大连理工大学

专业学位硕士研究生学位论文开题报告

论文题目：基于 Spark 的日志分析系统的设计与实现

姓　　名：姚晗

学　　号：31617015

专业/领域：软件工程

培养类型：全日制

指导教师：李凤岐

实践导师：

入学日期：2016.9

报告日期：

报告地点：

研究生院制表

目　　录

1 绪论

1.1 选题背景

随着互联网技术的快速发展，信息技术和信息系统已经应用到各行各业，用户每一次在信息系统的操作都会留下痕迹，这就形成了日志。日志是信息系统记录系统和用户操作最常用的方式，所有的信息系统都会产生日志数据，日志数据包含一系列正常或异常的数据操作。通过对日志数据进行分析，可以获取有用的信息，为改善系统设计提升系统质量提供重要依据。

目前，互联网上的数据呈现爆炸式增长，许多企业每天收集到的数据量十分巨大，单一服务器很难满足数据存储与数据分析的需求。而分布式计算具有高效性，经济性与可扩展性，更适合目前的大数据运算场景，能够更好的满足企业的业务需求。Spark 是专为大规模数据处理而设计的快速通用的计算引擎，由 UC Berkeley AMP lab 开发，采用 RDD(Resilient Distributed Datasets)弹性分布式数据集，将中间输出结果保存在内存中，不需要读写 HDFS(Hadoop Distributed File System)分布式文件系统，运行速度更快，更适用于迭代算法。

本文以公开数据集中的实际监控日志为基础，深入分析日志处理应用的特点和需求，综合利用 Spark、Hadoop、YARN 等技术构建一个弹性可扩展的，支持多计算框架，具备多种计算能力的日志分析系统，

1.2 研究内容

设计并实现基于 Spark 的日志分析系统，针对不同信息系统所产生的日志记录提供一个通用的日志分析系统，通过现有机器学习方法进行大数据分析，以可视化的方法将分析结果展现出来，并能够根据用户需求调整结果的展现形式。

研究内容分为以下几个步骤。

1) 搭建 Spark 开发环境，熟悉分布式计算的开发流程。

2) 获取日志数据的公开数据集，分析多种信息系统产生的日志数据的类型，进行分类汇总，总结出日志常用的数据格式。将所有日志数据转化为通用的数据格式。

3) 利用机器学习算法对日志数据进行处理，形成分析报告，并对算法进行改进。

4) 利用可视化技术，将分析报告以图表的形式展现出来。

其中，关键问题是日志收集，日志分析与日志展示三部分。

2 相关技术

2.1 日志分析

日志是指系统所指定对象的某些操作和其操作结果按时间有序的集合，日志数据由日志记录组成，每条日志记录描述一次单独的系统事件，包含系统一系列正常或异常的操作信息，如系统的关键操作记录、程序出错信息、用户操作记录等。日志分析就是通过对日志数据进行处理，从而获取有用信息，如通过错误信息追踪问题出现的原因，然后进行针对性的解决；通过分析系统访问量等相关指标，分析系统的运营状况；通过分析用户的操作记录，可以分析出用户的喜好，行为习惯等信息。通过这些有用信息，为改善系统设计，提高系统的质量提供重要依据。

目前已有很多日志分析的工具和方法，最简单常用的方式是直接利用 Shell、Python 等脚本语言，以脚本的方式进行处理，脚本方式灵活便捷，但不易重用，可读性较差。随着互联网的普及，已有许多免费的单机或在线日志分析工具，如 Webalizer、Awstats、Google Analytics，百度统计等都是专门分析日志的工具，可以通过分析用户站点的日志，实时统计和显示当前系统的运行状况，提供丰富的功能和各式的统计报表。

