ORACLE

大数据解决方案

段建民： James.duan@oracle.com

2013.05.18

一、大数据特点
二、传统DW处理方式的挑战
三、Hadoop 技术简述
四、Oracle 面向大数据的集成解决方案

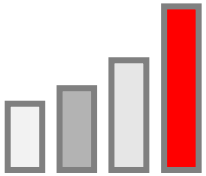
■ 以下内容仅供参考，不可纳入任何合同。该内容不构成提供任何材料, 代码或功能的承诺, 并且不应该作为制定购买决策的依据. 所描述的有关 Oracle 产品的任何特性或功能的开发、发布和时间安排均由 Oracle 自行决定。

一、大数据特点

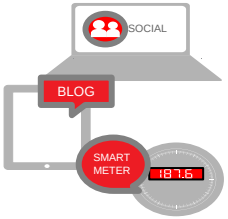
1. 大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合”—维基百科
2. 三大特征(3V)
   1. Volume：数量大（Twitter1.75亿用户每天创建9500万条微博；Facebook每天在30万台服务器上处理25Tb数据；YouTube每天上传168Tb视频）
   2. Velocity：时效性要求高（搜索引擎要求几分钟前的新闻能够被用户查询到）
   3. Variety：种类和来源多样化（结构化/半结构化/非结构化；关系数据库/数据仓库/互联网网页等）
3. 通常用于分析型的应用场景，如搜索引擎网页处理、用户行为分析、商业智能（BI）等

Oracle 对大数据的理解—4V特征

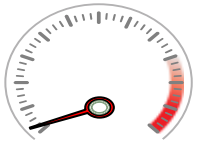
具有4V特性的数据称为大数据



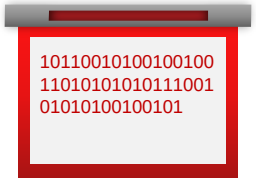
Volume



Variety



Velocity



Value

- **巨大的数据量 Volume**
  - 集中储存/集中计算已经无法处理巨大的数据量

新浪微博

3亿用户，每天上亿条微博

中型城市每月数十亿智能电表数据

2015年全球移动终端产生的数据量6300PB

- **多结构化数据 Variety**
  - 文本/图片/视频/文档等
- **增长速度很快 Velocity**
  - 海量数据的及时有效分析
  - 用户基数庞大/设备数量众多/实时海量/数据指数级别增长
- **价值密度低Value**
  - 单条数据并无太多价值，但庞大的数据量蕴含巨大财富

Why Oracle ?

Volume

Velocity

Variety

Unstructured Data

CORRELATE

VISUALIZE

Stream | Acquire | Organize

Oracle Exadata

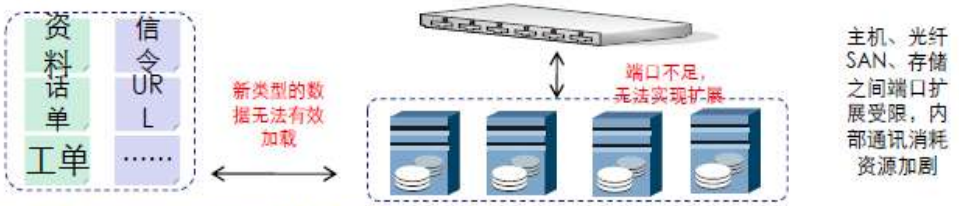
Oracle Exalytics

Value

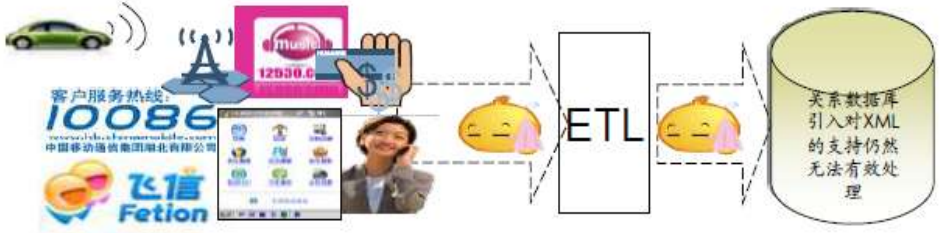
二、传统DW数据处理方式的挑战

海量数据的出现、数据结构的改变，对数据管理及分析带来挑战

- 传统数据仓库无法有效存储日益增长的业务数据
  - 随着业务发展数据量的增加，随着应用复杂导致的数据量增加，这些数据量导致了数据存储和处理压力； 数据仓库无法线性扩容，管理难度加大，成本高扩容压力大，效率下降等

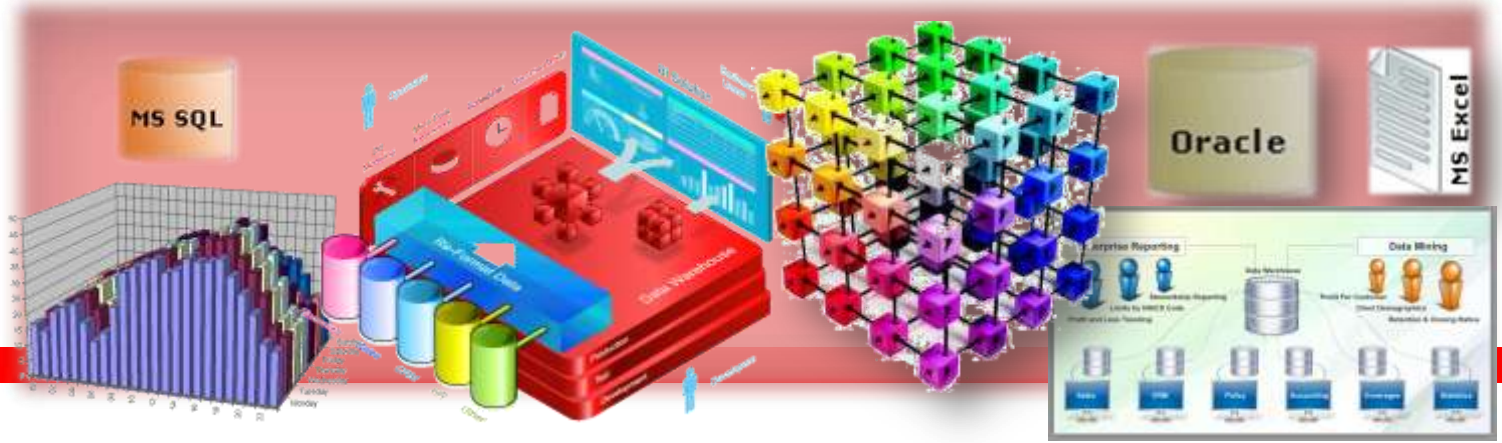


- 传统数据仓库无法有效处理新型的业务数据
  - 公司在移动互联网和物联网上需要有新领域的突破，不同于传统通信业务分析特点，需要对内容等非结构化、大容量信息进行有效分析，传统的架构处理吃力 ；



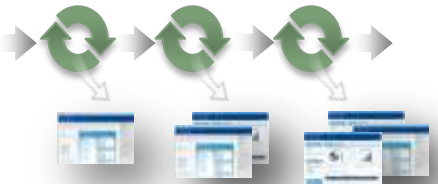
传统数据源 VS 新数据源

非结构化
半结构化
数据



结构化
数据

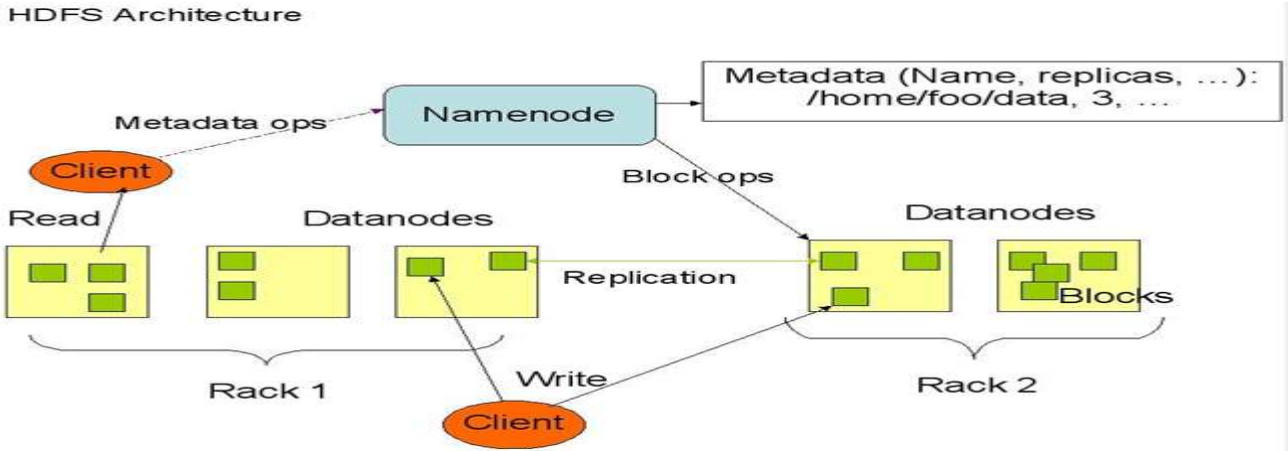
数据量、复杂的数据种类剧增带来新的挑战

更多的多样化数据		结构化 和非结构化的内外部数据快速增长
更多的变化和不确定性		预定义的模型、信息板和报告无法满足意外业务需求
更多的意外问题		能够根据需要以自助方式挖掘数据、添加新数据和构建分析

需要革新的技术手段 *Hadoop技术*

Hadoop几乎成为大数据处理的事实标准

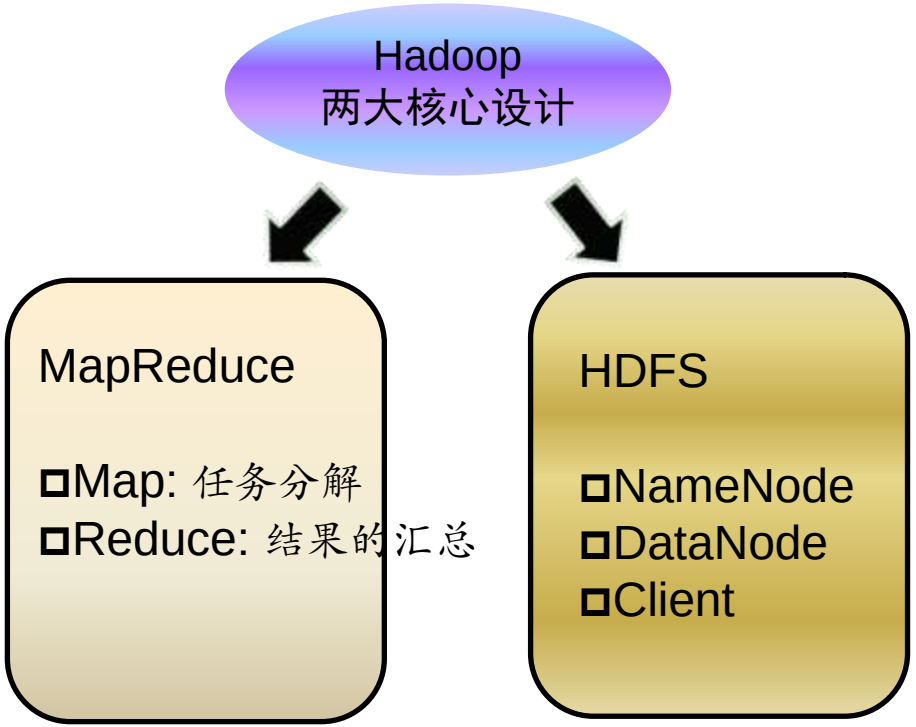
- 海量数据“分而治之”------批量分布式并行计算Hadoop
- 海量数据“灵活多变”------实时分布式高吞吐高并发数据存取处理NoSQL
- 海量数据“跨越鸿沟”------大数据超高速装载进数据库



