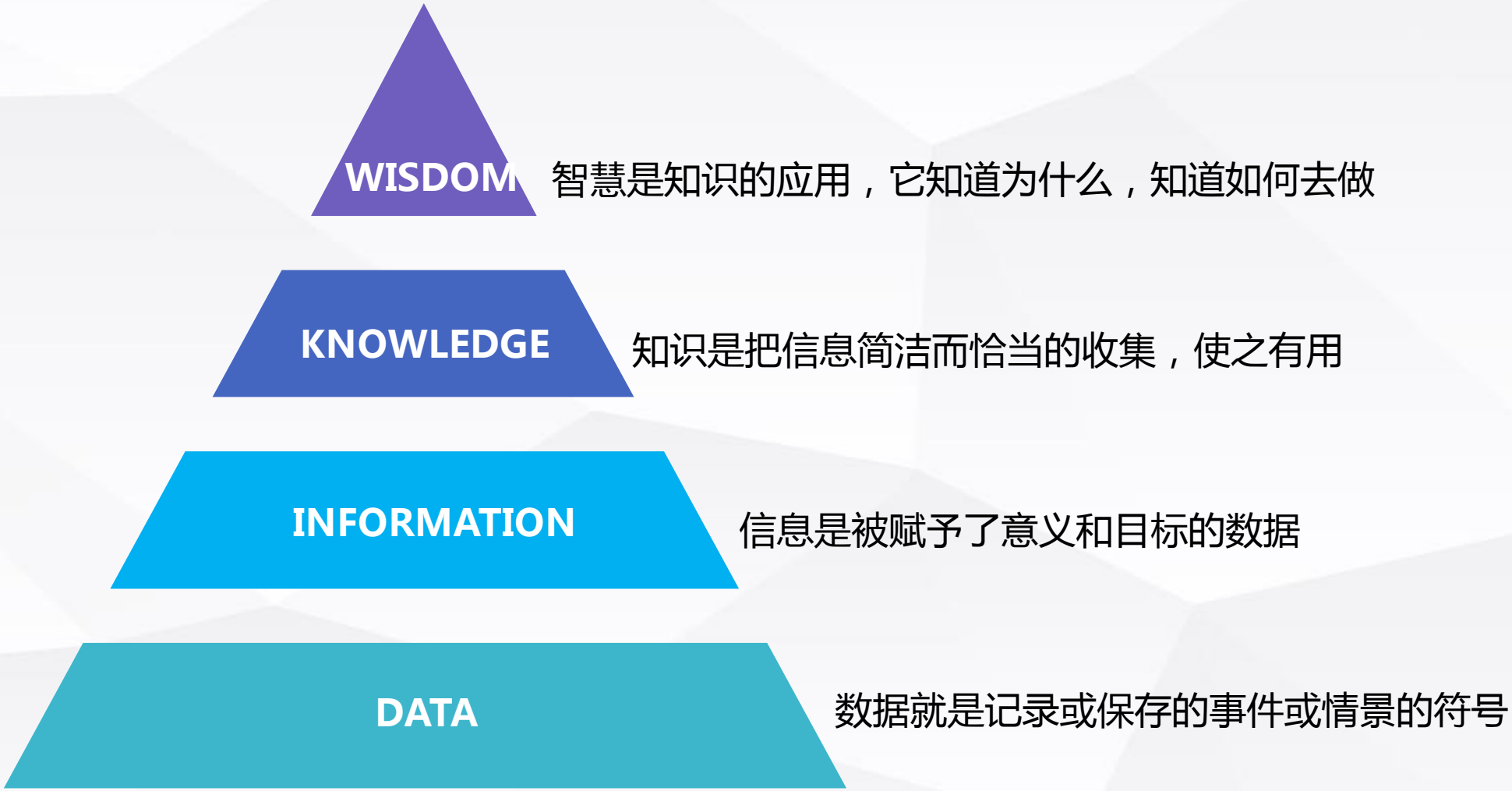
探寻数据价值之路

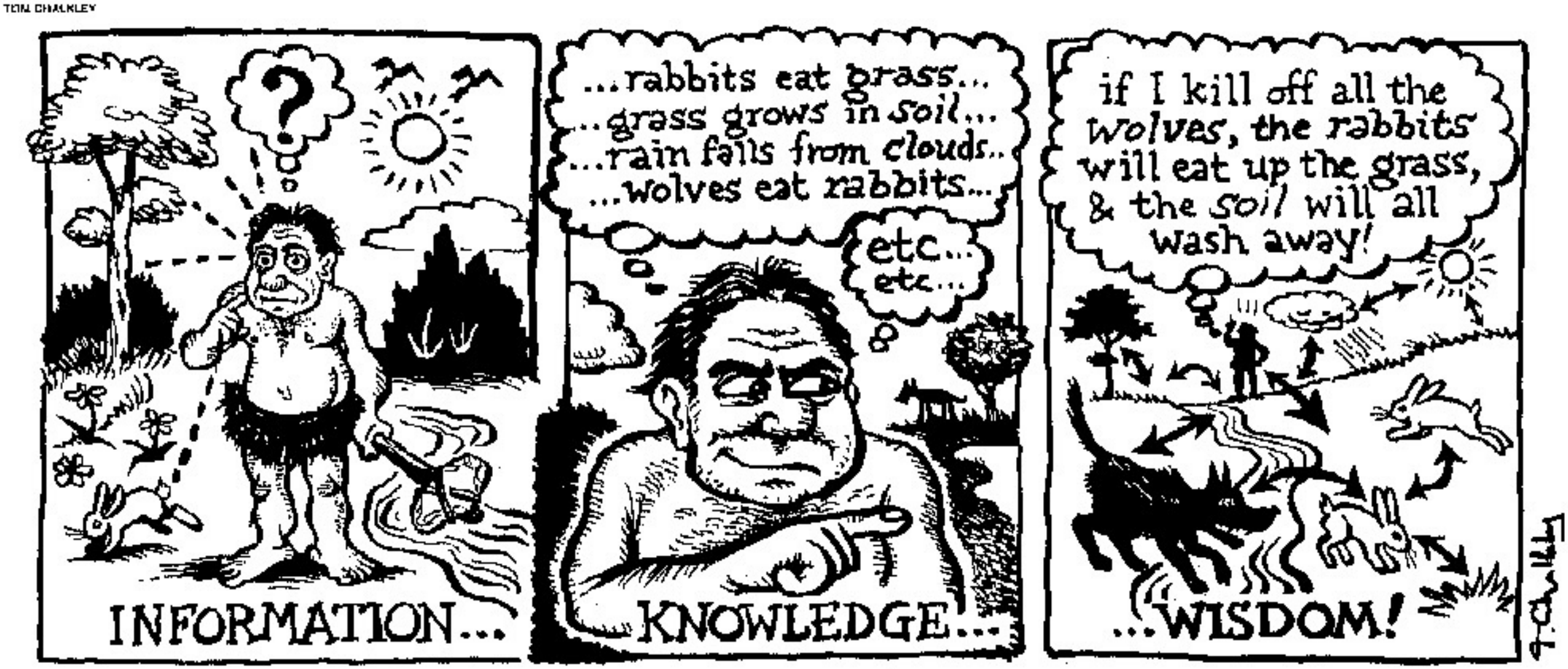
Data Cloud智能数据平台的实践

主讲人：王旭鹏

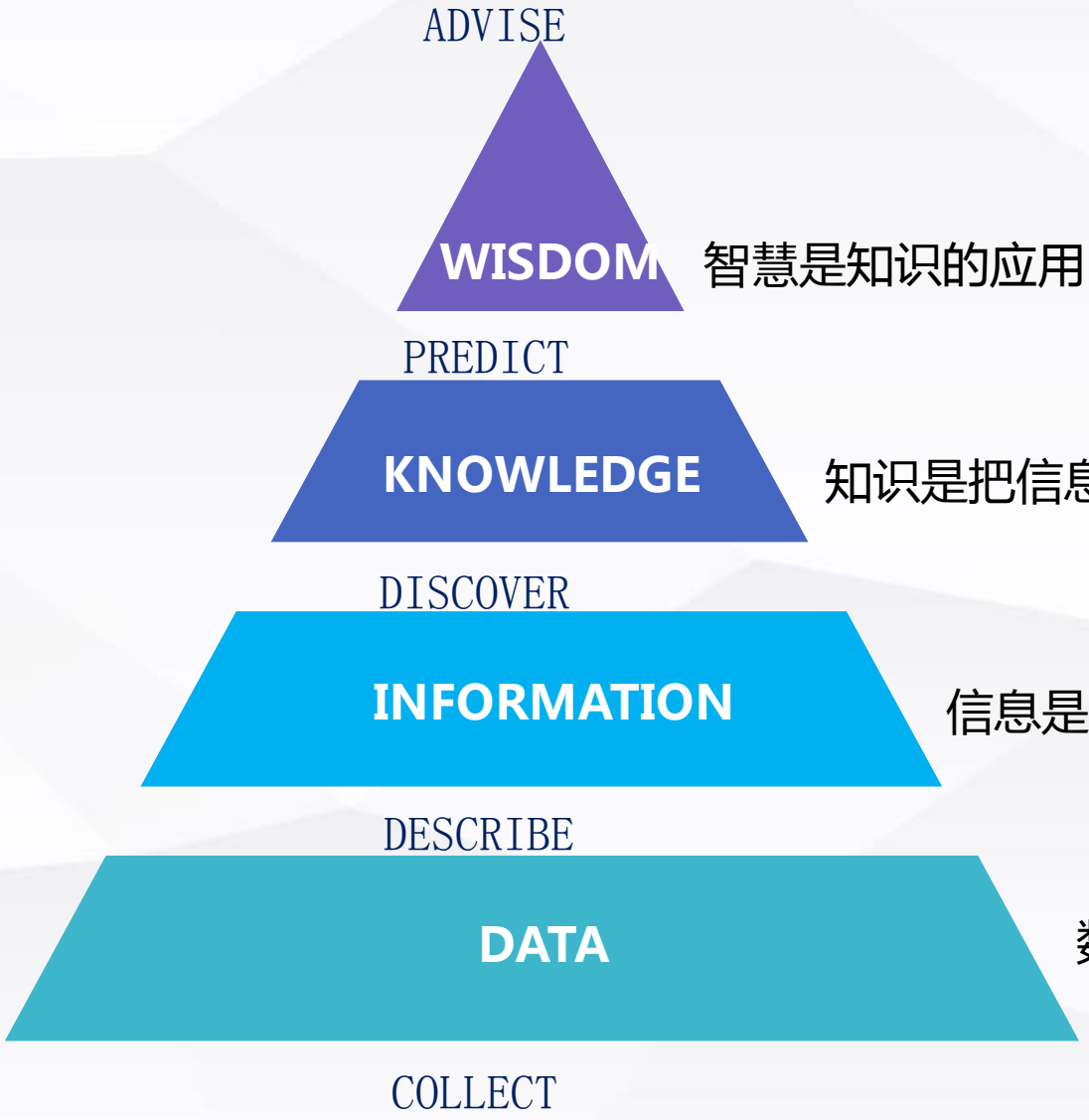
DataCloud 数据价值模型



DataCloud 数据价值模型



DataCloud 数据价值模型



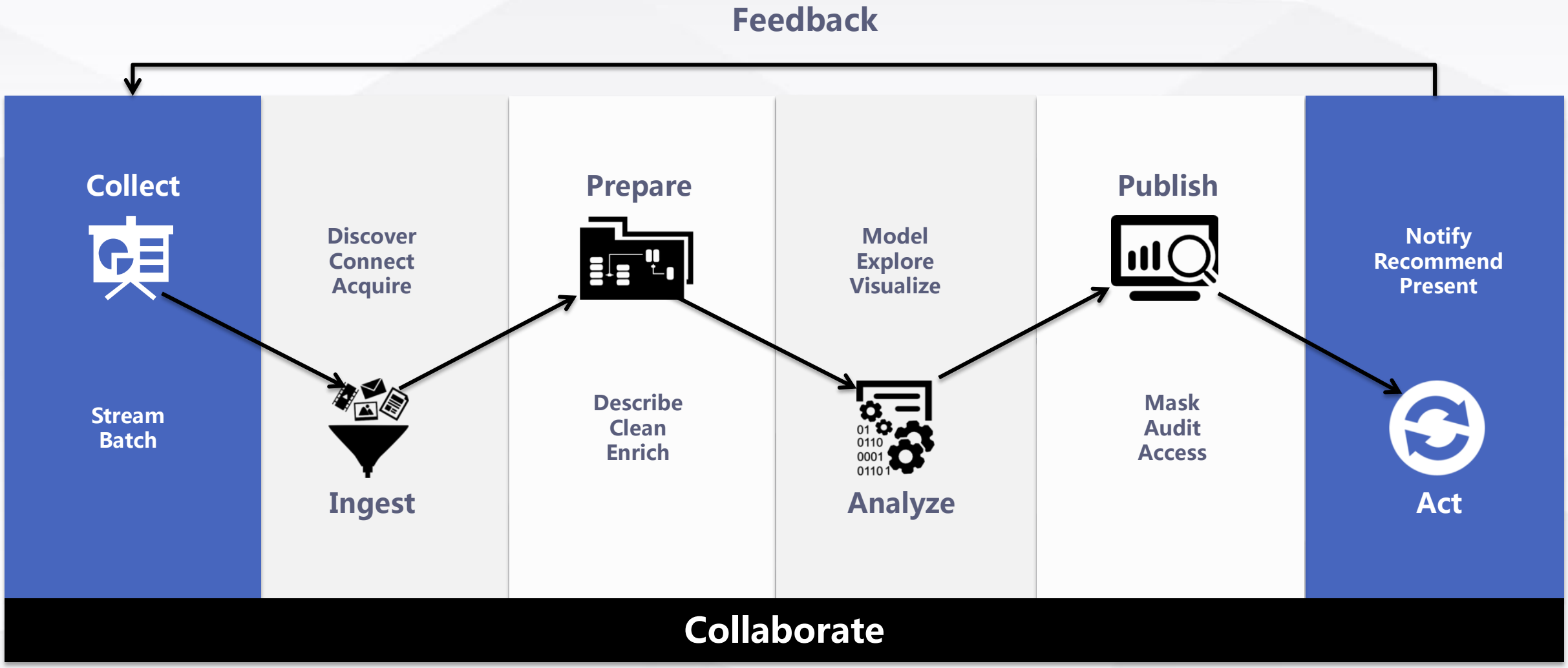
智慧是知识的应用，它知道为什么，知道如何去做

知识是把信息简洁而恰当的收集，使之有用

信息是被赋予了意义和目标的数据

数据就是记录或保存的事件或情景的符号

DataCloud 数据处理模型



数据

Data

多源数据填补核心资产

平台

Platform

健壮平台支撑掌控能力

应用

Application

实用的应用打开思维脑洞

大数据价值体现

平台需要解决的问题

如何把数据汇集起来？

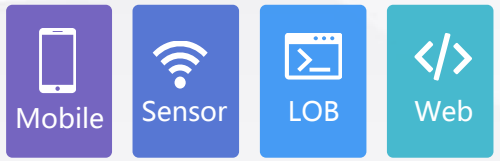
如何把数据管理起来？

如何把数据用起来？

DataCloud 面临的数据量级

覆盖体量	**40亿** 累计设备 Device Coverage	**2.5亿** 日活设备 Daily Active Devices	**6.5亿** 月活设备 Monthly Active Devices
数据吞吐	**14T** 每天新增数据 Daily Ingested Data	**34亿** 每天交互会话 Daily Sessions	**370亿** 每天处理事件 Daily Events

