



(12)发明专利

(10)授权公告号 CN 103678339 B

(45)授权公告日 2017.05.17

(21)申请号 201210328490.9

(56)对比文件

(22)申请日 2012.09.06

CN 102141907 A, 2011.08.03,

(65)同一申请的已公布的文献号

US 2003037037 A1, 2003.02.20,

申请公布号 CN 103678339 A

CN 1317882 A, 2001.10.17,

(43)申请公布日 2014.03.26

审查员 胡平

(73)专利权人 阿里巴巴集团控股有限公司

地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 李庆丰

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319

代理人 苏培华

(51)Int.Cl.

G06F 17/30(2006.01)

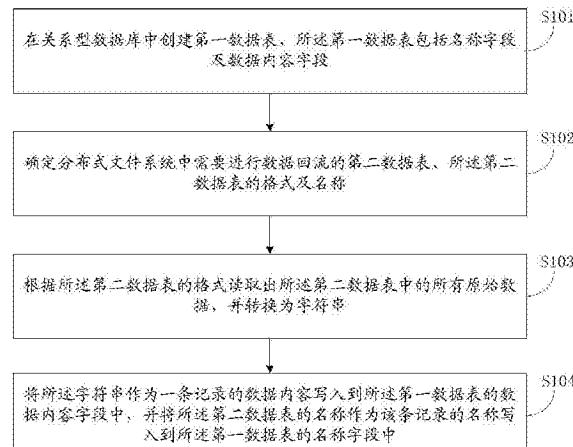
权利要求书3页 说明书12页 附图2页

(54)发明名称

数据回流、关系型数据库中的数据访问方法及系统

(57)摘要

本申请公开了数据回流、关系型数据库中的数据访问方法及系统，所述数据回流方法包括：在关系型数据库中创建第一数据表，所述第一数据表包括名称字段及数据内容字段；确定分布式系统中需要进行数据回流的第二数据表、所述第二数据表的格式及名称；根据所述第二数据表的格式读取出所述第二数据表中的所有原始数据，并转换为字符串，在所述字符串中，根据各个原始数据在所述第二数据表中所处的行与列的不同，利用预置的分隔符进行分隔，所述分隔符包括行分隔符及列分隔符；将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中，并将所述第二数据表的名称作为该条记录的名称写入到所述第一数据表的名称字段中。



1. 一种数据回流方法，包括：

在关系型数据库中创建第一数据表，所述第一数据表包括名称字段及数据内容字段；

确定分布式系统中需要进行数据回流的第二数据表、所述第二数据表的格式及名称；

根据所述第二数据表的格式读取出所述第二数据表中的所有原始数据，并转换为字符串，在所述字符串中，根据各个原始数据在所述第二数据表中所处的行与列的不同，利用预置的分隔符进行分隔，所述分隔符包括行分隔符及列分隔符；

将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中，并将所述第二数据表的名称作为该条记录的名称写入到所述第一数据表的名称字段中。

2. 根据权利要求1所述的方法，所述将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中包括：

将所述字符串按照指定的格式进行压缩后，作为一条记录的数据内容写入到所述第一数据表的数据内容字段中。

3. 根据权利要求1所述的方法，所述将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中包括：

按照指定的输出格式，将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中。

4. 根据权利要求1所述的方法，所述第一数据表还包括日期字段；所述方法还包括：

将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中的同时，将当前日期作为该条记录的日期写入到所述第一数据表的日期字段中。

5. 根据权利要求1所述的方法，还包括：

监控所述第一数据表中记录条数的变化；

当所述记录条数达到预置阈值时，为所述第一数据表添加索引字段，每一条索引对应所述预置阈值条数的记录。

6. 一种关系型数据库中的数据访问方法，所述关系型数据库中保存有第一数据表，所述第一数据表包括名称字段及数据内容字段，所述第一数据表中每条记录的名称字段用于保存第二数据表的名称，数据内容字段用于保存第二数据表中的所有原始数据，所述原始数据在存入所述数据内容字段之前被转换为字符串，在所述字符串中，根据各个原始数据在所述第二数据表中所处的行与列的不同，利用预置的分隔符进行分隔，所述分隔符包括行分隔符及列分隔符；所述方法包括：

接收查询请求，根据所述查询请求与所述第一数据表的名称字段的匹配情况，确定目标记录条目；

提取所述目标记录条目的数据内容字段中的字符串，并按照所述行分隔符及列分隔符对所述字符串进行拆分，还原成与第二数据表对应的二维数组；

确定所述第二数据表中各字段的含义；

按照所述各字段的含义将所述二维数组输出为二维数据表文件并返回。

7. 根据权利要求6所述的方法，所述将所述二维数组输出为二维数据表文件并返回，包括：

提供所述二维数据表文件的访问接口供调用；

或者，

将所述二维数据表文件输出到网页供查看或下载。

8. 根据权利要求6所述的方法,还包括:

确定所述第二数据表中各字段的输出格式;

所述按照所述各字段的含义将所述二维数组输出为二维数据表文件并返回,包括:

按照所述各字段的含义及输出格式将所述二维数组输出为二维数据表文件并返回。

9. 根据权利要求6所述的方法,所述字符串在被存入所述第一数据表的数据内容字段之前被按照指定的格式压缩;所述提取所述目标记录条目的数据内容字段中的字符串包括:

提取所述目标记录条目的数据内容字段中的数据并根据所述指定的格式进行解压得到字符串。

10. 根据权利要求6所述的方法,将所述字符串被按照指定的输出格式写入到所述第一数据表的数据内容字段中,所述提取所述目标记录条目的数据内容字段中的字符串包括:

将所述目标记录条目的数据内容字段中的字符串按照所述输出格式输出为标准文件。

11. 根据权利要求6所述的方法,所述第一数据表还包括日期字段,所述日期字段用于保存将所述第二数据表回流到所述第一数据表时的日期;所述根据所述查询请求与所述第一数据表的名称字段的匹配情况,确定目标记录条目包括:

