大数据让企业实现“大智能”

——用友数据处理平台介绍

用友集团iUAP中心

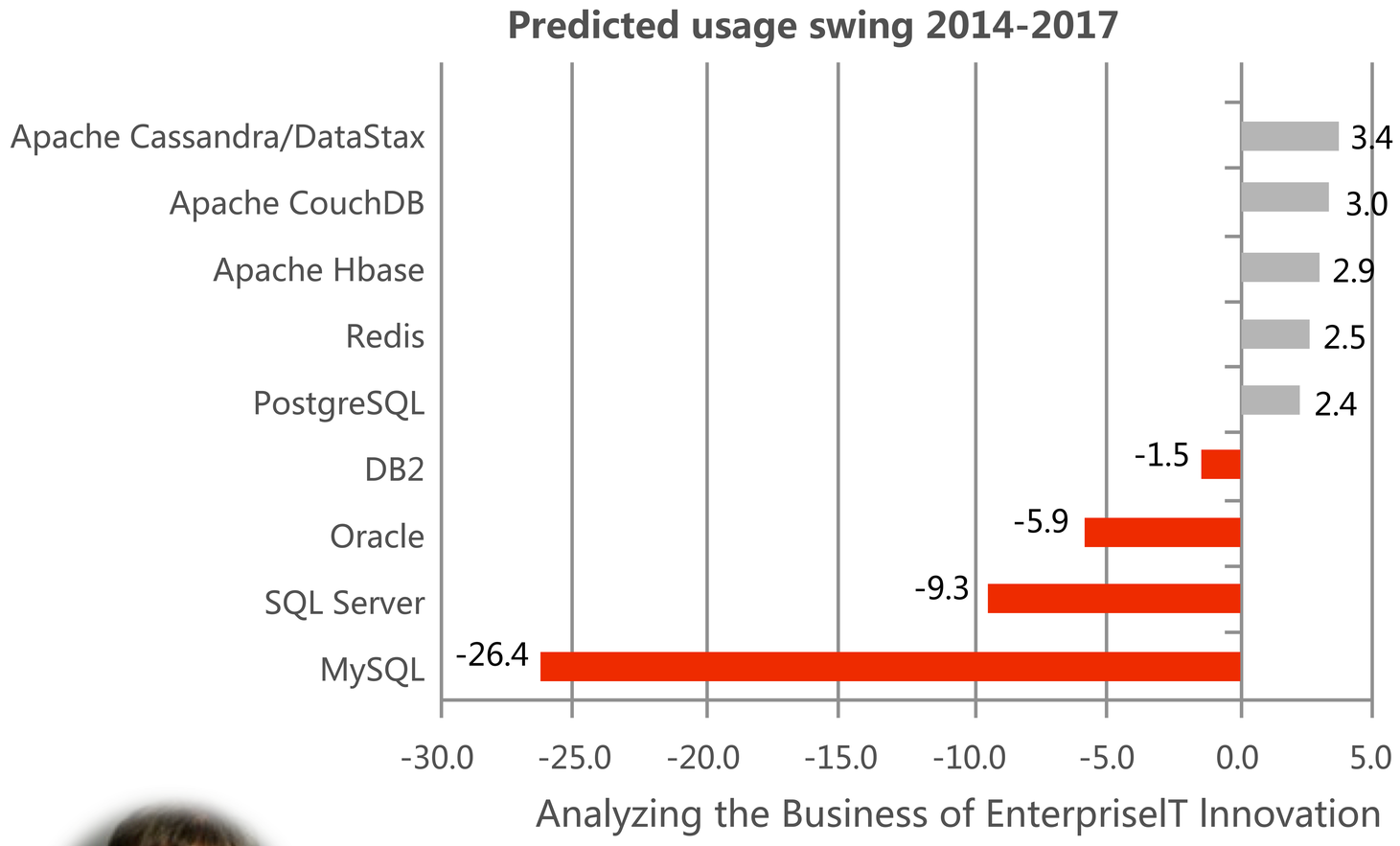
2015年5月5日

yonyou Network Technology Co., Ltd..

数据范畴的延伸

GPS
电子支付
协同制造
公共服务
第三方电商平台
运营商
阿里巴巴
Alibaba.com
Baidu 百度
外包制造企业
分销中心
经销商
消费者
供应商
核心企业
下属企业
直营店
供应商
内部一体化
新浪微博
第三方物流
KA商超
第三方物流
消费者
供应商
供应商
网上商城
售后服务站
支付宝
外包制造企业
企业客户
供应商
产业链协同
仓储
仓储
中国建设银行
China Construction Bank
社会化商业协同
www.51job.com
前程无忧
淘宝网
腾讯

数据基础架构的演进

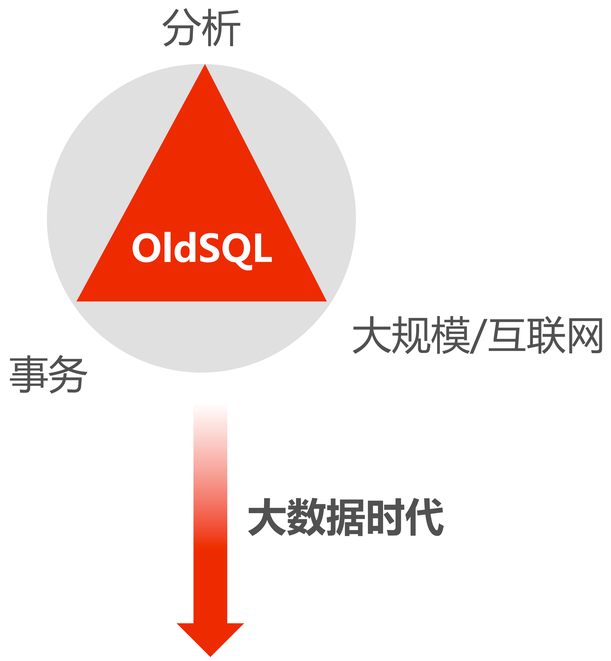


数据基础架构思路应该由“**一种架构支持所应用**转变成**多种架构支持多类应用**”

