ICS 35.020
CCS L80

DB3704

枣　庄　市　地　方　标　准

DB3704/T 0041-2024

一体化大数据平台数据汇聚治理规范

2024－ 03－27 发布　　2024－04 －27 实施

枣庄市市场监督管理局　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。

本文件由枣庄市大数据局提出、归口并组织实施。

本文件起草单位：枣庄市大数据中心、浪潮云信息技术股份公司。

本文件主要起草人：赵瑞欣、王辉、陈夫真、陈亚楠、郗金鑫、王振、王俐、王延朔、刘荣栋、马俊华、张旋。

本文件为首次发布。

一体化大数据平台数据汇聚治理规范

1 范围

本文件规定了一体化大数据平台数据汇聚和治理的总体架构、总体要求、汇聚治理流程、数据汇聚要求、数据治理要求和数据安全保护要求。

本文件适用政务数据、公共数据以及社会数据通过枣庄市一体化大数据平台进行数据汇聚治理的实施和管理。

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 18391.1-2009 信息技术 元数据注册系统(MDR) 第1部分：框架

GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求

GB/T 34960.5-2018 信息技术服务 治理 第5部分：数据治理规范

GB/T 35273-2020 信息安全技术 个人信息安全规范

GB/T 35295-2017 信息技术 大数据 术语

GB/T 36344 信息技术 数据质量评价指标

GB/T 38664.1-2020 信息技术 大数据 政务数据开放共享 第1部分：总则

GB/T 38664.2-2020 信息技术 大数据 政务数据开放共享 第2部分：基本要求

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

DB37/T 4646.1-2023 公共数据 数据治理规范 第1部分：数据归集

3 术语和定义

GB/T 35295-2017 界定的以及下列术语和定义适用于本文件。

3.1

政务数据

各级政务部门及其技术支撑单位在履行职责过程中依法采集、生成、存储、管理的各类数据资源。

注：根据可传播范围,政务数据一般包括可共享政务数据、可开放公共数据及不宜开放共享政务数据。

[来源：GB/T 38664.1-2020，3.1]

3.2

数据汇聚

大数据业务主管部门根据数据管理和共享服务需求采集各类数据资源的活动。

3.3

数据治理

数据资源及其应用过程中相关管控活动、绩效和风险管理的集合。

[来源：GB/T 34960.5-2018，3.1]

3.4

数据管理

数据资源获取、控制、价值提升等活动的集合。

[来源：GB/T 34960.5-2018，3.2]

3.5

元数据

定义和描述其他数据的数据。

[来源：GB/T 18391.1-2009，3.2.16]

3.6

数据生命周期

数据获取、存储、治理、整合、分析、应用、归档和销毁等各种生存形态变化的过程。

3.7

数据提供方

在数据资源汇聚、治理、应用过程中，提供数据资源的数据权属单位。

3.8

数据需求方

在数据资源共享开放和应用过程中，提出使用需求或者申请使用数据的单位。

4 总体要求

本标准针对一体化大数据平台数据汇聚治理提出要求，总体上应满足以下要求：

a） 数据汇聚治理安全应符合 GB/T 38664.2-2020 的要求。

b） 应根据数据不同的业务更新周期建立高速及时的汇聚通道，确保数据的及时性。

c） 应记录并保留汇聚治理过程中历史数据的变化和移动情况，确保数据生命周期的可追溯性。

d） 数据汇聚治理过程中不应造成数据的缺失和遗漏，确保数据的完整性。

e） 应如实准确的处理数据，不应虚构或篡改数据，应准确记录数据不应存在异常或错误数据，确保数据的准确性。

f） 应依据国家、行业或地方数据标准对数据进行治理，确保数据治理的规范性。

5 总体架构

枣庄市数据汇聚治理的总体架构分为三层，分为数源层、市级平台层和省级枢纽层，数源层由各区县节点和市直各部门提供政务数据资源，同时接入公共数据和社会数据，市级平台依托枣庄市一体化大数据平台实现各类数据资源的汇聚、通过数据治理实现数据清洗和数据质量检测形成数据资源库，省级枢纽为上级平台，市级平台按照省级要求实现数据业务的互联互通。（总体框架见图1）

