(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111949710 B
(45)授权公告日 2024.03.22

(21)申请号 202010825330.X

(22)申请日 2020.08.17

(65)同一申请的已公布的文献号
申请公布号 CN 111949710 A

(43)申请公布日 2020.11.17

(73)专利权人 北京锐安科技有限公司
地址 100044 北京市海淀区西小口路66号中关村东升科技园北领地B-2号楼七层

(72)发明人 任丽超 谢永恒 程强

(74)专利代理机构 北京品源专利代理有限公司 11332
专利代理师 孟金喆

(51)Int.Cl.
G06F 16/2458(2019.01)
G06F 16/27(2019.01)
G06F 16/182(2019.01)

(56)对比文件
CN 106033452 A,2016.10.19
CN 109543463 A,2019.03.29
CN 111382123 A,2020.07.07

审查员 阚子雄

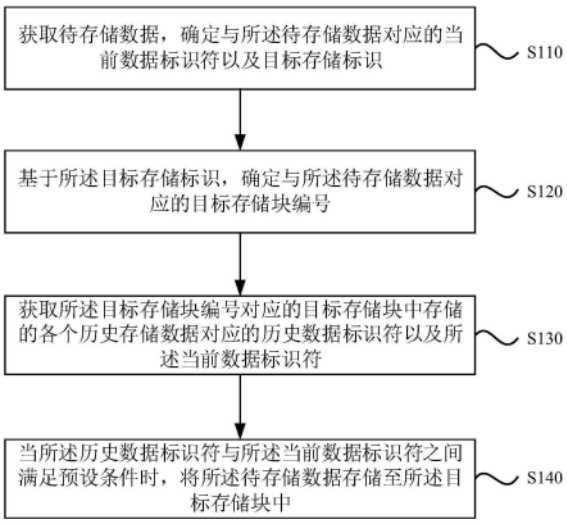
权利要求书2页 说明书9页 附图4页

(54)发明名称

数据存储方法、装置、服务器及存储介质

(57)摘要

本发明公开了一种数据存储方法、装置、服务器及存储介质。其中,该方法包括:获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识;基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号;获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符;当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中。本发明实施例能够实现海量结构化数据存储,并且实现了海量结构化数据的快速去重归并。



CN 111949710 B

1.一种数据存储方法,其特征在于,包括:

获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识;

基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号;

获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符;

当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中;

所述获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识,包括:

针对每一条所述待存储数据,确定与至少一个去重字段对应的关键信息;

采用哈希算法对所述关键信息进行处理,确定与所述待存储数据对应的所述当前数据标识符;

基于所述待存储数据所属数据集编号以及所述当前数据标识符,确定与所述待存储数据对应的所述目标存储标识。

2.根据权利要求1所述的方法,其特征在于,基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号,包括:

根据所述目标存储标识以及预先设置的目标函数,得到第一处理结果值;

根据所述第一处理结果值以及预先设置的存储块数量,确定与所述待存储数据对应的目标存储块编号。

3.根据权利要求1所述的方法,其特征在于,所述获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符,包括:

确定与所述目标存储块编号对应的目标存储块;

调取所述目标存储块中各个历史存储数据所对应的关联信息,以基于所述关联信息确定与各个历史存储数据对应的历史数据标识符。

4.根据权利要求1所述的方法,其特征在于,所述当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中,包括:

当所述历史数据标识符与所述当前数据标识符不一致,则将所述待存储数据缓存至所述目标存储块中,并建立与所述待存储数据相对应的关联信息。

5.根据权利要求1所述的方法,其特征在于,所述当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中,包括:

当所述历史数据标识符与所述当前数据标识符相一致时,则分别获取所述待存储数据所属的当前数据集标识,以及所述历史数据标识符对应的历史存储数据所属的历史数据集标识;根据所述当前数据集标识以及所述历史数据集标识,将所述待存储数据存储至所述目标存储块中;

当所述历史数据标识符与所述当前数据标识符不一致时,则将所述待存储数据按照预设格式存储至所述目标存储块中,并更新与所述待存储数据对应的关联信息。

6.根据权利要求5所述的方法,其特征在于,所述根据所述当前数据集标识以及所述历史数据集标识,将所述待存储数据存储至所述目标存储块中,包括:

当所述当前数据集标识与所述历史数据集标识相一致时,则将所述待存储数据删除,并基于所述待存储数据更新与所述历史数据标识符相对应的历史存储数据的关联信息;

当所述当前数据集标识与所述历史数据集标识不一致时,则将所述待存储数据按照预设格式存储至所述目标存储块中,并更新与所述待存储数据对应的关联信息。

7.一种数据存储装置,其特征在于,包括:

标识确定模块,用于获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识;

存储块编号确定模块,用于基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号;

标识获取模块,用于获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符;

数据存储模块,用于当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中;

所述标识确定模块包括:关键信息确定单元、当前数据标识符确定单元和目标存储标识确定单元,其中:

所述关键信息确定单元,用于针对每一条所述待存储数据,确定与至少一个去重字段对应的关键信息;

所述当前数据标识符确定单元,用于采用哈希算法对所述关键信息进行处理,确定与所述待存储数据对应的所述当前数据标识符;

所述目标存储标识确定单元,用于基于所述待存储数据所属数据集编号以及所述当前数据标识符,确定与所述待存储数据对应的所述目标存储标识。

8.一种服务器,其特征在于,所述服务器包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-6中任一所述的数据存储方法。

