

(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 11 月 9 日 (09.11.2017)



(10) 国际公布号
WO 2017/190469 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2016/101058
- (22) 国际申请日: 2016 年 9 月 30 日 (30.09.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610290381.0 2016年5月4日 (04.05.2016) CN
- (71) 申请人: 乐视控股 (北京) 有限公司 (LE HOLDINGS (BEIJING) CO., LTD.) [CN/CN]; 中国北京市朝阳区姚家园路105号3号楼10层1102, Beijing 100025 (CN)。乐视网信息技术 (北京) 股份有限公司 (LE SHI INTERNET INFORMATION TECHNOLOGY CORP. BEI JING) [CN/CN]; 中国北京市朝阳区姚家园路105号观湖国际大厦10层, Beijing 100025 (CN)。
- (72) 发明人: 刘宏斌 (LIU, Hongbin); 中国北京市朝阳区姚家园路105号观湖国际大厦10层, Beijing 100025 (CN)。国铁龙 (GUO, Tielong); 中国北京市朝阳区姚家园路105号观湖国际大厦10层, Beijing 100025 (CN)。杨海乐 (YANG, Haile); 中国北京市朝阳区姚家园路105号观湖国际大厦10层, Beijing 100025 (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,

(54) Title: DATA OPTIMISATION METHOD AND APPARATUS IN BIG DATA PROCESSING

(54) 发明名称: 大数据处理中的数据优化方法和装置

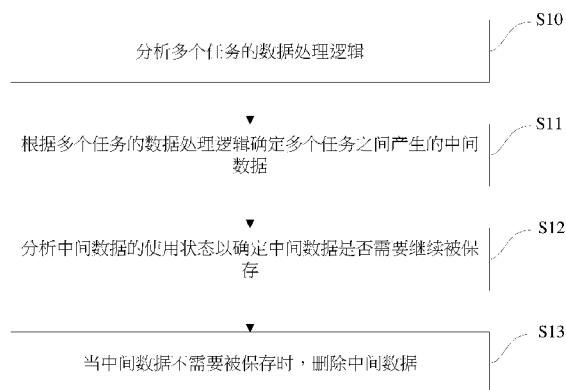


图 1

S10 ANALYSING A DATA PROCESSING LOGIC OF A PLURALITY OF TASKS
S11 DETERMINING, ACCORDING TO THE DATA-PROCESSING LOGIC OF THE PLURALITY OF TASKS, INTERMEDIATE DATA GENERATED BETWEEN THE PLURALITY OF TASKS
S12 ANALYSING THE USAGE STATE OF THE INTERMEDIATE DATA SO AS TO DETERMINE WHETHER THE INTERMEDIATE DATA NEEDS TO BE CONSERVED CONTINUOUSLY
S13 DELETING THE INTERMEDIATE DATA WHEN THE INTERMEDIATE DATA DOES NOT NEED TO BE CONSERVED

(57) Abstract: A data optimisation method and apparatus in big data processing. The method comprises: analysing a data processing logic of a plurality of tasks (S10); determining, according to the data processing logic of the plurality of tasks, intermediate data generated between the plurality of tasks (S11); analysing the usage state of the intermediate data, so as to determine whether the intermediate data needs to be conserved continuously (S12); and deleting the intermediate data when the intermediate data does not need to be conserved (S13). Thereby, unnecessary intermediate data is removed, and the storage space of a data warehouse is saved.

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, 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))。

(57) 摘要: 一种大数据处理中的数据优化方法和装置, 该方法包括: 分析多个任务的数据处理逻辑 (S10); 根据所述多个任务的数据处理逻辑确定多个任务之间产生的中间数据 (S11); 分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保存 (S12); 当所述中间数据不需要被保存时, 删除所述中间数据 (S13)。清除了不必要的中间数据, 从而节省了数据仓库的存储空间。

大数据处理中的数据优化方法和装置

交叉引用

5 本申请引用于 2016 年 05 月 04 日递交的名称为“大数据处理中的数据优化方法和装置”的第 201610290381.0 号中国专利申请，其通过引用被全部并入本申请。

技术领域

10 本发明属于计算机领域，具体地说，涉及一种大数据处理中的数据优化方法和装置。

背景技术

15 随着互联网的快速发展，很多互联网公司都已积累了 TB 量级的数据。数据仓库每天都在接收来自不同生态的数据，例如来自手机、智能电视、视频网站的用户数据记录等，做为大数据资源的一部分。

