(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 114297204 A
(43)申请公布日 2022.04.08

(21)申请号 202111677171.4

(22)申请日 2021.12.31

(71)申请人 奇安信科技集团股份有限公司
地址 100032 北京市西城区新街口外大街28号102号楼3层332号
申请人 网神信息技术(北京)股份有限公司

(72)发明人 高羽

(74)专利代理机构 北京鼎佳达知识产权代理事务所(普通合伙) 11348
代理人 刘铁生 孟阿妮

(51)Int.Cl.
G06F 16/22(2019.01)
G06F 16/245(2019.01)
G06F 16/25(2019.01)
G06F 16/2457(2019.01)

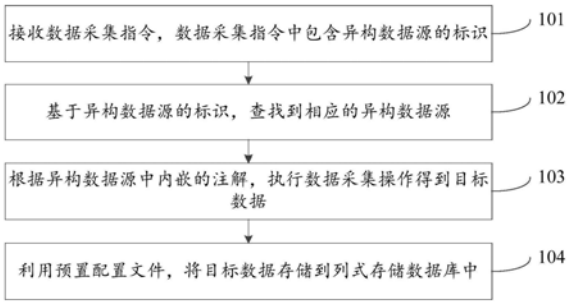
权利要求书2页 说明书11页 附图5页

(54)发明名称

一种异构数据源的数据存储、检索方法及装置

(57)摘要

本发明公开了一种异构数据源的数据存储、数据检索方法及装置,涉及异构数据源处理技术领域,能够优化对异构数据源的采集、存储和检索操作,降低处理成本,提高处理效率,也便于后续维护管理。本发明的主要技术方案为:接收数据采集指令,所述数据采集指令中包含异构数据源的标识;基于所述异构数据源的标识,查找到相应的异构数据源,其中,所述异构数据源中内嵌有注解,所述注解用于配置所述异构数据源的数据采集方式;根据所述异构数据源中内嵌的注解,执行数据采集操作得到目标数据;利用预置配置文件,将所述目标数据存储到列式存储数据库中,进而以利用这样的列式存储数据库提供对异构数据源的检索操作。



CN 114297204 A

1.一种异构数据源的数据存储方法,其特征在于,所述方法包括:

接收数据采集指令,所述数据采集指令中包含异构数据源的标识;

基于所述异构数据源的标识,查找到相应的异构数据源,其中,所述异构数据源中内嵌有注解,所述注解用于配置所述异构数据源的数据采集方式;

根据所述异构数据源中内嵌的注解,执行数据采集操作得到目标数据;

利用预置配置文件,将所述目标数据存储到列式存储数据库中。

2.根据权利要求1所述的方法,其特征在于,所述注解中包括所述异构数据源的接口和数据转换格式,所述根据所述异构数据源中内嵌的注解,执行数据采集操作得到目标数据,包括:

从所述注解中获取所述异构数据源的接口和数据转换格式;

根据所述接口从所述异构数据源中采集数据;

根据所述数据转换格式,对采集到的数据进行统一格式转换处理得到目标数据。

3.根据权利要求1所述的方法,其特征在于,所述利用预置配置文件,将所述目标数据存储到列式存储数据库中,包括:

从所述预置配置文件中获取多个预置属性字段、多个存储量阈值和多个存储地址,其中,每个所述存储量阈值关联了不同的存储地址;

从所述多个存储量阈值中获取大于所述目标数据的数据量所对应的存储量阈值,作为目标存储量阈值;

根据所述目标存储量阈值,从所述多个存储地址中确定对应关联的目标存储地址;

根据所述多个预置属性字段,在所述目标存储地址对应的存储空间上创建列式存储数据库;

将所述目标数据存储到所述列式存储数据库中。

4.根据权利要求3所述的方法,其特征在于,所述方法还包括:

从多个所述预置属性字段中获取目标属性字段;

设置所述目标属性字段对应的数据存储格式,以用于在所述目标属性字段内以所述数据存储格式存储数据信息;和/或,

若所述目标属性字段为多个,则通过对多个所述目标属性字段进行解析,得到所述多个目标属性字段之间存在的关联关系,以用于在将所述目标数据存储到所述列式存储数据库中时,优先向具有所述关联关系的目标属性字段内存储数据信息。

5.根据权利要求4所述的方法,其特征在于,所述将所述目标数据存储到所述列式存储数据库中,包括:

在利用所述多个预置属性字段存储所述目标数据的过程中,按照所述数据存储格式向所述目标属性字段存储相应的目标数据;

若所述目标属性字段为多个,则根据所述多个目标属性字段之间存在的关联关系,向所述多个目标属性字段存储相应的目标数据。

6.一种异构数据源的数据检索方法,其特征在于,应用于如权利要求1至5中任一项所述的异构数据源的数据存储方法所得到的列式存储数据库,所述方法包括:

接收数据检索指令,所述数据检索指令携带检索信息;

通过遍历所述列式存储数据库中的各个属性列,查找与所述检索信息匹配的检索结

果。

7.根据权利要求6所述的方法，其特征在于，所述通过遍历所述列式存储数据库中的各个属性列，查找与所述检索信息匹配的检索结果，包括：

从所述检索信息中解析出检索条件和检索关键字；

根据所述检索关键字，逐个遍历所述列式存储数据库内每个属性字段下的属性信息，查找匹配的目标属性信息；

根据所述目标属性信息，确定对应归属的目标属性字段；

在所述目标属性字段下，查找与所述检索条件匹配的检索结果。

8.一种异构数据源的数据存储装置，其特征在于，所述装置包括：

接收单元，用于接收数据采集指令，所述数据采集指令中包含异构数据源的标识；

查找单元，用于基于所述异构数据源的标识，查找到相应的异构数据源，其中，所述异构数据源中内嵌有注解，所述注解用于配置所述异构数据源的数据采集方式；

采集单元，用于根据所述异构数据源中内嵌的注解，执行数据采集操作得到目标数据；

存储单元，用于利用预置配置文件，将所述目标数据存储到列式存储数据库中。

9.一种异构数据源的数据检索装置，其特征在于，所述装置包括：

接收单元，用于接收数据检索指令，所述数据检索指令携带检索信息；

查找单元，用于通过遍历所述列式存储数据库中的各个属性列，查找与所述检索信息匹配的检索结果。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求1-5中任一项所述的异构数据源的数据存储方法；

或所述计算机程序被处理器执行时实现如权利要求6或7所述的异构数据源的数据检索方法。

11.一种电子设备，其特征在于，包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如权利要求1-5中任一项所述的异构数据源的数据存储方法；

或所述计算机程序被处理器执行时实现如权利要求6或7所述的异构数据源的数据检索方法。

一种异构数据源的数据存储、检索方法及装置

技术领域

[0001] 本发明涉及异构数据源处理技术领域，尤其涉及一种异构数据源的数据存储、数据检索方法及装置。

背景技术

[0002] 数据的检索操作可以横跨多种数据源，数据源即数据的来源之处，不同数据源包含了各行各业的业务活动产生的大量数据(例如，日志数据)，这些来源之处不同的大量数据会在数据结构、存储格式等等方面存在差异，构成了异构数据源，它所包含的数据量是巨大的，可以是百亿级的。

