# 数据仓库数据质量的治理及体系构建

中国建设银行股份有限公司信息技术管理部厦门开发中心　程大庆　郑承满

在信息化应用不断深入的背景下，数据资源优势挖掘，基于数据治理提升业务响应能力等问题已经成为金融行业关注的焦点。本文主要讨论在大型银行数据仓库中构建数据质量治理体系的方法。

## 一、数据质量治理的基本内容

1.数据质量检核

数据质量检核是指通过技术手段，以数据质量指标（包含技术指标和业务指标）为标准进行检核、监控，以发现数据质量问题。以数据质量6Σ标准为基础制定数据质量指标集，再针对具体的数据集编写数据质量检核规则，即可进行数据质量检核。对于每一个数据质量指标，均可衍生多个数据质量检核规则。

如图1所示，以数据质量6Σ标准的6个特性为基础，根据数据仓库中的信息特征，划分为多个指标集如“完整性_主键重复”、“完整性_拉链错误”等；在这些指标集下，针对不同的实体和属性，形成不同的可实施的数据质量检核规则如“完整性_主键重复_客户信息表”。

通常情况下，数据质量检核规则是以制定好的数据质量指标集为基础，逐层逐块的进行质量检核；但是在实施中，经常会根据数据使用中发现的数据质量关键点或者某些业务需求，进行数据质量专项治理。

（1）数据质量指标集的制定

数据质量指标集的制定，需要考虑数据质量6Σ标准在不同数据集中的信息特征，以及数据仓库的数据架构和数据流向。在不同的数据架构和数据流向下，数据集中的数据质量信息特征不尽相同，数据质量检核重点不同，由此制定的数据质量指标集也不相同。

由于数据量、数据加载工具的处理能力和数据库引擎的处理能力不同，不同数据仓库的数据处理顺序是不同的，主要分为ETL和ELT两种模式。一般来说，数据库引擎厂商主推的是ELT模式，在数据加载入数据仓库后进行数据转换，如Teradata、Oracle；专业的ETL工具厂商主推ELT模式，如Informatica。两种模式下的数据

