Microsoft

Microsoft Parallel Data Warehouse
微软并行数据仓库

The turnkey modern data warehouse appliance
现代化数据仓库一体机

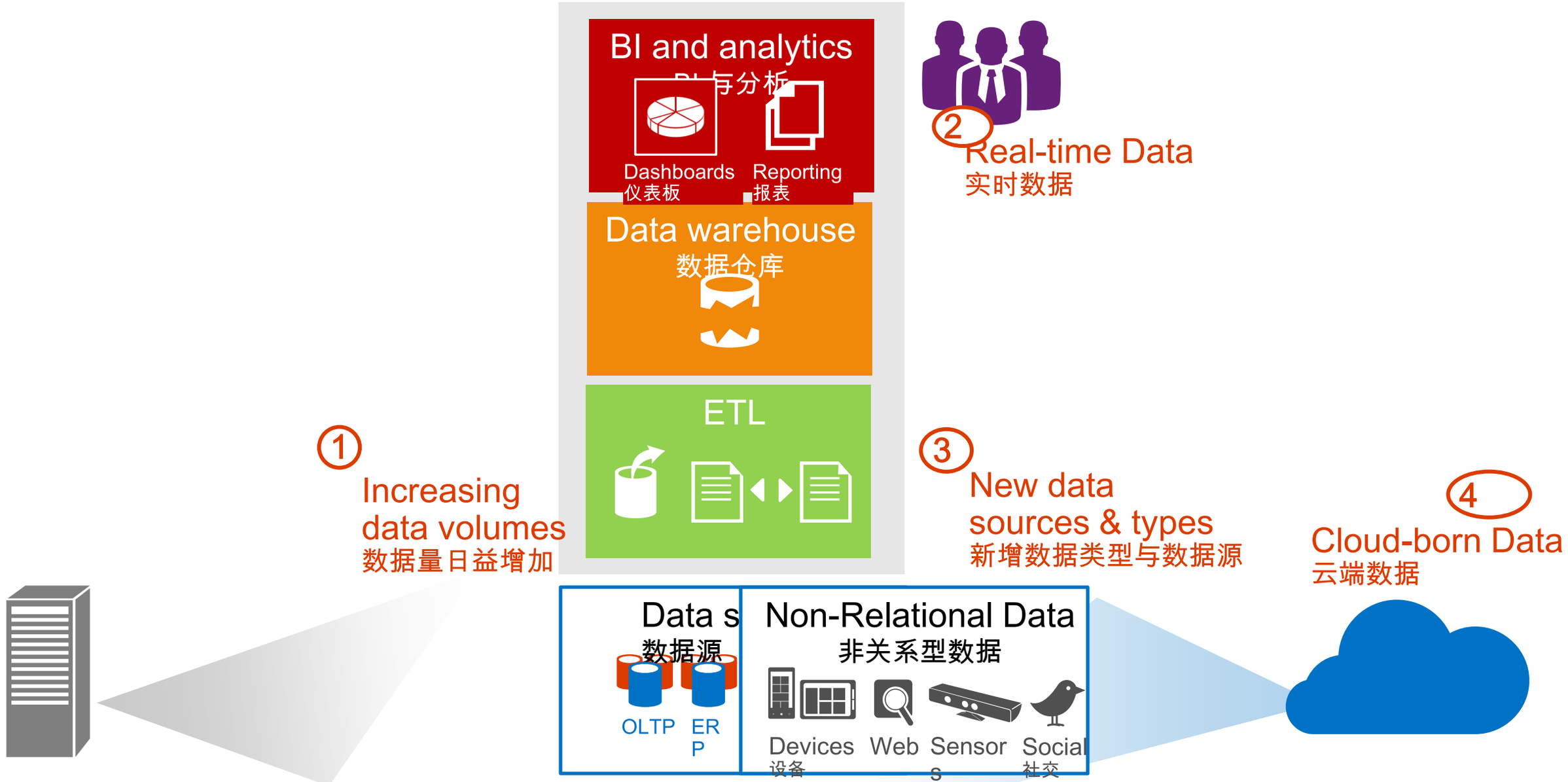
Mark Jewett
Worldwide Director, Server Appliances

