(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108920516 A
(43)申请公布日 2018.11.30

(21)申请号 201810555976.3

(22)申请日 2018.05.31

(71)申请人 北京字节跳动网络技术有限公司
地址 100041 北京市石景山区实兴大街30号院3号楼2层B-0035房间

(72)发明人 沈维海 高俊秀 谭奇

(74)专利代理机构 北京中原华和知识产权代理有限责任公司 11019
代理人 寿宁 张华辉

(51)Int.Cl.
*G06F 17/30*(2006.01)

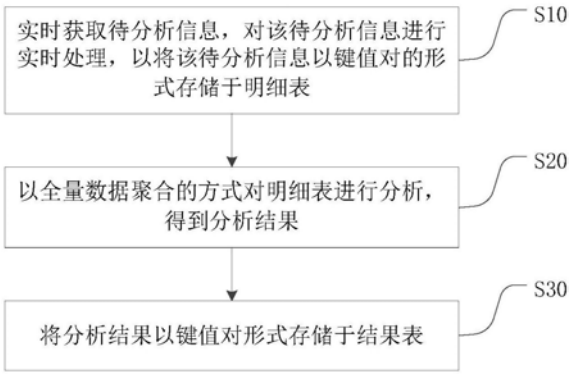
权利要求书3页 说明书12页 附图5页

(54)发明名称

实时分析方法、系统、装置及计算机可读存储介质

(57)摘要

本公开涉及一种实时分析方法、系统、装置及计算机可读存储介质，该方法包括：实时获取待分析信息，对所述待分析信息进行实时处理，以将所述待分析信息以键值对形式存储于明细表；以全量数据聚合的方式对所述明细表进行分析，得到分析结果；将所述分析结果以键值对形式存储于结果表，以供查询。




CN 108920516 A

1.一种实时分析方法,所述方法包括:

实时获取待分析信息,对所述待分析信息进行实时处理,以将所述待分析信息以键值对形式存储于明细表;

以全量数据聚合的方式对所述明细表进行分析,得到分析结果;

将所述分析结果以键值对形式存储于结果表,以供查询。

2.根据权利要求1所述的实时分析方法,其中,所述待分析信息包含指纹信息和明细信息,所述指纹信息包括设备标识、应用程序标识或用户标识的一种或多种,所述明细信息包括一种或多种维度类别以及每种所述维度类别所取的维度值。

3.根据权利要求2所述的实时分析方法,其中,所述的将所述待分析信息以键值对形式存储于明细表包括:

将所述指纹信息记录于所述明细表中一个键值对的键字段,将与所述指纹信息对应的所述明细信息记录于所述明细表中同一个键值对的值字段。

4.根据权利要求3所述的实时分析方法,其中,所述的以全量数据聚合的方式对所述明细表进行分析,得到分析结果包括:

在所有的筛选条件下,以全量数据聚合的方式对所述明细表进行分析,得到每个所述筛选条件下的分析结果;其中,每个所述筛选条件包含一项或多项条件项,所述条件项包括任意的所述维度值。

5.根据权利要求4所述的实时分析方法,其中,所述的以全量数据聚合的方式对所述明细表进行分析包括:

将所述明细表划分为多个第一数据集群,且将具有相同所述指纹信息的所述待分析信息划分在同一个所述第一数据集群;

分布式并发地对每个所述第一数据集群进行第一聚合,得到每个所述第一数据集群的第一聚合结果,用以在进行聚合的同时对所述具有相同指纹信息的所述待分析信息进行去重;

对所有的所述第一聚合结果进行第二聚合以得出分析结果。

6.根据权利要求5所述的实时分析方法,其中,

所述的将所述待分析信息以键值对形式存储于明细表还包括:将第一数据集群标识记录于所述明细表中的键字段,所述第一数据集群标识为所述指纹信息对第一数据集群总数取模的结果;

所述的将所述明细表划分为多个第一数据集群,且将具有相同指纹信息的所述待分析信息划分在同一个所述第一数据集群包括:根据所述第一数据集群信息将所述明细表中的所述键值对划分到多个第一数据集群中。

7.根据权利要求4所述的实时分析方法,其中,所述的将所述分析结果以键值对形式存储于结果表包括:

将所述分析结果记录于所述结果表中一个键值对的值字段,将对应的所述筛选条件记录于所述结果表中同一个键值对的键字段。

8.根据权利要求7所述的实时分析方法,其中,所述的将所述分析结果以键值对形式存储于结果表还包括:

将第二数据集群信息记录于所述结果表中的键字段,用以根据所述第二数据集群信息

将所述分析结果分散到多个第二数据集群中。

9.根据权利要求7所述的实时分析方法，其中，所述的将所述分析结果以键值对形式存储于结果表还包括：

将所述指纹信息之中的一种或多种记录于所述结果表中的键字段。

10.根据权利要求1所述的实时分析方法，其中：

获取的所述待分析信息还包括时间信息；

所述的将所述待分析信息以键值对形式存储于明细表包括，将所述时间信息记录于所述明细表中的键字段。

11.根据权利要求10所述的实时分析方法，其中，所述的将所述待分析信息以键值对形式存储于明细表还包括：

将所述待分析信息存储到具有不同时间跨度的多个明细表之中。

12.根据权利要求10所述的实时分析方法，其中，

所述的以全量数据聚合的方式对所述明细表进行分析，得到分析结果包括：对于一个所述明细表，按照多种时间跨度进行所述分析，以得到与多种时间跨度对应的多个所述分析结果；

所述的将所述分析结果以键值对形式存储于结果表包括：将所述的与多种时间跨度对应的多个分析结果分别存储于多个所述结果表，和/或将时间跨度信息存储于所述结果表中的键字段。

13.根据权利要求10所述的实时分析方法，其中，所述的以全量数据聚合的方式对所述明细表进行分析，得到分析结果包括：

在多个时间窗口，对所述明细表进行所述分析，以得到分时段的多个分析结果。

14.根据权利要求1所述的实时分析方法，其中，所述的对所述待分析信息进行实时处理包括：将所述待分析信息上报到实时流，利用实时流数据处理框架对所述实时流进行实时处理。

15.根据权利要求1所述的实时分析方法，还包括：

获取待查询的筛选条件；

根据所述待查询的筛选条件对所述结果表进行查询，得到与所述待查询的筛选条件对应的分析结果作为查询结果。

16.一种实时分析系统，所述系统包括：

明细表确定模块，用于实时获取待分析信息，对所述待分析信息进行实时处理，以将所述待分析信息以键值对形式存储于明细表；

分析模块，用于以全量数据聚合的方式对所述明细表进行分析，得到分析结果；

结果表确定模块，用于将所述分析结果以键值对形式存储于结果表，以供查询。

17.根据权利要求16所述的实时分析系统，所述系统还包括执行权利要求2到15中任一权利要求所述步骤的模块。

18.一种实时分析装置，包括：

存储器，用于存储非暂时性计算机可读指令；以及

处理器，用于运行所述计算机可读指令，使得所述计算机可读指令被所述处理器执行时实现根据权利要求1到15中任意一项所述的实时分析方法。

19.一种计算机可读存储介质，用于存储非暂时性计算机可读指令，当所述非暂时性计算机可读指令由计算机执行时，使得所述计算机执行权利要求1到15中任意一项所述的实时分析方法。

