

SAP携手Intel, 构建实时大数据平台

朱 旻(Benjamin Zhu), SAP HANA中国业务拓展总监
程从超, Intel云创新中心解决方案架构师
Jun 17, 2014

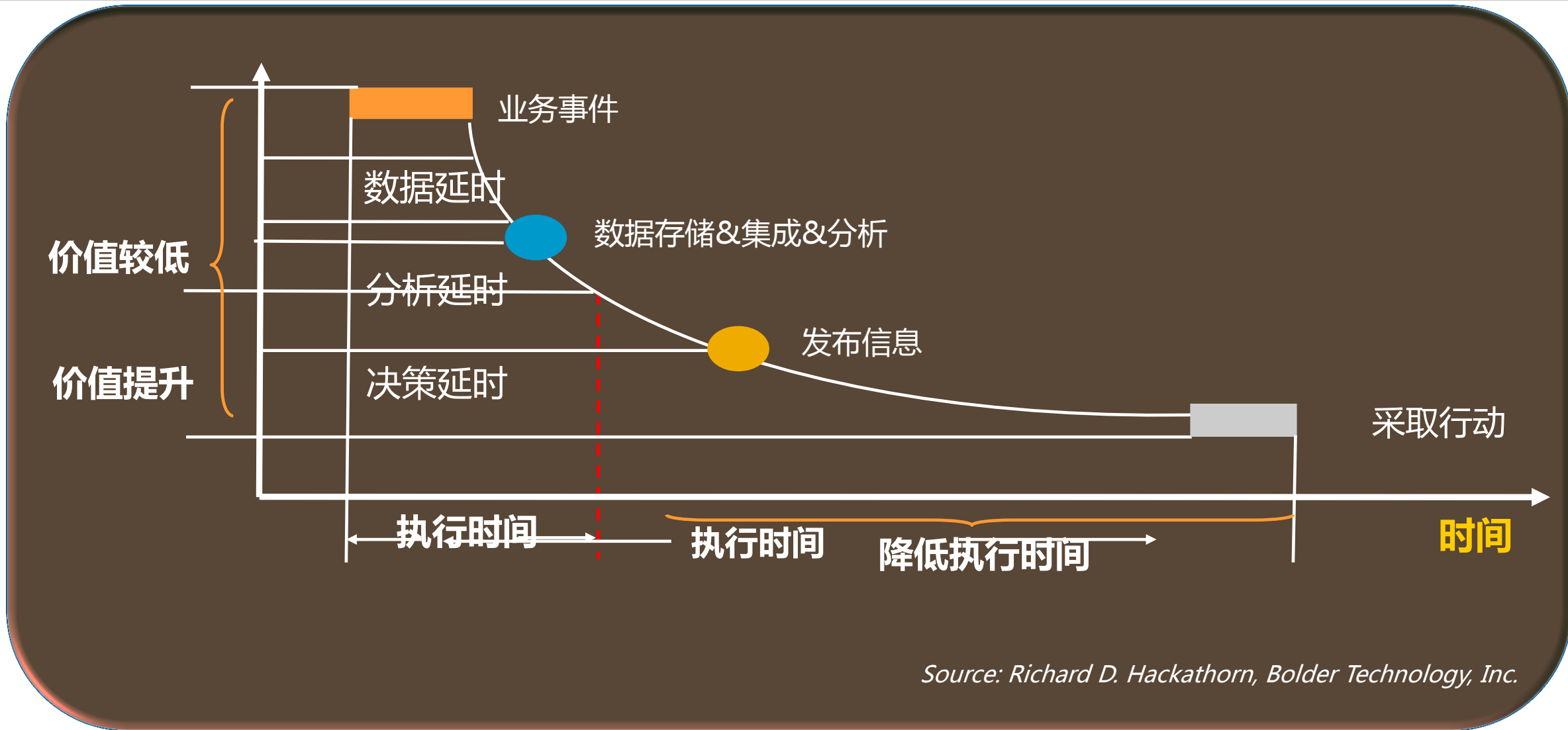


世界杯里的大数据

SAP和德国足协将大数据转变为智能决策，以提高球员的表现现在巴西世界杯。



价值—时间曲线

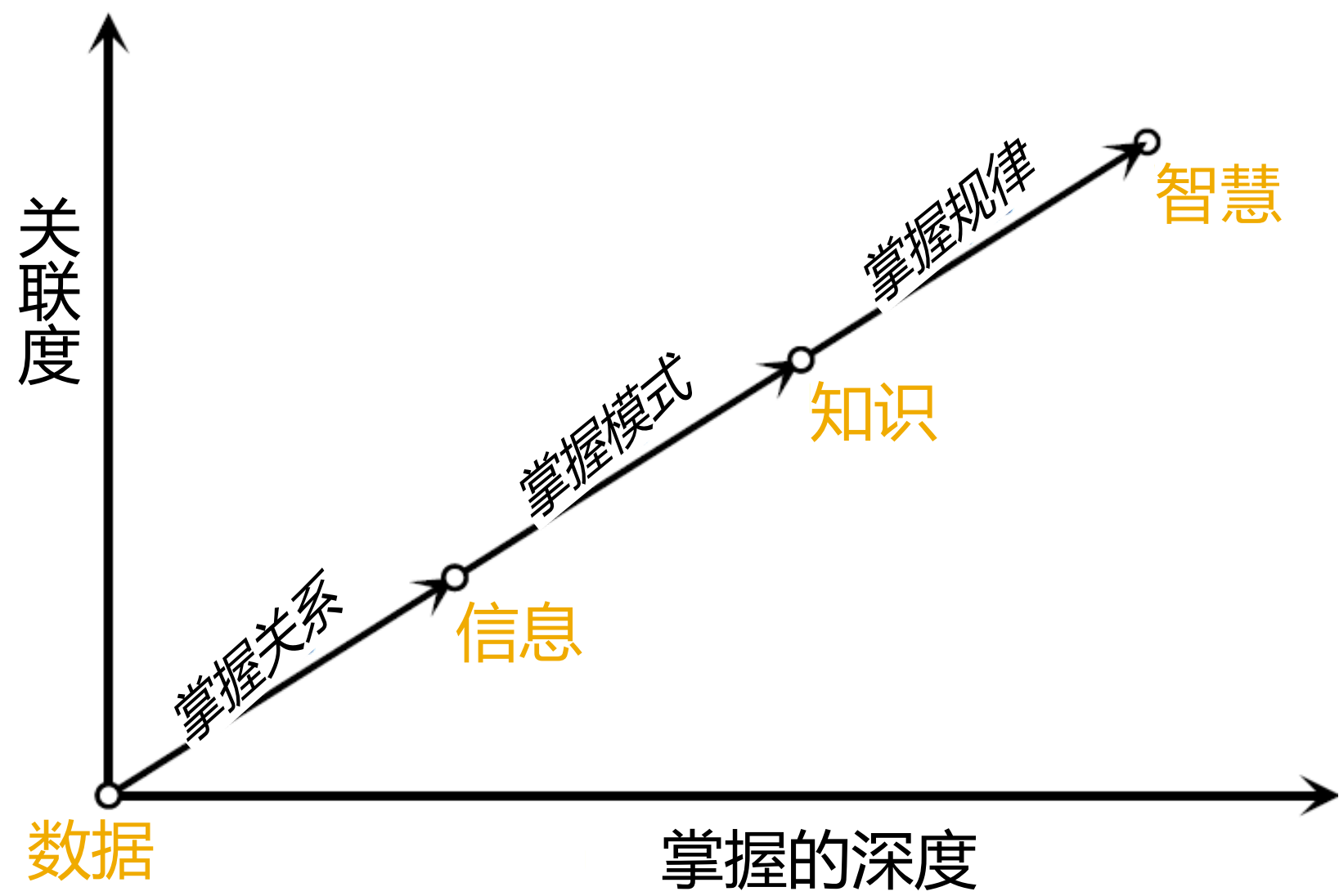


Source: Richard D. Hackathorn, Bolder Technology, Inc.

内容

- ❑ 大数据时代的业务挑战和技术趋势
- ❑ SAP实时大数据平台核心 **HANA** 介绍
- ❑ 成功客户案例分享
- ❑ Intel大数据解决方案

从数据、信息、到 知识和智慧的进阶



大数据在5个方面创造价值

- **先见之明** —— 通过已经发生的、正在发生的事件或实验结果发现或预测需求，洞察变化倾向
- **英明决策** —— 自动算法代替/支持人类的决策
- **一目了然** —— 发现数据之间的关系
- **有的放矢** —— 细分人群，定制行动
- **推陈出新** —— 创新的商业模式、产品和服务