9.一种包含计算机可执行指令的存储介质,所述计算机可执行指令在由计算机处理器执行时用于执行如权利要求1-6中任一所述的数据存储方法。

数据存储方法、装置、服务器及存储介质

技术领域

[0001] 本发明实施例涉及数据处理技术,尤其涉及一种数据存储方法、装置、服务器及存储介质。

背景技术

[0002] 随着互联网应用的广泛普及,海量数据的存储成为了系统设计中不可或缺的一部分。

[0003] 目前海量数据存储的过程是在数据存储时,将对象数据转成字符串,并以特定文件格式存储到HDFS(Hadoop分布式文件系统)上。为了便于后期批量进行任务时寻找数据,数据存储时按业务、日期创建目录。接下来,运行MapReduce(超大机群上的简单数据处理)离线任务,即按业务要求读取数据,将全部海量数据加载到内存中,在内存中完成按归并维度进行归并、计数等操作,最终,输出全量数据,更新HBase(分布式面向列的开源数据库)等用于存储海量数据的数据库。

[0004] 面对千亿级海量结构化数据时,此种海量数据存储方式将全部海量数据加载到内存中,将会导致内存占用过大。当对所有数据进行计算时,对内存、CPU等硬件的要求过高,并且处理时间较长,处理效率较低。

发明内容

[0005] 本发明提供一种数据存储方法、装置、服务器及存储介质,解决海量数据存储占用存储空间大,数据处理效率低的问题,以实现数据的快速存储,并且降低了对硬件环境的要求,进一步提升数据处理效率。

[0006] 第一方面,本发明实施例提供了一种数据存储方法,其特征在于,包括:

[0007] 获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识;

[0008] 基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号;

[0009] 获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符;

[0010] 当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中。

[0011] 第二方面,本发明实施例还提供了一种数据存储装置,其特征在于,包括:

[0012] 标识确定模块,用于获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识;

[0013] 存储块编号确定模块,用于基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号;

[0014] 标识获取模块,用于获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符;

[0015] 数据存储模块,用于当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中。

[0016] 第三方面,本发明实施例还提供了一种服务器,其特征在于,所述服务器包括:

[0017] 一个或多个处理器;

[0018] 存储装置,用于存储一个或多个程序,

[0019] 当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现任一实施例所述的数据存储方法。

[0020] 第四方面,本发明实施例还提供了一种包含计算机可执行指令的存储介质,所述计算机可执行指令在由计算机处理器执行时用于执行任一实施例所述的数据存储方法。

[0021] 本发明实施例提供的技术方案,通过获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识;基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号;获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符;当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中,解决了存储过程中因数据量大而导致的数据处理效率低的问题,实现了数据的分块存储,避免了存储空间占用过大、内存占用过高的情况,进一步提升了数据处理效率。

附图说明

[0022] 图1为本发明实施例一所提供的一种数据存储方法的流程示意图;

[0023] 图2为本发明实施例二所提供的一种数据存储方法的流程示意图;

[0024] 图3为本发明实施例三所提供的一种数据存储装置结构框图;

[0025] 图4为本发明实施例四所提供的一种服务器结构示意图。

具体实施方式

[0026] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是,此处所描述的具体实施例仅仅用于解释本发明,而非对本发明的限定。另外还需要说明的是,为了便于描述,附图中仅示出了与本发明相关的部分而非全部结构。

[0027] 实施例一

[0028] 图1为本发明实施例提供的一种数据存储方法的流程示意图,本实施例可适用于海量结构化数据去重、归并和存储的情况,该方法可以由存储芯片的管理系统来执行,该系统可以通过软件和/或硬件的形式实现。

[0029] 在介绍本实施例技术方案之前,先简单介绍下应用场景。例如,可以将该方法应用到离线和/或在线的任务中,即离线或者在线对数据去重归并的情形。

[0030] 如图1所示,该方法具体包括如下步骤:

[0031] S110、获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识。

[0032] 其中,待存储数据可以是需要进行结构化去重的海量结构化数据,其中,也可以是不同数据集中的结构化数据。待存储数据可以是通过Kafka接入的,进一步,可以使用Flink开源框架进行实时数据处理。当前数据标识符是同一数据集中每一条待存储数据的特定标

识,可以是通过不同去重策略确定的标识。

[0033] 在数据的获取过程中,可以使用不止一个数据采集设备,同一采集设备也可以获取不同的数据。简言之,待存储数据所属的数据集不止一个。因此,为了区分不同数据集中的数据,每个数据集有其对应的唯一数据集编号。所述数据集编号可以是整型,字符串型等数据类型。目标存储标识是每一条数据的当前数据标识与所属数据集编号拼接后的标识,可以是拼接后经过数据处理的标识符。

[0034] S120、基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号。

[0035] 需要说明的是,为了采用分而治之的策略针对海量结构化数据进行存储,也就是将海量结构化数据分块存储,可以预先将存储空间划分为至少一个存储块,并对每个存储块进行编号,以便将待存储数据存储至相应的存储块中。示例性的,存储块数量可以设置为默认1000,那么,存储块编号为0~999。具体存储块数量的设置需要根据具体接入的数据量大小,以及任务集群主机情况而定。如果接入数据量大,且任务集群的主机内存比较小,则可以将存储块数量设置多一些,这样后续分块读取的数据量少,内存占用低。

[0036] 其中,目标存储块编号指的是存储待存储数据时所对应的存储块编号。目标存储块编号的确定可以是根据目标存储标识与存储块编号的映射关系来确定的。其中,映射关系可以是函数关系,也可以是人为设定的对应关系。

