ICS 35.020
L 70

DB15

内 蒙 古 自 治 区 地 方 标 准

DB15/T 1873—2020

大数据平台 数据接入质量规范

Data access quality specification for big data platform

2020-04-03 发布　　2020-05-03 实施

内蒙古自治区市场监督管理局　　发布

目　　次

前　言

本标准按照GB/T 1.1—2009给出的规则起草。

本标准由内蒙古自治区大数据发展管理局提出。

本标准由内蒙古自治区大数据发展管理局归口。

本标准起草单位：内蒙古自治区大数据发展管理局、新华三技术有限公司、中国电子技术标准化研究院、内蒙古自治区大数据与云计算标准化委员会、内蒙古自治区标准化院、内蒙古银保监局、内蒙古自治区地图院、内蒙古自治区电子信息产品质量检验院、内蒙古大学、浪潮软件集团有限公司、内蒙古跃晨科技有限公司、北京东方金信科技有限公司、北京东方国信科技股份有限公司、天帆创新（北京）科技发展有限公司、同方知网（北京）技术有限公司、内蒙古纵横云技术有限公司。

本标准主要起草人：张建军、崔连伟、孙卫、石彦龙、周佳琪、李向前、石建军、巩韶飞、顾君、武茂春、卫凤林、马学彬、徐小强、万磊、张晓磊、王楠、李建文、刘玉坤、冯国忠。

大数据平台　数据接入质量规范

1　范围

本标准规定了大数据平台数据接入过程中数据质量保障的规范及技术要求。

本标准适用于内蒙古自治区各数据提供单位接入大数据平台过程中的数据质量保障工作。

2　规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件；

GB/T 5271.1　信息技术词汇 第1部分：基本术语

GB/T 36344　信息技术 数据质量评价指标

3　术语和定义

GB/T 5271.1、GB/T 36344和SY/T 6227-2005界定的术语和定义适用于本文件。为了便于使用，以下重复列出以上标准的一些术语和定义。

3.1

数据　data

信息的可再解释的形式化表示，以适用于通信、解释或处理。

注：可以通过人工或自动手段处理数据。

[GB/T 5271.1-2000，定义01.01.02]

3.2

元数据　metadata

关于数据或数据元素的数据（可能包括其数据描述），以及关于数据拥有权、存取路径、访问权限和数据易变性的数据。

[GB/T 5271.1-2000，定义17.06.05]

3.3

数据质量　data quality

在指定条件下使用时，数据的特性满足明确的和隐含的要求程度。

[GB/T 36344-2018，定义2.3]

3.4

原始数据　raw data

终端用户所存储使用的各种未经过处理或简化的数据。

注：原始数据有多种存在形式，如文本数据、图像数据、音频数据或者几种数据混合存在。

[GB/T 36344-2018，定义2.4]

3.5

数据生命周期　data life cycle

将原始数据转化为可用于行动的知识的一组过程。

[GB/T 36344-2018，定义2.5]

3.6

数据集　data set

具有一定主题，可以标识并可以被计算机化处理的数据集合。

[GB/T 36344-2018，定义2.6]

3.7

数据标准　data standard

数据的命名、定义、结构和取值规范方面的规则和基准。

[GB/T 36344-2018，定义2.8]

3.8

检核任务　inspection task

数据核查的最小调度单位。

4　概述

大数据平台支持结构化数据、半结构化数据和非结构化数据等异构数据源采集数据，实现各类离线数据、实时数据的采集与接入。针对大数据平台的数据接入，执行相应的质量评价标准，保证数据质量，为各数据使用单位提供优质的数据服务。

5　数据质量评价维度

数据质量是保证数据应用和提供优质数据服务的基础，数据质量的评估标准主要包括八个维度：完整性、规范性、一致性、准确性、唯一性、关联性、时效性、可访问性，本标准从以上八个维度评价数据质量，详细请参考附录A。

6 数据接入质量技术要求

大数据平台为了保障数据质量，须从四个层面进行数据质量控制，数据标准体系质量控制、数据采集质量控制、数据存储质量控制、数据使用质量控制进行全流程数据质量管控。数据接入质量整体框架图见图1：