[0003] 目前，通常需要针对每种数据源，进行定制化的编写脚本，以实现数据解析和采集所需数据的功能，以及还需要存储采集到的数据并对外提供检索能力。然而，这对程序性能和可靠性有极高要求的，这相当于是需要进行定制化的开发，造成开发成本高，且不利于后续维护管理。

发明内容

[0004] 有鉴于此，本发明提供一种异构数据源的数据存储、数据检索方法及装置，主要目的在于利用通用方式实现对异构数据源的采集、存储和检索操作，优化了对异构数据源的相关数据处理方法，降低数据处理成本，提高数据处理效率，也有利于后续维护管理。

[0005] 本申请第一方面提供了一种异构数据源的数据存储方法，该方法包括：

[0006] 接收数据采集指令，所述数据采集指令中包含异构数据源的标识；

[0007] 基于所述异构数据源的标识，查找到相应的异构数据源，其中，所述异构数据源中内嵌有注解，所述注解用于配置所述异构数据源的数据采集方式；

[0008] 根据所述异构数据源中内嵌的注解，执行数据采集操作得到目标数据；

[0009] 利用预置配置文件，将所述目标数据存储到列式存储数据库中。

[0010] 在本申请第一方面的一些变更实施方式中，所述根据所述异构数据源中内嵌的注解，执行数据采集操作得到目标数据，包括：

[0011] 从所述注解中获取所述异构数据源的接口和数据转换格式；

[0012] 根据所述接口从所述异构数据源中采集数据；

[0013] 根据所述数据转换格式，对采集到的数据进行统一格式转换处理得到目标数据。

[0014] 在本申请第一方面的一些变更实施方式中，所述利用预置配置文件，将所述目标数据存储到列式存储数据库中，包括：

[0015] 从所述预置配置文件中获取多个预置属性字段、多个存储量阈值和多个存储地址，其中，每个所述存储量阈值关联了不同的存储地址；

[0016] 从所述多个存储量阈值中获取大于所述目标数据的数据量所对应的存储量阈值，作为目标存储量阈值；

[0017] 根据所述目标存储量阈值，从所述多个存储地址中确定对应关联的目标存储地

址；

[0018] 根据所述多个预置属性字段，在所述目标存储地址对应的存储空间上创建列式存储数据库；

[0019] 将所述目标数据存储到所述列式存储数据库中。

[0020] 在本申请第一方面的一些变更实施方式中，所述在根据所述多个预置属性字段，在所述目标存储地址对应的存储空间上创建列式存储数据库之后，所述方法还包括：

[0021] 从多个所述预置属性字段中获取目标属性字段；

[0022] 设置所述目标属性字段对应的数据存储格式，以用于在所述目标属性字段内以所述数据存储格式存储数据信息；和/或，

[0023] 若所述目标属性字段为多个，则通过对多个所述目标属性字段进行解析，得到所述多个目标属性字段之间存在的关联关系，以用于在将所述目标数据存储到所述列式存储数据库中时，优先向具有所述关联关系的目标属性字段内存储数据信息。

[0024] 在本申请第一方面的一些变更实施方式中，所述将所述目标数据存储到所述列式存储数据库中，包括：

[0025] 在利用所述多个预置属性字段存储所述目标数据的过程中，按照所述数据存储格式向所述目标属性字段存储相应的目标数据；

[0026] 若所述目标属性字段为多个，则根据所述多个目标属性字段之间存在的关联关系，向所述多个目标属性字段存储相应的目标数据。

[0027] 本申请第二方面提供了一种异构数据源的数据检索方法，应用于如上所述的异构数据源的数据存储方法所得到的列式存储数据库，该方法包括：

[0028] 接收数据检索指令，所述数据检索指令携带检索信息；

[0029] 通过遍历所述列式存储数据库中的各个属性列，查找与所述检索信息匹配的检索结果。

[0030] 在本申请第二方面的一些变更实施方式中，所述通过遍历所述列式存储数据库中的各个属性列，查找与所述检索信息匹配的检索结果，包括：

[0031] 从所述检索信息中解析出检索条件和检索关键字；

[0032] 根据所述检索关键字，逐个遍历所述列式存储数据库内每个属性字段下的属性信息，查找匹配的目标属性信息；

[0033] 根据所述目标属性信息，确定对应归属的目标属性字段；

[0034] 在所述目标属性字段下，查找与所述检索条件匹配的检索结果。

[0035] 本申请第三方面提供了一种异构数据源的数据存储装置，该装置包括：

[0036] 接收单元，用于接收数据采集指令，所述数据采集指令中包含异构数据源的标识；

[0037] 查找单元，用于基于所述异构数据源的标识，查找到相应的异构数据源，其中，所述异构数据源中内嵌有注解，所述注解用于配置所述异构数据源的数据采集方式；

[0038] 采集单元，用于根据所述异构数据源中内嵌的注解，执行数据采集操作得到目标数据；

[0039] 存储单元，用于利用预置配置文件，将所述目标数据存储到列式存储数据库中。

[0040] 在本申请第三方面的一些变更实施方式中，采集单元包括：

[0041] 获取模块，用于从所述注解中获取所述异构数据源的接口和数据转换格式；

[0042] 采集模块，用于根据所述接口从所述异构数据源中采集数据；

[0043] 处理模块，用于根据所述数据转换格式，对采集到的数据进行统一格式转换处理得到目标数据。

[0044] 在本申请第三方面的一些变更实施方式中，所述存储单元包括：

[0045] 获取模块，用于从所述预置配置文件中获取多个预置属性字段、多个存储量阈值和多个存储地址，其中，每个所述存储量阈值关联了不同的存储地址；

[0046] 所述获取模块，还用于从所述多个存储量阈值中获取大于所述目标数据的数据量所对应的存储量阈值，作为目标存储量阈值；

[0047] 确定模块，用于根据所述目标存储量阈值，从所述多个存储地址中确定对应关联的目标存储地址；

[0048] 创建模块，用于根据所述多个预置属性字段，在所述目标存储地址对应的存储空间上创建列式存储数据库；

[0049] 存储模块，用于将所述目标数据存储到所述列式存储数据库中。

[0050] 在本申请第三方面的一些变更实施方式中，所述存储单元还包括：

[0051] 所述获取模块，还用于从多个所述预置属性字段中获取目标属性字段；

[0052] 设置模块，用于设置所述目标属性字段对应的数据存储格式，以用于在所述目标属性字段内以所述数据存储格式存储数据信息；

[0053] 建立模块，用于当所述目标属性字段为多个时，通过对多个所述目标属性字段进行解析，得到所述多个目标属性字段之间存在的关联关系，以用于在将所述目标数据存储到所述列式存储数据库中时，优先向具有所述关联关系的目标属性字段内存储数据信息。

