ICS XXXXXXX
CCS X XX

DB65

新 疆 维 吾 尔 自 治 区 地 方 标 准

DB65/T XXXX-2024

一体化数据资源体系数据汇聚技术规范

（征求意见稿）

2024-XX-XX 发布　　2024-XX-XX 实施

新疆维吾尔自治区市场监督管理局　　发 布

目　次

前言 III
1　范围 1
2　规范性引用文件 1
3　术语与定义 1
4　总体要求 1
5　网络要求 2
6　汇聚流程 2
6.1　数据库类型 2
6.2　文件类型 2
6.3　服务类型 3
7　数据清洗加工 3
7.1　数据清洗加工目的 3
7.2　数据清洗加工策略 4
7.3　数据清洗加工流程 4
8　数据更新 5
9　数据汇聚安全 5
参考文献 6

前　言

本标准按照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》的规则编制。

本标准由新疆维吾尔自治区数字化发展局提出并归口。

本标准主要起草单位：新疆维吾尔自治区数字化发展局。

本标准主要起草人：

一体化数据资源体系数据汇聚技术规范

1 范围

本文件规定了一体化数据资源体系数据汇聚技术规范的总体要求、网络要求、汇聚流程、数据清洗加工、数据更新、数据汇聚安全的要求。

本文适用于指导一体化数据资源体系管理中数据汇聚的规划、实施、运行管理。

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 35273-2020 信息安全技术 个人信息安全规范

GB/T 37973-2019 信息安全技术 大数据安全管理指南

GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求

GB/T 35295-2017 信息技术 大数据 术语

3 术语与定义

GB/T 35295-2017 界定的以及下列术语和定义适用于本文件。

3.1

数据汇聚 data aggregation

将不同数据源的数据按照一定的规则与更新频率，进行采集、清洗加工和整合的行为和过程。

3.2

数据清洗加工 data cleaning and processing

运用一定方法修正识别到的数据问题，提高数据质量的过程。

3.3

全量更新 full update

使用新的数据对历史数据进行完全覆盖。

3.4

增量更新 incremental update

将两次更新间隔发生变更的数据同步到存储区域。

4 总体要求

数据汇聚工作应符合以下总体要求：

安全性：加强数据汇聚过程安全防护保障措施，包括数据的加密、权限控制、防止数据泄露等，同时，也需要对汇聚后的数据进行安全防护，确保汇聚数据不会受到未经授权的访问、篡改或泄露。

合规性：应按照法定职责和采集标准，以符合国家法律、法规的途径汇聚数据，不应汇聚、披露法律法规禁止的数据。

完整性：数据汇聚过程应明确数据归集的范围、内容及目标，确保汇聚数据的准确和完整，不应造成数据的缺失和遗漏。

时效性：应建立明确的数据更新机制，设置合理的数据更新频率和方式，确保数据时效性的同时，能够满足使用需求。

兼容性：数据汇聚应支持主流数据库与国产数据库，综合考虑国产数据库的性能、稳定性、可扩展性，制定合理的汇聚方案和办法。

5 网络要求

来源于不同网络环境的原始数据应根据以下要求进行归集：

a） 原始数据位于政务外网环境的，对数据直接归集；

b） 原始数据位于专网环境的，应建立与政务外网之间的安全传输通道，采取必要的安全措施保障数据传输安全性，对专网环境数据的归集；

c） 原始数据位于互联网环境的，应建立与政务外网之间的安全传输通道，采取必要的安全措施保障数据传输安全性，对互联网环境数据的归集。

6 汇聚流程

结合汇聚数据的范围、数据传输要求等，确定数据的汇聚方式。汇聚方式主要包括库表、文件、服务接口三种，其中库表、文件方式适用于对数据传输速度和实时性无特殊要求的情况，服务接口方式适用于对数据传输速度和实时性有较高要求的情况。