根据所述查询请求与所述第一数据表的名称字段及日期字段的匹配情况,确定目标记录条目。

12. 一种数据回流系统,包括:

创建单元,用于在关系型数据库中创建第一数据表,所述第一数据表包括名称字段及数据内容字段;

信息获取单元,用于确定分布式系统中需要进行数据回流的第二数据表、所述第二数据表的格式及名称;

数据转换单元,用于根据所述第二数据表的格式读取出所述第二数据表中的所有原始数据,并转换为字符串,在所述字符串中,根据各个原始数据在所述第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,所述分隔符包括行分隔符及列分隔符;

数据写入单元,用于将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中,并将所述第二数据表的名称作为该条记录的名称写入到所述第一数据表的名称字段中。

13. 一种关系型数据库中的数据访问系统,所述关系型数据库中保存有第一数据表,所述第一数据表包括名称字段及数据内容字段,所述第一数据表中每条记录的名称字段用于保存第二数据表的名称,数据内容字段用于保存第二数据表中的所有原始数据,所述原始数据在存入所述数据内容字段之前被转换为字符串,在所述字符串中,根据各个原始数据在所述第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,所述分隔符包括行分隔符及列分隔符;所述系统包括:

目标记录条目确定单元,用于接收查询请求,根据所述查询请求与所述第一数据表的名称字段的匹配情况,确定目标记录条目;

拆分单元,用于提取所述目标记录条目的数据内容字段中的字符串,并按照所述行分隔符及列分隔符对所述字符串进行拆分,还原成与第二数据表对应的二维数组;

字段含义确定单元，用于确定所述第二数据表中各字段的含义；
返回单元，用于按照所述各字段的含义将所述二维数组输出为二维数据表文件并返回。

数据回流、关系型数据库中的数据访问方法及系统

技术领域

[0001] 本申请涉及数据处理技术领域,特别是涉及数据回流、关系型数据库中的数据访问方法及系统。

背景技术

[0002] 互联网行业产生的数据量非常大,其运算的量一般需要在Hadoop等大型的分布式系统中才能完成,例如,相关的日志数据、浏览数据、用户数据、交易数据、商品数据等等全部会通过Hadoop完成相关计算。

[0003] Hadoop充分利用集群的威力高速运算和存储,因此,对大数据量的运算非常有优势。但是,由于Hadoop处理后的可用数据往往会被分散存放在不同的服务器上,并且一般只提供命令行的方式进行读取,在用户访问和数据获取方面不是很友好。因此,一般会将Hadoop上对大数据进行处理之后的可用数据进行回流,在回流到关系型数据库之后,可以方便的做成各种程序接口(API)供调用,然后可视化的方式提供给访问者。

[0004] 传统的数据回流方法中,每当在Hadoop上产生一个新的数据报表,都会在对应的关系型数据库中建立同样表结构(表字段数目及含义完全一致)的表,然后通过程序将Hadoop上的数据读出并写入关系型数据库的表中,从而达到回流的目的。

[0005] 但是,在这种传统的方式中,由于每产生一张Hadoop的表,都要在关系型数据库中建立同样的表,每次都需要走数据库的建表流程,相对繁琐和冗长,并且关系型数据库中每产生一个新的表,都需要编写相应的代码以便访问表中的数据,工作量比较大。

发明内容

[0006] 本申请提供了数据回流方法及系统,能够简化数据回流的流程。本申请还提供了关系型数据库中的数据访问方法及系统。

[0007] 本申请提供了如下方案:

[0008] 一种数据回流方法,包括:

[0009] 在关系型数据库中创建第一数据表,所述第一数据表包括名称字段及数据内容字段;

[0010] 确定分布式系统中需要进行数据回流的第二数据表、所述第二数据表的格式及名称;

[0011] 根据所述第二数据表的格式读取出所述第二数据表中的所有原始数据,并转换为字符串,在所述字符串中,根据各个原始数据在所述第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,所述分隔符包括行分隔符及列分隔符;

[0012] 将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中,并将所述第二数据表的名称作为该条记录的名称写入到所述第一数据表的名称字段中。

[0013] 可选地,所述将所述字符串作为一条记录的数据内容写入到所述第一数据表的数

据内容字段中包括：

[0014] 将所述字符串按照指定的格式进行压缩后，作为一条记录的数据内容写入到所述第一数据表的数据内容字段中。

[0015] 可选地，所述将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中包括：

[0016] 按照指定的输出格式，将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中。

[0017] 可选地，所述第一数据表还包括日期字段；所述方法还包括：

[0018] 将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中的同时，将当前日期作为该条记录的日期写入到所述第一数据表的日期字段中。

[0019] 可选地，还包括：

[0020] 监控所述第一数据表中记录条数的变化；

[0021] 当所述记录条数达到预置阈值时，为所述第一数据表添加索引字段，每一条索引对应所述预置阈值条数的记录。

[0022] 一种关系型数据库中的数据访问方法，所述关系型数据库中保存有第一数据表，所述第一数据表包括名称字段及数据内容字段，所述第一数据表中每条记录的名称字段用于保存第二数据表的名称，数据内容字段用于保存第二数据表中的所有原始数据，所述原始数据在存入所述数据内容字段之前被转换为字符串，在所述字符串中，根据各个原始数据在所述第二数据表中所处的行与列的不同，利用预置的分隔符进行分隔，所述分隔符包括行分隔符及列分隔符；所述方法包括：