# 实时分析方法、系统、装置及计算机可读存储介质

## 技术领域

[0001] 本公开涉及计算机技术领域，特别是涉及一种实时分析方法、系统、装置及计算机可读存储介质。

## 背景技术

[0002] 在实时分析领域，现有的实现框架或者系统存在着诸多问题，例如：无法对维度进行细粒度拆分分析；对日登录用户数(Daily Login User，简称为DLU)、日活跃用户数(Daily Activation User，简称为DAU)、日新增用户数(Daily New User，简称为DNU)等指标的统计分析基于基数估算方法，无法得到精确的分析结果；对于多维交叉查询的局限太多。

## 发明内容

[0003] 本公开的目的在于提供一种新的实时分析方法、系统、装置及计算机可读存储介质。

[0004] 本公开的目的是采用以下的技术方案来实现的。依据本公开提出的实时分析方法，包括以下步骤：实时获取待分析信息，对所述待分析信息进行实时处理，以将所述待分析信息以键值对形式存储于明细表；以全量数据聚合的方式对所述明细表进行分析，得到分析结果；将所述分析结果以键值对形式存储于结果表，以供查询。

[0005] 本公开的目的还可以采用以下的技术措施来进一步实现。

[0006] 前述的实时分析方法，其中，所述待分析信息包含指纹信息和明细信息，所述指纹信息包括设备标识、应用程序标识或用户标识的一种或多种，所述明细信息包括一种或多种维度类别以及每种所述维度类别所取的维度值。

[0007] 前述的实时分析方法，其中，所述的将所述待分析信息以键值对形式存储于明细表包括：将所述指纹信息记录于所述明细表中一个键值对的键字段，将与所述指纹信息对应的所述明细信息记录于所述明细表中同一个键值对的值字段。

[0008] 前述的实时分析方法，其中，所述的以全量数据聚合的方式对所述明细表进行分析，得到分析结果包括：在所有的筛选条件下，以全量数据聚合的方式对所述明细表进行分析，得到每个所述筛选条件下的分析结果；其中，每个所述筛选条件包含一项或多项条件项，所述条件项包括任意的所述维度值。

[0009] 前述的实时分析方法，其中，所述的以全量数据聚合的方式对所述明细表进行分析包括：将所述明细表划分为多个第一数据集群，且将具有相同所述指纹信息的所述待分析信息划分在同一个所述第一数据集群；分布式并发地对每个所述第一数据集群进行第一聚合，得到每个所述第一数据集群的第一聚合结果，用以在进行聚合的同时对所述具有相同指纹信息的所述待分析信息进行去重；对所有的所述第一聚合结果进行第二聚合以得出分析结果。

[0010] 前述的实时分析方法，其中，所述的将所述待分析信息以键值对形式存储于明细

表还包括:将第一数据集群标识记录于所述明细表中的键字段,所述第一数据集群标识为所述指纹信息对第一数据集群总数取模的结果;所述的将所述明细表划分为多个第一数据集群,且将具有相同指纹信息的所述待分析信息划分在同一个所述第一数据集群包括:根据所述第一数据集群信息将所述明细表中的所述键值对划分到多个第一数据集群中。

[0011] 前述的实时分析方法,其中,所述的将所述分析结果以键值对形式存储于结果表包括:将所述分析结果记录于所述结果表中一个键值对的值字段,将对应的所述筛选条件记录于所述结果表中同一个键值对的键字段。

[0012] 前述的实时分析方法,其中,所述的将所述分析结果以键值对形式存储于结果表还包括:将第二数据集群信息记录于所述结果表中的键字段,用以根据所述第二数据集群信息将所述分析结果分散到多个第二数据集群中。

[0013] 前述的实时分析方法,其中,所述的将所述分析结果以键值对形式存储于结果表还包括:将所述指纹信息之中的一种或多种记录于所述结果表中的键字段。

[0014] 前述的实时分析方法,其中,获取的所述待分析信息还包括时间信息;所述的将待分析信息以键值对的形式存储于明细表包括,将所述时间信息记录于所述明细表中的键字段。

[0015] 前述的实时分析方法,其中,所述的将所述待分析信息以键值对形式存储于明细表包括:将所述待分析信息存储到具有不同时间跨度的多个明细表之中。

[0016] 前述的实时分析方法,其中,所述的以全量数据聚合的方式对所述明细表进行分析,得到分析结果包括:对于一个所述明细表,按照多种时间跨度进行所述分析,以得到与多种时间跨度对应的多个所述分析结果;所述的将所述分析结果以键值对形式存储于结果表包括:将所述的与多种时间跨度对应的多个分析结果分别存储于多个所述结果表,和/或将时间跨度信息存储于所述结果表中的键字段。

[0017] 前述的实时分析方法,其中,所述的以全量数据聚合的方式对所述明细表进行分析,得到分析结果包括:在多个时间窗口,对所述明细表进行所述分析,以得到分时段的多个分析结果。

[0018] 前述的实时分析方法,其中,所述的对所述待分析信息进行实时处理包括:将所述待分析信息上报到实时流,利用实时流数据处理框架对所述实时流进行实时处理。

[0019] 前述的实时分析方法,还包括:获取待查询的筛选条件;根据所述待查询的筛选条件对所述结果表进行查询,得到与所述待查询的筛选条件对应的分析结果作为查询结果。

[0020] 本公开的目的还采用以下技术方案来实现。依据本公开提出的实时分析系统,包括:明细表确定模块,用于实时获取待分析信息,对所述待分析信息进行实时处理,以将所述待分析信息以键值对形式存储于明细表;分析模块,用于以全量数据聚合的方式对所述明细表进行分析,得到分析结果;结果表确定模块,用于将所述分析结果以键值对形式存储于结果表,以供查询。

[0021] 本公开的目的还可以采用以下的技术措施来进一步实现。

[0022] 前述的实时分析系统,其中,所述待分析信息包含指纹信息和明细信息,所述指纹信息包括设备标识、应用程序标识或用户标识的一种或多种,所述明细信息包括一种或多种维度类别以及每种所述维度类别所取的维度值。

[0023] 前述的实时分析系统,其中,所述明细表确定模块包括第一记录子模块,用于将所

述指纹信息记录于所述明细表中一个键值对的键字段，将与所述指纹信息对应的所述明细信息记录于所述明细表中同一个键值对的值字段。

[0024] 前述的实时分析系统，其中，所述分析模块具体用于：在所有的筛选条件下，以全量数据聚合的方式对所述明细表进行分析，得到每个所述筛选条件下的分析结果；其中，每个所述筛选条件包含一项或多项条件项，所述条件项包括任意的所述维度值。

[0025] 前述的实时分析系统，其中，所述分析模块包括：第一数据集群划分单元，用于将所述明细表划分为多个第一数据集群，且将具有相同所述指纹信息的所述待分析信息划分在同一个所述第一数据集群；第一聚合单元，用于分布式并发地对每个所述第一数据集群进行第一聚合，得到每个所述第一数据集群的第一聚合结果，用以在进行聚合的同时对所述具有相同指纹信息的所述待分析信息进行去重；第二聚合单元，用于对所有的所述第一聚合结果进行第二聚合以得出分析结果。

[0026] 前述的实时分析系统，其中，所述的明细表确定模块还包括第二记录子模块，用于将第一数据集群标识记录于明细表中的键字段，所述第一数据集群标识为所述指纹信息对第一数据集群总数取模的结果；所述的第一数据集群划分单元具体用于，根据所述第一数据集群信息将所述明细表中的所述键值对划分到多个第一数据集群中。

