

多源异构大数据平台的建设及应用

刘 琦

(河南警察学院网络安全系, 河南 郑州 450000)

✉569797767@qq.com



摘 要: 针对公安机关使用的信息系统普遍存在“数据壁垒”, 不利于侦查办案的实际问题, 设计并实现了多源异构大数据平台。该平台能够实现侦查、治安、交管等多类业务系统结构化、非结构化、半结构化异构数据的有效汇集整合, 通过数据处理、目录管理、分析研判等功能, 解决由于数据无法有效利用而导致“信息孤岛”、无法服务于侦查破案的问题。该平台应用结果表明, 平台的建设能够有效地汇聚数据资源, 实现多源异构数据的分析处理及态势预测, 具有较高的行业应用价值。

关键词: 多源异构数据; 大数据平台; SolrCloud

中图分类号: TP391 **文献标识码:** A

Construction and Application of Multi-source Heterogeneous Big Data Platform

LIU Qi

(Network Security Department, Henan Police College, Zhengzhou 450000, China)

✉569797767@qq.com

Abstract: Data barriers widely exist in the information systems of public security organs, which are not conducive to investigating and handling cases. Aiming at this problem, this paper proposes a multi-source heterogeneous big data platform. The platform can realize various business systems such as investigation, public security, and traffic management, and effectively collect and integrate structured, unstructured, and semi-structured heterogeneous data. Functions of data processing, catalog management, analysis and judgment are performed to solve problems of information islands and the inability to serve investigations and resolution of cases, due to ineffective use of data. The platform application results show that the platform construction can effectively gather data resources, realize the analysis and processing of multi-source heterogeneous data and forecast the situation. It has high industry application value.

Keywords: multi-source heterogeneous data; big data platform; SolrCloud

1 引言(Introduction)

随着经济和科技的高速发展, 伴随而来的各种新型犯罪活动频率高并且难以侦破^[1]。经过多年持续不断的建设, 公安信息化已经步入智慧公安时代, 其发展重点逐步从基础设施建设向管理与应用转换。但是, 由于技术、管理等多方面因素限制, “数据壁垒”问题尤为突出^[2], 各业务警种各自占据

不同的数据资源, 却无法甚至不愿意进行资源共享, 使得数据、信息流转不畅, 对侦查办案支撑不力。为解决上述棘手问题, 各地公安机关纷纷通过建设公安大数据中心、搭建公安大数据平台的方式, 实现数据汇聚、资源共享并完成研判分析等工作^[3]。

随着公安大数据中心建设的不断推进, 根据工作需要,

公安机关数据中心纷纷建设多源异构大数据平台,组建与之匹配的机构,分配专职工作人员,用以满足业务需要。公安大数据平台(以下简称“平台”)逐步成为提升公安机关侦查破案效率的有力工具^[4]。平台能够为授权用户安全获取相应数据,深度利用数据取得情报奠定基础^[5];平台也能够为满足公安数据考核要求、公安情报工作、公安业务开展,提供公共、基础、权威、完善的数据资源支撑服务。

2 多源异构大数据平台研究现状分析(Research status analysis of multi-source heterogeneous big data platforms)

平台建设的目标是汇集整合各类业务系统数据,有效解决各业务系统间存在的信息孤岛、信息壁垒等客观问题^[6],构建形成一套上下级联合、横向贯通、逻辑一体化的数据服务体系,建设一套科学理论来实现数据的科学治理^[7]。但是,目前各级公安机关仅仅实现了各业务系统数据的汇集和数据标准化,数据间蕴含的价值无法体现;各种类型数据间关系无法进行关联,结构化与非结构化的数据无法实现横向扩展,历史图像、音频、视频等也不具备再次分析利用的条件,整个数据循环无法达到闭环传输利用的效果^[8];从海量动态数据中快速、精准地发现、排查可疑信息的能力提升空间较大;重大案事件的预测预警预防精度不够,数据治理方法不科学,治理效果堪忧。基于上述情况,本文提出公安多源异构大数据平台的设计方案,希望为解决公安大数据平台中存在的诸多问题起到借鉴作用。

3 公安多源异构大数据平台的设计(Design of public security multi-source heterogeneous big data platform)

