
八爪鱼采集系统应用案例

——政府税务部门案例

一、项目背景

随着各行业信息化程度的飞速发展，互联网应用已成为各行业生产经营的主要因素之一，从而形成了大量的互联网涉税信息，对互联网大数据的挖掘应用已成为“互联网+税务”应用的主题。

在这样的时代背景下，2015年税务总局决定在全国税务系统实施“互联网+税务”行动计划，把互联网的创新成果与税收工作深度融合，拓展信息化应用领域，推动效率提升和管理变革，为实现税收现代化提供了有力的支撑与保障。

二、用户需求

·随着互联网的发展和信息数据的爆炸式增长，税务部门对互联网信息的采集、各方数据的深入调和匹配、内外数据的关联和互动提出了更高的要求，因此如何获取互联网和第三方的涉税信息并挖掘其中的价值，成为了税务部门较为迫切的需求。

·通过综合治税合作部门提供的涉税信息有限，为有效解决征纳双方信息不对称问题，税务部门急需拓展互联网信息采集渠道。

总的来说，税务部门各项工作的开展需要强有力的数据支撑，需要全方位多维度识别漏征漏管户，从而提高补查税款能力，提高涉税风险识别、风险监控、风险预警、风险趋势分析能力。

三、解决方案

1、技术架构

八爪鱼数据采集系统采用分布式云计算构架，大规模高速并发采集，能够对互联网目标涉税信息源（电商、房产、土地、商演、企业经营、行政监管、金融、本地生活、社交网站，新闻、采购、税务政策等数据）进行各类碎片化信息的实时采集。采集信息源覆盖各类涉税相关网站，各类公开数据源，税务指定网站，频道，页面的税务领域的信息。

整个系统包含主程序、监控程序、采集规则配置客户端、分布式采集集群和存储集群。

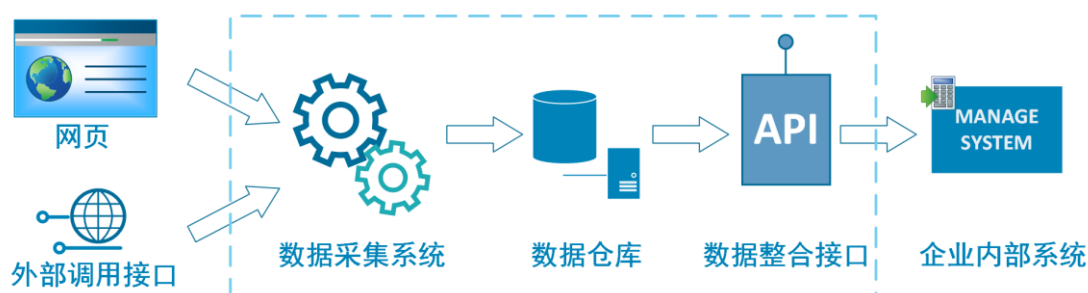


2、实施过程

八爪鱼数据采集系统主程序负责实现任务负载均衡分发，采集集群实现多类别、多站点同时并发采集。存储集群可存储半结构化和非结构化的数据，具有高容错性、高可用性。监控程序提供一个“一站式”人机交互界面供系统管理人员使用。功能包括服务资源管理、节点资源管理、任务控制及监视。

采集过程中需要采集规则制作，规则配置客户端通过模拟人工网页浏览操作，全程可视化制作，无须编写代码，任何人都可以掌握。同时云采集功能可以实现 7×24 小时循环信息数据的采集、增量的信息采集，全网采集、定向采集等，多种智能采集形式相结合。

采集完成后八爪鱼提供数据导出接口，轻松实现数据导出，并与企业内部系统无缝链接。



四、客户效益

目前，八爪鱼已为多家政府税务部门提供强有力的数据源支持，提高税务部门风险识别、监控、预警、分析能力，帮助政府建立一个高效的税务管理系统。具体效益如下：

1、全面采集互联网涉税信息，解决征纳双方信息不对称问题，提升互联网税源管理能力。

2、利用大数据云计算技术，主动采集各类互联网涉税信息，推进税务工作效率提升和涉税信息来源管理变革。

3、落实“互联网+税务”行动计划，拓展税务信息化应用。