[0023] 接收查询请求，根据所述查询请求与所述第一数据表的名称字段的匹配情况，确定目标记录条目；

[0024] 提取所述目标记录条目的数据内容字段中的字符串，并按照所述行分隔符及列分隔符对所述字符串进行拆分，还原成与第二数据表对应的二维数组；

[0025] 确定所述第二数据表中各字段的含义；

[0026] 按照所述各字段的含义将所述二维数组输出为二维数据表文件并返回。

[0027] 可选地，所述返回给所述访问者包括：

[0028] 提供所述二维数据表文件的访问接口供调用；

[0029] 或者，

[0030] 将所述二维数据表文件输出到网页供查看或下载。

[0031] 可选地，还包括：

[0032] 确定所述第二数据表中各字段的输出格式；

[0033] 所述按照所述各字段的含义将所述二维数组输出为二维数据表文件，返回给所述访问者包括：

[0034] 按照所述各字段的含义及输出格式将所述二维数组输出为二维数据表文件并返回。

[0035] 可选地，所述字符串在被存入所述第一数据表的数据内容字段之前被按照指定的格式压缩；所述提取所述目标记录条目的数据内容字段中的字符串包括：

[0036] 提取所述目标记录条目的数据内容字段中的数据并根据所述指定的格式进行解

压得到字符串。

[0037] 可选地,将所述字符串被按照指定的输出格式写入到所述第一数据表的数据内容字段中,所述提取所述目标记录条目的数据内容字段中的字符串包括:

[0038] 将所述目标记录条目的数据内容字段中的字符串按照所述输出格式输出为标准文件。

[0039] 可选地,所述第一数据表还包括日期字段,所述日期字段用于保存将所述第二数据表回流到所述第一数据表时的日期;所述根据所述查询请求与所述第一数据表的名称字段的匹配情况,确定目标记录条目包括:

[0040] 根据所述查询请求与所述第一数据表的名称字段及日期字段的匹配情况,确定目标记录条目。

[0041] 一种数据回流系统,包括:

[0042] 创建单元,用于在关系型数据库中创建第一数据表,所述第一数据表包括名称字段及数据内容字段;

[0043] 信息获取单元,用于确定分布式系统中需要进行数据回流的第二数据表、所述第二数据表的格式及名称;

[0044] 数据转换单元,用于根据所述第二数据表的格式读取出所述第二数据表中的所有原始数据,并转换为字符串,在所述字符串中,根据各个原始数据在所述第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,所述分隔符包括行分隔符及列分隔符;

[0045] 数据写入单元,用于将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中,并将所述第二数据表的名称作为该条记录的名称写入到所述第一数据表的名称字段中。

[0046] 一种关系型数据库中的数据访问系统,所述关系型数据库中保存有第一数据表,所述第一数据表包括名称字段及数据内容字段,所述第一数据表中每条记录的名称字段用于保存第二数据表的名称,数据内容字段用于保存第二数据表中的所有原始数据,所述原始数据在存入所述数据内容字段之前被转换为字符串,在所述字符串中,根据各个原始数据在所述第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,所述分隔符包括行分隔符及列分隔符;所述系统包括:

[0047] 目标记录条目确定单元,用于接收查询请求,根据所述查询请求与所述第一数据表的名称字段的匹配情况,确定目标记录条目;

[0048] 拆分单元,用于提取所述目标记录条目的数据内容字段中的字符串,并按照所述行分隔符及列分隔符对所述字符串进行拆分,还原成与第二数据表对应的二维数组;

[0049] 字段含义确定单元,用于确定所述第二数据表中各字段的含义;

[0050] 返回单元,用于按照所述各字段的含义将所述二维数组输出为二维数据表文件并返回。

[0051] 根据本申请提供的具体实施例,本申请公开了以下技术效果:

[0052] 通过本申请提供的数据回流方法,只需要在关系型数据库中建立一张表,可以快速的输入分布式系统(如Hadoop集群等)上的任何数据表,而不需要每次在Hadoop上产生的数据报表都在关系型数据库上建立同样的表单,节省了存储空间,减少了中间环节。

[0053] 另外,在实现过程中,支持各种个性化的配置,可以根据Hadoop上的数据报表的不

同来配置各种不同的输入输出规则,具有很强的灵活性。

[0054] 通过本申请提供的数据访问方法,可以针对特殊结构的一个数据表形成统一的输出接口,而不需要针对数据库中的多个数据表都编写数据访问代码,简化了实现的流程。

[0055] 当然,实施本申请的任一产品并不一定需要同时达到以上所述的所有优点。

附图说明

[0056] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例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申请的一些实施例,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。

[0057] 图1是本申请实施例提供的数据回流方法的流程图;

[0058] 图2是本申请实施例提供的数据访问方法的流程图;

[0059] 图3是本申请实施例提供的数据回流系统的示意图;

[0060] 图4是本申请实施例提供的数据访问系统的示意图。

具体实施方式

[0061] 下面将结合本申请实施例中的附图,对本申请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申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员所获得的所有其他实施例,都属于本申请保护的范围。

[0062] 首先,本申请实施例提供了一种数据回流方法,参见图1,该方法可以包括以下步骤:

[0063] S101:在关系型数据库中创建第一数据表,所述第一数据表包括名称字段及数据内容字段;

[0064] 关系型数据库可以是MySql、Oracle、DB2等。在本申请实施例中,在关系型数据库中创建第一数据表的操作是在进行具体的数据回流之前进行的。也就是说,与现有技术不同,在数据库中创建数据表时,不需要考虑Hadoop上的数据表的结构、字段含义等,而是直接按照预先定义好的结构进行创建即可,并且,不需要针对每个需要回流的数据表都在关系型数据库中重新创建与之对应的数据表,本申请实施例在关系型数据库中创建第一数据表的操作只需进行一次即可,Hadoop上产生的所有第二数据表中的数据都可以统一回流到这一张数据库表中。

[0065] 具体在创建该第一数据表时,可以包括名称(name)字段及数据内容(content)字段,其中,名称字段用于保存Hadoop中每个第二数据表的名称,数据内容字段用于保存Hadoop中每个第二数据表的全部数据。当然,在实际应用中,该第一数据表中还可以有其他的字段,后续会有相应的介绍。

[0066] S102:确定分布式系统中需要进行数据回流的第二数据表、所述第二数据表的格式及名称;

[0067] 在关系型数据库中创建了第一数据表之后,就可以针对Hadoop上具体的第二数据表进行数据回流操作。在实际应用中,这里的第二数据表可以是通过编写Map/Reduce程序

