大数据的实践及应用

Big Data in Action

孙巍
高级项目经理
微软云计算中心

问题
Questions

什么是大数据？
What is Big Data？

多大的数据才是大数据？
How big is Big Data？

你想从大数据里得到什么？
What do you want to get out of Big Data？

议程
Agenda

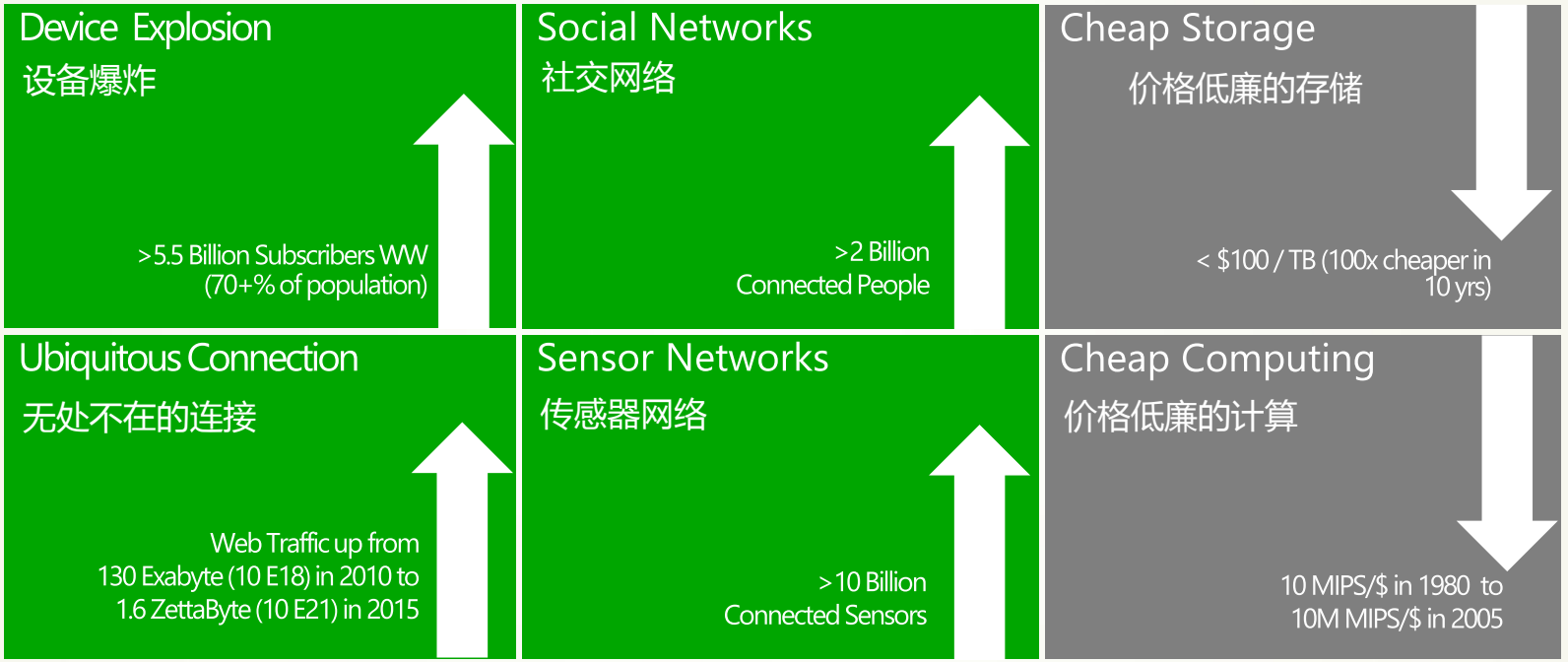
大数据简介
Big Data
Overview

大数据思考
Big Data
Rethinking

实施参考
Reference
Implementation

实施场景
Scenario & Reference

主要趋势
Key Trends



什么是大数据
What Is Big Data?

一系列新问题
A New Set Of Questions

SOCIAL & WEB ANALYTICS
社交网络和互联网分析

What's the social sentiment for my brand or products ?
我的品牌或产品情绪

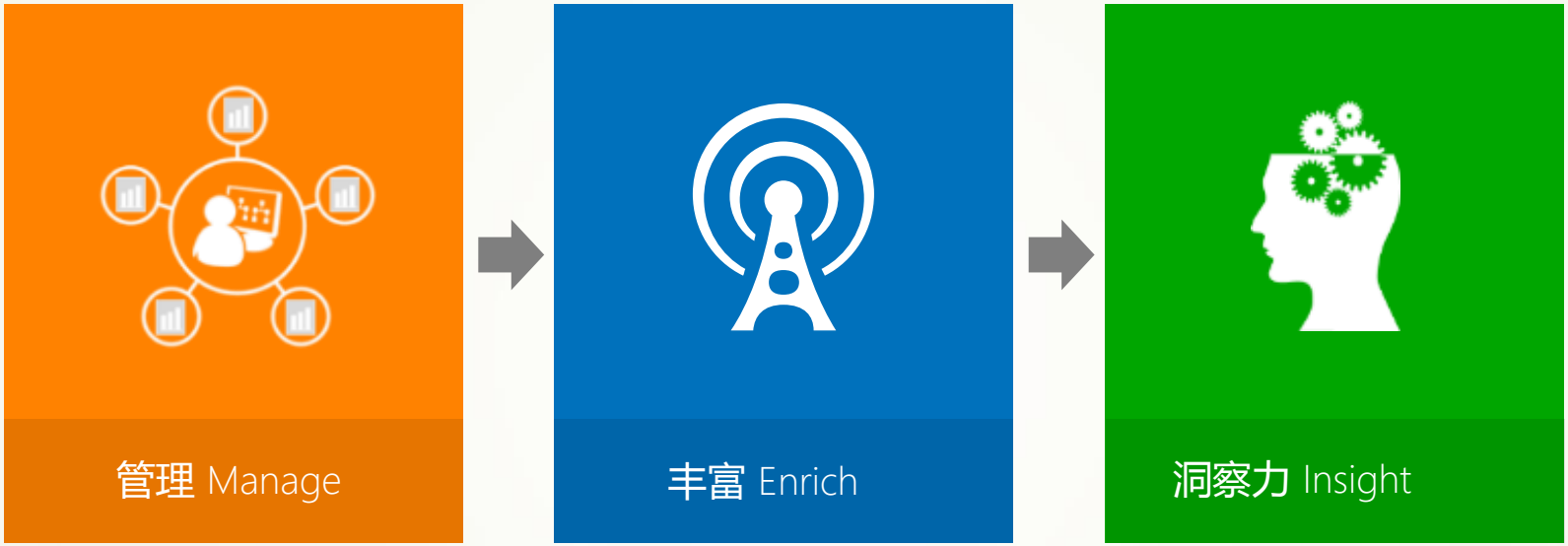
LIVE DATA FEEDS
实时数据源

How do I optimize my fleet based on weather and traffic patterns?
如何优化我的车队运行 (基于天气和交通趋势)

ADVANCED ANALYTICS
高级分析功能

How do I better predict future outcomes?
如何更好预测未来结果？

大数据生命周期
The Big Data Lifecycle



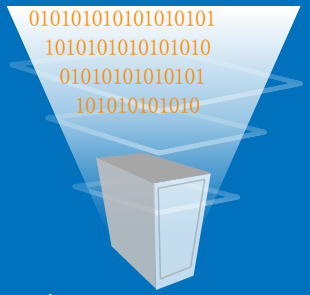
管理任何种类、大小、 来源的数据 DTCC2013
Manage Any Data, Any Size, Anywhere

Microsoft SQL Server 2012

统一监控、 管理和安全
Unified Monitoring, Management & Security

关系型
Relational

非关系型
Non-Relational



数据流
Streaming

数据移动 Data Movement

HADOOP 集成
HADOOP Integration

Non-Relational

Microsoft SQL Server

Windows Azure

Hadoop-based distribution on premises
Hadoop-based service in the cloud

企业级安全，高可靠性，管理 Enterprise class security, HA & management
与微软商业智能工具无缝集成 Seamlessly integrated with Microsoft BI tools
SQL Server 数据平台的一部分 Delivered as part of the SQL Server Data Platform
在Windows Azure上几分钟内完成部署 Provisioned in minutes on Windows Azure