Hadoop 包括两个部分：

1. HDFS
(Hadoop分布式文件系统)
Hadoop Distributed File System
2. MapReduce 的实现

三、Hadoop 技术简述

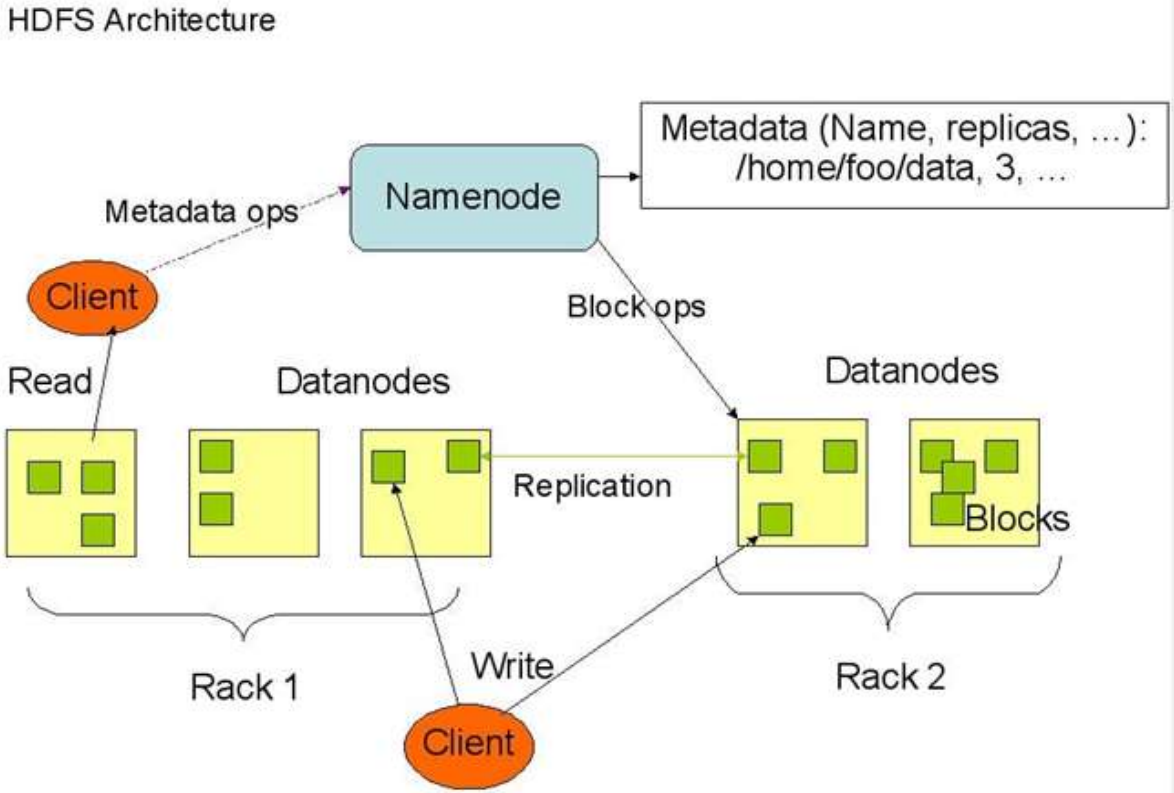


分布式文件系统

MapReduce 编程范式

高度可伸缩的数据处理能力

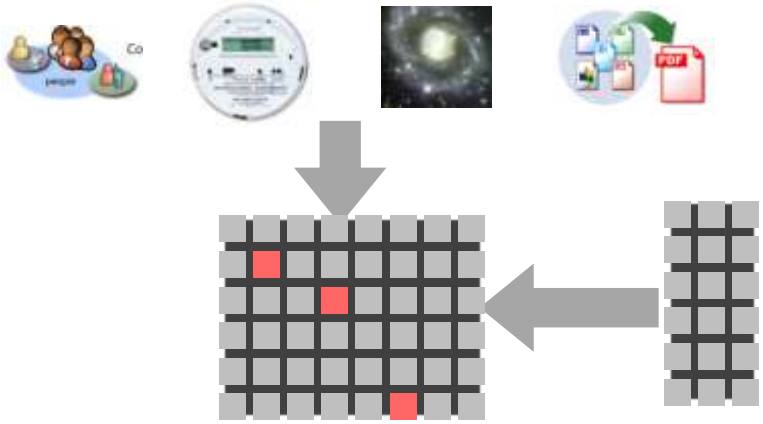
Hadoop 架构论述



Hadoop 包括两个部分：

1. HDFS (Hadoop分布式文件系统)
   Hadoop Distributed File System
2. MapReduce 的实现

HDFS 简述



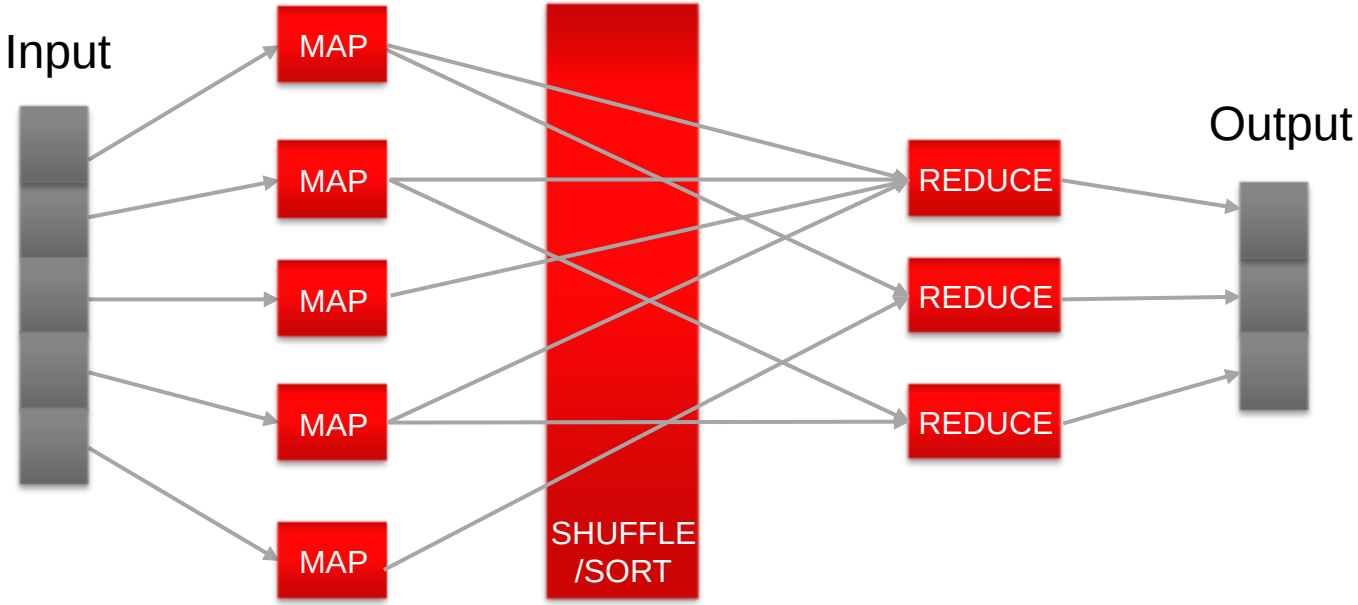
- 将数据分布在集群上
- 多个副本
- 通过添加节点实现扩展

HDFS 用例：

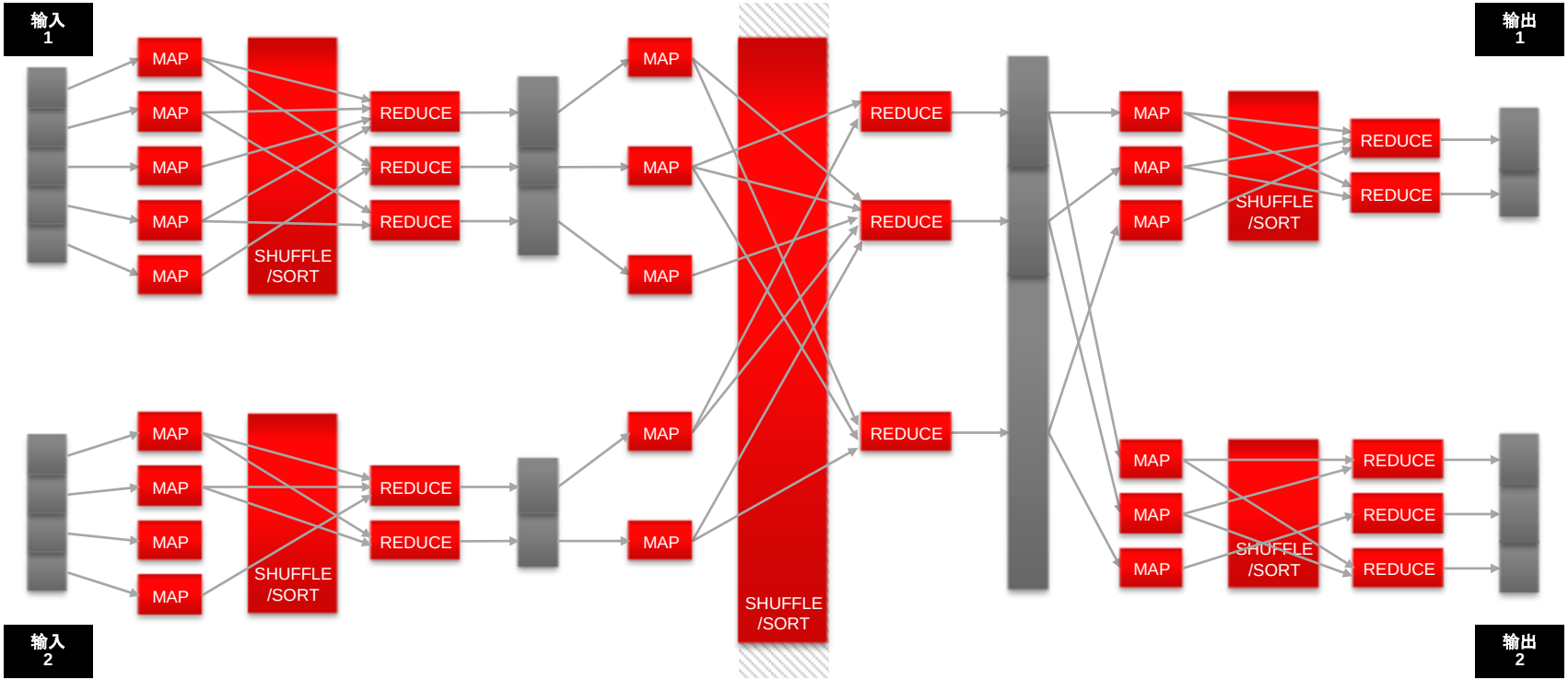
- 点击流存储和分析
  - 持续时间超过 X 分钟的 Web 会话数
  - 浏览频率最高/最低的页面
  - 按钟点和源位置进行会话时间分组
- 舆情分析
  - 多少个评论包含单词或词组
- 关系发现
  - 哪些项目看似在时间或相近性方面相关
  - X 和 Y 有多少次相近

