(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 106326317 A
(43)申请公布日 2017.01.11

(21)申请号 201510400093.1

(22)申请日 2015.07.09

(71)申请人 中国移动通信集团山西有限公司
地址 030032 山西省太原市经济技术开发区武洛街25

(72)发明人 卢山

(74)专利代理机构 北京派特恩知识产权代理有限公司 11270
代理人 李梅香 张颖玲

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

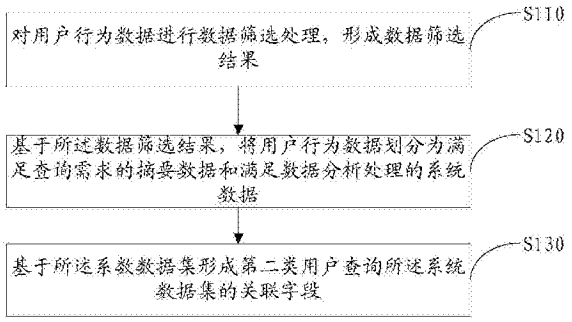
权利要求书2页 说明书15页 附图4页

(54)发明名称

数据处理方法及装置

(57)摘要

本发明公开了一种数据处理方法及装置，所述方法包括：对用户行为数据进行数据筛选处理，形成数据筛选结果；基于所述数据筛选结果，将用户行为数据划分为满足查询需求的摘要数据和满足数据分析处理的系统数据；其中，所述摘要数据归属于用户列表集；所述系统数据和所述摘要数据均属于系统数据集；基于所述系数数据集形成查询所述系统数据集的关联字段；其中，所述关联字段归属于用户明细集。



CN 106326317 A

1.一种数据处理方法,其特征在于,所述方法包括:

对用户行为数据进行数据筛选处理,形成数据筛选结果;

基于所述数据筛选结果,将用户行为数据划分为满足查询需求的摘要数据和满足数据分析处理的系统数据;其中,所述摘要数据归属于用户列表集;所述系统数据和所述摘要数据均属于系统数据集;

基于所述系数数据集形成查询所述系统数据集的关联字段;其中,所述关联字段归属于用户明细集。

2.根据权利要求1所述的方法,其特征在于,

所述方法还包括:

基于所述系统数据集,建立与所述关联字段关联的主表。

3.根据权利要求1所述的方法,其特征在于,

所述用户列表集还包括摘要字段;

其中,所述摘要字段与所述摘要数据具有映射关系,能够用于查询所述摘要数据。

4.根据权利要求3所述的方法,其特征在于,

所述摘要字段包括用户标识及查询时间。

5.根据权利要求3所述的方法,其特征在于,

所述方法还包括:

基于所述用户列表集及所述用户明细集,建立索引表;

其中,所述索引表的查询索引包括关联字段以及所述摘要字段。

6.根据权利要求5所述的方法,其特征在于,

所述索引表还包括所述摘要数据。

7.根据权利要求5所述的方法,其特征在于,

所述方法还包括:

接收基于用户输入形成的查询标签;

将所述查询标签与所述索引表中的字段进行匹配;

若所述查询标签与所述摘要字段相匹配,则基于所述摘要字段查询所述摘要数据,并返回所述摘要数据;

若所述查询标签与所述关联字段相匹配,则基于所述关联字段,查询所述主表并返回查询结果。

8.一种数据处理装置,其特征在于,所述装置包括:

筛选单元,用于对用户行为数据进行数据筛选处理,形成数据筛选结果;

划分单元,用于基于所述数据筛选结果,将用户行为数据划分为满足查询需求的摘要数据和满足数据分析处理的系统数据;其中,所述摘要数据归属于用户列表集;所述系统数据和所述摘要数据均属于系统数据集;

生成单元,用于基于所述系数数据集形成查询所述系统数据集的关联字段;其中,所述关联字段归属于用户明细集。

9.根据权利要求8所述的装置,其特征在于,

所述装置还包括:

第一建立单元,用于基于所述系统数据集,建立与所述关联字段关联的主表。

10. 根据权利要求 9 所述的装置，其特征在于，

所述用户列表集还包括摘要字段；

其中，所述摘要字段与所述摘要数据具有映射关系，能够用于查询所述摘要数据。

11. 根据权利要求 10 所述的装置，其特征在于，

所述摘要字段包括用户标识及查询时间。

12. 根据权利要求 10 所述的装置，其特征在于，

所述装置还包括：

第二建立单元，用于基于所述用户列表集及所述用户明细集，建立索引表；

其中，所述索引表的查询索引包括关联字段以及所述摘要字段。

13. 根据权利要求 12 所述的装置，其特征在于，

所述索引表还包括所述摘要数据。

14. 根据权利要求 12 所述的装置，其特征在于，

所述装置还包括：

接收单元，用于接收基于用户输入形成的查询标签；

匹配单元，用于将所述查询标签与所述索引表中的字段进行匹配；

第一查询单元，用于若所述查询标签与所述摘要字段相匹配，则基于所述摘要字段查询所述摘要数据，并返回所述摘要数据；

第二查询单元，用于若所述查询标签与所述关联字段相匹配，则基于所述关联字段，查询所述主表并返回查询结果。

数据处理方法及装置

技术领域

[0001] 本发明涉及数据处理领域，尤其涉及一种数据处理方法及装置。

背景技术

[0002] 随着信息技术和电子技术的发展，出现了大数据的概念和使用。大数据能够更好的实现数据共享。然而在现有技术中发现，目前数据处理中依然存在着大量的数据查询慢、数据处理效率低及消耗了大量的数据处理资源等问题。

[0003] 比如，基于 Hbase 数据库的数据处理，以行关键字 (RowKey) 进行查询时，速度快且效率高，但是以非行关键字进行查询时，通常会出现速率慢及处理效率低等问题。所述 Hbase 数据库为是一个分布式的、面向列的开源数据库，不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。另一个不同的是 HBase 基于列的而不是基于行的模式的数据库。

[0004] 故在现有技术，提出一种数据处理效率高且查询速度快的数据处理方法，是亟待解决的问题。

发明内容

[0005] 有鉴于此，本发明实施例期望提供一种数据处理方法及装置，能够至少部分解决数据处理效率低或查询速度慢的问题。

[0006] 为达到上述目的，本发明的技术方案是这样实现的：本发明实施例第一方面提供了一种数据处理方法，所述方法包括：

[0007] 对用户行为数据进行数据筛选处理，形成数据筛选结果；

[0008] 基于所述数据筛选结果，将用户行为数据划分为满足查询需求的摘要数据和满足数据分析处理的系统数据；其中，所述摘要数据归属于用户列表集；所述系统数据和所述摘要数据均属于系统数据集；

