DTCC 2014中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2014
大数据技术探索和价值发现

数据治理
大数据平台设计

万振龙

议程

- 数据治理的背景和现状
- 数据治理策略
- 元数据管理
- 主数据管理
- 数据质量管理
- 大数据平台设计

数据治理背景

数据治理

1. 大数据时代凸现数据重要性
2. 数据治理是大数据的基础
3. 信息孤岛现象严重
4. 数据质量问题严重
5. 数据应用未得到有效管理
6. 数据安全问题日益严峻

www.themegallery.com

1. 意识到了问题的严重
2. “维持”代替“管理”
3. 历史“包袱”沉重
4. 相关方利益交织，协调困难
5. 方案规划容易，落地困难
6. 过度依赖技术工具
7. 对于数据没有明确区分

议程

- 数据治理的背景和现状
- 数据治理策略
- 元数据管理
- 主数据管理
- 数据质量管理
- 大数据平台设计

数据治理要素

组织

Organization

流程、活动与机制

Process & Activities & Mechanism

技术平台与工具

Platform & Tools

计划、制度与标准规范

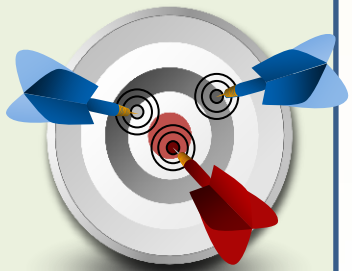
Plan & Rule & Standards

数据治理策略

获得支持

引入外援

找到“痛点”



确定“起点”

