(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 109299183 A
(43)申请公布日 2019.02.01

(21)申请号 201811386474.9

(22)申请日 2018.11.20

(71)申请人 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
申请人 青海省公安厅

(72)发明人 火一莽 许山川 王生玉

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
G06F 16/25(2019.01)
G06F 16/215(2019.01)

权利要求书2页 说明书13页 附图8页

(54)发明名称

一种数据处理方法、装置、终端设备和存储介质

(57)摘要

本发明公开了一种数据处理方法、装置、终端设备和存储介质。该方法包括:获取两个或两个以上的原始数据,该原始数据包括字段信息和数据内容;将原始数据转换为字段标识,并且从字段信息中提取主关键字;原始数据对应的字段标识与原始数据库中的字段标识无交集,则将原始数据写入原始数据库;将写入原始数据库的原始数据进行标准化预处理,得到清洗数据;清洗数据的主关键字与清洗数据库中的主关键字无交集,则将清洗数据写入清洗数据库。本发明在数据处理过程中,无需将原始数据库中所有数据对应的主关键字和清洗数据库中的主关键字进行比对,从而提高了数据处理效率。

CN 109299183 A

1.一种数据处理方法,其特征在于,包括:

获取两个或两个以上的原始数据,所述原始数据包括字段信息和数据内容;

将所述原始数据转换为字段标识,并且从所述字段信息中提取主关键字;

所述原始数据对应的字段标识与原始数据库中的字段标识无交集,则将所述原始数据写入所述原始数据库;

将写入所述原始数据库的原始数据进行标准化预处理,得到清洗数据;

所述清洗数据的主关键字与清洗数据库中的主关键字无交集,则将所述清洗数据写入所述清洗数据库。

2.根据权利要求1所述的数据处理方法,其特征在于,在所述获取两个或两个以上的原始数据之前,还包括:

获取原始数据文件,并对所述原始数据文件进行格式判断;

所述原始数据文件为JSON文件,对所述原始数据文件进行解析,以得到JSON数据格式的原始数据。

3.根据权利要求2所述的数据处理方法,其特征在于,在所述将写入所述原始数据库的原始数据进行标准化预处理之前,还包括:

获取原始数据写入原始数据库的写入时间或原始数据文件的创建时间;

将所述写入时间或创建时间作为原始数据的批次标识。

4.根据权利要求3所述的数据处理方法,其特征在于,所述将写入所述原始数据库的原始数据进行标准化预处理,包括:

查询所述原始数据库中原始数据对应的批次标识;

对最新批次标识对应的原始数据进行标准化预处理。

5.根据权利要求1所述的数据处理方法,其特征在于,在所述将所述清洗数据写入所述清洗数据库之后,还包括:

获取清洗数据的最近一次推送时间;

将大于所述最近一次推送时间且小于当前系统时间的清洗数据推送至所关联的应用平台中。

6.根据权利要求1所述的数据处理方法,其特征在于,所述将所述原始数据转换为字段标识,具体为:

将所述原始数据转换为对应的哈希值。

7.一种数据处理装置,其特征在于,包括:

第一获取模块,用于获取两个或两个以上的原始数据,所述原始数据包括字段信息和数据内容;

转换提取模块,用于将所述原始数据转换为字段标识,并且从所述字段信息中提取主关键字;

第一写入模块,用于所述原始数据对应的字段标识与原始数据库中的字段标识无交集,则将所述原始数据写入所述原始数据库;

预处理模块,用于将写入所述原始数据库的原始数据进行标准化预处理,得到清洗数据;

第二写入模块,用于所述清洗数据的主关键字与清洗数据库中的主关键字无交集,则

将所述清洗数据写入所述清洗数据库。

8.根据权利要求7所述的数据处理装置，其特征在于，所述装置还包括：

格式判断模块，用于获取原始数据文件，并对所述原始数据文件进行格式判断；

解析模块，用于所述原始数据文件为JSON文件，对所述原始数据文件进行解析，以得到JSON数据格式的原始数据。

9.一种终端设备，其特征在于，包括：存储器以及一个或多个处理器；

所述存储器，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中任一所述的数据处理方法。

10.一种包含计算机可执行指令的存储介质，其特征在于，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中任一所述的数据处理方法。

