数据质量是数据分析结论有效性和准确性的基础。本文为您介绍数据质量保障教程的业务场景以及如何衡量数据质量的高低。

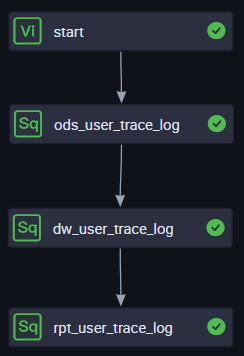
**前提条件**

在开始本教程前，请您首先完成[搭建互联网在线运行分析平台](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。)教程，详情请参见[业务场景与开发流程](https://help.aliyun.com/document_detail/122338.html#concept-711325)。

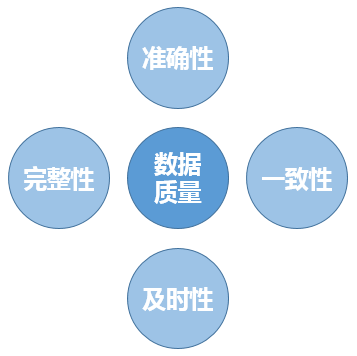
**业务场景**

要保证业务数据质量，首先您需要明确数据的消费场景和加工链路。

本教程使用的数据来源于某网站上的HTTP访问日志。基于这份网站日志，您可以统计并展现网站的浏览次数（PV）和独立访客（UV），并能够按照用户的终端类型（如Android、iPad、iPhone、PC等）和地域分别统计。

在整体数据链路的处理过程中，为保证最终产出数据的质量，您需要对数据仓库ODS、CDM和ADS层的数据分别进行监控。数据仓库分层的定义请参见[数仓分层](https://help.aliyun.com/document_detail/114447.html" \l "concept-184862" \o "在阿里巴巴的数据体系中，我们建议将数据仓库分为三层，自下而上为：数据引入层（ODS，Operation Data Store）、数据公共层（CDM，Common Data Model）和数据应用层（ADS，Application Data Service）。)。本教程基于[搭建互联网在线运行分析平台](https://help.aliyun.com/document_detail/122338.html#concept-711325)教程，ods\_user\_trace\_log、dw\_user\_trace\_log、rpt\_user\_trace\_log分别代表数据仓库的ODS、CDM和ADS层，详情请参见[设计工作流](https://help.aliyun.com/document_detail/122353.html" \l "task-711342" \o "通过设计工作流，您可以明确在整体数据开发过程中各任务节点的排布。对于本教程中这种较为简单的单数据流场景，您可以选择每个数据表（数仓层次）对应一个工作流。)。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/6435932951/p50680.png)

**数据质量的评估**

数据质量可以从完整性、准确性、一致性和及时性共四个角度进行评估，详情请参见[数据质量评估标准](https://help.aliyun.com/document_detail/116880.html#concept-221766)。[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/6435932951/p50681.png)在本教程中，您将学会通过数据质量风险监控，保证数据的完整性、准确性、一致性；通过数据及时性监控，保证数据的及时性。

* 完整性

完整性是指数据的记录和信息是否完整、不缺失。数据的缺失包括数据记录的缺失（表行数异常）和记录中某字段信息的缺失（字段出现空值）。在本教程中，您需要重点关注数据的生产环节（MaxCompute外部表引用的表格存储数据）和加工环节（数据仓库CDM及ADS层）中表行数是否大于0、表行数波动是否正常以及字段是否出现空值或重复的情况。

* 准确性

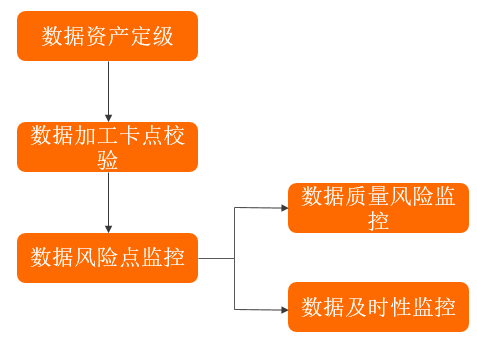
准确性是指数据记录中信息和数据是否准确、不存在错误或异常。例如，在本教程中，如果UV、PV数值小于0，则明显是错误数据。

* 一致性

对于不同的业务流程和节点，同一份数据必须保持一致性。例如表province字段中如果有**浙江**、**ZJ**两种表述，在您**group by province**时会出现两条记录。

* 及时性

及时性主要体现在最终ADS层的数据可以及时产出。为保证及时性，您需要确保整条数据加工链路上的每个环节都可以及时产出数据。本教程将利用DataWorks智能监控功能保证数据加工每个环节的及时性。

[[](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/9435932951/p96996.png)](https://static-aliyun-doc.oss-cn-hangzhou.aliyuncs.com/assets/img/zh-CN/9435932951/p96996.png)

数据质量管理的流程说明如下：

1. 分析业务场景，对数据流转链路上的整个依赖关系，进行资产定级。详情请参见[数据资产定级](https://help.aliyun.com/document_detail/123971.html?spm=a2c4g.11186623.2.11.af14415ckboNHR#concept-977415)。
2. 在业务系统的数据生成过程中进行卡点校验。详情请参见[离线数据加工卡点](https://help.aliyun.com/document_detail/123972.html?spm=a2c4g.11186623.2.12.af14415ckboNHR#concept-977416)。
3. 对数据风险点进行监控，包括数据的质量风险和及时性。详情请参见：
   * [数据质量风险监控](https://help.aliyun.com/document_detail/123975.html?spm=a2c4g.11186623.2.13.af14415ckboNHR#task-977419)
   * [数据及时性监控](https://help.aliyun.com/document_detail/123974.html?spm=a2c4g.11186623.2.14.af14415ckboNHR#task-977418)