DataCloud 功能架构



内部自有数据

外部自有数据

第三方数据

数据接入

- 数据连接器
- 数据同步器
- 数据源管理

数据目录

- 数据概要
- 数据安全
- 数据质量
- 数据标签
- 元数据管理

数据工厂

- 可视化加工
- 交互式探索
- 智能调度
- 智能清洗规则
- 自定义算子

数据发布

- 敏捷发布
- 访问审计
- 智能脱敏

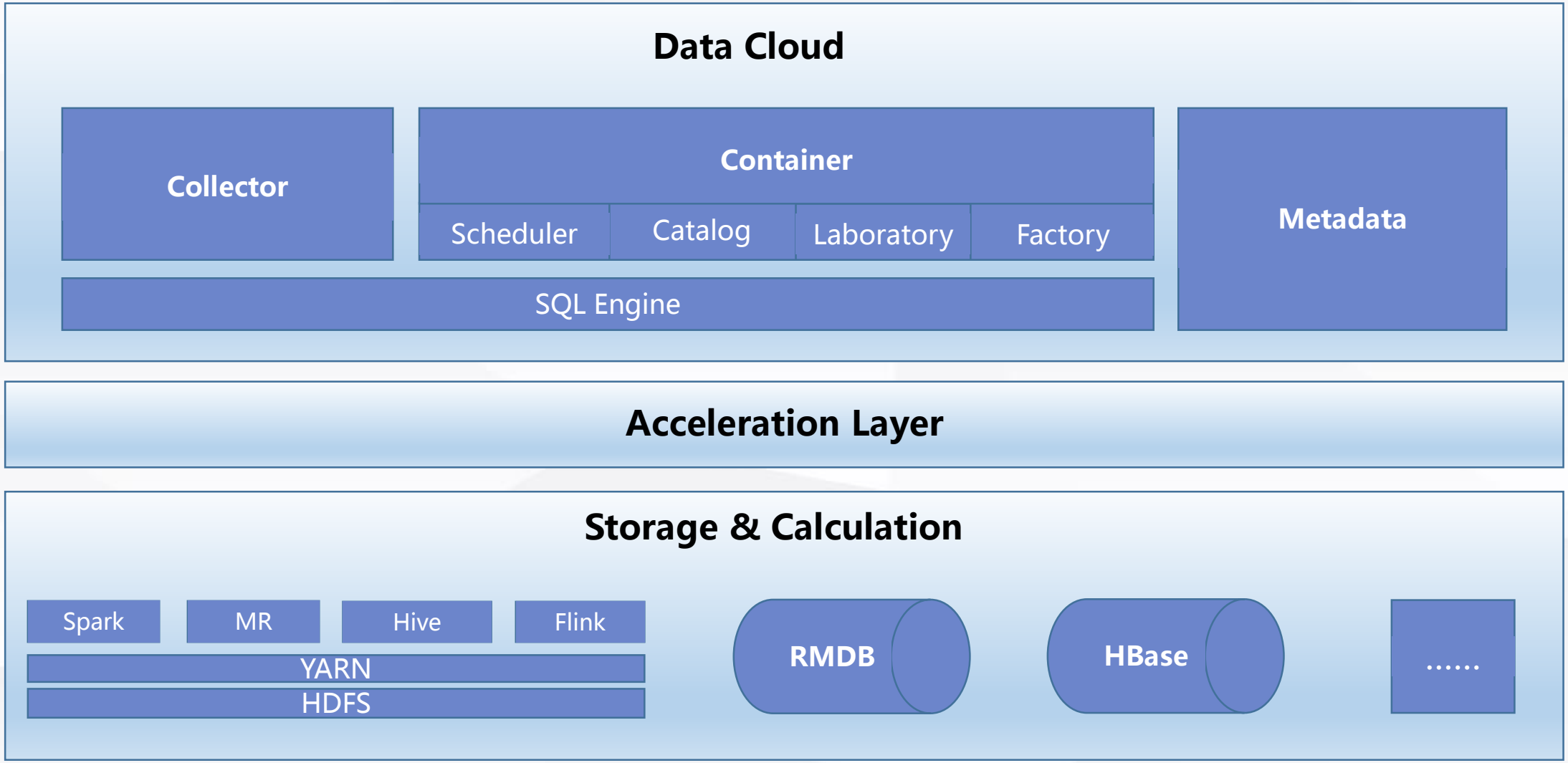
通用统计分析

移动广告监测

移动观象台

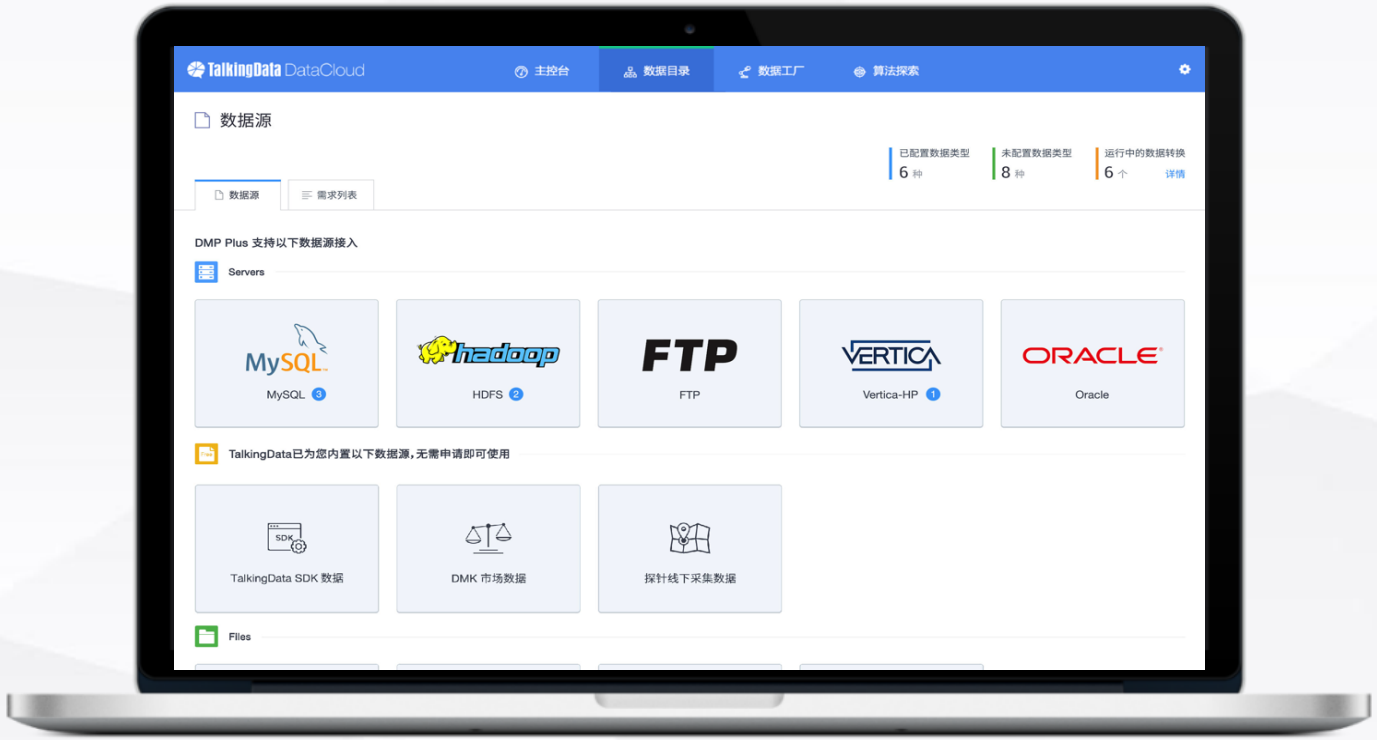
更多

DataCloud 系统架构



DataCloud 数据接入特点

- 支持多种数据源的接入
- 支持TD数据市场数据接入
- 支持外挂、同步数据源



DataCloud

数据同步的实践

- **同步数据（实时性低、最终一致性）**

  查询数据 -> 可能需要做一些转换 -> 然后插入数据

- **数据源发生变化**

  增加UpdateTime、IsDeleted字段

- **选取数据的方法**

```
SELECT * FROM User WHERE UpdateTime >= DATE_SUB(Now(),INTERVAL 5 MINUTE)
SELECT * FROM User WHERE UpdateTime >= #LastRunTime#
SELECT * FROM User WHERE UpdateTime >= #LastRowTime#
```