Ganglia 是由加州大学伯克利分校开发维护的一款分布式开源软件，帮助用户监控系统与各个节点的性能数据，并将历史数据以曲线形式通过 php 页面呈现。Nagios 是一款优秀的开源监视软件，能够提供关于网络与系统运行状态等的信息数据，同时提供异常通知功能。ChuKwa 由 Yahoo 贡献，基于 Hadoop 框架，可以用它来分析和收集系统中的数据。Scribe 是 Facebook 开源的收集系统，将各个数据源的数据汇总到中央存储系统中。Hitune 是由英特尔亚太研发中心云计算组开发维护的基于 Hadoop 的性能收集分析工具，建立在 ChuKwa 之上，通过 Excel 展示收集分析结果。ClusterProbe 是由香港大学开发的分布式集群采样监控系统。

现有的日志分析系统，应用于特定的领域，设计方式与特色各有不同，目前仍存在以下几个问题。

1) 依赖程序多。采用了较多的开源程序，所以在安装和配置方面限制较大。
2) 占用资源大。许多日志分析系统设计不合理，负载很重，占用大量系统资源，影响了作业程序。
3) 缺乏备份与历史查询。目前主流的日志分析系统都是实时分析的，只能对某些指标进行分析结果展示，无法保存以及历史查询。

综上所述，国内外已有一些日志分析系统的解决方案与产品，它们有着各自的特性与优点，同时也存在着一些不足。本文结合现有系统的优点，设计并实现一款通用的日志分析系统，能够对目前主流系统的日志记录进行分析，生成可视化分析报告。

2.2 Spark

随着数据量的爆炸增长，单机或在线数据分析工具已经不能满足日志分析的需要，随着分布式计算的不断普及，基于 Hadoop 的技术方案逐渐被应用于日志数据处理中，但由于 Hadoop 自身的一些限制，不能满足大数据日志的分析。Spark 是由 UC Berkeley AMP lab 开发的一个开源分布式集群计算引擎，采用 RDD 的方式将中间结果保存在内存中，基于内存进行集群计算，实现快速处理。

传统的 Map/Reduce 方法基于无循环的数据流，先将数据从 HDFS 文件系统中读出，经过处理后又写入 HDFS 系统中，这种方式在需要数据重复利用时表现不佳，每次数据读取和操作都需要大量的 I/O 操作，降低了执行效率。Spark 针对这种应用场景，提供了基于内存的分布式计算，根据需要将重复利用的数据缓存至内存，通过这种方式，效率可以提高几十倍。

图 2.1 Spark 系统架构

如图 2.1 所示为 Spark 的系统架构图，包括多种高级组件。其中 Spark Core 实现了 Spark 的基本功能，包含任务调度，内存管理，错误恢复，与存储系统交互等模块，以及对 RDD 的 API 定义。RDD 表示分布在多个计算节点上可以并行操作的元素集合，是 Spark 主要的编程抽象。Spark SQL 是 Spark 用来操作结构化数据的程序包，通过 Spark SQL，我们可以使用 SQL 来进行数据查询。Spark Streaming 是 Spark 提供的对实时数据进行流式计算的组件，提供用来操作数据流的 API，并与 Spark Core 中的 API 高度对应。MLlib 提供了常见的机器学习功能的程序库，包含分类，回归，聚类，协同过滤等算法，还提供了模型评估，数据导入等额外功能支持。GraphX 是用来操作图的程序库，可以进行并行的图计算，还支持针对图的各种操作，以及一些常用图算法。独立调度器是

Spark 自带的一个简易调度器，让 Spark 可以高效的在一个计算节点到数千个计算节点之间伸缩计算。同时支持 Hadoop YARN、Apache Mesos 等常用的调度器。

Spark 目前广泛应用于国内外各大公司，如国外的 Google，Amazon，Yahoo，Microsoft 等以及国内的百度，腾讯，阿里巴巴等，并在很多具体的大数据环境中得到了广泛的应用。如阿里巴巴将 Spark 应用在双十一购物节中，处理大量产生的实时数据，百度也利用 Spark 进行大数据量的网页搜索优化实践，随着各行业数据量的增加，Spark 会应用到越来越多的实际应用中。

