

数据质量监控

51信用卡-曹传宇



分享提纲

- ◆ 数据质量监控的必要性
- ◆ 主要功能特性
- ◆ 设计和实现概述
- ◆ 遇到的问题

数据质量监控的必要性

- ◆ 数据问题的严重性
- ◆ 问题的滞后性

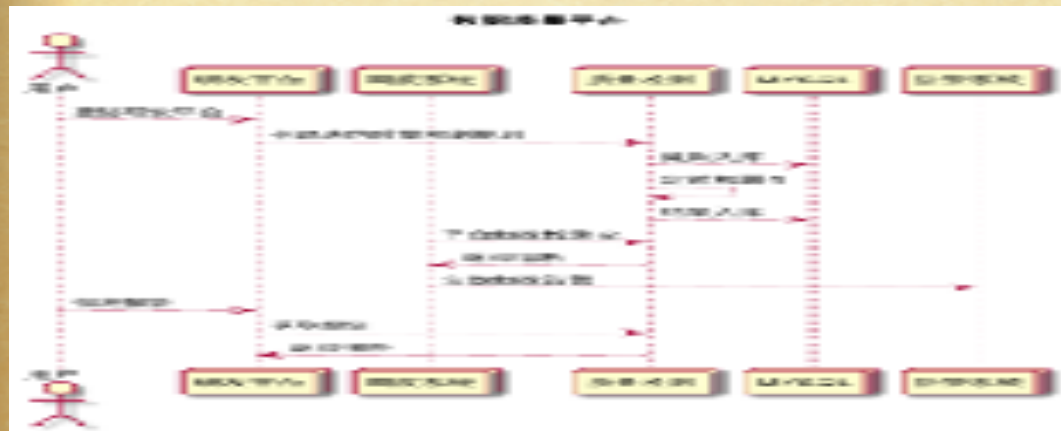
几个基本的名称

- ◆ Hadoop/hive平台
- ◆ 在线->离线->在线
- ◆ 调度系统、任务、表
- ◆ 数据仓库的分层

主要功能特性

- ◆ 支持表纬度、字段纬度的检测
- ◆ 表级别支持：行数与固定值和波动率的检测
- ◆ 字段级别支持：非空、主键、枚举、汇总、列和列之间的计算关系
- ◆ 比较类型支持：>,<,>=,<=,=,<>
- ◆ 比较范围支持：固定值、1/7/30天的环比
- ◆ 支持强、弱两种规则，强规则失败后中断下游任务
- ◆ 检测失败进行告警

设计-系统交互



实现-基本流程

- ◆ 创建原子规则/实体规则
- ◆ 运行检测
 - ◆ 根据实体生成sql执行语句
 - ◆ 提交任务到hive/yarn
 - ◆ 异步获取结果
 - ◆ 结果与预期值比较
 - ◆ 检测结论
 - ◆ 告知调度系统，进行下一步动作

存储

- ◆ 原子规则
- ◆ 实体规则
- ◆ 检测结果
- ◆ 历史数据

Field

ruleId

ruleName

ruleDesc

type

expression

source

owner

Field

entityId

entityName

ruleId

plusId

jobId

projectId

projectName

database

tableName

columnName

type

operator

weight

threshold

day

create

modify

owner

update

status

do

valid

卡片

实现-外部依赖

- ◆ Hive/yarn
- ◆ 元仓统计的表count数据
- ◆ 表的字段信息
- ◆ 调度做依赖触发
- ◆ 告警服务

前后端

- ◆ 服务端 base框架 + MyBatis + 配置服务
- ◆ 前端vue全家桶

遇到的问题

- ◆ 不同的分区格式
- ◆ 调度触发还是定时触发
- ◆ 究竟要不要阻断下游任务
- ◆ yarn资源问题