(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 108153852 A
(43)申请公布日 2018.06.12

(21)申请号 201711399172.0

(22)申请日 2017.12.22

(71)申请人 中国平安人寿保险股份有限公司
地址 518000 广东省深圳市福田区福华三路星河发展中心办公9、10、11层

(72)发明人 杨宏伟

(74)专利代理机构 深圳众鼎专利商标代理事务所(普通合伙) 44325
代理人 谭果林

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

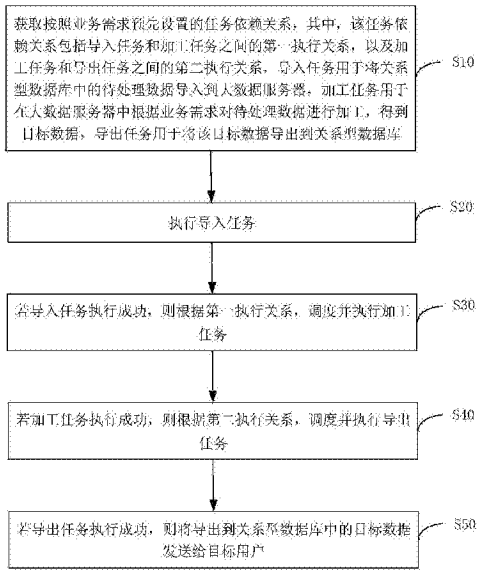
权利要求书2页 说明书11页 附图5页

(54)发明名称

一种数据处理方法、装置、终端设备及存储介质

(57)摘要

本发明公开了数据采集技术领域，提供了一种数据处理方法、装置、终端设备及存储介质，所述方法包括：获取按照业务需求预先设置的任务依赖关系，其中，任务依赖关系包括导入任务和加工任务之间的第一执行关系，以及加工任务和导出任务之间的第二执行关系；执行导入任务；若导入任务执行成功，则根据第一执行关系，调度并执行加工任务；若加工任务执行成功，则根据第二执行关系，调度并执行导出任务；若导出任务执行成功，则将导出到关系型数据库中的目标数据发送给目标用户。本发明的技术方案实现了在大数据环境下对待处理数据的自动加工以及目标数据的自动导出，减少人工干预，提升了数据下发的时效性。



CN 108153852 A

1.一种数据处理方法,其特征在于,所述数据处理方法包括:

获取按照业务需求预先设置的任务依赖关系,其中,所述任务依赖关系包括导入任务和加工任务之间的第一执行关系,以及所述加工任务和导出任务之间的第二执行关系,所述导入任务用于将关系型数据库中的待处理数据导入到大数据服务器,所述加工任务用于在所述大数据服务器中根据所述业务需求对所述待处理数据进行加工,得到目标数据,所述导出任务用于将所述目标数据导出到关系型数据库;

执行所述导入任务;

若所述导入任务执行成功,则根据所述第一执行关系,调度并执行所述加工任务;

若所述加工任务执行成功,则根据所述第二执行关系,调度并执行所述导出任务;

若所述导出任务执行成功,则将导出到所述关系型数据库中的所述目标数据发送给目标用户。

2.如权利要求1所述的数据处理方法,其特征在于,所述执行所述导入任务包括:

获取待连接的目标数据库的连接信息;

根据所述连接信息连接所述目标数据库;

创建用于在大数据环境中存放导入数据的第一数据表;

使用sqoop导入工具将所述待处理数据导入到所述第一数据表中。

3.如权利要求2所述的数据处理方法,其特征在于,所述若所述导入任务执行成功,则根据所述第一执行关系,调度并执行所述加工任务包括:

创建用于在所述大数据环境中存放加工后的所述目标数据的第二数据表;

根据所述业务需求,对所述第一数据表中的数据进行加工,得到所述目标数据;

将所述目标数据保存在所述第二数据表中。

4.如权利要求3所述的数据处理方法,其特征在于,所述若所述加工任务执行成功,则根据所述第二执行关系,调度并执行所述导出任务包括:

创建用于在关系型数据库中存放所述目标数据的第三数据表;

使用sqoop导出工具将所述第二数据表中保存的所述目标数据导出到所述第三数据表中;

根据所述导出任务的标识信息,在预存的导出标识表中将该标识信息对应的完成状态设置为导出完成,其中,所述导出标识表包含导出任务的标识信息和完成状态,所述标识信息用于唯一标识所述导出任务,所述完成状态用于标识所述导出任务是否完成。

5.如权利要求4所述的数据处理方法,其特征在于,所述若所述导出任务执行成功,则将导出到所述关系型数据库中的所述目标数据发送给目标用户包括:

定期读取所述导出标识表中所述导出任务的标识信息对应的完成状态;

若所述完成状态为导出完成,则确认所述导出任务执行完成,并将所述第三数据表中的目标数据发送给所述目标客户。

6.一种数据处理装置,其特征在于,所述数据处理装置包括:

依赖关系获取模块,用于获取按照业务需求预先设置的任务依赖关系,其中,所述任务依赖关系包括导入任务和加工任务之间的第一执行关系,以及所述加工任务和导出任务之间的第二执行关系,所述导入任务用于将关系型数据库中的待处理数据导入到大数据服务器,所述加工任务用于在所述大数据服务器中根据所述业务需求对所述待处理数据进行加

工，得到目标数据，所述导出任务用于将所述目标数据导出到关系型数据库；

任务导入模块，用于执行所述导入任务；

任务加工模块，用于若所述导入任务执行成功，则根据所述第一执行关系，调度并执行所述加工任务；

任务导出模块，用于若所述加工任务执行成功，则根据所述第二执行关系，调度并执行所述导出任务；

数据发送模块，用于若所述导出任务执行成功，则将导出到所述关系型数据库中的所述目标数据发送给目标用户。

7.如权利要求6所述的数据处理装置，其特征在于，任务导入模块包括：

连接信息获取子模块，用于获取待连接的目标数据库的连接信息；

数据库连接子模块，用于根据所述连接信息连接所述目标数据库；

第一创建子模块，用于创建用于在大数据环境中存放导入数据的第一数据表；

数据导入子模块，用于使用sqoop导入工具将所述待处理数据导入到所述第一数据表中。

8.如权利要求7所述的数据处理装置，其特征在于，所述任务加工模块包括：

第二创建子模块，用于创建用于在所述大数据环境中存放加工后的所述目标数据的第二数据表；

数据加工子模块，用于根据所述业务需求，对所述第一数据表中的数据进行加工，得到所述目标数据；

数据保存子模块，用于将所述目标数据保存在所述第二数据表中。

9.一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至5任一项所述数据处理方法的步骤。

10.一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至5任一项所述数据处理方法的步骤。

一种数据处理方法、装置、终端设备及存储介质

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种数据处理方法、装置、终端设备及存储介质。

背景技术

[0002] 传统的关系型数据库能够将采集到的目标数据自动导出，并通过邮件系统自动发送给用户，但是，随着数据量的不断增大，越来越多的数据采集平台基于大数据环境进行数据采集，而在大数据的环境中，常使用HIVE工具对数据进行加工，现有技术中，通过HIVE工具加工得到的目标数据无法直接自动导出，需人工手动导出后再发送给用户，影响了数据发送的及时性，导致数据下发的时效性较低。