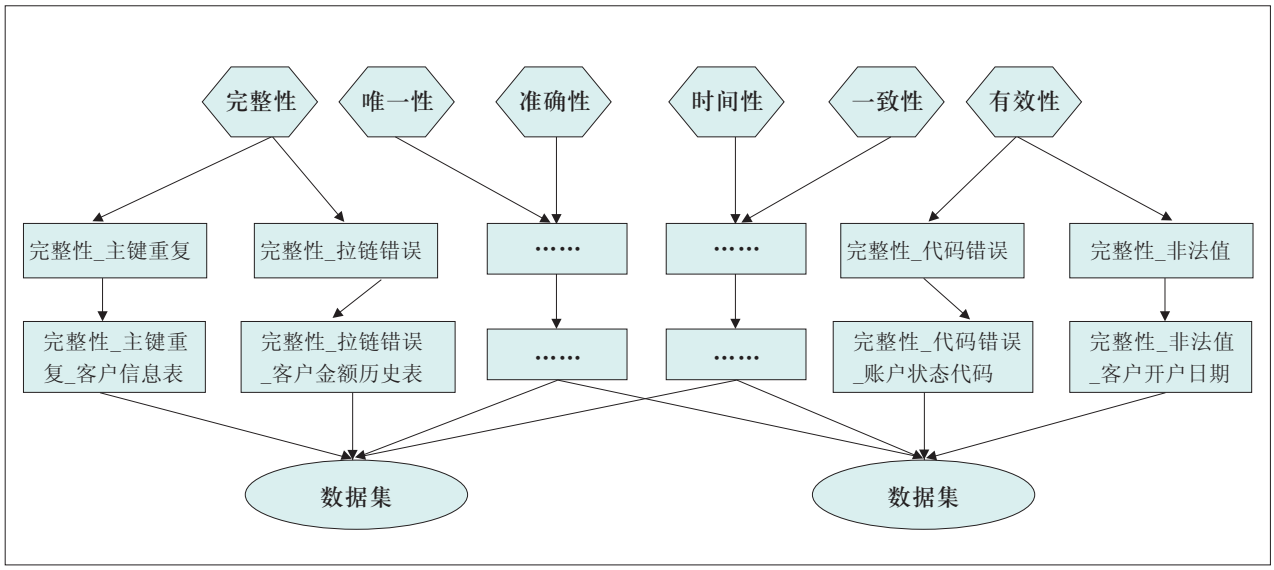


图1　基于数据集的数据质量核查规则

架构最大的不同点，在于ELT模式存在数据缓冲层，而ETL模式无数据缓冲层。数据缓冲层一般与数据源同构，用于缓冲放置数据仓库从异构数据源中获取的数据。

下面以ELT模式下的数据仓库为例讨论数据指标集的制定，该模式下数据仓库的ETL过程分为3个环节：从数据源到仓库的ETL过程，仓库内部的ETL过程,仓库到目标的ETL过程。在整个ETL过程中，数据仓库的数据架构共分为四层：缓冲层、基础层、汇总层和应用集市层。

以数据质量6Σ标准为基础，结合数据仓库的逻辑数据模型，遵循ETL过程和数据流向，分层级制定数据质量指标集即可进行立体的、全方位的数据质量检核（如图2所示）。

数据质量指标集的制定，可采用三层级制定方法：以数据质量6Σ标准的6个特性为基础，作为第一层级；将数据质量6Σ标准按数据仓库的数据层级划分，作为

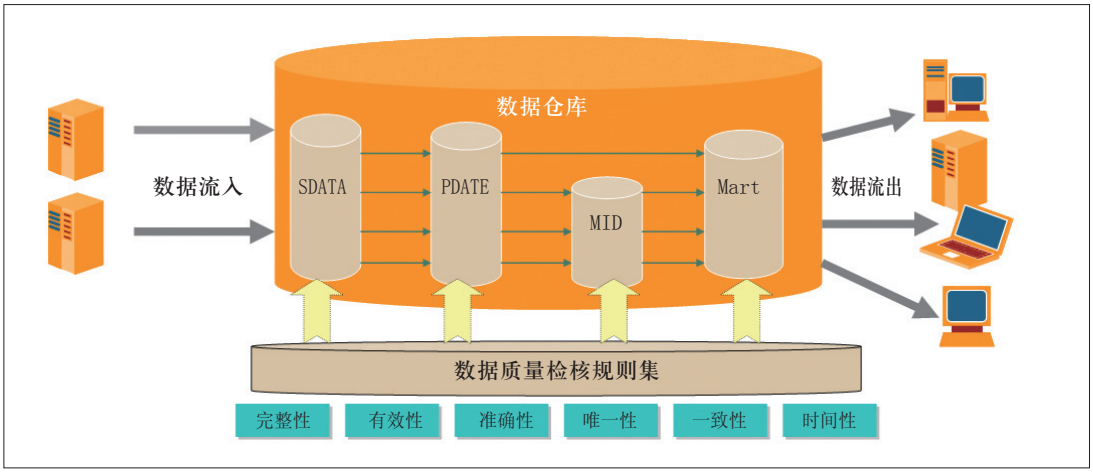


图2　数据质量检核分层级制定示意

第二层级；在数据层级上，再根据每个特性的分割，制定各层级的数据质量指标集，即第三层级。以下为某大型银行的数据质量指标集实施实例。

首先对各数据集缩写定义见表1。根据每个数据层集的数据特征，再对每个数据特性分子类，制定可实施的指标集。指标名称定义如下：数据特性_数据集缩写_特性子名称。表2为一个较为完整的、可实施的数据质量指标集。

表1　各数据集缩写定义

| 数据集中文名 | 数据集英文名 | 数据集缩写 |
|---|---|---|
| 缓冲层 | SDATA | SD |
| 基础层 | PDATA | PD |
| 汇总层 | MID | MI |
| 集市层 | MART | MA |

