(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104199820 A
(43)申请公布日 2014.12.10

(21)申请号 201410315763.5

(22)申请日 2014.07.03

(71)申请人 浙江大学
地址 310027 浙江省杭州市浙大路 38 号浙大计算机学院曹光彪东楼 505

(72)发明人 吴朝晖 何延彰 姜晓红 陈英芝 毛宇

(74)专利代理机构 杭州裕阳专利事务所（普通合伙）33221
代理人 应圣义

(51)Int.Cl.
G06F 17/30(2006.01)

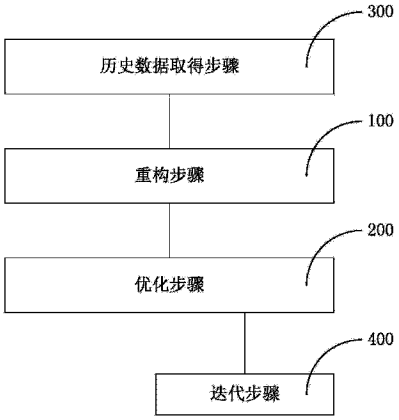
权利要求书1页 说明书5页 附图1页

(54)发明名称

云平台 MapReduce 工作流调度优化方法

(57)摘要

本发明涉及大数据计算，公开了一种云平台 MapReduce 工作流调度优化方法，大致地包括对现有的工作流进行重构得到新的工作流的重构步骤，应用遗传算法对工作流进行优化的优化步骤以及通过历史数据记录或者在建立相应的回归模型后记录回归模型的相关数据的方式保留上述历史数据的历史数据取得步骤，从而在优化步骤中，可以通过部分的历史数据生成不同的个体。本发明的优点在于，不仅考虑到了工作流作业的运行时间，同时考虑在云平台上计算过程中集群租用所需要的花费，优化效果好，可以确实地解决大型云计算平台上工作流调度过程中效率不高的问题。



CN 104199820 A

1. 一种云平台 MapReduce 工作流调度优化方法，其特征在于，包括以下具体步骤：

重构步骤 (100)：将用户提交的至少包括一个作业的工作流 W 重构为一个新的工作流 G，所述重构包括：

工作流中的新作业节点组成集合 V，以新作业为节点的有向无环图的节点之间的有向边组成集合 E，所述新作业包括开始作业 J_{Entry}、同步作业 J_{Syn}、结束作业 J_{Exit} 以及分支作业 J_{Bran}，所述开始作业 J_{Entry} 是指工作流 W 中没有任何父节点的作业，所述结束作业 J_{Exit} 是指工作流 W 中没有任何子节点的作业，所述同步作业 J_{Syn} 同时拥有父节点和子节点，并且具备父节点数量大于等于两个或者子节点数量大于等于两个的性质，所述分支作业 J_{Bran} 是指相互依赖的简单作业 J_{Sim} 的集合，所述相互依赖是指不同的简单作业 J_{Sim} 之间的有向边能够连通分支作业内所有的简单作业，所述简单作业 J_{Sim} 是指工作流 W 中只有一个父节点和一个子节点的作业；

计算工作流 G 中所有作业的输入数据集的大小，并将所述输入数据集的大小组成向量 S；

优化步骤 (200)：产生初始种群，所述初始种群是指通过对工作流 G 中的作业赋随机初值得到的个体的集合；通过产生新的个体的方式扩大初始种群内个体的数量得到下一代种群，所述新的个体是指由随机点交叉和 / 或随机点变异的方式所产生的新的个体；分别计算所述下一代种群中所有个体的工作时间，选取所述下一代种群中的至少一个个体作为优化结果输出。

2. 根据权利要求 1 所述的云平台 MapReduce 工作流调度优化方法，其特征在于，还包括历史数据取得步骤 (300)；

所述历史数据取得步骤 (300) 包括：任意选择一个用户提交的工作流 W；

以不同的作业配置参数以及集群虚拟机节点数目分别运行所述工作流 W 中的作业；将所述工作流 W 中的作业的运行结果进行保存。

3. 根据权利要求 2 所述的云平台 MapReduce 工作流调度优化方法，其特征在于，所述历史数据取得步骤 (300) 还包括：对运行于不同的作业配置参数以及集群虚拟机节点数目下的所述工作流 W 中的作业的运行结果进行拟合，得到拟合后曲线的参数。

