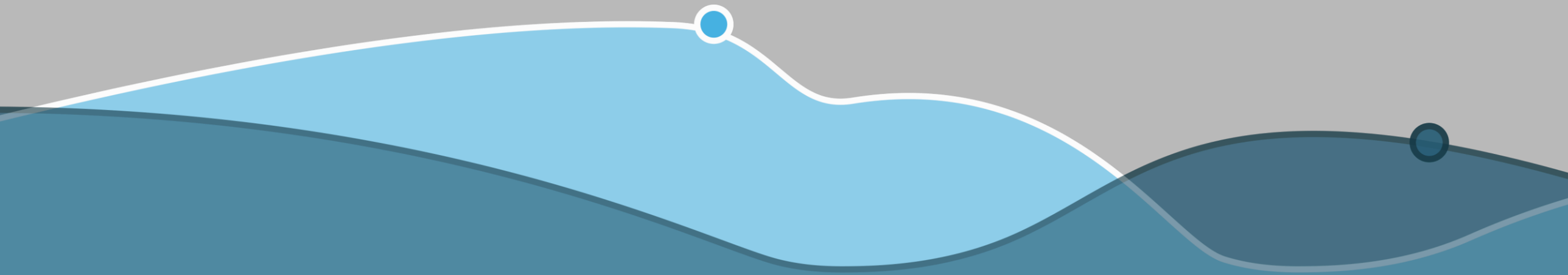
Microsoft

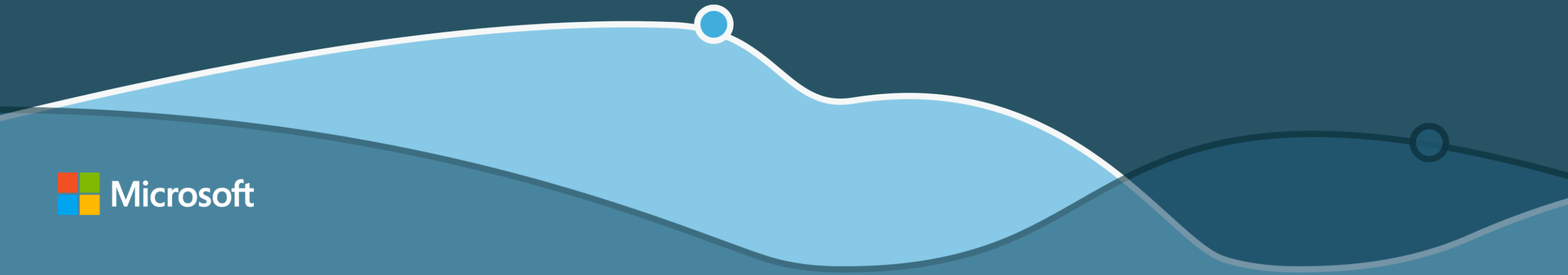
Cortana Analytics Workshop

Sept 10 – 11, 2015 • MSCC



Azure Data Lake

Rajesh Dadhia
Group Program Manager
Microsoft Big Data Platform



Microsoft