[0009] 基于所述系数数据集形成查询所述系统数据集的关联字段；其中，所述关联字段归属于用户明细集。

[0010] 基于上述方案，所述方法还包括：

[0011] 基于所述系统数据集，建立与所述关联字段关联的主表。

[0012] 基于上述方案，所述用户列表集还包括摘要字段；

[0013] 其中，所述摘要字段与所述摘要数据具有映射关系，能够用于查询所述摘要数据。

[0014] 基于上述方案，所述摘要字段包括用户标识及查询时间。

[0015] 基于上述方案，所述方法还包括：

[0016] 基于所述用户列表集及所述用户明细集，建立索引表；

[0017] 其中，所述索引表的查询索引包括关联字段以及所述摘要字段

[0018] 基于上述方案，所述索引表还包括所述摘要数据。

[0019] 基于上述方案，所述方法还包括：

[0020] 接收基于用户输入形成的查询标签；

[0021] 将所述查询标签与所述索引表中的字段进行匹配；

[0022] 若所述查询标签与所述摘要字段相匹配，则基于所述摘要字段查询所述摘要数据，并返回所述摘要数据；

[0023] 若所述查询标签与所述关联字段相匹配，则基于所述关联字段，查询所述主表并返回查询结果。

[0024] 本发明实施例第二方面提供一种数据处理装置，所述装置包括：

[0025] 筛选单元，用于对用户行为数据进行数据筛选处理，形成数据筛选结果；

[0026] 划分单元，用于基于所述数据筛选结果，将用户行为数据划分为满足查询需求的摘要数据和满足数据分析处理的系统数据；其中，所述摘要数据归属于用户列表集；所述系统数据和所述摘要数据均属于系统数据集；

[0027] 生成单元，用于基于所述系数数据集形成查询所述系统数据集的关联字段；其中，所述关联字段归属于用户明细集。

[0028] 基于上述方案，所述装置还包括：

[0029] 第一建立单元，用于基于所述系统数据集，建立与所述关联字段关联的主表。

[0030] 基于上述方案，所述用户列表集还包括摘要字段；

[0031] 其中，所述摘要字段与所述摘要数据具有映射关系，能够用于查询所述摘要数据。

[0032] 基于上述方案，所述摘要字段包括用户标识及查询时间。

[0033] 基于上述方案，所述装置还包括：

[0034] 第二建立单元，用于基于所述用户列表集及所述用户明细集，建立索引表；

[0035] 其中，所述索引表的查询索引包括关联字段以及所述摘要字段

[0036] 基于上述方案，

[0037] 所述索引表还包括所述摘要数据。

[0038] 基于上述方案，所述装置还包括：

[0039] 接收单元，用于接收基于用户输入形成的查询标签；

[0040] 匹配单元，用于将所述查询标签与所述索引表中的字段进行匹配；

[0041] 第一查询单元，用于若所述查询标签与所述摘要字段相匹配，则基于所述摘要字段查询所述摘要数据，并返回所述摘要数据；

[0042] 第二查询单元，用于若所述查询标签与所述关联字段相匹配，则基于所述关联字段，查询所述主表并返回查询结果。

[0043] 本发明实施例所述的数据处理方法及装置，将形成用户列表集、用户明细集和系统数据集这三个数据集，用户列表集是用户通常会查询到的数据，放在用户列表集中，这样在进行一般数据查询时，用户列表集中的数据是小于所用用户行为数据的，从而减少检索量，从而提高了查询速度。同时形成了用户明细集，用户明细集内形成有关联字段，能够查询到系统数据集中不常查询的数据，且系统数据集中的数据方便进行系统分析处理，实践证明，数据冗余度小且冗余度可以根据需要通过调整数据集所包括数据实现冗余度的可控，从而减少了数据占用大量的存储和系统维护运行资源的现象。

附图说明

[0044] 图 1 为本发明实施例所述的一种数据处理方法的流程示意图；

[0045] 图 2 为本发明实施例所述的一种数据处理方法的局部流程示意图；

[0046] 图 3 为本发明实施例所述的数据处理装置的结构示意图之一；

[0047] 图 4 为本发明实施例所述的数据处理装置的结构示意图之二；

[0048] 图 5 为本发明实施例所述的数据处理方法中数据划分的流程示意图；

[0049] 图 6 为本发明实施例所述的三种数据集之间的关系示意图；

[0050] 图 7 为本发明实施例所述的主表和索引表的效果示意。

具体实施方式

[0051] 以下结合说明书附图及具体实施例对本发明的技术方案做进一步的详细阐述。

[0052] 方法实施例：

[0053] 如图 1 所示，本实施例提供了一种数据处理方法，所述方法包括：

[0054] 步骤 S110：对用户行为数据进行数据筛选处理，形成数据筛选结果；

[0055] 步骤 S120：基于所述数据筛选结果，将用户行为数据划分为满足查询需求的摘要数据和满足数据分析处理的系统数据；其中，所述摘要数据归属于用户列表集；所述系统数据和所述摘要数据均属于系统数据集；

[0056] 步骤 S130：基于所述系数数据集形成查询所述系统数据集的关联字段；其中，所述关联字段归属于用户明细集。

[0057] 在本实施例中所述步骤 S110 中可以根据数据处理需求来进行所述数据筛选处理。通常需要满足普通用户查询需求的数据则应该属于一类的数据。这里满足用户查询需求可包括：满足用户对指定时间内发生的用户行为数据实时查询需求的数据。这里的指定时间可以为从当前时间开始，向前退一段时间内的数据。所述指定时间可为最近一个月内的用户行为数据。而所述系统数据集中的系统数据的话，可能是用户查询的概率较小的数据，具体如，根据对查询统计结果，将用户查询概率小于阈值或查询概率从高到低靠后的数据作为系统数据归属到系统数据集中。

[0058] 值得注意的是在本实施例中所述步骤 S110 中所述数据筛选处理，可认为是数据存储之前依据数据存储规则进行的数据分析和抽象，在存储空间上各个集合的数据都可以存储在一起，可以在存储逻辑上，这些数据归属于不同的集合。这里的集合可包括用户列表集和系统数据集等。数据在存储逻辑上的划分，可以通过数据指针以及数据标签等方式来实现。

[0059] 以 HBase 数据库中存储的数据为例，在主表中存储有 P1 列数据，在步骤 S110 中筛选出 P2 列作为所述摘要数据中的数据；所述 P2 为小于所述 P1 的正整数。在步骤 S120 中进行数据划分的过程中，还包括生成所述摘要数据的查询索引的步骤。实质上摘要数据相当于一个可查询的表，查询该表需要查询索引，该查询索引能够获取该摘要数据。如所述 P2 列数据中的每一行数据都对应一个查询的索引，该查询索引在所述 HBase 数据库中可称为 RoWKey。

[0060] 当然，进行数据查询的用户可分为多个类别，具体如包括两个类别。在步骤 S120 中所述摘要数据可能是能够满足第一类用户查询需求的数据。这里的普通用户即为所述第一类用户，通常所述第一类用户即为权限较低的用户，可能某些数据不对这些用户开放，或

者这些用户对某些数据不感兴趣，不会要求查询对应的数据。这里的某些数据即包括所述系统数据。譬如，所述用户行为包括用户 A 的上网行为。用户 A 感兴趣的数据可能是自己访问了哪些网页、访问网页所产生的数据流量等数据。但是用户 A 可能不感兴趣的是，被自己访问的网页采用的通信协议及资源 IP 地址等。

