Microsoft Azure

Big data processing with Azure Databricks

Chris Campbell
Data and AI Architect
Microsoft Technology Center, Detroit

Sam Istephan
National AI Architect
Microsoft Technology Centers

WiFi: MSFTGuest
Click on Event Attendee Code

Download Lab File From:
https://github.com/ckcampbell248/ADB_MTC_Workshop

Microsoft Azure

Productive + Hybrid + Intelligent + Trusted

Azure

Dev Tools + DevOps

Containers + Serverless

Internet of Things

Data

Artificial Intelligence

Azure: Trusted

Global

- ISO 27001
- ISO 27018
- ISO 27017
- ISO 22301
- SOC 1 Type 2
- SOC 2 Type 2
- SOC 3
- CSA STAR Self-Assessment
- CSA STAR Certification
- CSA STAR Attestation

Regional

- Argentina PDPA
- EU Model Clauses
- UK G-Cloud
- China DJCP
- China GB 18030
- China TRUCS
- Singapore MTCS
- Australia IRAP/CCSL
- New Zealand GCIO
- Japan My Number Act
- ENISA IAF
- Japan CS Mark Gold
- Spain ENS
- Spain DPA
- India MeitY
- Canada Privacy Laws
- Privacy Shield
- Germany IT Grundschutz workbook

Industry

- PCI DSS Level 1
- CDSA
- MPAA
- FACT UK
- Shared Assessments
- FISC Japan
- HIPAA/ HITECH Act
- HITRUST
- GxP 21 CFR Part 11
- MARS-E
- IG Toolkit UK
- FERPA
- GLBA
- FFIEC

Us Gov

- Moderate JAB P-ATO
- High JAB P-ATO
- DoD DISA SRG Level 2
- DoD DISA SRG Level 4
- DoD DISA SRG Level 5
- SP 800-171
- FIPS 140-2
- Section 508 VPAT
- ITAR
- CJIS
- IRS 1075

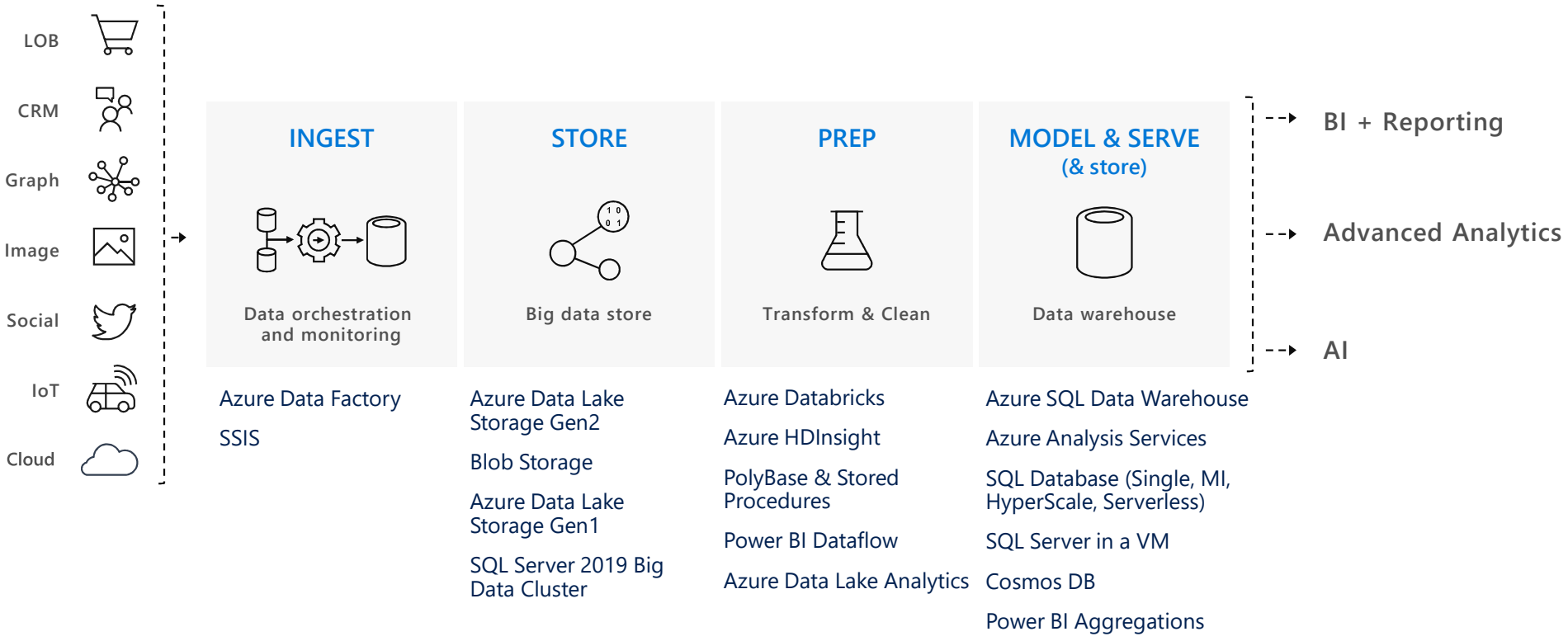
54

Azure regions

More than AWS &
Google combined

Azure for Big Data and Advanced Analytics

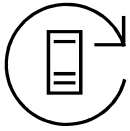
Modern Data Warehouse (possible products by four areas)



Note: Those products that span more than one area are listed in their primary area

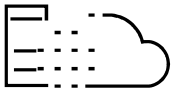
Azure data + AI

Data



Modernization on premises

SQL Server 2017



Hybrid data estate

Azure SQL Database
Azure DB for MySQL & PostgreSQL

Cloud Native Apps

Azure Cosmos DB

Cloud Scale Analytics

Azure SQL DW
Azure Databricks

AI

Custom AI

Azure Databricks
Azure ML service (Preview)

Pre-Built AI

Cognitive Services

Conversational AI

Bot Services

Seamlessly integrate with the Azure portfolio

Azure active directory | ExpressRoute | OMS | KMS | Visual studio tools

Machine Learning on Azure

Domain specific pretrained models
To reduce time to market


Vision


Speech


Language


Search

Familiar Data Science tools
To simplify model development


PyCharm


Jupyter


Visual Studio Code


Command line

Popular frameworks
To build advanced deep learning solutions


Pytorch


TensorFlow


Scikit-Learn


Onnx

Productive services
To empower data science and development teams


Azure Databricks


Azure Machine Learning


Machine Learning VMs

Powerful infrastructure
To accelerate deep learning

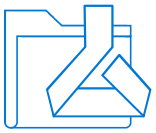

CPU


GPU


FPGA


From the Intelligent Cloud to the Intelligent Edge


Azure ML service



Workspace

Models

Experiments

Pipelines

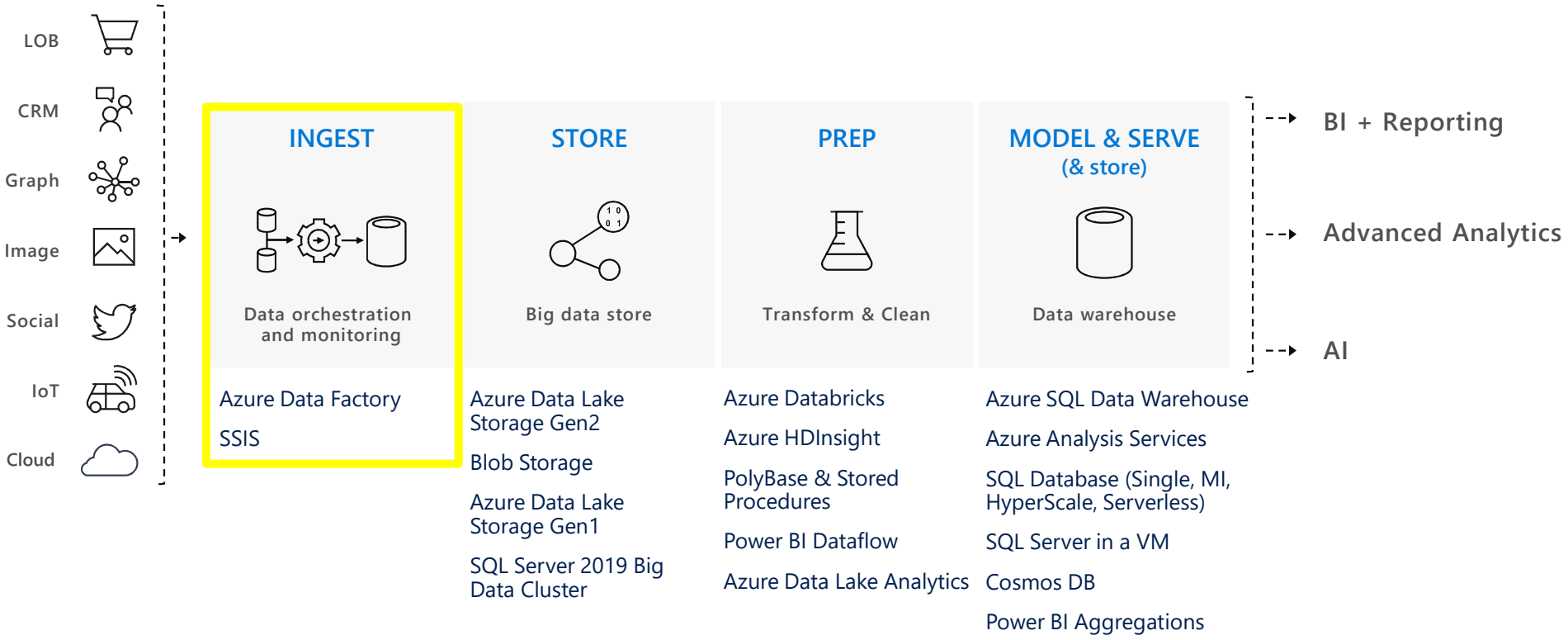
Compute Target

Images

Deployment

Data Stores

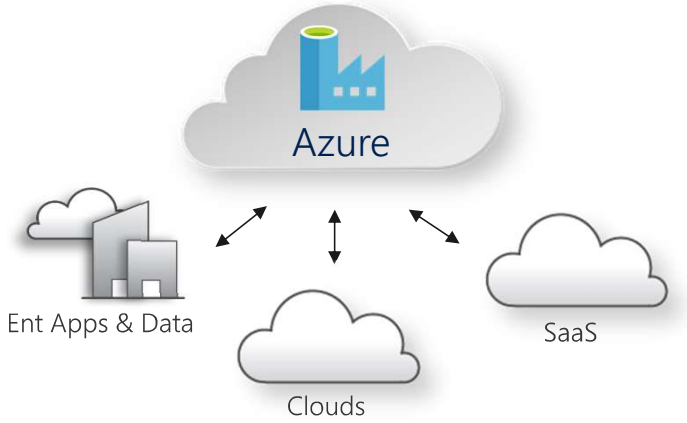
Modern Data Warehouse



Note: Those products that span more than one area are listed in there primary area

Azure Data Factory

Data Integration Service: Serverless, Scalable, Hybrid



Hybrid Pipeline Model

Seamlessly span: on prem, Azure, other clouds & SaaS
Run on-demand, scheduled, data-availability or on event

