基于 Kalman 算法的大数据存储架构可扩展性优化算法

韩镇阳，张　磊，任　冬
（武警陕西省总队，陕西　西安　710116）

摘　要：为了优化大数据存储架构可扩展性能，提高大数据架构资源利用率，通过引入 Kalman 算法设计了一种大数据存储架构可扩展性优化算法。首先，综合考虑大数据存储架构与多核环境内存布局之间的兼容性，设计架构内存布局。其次，设计分布式共享内存协议，确保各个进程在访问共享内存时能够正确地协同工作，提高存储架构的容错性。在此基础上，利用 Kalman 算法，动态调整存储节点的负载，进而优化大数据存储架构，以提高其可扩展性。实验结果表明，应用该算法后，大数据存储架构的资源利用率始终高于对照组，均达到了 96% 以上，最高达到了 98%，架构可扩展性优化效果显著，服务器资源利用更充分，大规模数据处理更高效。

关键词：Kalman 算法；大数据存储架构；可扩展性优化；共享内存协议；节点负载
中图分类号：TP311　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2023.11.005
引用格式：韩镇阳，张磊，任冬. 基于 Kalman 算法的大数据存储架构可扩展性优化算法［J］. 网络安全与数据治理，2023，42(11)：25-28.

A scalability optimization algorithm for big data storage architecture based on Kalman algorithm

Han Zhenyang, Zhang Lei, Ren Dong
(Shanxi Provincial Corps of the Chinese People's Armed Police Force, Xi'an 710116, China)

Abstract: In order to optimize the scalability performance of big data storage architecture and improve the resource utilization of big data architecture, a Kalman algorithm was introduced to design a scalability optimization algorithm for big data storage architecture. Firstly, considering the compatibility between big data storage architecture and multi core environment memory layout, design the architecture memory layout. Secondly, design a distributed shared memory protocol to ensure that various processes can work together correctly when accessing shared memory, and improve the fault tolerance of the storage architecture. On this basis, the Kalman algorithm is used to dynamically adjust the load of storage nodes and optimize the big data storage architecture to improve its scalability. The experimental results show that the resource utilization rate of the big data storage architecture is consistently higher than that of the control group, reaching over 96%, with a maximum of 98%. The scalability optimization effect of the architecture is significant, and the utilization of server resources is more sufficient, enabling more efficient processing of large-scale data.

Key words: Kalman algorithm; big data storage architecture; scalability optimization; shared memory protocol; node load

0　引言

大数据存储架构是指在存储、处理和分析大规模数据时所采用的技术架构。从广义角度分析，大数据存储架构是用于提取和处理海量数据并针对业务目的进行分析整理的整体系统，可视作基于机构业务需求的大数据解决方案的蓝图[1]。大数据存储架构通常包括以下几个主要组成部分：数据存储层、数据处理层、数据分析层和数据可视化层。随着大数据时代的来临，信息资源数据的体量越来越庞大，大数据存储架构面临着巨大的挑战[2]。传统的大数据存储架构通常采用中央式存储方式，这种方式在处理大规模数据时存在着很多局限性，例如可扩展性差、容错能力低等问题[3]。为了应对挑战，研究者们提出了大数据存储架构可扩展性优化算法，对大数据存储架构进行优化，以提高其性能和可扩展性。当前，传统的大数据存储架构可扩展性优化算法在实际应

用中以批处理为主，缺乏实时的支撑。面对需要快速响应和处理的应用场景，如实时分析、实时推荐等，仍然存在缺陷，且对业务支撑的灵活度效果不佳[4]。

Kalman 算法是一种优秀的估计算法，它具有很好的自适应性和鲁棒性，能够对复杂系统进行准确的估计和预测[5]。在大数据存储架构中，Kalman 算法可以用于数据的优化和预测，采用分布式存储方式，通过将数据分散到多个节点上进行存储和处理，提高数据的可扩展性和容错能力，提高数据存储和处理的效率。基于此，本文引入 Kalman 算法来开展大数据存储架构可扩展性优化算法研究。

1　大数据存储架构可扩展性优化算法研究

1.1　大数据存储架构内存布局设计

内存布局对后续架构可扩展性优化起到了至关重要的作用。首先，大数据存储架构内存布局设计中，需要综合考虑架构与多核环境内存布局之间的兼容性。本文设计的大数据存储架构内存布局示意图如图 1 所示。

