Datacleaner功能扩展

设计文档

目录

[1 datacleaner简介 1](#_Toc68859263)

[1.1 什么是数据质量 1](#_Toc68859264)

[1.2 datacleaner 1](#_Toc68859265)

[2 客户端质控组件设计 2](#_Toc68859266)

[2.1 主附表数据关联性 2](#_Toc68859267)

[2.2 数据连续性 2](#_Toc68859268)

[3 web端首页设计 3](#_Toc68859269)

[3.1 首页原型草图 3](#_Toc68859270)

[3.2 事件流 3](#_Toc68859271)

[3.3 约束条件 3](#_Toc68859272)

# datacleaner简介

## 什么是数据质量

数据质量需要保证的便是数据的使用者的需求。

狭义上讲，衡量数据质量至少应该包含以下几个方面：

1. 完整性 Completeness：完整性用于度量哪些数据丢失了或者哪些数据不可用。

2. 规范性 Conformity：规范性用于度量哪些数据未按统一格式存储。

3. 一致性 Consistency：一致性用于度量哪些数据的值在信息含义上是冲突的。

4. 准确性 Accuracy：准确性用于度量哪些数据和信息是不正确的，或者数据是超期的。

5. 唯一性 Uniqueness：唯一性用于度量哪些数据是重复数据或者数据的哪些属性是重复的。

6. 关联性 Integration：关联性用于度量哪些关联的数据缺失或者未建立索引。

7．连续性：用于检验数据的产生或者交换过程中是否有中断。

## datacleaner

DataCleaner是一个数据质量分析,比较,验证和监督的软件.DataCleaner包括一个独立的图形用户界面分析,比较和验证,并进行监测web应用。

其中客户端程序中包含各类质控组件，涵盖数据治理、分析、改写等功能；按分类包含转换、改善、分析、读写类共计百余种各种功能的组件，结合数据质控在实际业务中的使用需求，目前组件中对数据的关联性检查存在缺陷、缺失数据连续性检查功能。在此情况下，需要对质控组件功能进行功能拓展，全面优化提升关联性检查组件，新增联系行检查组件。

web端作为质控任务调度配置存在，缺乏一个总的监控页面，很多需要多步操作才能看到，因此在首页增加任务监控总览十分有必要。

# 客户端质控组件设计

客户端组件开发原则上使用程序现有框架和类。

## 主附表数据关联性

输入数据：主表字段、附表路径、字段关联关系、主表时间字段、时间范围

输出数据：主表总数据量、附表总数据量、主表时间范围内总数据量、主表成功数据量、主表失败数据量。

输出样式：表格、饼图

## 数据连续性

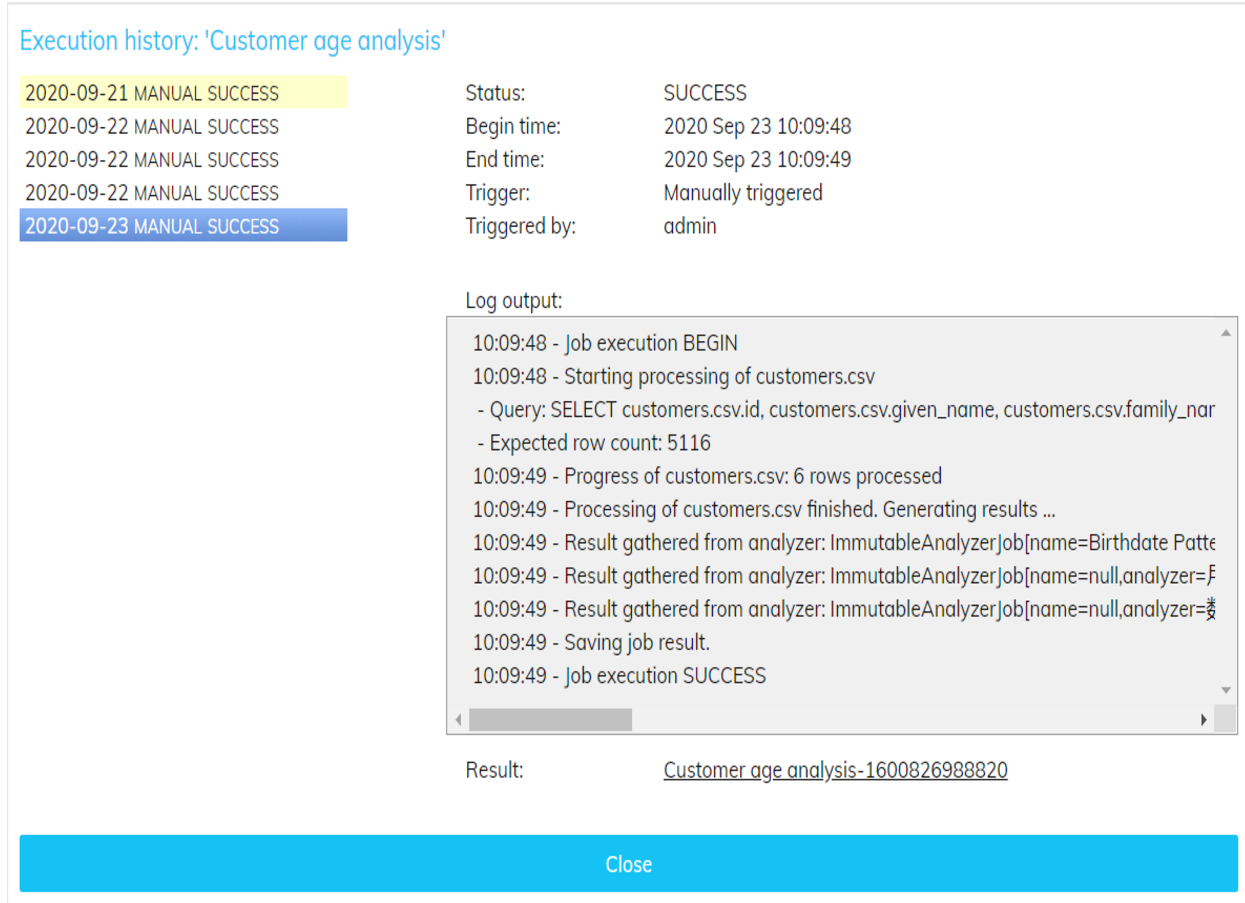
输入数据：表日期字段（可多选）、时间范围

输出数据：时间范围内数据总量、有数据的天数、没有数据的天数

输出样式：表格

# web端首页设计

## 首页原型草图



## 事件流

1、点击首页，系统加载首页页面，页面展示数据项包括：总体情况（任务总数、数据源、时间线分析），今日运行（运行任务、运行结果），任务执行情况（任务名称、最近执行状态、最近运行时间、最近结束时间、耗时）

## 约束条件

1、今日运行>>运行任务：统计当日0:00:00到当前时间之间执行过的任务总数（按任务执行开始时间统计），如果一个任务在此时间段内执行过多次也只算一个任务。

2、今日运行>>运行结果：统计当日0:00:00到当前时间之间由于任务执行而产生的运行结果（按结果生成时间统计）。

3、进入系统后首页为默认页面。