伴随数据类型、数据来源的不断丰富以及数据量的飞速增长，企业面临数据质量问题的概率正在显著增加。而不论做了多少数据质量预防措施，实施多严格的数据质量过程控制，只要涉及人为干预，总会存在数据质量的问题。为了避免或降低数据质量对业务的影响，数据质量问题的发现既可以主动实施数据质量监控，也可以通过下游环节反馈问题来识别。如果主动发现、制定解决方案、采取行动，比被动采取补救措施效果更好，并且代价更小。

大数据应用必须建立在高质量可靠的数据之上才有意义。

今天，我们将结合
寄云NeuSeer工业数据管理平台的数据质量功能模块
，重点分享探讨下数据质量的设计思路与实践，助力工业企业实现高效数据治理。

NeuSeer工业数据管理
平台提供的数据质量功能主要是
正向主动监控数据质量，将数据中的问题数据识别出来，形成数据质量报告，从源头上进行彻底整改，提升数据质量支持运营与决策。

数据质量是指数据满足
用户使用要求的程度。寄云认为
数据质量
不是追求100%，主
要从用户角度满足业务需要即为好数据
。通常从完整性、及时性、准确性、一致性、唯一性、有效性六个维度来描述数据的质量。

- 完整性：
  指数据在创建、传递过程中无缺失和遗漏，包括实体完整、属性完整、记录完整和字段值完整四个方面。完整性是数据质量最基础的一项，例如员工工号不可为空。
- 及时性：
  指及时记录和传递相关数据，满足业务对信息获取的时间要求。数据交付要及时，抽取要及时，展现要及时。数据交付时间过长可能导致分析结论失去参考意义。
- 准确性：
  指真实、准确地记录原始数据，无虚假数据及信息。数据要准确反映其所建模的“真实世界”实体。例如员工的身份信息必须与身份证件上的信息保持一致。

- 一致性：
  指遵循统一的数据标准记录和传递数据和信息，主要体现在数据记录是否规范、数据是否符合逻辑。例如同一工号对应的不同系统中的员工姓名应该一致。
- 唯一性：
  指同一数据只能有唯一的标识符。体现在一个数据集中，一个实体只出现一次，并且每个唯一实体有一个键值且该键值只指向该实体。例如员工有且仅有一个有效工号。
- 有效性：
  指数据的值、格式和展现形式符合数据定义
  和业务定义的要求。例如员工的国籍必须是国家基础数据中定义的允许值。

数据质量的改进是一个持续的过程，通过关注业务对象、业务规则、业务过程、业务结果等数据，以数据清洗为目标，以业务需求为驱动，通过PDCA的循环，提升数据质量，达到数据质量满意的结果。

一、数据质量的设计逻辑

基于此，NeuSeer工业数据管理平台的数据质量模块整体设计逻辑：结合业务的数据需求，在规则模板中创建符合业务要求的模板，然后在质量规则中以质量模型为载体创建校验数据的规则，创建质量规则时选择平台内置的规则模板或者用户自定义的规则模板，在规则模板的引到下为规则指定参数，形成一个完成的SQL语句
，经过校验生效之后保存，这样，质量模型的一个规则就创建完成。基于质量监控目标按照同样方式创建质量规则，直至质量模型的规则全部创建完成。

在质量规则创建完成后关联质检任务，其中关联质检任务要求同一质量模型的质量规则的权重为100%，关联之后根据质量模型中的数据量再结合过往的处理经验预估一下处理时间，确定定时任务的执行频率，系统支持月、周、日等多种方式，设定好定时任务之后启用定时任务，待定时任务处理完数据之后，通过质量报告了解当前质量模型的数据质量情况，同时，通过导出质量报告的明细数据给到业务系统，从数据产生的源头进行彻底整改，保证数据修正一步到位。

