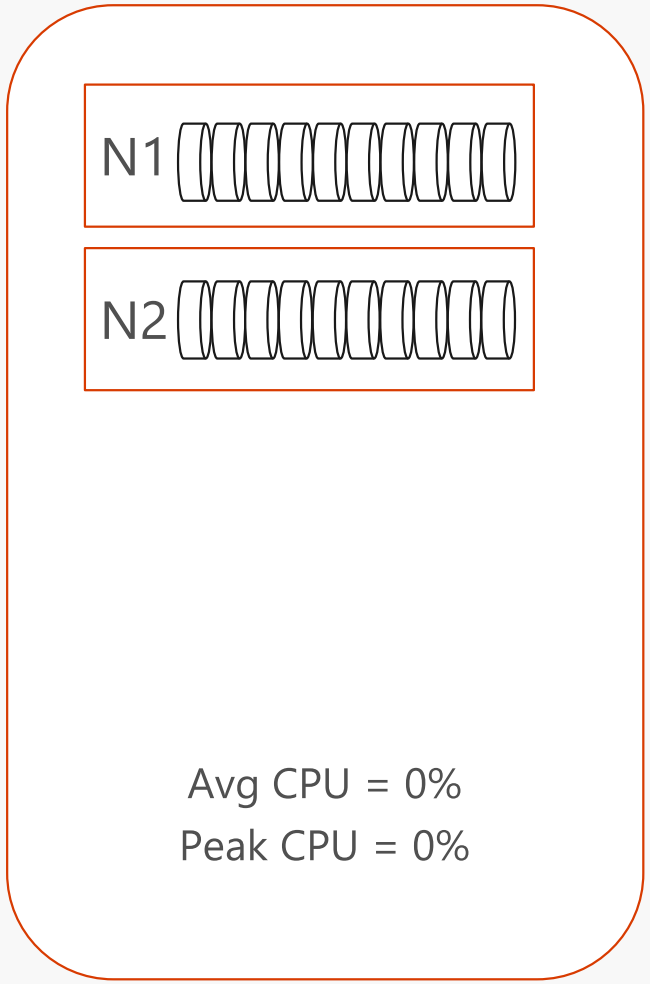
Microsoft

Big Data as a Service

Guido Jacobs
Technology Solution Professional – Big Data
Guido.Jacobs@microsoft.com