2014 年 4 月 10 日

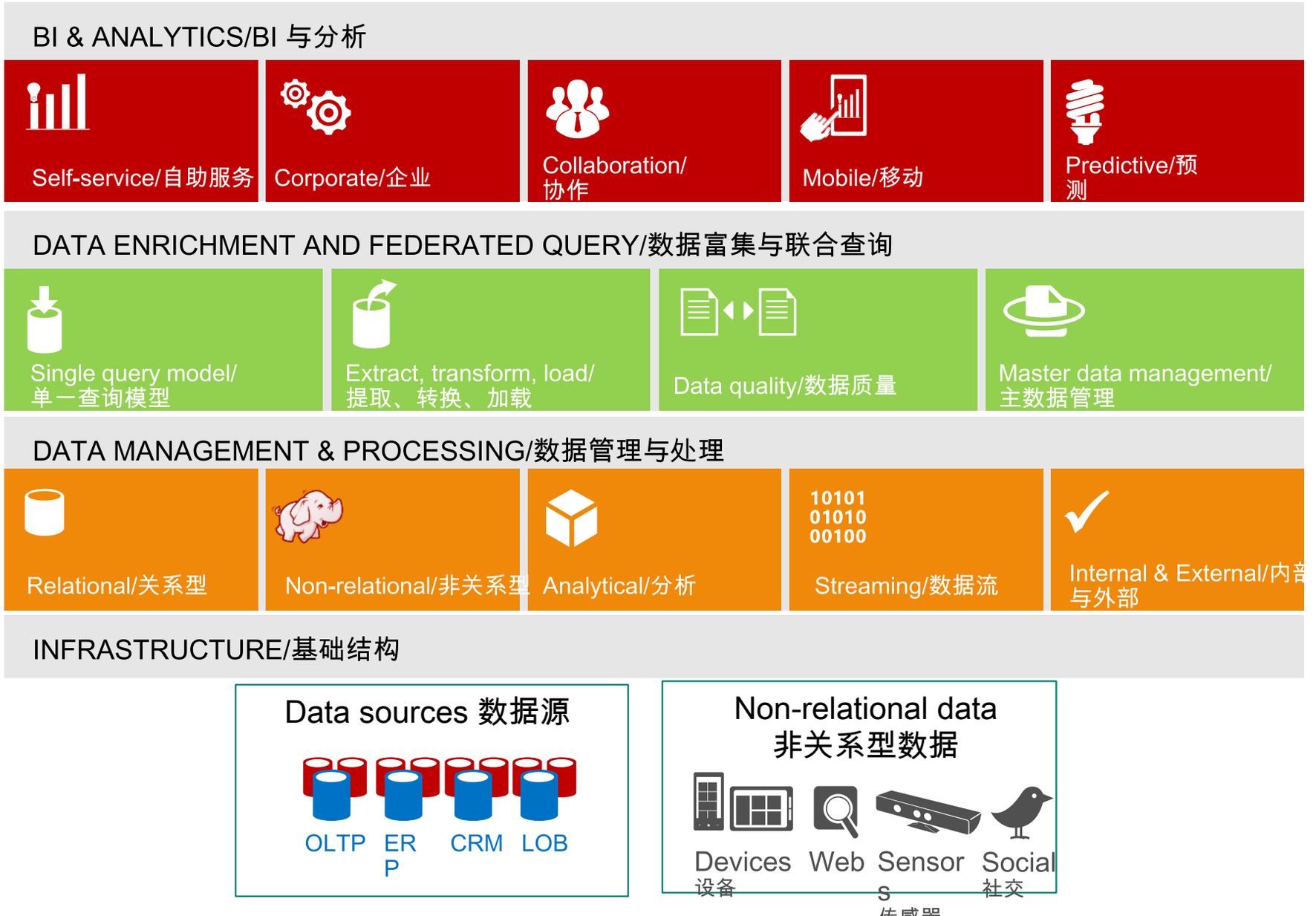
The traditional data warehouse传统数据仓库

> … data warehousing has reached the most significant tipping point since its inception. The biggest, possibly most elaborate data management system in IT is changing.
> …数据仓库自诞生而来终于到达了临界点。IT 领域规模最大，最精致的数据管理系统正在发生变化。
>
> – Gartner, “The State of Data Warehousing in 2012”

The traditional data warehouse 传统数据仓库



The modern data warehouse 现代化数据仓库



Gain knowledge of all your existing data 通过任何数据获得知识

Enrich and optimize your data from non-traditional sources 通过非关系型数据源富集并优化数据

A city wanted better insights into service effectiveness. They improved services by using social, service logs, devices and GPS to improve safety and enhance services and community.
某座城市希望改善针对服务效力所获得的洞察力。他们通过使用社交、服务日志、设备及 GPS 数据改善安全性，构建更好的服务与社区。