—— 摘自 麦肯锡 《大数据：创新、竞争和提高生产率的下一个新领域》

如果我们可以把新的信号转化为商业价值



品牌感知



预测性维修



网络优化



内部威胁实时发
觉



资产跟踪



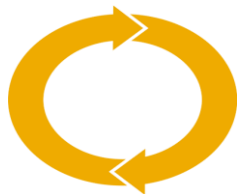
个性化关怀



产品推荐



风险缓释



流失倾向



实时供需预测



360° 客户视图



欺诈防范

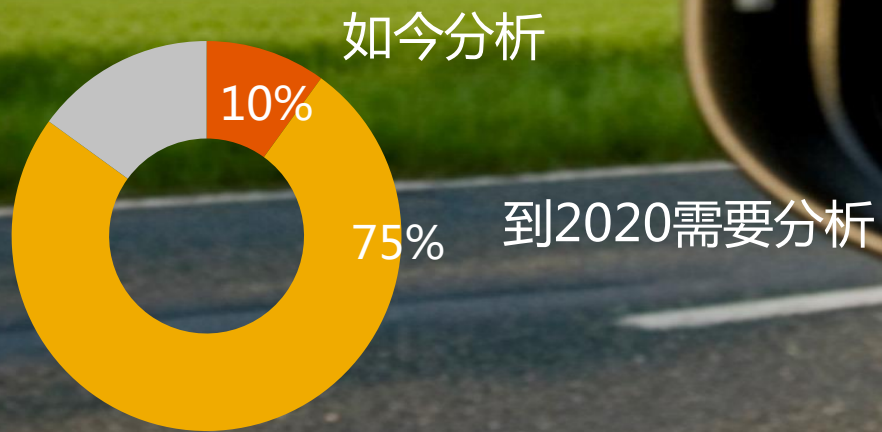
许多企业仍在忽视这些新的信号

\$2.01B

如果位于财富1000强的中游企业能把数据可用性仅提高10%，年营业额就可能增长20.1亿美元。

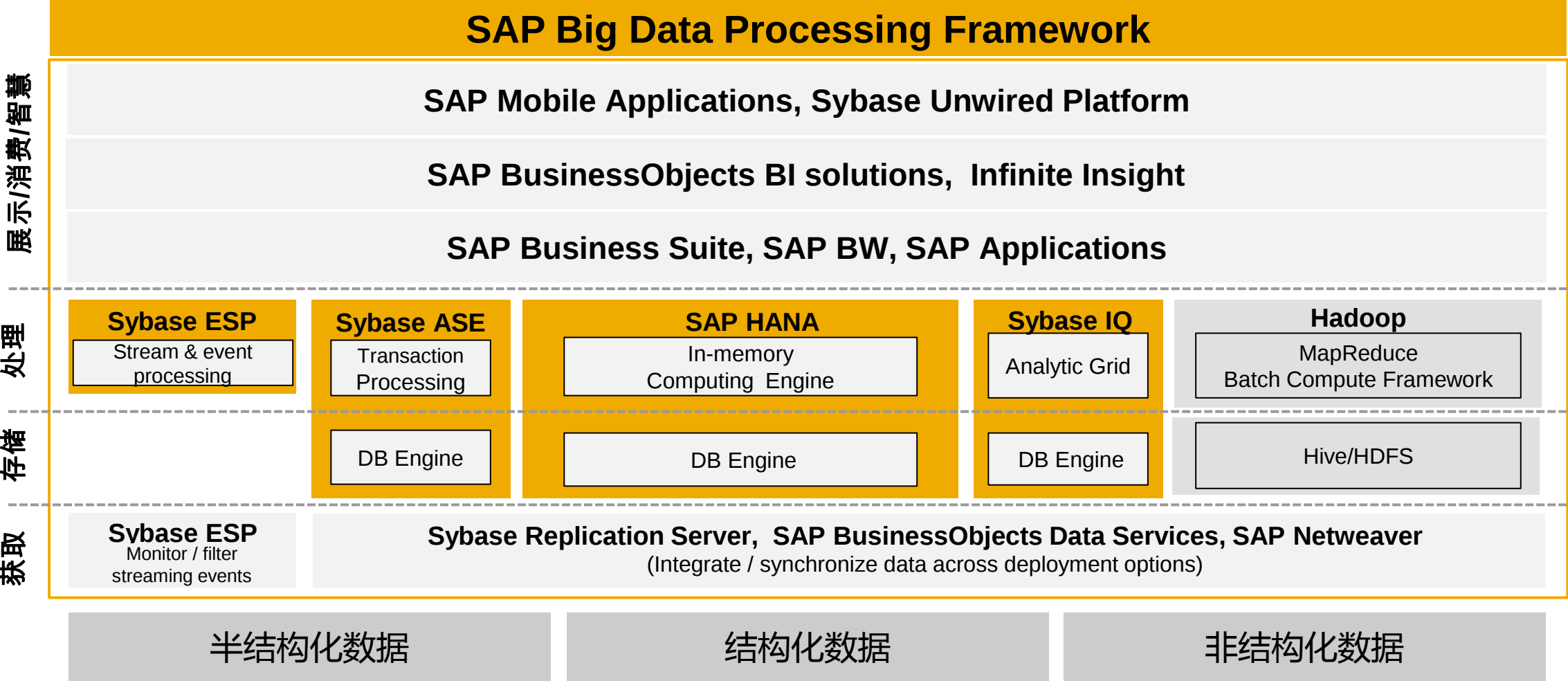
1,000%

每投入\$1到数据分析的ROI会达到\$1000。



最受认可的
KPI

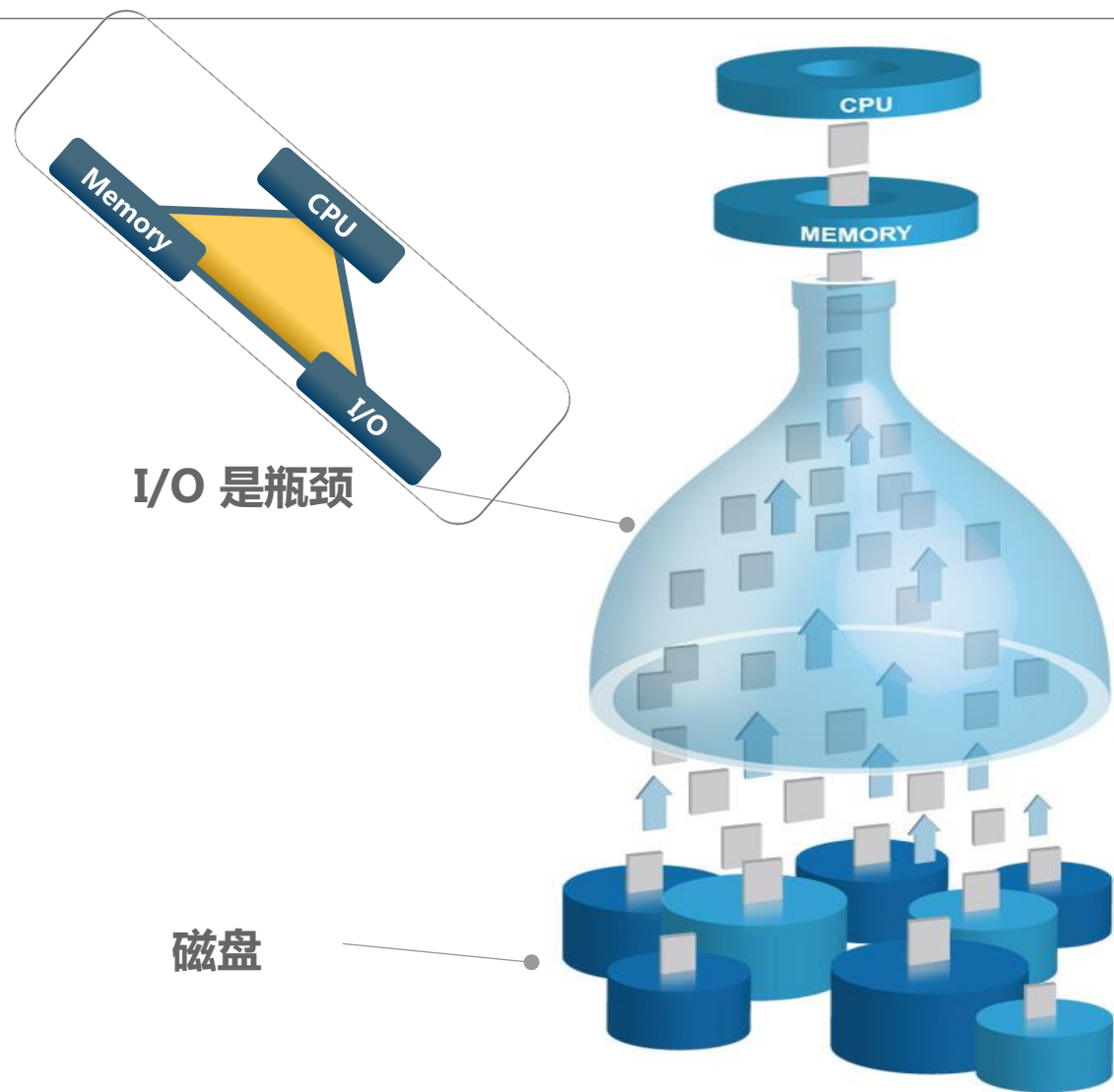
SAP大数据处理框架



内容

- ❑ 大数据时代的业务挑战和技术趋势
- ❑ SAP实时大数据平台核心**HANA** 介绍
- ❑ 成功客户案例分享
- ❑

传统RDBMS不具备实时业务的洞察力，在大数据时代这一问题尤为突出



以磁盘为中心的计算造成巨大的计算瓶颈

I/O 瓶颈导致延迟及CPU的低效利用

关系型数据库为OLTP设计的，而单一的数据存储的方式无法高效满足分析需求

80% 的 OLAP查询使用不到20%的获取的行数据

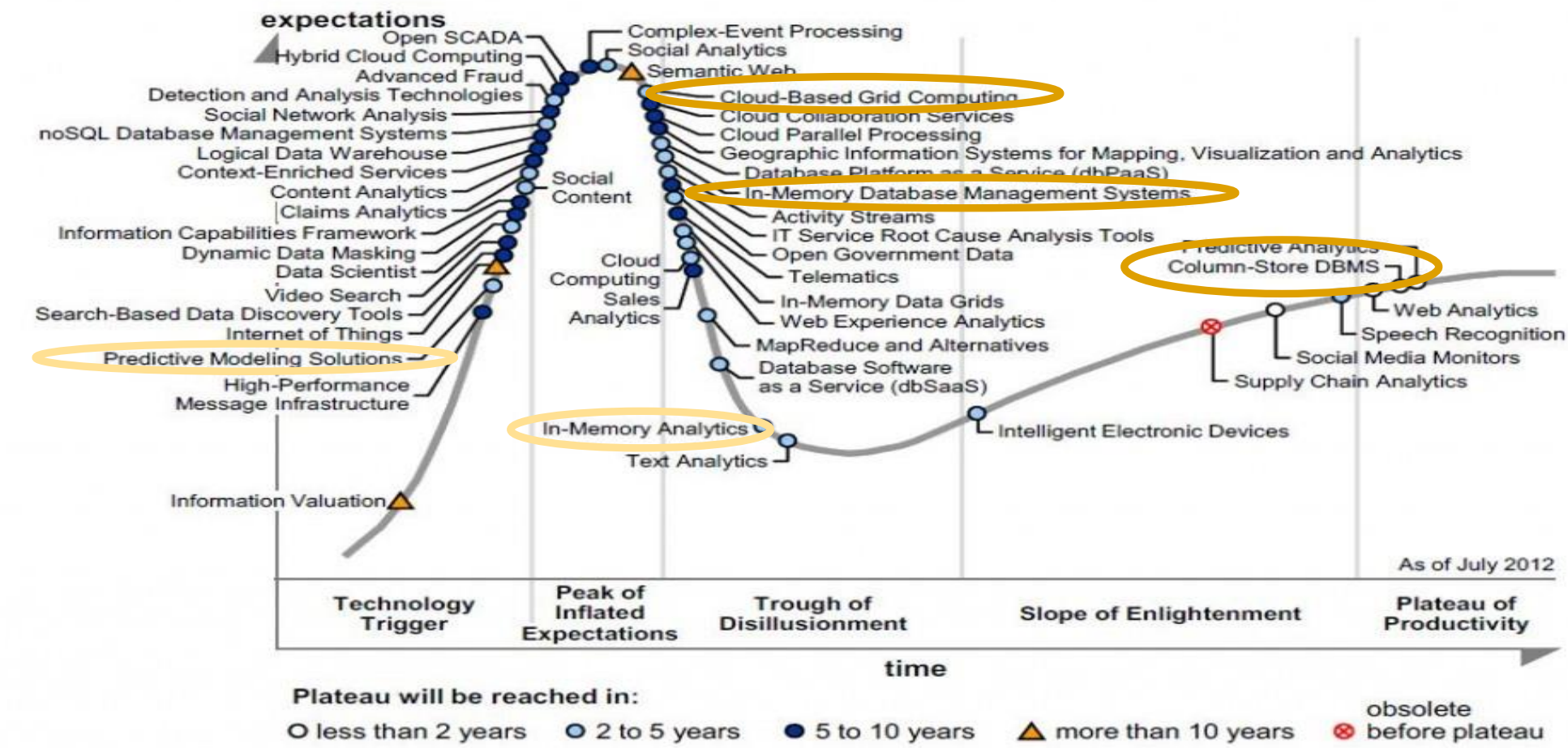
...引起复杂的开发策略

...造成低效的部署方案

...带来不好的用户体验

Gartner2012报告：围绕处理大数据的技术如雨后春笋

Figure 1. Hype Cycle for Big Data, 2012



Gartner 2012年7月的分析报告认为：“列式存储数据库(Column-Store DBMS)、云计算和内存数据库(In-Memory DBMS)在未来的5年内将是3个最重要的技术”。

Source: Gartner (July 2012)

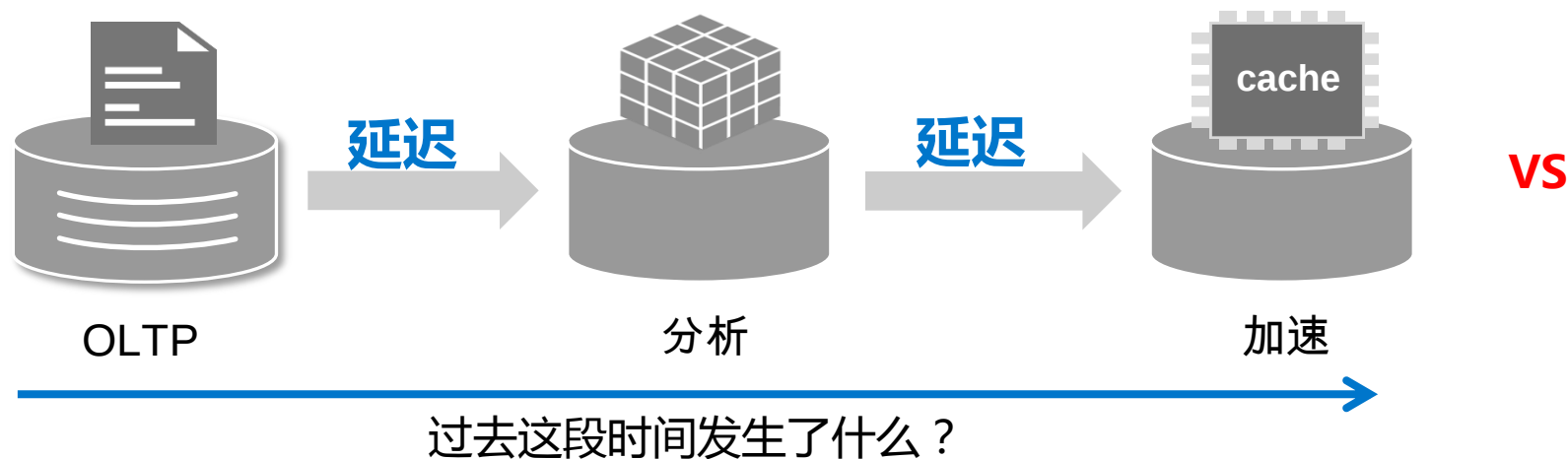
梦想成为现实

立足于创新并以客户为中心，实现业务转型



使用列式和行式混合存储的内存计算技术，实现 OLTP 和 OLAP 的通用数据库平台

~ 哈索·普拉特纳 (Hasso Plattner)

