

部委级数据治理项目 建设经验分享



彭文华





前言

大数据技术现在虽然很成熟，但是数据质量一直被人诟病。究其原因，主要是因为数据量的暴增、业务方对数据需求的渴求，导致没有时间做数据治理的工作。以至于大多数大数据场景中，只要能稳定的出数据就好的情况。

绝大多数互联网公司没时间建模、治理，直接拖宽表，这也是导致大数据环境中数据质量的低下的根本原因。

而在部委中，时间相对充裕一些，标准更规范一些，但是同样面临数据部、省之间、各系统之间数据交换、对齐的问题。

因此，在不同的环境中，数据治理的重点和偏向都是完全不一样的。



原目录/Contents

1. 数据治理目标
2. 数据治理架构
3. 数据质量管理方案
4. 大数据治理平台建设方案
5. 数据治理组织建设



新目录/Contents

1. 部委级数据治理困境
2. 不想治？有招！
3. 不会治！有招！
4. 没啥用~有招！
5. 部委级数据治理经验复制





定标准、做执行、强监督
PDCA戴明环

下文件、搞培训、做排名



你以为是部委

数聚大咖



DAMA中国数据大讲堂系列活动

中华人民共和国 [redacted] 部

[redacted] 297 号

关于全面推进 [redacted] 建设的指导意见

各省、自治区、直辖市

为深入贯彻落实中
安工作现代化水平的部
深度融合，全面提高消防工作科技化、信息化、智能化水平，实

新闻发布会

Press Conference of the Joint Prevention and
Control Mechanism of the State

开个会、发个文件就行了





甲方疯狂接电话



乙方疯狂赶工期





ICS 01.140.10:01.075
A 14



中华人民共和国国家标准

GB/T 39872—2021

标准文献技术指标揭示数据规范

5.1.2.20 专业领域代码

定义：本体类/揭示词所属的专业领域代码。

英文名称：SystemTradeCode

数据类型：字符型

值域：见 A.3

短名：syTCd

约束/条件：必选项

最大出现次数：N

5.1.2.21 本位体系代码

定义：本体类/揭示词所属的标准体系表的体系类目的代码。

英文名称：naturalSystemCode

数据类型：字符型

把标准制定好，数据质量自然就高了





实际上的部委

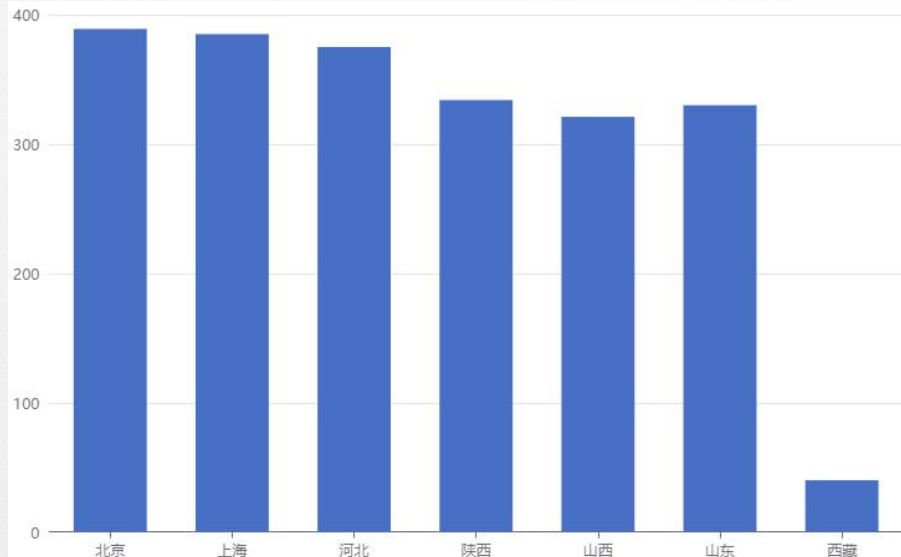
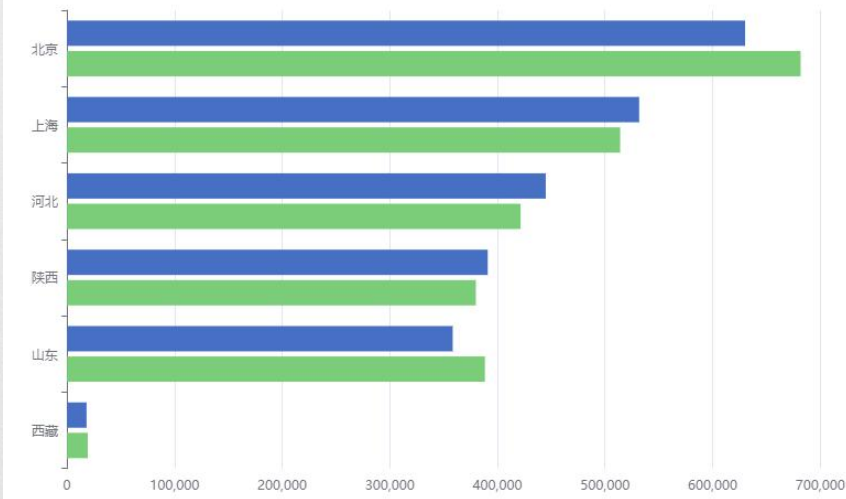
数聚大咖



