(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
2017 年 9 月 8 日 (08.09.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/148297 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2017/074177

(22) 国际申请日: 2017 年 2 月 20 日 (20.02.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610118167.7 2016 年 3 月 2 日 (02.03.2016) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人; 及

(71) 申请人 (仅对美国): 吴炜 (WU, Wei) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (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, 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, 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)。

本国际公布:
— 包括国际检索报告(条约第 21 条(3))。

(54) Title: METHOD AND DEVICE FOR JOINING TABLES

(54) 发明名称 : 数据表连接方法及装置

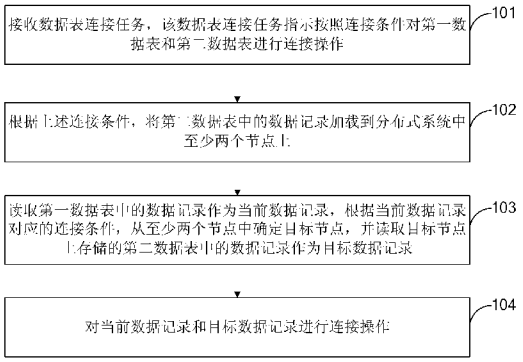

图 1

101 RECEIVE A TABLE JOINING TASK INSTRUCTING TO JOIN ACCORDING TO A JOINING REQUIREMENT A FIRST TABLE AND A SECOND TABLE
102 LOAD, ACCORDING TO THE JOINING REQUIREMENT, A DATA ENTRY IN THE SECOND TABLE TO TWO OR MORE NODES IN A DISTRIBUTED SYSTEM
103 READ A DATA ENTRY IN THE FIRST TABLE AS A CURRENT DATA ENTRY; DETERMINE, FROM THE TWO OR MORE NODES, ACCORDING TO A JOINING REQUIREMENT CORRESPONDING TO THE CURRENT DATA ENTRY, A TARGET NODE; AND READ THE DATA ENTRY OF THE SECOND TABLE AND STORED IN THE TARGET NODE AS A TARGET DATA ENTRY
104 PERFORM A JOINING OPERATION ON THE CURRENT DATA ENTRY AND THE TARGET DATA ENTRY

(57) Abstract: A method and device for joining tables. The method comprises: receiving a table joining task instructing to join according to a joining requirement a first table and a second table (101); loading, according to the joining requirement, a data entry in the second table to two or more nodes in a distributed system (102); reading a data entry in the first table as a current data entry, determining, from the two or more nodes, according to a joining requirement corresponding to the current data entry, a target node, and reading the data entry of the second table and stored in the target node as a target data entry (103); and performing a joining operation on the current data entry and the target data entry (104). The method can decrease computation resources consumed during performing the joining operation on tables.

(57) 摘要:

[见续页]

WO 2017/148297 A1

一种数据表连接方法及装置。方法包括：接收数据表连接任务，数据表连接任务指示按照连接条件对第一数据表和第二数据表进行连接操作（101）；根据连接条件，将第二数据表中的数据记录加载到分布式系统中至少两个节点上（102）；读取第一数据表中的数据记录作为当前数据记录，根据当前数据记录对应的连接条件，从至少两个节点中确定目标节点，并读取目标节点上存储的第二数据表中的数据记录作为目标数据记录（103）；对当前数据记录和目标数据记录进行连接操作（104）。该方法可以降低数据表连接操作消耗的计算资源。

数据表连接方法及装置

【技术领域】

本申请涉及数据库技术领域，尤其涉及一种数据表连接方法及装置。

【背景技术】

随着互联网的发展，数据呈现爆发式增长，数据结构也开始多元化，数据含有的信息量越来越多，数据仓库在这样的背景下发挥着巨大的作用。由于大数据时代的降临，数据仓库转成为分布式架构，以满足爆发式增长的计算及存储的需求。分布式数据仓库一般使用列式存储，并以文件的形式保存数据，因此，采用分布式数据仓库可提高大数据的存储及计算性能。

在分布式数据仓库的查询过程中，经常需要进行数据表之间的连接(Join)计算。现有技术在处理数据表之间的 Join 计算时，一般都是先将所有待 Join 的数据表通过 MapReduce 的方式做洗牌（shuffle）排序，然后在 Reducer 端对已经排过序的数据表进行归并操作。shuffle 排序实际上是指将 Map 端各个待 Join 的数据表按照 Join 条件进行分区并分配到不同 Reducer 端的过程。

在典型的“星型”Join 场景下，假设待 Join 数据表包括一个主表和 n 个辅表，主表包含 M 条数据记录，那么在对主表和 n 个辅表进行 Join 计算时，shuffle 排序需要处理的总数据量包括 shuffle 主表需要处理的数据量即 M*n 和 shuffle n 个辅表需要处理的数据量，这会消耗很多计算资源。

【发明内容】

本申请的多个方面提供一种数据表连接方法及装置，用以降低数据表连接操作消耗的计算资源。

本申请的一方面，提供一种数据表连接方法，包括:

接收数据表连接任务，所述数据表连接任务指示按照连接条件对第一

数据表和第二数据表进行连接操作;

根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上;

读取所述第一数据表中的数据记录作为当前数据记录，根据所述当前数据记录对应的连接条件，从所述至少两个节点中确定目标节点，并读取所述目标节点上存储的所述第二数据表中的数据记录作为目标数据记录;