[0037] 示例性的,假设目标存储标识为正整数N,存储块数量为1000,映射关系为BlockNum=N%1000。那么,当目标存储标识N=32709时,BlockNum=709,即待存储数据对应的目标存储块编号为709。

[0038] S130、获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符。

[0039] 其中,历史存储数据是依据本实施例所提供的数据存储方法存储至相应存储块中的数据。数据存储块中包括至少一条历史存储数据,相应的,也包括与每一条历史存储数据对应的历史数据标识符。

[0040] 需要说明的是,向各个存储块中存储历史数据的存储方式以及存储格式可以是:历史数据存储路径可以为:日期/存储块编号,数据存储为SequenceFile文件,其中Key值为数据集编号、历史数据标识符、首次采集时间、最近采集时间、发现次数、累计发现天数等信息以逗号拼接;Value的值为空。为了降低各存储块存储的历史数据所占用的存储空间,可以使用Snappy压缩算法(快速无损数据压缩算法)对存储文件进行压缩处理。

[0041] S140、当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中。

[0042] 其中,预设条件是用来判断如何将待存储数据存储至相对应的目标存储块中的条件。

[0043] 其中,将待存储数据存储至目标存储块中的存储方式和格式可以是:将待存储数据存储为SequenceFile文件,其中,Key的值为数据集编号、存储块编号和当前数据标识符,以逗号拼接;Value的值为待存储数据的Protobuf格式字节数组。进一步的,可以使用Snappy压缩算法对存储文件进行压缩来降低存储空间。

[0044] 需要说明的是,执行本实施例所提供的数据存储方法,可以是在预设时间点执行。预设时间点包括相对时间点和绝对时间点,相对时间点指的是接收到消息的时间点,绝对

时间点可以是预先设置的时刻,例如,每天的零点。

[0045] 本发明实施例通过确定待存储数据的当前数据标识符以及目标存储标识,进而根据目标存储标识确定目标存储块编号,根据预设条件将待存储数据存储至目标存储块中,实现了海量数据的分块存储,达到了降低内存占用率以及提高数据处理效率的目的。

[0046] 实施例二

[0047] 图2是本发明实施例二提供的一种数据存储方法的流程图,本实施例是在实施例一的基础上的进一步优化。如图2所示,该方法具体包括:

[0048] S201、针对每一条待存储数据,确定与至少一个去重字段对应的关键信息。

[0049] 其中,待存储数据的数据量级一般在千亿级,空间占用可能达到几十T甚至上百T。因此,需要针对每一个数据集中的每一条待存储数据都进行去重、归并和存储。

[0050] 其中,本实施例首先可以根据去重策略对数据进行处理,取出至少一个去重字段中的关键信息。可以理解的是,去重策略是确定待去重的字段,即需要去重的字段,该策略可以是根据对数据的去重需求设定的。当去重字段存在多个时,可以将多个字段中的数值进行拼接。

[0051] 示例性的,假设对某一数据集中的数据进行存储,该数据集包含10个字段,记为[Field1,Field2,…,Field10],去重策略是根据Field1,Field3和Field5进行处理,可以得到的是该数据集中第i条数据经去重策略处理后的信息为Field1_i,Field3_i和Field5_i拼接后的信息,记为Field_i。

[0052] S202、采用哈希算法对所述关键信息进行处理,确定与所述待存储数据对应的当前数据标识符。

[0053] 其中,将待去重字段中的关键信息进行MD5(Message-Digest Algorithm 5,信息-摘要算法5)操作,获取其MD5值作为当前数据标识符。其中,MD5是在计算机领域被广泛使用的一种哈希算法,用来对信息进行完整性保护,典型的应用是针对一段信息产生信息摘要。

[0054] 示例性的,假设对某一数据集中的数据进行存储,待存储数据的去重字段拼接信息为Field。将Field进行MD5操作得到的MD5值就是待存储数据的当前数据标识符。

[0055] S203、基于所述待存储数据所属数据集编号以及所述当前数据标识符,确定与所述待存储数据对应的目标存储标识。

[0056] 本实施例首先获取待存储数据所属数据集编号,由于待存储数据中有不止一个数据集的数据,为了区分不同数据集中的数据,每个数据集有其对应的唯一数据集编号。所述数据集编号可以是整型,字符串型等数据类型。其次,获取待存储数据的当前数据标识符。最后,将待存储数据所属数据集编号和当前数据标识符进行拼接,通过Hash函数对拼接后的标识符进行处理,得到目标存储标识,其中,哈希值是一段数据唯一且极其紧凑的表示形式。

[0057] 示例性的,假设某一待存储数据的数据集编号为Dataset,当前数据标识符为Field。则针对数据集编号和当前数据标识符拼接后的信息取哈希值,即为目标存储标识H=(Dataset+Field).hashCode()。

[0058] S204、根据所述目标存储标识以及预先设置的目标函数,得到第一处理结果值。

[0059] 其中,预先设置的目标函数可以是针对目标存储标识进行数值化的函数,也可以是将目标存储标识转化为正整型数值的函数。

[0060] 示例性的,假设待存储数据的目标存储标识为H,预先设置的目标函数为:Num＝H&Integer.MAX_VALUE,其中,Num为第一处理结果值。其中,哈希值和Integer.MAX_VALUE进行与操作可以保证结果是非负数,便于后续得到分块编号。