[0027] 前述的实时分析系统，其中，所述的结果表确定模块还包括第三记录子模块，用于：将所述分析结果记录于所述结果表中一个键值对的值字段，将对应的所述筛选条件记录于所述结果表中同一个键值对的键字段。

[0028] 前述的实时分析系统，其中，所述的结果表确定模块还包括第四记录子模块，用于将第二数据集群信息记录于结果表中的键字段，用以根据所述第二数据集群信息将所述分析结果分散到多个第二数据集群中。

[0029] 前述的实时分析系统，其中，所述的结果表确定模块还包括第五记录子模块，用于将所述指纹信息之中的一种或多种记录于结果表中的键字段。

[0030] 前述的实时分析系统，其中，获取的所述待分析信息还包括时间信息；所述的明细表确定模块还包括第六记录子模块，用于将所述时间信息记录于所述明细表中的键字段。

[0031] 前述的实时分析系统，其中，所述的明细表确定模块还包括第七记录子模块，用于将所述待分析信息存储到具有不同时间跨度的多个明细表之中。

[0032] 前述的实时分析系统，其中，所述的分析模块包括第一分析子模块，用于对于一个所述明细表，按照多种时间跨度进行所述分析，以得到与多种时间跨度对应的多个分析结果；所述的结果表确定模块具体用于，将所述的与多种时间跨度对应的多个分析结果分别存储于多个结果表，和/或将时间跨度信息存储于结果表中的键字段。

[0033] 前述的实时分析系统，其中，所述的分析模块包括第二分析子模块，用于：在多个时间窗口，对所述明细表进行所述分析，以得到分时段的多个分析结果。

[0034] 前述的实时分析系统，其中，所述的明细表确定模块包括实施流处理子模块，用于将所述待分析信息上报到实时流，利用实时流数据处理框架对所述实时流进行实时处理。

[0035] 前述的实时分析系统，还包括：查询条件获取模块，获取待查询的筛选条件；查询模块，用于根据所述待查询的筛选条件对所述结果表进行查询，得到与所述待查询的筛选条件对应的分析结果作为查询结果。

[0036] 本公开的目的还采用以下技术方案来实现。依据本公开提出的一种实时分析装

置，包括：存储器，用于存储非暂时性计算机可读指令；以及处理器，用于运行所述计算机可读指令，使得所述处理器执行时实现前述任意一种实时分析方法。

[0037] 本公开的目的还采用以下技术方案来实现。依据本公开提出的一种计算机可读存储介质，用于存储非暂时性计算机可读指令，当所述非暂时性计算机可读指令由计算机执行时，使得所述计算机执行前述任意一种实时分析方法。

[0038] 本公开的目的还采用以下技术方案来实现。依据本公开提出的一种终端设备，包括前述任意一种实时分析系统。

[0039] 上述说明仅是本公开技术方案的概述，为了能更清楚了解本公开的技术手段，而可依照说明书的内容予以实施，并且为让本公开的上述和其他目的、特征和优点能够更明显易懂，以下特举较佳实施例，并配合附图，详细说明如下。

## 附图说明

[0040] 图1是本公开一个实施例的实时分析方法的流程框图。

[0041] 图2是本公开一个实施例提供的以全量数据分布式聚合的方式对明细表进行分析的流程示意图。

[0042] 图3是本公开一个实施例提供的以全量数据分布式聚合的方式对明细表进行分析的流程框图。

[0043] 图4是本公开一个实施例的实时分析系统的结构框图。

[0044] 图5是本公开一个实施例提供的分析模块的结构框图。

[0045] 图6是本公开一个实施例的实时分析装置的硬件框图。

[0046] 图7是本公开一个实施例的计算机可读存储介质的示意图。

[0047] 图8是本公开一个实施例的终端设备的结构框图。

## 具体实施方式

[0048] 为更进一步阐述本公开为达成预定发明目的所采取的技术手段及功效，以下结合附图及较佳实施例，对依据本公开提出的实时分析方法、系统、装置及计算机可读存储介质的具体实施方式、结构、特征及其功效，详细说明如后。

[0049] 图1为本公开的实时分析方法一个实施例的示意性流程框图。请参阅图1，本公开示例的实时分析方法，主要包括以下步骤：

[0050] 步骤S10，实时获取待分析信息，对该待分析信息进行实时处理，以将该待分析信息以键值对(key-value对，简称kv对)的形式存储于明细表。此后，处理进到步骤S20。

[0051] 步骤S20，以全量数据聚合的方式对明细表进行分析，得到分析结果。此后，处理进到步骤S30。

[0052] 步骤S30，将分析结果以键值对形式存储于结果表，以供查询。

[0053] 本公开提出的实时分析方法，通过以全量聚合的方式对明细表中键值对形式的数据进行实时统计分析，能够大大提高实时分析的准确性和效率。

[0054] 值得注意的是，不同于以往的关系型的数据存储形式(例如MySQL数据库)，键值对形式的数据存储形式无法支持诸如条件查询的各种复杂操作。因此需要对键值对形式的明细表和结果表进行精心设计，以能够在仅使用简单操作的情况下来模拟关系类数据库进行

的复杂统计操作，得到准确的统计分析结果并支持多维交叉查询。

[0055] 具体地，该待分析信息中的每条信息包含指纹信息以及明细信息。其中，该指纹信息包括设备标识(device id)、应用程序标识(app id)或用户标识(user id)之中的一种或多种。事实上，该指纹信息可以仅是上述的多种标识中的一种，可以是由多种标识组合而成，也可以是根据多种标识中的一个或多个并利用特定算法而生成的。另外需要注意的是，待分析信息可以同时包含设备标识、应用程序标识和用户标识，但仅将其中的一个作为指纹信息，例如仅将设备标识作为指纹信息。该明细信息包括一种或多种维度类别以及每种维度类别所取的维度值。在一种示例中，待分析信息中的指纹信息包含设备标识，明细信息包含设备上报的属性数据，例如激活渠道、网络运营商、地理位置等维度类别的具体取值。

[0056] 在本公开的一些实施例中，根据待分析信息来确定键值对形式的明细表包括：将指纹信息记录于明细表中一个键值对的键字段(key字段)，将该指纹信息对应的明细信息记录于明细表中同一个键值对的值字段(value字段)。

[0057] 需要说明的是，不需要在明细表中保存所有历史数据，可以在明细表中仅保存一段时间内的数据，例如仅保存当天的数据。而每天的数据以最后一次上报的数据为准。

[0058] 在本公开的一些实施例中，步骤S20的对明细表进行分析的过程包括：在所有的筛选条件下，以全量数据聚合的方式对明细表进行分析，得到每个筛选条件下的分析结果。其中，每个筛选条件由一项或多项条件项构成。任意一个维度值都可以作为一个筛选条件中的一项条件项，或者说一个筛选条件包含由任意维度值的交叉而形成的多维度交叉条件。例如“activation_channel:甲厂家#brand:甲品牌#os:乙系统”这个筛选条件就是由激活渠道为“甲厂家”、品牌为“甲品牌”和操作系统为“乙系统”这三项条件项组成的，这个筛选条件是一个多维度交叉条件，每个条件项都是一个维度值。值得注意的是，筛选条件不仅包含多维度交叉条件，例如，指纹信息也可以作为一个筛选条件之中的一项条件项；从另一个角度来说，事实上指纹信息也可以作为一种维度类别。