[0061] 总之，所述用户行为数据根据预定的数据划分策略划分称为摘要数据和系统数据。系统数据为能够满足系统分析需求的数据。

[0062] 在本实施例中为了提高用户查询摘要速率，通过步骤 S110 中将用户不感兴趣或不允许查询的系统数据与所述摘要数据分离出来。这样的话用户在查询数据时，就不用到所有用户行为数据中去查询，从而减少了数据比对匹配查询的量，从而能够提高查询速度。

[0063] 当然为了慢速分析处理需求，本实施例中摘要数据和系统数据都归属到了系统数据集中，方便数据分析装置对数据进行分析，这样的话，也保证了系统需要进行数据分析时的数据分析处理效率。与此同时，为了方便对所有用户行为数据的查询，在本实施例中还引入了用户明细集，在用户明细集里面形成并存储有关联字段，这些关联字段可以作为查询所述系统数据集中各个用户行为数据的查询索引。

[0064] 这样的话，也同时满足了对系统数据的查询需求；且采用这种数据处理结构在进行数据查询时，实践证明速度也是毫秒级别的。

[0065] 作为本实施例的进一步改进，所述方法还包括：

[0066] 基于所述系统数据集，建立与所述关联字段关联的主表。

[0067] 本实施例所述主表能够包括各个用户行为数据，在本实施例中所述主表可为按时间分布形成的表；通常以预定的时间增量周期，更新所述主表。

[0068] 所述关联字段与所述主表关联，采用所述关联字段可以在所述主表中进行查询，能够返回对应的数据；这样就能够简便的实现主表中数据的查询，尤其方便了第二类用户对主表中系统分析处理之后的处理结果的查询。譬如，通过对用户行为数据的分析，产生了分析结果，该分析结果可能是对某一个网站在指定时间内的访问频次，这些数据可能涉及商业密码，对具有较低权限的第一类用户就不开放，但是对于第二类用户（如网站分析维护工作人员）就开放，则这个通过系统分析处理产生基于用户行为产生的数据，能够通过所述用户明细列表中的关联字段查询到。

[0069] 作为本实施例的进一步改进，所述用户列表集还包括摘要字段；

[0070] 其中，所述摘要字段与所述摘要数据具有映射关系，能够用于查询所述摘要数据。所述摘要字段包括用户标识及查询时间。所述用户标识为可以标识用户的任意信息，具体如用户账号或用户身份序列号等信息。所述查询时间为用户指定查询的时间范围，通常为当前时间或当前时间以前的某一个时刻或时间段。

[0071] 不同的用户对应用不同的用户标识。所述查询时间可理解为用户想用查询该段时间内产生的用户行为数据。这里的用户标识至少包括第一类用户标识。

[0072] 从本实施例可知在所述用户列表集中的数据是按用户标识进行排列的，针对于每一用户都记录着其用户行为产生可供查询的用户行为数据，如访问网站的时间、访问了哪些网站、访问这些网站产生的数据流量以及访问这些网站产生的流量类型。所述流量类型可包括第二代移动通信 2G 流量、第三代移动通信 3G 流量或第四代移动通信 4G 流量等。

[0073] 所述摘要字段相当于查询所述摘要数据的索引，当数据处理装置接收到对应的摘

要字段时，可根据摘要字段的匹配等处理，查询到对应的摘要数据。在具体实现时，为了对所述摘要字段的管理，减少索引失效的问题，通常所述摘要字段还归属于所述用户明细集，由用户明细集对进行数据查询的索引（包括摘要字段和关联字段）同一进行管理和维护。

[0074] 所述方法还包括：

[0075] 基于所述用户列表集及所述用户明细集，建立索引表；

[0076] 其中，所述索引表的查询索引包括关联字段以及所述摘要字段。

[0077] 在本实施例中所述方法还包括索引表。根据前述技术方案可知，查询数据可利用所述摘要字段和所述关联字段，在本实施例中为了方便对这些具有数据查询字段的同一关联，避免形成多个索引表和多种索引类型，在本实施例中将关联字段和摘要字段都存到索引表中，减少了索引类型，避免了索引类型混乱及索引存储占用存储空间大的问题，减少了数据冗余。

[0078] 所述索引表还包括所述摘要数据。

[0079] 在本实施例中所述索引表还直接包括摘要数据，利用所述摘要字段就能直接查询所述索引表中的摘要数据，这样的话，所述摘要字段相当于非关联索引。

[0080] 在本实施例中所述索引表中的关联字段，还可以关联到主表中进行主表中数据的查询，相当于关联索引。这就实现了非关联索引和关联索引的同一管理和处理。

[0081] 当然摘要数据也可以复用为所述关联字段。比如，第一类用户可能会查询之前自己访问过的网站，但是第二类用户可能会查询第一类用户访问过的网站的累积访问数等信息。这个实收第一类用户访问过的网站是可供第一类用户查询的数据，作为摘要数据存储在用户列表集中。当时第一类用户访问过的网站也作为关联字段，存储在用户明细集中，这个时候为了进一步减少数据冗余，可以将用户明细集和用户列表集中的数据共同形成前述索引表。在索引表中有些输入即作为摘要数据，也作为关联字段，这样就能尽可能减少数据冗余，减少数据存储及维护占用的资源。

[0082] 在具体的实现过程中，在所述索引表中的每一行元素中可至少包括一个所述摘要字段和一个所述关联字段。当然所述索引表中每一行元素还可包括是否做主表关联查询参数的属性，这个属性对应的参数值为假，则表示该行元素中的查询索引为摘要字段，可以直接返回该摘要字段对应的摘要数据即可。通常该摘要数据也存储在该行元素中。当所述是否做主表表关联参数的属性为真时，表示该行元素对应的查询索引为所述关联字段，是可用于关联到主表进行查询的。此外，索引表中还可包括关联元素个数的属性，若关联元素个数为0，也表示对应的摘要字段，否则为关联字段。

[0083] 总之所述索引表中的每一个数据或字段都可以作为用户查询数据的索引，但是对应的字段具体为摘要字段还是关联字段，则可以通过设定却分两种字段的参数来表示，如关联元素个数或是否关联属性的参数。

[0084] 如图2所示，所述方法还包括：

[0085] 步骤S210：接收基于用户输入形成的查询标签；

[0086] 步骤S220：将所述查询标签与所述索引表中的字段进行匹配；

[0087] 步骤S230：若所述查询标签与所述摘要字段相匹配，则基于所述摘要字段查询所述摘要数据，并返回所述摘要数据；

[0088] 步骤S240：若所述查询标签与所述关联字段相匹配，则基于所述关联字段，查询

所述主表并返回查询结果。

[0089] 所述查询标签可为用户输入的查询索引，或电子设备结合用户输入的信息和用户标识等信息生成的可用于与用户明细列表中的数据进行匹配的数据。

[0090] 显然本实施例中提供了一种数据查询方法，能够提供统一的数据查询接口，将查询标签与索引表中的字段的匹配，可以快速通过非关联的方式返回摘要数据，同时通过关联的方式返回出主表中的数据，具有数据处理效率快、数据冗余度小及索引管理简便的特点。

