数据质量风险监控

更新时间：2020-06-17 20:02:49

[**本页目录**](javascript:void(0))

* [前提条件](https://help.aliyun.com/document_detail/123975.html?spm=a2c4g.11186623.6.637.1c6b415cuob0Yo#h2-url-1)
* [背景信息](https://help.aliyun.com/document_detail/123975.html?spm=a2c4g.11186623.6.637.1c6b415cuob0Yo#h2-url-2)
* [操作步骤](https://help.aliyun.com/document_detail/123975.html?spm=a2c4g.11186623.6.637.1c6b415cuob0Yo#h2-url-3)

数据质量风险监控主要针对数据的准确性、一致性和完整性。本教程使用DataWorks数据质量（DQC）功能，完成数仓各层次的数据质量监控。

**前提条件**

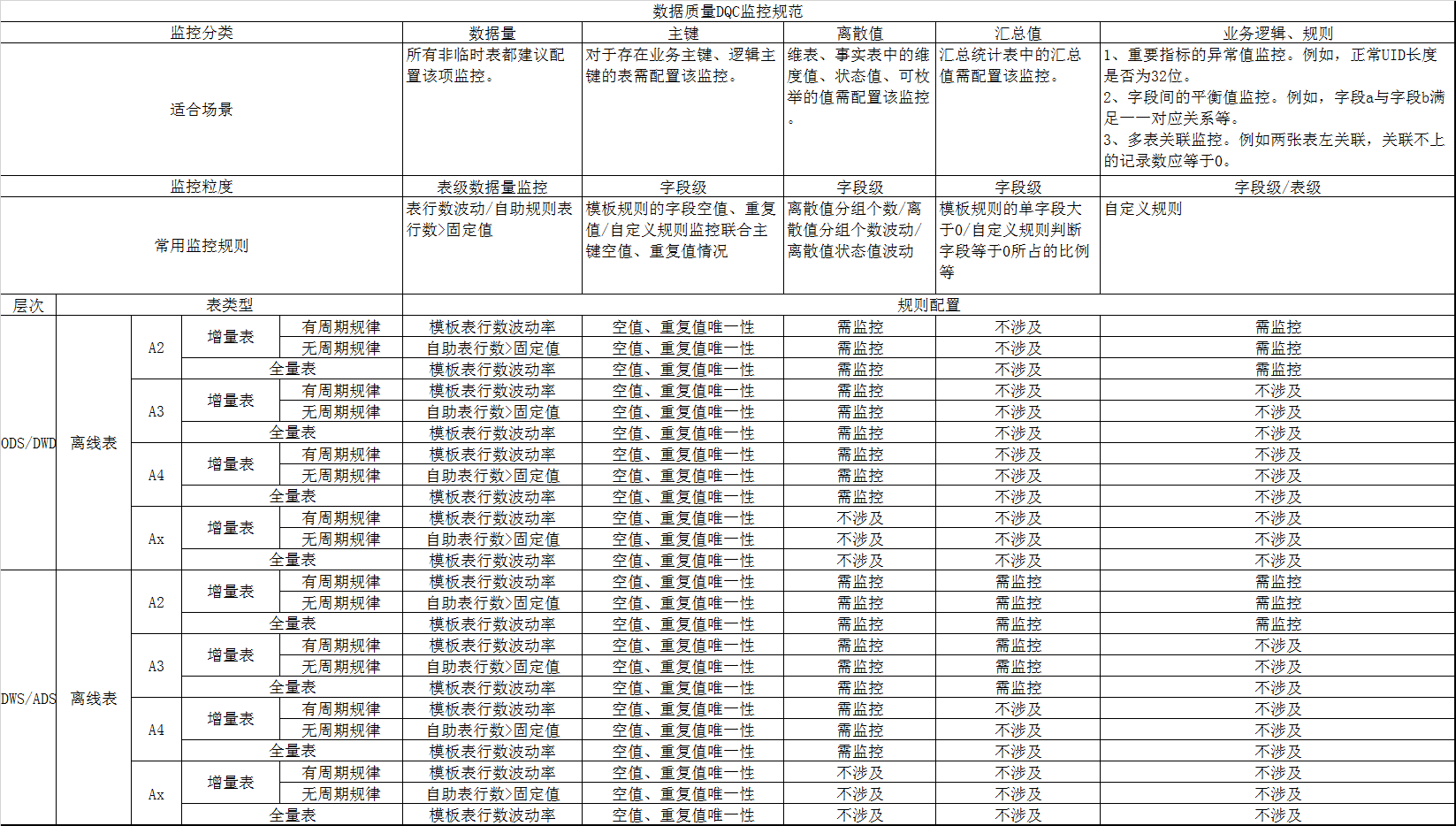
首先您需要完成教程[搭建互联网在线运营分析平台](https://help.aliyun.com/document_detail/122338.html" \l "concept-711325" \o "本教程基于大数据时代在线运营分析平台的基础需求，为开发者提供从数据高并发写入存储、便捷高效的数据加工处理到数据分析与展示的全链路解决方案。本教程帮助您了解并操作阿里云的大数据产品，完成在线运营分析平台的搭建。)，并保证您的DataWorks工作空间创建区域为华东2（上海），详情参见[业务场景与开发流程](https://help.aliyun.com/document_detail/122338.html" \l "concept-711325" \o "本教程基于大数据时代在线运营分析平台的基础需求，为开发者提供从数据高并发写入存储、便捷高效的数据加工处理到数据分析与展示的全链路解决方案。本教程帮助您了解并操作阿里云的大数据产品，完成在线运营分析平台的搭建。)。您需要完成数据资产定级，本教程中定义为A2，详情请参见[数据资产定级](https://help.aliyun.com/document_detail/123971.html" \l "concept-977415" \o "数据的资产等级，可以根据数据质量不满足完整性、准确性、一致性、及时性对业务的影响程度进行划分。)。