4. 根据权利要求 2 所述的云平台 MapReduce 工作流调度优化方法，其特征在于，所述作业配置参数包括 Mapper 数量 N_M、Reducer 数量 N_R、输入数据的大小 S_{Input} 以及集群虚拟机节点的数目 $N_{Cluster}$，所述输入数据以分块的形式进行存储，所述 N_M、N_R 或 $N_{Cluster}$ 不大于 N_{Block}，所述 N_{Block} 是指所述输入数据的分块数目。

5. 根据权利要求 4 所述的云平台 MapReduce 工作流调度优化方法，其特征在于，所述输入数据以 64MB 每块的大小进行分块。

6. 根据权利要求 1 所述的云平台 MapReduce 工作流调度优化方法，其特征在于，所述初始种群或者下一代种群的染色体为

$\{N_{M1}, N_{R1}, N_{Cluster1}, N_{M2}, N_{R2}, N_{Cluster2}, \cdots, N_{M(k+1)}, N_{R(k+1)}, N_{Cluster(k+1)}\}$。

7. 根据权利要求 6 所述的云平台 MapReduce 工作流调度优化方法，其特征在于，所述染色体中每个元素包含两位数字。

云平台 MapReduce 工作流调度优化方法

技术领域

[0001] 本发明涉及大数据计算领域，特别涉及一种云平台 MapReduce 工作流调度优化方法，有效地提高了云平台上的工作流调度的优化效率。

背景技术

[0002] 随着以物联网、社交网站 SNS、生物信息学为代表的新型信息发布方式的产生和发展，人类社会的数据种类和数量正在以爆炸式的速度增长，大数据时代已经到来。目前，对于大数据尚未有一个公认的定义，它与传统的“海量数据”、“超大规模数据”等概念的区别，主要体现在大数据需要具备以下三个特点：规模性 (volume)、多样性 (variety) 和高速性 (velocity)。据统计，纽约证券交易所每天产生约 1TB 的交易数据，百度公司每天要处理的数据达到 10 ～ 100PB。大数据计算可以分为单作业单步计算、单作业迭代计算和多作业工作流计算等，每个作业可以由多个任务并行计算来加快运行的速度，即每个作业可以由若干数据并行的任务构成。

[0003] “云计算”为大数据提供了计算平台，它指通过互联网向用户提供的服务，包括基础设施即服务 (Infrastructure as a Service)、平台即服务 (Platform as a Service) 和软件即服务 (Software as a Service)。“云计算”通过网络，以付费即用 (Pay-as-you-go) 的方式，为全世界的用户提供基于效用的信息服务。

[0004] 按处理模式的不同分，处理大数据的框架可以分为流处理 (stream processing) 框架和批处理 (batch processing) 框架。批处理是先把数据存储后再处理 (store-then-process)，而流处理则是在数据产生后直接处理 (straight through processing)，在流处理中，数据的价值会随着时间的流逝骤减。大数据工作流可以由批处理作业或者流处理作业构成，现有的大数据处理优化方法只针对单一作业，并没有考虑在云平台运行时的集群租用费用。

[0005] 鉴于上述问题，在本发明中，我们拟针对云平台上大数据批处理工作流的性能和费用进行优化，以期能够研发一种可以更为有效地在调度过程中维持原 有条件下的优化效率的新型调度优化方法。

发明内容

[0006] 本发明针对现有技术中，优化方法依赖于初始条件，其优化效果会随时间而变化甚至减弱的缺点，提供了一种云平台 MapReduce 工作流调度优化方法，可以提供更为稳定的优化效果，有效地提高了工作流调度的优化效率。

