DQMiS

第二届数据质量管理国际峰会

The 2nd Data Quality Management International summit

华矩科技

大数据时代的数据治理挑战及应对策略

主讲人：华矩科技董事长CEO 谭海华

2018年9月

DQMiS

目录

CONTENTS

01

数据治理在当下大数据内外环境下的发展

大数据的冰山与金山

企业拥有的巨量数据
都是一个沉睡的金矿

数据的价值则如冰山，
您所知道的只是显露的一角

企业大数据价值变现的阶梯发展

大环境对数据治理提出了更高的要求

目前数据应用项目非常多，但真正取得预期效果的项目少之又少，而且开发过程困难重重，其中的一个重要原因就是数据质量问题导致许多预期需求无法实现。如果没有数据治理，再多的业务和技术投入都是徒劳的，因为很经典的一句话：Garbage in Garbage out。数据治理是保证数据质量的必需手段，从全球范围来看，加强数据治理提升数据质量已成为企业提升管理能力的重要任务。

全球的大数据发展都对数据治理提出了更高的要求

中国

《银行业金融机构数据治理指引（征求意见稿）》

欧盟

《通用数据保护条例》(GDPR)

美国

《2018年加州消费者隐私法案》(CCPA)

联邦

2018年7月18日联邦贸易委员会(FTC)发布了在众议院能源和贸易小组委员会的作证文本。根据证词，消费者隐私和数据安全将继续成为FTC的执法优先事项。

数据治理成企业当前大数据发展的核心策略

> 数据治理正迅速发展成一种企业核心策略！

如今，企业对于全面数据治理的需求从未如此强烈。监管机构希望企业能更加清晰地了解数据，对它进行有效的管控；企业管理层希望理清数据资产，降低数据应用的复杂性，对企业进行更高效的管理；企业每名员工都认识到数据的重要性，更多地采用数据驱动的方式来开展工作。

02

企业数据治理建设面对挑战

数据治理中常见的几个问题

“传统的数据治理的主要问题是关注于数据本身，而没有首先关注业务价值，数据只有创造业务价值对于企业而言才有意义。”

数据治理的成果及目标

一个系统？　　优化的数据？

一个平台？　　改善的报告？

数据集成？　　实时数据展现？

数据仓库？　　精准用户画像？

一套标准？

“传统的数据治理的主要问题是关注于数据本身，而没有首先关注业务价值，数据只有创造业务价值对于企业而言才有意义。”

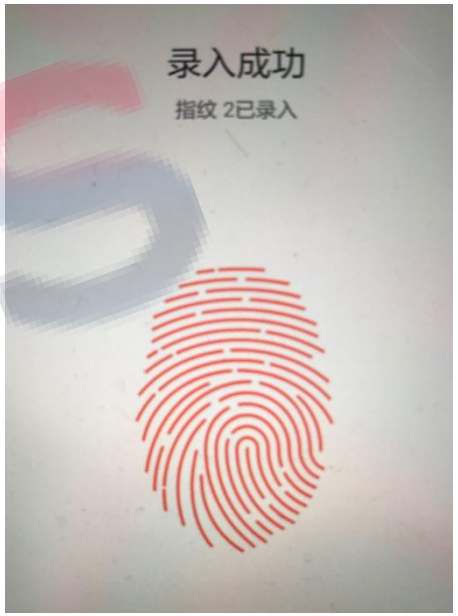
数据治理目标（示例）

DQMiS

数据治理目标

- 搭建数据质量管理机制，快速诊断并优化数据质量，确保数据应用的高效运行。
- 梳理数据资源，为数据资源整合和元数据管理提供可靠依据。
- 整合数据资源，搭建可扩展、高性能、高可用的ODS/EDW数据仓库系统，实现数据资源集成共享。
- 基于高质量的数据和整合的ODS数据，实现主数据管理，实现客戶单一视图和客户分群，搭建客戶标签体系，推动精确营销业务。
- 建立数据资产化管理和数据治理保障机制，加强数据管控，健全数据管理体系。
- 构建数据服务和应用体系，不断优化数据管理体系，实现真正的数据驱动业务。

数据与真相的关系



- 对的数据不代表真相
- 错的数据一定不能反映真相
- 大数据时代，**正确的，高质量的数据能让你无限接近真相**

数据质量的相关指标（一）

指标类型	说明	衡量标准	备注
完整性（Completeness）	不存在或缺失字段的数据的百分比。即实体的每个属性都有明确的值，不存在“空”或“未知”的属性。	字段的空值率	= 空值记录总数/总记录数
相关性（Reliability）	满足外键参照完整性数据的百分比。对于数据库中的某些实体，它们的存在可能要依赖于其它的实体。	外键无对应主键的记录比率	= 外键无对应主键的记录总数/总记录数
唯一性（Uniqueness）	满足主键唯一性约束数据的百分比。即一个表中的一组属性的值是唯一的。	1 - 主键的重复率	=1 - 主键重复的记录总数/总记录数
有效性（Validity）	满足域和数据有效范围定义的数据的百分比，即实体属性的值要在用户定义的有效范围之内。	1 - 异常值比率	= 1 - 超出值域的异常值记录总数/总记录数
及时性（Timeliness）	是否满足业务应用对数据的时间要求。	满足时间要求的比率	= 满足时间要求的数据总数/总数据数
非重复记录（Non-duplicate records）	是否存在多个记录表现同一个实体的现象。	样本数据非重复记录比率	= 1 - 样本数据重复记录总数/样本数据总记录数