发明内容

[0003] 本发明实施例提供一种数据处理方法，以解决现有技术中在大数据环境中加工后的目标数据无法直接自动导出，导致数据下发的时效性低的问题。

[0004] 第一方面，本发明实施例提供一种数据处理方法，包括：

[0005] 获取按照业务需求预先设置的任务依赖关系，其中，所述任务依赖关系包括导入任务和加工任务之间的第一执行关系，以及所述加工任务和导出任务之间的第二执行关系，所述导入任务用于将关系型数据库中的待处理数据导入到大数据服务器，所述加工任务用于在所述大数据服务器中根据所述业务需求对所述待处理数据进行加工，得到目标数据，所述导出任务用于将所述目标数据导出到关系型数据库；

[0006] 执行所述导入任务；

[0007] 若所述导入任务执行成功，则根据所述第一执行关系，调度并执行所述加工任务；

[0008] 若所述加工任务执行成功，则根据所述第二执行关系，调度并执行所述导出任务；

[0009] 若所述导出任务执行成功，则将导出到所述关系型数据库中的所述目标数据发送给目标用户。

[0010] 第二方面，本发明实施例提供一种数据处理的装置，包括：

[0011] 依赖关系获取模块，用于获取按照业务需求预先设置的任务依赖关系，其中，所述任务依赖关系包括导入任务和加工任务之间的第一执行关系，以及所述加工任务和导出任务之间的第二执行关系，所述导入任务用于将关系型数据库中的待处理数据导入到大数据服务器，所述加工任务用于在所述大数据服务器中根据所述业务需求对所述待处理数据进行加工，得到目标数据，所述导出任务用于将所述目标数据导出到关系型数据库；

[0012] 任务导入模块，用于执行所述导入任务；

[0013] 任务加工模块，用于若所述导入任务执行成功，则根据所述第一执行关系，调度并执行所述加工任务；

[0014] 任务导出模块，用于若所述加工任务执行成功，则根据所述第二执行关系，调度并执行所述导出任务；

[0015] 数据发送模块，用于若所述导出任务执行成功，则将导出到所述关系型数据库中的所述目标数据发送给目标用户。

[0016] 第三方面，本发明实施例提供一种终端设备，包括存储器、处理器以及存储在所述存储器中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所述数据处理方法的步骤。

[0017] 第四方面，本发明实施例提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现所述数据处理方法的步骤。

[0018] 本发明实施例与现有技术相比具有如下优点：本发明实施例所提供的数据处理方法、装置、终端设备及存储介质中，按照预先设置的任务依赖关系，首先执行导入任务，将关系型数据库中的待处理数据导入到大数据服务器后，然后调度并执行加工任务，在大数据服务器中根据业务需求对导入的待处理数据进行加工，得到目标数据，再执行导出任务，将目标数据导出到关系型数据库后，将关系型数据库中的目标数据直接发送给目标用户，从而实现在大数据环境下对待处理数据的自动加工以及目标数据的自动导出，减少人工干预，提升了数据下发的时效性。

附图说明

[0019] 为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例的描述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发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0020] 图1是本发明实施例1提供的数据处理方法的实现流程图；

[0021] 图2是本发明实施例1提供的数据处理方法中步骤S20的实现流程图；

[0022] 图3是本发明实施例1提供的数据处理方法中步骤S30的实现流程图；

[0023] 图4是本发明实施例1提供的数据处理方法中步骤S40的实现流程图；

[0024] 图5是本发明实施例1提供的数据处理方法中步骤S50的实现流程图；

[0025] 图6是本发明实施例2提供的数据处理装置的示意图；

[0026] 图7是本发明实施例4提供的终端设备的示意图。

具体实施方式

[0027] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0028] 实施例1

[0029] 请参阅图1，图1示出了本实施例提供的数据处理方法的实现流程。该数据处理方法应用在数据采集平台中，用于实现对数据的自动加工及导出。详述如下：

[0030] S10：获取按照业务需求预先设置的任务依赖关系，其中，该任务依赖关系包括导入任务和加工任务之间的第一执行关系，以及加工任务和导出任务之间的第二执行关系，导入任务用于将关系型数据库中的待处理数据导入到大数据服务器，加工任务用于在大数

据服务器中根据业务需求对待处理数据进行加工，得到目标数据，导出任务用于将该目标数据导出到关系型数据库。

[0031] 在本发明实施例中，根据业务需求预先确定导入任务、加工任务和导出任务，并按照业务需求设置导入任务、加工任务和导出任务之间的任务依赖关系，包括导入任务和加工任务之间的第一执行关系，以及加工任务与导出任务之间的第二执行关系。

[0032] 导入任务、加工任务和导出任务均可包含多个子任务，具体可根据业务需求确定。例如，某业务需求涉及到三张数据表，则导入任务可以包括三个子任务，每个子任务针对一张数据表进行数据导入。

[0033] 第一执行关系是指加工任务的执行依赖导入任务的成功完成，当导入任务包含多个导入子任务时，需要在该导入任务包含的导入子任务全部执行完成后才执行加工任务。

[0034] 第二执行关系是指导出任务的执行依赖加工任务的成功完成，当加工任务包含多个加工子任务时，需要在该加工任务包含的加工子任务全部执行完成后才执行导出任务。

[0035] 进一步地，导入任务可以是sqoop导入任务，导出任务可以是sqoop导出任务，加工任务可以使HIVE加工任务，sqoop导入任务、sqoop导出任务，以及HIVE加工任务均以脚本文件的形式保存。

[0036] sqoop是一款开源工具，主要用于在大数据的分布式文件系统与传统的关系型数据库之间进行数据的传递，其可以将关系型数据库中的数据导入到大数据文件系统中，也可以将大数据文件系统中的数据导出到关系型数据库中。其中，分布式文件系统(Distributed File System,DFS)可以是Hadoop、Mortar Data、Continuuity等应用程序构建，例如HDFS(Hadoop Distributed File System)，关系型数据库可以是MySQL、Oracle、Postgres等数据库。

[0037] 可以理解的是，任务依赖关系具体可以通过依赖关系表体现，在依赖关系表中包括任务名称和被依赖任务名称两个字段，例如任务A依赖任务B和任务C，则在依赖关系表中增加A依赖B和A依赖C的两条记录。

[0038] 导入任务将关系型数据库中的待处理数据导入到大数据服务器，加工任务在大数据服务器中根据业务需求对待处理数据进行加工，得到目标数据，导出任务将该目标数据导出到关系型数据库，通过预先配置导入任务、加工任务，以及导出任务之间的依赖关系，并按照该依赖关系对数据进行有序的导入、加工和导出，确保不同的任务之间能够按照依赖关系顺序执行，从而提高数据获取的准确性。

