聊聊大数据质量监控的那些事

原创：诸葛子房 诸葛子房的博客 5天前

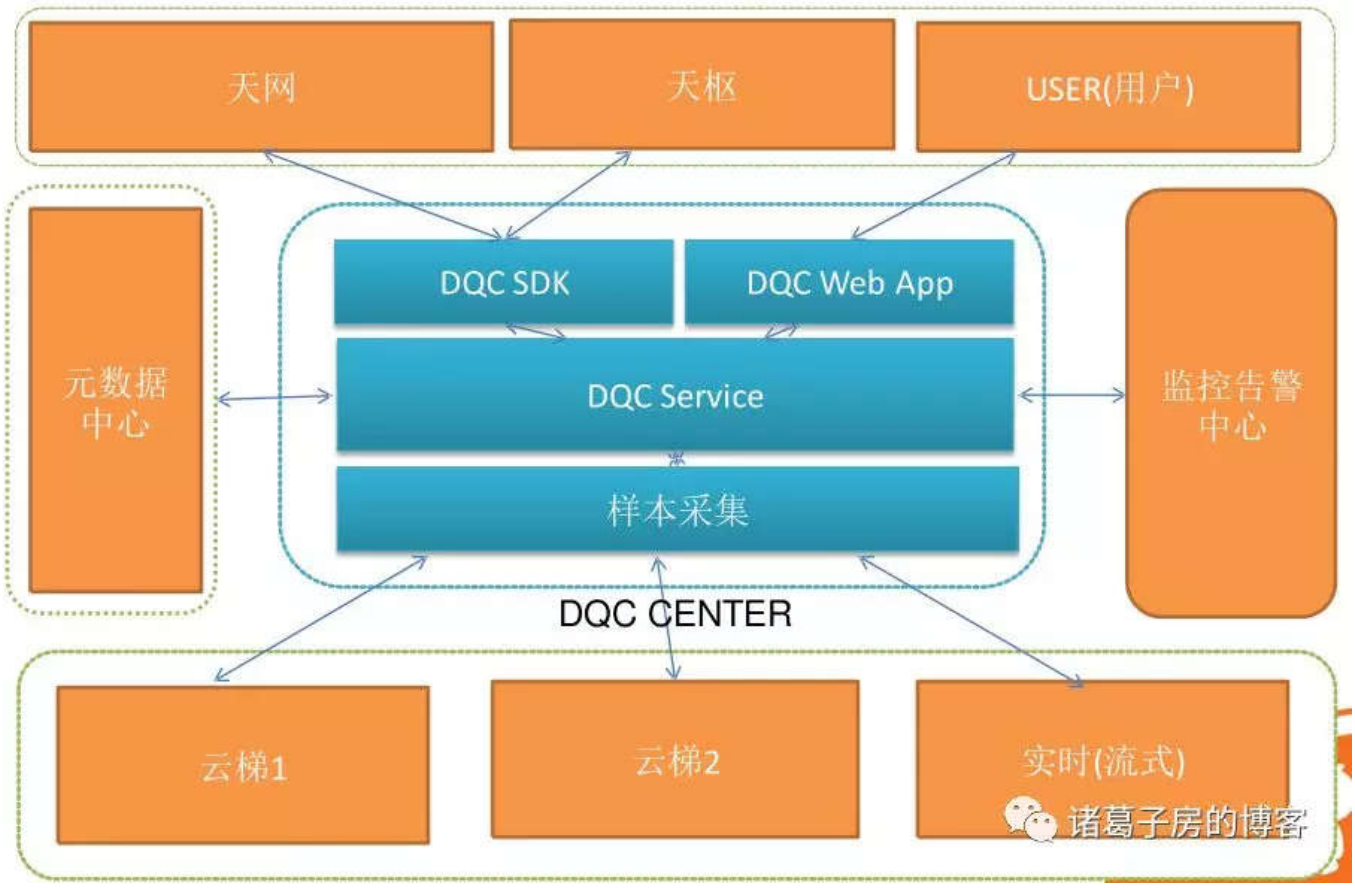
在这个信息化时代，你用手机打开微信聊天、打开京东app浏览商品、访问百度搜索、甚至某些app给你推送的信息流等等，数据无时无刻不在产生。