一种数据处理方法、装置、终端设备和存储介质

技术领域

[0001] 本发明实施例涉及数据处理技术，尤其涉及一种数据处理方法、装置、终端设备和存储介质。

背景技术

[0002] ETL是英文Extract-Transform-Load的缩写，中文名称为数据抽取、转换和加载。ETL是构建数据仓库的重要一环，ETL是用来将从异构数据源抽取出的数据，经过清洗和转换，将数据加载到目的数据仓库(清洗数据库)中，作为联机分析处理、数据挖掘的基础。

[0003] 在ETL处理过程中，当源数据仓库有大批量数据接入时，在常规的ETL工作中，直接将源数据仓库的数据接入，然后通过清洗转换环节，将数据更新插入至目的数据仓库中。但当面对大数据量的数据更新，并且在要求时效性的应用场景时，上述处理方案容易成为数据及时性的瓶颈，从而无法及时地将源数据仓库中的数据更新至目的数据仓库中。

发明内容

[0004] 有鉴于此，本发明提供一种数据处理方法、装置、终端设备和存储介质，以提高数据处理效率。

[0005] 第一方面，本发明实施例提供了一种数据处理方法，包括：

[0006] 获取两个或两个以上的原始数据，所述原始数据包括字段信息和数据内容；

[0007] 将所述原始数据转换为字段标识，并且从所述字段信息中提取主关键字；

[0008] 所述原始数据对应的字段标识与原始数据库中的字段标识无交集，则将所述原始数据写入所述原始数据库；

[0009] 将写入所述原始数据库的原始数据进行标准化预处理，得到清洗数据；

[0010] 所述清洗数据的主关键字与清洗数据库中的主关键字无交集，则将所述清洗数据写入所述清洗数据库。

[0011] 进一步的，在所述获取两个或两个以上的原始数据之前，还包括：

[0012] 获取原始数据文件，并对所述原始数据文件进行格式判断；

[0013] 所述原始数据文件为JSON文件，对所述原始数据文件进行解析，以得到JSON数据格式的原始数据。

[0014] 进一步的，在所述将写入所述原始数据库的原始数据进行标准化预处理之前，还包括：

[0015] 获取原始数据写入原始数据库的写入时间或原始数据文件的创建时间；

[0016] 将所述写入时间或创建时间作为原始数据的批次标识。

[0017] 进一步的，所述将写入所述原始数据库的原始数据进行标准化预处理，包括：

[0018] 查询所述原始数据库中原始数据对应的批次标识；

[0019] 对最新批次标识对应的原始数据进行标准化预处理。

[0020] 进一步的，在所述将所述清洗数据写入所述清洗数据库之后，还包括：

[0021] 获取清洗数据的最近一次推送时间；

[0022] 将大于所述最近一次推送时间且小于当前系统时间的清洗数据推送至所关联的应用平台中。

[0023] 进一步的，将所述原始数据转换为字段标识，具体为：

[0024] 将所述原始数据转换为对应的哈希值。

[0025] 第二方面，本发明实施例还提供了一种数据处理装置，包括：

[0026] 第一获取模块，用于获取两个或两个以上的原始数据，所述原始数据包括字段信息和数据内容；

[0027] 转换提取模块，用于将所述原始数据转换为字段标识，并且从所述字段信息中提取主关键字；

[0028] 第一写入模块，用于所述原始数据对应的字段标识与原始数据库中的字段标识无交集，则将所述原始数据写入所述原始数据库；

[0029] 预处理模块，用于将写入所述原始数据库的原始数据进行标准化预处理，得到清洗数据；

[0030] 第二写入模块，用于所述清洗数据的主关键字与清洗数据库中的主关键字无交集，则将所述清洗数据写入所述清洗数据库。

[0031] 进一步的，所述数据处理装置还包括：

[0032] 格式判断模块，用于在获取两个或两个以上的原始数据之前，获取原始数据文件，并对所述原始数据文件进行格式判断；

[0033] 解析模块，用于所述原始数据文件为JSON文件，对所述原始数据文件进行解析，以得到JSON数据格式的原始数据。

[0034] 进一步的，所述数据处理装置，还包括：

[0035] 第二获取模块，用于在将写入所述原始数据库的原始数据进行标准化预处理之前，获取原始数据写入原始数据库的写入时间或原始数据文件的创建时间；

[0036] 确定模块，用于将所述写入时间或创建时间作为原始数据的批次标识。

[0037] 进一步的，所述预处理模块，包括：