[0039] S20:执行导入任务。

[0040] 在本发明实施例中，在获取任务依赖关系后，首先执行该任务依赖关系中确定的导入任务。

[0041] 具体地，通过sqoop导入的方式将关系型数据库中的待处理数据导入到大数据环境中的分布式文件系统中，导入任务包含的每个导入子任务均以sqoop导入脚本文件的形式保存。

[0042] S30:若导入任务执行成功，则根据第一执行关系，调度并执行加工任务。

[0043] 在本发明实施例中，根据第一执行关系，即加工任务的执行依赖导入任务的成功完成，在确认导入任务执行成功后，调度并执行任务依赖关系中确定的加工任务。

[0044] 具体地，采用HIVE工具对导入到分布式文件系统中的待处理数据进行加工，得到

目标数据，加工任务包含的每个加工子任务均以HIVE脚本文件的形式保存。需要说明的是，若导入任务执行失败，则无法根据第一执行关系调度加工任务，此时，可以在预设的导入次数内重复执行该导入任务，直到达到预设的最大导入次数或者该导入任务执行成功为止。

[0045] 若达到预设的最大导入次数时该导入任务依然执行失败，则该业务需求的数据处理失败，流程结束。

[0046] S40：若加工任务执行成功，则根据第二执行关系，调度并执行导出任务。

[0047] 在本发明实施例中，根据第二执行关系，即导出任务的执行依赖加工任务的成功完成，在确认加工任务执行成功后，调度并执行任务依赖关系中确定的导出任务。

[0048] 具体地，通过sqoop导出的方式将加工后的目标数据从大数据环境中的分布式文件系统导出到关系型数据库，导出任务包含的每个导出子任务均以sqoop导出脚本文件的形式保存。需要说明的是，若加工任务执行失败，则无法根据第二执行关系调度导出任务，此时，可以在预设的加工次数内重复执行该加工任务，直到达到预设的最大加工次数或者该加工任务执行成功为止。

[0049] 若达到预设的最大加工次数时该加工任务依然执行失败，则该业务需求的数据处理失败，流程结束。

[0050] S50：若导出任务执行成功，则将导出到关系型数据库中的目标数据发送给目标用户。

[0051] 在本发明实施例中，在确认导出任务执行成功后，根据预先设置的发送方式，将导出到关系型数据库中的目标数据发送给该目标用户。

[0052] 可以理解的是，该预先设置的发送方式可以是邮件发送的方式或者即时消息发送的方式，还可以是其他设置的方式，此处不做限制。以邮件发送方式为例，在导出任务执行完成后，根据预先保存的邮箱地址将目标数据发送给目标用户。

[0053] 需要说明的是，若导出任务执行失败，则可以通过预先设置重复导出次数的方式，在该重复导出次数内重复执行该导出任务，直到达到预设的最大导出次数或者该导出任务执行成功为止。

[0054] 若达到预设的最大导出次数时该导出任务依然执行失败，则该业务需求的数据处理失败，流程结束。

[0055] 在图1对应的实施例中，通过获取按照业务需求预先设置的任务依赖关系，并按照该任务依赖关系，首先执行导入任务，将关系型数据库中的待处理数据导入到大数据服务器后，然后调度并执行加工任务，在大数据服务器中根据业务需求对导入的待处理数据进行加工，得到目标数据，再执行导出任务，将目标数据导出到关系型数据库后，将关系型数据库中的目标数据直接发送给目标用户，从而实现在大数据环境下对待处理数据的自动加工以及目标数据的自动导出，减少人工干预，提升数据下发的时效性；同时，通过预先配置导入任务、加工任务，以及导出任务之间的依赖关系，并按照该依赖关系对数据进行有序的导入、加工和导出，确保不同的任务之间能够按照依赖关系顺序执行，从而提高数据获取的准确性。

[0056] 接下来，在图1对应的实施例的基础之上，下面通过一个具体的实施例对步骤S20中提及的执行导入任务的具体实现方法进行详细说明。

[0057] 请参阅图2，图2示出了本发明实施例提供的步骤S20的具体实现流程，详述如下：

[0058] S201:获取待连接的目标数据库的连接信息。

[0059] 在本发明实施例中,待连接的目标数据库为待处理数据所在的关系型数据库,从目标数据库采集待处理数据前,首先须与该目标数据库建立连接关系。其中,待处理数据是数据采集具体的业务需求涉及到的数据,包括数据表,以及数据表中的目标字段等。

[0060] 目标数据库的连接信息包括连接目标数据库的用户名、密码,以及连接串信息。例如,用户名为gbdsqp,密码为paic0987,连接串信息为jdbc:oracle:thin:@192.168.1.1:1526:test,其中,test为数据库名。

[0061] S202:根据待连接的目标数据库的连接信息连接该目标数据库。

[0062] 具体地,根据步骤S201获取到的连接信息连接目标数据库。

[0063] 例如,若待连接的目标数据库的名称为test,连接信息中的用户名为gbdsqp,密码为paic0987,连接串信息为jdbc:oracle:thin:@192.168.1.1:1526:test,则使用该连接信息连接目标数据库test的方法如下:

[0064] oracle_connection:::::jdbc:oracle:thin:@192.168.1.1:1526:test

[0065] oracle_username:::::gbdsqp

[0066] oracle_password:::::paic0987

[0067] 需要说明的是,若目标数据库连接成功,则继续执行步骤S203,完成导入任务的执行;若目标数据库连接失败,则确认导入任务执行失败,不再继续执行后面的步骤,直接输出数据处理失败,以及失败提示信息,该失败提示信息包括导入任务执行失败的提示信息,以及连接失败的数据库的标识信息等

[0068] S203:创建用于在大数据环境中存放导入数据的第一数据表。

[0069] 在本发明实施例中,第一数据表用于在大数据环境中存放从关系型数据库中导入的待处理数据。第一数据表具体可以是HIVE数据表,该HIVE数据表以文本文件的形式保存。

[0070] S204:使用sqoop导入工具将待处理数据导入到第一数据表中。

[0071] 具体地,根据步骤S203创建的第一数据表,以及步骤S202连接成功的目标数据库,使用sqoop导入工具将目标数据库中的待处理数据导入到第一数据表中。

[0072] 在图2对应的实施例中,根据待连接的目标数据库的连接信息成功连接目标数据库,创建用于在大数据环境中存放导入的待处理数据的第一数据表,并使用sqoop导入工具将待处理数据从目标数据库导入到第一数据表中,从而实现将将关系型数据库中的待处理数据自动准确的导入到大数据环境中的分布式文件系统中。

[0073] 在图2对应的实施例的基础之上,下面通过一个具体的实施例对步骤S30中提及的若导入任务执行成功,则根据第一执行关系,调度并执行加工任务的具体实现方法进行详细说明。

[0074] 请参阅图3,图3示出了本发明实施例提供的步骤S30的具体实现流程,详述如下:

[0075] S301:创建用于在大数据环境中存放加工后的目标数据的第二数据表。

[0076] 在本发明实施例中,第二数据表用于在大数据环境中存放对导入的待处理数据进行加工后的目标数据。第二数据表具体可以是HIVE数据表,该HIVE数据表以文本文件的形式保存。

[0077] S302:根据业务需求,对第一数据表中的数据进行加工,得到目标数据。

[0078] 具体地,在按照业务需求预先设置的任务依赖关系中确定了加工任务的加工内

容，根据该加工内容，采用HIVE工具对第一数据表中导入的待处理数据进行加工，得到目标数据。

[0079] S303：将目标数据保存在第二数据表中。

[0080] 具体地，将步骤S302加工得到的目标数据保存在步骤S301创建的第二数据表中。

[0081] 下面以面向保险业务员的某个应用app为例进行说明：

[0082] 假设业务需求为获取预设时间内使用该应用app的保险业务员的机构信息。第一数据表包括登录记录表和业务员信息表，其中，登录记录表记录了保险业务员编号、应用使用时间和该保险业务员具体使用的页面等登录数据，当保险业务员登录该应用app时，该保险业务员的登录数据将被记录在该登录记录表中，业务员信息表记录了保险业务员的个人信息及其所属的业务机构的信息，其中，个人信息包括保险业务员编号、业务员姓名，以及入职时间等，业务机构的信息包括二级机构名称、二级机构代码、三级机构名称、三级机构代码、营业区名称、营业区代码、营业部名称、营业部代码、营业组名称，以及营业组代码等。

[0083] 因此，创建的存放加工后的目标数据的第二数据表包含的字段包括二级机构名称、二级机构代码、三级机构名称、三级机构代码、营业区名称、营业区代码、营业部名称、营业部代码、营业组名称，以及营业组代码等。根据业务需求确定的加工任务为从登录记录表中获取应用使用时间在预设时间范围内的保险业务员编号，再根据该保险业务员编号从业务员信息表中获取对应的机构信息。按照该加工任务对登录记录表和业务员信息表中的相关数据进行加工后，得到预设时间内使用该应用app的保险业务员的机构信息，并将该机构信息保存在第二数据表中。

[0084] 在图3对应的实施例中，创建用于在大数据环境中存放目标数据的第二数据表，对第一数据表中的待处理数据进行加工，将加工得到的目标数据保存在第二数据表中，实现了对导入到分布式文件系统中的待处理数据自动进行HIVE加工，得到目标数据，通过HIVE加工的方式能够有效提高在大数据环境中数据加工的效率。

[0085] 在图3对应的实施例的基础之上，下面通过一个具体的实施例对步骤S40中提及的若加工任务执行成功，则根据第二执行关系，调度并执行导出任务的具体实现方法进行详细说明。

[0086] 请参阅图4，图4示出了本发明实施例提供的步骤S40的具体实现流程，详述如下：

[0087] S401：创建用于在关系型数据库中存放目标数据的第三数据表。

[0088] 在本发明实施例中，第三数据表用于在关系型数据库中存放导出后的目标数据，该第三数据表中包含目标数据的字段。

[0089] 需要说明的是，该第三数据表只是一份临时数据表，当目标数据从该第三数据表中完全导出成功后，该第三数据表将被删除。采用临时数据表能够有效减少对存储空间的占用，提高存储空间的利用率。

[0090] S402：使用sqoop导出工具将第二数据表中保存的目标数据导出到第三数据表中。

[0091] 具体地，根据在步骤S303中第二数据表保存的目标数据，通过sqoop导出工具将该目标数据导出到步骤S401创建的第三数据表中。

[0092] S403：根据导出任务的标识信息，在预存的导出标识表中将该标识信息对应的完成状态设置为导出完成，其中，该导出标识表包含导出任务的标识信息和完成状态，导出任务的标识信息用于唯一标识该导出任务，导出任务的完成状态用于标识该导出任务是否完

成。

[0093] 在本发明实施例中，每个导出任务对应一个标识信息，该标识信息具体可以是一个唯一的随机序列，在生成导出任务的时候同步生成该随机序列，用于唯一标识该导出任务。

[0094] 具体地，通过预先创建的导出标识表，采用写标签的方式记录导出任务的完成状态。导出标识表包含导出任务的标识信息和完成状态，当生成导出任务时，在该导出标识表中新增一条记录，保存该导出任务的标识信息，并将该条记录中的完成状态设置为导出未完成。当在步骤S402中的导出任务执行成功后，在导出标识表中查询该导出任务的标识信息对应的记录，并将该记录中的完成状态设置为导出完成。

[0095] 通过读取导出标识表中记录的导出任务的完成状态，能够及时捕获导出任务是否执行完成。

[0096] 在图4对应的实施例中，创建用于在关系型数据库中存放导出后的目标数据的第三数据表，利用sqoop导出工具将第二数据表保存的加工后的目标数据导出到第三数据表中，实现了将加工后的目标数据自动准确的从大数据环境中的分布式文件系统导出到关系型数据库中。并且，使用导出标识表，采用写标签的方式记录导出任务的完成状态，从而通过读取导出标识表中记录的导出任务的完成状态，能够及时捕获导出任务是否执行完成。

[0097] 在图4对应的实施例的基础之上，下面通过一个具体的实施例对步骤S50中提及的若导出任务执行成功，则将导出到关系型数据库中的目标数据发送给目标用户的具体实现方法进行详细说明。

[0098] 请参阅图5，图5示出了本发明实施例提供的步骤S50的具体实现流程，详述如下：

[0099] S501：定期读取导出标识表中导出任务的标识信息对应的完成状态。

[0100] 具体地，数据采集平台通过定期循环读取的方式，读取导出标识表中记录的每个导出任务的标识信息对应的完成状态，根据该完成状态判断该导出任务是否完成。

[0101] S502：若导出任务的标识信息对应的完成状态为导出完成，则确认该导出任务执行完成，并将第三数据表中的目标数据发送给目标客户。

[0102] 具体地，若在步骤S501中读取到导出标识表中某个导出任务的标识信息对应的完成状态为导出完成，则根据该导出任务的标识信息获取该导出任务，并将该导出任务在第三数据表中保存的目标数据发送给目标客户。

[0103] 可以理解的是，导出标识表中记录了多个不同导出任务的标识信息和完成状态，不同的导出任务对应的业务需求不同，数据采集平台通过定期循环读取导出标识表中的完成状态，当发现完成状态为导出完成时，获取该完成状态对应的导出任务，并将该导出任务在第三数据表中保存的目标数据发送给目标客户，从而能够及时获知业务需求是否处理完成，并自动将导出后的目标数据发送给目标客户，减少人工干预，有效提升数据下发的时效性。