DAMA中国数据大讲堂系列活动

各地数据上报质量

■ 及时性 ■ 准确性



各地基础条件不一样，人员能力不一样
中文语义非常广，不同语境理解不一样



你以为是部委

数聚大咖

DAMA中国数据大讲堂系列活动



一声令下，山呼海啸





一声令下，应者寥寥

每个人的事都非常多，肯定优先处理着急的事情





治好了没政绩，**没啥用**

数据治理太难了，**不会治**

耽误工夫，**不想治**



为啥？





定标准、做执行、强监督

下文件、搞培训、做排名

领导视角



兄弟视角



耽误工夫，**不想治**

兄弟不是给你**派活**的，兄弟是给你**助攻**的

举例子+挂欲望+给希望





耽误工夫，**不想治**

举例子：部里做了治理，X处省了很多事

挂欲望：你节省这些时间他不香吗？

给希望：国家主动推，机会难得！

将“**要我治**”，转为“**我要治**”



数据治理太难了，**不会治**

兄弟不是给你**出难题**，兄弟来给你**解决方案**



数据治理太难了，**不会治**

数据质量差，你也头疼，兄弟帮你**根治**

一致性校验
阈值校验
质量闭环

厂商难协调，你也烦恼，兄弟帮你**弄他**

厂商约谈
准入标准

质量缺监控，你也茫然，兄弟帮你**理清**

质量监控
质量量化

提供全套解决方案，降低执行**难度**





治好了没政绩，**没啥用**

兄弟不是来**忽悠你**，兄弟来给你**送功劳**





治好了没政绩，**没啥用**

荣誉要给到位！

各种排名要弄上
抓两头，放中间
选试点，比学赶帮超

应用要规划好！

规划先行
跨省应用
典型案例

抓手要准备上！

向下管理
向上汇报
多级应用

规划先行，设计动力系统



其实我有个疑问也想跟大家请教一下：我今年在公司提出建立数据管理体系的意见，希望公司成立一个数据管理体系小组，要建立数据管理SOP，盘点数据资产，梳理数据血缘，规范数据的采集、录入、调用、分享等过程，全面提升数据管控力度和数据质量。但是被同事否定说这些事情如果上ERP和SAP系统就能全部解决，我想请问一下各位，是这样吗？

想请教一下我应该怎么去反驳他呢？感觉别人经验很丰富，说服力很强，我不知道怎么去反驳

其实我们公司就是这么一个状态，我们是很传统的公司，信息系统很少，公司现在想做，我以前做数据分析，但是我觉得制约我数据分析的关键在于我们数据的质量和各部门数据口径的统一，调取的困难，使得分析很难进行，所以我就一直在想能不能建立一个好的系统性的管理体系，我们再上系统，这样各个系统之间逻辑口径都统一，以后才好分析。但是现在每个业务部门都想马上上系统，立马提升效率，觉得这种建体系的工作看不到实际效果，就阻力很大

光一个one id就很不容易

上SAP公司的ERP系统，或者说ERP实施项目，通常关注的是实现业务逻辑，同时基于一定的数据标准去实施，只能说是数据治理中的一部分内容，但不能解决您说的那些内容。

现在为啥都在提数据治理，是因为经过这么多年信息化建设，搭建了很多系统，数据也很多，那么导致系统间的集成应用和数据利用越来越不容易，所以需要治理一下。

嗯嗯，先不说别的，如果没有把主数据管理好，集成在一起的系统的数据库是混乱的，比如同一家供应商在不同的系统有多个编码，那么过账到ERP或者财务模块，账就是错误的，生产数据都是错误的，那数据分析无从谈起了。



