ICS 25.040.40
CCS L 70

DB21

辽　宁　省　地　方　标　准

DB21/T 3893—2023

工业数据流通　数据清洗规范

2023-12-30 发布　　2024-01-30 实施

辽宁省市场监督管理局　发布

目　　次

前　言

本文件按照GB/T 1.1-2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由辽宁省工业和信息化厅提出并归口。

本文件起草单位：沈阳华睿博信息技术有限公司、国家计算机网络应急技术处理协调中心辽宁分中心、辽宁艾特斯智能交通技术有限公司、辽宁职业学院、东北大学、上海数据交易所、辽宁省大数据管理中心、北京赛迪时代信息产业股份有限公司、辽宁省先进装备制造业基地建设工程中心。

本文件主要起草人：邵华、李凯、黄书鹏、王宇飞、宋宪辉、王义刚、申翔宇、谭振华、杨成实、张翔宇、魏国伟、刘洋。

本文件发布实施后，任何单位和个人如有问题和意见建议，均可以通过来电和来函等方式进行反馈，我们将及时答复并认真处理，根据实际情况依法进行评估及复审。

归口管理部门通信地址：沈阳市辽宁省沈阳市皇姑区北陵大街45-2号。

归口管理部门联系电话：024-86913384。

文件起草单位通讯地址：辽宁省沈阳市和平区青年大街386号华阳国际大厦2396。

文件起草单位联系电话：18698849086。

工业数据流通　数据清洗规范

1　范围

本文件规定了工业数据清洗的过程要求、环境要求和质量要求。

本文件适用于数据流通中的工业数据清洗。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 22239-2019　信息安全技术　网络安全等级保护基本要求

GB/T 35274-2017　信息安全技术　大数据服务安全能力要求

GB/T 35295-2017　信息技术　大数据　术语

GB/T 37973-2019　信息安全技术　大数据安全管理指南

GB/T 39477-2020　信息安全技术　政务信息共享　数据安全技术要求

GB/T 42128-2022　智能制造　工业数据　分类原则

DA/T 82-2019　基于文档型非关系型数据库的档案数据存储规范

3　术语和定义

下列术语和定义适用于本文件。

3.1

数据清洗　data cleaning

运用一定方法识别并修正数据问题，提高数据质量的过程。

3.2

工业数据　industrial data

在工业领域中，涉及企业的所有生产活动和服务所产生的数据。

[来源:GB/T 42128-2022，3.1.1]

3.3

结构化数据　structured data

一种数据表示形式，按此种形式，由数据元素汇集而成的每个记录的结构都是一致的并且可以使用关系模型予以有效描述。

[来源:GB/T 35295-2017，2.2.13]

3.4

非结构化数据　unstructured data

不具有预定义模型或未以预定义方式组织的数据。

[来源:GB/T 35295-2017，2.1.25]

3.5

半结构化数据 semi-structured data

具有结构性，但结构变化大，且难以用结构化数据的处理方法将其放进二维表的数据。

示例：XML文档内容，每项都被一对标记封起来，如<title></title>，表面上看是结构化数据，但<title></title>之间的数据却是千变万化，这是典型的半结构化数据。

[来源:DA/T 82-2019，2.8]

3.6

表结构 table structure

为主体层内容提供表示语义的一种存储范例。

[来源:GB/T 35295-2017，2.2.14]

3.7

敏感数据 sensitive data

由权威机构确定的受保护的信息数据。

注:敏感信息数据的泄露、修改、破坏或丢失会对人或事产生可预知的损害。

[来源:GB/T 35295-2017，2.2.14]

4 缩略语

下列缩略语适用于本文件。

ETL：数据的抽取、转换、加载（Extract Transform Load）

5 清洗目的

工业数据涉及到各种传感器、监测设备和生产设备，具有多样性和异构性，由于传感器和设备的不稳定性、及环境变化等因素影响，工业数据中存在大量错误数据、缺失数据和异常数据。