或者HIVE SQL脚本在Hadoop集群上运行(主要进行数据的清洗、运算、统计等,比如说要查看昨天访问某网站的用户有多少、成交额是多少等等,需要用交易表、用户表、日志表等进行汇总、过滤、计算)产生的数据表(二维表),也可以是Hadoop集群上已经生成好的各种格式的数据表。在实际应用中,Hadoop上的第二数据表一般是很多个,需要对哪个或者哪些具体的第二数据表进行数据回流需要指定。具体实现时,可以提供一配置界面,由配置人员在配置界面上填写需要回流的第二数据表的名称,这样就可以确定出需要对哪个第二数据表中的数据进行数据回流。

[0068] 在确定了需要进行数据回流的第二数据表之后,由于后续的步骤中需要从第二数据表中进行数据读取,而不同的表格式对应着第二数据表中不同的文件内容格式(例如,有的是用“\n”对文件内容进行分隔,还有的是用“,”,还有的是用空格等等),只有在获知了一个数据表的表格式,才能正确地从读取出数据,因此,从第二数据表中读取数据之前需要获知第二数据表的表格式。具体的,如果Hadoop上的所有第二数据表都采用相同的表格式,则关于第二数据表的表格式信息可以是预先获知的,统一按照该表格式从第二数据表中进行数据读取即可。但实际应用中,Hadoop上的各个第二数据表可能会具有不同的表格式,具体是何种表格式,配置人员是可以知晓的,因此,还可以在前述配置界面中提供第二数据表的表格式配置入口,由配置人员在输入第二数据表的名称的同时,输入第二数据表的表格式,这样就可以获知第二数据表的表格式,然后按照该表格式从第二数据表中提取数据即可。

[0069] 当然,在前述配置界面中,还可以提供其他的配置入口。例如,输出格式配置入口,如果配置人员需要指定第二数据表中的数据输出到第二数据表时采用的输出格式,则可以通过该入口进行配置,例如,配置为json格式、文本格式、xml格式等等;当然如果配置人员没有指定输出格式,则可以采用默认的输出格式进行输出,例如,json格式。又如,还可以包括压缩方式配置入口,为了节省存储空间,在将第二数据表中的数据保存到第一数据表的数据内容字段之前,还可以进行压缩,配置人员可以通过该入口指定具体的压缩格式,例如zip等。当然,如果配置人员没有指定,则可以不进行压缩,或者按照默认的格式进行压缩,等等。

[0070] S103:根据所述第二数据表的格式读取出所述第二数据表中的所有原始数据,并转换为字符串,在所述字符串中,根据各个原始数据在所述第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,所述分隔符包括行分隔符及列分隔符;

[0071] 由于步骤S102中已经获知了第二数据表的表格式,因此,就可以按照该表格式从该第二数据表中一次性读取出所有的原始数据。然后,在本申请实施例中,可以将这些原始数据转换成一个大的字符串。由于第二数据表中也存在行、列的概念,原始数据分布在第二数据表的各行各列中,因此,在转换的过程中,可以采用逐条逐字段地追加写入的方式,同时,根据各个原始数据在第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,这里的分隔符包括行分隔符及列分隔符。例如,列分隔符采用“,”,行分隔符用“[]”,如,某第二数据表中的原始数据如表1所示:

[0072] 表1

[0073]

15	2333	123	56457444.12	12323
16	22	12	123123.14	12

18	5555	444	231932423.22	343254
----	------	-----	--------------	--------

[0074] 则转换之后得到的字符串可以为: [15,2333,123,56457444.12,12323], [16,22,12,123123.14,12], [18,5555,444,231932423.22,343254]。

[0075] 另一个第二数据表中的原始数据假设如表2所示:

[0076] 表2

[0077]

20120427	22.23	LIST	120.11
20120427	20.11	SEARCH	130.22

[0078] 则转换之后得到的字符串可以为:

[0079] [20120427,22.23,LIST,120.11], [20120427,20.11,SEARCH,130.22]。

[0080] 当然,如果需要进行压缩,则在转换得到上述字符串之后,还可以根据一定的压缩格式进行数据压缩,如前文所述,该压缩格式可以是默认的某种格式,也可以是由配置人员指定的某种格式。

[0081] S104:将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中,并将所述第二数据表的名称作为该条记录的名称写入到所述第一数据表的名称字段中。

[0082] 在将一个第二数据表中的所有原始数据转换为一个字符串之后,就可以将该字符串作为一条记录的数据内容写入到第一数据表的数据内容字段中,同时,可以将该第二数据表的名称作为该条记录的名称写入到第一数据表的名称字段中。也就是说,一个第二数据表中的原始数据,回流到关系型数据库中之后,会成为第一数据表中的一条记录中的一个字段,而不是单独的一张数据表。例如,对于前述表1及表2中所示的两个第二数据表,分别回流到第一数据表之后,可以如表3所示:

[0083] 表3

[0084]

Name	Content
QUERY1	[15, 2333, 123, 56457444.12, 12323], [16, 22, 12, 123123.14, 12], [18, 5555, 444, 231932423.22, 343254]
SEARCH	[20120427, 22.23, LIST, 120.11], [20120427, 20.11, SEARCH, 130.22]

[0085] 其中,“name”字段中的具体值可以是由配置人员在前述配置界面中输入的。从表3可以看出,虽然表1与表2是Hadoop中的两个不同的数据表,但是回流到关系型数据库之后,却成为一个大表中的两条记录,并且,关系型数据库中的第一数据表的结构、字段含义等完全与表1、表2的结构及字段含义无关。

[0086] 在实际应用中,经常会出现以下情况:对应Hadoop上的同一数据表而言,可能会经常发生数据内容的更新,数据回流的操作也一般是按照一定的时间间隔周期性地进行,例如,每天回流一次,或者每周回流一次,等等。因此,回流到第一数据表中不同记录条目的数据可能会对应同一个第二数据表,但是访问者在访问的时候,可能会访问具体到某一天的

