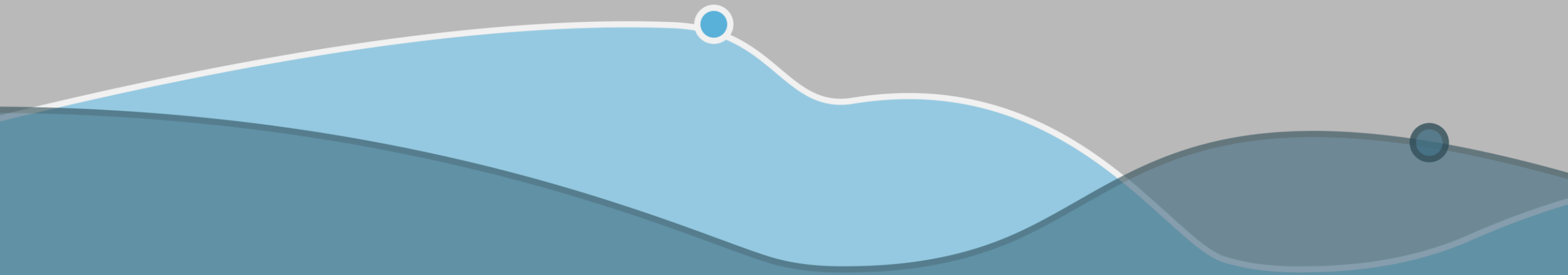
Microsoft

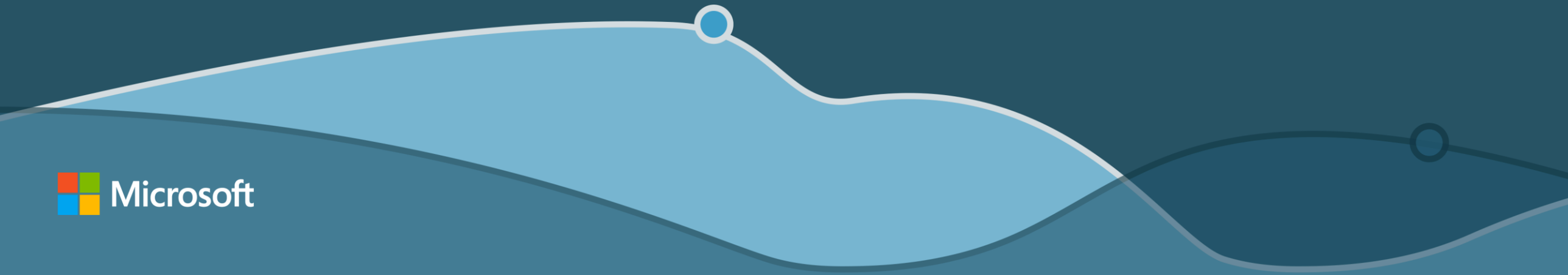
Cortana Analytics Workshop

Sept 10 – 11, 2015 • MSCC



Big Data @ Microsoft

Raghu Ramakrishnan
Technical Fellow
Head, Big Data Engineering and CISL



Microsoft


events
data
intelligence
people
insights
action
rules
automation

From data to decisions to actions

Trends

Data Storage Scarcity	→	Data Storage Abundance
Operational Data	→	Operational and Observational Data
Highly Modeled Schema	→	Flexible storage, Exploratory Analysis
Reporting	→	Insight, Predictions, Actions

Big Data

Store **any data**

Files, relations, docs, logs, graphs, multimedia …

Do **any analysis**

SQL queries, ML, image processing, log analytics …

Hive, Map-Reduce, Spark-ML, …

At **any speed**

Batch, interactive, streaming

Hive, Spark, Storm, …

At **any scale**

"Terabytes or more" or "more than your old system"