[0054] 在本申请第三方面的一些变更实施方式中，所述存储模块还具体用于：

[0055] 在利用所述多个预置属性字段存储所述目标数据的过程中，按照所述数据存储格式向所述目标属性字段存储相应的目标数据；

[0056] 当所述目标属性字段为多个时，根据所述多个目标属性字段之间存在的关联关系，向所述多个目标属性字段存储相应的目标数据。

[0057] 本申请第四方面提供了一种异构数据源的数据检索装置，该装置包括：

[0058] 接收单元，用于接收数据检索指令，所述数据检索指令携带检索信息；

[0059] 查找单元，用于通过遍历所述列式存储数据库中的各个属性列，查找与所述检索信息匹配的检索结果。

[0060] 在本申请第四方面的一些变更实施方式中，所述查找单元包括：

[0061] 解析模块，用于从所述检索信息中解析出检索条件和检索关键字；

[0062] 查找模块，用于根据所述检索关键字，逐个遍历所述列式存储数据库内每个属性字段下的属性信息，查找匹配的目标属性信息；

[0063] 确定模块，用于根据所述目标属性信息，确定对应归属的目标属性字段；

[0064] 所述查找模块，还用于在所述目标属性字段下，查找与所述检索条件匹配的检索结果。

[0065] 本申请第五方面提供了一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现如上所述的异构数据源的检索方法，以及实现如上所述的异构数据源的数据检索方法。

[0066] 本申请第六方面提供了一种电子设备，包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如上所述的异构数据源的检索方法，以及实现如上所述的异构数据源的数据检索方法。

[0067] 借由上述技术方案，本发明提供的技术方案至少具有下列优点：

[0068] 本发明提供了一种数据存储、数据检索方法及装置，本发明是预先向待采集的异构数据源内添加注解，继而在接收到数据采集指令时，根据异构数据源内嵌的注解，执行数据采集操作得到目标数据，然后在利用预置配置文件将目标数据存储到列式存储数据库中，据此以通用方式实现了对每个异构数据源的采集和存储操作。以及在接收到检索指令时利用该列式存储数据库能够反馈相应的检索结果。相较于现有技术，解决了因针对每个异构数据源定制化开发提供采集、存储和检索功能导致成本高、不利于后续维护管理的问题，本发明优化对异构数据源的采集和存储操作，无需针对每个异构数据源定制化开发去达到采集和存储的功能，即无需侵入性的修改编写代码，而是以通用方式实现采集和存储功能，同时相应地也提供了通用的检索功能，据此通用化处理方式，降低处理成本，效率高，也利于后续维护管理。

[0069] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0070] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0071] 图1为本发明实施例提供的一种异构数据源的数据存储方法流程图；

[0072] 图2为本发明实施例提供的另一种异构数据源的数据存储方法流程图；

[0073] 图3为本发明实施例提供的一种异构数据源的数据检索方法流程图；

[0074] 图4为本发明实施例提供的另一种异构数据源的数据检索方法流程图

[0075] 图5为本发明实施例提供的一种异构数据源的数据存储装置的组成框图；

[0076] 图6为本发明实施例提供的另一种异构数据源的数据存储装置的组成框图；

[0077] 图7为本发明实施例提供的一种异构数据源的数据检索装置的组成框图；

[0078] 图8为本发明实施例提供的另一种异构数据源的数据检索装置的组成框图。

具体实施方式

[0079] 下面将参照附图更详细地描述本发明的示例性实施例。虽然附图中显示了本发明的示例性实施例，然而应当理解，可以以各种形式实现本发明而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本发明，并且能够将本发明的范围完整的传达给本领域的技术人员。

[0080] 本发明实施例提供了一种异构数据源的数据存储方法，如图1所示，该方法是针对不同异构数据源以通用方式实现了采集和存储功能，对此本发明实施例提供以下具体步骤：

[0081] 101、接收数据采集指令，数据采集指令中包含异构数据源的标识。

[0082] 其中，数据源即数据的来源之处，不同数据源包含了各行各业的业务活动产生的大量数据(例如，日志数据)，这些来源之处不同的大量数据会在数据结构、存储方式等等方面存在差异，构成了异构数据源。

[0083] 在本发明实施例中，接收到的数据采集指令中可以包含一个或多个数据源的标识的，该标识用于指示数据采集对象。

[0084] 102、基于异构数据源的标识，查找到相应的异构数据源。

[0085] 103、根据异构数据源中内嵌的注解，执行数据采集操作得到目标数据。

[0086] 其中，异构数据源中内嵌有注解，注解用于配置异构数据源的数据采集方式，具体的，注解可理解为代码里的特殊标记，这些标记可以在编译，类加载，运行时被读取，并执行相应的处理，通过注解开发人员可以在不改变原有代码和逻辑的情况下在源代码中嵌入补充信息。

[0087] 在本发明实施例中，针对不同异构数据源，注解都是通用的，而不是针对每个数据源实现某些指定操作而定制开发编写的脚本，在确定采集对象(即数据源)之后，注解是被预先写入到数据源中的。

[0088] 具体的，在代码层面，可以是利用一个软件开发工具包(Software Development Kit，SDK)将注解写入到数据源的起始端，注解相当于是代码里的特殊标志，这些标志可以在编译，类加载，运行时被读取，并执行相应的处理，以便于其他工具补充信息或者进行部署。对于本发明实施例中，就是通过运行数据源内嵌的注解执行数据采集操作，得到目标数据。

[0089] 104、利用预置配置文件，将目标数据存储到列式存储数据库中。

[0090] 其中，预置配置文件至少包含了预置属性字段和存储地址，预置属性字段用于指代利用哪些属性字段对目标数据进行存储；存储地址用于指代将目标数据存储在哪里，例如内存或服务器。以及此外如果存储位置需要授权登录，那么在预置配置文件中还可以存储相应的账号和密码。

[0091] 其中，列式存储数据库是指按照列存储索引存储采集到的目标数据，列式属性是以关系数据库中的属性(列)为单位进行数据存储的，在数据表中将同一属性的数据信息存储在一起，而一条记录中不同的属性的属性信息则分别存放在不同的存储单元中。

[0092] 在本发明实施例中，预置配置文件相当于是给出了存储目标数据所依据的标准。以及结合这样的存储标准，采用列式存储数据库实施存储目标数据的作用为：对于多个异构数据源而言，它们的数据量可以是百亿级的，采集这些异构数据源所得到的属性信息的种类数量也是很大的，继而利用创建的数据库存储这些属性信息所需要的属性字段数量也是很大的。如果按照行式存储方式存储数据信息，那么在检索操作时就会遍历查找大量的属性字段内的数据信息，这将浪费太多检索处理成本的，导致检索效率低。但是如果采用列式存储方式存储数据信息，在检索操作时基于列存储索引，即遍历一个属性字段内的属性信息，才会查找到下一个属性字段，因此无需遍历所有属性字段，就能够查找到匹配的检索结果了。

[0093] 以上，本发明实施例提供了一种数据存储方法，本发明实施例是预先向待采集的异构数据源内添加注解，继而在接收到数据采集指令时，根据异构数据源内嵌的注解，执行

数据采集操作得到目标数据，然后在利用预置配置文件将目标数据存储到列式存储数据库中，据此以通用方式实现了对每个异构数据源的采集和存储操作。相较于现有技术，解决了因针对每个异构数据源定制化开发提供采集、存储功能导致成本高、不利于后续维护管理的问题，本发明优化对异构数据源的采集和存储操作，无需针对每个异构数据源定制化开发去达到采集和存储的功能，即无需侵入性的修改编写代码，而是以通用方式实现采集和存储功能，降低数据采集和存储的处理成本，提高数据处理效率，也利于后续维护管理。