[0061] S205、根据所述第一处理结果值以及预先设置的存储块数量,确定与所述待存储数据对应的目标存储块编号。

[0062] 其中,预先设置的存储块数量可以设置为默认1000,其中,存储块编号为0～999。具体存储块数量的设置需要根据具体接入的数据量大小,以及任务集群主机情况而定。如果接入数据量大,且任务集群的主机内存比较小,则可以将存储块数量设置多一些,这样后续分块读取的数据量少,内存占用低。

[0063] 进一步的,根据所述第一处理结果值与所述存储块数量,确定与所述第一处理结果值对应的余数值;基于所述余数值,确定与所述待存储数据对应的目标存储块编号。

[0064] 示例性的,假设待存储数据的第一处理结果值为Num,预设存储块数量为1000。那么,目标存储块编号BlockNum＝Num％1000,进而,对每一条待存储数据都确定一个对应的目标存储块编号。

[0065] 计算出存储块编号后,将待存储数据存储到HDFS的指定路径,其中,HDFS的路径可以是:日期/存储块编号/任务编号/数据文件编号。其中,任务编号为同时处理多个任务时,为每一个任务分配的编号,对于任务编号的形式没有具体限制。其中,数据文件编号表示暂存数据的文件编号,对于数据文件编号的形式没有具体限制。所述数据文件编号可以为分桶文件编号,其中,分桶文件是指将海量数据存储至相应的存储块内时,是预先存储到多个桶文件中的。其中,默认设置的是当桶文件大小达到384M或者数据存储时间超过3小时时,关闭桶文件,同时,开启下一个桶文件。

[0066] S206、调取所述目标存储块中各个历史存储数据所对应的关联信息,以基于所述关联信息确定与各个历史存储数据对应的历史数据标识符。

[0067] 其中,历史存储数据关联信息应当包含与数据去重归并相关的信息,可选的,历史存储数据关联信息包括:数据集编号、历史数据标识符、首次采集时间、最近采集时间、发现次数、累计发现天数等信息。其中,数据集编号表示数据所属的数据集编号,历史数据标识符可以是通过去重策略确定的标识,首次采集时间表示第一次采集到数据集编号和历史数据标识符所对应的数据的时间,最近采集时间表示最近一次采集到数据集编号和历史数据标识符所对应的数据的时间,发现次数表示采集到数据集编号和历史数据标识符所对应的数据的总次数,累计发现天数表示采集到数据集编号和历史数据标识符所对应的数据持续的天数。根据上述历史存储数据的关联信息可以确定各个与历史存储数据对应的历史数据标识符。

[0068] S207、判断所述历史数据标识符与所述当前数据标识符是否一致,若是则执行S208,若否则执行S209。

[0069] 将待存储数据存储至目标存储块中时,需要判断所述历史数据标识符与所述当前数据标识符是否一致,即判断待存储数据是否需要在存储时进行去重归并。

[0070] 当历史数据标识符与所述当前数据标识符相一致时,表明待存储数据在目标存储块中是可能是存在的,需要进行进一步的判断,故执行S208;

[0071] 当历史数据标识符与所述当前数据标识符不一致时,表明待存储数据在目标存储

块中是不存在的,故执行S209。

[0072] S208、判断所述当前数据集标识与所述历史数据集标识是否一致,若是则执行S210,若否则执行S209。

[0073] 当历史数据标识符与所述当前数据标识符相一致时,表示当前数据可能已存在于目标存储块中,但是,由于待存储数据与历史存储数据所属的数据集可能不同,因此需要进行下一步的比较分析。

[0074] 进一步的,通过判断所述当前数据集标识与所述历史数据集标识是否一致,来决定是将所述待存储数据直接存储至目标存储块中,还是将所述待存储数据与所述历史存储数据进行去重归并后存储值目标存储块中。

[0075] 当所述当前数据集标识与所述历史数据集标识相一致时,表明目标存储块中存在与待存储数据对应的历史存储数据,故执行S210,将待存储数据存储至目标存储块中;

[0076] 当所述当前数据集标识与所述历史数据集标识不一致时,表明待存储数据在目标存储块中是不存在的,故执行S209,将待存储数据存储至目标存储块中。

[0077] S209、将所述待存储数据删除,并基于所述待存储数据更新与所述历史数据标识符相对应的历史存储数据的关联信息。

[0078] 当目标存储块中存在与待存储数据对应的历史存储数据时,需要对数据进行去重归并,即将待存储数据与对应的历史存储数据进行合并,更新历史存储数据的关联信息,例如:最近采集时间、发现次数、累计发现天数等信息。

[0079] S210、将所述待存储数据按照预设格式存储至所述目标存储块中,并更新与所述待存储数据对应的关联信息。

[0080] 当目标存储块中不存在待存储数据时,需要将待存储数据存储至目标存储块中,并且更新与所述待存储数据对应的关联信息。关联信息中的数据集编号为待存储数据的数据集编号,历史数据标识符为待存储数据的当前数据标识符。并且,需要补充关联信息,例如:首次采集时间、最近采集时间、发现次数、累计发现天数等。

[0081] 优选的,在读取待存储数据时,根据存储块编号进行读取;在读取历史储数据时,根据存储块编号进行读取。进一步的,可以将读取到的待存储数据存储到Treeset中;将读取到的历史存储数据存储到Treeset中。

[0082] 优选的,在Treeset中存储待存储数据时,按当前数据标识符升序存储;在Treeset中存储历史存储数据时,按历史数据标识符升序存储。这种存储方式便于判断历史数据标识符与当前数据标识符之间是否满足预设条件,提升了数据处理效率。