3.1 总体架构

根据业务需要,平台的建设需要汇聚各类业务系统数据,涉及公安、交通、发改(发展改革)、统计、工商、地税、互联网、银行、社保等各类多源异构数据。获取的数据需要经过清洗及抽取,内容过滤、存储,消息服务,数据统计等预处理,汇入数据资源池。入池后数据经过再清洗,标记,分类分级,索引、存储、更新等数据处理过程后,实现数据目录管理,如将数据分别放置于内存数据库、基础数据库、专项数据库、关联数据库、共享数据库、决策数据库、资源数据库等目录中,以备之后的数据研判与挖掘。数据研判与挖掘子系统实现实时数据流挖掘,关联分析、统计、反演,态势预测分析、可视化展示等功能,并推向需要的业务系统中的应用。最终,通过综合分析研判,领导、专家会商等,支持最终决策判定。其总体架构如图1所示。

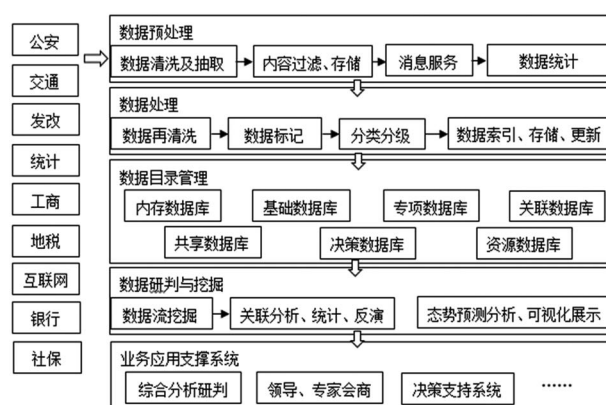


图1 多源异构大数据平台总体架构

Fig.1 Overall architecture of multi-source heterogeneous big data platform

3.2 功能架构

融合平台包括基础设施层、信息资源层、应用支撑层、应用层四个层面。基础设施层主要包括网络、主机、存储、系统软件、大数据软硬件等设备,完成融合平台建设需要的设施保障。信息资源层包括基础地理信息库、治安库、警综库、交警库、出入境库、信息资源服务总线等,实现原始库、历史库、共享库、日志库之间的信息资源流转。应用支撑层提供相关应用的支撑,包括权限控制、系统监控、采集系统、目录管理、资源发布、交换共享服务、其他服务、数据交换、数据采集、安全监控、统计分析、数据整合、数据审核、综合查询、统计共享、元数据管理及其他组件等。应用层提供平台门户、搜索引擎、预警发布、业务系统、决策分析、查询与统计等应用。同时,系统需要以运维管理体系、安全防护体系、标准体系及容灾备份系统做支撑,如图2所示。

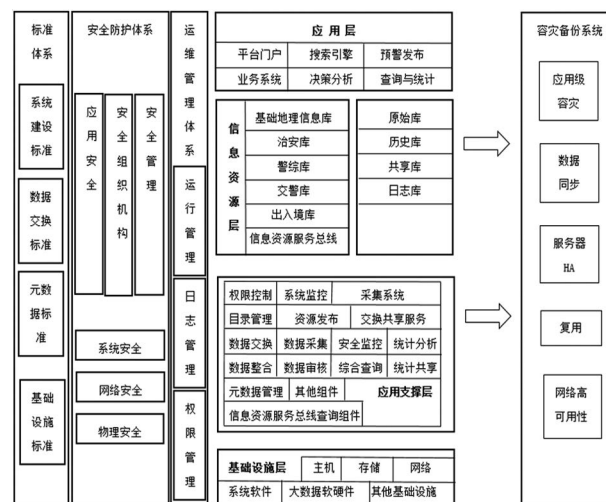


图2 融合平台功能架构

Fig.2 Functional architecture of integration platform

融合平台设计与实现的关键在于数据接入整合、数据标

准管理、数据资源目录、数据融合、数据汇聚库、基础数据库、动态数据库、数据存储管理、数据质量管理、数据资源监控等功能的实现,下面将分述之。

(1)数据接入整合

汇聚整合数据来源包括警务基础工作平台数据、情报综合应用平台数据、地理信息系统数据、视频监控平台等系统数据;公安部及省厅下发数据、平行公安机关协作交换数据,同时接入电子政务数据、社会信息数据及移动互联网数据。汇聚整合数据种类包括数字、文本、图表、视频、音频、地理空间等结构化、半结构化、非结构化数据。数据接入要解决的关键问题是多源数据问题,如数据跨网传输、数据存储不统一、数据结构不规范、动态静态数据同时存在等情况。