Social and web analytics
社交与 Web 分析

A building management company wanted to integrate and analyze data from sensors and equipment to improve efficiency and lower energy costs by 20%.
一家建筑物管理公司希望集成来自传感器与仪器的数据，并对其进行分析，借此改善效率，并将能耗成本降低了 20%

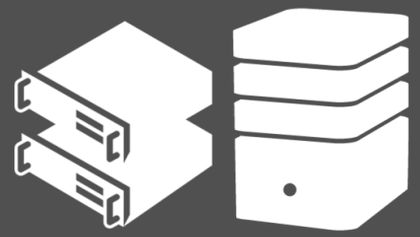
Live data feeds
实时数据源

A technical university needed on-demand computing in the cloud for DNA sequencing to accelerate access, discovery, and analysis.
一家技术学院需要通过云环境的按需计算能力处理 DNA 序列数据，提高访问、发现和分析工作的速度。

Advanced analytics
高级分析

Roadblocks to evolving to a modern data warehouse
现代化数据仓库面临的发展障碍

Keep legacy investment
保持原有投资



Limited scalability & ability to handle new data types
扩展性有限，无法处理新型数据

Acquire big data solution
获取大数据解决方案

Significant training & still siloed
需要大量培训，不同系统相互隔离

Buy new tier one hardware appliance
购买第一层硬件设备

High acquisition/ migration Costs
购买/迁移成本高

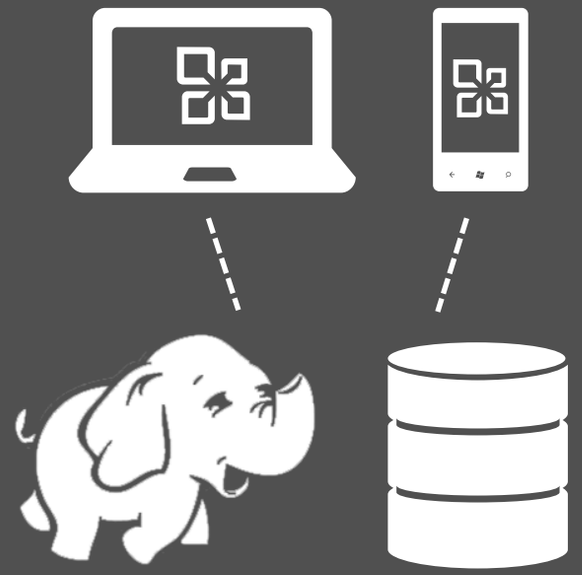
Acquire business intelligence
获取商业智能

Complex with low adoption
过于复杂，接受度低

Microsoft Parallel Data Warehouse 微软并行数据仓库

The turnkey modern data warehouse appliance 现代化数据仓库设备一体机

Enterprise-ready big data
面向企业的大数据



Next-generation performance at scale
下一代性能与规模

Engineered for optimal value
以提供更优化价值为设计目标

Hadoop alone is not the answer to all big data challenges
Hadoop 自身并非所有大数据问题的终极答案

Steep learning curve, slow and inefficient

Hadoop ecosystem
陡峭的学习曲线，缓慢低效的 Hadoop 生态系统

Learn new skills/
学习新技能

MapReduce

“New” data sources
“新”数据源

Devices 设备 | Web | Sensor 传感器 | Social 社交

HDFS (Hadoop)