[0083] 优选的,执行任务是可以根据任务设置的开始存储块编号和结束存储块编号进行批量处理,其中开始存储块编号和结束存储块的设定需要根据执行任务的设备自身性能确定,即根据设备性能确定批量处理的存储块数量。

[0084] 本发明实施例提供了一种数据存储方法,通过当前数据标识符与历史数据标识符的比较以及当前数据集标识与所述历史数据集标识的比较,进一步将待存储数据去重归并存储至目标存储块中,避免了存储空间浪费和内存占用过高的情况,实现了数据处理效率的提升。

[0085] 实施例三

[0086] 图3是本发明实施例三提供的一种数据存储装置结构框图。该装置用于执行上述

任意实施例所提供的一种数据存储方法,具备执行方法相应的功能模块和有益效果。该装置包括:标识确定模块310、存储块编号确定模块320、标识获取模块330和数据存储模块340。

[0087] 标识确定模块310,用于获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识;存储块编号确定模块320,用于基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号;标识获取模块330,用于获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符;数据存储模块340,用于当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中。

[0088] 可选的,标识确定模块,还用于针对每一条待存储数据,确定与至少一个去重字段对应的关键信息;采用哈希算法对所述关键信息进行处理,确定与所述待存储数据对应的当前数据标识符;基于所述待存储数据所属数据集编号以及所述当前数据标识符,确定与所述待存储数据对应的目标存储标识。

[0089] 可选的,存储块编号确定模块,还用于根据所述目标存储标识以及预先设置的目标函数,得到第一处理结果值;根据所述第一处理结果值以及预先设置的存储块数量,确定与所述待存储数据对应的目标存储块编号。

[0090] 可选的,标识获取模块,还用于确定与所述目标存储块编号对应的目标存储块;调取所述目标存储块中各个历史存储数据所对应的关联信息,以基于所述关联信息确定与各个历史存储数据对应的历史数据标识符。

[0091] 可选的,数据存储模块,还用于当历史数据标识符与当前数据标识符不一致,则将待存储数据缓存至目标存储块中,并建立与待存储数据相对应的关联信息。

[0092] 可选的,数据存储模块,还用于当所述历史数据标识符与所述当前数据标识符相一致时,则分别获取所述待存储数据所属的当前数据集标识,以及所述历史数据标识符对应的历史存储数据所属的历史数据集标识;根据所述当前数据集标识以及所述历史数据集标识,将所述待存储数据存储至所述目标存储块中;当所述历史数据标识符与所述当前数据标识符不一致时,则将所述待存储数据按照预设格式存储至所述目标存储块中,并更新与所述待存储数据对应的关联信息。

[0093] 可选的,数据存储模块,还用于当所述当前数据集标识与所述历史数据集标识相一致时,则将所述待存储数据删除,并基于所述待存储数据更新与所述历史数据标识符相对应的历史存储数据的关联信息;当所述当前数据集标识与所述历史数据集标识不一致时,则将所述待存储数据按照预设格式存储至所述目标存储块中,并更新与所述待存储数据对应的关联信息。

[0094] 本实施例提供的数据存储装置,通过当前数据标识符与历史数据标识符的比较以及当前数据集标识与所述历史数据集标识的比较,进一步将待存储数据去重归并存储至目标存储块中,避免了存储空间浪费和内存占用过高的情况,实现了数据处理效率的提升。

[0095] 本发明实施例所提供的数据存储装置可执行本发明任意实施例所提供的数据存储方法,具备执行方法相应的功能模块和有益效果。

[0096] 值得注意的是,上述装置所包括的各个单元和模块只是按照功能逻辑进行划分的,但并不局限于上述的划分,只要能够实现相应的功能即可;另外,各功能单元的具体名

称也只是为了便于相互区分,并不用于限制本发明实施例的保护范围。

[0097] 实施例四

[0098] 图4为本发明实施例四提供的一种服务器的结构示意图。图4示出了适于用来实现本发明实施例实施方式的示例性服务器40的框图。图4显示的服务器40仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0099] 如图4所示,服务器40以通用服务器的形式表现。服务器40的组件可以包括但不限于:一个或者多个处理器或者处理单元401,系统存储器402,连接不同系统组件(包括系统存储器402和处理单元401)的总线403。

[0100] 总线403表示几类总线结构中的一种或多种,包括存储器总线或者存储器控制器,外围总线,图形加速端口,处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说,这些体系结构包括但不限于工业标准体系结构(ISA)总线,微通道体系结构(MAC)总线,增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0101] 服务器40典型地包括多种计算机系统可读介质。这些介质可以是任何能够被服务器40访问的可用介质,包括易失性和非易失性介质,可移动的和不可移动的介质。

[0102] 系统存储器402可以包括易失性存储器形式的计算机系统可读介质,例如随机存取存储器(RAM)404和/或高速缓存存储器405。服务器40可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例,存储系统406可以用于读写不可移动的、非易失性磁介质(图4未显示,通常称为“硬盘驱动器”)。尽管图4中未示出,可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器,以及对可移动非易失性光盘(例如CD-ROM,DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下,每个驱动器可以通过一个或者多个数据介质接口与总线403相连。存储器402可以包括至少一个程序产品,该程序产品具有一组(例如至少一个)程序模块,这些程序模块被配置以执行本发明各实施例的功能。

[0103] 具有一组(至少一个)程序模块407的程序/实用工具408,可以存储在例如存储器402中,这样的程序模块407包括但不限于操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块407通常执行本发明所描述的实施例中的功能和/或方法。

