Microsoft technológie pre BigData

Ľubomír Goryl

Solution Professional

Agenda

Súčasný stav

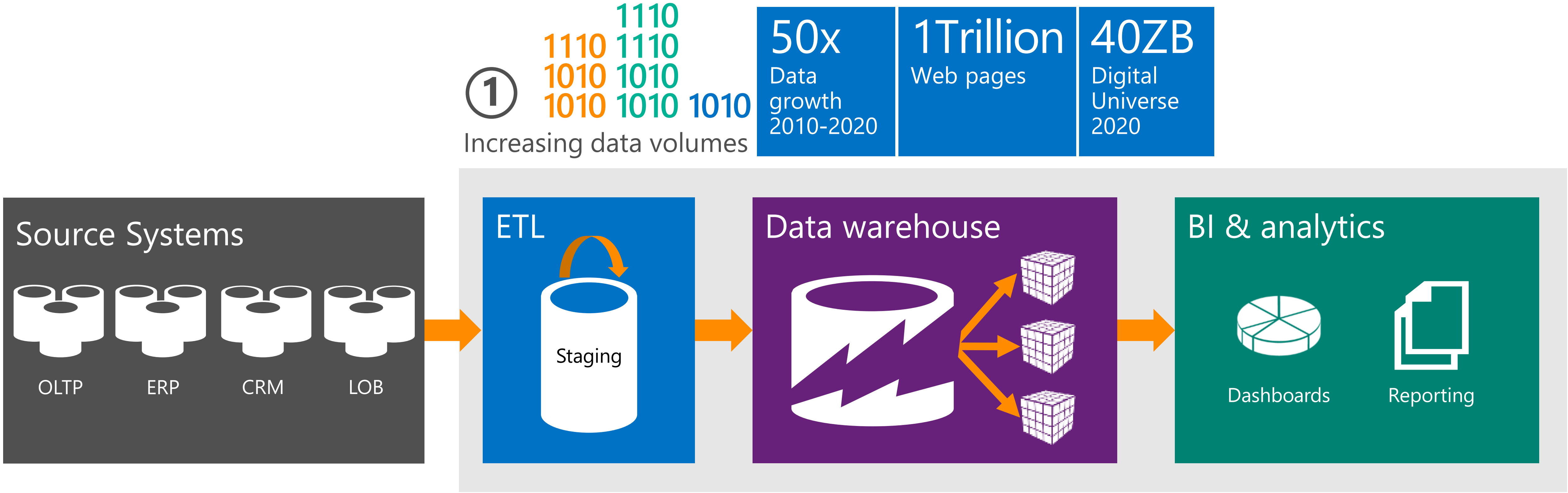
Hadoop = **HDInsight**

Moderný DWH = **APS**

Analytika = **AzureML, PowerBI**

Scenáre pre BigData

Tradičný prístup



1110
1110 1110
1010 1010
1010 1010 1010

①

Increasing data volumes

50x
Data growth 2010-2020

1Trillion
Web pages

40ZB
Digital Universe 2020

Source Systems

OLTP
ERP
CRM
LOB

ETL

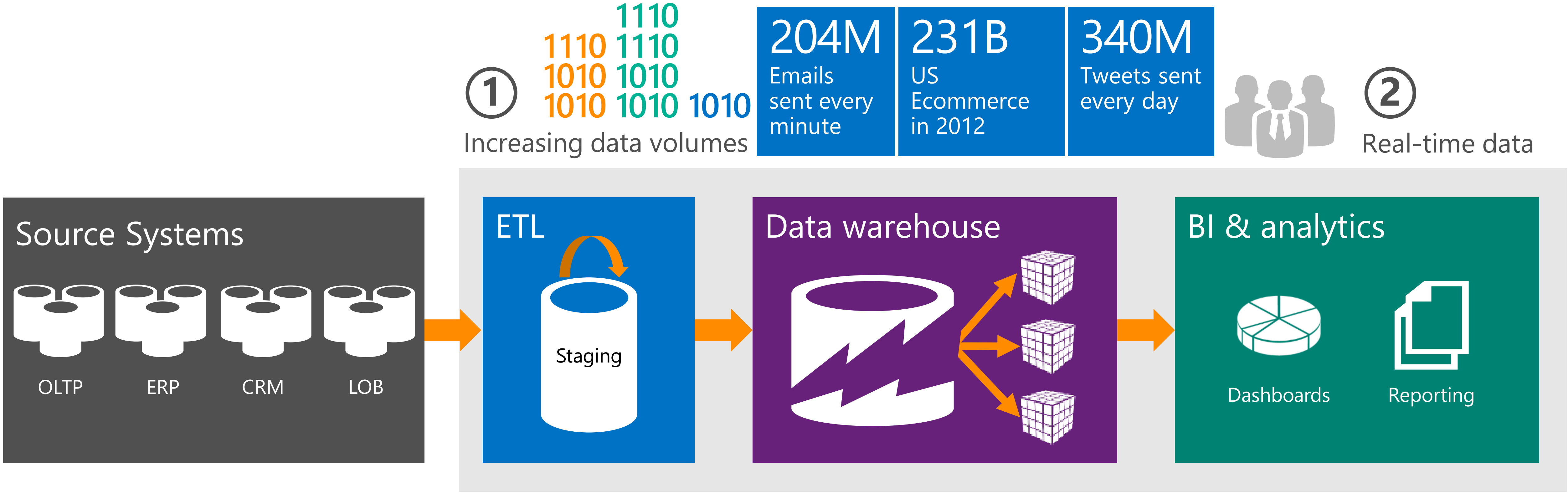
Staging

Data warehouse

BI & analytics

Dashboards
Reporting

Breaking points of traditional approach



1110
1110 1110
1010 1010
1010 1010 1010

① Increasing data volumes

204M Emails sent every minute

231B US Ecommerce in 2012

340M Tweets sent every day

② Real-time data

Source Systems

OLTP

ERP

CRM

LOB

ETL

Staging

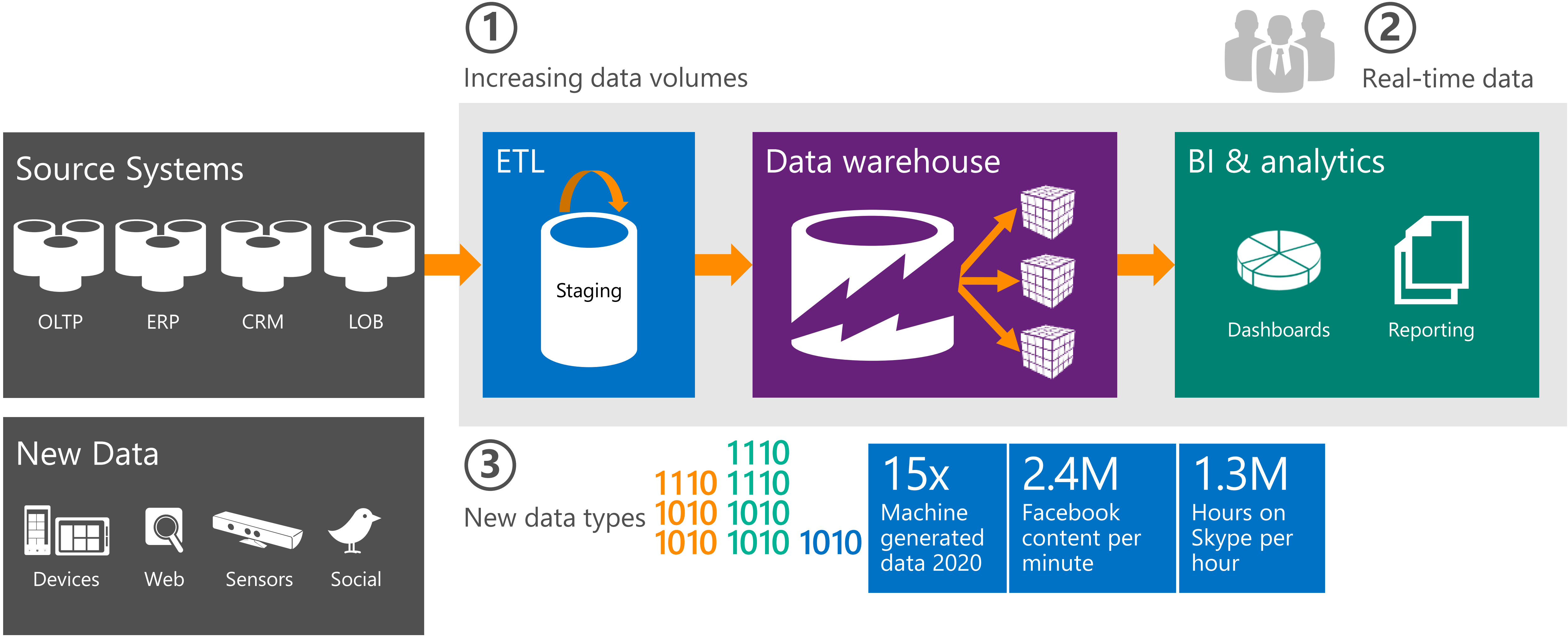
Data warehouse

BI & analytics

Dashboards

Reporting

Breaking points of traditional approach