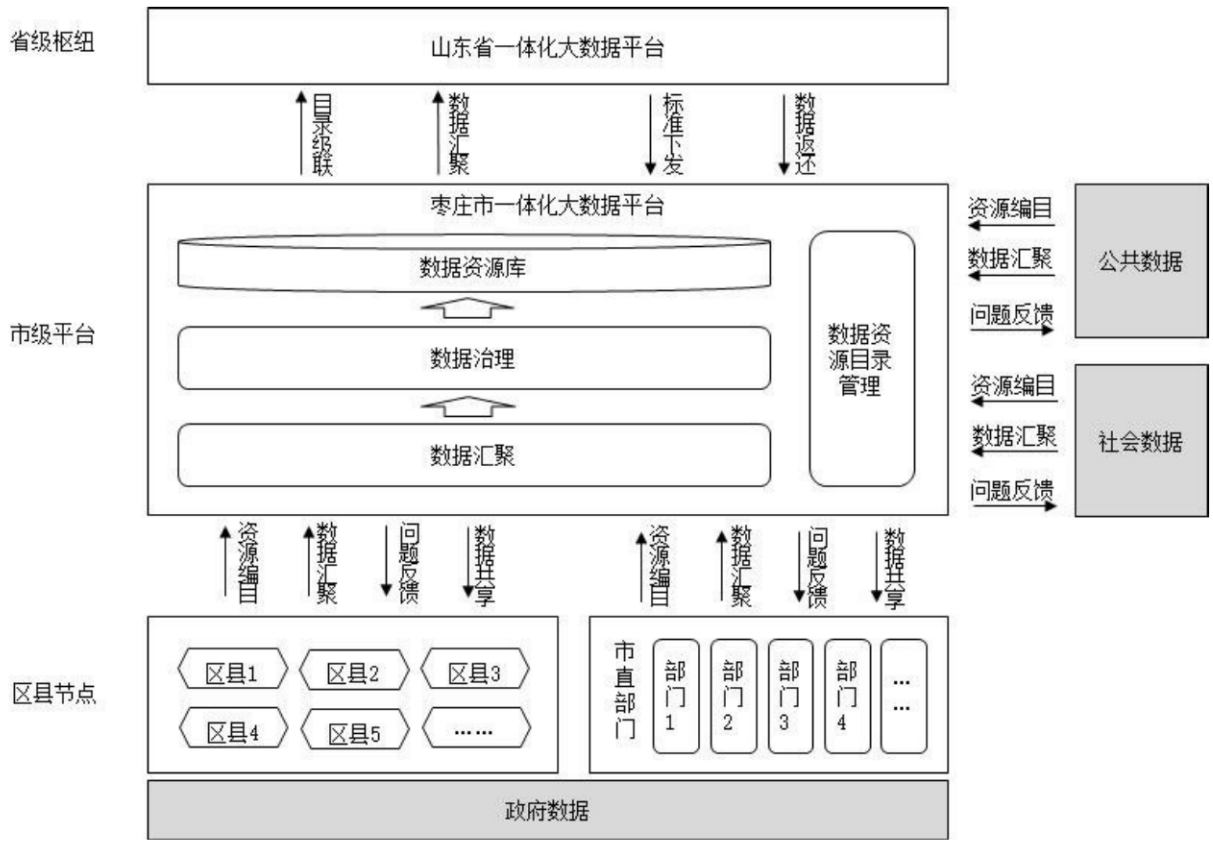

图 1　数据汇聚治理总体构图

6　汇聚治理流程

数据汇聚治理总体流程见图2，流程包含以下内容：

a)　将原始数据进行数据汇聚并存放在一体化大数据平台原始库中；

b)　对原始库数据进行数据治理包括数据清洗、数据质量检测等，使其统一标准规范并且能够满足数据共享应用的质量要求，治理后的数据存放在一体化大数据平台标准库中，对于数据治理过程中发现的问题数据存放在一体化大数据平台问题库中，其中明确数据来源单位的问题数据反馈回数据源端，无法反馈的问题数据根据业务要求进行存储或者销毁；

