云海佳信科技
MacroData Technology

MacroData Data Quality Platform

云海睿鉴数据质量平台

白皮书

北京云海佳信科技有限公司

目录

背景及定位

背景

大数据的时代，数据资源及其价值利用能力逐渐成影响企业核心竞争力的关键因素；大数据应用必须建立在质量可靠的数据之上，建立在低质量甚至错误数据之上的应用往往与其初衷南辕北辙、背道而驰。而今，数据质量已成为企业数据应用的瓶颈，高质量的数据可以决定数据应用的上限，而低质量的数据则必然拉低数据应用的下限。

数据质量是数据资源的关键指标，数据质量好坏直接影响数据应用成效。多年来，企事业单位的由于机构变动、职能调整、重视程度等因素，数据管理责权体系不完善，质量管理机制不健全，导致在数据集成整合过程中，各数据来源的数据质量参差不齐，数据二次利用价值无法充分体现。如何提升数据质量，为数据应用提供高质量数据支撑，是一个亟待解决问题。

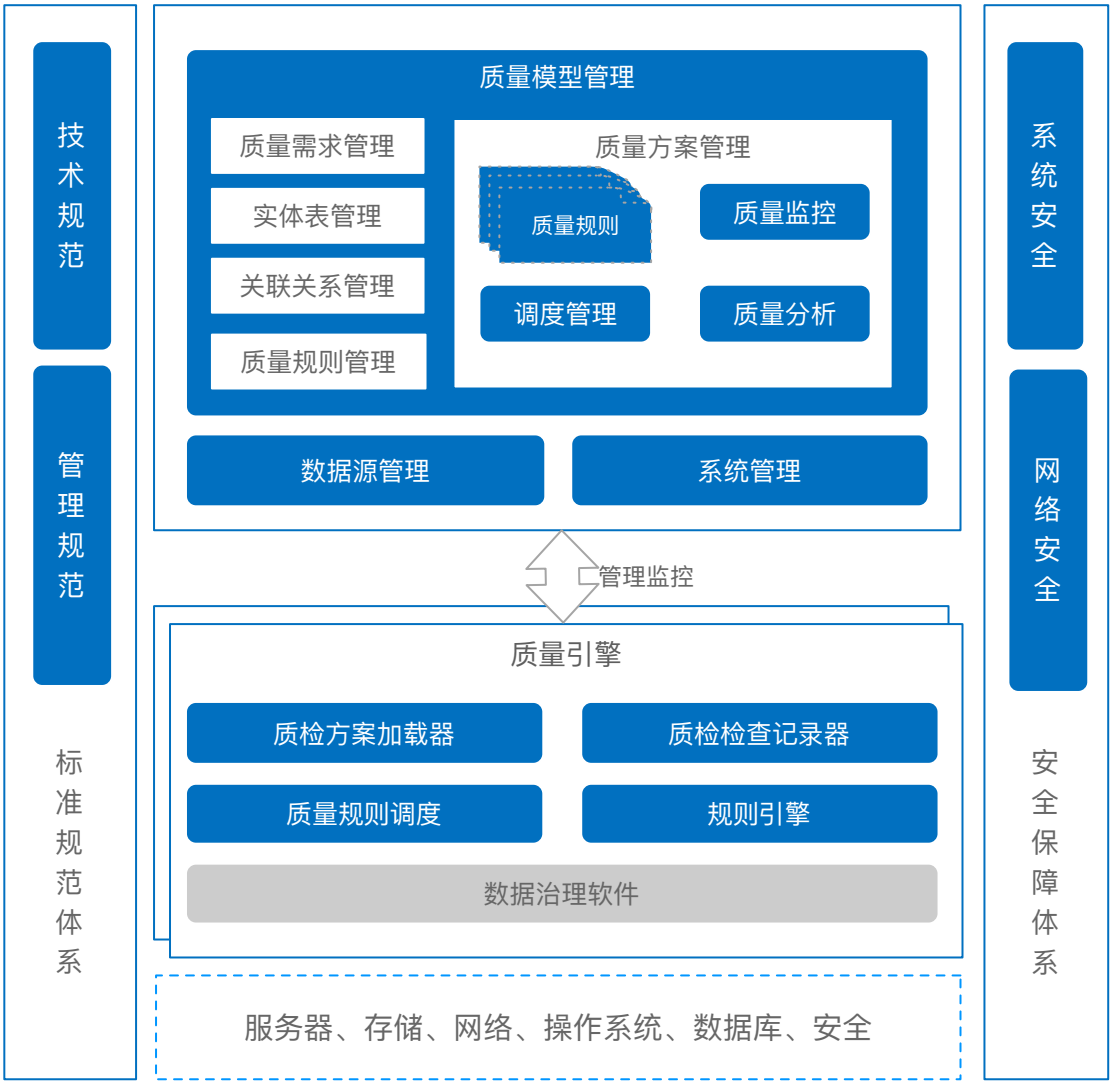
产品定位