MapReduce 的简单示例

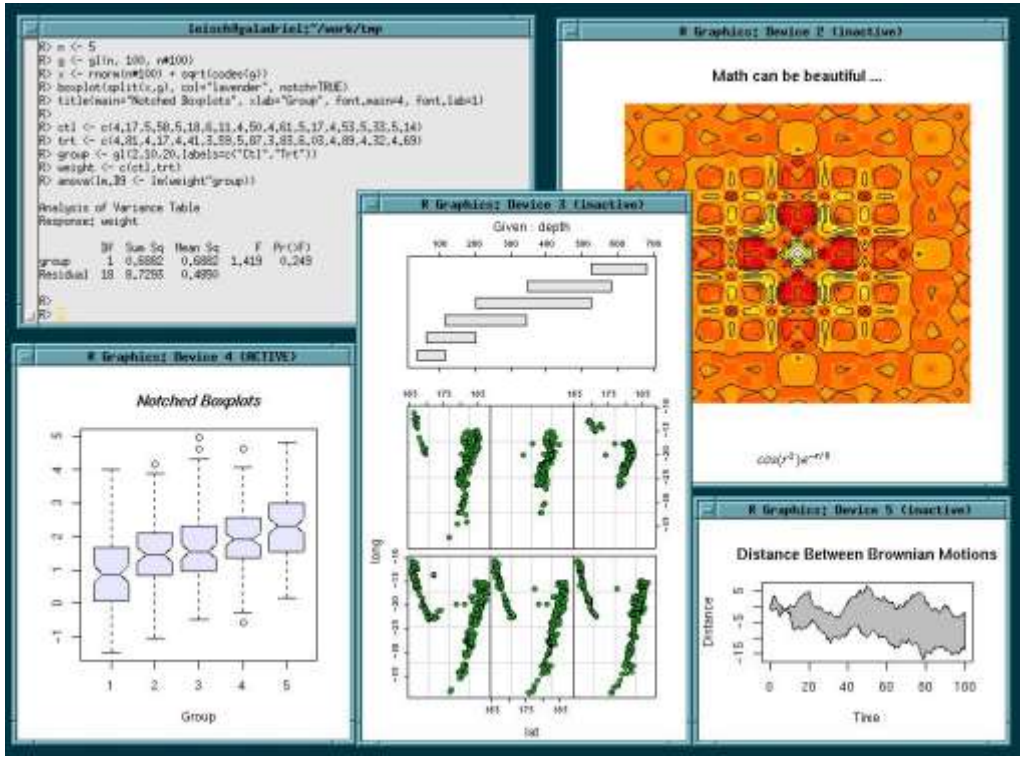
输入- Map – shuffle – Reduce - 输出



使用 Map/Reduce 扫描所有数据



R 统计编程语言



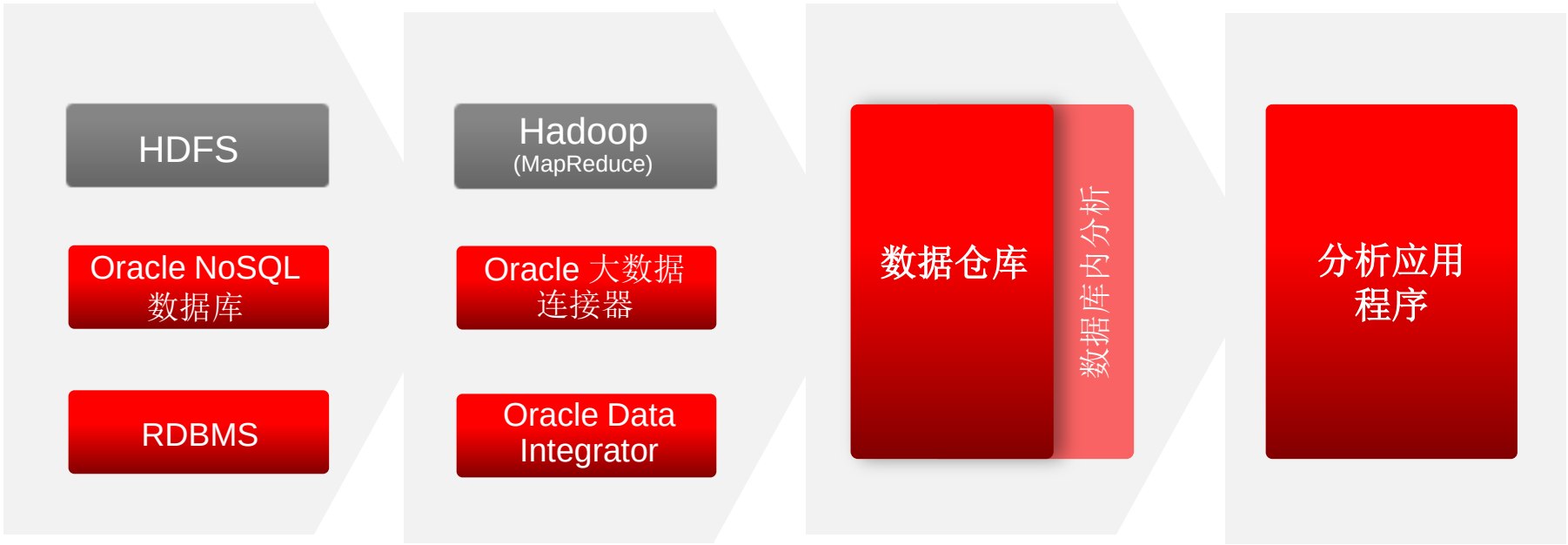
开源语言和环境

用于统计计算和统计绘图

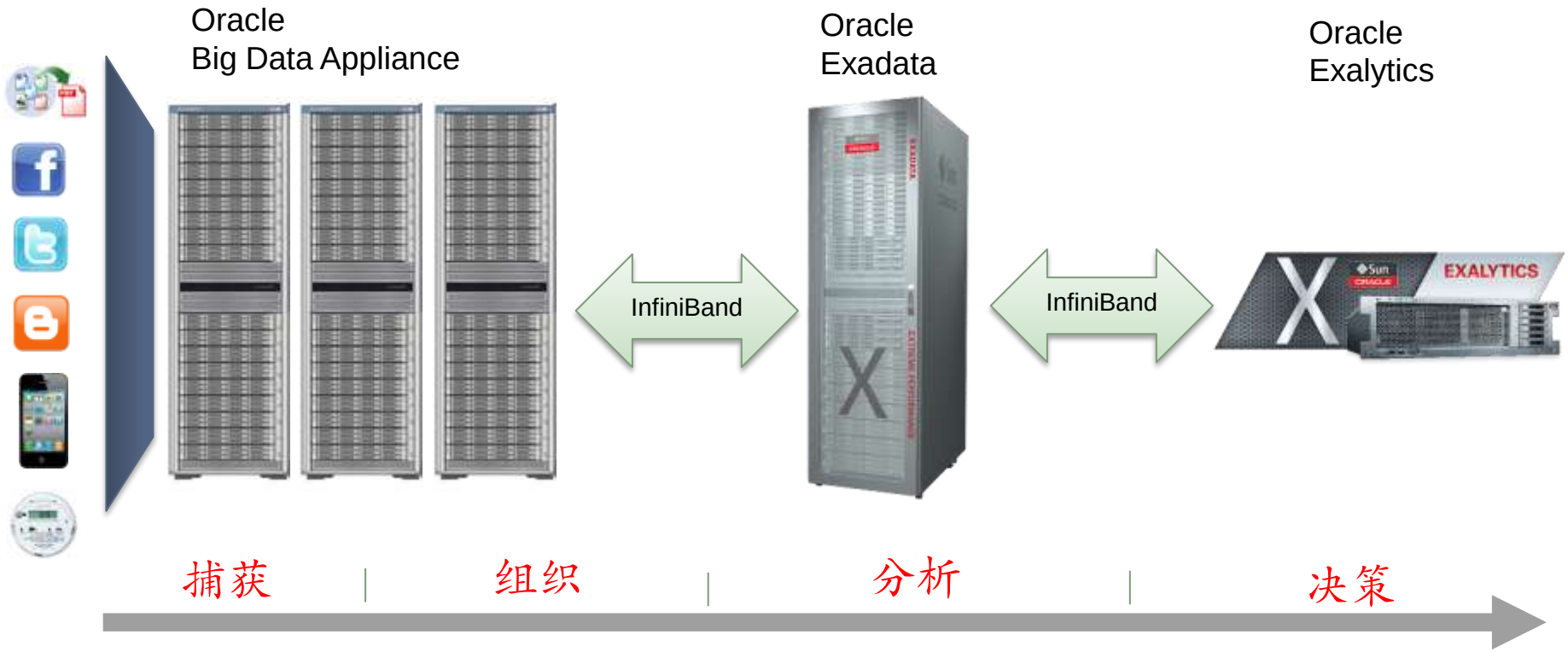
能够轻松制作出版级高质量图表

高度可扩展

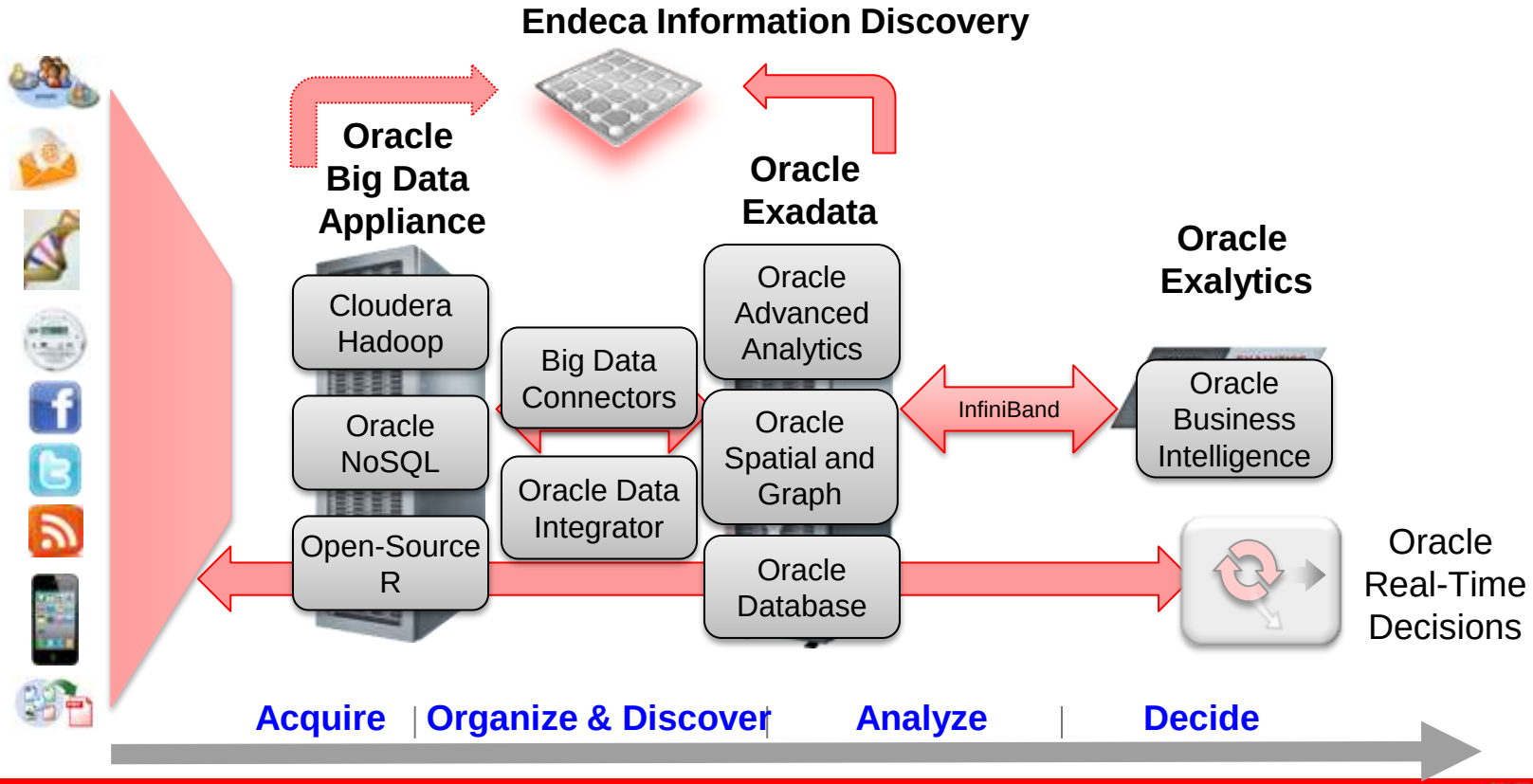
四、Oracle 面向大数据的解决方案体系



软硬一体优化集成的Oracle大数据综合解决方案

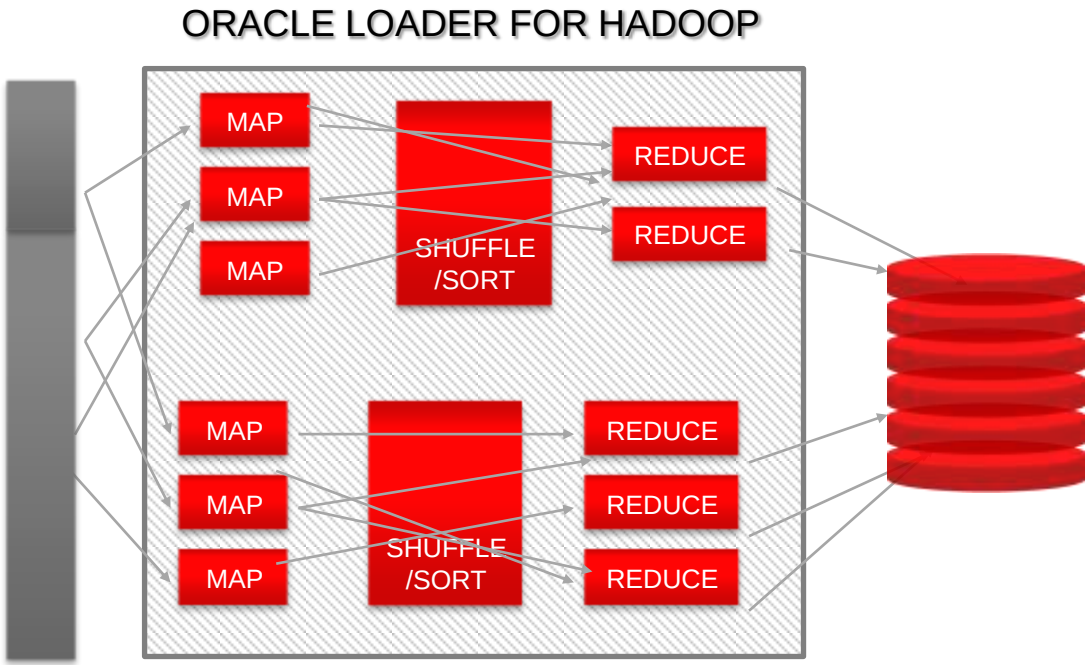


Oracle Big Data solution



Oracle Loader for Hadoop

使用集群技术



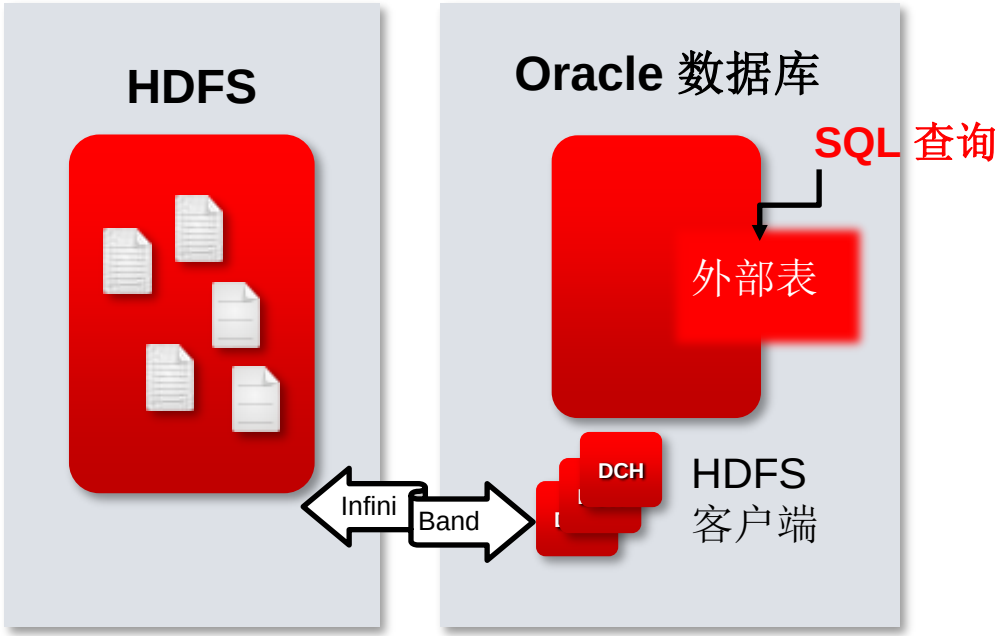
MapReduce 工作流的最后阶段

