基于ES的纪检监察大数据搜索引擎的研究与实现

摘要

第1章 绪论

1.1 本论文的背景和意义

1.2 国内外研究现状

1.3 主要研究内容

1.4 论文组织结构

第2章 相关技术简介

**2.1 搜索引擎技术**

2.1.1 搜索引擎定义

2.1.2 搜索引擎分类

2.1.3 搜索引擎的工作原理

（数据抓取、分词及建立倒排索引、搜索结果排序）

2.1.4 elasticsearch介绍

**2.2 ETL技术**

Logstash、kafka

**2.3 数据存储技术**

TiDB、Redis、Hibernate、Spring Data JPA

**2.4 后端技术**

Java、Spring Boot、Spring Cloud

# 第3章 纪检大数据搜索引擎系统需求分析

## 3.1 纪检大数据的特点

（1）数据来源广泛

纪检业务的数据来源十分广泛。为了满足纪检业务的需要，纪检部门通常会收集汇总及其直属及下属各地区各部门的电子政务数据，而下属部门的数据又是汇总自其下属乡镇、街道委员会相关单位，整个数据来源根据职能单位的级别近似形成了一个树形的结构。以沈阳市纪委为例，截至目前，该部门拥有从市、区县（市）两级共计835家政府职能部门[1]采集的数据22亿余条，涉及数据库总量达上千余个。这些数据通过人工拷贝、专网接口、专线网闸等方式流入纪委内部数据系统。

（2）数据类型多样

纪检数据汇集了多类异构的数据，其既有关系性数据库的结构化数据，也有类似XM形式的半结构化数据以及图片等非结构化数据，数据类型多种多样。以沈阳市纪委为例，其现有数据的类型涉及MySQL、Oracle、Access、Microsoft SQLServer等几乎市面上常用的各种数据库以及Excel、Word、GIS、报表等非结构化数据。

（3）数据结构散乱

大多数政务部门并不具备传统互联网公司那样大规模且专业的技术团队，其数据库表、字段的命名方式多种多样，例如使用拼音首字母、使用字母数字编码等等，这给阅读数据的人员带来很大的阻碍。而且，很多数据都是由Excel汇总导入至数据表，由于Excel在数据校验方面的不足，使得大量数据未经检查便被导入数据库，数据的格式不能保证统一。例如“2019-03-01”写法有“2019年3月1号”、“3-1-2019”、“二零一九年三月一日”等等。这些数据在存储时，便不能按照数字类型、日期类型等去存储，只能使用字符串类型，这也增加了数据错误、数据重复现象出现的概率。

（4）数据更新方式单一

## 3.2 系统用户的特点

## 3.3 系统所面临的挑战

（1）数据存储——大规模库表下的索引设计

（2）数据同步——非规范数据表与搜索引擎之间的数据同步

（3）数据搜索——海量数据下的搜索方案

（4）数据展示——大规模库表下搜索结果的展示

（5）搜索管理——功能完备的搜索管理模块的设计

## 3.4 系统需要解决的问题

### 3.4.1 数据索引——大规模库表下的索引设计

Elasticsearch的索引为分布式的分片结构，每个分片相对独立，因此搜索请求会广播到索引的每个分片，每个分片并发的查询自己分片内的数据。各分片的查询结果有Elasticsearch汇总后返回给用户。

在传统的互联网应用中，索引的设计方式多为针对每个实体类（数据表）建立一个索引，索引各字段的名称与实体类（数据表）中的字段名称一一对应。这种查询方式适合单索引的查询，即每次查询请求都是在某一个具体的索引下进行。

而纪检数据面对的是成千上万的数据表，且表与表之间的结构、字段各不相同。查询方式为全库全文搜索，即每次查询针对的是所有库表所有字段进行搜索。在这种情况下，若使用传统的“一表一索引”的方式，假设库中有5000个数据表，每个表2个分片。当查询请求到来时，Elasticsearch会将请求广播至10000个分片中，这10000个分片会去一同竞争使用处理器的资源，使得处理器将大量时间花费在线程上下文的过程中，从而严重拖垮处理器及Elasticsearch的性能。

