申请上海交通大学工程硕士学位论文

**基于计算特性的Spark内存自适应管理策略研究**

|  |  |
| --- | --- |
| 学校代码： | 10248 |
| 作者姓名： | 赵耀 |
| 学 号： | 1130379091 |
| 第一导师： | 胡飞 |
| 第二导师： | 陈昊鹏 |
| 学科专业： | 软件工程 |
| 答辩日期： | 年 月 日 |

上海交通大学软件学院

年 月

A Dissertation Submitted to Shanghai Jiao Tong University

for Master Degree of Engineering

**Adaptive Memory Management Research Based on In-memory Computing Characteristics In Spark**

|  |  |
| --- | --- |
| University Code： | 10248 |
| Author： | Yao Zhao |
| Student ID: | 1130379091 |
| Mentor 1： | Fei Hu |
| Mentor 2: | Haopeng Chen |
| Field： | Software Engineering |
| Date of Oral Defense： |  |

School of Software

Shanghai Jiaotong University

Jun, 2008

**上海交通大学**

**学位论文原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

**上海交通大学**

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

**保密**□，在 年解密后适用本授权书。

本学位论文属于

**不保密**□。

（请在以上方框内打“**√**”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

**基于计算特性的Spark内存自适应管理策略研究**

摘 要

伴随着信息时代到来的是海量的数据信息，在被称为“大数据”时代的21世纪，大数据的处理成为了这个时代的难题。目前，分布式计算机技术是处理大数据的主流方式，通过建立分布式集群，能让分布式系统获得堪比超级计算机的运算能力，而且分布式系统的计算和存储能力还会随着集群的扩展而不断提升。但是，内存容量不足一直是制约分布式系统性能的一大因素，但随着内存制造工艺的发展提高，一种新型的基于内存计算技术的分布式框架——Spark，应运而生。Spark在处理迭代式机器学习运算和交互式实时查询操作时，所获得的性能远超其他分布式框架。然而，内存容量始终要比数据的规模小得多，在内存不足时，Spark会遭遇到性能瓶颈。

如何高效地使用内存，成为了提升Spark系统性能的关键问题。为了解决这个难题，本文设计了一种自适应的内存调优策略。这种自适应调优策略主要分为三个部分：首先，是针对数据序列化算法的自适应优化策略。数据序列化是分布式系统中常用的优化方式，与系统性能的优劣密切相关。序列化方法在节省存储空间的同时，还能减轻系统垃圾回收的压力。同时，数据要在分布式集群各个节点间传输，也必须要进行序列化操作。序列化自适应策略通过当前系统的资源消耗情况，来选择合适的序列化算法。其次，是数据压缩算法的的自适应优化策略。压缩算法可将数据压缩至原来的几分之一，甚至十几分之一，相比于序列化方法，更能节省系统存储空间。但是，每种不同的压缩算法之间，也是存在差异的。自适应算法会根据系统情况选择合适的压缩算法。最后，是垃圾回收的自适应优化策略。对于运行在JVM虚拟机上的Spark分布式框架来说，JVM的性能优劣直接关系到了整个系统的性能好坏。自适应算法通过收集和分析系统当前系统信息，进行垃圾回收的调优。

在设计部分，通过对Spark源代码的添加和修改，本文在原生的Spark系统上实现了SATS（Spark Adaptive Tuning Strategy）子系统，系统分为三个模块：运行时数据收集模块、自适应决策模块和参数优化模块。实现部分对着三个模块的实现细节进行了详细描述。在实验部分，详细的分析了实验结果，并验证了条有策略的有效性。

关键词 大数据，分布式计算系统，Spark， 内存计算，自适应

**Adaptive Memory Management Research Based on In-memory Computing Characteristics In Spark**

ABSTRACT

With the advent of the information age there comes the huge number of data. In the 21st century which is called the "big data" era, how to process big data has become a big problem. Currently, the distributed computer technology is a mainstream way to deal with big data. By establishing distributed cluster, distributed system can obtain great computing power like supercomputer and huge storage capacity. Furthermore, these abilities can be improved with the expansion of the cluster. However, the shortage of memory capacity is a main factor which restricts the performance of distributed systems. In recent years with the improvement of memory manufacturing process, Spark, which is a new kind of distributed framework based on in-memory computation , comes out. When facing iterative machine learning computation and interactive query, the performance of Spark exceeds other distributed frameworks. However, the memory capacity is always much smaller than the data volume. In this case Spark will encounter performance bottlenecks.

How to make better use of memory is a key issue to enhance Spark performance. To solve this problem, we design an adaptive memory tuning strategy. This strategy consists of three adaptive tuning algorithms: The first algorithm is the serialization adaptive algorithm. Data serialization is an optimization approach which is commonly used in distributed systems. Whether to use proper data serialization has a great influence on system performance. Serialization can save storage space. At the same time, it can reduce the pressure of the system garbage collection. Meanwhile, the data to be transmitted between the nodes in the cluster, also need to be serialized. Serialization adaptive strategy will select an appropriate serialization algorithm based on the system resource consumption. The second algorithm is the compression adaptive algorithm. Compression algorithm can compress data to a fraction, or even a dozen of the original data volume. Compared to the serialization algorithm, the system can save much more storage space. However, different compression algorithms perform differently. Compression adaptive algorithm will choose the appropriate compression algorithm according to system information. Finally, the last algorithm is the garbage collection adaptive algorithm. Spark is a distributed framework which runs on JVM virtual machine. So the performance of JVM is directly related to the performance of the entire system. Garbage collection adaptive algorithm tunes garbage collection by collecting and analyzing the information of the current system.

In the design section, through adding and modifying Spark source code, we implements SATS (Spark Adaptive Tuning Strategy) subsystem on Spark. The subsystem is divided into three modules: run-time data collection module, adaptive decision module and parameter optimization module. The implementation section describes the details of these three modules. In the experimental part, we analyze the experiment results in detail and verify the effectiveness of adaptive tuning strategy.

**Keywords** big data, distributed computer system, Spark, in-memory computation, adaptive

目录

[摘 要 i](#_Toc439620989)

[ABSTRACT iii](#_Toc439620990)

[目录 v](#_Toc439620991)

[第一章 绪 论 1](#_Toc439620992)

[1.1 研究背景与意义 1](#_Toc439620993)

[1.2 研究目的和内容 2](#_Toc439620994)

[1.3 研究意义 3](#_Toc439620995)

[1.4 论文结构 4](#_Toc439620996)

[第二章 分布式系统相关技术研究与分析 5](#_Toc439620997)

[2.1 分布式计算技术概述 5](#_Toc439620998)

[2.1.1 MapReduce计算框架 5](#_Toc439620999)

[2.1.2 Tez计算框架 6](#_Toc439621000)

[2.1.3 Spark计算框架 6](#_Toc439621001)

[2.2 Spark框架研究 6](#_Toc439621002)

[2.2.1 内存抽象RDD 7](#_Toc439621003)

[2.2.2 Spark的Job生成 8](#_Toc439621004)

[2.2.3 Spark任务执行流程 9](#_Toc439621005)

[2.3 分布式框架优化技术研究 10](#_Toc439621006)

[2.3.1 参数空间优化法 10](#_Toc439621007)

[2.3.2 历史数据优化法 10](#_Toc439621008)

[2.3.3 融入机器学习的混合优化法 11](#_Toc439621009)

[2.3.4 Spark优化策略 12](#_Toc439621010)

[2.4 本章小结 12](#_Toc439621011)

[第三章 系统设计 13](#_Toc439621012)

[3.1 SATS架构设计 13](#_Toc439621013)

[3.1.1 SATS系统架构 13](#_Toc439621014)

[3.1.2 SATS主要模块 14](#_Toc439621015)

[3.2 运行时数据收集模块设计 15](#_Toc439621016)

[3.2.1 运行时数据收集 15](#_Toc439621017)

[3.2.2 SATS运行时数据收集模块设计 16](#_Toc439621018)

[3.3 自适应决策模块设计 17](#_Toc439621019)

[3.3.1 序列化自适应策略设计 17](#_Toc439621020)

[3.3.2 压缩算法自适应策略设计 21](#_Toc439621021)

[3.3.3 垃圾回收自适应策略设计 23](#_Toc439621022)

[3.4 参数优化模块设计 27](#_Toc439621023)

[3.4.1 参数值的初始设定 27](#_Toc439621024)

[3.4.2 优化参数说明 27](#_Toc439621025)

[3.4.3 模拟退火算法的使用 30](#_Toc439621026)

[3.4.4 参数优化模块设计 31](#_Toc439621027)

[3.5 本章小结 33](#_Toc439621028)

[第四章 系统实现 34](#_Toc439621029)

[4.1 系统实现概述 34](#_Toc439621030)

[4.2 运行时数据收集模块实现 36](#_Toc439621031)

[4.3 自适应决策模块的实现 38](#_Toc439621032)

[4.4 参数优化模块的实现 41](#_Toc439621033)

[4.5 本章小结 42](#_Toc439621034)

[第五章 系统测试与结果分析 43](#_Toc439621035)

[5.1 实验方法和目的 43](#_Toc439621036)

[5.2 实验环境搭建 43](#_Toc439621037)

[5.2.1 实验环境 43](#_Toc439621038)

[5.2.2 集群配置 44](#_Toc439621039)

[5.3 自适应决策模块测试 45](#_Toc439621040)

[5.3.1 序列化自适应算法测试与分析 45](#_Toc439621041)

[5.3.2 压缩自适应算法测试与分析 47](#_Toc439621042)

[5.3.3 垃圾回收自适应算法测试与分析 48](#_Toc439621043)

[5.3.4 模块集成测试 50](#_Toc439621044)

[5.4 参数优化模块测试 51](#_Toc439621045)

[5.5 系统性能整体测试 52](#_Toc439621046)

[5.6 本章小结 53](#_Toc439621047)

[第六章 总结和展望 54](#_Toc439621048)

[6.1 全文总结 54](#_Toc439621049)

[6.2 未来展望 55](#_Toc439621050)

[参考文献 56](#_Toc439621051)

[致谢 60](#_Toc439621052)

[攻读学位期间发表的学术论文目录 61](#_Toc439621053)

# 绪 论

## 研究背景与意义

互联网和计算机技术的飞速发展，使得随之产生的数据信息爆炸式增长，21世纪也被称为“大数据”时代。大数据的特点有：数据量大（Volume），新浪微博、facebook等大型交互式网站每天要处理千万甚至上亿条的评论，而企业和科研机构产生在数据在日积月累下也逐渐达到了TB、PB甚至是ZB级别；种类繁多（Variety），既包括了传统的结构化数据，又包括了文本、音频、视频等非结构化数据，而且非结构化的数据比例还在不断增加；价值密度低（Value），往往在大量的数据上才能提取出一点有用的价值，而且预测分析、运营智能和决策支持上难以有效进行；处理速度慢（Velocity），数据处理的时效性难以保证，而实时的处理对商业决策等实时性要求高的数据服务至关重要。

要处理规模庞大的大数据，单靠少数几台计算机的运算能力远远不够，即便是使用超级计算机也不一定能够满足要求，而且由于价格高昂，也不是一种能够普遍使用的方法。分布式计算机技术将处于计算机网络中的计算资源整合成为一个计算机集群来使用，相当于获取了一台超级计算机，而且集群的计算和存储能力会随着新节点的加入而不断增加。因而，分布式技术成为了当前处理大数据性价比最高的技术而被广泛使用。Google推出的有关大数据处理的GFS[1]（Google File System）、Bigtable[2]和MapReduce[3]三篇著名的论文，为分布式平台的开发提供了理论依据。

传统的分布式计算平台，如Hadoop[4]，拥有高性能读写、负载均衡、高容错性等特点，用户在使用时，不必关心底层的实现就可以获得这些良好的系统特性。然而，使用传统的分布式框架处理来处理迭代式运算和交互式查询操作时，运算性能会大大降低[5]。究其原因，这两种任务在运行中会产生很强的数据局部依赖性，即下一次的运算对上一次结果有很强的依赖性，而Hadoop的MapReduce模型会产生大量的中间结果，这些中间结果会被转存到磁盘或是其他外部存储系统中，当下一次的迭代或是查询进行的时候，系统又会将这些中间结果从外部存储系统中取出。由于磁盘的I/O操作代价远远大于内存的存取时间，导致了整个任务的绝大部分运行时间都消耗在了无意义的I/O操作上。那么如果能将这些中间结果存储到内存中，就节省了从外部存储系统读取数据的时间开销。

近几年来，随着变相存储器（phase change memory, PCM）[7]等非易失性随机存储介质技术的发展，产生了新型的混合内存体系[8]，使得内存的性价比不断提升。内存计算技术[9]和存储技术[6]在此前提下逐步发展壮大起来，文献[20]中提出了一种进行分布式集群的内存管理的系统——Tachyon。Spark[10]分布式框架是一种基于内存计算的框架，它能够将任务中产生的中间结果缓存在内存中，节省了从外部存储系统存取数据的时间。Spark相较于分布式内存管理系统[11]具有更加高效的计算速度、容错恢复等特性，并且能够很好地满足机器学习[12]、图计算[13]的大量迭代计算的请求，能够实时地响应交互式查询。使用了内存计算特性的Spark平台，在处理迭代式机器学习算法和交互式查询的任务时，运行速度比Hadoop要高出40倍左右[15]。

然而，由于单个节点的内存资源是有限的，如果任务的规模过于庞大，产生的中间结果大部分都无法缓存在内存中，Spark的运算性能就会受到严重影响。目前Spark针对这个问题并没有对应的优化方案，只提供了一些参数，让使用者针对具体的任务进行调整，这无疑增加了Spark的使用难度。那么，让Spark针对不同的任务自动进行调整以获取最佳性能，推动内存计算技术的发展，是一个很有意义的尝试。

因此，在这一研究背景下，通过对常见任务数据的收集分析，针对不同类型的任务，让Spark在运行时通过动态的收集和分析系统信息，根据分析结果分别采取不同的优化策略以获取最佳的运行性能，这是一个具有广泛的实际应用和商业及科研价值的研究。

