(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 5 月 14 日 (14.05.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/094064 A1

(51) 国际专利分类号:
G06F 16/182 (2019.01)

(21) 国际申请号: PCT/CN2019/116024

(22) 国际申请日: 2019 年 11 月 6 日 (06.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811323508.X 2018年11月7日 (07.11.2018) CN

(71) 申请人: 中兴通讯股份有限公司 (ZTE CORPORATION) [CN/CN]; 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。

(72) 发明人: 胡晓东 (HU, Xiaodong); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。 张东涛 (ZHANG, Dongtao); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。 辛丽华 (XIN, Lihua); 中国广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦, Guangdong 518057 (CN)。

(74) 代理人: 北京聿宏知识产权代理有限公司 (YUHONG INTELLECTUAL PROPERTY LAW FIRM); 中国北京市西城区宣武门外大街6号庄胜广场第一座西翼713室吴大建/霍玉娟, Beijing 100052 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,

(54) **Title:** PERFORMANCE OPTIMIZATION METHOD, DEVICE, APPARATUS, AND COMPUTER READABLE STORAGE MEDIUM

(54) **发明名称**: 性能优化方法、装置、设备及计算机可读存储介质

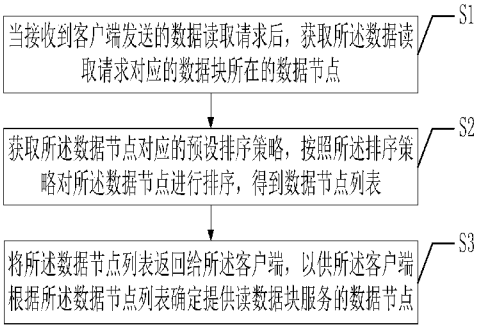

图 3

S1 Upon receiving a data read request sent by a client, acquire data nodes having a data block corresponding to the data read request

S2 Acquire a pre-determined sorting policy corresponding to the data nodes, sort the data nodes according to the sorting policy, and obtain a data node list

S3 Return the data node list to the client, such that the client determines, according to the data node list, a data node to provide a data block reading service

(57) Abstract: A performance optimization method, a device, an apparatus, and a computer readable storage medium. The method comprises: upon receiving a data read request sent by a client, acquiring data nodes having a data block corresponding to the data read request (S1); acquiring a pre-determined sorting policy corresponding to the data nodes, sorting the data nodes according to the sorting policy, and obtaining a data node list (S2); and returning the data node list to the client, such that the client determines, according to the data node list, a data node to provide a data block reading service (S3).

(57) 摘要: 一种性能优化方法、装置、设备及计算机可读存储介质，所述方法包括：当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点（S1）；获取所述数据节点对应的预设排序策略，按照所述排序策略对所述数据节点进行排序，得到数据节点列表（S2）；将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点（S3）。

WO 2020/094064 A1

[见续页]

NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

— 发明人资格(细则4.17(iv))

本国际公布:

— 包括国际检索报告(条约第21条(3))。

性能优化方法、装置、设备及计算机可读存储介质

本公开要求享有 2018 年 11 月 07 日提交的名称为“性能优化方法、装置、设备及计算机可读存储介质”的中国专利申请 CN201811323508.X 的优先权，其全部内容通过引用并入本文中。

技术领域

本公开涉及通信技术领域，尤其涉及一种性能优化方法、装置、设备及计算机可读存储介质。

背景技术

Hadoop 是一个开源分布式计算平台，Hadoop 分布式文件系统（Hadoop Distributed File System，HDFS）是 Hadoop 的一个核心组成部分，目前广为大数据服务所应用。HDFS 在 Hadoop 中主要负责存储文件数据。HDFS 上的文件按照数据块进行存储。数据块是一个抽象概念，它是文件存储处理的逻辑单元。一个数据块通常有多个副本以增加数据安全性，数据块的多个副本通常被存放在不同的数据节点中，这些数据节点可能在同一个机架中，也可能在不同机架中。当 HDFS 中的客户端要读取文件时，通常优先读取与其距离最近的数据节点中的数据块副本，如读取与其在同一机架中的数据节点中的数据块副本。此时，客户端总是访问与其距离最近的数据节点，与客户端距离最近的数据节点会压力过大，相对较远的数据节点则压力偏小，从而导致 HDFS 压力分布不均匀，整个 HDFS 的读性能下降。

发明内容

本公开的主要目的在于提供一种性能优化方法、设备及计算机可读存储介质，旨在解决在 HDFS 中由于客户端总是从与其距离最近的数据节点中读取数据，导致的 HDFS 压力分布不均匀，整个 HDFS 的读性能下降的技术问题。

为实现上述目的，本公开提供一种性能优化方法，所述性能优化方法包括步骤：当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点；获取所述数据节点对应的预设排序策略，按照所述排序策略对所述数据节点进行排序，得到数据节点列表；将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点。

此外，为实现上述目的，本公开还提供一种性能优化装置，其中，所述性能优化装置

包括：获取模块，用于当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点；获取所述数据节点对应的预设排序策略；排序模块，用于按照所述排序策略对所述数据节点进行排序，得到数据节点列表；数据返回模块，用于将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点。

此外，为实现上述目的，本公开还提供一种性能优化设备，所述性能优化设备包括存储器、处理器和存储在所述存储器上并可在所述处理器上运行的性能优化程序，所述性能优化程序被所述处理器执行时实现如上所述的性能优化方法的步骤。

此外，为实现上述目的，本公开还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有性能优化程序，所述性能优化程序被处理器执行时实现如上所述的性能优化方法的步骤。

附图说明

图1本公开实施例方案涉及的硬件运行环境的结构示意图；

图2本公开实施例方案涉及的HDFS中数据读的流程图；

图3为本公开性能优化方法较佳实施例的流程示意图；

图4为本公开实施例方案涉及的一种将数据节点按压力大小排序的排序图；

图5为本公开实施例方案涉及的一种将数据节点按与客户端之间的距离排序的排序图；

图6为本公开实施例方案涉及的一种将数据节点按与客户端之间的距离以及压力排序的排序图；

图7为本公开性能优化装置较佳实施例的功能示意图模块图。

本公开目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本公开，并不用于限定本公开。

由于目前存在 HDFS 中客户端总是从与其距离最近的数据节点中读取数据，导致的 HDFS 压力分布不均匀，整个 HDFS 的读性能下降的技术问题，本公开提供一种解决方案，通过当接收到客户端发送的数据读取请求后，获取该数据读取请求对应的数据块所在的数据节点；获取数据节点对应的预设排序策略后，按照该排序策略对该数据节点进行排序，