数据，已经成为互联网企业非常依赖的新型重要资产。数据质量的好坏直接关系到信息的精准度，也影响到企业的生存和竞争力。Michael Hammer（《Reengineering the Corporation》一书的作者）曾说过，看起来不起眼的数据质量问题，实际上是拆散业务流程的重要标志。 数据质量管理是测度、提高和验证质量，以及整合组织数据的方法等一套处理准则，而体量大、速度快和多样性的特点，决定了大数据质量所需的处理，有别于传统信息治理计划的质量管理方式。

本文主要探讨了一二线互联网公司数据质量监控平台。

一、Data Quality Center(DQC阿里巴巴数据质量监控平台)

1.系统架构图

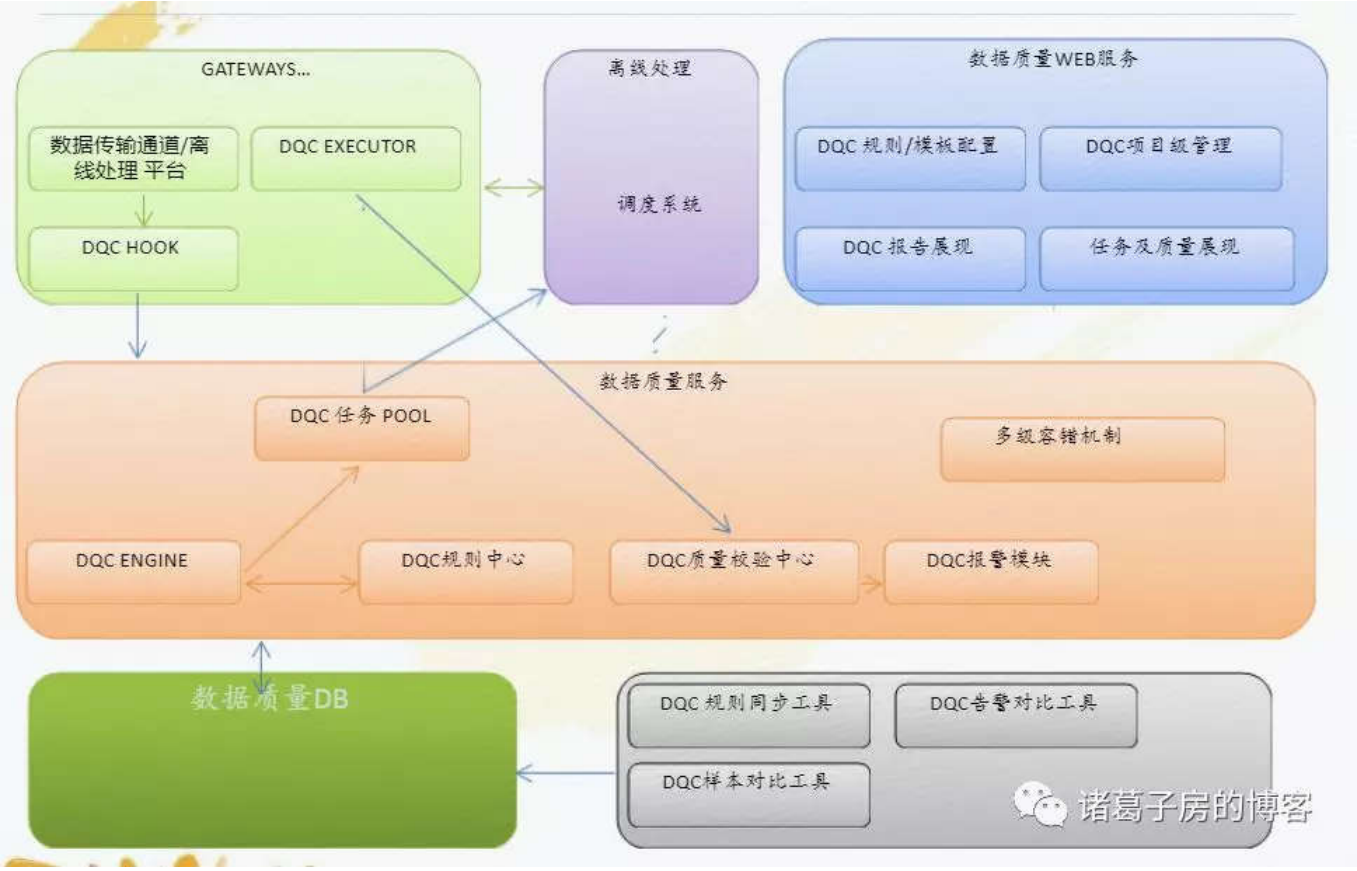


(1)基于线上业务数据，进行数据采集

(2)基于监控规则库，执行SQL任务，进行计算处理

(3)基于用户规则，发送数据报警(短信、邮件)

2.系统流程图



(1)用户进行规则配置

(2)通过定时的调度任务触发检查任务执行

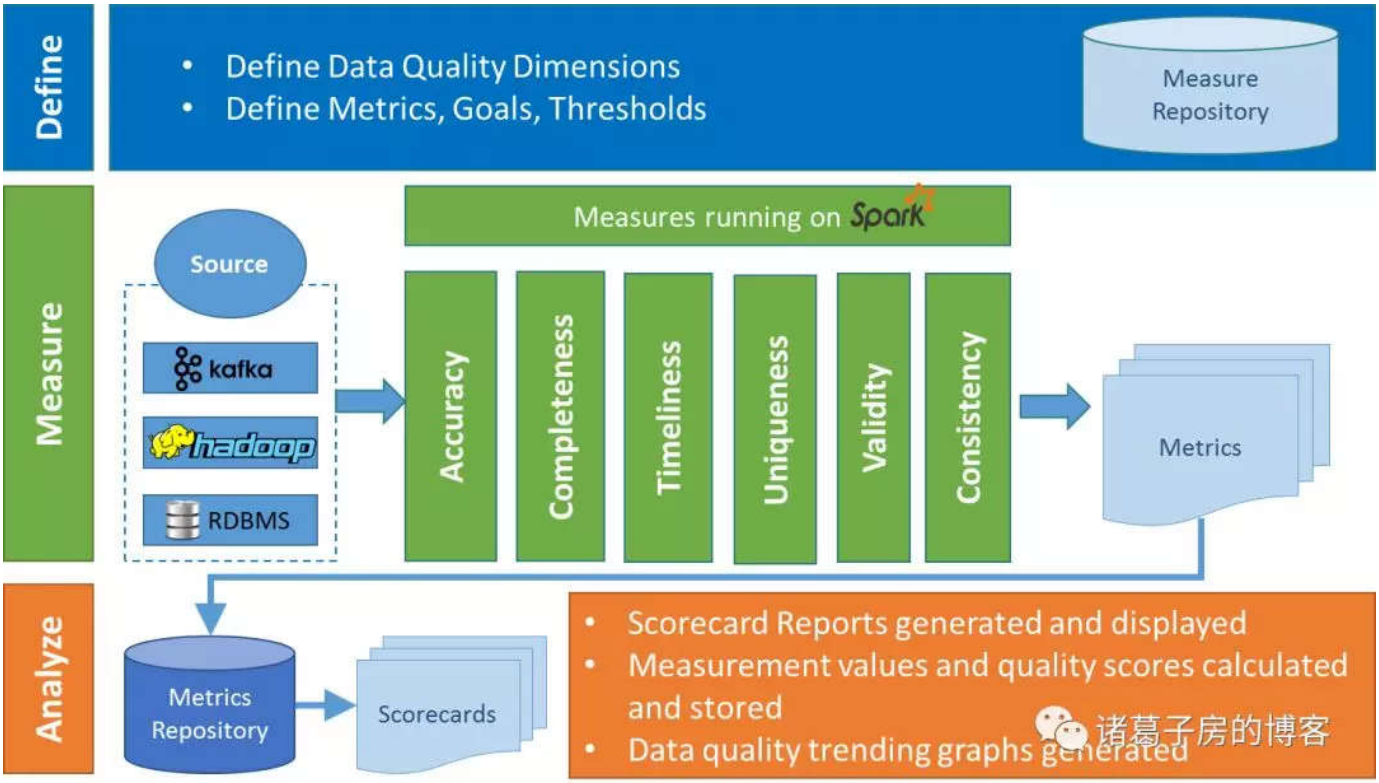
(3)基于任务配置，获取样本数据

(4)基于计算返回检验结果

(5)调度根据检验结果，决定是否阻断干预(强依赖、弱依赖)