## 研究目的和内容

在目前针对Hadoop、Spark等大数据平台的研究中，主要是针对如何提高系统的计算性能和将各种机器学习算法用这些平台提供的API进行重写。目前基于Spark资源管理的研究较少，主要集中在各个节点的数据负载均衡，而对Spark内存优化的研究则更少。内存作为系统中的稀缺资源理应得到合理的应用，特别是对于Spark等基于内存运算的分布式平台，内存的合理应用尤为重要。目前所采用的优化方案是通过手工进行调整，这样的做法不仅过程复杂耗时，而且还需要有经验的专业人员来进行操作，最后得到的结果也不尽如人意。这一点给Spark的使用带来了不便。本文的研究目的是要找出一种比较合理的方法，将内存优化这件事交给计算机来做，而不需要人为参与，并且保证面对不同的工作集的时候，系统能够选择适当的调优方案。

在任务的运行过程中，系统的资源分配是在不断发生变化的，不同时刻消耗的系统资源，如内存资源、CPU资源等，是不同的[17]。因此，一种事先选定的优化方案在针对比较复杂的工作集时不会发挥太大作用，有时候甚至在经过初始时的性能提升之后反而还会造成性能的下降。由此本文的思路是，需要按照不同时间段系统资源的情况对初始时所采用的优化策略进行修改。本课题对Spark自适应调优策略进行了相关研究，并且设计与实现了Spark自适应调优系统（SATS，Spark Adaptive Tuning System），从而使得Spark能够按照系统资源的分配情况来进行自动调优。

要设计一个能够适时调整系统优化方法的算法，首先要进行数据的实时收集工作，并且对这些收集到的信息进行实时地分析，将分析的结果反馈到系统中。这样的做法有两个难点：第一是实时收集的数据也会占用一定的空间给系统的性能造成一些损耗；第二是从收集数据到对数据分析完成需要消耗一定的时间，这段时间造成反馈信息的非实时性，即当系统得到分析结果时，当前系统的状态又发生了转变。因此，要保证自适应算法所带来的空间消耗和反馈的滞后性。其次就是要分析具体的优化方式，Spark提供的优化方案主要有三个方面：第一是进行数据的序列化，序列化不仅能够节省内存，而且还能够加快系统垃圾回收的速度；第二是数据的压缩存储，压缩后的数据往往滞后原来数据的几分之一，甚至十几分之一，这将大大节省内存，但却会消耗更多的CPU资源；第三是优化系统的垃圾回收机制，因为Spark使用scala语言进行编写，而scala语言又运行在JVM上，这个为题就转换为了在Spark环境下针对JVM的垃圾回收优化的问题。

## 研究意义

本文所提出的自适应方法的研究意义主要体现在两个方面。首先是将自适应调优的思路应用到Spark的内存优化上，改变了传统的单纯依赖手工进行调优的方式；其次是提升Spark系统性能。一直以来，对大型系统的优化是业内难以解决且避无可避的问题。分布式系统的设计要考虑到性能、负载均衡、容错性等等诸多问题，是一个复杂的大型系统，而且会随着集群的扩大而变得越来越复杂。越复杂的系统优化起来越困难，有时候优化反而会造成系统性能的下降！Spark在诸多分布式框架中算是一个轻量级框架——核心代码只有5万行左右，而且其最主要的数据结构RDD的设计遵循了简单即美的原则。这些特点减少了Spark内存优化的复杂度，使得对其的优化变得相对简单了一些。Spark的内存计算特性使得其性能和内存的利用率有很大联系，系统的性能会随着缓存在内存中的数据的减少而下降。如何高效利用内存空间是提升Spark性能一个关问题。与Hadoop类似，Spark中的shuffle操作会产生很多的临时对象，这些临时对象占用了大量的内存空间。数据序列化和压缩算法能够为Spark节省大量内存空间，但是如何恰当地使用它们，需要进行精心地设计。本文提出了基于Spark计算特性的自适应算法，算法会在任务运行时收集数据，根据对这些数据进行分析来优化系统性能。

## 论文结构

本文的结构如下：第一章介绍研究背景、研究目的和内容和研究意义；第二章主要介绍一些常见的大数据平台、针对这些平台的优化技术和Spark内存优化技术的国内外研究现状；第三章阐述了Spark的自适应内存优化算法的架构设计，主要是针对数据序列化、数据压缩和垃圾回收优化三个方面的自适应算法设计；第四章将介绍介绍系统的具体实现，包括将Spark的序列化单位从一个RDD细化到一个分区；第五章将描述系统试验部分并对实验的结果进行分析；最后，第六章对全文做一个总结并对未来要进行的工作做一个展望。

# 分布式系统相关技术研究与分析

本章主要研究Spark内存优化技术，旨在解决内存不足的情况下，Spark能够根据运行时的数据信息和内存使用情况，能够自动地调整相关策略，提高内存的使用效率。要想分析上述问题，首先要了解一些常见的分布式框架的特点。其次，是针对这些框架常用的优化策略以及它们的性能和适用情况。最后，阐述了针对Spark的内存优化技术。

## 分布式计算技术概述

### MapReduce计算框架

MapReduce的概念源自于函数式编程语言，它将任务的处理过程分成两个部分：Map（映射）和Reduce（规约）。首先，Map操作会将一个指定的操作运用到待处理数据集中的每一个数据上。Map操作会产生一个新的数据集来存放处理后的结果，因而Map操作是会产生很多的小文件。多个Map操作是可以并行进行的，也就是说可以将一个数据集很大的任务分解成为多个数据量较小的任务来进行处理。在多核环境中，可以让这些任务并行执行，从而加快处理速度。其次，执行Reduce任务的线程在获取Map操作的结果后，会对这些结果执行“规约”操作，即按照指定的操作将接收到的结果进行归并。同样的，Reduce操作也是可以并行执行的，即可以让多个Reduce任务接收多个Map任务的结果。

MapReduce计算框架充分利用了系统多核资源，大大加快了任务的执行速度。根据MapReduce思想实现的Hadoop框架提供了高性能，高并发，高可靠性，良好的负载均衡机制和容错机制。目前，MapReduce计算框架被广泛应用到大数据处理领域中，并取得了巨大的成功，成为当下最流行的分布式编程模型之一。并且，针对实现了MapReduce模型的Hadoop大数据平台还在不断的发展完善之中，例如，Lizhe Wang等人设计的G-Hadoop[33]系统将Hadoop的计算特性从单集群发展到了多集群，Ablimit Aji和Fusheng Wang等人通过研究数据的空间局部性，来进一步提高Hadoop的性能[34]，Hui He、Zhonghui Du等人针对MapReduce中最为消耗系统资源的shuffle操作进行了优化[35]。但MapReduce框架在处理迭代式运算和交互式查询时，性能稍显不足。同时，MapReduce的编程接口过于严苛死板，缺乏灵活性。

### Tez计算框架

Tez[32]是针对Hadoop数据处理应用程序的分布式执行框架。与MapReduce不同，Tez采用DAG作业调度的方式，将多个有依赖的作业转换成为一个作业从而提升DAG作业的性能。在一些不适合应用MapReduce的场景，例如机器学习算法，Hadoop可以借助Tez来完成。Tez的目的是高度订制化，即让人们不必借助外部方式就能够顺利完成自己的工作。为了达到这个目的，Tez提供了更具表现力的数据流API，让用户能够形象的描述自己所要运行的DAG作业图。

Tez灵活的输入—处理器—输出的运行模式对比与MapReduce的映射—规约，显得更加灵活。此外，Tez的数据类型无关、动态图重新配置、部署的简单性等特性，大大拓展了其应用领域。

### Spark计算框架

在大数据处理领域，根据处理时效性的不同，可将大数据框架划分了两种不同的类型：脱机型和实时型。Hadoop的MapReduce和Tez都属于前者，它们处理大数据的时效性都不高，无法为使用者提供实时的服务。Hadoop虽然高效但依然存在一些问题，在处理实时性任务的性能上远远不如Spark。为了获得高性能，Spark简化了内存数据结构，通过将任务处理的中间结果缓存在内存中的方式，在处理迭代式运算和交互式查询的任务上，获得了超过Hadoop框架40倍以上的速度。此外，巧妙的设计还让Spark在高可靠性和高容错性上也不逊色于其他的分布式计算框架。

目前，Spark提供的Mlib[26]所支持的机器学习算法越来越多，成为了众多企业和研究机构所依赖的机器学习算法工具。

## Spark框架研究

Spark是典型的master-slaver分布式架构，其集群架构图如图2-1所示。集群中有一个driver（master）和多个worker（slaver）节点，driver节点负责将任务分发到各个worker节点，并接受worker节点发送回来的执行结果；worker节点则执行分配的任务，并将执行结果发送到driver节点。

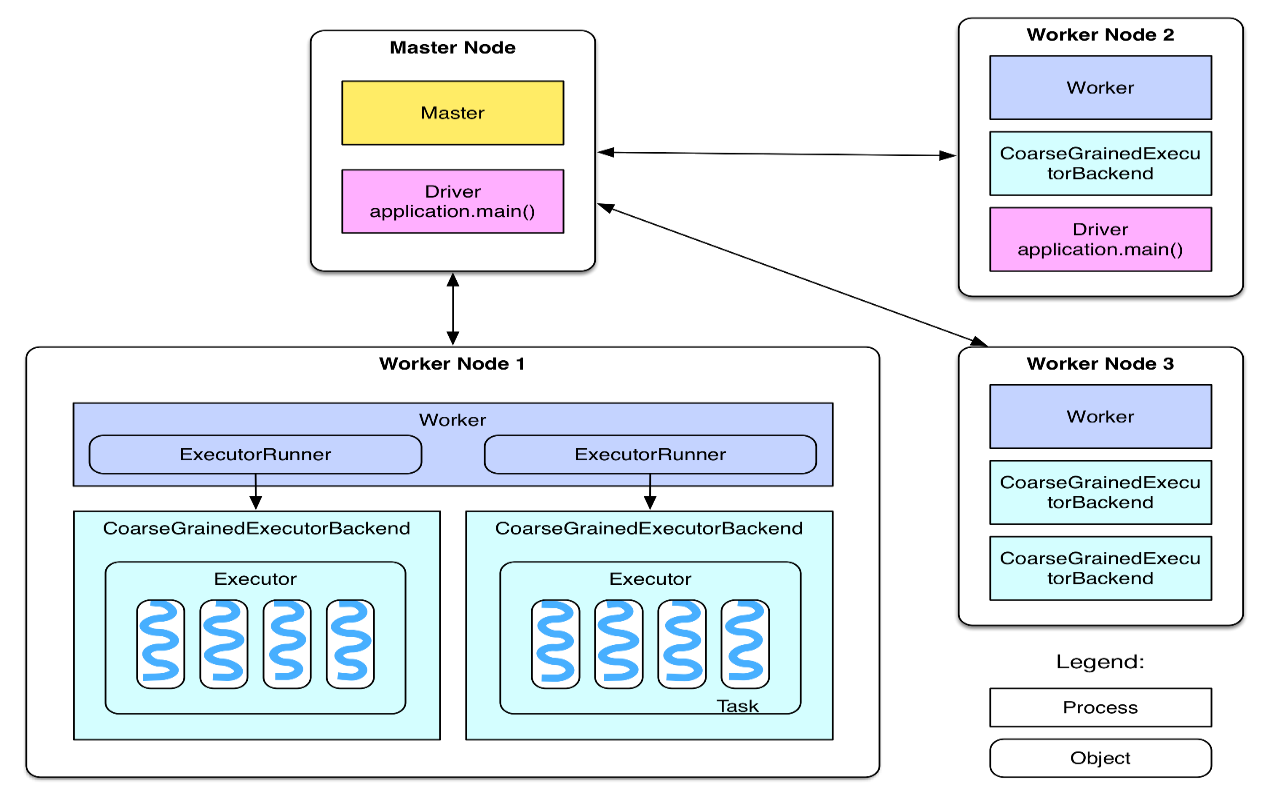


图 2‑1 Spark部署图[39]

Figure 2-1 Spark deployment diagram

### 内存抽象RDD

RDD（Resilient Distributed Dataset）是Spark内存计算框架的核心。RDD的设计遵从了简单即美的原则，它只支持一些粗粒度的操作，如批量读、查询等。RDD是不支持写操作的，这样的设计巧妙地避开了数据一致性问题。RDD中包含了用于存放数据的数据分区（partition）、元数据信息和对父RDD或是原始输入数据的一组依赖（Lineage）。其中，根据数据依赖（Lineage）生成的DAG图，是Spark任务执行的基础。RDD的数据依赖分为两种类型：NarrowDependency和ShuffleDependency，这两种类型描述了两个RDD中partition的相关关系，两者的区别如图2-2所示。从图中可以看出，NarrowDependency关系是整个partition的相关性，而ShuffleDependency是partition部分相关关系。

Hao Zhang等人对RDD的在内存计算中的性能进行了实验分析，验证了RDD在内存运算上相对于其他内存模型更为高效[16]。Spark应用程序开发主要是针对RDD的操作来进行的。在Spark的官方文档中，RDD的操作可以分为两类：transformation和action。每个RDD或是输入的原始数据都可以通过一个transformation操作生成新的RDD，但是这个过程并没有触发任务的执行，也就是说通过转换操作得到的RDD中只包含了一些元数据信息和依赖关系，缺少数据分区。而actions操作则会触发任务的执行。



图 2‑2 RDD依赖类型

Figure 2-2 RDD dependency type

### Spark的Job生成

Spark应用程序开发主要是针对RDD的操作来进行的。在Spark的官方文档中，RDD的操作可以分为两类：transformation和action。每个RDD或是输入的原始数据都可以通过一个transformation操作生成新的RDD，但是这个过程并没有触发任务的执行，也就是说通过transformation操作得到的RDD中只包含了一些元数据信息和依赖关系，缺少数据分区。而action操作则会触发任务的执行，得到执行的结果数据信息。transformation操作建立起了每一个RDD之间的依赖关系，这些相关的RDD组合在一起构成了一张任务执行DAG（Directed Acyclic Graph）图，该图描述了这个任务的执行流程，这个过程称为Spark的逻辑执行流程。action操作触发了真正的Spark任务执行，计算得到了结果数据信息，这个过程称为Spark的物理执行流程。