SSIS Package Execution

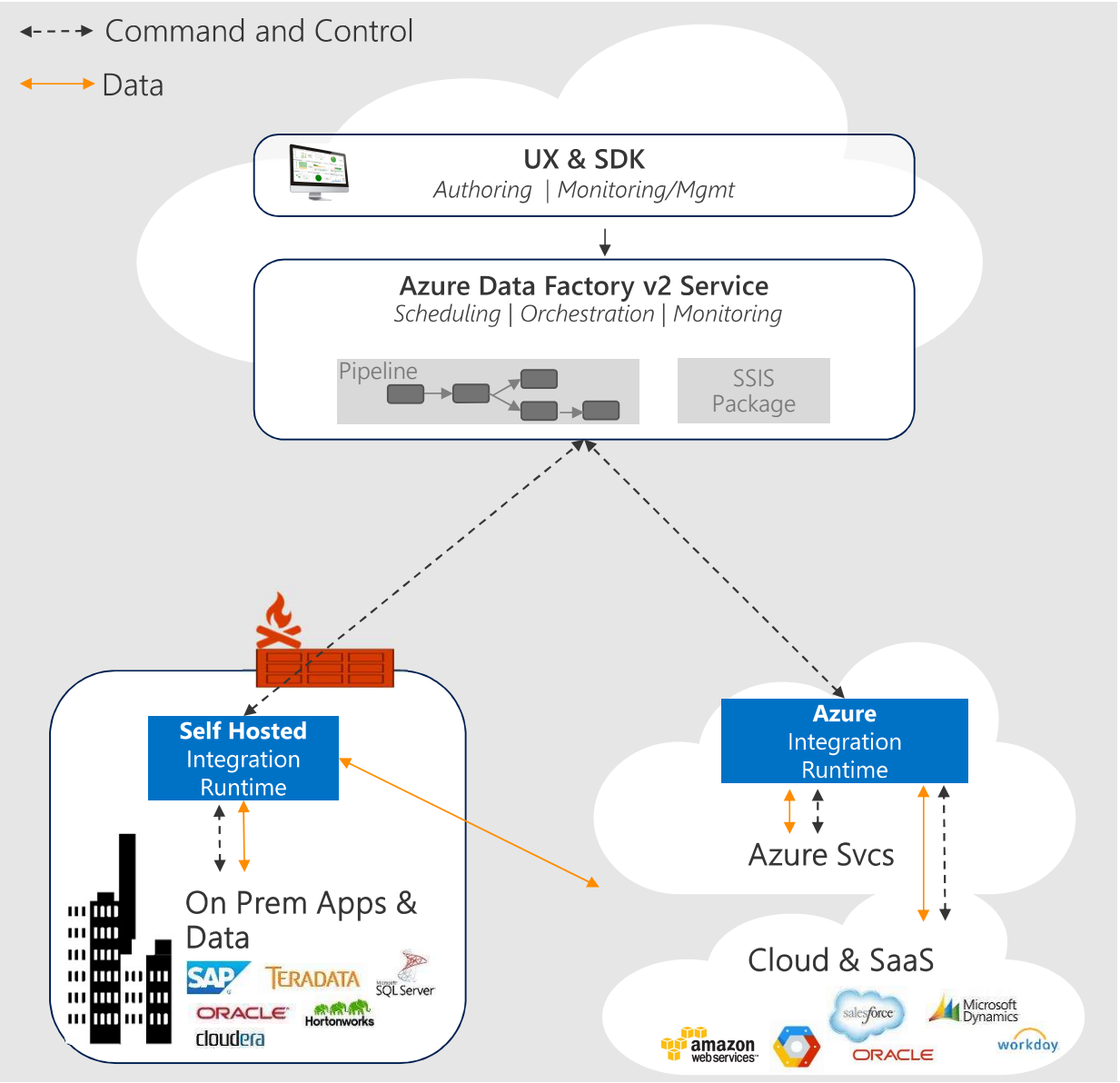
Lift existing SQL Server ETL to Azure
Use existing tools (SSMS, SSDT)

Data Movement @Scale

Cloud & Hybrid w/ 75+ connectors provided
Up to 1 GB/s

Author & Monitor

Programmability w/ multi-language SDK
Visual Tools



Data Factory

A data integration account.

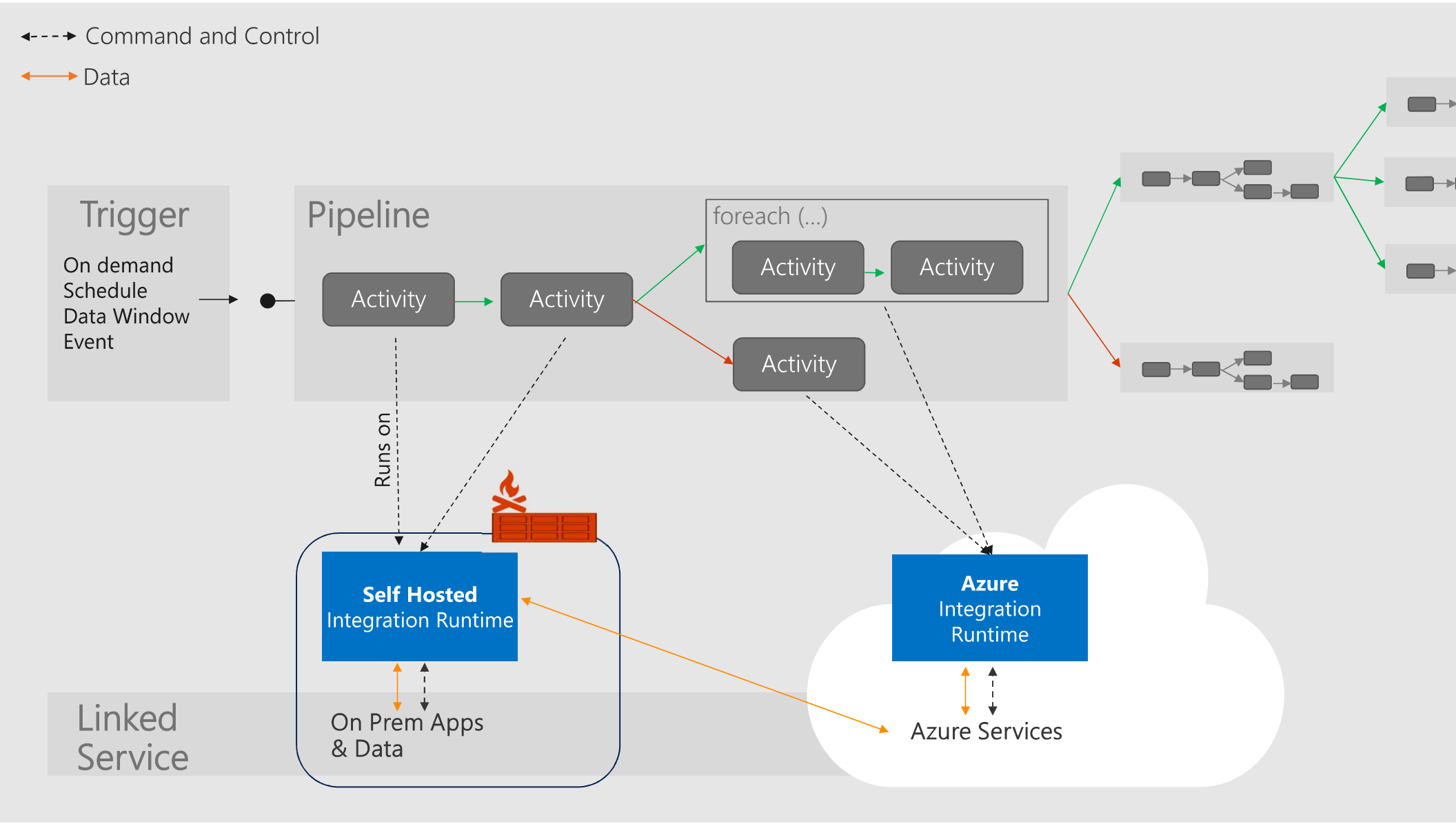
Location of orchestration, service metadata

Integration Runtime (IR)

ADF's execution engine

Three core capabilities:

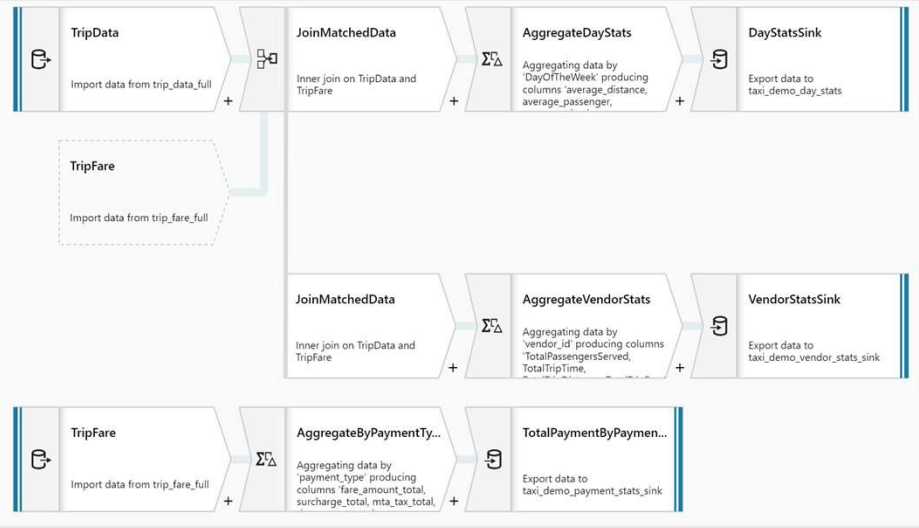
- data movement
- pipeline activity execution
- SSIS package execution


Command and Control
Data
Trigger
On demand
Schedule
Data Window
Event
Pipeline
Activity
Activity
foreach (...)
Activity
Activity
Activity
Runs on
Self Hosted
Integration Runtime
On Prem Apps
& Data
Azure
Integration
Runtime
Azure Services
Linked
Service