c)　基于数据资源库建设和数据创新应用需求，从业务维度对标准数据进行数据融合，融合后的数据存放在一体化大数据平台主题库中；

d)　当原始数据发生更新时，应依照数据汇聚、数据治理、数据融合的步骤进行数据处理；

e)　数据流转各个环节能够对数据进行溯源。

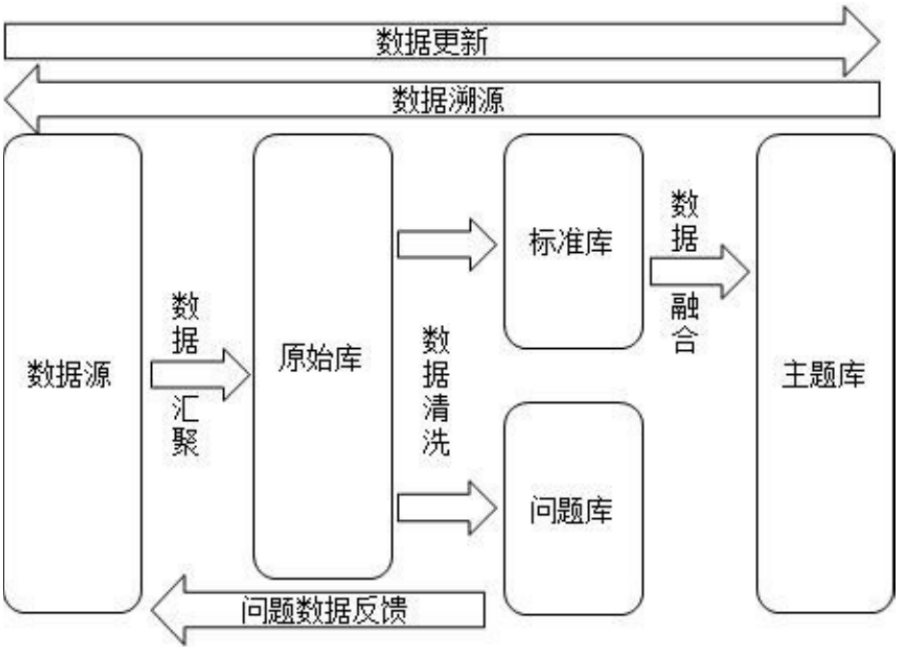

图 2　数据汇聚治理流程

7　数据汇聚要求

7.1　数据汇聚范围

大数据业务主管部门应基于数据资源管理和共享开放的需求，按照“按需归集，应归尽归”的原则将各类数据资源汇聚至市一体化大数据平台。

a)　枣庄市一体化大数据平台数据汇聚范围应包括政务数据、公共数据以及社会数据。

b)　公共数据汇聚应符合 DB37/T 4646.1-2023 的要求。

7.2　汇聚数据类型

a)　采集数据包括结构化数据、半结构化数据、非结构化数据；

b)　接入形式包括数据库表、文件、服务接口。

7.2.1　数据库表汇聚

a)　为保证数据汇聚的准确和及时，应优先采用库表方式进行汇聚。

b)　应采用前置库方式进行数据汇聚，前置库应支持国产数据库。

c)　前置数据库字符编码应为 UTF-8。

d)　数据提供方提供的业务数据表除业务字段外应包含记录 ID、批次号、业务操作标识、更新时间字段等扩展信息,业务数据表的示例见 A1.1。

e)　数据库表名称应规范统一，一般为“机构简称首字母缩写+数据资源名称首字母缩写”，数据资源名称应和数据资源目录的数据资源名称保持一致。

f)　数据库表字段应与数据目录的数据项保持一致，且必须设有主键字段，并在数据表库中创建主键约束。

g） 数据库表结构应保持稳定，不应随意更改。

h） 数据提供方应提供必要的字段说明文档和对应的全部代码表，确保所有数据内容可被正确理解。

i） 数据提供方应提供数据对账表，包含数据条数等信息，一体化大数据平台数据汇聚通过数据对账保证数据无缺失和遗漏。

7.2.2 文件汇聚

a） 结构化文件格式包括 CSV、TXT、XLS、XLSX 等，非结构化文件格式包括 PDF、DOC、DOCX、WPS、ZIP 以及图片、音频、视频等资源的常用格式。CSV、XLS、XLSX 的首行数据应为列名，字符编码应为 UTF-8 且不应存在特殊字符。