[0104] 在图5对应的实施例中，采用读标签的方式定期读取导出标识表中导出任务的标识信息对应的，判断导出任务是否执行完成，并在确定导出任务执行完成时将第三数据表中的目标数据及时发送给目标用户，这种读标签的方式能够及时获知业务需求是否处理完成，并及时将导出后的目标数据自动发送给目标客户，减少人工干预，有效提升数据下发的时效性。

[0105] 应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本发明实施例的实施过程构成任何限定。

[0106] 实施例2

[0107] 对应于实施例1所述的数据处理方法，图6示出了与实施例1提供的数据处理方法一一对应的数据处理装置，为了便于说明，仅示出了与本发明实施例相关的部分。

[0108] 请参阅图6，该数据处理装置包括：依赖关系获取模块61、任务导入模块62、任务加工模块63、任务导出模块64和数据发送模块65，各功能模块详细说明如下：

[0109] 依赖关系获取模块61，用于获取按照业务需求预先设置的任务依赖关系，其中，任务依赖关系包括导入任务和加工任务之间的第一执行关系，以及加工任务和导出任务之间的第二执行关系，导入任务用于将关系型数据库中的待处理数据导入到大数据服务器，加工任务用于在大数据服务器中根据业务需求对待处理数据进行加工，得到目标数据，导出任务用于将目标数据导出到关系型数据库；

[0110] 任务导入模块62，用于执行导入任务；

[0111] 任务加工模块63，用于若导入任务执行成功，则根据第一执行关系，调度并执行加工任务；

[0112] 任务导出模块64，用于若加工任务执行成功，则根据第二执行关系，调度并执行导出任务；

[0113] 数据发送模块65，用于若导出任务执行成功，则将导出到关系型数据库中的目标数据发送给目标用户。

[0114] 进一步地，任务导入模块62包括：

[0115] 连接信息获取子模块621，用于获取待连接的目标数据库的连接信息；

[0116] 数据库连接子模块622，用于根据连接信息连接目标数据库；

[0117] 第一创建子模块623，用于创建用于在大数据环境中存放导入数据的第一数据表；

[0118] 数据导入子模块624，用于使用sqoop导入工具将待处理数据导入到第一数据表中。

[0119] 进一步地，任务加工模块63包括：

[0120] 第二创建子模块631，用于创建用于在大数据环境中存放加工后的目标数据的第二数据表；

[0121] 数据加工子模块632，用于根据业务需求，对第一数据表中的数据进行加工，得到目标数据；

[0122] 数据保存子模块633，用于将目标数据保存在第二数据表中。

[0123] 进一步地，任务导出模块64包括：

[0124] 第三创建子模块641，用于创建用于在关系型数据库中存放目标数据的第三数据表；

[0125] 数据导出子模块642，用于使用sqoop导出工具将第二数据表中保存的目标数据导出到第三数据表中；

[0126] 状态设置子模块643，用于根据导出任务的标识信息，在预存的导出标识表中将该标识信息对应的完成状态设置为导出完成，其中，导出标识表包含导出任务的标识信息和

完成状态，该标识信息用于唯一标识导出任务，该完成状态用于标识导出任务是否完成。

[0127] 进一步地，数据发送模块65包括：

[0128] 信息读取模块651，用于定期读取导出标识表中导出任务的标识信息对应的完成状态；

[0129] 数据发送模块652，用于若完成状态为导出完成，则确认导出任务执行完成，并将第三数据表中的目标数据发送给目标客户。

[0130] 本实施例提供的一种数据处理装置中各模块实现各自功能的过程，具体可参考前述实施例1的描述，此处不再赘述。

[0131] 实施例3

[0132] 本实施例提供一计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器执行时实现实施例1中的数据处理方法，或者，该计算机程序被处理器执行时实现实施例2中数据处理装置中各模块的功能，为避免重复，这里不再赘述。

[0133] 实施例4

[0134] 图7是本发明一实施例提供的终端设备的示意图。如图7所示，该实施例的终端设备70包括：处理器71、存储器72以及存储在存储器72中并可在处理器71上运行的计算机程序73，例如数据处理程序。处理器71执行计算机程序73时实现上述各个数据处理方法实施例中的步骤，例如图1所示的步骤S10至步骤S50。或者，处理器71执行计算机程序73时实现上述各装置实施例中各模块的功能，例如图6所示模块61至模块65的功能。

[0135] 示例性的，计算机程序73可以被分割成一个或多个模块，一个或者多个模块被存储在存储器72中，并由处理器71执行，以完成本发明。一个或多个模块可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述计算机程序73在终端设备70中的执行过程。例如，计算机程序73可以被分割成依赖关系获取模块、任务导入模块、任务加工模块、任务导出模块和数据发送模块，各模块具体功能如下：

[0136] 依赖关系获取模块，用于获取按照业务需求预先设置的任务依赖关系，其中，任务依赖关系包括导入任务和加工任务之间的第一执行关系，以及加工任务和导出任务之间的第二执行关系，导入任务用于将关系型数据库中的待处理数据导入到大数据服务器，加工任务用于在大数据服务器中根据业务需求对待处理数据进行加工，得到目标数据，导出任务用于将目标数据导出到关系型数据库；

[0137] 任务导入模块，用于执行导入任务；

[0138] 任务加工模块，用于若导入任务执行成功，则根据第一执行关系，调度并执行加工任务；

[0139] 任务导出模块，用于若加工任务执行成功，则根据第二执行关系，调度并执行导出任务；

[0140] 数据发送模块，用于若导出任务执行成功，则将导出到关系型数据库中的目标数据发送给目标用户。

[0141] 进一步地，任务导入模块包括：

[0142] 连接信息获取子模块，用于获取待连接的目标数据库的连接信息；

[0143] 数据库连接子模块，用于根据连接信息连接目标数据库；

[0144] 第一创建子模块，用于创建用于在大数据环境中存放导入数据的第一数据表；

[0145] 数据导入子模块，用于使用sqoop导入工具将待处理数据导入到第一数据表中。

[0146] 进一步地，任务加工模块包括：

[0147] 第二创建子模块，用于创建用于在大数据环境中存放加工后的目标数据的第二数据表；

[0148] 数据加工子模块，用于根据业务需求，对第一数据表中的数据进行加工，得到目标数据；

[0149] 数据保存子模块，用于将目标数据保存在第二数据表中。

[0150] 进一步地，任务导出模块包括：

[0151] 第三创建子模块，用于创建用于在关系型数据库中存放目标数据的第三数据表；

[0152] 数据导出子模块，用于使用sqoop导出工具将第二数据表中保存的目标数据导出到第三数据表中；

[0153] 状态设置子模块，用于根据导出任务的标识信息，在预存的导出标识表中将该标识信息对应的完成状态设置为导出完成，其中，导出标识表包含导出任务的标识信息和完成状态，该标识信息用于唯一标识导出任务，该完成状态用于标识导出任务是否完成。