[0059] 需要注意的是，根据所确定的指标类型的不同，进行聚合时进行的具体统计分析会有所不同。例如，对明细表进行的分析会根据是为了统计日登录用户数还是为了统计日新增用户数而有所区别。另外，对于不同的指标类型，可以生成对应的多个结果表，例如根据明细表生成日登录用户数结果表、日活跃用户数结果表、日新增用户数结果表等。

[0060] 在本公开的一些实施例中，根据分析结果来确定键值对形式的结果表包括：将分析结果记录于结果表中一个键值对的值字段(value字段)，将对应的分析结果属性信息记录于结果表中同一个键值对的键字段(key字段)。其中，一个分析结果的属性信息包括该分析结果对应的筛选条件。在一种示例中，将多维度交叉条件记录于结果表中的键值对的键字段。另外，还可以将指纹信息之中的一种或多种信息记录于结果表中的键值对的键字段。需要说明的是，即使未将某种指纹信息作为筛选条件，也可以将该种指纹信息记录于结果表中的键值对的键字段。

[0061] 本公开通过实时分析得到在所有可能的筛选条件下的分析结果，并将筛选条件记录于结果表中一个键值对的键字段、将分析结果记录于结果表中同一键值对的值字段，从而在查询时，通过查询结果表中的键字段就可以得到与待查询的筛选条件对应的分析统计结果，并能够支持以前只有关系数据库才能支持的多维度交叉的功能。

[0062] 图2为本公开的实时分析方法一个实施例提供的以全量数据分布式聚合的方式对

明细表进行分析的示意性流程图。请参阅图2，由于明细表数据量非常巨大，每天的数据量就可达到百亿级别，为了便于海量数据的实时分析，在本公开的一些实施例中，在对明细表进行分析的过程包括，根据指纹信息将明细表中所有的键值对形式的待分析数据切分为n份(切分的份数可按照待分析数据量的大小而定，例如可以将n取为1000)，以将海量数据打散成n个数据集群，然后启动n个执行器(executor)对n份数据进行分布式并发聚合(group by)，再将所得的n份结果合并(merge，事实上也是进行聚合group by操作)，以得到完整的全量聚合分析结果，并将聚合分析结果传递给驱动器(driver)，最后驱动器将聚合分析结果写入到对应的结果表中。利用此分布式聚合分析方法对明细表进行分析，对于每个执行器(executor)的压力都非常小，可以保证数据在很短的时间内完成数据的分析统计；并且具有非常高的扩展性，如果数据继续增多，只需适当增加数据集群总数n的数量即可。

[0063] 图3为本公开的实时分析方法一个实施例提供的以全量数据分布式聚合的方式对明细表进行分析的示意性流程框图。请参阅图3，在本公开的一种实施例中，以全量数据分布式聚合的方式对明细表进行分析的具体过程包括：

[0064] 步骤S21，将明细表划分为多个数据集群作为第一数据集群，且将具有相同指纹信息的待分析信息划分在同一个第一数据集群之中。需要说明的是，这里所说的具有相同的指纹信息，并非是指必须各种指纹信息均相同，而是可以按照多种指纹信息之中的至少一种对待分析信息进行划分，将具有相同设备标识和/或相同应用程序标识和/或相同用户标识划分在同一个第一数据集群，例如，可以仅按照设备标识对待分析信息进行划分。需要说明的是，第一数据集群的总数是可以设置的，可以根据待分析信息数据量的实际情况来调整第一数据集群总数的具体取值。

[0065] 步骤S22，分布式并发地对每个第一数据集群进行聚合(不妨称为第一聚合)，得到每个第一数据集群的第一聚合结果。

[0066] 由于具有相同指纹信息的待分析信息已划分在同一个第一数据集群中，因此第一聚合能够在进行聚合累加的同时对具有相同指纹信息的待分析信息进行去重。

[0067] 步骤S23，对所有的第一聚合结果再进行聚合(不妨称为第二聚合)，以得出全量数据的分析结果。

[0068] 需要说明的是，前述的第一聚合和第二聚合均是基于筛选条件的聚合分析，得出的是筛选条件下的分析结果，从而根据每个筛选条件分别按照步骤S21到步骤S23对明细表进行分布式聚合分析，就可以得到所有筛选条件对应的各个分析结果。

[0069] 进一步地，可以将第一数据集群的划分情况记录在明细表中的键值对的键字段。在本公开一种实施例中，步骤S10还包括将第一数据集群标识记录于明细表中的键字段，该第一数据集群标识为该指纹信息对第一数据集群总数取模而得到的结果。从而在对明细表进行分布式聚合分析时，仅需按照该第一数据集群信息对明细表中的键值对进行划分，就能将待分析信息划分为多个第一数据集群，且将具有相同指纹信息的待分析信息划分在同一个第一数据集群之中。

[0070] 在本公开的一些实施例中，步骤S30还包括将第二数据集群信息记录于结果表中的键字段，用以根据该第二数据集群信息将分析结果分散到多个第二数据集群中。通过将分析结果分散到多个数据集群，可以减小存储压力。可选地，该第二数据集群信息为结果表键字段所记录的筛选条件中的多个条件项的哈希值。

[0071] 在本公开的一些实施例中，获取的待分析信息还包括时间信息。该时间信息可以包括待分析信息的获取时间，以及其他的时间信息，例如在一种示例中，该时间信息包括设备登录某个应用程序的时间。需要注意的是，该时间信息也可作为一个筛选条件中的一项，甚至可以视为一种维度类别。

[0072] 在本公开的一些实施例中，步骤S10的将待分析信息以键值对的形式存储于明细表的过程包括：将待分析信息存储到具有不同时间跨度的多个明细表之中，以对实时获得的待分析信息进行多种时间跨度的记录。具体地，可以将待分析信息依次写入到十分钟级明细表、小时级明细表和天级明细表。

[0073] 在本公开一些实施例中，步骤S20包括，在多个时间窗口(或者称为时间段)对明细表通过第一聚合和第二聚合进行去重处理以得到分时段的多个分析结果。作为一种可选示例，该多个时间窗口为连续的多个跨度相同的时间段。值得注意的是，多个不同的时间窗口可能会包含具有重复指纹信息的待分析信息。例如在一种示例中，每十分钟按照前述步骤S21到步骤S23所示方法对明细表进行分析并同时对设备标识信息进行去重处理，而在两个不同的十分钟区间可能会包含具有重复设备标识信息的待分析信息。因此通过对不同时间窗口进行不同存储，能够在进行分析时得到准确的分时段的分析结果。

[0074] 在本公开的一些实施例中，步骤S20的以全量数据聚合的方式对明细表进行分析的过程包括：对于一个明细表，按照多种时间跨度进行分析以得到与多种时间跨度对应的多个分析结果。然后可以将所得到的与多种时间跨度对应的多个分析结果分别存储于多个结果表，和/或可以将分析结果对应的时间跨度信息存储于结果表中的键字段。

[0075] 具体地，可以根据一个天级明细表得到天级的分析结果、小时级的积累值分析结果、十分钟级的积累值分析结果等多个分析结果，进而得到天级的分时段结果表、小时级的积累值结果表、十分钟级的积累值结果表这三个结果表。

