HDFS数据副本均衡放置策略的改进

袁丽娜
(广州大学华软软件学院软件工程系 广州 510990)

摘 要 HDFS默认的数据副本放置策略仅仅只根据磁盘空间使用单个指标进行负载衡量,无法实现各节点真正的负载均衡。提出了一种基于性能的副本负载均衡放置改进策略,从磁盘空间负载能力、CPU处理能力、内存处理能力、磁盘读写处理能力、带宽等5个方面考究节点实际工作负载,并定义了一个负载能力模型。实验结果表明,该改进策略比默认策略能更好地实现副本的均衡放置。

关键词 HDFS,副本均衡放置策略,负载能力模型
中图法分类号 TP391 **文献标识码** A

Improvement of HDFS Balanced Placement Strategy

YUAN Li-na
(Department of Software Engineering, South China Institute of Software Engineering, Guangzhou 510990, China)

Abstract The default data replica placement policy for HDFS is only measured using a single metric based on disk space, the true load balancing of each node cannot be realized. This paper proposed an improvement strategy of the load balancing based on performance through five aspects, such as disk space, CPU processing power, memory processing power, disk read/write, bandwidth, and a load capacity model was defined. Experimental results indicate that the improvement strategy is better than the default policy.

Keywords HDFS, Balanced placement strategy, Load capacity model

随着计算机技术及互联网技术的高速发展,大数据呈爆炸式增长,Hadoop的应用也日益广泛。Hadoop的核心是分布式计算框架MapReduce和分布式文件系统HDFS。HDFS仍处于一个发展和完善的阶段,现有的数据管理策略依然存在许多不足之处。目前,Hadoop2.X版本中,data node数据副本存放磁盘的选择策略有两种方式,一种是磁盘目录轮询方式,另一种方式是选择磁盘空间足够大的磁盘进行存储。轮询方式存储数据副本能够保证所有磁盘都被使用,但是经常会出现各个磁盘数据存储不均衡的问题。根据磁盘空间剩余量来选择磁盘存储数据副本,只能通过磁盘空间使用率单个指标进行衡量,并不能准确反映各节点实际工作负载,不能实现真正的负载均衡。因此现有的副本放置策略仍然存在集群负载不均衡、性能不高等问题。

本文针对默认数据副本存放策略存在的不足进行了改进,提出了一种基于性能的副本负载均衡放置改进策略,从磁盘空间负载能力、CPU处理能力、内存处理能力、磁盘读写处理能力、带宽等5个方面考究节点实际工作负载,并定义了一个负载能力函数模型。实验结果表明,该改进策略比默认策略更好地实现了各节点的负载均衡,并提高了效率。

1 HDFS及副本技术概述

1.1 HDFS简介

分布式文件系统HDFS是基于Hadoop的分布式存储架构中基础数据存储。HDFS采用主/从架构,对整个分布式文件系统进行管理。一个HDFS集群分别是由一个名称节点(NameNode)和若干个数据节点(DataNode)组成。Hadoop2.x中HDFS的体系结构如图1所示。

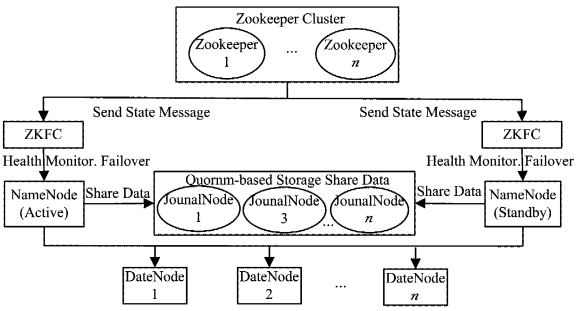

图1 HDFS体系结构

1.2 副本技术

为了提高系统的可靠性,HDFS中引入了副本冗余机制,即当集群中的某个节点出现损坏,系统可以从其他节点获取冗余备份数据。副本是一种能够提高数据访问效率和容错性能的技术。高效的副本冗余策略不但能够提高集群的可靠性,提升集群的负载均衡能力,同时也能够提高系统MapReduce并行计算的性能。