数据接入平台由两部分组成:数据采集工具和数据转换工具。数据采集工具KETTLE的功能和技术特点主要体现在:数据联邦技术、数据缓存技术、数据优化技术、高效跨库数据加载技术和数据复制技术。其数据处理流程图如图3所示。

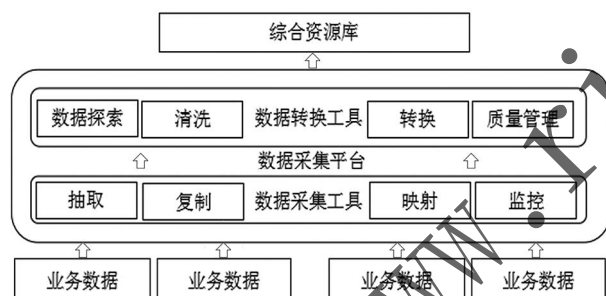


图3 数据处理流程图

Fig.3 Data processing flowchart

(2)数据标准管理

公安信息标准化的基础是数据的标准化。公安数据中心建设旨在整合共享最大化,数据共享的最佳途径是数据标准化。基于数据中心建设的总体思路,依据统一数据规范建设数据标准基础库、数据标准管理库、数据标准应用库和专题分析库。数据标准管理主要包含数据元管理、代码项管理、数据项管理、同义(近义)词管理、术语管理、质量自检、标准检索统计等,结合标准注册工具、标准监测工具,实现对数据标准的注册、修改、发布、审核、启用、停用、下线等全流程管理,为数据标准化提供依据支撑,同时数据标准也要符合公安部数据标准规范。

(3)数据资源目录

数据资源种类繁多,需要根据一定的标准进行科学编目。数据资源目录包含数据资源注册、数据资源发布、数据资源维护、数据资源查询等。数据资源注册时,关键信息包含数据资源编号、数据资源名称、所属目录编号、数据资源描述、数据资源提供单位_公安机关机构代码、数据资源事权单位_事权单位代码、数据资源及其更新方式代码、更新周期代码、公安业务分类代码、行业属性代码、数据资源共享范围代码、数据资源共享方式代码、数据资源共享地区代码、数据资源共享部门代码、数据资源要素分类代码、数据资源属性分类代码等。

(4)数据融合

数据融合是基于公安信息网、各类专网、互联网等不同数据源采集的海量异构数据,根据人员、地址、车辆、资讯信息、虚拟身份、警情、案事件等要素建立若干个相互独立而又逻辑贯通的数据库。

(5)数据汇聚库

数据汇聚库是对元数据中结构化数据和非结构化数据通过数据文件交换、数据接口同步后进行存储,实现数据预处理(格式转化、错误纠正、去重、标准化等工作),再实现数据关联,通过分类整合,形成基础数据或动态数据。

(6)基础数据库

将基础数据从数据汇聚库中抽取出来,储存到基础数据库中。此类信息具有相对固定、变化频率不高、变化程度低的特点。

(7)动态数据库

将动态数据从数据汇聚库中抽取出来,储存到动态数据库中,如卡口车辆通过信息、网吧上网信息、住宿信息、出租车信息、火车订票信息、出入境记录信息等。此类信息具有实时变化、每日数据增量较高的特点。

(8)数据存储管理

数据存储管理实现对计算集群、数据存储、文件系统及数据库的有效管理、任务调度、服务监控,具有高可靠的架构设计。基础数据、动态数据在存储入库后,将所有数据按照公安五要素进行逻辑上的分类。

(9)数据质量管理

数据质量管理贯穿数据采集、汇聚、整合和综合应用全过程。数据质量管理目标是建立一体化的数据质量监控体系,构建平台化、智能化的数据质量监控分析模块,包括业

务逻辑校验、数据质量监控和问题数据纠错等应用功能。

(10)数据资源监控

数据资源监控能够及时掌握数据资源汇聚的规模、数据质量情况、数据传输情况。通过对数据资源涉及的整合情况、数据交换情况、数据资源情况、数据质量情况、数据实时传输情况、硬件设备资源情况等进行实时监控，可视化地展示给民警，并能进行及时告警。

4 多源异构大数据平台的应用示例——大数据检索的实现(Application example of multi-source heterogeneous big data platform—implementation of big data retrieval)

