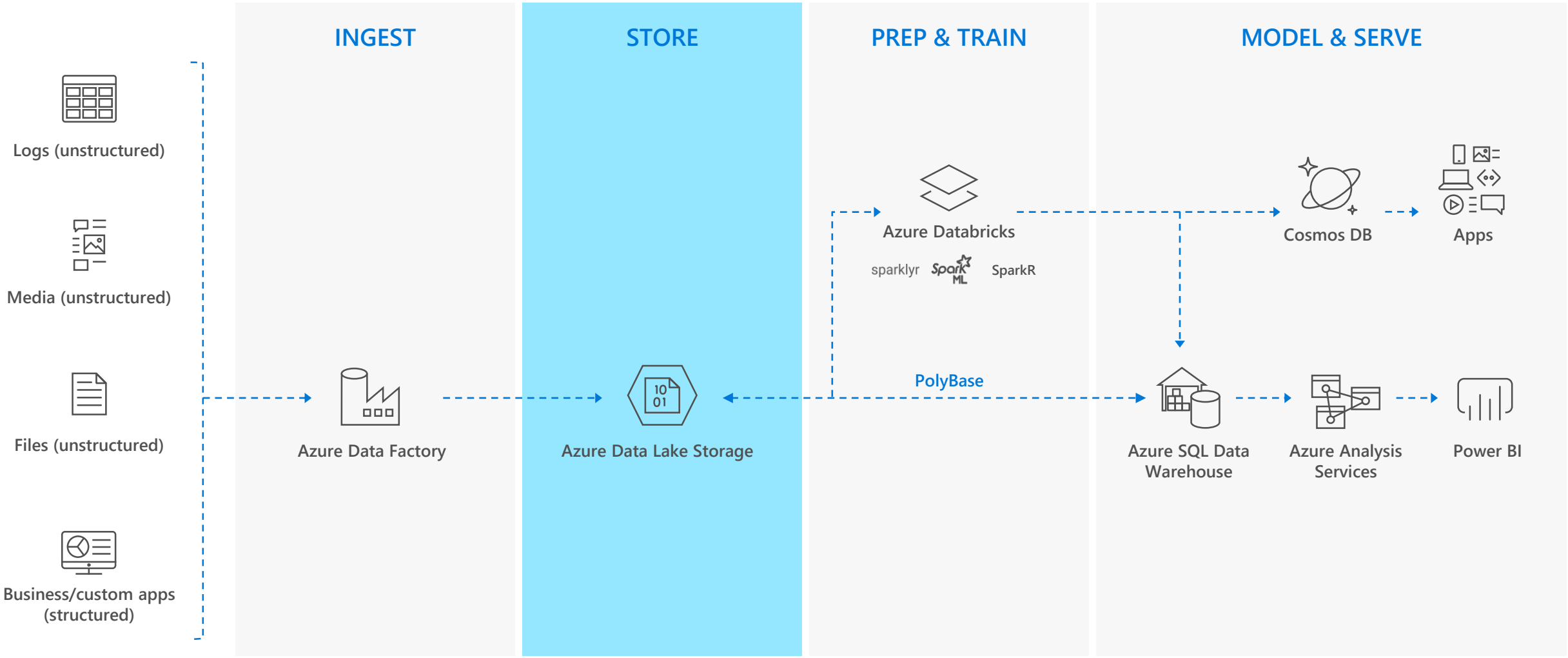
With a WebHDFS endpoint Azure Data Lake Store is a Hadoop-compatible file system that integrates seamlessly with Azure HDInsight