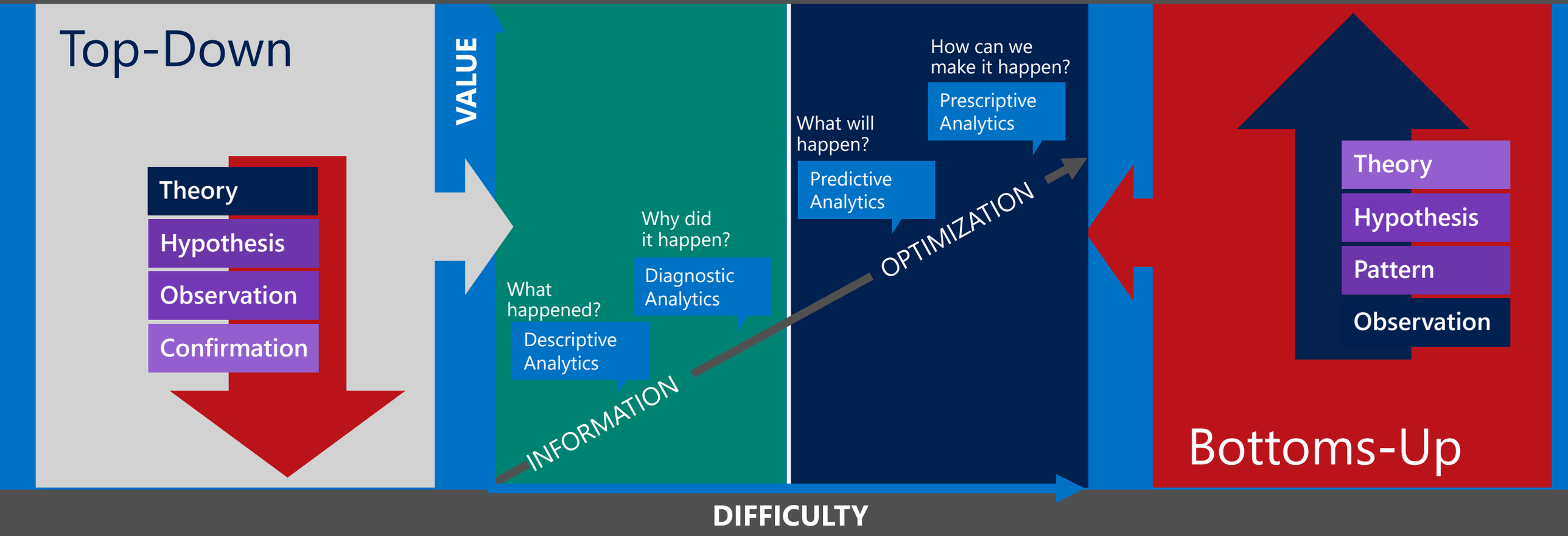
Cortana Analytics Suite
Big Data & Advanced Analytics

What is Azure Data Lake?

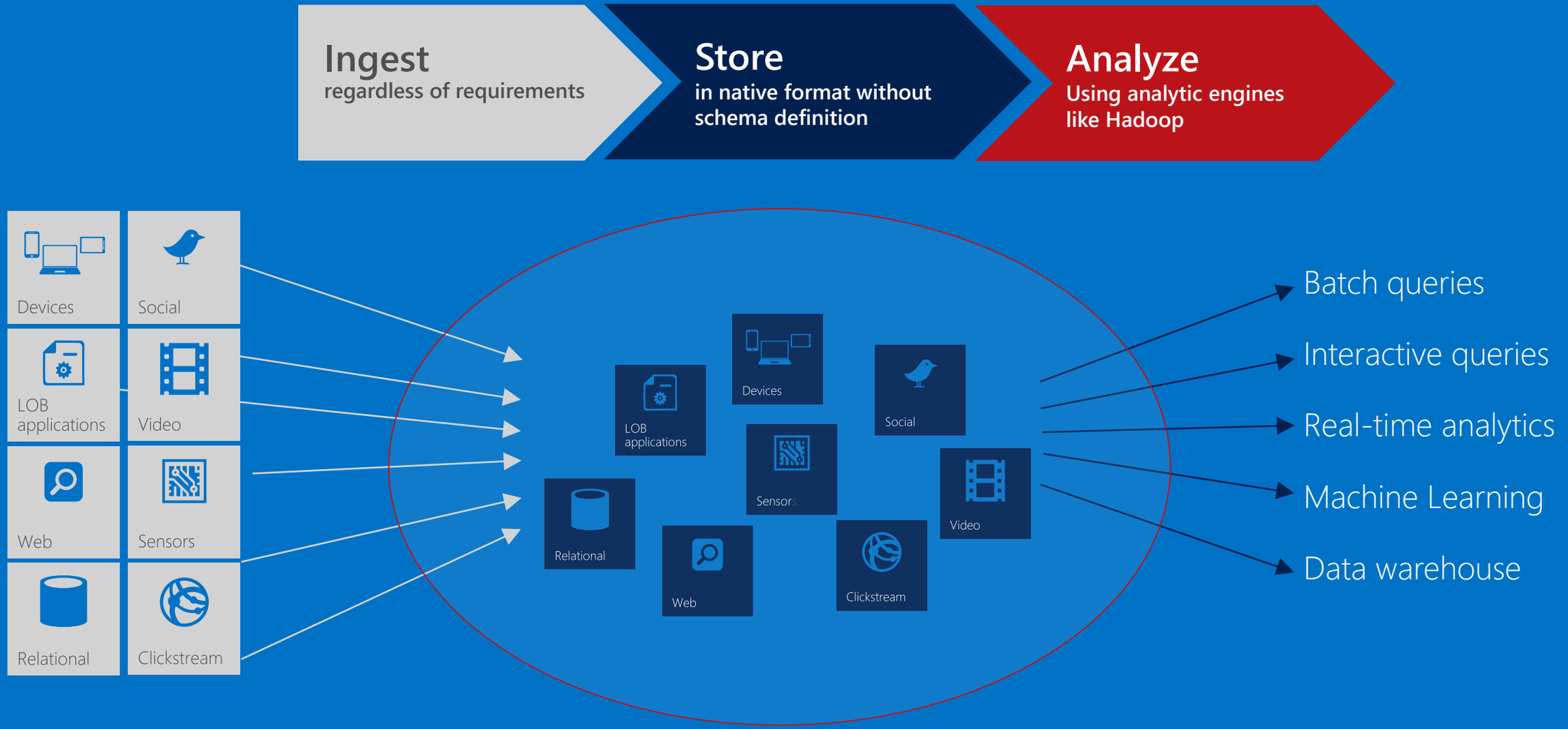
A hyper scale repository for any data, optimized for big data analytic workloads.