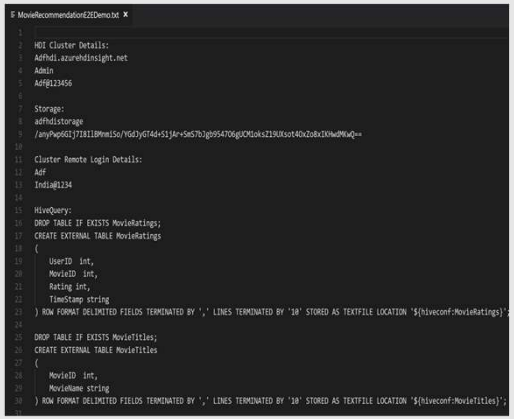
Mapping Dataflow

No Code Data Transformation @Scale

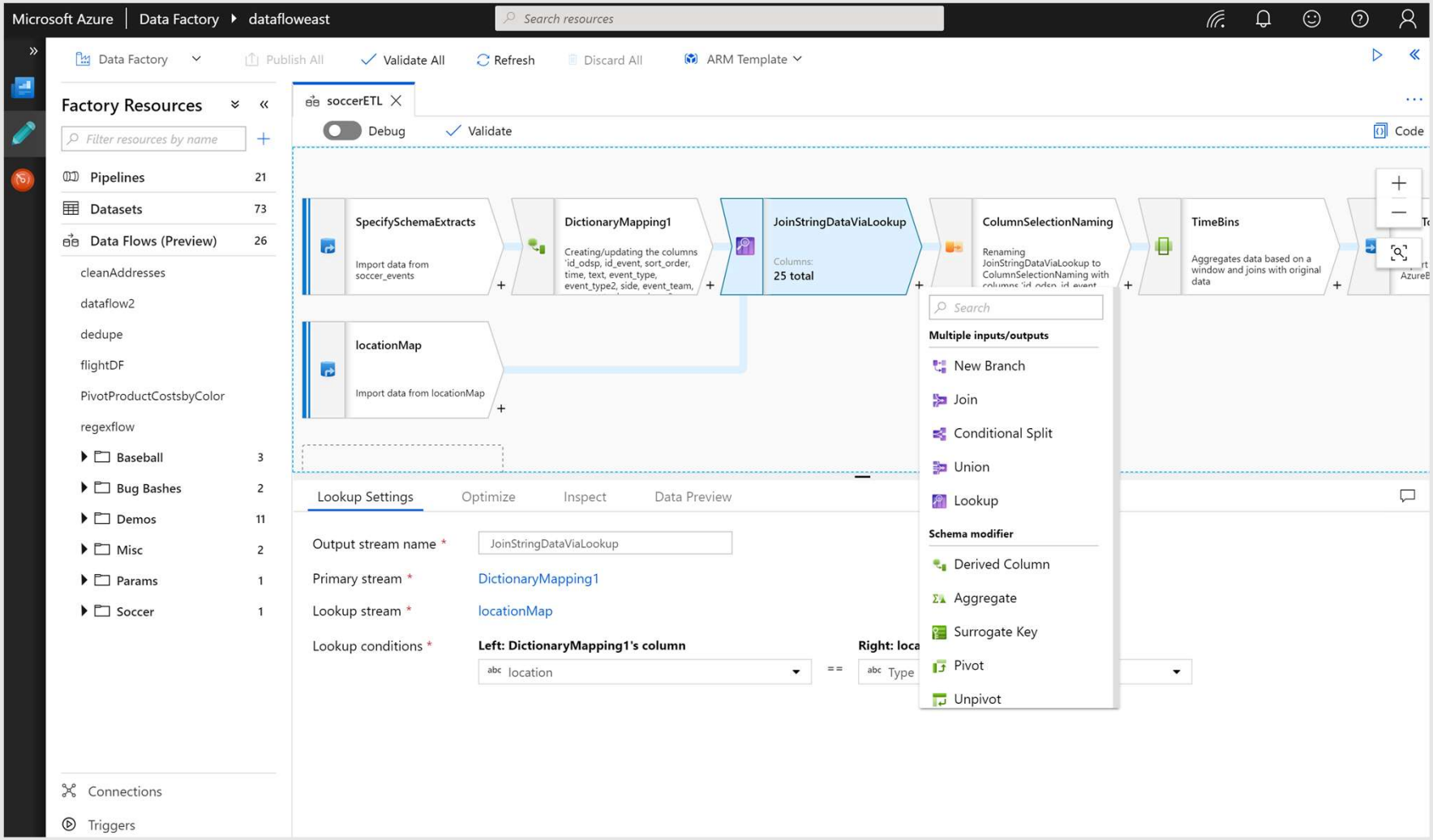
- **Data cleansing, transformation, aggregation, conversion, etc.**
- **Cloud scale via Spark execution**
- **Easily build resilient data flows**



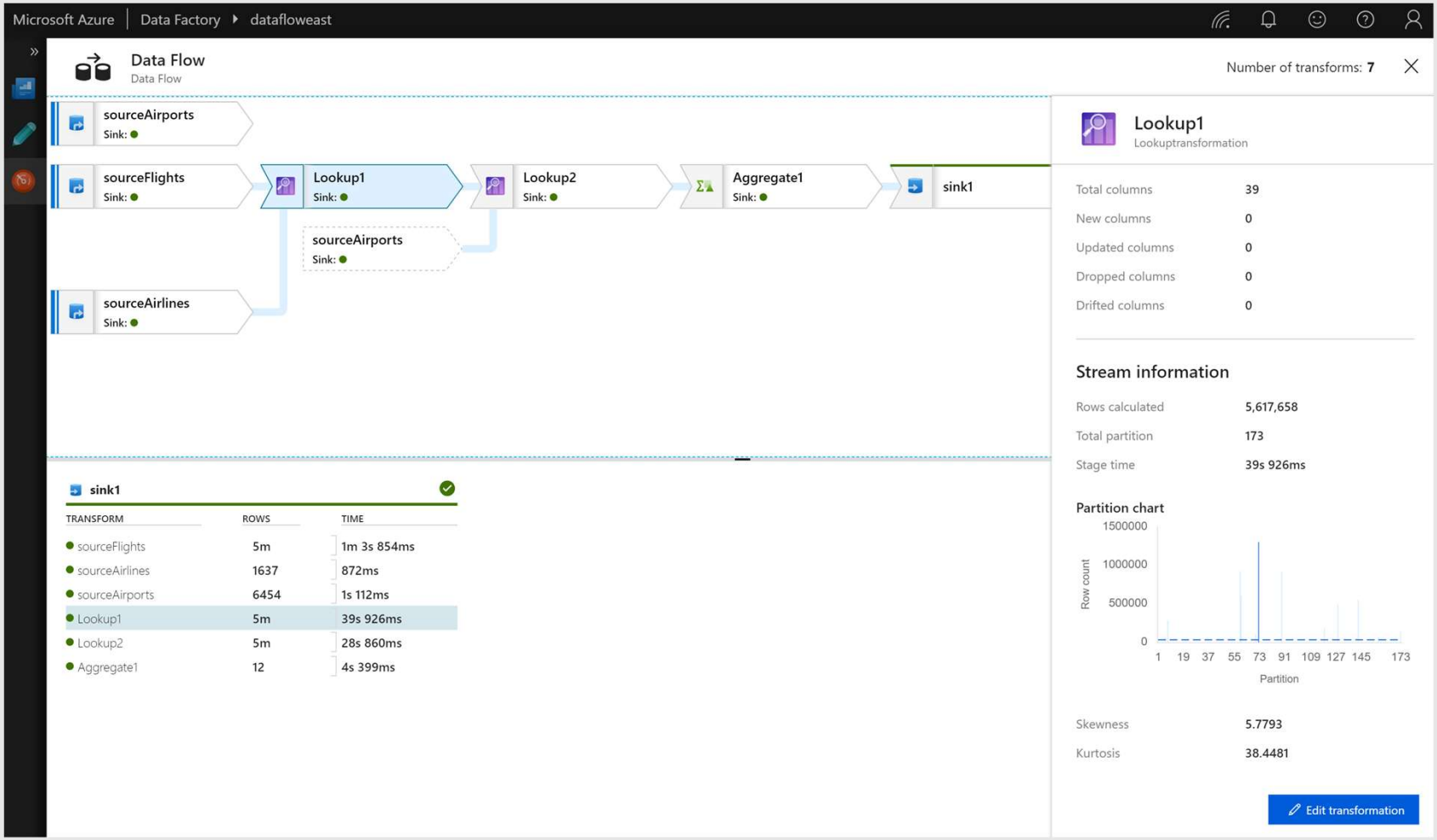
... not



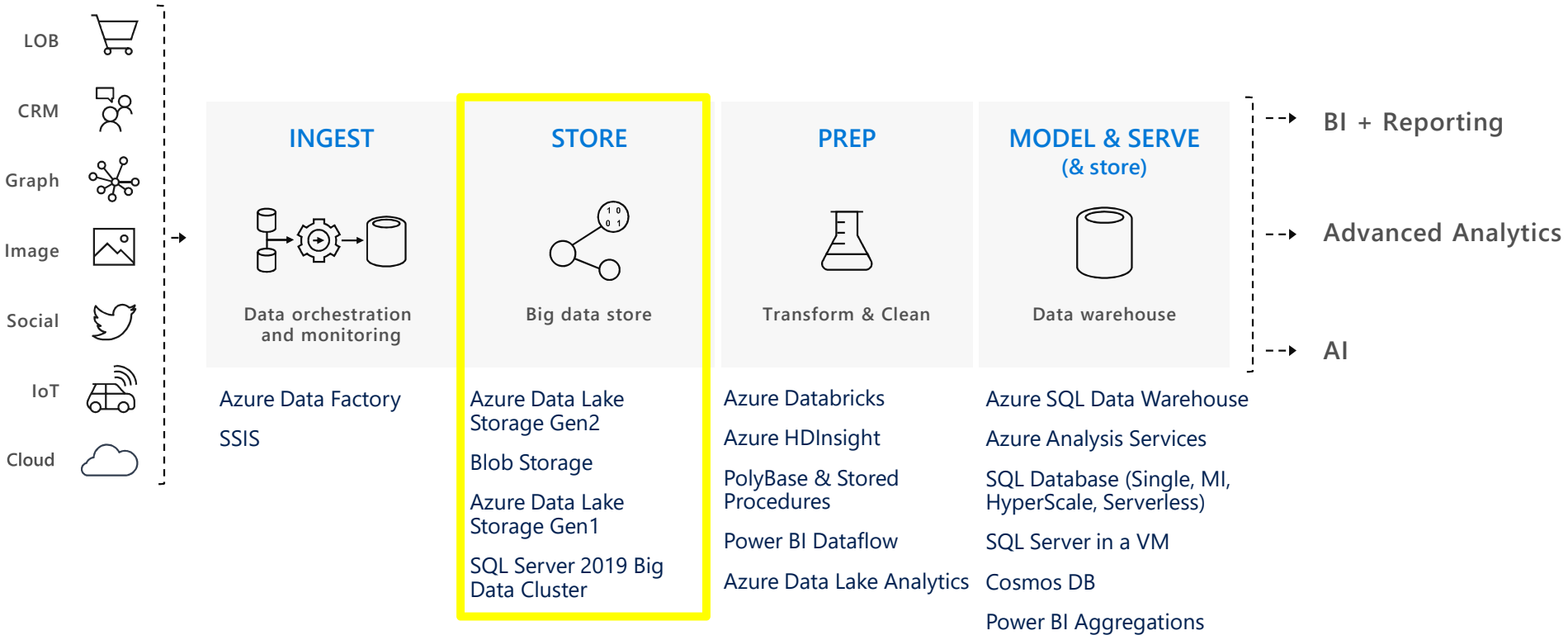
Guided experience to build data flows



Monitor & Manage Dataflows from a single pane of glass



Modern Data Warehouse



Note: Those products that span more than one area are listed in there primary area

AZURE DATA LAKE STORAGE GEN2

A "no-compromises" Data Lake: secure, performant, massively-scalable Data Lake storage that brings the cost and scale profile of object storage together with the performance and analytics feature set of data lake storage

SECURE

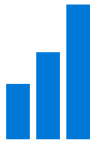
- ✓ Support for fine-grained ACLs, protecting data at the file and folder level
- ✓ Multi-layered protection via at-rest Storage Service encryption and Azure Active Directory integration

MANAGEABLE

- ✓ Automated Lifecycle Policy Management
- ✓ Object Level tiering

FAST

- ✓ Atomic file operations means jobs complete faster



SCALABLE

- ✓ No limits on data store size
- ✓ Global footprint (50 regions)