得到数据节点列表；将该数据节点列表返回给该客户端，以供该客户端根据该数据节点列表确定提供读数据块服务的数据节点。避免了客户端总是从与其距离最近的数据节点中读取数据块，减小了与客户端距离最近的数据节点的压力，避免了 HDFS 压力分布不均匀、整个 HDFS 的读性能下降的问题。

本公开提供了一种性能优化设备，参照图 1，图 1 是本公开实施例方案涉及的硬件运行环境的结构示意图。

需要说明的是，图 1 即可为性能优化设备的硬件运行环境的结构示意图。本公开实施例性能优化设备可以是 PC、服务器，例如 HDFS 的元数据服务器，也可以是智能手机、平板电脑、便携计算机等可移动式终端设备。

如图1所示，该性能优化设备可以包括：处理器1001，例如CPU，网络接口1004，用户接口1003，存储器1005，通信总线1002。其中，通信总线1002用于实现这些组件之间的连接通信。用户接口1003可以包括显示屏（Display）、输入单元比如键盘（Keyboard），可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口（如WI-FI接口）。存储器1005可以是高速RAM存储器，也可以是稳定的存储器（non-volatile memory），例如磁盘存储器。存储器1005可选的还可以是独立于前述处理器1001的存储装置。

在一个实施例中，性能优化设备还可以包括、摄像头、RF（Radio Frequency，射频）电路，传感器、音频电路、WiFi模块等等。本领域技术人员可以理解，图2中示出的性能优化设备结构并不构成对性能优化设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

如图1所示，作为一种计算机存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及性能优化程序。

在图 1 所示的性能优化设备中，网络接口 1004 主要用于连接其他数据节点，名字节点或客户端；HDFS 运维人员可通过用户接口 1003 触发设置指令；而处理器 1001 可以用于调用存储器 1005 中存储的性能优化程序，并执行以下操作：当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点；获取所述数据节点对应的预设排序策略，按照所述排序策略对所述数据节点进行排序，得到数据节点列表；将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点。

在一个实施例中，当所述排序策略为第一排序策略时，所述按照所述排序策略对所述数据节点进行排序，得到数据节点列表的步骤包括：获取所述数据节点对应的压力值；根据所述压力值确定所述数据节点对应的压力，将所述数据节点按照所述压力从小到大的顺序排序，得到数据节点列表。

在一个实施例中，所述获取所述数据节点对应的压力值的步骤包括：获取所述数据节点的压力数据；根据所述压力数据和预设的压力数据分值标准，得到所述数据节点的压力数据分值；根据所述压力数据分值和对应预设的压力数据权重值，计算得到所述数据节点对应的压力值。

在一个实施例中，当所述排序策略为第二排序策略时，所述按照所述排序策略对所述数据节点进行排序，得到数据节点列表的步骤包括：将所述数据节点按照与所述客户端的距离由近到远的顺序排序，得到预处理数据节点列表；获取所述数据节点对应的压力值，检测所述数据节点对应的压力值是否满足预设条件；当检测到所述数据节点对应的压力值满足预设条件时，将所述压力值满足预设条件的数据节点移动到所述预处理数据节点列表的末端，得到处理后的数据节点列表。

在一个实施例中，所述将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点的步骤包括：将所述数据节点列表返回给所述客户端，以供所述客户端将所述数据节点列表中排在最前面的数据节点确定为提供读数据块服务的数据节点。

在一个实施例中，所述当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点的步骤之前，处理器 1001 可以调用存储器 1005 中存储的性能优化程序，还执行以下操作：当接收到设置所述排序策略的设置请求后，根据所述设置请求设置所述数据节点对应的排序策略。

基于上述的硬件结构，提出本公开性能优化方法的各个实施例。在本公开性能优化方法的各个实施例中，为了便于描述，以 HDFS 的元数据服务器名字节点为执行主体进行阐述各个实施例。HDFS 体系结构中主要有两类节点，一类是名字节点，一类是数据节点。名字节点是 HDFS 的元数据服务器，用于管理并协调数据节点的工作，其内存中保存整个 HDFS 的两类元数据：（1）文件系统的名字空间，即文件目录树；以及文件的数据块索引，即每个文件对应的数据块列表；（2）数据块与数据节点的映射，即数据块存储在哪个数据节点上。从名字节点中可以获得每个文件的每个数据块所在的数据节点。数据节点均对应一个端口号和 IP（Internet Protocol，网络协议）地址，根据该端口号或 IP 地址可唯一识别一个数据节点。以下各实施例中为方便描述，用阿拉伯数字给数据节点命名，以区分不同的数据节点，如当副本数为 3 个时，一个数据块可能被存放在数据节点 1，数据节点 2 和数据节点 3 上，这个映射关系被保存在名字节点中。数据节点负责存储实际的文件数据块，被客户端和名字节点调用，同时，它会通过心跳定时向名字节点发送所存储的数据块信息。需要说明的是，在 HDFS 中，通常一个节点是一个机器，在本公开各实施例中，将要读取数据或文件的机器均称作客户端，客户端可能是一个数据节点，也可能是一个名字节点，

或者其他个人计算机、智能手机等终端或设备。因此数据块副本所在的数据节点与客户端，以及数据节点之间可能在同一机架，也可能在不同机架，数据节点与客户端也可能是同一台机器。HDFS 中数据读的流程如图 2 所示：

1、客户端向名字节点发起读数据请求，该读数据请求可以是读取文件。

2、名字节点根据数据块索引找到客户端要读取的数据对应的数据块列表，再根据数据块与数据节点的映射，找到每个数据块的各个副本所在的数据节点，并将这些数据节点返回给该客户端。如图 2 所示，名字节点将数据块副本所在的数据节点 1、2 和 3 返回给客户端。

3、客户端从名字节点返回的数据节点中，确定一个为其提供读服务的数据节点，向该数据节点发送读数据块请求。如图 2 所示，客户端向该数据节点 1 发送读数据块请求。

4、数据节点 1 接收到读数块请求后，将存储在其上的数据块副本发送给该客户端。

参照图 3，本公开性能优化方法较佳实施例提供一种性能优化方法，需要说明的是，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。所述性能优化方法包括：