对所述当前数据记录和所述目标数据记录进行连接操作。

本申请的另一方面，提供一种数据表连接装置，包括:

接收模块，用于接收数据表连接任务，所述数据表连接任务指示按照连接条件对第一数据表和第二数据表进行连接操作;

加载模块，用于根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上;

读取模块，用于读取所述第一数据表中的数据记录作为当前数据记录，根据所述当前数据记录对应的连接条件，从所述至少两个节点中确定目标节点，并读取所述目标节点上存储的所述第二数据表中的数据记录作为目标数据记录;

连接模块，用于对所述当前数据记录和所述目标数据记录进行连接操作。

在本申请中，在处理数据表连接任务时，首先根据其中的连接条件，将第二数据表中的数据记录加载到分布式系统中至少两个节点上，之后，可以直接读取第一数据表中的数据记录，并根据所读取的第一数据表中的数据记录对应的连接条件，从相应节点上读取所需的第二数据表中的数据记录，之后对读取到的两个数据表中的数据记录进行连接操作。由此可见，本申请只需将第二数据表按照连接条件分布到不同节点上，不需要将第一数据表分布到不同节点上，减少了 shuffle 排序需要处理的数据量，有利于降低连接操作所消耗的计算资源。

【附图说明】

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 为本申请一实施例提供的数据表连接方法的流程示意图；

图 2 为本申请另一实施例提供的分布式系统的架构示意图；

图 3 为本申请又一实施例提供的数据表连接装置的结构示意图；

图 4 为本申请又一实施例提供的数据表连接装置的结构示意图。

【具体实施方式】

为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

在分布式数据仓库的查询过程中，经常需要进行数据表之间的连接(Join)计算。现有技术在处理数据表之间的 Join 操作时，由于待 Join 的数据表比较大，所以一般都是先将所有待 Join 的数据表通过 MapReduce 的方式做 shuffle 排序，然后在 Reducer 端对已经排过序的数据表进行归并操作。shuffle 排序实际上是指将 Map 端各个待 Join 的数据表按照 Join 条件进行分区并分配到不同 Reducer 端的过程。由于需要对所有待 Join 的数据表进行 shuffle 排序，消耗的计算资源较多。

针对上述技术问题，本申请提供一种解决方案，即通过将第二数据表分布存储到多个节点上，成为一个分布式的缓存，对第一数据表处理时，通过网络获取远程节点上存储的第二数据表中的数据记录，从而进行分布式的哈希映射连接（Hash map Join），使得无需对主表进行 shuffle 排序，

这样可以节约对第一数据表进行 shuffle 排序消耗的计算资源。

图 1 为本申请一实施例提供的数据表连接方法的流程示意图。如图 1 所示，该方法包括:

101、接收数据表连接任务，该数据表连接任务指示按照连接条件对第一数据表和第二数据表进行连接操作。

102、根据上述连接条件，将第二数据表中的数据记录加载到分布式系统中至少两个节点上。

103、读取第一数据表中的数据记录作为当前数据记录，根据当前数据记录对应的连接条件，从至少两个节点中确定目标节点，并读取目标节点上存储的第二数据表中的数据记录作为目标数据记录。

104、对当前数据记录和目标数据记录进行连接操作。

本实施例提供一种数据表连接方法，可由数据表连接装置来执行，用以进行数据表之间的 Join 操作，同时尽量降低所消耗的计算资源。本实施例提供的方法适用于分布式系统，这里的分布式系统中的不同机器可以分别作为一个节点。本实施例并不限制分布式系统的实现架构，例如可以是但不限于 MapReduce 架构。

当需要进行数据表之间的 Join 操作时，可以向数据表连接装置发送数据表连接任务；数据表连接装置接收数据表连接任务。该数据表连接任务指示按照连接条件对第一数据表和第二数据表进行 Join 处理。这里的第一数据表和第二数据表实际上是待连接的数据表。

在具体实现上，该数据表连接任务携带有连接条件、第一数据表的标识、第二数据表的标识、第一数据表的存储位置、以及第二数据表的存储位置等信息。其中，数据表连接装置可以对数据表连接任务进行解析，获取连接条件、第一数据表的标识、第二数据表的标识、第一数据表的存储位置、以及第二数据表的存储位置等信息，并根据第一数据表的标识和第二数据表的标识确定需要进行 Join 操作的数据表，另外，可以根据第一数据表的存储位置和第二数据表的存储位置读取第一数据表和第二数据表。

在一种实际应用中，第一数据表可以作为主表，第二数据表可以作为

辅表实现。其中，辅表的数量可以是一个或多个。

数据表连接装置接收到数据表连接任务之后，可以获知需要按照连接条件对第一数据表和第二数据表进行 Join 操作。之后，在执行 Join 操作之前，首先根据连接条件，将第二数据表中的数据记录加载到分布式系统中至少两个节点上，实现分布式存储。

优选的，位于至少两个节点中每个节点上的第二数据表中的数据记录，其数据量小于单个节点的内存限制，也就是说，分布到至少两个节点中每个节点上的第二数据表中的数据记录，均能够全部放到相应节点的存储空间（优选为内存）中。

在一可选实施方式中，上述连接条件包括连接所需的至少一个目标键，这里的目标键实际上就是键值对（key-value）中的键（key）。基于此，数据表连接装置具体可以分别对至少一个目标键中的各目标键进行哈希运算，以获取各目标键的哈希值；根据各目标键的哈希值和上述用于存储第二数据表中的数据记录的至少两个节点的数量，确定各目标键对应的节点；将第二数据表中对应于各目标键的数据记录分别加载到各目标键对应的节点上。

