

大数据的实时分析与应用案例分享



上海云人信息科技有限公司

个人简介

- ▶ 吴朱华，专注于云计算和大数据这两个方向，之前曾在 IBM 中国研究院参与过多款云计算操作系统的开发工作，包括 PureSystem 的原型机，同济本科，并曾在北京大学读过硕士，在 2010 年底组建上海云人科技团队，在 2011 年中发表业界最好的两本云计算书之一《云计算核心技术剖析》，在 2013 年的 3 月被福布斯评为中国 30 岁以下 30 位创业者。

《云计算核心技术剖析》

云计算核心技术剖析

“本书深入浅出，详细分析了云计算各个层面的核心技术。我强烈向广大工程技术人员以及在校相关专业的研究生和高级本科生推荐这本书。”

——陈怀德，“弯曲评论”创办人和首席科学家

“吴朱华的书绝对是迄今国内出版的云计算图书中最专业的一本。”

——阿朱，《走出软件作坊》作者

云计算是新一代IT计算模式，它运用先进的分布式计算及存储架构为用户提供方便、安全的体验并降低使用成本。本书首先介绍了云计算理论方面的知识，接着剖析了多个顶尖云计算产品（比如Google App Engine和Salesforce Force.com）的实现，介绍了非常重要的系统虚拟化技术和安全方面的机制，然后以云的核心模块之一——分布式数据库为实践方向，并以YunTable这个云时代的BigTable为例，演示了如何手动编写和设计一个分布式数据库，最后对云计算的未来发展做了展望。



吴朱华

曾在IBM中国研究院参与过多个云计算产品的开发工作，人云科技信息技术有限公司（<http://peopleyun.com>）创始人，官方微博为<http://t.sina.com.cn/peopleyun>，专注于YunTable和YunEngine研发。

图灵网站：www.turingbook.com 热线：(010)51095186转604
反馈/投稿/推荐信箱：contact@turingbook.com
有奖勘误：debug@turingbook.com

分类建议 计算机/程序设计/云计算

人民邮电出版社网址：www.ptpress.com.cn



ISBN 978-7-115-25219-7
定价：49.00元

TURING

TURING 图灵程序设计丛书

云计算核心技术剖析

吴朱华 编著

人民邮电出版社

云计算 核心技术剖析

吴朱华 编著

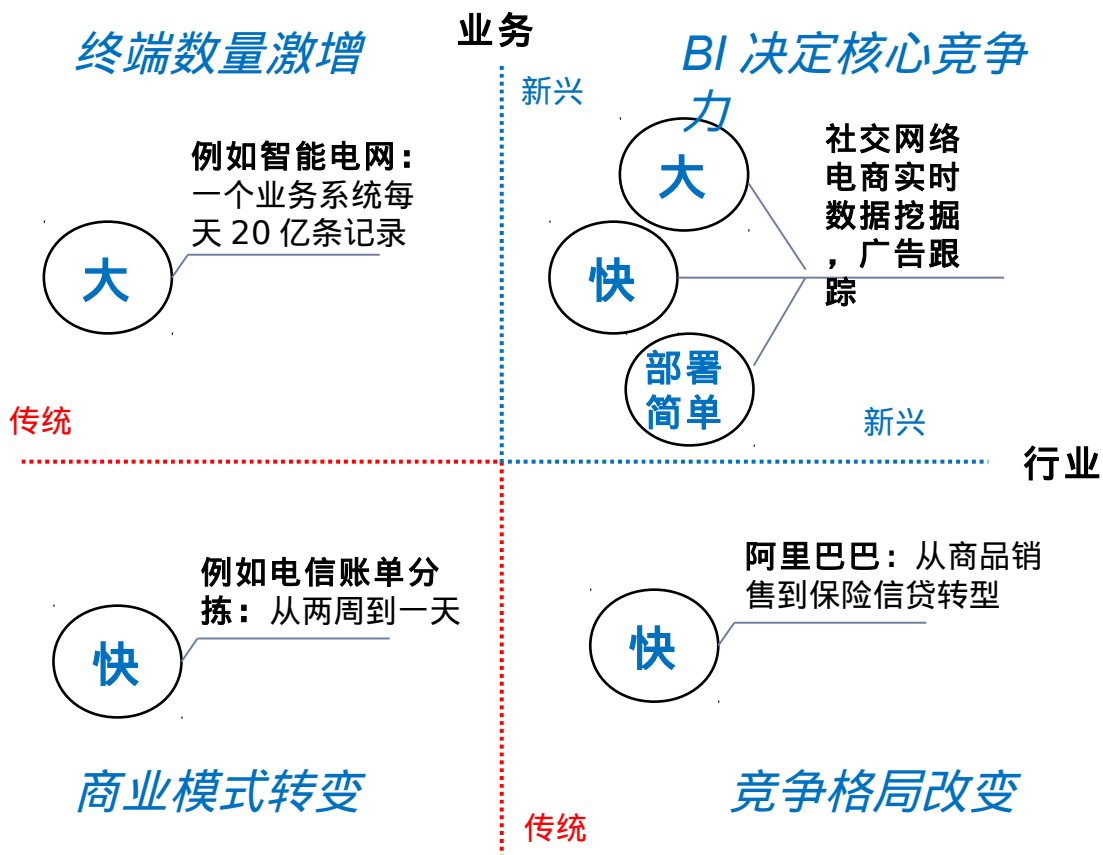
- 一线云计算专家倾情奉献
- 深入分析多种云计算核心技术
- 通过实例教你编写简化版的BigTable