二、数据质量的四大功能

NeuSeer工业数据管理

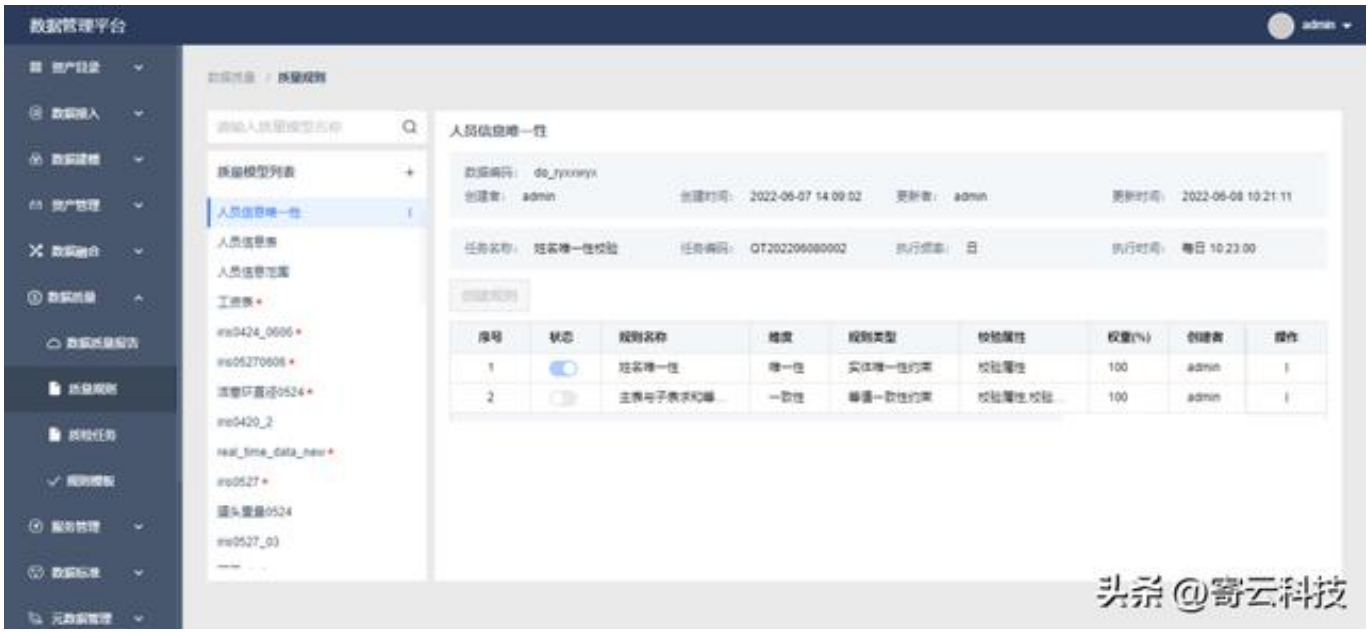
系统数据质量的四大功能，包括规则模板、质量规则、质检任务、质量报告。

1、规则模板

即将一定范围之内通用的规则进行抽象，并借助占位符预留参数，经过多次反复验证有效之后，以模板的形式固定并沉淀下来，进而在质量规则中复用的模板。一个模板，用的越多就越可靠，越可靠就越不用去复查，直接可用，实现了高效率（节约人的时间）和高可靠（节约人的时间，提升交付物/产品的性能）。