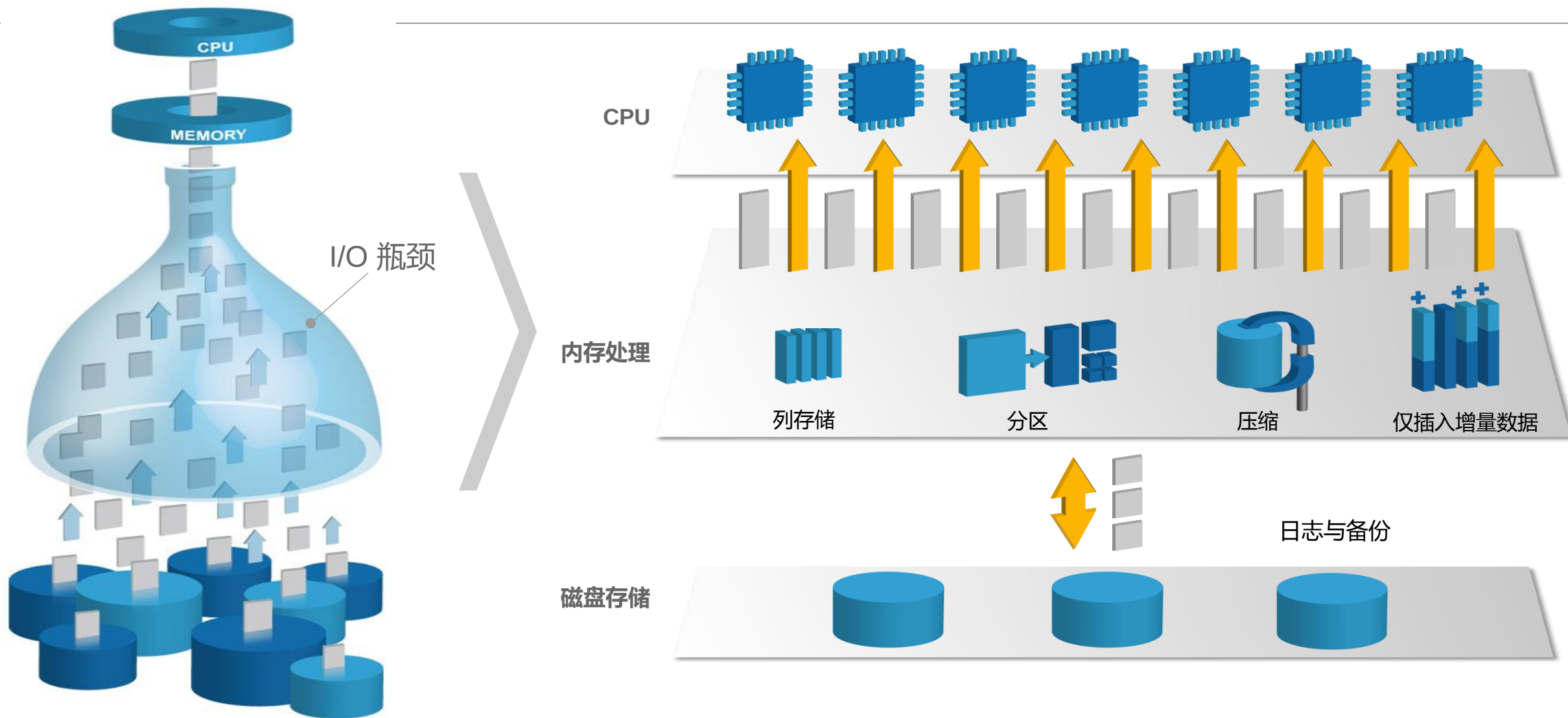


VS



打破传统IT发展的瓶颈

利用内存技术，降低延迟时间



SAP HANA实时数据计算平台

- 软件上的变化

实时

- 业务无延迟，实时计算
- 海量数据，瞬间得到结果

迅捷

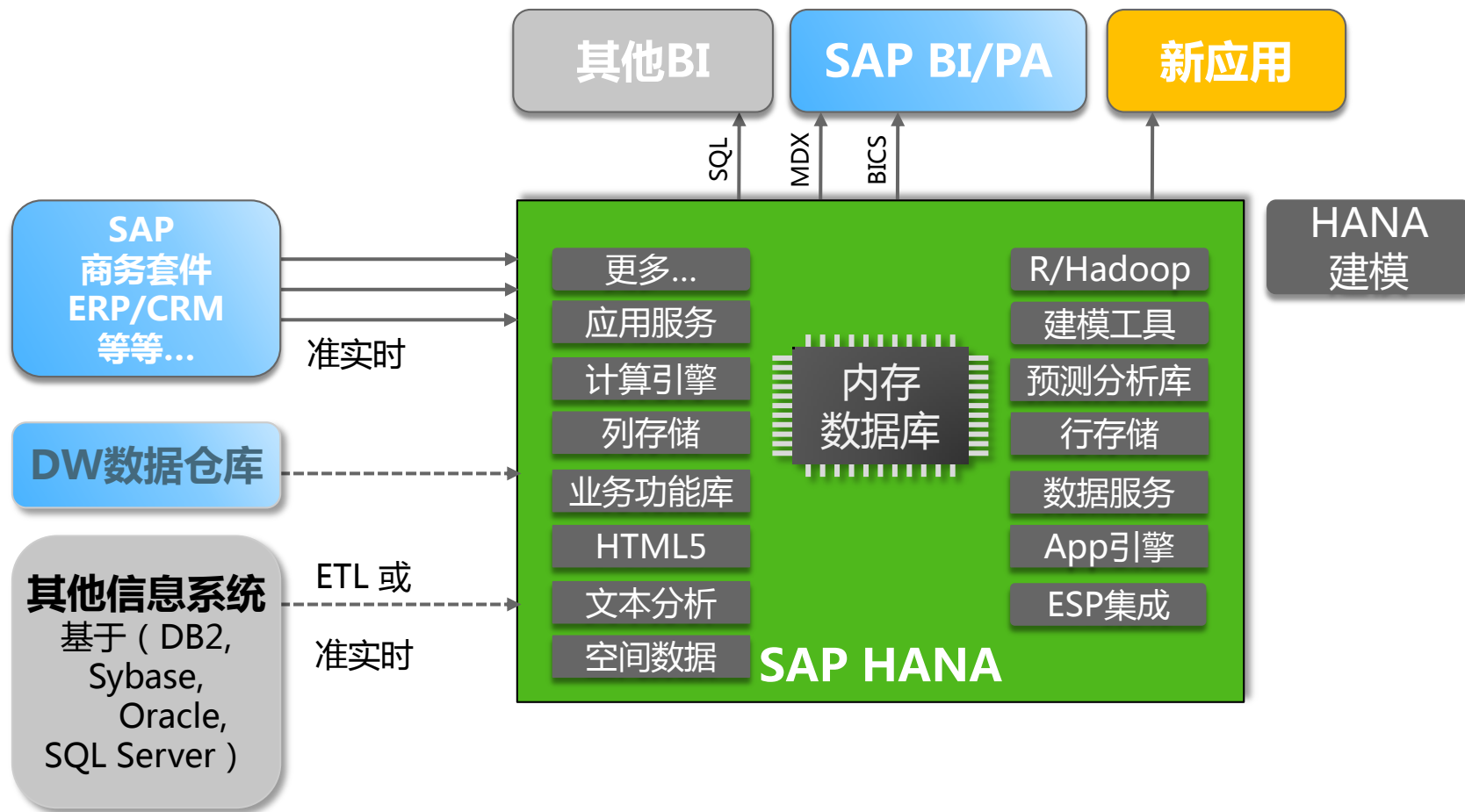
- 多处理器，海量内存
- 实时的计算引擎
- 支持OLTP+OLAP混合负载

开放式硬件架构

- x86架构的一体机
- 内置BAE (业务分析引擎), BFL、PAL库，集成R。

开放式软件架构

- SQL/ODBC/JDBC/ODBO..
- Python/OData/WebService



SAP HANA最简单的理解



商务智能软件-BI

所有的I/O，计算都在内存

数据

写入

内存（数据）

来自任何DB
来自任何系统

同步写入

闪存-Flash（日志）

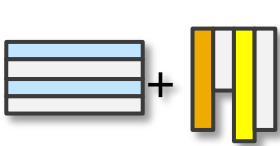
重启之后

磁盘

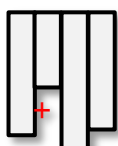
异步



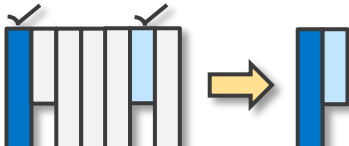
SAP HANA 技术特性



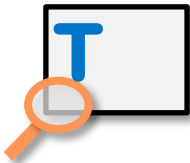
Column and row store



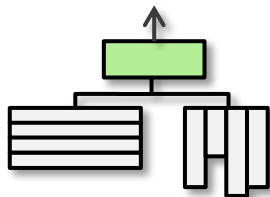
Insert only on change



Minimal projections



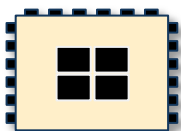
Text Retrieval & Exploration



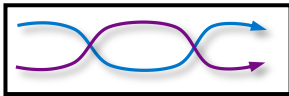
No aggregates



Analytics on historical data



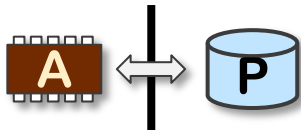
Multi-core/parallelization



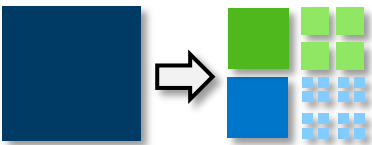
Multi-threading within nodes



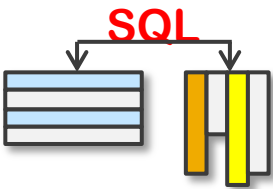
In-memory Compression



Active/passive & data aging



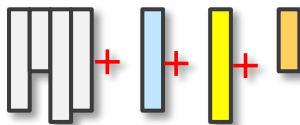
Partitioning



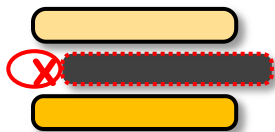
SQL interface on columns & rows



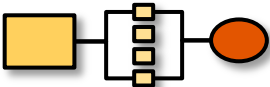
Single and multi-tenancy



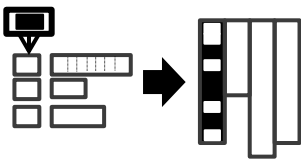
Dynamic Extensibility



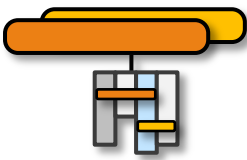
Reduction of tiers / layers



Map reduce



Group Key



In-memory Apps



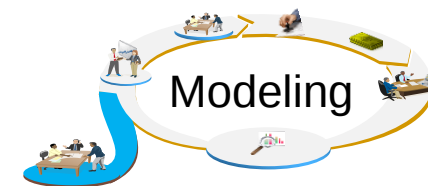
Bulk load



Geo-data

...