① Increasing data volumes

② Real-time data

Source Systems

OLTP ERP CRM LOB

ETL

Staging

Data warehouse

BI & analytics

Dashboards Reporting

New Data

Devices Web Sensors Social

③ New data types

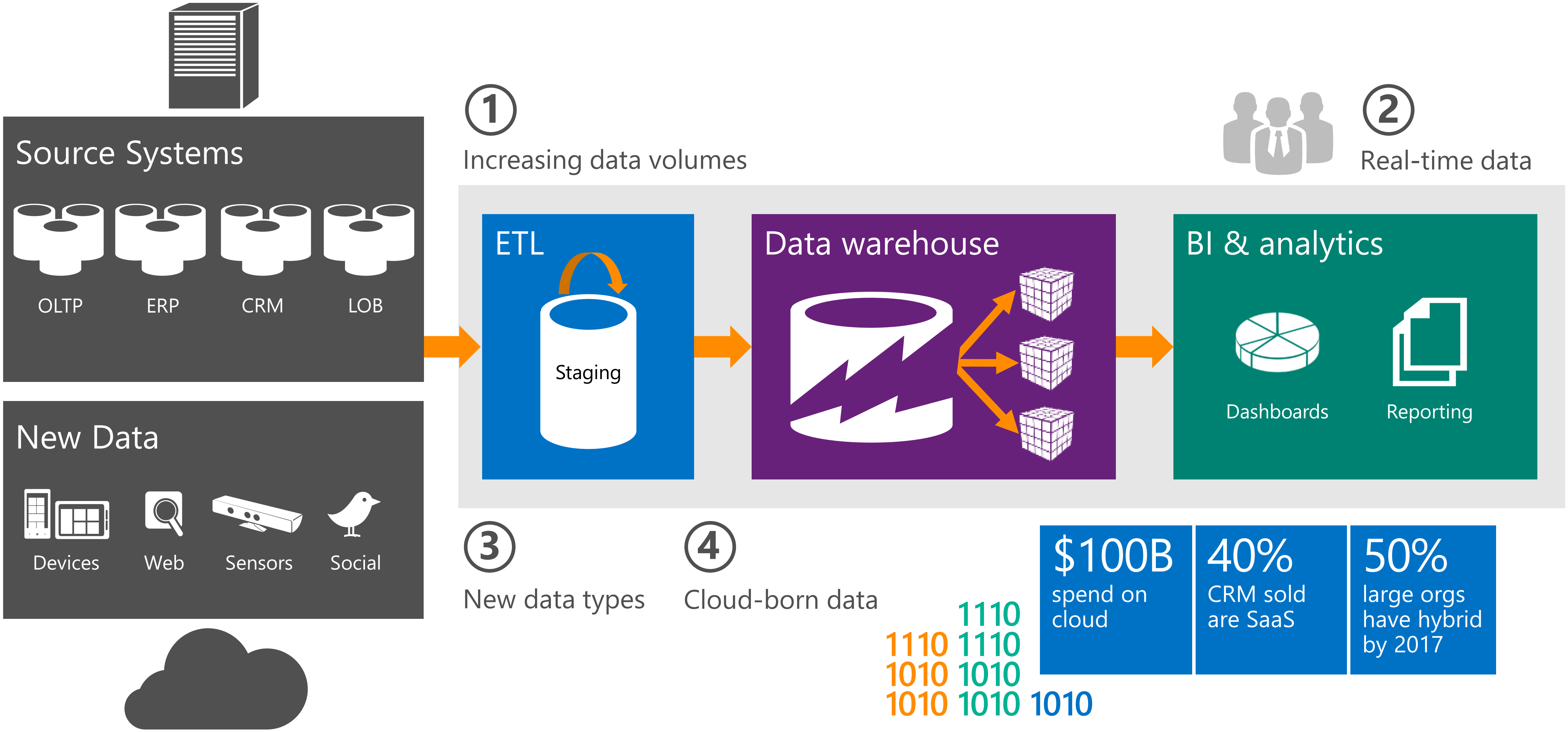
1110
1110 1110
1010 1010
1010 1010 1010

15x
Machine generated data 2020

2.4M
Facebook content per minute

1.3M
Hours on Skype per hour

Breaking points of traditional approach



Source Systems

OLTP ERP CRM LOB

New Data

Devices Web Sensors Social

① Increasing data volumes

② Real-time data

ETL

Staging

Data warehouse

BI & analytics

Dashboards Reporting

③ New data types

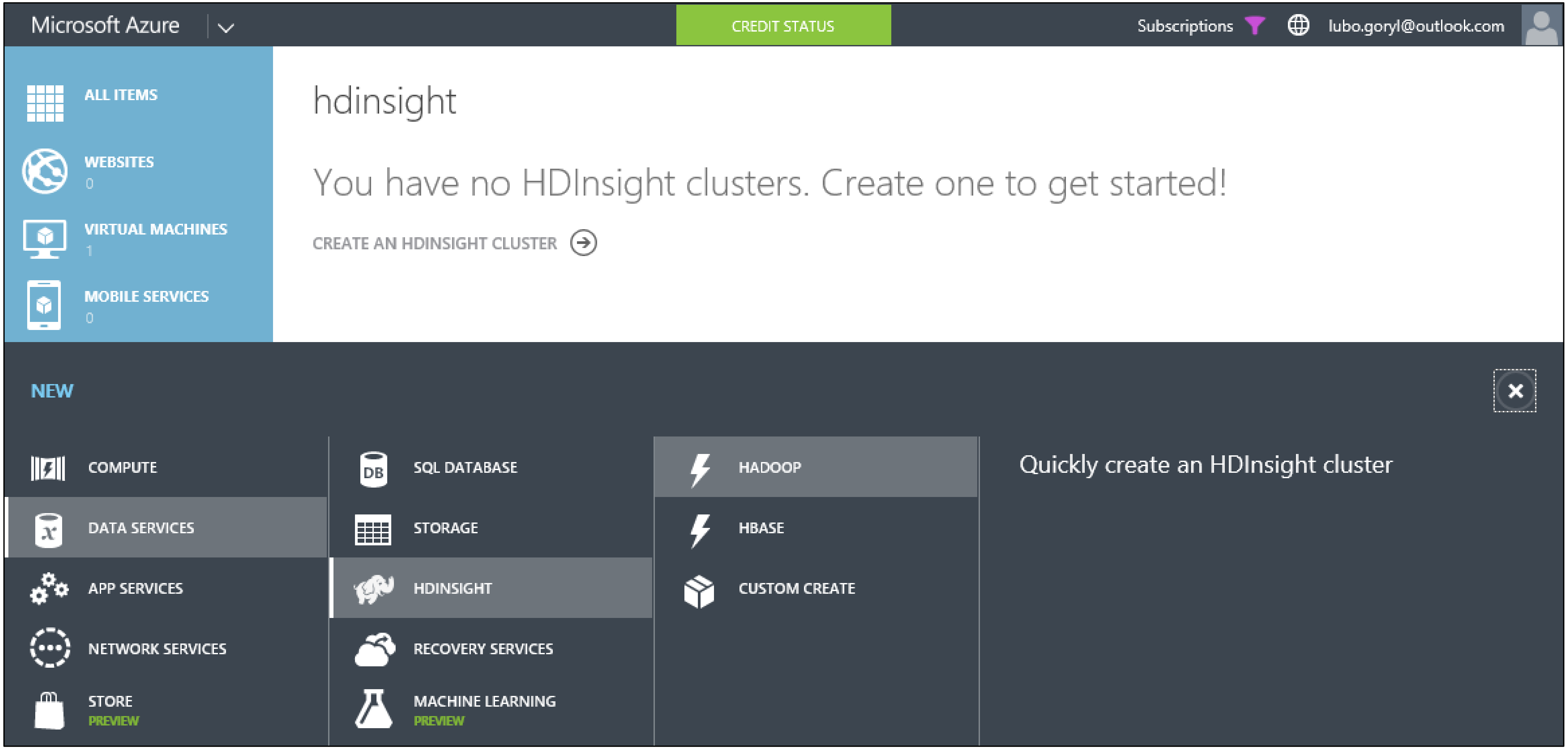
④ Cloud-born data

$100B spend on cloud

40% CRM sold are SaaS

50% large orgs have hybrid by 2017

HDInsight v Azure (Haddop)



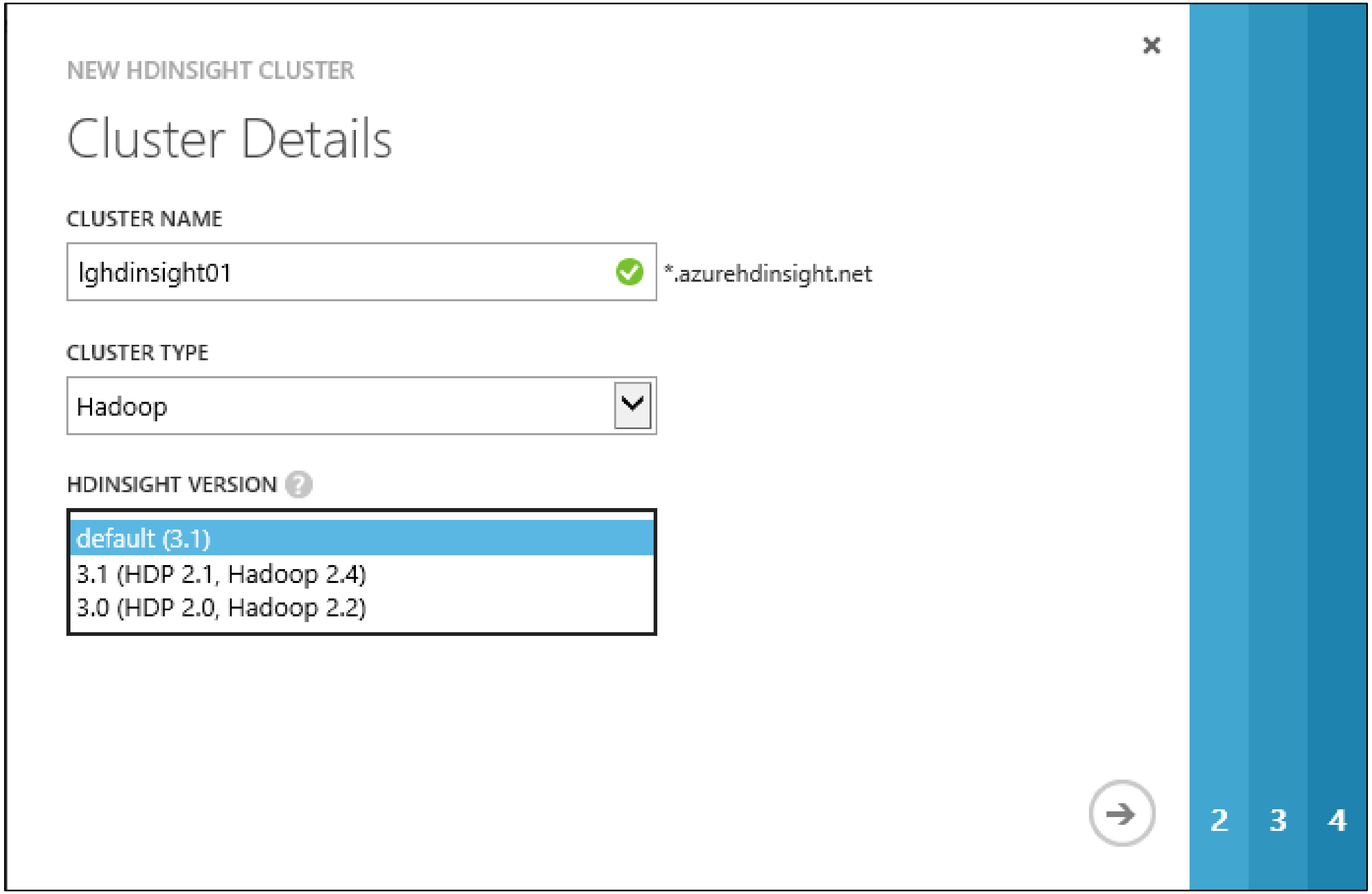
hdinsight

You have no HDInsight clusters. Create one to get started!