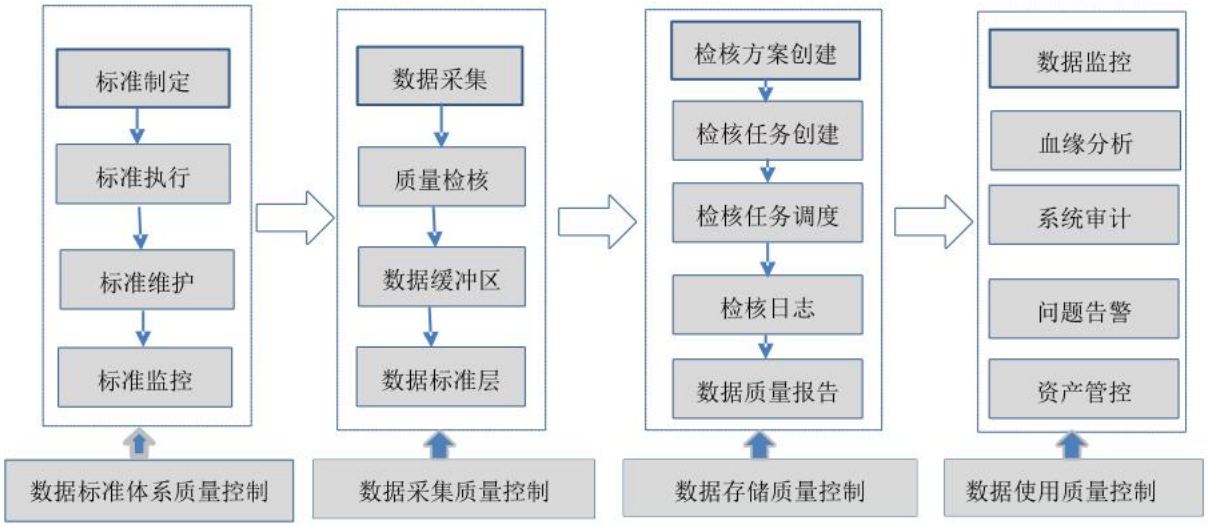

图 1 整体流程图 6.1 数据标准体系质量控制

6.1.1 数据标准制定

数据标准的制定按照数据标准管理的业务分类和定义规范指导要求，基于行业数据管控需求，进行数据标准规范的制定,要求大数据平台按照该标准规范进行统一的数据管理。

数据标准制定包括数据标准的编制、数据标准的审核、数据标准的发布。数据标准化管理组织将数据标准以正式发文的方式在内部进行发布，并在发布后将数据标准、版本说明保存备案。最终将发布的数据标准更新至数据标准管理模块中，数据标准制定流程见图2：