预测分析/数据挖掘最常用的业务场景



■ 分类

哪些客户在（下周/下月/明年）最有可能发生（流失/欺诈/购买）行为？

■ 预测

一位客户在（下周/下月/明年）可能购买多少个产品？

明年的每个月会产生多少（收入/客户流失）

■ 客户细分

哪些客户具有相似的（行为/偏好/...）？

■ 社交网络分析

分析交互联系以识别（社区/影响者/介绍人/...）

■ 相关性规则

通过分析交易历史识别最可能同时发生的事件



客户购买预测

模拟与预测

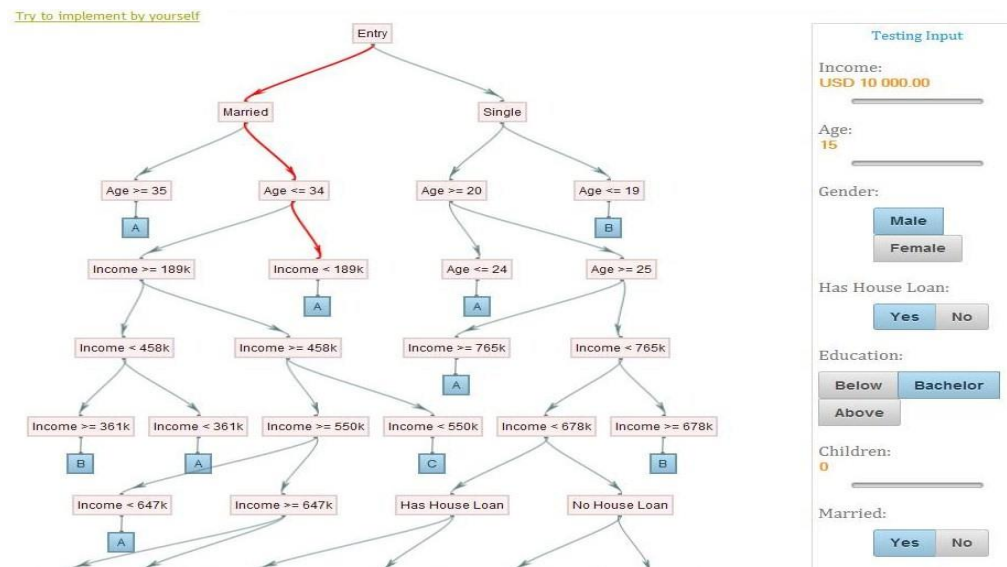
概要：如何识别 VIP 客户以及如何确定影响其购买决定的重要因素。

功能

- 历史销量和销量预测
 - 历史销量报告
 - 销量预测报告
- 识别 VIP 和预测潜在购买者偏好
 - 识别最有价值的老客户 (ABC)
 - 预测潜在购买者的偏好模型 (KNN)
- 分析决定因素和推荐交叉销售
 - 使用假设分析来分析影响购买决定的最重要因素
 - 购买此产品的人也购买了其他产品

主要益处

- 预测和期望 VIP 客户可能会购买哪种型号的商品
- 提供量身打造个性化服务来吸引 VIP 客户



SAP HANA 预测平台

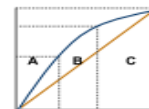
预测分析算法库 (PAL)

信息挖掘算法：

K-means聚类



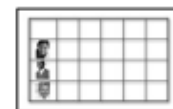
ABC 分类



关联分析(购物篮分析)