CREATE AN HDINSIGHT CLUSTER

Quickly create an HDInsight cluster

HDInsight – krok 1/3



NEW HDINSIGHT CLUSTER

Cluster Details

CLUSTER NAME

lghdinsight01 *.azurehdinsight.net

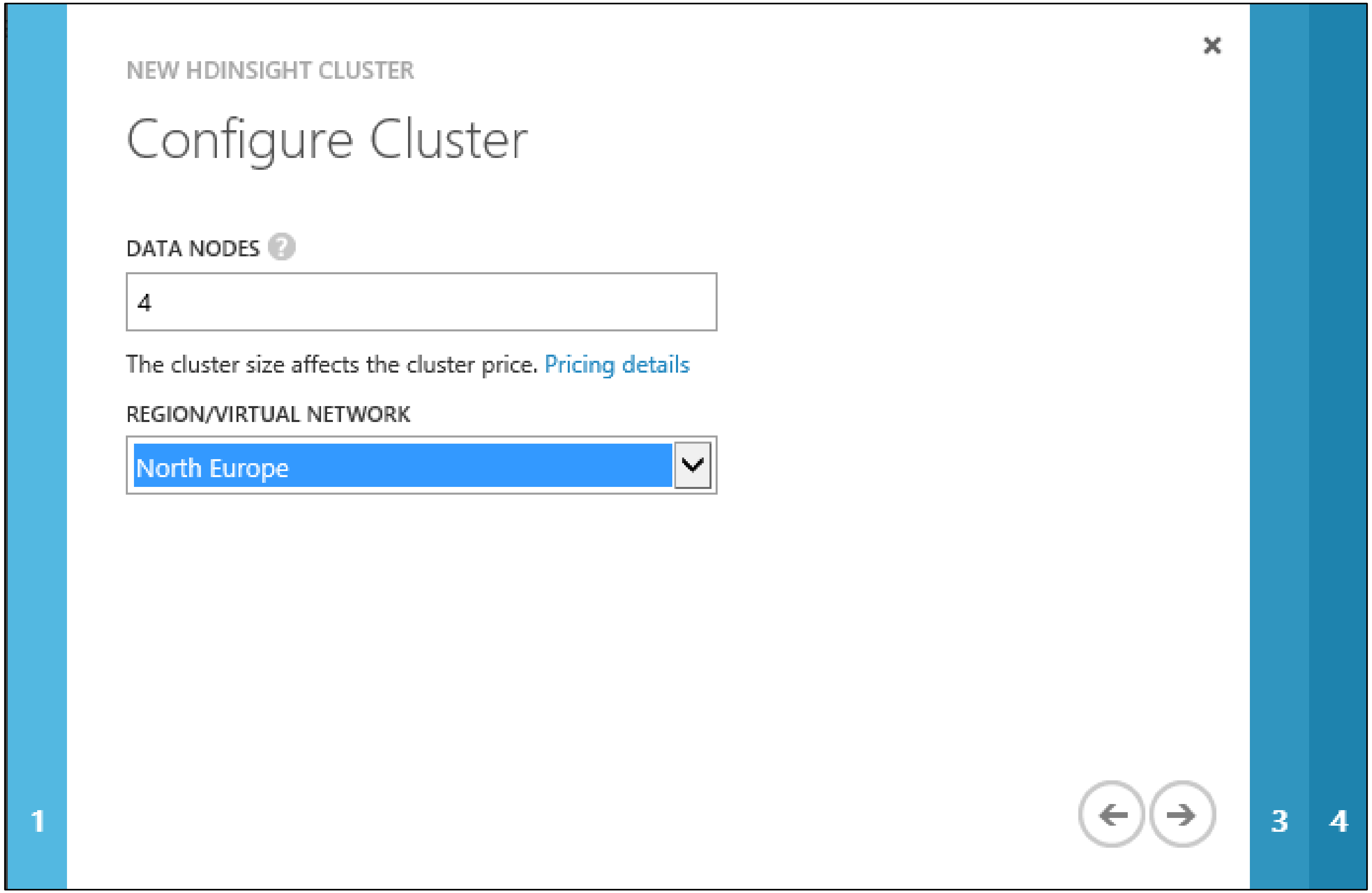
CLUSTER TYPE

Hadoop

HDINSIGHT VERSION

default (3.1)
3.1 (HDP 2.1, Hadoop 2.4)
3.0 (HDP 2.0, Hadoop 2.2)

HDInsight – krok 2/3



NEW HDINSIGHT CLUSTER

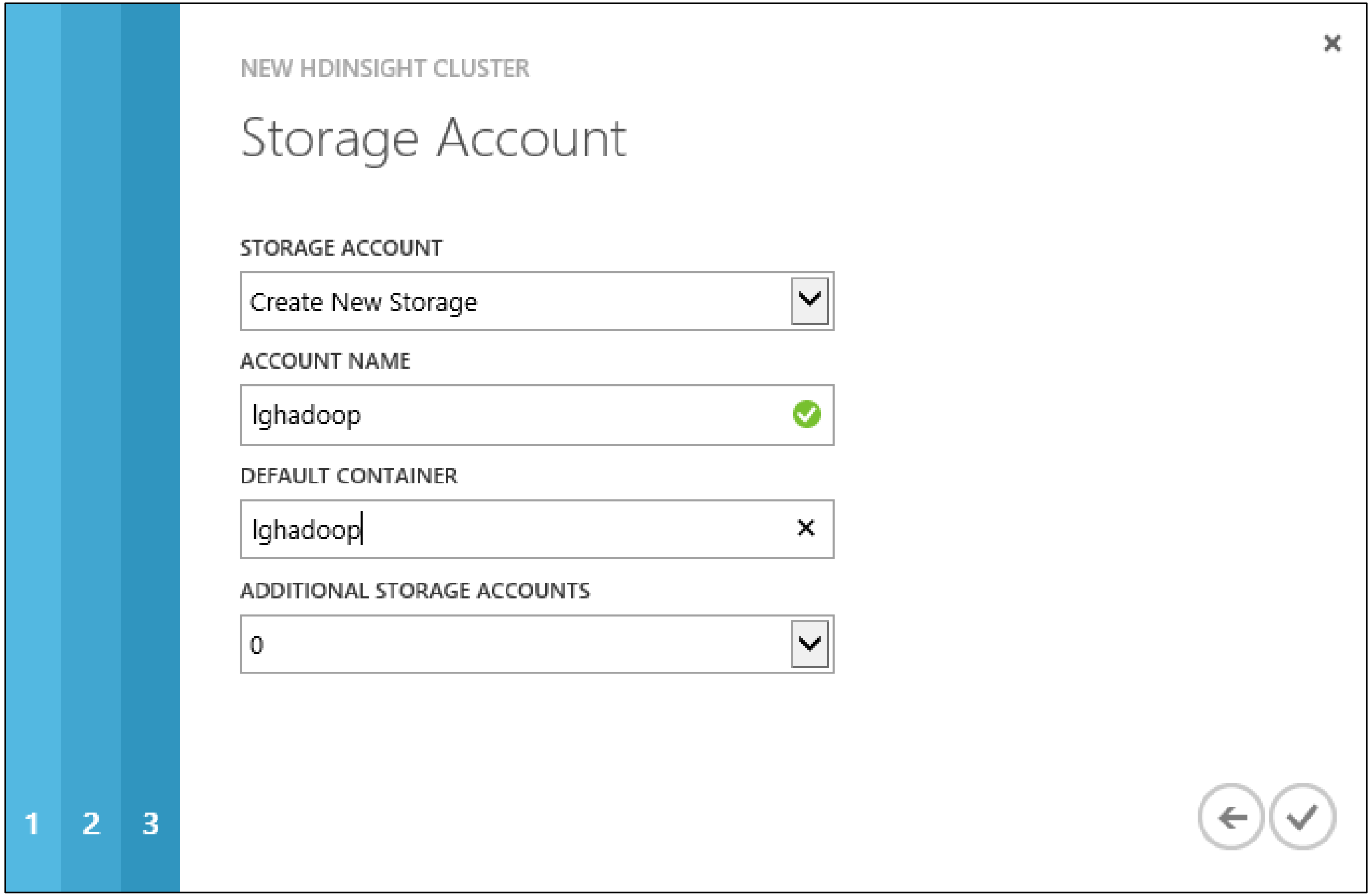
Configure Cluster

DATA NODES: 4

The cluster size affects the cluster price. Pricing details

REGION/VIRTUAL NETWORK: North Europe

HDInsight – krok 3/3



NEW HDINSIGHT CLUSTER

Storage Account

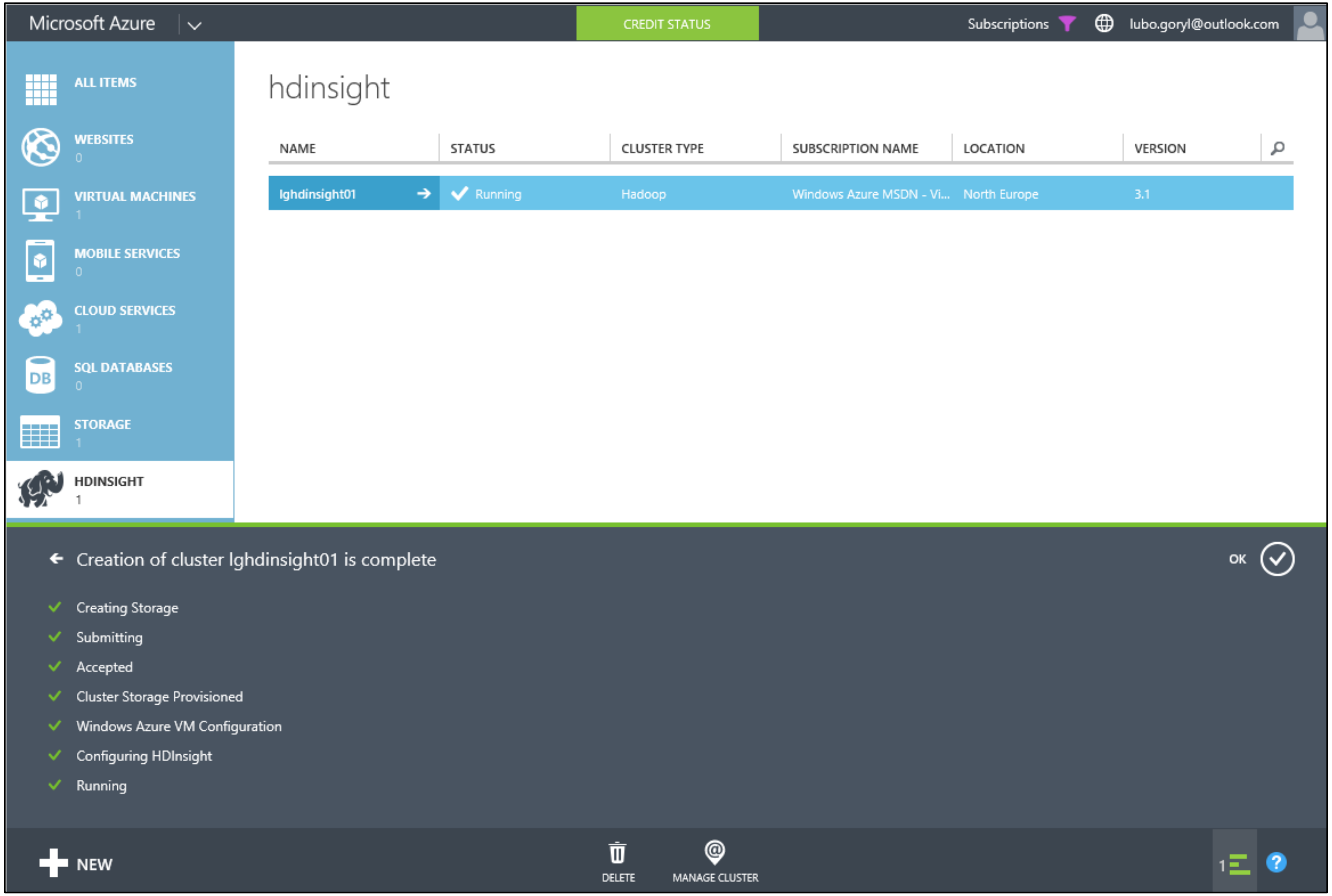
STORAGE ACCOUNT: Create New Storage

ACCOUNT NAME: lghadoop ✓

DEFAULT CONTAINER: lghadoop

ADDITIONAL STORAGE ACCOUNTS: 0

15' …. Hadoop cluster - running



hdinsight

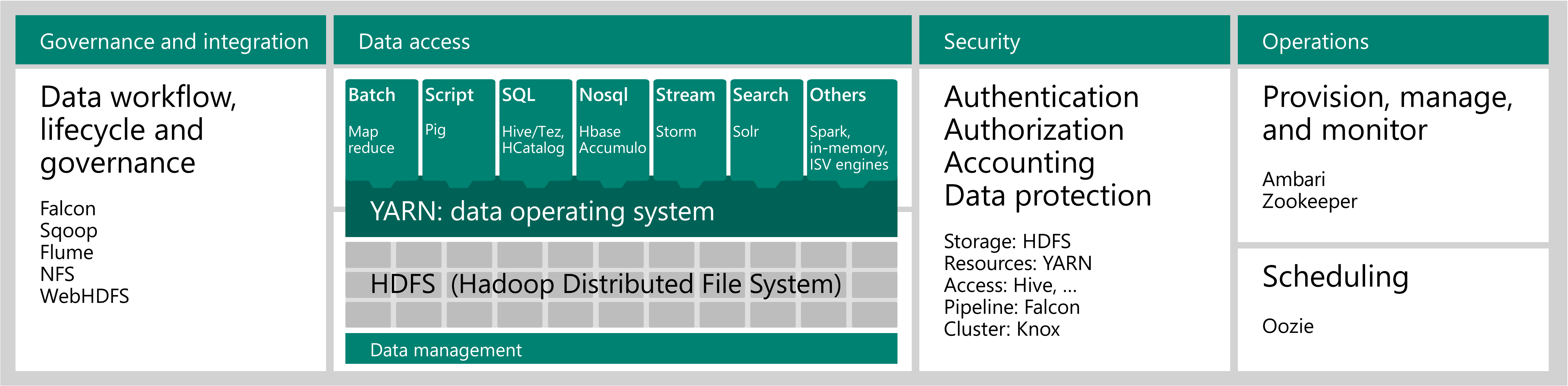
NAME	STATUS	CLUSTER TYPE	SUBSCRIPTION NAME	LOCATION	VERSION
lghdinsight01	✓ Running	Hadoop	Windows Azure MSDN - Vi...	North Europe	3.1