### Spark任务执行流程

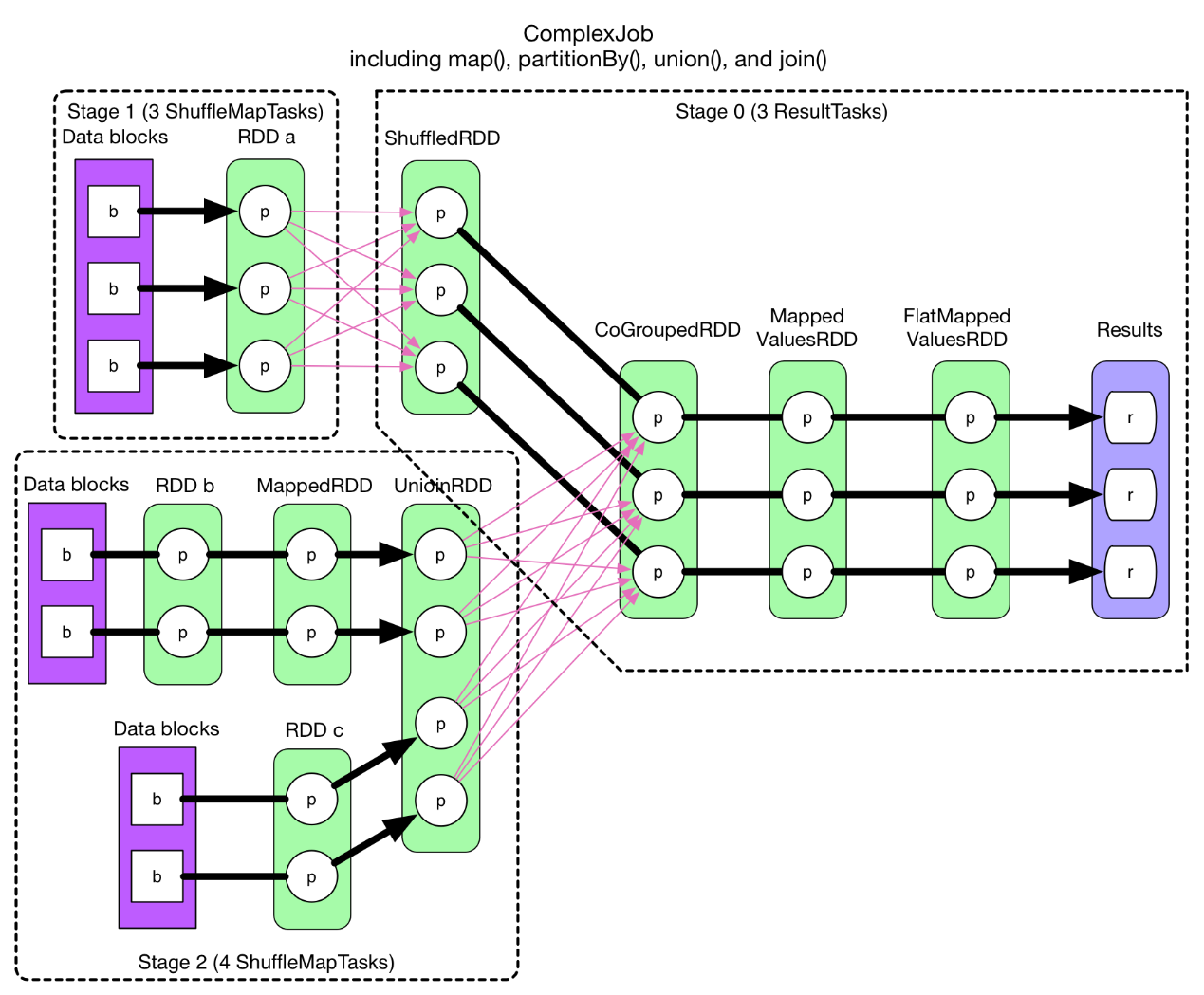


图 2‑3 DAG任务执行图[40]

Figure 2-3 DAG job execution

Spark的执行流程是如何执行由transformation生成的任务的过程。Spark调度器会将DAG图切分成多个小的子图，这些子图被称为stage。那么如何来划分stage呢？Spark采用了从后往前进行划分的方法，即从最后一个RDD开始向前寻找它所依赖的parent RDD。如果它们之间的依赖关系是NarrowDependency，那么继续往前寻找；如果是ShuffleDependency依赖就断开，断开的部分组成一个stage。按照同样的方法继续寻找，直到所有的RDD都加入了一个stage为止。如图2-3是一个DAG图划分stage的例子。

Spark在进行任务提交的时候，也是从后往前依次提交stage。stage中最后一个RDD称为Final RDD，在计算Final RDD时，任务会向前找它所依赖的parent RDD获取数据，如果它的parent RDD中的数据也没有准备好，该parent RDD同样会向前寻找依赖的RDD获取数据，如此类推。当前一个RDD的数据准备好了时候就可以通过运算得到它的child RDD中的数据了。每个stage所产生的任务数量和其Final RDD的数量一致。当程序执行到action操作的时候，就会触发这些任务的执行。

## 分布式框架优化技术研究

每当系统的规模扩大10倍的时候，原来的系统就已经不能正常运行，需要进行重新设计。这意味着，随着系统的不断扩大，需要对平台进行优化，才能保证系统能够发挥最大的性能。文献[24]中描述了用来进行大数据调优的主要方法。分布式系统的优化方式有很多种，但基本上都是通过对参数的调整来实现，本节主要讨论当前一些分布式平台的调优方案。

### 参数空间优化法

比较常见的调优方法是通过待调优的大数据平台所提供的参数来实现，这样的调优方法称为参数空间优化法。具体来说，为了能够使系统在针对具有不同特性的工作集的时候，能够发挥其最佳性能，主流的大数据平台都提供了大量的参数供开发者进行选择。这些参数的所有可能组合组成的空间称之为参数空间（configuration space），举个例子来说，假设系统提供的可调整参数有N个，每个参数可能的值有M种，那么这个平台的参数空间中有个元素，每个元素代表一组参数配置组合。得到这个参数空间之后，就可以针对某个具体的工作集来得到最优的参数配置方案：将参数空间中的每一种配置分别测试一遍，性能最优的结果所采用的配置就是最优配置。文献[27]、[30]和[31]中描述了如何进行系统参数空间的配置优化，以及进行参数空间优化后所得到的效果。当处理相同或类似的工作集的时候就可以使用之前找到的参数配置，从而达到调优的目的。使用这种方法能够针对不同类型的工作集得到性能最优的参数配置，但却耗时巨大。试想当一个新的工作集到达的时候，要用参数空间的每一种配置进行一一测试，这种做法的时效性无疑是非常低下的。因而这种方法通常会结合其他方法一起使用。

### 历史数据优化法

在任务的运行中会产生大量的任务日志，除此之外开发者也可以根据自己的需求收集一些特定的信息。利用这些历史信息来进行系统的优化是另一种比较常见的大数据优化方法。举例来说，在Quan Chen等提出的HAT[25]系统中，会给每一个任务分配一个权值，再通过收集到的历史信息对任务的权值进行调整，最后根据权值的大小来确定哪些任务是会最终影响系统运行总时间的，并通过加速这些任务的执行速度来提升总体的运行速度。Fuller N C，Li M等人提出了在Hadoop框架中如何根据前一次任务运行的信息，来决定为接下来的任务分配多少内存，从而达到性能优化的目的[19]。Spark生态圈中的Shark[14]系统也提出了一种利用历史数据来加速系统运行的技术——PDE（Partial DAG Execution），该技术利用在先前执行的任务中收集到的相关信息，例如数据输入大小、输出大小、计算时间等，来优化当前的查询任务。虽然Shark已经被Spark SQL[18]淘汰，但是它在针对Spark进行的优化技术上，还是值得借鉴的。历史数据优化法能够根据历史信息来预测任务的执行特征，根据这些特征来进行调优。其缺点就是根据历史数据来分析当前执行的任务特征有时候会存在很大的偏差，因为即便是相同的任务运行多次之后的结果也不尽相同。这是因为不同时刻系统资源的分配可能会有些差异，这会影响到计算框架中使用的算法对任务的划分和执行。

### 融入机器学习的混合优化法

在前面提到的参数空间优化法中，往往某一个分布式计算框架提供的参数是非常多的，直接导致了参数空间过于巨大。如果使用一一代入测试的方法，会消耗很多的时间，而且每当一个新的任务到达的时候都要使用这样的代入法进行测试，这对于一些实时性要求高的任务和资源稀缺的集群来说会造成很大的负担。要解决这个问题可以从参数空间入手。仔细分析就会发现，对于某一个具体的任务，参数空间中的很多配置方案明显都是不可行的。如何排除这些明显不符合条件的配置方案成为了简化参数空间的关键。

通常情况下有两种方法可以用于解决这个问题。其一是使用人工筛选的方法，针对常见的任务特征制定筛选条件。这样的做法对筛选条件的制定有很高的要求，如果筛选条件太少，那么能够甄别出来的任务特征差异过少，会导致筛选后的参数空间依然很大；如果筛选条件过多，那么还要考虑到这些筛选条件之间会不会产生冲突，以及由于筛选的条件过多过于苛刻而导致最优配置方案可能被剔除掉，造成过度过滤的后果。而且随着任务的运行，会产生越来越多的任务特征，还要针对这些新产生的特征指定新的筛选条件，这无疑会增加系统的维护成本。另外一种方式就是使用机器学习算法来解决这个问题，即选择一种合适的机器学习算法来快速找到参数空间中的最优配置。在Min Li等人提出的MRONLINE[23]系统中，就通过采用一种智能爬山算法来快速找到MapReduce任务最优的配置方案。文献[28]和[29]提出了使用搜索算法来寻找最有参数配置的方法。

### Spark优化策略

Spark是新型的基于内存计算的分布式框架，与传统分布式平台的优化方式不同的是，针对Spark的优化更加注重对内存使用的优化，能否高效地利用内存资源是决定Spark系统性能优劣的关键。首先，需要了解的是内存计算的一些独特性质，Tao Jiang等人分析了内存运行的特点，并对比了不同框架在内存运算上的性能表现[17]。Kaur M、Verma N等人研究了内存分配对系统性能的影响，阐述了内存动态分配技术对提高内存运算性能的重要性[21]。

根据目前的研究结果，可以从两个方面来考虑Spark的优化方案，第一是减少运行时所需数据块的缺失所造成的“抖动”，第二是是避免频繁的GC操作影响系统性能。常用的优化手段有采用高效的序列化算法序列化数据，采用高效的压缩算法压缩数据，以及结合Spark系统自身的状态对JVM过程进行优化。序列化方法在减少数据占用空间的同时，还保证了数据在网络上的稳定传输，以及减少了GC操作次数。目前，Spark继承了两种序列化方法：Java序列化和Kryo[22]序列化方法。压缩数据使得要存储的数据降低到了原来的几分之一，甚至十几分之一，是节省内存空间的最有效的手段。可以根据Spark提供的系统参数来实现对以上三个优化方案。

## 本章小结

本章首先介绍了Spark内存运算的过程和特点，其次详细讨论了序列化、压缩和内存垃圾回收三种主要的优化方法。最后分析了人工手动调优存在的弊端，提出了自适应调优的方法。下一章将针对本章提到的序列化、压缩和垃圾回收三种优化方法，设计出相对应的自动调优方法。

# 系统设计

自适应调优摒弃了手动调优的传统调优方式，在Spark运行过程中动态收集系统信息，根据对这些信息的分析结果对系统中的序列化方法、压缩算法和垃圾回收策略进行自动调整。本章首先阐述SATS系统的结构，并对各个模块进行说明。在之后的3.2、3.3和3.4节中将分别介绍系统各个模块的设计细节。

## SATS架构设计

本节从两方面介绍自适应优化系统架构。3.1.1将介绍整个SATS系统架构，3.1.2将描述架构的主要模块。

### SATS系统架构



图 3‑1 SATS系统架构图

Figure 3-1 the system architecture of SATS

本小节将介绍Spark自适应优化系统（SATS）框架设计。SATS是一个分层结构系统，系统架构图如图3-1所示。整个SATS系统分为三个模块，分别是：运行时数据收集模块（RDCM，Runtime Data Collection Module）、自适应决策模块（ADM，Adaptive Decision Module）和参数优化模块（POM，Parameter Optimization Module）。系统在Spark应用程序任务运行的时候开始运行，当整个Spark应用程序的任务运行完成时，SATS系统不会停止运行，而是会根据Spark任务运行过程中收集到的数据对SATS系统的自身和部分系统参数进行优化。当下一个Spark应用程序的任务开始运行时，SATS会停止当前的参数优化活动，并保存当前的计算结果，然后开始处理新的Spark任务。

### SATS主要模块

本节将对SATS系统的三个主要模块——RDCM、ADM和POM进行详细介绍，如下：

RDCM主要负责在Spark任务运行时收集系统和任务相关信息，例如系统CPU消耗、内存消耗、RDD的大小和计算耗时等等。这些收集到的数据必然会占用一定的存储空间，这对基于内存计算的Spark框架来说是一个不小的开销。如果将这些收集到的数据存放到外部存储设备当中，那么存取这些数据会造成额外的时间开销，而且这些数据的时间局部性相关性非常大（收集到的数据会在之后的数据分析模块中被用到）。因而SATS系统会将这些数据存储在内存中，为了节省内存的开销，系统会使用循环存储结构来存放数据。

ADM是SATS系统的主要模块，该模块负责分析运行时数据收集模块收集到的数据，并通过一套自适应决策算法根据收集到的数据对当前系统的存储策略进行调整。模块中又包含了三个子模块：序列化算法自适应模块、压缩算法自适应模块和垃圾回收自适应模块，这三个子模块会在任务运行时，动态调整系统的数据序列化算法、压缩算法和垃圾回收策略。ADM模块会接收来自RDCM的数据信息，将这些数据进行处理后作为算法的输入。

POM会对运行时收集到的数据进一步进行分析，通过分析的结果对自适应决策算法中使用到的阈值参数，以及Spark系统参数和JVM相关参数进行调整。POM模块主要针对三种类型的参数进行调优工作，首先是JVM参数，Spark框架运行JVM虚拟机上，因而，JVM的性能和Spark的系统性能息息相关。其次是Spark系统参数，这类参数的初始默认值在针对不同的应用程序时，不一定是最好的选择，POM模块将根据系统信息找到最适合当前任务的配置。最后是对ADM模块中的阈值参数进行调优，ADM模块中的阈值也是人为手动设定的，同样需要针对不同的应用进行择优选择不同的参数配置。

这三个模块在Spark任务运行过程中相互协作，共同完成对Spark系统自适应优化操作。

## 运行时数据收集模块设计

本节将介绍SATS系统运行时数据收集模块的设计。运行时数据收集模块主要负责在Spark任务运行过程中收集相关数据，这些相关数据的收集和存储涉及到之后的自适应决策模块是否能够获取运行所需的数据。