二、 Apache Griffin(Ebay开源数据质量监控平台)

1.系统架构



(1)从准确性、完整性、时效性、唯一性等多个维度进行监控

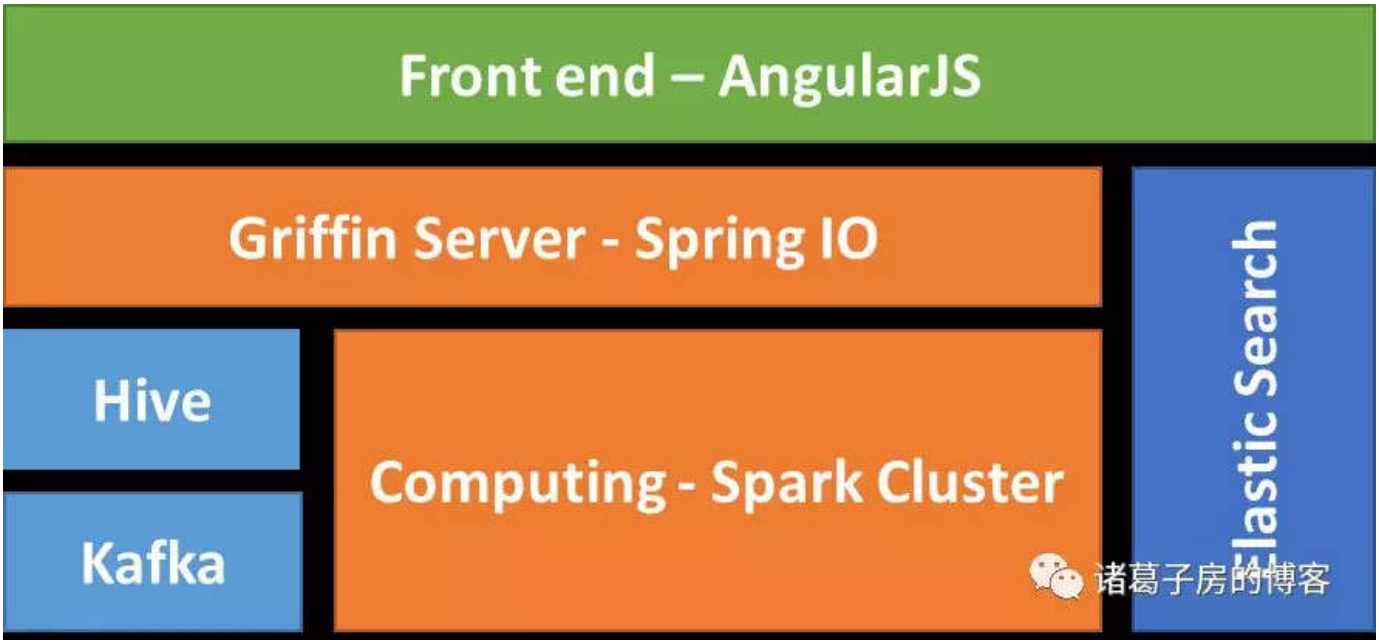
(2)计算结果存储至ES、HDFS

