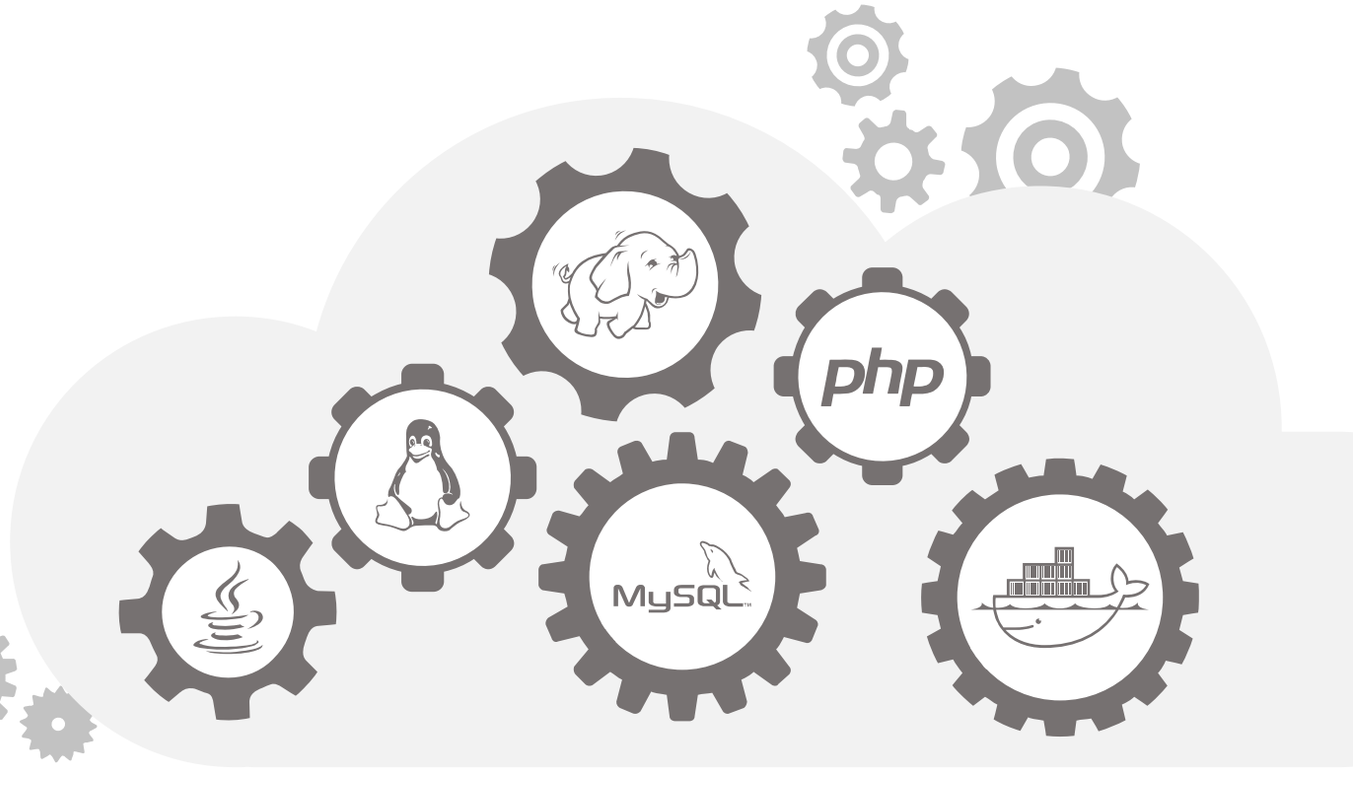
Data Lake Organization

A Hadoop Eco-System



Jan Cordtz, Microsoft Denmark

jcordtz@Microsoft.com

Cloud Solution Architect

Hyper scale Infrastructure

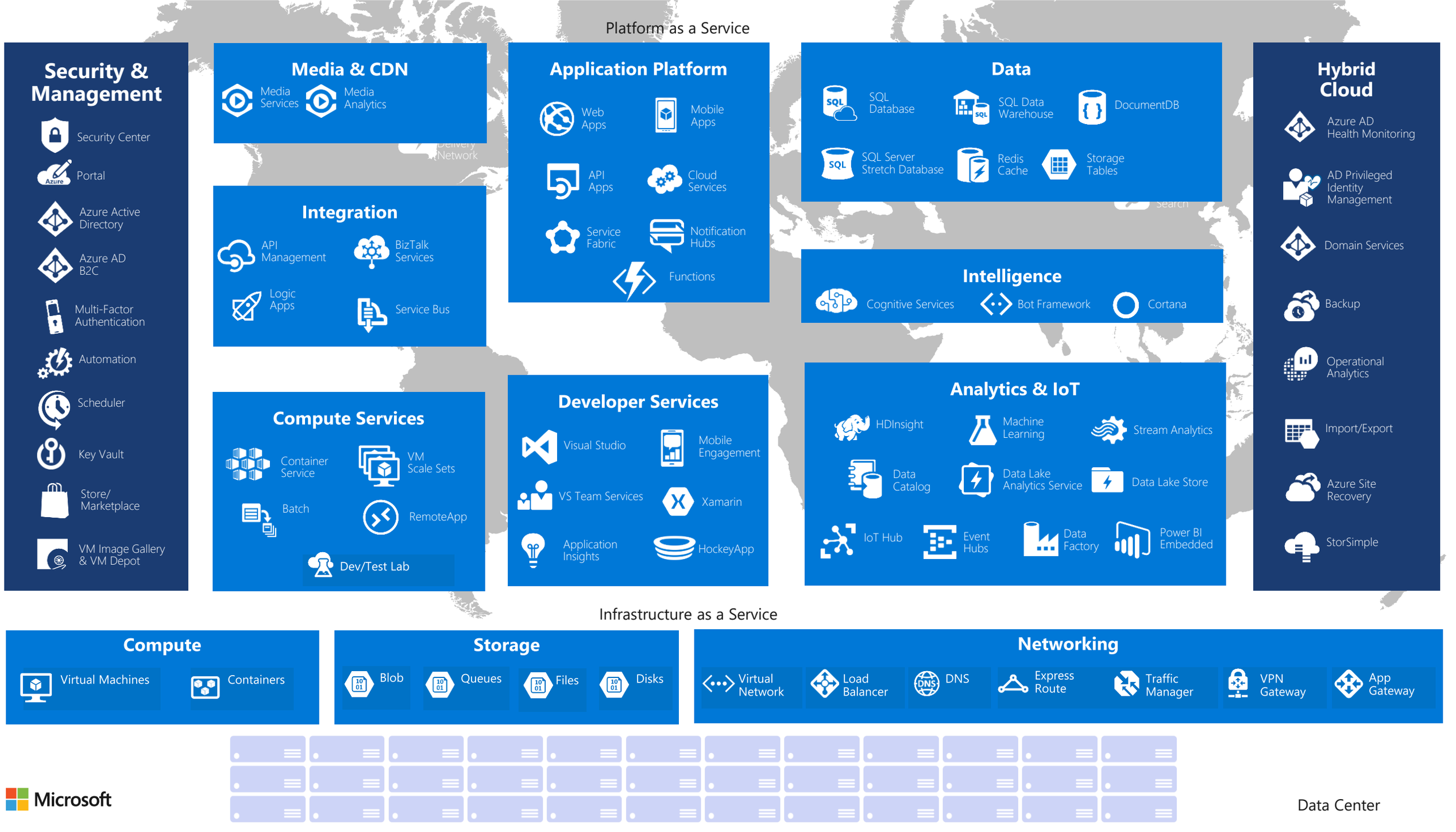
100+ Data-centers across 42 Regions Worldwide

Top 3 networks in the world

Learn more:
Microsoft.com/datacenter

42 Azure regions

Microsoft


Platform as a Service
Security & Management
Security Center
Portal
Azure Active Directory
Azure AD B2C
Multi-Factor Authentication
Automation
Scheduler
Key Vault
Store/ Marketplace
VM Image Gallery & VM Depot
Media & CDN
Media Services
Media Analytics
Integration
API Management
BizTalk Services
Logic Apps
Service Bus
Compute Services
Container Service
VM Scale Sets
Batch
RemoteApp
Dev/Test Lab
Application Platform
Web Apps
Mobile Apps
API Apps
Cloud Services
Service Fabric
Notification Hubs
Functions
Developer Services
Visual Studio
Mobile Engagement
VS Team Services
Xamarin
Application Insights
HockeyApp
Data
SQL Database
SQL Data Warehouse
DocumentDB
SQL Server Stretch Database
Redis Cache
Storage Tables
Intelligence
Cognitive Services
Bot Framework
Cortana
Analytics & IoT
HDInsight
Machine Learning
Stream Analytics
Data Catalog
Data Lake Analytics Service
Data Lake Store
IoT Hub
Event Hubs
Data Factory
Power BI Embedded
Hybrid Cloud
Azure AD Health Monitoring
AD Privileged Identity Management
Domain Services
Backup
Operational Analytics
Import/Export
Azure Site Recovery
StorSimple
Infrastructure as a Service
Compute
Virtual Machines
Containers
Storage
Blob
Queues
Files
Disks
Networking
Virtual Network
Load Balancer
DNS
Express Route
Traffic Manager
VPN Gateway
App Gateway
Microsoft
Data Center

