# 本科论文检测报告（简明版）

报告编号：PL-20190504-DCE4CF38-JM　　检测时间：2019-05-04 23:48:58

题　　名：15310320108潘小宇日志采集存储系统的设计与实现
作　　者：潘小宇
检测范围：
☑中国学术期刊数据库　☑中国学位论文全文数据库　☑中国学术会议论文数据库
☑中国学术网页数据库　☑中国专利文献数据库　☑中国优秀报纸数据库

## 检测结果

总相似比：4.25%

检测字数：8557
参考文献相似比：0.00%　/　排除参考文献相似比：4.25%
可能引用本人已发表论文相似比：0.00%　/　辅助排除本人已发表论文相似比：4.25%
单篇论文最大相似比：0.84%

相似片段分布图


注：绿色区域为参考文献相似部分，蓝色区域为本人已发表论文相似部分，黄色区域为本人学位论文相似部分，红色区域为其他文献相似部分

相似文献列表

| 序号 | 相似比 | 题名 | 作者 | 文献类型 | 来源 | 发表时间 | 是否引用 |
|---|---|---|---|---|---|---|---|
| 1 | 0.84% | 时序数据库 | | 学术网文 | 百度百科 | 1900-01-01 | 否 |
| 2 | 0.68% | 基于过程管理的硕士学位论文质量保障制度研究——以N大学为例 | 仇雪萍 | 学位论文 | 江西师范大学 | 2018-05-01 | 否 |
| 3 | 0.54% | 家校沟通信息平台的应用现状分析与对策研究——以日照市岚山区S小学为例 | 黄丽丽 | 学位论文 | 曲阜师范大学 | 2017-12-10 | 否 |
| 4 | 0.43% | 实验室预约管理系统设计与实现 | 王爱春 | 学位论文 | 江苏科技大学 | 2018-06-01 | 否 |
| 5 | 0.26% | 村级治理分权化改革策略——浙江省溪村分权治理模式案例研究 | 杨利登 等 | 会议论文 | 浙江省公共管理学会2008年年会 | 2008-11-14 | 否 |
| 6 | 0.23% | 山的启示：美国自然文学中山的原型、美学价值及伦理意蕴 | 魏薇 | 学位论文 | 首都经济贸易大学 | 2018-05-11 | 否 |
| 7 | 0.23% | 某核电研究所共性项目管理系统设计与实现 | 刘晓满 | 学位论文 | 华中科技大学 | 2017-12-21 | 否 |
| 8 | 0.20% | 瑞翔能源公司客户关系管理研究 | 齐炜煜 | 学位论文 | 华北电力大学(北京) | 2018-03-01 | 否 |
| 9 | 0.18% | 201410290387.9 | 广东睿江科技有限公司 | 专利论文 | 中文专利全文数据库 | 2014-06-22 | 否 |
| 10 | 0.18% | 视频网站自制综艺节目社会化传播过程及策略研究——以《奇葩说》为例 | 李夏薇 | 学位论文 | 成都理工大学 | 2017-05-01 | 否 |
| 11 | 0.18% | 国内科技类图书评价机制构建研究 | 郝玉龙 等 | 期刊论文 | 《科技与出版》 | 2015-02-08 | 否 |
| 12 | 0.16% | 视频点播系统的研究和实现 | 李宁 | 学位论文 | 北京邮电大学 | 1999-03-01 | 否 |
| 13 | 0.15% | 支持医疗过程管理的电子病历系统研发及有意义应用研究 | 张小光 | 学位论文 | 浙江大学 | 2012-06-03 | 否 |

原文

成都东软学院本科毕业设计报告系别：　信息与软件工程系专业：　软件工程
班级：　软件工程15201

学生姓名： 潘小宇
学生学号： 15310320108
指导教师： 吴平贵
二〇一九年六月

摘要

随着互联网的发展，网络基础设置的改善，人们对于网络应用提出了更高的要求，于是各大互联网公司纷纷采用分布式架构以支持越来越高的并发量。由分布式架构引出来的日志问题也越来越突出。在单机环境下可以正常使用的日志解决方案在分布式环境下表现的有心无力。因此本系统的目的在于提供一个更加友好的日志采集和存储方案。