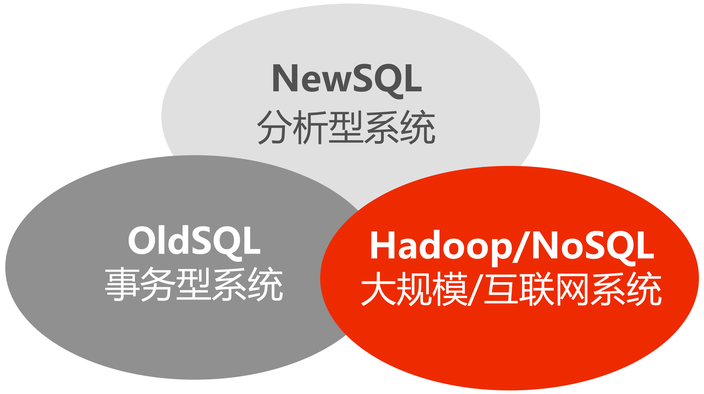
— Michael Stonebraker

现代数据库的基础架构奠基人，图灵奖获得者

一种架构支持多类应用



多种架构支持多类应用



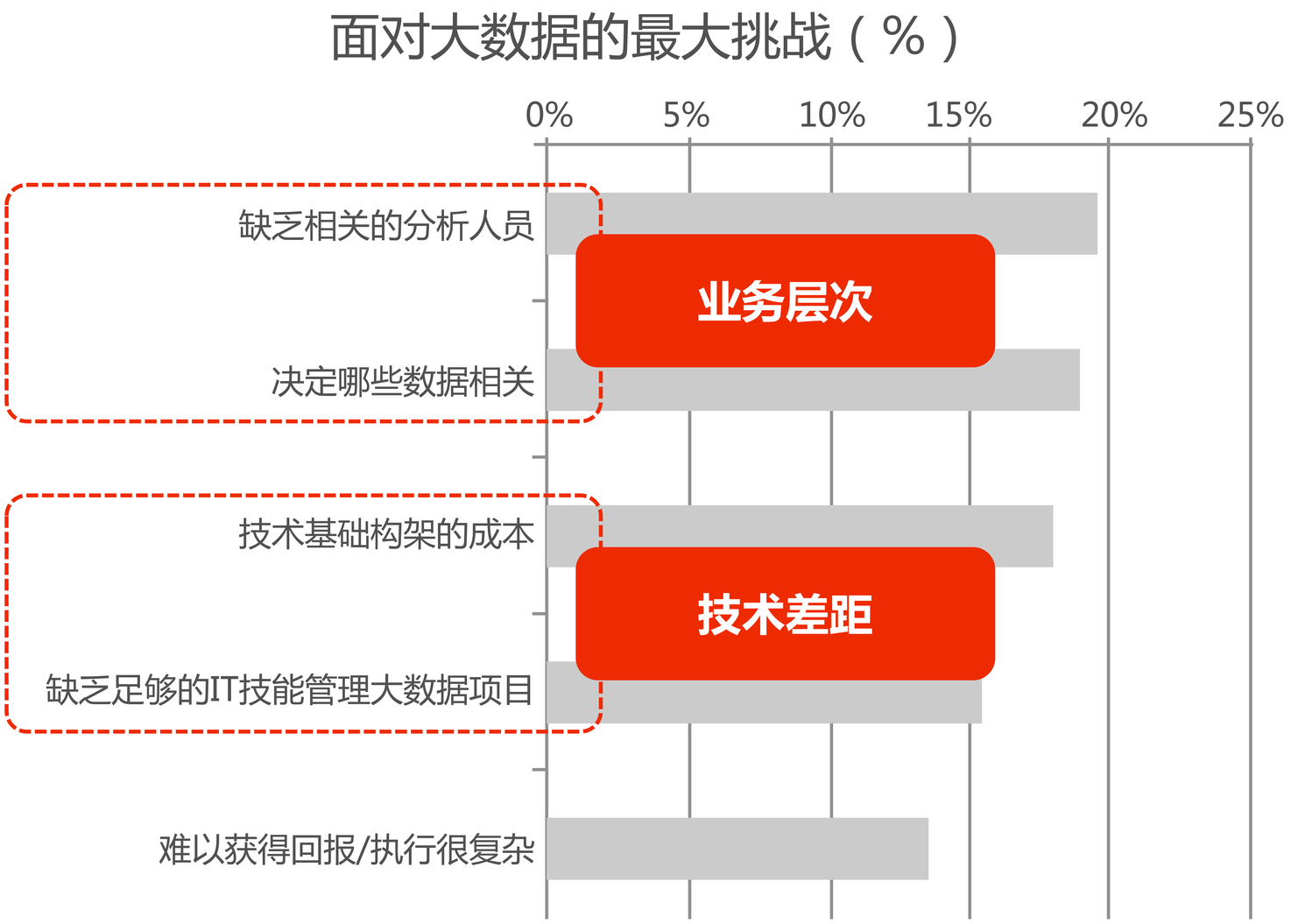
应用趋势与大家的挑战

自动化决策

分析挖掘

组织与管理

基础架构



来源：IDC软件市场最终用户调研 2014（n=750,中国=100）

数据处理技术的阶段

yonyou iUAP

数据源

数据采集

iUAP DI（数据集成）

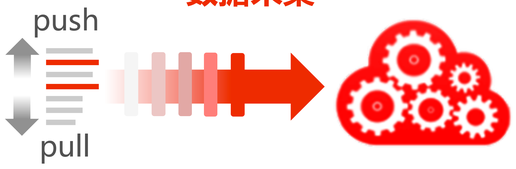
Data Integrator

数据质量	模型管理	模型管理
数据转换组件		
文件	JDBC	ODBC	JMS	SOAP	Others
插件化，可扩展的数据类型支持		

iUAP CDC（实时数据同步）

Change Data Capture

数据采集



组织存储

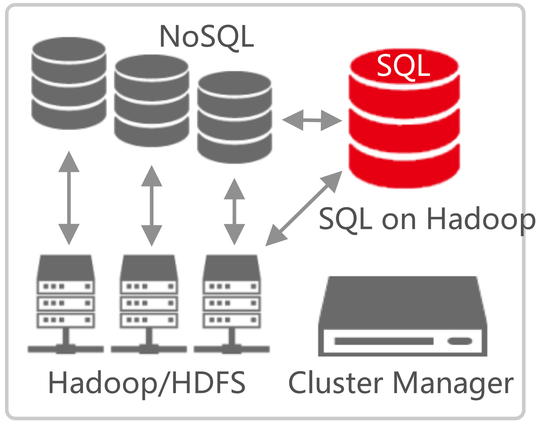
Yonyou AE（加速引擎）

Accelerstion Engine

海量结构化数据分析

iUAP UDH（大数据引擎）

iUAP Distribution for Hadoop



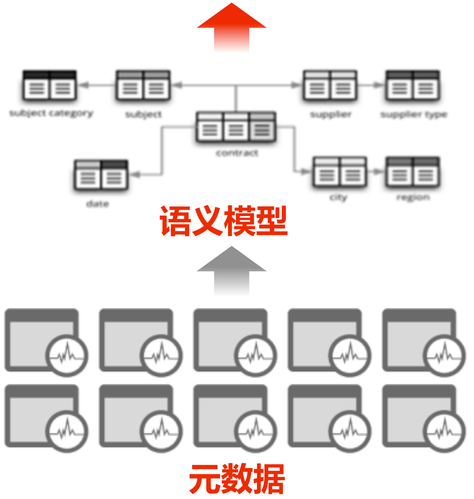
探索分析

统一建模与管理

Data Studio　调度　用户

元数据仓库　日志　监控

OLAP模型　**挖掘模型**



语义模型

元数据

可视化决策

iUAP DataInsight

Yonyou BQ

BQ Mobile

3rd

tableau　QlikView

我们能够做到

- 支持**10倍**以上的分析效率提升
- 支持**10倍**以上的数据空间压缩
- 支持与分析应用的透明隔离，**零成本**

- 支持仓库，分析，挖掘的可视化建模
- 支持模型的**统一管理**

数据整合

- 支持多种类型数据来源的数据集成
- 支持基于DB日志解析的增量数据**实时**同步
- 支持企业外部数据（互联网）的采集获取

分析加速

海量处理

- 支持结构化和非结构化数处理
- 支持**PB级**数据量的分布式并行处理
- 支持**离线批量**和**在线实时**的数据处理

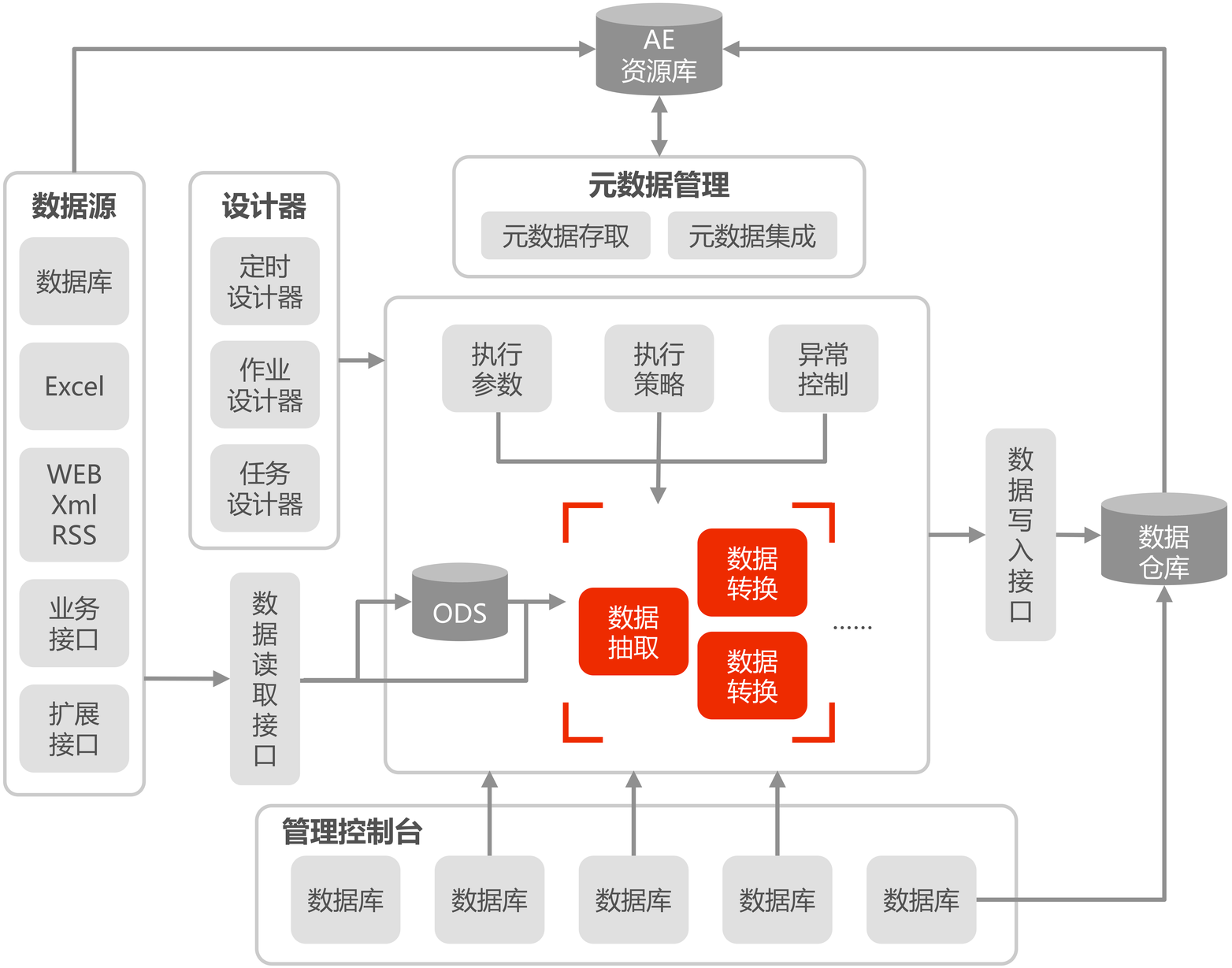
统一建模

挖掘预测

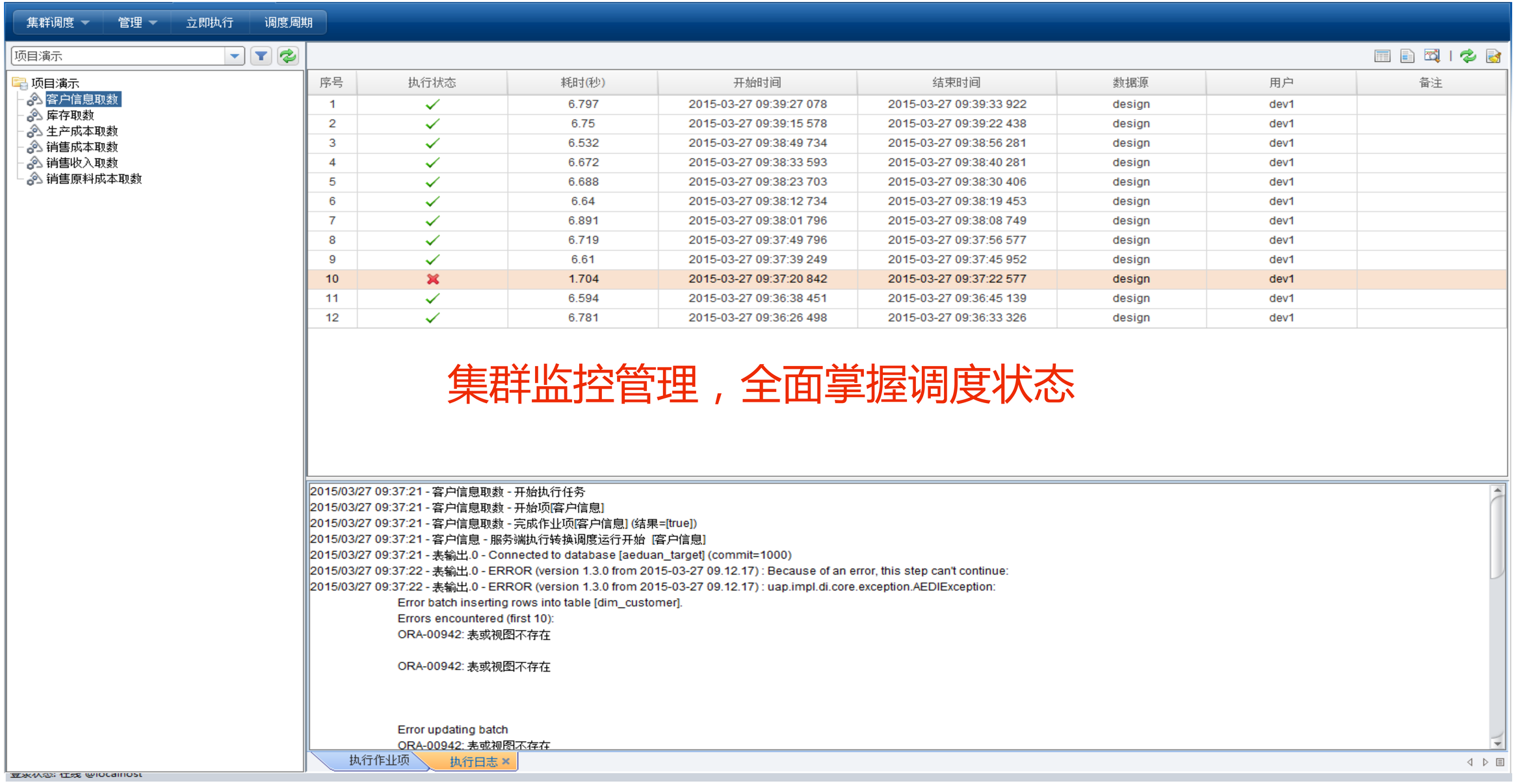
- 支持广泛的数据源类型和算法
- 支持**分布式并行**架构，提升计算效率
- 支持CRISP-DM和PMML标准规范
- 支持丰富的面向主题的模型和结果可视化组件
- 支持**R语言**

异构数据源集成

- 支持常见的数据源类型和专有定制来源
- 丰富的数据处理组件，通过流程设计定制处理逻辑
- 支持集群调度管理，提供完善的日志监控服务
- 集成仓库模型，分析模型，提供仓库全局视图，提升建模效率
- 支持对过程质量和结果质量进行监控，评估。

