软件发掘数据价值

GBASE

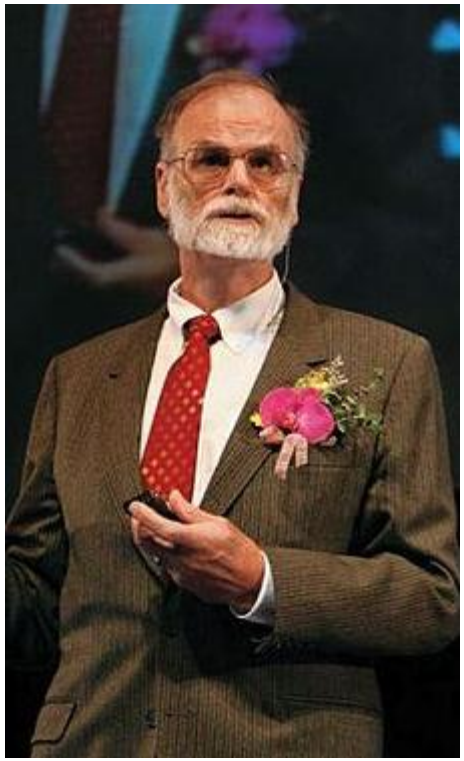
# MPP数据库技术及其在电信、金融行业的大数据应用

南大通用数据技术股份有限公司

2014年9月13日

# 目录

# MPP并行数据库的理论基础



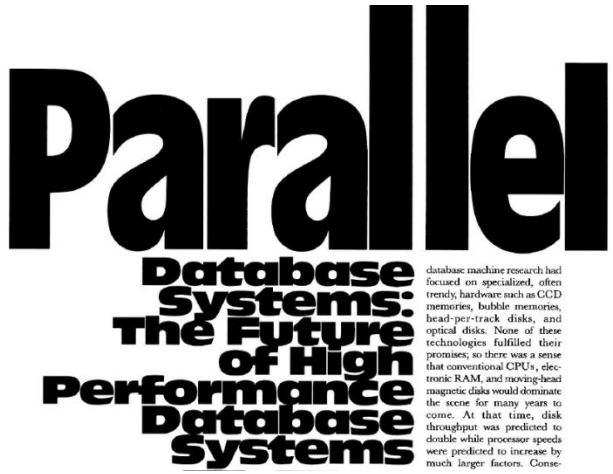
Jim Gray
在数据库和事务处理领域的持续研究成就
1998年获图灵奖

- 大数据时代来临的预言家：

“未来每18个月产生的数据量等于有史以来的数据量之和”

Jim Gray 1998年图灵奖获奖演说

- 数据库技术对大数据需求的解决之道：


Parallel
Database Systems: The Future of High Performance Database Systems

database machine research had focused on specialized, often trendy, hardware such as CCD memories, bubble memories, head-per-track disks, and optical disks. None of these technologies fulfilled their promises; so there was a sense that conventional CPUs, electronic RAM, and moving-head magnetic disks would dominate the scene for many years to come. At that time, disk throughput was predicted to double while processor speeds were predicted to increase by much larger factors. Conse-

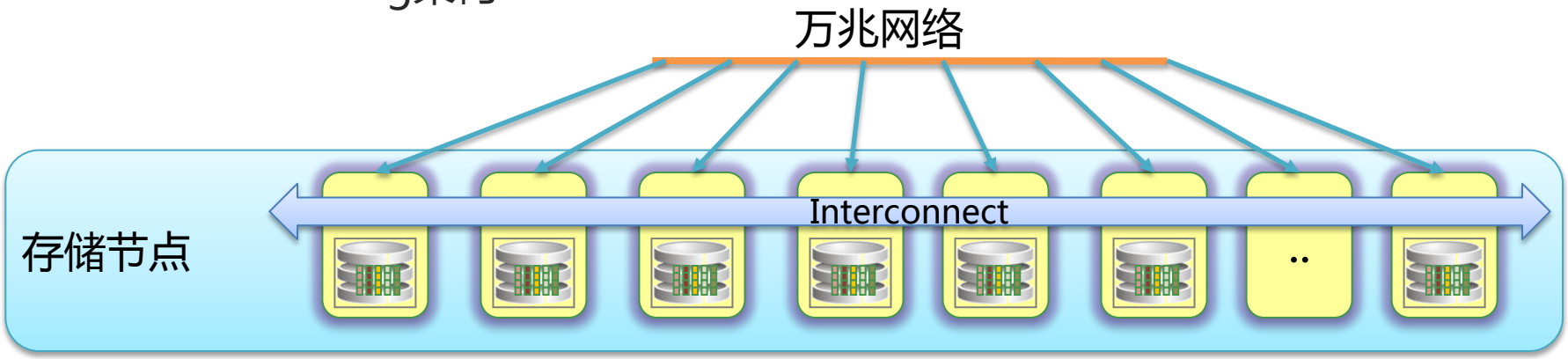
《Parallel Database Systems:
The Future of High Performance Database Systems》

1992 By David Dewitt and Jim Gray

# 什么是MPP？

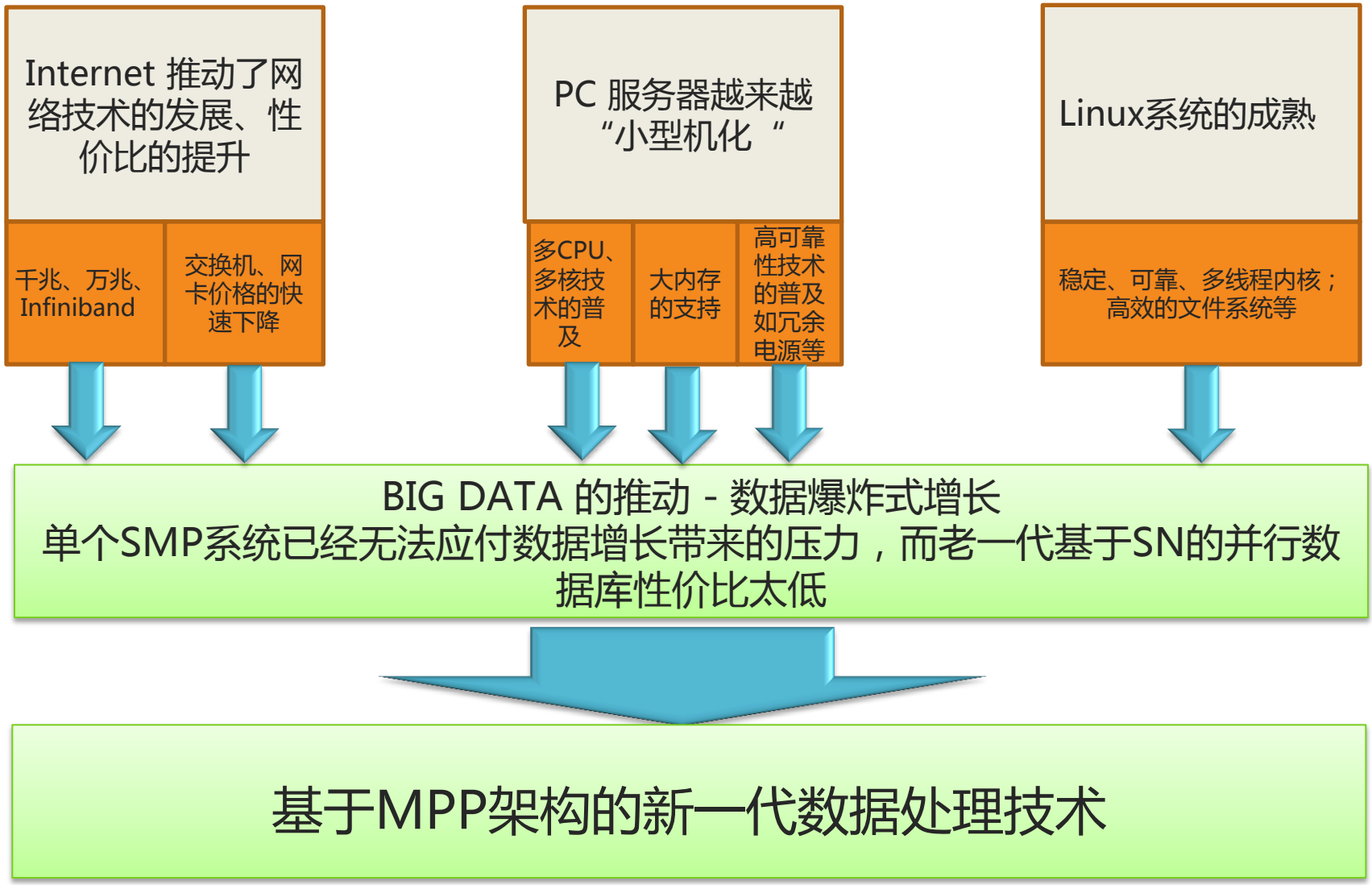
**MPP (Massively Parallel Processing)**：大规模并行处理系统，系统由许多松耦合处理单元组成的。每个单元内的CPU都有自己私有的资源，如总线、内存、硬盘等。在每个单元内都有操作系统和管理数据库的实例副本。这种结构最大的特点在于不共享资源。

