

## (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  
2014 年 7 月 3 日 (03.07.2014)



(10) 国际公布号  
WO 2014/101520 A1

- (51) 国际专利分类号:  
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2013/084860
- (22) 国际申请日: 2013 年 10 月 9 日 (09.10.2013)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:  
201210580817.1 2012 年 12 月 27 日 (27.12.2012) CN
- (71) 申请人: 腾讯科技(深圳)有限公司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。
- (72) 发明人: 张书彬 (ZHANG, Shubin); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。 田万鹏 (TIAN, Wanpeng); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。 肖品 (XIAO, Pin); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。 鲍春健 (BAO, Chunjian); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。 郭玮 (GUO, Wei); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。

(74) 代理人: 北京市柳沈律师事务所 (LIU, SHEN & ASSOCIATES); 中国北京市朝阳区北辰东路 8 号汇宾大厦 A0601, Beijing 100101 (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, 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, KZ, LA, LC, LK, LR, LS, LT, 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, 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 SYSTEM FOR ACHIEVING ANALYTIC FUNCTION BASED ON MAPREDUCE

(54) 发明名称: 基于 MapReduce 实现分析函数的方法及系统

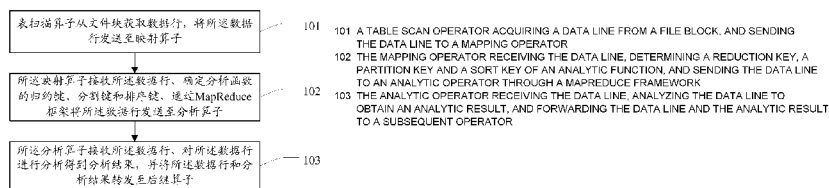


图 1 / Fig. 1

(57) Abstract: Provided are a method and system for achieving an analytic function based on MapReduce. The method comprises: a table scan operator acquiring a data line from a file block, and sending the data line to a mapping operator; the mapping operator receiving the data line, determining a reduction key, a partition key and a sort key of an analytic function, and sending the data line to an analytic operator through a MapReduce framework; and the analytic operator receiving the data line, analyzing the data line to obtain an analytic result, and forwarding the data line and the analytic result to a subsequent operator. The present disclosure can achieve an analytic function in a distributed data warehouse of the MapReduce framework, thereby solving the problem that data analysis processing cannot be performed using the analytic function in the distributed data warehouse based on the MapReduce framework.

(57) 摘要: 本公开提供了一种基于 MapReduce 实现分析函数的方法及系统。所述方法包括: 表扫描算子从文件块获取数据行, 将所述数据行发送至映射算子; 所述映射算子接收所述数据行, 确定分析函数的归约键、分割键和排序键, 通过 MapReduce 框架将所述数据行发送至分析算子; 所述分析算子接收所述数据行, 对所述数据行进行分析得到分析结果, 并将所述数据行和分析结果转发至后继算子。本公开可以在 MapReduce 框架的分布式数据仓库中实现分析函数, 从而解决在基于 MapReduce 框架的分布式数据仓库中无法使用分析函数进行数据处理的问题。



WO 2014/101520 A1

## 基于 MapReduce 实现分析函数的方法及系统

### 技术领域

本公开涉及数据仓库领域，尤其涉及一种基于 MapReduce 实现分析函数的方法及系统。

### 背景技术

数据仓库（Data Warehouse）是按照数据结构来组织、存储和管理数据的仓库。随着计算机的推广，数据仓库已经广泛的应用于工作和生活中。目前，随着互联网及信息技术的快速发展，数据仓库不仅仅是存储和管理数据，且具备了较强的分析数据的能力。常用的数据库，例如 ORACLE、PostgreSQL 等，均提供了多个分析函数，可以根据用户需求对数据进行分析，向用户提供分析结果。分析函数用于计算基于数据组的某种聚集值，与聚集函数不同，分析函数对数据组进行处理后返回多行数据，而聚集函数对数据组进行处理后返回一行数据。

MapReduce 是一种编程模型，用于大规模数据集的并行运算。目前，基于 MapReduce 框架的分布式数据仓库（例如 Hive 数据仓库）无法进行实现分析函数进行数据处理，在数据库的使用过程中带来诸多不便。

### 发明内容

本公开的实施例提供一种基于 MapReduce 实现分析函数的方法及系统，能够解决基于 MapReduce 框架的分布式数据库无法实现分析函数进行数据处理的问题。

为达到上述目的，本公开的实施例采用如下技术方案。

第一方面，本公开实施例提供了一种基于 MapReduce 实现分析函数的方法，所述方法包括：表扫描算子从文件块获取数据行，将所述数据行发送至映射算子；所述映射算子接收所述数据行，确定分析函数的归约键（reduce key）、分割键（partition key）和排序键（sort key），通过 MapReduce 框架将所述数据行发送至分析算子，所述分析算子属于所述 MapReduce 框架的 Reduce 端；所述分析算子接收所述数据行，对所述数据行进行分析得到分析结果，并将所述数据行和分析结果转发至后继算子。

第二方面，本公开实施例还提供了一种基于 MapReduce 实现分析函数的系统，所述系统包括扫描算子模块、映射算子模块和分析算子模块，其中：所述扫描算子模块被配置为从文件块获取数据行，将所述数据行发送至映射算子模块；所述映射算子模块被配置为接收所述数据行，确定分析函数的归约键、分割键和排序键，通过 MapReduce 框架将所述数据行发送至分析算子模块，所述分析算子模块属于所述 MapReduce 框架的 Reduce 端；所述分析算子模块被配置为接收所述数据行，对所述数据行进行分析得到分析结果，并将所述数据行和分析结果转发至后继算子模块。

本公开实施例提供的基于 MapReduce 实现分析函数的方法及系统，能够应用于基于 MapReduce 框架的分布式数据库（例如腾讯分布式数据仓库、Hive 数据库等）以实现数据分析，增加基于 MapReduce 框架的分布式数据库的功能，从而使得用户能够在基于 MapReduce 框架的分布式数据库中进行数据分析。

## 附图说明

为了更清楚地说明本公开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简单地介绍，显而易见地，下面描述中的附图仅仅是本公开的一些实施例，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 为根据本公开实施例一的基于 MapReduce 实现分析函数的方法的流程示意图；