本文分几部分阐述了基于Java开发，涉及消息中间件，socket短连接等技术的日志采集系统的结构和设计实现过程，实现了日志发送、日志存储、日志备份、日志实时浏览等功能。支持横向扩展，可以轻松实现分布式部署。同时充分考虑了宕机情况的出现，尽可能地保证系统核心功能的正常运行。

根据本文设计思路，最终开发出一个健壮、稳定的日志采集系统。

关键词：日志，消息中间件，Java，Jetty

Abstract

With the development of the Internet and the improvement of network infrastructure, people put forward higher requirements for network applications, so major Internet companies have adopted distributed architecture to support higher and higher concurrency. The problem of log caused by distributed architecture is becoming more and more prominent. Logging solutions that can be used normally in stand-alone environments are powerless in distributed environments. Therefore, the purpose of this system is to provide a more friendly log collection and storage scheme.

This paper describes the structure and design process of the log acquisition system based on Java, which involves message middleware and socket short connection technology. It realizes the functions of log sending, log storage, log backup and real-time log browsing. Horizontal scaling is supported and distributed deployment can be easily implemented. At the same time, the outage situation is fully considered to ensure the normal operation of the core functions of the system as far as possible. According to the design idea of this paper, a robust and stable log acquisition system is finally developed.

Key Words: log, JMS, Java, Jetty

目录

第 1 章绪论

目前各大互联网公司普遍采用分布式架构。这种架构的有点在于可以提供更加健壮、容量更大的服务。但是这种架构造成了各种服务分布在不同的机器上，一次请求可能被拆散成若干部分，不同部分在不同的机器上执行，因此产生的日志也七零八落，这给流程追踪、项目调试增加了不小的难度。因此，本系统旨在建立统一的日志采集存储平台，使得开发者可以便利的查看、下载日志信息。

第 2 章日志系统概述

2.1.日志的定义

日志，软件或**硬件**在**运行过程中产生的事件记录**，一般包括**日期、时间、使用者**、**操作**描述**等信息。**

2.2.日志的作用及重要意义

日志在各种软件中广泛存在，可以用于数据恢复、软件调试、数据挖掘、安全防护等方面。当软件或服务宕机之后，运维人员可以根据宕机前的日志，迅速将服务恢复到宕机前的状态，减少损失；开发人员通过追踪日志，可以发现软件的各种bug；通过对日志记录的分析，可以提供诸如“你可能喜欢。。”等推荐功能；

2.3.日志系统的目的

由于目前的软件架构，一次请求中产生的日志可能分布在不同的机器上。这给日志追踪、存储带来了不小的苦难。因此，本系统旨在提供一个简单、轻便的日志采集存储方案，可以将不同机器产生的日志统一存储，并且提供实时浏览和下载功能。

2.4.相关研究

目前市面上主流的日志解决方案为ELK，ELK是三个开源软件的缩写，分别表示：Elasticsearch , Logstash, Kibana。ELK支持日志的采集、统一存储、日志分析、报警等功能，支持多种数据源，可采用多种架构搭建。同时由于组件之间的隔离性，可以轻松实现分布式和集群部署。

ELK的缺点在于：一、过于复杂，相关组件配置难度较高；二、过于重量级，其中的Logstash占用内存和cpu过高；三、对于Java项目支持不够友好。

第 3 章系统概述

3.1.原理

软件客户端发送日志消息到消息中间件，采集模块从消息中间件中拿取数据，存入时序数据库，并且定时备份日志文件到本地文件系统，同时利用WebSocket等技术实现日志的实时浏览。