步骤 S1，当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点。

客户端向名字节点发起数据读取请求，名字节点在接收到该数据读取请求后，根据数据块索引获取客户端要读取的数据对应的数据块列表，如客户端要读取的数据被分为三个数据块存储，获取到的数据块列表为数据块 1、数据块 2 和数据块 3，每个数据块分别有三个副本。名字节点根据数据块与数据节点的映射，获取数据块列表中的每个数据块的各个副本所在的数据节点，如数据块 1 的三个副本分别存储在数据节点 1、2 和 3 上。为方便描述，以下各实施例中，按照数据块个数为 1，数据块副本数为 3 进行描述。

步骤 S2，获取所述数据节点对应的预设排序策略，按照所述排序策略对所述数据节点进行排序，得到数据节点列表。

名字节点中预先设置有对数据节点进行排序的排序策略，排序策略可以是按照数据节点与客户端的距离进行排序的策略，也可以是按照数据节点的压力大小进行排序的策略等。当名字节点获取到数据块所在的数据节点后，获取该排序策略，根据该排序策略对数据节点进行排序，得到数据节点列表。可以理解的是，排序后的数据节点组成的列表即为数据节点列表。如对数据节点 1、2、3 进行排序，得到的数据节点列表为，数据节点 1、数据节点 3、数据节点 2。

步骤 S3，将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点。

名字节点将根据该排序策略排序后得到的数据节点列表返回给客户端，客户端在接收

到该数据节点列表后，从该数据节点列表中选择一个数据节点，将其确定为提供读数据块服务的数据节点，并向该数据节点发送读数据块请求。当该数据节点接收到读数据块请求后，会将对应的数据块发送给客户端。需要说明的是，客户端可以选择排在数据节点列表第一位的数据节点，也可以选择排在第二为的数据节点，或者优先选择排在前二位的数据节点，或者随机选择一个数据节点等。

在一个实施例中，为了客户端快速确定读取数据块的数据节点，减少客户端的计算量，步骤 S3 包括：

步骤 a，将所述数据节点列表返回给所述客户端，以供所述客户端将所述数据节点列表中排在最前面的数据节点确定为提供读数据块服务的数据节点。

客户端在接收到该数据节点列表后，选择数据节点列表中排在第一的数据节点，将其确定为提供读数据块服务的数据节点，并向数据节点列表中排在第一的数据节点发送读数据块请求，以获得要读取的数据块。

在一个实施例中，步骤 S1 之前，还包括：

步骤 b，当接收到设置所述排序策略的设置请求后，根据所述设置请求设置所述数据节点对应的排序策略。

名字节点中预设有多种排序策略可供 HDFS 运维人员选择，运维人员也可以在名字节点中设置新的排序策略。即运维人员可以根据具体情况，设置不同的排序策略，以应对不同的 HDFS 运行环境。当名字节点接收到设置排序策略的设置请求后，根据该设置请求设置排序策略，在此之后，当要对数据节点进行排序时，就按照根据该设置请求设置的排序策略对数据节点进行排序。

在一个实施例中，可以通过 HDFS 配置文件对排序策略进行管理。运维人员可以在名字节点或者专门设置的管理节点中修改配置文件，如修改配置文件中数据节点的排序策略，或者设置新的排序策略。配置文件一经修改，将同步到 HDFS 的各个名字节点和各个数据节点中。名字节点可从 HDFS 配置文件中获取排序策略。

本实施例通过当接收到客户端发送的数据读取请求后，获取该数据读取请求对应的数据块所在的数据节点；获取数据节点对应的预设排序策略后，按照该排序策略对该数据节点进行排序，得到数据节点列表；将该数据节点列表返回给该客户端，以供该客户端根据该数据节点列表确定提供读数据块服务的数据节点。由于客户端根据该排序策略排序后所得的数据节点列表确定提供读数据块服务的数据节点，不再总是将与客户端距离最近的数据节点确定为提供读数据块服务的数据节点，从而避免了客户端总是从与其距离最近的数据节点中读取数据块，减小了与客户端距离最近的数据节点的压力，避免了 HDFS 压力分布不均匀，提高了整个 HDFS 的读性能。

在一个实施例中，基于上述第一实施例，本公开性能优化方法第二实施例提供一种性能优化方法。在本实施例中，当名字节点获取到的排序策略为第一排序策略时，步骤 S2 中的按照所述排序策略对所述数据节点进行排序，得到数据节点列表的步骤包括：

步骤 c，获取所述数据节点对应的压力值。

名字节点在获取到数据块所在的数据节点后，首先获取每个数据节点当前的压力值。数据节点当前的压力值可由该数据节点根据其当前的压力数据，以及配置文件中预设的压力值计算方法计算得出，此时，名字节点从数据节点中获取其当前的压力值即可。其中配置文件中预设的压力值计算方法可由运维人员在名字节点或专门设置的管理节点中进行设置，如将各个压力数据直接相加即得到压力值。

数据节点的压力值也可以由名字节点根据从数据节点中获取到的该数据节点当前的压力数据，和配置文件中预设的压力值计算方法计算得出，此时，名字节点需要先从数据节点获取该数据节点的当前的压力数据。

压力数据包括但不限于磁盘 IO 速率（磁盘读写速率），内存使用率，CPU（Central Processing Unit，中央处理器）使用率和网络 IO 速率（网络输入输出速率）。数据节点可通过设置监控进程，实时监控其压力数据，如监控进程监控到该数据节点当前的磁盘 IO 速率为 100 兆每秒，内存使用率为 20%，CPU 使用率为 40%，网络 IO 速率为 50M 每秒。其中监控进程监控到的磁盘 IO 速率和网络 IO 速率也可以是百分比的形式，即将磁盘 IO 速率和网络 IO 速率转换成了百分比，如磁盘 IO 速率为 30%。