加权评分表

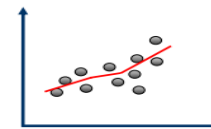


预测算法：

基因算法：规划与优化



回归分析

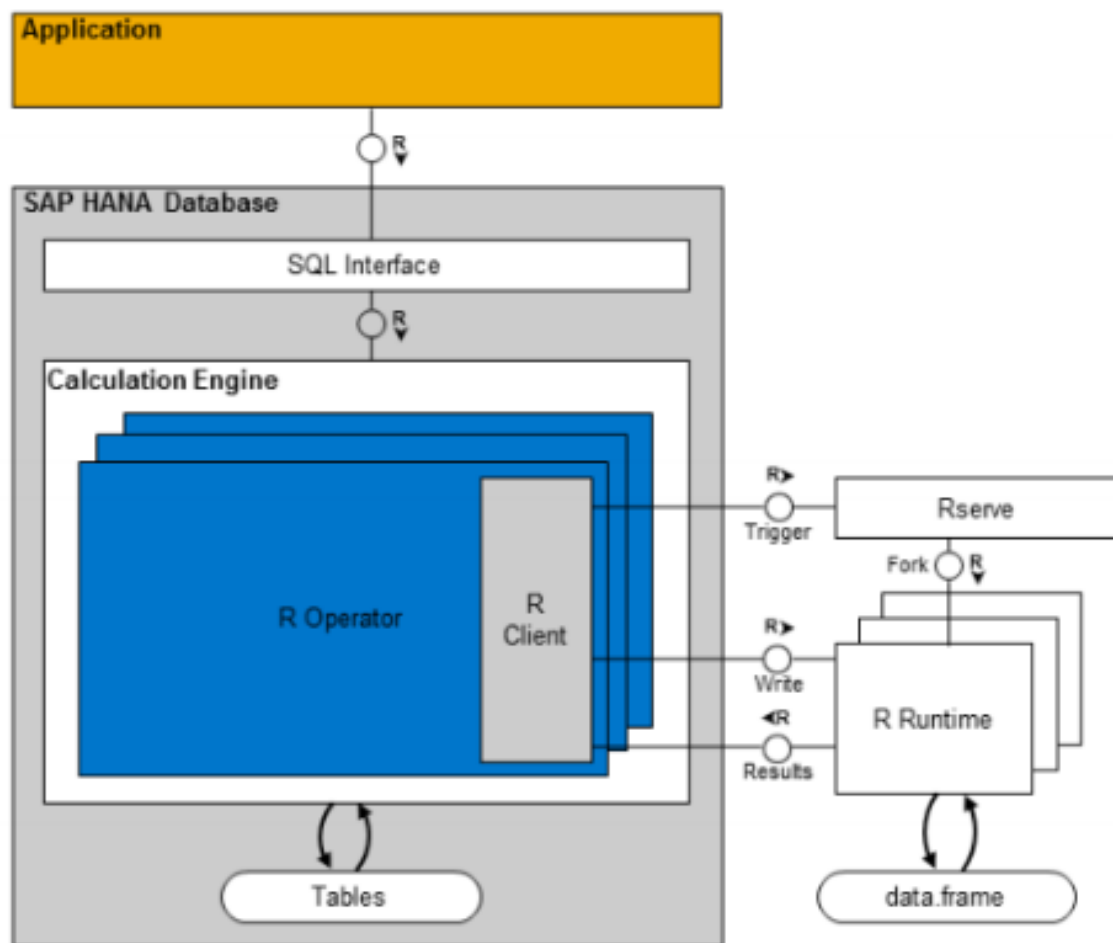


C4.5 决策树



SAP HANA 预测平台

与R(最流行的数据挖掘语言)集成



- 当HANA计算引擎识别出R操作符，便由R Client发出一个请求到Rserve，并发送相关的参数，触发执行
- 执行完毕之后，结果数据框被发送至计算引擎。

InfiniteInsight(KXEN) 建模模块功能

回归 / 分类

- 定义：理解预测客户行为（连续&二值分类目标）
- 应用：目标市场，欺诈检测，信用评分，流失

分群/ 聚簇

- 定义：聚类相似群体（监督/无监督学习）
- 应用：市场分析，渠道、信息管理

时间序列 / 预测

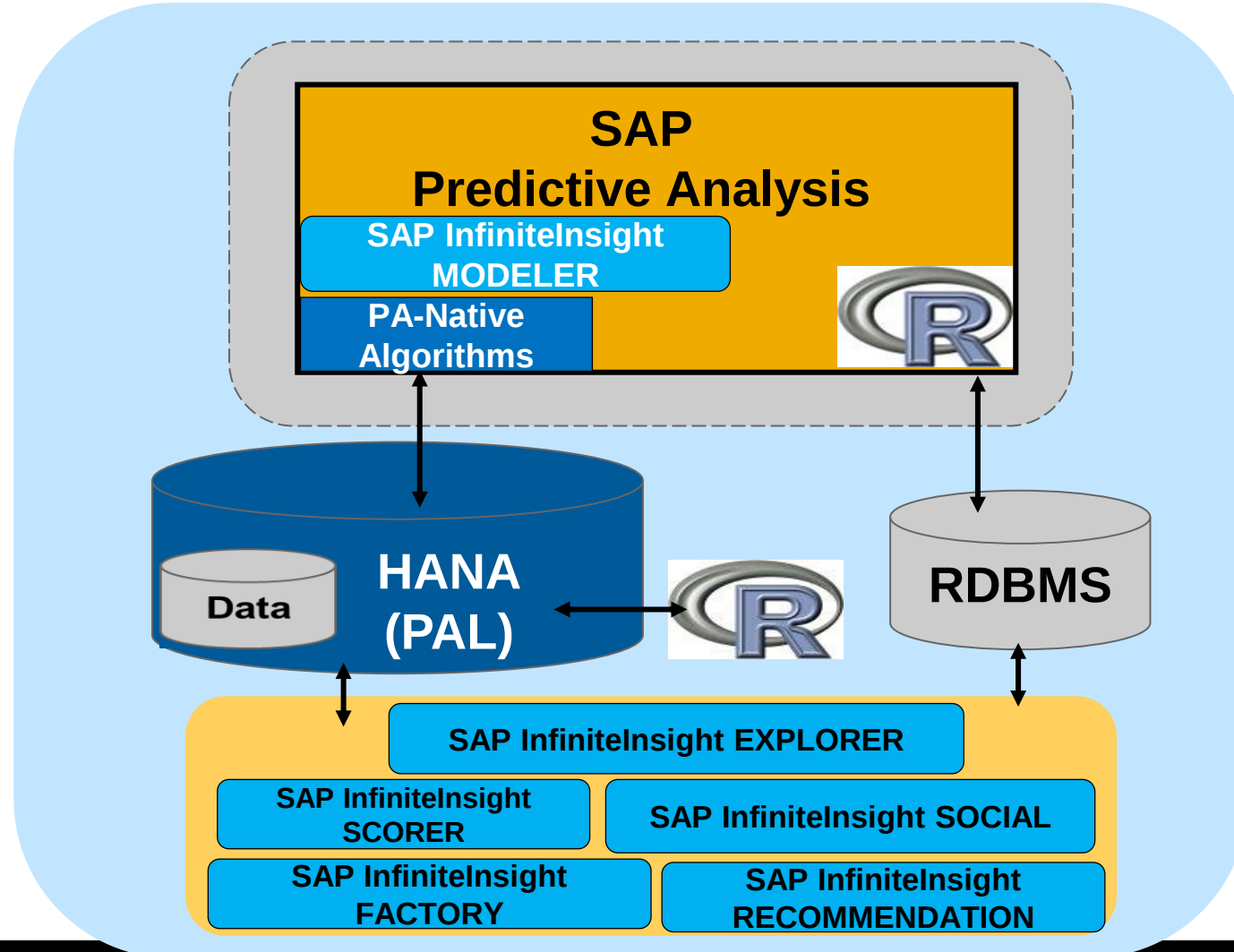
- 定义：分析预测与时间有关数据（趋势，周期等）
- 应用：证券，利率

关联规则/ 市场篮分析

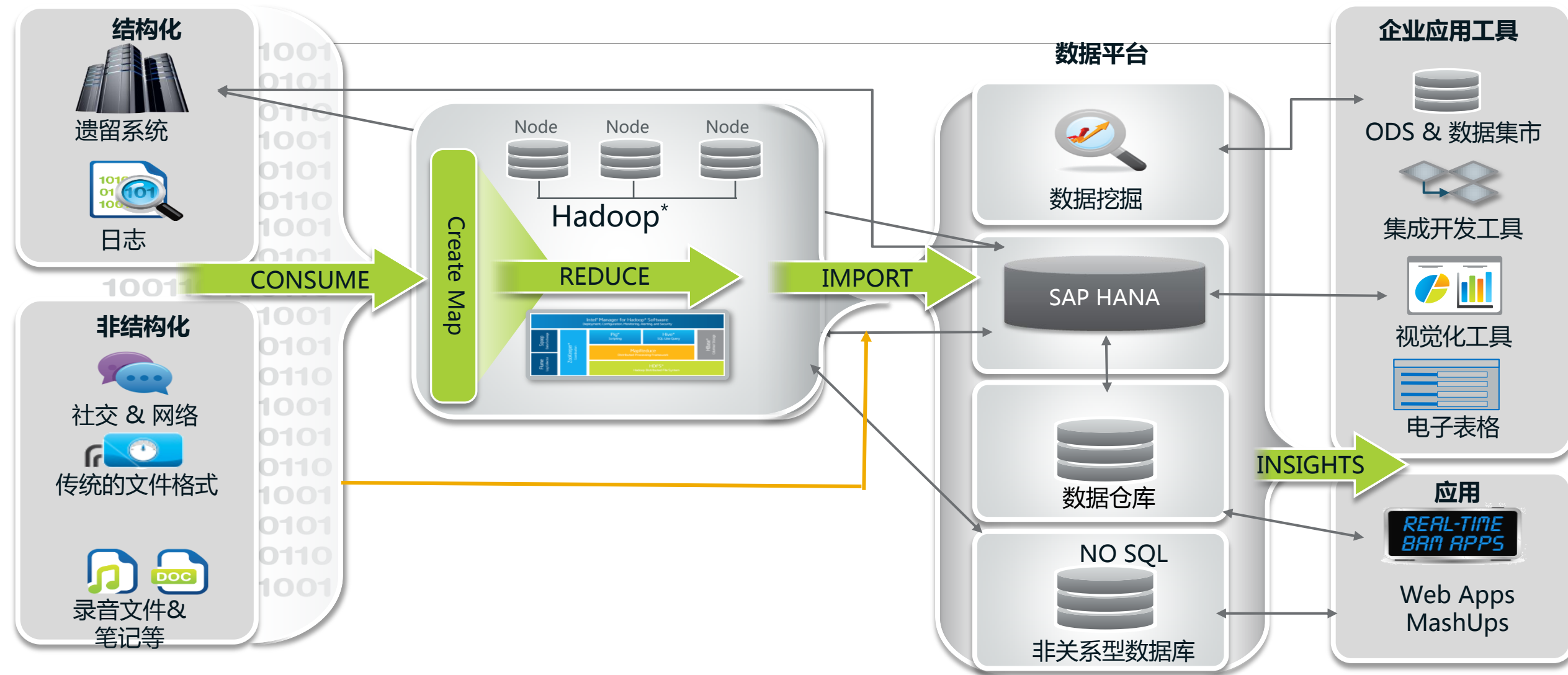
- 定义：发现产品和服务内在关联
- 应用：零售业，电信业

KXEN
一致编码

KXEN + HANA



SAP HANA与Hadoop集成



利用SAP HANA实现实时预测分析驱动更新的商业活力，变现大数据价值

世界在飞快地变化着。这已经不再是以大胜小的时代。而是以快胜慢的时代。

—— Rupert Murdoch

内容

- ❑ 大数据时代的业务挑战和技术趋势
- ❑ SAP实时大数据平台核心 **HANA** 介绍
- ❑ 成功客户案例分享
- ❑ Intel大数据解决方案

实时大数据在行业里的应用

媒体行业应用案例-分析并预测网游用户行为



5,000 个事件在每秒
钟载入 SAP HANA
(以前无法实现)



10-30% 年收入增长



交互性 通过数据分
析改善设计思路以及游
戏策划

业务挑战

- 提高从免费 → 付费玩家的转变率
- 提高每个付费玩家的平均收益
- 减少玩家流失 – 让付费玩家长期参与

技术挑战

- 利用 SAP HANA 中的实时数据处理以及带有针对 SAP HANA 集成的 R 的分类算法，为玩家提供个性化的与背景相关的建议
- 分析超大量历史和交易数据，预测玩家行为模式

优势

- 实时洞察力
- 通过对每个玩家的利润率分析，进一步了解玩家行为



在 Bigpoint 的 Battlestar Galactica Online 游戏中，我们每秒的游戏中事件有 5,000 多个，我们必须将这些事件载入 SAP HANA 环境并进行处理，为玩家创造出个性化的游戏环境。在这个与 SAP HANA 协作展开的共同创新计划中，借助实时建议管理，我们收入至少人提升 10-13%。

Claus Wagner, Bigpoint SAP 技术高级副总裁

实时大数据在行业里的应用

快速消费行业应用案例-物流与运输优化



解决方案概述:

我们关注了许多中国客户，他们在交通运输的开销占到运营成本的80%。如何降低流通成本，提高产品流通效率正迫在眉睫。客户迫切需要为扩张和增加产品销售而建立新工厂和进入新地区的指导，并以高效的产品流通来降低成本、扩大市场份额。未来的扩展策略离不开专业、可靠的物流计划系统。

客户

- 农夫山泉: 中国领先的饮用水生产企业。拥有10个生产基地和20个配送中心，服务于5768个零售代理，1070779个零售商店，和1530家超市。
- SAP 在物流和运输解决方案上和农夫山泉共同创新合作。
- 这个解决方案适应于所有拥有运输与物流的公司。

每年的运费占到了农夫山泉15%的销售额，使用了SAP HANA的**供应网络模拟和配送计划**优化的功能，使农夫能比以前**降低35%**的运费，并从原来的业务需要**1天**的计算，到现在只需用**3.7秒**。



实时大数据在行业里的应用

零售行业应用案例-自动建议订单



客户:

巴西食品公司-是世界最大的食品商之一,出口到140多个国家, 出口的动物蛋白占全球的 9%, 年收入130亿 USD。61个工厂, 119,000个员工。

业务挑战:

- 需要知道, 哪些产品需要捆绑销售而增加销量。
- 以往只是用Excel表对短期的历史数据、在没有科学指导原则下分析和决策。
- 没有一个科学的统计方法, 以支持商业智能。

HANA价值:

- 事实证明HANA对数十TB的数据进行数据挖掘易如反掌
- 引入科学原则到获得用户分析、简化架构、易用的移动接入等创新能力
- 实现为客户建议订单, 从而为BRF增加了大量收入。



业务用例-自动建议销售订单

Data Size:

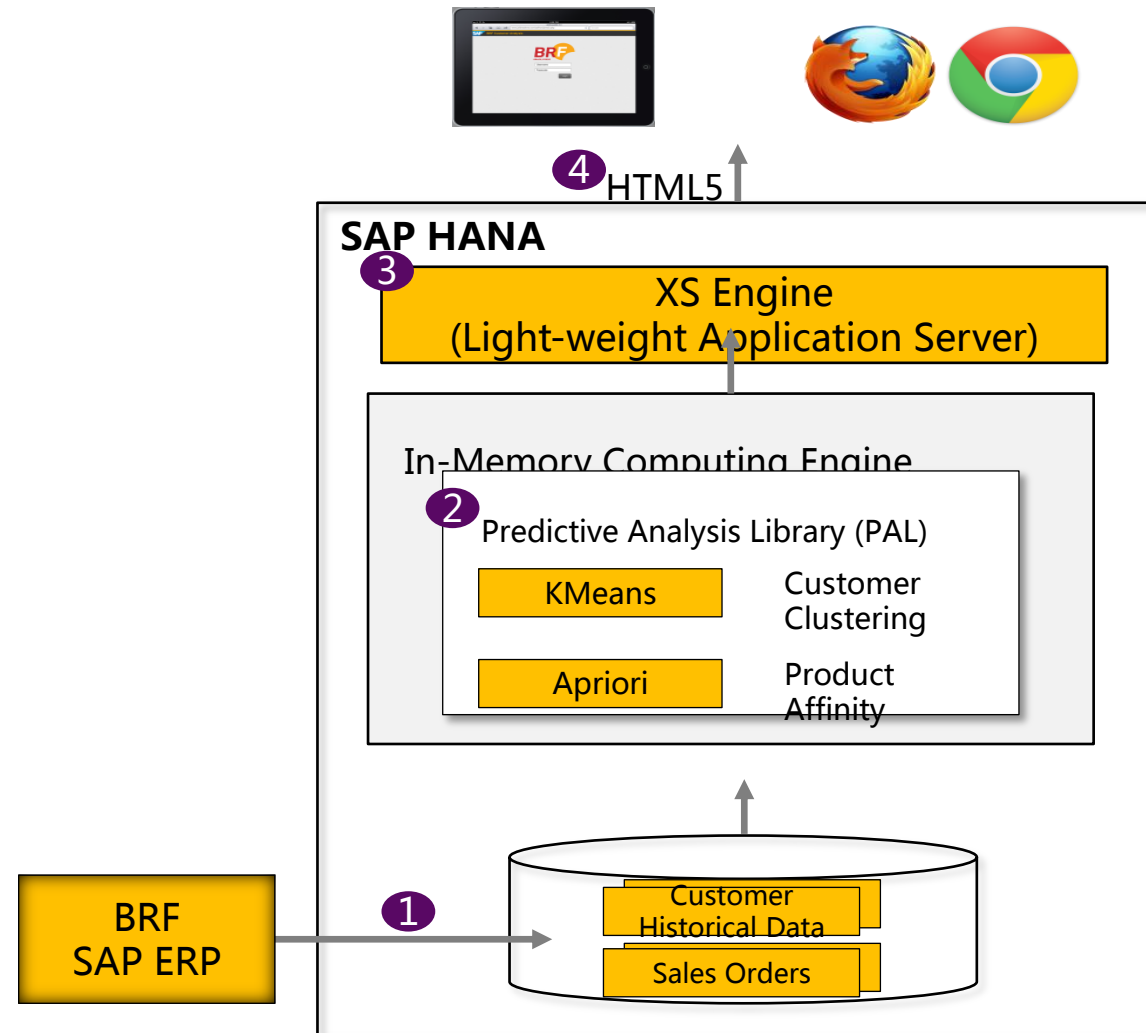
- 装载 最近3个月的销售数据 到 HANA
- 销售订单分析数据模型和客户历史数据 (3个月)

使用的科学原则(Scientific Principles) :

- 用预测分析库(PAL: Predictive Analysis Library)中的 K-means 进行客户聚类
- 用PAL中的Apriori 算法进行产品相关性分析
- HANA + R 用于销售预测

场景举例:

- 用聚类算法按Top销量产品类别将门店分为不同组
- 在某类门店中的个别门店，BRF可能发现某些产品类别的销售情况低于平均
- 用聚类关联规则为这个门店建议销售订单



实时大数据在行业里的应用

电商行业应用案例-初期信号探测



公司名称

eBay

总部

San Jose, California

行业

互联网服务

产品与服务

Online Marketplace

员工

31,500 (2012)

营收

US\$ 141亿 (2012)

Web Site

eBay.com



业务挑战

- 提高从噪音当中识别信号、并确定有利于于eBay的市场变化方向
- 提高eBay预见和预测能力
- 获得明察秋毫和发现原因的洞察力

技术挑战

- 从eBay 100PBs的EDW中发现关键的信号
- 高度手动过程，因为一个模型并不适合所有的指标，因此需要分析师干预

核心利益

- 基于SAP HANA预测分析的自动信号探测系统会自动选择最好的度量模型，提高预测的准确性
- 由可靠和可扩展的系统提供实时的洞察，让数据分析师，专注于战略任务
- 决策树的逻辑和可按场景调整灵活性允许eBay选择最佳模式以适应他们的数据的



**判断的准确率达
100%**

当信号是真阳性的，在97%的置信水平下



初期信号自动探测

系统由SAP HANA驱动

“HANA有非常有价值的，它加快了洞察的速度。HANA充分利用了内存、多核、快速、大量数据等要素，所有这些混合在一起，我认为这才会是分析的广阔未来。”

David Schwarzbach, VP&CFO eBay North America at eBay Inc.

“HANA系统可以即刻释放所有的潜力，洞察现在正在发生什么。用户只需提供他们的度量，不必担心哪个算法是最好的，就能够使用该系统，因为它天生就是智能的和可配置的。”

Gagandeep Bawa, Manager, North America FP&A at eBay Inc.