Why do we need Data Lakes?

Two Approaches to Analytics



The “data lake” Uses Bottoms-Up Approach



Introducing Azure Data Lake

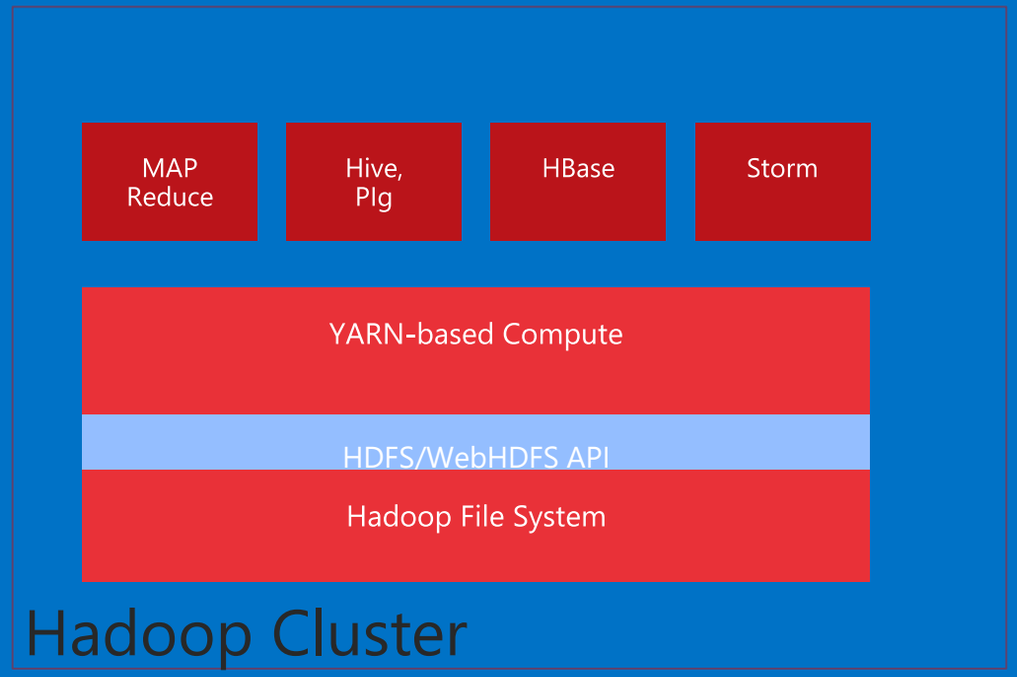
A hyper scale repository for big data analytics workloads