数据质量的相关指标（二）

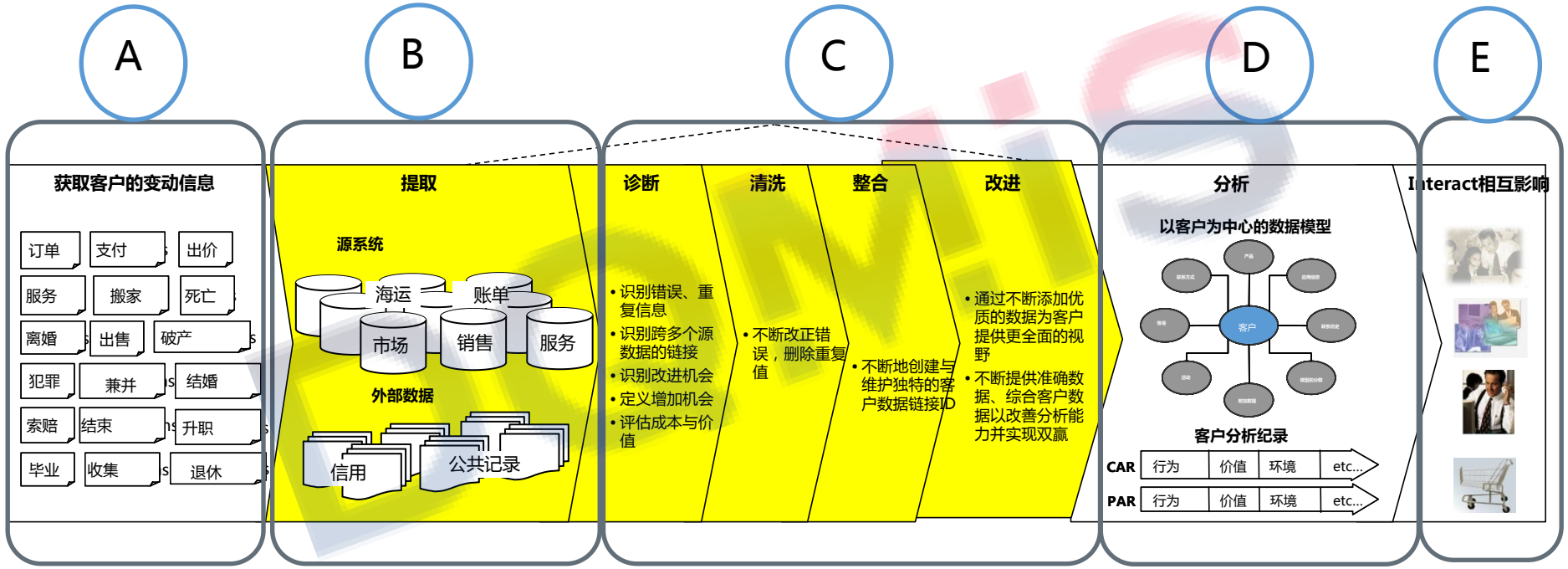
指标类型	说明	衡量标准	备注
真实性（Facility）	真实数据的百分比。真实性是指数据库中实体必须与对应的现实世界中的对象是一致的。	样本数据真实数据比率	= 1 - 样本数据中失真记录总数/样本数据总记录数
精确性（Accuracy）	指数据精度是否符合业务需要。	样本数据满足业务对精度需求的比率	= 样本数据中满足业务精度需求的记录总数/样本数据总记录数
一致性（Consistency）	与其它系统（或者系统内部）一致的数据的百分比。	样本数据不同存储的匹配率	= 1 - 样本数据中不同存储不同意义记录总数/样本数据总记录数
可理解性（Understandable）	含义明确和易于理解的数据的百分比。即数据本身的含义是否简单、明确。	样本数据易于理解的记录比率	= 1 - 样本数据中费解的记录总数/样本数据总记录数
可用性（Availability）	数据是否可获得，满足业务使用。	样本数据可获得记录的比率	= 样本数据可获得记录总数/样本数据记录总数