进一步，数据表连接装置可以利用各目标键的哈希值对上述用于存储第二数据表中的数据记录的至少两个节点的数量取模，根据取模结果确定各目标键对应的节点。具体的，可以将取模结果代表的节点作为目标键对应的节点。或者，

数据表连接装置可以根据上述用于存储第二数据表中的数据记录的至少两个节点的数量和目标键的数量，将各目标键均分到各节点上，在均分过程中，可以根据各目标键的哈希值，将哈希值相近的目标键分到相同节点。这里的哈希值相近可以是指哈希值之差小于预设门限，但不限于此。

进一步，在上述加载第二数据表中的数据记录到至少两个节点上的过程中，具体可以将第二数据表中的数据记录加载到至少两个节点的内存中。第二数据表中的数据记录存储在节点的内存中，可以随时读取，读取速度较快，有利于提高 Join 操作的效率。

值得说明的是，优选的，可以将第二数据表中的数据记录加载到上述

至少两个节点的内存中，但并不限于内存，还可以是节点的固态硬盘（Solid State Drives，SSD）或者其他存储介质中。

在一可选实施方式中，在根据连接条件，将第二数据表中的数据记录加载到分布式系统中的至少两个节点上之前，可以判断第二数据表的数据量是否大于单个节点的内存限制；若判断结果为是，即第二数据表的数据量大于单个节点的内存限制，这意味着第二数据表中的数据记录不能全部放在单个节点的内存中，因此可以根据连接条件，将第二数据表中的数据记录加载到至少两个节点中，使得分布到每个节点上的第二数据表中的数据记录均能全部放到相应节点的内存中，实现分布式存储。简单的说，分布到每个节点上的第二数据表中的数据记录，其数据量小于单个节点的内存限制。

若上述判断结果为否，即第二数据表的数据量小于或等于单个节点的内存限制，这意味着第二数据表中的数据记录可以全部放在单个节点的内存中，较为优选的，可以将第二数据表的数据记录全部放到单个节点的内存中，从而节省对第二数据表的数据记录进行 shuffle 排序，节约计算资源。

上述将第二数据表中的数据记录加载到分布式系统中至少两个节点上，这相当于将第二数据表变成了多个小表，每个小表可以全部在相应节点的内存中放得下，形成分布式的 KV 存储，使得可以做分布式的 Hash map Join，而不需要做排序合并连接（Sort Merge Join）。做分布式的 Hash map Join，不需要对第一数据中的数据记录进行排序，可以直接读取第一数据表中的数据记录，并根据所读取的第一数据表中的数据记录对应的连接条件，从相应节点上读取所需的第二数据表中的数据记录，之后对读取到的两个数据表中的数据记录进行 Join 操作。

其中，本实施例中分布式的 Hash map Join 与现有 Hash map Join 的区别在于：在对第一数据表处理时，不是在本地内存中查找第二数据表中的数据记录，而是通过网络获取远程节点上存储的第二数据表中的数据记录。

具体的，在将第二数据表中的数据记录加载到分布式系统中至少两个节点之后，数据表连接装置可以到第一数据表的存储位置读取第一数据表

中的数据记录，将读取到的数据记录作为当前数据记录，根据当前数据记录对应的连接条件，从上述至少两个节点中确定目标节点，这里的目标节点是指与当前数据记录进行Join操作所需的第二数据表中的数据记录所在的节点，然后读取目标节点上存储的第二数据表中的数据记录作为目标数据记录，这里的目标数据记录是指与当前数据记录进行Join操作所需的第二数据表中的数据记录。

在读取到当前数据记录以及与当前数据记录进行 Join 操作所需的目标数据记录之后，对当前数据记录与目标数据记录进行Join操作。由于如何对当前数据记录与目标数据记录进行Join操作不是本申请的重点，在此不再详述，可参考现有技术中有关Join操作的处理流程。

在一可选实施方式中，考虑到数据表连接装置的本地缓存中可能会存在与当前数据记录进行Join操作所需的目标数据记录，基于此，在根据当前数据记录对应的连接条件，从至少两个节点中确定目标节点，并读取目标节点上存储的第二数据表中的数据记录作为目标数据记录之前，可以根据当前数据记录对应的连接条件，判断本地缓存中是否存在目标数据记录，若判断结果为否，则执行根据当前数据记录对应的连接条件，从至少两个节点中确定目标节点，并读取目标节点上存储的第二数据表中的数据记录作为目标数据记录的操作；若判断结果为是，则可以从本地缓存中获取目标数据记录，这样可以更加快速的获取目标数据记录，节约获取目标数据记录所消耗的网络资源，提高Join操作的效率。

进一步，上述当前数据记录对应的连接条件可以是目标键，则一种根据当前数据记录对应的连接条件，从至少两个节点中确定目标节点的实施方式包括:

对当前数据记录对应的目标键进行哈希运算，以获得当前数据记录对应的目标键的哈希值；根据当前数据记录对应的目标键的哈希值和上述至少两个节点的数量，确定当前数据记录对应的目标键对应的节点作为目标节点。

