



(12) 发明专利

(10) 授权公告号 CN 104298475 B

(45) 授权公告日 2015. 11. 11

(21) 申请号 201410538144. 2

(22) 申请日 2014. 10. 13

(73) 专利权人 合一网络技术(北京)有限公司

地址 100080 北京市海淀区海淀大街 8 号中
钢国际广场 A 座 5 层 A、C 区

(72) 发明人 肖士锋 单明辉 卢学裕 姚健
潘柏宇 卢述奇

(74) 专利代理机构 北京市天玺沫泽专利代理事
务所(普通合伙) 11532

代理人 鲍晓

(51) Int. Cl.

G06F 3/06(2006. 01)

审查员 王晓渊

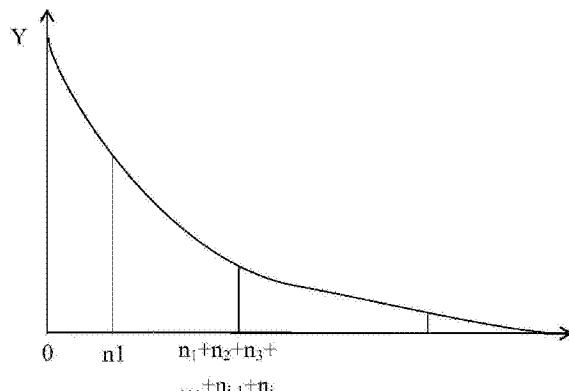
权利要求书1页 说明书4页 附图1页

(54) 发明名称

一种数据存储的优化方法

(57) 摘要

一种数据存储的优化方法，按照响应时间从短到长，单位存储成本从高到低，将存储分为了若干级，对于第 i 级存储， r_i 表示在该级存储中的响应时间， c_i 表示该级存储的单位存储的成本， n_i 表示该级存储的存储容量， N_i 表示所述第 i 级中 n_i 个数据的总访问次数，多级存储和每一级存储中按照访问次数对数据进行按序存储，可以得到平均响应时间和总存储成本的计算公式，利用上述公式作为约束条件，求得最优解，可以优化平均响应时间或总存储成本。本发明使得数据存储在合适的存储上，以平衡数据访问的平均响应时长与总的存储成本，满足业务需求。



1. 一种数据存储的优化方法,按照响应时间从短到长,单位存储成本从高到低,将存储分为了若干级,对于第一级存储, r_1 是响应时间, n_1 是存储的数据量, N_1 为第一级的 n_1 个数据的总访问次数,对于第 i 级存储, r_i 表示在该级存储中的响应时间, c_i 表示该级存储的单位存储的成本, n_i 表示该级存储的存储容量, N_i 表示所述第 i 级中 n_i 个数据的总访问次数,多级存储中按照访问次数对数据进行按序存储,将访问次数较高的数据按顺序放在访问响应时长较短的存储,在每一级存储中也按照访问次数高低进行排序,将访问次数高的数据放在前面,得到如下公式:

$$\text{平均响应时间 } \bar{r} = (N_1 * r_1 + N_2 * (r_2 + r_1) + N_3 * (r_3 + r_2 + r_1) + \dots + N_i * (r_i + r_{i-1} + \dots + r_1)) / N = f(n) \quad \text{公式 (1)}$$

$$\text{总存储成本 } C = n_1 * c_1 + n_2 * c_2 + \dots + n_i * c_i = g(n) \quad \text{公式 (2)}$$

其中, N 表示所有数据的总访问次数,

利用公式 (1) 和公式 (2) 作为约束条件,求得最优解,优化平均响应时间或总存储成本。

2. 根据权利要求 1 所述的数据存储的优化方法,其特征在于:

当平均响应时长上限限时,求得一组 n_i 最优解,使得 $g(n)$ 最小,即总存储成本最小。

3. 根据权利要求 1 所述的数据存储的优化方法,其特征在于:

当总存储成本上限限时,求得一组 n_i 最优解,使得 $f(n)$ 最小,即平均响应时长最小。

4. 根据权利要求 2 或 3 所述的数据存储的优化方法,其特征在于:

在对公式最优化求解时,采用线搜索全局或动态规划的计算方法来求出最优解。

5. 一种对数据进行存储的方法,其特征在于,利用权利要求 1-4 中任一项所述的优化结果,对数据进行存储。

一种数据存储的优化方法

技术领域

[0001] 本申请涉及大数据存储领域,特别的,涉及一种数据存储的优化方法,以及一种对数据进行存储的方法。

背景技术

[0002] 随着互联网技术的发展,网络所使用的数据越来越多。如何对数据进行保存,特别是,如何对海量数据进行保存成为现有技术亟需解决的问题。

[0003] 目前按照存储的访问方式或者访问存储的响应时长可以将存储分成若干级,比如内存级,响应时长最短,但是单位存储成本最高;memcache 集群,响应时长较短,单位存储成本较高;hbase 集群,将数据持久化在硬盘上,响应时长较长,单位存储成本较低,等等。