需要说明的是，数据节点可以是在检测到配置文件中的数据节点的排序策略为第一排序策略后，才增设监控进程对其压力数据进行监控。

步骤 d，根据所述压力值确定所述数据节点对应的压力，将所述数据节点按照所述压力从小到大的顺序排序，得到数据节点列表。

当获取到每个数据节点对应的压力值后，由于压力值大小表示该数据节点压力的大小，因此可以根据每个数据节点对应的压力值，确定每个数据节点的压力，将数据节点按照压力从小到大的顺序排列，得到数据节点列表，此时，压力最小的数据节点被排在数据节点列表的最前面。此处可包含两种可能的情况，一种是压力值越大代表数据节点的压力越大，一种是压力值越小代表数据节点的压力越小，这两种情况是根据计算数据节点的压力值时所采用的计算方法不同而导致的。无论是上述哪一种情况，都是将压力小的数据节点排在数据列表的前面。如图 4 所示，数据节点的压力值越大，表示该数据节点的压力越小，将压力值最大的数据节点 2 排在数据节点列表的最前面，压力值最小的数据节点 1 排在最后面。

在一个实施例中，步骤 c 包括：

步骤 e，获取所述数据节点的压力数据。

名字节点从数据节点中获取该数据节点当前的压力数据。

步骤 f，根据所述压力数据和预设的压力数据分值标准，得到所述数据节点的压力数据分值。

配置文件中预设有数据节点各个压力数据的压力数据分值标准，压力数据分值标准反映压力数据与压力数据分值之间的映射关系，HDFS 运维人员在设置排序策略时，可以根据具体情况设置压力数据分值标准，例如可将磁盘 IO 速率分值标准设置为表 1 所示的磁盘 IO 速率分值标准，反映磁盘 IO 速率与磁盘 IO 速率分值之间的映射关系，类似地，表 2 为 CPU 使用率分值标准，表 3 为内存使用率分值标准，表 4 为网络 IO 速率分值标准。应当理解的是，压力数据分值标准不限于表中所示的各个分值标准。

磁盘IO速率	磁盘IO速率分值
0-10%	10
11%-20%	9
21%-30%	8
31%-40%	7
41%-50%	6
51%-60%	5
61%-70%	4
71%-80%	3
81%-90%	2
91%-100%	1

表 1

CPU使用率	CPU使用率分值
0-10%	10
11%-20%	9
21%-30%	8
31%-40%	7
41%-50%	6
51%-60%	5
61%-70%	4
71%-80%	3
81%-90%	2
91%-100%	1

表 2

内存使用率	内存使用率分值
0-10%	10
11%-20%	9
21%-30%	8
31%-40%	7
41%-50%	6
51%-60%	5
61%-70%	4
71%-80%	3
81%-90%	2
91%-100%	1

表 3

网络IO速率	网络IO速率分值
0-10%	10
11%-20%	9
21%-30%	8
31%-40%	7
41%-50%	6
51%-60%	5
61%-70%	4
71%-80%	3
81%-90%	2
91%-100%	1

表 4

名字节点从配置文件中获取该压力数据分值标准，将数据节点的各个压力数据与对应的分值标准进行比对，得到各个压力数据分值。如当数据节点当前的磁盘 IO 速率为 20%，CPU 使用率为 30%，内存使用率为 40%，网络 IO 为 20% ，则根据表 1-4 所示的各个分值标准，得到该数据节点的磁盘 IO 速率分值为 9，CPU 使用率分值为 8，内存使用率分值为 7，网络 IO 分值为 9。

步骤 g，根据所述压力数据分值和对应预设的压力数据权重值，计算得到所述数据节点对应的压力值。

配置文件中预设有数据节点各个压力数据的权重值，HDFS 运维人员在设置排序策略时，可以根据具体情况设置各个压力数据的权重值，例如可将磁盘 IO 速率权重值设置为 10，CPU 使用率权重值设置为 5，内存使用率权重值设置为 5，网络 IO 速率权重值设置为 8。

名字节点从配置文件中获取到各个压力数据的权重值后，将各个压力数据分值与对应的压力数据权重值相乘后相加，即得到该数据节点对应的压力值。如根据上述具体例子中的各个压力数据分值以及各个压力数据的权重值，计算得到该数据节点的压力值为 9*10+8*5+7*5+9*10=255。

需要说明的是，在由数据节点计算其当前的压力值时，计算过程也与上述名字节点计算压力值的过程相同。

在本实施例中，通过按照数据节点的压力从小到大的顺序给数据节点进行排序，使得数据节点列表中最前面的数据节点为压力最小的数据节点，从而避免了总是将与客户端最近的数据节点排在最前面，使得该距离最近的数据节点压力过大、HDFS 压力分布不均匀的问题，提高了整个 HDFS 的读性能。

在一个实施例中，基于上述第一或第二实施例，本公开性能优化方法第三实施例提供一种性能优化方法。在本实施例中，当名字节点获取到的排序策略为第二排序策略时，步骤 S2 中的按照所述排序策略对所述数据节点进行排序，得到数据节点列表的步骤包括：

步骤 h，将所述数据节点按照与所述客户端的距离由近到远的顺序排序，得到预处理数据节点列表。

当数据节点与客户端是同一台机器时，该数据节点与客户端的距离最近，当数据节点与客户端在同一机架的不同机器上时，距离较远，当数据节点与客户端在不同机架时，距离更远。数据块所在的各个数据节点与客户端的距离可能相同，也可能不相同。名字节点在获取到数据块所在的数据节点后，按照数据节点与客户端的距离由近及远的顺序给数据节点进行排序，两个与客户端距离相同的数据节点，可任一排序，得到预处理数据节点列表。如名字节点将数据节点 1、2、3 按照与客户端距离由近到远的顺序排序，得到图 5 所示的预处理数据节点列表。

步骤 i，获取所述数据节点对应的压力值，检测所述数据节点对应的压力值是否满足预设条件。

名字节点获取数据节点对应的压力值的过程与第二实施例中的步骤 a 所述过程相同，在此不再详细赘述。当名字节点获取到数据节点对应的压力值后，遍历预处理数据节点列表，检测每个数据节点的压力值是否满足预设条件。其中，预设条件可根据具体情况进行设置，如当数据节点的压力值越大表示其压力越大时，可以设置为当数据节点的压力值大于预设压力值时即为满足预设条件；当数据节点的压力值越大表示其压力越小时，可以设置为当数据节点的压力值小于预设压力值时即为满足预设条件。预设压力值可根据具体情况设置。