2 HDFS默认副本放置策略

当集群内新增、删除节点,或者某个节点机器内硬盘存储达到饱和值时,Hadoop的HDFS集群非常容易出现机器与机器之间磁盘利用率不平衡的情况。当HDFS负载不均衡时,需要对HDFS进行数据的负载均衡调整,即对各节点机器上数据的存储分布进行调整,进而让数据均匀地分布在各个DataNode上,均衡IO性能,防止热点的发生。数据副本放

袁丽娜 女,硕士,讲师,主要研究方向为数据库技术、大数据处理、数据仓库与数据挖掘,E-mail:18971656@qq.com。

置策略的核心是一个数据均衡算法，该数据均衡算法将不断迭代数据均衡逻辑，直至集群内数据均衡为止。该数据均衡算法每次迭代的逻辑如图 2 所示。

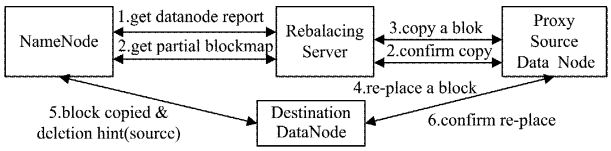

图 2 HDFS 负载均衡原理

数据均衡服务(Rebalancing Server)首先要求 NameNode 生成 DataNode 的数据分布分析报告，获取每个 DataNode 的磁盘使用情况；负载均衡器(Rebalancing Server)汇总需要移动的数据分布情况，计算具体数据块迁移路线图，确保网络内最短路径；开始数据块迁移任务，代理源节点(Proxy Source Data-Node)复制一块需要移动的数据块，将复制的数据块复制到目标 DataNode 上，删除原始数据块；目标 DataNode 向代理源节点确认该数据块迁移完成；代理源节点向负载均衡器确认本次数据块迁移完成。然后继续执行该过程，直至集群达到数据均衡。

3 基于性能的副本负载均衡放置改进策略

目前 HDFS 默认的负载均衡方法策略对节点负载量只采用了磁盘空间使用率这个单一的衡量指标。为进一步提高 HDFS 集群负载均衡的效果，充分考虑各节点的性能，针对 HDFS 默认的负载均衡方法策略的不足，提出了一种副本均衡放置改进策略，该策略除了考虑磁盘空间使用率，还综合考虑了各个节点的 CPU 处理能力、磁盘读写速度、内存处理能力和网络带宽。

3.1 性能衡量指标

本文通过对 Datanode 节点的负载性能的分析，根据分析获得的权值来选择最优的节点进行副本放置，从而实现集群的负载均衡。本文衡量节点的负载性能指标主要包括以下5 个。

(1)磁盘空间的负载能力

节点磁盘空间剩余量越大，剩余率越高，表示该节点磁盘空间负载能力越强。系统需将副本优先存放在磁盘空间负载能力强的数据节点上，让集群整体的磁盘负载更加均衡。磁盘空间剩余率用 P_{disk} 表示，$P_{disk}=1-disk_{use}$。

(2)CPU 负载能力

节点 CPU 的使用率越低则表示 CPU 的负载越小，该节点的 CPU 负载能力越强。系统会优先选择 CPU 负载能力较强的数据节点进行副本的存放。CPU 负载能力 $P_{CPU}=CPU_{主频}*CPU_{核心数}*(1-CPU_{使用率})$。其中，$CPU_{使用率}$ 表示 CPU 执行系统非空闲的进程的时间占 CPU 运行总时间的比例。

(3) 内存负载能力

节点的内存占有率越低则表示该节点的内存负载越小，内存负载能力越强。系统会优先选择内存负载能力较强的数据节点进行副本存放。内存负载能力 $P_{MEM}=MEM_{内存大小}*(1-MEM_{使用率})$，$MEM_{使用率}$ 表示数据节点运行系统进程所使用的内存占节点总物理内存的百分比。