Azure Data Lake

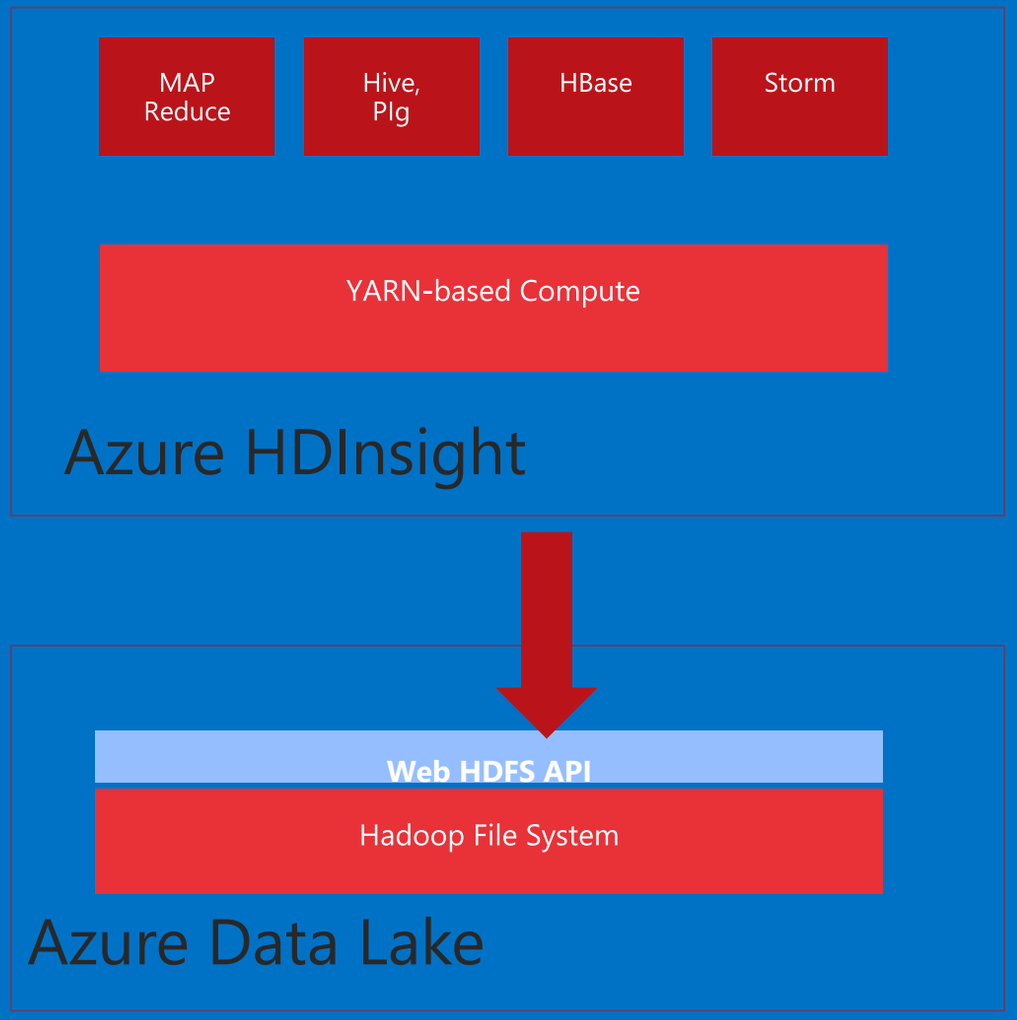
- Store any data in its native format
- Hadoop File System (HDFS) for the cloud
- Enterprise grade
- No limits to scale
- Optimized for analytic workload performance