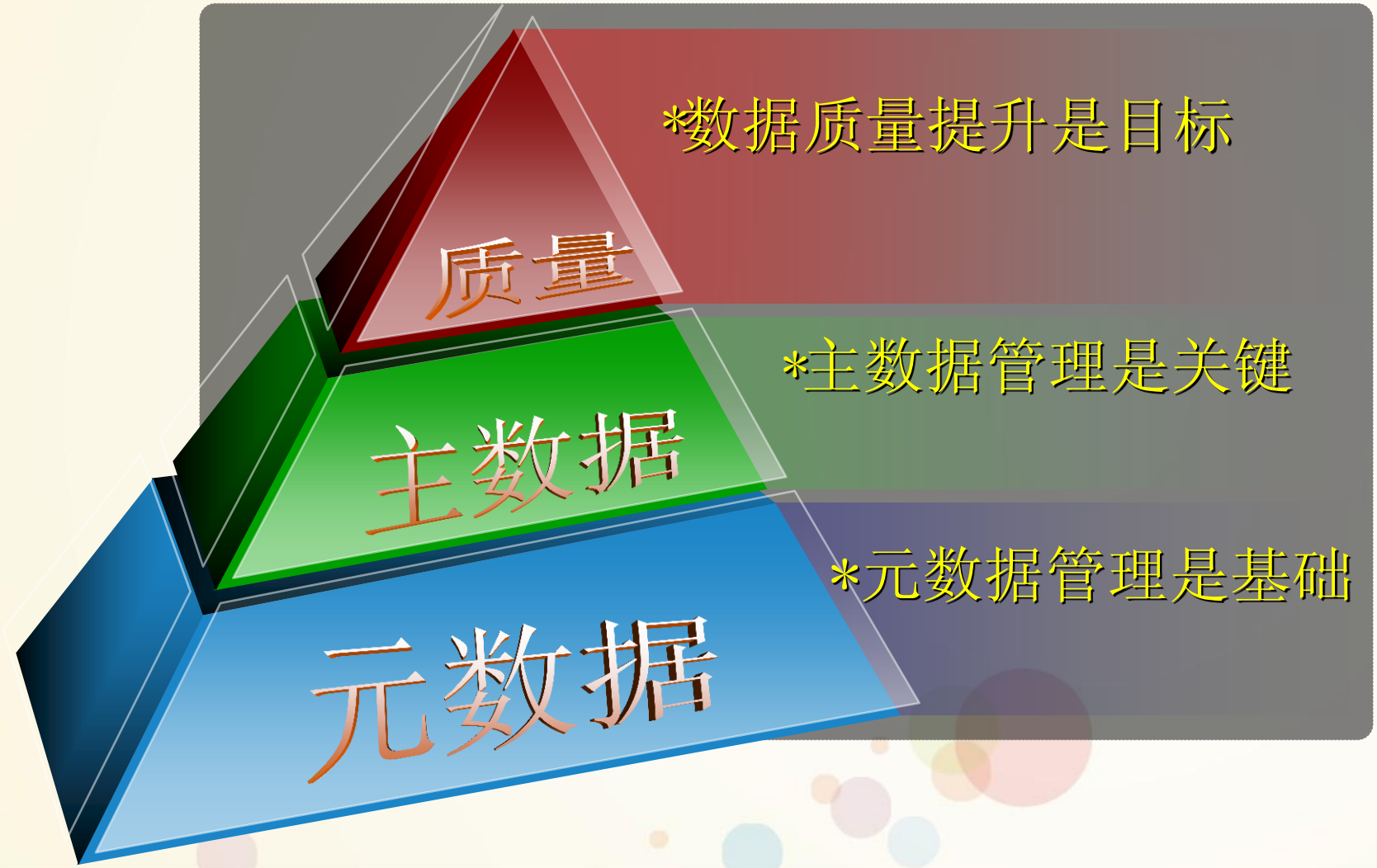
责任到人

持之以恒

绩效评估

经验总结

实施建议



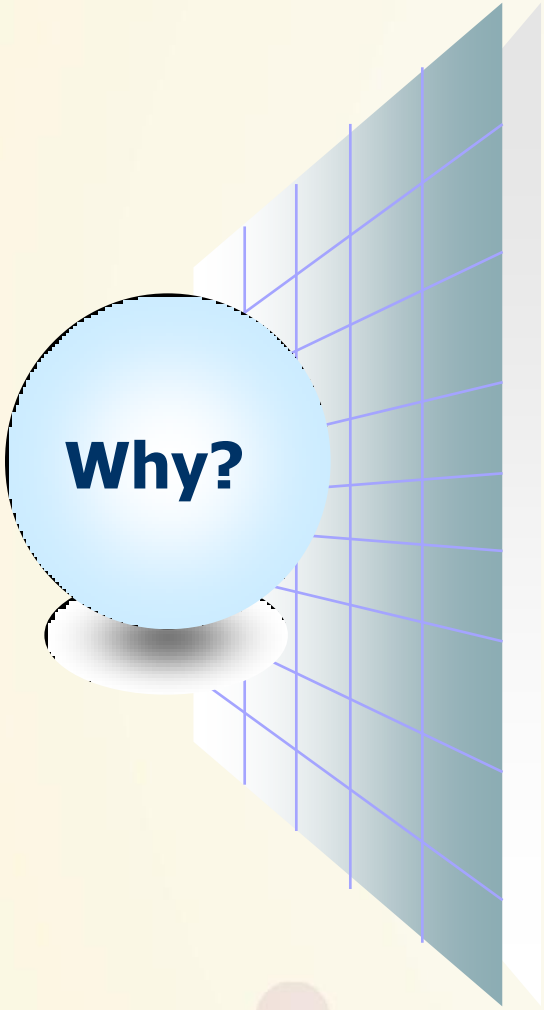
议程

- 数据治理的背景和现状
- 数据治理策略
- 元数据管理
- 主数据管理
- 数据质量管理
- 大数据平台设计

什么是元数据

- 元数据的定义
  - 技术元数据
  - 业务元数据
  - 操作元数据

为什么要进行元数据管理

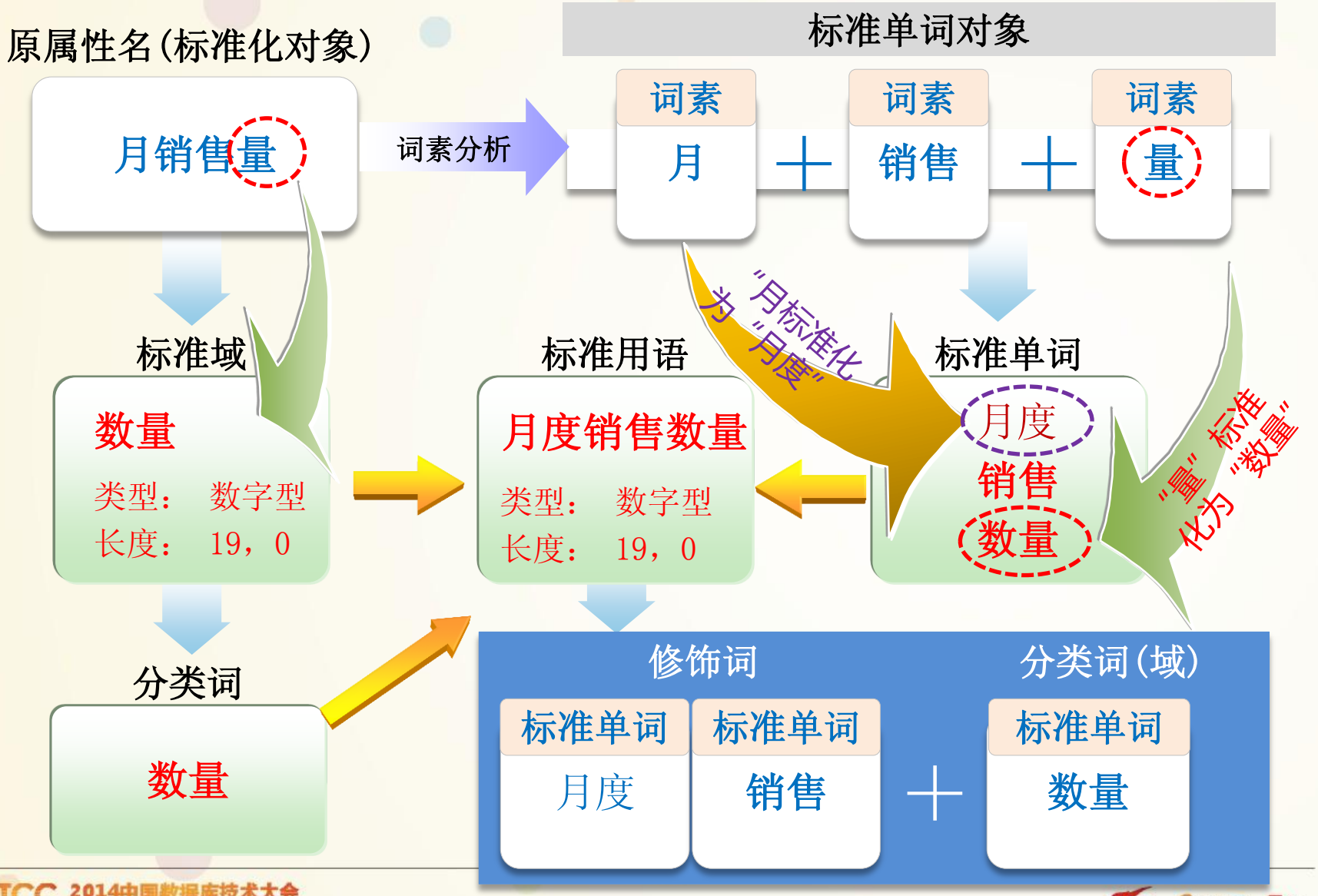


1. 数据的参考框架
2. 解决数据模糊性
3. 可视化数据流动
4. 影响和血缘分析
5. 推进标准化建设