[0091] 设备实施例：

[0092] 如图 3 所示，本实施例提供一种数据处理装置，所述装置包括：

[0093] 筛选单元 110，用于对用户行为数据进行数据筛选处理，形成数据筛选结果；

[0094] 划分单元 120，用于基于所述数据筛选结果，将用户行为数据划分为满足查询需求的摘要数据和满足数据分析处理的系统数据；其中，所述摘要数据归属于用户列表集；所述系统数据和所述摘要数据均属于系统数据集；

[0095] 生成单元 130，用于基于所述系数数据集形成查询所述系统数据集的关联字段；其中，所述关联字段归属于用户明细集。

[0096] 本实施例中所述的数据处理装置可对应于各种形式的能够进行数据处理的电子设备，如台式电脑、笔记本电脑、服务器或服务器平台等。

[0097] 所述筛选单元 110、划分单元 120 及生成单元 130 所对应的结构可包括处理结构和存储介质。所述处理结构可包括处理器和处理电路。所述处理器可包括应用处理器 AP、数字信号处理器 DSP、可编程阵列 PLC、数字信号处理器 DSP 或微处理器 MCU 等结构。所述处理电路可包括专用集成电路 ASIC。所述存储介质可以通过总线接口等连接结构与所述处理结构相连。所述存储介质上存储有可执行代码，所述处理结构可以通过执行所述可执行代码实现上述单元的功能。

[0098] 上述任意两个单元可分别对应不同的处理结构，也可以集成对应于同一个所述处理结构。所述处理结构集成对应多个单元时，所述处理结构采用时分复用或并发线程的方式，分别完成不同单元的操作。

[0099] 在本实施例中所述摘要数据和所述系统数据的区分可以详细参见对应方法实施例，在此就不重复了。

[0100] 如图 4 所示，所述装置还包括：

[0101] 第一建立单元 130，用于基于所述系统数据集，建立与所述关联字段关联的主表。

[0102] 本实施例所述第一建立单元可包括存储介质，所述存储介质用于存储所述主表。所述第一建立单元用于根据主表的建立函数或策略，基于所述系统数据集中的数据，形成所述主表。

[0103] 所述主表与关联字段之间有关联关系，这种关联关系可以用于第二类用户查询到所述主表内的数据。

[0104] 所述用户列表集还包括摘要字段；其中，所述摘要字段与所述摘要数据具有映射关系，能够用于查询所述摘要数据。

[0105] 在本实施例中所述用户列表中直接包括能够第一类用户查询的摘要数据和摘要字段。这样所述数据处理装置通过摘要字段能过在毫秒级内快速查询到所述摘要数据。

[0106] 在本实施例中，所述摘要字段包括用户标识及查询时间；所述摘要数据是基于用户标识分布和形成的数据，这样方便快速查询。

[0107] 作为本实施例的进一步改进，所述装置还包括：第二建立单元150，用于基于所述用户列表集及所述用户明细集，建立索引表；其中，所述索引表的查询索引包括关联字段以及所述摘要字段。

[0108] 在本实施例中所述第二建立单元150的具体结构与所述第一建立单元的结构类似，不同的是：所述第二建立单元150是用于建立索引表的结构。在所述索引表中包括关键字段和关联字段，这样通过索引表统一对索引进行管理，能够避免索引混乱等各种问题。

[0109] 所述索引表还包括所述摘要数据。在本实施例中直接将所述摘要数据在所述索引表中进行维护，这样通过与索引表中的字段匹配，确定了对应查询标签与关键字段匹配，就直接将维护在索引表中的摘要数据返回，这样就能最大限度的提高查询速率；且相对于摘要字段和摘要数据分别存储映射，最大限度的减少了数据冗余，减少数据占用的存储资源和数据维护资源。

[0110] 作为本实施例的进一步改进，如图4所示，所述装置还包括：

[0111] 接收单元210，用于接收基于用户输入形成的查询标签；

[0112] 匹配单元220，用于将所述查询标签与所述索引表中的字段进行匹配；

[0113] 第一查询单元230，用于若所述查询标签与所述摘要字段相匹配，则基于所述摘要字段查询所述摘要数据，并返回所述摘要数据；

[0114] 第二查询单元240，用于若所述查询标签与所述关联字段相匹配，则基于所述关联字段，查询所述主表并返回查询结果。

[0115] 本实施例所述的接收单元210的具体结构可包括通信接口，用于接收基于用户输入形成的查询标签。通常情况下，查询摘要数据的标签可为用户标识加上查询事件。若是需要查询所述主表中的数据，通常可能是查询时间加上用户标识，再加上业务字段等信息。

[0116] 所述第一查询单元230和第二查询单元240的具体结构都可为对应于具有信息查询功能的处理结构；所述处理结构的具体描述可以参见前述部分。总之，本实施例所述的装置在查询数据时，结合使用关联索引和非关联索引，能够快速的查询到用户想要的数据，同时具有数据的冗余度小，占用的系统资源少等特点。

[0117] 以下结合上述任意实施例提供几个具体示例。

[0118] 示例一：

[0119] 图5所示的为基于前述实施例中对数据进行划分的流程。

[0120] 将HBase中的用户行为数据进行划分，划分为：满足用户实时查询需求的数据和满足系统处理需求的数据。再进一步将满足用户实施查询需求的数据划分归属于用户列表集的数据和归属于用户明细集的数据。而仅满足系统处理需求的数据则将全部归属于系统数据集。

[0121] 所述用户行为数据划分的过程可如下：

[0122] 第一步：满足用户行为实时查询的需求抽象：

[0123] 因为用户行为类数据放在HBase里的原因主要是由于用户行为实时查询的需求引起的。所以首先我们对用户行为实时查询数据的特征进行了抽象分析。

[0124] 由于用户的注意力范围有限有限，一个用户通常不可能同时关注大量的数据内

容。所以用户喜欢把浏览数据的过程分为两步来执行。我们根据这样的行为特点将数据展示分为两个步骤：展示摘要的数据和展示明细数据。

[0125] 我们基于这样一种数据展示方式，把将来用来让用户查询的数据分为两类的记录集，一个是用户列表集，一个是用户明细集。用户摘要集对应的是用户浏览大批数据的摘要需求，用户明细集对应的是查询明细的需求。这两个数据集合都是全部用户行为数据的子集。

[0126] 那么这两个数据集是怎么从全部数据中筛选出来的呢？因为用户只会关心自己的数据，而且很有可能是某一段时间的。所以这两个数据集是通先按用户标识字段过滤，然后再按时间字段过滤，就得到了所需要的数据集。

