

工业互联网综合平台 数据质量管理要求

2022-08-18 发布

2022-11-18 实施

目 次

前言 II

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 缩略语 1

5 数据分类 2

6 数据质量特性 2

7 数据质量管理 3

 7.1 数据质量管理框架 3

 7.2 数据质量管理流程 3

 7.3 数据质量管理工具 4

参考文献 7

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。

本文件由山西省工业和信息化厅提出、组织实施和监督检查。

山西省市场监督管理局对标准的组织实施情况进行监督检查。

本文件由山西省物联网和人工智能标准化技术委员会归口。

本文件起草单位：山西大数据产业发展有限公司、华为技术有限公司、浪潮工业互联网股份有限公司、新华三技术有限公司、精英数智科技股份有限公司、山西中科同昌智能科技有限公司。

本文件主要起草人：阳军、王文逾、吉凤鸣、刘耀宏、刘泱、张国华、张尚礼、高巍、王盾、刘品杰、刘赞、吕洪、龚大立、袁铎、赵存会、王延辉、王学斌、吴喆峰、朱晓宁、杜石磊、杨少封、林薇、彭祖炜、伍云山、李志国、郭晓军、王俊、李鹏飞、杨向东、张志芳、余小占、韩丽娟、路旭、张晶亮、李炳增、李峰、张汉超、顾朝阳。

工业互联网综合平台 数据质量管理要求

1 范围

本文件规定了工业互联网综合平台数据的分类、质量特性以及质量管理的框架、流程和工具。
本文件适用于工业互联网综合平台数据质量的管理、检测、分析和提升。

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 36344—2018 信息技术 数据质量评价指标
GB/T 39400—2020 工业数据质量 通用技术规范

3 术语和定义

GB/T 39400—2020界定的以及下列术语和定义适用于本文件。

3.1

工业互联网综合平台 industrial internet integrated platform

以工业互联网平台为基础，支持汇聚数据、服务、用户等各类资源，具备数据集成分析、应用支撑能力和基础应用能力，支持省域内特色型、专业型、企业级等平台的集成和接入，支撑制造资源泛在连接、弹性供给、高效配置和协同创新的载体。

3.2

数据质量 data quality

数据的一组固有特性满足要求的程度。

注：固有特性一般指永久性的特性。

[来源：GB/T 39400—2020, 3.1]

3.3

数据质量管理 data quality management

指导和控制某机构数据质量的协调活动。

[来源：GB/T 39400—2020, 3.2]

3.4

数据质量管理工具 data quality management tool

运用信息化技术和管理方法来规划、实施与控制数据质量的识别、评估、提升等一系列活动。

4 缩略语

下列缩略语适用于本文件。

App：应用程序（Application）

SQL: 结构化查询语言 (Structured Query Language)

5 数据分类

工业互联网综合平台数据包括但不限于煤炭、制造、物流运输、电力、燃气、建筑、水务、现代服务、化工、冶金等各个行业领域产品和服务全生命周期产生和应用的数据，可分为以下类型：

- a) 研发数据：包括研发设计数据、开发测试数据等；
- b) 生产数据：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、交易金额、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) 金融服务数据：包括信贷服务数据、融资租赁服务数据、征信服务数据等；
- h) 平台运营数据：接入的设备数据、工业模型库数据、工业 App 数据、平台运行数据等；
- i) 外部数据：与其他主体共享的数据。

6 数据质量特性

根据GB/T 36344—2018中第3章和第5章，工业互联网综合平台数据质量特性包括：

- a) 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；
 - 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。
- 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) 关联性：度量相关联的数据是否缺失或建立索引；
- g) 可访问性：数据能被访问的程度，包括可访问和可用性。

7 数据质量管理

7.1 数据质量管理框架

数据质量管理框架主要由数据质量管理流程和数据质量管理工具构成，以数据质量管理流程为指导，构建数据质量管理工具，见图1。数据质量管理流程主要从数据质量描述、数据质量识别、数据质量评估和数据质量控制四方面进行搭建；数据质量管理工具主要从数据标准管理、数据质量规则、质量检测任务、数据质量监控、数据质量报警、质量检测报告、检测报告分析、数据质量整改、数据质量跟踪方面进行规定。

