

ESENSOFT



小亿直播

数据治理之数据质量管理

汇报人：毛序

时间：2020.07.02



小亿直播

亿信华辰 ESENSOFT

目 录

CONTENTS

I

数据质量问题分析

II

数据整改方法建议

III

亿信数据质量管理

IV

数据质量成功案例

亿信华辰 ESENSOFT

I 数据质量问题分析





什么是数据质量管理?

数据质量管理 (Data Quality Management) , 是指对数据从计划、获取、存储、共享、维护、应用、消亡生命周期的每个阶段里可能引发的各类数据质量问题, 进行识别、度量、监控、预警等一系列管理活动, 并通过改善和提高组织的管理水平使得数据质量获得进一步提高。

数据质量管理是循环管理过程, 其终极目标是通过可靠的数据提升数据在使用中的价值, 并最终为企业赢得经济效益。

大数据时代数据的核心不是“大”，而在于“有价值”，而有价值的关键在于“质量”。

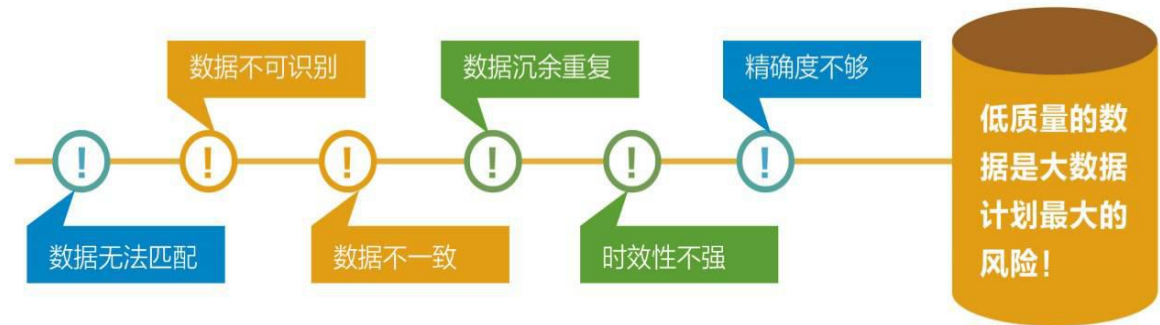
但现实是，数据往往存在**很多问题**。

45%

Collibra 一项调查发现，45% 的数据使用者称其报告基于的数据半数以上质量不够理想。

960万美元

Gartner 2016 年的一项研究发现，由于数据质量差，受访组织平均每年将损失 960 万美元。



数据要能发挥其价值，关键在于其数据的质量的高低，高质量的数据是一切数据应用的基础。

企业做数据质量管理的目标总结起来是要得到可信、可用的数据。

可信

- ✓ 实用性
- ✓ 准确性
- ✓ 及时性
- ✓ 完整性
- ✓ 有效性

可用

- ✓ 规范性
- ✓ 可读性



数据质量问题是数据治理和商业智能的头号障碍

技术因素

主要表现为由于具体数据处理的各技术环节异常造成的数据质量问题。

数据质量问题的产生环节主要包括数据创建、数据获取、数据传输、数据装载、数据使用、数据维护等方面的内容。

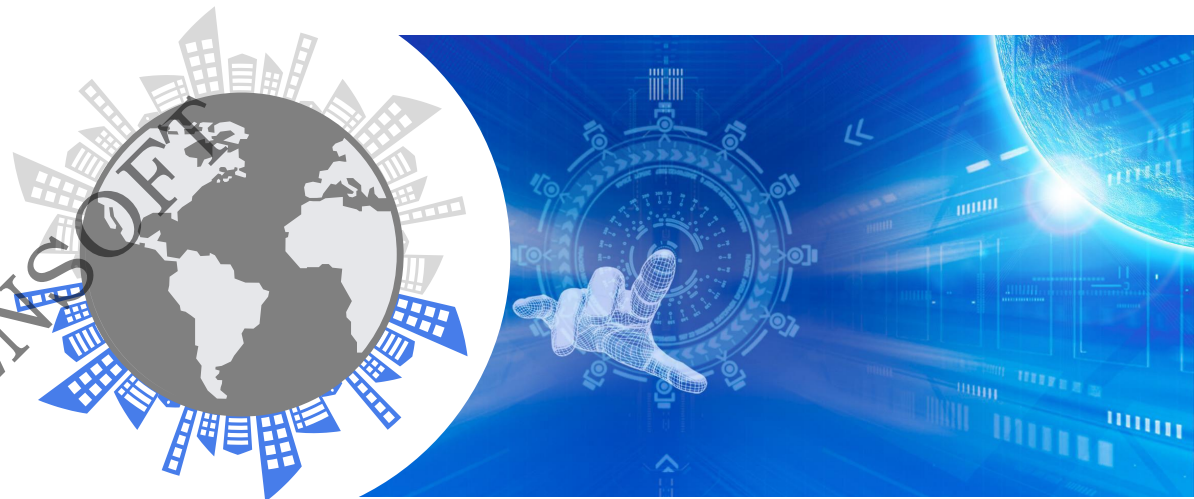
业务因素

主要表现为由于系统作业流程和人工操作流程设置不当而造成的数据质量问题，主要来源于系统数据的创建流程、传递流程、装载流程、使用流程、维护流程和稽核流程等各环节。

管理因素

主要表现为由于人员素质及管理机制方面的原因造成的数据质量问题。

如人员培训、人员管理、培训或者奖惩措施不当导致的管理缺失或者管理缺陷。





数据质量问题

- 源系统业务人员录入不规范，且系统逻辑校核不严谨，导致 **大量低质量、空值数据**。
- **源系统设计缺陷** 导致数据质量问题：跨表或跨系统有冗余字段但没有同步机制；逻辑设计问题导致特定情况下出现错误数据；关联性较强的流程设计缺乏有效衔接。
- 数据上报不及时。
- 历史数据迁移时操作不规范，导致数据不一致。
- 系统升级改造，导致历史数据错乱。
- 各源系统相同含义数据标准不一。
- 数据增值过程中，每次加工转换产生了新的质量问题。
-

II 数据质量整改建议

亿信华辰 ESENSOFT

1、建立统一的数据标准

有了统一的数据标准后，数据录入采集、加工处理等过程都将按照标准要求进行，数据混乱出错的情况将大为减少。需要整理发布的基础性数据标准包括业务元数据的标准和相关代码的标准。