6.1 数据库类型

6.1.1 数据库类型方式汇聚流程

a） 资源提供方在保证业务数据传输安全的前提下，访问自治区一体化数据资源服务平台提供的部门前置库，建立对应的数据库表结构，并将数据资源推送至部门前置库；

b） 自治区一体化数据资源服务平台将部门前置库的数据汇聚至中心前置库中并完成数据清洗加工。

6.1.2 数据库类型规范要求

资源提供方在一体化资源资源服务平台前置库创建数据库表应遵循以下规范要求：

a） 数据库表命名：数据库表命名规则应为“T_”+“数据资源编码”，数据资源编码应符合自治区《一体化数据资源体系数据资源目录体系规范》中的数据资源编码规则要求；

b） 数据库表和字段注释：资源提供方应对数据库表及其业务字段添加注释信息，如记录 ID、批次号、业务操作标识(如更新、修改、删除等)、更新时间等，便于理解业务含义；

c） 设置主键：资源提供方应对数据库表设置主键，业务字段中存在作为此表主键的字段，应设置此字段作为主键；业务字段中不存在作为此表主键的字段，应使用 GUID 字段作为主键，使用时需要在 GUID 字段中上传能代表数据唯一标识的主键值。

6.2 文件类型

6.2.1 文件类型方式汇聚流程

a） 资源提供方将文件资源上传至部门前置机指定目录；

b） 自治区一体化数据资源服务平台工具通过连接部门前置机指定目录自动下载新增文件，并把数据采集至中心库中，同时将具有固定结构的文件解析入库。

6.2.2 文件类型规范要求

资源提供方应遵循以下文件类型规范要求：

a） 文件存储路径：根据前置节点服务器操作系统种类，应明确交换文件的存放路径；

b） 文件命名规范：文件命名规则应为“数据资源编码”+“文件生成时间”+“序号码”+“.”+“文件类型后缀”，其中“文件生成时间”格式应为“yyyyMMdd”，“序号码”长度为三位，范围为 001-999。

6.3 服务类型

6.3.1 服务类型方式数据汇聚流程

a） 资源提供方提供服务接口，实现通过接口调取业务应用数据库中的数据；

b） 自治区一体化数据资源服务平台工具通过资源提供方提供的服务地址、用户名密码、传入参数进行服务调用获取数据，并把数据采集至前置库中；

c） 一体化数据资源服务平台将前置库中的数据采集至中心前置库中。

6.3.2 服务类型规范要求

资源提供方在发布服务接口时应遵循如下要求：

a） 接口协议：资源提供方应满足 HTTP Open API、Restful API、Web Service 三种接口协议中的一种；

b） 字符编码：应遵循 UTF-8 编码规则；

c） 接口服务接入信息：接口接入信息应包含表 1 所列信息。

表 1 接口服务接入信息

参数	参数说明
接口名称	接口名称
功能说明	接口功能描述
URL 样式	调用接口的 URL 样式
接入接口协议	支持的协议，如：HTTP Open API、RESTful API、Web Service 等
请求方式	支持 POST、GET、POST/GET 等方式
接口参数	参数名称、参数说明、参数数据类型、参数传值要求等
提交数据举例	对于特定格式（json、xml 格式等）参数应举例说明
返回 HTTP 状态	返回 HTTP 状态所对应的状态说明
返回数据参数	返回参数的名称、参数说明等
返回数据格式	字符串/json/xml 或其他格式应举例说明
返回数据举例	对特定格式（json、xml、自定义格式等）的返回参数应提供示例说明
错误代码	对错误信息进行定义，并进行详细说明