**说明** 数据质量风险监控理论规范，请参见[数据风险点监控](https://help.aliyun.com/document_detail/116897.html" \l "concept-221932" \o "本文为您介绍在线数据风险点监控和离线数据风险点监控。)。

**背景信息**

数据质量监控和数据资产等级对应，您可以根据以下因素细化您的监控配置，数据质量的详情请参见[概述](https://help.aliyun.com/document_detail/73660.html?spm=a2c4g.11186623.2.17.72412109FeWv5r#concept-zsz-44h-r2b)。

* 监控分类：数据量、主键、离散值、汇总值、业务规则和逻辑规则。
* 监控粒度：字段级别、表级别。
* 监控层次：ODS、CDM、ADS三层数据，其中ODS和DWD层主要偏重数据的完整性和一致性。DWS和ADS层数据量较小、逻辑复杂，偏重数据的准确性。

以下为不同数据资产等级和数仓层次数据的数据质量监控建议，仅供参考。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/6635932951/p50862.png)

**操作步骤**

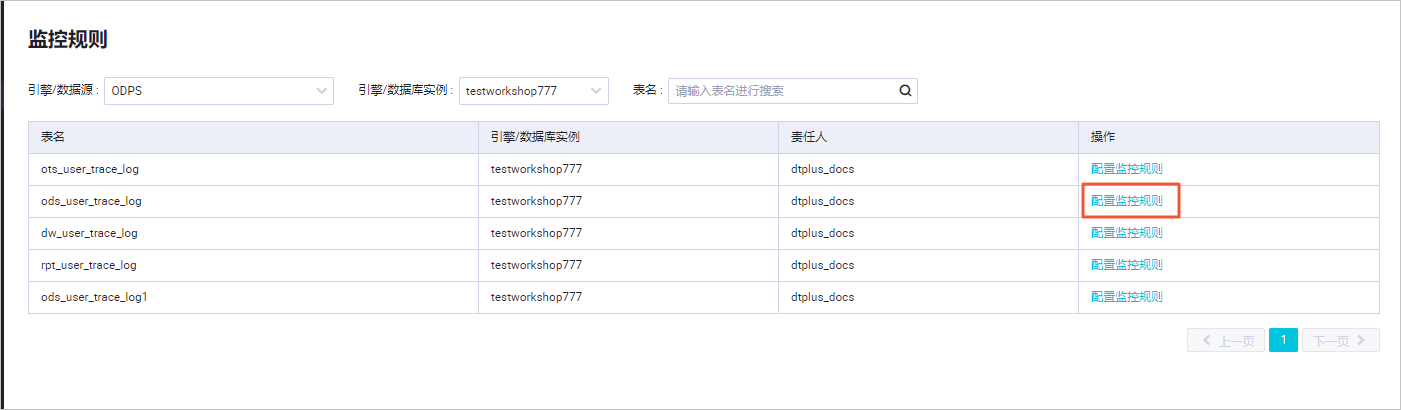
1. ODS层数据质量监控。

ODS层表中的数据来源于OSS上的日志文件，作为源头表，您需要尽早判断此表分区中是否有数据。如果这张表中没有数据，则后续任务运行无意义，需要阻止后续任务运行。

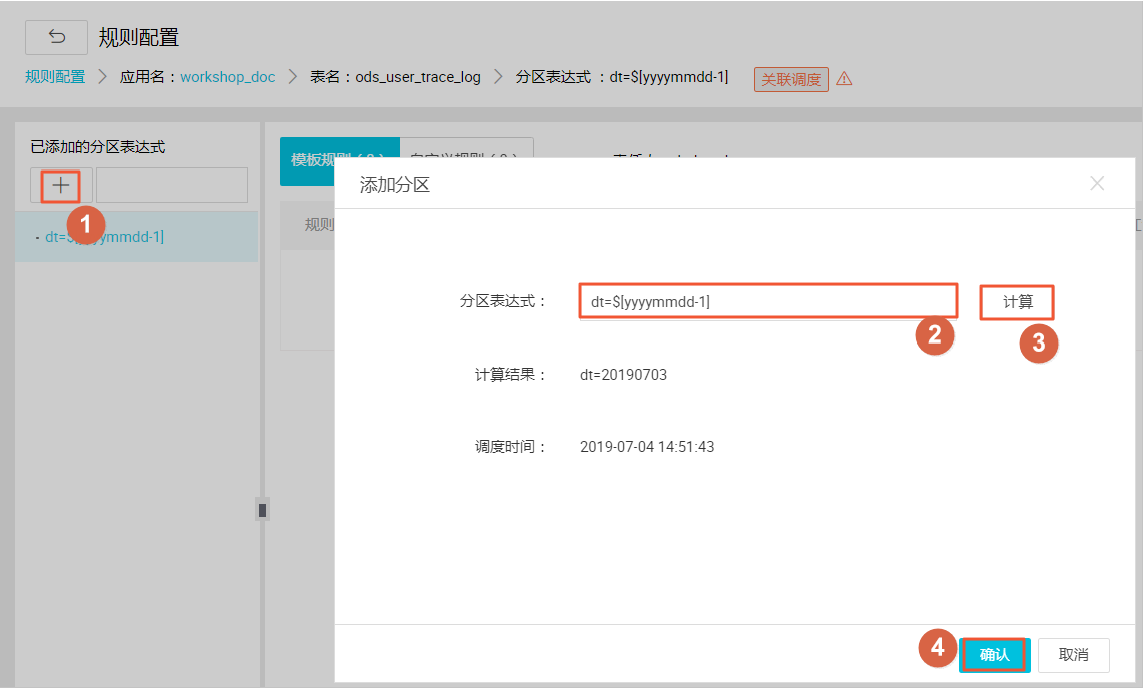
* 1. 进入**数据质量**页面。

在**数据开发**页面，单击左上角图标，选择**数据质量**。

* 1. 进入ods\_user\_trace\_log监控规则页面。

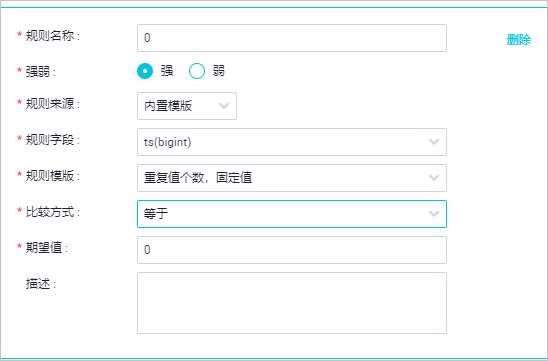
单击左侧导航栏上的**监控规则**，在**监控规则**页面找到代表外部数据源的ODS层表ods\_user\_trace\_log，单击其后的**配置监控规则**。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/6635932951/p93169.png)

* 1. 添加分区。

[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/6635932951/p50857.png)