开放和灵活
Open & Flexible

与Apache Hadoop 100%兼容
100% compatible with Apache Hadoop

工具由丰富的合作伙伴生态系统提供
Tools from a rich ecosystem of partners

与社区的紧密合作
Built with close community collaboration

hortonworks

Accelerating the delivery of Hadoop for Windows

HSTREAMING

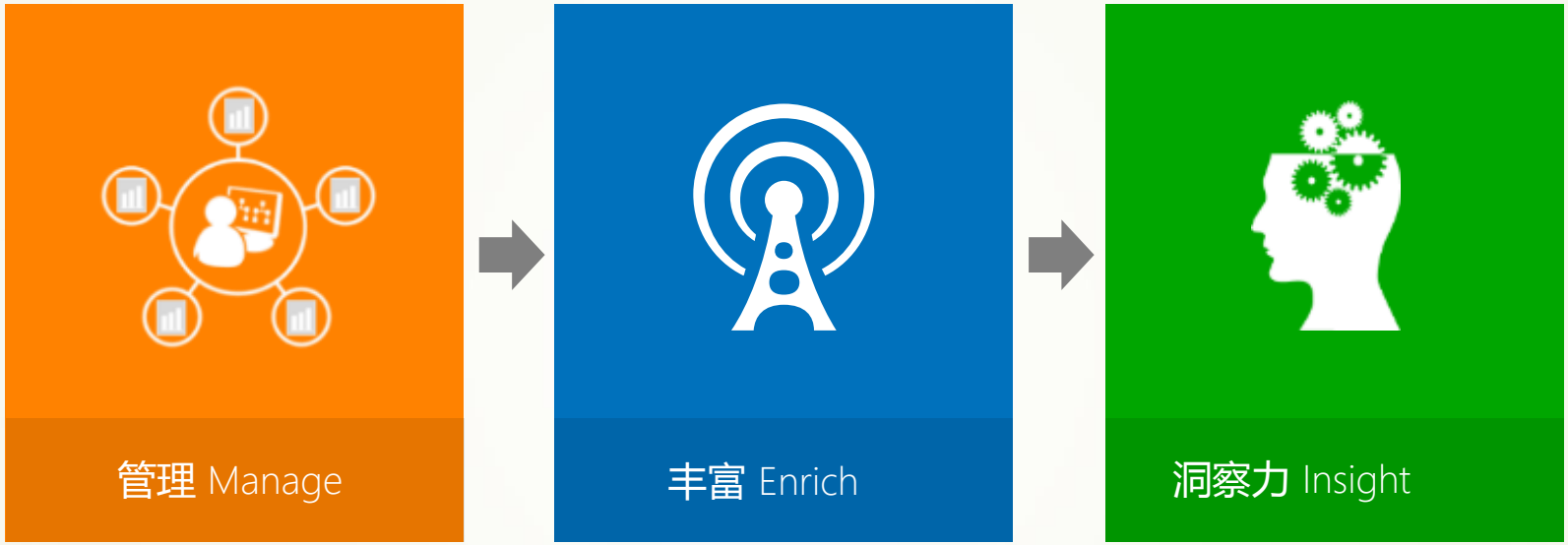
Datameer
Powerfully Simple™

KARMASPHERE

The Apache Software Foundation

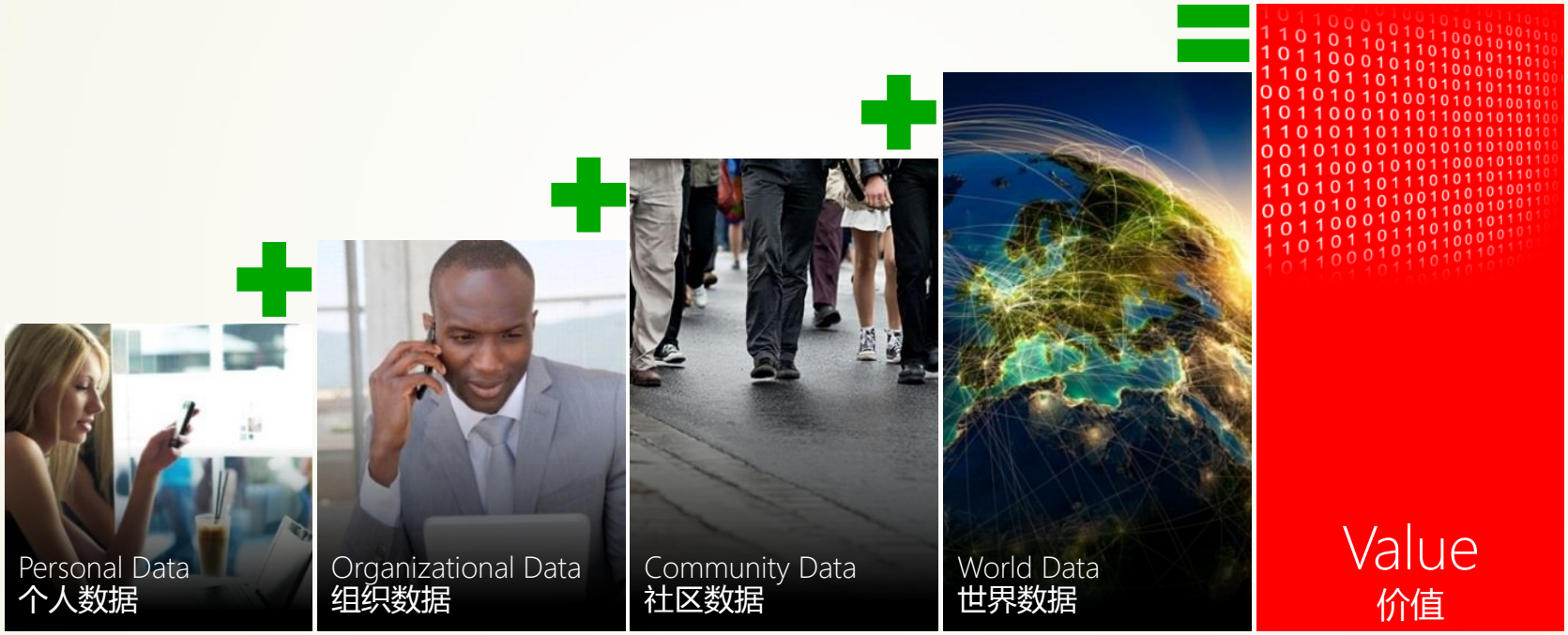
Hadoop for Windows JavaScript libraries
Hive ODBC drivers

大数据生命周期
The Big Data Lifecycle



连接数据集市产生更多价值
Enrich By Connecting To The Worlds Data

数据整合带来的价值
Power Of Combining The Worlds Data



数据集市
Data market – Windows Azure Marketplace

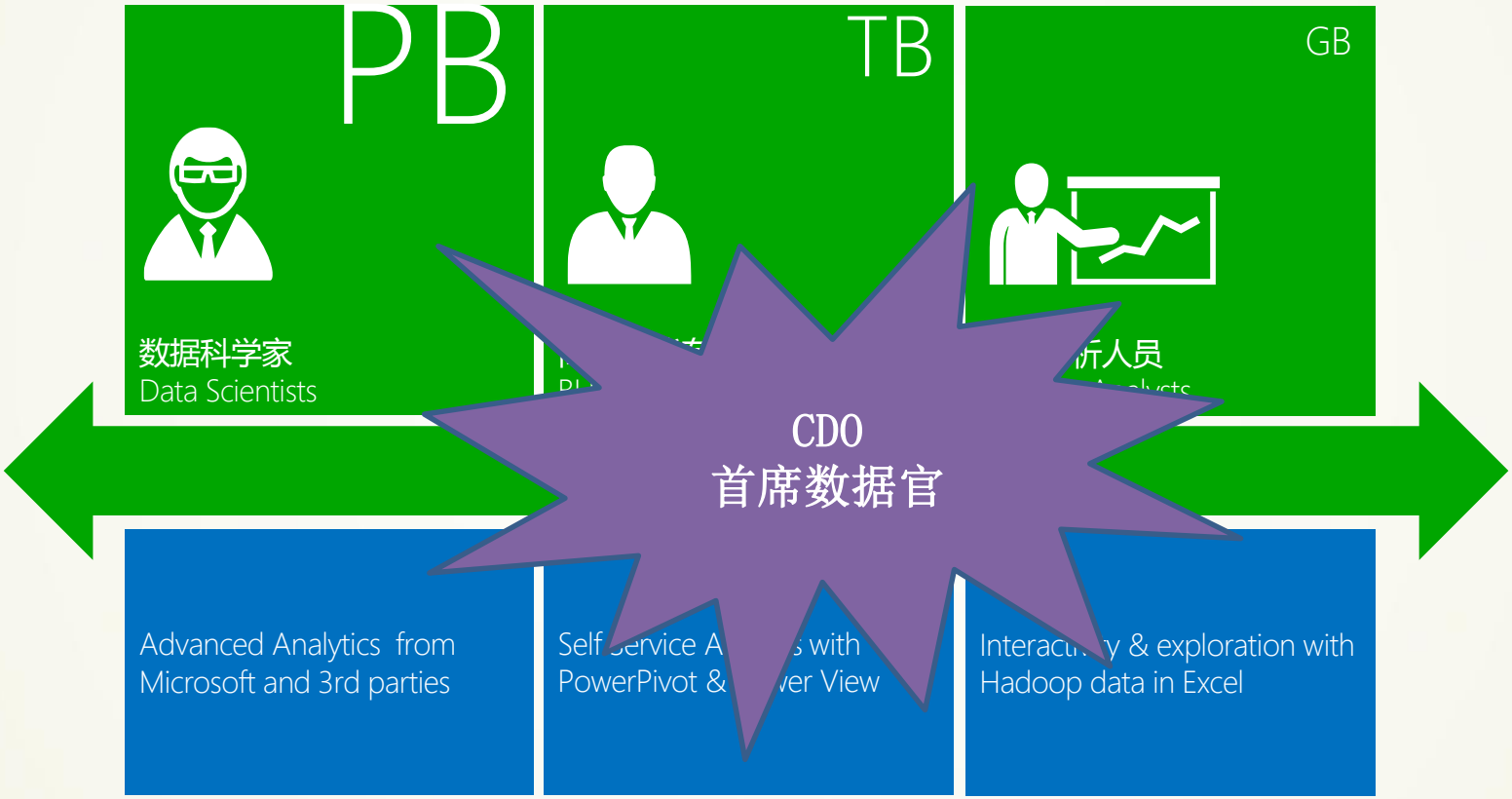
内容提供商 Content Providers	独立软件开发商及开发者 ISVs and Devs	信息工作者 Information Workers
• Global reach • Unified billing & provisioning platform • Easy content onboarding • Data security / authorization model • Flexible pricing, auditing, logging	• Consistent, flexible , context optimized APIs - OData • Single Contract – One Stop shop for data • Easy access to premium data • Unified billing and provisioning platform	• Easy of Discovery and Rich Apps to Consume Data • Microsoft Office, Dynamics, Bing + 3rd party ISV Applications • Ability to mash up public and private data • Flexible pricing – pay as you go