人民邮电出版社
POSTS & TELECOM PRESS

大数据的时代

- ▶ 来自麦肯锡的报告，未来的 10 年里，数据和内容将增长 **44 倍**，并且这些数据有**无法估量的价值**；
- ▶ 对于很多以数据为资产的行业，BI 决定公司的核心竞争力。比如互联网广告，金融机构，大数据实时分析工具对他们而言，就等同于竞争武器，快或慢一秒钟，往往就意味着财富的得与失。
- ▶ 对传统行业来说，大数据的冲击来自三个方面：数据终端数量的增长，例如智能电网和物联网；数据维度的变化，例如消费行为与社交网络的关联；商业模式和管理模式的变化：例如从产品消费到信用营销，从经验和直觉决策到数据智能决策。三个因素组织在一起，使大数据发生了几何级数的增长。

大数据需求



大数据的阶段

- ▶ **第一个阶段**：自身业务需求产生大量数据，利用这些数据，通过深入证析，优化相关业务；
 - ▶ **第二个阶段**：搜集与目标业务直接或间接关联的大量异质数据，建立复杂的分析和预测模型，产生针对目标业务的输出；
 - ▶ **第三个阶段**：随着整体数据相关的法律不断补充，以及技术不断成熟，形成一个完善的数据生态，包括数据市场，数据运营商和数据商店等。
-
- ▶ 从技术角度而言，趋势是更实时，越快越好，更全面数据分析需求，包括 SQL、挖掘算法，以及以 Deep Learning 为代表机器学习技术。

大数据实时分析的目的

- ▶ 实时决策能力；
- ▶ 提高业务效率；
- ▶ 快速智能发现新观点和商业机会；
- ▶ 提供业务产出；
- ▶ 提升 IT 效率；

大数据头的分析所需的技术支撑

- ▶ 大数据秒级，甚至毫秒级的处理；
- ▶ 上千人的并发访问；
- ▶ 支持 SQL 标准，特别是 OLAP 相关的语句；
- ▶ 数据的安全和集群的稳定型；

大数据实时分析的技术选型

- ▶ Hadoop 系列： Hive ， Impala ；
- ▶ NoSQL 类别： MongoDB ， HBase ；
- ▶ 传统关系型数据库： Oracle ， DB2 ， MySQL ；
- ▶ 传统列式数据库： Infobright ， Monet DB ；
- ▶ 新一代基于内存计算的数据库？

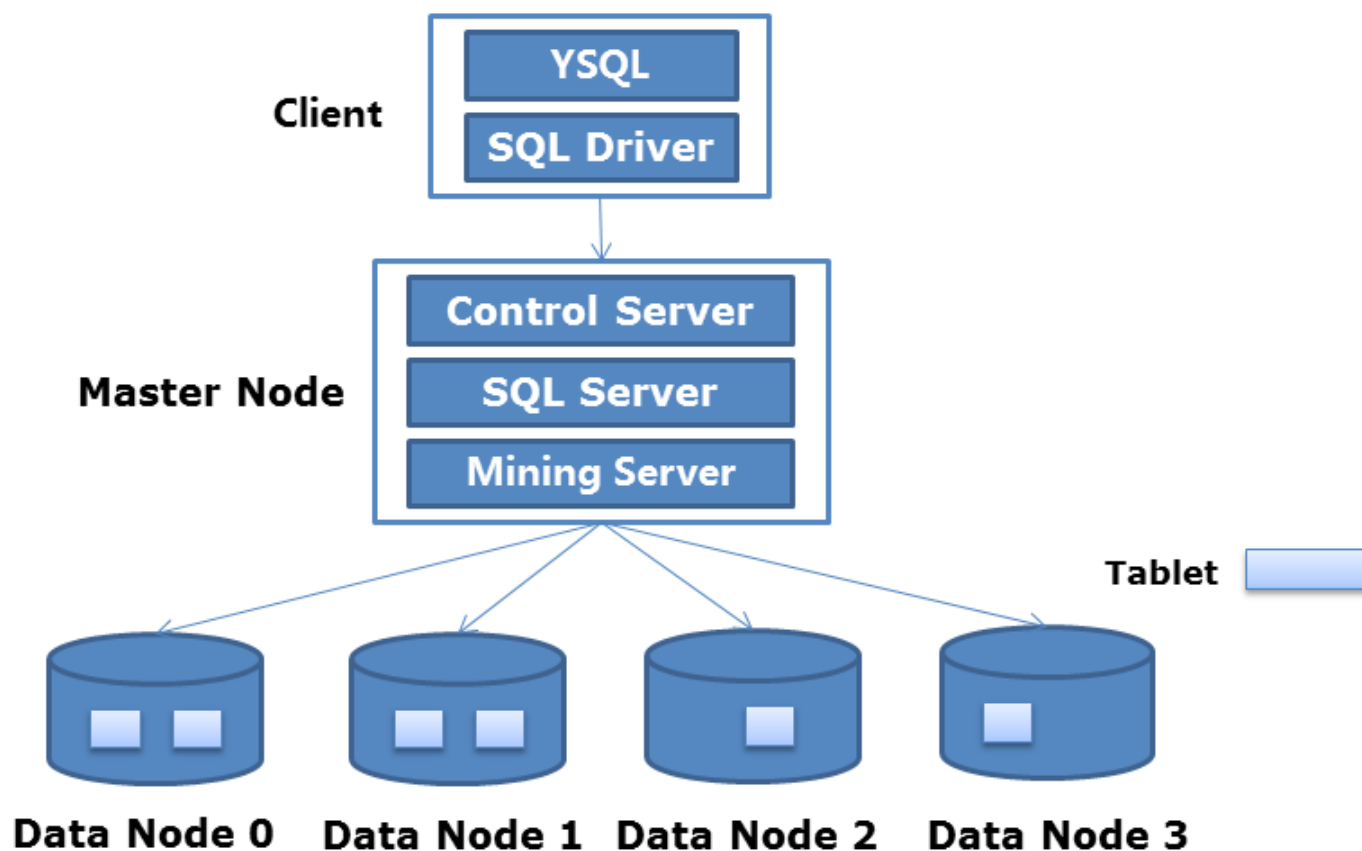
技术选型的对比图

	秒级处理	并发	SQL 支持	安全和稳定
Hadoop	No	Depends	Depends	Yes
NoSQL	Yes	Yes	Depends	Depends
传统关系型数据库	Depends	Yes	Yes	Yes
传统列式数据库	Yes	Depends	Yes	Depends
基于内存技术的新一代数据库	?	?	?	?