更进一步，在通过目标键确定从某个节点上取目标数据的过程中，若目标键有多个，则可以进行批量操作，这样可以充分发挥分布式系统的优

势，提高处理性能。

由上述分析可见，本实施例在处理数据表连接任务时，首先根据其中的连接条件，将第二数据表中的数据记录加载到至少两个节点上，这相当于变成了一个分布式的 KV 存储（即会有分布式的哈希表），这样不需要做 Sort Merge Join，使得可以做分布式的 Hash map Join，即不需要对第一数据中的数据记录进行排序，而是可以直接读取第一数据表中的数据记录，并根据所读取的第一数据表中的数据记录对应的连接条件，从相应节点上读取所需的第二数据表中的数据记录，之后对读取到的两个数据表中的数据记录进行 Join 操作。由此可见，本实施例只需将第二数据表按照连接条件分布到不同节点上，不需要将第一数据表分布到不同节点上，减少了 shuffle 排序需要处理的数据量，有利于降低连接操作所消耗的计算资源。

下面通过对比 Sort Merge Join 与分布式的 Hash map Join 消耗的计算资源，以说明本申请技术方案带来的优势。

假设主表是 A，其数据大小是 100T，假设辅表有 2 个分别是 B 和 C，辅表 B 的数据大小为 10G，辅表 C 的数据大小为 100G。

若采用现有 Sort Merge Join，其 shuffle 排序阶段需要将主表 A 和辅表 B 进行一次排序处理，还需要将主表 A 和辅表 C 进行一次排序处理，每次排序处理包括通过网络 IO 读数据表以及通过 CPU 进行排序，所以每次排序处理的资源消耗包括：排序所占 CPU 和读表所占的网络 IO。为便于描述，通过处理的数据量表示资源消耗，在这里，考虑到 CPU 排序处理的数据量也就是通过网络 IO 读取的数据量，故以一份数据量来表示每次排序处理的资源消耗，则 shuffle 排序阶段需要总资源消耗为：（100T+10G）+（100T+100G）=2 * 100T + 10G + 100G。

若采用本申请分布式的 Hash map Join，其 shuffle 排序阶段需要将辅表 B 分布到不同节点上，还需要将辅表 C 分布到不同节点上，每次将某个表分布到不同节点上包括通过网络 IO 读数据表以及通过 CPU 进行排序，所以将某个表分布到不同节点上的资源消耗同样包括：排序所占 CPU 和读表所占的网络 IO。为便于描述，通过处理的数据量表示资源消耗，在这里，

考虑到 CPU 排序处理的数据量也就是通过网络 IO 读取的数据量，故以一份数据量来表示每次排序处理的资源消耗，则 shuffle 排序阶段需要总资源消耗为：10G + 100G。

由上述可见，由于本申请技术方案只需将第二数据表按照连接条件分布到不同节点上，不需要将第一数据表分布到不同节点上，减少了 shuffle 排序需要处理的数据量，有利于降低连接操作所消耗的计算资源。

图 2 为本申请另一实施例提供的分布式系统的架构示意图。如图 2 所示，该分布式系统包括：控制节点 21、调度节点 22、以及至少两个计算节点 23。进一步，如图 2 所示，计算节点 23 至少包括缓存模块和处理模块。

值得说明的是，图 2 所示分布式系统仅为一种示例，并不限于此，例如可以将图 2 中的调度节点 22 省略从而获得一种更为简单的分布式系统。

下面将基于图 2 所示分布式系统，对本申请技术方案进行详细说明。

控制节点 21 负责接收数据表连接任务，根据数据表连接任务获知需要按照连接条件对第一数据表和第二数据表进行连接操作。

控制节点 21 可以根据数据表连接任务向调度节点 22 发送调度指令，控制调度节点 22 调度分布式系统中可用的计算节点 23。调度节点 22 具体接收控制节点 21 的调度指令，根据调度指令调度分布式系统中的计算节点 23。

在上述调度分布式系统中的计算节点 23 的过程中，控制节点 21 通过调度节点 22 向计算节点 23 提供后续加载第二数据表中的数据记录所需的配置文件，该配置文件记载有第二数据表的标识、存储位置、以及需要加载的数据记录的标识信息等。

在分布式系统中各计算节点 23 上部署有加载进程，该加载进程主要根据配置文件，将第二数据表中的数据记录加载到缓存模块中。具体的，调度模块 22 激活各计算节点 23 上的加载进程，加载进程根据配置文件，到相应存储位置读取第二数据表中的相应数据记录，将其所读取的数据记录加载到缓存模块中。值得说明的是，第二数据表可以存储在分布式系统

之外的空间中，但并不限于此。

当所有计算节点 23 上的加载进程执行完加载操作，即均进入监听端口状态时，通过调度节点 22 向控制节点 21 返回一个加载结束指令。控制节点 21 根据该加载结束指令，可以获知各计算节点 23 已经将第二数据表中的数据记录加载到缓存模块中。

控制节点 21 向调度节点 22 发送激活指令，使得调度节点 22 激活各计算节点 23 上的处理进程。在各计算节点 23 上部署有处理进程，处理进程主要用于读取第一数据表中的数据记录作为当前数据记录，根据所读取的当前数据记录对应的 key，确定该 key 对应的第二数据表中的数据记录所在的计算节点 23，读取所确定的计算节点 23 上存储的第二数据表中的数据记录作为目标数据记录，对所读取的当前数据记录和目标数据记录进行 Join 操作。值得说明的是，第一数据表可以存储在分布式系统之外的空间中，但并不限于此。