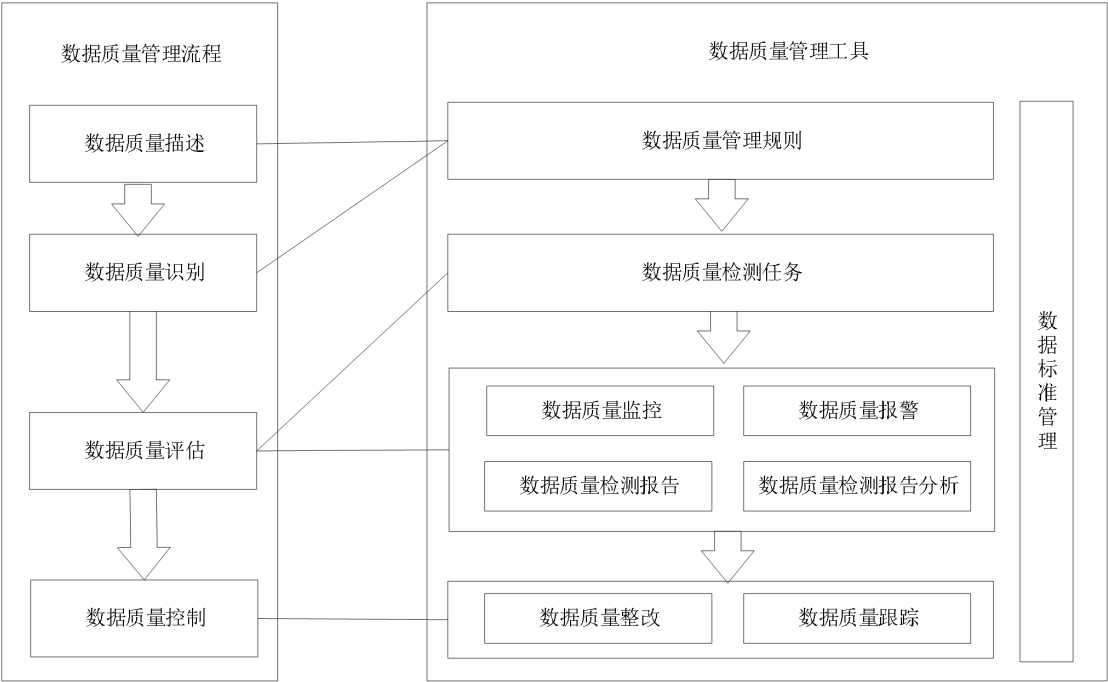


图1 数据质量管理框架

7.2 数据质量管理流程

7.2.1 数据质量管理模型

应符合GB/T 39400—2020中4.2的要求。

7.2.2 数据质量描述

数据质量描述要求包括但不限于：

- a) 应根据平台企业管理的需求，参考外部监管的要求，确定平台企业数据质量管理目标；
- b) 应根据平台企业数据质量管理目标、数据标准规范和数据质量特性，确定平台企业数据质量评估维度。

7.2.3 数据质量识别

数据质量识别要求包括但不限于：

- a) 应根据平台企业业务发展的需求及常见数据问题，确定平台数据质量管理范围、各类数据的优先级以及质量需求；

- b) 应根据平台企业数据质量管理目标及需求,定义各类数据的质量评价指标、校验规则与方法;
- c) 应根据业务发展需求及数据质量检查分析结果对数据质量规则进行持续维护与升级。

7.2.4 数据质量评估

7.2.4.1 数据质量检查

应根据数据质量规则中的有关技术指标和业务指标、校验规则与方法,对平台企业的数据质量情况进行实时监控,若发现数据质量问题应及时向数据管理人员进行反馈。

7.2.4.2 数据质量分析

数据质量分析要求包括但不限于:

- a) 应对数据质量检查过程中发现的数据质量问题及相关信息进行分析,并找出影响数据质量的原因;
- b) 应定义数据质量问题的优先级。

7.2.5 数据质量控制

数据质量控制要求包括但不限于:

- a) 应根据数据质量分析的结果制定、实施数据质量改进方案,包括错误数据更正、业务流程优化、应用系统问题修复等;
- b) 应制定数据质量问题预防方案。

7.3 数据质量管理工具

7.3.1 数据标准管理

数据标准管理要求包括但不限于:

- a) 应支持根据国家标准、行业标准、山西省地方标准等进行数据标准制定,包括数据标准的编制、审核、发布等;
- b) 应支持数据标准执行,包括将标准要求提供给数据质量管理、根据已录入的数据标准形成质量检测规则等;
- c) 应支持数据标准维护,包括标准内容变更和版本管理等;
- d) 应对数据标准的执行情况、实施效果进行监控,及时发现问题;
- e) 应根据执行情况、实施效果和存在的问题,修订或重新制定已不适用的数据标准。