2、提升操作人员数据应用水平

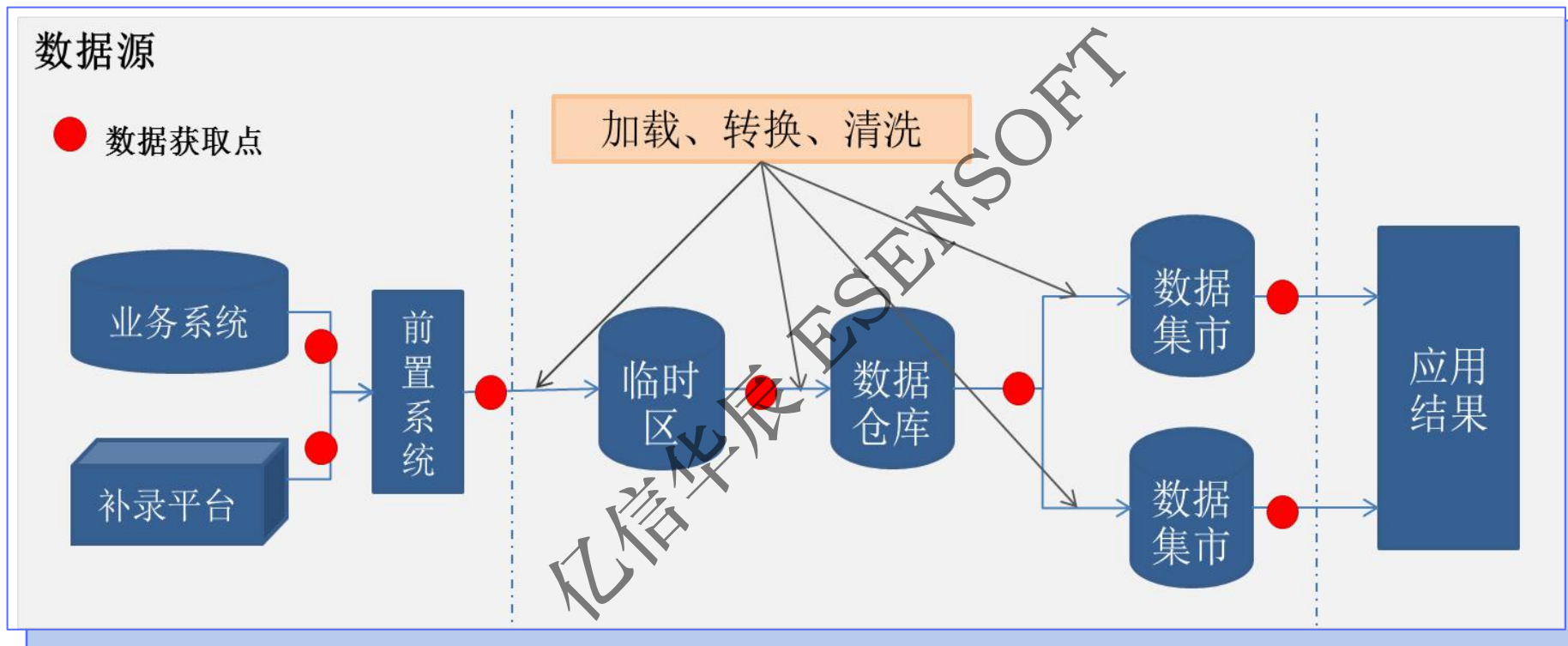
在培训模式上，建议根据岗位需要，分批分岗位培训，集中整理特殊环节、特殊业务处理的注意事项，尽快下发系统操作的帮助文档。实际应用中其实并不需要太多懂得全流程的通才，更需要熟悉自己岗位所涉及模块的专才，只要每个岗位都能熟练操作，那整个系统才能发挥应有作用。

3、完善数据质量管理机制

为了保证业务数据准确、及时、完整一致，必须要有专门的机构和岗位负责业务数据质量的监督管理。专门设置数据质量管理岗，或将数据管理的职能详细明确到相关岗位，牵头负责企业业务数据的质量管理，使得业务数据质量管理工作制度化、常规化，并量化成可考核的指标。

4、提供数据质量管理工具

提供数据质量管理工具，减低质量管理的门槛，减少沟通成本，及时发现问题并整改。



数据中心建设各个环节数据治理检查和监控

四、处理

- 将改进措施落实责任部门，完成改进
- 总结改进方式，准备下一个循环

三、检查

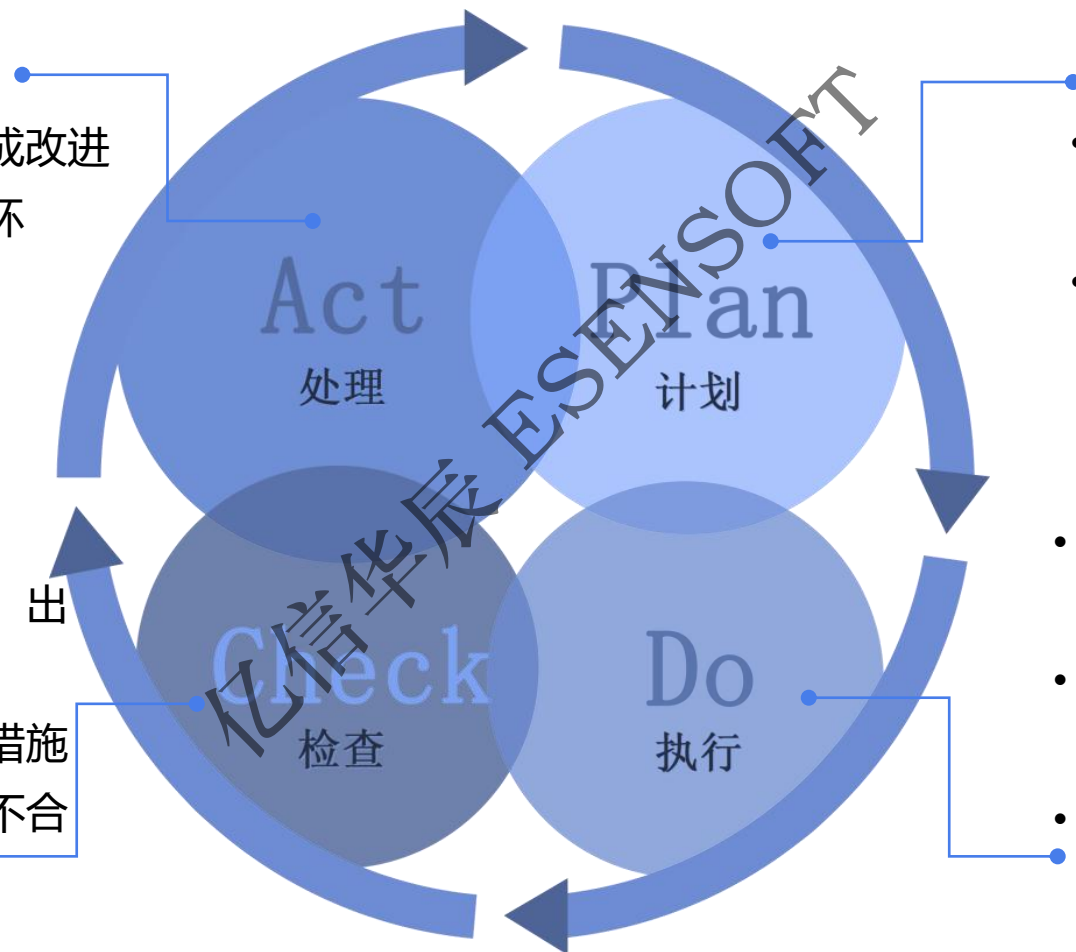
- 对处理后的数据进行再次质检，出具数据质量的报告
- 对比处理前后效果，总结改进措施
- 检查数据质量是否合格，分析不合格原因

一、计划

- 建立数据质量流程和规范，明确质量建设目标
- 明确质量管理组织和职责

二、执行

- 建立质量规则，分系统进行数据质量规则的定义
- 执行数据质量检查方案、出具数据质检报告
- 数据分类处理
 - a. 退回源系统
 - b. 数据总线完成数据清洗