6. 规范化数据审计

经验分享

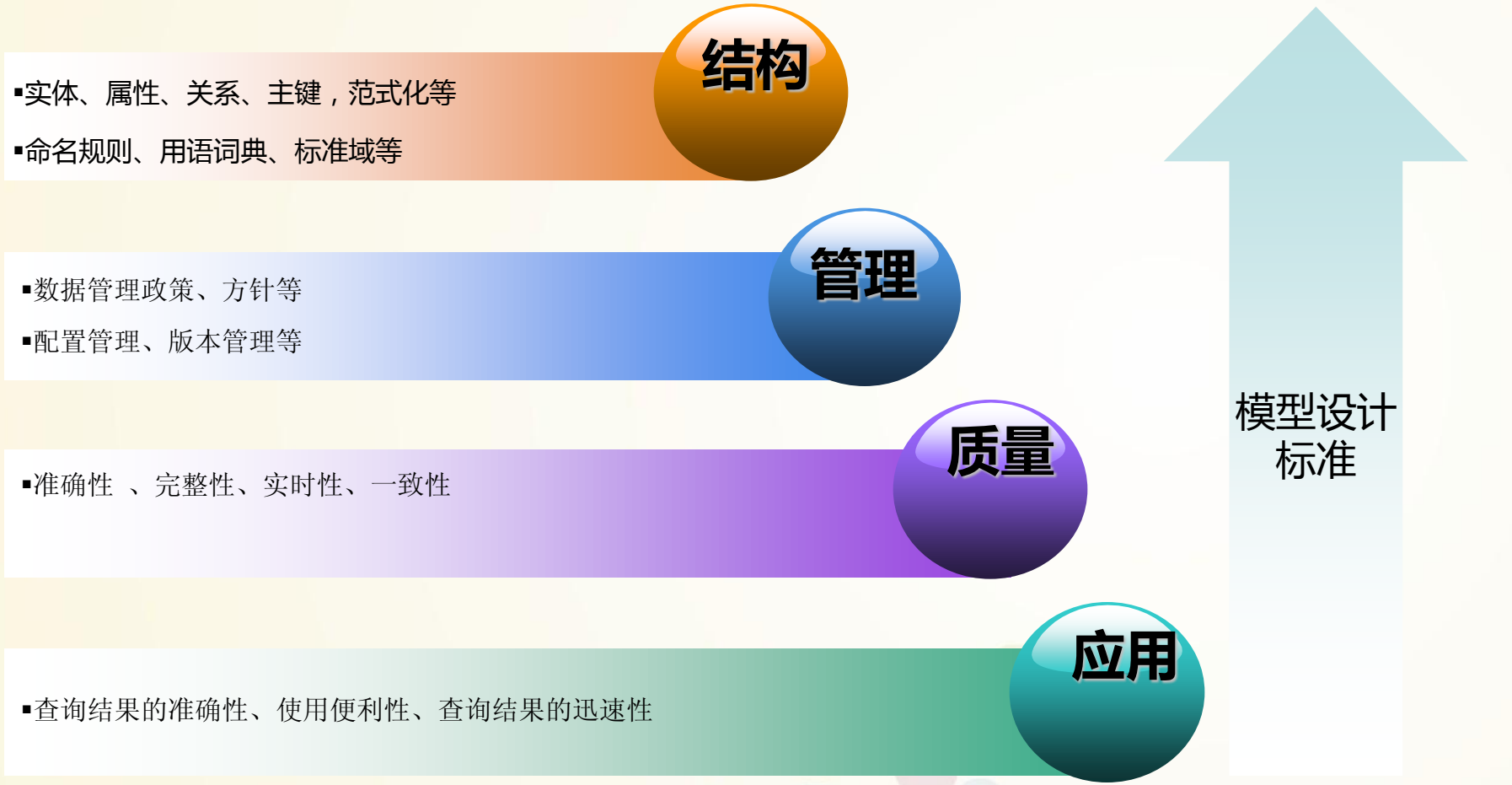
1. 标准先行
2. 全局治理
3. 尽快见效
4. 高层支持
5. 业务参与
6. 奖惩机制

数据定义标准化

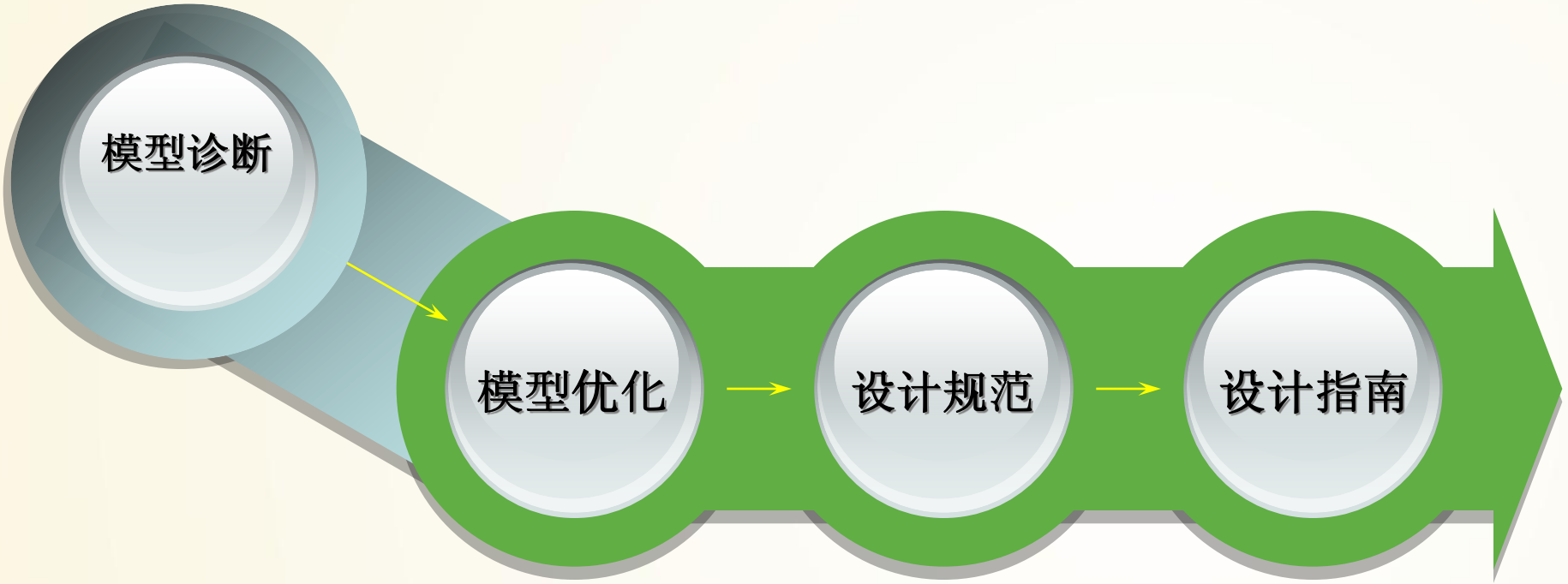

原属性名(标准化对象)
月销售量
词素分析
标准单词对象
词素
月
词素
销售
词素
量
"月"标准化为"月度"
"量"标准化为"数量"
标准域
数量
类型： 数字型
长度： 19，0
标准用语
月度销售数量
类型： 数字型
长度： 19，0
标准单词
月度
销售
数量
分类词
数量
修饰词
标准单词
月度
标准单词
销售
分类词(域)
标准单词
数量