* + 1. 单击**+**，选择**分区表达式**为**dt=$[yyyymmdd-1]**，对应表ods\_user\_trace\_log的分区格式为**${bdp.system.bizdate}**（即获取到前一天的日期）。分区表达式的详细信息请参见[调度参数](https://help.aliyun.com/document_detail/137548.html" \l "concept-t2q-jmq-p2b" \o "常见的数据研发场景中，各类型节点的代码需要根据调度周期动态赋值（例如日期时间），替换变量后再进行计算。DataWorks的调度参数配置功能可以满足该业务场景的需求。)。如果表中无分区列，可以配置无分区。
    2. 单击**计算**，验证计算结果是否正确。
    3. 单击**确认**，完成分区的添加。
  1. 创建规则确保ODS层表分区内存在数据。
     1. 单击**创建规则**。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/7635932951/p93173.png)
     2. 单击**模板规则 > 添加监控规则**。
     3. 输入配置参数。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/7635932951/p93174.png)

| **参数** | **描述** |
| --- | --- |
| **规则名称** | 请输入规则名称。您可以自定义。 |
| **强弱** | 设置为强规则。强弱规则说明如下：   * + - * 如果设置强规则，红色异常报警并阻塞下游任务节点，橙色异常报警不阻塞。       * 如果设置弱规则，红色异常报警不阻塞下游任务节点，橙色异常不报警不阻塞。 |
| **规则来源** | 选择**内置模板**。 |
| **规则字段** | 选择**表级规则**。 |
| **规则模板** | 选择**表行数，固定值**。 |
| **比较方式** | 选择**大于**。 |
| **期望值** | 设置为0。 |

* 1. 监控重复数据。
     1. 单击**添加监控规则**。
     2. 输入配置参数。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/7635932951/p93176.png)

| **参数** | **描述** |
| --- | --- |
| **规则名称** | 请输入规则名称。您可以自定义。 |
| **强弱** | 设置为强规则。强弱规则说明如下：   * + - * 如果设置强规则，红色异常报警并阻塞下游任务节点，橙色异常报警不阻塞。       * 如果设置弱规则，红色异常报警不阻塞下游任务节点，橙色异常不报警不阻塞。 |
| **规则来源** | 选择**内置模板**。 |
| **规则字段** | 选择**ts(bigint)**。**ts(bigint)**值为用户时间戳，目的是避免ODS层出现重复的数据。 |
| **规则模板** | 选择**重复值个数、固定值**。 |
| **比较方式** | 选择**等于**。 |
| **期望值** | 设置为0。 |

* 1. 监控空值数据。
     1. 单击**添加监控规则**。
     2. 输入配置参数。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/7635932951/p93177.png)

| **参数** | **描述** |
| --- | --- |
| **规则名称** | 请输入规则名称。您可以自定义。 |
| **强弱** | 设置为强规则。强弱规则说明如下：   * + - * 如果设置强规则，红色异常报警并阻塞下游任务节点，橙色异常报警不阻塞。       * 如果设置弱规则，红色异常报警不阻塞下游任务节点，橙色异常不报警不阻塞。 |
| **规则来源** | 选择**内置模板**。 |
| **规则字段** | 选择**uid(string)**。**uid(string)**值为用户ID，目的是避免出现用户ID为空值的脏数据。 |
| **规则模板** | 选择**空值个数、固定值**。 |
| **比较方式** | 选择**等于**。 |
| **期望值** | 设置为0。 |