[0127] 第二步：满足数据分析处理需求的抽象：

[0128] 刚刚我们分析了用户实时查询的需求，但是在一个大数据共享平台中，用户行为类数据不可能只为用户实时查询所使用。还有可能会被系统内部的数据分析和处理。

[0129] 下面分析一下系统内部的数据分析和处理的业务特征。

[0130] 系统内的数据分析和处理的场景会根据具体分析的内容有所不同，不同的分析所关心的数据列也不同，这也是列存储的由来，所以从理论上来说数据分析和处理过程中需要表中任何一个字段都有可能。同时在大数据共享平台的分析处理中，通常不会关心某一个用户的数据，而是关心大量用户的数据，但基本都是按时间周期进行增量处理。不管是使用多表联的 SQL 处理，还是 MepReduce 处理都是这样。针对于这样的系统内部的数据分析和处理需求特征，我们得到了这样的一个“系统数据集合”，这个数据集包含了所有的数据字段，它通过时间字段来从全部数据中过滤出来。

[0131] 图 6 所示的可为根据图 5 所示方法，划分后的数据分布示意图。从图 6 可知，可能所有的用户行为数据都满足系统分析处理需求的数据，故所有的用户行为数据都归属于系统数据集。而满足用户实时查询需求的数据可包括归属于用户列表集和用户明细集中的数据。用户列表集中的数据包括摘要数据和关键字段，所述关键字段为查询所述摘要数据的查询索引。在图 6 中为了方便对索引的统一管理。所述用户明细集中除了包括对系统数据集中数据查询的关联字段以外，还包括用户明细集中的摘要字段。

[0132] 下表为比对三种数据集的异同。

[0133]

数据集合	包括字段	记录数	检索字段
系统数据集	全部字段	最大	时间+业务字段
户明细集	系统数据集合的子集	极少	用户标识+时间
用户列表集	系统数据集合的子集	介于另外两者之间	用户标识+时间

[0134] 显然从上表可知，用户列表集的数据可能介于用户明细集和系统数据集之间，但是用户明细集和用户列表集都是系统数据集的子集。所述检索字段为查询对应数据集中的索引。在本示例中查询用户列表集中的数据都可以采用用户标识加上时间的检索字段。这里的时间即为前述的查询时间。

[0135] 基于上述数据划分，进行数据结构的设置。因为系统数据集合包括全部的字段，所以适合用主表存储相关数据。用户明细集合有返回记录数极小的特点，非常适合通过对主表建一个关联索引（这里的关联索引即为前述的关联字段）来实现。因为 HBase 关联二级索引在记录较少时性能很高，空间占用也较少。用户列表集合因为返回记录数较大，使用 HBase 关联二级索引得到最好的性能，但因为这个集合的字段比其他两个集合都要小，比较适合使用 HBase 非关联二级索引实现。也就是索引中保存全部需要的摘要数据，查询时不作与主表的关联查询。这样性能较高，但是会占用一些空间，因为这个集合字段最少，相当于牺牲一点冗余换取最佳的性能。

[0136] 通过上述的设计，使得在各种查询的需求下都有较高的查询性能，同时数据存在了一些数据冗余，但冗余量很小。

[0137] 按如上述，将形成一个主表和两个索引。这里的两个索引即相当于前述实施例中的摘要字段和关联字段。摘要字段相当于非关联索引，而关联字段相当于关联索引。

[0138] 接下来把这两个索引进行合并。在具体实现时，非联合索引的内容包括了联合索引的内容，合并方法是以用户列表集合的索引内容作为合并后索引的内容，通过给索引器传递“是否作主表关联查询参数”，当参数为假时，直接返回索引表数据，和“关联元素个数”两个参数即可。

[0139] 在大数据共享平台的分析处理中，通常不会关心某一个用户的数据，而是关心大量用户的数据，但基本都是按时间周期进行增量处理。不管是使用多表联的 SQL 处理，还是 MepReduce 处理。

[0140] 系统数据集合主要针对内部分析处理的，通过主表以能较好的解决系统内部分析和处理时间性能问题。主表使用时间加上业务字段作为查询索引，也就是说主表中的数据分布是按时间在 HBase 中进行分布的，这样时我们在作系统内部分析处理时，能高效的定位到新增数据所在的位置。只取需要处理的数据。能够提高数据处理过程中对数据查询和

提取的速率。

[0141] 如图 7 所示，主表中的数据是按时间顺序组织排列的；而索引表中的数据是按用户账号进行组织排列的。当然用户账号为用户标识的其中一种。

[0142] 在索引表中包括摘要字段和摘要数据，其中，图 7 中显示的“5-6”、“6-5xxx”都可作为所述摘要字段映射的摘要数据。在图 7 中还显示索引值，此处的索引值相当于前述实施例中的关联字段，用于关联到主表进行对主表中数据的查询。

[0143] 在主表中存储着各种数据，这些数据当然也包括用户账号、查询时间、Fa 以及 Fb 组成的用户查询检索的业务字段等。

[0144] 索引表将用户 ID+ 时间戳 + 其他摘要字段串成一个字符串作产需行关键字，中间用分割符分割，用户查询摘要数据时，将自己的查询的用户 ID、时间范围及其他查询条件变成一种行关键字查询。通过这个行关键字查询快速定位数据位置。然后直接取出索引表数据，前面叙述过了索引表包括了所有摘要字段，所以直接从索引表返回数据，不用关联主表查询。这里的用户 ID 即为前述用户标识的一种。

[0145] 所以整个过程其实就是一个索引表的行关键字范围查询，因此效率较高。

[0146] 索引表与主表关联，索引表的索引值前半部分与主表行关键字相同，从使索引表与主表关联通关联起来，索引表的索引值后半部分是其它的用户摘要数据。这样的目的是因为索引表的所索引值有本身有两种用途，第一是与主表的行关键字关联，第二是能承载用户摘要数据。这样就通过索引值实现了关联索引与非关联索引进行合并的结果。比如这时索引值包括第一部分和第二部分；第一部分用于与摘要数据映射，第二部分用于与主表映射，这样第一部分相当于摘要字段，而第二部分相当于关联字段。这样的话一个索引值，可以同时用于查询到摘要数据和系统数据集中的数据。当然这种方式，是区分于前述设置特别的参数来区分关联字段和摘要字段的。相当于一个索引值对应两个字段，一个是关联还是非关联，另一个是关联元素个数。这样可实现数据保存一份，用于两种用途。

[0147] 当用户查询明细数据时，通过索引表关联主表。整个过程其实就是执行（返回记录数 *2) 次行关键字查询，因为明细数据查询的特点是查询少量数据，但有较多字段。所以关联时可以有较高的性能。

[0148] 在数据处理分析的场景里，处理过程通常不是以单个用户为中心的。基本都是按时间作增量处理，比如定位处理前一小时数据。不断新增数据不断作处理。

[0149] 数据分析处理系统获取数据，可以从主表中按行关键字查询通过 Hive、impala、spark 引擎查询及处理。或使用 MapReduce 直接通过 Rowkey 定位要处理的数据区间，避免全表扫描。

[0150] 示例二：

[0151] 下面通过一个示例来介绍本发明：

[0152] 将 A 口的用户上网行为数据进存储。支撑业务使用数据的需求。该数据通常有两类用途：

[0153] 用户查询上网记录。

[0154] 系统用该数据作用户上网行为分析。

[0155] 元数据如下：

[0156]