[0038] 查询单元，用于查询所述原始数据库中原始数据对应的批次标识；

[0039] 预处理单元，用于对最新批次标识对应的原始数据进行标准化预处理。

[0040] 进一步的，所述数据处理装置，还包括：

[0041] 第三获取模块，用于在将所述清洗数据写入所述清洗数据库之后，获取清洗数据的最近一次推送时间；

[0042] 推送模块，用于将大于所述最近一次推送时间且小于当前系统时间的清洗数据推送至所关联的应用平台中。

[0043] 进一步的，所述将所述原始数据转换为字段标识，具体用于：

[0044] 将所述原始数据转换为对应的哈希值。

[0045] 第三方面，本发明实施例还提供了一种终端设备，包括：存储器以及一个或多个处理器；

[0046] 所述存储器，用于存储一个或多个程序；

[0047] 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

器实现如第一方面所述的数据处理方法。

[0048] 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如第一方面所述的数据处理方法。

[0049] 本发明通过获取两个或两个以上的包含有字段信息和数据内容的原始数据；将原始数据转换为字段标识，并且从字段信息中提取主关键字；若原始数据对应的字段标识与原始数据库中的字段标识无交集，将原始数据写入原始数据库；将写入原始数据库的原始数据进行标准化预处理，得到清洗数据；若清洗数据的主关键字与清洗数据库中的主关键字无交集，将清洗数据写入清洗数据库，在数据处理过程中，无需将原始数据库中所有数据对应的主关键字和清洗数据库中的主关键字进行比对，从而提高了数据处理效率。

附图说明

[0050] 图1是本发明实施例一提供的一种数据处理方法的流程图；

[0051] 图2是本发明实施例一提供的一种原始数据写入原始数据库的显示示意图；

[0052] 图3是本发明实施例一提供的一种清洗数据写入清洗数据库的显示示意图；

[0053] 图4是本发明实施例二提供的一种数据处理方法的流程图；

[0054] 图5是本发明实施例二提供的一种数据处理过程的示意图；

[0055] 图6是本发明实施例二提供的一种确定批次标识的显示示意图；

[0056] 图7是本发明实施例二提供的一种数据处理系统的结构框图；

[0057] 图8是本发明实施例三提供的一种数据处理方法的流程图；

[0058] 图9是本发明实施例三提供的一种数据处理的组件连接示意图；

[0059] 图10是本发明实施例四提供的一种数据处理装置的结构框图；

[0060] 图11是本发明实施例五提供的一种终端设备的结构示意图。

具体实施方式

[0061] 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。

[0062] 在此需要说明的是，本方案中数据处理方法的所有实施例，是通过水壶工具集Kettle中组件来实现的ETL数据处理过程。其中，Kettle是一款ETL工具，允许对来自不同数据库的数据进行管理，并通过提供一个图形化的用户环境来描述所要进行的数据处理过程。

[0063] 实施例一

[0064] 图1是本发明实施例一提供的一种数据处理方法的流程图，本实施例中提供的数据处理方法可以由终端设备执行，该终端设备可以通过软件和/或硬件的方式实现，该终端设备可以是两个或多个物理实体构成，也可以是一个物理实体构成。本实施例中终端设备为服务器，用来对原始数据进行处理，为服务器所关联的应用平台提供性能支撑。

[0065] 参考图1，该数据处理方法具体包括如下步骤：

[0066] S110、获取两个或两个以上的原始数据。

[0067] 其中，原始数据包括字段信息和数据内容。

[0068] 在实施例中，原始数据可以理解为从异构数据源获取的不同数据信息。其中，异构数据源指的是不同的数据库管理系统之间的数据。需要理解的是，服务器在同一时间可从不同的数据源获取原始数据，在原始数据中包含有多个数据表，并且在原始数据的每个数据表中均包含有字段信息和数据内容，其中，字段信息可以理解为每个数据表中的各个字段名，而数据内容可以理解为数据表中各个字段名中对应的数据信息。可以理解为，每个字段名均对应有不同的数据内容。其中，原始数据是未经过字段筛选、格式转换等标准化预处理的数据信息。