数据定义标准体系

数据模型标准化

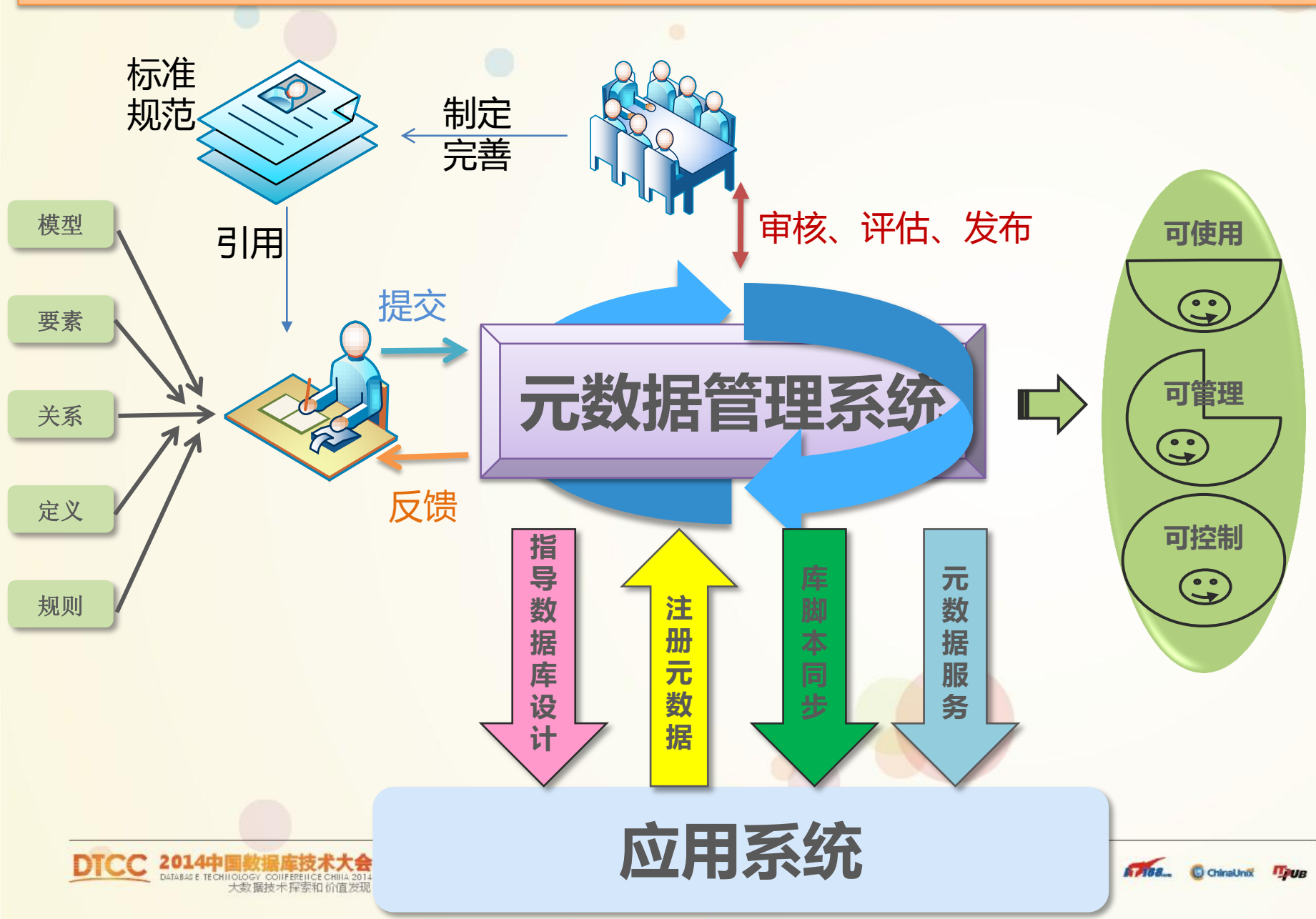

结构
▪实体、属性、关系、主键，范式化等
▪命名规则、用语词典、标准域等
管理
▪数据管理政策、方针等
▪配置管理、版本管理等
质量
▪准确性 、完整性、实时性、一致性
应用
▪查询结果的准确性、使用便利性、查询结果的迅速性
模型设计
标准