[0094] 为了对上述实施例做出更加详细的说明，本发明实施例还提供了另一种异构数据源的数据存储方法，如图2所示，该方法是对上述实施例的细化解释说明，对此本发明实施例提供以下具体步骤：

[0095] 201、接收数据采集指令，数据采集指令中包含异构数据源的标识。

[0096] 在本发明实施例中，本步骤解释说明，参见步骤101，此处不再赘述。

[0097] 202、基于异构数据源的标识，查找到相应的异构数据源。

[0098] 203、根据异构数据源中内嵌的注解，执行数据采集操作得到目标数据。

[0099] 其中，异构数据源中内嵌有注解，注解用于配置异构数据源的数据采集方式，注解中包括异构数据源的接口和数据转换格式。

[0100] 在本发明实施例中，利用异构数据源中内嵌的注解实现数据采集的具体实施方法包括如下：

[0101] 首先是，从注解中获取异构数据源的接口和数据转换格式。

[0102] 其次是，根据接口从异构数据源中采集数据，根据数据转换格式，对采集到的数据进行统一格式转换处理得到目标数据。

[0103] 需要说明的是，对于多个异构数据源，它们在数据结构、存储格式等等方面存在差异，因此需要利用本发明实施例提及的数据转换格式，实现对从多个异构数据源内采集到的数据信息进行标准化格式处理，从而得到标准统一的数据信息，作为目标数据，用于后续存储操作。

[0104] 204、从预置配置文件中获取多个预置属性字段、多个存储量阈值和多个存储地址，其中，每个存储量阈值关联了不同的存储地址。

[0105] 其中，预置属性字段是从多个技术领域的数据库整合得到的标准化通用的属性字段。具体的，对于本发明实施例从异构数据源采集到的目标数据，由于是来自于不同异构数据源的，且多个异构数据源所包含的属性种类数量也是很大的，因此，本发明实施例为了实现对更多样性的数据信息进行有效存储，可以从这些异构数据源中获取属性字段来构建一个标准化通用的属性字段。

[0106] 例如，通过对多个异构数据源内包含的属性字段进行取交集处理，得到公共属性字段；其次是，根据公共属性字段，衍生属性字段。具体的，对于衍生属性字段，可以采用语义分析方法，例如：对相似语义的属性字段进行合并，或者基于高频使用的属性字段扩展更多下位概念细化分支的属性字段。进一步的说明，多个异构数据源包含的属性种类是多样的且数量大，对于那些很低频出现的属性种类的数据，它的数据价值不高，可以将其从采集到的目标数据中数据清洗掉，而不存储到列式存储数据库中，具体的，可以不在数据库内设置这些很低频的属性字段，以避免数据库内存储过多无用冗余数据而占用存储资源了。

[0107] 其中，存储量阈值是指存储空间的数据容量阈值，存储地址是指存储空间的地址，

每个存储量阈值关联了不同的存储地址，以便于根据采集到目标数据的数据量大小，选择相应的存储空间，需要说明的是，对于采集到的目标数据，可以包括动态数据和静态数据，动态数据是常常变化，直接反映事务过程的数据，比如，网站访问量、在线人数、日销售额等等，由于动态数据更新周期短且数据量大，因此优化处理方式，是将动态数据和静态数据分开处理。

[0108] 在本发明实施例中，利用预置配置文件中的预置属性字段、存储量阈值和存储地址，能够自动完成对目标数据创建相应数据库完成存储。

[0109] 205、从多个存储量阈值中获取大于目标数据的数据量所对应的存储量阈值，作为目标存储量阈值。

[0110] 206、根据目标存储量阈值，从多个存储地址中确定对应关联的目标存储地址。

[0111] 在本发明实施例中，对于步骤205-206解释说明，预置配置文件中存储了多个存储量阈值，每个存储量阈值对应关联一个存储地址。为了避免出现存储地址对应的存储空间不足以容纳存储目标数据的情况，因此可以优先将目标数据的数据量与不同存储量阈值进行比较，查找大于目标数据的数据量的存储量阈值所对应的存储地址，当然若这样的存储地址有多个，则可以从中选择任意一个即可，从而利用存储地址获取到存储空间，去完成存储目标数据。其中，每个存储地址都可以包括：成对存储的起始地址和结束地址。

[0112] 在另外的实施例中，为了提升存储空间的利用率，在查找到大于目标数据的数据量的存储量阈值所对应的多个存储地址后，可以从多个存储地址中选择最小的存储量阈值对应的存储地址，从而确保相应的地址空间尽可能多地被占用，防止资源浪费。

[0113] 207、根据多个预置属性字段，在目标存储地址对应的存储空间上创建列式存储数据库。

[0114] 在本发明实施例中，主要是根据预置配置文件中的预置属性字段去创建列式存储数据库的，那么该列式存储数据库内就包含了这些预置属性字段，进一步的，还可以对这些预置属性字段做出优化处理，具体包括如下：

[0115] 从多个预置属性字段中获取目标属性字段，即需要优化处理的预置属性字段，简称为目标属性字段，对目标属性字段优化处理方式为：

[0116] 例如，通过对目标属性字段进行预处理，设置目标属性字段对应的数据存储格式。

[0117] 示例性的，在一些特定领域有特殊含义的字段，或者例如根据实际业务需求，对于省、市、区、县的名称存储时，需要同时存储文字名称和其映射的编码。

[0118] 又或者，作为并列方案，若目标属性字段为多个，则通过对多个目标属性字段进行解析，得到多个目标属性字段之间存在的关联关系。基于这种关联关系，可以优先向具有关联关系的目标属性字段内存储数据信息，例如，在向一个目标属性字段存储数据信息之后，若该目标属性字段与另一个目标属性字段存在关联关系，则基于这种关联关系，优先向该另一个目标属性字段内存储数据信息，然后再根据其他属性字段去完成对数据库的数据存储操作。

[0119] 需要说明的，基于关联关系进行数据存储，不仅在存储时，能有效提升数据存储效率，在后续检索时，基于一个目标属性字段，能够带出与其存在关联关系的其他目标属性字段，从而有效提升数据的查询效率。

[0120] 示例性的，如上提及的省、市、区、县分别对应不同属性字段，可以预先按照一定排

序建立这些属性字段之间的关联关系，从而基于该关联关系，按照这个排序依次向省、市、区、县分别对应不同属性字段内存储数据信息，由于采集到的目标数据也可能是存在关联的，那么基于具有关联关系的目标属性字段去完成优先存储操作，提高了数据存储效率。

[0121] 208、将目标数据存储到列式存储数据库中。

[0122] 在本发明实施例中，通过对采集得到的目标数据进行数据解析处理，确定需要存储到的属性字段下，从而完成存储操作，但是需要对步骤207预先设定的目标属性字段进行说明如下：