分区表和未分区表

在线和离线加载

Oracle Direct Connector for HDFS

从 Oracle 数据库直接访问

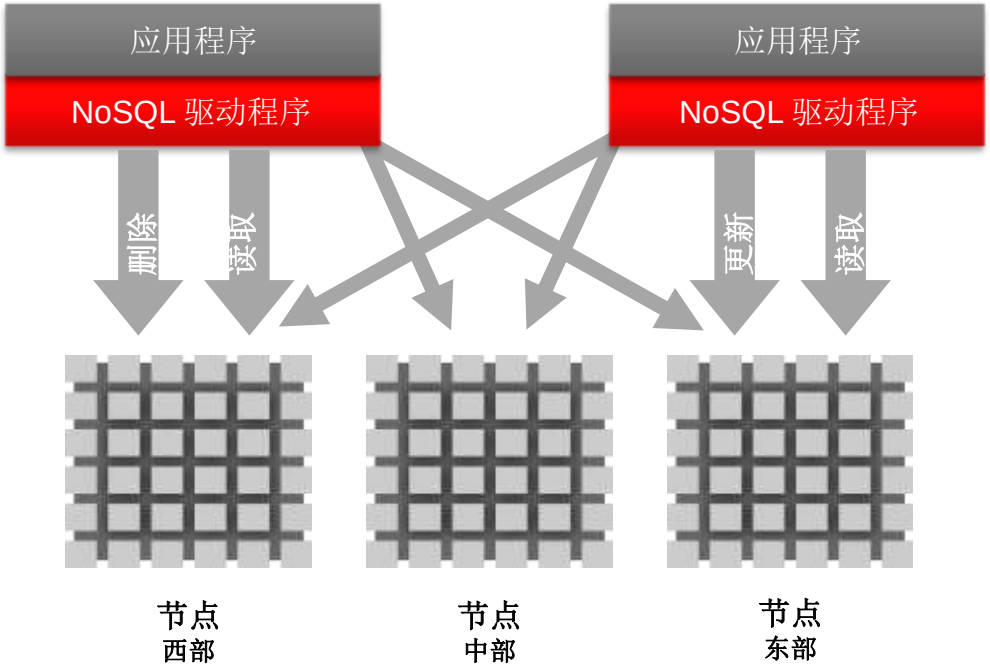


对 HDFS 的 SQL 访问

外部表视图

数据查询或导入

Oracle NoSQL 数据库



分布式键值对数据库

简单编程模型

可伸缩的吞吐量

商业软件和支持

易于管理

Oracle的这个NoSQL Database，
是在2011年10月4号的甲骨文全球大全上发布的 Big Data Appliance的其中一个组件，
Big Data Appliance是一个集成了Hadoop、NoSQL Database、Oracle数据库Hadoop适配器、
Oracle数据库Hadoop装载器及R语言的系统。

Oracle NoSQL 数据库主要特性

简单数据模型：

- 简单数据模型 — 键值对（主键 + 次键模式）
- 简单操作 — 读取/插入/更新/删除，RMW 支持
- 事务范围 — 主键内的记录、单一 API 调用
- 无序扫描所有数据（非事务）