何为反映真相的优质数据

典型的数据治理建设路径

DQMiS

客户数据管理过程



- 各家客户数据纬度有限，无法靠自家数据形成丰满的360度客户画像

- 历史及异构数据原因，直接导致元数据管、主数据管理困难。
- 数据质量差
- ETL不堪重负

- 无法发现核心处理规则
- IT与业务人员无法有效互动
- 工作量巨大无法胜任
- 无法形成精准的统一视图
- ”半吊子“工程

- 没有高质量数据输出，无法调试建立更优模型
- 分析报告误差率高，但无法判断是数据质量问题还是模型问题

- 应用
- 反馈

03

如何提升企业数据质量

数据治理的核心：全面了解您的数据

- 易受控制的风险
- 易处理的业务规则
- 清晰的期望值
- 业务用户参与程度高

- 不易受控制的风险
- 难以知晓的业务规则
- 低于预期的期望值
- 业务用户参与程度低

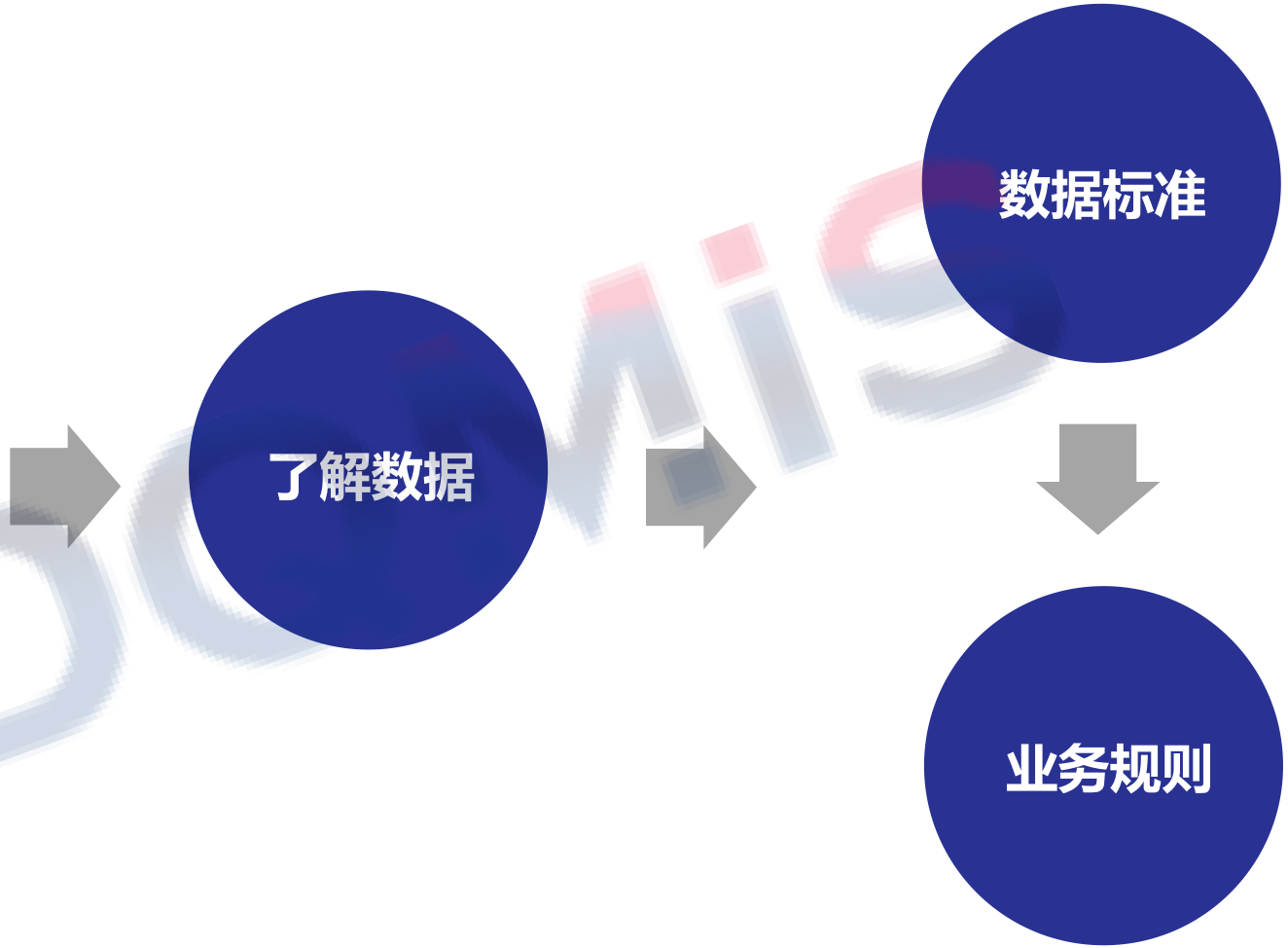
了解数据

DQMiS

数据本质

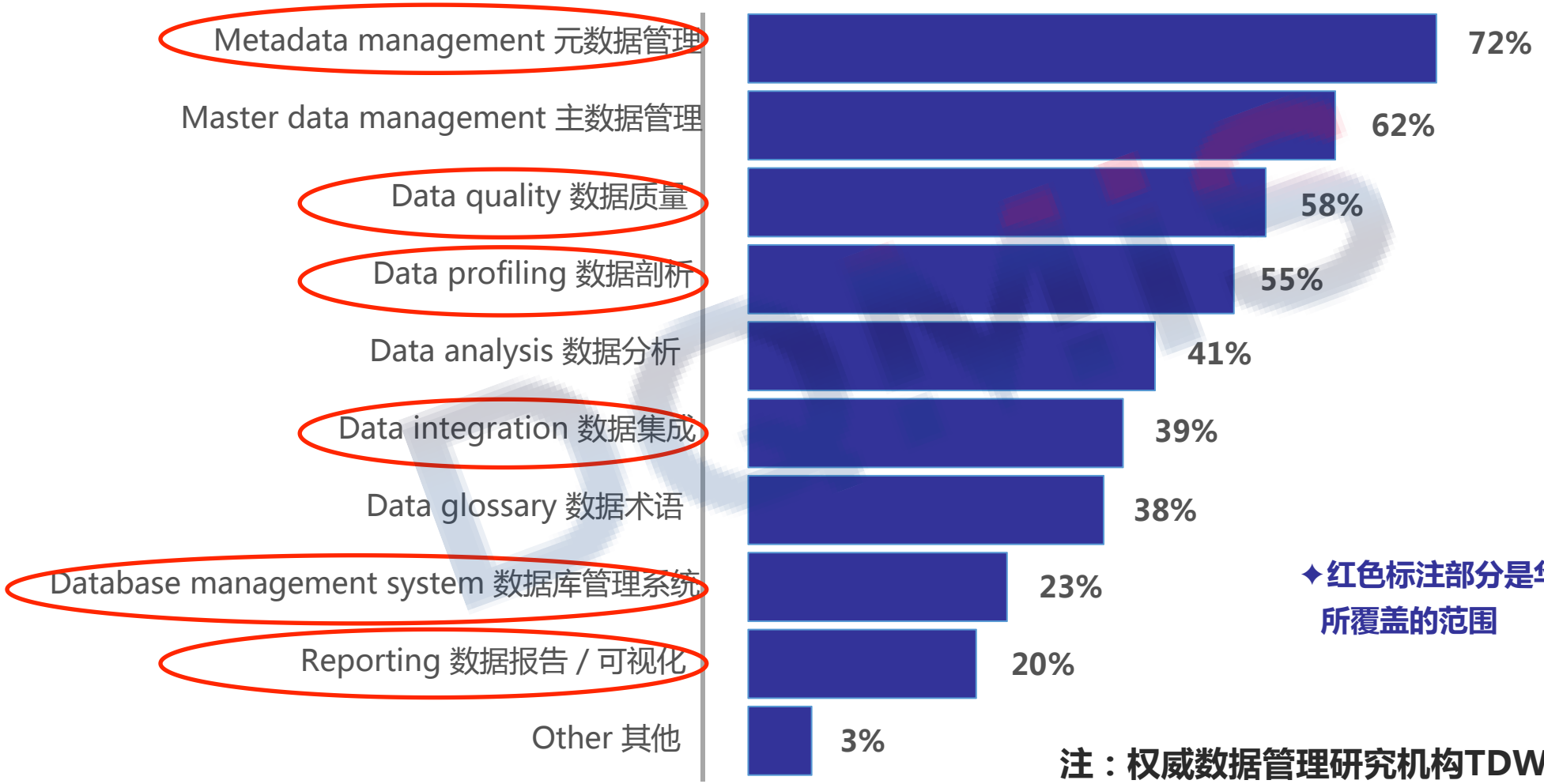
业务信息

设计信息



数据治理几个关键技术的重要性

DQMiS



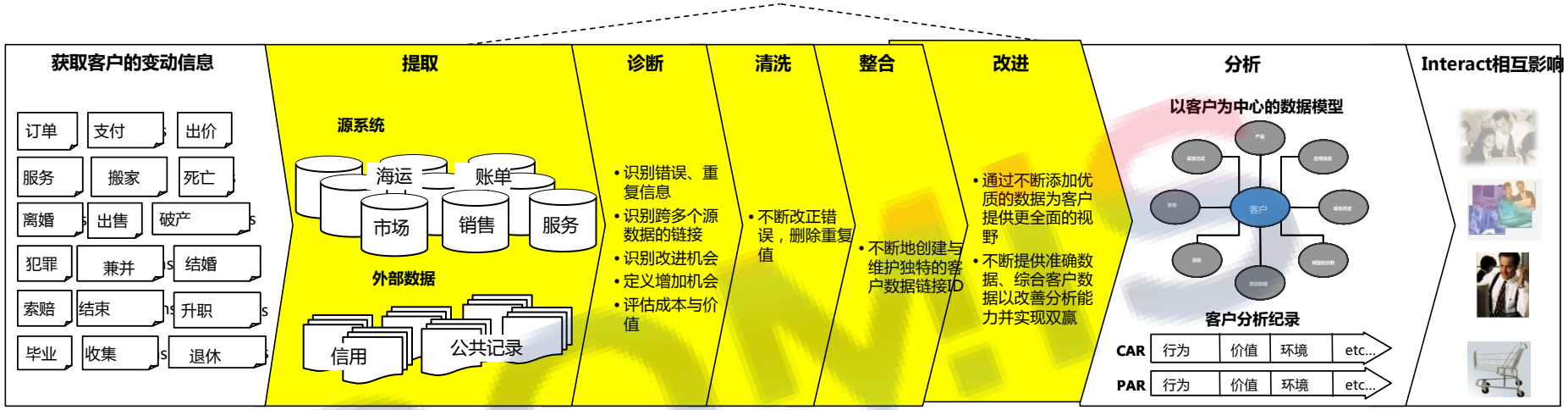