YunTable

- ▶ YunTable 是在从分布式数据库的基础上发展而来，同时加入一些 NoSQL 的基因的新一代大数据实时分析数据库，并且支持内存计算，比较接近 SAP HANA。

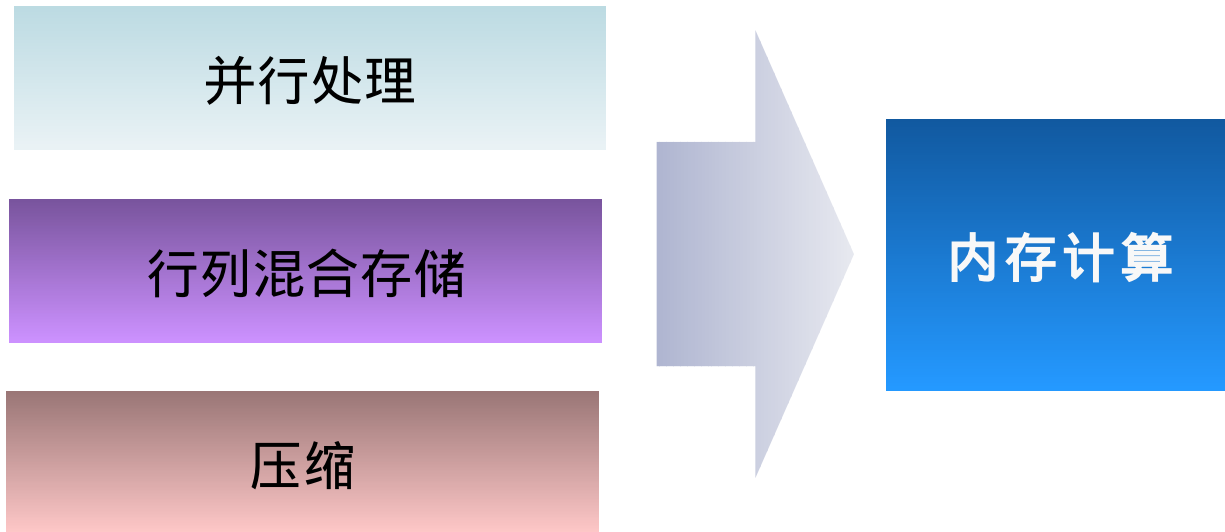
系统架构



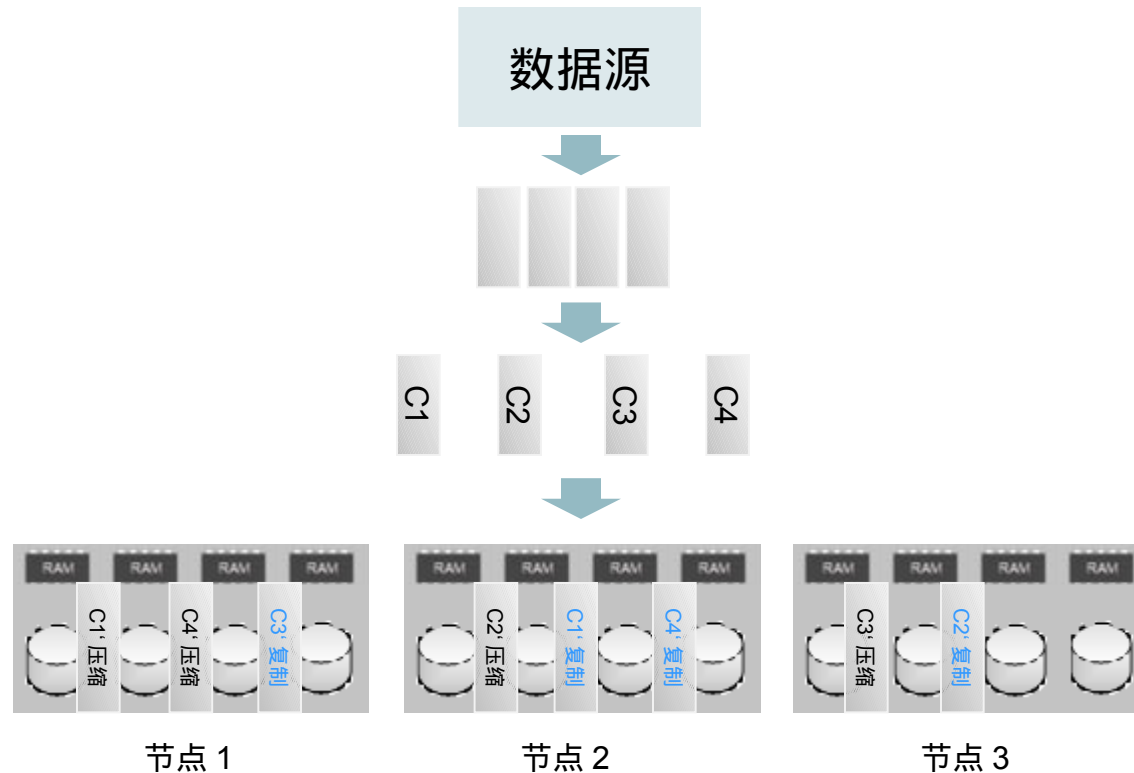
核心特性

- ▶ 大数据，秒级内存计算；
- ▶ 采用廉价的 x86 硬件；
- ▶ 自动线性动态扩展至数百台集群；
- ▶ 每秒 GB 级别吞吐量，PB 级别存储量；
- ▶ SQL92 特性覆盖，并提供多平台的 SQL 驱动，还支持 R；