(4)磁盘读写能力

磁盘的读写能力对任务执行时间有着直接的影响，系统优先选择读写能力强的数据节点。磁盘的读写能力是副本放置策略的重要指标，用参数 T 表示，Tw 表示磁盘的写速度，Tr 表示磁盘的读速度，α 为写操作相对于所有的读写操作所占的比率。磁盘读写能力表示为：$P_{I/O}=\alpha Tw+(1-\alpha)Tr$。

(5)带宽剩余率

网络带宽也会影响集群的负载均衡，带宽使用率指集群中所有的节点实际使用集群运行的带宽大小与总带宽大小的百分比，使用率越低说明集群越空闲，反之越繁忙。随机取两个时间点 M_1 和 M_2，即可以计算出此段时间内总的数据流量 $S=D_i*(M_2-M_1)(i\in[1,n])$，其中 D_i 指 i 节点的网络带宽，n 表示集群中数据节点的数量。某段时间内，数据节点 i 的网络带宽剩余率 $P_{Bandw}=1-S_{dj}/S$，S_{dj} 表示实际的数据流量。

3.2 负载能力模型

本文使用线性加权法来进行负载能力函数的建模。由于各个度量指标对于总目标所占据的重要程度有所不同，因此可以根据各个度量的各自重要性分别为其设定系数，并且将这些各自带有系数的度量指标值进行相加，从而得到总目标的值。可以得到改进的基于性能指标的负载能力函数如下：

$$P_i=A_1*P_{disk}+A_2*P_{cpu}+A_3*P_{mem}+A_4*P_{I/O}+A5*PB_{andw}$$

其中，P_i 表示负载综合剩余能力，A_1,A_2,A_3,A_4,A_5 分别为 5 个负载衡量指标的权值系数($\Sigma A_i=1$)，用于表示各指标对于数据节点负载能力影响的重要程度，此函数的权向量为 $\omega=(A_1,A_2,A_3,A_4,A_5)$。

3.3 计算权向量

根据权数产生方法的不同，多指标综合评价方法可分为主观赋权评价法和客观赋权评价法两大类，其中主观赋权评价法采取定性的方法由专家根据经验进行主观判断而得到权数，然后再对指标进行综合评价，如模糊评价法、层次分析法、指数加权法等，此方法有一定的主观随意性，受人为因素的干扰较大，在评价指标较多时难以得到准确的评价。而客观赋权评价法则根据指标之间的相关关系或各项指标的变异系数来确定权数进行综合评价，如变异系数法、神经网络分析法、熵值法等，该方法综合考虑了各指标间的相互关系，根据各指标所提供的初始信息量来确定权数，能使评价结果更加精确。本文主要采用变异系数法来计算权向量的初始值。

变异系数法是直接利用各项指标所包含的信息，通过计算得到指标的权重，是一种客观赋权的方法。该方法的基本做法是：在评价指标体系中，指标取值差异越大的指标，也即越难以实现的指标，越能反映被评价单位的差距。由于评价指标体系中的各项指标的量纲不同，不宜直接比较其差别程度，为了消除各项评价指标的量纲不同的影响，需要用各项指标的变异系数来衡量各项指标取值的差异程度。各项指标的变异系数公式如下：

$$V_i=\frac{\sigma_i}{\bar{x}_i},i=1,2,\cdots,n$$

其中，V_i 是第 i 项指标的变异系数，也称为标准差系数；σ_i 是第 i 项指标的标准差；$\bar{x}_i$ 是第 i 项指标的平均数。各项指标的权重为：

$$W_i=\frac{V_i}{\sum_{i=1}^{n}V_i}$$

计算各权值系数的仿真实验平台采用 Hadoop 集群，总共由 16 台 PC 机组成，具体硬件参数见 4.1 节实验环境，同时分别使用测试文件进行各种操作。由于涉及较多数据，在此并未详细列出，只给出计算步骤及结果。在操作过程中，查看并记录每次操作中各台机器的磁盘使用率、CPU 使用情况、内存使用情况、带宽使用情况等，然后根据 3.1 节性能衡

量指标中提及的磁盘空间负载能力公式、CPU 负载能力公式等，计算出每次操作中磁盘空间负载能力、CPU 处理能力、内存处理能力、磁盘读写能力和带宽等 5 个性能衡量指标，并计算出平均数和标准差，得到表 1 第 2—3 行数据。