ACID 事务：

- 按操作逐个指定，应用程序设置默认值
- 可配置的持久性策略
  - 同步策略 + 副本确认策略
- 可配置的一致性策略

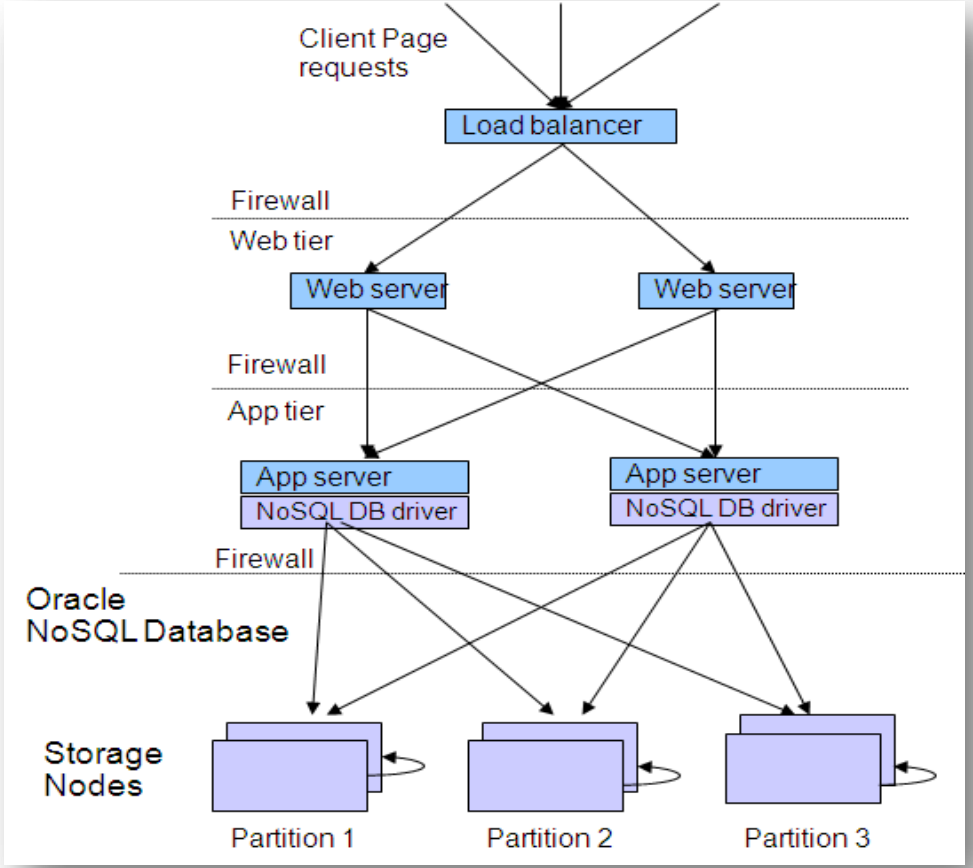
独特优势：

- 与 Oracle 体系无缝集成
- 商业级
- 可伸缩
- 简单编程模型
- 易于管理

Oracle NoSQL 数据库

企业拓扑

- 复制了应用服务器
- 驱动程序链接到每个应用程序中
- 数据节点保持最新
- 存储节点跨多个数据中心
- 自动处理存储节点故障
  - 优雅降级
  - 自动发现
- 无单点故障



Oracle NoSQL 数据库用例

- 数据捕获
  - 传感器数据捕获（即信息家电、智能电网、地球科学、生物医学科学）
  - 统计信息和网络捕获（QOS 网络管理）
  - Web 应用（一路点击式捕获）
  - 针对移动设备的备份服务
- 数据服务
  - NoSQL 数据共享（地球科学、生物医学）
  - 可伸缩的身份验证
  - 实时通信（MMS、SMS、路由）
  - 社交网络、个性化

从 Oracle 数据库访问 Hadoop 数据

Oracle Loader for Hadoop	用例特性
通过 JDBC 在线加载	最简单的未分区表用例
通过直接路径在线加载	分区表的快速在线加载
通过 datapump 文件离线加载	外部表的最快加载方法数据库服务器上的加载较少

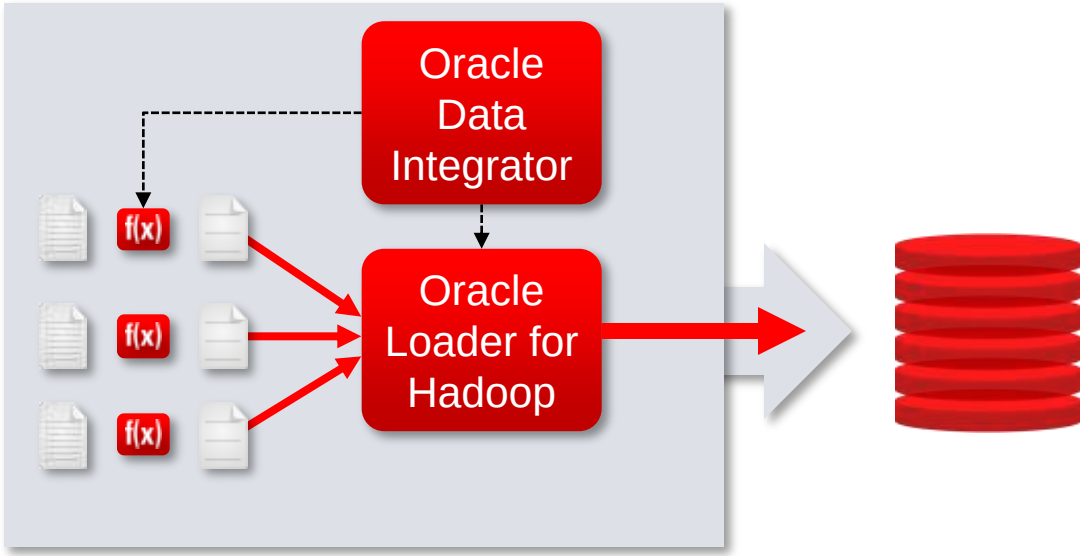
Oracle Direct Connector for HDFS	
从 Oracle 数据库对 HDFS 进行 SQL 访问	数据留在 HDFS 上 从数据库并行访问
与 Oracle Loader for Hadoop 联用	访问由 OLH 创建的文件或导入 Oracle 表