20 数据从数据仓库的入口机进入数据仓库以及在数据仓库内部进行分层，都需要进行数据处理，每次数据处理过程都是多个任务的集合，每个任务都具有内在的处理逻辑，例如任务 1 是将 A 表中的部分字段的数据读取再写入到 B 表。有时，当很多数据工程师都需要某些数据时，不同的数据工程师利用现有数据获取到所需要数据方法路径可能就会不同，此时遗留下一些中间数据，久而久之还会出现很多重复的数据，而且其中的很多数据在今后都不会再被用到。

25 上述问题是由于对任务的内在处理逻辑分析不到位而造成的，导致了很多存储资源的浪费，减少了数据仓库的有效存储空间。

发明内容

有鉴于此，本发明实施例提供了一种大数据处理中的数据优化方法和装

置，用以解决现有技术中由于对任务的内在处理逻辑分析不到位而导致浪费存储资源的技术问题。

本发明实施例还提供一种电子设备、一种非暂态计算机存储介质以及一种计算机程序产品。

5 为了解决上述技术问题，本发明公开了一种大数据处理中的数据优化方法，包括：分析多个任务的数据处理逻辑；根据所述多个任务的数据处理逻辑确定多个任务之间产生的中间数据；分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保存；当所述中间数据不需要被保存时，删除所述中间数据。

10 为了解决上述技术问题，本发明还公开了一种大数据处理中的数据优化装置，包括：第一分析模块，用于分析多个任务的数据处理逻辑；第一确定模块，用于根据所述多个任务的数据处理逻辑确定多个任务之间产生的中间数据；第二确定模块，用于分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保存；第一删除模块，用于当所述中间数据不需要被保存
15 时，删除所述中间数据。

本发明实施例还提供一种非暂态计算机存储介质，存储有计算机可执行指令，所述计算机可执行指令用于执行本申请上述任一项大数据处理中的数据优化方法。

20 本发明实施例提供一种电子设备，包括：至少一个处理器；以及，与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述一个处理器执行的指令，所述指令被被所述至少一个处理器执行，以使所述至少一个处理器能够：

分析多个任务的数据处理逻辑；

根据所述多个任务的数据处理逻辑确定多个任务之间产生的中间数据；

25 分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保存；

当所述中间数据不需要被保存时，删除所述中间数据。

本发明实施例还提供了一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，使所述计算机执行本申请上述任一项大数据处理中的数据优化方法。

- 5 与现有技术相比，本发明实施例提供的大数据处理中的数据优化方法和装置，会对各个任务之间产生的中间数据进行检测，以判定其是否还会被利用，如果判定其没有被使用，则会将该中间数据删除，清除了不必要的中间数据，从而节省了数据仓库的存储空间。

10 附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，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发明的一些实施例，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。

- 15 图 1 是本发明实施例提供的一种大数据处理中的数据优化方法的流程图；

图 2 是本发明实施例提供的一种大数据处理中的数据优化方法的流程图；

图 3 是本发明实施例提供的一种大数据处理中的数据优化装置的框图；

- 20 图 4 是本发明实施例提供的一种大数据处理中的数据优化装置的框图；

图 5 为本发明实施例提供的一种电子设备的框图。

具体实施方式

- 25 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明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发明一部分实施例，而不是全部的实施例。基于

本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

5 本发明实施例中，针对数据仓库中的计算任务进行分析，分析每个任务的数据处理逻辑，通过数据处理逻辑来找到各个任务之间的逻辑关系以及数据依赖关系，对各个任务之间产生的中间数据以及任务的执行情况进行分析，找到不再被使用的可以进行优化的数据，删除不再被使用的中间数据，节省了数据仓库的存储空间，例外对相应的任务进行适当的合并，从而节省数据仓库的计算资源并提高任务的执行效率。

10 图 1 是本发明实施例提供的一种大数据处理中的数据优化方法，适用服务器，该方法包括以下步骤。

S10，分析多个任务的数据处理逻辑。

15 数据处理逻辑包括处理对象和计算方法。处理对象包括源数据、目标数据等，例如任务 T01 是从表 A 读取三个字段的数据并写入表 B。计算方法是指利用源数据生成目标数据的方法，如果是直接从表 A 读取数据并写入表 B 则不存在计算方法，而如果对从表 A 读取的数据进行计算之后将结果写入表 B，在该任务存在表 A 与表 B 之间存在计算方法。

S11，根据多个任务的数据处理逻辑确定多个任务之间产生的中间数据。

20 从多个任务的数据处理逻辑中，找出多个任务的之间的逻辑关系。例如，任务 T01 从表 A 读取三个字段的数据并写入表 B，任务 T02 对表 B 中的三个字段的数据进行筛选，筛选出满足预设条件的数据并写入表 C，任务 T03 读取表 C 的数据并添加到表 D 中。可以看出任务 T01 至 T03 是按照彼此之间的逻辑关系依次进行的。找到多个任务之间的逻辑关系后，就可以确定各个任务之间都产生了哪些中间数据，上例中的表 B 和表 C 即可以被确定为中间数据。

