北京交通大学

**硕士学位研究生选题报告及工作计划**

**研究生**  赵晋萍 **学 号**  13126176

**学 院** 软件学院

**专 业**  软件工程

**研究方向**  嵌入式

**指导教师** 冯凤娟

**入学日期**  2013年09月01日

填表日期 2014 年 11 月 10 日

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 选 题 名 称 | | 基于LAMP框架的通用数据测试平台的设计与实现 | | |
| 选题来源（公司实际项目、导师科研项目、自选）与项目背景说明  题目来源于本人在百度公司实习参与的DHC数据体检中心平台项目。该平台主要是为支撑互联网数据测试、修复和监控等质量保证业务，而基于LAMP框架（Linux + Apache + MySQL + Python/PHP Yii）搭建的。在百度实习期间，本人主要负责DHC数据体检中心平台前后端开发工作，其中部分参与了数据测试和数据修复模块的开发工作，以及独立负责了数据质量监控（数据文件、覆盖率和错误率三个维度的监控）模块的设计与实现工作。  在物联网、云计算、移动互联网等技术的推动下，数据规模正爆炸式增长，数据来源、结构和应用特征繁多复杂，当今社会已进入大数据时代，数据成为信息企业业务活动的副产品和重要资产。企业可快速获取、分析并提炼数据中隐含的、有价值的信息，从而做出更加精准、更加符合市场和客户需求的决策，实现高效运营和效益最大化。然而要想充分发挥大数据所赋予的机遇和优势，前提是必须拥有可靠、准确、及时的高质量的数据，否则大数据的优势将化为泡影。此外，传统软件测试技术侧重于发现代码质量问题，数据质量方面较为薄弱。而问题数据同样会给软件系统带来直接或潜在的危害（如用户关注度较高的数据存在明显错误或更新不及时等不同程度的线上问题），影响产品用户体验和企业信誉形象。因此，大数据时代背景下如何保证数据质量是需要持续探索和不断积累的技术领域。 | | | | |
| 论文类型：在所属类型前划“√”（只可填写一项） | | | | |
| √ | （1）工程项目开发 | |  | （2）软件策划与营销 |
|  | （3）其它 | |  |  |

|  |  |
| --- | --- |
| 论和  文拟  需达  解到  决的  的目  问标  题 | 1. 论文需解决的问题   1）数据测试方面：传统的数据测试方法，耗费人力且不全面。要么是从全量数据中随机抽样一部分，然后人工逐条review抽样结果。这种方法，很容易遗漏数据中的重要问题。要么是针对一份特定的数据，开发程序或脚本进行检测。随着数据种类的不断增多，往往会有大量的重复工作和维护负担。我们急需一套高效全面通用的数据测试方法和工具，最大程度的自动化和平台化成为我们的必然选择。  2）数据修复方面：传统的测试思维主要局限于正确性验证，QA积累的数据检测经验在错误修复方面没有得到充分利用。  3）数据质量监控方面：数据通常是不断更新的，数据质量也会相应变化而带来一定的风险。如数据来源发生变化时，原有数据抓取模板将会失效而抓取到错误的数据或抓不到数据。或数据产出方面的不稳定因素，也可能使数据不能及时正确地上线生效。因此，自动跟踪数据质量变化并进行有效预警，对于指导数据产出方及早发现、准确定位和快速解决问题，从而规避风险具有重要意义。   1. 拟达到的目标   1）数据测试方面，避免繁杂琐碎的配置，只要提交数据源地址，即可自动探测出数据编码、类型和结构（支持通用二维表数据和xml/json树状结构数据），同时通过自动学习得出各列数据类型，然后根据类型匹配相应规则（预置、领域或用户扩展规则），从而逐行计算得到各项统计指标和明细数据。最后将数据检测结果以web页面形式渲染出来，实现数据及其测试结果的可视化。同时支持开放API，提供数据检测任务提交和结果获取的通用服务。  2）数据修复方面，根据不断积累的数据检测经验，总结得出某些特定类型及其对应的修复规则，从而在识别出相应类型后自动进行数据修复。同时抽象出一种数据修复机制，通过事先根据数据列结构配置修复规则的方式也可达到数据修复和清洗的目的。  3）数据监控方面，通过事先配置监控关注项和告警阈值，根据数据测试计算结果进行二次计算，判断数据文件、覆盖率或错误率相关监控指标是否发生告警。若告警，则发出告警邮件和短信通知数据负责人。同时提供数据监控项实时报表和趋势图表等服务，便于整体把握数据质量状况。此外，还可将数据监控结果在API中输出，根据监控结果判断，有效阻止错误数据进入后续流程。 |
| 国品  内研  外发  同现  类状  产  （可选） |  |
| 问方  题案  解简  决述 | 以软件工程思路来研究和解决数据测试自动化和平台化问题，首先，调研和分析数据测试业务相关的功能和非功能需求，然后进行架构、数据库和接口的概要设计以及数据测试、修复和监控各功能模块的详细设计，并进行系统实现：一次数据检测作为一个数据体检任务，系统用户交互方面功能基于PHP Yii框架（一种MVC框架）并配合前端技术（HTML+CSS+Jquery等）实现，包括预处理、类型学习、规则计算（含数据修复）和数据监控在内的后端核心功能使用Python脚本语言实现，并将计算结果写入MySQL数据库，UI方面PHP取数据渲染任务报告即可，也可通过API获取JSON格式的结果数据。最后验证方案设计和系统效果满足数据测试业务需求，推广上线应用即可。 |
| 项价  目值  的和  应意  用义 | 1. 自动通用的二维表和树状数据检测服务，从容应对爆炸式增长和繁多复杂的数据，释放传统数据测试和标注所需的人力。同时不漏掉任何一行数据，告别大数据抽样人工检测带来的数据质量隐患。 2. 提供数据质量保证全套服务，不仅可以自动发现数据中隐含的问题，还可积累宝贵经验，用于数据修复，直接提升数据质量。数据产出人员与QA共建体检规则模板，实现测试前置，共建数据生态。 3. 面对不断更新的数据，有较好的质量监控能力，每期数据自动跟踪预警质量变化，让文件、覆盖率和错误率相关指标的变化一目了然，帮助数据负责人第一时间定位质量风险，从而为线上产品提供及时、准确的数据或帮助产品运营方有效决策，实现效益最大化。 |
| 本些  人准  做备  过工  哪作  ？ | 1. 收集、参阅和分类整理了大数据质量、测试技术相关中英文文献及百度DHC数据体检中心项目相关文档资料； 2. 学习软件工程专业基础知识（需求、设计、实现、测试等）和本项目相关技术（Python+PHP Yii框架+Jquery等），充分调研和理解用户需求基础上，着手设计和实现目标系统； 3. 分别与企业和校内导师沟通实习项目和论文工作情况，确认论文计划可行 |

