Microsoft Azure
TLV Cloud Workshops

18.11.2018 | Hilton Tel-Aviv

Microsoft

Cloudera on Azure

Oshik Avioz
Cloud Solution Architect | Data & AI
osavioz@microsoft.com
@OshikAvioz
www.linkedin.com/oshikavioz
https://github.com/oavioz

Session objectives and takeaways

❑ Cloudera on Azure – Decouple data bus

❑ Azure Data Lake Store - ADLS

❑ Why Cloudera on ADLS

❑ How/Where to use Cloudera on Azure

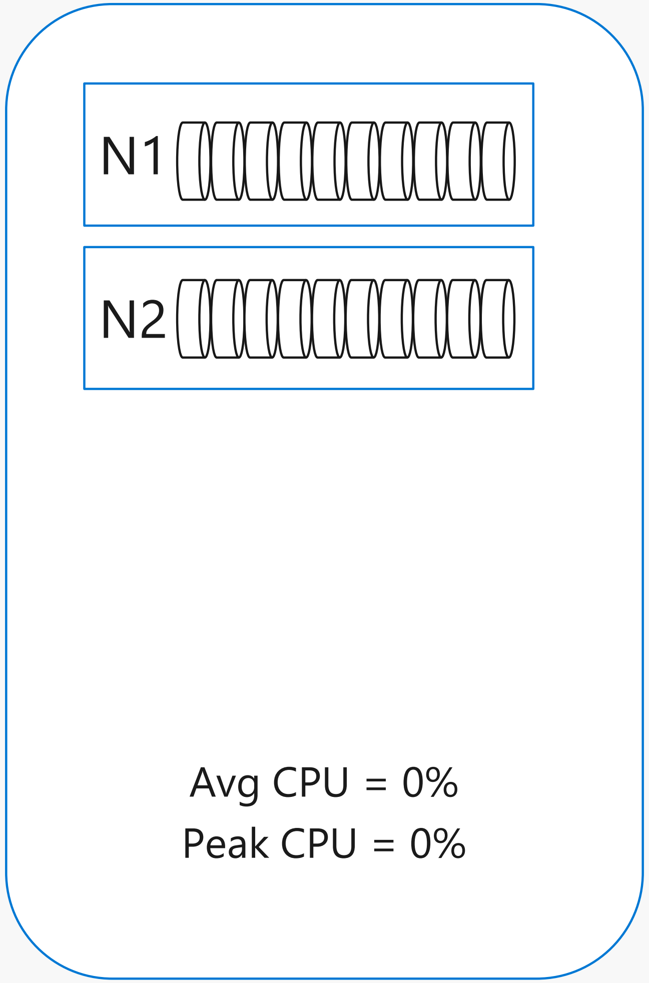
❑ Demo

Microsoft

Insert session code here

Cloudera on Azure – Decouple data bus

Traditional Hadoop



Scope out a cluster size

Build it