图 1　大数据存储架构内存布局示意图

47 位地址以下部分均为用户态地址空间，按照应用性能的不同，将其划分为了 7 个不同的地址，其中，内存映射区域地址与堆地址之间的一段地址未被存储架构使用，因此将其标记为内存空洞[6]。从大数据存储架构配置文件中读取相关的用户配置信息，定义架构分布式共享内存的物理资源与协议[7]。其次，基于线程信息的地址段，存储架构中所有线程的运行状态信息（以及 SNOP 运行过程中产生的日志）提供大数据存储架构运行所需的同步原语状态信息，在不同机器上访问堆和栈内存。

1.2　设计分布式共享内存协议

为提高架构的容错性，在设计分布式共享内存协议前，需要设计内存协议本身所使用的库名与代码，避免协议运行中出现数据存储无限递归现象。共享内存协议使用库名及代码如表 1 所示。

表 1　分布式共享内存协议使用库名及代码

编号	库名	代码
（1）	大数据存储框架为用户程序提供的函数库	Libsthread
（2）	提供远程内存直接访问的通信管理	Librdmacm
（3）	用于调试的信息库	Libdwarf
（4）	C 函数库	Libc
（5）	提供大数据存储框架使用的常用接口	Libsnop
（6）	用于栈回溯的库	Libunwind
（7）	POSIX 线程库	Libpthread

通过表 1 获取到大数据存储架构分布式共享内存协议本身所使用的库，避免协议在机器同步过程中引发无限递归。在此基础上，设计大数据存储架构分布式共享内存协议，如表 2 所示。

表 2　大数据存储架构分布式共享内存协议

类型	说明	协议权限
私有	当前机器单独占有内存页	读写协议权限
共享	当前机器与其他机器共有内存页	只读协议权限
无效	当前机器上没有内存页	无协议权限

按照表 2 所示的协议权限对大数据存储架构内存页进行共享操作，展现协议的基本逻辑。基于同步操作机制，以确保各个进程在访问共享内存时能够正确地协同工作，即便节点发生故障仍然能够正确地访问和更新共享内存中的数据，提高大数据存储架构的容错性[8]。

1.3　基于 Kalman 算法优化大数据存储架构可扩展性

在提高了大数据存储架构容错性的基础上，利用 Kalman 算法对大数据存储架构的可扩展性进行全方位的优化。

首先，收集大数据存储架构存储节点的负载数据，包括 CPU 利用率、内存利用率、磁盘 IO 等。根据收集到的架构历史数据，利用 Kalman 算法，建立存储节点负载预测模型，对大数据存储架构的参数与状态作出估计[9]。Kalman 算法是一种线性动态系统的最优估计方法，其利用系统各时刻的测量值求得系统的状态值，并不断更新。Kalman 算法主要由两个步骤组成：预测步骤和更新步骤。

利用 Kalman 算法估计大数据存储架构的参数与状态的过程如下：

（1）定义系统模型。首先，定义一个用于描述存储节点负载的系统模型，通常包含一些状态变量，包括 CPU 利用率、内存利用率、磁盘 IO 等。

（2）初始化状态估计和协方差矩阵。在开始循环之前，初始化状态估计向量（即系统的初始状态）和协方差矩阵。其中，协方差矩阵用于描述系统状态估计的

误差。

（3）预测。在每个时间步内，使用系统模型和当前状态估计来预测下一个时间步的状态。在大数据存储架构中，这个步骤可能涉及对存储节点负载的预测。预测状态估计如下：

$$X(k+1|k)=\boldsymbol{F}(k)\times X(k|k)+\boldsymbol{G}(k)\times U(k) \tag{1}$$

其中，$X(k+1|k)$ 表示在 $k+1$ 时间步内基于 k 时间步信息的预测状态，$\boldsymbol{F}(k)$ 表示状态转移矩阵，$\boldsymbol{G}(k)$ 表示控制矩阵，$U(k)$ 表示控制输入。

（4）更新步骤。使用测量值（即实际负载）来更新对系统状态的估计。这涉及将预测状态与实际测量值进行比较，然后根据比较结果调整状态估计。调整的幅度取决于测量误差协方差矩阵和过程噪声协方差矩阵。

更新状态估计如下：

$$X(k+1|k+1)=X(k+1|k)\times(1-H(k+1))+K(k+1)\times\boldsymbol{Z}(k+1) \tag{2}$$

更新协方差矩阵如下：

$$\boldsymbol{P}(k+1|k+1)=(\boldsymbol{I}-K(k+1)\times H(k+1))\times\boldsymbol{P}(k+1|k) \tag{3}$$

其中，$H(k)$、$Q(k)$、$K(k)$ 均表示系统模型和噪声模型的参数，$\boldsymbol{Z}(k)$ 表示测量值，$\boldsymbol{P}(k+1|k+1)$ 和 $\boldsymbol{P}(k+1|k)$ 表示协方差矩阵，I 表示单位矩阵。