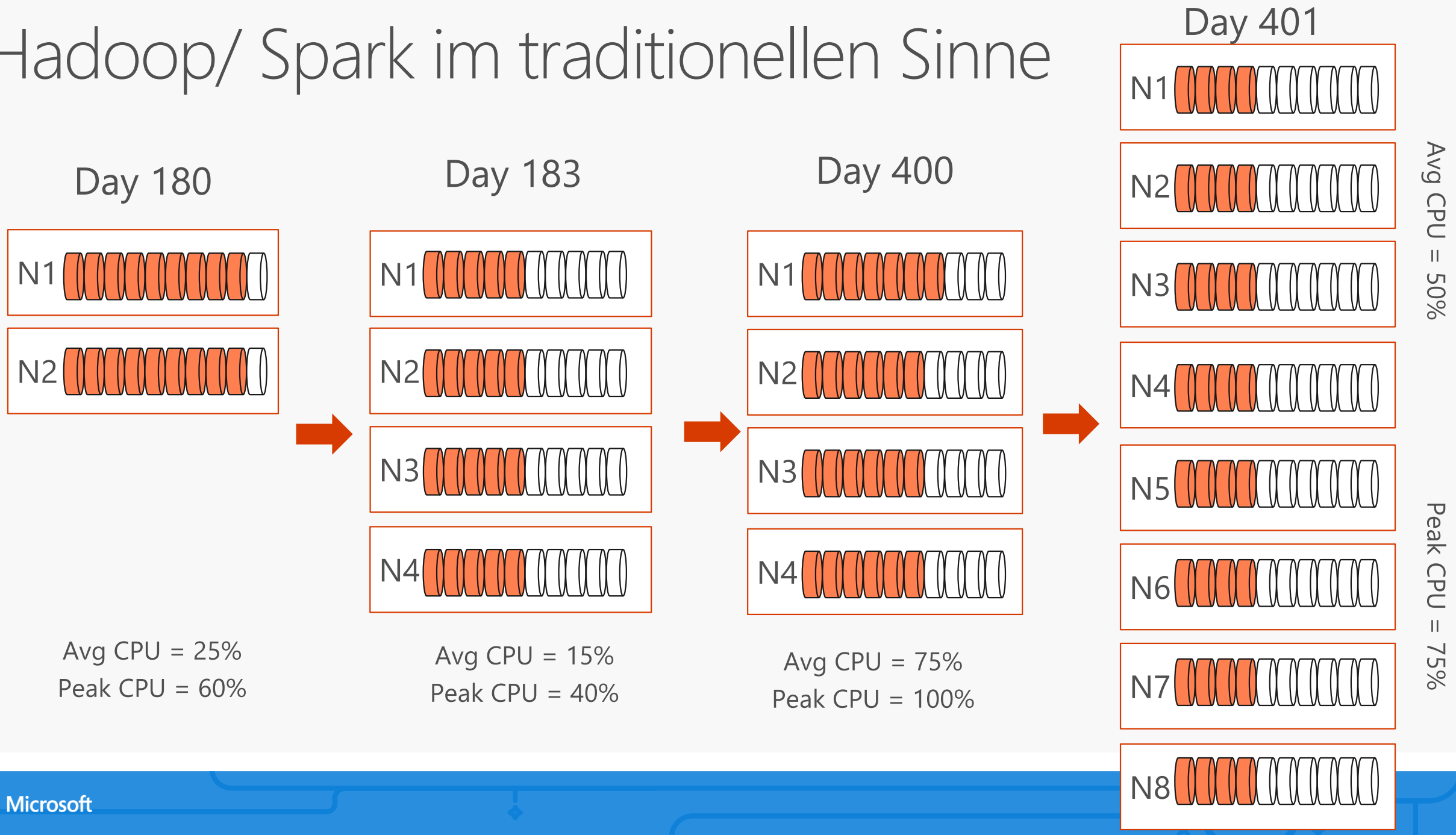
Microsoft Deutschland GmbH

Hadoop/ Spark im traditionellen Sinne

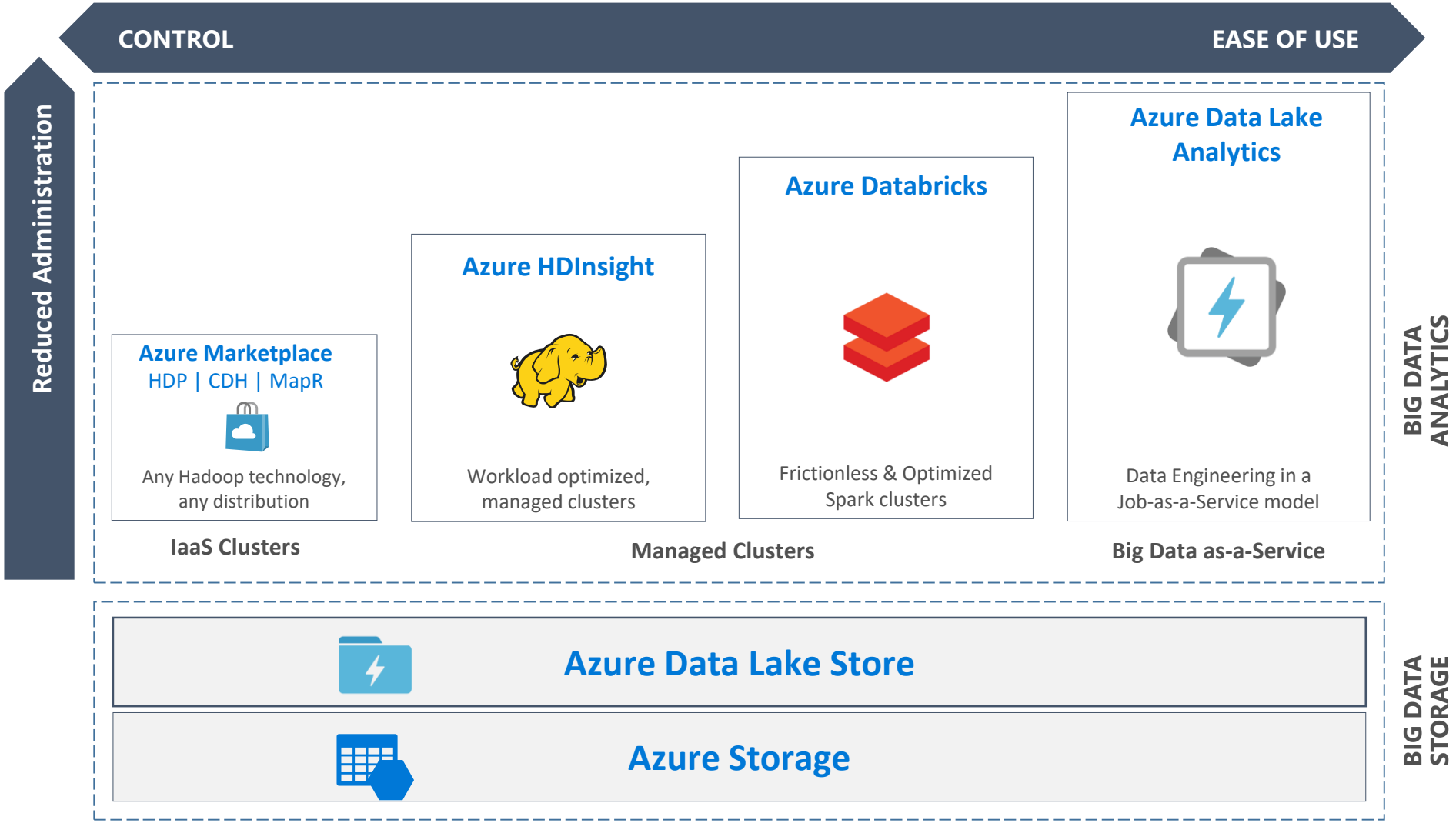


Ermittlung der Cluster Größe
Erstellung
Konfiguration
Administration
Beginn der Datenladung
Beginn der Entwicklung von Abfragen

Hadoop/ Spark im traditionellen Sinne



Flexibilität in der Cloud



Azure Data Lake Store

A hyper-scale repository for Big Data analytics workloads

Hadoop File System (HDFS) for the cloud

No limits to scale

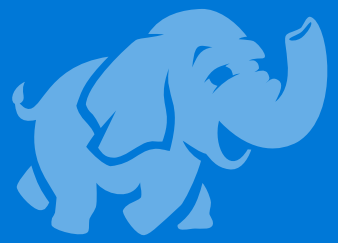
Store **any data** in its native format

Enterprise-grade access control, encryption at rest

Optimized for analytic workload **performance**

Azure HDInsight

Hadoop and Spark as a Service on Azure



Fully-managed Hadoop and Spark for the cloud

100% Open Source Hortonworks data platform

Clusters up and **running in minutes**

Managed, monitored and supported by Microsoft with the **industry's best SLA**

Familiar **BI tools for analysis**, or open source notebooks for **interactive data science**

63% lower TCO than deploy your own Hadoop on-premises*

*IDC study "The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight"

Azure Data Lake Analytics

A new distributed analytics service

Distributed analytics service built on Apache YARN

Elastic scale per query lets users focus on business goals—not configuring hardware

Includes U-SQL—a language that unifies the **benefits of SQL with the expressive power of C#**

Integrates with Visual Studio to develop, debug, and tune code faster

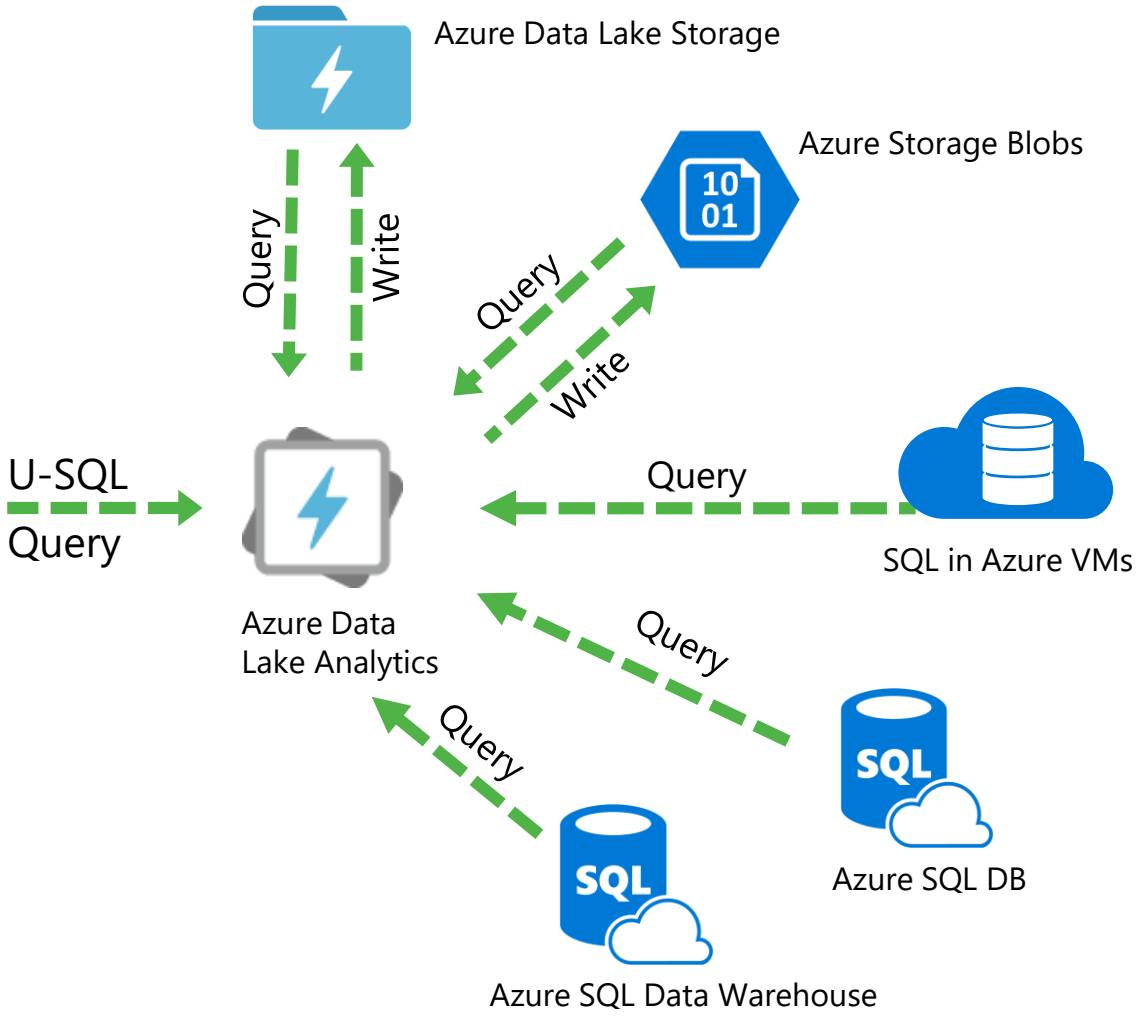
Federated query across Azure data sources