Creation of cluster lghdinsight01 is complete

- ✓ Creating Storage
- ✓ Submitting
- ✓ Accepted
- ✓ Cluster Storage Provisioned
- ✓ Windows Azure VM Configuration
- ✓ Configuring HDInsight
- ✓ Running

Hadoop is a platform with portfolio of projects

- Governed by Apache Software Foundation (ASF)
- Comprises core services of MapReduce, HDFS, and YARN
- In addition to the core, includes functions across:
  - Data services which allow you to manipulate and move data (Hive, HBase, Pig, Flume, Sqoop)
  - Operational services which help manage the cluster (Ambari, Falcon, and Oozie)



Governance and integration

Data workflow, lifecycle and governance

Falcon
Sqoop
Flume
NFS
WebHDFS

Data access

Batch Map reduce
Script Pig
SQL Hive/Tez, HCatalog
Nosql Hbase Accumulo
Stream Storm
Search Solr
Others Spark, in-memory, ISV engines

YARN: data operating system

HDFS (Hadoop Distributed File System)

Data management

Security

Authentication
Authorization
Accounting
Data protection

Storage: HDFS
Resources: YARN
Access: Hive, …
Pipeline: Falcon
Cluster: Knox

Operations

Provision, manage, and monitor

Ambari
Zookeeper

Scheduling

Oozie

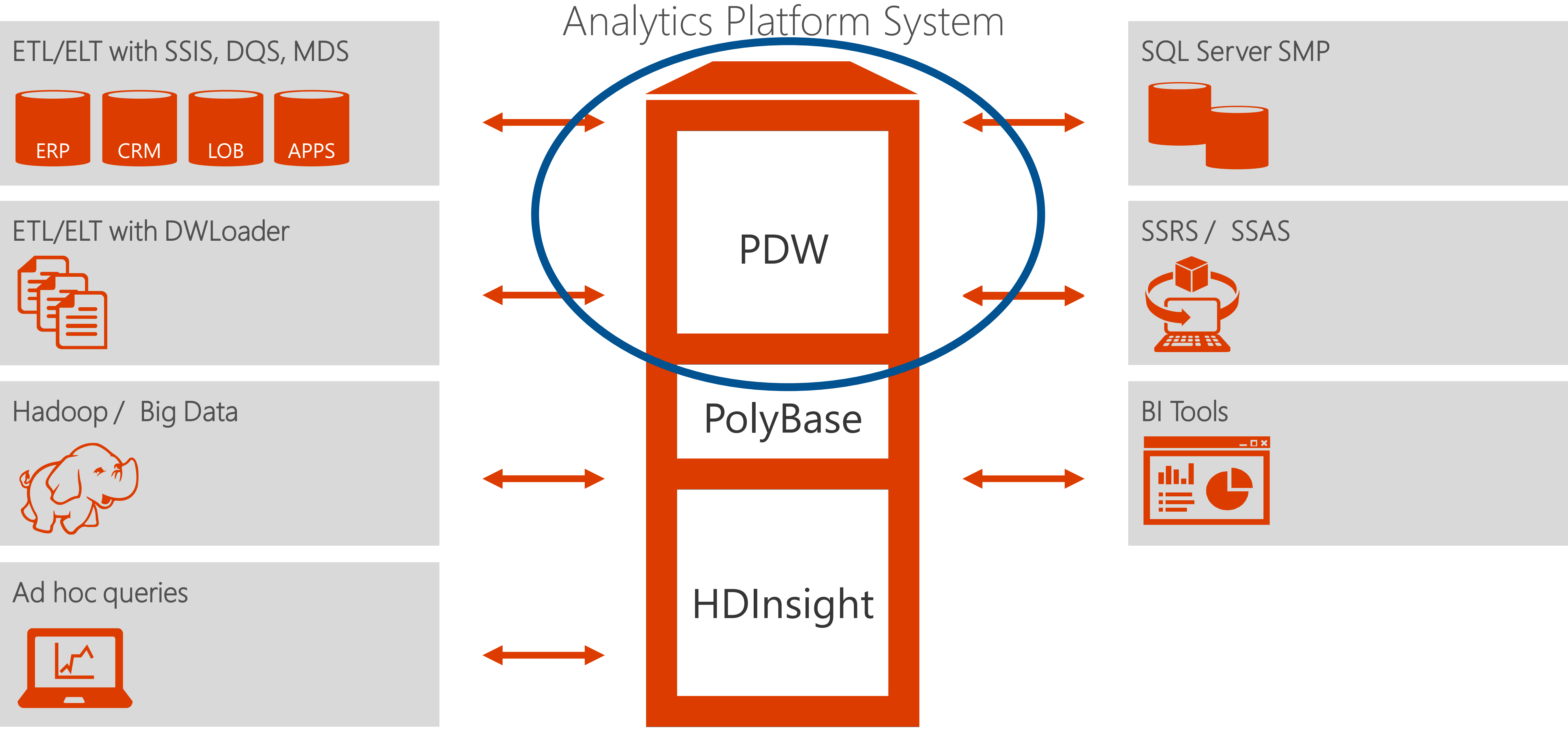
A Hadoop distribution is a package of projects

Tested for consistency across entire package

	Hadoop and YARN	Tez	Pig	Hive and HCatalog	HBase	Phoenix	Accumulo	Storm	Mahout	Solr	Falcon	Sqoop	Flume	Ambari	Oozie	Zookeeper	Knox
HDP 2.1 April 2014	2.4.0	0.4.0	0.12.1	0.13.0	0.98.0	4.0.0	1.5.1	0.9.1	0.9.0	4.7.2	0.5.0	1.4.4	1.4.0	1.5.1	4.0.0	3.4.5	.0.4.0
HDP 2.0 October 2013	2.2.0		0.12.0	0.12.0	0.96.1				0.8.0			1.4.4	1.3.0	1.4.4	3.3.2	3.4.5	.0.4.0
HDP 1.3 May 2013	1.1.2		011.0	0.11.0	0.94.6				0.7.0			1.4.3	1.3.1	1.2.5	3.3.2	3.4.5	.0.4.0
	Data management	Data access									Governance and integration			Operations			Security

Microsoft Analytics Platform System

Appliance pre moderný Datawarehouse



ETL/ELT with SSIS, DQS, MDS

ERP

CRM

LOB

APPS

ETL/ELT with DWLoader

Hadoop / Big Data

Ad hoc queries

Analytics Platform System

PDW

PolyBase

HDInsight

SQL Server SMP

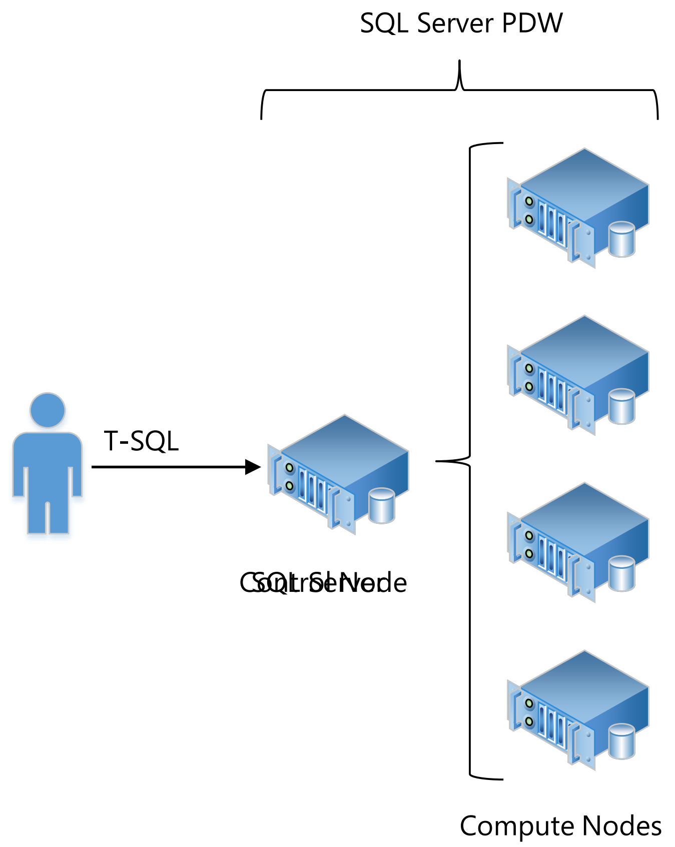
SSRS / SSAS

BI Tools

Čo je Parallel Data Warehouse?

- Shared-nothing parallel database system
  - » Massively parallel processing (MPP)
  - » A “Control” server that accepts user queries, generates a plan, and distributes operations in parallel to compute nodes
  - » Multiple “Compute” servers running SQL Server
  - » A “Management” server for administering the system
  - » A “Data Movement Service” that facilitates parallel SQL operations
- Delivered as an appliance
  - » Balanced and pre-configured software and industry standard hardware from ***HP***
  - » Single Call Support
  - » Fastest Time to Market
  - » Scales from 2 to 56 Nodes