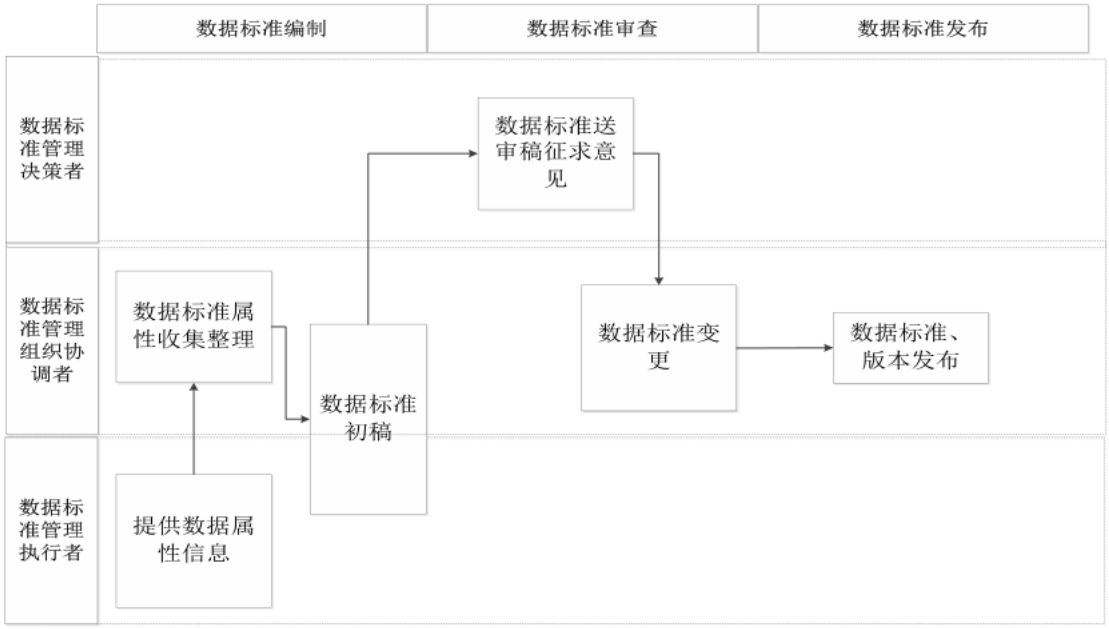

图 2　数据标准制定流程

数据标准制定流程描述如下：

a）　数据标准管理组织协调者组织数据提供者和执行者参与数据标准属性的收集和整理工作，并按照协商一致的原则形成数据标准初稿；

b）　数据标准初稿进行多次的讨论和丰富后，形成数据标准送审稿提交给数据标准管理决策者；

c）　经过数据标准管理决策者的讨论审核后，由数据标准管理组织协调者再次进行数据标准的修改完善，并完成数据标准的发布。

6.1.2 数据标准执行

数据标准管理执行流程见图 3。

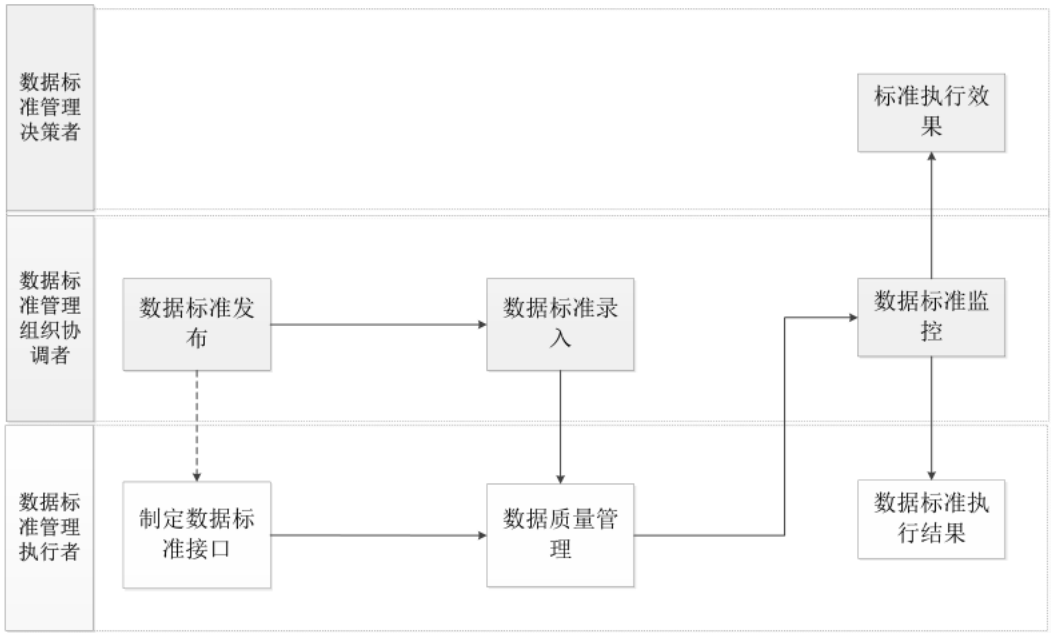

图 3　数据标准执行流程

数据标准执行的流程步骤描述如下：

a) 数据标准制定发布后，将数据标准录入到数据标准管理模块；
b) 数据标准管理执行者按照发布的数据标准，制定和发布数据标准接口；
c) 数据标准管理模块将标准要求提供给数据质量管理，根据已录入系统的数据标准形成稽查规则，对数据标准管理执行者制定和发布的数据标准接口中的内容进行相关的标准稽核监控；
d) 将标准稽核结果发送给数据标准管理模块，并反馈给数据标准管理决策者和数据标准管理执行者。