Enterprise-grade **role based access control**

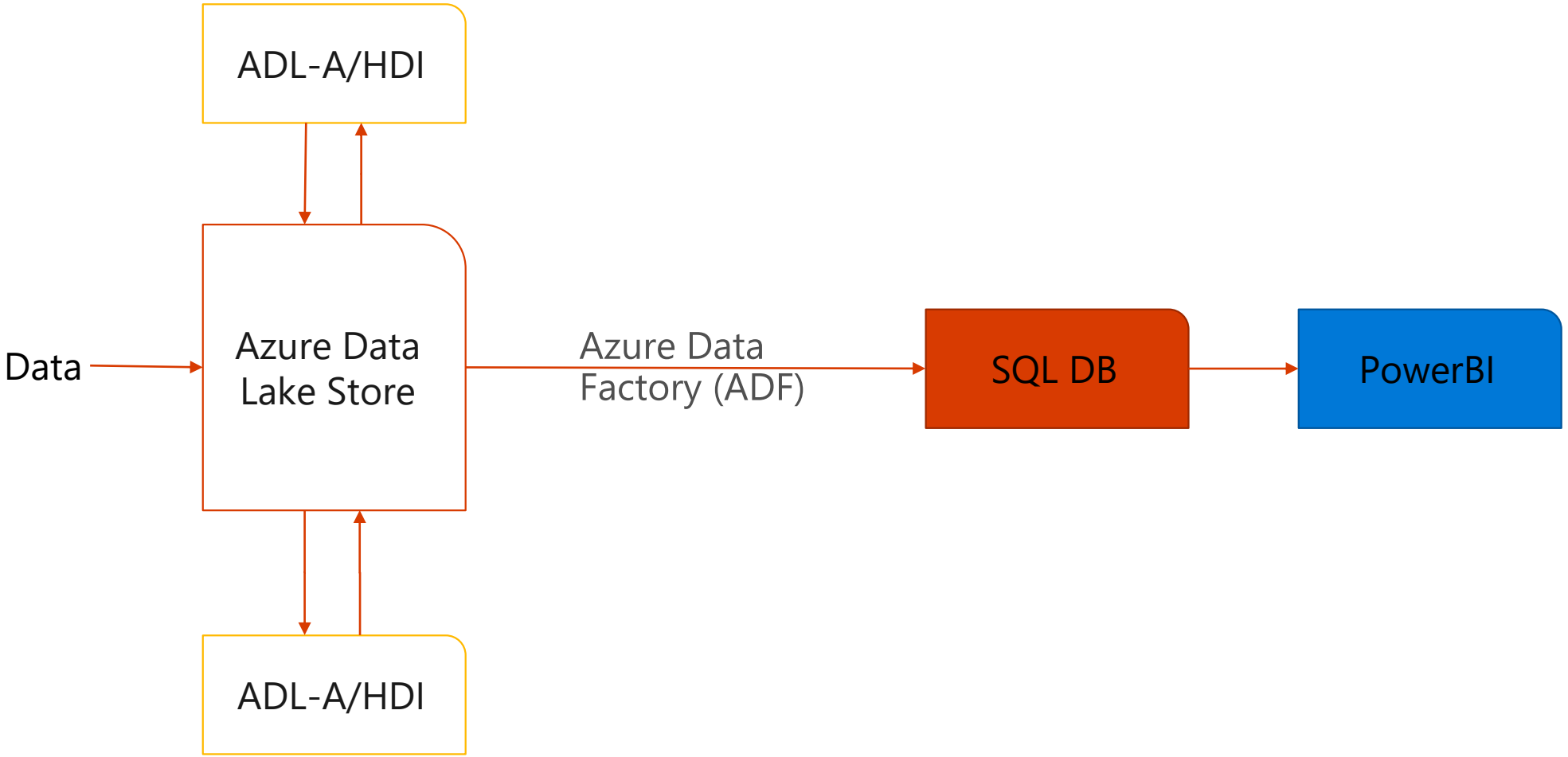
Daten einfach kombinieren

Vorteile

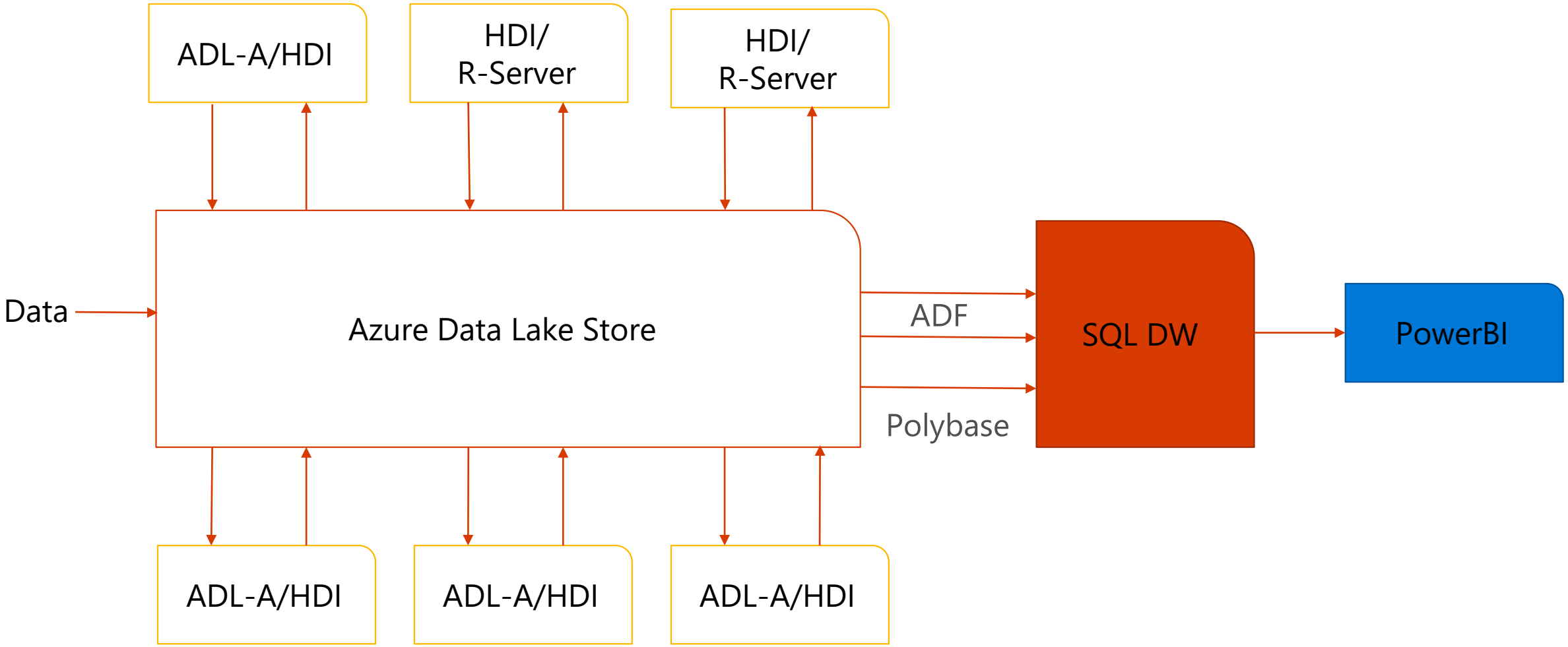
- Große Datenmengen müssen NICHT zwischen unterschiedlichen Speichern verschoben werden
- Einheitliche Sicht auf Daten unabhängig vom physikalischen Speicherplatz
- Verringerung der Datenpflege durch weniger Kopien
- Eine Abfragesprache für ALLE Datenquellen
- Jeder Datenspeicher behält seine Souveränität
- Bedarfsbezogenen Lösungsdesign
- SQL Prädikate werden an die SQL-Quellen gesendet
  - Filters
  - Joins



Trennung von Storage & Compute



Trennung von Storage & Compute (2)

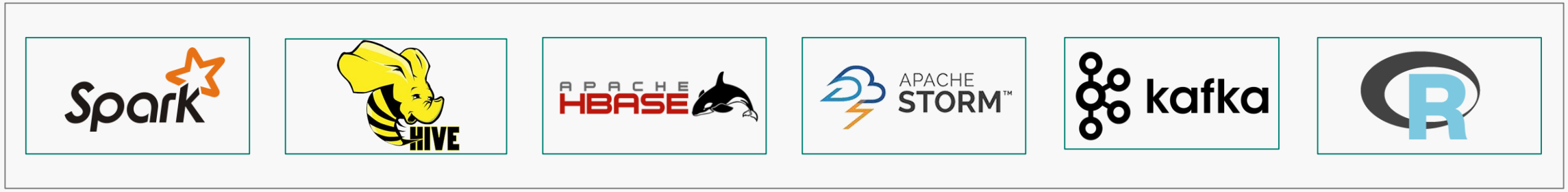


What is HDInsight?