实施路线



按照模型设计规范和指南统一设计企业内部数据模型

标准化体系（数据定义&模型设计）



元数据管理工具的选择

- 元模型易于扩展
- 界面友好
- 安全和系统管理
- 配置管理
- 发布、查询、报表功能
- 平台开放
- 提前试用

议程

- 数据治理的背景和现状
- 数据治理策略
- 元数据管理
- 主数据管理
- 数据质量管理
- 大数据平台设计

什么是主数据

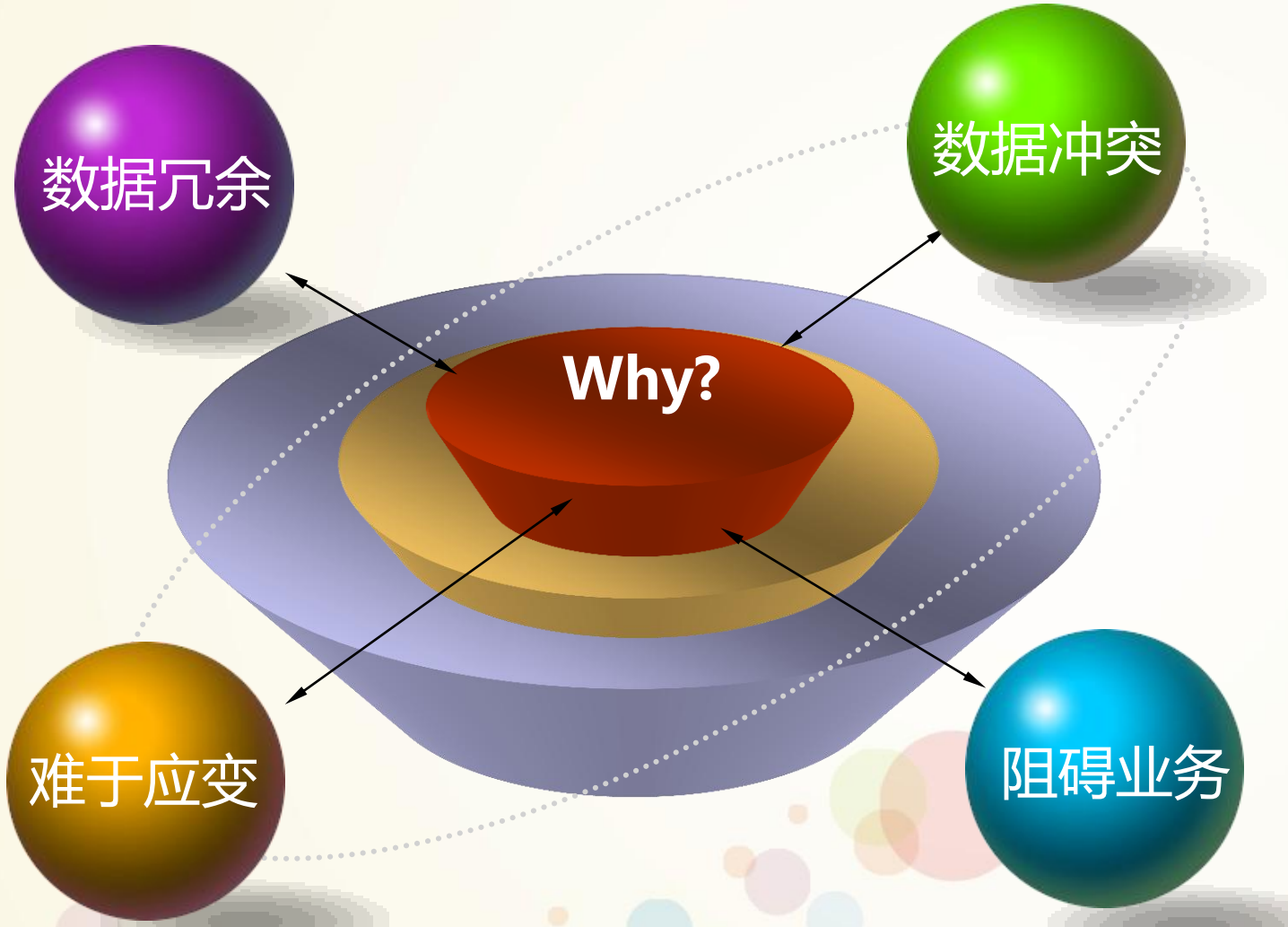
- 企业主数据分散存储在企业各系统内，对企业至关重要的核心业务实体的数据，比如客户、合作伙伴、员工等
  - 关键
  - 分散
  - 缓慢
  - 共享