25 不同的数据工程师对得到目标数据而设置的计算方式会有所不同，有时还会根据其所负责的业务的实际需求来得到一些中间数据供进行其他计算使用。因此，需要进一步判断这些中间数据会被使用，也就是判断这些中间数据是否有必要进行保存。

S12, 分析中间数据的使用状态以确定中间数据是否需要继续被保存。

使用状态包括该中间数据是否会被用于其它计算, 以及该中间数据本身是否是其他任务链的最终结果。因此, 对于中间数据是否需要保存的判定, 可以通过多种方式进行。

5 在一个实施例中, 该步骤 S12 可进一步被实施为以下步骤。

S120, 根据业务需求分析中间数据是否在业务中被使用。

10 业务需求包括该数据是否用于其它业务数据的计算以及该中间数据在业务中是否也是需要的最终结果。例如, 中间数据 B 记录了上海的各门店在 2016 年 1 月至 3 月的智能电视销量, 如果业务中还需要进一步筛选出销量排名前五位的门店, 则代表该中间数据 B 还会被使用; 或者, 该中间数据 B 本身就是一个统计上海市在 2016 年 1 月至 3 月的智能电视销量的任务链的最终结果, 则代表该中间数据也需要被使用。

S121, 当中间数据在业务中不被使用时, 确定中间数据不需要继续被保存。

15 实现了根据预设的业务逻辑中对数据的实际需求来判定任务链的中间数据是否需要被保存。

在另一实施例中, 该步骤 S12 还可以进一步被实施为以下步骤。

S122, 统计中间数据的未被使用的累积时长, 当累积时长达到预设门限时, 标记中间数据为不被使用的数据。

20 对于被判定为任务链中的中间数据, 会统计该中间数据未被使用的累积时长, 例如, 只要没有出现针对中间数据 B 的读取操作, 就说明该中间数据 B 没有被使用, 当中间数据 B 被读取时, 累积时长将被清零并重新开始计时, 如果在预设时长 (例如 12 小时) 都没有针对中间数据 B 的读取操作, 则标记该中间数据 B 为不被使用的数据。

25 为了降低发生误判的概率, 还会进一步对该中间数据被标记为不被使用的数据的次数进行统计。如果在接下来的预设时长该数据仍然没有被使用, 则再一次标记该中间数据为不会被使用的数据。

S123, 当中间数据被标记为不被使用的数据的次数大于或等于预设门限时, 确定中间数据不需要继续被保存。

例如, 中间数据 B 已连续 10 次被标记为不被使用的数据, 则可以认为该数据不需要被继续保存。

- 5 这种不会被使用的中间数据的出现往往都是由于不同的数据工程师通过不同的方式获取目标数据时而人为配置的, 随意性会比较强又不会被其他其他数据工程师所利用。

S13, 当中间数据不需要被保存时, 删除中间数据。

- 10 如上例中, 如果表 B 被判定为不需要保存的中间数据, 则删除该表 B; 如果表 C 被判定为不需要保存的中间数据, 则删除该表 C; 如果表 B 和表 C 都被判定为不需要保存的中间数据, 则将表 B 和表 C 全部删除。

- 15 在多个任务组成的任务链中, 会对各个任务之间产生的中间数据进行检测, 以判定其是否还会被利用, 如果根据业务逻辑判定其不会被使用或者通过计时判定其很长时间都没有被使用, 则会将该中间数据删除, 清除了不必要的中间数据, 从而节省了数据仓库的存储空间。

在一个实施例中, 该大数据处理中的数据优化方法进一步包括以下步骤。

S14, 根据数据处理逻辑将多个任务合并为一个任务。

- 20 在删除不需要保存的中间数据后, 对于生成这些被删除的中间数据的任务也可以随之进行相应调整, 将原来的多个任务合并成一个任务, 也就避免了再次产生中间数据, 同时还能节约数据仓库的计算资源, 提高数据仓库的处理效率。如上例中, 如果表 B 被判定为不需要保存的中间数据, 则根据数据处理逻辑将任务 T01 和 T02 合并为 T12, 合并后的任务 T12 的处理对象就是表 A 和表 C, 计算方法也相应合并为从表 A 中读取三个字段的 data 并根据
25 预设条件进行筛选, 将筛选结果写入表 B。如果表 C 被判定为不需要保存的中间数据, 则根据数据处理逻辑将任务 T02 和 T03 合并为 T23, 合并后的任务 T23 的处理对象就是表 B 和表 D, 计算方法也相应合并为对表 B 中三个字