开发 MapReduce 所需的技能

- Java
- Hadoop 框架
- 并行算法

Oracle Data Integrator

简化 MapReduce

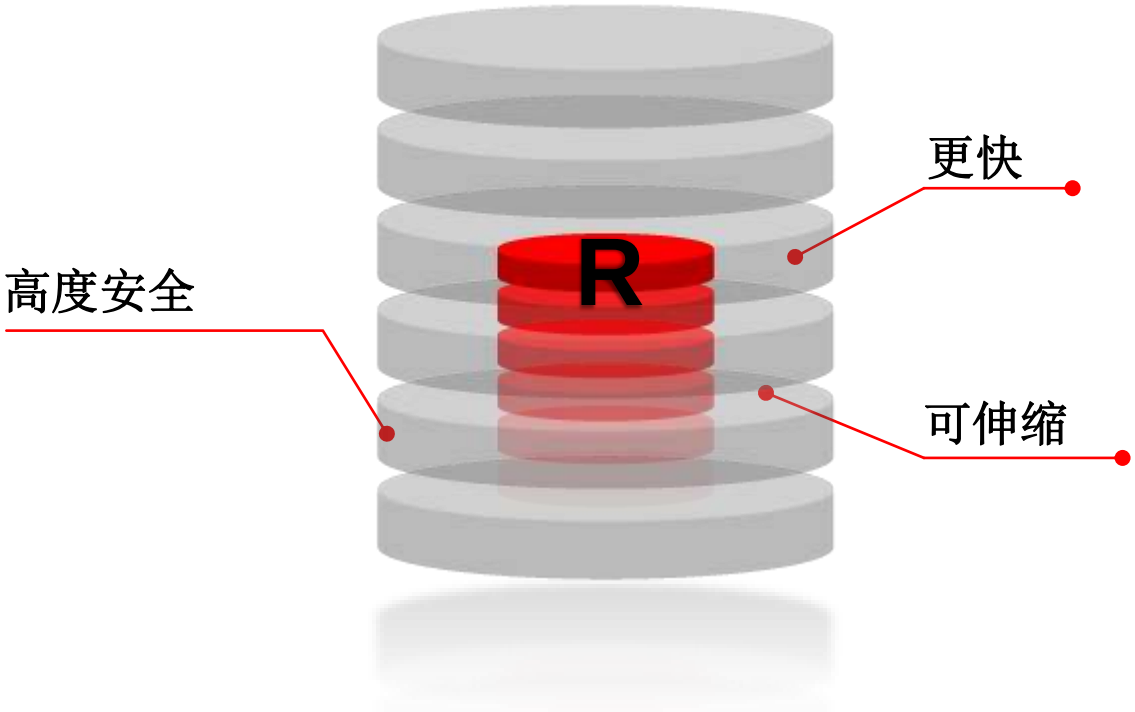


自动生成 MapReduce 代码

管理进程

加载到数据仓库

Oracle R Enterprise



在数据库中运行模型

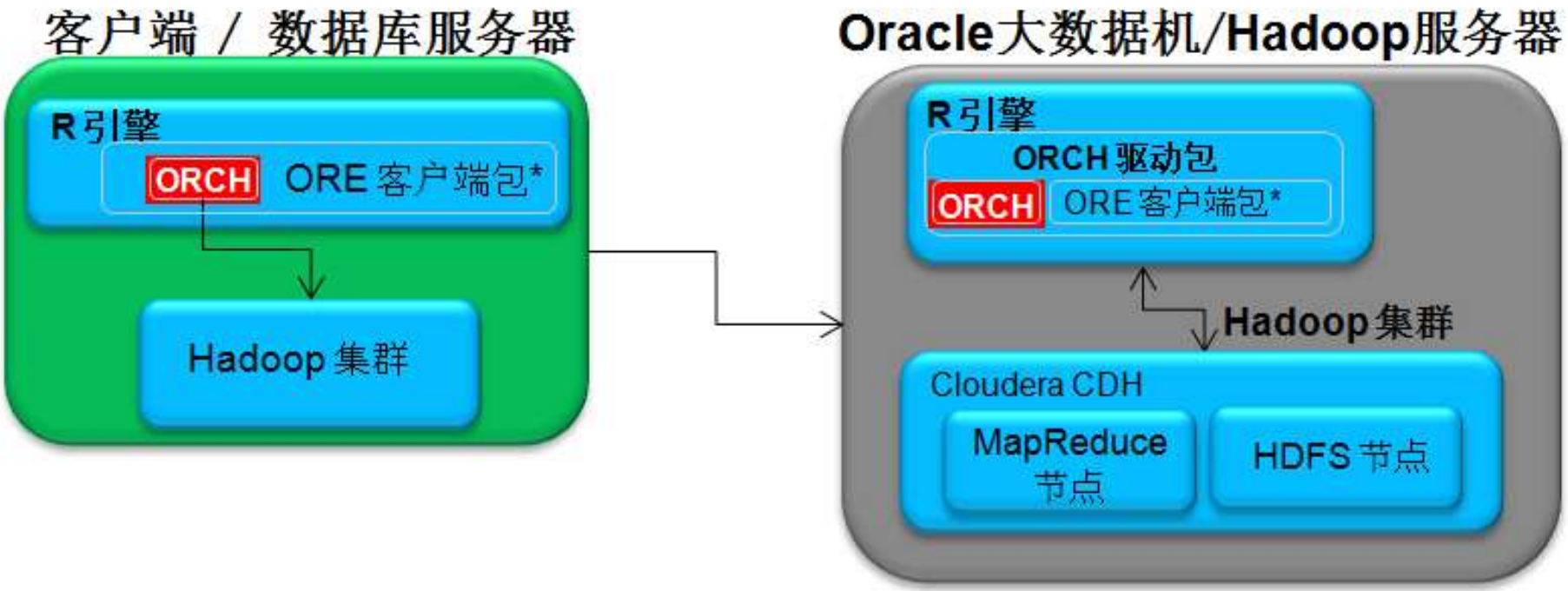
可处理大型数据集

发挥 Oracle Database 11g 和 Exadata 的强大能力

代码相同，而速度更快

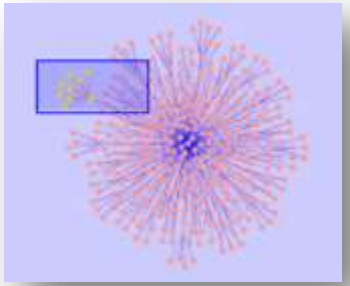
Oracle Advance Analytics

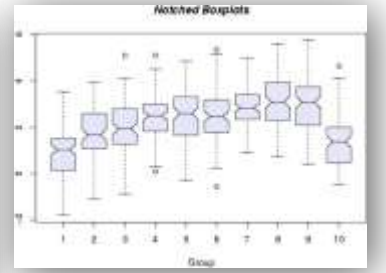
Oracle R Connector for Hadoop

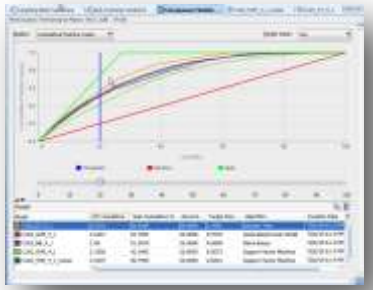


Oracle 数据库强大分析平台

新增 Oracle Advanced Analytics







统计

数据挖掘

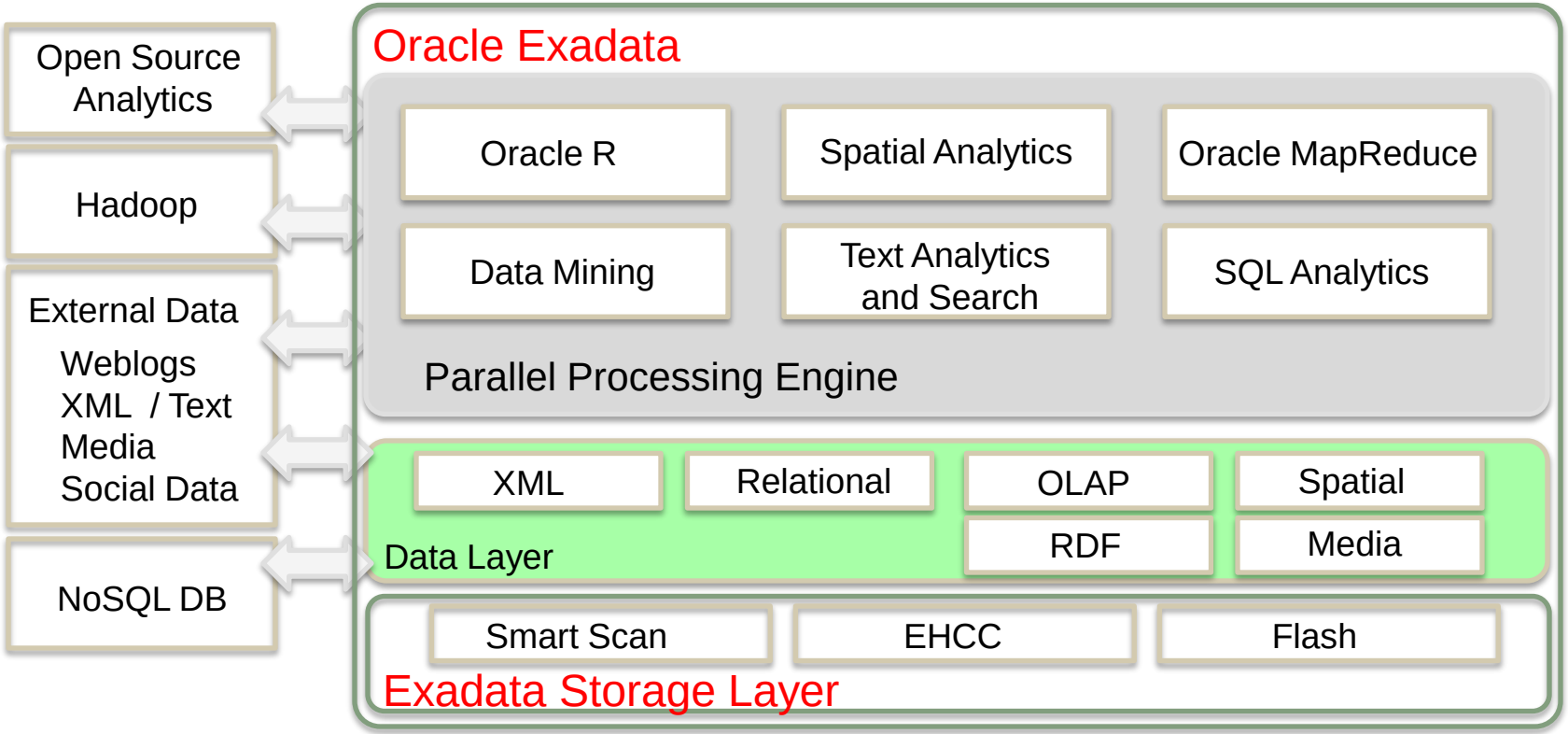
文本

图形

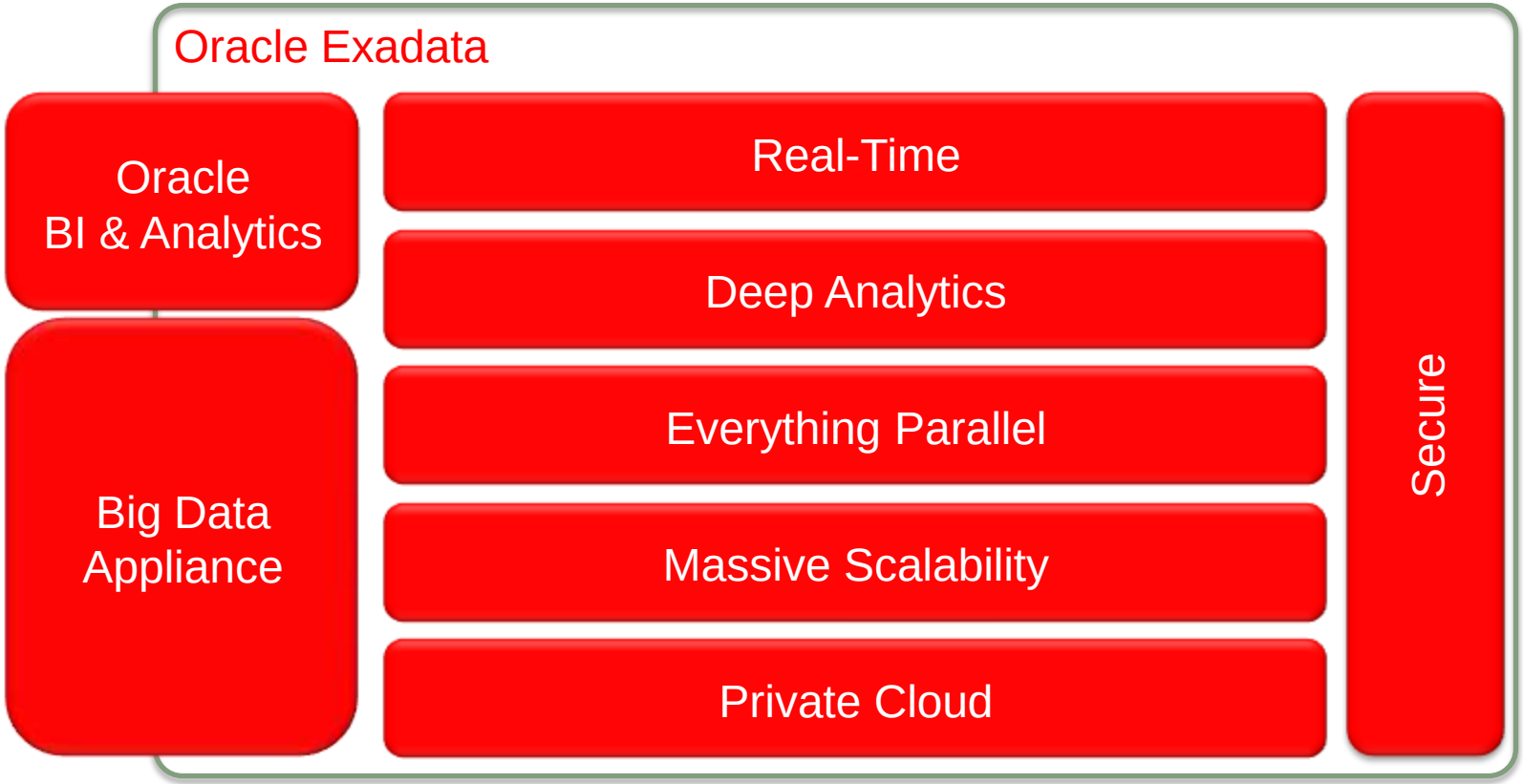
空间

语义

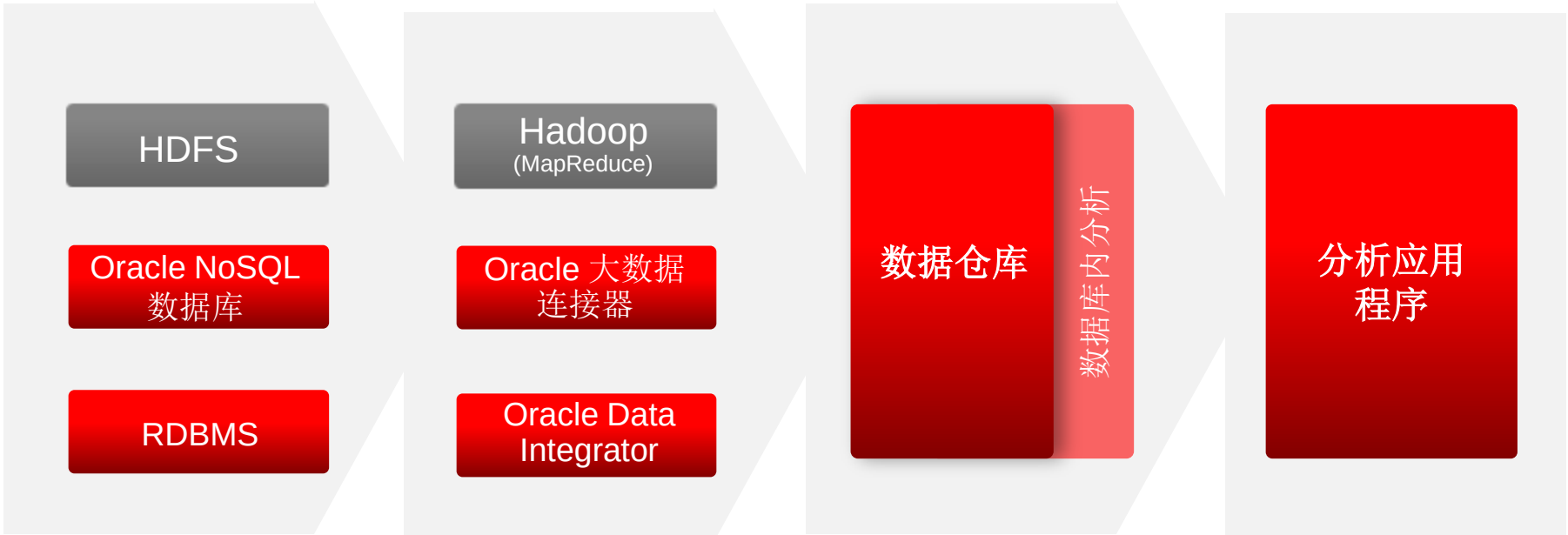
Oracle大数据增强Exadata数据分析能力



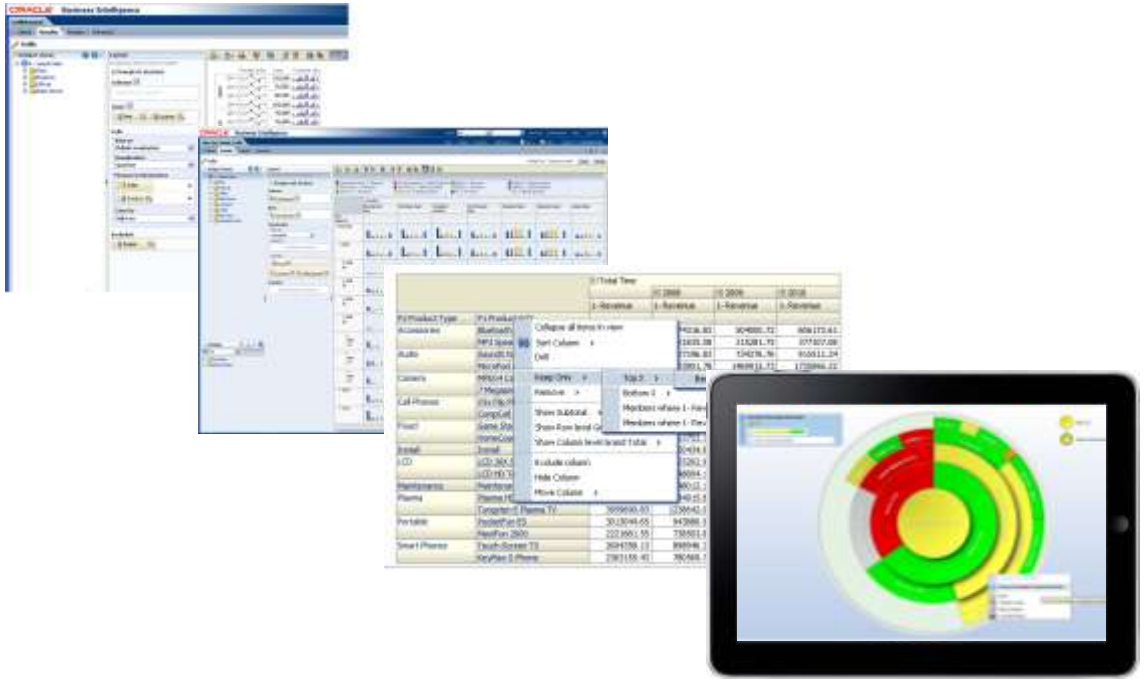
Oracle大数据增强Exadata数据分析能力



Oracle 面向大数据的集成解决方案体系



商务智能辅助决策—快如闪念的交互式分析



交互式分析

自由挖掘

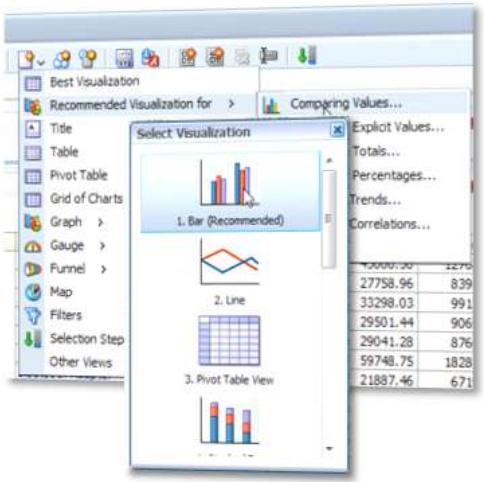
密集可视化

