文章编号：2095-2163(2023)07-0123-05 中图分类号：TN929.5 文献标志码：A

基于 Storm 的涉油气犯罪实时检测系统的研究与应用

张丽华[1]，李宏博[1]，王 健[1]，张伟民[2]，李欣欣[1]，李 娟[1]，邵国强[1]
（1 大庆师范学院 计算机科学与信息技术学院，黑龙江 大庆 163712；
2 大庆油田有限责任公司试油试采分公司，黑龙江 大庆 163712）

摘 要：针对重点涉油气企业犯罪人群分布特性，构建一个分析模型至关重要。通过对比分析、经验总结及仿真实验等研究方法来全局掌握涉犯罪人群热度指数、分布密度等因素，并借助服务器和 Storm 等相关技术搭建数据处理平台，再基于 Logstash 构建热力图，对数据具体位置结果进行判定，实现最终的设计及体系架构。经过本次构建研究，可以更精准地监测犯罪者，预判涉油气行业，尤其是在未来涉油气犯罪的总体趋势，充分发挥了数据处理的实时性，为涉油气企业资源犯罪的防控提供了相应的参考依据与保障。

关键词：大数据处理；Storm 技术；Logstash；热力图；实时检测

Research and application of monitoring system for oil and gas related crimes based on Storm

ZHANG Lihua[1], LI Hongbo[1], WANG Jian[1], ZHANG Weimin[2], LI Xinxin[1], LI Juan[1], SHAO Guoqiang[1]
(1 School of Computer Science and Information Technology, Daqing Normal University, Daqing Heilongjiang 163712, China;
2 Daqing Oilfield Co., Ltd., Oil Testing and Production Test Branch, Daqing Heilongjiang 163712, China)

【Abstract】 It is very important to construct an analysis model according to the distribution characteristics of key criminal groups involved in oil and gas enterprises. Through comparative analysis, experience summary, simulation experiment and other research methods, crime related population heat index, distribution density and other factors are globally grasped, and with the help of the server and related technologies such as Storm, a data processing platform is built, after that based on Logstash to build heat map, the specific location results of the data could be determined, therefore the final design and architecture is realized. Through this construction research, the criminals can be more accurately monitored and predicted, especially the general trend of oil and gas related crimes in the future, which could give full play to the real-time data processing and provide the corresponding reference basis and guarantee for the prevention and control of resource crimes related to oil and gas enterprises.

【Key words】 big data processing; Storm technology; Logstash; thermal map; real-time detection

0 引 言

近年来，涉油气犯罪的发生以及所带来的严重后果已经在全球范围内引起广泛关注，涉油气犯罪是影响涉油气企业生产秩序的最主要危害之一，如何结合国内相关部门的整治措施，深入分析涉油气资源犯罪的基本特征，探究其形成机制已刻不容缓。在此基础上，如何将“预防”与“治理”二者紧密结合，探讨和分析涉油气企业和国家如何对防控资源进行合理有效的配置是十分必要的。十九大报告明确要求，要加快社会治安防控体系建设，打造共建共享共治社会治理格局，提高防范和抵御安全风险能力。随着“5G+大数据”与“物联网+安全防范”的梦幻联动，涉油气行业应用数据逐渐呈现出高实时性，对提升处理平台的耗时及资源请求响应速率也提出更高要求。

本文首先对主要技术进行了整体阐述，然后对搭建在 Storm 平台上的 Flume-Kafka 高可用数据采

基金项目：大庆市哲学社会科学规划研究项目(DSGB2023102)；黑龙江省教育教学改革一般项目(SJGY20220630)；黑龙江省哲学社会科学规划研究项目(22TQE428)。

作者简介：张丽华(1980-)，女，讲师，主要研究方向：网络安全及大数据；李宏博(1982-)，男，博士，副教授，主要研究方向：大数据管理；王健(1971-)，男，教授，主要研究方向：数据库开发。

通讯作者：李宏博 Email：lihongbo@dqnu.edu.cn

收稿日期：2023-01-30

集缓存方案进行设计,重点讨论了热力点的获取方法,系统使用 Logstash 技术获取涉犯罪人数据源,并将采集到的涉犯罪人信息传输到系统服务器的数据库进行存储,最后使用热力图技术实现数据可视化。经过测试,系统运行正常有效,表明基于 Storm 的数据实时流数据检测平台能够满足涉油气企业的要求,可以对涉油气犯罪人的防控对策提供有价值的数据。