图 2 为根据本公开实施例二的基于 MapReduce 实现分析函数的方法的流程示意图；

图 3 为根据本公开实施例二的分析算子缓冲区的结构示意图；

图 4 为根据本公开实施例二的分析器缓冲区的结构示意图；

图 5(a) - (d) 和图 6(a) - (d) 分别为根据本公开实施例二的窗口模式的示意图；

图 7 为根据本公开实施例三的基于 MapReduce 实现分析函数的系统的结构示意图；

图 8 为图 7 所示的分析算子模块 53 的结构示意图。

## 具体实施方式

下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本公开一部分实施例，而不是全部的实施例。基于本公开中的实施例，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，都属于本公开保护的范围。

### 实施例一

本公开实施例提供了一种基于 MapReduce 实现分析函数的方法，适用于基于 MapReduce 框架的分布式数据仓库进行数据分析，如图 1 所示，所述方法包括以下步骤。

10 步骤 101、表扫描算子（TableScanOperator）从文件块获取数据行，将所述数据行发送至映射算子。

步骤 102、所述映射算子（ReduceSinkOperator）接收所述数据行，确定分析函数的归约键、分割键和排序键，通过 MapReduce 框架将所述数据行发送至分析算子，所述分析算子属于所述 MapReduce 框架的 Reduce 端。

15 步骤 103、所述分析算子（AnalysisOperator）接收所述数据行，对所述数据行进行分析得到分析结果，并将所述数据行和分析结果转发至后继算子。

其中，后继算子可以根据具体情况需要的操作进行确定，例如：聚合算子、过滤算子、或写文件算子等，但不仅限于此。

20 本公开实施例提供的基于 MapReduce 实现分析函数的方法，能够应用于基于 MapReduce 框架的分布式数据仓库（例如腾讯分布式数据仓库、Hive 数据仓库等）进行数据分析分析函数，增加基于 MapReduce 框架的分布式数据库的功能，从而使得在基于 MapReduce 框架的分布式数据库中使用分析函数进行数据分析。

### 25 实施例二

本公开实施例提供了一种基于 MapReduce 实现分析函数的方法，适用于基于 MapReduce 框架的分布式数据库进行数据分析，如图 2 所示，所述方法包括以下步骤。

30 步骤 201、表扫描算子从文件块获取数据行，将所述数据行发送至映射算子。

值得说明的是，在本实施例提供的方法中，可以预设多种不同的分析函

数对数据进行分析，常用的分析函数例如可以包括 LAG、LEAD、RANK、DENSE\_RANK、ROW\_NUMBER、SUM、COUNT、AVG、MAX、MIN、RATIO\_TO\_REPORT 等。可选的，在本实施例提供的方法中，可以根据用户需要添加新的分析函数。

- 5       步骤 202、所述映射算子接收所述数据行，确定分析函数的归约键、分割键和排序键，通过 MapReduce 框架将所述数据行发送至分析算子，所述分析算子属于所述 MapReduce 框架的 Reduce 端。

例如，所述映射算子可以通过如下方法确定分析函数的归约键、分割键和排序键，具体可以包括：

- 10       （1）当所述分析函数包含分区子句和/或排序子句时，可以将所述分析函数的分区子句中的列和/或排序子句中的列作为归约键，或者

当所述分析函数没有排序子句但有 distinct 关键字时，可以以 distinct 列作为归约键，或者

- 15       当分析函数不含分区子句、排序子句，也不含 distinct 关键字时，可以指定任意常量作为归约键；

（2）当所述分析函数包含分区子句时，可以将所述分析函数的分区子句中的列作为分割键，或者

当所述分析函数不含分区子句时，可以以与归约键相同的常量作为分割键。

- 20       （3）当所述分析函数含有排序子句时，可以以排序子句中的列作为排序键。

步骤 203、所述分析算子接收所述数据行，将所述数据行存储于分析算子缓冲区以供所有分析器使用。

- 25       为了实现数据共享，在分析算子（具体地，该分析算子形成的分析算子模块）中可以提供一个分析算子缓冲区 AnalysisBuffer，该缓冲区具备以下特点：a. 允许指定长度的数据保存在内存中；b. 当长度超出限定值后，将原内存缓冲区中的一半内容溢出到硬盘；c. 允许用户按照索引访问其中的元素；d. 允许用户从头开始删除其中已转发的元素。

- 30       具体地，如图 3 所示，分析算子缓冲区可以包括内存缓冲区和磁盘缓冲区（其可以位于图 4 所示的磁盘中）。在所述分析算子缓冲区中，可以优先将接收的新数据行放入内存缓冲区；如果内存缓冲区已满，则可以将内存缓冲

区中较旧的数据行存入所述磁盘缓冲区，以释放内存缓冲区的存储空间，然后可以将接收的新数据行放入内容缓冲区。

步骤 204、所述分析算子解析出所述数据行的分区字段和排序字段，判断所述数据行是否属于当前分区，其中，所述当前分区是所述分析算子接收到的上一数据行所属的分区；若是，执行步骤 205；若否，执行步骤 206。

步骤 205、所述分析算子调用分析函数对应的分析器对所述数据行进行分析，得到分析结果，将所述分析结果存储于分析器缓冲区。

值得说明的是，一个分析函数可以对应一个分析器，每个分析器可以对应一个分析器缓冲区，用于存储与每一数据行相关的分析结果、中间结果或总的聚合结果。如图 4 所示，所述分析器缓冲区可以包括内存缓冲区和磁盘缓冲区（其可位于图 4 所示的磁盘中），所述内存缓冲区可以包括输出缓冲区和输入缓冲区。

所述分析器缓冲区用于对分析结果进行缓冲和更新。具体地，当所述分析器缓冲区对分析结果进行缓冲时：可以将所述分析结果存储于所述输出缓冲区；如果所述输出缓冲区已满，则可以将所述输出缓冲区中的内容存入所述磁盘缓冲区，以释放所述输出缓冲区的存储空间。当所述分析器缓冲区对分析结果进行更新时：如果待更新行存储于输出缓冲区，则可以直接根据所述输出缓冲区中的待更新行和接收到的新数据行对分析结果进行更新；如果待更新行存储于输入缓冲区，则可以直接根据所述输入缓冲区中的待更新行和接收到的新数据行对分析结果进行更新；如果待更新行存储于磁盘（即，磁盘缓冲器），则可以将所述输入缓冲区中的内容存储到所述磁盘，并将所述磁盘中的待更新行所在的缓冲块读入所述输入缓冲区，以使得根据所述输入缓冲区中的待更新行和接收到的新数据行对分析结果进行更新。

步骤 206、所述分析算子结束对所述当前分区的分析，将所述分析算子缓冲区中存储的当前分区的所有数据行、以及所述分析器缓冲区中存储的当前分区的所有分析结果汇总成新的数据行转发至后继算子。

值得说明的是，如果所述分析函数不需要累计，那么在所述调用分析函数对应的分析器对所述数据行进行分析，得到分析结果之后，可以直接将所述数据行和分析结果汇总转发至后继算子，无需对所述数据行和分析结果进行缓存。

为了便于理解，本实施例提供了 11 种常见的分析函数的示例性算法概

述，具体如下。

算法 1: LAG 算法概述:

假设调用的分析函数为 lag(col, offset) over(...).