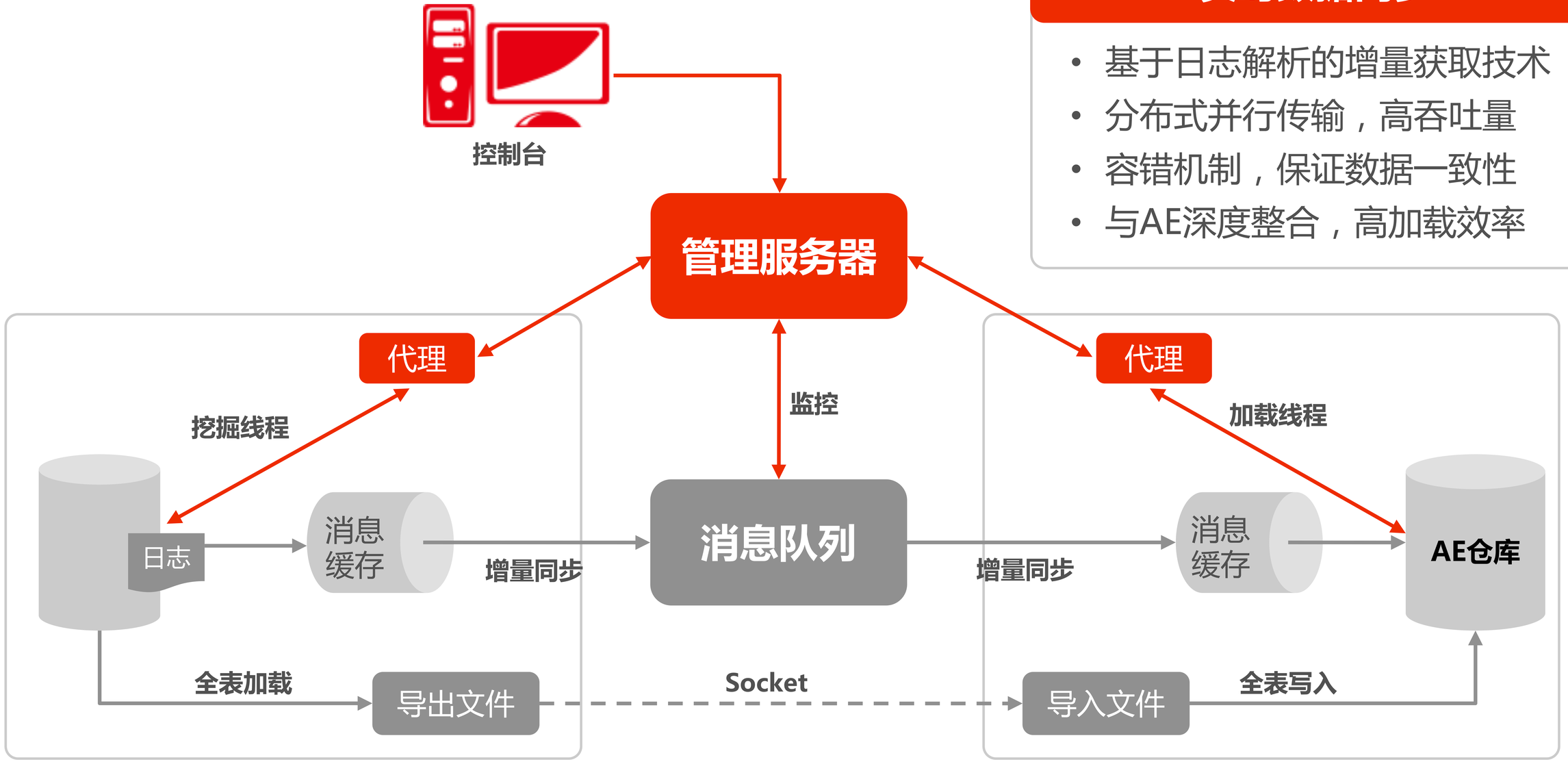


iUAP DI(数据集成）

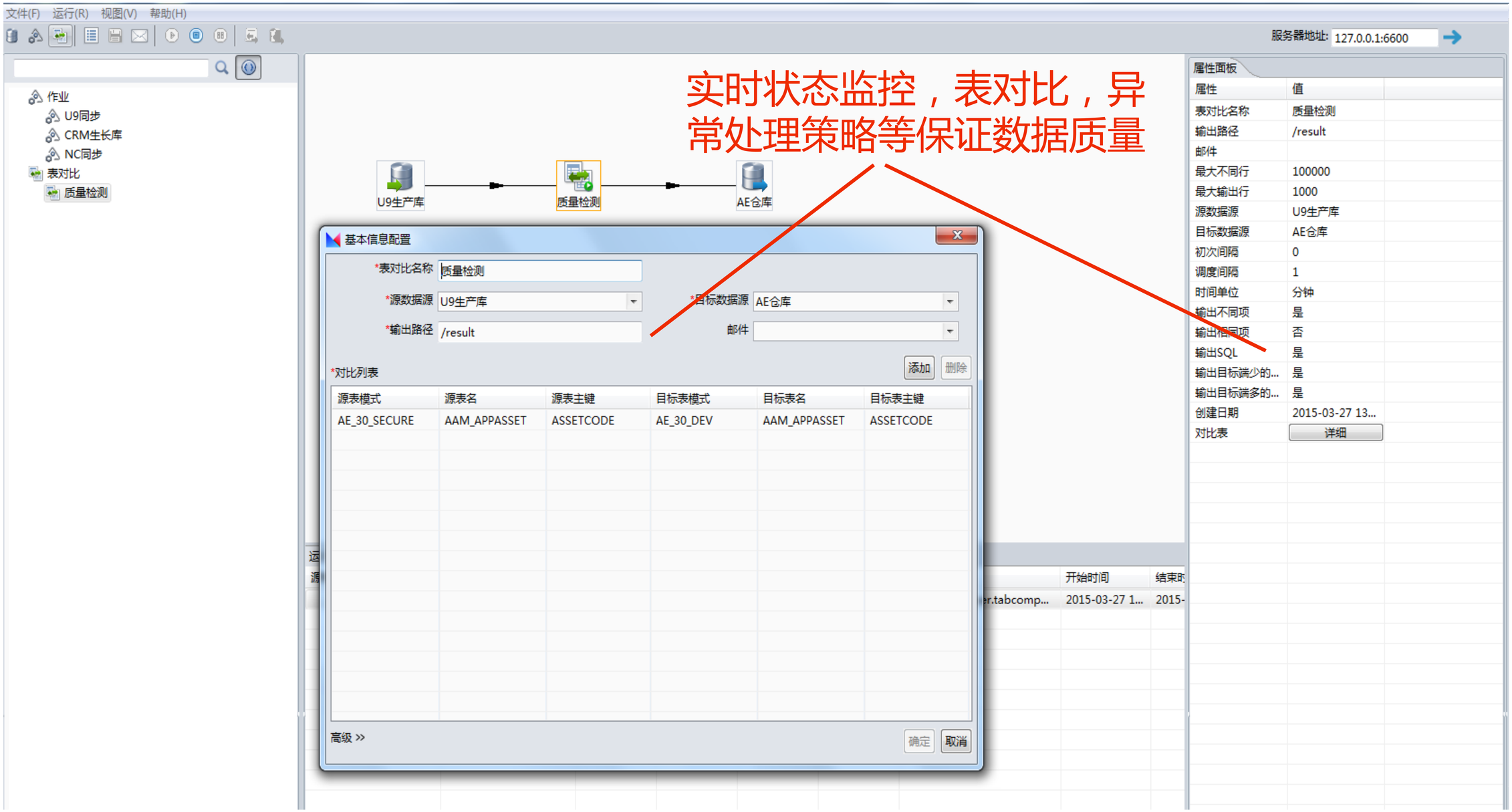

集群调度
管理
立即执行
调度周期
项目演示
项目演示
客户信息取数
库存取数
生产成本取数
销售成本取数
销售收入取数
销售原料成本取数
序号 执行状态 耗时(秒) 开始时间 结束时间 数据源 用户 备注
1 6.797 2015-03-27 09:39:27 078 2015-03-27 09:39:33 922 design dev1
2 6.75 2015-03-27 09:39:15 578 2015-03-27 09:39:22 438 design dev1
3 6.532 2015-03-27 09:38:49 734 2015-03-27 09:38:56 281 design dev1
4 6.672 2015-03-27 09:38:33 593 2015-03-27 09:38:40 281 design dev1
5 6.688 2015-03-27 09:38:23 703 2015-03-27 09:38:30 406 design dev1
6 6.64 2015-03-27 09:38:12 734 2015-03-27 09:38:19 453 design dev1
7 6.891 2015-03-27 09:38:01 796 2015-03-27 09:38:08 749 design dev1
8 6.719 2015-03-27 09:37:49 796 2015-03-27 09:37:56 577 design dev1
9 6.61 2015-03-27 09:37:39 249 2015-03-27 09:37:45 952 design dev1
10 1.704 2015-03-27 09:37:20 842 2015-03-27 09:37:22 577 design dev1
11 6.594 2015-03-27 09:36:38 451 2015-03-27 09:36:45 139 design dev1
12 6.781 2015-03-27 09:36:26 498 2015-03-27 09:36:33 326 design dev1
集群监控管理，全面掌握调度状态
2015/03/27 09:37:21 - 客户信息取数 - 开始执行任务
2015/03/27 09:37:21 - 客户信息取数 - 开始项[客户信息]
2015/03/27 09:37:21 - 客户信息取数 - 完成作业项[客户信息] (结果=[true])
2015/03/27 09:37:21 - 客户信息 - 服务端执行转换调度运行开始 [客户信息]
2015/03/27 09:37:21 - 表输出.0 - Connected to database [aeduan_target] (commit=1000)
2015/03/27 09:37:22 - 表输出.0 - ERROR (version 1.3.0 from 2015-03-27 09.12.17) : Because of an error, this step can't continue:
2015/03/27 09:37:22 - 表输出.0 - ERROR (version 1.3.0 from 2015-03-27 09.12.17) : uap.impl.di.core.exception.AEDIException:
Error batch inserting rows into table [dim_customer].
Errors encountered (first 10):
ORA-00942: 表或视图不存在
ORA-00942: 表或视图不存在
Error updating batch
ORA-00942: 表或视图不存在
执行作业项
执行日志