Azure Data Lake & Azure HDInsight

On-Premises



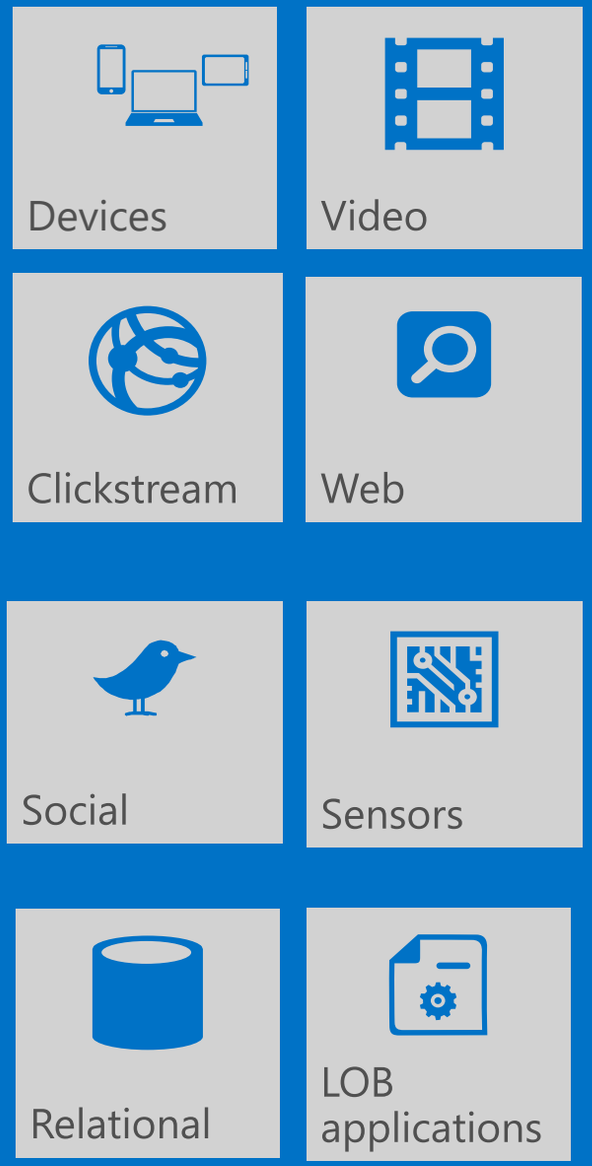
Azure Cloud



Any Data

Any Data

- Unstructured
- Semi-structured
- Structured



In it's native form…

Ingress

Azure SQL

Azure Tables

Azure SQL DW

Server Logs

Azure Data Factory
Sqoop

On-prem Databases

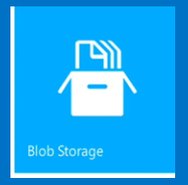
Azure Data Factory

LOB & other Apps

Azure Data Factory
Third-party tools

Azure Data Lake

ADL built-in Copy Service
Azure Data Factory



Azure Storage Blobs

ASA

Azure Web Portal via Browser
Azure PowerShell
.NET SDK
JavaScript CLI

Client Machines

Egress

Azure SQL

Azure Tables

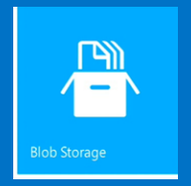
Azure SQL DW

Server Logs

Azure Data Factory
Sqoop

Azure Data Lake

ADL built-in Copy Service
Azure Data Factory



Azure Storage Blobs

On-prem Databases

Azure Data Factory

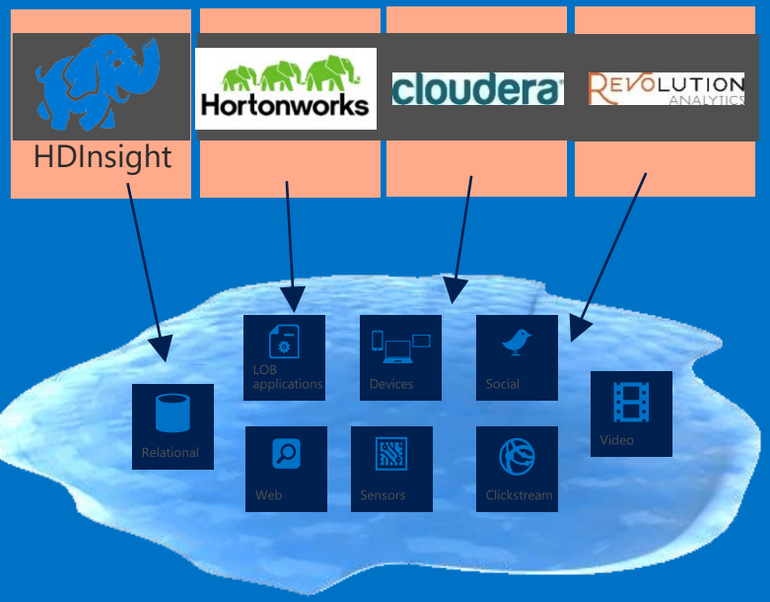
Azure Web Portal via Browser
Azure PowerShell
.NET SDK
JavaScript CLI