2.3 可视化

可视化是指将晦涩难懂的数据进行更友好的图形图像的表示，不仅限于视觉上，更多的是指易于理解,把复杂的，不直观的，不清晰而难以理解的事物变得通俗易懂且一目了然，以便于传播，交流和沟通，以及进一步的相应研究等。

目前已经有部分企业进行可视化的应用实践，如 Google Analytics，百度统计等工具，可以分析用户站点的日志，提供丰富的功能和各式的统计报表，为企业提供数据支持，以便于企业更加了解用户的使用习惯，改善系统设计，提高系统的质量。

本文使用可视化技术，将日志数据的分析结果以图表的形式展现出来，直观的展现出系统的性能数据，为企业了解系统运行状况，进行系统改进，制定发展规划提供参考依据。

3 系统设计

3.1 研究方法

按照软件工程理论，首先确定系统的实现目标，然后对系统进行需求分析，再针对需求进行概要设计，完成模块划分，针对每个模块进行详细设计，然后进行编码实现，最后对整个系统进行测试，修复 BUG。针对测试过程中暴露出的问题进行二次设计并开发，完善系统。

开发过程中所用到的算法，可以采用文献法，根据现有文献资料进行调查研究，熟悉并重现经典算法，然后根据项目需求对经典算法进行针对性的改进。

3.2 技术路线

本系统将按照软件工程原则，严格按照项目计划，需求分析，概要设计，详细设计，编码实现，测试维护等步骤进行，具体技术路线如下所示。

1) 需求分析。根据系统的实现目标，制定项目计划，并进行需求分析。

2) 概要设计。将整个系统划分为数据采集模块，数据分析模块与数据展示模块。

3) 详细设计。数据采集模块可利用现有的日志收集软件，对其进行相应的改进，并将不同日志格式进行统一。数据分析模块根据用户需求调用相应的机器学习算法对统一后的日志数据进行数据分析，输出数据分析结果。数据展示模块将数据分析结果以可视化的方式，用图表进行展示，并根据用户需求调整展示效果。

4) 编码实现。数据采集模块和数据分析模块，采用 Spark 计算引擎，搭建集群进行分布式计算，使用 Scala 语言将日志数据进行统一，并调用 MLlib 中的算法进行数据分析。数据展示模块，采用现有的 javascript 插件，将数据分析结果以图表的形式进行可视化显示，

5) 最终展现形式为 web 网页，用户导入日志文件或者配置服务器参数进行实时日志数据采集，并选择数据分析策略，在后台对日志文件进行数据分析，然后以图表的形式将数据分析结果展现出来，用户根据需要调整参数修改显示结果。

3.3 关键技术

3.3.1 数据采集

日志数据不仅数量巨大，而且不同方式采集到的日志格式也有所不同，如操作系统系统目录下记录系统状态或时间的系统日志，包括硬盘管理产生的硬盘读写事件、网络

管理器产生的网络情况、内存管理工具产生的内存访问情况等。由于格式不一，对系统日志进行具体分析之前，必须对日志进行识别和转换，格式化为统一的结构，方便对其进行处理。另外系统的同一个状态，往往会被不同组件记录到日志中，造成日志信息冗余的问题，所以在进行日志收集时需要对其进行重复记录过滤操作。

日志数据一般由少数输出语句统一输出，可以根据日志类型将由同一语句输出的日志划分为同一类，对其进行聚类操作，并提取其特征形成特征库，新产生的日志就可以依据该特征库进行类别标记。

对日志进行收集，要考虑关系型数据库数据，文本文件，检测日志数据，统计数据与告警数据等多种类型的日志收集，可以按照业务数据的自定义格式进行自定义加载。

3.3.2 数据分析

数据分析是整个系统的核心业务。数据分析基于规则对日志进行实时匹配，分析方法可以分为关联分析、单位时间频率分析、单条日志特征分析等。

单条日志特征分析是接收到规范化的日志后对每一条日志进行检索是否符合某一特征，如果符合某一特征则形成一个事件。特征分析是单位时间频率分析和关联分析的基础。

单位时间频率分析是指在一个时间窗内符合某个条件的某种单条日志特征发生的次数超过临界值，则发出相应的警告信息。时间窗是一个时间范围，一个可滑动的时间窗口。将一段时间内的日志组合为一个日志集合，称为一个时间窗内的日志集，简称时间窗。