可选的，在一种具体实现方式中，上述各计算节点 23 可以采用服务端/客户端的方式实现。例如，各计算节点 23 的缓存模块可以作为缓存服务端（CacheService）实现，该缓存服务端还包括一个缓存管理者（CacheManager），各缓存模块对应一个缓存节点（CacheNode）；相应的，各计算节点 23 的处理模块作为缓存客户端（CacheClient）实现。

具体的，CacheManager 协调管理所有 CacheNode。CacheNode 负责加载数据到内存，并提供服务。可选的，第二数据表可以用 shard 文件的形式进行存储管理，采用 shard 文件的目的是因为在故障（failover）时，一但 CacheNode 重启，只需再次读入 shard 文件，使得处理相对简单。

CacheClient 访问 CacheService，通过对 key 进行 hash 计算，并根据计算结果从其中某个 CacheNode 读取数据。此外，CacheClient 中应该有一部分本地缓存，通常会使用近期最少使用算法（Least Recently Used，LRU）等缓存算法将部分已经读取的数据保存在本地缓存中，这样 CacheClient 可以优先从本地缓存中读取所需的数据，如果在本地缓存中读取到所需的数据，可以节约通过网络从 CacheNode 读取数据的操作，有利于提高效率、节约资源。

由上述分析可见，本实施例在处理数据表连接任务时，首先根据其中的连接条件，将第二数据表中的数据记录加载到至少两个节点上，实现分布式存储，使得可以直接读取第一数据表中的数据记录，并根据所读取的第一数据表中的数据记录对应的连接条件，从相应节点上读取所需的第二数据表中的数据记录，之后对读取到的两个数据表中的数据记录进行 Join 操作，实现分布式的 Hash map Join。由此可见，本实施例只需将第二数据表按照连接条件分布到不同节点上，不需要将第一数据表分布到不同节点上，减少了 shuffle 排序需要处理的数据量，有利于降低连接操作所消耗的计算资源。

需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本申请所必须的。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

图 3 为本申请又一实施例提供的数据表连接装置的结构示意图。如图 3 所示，该装置包括：接收模块 31、加载模块 32、读取模块 33 和连接模块 34。

接收模块 31，用于接收数据表连接任务，该数据表连接任务指示按照连接条件对第一数据表和第二数据表进行连接操作。

加载模块 32，用于根据连接条件，将第二数据表中的数据记录加载到分布式系统中至少两个节点上。

读取模块 33，用于读取第一数据表中的数据记录作为当前数据记录，根据当前数据记录对应的连接条件，从至少两个节点中确定目标节点，并读取目标节点上存储的第二数据表中的数据记录作为目标数据记录。

连接模块 34，用于对当前数据记录和目标数据记录进行连接操作。

优选的，位于至少两个节点中每个节点上的第二数据表中的数据记录，其数据量小于单个节点的内存限制，也就是说，分布到至少两个节点中每个节点上的第二数据表中的数据记录，均能够全部放到相应节点的存储空间（优选为内存）中。

进一步，如图 4 所示，该装置还包括：第一判断模块 35。

第一判断模块 35，用于判断第二数据表的数据量是否大于单个节点的内存限制，以及在判断结果为是时触发加载模块 32 执行根据连接条件，将第二数据表中的数据记录加载到分布式系统中至少两个节点上的操作。

更进一步，如图 4 所示，该装置还包括：第二判断模块 36。

第二判断模块 36，用于根据当前数据记录对应的连接条件，判断本地缓存中是否存在所述目标数据记录，以及在判断结果为否时触发读取模块 33 执行根据当前数据记录对应的连接条件，从至少两个节点中确定目标节点，并读取目标节点上存储的第二数据表中的数据记录作为目标数据记录的操作。

在一可选实施方式中，上述连接条件包括连接所需的至少一个目标键。这里的目标键实际上是键值对中的键。

基于上述，加载模块 32 具体用于：

分别对所述至少一个目标键中的各目标键进行哈希运算，以获取各目标键的哈希值；

根据各目标键的哈希值和所述至少两个节点的数量，确定各目标键对应的节点；

将所述第二数据表中对应于各目标键的数据记录分别加载到各目标键对应的节点上。

在一可选实施方式中，加载模块 32 具体用于：

根据所述连接条件，将所述第二数据表中的数据记录加载到所述至少两个节点的内存中。第二数据表中的数据记录存储在节点的内存中，可以随时读取，读取速度较快，有利于提高 Join 操作的效率。

值得说明的是，优选的，可以将第二数据表中的数据记录加载到上述

至少两个节点的内存中，但并不限于内存，还可以是节点的SSD或者其他存储介质中。

本实施例提供的数据表连接装置，在处理数据表连接任务时，首先根据其中的连接条件，将第二数据表中的数据记录加载到分布式系统中至少两个节点上，这相当于变成了一个分布式的KV存储，这样不需要做Sort Merge Join，使得可以做分布式的Hash map Join，即不需要对第一数据中的数据记录进行排序，而是可以直接读取第一数据表中的数据记录，并根据所读取的第一数据表中的数据记录对应的连接条件，从相应节点上读取所需的第二数据表中的数据记录，之后对读取到的两个数据表中的数据记录进行连接操作。由此可见，采用本实施例提供的数据表连接装置，只需将第二数据表按照连接条件分布到不同节点上，不需要将第一数据表分布到不同节点上，减少了shuffle排序需要处理的数据量，有利于降低连接操作所消耗的计算资源。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在