[0004] 如上所述,互联网服务,例如设计在线接口服务时,通常面临存储数据量大的问题,需要为数据选择合适的存储方式,以求平均响应时长能满足业务需求同时总的存储成本最低,或者在存储成本有限的时,平均响应时长最低。大数据存储通常需要规划多级存储来满足业务需求,将访问频率高的数据存储在成本高响应时长短的存储上,访问频率相对较低的存储在成本较高响应时长较短的存储上,访问频率最低的数据存储在成本最低但是响应时长最大的存储上。

[0005] 例如:某视频网站有超过数亿的视频资源,在设计根据视频 ID 获取视频相关信息这个服务时面临着如何优化多级存储的问题,如果选择三级存储,可以将过去一天或者一周内访问次数最高的一部分视频相关信息放在内存;访问次数较高的一部分视频相关信息存储在 memcache 集群;访问次数相对较低的存储在 hbase 集群。但是,如何合理选择每部分的比例以达到平均响应时长与总的存储成本的最优化,或者说,如何对数据存储进行优化,以用于降低数据的平均响应时间,进一步的,达到降低数据存储成本,成为现有技术亟需解决的技术问题。

发明内容

[0006] 本发明的目的在于提出一种数据存储的优化方法,以及利用该优化结果对数据进行存储的方法,通过该方法,可以降低数据的平均响应时间,进一步的,达到降低数据存储成本。

[0007] 为达此目的,本发明采用以下技术方案:

[0008] 一种数据存储的优化方法,按照响应时间从短到长,单位存储成本从高到低,将存储分为了若干级,对于第一级存储,r₁是响应时间,n₁是存储的数据量,N₁为第一级的n₁个数据的总访问次数,对于第 i 级存储,r_i表示在该级存储中的响应时间,c_i表示该级存储的单位存储的成本,n_i表示该级存储的存储容量,N_i表示所述第 i 级中 n_i 个数据的总访问次数,多级存储中按照访问次数对数据进行按序存储,将访问次数较高的数据按顺序放在访问响应时长较短的存储,在每一级存储中也按照访问次数高低进行排序,将访问次数高的数据放在前面,可以得到如下公式:

[0009] 平均响应时间 $\bar{r} = (N_1 * r_1 + N_2 * (r_2 + r_1) + N_3 * (r_3 + r_2 + r_1) + \dots + N_i * (r_1 + r_2 + \dots + r_{i-1})) / N = f(n)$ 公式 (1)

[0010] 总存储成本 $C = n_1 * c_1 + n_2 * c_2 + \dots + n_i * c_i = g(n)$ 公式 (2)

[0011] N 表示所有数据的总访问次数,

[0012] 利用公式 (1) 和公式 (2) 作为约束条件, 求得最优解, 可以优化平均响应时间或总存储成本。

[0013] 优选地, 当平均响应时长上限限时, 可求得一组 n_i 最优解, 使得 $g(n)$ 最小, 即总存储成本最小。

[0014] 优选地, 当总存储成本上限限时, 可求得一组 n_i 最优解, 使得 $f(n)$ 最小, 即平均响应时长最小。

[0015] 优选地, 在对公式最优化求解时, 可以采用线搜索全局或动态规划的计算方法来求出最优解。

[0016] 本发明还公开了一种对数据进行存储的方法, 其特征在于, 利用上述的优化结果, 对数据进行存储。

[0017] 因此, 本发明在选择多级存储时, 基于数据的历史访问规律, 建立访问响应时间和存储成本的表达公式, 并进行优化, 使得数据存储在合适的存储上, 以平衡数据访问的平均响应时长与总的存储成本, 满足业务需求。

附图说明

[0018] 图 1 是根据本发明的多级数据存储的示意图;

[0019] 图 2 是根据本发明的数据访问次数与多级数据存储的关系图。

具体实施方式

[0020] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是, 此处所描述的具体实施例仅仅用于解释本发明, 而非对本发明的限定。另外还需要说明的是, 为了便于描述, 附图中仅示出了与本发明相关的一部分而非全部结构。

[0021] 在现有技术中, 通常单位容量的存储成本与响应时间成反比, 也就是说单位存储成本越大, 响应时间越小; 单位存储成本越小, 响应时长越大。比如内存与硬盘, 内存单位存储成本高, 但响应时长短, 硬盘单位存储成本低, 但响应时长长。

[0022] 大数据存储中通常采用多级存储的方式, 参见图 1, 示出了根据本发明的多级数据存储的示意图。其中第一级存储通常存储访问最频繁的数据, 此时, 响应时间最短, 单位存储成本最高, 二级存储, 响应时长较短, 单位存储成本较高, 以此类推。