步骤 j，当检测到所述数据节点对应的压力值满足预设条件时，将所述压力值满足预设条件的数据节点移动到所述预处理数据节点列表的末端，得到处理后的数据节点列表。

当检测到数据节点对应的压力值满足预设条件时，将该压力值满足预设条件数据节点移动到预处理数据节点列表的末端。当遍历完所有数据节点后，得到处理后的，即最终的数据节点列表。此时，排在最前面的数据节点压力相对较小，与客户端的距离相对较近。如图 6 所示，是将图 5 所示的预处理数据节点列表中，压力值满足预设条件的数据节点 1 移动到末端后，得到的数据节点列表。

需要说明的是，如图 5 所示，若名字节点将此按照数据节点与客户端的距离排过序的预处理数据节点列表返回给客户端，并且总是按照数据节点与客户端的距离排序，数据节点 1 将会经常排在最前面，而客户端优先选择排在最前面的数据节点读取数据块时，排在前面的数据节点 1 由于经常被客户端访问，会变得压力过大，而后面的数据节点 2 和 3 则压力偏小，从而可能造成 HDFS 压力分布不均。因此，在本实施例中，通过将数据节点先按照与客户端的距离由近到远的顺序排序，再将数据节点满足预设条件，即压力超过预设压力的数据节点排到所有数据节点之后，使得排在数据节点列表最前面的数据节点为压力相对较小，与客户端距离相对较近的数据节点，从而避免了总是将与客户端最近的数据节点排在最前面，使得该距离最近的数据节点压力过大的问题。

在一个实施例中，基于上述第一、第二或第三实施例，本公开性能优化方法第四实施例提供一种性能优化方法。在本实施例中，当名字节点获取到的排序策略为第三排序策略时，步骤 S2 中的按照所述排序策略给所述数据节点排序，得到数据节点列表的步骤包括：

步骤 k，将所述数据节点进行随机排序，得到数据节点列表。

名字节点在获取到数据块所在的数据节点后，将数据节点进行随机排序，得到数据节点列表。随机排序的方法可以是任何能够将数据进行随机排序的方法。当数据块所在的各个数据节点在同一机架中时，也即各个数据节点与客户端的距离相同时，不需要考虑数据节点与客户端的距离，此时按照随机排序的策略给数据节点进行排序，使得数据节点被客户端访问的几率相同，因此避免了由于某一个数据节点压力过大，导致的 HDFS 压力分布不均问题。

此外，参照图 7，本公开还提供一种性能优化装置，所述性能优化装置包括：获取模块 10，用于当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点；获取所述数据节点对应的预设排序策略；排序模块 20，用于按照所述排序策略对所述数据节点进行排序，得到数据节点列表；数据返回模块 30，用于将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点。

在一个实施例中，当所述排序策略为第一排序策略时，所述排序模块 20 包括：第一获取单元，用于获取所述数据节点对应的压力值；第一排序单元，用于根据所述压力值确定所述数据节点对应的压力，将所述数据节点按照所述压力从小到大的顺序排序，得到数据节点列表。

在一个实施例中，所述第一获取单元还包括：获取子单元，用于获取所述数据节点的压力数据；计算子单元，用于根据所述压力数据和预设的压力数据分值标准，得到所述数

据节点的压力数据分值；还用于根据所述压力数据分值和对应预设的压力数据权重值，计算得到所述数据节点对应的压力值。

在一个实施例中，当所述排序策略为第二排序策略时，所述排序模块 20 还包括：第二排序单元，用于将所述数据节点按照与所述客户端的距离由近到远的顺序排序，得到预处理数据节点列表；第二获取单元，用于获取所述数据节点对应的压力值；检测单元，用于检测所述数据节点对应的压力值是否满足预设条件；所述第二排序单元还用于当检测到所述数据节点对应的压力值满足预设条件时，将所述压力值满足预设条件的数据节点移动到所述预处理数据节点列表的末端，得到处理后的数据节点列表。

在一个实施例中，当所述排序策略为第三排序策略时，所述排序模块 20 还包括：第三排序单元，用于将所述数据节点进行随机排序，得到数据节点列表。

在一个实施例中，所述数据返回模块 30 还用于将所述数据节点列表返回给所述客户端，以供所述客户端将所述数据节点列表中排在最前面的数据节点确定为提供读数据块服务的数据节点。

在一个实施例中，所述性能优化装置还包括：设置模块，用于当接收到设置所述排序策略的设置请求后，根据所述设置请求设置所述数据节点对应的排序策略。

需要说明的是，性能优化装置的各个实施例与上述性能优化方法的各实施例基本相同，在此不再详细赘述。

此外，本公开实施例还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有性能优化程序，所述性能优化程序被处理器执行时实现如上所述性能优化方法的步骤。

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者系统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

上述本公开实施例序号仅仅为了描述，不代表实施例的优劣。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本公开的技术方案本质上或者说对一些情况做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，

计算机，服务器，空调器，或者网络设备等)执行本公开各个实施例所述的方法。

避免了客户端总是从与其距离最近的数据节点中读取数据块，减小了与客户端距离最近的数据节点的压力，避免了 HDFS 压力分布不均匀，提高了整个 HDFS 的读性能。

本公开通过当接收到客户端发送的数据读取请求后，获取该数据读取请求对应的数据块所在的数据节点，以及获取数据节点对应的预设排序策略，按照该排序策略对该数据节点进行排序，得到数据节点列表；将该数据节点列表返回给该客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点。由于客户端根据该排序策略排序后所得的数据节点列表确定提供读数据块服务的数据节点，不再总是将与客户端距离最近的数据节点确定为提供读数据块服务的数据节点，从而避免了客户端总是从与其距离最近的数据节点中读取数据块，减小了与客户端距离最近的数据节点的压力，避免了 HDFS 压力分布不均匀，提高了整个 HDFS 的读性能。

以上仅为本公开的优选实施例，并非因此限制本公开的专利范围，凡是利用本公开说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本公开的专利保护范围内。