[0076] 在查询时，直接利用结果表就能得到查询结果，具体的查询过程包括：获取查询条件，根据该查询条件对结果表中的键字段进行查找，根据查找到的键值对的值字段确定与该查询条件对应的分析结果作为查询结果。事实上，该查询条件是一个筛选条件，因此也可以将查询条件称为待查询的筛选条件。由于结果表中记录了预先分析得到的所有筛选条件以及对应的分析结果，因此通过查询结果表就能够找到所需的查询结果。如果确定了多个结果表，例如根据十分钟级明细表、小时级明细表和天级明细表生成了天级的分时段结果表、小时级的分时段结果表、小时级的积累值结果表、十分钟级的分时段结果表以及十分钟级的积累值结果表，则查询过程还包括，根据该查询条件确定待查询的结果表，然后从该待查询结果表中查找查询结果。

[0077] 在本公开的一些实施例中，可以利用hbase数据库存储明细表和结果表。hbase是一种键值对形式的数据存储形式，可以非常优雅的支持海量的数据存储，同时有着高效的查询速度。

[0078] 在本公开的一些实施例中，可以将实时获取的待分析信息上报到实时流，并利用实时流数据处理框架对该实时流形式的待分析信息进行实时处理。具体地，可以利用诸如Kafka和Storm这些流处理平台对待分析信息进行实时处理，以将待分析信息实时地写入明细表。例如，在一种具体示例中，将获取的待分析信息写入Kafka实时流，然后利用Storm实时消费Kafka数据，将处理后的待分析信息写入hbase明细表。

[0079] 在本公开的一些实施例中，可以利用诸如Spark的统计框架轮询查询明细表，按照前述实施例中的具体步骤以全量数据分布式聚合的方式计算出所有可能的分析结果，再将分析结果写入到对应的结果表中。

[0080] 在查询时，Web系统通过直接读取hbase结果表就能够得到查询结果，并通过Web系统对查询结果做可视化展现。

[0081] 在本公开的一个实施例中，hbase明细表中的键字段(row_key)的格式为：

[0082] {salt1}#{date_format}#{app_id}#{device_id}。

[0083] 其中，该date_format为时间信息，对于不同时间跨度的明细表具有不同的具体形式，例如天级表的date_format的取值的具体形式可以为某年某月某日，小时级表的date_format的取值的具体形式可以为某年某月某日某时，而十分钟级表的date_format的取值的具体形式可以为某年某月某日某时第某个十分钟。该app_id为应用程序标识。该device_id为设备标识。该salt1为前述的第一数据集群标识，salt1的具体取值可以为device_id%1000，以通过将设备标识取模1000，而将待分析信息划分为1000份且将具有相同设备标识的待分析信息划分在一起。

[0084] hbase明细表中的值字段(value)的格式为：

[0085] dimension_key:dimension_value,dimension_key:dimension_value,...其中，dimension_key为维度类别，dimension_value为维度值。例如，hbase明细表中一个键值对中的值字段可以是：“brand:甲品牌,os:乙系统,os_version:0.12”。

[0086] 在本公开的一个实施例中，hbase结果表中的键字段(row_key)的格式为：

[0087] {salt2}#{date_format}#{app_id}#{dimension_whence_str}#{optional_time}。

[0088] 其中，该date_format为时间信息，与hbase明细表类似地，对于不同时间跨度的结果表具有不同的具体形式。该app_id为应用程序标识。该dimension_whence_str由一个筛选条件中的所有维度值的字符串拼接而成，例如，如果一个筛选条件为品牌(brand)是“甲品牌”且操作系统(os)是“乙系统”，则dimension_whence_str为“brand:甲品牌#os:乙系统”。值得注意的是，dimension_whence_str中的各段字符串可以是按照维度类别有序排列的。该optional_time记录的是较date_format更加细化的时间信息，例如在根据天级明细表分析得出十分钟级结果表和小时级结果表的示例中，十分钟级结果表和小时级结果表的date_format均与天级明细表的date_format一致，而十分钟级结果表的optional_time与小时级结果表的optional_time并不相同，十分钟级结果表的optional_time记录的是分析结果对应的是哪个十分钟，而小时级结果表的optional_time记录的是分析结果对应的是哪个小时。因此，对于不同时间跨度的结果表optional_time具有不同的具体形式，例如，如果将时间跨度分为天级、小时级和十分钟级三个级别，天级表不需要optional time或者说天级表的optional time的具体取值为None，小时级表的optional time的取值的具体形式可以为某时，而十分钟级表的optional time的取值的具体形式可以为某时第某个十分钟。该salt2为前述的第二数据集群标识，salt2的具体取值可以为hash(date_format+app_id+dimension_whence_str)%10，通过设置第二数据集群标识可以优化查询，根据查询条件直接定位salt分区。

[0089] hbase结果表中的值字段(value)的格式为：

[0090] 统计类别:{统计类别的取值}。

[0091] 例如,当统计分析的类别为日登录用户数(简称为dlu)时,hbase结果表中的值字段的格式为dlu:{dlu的取值},当统计分析的类别为日新增用户数(简称为dnu)时,hbase结果表中的值字段的格式为dnu:{dnu的取值}。

[0092] 在本公开的一个实施例中,根据结果表确定查询结果的过程包括:

[0093] 获取待查询的筛选条件;

[0094] 根据待查询筛选条件中的待查询的指标类型(例如,是查询dlu还是查询dnu)、待查询的时间跨度、查询的是积累值还是分时段值等条件项,确定待查询的结果表;

[0095] 根据待查询筛选条件中的待查询的维度值、指纹信息、时间信息确定待查询的键字段;在利用hbase数据库时,就是根据待查询筛选条件拼凑出完整的hbase键字段前缀(key prefix),在一种示例中,键字段前缀的形式为{salt2}#{date_format}#{app_id}#{dimension_whence_str}#;

[0096] 利用所确定的该待查询键字段在所确定的该待查询结果表中查找出相关的分析结果作为查询结果。

[0097] 图4为本公开的实时分析系统100一个实施例的示意性结构图。请参阅图4,本公开示例的实时分析系统100,主要包括:

[0098] 明细表确定模块110,用于实时获取待分析信息,对该待分析信息进行实时处理,以将该待分析信息以键值对的形式存储于明细表;

[0099] 分析模块120,用于以全量数据聚合的方式对明细表进行分析,得到分析结果;

[0100] 结果表确定模块130,用于将分析结果以键值对形式存储于结果表,以供查询。

[0101] 具体地,明细表确定模块110所获取的待分析信息中的每条信息包含指纹信息以及明细信息。

[0102] 在本公开的一些实施例中,明细表确定模块110包括第一记录子模块(图中未示出),用于:将指纹信息记录于明细表中一个键值对的键字段(key字段),将该指纹信息对应的明细信息记录于明细表中同一个键值对的值字段(value字段)。

[0103] 在本公开的一些实施例中,分析模块120具体用于在所有的筛选条件下,以全量数据聚合的方式对明细表进行分析,得到每个筛选条件下的分析结果。

[0104] 图5为本公开一个实施例提供的分析模块120的示意性结构图。请参阅图5,在本公开的一种实施例中,分析模块120具体包括:

[0105] 第一数据集群划分单元121,用于将明细表划分为多个数据集群作为第一数据集群,且将具有相同指纹信息的待分析信息划分在同一个第一数据集群之中;

[0106] 第一聚合单元122,用于分布式并发地对每个第一数据集群进行聚合(不妨称为第一聚合),得到每个第一数据集群的第一聚合结果;

[0107] 第二聚合单元123,用于对所有的第一聚合结果再进行聚合(不妨称为第二聚合),以得出分析结果。

[0108] 进一步地,在本公开一种实施例中,明细表确定模块110还包括第二记录子模块(图中未示出),用于将第一数据集群标识记录于明细表中的键字段,该第一数据集群标识为该指纹信息对第一数据集群总数取模而得到的结果。并且,第一数据集群划分单元121具体用于:按照该第一数据集群信息对明细表中的键值对进行划分。这样就能将待分析信息