表2　数据质量指标集

| 指标名称 | 指标说明 |
|---|---|
| 完整性_SD_非空 | 加载入仓库缓冲层的源表非空 |
| 完整性_SD_数量 | 加载入仓库缓冲层的源表的数量正确、稳定 |
| 完整性_SD_主键 | 加载入仓库缓冲层的源表与上游源表的同时点主键值相同 |
| 准确性_SD_属性 | 加载入仓库缓冲层的源表与上游源表的同时点属性信息值相同 |
| 一致性_SD_关联 | 加载入仓库缓冲层的各张源表满足主外键、包含关系 |
| 有效性_SD_代码 | 加载入仓库缓冲层的源表的代码值符合范围 |
| 有效性_SD_属性 | 加载入仓库缓冲层的源表的属性值符合范围，如日期字段符合日期范围，金额字段符合金额范围 |
| 唯一性_SD_主键 | 加载入仓库缓冲层的源表是否主键重复 |
| 时间性_SD_时长 | 加载入仓库缓冲层的源表数据保留时长是否正确 |
| 时间性_SD_时点 | 加载入仓库缓冲层的源表数据时间戳是否与业务时点符合 |
| 完整性_PD_数量 | 基础层实体当日发生变化记录数量，是否与相应的源表中数量相同 |
| 准确性_PD_属性 | 基础层实体的关键属性值，是否与相应的源表中该属性值相同 |
| 一致性_PD_关联 | 基础层各实体间是否满足系统内和系统间的关联关系 |
| 有效性_PD_代码 | 基础层实体的代码值，是否符合定义范围 |
| 有效性_PD_属性 | 基础层实体的属性值，是否符合定义范围 |
| 有效性_PD_拉链 | 基础层采用拉链存储策略的实体，拉链是否正确 |
| 有效性_PD_离线 | 基础层实体离线数据的正确性 |
| 唯一性_PD_主键 | 基础层实体的主键是否重复 |
| 一致性_MI_关联 | 汇总层各实体间是否满足系统内和系统间的关联关系 |
| 唯一性_MI_主键 | 汇总层实体的主键是否重复 |
| 准确性_MA_业务 | 集市层数据口径是否准确反应该时点业务状况 |

根据制定的数据质量指标集，作用到各数据集中的实体以及实体中的各属性，即形成系统的、可实施的检核规则。

（2）数据质量专项检核

数据质量专项治理的针对性较强，在实施中均能带来很好的效果。下面介绍四种在数据仓库实施中常用的数据质量专项检核。

①代码检核

代码检查，即数据中的代码值是否符合代码定义的标准范围。根据制定标准的不同，代码分为两大类：全行标准代码，由信息管理部门主导、各信息系统共同参与制定的代码定义。如币种、全行机构编码等。此类代码一般都是由各源系统转换，或者由操作型数据存储系统进行标准化后下发到各个系统。对于此类代码的检查，一般是在缓冲层进行。部门标准代码，由数据仓库定义，对某类业务含义代码进行定义。此类代码的检查，一般是在基础层进行。

由于代码字段具有重要的业务含义，且在各系统的数据库表设计中均大量使用，因此是数据质量检核工作中非常重要的一环。

②总分检查

由于数据大集中的规划，同时基于集中核算、集中稽核、集中结算、集约经营的目的，目前的各业务系统，均是以核心业务系统为中心，外围的交易系统和管理系统围绕核心业务系统实现连动的业务流程和业务操作。

在此模式下，作为账务数据中心和交易处理中心的核心业务系统，外围交易系统中的分户账与核心业务系统则中的内部账、总账，核心业务系统中的分户账和总账，都是账务平衡的关系。

基于该原理，可将各系统的分户数据，按入账网点机构、科目、币种的粒度进行汇总后与核心业务系统中的总账进行总分平衡核对。

③JOIN检查

在数据仓库基础层模型设计中，基本的原则是遵循三范式理论。多个表按照一定的规则进行组合，可以非常清晰地描述企业中某类业务运行方式。但是在实际的项目开发过程中，由于连接较为损耗系统资源，因此模型的物理化往往并不严格遵循三范式，而是做适度的冗余，以减少复杂的关联。

我们在数据质量检查中发现，很多表的数据错误，其根源在于该表的进数脚本中多表关联部分的编写错误导致。通过对结果表的纯数据检查，或许可以发现此类错误，但是往往由于没有较为完善准确的检核规则，使得错误难以迅速发现。

JOIN检查是以类似于脚本测试的方式，对脚本进行批量的排查，以发现多表关联部分的编写错误。

④系统间信息核对

系统间信息核对的目的是验证数据仓库从源系统获取到的数据的过程质量。核对要点时要根据数据线的流程，以上游系统的信息为参照系，检核直接下游系统的信息是否与参照系一致，顺着数据流向，依次做系统间的检核；对于数据库操作而言，检核上游系统的各信息档的新增、修改、删除三个操作，是否有实时准确的传输到下游系统；对于实时性，则通过提取数据时点来控制；对于具体数据核对，主要围绕完整性和准确性两方面进行检核。