[0069] S120、将原始数据转换为字段标识，并且从字段信息中提取主关键字。

[0070] 其中，字段标识用来表示是否对原始数据所在文件的数据内容进行修改，若发生修改，则字段标识发生变化；反之，若没有发生修改，则字段标识也不发生变化。示例性地，字段标识可以为哈希值，也可为MD5消息摘要算法(MD5 Message Digest Algorithm)值。其中，哈希值也是一种散列函数，用于把消息或数据压缩成摘要，使得数据量变小，将数据的格式固定下来；MD5是一种被广泛使用的密码散列函数，可以产生一个128位的散列值，用于确保信息传输完整一致。在实施例中，将原始数据转换为字段标识，是为了检测原始数据所在文件的数据内容是否发生修改，而不需对原始数据中的数据内容进行一一比对分析，从而加快了原始数据的检测速度。

[0071] 在此需要说明的是，在原始数据中存在有多个数据表，在每个数据表中都有多条记录，为了唯一地标识数据表中的某一条记录，可从数据表中提取出一个或多个字段作为该数据表的主关键字，从而在查找某个数据表时，通过主关键字可加快数据库的操作速度。

[0072] S130、原始数据对应的字段标识与原始数据库中的字段标识无交集，则将原始数据写入原始数据库。

[0073] 其中，原始数据库可以理解为用来存储原始数据的数据库。在实施例中，每个原始数据所在的文件在原始数据库中都保存有唯一的字段标识。若原始数据对应的字段标识与原始数据库中的字段标识相同，则表明该原始数据中的数据内容未发生修改，即该原始数据中的数据内容和原始数据库中的数据内容是重复的，则不需将该原始数据写入原始数据库中；反之，若原始数据对应的字段标识与原始数据库中的字段标识没有交集，即原始数据中的数据内容和原始数据库中的数据内容是不同的，则将该原始数据写入原始数据库中，以为服务器所关联的应用平台提供数据性能支撑。图2是本发明实施例一提供的一种原始数据写入原始数据库的显示示意图。参考图2，假设获取到四个原始数据，则将每个原始数据转换为对应的字段标识，分别为字段标识1、字段标识2、字段标识3和字段标识4，然后对四个原始数据的字段标识与原始数据库中的字段标识进行比对分析，发现字段标识1和字段标识3已在原始数据库中存在，则将字段标识1和字段标识3分别对应的原始数据删除，而只将字段标识1和字段标识4对应的原始数据写入原始数据库中。

[0074] 需要注意的是，在本方案中，采用原始数据对应的字段标识和原始数据库中的字段标识进行比对分析，以确定是否将原始数据写入原始数据库中，而并非采用原始数据对应的主关键字，是为了便于追溯原始数据。可以理解为，当需要追溯原始数据，以查找到对应的历史版本记录时，若采用原始数据的主关键字和原始数据库中的主关键字进行比对，当原始数据的主关键字未发生变化，而其它字段对应的数据内容发生变化时，不将该字段对应的数据内容写入原始数据库中，从而造成了原始数据的遗漏，无法查找到对应的历史

版本记录;而若采用本方案中原始数据对应的字段标识和原始数据库中的字段标识进行比对分析,若原始数据对应的任一数据内容发生修改时,则该字段标识也发生变化,从而可将该原始数据写入原始数据库中,保证了原始数据的完整性,并可查找到对应的历史版本记录和原始数据。

