ICS

**DB52**

贵 州 省 地 方 标 准

DB 52/T XXX—XXXX

贵州省数据清洗加工规范

Big Data Cleaning Specifications for Guizhou

2018 - XX - XX 发布 2018 - XX - XX 实施

贵州省市场监督管理局 发

目 次

[前言 1](#_Toc529456610)

[1 范围 2](#_Toc529456611)

[2 术语和定义、缩略语 2](#_Toc529456612)

[2.1 术语和定义 2](#_Toc529456613)

[2.2 缩略语 3](#_Toc529456620)

[3 总则 3](#_Toc529456622)

[3.1 数据的范围 3](#_Toc529456623)

[3.2 数据清洗加工原则 3](#_Toc529456624)

[3.3 数据清洗加工与数据质量的关系 3](#_Toc529456628)

[4 数据清洗加工流程 4](#_Toc529456629)

[4.1 数据加工全流程 4](#_Toc529456630)

[4.2 数据预处理 5](#_Toc529456631)

[4.3 数据清洗加工 6](#_Toc529456636)

[5 数据清洗加工过程管理 10](#_Toc529456645)

[5.1 数据清洗角色定义 10](#_Toc529456646)

[5.2 数据提供者 10](#_Toc529456647)

[5.3 数据管理者 10](#_Toc529456648)

[5.4 数据审核者 10](#_Toc529456649)

[6 数据清洗加工质量管理 10](#_Toc529456650)

[6.1 质量控制 10](#_Toc529456651)

[6.2 质量要素 11](#_Toc529456652)

[附录A 12](#_Toc529456653)

[（资料性附录） 12](#_Toc529456654)

前言

本标准按照GB/T 1.1-2009《标准化工作导则 第1部分：标准的结构和编写》给出的规则起草。

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这些专利的责任。

本标准由贵州省大数据发展管理局提出。

本标准由贵州省大数据标准化技术委员会归口。

本标准起草单位：贵州中软云上数据技术服务有限公司、云上贵州大数据产业发展有限公司。

本标准主要起草人：秦晓东、杨建国、王仕品、XXX、XXX。

1. 范围

本标准规定了大数据清洗加工的流程、过程管理和质量管理。

本标准适用于贵州全省范围内政务大数据应用主管单位、设计单位、建设单位、实施单位及评估单位等，用于指导政府数据的清洗加工工作。

本标准针对实例层数据质量问题提供清洗加工指导。

1. 术语和定义、缩略语
   1. 术语和定义

下列术语和定义适用于本文件。

* + 1. 政务大数据

政务大数据 Government big data

政府中无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能获得更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的政府信息资产。

* + 1. 脏数据

脏数据 Dirty data

系统中的数据不在给定的范围内或对于实际业务毫无意义，或是数据格式非法，以及在系统中存在不规范的编码和含糊的业务逻辑的数据。

* + 1. 数据清洗

数据清洗 Data cleaning

发现并纠正数据资源中可识别的错误的最后一道程序。利用现有的数据挖掘手段和方法清洗“脏数据”，将“脏数据”转化为满足数据质量要求或应用要求的数据的过程称为数据清洗。

* + 1. 结构化数据

结构化数据 Structural data

能够用统一的结构加以表示的数据，如数字、符号等可以存储在数据库里,可以用二维表结构来逻辑表达实现的数据。

* + 1. 非结构化数据

非结构化数据 Unstructured data

无法用统一的结构表示的数据，如文本、图像、声音、视频、网页等不使用数据库二维逻辑表来表现的数据。

* + 1. 噪声数据

噪声数据 Noisy data

无意义的数据。

* 1. 缩略语
     1. ETL

ETL Extract-Transform-Load

将数据从来源端经过抽取（extract）、交互转换（transform）、加载（load）至目的端的过程。

1. 总则
   1. 数据的范围

本标准中的数据指的是政务数据资源，即政府部门在依法履行职能过程中生成、采集的，以一定形式记录、保存的各类数据资源，包括政府部门直接或通过第三方依法采集的、依法授权管理的和因履行职责需要依托政务信息系统形成的数据资源等。

本标准就实例层数据提供清洗加工指导。

* 1. 数据清洗加工原则
     1. 科学性原则

数据清洗加工工作应采用科学合理的规范标准，要满足现阶段科学数据资源的使用需求，也要考虑将来一定时间内由于科技快速发展等原因可能产生的数据资源应用需求，保证清洗加工后的数据资源更有生命力。