### 运行时数据收集

在单个节点中，系统的存储能力和计算能力是主要的系统资源。对于Spark这类基于内存计算的分布式框架来说，内存的容量是限制系统性能的重要瓶颈，外部存储系统的大小虽然也对系统性能有所影响，但远远不如内存的影响大。这是因为CPU访问内存的速度比访问外部存储设备的速度通常要快100倍以上。在SATS系统中主要考虑CPU和内存资源对Spark系统性能的影响，因此，CPU和内存的实时消耗情况就成为了主要收集的数据信息。RDCM主要收集的数据如下所示：

1. 内存数据收集

Spark运行在JVM上，因此它所使用的内存只是JVM所管理的内存中的一部分。具体来说JVM运行时数据区分为五个部分，如图3-2所示，Spark用来缓存数据的区域是JVM堆区，参数spark.storage.memoryFraction用来指定用来缓存数据的JVM堆空间所占的比例。该参数的默认值为0.6，表示使用60%的JVM堆空间来缓存数据，另外的40%用来存放JVM运行时产生的java对象。SATS系统将内存的使用情况分为两类：一类是用来存放java运行时产生变量的JVM堆空间的使用情况，另一类是JVM堆空间中用来缓存数据的空间的使用情况。这两类空间都会对Spark的性能造成影响，当用于存放运行时java对象的JVM堆的空间使用率过高的时候，Minor GC操作就会频繁发生（GC操作只会发生在用来存放java对象的JVM堆空间上），Full GC操作发生的概率也会增大，从而导致系统运行速度变慢；当用来缓存数据的JVM堆空间使用比例过高时，发生Spark任务运行数据缺失的概率就会增大，Spark会采用一定的策略（默认是LRU）对JVM堆中的数据进行替换，这就造成了内存数据的“抖动”，同样也会影响系统的运行性能。因而，RDCM模块要收集这两类内存的使用情况。

1. CPU数据收集

之所以要收集CPU资源的使用情况情况，主要是为了在之后的自适应决模块中为序列化自适应算法和压缩策略自适应算法提供数据。在这两个自适应算法中，CPU使用情况是用来进行算法选择决策的一种重要参数。



图 3‑2 JVM运行时数据区图

Figure 3-2 JVM runtime data area diagram

1. JVM信息收集

Spark运行在JVM上，JVM的性能调优对Spark系统性能的提升影响非常大。对JVM的调优主要包括两个方面，首先是通过JVM的运行信息来调整用于缓存和用于运行时创建对象的JVM堆空间的比例（通过参数spark.storage.memoryFraction来实现）；其次是对JVM中的参数进行调整，具体来说是调整JVM新生代和老生代所占的比例，以及新生代中各区域所占的比例。在Spark中可以通过JVM相关配置参数-verbose:gc –XX:+PrintGCDetails –XX:+PrintGCTimeStamps将有关的GC信息打印出来。

### SATS运行时数据收集模块设计

Spark任务执行过程是在Executor类中完成的，SATS系统收集收据的开始时间就是Executor开始执行第一个任务的时间，直到最后一个任务结束，数据收集的过程才结束。为了设计上的简易性，SATS将收集数据过程看做是一个特殊的任务，当Executor开始执行第一个任务时，这个特殊任务会占用一个线程而启动；当最后一个任务结束时，该任务释放掉持有的线程，结束数据收集。这样的做法使得不仅让系统的资源消耗到达最低（只占用了一个线程），而且还大大地简化了数据收集模块的设计，即不用开辟单独的接口专门来进行数据收集，使用Executor中的接口就可以了。

## 自适应决策模块设计

本节将介绍SATS主要模块——自适应决策模块的设计。该模块涉及到三个方面的优化设计：数据序列化、数据压缩和垃圾回收。前两个方面在Spark系统中同属于与存储相关的层次的内容，因而在自适应算法的设计上有一定的联系性。

### 序列化自适应策略设计

1. Spark序列化技术分析

图 3‑3 Java序列化算法与Kryo序列化算法压缩效果对比图（压缩大小）

Figure 3-3 Java and Kryo serialization algorithm performance comparison on data size

序列化主要有两个方面的作用。首先，java对象在网络上的传输需要通过序列化来实现。序列化能够将对象转换成为字节流在网络上进行传输，接收端接收到传输的字节流之后通过反序列化对传输的对象进行重构。其次，序列化对于节省内存空间和提高垃圾回收效率有重要的意义。多个同一类型的java对象经过序列化之后可以转换成为一个字节数组，这种方式不仅节省了内存空间，而且减少了垃圾回收时所需要遍历的java对象个数。因此，正确地应用序列化技术可以对Spark性能带来很大的提升。

Spark内置了两种序列化算法：Java和Kryo，Java序列化算法是java语言自带的，使用起来比较灵活，只需让需要被序列化的类继承Serializable 接口即可。Java序列化算法虽然使用起来比较灵活，但其缺点就是序列化的速度过慢，现在主流的序列化方法在序列化速度和压缩率上普遍都比Java序列化方法要好。Kryo序列化算法要比Java序列化算法高效很多，但是在使用的时候，需要对要被序列化的对象所属的类进行注册，如果没有注册的话，有可能会发生异常。目前，Spark已经自动将本身类在Kryo上进行了注册。两者在压缩大小的上的性能比较如图3-3所示，图3-4展示两者在序列化时间上的性能差异。

图 3‑4 Java序列化算法和Kryo序列化算法性能对比图（序列化时间）

Figure 3-4 Java and Kryo serialization algorithm

performance comparison on cost time

但是，并不支持对用户自定义类的注册，在使用时，用户需要自己进行注册。Kryo存在的问题就是比较消耗CPU资源。Spark在默认情况下是使用Java序列化算法来对数据进行序列化，这里的序列化包括了两个方面：首先是闭包序列化，目前Spark只提供了Java序列化方法对闭包进行序列化；在经过闭包序列化之后，Spark会对序列化后的数据闭包再进行一次序列化，在这个过程中有两个序列化方法可以选择，即Java序列化方法和Kryo序列化方法。由此可见，目前所能做的优化集中在了第二次序列化过程中，也就是在什么情况下对两种序列化方法进行切换。

进行序列化调优的目的是为了加快任务的运行速度，Kryo序列化方法虽然比较节省空间，但是相对的，却增大了对CPU的消耗，而CPU的消耗会降低任务的运行速度；另一方面，Java序列化方法虽然对CPU的消耗很小，但其序列化后的数据大小却比Kryo序列化后的数据大得多，导致能够被缓存在内存中的数据量下降，增大了系统从外部存储系统存取数据的概率，也会影响到任务的执行速度。因而，对于如何选择合适的序列化方法，就必须综合考虑CPU消耗和内存消耗两方面的情况。

1. 阈值设置分析

那么应该采取什么样的策略进行调优呢？首先来考虑序列化算法切换的条件问题。通过刚刚的讨论可以看出，Java序列化算法的优势在于其对CPU消耗比较少，而Kryo算法的优势在于序列化后的数据所占用的空间少。在Spark官方文档给出的调优方案中提到：节省内存空间能够更多地缓存数据，而其根本目的是为了加快任务的运行速度。显然节省空间只是一种手段，而加快任务的执行速度才是调优的目的。CPU的高消耗会直接性地导致任务的运行速度变慢。相较而言，缓存数据变少给Spark任务运行带来的影响并不是直接性的。

具体来说，Spark的缓存策略使用的是LRU（Least Recently Used）算法：当Spark任务运行所需的数据不在内存中的时候，LRU算法会将缓存在内存中最近未被使用的数据块清理掉，替换成为从外部存储系统中刚刚读取到的（或是重新计算得到）任务所需数据块。当系统使用Kryo序列化算法时，内存中能够缓存的数据块要比Java序列化算法多得多。这样就大大降低了数据块的Miss率，减少了内存的抖动，从而加快了任务的执行效率。由此分析可知，缓存数据的减少只是导致了数据Miss率的增多，任务所消耗的时间有很大的概率会增加，但却不会直接性地导致任务的运行速度变慢。由此可以得出的结论是：在进行序列化调优时，首先要考虑到CPU的消耗情况，其次再考虑能够缓存数据的多少。

其次要考虑的是资源阀值设置问题。算法初次运行时，将系统的CPU阀值初始设定为90%，当CPU的使用率超过这个值的时候，首先判断当前系统所使用的序列化算法是不是Java序列化算法，如果不是，则将当前的序列化算法切换为Java序列化算法。缓存阀值的设定会稍微复杂一些，Spark将JVM堆空间分成两个部分来使用，一个部分用于缓存RDD数据，另一部分用来存储系统运行时产生的对象。这两个部分的比例通过参数spark.storage.memoryFraction来进行调整，因此，在此设置了两个内存阀值，一个表示整个JVM堆使用阀值，一个表示JVM堆中用来缓存RDD数据内存容量的阀值。为了算法的简洁性，将这两个阀值上限都设置为90%。同时将阀值的下限设置为40%，当内存使用低于这个值得时候将当前的序列化算法设置为Java序列化算法。

为了增加系统的灵活性以及提升对系统存储空间的利用率，SATS系统对Spark原来的数据序列化和数据压缩基本单位进行了修改：将数据操作的基本单元由一个RDD改为了一个Partition。采取这样的做法使得当一个RDD中包含多个Partition的时候，SATS系统可以使用不同的数据序列化算法或者数据压缩算法对这个RDD中不同Partition进行处理，从而提高了SATS系统对存储空间的支配能力。

1. 序列化自适应算法设计

在收集完数据后ADM模块会根据收集到的数据，采用如图3-5所示的算法，对Spark所使用的序列化方法进行调整。

|  |
| --- |
| **Algorithm 1**: Adaptive Serialization  **Input**: cpuUsg，memUsg，memCacheSize  **Output**: curSer |
| 1. if cpuUsg > ***cpuThr*** do 2. ***cpuOTIMES*** += 1 3. if ***cpuOTIMES*** == 3 do 4. curSer = JavaSer 5. ***cpuOTIMES*** = 0 6. endif 7. else 8. ***cpuOTIMES*** = 0 9. if memUsg > ***memRUpperThr*** || memCacheSize >***memCacheUpperThr*** do 10. curSer = KryoSer 11. elif memUsg<***memRLowerThr*** || memCacheSize<***memCacheLowerThr*** do 12. curSer = JavaSer 13. endif 14. endif |

图 3‑5 序列化自适应算法

Figure 3-5 Serialization Adaptive algorithm

* 参数说明：

cpuThr——CPU的使用阈值

cpuOTIMS——记录CPU连续超过阈值的次数

memRUpperThr——非缓存RDD内存使用上限

memRLowerThr——非缓存RDD内存使用下限

memCacheUpperThr——缓存RDD的内存使用上限

memCacheLowerThr——缓存RDD的内存使用下限

cpuUsg——当前系统CPU的利用率

memUsg——已使用的非缓存RDD的内存大小

memCacheSize——已使用的缓存RDD的内存大小

curSer——当前使用的序列化方法

JavaSer——Java序列化算法

KryoSer——Kryo序列化算法

算法中考虑了两个主要影响系统性能的因素：CPU使用率和内存使用率。算法每隔1s，会根据当前的系统资源状况对序列化算法进行调整。算法首先考虑CPU的使用情况，当CPU的使用率超过设定的阈值cpuThr时，将系统当前的序列化方法改为Java序列化方法，以降低CPU的消耗。若CPU没有超过阈值，则查看内存的使用情况。值得注意的是，由于环境和系统因素，在某一时刻，CPU的使用率会急剧升高，这种情况并不是CPU使用率过高，因而不必对算法进行切换。为了避免这种状况，算法默认在侦测到CPU使用率连续3次超过阈值的时候，才会对序列化算法进行切换（参数cpuOTIMES用来记录CPU利用率连续超过阈值的次数）。

根据JVM堆使用类型的不同，算法设置了两种内存阈值，即用于缓存RDD的内存阈值和用于存放运行过程中产生的临时变量的阈值。这两个类型的内存阈值都有上限和下限，当两种类型的内存使用率有一个超过阈值上限的时候，将当前的序列化方法切换为Kryo，以节省内存；否则，当两种类型的内存使用率有一个低于阈值下限的时候，将当前的序列化方法调整为Java序列化方法。

算法中的阈值，在参数优化模块中会进行调整。因此，对于阈值的初始值，开发人员可以自行设定也可以使用默认的初始值。

### 压缩算法自适应策略设计

压缩算法能够将数据压缩到原来的几分之一甚至十几分之一，正确地应用压缩算法能够很大地节省空间，使得能够存放到内存中的数据量大大增加，从而提高任务的执行效率。本节讨论如何在Spark任务运行中采用自适应策略选择适当的压缩算法来节省内存空间，从而提高任务的执行效率。

1. Spark压缩算法分析

Spark一共内置了三种压缩算法：Snappy、LZF和LZ4。有关这三种压缩算法的对比如表3-1所示。这三种压缩算法各有优缺点。Snappy压缩算法着重于加快压缩速度，而不考虑压缩率（压缩后的文件体积和原文件体积的比例）和兼容

表 3‑1 压缩算法对照表

Table 3-1 comparison of compression algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| 算法名称 | 压缩速度 | 解压速度 | 特点 |
| Snappy[36] | 250MB/s | 500MB/s | 速度快，压缩率高 |
| LZF[37] | 150MB/s | 200MB/s | 速度最慢，压缩率最低 |
| LZ4[38] | 400MB/s | 1GB/s | 速度最快，压缩率最高 |

注：，压缩率越低节省空间越多

性。其单核压缩速度达到了250MB/s，解压缩速度为500MB/s。LZF算法着重于减小压缩后数据的体积，因而具有很高的压缩效率，相比于Snappy能够节省更多的空间。但是，LZF的压缩速度却不是很快，单核压缩速度为150MB/s，解压速度为200MB/s。LZ4在这三种压缩算法中，是压缩品质最好且压缩速度最快的。为了保证压缩的高品质，LZ4的压缩效率并是不很高。压缩算法对不同的数据集所表现出来的性能大不相同，那么就要根据数据类型的不同采取不同的压缩算法，以达到最优效果。