大数据生命周期
The Big Data Lifecycle

对任何种类、大小、来源数据的洞察力

Insights On Any Data, All Users, Whatever They Are

通过熟悉的工具，为所有用户提供对数据的洞察力
Insights For All Users Through Familiar Tools



客户示例

KLOUT

Connects to more than 1 billion signals
连接到超过 10 亿的信号/数据源
Across 15 leading social networks, including Facebook
排名前15位的社交网络，包括Facebook
Generates a 'Klout' score for individual people, brands & partners
为个人、 品牌及合作伙伴生成一个 'Klout' 分数
Enables analysis, targeting and social graphs
提供分析、 目标和社交图

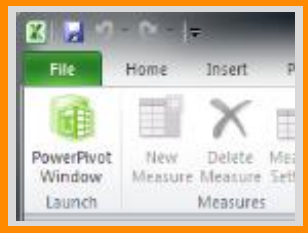
"When it comes to business intelligence, Microsoft SQL Server 2012 demonstrates that the platform has continued to advance and keep up with the innovations that are happening in big data.
在商业智能领域，Microsoft SQL Server 2012平台持续发展，支持不断创新的大数据平台。"

David Mariani
Vice President of Engineering
工程副总裁

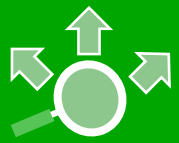
端到端的大数据解决方案
Big Data Requires An End-To-End Apporoach

洞察力
INSIGHTS

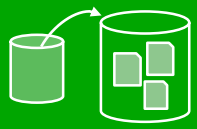
- SELF-SERVICE
- COLLABORATIVE
- MOBILE
- REAL-TIME