* + 1. 可操作性原则

数据清洗加工工作应考虑数据资源单位的数据资源特点以及工作的复杂程度，应切实以用户需求为导向，以应用为目标，制定可操作性的流程规范进行清洗加工。

* + 1. 一致性原则

数据清洗加工工作应统一决策，同一数据库范围内工作方法统一，技术指标统一，从而达成数据产品的一致性。

1. 数据清洗加工流程
   1. 数据加工全流程

数据清洗加工全流程如图2。

图2 数据清洗加工全周期

数据抽取

数据过滤

数据转换

数据加载

筛选阶段

结束

通过

不通过

**数据**

**预处理**

**数据**

**清洗**

**加工**

数据审核

开始

分析阶段

修正阶段

* 1. 数据预处理
     1. 数据抽取

数据抽取是从数据源中抽取数据的过程。数据抽取最常用的是ETL工具，具体数据抽取工具种类繁多，可根据实际业务数据的特点进行选择。

从数据库中抽取数据有以下两种方式：

1）全量抽取：类似于数据镜像或数据复制，它将数据源中的表或视图的数据原封不动的从数据库中抽取出来。该方法主要用于系统数据初始化时。

2）增量抽取（更新）：在上次抽取完成后，对数据库中新增或修改的数据的抽取。该方法主要用于抽取数据量庞大的系统数据。

* + 1. 数据过滤

实现对业务数据中不符合应用规则或者无效的数据进行过滤操作，使得抽取的数据标准统一。

* + 1. 数据转换

对数据的格式、信息代码、值的冲突进行转换。常见的业务数据转换规则详见附录A。

* + 1. 数据加载

将经过预处理的干净数据及脏数据分别插入到不同的数据表中。对于数据加载工作，一般会搭建数据库环境，对于数据量庞大(千万级以上)，宜使用文本文件存储结合脚本程序处理进行操作。

* 1. 数据清洗加工
     1. 数据清洗加工阶段

数据清洗加工阶段流程如图3。

否

否

是

是

定义检核规则

数据清洗加工

获取干净回流数据

结束

是

否

否

开始

分析原始数据源

是否严重脏数据

过滤去除

是

**筛选阶段**

**分析阶段**

**修正阶段**

分析数据是否需要清洗

图3 数据清洗加工阶段流程图

* + 1. 筛选阶段

在筛选阶段需要对数据的完整性和一致性进行检核，主要检查数据的数量是否满足最低要求，字段值内容是否与调查要求一致，还包括利用一些描述性统计分析检查各个字段类型、字段的最大值、最小值、平均值、中位数、缺失值个数等，进而筛选出存在的脏数据。

常见的脏数据特征为：

1. 数据缺失：缺一些记录，或者一条记录里缺一些值（空值），或者两者都缺。原因可能有很多种，系统导致的或人为导致的可能性都存在。
2. 数据重复：相同的记录出现多条。
3. 数据错误：数据没有严格按照规范记录。包括异常值、格式错误、数据不统一等。
4. 数据不可用：数据正确，但不可用。

常见的数据检核规则包括：

1. 非空检核：要求字段为非空的情况下，需要对该字段数据进行检核。
2. 主键重复检核：多个业务系统中同类数据经过清洗后，在统一保存时，为保证主键唯一性，进行检核工作。
3. 非法代码、非法值检核：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。
4. 数据格式检核：通过检查表中属性值的格式是否正确来衡量其准确性，如时间格式、币种格式、多余字符、乱码。
5. 记录数检核：各个系统相关数据之间的数据总数检核或者数据表中每日数据量的波动检核。

f） 业务约束检核:需要在项目实施过程中与业务人员共同确定，业务人员提出检核规则，从业务的角度考虑数据的正确性、一致性、有效性等。如：建档日期、入学日期、民族信息等的有效性检核。

* + 1. 分析阶段

该阶段是通过分析脏数据的产生原因和存在形式，利用现有的数据挖掘技术和方法清洗脏数据，将脏数据转化为满足数据质量要求或应用要求的数据，从而提高数据集的质量，满足现阶段数据分析的需求。

分析脏数据的两种常用方法：

1. 统计分析：统计分析可得到关于属性的信息，如：数据类型、长度、取值空间、离散值、出现频率和不同值的个数等，通过应用统计技术，计算出属性间的平均值，中间值、标准差等统计量，比对各属性的统计量可分析脏数据产生的原因和存在形式。
2. 数据挖掘：在大型数据集中发现特定的数据模式，通过数据挖掘来发现属性间的一些完整性约束，如：函数依赖和行业规则，从而对脏数据进行分析。
   * 1. 修正阶段
        1. 数据清洗路径