核心技术



核心技术（一）：并行处理



并行处理：数据复制分布存储在不同的节点上并行处理

内存本地化：把大数据量和计算量分散到不同处理器

高可用性：任何节点宕机将不影响数据完整和业务连续性

核心技术（二）：行列混合存储

- ▶ 行分区
 - ▶ 保留数据关联
- ▶ 列式数据组织
 - ▶ 高效的数据压缩
 - ▶ 快速的数据聚合
 - ▶ 优化的数据上传到中央处理器
- ▶ 专利的索引结构

核心技术（三）：高效压缩

- ▶ 多种无损压缩算法；
- ▶ 列式数据组织，整体压缩率高达 10~20 倍以上

核心技术（四）：内存计算

- ▶ 硬件性能的提升
 - ▶ X86 多核技术
 - ▶ 64 位地址空间 — 单台服务器内存容量可达 2 TB

- ▶ 软件技术创新
 - ▶ 行列混合存储
 - ▶ 高效压缩
 - ▶ 数据分片
 - ▶ 高效索引
 - ▶ 增量插入

硬件性能提升结合 YunTable 软件技术创新，使原来通过大量磁盘读写处理的海量数据，可以在服务器的主内存中实时处理，提供实时统计分析结果！

具体实时分析场景

目标市场



大数据资产

实时分析案例：互联网

主要业务应用： 电商交易分析，社交网络，位置信息服务，广告交易、跟踪分析等

典型用户： 某互联网广告公司广告投放效果实时监测

数据规模： 100 亿条记录

关系型数据库的问题： 不能满足 10 亿条以上记录的存储和查询要求

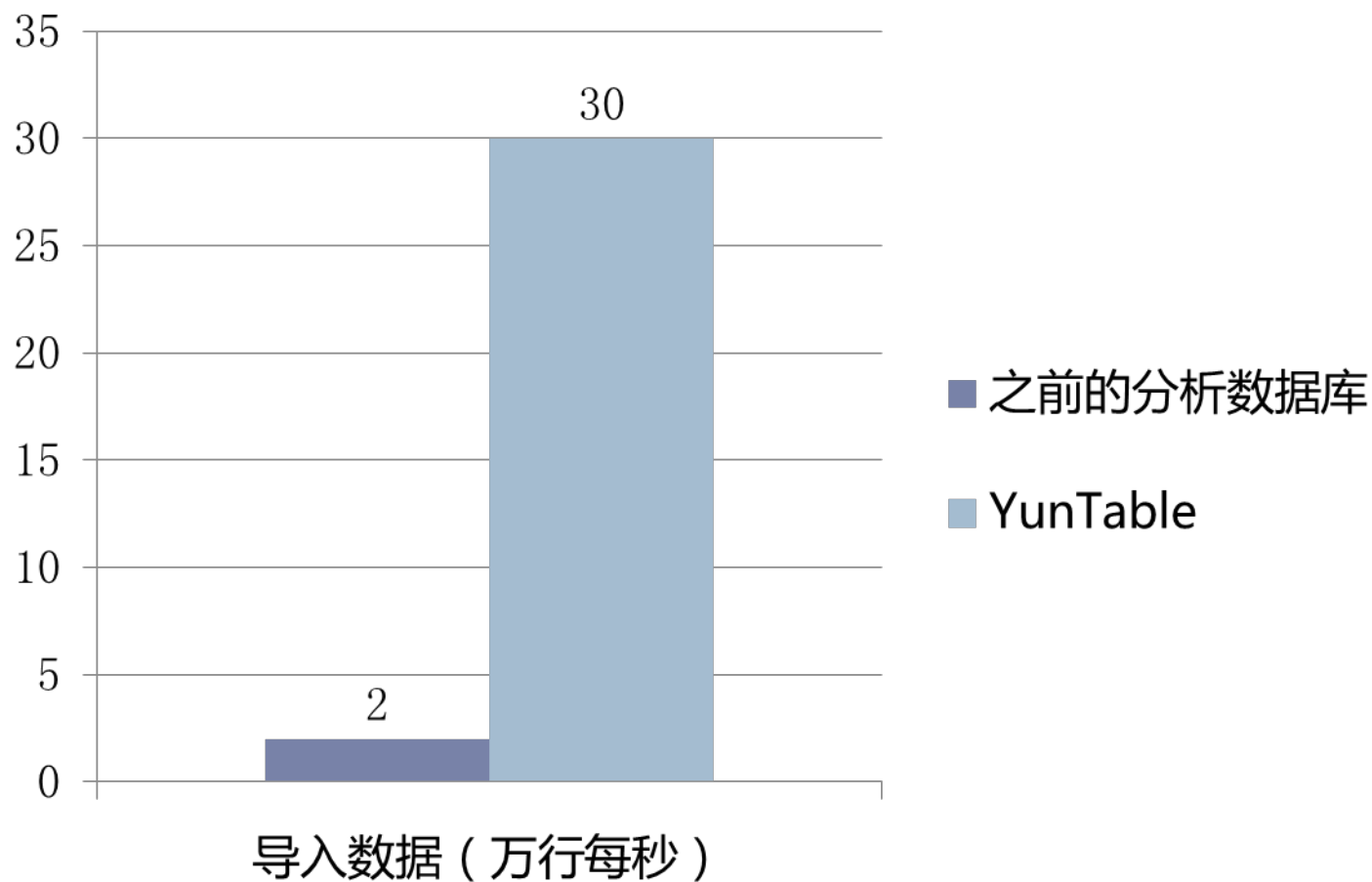
Hadoop 的问题： 不能满足结构化数据的存储和实时查询要求

解决方案：

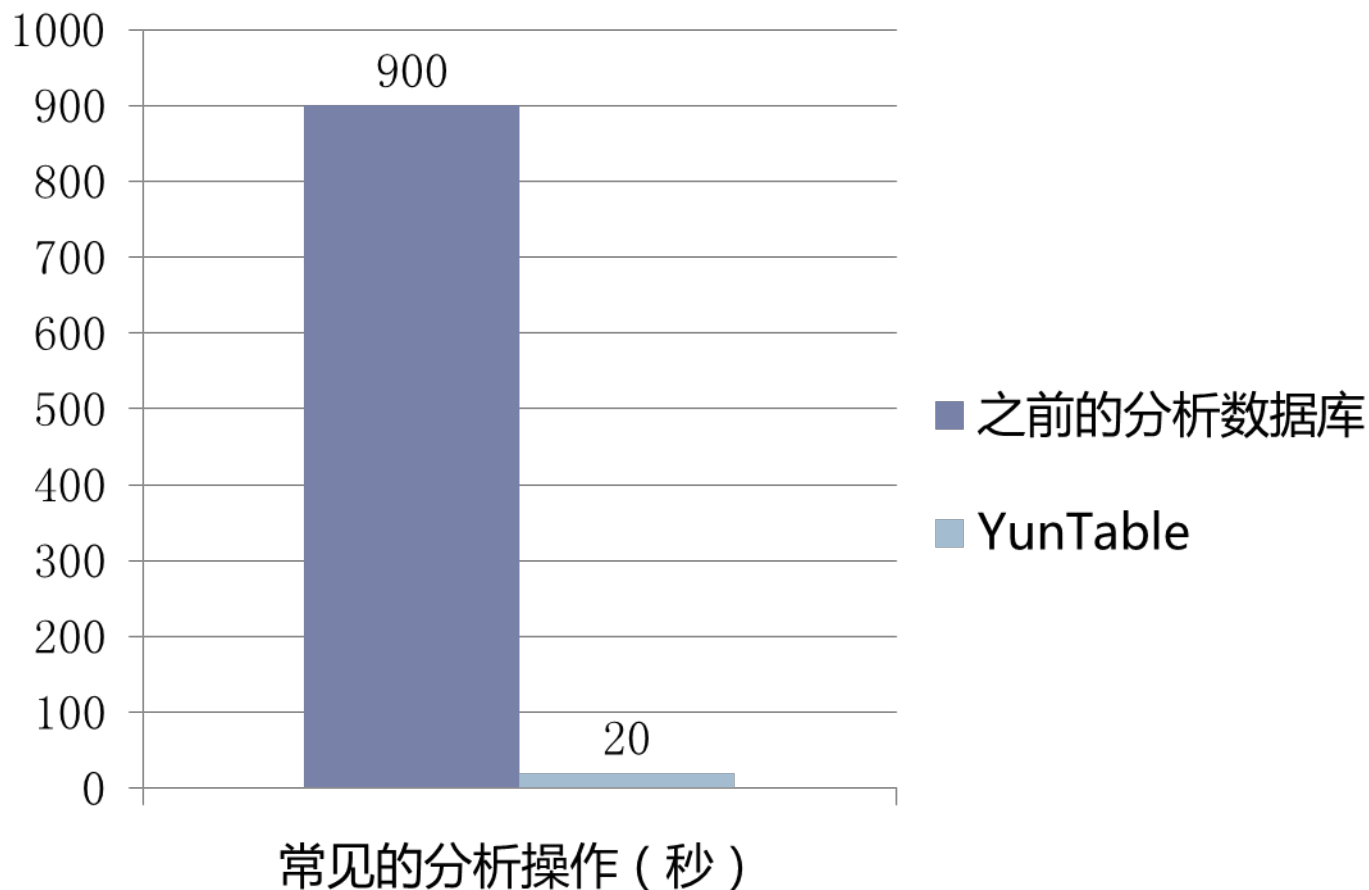
1、技术团队自行开发（例如淘宝，腾讯，新浪微博），优点：可以根据业务流程进行模型优化，获得良好性能；缺点：对技术团队开发水平和人员数量要求高，总体维护成本很高；