丰富数据
DATA
ENRICHMENT



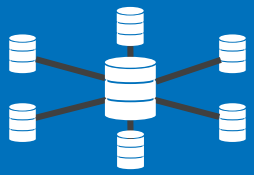
DISCOVER
AND RECOMMEND



TRANSFORM
AND CLEAN

SHARE
AND GOVERN

数据管理
DATA
MANAGEMENT

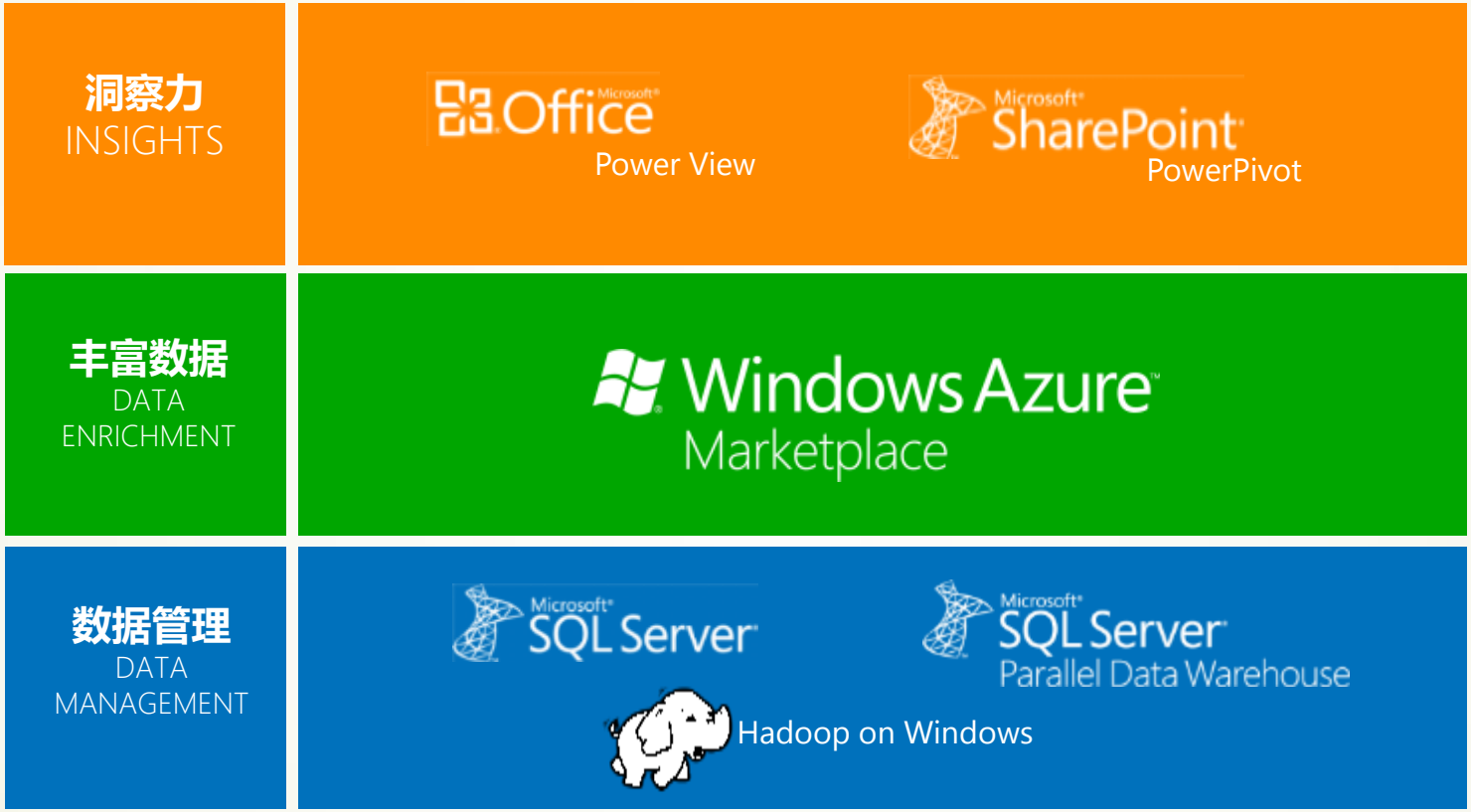


RELATIONAL

NON-RELATIONAL

STREAMING

微软大数据
Microsoft Big Data



议程
Agenda

大数据简介
Big Data
Overview

大数据思考
Big Data
Rethinking

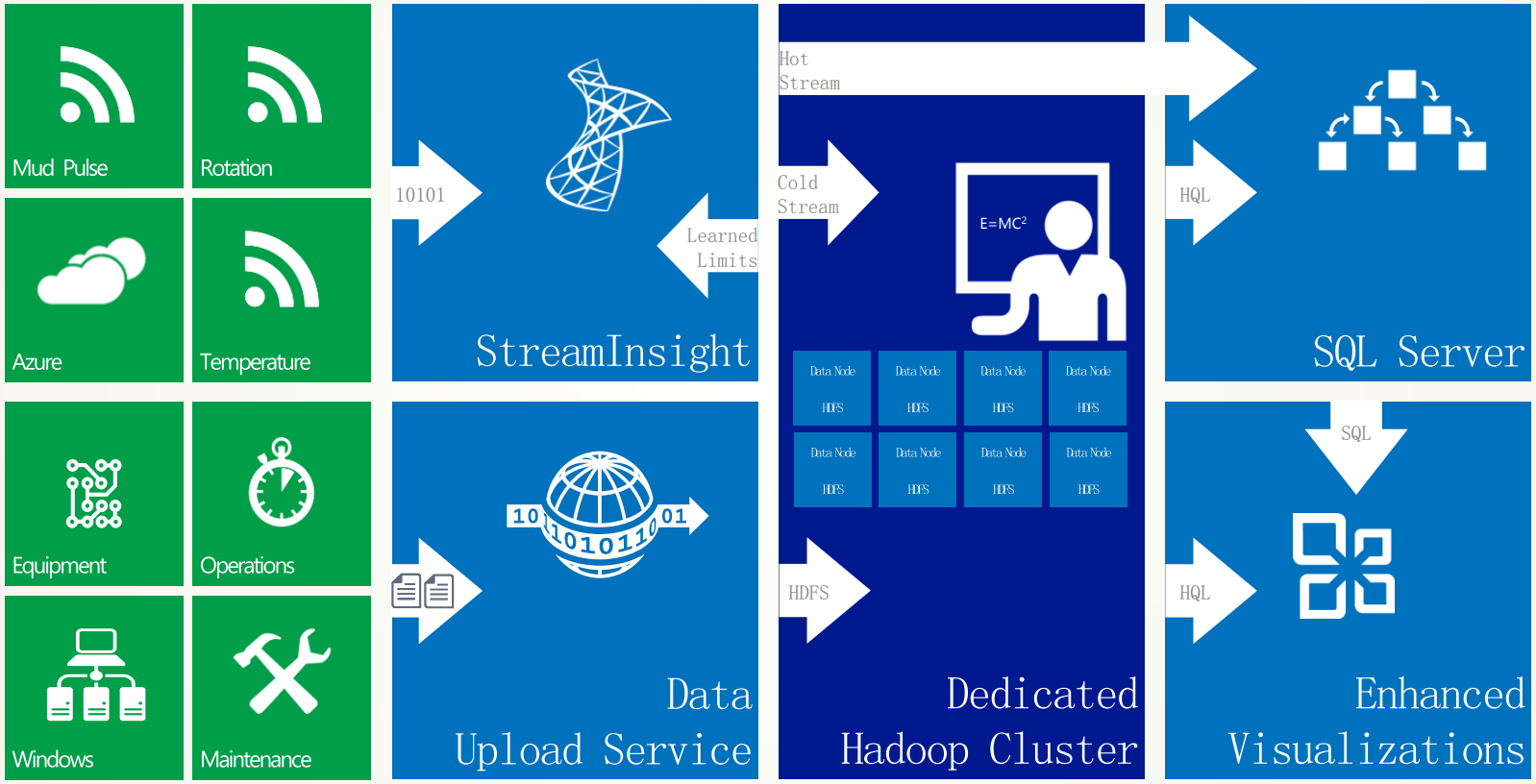
实施参考
Reference
Implementation

实施场景
Scenario & Reference

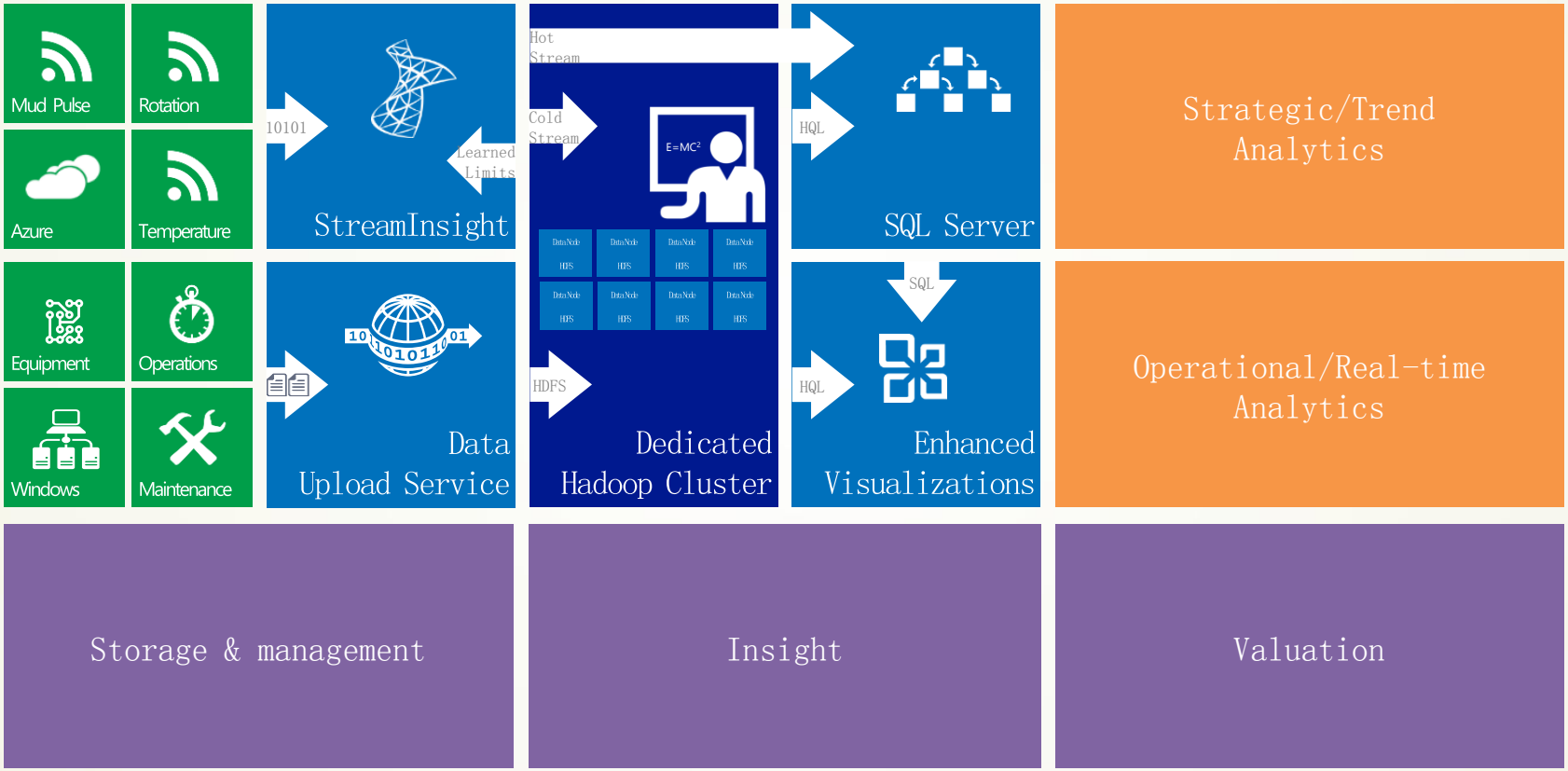
大数据的再思考
Re-thinking BIG DATA

大数据定位 The Big Data Positioning	A New Era with new data technology and technique that manage, analyze and create value with data of modern characteristics (the "V"s)
大数据数量 The Big Data Volume	Big Data is not defined by volume only, but by any of the "V" characteristics. And volume is as large as you want it to be, or you can afford it to be.
大数据目的 Why Big Data	Big Data is about using new technology and technique to transform, and through intelligence from data, explore new value

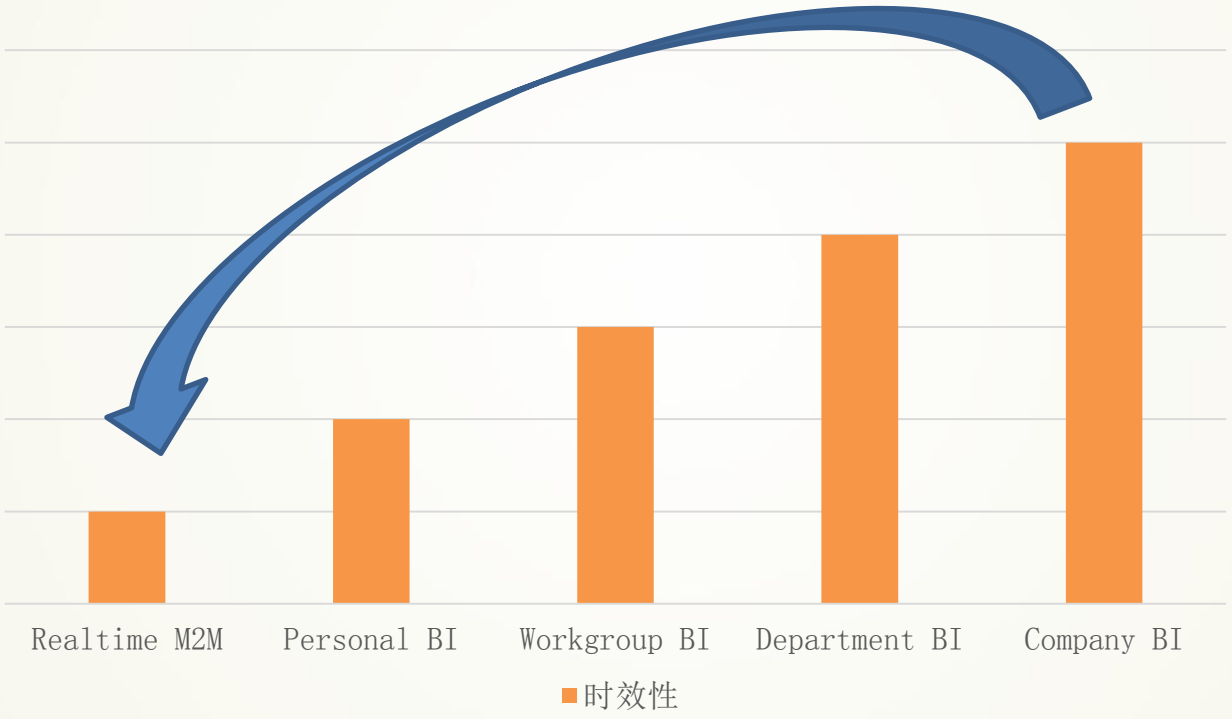
典型大数据数据分析场景
Typical Big Data End-to-End Analytics