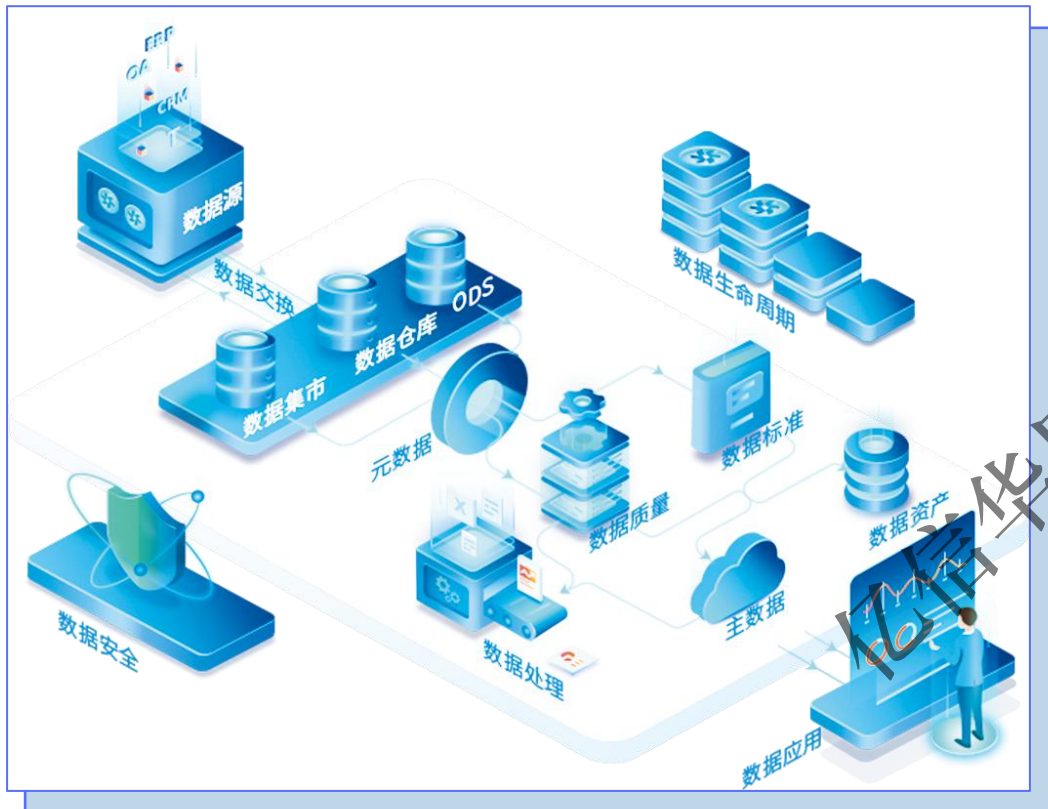


数据质量持续改进

III 亿信数据治理管理

亿信华辰 ESENSOFT





睿治数据治理平台是我公司**完全自主研发的、开创性的、一站式综合数据治理整体解决方案**。睿治是**全国领先，实现了数据治理场景全覆盖的突破性产品**，九大核心模块：元数据、数据标准、数据质量、主数据、数据资产、数据安全、数据交换、数据处理、数据生命周期等，以创新的方式保证了企业的业务数据在采集、汇总、转换、存储、应用整个过程中的完整性、准确性、一致性和时效性，**全面为客户量身打造符合自身特征的数据治理体系**。



图形化操作界面

“零” 编码，易操作、易阅读、易维护，适用于所有的业务人员和技术人员

一站式质检全过程



提供从标准定义、质量监控、绩效评估、质量分析、质量报告、重大问题及时告警、流程整改发起、系统管理等数据质量管理全过程的功能



全方位质检报告

多维度质量分析报表，内置丰富的统计分析报告及多种质检结果主题，同时支持自定义扩展；多角度质量绩效评分，支持用户定义评分依据和权重

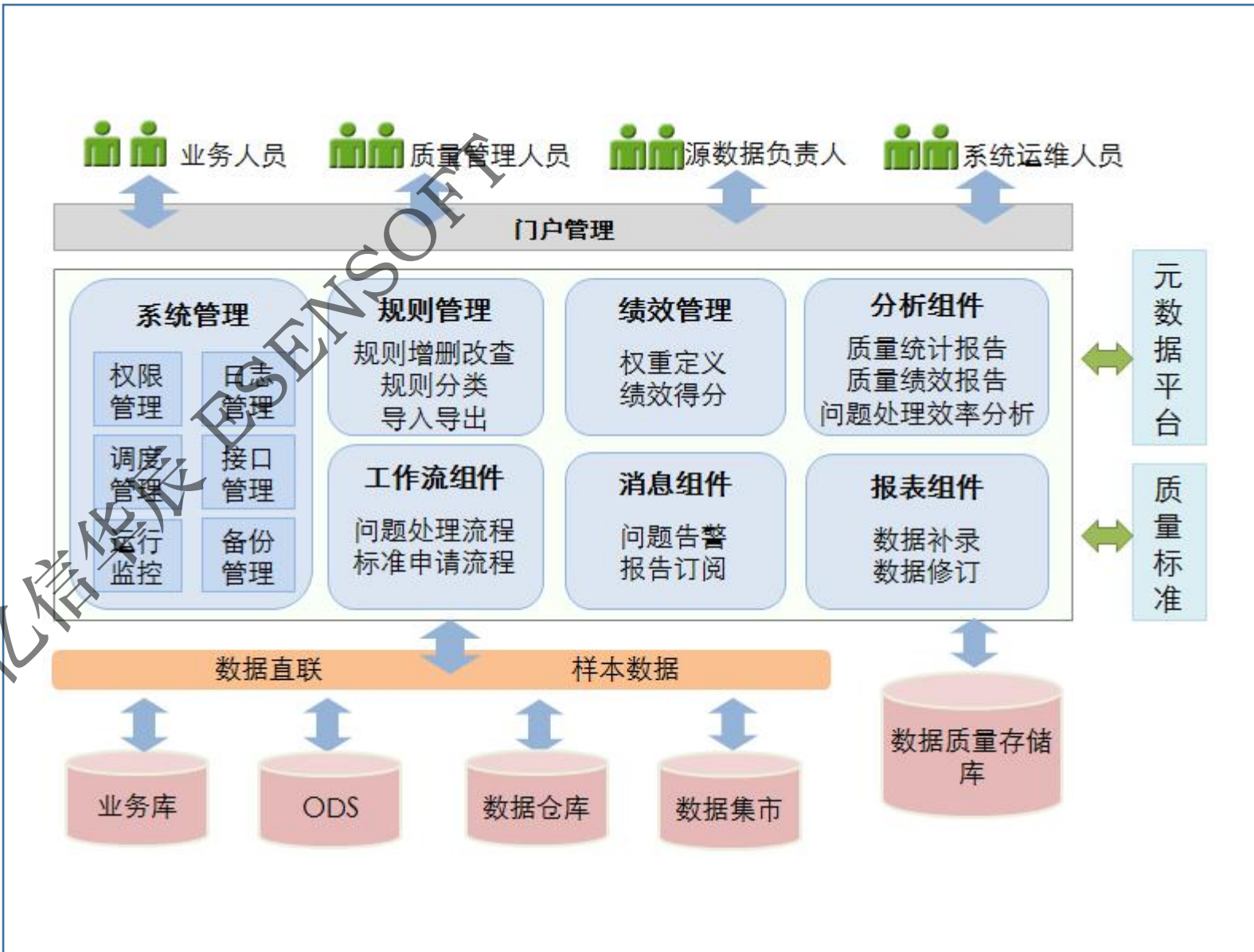
卓越的质检性能



数据质量管理平台具有出色的跨平台运行能力，兼容多种操作系统，支持所有符合JDBC2.0规范的数据库。系统的质量规则检查支持多线程并发执行，百万级数据20条规则的质量检查只需2分30秒即可完成