COST EFFECTIVE

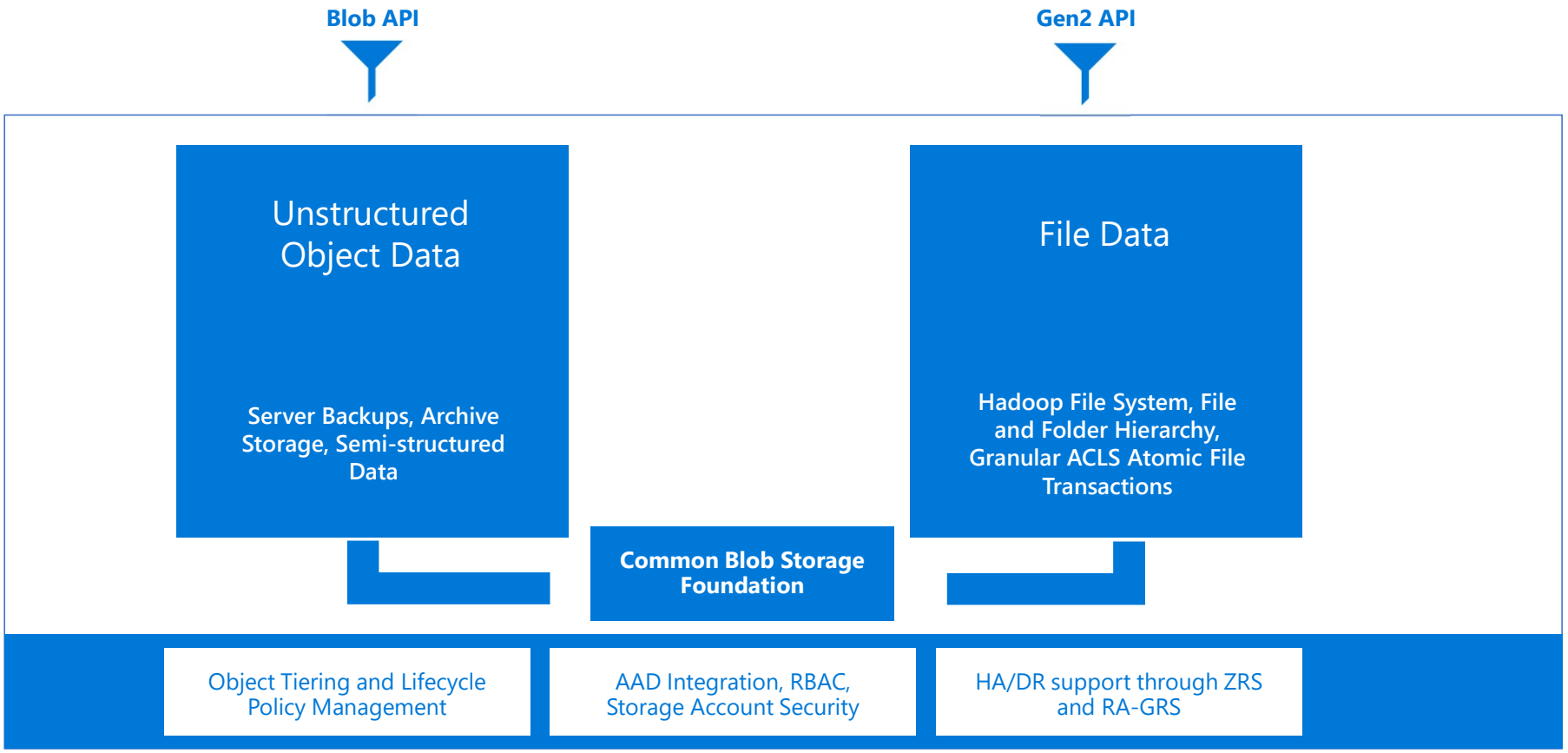
- ✓ Object store pricing levels
- ✓ File system operations minimize transactions required for job completion

INTEGRATION READY

- ✓ Optimized for Spark and Hadoop Analytic Engines
- ✓ Tightly integrated with Azure end to end analytics solutions

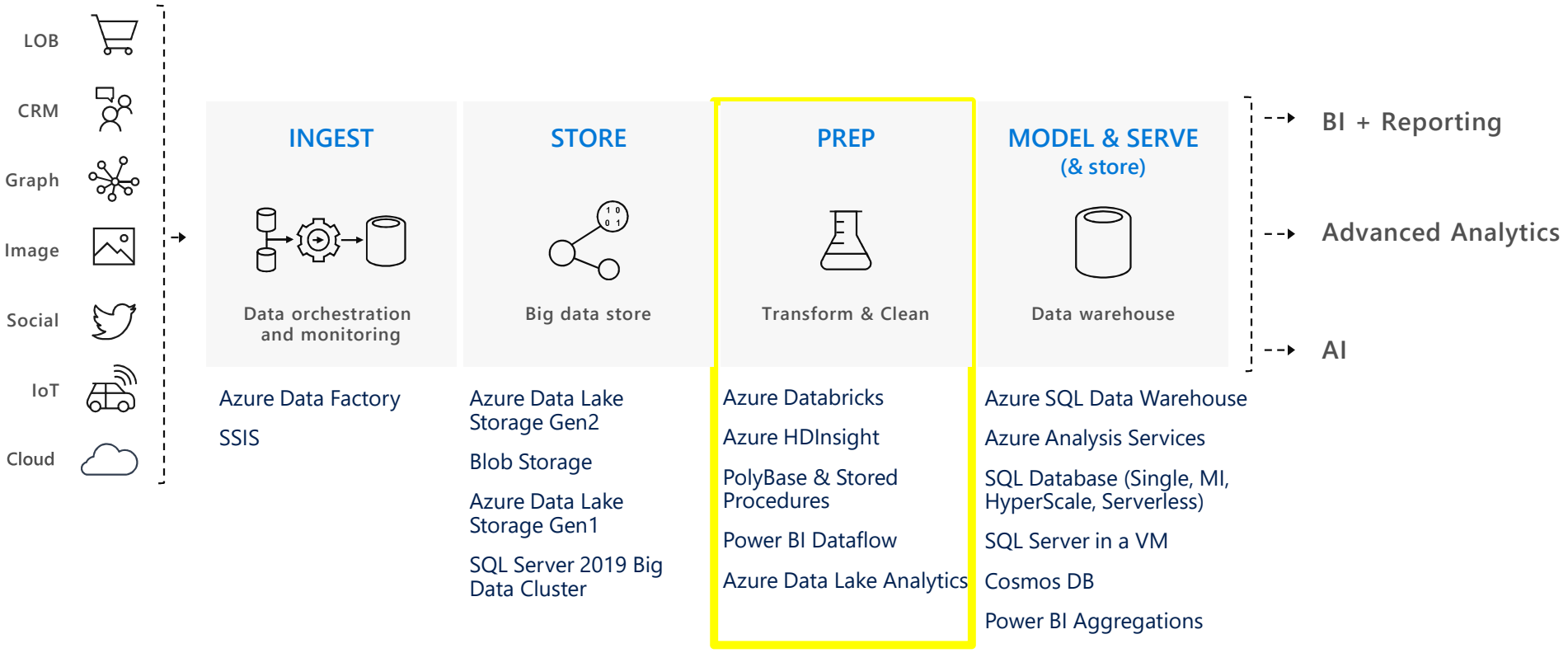
AZURE DATA LAKE STORAGE GEN2

ADLS Gen2 adds a high performance HDFS Endpoint to Azure Blob Storage and inherits the rich feature set of Azure Blob Storage *



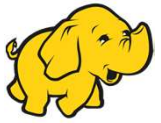
** Full Blob and HNS Interoperability available in post-GA*

Modern Data Warehouse



Note: Those products that span more than one area are listed in there primary area

Azure HDInsight



A Cloud Spark and Hadoop service for the Enterprise

Reliable
with an industry leading SLA

Enterprise-grade
security and monitoring

Productive platform
for developers and scientists

Easy
for administrators to manage

Cost effective
cloud scale

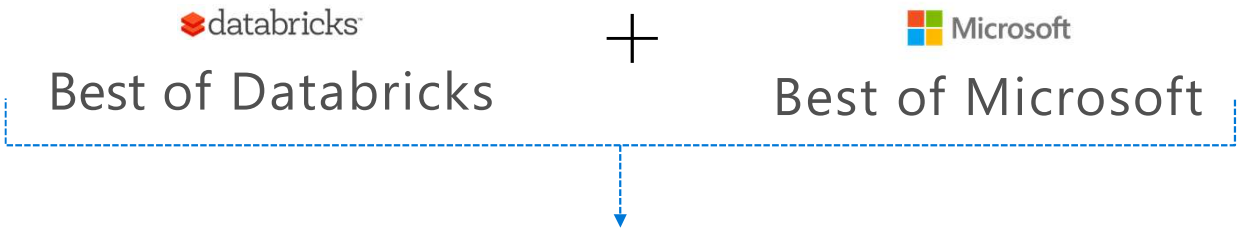
Integration
with leading ISV applications

63% lower TCO
than deploy your own Hadoop on-premises*

*IDC study "The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight"