Build
Integrate
Manage
Maintain
Support
构建集成式管理的维护支持

Move HDFS into the warehouse before analysis
分析前需将 HDFS 迁移至仓库

“New” data sources
“新”数据源

Devices 设备 | Web | Sensor 传感器 | Social 社交

TSQL

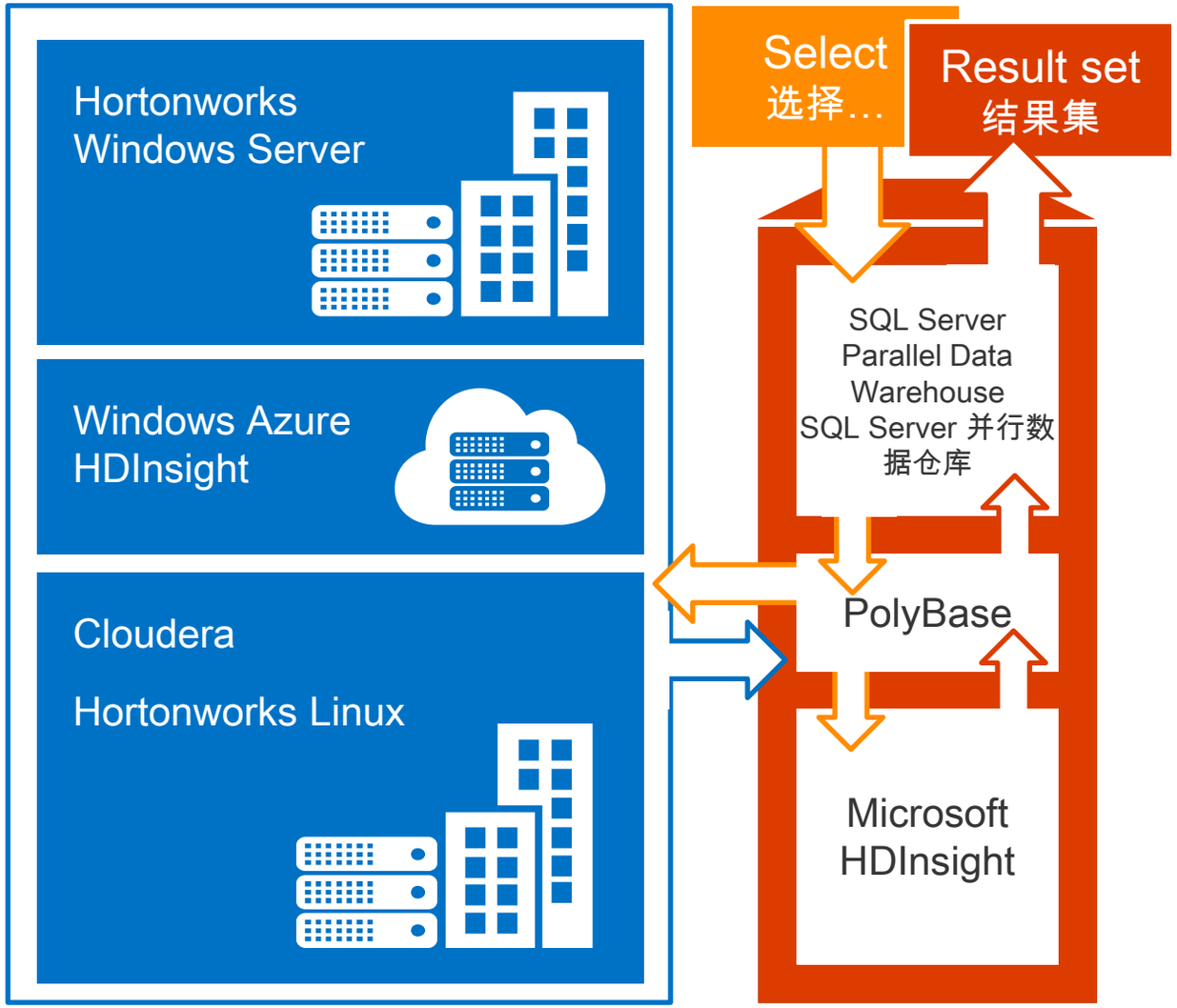
HDFS (Hadoop)

ETL

Warehouse
仓库

Connecting islands of data with PolyBase 使用 PolyBase 连接信息孤岛

Bringing Hadoop point solutions and the data warehouse together for users and IT
将 Hadoop 单点解决方案与数据仓库打包供用户与 IT 使用



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

PDW 与 Hadoop 可使用通用的 T-SQL 查询模型，并具备丰富的 T-SQL 功能，包括无需 ETL 的连接

Leverages the power of MPP to enhance query execution performance
利用大规模并行处理的强大运算能力改善查询性能

Supports Windows Azure HDInsight to enable new hybrid cloud scenarios
支持 Windows Azure HDInsight，实现全新混合云场景