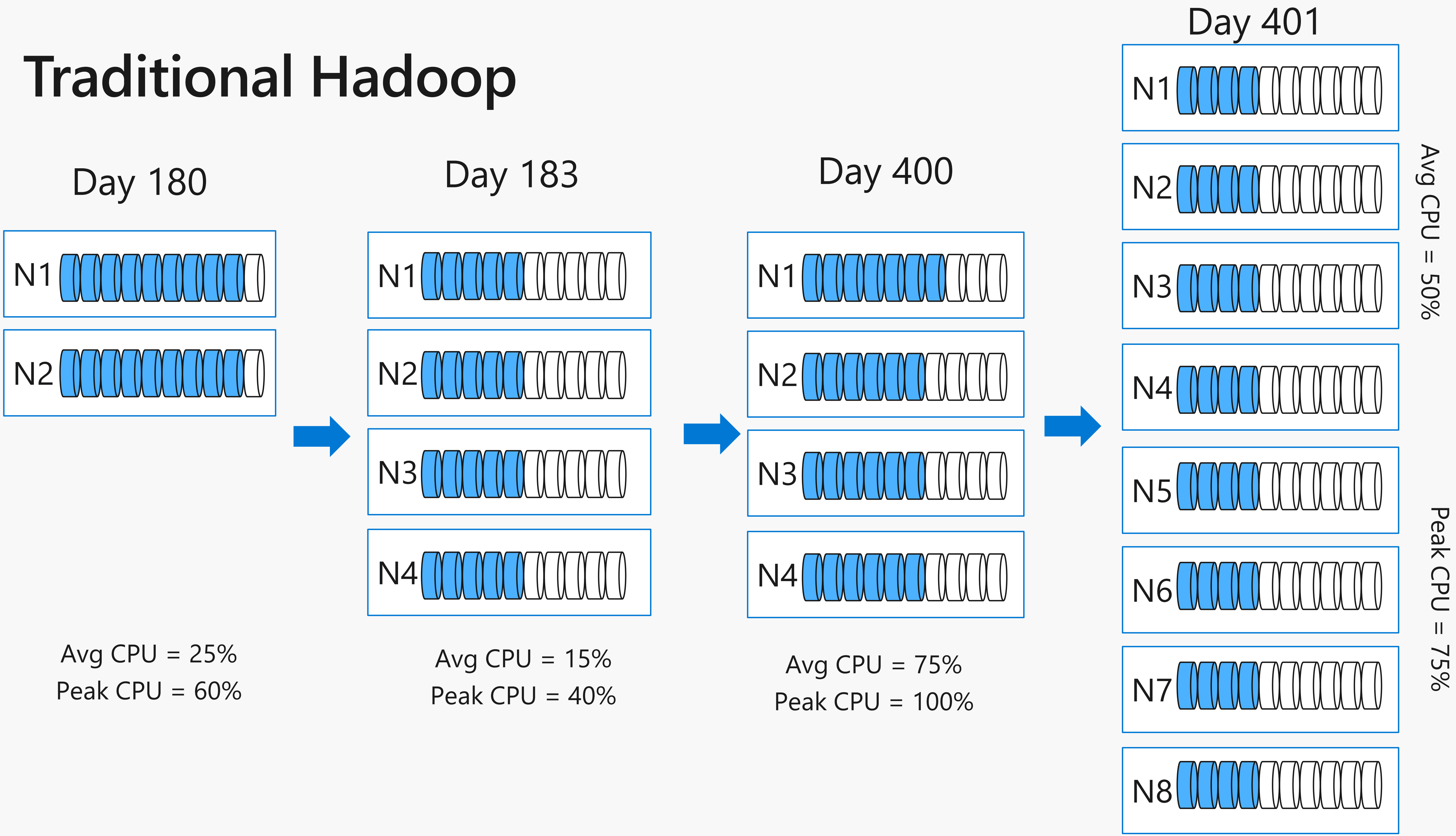
Configure it

Administer it

Start adding data

Start developing queries

Traditional Hadoop



Day 180

N1

N2

Avg CPU = 25%
Peak CPU = 60%

Day 183

N1

N2

N3

N4

Avg CPU = 15%
Peak CPU = 40%

Day 400

N1

N2

N3

N4

Avg CPU = 75%
Peak CPU = 100%

Day 401