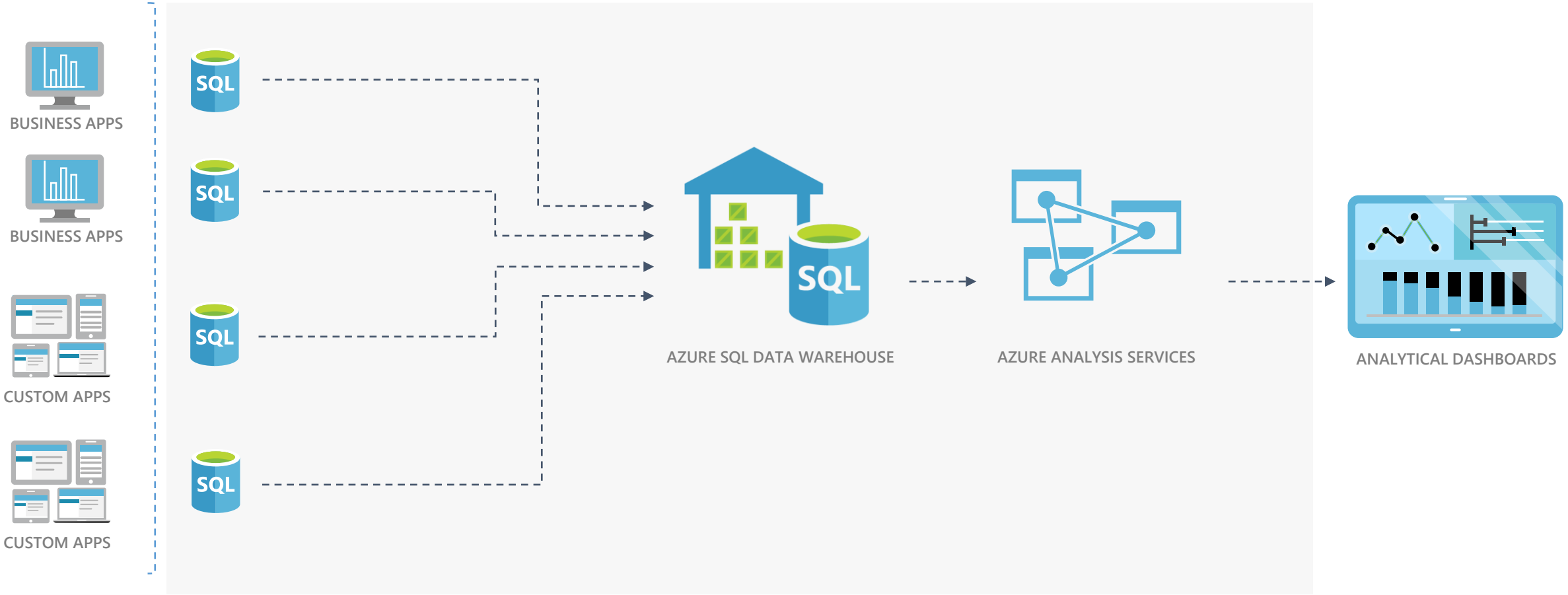
Reliable Open Source analytics with an Industry leading SLA

HDInsight allows you to easily spin up open source cluster types guaranteed with the industry's best **99.9% enterprise-grade SLA and 24/7 support** within minutes. We guarantee this SLA for the entire big data solution, not just the VM instances. HDInsight is architected for **full redundancy** and **high availability** including **head node replication**, **data geo-replication**, and it comes with **built-in enterprise level security and monitoring**

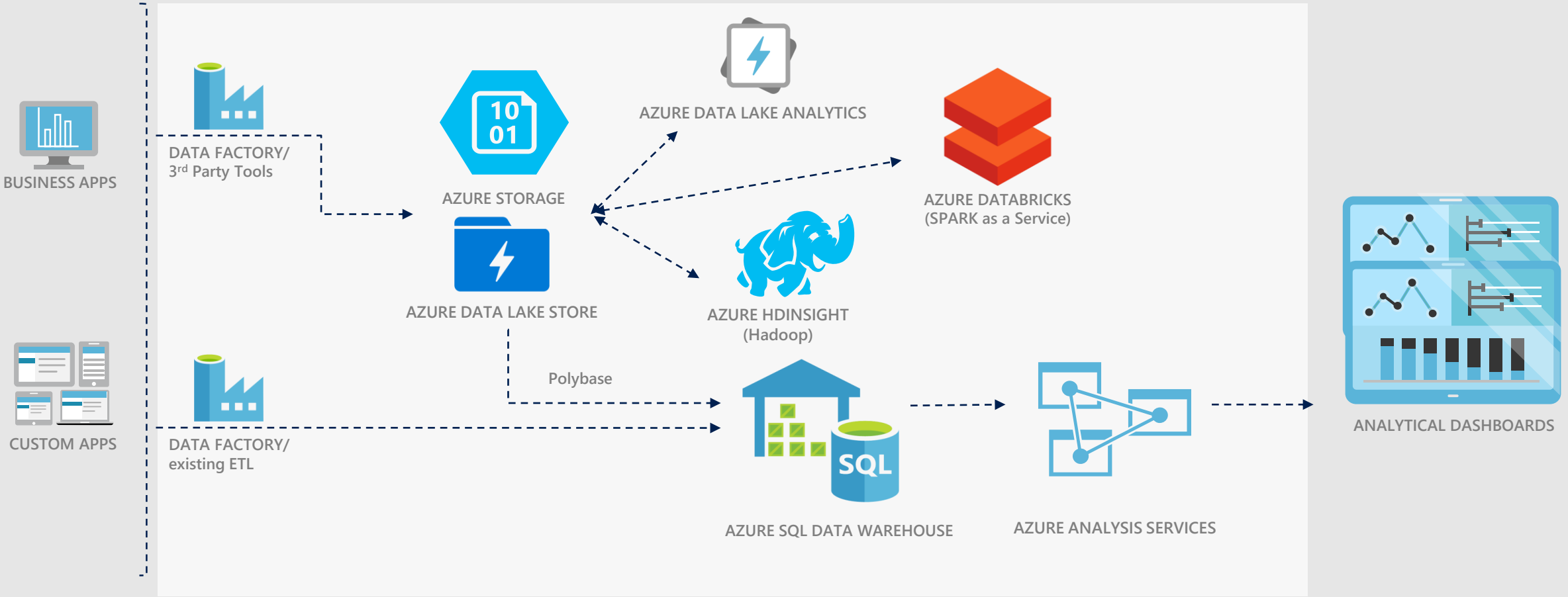
- Enterprise Ready + Secure
- Reliable + Fault Tolerant
- Engineered to scale
- Easy to use
- Productivity Applications
- Low Cost



Traditionelles DWH



Modern Data Warehouse (Data Lake driven)



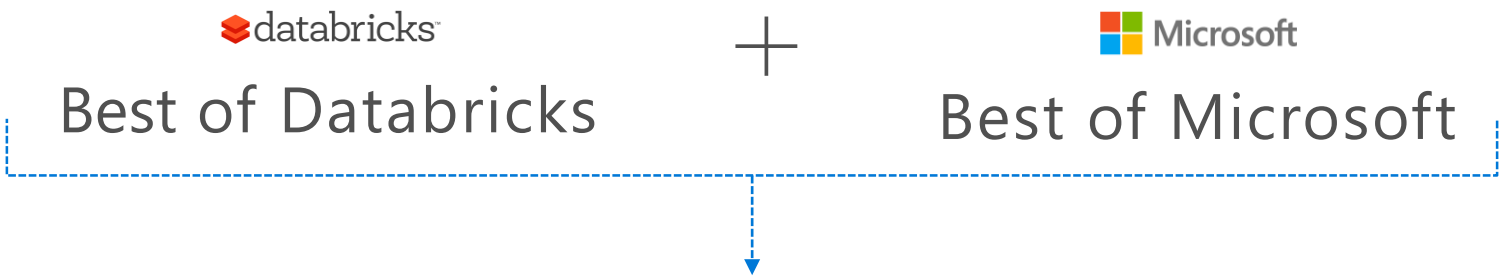
Azure Databricks

Why Spark?