b） 文件传输方式可采用 FTP、SFTP 等。

c） 应设置文件大小阈值，文件超出阈值应拆分为多个小文件。

d） 文件名称应规范统一，应与数据目录的数据资源名称保持一致。

e） 文件的存储路径应规范统一，可根据更新的频度和检索效率建立子文件夹，不应随意更改路径。

f） 特殊类文件应提供必要的说明文档，确保所有文件内容可被正确理解。

g） 结构化文件中除业务字段外,应包含记录 ID、批次号、业务操作标识和更新时间等字段。

h） 数据提供方应提供文件对账表，明确所汇聚电子文件包含的内容和数量等信息。

7.2.3 服务接口汇聚

a） 服务接口资源采用 Schema 架构说明的标准 XML 文件方式进行描述,其中编码方式为 UTF-8，服务接口示例见 A.1.1。

b） 服务接口主要分为增删改类数据服务接口和只增类数据服务接口。

c） 通过服务接口汇聚，数据提供方应提供详细的数据接口服务说明文档。

d） 一个数据接口服务一般应且只对应一项数据资源。

e） 服务应是无状态的，两次请求之间无须状态和会话的保持。

f） 服务地址和参数不应随意变更。

7.3 数据对账要求

a） 数据资源提供方应对各类数据资源汇聚时提供对账表，明确所汇聚资源的数量和内容等信息，数据对账表示例见 A1.1.3.

b） 一体化大数据平台汇聚任务完成后应根据对账表对已汇聚数据进行核对，确保数据提供方提供的数据与已汇聚的数据保持一致。

c） 数据对账出现异常，应及时进行数据汇聚任务核查并进行纠正，保证数据无重复无遗漏。

7.4 数据更新要求

7.4.1 更新方法

a） 对存在更新标识的数据应支持增量更新。

b） 对不存在更新标识的数据应支持全量更新。

7.4.2 更新策略

a） 对产生呈现周期性规律的数据应支持定时更新策略。

b） 对产生由特定事件触发的数据应支持事件触发更新策略。

c） 对产生无特定规律的数据应支持手动更新策略。

7.4.3 更新频率

a） 根据数据变化情况，数据应进行及时和持续更新。

b） 实时产生且实时性要求高的数据应进行实时更新。

c） 实时产生且实时性要求低的数据应采用定时更新。

8 数据治理要求

8.1 数据治理规划

数据治理规划的基本内容包括但不限于：

a） 建立一体化大数据平台数据治理规划组织架构，明确数据治理管理制度和职责。

b） 开展需求调研，调研一体化大数据平台数据治理现状、治理环境，明确数据治理需求和目标，形成数据治理需求调研报告。

c） 进行需求分析，对数据治理调研结果进行分析，梳理数据治理需求，包括数据模型、数据标准、数据关系、业务视图、技术视图、数据分类分级等，确认影响业务的关键数据指标，分析关键业务的数据质量，形成数据治理需求分析报告。