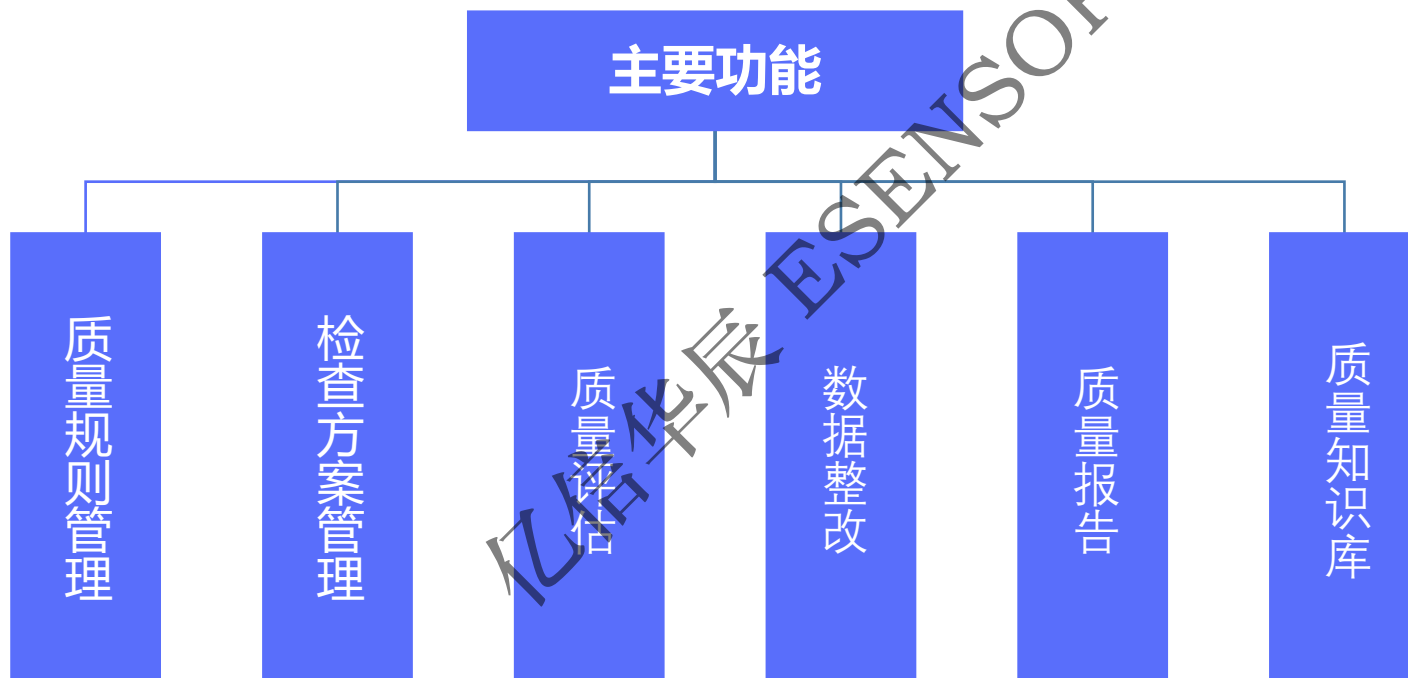
它具有以下特色

- **良好的交互体验**: 无需数据库工程师参与, 用户即可完成质检方案的各项操作。
- **全面的规则体系**: 支持13类规则评价方式, 全方位满足行业个性化的质检需求。
- **科学的绩效评估**: 内置绩效评估系统, 一键式触发, 快速得到评估结果。
- **丰富的质检报告**: 提供图文并茂多维度的数据质量分析及绩效情况展现。
- **完备的数据整改**: 自定义数据整改流程, 整改问题自动分发, 实时监控整改进度。
- **多样化的评分卡**: 自定义多套评分卡, 任意切换满足不同场景数据质量评分需求。
- **智能的问题修复**: 支持自定义智能修复策略, 实现数据质量问题的一键修复。



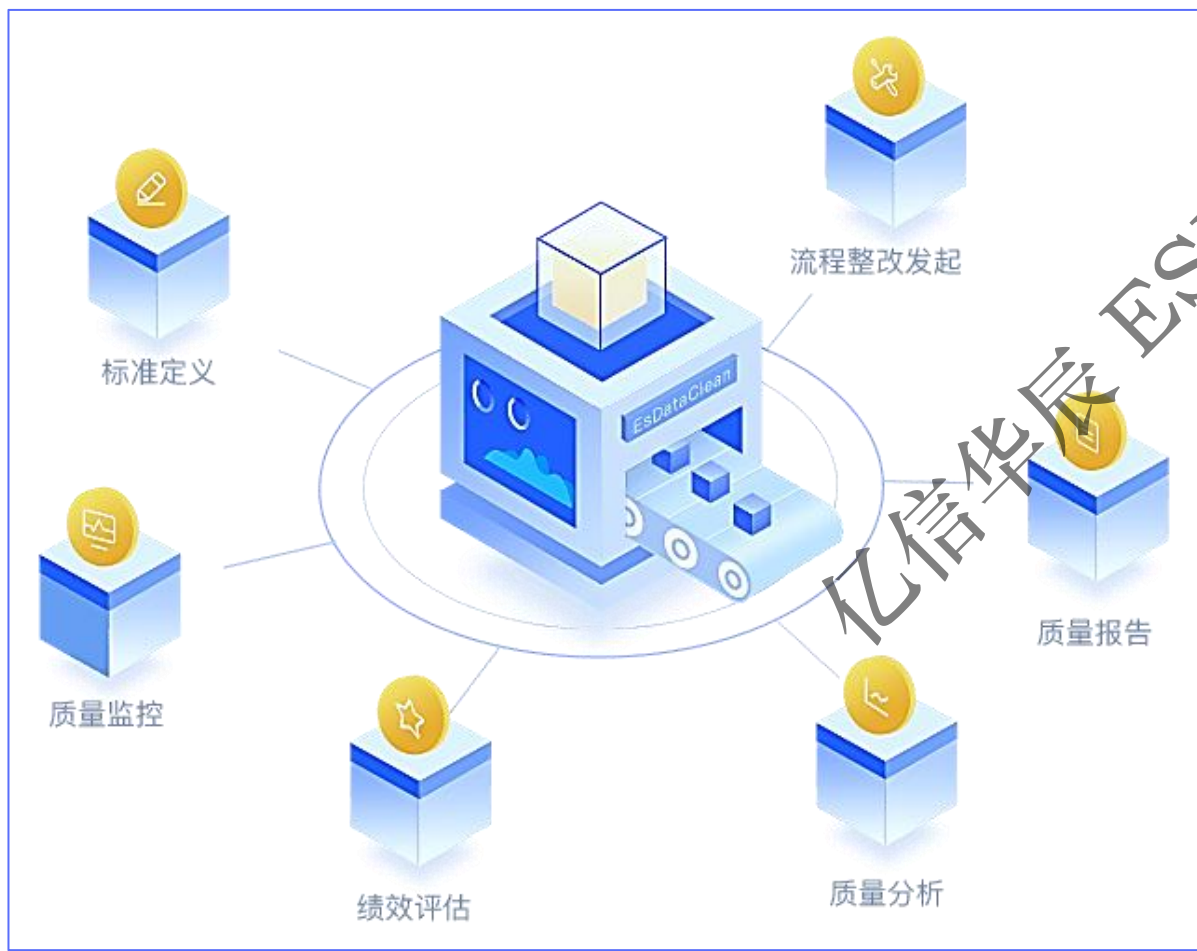
价值主张：让数据清澈如水

特点：零编码、覆盖质量管理全过程



面向对象：数据治理人员、业务人员

亿信数据质量管理平台 (EsDataClean) 提供从标准定义、质量监控、绩效评估、质量分析、质量报告、重大问题及时告警、流程整改发起、系统管理等数据质量管理全过程的功能。