3.2.功能

3.2.1 日志发送

软件客户端通过本系统提供的工具发送日志消息到指定的消息中间件，且在一定程度上保证有序。

3.2.2 日志收集

本系统通过监听消息中间件，即使拿取日志消息，并且按照一定的规则进行持久化存储。

3.2.3 服务监控

本系统提供Web模块以监控采集模块以及相关服务运行状况。

3.2.4 日志备份

本系统可以定期将日志持久化到文件，存储在本地文件系统。

3.2.5 日志文件下载

本系统提供接口以下载各个日志采集模块备份的日志文件，同时记录所有下载操作以备后期查看。

3.2.6 实时日志浏览

本系统提供接口以实时查看消息中间件接收到的日志消息，同时支持简单的规则过滤，并保证在一定程度下日志消息有序，以便开发人员调试。

3.2.7 自定义消息中间件

系统提供对应的解决方案以替换系统默认的消息中间件依赖，以尽可能减少耦合，贴近使用方的业务需求。

3.2.8 登录日志记录

系统将会记录每一次管理员登录日志，尽可能防范社会工程学攻击。

3.2.9 报警功能

系统对于部分情况，例如短时间内大量error日志的产生，提供相应的邮件报警功能，以便运维人员能及时发现系统漏洞。

3.2.10 数据统计

系统提供基础的数据统计功能，记录系统运行过程中收集到的日志情况；以便开发运维人员关注日志产生方的相关运行情况，及时做出相应的技术调整，如降低日志输出级别以减少日志量。

3.3.系统开发环境

Jdk1.8+,ActiveMQ5.15.8,maven3.3.9,Influxdb,IntelliJ IDEA-2018

第 4 章系统设计

4.1.关键技术介绍

4.1.1 消息中间件

用在应用软件和底层操作系统之间的一种软件服务，主要用于多个软件服务之间的消息通讯，实现跨进程通讯，同时可降低各个软件服务的耦合度。一般遵循JMS（Java Message Service）规范。

常见的消息中间件有ActiveMQ、RocketMQ、RabbitMQ、Kafka、Redis等。

本系统默认采用ActiveMQ，同时提供可替换成其余消息中间件的解决方案。

4.1.2 Vert.x

Vert.x是一套异步框架，提供诸如socket、Web等功能，拥有较高的性能。

本系统主要用于采集模块和Web模块之间的简单通讯。

4.1.3 Jetty

Jetty 是一个开源的servlet容器，它为基于Java的Web应用——例如JSP和servlet——提供运行环境。可以支持嵌入到普通Java应用中。

本系统用于Web模块的servlet容器，以嵌入式Web容器的方式减少系统的部署复杂度。

4.1.4 Spring

Spring是一个开放源代码的设计层面框架，它解决的是业务逻辑层和其他各层的松耦合问题，因此它将面向接口的编程思想贯穿整个系统应用。

本系统大多数模块均是基于Spring构建。

4.1.5 Maven

Maven是Java的包管理器，提供依赖管理、项目打包功能。

本系统使用Maven管理依赖、打包项目，以尽可能减小项目体积。

4.1.6 Quartz

Quartz是一套Java的开源任务调度框架，提供定时任务功能。

本系统用于心跳包管理、服务监控、日志文件备份等功能。

4.1.7 时序数据库

时序数据库全称为时间序列数据库，英文缩写为TSDB。时间序列数据库主要用于指处理带时间标签（按照时间的顺序变化，即时间序列化）的数据，带时间标签的数据也称为时间序列数据。