完全移动

快如闪念的交互式最终用户体验

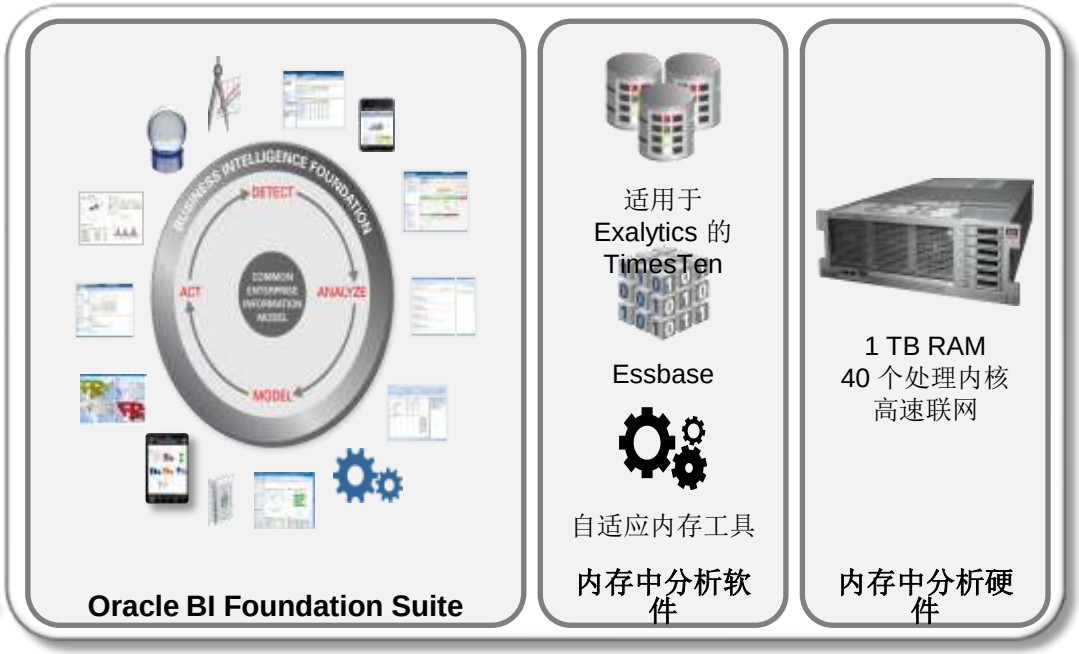
- 高度交互式分析
- 自由格式数据挖掘
- 高密度可视化

- 视图自动建议
- 上下文相关的操作
- 全面支持移动



Oracle Exalytics 智能分析服务器

- 首个集成设计的分析系统
- 无限制的可视分析
- 更智能的分析应用程序



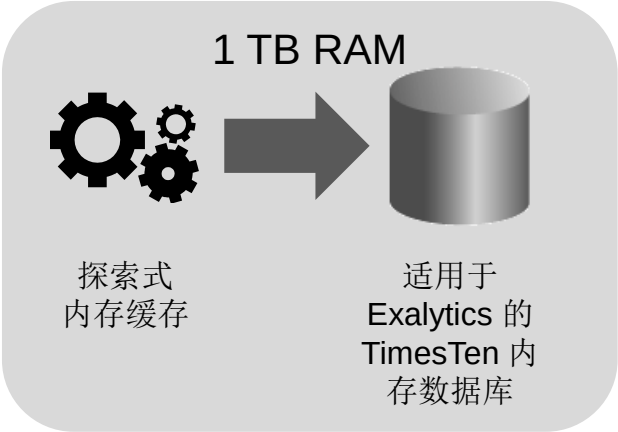
Oracle Exalytics 内存中分析

探索式自适应内存缓存

- 确定在内存中存放哪些内容
- 自我调适以适应分析负载的变化

内存数据库

- 并行 TimesTen 数据库
- 并行 Essbase
- 高级列压缩
- 内存中分析功能



Oracle Exalytics

卓越的洞察力带来更高的业务绩效

首个面向分析的集成设计系统

性能卓越且总拥有成本更低

无限制的可视分析

随时随地查看想要的分析

更智能的分析应用程序

降低采用风险，提供新的机遇

Oracle Exalytics上运行 80 多个分析应用程序

无需更改应用程序

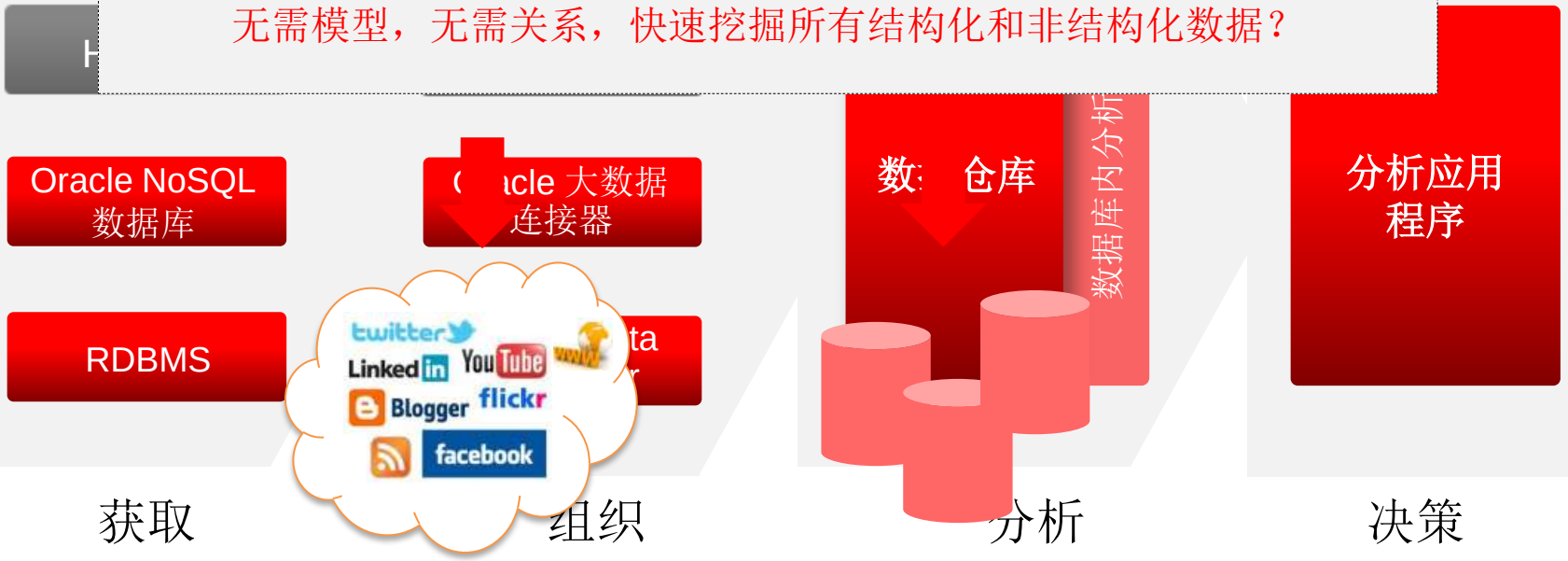
- 财务、HR
- 销售、市场营销
- 计划、预测
- 多个行业

Oracle Exalytics 分析大数据

- 全面
- 满足企业级需求
- 集成设计、卓越性能
- 经过优化，实现卓越分析

Oracle 面向非结构化数据直接数据发现

无需模型，无需关系，快速挖掘所有结构化和非结构化数据？



Oracle Endeca Information Discovery

适用于企业范围数据发现的应用平台

Oracle Endeca Information Discovery 可帮助组织快速挖掘所有相关数据

- 组合来自不同系统的结构化和非结构化数据
- 自动组织信息以便搜索、发现和分析
- 快速组装易于使用的分析应用

什么是数据发现?