DataCloud 数据管理特点

- 便捷的元数据管理
- 自动生成数据世系拓扑
- 自定义的数据资产视图
- 数据质量管理

DataCloud

数据质量评估标准

及时性

数据从产生到可用是否延时过长。

完整性

数据或元数据是否存在缺失。

一致性

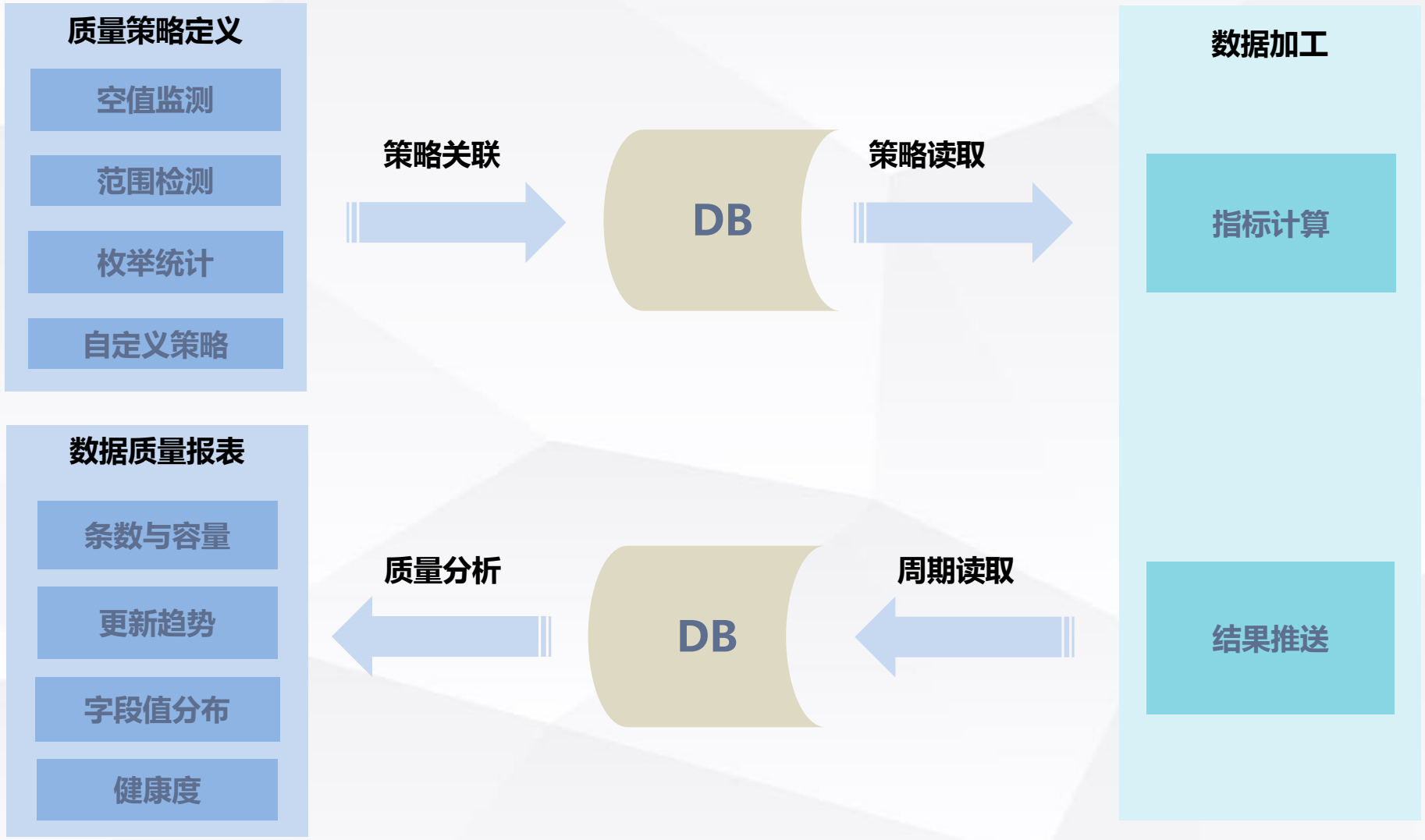
数据是否遵循了统一的规范，是否符合逻辑。

准确性