- **MPP架构数据库应具有的特征**
  - 任务并行执行
  - 数据分布式存储
  - 分布式计算
  - 私有资源
  - 横向扩展
  - Shared Nothing架构
- **MPP架构数据库**
  - OldSQL
  - NewSQL
  - NoSQL



# MPP在大数据技术宏观架构中的定位

# 为什么新一代数据处理技术都是MPP架构？



# MPP数据库核心技术

1. 关系模型和CAP问题

2. 并行计算

3. 横向扩展技术

4. 动态扩展和资源管理技术

# MPP数据库核心技术：关系模型和CAP问题

## 关系模型

- SQL92 的支持
  - 必须支持 + SQL99 的一些特性（BI 函数）
- 存储过程
  - 必须支持，最好兼容Pl/SQL
- ACID 的变化和数据的强一致性
  - 事务的完整性： 类似2-P Commit 机制（简化的Paxos）
  - Read committed： 集群层面的MVCC？是否需要 ？
  - Redo、undo 的问题： AB 切换 vs redo log

## CAP问题

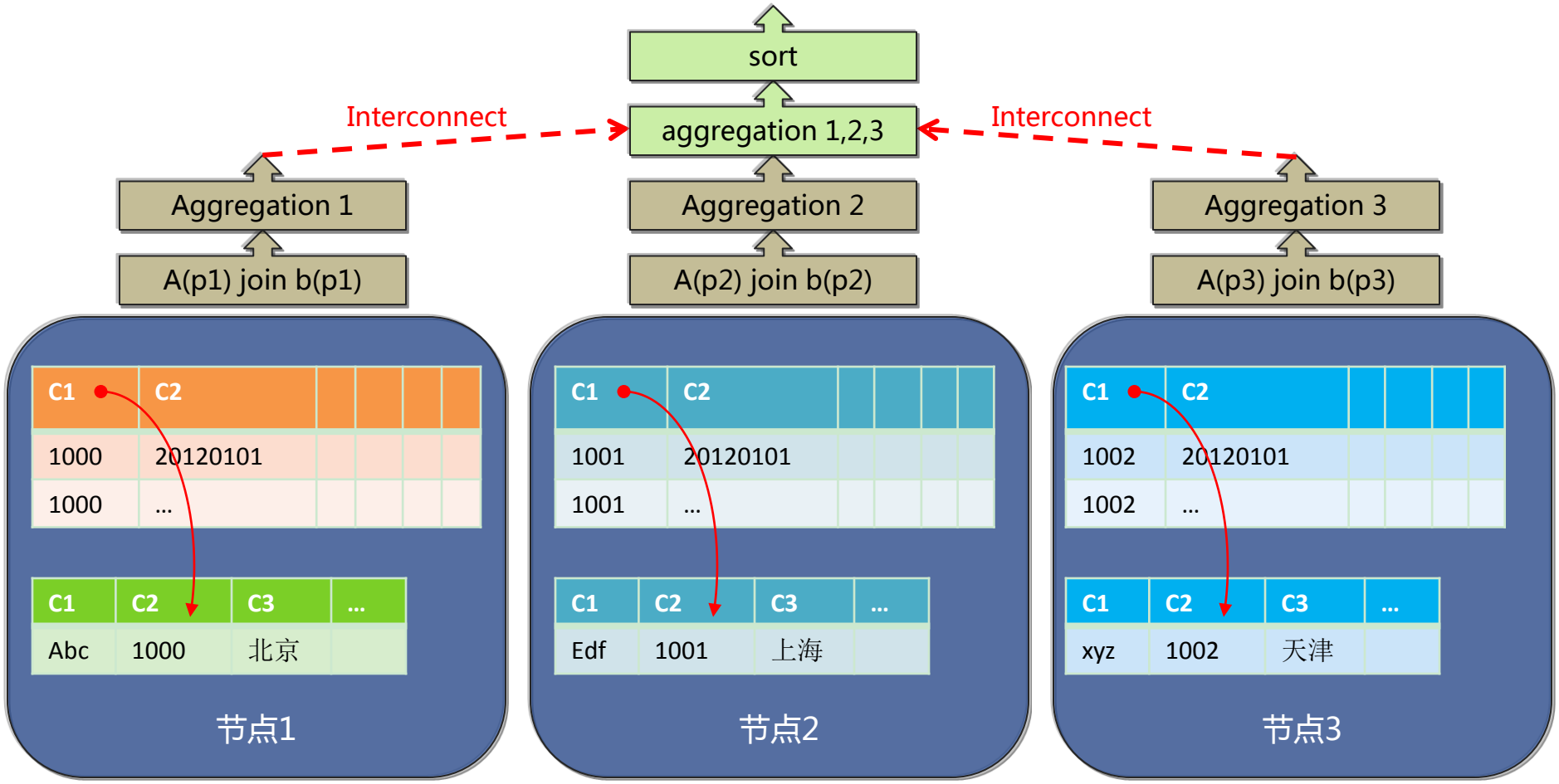
- CAP 是否可怕 ？ 3选2 ？
  - Consistency，Availability，Partition-Tolerance
- CAP 其实是完全可用的
  - 集群出现拓扑变化时，CAP才介入
  - CAP要求我们必须取舍： MPP数据库一般选择CP，牺牲A（并不意味集群不可用，只是暂时把高可用水平降低了）
  - 当集群分裂、节点故障、网络故障时，一般主动选择部分节点或功能暂时off line, 然后同步、修复失败的节点。
  - 这样的策略分为三个步骤：探知分区发生，进入显式的分区模式以限制某些操作，启动恢复过程以恢复数据一致性并补偿分区期间发生的错误。
- 在实践中，每个操作都要考虑CAP效应，软件故障频率远大于硬件故障

# MPP数据库核心技术：并行计算

- 并行计算的效率取决于数据分布特征和SQL算子
  - Hash 分布是最常用的方法
  - 注意数据倾斜问题
  - 静态hash与动态hash的结合（多表关联）
  - 最终实现本地join是核心
- 并不是所有的算法都能很好的线性扩展：
  - Select count（distinct x）…
  - OLAP functions
  - 复杂SQL
- 正确评估分布式执行计划的成本是执行器的核心问题
  - 数据在节点间动态迁移是不可避免的
  - 网络速度、数据动态重分布效率、pipelining是核心