2、选用 Yuntable 和 Exadata、Hana、Greenplum，优点：使用和管理简单；缺点：后三家购置成本高，性价比不高

导入操作的性能比较



查询操作的性能比较



具体的性能测试结果

数据场景：2.3 亿条互联网用户访问记录数据

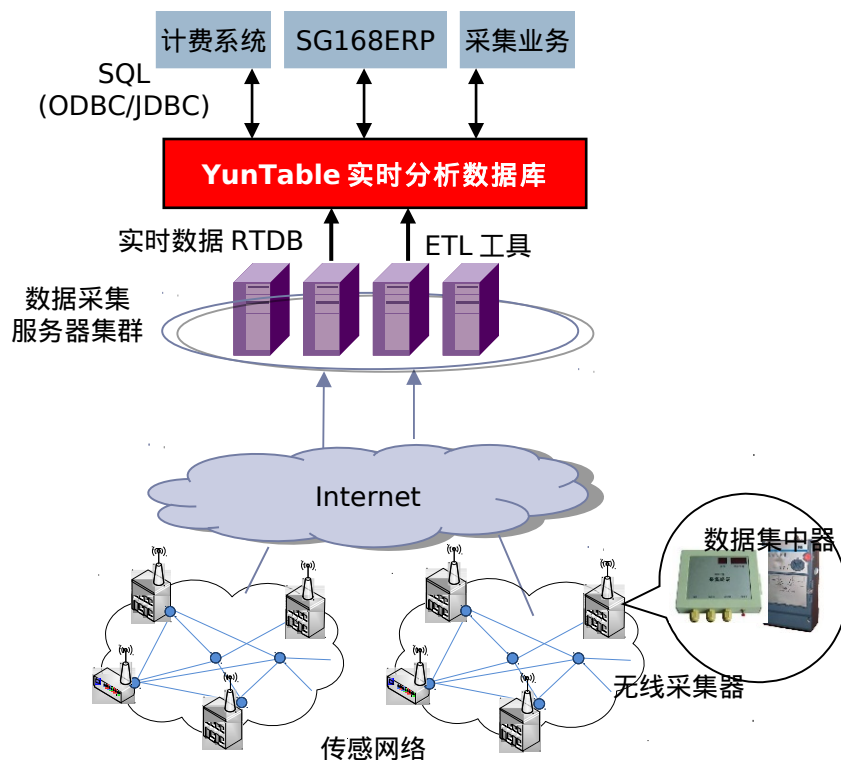
测试环境：YunTable 3台 4核 64G 内存 Dell 服务器

项目	YunTable 指标 (秒)
频次分析	9.492
重合度分析	16.625
多维度分析	11.408