Elastic capacity

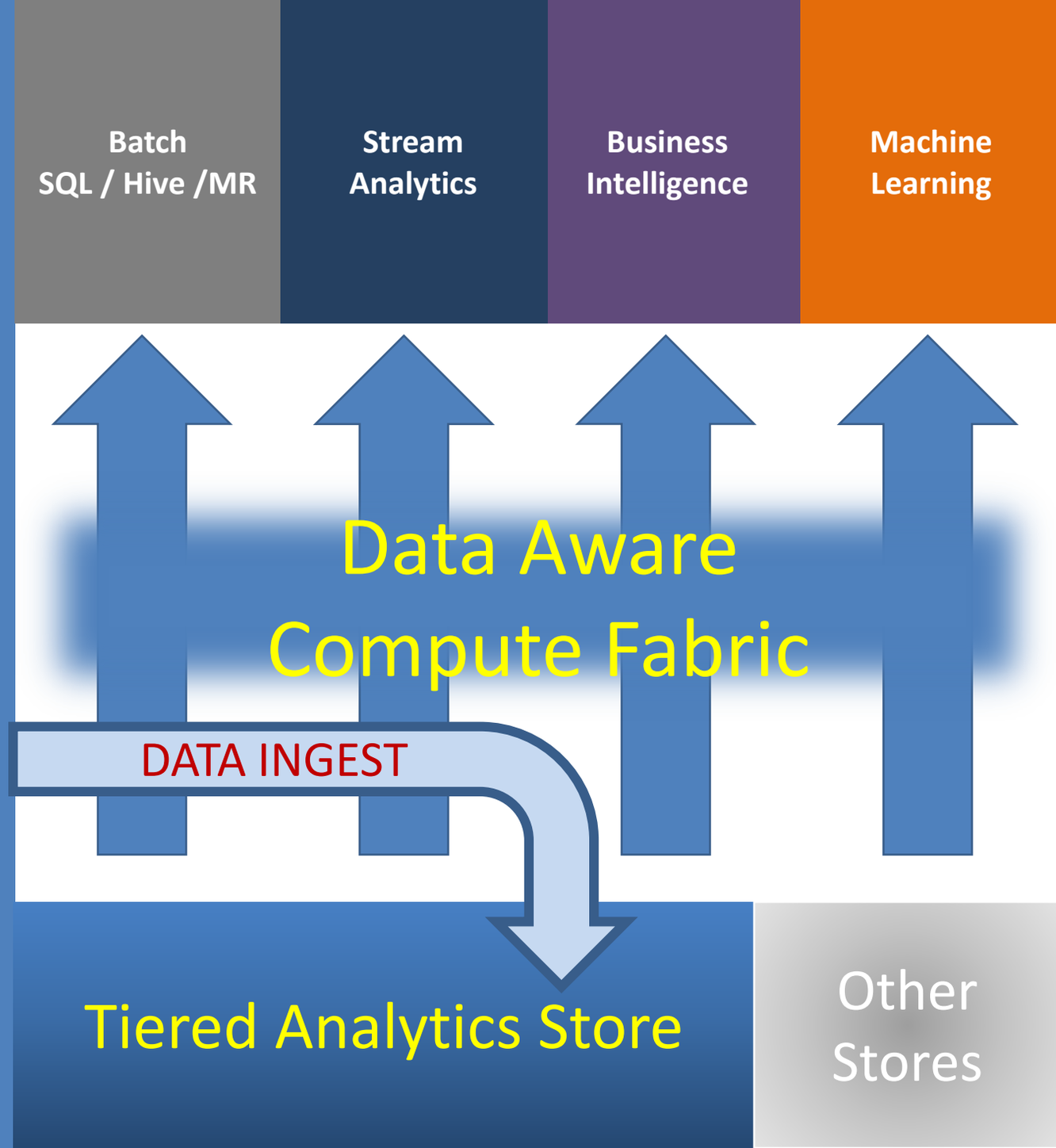
Big Data Overview

Capture any data, react instantaneously, mix with data stored anywhere

- Tiered storage management
- Federated access within and across clouds

Use any analysis tool (anywhere, mix and match, interactively)