一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

上述以软件功能单元的形式实现的集成的单元，可以存储在一个计算机可读取存储介质中。上述软件功能单元存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）或处理器（processor）执行本申请各个实施例所述方法的部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（Read-Only Memory，ROM）、随机存取存储器（Random Access Memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

最后应说明的是：以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

权 利 要 求 书

1、一种数据表连接方法，其特征在于，包括:

接收数据表连接任务，所述数据表连接任务指示按照连接条件对第一数据表和第二数据表进行连接操作;

根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上;

读取所述第一数据表中的数据记录作为当前数据记录，根据所述当前数据记录对应的连接条件，从所述至少两个节点中确定目标节点，并读取所述目标节点上存储的所述第二数据表中的数据记录作为目标数据记录;

对所述当前数据记录和所述目标数据记录进行连接操作。

2、根据权利要求 1 所述的方法，其特征在于，所述根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上之前，包括:

判断所述第二数据表的数据量是否大于单个节点的内存限制;

若判断结果为是，则执行根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上的操作。

3、根据权利要求 1 所述的方法，其特征在于，所述根据所述当前数据记录对应的连接条件，从所述至少两个节点中确定目标节点，并读取所述目标节点上存储的所述第二数据表中的数据记录作为目标数据记录之前，包括:

根据所述当前数据记录对应的连接条件，判断本地缓存中是否存在所述目标数据记录;

若判断结果为否，则执行根据所述当前数据记录对应的连接条件，从所述至少两个节点中确定目标节点，并读取所述目标节点上存储的所述第二数据表中的数据记录作为目标数据记录的操作。

4、根据权利要求 1 所述的方法，其特征在于，所述连接条件包括连

接所需的至少一个目标键;

所述根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上，包括:

分别对所述至少一个目标键中的各目标键进行哈希运算，以获取各目标键的哈希值;

根据各目标键的哈希值和所述至少两个节点的数量，确定各目标键对应的节点;

将所述第二数据表中对应于各目标键的数据记录分别加载到各目标键对应的节点上。

5、根据权利要求 1-4 任一项所述的方法，其特征在于，所述根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上，包括:

根据所述连接条件，将所述第二数据表中的数据记录加载到所述至少两个节点的内存中。

6、一种数据表连接装置，其特征在于，包括:

接收模块，用于接收数据表连接任务，所述数据表连接任务指示按照连接条件对第一数据表和第二数据表进行连接操作;

加载模块，用于根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上;

读取模块，用于读取所述第一数据表中的数据记录作为当前数据记录，根据所述当前数据记录对应的连接条件，从所述至少两个节点中确定目标节点，并读取所述目标节点上存储的所述第二数据表中的数据记录作为目标数据记录;

连接模块，用于对所述当前数据记录和所述目标数据记录进行连接操作。

7、根据权利要求 6 所述的装置，其特征在于，还包括:

第一判断模块，用于判断所述第二数据表的数据量是否大于单个节点

的内存限制，以及在判断结果为是时触发所述加载模块执行根据所述连接条件，将所述第二数据表中的数据记录加载到分布式系统中至少两个节点上的操作。

8、根据权利要求 6 所述的装置，其特征在于，还包括:

第二判断模块，用于根据所述当前数据记录对应的连接条件，判断本地缓存中是否存在所述目标数据记录，以及在判断结果为否时触发所述读取模块执行根据所述当前数据记录对应的连接条件，从所述至少两个节点中确定目标节点，并读取所述目标节点上存储的所述第二数据表中的数据记录作为目标数据记录的操作。

9、根据权利要求 6 所述的装置，其特征在于，所述连接条件包括连接所需的至少一个目标键;

所述加载模块具体用于:

分别对所述至少一个目标键中的各目标键进行哈希运算，以获取各目标键的哈希值;

根据各目标键的哈希值和所述至少两个节点的数量，确定各目标键对应的节点;

将所述第二数据表中对应于各目标键的数据记录分别加载到各目标键对应的节点上。

10、根据权利要求 6-9 任一项所述的装置，其特征在于，所述加载模块具体用于:

根据所述连接条件，将所述第二数据表中的数据记录加载到所述至少两个节点的内存中。

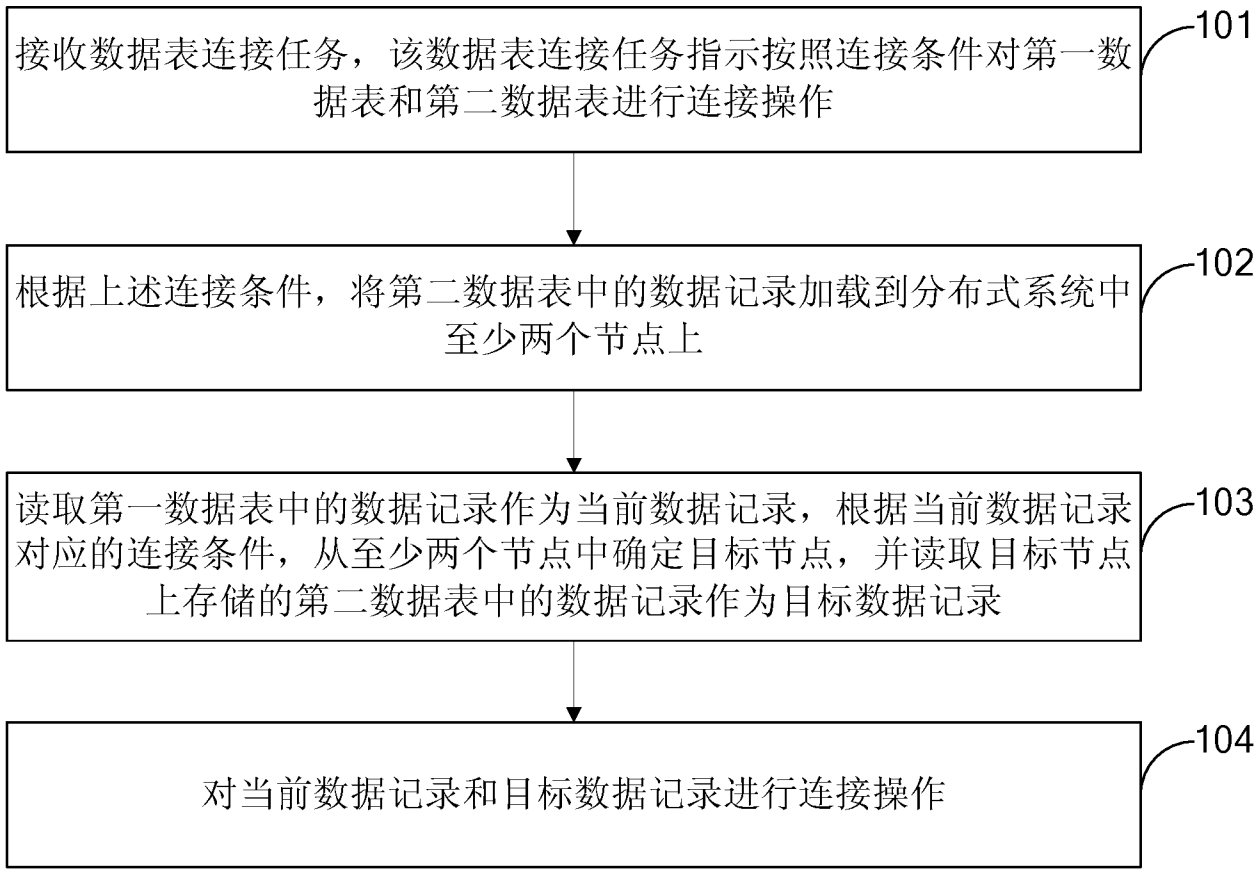
接收数据表连接任务，该数据表连接任务指示按照连接条件对第一数据表和第二数据表进行连接操作
101
根据上述连接条件，将第二数据表中的数据记录加载到分布式系统中至少两个节点上
102
读取第一数据表中的数据记录作为当前数据记录，根据当前数据记录对应的连接条件，从至少两个节点中确定目标节点，并读取目标节点上存储的第二数据表中的数据记录作为目标数据记录
103
对当前数据记录和目标数据记录进行连接操作
104

图 1

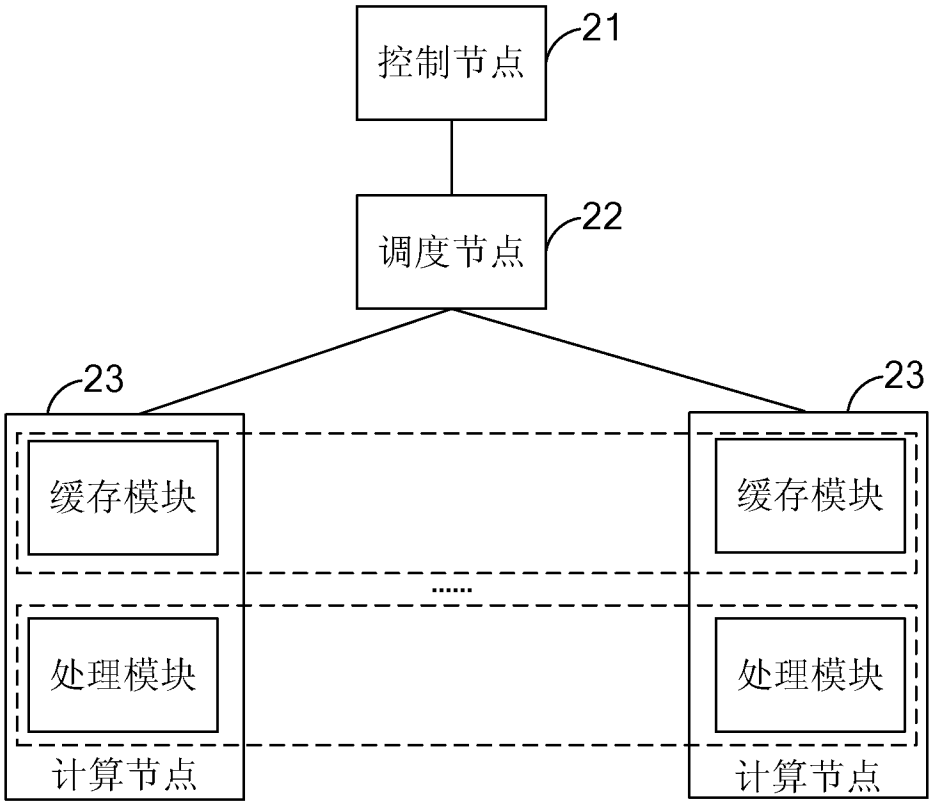
控制节点
21
调度节点
22
23
23
缓存模块
缓存模块
.....
处理模块
处理模块
计算节点
计算节点