1 相关技术

1.1 Storm 框架分析

Storm 拓扑结构如图 1 所示。由图 1 可知,主要功能在于可以构建一个具有高稳定性、高可靠性、分布式的数据实时计算系统,该系统主要被应用于数据资源的实时数据分析领域。Storm 结构中数据基本都以基元组做单元处理,其在注入 Storm 数据处理平台后,须严格按照 Tuple 格式重新进行组装,最后形成一条 Tuple 数据流,进入 Storm 结构的整个数据拓扑模型进行处理。Storm 系统中使用的组件 Topology 是由 2 个 Spout 与 2 个 Bolt 构成。其中,Spout 负责读取、封装与处理数据,再对数据进行转发,Bolt 则用于优化后续的业务逻辑。

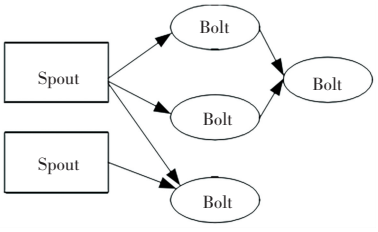

图 1 Storm 拓扑结构图

Fig. 1 Storm topology diagram

1.2 热力图原理

热力图主要采用的 3 种形式分别是基于鼠标单击位置的热力图、基于鼠标轨迹方向的热力图以及基于鼠标内容的单击轨迹热力图等[1]。本研究主要基于最后一种形式,通过该形式可实时记录并显示当前浏览用户的行走轨迹,实时优化整个网页及内容布局[2]。通过利用热力图上绘制的 3 种颜色来区分渲染图的不同覆盖区域,与图里实际的涉犯罪人数分别相对应,从而分析算出实际涉犯罪人群密度,再根据此密度规划出涉犯罪人分布的运动移动图[3]。

1.3 Flume-Kafka 数据收集缓存

Flume 是一种可靠性很高的分布式数据采集、传输工具, 能将数据转化成数据流进行控制,而 Kafka 是一种发布-订阅模式的消息系统,除了具有高吞吐量外,也支持在线实时和离线的数据处理。Kafka 按主题对数据进行分类存储,每个主题可扩展成多个分区,并能将其部署在集群的各个服务器上,以确保数据安全性。Kafka 还支持副本模式,能够设置分区的副本数,通过将 Flume 收集的模拟数据送入已经创建的日志主题中,可指定主题分区数和副本数。

2 实时检测系统架构功能及数据库设计

2.1 系统总体功能设计

系统总体功能架构如图 2 所示,按照功能模块可划分为数据获取、数据分析整理、热力图显示等。

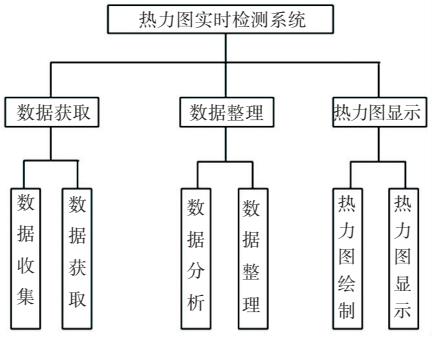

图 2 系统总体功能架构图

Fig. 2 Overall functional architecture diagram of the system

2.2 数据实时分析整理功能

数据分析整理功能设计流程如图 3 所示。图 3 中,将获取的数据源信息发送到 Kafka 集群中进行缓存,依据主题 Kafka 对数据进行维护管理,等待 Storm 集群主动拉取数据进行统计分析,其中,Zookeeper 集群主要负责对 Storm 集群及 Kafka 集群的状态进行维护管理。ZK 中的数据要对接 Storm 集群实时处理框架,Storm 接收到数据后通过 bolt 来进行相应的处理,处理后的数据写入数据库、即 DB 中。