6.1.3 数据标准维护

数据标准的维护指数据标准建立后，根据业务需求的发展变化或外部数据标准要求不一致时，对数据标准的内容进行变更和版本管理，见图 4：

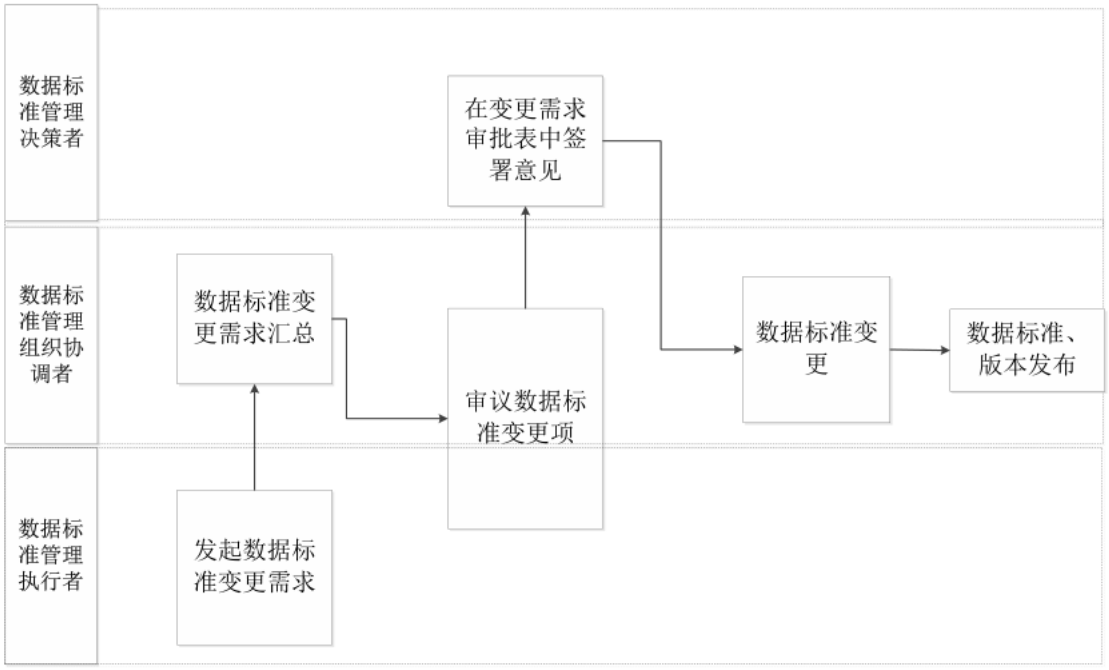

图 4 数据标准维护流程

数据标准维护流程描述如下：

a) 对执行的相关数据标准进行变更请求的申请，组织该数据标准相关执行层和各数据运维者进行讨论和变更需求汇总；
b) 由数据标准管理组织协调机构进行标准变更的审核；
c) 讨论审议数据标准项的变更内容，并形成标准变更需求审批表提交给数据标准管理决策层进行审批；
d) 决策层将审批结果反馈给标准管理组织协调者，并由其进行数据标准发布及版本维护。

6.1.4 数据标准监控

数据标准监控实现对数据标准执行过程的监控，包括对数据标准的执行、效果、问题进行监控管理，为后期数据标准维护管理提供依据。数据标准的监控通过数据标准管理和元数据管理、数据质量管理协同实现落地。

6.2 数据采集质量控制

为了保证数据质量，根据数据仓库建设的总体目标和设计对数据的采集阶段提出严格、明确的质量要求及必要的标准，具体要求如下：

a) 待采集数据字段长度、精度、类型等应优先遵循遵循国家标准、行业标准的约定；
b) 采集过程支持元数据的保留，包含技术元数据和业务元数据；
c) 支持对元数据的监控，控制数据库和表结构的异常修改，保证数据质量；
d) 支持采集阶段初步数据检核；