- LAG 的分析器缓冲区中仅有一个行号计数器 p (初始值为-1)。当分析新  
5 的一行时，将 p 加 1，如果  $p \geq \text{offset}$ ，则将 p 所指向的行的该列设为 p-offset  
行 col 列的内容，并指示 p-offset 行及之前的行的内容可以转发；否则，将当  
前行的结果设为 null，所有行都不得转发。

算法 2: LEAD 算法概述:

假设调用的分析函数为 lead(col, offset) over(...).

- 10 LEAD 的分析器缓冲区中有两个指针，指针 p1 指向当前尚未处理的最小  
行，指针 p2 指向当前行。当分析新的一行时，将指针 p2 加 1，此时，如果  
 $p2 - p1 \geq \text{offset}$ ，则将 p1 所指向行的结果设为 p2 所指行 col 列的内容，且  $p1++$ ，  
行号小于等于 p1 的行均可转发。

算法 3: RANK 算法概述:

- 15 RANK 的分析器缓冲区中有当前序号 rank，当前序号对应的值 value，  
具有当前序号的行数 number。当分析新的一行时，如果新的一行的值与 value  
相等，则将该行的 rank 列设为 rank，分析器缓冲区中的  $\text{number}++$ ；否则，  
将 rank 列设为  $\text{rank} + \text{number}$ ，同时将分析器缓冲区中的 rank 设为  
 $\text{rank} + \text{number}$ ，value 设为新行的指定值，number 设为 1。当前处理后的所有  
20 行均可转发。

算法 4: DENSE\_RANK 算法概述:

- DENSE\_RANK 的分析器缓冲区中有当前序号 rank，当前序号对应的值  
value，具有当前序号的行号 number。当分析新的一行时，如果新的一行的值  
与 value 相等，则将该行的 rank 列设为 rank，分析器缓冲区中的  $\text{number}++$ ；  
25 否则，将 rank 列设为  $\text{rank} + 1$ ，同时将分析器缓冲区中的 rank 设为  $\text{rank} + 1$ ，  
value 设为新行的指定值，number 设为 1。当前处理后的行均可转发。

算法 5: ROW\_NUMBER 算法概述:

- ROW\_NUMBER 的分析器缓冲区中只有一个 rownumber 值 (初始值为  
-1)。当分析新的一行的时候，将新行的 rownumber 列设为  $\text{rownumber} + 1$ ，同  
30 时将分析器缓冲区中的 rownumber 设为  $\text{rownumber} + 1$ 。当前处理后的行均可  
转发。

算法 6: SUM 算法概述:

在 SUM 的分析器缓冲区中, 保存一个变量, 即当前总和 sum。当分析新的一行时, 将 sum 的值加上新行的指定表达式值(需非空)存入 sum 即可。

在整个分区分析完成前不得转发。分区分析完成后, 将 sum 值作为每一行的计算结果即可。

算法 7: COUNT 算法概述:

COUNT 的分析器缓冲区中只有一个 count 计数器。每分析一个新行, 如果待分析列的值非空, 就将该计数器加一。

在整个分区分析完成前不得转发。分区分析完成后, 将 count 值作为每一行的计算结果即可。

算法 8: AVG 算法概述:

AVG 的分析器缓冲区中有两个计数器值, 一个是 sum (初始值为 0), 一个是 count (初始值为 0)。当分析新的一行时, 如果表达式为非空值, count++, sum 设为 sum+新行的表达式值。

在整个分区分析完成前不得转发任一行。分区分析完成后, 如果 count != 0, 将 sum / count 值作为每一行的计算结果即可; 否则, 将 null 作为每一行的分析结果。

算法 9: MAX 算法概述:

MAX 的分析器缓冲区中只有一个 max 值。分析新行时, 将新行的表达式(非空)与 max 比较, 如果比 max 大则更新 max。在分析完分区时, 将所有的行的指定列设为 max 即可。

在整个分区分析完成前不得转发。

算法 10: MIN 算法概述:

MIN 的分析器缓冲区中只有一个 min 值。分析新行时, 将新行的表达式(非空)与 min 比较, 如果比 min 小则更新 min。在分析完分区时, 将所有的行的指定列设为 min 即可。

在整个分区分析完成前不得转发。

算法 11: RATIO\_TO\_REPORT 算法概述:

RATIO\_TO\_REPORT 类的分析器缓冲区中只有一个 sum 值。分析新行时, 将新行的表达式(非空)与 sum 相加设为 sum 的值。在分析完分区时, 用所有的行的指定列分别除以 sum 设为该列的值即可, 如果 sum 为 0, 则均

置为 null。

在整个分区分析完成前不得转发。

值得说明的是，分析函数是基于一组记录（例如多个数据行）为每一行数据计算聚集值得到分析结果的，所基于的这一组记录称之为“窗口”（window）。对于每一行记录，都有一个窗口，用它来指定分析函数执行聚集运算的记录集。针对带窗口子句的情况，本实施例提供了如下 8 种模式（即，窗口模式，具体地，设置窗口位置的模式）以供参考：

模式 1，在图 5（a）中示出：

10 该模式的代表语句为：

Rows between window.lag preceding and window.lead following //位于当前行之之前 window.lag 行和之后 window.lead 行的范围内；

Range between window.lag preceding and window.lead following //比当前值小(或大)window.lag 和比当前值大（或小）window.lead 的范围内。

15 模式 2，在图 5（b）中示出：

该模式的代表语句为：

Rows between window.lag preceding and window.lead preceding //位于当前行之之前 window.lag 行和 window.lead 行的范围内；

20 Range between window.lag preceding and window.lead preceding //比当前值小（或大）window.lag 和 window.lead 的范围内。

模式 3，在图 5（c）中示出：

该模式的代表语句为：

25 Rows between window.lag following and window.lead following //位于当前行之之后 window.lag 行和 window.lead 行的范围内；

Range between window.lag following and window.lead following //比当前值大（或小）window.lag 和 window.lead 的范围内。

模式 4，在图 5（d）中示出：

30 该模式的代表语句为：

Rows between unbounded preceding and window.lead following //从最开

始到当前行之后 window.lead 行的范围内;

Range between unbounded preceding and window.lead following //从最开始到比当前值大(或小) window.lead 的范围内。

5 模式 5, 在图 6(a) 中示出:

该模式的代表语句为:

Rows between window.lag preceding and unbounded following //从当前行之前 window.lag 行到最后的范围内;

10 Range between window.lag preceding and unbounded following //从比当前值小(或大)window.lag 到最后的范围内。

模式 6, 在图 6(b) 中示出:

该模式的代表语句为:

15 Rows between unbounded preceding and unbounded following //从开始到最后;

Range between unbounded preceding and unbounded following //从开始到最后。

模式 7, 在图 6(c) 中示出:

20 该模式的代表语句为:

Rows between unbounded preceding and window.lead preceding //从开始到 window.lead 行之前的范围内;

Range between unbounded preceding and window.lead preceding //从开始到比当前值小(或大)window.lead 的范围内。

25

模式 8, 在图 6(d) 中示出:

该模式的代表语句为:

Rows between window.lag following and unbounded following //从当前行之后 window.lag 行到最后的范围内;

30 Range between window.lag following and unbounded following //从比当前值大(或小)window.lag 到最后的范围内。

根据上述 8 种模式，可以很容易的实现出相应的分析函数处理算法。

本公开实施例提供的基于 MapReduce 实现分析函数的方法，能够应用于基于 MapReduce 框架的分布式数据库（例如腾讯分布式数据仓库、Hive 数据仓库等）以实现数据分析，增加基于 MapReduce 框架的分布式数据库的功能，从而使得在基于 MapReduce 框架的分布式数据库中进行数据分析。

### 实施例三

本公开实施例提供了一种基于 MapReduce 实现分析函数的系统，能够实现上述方法实施例。如图 6 所示，所述系统可以包括扫描算子 51、映射算子 52 和分析算子 53。所述扫描算子 51 可以形成扫描算子模块或被包括在扫描算子模块中，在本实施例中，可互换地使用术语“扫描算子”和“扫描算子模块”。所述映射算子 52 可以形成映射算子模块或被包括在映射算子模块中，在本实施例中，可互换地使用术语“映射算子”和“映射算子模块”。所述分析算子 53 可以形成分析算子模块或被包括在分析算子模块中，在本实施例中，可互换地使用术语“分析算子”和“分析算子模块”。所述系统还可以包括分析算子缓冲区（图中未示出），它们与在上文中描述的分析算子缓冲区相同，因此在这里省略其详细描述。

所述扫描算子 51 用于从文件块获取数据行，将所述数据行发送至映射算子 52；

所述映射算子 52 用于接收所述数据行，确定分析函数的归约键、分割键和排序键，通过 MapReduce 框架将所述数据行发送至分析算子 53，所述分析算子 53 属于所述 MapReduce 框架的 Reduce 端；

所述分析算子 53 接收所述数据行，对所述数据行进行分析得到分析结果，并将所述数据行和分析结果转发至后继算子。

可选的，所述映射算子 52 可以具体用于当所述分析函数包含分区子句和/或排序子句时，将所述分析函数的分区子句中的列和/或排序子句中的列作为归约键，或者所述映射算子 52 还可以用于当所述分析函数没有排序子句但有 distinct 关键字时，以 distinct 列作为归约键，或者所述映射算子 52 还可以用于当分析函数不含分区子句、排序子句，也不含 distinct 关键字时，指定任意常量作为归约键。

所述映射算子 52 还可以用于当所述分析函数包含分区子句时，将所述分

析函数的分区子句中的列作为分割键,或者所述映射算子 52 还可以用于当所述分析函数不含分区子句时, 以与归约键相同的常量作为分割键。

所述映射算子 52 还可以用于当所述分析函数含有排序子句时,以排序子句中的列作为排序键。

5 进一步的, 如图 7 所示, 所述分析算子 53 可以包括:

存储模块 531, 其可以用于接收所述数据行, 将所述数据行存储于分析算子缓冲区以供所有分析器使用;

判断模块 532, 其可以用于解析出所述数据行的分区字段和排序字段, 判断所述数据行是否属于当前分区, 所述当前分区是所述分析算子接收到的  
10 上一数据行所属的分区, 若是, 则所述分析算子 53 可以调用分析函数对应的分析器对所述数据行进行分析, 得到分析结果, 将所述分析结果存储于分析器缓冲区, 若否, 则所述分析算子 53 可以结束对所述当前分区的分析, 将所述分析算子缓冲区中存储的当前分区的所有数据行、以及所述分析器缓冲区中存储的当前分区的所有分析结果汇总成新的数据行转发至后继算子 (即,  
15 算子模块)。所述分析器和分析器缓冲区与上文所述相同, 它们可以位于根据本发明实施例三的系统, 也可以位于所述系统之外并且可操作地耦接到所述系统。

可选的, 如果所述分析函数不需要累计, 那么所述分析算子 53 可以在得到分析结果之后, 直接将所述数据行和分析结果汇总转发至后继算子 (即,  
20 算子模块), 无需对所述数据行和分析结果进行缓存。

本公开实施例提供的基于 MapReduce 实现分析函数的系统, 能够应用于基于 MapReduce 框架的分布式数据库 (例如腾讯分布式数据仓库、Hive 数据库等) 以实现数据分析, 增加基于 MapReduce 框架的分布式数据库的功能, 从而使得在基于 MapReduce 框架的分布式数据库中实现分析函数进行数据  
25 分析。

通过以上的实施方式的描述, 所属领域的技术人员可以清楚地了解到本公开可借助软件加必需的通用硬件的方式来实现, 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现有技术做出贡献的部分可以以软件产品的形式体现出来, 该  
30 计算机软件产品存储在可读取的存储介质中, 如计算机的软盘, 硬盘或光盘等, 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)执行本公开各个实施例所述的方法。

以上所述, 仅为本发明的具体实施方式, 但本发明的保护范围并不局限于此, 任何熟悉本技术领域的技术人员在本发明揭露的技术范围内, 可轻易想到变化或替换, 都应涵盖在本发明的保护范围之内。因此, 本发明的保护

5 范围应以所述权利要求的保护范围为准。

## 权利要求书

1、一种基于 MapReduce 实现分析函数的方法，包括：

表扫描算子从文件块获取数据行，将所述数据行发送至映射算子；

- 5        所述映射算子接收所述数据行，确定分析函数的归约键、分割键和排序键，通过 MapReduce 框架将所述数据行发送至分析算子，所述分析算子属于所述 MapReduce 框架的 Reduce 端；