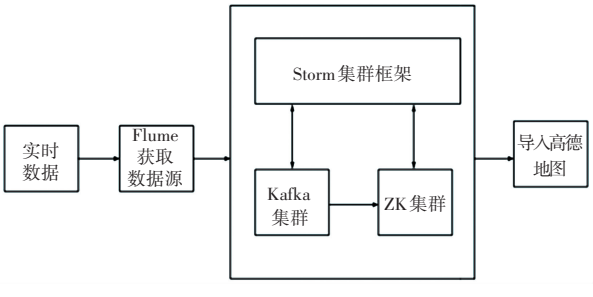

图 3 数据分析整理功能设计流程图

Fig. 3 Flow chart of data analysis and arrangement function design

2.3　热力图的显示功能

热力图显示功能设计流程如图 4 所示。在图 4 中，先对获取的热力图信息进行检测，在获得数据热力点信息及运动目标坐标具体位置后绘制热力图，再在高德地图上创建热力图显示，就可以实时查询区域内的涉犯罪人流量情况[4]。

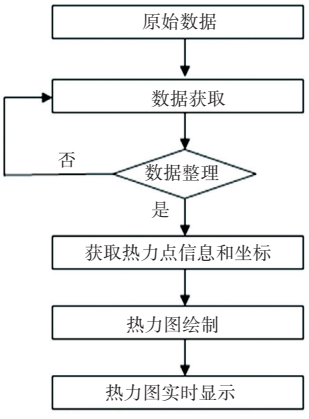

图 4　热力图显示功能设计流程图

Fig. 4　Flow chart of thermal map display function design

2.4　系统数据库设计

2.4.1　数据库 E-R 图设计

E-R 关系图如图 5 所示。MySQL 数据库系统可根据平台数据信息将数据信息进行实体划分，并赋予信息之间相互的联系。

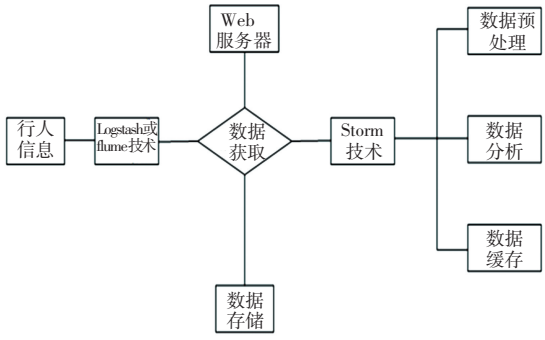

图 5　E-R 关系图

Fig. 5　E-R diagram

2.4.2　数据库表结构设计

主数据表见表 1。由 E-R 关系图，建立与主功能模块相关的包括系统登录数据表、数据信息抓取数据表、数据信息整理数据表、数据推送数据表等在内的数据表。

表 1　主数据表

Tab. 1　Master data table

编号	属性	字段类型	主键	是否空	说明
1	Data_time	Bigint	是	否	获取时间
2	Data_Thermal point	Varchar	否	是	数据整理
3	Data_coordinates	Varchar	否	是	数据分析
4	Data_Thermal point	Varchar	否	是	热力点信息
5	Data_coordinates	Varchar	否	是	热力点坐标
6	Longitude information	Varchar	否	是	经度信息
7	Latitude information	Varchar	否	是	纬度信息
8	Thermodynamic_display	Varchar	否	是	热力图显示

3　实时检测系统的实现

3.1　热力图生成方法

3.1.1　热力点坐标获取

Logstash 信息获取框架如图 6 所示，Logstash 是一个开源的服务器端数据处理管道，可以从不同数据源获取数据，能进行转换并将数据发送到 Elasticsearch 中[5]。数据获取模块主要是对本系统进行人流数据信息的获取，通过 Logstash 技术获取涉犯罪人数据源，再将采集到的涉犯罪人信息传输到系统服务器的数据库进行存储[6]。

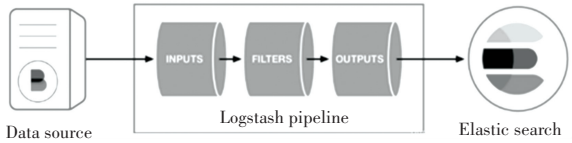

图 6　Logstash 信息获取框架

Fig. 6　Logstash information acquisition framework

3.1.2　热力点的处理

采集到涉犯罪人信息后，将涉犯罪人信息中的数据信息整合到二维坐标点 x，y。因数据信息变化较大，数据间的信息分布并不是很均匀，所以第一时间要对全部热力点数据做预处理[7]，具体如下：