本系统中主要用于临时存储日志消息，以保证多日志采集模块时日志消息备份时序性。

4.2.设计指导思想

充分利用设计模式和aop思想，以达到低耦合、高复用、可自定义的效果。

4.3.系统架构

图 4.1 整体架构图

图 4.2 程序流程图

4.3.1 Client

供日志产生端使用的工具模块。思路为自定义日志组件消息通道，目前支持log4j框架；以及直接读入程序日志文件。该模块需要对现有Java项目进行一定程序的修改。

4.3.2 MQ

消息中间件依赖。用于中转日志以及服务状态监控等数据通信领域。

4.3.3 Server

日志采集模块。负责监听消息中间件，将日志消息持久化到TSDB，同时负责按照一定的周期备份日志文件到本地文件系统或其他存储系统。

该模块支持横向扩展，只需要多台机器保持消息中间件等相关配置一致即可。

4.3.4 Web

图形化界面模块。B/S架构。该模块以Web界面方式提供监控Server模块状态、日志文件下载、系统报警和实时日志浏览等功能。

该模块非必需模块，缺失或宕机不影响日志采集存储功能。

4.3.5 TSDB

时序数据库。用于临时存储日志消息，以解决使用消息中间件通讯中由于网络抖动等原因导致的顺序错位。

4.4.通信方案

本系统使用了三种通信方式

1、JMS——主要用于Client-MQ-Server。

2、HTTP协议——Web模块对用户提供服务。

3、Socket——Web和Server直接通信，主要用于文件传输。

4、WebSocket——用于Web模块浏览实时日志

图 4.3 Client-MQ-Server

图 4.4 Web-User

图 4.5 Client-Web

第 5 章系统可行性分析

5.1.技术可行性分析

5.1.1 日志发送方式

引入消息中间件作为通信桥梁，使用队列模式　优点 日志产生方和日志采集方解耦 日志采集方可以横向扩展 缺点 采集的日志无法保证有序 备份的日志文件依旧散乱

5.1.2 已有Java程序如何接入本系统

系统提供基于slf4j的工具类，只需要引入工具，修改项目配置文件即可接入系统 系统提供基于文件监控的工具，只需要引入工具，修改工具配置即可接入本系统

5.1.3 如何保证日志消息有序

（1） 难点

由于网络为不可靠资源，无法保证到达目的地顺序和发送顺序一致 且日志采集方采用横向扩展方案，代表同一段时间内的消息会被分散到多个机器，无法使用内存排序

（2） 解决方案

引入时序数据库，日志采集方将采集到的日志存储到时序数据库，由时序数据库负责排序；日志采集方间隔一段时间再从

数据库获取数据；以此保证在一定时间段内的日志消息有序。

5.1.4 如何实现实时日志浏览

（1） 难点

目前日志采用的是消息中间件发送，使用queue模式；一旦日志消息被消费，则不能再次利用 且Web模块采用B/S架构，常规B/S架构为半双工通信，无法实现S端推送消息到B端

（2） 解决方案

日志采集方消费日志后再次将日志消息以topic模式发送给消息中间件 Web模块前端采用WebSocket与后端连接，实现全双工通信 Web模块订阅消息中间件，一旦接收到日志消息，则通过WebSocket广播到前端

5.1.5 如何备份及下载日志文件

日志采集方定期从时序数据库抓取数据备份到本地服务器 Web模块通过socket与日志采集方通信，获取文件列表及文本二进制流

5.1.6 如何降低耦合

（1） 耦合来源

对于消息中间件的依赖——如果实现中大量使用ActiveMQ独有代码，将会导致无法替换消息中间件 对于时序数据库的依赖

（2） 解决方案

消息中间件依赖 采用工厂模式，定义消息中间件发送、接收等操作等接口规范；只要提供实现了该规范的依赖文件，即可实现消息中间件替换 时序数据库依赖 时序数据库遵循JDBC标准，只要代码中不涉及某个时序数据库独有的操作，只需要修改配置信息，即可轻松替换时序数据库