SQL Server Parallel Data Warehouse



SQL Server PDW

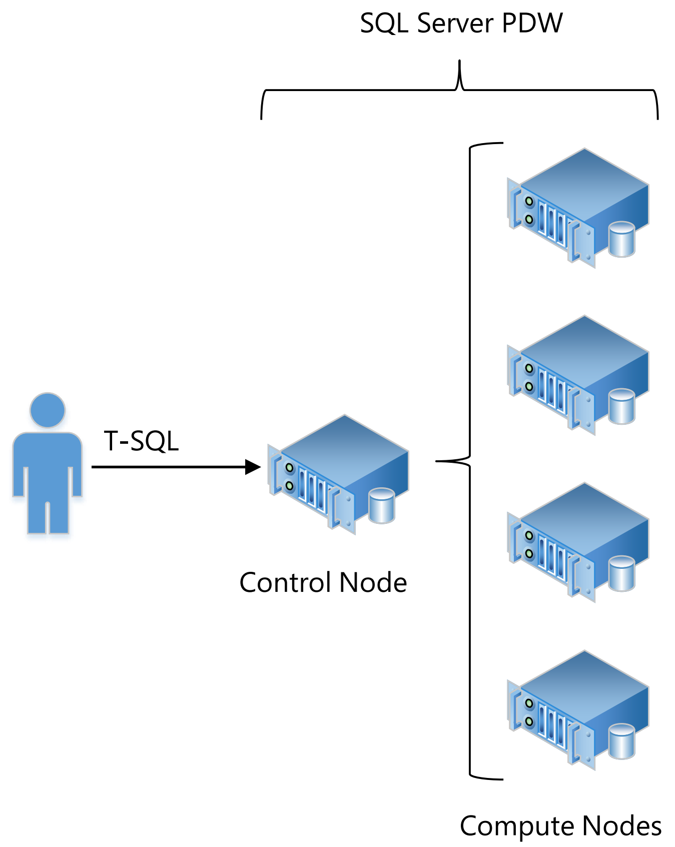
T-SQL

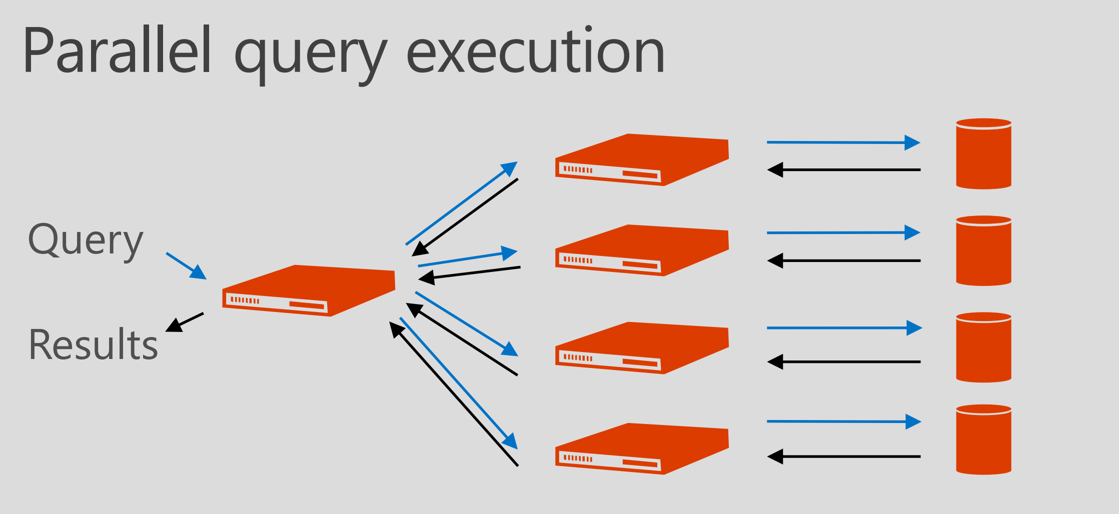
Control Node

SQL Server

Compute Nodes

SQL Server Parallel Data Warehouse





HP ConvergedSystem 300 for Microsoft Analytics Platform **Base unit**

FDR InfiniBand
2 x HP 5120 EI switches

InfiniBand (data network) and Ethernet (management network) connectivity

Passive server (#1)
Passive server (#2)
Optional passive server (#3)

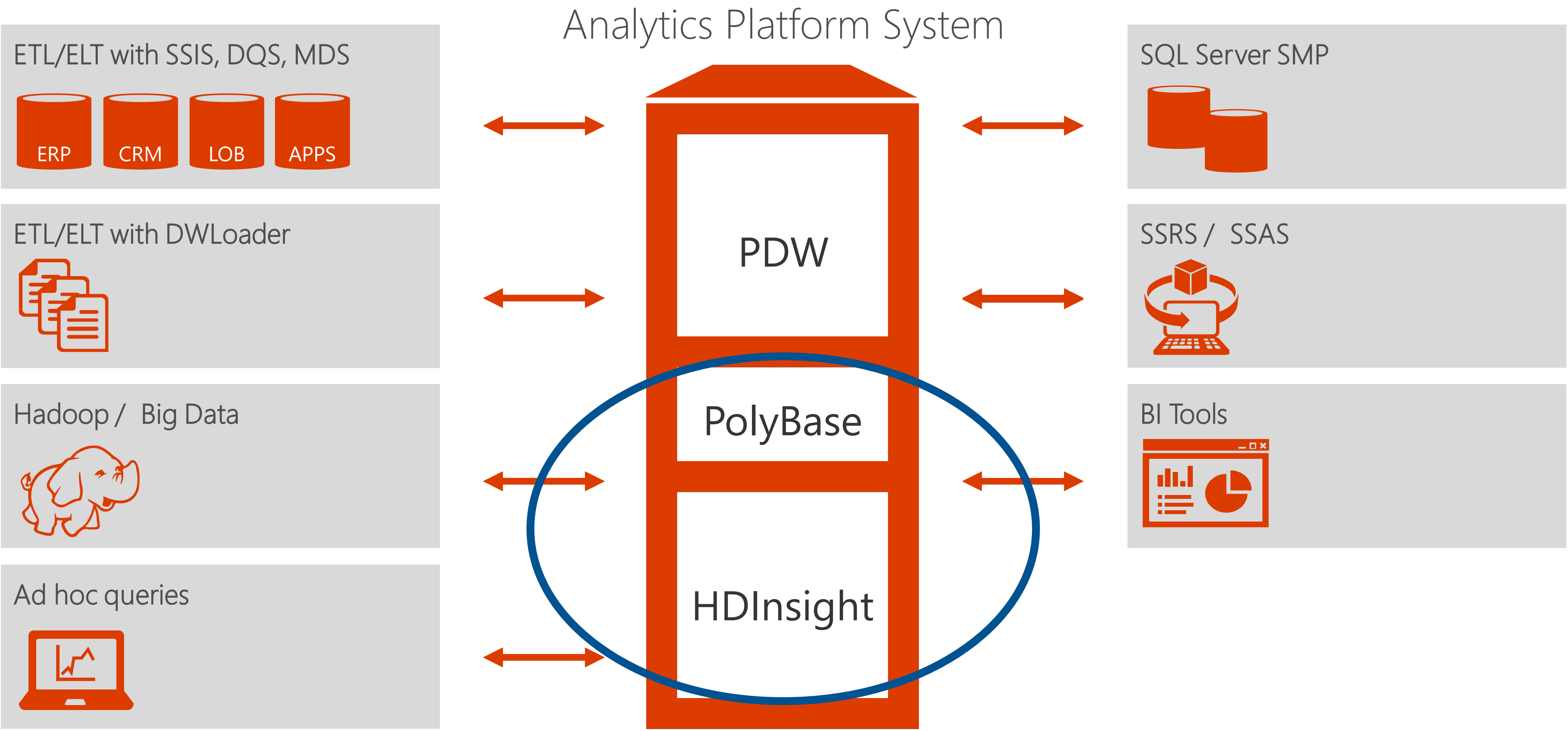
Passive Server Block (2): Virtualized control and management node; failover node for HA

Base unit

Active Server Block (2) and Storage Block (1)