1. 压缩算法自适应策略设计

|  |
| --- |
| **Algorithm 2:** Adaptive Compression  **Input**: memUsg  **Output**: curCodec |
| 1. if memUsg < ***memLowerThr*** do 2. curCodec = “” 3. elif *memUsg* >***memLowerThr*** && *memUsg* <***memMiddleThr*** do 4. curCodec = “lz4” 5. elif memUsg > ***memMiddleThr*** && memUsg < ***memUpperThr*** do 6. curCodec = “snappy” 7. else 8. curCodec = “lzf” 9. endif |

图 3‑6 压缩自适应算法

Figure 3-6 Compression Adaptive Algorithm

* 参数说明：

memUsg——已使用内存大小

memLowerThr——内存使用低，

memMiddleThr——内存使用高，

memUpperThr——内存使用上限，表示系统对内存的消耗达到瓶颈

curCodec——当前系统所使用的压缩算法

压缩算法自适应策略如图3-6所示，算法中设置了三个阈值：memLowerThr、memMiddleThr和memUpperThr，用来判断内存的使用级别。内存使用量没有超过memLowerThr，表示内存使用量低，这时候没有必要使用压缩算法对缓存数据进行压缩存储。超过了memLowerThr，但低于memMiddleThr表示缓存的数据开始增多，这时候使用LZ4算法，该算法能在节省能存的同时保证较快的压缩速度。当内存使用值超过memMiddleThr，低于memUpperThr时，表示内存的使用已经达到一个较高的程度，这时候需要进一步节省内存空间，在这个阶段采用压缩率更高的Snappy算法。当内存使用量超过memUpperThr时，表示内存使用已经达到了瓶颈，在这个阶段内存嫉妒消耗，GC操作频繁发生，还有产生内存溢出的可能，这时候使用压缩率最低的LZF算法。

在内存比较充足的情况下，首先考虑采用LZ4压缩算法，不仅能够获取较高的压缩速度，还能够保证数据的正确率。随着Spark任务的运行，用于缓存RDD的内存空间逐渐减少。这时候就可以使用Snappy算法，在保证一定压缩速度的同时，减少压缩数据所占用的空间。当内存资源紧缺，频繁发生GC操作，甚至是内存溢出的时候，可以将压缩算法切换为LZF，进一步减小压缩后的数据体积。这三个阈值的初始值，在设定之后，会在参数优化模块中进行优化调整，开发人员只需根据自己的经验进行设置，或是直接使用默认的初始值即可。

### 垃圾回收自适应策略设计

相较于Hadoop等传统大数据处理框架，Spark的最大的优势在于能将计算过程中产生的中间结果存储在内存中，以便于对数据进行重复使用，减少从磁盘取数据的I/O开销。但当内存不足以存放所有的中间结果的时候，JVM会对内存中的对象进行回收。JVM的垃圾回收策略将近期不使用的java对象释放掉，这意味着JVM要将所有的对象都要遍历一遍。由此可见，过于频繁的垃圾回收会导致系统性能下降。本节将讨论Spark的内存回收策略。

Spark使用Scala语言编写，Scala语言运行在JVM上，那么针对Spark的垃圾回收优化与JVM的垃圾回收优化有很大的关系。本节主要讨论如何使用自适应的方法针对Spark垃圾回收进行优化。

1. JVM垃圾回收策略分析

Spark中垃圾回收调优的目的是要将长期存活的对象存放到老生代中，而且要尽量保证新生代区足够大以存放新生的对象，这样可以避免任务运行时由于回收临时变量导致的Full GC操作。为了达到这个目的，较为常用的调优手段有：

1）通过运行时收集GC信息，判断垃圾回收是否频繁发生。如果再一次任务开始之前发生过了过多的Full GC，说明JVM堆中用来存放临时变量的空间过少，那么可以将这部分空间的比例调大。

2）在收集到的GC信息中，如果老生代区域剩余空间不多，那么可以降低缓存数据所占空间的比例。因为获取较高的执行效率才是调优的目的，而缓存数据只是实现这个目的的一种手段。在java运行时状态区中，JVM堆的新生代中存放存活时间比较短的java对象，老生代中存放存活时间比较长的对象。而新生代区进一步又可以分为Eden，Survivor1，Survivor2三个更小的区域。JVM运行过程中产生的新生对象会被存放到Eden与Survivor1中，当这两个区域的空间不足以存放所有的新生对象时，会发生一次Minor GC操作。Minor GC会将新生代前两个区域中的存活对象拷贝到Survivor2中，然后交换Survivor1和Survivor2中的内容。在此过程中，如果Survivor2的空间不能存放前两者的对象，那么这些对象会被存储到老生代区中，如果老生代中剩余的空间也不能存放这些对象，那么将发生一次Full GC。相对于Minor GC，Full GC是相当耗时的，因为在进行Full GC时，JVM虚拟机会将整个内存中的对象都遍历一遍，以找到不活跃的对象进行回收。

3）如果发生过多的Minor GC，而鲜有full GC发生，则意味着新生代中Eden区域所占空间过小。这种情况下，可以将Eden区域所占的空间比例增大，以减缓Minor GC的发生。Eden区域到底调整为多大比较合适，可以根据运行时的任务数量进行估算，例如，当从HDFS上读取数据时，如果需要执行4个任务，每个任务占用一个数据块，由于从HDFS上读取到的数据经过解序列化和解压缩后的大小为原来的2~3倍，每一个数据块的大小为64MB，那么执行这四个任务所需要的空间大小约为：4×3×64MB。自适应算法流程图，如图3-6所示。

1. 垃圾回收自适应算法设计

垃圾回收自适应算法对会对Spark系统参数spark.storage.memFraction和JVM中的控制新生代老生代所占空间比例的参数——NewRatio，和控制新生代中Eden区与survivor区所占内存比例的参数——SurvivorRatio进行调整。保证在运行时JVM堆空间能被高效地使用。

|  |
| --- |
| **Algorithm 3**: Adaptive Garbage Collection  **Input**: totalGCVec, fullGCVec, minorGCVec |
| 1. if totalGCVec > ***gcUpperThr*** do 2. reduceCacheRatio() 3. elif totalGCVec < ***gcLowerThr*** do 4. increaseCacheRatio() 5. endif 6. if fullGCVec > ***fullGCUpperThr*** do 7. increaseNewRatio() 8. elif fullGCVec < ***fullGCLowerThr*** do 9. reduceNewRatio() 10. endif 11. if minorGCVec > ***minorGCUpperThr*** do 12. increaseSurvivorRatio() 13. elif minorGCVec < ***minorGCUpperThr*** do 14. reduceSurvivorRatio() 15. endif |

图 3‑7 垃圾回收自适应算法

Figure 3-7 GC Adaptive Alogorithm

* 参数说明：

totalGCVec——总GC操作速度（次/秒）

fullGCVec——full GC操作速度（次/秒）

minorGCVec——minor GC操作速度（次/秒）

gcUpperThr——GC操作上限，超过表示GC操作过于频繁

gcLowerThr——GC操作下限

fullGCUpperThr——full GC操作上限

minorGCUpperThr——minor GC操作下限

* 函数说明：

increaseCacheRatio——增大参数spark.storage.memoryFraction的值

reduceCacheRatio——降低参数spark.storage.memoryFraction的值

increaseNewRatio——增大JVM参数-XX:NewRatio（增大老生代所占比例）

reduceNewRatio——降低-XX:NewRatio（增大新生代所占比例）

inceaseSurvivorRatio——增大新生代中Eden区所占比例

reduceSurvivorRatio——降低新生代中Eden区所占比例

垃圾回收自适应策略如图3-7所示。在之前的设计分析中得出了两个关于调优的结论：CPU过高的消耗会直接导致任务运行速度变慢，而Spark缓存在内存的数据量变少则是增大了数据块的miss率，会间接地导致任务执行速度变慢。因而，垃圾回收自适应策略首先要考虑的就是GC操作的信息，如果GC操作过于频繁，说明JVM堆中用于存放运行时临时变量的空间过少，应将其比例增大，在算法中通过函数reduceCacheRatio降低spark.storage.memoryFraction的值来实现；其次才考虑用于缓存数据的JVM堆空间大小，如果GC操作不是那么的频繁，则算法尝试增大用于缓存数据的堆空间大小，在算法中通过函数increaseCacheRatio增大spark.storage.memoryFraction的值来实现。

在之前的垃圾回收策略中提到提到，JVM堆分为新生代区和老生代区，在进行调优的时候要尽量保证存活时间长的对象被存放到老生代区中。在任务运行时，如果full GC发生得过于频繁表示老生代区过小，可通过increaseNewRatio来增大其比例；反之，通过reduceNewRatio来降低其比例。而新生代区又可以分为Eden、survivor1和survivor2三个区域，在进行GC操作的时候，Eden和survivor1中到达一定生存时间的对象会被转存到survivor2中（如果survivor2中存放不下则转存到老生代区中）。之后交换survivor区的名字，即原来的survivor1变为了survivor2，原来的survivor2变成了survivor1。在这里需要说明的是，如果Eden区过大而survivor区过小（两个survivor区的大小是相等的），那么在进行GC操作时，如果Eden区中幸存的对象由于体积过大不能存放到survivor区中，则会被转存到老生代中。这会导致老生代区很快被填满，进而发生非常耗时的full GC操作，导致系统性能下降，这显然不是开发者想要见到的。因而，如果minor GC频繁发生，可以将Eden区占整个新生代区的比例增大，在算法中通过increaseSurvivorRatio函数来实现；反之，降低Eden区所占比例，通过算法中的reduceSurvivorRatio函数来实现。

与序列化自适应算法和压缩自适应算法类似，垃圾回收自适应算法中阈值参数的初始值，在设置后，会在之后的参数优化模块中进行调整。因而，开发人员只需设置合理的初始值，或是直接使用默认初始值即可。

## 参数优化模块设计

本节主要介绍参数优化模块的设计，该模块与SATS系统性能的优劣有着密切的关系。本文采用混合优化方式对Spark系统中的参数进行优化，将使用模拟退火算法来对参数空间进行寻优，找到近似最优的参数配置。该算法会在3.4.3中进行介绍。

### 参数值的初始设定

在自适应决策模块中，SATS使用三个优化算法来对存储策略进行调整，这些算法中会涉及到一些常量参数，这些参数的取值会很大地影响算法的性能。参数优化模块主要负责的是，利用运行时收集到的数据对自适应决策模块中所用到的一些算法的参数进行优化。在SATS系统第一次启动的时候，因为还没有可利用的历史数据来对调优算法中的参数进行优化，因此这个时候系统所采用的参数是根据Spark应用程序员开发过程中积累下来的经验得到的参数值。显而易见，这些通过经验得到的参数值在面对不同的工作集的时候不一定能够提升系统性能，有可能还会让系统的性能降低。因此，参数的优化工作显得尤为重要，根据不同的工作集可以得到不同的优化参数。SATS系统会收集这些参数的值，当相同或是类似的任务达到的时候，选择对应的优化参数就能够到达优化的效果。

### 优化参数说明

表3-2是Spark中有关内存和CPU的可调控参数，这些参数对Spark性能的影响很大，之后会对这些参数进行说明。

1. 内存参数

首先来说明与shuffle过程相关的参数，主要有两个参数：spark.shuffle.file.buffer和spark.shuffle.memoryFraction。前者用来指用来存放shuffle操作产生的临时文件的内存大小，如果该参数设置得过小，则系统会将存放不下的数据转存到磁盘中，如果设置得过大，则浪费了系统的内存空间。一般来说，把这个参数设置为比产生的临时文件中占用空间最大的文件要大。spark.shuffle.memoryFraction参数用来指定用于缓存shuffle临时文件的内存空间占整个JVM堆空间的比例。其默认值是0.2，即使用20%的内存空间来存放shuffle临时文件，多出来的数据将会被转存到磁盘中。该参数的设置还关联到spark.shuffle.spill参数，如果后者被设置为true，表示当临时文件无法存放在内存中时，会将它们存放到磁盘中；如果设置为false，系统不会使用磁盘来存储临时文件，而会对内存空间进行回收，这往往会造成内存溢出错误。

表 3‑2 Spark配置参数

Table 3-2 Spark configuration parameter

|  |  |
| --- | --- |
| **配置参数** | **默认值** |
| **内存配置参数** |  |
| spark.driver.maxResultSize | 1g |
| spark.reducer.maxSizeInFlight | 48m |
| spark.shuffle.file.buffer | 32k |
| spark.shuffle.memoryFraction | 0.2 |
| spark.io.compression.lz4.blockSize | 32k |
| spark.io.compression.snappy.blockSize | 32k |
| spark.kryoserializer.buffer.max | 64m |
| spark.kryoserializer.buffer | 64k |
| spark.serializer.objectStreamReset | 100 |
| spark.broadcast.blockSize | 4m |
| spark.storage.memoryFraction | 0.6 |
| spark.storage.memoryMapThreshold | 2m |
| spark.storage.unrollFraction | 0.2 |
| **CPU配置参数** |  |
| spark.driver.cores | 1 |
| spark.executor.cores | 1 |
| spark.cores.max | NULL |
| spark.task.cpus | 1 |

因此，为了安全起见，会把spark.shuffle.spill设置为true（其默认值是true）。在设置spark.shuffle.memoryFraction参数时，要考虑到spill操作是否频繁，如果spill操作过于频繁，可以适当地将该参数值调高，否则，可以降低该参数值。但是，对JVM堆空间的分配会影响到用于缓存RDD的内存空间，因此在调整该参数的时候，还需要考虑到spark.storage.memoryFraction参数的设置。

其次，说明与数据序列化和压缩有关的参数。主要参数有：

1）spark.io.compression.lz4.blockSize

2）spark.io.compression.snappy.blockSize

3）spark.kryoserializer.buffer.max

4）spark.kryoserializer.buffer

5）spark.serializer.objectStreamReset

1）和2）指定lz4和snappy算法输出数据块的大小，如果设置得过大则会浪费空间；过小，很可能造成原来只需要一个数据块来存放的压缩数据，需要使用跟多的数据快来存放。当shuffle操作产生的临时文件过多时，可以将以上两个参数调小，用来减少临时文件占用的空间。3）和4）分别用来指定存放经过Kryo序列化后的数据所使用缓存空间的最大值和初始值。在Spark中每个CPU会分配一个spark.kryoserializer.buffer。如果运行中发生“buffer limit exceeded”异常时，可以将3）的值调大。5）是和Java序列化相关的参数。采用Java序列化方式时，系统会将序列化后的对象进行缓存，作为灾备恢复数据。这些被缓存的序列化对象对数量，会在达到参数5）设定的数值时进行回收。该参数如果设置得过大，会占用过多内存空间，影响JVM的垃圾回收，有可能造成内存溢出错误；设置得过小，会导致回收对象的操作过于频繁，使得系统性能下降。