Client Machines

Cloud HDFS

Hadoop Distributed File System (HDFS) For The Cloud

``



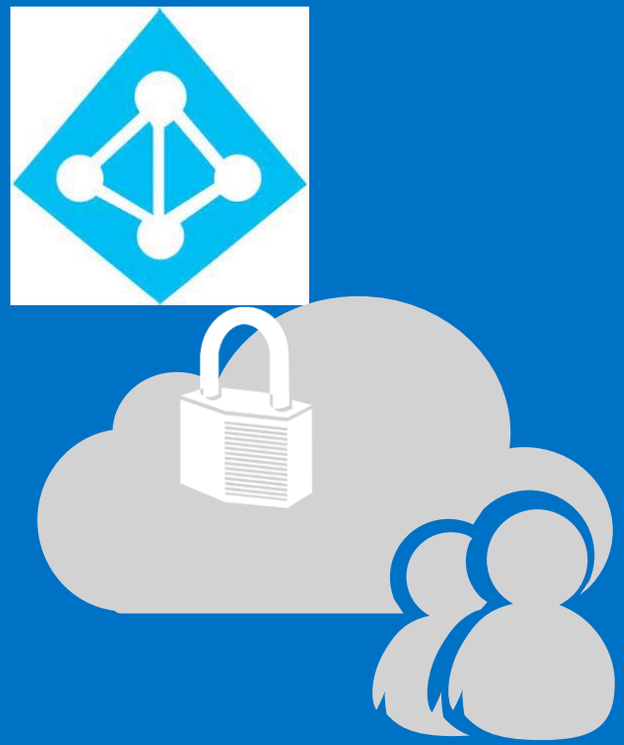
Microsoft Azure Data Lake

- Built from the ground-up as a Hadoop File System
- Support for file/folder objects and operations
- Integrated w/ HDInsight, Hortonworks, Cloudera
- Accessible to all HDFS compliant projects (Spark, Storm, Flume, Sqoop, Kafka, R, etc.)

Built using open standards

Enterprise Grade

Manage and Secure Your Data Assets



- Azure Active Directory integration
- File and folder level access control
- Audit data access
- Encryption of data-at-rest

Protect your data assets

Security

Access Control

- Secure Files and Folders
- POSIX compliant ACLs
  - Minimal (octet) and enhanced ACLs
- Based on Azure AD principals