数据质量规则一站式管理

新建规则-请选择规则类型

- 空值检查
用于检查字段是否非空
- 值域检查
用于检查关键指标取值范围
- 规范检查
用于检查一个字符型字段的类型...
- 逻辑检查
多个指标的逻辑检查
- 重复数据检查
检查一张表内的重复数据
- 及时性检查
用于检查数据上报的及时性
- 记录缺失检查
用于检查相对比照表数据是否完整
- 引用完整性检查
用于检查相对比照表数据是否有效
- 平衡性检查
用于检查数据的误差是否在允许...
- SQL脚本
用于执行SQL检查数据
- 数据集检查
检查当前事实表的结果与另外...

质量规则类别

EsDataClean 数据质量管理平台

系统首页 数据源 数据集 质量管理 workflow

demo,您好!

新建 新建分组 批量编辑 移动 重新计算权重 启用禁用 导入导出 标准同步 克隆 删除 刷新

代号/主题表/描述

代号	描述	类型	问题级别	全局权重/权重	主题表	创建时间
LOGIC_0	出院时间一定是在入院时间之后的	逻辑检查	☆☆☆	10.0/10	CYBR(患者...	2019-12
LOGIC_1	实际住院天数=出院时间-入院时间+1	逻辑检查	☆☆☆	20.0/20	CYBR(患者...	2019-12
LOGIC_2	年龄在13岁到18岁(含)之间, 疾病代码不允许为C50...	逻辑检查	☆☆☆	10.0/10	CYBR(患者...	2019-12
NULL_0	患者现住址不得为空	空值检查	☆☆☆	10.0/10	CYBR(患者...	2019-12
NULL_1	主治医师不得为空	空值检查	☆☆☆	10.0/10	CYBR(患者...	2019-12
REPEAT_0	同一期的出院病人, 姓名、性别、身份证号不应该重...	重复数据检查	☆☆☆	20.0/20	CYBR(患者...	2019-12
STANDARD_0	患者身份证号须符合规范	规范检查	☆☆☆	30.0/30	CYBR(患者...	2019-12
WAVE_0	手术费异常波动, 应在上下20%之间。	波动检查	☆☆☆	10.0/10	CYBR(患者...	2019-12

质量规则管理

亿信数据质量管理平台（EsDataClean）包含丰富的质量评价方法，并且易于扩展。系统支持数十种质量评价算法技术，满足业务系统运行、数据中心建设、数据治理过程中各类规则的定义，并可实现跨数据源的对比分析；支持通过XML扩展，可完全适应企业未来的数据质量管理需求的变化。

评估指标、分值权重

规则代码	路径	规则类型	权重	问题级别	创建时间	创建人员	操作
LOGIC_8_1	/	逻辑检查	1.00*10	一般	2014-09-18 09:44:08	超级管理员	编辑 删除 禁用
LOGIC_8_2	/	逻辑检查	1.00*10	一般	2014-09-18 09:44:52	超级管理员	编辑 删除 禁用
LOGIC_8_3	/	逻辑检查	1.00*10	一般	2014-09-18 09:49:37	超级管理员	编辑 删除 禁用
LOGIC_8_4	/	逻辑检查	1.00*10	一般	2014-09-18 09:51:27	超级管理员	编辑 删除 禁用

系统质量绩效

序号	质检结果	数据期	数据量	错误数	绩效	质检方式	结果处理方式	耗时	执行者	质检时间
1	成功	2014年8月	665	287	92.81	计划任务	更新	8秒797毫秒	system	2014-09-20 03:00:00
2				277	92.84	手动	覆盖	3秒	admin	2014-09-18 09:52:50

各主体绩效考核

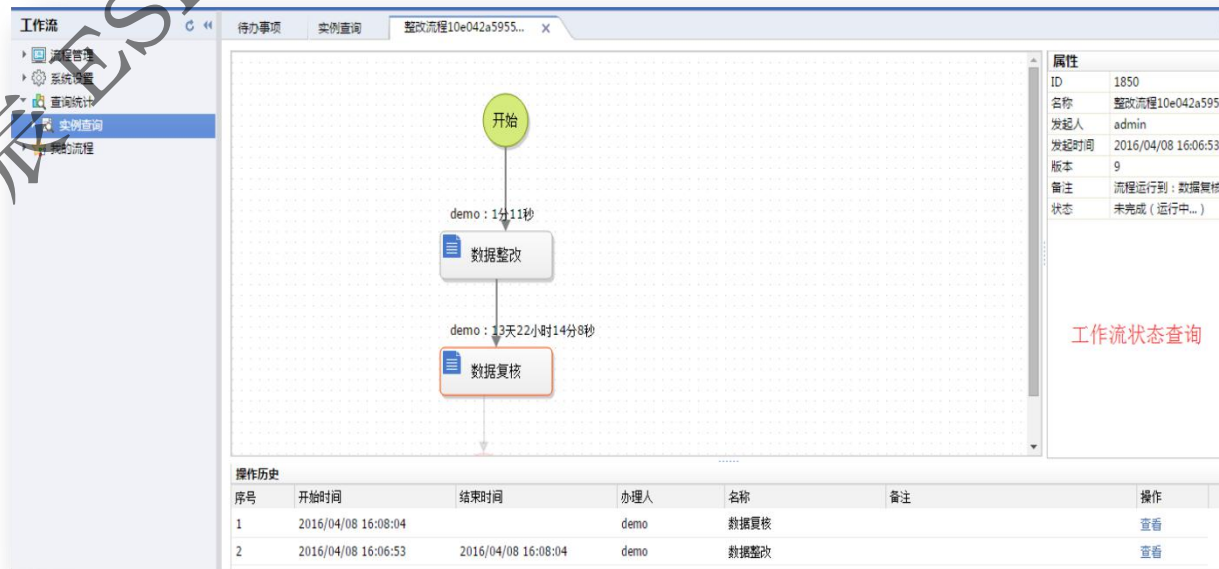
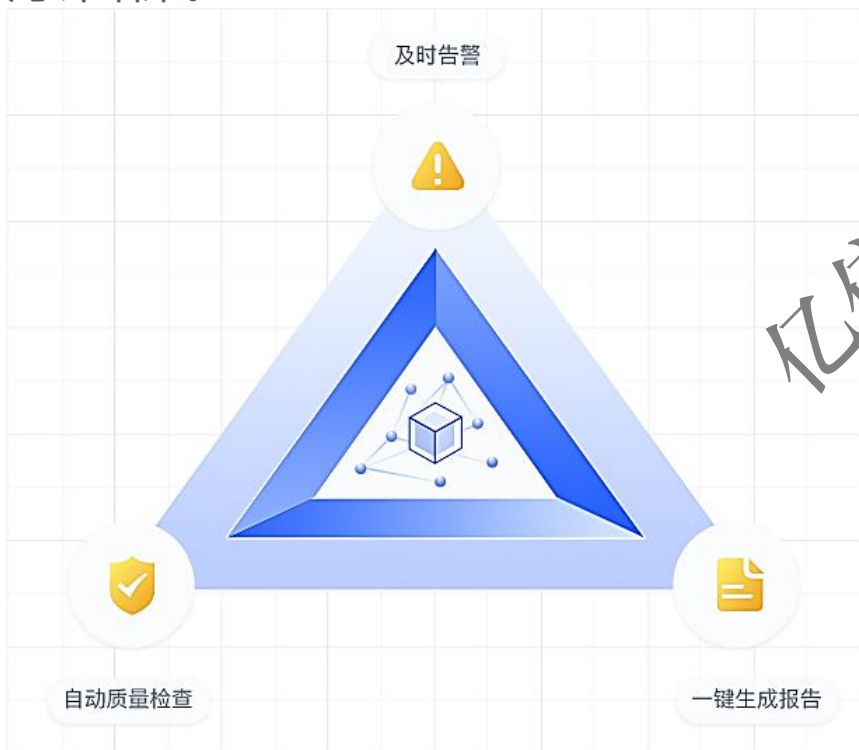
数据期：2013年1月 数据级次：湖北省410000