主数据类型

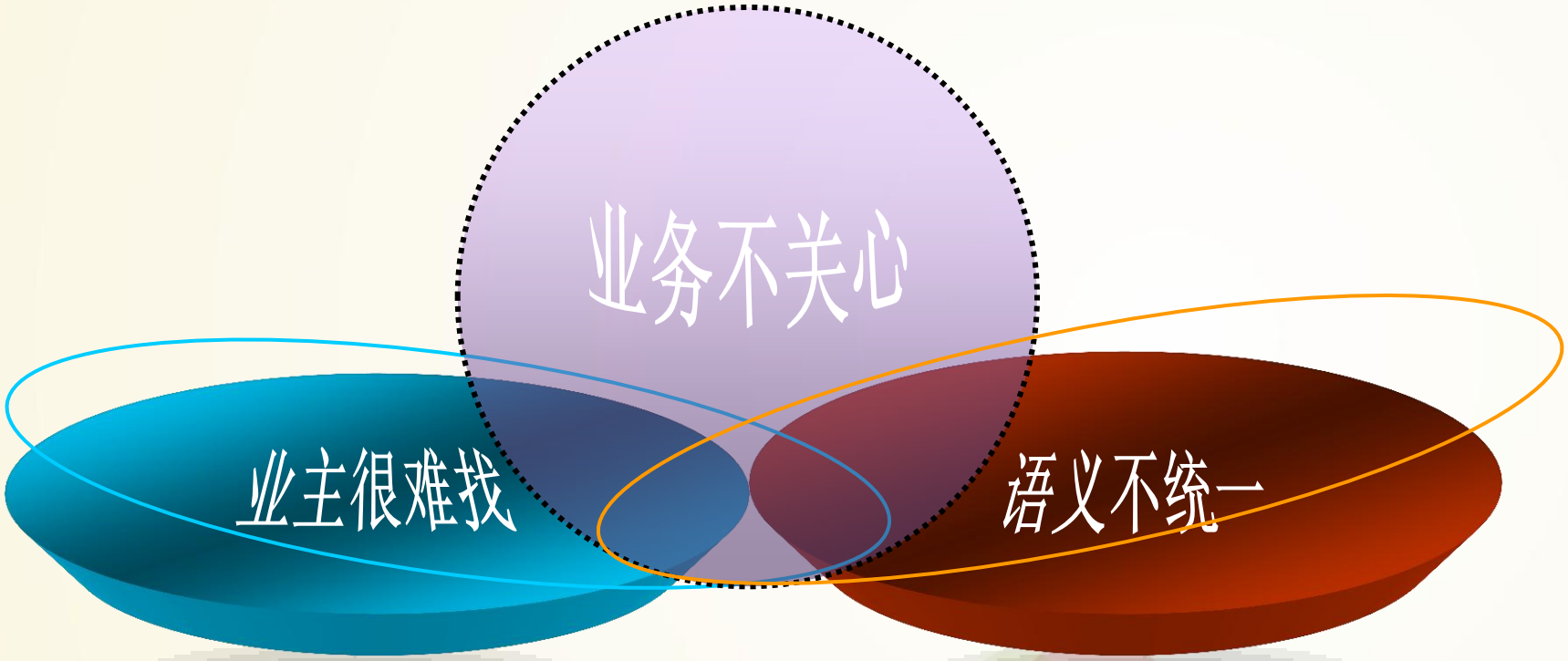
主数据与参考数据

- 参考数据可以是主数据，但不一定是主数据

为什么要作主数据管理



如何做好主数据管理



如何做好主数据管理

识别并管理主数据相关方

整理并分析主数据的生命周期

识别主数据含义、上下文、类型

主数据实施流程

主数据管理体系

提升数据质量　　统一数据共享　　强化决策支持

标准规范

组织机构

主数据管理系统

访问服务

数据导入

通知　注册　准入　申请

废弃　审批　维护

匹配查重　查　询　数据校验　版本管理

安全管理

数据分发

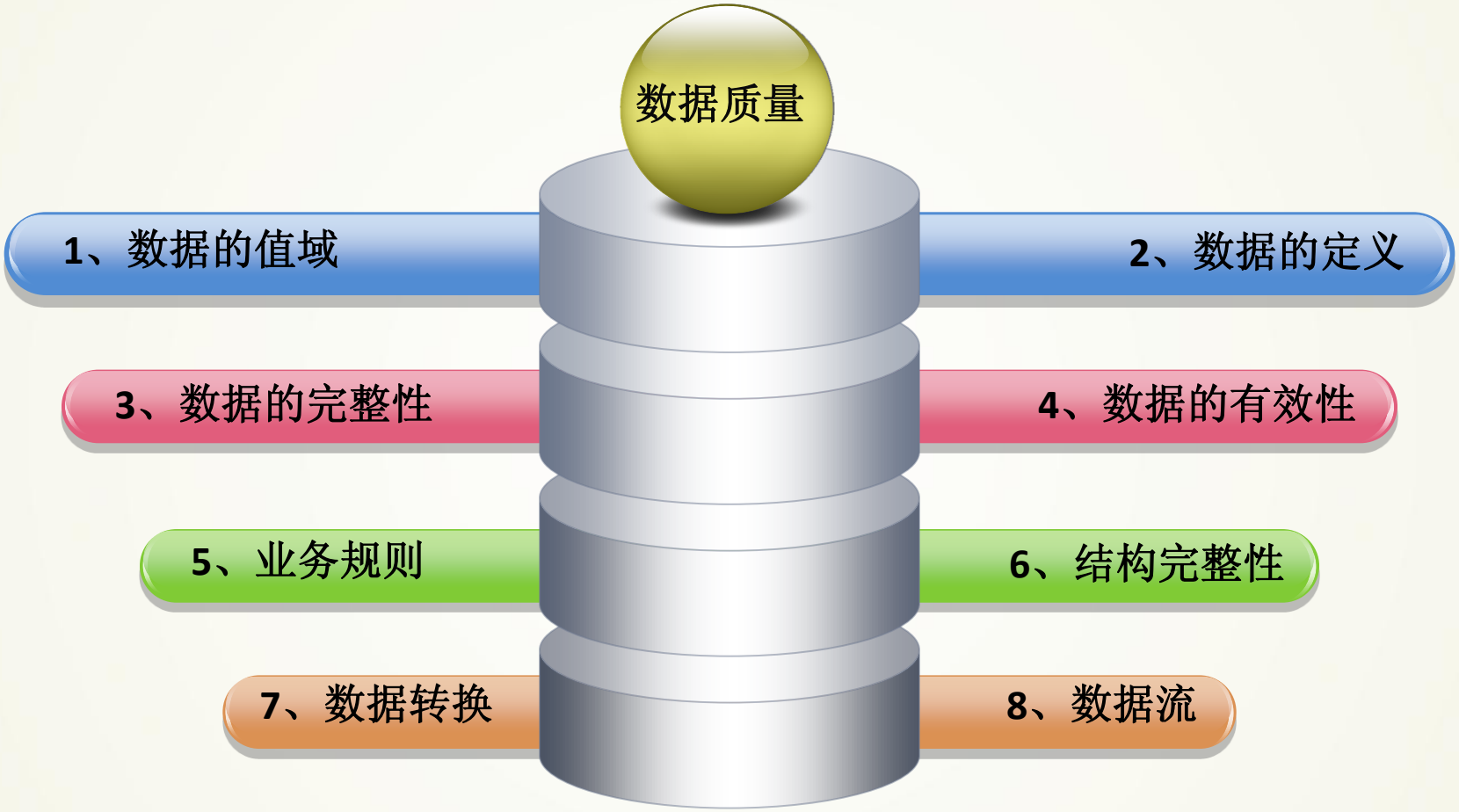
管理流程

ERP　CRM　人事　财务　……

议程

- 数据治理的背景和现状