对立视角



合作视角

你这玩意没做好，你说的根本不对！你那玩意满足不了我需求！数据质量这么烂，我怎么可能分析出好结果？

把要我做变成我要做：举例子、挂欲望、给希望

谁家数据治理做的很好，他们IT都不管数据的事情；SAP属于业务系统，我这属于分析系统。以前数据异常总得找你人工解决，以后系统自己就发现了。而且分析的结果更准确

提供全套解决方案，降低执行难度：数据治理有方法

数据治理有系统，有方法。我这里的高手，可以来帮忙。不会给你带来更多麻烦的

规划先行，设计动力系统：做好了，你省心啊！

要是你不做，那我这些分析需求都得找你呀。你看我现在分析中有各种数据孤岛问题。治理好了，你也省心啊！



原目录/Contents

1. 数据治理目标
2. 数据治理架构
3. 数据质量管理方案
4. 大数据治理平台建设方案
5. 数据治理组织建设



第一部分

数据治理目标



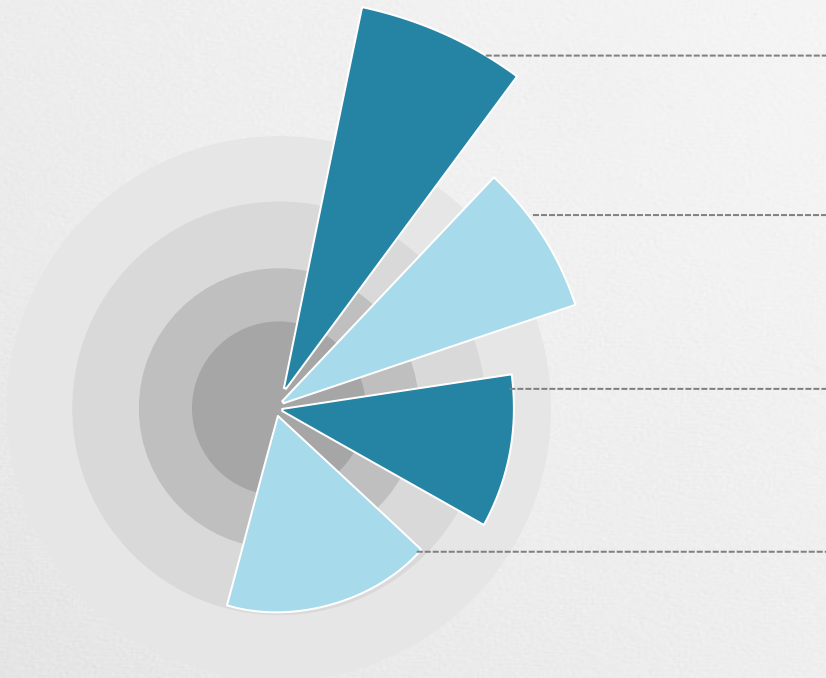


数据治理目标

数聚大咖



DAMA中国数据大讲堂系列活动



■ 提升数据质量

从数据质量的6个维度，全面提升数据质量

■ 规范数据管理

制定数据规范，持续的保障数据质量

■ 统一数据出口

建立数据权威，统一数据出口，规避数出多门的情况

■ 数据资产化

将数据资产化，为数字化转型奠定基础





■ 数据治理核心内容

战略
组织与角色
政策和标准
项目和服务
问题
估值

■ 数据治理成功实施要点

战略
组织
规范
工具
执行





常规现状问题分析

数聚大咖



DAMA中国数据大讲堂系列活动

问题1：权限散

每个人都能过来定义指标，数据部门沦为提数工具。
数据权限管理散乱，越级访问等情况时有发生

缺乏归口管理

问题2：环节多

数据产生的环节多，参与人员多，事情杂乱。数据流转程序多，控制节点多，治理难度大

缺乏制度管理

问题3：使用乱

各部门数据使用混乱，缺乏统一管控

问题4：质量差

人员管控、系统缺位、人性使然，导致人为错填、漏填、隐瞒等情况

缺乏考核管理

问题5：团队弱

缺少专业的数据治理团队

缺乏组织支撑



问题1：系统多

每个部委都有无数的信息系统，建设年代各异，使用技术多种多样，异构复杂

问题2：层级多

通常会有部、省、市三级，部分还有县、单位，一共五级