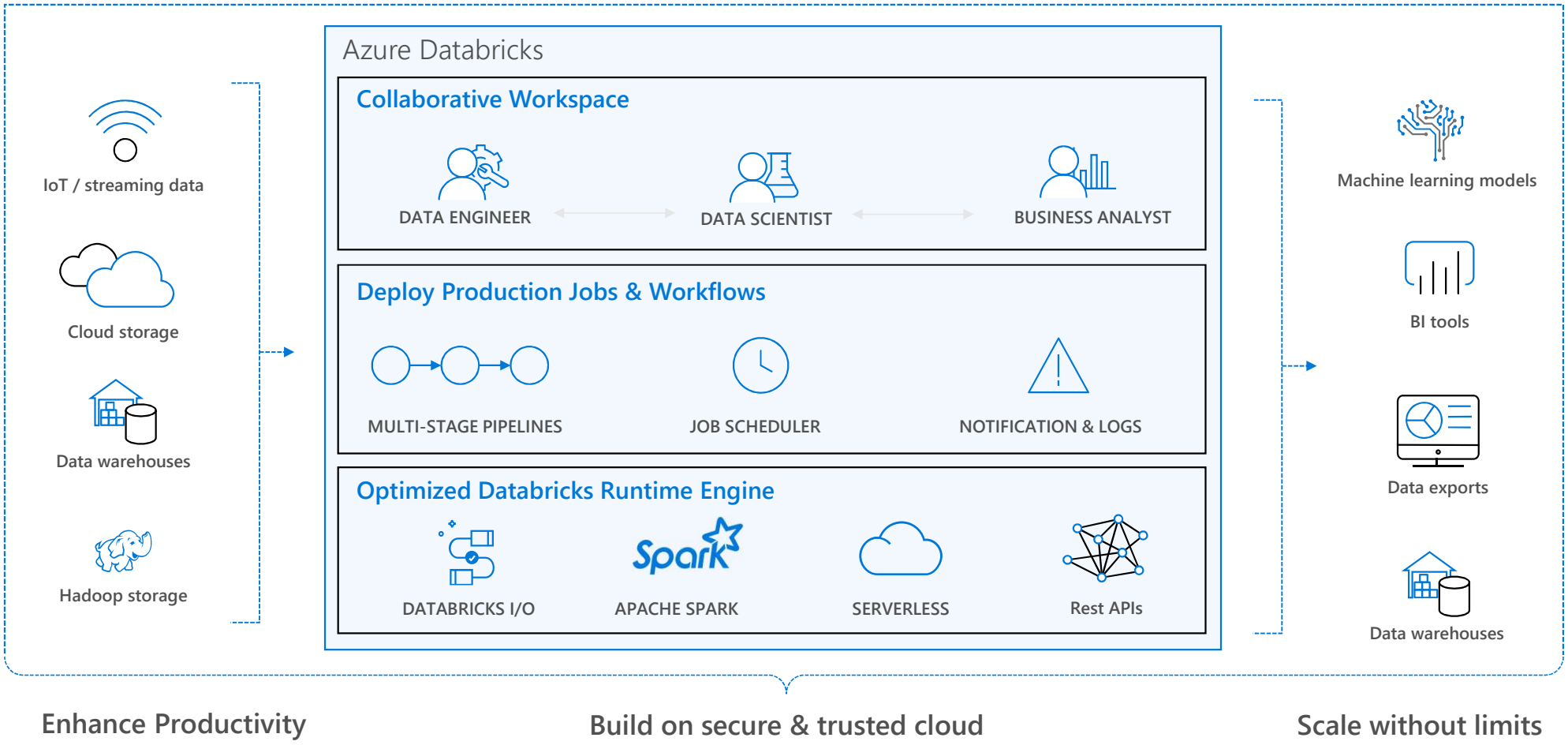
INTRODUCING, AZURE DATABRICKS

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure



- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

AZURE DATABRICKS

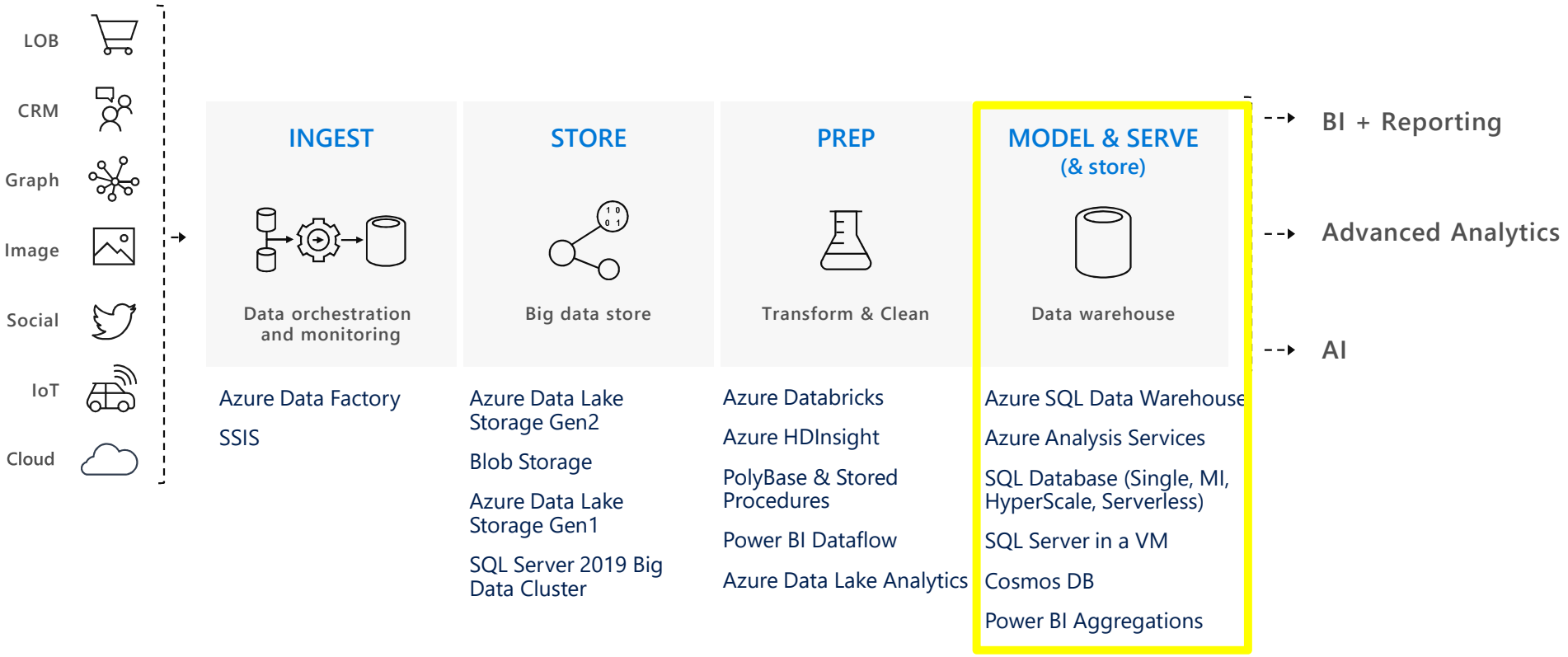


Enhance Productivity

Build on secure & trusted cloud

Scale without limits

Modern Data Warehouse



Note: Those products that span more than one area are listed in there primary area

AZURE SQL DATA WAREHOUSE

INDUSTRY LEADING PERFORMANCE & SECURITY AT SCALE IN THE CLOUD

Unrivaled power

Up to 10x more query performance than traditional storage

Unlimited scale

Independent and limitless scaling of storage and compute

Fast time to value

Seamless integration with Microsoft and 3rd party services

Trusted & reliable

Enhanced security with encryption, audit, VNET and leading compliance

Azure SQL Data Warehouse

Best in class price-performance

Up to 94% less expensive than competitors

Industry-leading security

Defense-in-depth security and 99.9% financially backed availability SLA

Intelligent workload management

Separation of compute and storage

Prioritize resources for the most valuable workloads

Data flexibility

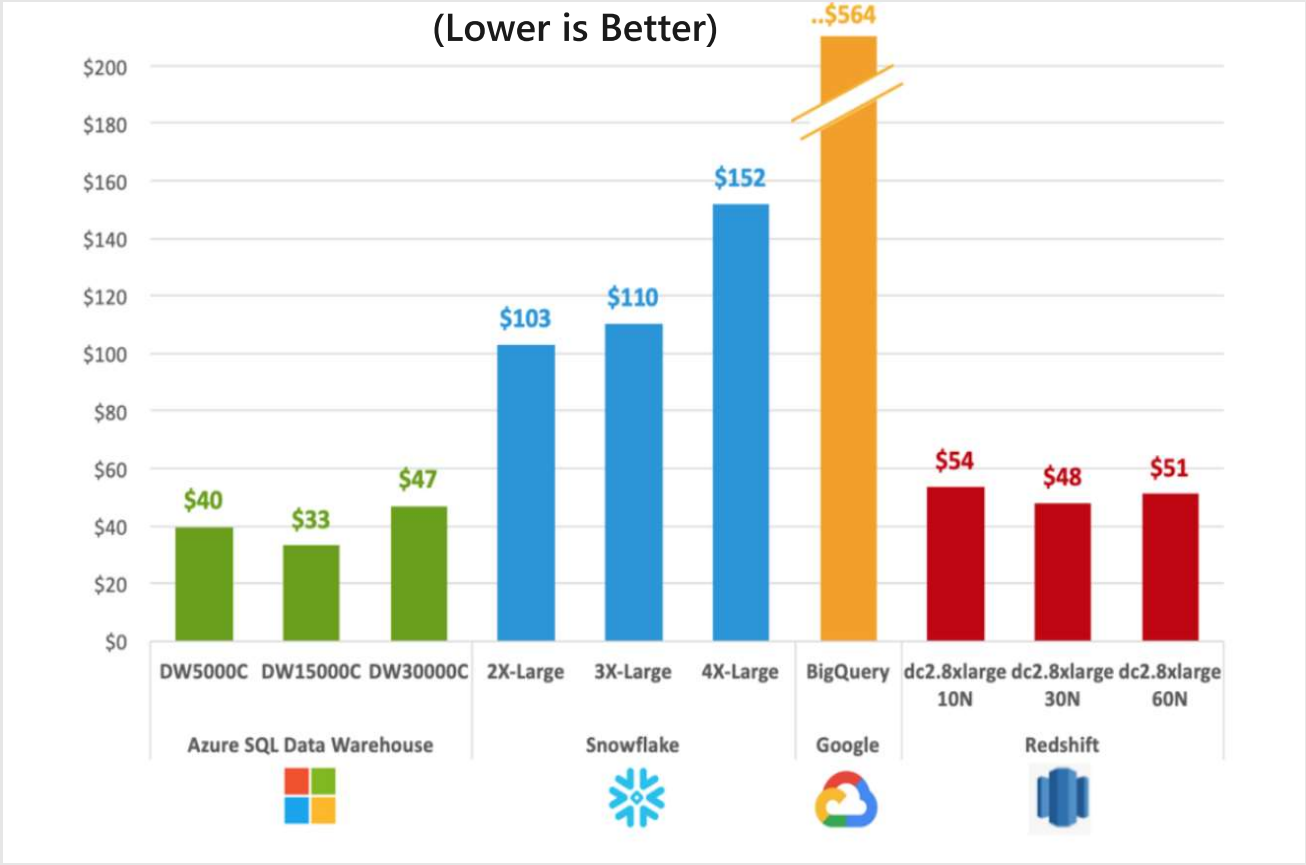
Query directly over the Data Lake

Support for structured and semi-structured data

Developer productivity

Enterprise class application lifecycle management

Industry-leading price performance



Concurrency limits

Overview

Concurrent query limit determined by SLO.

Each SLO has a set amount of concurrency slots.

Resource classes assign concurrency slots to queries.

Benefits

Fine-grained control over resource usage

Flexible assignment with dynamic and static classes

Assign more resources to compute-heavy jobs

Example Concurrency at DW1000c (32 slots)

32 *smallrc* queries (1 slot each)

16 *staticrc20* queries (2 slots each)

8 *mediumrc* queries concurrently (4 slots each)

2 *staticrc50* queries concurrently (16 slots each)

1 *staticrc50* + 2 *mediumrc* + 2 *staticrc20* + 4 *smallrc* queries (32 total)

Azure SQL Data Warehouse SLOs and Concurrency Slots

Service Level	Max. Concurrent Queries	Max. Concurrency Slots
DW100c	4	4
DW500c	20	20
DW1000c	32	40
DW3000c	64	120
DW7500c	128	300
DW10000c	128	400
DW15000c	128	600
DW30000c	128	1200

Source: https://docs.microsoft.com/en-us/azure/sql-data-warehouse/memory-and-concurrency-limits

Azure Analysis Services

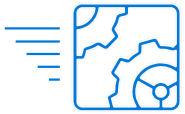
Enterprise grade analytics engine as a service