✦红色标注部分是华矩数据治理解决方案所覆盖的范围

注：权威数据管理研究机构TDWI对数百家国际企业的调研结果。

核心的数据治理技术

DQMiS

数据治理路径



调查

数据提取和调查

- 数据库连接
- 域
- 频率
- 模式
- 依赖性
- 外键关系
- 辨识数据值

丰富标准化

数据集成&标准化

- 价值诠释
- 属性赋值
- 数据解析
- 名字/地址
- 标准化
- 地址验证
- 附加数据

联系

关系匹配和数据整合

- 商业匹配
- 第三方数据匹配
- 最好的数据代

集成

实时数据集成

- 控制DB升级
- CRM, ETL, ERP
- RT验证和匹配
- 网页服务

数据治理平台交付的数据治理技术服务

DQMiS

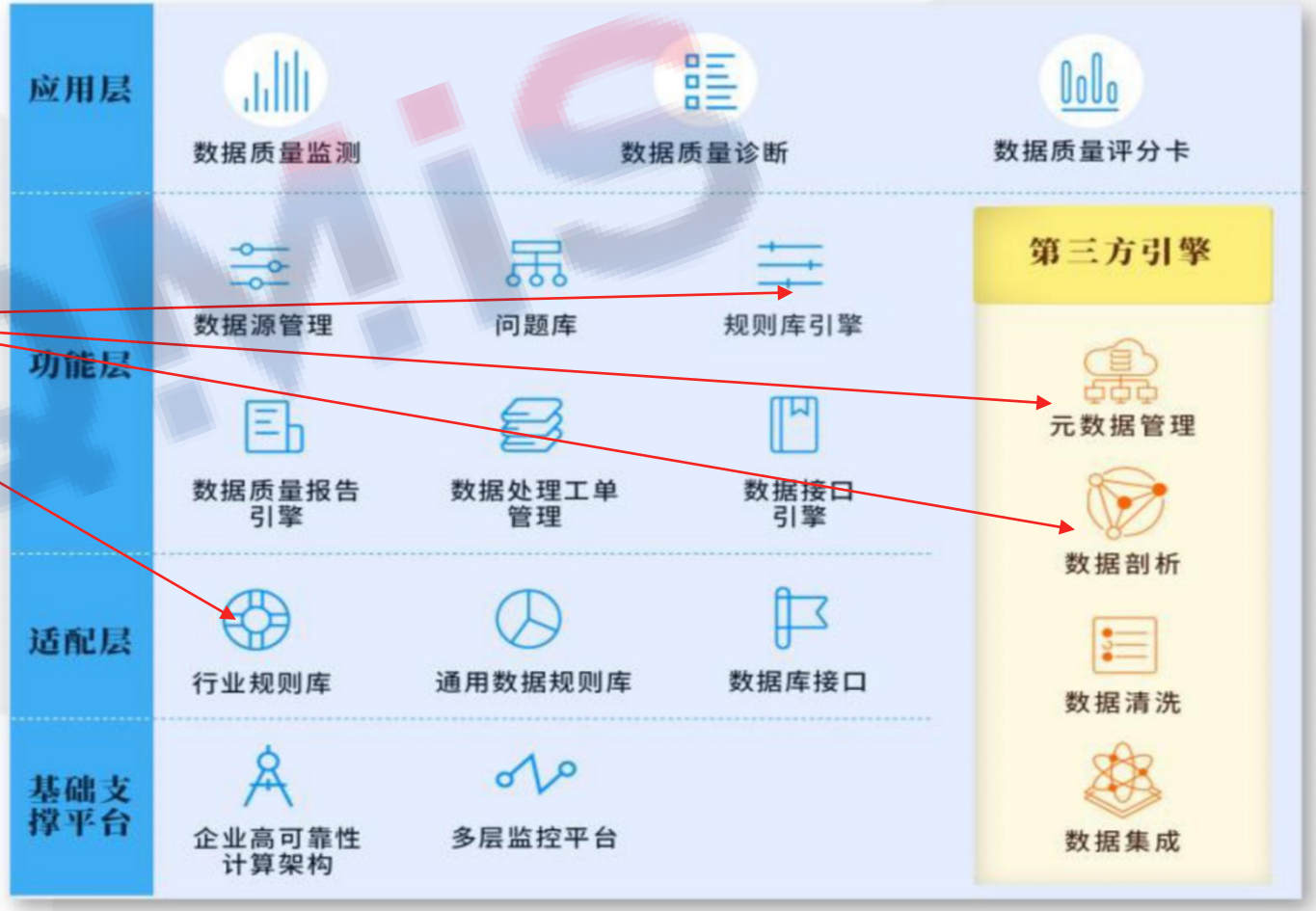
数据治理平台功能概述（示例）

DQMiS

数据治理平台

DQMiS

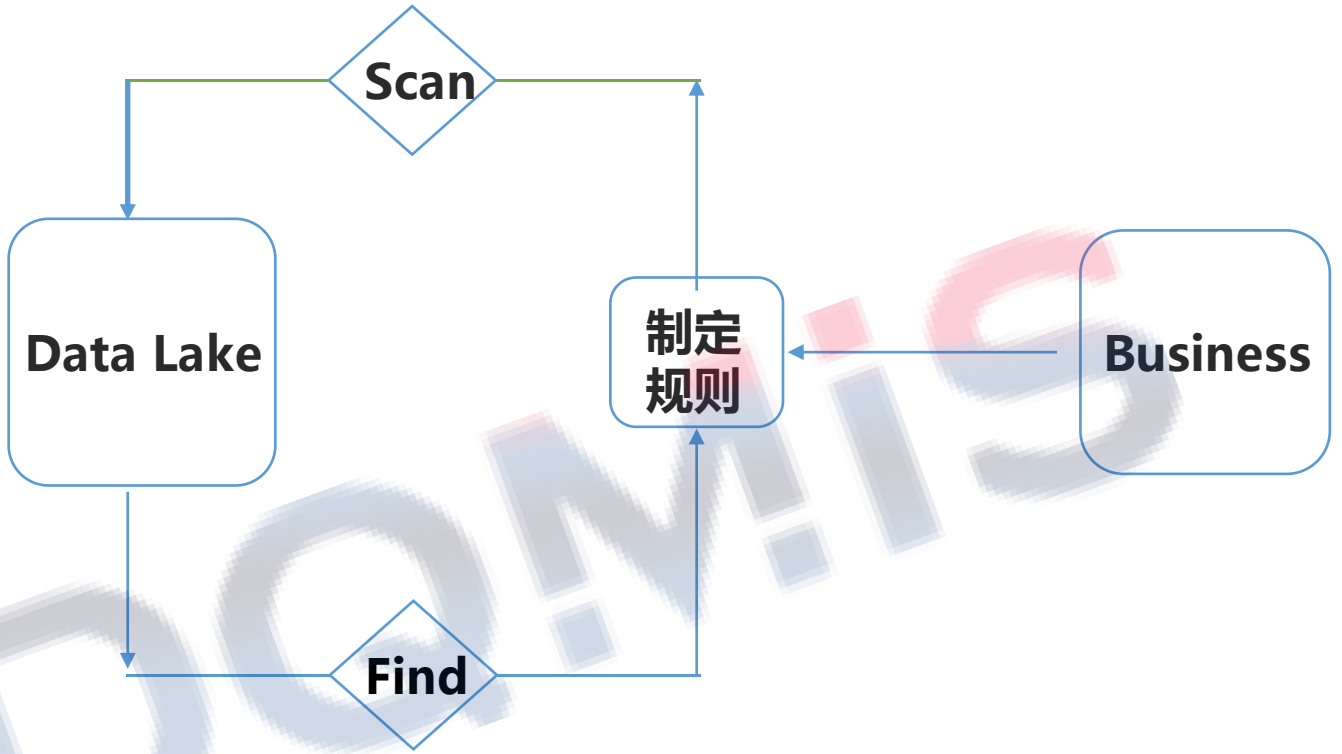
华矩数据治理平台



数据质量优化体系

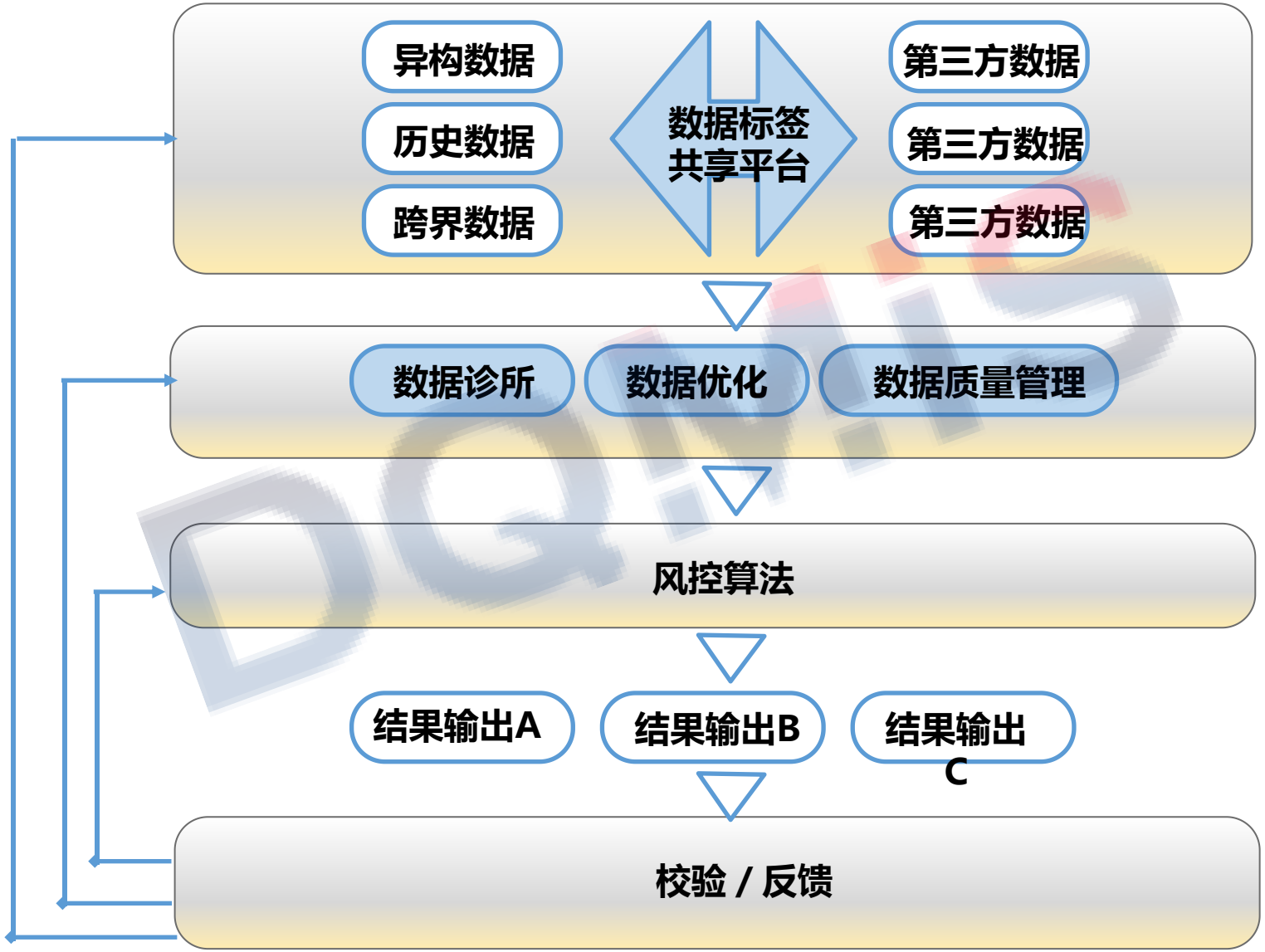
DQMiS

路径？

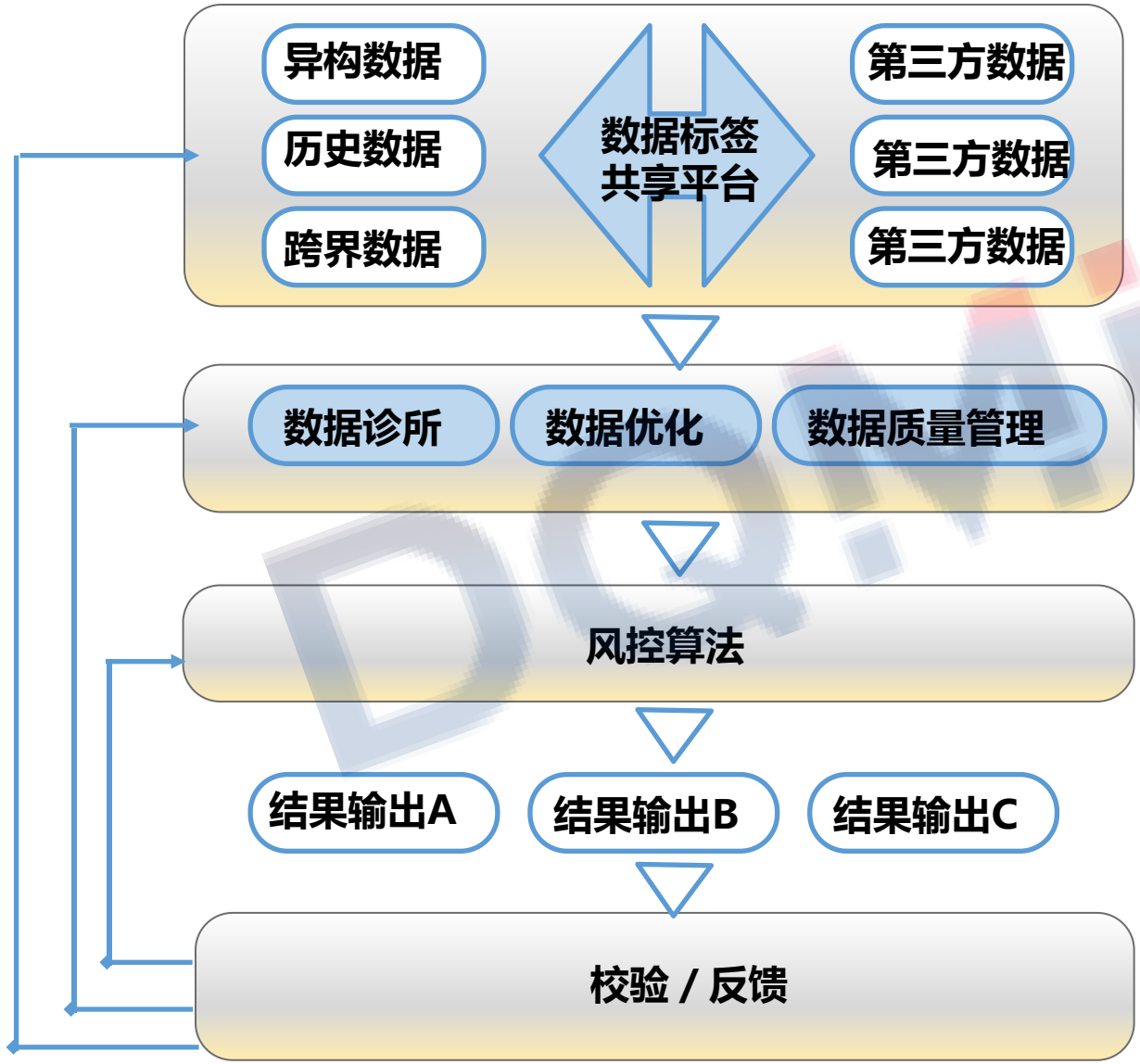