实时数据同步

- 基于日志解析的增量获取技术
- 分布式并行传输，高吞吐量
- 容错机制，保证数据一致性
- 与AE深度整合，高加载效率



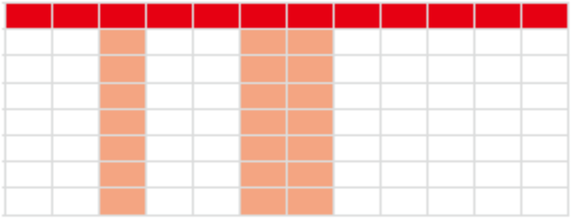
iUAP CDC（数据同步工具）



结构化数据引擎

- I/O降低到传统数据库的1/10以下;
- 缩小比至少1/10;
- 复杂查询和即席查询性能是传统数据库的10倍以上；
- 简单易用，降低使用成本

列存



1TB → 100GB

自适应压缩

100GB → 10GB

水平分区

10GB → 1GB

智能索引

1GB → 100MB

并行

并行加载

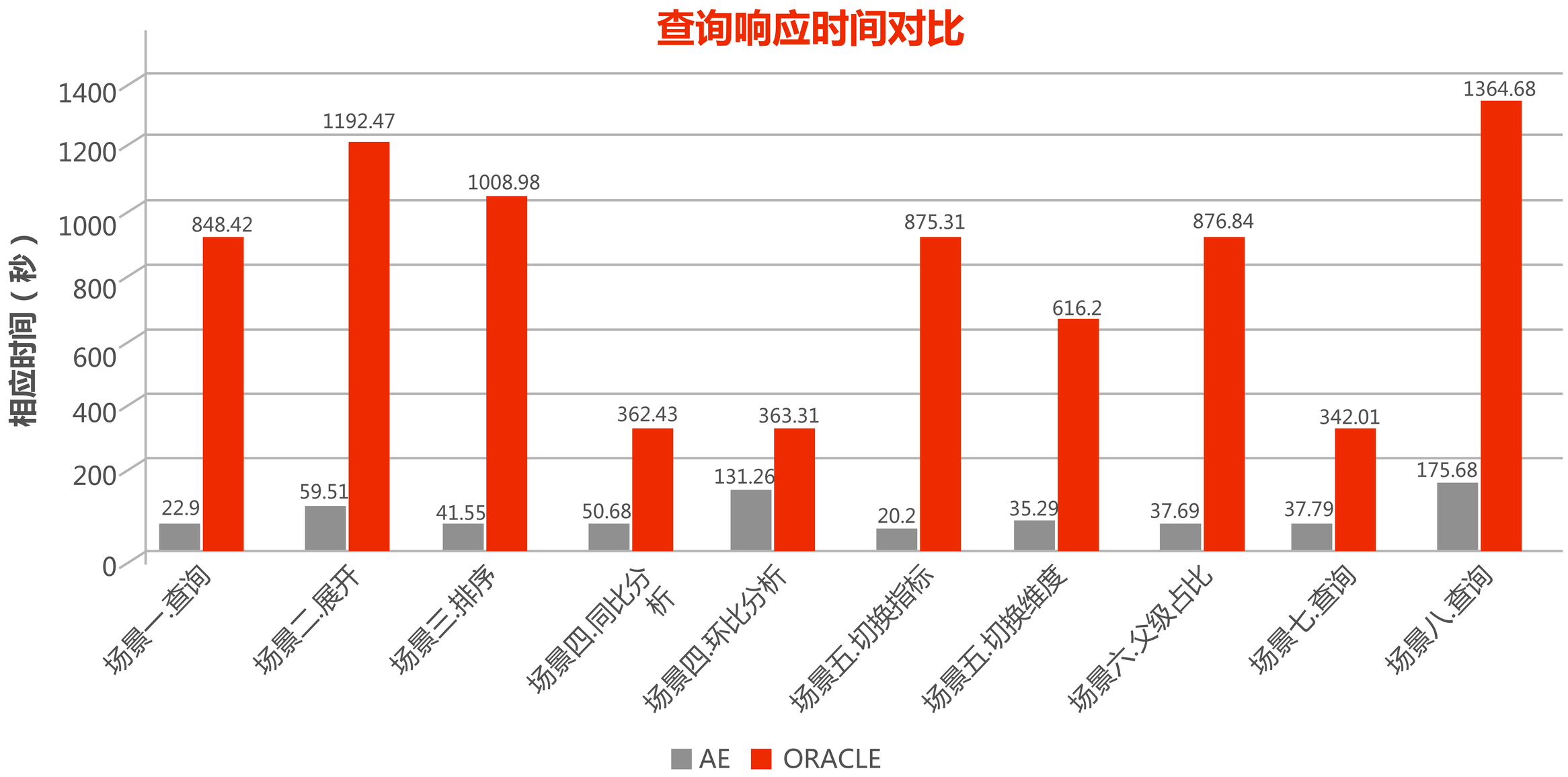
单表200GB/小时

多表500GB/小时