5.1.7 如何监控系统运行状况

日志采集模块定时发送心跳包到消息中间件，Web模块通过订阅消息查看日志采集模块状态

5.2.安全性分析

本系统面向开发运维人员设计，并最终运行在内网环境中。考虑到用户本身的专业性，因此不多考虑安全性问题。

5.2.1 XSS攻击

使用Servlet规范以及装饰模式防范XSS攻击，同时尽可能减少HTTP协议使用场景。

5.2.2 DDOS攻击

本系统设计思路为面向内网环境，互联网中的DDOS攻击已经被外界环境隔离。因此主要考虑当日志采集对象产生的日志过多导致的DDOS。

主要解决思路为：

1、消息中间件需要支持集群模式，以支持更高的流量，ActiveMQ已经实现该功能。

2、日志采集模块必须支持横向扩容，以充分消化来自中间件的海量消息。

5.2.3 社会工程学攻击

对于系统使用方管理不慎导致的账号密码泄露问题，系统提供登录日志记录以及日志文件下载记录。尽管不能防范社会工程学攻击，但是能为后期排查提供一定的帮助。

5.3.健壮性分析

5.3.1 服务宕机

（1） 消息中间件

消息中间件宕机将会影响Client发送日志，此时Client需要将日志缓存到本地，等待消息中间件恢复之后将日志消息重新发送。

(2) Client

Client宕机即代表终端服务宕机，此时不会产生日志，且终端服务宕机应该有另外的报警方案，不在本系统设计考虑范围内。

(3) Server

Server支持横向扩展，单一机器宕机只会影响日志采集效率，不会导致系统全线崩溃。

(4) Web

该模块为非必需模块，宕机不会影响日志采集功能，因此不在考虑范围内；建议使用额外的方案以监听该模块状态，例如定时发送HTTP请求即可判断模块运行状态。

(5) TSDB

时序数据库宕机将会导致Server无法持久化，此时Server将会把采集到的日志返还到消息中间件，不采用缓存方案，防止此时Server宕机导致缓存的日志丢失。

5.3.2 Bug反馈

项目托管在github，可通过github反馈Bug。

第 6 章系统实现

6.1.数据库设计

图 6.1 Web模块

图 6.2 Server模块

图 6.3 时序数据库

6.2.关键代码

6.2.1 日志采集

(1) Client

图 6.4

(2) Server

图 6.5

6.2.2 心跳维持

图 6.6

6.2.3 文件传输

图 6.7

6.2.4 日志备份

图 6.8

6.3.源码地址

本系统已托管在github，项目地址为：https：//github.com/inkroom/log-colleage/

第 7 章结论

经过上述的论述，本系统可以运行于Windows、MacOs、Linux操作系统中，可以轻松结合到现有项目中，可以稳定、持续、健壮地采集日志。同时在系统运维、项目调试等方面提供一定的帮助。

本系统为降低耦合付出了不少努力，因此本系统具有一定的扩展性，允许替换部分组件，如消息中间件；以便符合现在企业开发中要求使用的工具与自身技术栈以及业务相契合的现况。

但是本系统依然有许多不足和缺点，例如日志下载无法支持高并发，实时日志不能分类查看，邮件报警过于死板。

尽管如此，对于规模不大，要求不高的项目，本系统足以应用于生产环境。

致谢

我之所以设计本系统来源于我的学习和工作经历。最早我在实验室学习时，就曾经部署过分布式项目。部署成功让我非常开心，但旋即我就发现了一个很严重的问题——我没办法看到一个完整的日志消息。于是我通过上网查资料了解到了ELK，并尝试在本地环境中搭建。尽管最终搭建成功了，但是也因此发现了一些问题。由于那些问题解决起来比较麻烦，我于是萌生了自己写一个日志采集系统的想法。

我在大三的时候就动手写了一个日志采集工具，算是毕业设计的前身。那是一个简单到不能称之为系统的程序。在又经过一段时间的学习之后，终于设计出了一个较为成熟的设计。这次完成毕业设计的过程中，充分实践了我在大学中学习到的各种知识，包括但不限于设计模式、任务流程、项目管理。设计过程中也参考了许多成熟的开源项目，从中学习到了很多

系统架构的知识，对于我以后的学习和工作有着不小的帮助。

当然，由于我知识上的局限性，本设计不可避免的有着一些缺点和不足。对此我打算继续充实自己，并在以后的学习和工作中继续优化和改进本设计。

最后，我想对在整个毕业设计过程中，给予我悉心指导和耐心帮助的吴平贵老师和罗频捷老师表示感谢。同时，也感谢本设计中使用到的开源项目——Spring、Jetty、ActivmeMQ等等——的作者们。

参考文献

[1] 程序人生ly. Web项目嵌入Jetty运行的两种方式(Jetty插件和自制Jetty服务器)[EB/OL]. https://www.cnblogs.com/lylife/p/5670396.html, 2016-07.