[0023] 本发明的数据存储的优化方法, 将存储单元按照响应时间从短到长, 单位存储成本从高到低, 将存储分为了若干级, 对于第一级存储, r_1 是响应时间, n_1 是存储的数据量, n_1 个数据的总访问次数 N_1 , 对于第 i 级存储, r_i 表示在该级存储中的响应时间, c_i 表示该级存储的单位存储的成本, n_i 表示该级存储的存储容量, N_i 表示所述第 i 级中 n_i 个数据的总访问次数。显然, N_i 所表示总访问次数, 应当是这 n_i 个数据的访问次数之和。

[0024] 分析过去一段时间内数据的访问规律, 内容按照访问次数由高到低排序, 访问次数最多的数据编号为 1, 第二多的编号为 2, … 编号为 k 的数据总访问次数为 f_k , 依次类推。如图 2 所示, 横坐标为数据编号, 纵坐标为访问次数。将访问次数高的数据按顺序放在

访问响应时长短的存储,将访问次数较高的数据按顺序放在访问响应时长较短的存储,在每一级存储中也按照访问次数高低进行排序,将访问次数高的数据放在前面,依次类推。具体而言,将编号在 1-n1 的数据存储在第一级存储,编号为 $\sum_{k=1}^{i-1} n_k + 1 - \sum_{k=1}^i n_k$ 的数据

存储在第 i 级存储。横坐标为每个数据的编号,编号已经按照数据的访问次数排序,编号为 1 的数据访问次数最多,编号为 i 个数据访问次数第 i-1 多;纵坐标 Y 为访问次数。每个数据的总访问次数可以通过经验得到。或者,对于某些已有待存储数据,存在着在过去一段时间内的访问规律文件,这些文件里记载数据的每一次访问以及访问时间等信息,从这些文件里统计每个数据的访问次数,从而得到每个数据的总访问次数,以及每级存储的总访问次数。

[0025] 访问数据时,根据数据的 ID 首先从第一级存储查询,若查询到则返回数据相关信息,查询不到再查询第二级,直到访问到位置。其中,第 i 级数据的访问时长为 $r_1+r_2+\dots+r_i$,

[0026] 因此,可以得到如下公式:

[0027] 平均响应时间 $\bar{r} = (N_1 * r_1 + N_2 * (r_2 + r_1) + N_3 * (r_3 + r_2 + r_1) + \dots + N_i * (r_1 + r_2 + \dots + r_{i-1})) / N = f(n)$ 公式 (1);

[0028] 总存储成本 $C = n_1 * c_1 + n_2 * c_2 + \dots + n_i * c_i = g(n)$ 公式 (2)。

[0029] N 表示所有数据的总访问次数。

[0030] 利用公式 (1) 和公式 (2) 作为约束条件,求得最优解,可以优化平均响应时间或总存储成本。

[0031] 例如,当平均响应时长上限限时,可求得一组 n_i 最优解,使得 $g(n)$ 最小即总存储成本最小。

[0032] 当总存储成本上限限时,可求得一组 n_i 最优解,使得 $f(n)$ 最小,即平均响应时长最小。

[0033] 在对公式最优化求解时,是典型的等式约束最优化问题,可以采用线搜索全局,动态规划等其它的计算方法求出最优解。

[0034] 因此,通过上述的方法,可以对数据存储进行优化。

[0035] 实施例 1:

[0036] 有 0, 1, 2 三个数据,分析他们在过去段时间的访问次数分别为 300, 100, 50 次,

[0037] 现有 3 级存储,单位存储成本分别为 100, 10, 5;访问响应时长分别为 1, 10, 100

[0038] 假设业务需求要求平均响应时长不得超过 36,则

[0039] 公式 1:访问响应时间 $r = ((N_1 * 1 + N_2 * (1 + 10) + N_3 * (1 + 10 + 100)) / 450$

[0040] 公式 2:总存储成本 $C = n_1 * 100 + n_2 * 10 + n_3 * 5$

[0041] 可以得出 $r \leq 36$ 时, $n_1 = 1, n_2 = 1, n_3 = 1$, 总成本 C 可取最小值为 115。从而分别在第一级至第三级存储中分别存储 0, 1, 2。

[0042] 进一步的,本发明还公开了数据存储的方法,利用上述数据存储优化的结果进行数据存储。

[0043] 因此,本发明提供了一种数据存储的优化方法,以及一种对数据进行存储的方法,在选择多级存储时,基于数据的历史访问规律,建立访问响应时间和存储成本的表达公式,并进行优化,使得数据存储在合适的存储上,以平衡数据访问的平均响应时长与总的存储