（5）循环执行。在每个时间步重复执行预测步骤和更新步骤，直到得到最优估计值。

根据最优估计值，衡量 Kalman 算法估计结果的好坏程度，估计值与实际值越接近越好，利用该值代替大数据存储架构的实际值[10]。在此基础上，提取与大数据存储结构存储节点负载相关的特征，描述存储节点负载的变化规律和趋势。根据 Kalman 算法预测结果，对存储节点的负载进行动态调整，以实现负载均衡。若发现负载不均或者出现其他问题，及时进行干预和处理。基于存储节点负载动态调整结果，对大数据存储架构进行优化，以提高其可扩展性。优化中包括增加架构存储节点、优化数据存储策略、改进数据处理流程等。定期重复以上步骤，以实现大数据存储架构的可扩展性持续优化。

2 实验与结果分析

2.1 实验准备

为了验证上述提出的基于 Kalman 算法的大数据存储架构可扩展性优化算法的可行性及可扩展性优化效果，开展了如下实验测试分析。

首先，基于该算法的运行需求及运行特征，创建实验所需的测试环境。实验环境配置如表 3 所示。

表 3 大数据存储架构可扩展性优化算法实验环境配置

编号	项目	配置
（1）	服务器	处理器：Intel Xeon E5 - 2680 v4@2.40 GHz；内存：64 GB DDR4 ECC 内存；硬盘：1 TB SSD 硬盘（系统盘）+4 TB SATA 硬盘（数据盘）；网络：1 Gb/s 以太网接口。
（2）	网络交换机	1 台支持 1 Gb/s 以太网接口的交换机
（3）	操作系统	与 CentOS 7.9 同等性能的 Linux 操作系统
（4）	Kalman 算法库	Python Kalman Filter 库
（5）	大数据处理框架	Apache Hadoop 3.3.0

服务器节点之间通过以太网交换机连接，形成一个星型拓扑结构，服务器节点与交换机之间的网络连接采用双绞线或光纤连接，以确保数据传输的稳定性和可靠性。创建好优化算法实验测试环境后，选取实验所需的工具，包括数据预处理工具与数据可视化工具两种。其中，数据预处理工具采用 Python Pandas 库；数据可视化工具采用 Matplotlib。完成实验测试准备后，应用上述本文提出的基于 Kalman 算法的大数据存储架构可扩展性优化算法，在进行实验之前，确保所有设备和软件都已正确安装和配置后，开展实验测试分析。

2.2 优化结果分析

完成以上实验测试准备后，接下来，对大数据存储架构可扩展性优化结果作出全方位、多维度的分析。使用公开可用的 ImageNet 数据集，该数据集是一个宝贵的大数据资源，包含了超过 1 400 万的图像，涵盖了 2 万多个类别。

为了增强可扩展性优化结果的说服力，将上述基于 Kalman 算法的大数据存储架构可扩展性优化算法设置为实验组，将文献［1］中提出的基于分布 K-means 算法的可扩展性优化算法、文献［2］中提出的基于 ARM 架构的均衡计算型服务器数据存储系统优化算法分别设置为对照组 1 与对照组 2，以对比分析的形式，判断本文提出的算法是否可行。

三种优化算法的大数据处理流程为：将大数据集内的海量数据分散到 4 个配置相同的高性能服务器上，以分散的方式进行处理，使用三种优化算法对应的数据处理模式，对大数据进行分析和计算。

选取大数据存储架构可扩展性优化后的资源利用率作为此次实验测试的评价指标，其计算公式如下所示：

$$Q = \frac{M_P}{M} \times 100\% \tag{4}$$

其中，M_P 表示大数据存储架构运行过程中实际使用的资源量；M 表示大数据存储架构运行过程中所能使用的最大资源总量。大数据架构资源利用率越高，说明系统对服务器资源的利用越充分，可扩展性能越好，能够更加高效地处理大规模数据，反之同理。

为避免实验测试结果存在偶然性，进行了 6 组实验，将 6 组实验分别标号为 A ~ F。测定三种优化算法应用后大数据存储架构资源利用率并作出客观对比，结果如图 2 所示。