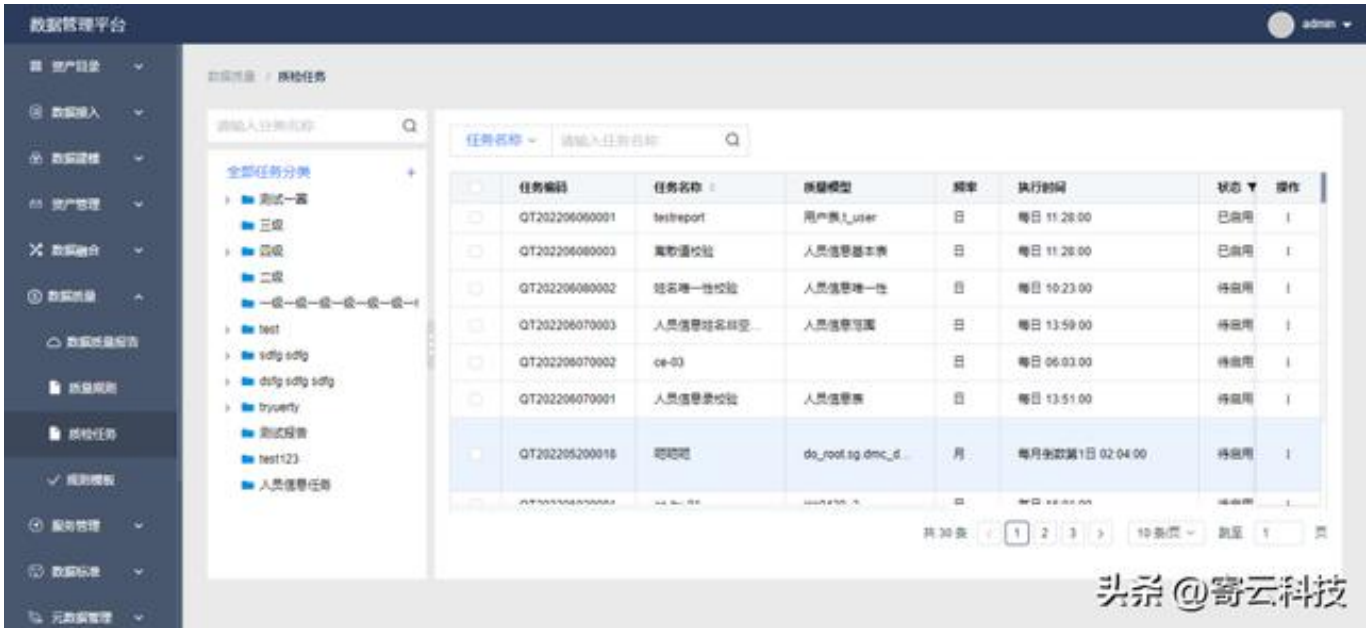
2、质量规则

判断数据是否符合数据质量要求的逻辑约束。质量规则的好坏直接影响监控数据质量的效果。通常，从描述质量的六大维度，同时兼顾影响范围来创建质量规则。创建质量规则是以治理模型为载体，结合其数据需求来创建检测数据质量的规则。通过定时任务有计划的触发质量规则对数据质量模型中存储的数据进行检测，发现引起不合格或不满意的因素，发现的问题以数据质量报告呈现出来，提供数据源头进行整改处理。



3、质检任务

依据用户制定的质量改善计划周期性触发质量规则运行来发现质量模型中的异常数据
的定时任务。质检任务主要提供对质量模型的增量数据实施数据质量监控，其目的就是致力于满足数据质量要求，消除或减少异常数据。质检任务可以在数据的生命周期
内的不同时点实施数据质量监控。为了有效管理质检任务，平台了维护质检任务分类的功能，实现从企业视角规划、管理质检任务，提升平台运维能力。



4、质量报告

以天为单位的
对质量模型实施数据质量监
控形成的数据统计报告
。主要包括三大块信息：基本统计信息、综合质量得分与质量评分历史趋势、以及质量评分明细。

质量报告基本统计信息

主要包括质检任务总数、质检模型总数、质检规则总数、数据总量、不合规数据总量以及不合规比率。

综合质量得分与质量评分历史趋势

1、综合质量得分概览：
通过仪表盘+得分的形式展示当前数据质量满足用户使用要求的程度。用户根据企业的实际情况设定数据质量的衡量标准，分为四个等级：差、中、良、优，采用百分制展示得分。

特别说明：（A）如果同一天同一个质量模型所属的质检任务执行多次，这个质检任务下的质量模型得分取最新一次的得分。（B）如果同一天同一个质量模型在不同任务下执行后的质量模型得分需要在不同任务下的分别计算得分。（C）如果筛选条件为质量模型，综合质量评分则 =
∑(质量模型在不同任务下得分) / 质量模型同一天所在质检任务个数

2、综合质量得分评分维度：
从描述数据质量的六大维度进行评估，向用户展示数据质量在这些维度下的得分。

3、质量评分历史趋势：以面积图
形式向用户展示数据监控的效果，通过横向比较更直观显示数据质量改进的效果。系统提供近7日、近30日、近1年三个时段，均以日的粒度展示，默认展示近7日。

质量评分明细

以细粒度向用户展示数据质量监控发现的问题，方便用户告知上游业务系统从源头上进行整改。同时，质量评分明细还支持导出，方便用户提交数据源头更

精准定位问题。

?

数据是未来企业核心价值与竞争力。聚焦工业数据管理，寄云NeuSeer工业数据管理平台以数据智能为核心理念，借助人工智能技术，轻松实现对企业多源、异构、海量数据的全生命周期管理
，通过构建数据标准规范体系，提升数据质量，让数据增值，实现数据流动下的
智能
制造，更
好实现工业企业客
户数字价值的高效呈现，为工业企业实现数字化转型奠定数据基础。

扫描下方二维码，欢迎体验试用—全新寄云NeuSeer工业互联网平台 3.0

寄云工业互联网平台 3.0
www.neuseer.com
欢迎访问 联系试用
头条 @寄云科技