成本,满足业务需求。

[0044] 以上内容是结合具体的优选实施方式对本发明所作的进一步详细说明,不能认定本发明的具体实施方式仅限于此,对于本发明所属技术领域的普通技术人员来说,在不脱离本发明构思的前提下,还可以做出若干简单的推演或替换,都应当视为属于本发明由所提交的权利要求书确定保护范围。

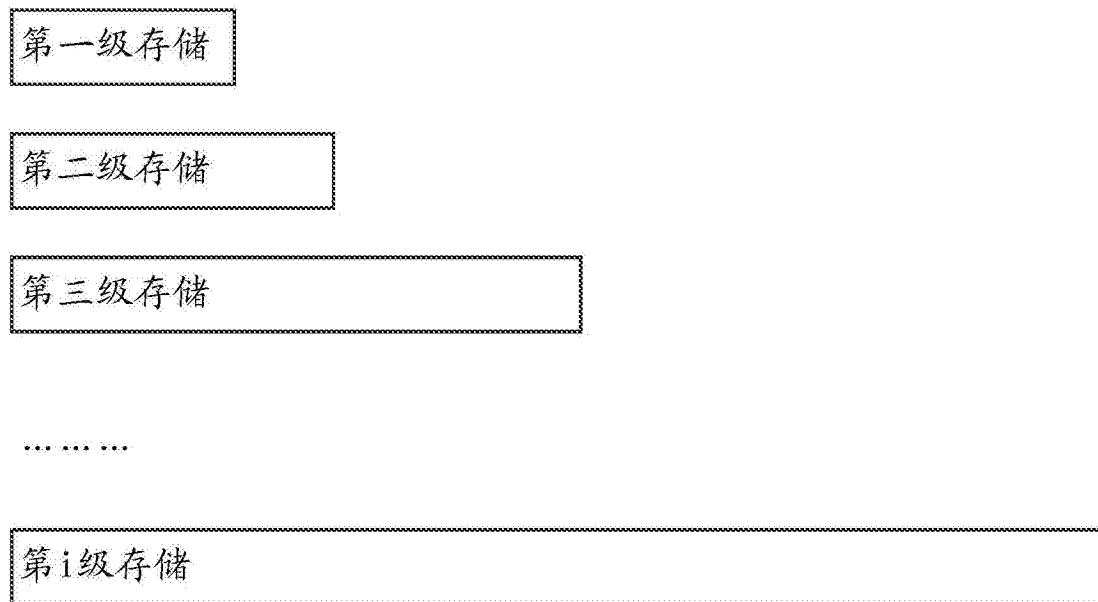


图 1

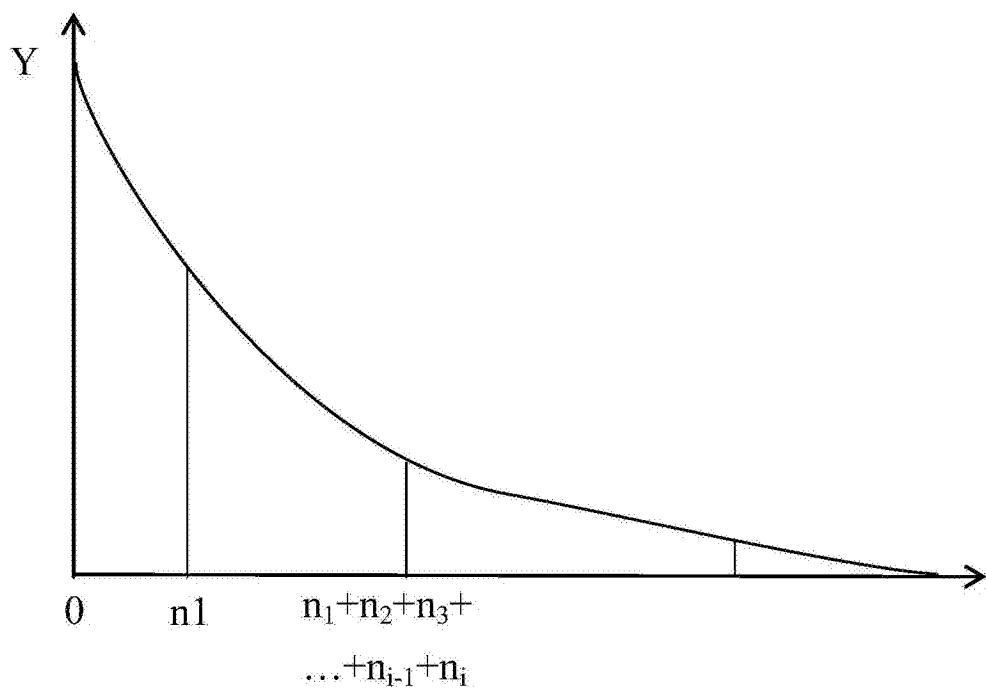


图 2