(3)计算结果metrics展示

(4)支持实时和离线

(5)优势：开源

2.系统技术路线

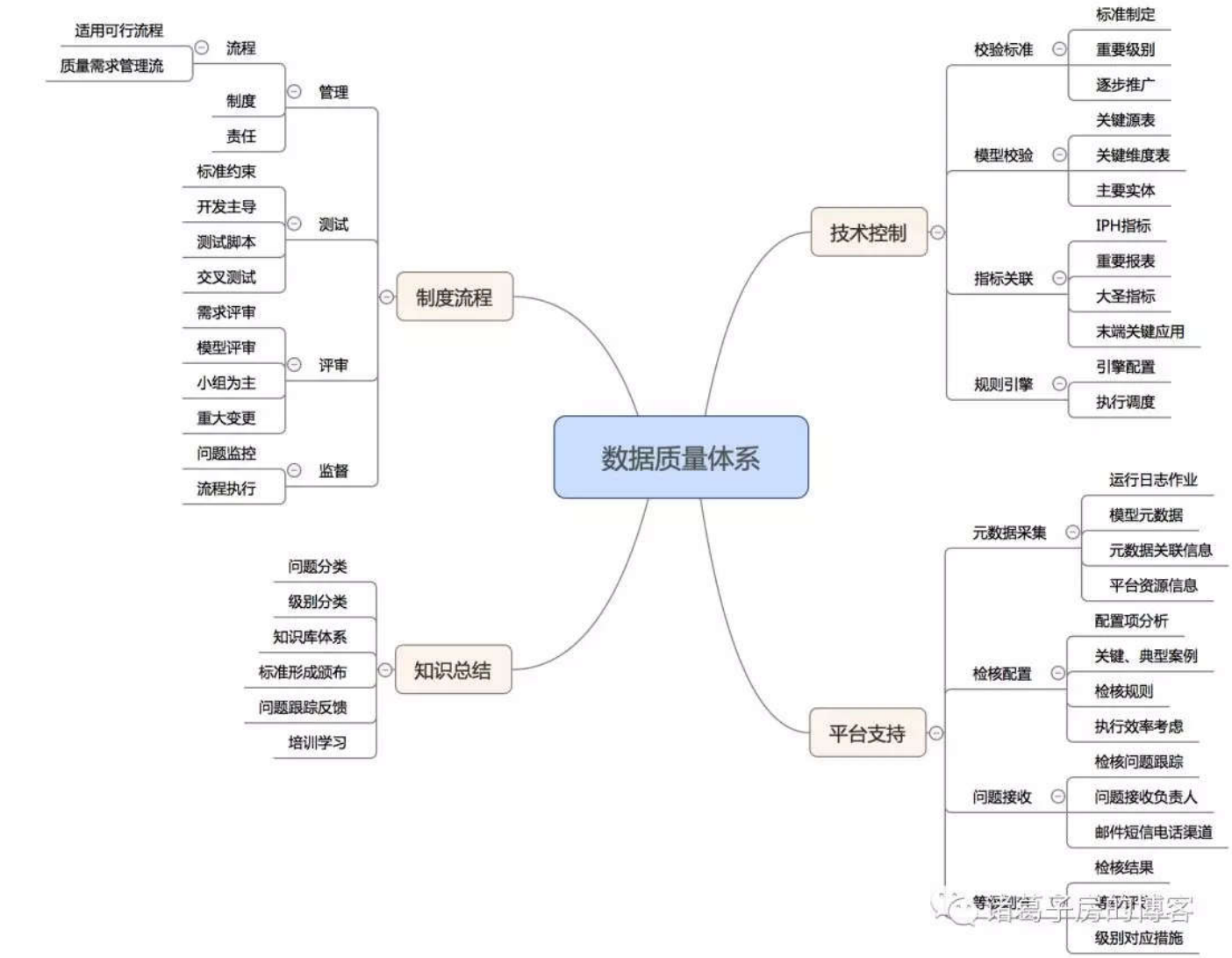


3.metrics展示

三、 DataMan(美团点评数据质量监控平台)