[0123] 在利用多个预置属性字段存储目标数据的过程中，按照数据存储格式向目标属性字段存储相应数据。以及若目标属性字段为多个，则根据多个目标属性字段存在的关联关系，向多个目标属性字段存储相应数据。

[0124] 进一步的，在本发明实施例提供的异构数据源的数据存储方法得到的列式存储数据库基础之上，本发明实施例还提供了一种异构数据源的数据检索方法，如图3所示，对此本发明实施例提供以下具体步骤：

[0125] 301、接收数据检索指令，数据检索指令携带检索信息。

[0126] 302、通过遍历列式存储数据库中的各个属性列，查找与检索信息匹配的检索结果。

[0127] 在本发明实施例中，列式存储数据库中存储了来自于多个异构数据源内的数据信息，这相当于是通过数据采集和存储的方式对多个异构数据源内数据信息进行了整合预处理。那么对于接收到检索指令，它可以间接相当于是面向这些异构数据源的检索操作，因此通过检索列式存储数据库查找到检索结果，等同了向原异构数据源发起检索指令得到的检索结果。

[0128] 对于本发明实施例中，基于检索该列式存储数据库得到检索结果，替代了原本各自向每个异构数据源发起检索指令去获取相关数据信息，大大提高了检索效率和检索通用性。

[0129] 进一步的，为了对检索操作的细化解释说明，本发明实施例还提供了另一种异构数据源的数据检索方法，如图4所示，对此本发明实施例提供以下具体步骤：

[0130] 401、接收数据检索指令，数据检索指令携带检索信息。

[0131] 402、从检索信息中解析出检索条件和检索关键字。

[0132] 403、根据检索关键字，逐个遍历列式存储数据库内每个属性字段下的属性信息，查找匹配的目标属性信息。

[0133] 本发明实施例利用列式存储数据库可以将采集这些异构数据源所得到的大量种类的属性信息进行存储，即列式存储数据库内属性字段数量也是很多的，那么基于列式存储数据库的列式存储特点，也会在数据读取时能够实现整列读取数据，即检索到一个属性字段，能够读取整个属性列的数据信息，如此检索，可以避免大量属性字段被遍历，而是当确定检索关键字所在属性字段之后，不需要再遍历其他属性字段了，大大提高检索效率。

[0134] 404、根据目标属性信息，确定对应归属的目标属性字段。

[0135] 405、在目标属性字段下，查找与检索条件匹配的检索结果。

[0136] 在本发明实施例，在根据检索关键字，确定匹配的目标属性信息之后，再进一步的根据检索条件在该目标属性字段下完成检索操作，高效的获取到检索结果。

[0137] 进一步的，作为对上述图1、图2所示方法的实现，本发明实施例提供了一种异构数据源的数据存储装置。该装置实施例与前述方法实施例对应，为便于阅读，本装置实施例不再对前述方法实施例中的细节内容进行逐一赘述，但应当明确，本实施例中的装置能够对应实现前述方法实施例中的全部内容。该装置应用于针对不同异构数据源以通用方式实现了采集和存储操作，具体如图5所示，该装置包括：

[0138] 接收单元31，用于接收数据采集指令，所述数据采集指令中包含异构数据源的标识；

[0139] 查找单元32，用于基于所述异构数据源的标识，查找到相应的异构数据源，其中，所述异构数据源中内嵌有注解，所述注解用于配置所述异构数据源的数据采集方式；

[0140] 采集单元33，用于根据所述异构数据源中内嵌的注解，执行数据采集操作得到目标数据；

[0141] 存储单元34，用于利用预置配置文件，将所述目标数据存储到列式存储数据库中。

[0142] 进一步的，如图6所示，采集单元33包括：

[0143] 获取模块331，用于从所述注解中获取所述异构数据源的接口和数据转换格式；

[0144] 采集模块332，用于根据所述接口从所述异构数据源中采集数据；

[0145] 处理模块333，用于根据所述数据转换格式，对采集到的数据进行统一格式转换处理得到目标数据。

[0146] 进一步的，如图6所示，所述存储单元34包括：

[0147] 获取模块341，用于从所述预置配置文件中获取多个预置属性字段、多个存储量阈值和多个存储地址，其中，每个所述存储量阈值关联了不同的存储地址；

[0148] 所述获取模块341，还用于从所述多个存储量阈值中获取大于所述目标数据的数据量所对应的存储量阈值，作为目标存储量阈值；

[0149] 确定模块342，用于根据所述目标存储量阈值，从所述多个存储地址中确定对应关联的目标存储地址；

[0150] 创建模块343，用于根据所述多个预置属性字段，在所述目标存储地址对应的存储空间上创建列式存储数据库；

[0151] 存储模块344，用于将所述目标数据存储到所述列式存储数据库中。

[0152] 进一步的，如图6所示，所述存储单元34还包括：

[0153] 所述获取模块341，还用于从多个所述预置属性字段中获取目标属性字段；

[0154] 设置模块345，用于设置所述目标属性字段对应的数据存储格式，以用于在所述目标属性字段内以所述数据存储格式存储数据信息；

[0155] 建立模块346，用于当所述目标属性字段为多个时，通过对多个所述目标属性字段进行解析，得到所述多个目标属性字段之间存在的关联关系，以用于在将所述目标数据存储到所述列式存储数据库中时，优先向具有所述关联关系的目标属性字段内存储数据信息。

[0156] 进一步的，如图6所示，所述存储模块344还具体用于：

[0157] 在利用所述多个预置属性字段存储所述目标数据的过程中，按照所述数据存储格式向所述目标属性字段存储相应的目标数据；

[0158] 当所述目标属性字段为多个时，根据所述多个目标属性字段之间存在的关联关

系，向所述多个目标属性字段存储相应的目标数据。

[0159] 进一步的，作为对上述图3、图4所示方法的实现，本发明实施例提供了一种异构数据源的数据检索装置。该装置实施例与前述方法实施例对应，为便于阅读，本装置实施例不再对前述方法实施例中的细节内容进行逐一赘述，但应当明确，本实施例中的装置能够对应实现前述方法实施例中的全部内容。该装置应用于完成针对包含异构数据源的数据库的检索操作，具体如图7所示，该装置包括：

[0160] 接收单元41，用于接收数据检索指令，所述数据检索指令携带检索信息；

[0161] 查找单元42，用于通过遍历所述列式存储数据库中的各个属性列，查找与所述检索信息匹配的检索结果。

[0162] 进一步的，如图8所示，所述查找单元42包括：

[0163] 解析模块421，用于从所述检索信息中解析出检索条件和检索关键字；

[0164] 查找模块422，用于根据所述检索关键字，逐个遍历所述列式存储数据库内每个属性字段下的属性信息，查找匹配的目标属性信息；

[0165] 确定模块423，用于根据所述目标属性信息，确定对应归属的目标属性字段；

[0166] 所述查找模块422，还用于在所述目标属性字段下，查找与所述检索条件匹配的检索结果。

[0167] 综上所述，本发明实施例提供了一种数据存储、数据检索方法及装置，本发明是预先向待采集的异构数据源内添加注解，继而在接收到数据采集指令时，根据异构数据源内嵌的注解，执行数据采集操作得到目标数据，然后在利用预置配置文件将目标数据存储到列式存储数据库中，据此以通用方式实现了对每个异构数据源的采集和存储操作。以及在接收到检索指令时利用该列式存储数据库能够反馈相应的检索结果。相较于现有技术，解决了因针对每个异构数据源定制化开发提供采集、存储和检索功能导致成本高、不利于后续维护管理的问题，本发明实施例优化对异构数据源的采集和存储操作，无需针对每个异构数据源定制化开发去达到采集和存储的功能，即无需侵入性的修改编写代码，而是以通用方式实现采集和存储功能，同时相应地也提供了通用的检索功能，据此通用化处理方式，降低处理成本，效率高，也利于后续维护管理。