划分为多个第一数据集群，且将具有相同指纹信息的待分析信息划分在同一个第一数据集群之中。

[0109] 在本公开的一些实施例中，结果表确定模块130包括第三记录子模块(图中未示出)，用于：将分析结果记录于结果表中一个键值对的值字段(value字段)，将对应的分析结果属性信息记录于结果表中同一个键值对的键字段(key字段)。

[0110] 在本公开的一些实施例中，结果表确定模块130还包括第四记录子模块(图中未示出)，用于将第二数据集群信息记录于结果表中的键字段，用以根据该第二数据集群信息将分析结果分散到多个第二数据集群中。

[0111] 在本公开的一些实施例中，结果表确定模块130还包括第五记录子模块(图中未示出)用于：将指纹信息之中的一种或多种记录于结果表中的键字段。

[0112] 在本公开的一些实施例中，明细表确定模块110所获取的待分析信息还包括时间信息。明细表确定模块110还可包括第六记录子模块(图中未示出)，用于将该时间信息存储于明细表中的键值对的键字段，和/或结果表确定模块130还可包括一个子模块(图中未示出)，用于将该时间信息储存于结果表中的键值对的键字段。

[0113] 在本公开的一些实施例中，明细表确定模块110还包括第七记录子模块，用于将待分析信息存储到具有不同时间跨度的多个明细表之中，以对实时获得的待分析信息进行多种时间跨度的记录。

[0114] 在本公开的一些实施例中，分析模块120包括第一分析子模块，用于：对于一个明细表，按照多种时间跨度进行分析以得到与多种时间跨度对应的多个分析结果。而结果表确定模块130可以具体用于将所得到的与多种时间跨度对应的多个分析结果分别存储于多个结果表，和/或结果表确定模块130可以具体用于将分析结果对应的时间跨度信息存储于结果表中的键字段。

[0115] 在本公开一些实施例中，分析模块120包括第二分析子模块，用于：在多个时间窗口对明细表通过第一聚合和第二聚合进行去重处理以得到分时段的多个分析结果。

[0116] 在一些实施例中，本公开的实时分析系统100还包括：查询条件获取模块(图中未示出)，用于获取查询条件，该查询条件事实上也就是待查询的筛选条件；查询模块(图中未示出)，用于根据该查询条件对结果表中的键字段进行查找，根据查找到的键值对的值字段确定与该查询条件对应的分析结果作为查询结果。

[0117] 在本公开的一些实施例中，明细表确定模块110包括实施流处理子模块(图中未示出)，用于将实时获取的待分析信息上报到实时流，并利用实时流数据处理框架对该实时流形式的待分析信息进行实时处理。

[0118] 图6是图示根据本公开的实施例的实时分析装置的硬件框图。如图6所示，根据本公开实施例的实时分析装置200包括存储器201和处理器202。实时分析装置200中的各组件通过总线系统和/或其它形式的连接机构(未示出)互连。

[0119] 该存储器201用于存储非暂时性计算机可读指令。具体地，存储器201可以包括一个或多个计算机程序产品，该计算机程序产品可以包括各种形式的计算机可读存储介质，例如易失性存储器和/或非易失性存储器。该易失性存储器例如可以包括随机存取存储器(RAM)和/或高速缓冲存储器(cache)等。该非易失性存储器例如可以包括只读存储器(ROM)、硬盘、闪存等。

[0120] 该处理器202可以是中央处理单元(CPU)或者具有数据处理能力和/或指令执行能力的其它形式的处理单元，并且可以控制实时分析装置200中的其它组件以执行期望的功能。在本公开的一个实施例中，该处理器202用于运行该存储器201中存储的该计算机可读指令，使得该实时分析装置200执行前述的本公开各实施例的实时分析方法的全部或部分步骤。

[0121] 图7是图示根据本公开的实施例的计算机可读存储介质的示意图。如图7所示，根据本公开实施例的计算机可读存储介质300，其上存储有非暂时性计算机可读指令301。当该非暂时性计算机可读指令301由处理器运行时，执行前述的本公开各实施例的实时分析方法的全部或部分步骤。

[0122] 图8是图示根据本公开实施例的终端设备的硬件结构示意图。终端设备可以以各种形式来实施，本公开中的终端设备可以包括但不限于诸如移动电话、智能电话、笔记本电脑、数字广播接收器、PDA(个人数字助理)、PAD(平板电脑)、PMP(便携式多媒体播放器)、导航装置、车载终端设备、车载显示终端、车载电子后视镜等等的移动终端设备以及诸如数字TV、台式计算机等等的固定终端设备。

[0123] 如图8所示，终端设备1100可以包括无线通信单元1110、A/V(音频/视频)输入单元1120、用户输入单元1130、感测单元1140、输出单元1150、存储器1160、接口单元1170、控制器1180和电源单元1190等等。图8示出了具有各种组件的终端设备，但是应理解的是，并不要求实施所有示出的组件。可以替代地实施更多或更少的组件。

[0124] 其中，无线通信单元1110允许终端设备1100与无线通信系统或网络之间的无线电通信。A/V输入单元1120用于接收音频或视频信号。用户输入单元1130可以根据用户输入的命令生成键输入数据以控制终端设备的各种操作。感测单元1140检测终端设备1100的当前状态、终端设备1100的位置、用户对于终端设备1100的触摸输入的有无、终端设备1100的取向、终端设备1100的加速或减速移动和方向等等，并且生成用于控制终端设备1100的操作的命令或信号。接口单元1170用作至少一个外部装置与终端设备1100连接可以通过的接口。输出单元1150被构造为以视觉、音频和/或触觉方式提供输出信号。存储器1160可以存储由控制器1180执行的处理和控制操作的软件程序等等，或者可以暂时地存储已经输出或将要输出的数据。存储器1160可以包括至少一种类型的存储介质。而且，终端设备1100可以与通过网络连接执行存储器1160的存储功能的网络存储装置协作。控制器1180通常控制终端设备的总体操作。另外，控制器1180可以包括用于再现或回放多媒体数据的多媒体模块。控制器1180可以执行模式识别处理，以将在触摸屏上执行的手写输入或者图片绘制输入识别为字符或图像。电源单元1190在控制器1180的控制下接收外部电力或内部电力并且提供操作各元件和组件所需的适当的电力。

[0125] 本公开提出的实时分析方法的各种实施方式可以以使用例如计算机软件、硬件或其任何组合的计算机可读介质来实施。对于硬件实施，本公开提出的实时分析方法的各种实施方式可以通过使用特定用途集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理装置(DSPD)、可编程逻辑装置(PLD)、现场可编程门阵列(FPGA)、处理器、控制器、微控制器、微处理器、被设计为执行这里描述的功能的电子单元中的至少一种来实施，在一些情况下，本公开提出的实时分析方法的各种实施方式可以在控制器1180中实施。对于软件实施，本公开提出的实时分析方法的各种实施方式可以与允许执行至少一种功能或操作的单独的软

件模块来实施。软件代码可以由以任何适当的编程语言编写的软件应用程序(或程序)来实施,软件代码可以存储在存储器1160中并且由控制器1180执行。

[0126] 以上,根据本公开实施例的实时分析方法、系统、装置、计算机可读存储介质以及终端设备,通过将待分析数据记录为键值对形式的明细表,并以全量聚合的方式对明细表中的数据进行实时统计分析,能够大大提高实时分析的准确性和效率。