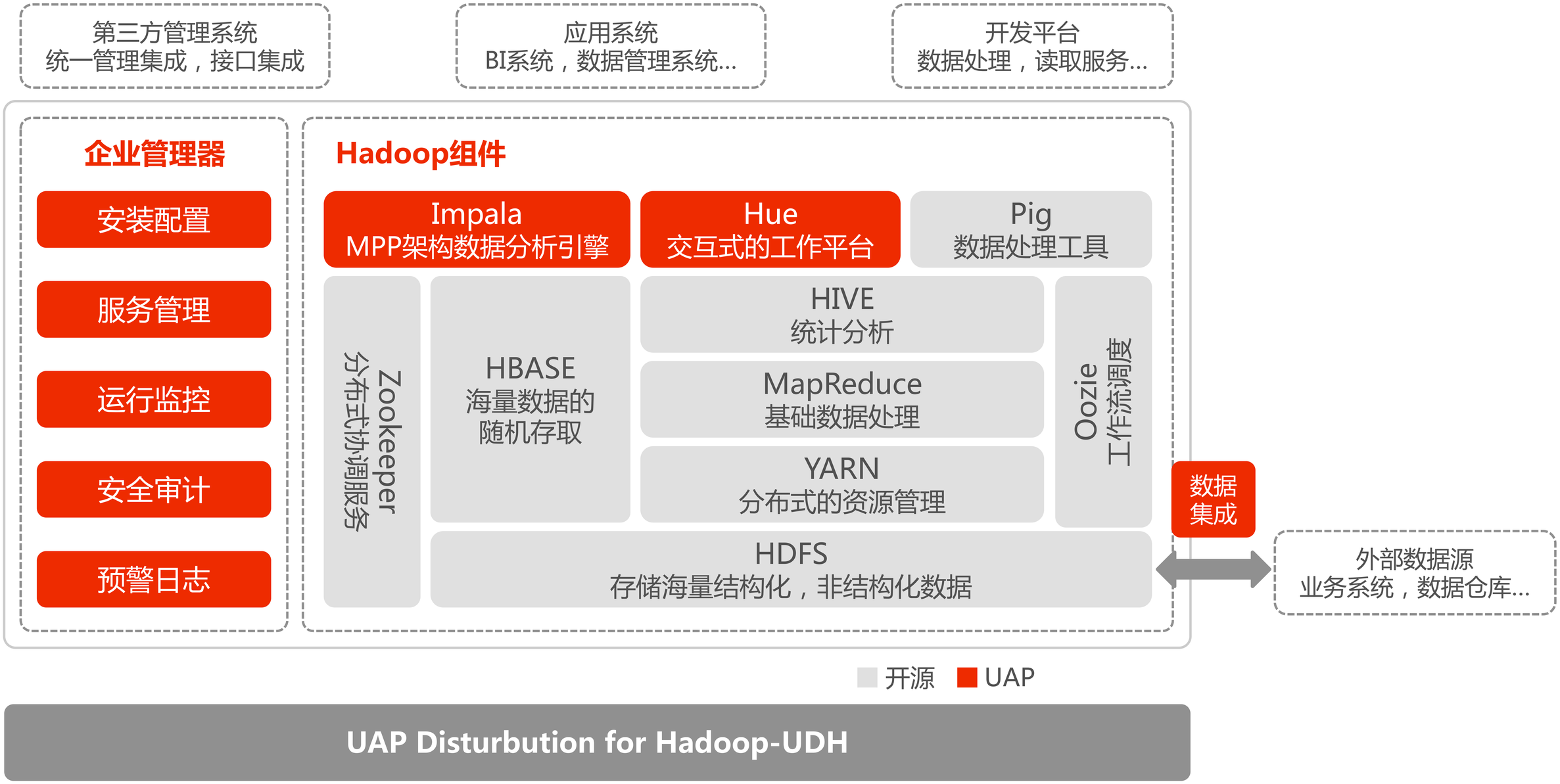
群集1TB/小时

yonyou Network Technology Co., Ltd..

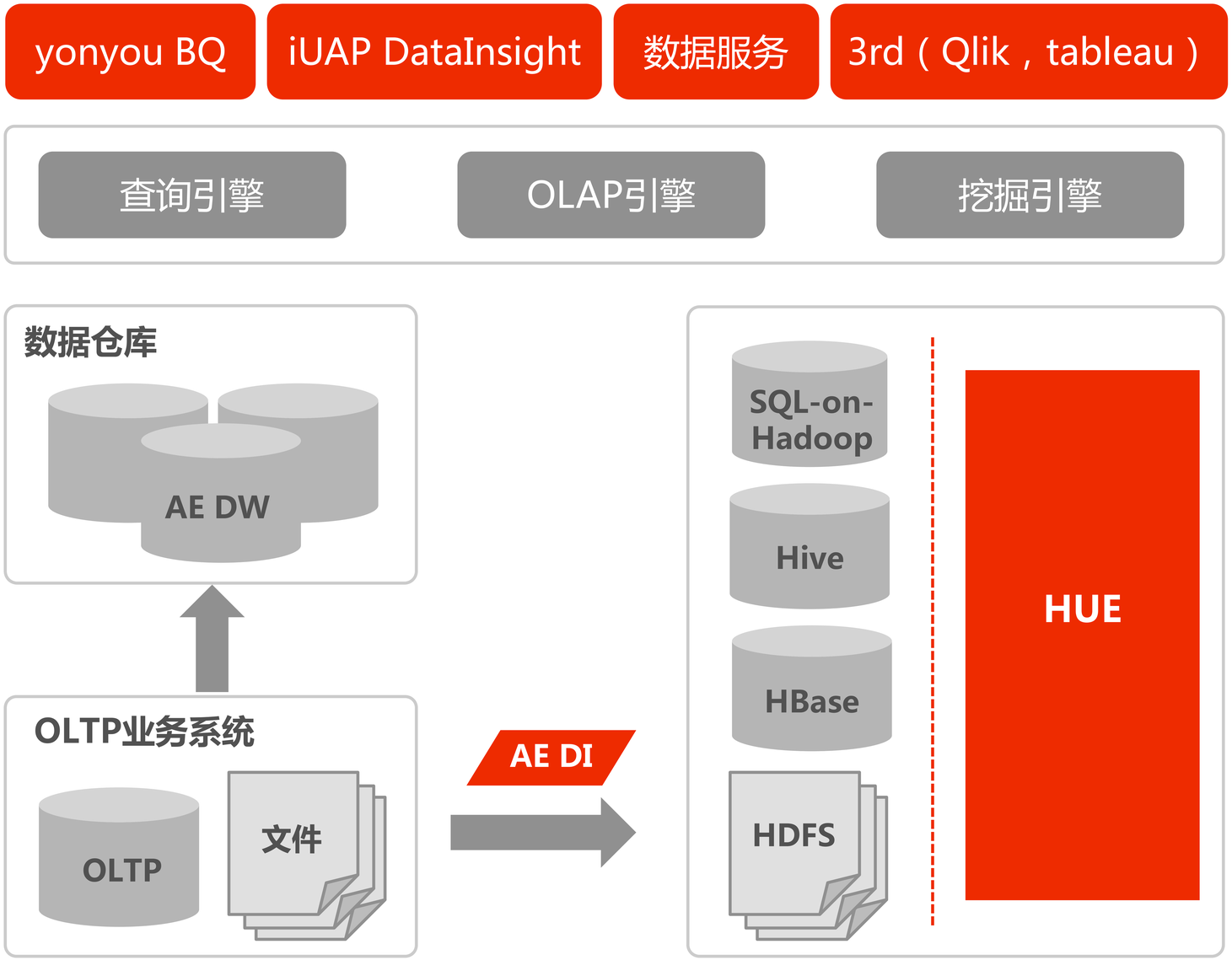
iUAP AE(结构化数据引擎）性能提升效果



大数据引擎UDH



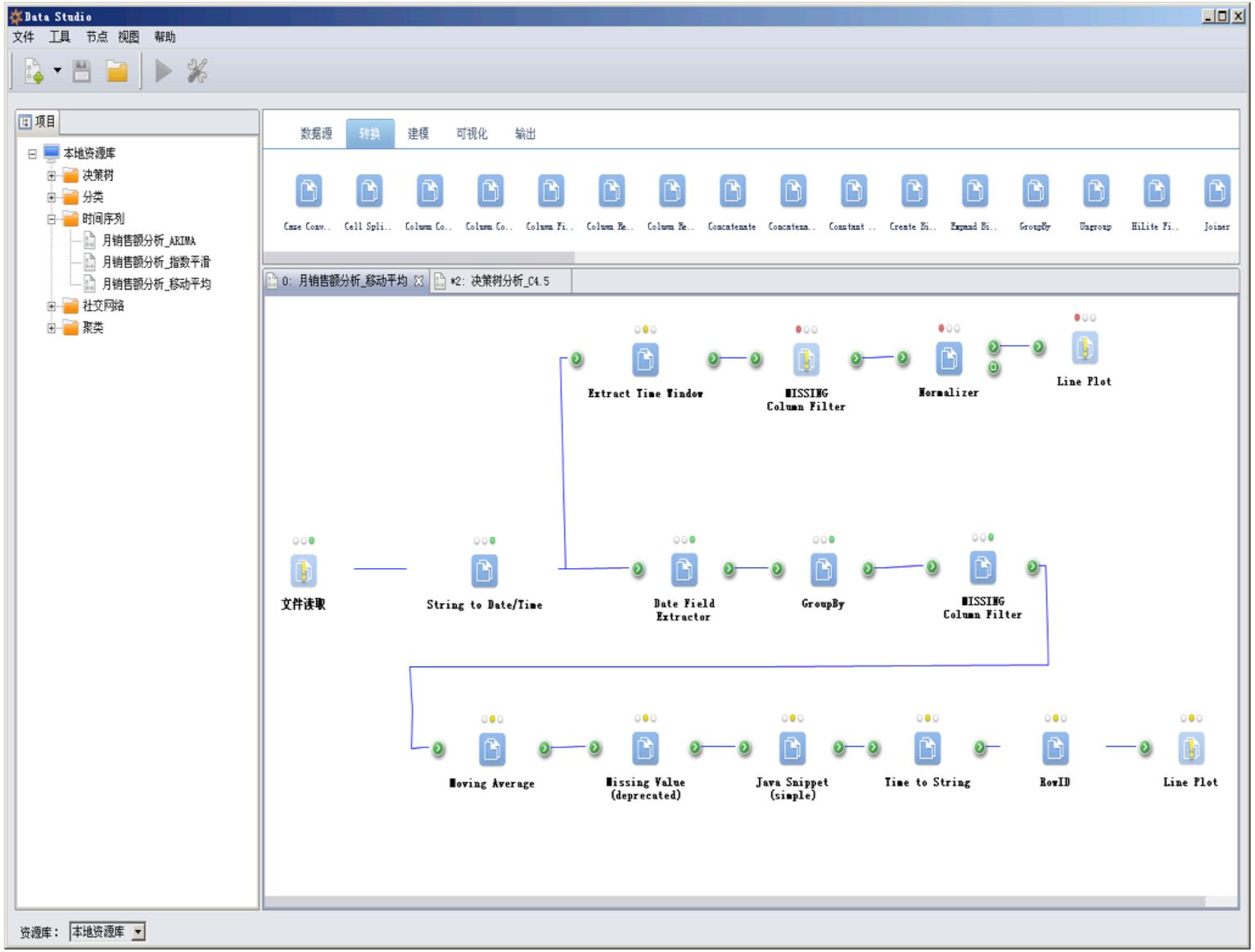
UAP Disturbution for Hadoop-UDH



海量数据的深度分析

- 面向结构化和非结构化数据处理的完整解决方案
- 利用UDH处理组件的横向扩展，并行处理的能力提升大数据处理的效率
- 统一的SQL查询引擎，屏蔽底层数据架构差异
- 结合yonyou BQ，iUAP DataInsight或第三方工具进行海量数据的探索，实现深度分析