Build rich semantic models

Transform complex data into business user friendly semantic models

Proven technology

Based on powerful, proven SQL Server Analysis Services



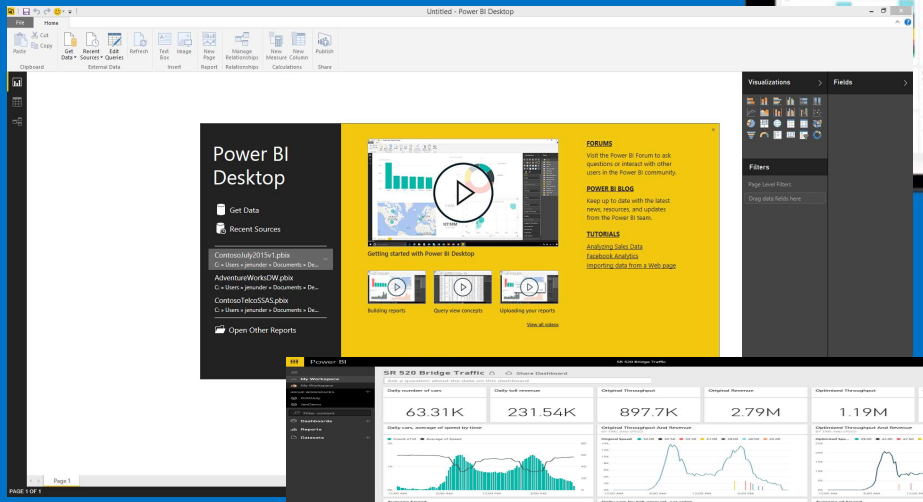
Gain insights at the speed of thought

Gain instant insights with in-memory cache using your preferred visualization tools

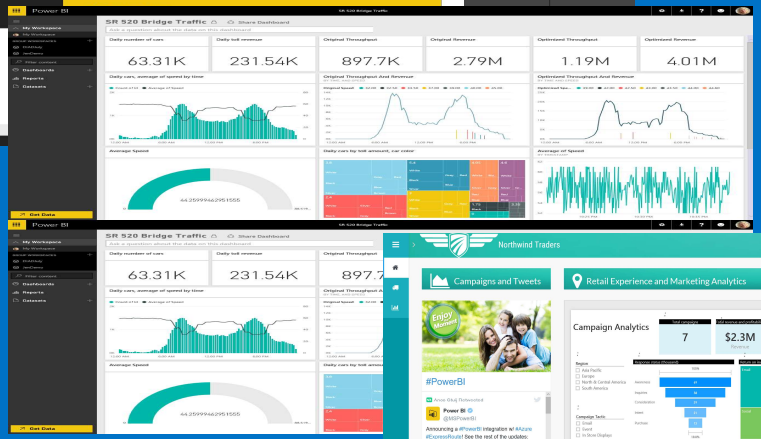
Provision and scale with ease

Easy to deploy, scale, and manage as platform-as-a-service

Power BI

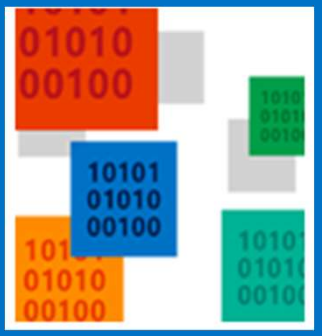


Power BI Desktop

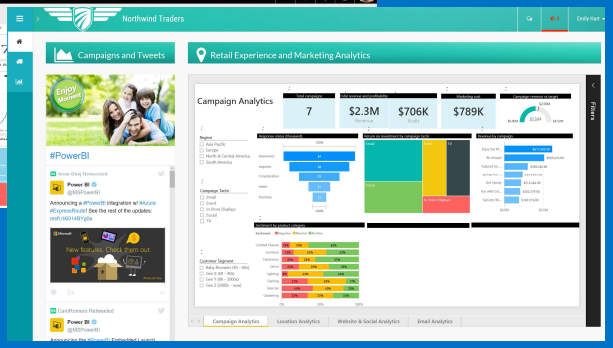


PowerBI.com/Power BI Premium

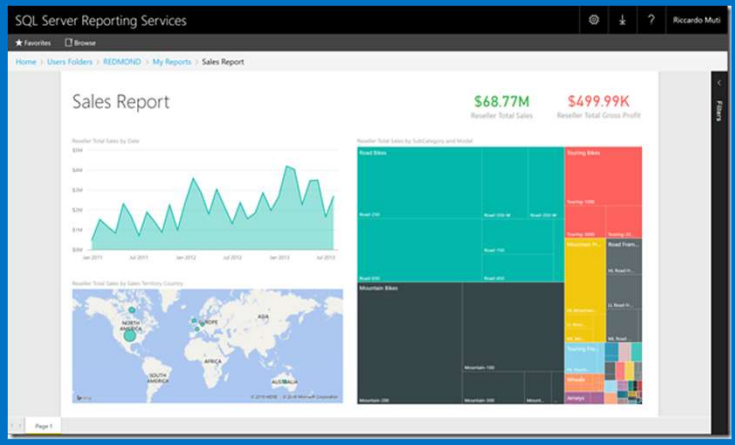
Mobile Apps for iOS, Android and Windows