所述分析算子接收所述数据行，对所述数据行进行分析得到分析结果，并将所述数据行和分析结果转发至后继算子。

- 10       2、根据权利要求 1 所述的方法，其中，所述确定分析函数的归约键、分割键和排序键，包括：

当所述分析函数包含分区子句和/或排序子句时，将所述分析函数的分区子句中的列和/或排序子句中的列作为归约键，或者

- 15       当所述分析函数没有排序子句但有 distinct 关键字时，以 distinct 列作为归约键，或者

当分析函数不含分区子句、排序子句，也不含 distinct 关键字时，指定任意常量作为归约键；

当所述分析函数包含分区子句时，将所述分析函数的分区子句中的列作为分割键，或者

- 20       当所述分析函数不含分区子句时，以与归约键相同的常量作为分割键；  
当所述分析函数含有排序子句时，以排序子句中的列作为排序键。

3、根据权利要求 1 或 2 所述的方法，其中，所述分析算子接收所述数据行，对所述数据行进行分析得到分析结果，并将所述数据行和分析结果转发至后继算子，包括：

- 25       所述分析算子接收所述数据行，将所述数据行存储于分析算子缓冲区以供所有分析器使用；

所述分析算子解析出所述数据行的分区字段和排序字段，判断所述数据行是否属于当前分区，所述当前分区是所述分析算子接收到的上一数据行所属的分区，

- 30       若是，则调用分析函数对应的分析器对所述数据行进行分析，得到分析结果，将所述分析结果存储于分析器缓冲区，

若否，则结束对所述当前分区分析，将所述分析算子缓冲区中存储的当前分区的所有数据行、以及所述分析器缓冲区中存储的当前分区的所有分析结果汇总成新的数据行转发至后继算子。

4、根据权利要求3所述的方法，其中，如果所述分析函数不需要累计，  
5 那么在所述调用分析函数对应的分析器对所述数据行进行分析，得到分析结果之后，直接将所述数据行和分析结果汇总转发至后继算子，无需对所述数据行和分析结果进行缓存。

5、根据权利要求3所述的方法，其中，所述分析算子缓冲区包括内存缓冲区和磁盘缓冲区，所述分析算子缓冲区优先将接收的新数据行放入内存缓冲区，  
10 如果内存缓冲区已满，则将内存缓冲区中较旧的数据行存入所述磁盘缓冲区，以释放内存缓冲区的存储空间。

6、根据权利要求3所述的方法，其中，所述分析器缓冲区包括内存缓冲区和磁盘缓冲区，所述内存缓冲区包括输出缓冲区和输入缓冲区，所述分析器缓冲区用于对分析结果进行缓冲和更新；

15 所述分析器缓冲区对分析结果进行缓冲时，将所述分析结果存储于所述输出缓冲区，如果所述输出缓冲区已满，则将所述输出缓冲区中的内容存入所述磁盘缓冲区，以释放所述输出缓冲区的存储空间；

所述分析器缓冲区对分析结果进行更新时：

如果待更新行存储于输出缓冲区，则直接根据所述输出缓冲区中的  
20 待更新行和接收到的新数据行对分析结果进行更新，

如果待更新行存储于输入缓冲区，则直接根据所述输入缓冲区中的待更新行和接收到的新数据行对分析结果进行更新，

如果待更新行存储于磁盘缓冲区，则将所述输入缓冲区中的内容存储到所述磁盘缓冲区，并将所述磁盘缓冲区中的待更新行所在的缓冲块读入  
25 所述输入缓冲区，以使得根据所述输入缓冲区中的待更新行和接收到的新数据行对分析结果进行更新。

7、一种基于 MapReduce 实现分析函数的系统，包括扫描算子模块、映射算子模块和分析算子模块，其中：

所述扫描算子被配置为从文件块获取数据行，将所述数据行发送至映射  
30 算子；

所述映射算子被配置为接收所述数据行，确定分析函数的归约键、分割

键和排序键，通过 MapReduce 框架将所述数据行发送至分析算子，所述分析算子属于所述 MapReduce 框架的 Reduce 端；

所述分析算子被配置为接收所述数据行，对所述数据行进行分析得到分析结果，并将所述数据行和分析结果转发至后继算子模块。

5        8、根据权利要求 7 所述的系统，其中，所述映射算子模块被配置为：

当所述分析函数包含分区子句和/或排序子句时，将所述分析函数的分区子句中的列和/或排序子句中的列作为归约键，或者

当所述分析函数没有排序子句但有 distinct 关键字时，以 distinct 列作为归约键，或者

10        当分析函数不含分区子句、排序子句，也不含 distinct 关键字时，指定任意常量作为归约键；

所述映射算子模块还被配置为：

当所述分析函数不含分区子句时将所述分析函数的分区子句中的列作为分割键，或者

15        当所述分析函数不含分区子句时，以与归约键相同的常量作为分割键；

所述映射算子还被配置为当所述分析函数含有排序子句时，以排序子句中的列作为排序键。

9、根据权利要求 7 或 8 所述的系统，其中，所述分析算子模块包括：

20        存储模块，被配置为接收所述数据行，将所述数据行存储于分析算子缓冲区以供所有分析器使用；

判断模块，被配置为解析出所述数据行的分区字段和排序字段，判断所述数据行是否属于当前分区，所述当前分区是所述分析算子接收到的上一数据行所属的分区，

25        若是，则所述分析算子调用分析函数对应的分析器对所述数据行进行分析，得到分析结果，将所述分析结果存储于分析器缓冲区；

若否，则所述分析算子结束对所述当前分区的分析，将所述分析算子缓冲区中存储的当前分区的所有数据行、以及所述分析器缓冲区中存储的当前分区的所有分析结果汇总成新的数据行转发至后继算子模块。