段数据进行筛选并将筛选结果添加至表 D。如果表 B 和表 C 都被判定为不需要保存的中间数据,则根据数据处理逻辑将任务 T01、T02 和 T03 合并为 T13,合并后的任务 T13 的处理对象就是表 A 和表 D,计算方法也相应合并为从表 A 中读取三个字段的数据并按照预设条件进行筛选,将筛选结果添加到表 D。

- 5 也就是说,如果两个任务之间存在不会被使用的中间数据,则可以这两个任务合并一个任务,如果连续出现多个不会被使用的中间数据,则可以将多个任务合并为一个任务,从而减少了数据仓库中需要执行的计算任务数量,节约了计算资源,有助于提高数据仓库的处理效率。

10 在一个实施例中,如图 2 所示,上述大数据处理中的数据优化方法可进一步包括以下步骤。

S15, 根据数据处理逻辑判断是否同时存在多个任务能够产生相同的中间数据。

- 15 S16, 当同时存在多个任务能够产生相同的中间数据时,在相同的中间数据中保留一份副本并删除其他相同的中间数据,后续任务都从保留的副本中读取需要的数据。

- 20 该多个能够产生相同中间数据的任务来自于不同数据工程师的配置。例如,大家都已知存在表 A,甲需要提取表 A 中三个字段的数据并写入表 B,对表 B 的数据进行预测分析,输出分析结果至表 C;而乙需要提取表 A 中相同的三个字段的数据并写入表 D,对表 D 的数据进行筛选并将结果输出到表 E。可见此时存在两个从表 A 读取三个字段数据的任务,并分别将读取到的数据写入表 B 和表 D。那么此时可以在表 B 和表 D 中保留任意一个并删除另外一个,例如保留表 B 同时删除表 D,并将乙配置的把从表 A 读取的数据写入表 D 的任务以及从表 D 读取数据进行筛选的任务都重定向至表 B,从而使乙配置的任务会把从表 A 读取的数据写入表 B 同时会从表 B 读取数据进行筛选。这样即可将重复的中间数据进行删除,只保留一份副本来满足其他任务的数据读写需求,进一步节省了数据仓库的存储资源。
- 25

此外,在另一个实施例中,还可以进一步将同时产生相同中间数据的多个任务合并为一个任务,如上例中在删除表 D 之后可以进一步将甲配置的提

取表 A 中三个字段的数据并写入表 B 的任务与甲配置的提取表 A 中三个字段的数据并重定向写入表 B 的任务合并为一个任务。合并后，甲乙配置的其他后续任务共同利用该合并后的任务的输出结果。

对同时产生相同中间数据的多个任务进行合并，可以进一步减少计算任务的数量，节约计算资源。

下面是本发明的装置实施例，用于执行本发明的上述方法实施例。

图 3 是本发明实施例提供的一种大数据处理中的数据优化装置，包括：

第一分析模块 30，用于分析多个任务的数据处理逻辑；

10 第一确定模块 31，用于根据多个任务的数据处理逻辑确定多个任务之间产生的中间数据；

第二确定模块 32，用于分析中间数据的使用状态以确定中间数据是否需要继续被保存；

第一删除模块 33，用于当中间数据不需要被保存时，删除中间数据。

在一个实施例中，该第二确定模块 32 进一步包括：

15 第一分析子模块，用于根据业务需求分析所述中间数据是否在业务中被使用；

第一确定子模块，用于当所述中间数据在业务中不被使用时，确定所述中间数据不需要继续被保存。

在一个实施例中，该第二确定模块 32 进一步包括：

20 标记子模块，用于统计所述中间数据的未被使用的累积时长，当所述累积时长达到预设门限时，标记所述中间数据为不被使用的数据；

第二确定子模块，用于当所述中间数据被标记为不被使用的数据的次数大于或等于预设门限时，确定所述中间数据不需要继续被保存。

在一个实施例中，该装置进一步包括：

25 合并模块，用于根据数据处理逻辑将多个任务合并为一个任务。

在一个实施例中，如图 4 所示，该装置进一步包括：

判断模块 34, 用于根据数据处理逻辑判断是否同时存在多个任务能够产生相同的中间数据;

第二删除模块 35, 用于当同时存在多个任务能够产生相同的中间数据时, 在相同的中间数据中保留一份副本并删除其他相同的中间数据, 后续任务都从保留的副本中读取需要的数据。

此外, 本发明实施例中可以通过硬件处理器 (hardware processor) 来实现上述各个功能模块。

图 5 为本发明实施例提供的一种电子设备的框图, 本实施例所述设备可以为大数据处理中的数据优化服务器或大数据处理中的数据优化服务器中的一部分, 该设备可以包括:

一个或多个处理器 501 以及存储器 502, 图 5 中以一个处理器 501 为例。

电子设备还可以包括: 输入装置 503 和输出装置 504。

处理器 501、存储器 502、输入装置 503 和输出装置 504 可以通过总线或者其他方式连接。

存储器 502 作为一种非暂态计算机可读存储介质, 可用于存储非暂态软件程序、非暂态计算机可执行程序以及模块, 如本发明实施例中的大数据处理中的数据优化方法对应的程序指令/模块。处理器 501 通过运行存储在存储器 502 中的非暂态软件程序、指令以及模块, 从而执行服务器的各种功能应用以及数据处理, 即实现上述方法实施例大数据处理中的数据优化方法。

存储器 502 可以包括存储程序区和存储数据区, 其中, 存储程序区可存储操作系统、至少一个功能所需要的应用程序; 存储数据区可存储根据大数据处理中的数据优化装置的使用所创建的数据等。此外, 存储器 502 可以包括高速随机存取存储器, 还可以包括非暂态存储器, 例如至少一个磁盘存储器件、闪存器件、或其他非暂态固态存储器件。在一些实施例中, 存储器 502 可选包括相对于处理器 501 远程设置的存储器, 这些远程存储器可以通过网络连接至列表项操作的处理装置。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

输入装置 503 可接收输入的数字或字符信息，以及产生与大数据处理中的数据优化装置的用户设置以及功能控制有关的键信号输入。输出装置 504 可包括显示屏等显示设备。

5 所述一个或者多个模块存储在所述存储器 502 中，当被所述一个或者多个处理器 501 执行时，执行上述任意方法实施例中的大数据处理中的数据优化方法。

上述产品可执行本发明实施例所提供的方法，具备执行方法相应的功能模块和有益效果。未在本实施例中详尽描述的技术细节，可参见本发明实施例所提供的方法。

10 本发明实施例的电子设备以多种形式存在，包括但不限于：

(1)移动通信设备：这类设备的特点是具备移动通信功能，并且以提供话音、数据通信为主要目标。这类终端包括：智能手机(例如 iPhone)、多媒体手机、功能性手机，以及低端手机等。

15 (2)超移动个人计算机设备：这类设备属于个人计算机的范畴，有计算和处理功能，一般也具备移动上网特性。这类终端包括：PDA、MID 和 UMPC 设备等，例如 iPad。

(3)便携式娱乐设备：这类设备可以显示和播放多媒体内容。该类设备包括：音频、视频播放器(例如 iPod)，掌上游戏机，电子书，以及智能玩具和便携式车载导航设备。

20 (4)服务器：提供计算服务的设备，服务器的构成包括处理器、硬盘、内存、系统总线等，服务器和通用的计算机架构类似，但是由于需要提供高可靠的服务，因此在处理能力、稳定性、可靠性、安全性、可扩展性、可管理性等方面要求较高。

(5)其他具有数据交互功能的电子装置。

25 相应地，本发明实施例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中存储有用于执行上述实施例方法的程序。

以上所描述的装置实施例仅仅是示意性的，其中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模块来实现本实施例5 方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

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各实施方式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。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分10 可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如 ROM/RAM、磁碟、光盘等，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）执行各个实施例或者实施例的某些部分所述的方法。

最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其15 限制；尽管参照前述实施例对本发明进行了详细的说明，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发明各实施例技术方案的精神和范围。

权 利 要 求 书

1、一种大数据处理中的数据优化方法，其特征在于，包括：

分析多个任务的数据处理逻辑；

根据所述多个任务的数据处理逻辑确定多个任务之间产生的中间数据；

5 分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保存；

当所述中间数据不需要被保存时，删除所述中间数据。

2、根据权利要求 1 所述的方法，其特征在于，所述分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保存包括：

10 根据业务需求分析所述中间数据是否在业务中被使用；

当所述中间数据在业务中不被使用时，确定所述中间数据不需要继续被保存。

3、根据权利要求 1 所述的方法，其特征在于，所述分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保存包括：