- Shared compute fabric
- Extensible suite of tools



Principles: Data as a Core Enterprise Asset

Cloud first
All data accessible
Choice & open standards (HDFS, YARN)
Storage and processing scale independently
Secure and compliant
Simple to use, productive from Day 1

Big Data Services @ Microsoft

Cosmos – Exascale Big Data (Internal to Microsoft)

Azure Data Lake – Managed store for analytics

Azure HDInsight – Managed Hadoop Clusters

Azure SQL DW – Managed Relational Warehouse

Cosmos/Scope

The Big Data service Microsoft runs on

internal developers: 1000s

daily jobs: 100s of Ks

daily I/O: >100 PBs

data managed: EBs

cluster sizes: 10s of Ks

machines: 100s of Ks

Azure Data Lake

Fully managed cloud data store designed for analytics

Supports HDFS compliant analytics applications and tools

Enterprise grade security, compliance & management

Petabyte files, unlimited account size

High throughput for analytics performance

Low latency ingestion with read as you write

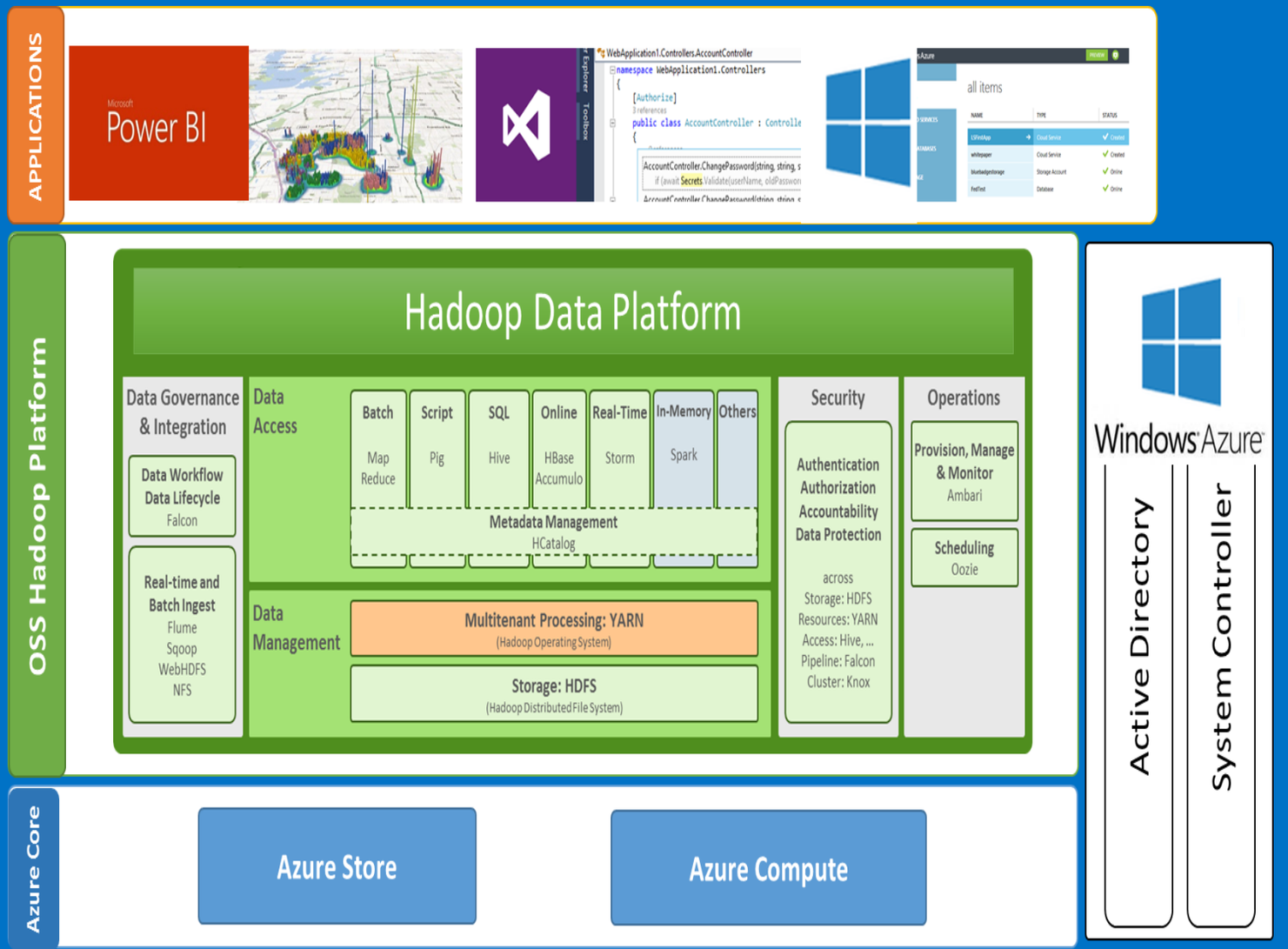
AAD-based authentication, access auditing