采集阶段具体流程图见图5：

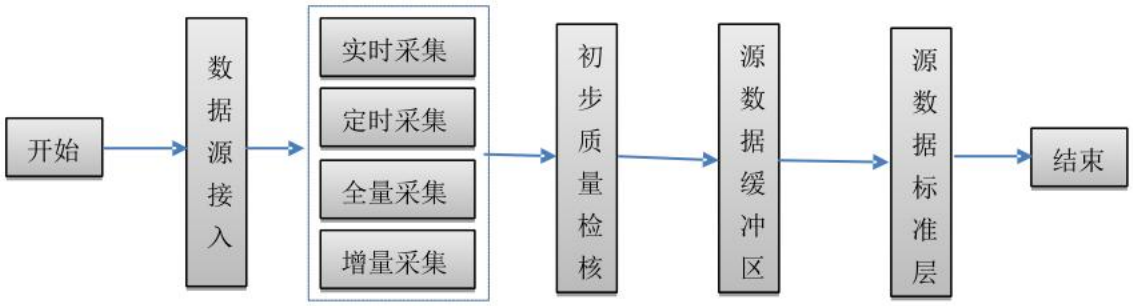

图 5 采集阶段流程

6.3 数据存储质量控制

在数据存储阶段需通过多种检核规则及任务调度方式对数据进行检核，数据存储阶段要求如下：

a) 支持从 5 个维度、7 种检核规则和自定义检核规则对数据进行数据质量检核；
   1) 5 个维度包含完整性、规范性、准确性、唯一性、关联性；
   2) 7 种检核规则包含空值校验、值域校验、格式校验、长度校验、精度校验、唯一性约束校验、主外键校验；
   3) 自定义检核规则指根据具体业务场景，用户可以通过自定义 SQL 语句的方式完成对数据质量的检核；
b) 支持检核任务的创建，检核规则的设定；
c) 支持检核任务的创建和检核任务调度方式的设定；
d) 检核任务调度支持自动调度和手动调度；
e) 支持对数据质量报告的查看的下载；
f) 支持数据的全生命周期管理；
g) 支持对元数据的版本管理。

基于检核规则对数据检核流程图见图 6：

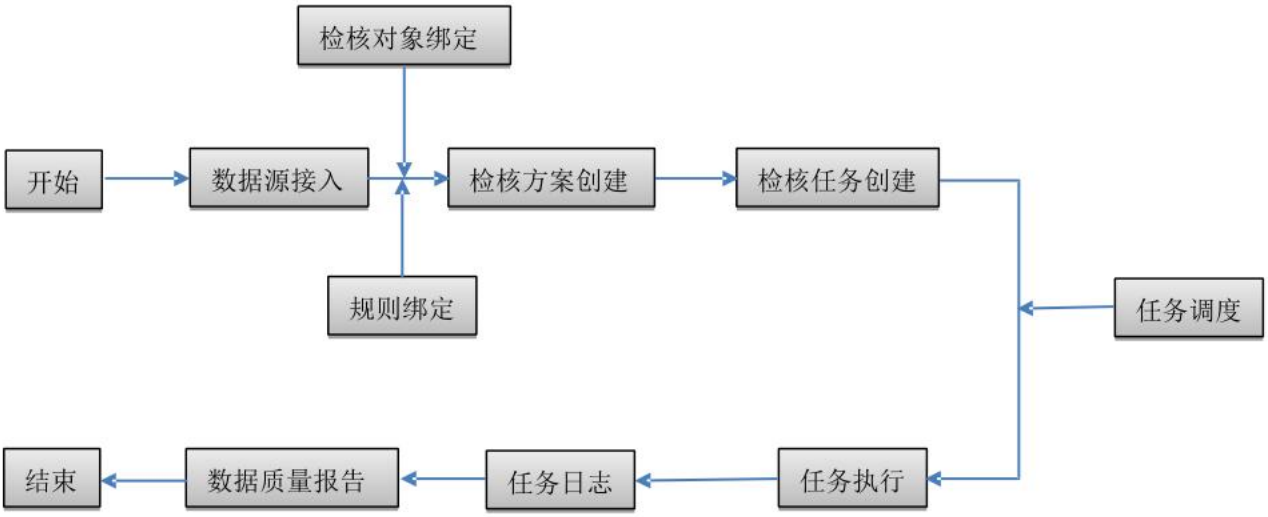

图 6　数据检核

质量检核流程说明：

a）　接入待检核数据源；

b）　创建检核方案；

1）　针对系统内置的检核规则，选择检核规则，具体包含空值校验、值域校验、格式校验、长度校验、精度校验、唯一性约束校验、主外键校验；针对自定义检核规则，通过自定义 sql 语句实现检核规则；