Microsoft Analytics Platform System

Appliance pre moderný datawarehouse



Analytics Platform System

ETL/ELT with SSIS, DQS, MDS

ERP

CRM

LOB

APPS

ETL/ELT with DWLoader

Hadoop / Big Data

Ad hoc queries

PDW

PolyBase

HDInsight

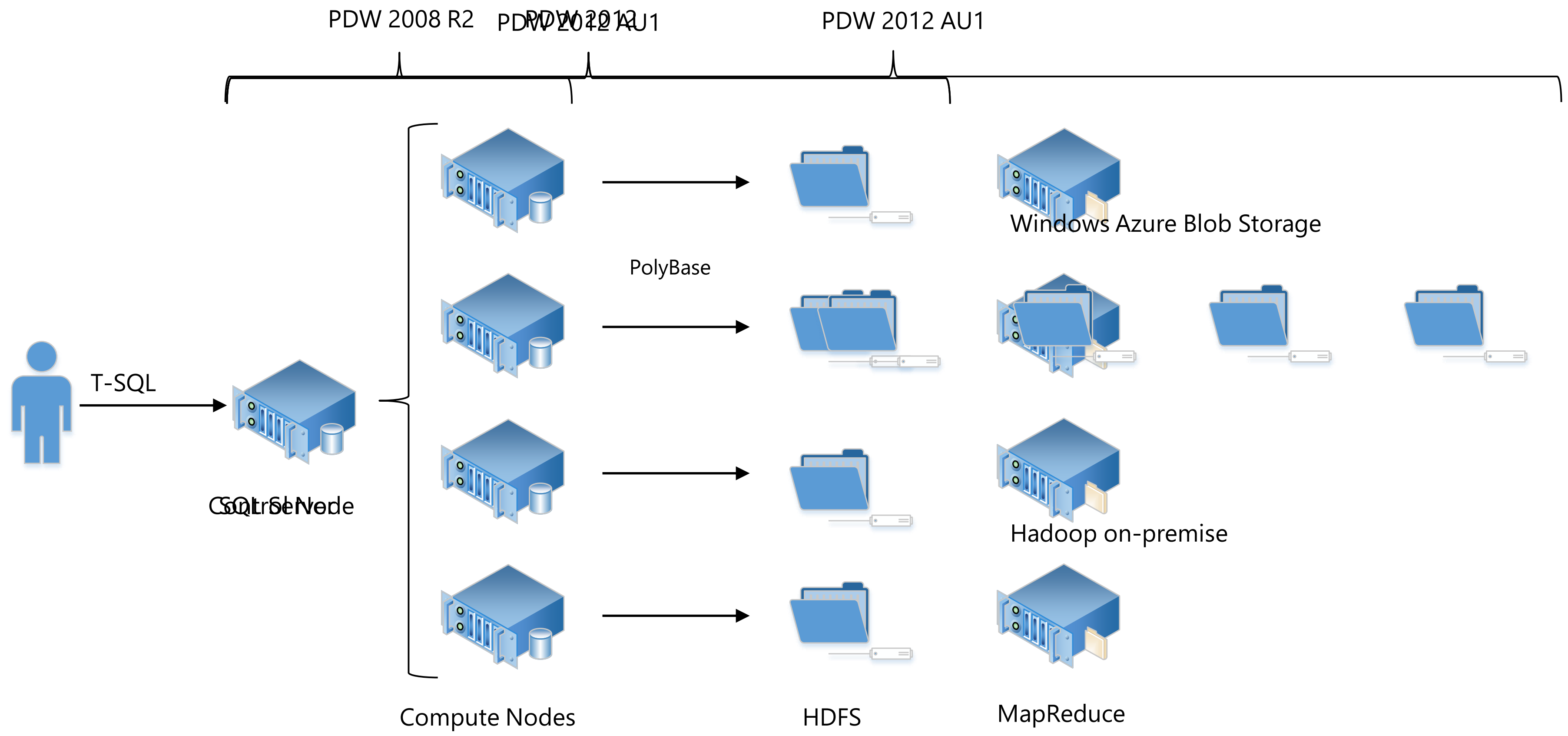
SQL Server SMP

SSRS / SSAS

BI Tools

SQL Server Parallel Data Warehouse – Appliance Update 1

Extending the distributed Data Warehouse further



PDW 2008 R2

PDW 2012

PDW 2012 AU1

PDW 2012 AU1

T-SQL

Control Node

SQL Server

PolyBase

Windows Azure Blob Storage

Hadoop on-premise

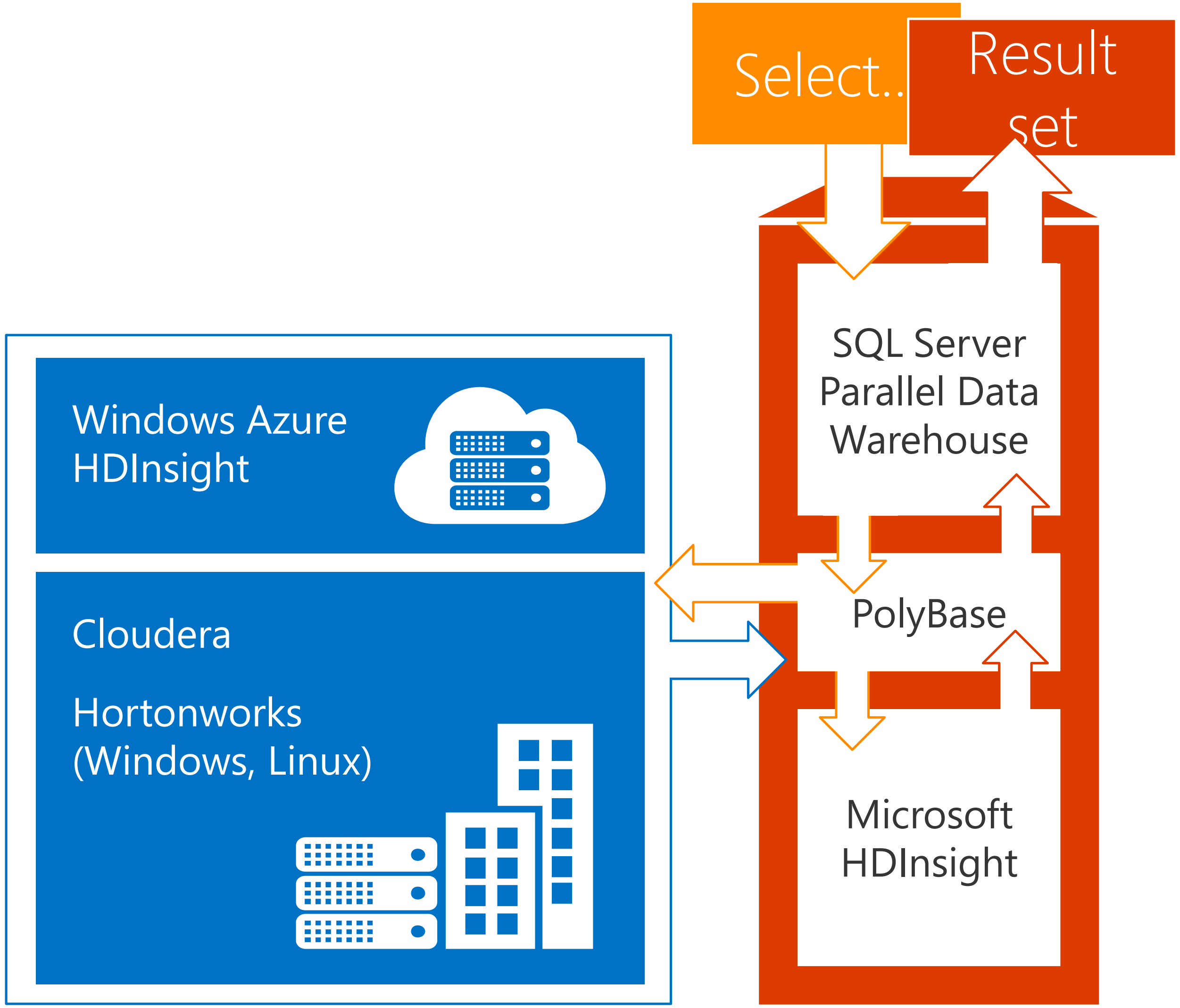
Compute Nodes

HDFS

MapReduce

Query Hadoop data with T-SQL using PolyBase

Bringing the worlds or big data and the data warehouse together for users and IT


Select..

Result set

Windows Azure HDInsight

Cloudera

Hortonworks (Windows, Linux)

SQL Server Parallel Data Warehouse

PolyBase

Microsoft HDInsight

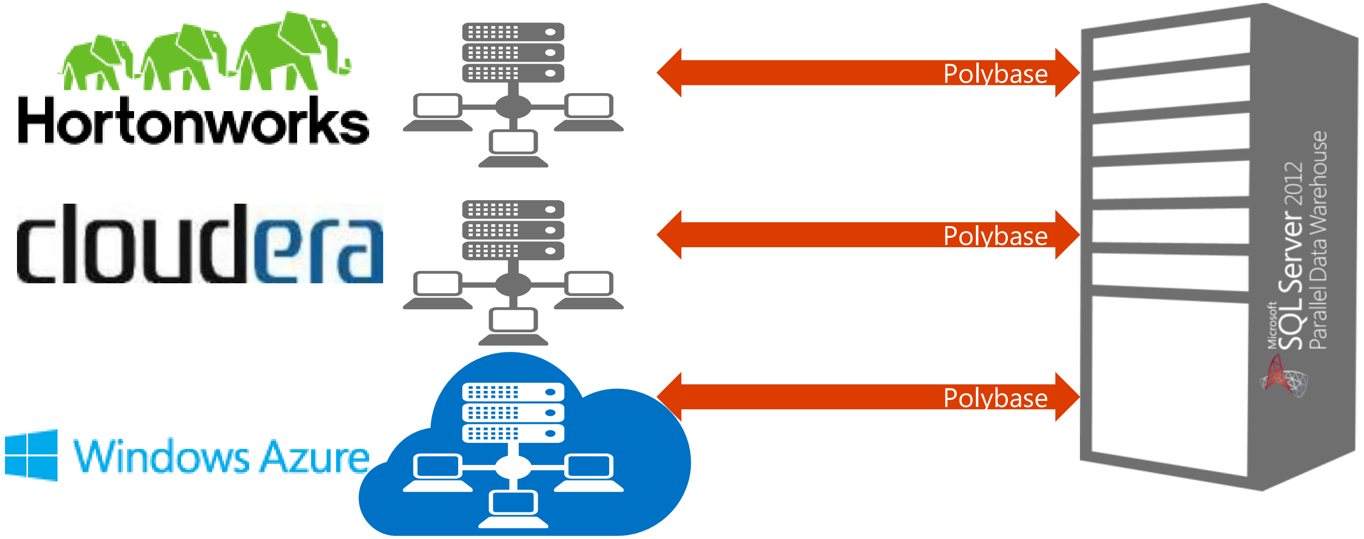
Single T-SQL query model for PDW and Hadoop with rich features of T-SQL including joins without ETL

Leverages the power of MPP to enhance query execution performance

Supports Windows Azure HDInsight to enable new hybrid cloud scenarios

Query non-Microsoft Hadoop distributions such as Hortonworks and Cloudera

Access Hadoop on different cluster (cloud or on premise)



Hortonworks Data Platform (HDP) on Windows or Linux

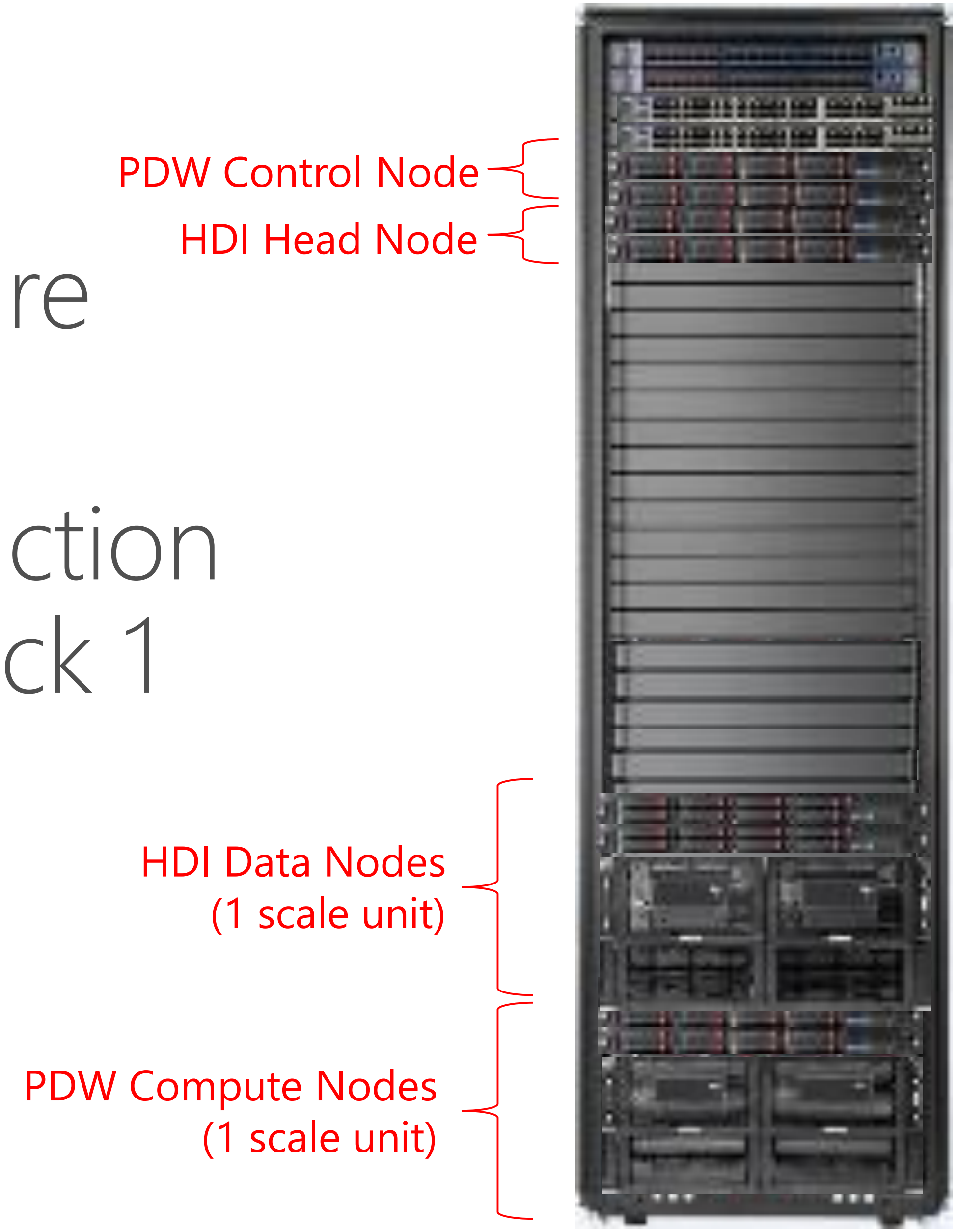
HDInsight on Azure (rebranded HDP)

HDInsight is the Microsoft branded Hortonworks Data Platform

- **We made it work on Windows**
- **We brougth SystemCenter support to it**

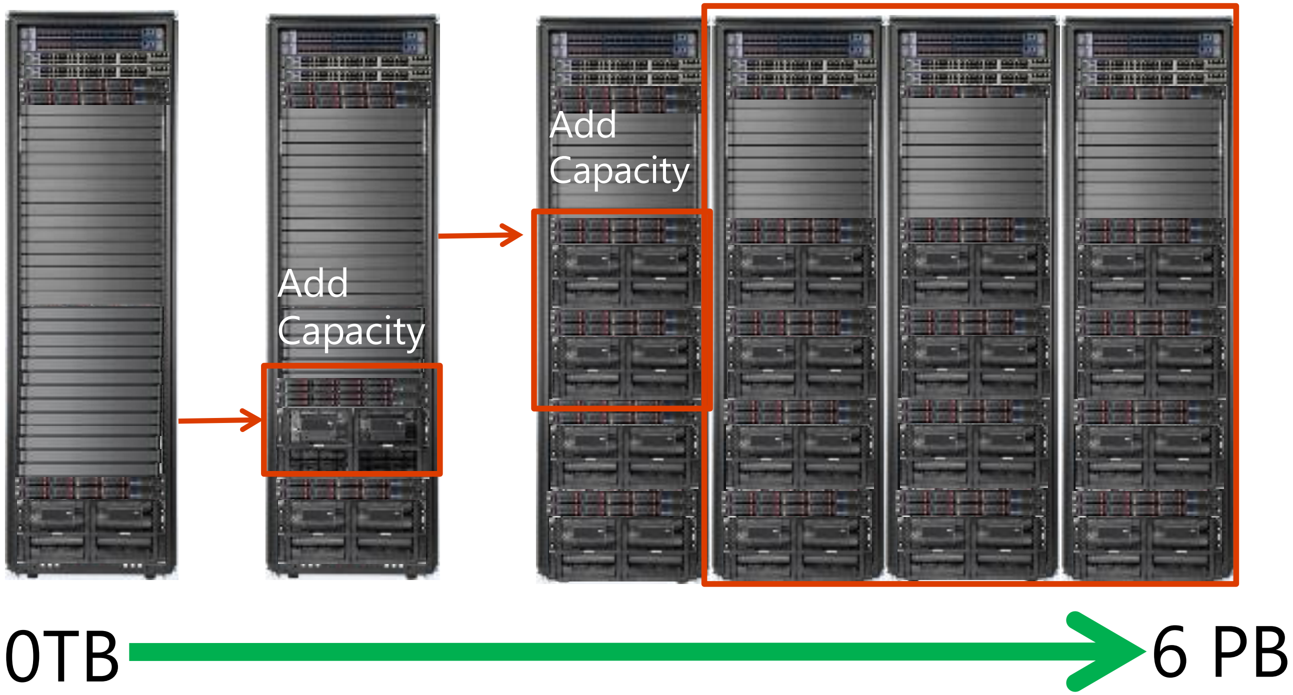
Hardware topology overview

- Uses the same PDW hardware and topology
- The key difference is introduction of 2 additional servers on rack 1 for the HDI Head Node
  - 1 active server and 1 fail over server.