问题3：情况多

基数大，导致各种超出预想的数据质量情况比较复杂，比如重名且无身份证号

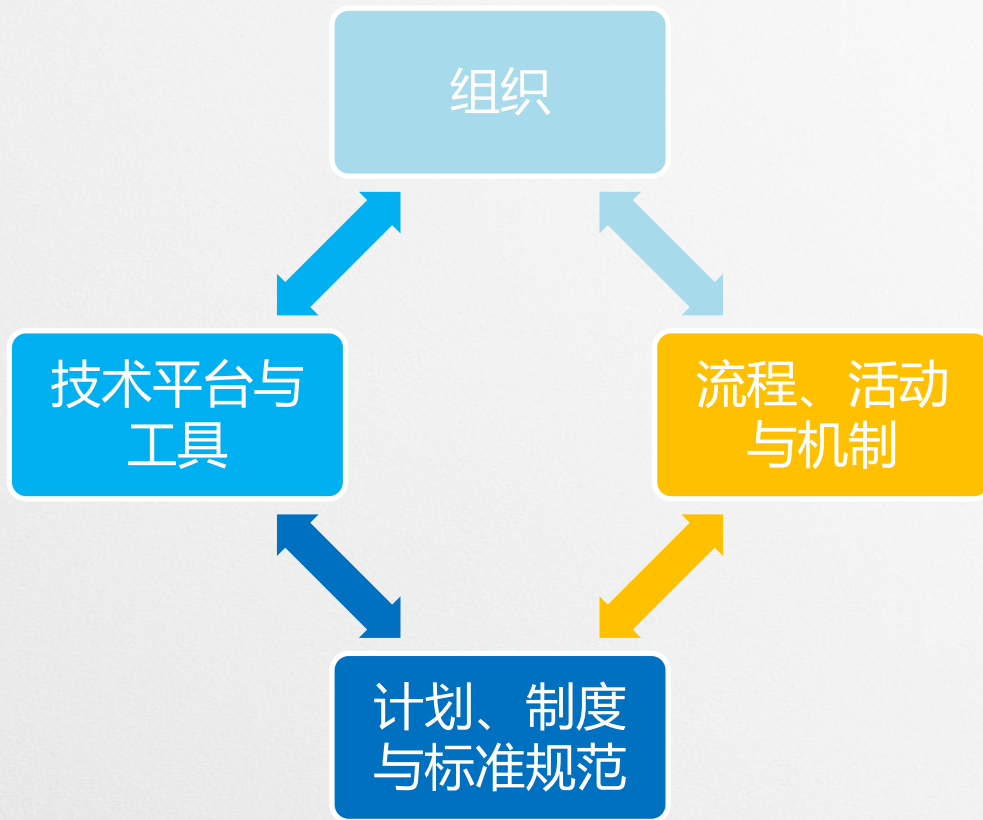
问题4：厂商多

通常一事情会扯出十多个厂商，每个厂商都是在赔本做项目，因此推动起来非常困难

问题5：奇葩多

部委弱势，常有人因为技术之外的事情闹事、上访





第二部分

数据治理架构



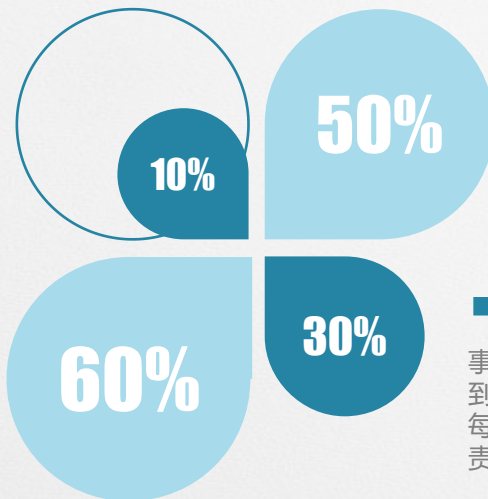


■ 设立数据治理管委会

把老板拉进来，把业务负责人拉进来，建立组织势能

■ 制定管理办法

落地执行，必须先成文，后培训，再检查，重复盘。



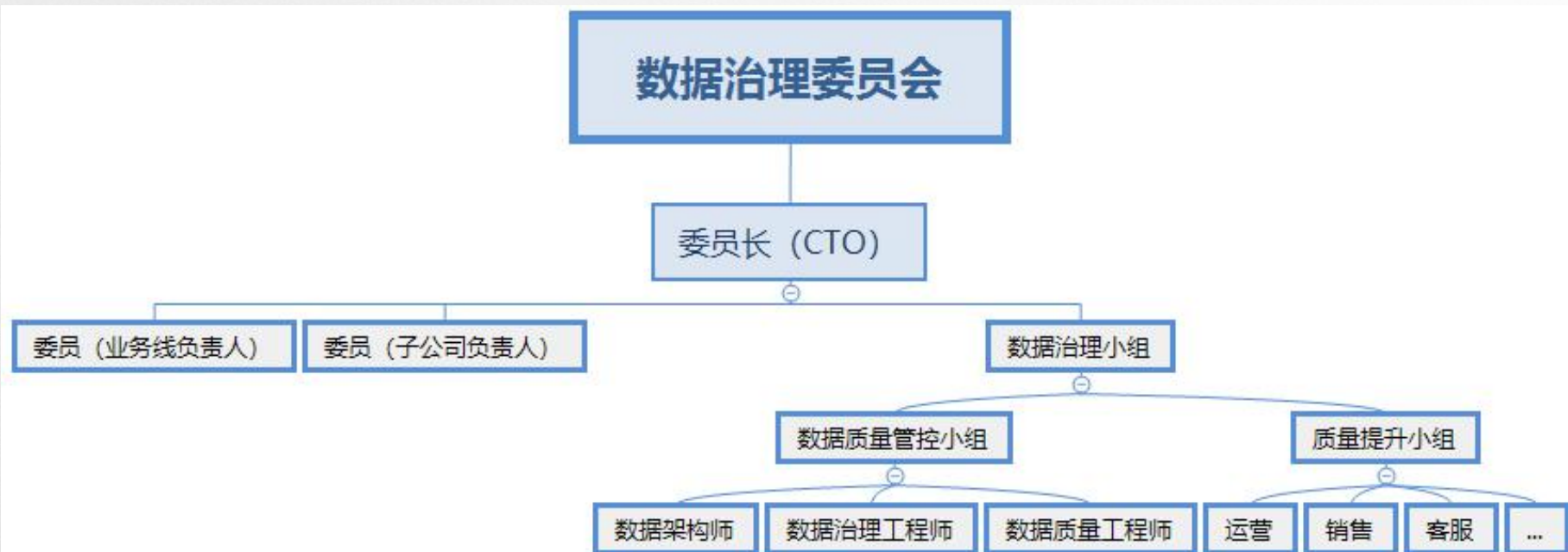
■ 建立数据治理团队

数据治理不仅仅是数据团队的任务。需要组织协同，增强其他团队的参与度。建议建立虚拟团队

■ 责任到人，落实到位

事情落实到位的前提是任务到人、责任到人。必须要给每个人分配对应的任务，且责权利对等











第三条 XXXX 数据治理应遵循真实性、完整性、安全性和合规性原则。

第三章 组织保障

第四条 XXXX 数据治理工作由 XXXX 服务部牵头，销售、客服、运营条线协助配合，具体职责如下：

第五条 XXXX 数据治理工作应建立专兼职队伍，其主要构成如下：

-  数据安全管理制度.doc
-  数据一致性校验方案.doc
-  数据质量管理流程.doc
-  数据治理管理办法.docx





制定计划

数聚大咖

DAMA 中国数据大讲堂系列活动



平台数据治理工作执行计划表

任务	工作包拆解	开始时间	结束时间	负责人
建制度	CTO牵头建立数据管理方案，指明工作方向和分阶段目标			
	明确数据管理职能，包括数据校验规则、数据质量衡量指标、治理项目沟通机制			
	分解任务到岗，责任到人。明确每一个环节的数据治理人员职责和责任			
搭团队	技术团队建立专职数据治理团队			
	业务线/子公司建立虚拟数据治理团队			
定标准	设立数据治理考核目标			
	建立数据质量指标体系			
	确定数据标准阈值			
搞建设	搭建大数据治理平台底座			
	构建数据质量监控平台			
	进行数据一致性校验			
	进行数据清洗			