2.数据质量问题处理

数据质量问题处理是对数据质量问题进行分析，通过相关的程序和数据修改，使得已发现的数据质量问题得到改正，并且在管理、设计、开发、维护等流程和规范上进行改进，以使同类型数据质量问题在后续得到杜绝，或者针对同类型问题形成有效的预防和监控措施。通常情况下，数据质量问题的处理被更狭义的认为是对质量问题的技术处理和数据处理。实际上，在数据仓库实施过程中，针对数据质量问题产生根源所做的在工作规范和工作流程上的改进，对于预防和杜绝同类型数据

质量问题具有决定性作用。

由于数据的绝对数据质量往往是由业务系统和管理系统产生，对于这类问题，数据仓库并不需要关心问题处理的技术细节，只需关注数据表现和处理结果。因此，在数据质量问题的处理流程中，对于非数据仓库错误问题的处理方法，不必过多关注细节，数据仓库项目以管理、协作的角色出现。

数据质量问题的处理过程中一般包含以下三个内容。

（1）问题发起和分析，检核人员把数据质量问题报送数据质量管理人员，之后由专业的数据质量分析人员根据数据特征、数据模型、ETL日志、运维日志、相关的业务和技术资料，综合分析以确定问题的来源；若问题涉及外部项目组，则还需通过商定的工作模式，请外项目组协作分析。

（2）问题处理，根据问题的处理可行性及处理必要性，有以下三种处理模式：对于可解决的问题，则由责任方进行程序修改、数据修复、相应的工作流程和工作规范调整、数据修正，同时整条数据流线上的系统，均应进行数据修复；对于不需解决或是无法解决的数据质量问题记录在案，根据需要在数据上打标记；对于需要解决但是暂时无法查找出原因的数据质量问题进行监控，待情况重现后进一步分析处理。

（3）问题验证总结，每个问题处理完成后，应进行结果的正确性检验。并提取出重要的检核规则，进行日常监控。

一个典型的数据仓库数据质量处理流程，由以下四个环节组成：问题发起、问题分析、问题处理、问题验证。由数据质量管理人员总体控制、流转，协调各方进行相关工作，参与每一个环节（如图3所示）。

在处理过程中，为了便于数据质量问题的处理、流转和记录，一般需要建立以下文档：数据质量检核结果登记簿，用于数据质量检核人员登记检核结果；工作联系单，用于项目组内或项目组间关于数据质量问题进行沟通、问题流转；数据质量问题跟踪单，用于记录数据质量问题的整个发起、分析、处理、验证的完整过程信息；数据质量问题的处理计划和处理的方案，当责任方为外项目组时，一般不关注处理方案的细节。

3.数据质量知识库

数据质量知识库是指通过在数据质量检核工作、数据质量问题处理工作中的知识提取和规则发现，形成数据质量知识库，再以数据质量知识库来指导数据质量治理工作的调整和延拓。数据质量知识库以规范化的自然语言或形式化语言编写，也可以包含可执行的程序语言。主要包含业务规则、技术规则、典型的分析方法、