N1

N2

N3

N4

N5

N6

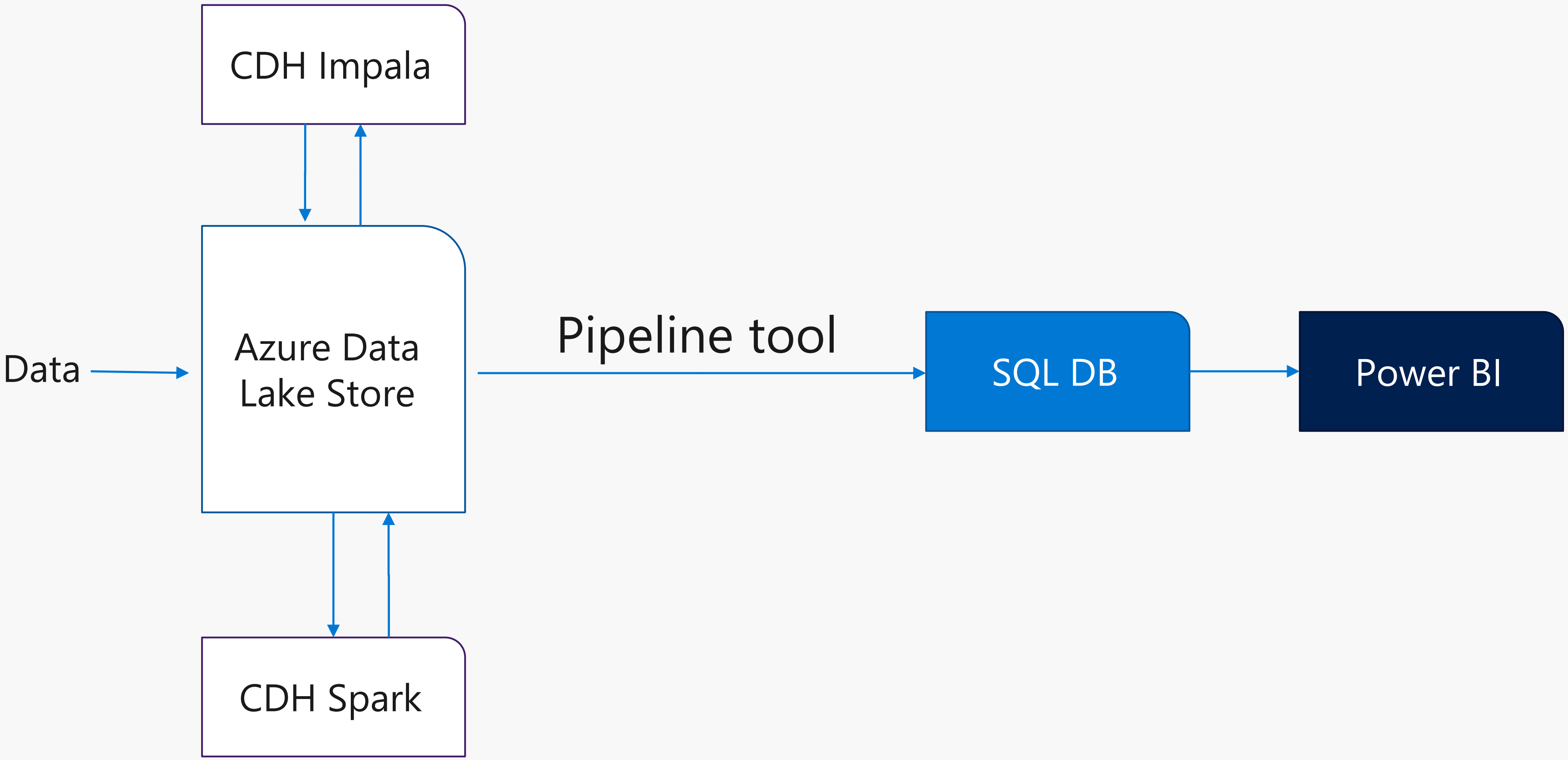
N7

N8

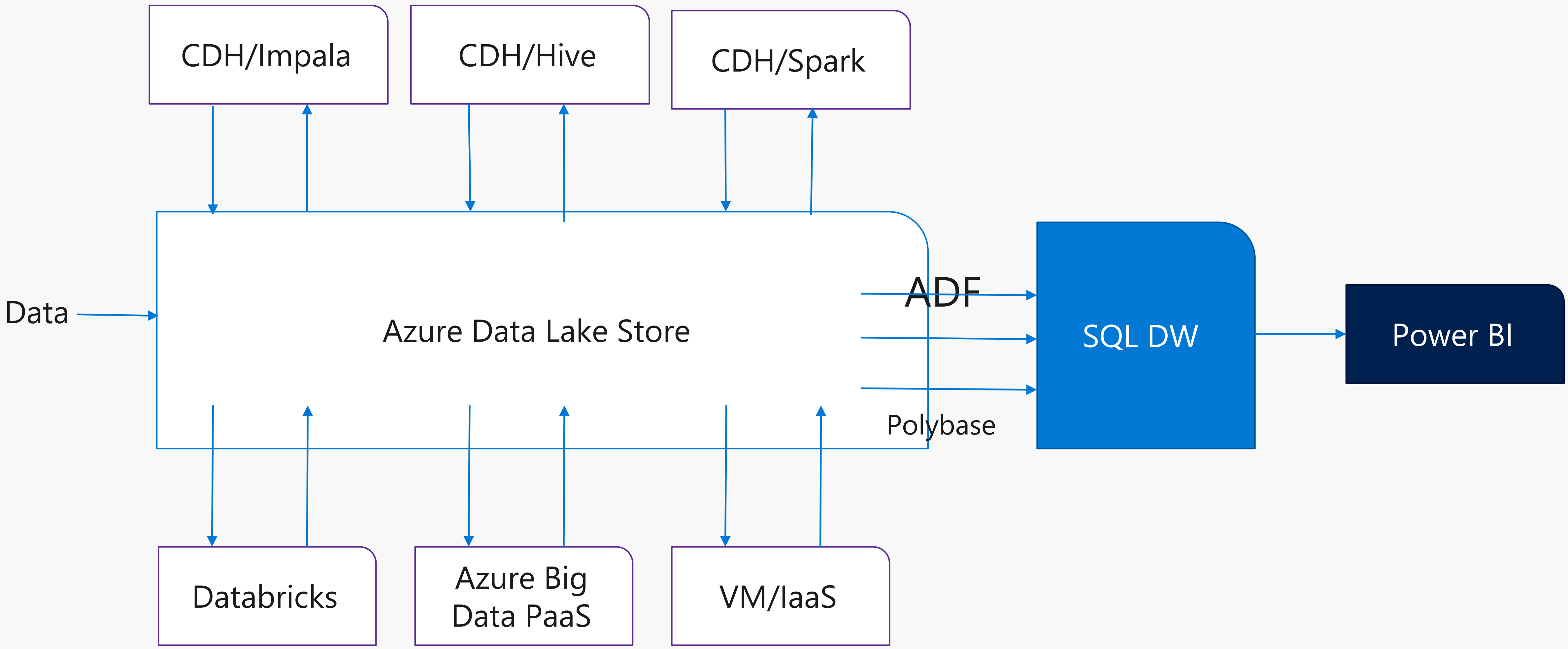
Avg CPU = 50%

Peak CPU = 75%

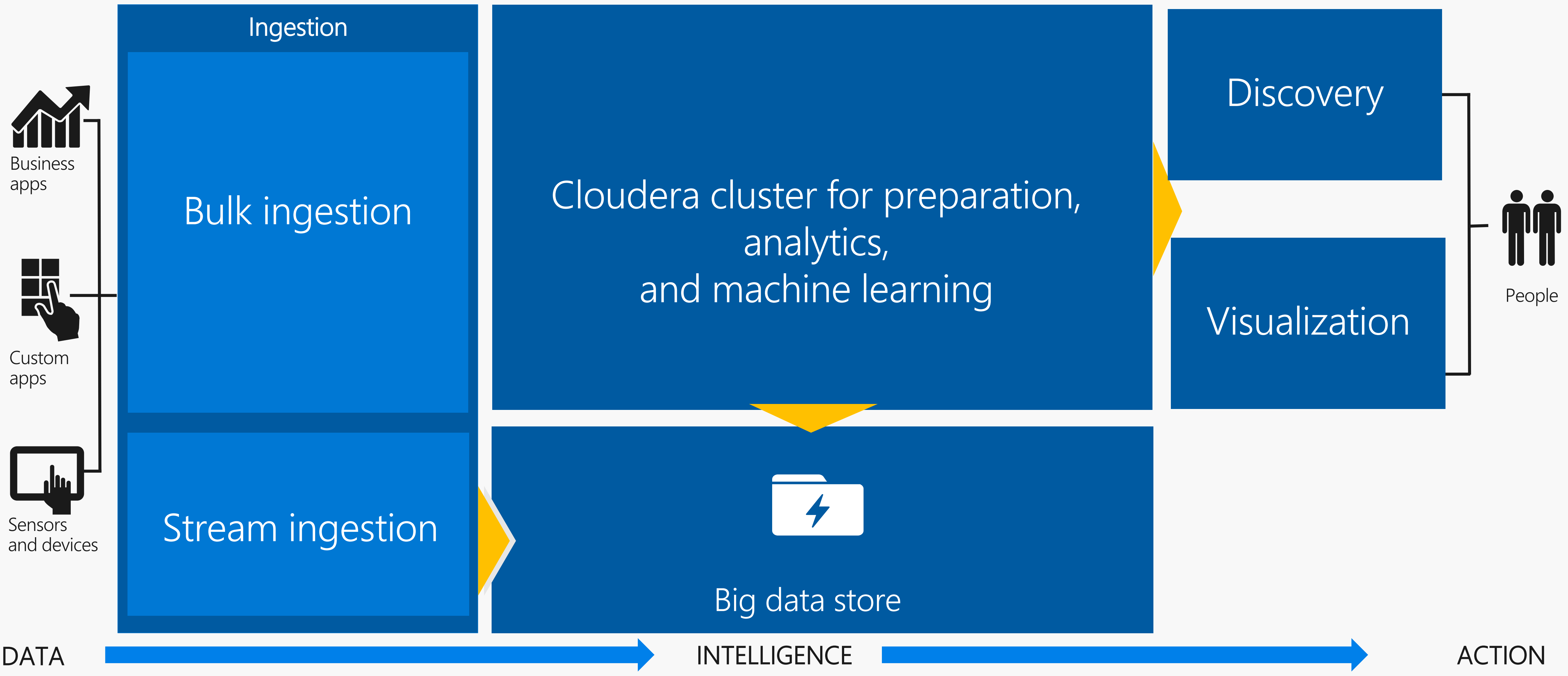
Decoupling data from compute



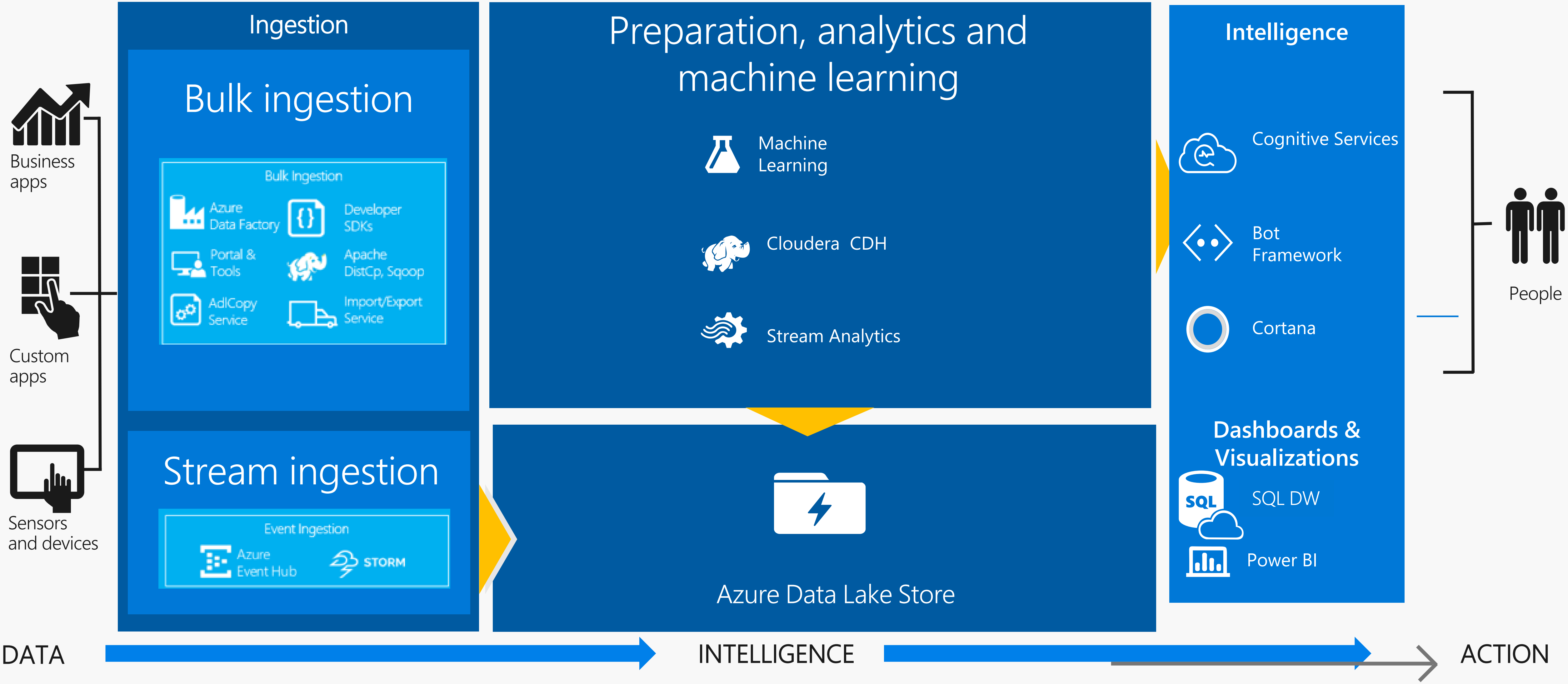
Decoupling data from compute (2)