第三部分

数据质量管理方案





■ 数据生产环节

梳理所有数据生产环节，对所有环节中的关键数据进行有效性校验。

对录入人员进行定点培训，加强管控，增设KPI考核机制

■ 数据加工环节

加强数据部门权威性，由数据治理委员会指定组织统一指标口径，统一数据出口



■ 数据传输环节

数据传输全链条中，增设一致性校验机制

增加系统监控，发现问题数据及时报警

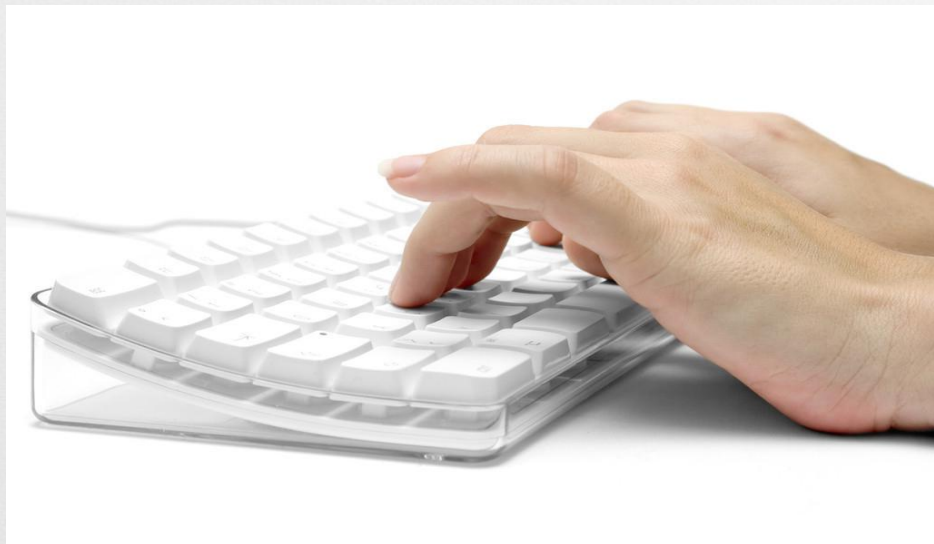




清洗篇

数聚大咖

DAMA 中国数据大讲堂系列活动



数据清洗如同米中淘沙，重点在于异常数据的区分。对关键数据进行监控，设定异常数据排查规则是数据清洗的重点

对于异常数据，应该分门别类处理，可以自动化填补的，可以进行修复；无法处理的，进行回退，由业务系统进行修正。

同时做好数据质量量化管理，督促各部门进行数据质量提升





01

已有数据的数据清洗，做一次全量排查即可。系统任务轻，但是人工工作量大。

核心在于如何协调各部门通力配合。部委向下推动，主要靠行政命令，但是命令需要合理，否则容易造成不良影响。

02

增量数据的治理，需要全面的监控，以及数据质量的量化管理，对系统要求较高。

人力工作量是持续投入的，需要各单位设置兼职的数据治理人员。



存量、增量场景不同，需要分阶段治理





协调篇

数聚大咖

DAMA中国数据大讲堂系列活动



数据治理不是信息中心/数据部门的事情，
而是整个单位的事情，需要集合全部力量

由信息中心牵头，定好计划、步骤和说明，对各省厅、
厂商进行试点和培训，做好组织保障

提前制定数据规范、标准，确定数据治理制度、岗位
要求，从管理制度层面做好保障

数据治理需要协调各厂商工作，需要耗费人力物力，
需要提前预留专项资金





考核是核心推动力

没有考核，就没有执行力。

行政命令是一回事，执行结果好坏是另外一回事



数据质量排行榜、错误数据未修复情况曝光等方法是推动数据质量提升的有效手段。

年度数据质量盘点列入工作报告中。



第四部分

治理平台建设





管理前期

制定数据质量管理的标准及规范

管理中期

对数据质量进行量化，对问题数据进行处理及跟踪

管理后期

对数据处理的结果进行统计，方便查阅





- 1、收集现有可参考标准及规范；
- 2、讨论并确定所有数据处理环节，确定数据监控节点；
- 3、收集所有关键监控事项，确认每个监控节点的监控内容；
- 4、讨论并确定数据质量量化指标；
- 5、编制《数据质量监控需求》；
- 6、编制《数据校验规则》；
- 7、开发数据质量监控程序；
- 8、设计数据质量流转流程，并开发数据质量控制前端页面；
- 9、试采数据，测试数据质量管理流程是否通畅
- 10、选择试点省份，进行试点，确认无误后，全国铺开。