30        10、根据权利要求 9 所述的系统，其中，如果所述分析函数不需要累计，那么所述分析算子在得到分析结果之后，直接将所述数据行和分析结果汇总

转发至后继算子，无需对所述数据行和分析结果进行缓存。

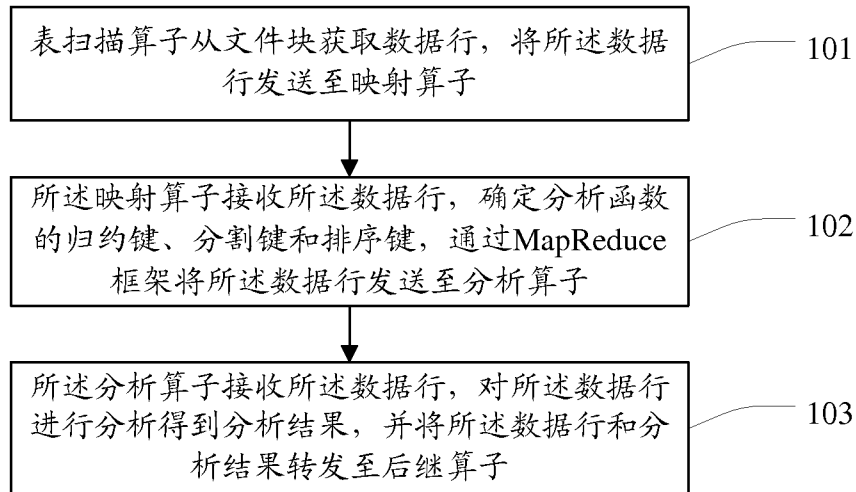


图 1

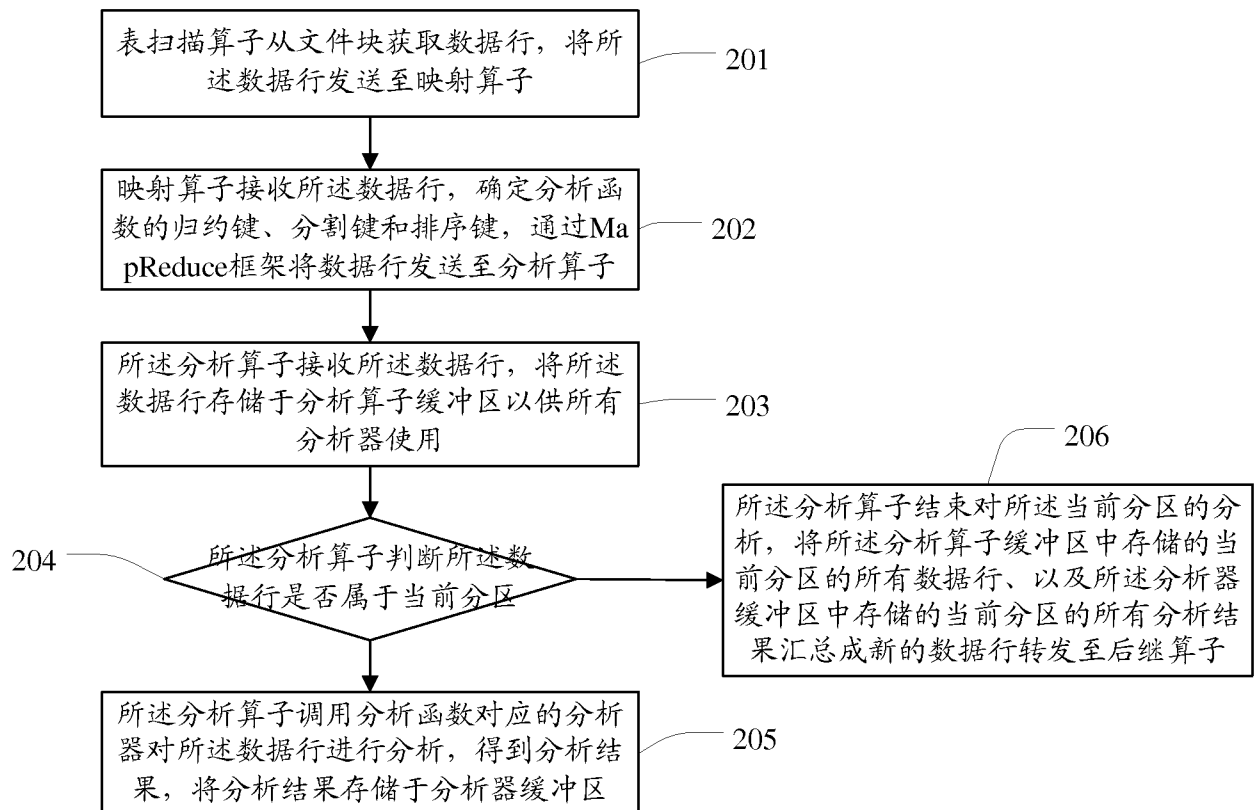


图 2

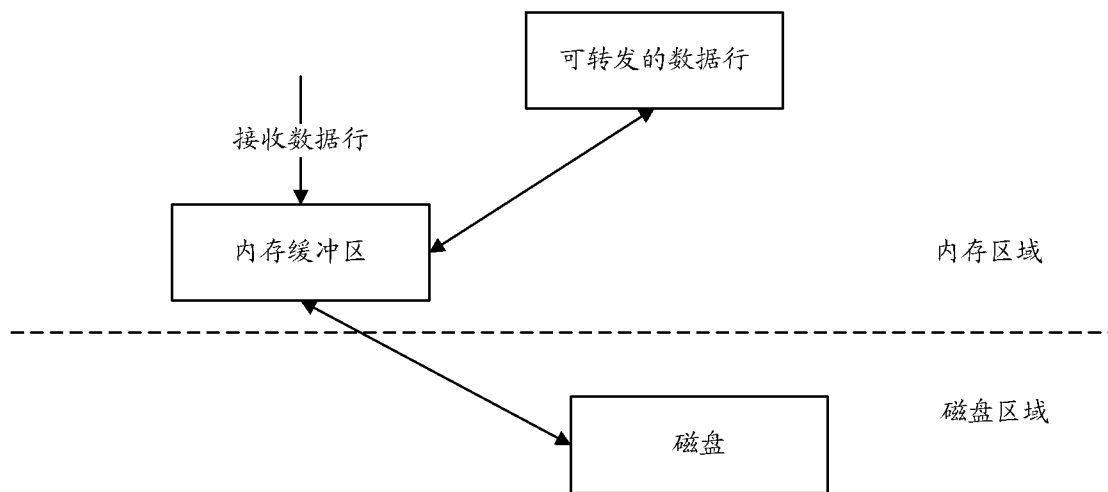


图 3

