

# 大数据拥抱开源

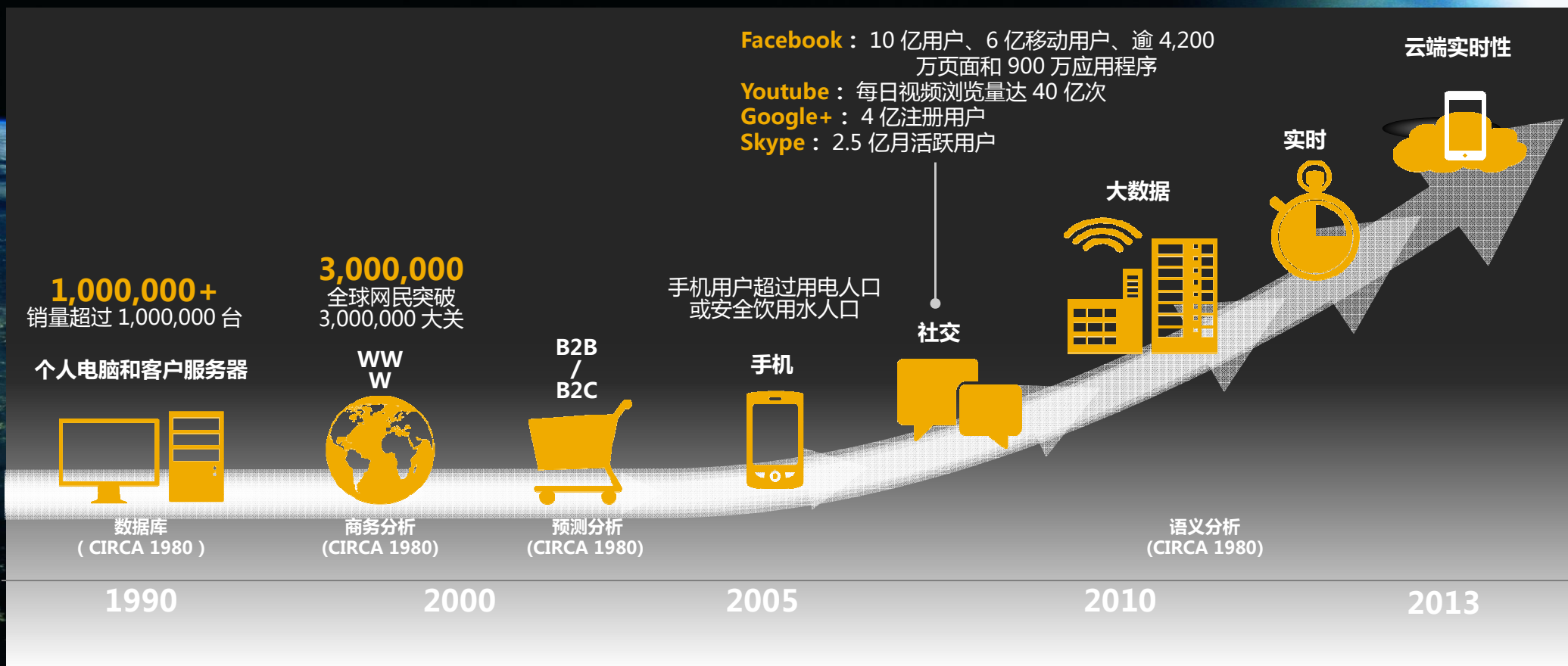
宋一平 Peter Song

售前总监 | 数据库及技术平台部

SAP China

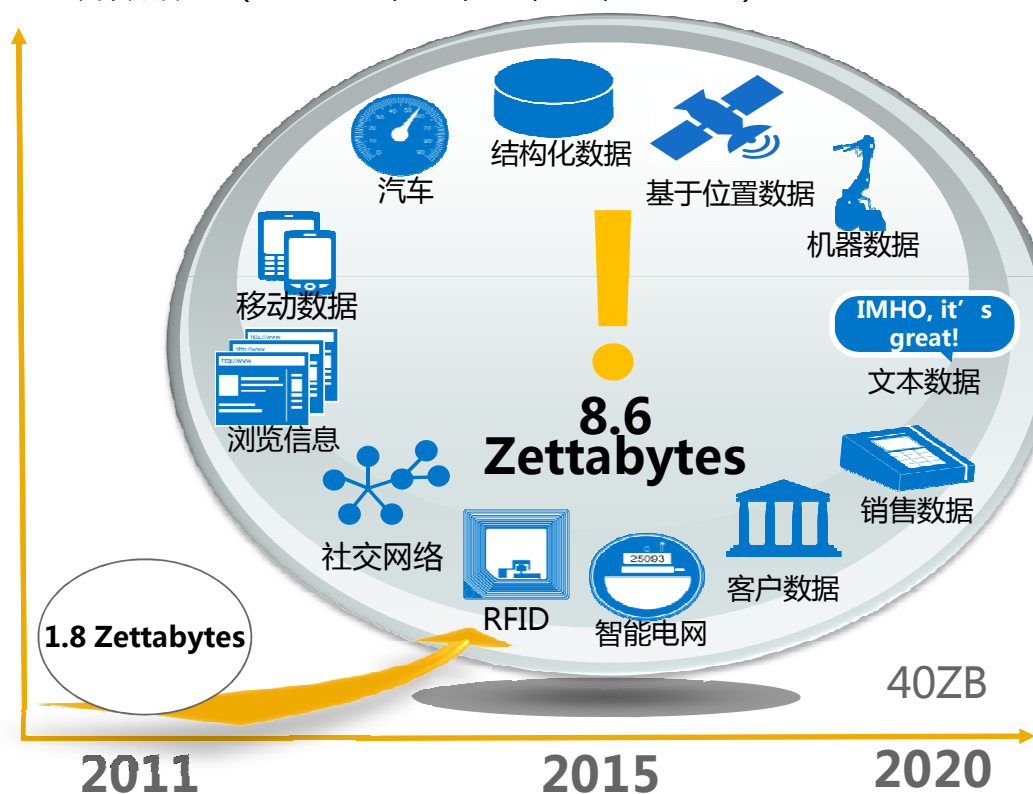
The SAP logo is located in the bottom left corner of the slide. It consists of the letters 'SAP' in a bold, white, sans-serif font, set against a blue triangular background that points towards the top right.

# 顺应大数据发展趋势



# 我们正处在一个信息爆炸的时代

全世界数据量 (1ZB = 1,000,000,000,000 GB)