Big data pipeline and workflow



Big data pipeline and data flow in Azure



Business apps

Custom apps

Sensors and devices

Ingestion

Bulk ingestion

Bulk Ingestion

- Azure Data Factory
- Developer SDKs
- Portal & Tools
- Apache DistCp, Sqoop
- AdlCopy Service
- Import/Export Service

Stream ingestion

Event Ingestion

- Azure Event Hub
- STORM

Preparation, analytics and machine learning

- Machine Learning
- Cloudera CDH
- Stream Analytics

Azure Data Lake Store

Intelligence

- Cognitive Services
- Bot Framework
- Cortana

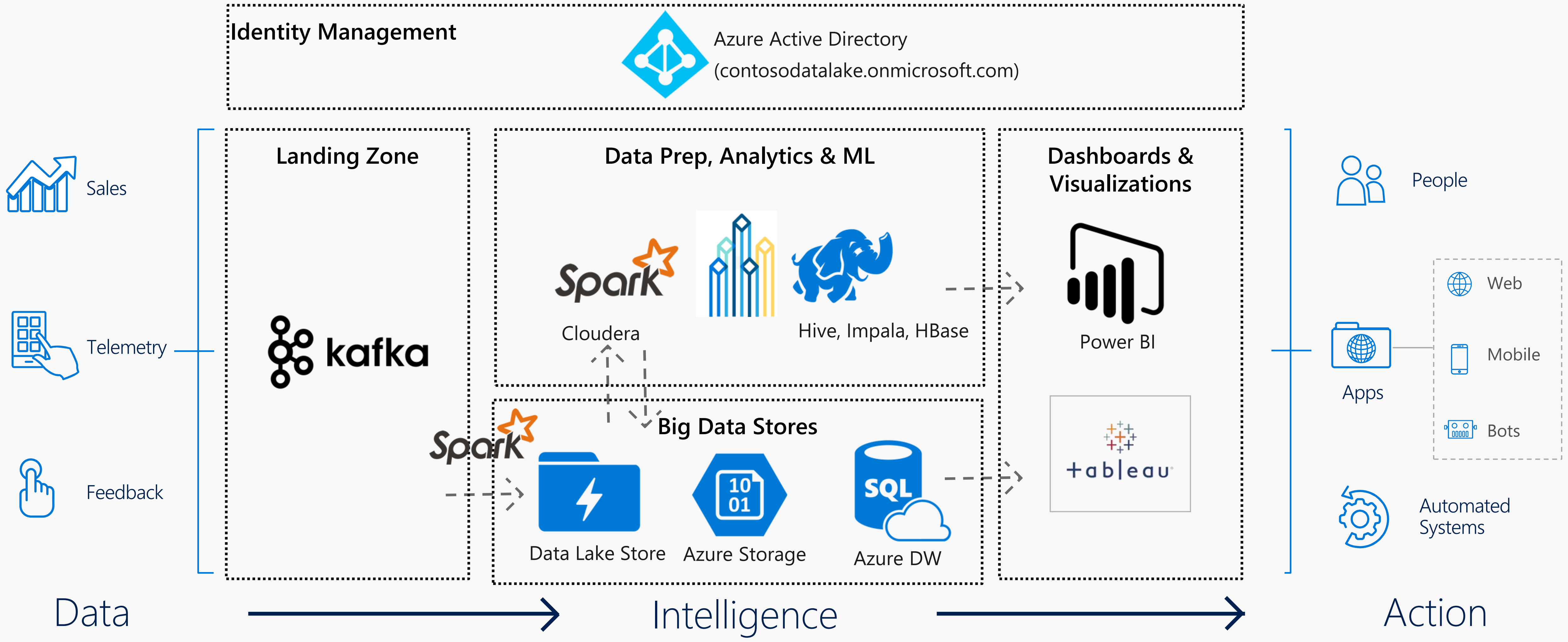
Dashboards & Visualizations

- SQL DW
- Power BI

People

DATA → INTELLIGENCE → ACTION

Cloudera on Azure Big Data Pipeline



Azure
Data Lake Store

A hyper-scale repository for big data analytics workloads

Hadoop File System (HDFS) for the cloud

No limits to scale

Store any data in its native format