Seamlessly add capacity

Scale from a Quarter Rack with 2 Compute Nodes up to 56 Compute Nodes!



Smallest (0TB) To Largest (6PB)

- Start small with a few Terabyte warehouse
- From 2 compute nodes to 56 compute nodes
- 1 quarter rack up to 7 full racks
- Add capacity up to 6 Petabytes

Start Small And Grow

Largest Warehouse

PB

Minimal Downtime

Bringing Hadoop to a billion people

Excel as the BI tool for everyone

1 Billion
Microsoft Office users

- Connect to HDInsight
- Analyze
- Visualize

Power BI for collaboration & new experiences

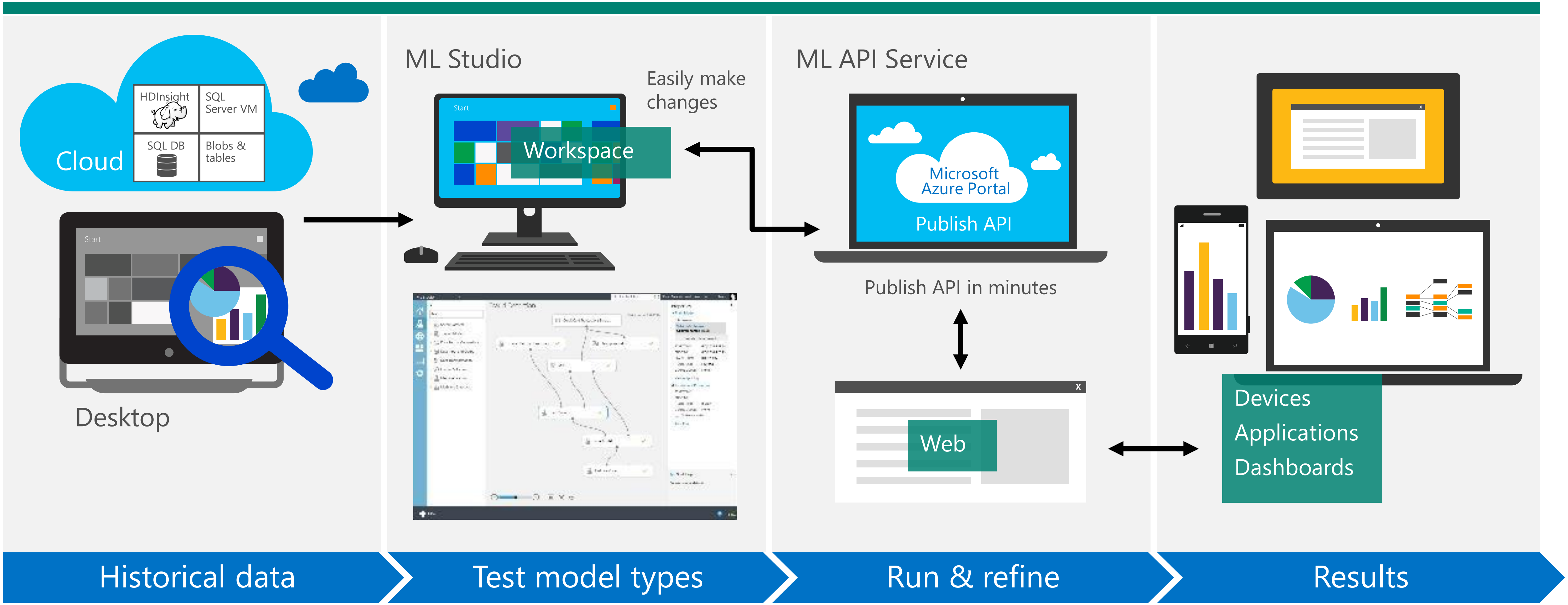
Office 365 is our fastest-growing commercial product ever

- Share
- Ask
- Access

Scalable, manageable, trusted

Making advanced analytics accessible to Hadoop

Microsoft Azure Machine Learning



Cloud

HDInsight

SQL Server VM

SQL DB

Blobs & tables

Desktop

ML Studio

Workspace

Easily make changes

ML API Service

Microsoft Azure Portal

Publish API

Publish API in minutes

Web

Devices

Applications

Dashboards

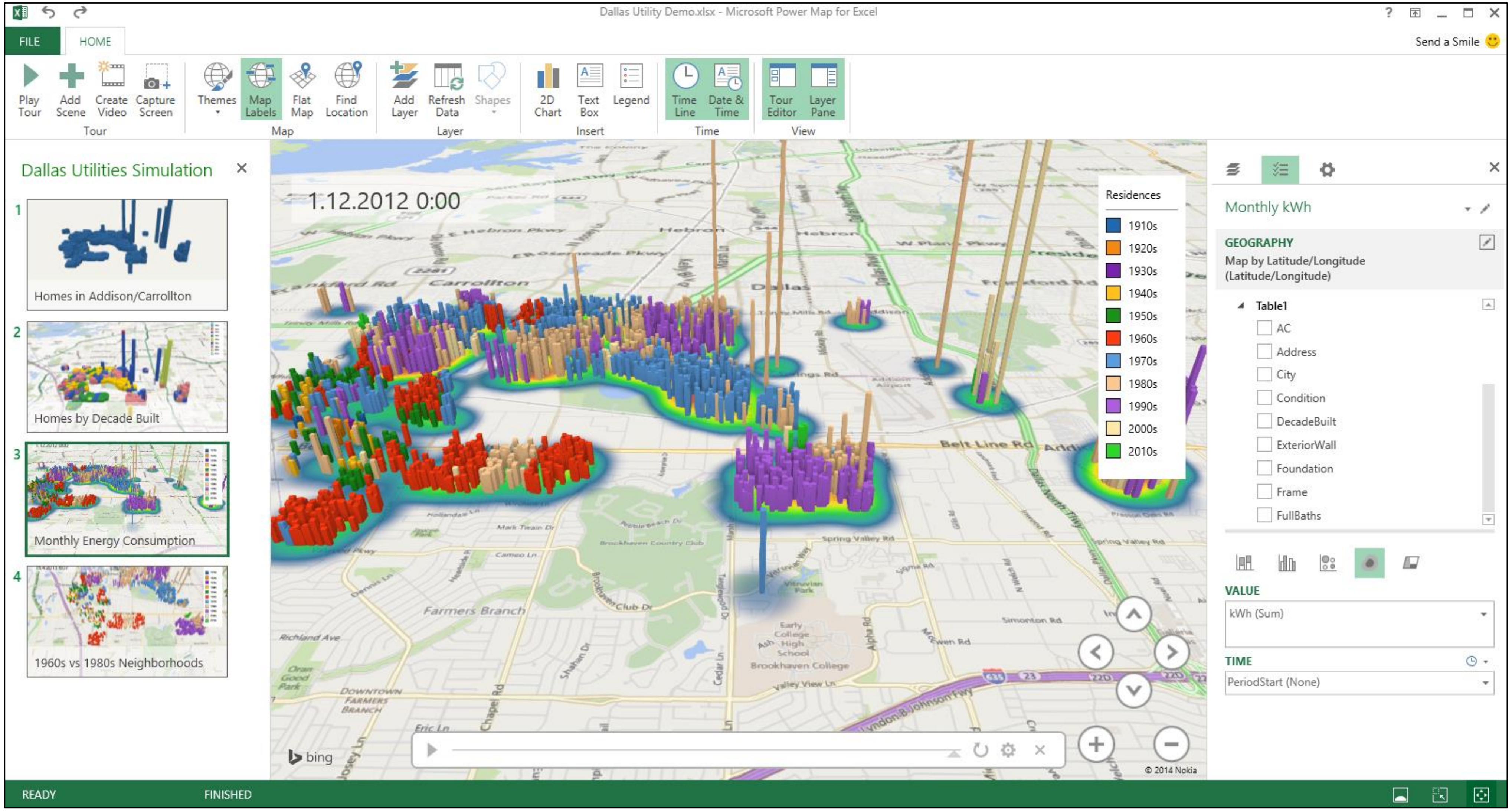
Historical data

Test model types

Run & refine

Results

PowerBI, Excel PowerMap, PowerQuery, …



Dallas Utility Demo.xlsx - Microsoft Power Map for Excel

Monthly kWh

GEOGRAPHY
Map by Latitude/Longitude
(Latitude/Longitude)

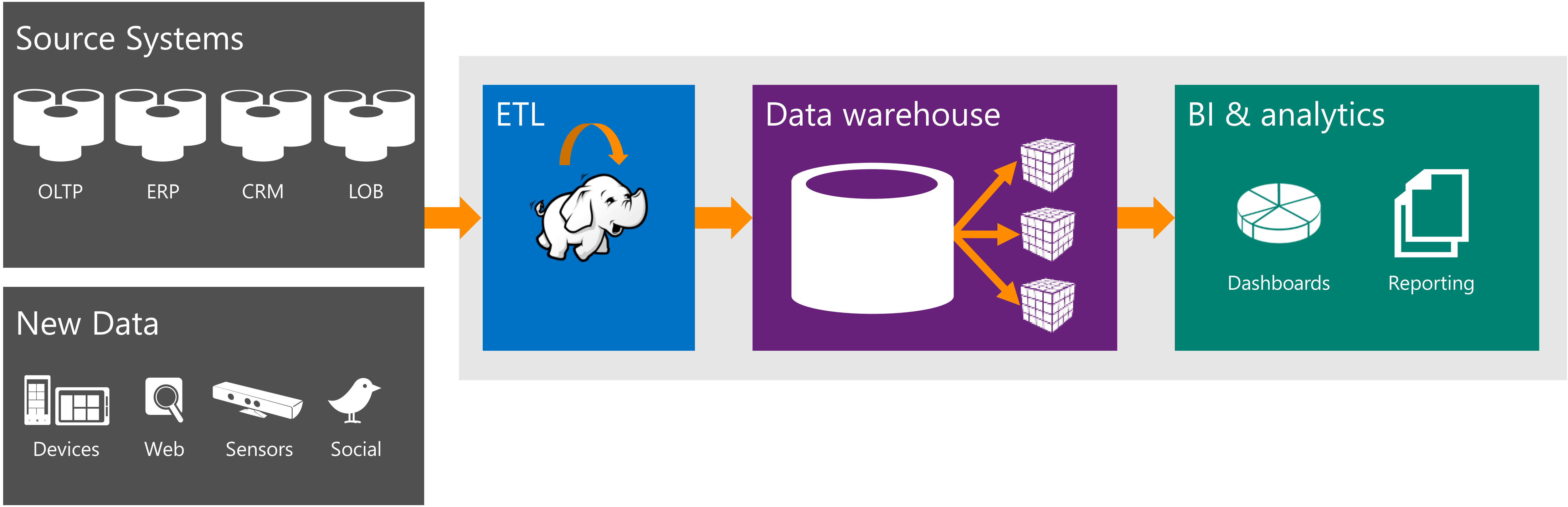
VALUE
kWh (Sum)