- Open-source data processing engine built around **speed, ease of use, and sophisticated analytics**
- In memory engine that is up to **100 times faster than Hadoop**
- **Largest open-source data project** with 1000+ contributors
- **Highly extensible** with support for Scala, Java and Python alongside Spark SQL, GraphX, Streaming and Machine Learning Library (Mllib)

What is Azure Databricks?

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure



- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

Differentiated experience on Azure

ENHANCE PRODUCTIVITY

Get started quickly by launching your new Spark environment with one click.

Share your insights in powerful ways through rich integration with Power BI.

Improve collaboration amongst your analytics team through a unified workspace.

Innovate faster with native integration with rest of Azure platform

BUILD ON THE MOST COMPLIANT CLOUD

Simplify security and identity control with built-in integration with Active Directory.

Regulate access with fine-grained user permissions to Azure Databricks' notebooks, clusters, jobs and data.

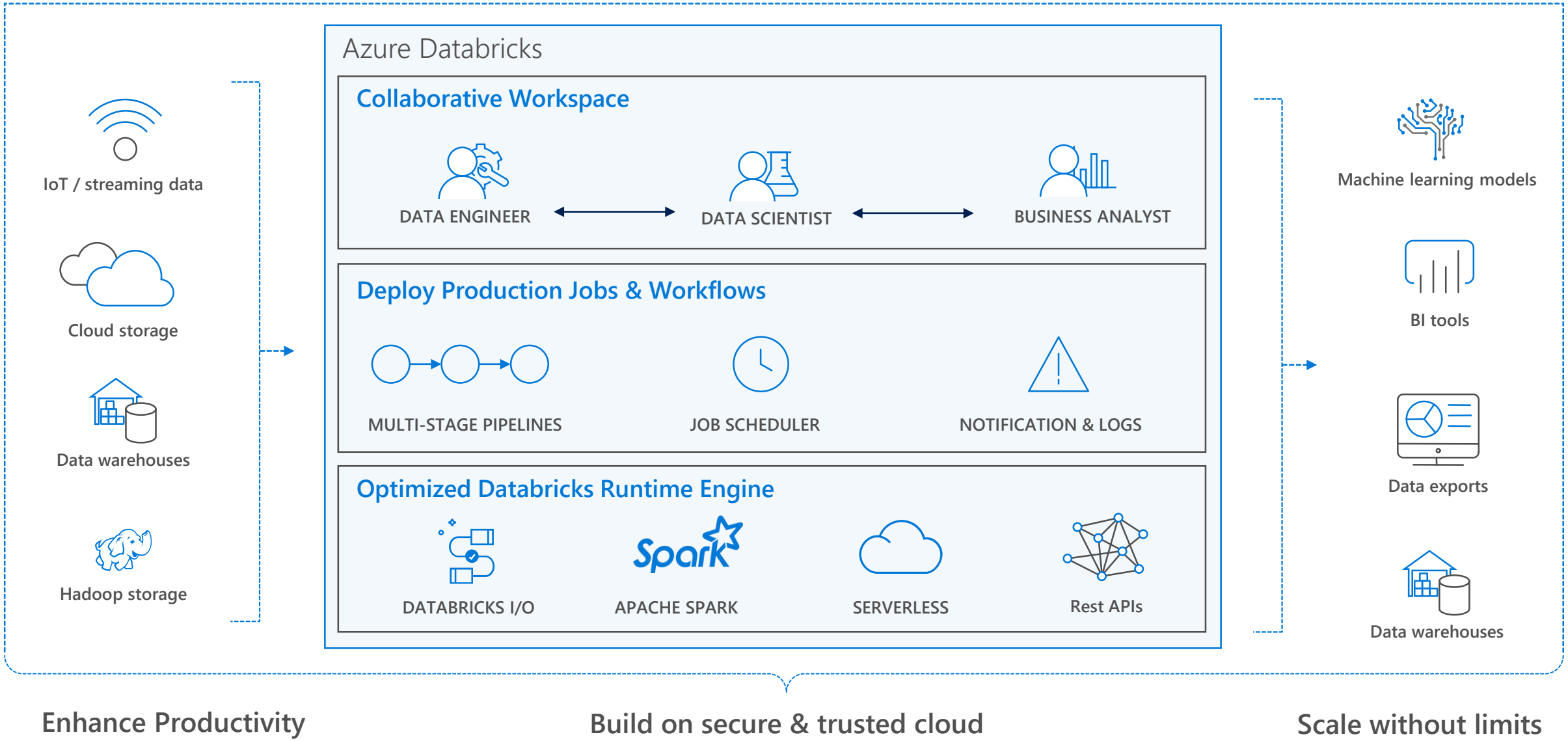
Build with confidence on the trusted cloud backed by unmatched support, compliance and SLAs.

SCALE WITHOUT LIMITS

Operate at massive scale without limits globally.

Accelerate data processing with the fastest Spark engine.