实时分析案例：物联网

主要业务应用：海量数据终端信息采集与用户行为分析

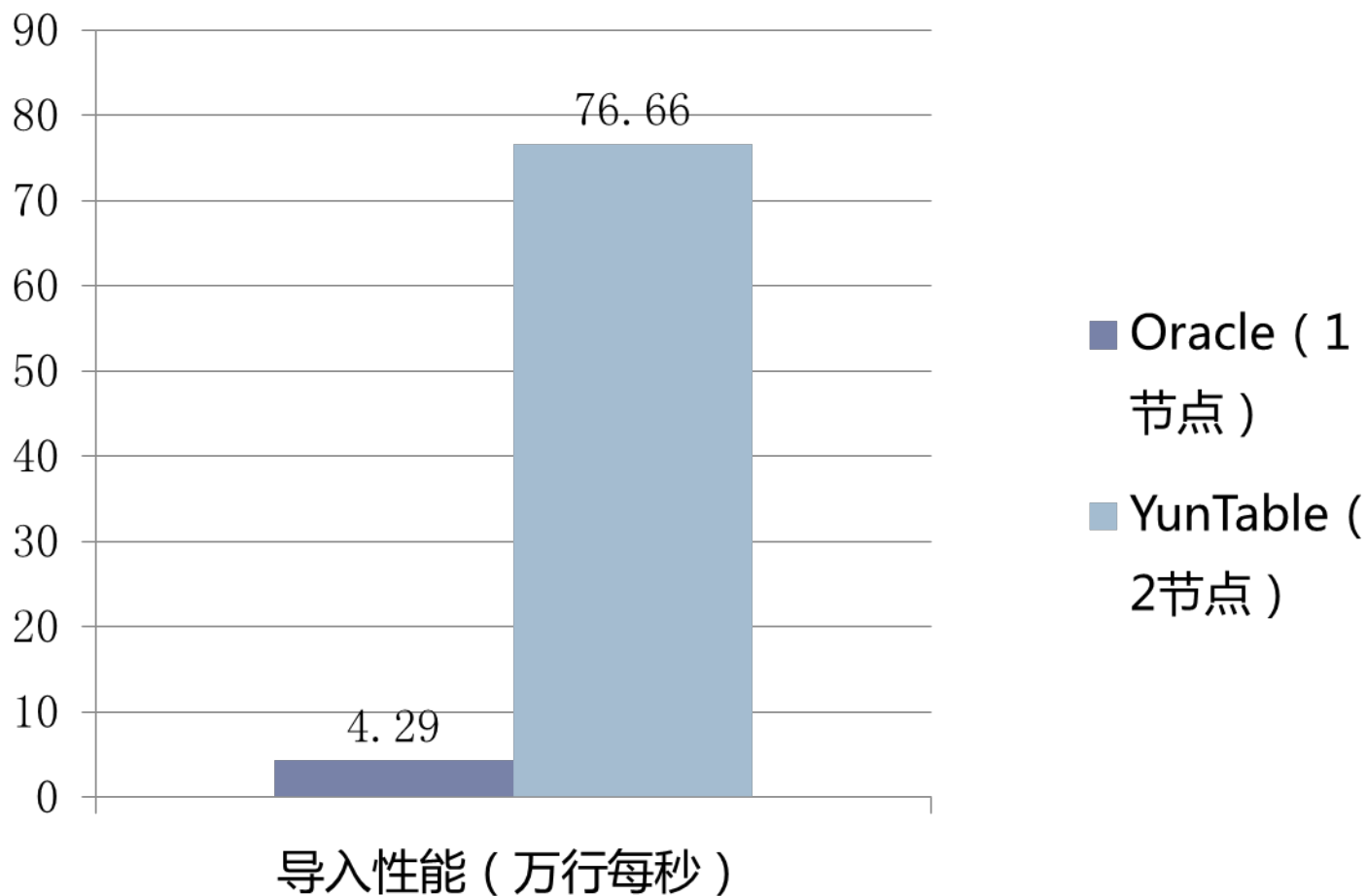
典型应用场景：智能电网用电信息采集（子系统）



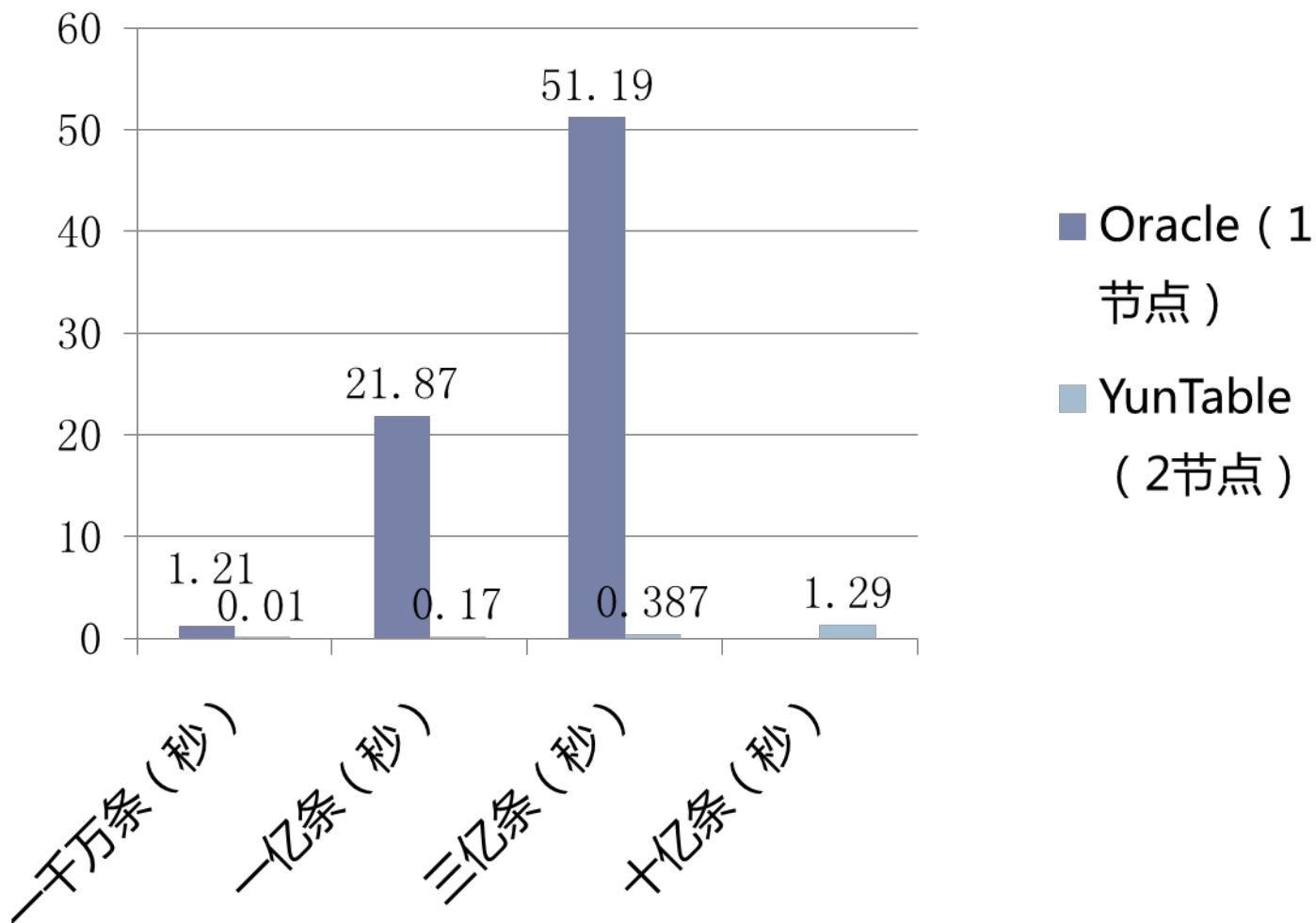
商业建设的案例 - 某物联网企业

2012 年底，我们团队参与了某核心企业大数据实验室的建设，并且建设过程中，我们在性能方面与 Oracle 数据库进行了正面的 PK。在本次 PK 中，我们无论在导入和分析等性能方面，**都远胜 Oracle**。