- 数据治理策略
- 元数据管理
- 主数据管理
- 数据质量管理
- 大数据平台设计

数据质量问题

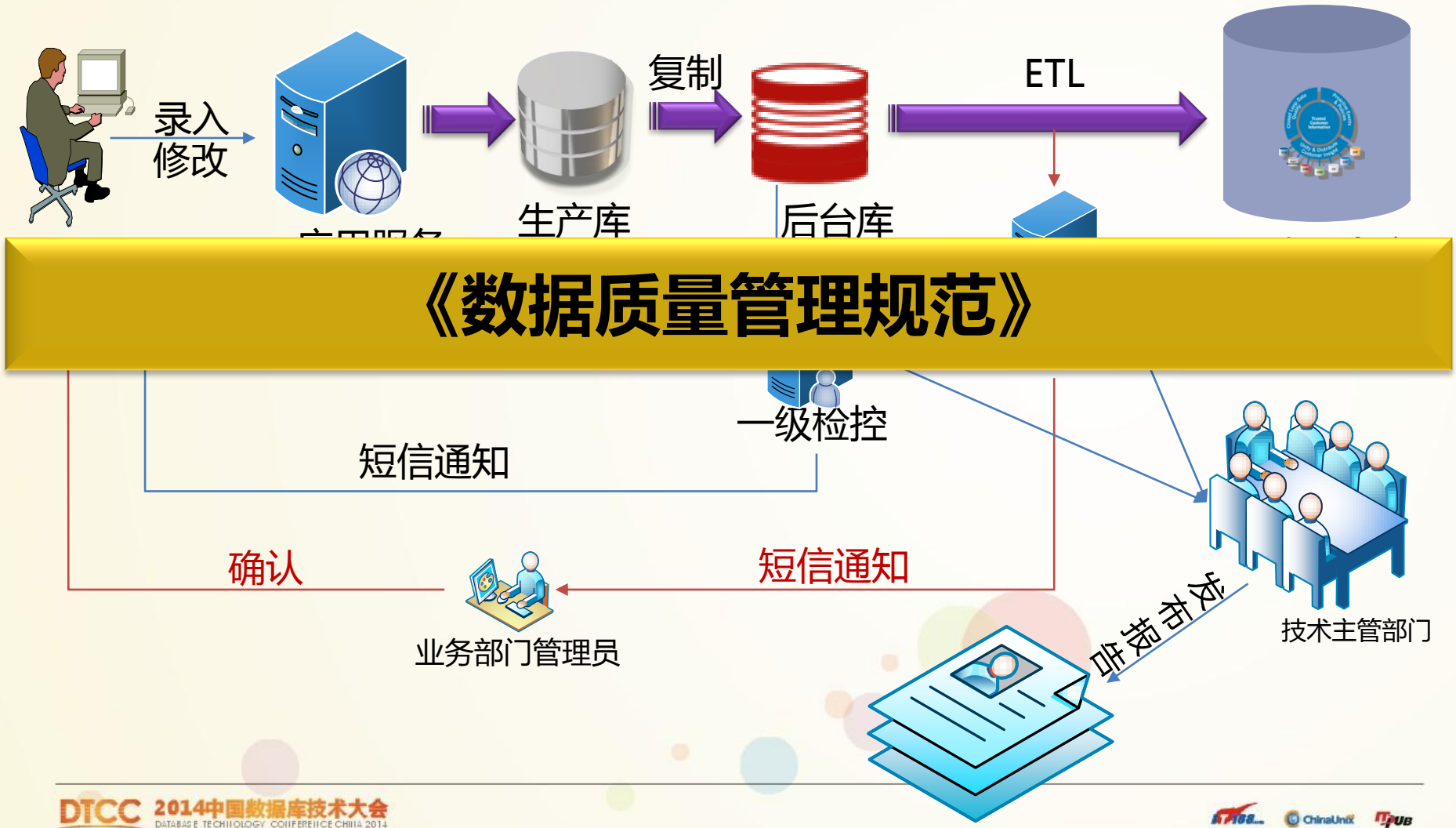

数据质量
1、数据的值域
2、数据的定义
3、数据的完整性
4、数据的有效性
5、业务规则
6、结构完整性
7、数据转换
8、数据流

组织架构设计

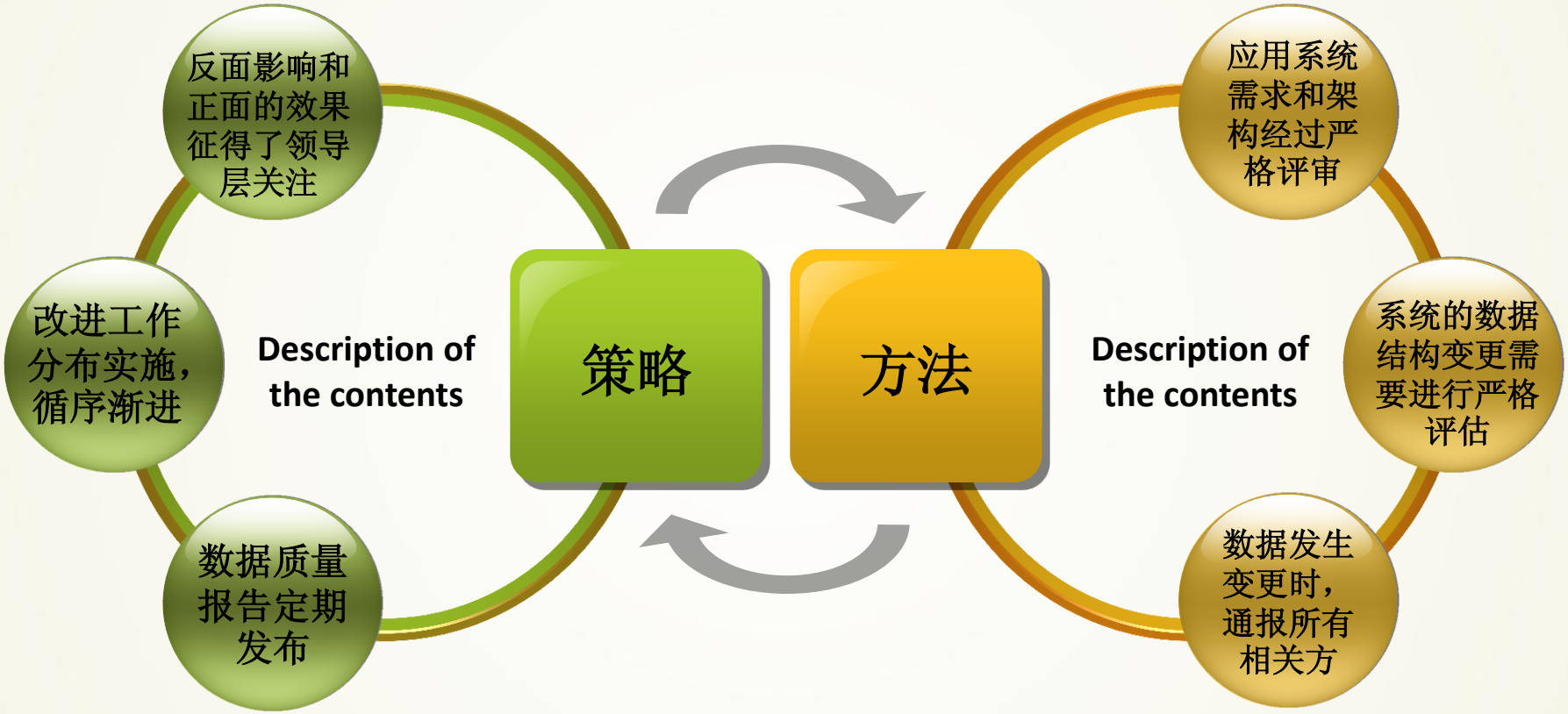
- 业务与技术部门各司其职，共同做好数据质量管理工作

业务部门	技术主管部门	评审委员会
统计部门（业务部门）负责业务规则的制定，在业务层面统管数据质量和安全。	技术部门负责数据集成、使用等过程中的数据质量，并对数据质量报告进行定期发布。	技术部门设置评审委员会，对数据方面的变更进行管控，具备技术方案否决权。