* 1. 批量保存规则。完成上述操作后，单击**批量保存**。
  2. 规则试跑。

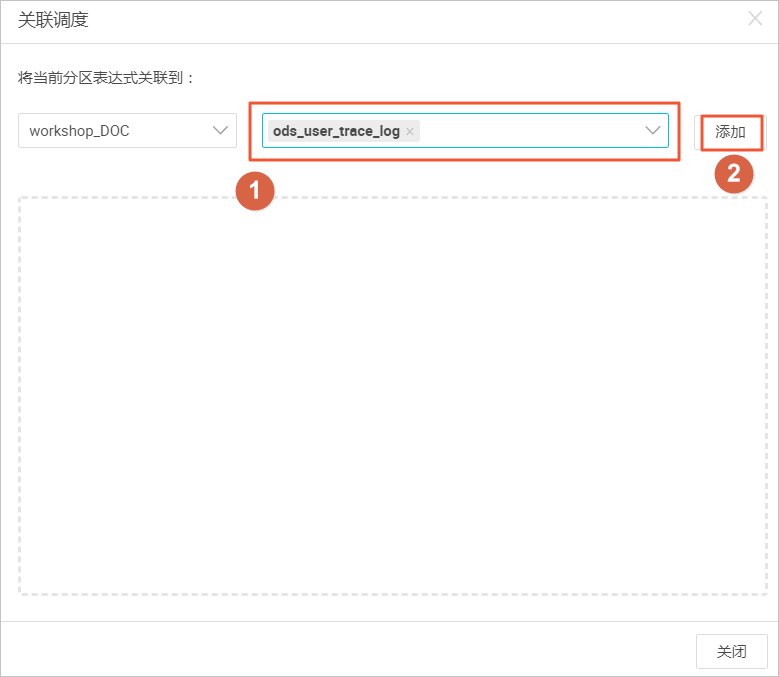
单击**试跑**，进行数据质量的校验规则。

* 1. 查看试跑结果。

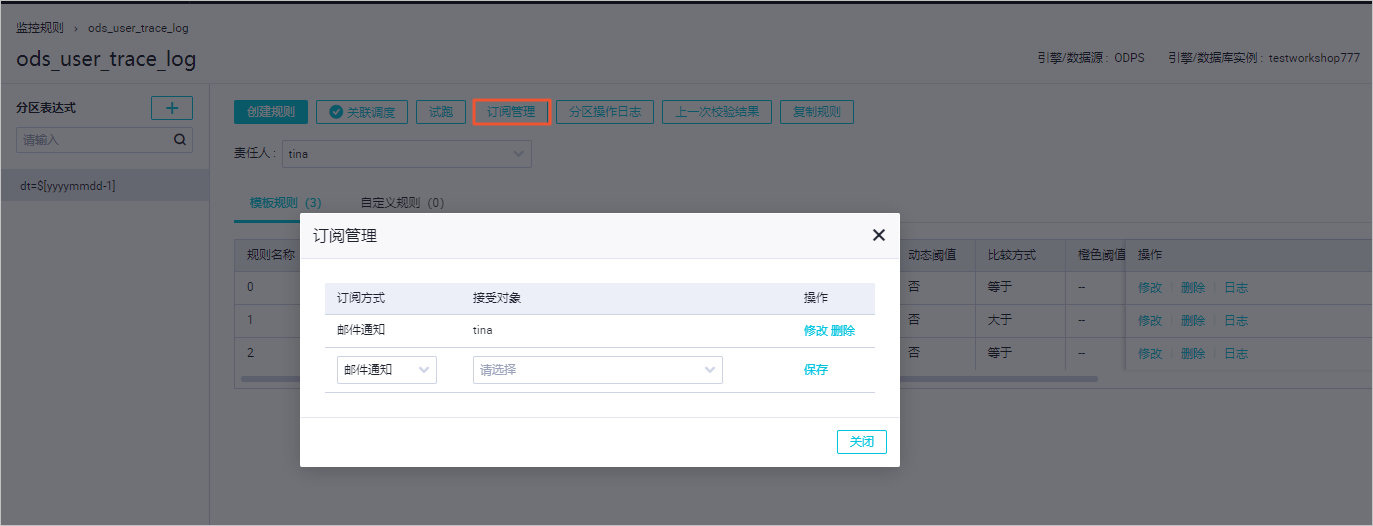
试跑完成后，单击**试跑成功！点击查看试跑结果**查看试跑结果。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/7635932951/p50914.png)在弹出的页面中，您可以查看表数据是否已符合您的规则。根据试跑结果，可以确认此次任务产出的数据是否符合预期。建议每个表规则配置完毕后，都进行一次试跑操作，以验证表规则的适用性。

* 1. 关联调度。

在规则配置完毕，且试跑成功的情况下，您需要将表和其产出任务进行关联，这样每次表的产出任务运行完毕后，都会触发数据质量规则的校验，以保证数据的准确性。

* + 1. 在表监控规则页面，单击**关联调度**，配置规则与任务的绑定关系。
    2. 在**关联调度**弹框中输入您需要关联的任务节点名称，单击**添加**。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/7635932951/p50917.png)
    3. 单击**关闭**退出**关联调度**页面。如下图所示，关联调度配置成功。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/8635932951/p93179.png)
  1. 配置任务订阅。