**59%** 全球数据量的复合增长率  
- Gartner, 2013

**64%** 的企业正考虑大数据项目  
- Gartner, 2013

**238亿** 美元的市场 (2016年)  
- IDC, 2013

# 大挑战 → 大机遇



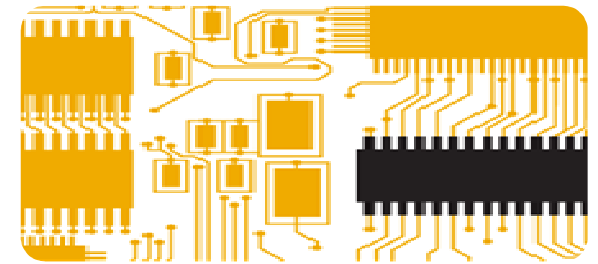
## 业务需求 BUSINESS NEEDS

社交媒体数据、空间地理数据和事件数据的出现  
多渠道体验(移动体验)  
从海量数据中做出快速的洞察力和正确的行动  
对信息质量和治理的信赖



## 数据特征 DATA CHARACTERISTICS

数据量激增  
不断增加的数据源和多样性, 包括社交媒体和机器产生的数据  
加速数据处理速度



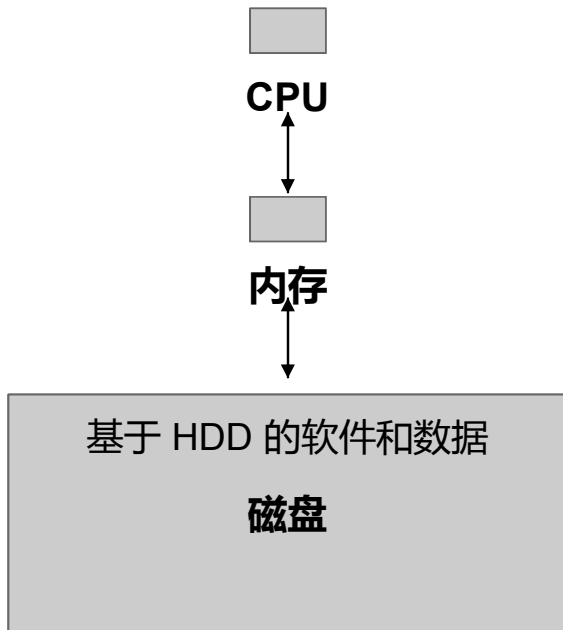