Query non-Microsoft Hadoop distributions such as Hortonworks and Cloudera

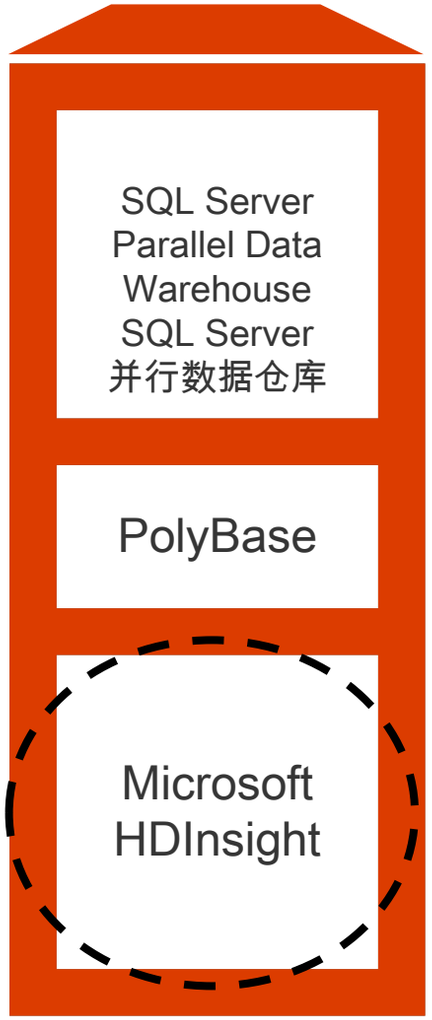
查询非微软 Hadoop 发行版，例如 Hortonworks 与 Cloudera

Coming Soon:

PDW delivers enterprise-ready Hadoop with HDInsight

即将发布：PDW 面向企业提供带有 HDInsight 的 Hadoop

Manageable, secured and highly available Hadoop integrated into the appliance可管理，安全，高可用的 Hadoop 直接集成于一体机

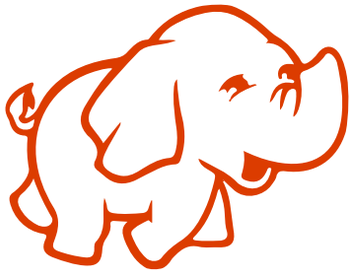


High performance tuned within the appliance

装置进行高性能优化

End-user authentication with Active Directory

最终用户使用 Active Directory实现身份验证



100% Apache Hadoop

Managed and monitored using System Center

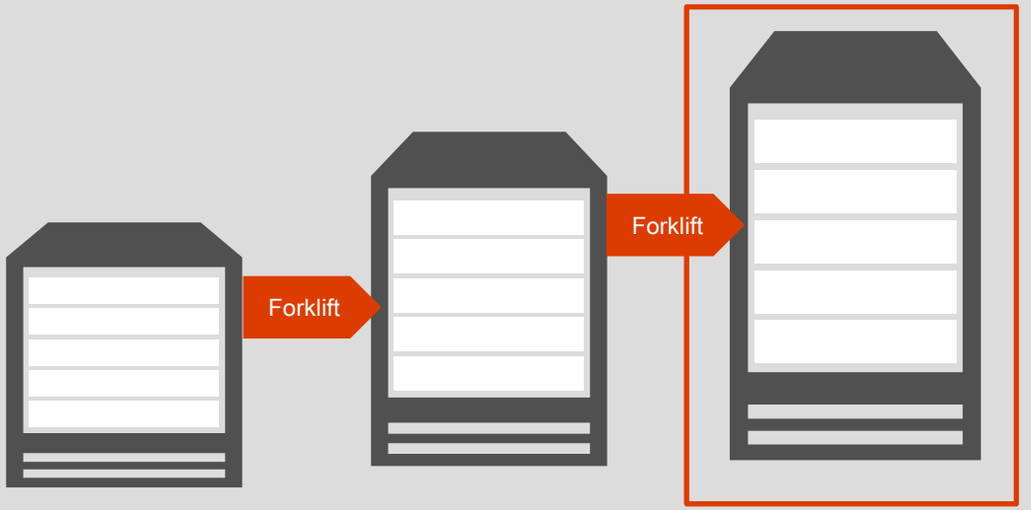
使用 System Center 进行管理与监控

Accessible insights for everyone with Microsoft BI tools

所有人可通过微软 BI 工具访问洞察力

Performance limitations and scale with a traditional data warehouse
传统数据仓库的性能与扩展性局限

Scale up
纵向扩展



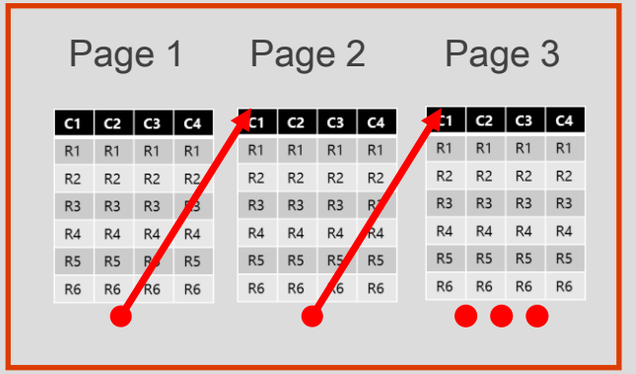
Diminishing scale as requirements grow
按需逐层扩展