15 统计所述中间数据的未被使用的累积时长，当所述累积时长达到预设门限时，标记所述中间数据为不被使用的数据；

当所述中间数据被标记为不被使用的数据的次数大于或等于预设门限时，确定所述中间数据不需要继续被保存。

4、根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

20 根据所述数据处理逻辑将所述多个任务合并为一个任务。

5、根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

根据数据处理逻辑判断是否同时存在多个任务能够产生相同的中间数据；

25 当同时存在多个任务能够产生相同的中间数据时，在相同的中间数据中保留一份副本并删除其他相同的中间数据，后续任务都从保留的副本中读取需要的数据。

6、一种大数据处理中的数据优化装置，其特征在于，包括：

第一分析模块，用于分析多个任务的数据处理逻辑；

第一确定模块，用于根据所述多个任务的数据处理逻辑确定多个任务之间产生的中间数据；

5 第二确定模块，用于分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保存；

第一删除模块，用于当所述中间数据不需要被保存时，删除所述中间数据。

7、根据权利要求6所述的装置，其特征在于，所述第二确定模块包括：

10 第一分析子模块，用于根据业务需求分析所述中间数据是否在业务中被使用；

第一确定子模块，用于当所述中间数据在业务中不被使用时，确定所述中间数据不需要继续被保存。

8、根据权利要求6所述的装置，其特征在于，所述第二确定模块包括：

15 标记子模块，用于统计所述中间数据的未被使用的累积时长，当所述累积时长达到预设门限时，标记所述中间数据为不被使用的数据；

第二确定子模块，用于当所述中间数据被标记为不被使用的数据的次数大于或等于预设门限时，确定所述中间数据不需要继续被保存。

9、根据权利要求6所述的装置，其特征在于，所述装置还包括：

20 合并模块，用于根据所述数据处理逻辑将所述多个任务合并为一个任务。

10、根据权利要求6所述的装置，其特征在于，所述装置还包括：

判断模块，用于根据数据处理逻辑判断是否同时存在多个任务能够产生相同的中间数据；

25 第二删除模块，用于当同时存在多个任务能够产生相同的中间数据时，在相同的中间数据中保留一份副本并删除其他相同的中间数据，后续任务都

从保留的副本中读取需要的数据。

11、一种电子设备，其特征在于，包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

5 所述存储器存储有可被所述一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够：

分析多个任务的数据处理逻辑；

根据所述多个任务的数据处理逻辑确定多个任务之间产生的中间数据；

分析所述中间数据的使用状态以确定所述中间数据是否需要继续被保

10 存；

当所述中间数据不需要被保存时，删除所述中间数据。

12、一种非暂态计算机可读存储介质，其特征在于，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令用于使所述计算机执行权利要求 1-5 任一所述方法。

15 13、一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，使所述计算机执行权利要求 1-5 任一所述方法。