## 技术趋势 TECHNOLOGY TRENDS

存储 / 内存 / CPU  
增强内存计算  
EDW / 分布式 MPP /  
Hadoop  
数据挖掘 / 预测分析  
实时数据访问和事件管理

# 内存计算：速度提高 10,000,000 倍

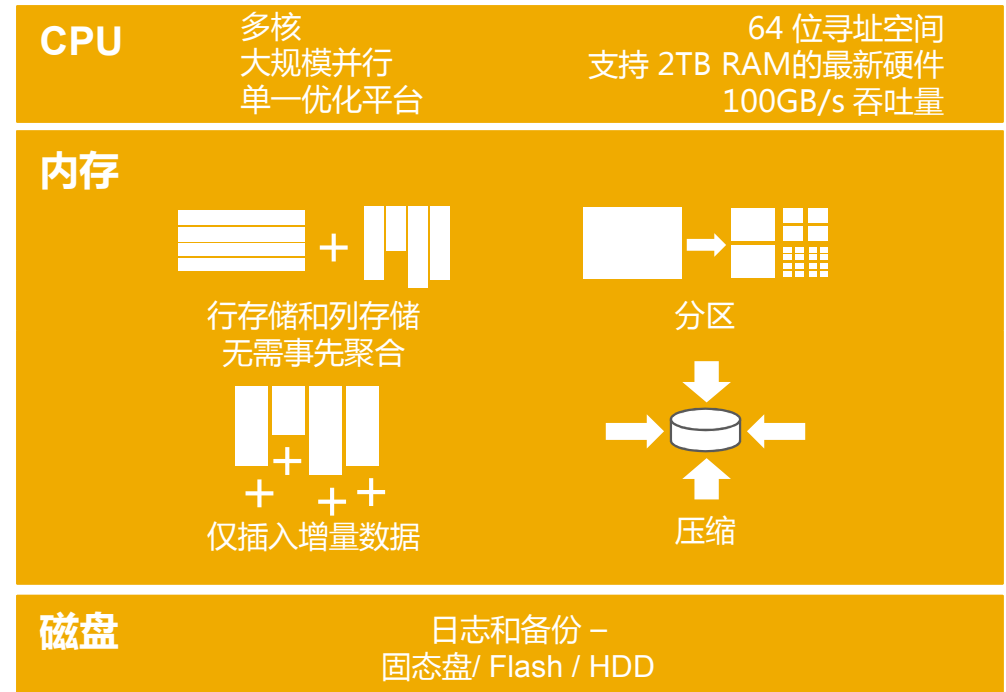
## 新思路

过去



- IO 限制
- 能支持多平台
- 但未经专门优化

今天



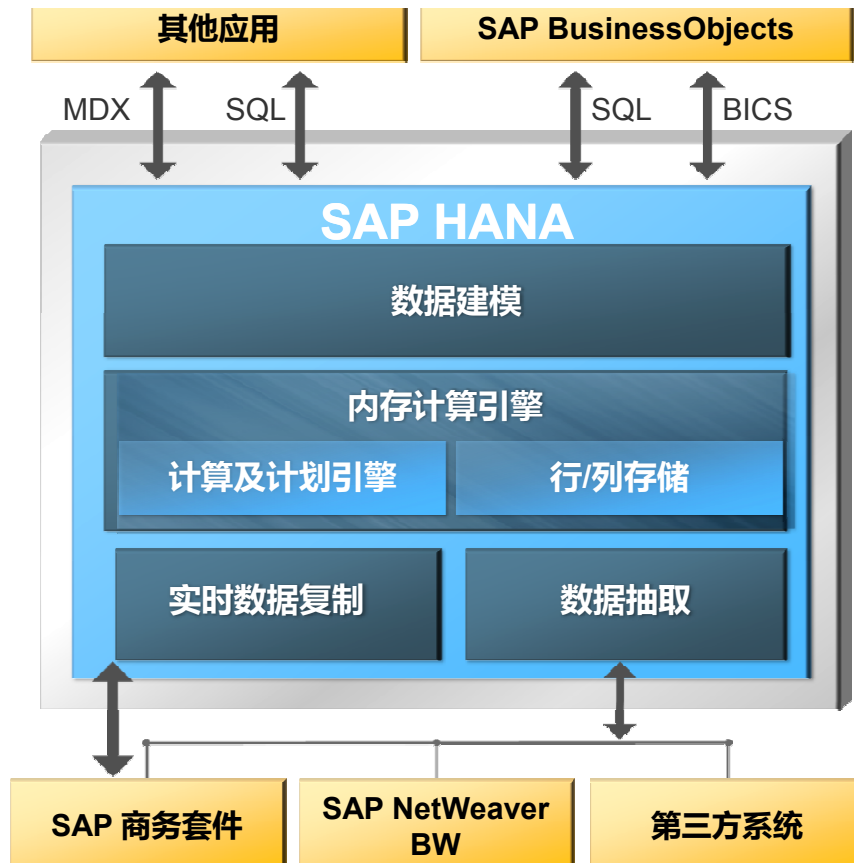
- 利用最新的硬件优势
- 最小化 IO 时间
- 针对 x86 平台专门优化

# SAP HANA

(High-Performance **A**Nalytic **A**ppliance)