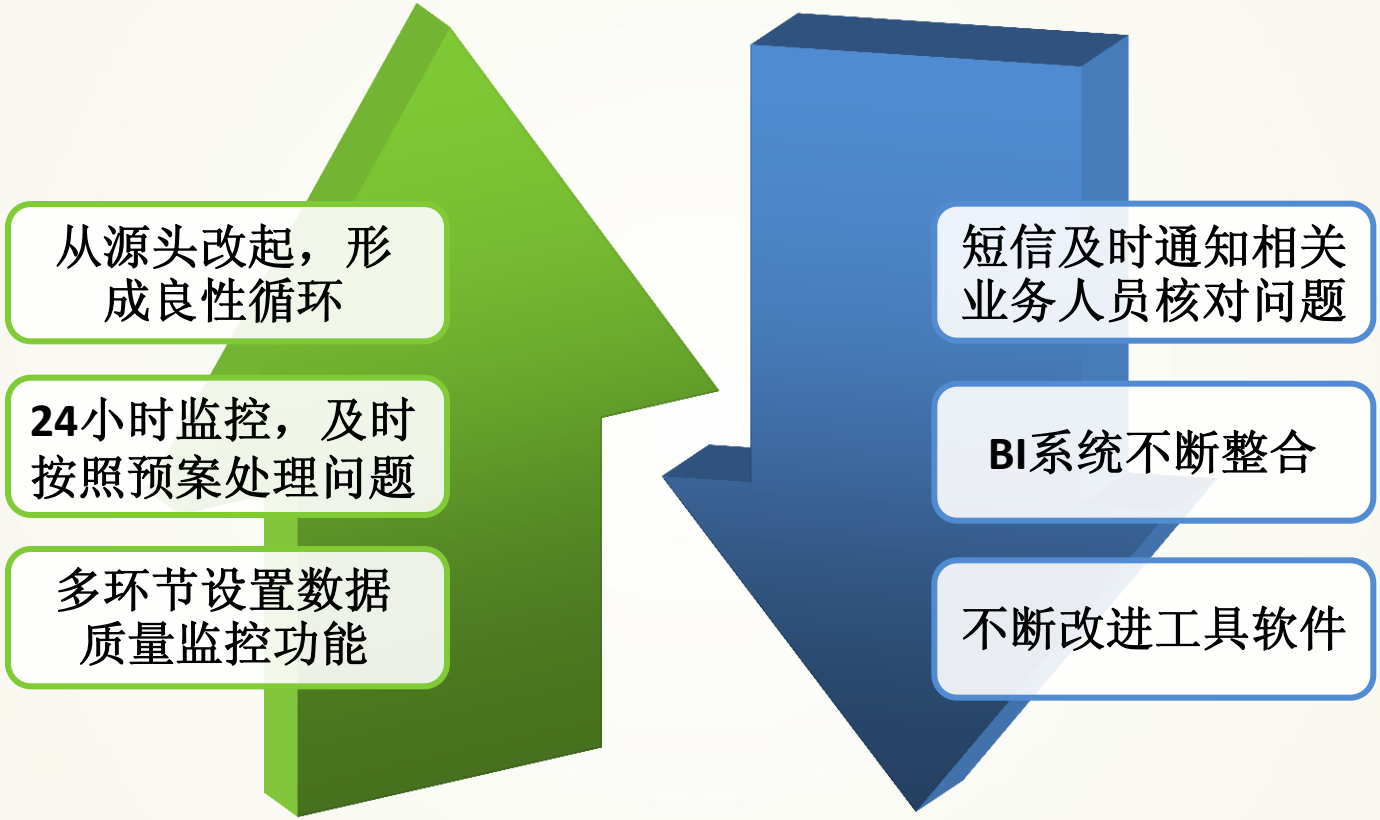
数据质量治理流程


录入
修改
复制
ETL
生产库
后台库
《数据质量管理规范》
一级检控
短信通知
确认
短信通知
业务部门管理员
发现异常
技术主管部门

策略和方法


反面影响和正面的效果征得了领导层关注
改进工作分布实施，循序渐进
数据质量报告定期发布
Description of the contents
策略
方法
Description of the contents
应用系统需求和架构经过严格评审
系统的数据结构变更需要进行严格评估
数据发生变更时，通报所有相关方

技术手段


从源头改起，形成良性循环
24小时监控，及时按照预案处理问题
多环节设置数据质量监控功能
短信及时通知相关业务人员核对问题
BI系统不断整合
不断改进工具软件

最佳实践

从数据剖析（Profiling）开始

尽量使用工具进行数据剖析

数据剖析工作需要持续开展

数据集成过程也需要进行数据剖析

数据质量评估和改进需要被动和主动两种方式

最佳实践

得到高层的支持

关键数据先行，渐进开展

在数据的“上游”解决质量问题

“防患于未然”优于“后期治疗”

数据质量报告要大范围发布

议程

- 数据治理的背景和现状
- 数据治理策略
- 元数据管理
- 主数据管理
- 数据质量管理
- 大数据平台设计

关于大数据的几个问题

- 什么是大数据
- 大数据与传统数据仓库是什么关系
- Hadoop与MPP数据库

传统数据仓库


数据采集
数据存储计算
数据展现
源数据
数据转换
数据仓库
数据展示
业务用户
ERP
CRM
财务
人事
ODS
抽取
转换
清洗
加载
EDW
数据集市
报表展现
数据挖掘
多维分析
分析人员
管理员

Q&A

THANKS