权利要求书

1、一种性能优化方法，其中，所述性能优化方法包括以下步骤：

当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点；

获取所述数据节点对应的预设排序策略，按照所述排序策略对所述数据节点进行排序，得到数据节点列表；

将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点。

2、如权利要求 1 所述的性能优化方法，其中，当所述排序策略为第一排序策略时，所述按照所述排序策略对所述数据节点进行排序，得到数据节点列表的步骤包括：

获取所述数据节点对应的压力值；

根据所述压力值确定所述数据节点对应的压力，将所述数据节点按照所述压力从小到大的顺序排序，得到数据节点列表。

3、如权利要求 2 所述的性能优化方法，其中，所述获取所述数据节点对应的压力值的步骤包括：

获取所述数据节点的压力数据；

根据所述压力数据和预设的压力数据分值标准，得到所述数据节点的压力数据分值；

根据所述压力数据分值和对应预设的压力数据权重值，计算得到所述数据节点对应的压力值。

4、如权利要求 1 所述的性能优化方法，其中，当所述排序策略为第二排序策略时，所述按照所述排序策略对所述数据节点进行排序，得到数据节点列表的步骤包括：

将所述数据节点按照与所述客户端的距离由近到远的顺序排序，得到预处理数据节点列表；

获取所述数据节点对应的压力值，检测所述数据节点对应的压力值是否满足预设条件；

当检测到所述数据节点对应的压力值满足预设条件时，将所述压力值满足预设条件的数据节点移动到所述预处理数据节点列表的末端，得到处理后的数据节点列表。

5、如权利要求 1 所述的性能优化方法，其中，当所述排序策略为第三排序策略时，

所述按照所述排序策略对所述数据节点进行排序，得到数据节点列表的步骤包括：

将所述数据节点进行随机排序，得到数据节点列表。

6、如权利要求 1 所述的性能优化方法，其中，所述将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点的步骤包括：

将所述数据节点列表返回给所述客户端，以供所述客户端将所述数据节点列表中排在最前面的数据节点确定为提供读数据块服务的数据节点。

7、如权利要求 1 至 6 任一项所述的性能优化方法，其中，所述当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点的步骤之前，还包括：

当接收到设置所述排序策略的设置请求后，根据所述设置请求设置所述数据节点对应的排序策略。

8、一种性能优化装置，其中，所述性能优化装置包括：

获取模块，用于当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点；获取所述数据节点对应的预设排序策略；

排序模块，用于按照所述排序策略对所述数据节点进行排序，得到数据节点列表；

数据返回模块，用于将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点。

9、一种性能优化设备，其中，所述性能优化设备包括存储器、处理器和存储在所述存储器上并可在所述处理器上运行的性能优化程序，所述性能优化程序被所述处理器执行时实现如权利要求 1 至 7 中任一项所述的性能优化方法的步骤。

10、一种计算机可读存储介质，其中，所述计算机可读存储介质上存储有性能优化程序，所述性能优化程序被处理器执行时实现如权利要求 1 至 7 中任一项所述的性能优化方法的步骤。

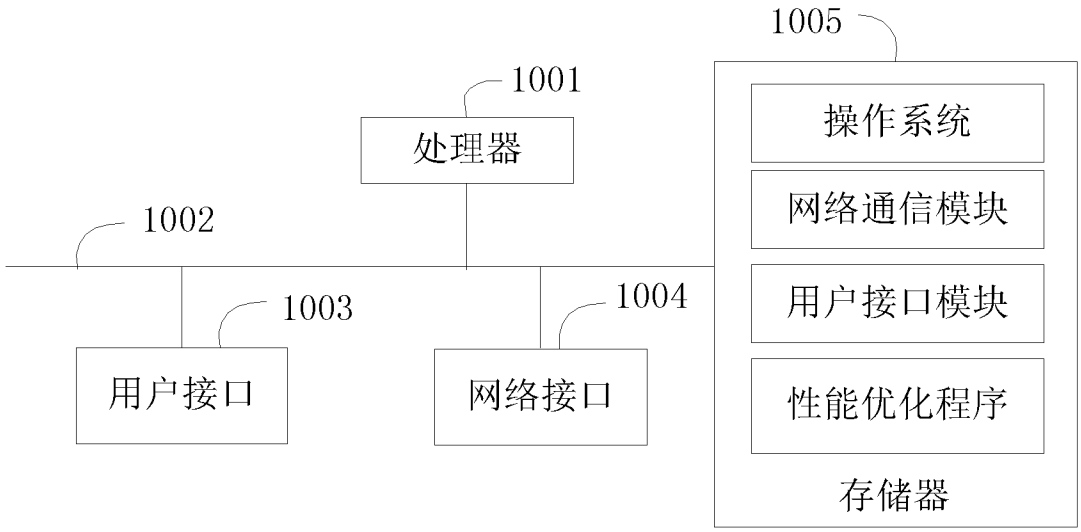
1005
1001
处理器
操作系统
网络通信模块
用户接口模块
性能优化程序
1002
1003
1004
用户接口
网络接口
存储器

图 1

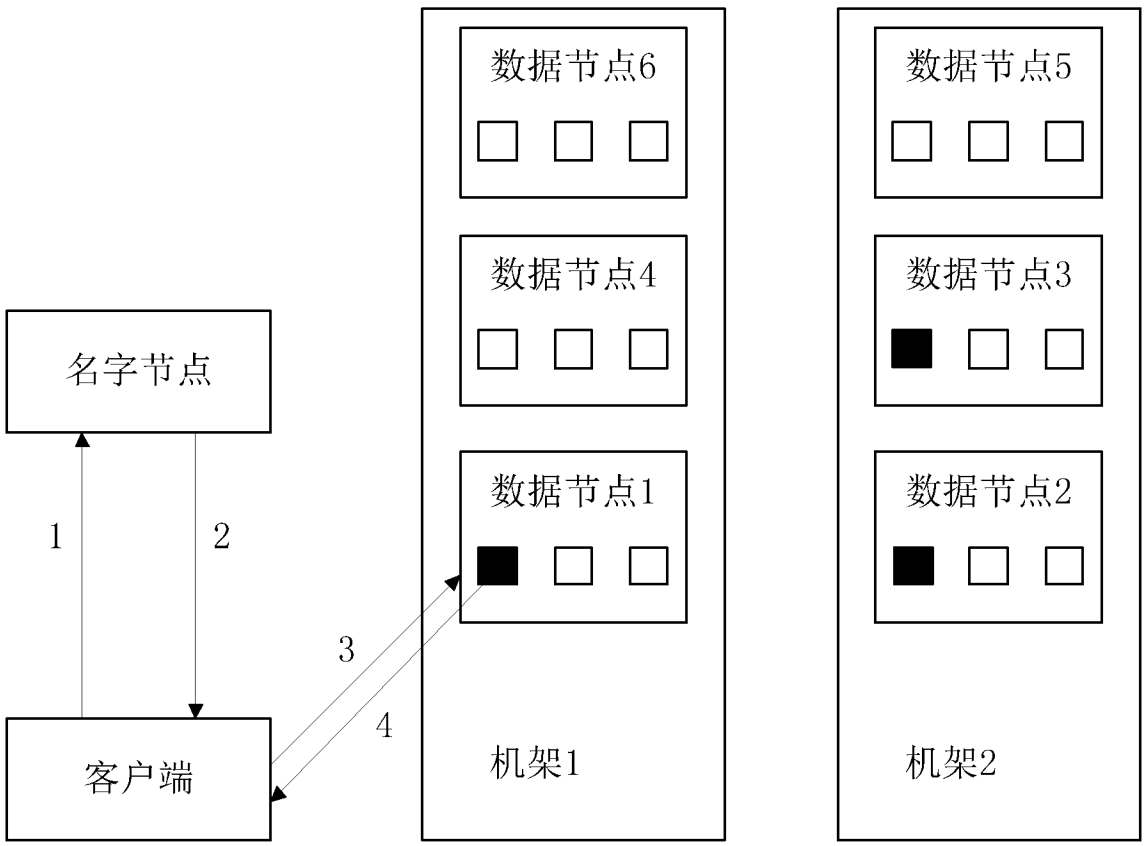
数据节点6
数据节点5
数据节点4
数据节点3
名字节点
数据节点1
数据节点2
1
2
3
4
客户端
机架1
机架2