关联分析是指多种设备日志间的联动，以上两种分析方法是针对同一类日志来源进行分析，而关联分析则是将多种日志来源的日志组合到一起进行分析，扩大分析范围。

3.3.3 数据展示

数据展示模块使用计算机图形学以及图像处理技术，将庞杂的数据转换为方便查看的图形或者图像，并能够进行交互操作，将抽象的、不规则的数据以可交互的视觉表达方式展现出来，达到对数据本质深入认识。数据展示模块负责将数据分析返回的数据进行直观化展现，以可视化图表形式用网页进行展示。

数据分析的结果一般都是高维复合数据，每条记录不只一个属性，可以利用一定的通道对数据进行转换，然后将高维数据 在二维平面中进行可视化，利用散点图和平行坐标轴进行数据展示，方便用户直观的发现数据信息的相似性和差异性，从而利用数据之间的关系，做出结果分析并制定相应的措施。

在本系统中，基于 AugularJS 框架构建前端平台，用于网页上展现可视化结果。AngularJS 由 google 提出的前段 mvc 框架，其组件化、宣告式语法、依赖注入等特性可以方便的进行网站的快速开发。同时使用 d3.js 库协助实现可视化效果。d3 可以将数据与数据对象模型(DOM)相结合，对文档进行数据驱动的操作和交互，同时能够支持对大数据集进行操作。

4 研究规划

4.1 现有基础

目前已经在 Ubuntu16.04 上安装 Spark 开发环境，能够进行初步的数据处理与分析。目前已经获取到一定量的开源数据集，包括 youtube 一段时间内的观看记录，维基百科一段时间内的浏览记录等。

4.2 研究计划

2017 年 4 月 搭建 Spark 开发环境并在本机构建集群。

2017 年 5 月 下载开源数据集，分析不同系统产生的日志，提取共同点。

2017 年 6 月 研究现有的日志分析系统，深入了解现有系统的特性与存在问题，设计本文的日志分析系统

2017 年 7 月 开发第一版的日志分析系统。

2017 年 8 月 完善第一版的日志分析系统，针对某个特定数据集进行分析，形成可视化的分析报告。

2017 年 9 月 进行系统测试，总结第一版系统存在的问题，需要改进的方向，进行第二版系统的设计。

2017 年 10 月 研究常用的机器学习算法，总结用户常用的数据分析策略，实现常用算法的调用模块

2017 年 11 月 开发第二版系统。

2017 年 12 月 完善第二版系统，可以对不同的日志数据进行分析，并能够选择分析策略，形成个性化的分析报告。

2018 年 1 月 进行系统测试，总结第二版系统存在的问题，进行针对性的改进。

2018 年 2 月 继续完善系统，根据开发过程中的文档记载进行汇总，开始论文编写工作。

2018 年 3 月 继续完善系统，完成论文的编写工作，不断完善。

4.3 预期成果

一个通用的日志分析系统,将日志数据导入系统或者配置参数采集实时的日志数据，选择数据分析策略，得到可视化的数据分析结果。使用户了解系统性能瓶颈，了解用户的使用习惯，为改善系统设计，提高服务质量提供数据依据。

参考文献

[1] Khosla, Shivkumar, and Varunakshi Bhojane. "Performing Web Log Analysis and Predicting Intelligent Navigation Behavior Based on Student Accessing Distance Education System." Advances in Computing, Communication, and Control. Springer Berlin Heidelberg, 2013. 70-81.

[2] Karau, Holden. Fast Data Processing With Spark. Packt Publishing Ltd, 2013.

[3] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.

[4] Jagadish, H. V., et al. "Big data and its technical challenges." Communications of the ACM 57.7 (2014): 86-94.

[5] Zaharia, Matei, et al. "Spark: Cluster Computing with Working Sets." HotCloud 10.10-10 (2010): 95.