[0168] 所述异构数据源的数据存储装置包括处理器和存储器，上述接收单元、查找单元、采集单元和存储单元等均作为程序单元存储在存储器中，由处理器执行存储在存储器中的上述程序单元来实现相应的功能。

[0169] 所述异构数据源的数据检索装置包括处理器和存储器，上述接收单元、和查找单元等均作为程序单元存储在存储器中，由处理器执行存储在存储器中的上述程序单元来实现相应的功能。

[0170] 处理器中包含内核，由内核去存储器中调取相应的程序单元。内核可以设置一个或以上，通过调整内核参数以利用通用方式实现对异构数据源的采集、存储和检索操作，优化了对异构数据源的相关数据处理方法，降低数据处理成本，提高数据处理效率，也有利于后续维护管理。

[0171] 本发明实施例提供了一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现如上所述的异构数据源的数据存储方法，或异构数据源的数据检索方法。

[0172] 本发明实施例提供了一种电子设备，包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如上所述的异构数据源的数据存储方法，或异构数据源的数据检索方法。

[0173] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0174] 在一个典型的配置中，设备包括一个或多个处理器(CPU)、存储器和总线。设备还可以包括输入/输出接口、网络接口等。

[0175] 存储器可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)，存储器包括至少一个存储芯片。存储器是计算机可读介质的示例。

[0176] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0177] 还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0178] 本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0179] 以上仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同插入、改进等，均应包含在本申请的权利要求范围之内。

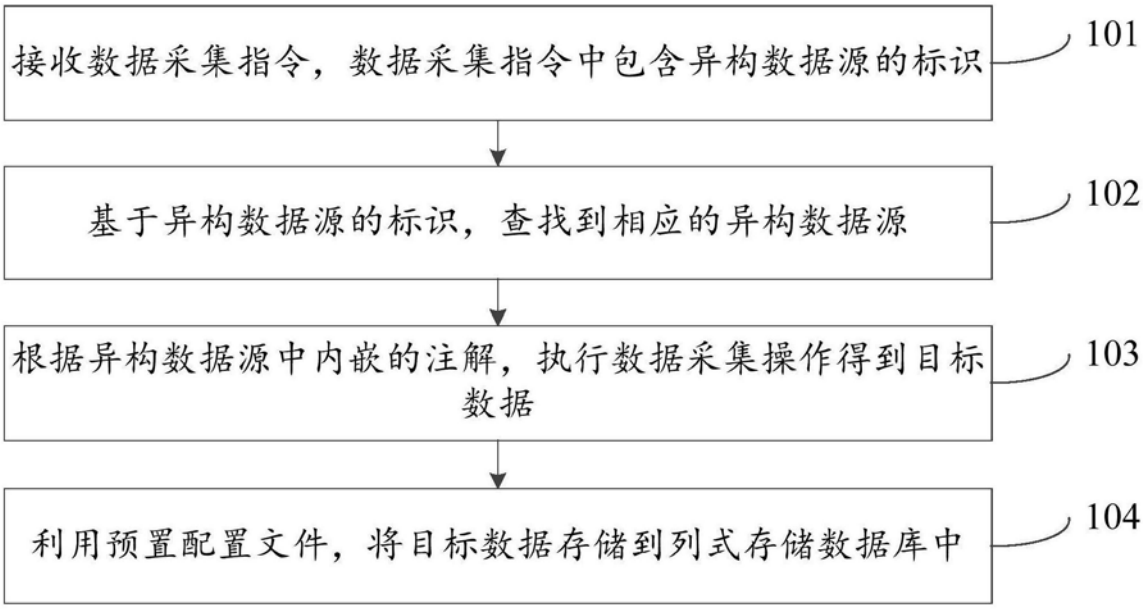

接收数据采集指令，数据采集指令中包含异构数据源的标识
101
基于异构数据源的标识，查找到相应的异构数据源
102
根据异构数据源中内嵌的注解，执行数据采集操作得到目标数据
103
利用预置配置文件，将目标数据存储到列式存储数据库中
104