TIME
PeriodStart (None)

1.12.2012 0:00

Hadoop Scenario 1: pre-process ETL

Shift the pre-processing of ETL in staging data warehouse to Hadoop
Shifts high cost data warehousing to lower cost Hadoop clusters



Source Systems

OLTP
ERP
CRM
LOB

New Data

Devices
Web
Sensors
Social

ETL

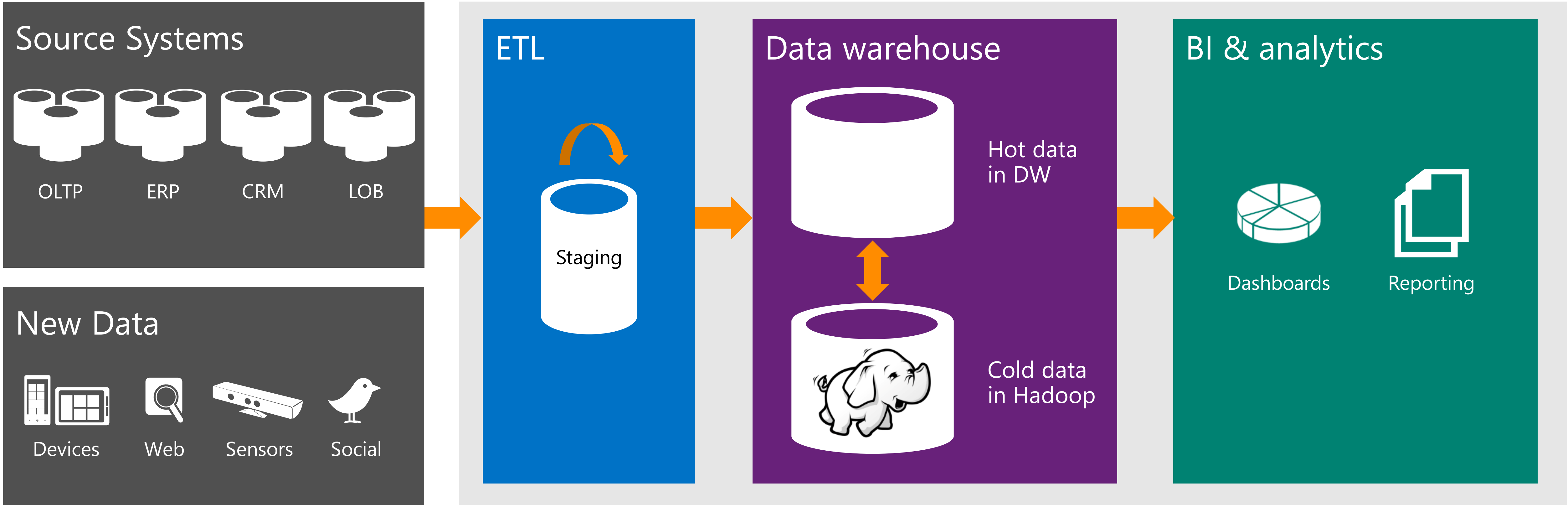
Data warehouse

BI & analytics

Dashboards
Reporting

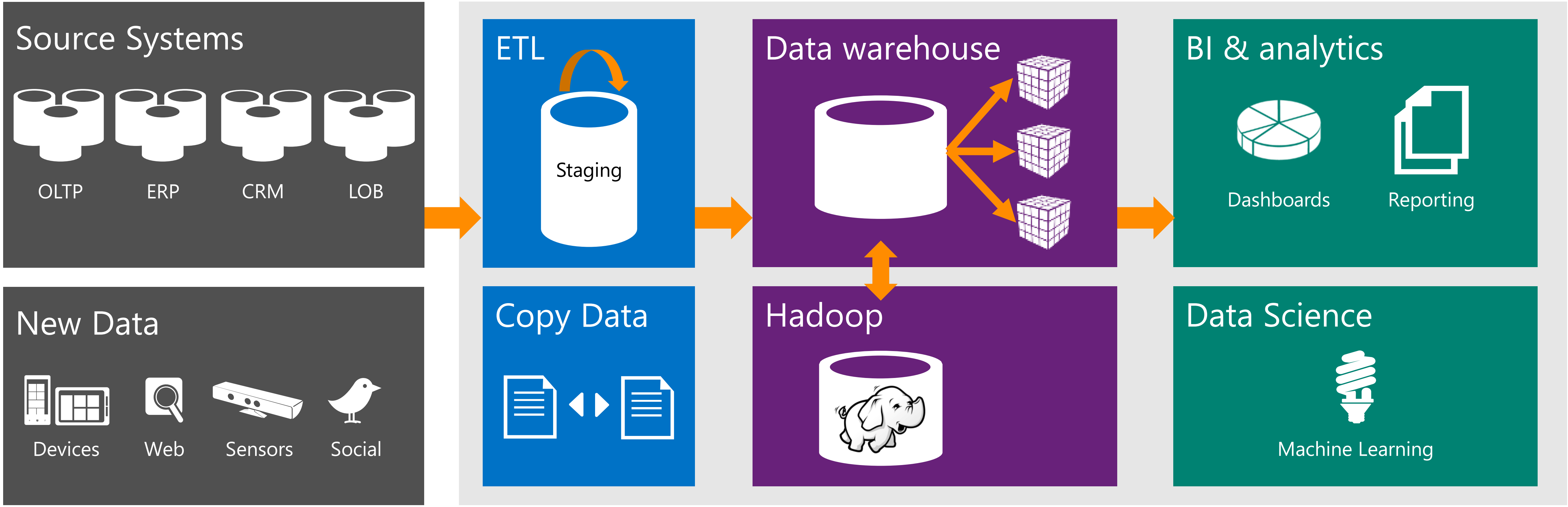
Hadoop Scenario 2: hot and cold storage

Offloading large volume of historical data into cold storage with Hadoop
Keep data warehouse for hot data to allow BI and analytics
When data from cold storage is needed, it can be moved back into the warehouse



Hadoop Scenario 3: true data discovery

Keep data warehouse for operational BI and analytics
Allow data scientists to gain new discoveries on raw data (no format or structure)
Operationalize discoveries back into the warehouse



Source Systems

OLTP
ERP
CRM
LOB

New Data

Devices
Web
Sensors
Social

ETL

Staging

Copy Data

Data warehouse

Hadoop

BI & analytics

Dashboards
Reporting

Data Science

Machine Learning

Industry Use Cases of Hadoop

Financial services

New account risk screens
Fraud prevention
Trading risk
Maximize deposit spread
Insurance underwriting
Accelerate loan processing

Retail

360° view of the customer
Analyze brand sentiment
Localized, personalized promotions
Website optimization
Optimal store layout

Telecom

Call detail records (CDRs)
Infrastructure investment
Next product to buy (NPTB)
Real-time bandwidth allocation
New product development

Manufacturing

Supplier consolidation
Supply chain and logistics
Assembly line quality assurance
Proactive maintenance
Crowd source quality assurance

Healthcare

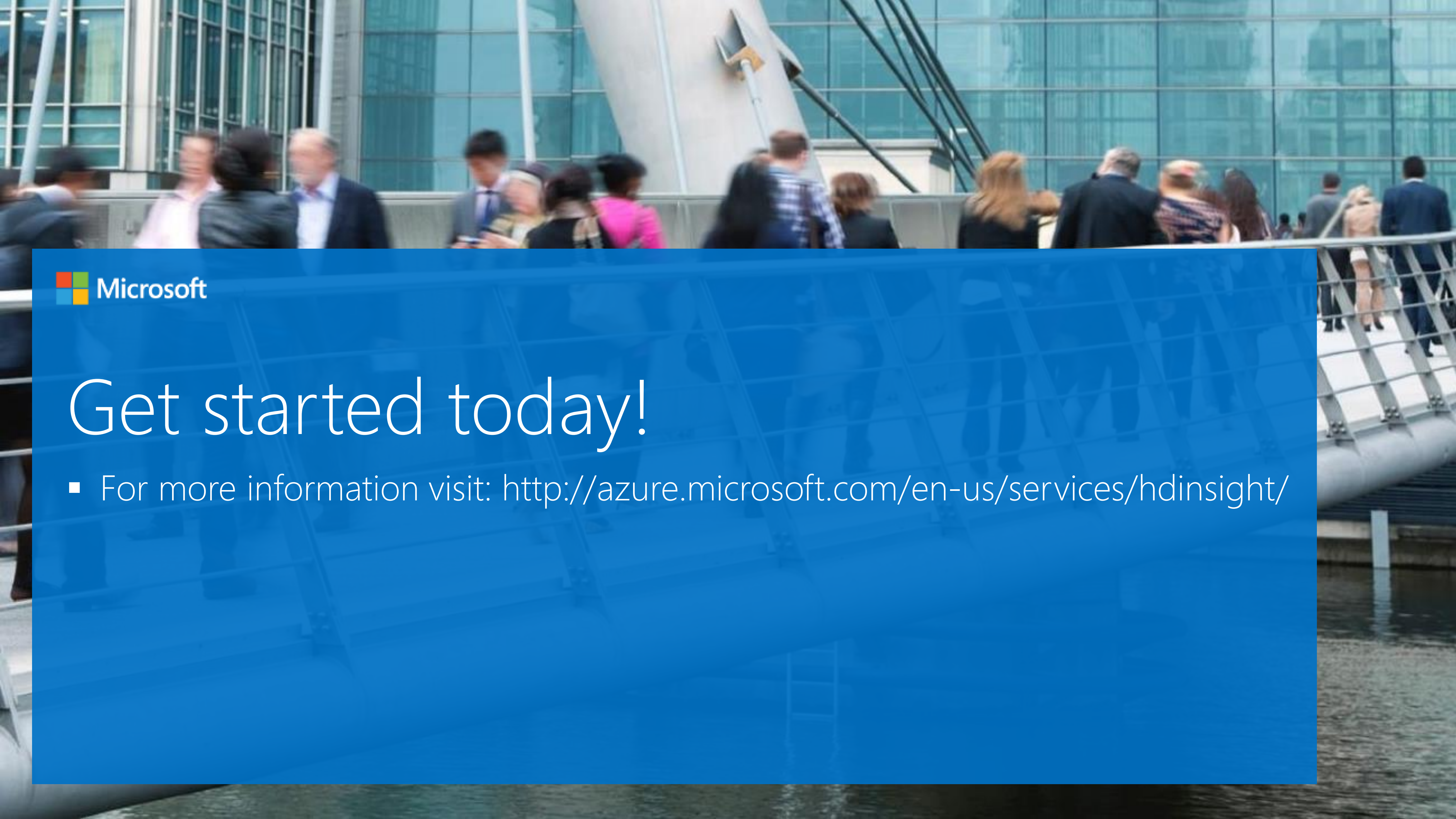
Genomic data for medical trials
Monitor patient vitals
Reduce re-admittance rates
Store medical research data
Recruit cohorts for pharmaceutical trials

Utilities, oil and gas

Smart meter stream analysis
Slow oil well decline curves
Optimize lease bidding
Compliance reporting
Proactive equipment repair
Seismic image processing

Public sector

Analyze public sentiment
Protect critical networks
Prevent fraud and waste
Crowd source reporting for repairs to infrastructure
Fulfill open records requests



Microsoft

Get started today!

- For more information visit: http://azure.microsoft.com/en-us/services/hdinsight/

Microsoft