脏数据处理具体步骤如图4。

图4 数据清洗路径图

* + - 1. 缺失值处理

在数据集中，若某记录的属性值被标记为空缺，则认为该记录存在缺失值，是不完整的数据。

缺失值是最常见的数据问题，处理缺失值按照以下五个步骤进行：

1. 确定缺失值范围：对每个字段都计算其缺失值比例，然后按照缺失比例和字段重要性，分别制定策略。策略制定参考图5：

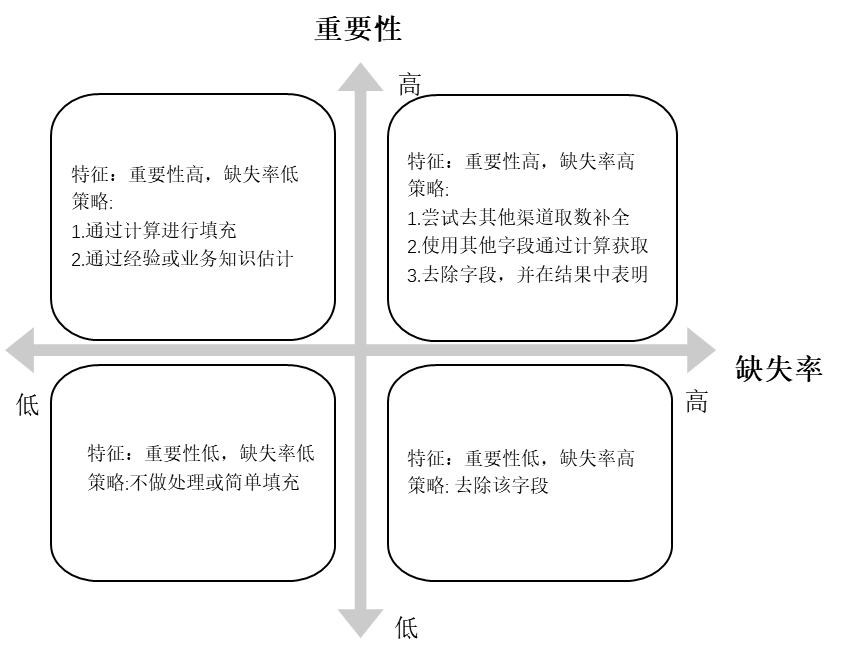


图5 缺失值范围图

1. 对于一些指标重要缺失值较低的缺失值进行填补，可根据经验或业务知识估计，也可通过计算。
2. 如果某些指标非常重要又缺失率高，那需要和取数人员或业务人员了解是否有其他渠道可以取到相关数据，必要时进行重新采集。若无法取得相关数据，则需要对缺失值进行填补。
3. 对于重要性低并且缺失率也低的缺失值可只进行简单填充或不作处理。
4. 一些重要性低缺失率低的缺失值则可去除不需要的字段：备份当前数据，直接删掉不需要的字段。

填补空缺值的方法有以下三种：

1. 以业务知识或经验推测填充缺失值；
2. 以同一指标的计算结果(均值、中位数、众数等)填充缺失值；
3. 以不同指标的计算结果填充缺失值；比如：年龄字段缺失，但是具有公民身份证号，可以从公民身份证号提取年龄数据。
   * + 1. 错误数据处理

错误数据包含格式内容问题数据和逻辑问题数据两类。

1. 格式内容问题数据有以下三类：
2. 时间、日期、数值、全半角等显示格式不一致：将其处理成一致的某种格式。
3. 内容中有不该存在的字符：需要以半自动校验半人工方式来找出可能存在的问题，并去除不需要的字符。
4. 内容与该字段应有内容不符：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的问题等）后再处理。
5. 逻辑问题数据处理采用逻辑处理的方法，可以去掉一些使用简单逻辑推理即可直接发现问题的数据，防止分析结果错误。

主要包含以下三个步骤：

1. 去重：主要使用字段相似度来识别判断重复值，删除重复值并保留唯一值。
2. 离群值（异常值）：

识别离群值的方法有如下几种：

* 数据分布特征及箱型图方法

利用数据自身分布特征可采用直观的箱型图方法可视化识别离群值。

* 基于欧几里德距离的聚类方法

通过对数据之间的距离测量识别数据的离群值。

识别离群值后，操作人员需要按照经验和业务流程判断其值的合理性，若此数值合理，则保留该数值；若不合理，则按照其重要性考虑是否需要重新采集。对于重要性较高而又无法重新采集的数值，按照缺失值办法处理。对于重要性较低数值，可直接去除。