单位代号	单位简称	错误数	警告数	例外数	已修复数	绩效	操作
420010	业务库1	23	13	11	12	87	发起整改
420020	业务库2	31	12	12	5	86	办理 查看
420030	业务库3	13	20	85	20	93	

智能数据质量检查调度：通过事先定义好的规则、调度时间、工作流程，自动完成数据的质量检查，极大的减少人力的投入和过程干预，提升效率，减少误差。

重大问题及时告警：对质量检查的结果提供多方式（界面、邮件、短信）告警，让用户及时了解到系统检查结果，避免重大问题的延误。

一键生成质量报告和评估结果：系统通过数理统计、数据分析等技术，根据事先定义好的模板，自动生成质量报告和绩效考评结果。



序号	开始时间	结束时间	办理人	名称	备注	操作
1	2016/04/08 16:08:04		demo	数据复核		查看
2	2016/04/08 16:06:53	2016/04/08 16:08:04	demo	数据整改		查看

批准、退回、转办、催办、抄送、流程监控.....

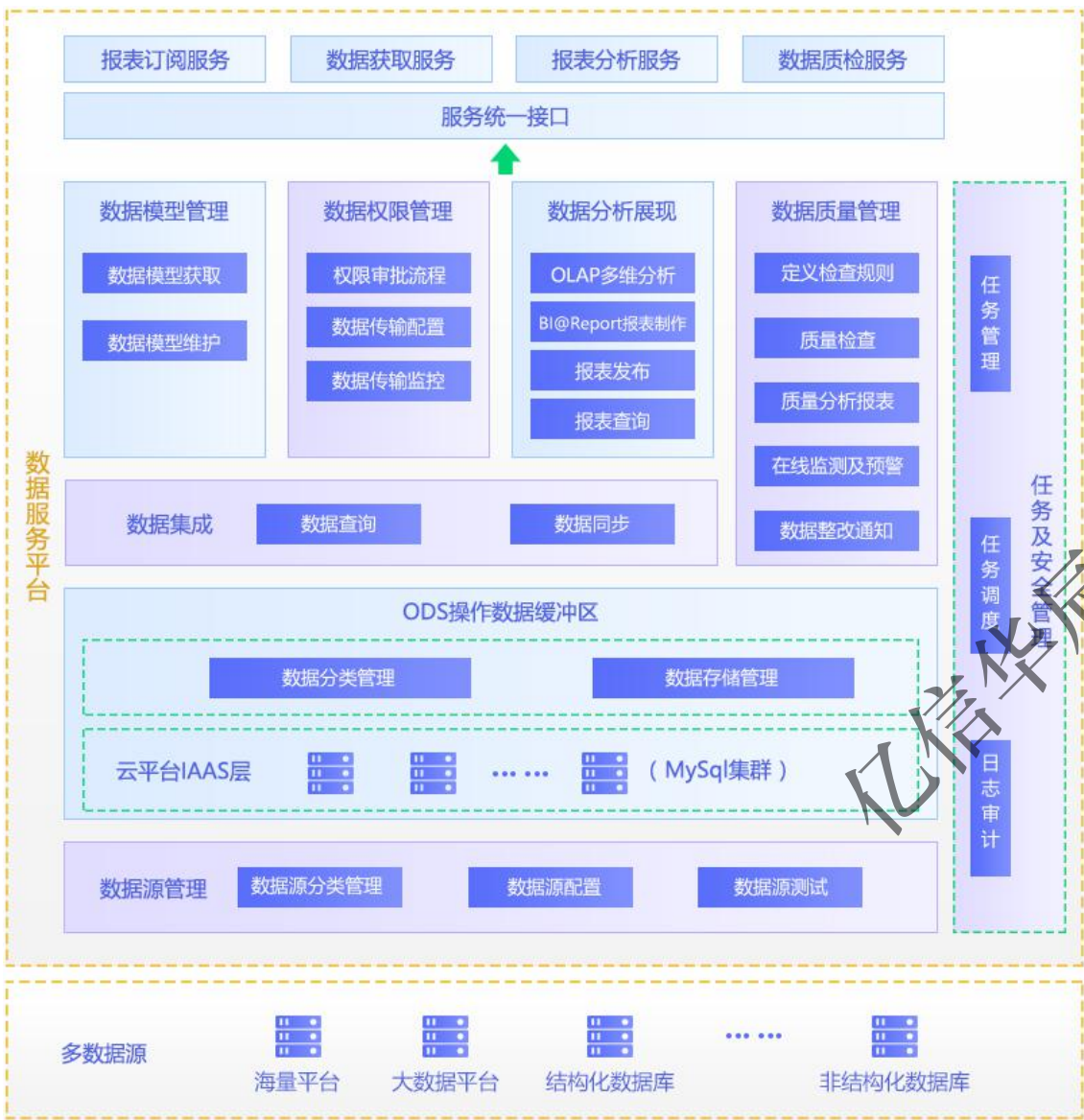
- ✓ 多维度质量分析报表，辅助用户对问题数据进行质量分析，以便用户进行有针对性的质量改进；
- ✓ 内置丰富的统计分析报告及多种质检结果主题，同时支持自定义扩展符合行业需求的质检结果主题，满足各行业用户的个性化需求。
- ✓ 多角度质量绩效评分，EsDataClean支持用户定义评分依据和权重，并可按照字段、表、规则类别、关键字等粒度生成质量评估结果。



IV 数据质量成功案例

亿信华辰 ESENSOFT





项目背景:

电力公司经过多年信息化系统建设和应用，积累了海量的数据信息，这些数据已然成为电力公司的核心资源之一。故构建数据服务平台，将企业运营数据以主题的形式整理，以服务的方式向数据使用方提供数据服务，促进数据在电力企业内部共享流通，外部开放，提高数据利用率。

为快速解决现有系统中存在的飞码异常、表码倒走、台区/用户漏采、日超容等问题，在数据服务平台中规划数据质量管理平台，快速定位问题数据，对数据进行定期巡检，提升数据质量，为后续的数据分析提供有效的数据基础。

项目方案:

数据服务平台分期建设数据源管理平台，数据集成平台，数据权限管理平台，数据质量管理平台，数据缓冲区，数据分析平台，统计对外服务接口等多个模块。多模块协同整合多源数据，提升数据质量，面向不同用户提供相关数据支持服务

其中数据质量管理部分则借助亿信EsDataClean数据质量管理工具，全面提升数据质量，减少因数据不可靠导致的用户决策偏差和损失。

亿信EsDataClean按照指定检测数据源，定义之间规则，定期巡检，推送质检报告的流程为用户提供全面完善的质检方案。

亿信EsDataClean以全面的质量管理PDCA循环管理方法为指导，根据实际数据情况定义规则检测相关数据。例如：

1) 飞码异常

对用电采集系统中的电能表进行异常判别。

判别规则：如当日的抄见电量大于该表上个月平均电量的X倍（X阈值，暂定为10倍），则该数据属于异常数据。

2) 表码倒走

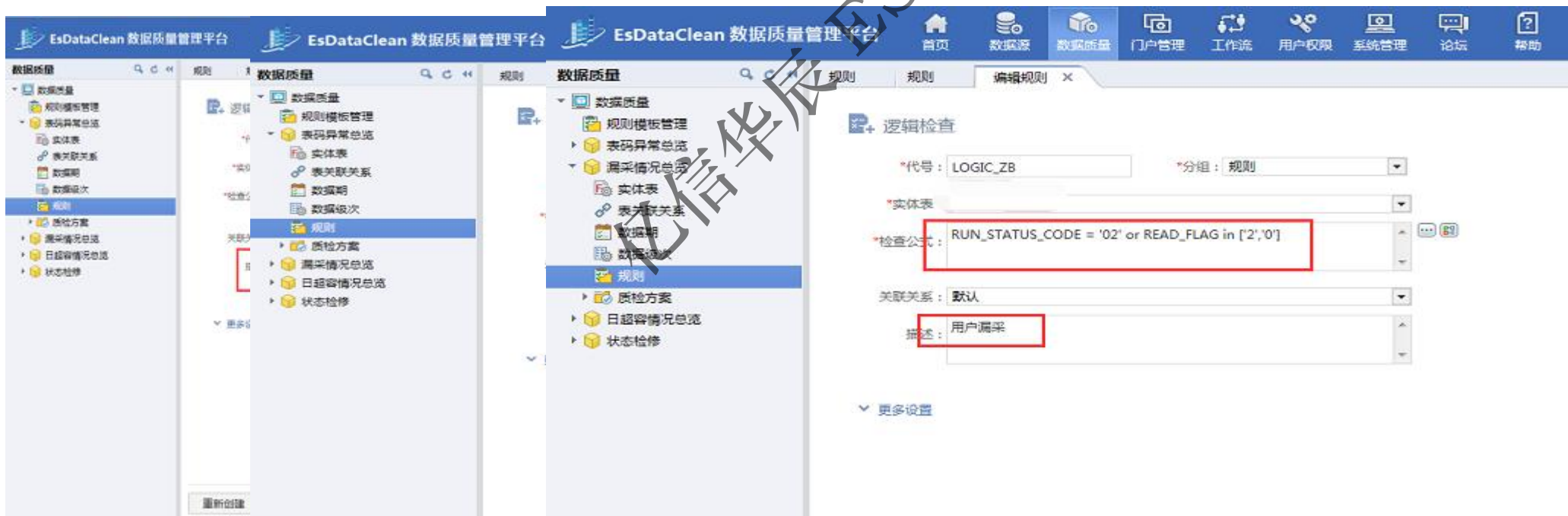
对用电采集系统中的电能表进行异常判别。

判别规则：如当日的止码小于上一日的止码则该数据属于异常数据。

3) 台区/用户漏采

对公变台区、专变用户进行异常判别。

判别规则：对用户\台区的供电电量明细数据表中的表计读数状态、表计运行状态进行判别，如标记为漏采的则表示该数据异常。



项目概述

2014年3月，启动“一门式”改革，吹响了政务大数据改革的号角；2016年7月，区数据统筹局正式挂牌成立，在区大数据发展决策委员会的指导下，构建“一局、一中心、一公司”的组织体系，推进数统统筹发展的共享、共建、共用，实现“用数据决策、用数据监管、用数据创业”的数据统筹发展运行机制。

区大数据统筹局于2018年开展数据治理系统建设工作，通过构建模型、配置规则、人工智能识别等方法对区数据进行数据核查、数据清洗、数据更新及数据修复，并输出数据分析报告，提出改进建议，为区各个部门数据源提供数据标准规范及改进建议，为政府部门、事业单位、企业、市民等数据使用者提供更优质、更高效的数据支持。

项目价值

以治理自然人库数据为主，重点治理卫计、工商、流管三个部门的数据，为部门及数据应用提供高质量数据支撑。支持自助填表系统、教育无纸化两大应用，通过数据比对功能，返回比对结果，提供错误修改建议，提高填报成功率。



建立质检规则



该区大数据统筹局在数据质量管理模块出现的问题：
数据质量低，各部门数据标准和格式不一致，相似情况重复发生的问题等等

整理质检规则及清洗方案

序号	字段名称	中文名称	质检	质检规则编号	清洗
1.	F29E1000_23AB011001	唯一标识	不能为空 不能重复 如果不为空，则应大于等于两个汉字	NULL_36 REPEAT_9 LOGIC_23	
2.	F29E1000_23AB011002	子女的姓名	如果不为空，则不能含有空格及不可见字符	LOGIC_70	转半角； 去空格；
3.	F29E1000_23AB011003	性别	不能为空 符合码表 如果身份证号不为空，则性别应与身份证中表示性别的位数保持一致（18位中的第17位，15位中的第15位，奇数为男性，偶数为女性）	NULL_35 QUOTE_34 LOGIC_18	
	F29E1000_23AB011004	出生日期	不能为空 如果身份证号不为空，则应与18位身份证中的第7-14位相符，与15位身份证中的第9-14位相符	NULL_11	
	F29E1000_23AB011005	政策生育编码	如果为空，则应符合码表	LOGIC_8	
6.	F29E1000_23AB011006	统计批次	如果为空，则应符合码表	QUOTE_35	
			如果为空，且证件类型编码为身份证，则校验身份证号码规范	QUOTE_16	
			如果为空，则不能重复	STANDARD_5 REPEAT_1	
7.	F29E1000_23AB011007	身份证号码	如果为空，身份证号长度18位且字母需为大写	LOGIC_65	转大写； 转半角； 去空格； 15位转18位
8.	F29E1000_23AB011008	子女死亡原因编码	如果为空，则应符合码表	QUOTE_36	
9.	F29E1000_23AB011009	子女死亡日期	如果子女死亡日期不为空，则子女死亡原因编码不能为空	LOGIC_26	
10.	F29E1000_23AB011010	变化原因编码	如果子女死亡原因编码不为空，则子女死亡日期不能为空	LOGIC_17	
11.	F29E1000_23AB011011	现家庭母亲标识ID	如果为空，则应符合码表	QUOTE_37	
12.	F29E1000_23AB011012	现家庭父亲标识ID			
13.	F29E1000_23AB011013	数据共享日期	不能为空	NULL_40	