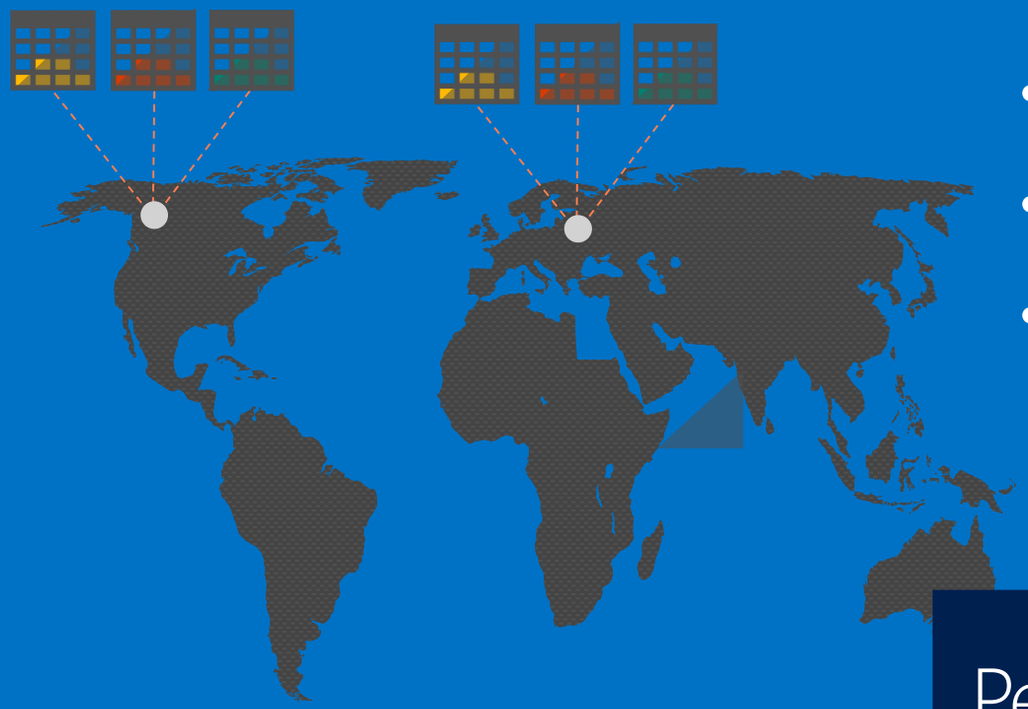
Auditing

- Audit logs for all operations
- Consumable via big data analytics

Encryption at Rest

- Transparent server-side encryption
- Azure Managed and Customer managed Keys
- Azure Key Vault Integration

Durable and Highly Available

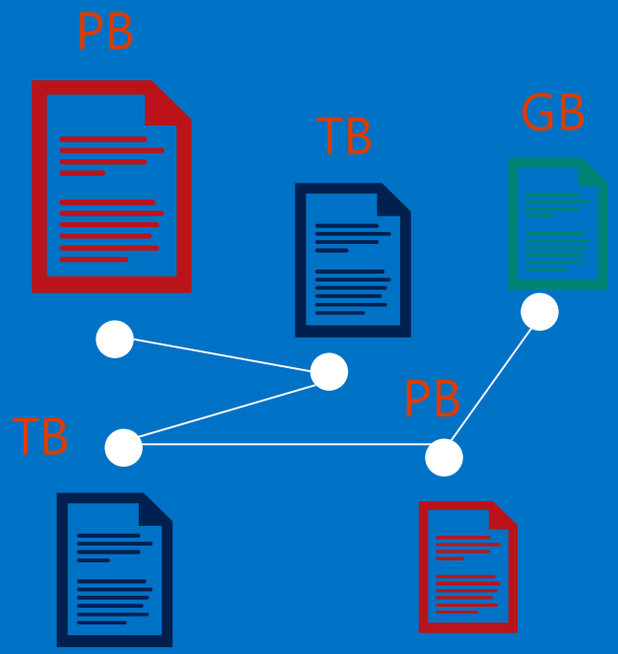


- Automatically replicates your data
- 3 copies within a single region
- Highly available

Peace of mind for data of high durability

No Limits To Scale & Optimized Performance

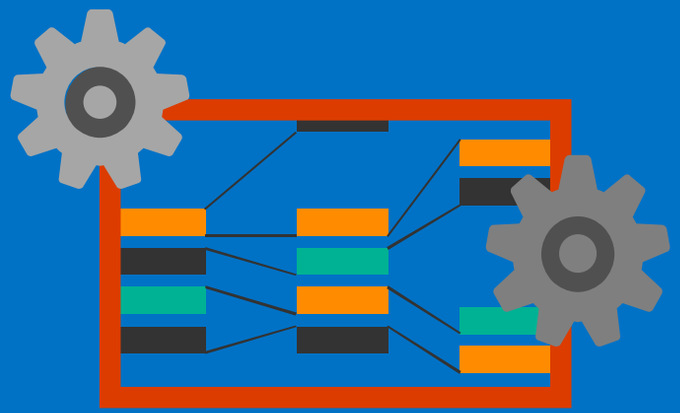
Unlimited Storage, Petabyte Files



- Unlimited account sizes
- Individual file sizes from GBs to PBs
- No limits to scale