[0154] 进一步地，数据发送模块包括：

[0155] 信息读取模块，用于定期读取导出标识表中导出任务的标识信息对应的完成状态；

[0156] 数据发送模块，用于若完成状态为导出完成，则确认导出任务执行完成，并将第三数据表中的目标数据发送给目标客户。

[0157] 终端设备70可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。终端设备可包括，但不仅限于，处理器71、存储器72。本领域技术人员可以理解，图7仅仅是终端设备70的示例，并不构成对终端设备70的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如终端设备还可以包括输入输出设备、网络接入设备、总线等。

[0158] 处理器71可以是中央处理单元(Central Processing Unit,CPU)，还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0159] 存储器72可以是终端设备70的内部存储单元，例如终端设备70的硬盘或内存。存储器72也可以是终端设备70的外部存储设备，例如终端设备70上配备的插接式硬盘，智能存储卡(Smart Media Card,SMC)，安全数字(Secure Digital,SD)卡，闪存卡(Flash Card)等。进一步地，存储器72还可以既包括终端设备70的内部存储单元也包括外部存储设备。存储器72用于存储计算机程序以及终端设备所需的其他程序和数据。存储器72还可以用于暂时地存储已经输出或者将要输出的数据。

[0160] 所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将所述装置的内部结构划分成不同的功能单元或模块，以完成以上

描述的全部或者部分功能。

[0161] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，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，也可以采用软件功能单元的形式实现。

[0162] 所述集成的模块如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明实现上述实施例方法中的全部或部分流程，也可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一计算机可读存储介质中，该计算机程序在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机程序包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读介质不包括是电载波信号和电信信号。

[0163] 以上所述实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围，均应包含在本发明的保护范围之内。

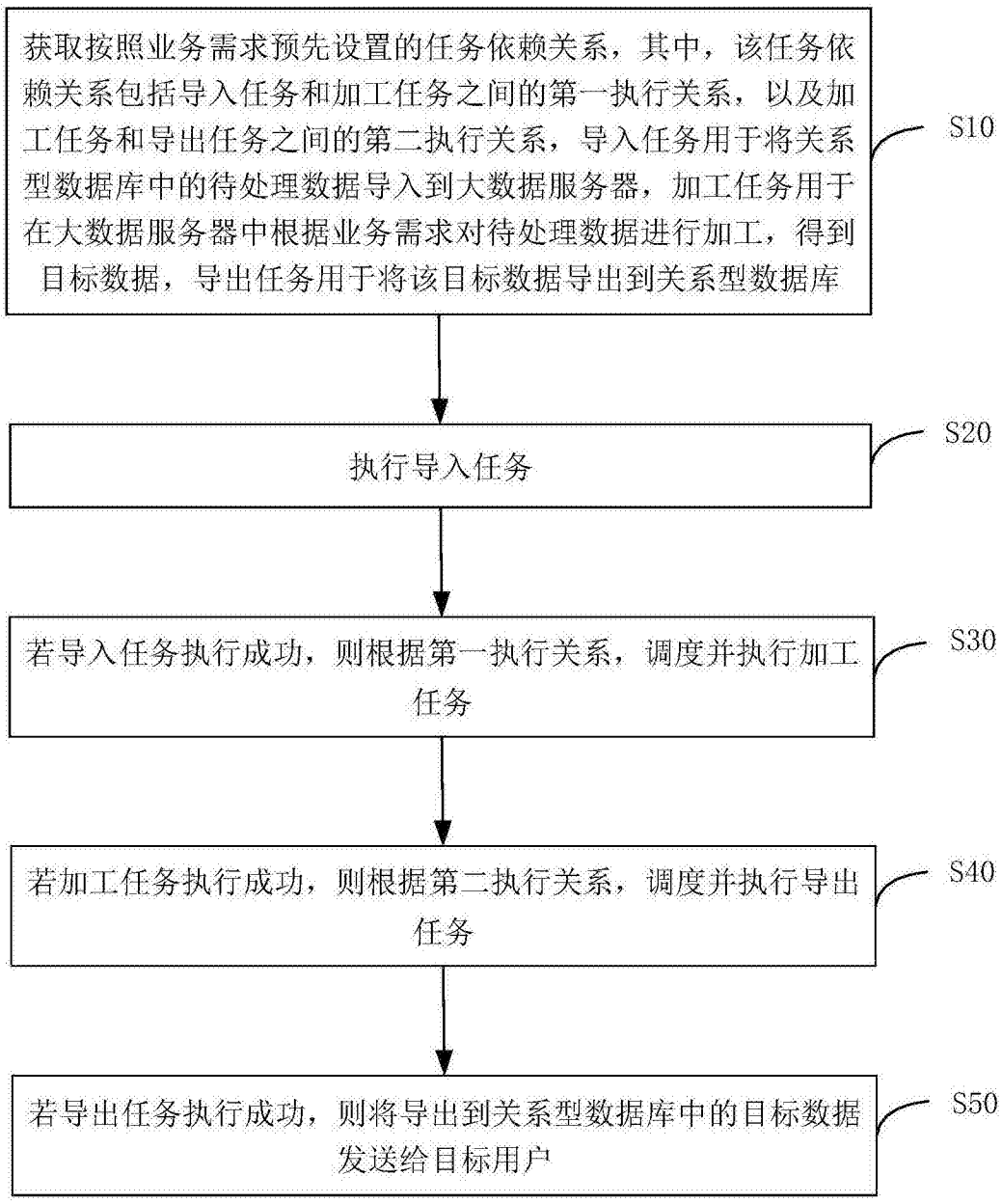

获取按照业务需求预先设置的任务依赖关系，其中，该任务依赖关系包括导入任务和加工任务之间的第一执行关系，以及加工任务和导出任务之间的第二执行关系，导入任务用于将关系型数据库中的待处理数据导入到大数据服务器，加工任务用于在大数据服务器中根据业务需求对待处理数据进行加工，得到目标数据，导出任务用于将该目标数据导出到关系型数据库
S10
执行导入任务
S20
若导入任务执行成功，则根据第一执行关系，调度并执行加工任务
S30
若加工任务执行成功，则根据第二执行关系，调度并执行导出任务
S40
若导出任务执行成功，则将导出到关系型数据库中的目标数据发送给目标用户
S50