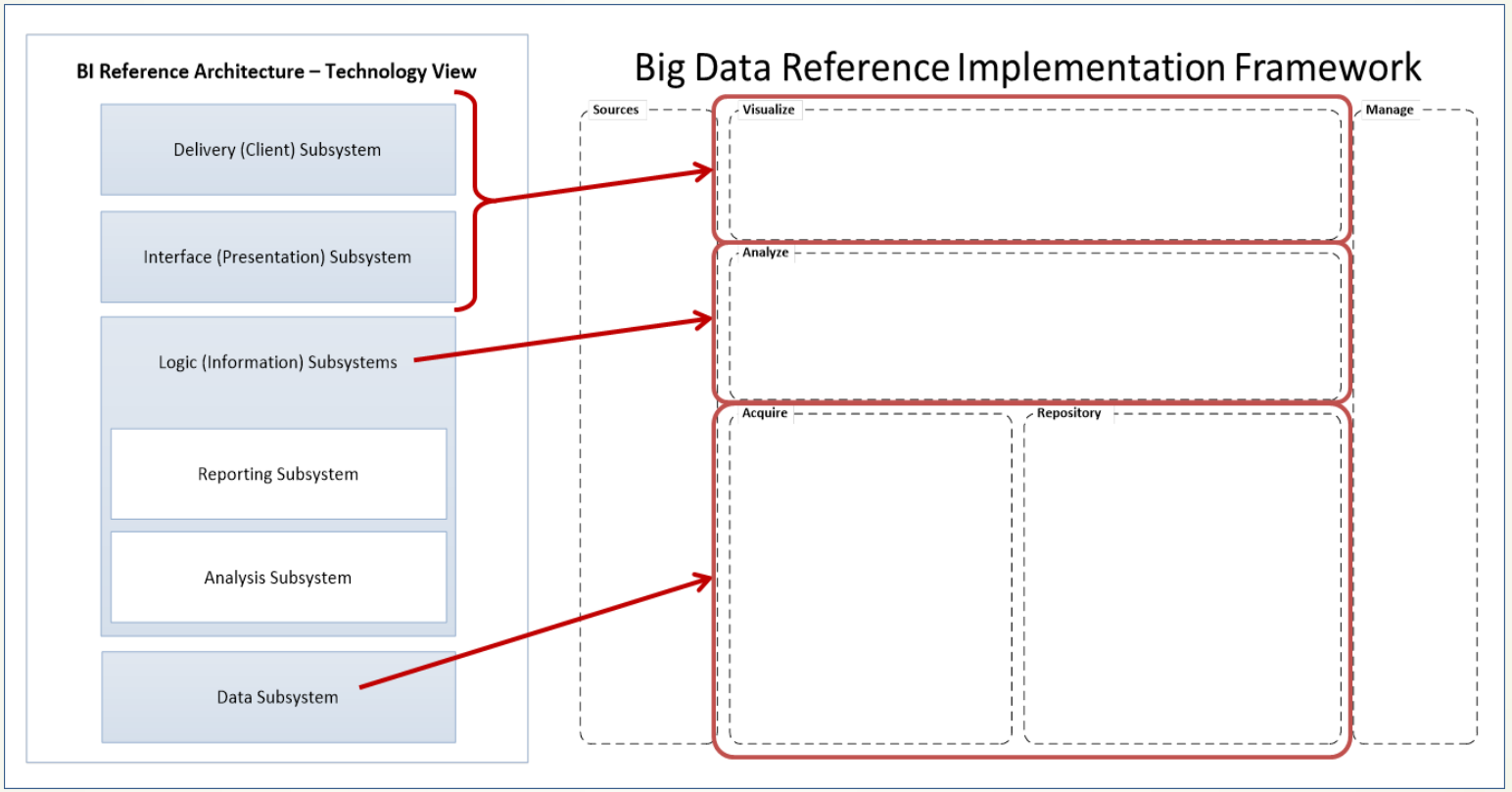
端到端的大数据生命周期
Typical Big Data End-to-End Analytics



大数据的时效性
New Thinking of Big Data



实施框架参考
Reference Implementation Framework

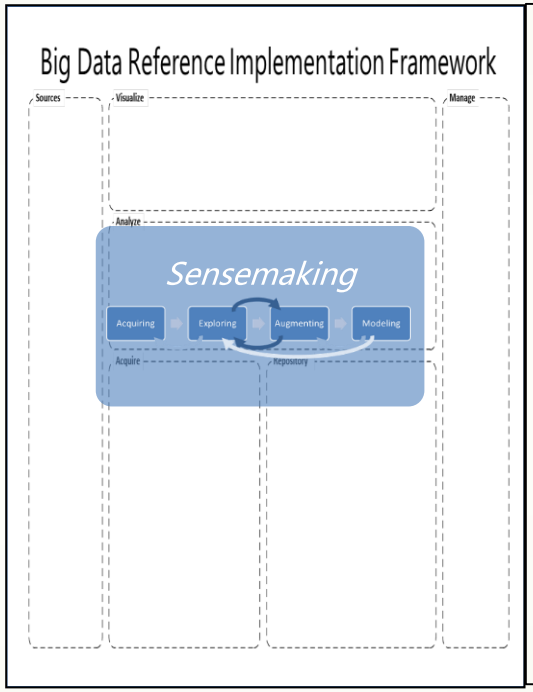


大数据和传统BI的差别
Big Data and Traditional BI Difference

Big Data

Schema on Read

- 数据架构模型在查询时动态定义
- 更具探索性，需要行业知识
- 目标是在环境数据中寻找新的价值
- *…You don’t know what you don’t know…*

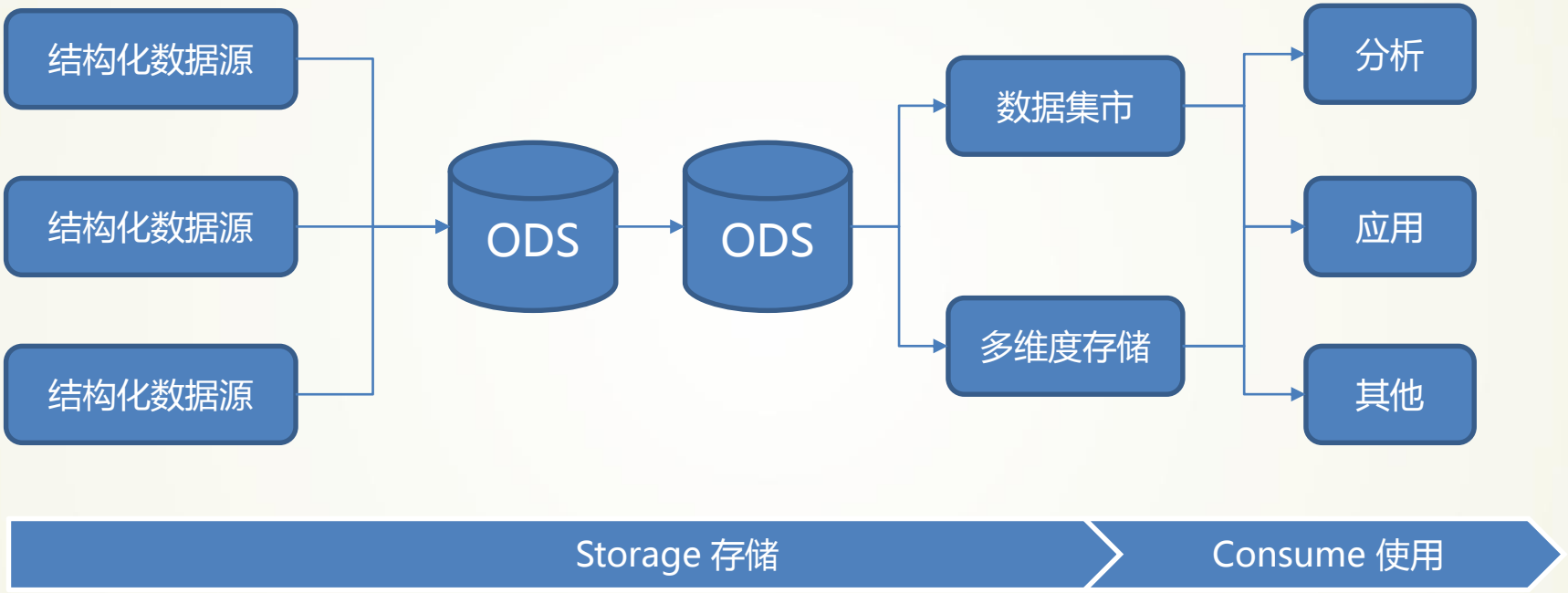


Traditional BI

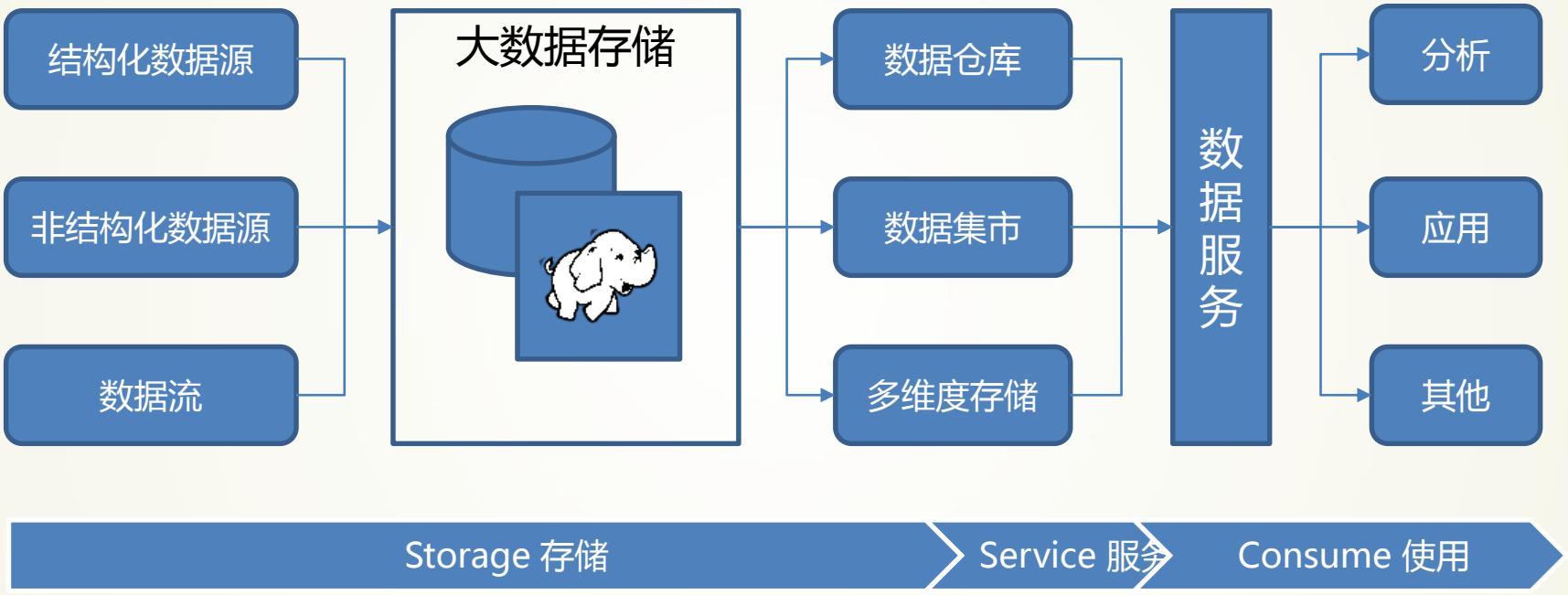
Schema on Write

- 数据架构模型在写入时已经定义
- 体现明确定义的标准及KPI
- 成熟的开发模式及丰富的实践经验
- *…Show me what I already know…*