**SAP HANA**是一项在本地内存中分析海量数据的技术，能够在刹那间获得复杂的分析与交易结果，实时完成业务决策，毫无延迟



## 什么是SAP HANA

- 预置的分析设备
- 基于内存的软件预装在硬件设备中
- 软件-SAP IMCE
- 包含数据建模、数据管理、安全管理及操作的工具
- 数据复制服务器、ETL 及SAP BOBJ协同工作
- 支持多种客户端应用
- 预置内容包（抽取器及数据模型）

## 功能

- 对海量数据进行的高速实时分析
- 基于历史以及实时数据，创建灵活的分析模型
- 减少数据重复
- 新一代应用的基础

# SAP HANA 支持平台

**SUSE Linux Enterprise 是SAP HANA推荐和支持的操作系统**

## SAP HANA 技术合作伙伴



### ■ SAP HANA Product Flavors

- SAP HANA Platform Edition
- SAP HANA Enterprise Edition
- SAP HANA Runtime Edition

## SAP HANA 硬件合作伙伴



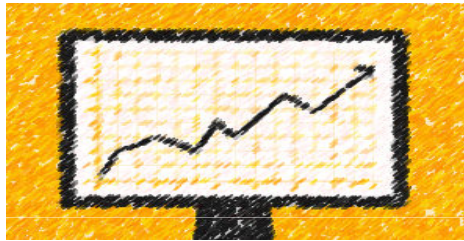
# 实时数据平台 市场领先的“五虎上将”

Sybase ESP,  
Replication Server,  
PowerDesigner, + SAP EIM



统一的实时数据管  
理平台

SAP Sybase ASE



交易型数据库  
Best 700

SAP Sybase IQ



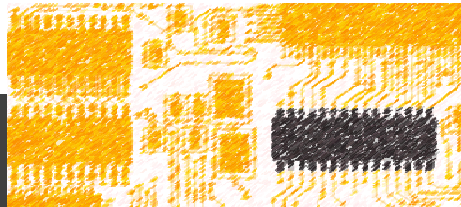
分析型数据库 Best  
700

SAP Sybase  
SQL Anywhere



移动及嵌入式数  
据库

SAP HANA

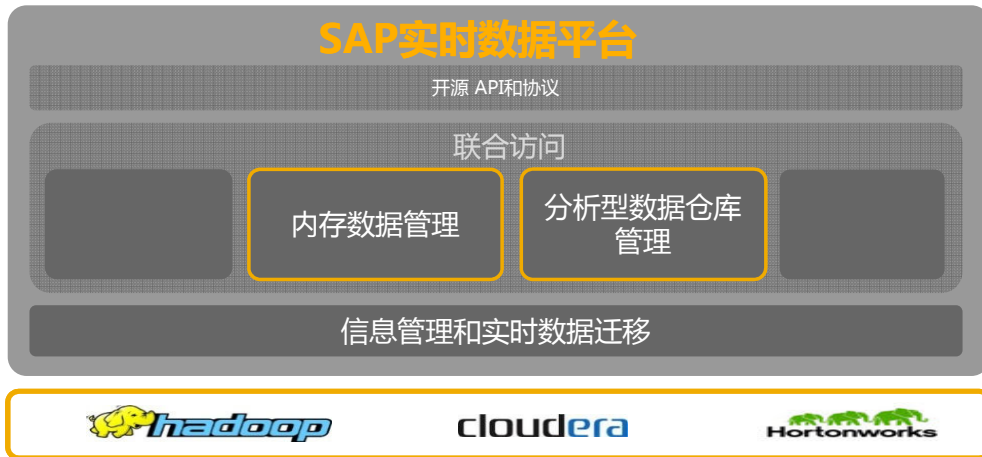


SAP 实时数据平台

对合作伙伴开放



# 大数据管理解决方案



## Sybase IQ：分析型数据仓库管理

- 全球第一列式数据库
- 提供高性能低价比的P级数据存储和管理
- 高性能、高扩展性、强大压缩：文本和二进制数据
- 支持不同SLA的数以千计的用户，具有大规模并行处理的网格计算技术
- 支持高级数据分析的内置数据库