1/3

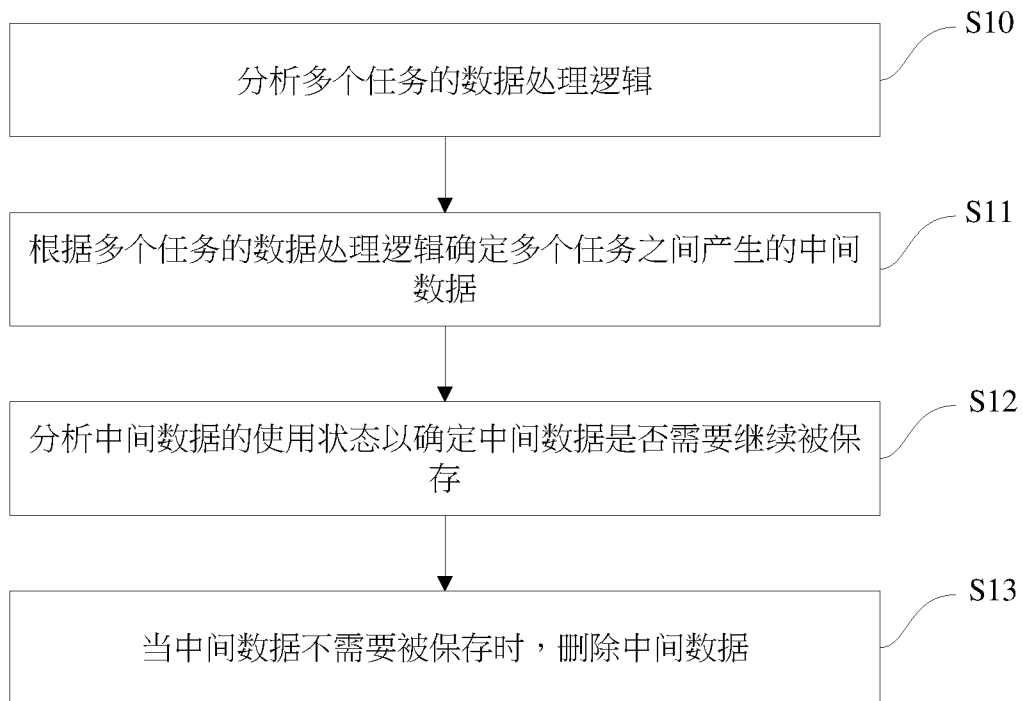


图 1

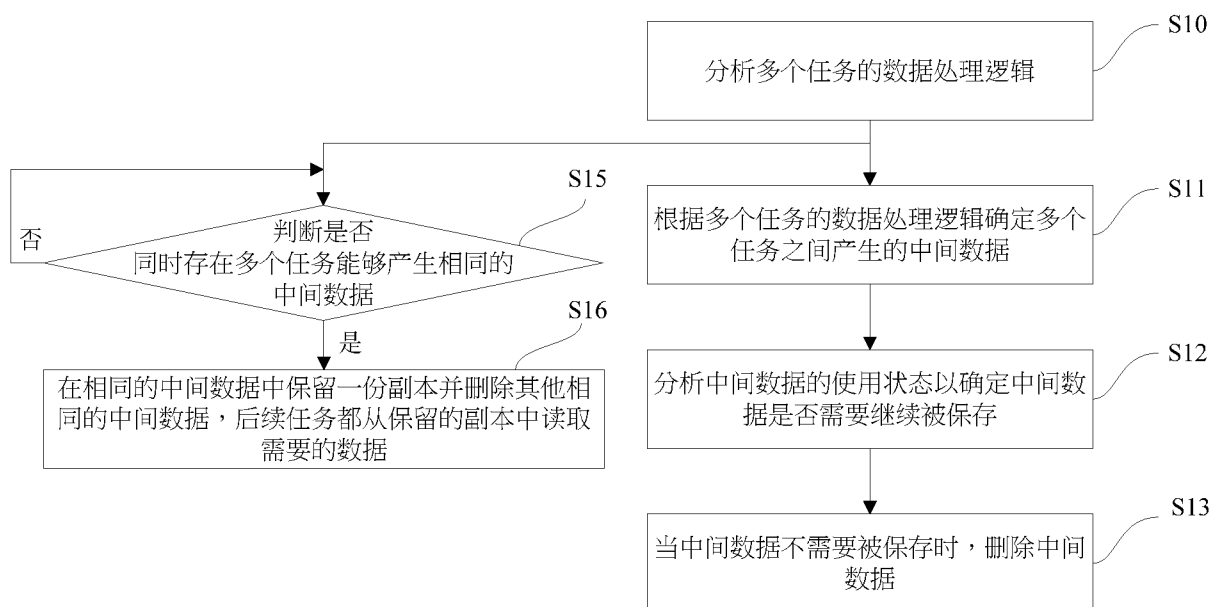


图 2

2/3

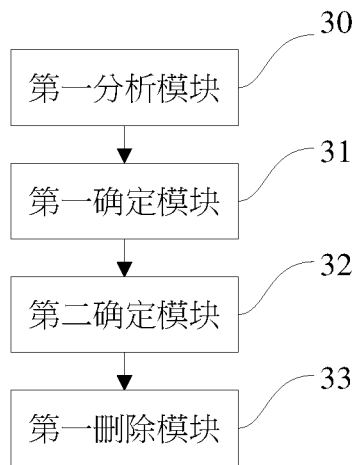


图 3

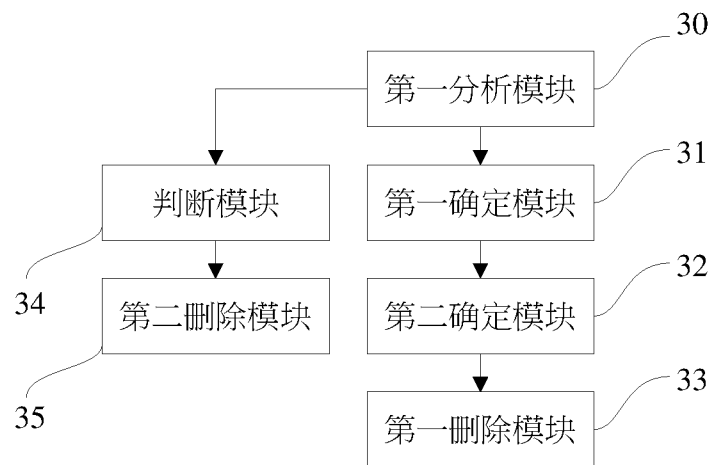


图 4

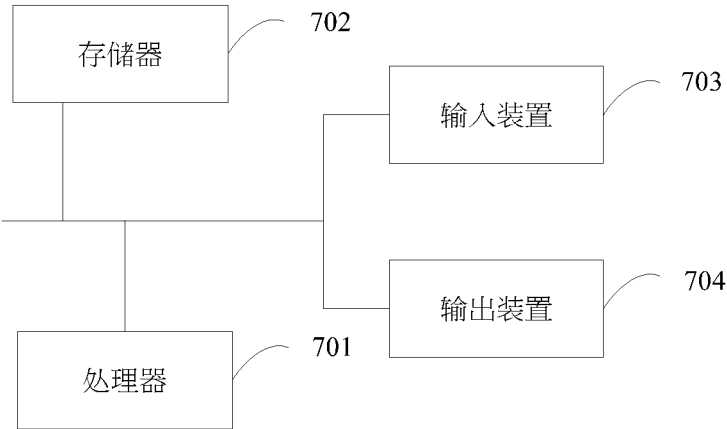


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/101058

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)