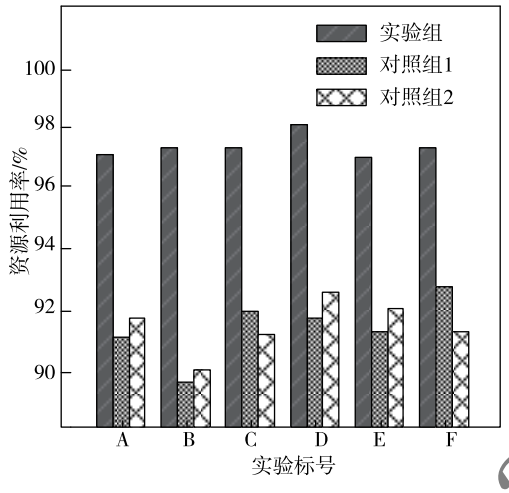

图 2　大数据存储架构资源利用率对比结果

在 6 组实验测试中，应用本文提出的基于 Kalman 算法的大数据存储架构可扩展性优化算法后，大数据存储架构的资源利用率始终高于另外两种算法，均达到了 96% 以上，最高达到了 98%，这意味着大数据存储架构在处理或管理数据方面效率更高，能够处理更多的工作负载。由此可知，本文提出的优化算法具有较高的可行性，架构可扩展性优化效果优势显著，对服务器资源的利用更加充分，能够更高效地处理大规模数据。

产生上述结果主要有如下三点原因：

（1）优化内存布局：本文算法综合考虑了大数据存储架构与多核环境内存布局之间的兼容性，通过合理设计架构内存布局，可以更好地满足多核环境下的数据存储和访问需求，从而提高了内存利用率。

（2）设计分布式共享内存协议：本文算法设计了一种分布式共享内存协议，确保各个进程在访问共享内存时能够正确地协同工作，避免了冲突和数据不一致性问题，提高了存储架构的容错性和整体性能。

（3）动态调整存储节点负载：利用 Kalman 算法可以动态地调整存储节点的负载，根据实际需求和系统状态来合理分配存储资源，避免了资源浪费和资源瓶颈的产生，进一步提高了资源利用率和系统性能。

3　结束语

本文通过引入优秀的 Kalman 算法，有效地提高了大数据存储架构的性能和可扩展性。该优化算法利用 Kalman 算法的自适应性和鲁棒性，对海量大数据进行优化和预测，提高了数据处理的效率和准确性。同时，采用分布式存储方式，将数据分散到多个节点上进行存储和处理，提高了数据的可扩展性和容错能力。该算法的研究和应用，为大数据存储和处理领域的发展提供了新的思路和方法，具有较为重要的理论意义和实践价值。

参考文献

[1] 莫理，柳本林，张树保，等. 基于分布式 K－means 算法的水电厂光纤测温系统可扩展性优化 [J]. 电子设计工程，2023，31（16）：107－111.

[2] 沈桂泉，沈伍强，张金波，等. 基于 ARM 架构的均衡计算型服务器数据存储系统 [J]. 自动化应用，2023，64（8）：210－212，217.

[3] 连彦泽，李鹏程，赵雷，等. 运载火箭试验大数据存储架构设计与应用 [J]. 遥测遥控，2022，43（6）：78－88.

[4] 方圆，张亮，盛剑桥，等. 基于微服务架构的电力数据存储自动加密系统设计 [J]. 自动化与仪器仪表，2022（1）：189－192，196.

[5] 陈学渊，梁璟，倪辰辰，等. 跨地域环境下的电磁数据分布式存储架构 [J]. 电子信息对抗技术，2021，36（6）：114－118.

[6] 张建军，吕琳，韩明，等. 一种基于干涉测角衰减记忆 Kalman 算法的仿真应用 [J]. 软件，2021，42（10）：57－59.

[7] 孙方圆，庞光垚. 云存储架构在创新创业大数据智慧服务平台的应用研究 [J]. 现代信息科技，2021，5（16）：5－9.

[8] 毛安琪，汤小春，丁朝，等. 集中式集群资源调度框架的可扩展性优化 [J]. 计算机研究与发展，2021，58（3）：497－512.

[9] 刘昕林，邓巍，黄萍，等. 基于 Hadoop 和 Spark 的可扩展性大数据分析系统设计 [J]. 自动化与仪器仪表，2020（3）：132－136.

[10] 王红心，龙文佳. 网络关联大数据的存储与计算架构设计 [J]. 数据通信，2020（1）：35－39.

（收稿日期：2023－10－13）

作者简介：

韩镇阳（1991－），通信作者，男，硕士，工程师，主要研究方向：计算机视觉、深度学习、数据融合分析与应用等。

张磊（1984－），男，硕士，工程师，主要研究方向：网络安全、大数据等。

任冬（1989－），男，硕士，工程师，主要研究方向：物联网、大数据、人工智能等。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部