Rowstore
行存储

Data数据

C1	C2	C3	C4
R1	R1	R1	R1
R2	R2	R2	R2
R3	R3	R3	R3
R4	R4	R4	R4
R5	R5	R5	R5
R6	R6	R6	R6

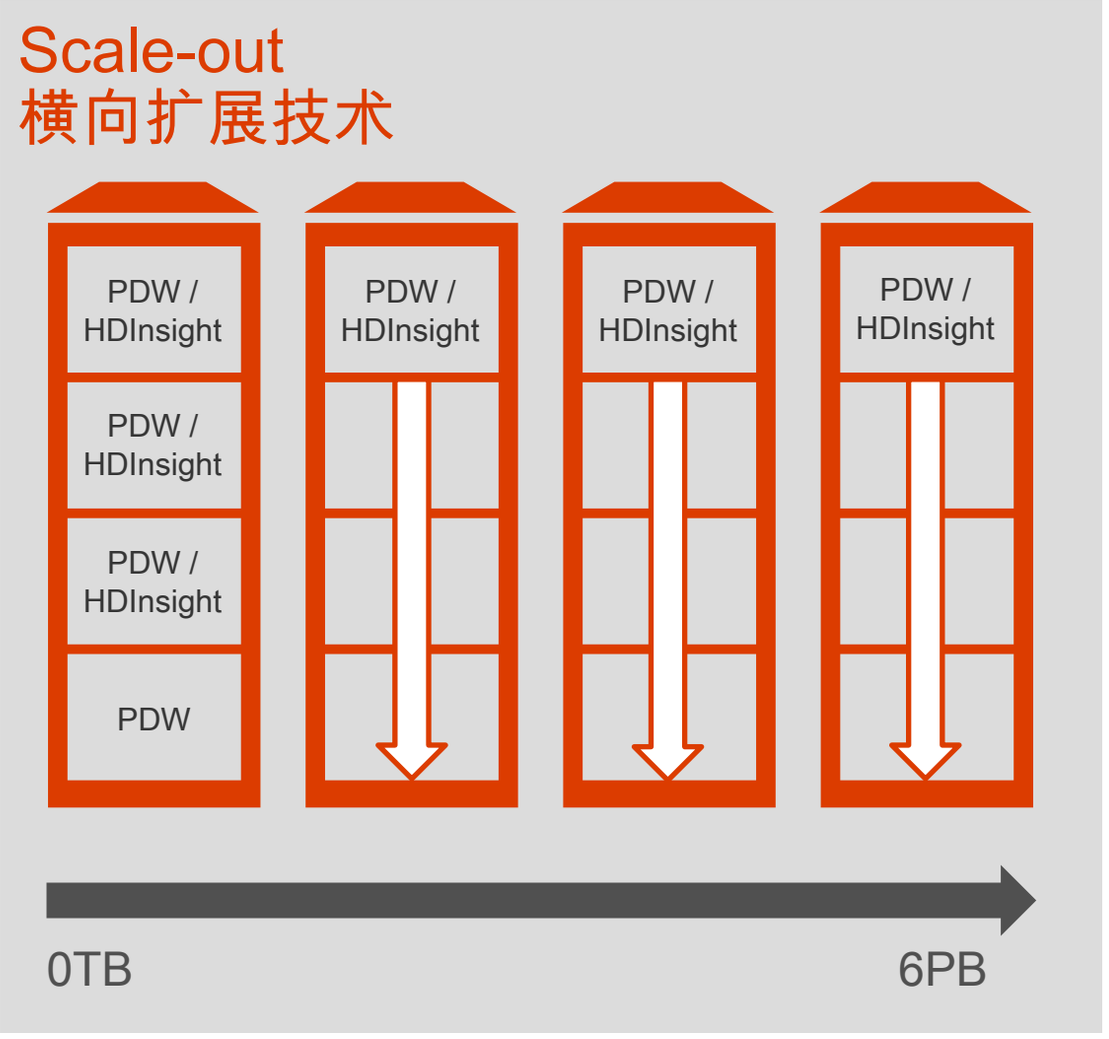
Querying data by row 逐行查询数据



Sub-optimal performance for many data warehouse queries
很多数据仓库查询无法获得最优性能

Scaling out your data to petabytes 横向扩展数据至 PB 级别

Scale-out technologies in the Parallel Data Warehouse 并行数据仓库中的横向扩展技术



Multiple nodes with dedicated CPU, memory, and storage
使用专用 CPU、内存及存储的多个节点

Ability to incrementally add hardware for near-linear scale to multiple petabytes
可逐渐添加硬件，用近乎线性的方式扩展至数 PB 级别