数据,因此,在本申请实施例中,还可以在第一数据表中增加一日期字段(date),用于保存某条记录产生的日期,也就是将某第二数据表回流到第一数据表的日期。具体实现时,具体日期取值可以是在回流时根据当前的系统时间获取。也即,在将字符串作为一条记录的数据内容写入到第一数据表的数据内容字段中的同时,可以将当前日期作为该条记录的日期写入到第一数据表的日期字段中。这样,在访问者想要访问某第二数据表在某一天的数据内容时,就可以在输入查询请求时,同时输入该第二数据包的名称及回流日期,这样就可以得到精确的访问结果。例如,表1中的数据是2012年4月26日回流到第一数据表中的,而表2中的数据是2012年4月26日回流到第一数据表中的,则第一数据表如表4所示:

[0087] 表4

[0088]

Name	Content	Date
QUERY1	[15, 2333, 123, 56457444.12, 12323], [16, 22, 12, 123123.14, 12], [18, 5555, 444, 231932423.22, 343254]	20120426
SEARCH	[20120427, 22. 23, LIST, 120. 11], [20120427, 20. 11 , SEARCH, 130. 22]	20120428

[0089] 另外,由于本申请实施例中,将众多第二数据表中的数据内容都回流到同一个第一数据表中,随着回流的第二数据表数目的增加,第一数据表中的记录条目也在不断地增加,假设有几十万个第二数据表的数据都回流到该第一数据表中,则该第一数据表中就会有几十万条记录,这样,第一数据表中每个字段中需要保存的数据量就会非常大,尤其是数据内容字段。但是,一个数据表中一个字段的容量一般是有限制的,例如不能超过30M,因此,当数据非常多时,可能无法存放在同一个字段里面。因此,在本申请实施例中,优选地,还可以监控第一数据表中记录条数的变化,当记录条数达到某预置阈值时,就可以为第一数据表添加索引字段(index),每一条索引对应该预置阈值条数的记录。也就是说,通过添加索引字段的方式,可以实现一种自动分隔,例如,每一条记录做成一个index,等等。

[0090] 以上所述介绍了本申请实施例提供的数据回流方法,而数据回流的目的就在于方便访问者的访问及使用,因此,本申请实施例还提供了相应的关系型数据库中的数据访问方法。在该方法中,关系型数据库中的数据表可以是在前文所述的数据回流方法中产生的,与现有技术中回流会产生的数据表有所不同,关系型数据库中的这种第一数据表包括名称字段及数据内容字段,第一数据表中每条记录的名称字段用于保存第二数据表的名称,数据内容字段用于保存第二数据表中的所有原始数据,这些原始数据在存入数据内容字段之前会被转换为字符串,在这种字符串中,根据各个原始数据在第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,这种分隔符包括行分隔符及列分隔符。当然,只要关系型数据库中的数据表具有上述这些特点,都可以使用本申请实施例提供的以下数据访问方法。参见图2,该数据访问方法可以包括以下步骤:

[0091] S201:接收查询请求,根据所述查询请求与所述第一数据表的名称字段的匹配情况,确定目标记录条目;

[0092] 当一个访问者需要查看、下载或者调用某个二维数据表中的数据时,就可以向关

系型数据库发起查询请求。在发起查询请求时,可以携带需要查询的第二数据表的名称,这样,在接收到查询请求之后,就可以将查询请求中携带的名称与第一数据表中name字段中的各个名称进行匹配,匹配成功之后,就可以将对应的记录条目确定为目标记录条目。例如,如果某查询请求中携带的是“SEARCH”,则表3中的第二条记录就与查询请求匹配成功,该第二条记录就是目标记录条目。

[0093] 当然,如果第一数据表中还包括日期字段,并且访问者需要查询某指定日期某第二数据表的数据,则在查询请求中就可以携带名称以及日期这两方面的信息,接收到查询请求之后,需要与第一数据表中的名称字段以及日期字段同时进行匹配,只有当某条记录同时满足这两个条件时,才匹配成功。例如,需要查询2012年4月28日“SEARCH”中的数据,则表4中的第二条记录就是相匹配的目标记录条目。

[0094] S202:提取所述目标记录条目的数据内容字段中的字符串,并按照所述行分隔符及列分隔符对所述字符串进行拆分,还原成与第二数据表对应的二维数组;

[0095] 在查找到目标记录条目之后,就可以从该记录的数据内容字段中提取出其中的字符串,然后按照回流过程中使用的行分隔符及列分隔符,再将字符串进行拆分,还原成与第二数据表对应的二维数组。例如,表4中的第二条记录是与查询请求相匹配的目标记录,则就可以将该条记录中的数据内容字段[20120427,22.23,LIST,120.11],[20120427,20.11,SEARCH,130.22]提取出来,然后,由于已知行分隔符是“[]”,列分隔符是“,”,因此,进行拆分后就可以还原成二维数组状态,显然,还原出的二维数组实际上对应着一个第二数据表。

[0096] 当然,如果在数据回流的过程中,在转换成字符串之后,还进行了数据压缩,则在进行拆分之前,还需要先将提取出的数据进行解压,之后才能得到原始字符串。如果回流时的数据压缩格式是配置人员指定的某种压缩格式,则解压时也需要按照与之对应的解压方式进行解压。

[0097] 另外,如果在进行数据回流时,字符串是按照配置人员指定的某种输出格式写入到第一数据表的数据内容字段中的,则在提取数据内容字段的数据时,也可以按照该输出格式,将字符串输出成某种标准文件,之后再在该标准文件中对字符串进行拆分操作。如果配置人员没有指定输出格式,则也可以按照默认的输出格式(例如json)将字符串输出成标准文件,之后再在该标准文件中对字符串进行拆分操作。

[0098] S203:确定所述第二数据表中各字段的含义;