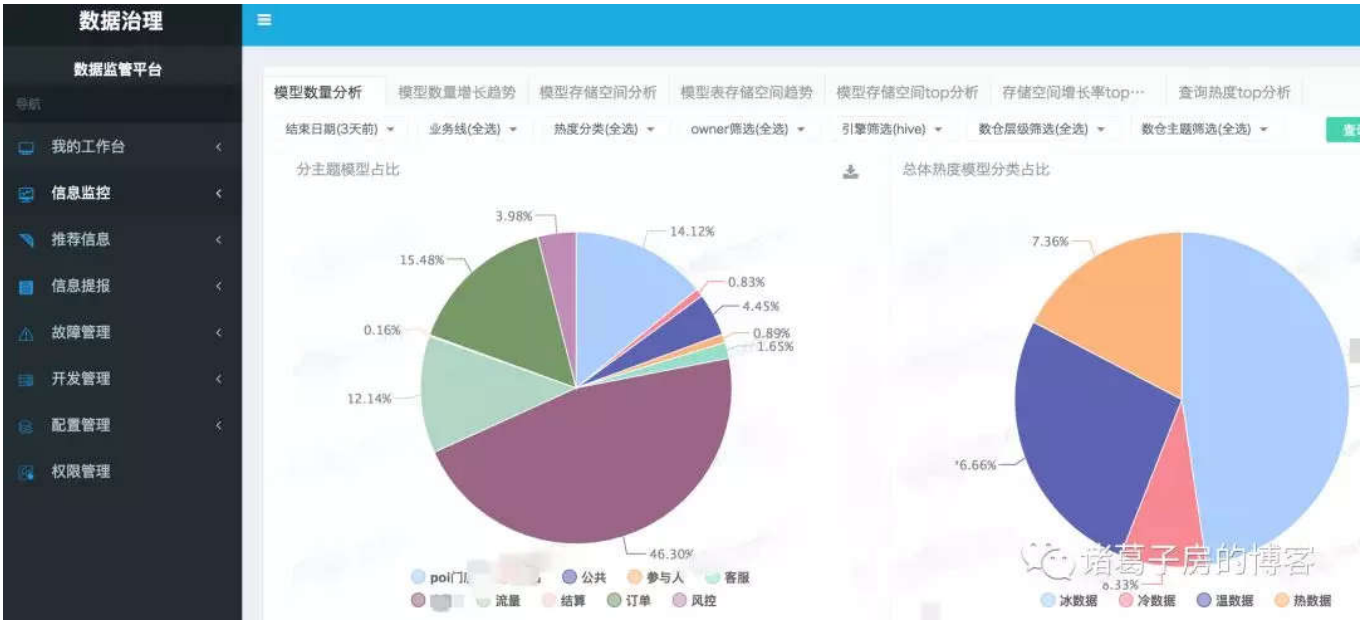
1.系统架构

DataMan系统建设总体方案基于美团的大数据技术平台。自底向上包括：检测数据采集、质量集市处理层；质量规则引擎模型存储层；系统功能层及系统应用展示层等。整个数据质量检核点基于技术性、业务性检测，形成完整的数据质量报告与问题跟踪机制，创建质量知识库，确保数据质量的完整性（Completeness）、正确性（Correctness）、当前性（Currency）、一致性（Consistency）。


数据质量体系
制度流程
管理
流程
适用可行流程
质量需求管理流
制度
责任
测试
标准约束
开发主导
测试脚本
交叉测试
评审
需求评审
模型评审
小组为主
重大变更
监督
问题监控
流程执行
技术控制
校验标准
标准制定
重要级别
逐步推广
模型校验
关键源表
关键维度表
主要实体
指标关联
IPH指标
重要报表
大圣指标
末端关键应用
规则引擎
引擎配置
执行调度
知识总结
问题分类
级别分类
知识库体系
标准形成颁布
问题跟踪反馈
培训学习
平台支持
元数据采集
运行日志作业
模型元数据
元数据关联信息
平台资源信息
检核配置
配置项分析
关键、典型案例
检核规则
执行效率考虑
问题接收
检核问题跟踪
问题接收负责人
邮件短信电话渠道
检核结果
级别对应措施
诸葛子房的博客