工业数据清洗目的是清除或修正错误数据、缺失数据、异常数据或其他有问题的数据，提高工业数据在建模分析、应用开发、资源调度和监测管理等方面的应用价值，保障流通的工业数据质量。

6 清洗范围

清洗范围涵盖工业领域产品和服务全生命周期产生和应用的数据，包括但不限于工业企业在研发设计、生产制造、供应链物流、营销、运维、管理及金融等环节中生成和使用的数据，以及工业互联网平台企业在设备接入、平台运行、工业应用程序使用等过程中生成和使用的数据。各类型数据说明如下：

a） 研发设计数据：包括研发设计数据、开发测试数据等；

b） 生产制造数据：包括控制信息、工况状态、工艺参数、系统日志、生产质量数据、生产实绩数据等；

c） 供应链物流数据：包括供需计划数据、仓储物流数据等；

d） 营销数据：包括投标次数、订单数量、交易金融、客户异议数据等；

e） 运维数据：包括产品运行状况数据、产品售后服务数据等；

f） 管理数据：包括客户基本信息、业务合作数据、人事财务数据、系统设备资产信息、产品基本信息、项目进度数据、业务统计数据（如资源量数据、能耗监测数据等）；

g） 金融数据：包括信贷数据、融资租赁数据、征信数据等；

h） 平台运营数据：接入的设备数据、工业模型数据、工业应用程序数据、平台运行数据等。

7 过程要求

7.1 清洗流程

工业数据清洗流程包括数据抽取、定义规则、数据过滤、数据校验、错误标识、修正处理、数据转换、结果检验及数据加载等环节。工业数据清洗可采取ETL流程。工业数据清洗ETL流程图见图1。

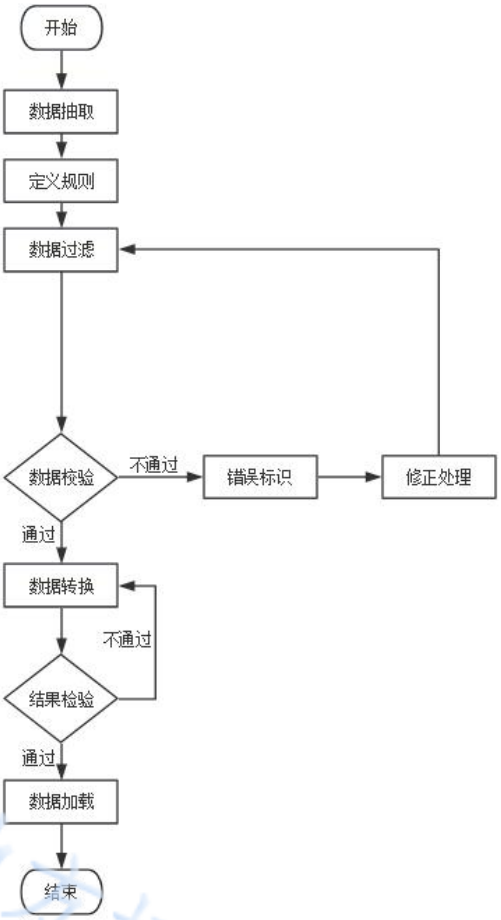

图1 工业数据清洗ETL流程图

7.2 数据抽取

数据抽取应符合以下要求：

a） 应具备全量抽取和增量抽取两种方式；

b） 数据抽取来源应能支撑抽取操作，使用生产库，或通过前置库等方式进行抽取；

c） 应支持结构化数据、半结构化数据和非结构化数据等不同类型数据的抽取；

d） 数据抽取目的地的存储容量应能支持数据抽取来源的数据总量，数据抽取目的地的表结构应与数据抽取来源的表结构保持一致；

e） 增量抽取应确定增量更新的方式，抽取的数据应有字段可区分，如更新时间等。

7.3 定义规则

应分析抽取目标数据的范围、体量、类型、内容、关系、质量等信息，全面认识数据情况。数据清洗针对的对象主要有缺失值、异常值、重复值和无用值，针对不同对象的不同形式，结合应用需求，确定数据清洗目标和规则，从而得到期望的数据。