图 2

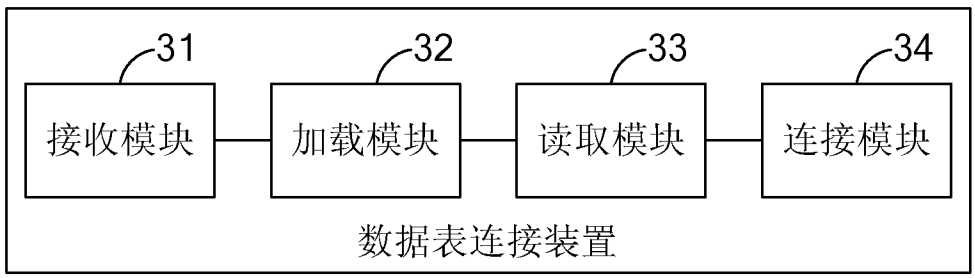
31
32
33
34
接收模块
加载模块
读取模块
连接模块
数据表连接装置

图 3

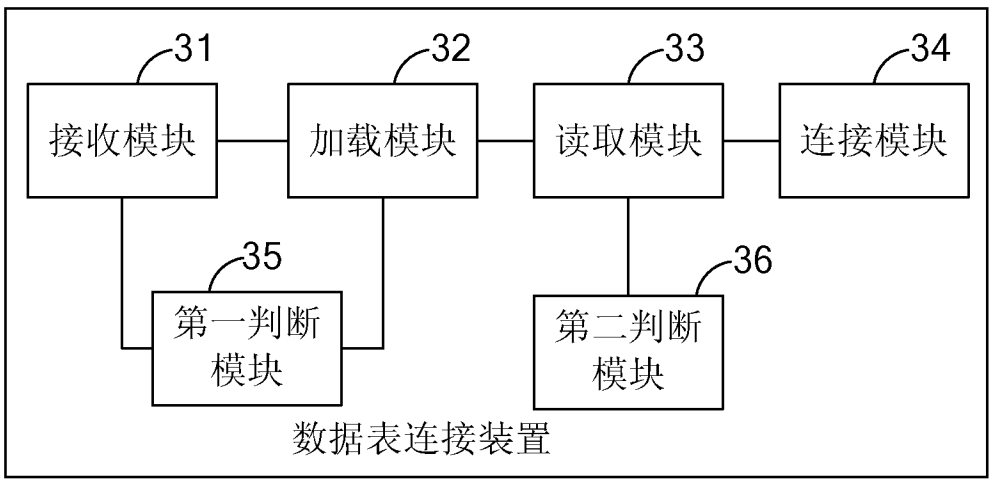
31
32
33
34
接收模块
加载模块
读取模块
连接模块
35
36
第一判断模块
第二判断模块
数据表连接装置

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/074177

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) 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)

CNKI, CNPAT, WPI, EPODOC, IEEE: key value, hash, map, side, join, KV, memory, distribut+, node

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104504114 A (HANGZHOU HUAWEI DIGITAL TECHNOLOGY CO., LTD.) 08 April 2015 (08.04.2015) description, paragraphs [0044]-[0068] and [0081]-[0083]	1-10
A	CN 105183880 A (INSPUR GROUP CO., LTD.) 23 December 2015 (23.12.2015) the whole document	1-10
A	CN 104391957 A (INSPUR ELECTRONIC INFORMATION INDUSTRY CO., LTD.) 04 March 2015 (04.03.2015) the whole document	1-10
A	CN 102214176 A (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY, PLA) 12 October 2011 (12.10.2011) the whole document	1-10
A	US 7085769 B1 (NCR CORP.) 01 August 2006 (01.08.2006) the whole 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
01 April 2017	28 April 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer LI, Duo Telephone No. (86-10) 62414054

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/074177

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104504114 A	08 April 2015	None	
CN 105183880 A	23 December 2015	None	
CN 104391957 A	04 March 2015	None	
CN 102214176 A	12 October 2011	None	
US 7085769 B1	01 August 2006	None	

国际检索报告

国际申请号
PCT/CN2017/074177

A. 主题的分类

G06F 17/30(2006.01)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））

CNKI，CNPAT，WPI，EPODOC，IEEE：哈希，映射，连接，键值，存储，分布式，节点，内存，hash，map，side，join，KV，memory，distribut+，node

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 104504114 A (杭州华为数字技术有限公司) 2015年 4月 8日 (2015-04-08) 说明书第[0044]-[0068]，[0081]-[0083]段	1-10
A	CN 105183880 A (浪潮集团有限公司) 2015年 12月 23日 (2015-12-23) 全文	1-10
A	CN 104391957 A (浪潮电子信息产业股份有限公司) 2015年 3月 4日 (2015-03-04) 全文	1-10
A	CN 102214176 A (中国人民解放军国防科学技术大学) 2011年 10月 12日 (2011-10-12) 全文	1-10
A	US 7085769 B1 (NCR CORP.) 2006年 8月 1日 (2006-08-01) 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
2017年 4月 1日	2017年 4月 28日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	李铎
传真号 (86-10)62019451	电话号码 (86-10)62414054

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/074177

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 104504114 A	2015年 4月 8日	无	
CN 105183880 A	2015年 12月 23日	无	
CN 104391957 A	2015年 3月 4日	无	
CN 102214176 A	2011年 10月 12日	无	
US 7085769 B1	2006年 8月 1日	无	