属性编码	字段名称	字段描述	备注
1	ID	记录 id	
2	UserAccount	用户帐号	
3	AccessPoint	接入点	GPRS：代表 APN，如 CMWAP、CMNET
4	AccessType	接入方式	GPRS 代表 1：2G; 2：3G; 3：4G 0：未知;
5	IMEI	国际移动电话设备识别码	手机的国际移动电话设备识别码号
6	PositionFlag	位置标识	GPRS：代表 lac-ci,lac 和 ci 都是十进制，通过中划线相连
7	BearingLayer_Protocol	承载层协议	1:TCP 2:UDP
8	ApplicationLayer_Protocol	应用层协议编号	HTTP、FTP、SMTP、QQ、SIP 等
9	BusinessID	业务编号	详细的业务类型编号，例如，QQ，迅雷、新

[0157]

			浪微博、微信
10	BusinessEntrance_ID	业务入口编号	业务入口编号就是软件名称编号。从 UA 中获取应用软件信息
11	BusinessType_ID	业务类别编号	即时通信、微博等
12	BusinessSupplier_ID	业务提供商编号	
13	StartTime	开始时间	业务开始时间；URL 的记录代表访问时间
14	UploadPackage	上行包数	单位：个
15	DownloadPackage	下行包数	单位：个
16	UploadTraffic	上行流量	单位：字节（Byte）
17	DownloadTraffic	下行流量	单位：字节（Byte）
18	WebsiteID	网站名称编号	
19	WebstieSubID	网站频道编号	
20	ResourceID	资源 IP 地址	点分格式如：192.168.1.1 提供有网站的 IP 地址。 没有 URL 的记录，该字段为空
21	ResourceCarrier	资源归属运营商	根据 IP 地址来判别；区分运营商：

[0158]

			0: 移动 1: 电信 2: 联通 3: 铁通 4: 教育 6: 国内其他 7: 国外
22	ResourcdeProvince	资源归属省份	提供网络资源归属省份编码
23	URL	访问 URL	
24	UserAgent	用户代理商属性	

[0159] 定义数据集

[0160] 将字段按用户关注度排序，从高关注度向低关注度进行排序，进而选择出高关注度的字段作为摘要字段，并根据摘要字段确定出摘要数据集，再加上其余用户关心的数据作为明细数据集，全集就是系统分析处理数据集。这里的关注度可体现在用户查询该数据的频次上。排序结果如下：

[0161]

字段名称	字段描述
UserAccount	用户帐号
StartTime	开始时间
BusinessID	业务编号
BusinessType_ID	务类别编号
WebsiteID	站名称编号
WebstieSubID	站频道编号

[0162]

BusinessEntrance_ID	业务入口编号
BusinessSupplier_ID	业务提供商编号
UploadTraffic	上行流量
DownloadTraffic	下行流量
URL	访问 URL
UserAgent	用户代理商属性
AccessType	接入方式
ID	记录 id
AccessPoint	接入点
IMEI	国际移动电话设备识别码
ApplicationLayer_Protocol	应用层协议编号
…	系统字段（略）

[0163] 在上表形成的排序结果中，将用户账号、开始时间、业务编号、业务类别编号及网站名称编号作为摘要字段，这些摘要字段对应的数据作为摘要数据。将上表中从用户账号到应用层协议编号作为用户明细数据集中的数据字段。其他的系统字段归属于系统数据集。

[0164] 定义主表

[0165] 主表是用时间 + 业务字段作为行关键字的，在本例子中使用开始时间加上用户账号连起来作为主表的关键字。（时间作为行关键字 rowkey 前缀是固定的，后面部分可以根据业务需求来设计），全部数据都保存在主表里

[0166] 定义索引表

[0167] 将用户列表集内的字段联合起来建成索引。所述用户列表集内字段可包括：用户帐号、业务编号、业务类别编号、开始时间、网站名称编号。

[0168] 索引表是以用户账号为前缀的，服务于用户查询。用户查询数据时按该用户的用户账号作快速匹配。

[0169] 用户查询上网记录时。

[0170] 用户查询他某一时间段上网记录时，通过用户帐号和时间段（这里的时间段即

相当于前述的查询时间），以及选择业务类型（即时通信、微博等）等查询上网记录是数据摘要部分。用户可以看到相应的摘要数据。示例中他将会看到，他使用过什么软件(BusinessType_ID)访问了什么网站(WebsiteID)等信息。这些数据量可能有很多，有可能会有分页。系统内部是通过HBase二级索引进行行关键字匹配，完成快速查询。因为不关联主表，即使查询记录有几万条，也能保证在数秒内返回。

[0171] 用户查询明细数据时，用户看了摘要数据后，选择出他要进一步查看的明细数据。明细数据有更多的信息详情（字段），但数据记录数很小。系统内部是通过按用户帐号、时间段等对HBase二级索引进行行关键字查询，然后再关联到主表进行查询。这个关联过程比不关联性能要低，但是由于明细查询记录数较小，所以秒级返回也有保障。

[0172] 系统对用户行为数据进行分析处理。

[0173] 例如：要分析哪个网站用户访问数最多。系统不可能只分析一次，而是按时间窗口（比如一天）作增量分析。所以要求数据要按时间分布，方便只处理最近一天的数据。我们的主表是按时间和业务字段作rowkey的，也就是按时间分布的，同时主表包含所有字段。分析时直接对主表进行分析，而不用走索引表。从HBase取数据时，本质上是对主表进行范围的rowkey查询，从HBase中取出最近一天的数据进行汇总，而不用扫描以往的数据。

[0174] 在本申请所提供的几个实施例中，应该理解到，所揭露的设备和方法，可以通过其它的方式实现。以上所描述的设备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，如：多个单元或组件可以结合，或可以集成到另一个系统，或一些特征可以忽略，或不执行。另外，所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口，设备或单元的间接耦合或通信连接，可以是电性的、机械的或其它形式的。

[0175] 上述作为分离部件说明的单元可以是、或也可以不是物理上分开的，作为单元显示的部件可以是、或也可以不是物理单元，即可以位于一个地方，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全部单元来实现本实施例方案的目的。

[0176] 另外，在本发明各实施例中的各功能单元可以全部集成在一个处理模块中，也可以是各单元分别单独作为一个单元，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。

[0177] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于一计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：移动存储设备、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0178] 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，本发明的保护范围应以所述权利要求的保护范围为准。