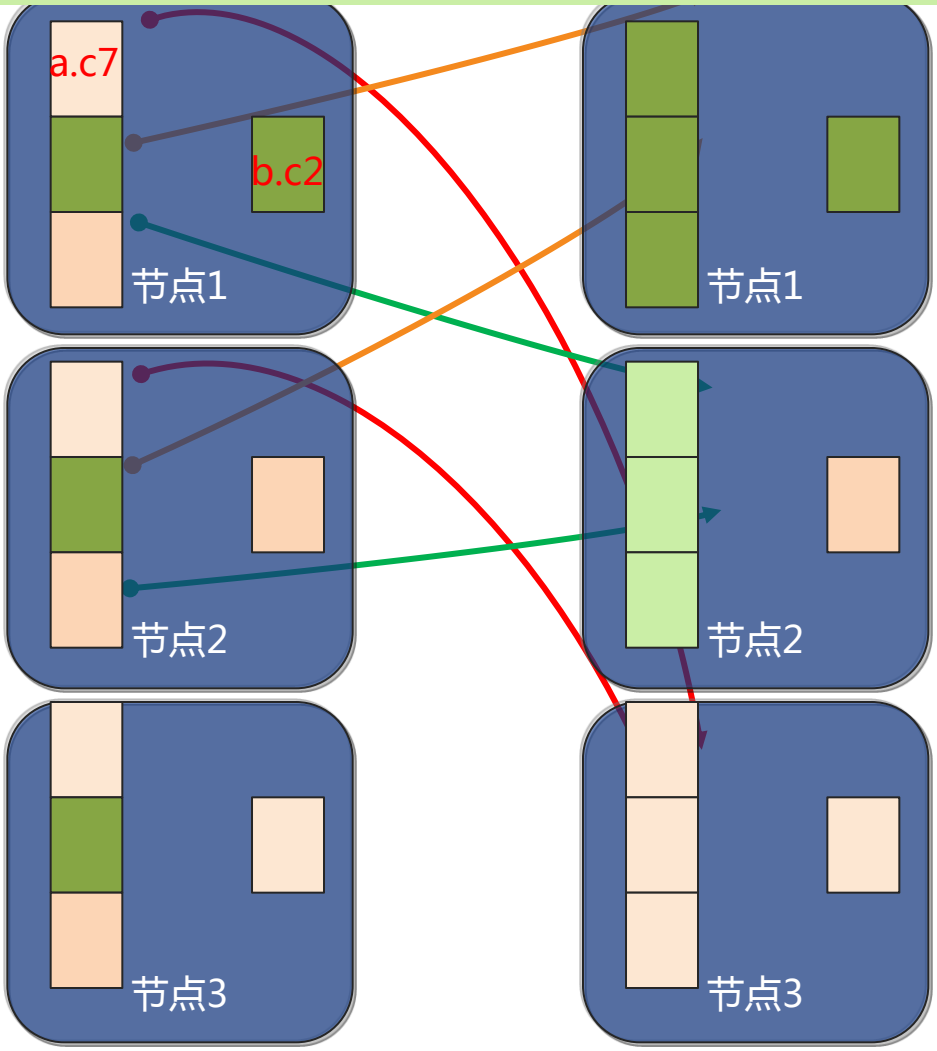
# MPP并行计算技术之：静态hash join

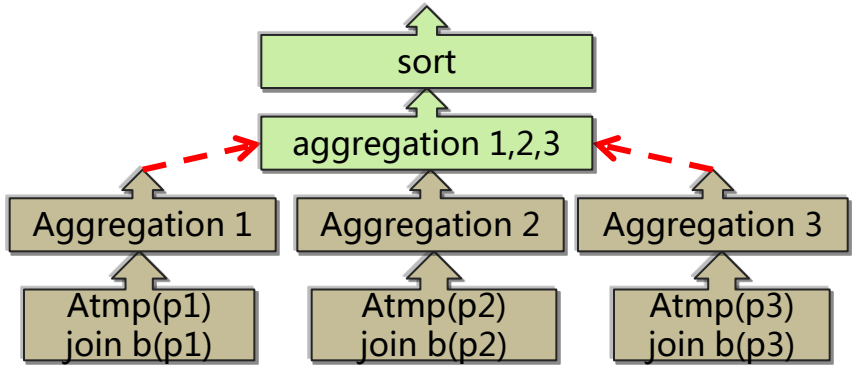
Select b.c3,sum(a.c5) from a, b where a.c1=b.c2 and ....  group by b.c3 order by ...



# MPP并行计算技术之：动态hash join—需广播数据

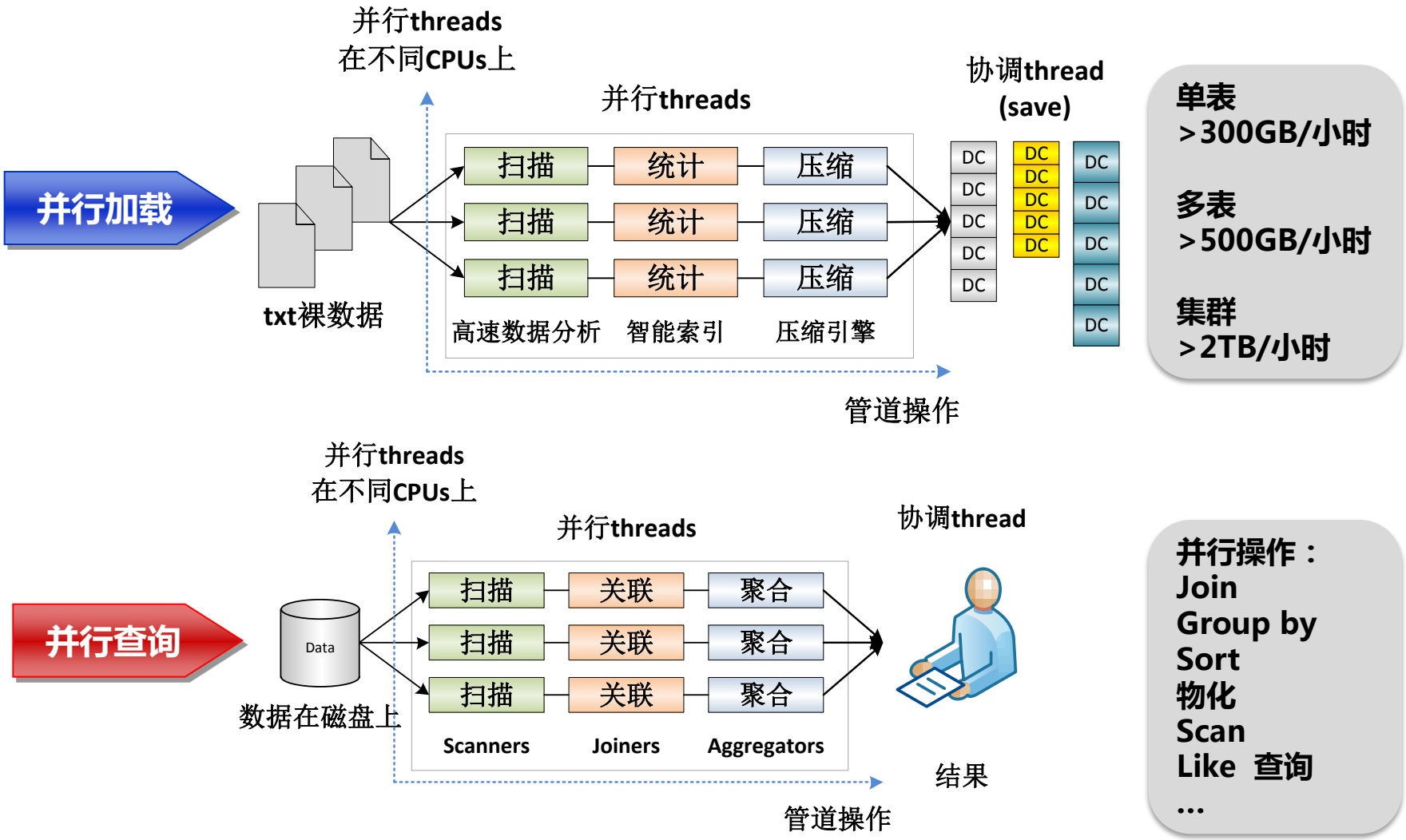
Select b.c3,sum(a.c5) from a, b where a.c7=b.c2 and …. group by b.c3 order by …