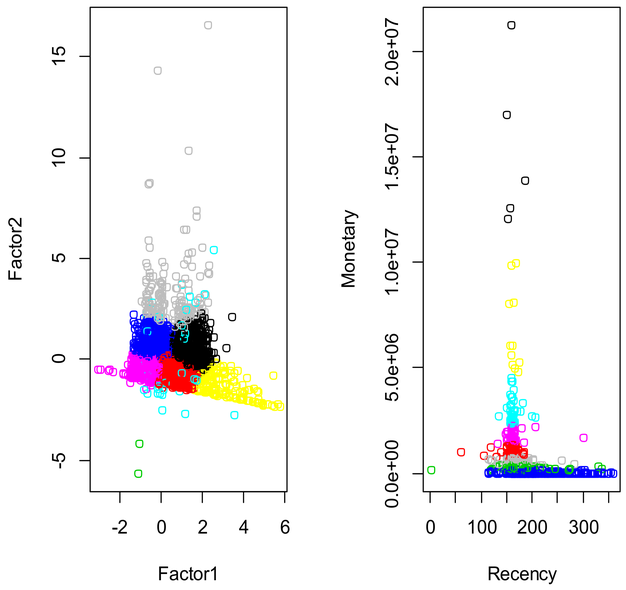
数据挖掘DataInsight

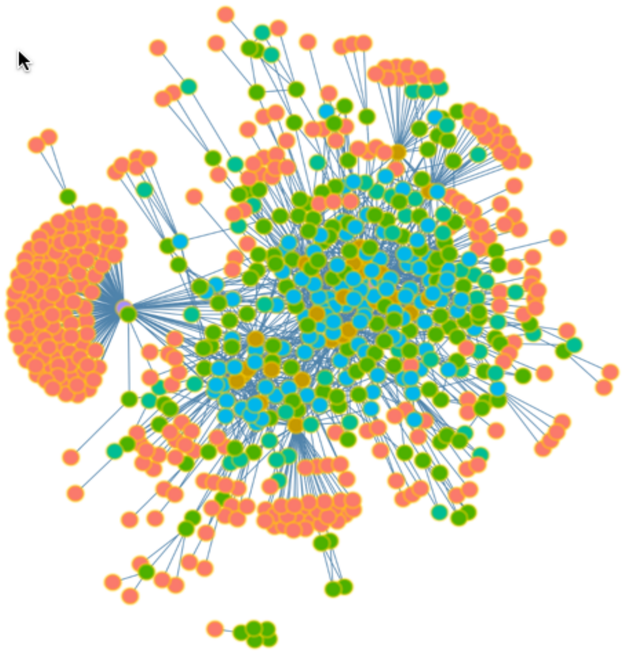


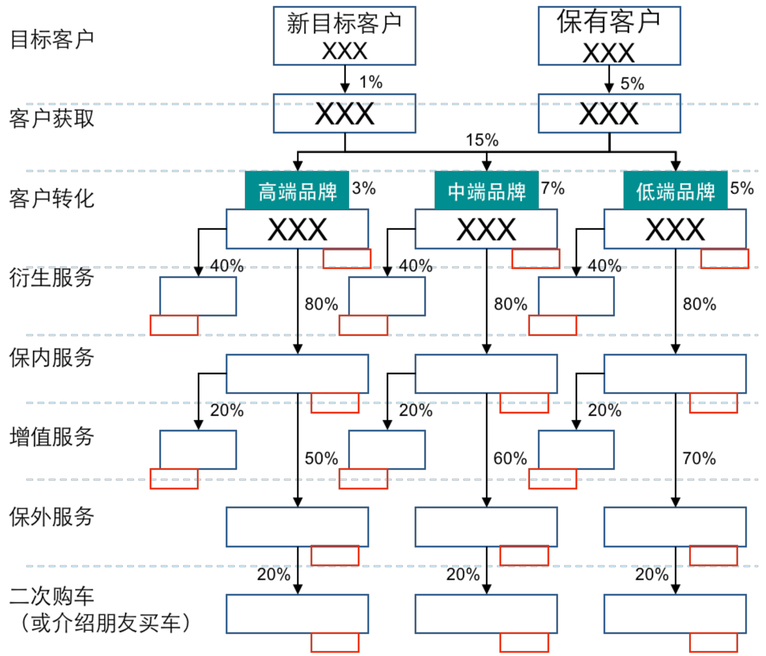
挖掘预测平台

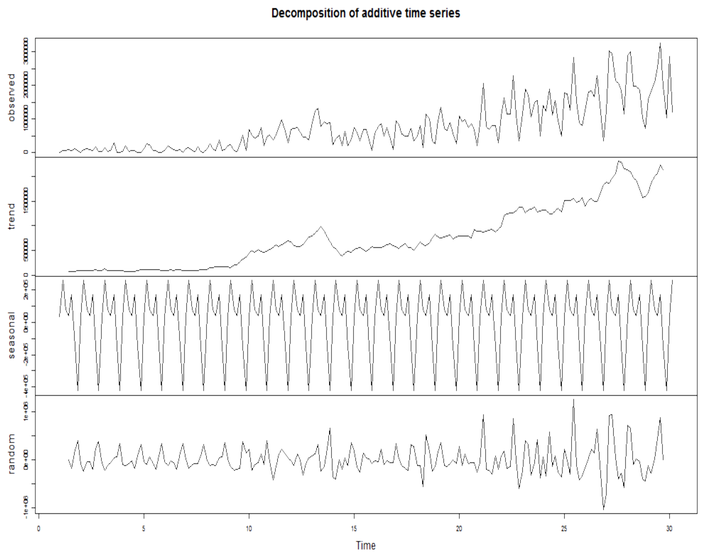
- 大量预处理组件，可视化拖拽设计，支持完整的挖掘建模流程
- 内置回归，分类，聚类，关联，时间序列等常见算法和分析模型，支持分布式并行计算，高性能的计算引擎
- 集成定制的R引擎，3-5倍的性能提升。
- 丰富的数据可视化组件和部署方式

丰富的数据可视化

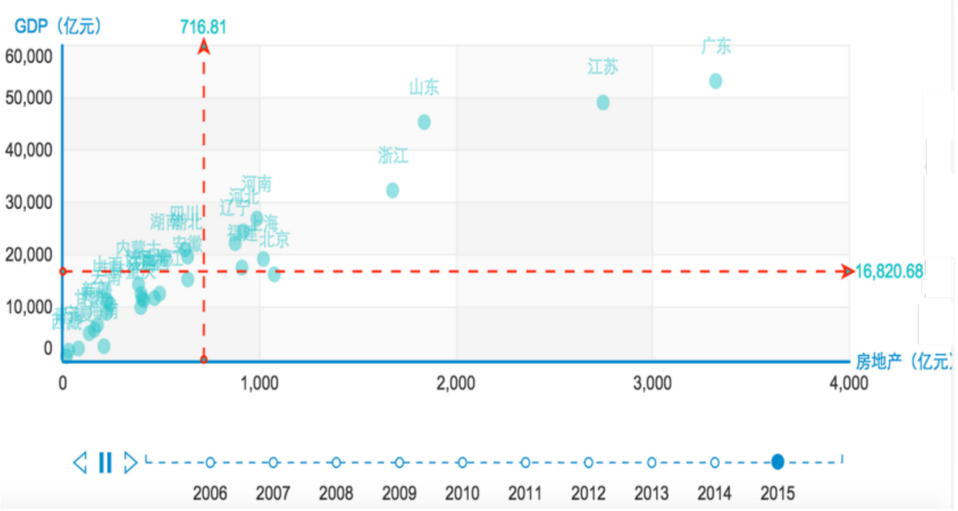

Factor2
Factor1
Monetary
Recency




目标客户
新目标客户
XXX
保有客户
XXX
1%
5%
客户获取
XXX
XXX
15%
客户转化
高端品牌
3%
中端品牌
7%
低端品牌
5%
XXX
XXX
XXX
衍生服务
40%
40%
40%
80%
80%
80%
保内服务
20%
20%
20%
增值服务
50%
60%
70%
保外服务
20%
20%
20%
二次购车
（或介绍朋友买车）


Decomposition of additive time series
observed
trend
seasonal
random
Time

红秀网络
金威海鲜
打豆豆
长理乐购
长理湘味馆
太和公鱼粉
九九鸭脖
汉唐华都酒店
酒醉粥饱
大排骨
五谷根竹筒饭
木桶饭
麻辣香锅
西门移动营业厅
绝味鸭脖
湘村发现
东门星期吧奶茶咖啡
湘村发现家菜馆
青春年华会所
思乐客快餐连锁店理工店
西门老食客砂锅煲
学友家菜馆
爱情公寓
铭达餐馆
抹茶啊hian
台湾水晶锅
心点小滋
卤大师
水木时光公寓酒店
麦乐司
0731飞速外卖
鑫蔓酒吧
湖园大碗饭
奥乐滋汉堡店
甜品小姐一号店
颐源堂足道养生会所


GDP（亿元）
716.81
60,000
50,000
40,000
30,000
20,000
10,000
0
16,820.68
0
1,000
2,000
3,000
4,000
房地产（亿元）
广东
江苏
山东
浙江
河南
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