表 1　指标平均数和标准差

指标	磁盘负载能力	CPU 处理能力	内存处理能力	磁盘读写能力	带宽剩余率	总和
平均数	0.55	0.52	0.49	0.59	0.71	—
标准差	0.22	0.11	0.06	0.20	0.03	—
变异系数	0.40	0.21	0.13	0.34	0.04	1.12
权重	0.36	0.18	0.12	0.30	0.04	1.00

再根据以上均值和标准差，通过变异系数公式计算各项指标的变异系数，并对所有变异系数求总和，得到表 1 的第 4 行数据。最后通过变异系数权重计算公式计算出各项指标的各个权重值，得到表 1 的第 5 行数据。由此可以得出最终的负载综合剩余能力评价函数为：

$$P_i = 0.36 * P_{disk} + 0.18 * P_{cpu} + 0.12 * P_{mem} + 0.3 * P_{I/O} + 0.04 * P_{Bandw}$$

4　实验结果及分析

本文主要提出了基于性能的副本负载均衡放置改进策略，为了测试改进策略的实际效果，对改进算法进行了实验验证，并与默认算法进行了比较。

4.1　实验环境

该仿真实验平台设置 Hadoop 集群包含 3 个机架 Rack1，Rack2 和 Rack3，每个机架包括 5 个 DataNode 节点。仿真实验平台 Hadoop 采用 Hadoop2.6.5 版本，JDK 采用 jdk1.7.0_45。

该仿真实验平台的 Hadoop 集群总共包括 16 台 PC 机，其网络拓扑结构如图 3 所示。

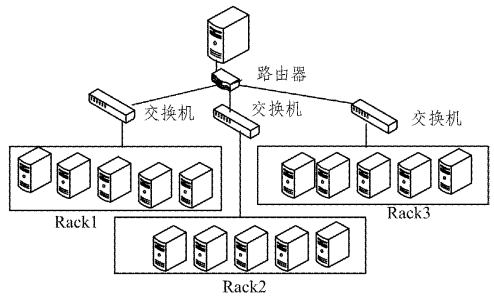

图 3　网络拓扑图

16 台 PC 机的主要硬件参数如表 2 所列。

表 2　硬件参数

	节点	CPU	内存/G	硬盘容量	I/O
	NameNode	i5	8	1T	7200
	DataNode1	i5	8	1T	7200
	DataNode2	i5	8	1T	7200
Rack1	DataNode3	i3	4	500G	7200
	DataNode4	i3	4	500G	7200
	DataNode5	i3	2	500G	5400
	DataNode1	i5	4	1T	7200
	DataNode2	i5	4	1T	7200
Rack2	DataNode3	i3	4	500G	7200
	DataNode4	i3	4	500G	7200
	DataNode5	i3	2	500G	5400
	DataNode1	i3	2	500G	7200
	DataNode2	i5	4	1T	7200
Rack3	DataNode3	i3	4	500G	7200
	DataNode4	i3	4	500G	7200
	DataNode5	i3	2	500G	5400

各节点带宽均为 100M。

4.2　实验结果分析

该仿真实验中，客户端分别对系统默认的副本放置策略和改进后的副本放置策略进行对比测试。通过客户端向服务器提交 2000 个数据块，默认副本冗余系数为 3，即系统总共需要处理 6000 个数据块。经统计得出，采用系统默认放置策略时，各个数据副本的分布情况如图 4 所示。