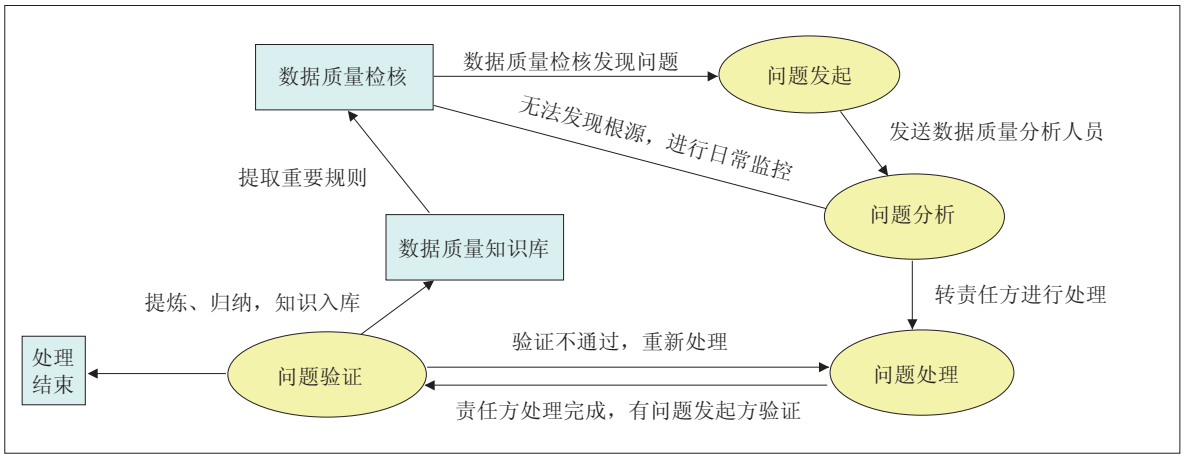


图3　数据质量处理流程

典型的程序错误和修改方式、典型的模型调整和数据处理方法等。

数据质量知识库的形式不限，一般由数据质量工作人员在数据质量问题的处理和验证通过后，进行归纳总结以发现知识和规则，扩充知识库。

4.数据质量评估

数据质量评估是对数据质量进行系统的量化评估，以评估某类数据集的数据质量整体情况。基于数据质量指标集，可以对数据质量进行定量的分析，进行系统级、主题级的评估，以对于数据质量治理工作做定量的事前评估和事后验收。数据质量评估基于以下的五元评估模型和评估公式。

（1）数据质量评估模型

该模型是一个五元组，即M＝<D,I,R,W，S>。D表示需要进行评估的数据集。对于关系数据库来讲，一个数据集相当于一类表或视图。I表示数据集D上需要进行评估的数据质量指标集。R表示与评估指标相对应的数据质量检核规则。W表示赋予规则R的权值（大于0的整数），描述了该规则在所有规则中所占的比重，一般实施中同一质量指标下的规则都赋与同一权值；S表示规则R对应的最终结果，一般取值为（正常数/总数）*100。

（2） 评估公式

在确定出指标集和权系数之后，评估公式为

$$SA=\frac{\sum_{i=1}^{n}W_iS_i}{\sum_{i=1}^{n}W_i}$$

（3） 评估步骤

数据质量评估中，一共以下4个步骤：确定评估数据集；选择评估数据质量指标集；制定规则集，并给出各规则的权值；根据评估模型和评估公式，计算规则结果得分。

## 二、数据质量治理体系的构建

从上文可以看出，数据质量检核是数据质量治理工作的基本驱动因素，通过检核发现数据质量问题后的问题处理和知识库的扩充，都是较为明确的、可规范化的工作过程。

根据上面给出的数据质量检核指标集，我们不难想到，最理想的工作方式就是根据数据质量检核指标集，编写针对全数据集的检核规则并进行日常监控，以尽可能的发现和预防数据质量问题。但是这种方法，将会极大的损耗系统资源和人力资源，在数据仓库的实际实施中并不可取。

正如数据仓库本质上是一个动态的建设过程，数据仓库的数据质量治理工作也是一个动态的工作过程。虽然始终是以6Σ标准为数据质量的检核基础，但是数据质量工作的实施人员、对象、方式、目标，都会随着数据仓库的建设发展而变化。

1.数据质量治理基本方法

一个常用的数据质量治理的基本方法为PDCA法，共分为四个步骤：P计划 →D执行 → C检查 →A行动。这四个步骤构成一个闭环，是一个逐步扩充、循环发展的治理方法。P表示在数据质量指标集中选择某一类指标准备改进。D表示在部分数据集上执行数据质量检核、数据质量评估、数据质量问题处理，以改进数据质量。C表示通过数据使用反馈和数据质量评估来验证质量改进效果。A表示制定数据质量改进指引，对所有相关数据集进行质量改进，并验证改进情况。

2.数据质量治理的三个发展阶段

在数据仓库的建设中，数据质量工作往往需经历以下三个阶段。

（1）被动处理问题阶段

该阶段往往处于数据仓库建设的初期，在数据仓库各数据集中各系统数据集成度不高，模型设计尚不完善，与各OLTP系统、ODS系统、DSS系统尚处于磨合期，基于本行IT系统群特征的工作规范和工作模式尚未完全成熟。由于在项目早期，各方面资源和人力主要用于数据仓库的开发、建设，对于数据质量工作，尚无足