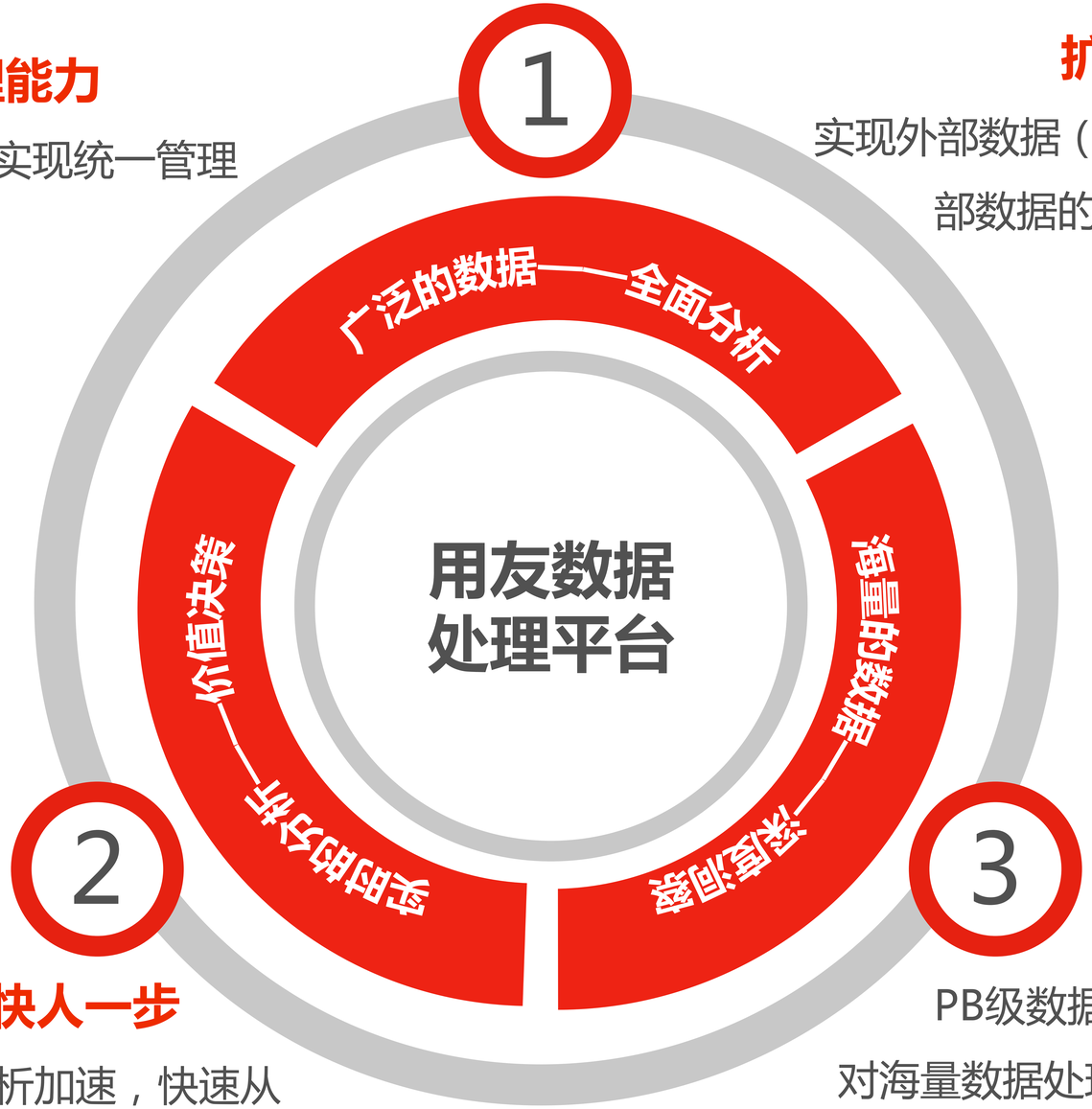
商业价值

提升企业的数据管理能力

多业务系统，不同数据类型数据，实现统一管理

扩大企业数据的边界

实现外部数据（客户，宏观，对手，舆情等）与内部数据的综合分析



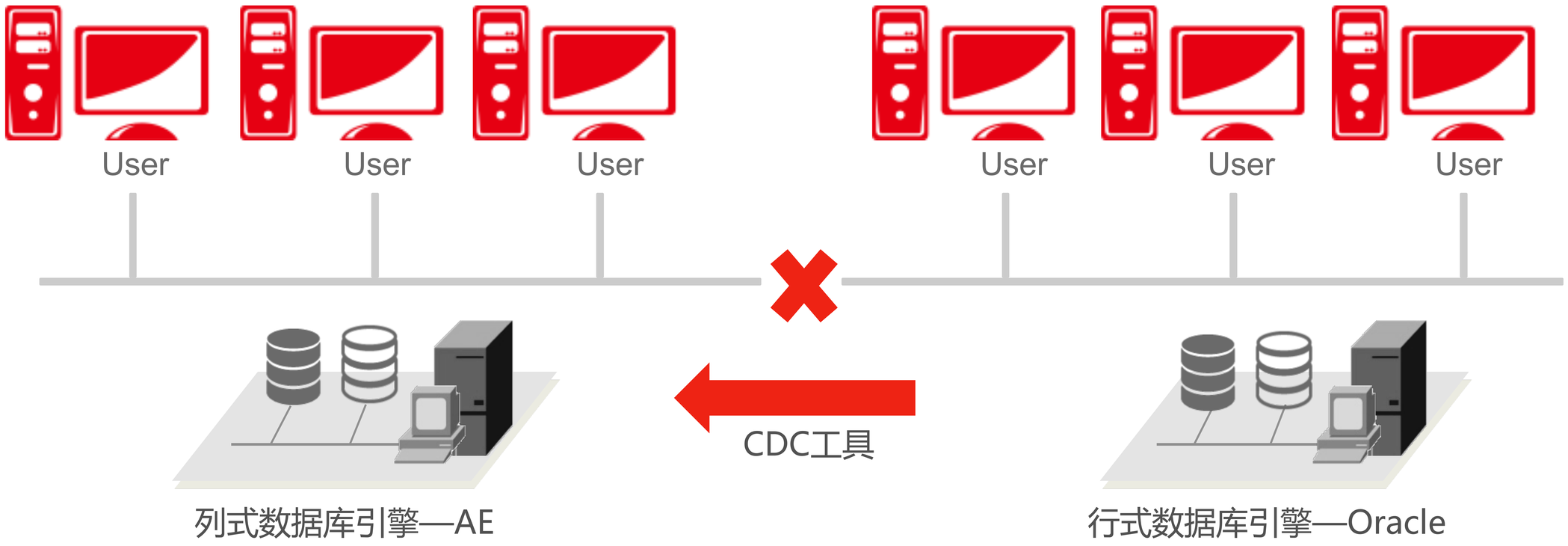
实时掌握经营状态，快人一步

秒级延迟的数据同步，百倍的分析加速，快速从海量信数据中提取有价值信息

洞察数据的价值

PB级数据存储，计算，分析能力，轻松应对海量数据处理，提供高效的数据挖掘引擎

应用案例（一）：即席查询应用加速



解决方案

- 通过CDC工具，实时增量同步数据到列式数据库引擎，剥离应用场景
- 利用采用高压缩比，并行处理，智能索引的列式数据库引擎AE，加速查询分析

即席查询应用特点

- 根据条件动态生成查询SQL，非预先定制
- 通常需要消耗大量的磁盘I/O导致整个业务系统性能下降

应用案例（一）：优化结果

SQL	期间	AE（优化前）	ORACLE（优化前）	数据结果
SQL（1）	查一个月	10.52 秒	60秒	均为1437行
	查七个月	29.17秒	20分钟以上	均为9400行
	查2010全年至今	40.15秒	未测试完成	18145行
SQL（2）	查一个月	35.21 秒	228秒	均为 236行
	查七个月	46.15秒	20分钟以上	均为1468行
	查2010全年至今	26.58秒	未测试完成	2801行
SQL（3）	查一个月	82.2秒	87.906秒	均为126485行
	查七个月	132.27秒	20分钟以上	均为728740行
	查2010全年至今	236秒	未测试完成	1292748行
SQL（4）	查一个月	120秒	66.687秒	均为63行
	查七个月	251秒	20分钟以上	均为398行
	查2010全年至今	5分16秒	未测试完成	740行