企业数据及商业智能平台的进化
Evolution of the BI/Data Platform



企业数据及商业智能平台的进化
Evolution of the BI/Data Platform



大数据时代的工作角色转变
Big Data Job Roles

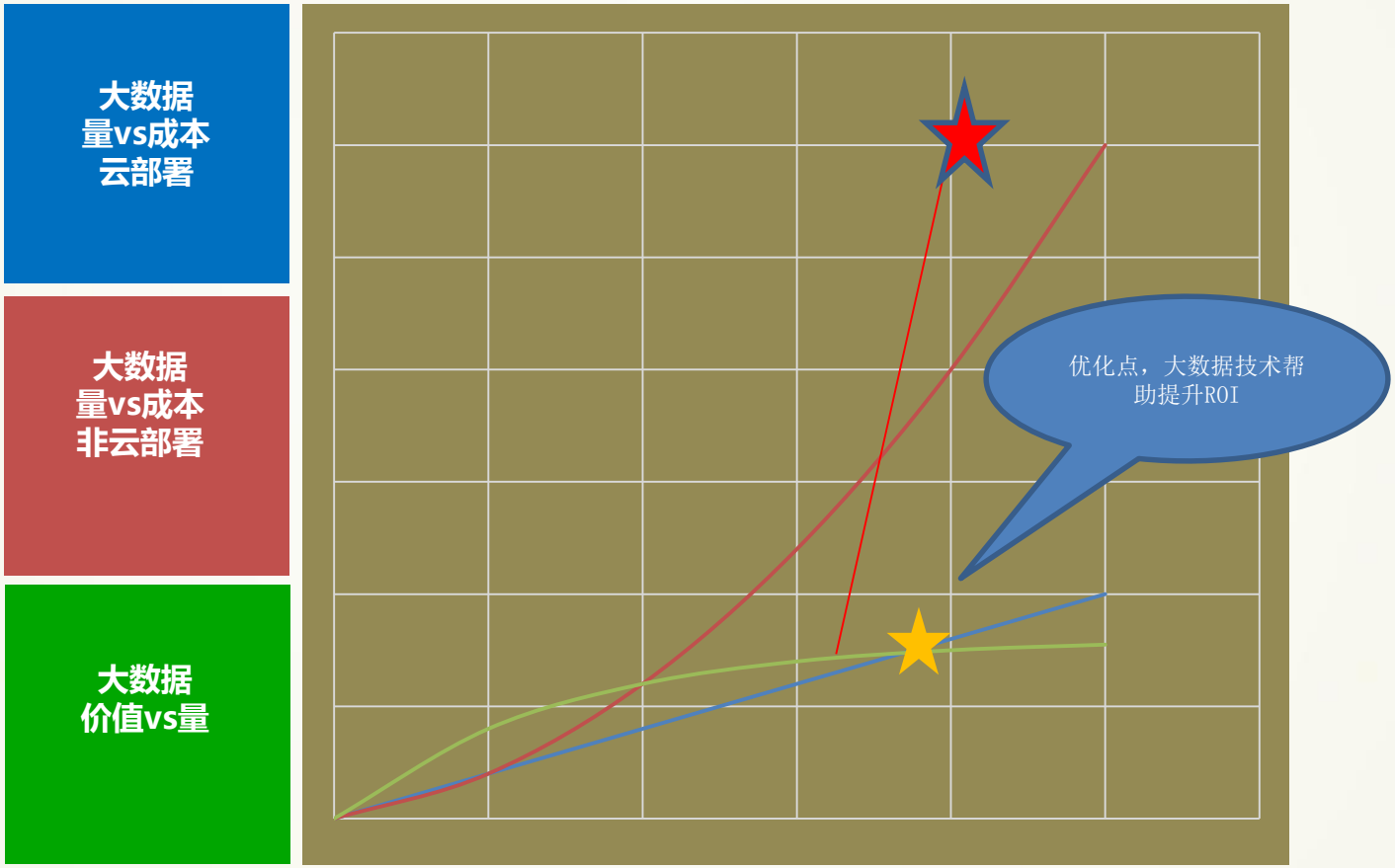
首席数据官
Chief Data Officer

数据分析师
Data Scientist

行业经验和大数据智能
Industry Experience vs. Data Intelligence

智能运维优化
Intelligent Operation Optimization

企业大数据的优化
Big Data ROI Optimization



议程
Agenda

大数据简介
Big Data
Overview

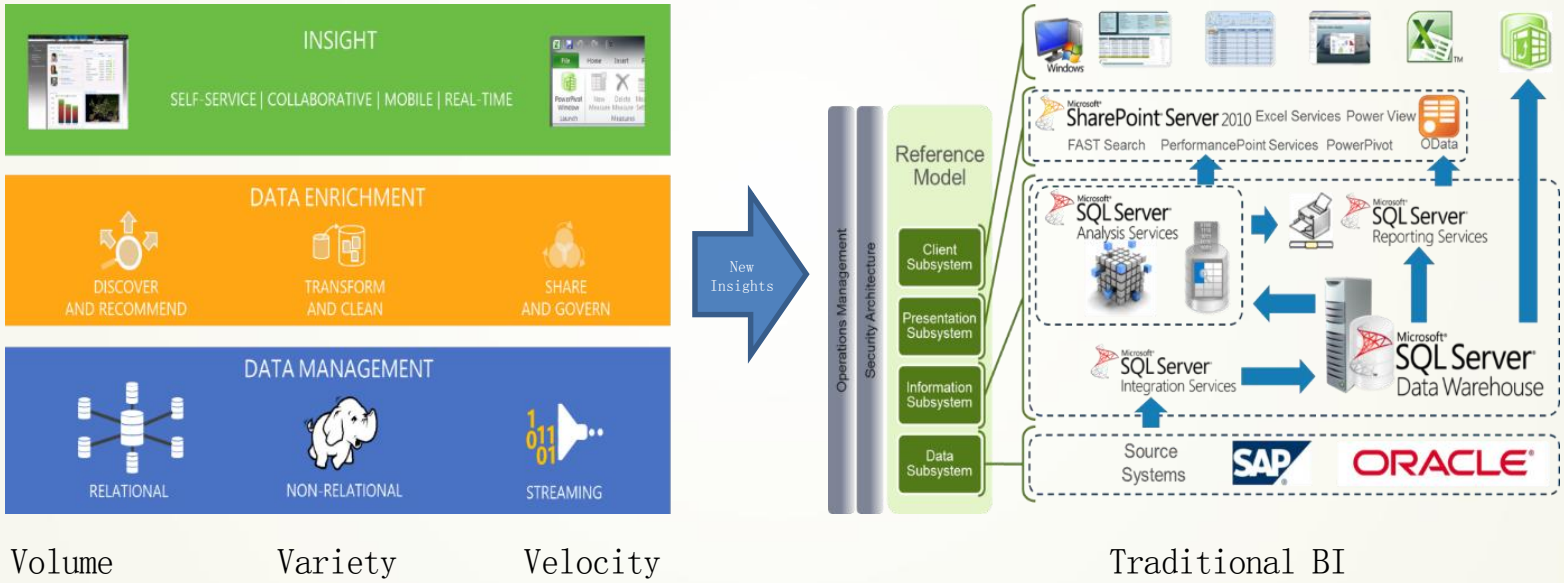
大数据思考
Big Data
Rethinking

实施参考
Reference
Implementation

实施场景
Scenario & Reference

大数据的新机遇
New Opportunities

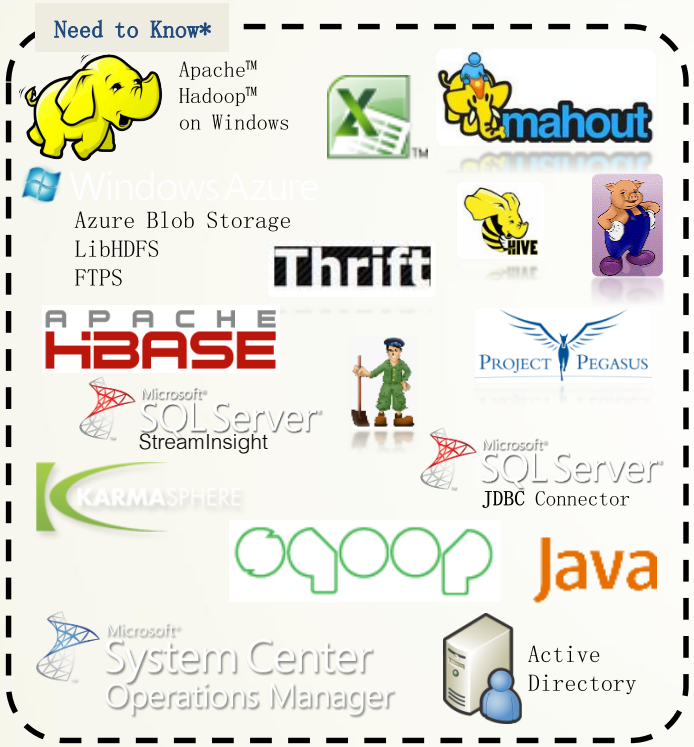
Data Scientist — Information Worker — Casual User