“传统的数据治理的主要问题是关注于数据本身，而没有首先关注业务价值，数据只有创造业务价值对于企业而言才有意义。”

数据治理与业务应用关系 - （示例）智慧金融风控模型



数据治理与业务应用关系 - （示例）智慧金融风控模型



- 大数据
- 多维度
- 跨界
- 标签丰满度
- 快速 / 实时数据整合

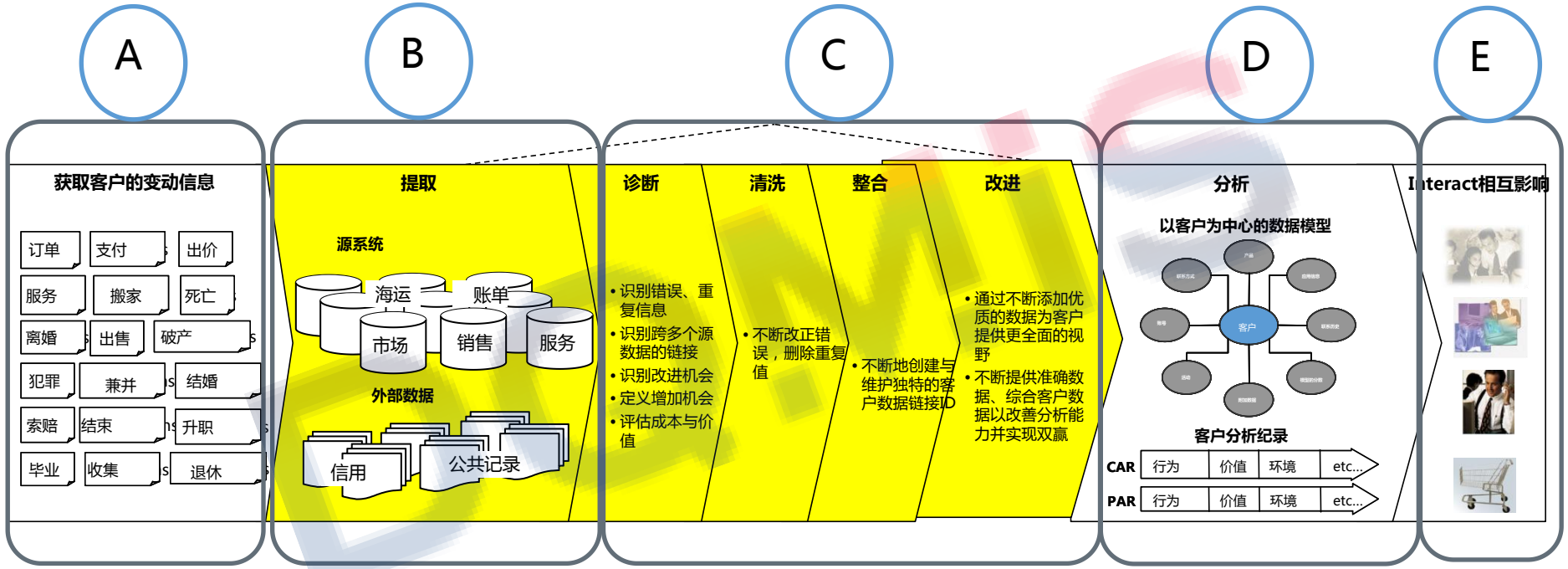
- 样本精准
- 数据质量高
- 控放自如的样本选择
- 可持续优化

- 校验及判断算法优化方向
（样本采集 / 数据质量 / 算法）
- 快速验证 / 反馈
- 算法自学习及调整
- 自动调节

典型的数据治理建设路径

DQMiS

客户数据管理过程



- 客户360度画像设计

- 数据源选择
- 元数据管理
- 主数据管理
- ODS/data mark 建设
- ETL

- 数据剖析
- 建立业务规则
- 数据质量诊断
- 建立客户统一视图
- 自动化处理的调试和部署

- 建模
- 数据分析
- 结果输出

- 应用
- 反馈

DQMiS

第二届数据质量管理国际峰会

The 2nd Data Quality Management International summit

华矩科技

Thank You!