北京理工大学

本科生毕业设计（论文）开题报告

**学 院：** 软件学院

**专 业：** 软件工程

**班 级：** 08111401

**姓 名：** 黄复贵

**指导教师：** 马锐

**校外指导教师：**

二○一八年三月二日

Contents

[1 选题内容 3](#_Toc507790247)

[2 研究方案 3](#_Toc507790248)

[2.1 选题主要任务 3](#_Toc507790249)

[2.2 技术方案分析与选择 3](#_Toc507790250)

[2.3 技术实施条件 4](#_Toc507790251)

[2.4 目前主要问题 5](#_Toc507790252)

[2.5 预期研究目标 5](#_Toc507790253)

[3 课题计划进度表 5](#_Toc507790254)

[4 参考文献 5](#_Toc507790255)

# 选题内容

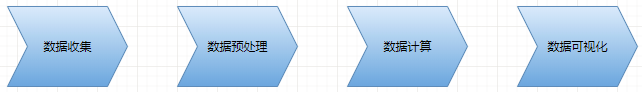
论文研究课题为《基于电商日志的大数据实时流处理平台设计与实现》，旨在解决当今移动互联网高速发展的今天，如何高效率、有深度地利用并分析大数据量级电商日志的价值问题，实现电商日志数据隐藏价值的最大化利用，从而为商业决策提供有前导性意义的支持，推动社会的可持续发展。

# 研究方案

## 选题主要任务

通过打造一个大数据实时流处理平台，根据已有的由一些购物网站提供的日志或自主生成的日志，模拟出实时产生的电商日志，对其中的信息进行统计分析并可视化输出，为商业决策提供依据。例如统计销售总金额、搜索热词排行、活跃城市分布等。

其中软件系统涉及的功能需求包括：数据收集功能，搭建数据流平台，实现实时日志数据的收集；数据预处理功能，根据后续进行数据分析和统计的需要，对原始数据进行清洗和过滤；数据计算功能，对数据进行有目的性的统计计算；数据可视化功能，将统计计算后的数据以可视化的方法进行展示。



## 技术方案分析与选择

伴随课题产生的电商日志大数据实时流处理平台，是一个带可视化界面的后台应用，整个应用会部署到云服务器上。前端可视化使用阿里云开源可视化工具dataV进行构建，后台使用Go进行搭建。软件系统整体架构如下：



在大数据的应用中，数据的采集是后续所有数据处理步骤的前提，因此，采集数据是重要的第一步。

实时数据是即时数据，是由电商应用实时产生的数据，距离当前时间可能只有1小时或1分钟，大多数应用产生的日志都会追加到文件里，因此数据的采集从文件开始。采集工具采用Cloudera提供的分布式海量日志收集组件Flume，以监听文件的方式进行采集，一旦有日志数据追加，程序就会收集。

由Flume收集到的日志数据会进入消息队列Kafka，Kafka主要用于解耦数据处理与数据采集过程，同时起到缓冲数据的作用，避免因为下游Spark Steaming临时出现故障导致错失数据的情况。

得益于Flume、Kafka都是天生的分布式组件，因此可以做到弹性扩容提高数据收集的能力，同时也提供了在个别节点故障下程序仍然正常运行的高可用能力。

进入Kafka的数据，会被Spark Streaming拉取进行有目的性的实时计算，例如统计销售总金额、消费男女比例、搜索热词排行、活跃城市分布等。计算程序采用Scala语言编写，计算作业会部署在Spark集群上，理论上可以做到计算能力无限横向扩展。

计算后得到的结果，会按照一定格式存储在Redis中，方便与前端可视化做交互。可视化dataV通过定制化api的方式访问Redis中计算好的数据，这样就能呈现最后计算得到的结果，输出数据的价值。

## 技术实施条件

该系统的实施需要软硬件的配合，软件需要JDK、Flume、ZooKeeper、Kafka、Spark、Redis等、硬件则需要至少3台（云）服务器的支持，具体版本如下：

|  |  |
| --- | --- |
| 名称 | 版本 |
| >=3台（云）服务器 | CentOS 7.3 64位 |
| JDK | java version "1.8.0\_151" |
| Flume | 1.8.0 |
| ZooKeeper | 3.4.10 |
| Kafka | 1.0.0 |
| Spark | 2.2.1 |
| Redis | 4.0.8 |

## 目前主要问题

首先是服务器资源缺乏，项目中会用到消耗内存比较多的组件，CPU计算能力、内存资源等得不到保证。

再者是对Spark Streaming大数据技术掌握还不到位。例Spark配置优化、Spark RDD使用优化、Spark 运行细节等，这样在编写Spark Streaming计算作业时可能会遇到问题。

最后是对阿里云开源可视化工具dataV的使用还不了解，https的api接入还没尝试。

## 预期研究目标

实现上述设计图中的大数据实时流处理平台，包括从数据采集、预处理，到统计计算入库并最后可视化呈现的全过程。与此同时，系统设计要遵循高内聚、低耦合，具备可扩展性的软件体系结构设计原则，全链路保证高性能、高可用的数据服务。

# 课题计划进度表

|  |  |
| --- | --- |
| 任务信息 | 时间节点 |
| 准备服务器，各个技术组件的初步使用 | 第1周 |
| 实时日志产生器的开发 | 第2周 |
| 使用Flume收集产生的日志数据至Kafka，Kafka终端测试消费数据 | 第3周 |
| Spark Streaming程序从Kafka消费日志数据，计算统计前五活跃城市与地理位置 | 第4、5周 |
| Spark Streaming程序计算统计交易总金额、消费男女比例、搜索热词排行 | 第6、7、8周 |
| 开发所有Spark Streaming计算结果的dataV数据api | 第9、10周 |
| 梳理总结、论文撰写 | 第11、12、13、14、15周 |

# 参考文献

暂无