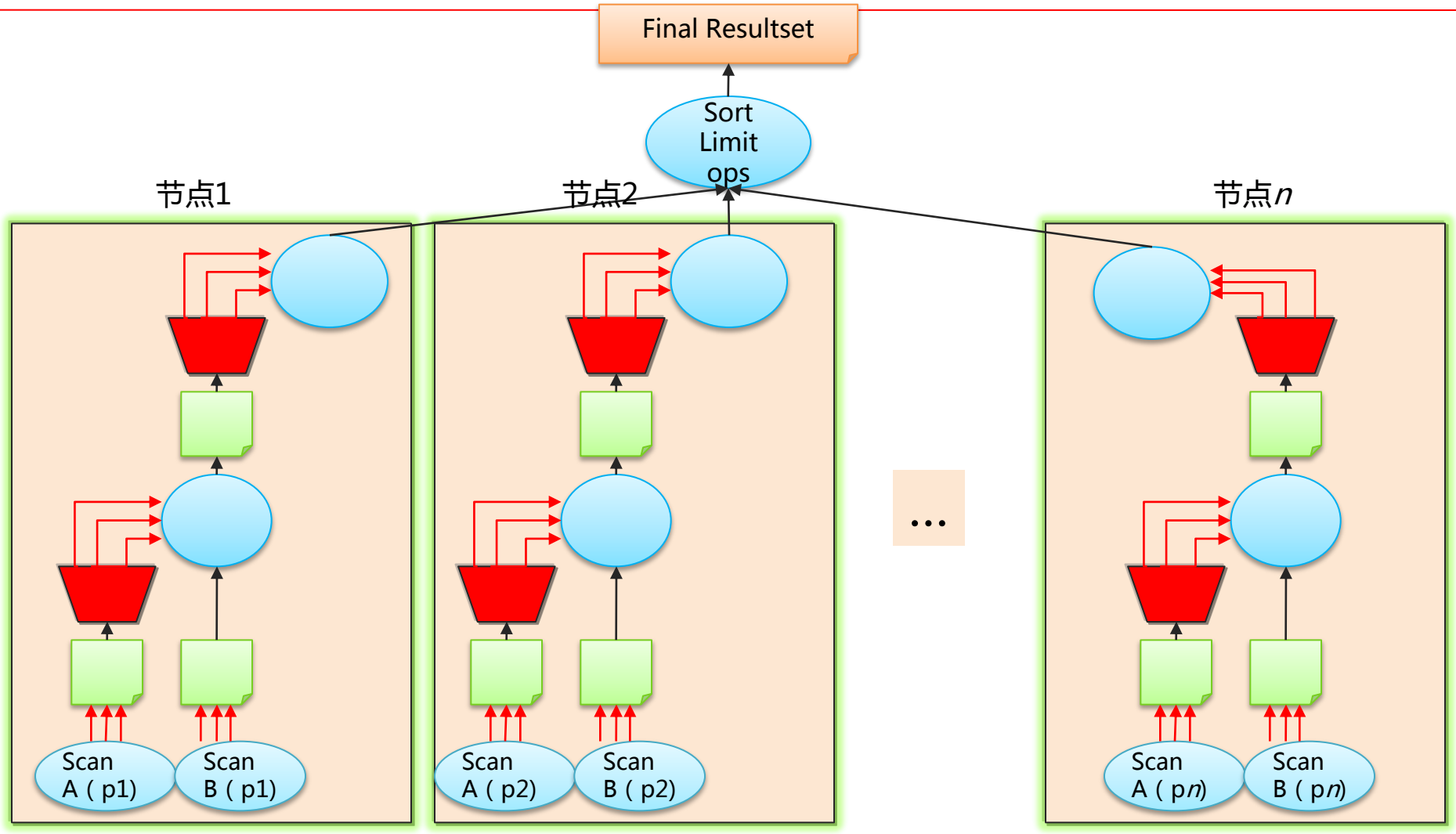




- 列存储减少节点间数据传输量
- 使用pipeling 并行技术提升效率
- 边传输、边计算
- CBO动态取样，评估网络成本
- 动态hash + 静态hash实现分布式算子、保证集群的扩展能力

