

**数据质量监控方案**

**目 录**

1. 概述 1

2. 数据质量保障原则 1

3. 实现方案 2

3.1. 规则定义 3

3.1.1. 通用规则定义 3

3.1.2. 表结构设计 4

3.2. 离线计算流程 6

3.3. 实时 6

3.4. 告警 6

4. 展示示例 7

4.1. 告警 7

4.2. 事前 7

4.3. 事中 8

4.4. 事后 8

# 概述

随着IT向DT时代的转变，数据已经成为企业的重要资产，为企业业务开展、战略决策提供有利支撑。数据质量是确保报表、分析、应用的准确的基础。

# 数据质量保障原则

评价数据质量好坏通常有几个方面，完整性，准确性、有效性、一致性、及时性、稳定性。

* **完整性：**描述数据是否存在缺失记录或缺失字段。

举例：某个字段不能为null或空字符。

* **准确性：**描述数据是否与其对应客观实体的特征一致。

举例：用户的住址是否准确；某个字段规定应该是英文字符，在其位置上是否存在乱码。

* **有效性：**描述数据是否满足用户定义的条件或在一定的取值范围内。

举例：年龄的值域在0~200之间。另一个枚举的有效性例子是银行的币种代码。

* **一致性：**描述同一实体同一属性的值在不同的系统中是否一致。

举例：男女是否在不同的库表中都使用同一种表述。例如在A系统中，男性表述为1，女性 表述为0；在B系统中，男性表述为M，女性表述为F。

* **及时性：**描述数据的产生和供应是否及时。

举例：生产数据必须在凌晨2：00入库到ODS（Operational Data Store，操作数据层）。

* **稳定性：**描述数据的波动是否稳定，是否在其有效范围内。

举例：产品质量抽样统计的合格率，不会有超过20%的波动范围。

为了全方位监测数据采集、计算过程，需要对数据处理进行事前、事中、事后三阶段监控，并持续跟踪数据质量事件，不断提升数据质量。

* **事前：**检测数据处理必备条件，以及输入数据的质量情况。

例如：集群资源是否足够，集群权限（如文件数）是否满足；输入数据总大小是否异常、文件数是否异常、字段数是否异常、字段值是否准确、有限、一致、数据是否延迟等。

* **事中：**检测数据处理过程是否异常。

例如：处理环节是否失败，处理时长是否异常、业务指标埋点统计等。

* **事后：**检测计算结果是否异常。

例如：结果数据文件大小是否异常、数据条数是否异常、关键指标是否异常、指标波动是否异常等。

# 实现方案

总体分事前、事中、事后三阶段监控，及时发现数据异常问题，并进行告警，辅助快速修复。

* **事前：**根据数据质量保证原则，从完整性，准确性、有效性、一致性、及时性、稳定性等方面定义监测指标，进行相关规则的配置，并进行数据质量统计，生成统计报表与异常告警。
* **事中：**进行计算步骤、计算开始时间、结束时间等计算基本信息以及业务指标埋点日志的记录，通过日志分析计算过程的有效性、及时性、稳定性，计算指标的完整性等。
* **事后：**从数据量监控计算结果是否完整、稳定，从关键指标波动情况监测指标计算是否异常。
* **告警：**汇聚数据处理异常情况，对异常问题进行告警，有条件可以进行邮件、短信告警。并进行告警分类统计、排名等，辅助关键问题排查。

## 规则定义

### 通用规则定义

通用性的规则可以固化，应用于多种业务监控。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 通用规则 | | | | 告警级别 |
| 准则 | 示例1 | 示例2 | 示例3 |
| 准确性 | 字段长度 | 字符串是否合法 | 类型 | 不准确 |
| 完整性 | 条数 | 列数 | 文件损坏 | 不完整 |
| 一致性 | 表述是否一致 |  |  |  |
| 有效性 | 空值 | 字段值范围 |  | 无效 |
| 唯一性 | 重复 |  |  |  |
| 及时性 | 数据推送是否及时 | 种类 | 配置更新（过期） | 无法计算 |
| 稳定性 | 统计值是否合理 | 历史数据波动幅度 | 文件大小 | 不稳定 |
| 连续性 | 时间是否连续 |  |  | 不完整 |
| 合理性 | 结束时间晚于开始时间 |  |  | 无法计算 |