最后，来说明与存储相关的参数。主要有三个：

1. spark.storage.memoryFraction
2. spark.storage.memoryMapThreshold
3. spark.storage.unrollFraction

1）在前中文也多次提到，表示用来缓存RDD的内存空间占整个JVM堆空间的比例。任务运行的过程中，如果GC操作过于频繁，可以把这个参数值调小；否则，可以将其调大。2）表示进行内存映射时，读取磁盘数据块的大小。如果这个参数设置得接近或者小于一个物理页的大小，那么会带来很高的系统开销。3）表示用来物化RDD数据的内存空间占用缓存RDD数据的JVM堆空间的比例。在Spark中，为了节省内存空间，当对一个RDD进行缓存的时候，系统只是将这个RDD的相关信息（大小，位置，元数据）等信息写入内存当中。当这个RDD真正被用到的时候，才会从磁盘中将其数据块抓取到内存中，这个过程称为“物化”。参数3）设置了使用多少缓存空间来物化RDD，如果物化所用空间不够，系统会动态回收内存空间。

1. CPU参数

Spark并没有提供类似于Hadoop用来精密控制内核虚拟化的参数，而只是粗略地提供了如表3-2所示的四个参数。因为CPU的调整优化不是本文主要的研究内容，所以在此不对CPU参数进行参数的调整，使用系统提供的默认值。

### 模拟退火算法的使用

在进行参数配置选择的时候，有经验的开发者可以根据自己的经验来进行恰当的配置。但是，相对于没有大数据框架开发经验的开发人员来说，想要找到一组最优配置，往往是比较困难的。一般的做法，是经过反复的测试找到合适的配置。如果想要获得最优配置，往往只有使用穷举法，即将参数空间中的所有参数配置组合都实验一遍之后，得到最优配置。这种做法的时效性是相当低下的。模拟退火算法是一种随机寻优算法，与贪心算法和爬山算法不同，它是一种启发式的搜索算法，有一定的概率能够找到全局最优解。模拟退火算法为全局择优提供了更好的时效性，如果参数设置得当，其效率是远远超过穷举法的。

本课题使用模拟退火算法作为参数调优中所使用的算法，算法首先要定义一个稳定的温度值和一个初始温度。一般来说初始温度设置的要大一些，以增加找到最优解的概率。当算法开始的时候，温度会由这个初始高温向给定低温变化，这个过程模拟了退火中的降温过程。温度的下降速度可以由一个函数来进行控制，本课题使用快速降温函数：

(公式 3‑1)

来进行温度下降的控制。设t时刻的温度为，在该时刻算法将采用拉丁超立方抽样方法从配置集合{}抽取出一个样本集合config[m]，其中m表示抽取出来的样本中的配置方案个数。从这些配置方案中选择最优的配置，即消耗时间最少的，作为当前最优配置。之后根据一个领域函数产生候选配置集合，该领域函数是以当前最优配置作为中心进行搜索。在得到的这个领域配置集合里面找出最优解作为候选配置，比较和两种配置下的系统性能。根据Metropolis准则进行选择，该准则如下所示：

(公式 3‑2)

E()表示在配置为时的系统性能，当时，将配置赋给。从公式中可以看出，即便候选配置的性能没有最优配置的好，算法还是有一定概率选择候选配置作为最优配置。这一步是模拟退火算法区别于贪心算法之处，通过这一步算法跳出了局部最优解的限制，使得算法有一定概率能够获得全局最优解。这样的处理过程在温度时刻会循环进行多次，次数可以进行指定（比如5次）。循环的次数越多，得到全局最优解的概率越大，但同时，消耗的时间也越多。

当温度下降到时，表示这时系统处于一个稳定的低温状态，这个状态就是算法的结束状态，在该状态下的即为当前最优配置。

### 参数优化模块设计

|  |
| --- |
| **Algorithm** 4: Simulated Annealing  **Input**: {}  **Output**: |
| 1. *local\_search* = 1 2. config[m] = LHS(m) 3. t = 0 4. = best(config[m]) 5. while T > do 6. while *local\_search* < *local\_search\_Thres* do 7. = adjust\_neighbor() 9. if < 0 do 10. = 11. else 12. p = exp() 13. c = random(0,1) 14. if c < p do 15. = 16. endif 17. endif 18. endwhile 20. t ++ 21. *local\_search* = 1 22. endwhile |

图 3‑8 参数优化算法

Figure 3-8 parameter tuning algorithm

* 参数说明：

local\_search——本地搜索次数

local\_search\_Thres——本地搜索次数阈值

config[m]——包含了m个配置组合的集合

——最优配置方案

——候选配置方案

p——Metropolis值

——初始温度

t——计时器

* 函数说明：

LHS——拉丁超立方抽样方法

best——找到最优配置

cost——用来评估一个配置方案的消耗

adjust\_neighbor()——邻域函数，根据现有的配置产生新的配置

random(a,b)——产生a到b之间的随机数，其中a < b

算法3-8是模拟退火算法的伪代码，算法主要有三个优点：1）保证一定概率能够获得全局最优解，2）有一定的抗噪声干扰能力，3）采用了比较高效的LHS抽样算法。下面对算法进行详细说明：

算法分成两层循环，外层循环用来控制降温过程，为了增强系统的时效性，这里使用的是快速降温方式，内存循环用来找最优解。寻找最优解的过程分为两步。第一步采用超立方拉丁抽样方法从参数配置样本集中抽取出m个样本，再从这m个配置方案中选取能让任务运行最快的配置，设为 ，此时得到的这个配置是局部最优解。进入内层循环后，使用邻域函数根据当前的最优配置得到新的候选配置方案，设为配置候选集。计算与 两个配置方案的性能差，若表示是更优的配置方案，将的值赋给。否则，根据Metropolis准则产生参数p（取值范围[0,1]），并利用随机函数生成一个值在[0,1]范围的随机数r，对比r与p的大小。若r < p，则接受作为最优解；否则，放弃配置。当内循环达到限定值的时候（例如，5次），内循环退出。在外循环中，对T进行降温，若T达到了设定的温度下限，外循环退出，整个算法结束。

## 本章小结

本章主要介绍了SATS系统的设计思想和各个模块的思路，通过运行时数据收集模块收集到的数据，利用这些收集到的数据，自适应决策模块根据序列化优化算法、数据压缩优化算法和垃圾回收优化算法，对系统的存储策略进行调整。自适应决策模块的三个算法所用到的参数，可以在Spark程序运行完成之后，通过参数优化模块进行优化，已得到最优的算法参数配置。下一次，当相同或类似的Spark应用程序到达的时候，SATS系统选择这个最优化参数就可以获得更高的运行性能。

# 系统实现

本章将介绍SATS系统的具体实现细节和原理。SATS系统分为三个子模块：运行时数据收集模块，自适应决策模块和参数优化模块。其中数据收集模块和自适应决策模块是在Spark任务运行时开始运行，而参数优化模块是在整个Spark分布式框架运行期间都在运行。自适应优化模块和参数优化模块是SATS系统的主要模块，将会进行重点阐述。

## 系统实现概述

Spark是一个开源的分布式系统架构，可运行在Windows、Linux、Unix等多个系统上。Spark基于Hadoop发展而来，其集群管理方式有：Standalone、YARN和Mesos等三种方式，开发人员可根据需求的不同选择不同的集群管理模式。其中，YARN和Mesos模式都提供了良好的集群资源管理功能。在本实验中，采用了Spark-1.5.1版本，并且为了凸显出本文的工作效果，采用了Spark的Standalone集群管理模式。该模式没有使用自动化的资源管理功能，在一定程度上排除了管理环境对实验结果的干扰。在操作系统方面，Ubuntu14.10是一款稳定的操作系统，能良好地支持Spark的安装与运行，因此，本文选择它作为实验集群的操作系统。在开发语言方面，Spark支持Java、Python和Scala三种语言，Spark的核心代码使用Scala语言进行开发。Scala语言是运行在JVM上的高级编程语言，提供了比Java更加灵活的编程特性，使用Scala语言编写的程序比Java语言更加简洁、高效，例如Spark框架的所有代码只有5万行左右。因为本系统需要对Spark源码进行修改，所以系统的实现也采用Scala语言来完成。

SATS系统主要针对单个节点进行内存优化，因此将集群建立在一个虚拟内网中，从而排除网络因素的影响。对于数据的输入，本实验统一从HDFS中进行读取。HDFS是Hadoop用来进行数据存储管理的文件系统，具有高容错、高吞吐量等优点，非常适合用来存储大规模数据。本文采用的Hadoop版本为Hadoop-2.5.2。

SATS的三个模块——RDCM、ADM和POM，共同协作完成对Spark任务的自适应调优工作。这三个模块的时序图如图4-1所示，在图中可以看出，系统首先进行运行时数据的收集；之后将收集到的数据经过预处理之后输入都自适应决策模块，自适应决策模块会根据系统实时状态调整数据序列化、压缩算法和垃圾回收策略；最后参数优化模块进行参数优化。



图 4‑1 系统时序图

Figure 4-1 system sequence diagram

系统的工作模块图如图4-2所示，SATS系统一共有三个模块，对这三个模块的详细描述如下：

图 4‑2 系统工作模块图

Figure 4-2 System work module

* 运行时数据收集模块：负责在运行时收集和存储数据信息，通过Collector接口来实现。在系统资源方面需要收集系统的内存使用量，CPU使用情况；在数据存储方面，需要记录每一个RDD所使用的数据序列化和压缩算法。需要说明的是，SATS修改了Spark原来的数据序列化方式，具体来说，在原生Spark中，系统使用RDD作为数据序列化的基本单位，这样的方式使得只能使用一种序列化方法来序列化一个RDD。所带来的问题是，如果一个RDD非常大，当内存资源达到设置阈值后，不能切换使用更为高效的序列化算法来节省空间，而是要等到这个RDD序列化完成后，才能修改当前序列化算法，从而错过了切换算法的最佳时机。因此，SATS将原生Spark的序列化单位从RDD变为了partition，从而缩短了算法切换的响应时间，增加了系统的灵活性。与序列化不同的是，压缩算法的压缩单位依然是以系统设定的一个RDD为准，因为相对于序列化来说，压缩算法的速度是相当快的（压缩速度最慢的的LZ4的压缩速度为150MB/s）。收集到的数据会使用循环存储结构来进行存储，目的是为了节省内存空间。
* 自适应决策模块：这个模块主要负责，在系统的内存、CPU资源使用量以及GC操作的次数达到设置的阈值时，切换序列化算法、压缩算法，调整JVM堆中各个区域所占比例。所谓的自适应，就是系统能够根据自身资源的使用情况来调整所用的策略。这里使用了三个自适应策略：序列化自适应、压缩自适应和垃圾回收自适应。对于序列化和压缩算法的自适应策略，系统会根据当前内存和CPU使用情况来进行算法切换。对于垃圾回收自适应策略，主要包含两个方面：首先是调整用于缓存RDD的内存与用于运行时存储临时变对象的内存所占的比例，其次是调整老生代与新生代、survivor区与Eden区的比例。具体的切换策略在3.3中已经进行了详细地描述，在此不再赘述。
* 参数优化模块：主要负责优化系统的参数，这里的系统参数分为三个部分：Spark系统参数、JVM参数以及自适应决策模块中所使用的阈值。该模块使用模拟退火算法来实现，算法的具体说明见3.4.4节。在任务运行时，模块会对Spark和JVM参数进行优化调整，选出性能最优的配置组合；在Spark任务运行结束后，模块会对自适应决策模块中所使用的阈值参数进行优化调整。参数优化模块也使用到运行时数据收集模块收集到的数据。

## 运行时数据收集模块实现

RDCM模块的工作流程图，如图4-3所示，在Spark任务开始后，模块会收集所需的数据（参见3.2.1节），并将收集到的数据进行存储，这个过程会在任务运行期间一直进行，直到当前运行的Spark应用程序的所有任务结束后才终止。



图 4‑3 运行时数据收集模块工作流程图

Figure 4-3 The runtime data collection module workflow chart

RDCM模块的类图，如图4-4所示，由两个类构成。其中RuntimeData用来存放收集到的数据，而DataCollector类是一个继承了Runnable接口的类，用来在运行时收集数据。



图 4‑4 运行时数据收集模块类图

Figure 4-4 Runtime data collection module class diagram

* RuntimeData类：用来收集系统数据，所要收集的数据如图4-4所示，下面进行详细说明。blockId2Serc是一个Map类型的数据结构，用来存放blockId到用来序列化这个block的序列化类的hashCode的映射关系。注意这里存放的key-value键值对中value是所使用的序列化类的hashCode，而不是这个序列化类的对象。这样做是为了节省内存空间，并且设定blockId2Serc的大小为当前运行的DAG图中各个Stage中partition最多的RDD的分区数量之和，即。serc2Ser用来存储序列化类的hashCode到该序列化类名称的映射。当要查找一个block所使用的序列化方法时，只需在这两个表中进行查询即可。RDDId2Codec用来存放RDD到所使用压缩算法类的hashCode的映射，codec2Compresser存放该hashCode到压缩算法名称的映射，两者的作用和blockId2Serc与serc2Ser的作用相似。其他参数说明见3.2节。
* DataCollector类：该类实现了Runnable接口，是一个线程类，在任务运行时启动执行该类的一个线程。当Spark任务开始运行时，安插在Executor类中的监听代码（stub），会每隔一个固定的时间间隔（默认是1s，可进行设置），收集系统数据信息。当所有的Spark任务运行结束后，即一个Spark应用程序运行完成后，执行DataCollector类对象的线程结束运行，系统回收该线程。当新的Spark应用到达时，重复以上过程。

## 自适应决策模块的实现

自适应决策模块的类图，如图4-5所示。该模块由Data\_Receiver类、Adaptive\_Decision类和IAdaptive\_Decision接口组成。