Enterprise-grade access control, encryption at rest

Optimized for analytic workload performance

Azure Data Lake Store: No limits

- 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

Data Lake Store: Technical requirements

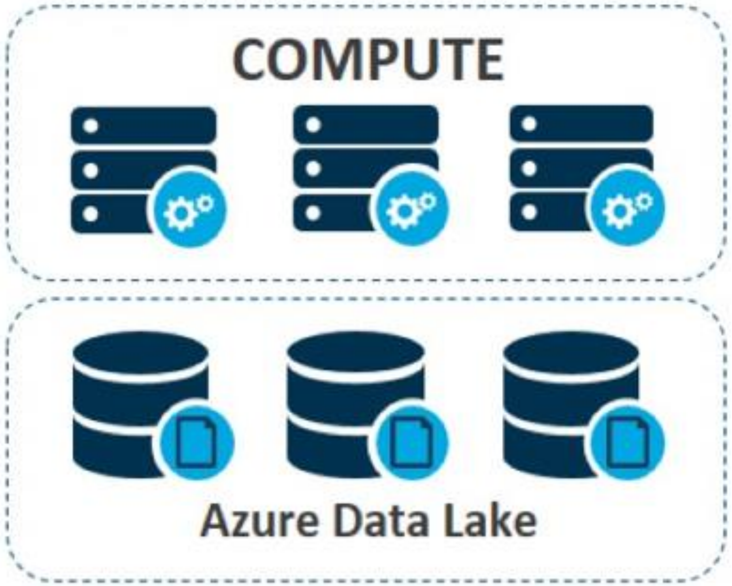
Secure	Must be highly secure to prevent unauthorized access (especially as all data is in one place)
Scalable	Must be highly scalable. When storing all data indefinitely, data volumes can quickly add up
Reliable	Must be highly available and reliable (no permanent loss of data)
Throughput	Must have high throughput for massively parallel processing via frameworks such as Hadoop and Spark
Low latency	Must have low latency for high-frequency operations
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

Microsoft

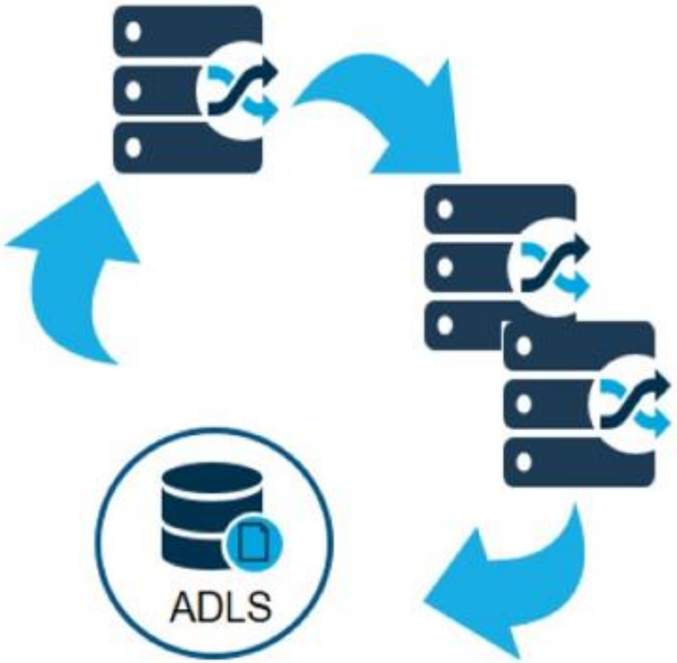
Cloudera on Azure Data Lake

Why Cloudera on Azure Data Lake Store?