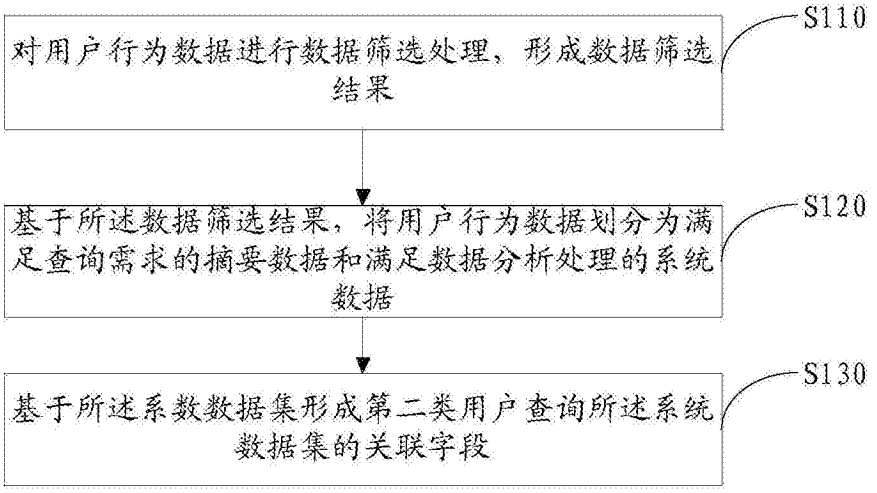

对用户行为数据进行数据筛选处理，形成数据筛选结果
S110
基于所述数据筛选结果，将用户行为数据划分为满足查询需求的摘要数据和满足数据分析处理的系统数据
S120
基于所述系数数据集形成第二类用户查询所述系统数据集的关联字段
S130

图 1

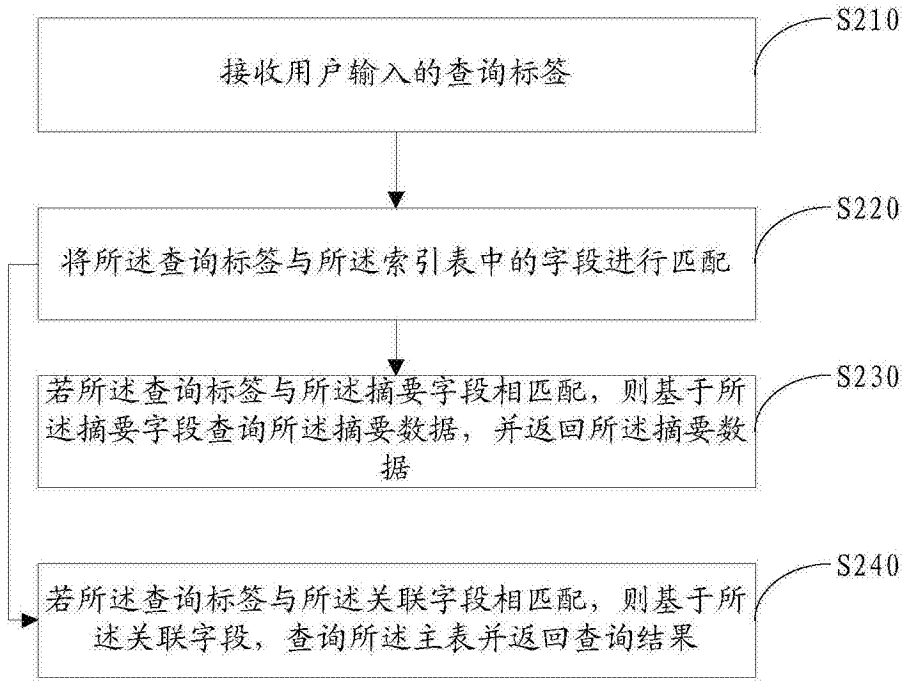

接收用户输入的查询标签
S210
将所述查询标签与所述索引表中的字段进行匹配
S220
若所述查询标签与所述摘要字段相匹配，则基于所述摘要字段查询所述摘要数据，并返回所述摘要数据
S230
若所述查询标签与所述关联字段相匹配，则基于所述关联字段，查询所述主表并返回查询结果
S240

图 2

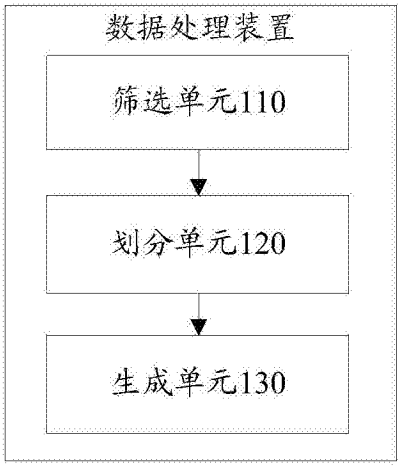

数据处理装置
筛选单元110
划分单元120
生成单元130

图 3

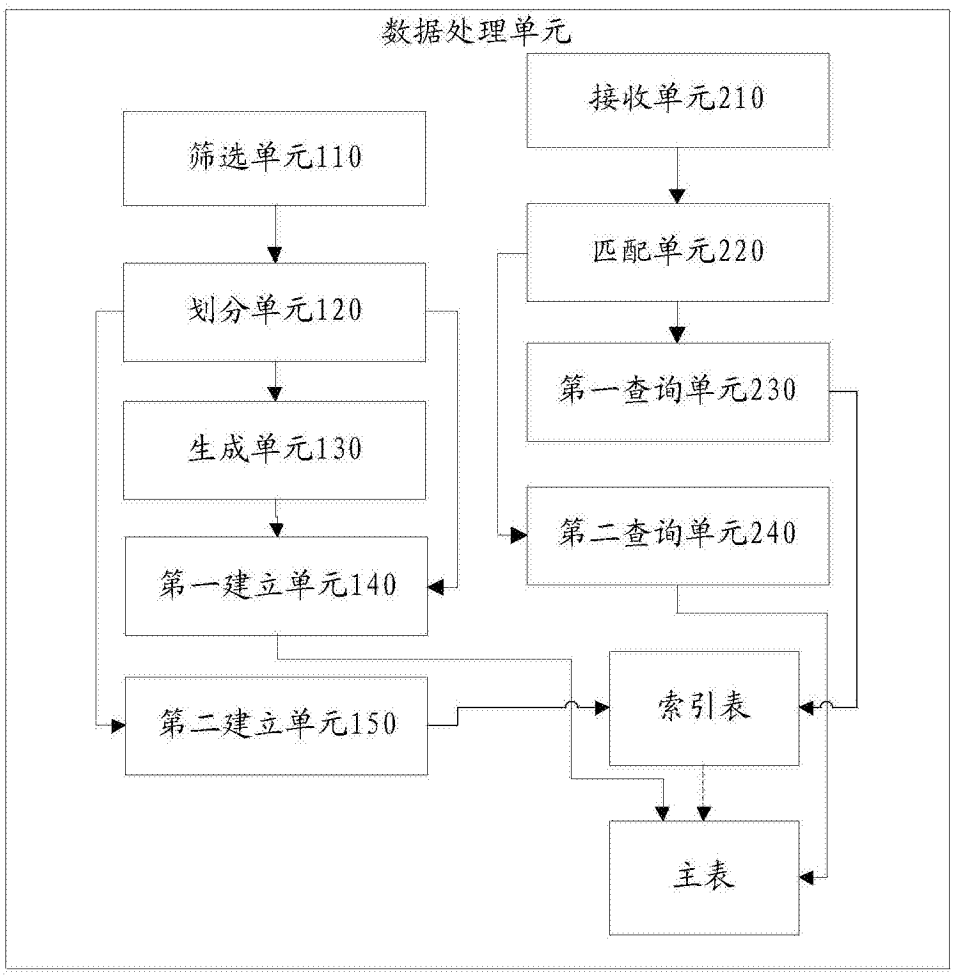

数据处理单元
接收单元210
筛选单元110
匹配单元220
划分单元120
第一查询单元230
生成单元130
第二查询单元240
第一建立单元140
索引表
第二建立单元150
主表