睿鉴数据质量平台（简称云海睿鉴）是云海佳信公司在多年行业研究及应用实践基础上，围绕数据真实性、数据准确性、数据唯一性、数据完整性、数据一致性、数据关联性、数据及时性等一系列数据质量指标，自主开发的一套质量管理系统，通过有效的数据质量控制手段，进行数据的管理和控制，消除数据质量问题，提升企业数据变现的能力。

产品围绕上述数据质量衡量指标，建立质量规则，构建数据质量方案，定期进行数据质量核查、度量、监控、预警等一系列管理活动，并生成质量报告，提出质量改进意见，帮助信息化部门改进数据质量，提升数据应用价值。

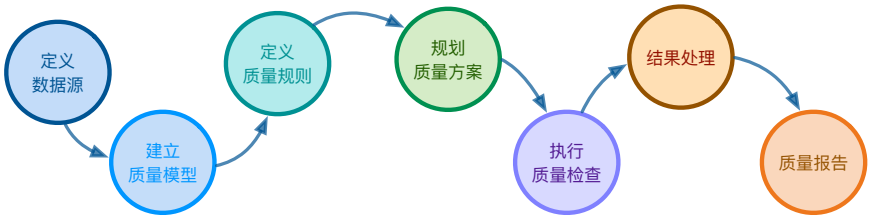
产品概览

系统架构

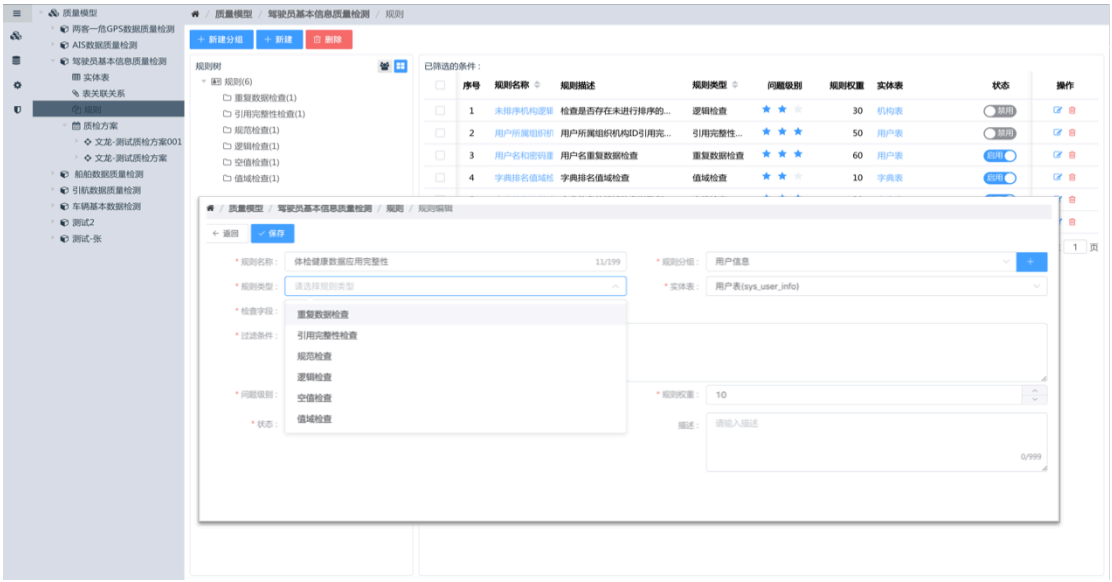


产品由质量管理中心和质量管理引擎两部分组成；

质量管理中心



提供基于浏览器的图形化人机交互界面，通过与质量管理引擎交互完成数据质量管理配置监控，包括数据源定义、建立质量模型、质量规则定义、质检方案规划、质量任务调度和监控分析以及质量分析报告。