# MPP并行计算技术之：SMP多核CPU双向并行技术



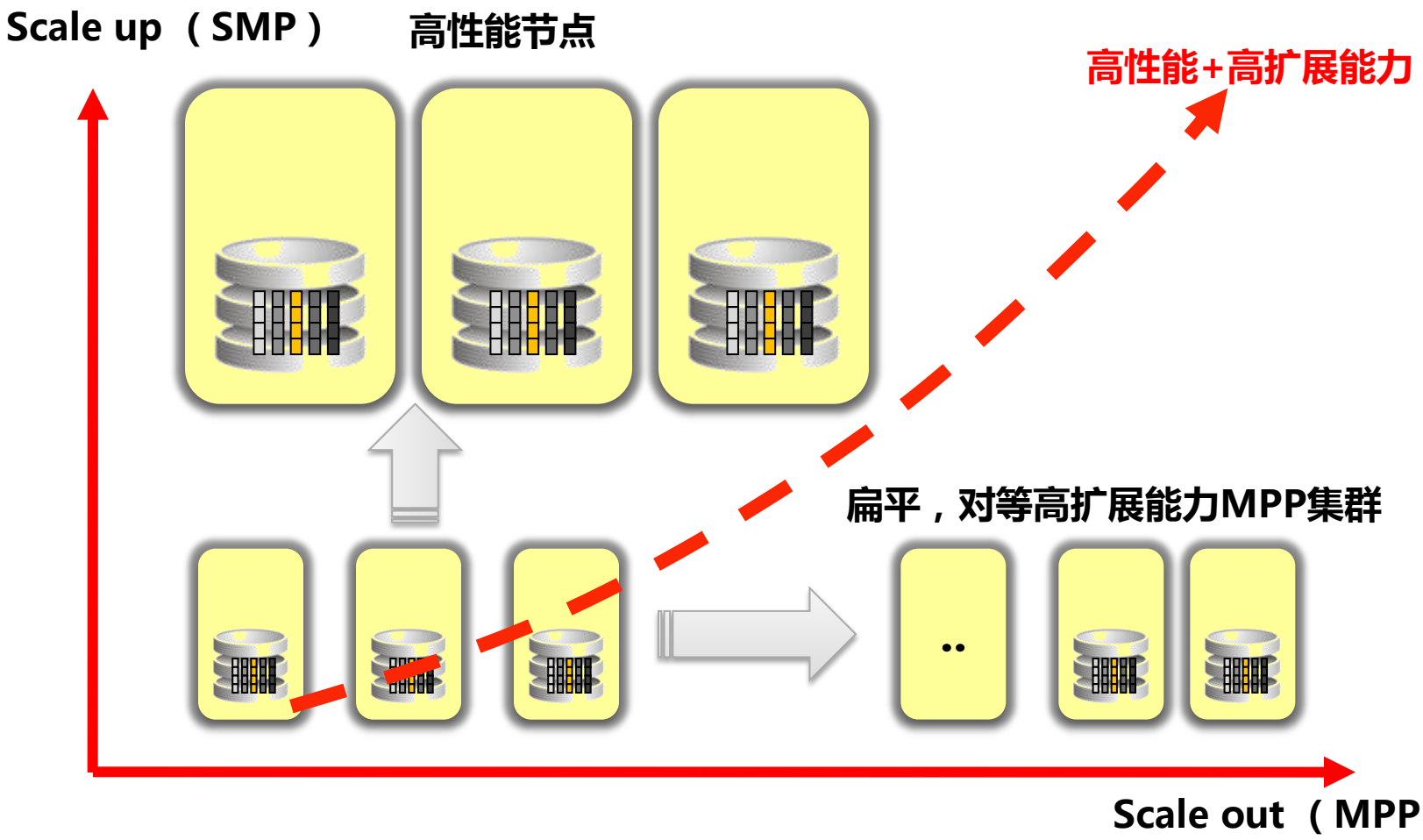
# MPP并行计算技术之：SMP + MPP 多层并行




同时可使用：计算能力>1千个核，I/O > 10GB/s，内存>10TB

# MPP数据库核心技术：横向扩展技术

**Shared Nothing + MPP集群性能随节点数增加呈近似线性关系**



# MPP数据库核心技术：动态扩展技术和资源管理技术

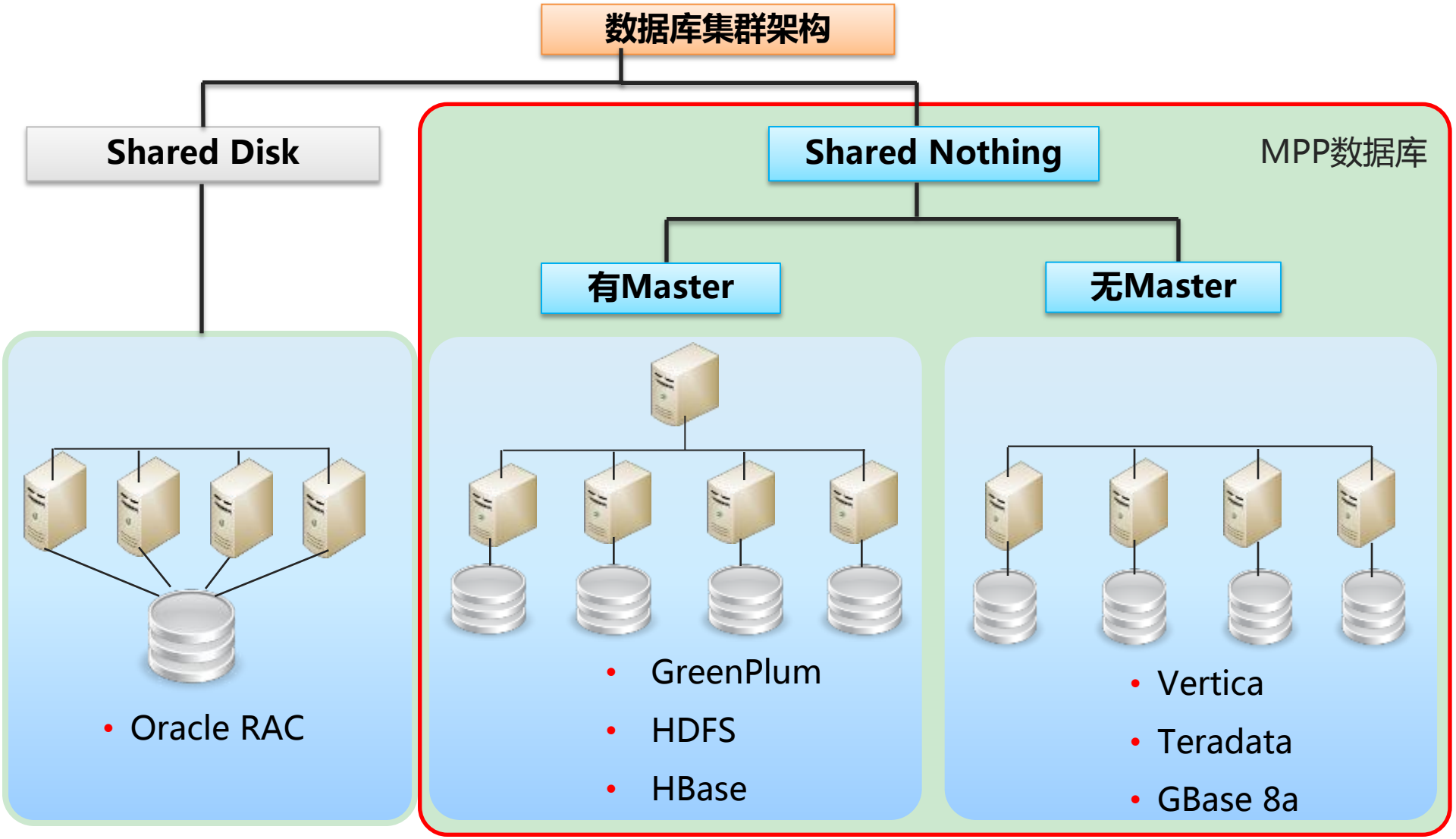
## 动态扩展技术

- **对于MPP数据库，在线扩展比DFS要困难**
  - 需要扩展的是hash后的数据，而不是文件块（HDFS，DFS）
  - 扩展后空间回收问题
  - 扩展过程中系统可提供部分服务
- **MPP数据库扩展中支持的模式**
  - 只读
  - 读 + 一定的DML
  - 读 + load
- **关键点**
  - 底层存储方式
  - 是否可使用副本
  - 数据Hash 方式

## 资源管理技术

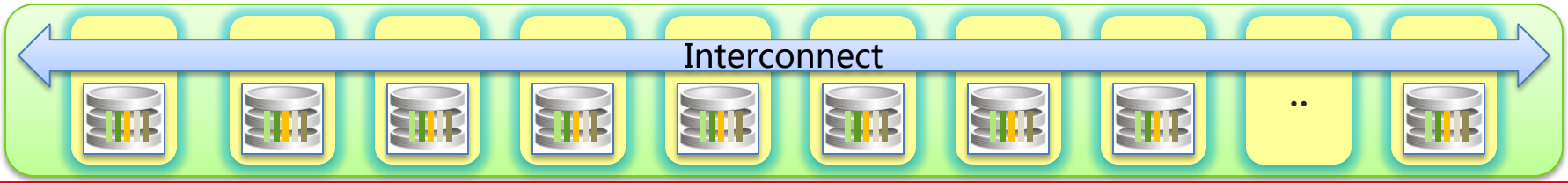
- **对于MPP，资源管理是一项必不可少的功能**
  - 内存
  - I/O
  - CPU
  - 网络
- **资源管理和优先级管理**
  - 需要一个动态可靠的资源评估机制
  - 任务需要排队
  - 内存的统一管理
  - 线程池的统一管理
  - 空间管理