3200+ HANA 客户 (25个进入万倍俱乐部, 5个进入10万倍俱乐部)



HANA硬件合作伙伴



SAP HANA 硬件合作伙伴



HITACHI



lenovo

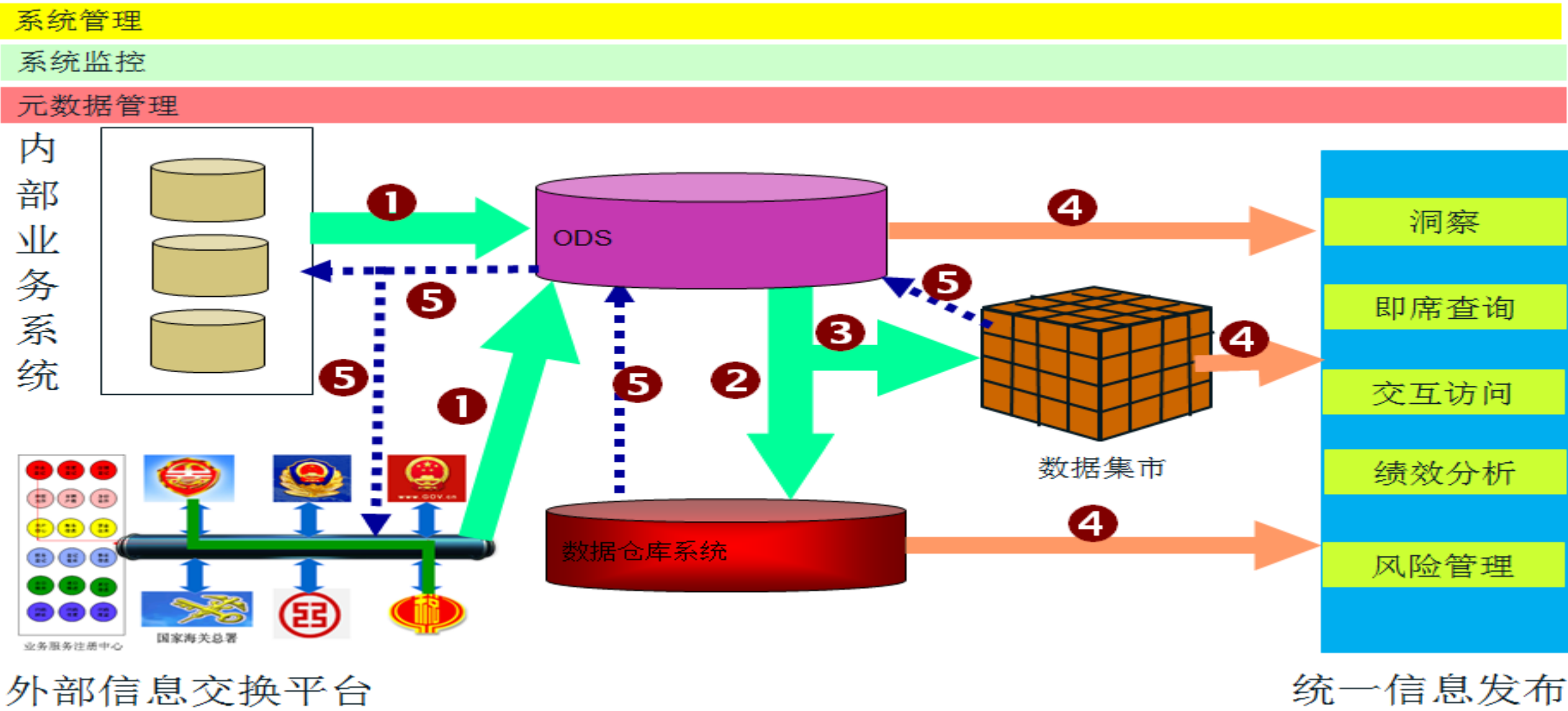
NEC



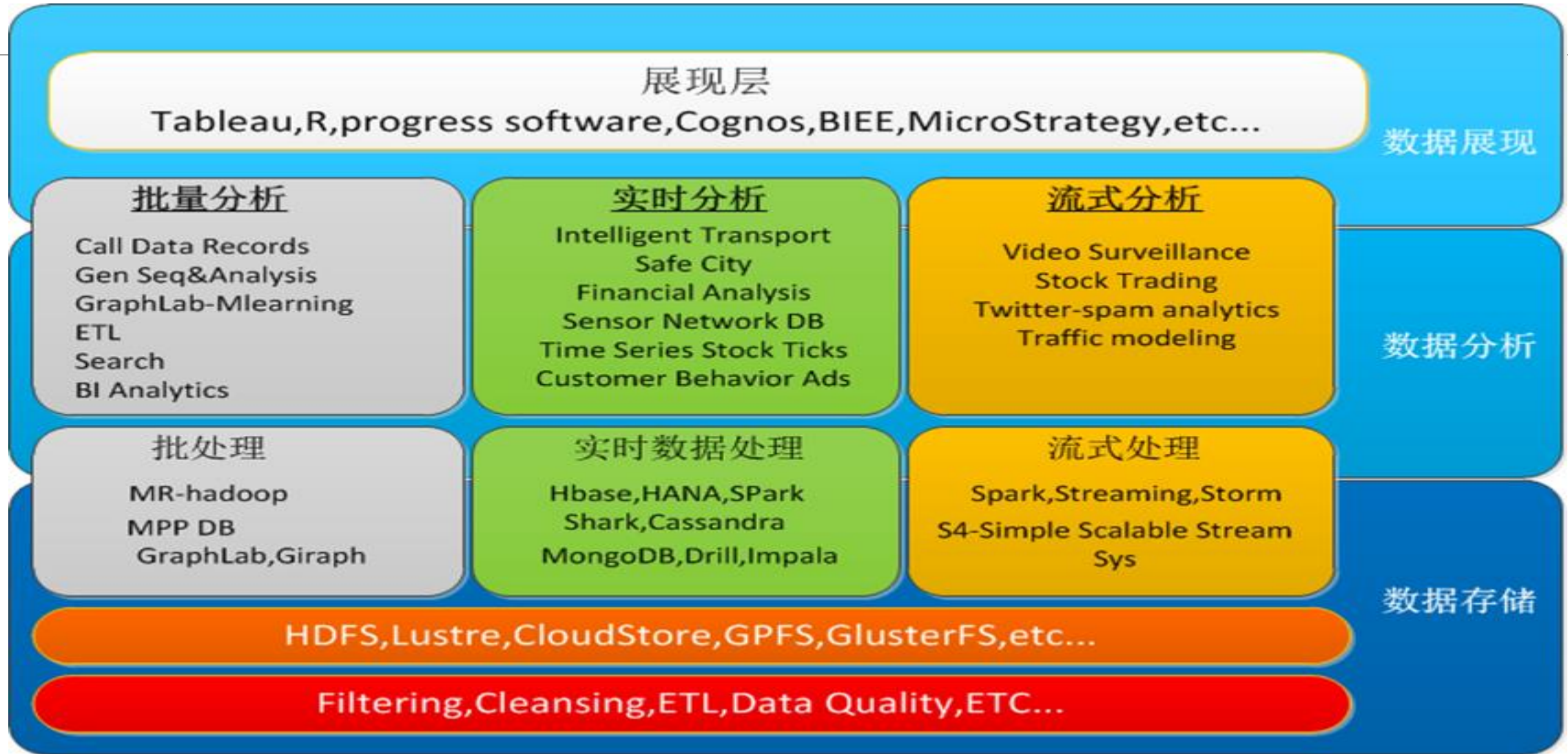
内容

- ❑ 大数据时代的业务挑战和技术趋势
- ❑ SAP实时大数据平台核心 **HANA** 介绍
- ❑ 成功客户案例分享
- ❑ **Intel大数据解决方案**