More certifications than any cloud provider

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

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

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

Microsoft

Microsoft

”Let’s Rethink” Data

How to (easily) disrupt a Data Warehouse

Digital transformation/disruption

Planning is dead

Build and run open source solutions

Open and hybrid

Applications

WordPress, Drupal, OpenShift Enterprise by Red Hat, Cloud Foundry

DevOps

Chef, Ansible, GitHub, puppet labs

Frameworks

Microsoft .NET, node.js, python, Java, php, R

Databases & middleware

cloudera (Ask Bigger Questions), Hortonworks, MariaDB, cassandra

Containers

docker, DC/OS, CoreOS, kubernetes

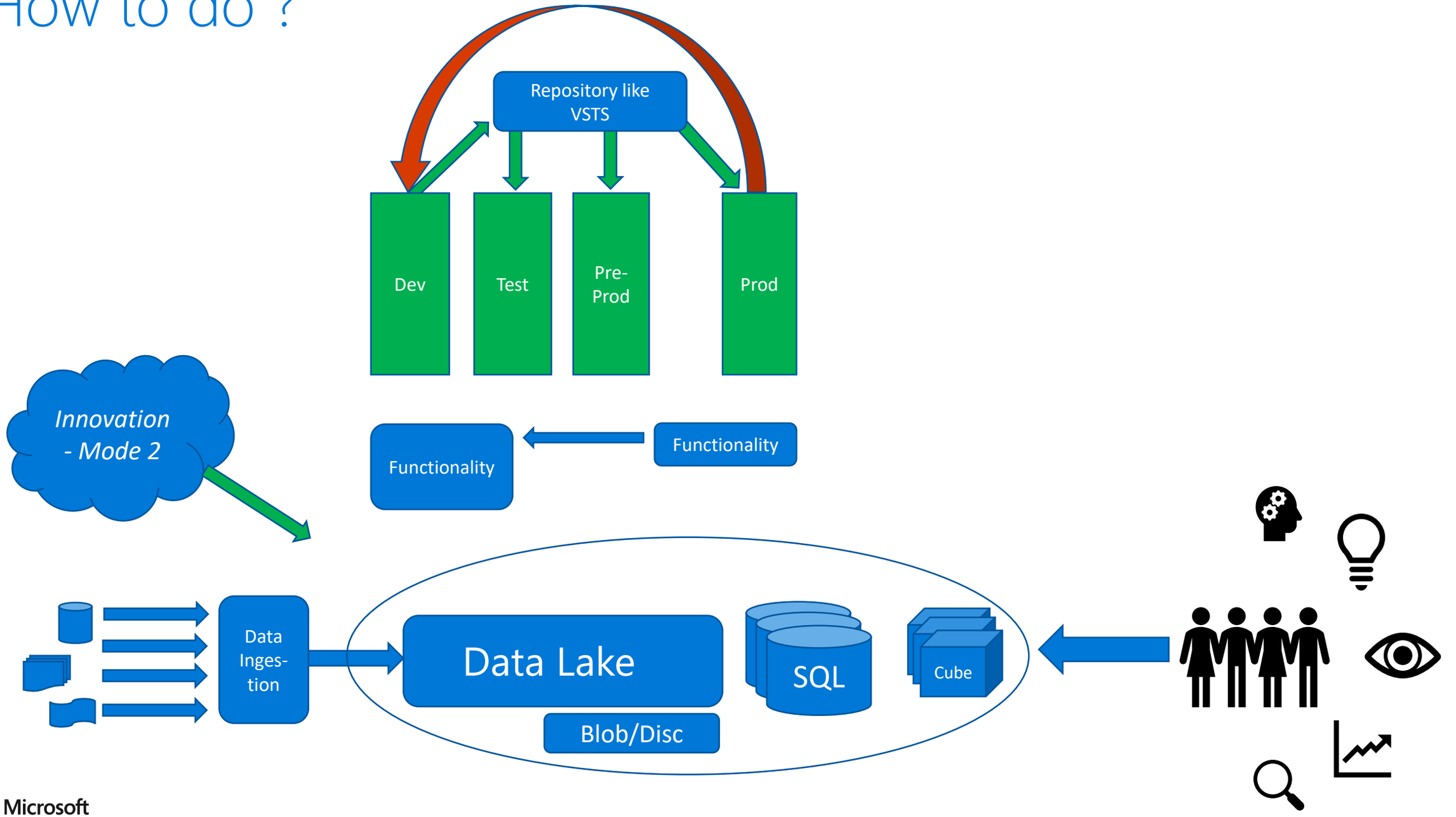
Infrastructure

DC/OS, redhat, CONSENSYS, blockstack, ubuntu, docker, FreeBSD

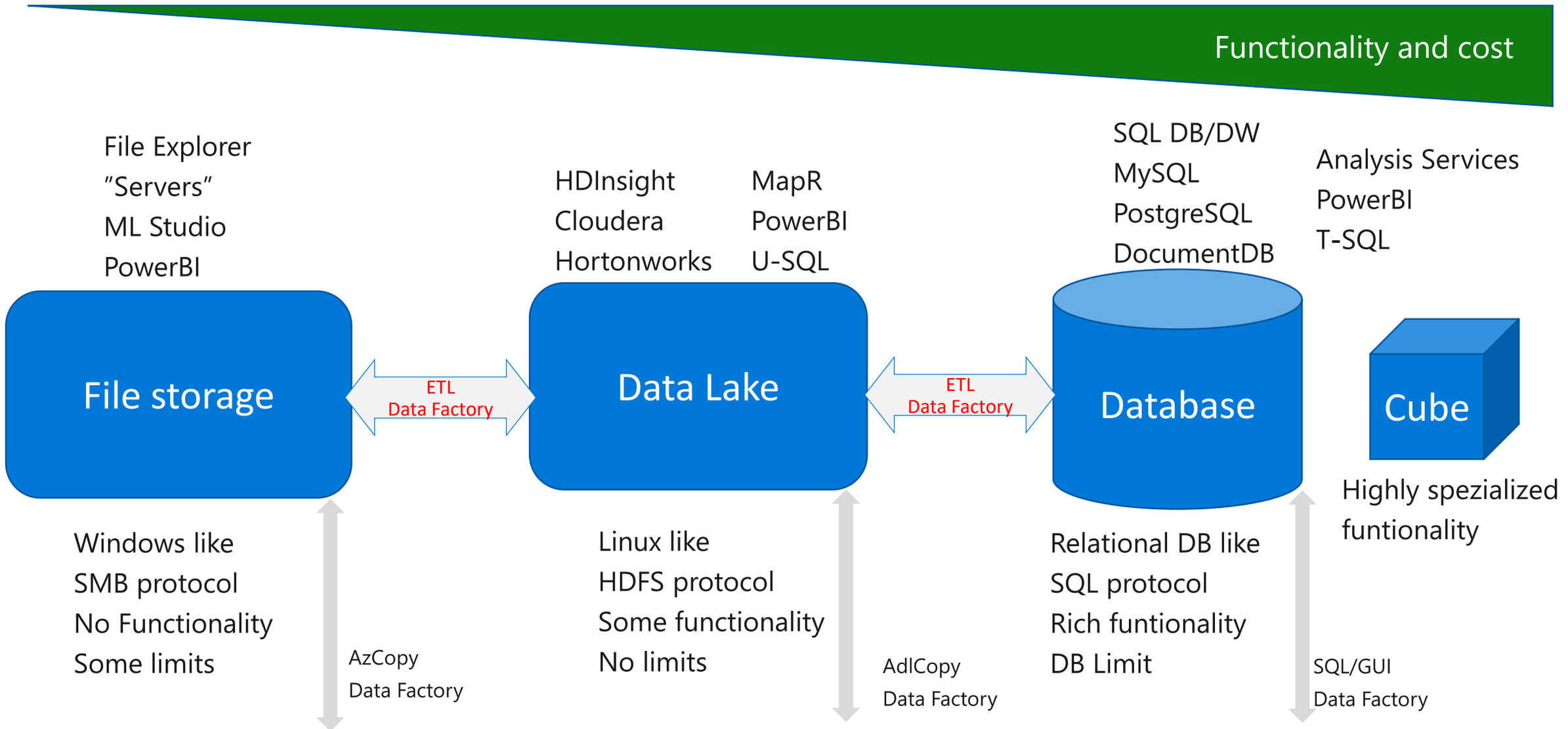
Microsoft

How to use and organize….

How to do ?


Repository like VSTS
Dev
Test
Pre-Prod
Prod
Innovation - Mode 2
Functionality
Functionality
Data Inges-tion
Data Lake
Blob/Disc
SQL
Cube

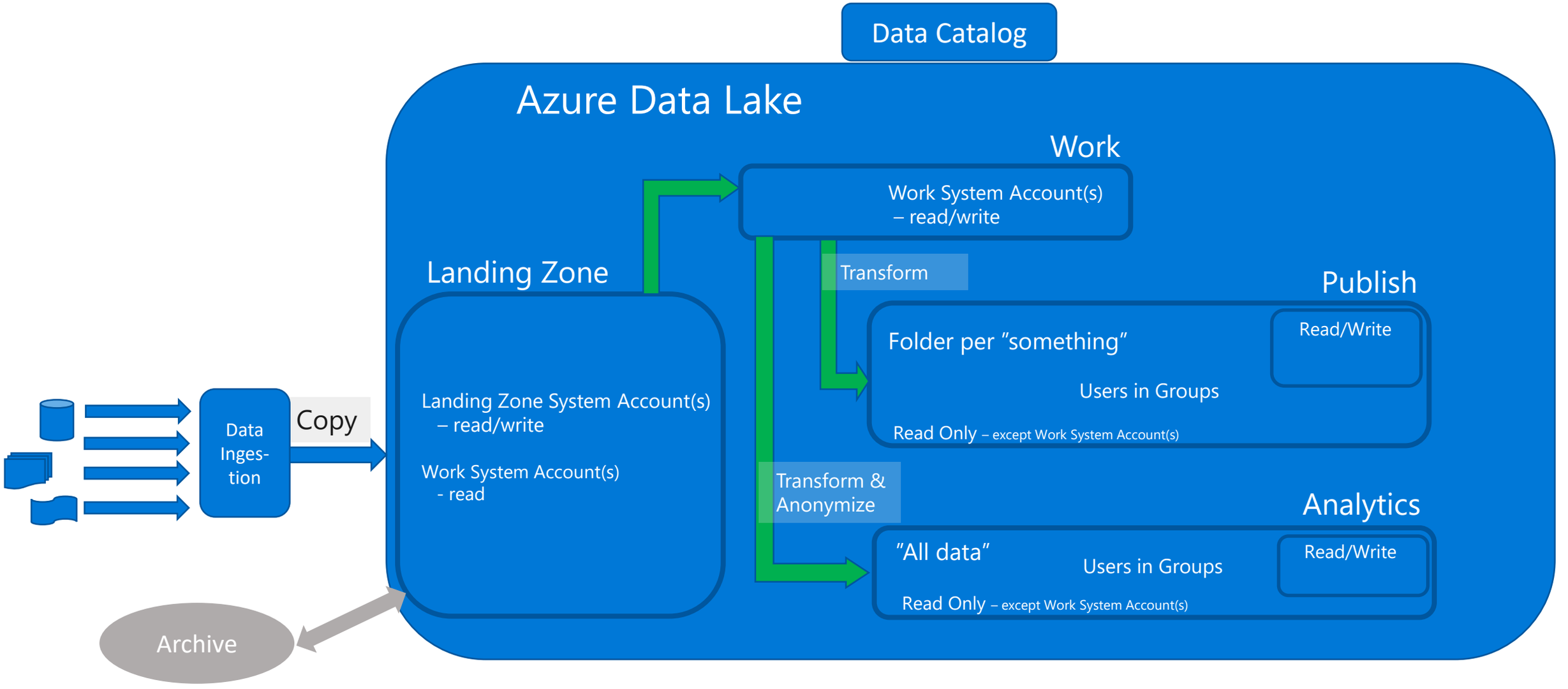
Storage – from a functionality point of view