由于数据的指数级增长趋势，从多个数据源采集、融合、分析、处理数据成为学术界、工程界的研究热点之一。使用大数据检索技术是解决海量数据中有价值信息快速提取的有效方式。基于多源大数据进行检索，提取有价值的信息深挖情报是公安机关侦查破案的常用方式。当今流行的搜索引擎有以下几种：全文搜索引擎、目录索引引擎和元搜索引擎。比较多种检索技术后，选择实用性强、重复利用性高的分布式全文检索系统，它的实现基于SolrCloud集群技术，以实现公安多源信息检索技术。

4.1 系统总体设计

多源异构大数据平台的典型应用——分布式全文检索系统总体设计根据现实需求分析，如图4所示。

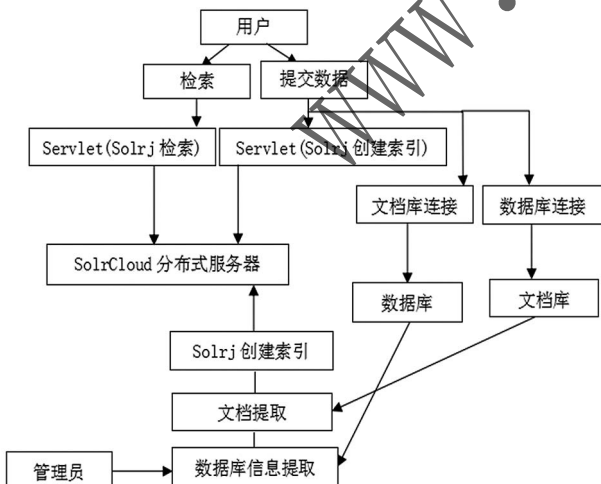


图4 分布式全文检索系统总体架构

Fig.4 Overall architecture of distributed full-text retrieval system

4.2 系统核心设计

分布式全文检索系统的核心设计为：用户输入搜索字

词，各索引分片上的节点分别进行条件检索，输出结果，再由最初受理的节点合并结果并将其排序，最终输出给用户，如图5所示。

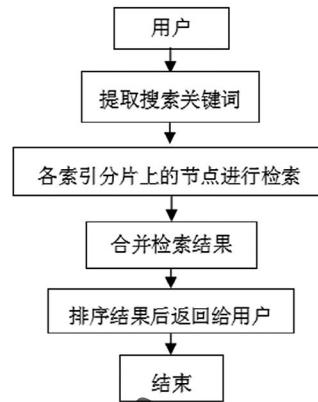


图5 分布式检索流程

Fig.5 Distributed retrieval process

4.3 系统部署与安装

第一步：环境准备，即服务器、3个centos 6.2。部署计划如表1所示。

表1 SolrCloud部署

Tab.1 SolrCloud deployment

IP地址	Solr	Zookeeper
***.16.5.1	yes	no
***.16.5.2	yes	yes
***.16.5.4	yes	no

Zookeeper：选择端口2181。

软件环境：JDK 1.7以上版本；Zookeeper管理配置信息和集群状态。下载JDK、Zookeeper、Solr。

第二步：安装相应环境，以便提供检索服务。安装完成后，能够访问部署的Solr集群中任意的端口服务。

5 结论(Conclusion)

随着智慧公安建设的不断推进，公安大数据必然发挥越来越重要的作用。公安大数据平台因其汇聚多源、异构、海量数据，并能够对这些数据进行分析研判，辅助领导决策而起到了“公安智慧大脑”的作用。本文阐述了公安大数据平台的架构、功能结构，并以大数据检索为例，阐述其部署及安装。

参考文献(References)

[1] 孟颖.基于云计算的大数据网络信息检索技术及扩展[J].科技资讯,2019,17(27):11-12.

- [2] 曾倩倩,张婷婷.基于大数据的图像检索技术在侦查中的应用[J].电子技术与软件工程,2018(21):175-176.
- [3] 朱维和,魏锋,高晓岚.公安大数据治理技术研究[J].警察技术,2018(06):32-35.
- [4] 邱明月,唐松泽.基于互联网资源的公安数据挖掘能力培养研究[J].信息与电脑(理论版),2018(19):158-159.
- [5] 徐茂春.网络安全分析中的大数据技术[J].电子技术与软件工程,2018(18):200.