WPI, EPODOC, CNPAT, GOOGLE, CNKI, IEEE: big data, business, duration, big, data, optimize, assignment, task, deal, middle, state, keep, hold, preserve, delete, use, time, period, unite, combine, copy

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 105975577 A (LETV HOLDING GROUP (BEIJING) CO., LTD. et al.), 28 September 2016 (28.09.2016), claims 1-10, and description, paragraphs 0066-0073	1-13
X	CN 102456031 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.), 16 May 2012 (16.05.2012), description, paragraphs 0060-0103	1-13
A	CN 103793530 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD.), 14 May 2014 (14.05.2014), the whole document	1-13
A	CN 104391748 A (INSUR ELECTRONIC INFORMATION INDUSTRY CO., LTD.), 04 March 2015 (04.03.2015), the whole document	1-13
A	CN 102932416 A (NEUSOFT CORPORATION), 13 February 2013 (13.02.2013), the whole document	1-13

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 19 January 2017 (19.01.2017)	Date of mailing of the international search report 15 February 2017 (15.02.2017)
Name and mailing address of the ISA/CN: 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 WANG, Fang Telephone No.: (86-10) 82245496

International application No.
PCT/CN2016/101058

Form PCT/ISA/210 (patent family annex) (July 2009)

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)) WPI, EPODOC, CNPAT, GOOGLE, CNKI, IEEE: 大数据, 数据, 优化, 任务, 业务, 处理, 中间, 状态, 保持, 删除, 使用, 时长, 时间, 合并, 副本, big, data, optimize, assignment, task, deal, middle, state, keep, hold, preserve, delete, use, time, period, unite, combine, copy																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 105975577 A (乐视控股北京有限公司 等) 2016年 9月 28日 (2016 - 09 - 28) 权利要求1-10, 说明书第0066-0073段</td> <td>1-13</td> </tr> <tr> <td>X</td> <td>CN 102456031 A (腾讯科技深圳有限公司) 2012年 5月 16日 (2012 - 05 - 16) 说明书第0060-0103段</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 103793530 A (北京京东尚科信息技术有限公司) 2014年 5月 14日 (2014 - 05 - 14) 全文</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 104391748 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</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 102932416 A (东软集团股份有限公司) 2013年 2月 13日 (2013 - 02 - 13) 全文</td> <td>1-13</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 105975577 A (乐视控股北京有限公司 等) 2016年 9月 28日 (2016 - 09 - 28) 权利要求1-10, 说明书第0066-0073段	1-13	X	CN 102456031 A (腾讯科技深圳有限公司) 2012年 5月 16日 (2012 - 05 - 16) 说明书第0060-0103段	1-13	A	CN 103793530 A (北京京东尚科信息技术有限公司) 2014年 5月 14日 (2014 - 05 - 14) 全文	1-13	A	CN 104391748 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	1-13	A	CN 102932416 A (东软集团股份有限公司) 2013年 2月 13日 (2013 - 02 - 13) 全文	1-13
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 105975577 A (乐视控股北京有限公司 等) 2016年 9月 28日 (2016 - 09 - 28) 权利要求1-10, 说明书第0066-0073段	1-13																		
X	CN 102456031 A (腾讯科技深圳有限公司) 2012年 5月 16日 (2012 - 05 - 16) 说明书第0060-0103段	1-13																		
A	CN 103793530 A (北京京东尚科信息技术有限公司) 2014年 5月 14日 (2014 - 05 - 14) 全文	1-13																		
A	CN 104391748 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	1-13																		
A	CN 102932416 A (东软集团股份有限公司) 2013年 2月 13日 (2013 - 02 - 13) 全文	1-13																		
☐ 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 2017年 1月 19日		国际检索报告邮寄日期 2017年 2月 15日																		
ISA/CN的名称和邮寄地址 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 传真号 (86-10) 62019451		受权官员 王芳 电话号码 (86-10) 82245496																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/101058

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	105975577	A	2016年 9月 28日	无	
CN	102456031	A	2012年 5月 16日	无	
CN	103793530	A	2014年 5月 14日	无	
CN	104391748	A	2015年 3月 4日	无	
CN	102932416	A	2013年 2月 13日	无	