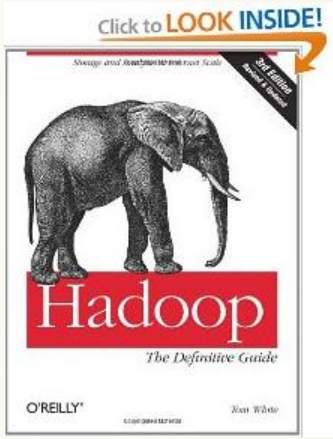


Volume Variety Velocity

Traditional BI

Reference Implementation – Products +





Hadoop: *The Definitive Guide* 3rd Ed.
– Tom White, O'Reilly Books

议程
Agenda

大数据简介
Big Data
Overview

大数据思考
Big Data
Rethinking

实施参考
Reference
Implementation

实施场景
Scenario & Reference

网站/社交网络场景
Web / Social

-'Creator of Hadoop uses SQL Server'
-Largest cube in the world @ 24TB w/ 2PB source
-Helping us performance test Hadoop to SSAS
-Plenty of PR

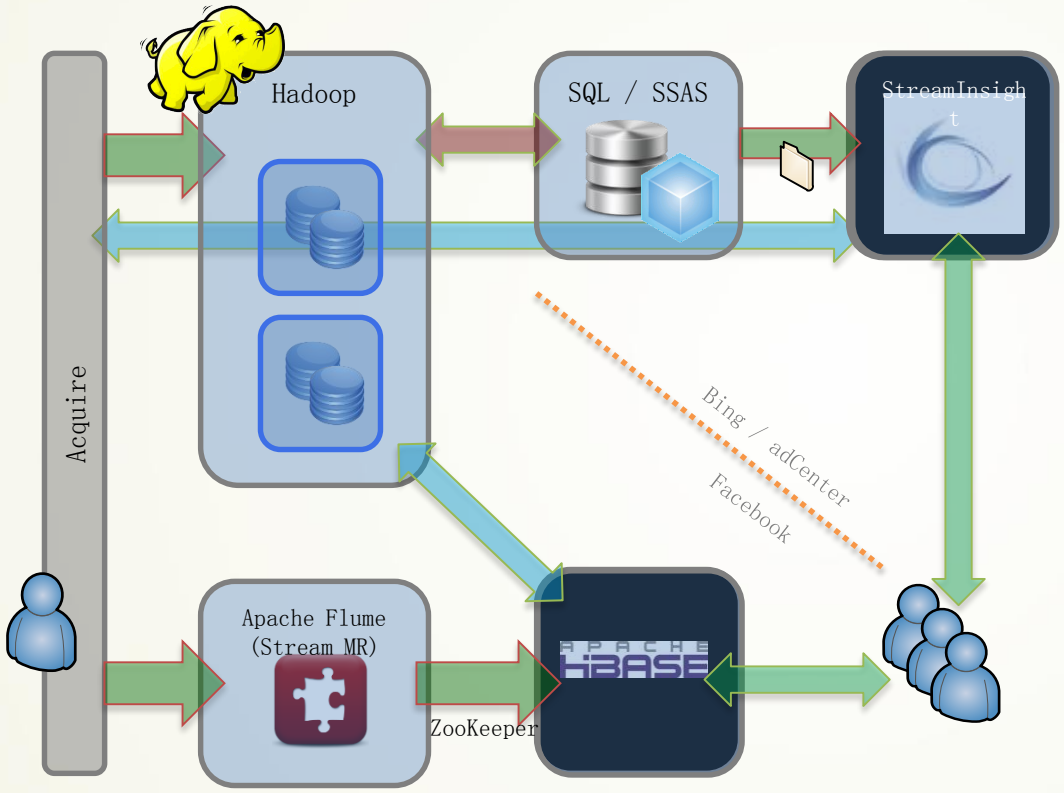
- Uber PR (Strata, blogs, CIO magazine and webcast, etc.)
- Hadoop to SSAS
- Social Media darling

webtrends

- Ultimate web analytics scenarios including RT
- Hadoop/HBase to SQL
- Will provide PR

- Ultimate OSS / build it here shop
- Talking to us because of Hadoop
- BI, Data Sharing, Knowledge Sharing scenarios

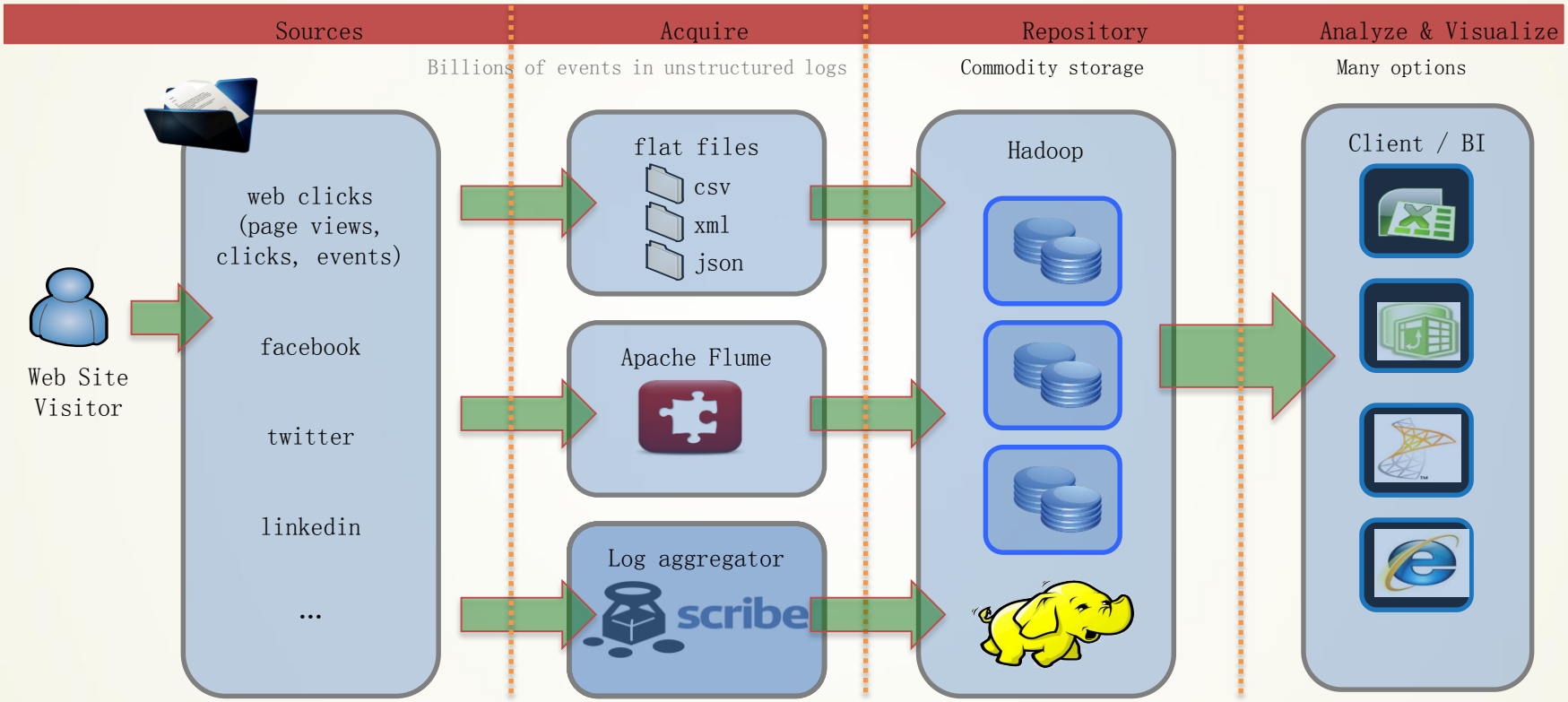
实时事态处理
Real Time Event Processing



Bing/adCenter Event Processing
- Display ads on msn.com
- Data goes into *Hadoop*
- ETL into SQL/SSAS
- Model for SI to use
- SI processes via model
- Updated display ad (latency <1min)
- Processing all 550B+ MSN users

Facebook Real Time Messaging
- Short set of volatile temporal data
- Continually growing dataset rarely accessed
- 20B events/day, 200,000 events/sec
- Latecy <30s