图1

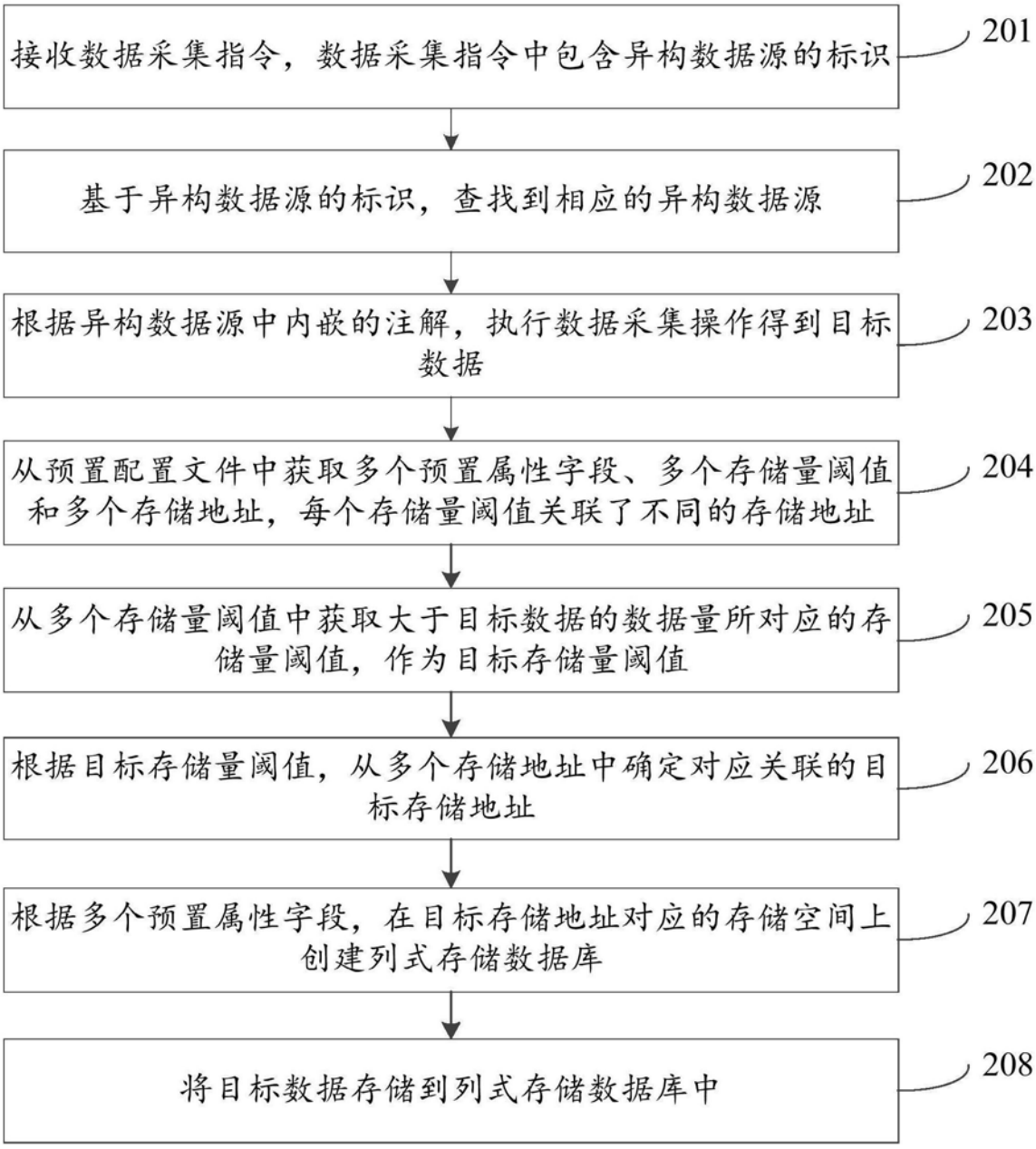

接收数据采集指令，数据采集指令中包含异构数据源的标识
201
基于异构数据源的标识，查找到相应的异构数据源
202
根据异构数据源中内嵌的注解，执行数据采集操作得到目标数据
203
从预置配置文件中获取多个预置属性字段、多个存储量阈值和多个存储地址，每个存储量阈值关联了不同的存储地址
204
从多个存储量阈值中获取大于目标数据的数据量所对应的存储量阈值，作为目标存储量阈值
205
根据目标存储量阈值，从多个存储地址中确定对应关联的目标存储地址
206
根据多个预置属性字段，在目标存储地址对应的存储空间上创建列式存储数据库
207
将目标数据存储到列式存储数据库中
208

图2

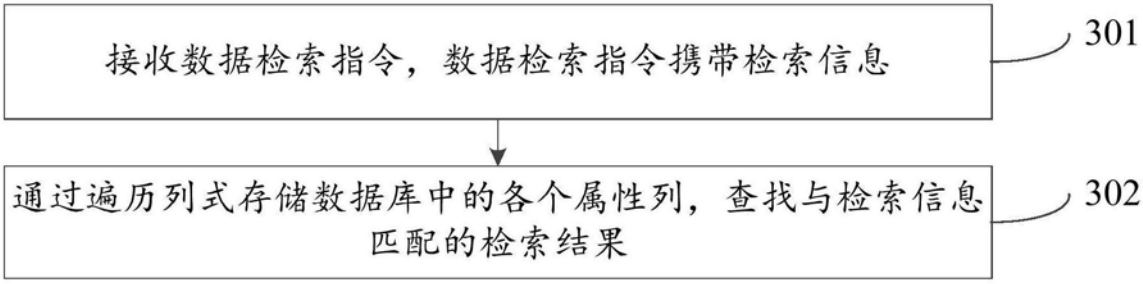

接收数据检索指令，数据检索指令携带检索信息
301
通过遍历列式存储数据库中的各个属性列，查找与检索信息匹配的检索结果
302

图3

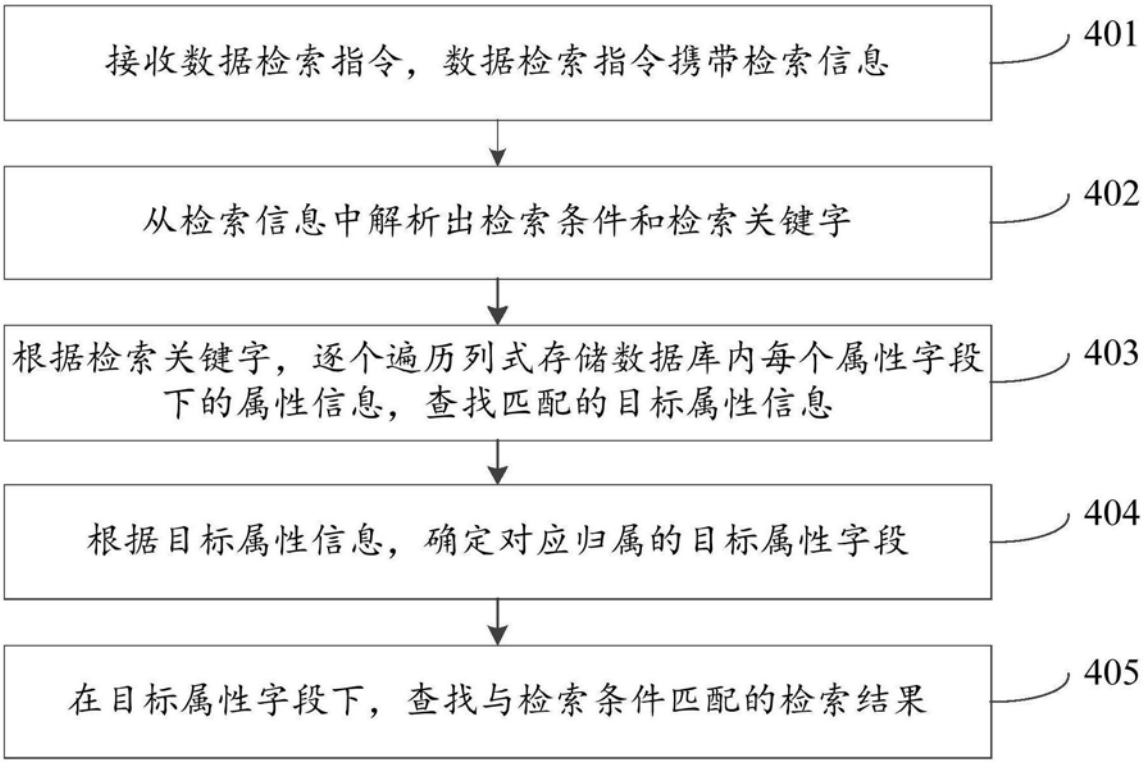

接收数据检索指令，数据检索指令携带检索信息
401
从检索信息中解析出检索条件和检索关键字
402
根据检索关键字，逐个遍历列式存储数据库内每个属性字段下的属性信息，查找匹配的目标属性信息
403
根据目标属性信息，确定对应归属的目标属性字段
404
在目标属性字段下，查找与检索条件匹配的检索结果
405