### 表结构设计

#### 数据源表配置表

tb\_metadata\_config\_src\_data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 中文名 | 字段类型 | 是否主键 | 是否可空 | 备注 |
| 1 | id | 序号 | int | 是 | 否 |  |
| 2 | province\_id | 省id | string |  | 否 |  |
| 3 | table\_name | 表名 | string |  | 否 |  |
| 4 | push\_time | 推送时间 | string |  | 否 |  |
| 5 | save\_format | 文件格式 | string |  | 否 |  |
| 6 | compression\_type | 压缩格式 | string |  | 否 |  |
| 7 | save\_path | 保存路径 | string |  | 否 |  |
| 8 | field\_num | 字段数 | int |  | 否 |  |
| 9 | delimiter | 分隔符 | string |  | 否 |  |
| 10 | header | 是否带标题 | bit |  | 否 |  |
| 11 | hasTimeSuffix | 是否带时间后缀 | bit |  | 否 |  |

#### 数据源字段配置表

tb\_metadata\_config\_field

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 中文名 | 字段类型 | 是否主键 | 是否可空 | 备注 |
| 1 | id | 序号 | int | 是 | 否 |  |
| 2 | province\_id | 省id | string |  | 否 |  |
| 3 | table\_name | 表名 | string |  | 否 |  |
| 4 | field\_name | 字段名 | string |  | 否 |  |
| 5 | index | 字段位置序号 | int |  | 否 | -1代表字段不存在 |
| 6 | field\_type | 字段类型 | string |  | 否 |  |
| 7 | nullable | 是否存在空 | string |  | 否 |  |
| 8 | default\_val | 默认值 | string |  | 否 |  |
| 9 | value\_range | 值域范围 | string |  | 否 |  |
| 10 | comment |  | string |  | 是 |  |

#### 告警门限配置表

tb\_metadata\_config\_alert\_threshold

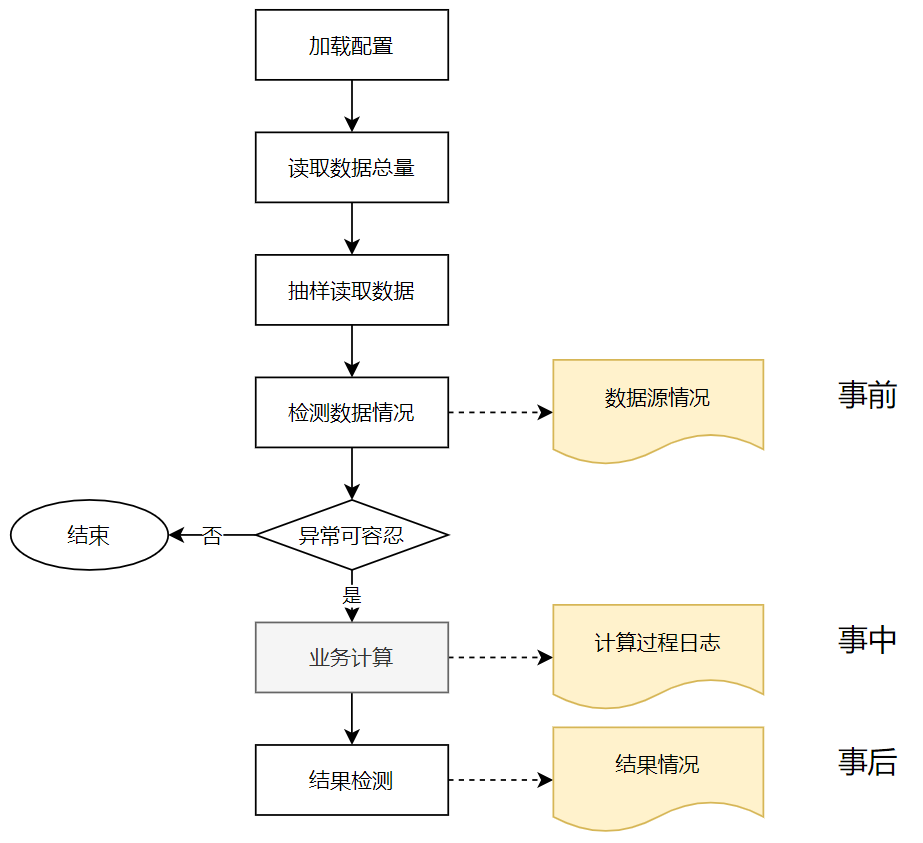
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 中文名 | 字段类型 | 是否主键 | 是否可空 | 备注 |
| 1 | id | 序号 | int | 是 | 否 |  |
| 2 | province\_id | 省id | string |  | 否 |  |
| 3 | table\_name | 表名 | string |  | 否 |  |
| 4 | field\_name | 字段名 | string |  | 否 |  |
| 5 | minimum\_valid\_percentage\_alert\_threshold | 合法告警门限比例 | float |  | 否 |  |