Principal regarding the Organization

- Is very simple to use for an end-user/application
- Is as cost-effective as sensible/possible.
- Do not compromise security.
- Fits well into a DevOps scenario
- Supports both of Gartner's mode 1 and mode 2 implementation scenarios.
- Have a well-defined path for the information needed to be able to support an effective auditing and logging process.

Organizing the Azure Data Lake



Data Ingestion

Standardization

Items like : Date formats (yyyymmdd), number formats (,. or .,)

Validation

"Is the content you are coming with in accordance with what we have agreed"

"Gatekeeper"

"Are you allowed to come in ?"

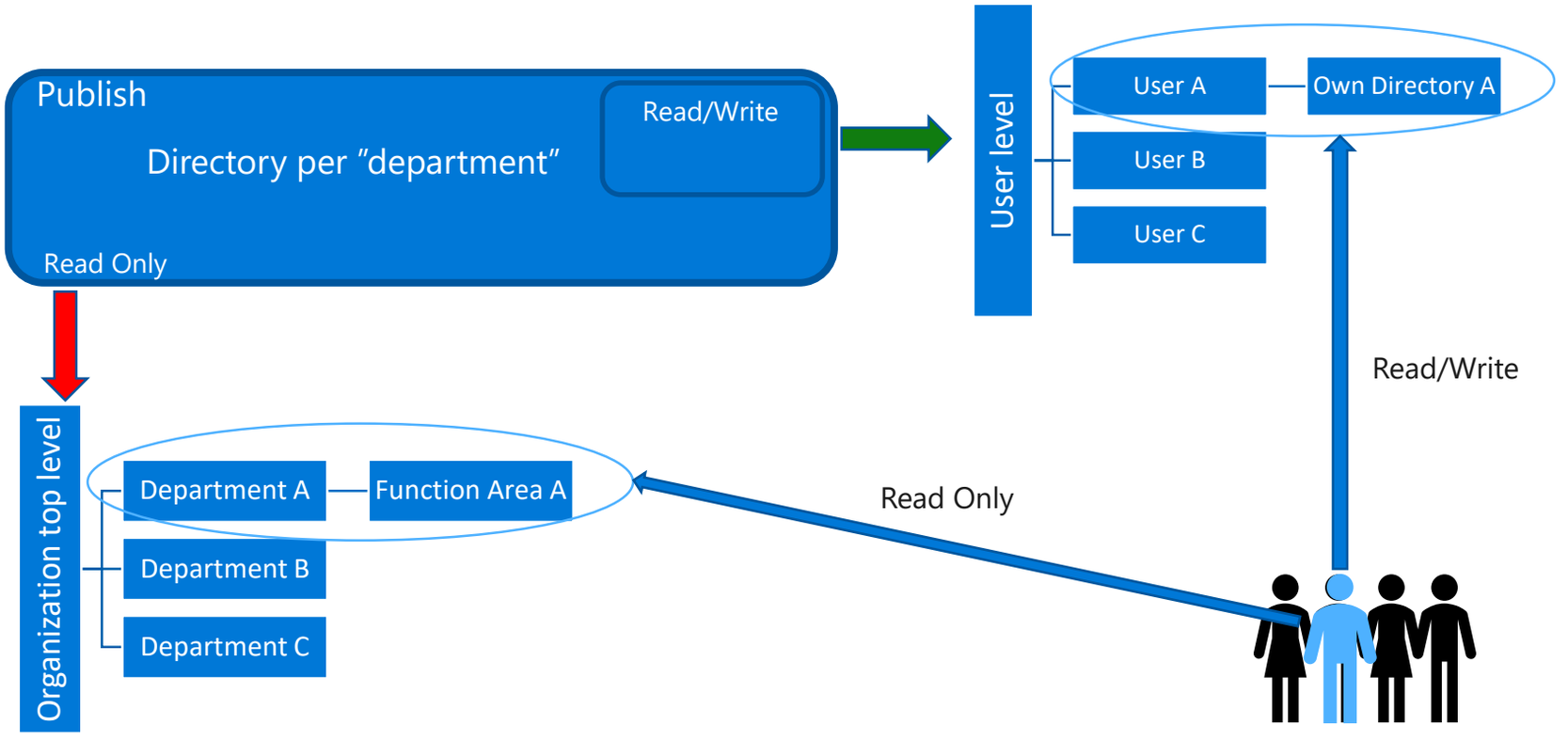
SSIS
Data Factory

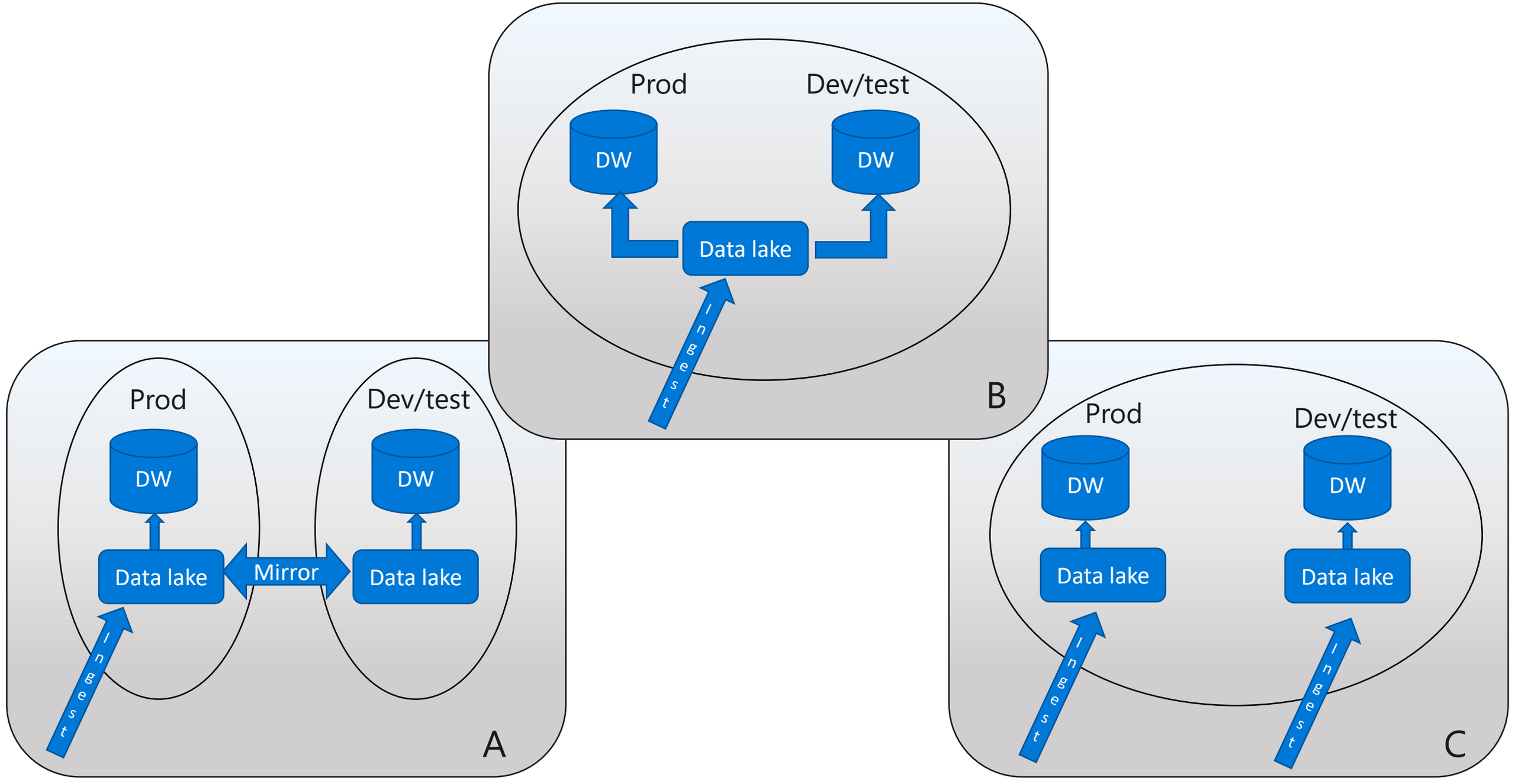
Database
FTP
File Storage

Firewall
AD control

Organize an Area in the lake

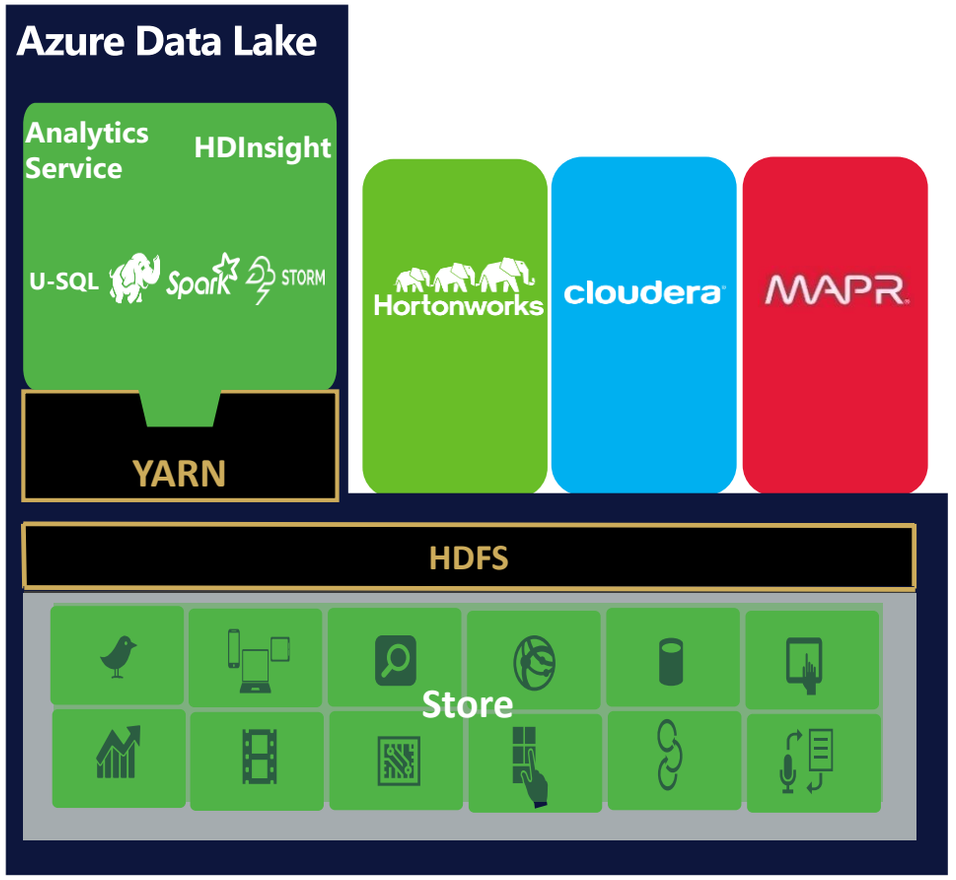
Example "by deparment"




Prod
Dev/test
DW
DW
Data lake
Ingest
B
Prod
Dev/test
DW
DW
Data lake
Mirror
Data lake
Ingest
A
Prod
Dev/test
DW
DW
Data lake
Data lake
Ingest
Ingest
C

What is Azure Data Lake then….