数据是否存在异常或错误。

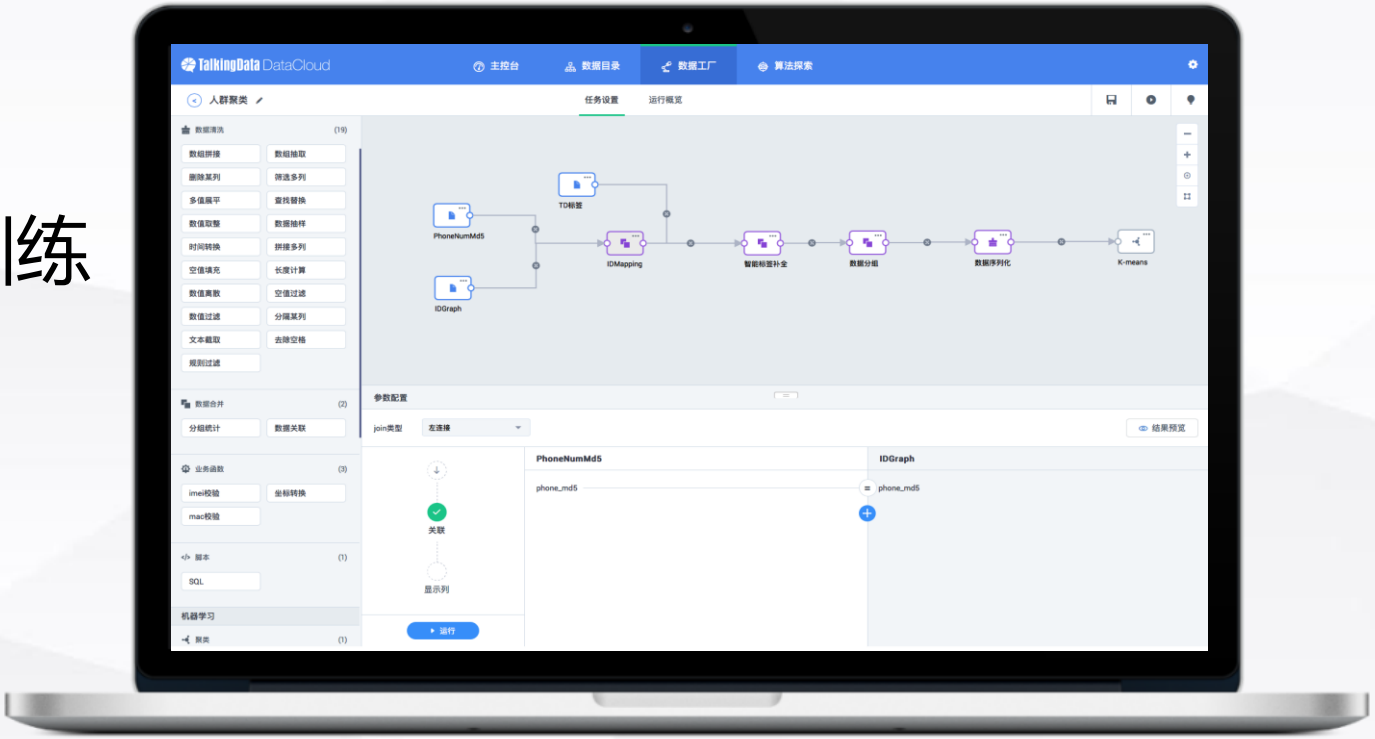
DataCloud

数据质量加工流程



DataCloud 数据加工特点

- 内置数据清洗加工算子
- 多种机器学习算法、模型训练
- 自定义算子、算法支持
- 交互式数据加工探索



人群聚类

任务设置 运行概览

数据集

源数据 目标数据

数据转换

数据清洗 (19)