d） 设计治理规划，根据数据治理需求开展数据治理规划，包括数据治理战略、制度、组织、标准、流程和技术架构等，形成数据治理规划方案。

8.2 数据治理实施

数据治理的实施包括但不限于数据标准管理、元数据管理、数据清洗、数据质量检测、问题数据处理。

8.2.1 数据标准管理

a） 大数据业务主管部门结合枣庄市实际，统筹管理枣庄市数据相关标准规范。

b） 一体化大数据平台应能与省级枢纽对接获取省级管理的标准规范。

c） 应根据相关国家标准、行业标准、地方标准，按照“一数一标准”原则规范数据资源管理工作。

d） 应基于标准规范中对数据元的规范要求建立数据治理规则，对数据资源进行治理实施和处理。

8.2.2 元数据管理

a） 应根据元数据的管理范围构建元数据库。

b） 应建立元数据管理体系，保障采集数据的质量。

c） 建立元数据创建、维护、整合、存储、分发、查询、报告和分析机制。

d） 应根据法律和政策要求，负责触发数据或数据集的可访问更新。

e） 应提供元数据的生存周期管理能力。

8.2.3 数据清洗

a） 制定数据过滤策略，应对确定的无效数据、干扰数据进行数据过滤操作。

b） 根据数据相关业务的合理性，应设置重复数据判定规则，基于唯一标识符或者关键字段进行判断，去除数据集中重复记录。

c） 对于来源于不同层级、不同业务系统的数据存在数据格式和数据内容不符合数据标准时，应对数据资源进行数据关系梳理，确定数据资源整体的统一数据视图；根据数据标准进行数据转换与加载，包括但不限于代码转换、从前往后截断、从后往前截断、日期格式转换、时间格式转换、IP 地址转换、身份证号码归一化、手机号码归一化、MAC 地址转换、全角数据转换为半角数据、繁体字符转换为简体字符等。

d） 对于数据资源目录要求必填项目进行检验，对于关键字段缺失的情况，查找源头数据填充缺失值，将数据对已有权威信息的值进行识别，与数源部门确认数据补全的规则后进行补全。