[0075] S140、将写入原始数据库的原始数据进行标准化预处理,得到清洗数据。

[0076] 其中,标准化预处理包括字段筛选、格式转换和数据校验等一系列的过程。在实施例中,在将原始数据写入原始数据库之后,对原始数据中的字段信息进行筛选,然后将筛选后得到的字段信息对应的数据内容转换为预先设定的格式,并将格式转换之后的数据内容进行数据校验,以得到清洗数据。其中,字段筛选可以理解为将原始数据中不符合原始数据库预先设定的字段信息筛选出来,以得到筛选后的原始数据。比如,原始数据中包含有字段D,但原始数据库中预先设定的字段信息中未设置有字段D,则在将原始数据中的其它数据内容写入原始数据库中时,对字段D对应的数据内容进行筛选处理。然后对筛选处理之后的原始数据进行格式转换,以转换为预先设定格式的数据信息。比如,预先设定的数据长度阈值为300,数据行数阈值为100;若原始数据的数据长度为200,数据行数为200,则需对原始数据进行拆分处理,以得到小于或等于数据长度阈值和数据行数阈值的原始数据;然后对格式转换之后的原始数据进行数据校验,以保证数据的合法性。具体地,对格式转换之后的原始数据中的字符串进行校验,以筛选出格式转换之后的原始数据中的非法字符串,最后得到清洗数据。

[0077] S150、清洗数据的主关键字与清洗数据库中的主关键字无交集,则将清洗数据写入清洗数据库。

[0078] 其中,清洗数据库可以理解为用来存储清洗数据的数据仓库,也可理解为清洗数据库为目的数据仓库。在实施例中,为了提高数据写入效率,可直接对清洗数据的主关键字和清洗数据库中的主关键字进行比对分析,若清洗数据的主关键字和清洗数据库中的主关键字没有重复,表明当前的清洗数据库中未存储有步骤S140得到的清洗数据,需将该清洗数据写入清洗数据库中,以对清洗数据库中的数据信息进行更新处理;反之,若清洗数据的主关键字和清洗数据库中的主关键字重复,表明当前的清洗数据库中存储有步骤S140得到的清洗数据,可直接对该清洗数据进行筛选处理。

[0079] 图3是本发明实施例一提供的一种清洗数据写入清洗数据库的显示示意图。图3是在图2的基础上对清洗数据写入清洗数据库的过程进行说明。参考图3,在将字段标识2和字段标识4分别对应的原始数据写入原始数据库之后,对原始数据进行标准化预处理,以得到清洗数据,然后将查找到对应的主关键字,分别为主关键字2和主关键字4,将这两个主关键字和清洗数据库中的主关键字进行比对分析,发现主关键字4已在清洗数据库中存在,则将主关键字4对应的清洗数据删除,只将主关键字2对应的清洗数据写入清洗数据库中。其中,字段标识2和主关键字2所对应的原始数据是相同的;字段标识4和主关键字4所对应的原始数据也是相同的。

[0080] 本实施例的技术方案,通过获取两个或两个以上的包含有字段信息和数据内容的原始数据;将原始数据转换为字段标识,并且从字段信息中提取主关键字;若原始数据对应的字段标识与原始数据库中的字段标识无交集,将原始数据写入原始数据库;将写入原始数据库的原始数据进行标准化预处理,得到清洗数据;若清洗数据的主关键字与清洗数据

库中的主关键字无交集，将清洗数据写入清洗数据库，在数据处理过程中，无需将原始数据库中所有数据对应的主关键字和清洗数据库中的主关键字进行比对，从而提高了数据处理效率。

[0081] 实施例二

[0082] 图4是本发明实施例二提供的一种数据处理方法的流程图。本实施例是在上述实施例的基础上，对数据处理方法作进一步的具体化。图5是本发明实施例二提供的一种数据处理过程的示意图。在此需要说明的是，为了便于对数据处理过程进行说明。在实施例中，只获取5个原始数据文件，分别为原始数据文件1、原始数据文件2、原始数据文件3、原始数据文件4和原始数据文件5，如图5所示。

[0083] 参照图4，该数据处理方法具体包括如下步骤：

[0084] S201、获取原始数据文件，并对原始数据文件进行格式判断。

[0085] 其中，原始数据文件可以理解为存储原始数据的文件。在实施例中，通过ETL中的Kettle获取与预先设定数据格式一致的原始数据。可以理解为，在获取到原始数据文件之后，对原始数据文件中原始数据的数据格式进行初步筛选，以得到符合预先设定数据格式的原始数据。其中，原始数据文件中原始数据的数据格式可为EXCEL、JSON、文本等数据格式，对此并不进行限定。可以理解为，从异构数据源获取到存储有原始数据的原始数据文件之后，对原始数据文件中的原始数据进行格式判断，以识别出符合预先设定数据格式的原始数据，以及原始数据文件。需要说明的是，原始数据文件的格式与原始数据的数据格式是相同的。比如，原始数据文件的格式一般为普通文件夹的格式，但该原始数据文件中存储有原始数据的各个文档的格式与原始数据的数据格式是相同的。可以理解为，每个原始数据文件中可包含有多个存储有原始数据的文档。比如，原始数据文件中包含有三个文档，并且三个文档中原始数据的数据格式均为EXCEL格式，则这三个文档的格式就为EXCEL格式。