[0127] 以上结合具体实施例描述了本公开的基本原理,但是,需要指出的是,在本公开中提及的优点、优势、效果等仅是示例而非限制,不能认为这些优点、优势、效果等是本公开的各个实施例必须具备的。另外,上述公开的具体细节仅是为了示例的作用和便于理解的作用,而非限制,上述细节并不限制本公开为必须采用上述具体的细节来实现。

[0128] 本公开中涉及的器件、装置、设备、系统的方框图仅作为例示性的例子并且不意图要求或暗示必须按照方框图示出的方式进行连接、布置、配置。如本领域技术人员将认识到的,可以按任意方式连接、布置、配置这些器件、装置、设备、系统。诸如“包括”、“包含”、“具有”等等的词语是开放性词汇,指“包括但不限于”,且可与其互换使用。这里所使用的词汇“或”和“和”指词汇“和/或”,且可与其互换使用,除非上下文明确指示不是如此。这里所使用的词汇“诸如”指词组“诸如但不限于”,且可与其互换使用。

[0129] 另外,如在此使用的,在包含“至少一个”、“一个或多个”、“一种或多种”的项的列举中使用的“或”指示分离的列举,以便例如“A、B或C的至少一个”或“A、B或C的一种或多种”的列举意味着A或B或C,或AB或AC或BC,或ABC(即A和B和C)。此外,措辞“示例的”不意味着描述的例子是优选的或者比其他例子更好。

[0130] 还需要指出的是,在本公开的系统和方法中,各部件或各步骤是可以分解和/或重新组合的。这些分解和/或重新组合应视为本公开的等效方案。

[0131] 可以不脱离由所附权利要求定义的教导的技术而进行对在此所述的技术的各种改变、替换和更改。此外,本公开的权利要求的范围不限于以上所述的处理、机器、制造、事件的组成、手段、方法和动作的具体方面。可以利用与在此所述的相应方面进行基本相同的功能或者实现基本相同的结果的当前存在的或者稍后要开发的处理、机器、制造、事件的组成、手段、方法或动作。因而,所附权利要求包括在其范围内的这样的处理、机器、制造、事件的组成、手段、方法或动作。

[0132] 提供所公开的方面的以上描述以使本领域的任何技术人员能够做出或者使用本公开。对这些方面的各种修改对于本领域技术人员而言是非常显而易见的,并且在此定义的一般原理可以应用于其他方面而不脱离本公开的范围。因此,本公开不意图被限制到在此示出的方面,而是按照与在此公开的原理和新颖的特征一致的最宽范围。

[0133] 为了例示和描述的目的已经给出了以上描述。此外,此描述不意图将本公开的实施例限制到在此公开的形式。尽管以上已经讨论了多个示例方面和实施例,但是本领域技术人员将认识到其某些变型、修改、改变、添加和子组合。

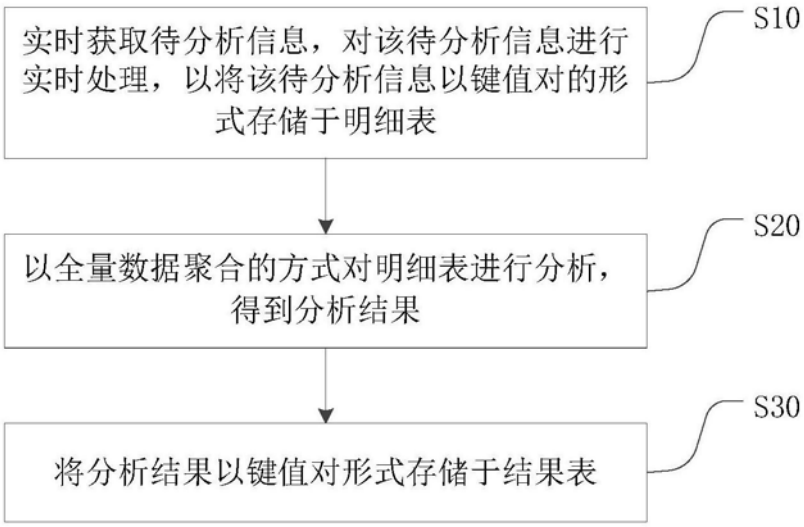

实时获取待分析信息，对该待分析信息进行实时处理，以将该待分析信息以键值对的形式存储于明细表
S10
以全量数据聚合的方式对明细表进行分析，得到分析结果
S20
将分析结果以键值对形式存储于结果表
S30