7 数据清洗加工

7.1 数据清洗加工目的

对汇聚数据按要求进行数据清洗加工，提升汇聚数据的数据质量。常见的需数据清洗加工的内容包括：

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：出现多条相同的记录或多条记录代表同一实体。

7.2 数据清洗加工策略

7.2.1 残缺数据处理

按照字段缺失比例和字段重要性，制定不同的数据处理策略：

a） 对重要性高、缺失率高的残缺数据：尝试通过其他渠道取数据补全；使用其他字段通过计算获取；去除该字段，并在结果中标明。

b） 对重要性低、缺失率高的残缺数据：去除该字段。

c） 对重要性高、缺失率低的残缺数据：尝试通过其他渠道取数据补全；使用其他字段通过计算获取；由业务专家根据实际情况将残缺数据主观经验值和估计值。

d） 对重要性低、缺失率低的残缺数据：不做处理进行简单填充。

7.2.1 错误数据处理

根据数据内容格式、事实业务逻辑、合规性等方面考虑，对错误数据进行删除或修正处理。

7.2.1.1 删除处理

采取逻辑推理法，了解数据的事实业务逻辑，对于不符合事实业务逻辑的数据，可直接删除字段内容，作为缺失值处理。

可设置合规性要求，对于不满足合规性要求的数据，可直接删除字段内容，作为缺失值处理。

7.2.1.1 修正处理

对于数据格式异常，或数据格式不统一的问题数据，应按照数据清洗规则，通过自动化或人工处理的方式将数据进行修正处理。

对于数据内容中存在多余无效违规字符的数据，可通过自动化或人工处理的方式将数据进行违规字符删除的修正处理。

7.2.2 重复数据处理

重复问题处理步骤如下：

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。

7.3 数据清洗加工流程

数据清洗加工流程如图 1 所示：

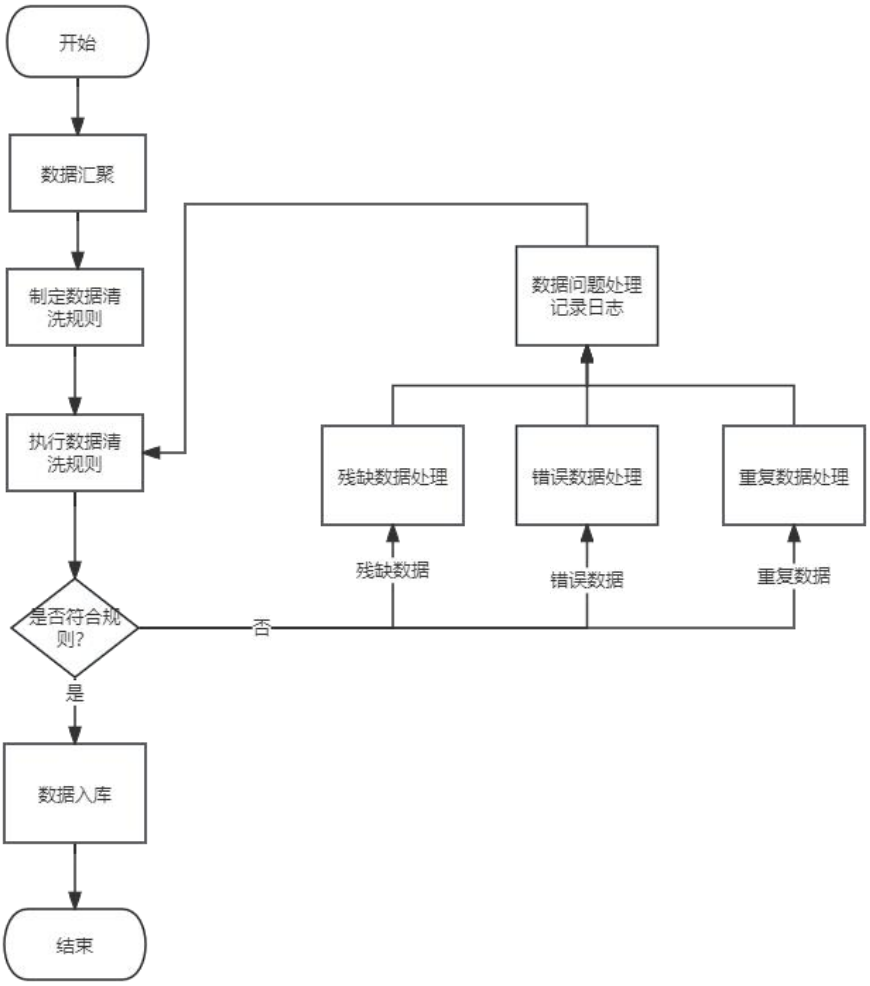

图 1 数据清洗加工流程

8 数据更新

8.1 数据更新方法

对存在更新标识的数据应支持增量更新；对不存在更新标识的数据应支持全量更新。

8.2 数据更新策略

确定数据更新的策略和周期，实时性和频率需根据实际需求确定。根据公共数据平台的使用情况和业务需求，制定数据更新的频率，可以是实时更新、每日更新、每周更新等。此外，也要考虑相关数据源的更新频率和数据变动性，建立汇聚前后的对账机制，用于比对数据的更新情况，确保及时获取最新数据。

9 数据汇聚安全

数据归集安全应符合 GB/T 22239-2019 中等级保护三级的要求，个人信息安全应符合 GB/T 35273-2020 要求，其他安全要求应符合 GB/T 37973-2019。

参 考 文 献

[1] 《国务院关于印发政务信息资源共享管理暂行办法》（国发〔2016〕51号）
[2] 《新疆维吾尔自治区公共数据管理办法（试行）》
[3] 《新疆维吾尔自治区标准化条例》