导入操作的性能比较



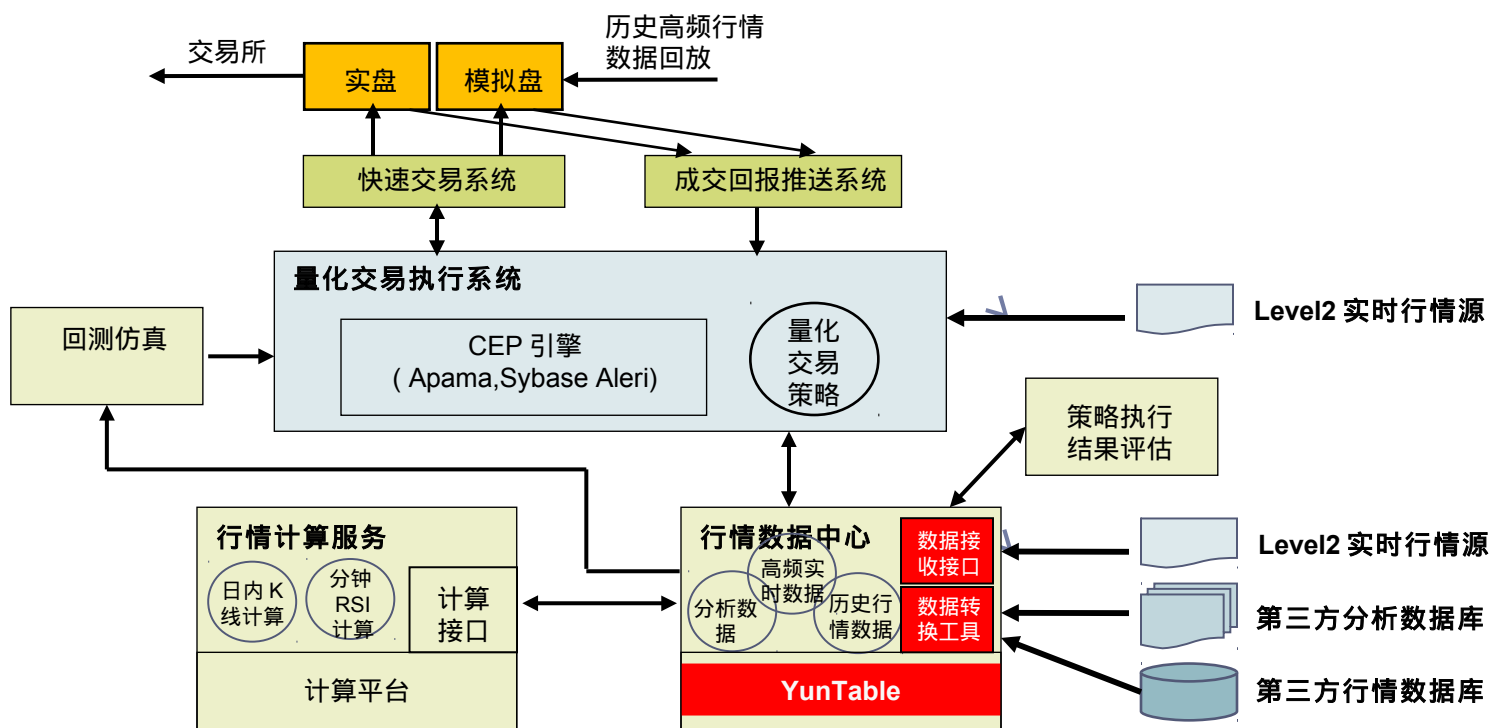
分析操作的性能比较



实时分析案例：金融

主要业务应用：量化交易，高频交易

典型场景：证券公司量化交易平台及各子系统



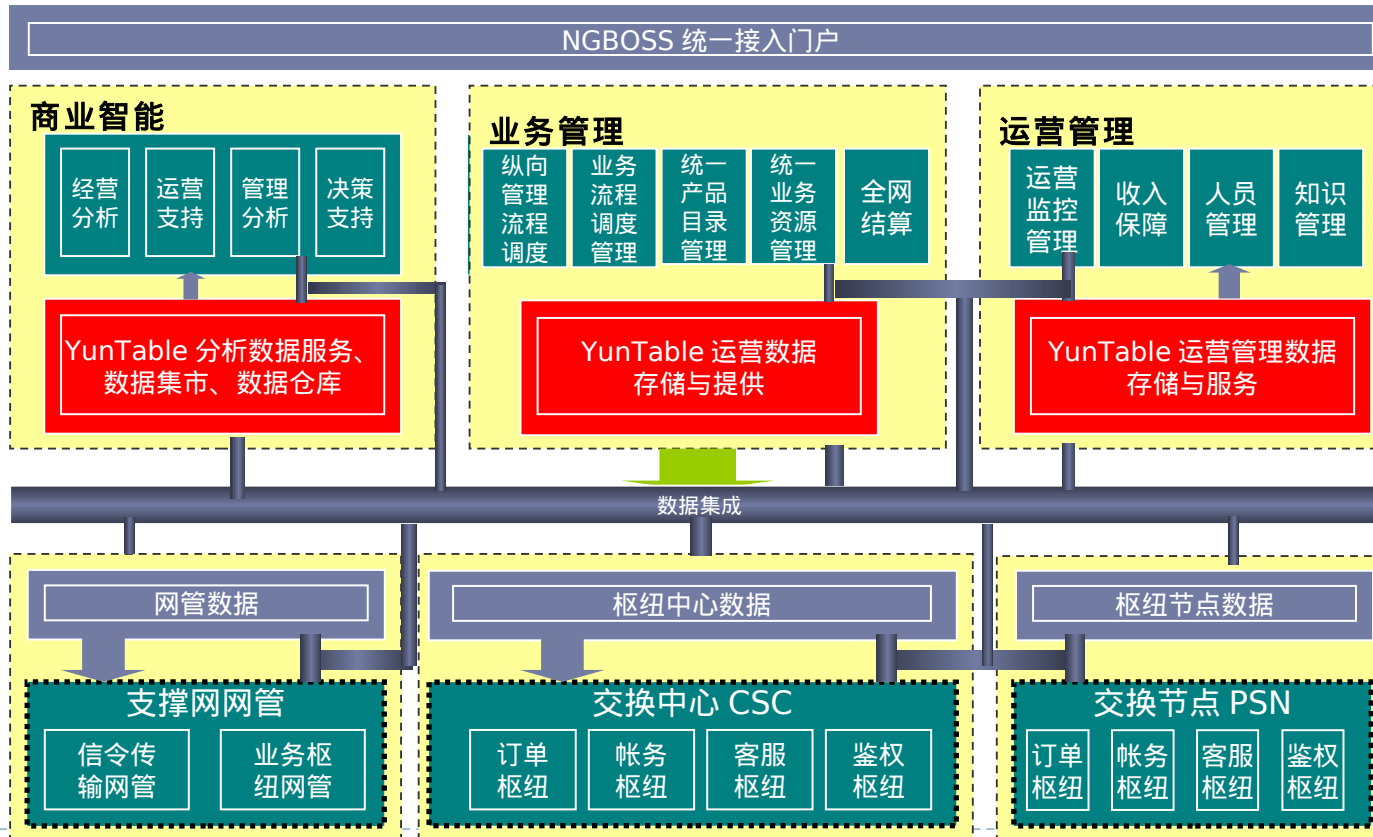
具体性能表现

	查询时间
单日业务数据统计	0.36 秒
单周业务数据统计	0.58 秒
单月业务数据统计	1.25 秒
单日股票代码汇总分析	2.27 秒
单日多列汇总分析	2.71 秒
单日账户汇总分析	4.43 秒
单月股票代码汇总分析	3.86 秒
单月多列汇总分析	5.09 秒
单月账户汇总分析	8.12 秒

实时分析案例：电信运营商

主要业务应用：BOSS/NGBOSS 系统及各子系统

典型应用场景：NGBOSS 业务运营支撑系统及各子系统



Q & A

Thank You !