针对不同的清洗对象，清洗规则包括：

a） 缺失值清洗：按照缺失比例和缺失字段重要性，制定清洗策略；

b） 异常值清洗：针对取值错误、格式错误和逻辑错误制定不同的清洗策略；

c） 重复值：重复数据可以去重或作出标记；

d） 无用值：无用数据字段可以直接进行删除。但在进行该过程的时候，要注意备份原始数据。

7.4 数据过滤

数据过滤应包括以下操作：

a） 将非结构化数据和半结构化数据转换为结构化数据；

b） 对噪声数据进行删除；

c） 对业务数据中不符合应用规则的数据进行删除；

d） 过滤删除掉的数据应存入问题数据库表，便于后续查证或重新使用。

7.5 数据校验

7.5.1 基本要求

应对工业数据进行非空校验、长度校验、数据量校验、数据类型和值校验。当不满足校验要求时，应进行数据错误标识；当满足校验要求时，直接进行数据转换。

7.5.2 非空校验

应在字段为非空的情况下，对该字段数据进行校验，数据不能为空值。

7.5.3 长度校验

数据长度应满足转换要求的字段长度。

7.5.4 数据量校验

过滤后的数据总量应与原始抽取的数据总量吻合。

7.5.5 数据类型和值校验

数据类型和值应能支持后续数据转换过程，如后续根据定义规则需要将时间字符串数据转换成时间类型时，还需检验时间字符串类型的数据是否符合时间格式。

7.6 错误标识

7.6.1 错误类型

工业数据的错误类型包括但不限于：

a） 残缺数据：缺一些记录，或一条记录里缺一些值（空值），或两者都缺；

b） 错误数据：数据没有严格按照规范记录，包括格式内容错误、逻辑错误、不合规等；

c） 重复数据：相同的记录出现多条或多条记录代表同一实体。

7.6.2 识别方法

可采用统计学方法、数据挖掘、基于聚类的方法、基于距离的方法、基于分类的方法、基于关联规则的方法、业务区分等方式分析数据，从而识别出数据的错误类型。

7.6.3 标识步骤

错误标识步骤如下：

a） 按 7.6.2 推荐的识别方法，分析筛选出工业数据资源中存在的数据问题；

b） 按 7.6.1 给出的错误类型，对数据问题进行分类，标识错误。

7.7 修正处理

7.7.1 残缺数据处理

7.7.1.1 处理策略

残缺数据按照字段缺失率和字段重要性，分别制定处理策略。残缺数据处理策略制定应满足以下内容：

a） 重要性高、缺失率低：通过计算进行填充；通过经验或业务知识估计；

b） 重要性高、缺失率高：尝试从其他渠道取数补全；使用其他字段通过计算获取；去除字段并在结果中标明；

c） 重要性低、缺失率低：不做处理或简单填充；

d） 重要性低、缺失率高：去除该字段。

7.7.1.2 去除字段处理

数据中如有多余字段，应备份当前数据，直接删除掉不需要的字段。

7.7.1.3 填充缺失内容处理

应采用以下方式填充缺失内容：

a） 同指标的计算结果填充：通过数据项与数据项之间的逻辑联系，采取一定的列项拆分、列计算等方式得到缺失内容；

b） 同一指标的计算结果填充：采取均值、中位数、众数等方式进行填充；

c） 重新获取：当缺失率高且非常重要的数据项，应采取重新抽取不同数据源的数据进行关联对比填充。

7.7.1.4 取数补全处理

应通过线下收集、业务知识或经验推测补全缺失值。

7.7.2 错误数据处理

7.7.2.1 格式内容问题处理

格式内容问题数据处理应采用以下方法：

a） 全、半角处理：通过正则表达式将全、半角符号按照事先定义的规则进行全、半角符号统一；

b） 有不该存在的字符：以半自动校验结合半人工方式来找出存在的问题，自动去除不需要的字符，将数据自动化统一或人工修正为正确字符；

c） 内容与字段不匹配：详细识别问题类型，如人工填写错误、前端没有校验、导入数据时部分或全部存在列没有对齐、数据源端业务系统缺陷等，不能直接删除，应按照清洗规则，采取加入更多数据源进行数据关联，找到匹配的相应字段进行填补。