Built on Open Standards



Built on YARN

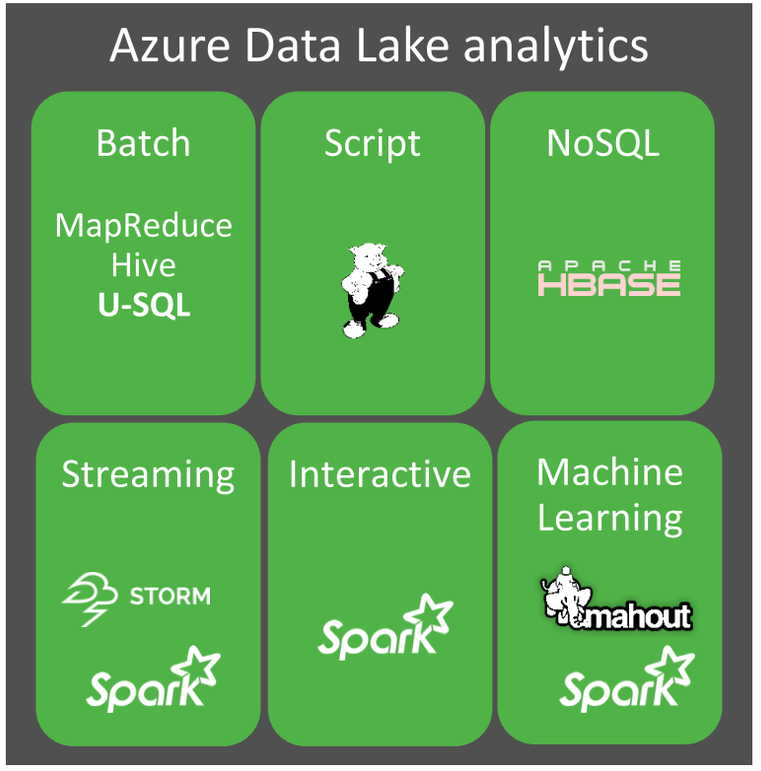
Store lets all HDFS compliant analytic applications connect to it like Hortonworks, Cloudera, and MapR

HDInsight is 100% Apache Hadoop

Microsoft continues to contribute tens of thousands of code and engineering hours to open source

Built using open standards

Any type of analytics: batch, streaming, interactive



Batch, interactive, streaming, machine learning

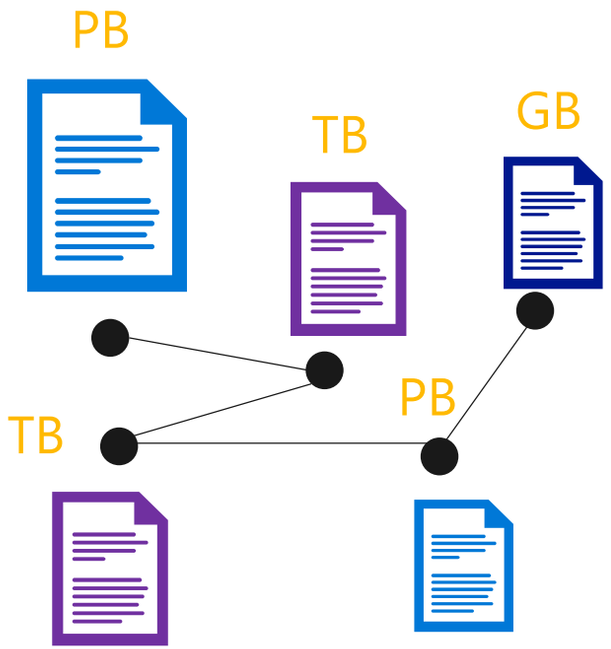
Allows for exploratory analytics over your data