File and folder-level ACLs

Encryption at rest

Azure HDInsight

- 100% open source Apache Hadoop
- Full Hadoop ecosystem as a managed service, supported and backed by Microsoft on
  - Linux
  - Windows
- Harness .Net or Java to write customer extensions
- Supports broad ecosystem of ISVs (Hadoop and Traditional)



Azure HDInsight

The Best of Hadoop

- Batch
  - MapReduce, PIG, Hive, Spark
- Interactive SQL
  - Hive (Tez), SparkSQL
- Stream Analytics
  - Storm, SparkStreaming
- Machine Learning
  - SparkML, Mahout
- Table Serving
  - Hbase
- Exploratory Visualization
  - Jupyter, Zeppelin

Made Better with Azure

- Interactive SQL
  - SQL DW
- Stream Analytics
  - Azure Stream Analytics
- Machine Learning
  - Azure ML
- Table Serving
  - Azure SQL DB
- Exploratory Visualization
  - Power BI

Azure SQL DW

Fully managed relational data warehouse-as-a-service
First elastic cloud data warehouse with proven SQL Server capabilities
Support your smallest to your largest data storage needs

Elastic scale & performance

Scales to petabytes of data

Massively Parallel Processing

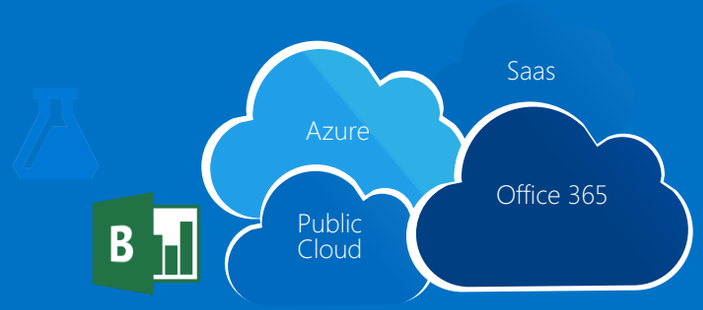
Instant-on compute scales in seconds

Query Relational / Non-Relational

Powered by the Cloud