(上接第62页)

(5)生产车间管理

生产车间管理主要包含车间动态看板、工资看板、派工看板、报工看板四个子页面。车间动态看板实现对各个生产车间中生产任务的动态监控,并将各个车间的生产任务分为等待开工、正在进行、已完成三类进行显示;工资看板查看对应员工的工资详细信息;派工看板可对生产计划管理模块所生成的生产计划表进行显示,并可选择相应的车间、班组、员工、设备进行派工;当所派工作完成时,可进行报工操作,报工看板可查看已报工的工序的详细信息。

4.4.2 移动端设计与实现

移动端功能设计和PC端类似,不同点在于移动端采用uni-app框架,使用uView组件库进行移动端静态页面编写,图表采用uniapp插件市场所发布的组件和uCharts组件库进行实现。

5 结论(Conclusion)

本文通过选用Spring Cloud、Vue等相关开发技术,设计并实现了基于Spring Cloud的工厂可视化管理系统,最终设计出的系统能够实现对工厂生产过程的可视化管控。经过测试、部署和试用,该系统运行稳定,使用体验良好,满足工厂生产管理需求,实现工厂管理的便捷化、信息化及可视化,对于工厂管理具有较大的应用价值。

参考文献(References)

- [1] 严聪蓉.制造业数字化转型发展的行动者网络建构与运行机制研究[D].西安:西安电子科技大学,2020.
- [2] 师宁,刘辉.基于工业4.0的数字化工厂监控系统设计[J].机电产品开发与创新,2020,33(06):71-73.
- [3] 王健,李冬睿.从单一模式系统架构往微服务架构迁移转化技术研究[J].科教导刊(下旬),2016(27):43-44.
- [4] 崔海涛,章程,丁翔,等.面向微服务架构的开发组织适应性评估框架[J].软件学报,2021,32(05):1256-1283.
- [5] 郑明钊,张建强.基于微服务的大平台系统架构演进探讨[J].软件,2017,38(12):165-169.
- [6] 王琪,卢军.基于Dubbo的分布式能效管理系统的设计与实现[J].电子设计工程,2018,26(11):174-178.

- [6] 葛文.视频大数据在公安行业的应用分析[J].信息记录材料,2018,19(10):74-75.
- [7] 党玉龙.公安资源服务平台异类数据融合设计与高效实现[J].警察技术,2018(05):91-94.
- [8] LI F H, ZHAO Z F, WANG Z Y. Hierarchical clustering based on single-pass for breaking topic detection and tracking[J]. High Technology Letters, 2018, 24(4):369-377.

作者简介:

刘琦(1978-),女,博士,副教授.研究领域:公安信息技术。

- [7] SUN J N. Design and implementation of dormitory repair management system based on Spring Cloud Microservices[J]. International Core Journal of Engineering, 2021, 7(1):201-204.
- [8] 孔晓阳,代真虎.基于WebSocket与Redis的高性能Web组态系统设计[J].信息技术与标准化,2021(03):51-56.
- [9] 李成仁.基于Vue.js的单页面WebGIS可视化框架研究与实现[J].地理空间信息,2020,18(05):83-86,98,7.
- [10] 许溜溜.基于HBuilder快速开发移动端APP的设计与实现[J].电脑知识与技术,2020,16(10):74-75.
- [11] 朱渔,罗来曦.浅议自动测试系统中MYSQL数据库的应用与分析[J].信息通信,2014(06):157.
- [12] 牟林钢.一种微服务架构下的负载均衡策略优化方法[J].电子设计工程,2020,28(07):87-90,94.
- [13] 王方旭.基于Spring Cloud和Docker的微服务架构设计[J].中国信息化,2018(03):53-55.
- [14] 崔灿,周伟.基于微服务的工地安全管理系统架构设计[J].软件导刊,2021,20(02):159-164.
- [15] 洪华军,吴建波,冷文浩.一种基于微服务架构的业务系统设计与实现[J].计算机与数字工程,2018,46(01):149-154.
- [16] 付曜华.基于微服务架构的数据分析平台的设计与实现[D].北京:北京邮电大学,2020.
- [17] 周永圣,侯峰裕,孙雯,等.基于SpringCloud微服务架构的进销存管理系统的设计与实现[J].工业控制计算机,2018,31(11):129-130,133.
- [18] 杨泰.基于微服务架构的云台服务端的设计与实现[D].成都:电子科技大学,2020.
- [19] 林红,华韵子.基于Redis的气象数据分发管理系统的设计与实现[J].软件工程,2019,22(03):47-50.

作者简介:

黄静(1965-),女,博士,教授.研究领域:通信工程,大数据,深度学习。

吴涵(1997-),男,硕士生.研究领域:软件工程,计算机技术。