Ability to handle query complexity and concurrency at scale
可并行处理大规模复杂查询

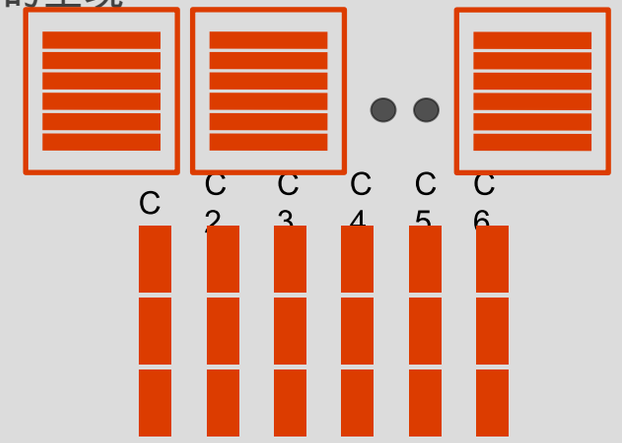
No “forklift” of prior warehouse to increase capacity
仓库扩容无需提前投入较高成本

Ability to scale out HDInsight and PDW
HDInsight 与 PDW 可横向扩展

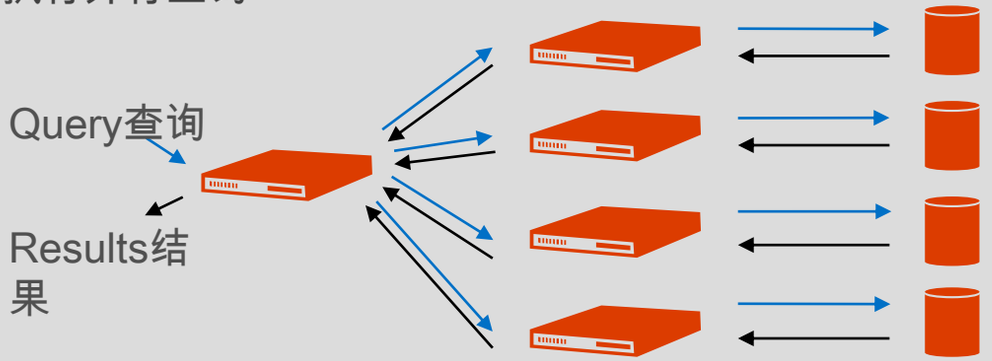
Blazing fast performance超快性能

MPP and In-memory columnstore for next-generation performance 大规模并行处理与列存储技术可提供下一代性能