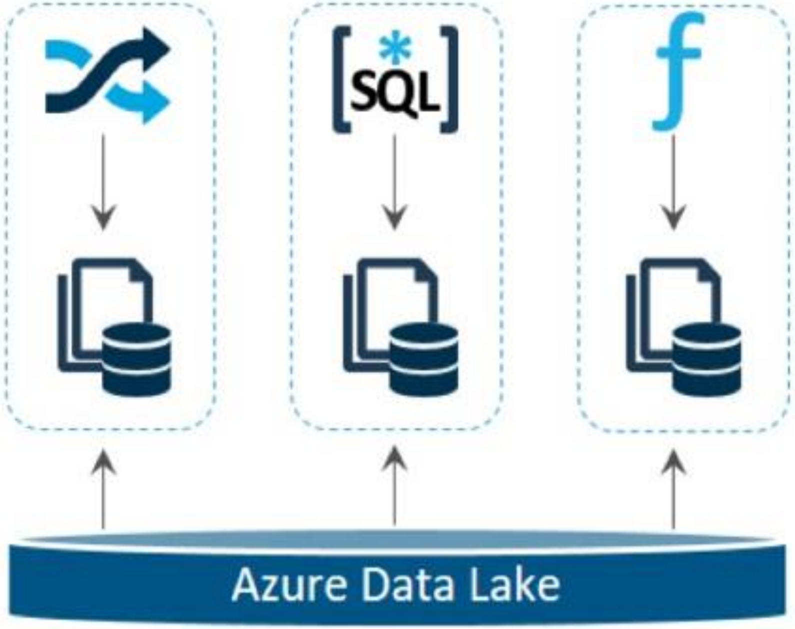
Separation of Compute & Storage



Transient clusters for flexibility, lower TCO



Shared storage for many optimized clusters



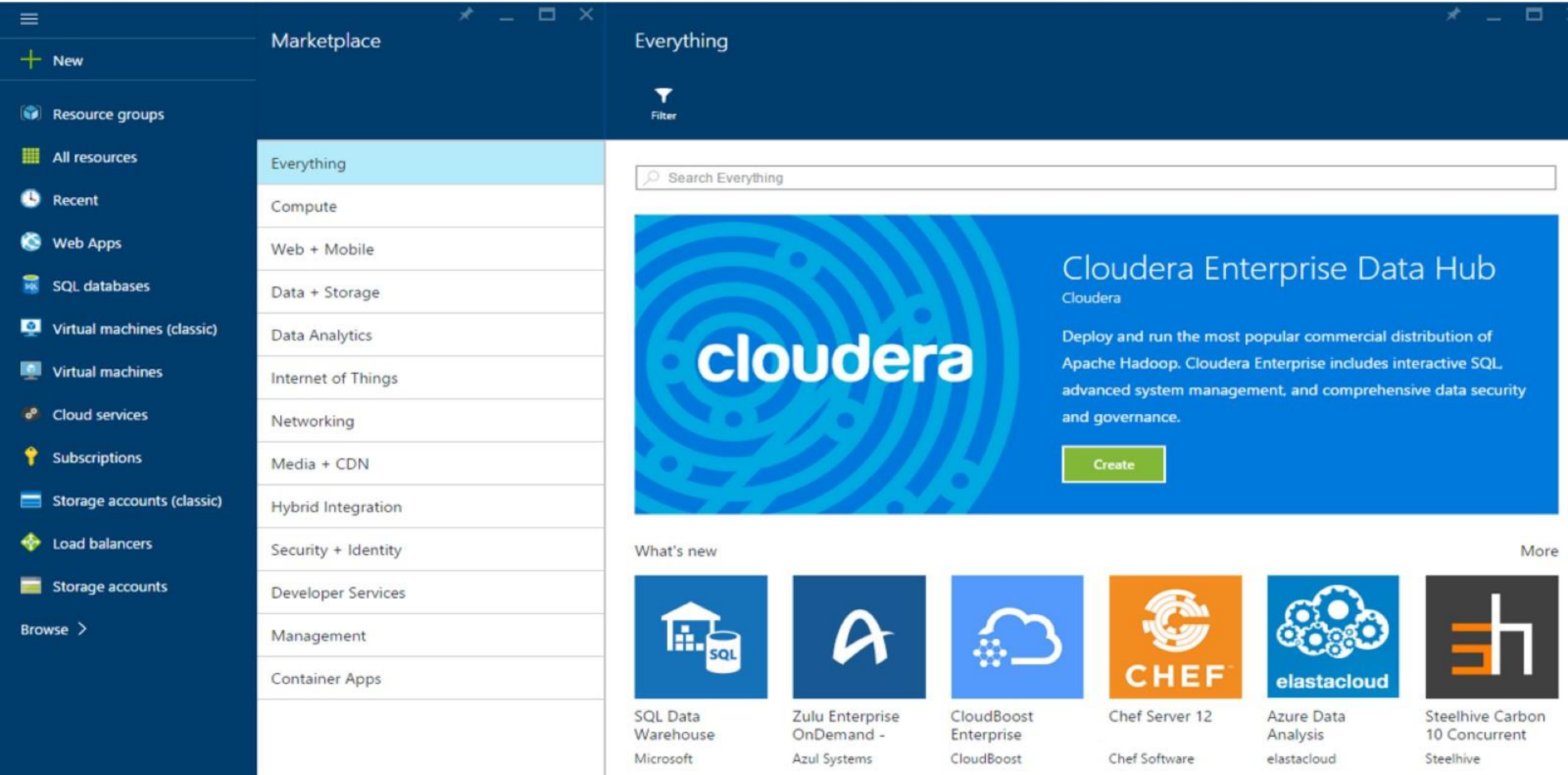
Microsoft

Cloudera/ADLS -Demo

Microsoft

How/Where to use Cloudera on Azure

Cloudera on Azure Marketplace



Cloudera Enterprise Data Hub

Cloudera

Deploy and run the most popular commercial distribution of Apache Hadoop. Cloudera Enterprise includes interactive SQL, advanced system management, and comprehensive data security and governance.