2）　确定待检核的对象，即选取待检核的数据库、待检核的表以及字段；

c）　关联检核方案，创建检核任务；

d）　配置检核任务的调度方式，可通过定时器实现自动调度，亦可通过人为实现手工调度；

e）　任务被调度后是查看任务执行日志；

f）　支持生成数据质量报告并提供下载功能，基于质量报告可实现异常数据发现并处理。

6.4　数据使用质量控制

数据使用要求如下：

a）　支持对数据进行监控，明确数据的来源和去向；

b）　支持数据地图、血缘分析、影响分析等方式的数据展现；

c）　支持对数据资产的安全性管控；

d）　支持对操作日志以及用户登录、退出的日志审计；

e）　支持数据质量报告分析的查看及下载；

f）　支持问题数据的告警；

g）　支持对问题数据进行整改分析。

数据使用阶段流程图见图7：

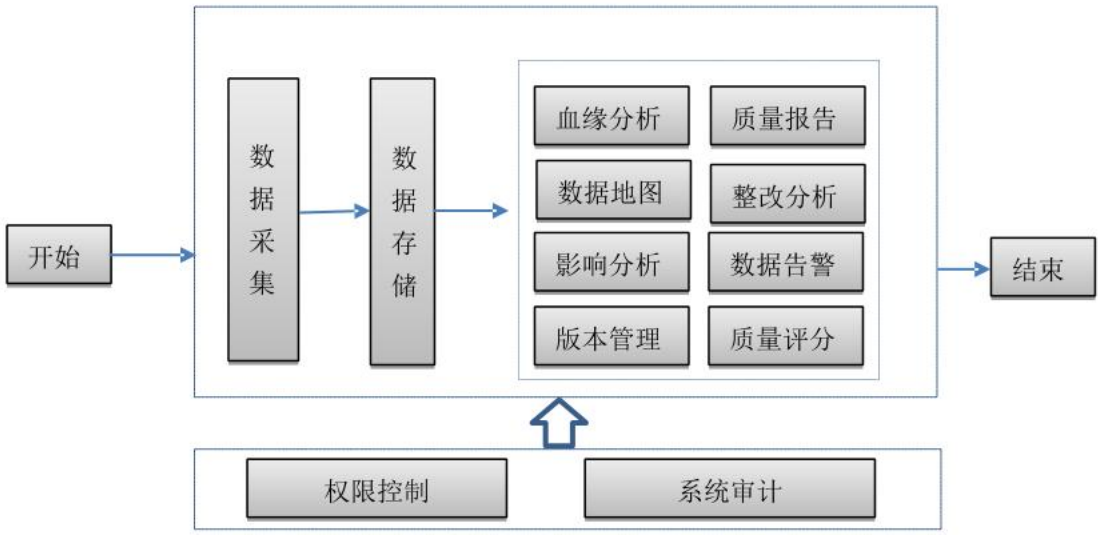

图 7　数据全流程监控流程图

数据使用阶段流程说明：

a) 在数据采集和数据存储基础上，实现血缘分析、数据地图、影响分析、版本管理、质量报告、问题数据整改分析、数据告警、质量评分；

b) 使用阶段支持数据及功能的权限控制；

c) 支持系统审计，包含操作日志审计和登录登出日志审计；

7　数据质量评分方法

7.1　规则评分

$$R = (1 - I/D)*(100\text{-W}) \cdots\cdots (1)$$

式中：

R ——每个检核规则得分；

I ——数据集D上被检核出的异常数据总量；

D ——需要进行检核的数据集，对于关系型数据库来说，一个数据集由若干条记录组成；

W ——规则对应权重值，需人工设置每个规则的权重。

7.2　任务评分

$$T = \frac{\left[\sum_{i}^{n}\left(1-\frac{I_i}{D_i}\right)*(100-W_i)\right]}{n} \qquad n = (1,2,3,4,5 \ldots\ldots) \cdots\cdots (2)$$