* DataReceiver类：负责从运行时数据收集模块收集自适应算法所需的数据。
* AdaptivSerialization类：AdaptivSerialization类实现IAdaptive接口，主要对系统进行序列化算法自适应调整的任务。该类中的参数说明，参见3.3.1节，在此不再赘述。序列化自适应算法的流程图如图4-6所示，算法首先判断CPU的使用是否达到了设定的阈值，若没有超过，则会判断内存的使用量是否超过内存使用上限或是低于内存使用下限。在此，主要说明阈值初始值的设定方案。初始的时候，会默认将CPU的使用阈值设置为0.9，即当CPU消耗达到90%，且这种状态持续一段时间的之后，系统会切换当前的序列化算法。初始的内存上限设置为0.9，下限为0.6，含义同CPU阈值设定。由于之后的参数优化模块会对这些设置的阈值进行调整，因此，不必担心因为初始



图 4‑5 自适应决策模块类图

Figure 4-5 Adaptive decision module class diagram



图 4‑6 序列化自适应算法工作流程图

Figure 4-6 Serialization adaptive algorithm workflow chart

阈值的设置不恰当而造成系统性能下降的情况，开发人员只需设定一个自己认为合理或者直接使用默认的阈值即可。



图4‑7 压缩自适应算法工作流程图

Figure 4-7 Compression adaptive algorithm workflow chart

* AdaptiveCompression类：该类也实现了IAdaptive接口，主要是对系统进行数据压缩算法自适应调整的任务。类中的参数说明，参见3.3.2节。算法的流程图如图4-7所示。与序列化自适应算法类似，参数优化模块会对数据压缩自适应算法中使用到的参数进行调整，因而，开发人员只需自定义或是使用默认的阈值参数即可。
* AdaptiveGC类：这个类主要实现了对垃圾回收自适应算法的调整工作，同样的，也实现了IAdaptive接口。垃圾回收自适应算法的工作流程图如图4-8所示。当系统的total GC、Full GC和Minor GC达到指定阈值的时候，会通过定义的6参数调整函数进行参数调整。举例说明，当total GC频繁发生的时候，这时候说明JVM堆中用于存放运行时产生的临时变量的堆空间过少，可以通过reduceCacheRatio函数来增加该类空间所占比例；反之，可以通过increaseCacheRatio函数来减少该空间的比例。在这里需要解决的问题，如何设置减少或增加比例的幅度。同样地，这个问题交给了参数优化模块来处理，该模块会找出合适的幅度值，具体过程将会在4.3节中进行讨论。



图 4‑8 垃圾回收自适应算法工作流程图

Figure 4-8 GC adaptive algorithm workflow chart

## 参数优化模块的实现



图 4‑9 参数优化模块类图

Figure 4-9 Parameter tuning module class diagram

参数优化模块的类图如图4-9所示，参数优化模块由三个部分构成，分别是DataReceiver类、ParameterTuner类和TuningAlgorithm接口。

* DataSender类：向自适应决策模块发送优化后的配置参数。
* ParameterTuner类：实现了TuningAlgorithm接口，用来从参数空间中选择最优的参数配置组合。参数优化模块进行调优的参数可以分为三类，第一类是JVM相关的参数，第二类是与Spark相关的系统参数，第三类是自适应决策模块中的参数。调优的过程就是，要从这三类参数空间中找到最优参数配置。比较常见的寻优方法有三种，第一种采用穷举法，即将参数空间中的所有参数组合都测试一遍，找到使任务性能最好的配置作为最优配置。这种方法虽然能够得到最优配置，但是却会消耗大量额时间，一般不会使用。第二种是历史信息择优法，即从收集到的历史数据中选择最优性能的配置。这种方法的缺点是，所获取的最优配置是在特定时刻的最优配置，当系统环境发生变化的时候，该配置不一定就是最优配置，甚至有可能造成系统性能下降。第三种方法是使用机器学习算法，在任务运行过程中动态寻找最优配置。如果参数设置得当的话，这种方法的找到最优配置的收敛速度是很快的，在对时间的消耗上，是优于穷举法的。本文采用第三种方法进行参数优化，并选择模拟退火算法作为调优算法。
* ITuningAlgorithm接口：通过实现这个接口，开发人员可以根据自己的需求选择不同的参数优化算法。

## 本章小结

本章主要介绍SATS系统的实现和原理，首先在4.1节中对SATS系统的整体架构和其在Spark中的作用进行了描述。接下来的三节分别详细描述了运行时数据收集模块，自适应决策模块和参数优化模块。自适应决策模块是系统的主要模块，模块实现了对序列化算法、压缩算法和垃圾回收策略的自适应调整策略。4.4节介绍了参数优化模块，该模块主要针对系统中的配置参数进行调优。该采用了模拟退火算法，分为动态调优和静态调优两种调优类型。

# 系统测试与结果分析

## 实验方法和目的

本实验通过部署安装原生的Spark和实现了SATS子系统的Spark系统，对两者的性能进行对比。通过实验结果来验证基于Spark内存计算特性的自适应调优策略在Spark系统优化中的作用。采用控制变量法，分别进行单元测试、集成测试和系统测试。

基于内存计算特性的优化策略，其重点在于如何高效地利用有限的内存资源。本实验的目在于摒弃了传统的手动调优的方式，通过收集分析任务运行时的特征，让系统自动根据系统资源状态的变化来调整相关策略（通过修改配置参数来实现），保证通过自适应调优后的系统性能要高于简单的手工配置，并能够接近通过具有相当开发经验的开发人员手工调优后的系统性能。

## 实验环境搭建

### 实验环境

|  |
| --- |
| **参数类型 参数值** |
| CPU信息 Intel(R) i7-4770S 3.10GHz  CPU核数 8核  内存 16GB  硬盘 1TB  操作系统 ubuntu14.10-desktop |

图 5‑1 实验环境配置

Figure 5-1 Configuration of experiment environment

### 集群配置

本实验采用虚拟机集群来进行搭建，在主机上使用VMware Workstation 10建立四个虚拟机。虚拟机的配置如图5-2所示。

|  |
| --- |
| **节点名称 参数类型 参数值** |
| driver CPU信息 Intel(R) i7-4770S 3.10GHz  CPU核数 1核  内存 1GB  硬盘 128GB  操作系统 ubuntu14.10-server-amd64  slave CPU信息 Intel(R) i7-4770S 3.10GHz  CPU核数 2核  内存 2GB  硬盘 128GB  操作系统 ubuntu14.10-server-amd64 |

图 5‑2 实验集群配置

Figure 5-2 Configuration of experiment cluster

注：slave包含三个节点slave1、slave2和slave3，配置均相同

实验将原生的Spark和实现了SATS系统的Spark都安装在driver节点中。之后，通过NFS网络系统将以上两个系统的安装目录挂载到三个slave节点中。这样搭建的好处是，当要对系统进行修改的时候，只需要修改driver上的系统即可，省去了将修改后的系统再次在各个slave中安装、编译的过程，大大节省了开发时间。driver节点的spark.driver.memory配置为1GB，三个slave节点的spark.executor.memory配置为1GB。

本实验将HDFS作为外部存储系统进行数据的存放。HDFS是Hadoop中的一个子系统，因此，在此需要安装Hadoop。本实验选择的Hadoop版本为Hadoop-2.5.2。在HDFS系统中个，将driver节点作为NameNode，slave1 、slave2 和slave3作为DataNode。本实验采用的Spark版本为1.5.1，git地址：<https://github.com/apache/spark>。

## 自适应决策模块测试

本节将对自适应决策模块进行测试，分为三个单元测试：序列化、压缩和垃圾回收自适应算法的测试与分析；之后，会对整个模块进行集成测试。

### 序列化自适应算法测试与分析

|  |
| --- |
| **Algorithm: SimpleCount** |
| 1. ***Require****: FilePath* 2. ***Ensure****: top 10 most popular IP* 3. val output = sc.textFile(FilePath) 4. .map(line => line.split(“\t”)).map(line => (line(2),1)) 5. .cache() 6. .reduceByKey( + ,2).filter(x => x. 2 > 100) 7. .collect() 8. .sortWith( . 2 > . 2) 9. .take(10) 10. .map(x => (x. 1,x. 2)) |

图 5‑3 simplecount算法

Figure 5-3 simplecount algorithm

SimpleCount用来找出访问最频繁的10个IP地址，数据格式为：服务器IP客户端IP 访问日期。实验中原生的Spark使用了默认的序列化方法——Java序列化方法。算法的伪代码实现，如图5-3所示。

SimpleCount算法的实验结果如图5-4所示，从图中可以看出随着实验数据集的增大，采用了序列化自适应算法的系统性能与原生Spark性能差距越来越明显。当实验数据超过2GB的时候，使用了SATS的系统性能达到了原生Spark的2倍！

PageRank是常用的推介算法之一，是典型的图计算算法。算法通过图中的点边关系计算出最受欢迎的网站地址。在计算过程中，图中的每一个点根据相邻点的pagerank值来计算自身的pagerank。当所有的点都完成了自身数据的更新后，算法就完成了一次迭代计算过程。在经过多次迭代后，每个点的pagerank值变化会越来越小，当这个变化值达到指定的收敛值时，算法运行结束。从以上分析可以看，PageRank算法是典型的迭代式机器学习算法，即每一次计算得到的值，将会被下一次计算用到。

图 5‑4 序列化自适应Spark与原生Spark的simplecount性能对比图

Figure 5-4 simplecount performance comparison between Spark with serialization adaptive strategy and original Spark

图 5‑5 序列化自适应Spark与原生Spark的pagerank性能对比图

Figure 5-5 pagerank performance comparison between Spark with serialization adaptive strategy and original Spark

图5-5展示了实验结果，同实验一，原生的Spark采用的序列化方法为默认的Java序列化方法。从结果中可以看出实现SATS的系统性能相比于原生的Spark并没有多大的提升。这是因为PageRank算法在Spark中会产生大量shuffle操作，这些shuffle操作会产生大量的体积很小临时文件。对这么多的临时文件采用不同序列化方法进行序列化和解序列化，会消耗掉较多的系统资源，导致任务运行速度变慢。同时，JVM堆中用于存放shuffle数据的空间有限（默认仅使用RDD缓存空间的20%），多余的数据要被存放到磁盘中，增加了系统从磁盘读取数据的开销。

### 压缩自适应算法测试与分析

图 5‑6 压缩自适应Spark与原生Spark的simplecount性能对比图

Figure 5-6 simplecount performance comparison between Spark with compression adaptive strategy and original Spark

从图5-6中可以看出，在单纯消耗内存使用的情况下，自适应的压缩策略能够取得更高的运行效率。当数据集的大小超过内存大小的时候（数据集为2GB时），采用了压缩自适应算法的Spark的执行速度达到了原生的Spark的2倍左右。在运行过程中，压缩自适应算法会根据内存的使用情况，自动调整所使用的压缩算法，这相比于默认情况下使用snappy压缩算法的原生Spark，要灵活的多。

图5-7展示了pagerank的运行结果，从图中可以看出，与之前实现了序列化自适应算法的Spark在pagerank算法上的表现类似。实现了压缩自适应算法的Spark在pagerank上的性能并没有执行的更快，其原因与序列化自适应算法相似，都是因为pagerank在有过多的shuffle操作，产生了大量的体积较小的临时文件。对这些临时文件的压缩处理，在一定程度上是节省了内存空间，但是多次的压缩与解压缩操作也占用了大量的系统时间。因而，压缩自适应算法在pagerank算法上的应用并不如在SimpleCount上取得的效果好。

图 5‑7 压缩自适应Spark与原生Spark的pagerank性能对比图

Figure 5-7 pagerank performance comparison between Spark with compression adaptive strategy and original Spark

从这两个实验中可以看出，实现了压缩自适应算法的Spark比原生的Spark执行速度要快，在SimpleCount实验中获得了超过原生Spark一倍多的运行效率。但是在pagerank实验中，其性能和原生的Spark相差较小，这主要是shuffle操作的问题。针对shuffle操作的优化已经超出了本文的研究范畴，在此不做详细讨论。

### 垃圾回收自适应算法测试与分析

对于垃圾回收的调优操作，主要是建立在对JVM堆的垃圾回收优化上。图5-8显示了实现垃圾回收优化策略后的系统在SimpleCount上与原生系统的对比。从图中可以看出，垃圾回收自适应算法表现得并不是很理想，甚至在内存充足的时候（数据集大小为1GB时，没有超过内存的使用量1GB），其效率是低于原生Spark的。分析原因，是在进行JVM堆中各个区域所占空间进行动态调整的时候，会发生数据的迁移，虽然是内存—内存的数据迁移，速度很快，但是当数据量较大的时候也会带来不小的开销。

图 5‑8 垃圾回收自适应Spark与原生Spark的simplecount性能对比图

Figure 5-8 simplecount performance comparison between Spark with GC adaptive strategy and original Spark

图 5‑9 垃圾回收自适应Spark与原生Spark的pagerank性能对比图

Figure 5-9 pagerank performance comparison between Spark with GC adaptive strategy and original Spark

图5-9是pagerank用例的运行结果，结果与数据序列化和压缩自适应算法的结果相差不大。原因也是由于过得shuffle操作导致产生了过多的临时文件的问题，在此不再赘述。

### 模块集成测试

图 5‑10 自适应决策Spark与原生Spark的simplecount性能对比图

Figure 5-10 simplecount performance comparison between Spark with adaptive decision strategy and original Spark

图 5‑11 自适应决策Spark与原生Spark的pagerank性能对比图

Figure 5-11 pagerank performance comparison between Spark with adaptive decision strategy and original Spark

## 参数优化模块测试

本节针对参数优化模块进行单元测试，实验在屏蔽了自适应决策模块的情况下进行，因而只针对Spark参数和JVM参数进行调整。

图5-12展示了SimpleCount测试用例的实验结果，纵轴表示实现了自适应决策模块的Spark比原生的Spark的性能提高了多少，横轴表示PageRank的数据集大小（PageRank的迭代次数为10次）。数据点的值低于0表示，性能比原生Spark下降了。从结果中可以看出，经过参数优化模块进行调优之后，虽然获得的性能不是最优项，但是却高于了平均水平，至少不是性能表现最差的。图5-13展示了PageRank测试用例的结果，从图中可以看出经过40次迭代后的性能并不是最高的，但也不是最低的，而且PageRank的性能也有所提升。通过实验得出的结论是，参数优化模块起到了一定的优化作用，只是由于算法是在任务运行时进行的，占用了一定的系统资源，同时得到的参数配置有一定的之后性。因而，算法并不能总是获得最优配置，但性能至少不是最差的。