[0007] 为实现上述目的，本发明可采取下述技术方案：

[0008] 一种云平台 MapReduce 工作流调度优化方法，包括以下具体步骤：

[0009] 重构步骤：将用户提交的至少包括一个作业的工作流 W 重构为一个新的工作流 G，所述重构包括：

[0010] 工作流中的新作业节点组成集合 V，以新作业为节点的有向无环图的节点之间的

有向边组成集合E，所述新作业包括开始作业J_{Entry}、同步作业J_{Syn}、结束作业J_{Exit}以及分支作业J_{Bran}，所述开始作业J_{Entry}是指工作流W中没有任何父节点的作业，所述结束作业J_{Exit}是指工作流W中没有任何子节点的作业，所述同步作业J_{Syn}同时拥有父节点和子节点，并且具备父节点数量大于等于两个或者子节点数量大于等于两个的性质，所述分支作业J_{Bran}是指相互依赖的简单作业J_{Sim}的集合，所述相互依赖是指不同的简单作业J_{Sim}的有向边能够连通分支作业内所有的简单作业，所述简单作业J_{Sim}是指工作流W中只有一个父节点和一个子节点的作业；

[0011] 计算工作流G中所有作业的输入数据集的大小，并将所述输入数据集的大小组成向量S；

[0012] 优化步骤：产生初始种群，所述初始种群是指通过对工作流G中的作业赋随机初值得到的个体的集合；通过产生新的个体的方式扩大初始种群内个体的数量得到下一代种群，所述新的个体是指由随机点交叉和/或随机点变异的方式所产生的新的个体；分别计算所述下一代种群中所有个体的工作时间，选取所述下一代种群中的至少一个个体作为优化结果输出。

[0013] 于本发明的实施例中，还包括历史数据取得步骤；

[0014] 所述历史数据取得步骤包括：任意选择一个用户提交的工作流W；以不同的作业配置参数以及集群虚拟机节点数目分别运行所述工作流W中的作业；将所述工作流W中的作业的运行结果进行保存。

[0015] 于本发明的实施例中，所述历史数据取得步骤还包括：对运行于不同的作业配置参数以及集群虚拟机节点数目下的所述工作流W中的作业的运行结果进行拟合，得到拟合后曲线的参数。

[0016] 于本发明的实施例中，所述作业配置参数包括Mapper数量N_M、Reducer数量N_R、输入数据的大小S_{Input}以及集群虚拟机节点的数目$N_{Cluster}$，所述输入数据以分块的形式进行存储，所述N_M、N_R或$N_{Cluster}$不大于N_{Block}，所述N_{Block}是指所述输入数据的分块数目。

[0017] 于本发明的实施例中，所述输入数据以64MB每块的大小进行分块。

[0018] 于本发明的实施例中，所述初始种群或者下一代种群的染色体为$\{N_{M1}, N_{R1}, N_{Cluster1}, N_{M2}, N_{R2}, N_{Cluster2}, \cdots, N_{M(k+1)}, N_{R(k+1)}, N_{Cluster(k+1)}\}$。

[0019] 于本发明的实施例中，所述染色体中每个元素包含两位数字。

[0020] 本发明具有以下的显著技术效果：

[0021] 优化效果好，稳定性高，通过任务的轮廓分析方法，在不同配置信息和虚拟集群规模的情况运行同一个任务，得到运行时间。使用最小二乘法进行多元线性回归，使用模型预测在新的配置参数下的运行时间。随着运行的进行，历史数据的不断积累，可以有效地保证调度方法可以随不同的工作流进行调整，从而降低工作流的变化对于优化效率的影响，通过这种方法达到提高整体优化效率的目的。

[0022] 通过改进的遗传算法来得到每个任务配置信息的近似最优解，收敛速度快，计算速度快，本发明不仅考虑工作流作业的运行时间，还能够考虑在云计算时代基础设施的租用花费。