**工作计划**

|  |  |  |  |
| --- | --- | --- | --- |
| 何阶段 | 课题分析调研阶段 | 课题设计阶段 | 系统实现与论文定稿阶段 |
| 主  要  内  容 | 1. 收集并整理课题及项目技术相关的文献和资料 2. 完成论文开题工作 3. 分析系统需求，包括功能需求和非功能需求，完成系统需求分析报告 | 1. 完成系统架构设计 2. 完成系统接口设计 3. 完成系统数据库设计 4. 完成数据测试、数据修复和数据质量监控各模块详细设计 5. 完成论文初稿 | 1. 完成系统实现 2. 完成系统测试和上线 3. 完善和修改论文 4. 论文定稿 5. 准备论文答辩 |
| 主 作  要 法  工 和  作 拟  地 达  点 目  、 标  方  式  、 | 1. 主要工作地点：百度公司、学校 2. 方式和做法：收集整理资料，学习现有知识和技术基础上，分析和调研系统需求 3. 目标：完成论文开题工作，后续明确目标系统各项功能需求和非功能需求，完成系统需求分析报告 | 1. 主要工作地点：百度公司、学校 2. 方式和做法：根据系统需求分析结果，进行系统架构、接口和数据库设计以及各模块实现方案详细设计 3. 目标：给出系统概要设计和详细设计方案，完成论文初稿 | 1. 主要工作地点：百度公司、学校 2. 方式和做法：实现和测试系统各模块，编写用户使用手册，推广系统投入线上应用 3. 目标：完成系统开发、测试和上线，完善论文，并在导师指导下进行修改和定稿，准备论文答辩 |
| 起  止  时  间 | 2014年10月—2014年11月 | 2014年12月—2015年1月 | 2015年2月—2015年5月 |
| 备 注 |  |  |  |

|  |
| --- |
| 导师对选题报告（就研究生对项目背景的了解情况、解决方法、实现手段等）及工作计划的评语：  导师签名： 年 月 日 |
| 考核小组对该选题的意见：（工作是否具备条件、是否同意选定该课题等）  签名： 年 月 日 |
| 学院对选题报告和工作计划的意见：  负责人签名： 年 月 日 |