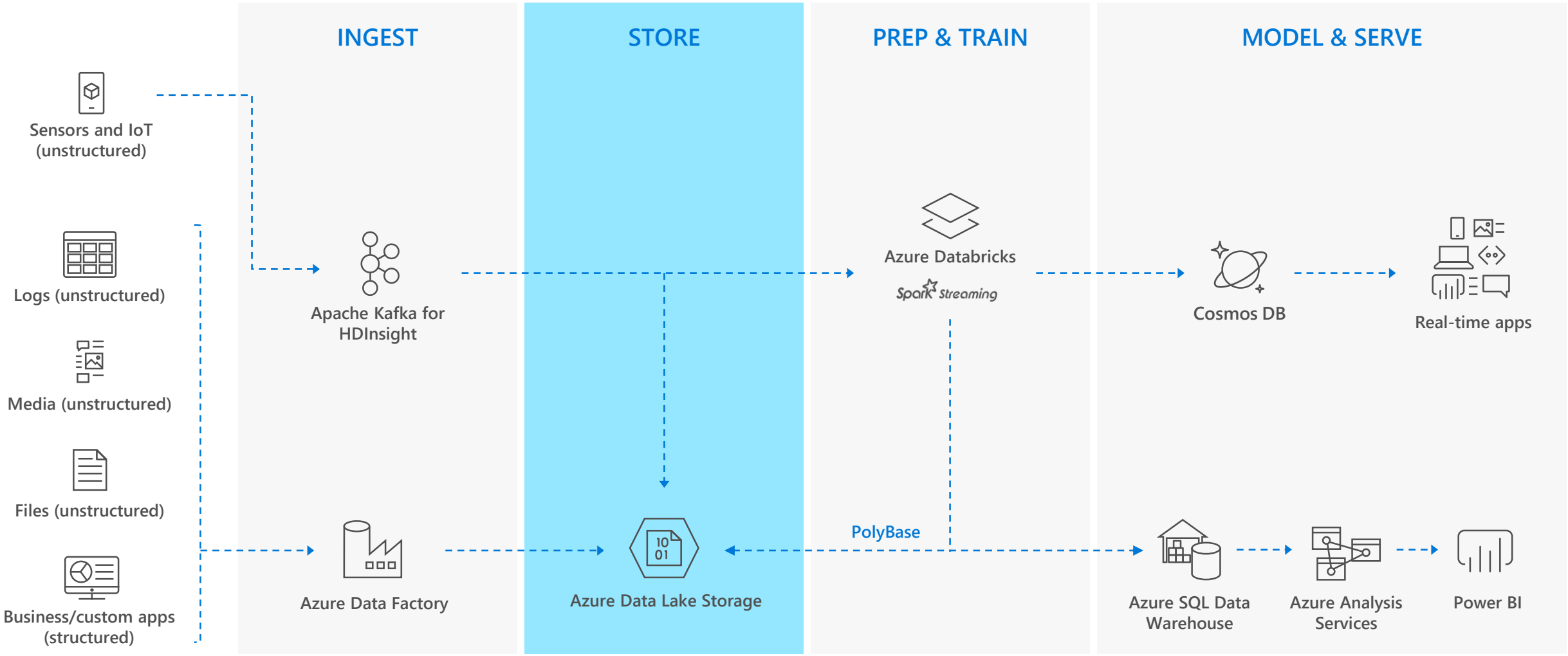
MODERN DATA WAREHOUSE



ADVANCED ANALYTICS



REAL-TIME ANALYTICS



Data Lake Storage Pricing Model

Data Lake Storage Pricing

Azure Blob Tiers	Hot	Cool	Archive	S3 Standard Storage	Big Query
	Frequently accessed data	Less frequently accessed data	Rarely accessed data		
PER TB PER MONTH	$18.40	$10.00	$2.00	$23.00	$20.00

- Same storage pricing as Azure Blob Storage!*
- Supports object level tiering
- Competitive with other cloud providers

** Transaction pricing model differences exist; Pricing for LRS availability*

RICH PARTNER NETWORK

Trusted ecosystem to accelerate time to value

DATA MIGRATION

Informatica

wanDISCO

ATTUNITY

striim

SPARK & HADOOP ENGINES

HORTONWORKS

MAPR

cloudera

Qubole

databricks