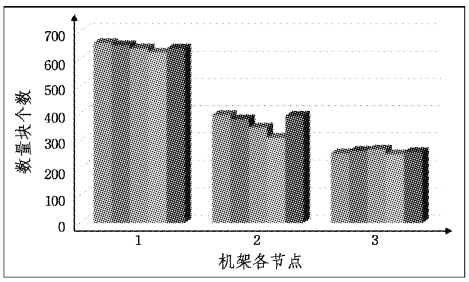

图 4　默认策略数据副本的分布情况

由图 4 的统计结果可知，若 HDFS 系统采用默认副本放置策略，则数据副本并未考虑数据节点的性能和负载的差别，而是较均匀地存放到机架各个节点上。

若采用改进后的基于性能的副本负载均衡放置改进策略，各个数据副本的分布情况如图 5 所示。

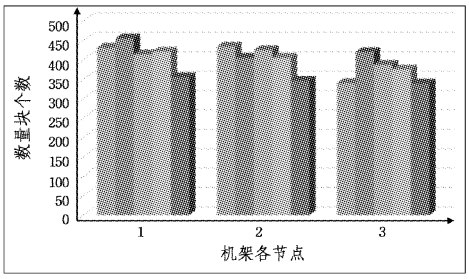

图 5　采用改进副本放置策略时数据副本的分布情况

由此可见，采用改进后的基于性能的副本负载均衡放置改进策略，更好地实现了系统的负载均衡。

在以上两种策略下将 30G 的输入文件上传到 HDFS 上，并且运行 wordcount 单词计数程序，统计其执行时间，对比图如图 6 所示。

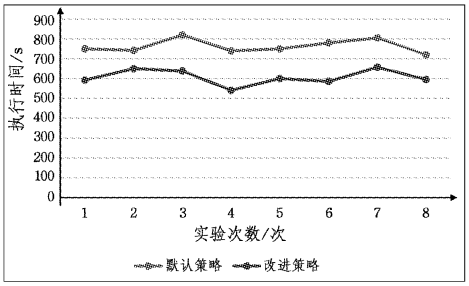

图 6　单词计数程序作业执行时间

由图 6 可知，改进之后的副本放置策略明显减少了作业执行时间，提升了集群的执行效率。

结束语　本文分析和研究了 HDFS 默认的副本放置策略，针对默认放置策略只使用磁盘空间使用率作为负载均衡的衡量指标这一不足，对其进行了简单改进，提出了一种基于性能的副本负载均衡放置改进策略，并通过一个负载能力模型进行了客观描述。负载能力模型从磁盘空间的负载能力、

（下转第 431 页）

表 3　测试样本均方根误差

方法	标准 LS-SVM	归一化 LS-SVM	TM-SVM	加权 LS-SVM
RMSE	4.9880	4.7770	4.7035	4.6715

结束语　本文提供了一种熵值法加权的最小二乘支持向量机的建模策略，主要针对多变量输入的回归分析问题。利用熵能够反映信息量的特性，根据每一维特征的离散程度，确定相应特征的权重大小；同时，对数据进行归一化处理，以消除量纲，避免各个特征之间的量级差异，将归一化后的数据与对应的特征权重相结合，得到能够充分反映特征信息的输入数据；最后，运用最小二乘支持向量机建立多输入的回归模型。仿真结果表明，所提方法取得了比常规方法更高的建模精度。

参 考 文 献

[1] HAYKIN S. Neural Networks: A Comprehensive Foundation [M]. Macmillan, 1998: 71-80.

[2] 焦李成. 神经网络系统理论[M]. 西安电子科技大学出版社, 1990.

[3] 张木想，马缚龙，肖国镇. 神经网络优化计算的新方法[J]. 电子学报，1993(7)：1-7.

[4] UKIL A. Support Vector Machine[J]. Computer Science, 2002, 1(4): 1-28.

[5] SUYKENS J A K, GESTEL T V, BRABANTER J D, et al. Least Square Support Vector Machine[J]. Euphytica, 2002, 2(2): 1599-1604.

[6] 阎威武，邵惠鹤. 支持向量机和最小二乘支持向量机的比较及应用研究[J]. 控制与决策，2003，18(3)：358-360.

[7] 李盼池，许少华. 支持向量机在模式识别中的核函数特性分析[J]. 计算机工程与设计，2005，26(2)：302-304.

[8] 巩知乐，张德贤，胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真，2009，26(7)：164-167.

[9] 何江平，文俊浩，邓恬洁，等. 基于支持向量机的图像识别[J]. 重庆大学学报(自然科学版)，2006，29(1)：57-60.

[10] 叶美盈，汪晓东，张浩然. 基于在线最小二乘支持向量机回归的混沌时间序列预测[J]. 物理学报，2005，54(6)：2568-2573.