附图说明

[0023] 图 1 为云平台 MapReduce 工作流调度优化方法的流程示意图。

具体实施方式

[0024] 下面结合实施例对本发明作进一步的详细描述。

[0025] 实施例 1

[0026] 一种云平台 MapReduce 工作流调度优化方法，如图 1 所示，包括以下具体步骤：

[0027] 重构步骤 100：本步骤的核心在于将用户提交的工作流 W 进行重构，从而生成一个具有全新结构，可以更好地适应遗传优化算法的新的工作流，具体而言，将用户提交的至少包括一个作业的工作流 W 重构为一个新的工作流 G，作为一种可选的方案，工作流 W 可以表示为 W(Γ, Λ, s, d)，Γ 为任务集，表示工作流 W 中所有作业的集合，此处，将一个作业作为一个任务，并视为该工作流 W 的有向无环图的节点，Λ 为有向变的集合，表示有向无环图中任意的两个节点之间的连接，s 表示工作流 W 的初始输入数据集的大小，d 表示工作流 W 的运行截止时间，即工作流 W 运行的结束时间。其中，所述重构包括以下具体步骤：

[0028] 工作流中的新作业节点组成集合 V，以新作业为节点的有向无环图的节点之间的有向边组成集合 E，所述新作业包括开始作业 J_{Entry}、同步作业 J_{Syn}、结束作业 J_{Exit} 以及分支作业 J_{Bran}，所述开始作业 J_{Entry} 是指工作流 W 中没有任何父节点的作业，所述结束作业 J_{Exit} 是指工作流 W 中没有任何子节点的作业，所述同步作业 J_{Syn} 同时拥有父节点和子节点，并且具备父节点数量大于等于两个或者子节点数量大于等于两个的性质，所述分支作业 J_{Bran} 是指相互依赖的简单作业 J_{Sim} 的集合，所述相互依赖是指不同的简单作业 J_{Sim} 的有向边能够连通分支作业内所有的简单作业，所述简单作业 J_{Sim} 是指工作流 W 中只有一个父节点和一节子节点的作业；

[0029] 计算工作流 G 中所有作业的输入数据集的大小，并将所述输入数据集的大小组成向量 S；上述重构由工作流调度器完成，该工作流调度器还包括计算近似最优解调度，并使用该近似最优解调度相对应的参数调度到云平台上运行。

[0030] 上述重构后的工作流 G 可以表示为 G(Γ, Λ, V, E, S, d)，为了减少作业之间数据传输产生的费用，在工作流 G 中，同一分支作业所包含的作业使用相同的虚拟机集群进行运行。

[0031] 优化步骤 200：产生初始种群，所述初始种群是指通过对工作流 G 中的作业赋随机初值得到的个体的集合；通过产生新的个体的方式扩大初始种群内个体的数量得到下一代种群，所述新的个体是指由随机点交叉和 / 或随机点变异的方式所产生的新的个体；分别计算所述下一代种群中所有个体的工作时间，选取所述下一代种群中的至少一个个体作为优化结果输出。此外，作为另一种可选择的方案，所述的优化结果也可以通过以下方式得到：计算最后生成的下一代种群的每个个体所对应调度策略的工作流运行所需要的租用费用，选取下一代种群中集群租用费用最小的个体作为最后的调度策略。

[0032] 作为一个可选的方案，通常地，设定初始种群的数目为 n，则通过随机点交叉和 / 或随机点变异的方式产生的下一代种群的数目为 2n。其中，

[0033] 随机点交叉是指采用以下的方式所生成的新的个体，这里所述的方式包括，在个体的染色体上设定至少一个的交叉点，该交叉点通常是随机设定的，然后依次将交叉点两侧的部分进行互换，生成一个新的染色体。

[0034] 随机点变异是指在染色体以二进制编码的系统中，它随机地将染色体的某一个基因由 1 变成 0，或由 0 变成 1。通过变异操作，可确保群体中遗传基因类型的多样性，以使搜索能在尽可能大的空间中进行，避免丢失在搜索中有用的遗传信息而陷入局部解，获得质量较高的优化解答。