系统展示层
质量监控 ……
推荐优化 ……
质量分析 ……
日志分析 ……
信息提报 ……
故障跟踪 ……
测量定级 ……
系统功能层
配置管理
过程监控
问题跟踪
实时同步
知识库创建
运行质量分析
……基线分析
……耗时性能
……依赖分析
……机器学习
模型监控
……热度资源
……存储空间
……模型变动
业务监控
……业务指标
……数据度量
……业务波动
……实时动态
存储模型层
配置管理引擎
Dataman-DB
质量结果库
检测数据源
DM-DQ数据质量集市层
元数据信息
SLA信息采集
DQC质量中心
实时数
DW
诸葛子房的博客

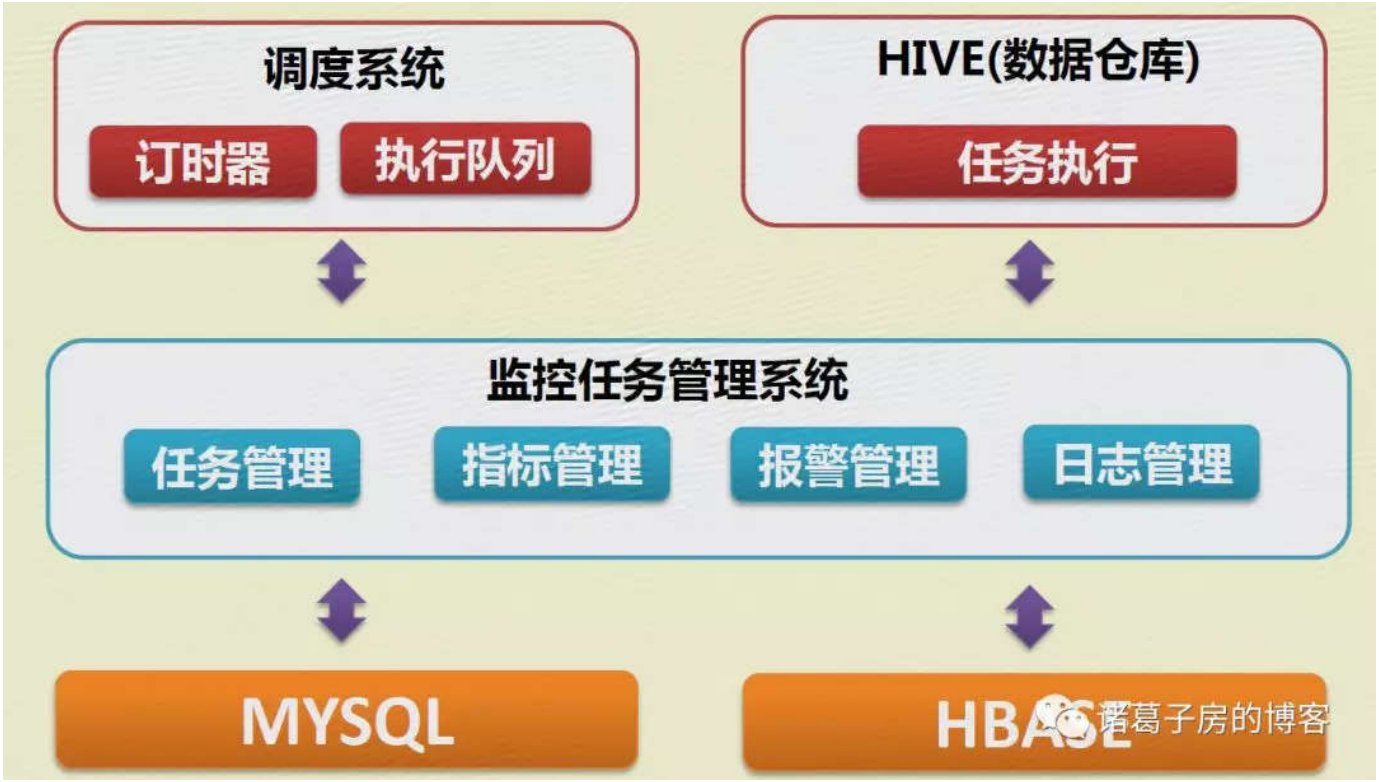
2.metric展示



四、 BDP(京东大数据质量监控平台)