[0086] S202、原始数据文件为JSON文件，对原始数据文件进行解析，以得到JSON数据格式的原始数据。

[0087] 在实施例中，原始数据文件中原始数据的预先设定数据格式为JSON数据格式。可以理解为，在得到原始数据文件之后，对原始数据文件进行格式判断，若原始数据文件为JSON格式的文件，则对该原始数据文件进行解析，以得到JSON数据格式的原始数据。当然，也可将原始数据的数据格式设定为其它数据格式，其可根据业务需求进行设定。如图5所示，对原始数据文件进行解析，以得到JSON数据格式的原始数据，分别为原始数据1、原始数据2、原始数据3、原始数据4和原始数据5。

[0088] S203、获取两个或两个以上的原始数据。

[0089] 其中，原始数据包括字段信息和数据内容。在实施例中，由于原始数据文件有5个，则所获取的原始数据对应的也有5个。

[0090] S204、将原始数据转换为字段标识，并且从字段信息中提取主关键字。

[0091] 在实施例中，将每个原始数据转换为对应的字段标识，该字段标识可为哈希值，也可为MD5值。同时每个原始数据的字段信息中提取出对应的主关键字。

[0092] S205、原始数据对应的字段标识与原始数据库中的字段标识无交集，则将原始数据写入原始数据库。

[0093] 具体来说，判断原始数据对应的字段标识和原始数据库中的字段标识是否有交

集，如图5所示，字段标识2和原始数据库中的字段标识有交集，则删除字段标识2所对应的原始数据，以实现原始数据库中的数据去重；而将字段标识1、字段标识3、字段标识4和字段标识5对应的原始数据写入原始数据库中。

[0094] S206、获取原始数据写入原始数据库的写入时间或原始数据文件的创建时间。

[0095] 其中，写入时间为将原始数据写入原始数据库时所处的系统时间；创建时间为对原始数据进行组装并形成原始数据文件时所处的系统时间。在实施例中，原始数据文件的创建时间早于原始数据写入原始数据库的写入时间。可以理解为，在从异构数据源获取到原始数据文件时，已经完成对原始数据文件的创建，为了便于统计原始数据文件的创建时间，可将从异构数据源中获取原始数据文件所处的系统时间作为原始数据文件的创建时间。在获取到原始数据文件之后，对原始数据文件进行字段标识的转换，以及字段标识的对比分析，然后将符合去重过滤规则的原始数据写入原始数据库中，将该写入原始数据库时所处的系统时间作为写入时间。其中，去重过滤规则可以理解为根据原始数据对应的字段标识与原始数据库中的字段标识进行对比分析，以过滤原始数据的规则。

[0096] S207、将写入时间或创建时间作为原始数据的批次标识。

[0097] 其中，批次标识用来标识原始数据在原始数据库中更新的前后顺序。在实施例中，为了可从原始数据库中尽快查找到最新的原始数据，对每次写入原始数据库中的原始数据设置一个对应的批次标识，可直接采用写入原始数据库的写入时间作为原始数据在原始数据库中的批次标识，也可采用原始数据文件的创建时间作为原始数据在原始数据库中的批次标识。为了更直观地根据批次标识确定原始数据在原始数据库中更新的前后顺序，批次标识可直接采用数值化的时间来设定。比如，假设原始数据文件的创建时间为2018年11月9日下午16点28分，则对应的批次标识为201811091628，又如，假设原始数据写入原始数据库中的写入时间为2018年11月9日下午18点6分，则对应的批次标识为201811091806。