[0035] 进一步地，作为一种生成下一代种群的个体选择方法，该方法包括：在下一代种群的数目为 2n 时，计算该下一代种群中每一个个体所对应的调度策略的加权有向无环图的关键路径作为工作流运行时间，选取运行时间最小的 n 个体成为下一代种群所包含的个体，并更新下一代种群，该处所记载的运行时间 ExecTime 可以使用下述的回归模型计算得到。此外，作为另外一种较为简化的生成下一代种群的个体选择方法，可以简单地通过计算每个个体对应的调度算法所需的运行时间 ExecTime，选取 ExecTime 最小的若干个体作为下一代种群的个体。

[0036] 进一步地，作为一种优化方案，所述优化步骤 200 还包括具有设定次数的迭代步骤 400，所述迭代步骤 400 包括：

[0037] 生成下一代种群后，计算该下一代种群中每个个体所对应的调度算法所需的运行时间 ExecTime，保留 ExecTime 的数值较小的若干个，优选为 n 个个体；以该下一代种群中的 n 个个体作为新的初始种群的个体以完成一次迭代过程。

[0038] 通常而言，上述迭代过程需要重复 6-10 次，以便于优化工作流 G 的运行。

[0039] 进一步地，为了能够更为准确地模拟最优的工作流调度，所述方法还包括历史数据取得步骤 300；

[0040] 所述历史数据取得步骤 300 包括：任意选择一个用户提交的工作流 W；以不同的作业配置参数以及集群虚拟机节点数目分别运行所述工作流 W 中的作业；将所述工作流 W 中的作业的运行结果进行保存。

[0041] 进一步地，对运行于不同的作业配置参数以及集群虚拟机节点数目下的所述工作流 W 中的作业的运行结果进行拟合，得到拟合后曲线的参数。进一步地，为了能够更为准确地反映工作流，所述作业配置参数还包括所述工作流 W 的输入数据集与输出数据集的大小比值 IORate，同样地作为运行结果进行保存。

[0042] 所述作业配置参数包括 Mapper 数量 N_M、Reducer 数量 N_R、输入数据的大小 S_{Input} 以及集群虚拟机节点的数目 $N_{Cluster}$，所述输入数据以分块的形式进行存储，所述 N_M、N_R 或 $N_{Cluster}$ 不大于 N_{Block}，所述 N_{Block} 是指所述输入数据的分块数目。该作业配置参数是与特定的作业 J 相关联的。

[0043] 作为一种可选的方案，上述拟合过程需要使用多元线性回归模型，优选地，可以采用以下的回归模型 $ExecTime = a\times N_M+b\times N_R+c\times N_{Cluster}+d\times S_{Input}+e$，其中，ExecTime 即所述工作流 W 中某个作业的运行时间。拟合过程中，在积累了足够多的历史数据后，可以计算得到上述回归模型中各个参数的取值或者取值范围；设定所述工作流 W 中输出数据集的大小与输入数据集的大小是成正比例关系的，由上述回归模型可以得到输出数据的大小 S_{Output}，令 $IORate = \frac{S_{Output}}{S_{Input}}$，最后，将作业 J 以及与其相关联的回归模型的参数以及比值 IORate 均进行保存，通常是存入一个名为 Job 的数据库表中。

[0044] 所述输入数据以 64MB 每块的大小进行分块，即分块数目 $N_{Block} = \left[\frac{S_{Input}}{64}\right]$。

[0045] 所述初始种群或者下一代种群的染色体为 $\{N_{M1}, N_{R1}, N_{Cluster1}, N_{M2}, N_{R2}, N_{Cluster2}, \cdots, N_{M(k+1)}, N_{R(k+1)}, N_{Cluster(k+1)}\}$，其中，(k+1) 是指工作流 W 中开始作业、同步作业、结束作业以及分支作业的数量的总和。