- 数组拼接
- 数组抽取
- 删除某列
- 筛选多列
- 多值展平
- 查找替换
- 数值取整
- 数据抽样
- 时间转换
- 拼接多列
- 空值填充
- 长度计算
- 数值离散
- 空值过滤
- 数值过滤
- 分隔某列
- 文本截取
- 去除空格
- 规则过滤

数据合并 (2)

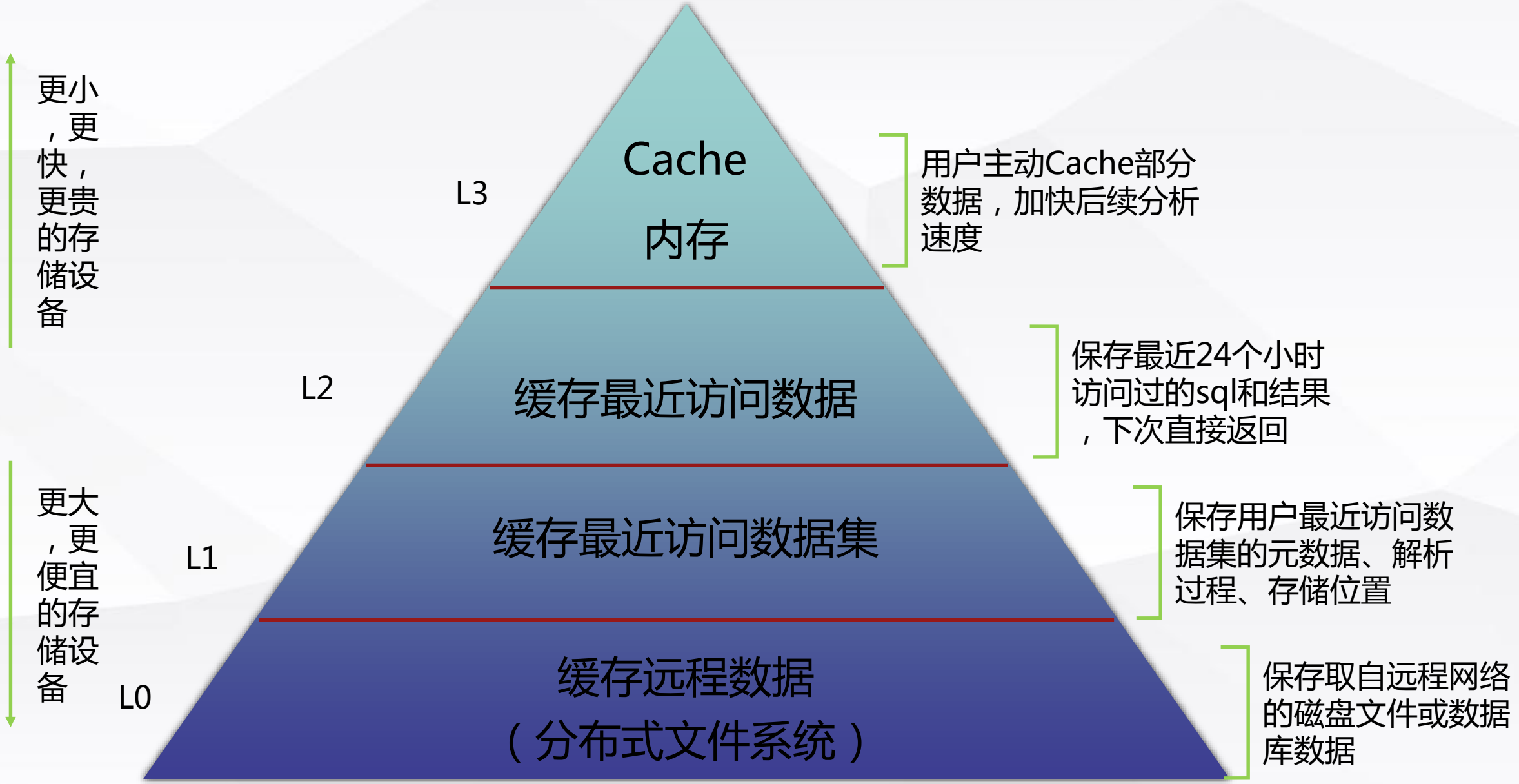
- 分组统计
- 数据关联

业务函数 (3)

- imei校验
- 坐标转换

参数配置

DataCloud 数据探索提速机制



DataCloud

更多挑战

- 不同角色之间的协同工作
- 屏蔽底层存储版本依赖带来的问题
- 冷热数据的智能存储分配

T11 2016

THANKS

Let’s rock data together!

hr@tendcloud.com

TalkingData