**密级:**



**硕士学位论文**

**大数据系统应用可靠性测试框架设计与实现**

**作者姓名： 郑莹莹**

**指导教师: 叶丹 研究员**

**中国科学院大学软件研究所**

**学位类别: 工学硕士**

**学科专业: 软件工程**

**培养单位: 中国科学院大学软件研究所**

**2017年 4月**

**Design and Implementation of a reliability benchmarking framework for big data system**

**By**

**Zheng Yingying**

**A Dissertation Submitted to**

**University of Chinese Academy of Sciences**

**In partial fulfillment of the requirement**

**For the degree of**

**Master of Computer Software and Theory**

**Institute of Software**

**University of Chinese Academy of Sciences**

**April, 2017**

**独创性声明**

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明。

签名：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**关于论文使用授权的说明**

本人完全了解中国科学院软件研究所有关保留、使用学位论文的规定，即：中国科学院软件研究所有权保留送交论文的复印件，允许论文被查阅和借阅；中国科学院软件研究所可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：\_\_\_\_\_\_\_\_\_\_ 导师签名：\_\_\_\_\_\_\_\_\_\_ 日期：\_\_\_\_\_\_\_\_\_\_

**大数据系统应用可靠性测试框架设计与实现**

**摘要**

当前，Spark、Flink等分布式处理框架被广泛应用于大数据的处理分析，然而在处理大数据时，经常出现内存溢出、IO异常、任务超时等运行时错误，这些错误会直接造成应用执行失败。然而，现有的测试基准主要针对系统的性能提供测试，没有考虑系统的可靠性问题；并且，这些测试基准通常使用常规数据以及固定的配置来进行测试，缺乏测试的多样性。针对大数据系统面临的多样及复杂的运行时错误以及现有测试基准的不足，论文设计并实现了面向大数据系统应用的可靠性测试基准框架，提供可视化的测试界面，支持自定义的数据生成和自动参数组合测试。

设计与实现大数据系统应用的可靠性测试基准框架，论文介绍了可靠性测试基准的设计，并针对以下关键技术进行研究。首先，本文提出了针对大数据应用的异常数据生成方法，定义了异常数据的概念和数据的多种随机分布形式；同时，通过分析应用程序特征，提出应用计算特性与数据异常特征对应关系，并给出了特定应用的异常数据的生成规则。其次，本文提出了针对大数据应用的参数组合测试方法，采用组合测试，并通过贪心算法对系统和应用参数进行组合空间削减测试；针对不满足参数独立性以及参数取值的相关性的，提出探测性参数验证方法，通过指数增长的慢启动方式来确定最差的资源占用的参数取值。

论文详细的介绍了大数据系统应用可靠性测试基准的设计与实现，并将该可靠性测试基准框架应用在Spark系统上。通过对Spark上的基准应用进行可靠性测试，验证了本文提出的异常数据生成以及参数组合测试方法的可用性，并在6个应用（Join、Mix、RandomForest、LogisticsRegression、ALS、PageRank）中发现了三种类型的错误（内存溢出错误、运行超时以及计算结果错误）。

**关键词：**可靠性；基准测试；异常；大数据系统；大数据应用；

**Design and Implementation of a reliability benchmarking framework for big data system**

**ABSTRACT**

**Keywords:** reliability, benchmark, error, large data system, large data application

目录

[1 绪论 3](#_Toc478388600)

[1.1 研究背景 3](#_Toc478388601)

[1.2 研究内容 4](#_Toc478388602)

[1.3 论文组织 5](#_Toc478388603)

[2 大数据系统及应用可靠性相关工作 6](#_Toc478388604)

[2.1 大数据系统及应用 6](#_Toc478388605)

[2.1.1 大数据系统 6](#_Toc478388606)

[2.1.2 大数据应用 9](#_Toc478388607)

[2.2 大数据系统及应用的可靠性问题 10](#_Toc478388608)

[2.2.1 可靠性定义 10](#_Toc478388609)

[2.2.2 可靠性问题分析 11](#_Toc478388610)

[2.3 测试基准框架研究现状 12](#_Toc478388611)

[2.4 测试基准需求 14](#_Toc478388612)

[3 可靠性测试基准设计 16](#_Toc478388613)

[3.1 基准应用 16](#_Toc478388614)

[3.1.1 应用类型 16](#_Toc478388615)

[3.1.2 工作负载 17](#_Toc478388616)

[3.2 测试数据 20](#_Toc478388617)

[3.3 基准执行 21](#_Toc478388618)

[3.4 测试度量 22](#_Toc478388619)

[4 可靠性测试基准关键技术 23](#_Toc478388620)

[4.1 数据生成方法 23](#_Toc478388621)

[4.1.1 异常特征 23](#_Toc478388622)

[4.1.2 数据概率分布 24](#_Toc478388623)

[4.1.3 数据生成 26](#_Toc478388624)

[4.2 参数组合测试方法 31](#_Toc478388625)

[4.2.1 参数配置 32](#_Toc478388626)

[4.2.2 组合测试 33](#_Toc478388627)

[4.2.3 参数组合空间削减 36](#_Toc478388628)

[5 可靠性测试基准框架设计与实现 40](#_Toc478388629)

[5.1 系统架构 40](#_Toc478388630)

[5.2 系统总体设计 41](#_Toc478388631)

[5.2.1 Web模块设计 42](#_Toc478388632)

[5.2.2 基准模块设计 43](#_Toc478388633)

[5.3 Web模块实现 44](#_Toc478388634)

[5.3.1 架构实现 44](#_Toc478388635)

[5.3.2 执行流程 47](#_Toc478388636)

[5.4 基准模块实现 48](#_Toc478388637)

[5.4.1 数据生成器实现 48](#_Toc478388638)

[5.4.2 组合参数发生器实现 53](#_Toc478388639)

[5.5 可靠性测试应用验证 55](#_Toc478388640)

[5.5.1 实验环境 56](#_Toc478388641)

[5.5.2 参数配置 56](#_Toc478388642)

[5.5.3 实例及分析 57](#_Toc478388643)

[6 结束语 63](#_Toc478388644)

[6.1 论文贡献 63](#_Toc478388645)

[6.2 未来工作展望 63](#_Toc478388646)

[参考文献 66](#_Toc478388647)

[发表文章 70](#_Toc478388648)

[致谢 72](#_Toc478388649)