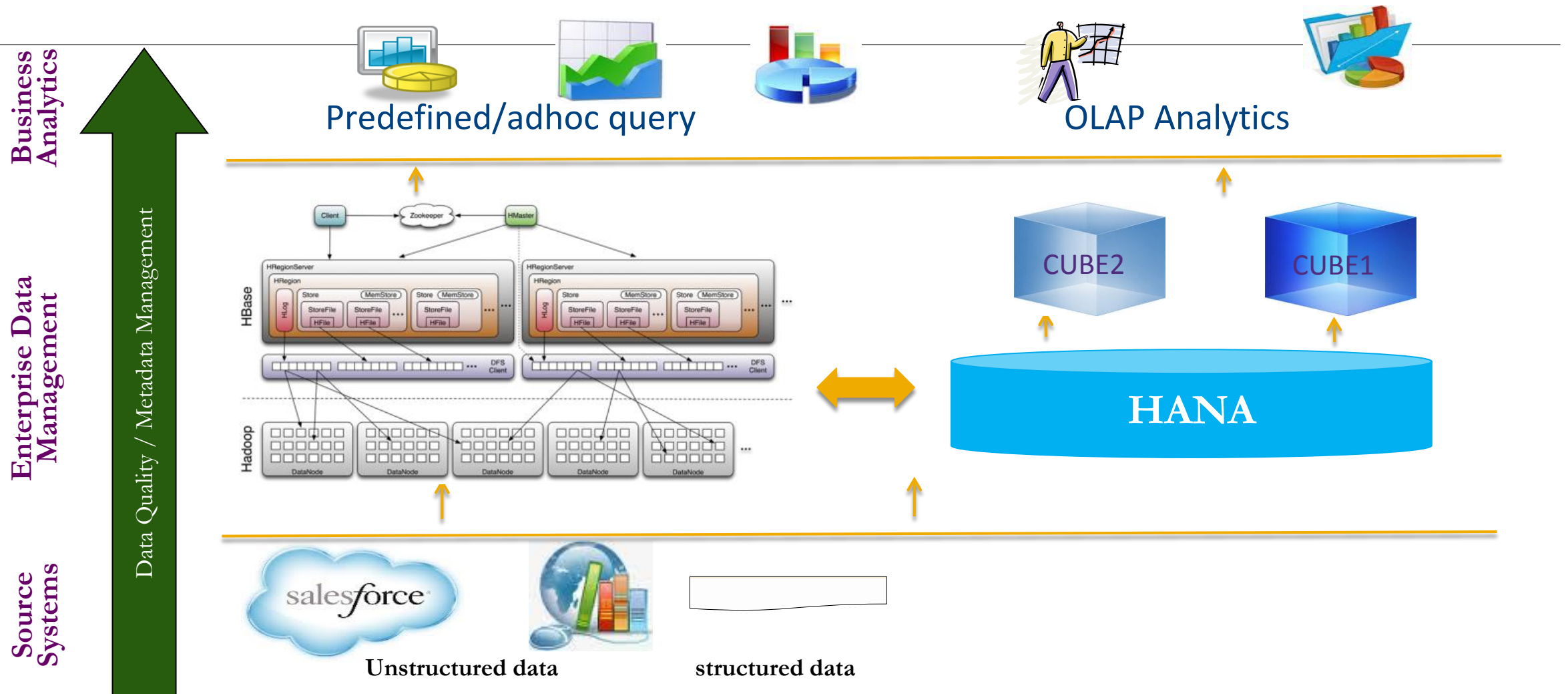
传统决策支持系统架构



大数据如何处理



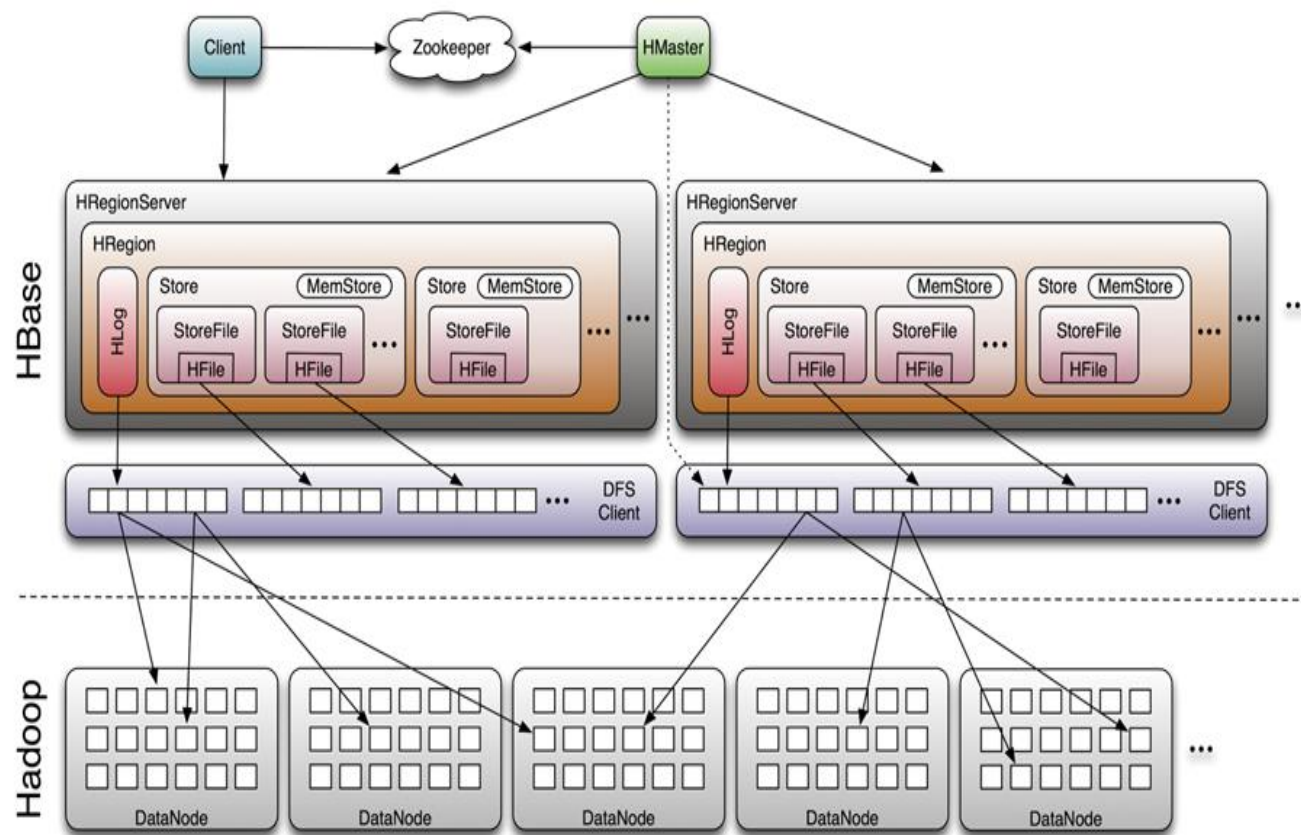
Solution1:HADOOP+HANA End2End Big Data



End2End IA based big data solution

结构化详单数据存储

- 分布式文件系统(HDFS)
 - 群集内完整的名字空间
 - 3个备份：数据自冗余
- 基于KV的NoSQL数据库HBase
 - 基于HDFS存储的查询系统
 - 系统的横向扩展能力好
 - 适合于单键查询系统