图1

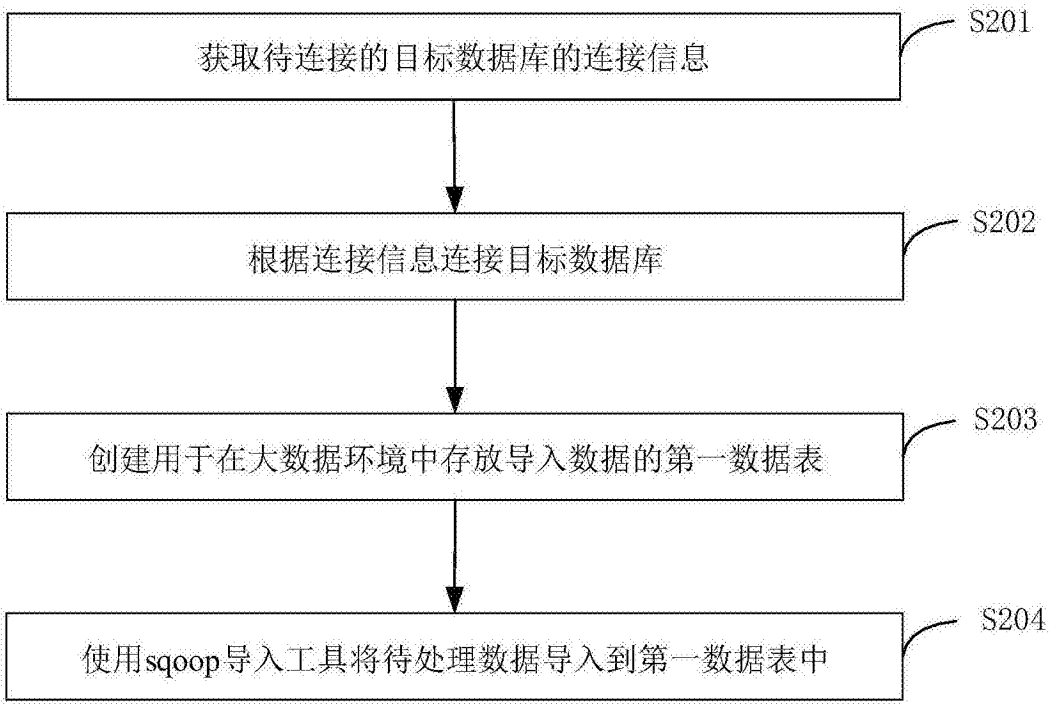

获取待连接的目标数据库的连接信息
S201
根据连接信息连接目标数据库
S202
创建用于在大数据环境中存放导入数据的第一数据表
S203
使用sqoop导入工具将待处理数据导入到第一数据表中
S204

图2

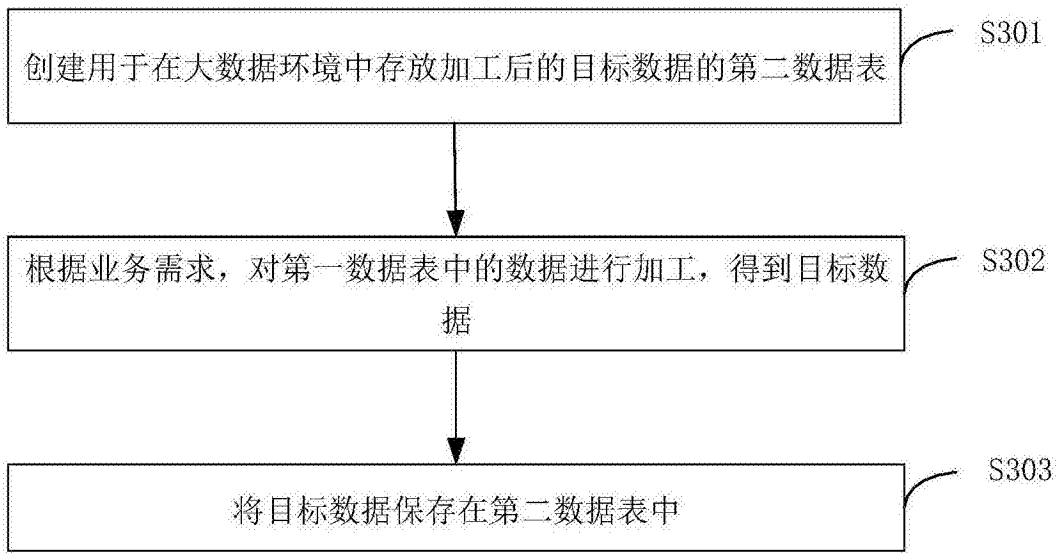

创建用于在大数据环境中存放加工后的目标数据的第二数据表
S301
根据业务需求，对第一数据表中的数据进行加工，得到目标数据
S302
将目标数据保存在第二数据表中
S303

图3

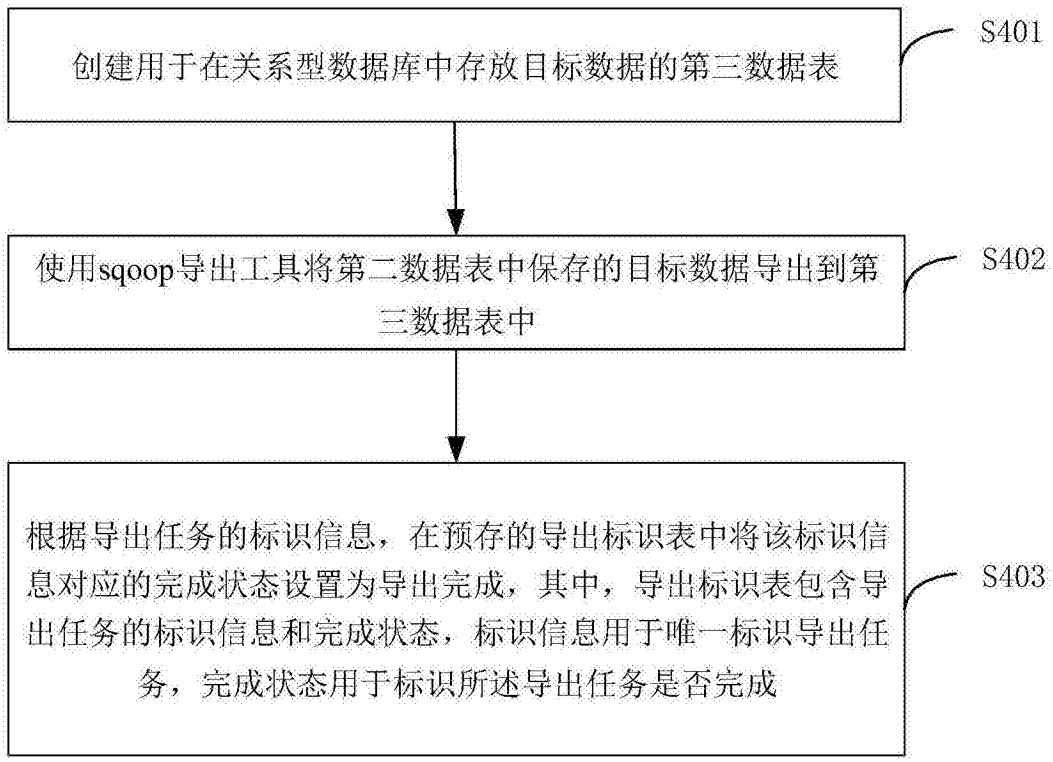

创建用于在关系型数据库中存放目标数据的第三数据表
S401
使用sqoop导出工具将第二数据表中保存的目标数据导出到第三数据表中
S402
根据导出任务的标识信息，在预存的导出标识表中将该标识信息对应的完成状态设置为导出完成，其中，导出标识表包含导出任务的标识信息和完成状态，标识信息用于唯一标识导出任务，完成状态用于标识所述导出任务是否完成
S403