(1)确认所要检测的数据,以及对应的各热力点区域信息的具体位置,查看所要检测数据的热力图区域信息数据信息与边界位置是否重合,并检查导出的数据热力图边界位置是否完全覆盖了此检测区域内所有热力点信息。

(2)对那些即将被检测的目标物体进行准确有效的区分,通过收集热力点等信息进行检测并制作出热力图。

(3)将所需要获取的数据信息发送到 Kafka 集群的缓存主题中,并对其进行实时数据信息缓存。

3.1.3 热力点的生成

得到二维坐标点 x,y 相对具体位置信息后,进一步运算生成热力图。研究给出的生成方法流程具体如下。

(1)获取区域的像元热力值人群聚集对比分析。通过 Storm 技术对所需要采集到的连续热力图大数据进行分析,准确求得出其平均值[8]。利用式(1)计算得出每个区域的热力均值:

$$H = \frac{\sum X_t}{T} \tag{1}$$

其中,H 表示涉油气区热力均值;T 表示所统计的时间段;X_t 表示不同时间涉油气区内像元热力值。

(2)人群聚集密度分析。人群密度对于进行涉油气犯罪监测评价尤为重要,过高的人群密度随时都可能会诱发各种风险,带来一系列潜在社会安全隐患。为了便于涉油气企业更精准分析国内涉油气区的人群规模及其热力点聚集情况,需要计算出不同涉油气区的人群热度指数分布[9]。此处需用到的数学公式如下:

$$P = \frac{H_t}{S} \times k \tag{2}$$

其中,P 表示人群热度指数;H_t 是根据式(1)计算出的不同时间段的热力均值;S 是涉油气区域像元面积;k 为计算系数。

3.2 系统的运行环境

研究开发环境具体见表 2,本检测系统主要是依据 Java 开发语言进行各功能的设计与实现的,是在戴尔 i7-8265U CPU,32 GB RAM 处理器搭载 Windows10 系统环境下进行的。

Java 开发语言具有高简洁性、高稳定性、高安全性、高适应性以及多平台高兼容性等优点,是计算机软件开发的首选语言,编程人员可以从系统直接调用一些常用的语句和函数,减少了软件编程的复杂性。

表 2 开发环境配置表

Tab. 2 Development environment configuration table

功能模块配置	使用工具
编译语言	Java 语言
编译环境	IDEA 2018
服务器	Apache Tomcat 8.5
数据存储库	MySQL8.0
操作系统	Windows10 系统
代码托管	Git 2.8

3.3 系统功能的实现

3.3.1 数据信息获取功能

数据信息获取实现如图 7 所示,本系统实现基础部分就是涉犯罪人数据的获取,该功能可以对本检测系统进行人流数据信息的获取,主要采用 Logstash 技术获取涉犯罪人数据源,再将采集到的信息传输到系统服务器数据库进行存储[10]。

图 7 数据信息获取功能的实现

Fig. 7 The realization of data information acquisition function

3.3.2 数据信息存储功能

数据信息存储实现如图 8 所示,Storm 接收到数据后通过 Bolt 进行相应的处理,将处理后的数据写入 MySQL 数据库中。

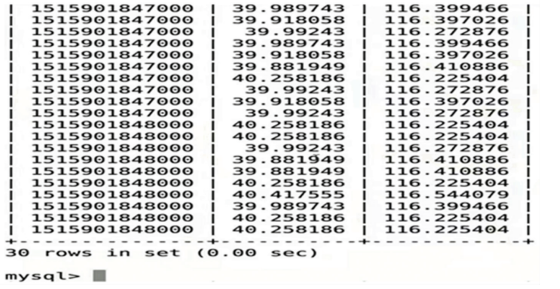

图 8 数据信息存储功能的实现

Fig. 8 The realization of data information storage function

3.3.3 数据实时分析整理功能

数据实时分析整理实现如图 9 所示,此功能主要利用 Storm 技术实现对数据信息的实时分析

整理。

```
mysql> select longitude,latitude ,count(1) from stat where time > unix_timestamp(date_sub(
current_timestamp(), interval 10 minute))*1000 group by longitude,latitude ;
+-----------+----------+----------+
| longitude | latitude | count(1) |
+-----------+----------+----------+
| 116.225404 | 40.258186 |        9 |
| 116.272876 |  39.99243 |        5 |
| 116.397026 | 39.918058 |        5 |
| 116.399466 | 39.989743 |        5 |
| 116.410886 | 39.881949 |        3 |
| 116.544079 | 40.417555 |        3 |
+-----------+----------+----------+
6 rows in set (0.00 sec)
```