京东数据质量监控系统（简称：数据质量系统） 是数据仓库、数据集市中表的数据变化进行监控。数据质量系统根据用户设定采集项配置、规则项配置、预警规则设置（枚举值），对用户指定的表进行每日定时数据采集、计算，并与历史数据或维表进行比对验证。最终将触发预警规则的异常数据以短信、邮件、App 等方式及时通知给用户。

1.系统架构图

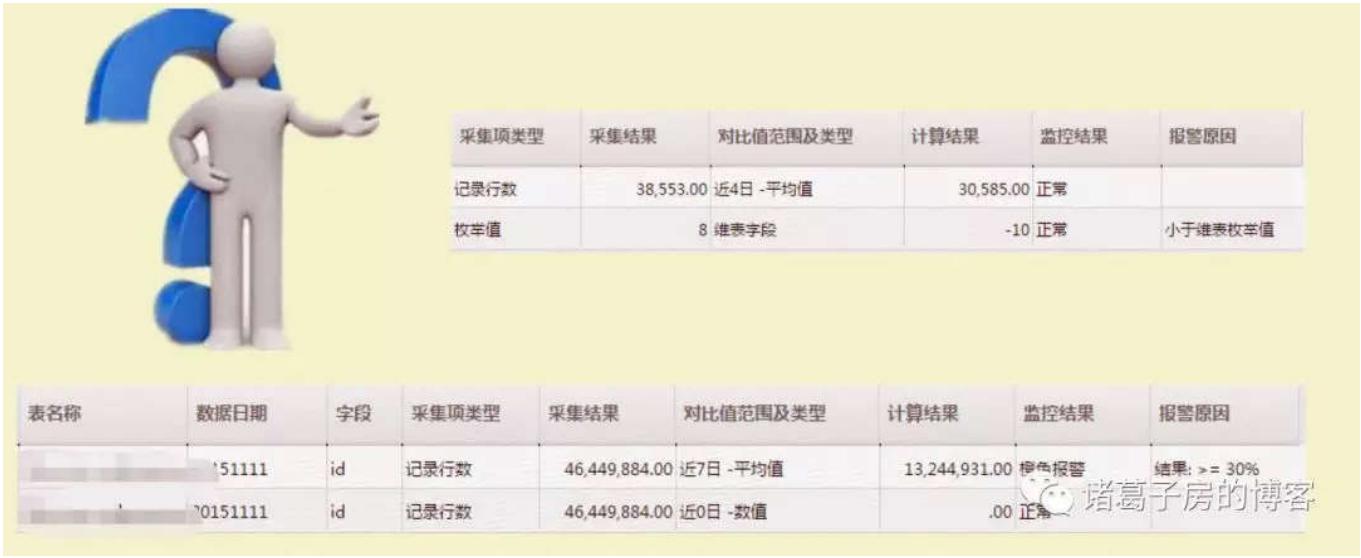


关系型数据库mysql和非关系型数据库HBase作为数据源，进行监控

2.系统流程图

(1)数据监控(2)运行日志(3)数据报警(4)规则配置

3.监控展示



上述主要分析了当前各大公司主要在使用或者开发的数据质量方面的平台，无论是离线数据监控还是实时数据监控，均有涉及。然而可能你的公司没有这么多的人力或者物力，但是由于数据量的增长，需要考虑数据治理方面的问题，就可以考虑采用开源的平台，在此基础上开发或者优化，毕竟站在前人的肩膀上才能看的更远，走的更快。

下面是我的公众号，如果想进入互联网行业，可协助帮忙内推，同时也欢迎对互联网行业感兴趣的同学一起交流学习。也想大家推荐一下hbase相关学习，大家工作学习遇到HBase技术问题，把问题发布到HBase技术社区论坛hbase.group，欢迎大家论坛上面提问留言讨论。想了解更多HBase技术关注HBase技术社区公众号:(hbasegroup)，非常欢迎大家积极投稿。

长按下面的二维码关注我的公众号

长按下面的二维码关注HBase技术社区公众号

参考资料：

1.美团点评技术专栏(DataMan-美团旅行数据质量监管平台实践)

2.开源中国(开源数据质量解决方案 Apache Griffin)