# MPP数据库集群架构技术

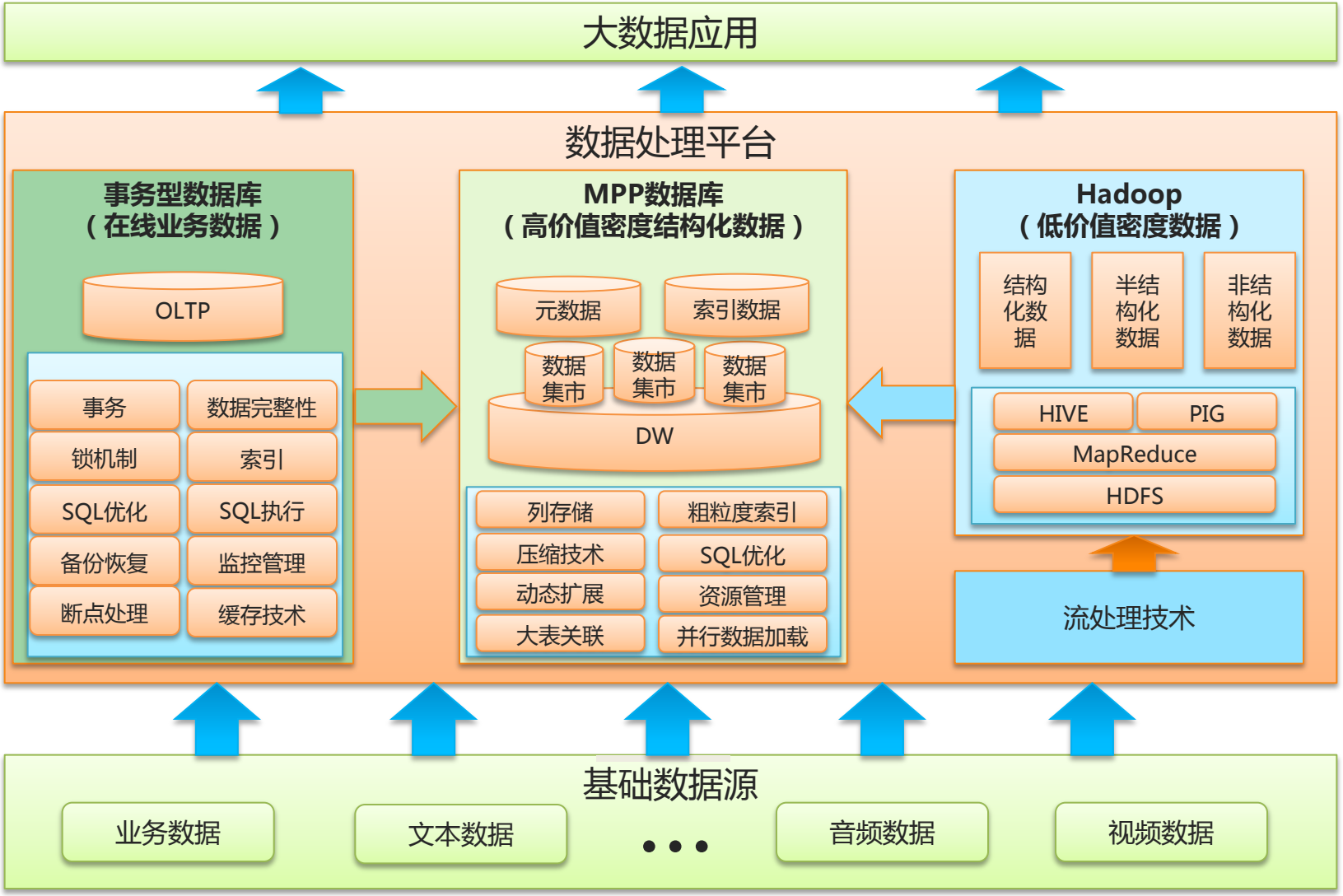


# MPP数据库应用方向

| | |
|---|---|
| 海量数据查询，统计、分析 | 互联网、移动互联网、金融、电信、物联网等：<br>PB支撑能力 – 海量数据边入库边使用 |
| 数据仓库支撑 | ODS，EDW，DM：<br>百TB支撑能力 – 千亿行多表join |
| ROLAP Cube | 基于星形、雪花模型的多维分析：<br>TB支撑能力 - TB级别的CUBE实时钻取 |
| 即席查询、统计分析 | 基于任何字段组合的随机查询、统计：<br>百TB支撑能力 – 几百列的宽表任意组合查询 |



# 大数据应用未来趋势——混搭架构



# 目录

# 电信行业大数据需求分析

**电信行业大数据，主要面临数据规模大、关联查询复杂、即席查询多三方面的问题。**

| | |
|---|---|
| 数据规模大 | • 中移动、中联通、中电信三大运营商，数据量均达到几十PB规模<br>• 数据管理支撑依靠硬件扩容，成本巨大 |
| 关联查询复杂 | • 结构化数据的复杂关联处理<br>• 结构化、非结构化的全数据关联分析<br>• 对复杂的任务调度进行有效管理 |
| 即席查询多 | • 临时性的统计分析，即席需求无法预知，预计80%以上为即席查询<br>• 数据仓库的混合负载管理效率低 |

# MPP数据库电信行业大数据应用场景

## 运营商

- 中国移动
- 中国联通
- 中国电信

## 业务类型

- 经分类
- 综分（信令监测）类
- 账单详单类
- 日志查询分析类

## 集群规模

- 百节点
- 10TB – PB

## 扩展能力

- PB级数据量
- 支持在线动态扩容

# 电信运营商某省云经分—项目背景

**传统经分系统面临的问题**

- **数据规模增长快速**
  - 现网用户量已达7200万，是中移动数据量最多的省分之一
  - 日均数据量超过2.3T，月数据50T，数据总规模超过500T，并快速增长，正向PB级迈进
- **现有系统响应慢**
  - 各业务部门业务需求不断增长，但因系统资源紧张和响应较慢，业务需求的支撑压力较大
- **流量分析不够精细化**
  - 流量分析相关的WAP日志、位置数据、数据业务话单、WLAN话单等并未实现有效整合和综合分析
  - 流量分析未到达事件级，无法实现精细化支撑的目标