[11] 柳小桐. BP 神经网络输入层数据归一化研究[J]. 机械工程与自动化，2010(3)：122-123.

[12] 刘冲，赵海滨，李春胜，等. 基于频带能量归一化和 SVM-RFE 的 ECoG 分类[J]. 仪器仪表学报，2011，32(3)：534-539.

[13] 常军，李祯，朱业玉，等. 基于支持向量机(SVM)方法的冬季温度预测[J]. 气象科技，2005(s1)：102-106.

[14] 黄安民，焦淑菲，任海青，等. 支持向量机结合近红外光谱法测定杉木木质素的含量[J]. 林产化学与工业，2009，29(5)：1-5.

[15] 郭显光. 熵值法及其在综合评价中的应用[J]. 财贸研究，1994(6)：56-60.

[16] PENG X, XU D. Twin Mahalanobis distance-based support vector machines for pattern recognition[J]. Information Sciences, 2012, 200(1): 22-37.

(上接第 399 页)

CPU 的处理能力、内存处理能力、磁盘读写处理能力、带宽等 5 个参数来反映实际工作负载，并根据变异系数法求出相关参数权重。实验结果表明，该副本放置改进策略比默认策略更好地实现了各节点的负载均衡，同时也提高了集群的工作效率。

参 考 文 献

[1] 周江，等. 面向大数据分析的分布式文件系统关键技术[J]. 计算机研究与发展，2014，51(2)：382-394.

[2] 罗鹏，等. HDFS 数据存放策略的研究与改进[J]. 计算机工程与设计，2014，35(4)：1127-1131.

[3] 孙知信，等. 基于云计算的数据存储技术研究[J]. 南京邮电大学学报，2014，34(4)：13-19.

[4] 王海荣，等. 基于 Hadoop 的海量数据存储系统设计[J]. 科技通报，2014，30(9)：127-130.

[5] 张柄虹，等. 空间高效的分布式数据存储方案[J]. 计算机应用研究，2015，32(5)：1508-1511.

[6] 马晓亭. 数字图书馆大数据分布式存储架构模式与策略研究[J]. 新世纪图书馆，2015(5)：43-46.

[7] 康承昆，等. 一种基于多衡量指标的 HDFS 负载均衡算法[J]. 四川大学学报，2014，51(6)：1163-1169.

[8] 熊安萍，等. 一种基于混合索引的 HDFS 小文件存储策略[J]. 重庆邮电大学学报，2015，27(1)：97-102.

[9] 英昌甜，等. 一种面向低延迟的内存 HDFS 数据存储策略[J]. 微电子学与计算机，2014，31(11)：160-166.

[10] 尹颖，等. HDFS 中高效存储小文件的方法[J]. 计算机工程与设计，2015，36(2)：406-409.

[11] 卢美莲，等. 基于 CMM 模型的 HDFS 负载均衡策略[J]. 北京邮电大学学报，2014，37(5)：20-25.

[12] 孟祥萍，等. 基于 hadoop 云平台的智能电网 HDFS 资源存储技术研究[J]. 电测与仪表，2014，51(19)：24-30.

[13] 张华伟，等. 基于多目标优化的云存储副本分布策略的研究[J]. 计算机科学，2015，42(4)：44-50.

[14] 肖达，苏丽裕，王俊龙，等. CSFS：云存储服务支撑的文件系统设计与实现[J]. 北京邮电大学学报，2015，38(6)：77-82.

[15] 罗芳，等. 基于多属性的海量 Web 数据关联存储及检索系统[J]. 计算机工程与科学，2014，36(3)：404-410.

[16] 宋宝燕，等. 基于范德蒙码的 HDFS 优化存储策略研究[J]. 计算机学报，2015，38(9)：1825-1837.

[17] 肖玉泽，等. HDFS 下海量小文件高效存储于索引方法[J]. 小型微型计算机系统，2015，36(10)：2218-2223.

[18] MISHNE G, DALTON J, LI Z H, et al. Fast data in the era of big data: Twitter's real-time related query suggestion architecture[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. 2013: 1147-1157.