Get started in minutes

Integrated with Azure ML, PowerBI & ADF

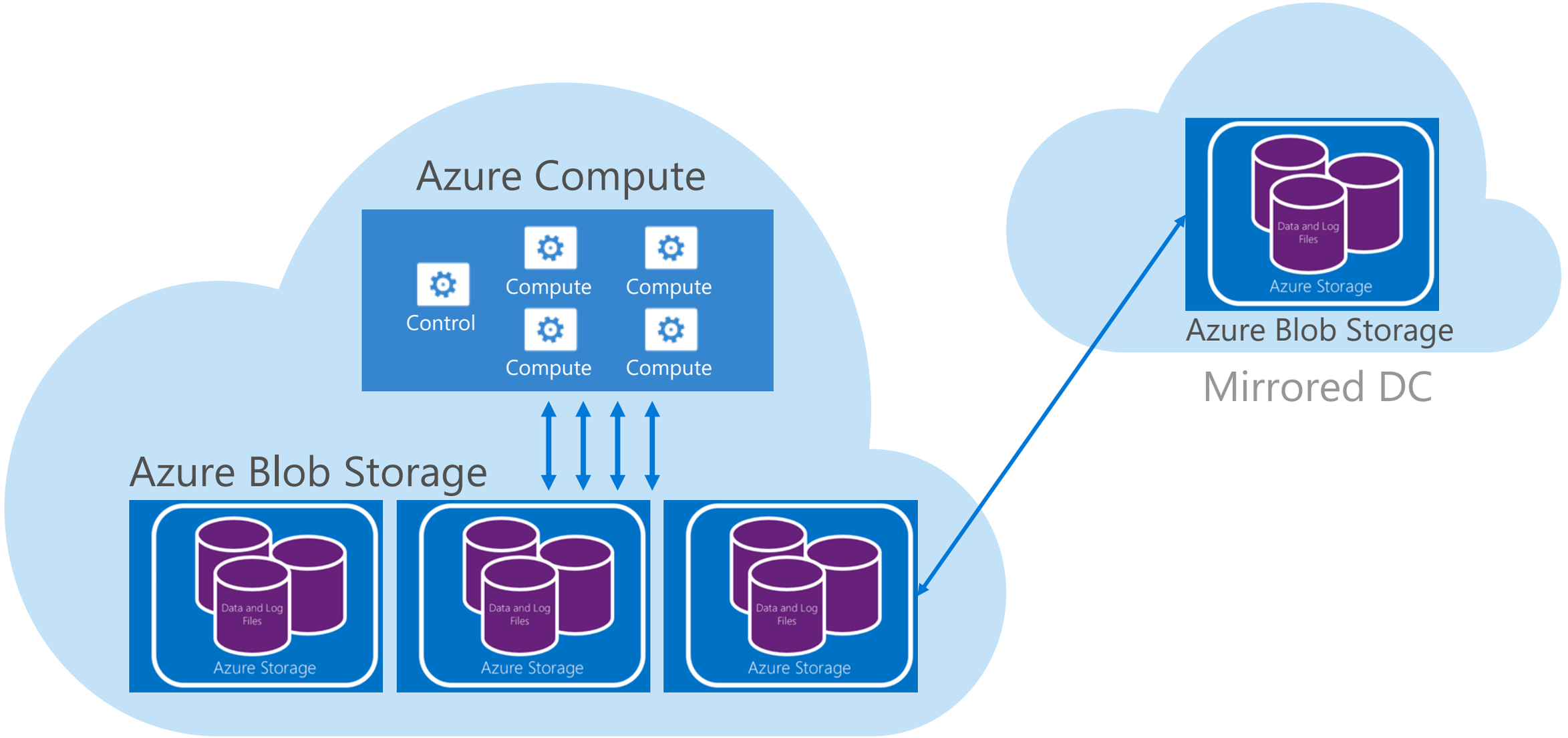


Market Leading Price & Performance

Simple billing compute & storage

Pay for what you need, when you need it with dynamic pause

Azure SQL-DW Compute and Storage



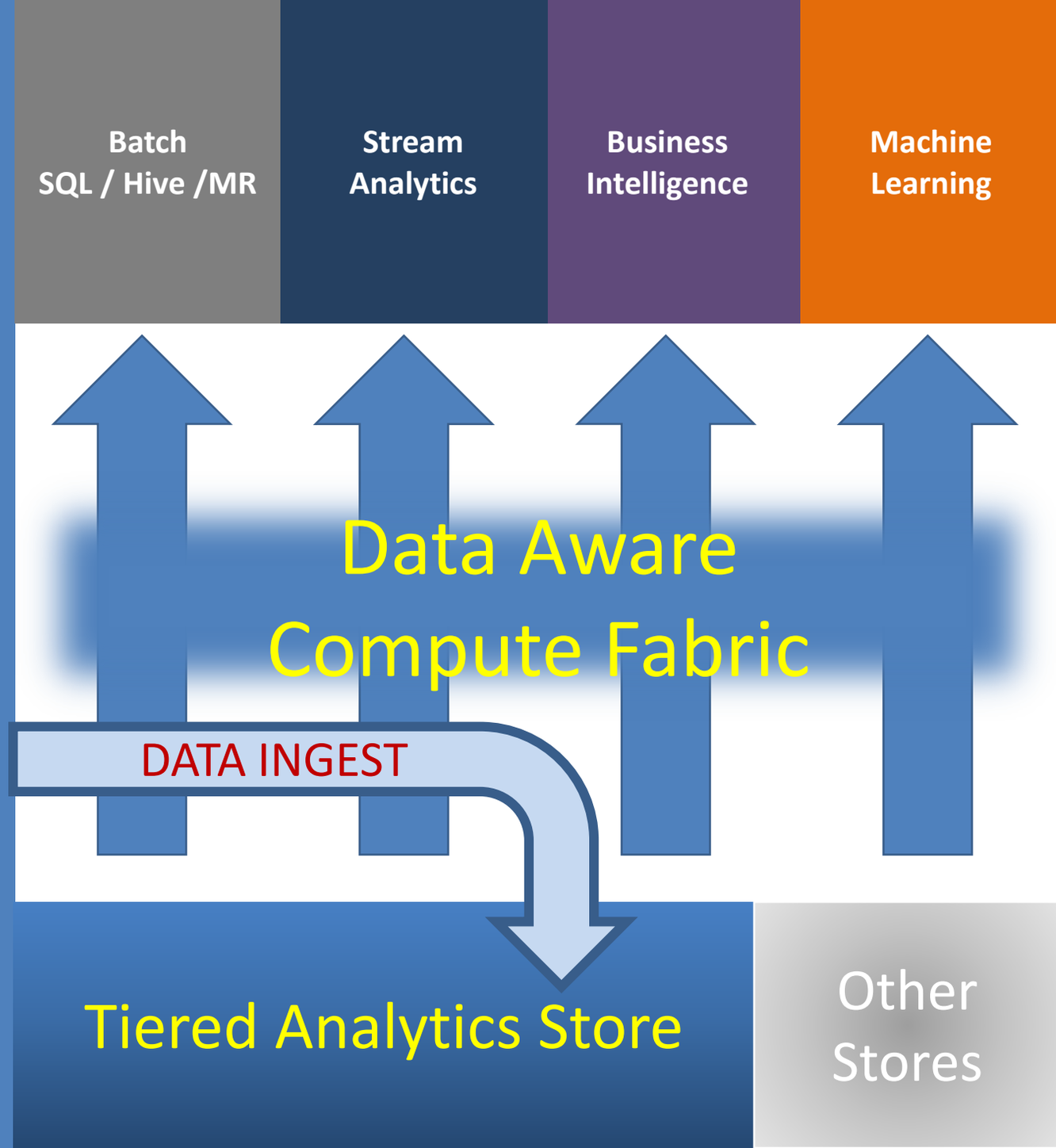
Big Data Overview

Capture any data, react instantaneously, mix with data stored anywhere

- Tiered storage management
- Federated access within and across clouds

Use any analysis tool (anywhere, mix and match, interactively)

- Shared compute fabric
- Extensible suite of tools



Microsoft