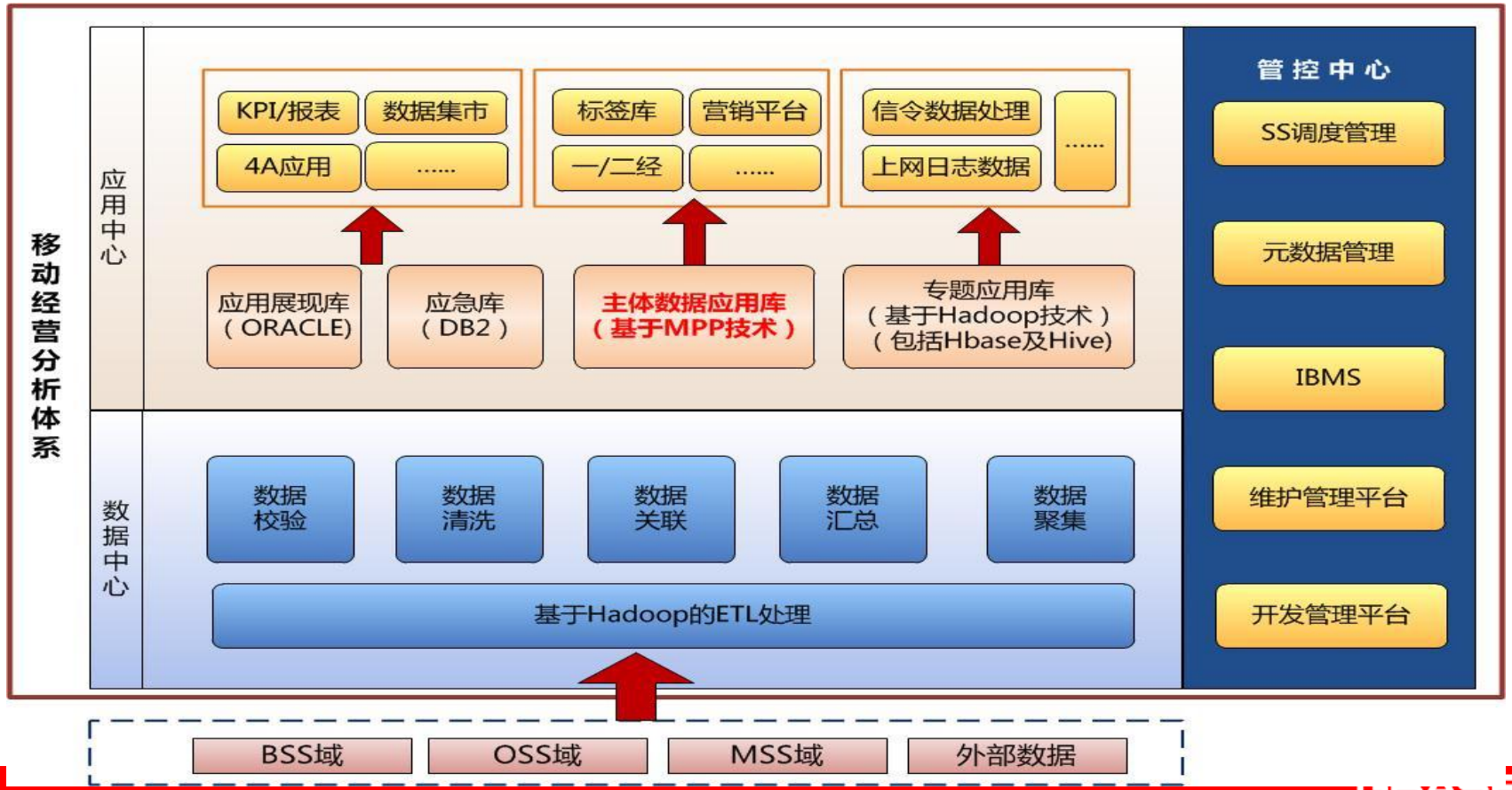
**项目建设内容**

- 去IOE，建设基于开放式、低成本的X86 PC Server集群云化架构
- 搭建基于ETL的数据分发平台和MPP集群数据库
- 以专题为单位将应用迁移到新的数据仓库平台

**构建大BI架构下“低成本，高效益，高性能”的云平台，支撑精细化运营管理和实时精确营销需求**

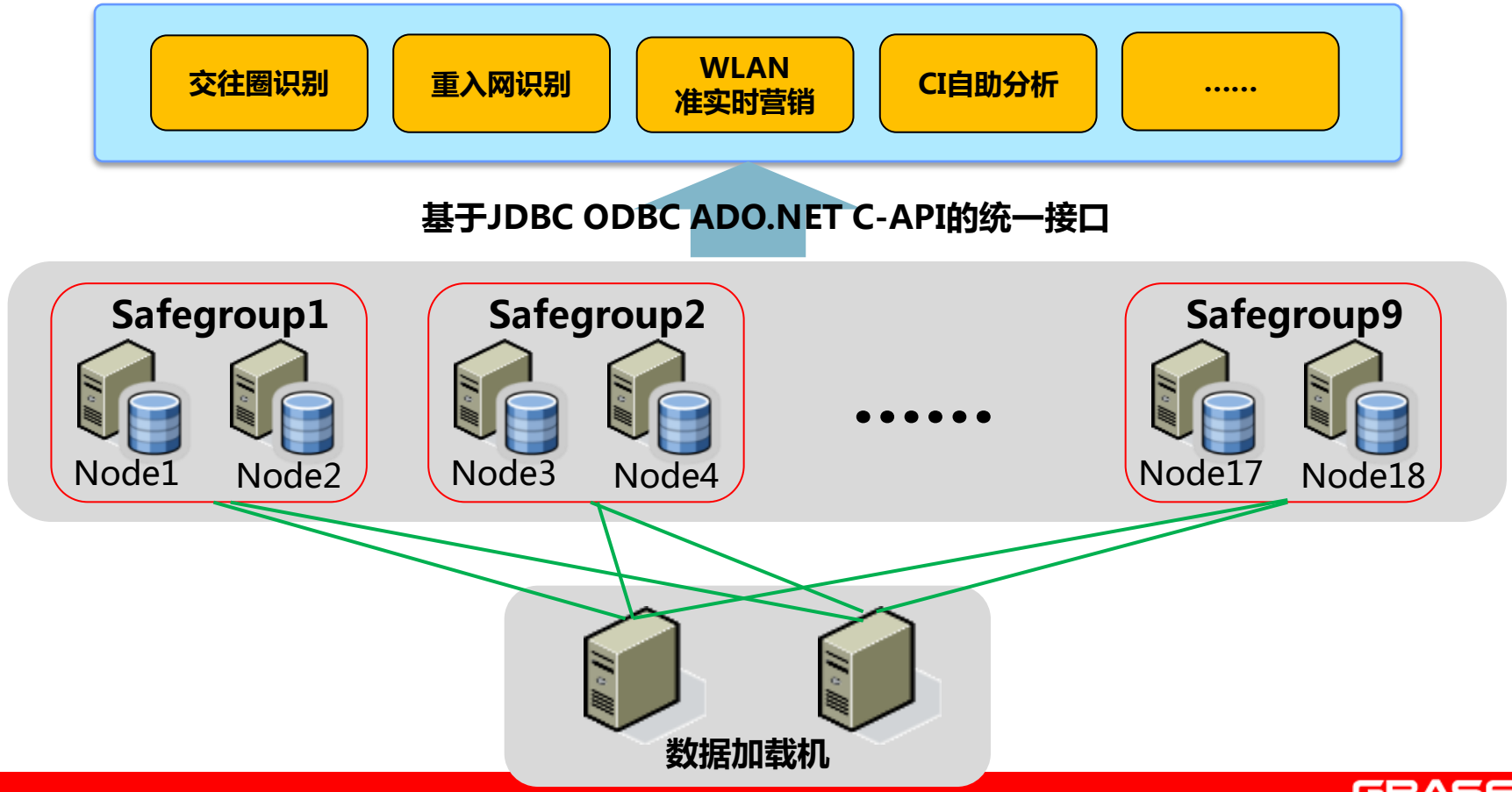
# 电信运营商某省云经分—总体架构

- 经分系统整体分为应用中心和数据中心两部分，应用中心采用MPP分布式并行数据库作为实现主体，数据中心基于Hadoop的HDFS完成统一数据存储。
- MPP数据库承载31220张表， 每天执行946条TCL程序，日均数据量超过2.3TB。
- MPP数据库集群承载KPI/报表生成、数据集市、营销平台、信令数据处理、上网日志分析等，应用涉及B域和O域的分析应用。



# 电信运营商某省云经分一解决方案

- 原系统使用3台IBM高端小型机+DB2，新系统使用18台x86服务器+MPP数据库集群
- 新系统整体成本降为原系统的1/30，但性能基本相当（日报执行时间都约为10小时）
- 系统拥有9个安全组，组内2个节点互为备份，节点发生故障仍能提供服务，具备高可用性



# MPP数据库应用于电信运营商某省云经分的价值

## 1 低成本、高性能

- MPP数据库运行于低成本X86 PC服务器，有效节省硬件投入成本
- 原系统投入3000万，新系统仅花费100万，新系统整体成本降为原来的**1/30**，大大节省项目投资，且性能与原系统相当

## 2 支撑海量数据、动态扩展

- MPP集群数据库能够有效处理PB级数据
- 将原有服务器的垂直纵向扩展模式改为依据数据量的水平横向扩展模式，动态扩展无须停止服务，保证服务连贯性

## 3 高可用性

- 通过合理配置能够有效实现均衡负载，充分发挥每一个节点的计算能力，提升整个系统的协同效率
- 基于安全组的备份策略，能够保证节点在发生故障时，不影响系统对外提供服务的连续性

## 4 实现深度精细化业务分析

- 高效的数据分析能力帮助客户应对复杂性强、效率及实时性要求高的场景
- 有效管理海量数据，实现对各类数据的多维深入分析，准确挖掘数据价值
- 已帮助客户实现交往圈和重入网识别、WLAN准实时营销、CI自助分析等主题应用场景