7.3.2 数据质量管理规则

数据质量管理规则要求包括但不限于:

- a) 应支持根据数据质量特性自定义质量维度;
- b) 应支持关联数据标准规范,进行基础规则定义。包括但不限于空值校验、值域校验、格式校验、长度校验、精度校验、唯一性约束校验、主外键校验等;
- c) 应支持根据基础规则和平台应用场景自定义业务规则;
- d) 应通过自定义 SQL 语句等方式完成对数据质量的检核。

7.3.3 数据质量检测任务

数据质量检测任务要求包括但不限于:

- a) 应支持质量检测任务的创建;

- b) 应支持检测任务调度规则的设定；
- c) 应支持手动执行与自动执行检测任务，支持手动、定时、轮询、每周、自定义等多种调度规则；
- d) 应支持质量检测任务运维，查看任务执行日志、任务调度次数等；
- e) 应支持问题数据的查询与导出；
- f) 应支持未达标数据的查询与导出。

注1：问题数据是指平台设置数据库表中合格率范围之外的数据。

注2：未达标数据是未达到平台设置的数据质量特性合格率的表中所有数据。

7.3.4 数据质量监控

数据质量监控要求包括但不限于：

- a) 应支持对数据进行监控，明确数据的来源和去向；
- b) 应支持对规则校验的结果进行监控和分析，校验结果应定位到原始数据项。

7.3.5 数据质量报警

应支持问题数据的报警。

7.3.6 数据质量检测报告

数据质量检测报告要求包括但不限于：

- a) 应支持对校验结果进行记录并生成检测报告，检测报告的内容包括但不限于：
 - 1) 历史检测记录的数据波动，以及本次质量检测执行时间、检测结果等基本信息；
 - 2) 对应数据的动态质量变化情况；
 - 3) 每项业务规则检测数据量、问题数据量、合格率等详细信息；
- b) 应支持在线查看每项业务规则问题数据的详情；
- c) 应支持打包下载对应的问题数据；
- d) 应支持数据质量检测报告的查看和下载。

7.3.7 数据质量检测报告分析

数据质量检测报告分析要求包括但不限于：

- a) 应具备对检测报告进行综合性分析的功能，包括但不限于以下内容：
 - 1) 应支持通过多维度的分析报告来展示质量检测总体结果，包含检测数据来源、检测时间、综合得分、检测数据量、合格数据量、规则数目、检测资源数目等基本信息，和各个质量维度下每项的得分，以及此次检测数据的短板问题所属维度；
 - 2) 应支持展示每项业务规则的详细检测内容，如每项的检测数据总数、问题数据数目，合格率、问题数据分布情况等；
 - 3) 应支持通过评分报告详情展示每项规则对应的得分情况及对应的比分权重，体现不同规则的重要程度，并对高权重质量规则进行着重分析处理；
 - 4) 应支持每个业务规则横向对比，提供宏观的检测报告分析，展示数据质量问题的分布。
- b) 应具备对检测报告进行统计分析的功能，包括但不限于以下内容：
 - 1) 应支持通过图表的形式展示数据质量概览信息；
 - 2) 应支持按照数据源目录层级，即分层分域原则，按照指定的业务层级如区域、部门、指定表等多种维度进行统计分析，并生成对应的质量趋势图。

7.3.8 数据质量整改

数据质量整改要求包括但不限于：

- a) 应支持根据质量分析结果生成对问题数据的质量提升建议；
- b) 应支持数据标准化、数据清洗、数据转换和数据整合等功能，对不符合质量要求的数据进行处理, 并纠正数据质量问题。

7.3.9 数据质量跟踪

数据质量跟踪要求包括但不限于：

- a) 应支持数据质量跟踪，记录数据质量事件的评估、初步诊断和后续行动等信息，验证数据质量提升的有效性；
- b) 应支持对系统问题进行修正、对数据标准进行完善。

参 考 文 献

- [1] GB/T 36073—2018 数据管理能力成熟度评估模型
 - [2] 工业互联网创新发展行动计划（2021-2023年）
 - [3] 山西省工业互联网创新发展行动计划（2021-2023年）
 - [4] 工业互联网标准体系（版本 3.0）
 - [5] 工业互联网综合标准化体系建设指南（2021版）
-