[2] 路慎强，苏卫，牟菁等. 胜利云平台日志数据采集方案设计与实现[J]. 电子技术与软件工程，2018，18：180～182.

[3] 刘尧，宁芊. 基于消息中间件的信息系统数据传输与同步设计[J]. 人民长江，2016，18：106～113

[4] 倪炜. 分布式消息中间件实践 [M]. 北京：电子工业出版社，2018：120-216.

[5] 千里码万里行，SpringMVC整合Thymeleaf模板[EB/OL]. https://blog.csdn.net/hustwht/article/details/79129682. 2018-01.

[6] 姚攀，马玉鹏，徐春香. 基于ELK的日志分析系统研究及应用[J]. 计算机工程与设计，2018，07：1000-7024.

[7] 张秀云，深入浅出ELK[J]. 网络安全和信息化，2018，08：62～67

[8] Joshua Bloch. Effective Java[M]. 北京：电子工业出版社，2018：193-222.

[9] Marcin Bajer. Building an IoT Data Hub with Elasticsearch, Logstash and Kibana[A]. 2017 5th International Conference on Future Internet of Things and Cloud Workshops: FiCloudW 2017, Prague, Czech Republic, 21-23 August 2017[C]. 2017:63-68.

[10] Ruangsak TRAKUNPHUTTHIRAK; Yen CHEUNG; Vincent C.S.LEE. Conceptualizing Mining of Firm’s Web Log Files[J]. Journal of Systems Science and Information，2019，06：489-510.

成都东软学院

毕业设计（论文）原创承诺书

1、本人承诺：所提交的毕业设计（论文）是认真学习理解学校的《毕业设计（论文）工作规范》后，在教师的指导下，独立地完成了任务书中规定的内容，不弄虚作假，不抄袭别人的工作内容。

2、本人在毕业设计（论文）中引用他人的观点和研究成果，均在文中**加以注释或以参考文献形式列出，对本文的研究工作做出重要贡献的个人和集体**均已在文中注明。

3、在毕业设计（论文）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

4、本人完全了解学校关于保存、使用毕业设计（论文）的规定，即：**按照学校要求提交论文和**相关材料**的印刷本**和**电子版本；同意学校保留**毕业设计（**论文**）**的复印件和电子版本，允许被查阅和借阅**；**学校可以采用影印、**缩印或其他复制手段保存毕业设计（论文），**可以公布**其中**的全部或部分内容。**

5、本人完全了解《毕业（设计）论文工作规范》关于“学生毕业设计（论文）出现购买、**他人代写、或者**抄袭、**剽窃等作假情形的，取消其学位申请资格；已经获得学位的，**依法撤销其学位。**取消学位申请资格或者撤销学位**者，**从处理决定之日起3**年内，学校不再接受学生学位申请”的规定内容。

6、本人完全了解《学生手册》中关于在“毕业设计（论文）等环节中被认定抄袭他人成果者”不授予学士学位，并且“毕业学年因违纪受处分影响学位的学生不授予学士学位，并且无学士学位申请资格”的规定内容。

以上承诺的法律结果、不能正常毕业及其他不可预见的后果由学生本人承担！

学生本人签字：年月日

说明：

1.送检文献总字数=送检文献的总字符数，包含汉字、非中文字符、标点符号、阿拉伯数字（不计入空格）

2.总相似比=送检论文与检测范围全部数据相似部分的字数/检测总字符数

3.参考文献相似比=送检论文与其参考文献相似部分的字数/检测总字符数

4.辅助排除参考文献相似比=总相似比-参考文献相似比

5.“单篇文献最大相似比”：送检文献与某一文献的相似比高于全部其他文献

6.“是否引用”：某一相似文献是否被送检文献列为其参考文献