

教育背景

中国科学院软件研究所 • 计算机软件与理论 (保送)

2013年9月至今

北京交通大学 • 软件工程学士 (GPA: 3.75 思测排名: 3/171 综合排名: 5/171)

2009年9月-2013年7月

项目及实习经历

HRS: 基于相似度的健康数据检索与摘要 • 中国科学院软件研究所(毕设课题)

2015年7月-2016年5月

项目概述: HRS 旨在利用历史健康数据(体检记录,门诊记录,住院记录等)对健康主体进行建模,利用机器学习并结合领域知识技术对该健康模型进行摘要,并且通过健康模型之间的相似度来进行检索。(项目参与人数:2人)

主要工作: 设计并实现系统架构;设计健康数据模型;基于 knowledge 和 data 两种方式来选取特征;相似度算法研究;

关键技术: 数据清洗,特征提取与融合,逻辑回归,相似度计算,信息检索,数据索引与缓存等。

海量数据定制查询与优化 • 国家脑卒中数据中心(国家自然科学基金子项目)

2015年5月-2015年7月

项目概述: 对基地医院的脑卒中筛查数据进行清理存储,提供基于 Hive SQL 的可定制的查询分析服务,可视化分析结果;

主要工作: 设计并实现权限机制(基于 RBAC)来隔离对服务的访问;Hive SQL 语法解析并将其参数化;制定参数化 SQL 语句的语法规则;多线程并发查询优化;设计两级缓存机制对查询分析结果进行缓存(一级缓存基于内存对象,二级缓存基于数据库);可视化分析结果;

关键技术: Hive, RBAC 权限, SQL 语法解析,正则表达式,java 弱引用,java 多线程,Reactor 模式;

Haflow: 大数据分析平台 • 中国科学院软件研究所(863子项目)

2014年8月-2015年2月

项目概述: Haflow 是一个一站式的大数据分析应用开发平台。支持多种大数据处理分析框架的接入,并借助其分布式存储和分布式计算能力,加快大数据分析的速度。构建基于 Web 的可视化,拖曳式的数据分析应用开发界面,同时将用户从平台搭建,工具集成,接口学习,算法管理,实验脚本编写和多任务调度等繁杂工作中抽离出来。(项目参与人数:7人)

主要工作: 设计基于组件以及模型驱动的数据分析业务流程处理框架;设计与实现基于嵌套图识别与分层的模型转换算法;基于 HDFS 的中间数据管理;设计自适应大数据平台层架构;系统服务管理;

关键技术: Hadoop, Hive, Sqoop, Spark, Oozie, Spring, Hibernate, java 注解,数据流模型转换

RMS: Resource Management System(原型系统) • (实习) Lenovo 北京研究院 • 开发实习生

2015年3月-2015年5月

项目概述: RMS 是 Lenovo 用来管理雇员工作量的资源管理系统,该系统原型实现了工作量上报,工作量统计,工作量审批和生成报表等核心业务模块。(开发人员:3人)

主要工作: 设计数据结构和关键算法;设计并实现系统基础架构(SSH);实现系统的两大主要模块:工作量上报和审批。

关键技术: MySQL, Spring MVC, Hibernate, Maven, Junit, JQuery, Ajax

MRTuner: Big Data Analytics Auto-Tuning • (实习) IBM 中国研究院(CRL) • Research Intern

2012年11月-2013年7月

项目概述: MRTuner 是用来对 MapReduce 作业进行全局优化的工具。它采用 PTC (Producer-Transporter-Consumer) 模型来对任务并行执行时的 tradeoffs 进行建模,并采用高效的搜索算法来找到最优的执行计划。(项目参与人数:3人)

主要工作: 研究 MapReduce 源码;设计流水线执行模型来建模任务之间关系;定义建模流水线执行计划的4种关键因素,并设计 PTC Cost Model 来预测作业运行时间;分析对性能敏感的参数之间的关系,设计快速搜索算法来找出最优执行计划;设计系统架构并实现原型系统;使用 HiBench (包含多种 workloads,数据量从 GB 到 TB 级别)在两种不同类型集群上进行实验评估。

关键技术: Hadoop MapReduce V1.x, Linux shell, Modeling, Ganglia, Starfish

发表论文

MRTuner: A Toolkit to Enable Holistic Optimization for MapReduce Jobs. PVLDB 7(13): 1319-1330 (2014)

合作发表在 VLDB 2014 (数据库领域国际顶级会议)。

获奖经历

- 2014-2015, 中国科学院大学 “研究生国家奖学金”(20000元,前1%);
- 2013-2015, 中国科学院大学 “三好学生”, “优秀学生干部”各2次;
- 2012-2013, 北京交通大学 “优秀毕业论文”, “优秀毕业生”; (前3%)
- 2009-2012, 北京交通大学 “三好学生”, “学习优秀奖学金”各4次;
- 2010-2011, 北京交通大学 “汉能李嘉宁专项奖学金”; (7000元,前1%)

竞赛经历

- 2014, Ow2 国际开源软件竞赛二等奖;
- 2012, 国家级创新实验评级优秀(最高等级);
- 2012, 北京交通大学 ACM 校队选拔赛一等奖;
- 2011, 北京交通大学数学建模二等奖;

专业技能

- 熟悉 java, 了解 C++ 语言, Linux; 熟悉 数据结构, 算法设计, 设计模式, 操作系统, 计算机网络; CET 6, 硕士英语免修
- 熟悉 Spring, Struts, Hibernate 等常用 java 开源框架, 熟悉 MySQL, Oracle 数据库;
- 熟悉 Hadoop 生态圈及 MapReduce 实现原理, 了解大数据相关工具如 HBase, Hive, Kafka, Zookeeper 等;