Do analytics with Hadoop and Microsoft solutions

Cortana Analytics Suite

Any type of analytics

Data stored of any size, optimized for high performance



Store has no fixed limits on file sizes

(PB sized files)

Ultra fast read/write access

No code rewrite as you increase size of data stored

Optimized for large analytic systems: with massive throughput

Optimized for IOT with high volume of small writes

Useful for very large data or for real-time

Be productive with U-SQL, a simple and powerful language

U-SQL

```
@t = EXTRACT date string
           , time string
           , author string
           , tweet string
     FROM "/Input/MyTwitterHistory.csv"
     USING Extractors.Csv();

@res = SELECT author AS author
            , COUNT(*) AS tweetcount
       FROM @t
       GROUP BY author;

OUTPUT @res TO "/Output/MyTwitterAnalysis.csv"
ORDER BY tweetcount DESC
USING Outputters.Csv();
```

Simple and familiar, easily extensible

Unifies declarative nature of SQL with expressive power of C#

Familiar syntax to millions of .NET developers

Empower SQL/.NET developers with big data

Manage and secure your data assets

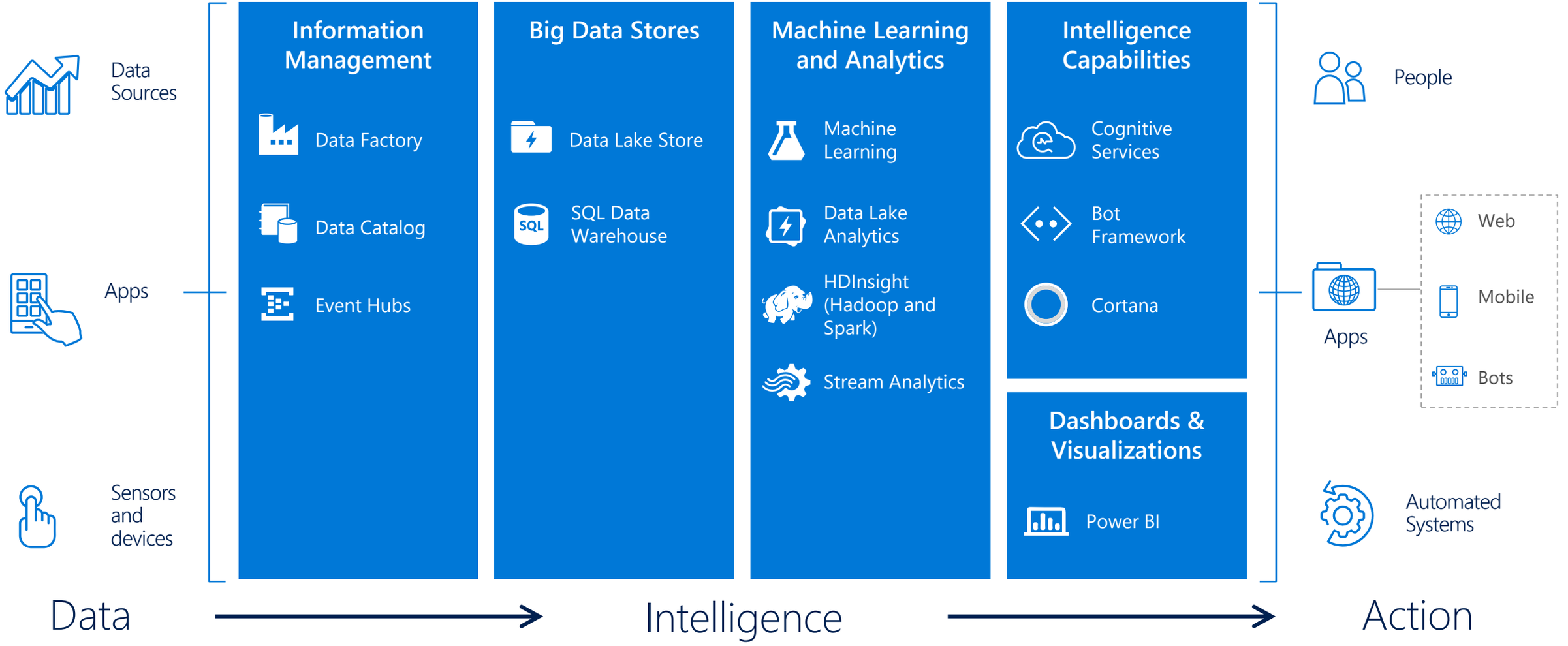
Auditing, alerting, access control - all from within a single web-based portal

Azure Active Directory integration for identity and access management

Use existing IT investment for security

Cortana Intelligence Suite

Microsoft Data solution Overview (CIS)



Microsoft

Thank you