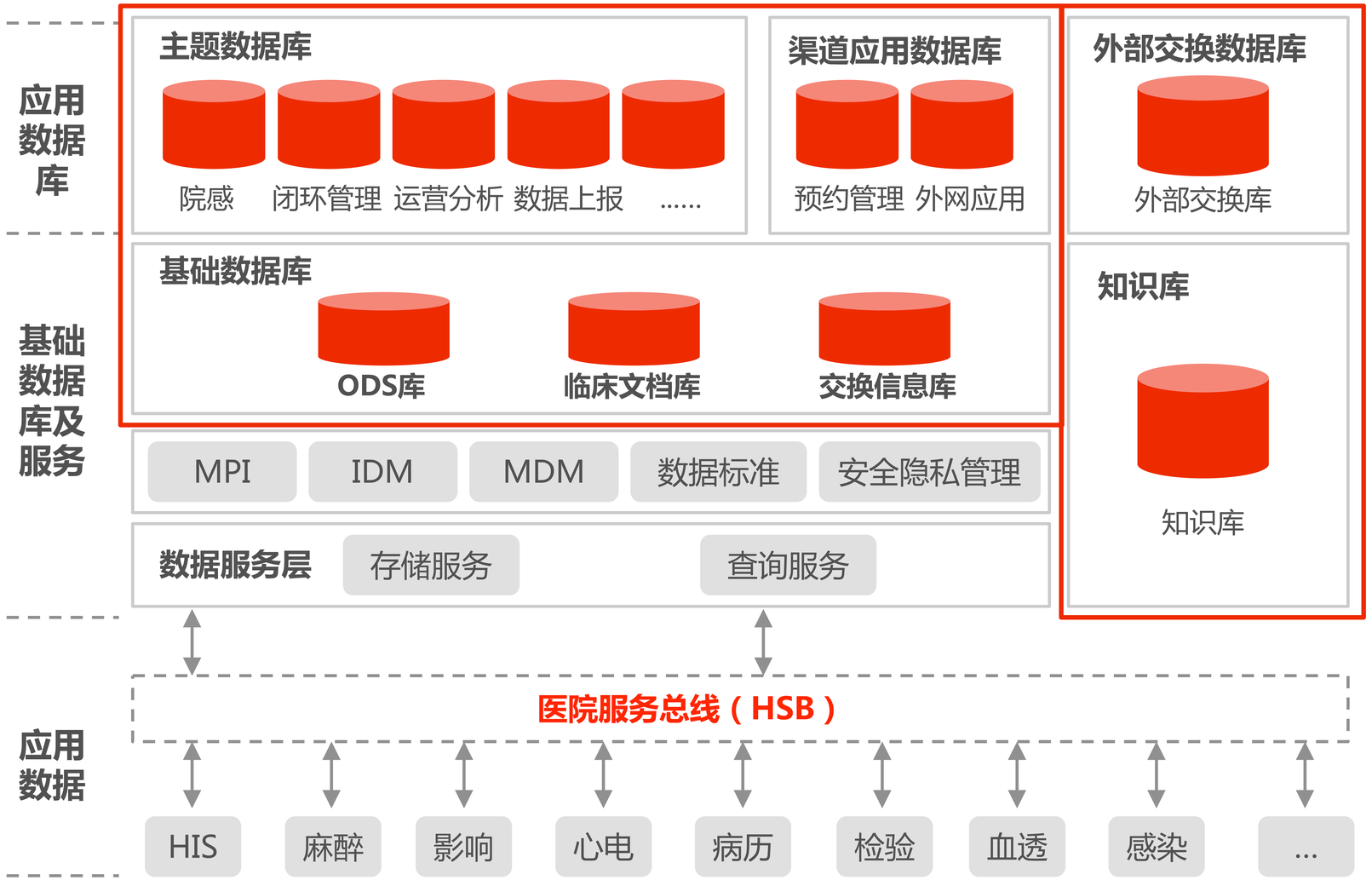
表明	数据记录行数
crm_customer	147113
pub_item_master	5406098
pub_item_type_explode	198695
sal_classify_customer	4324
sal_consign_detail	9197164
sal_consign_head	4680662
sal_org_customer_access_v	152541214
sal_return	96017
sal_return_detail	346041
sam_users	22493

单表1亿5千万行

结果分析

- 小范围查询期间场景性能相当
- 大范围查询期间（超过半年）AE优势明显
- 更大范围期间oracle完全无法支撑，AE能够从容应付

应用案例（二）：医疗诊断数据中心解决方案



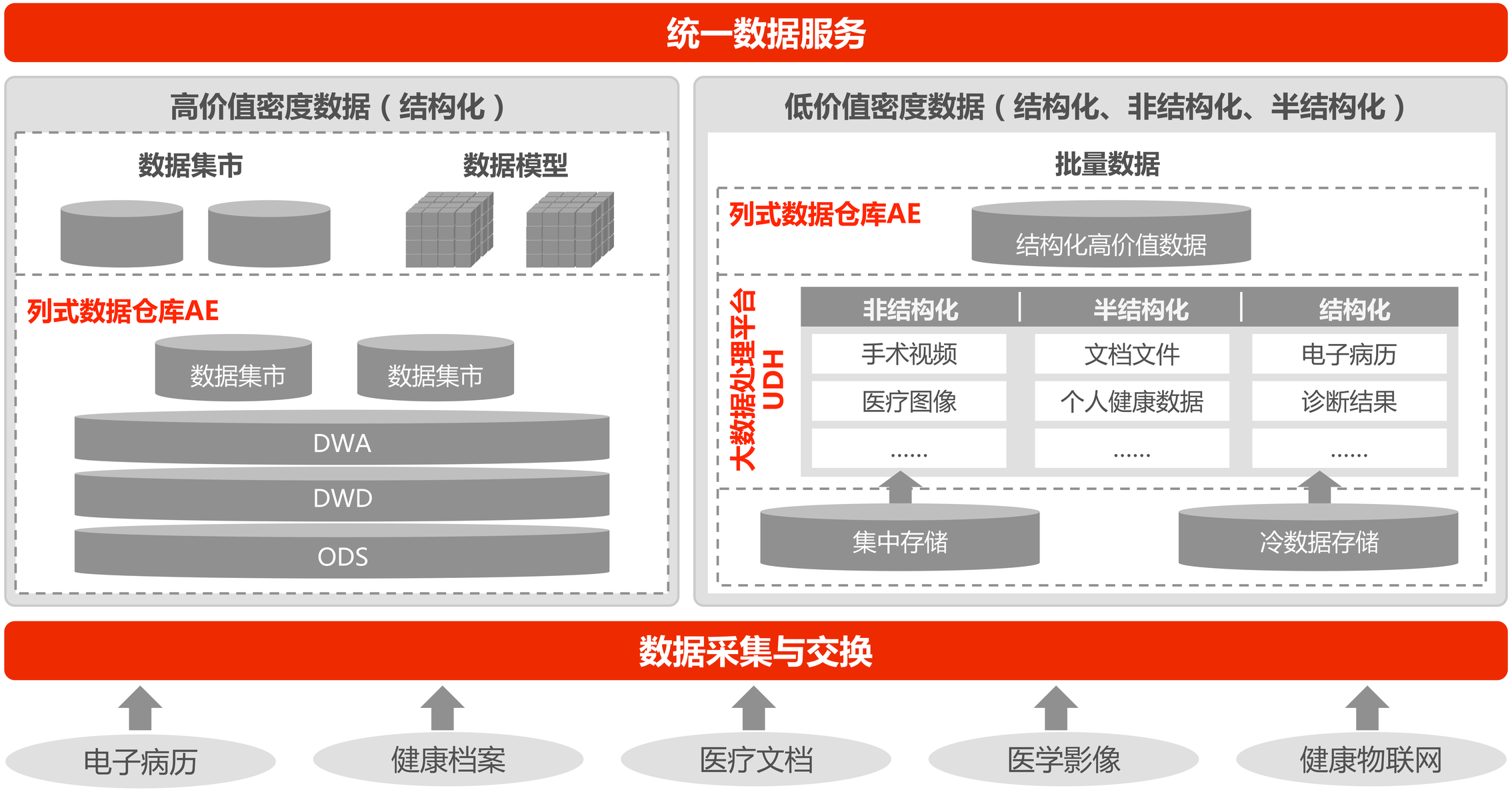
系统需求

- 结构化数据和非结构化数据（XML，JSON，文档等）混合存储，统一管理
- 同时支持业务应用与数据分析
- 支持大量读操作和少量直接写操作

解决方案

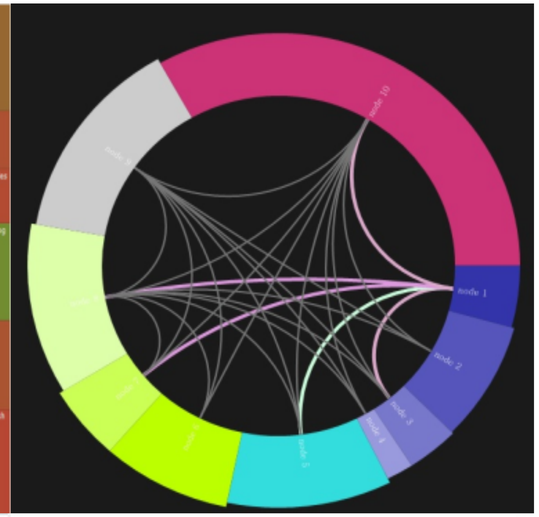
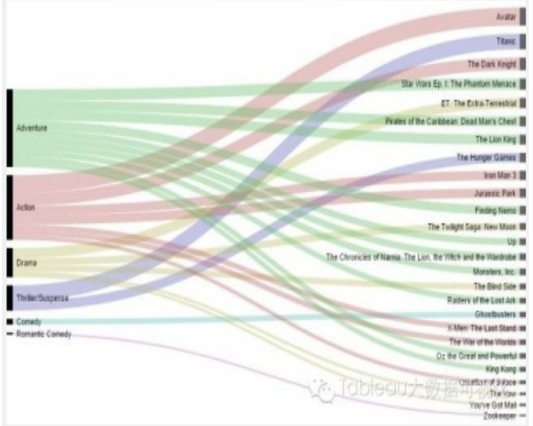
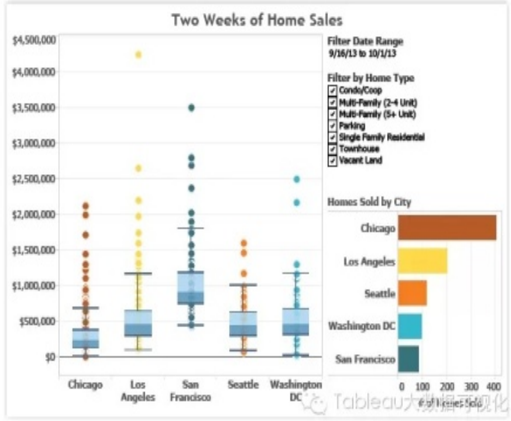
- 根据数据和业务场景，采用列式数据仓库AE与大数据处理引擎UDH的混合架构
- 通过统一查询层屏蔽底层数据差异

应用案例（二）：数据架构


统一数据服务
高价值密度数据（结构化）
数据集市
数据模型
列式数据仓库AE
数据集市
数据集市
DWA
DWD
ODS
低价值密度数据（结构化、非结构化、半结构化）
批量数据
列式数据仓库AE
结构化高价值数据
大数据处理平台
UDH
非结构化
半结构化
结构化
手术视频
文档文件
电子病历
医疗图像
个人健康数据
诊断结果
……
……
……
集中存储
冷数据存储
数据采集与交换
电子病历
健康档案
医疗文档
医学影像
健康物联网

应用案例（三）：行业分析

经济数据 | 行业指标数据 | 企业指标数据 | 行业资讯/动态 | 行业法律法规 | 行业报告

指标数据体系

可视化控件区 | 选择/切换可视化控件 | 可视化工具服务输出