够的资源投入。对数据质量的检核，主要由应用集市人员在数据使用中实施。本阶段的数据质量工作以解决影响各应用集市数据使用的问题为主，其工作目标是尽快解决已发生的影响应用集市区数据使用的问题。

（2）主动治理问题阶段

数据仓库建设的成熟期，模型设计成型，重要数据持续整合、集成入仓，各系统内和系统间的工作规范和工作模式已经成熟。后续应用集市对数据质量要求进一步提高，对数据质量的检核，主要由数据质量人员和应用集市人员实施，少部分由设计、分析、测试、维护人员实施发现。本阶段的数据质量工作，逐步由主要处理应用集市区的数据质量问题处理，调整为对数据缓冲区和基础区的数据质量问题主动发现和处理。本阶段的数据质量工作目标是在数据集市的数据使用受到影响前，发现并解决问题。

（3）预防出现问题阶段

数据仓库进入运行稳定期，重要数据均已入仓，基础数据区已基本稳定；模型设计、工作规范、工作模式经过回顾、验证、调整，已完全成型。越来越多的业务用户对数据仓库进行大量的动态查询，对数据仓库的数据质量要求，也越来越高。对数据质量的检核，主要由数据质量人员、设计分析人员、测试人员实施。本阶段的数据质量工作，逐步由生产环境的在线数据质量问题发现，调整为在设计、分析、开发、测试中的对数据质量问题的主动发现和处理。本阶段的数据质量工作目标是在数据质量发生于生产环境之前，发现并解决问题。

虽然三个阶段的数据质量工作重心、目标均不相同，但是各阶段的工作内容并非互斥、独立的，而是迭加扩张的。对于数据集市区的日常检核监控，在第三阶段仍会实施。对于基础区、缓冲区的数据检核，若有足够资源，也可在第一阶段同步实施。在进入新的阶段后，上一阶段的工作内容并不会完全取消，而是选择出重要部分，仍作为日常工作。但数据质量工作中的被动处理问题、主动发现问题、预防出现问题的发展过程，是符合数据仓库实施的建设趋势的。

以6Σ标准为基础，制定数据仓库的数据质量指标集，以PDCA法为数据质量治理基本思路，根据数据仓库的不同发展阶段，由不同的人员从数据指标集中选择不同的指标进行实施，并遵循规范的数据质量处理流程进行处理，即构建出了一个动态、有效、发展的数据仓库的数据质量治理体系（如图4所示）。

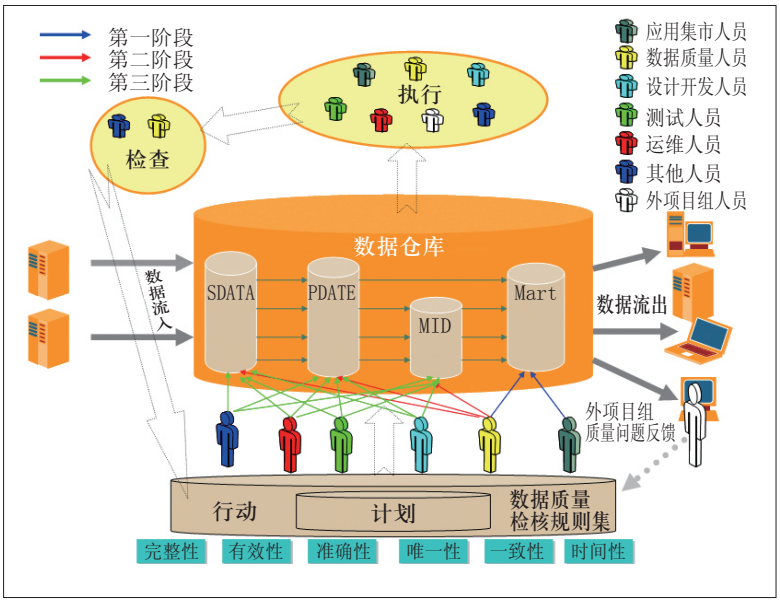


图4　数据质量治理体系

数据质量的改进是一个持续不断的过程，数据质量的确保，是数据仓库价值得到充分体现的确保。数据质量的提高，并非只是数据质量人员的职责，需要各环节各岗位的共同努力。FCC