图 2

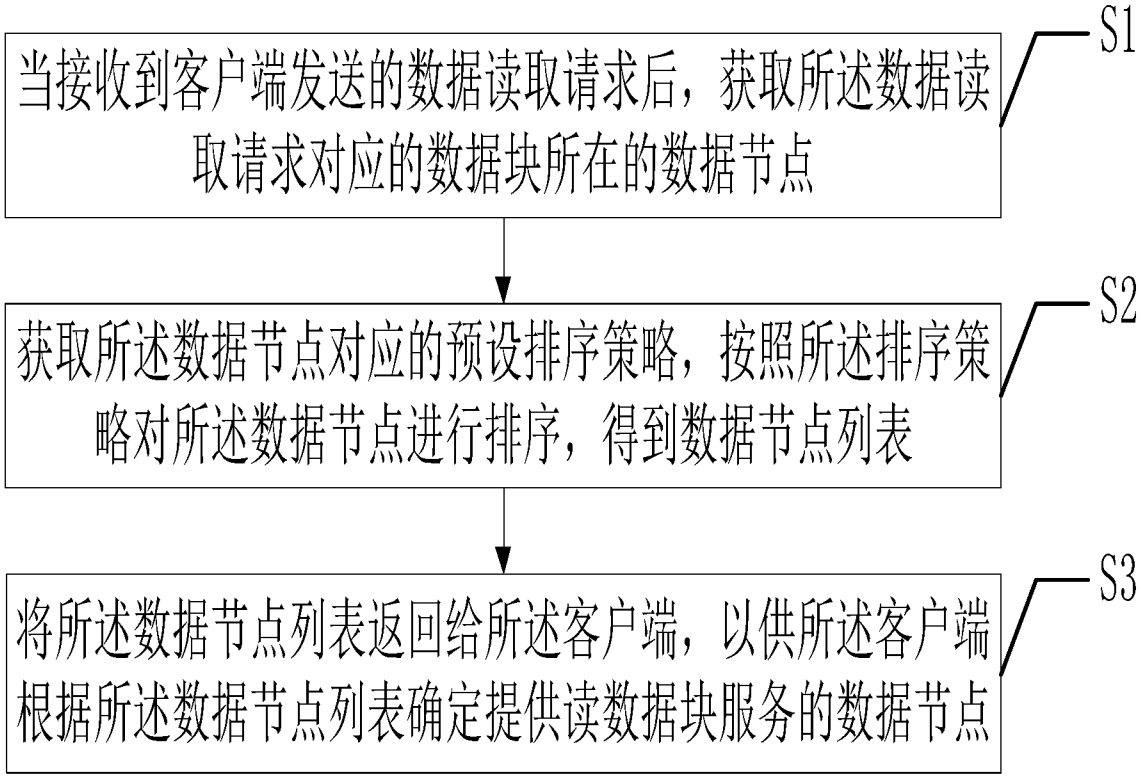
当接收到客户端发送的数据读取请求后，获取所述数据读取请求对应的数据块所在的数据节点
S1
获取所述数据节点对应的预设排序策略，按照所述排序策略对所述数据节点进行排序，得到数据节点列表
S2
将所述数据节点列表返回给所述客户端，以供所述客户端根据所述数据节点列表确定提供读数据块服务的数据节点
S3