数据治理架构



数据质量监测系统的系统总体架构设计主要分数据层、业务层和展现层

展现层：系统为B/S结构，以浏览器为展现工具。

业务逻辑层：包括业务应用支撑和业务应用服务，是数据质量监测的业务基础。

数据层：分为数据存储和数据处理两部分。数据存储部分由Oracle和HDFS组成。



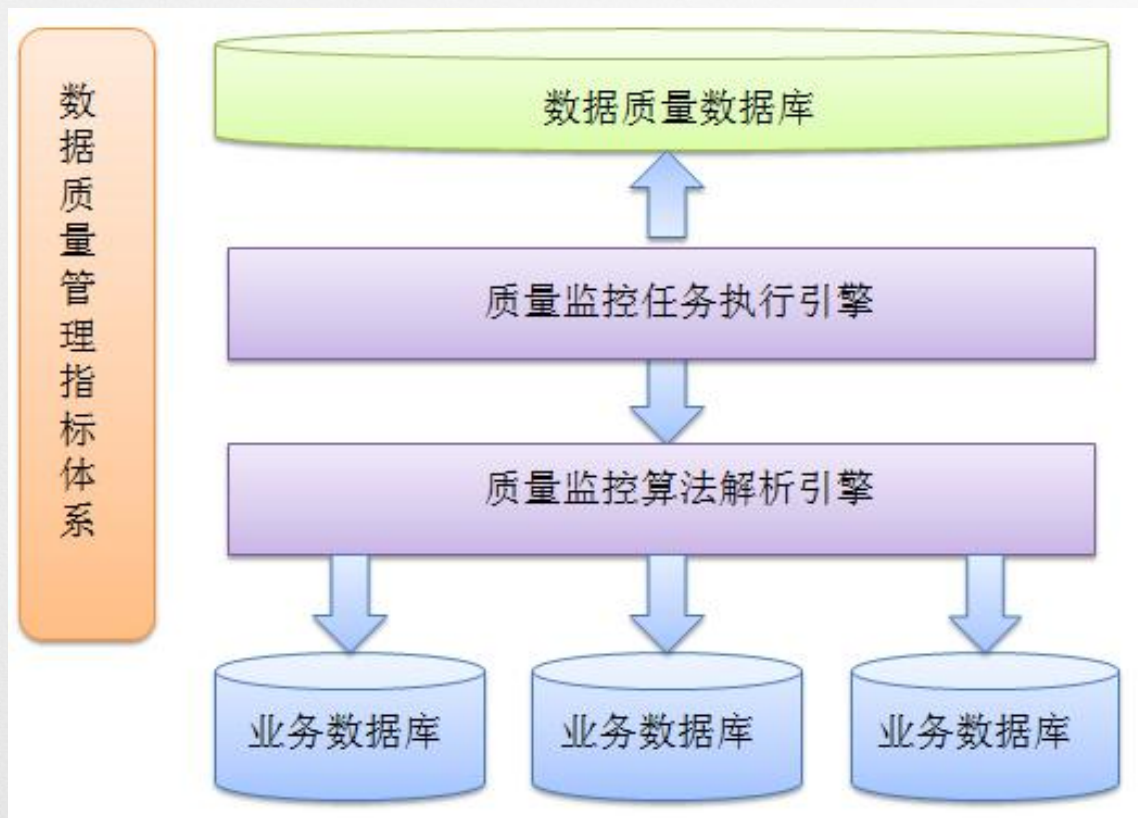


数据治理架构-数据质量监控体系

数聚大咖



DAMA 中国数据大讲堂系列活动



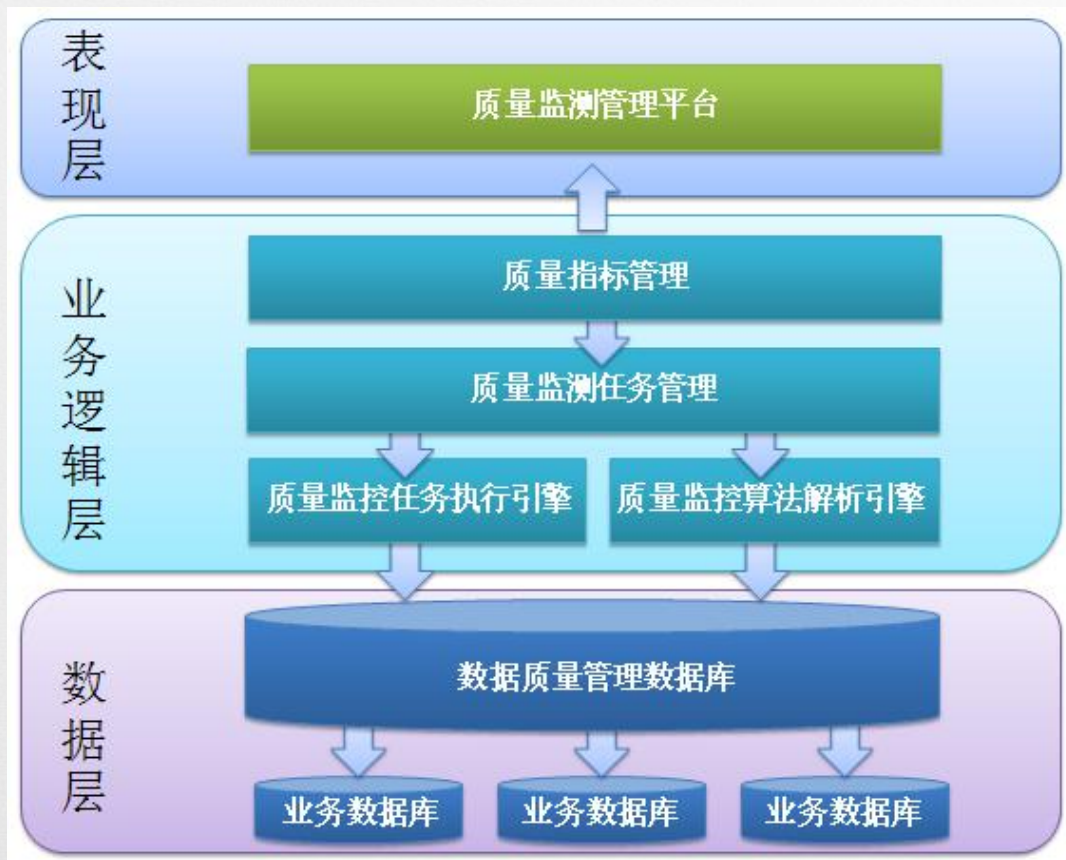


数据治理架构-数据质量监控体系

数聚大咖



DAMA中国数据大讲堂系列活动



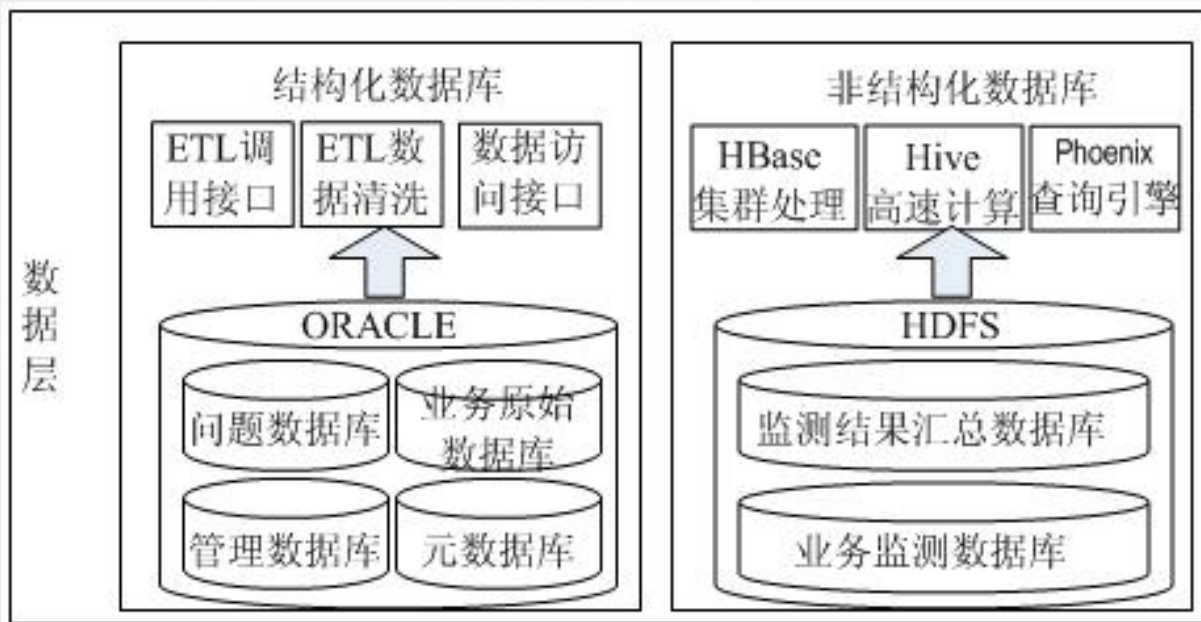


数据治理架构-数据处理层

数聚大咖



DAMA中国数据大讲堂系列活动



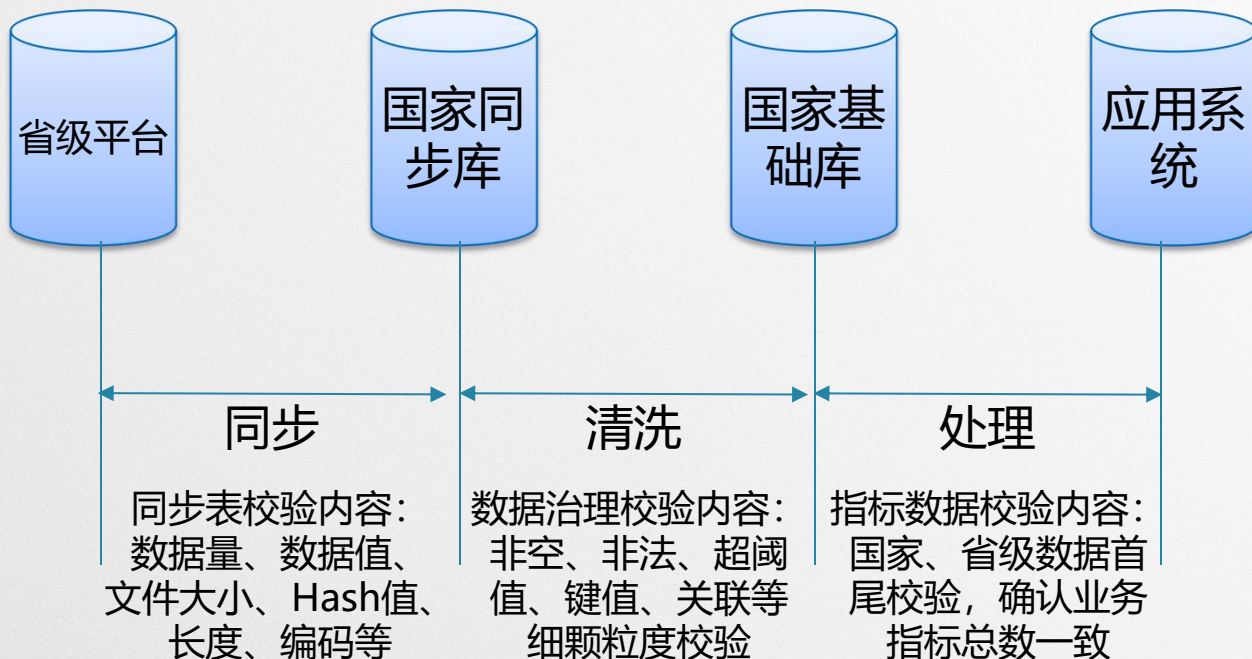


数据一致性校验

数聚大咖



DAMA中国数据大讲堂系列活动





同步表一致性校验内容

校验项	定义	校验对象	校验方案	通过标准
结构一致	对应表的表结构一致	元数据	确认表结构是否一致	一致
总量一致	对应表的记录条数一致	记录条数	全量表总记录数是否一致	一致
总数一致	对应表对应字段的值汇总结果一致	数值类字段	合计值是否一致	一致
数值一致	对应表对应字段码值对应一致	代码类字段	代码值一致	一致
Hash一致	文件、文本类数据Hash值一致	文件、文本类	Hash值是否一致	一致





业务指标一致性校验

校验项	定义	校验对象	校验方案	通过标准
指标定义一致	相同指标的内涵、外延定义保持一致	指标定义	是否按照部级标准设置	一致
统计口径一致	相同指标的统计口径、取数逻辑、范围是否一致	指标统计口径	是否按照部级标准设置	一致
统计结果一致	相同指标在相同限定范围内的汇总结果是否一致	指标值	指标计算结果一致	一致