SQL Data Warehouse	Zulu Enterprise OnDemand -	CloudBoost Enterprise	Chef Server 12	Azure Data Analysis	Steelhive Carbon 10 Concurrent
Microsoft	Azul Systems	CloudBoost	Chef Software	elastacloud	Steelhive

Cloudera Director: Cloud Elasticity

Use for Hyper-scale Cloud Platforms

Easy Administration

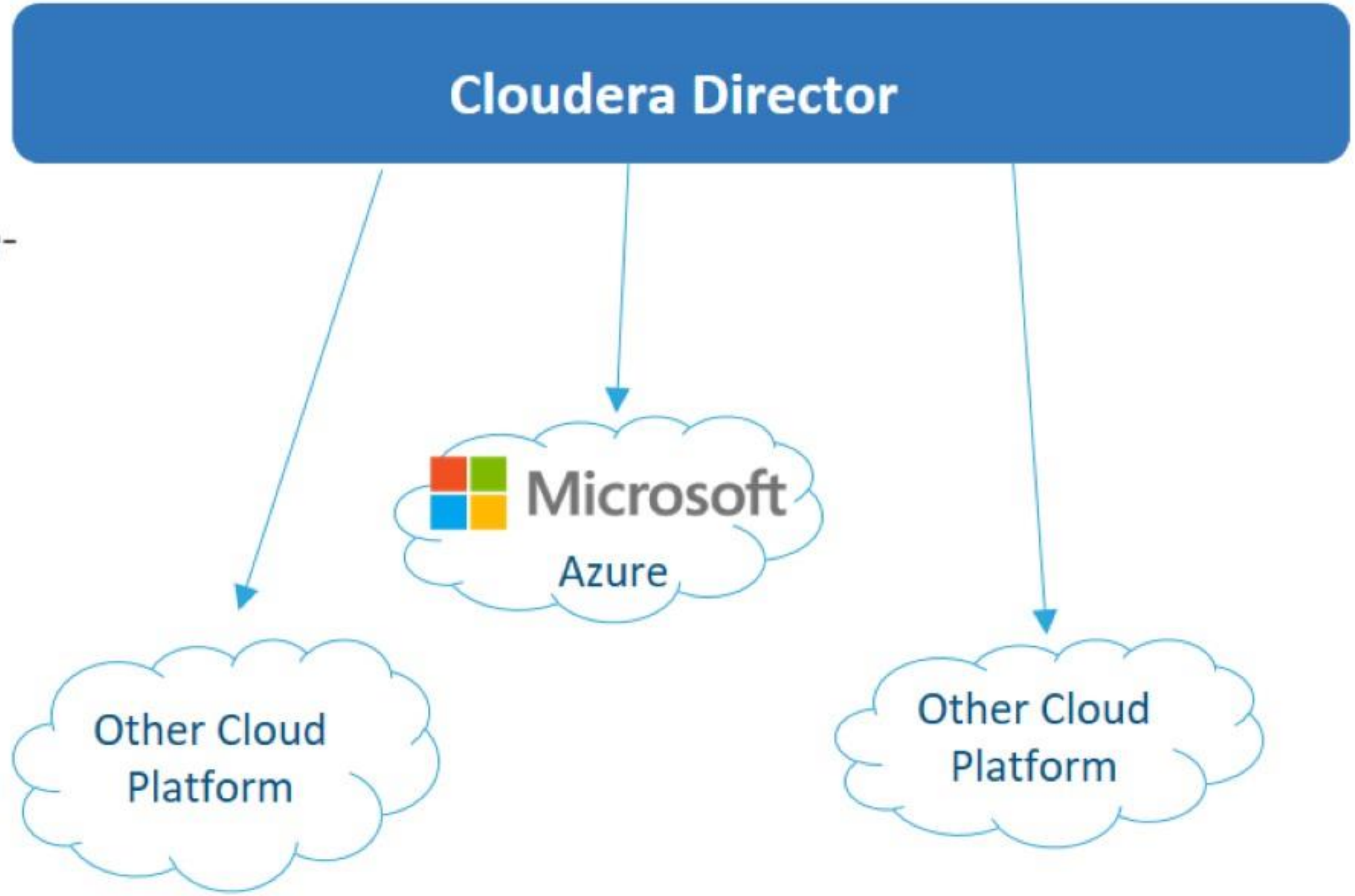
- Dynamic cluster lifecycle management
- Single pane of glass: multi-cluster view
- Create templates to run workloads in a pre-optimized manner

Flexible Deployments

- Multi public cloud
- Scaling of CDH clusters

Enterprise-grade

- Integration across Cloudera Enterprise
- Management of CDH deployments at scale



Q&A