[0099] 需要说明的是,一般而言,一个二维数据表的每一列代表一个字段,如果某数据表是提供给用户或者其他人员查看的,则数据表中会包含每个字段的名称,例如,表3及表4中的第一行,都是字段名称,包括name、content、date等,这一行并不是数据表中的具体数据,而是用于指明每一列数据的含义。例如,通过表4中的第一列第一行中的“name”,便可知该第一列的具体数值都代表名称,等等。但是,如表1及表2所示,这两个表格中并不存在字段名称这一行,也就是说,Hadoop上的第二数据表中并不包括各个字段的含义信息,其中的原始数据都只是每一条记录的具体数值,换言之,虽然同一字段的原始数据具有某种相同的含义,但是,从Hadoop的第二数据表中无法直接体现出来,进而,写入到第一数据表数据内容字段的字符串中也无法体现出来。而如果直接将不带有字段含义的信息返回给访问者,显然是不够友好的,会使得访问者只看到一些具体的数据,而不知道每一列数据代表的含义是什么。因此,在本申请实施例中,为了将第二数据表各字段的含义提供给访问者,还可

以确定出第二数据表各字段的含义。具体实现时,Hadoop中一般会包含对第二数据表的说明,前文所述的配置人员一般可以根据这种说明知晓第二数据表各个字段的含义;因此,可以为配置人员提供配置界面,配置人员可以通过该配置界面输入第二数据表中各个字段的含义,这样,就可以确定出第二数据表中各字段的含义。例如,表2中第一个字段的含义是日期,第二个字段的含义是CTR,第三个字段的含义是搜索类型,第四个字段的含义是客单价,等等。需要说明的是,该确定字段含义的步骤可以是在接收到具体的查询请求之后,针对查询的具体第二数据表进行字段含义的确定,或者,在另一种方式下,也可以是在数据回流操作完成之后,就分别确定出每个第二数据表中各个字段的含义,在收到查询请求之后,就可以直接根据之前已经确定出的字段含义返回结果即可,这样可以提高响应速度。

[0100] 另外,由于在将数据输出到一个二维数据表时,根据某个字段的数据输出格式的不同,显示到二维数据表中的样式可能会有所不同。例如,如果某字段的输出格式是日期,则该字段所在列的具体数值会自动以右对齐方式的方式显示,如果某字段的输出格式是金额,则该字段所在列的具体数值会自动精确到小数点后两位,并且整数部分从个位开始往前数,每三位之间会自动增加一个逗号,等等。因此,为了使得最终返回的二维数据表中的数据更加规范,还可以由配置人员在配置界面上配置第二数据表中各个字段的输出格式,包括文本、日期、数字等,这样,最终在输出成二维数据表时,就可以按照该配置好的输出格式将具体的数值输出到各个字段中。

[0101] S204:按照所述各字段的含义将所述二维数组输出为二维数据表文件并返回。

[0102] 在将字符串还原成二维数据之后,就可以按照之前已经确定出的对应第二数据表中各个字段的含义,将二维数据输出为二维数据表文件,该文件中不仅包含每行每列的具体数值,还包含有每一字段的含义,也即每一列的名称。例如,查询的是表2中的具体数据,则最终输出的二维数据表文件可以如表5所示:

[0103] 表5

[0104]

日期	CTR	搜索类型	客单价
20120427	22.23	LIST	120.11
20120427	20.11	SEARCH	130.22

[0105] 将该表5返回给访问者之后,访问者便可以很直观地了解该表中的具体数据内容及其含义。

[0106] 具体实现时,在将最终输出的二维数据表文件返回给访问者时,可以有以下实现方式:其中一种是,为访问者提供访问接口,使得外部系统可以通过该接口调用该二维数据表中的数据。另一种可以是,直接将该二维数据表文件输出到网页中,供访问者查看或者下载。其中,为访问者提供访问接口时,可以是提供一些具体的API(应用程序编程接口),具体的方法可以参见已有技术中的实现,这里不再赘述。

[0107] 至此,就完成了数据访问过程,在此过程中,由于并不是每个第二数据表都分别对应一个数据库中的表,因此,不需要针对关系型数据库中的多个表分别编写数据访问代码,以支持对数据的访问,使得整个流程得到简化。

[0108] 总之,在本申请实施例中,只需要在关系型数据库中建立一张表,而不需要每次在Hadoop上产生的数据报表都在关系型数据库上建立同样的表单,节省了存储空间,并且解

放了数据库管理人员和开发人员,减少了中间环节。并且,可以快速的输入Hadoop集群上的任何数据表,并且形成输出接口。另外,在实现过程中,支持各种个性化的配置,可以根据Hadoop上的数据报表的不同来配置各种不同的输入输出规则,具有很强的灵活性。

[0109] 与本申请实施例提供的数据回流方法相对应,本申请实施例还提供了一种数据回流系统,参见图3,该系统可以包括:

[0110] 创建单元301,用于在关系型数据库中创建第一数据表,所述第一数据表包括名称字段及数据内容字段;

[0111] 信息获取单元302,用于确定分布式系统中需要进行数据回流的第二数据表、所述第二数据表的格式及名称;

[0112] 数据转换单元303,用于根据所述第二数据表的格式读取出所述第二数据表中的所有原始数据,并转换为字符串,在所述字符串中,根据各个原始数据在所述第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,所述分隔符包括行分隔符及列分隔符;

[0113] 数据写入单元304,用于将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中,并将所述第二数据表的名称作为该条记录的名称写入到所述第一数据表的名称字段中。

[0114] 具体实现时,所述数据写入单元304可以包括:

[0115] 第一写入子单元,用于将所述字符串按照指定的格式进行压缩后,作为一条记录的数据内容写入到所述第一数据表的数据内容字段中。

[0116] 所述数据写入单元304也可以包括:

[0117] 第二写入子单元,用于按照指定的输出格式,将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中。

[0118] 由于有些数据报表需要周期性地回流到关系型数据库中,因此,所述第一数据表还可以包括日期字段;此时,所述系统还可以包括:

[0119] 日期字段写入单元,用于将所述字符串作为一条记录的数据内容写入到所述第一数据表的数据内容字段中的同时,将当前日期作为该条记录的日期写入到所述第一数据表的日期字段中。

[0120] 另外,该系统还可以包括:

[0121] 监控单元,用于监控所述第一数据表中记录条数的变化;