面对这样的情况，需要设计高效、通用的索引结构，使其不仅能够容纳不同类型的数据库表，还要确保高效的数据查询效率。

### 3.4.2数据同步——非规范数据表与搜索引擎之间的数据同步

纪检数据具有数据库表多、数据量大及结构不规范的特点。针对库表多、数据量大的特征，需要设计便捷的自动化数据同步方案，并结合优秀的开源数据抽取工具，提高数据读写效率，减少数据的丢失率。针对结构不规范的特点，需要充分考虑到数据库表中的主键缺失、格式不一的现象，尽可能降低搜索引擎中数据重复、数据错误情况的发生率。

### 3.4.3 数据搜索——海量数据下的搜索方案

纪检数据搜索引擎针对的是大规模库表的结构化数据，在设计搜索方案时，要需要考虑到方案的通用性，确保其适用于系统下的绝大多数库表结构。同时，数据各表中也存在一定数量的无用字段，例如业务的预留空字段、标志位字段、测试字段等等，这些字段在一定程度上会加重搜索的负担，影响搜索结果的准确率，因此，在搜索时，如何设计灵活的规则，使得搜索引擎可以跳过无需检索的字段，这也是一项需要解决的问题。

### 3.4.4 数据展示——大规模库表下搜索结果的展示

搜索展示模块是与用户直接接触的模块，也是用户最为关心的一个模块。搜索展示主要需要考虑两部分：一为搜索结果的排布，二为搜索结果的易读性。

针对搜索结果的排布，需要充分利用库、表、字段信息的导向作用，将搜索结果进行合理的分类汇总。由于面对的是结构化的数据，若是像百度搜索那样以“标题—内容”的方式将搜索结果一条条的罗列，将会使页面十分散乱。下图3-1是本人所在部早期搜索引擎的查询结果截图，可以看到，在搜索关键词“项目”时，表中的“发放对象”、“资金种类”、“地点”、“发放时间”四个字段及其字段值被拼成一个文本作为一条搜索结果的标题，而这763页搜索结果没有经过合理的分类排布，只是单纯的罗列出来，这对后面的搜索定位造成了很大的苦难。



图 3-1 搜索结果示例图

二为搜索结果的易读性，纪检数据中，库、表、字段的命名几乎都是以拼音首字母命名，对于不熟悉数据业务的人来说，根本无从知晓这些属性所代表的意思。因此，对于这些晦涩的命名，还需要设计一个“数据字典”映射模块，存放这些命名计它们所对应的中文名称。在搜索引擎返回搜索结果前，先从字典模块做一次映射操作，将易读性好的名称展现给用户。

### 3.4.5 搜索引擎管理——便捷的搜索引擎管理模块的设计

通过3.4.1—3.4.4小节的内容，可以了解到纪检搜索引擎涉及到索引管理、数据同步管理、搜索参数控制、数据字典管理等出错率高且耗时耗力的操作。因此，针对搜索引擎系统，需要设计一个操作便捷、功能实用的搜索引擎管理模块，向操作人员屏蔽到底层复杂的实现细节，将重复、繁琐的工作尽可能实现自动化，以提升系统的效率，减少错误的发生。

## 3.5 系统功能概述

按照模块化的思想，将系统

第4章 纪检大数据搜索引擎系统架构

4.1 系统架构图

4.2 数据库设计

第5章 纪检大数据搜索引擎系统设计

5.1 Eureka注册中心模块

5.2 Zuul网关模块

5.3 数据字典模块

5.4 搜索引擎索引管理模块

5.5 数据同步模块

5.6 搜索引擎模块

第6章 总结和展望

第7章 参考文献

[1]中国青年网.沈阳：用大数据深挖“不正之风” 管住身边腐败[EB/OL].https: //baijiahao.baidu.com/s?id=1618821229579984841&wfr=spider&for=pc,2018-12-03.