质量管理引擎

质量管理引擎负责完成质量检测执行，并详细记录质量问题，生成数据质量报告；质量管理引擎采用分布式集群设计，支持横向扩展。

功能介绍

数据源管理

定义待数据质量稽查的数据源，可对数据源进行新建、修改、删除、检索等管理。

质量模型管理

云海睿鉴定义了一套数据质量模型、通过该数据质量模型可从数据结构、关联关系、数据标准规范等纬度对数据质量进行规范和质量检查定义。

本功能主要完成数据质量模型的新建、修改、删除等管理操作。

实体管理

本功能完成数据源待质检数据的元数据建模，包括增加、修改、删除、数据预览等功能。

质量规则管理

本功能负责完成对数据质量规则的定义，系统内置多种质检规则模板：重复数据检查、引用完整性检查、规范性检查、逻辑性检查、空值检查、值域检查等。用户可利用这些内嵌的质检规则模板快速定义满足业务需求的质检规则；并可对质检规则进行自定义分组，以方便管理。各质量规则可设置问题级别、评分权重等质量分析指标。

质量方案管理

质检方案是质量检查的最小调度单位，一个质检方案可包含多个上述质量规则，并可设置调度策略（手动、Crontab）。

本功能包括质量方案的新建、修改、删除、调度执行等。

质量监控

质量监控可查看对应数据质检方案的各个批次的调度执行状态和结果，包括检查开始时间、结束时间、是否成功、检测数据量、问题数据量、综合评分等。

规则监控

规则监控可查看对应数据质检方案下各个质检规则的各个批次的调度执行状态和结果，包括检查开始时间、结束时间、检查目标实体、检测数据量、问题数据量、评分等。

质量报告

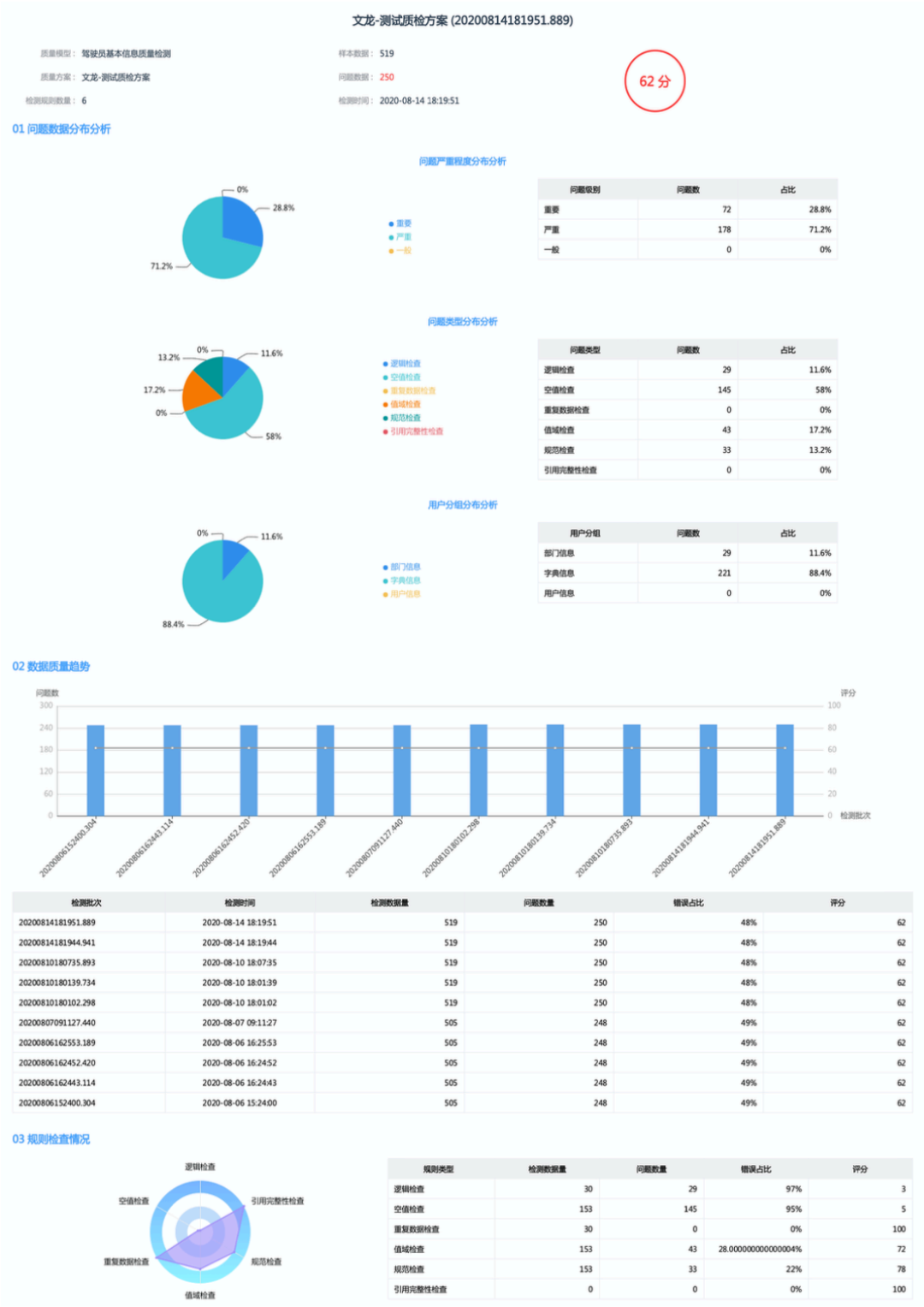
质量报告是对数据检测结果的一种多维度分析展现，并根据各个质量规则的权重进行质量综合评分。

质量分析纬度