[0122] 索引字段添加单元,用于当所述记录条数达到预置阈值时,为所述第一数据表添加索引字段,每一条索引对应所述预置阈值条数的记录。

[0123] 与本申请实施例提供的关系型数据库中的数据访问方法相对应,本申请实施例还提供了一种关系型数据库中的数据访问系统,其中,所述关系型数据库中保存有第一数据表,所述第一数据表包括名称字段及数据内容字段,所述第一数据表中每条记录的名称字段用于保存第二数据表的名称,数据内容字段用于保存第二数据表中的所有原始数据,所述原始数据在存入所述数据内容字段之前被转换为字符串,在所述字符串中,根据各个原始数据在所述第二数据表中所处的行与列的不同,利用预置的分隔符进行分隔,所述分隔符包括行分隔符及列分隔符;参见图4,所述系统可以包括:

[0124] 目标记录条目确定单元401,用于接收查询请求,根据所述查询请求与所述第一数据表的名称字段的匹配情况,确定目标记录条目;

[0125] 拆分单元402,用于提取所述目标记录条目的数据内容字段中的字符串,并按照所述行分隔符及列分隔符对所述字符串进行拆分,还原成与第二数据表对应的二维数组;

[0126] 字段含义确定单元403,用于确定所述第二数据表中各字段的含义;

[0127] 返回单元404,用于按照所述各字段的含义将所述二维数组输出为二维数据表文件并返回。

[0128] 具体实现时,所述返回单元404可以包括:

[0129] 接口提供子单元,用于提供所述二维数据表文件的访问接口供调用;

[0130] 或者,

[0131] 网页输出子单元,用于将所述二维数据表文件输出到网页供查看或下载。

[0132] 为了使得输出的二维数据表的格式更规范,该系统还可以包括:

[0133] 字段格式确定单元,用于确定所述第二数据表中各字段的输出格式;

[0134] 所述返回单元404具体用于:

[0135] 按照所述各字段的含义及输出格式将所述二维数组输出为二维数据表文件,返回给所述访问者。

[0136] 其中,所述字符串在被存入所述第一数据表的数据内容字段之前被按照指定的格式压缩;所述拆分单元402可以包括:

[0137] 第一提取子单元,用于提取所述目标记录条目的数据内容字段中的数据并根据所述指定的格式进行解压得到字符串。

[0138] 还可以将所述字符串被按照指定的输出格式写入到所述第一数据表的数据内容字段中,所述拆分单元402可以包括:

[0139] 第二提取子单元,用于将所述目标记录条目的数据内容字段中的字符串按照所述输出格式输出为标准文件。

[0140] 另外,如果所述第一数据表还包括日期字段,所述日期字段用于保存将所述第二数据表回流到所述第一数据表时的日期;则所述目标记录条目确定单元401具体可以用于:

[0141] 根据所述查询请求与所述第一数据表的名称字段及日期字段的匹配情况,确定目标记录条目。

[0142] 总之,在本申请实施例中提供的上述系统中,只需要在关系型数据库中建立一张表,而不需要每次在Hadoop上产生的数据报表都在关系型数据库上建立同样的表单,节省了存储空间,并且解放了数据库管理人员和开发人员,减少了中间环节。并且,可以快速的输入Hadoop集群上的任何数据表,并且形成输出接口。另外,在实现过程中,支持各种个性化的配置,可以根据Hadoop上的数据报表的不同来配置各种不同的输入输出规则,具有很强的灵活性。

[0143] 通过以上的实施方式的描述可知,本领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解,本申请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,该计算机软件产品可以存储在存储介质中,如ROM/RAM、磁碟、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)执行本申请各个实施例或者实施例的某些部分所述的方法。

[0144] 本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部

分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于系统或系统实施例而言,由于其基本相似于方法实施例,所以描述得比较简单,相关之处参见方法实施例的部分说明即可。以上所描述的系统及系统实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者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者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0145] 以上对本申请所提供的数据回流、关系型数据库中的数据访问方法及系统,进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的一般技术人员,依据本申请的思想,在具体实施方式及应用范围上均会有改变之处。综上所述,本说明书内容不应理解为对本申请的限制。