7.7.2.2 逻辑问题处理

逻辑问题数据处理应采用以下方法：

a） 了解数据潜在的逻辑规则，采取逻辑推理法，直接去掉一些使用简单逻辑推理即可发现问题的数据；

b） 对于不重要的不合理数据应过滤，形成错误数据集由数源部门进行确认是否可删除；

c） 通过字段间相互验证的方法修正矛盾内容，如根据字段的数据来源，判定哪个字段提供的信息更可靠，去除或重构不可靠字段；

d） 通过分箱、聚类、回归等方法识别离群值（异常值），按照经验和业务流程判断其合理性，若合理，则保留该数值；若不合理，对重要性较高而无法重新采集的数值，按缺失数据处理，对重要性较低的数值，可直接删除；

e） 对于复杂逻辑数据问题应咨询了解该数据的产生原因，按照协商的清洗加工规则进行处理。

7.7.2.3 不合规问题处理

不合规问题数据处理应采用以下方法：

a） 设定判定规则：设定强制合规条件，对于不在规则范围内的数据，应强制设置最大值及最小值，或剔除、判断为无效字段；

b） 设定警告规则：对于不在规则范围内的数据，应进行警告及人工处理。

7.7.3 重复数据处理

重复数据处理步骤如下：

a） 通过元数据血缘关系查询到重复数据的各个来源；

b） 通过数据主键或寻找相关信息识别重复数据的含义，不是相同含义的数据不能界定为重复数据进行去重处理，应分别保留；

c） 查询到确定的重复数据，根据权威性和应用场合，选择最恰当渠道来源的数据，或在不影响数据保真度和完整性的情况下进行合并处理。

7.8 数据转换

数据转换应符合以下要求：

a） 数据转换应在数据校验通过后开始；

b） 数据转换开始前应检查需要转换的数据规则和字段是否一致；

c） 应实现对数据的格式、信息代码、值的冲突进行转换；

示例 1：统一时间日期数据格式。

1） 将各类日期统一转换为八位的字符日期，如 YYYYMMDD。

2） 将各类时间统一转换为六位的字符时间，如 HHMMSS。

3） 将各类时序数据的时间日期统一转换为十四位的字符时间日期，如 YYYYMMDDHHMMSS。

示例 2：统一分类数据取值代码。

1） 将人员性别数据统一转换为国际性别信息代码。

2） 将组织地址数据统一转换为行政区划代码。

3） 将组织名称统一转换为统一社会信用代码。

d） 转换后的数据结构应与目标数据库的结构相兼容；

e） 数据向目标移动时，将其从源数据中移除，或数据复制到多个目标中；

f） 转换失败应立即停止，开始查找问题；

g） 长时间未转换结束，需仔细核查数据量、规则和字段是否一致，如有问题应立即停止；

h） 应在解决查找到的问题后再开始数据转换。

7.9 结果检验

7.9.1 检验内容

检验内容应包括：

a） 主键重复：检验多个业务系统中同类数据经过清洗后，在统一保存时，主键的唯一性；
b） 非法代码、非法值：检查个别字段出现的异常信息，包括非法代码、代码与数据标准不一致、取值错误、格式错误、多余字符、乱码等；
c） 数据格式：检验表中属性值的格式是否正确，衡量其准确性，如时间格式、币种格式、业务部门格式、物料格式等；
d） 记录数：检验各个系统相关数据之间的数据总数或检验数据表中每日数据量的波动；
e） 业务约束：应从业务的角度检验数据的正确性、一致性、有效性等，如出（入）库日期、客户（供应商）基本信息、设备运行信息等；
f） 标准约束：对照系统数据应符合的标准进行校验。