图4

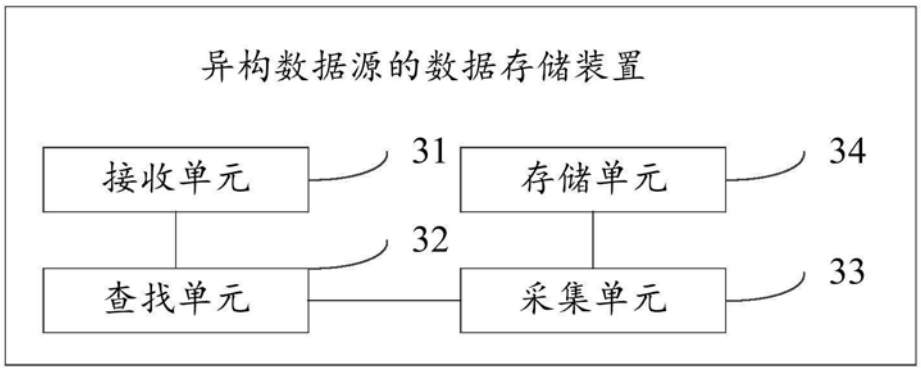

异构数据源的数据存储装置
接收单元
31
存储单元
34
查找单元
32
采集单元
33

图5

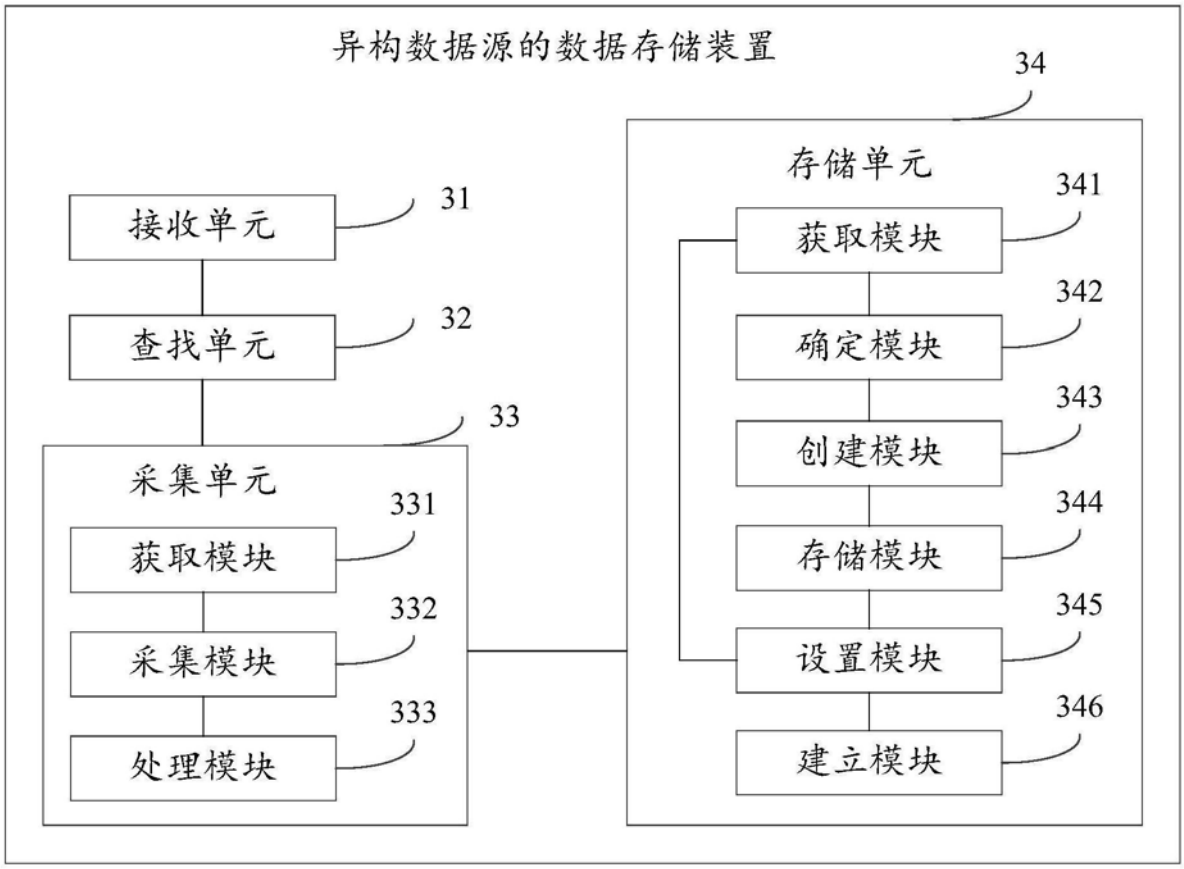

异构数据源的数据存储装置
34
存储单元
341
获取模块
342
确定模块
343
创建模块
344
存储模块
345
设置模块
346
建立模块
31
接收单元
32
查找单元
33
采集单元
331
获取模块
332
采集模块
333
处理模块

图6

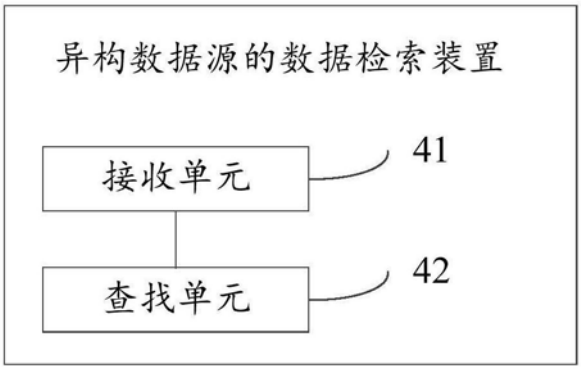

异构数据源的数据检索装置
41
接收单元
42
查找单元

图7

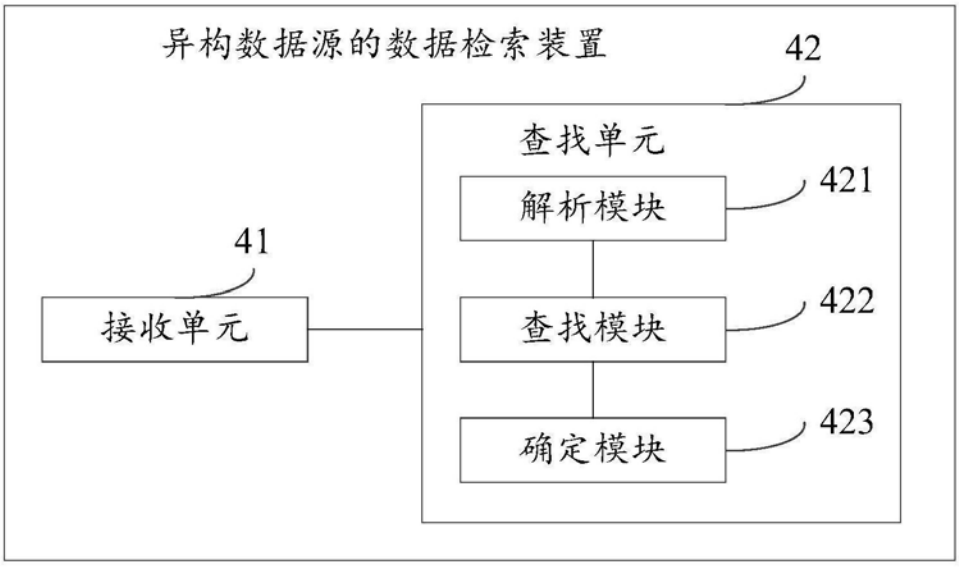


图8