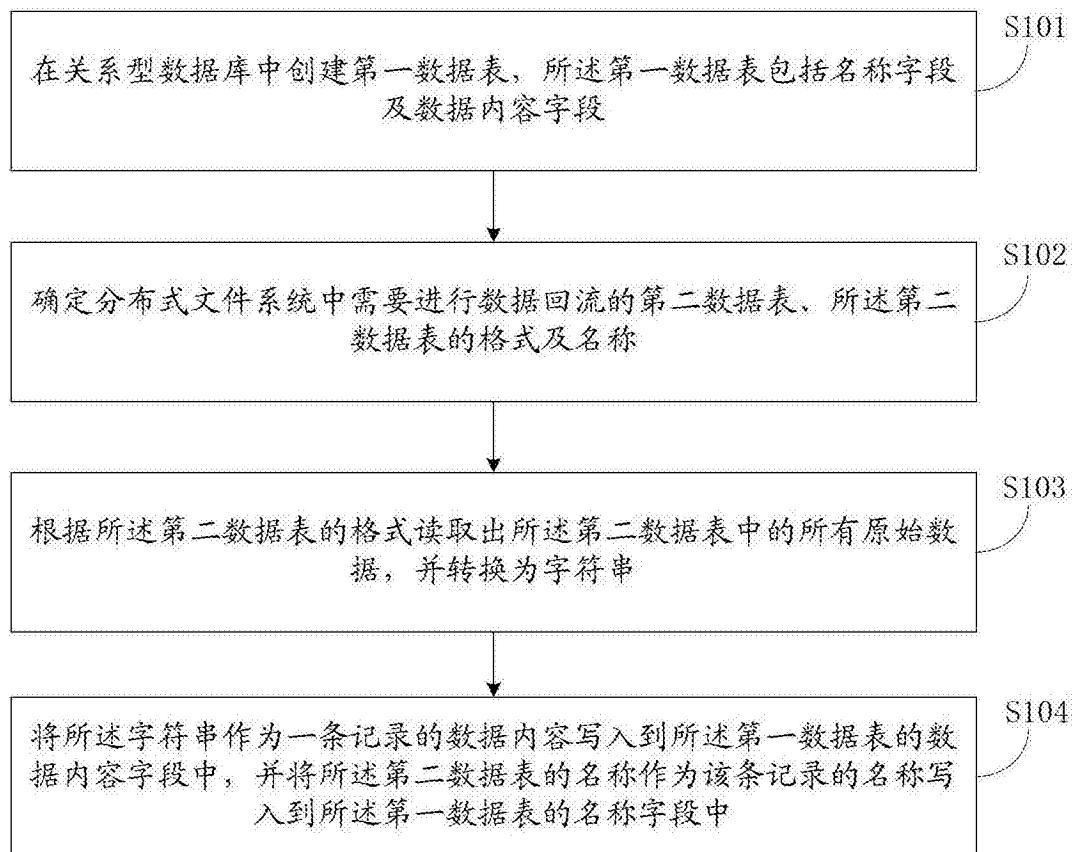


图1

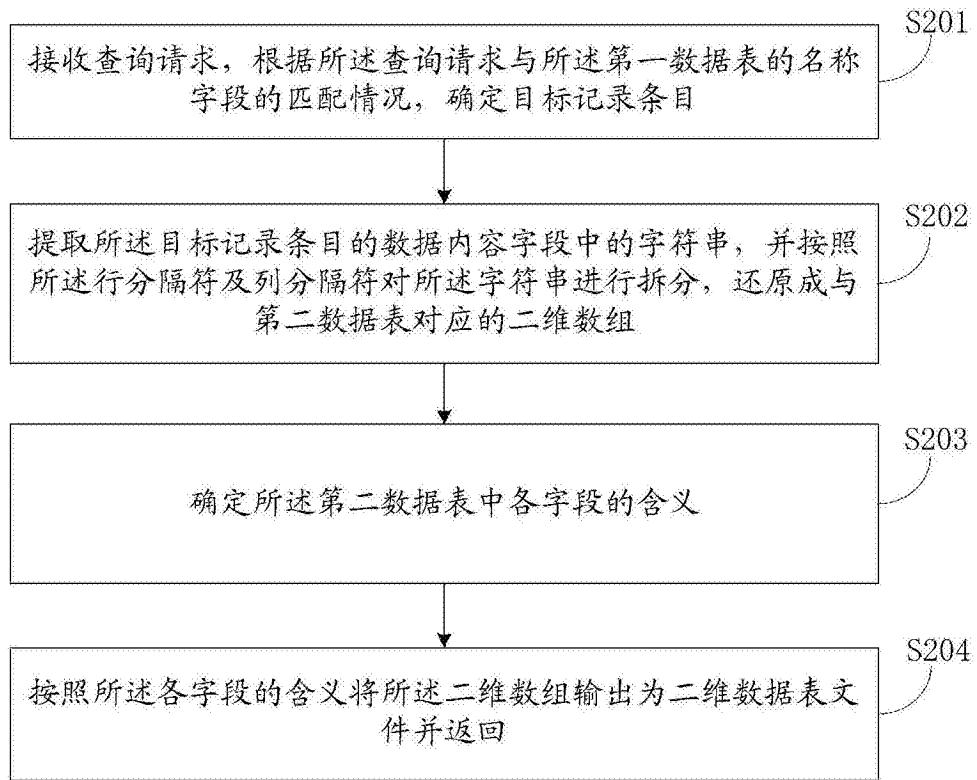


图2

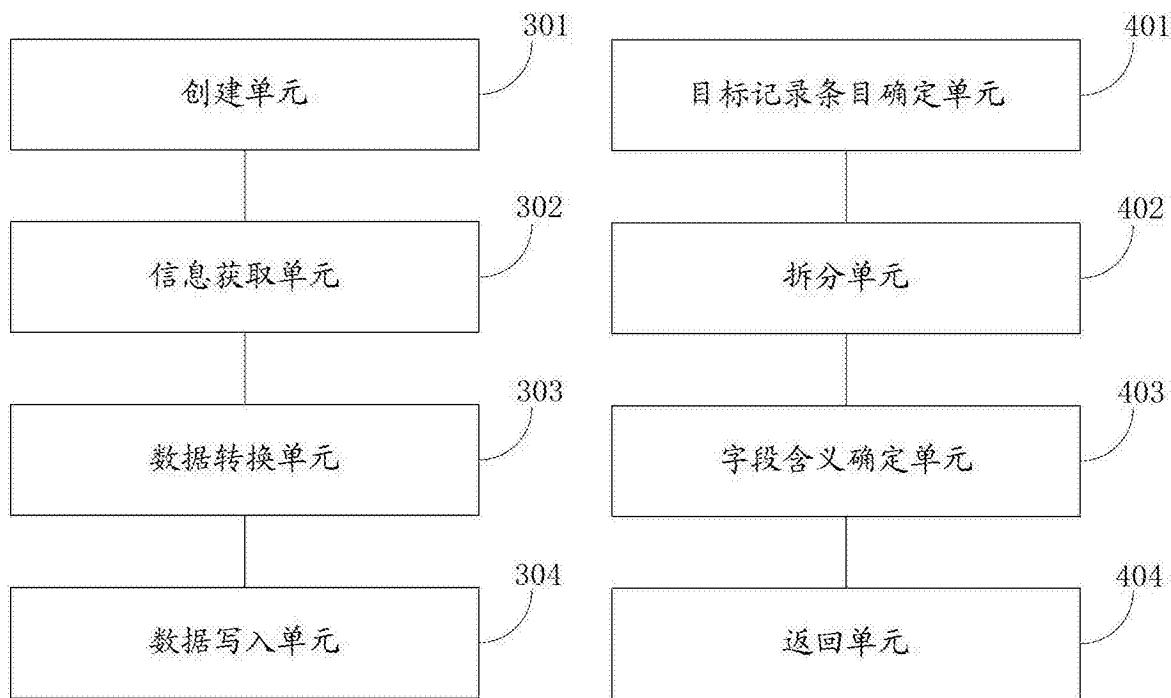


图3

图4