## SAP HANA：内存数据管理

- 在一个数据库中建立外部交易数据、分析数据和应用逻辑处理
- 对海量结构化和非结构化数据进行实时分析
- 在内存中以闪电般的速度解决复杂问题的答案
- 在数据库中进行预测和文本分析
- SAP HANA是基于内存计算技术创新的大数据应用技术基础平台

## Apache Hadoop：未知价值的数据进行长期存储和批处理

- 一个分布式系统架构，由Apache基金会开发。
- 用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力高速运算和存储。
- 未知价值数据的新存储和处理模型
- 任何数据易于存储的宽松结构
- 具有较快速度反馈查询所有存储数据而设计的一种大型结构
- 发展为企业级使用的工具和管理
- 具有Cloudera和 Hortonworks认证的结构; (但是其他分布式结构也是支持的)

# SAP HANA + Hadoop : 真正的实时大数据分析

集合了两者**实时分析**和**无限存储**的优势



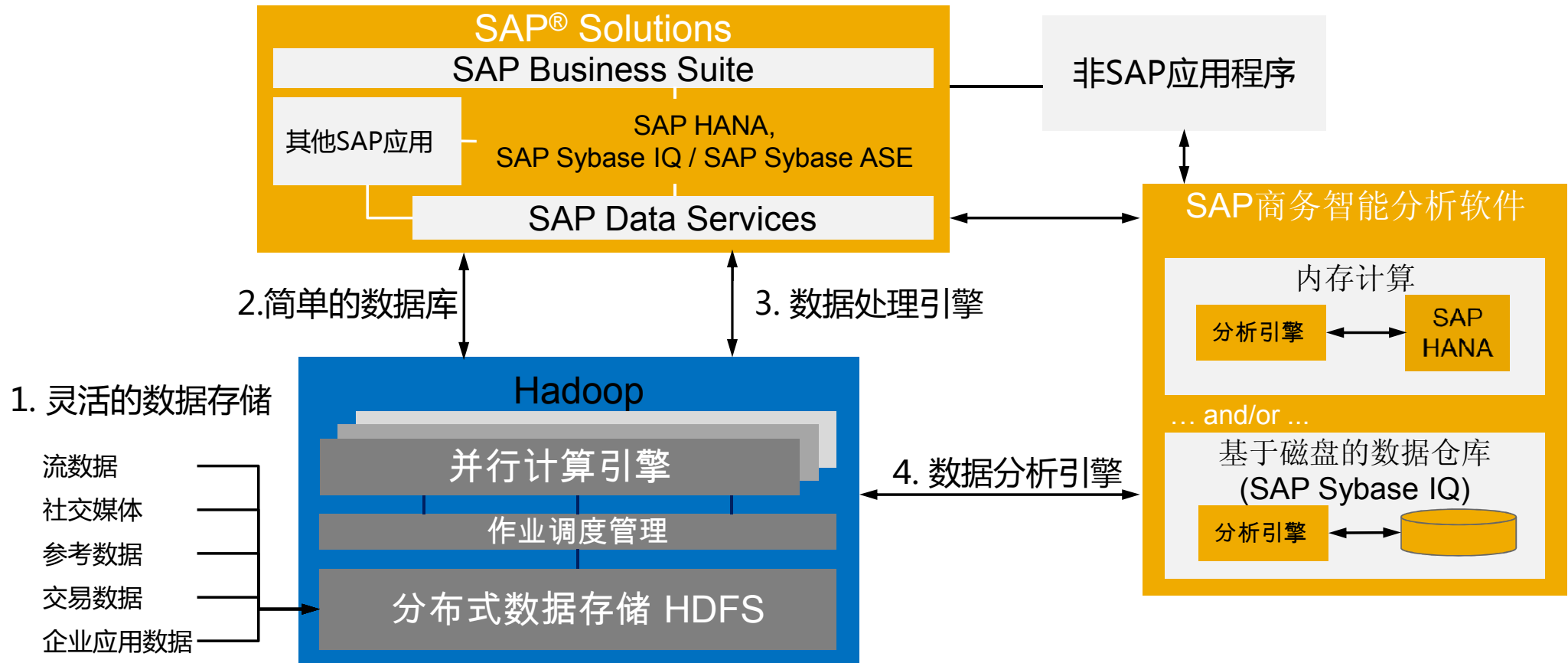
- 内存极速计算
- 实时的数据处理和分析
- 原生的预测和文本分析算法

+

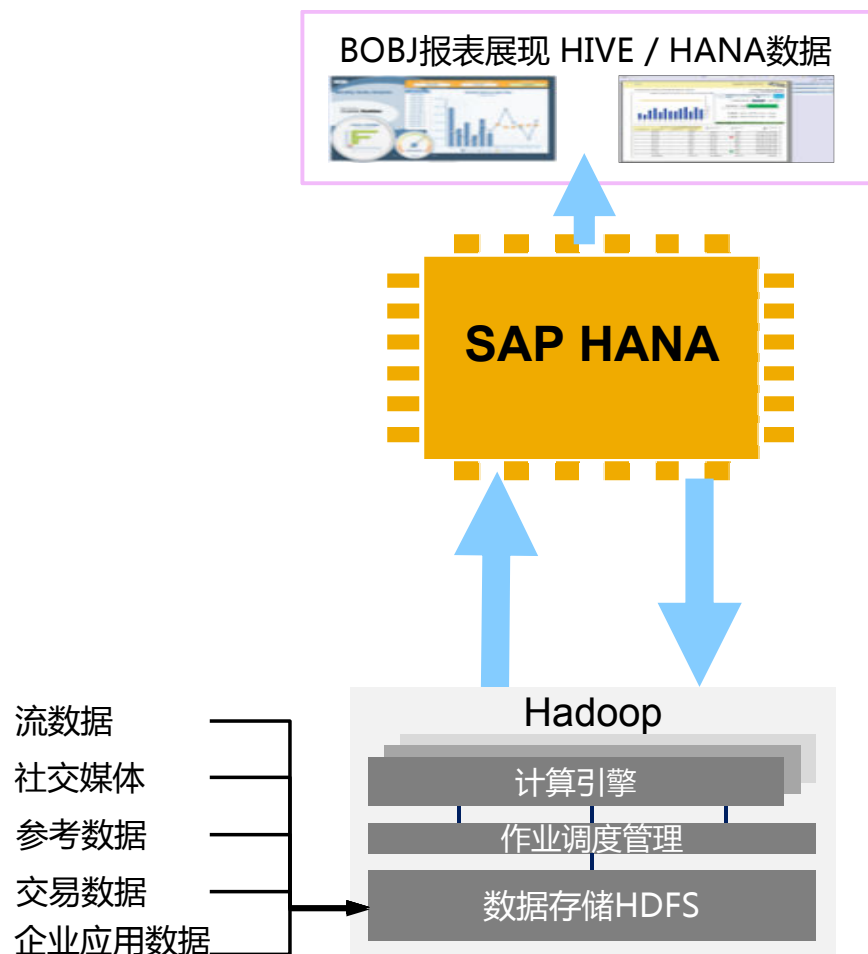


- 分布式磁盘文件系统
- 非结构化数据的无限存储
- No-SQL数据读取

# Hadoop与SAP集成 – 四个角色



# 应用场景一：Hadoop与SAP HANA优势互补



## ■ ETL层的集成整合

- Data Services 提供Hadoop的HIVE和HDFS数据连接，将整个数据抽取过程推送到Hadoop中作为一个MapReduce作业来执行

## ■ 从HANA直接到Hadoop的数据连接

- Proxy 表 (HANA SP6)，虚拟HANA表联邦查询Hadoop的Hive表
- Hcatalog集成 (HANA SP6)，利用Hadoop的元数据来提高查询性能