7.9.2 结果要求

7.9.2.1 规范性

数据的质量及存储标准应统一，源数据应在源头或备份表中能找到，数据在字段、记录内容或数据集内不应有重复值。

7.9.2.2 完整性

数据集合中应包含足够的数据来响应各种查询和支持各种计算。数据完整性体现在以下方面：

a） 元数据的完整性，例如：唯一性约束完整性、参照完整性等；
b） 数据条目完整性，例如：数据记录丢失或不可用会影响数据的完整性等；
c） 数据属性完整性，例如：数据属性空值情况等。

7.9.2.3 准确性

数据所指内容对数据所指对象的反应、表现应准确，数据形式对数据内容的表述、表达应准确。

7.9.2.4 一致性

数据一致性应符合以下要求：

a） 同一个数据在同一时刻在不同数据库、应用和系统中应只有一个值；
b） 数据字段内数据应与字段描述一致；
c） 最终结果数据的统计量应与预测一致；
d） 数据项应在取值范围、单位、精度等方面保持一致。

7.9.2.5 时效性

不同类型的应用对数据的时间特性有不同的要求，数据的时间特性应满足业务应用的要求，数据记录应根据时间特性及时更新。

7.9.2.6 可访问性

数据来源稳定，数据结果应支撑后续业务。

7.9.3 检验步骤

结果检验应包括以下步骤：

a） 按 7.9.1 规定检验清洗加工后的数据资源情况；

b） 按 7.9.2 要求核对数据资源达标情况；
c） 当数据资源未达到 7.9.2 要求，应返回再次进行数据转换；
d） 当数据资源达到 7.9.2 要求，应进行数据加载或结束数据清洗。

7.10 数据加载

数据加载应满足以下要求：
a） 数据价值方式应匹配数据抽取方式，包含全量加载、增量加载，如海量数据、数据变化比较规律、变化数据相对总量较小、业务系统能直接提供增量数据时，宜使用增量加载；
b） 数据加载环境应能支撑相应数据；
c） 数据加载工具具有高效的加载性能，应能至少满足业务需求；
d） 数据加载策略应考虑数据加载周期和数据追加策略；
e） 数据加载应记录日志，并按相关规定留存日志文件；
f） 数据加载过程可根据实际操作情况，在定义规则过程前进行。

8 环境要求

8.1 数据脱敏

8.1.1 脱敏流程

应在保证敏感信息不被泄露的环境下进行工业数据清洗，工业数据脱敏工作流程包括发现敏感数据、标识敏感数据、确定脱敏方法、定义脱敏规则、执行脱敏操作和评估脱敏效果等环节。

8.1.2 发现敏感数据

基于工业数据分类分级制度，在完整的数据范围内查找并发现敏感数据，并明确敏感数据结构化或非结构化的数据表现形态，如敏感数据固定的字段格式。

在发现敏感数据过程中，应满足以下内容：
a） 定义数据脱敏工作执行的范围，应在该范围内执行敏感数据的发现工作；
b） 应通过对数据表名称、字段名称、数据记录内容、数据表备注、数据文件内容等直接匹配或正则表达式匹配发现敏感数据；
c） 宜考虑数据引用的完整性，如保证数据库的引用完整性约束；
d） 数据发现手段应支持主流的数据库系统、数据仓库系统、文件系统，同时应支持云计算环境下的主流新型存储系统；
e） 宜利用自动识别工具执行数据发现工作，并降低该过程对生产系统的影响；
f） 数据发现工具应具有扩展机制，可根据业务需要自定义敏感数据的发现逻辑；
g） 应固化常用的敏感数据发现规则，例如身份证号、手机号等敏感数据的发现规则，避免重复定义数据发现规则。

8.1.3 标识敏感数据

在发现敏感数据后，应对敏感数据进行标识，包括标识敏感数据的位置、敏感数据的格式等信息。敏感数据的标识方法应确保敏感数据标识信息能够随敏感数据一起流动，并不易于删除和篡改，从而可以对敏感数据的访问、传输和处理进行跟踪和监督，以确保敏感数据的安全合规性。