Useful for scenarios with very large data

Optimized for Analytic Workload Peformance

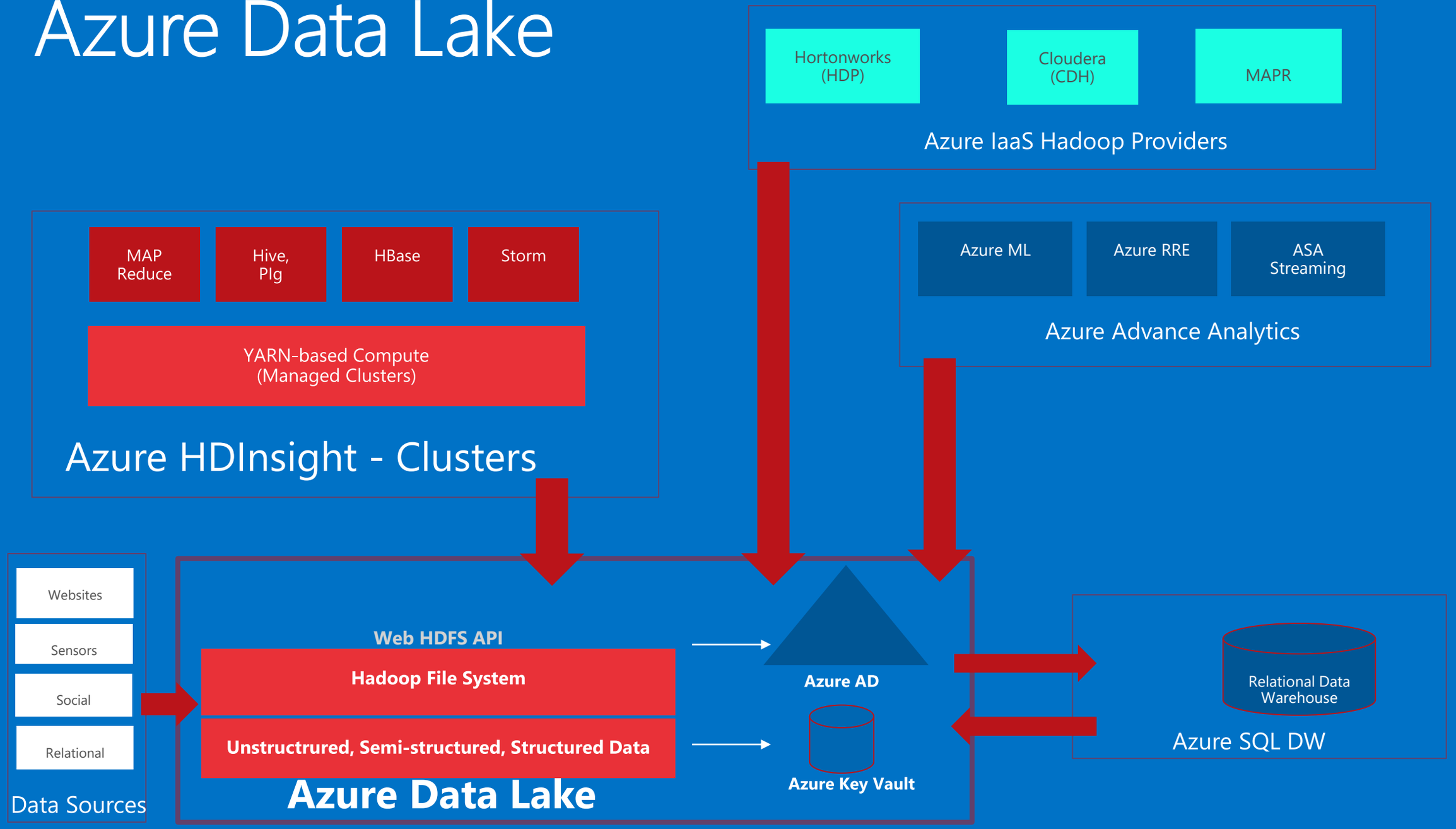


- Built for running large analytic systems that require massive throughput
- Optimized for parallel computation over PBs of data
- Automatically optimize for any throughput

Focus only on writing application logic

Architecture & Ecosystem

Azure Data Lake



To Summarize

Azure Data Lake

- Can store structured, semi-structured, unstructured data
- Can support all Hadoop applications
- Is built for the enterprise
- Can meet performance needs of big data applications

Microsoft