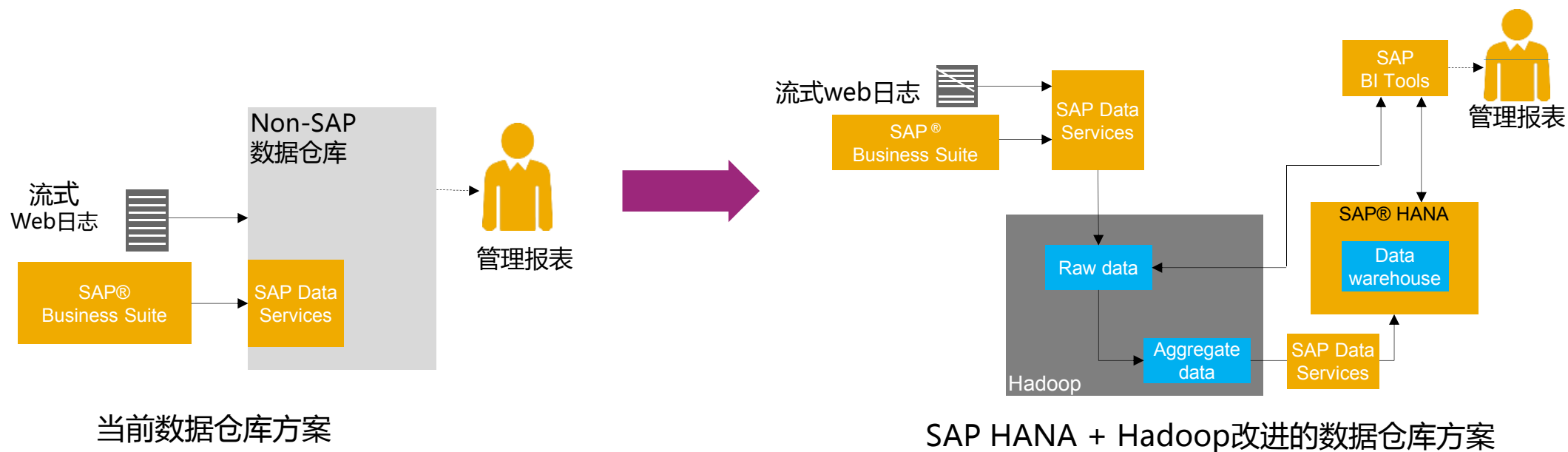
## ■ SAP BI工具的Hadoop数据连接

- SAP BOBJ 语义层直接读取Hadoop HIVE数据

## 应用场景二: 作为传统数据仓库的替代方案

### SAP HANA+HADOOP作为传统数据仓库的替代方案

- 低价值, 对实时要求不高的数据存储在Hadoop
- 高价值, 对实时要求高的数据存储在SAP HANA



## SAP HANA + HADOOP + R



**408,000x** 在做方案验证时，快于传统的磁盘系统



**216x** 更快的 DNA 分析结果 - 从 2-3 天 to 20 分钟

### 益处

- 减少诊断变体的 DNA
- 内存加速预测和相关分析
- 基于DNA突变，优化治疗计划
- 可长期研究基于 DNA 癌症治疗



Genomic DNA analysis in real-time will transform how we enable comprehensive patient care to fight against cancer. SAP HANA will be the mission critical and reliable data platform to make real-time cancer analytics into a reality. Separately, our internal technical comparison demonstrated that SAP HANA outperforms a traditional disk-based system by factor of 408,000 when performing other types of data analysis.

*Yukihisa Kato, Director & Executive Officer, CTO, Research and Development Center, MITSUI KNOWLEDGE INDUSTRY CO.,LTD.*

# 大数据分析工具



BusinessObjects  
BI Suite



SAP Predictive  
Analysis



SAP Visual  
Intelligence

第三方开源分析  
工具

R等

报告、即席查询

深层次商务分析

自定义可视化

实时答案

- SAP商务智能套件工具分析大数据
- 客户可以不用写任何代码，来实现报表、查询、例外分析
- 标准化的BI工具，分析Hadoop中解析数据
- 通过语义层访问Hadoop Hive
- 支持多个和单个语义层
- 可以分析多个数据源
- 由水晶报表、水晶易表、仪表盘、探索、分析等组成的套件

- 详细数据的实时分析工具
- 复杂预测模型的可视化设计
- 利用内置数据库处理、可以和R进行整合，进行大数据分析预测
- 可以有业务部门自己应用的简单易用的预测工具
- 不需要聚合和调优，可以保证高效率
- 从SAP HANA中获得信息的单一视图

- 创建自己的可视化工具
- 从各种数据源中获得数据
- 组合多个数据源，将数据以各种直观的方式展现出来
- 可视化的界面，几乎不需要培训就可以上手
- 可以使用的多个高级选项
- 可以利用PMML和其他的BI工具进行共享

- 利用已经存在的环境使用你的BI工具
- 运行常规的R开源算法或者SAP的3,500+ 算法库，在数据库中做高性能的模型建立和分析
- 在详细的大数据中进行实时的数据预测和分析
- 不用数据聚合和调优

# SAP BusinessObjects BI 解决方案

## 提供业务洞察力，实现有效分析

SAP BusinessObjects BI 解决方案使业务用户能够通过一整套商务智能工具访问统一的信息，从而使人员和团队能够在—个可扩展的商务智能平台上做出自信的决策并保持步伐—致