Developer APIs

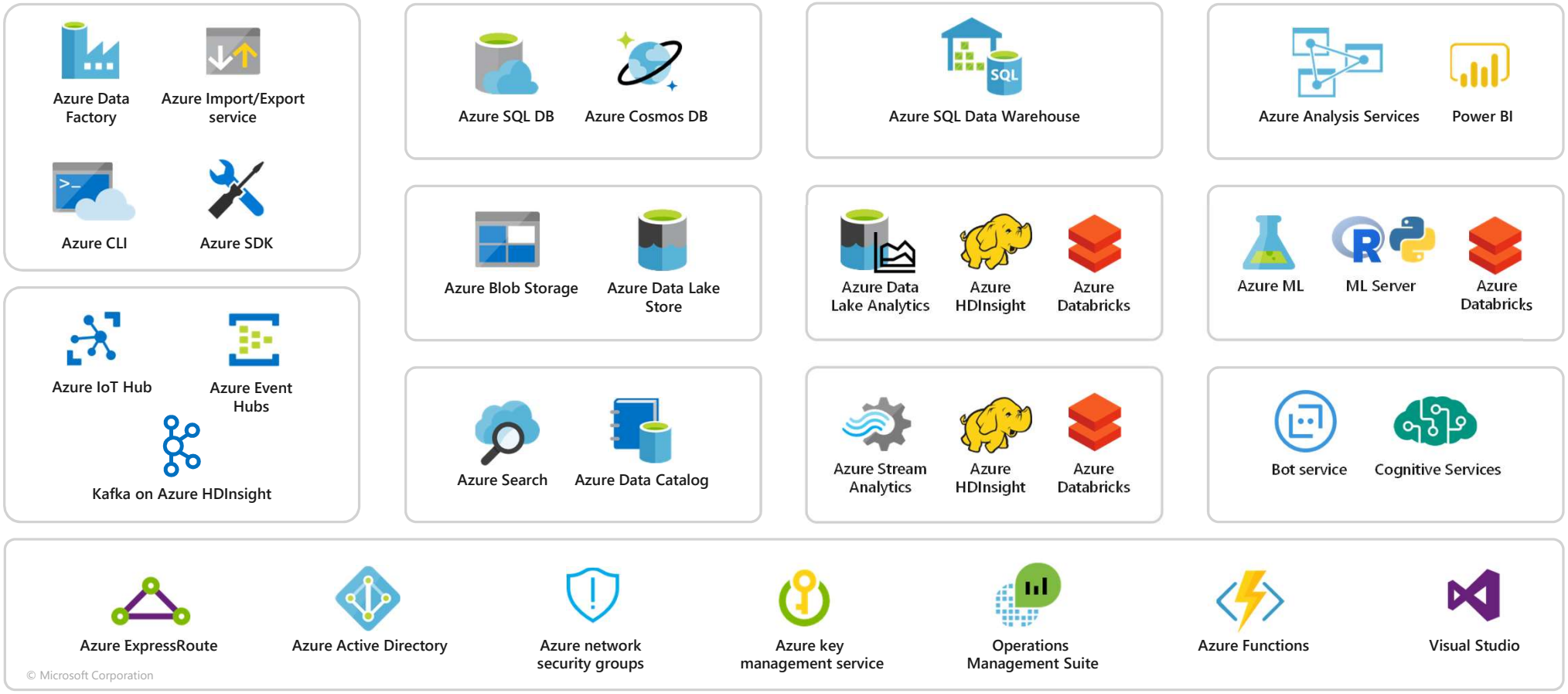


Power BI Premium Embedded

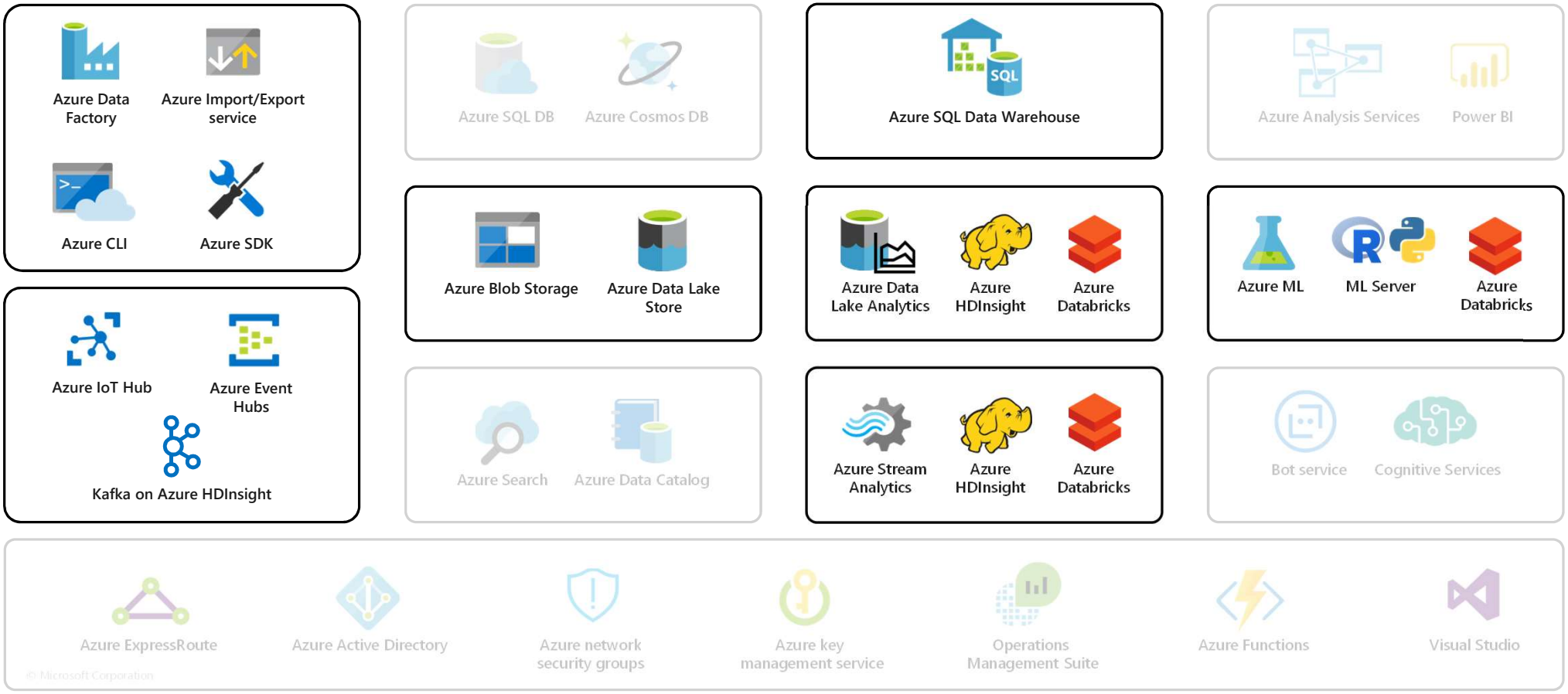


Power BI Report Server

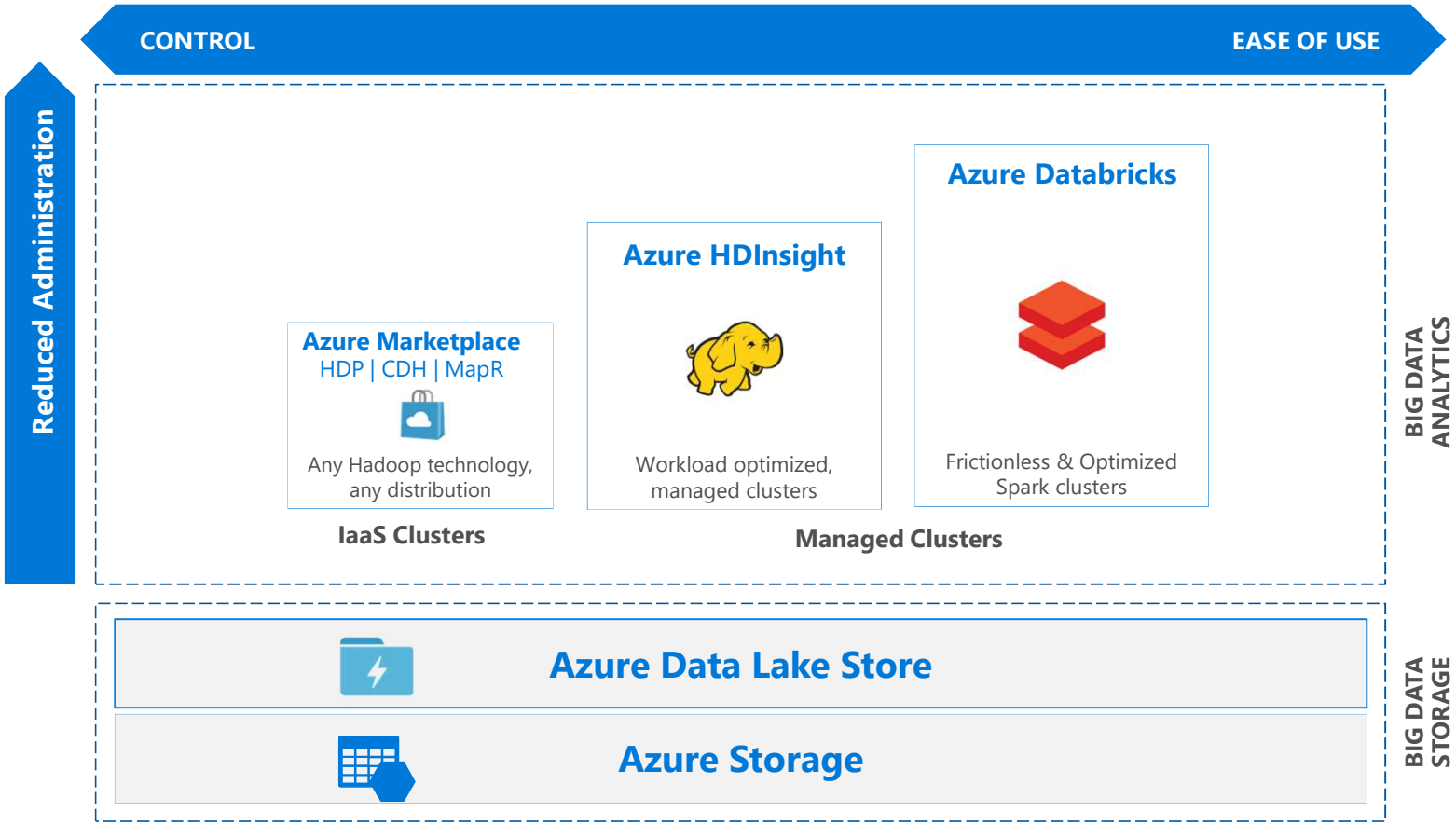
The Azure data landscape



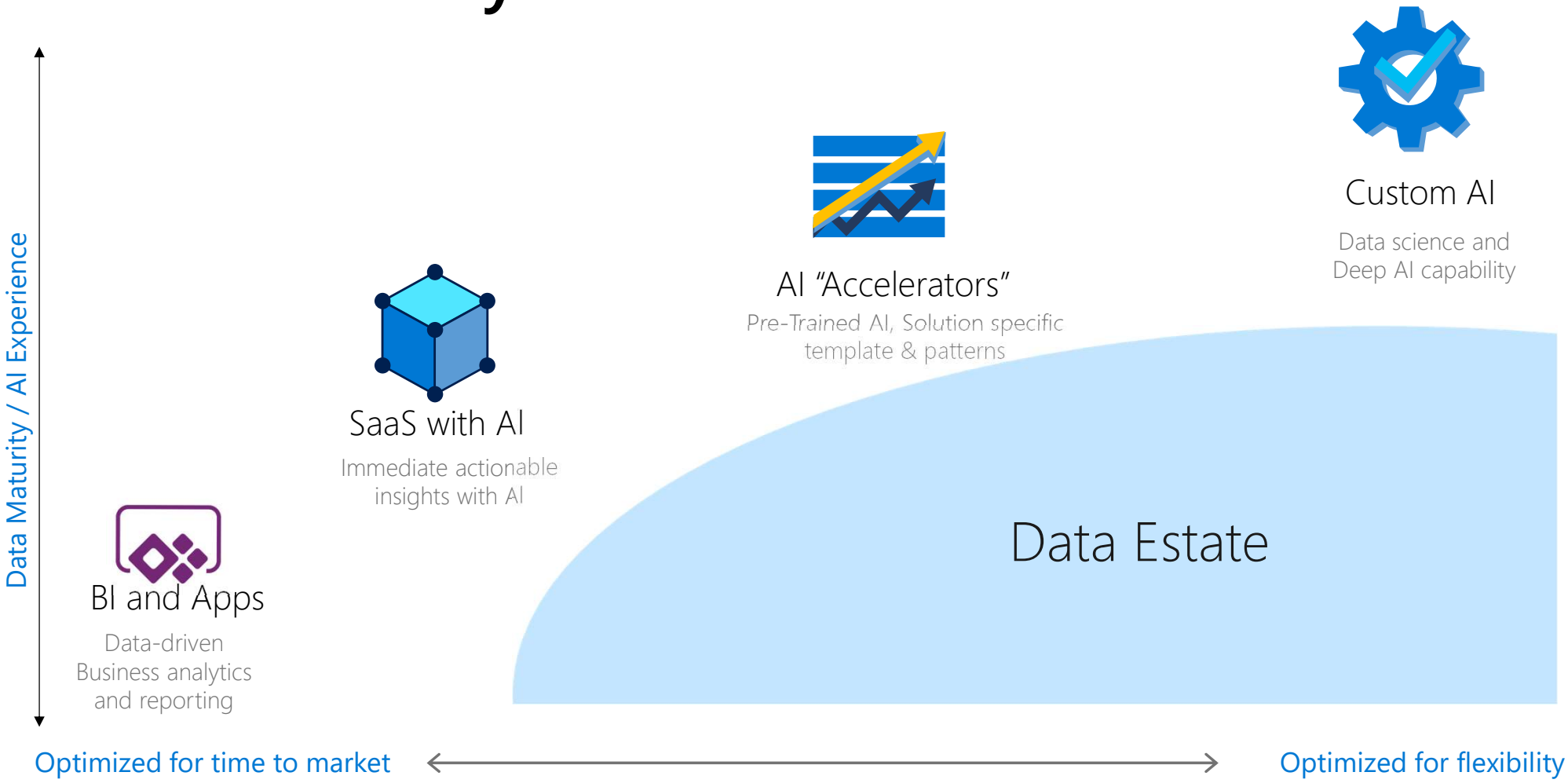
The Azure Big Data landscape



KNOWING THE VARIOUS BIG DATA SOLUTIONS



The AI Journey



Azure AI

Pre-built AI/Conversational AI
AI apps & agents

Azure Bot Service
Azure Cognitive Services

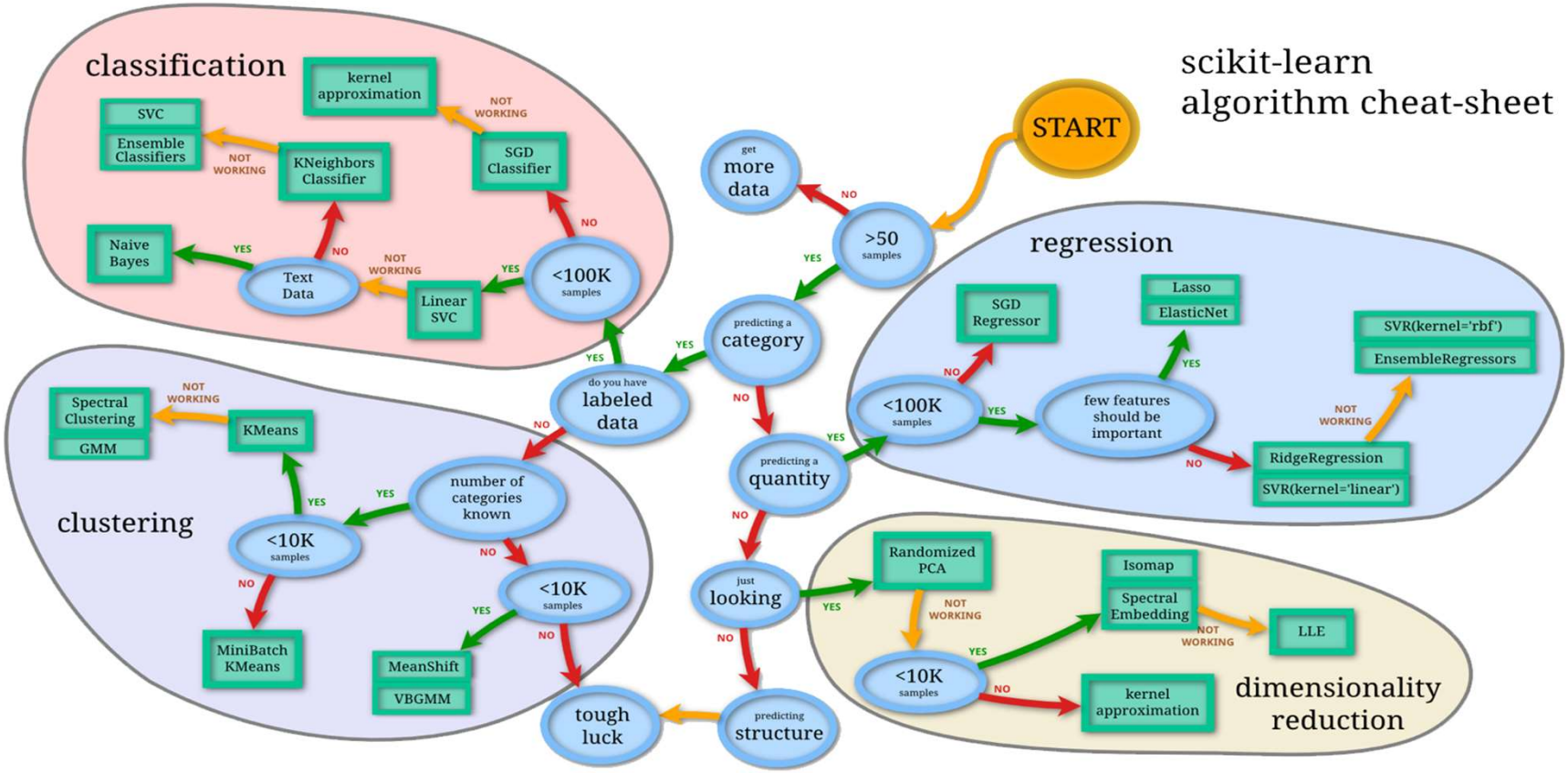
Knowledge mining

Azure Cognitive Search

Custom AI

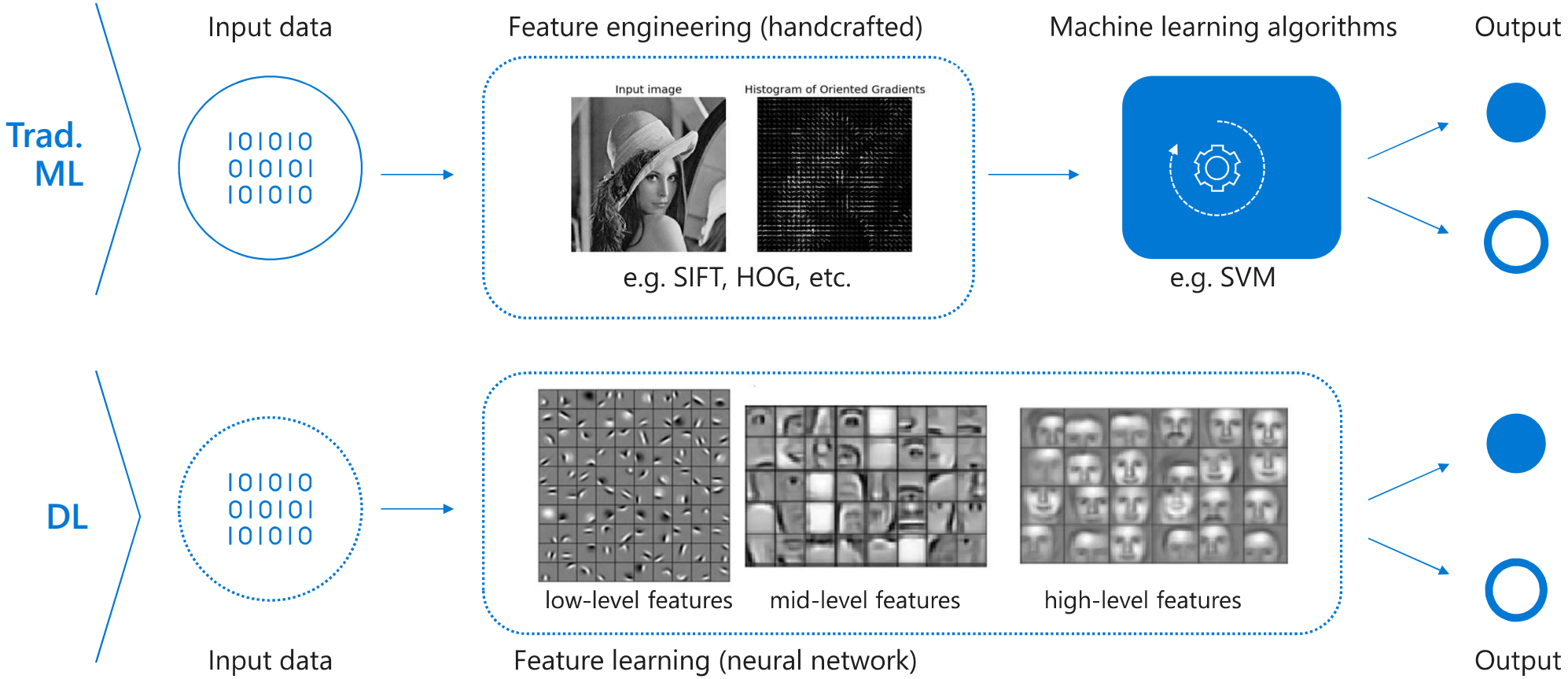
Azure Databricks
Azure Machine Learning
Azure AI infrastructure