Azure Databricks

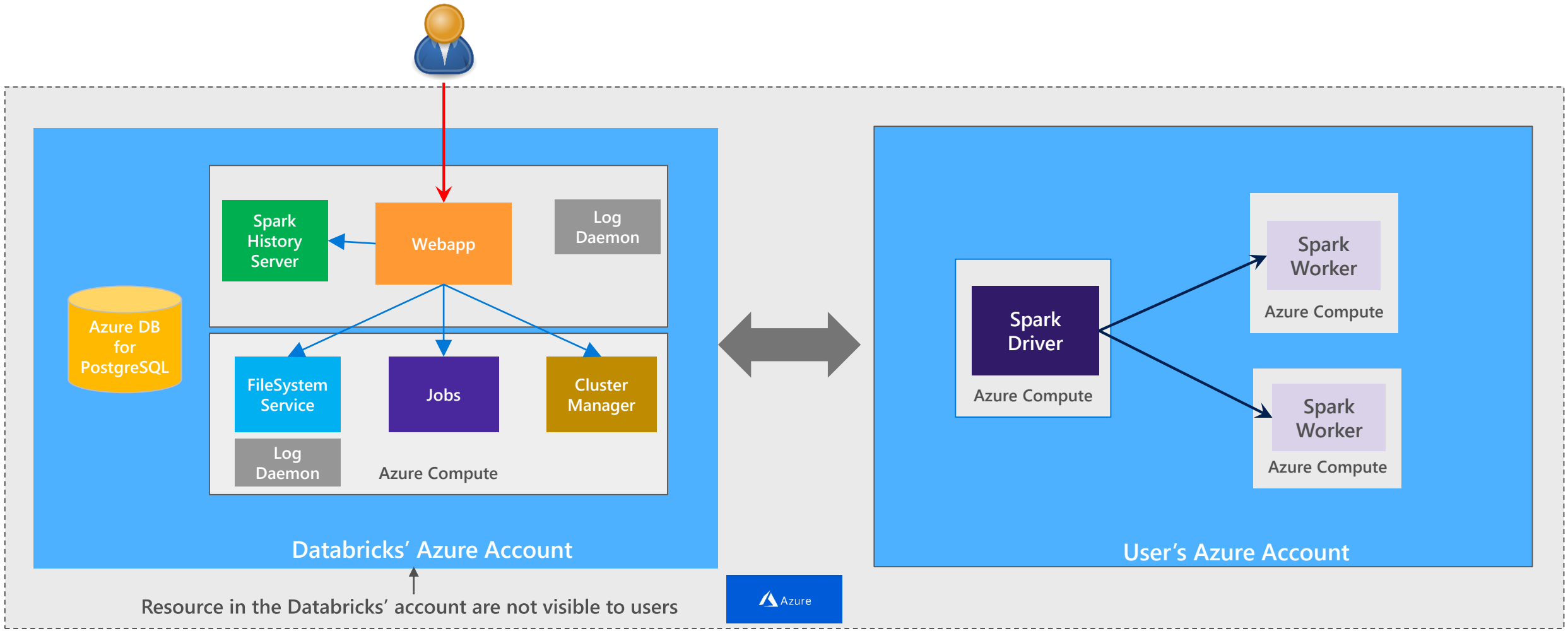


Enhance Productivity

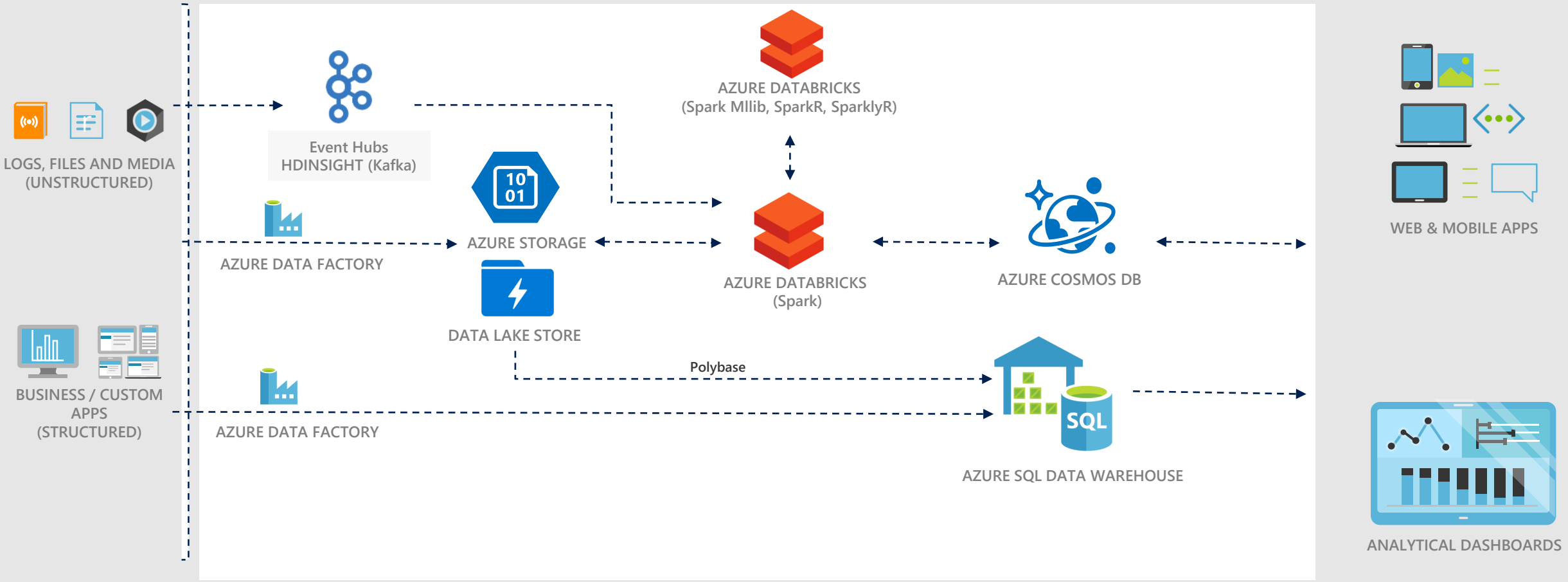
Build on secure & trusted cloud

Scale without limits

AZURE DATABRICKS CLUSTER ARCHITECTURE



Big Data Lambda Architecture



Azure Cosmos DB

Introducing Azure Cosmos DB

A globally distributed, massively scalable, multi-model database service

Global distribution

Automatically replicate all your data around the world –
across more regions than Amazon and Google combined

Introducing Azure Cosmos DB

A globally distributed, massively scalable, multi-model database service

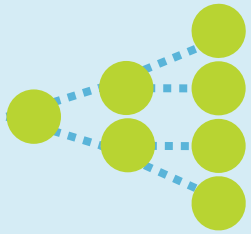
Global distribution

Multi-model + multi API

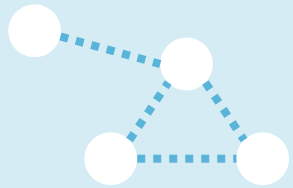
Use key-value, graph, and document with a schema-agnostic service that doesn't require any schema or secondary indexes

KEY-VALUE

COLUMN-FAMILY



DOCUMENT



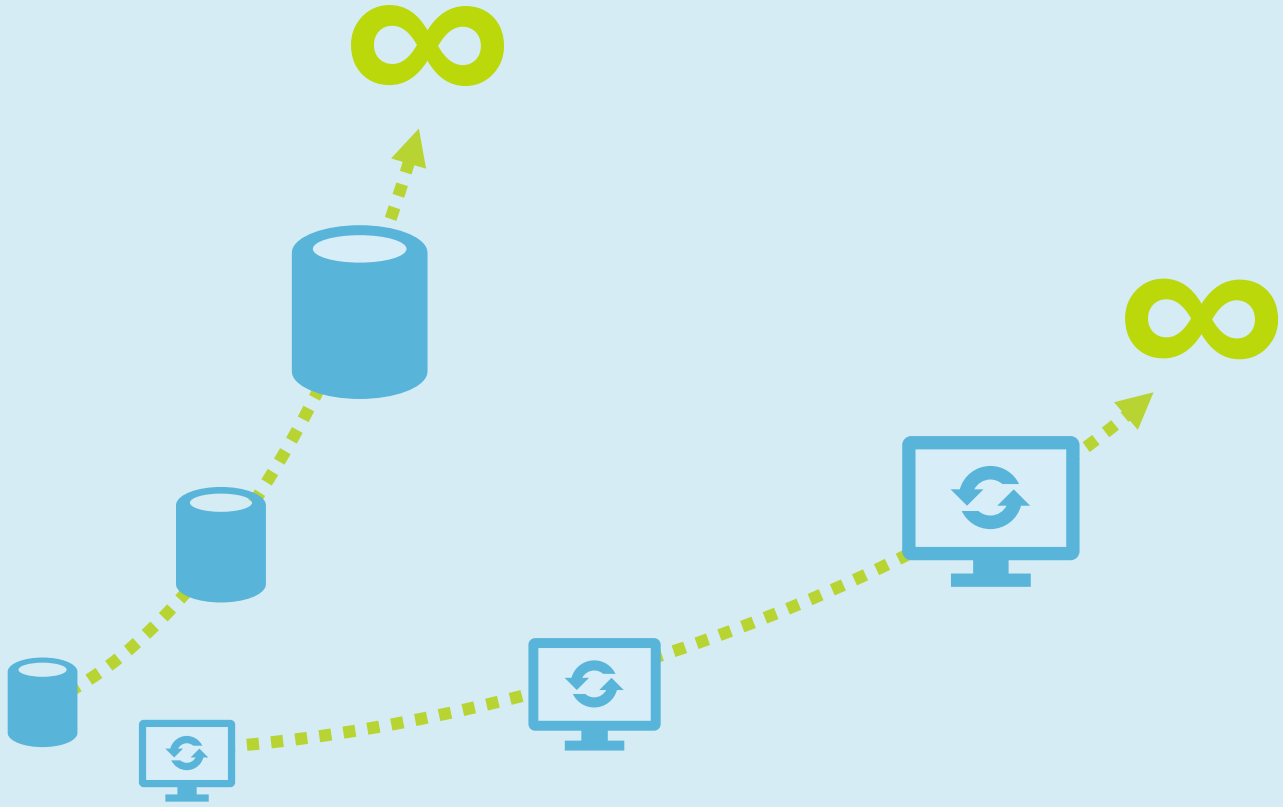
GRAPH

Introducing Azure Cosmos DB

A globally distributed, massively scalable, multi-model database service

- Global distribution
- Multi-model + multi API
- **Elastic scale-out**

Independently and elastically scale storage and throughput across regions

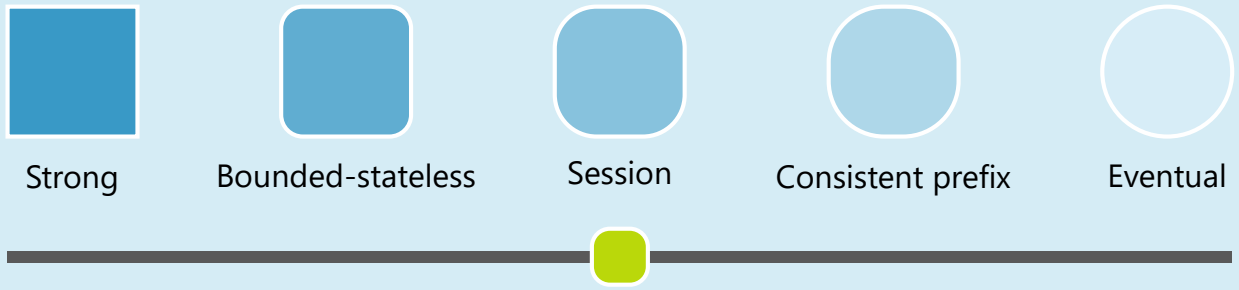


Introducing Azure Cosmos DB

A globally distributed, massively scalable, multi-model database service

- Global distribution
- Multi-model + multi API
- Elastic scale-out
- **Choice of consistency**