图 4

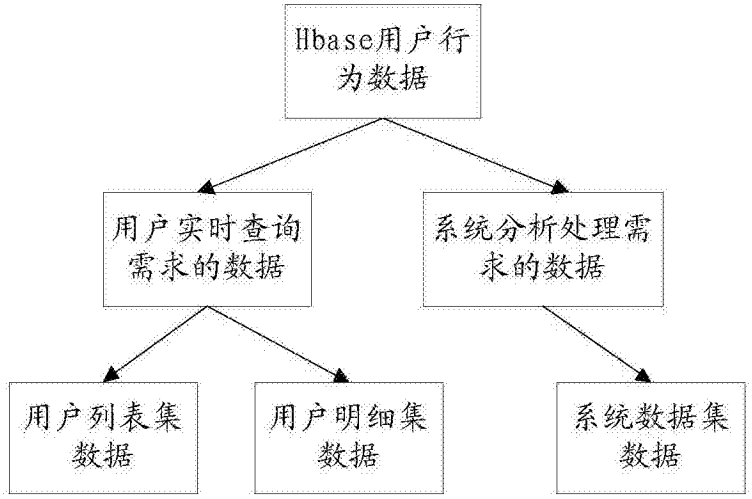

Hbase用户行为数据
用户实时查询需求的数据
系统分析处理需求的数据
用户列表集数据
用户明细集数据
系统数据集数据

图 5

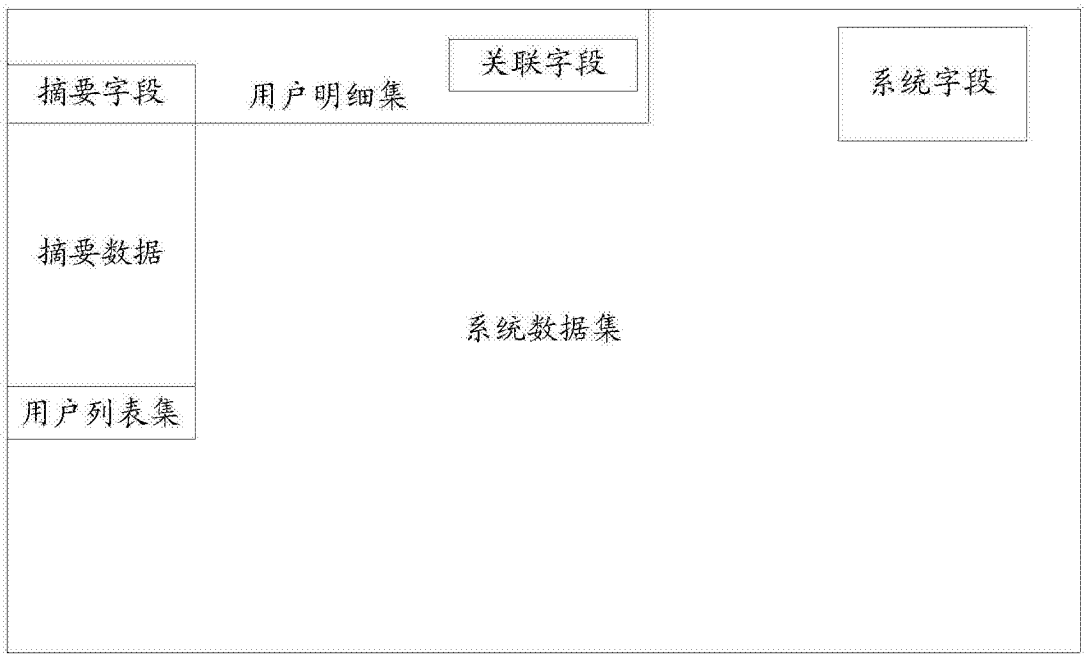

关联字段
摘要字段
用户明细集
系统字段
摘要数据
系统数据集
用户列表集

图 6

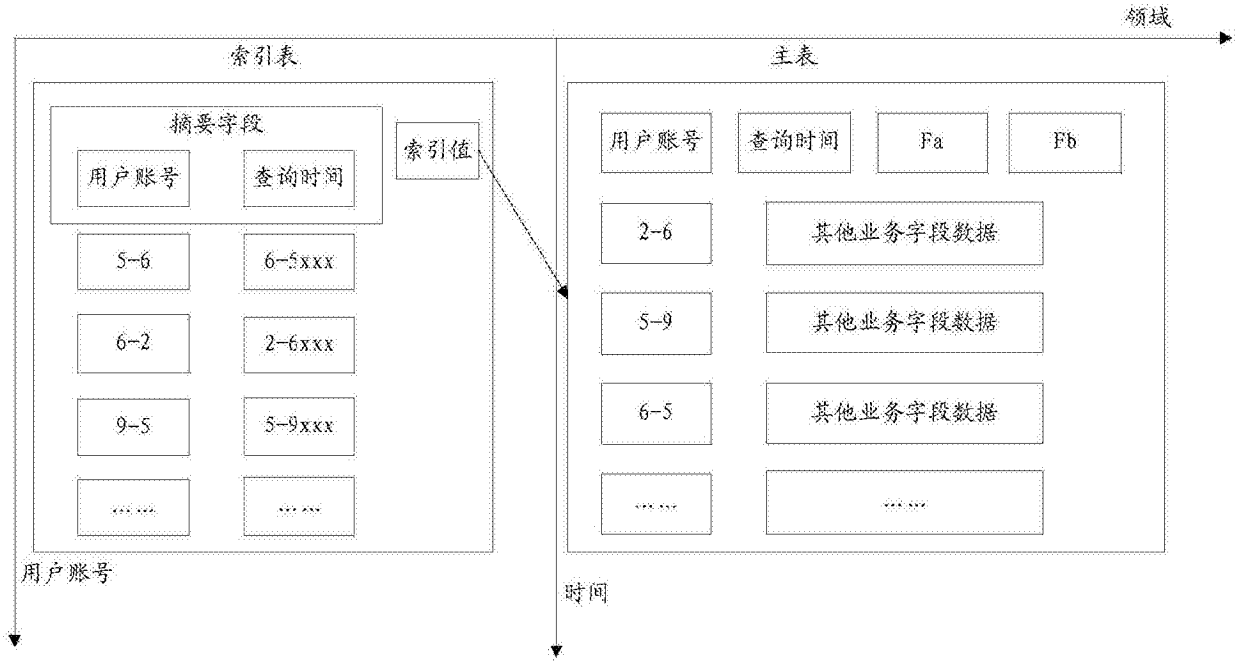

领域
索引表
主表
摘要字段
用户账号
查询时间
索引值
5-6
6-5xxx
6-2
2-6xxx
9-5
5-9xxx
……
……
用户账号
查询时间
Fa
Fb
2-6
其他业务字段数据
5-9
其他业务字段数据
6-5
其他业务字段数据
……
……
时间

图 7