### 报表

如何访问企业数据并将其转化为高度格式化的报表，从而增强洞察力？

### 仪表盘与可视化

如何可视化数据，以制定更明智的决策？

### 交互式分析

如何回答即席问题并与信息进行交互？

### 高级分析

如何根据复杂的历史数据确定发展趋势，并做出更准确的预测？

### 数据探索

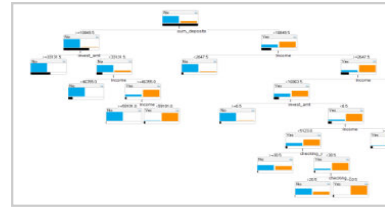
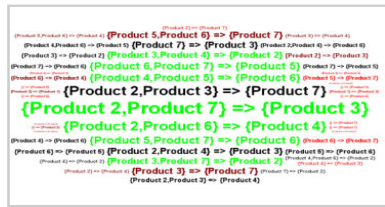
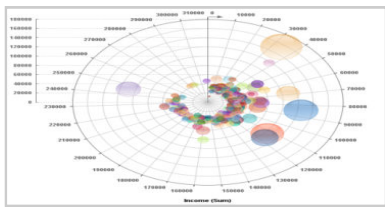
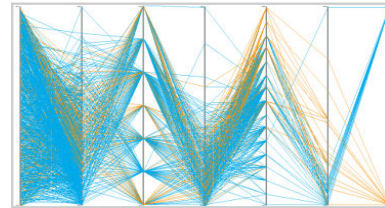
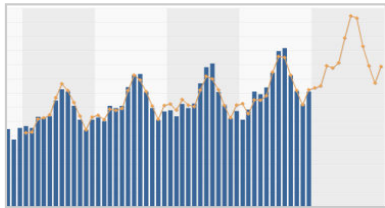
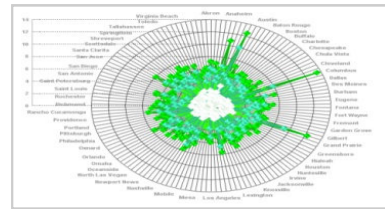
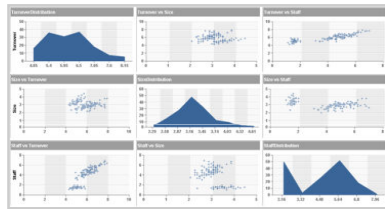
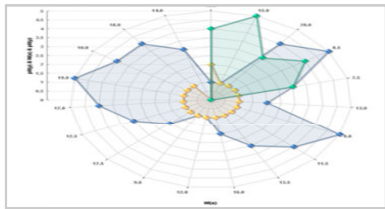
如何立即获得业务问题的答案？

统一的平台



# SAP BusinessObjects 预测分析

- 通过直观易用的方式来设计复杂的预测模型，从而将数据可视化，从数据中发现和分享隐藏的密码
  - 通过内存计算实现实时效能，内含建模工具
  - 支持SAP HANA中基于内存数据库的预测算法
  - 利用SAP HANA中集成的 3500多种开源的 R 预测算法



- Tree Map Chart
- Heat Map
- Pie Chart
- Pie with Variable Slice Depth
- Multiple Pie Chart
- Donut Chart
- Column Chart
- Bar Chart
- Column Chart with Dual Value Axes
- Line Chart
- Line with Dual Axes
- Surface Chart
- Combined Column and Line Chart
- Combined Column and Line Chart with Dual Value Axes
- Stacked Column Chart
- 100% Stacked Column
- Stacked Bar Chart
- 100% Stacked Bar
- 3D Column Chart
- Box Plot Chart
- Radar Chart
- Multiple Radar Chart
- Tag Cloud Chart

# 大数据有价值的场景分析

将SAP HANA中的实时分析数据在内存中进行分析和在Hadoop中的运营数据的批处理分析紧密结合到一起

- **营销优化:** 基于实时客户和市场数据的分析，优化营销战役
- **客户细分:** 基于客户的行为和活动，将客户进行细分
- **消费情感分析:** 了解客户的品牌喜好、满意度和客户忠诚度
- **客户流失分析:** 产生高风险客户的列表，研究客户维护项目计划
- **运行时间运营分析和改善:** 偶发事故模式、运营设备跟踪
- **欺诈行为发现:** 对于可疑行为分析和跟踪。

## 业务结果

- 更好的客户服务
- 降低高成本客户数量
- 改善网管绩效
- 优质运营效率

# SAP大数据处理框架

## SAP大数据处理框架

呈现应用

SAP移动应用、Sybase无线平台

SAP BusinessObjects 商务智能解决方案

SAP 商务套件、SAP BW 和 SAP 应用

处理

**Sybase ESP**

流和事件处理

**Sybase ASE**

交易处理

**SAP HANA**

内存计算引擎

**Sybase IQ**

分析网格

**Hadoop**

MapReduce  
批处理计算框架

存储

数据库引擎

数据库引擎

数据库引擎

Hive/HDFS

获取

**Sybase ESP**

监控过滤事件流

**Sybase Replication Server, SAP BusinessObjects Data Services**

(整合 / 同步跨系统数据)

半结构化数据

结构化数据

非结构化数据



**谢谢 !**

**Peter Song**  
**[Peter.song@sap.com](mailto:Peter.song@sap.com)**

**Presales Director**  
**Database and Technology / Business Analysis**  
**SAP China**