图1

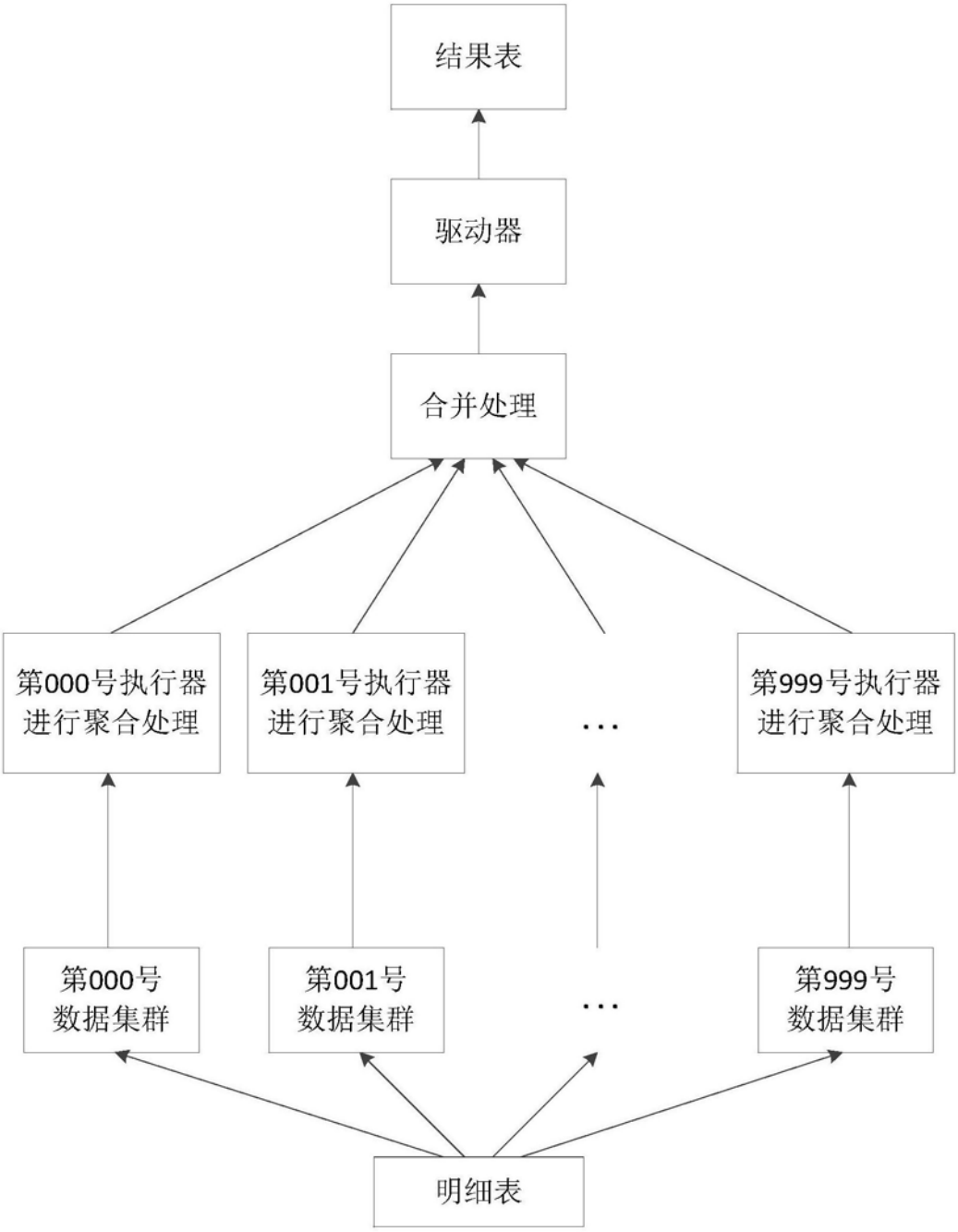

结果表
驱动器
合并处理
第000号执行器
进行聚合处理
第001号执行器
进行聚合处理
...
第999号执行器
进行聚合处理
第000号
数据集群
第001号
数据集群
...
第999号
数据集群
明细表

图2

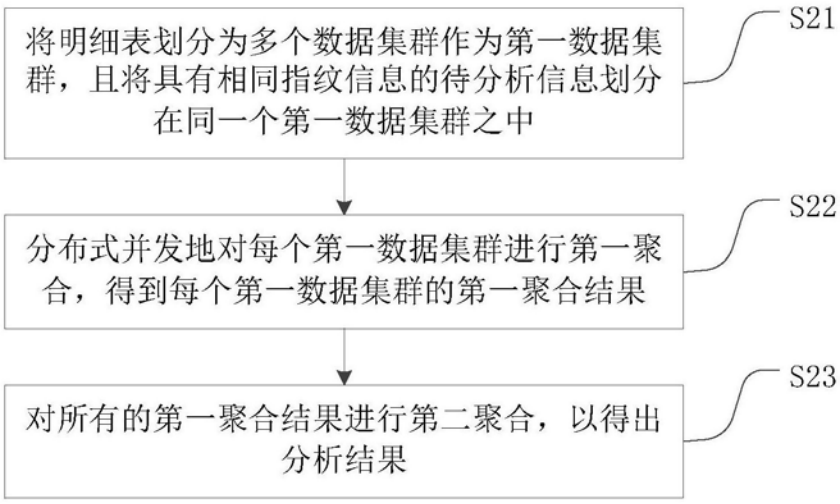

将明细表划分为多个数据集群作为第一数据集群，且将具有相同指纹信息的待分析信息划分在同一个第一数据集群之中
S21
分布式并发地对每个第一数据集群进行第一聚合，得到每个第一数据集群的第一聚合结果
S22
对所有的第一聚合结果进行第二聚合，以得出分析结果
S23

图3


实时分析系统100
明细表确定模块110
分析模块120
结果表确定模块130

图4


分析模块120
第一数据集群划分单元121
第一聚合单元122
第二聚合单元123

图5


实时分析装置200
存储器201
处理器202

图6

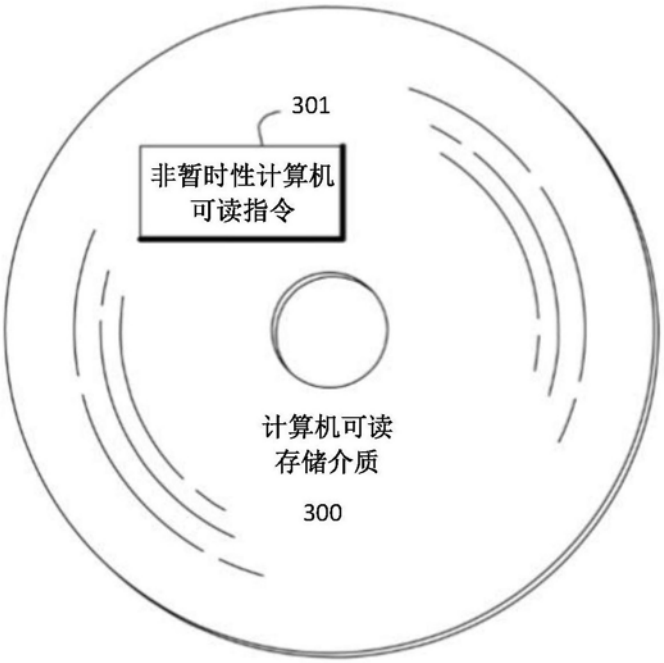

301
非暂时性计算机
可读指令
计算机可读
存储介质
300

图7

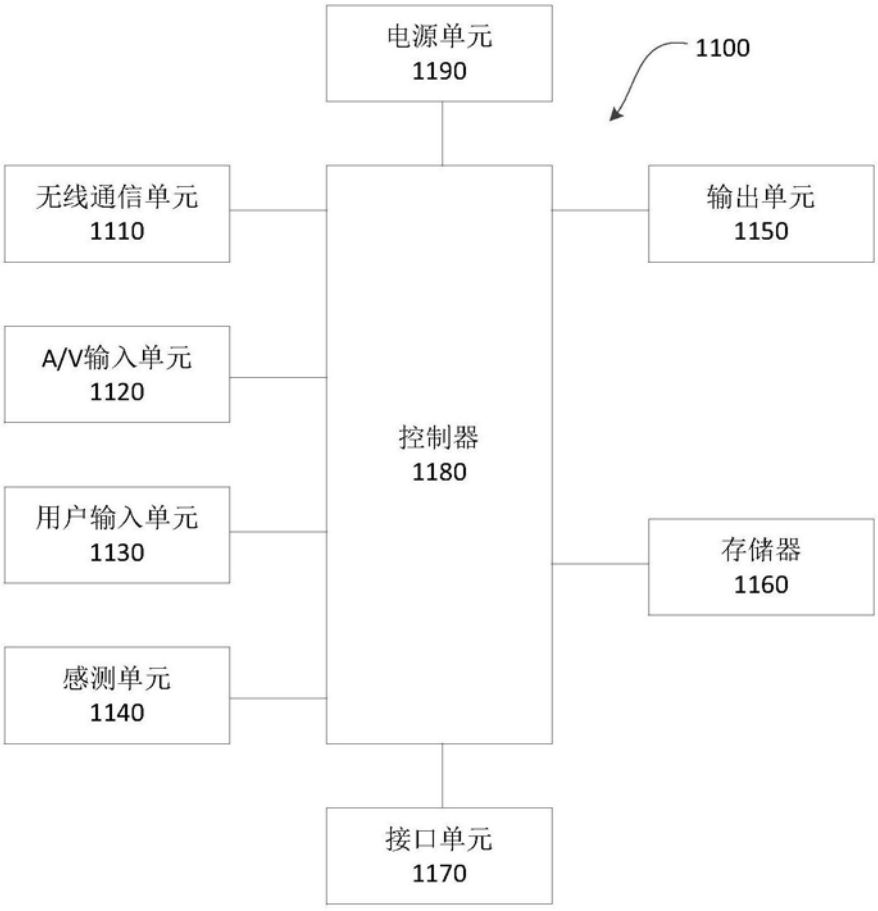

电源单元
1190
1100
无线通信单元
1110
输出单元
1150
A/V输入单元
1120
控制器
1180
用户输入单元
1130
存储器
1160
感测单元
1140
接口单元
1170

图8