图4

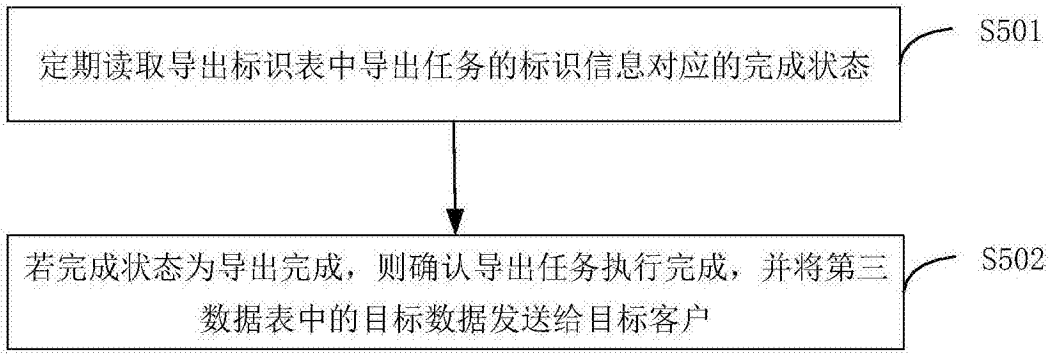

定期读取导出标识表中导出任务的标识信息对应的完成状态
S501
若完成状态为导出完成，则确认导出任务执行完成，并将第三数据表中的目标数据发送给目标客户
S502

图5

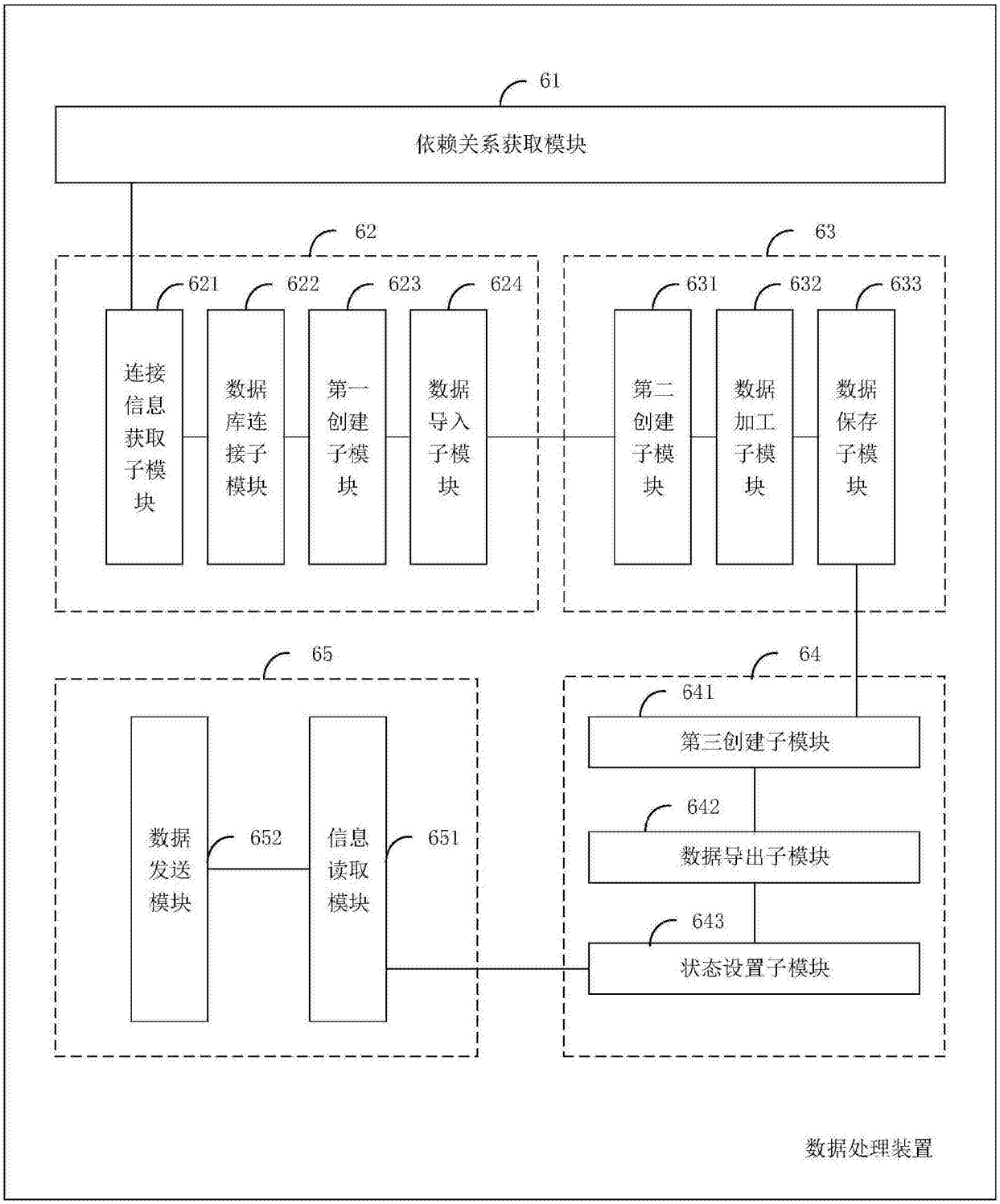

61
依赖关系获取模块
62
621
连接信息获取子模块
622
数据库连接子模块
623
第一创建子模块
624
数据导入子模块
63
631
第二创建子模块
632
数据加工子模块
633
数据保存子模块
65
数据发送模块
652
信息读取模块
651
64
641
第三创建子模块
642
数据导出子模块
643
状态设置子模块
数据处理装置

图6

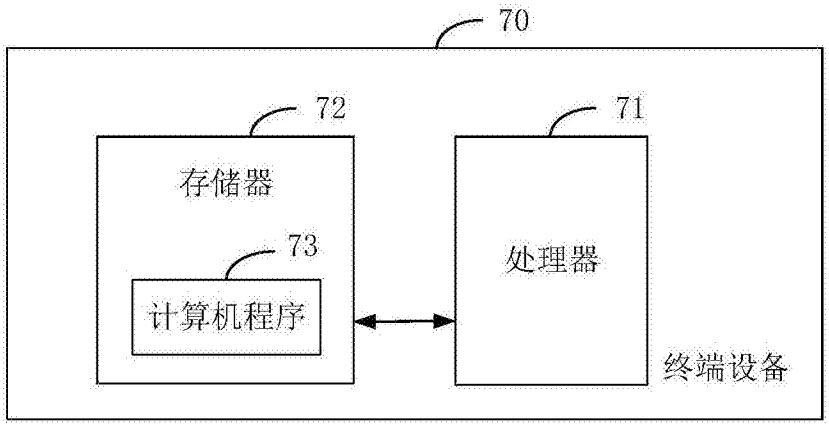


图7