- 问题数据分析：问题严重程度分布分析、问题类型分布分析、用户自定义分组分布分析
- 数据质量趋势分析：对同一质量方案多个检测批次结果进行趋势分析。
- 规则检查情况多维分析：按照质量规则类型多维统计分析

质检分析报告

云海睿鉴提供对质量稽查结果的可视化分析报告，并可以 PDF 形式导出下载。



系统管理

通过本功能可进行系统级的管理操作，如：机构管理、用户管理、角色管理、安全审计等。

产品特点

跨平台支持

数据集成平台基于 Java 技术开发，受益于 Java 的跨平台特性，数据集成平台支持跨平台，可部署运行在 Windows、Linux、MacOS、AIX 等操作系统中。

丰富便捷的质量规则

系统提供了引用完整性检查、规范性检查、逻辑性检查、重复性检查、值域、空值检查等多种数据检查规则定义方法，为用户提供全方位数据质检监测保障。

质量规则的定义完全可视化实现、灵编码，轻松完成所有规则的定义。

定制化质量评估体系

用户可自定义质量规则的问题级别、评分权重，满足不同领域、不同场景对不同质量检查纬度的侧重，从而给出更切合实际的质量评估报告。

全方位监控管理

系统支持数据质量检查方案的定义和管理，包括检查范围、检查时间、检查规则、评分规则等。调度上支持手工调度和自动调度。

调度结果实时监控，从质检方案到质检规则、实时掌握质量检查执行情况。

完善的质量分析报告

对每次质量检测都可以从问题严重程度、规则类型分布、质量趋势等多个纬度进行质量分析，并结合质量规则的权重进行综合评分，给出质量分析报告，

有效帮助用户提升数据质量。

高扩展、易集成

系统提供丰富的扩展点：质量规则、评价算法、脚本嵌入等层面都充分考虑二次开发的扩展需求，以满足特定场景需求的二次开发需要。

系统从配置、调度、监控各个环节都提供 Rest 接口 ,方便第三方深度集成。