[6] Lin, Xiuqin, Peng Wang, and Bin Wu. "Log analysis in cloud computing environment with Hadoop and Spark." Broadband Network & Multimedia Technology (IC-BNMT), 2013 5th IEEE International Conference on. IEEE, 2013.

[7] Wei, Jianwen, et al. "Analysis farm: A cloud-based scalable aggregation and query platform for network log analysis." Cloud and Service Computing (CSC), 2011 International Conference on. IEEE, 2011.

[8] Eickhoff, Carsten, et al. "Lessons from the journey: a query log analysis of within-session learning." Proceedings of the 7th ACM international conference on Web search and data mining. ACM, 2014.

[9] Kim, Gei-Young, et al. "The Detection Model of Malignant Query and Personal Information Leakage based on Log Analysis." International Journal of Multimedia and Ubiquitous Engineering 10.11 (2015): 105-114.

[10]张彬. 基于 Spark 大数据平台日志审计系统的设计与实现[D].山东大学,2015.

[11]肖东方. 基于 Hadoop 的运维日志采集分析平台的设计与实现[D].山东大学,2016.

[12]刘鹏. 基于 Spark 的数据管理平台的设计与实现[D].浙江大学,2016.

[13]薛瑞,朱晓民. 基于 Spark Streaming 的实时日志处理平台设计与实现[J]. 电信工程技术与标准化,2015,(09):55-58.

[14]吴雯祺. Spark 性能数据收集分析系统的设计与实现[D].哈尔滨工业大学,2015.

[15]王鹏. 云平台下日志分析系统的设计与实现[D].北京邮电大学,2013.

[16]刘季函(Liu,Chi Han). 基于 Spark 的网络日志分析系统的设计与实现[D].南京大学,2014.

[17]孔庆春. 基于 Spark 大数据平台日志审计系统的设计与实现[J]. 电脑知识与技术,2016,(15):10-11.

[18]周海靖. 日志大数据分析平台技术研究[D].山东大学,2015.

[19]阳小兰,钱程,赵海廷. Web 日志分析系统研究[J]. 计算机技术与发展,2011,(09):211-215.

[20]李文. 基于 Spark 可视化大数据挖掘平台[A]. 中国自动化学会系统仿真专业委员会、中国系统仿真学会仿真技术应用专业委员会、离散系统仿真专业委员会.系统仿真技术及其应用学术论文

集（第 15 卷）[C].中国自动化学会系统仿真专业委员会、中国系统仿真学会仿真技术应用专业委员会、离散系统仿真专业委员会:,2014:4.

附录

如图 1 所示为研究生期间的培养计划，图 2 为目前完成课程的成绩，预计第二学期能完成培养计划内所有课程的学习。

课程编号	课程名称	课程教学大纲	课程类别	是否必修课	学分	学时	开课学期	备注
2090430020	软件体系结构	未上传	专业基础课	是	3	48	秋	
2090530010	算法分析与设计 II	未上传	专业基础课	是	3	48	秋	
2090530031	网络科学	未上传	专业基础课	是	2	32	秋	
2070110062	知识产权	未上传	公共基础课	是	1	16	秋	
2070110072	信息检索	未上传	公共基础课	是	1	16	秋、春	
2070310013	中国特色社会主义理论与实践研究	未上传	公共基础课	是	2	36	秋、春	
2100010033	阅读与写作 I （基础读写技能）	未上传	公共基础课	是	2	32	春	
2120020013	矩阵与数值分析	未上传	公共基础课	是	3	48	秋	
2120020023	优化方法	未上传	公共基础课	是	2	32	秋	
2090140011	信息服务技术与标准	未上传	行业前沿课	是	1	16	秋、春	
2090440091	软件工程发展前沿	未上传	行业前沿课	是	1	16	春	
2090460012	企业专业实践	未上传	实验实践课	是	6	96	秋、春	
2090440040	人工智能	未上传	专业选修课	否	2	32	春	
2090440050	JAVA2和J2EE	未上传	专业选修课	否	2	32	春	
2090540070	移动技术	未上传	专业选修课	否	2	32	春	