1. 修正矛盾内容

有些字段可以互相验证。需要根据字段的数据来源，来判定哪个字段提供的信息更可靠，去除或重构不可靠字段。

逻辑错误除以上列举情况，还有很多其他情况，在实际操作中需根据实际情况处理。

* + - 1. 错误关联数据处理方法

如果数据有多个来源，有必要进行关联性验证。

错误关联数据清洗方法主要有以下方法：

a）统计学方法：将属性当做随机变量，通过置信区间来判断值的正误。

b）基于聚类的方法：根据数据相似度将数据分组，发现不能归 并到分组的孤立点。

c）基于距离的方法：使用距离度量来量化数据对象之间的相似性。

d）基于分类的方法：训练一个可以区分正常数据和异常数据的分类模型。

e）基于关联规则的方法：定义数据之间的关联规则，不符合规则的数据被认为是异常数据。

1. 数据清洗加工过程管理
   1. 数据清洗角色定义

数据清洗管理涉及的数据管理角色有提供者、管理者和审核者。提供者负责提供清洗的业务数据，管理者负责数据清洗系统的基本运行管理、数据清洗规则制定、数据清洗发起等。项目中提供者为数据接入方，管理者为项目建设方。审核者负责把控数据经过清洗加工后的质量，按照审核指标检查提交的数据，将不符合审核指标的数据退回。

* 1. 数据提供者

提供者应配合管理者根据接入数据指标规范与接入数据内容、接入数据流程要求，配置与部署接入服务，实现接入数据库的数据交换；提供者应该提供待清洗数据的数据结构；提供者应接收数据清洗系统的问题数据，及时修改，并通知管理者。

* 1. 数据管理者

管理者对数据清洗系统的管理要点包括:

a）管理者应负责协调并明确数据清洗规则;

b）管理者应负责构建清洗后数据库及问题数据库数据表的结构；

c）管理者应负责将问题数据库提交给提供者，并协调提供者进行数据完善。

* 1. 数据审核者

数据审核的目标是确保数据内容与被描述对象相一致，并且质量符合数据产品标准要求。

数据审核可以贯穿于整个数据资源加工过程之中，可以量化评价的内容包括数据来源质量评价、数据加工模型与算法质量评价、数据产品质量评价等。

数据审核可以由数据采集加工人员自检，也可由数据库主要承建单位专门进行。适宜时，数据审核宜采取计算机辅助方法进行。

数据库主要承建单位应明确审核所参照的评估模型和方法以及技术要求等。如果学科领域内已存在相关的数据质量管理国际、国家规范或行业标准，数据审核宜采用这些相关标准。

审核指标的设置应在符合实际的前提下尽可能不应与当前国际领先水平有太大差距。审核指标可以包括但不限于准确性，真实性误差等技术参数，特色数据和重点数据宜适当提高指标。

数据资源审核通过后方可正式对用户提供服务，未能通过审核的数据一般应返回到必要的流程进行修正或重新加工。

1. 数据清洗加工质量管理
   1. 质量控制

数据清洗的质量控制是对清洗后的数据的质量进行评估，主要对数据的形式、内容进行评估。

* 1. 质量要素

数据质量评估包含但不限于以下方面的评估指标：

1. 规范性: 数据的存在性、质量及存储标准的统一。反映信息表达的准确程度。
2. 重复性：考察数据在字段、记录内容或数据集内的重复值。
3. 准确性: 数据所指内容对数据所指对象的反应、表现是否准确，以及数据形式对数据内容的表述、表达是否准确。反映数据对业务对象描述的准确程度。
4. 完整性:数据集合中包含足够的数据来回答各种查询和支持各种计算。
5. 一致性: 同一个数据在同一时刻在不同数据库、应用和系统中只有一个值。反映数据资源描述要素、属性和它们间的相互关系符合逻辑规则的程度。
6. 时效性：考察数据的时间特性对业务应用的满足程度。不同类型的应用对数据的时间特性有不同的要求。反映数据资源更新的及时程度和对业务应用的支持程度。
7. 稳定性：数据来源稳定，考察数据的“可信赖的”或“可信任的”程度。
8. 可理解性: 阅读并理解数据资源编码的难易程度。反映数据资源描述的规范性以及数据内容的表达符合国内外相关规范、标准的程度。

附录A

（资料性附录）

常见业务数据转换规则表

| 转化规则 | 规则描述 |
| --- | --- |
| 统一时间日期数据格式 | 将各类日期统一为八位的字符日期，如YYYYMMDD； |
| 将各类时间统一为六位的字符时间，如HHMMSS； |
| 将各类事件日期统一为十四位的字符时间日期，如YYYYMMDDHHMMSS。 |
| 统一分类数据取值代码 | 将人员的性别数据统一转换为国标性别信息代码 |
| 将人员的民族数据统一转换为标准信息代码 |
| 将人员的户籍地址数据统一转换为行政区划代码 |
| 将人员的婚姻登记情况统一转换为标准的婚姻状况代码 |
| 将公民身份证统一转换为18位的身份证号 |