Choose from five defined consistency levels for low latency and high availability



Introducing Azure Cosmos DB

A globally distributed, massively scalable, multi-model database service

- Global distribution
- Multi-model + multi API
- Elastic scale-out
- Choice of consistency
- **Guaranteed single-digit latency**

Serve <10 ms read and <15 ms write requests at the 99th percentile from the nearest region while delivering data globally

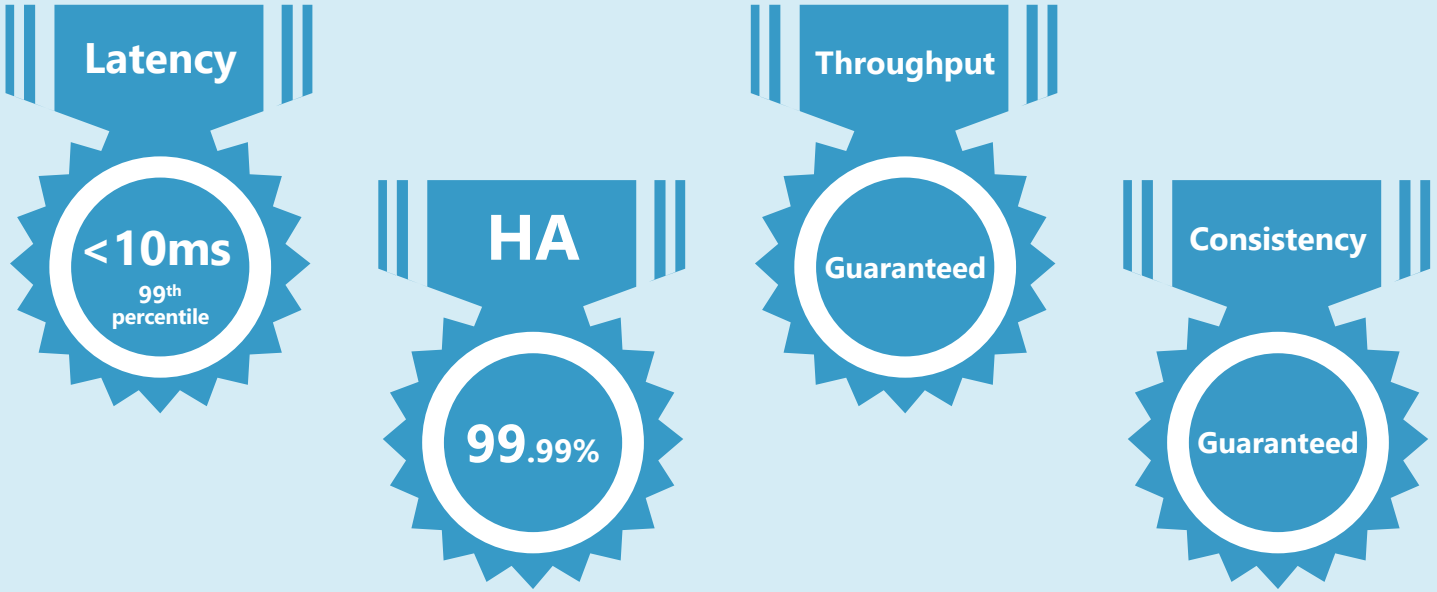
Guaranteed global millisecond latency at the 99th percentile

Introducing Azure Cosmos DB

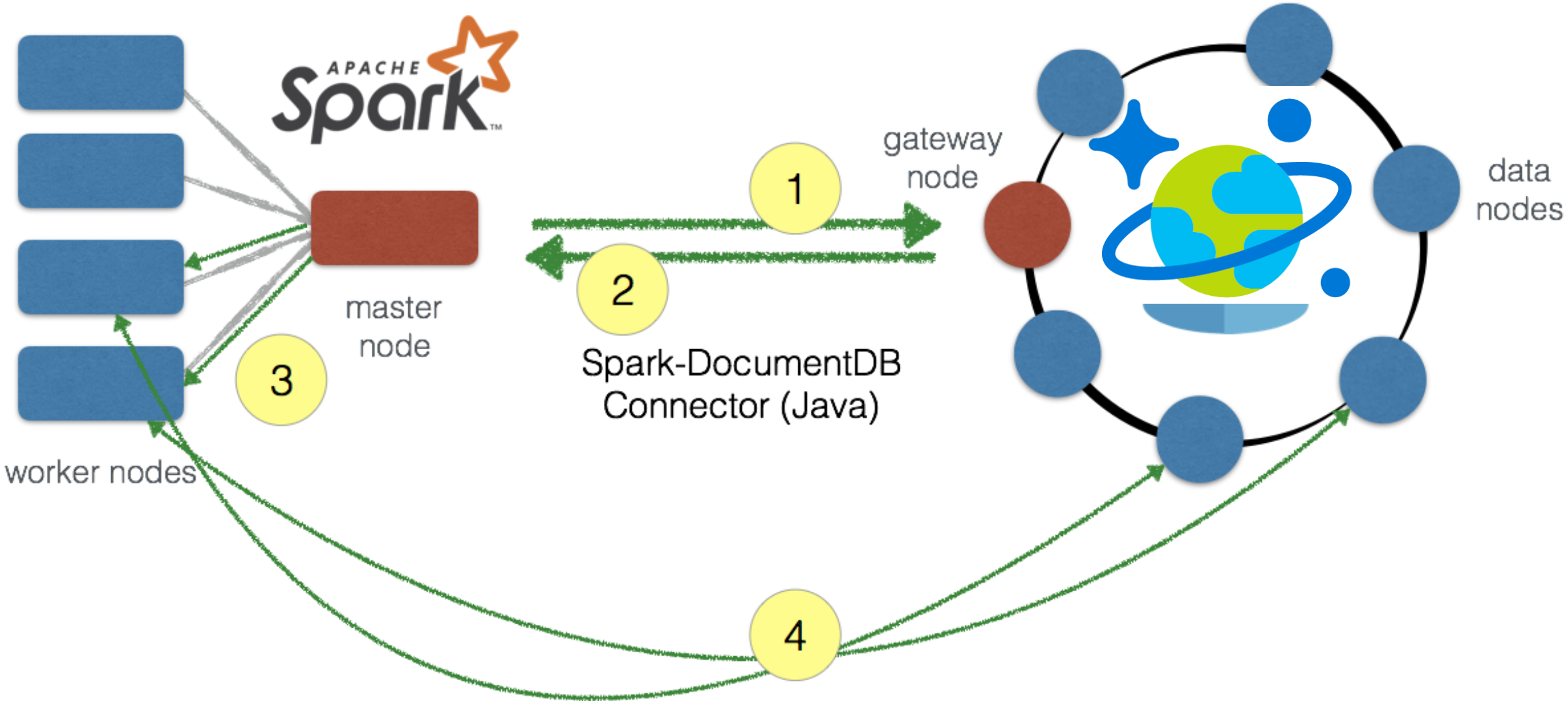
A globally distributed, massively scalable, multi-model database service

- Global distribution
- Multi-model + multi API
- Elastic scale-out
- Choice of consistency
- Guaranteed single-digit latency
- **Enterprise-level SLAs**

Only service with financially-backed SLAs for millisecond latency at the 99th percentile, 99.99% HA and guaranteed throughput and consistency