8.2.4 数据质量检测

a） 数据提供方应规范数据生产，在数据汇聚前依据相关标准规范完成数据质量自查。

b） 对于汇聚到市一体化大数据平台的数据应建立数据质量检测机制，依据标准规范进行全量数据质量检测。

c） 市一体化大数据平台应具备质量检测的相关功能，确保质量规则的应用。

d） 对于多来源的相同业务数据，应进行多源数据校核。

e） 数据质量检测的结果应按照总体情况、数据提供方、数据资源等各维度形成质量分析报告，并根据数据汇聚更新情况定期生成。

f） 数据质量检测应根据数据不断汇聚定期探查数据资源变化及时更新和优化检测规则。

8.2.5 问题数据处理

问题数据处理流程见图3，流程包含以下内容：

a） 数据治理过程中会产生按照规则不能处理的、不符合条件的各种数据，应将问题数据进行存储并根据治理的实施不断更新。

b） 应按照问题数据的来源单位及时将数据推送至数据提供方，并督促其核查完善。

c） 数据提供方接收到问题数据后，应及时进行修正并作为更新数据再次汇聚，形成闭环处置流程。

d） 通过数据质量问题及其相关处理经验的汇总、分析，逐步积累形成包含数据质量检测规则、质量问题描述、针对性解决方案的数据质量知识库。

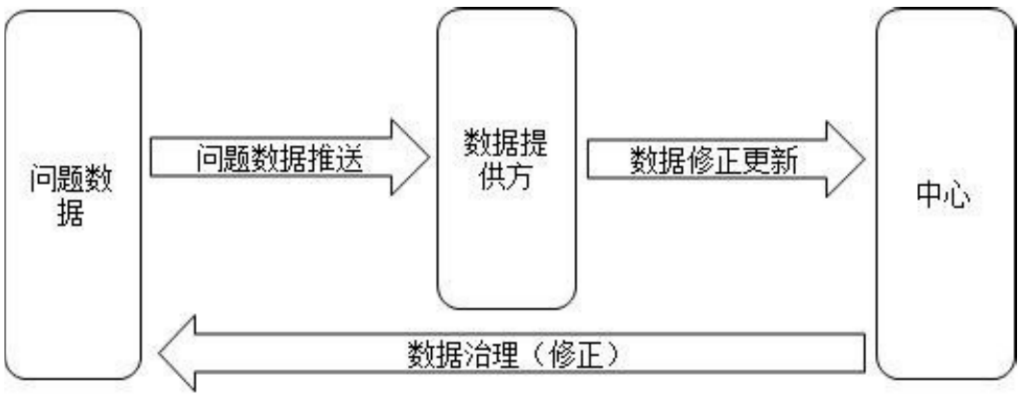

图 3 问题数据处理流程

8.2.6 数据治理结果评价

a） 应建立评价指标体系，明确评价范围、依据标准，根据不同周期、数据管理目标对评价指标进行适当调整，对数据治理结果进行综合评价；

b） 数据治理结果评价应围绕数据质量和数据安全两部分进行；数据质量评价维度包含完整性、准确性、规范性、一致性、时效性与可访问性六个维度，六个维度应符合 GB/T 36344 的规定，数据安全评价维度应包含数据采集安全、数据传输安全、数据存储安全、数据处理安全、数据交换安全、数据销毁安全六个维度。

c） 数据资源在提供数据服务和应用中，应根据服务的内容、应用的方向等不同场景，对所需的治理结果评价体系进行适当调整。

9 数据安全保护要求

a） 数据安全要求应符合 GB/T 39477-2020 的要求。

b） 数据汇聚治理安全应符合 GB/T 22239-2019 中等级保护三级的要求。

c） 个人信息安全应符合 GB/T 35273-2020 要求。

d） 对数据汇聚治理过程进行有针对性的保护，个人信息、敏感数据和重要数据应加强安全管控措施。

附 录 A
（资料性）
数据汇聚示例

A.1.1 数据库表示例

业务数据库表示例见表1

表1 业务数据库表实例

ID	业务字段 1	…	业务字段 N	批次号	操作标识	更新时间
业务主键	业务字段	…	业务字段	202312270000001	I	2023-12-23 17:23:36
业务主键	业务字段	…	业务字段	202312270000001	U	2023-12-23 17:23:36
业务主键	业务字段	…	业务字段	202312270000001	D	2023-12-23 17:23:36
业务主键	业务字段	…	业务字段	202312270000002	I	2023-12-23 18:17:09

A.1.2 服务接口示例

以test内容传输格式为例，定义数据模板，Schema验证模板标识为test，信息表名称为xinxi，XML编码为UTF-8，XML格式数据示例为：

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<test>
 <xs:schema ID="test" xmlns="" xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:msdata="urn:schemas-microsoft-com:xml-msdata">
  <xs:element name="test" msdata:IsDataSet="true" msdata:UseCurrentLocale="true">
   <xs:complexType>
    <xs:choice minOccurs="0" maxOccurs="unbounded">
     <xs:element name="xinxi">
      <xs:complexType>
       <xs:sequence>
        <xs:element name="ApeID" type="xs:string" minOccurs="0" />
        <xs:element name="mac" type="xs:string" minOccurs="0" />
        <xs:element name="ctxy" type="xs:string" minOccurs="0" />
        <xs:element name="dwID" type="xs:string" minOccurs="0" />
        <xs:element name="sjsj" type="xs:dateTime" minOccurs="0" />
       </xs:sequence>
      </xs:complexType>
     </xs:element>
    </xs:choice>
   </xs:complexType>
  </xs:element>
 </xs:schema>
<xinxi>
 <ApeID>1234567890oiuytrewq</ApeID>
 <mac>09876543219876543212</mac>
 <ctxy>56788.42</zdcq>
 <dwID>zzsswj</rdssID>
```

<sjsj>2023-12-27T13:09:58.99+08:00</cjsj>
</xinxi>
</test>

A.1.3 数据对账表示例

数据对账表示例见表2

表2 业务数据库表实例

ID	表名	批次号	批次条数	批次时间
UUID	业务表名	202312270000001	3699	2023-12-23 17:23:36
UUID	业务表名	202312270000002	12309	2023-12-23 18:17:09
UUID	业务表名	202312270000003	987	2023-12-23 20:30:02
UUID	业务表名	202312270000004	3301	2023-12-23 20:36:17