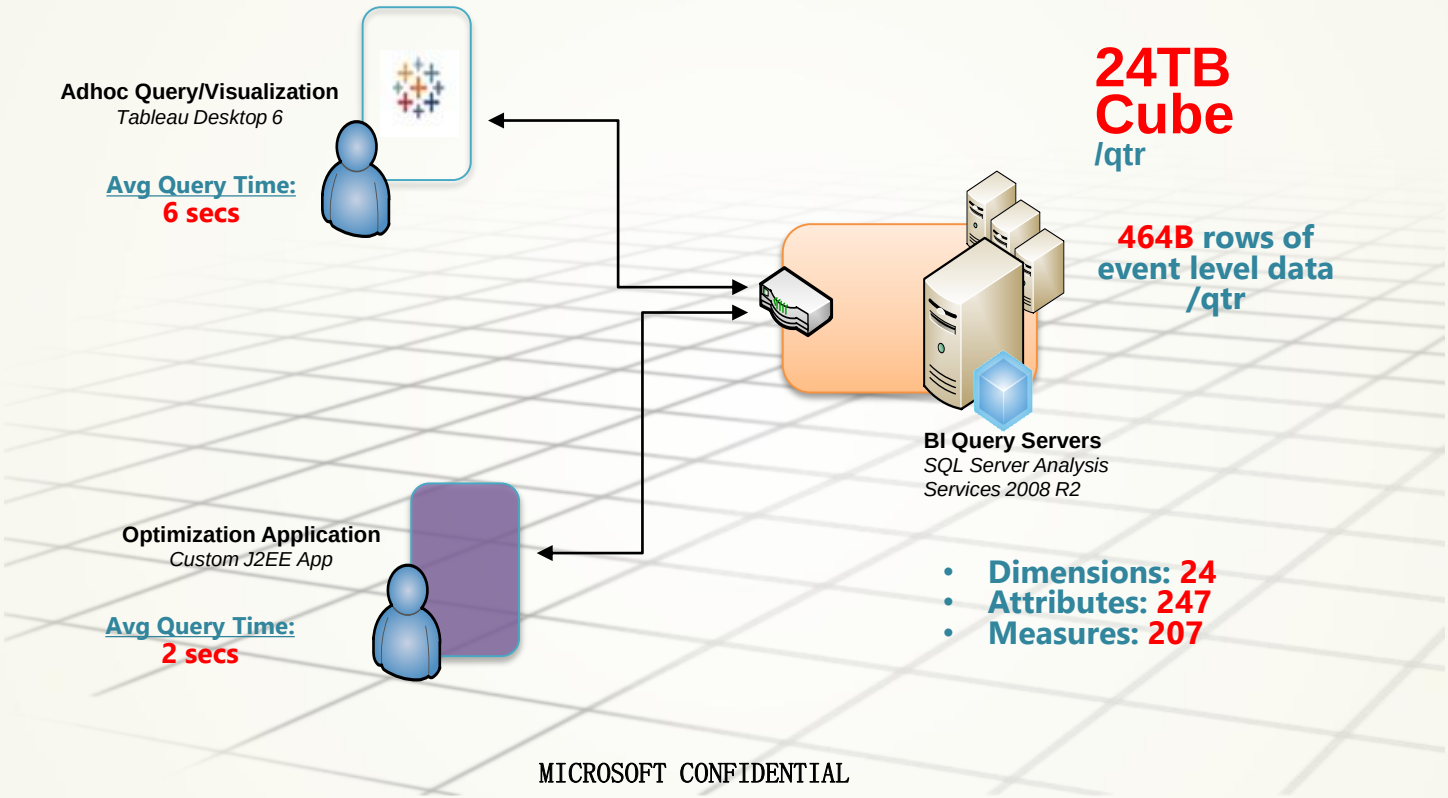
网站/社交网络场景
Web / Social



某全球著名互联网公司的大数据挑战 XYZ’s Big Data Problem

- 680,000,000 Visitors to XYZ Branded Sites
- 3,500,000,000 Ad impressions per day
- 35,000,000,000 Ad Impressions x Segments
- 464,000,000,000 Additional Rows per Quarter
- Hourly Refresh Frequency
- <6s Average Adhoc Query Time
- <2s Average Report Query Time

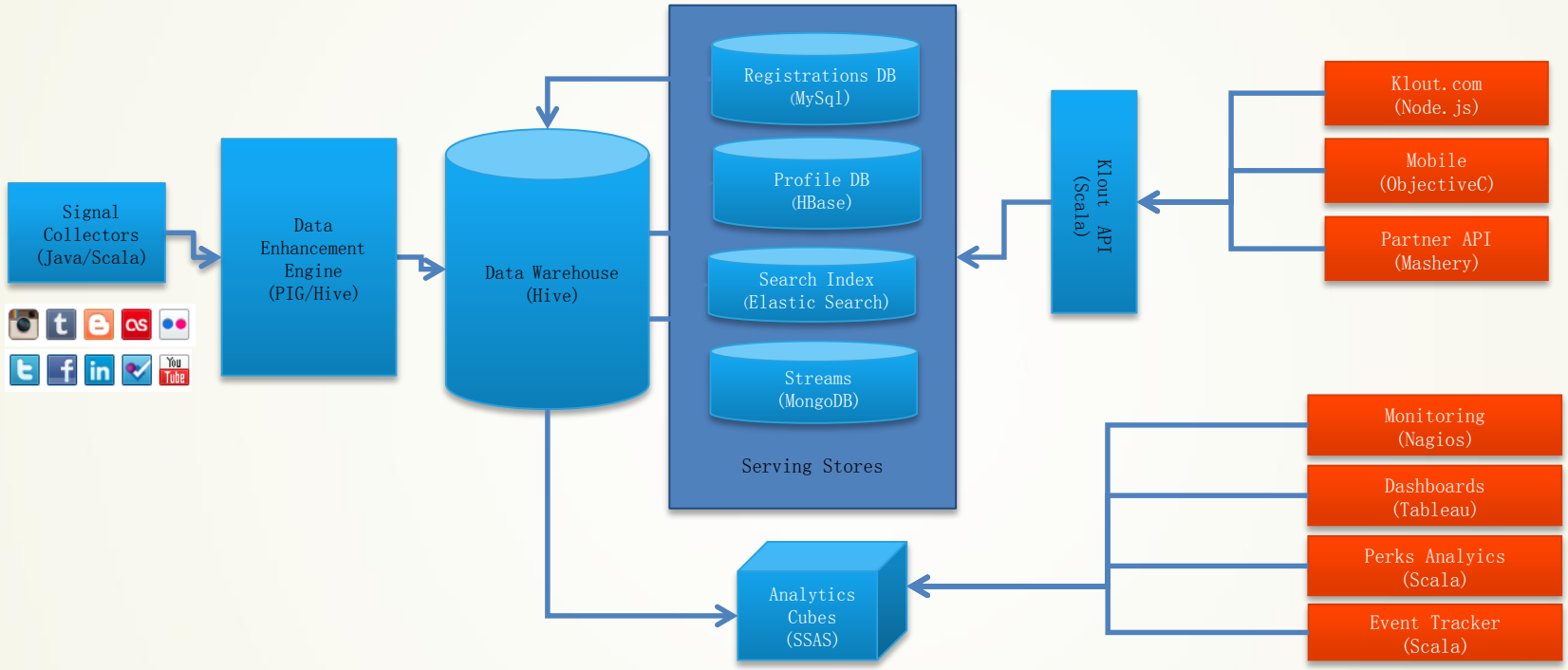
某全球著名互联网公司的大数据平台 XYZ’s Big Data Platform



Klout's Big Data Problem

- 15 Social Networks Processed Every Day
- 120 Terabytes of Data Storage
- 200,000 Indexed Users Added Every Day
- 140,000,000 Users Indexed Every Day
- 1,000,000,000 Social Signals Processed Every Day
- 30,000,000,000 API Calls Delivered Every Month
- 54,000,000,000 Rows of Data In Klout Data Warehouse

Klout Data Architecture



医疗卫生场景 Healthcare

- 临床试验：不只是审查现有药物的疗效，但也是潜在的偏差
  - 例如， 伟哥原先是为治疗低血压及心绞痛等病症研发的，但现在甚至用于新生儿肺动脉高压及高原反应
- 预测医疗保健的发病率问题
- 社交媒体药品广告的宣传效果
- 药品市场活动及广告效应分析
  - 为消费者建立分析模型进行行为分析，试图了解他们的用户行为（他们为什么要购买这种药物，他们如何看待他们的疾病，相关行为等）

医疗卫生场景
Healthcare

- 高新技术的采用相对迟缓
- 人体科学研究是一个例外，经常采用革命性的前沿技术
- 遗传因子等研究带来对人体科学更深入的认识
- 蛋白质结构的研究帮助研发为个人定制的药品
- 医疗病症的防治：心脏病突发，或者哮喘

政府及公用事业场景
Government / Utilities

- 评估消费者的决策和及针对绿色能源趋势的情绪
- 智能电网的负荷管理和有针对性的营销（如智能城市）
- 有针对性的市场营销和性能
- 公用事业市场

Government & Utilities

working closely with MS Federal team
- Government organizations were involved in the early prototypes of Hadoop
- They represent “Big Data” in so many ways
- MS Federal even have their own stamp/SKU for their own version of private cloud
- Prototypical surround strategy

CLP 中電

- Prototypical Chinese customer = long term relationship building
- As well, very innovative and willing to push boundaries
- Need more smart grid evidence against competitors
- Ned to work better with SAP (StreamInsight, BI, Big Data, etc.)

石油、天然气行业场景
Oil and Gas

- 地质数据处理
  - 大部分的数据处理采用20世纪50年代的地质研究的算法
  - Chevron雪佛龙公司拥有3000个节点的Linux集群来处理这个数据，有时间计算需要超过一年时间
  - Hadoop运行大规模的并行计算
- 新一代应用
  - WITSML数据处理（井场信息传输标准标记语言XML格式），通过Hive XML SerDe
  - 应用当前的BI工具，以了解和模拟数据
  - 使用 Stream Insight / Storm 实时出发
- 数据共享的场景

金融服务行业场景
Financial Service

- Financial Organizations have a lot of Consumer information
  - Customer Payment Information and Habits
  - Credit Reports
- How to mine the data itself – i.e. the Data is the IP
- Heavy SAS users but willing to switch to R
  - Willingness to go to Azure for Data Sharing scenarios
  - Private Cloud to share data with their partners
  - But Governance, Risk, Compliance scenarios are on top of their minds

其他金融行业场景
Other Financial Service Workloads

其他资源
Additional Resources

LEARN MORE

- Microsoft Big Data Solution: www.microsoft.com/bigdata
- Windows Azure: www.windowsazure.com/en-us/home/scenarios/big-data
- Microsoft BI blog: http://blogs.msdn.com/b/microsoft_business_intelligence1/

TRY NOW

- Preview of the Hadoop-based service for Windows Azure: https://www.hadooponazure.com

欢迎莅临

2013中国数据库技术大会