式中：

T ——检核任务评分；

I ——数据集D上被检核出的异常数据总量；

W ——规则对应权重值，需人工设置每个规则的权重；

D ——需要进行检核的数据集，对于关系型数据库来说，一个数据集由若干条记录组成；

n ——条检核任务中检核规则的个数。

注：一个检核任务可包含多条检核规则。

7.3 绩效评分

$$S_a=(1-\text{错误数据总量}/\text{检核数据总量})*W_a$$
$$S_b=(1-\text{错误数据表数量}/\text{检核数据表数量})*W_b$$
$$S_c=(1-\text{错误数据表数量}/\text{照管数据表数量})*W_c$$
$$P=(S_a+S_b+S_c+S_d)/4 \quad \cdots\cdots\cdots\cdots (3)$$

式中：

S_a ——评分计算指标之数据问题评分；

W_a ——评分计算指标----数据问题权重；

S_b ——评分计算指标之检核指标问题评分；

W_b ——评分计算指标----检查指标问题权重；

S_c ——评分计算指标之指标问题评分；

W_c ——评分计算指标----指标问题权重；

P ——照管人对应数据库评分；

S_d ——评分计算指标之自定义评分。

附录 A
（资料性附录）
数据质量评价维度

A.1 完整性

按照数据规则要求，数据元素被赋予数值的程度。即完整性指的是数据信息是否存在缺失的状况，数据缺失的情况可能是整个数据记录缺失，也可能是数据中某个字段信息的记录缺失。不完整的数据所能借鉴的价值会大大降低，完整性是数据质量评估标准的基础。

表 A.1 完整性评价指标

序号	指标名称	指标描述	计算方法
1	数据元素完整性	按照业务规则要求，数据集中应被赋值的数据元素的赋值程度。	计算公式：X=A/B 式中 A=被赋值的数据集中元素的个数； B=预期被赋值的数据集中元素的个数
2	数据记录完整性	按照业务规则要求，数据集中应被赋值的数据记录的赋值程度。	计算公式：X=A/B 式中 A=被赋值的数据集中元素的个数； B=预期被赋值的数据集中元素的个数

A.2 规范性

数据符合数据标准、数据模型、业务规则、元数据或权威参考数据的程度。

表 A.1 规范性评价指标

序号	指标名称	指标描述	计算方法
1	数据标准	数据符合数据标准的度量。 注1：评价数据质量时需要收集数据在命名、创建、定义、更新和归档时遵循的标准，包括国际标准、国家标准、行业标准、地方标准或相关规定等。 注2：和数据归档一样甚至更重要，在一个完整的数据规则中，旧数据的销毁一般也有一个比较详细且具有可行性的规定。	计算公式：X=A/B 式中 A=满足数据标准要求的数据集中元素的个数； B=被评价的数据集中元素个数
2	数据模型	数据符合数据模型的度量。 注1：数据模型是一种直观描述组织数据结构的手段，是数据表达的规范。 注 2：评价数据质量时需要检查是否存在清晰且可理解的数据模型定义以及这些数据的组织形式。	计算公式：X=A/B 式中 A=满足数据模型要求的数据集中元素的个数； B=被评价的数据集中元素个数

表 A.2 （续）

序号	指标名称	指标描述	计算方法
3	元数据	数据符合元数据定义的度量。 **注1：** 元数据标注、描述或刻画其他数据、以使检索或使用数据更容易。评价数据质量时需要检查是否提供可解读的元数据文档。	计算公式：X=A/B 式中 A=满足元数据定义的数据集中元素的个数； B=被评价的数据集中元素个数
4	业务规则	数据符合业务规则的度量。 **注1：** 业务规则是一种权威性原则或业务方针，用来描述业务交互，并建立行动和数据行为结果及完整性的规则。 **注2：** 评价数据质量时需要检查是否存在良好归档的业务规则。	计算公式：X=A/B 式中 A=满足业务规则的数据集中元素的个数； B=被评价的数据集中元素个数
5	权威参考数据	参考数据是系统、应用软件、数据库、流程、报告及交易记录和主记录用来参考的数值集合和分类表。 **注1：** 评价数据质量时需要收集参考数据列表。	计算公式：X=A/B 式中 A=满足参考数据规则的数据集中元素的个数； B=被评价的数据集中元素个数
6	安全规则	安全规则是安全和隐私方面的规则，包括数据权限管理，数据脱敏处理等。	计算公式：X=A/B 式中 A=满足安全规范的数据集中元素的个数； B=被评价的数据集中元素个数