Columnstore index representation
列存储索引的呈现

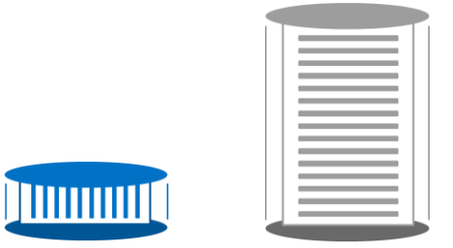


Parallel query execution
执行并行查询



Up to **100x**
faster queries
查询速度最高提速 100 倍

Up to **15x**
more compression
压缩率最高提升 15 倍



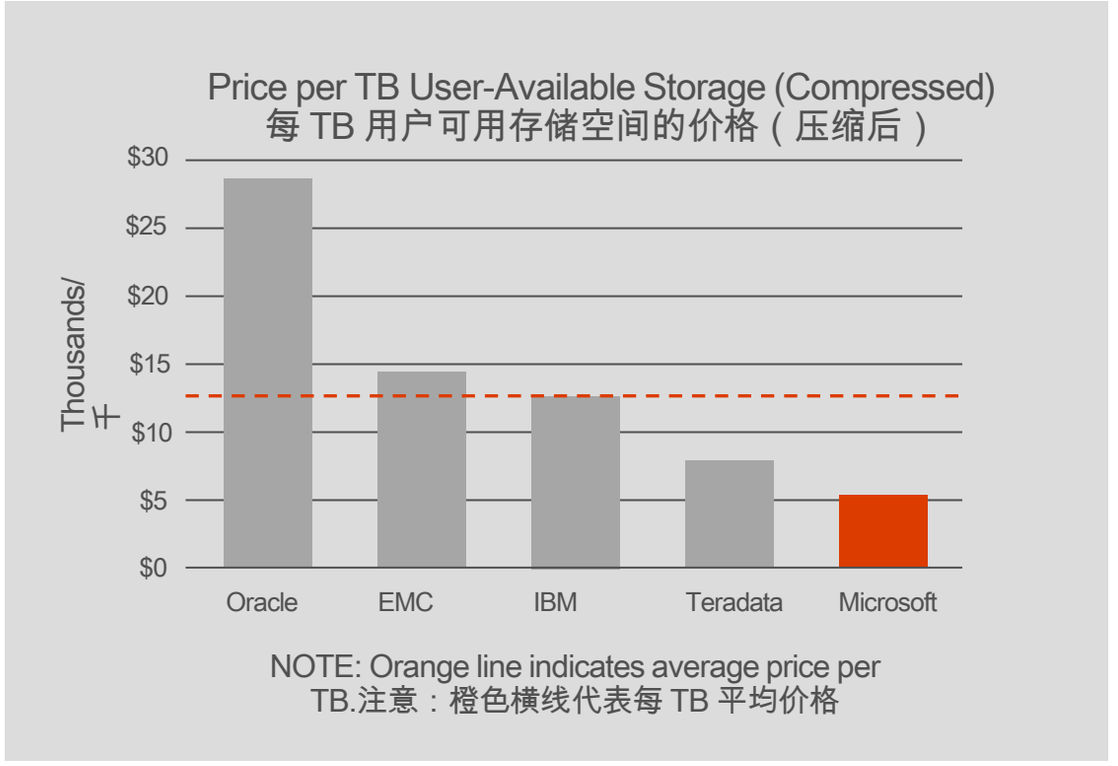
Updateable clustered columnstore vs. table with customary indexing
可更新的群集列存储相比使用传统索引的表

- Store data in columnar format for massive compression 用列格式存储数据，实现大规模压缩
- Load data into or out of memory for next-generation performance 通过内存读写数据，提供下一代性能
- Updateable and clustered for real-time trickle loading 可更新可群集，实现实时涓流加载

PDW provides the industry's lowest DW appliance price/TB
PDW 可提供业内每 TB 成本最低的数据仓库一体机

High performance using commodity hardware 使用市售硬件获得更高性能

Price per terabyte for leading vendors
主要供应商每 TB 成本



Significantly **lower price per TB** than the closest competitor
相比竞争对手大幅降低每 TB 价格

Lower storage costs with Windows Server 2012 Storage Spaces
通过 Windows Server 2012 存储空间大幅降低存储成本

Microsoft Parallel Data Warehouse微软并行数据仓库

The no-compromise modern data warehouse solution不妥协的现代化数据仓库解决方案

Meeting today's big data analytics requirements
认识当今的大数据分析需求

The modern data Warehouse
现代化数据仓库

Enterprise-ready Hadoop with HDInsight and the simplicity of PolyBase
带有 HDInsight 的企业级 Hadoop 及简化的 PolyBase

Enterprise-ready big data
企业级大数据

Optimized performance with MPP technology and In-Memory Columnstore
通过大规模并行处理与内存中列存储技术优化性能

Performance at scale
实现规模性能

Providing value with a low TCO
TCO 更低，价值更高

Optimal value
优化价值

Shinsegae Corporation, a major department store chain in Korea, needed better performance for customer data mining and basket purchase analysis. Shinsegae took advantage of Parallel Data Warehouse and Hadoop integrated together with combined data of 450 TB, Shinsegae was please to see PolyBase performing nearly twice as fast as their best Hive/Hadoop environment.

韩国大型仓储式连锁店 Shinsegae Corporation 需要改善挖掘客户数据及执行购物篮购买分析时的性能。Shinsegae 充分利用并行数据仓库与 Hadoop 技术的集成创建容量为 450TB 的数据仓库后，很高兴地发现 PolyBase 的运行速度是其他最先进 Hive/Hadoop 环境的两倍。

"We are really pleased with the performance of PolyBase to allow us to join relational and Hadoop data faster and easier"

我们对 PolyBase 的性能很满意，现在我们可以更快速简单地对关系型及 Hadoop 数据进行连接。

SHINSEGAE