图 9　数据实时分析整理功能的实现

Fig. 9　Realization of data real-time analysis and collation function

3.3.4　热力图实时显示功能

热力图实时显示功能主要是根据所获得的热力点信息以及运动目标具体位置去绘制热力图，然后在高德地图上创建热力图的显示。

4　系统测试分析

系统测试环境参见表 2，研究选用的测试方案具体如下。

(1)功能模块测试。数据获取功能模块测试见表 3，对系统中每个功能单元进行详细测试，检测每个需求功能是否能实现。

表 3　数据获取功能模块测试表

Tab. 3　Data acquisition functional module test table

测试情况	预期结果	实际结果
数据收集	利用 Logstash 收集行人数据	收集成功
数据抓取	系统成功显示行人数据信息	显示成功

(2)集成模块测试。数据实时分析整理功能模块测试见表 4，将各角色具有的功能单元模块放在一起，进行检查、测试。

表 4　数据实时分析整理功能模块测试表

Tab. 4　Functional module test table of data real-time analysis and collation

测试情况	预期结果	实际结果
Storm 实时统计	能进行对数据统计分析	符合预期结果
数据分析	能进行数据信息分析	符合预期结果
数据整理	能进行数字信息的分类整理	符合预期结果

(3)验收测试。热力图显示功能模块测试见表 5，将系统交付测试专业人员及用户进行共同测试、使用，测试系统整体运行、使用情况。

表 5　热力图显示功能模块测试表

Tab. 5　Heat map display functional module test table

测试对象	预期结果	实际结果
获取热力点信息	系统能够获取行人的热力点数据信息	符合预期结果
热力值计算	能进行热力值的信息计算	符合预期结果
热力图显示	成果渲染绘制热力图，显示热力图信息	符合预期结果

5　结束语

本文提出的基于 Storm 的涉油气犯罪实时检测系统的总体架构，利用热力图大数据方法对涉油气企业所在区域范围内的人群分布进行研究统计，系统分析了涉油气企业所在区域人群热度指数的问题，此人群热度指数可以很直观地反映区域内人群密度的分布情况，实现了热力值的实时显示。经系统测试，验证了此方法有助于提高数据处理的准确率以及效率，加快了系统的运行时间。

参考文献

[1] 宋燕妮. 基于热力图的展厅客流检测系统的设计与实现[D]. 石家庄：河北科技大学，2020.

[2] 刘霄，陈冲，李得海，等. 室内位置服务云构建及其人流监控应用[J]. 测绘科学，2020，45(10)：16-21.

[3] 张海林. 基于百度热力图的人口活动数量提取与规划应用[J]. 城市交通，2021，19(03)：103-111.

[4] 饶颖霞，李响. 基于人口热力图和土地利用分类实现人流量空间分布的精确提取[J]. 测绘与空间地理信息，2019，42(09)：36-39.

[5] 冉桂华，杨晔轩，殷浣益，等. 一种热力图的景区人流量动态监测方法[J]. 计算机与数字工程，2018，46(11)：2329-2332.

[6] 张蓝天，杨剑，王光霞，等. 一种室内空间结构约束的热力图生成方法[J]. 测绘科学技术学报，2018，35(05)：533-539.

[7] 李娟，李苗裔，龙瀛，等. 基于百度热力图的中国多中心城市分析[J]. 上海城市规划，2016(03)：30-36.

[8] 杨宇，徐万明. 基于 Storm 技术的实时数据处理平台研究与实现[J]. 电脑与电信，2021(Z1)：51-54.

[9] 周文捷. 基于 Storm 的 Web 日志分析系统的设计与实现[D]. 北京：北京邮电大学，2018.

[10] 杨振凯，李响，杨飞. 一种面向百万级数据的热力图生成算法[J]. 测绘科学，2018，43(08)：85-89.