快速挖掘所有相关数据

- ✓ 未定义或未知的关系
- ✓ 无需预定义的模型
- ✓ 快速、反复的变化

- ✓ 高级搜索
- ✓ 分面导航
- ✓ 分析

- ✓ 结构化
- ✓ 半结构化
- ✓ 非结构化
- ✓ 散乱数据也不例外
- ✓ 不仅限于数据仓库

何时需要数据发现?

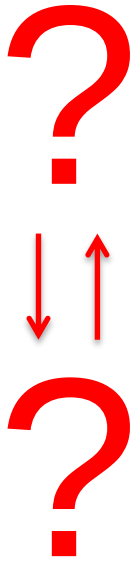
业务和 IT 可就以下最佳指标达成一致

1. 业务无法确定哪些问题至关重要，因为：

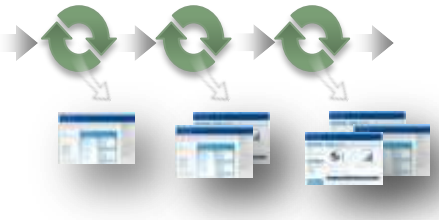
- 变量太多，无法给出所有组合
- 业务环境多变，无法完全预见

2. IT 无法确定哪个数据模型有效，因为：

- 源模式的多样性使得数据难以遵循单一模型，相关工作也非常耗时
- 源包括非结构化数据
- 模式经常变化，需要费钱费时的重复工作



数据量、种类的增加带来新的挑战

更多的多样化数据		结构化 和非结构化的内外部数据快速增长
更多的变化和不确定性		预定义的模型、信息板和报告无法满足意外业务需求
更多的意外问题		能够根据需要以自助方式挖掘数据、添加新数据和构建分析

从种类更多、量更大的更多中获取价值的挑战

挑战

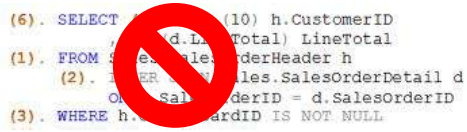
建模时间和成本

数据多样性使众多数据源愈发难以遵循单一模型，从而增加了项目人力成本

不确定性和变化

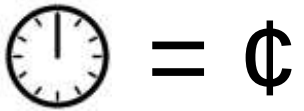
集成和可视化要求的变化导致必须在延迟和快速用户采用之间做出权衡

非技术用户



具备领域知识的业务用户提出适当的问题时无法编写查询来表达问题

Information Discovery 将克服这些障碍

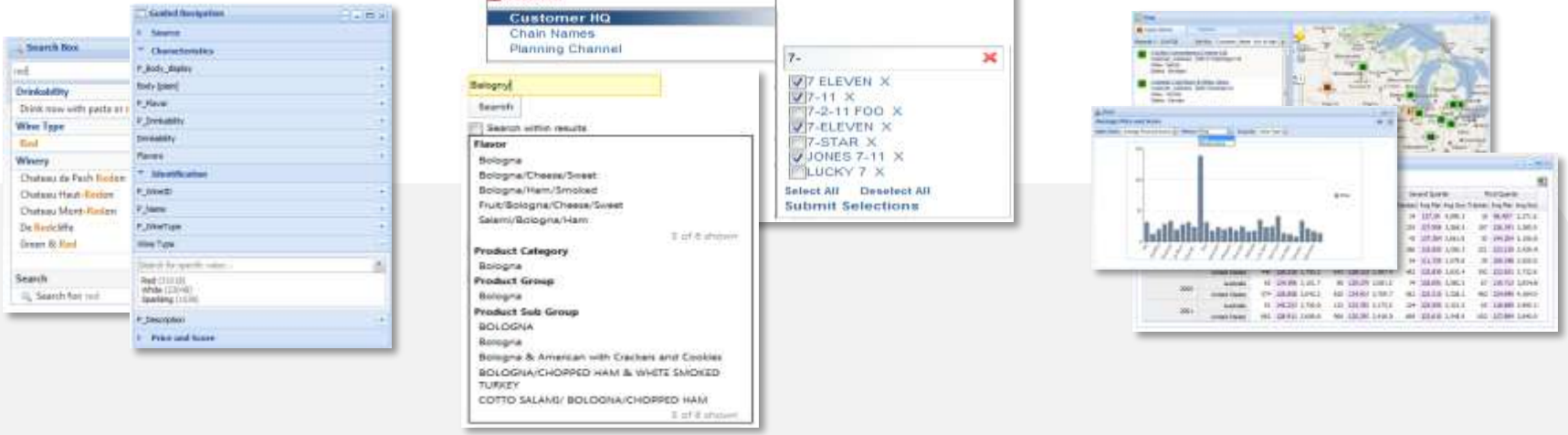
	挑战	解决方案
建模时间和成本 	数据多样性使众多数据源愈发难以遵循单一模型，从而增加了人力成本	从数据本身导出模型的技术自行完成许多建模工作
不确定性和变化 	集成和可视化要求的变化导致必须在延迟和快速用户采用之间做出权衡	与业务协作进行快速原型设计，以发现和满足潜在要求
非技术用户 	具备领域知识的业务用户提出适当的问题时无法编写查询来表达问题	无需培训的界面将消费者用户体验创新引入企业软件

搜索的简单性，BI 的强大功能

- 更易用
  - 源自 10 年的电子商务消费者用户经验
- 搜索 + 分面导航 + 可视分析
  - 搜索和选择属性，如网站
- 交互式挖掘
  - 响应迅速的 Endeca Server

交互式挖掘和发现

优化用户体验以促进发现



高级搜索

- 搜索前瞻
- 拼写纠正
- 数据驱动的筛选

+

分面导航

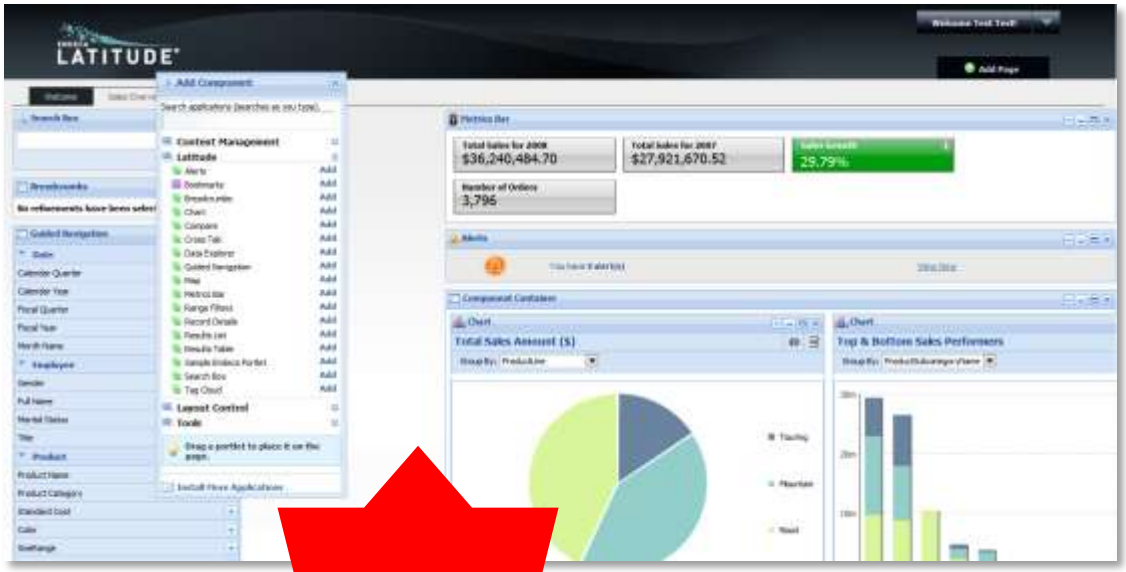
- 选择属性，如网站

+

可视分析

- 图表和交叉表
- 地理位置可视化
- 标签云

应用构建更像领域专家与数据交谈



- 促进发现以获得新洞察
- 快速扩展和变化
- 随着新数据的到达快速迭代
- 快速响应业务变化

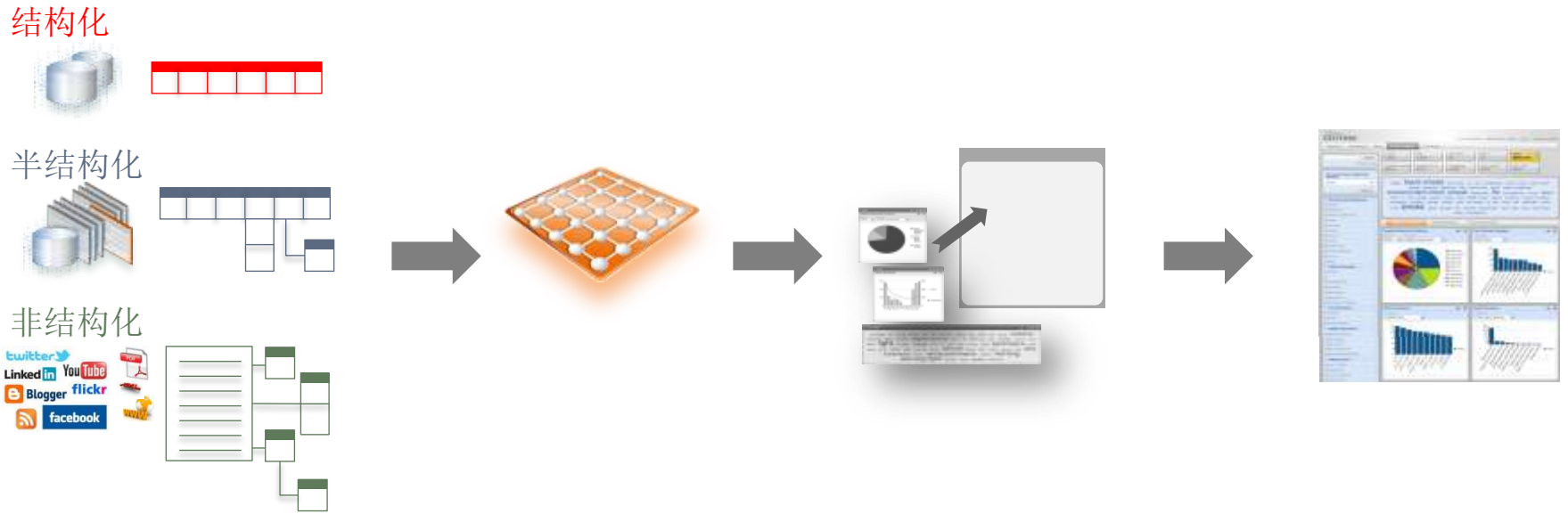
发现应用的生命周期

多种多样不断变化的信息

在 Oracle Endeca Server 中自动统一和丰富 — 无需预定义的模型

拖放式应用构建

通过交互式搜索、导航和分析实现挖掘和分析



Hardware and Software

ORACLE

Engineered to Work Together