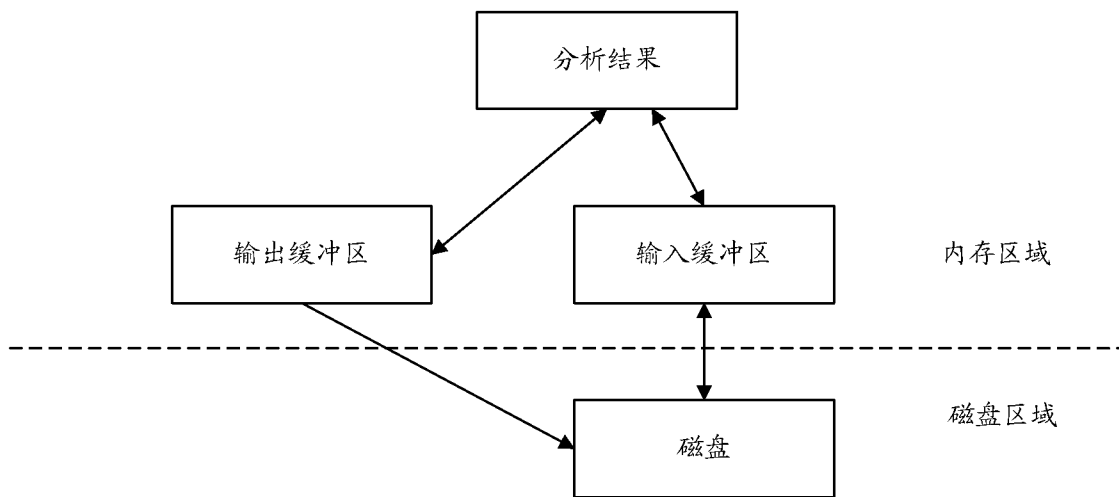


图 4

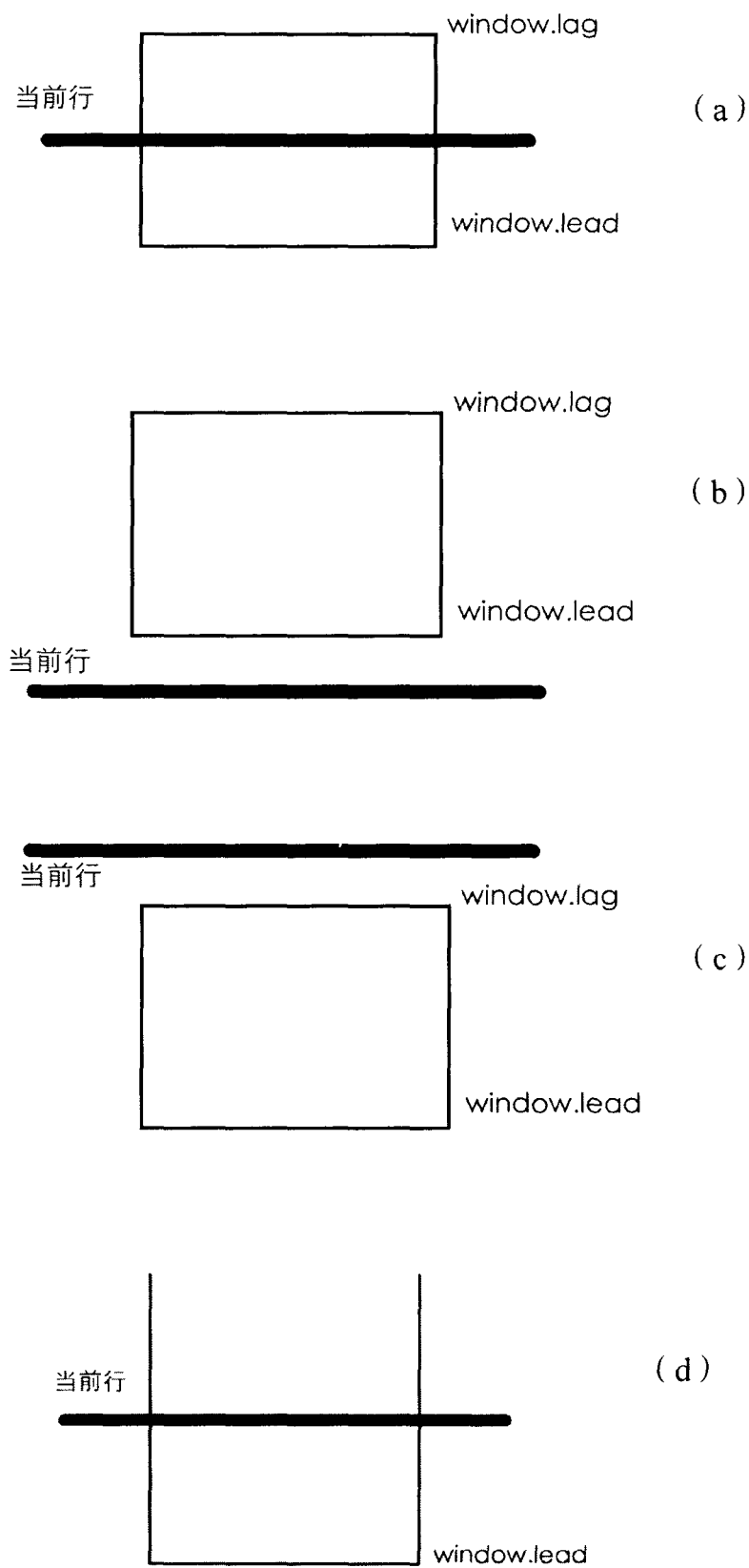


图 5

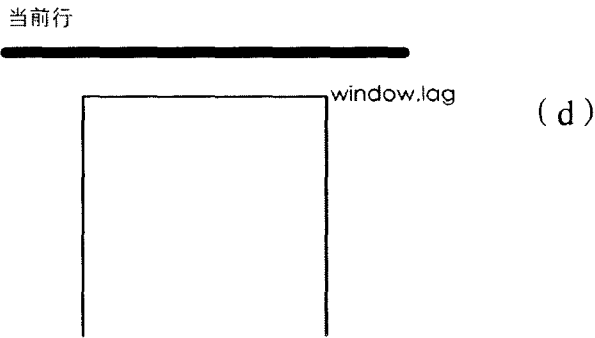
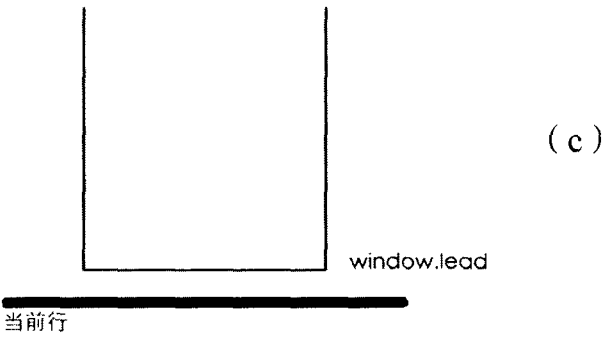
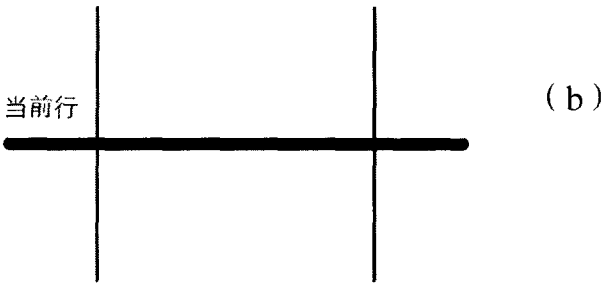
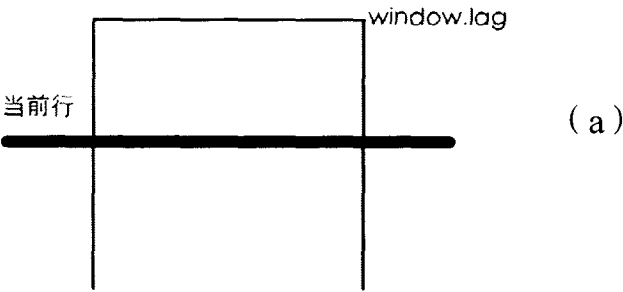