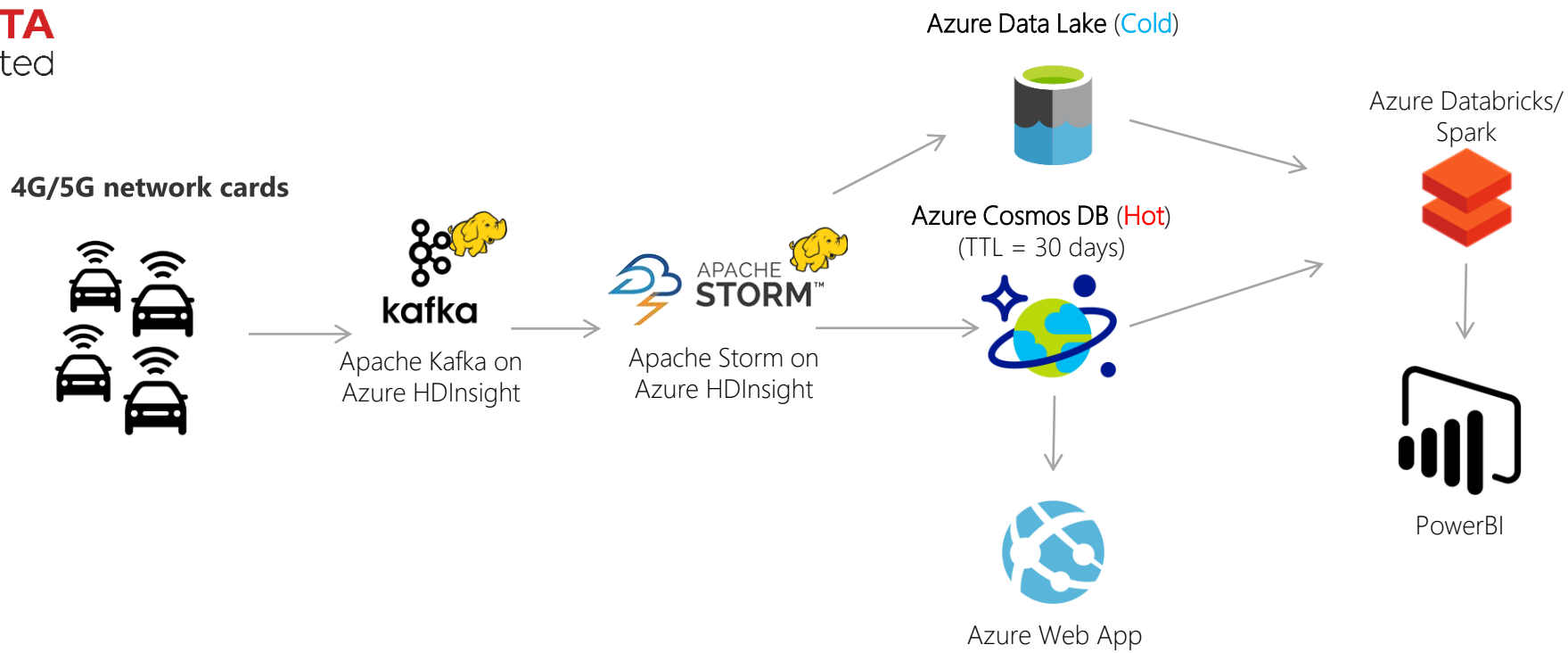


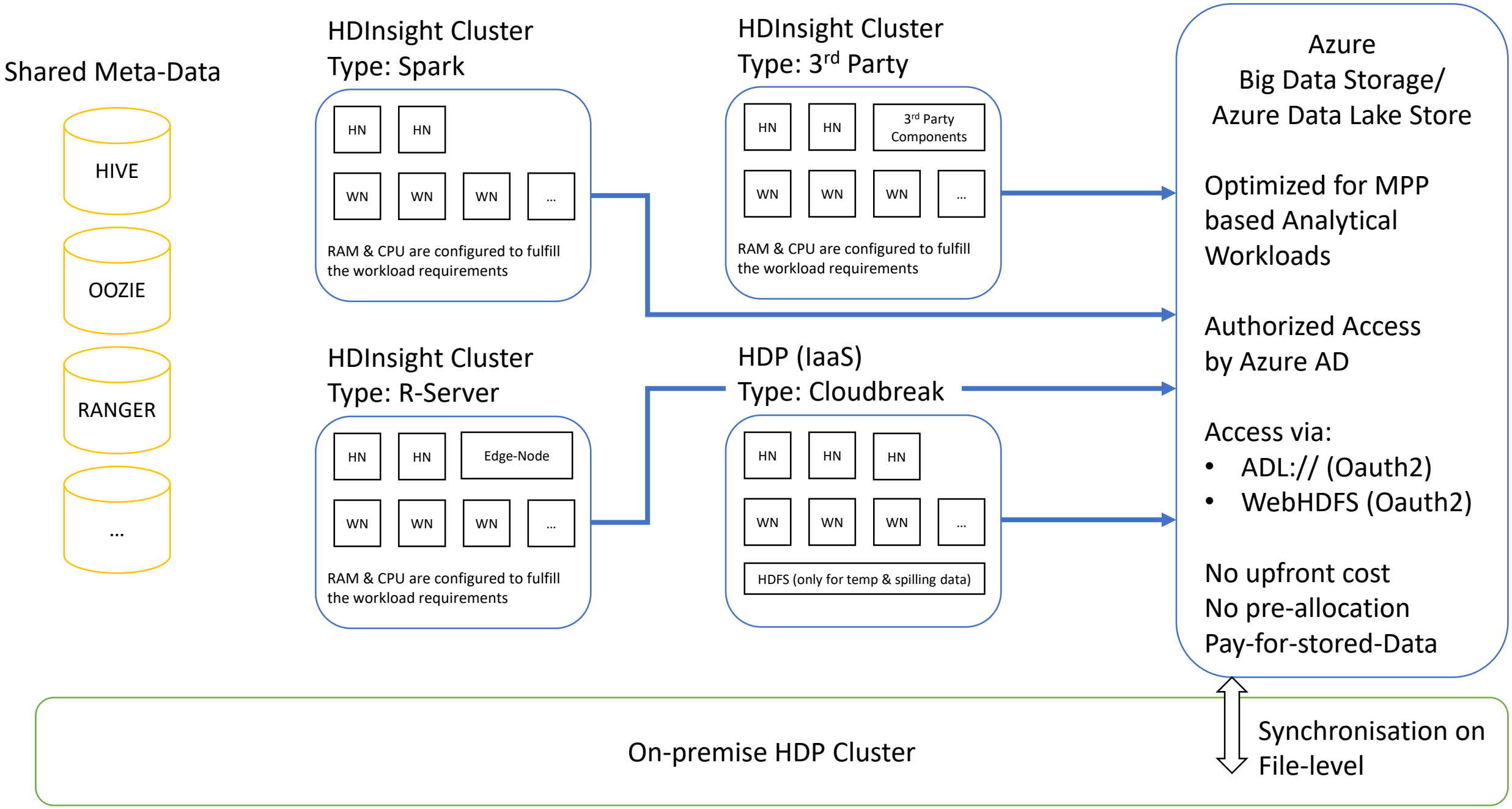
Spark - Cosmos DB Connector



Connected Car example

TOYOTA
connected




Shared Meta-Data
HIVE
OOZIE
RANGER
...
HDInsight Cluster
Type: Spark
HN
HN
WN
WN
WN
...
RAM & CPU are configured to fulfill the workload requirements
HDInsight Cluster
Type: 3rd Party
HN
HN
3rd Party Components
WN
WN
WN
...
RAM & CPU are configured to fulfill the workload requirements
HDInsight Cluster
Type: R-Server
HN
HN
Edge-Node
WN
WN
WN
...
RAM & CPU are configured to fulfill the workload requirements
HDP (IaaS)
Type: Cloudbreak
HN
HN
HN
WN
WN
WN
...
HDFS (only for temp & spilling data)
Azure
Big Data Storage/
Azure Data Lake Store
Optimized for MPP based Analytical Workloads
Authorized Access by Azure AD
Access via:
• ADL:// (Oauth2)
• WebHDFS (Oauth2)
No upfront cost
No pre-allocation
Pay-for-stored-Data
On-premise HDP Cluster
Synchronisation on File-level

© 2016 Microsoft Corporation. All rights reserved.