图 3

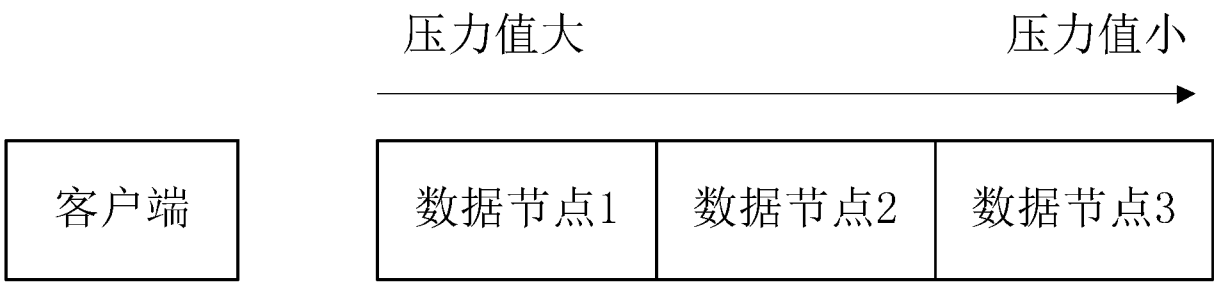
压力值大
压力值小
客户端
数据节点1
数据节点2
数据节点3

图 4

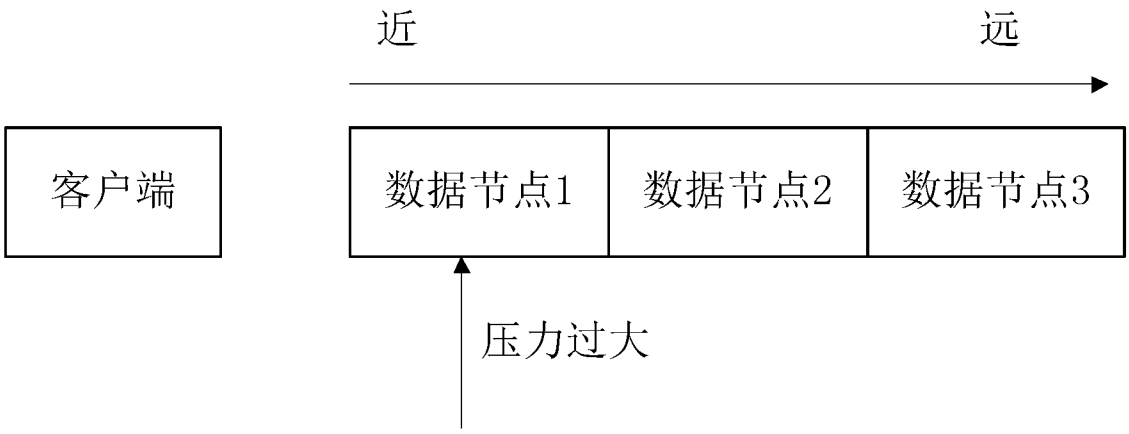
近
远
客户端
数据节点1
数据节点2
数据节点3
压力过大

图 5

图 6

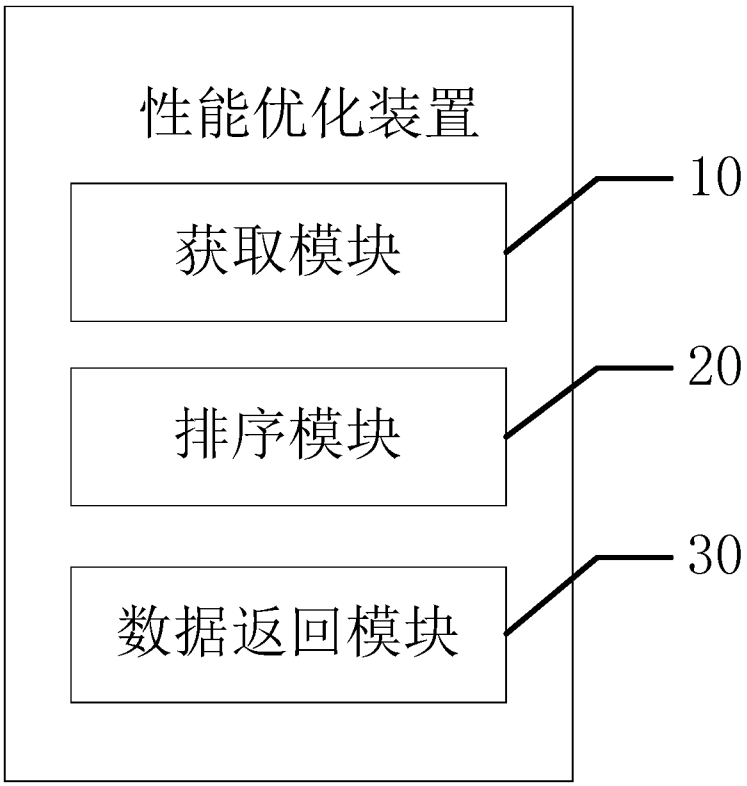

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/116024

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/182(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE, GOOGLE: 数据, 访问, 读取, 节点, 负载, 负载均衡, 压力, 距离, 排序, data, node, access, read, distance, pressure, load, balance, sequence

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104156381 A (SHENZHEN INSTITUTE OF INFORMATION TECHNOLOGY) 19 November 2014 (2014-11-19) description, paragraphs [0002], [0029]-[0046], [0053] and [0078]	1-10
A	CN 105550362 A (ZHEJIANG DAHUA TECHNOLOGY CO., LTD.) 04 May 2016 (2016-05-04) entire document	1-10
A	CN 108009260 A (XI'AN JIAOTONG UNIVERSITY) 08 May 2018 (2018-05-08) entire document	1-10
A	US 2017373977 A1 (PAYPAL, INC.) 28 December 2017 (2017-12-28) entire document	1-10
A	US 2018285167 A1 (OCIENT, INC.) 04 October 2018 (2018-10-04) entire document	1-10

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 December 2019	**05 February 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/116024

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 104156381 A	19 November 2014	None	
CN 105550362 A	04 May 2016	None	
CN 108009260 A	08 May 2018	None	
US 2017373977 A1	28 December 2017	WO 2018005550 A1	04 January 2018
		EP 3475710 A1	01 May 2019
		CN 109642923 A	16 April 2019
US 2018285167 A1	04 October 2018	WO 2018187229 A1	11 October 2018
		US 2018285414 A1	04 October 2018

国际检索报告

国际申请号
PCT/CN2019/116024

A. 主题的分类

G06F 16/182(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT,WPI,EPODOC,CNKI,IEEE,GOOGLE：数据，访问，读取，节点，负载，负载均衡，压力，距离，排序，data，node，access，read，distance，pressure，load，balance，sequence

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 104156381 A (深圳信息职业技术学院) 2014年 11月 19日 (2014-11-19) 说明书第[0002]，[0029]-[0046]，[0053]，[0078]段	1-10
A	CN 105550362 A (浙江大华技术股份有限公司) 2016年 5月 4日 (2016-05-04) 全文	1-10
A	CN 108009260 A (西安交通大学) 2018年 5月 8日 (2018-05-08) 全文	1-10
A	US 2017373977 A1 (PAYPAL, INC.) 2017年 12月 28日 (2017-12-28) 全文	1-10
A	US 2018285167 A1 (OCIENT, INC.) 2018年 10月 4日 (2018-10-04) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2019年 12月 25日	2020年 2月 5日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 孔昕 电话号码 86-(10)-53961371

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/116024

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	104156381	A	2014年 11月 19日	无			
CN	105550362	A	2016年 5月 4日	无			
CN	108009260	A	2018年 5月 8日	无			
US	2017373977	A1	2017年 12月 28日	WO	2018005550	A1	2018年 1月 4日
				EP	3475710	A1	2019年 5月 1日
				CN	109642923	A	2019年 4月 16日
US	2018285167	A1	2018年 10月 4日	WO	2018187229	A1	2018年 10月 11日
				US	2018285414	A1	2018年 10月 4日