[0104] 服务器40也可以与一个或多个外部设备409(例如键盘、指向设备、显示器410等)通信,还可与一个或者多个使得用户能与该服务器40交互的设备通信,和/或与使得该服务器40能与一个或多个其它计算设备进行通信的任何设备(例如网卡,调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口411进行。并且,服务器40还可以通过网络适配器412与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器412通过总线403与服务器40的其它模块通信。应当明白,尽管图4中未示出,可以结合服务器40使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0105] 处理单元401通过运行存储在系统存储器402中的程序,从而执行各种功能应用以及数据处理,例如实现本发明实施例所提供的数据存储方法。

[0106] 实施例五

[0107] 本发明实施例五还提供一种包含计算机可执行指令的存储介质,所述计算机可执行指令在由计算机处理器执行时用于执行实施例所提供的数据存储方法,该方法包括:

[0108] 获取待存储数据,确定与所述待存储数据对应的当前数据标识符以及目标存储标识;

[0109] 基于所述目标存储标识,确定与所述待存储数据对应的目标存储块编号;

[0110] 获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符;

[0111] 当所述历史数据标识符与所述当前数据标识符之间满足预设条件时,将所述待存储数据存储至所述目标存储块中。

[0112] 本发明实施例的计算机存储介质,可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0113] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0114] 计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括——但不限于无线、电线、光缆、RF等等,或者上述的任意合适的组合。

[0115] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明实施例操作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++,还包括常规的过程式程序设计语言——诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0116] 注意,上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解,本发明不限于这里所述的特定实施例,对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此,虽然通过以上实施例对本发明进行了较为详细的说明,但是本发明不仅仅限于以上实施例,在不脱离本发明构思的情况下,还可以包括更多其他等效实施例,而本发明的范围由所附的权利要求范围决定。

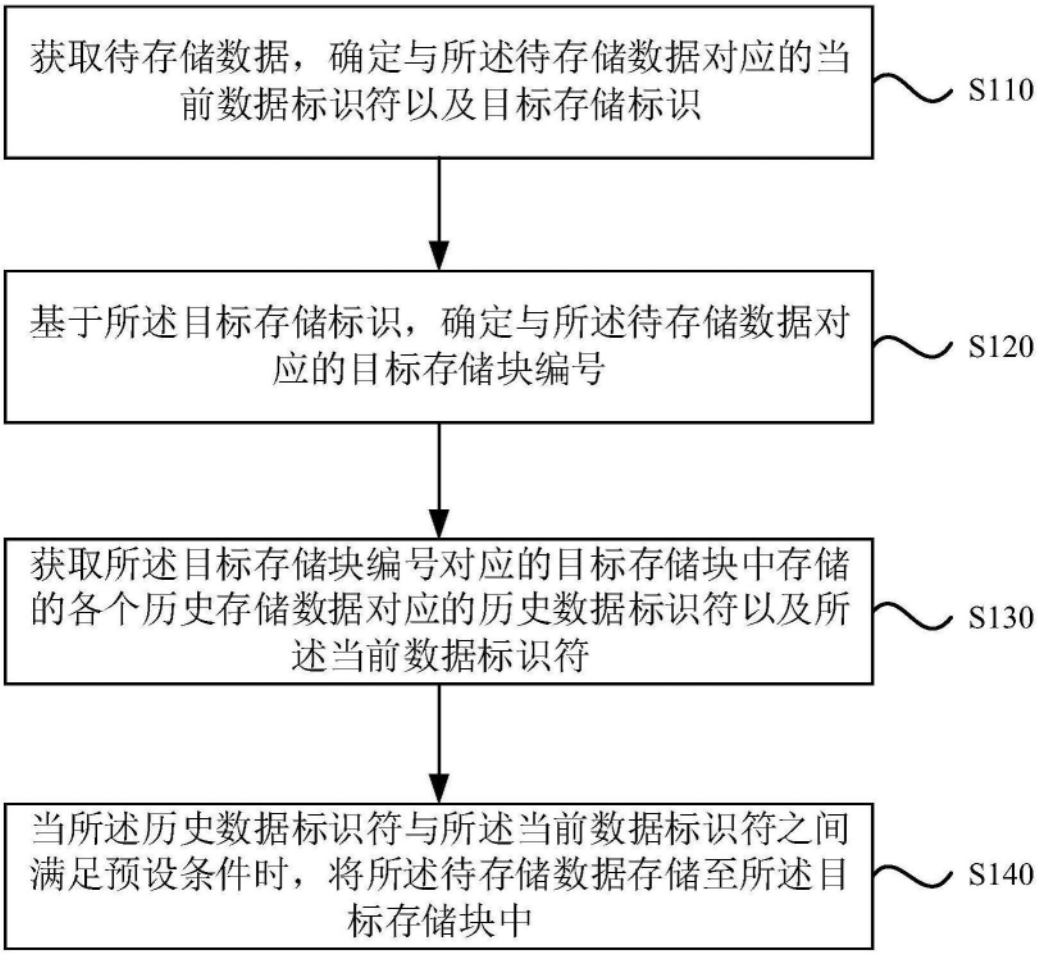

获取待存储数据，确定与所述待存储数据对应的当前数据标识符以及目标存储标识
S110
基于所述目标存储标识，确定与所述待存储数据对应的目标存储块编号
S120
获取所述目标存储块编号对应的目标存储块中存储的各个历史存储数据对应的历史数据标识符以及所述当前数据标识符
S130
当所述历史数据标识符与所述当前数据标识符之间满足预设条件时，将所述待存储数据存储至所述目标存储块中
S140