第五部分

组织建设





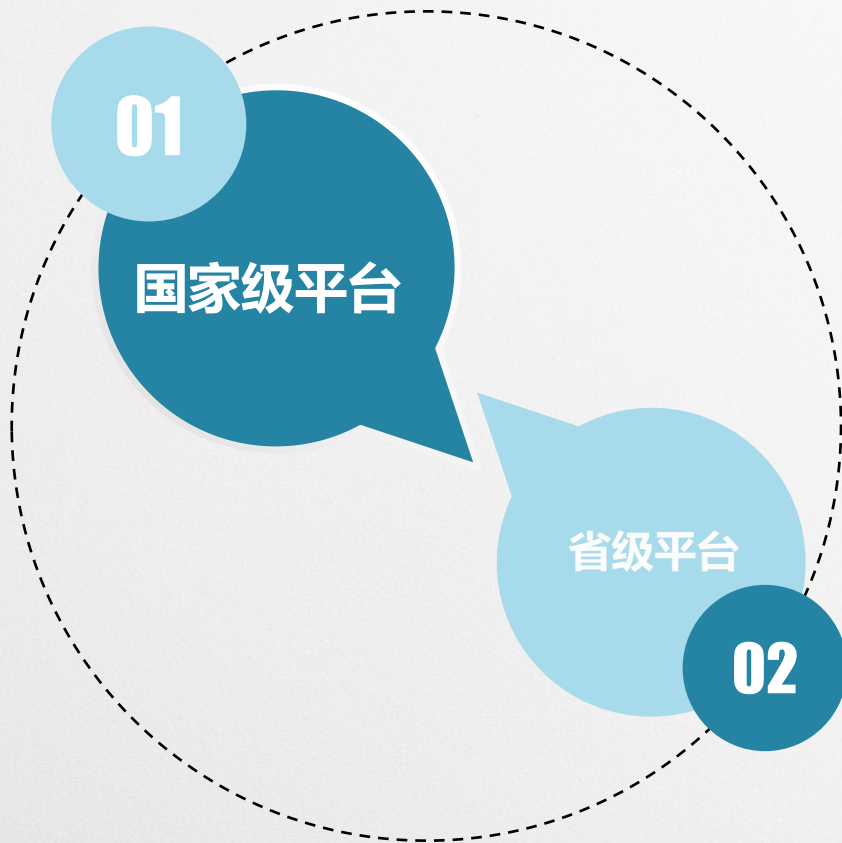
数据质量提升流程

数聚大咖



DAMA中国数据大讲堂系列活动

- 监控数据质量
- 过滤问题数据
- 下发问题数据



- 本省数据上传
- 本省数据质量监测
- 修正问题数据





数据质量提升组织建设

数聚大咖



DAMA中国数据大讲堂系列活动

01

事前发布标准

02

前期进行试点

03

加强培训

04

事中不断优化

标准先行，是部委级数据项目建设的重点。否则很容易乱成一团；
全国一盘棋不好操作，建议前期选择一个省份进行试点；

数据治理涉及多个系统、多个厂商，需要提前进行各种培训，保证上下一致；

数据治理不是一锤子买卖，需要持续优化。需要列入日常管控工作流程





胡萝卜要给到位

数据治理是精炼数据生产要素的重要工作，是数字化转型的重要基石。但是任务繁重，见效非常慢，因此需要一套有效的激励机制，引导数据治理工作顺利开展。

1	每月质量榜单
2	每年荣誉称号
3	年度报告邀请



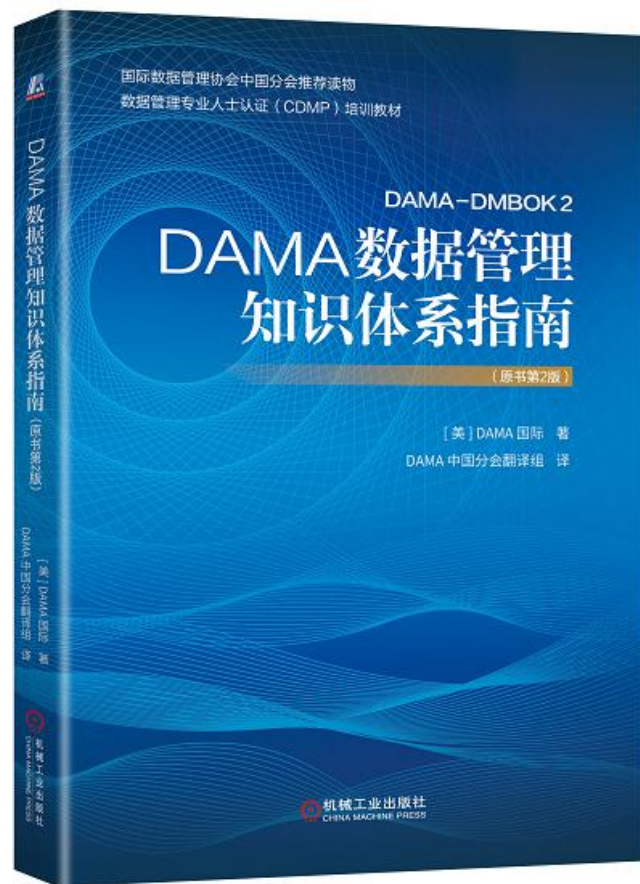
下一步： 如何挖掘数据价值？



彭文华

公众号：大数据架构师





DAMA中国分会翻译组译《DAMA数据管理知识体系指南（原书第2版）》是DAMA国际（DAMA International）组织专家对过去30多年数据管理领域知识和实践的总结，是一部综合了数据管理方方面面具有权威性的基础工具书。本书从数据治理、数据架构、数据质量、数据安全、主数据管理、参考数据管理、元数据管理、商务智能和数据参考管理、数据建模设计、数据存储和操作、数据集成和互操作、文档和内容管理、大数据、数据管理人员的道德要求等方面介绍了数据管理的知识体系。DMBOK已经成为数据管理领域的“圣经”，是指导个人知识体系完善和企业数据管理能力建设的重要文献。本书适合各种组织负责信息化和数字化转型的领导（如CIO、CDO），从事数据管理的各种技术及业务人员，工作中涉及数据的会计、法律、咨询、教育、政务等领域的人士阅读。同时，也可作为高校MBA和计算机专业教学用书。