[0046] 所述染色体中每个元素包含两位数字，换而言之，染色体中的每个元素仅仅只包括一个两位数字，便于后期进行染色体操作。单个的染色体的长度为 6(k+1)。

[0047] 总之，以上所述仅为本发明的较佳实施例，凡依本发明申请专利范围所作的均等变化与修饰，皆应属本发明专利的涵盖范围。

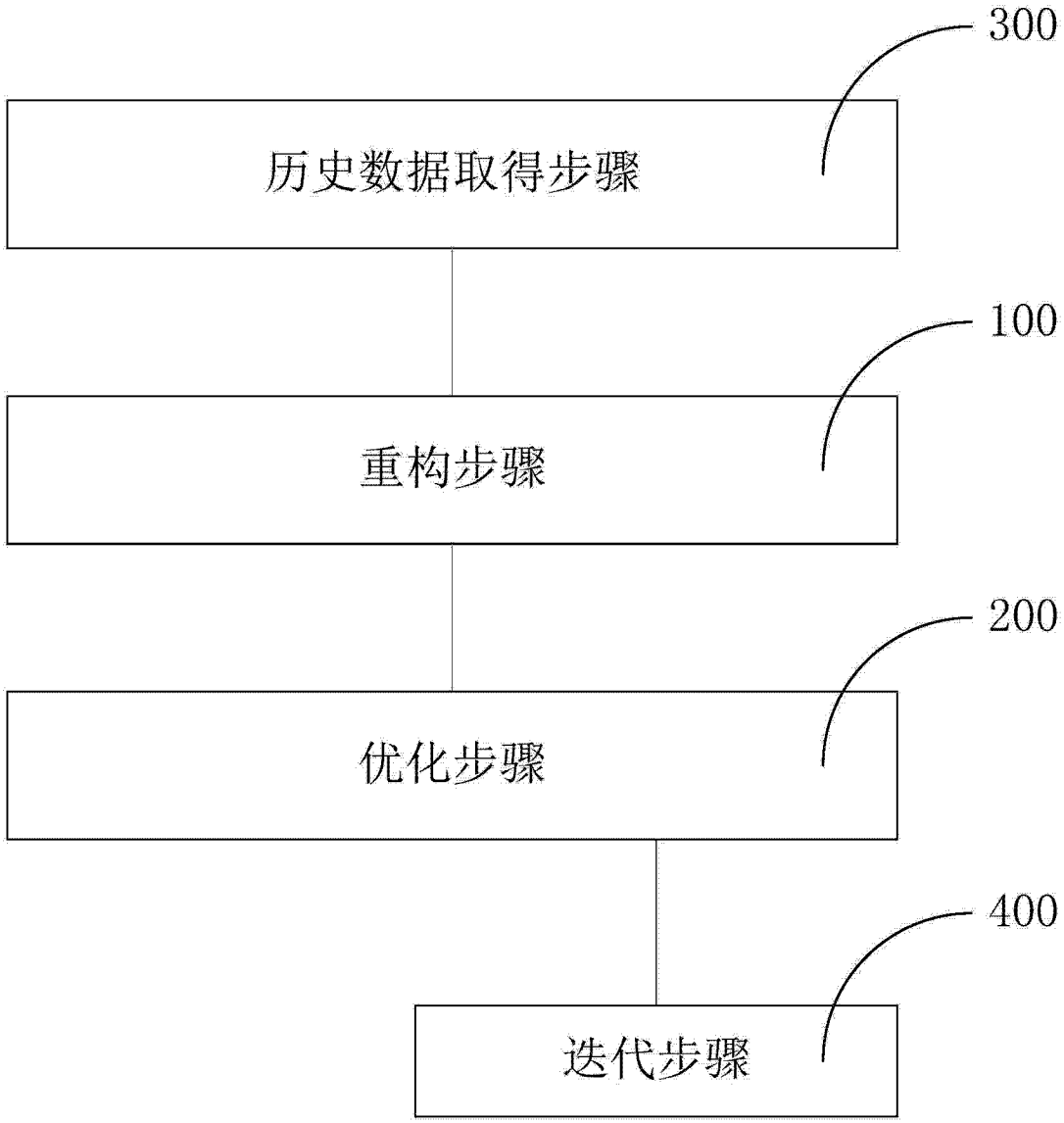


图 1