图1

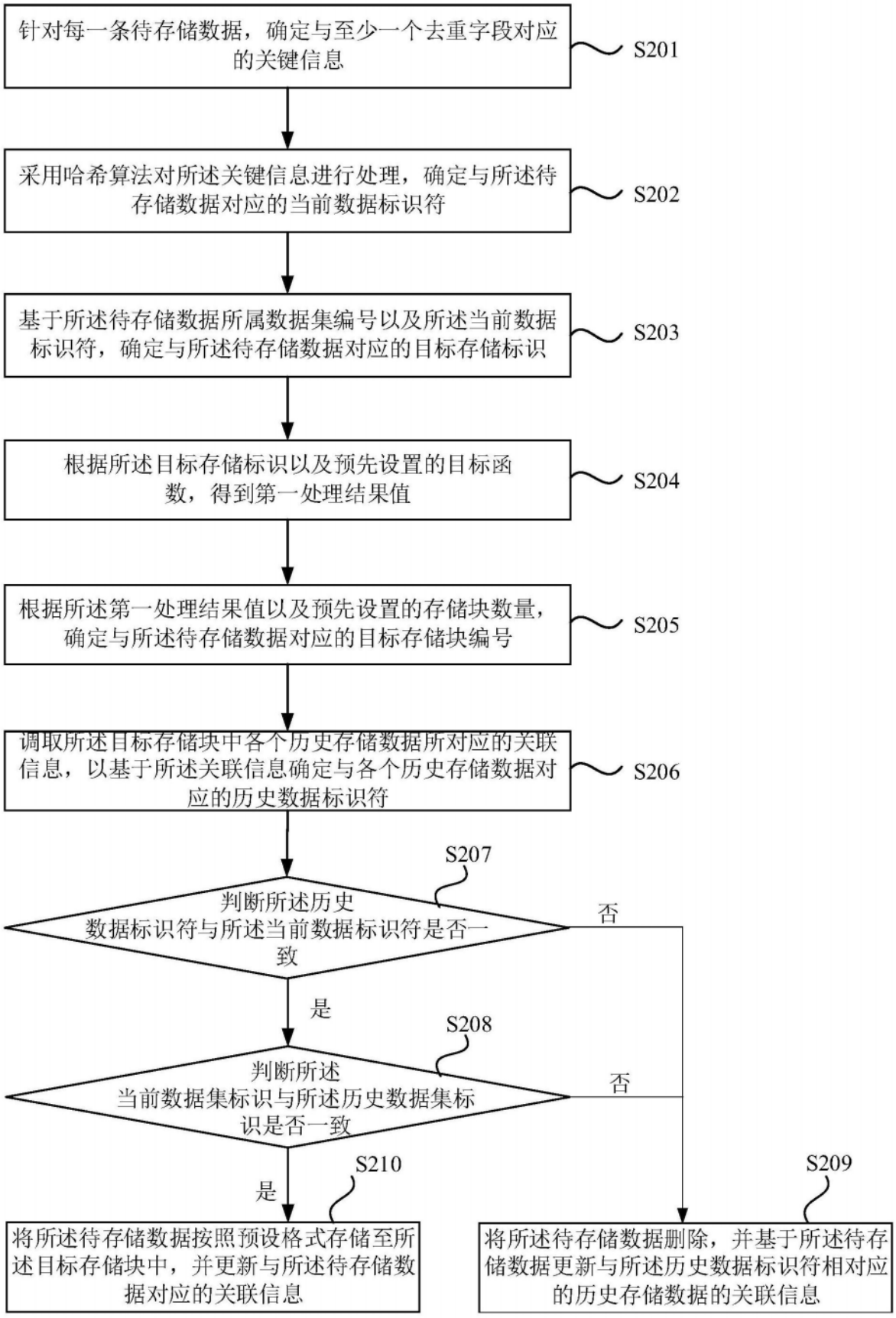

针对每一条待存储数据，确定与至少一个去重字段对应的关键信息
S201
采用哈希算法对所述关键信息进行处理，确定与所述待存储数据对应的当前数据标识符
S202
基于所述待存储数据所属数据集编号以及所述当前数据标识符，确定与所述待存储数据对应的目标存储标识
S203
根据所述目标存储标识以及预先设置的目标函数，得到第一处理结果值
S204
根据所述第一处理结果值以及预先设置的存储块数量，确定与所述待存储数据对应的目标存储块编号
S205
调取所述目标存储块中各个历史存储数据所对应的关联信息，以基于所述关联信息确定与各个历史存储数据对应的历史数据标识符
S206
S207
判断所述历史数据标识符与所述当前数据标识符是否一致
否
是
S208
判断所述当前数据集标识与所述历史数据集标识是否一致
否
是
S210
S209
将所述待存储数据按照预设格式存储至所述目标存储块中，并更新与所述待存储数据对应的关联信息
将所述待存储数据删除，并基于所述待存储数据更新与所述历史数据标识符相对应的历史存储数据的关联信息