HDFS - 分布式文件系统

设计目标

错误检测和快速自动恢复

- 硬件故障是常态而非异常

为流式数据访问优化

支持大数据集

- 单个文件大小有数GB或者TB
- 可提供高聚合带宽访问
- 可能够扩展至数千个节点

简化“一致性”模型

- 一次写入、多次读

移动“计算”比移动“数据”更便宜

主要特点

使用低成本存储和服务服务器构建

存放PB级别的海量数据

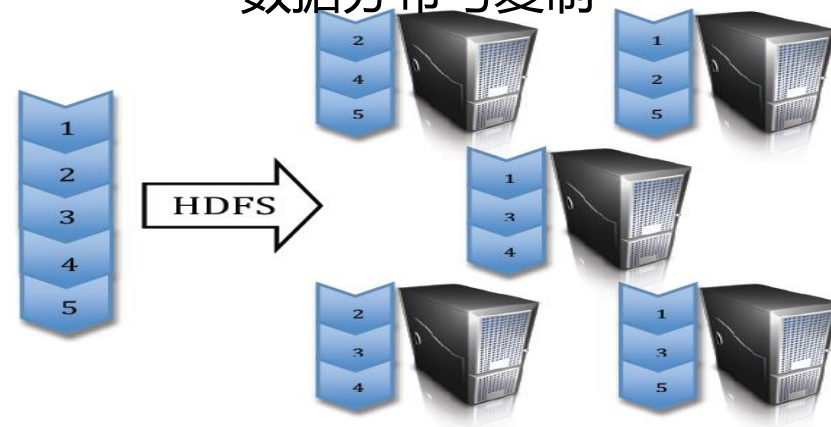
高可扩展性，实际生产环境扩展至4000个节点

高可靠性和容错性，数据自动复制，可自我修复

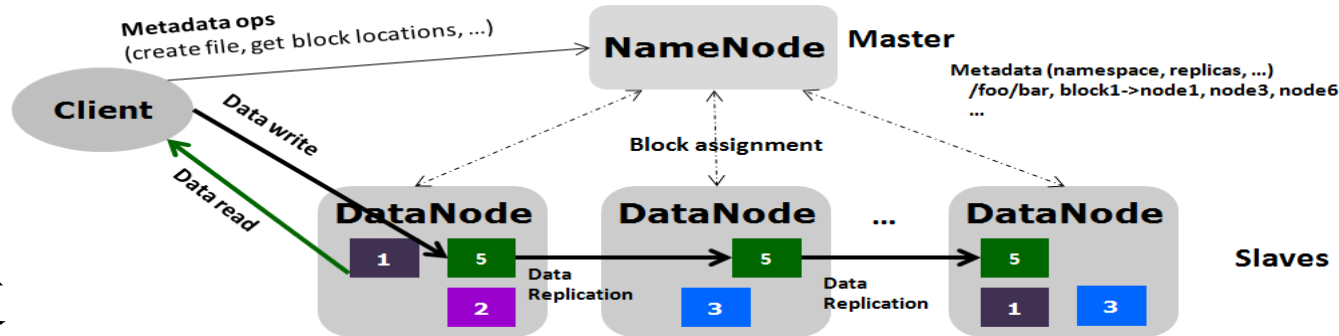
高带宽，高并发访问

高级均衡算法，提高系统扩展性，适合不同配置服务器组成的集群（只在intel版本中）

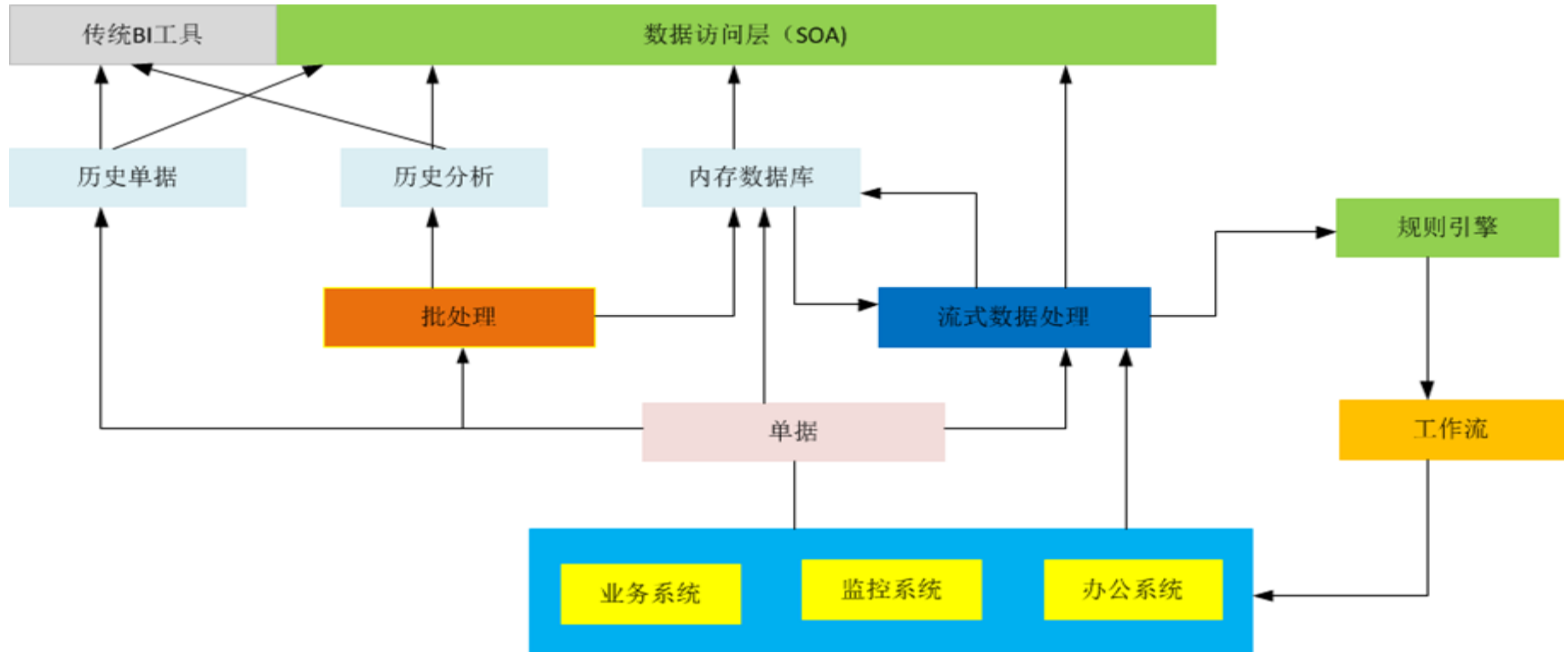
数据分布与复制



HDFS架构和读写操作

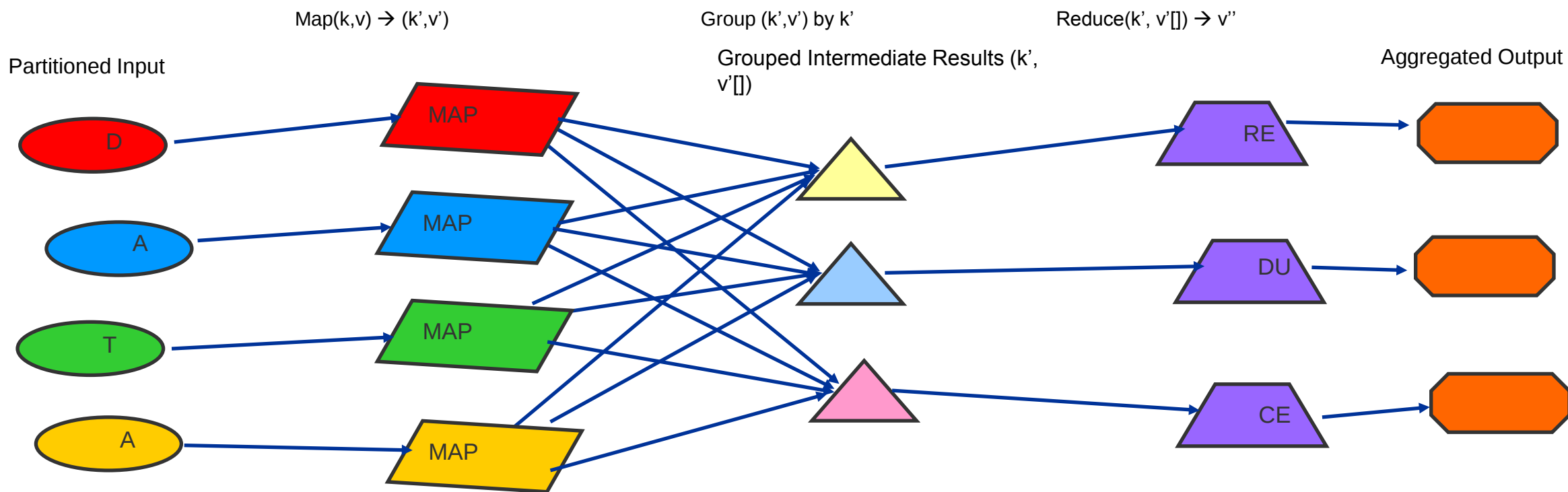


大数据处理方式

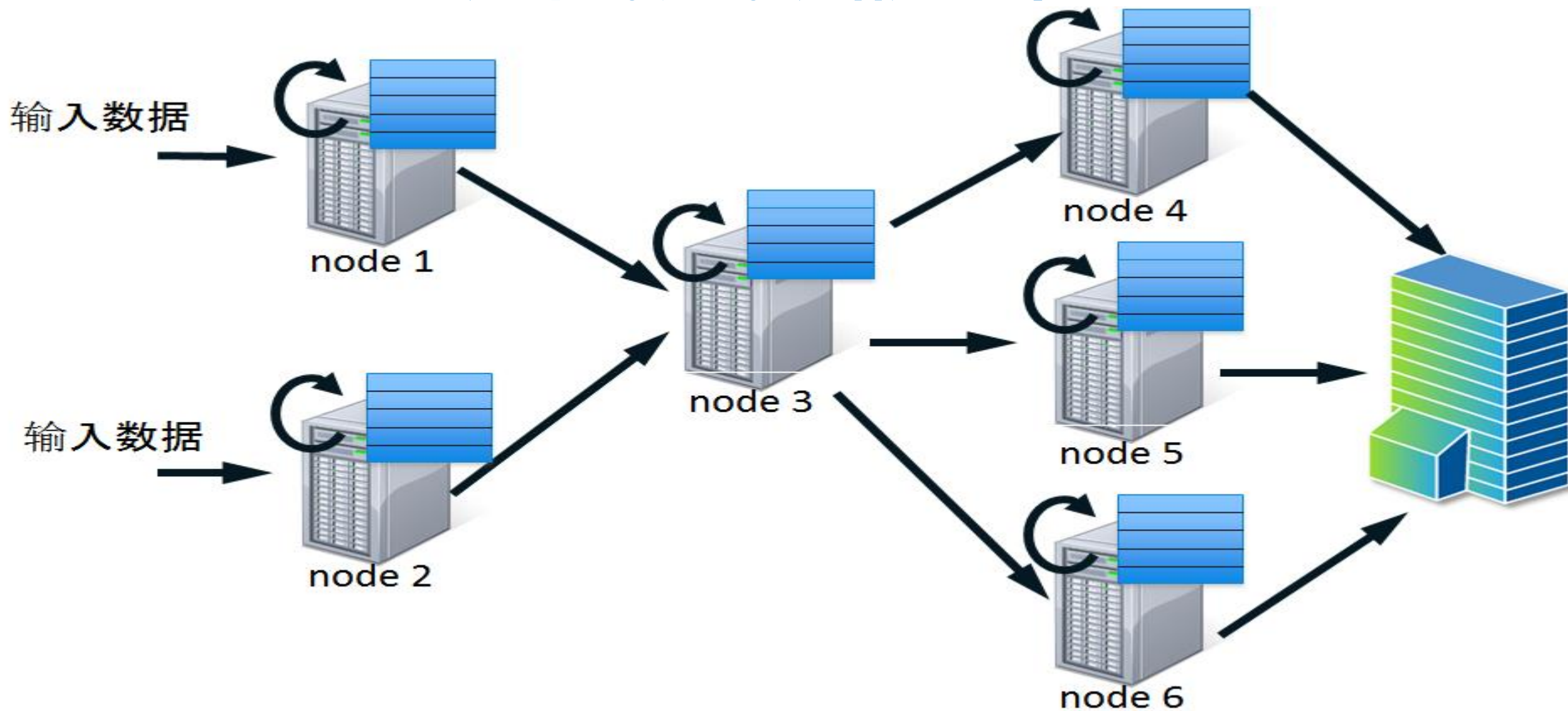


大数据分布式处理-批处理

- 为离线数据分析而设计，基本上是个利用数据并行性进行分布运算而后汇总结果的计算框架
 - 分析问题能够被并行化，且输入数据集可以被切分
 - 一个Map函数，在第一阶段计算<Key, Value>对
 - 一个Reduce函数，在第二阶段用于汇总Map函数的结果



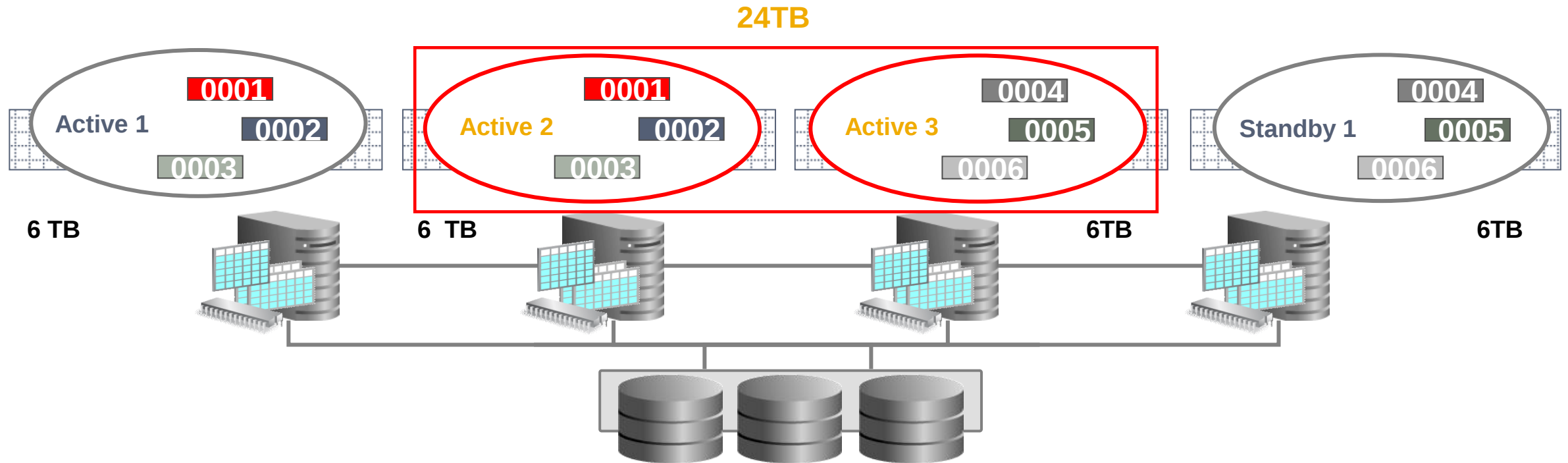
分布式流式数据处理架构



分层汇总数据存储

- 详单存储：HDFS或者HBASE
- 历史OLAP存储：Sybase IQ
- 实时查询：SAP HANA

Hot data/real-time data-SAP HANA



总结

- 技术不能搞定一切
- 不要试图采用一套方案搞定所有问题
- 建立和谐系统

英特尔与大数据



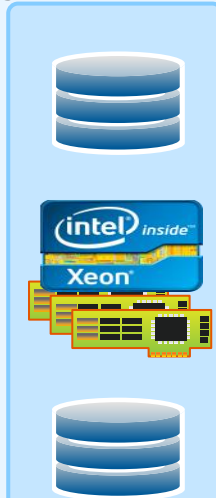
Performance and Scale Efficiencies



Data Delivery

Decision Support – CRM-ERP, OLTP, Batch

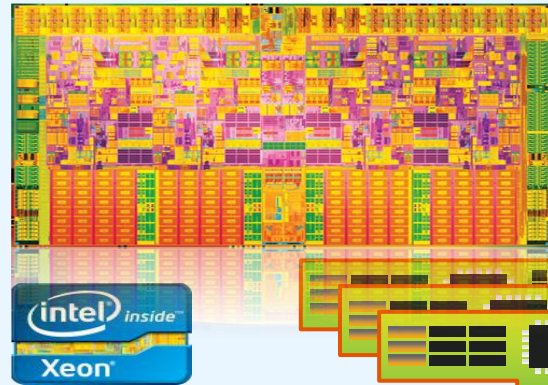
SOURCE – Big Data



ETL

ETL

Data Management



Data Usage

Business Intelligence

Business Strategy

KPI's

LOB Reporting



Rich Visualization

Reliability, Availability & Serviceability (RAS)

Power Management & Security

Intel大数据数据库架构配置建议

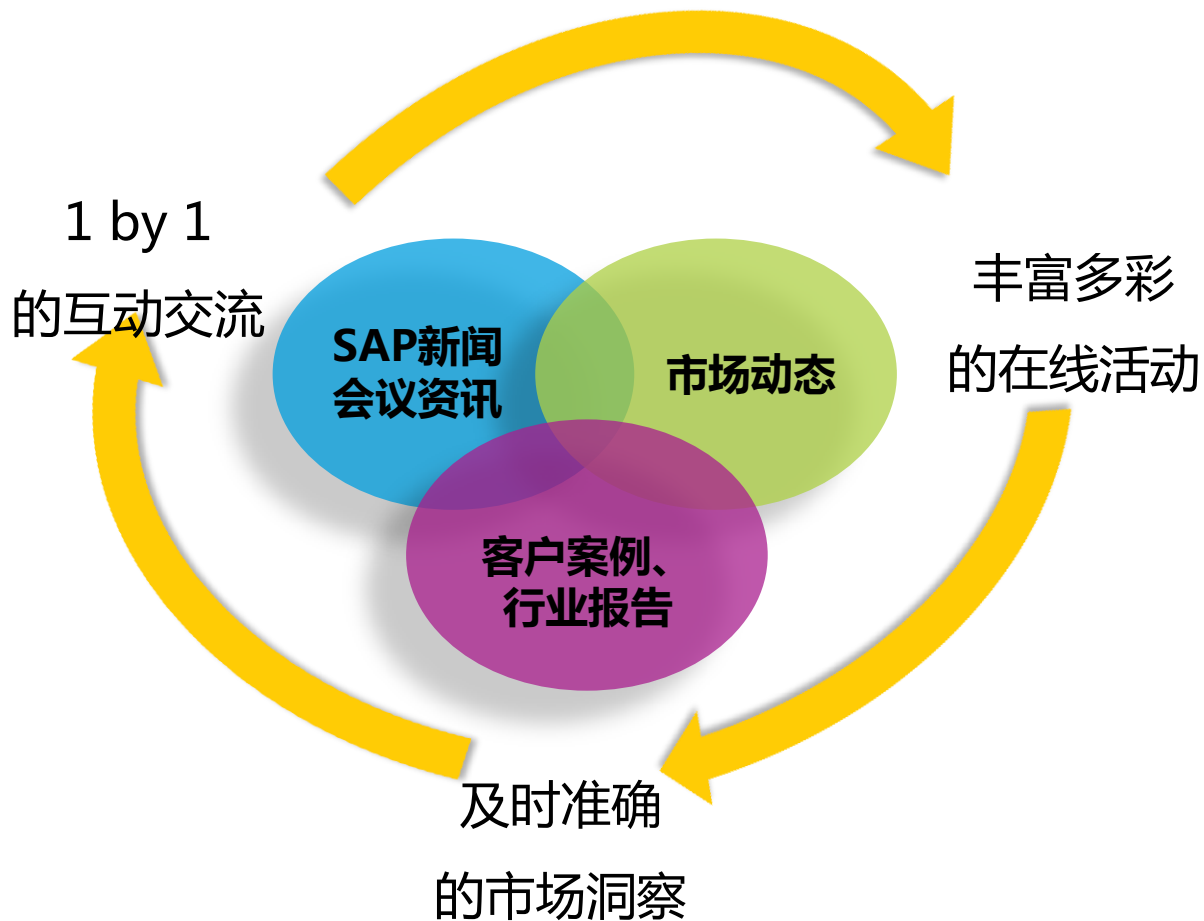
	CPU	网络	IO	内存
对象存储	2路E5(E3)	10GE	SATA	低
详单存储	2路E5-2630以上	10GE	SATA	中
数据处理	2路E5-2670以上	10GE	SAS	高
OLAP存储	4路 E7	10GE	SAS(SSD)	高
内存计算	4路 E7	10GE	SAS(SSD)	极高

Intel中国云创新中心



推动云计算数据中心方案创新，加速落地云实践

关注SAP官方微信：SAP天天事



尽在SAP官方微信：

SAP天天事



添加朋友中搜索：**sap**





第一时间分享：SAP新闻、市场动态、客户案例、行业
报告、线上线下会议等各种新鲜资讯

关注SAP官方微信：**SAP天天事**

感谢！



添加朋友→ 搜索：**sap**

有关 SAP 大数据 的详细信息，请访问：
<http://www.sapbigdata.com/>



感谢！

有关 SAP 大数据 的详细信息，请访问：
<http://www.sapbigdata.com/>