图 1 培养计划

课程	课程学分	选修学期	成绩
知识产权	1	1	84
信息检索	1	1	P
软件体系结构	3	1	77
算法分析与设计 II	3	1	89
网络科学	2	1	83
矩阵与数值分析	3	1	73

图 2 课程成绩

大连理工大学专业学位硕士研究生学位论文开题报告评审意见表

<table>
<tr><td>学　号</td><td>31617015</td><td>学生姓名</td><td>姚晗</td><td>专业/领域</td><td>软件工程</td></tr>
<tr><td colspan="3">第一次开题 ☑</td><td colspan="3">第二次开题 ☐</td></tr>
<tr><td colspan="6">实 践 导 师 信 息</td></tr>
<tr><td>姓　名</td><td></td><td>性　别</td><td></td><td>职称/职务</td><td></td></tr>
<tr><td>专　业</td><td></td><td>所在单位</td><td></td><td>联系电话</td><td></td></tr>
<tr><td>通讯地址</td><td colspan="3"></td><td>E_mail</td><td></td></tr>
<tr><td colspan="6">校内导师考核意见（对课程学习情况、校内实践实习情况、参加学术活动情况、学位论文工作及开题报告撰写情况、企业实践实习情况及计划、学生的学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：
1） 考核成绩：☐ 优秀，☑ 良好，☐ 中等，☐ 及格，☐ 不及格
2） 是否通过：☑ 通过，☐ 不通过
3） 关于开题报告撰写质量及学位论文工作的具体意见（可加页）：

同意开题

导师签字：李凤岐
2017 年 5 月 3 日</td></tr>
<tr><td colspan="6">实践导师考核意见（对学位论文工作及开题报告撰写情况、企业实践实习情况及计划、学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：
1） 考核成绩：☐ 优秀，☐ 良好，☐ 中等，☐ 及格，☐ 不及格
2） 是否通过：☐ 通过，☐ 不通过
3） 关于开题报告撰写质量及学位论文工作的具体意见（可加页）：

导师签字：
年　　月　　日</td></tr>
</table>

<table>
<tr><td rowspan="8">评议专家组</td><td></td><td>姓名</td><td>职称</td><td>学科专业</td><td>是否博导</td><td>签字</td></tr>
<tr><td>组长</td><td>李明楚</td><td>教授</td><td>网络工程</td><td>是</td><td>李明楚</td></tr>
<tr><td rowspan="6">成员</td><td>田园</td><td>副教授</td><td>网络工程</td><td>否</td><td>田园</td></tr>
<tr><td>高静</td><td>讲师</td><td>网络工程</td><td>否</td><td>高静</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="7">专家组评审意见（对课程学习情况、校内实践实习情况、参加学术活动情况、学位论文工作及开题报告撰写情况、企业实践实习情况及计划、学生的学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：
1） 选题是否属于本学科领域（含交叉学科）：☑ 是，☐ 不是（须重新开题）
2） 选题是否符合专业学位论文要求：☑ 是，☐ 不是（须重新开题）
3） 考核成绩：☐ 优秀，☑ 良好，☐ 中等，☐ 及格，☐ 不及格
4） 是否通过：☑ 通过，☐ 不通过
5） 关于开题报告撰写质量及学位论文工作的具体意见（可加页）：

该生的论文选题较好，具有较高的理论和实践价值，研究的前期准备较为充分，通过查阅有关文献，基本上了解了论文题目所涉及的理论知识，并根据文章的研究方向做了较为全面细致的梳理，研究内容较为充实，研究方法合理，研究的重难点比较明确，开题报告结构合理，内容完整，答辩时语言流畅，重点较突出，基本能正确的回答评审小组提出的问题
同意该课题开题

组长签字：李明楚
2017 年 5 月 4 日</td></tr>
<tr><td colspan="7">点长意见：

点长签字：
年 月 日</td></tr>
</table>