图2

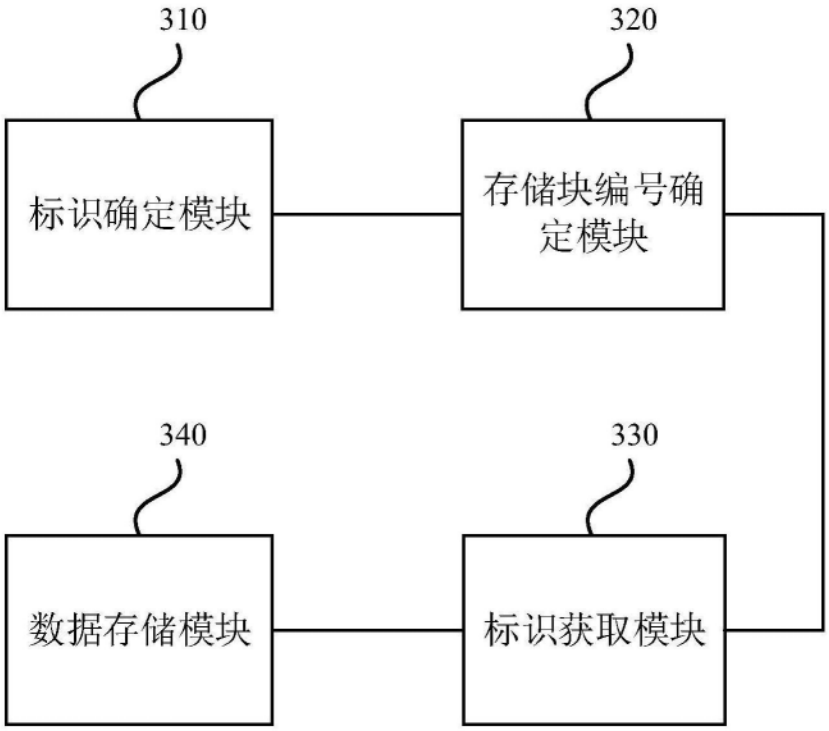


图3

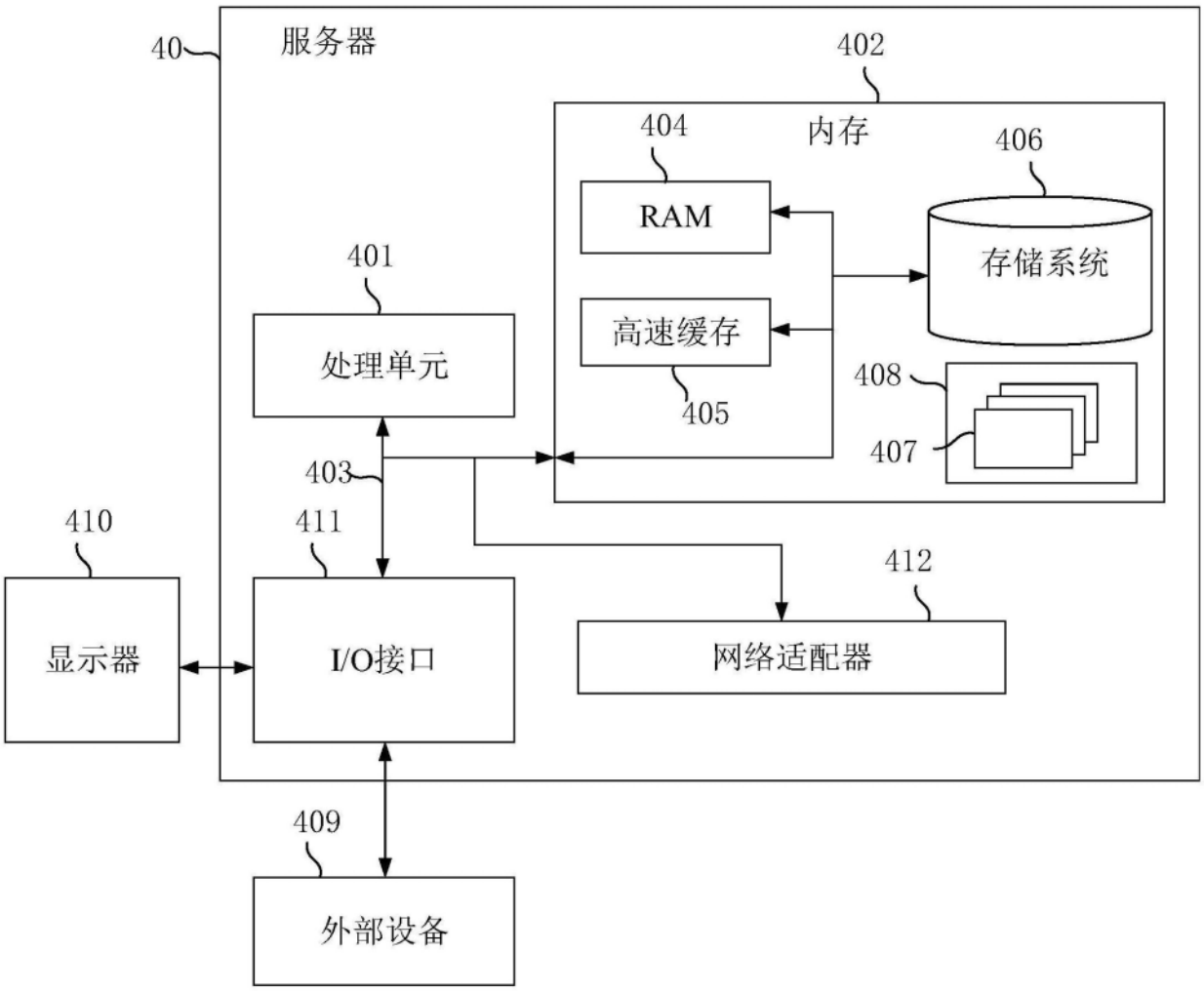

40
服务器
402
内存
404
RAM
406
存储系统
401
处理单元
高速缓存
405
408
407
403
410
411
412
显示器
I/O接口
网络适配器
409
外部设备

图4