图 5‑12 参数优化Spark与原生Spark的simplecount性能对比图

Figure 5-12 simplecount performance comparison between Spark with parameter optimization strategy and original Spark

图 5‑13 参数优化Spark与原生Spark的pagerank性能对比图

Figure 5-13 pagerank performance comparison between Spark with parameter optimization strategy and original Spark

## 系统性能整体测试

图5-14和图5-15展示系统整体测试的结果，从结果可以看，当数据集的大小超出系统可用内存时，实现了SATS系统的WordCount的性能是原生Spark的2倍以上；但在PageRank测试用例上，SATS-Spark的性能在内存不足时稍微比Spark要高一些，根本原因在之前的实验中已经分析过，在此不再赘述。

图 5‑14 SATS-Spark与原生Spark的simplecount性能对比图

Figure 5-14 simplecount performance comparison between SATS-Spark and original Spark

图 5‑15 SATS-Spark与原生Spark的pagerank性能对比图

Figure 5-15 pagerank performance comparison between SATS-Spark and original Spark

## 本章小结

本章通过实验验证了实现了SATS的Spark不仅解决了自动配置的问题，而且还在一定程度上提升了系统的性能。但是，对于shuffle过程的调优是基于Hadoop发展而来的分布式平台的一大难题。本文的重点在与如何自适应地根据系统历史信息和运行时系统状态，进行系统的调优工作，使得Spark系统能够获得比在naïve的配置方案下，更加良好的性能。从实验结果可以看出，本文提出的SATS-Spark基本实现了这个目的。

# 总结和展望

## 全文总结

本课题通过对Spark采用自适应的优化策略，改变了传统的通过手工进行系统优化的方式，增强了Spark的易用性。同时，通过自适应的优化方案，在一定程度上提升了系统的性能。课题创造性地提出了动态时调整系统的存储策略，即数据序列化、压缩和垃圾回收策略。并动态对系统运行时状态进行分析，根据分析结果转换系统所使用的策略。

本研究课题的贡献主要有：

1. 分析了手动调优方式的不足

本课题分析了手动进行Spark调优的不足之处，即不能针对系统资源实时的使用情况，对系统所使用的策略进行调整。当系统的内存资源消耗过多时，如果采用的序列化方案是Java序列化算法，那么就会造成内存的急剧消耗，从而导致系统性能的下降。如果在这时将系统的序列化方法转变为高效且节省空间的Kryo算法，那么系统的性能将会得到很大的提升。传统的手工调优方式，不能根据系统信息实施调整存储策略，这是它的不足之处。而且，依赖手动调优对没有调优经验的开发人员来说，比较困难，也增加了使用Spark的困难。

1. 设计了Spark自适应优化系统

本课题设计了SATS系统，将该系统作为Spark的一个子系统来运行。SATS系统由运行时数据收集模块，自适应决策模块和参数优化模块组成。首先，由数据收集模块动态的收集系统信息，然后将这些收集到的数据信息发送给自适应决策模块。自适应决策模块通过对数据的分析，对系统的存储策略进行调整。同时，参数优化模块会对系统以及自适应决策模块中的阈值参数进行寻优操作，找到最优的参数配置组合。通过上述过程，来对Spark进行调优工作。

1. 实现和验证了SATS系统

本研究使用Scala来完成对系统的开发工作，实现了SATS系统。通过搭建Spark实验集群来对SATS系统进行了测试验证工作。实验结果显示SATS系统能够带来一定的性能提升，在省去手动调优的同时，实现了整个Spark系统性能的优化，证明了本课题的研究价值和意义。

## 未来展望

Spark分布式框架发展到现在还是不很完善，还有很多方面有待提高。例如，目前的Spark只整合了两种数据序列化算法：Java和Kryo，其中Java序列化算法各方面的性能在众多的序列化算法中并不是很突出，甚至可以说很差。但因为其是Java语言本身自带的序列化算法，而被沿用到了Spark中。这样造成在序列化自适应算法中能够进行选择的序列化算法很少，没有完全开发出算法的优势。此外Kryo算法使得使用还需要经过注册，使用Kryo对未经注册的类的对象进行序列化，效果并不是太好。现在Spark并不支持Kryo对用户自定义类的注册功能。未来Spark会不断的发展完善，相信到时候系统可以支持多种功能更加丰富的序列化算法，未来的工作将会结合系统集成的这些序列化算法进行进一步的调优工作。

对于基于MapReduce编程模型发展而来的分布式计算框架，普遍存在的问题就是shuffle操作带来的高消耗。shuffle操作不仅消耗空间——会产生很多的临时文件，虽然这些临时文件很小，但是由于数量过多，用于存放文件基本信息的空间也必然增多；而且还很消耗时间——任务运行时，要打开大量的文件，必然会消耗大量的时间。针对shuffle操作进行系统优化，一直是这类系统所面临的一个难题。本文虽然稍微对涉及到shuffle操作的参数进行了调整，并未做深入的研究和分析，在下一步的工作中，将会重点研究shuffle操作的自适应优化问题。此外，还会对系统进行一些细节的改进和完善。

参考文献

1. Sanjay G, Howard G, Shun-Tak L. The Google file system[J]//ACM SIGOPS Operating Systems Review - SOSP '03, 2003, 37(5): 29-43.
2. Fay C, Jeffrey D, Sanjay G, et al. Bigtable: A Distributed Storage System for Structured Data[J]. ACM Transactions on Computer Systems (TOCS), 2008, 26(2): Article No. 4
3. Jeffrey D, Sanjay G. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM - 50th anniversary issue: 1958 – 2008, 2008, 51(1):107-113
4. Shvachko K, Hairong K, Radia S, et al. The Hadoop Distributed File System[C]// Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, 2010, 3(7):1-10.
5. Katarina G, Michael H, Wilson A H, et al. Challenges for MapReduce in Big Data[J]. Proceeding of the IEEE 10th 2014 World Congress on Services (SERVICES 2014), 2014.
6. Lee J, Kwon Y S, Farber F, et al. SAP HANA distributed in-memory database system: Transaction, session, and metadata management[C]//Data Engineering (ICDE), 2013 IEEE 29th International Conference on, IEEE, 2013: 1165-1173.
7. Moinuddin K Q, Vijayalakshmi S, Jude A R. Scalable high performance main memory system using phase-change memory technology[C]//ISCA '09 Proceedings of the 36th annual international symposium on Computer architecture, 2009, 37(3):24-33.
8. Dhiman G, Ayoub R, Rosing T. PDRAM: A hybrid PRAM and DRAM main memory system[C]//Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE, 2009, 26(31):664-669.
9. Joos H B, Cafer T, Christian M. Data management with SAPs in-memory computing engine[C]//EDBT '12 Proceedings of the 15th International Conference on Extending Database Technology, 2012, 542-544.
10. Matei Z, Mosharaf C, Tathagata D, et al. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing[C]//NSDI'12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012, 2-2.
11. Htway H H, Thein T A,Win A. A simple and effective Software Distributed Shared Memory System[C]//Electrical Engineering/Electronics Computer Telecommunications and Information Technology, 2008, 14(17):53-56.
12. Bu Y, Borkar V, Carey M J, et al. Scaling datalog for machine learning on big data[J]. arXiv preprint arXiv,2012:1203.0160.
13. Gonzalez J E, Xin R S, Dave A, et al. GraphX: Graph processing in a distributed dataflow framework[C]//Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2014.
14. Engle C, Lupher A, Xin R, et al. Shark: fast data analysis using coarse-grained distributed memory[C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012: 689-692.
15. Zaharia M, Chowdhury M, Das T, et al. Fast and interactive analytics over Hadoop data with Spark[C]//USENIX, 2012.
16. Zhang H, Tudor B M, Chen G, et al. Efficient In-memory Data Management: An Analysis[J]. Proceedings of the VLDB Endowment, 2014, 7(10).
17. Jiang T, Zhang Q, Hou R, et al. Understanding the behavior of in-memory computing workloads[C]//Workload Characterization (IISWC), 2014 IEEE International Symposium on. IEEE,2014.
18. Armbrust M, Xin R S, Lian C, et al. Spark SQL: Relational data processing in Spark[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015: 1383-1394.
19. Fuller N C, Li M, Meng S, et al. Dynamic Tuning of Memory in MapReduce Systems: U.S. Patent 20,150,309,731[P]. 2015-10-29.
20. Li H, Ghodsi A, Zaharia M, et al. Tachyon: Reliable, memory speed storage for cluster computing frameworks[C]//Proceedings of the ACM Symposium on Cloud Computing. ACM, 2014: 1-15.
21. Kaur M, Verma N. Dynamic Memory Allocation: Role in Memory Management[J]. International Journal of Current Engineering and Technology, 2014,vol.4:531-535.
22. kryo - fast, efficient java serialization and cloning*.http://code.google.com/p/kryo/. Accessed: 6 Jun 2012.*
23. Min Li, Liangzhao Zeng, Shicong Meng,et al. MRONLINE: MapReduce online performance tuning[C]//HPDC '14 Proceedings of the 23rd international symposium on High-performance parallel and distributed computing, 2014: 165-176.
24. Joshi S B. Apache hadoop performance-tuning methodologies and best practices[C]//Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering, ACM, 2012: 241-242.
25. Chen Q, Guo M, Deng Q, et al. HAT: history-based auto-tuning MapReduce in heterogeneous environments[J]. The Journal of Supercomputing, 2013, 64(3): 1038-1054.
26. Meng X, Bradley J, Yavuz B, et al. MLlib: Machine Learning in Apache Spark[J]. arXiv preprint arXiv:1505.06807, 2015.
27. Chiba T, Onodera T. Workload Characterization and Optimization of TPC-H Queries on Apache Spark[J]. 2015.
28. Liao G, Datta K, Willke T L. Springer Berlin Heidelberg Gunther: Search-based auto-tuning of mapreduce[M]//Euro-Par 2013 Parallel Processing, 2013: 406-419.
29. Chen C O, Zhuo Y Q, Yeh C C, et al. Machine Learning-Based Configuration Parameter Tuning on Hadoop System[C]//Big Data (BigData Congress), 2015 IEEE International Congress on. IEEE, 2015: 386-392.
30. Premchaiswadi W, Romsaiyud W. Optimizing and Tuning MapReduce Jobs to Improve the Large‐Scale Data Analysis Process[J]. International Journal of Intelligent Systems, 2013, 28(2): 185-200.
31. Kim J, Ashwin Kumar T K, George K M, et al. Performance evaluation and tuning for MapReduce computing in Hadoop distributed file system[C]//Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on, IEEE 2015: 62-68.
32. Saha B, Shah H, Seth S, et al. Apache tez: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data[C]//A unifying framework for modeling and building data processing applications, ACM, 2015: 1357-1369.
33. Wang L, Tao J, Ranjan R, et al. G-Hadoop: MapReduce across distributed data centers for data-intensive computing[J]. Future Generation Computer Systems, 2013, 29(3): 739-750.
34. Aji A, Wang F, Vo H, et al. Hadoop GIS: a high performance spatial data warehousing system over mapreduce[J]. Proceedings of the VLDB Endowment, 2013, 6(11): 1009-1020.
35. He H, Du Z, Zhang W, et al. Optimization strategy of Hadoop small file storage for big data in healthcare[J]. The Journal of Supercomputing, 2015: 1-12.
36. <https://github.com/google/snappy>

1. <https://github.com/ning/compress>
2. <https://github.com/Cyan4973/lz4>
3. <https://github.com/JerryLead/SparkInternals/blob/master/markdown/PNGfigures/deploy.png>
4. https://github.com/JerryLead/SparkInternals/blob/master/markdown/PNGfigures/ComplexJobStage.png

致谢

光阴似箭，岁月如梭，眨眼之间短短两年半的研究生生涯即将结束，回想研究生阶段的生活和学习，感慨良多，收获良多。求学路上的坎坷，以及克服困难、迎难而上的经历，让我受益匪浅。从中学到的，不只是知识，更有作为一个学子、一名研究者所需的坚持不懈与严谨思维，并认识到了在科学的道路上，没有捷径和坦途可寻，唯有奋发图强、勇攀险峰才会有所收获。在这两年半的时间里，我的成长，离不开我身边支持我的每一个人，有太多的人需要感谢。

首先要感谢的是我的导师胡飞教授。胡老师是一位一丝不苟的科学研究者，他严谨的作风深深地影响了我，使得我在学习和研究中更加努力，并时刻鞭策自己，遇到问题不放弃、不逃避，有了一种刨根问底，不明白不罢休的研究态度，这将让我终身受益。此外，胡老师时常关怀我们的生活与学习。

其次，我要感谢的是直接对我进行指导的陈昊鹏老师。陈老师是大数据研究领域的专家，对大数据领域的问题，有着独到的见解和眼光。每次当我的研究工作陷入困难之际，陈老师都会对我进行悉心的指导，提出宝贵的意见和建议，帮我重塑信心，一步步完成了学业。在平时的生活和学习中，陈老师也给与了我犹如家中长辈的关怀，并对我毕业后的发展给予帮助，再一次感谢陈老师。

另外，我要感谢实验室上一届的程书欣，王浩，周海航学长，特别是王浩学长。因为研究的领域相近，王浩学长经常和我探讨技术上的问题，并针对我的实验给出了具体的技术方案，他对一些关键问题的见解让我有一种茅塞顿开的顿悟。在此，非常感谢几位学长对我的帮助。

同时，我还要感谢同一届的张世晨，俞嘉栋，范嘉骅，朱维，杭存，何俊杰等实验室同学，在我们共同生活的两年半里，大家互相帮助，让我在艰苦的科研工作中，感受到了快乐。对于在交大能与你们的相遇，我深感幸运，并且非常怀念这段时光。

之后，我还要感谢实验室下一届的王臻，刘德霖，舒映等学弟学妹，在写论文的过程中，我或多或少地得到了你们的帮助或启发。

最后，我要感谢我的父母，感谢你们多年来对我学业的支持，让我能够心无旁骛地专心学习，感谢你们对我多年在外求学的理解，感谢你们在我受到挫折时给予我站起来再次迎接挑战的勇气和力量。

攻读学位期间发表的学术论文目录

1. 第一作者. An Adaptive Tuning Strategy on Spark Based on In-memory Computation Characteristics[C]//The 18th IEEE International Conference on Advanced Communications Technology, ICACT2016.[已录用]