## 离线计算流程

在业务计算开始前，首先进行业务计算基本条件的具备情况，然后对数据源进行健康度检测，输出数据源检测结果。

如果发现不可容忍的异常，直接终止计算，没有异常或者异常可容忍则继续进行业务计算，业务计算过程中输出业务埋点日志，供后续运行健康度分析。

业务计算完成后，对事前、事中、事后数据进行异常检测，生成告警。



## 实时

实时计算为了保障运算效率，不进行事前检测，只做业务计算过程中做埋点日志。通过对实时计算日志与输出结果进行异常检测。

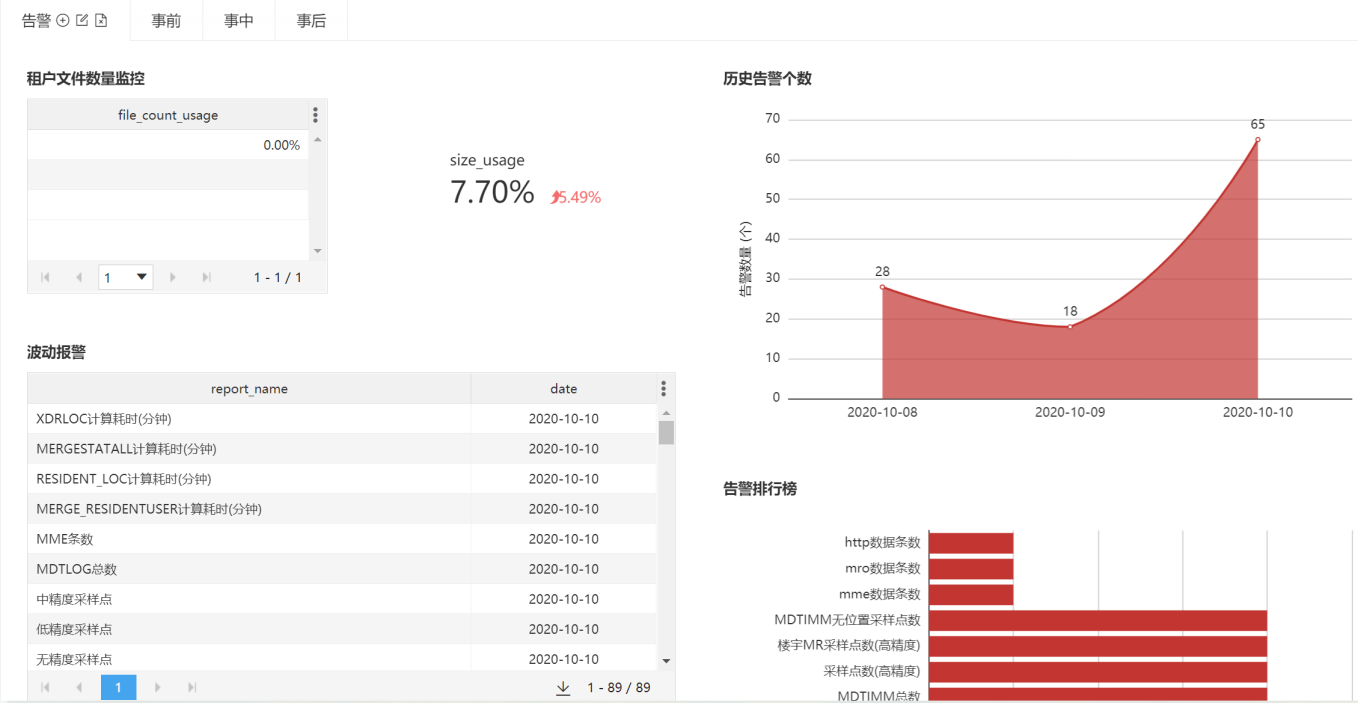
#### 

## 告警

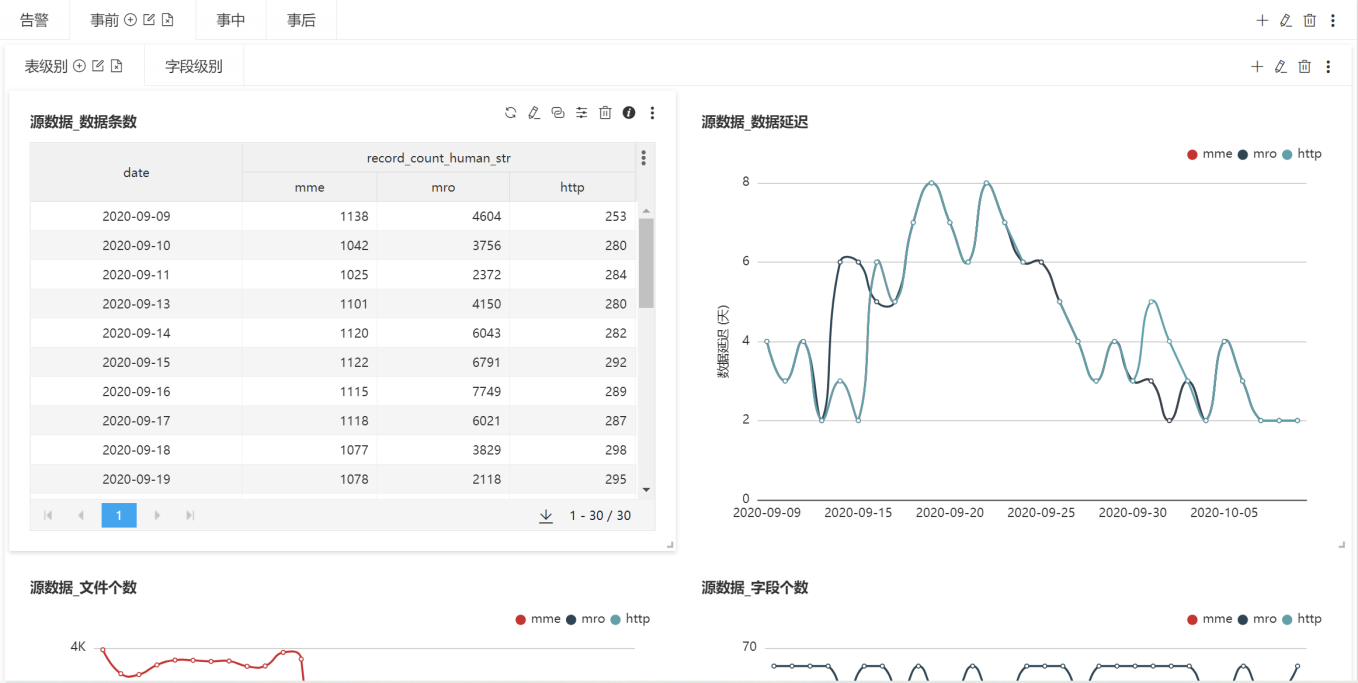
采用箱型图法检测各指标是否存在异常情况，异常指标生成告警。

# 展示示例

## 告警



## 事前



## 事中



## 事后