在标识敏感数据时，应满足以下内容：
a） 应尽早在数据的收集阶段就对敏感数据进行识别和标识，这样便于在数据的整个生命周期阶段对敏感数据进行有效管理；

b） 敏感数据的标识方法应考虑便捷性和安全性，使得标识后的数据很容易被识别，同时，要确保敏感数据标识信息不容易被恶意攻击者删除和篡改；

c） 敏感数据的标识方法应支持静态数据的敏感标识及动态流数据的敏感标识。

8.1.4 确定脱敏方法

可选的数据脱敏方法包括静态数据脱敏和动态数据脱敏。不同的数据脱敏方法对数据源的影响不同，脱敏的时效性也不一样。脱敏方法确定后，可选择对应的数据脱敏工具。

在确定数据脱敏方案时，应满足以下内容：

a） 静态数据脱敏方法是对原始数据进行一次脱敏，脱敏后的结果数据可以多次使用，适合使用场景比较单一的场合；

b） 动态数据脱敏方法是在敏感数据显示时，针对不同用户需求，对显示数据进行屏蔽处理的数据脱敏方式，它要求系统有安全措施确保用户不能够绕过数据脱敏层次直接接触敏感数据。动态数据脱敏适合用户需求不确定、使用场景复杂的情形。

8.1.5 定义脱敏规则

在敏感数据生命周期识别的基础上，应明确存在数据脱敏需求的业务场景，并结合行业法规的要求和业务场景的需求，制定相应业务场景下有效的数据脱敏规则。

在定义脱敏规则过程中，应满足以下内容：

a） 应遵循的个人隐私保护、数据安全保护等关键领域的国内外法规、行业监管规范或标准，以此作为数据脱敏规则必须遵循的原则；

b） 对已识别出的敏感数据执行全生命周期（产生、采集、使用、交换、销毁）流程的梳理，应明确在全生命周期各阶段，用户对数据的访问需求和当前的权限设置情况，分析整理出存在数据脱敏需求的业务场景。例如，在梳理过程中，会发现存在对敏感数据的访问需求和访问权限不匹配的情况（用户仅需获取敏感数据中部分内容即可，但却拥有对敏感数据内容全部的访权限），因此该业务场景存在敏感数据的脱敏需求；

c） 分析存在数据脱敏需求的业务场景，在“最小够用”的原则下明确待脱敏的数据内容、符合业务需求的脱敏方式，以及该业务的服务水平方面的要求，以便于脱敏规则的制定；

d） 数据脱敏工具应提供扩展机制，从而让用户可根据需求自定义脱敏的方法；

e） 通过数据脱敏工具选择数据脱敏方法时，脱敏工具中应对各类方法的使用进行详细的说明，说明应包括但不限于规则的实现原理、数据引用完整性影响、数据语义完整性影响、数据分布频率影响、约束和限制等，以支撑脱敏工具的使用者在选择脱敏方式时做出正确的选择；

f） 应固化常用的敏感数据脱敏规则，例如身份证号、手机号等的常用脱敏规则，避免数据脱敏项目实施过程中重复定义数据脱敏规则。

8.1.6 执行脱敏操作

数据脱敏操作可包括条数据脱敏和块数据脱敏。条数据脱敏是对单条数据根据脱敏规则实施脱敏，块数据脱敏是对聚合数据实施脱敏。在日常的脱敏工作中，监控分析数据脱敏过程的稳定性、以及对业务的影响性，同时对脱敏工作开展定期的安全审计，已发现脱敏工作中存在的安全风险。