Traditional Machine Learning



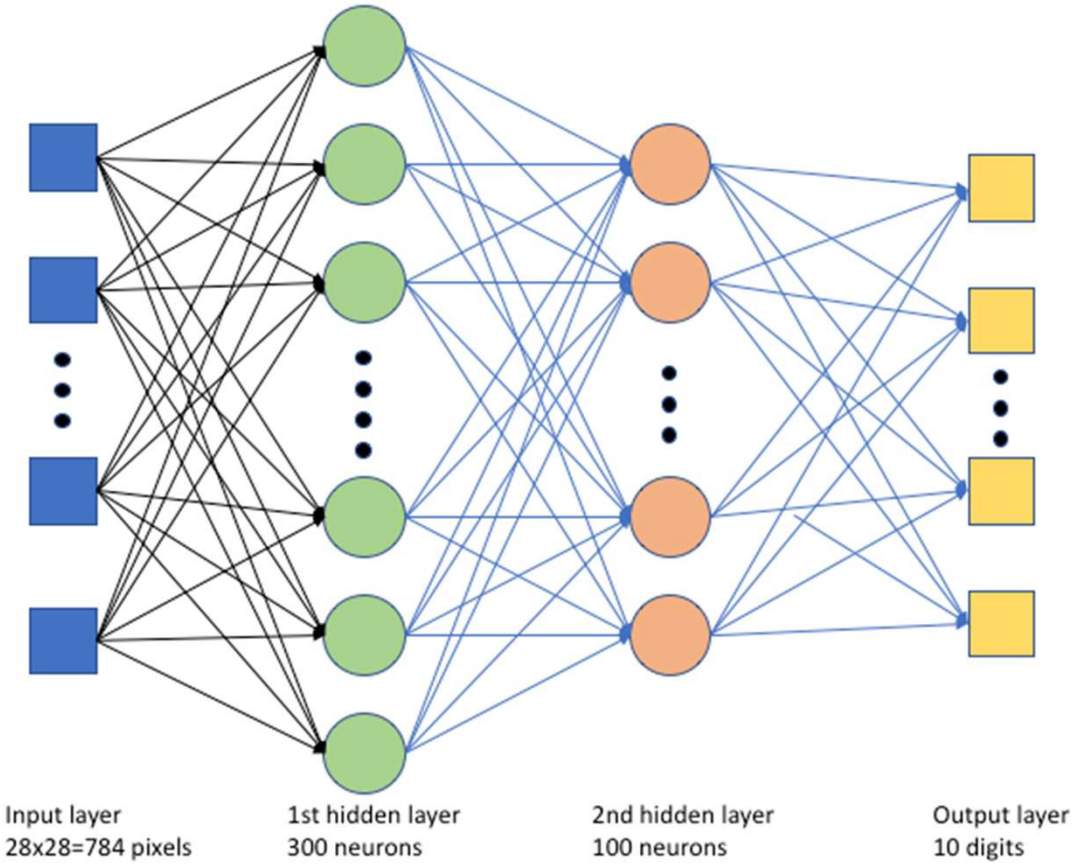
Source: http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Traditional ML versus DL



Top figure source;
Bottom figure from NVIDIA

What is a Artificial Neural Network?



Required Deep Learning Frameworks

Popular Deep Learning Frameworks


TensorFlow


PyTorch


Scikit-Learn


MXNet


Chainer


Keras


ONNX

Community project created by Facebook and Microsoft

Use the best tool for the job. Train in one framework and transfer to another for inference

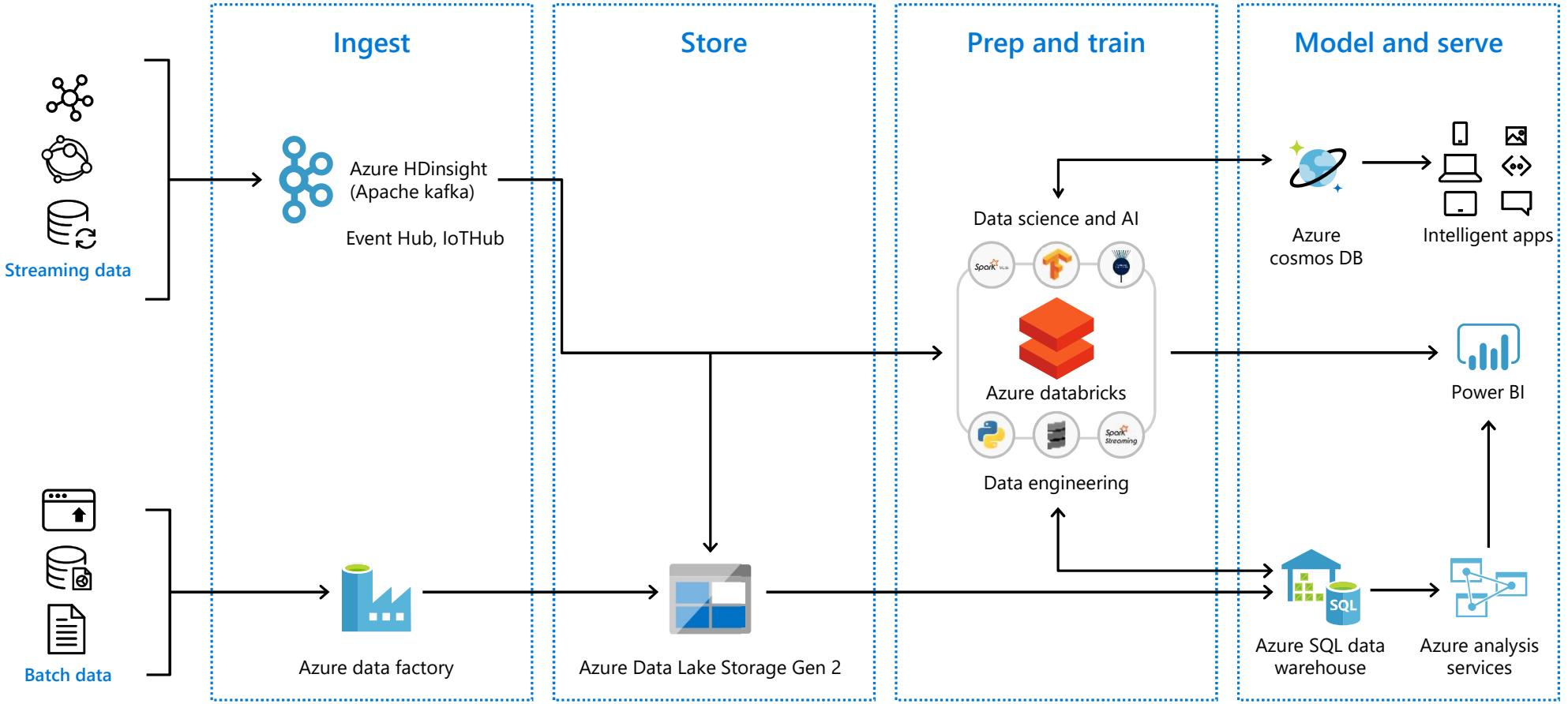









Microsoft has a recommended reference architecture



End to End Custom AI

Typical E2E Process