A.3 一致性

数据与其他特定上下文中使用的数据无矛盾的程度。即一致性是指数据是否遵循了统一的规范，数据集合是否保持了统一的格式。数据质量的一致性主要体现在数据记录的规范和数据是否符合逻辑。

表 A.2 一致性评价指标

序号	指标名称	指标描述	计算方法
1	相同数据一致性	同一数据在不同位置存储或被不同应用或用户使用时，数据的一致性，数据发生变化时，存储在不同位置的数据的同一数据被同步修改。	计算公式：X=A/B 式中 A=满足一致性要求的数据集中元素的个数； B=被评价的数据集中元素个数；
2	关联数据一致性	根据一致性约束规则检查关联数据的一致性。	计算公式：X=A/B 式中 A=满足一致性要求的数据集中元素的个数； B=被评价的数据集中元素个数；

A.4 准确性

数据准确表示其所描述的真实实体（实际对象）真实值得程度。即准确性是指数据记录的信息是否存在异常或错误。

表 A.3 准确性评价指标

序号	指标名称	指标描述	计算方法
1	数据内容正确性	数据内容是否是预期数据。	计算公式：X=A/B 式中 A=满足数据正确性要求的数据集中元素的个数。 B=被评价的数据集中元素个数；
2	数据格式合规性	数据格式包含（数据类型、数据范围、数据长度、精度等）是否满足预期要求。	计算公式：X=A/B 式中 A=满足格式要求的数据集中元素的个数。 B=被评价的数据集中元素个数；
3	数据重复率	特定字段、记录、文件或数据集意外重复的度量。	计算公式：X=A/B 式中 A=重复数据集中元素的个数 B=被评价的数据集中元素个数；
4	数据唯一性	特定字段、记录、文件或数据集唯一性的度量。	计算公式：X=A/B 式中 A=满足唯一性要求的数据集中元素的个数； B=被评价的数据集中元素个数；
5	脏数据出现率	正确字段、记录、文件或数据集之外无效数据的度量。	计算公式：X=A/B 式中 A=有脏数据出现的数据集中元素的个数； B=被评价的数据集中元素个数；

A.5 唯一性

数据唯一不重复。即唯一性是指度量哪些数据是重复数据或者数据的哪些属性是重复的。

A.6 关联性

数据的关联不可缺失的。即关联性是度量哪些关联的数据缺失或者未建立索引。

关联性评价因素：

a) 查找到的信息和主题不完全一致，但确是其中某一方面的阐述；
b) 查找到的信息集合多数在用户需要的检索主题内；
c) 提供的信息主题与用户检索主题相匹配；
d) 查找到的信息多数与用户需要的信息无关；
e) 信息必须和用户需求有相关性。

A.7 时效性

数据在时间变化中的正确程度。即及时性是指数据从产生到可以查看的时间间歇，也叫做数据的延时时长，及时性对数据分析本身要求并不高，但如果数据分析周期加上数据建立的时间过长，就可能导致分析出的结论失去借鉴意义。

表 A.4 时效性评价指标

序号	指标名称	指标描述	计算方法
1	基于时间段的正确性	基于日期范围内的记录数或频率分布符合业务需求的程度。	计算公式：X=A/B 式中 A=满足有效性要求的数据集中元素的个数； B=被评价的数据集中元素个数；
2	基于时间点的及时性	基于时间戳的记录数、频率分布或延时时间符合业务需求的程度。	计算公式：X=A/B 式中 A=满足及时性要求的数据集中元素的个数 B=被评价的数据集中元素个数；
3	时序性	数据集中同一实体的数据元素之间的相对时序关系。	计算公式：X=A/B 式中 A=满足时序性要求的数据集中元素的个数； B=被评价的数据集中元素个数；

A.8 可访问性

数据能被访问的程度。

表 A.5 可访问性评价指标

序号	指标名称	指标描述	计算方法
1	可访问	数据在需要时的可获取性。	计算公式：X=A/B 式中 A=满足可访问性要求的数据集中元素的个数 B=被评价的数据集中元素个数；
2	可用性	数据在设定有效生存周期内的可使用性。	计算公式：X=A/B 式中 A=满足可用性要求的数据集中元素的个数； B=被评价的数据集中元素个数