图 6

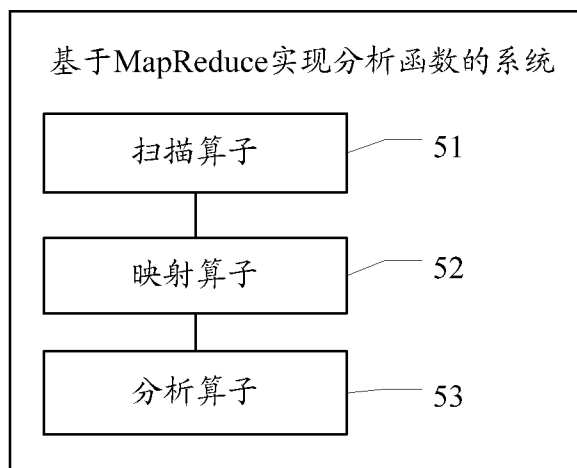


图 7

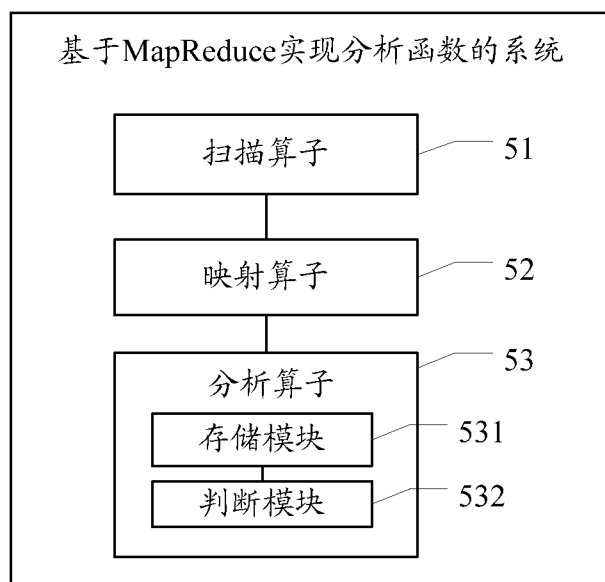


图 8

# INTERNATIONAL SEARCH REPORT

International application No.

**PCT/CN2013/084860**

## 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)

IPC: G06F, G06K

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; CNKI; WPI; EPODOC: protocol, mapreduce, map, reduce, partition?, sort+, analy+

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                              | Relevant to claim No. |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| X         | LANG, Weimin et al., MapReduce technology of cloud computing, TELECOMMUNICATIONS INFORMATION, 31 March 2012, no. 3, pages 3-5 and 12                            | 1, 7                  |
| A         |                                                                                                                                                                 | 2-6, 8-10             |
| A         | ZHANG, Shubin et al., Research on implementing spatial queries based on MapReduce, CHINESE HIGH TECHNOLOGY LETTERS, 31 July 2010, vol. 20, no. 7, pages 719-725 | 1-10                  |
| A         | CN 102129457 A (ZHEJIANG UNIVERSITY), 20 July 2011 (20.07.2011), see the whole document                                                                         | 1-10                  |
| A         | US 2012254193 A1 (GOOGLE INC.), 04 October 2012 (04.10.2012), see the whole document                                                                            | 1-10                  |

☐ 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<br><b>06 January 2014 (06.01.2014)</b>                                                                                                                   | Date of mailing of the international search report<br><b>16 January 2014 (16.01.2014)</b> |
| Name and mailing address of the ISA/CN:<br>State Intellectual Property Office of the P. R. China<br>No. 6, Xitucheng Road, Jimenqiao<br>Haidian District, Beijing 100088, China<br>Facsimile No.: (86-10) 62019451 | Authorized officer<br><b>HAN, Yan</b><br>Telephone No.: (86-10) <b>62411764</b>           |

**INTERNATIONAL SEARCH REPORT**  
Information on patent family members

International application No.

**PCT/CN2013/084860**

| Patent Documents referred<br>in the Report | Publication Date | Patent Family    | Publication Date |
|--------------------------------------------|------------------|------------------|------------------|
| CN 102129457 A                             | 20.07.2011       | None             |                  |
| US 2012254193 A1                           | 04.10.2012       | AU 2012236577 A1 | 24.10.2013       |
|                                            |                  | WO 2012135319 A1 | 04.10.2012       |

# 国际检索报告

国际申请号

**PCT/CN2013/084860**

## 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)

IPC: G06F, G06K

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; CNKI; WPI; EPODOC:映射, 规约, 分割, 排序, 分析, mapreduce, map, reduce, partition?, sort+, analy+

## 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    | 郎为民等, 云计算中的MapReduce技术, 电信快报, 31.03月2012, 第3期, 第3-5页, 第12页                  | 1,7      |
| A    |                                                                             | 2-6,8-10 |
| A    | 张书彬等, 基于 MapReduce 实现空间查询的研究, 高技术通讯, 31.07 月 2010, 第 20 卷第 7 期, 第 719-725 页 | 1-10     |
| A    | CN 102129457 A(浙江大学) 20.07 月 2011 (20.07.2011) 参见全文                         | 1-10     |
| A    | US2012254193 A1(GOOGLE INC) 04.10 月 2012 (04.10.2012) 参见全文                  | 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

**06.1 月 2014(06.01.2014)**

国际检索报告邮寄日期

**16.1 月 2014 (16.01.2014)**

ISA/CN 的名称和邮寄地址:

中华人民共和国国家知识产权局

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

**韩燕**

电话号码: (86-10) **62411764**

# 国际检索报告

关于同族专利的信息

国际申请号

**PCT/CN2013/084860**

| 检索报告中引用的<br>专利文件 | 公布日期       | 同族专利             | 公布日期       |
|------------------|------------|------------------|------------|
| CN 102129457 A   | 20.07.2011 | 无                |            |
| US2012254193 A1  | 04.10.2012 | AU 2012236577 A1 | 24.10.2013 |
|                  |            | WO2012135319 A1  | 04.10.2012 |