通过建设数据治理平台，**定义数据质量检核规则，制定数据质量检查方案**，对各部门数据进行绩效评分，并对问题数据进行整改，使进入治理库的数据质量得以提升。

质检规则（某业务版块）

针对数据质量问题定义了72条规则

数据完整性

- 姓名、身份证号等关键字段为空的情况
- 生育申请中的申请孩次信息为空的情况
- 生育结果检查单位为空的情况
- 被领养子女父母不在人口登记信息中
- 生育申请人不在人口登记信息中
- 孕产妇不在人口登记信息中

数据准确性

- 身份证号不符合规范的情况
- 姓名、年龄数据异常的情况
- 初婚日期、婚姻变更日期等日期出现时间逻辑错误的情况
- 领养儿童的父母登记为同一人的情况
- 人口登记信息中身份证号码重复的数据

数据一致性

- 出生日期与身份证信息中不一致的情况
- 性别与身份证信息中不一致的情况
- 国籍编码、政治面貌编码等数据与统一码表中的值不一致的情况

数据及时性

- 人口信息数据是否在规定时间内进行了上报



亿信华辰合作伙伴招募计划

携手共赢，开启智能数据新起点

亿信华辰 ESENSOFT



项目合作



产品/服务合作



OEM合作

数据治理专题培训第三期打卡学习心得

1、社区打卡跟帖谈谈你学完第三期课程后，对数据治理和亿信数据质量管理产品的理解

帖子入口：<http://bbs.esensoft.com/thread-136298-1-1.html>

打卡模板：#第三期数据质量——数据治理之数据整改方案作业打卡#+本期直播课具体的感受+你的疑问+其他的想法。



我们的培训绝不止于此

ESENTOFT



ESENTOFT 首页 产品 解决方案 案例 学习 服务 伙伴 关于 社区 400-0011-866

亿信华辰
智能数据全生命周期产品与服务提供商

提供数据采集、数据存储、数据治理、数据分析产品与服务

视频介绍 产品推荐 资料下载

在线客服 联系电话 社区交流

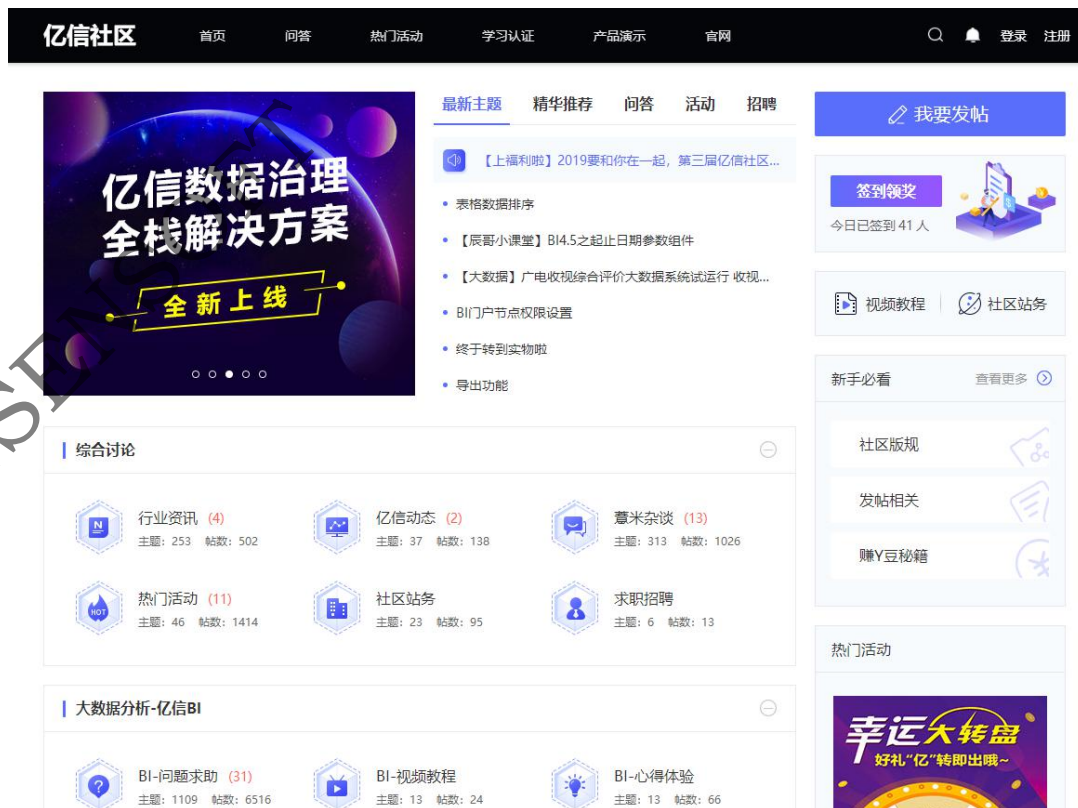
产品更新 PetaBase-s实时大数据平台发布 6大特性拥抱大数据生态	华为认证 亿信BI和查询与华为GaussDB 完成产品兼容性互认证	荣誉奖项 亿信华辰荣登CB Insights 中国企业服务竞争力榜单	数据资产管理白皮书 亿信华辰深度参与编写 为企业提升数据资产价值
---------------------------------------------------	------------------------------------------------	-------------------------------------------------	-----------------------------------------------

覆盖数据全生命周期的产品线

深耕大数据各环节，让数据驱动企业变革与进步

 数据存储	 数据处理	 数据治理	 数据分析	 运营监控
-------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------

亿信官网



亿信社区 首页 问答 热门活动 学习认证 产品演示 官网 登录 注册

最新主题 精华推荐 问答 活动 招聘

【上福利啦】2019要和你在一起，第三届亿信社区...

- 表格数据排序
- 【辰哥小课堂】BI4.5之起止日期参数组件
- 【大数据】广电收视综合评价大数据系统试运行 收视...
- BI门户节点权限设置
- 终于转到实物啦
- 导出功能

我要发帖

签到领奖
今日已签到 41人

视频教程 社区站务

新手必看 查看更多







社区版规 发帖相关 赚Y豆秘籍

热门活动




幸运大转盘
好礼“亿”转即出哦~

亿信数据治理全栈解决方案 全新上线

综合讨论

 行业资讯 (4) 主题: 253 帖数: 502	 亿信动态 (2) 主题: 37 帖数: 138	 薏米杂谈 (13) 主题: 313 帖数: 1026
 热门活动 (11) 主题: 46 帖数: 1414	 社区站务 主题: 23 帖数: 95	 求职招聘 主题: 6 帖数: 13

大数据分析-亿信BI

 BI-问题求助 (31) 主题: 1109 帖数: 6516	 BI-视频教程 主题: 13 帖数: 24	 BI-心得体验 主题: 13 帖数: 66
-------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------

亿信社区

我们的培训绝不止于此

ESENSOFT



亿信官微



ESENISOFT

感谢您的参与

Thank you for your participation

智能数据全生命周期产品与服务提供商