在执行脱敏操作过程中，应满足以下内容：

a） 支持从数据源克隆数据到新环境（例如从生产环境、备份库克隆数据到新环境），并在新环境中进行脱敏过程的执行，也支持在数据源端直接进行脱敏；

b） 对脱敏任务的管理，宜考虑采用自动化管理的方式提升任务管理效率，例如定时、条件设置的方式触发脱敏任务的执行；

c） 执行对脱敏任务的运行监控，宜考虑任务执行的稳定性以及脱敏任务对业务的影响；

d） 设置专人定期对数据脱敏的相关日志记录进行安全审计，发布审计报告，并跟进审计中发现的例外和异常，审计应重点关注高权限账号的操作日志和脱敏工作的记录日志。

8.1.7 评估脱敏效果

通过收集、整理数据脱敏工作执行的数据，例如相关监控数据、审计数据，对数据脱敏的前期工作开展情况进行反馈，从而优化相关规程，明确数据脱敏过程中应满足的内容。

在评估脱敏效果过程中，应满足以下内容：

a） 利用测试工具评估脱敏后数据对应用系统的功能、性能影响，从而明确对整体业务服务水平的影响，测试负载宜尽量保证与生产环境一致，宜尽量提供从生产环境克隆数据访问负载到脱敏系统进行回放测试的功能；

b） 应根据组织业务发展的情况和脱敏工作执行的反馈，优化数据脱敏工作开展的规程。

8.2 数据安全

应在与互联网隔绝的安全环境下清洗工业数据，环境应支持数据可存储、可转化，工业数据清洗应符合GB/T 22239-2019、GB/T 35274-2017和GB/T 37973-2019的相关要求，确保工业数据的保密性和完整性。

8.3 人员能力

工业数据清洗人员应经过相应的技术和安全培训，具有数据清洗的能力，取得相关业务领域的数据管理认证资格，并能按照数据安全管理相关制度完成工业数据清洗工作。

9 质量要求

清洗后的工业数据应符合数据流通的质量管理要求，工业数据质量特性包括：规范性、完整性、准确性、一致性、时效性及可访问性。各质量特性的说明如下：

a） 规范性：数据符合数据标准、数据模型、业务规则、元数据、权威参考数据及安全规范的程度。

1） 数据标准是数据的命名、定义、结构和取值规范方面的规则和基准；

2） 数据模型是对分析的图像和文本表述，该分析识别了组织为完成其使命、功能、目标、目的和战略，以及管理和评价组织所需要的数据；

3） 业务规则是一种权威性原则或指导方针，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；

4） 元数据是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、访问权和数据易变性的数据；

5） 权威参考数据是系统、应用软件、数据库、流程、报告或平台日志记录用来参考的特定字段的有效数据集合；

6） 安全规范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。

b） 完整性：按照数据规则要求，数据元素被赋予数值的程度。即数据信息是否存在缺失的状况，包括数据元素完整性和数据记录完整性；

c） 准确性：数据准确表示其所描述的真实实体(实体对象)真实值的程度，即数据记录的信息是否存在异常或错误，包括数据内容正确性、数据格式合规性、数据重复率、数据唯一性、脏数据出现率；

d） 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，主要体现在数据记录的规范和数据是否符合逻辑，包括相同数据一致性和关联数据一致性；

e） 时效性：数据在时间变化中的正确程度，包括基于时间段的正确性、基于时间点及时性、时序性；

f） 可访问性：数据能被访问的程度，包括可访问和可用性。

各质量特性的评价指标可参考GB/T 36344-2018第5章。

注：由于工业产品生命周期同一阶段的数据具有强关联性，如产品零部件组成、工况、设备状态、维修情况、零部件补充采购等，工业产品生命周期的研发设计、生产、服务等不同环节的数据之间也需要进行关联；工业企业生产、经营流程具有较强时序性。因此，工业数据质量特性评价宜侧重于一致性和时效性。

参 考 文 献

[1] GB/T 36344-2018 信息技术 数据质量评价指标
[2] GB/T 39400-2020 工业数据质量 通用技术规范
[3] DB14/T 2526-2022 工业互联网综合平台 数据质量管理要求
[4] DB52/T 1126-2016 政府数据 数据脱敏工作指南
[5] DB52/T 1540.3-2020 政务数据 政务数据 第3部分：数据清洗加工规范
[6] 《工业数据分类分级指南（试行）》
[7] 《工业和信息化领域数据安全管理办法（试行）》