# 目录

# 金融行业大数据面临的问题

| | |
|---|---|
| 数据规模大 | • 四大行的数据量均达到PB级规模 |
| 数据存储成本高 | • 采用国外产品受制于人 |
| 关联分析复杂 | • 跨业务领域分析设计多表关联<br>• 对复杂的任务调度进行有效管理 |
| 时间窗口紧 | • 日常数据加工时间窗口紧张 |

# MPP数据库金融行业大数据应用场景

| 客户 | 业务类型 | 集群规模 | 扩展能力 |
| --- | --- | --- | --- |
| 国有银行 | 数据中心 | 百节点 | •PB级数据量<br>•支持在线动态扩容 |
| 股份制银行 | 审计风控 | 10TB – PB |  |
| 地方商业银行 | 日志管理 | | |
| | 报表支撑 | | |

# 某国有大行数据仓库平台—项目背景

**某国有大行统计分析系统面临的问题**

- **原系统使用Sybase IQ遇到了性能瓶颈**
  - 无法及时完成数据加载和处理
  - 无法支撑更多的分析和访问
- **需求无法及时满足**
  - 行内对信用卡、贷记卡、网络银行、反洗钱、监管报表、内部审计等应用都有迫切的数据分析需求，却得不到及时响应
- **核心系统逐步完善，宝贵的历史数据亟需保存利用**
- **国内重要的商业银行都已经建成了数据仓库系统，并从中不断获取以往无法了解的信息**

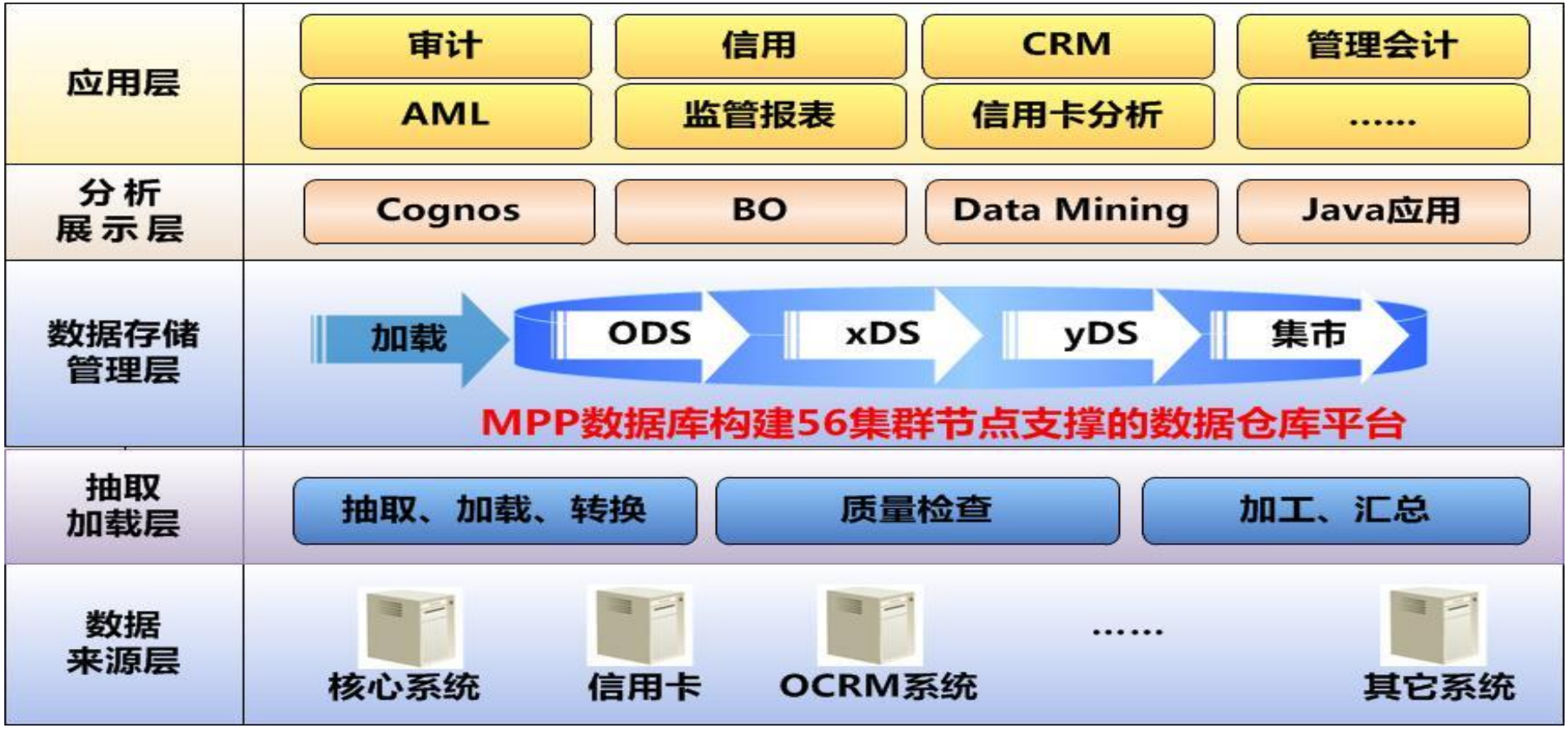
**项目建设内容**

- 以较低的成本，建设能够支撑500TB数据量的数据仓库平台，并具有不断扩容的能力
- 能够适应未来大数据平台方案的规划
- 避免其他行数据仓库建设中由于成本过高不可持续的问题

**构建“低成本，高性能，云架构”的数据仓库平台，支撑500TB数据高性能处理和大数据平台未来规划需求**

# 某国有大行数据仓库平台—总体架构

- 使用MPP数据库搭建56节点集群环境，承载500TB数据量，提供高性能分析查询应用。
- 目前系统每天处理4000个以上的复杂作业，每天数据增量大概为每节点800G。系统最大表已经超过1000亿行，该表数据增量为每天10亿。
- 支撑包括贷记卡、信用卡、网络银行、资金转移定价应用、零售业务、反洗钱、监管报表、个人金融、征信和风险等50多个数据源系统。



# MPP数据库应用于某国有大行数据仓库平台的价值

## 1 海量数据处理

- MPP数据库集群为用户提供了性价比很高的海量并行复杂数据处理平台，帮助客户形成500TB的业务数据单一视图，为客户提供及时高效的数据分析结果。

## 2 高性能

- 系统架构高可扩展，性能随着节点数的增加而提升，保证客户接入更全面的业务数据，满足市场营销、内部管理、内外监管的分析需求。

## 3 高压缩比

- 为用户提供完备压缩态存储管理海量数据的能力，进一步降低客户数据仓库建设的成本，并进一步提升系统性能。

## 4 国产化，去IOE

- 某国有大行是四大国有银行中首家在金融核心系统中完成国产化替代的银行。
- 该行数据仓库项目完成构建国产软件和PC服务器的大数据平台，实现国产金融大数据平台架构方案的成功落地。
- 使用国产数据库产品，代码自主可控，充分保证金融安全。

# 大数据技术创新大赛—期待你的参与

## 基于互联网大数据的日志类应用处理

- 时间：2014年8月20日—2014年12月13日
- 任务：基于互联网日志数据，开发出满足性能要求的数据库索引算法
- 奖项：一等奖10万元，二等奖3万元，三等奖1万元
- 详情：http://bigdatacontest.ccf.org.cn/gindex.html

# 世界级的市场必将培育出世界级的产品

# 谢谢

GBASE


天津：中国天津华苑产业区海泰发展六道6号海泰绿色产业基地J座
电　　话：022-58815678　　传　　真：022-58815679
北京：北京市朝阳区太阳宫中路9号太阳宫大厦10层1008室
电　　话：010-88866866　　传　　真：010-88864556
网　　址：http://www.gbase.cn　E-mail: info@gbase.cn
技术支持热线：400-817-9696