关联调度后，每次调度任务运行完毕，都会触发数据质量的校验。数据质量支持设置规则订阅，可以针对重要的表及其规则设置订阅，设置订阅后会根据数据质量的校验结果进行告警，从而实现对校验结果的跟踪。

单击**订阅管理**，设置接收人以及订阅方式。目前支持**邮件通知**、**邮件和短信通知**、**钉钉群机器人**和**钉钉群机器人@ALL**四种方式。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/8635932951/p93180.png)

订阅管理设置完毕后，单击左侧导航栏上的**我的订阅**进行查看及修改，建议您订阅所有规则。

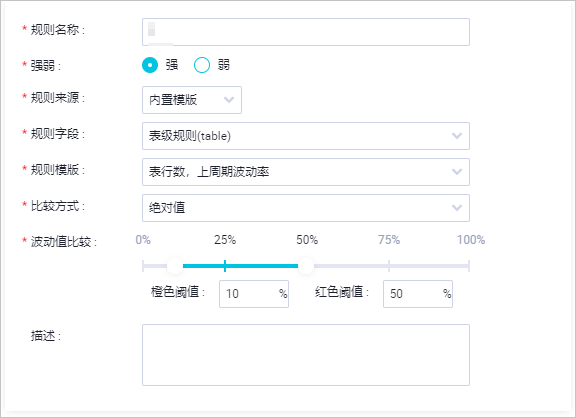
1. CDM层数据质量监控。

CDM层数据质量监控配置方法与ODS层相同，区别在于监控规则不同。

* 1. 添加分区表达式。

进入dw\_user\_trace\_log表的规则配置页面，与ODS层一样配置分区为**dt=$[yyyymmdd-1]**，确保分区内存在表数据。

* 1. 监控表行数及空值数据。表行数和空值数据的监控规则配置与ODS层相同。
  2. 监控表行数波动率。

[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/8635932951/p93184.png)

| **参数** | **描述** |
| --- | --- |
| **规则名称** | 请输入规则名称。您可以自定义。 |
| **强弱** | 设置为强规则。强弱规则说明如下：   * + 1. 如果设置强规则，红色异常报警并阻塞下游任务节点，橙色异常报警不阻塞。     2. 如果设置弱规则，红色异常报警不阻塞下游任务节点，橙色异常不报警不阻塞。 |
| **规则来源** | 选择**内置模板**。 |
| **规则字段** | 选择**表级规则（table）**。 |
| **规则模板** | 选择**表行数、上周期波动率**。 |
| **比较方式** | 选择**绝对值**。 |
| **波动值比较** | **橙色阈值**为10，**红色阈值**为50，代表当表行数波动率到达50%时，会产生红色报警。 |

* 1. 规则试跑并关联调度。方法和ODS层一致。

1. ADS层数据质量监控。

ADS层数据质量监控配置方法与ODS层相同，区别在于监控规则的不同。

* 1. 添加分区表达式。

进入rpt\_user\_trace\_log表的规则配置页面，同样配置分区为**dt=$[yyyymmdd-1]**。

* 1. 监控表行数、波动率及空值数据。

监控表行数、波动率和空值数据的监控规则配置与CDM层相同。由于在数仓分层中，越靠近应用层数据越少、约束性越低，**强弱**选择为**弱**。

* 1. 监控表异常PV。

您可以利用自定义规则功能监控ADS层的应用数据。

* + 1. 单击**自定义规则 > 添加监控规则**。
    2. 配置自定义规则参数。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/8635932951/p93183.png)

| **参数** | **描述** | |
| --- | --- | --- |
| **规则名称** | 请输入规则名称。您可以自定义。 | |
| **强弱** | 设置为弱规则。强弱规则说明如下：   * + - * 如果设置强规则，红色异常报警并阻塞下游任务节点，橙色异常报警不阻塞。       * 如果设置弱规则，红色异常报警不阻塞下游任务节点，橙色异常不报警不阻塞。 | |
| **规则字段** | 选择规则字段为**pv(bigint)**。 | |
| **采样方式** | 选择**sum**。 | |
| **校验类型** | 选择**数值型**。 | |
| **校验方式** | 选择**与固定值比较**。 | |
| **比较方式** | 选择**大于**。 |
| **期望值** | 设置为100。当PV和异常锐减到100时，您可以及时收到告警。 |

* + 1. 完成配置后，单击**批量保存**。
  1. 规则试跑并关联调度。方法与ODS层一致。