[0098] 图6是本发明实施例二提供的一种确定批次标识的显示示意图。假设以1分钟为间隔，在总共15分钟之内的数据统计量，如图6所示，原始数据文件5的创建时间为第两分钟；原始数据文件1的创建时间为第七分钟的第6秒；原始数据文件3的创建时间为第九分钟的第30秒；原始数据文件4的创建时间为第12分钟的第18秒，如图6所示的实线箭头；而这四个原始数据文件写入原始数据库的写入时间均为第十五分钟的第30秒，如图6所示的虚线箭头。

[0099] 当然，为了便于对原始数据的批次标识进行统一管理，必须对批次标识所采用的时间进行固定设定，可以理解为，若采用原始数据写入原始数据库的写入时间作为批次标识，需将所有原始数据的批次标识均以写入时间进行统计；同样地，若采用原始数据文件的创建时间作为批次标识，则需将所有原始数据的批次标识均以创建时间进行统计，不能将写入时间和创建时间进行混合统计。

[0100] S208、查询原始数据库中原始数据对应的批次标识。

[0101] 在实施例中，在将原始数据写入原始数据库之后，会将该原始数据对应的批次标识也写入原始数据库中，以便于后续根据批次标识可尽快从原始数据库中查找到对应的原始数据。其中，可采用数据查询语句从原始数据库中查找批次标识，比如，数据查询语句可采用结构化查询语言(Structured Query Language，SQL)、Oracle等数据库中的查询语句，当然，对此并不进行限定，可根据业务需求进行选择。在此需要说明的是，为了便于对批次

标识的查询，可将批次标识写入一个预先创建的临时数据表，并将该临时数据表存入原始数据库中。当然，在该临时数据表中写有原始数据和批次标识之间的关系，可直接通过批次标识得到原始数据库中对应的原始数据。

[0102] S209、对最新批次标识对应的原始数据进行标准化预处理，得到清洗数据。

[0103] 在实施例中，为了提高对原始数据库中的原始数据进行标准化处理的速度，只需对原始数据库中最新批次标识对应的原始数据进行标准化处理。可以理解为，通过数据查询语句获取当前原始数据库中最新的批次标识，并获取该最新批次标识对应的原始数据，然后通过字段筛选、格式转换和数据校验等一系列的标准化预处理之后，即可得到清洗过滤后的清洗数据。其中，对原始数据进行标准化预处理的具体过程可参见上述实施例的描述，在此不再赘述。如图5所示，字段标识1、字段标识3、字段标识4和字段标识5所对应的原始数据，其所对应的批次标识是原始数据库中最新的，则将这四个字段标识对应的原始数据均进行标准化预处理，其得到对应的清洗数据1、清洗数据3、清洗数据4和清洗数据5。

[0104] 在此需要说明的是，本实施例中采用原始数据写入原始数据库的写入时间作为原始数据的批次标识，则字段标识1、字段标识3、字段标识4和字段标识5所对应的批次标识是相同的。

[0105] S210、清洗数据的主关键字与清洗数据库中的主关键字无交集，则将清洗数据写入清洗数据库。

[0106] 如图5所示，清洗数据3所对应的主关键字3和清洗数据库中的主关键字重复，则删除清洗数据3，而只将清洗数据1、清洗数据4和清洗数据5写入清洗数据库中。

[0107] 本实施例的技术方案，在上述实施例的基础上，通过对原始数据文件进行格式判断，以得到JSON数据格式的原始数据，同时，对原始数据库中最新批次标识对应的原始数据进行标准化预处理，以得到清洗数据，并在清洗数据的主关键字与清洗数据库中的主关键字无交集时，将清洗数据写入清洗数据库中，实现了只对预先设定格式的原始数据进行获取，以及只对最新批次标识对应的原始数据进行标准化预处理，简化了数据处理过程，从而提高了数据处理速度。

[0108] 在上述实施例的基础上，为了对服务器所关联的应用平台上的数据进行及时更新，在步骤S210之后，还包括：

[0109] S211、获取清洗数据的最近一次推送时间。

[0110] 其中，最近一次推送时间可以理解为最近一次将清洗数据库中的清洗数据推送至服务器所关联的应用平台的时间。在实施例中，可直接采用数据查询语句对最近一次推送时间进行查询获取。具体地，在最近一次将清洗数据发送至所关联的应用平台之后，将最近一次推送时间进行统计并存储至预先设定的时间临时数据表中，以在后续调取使用。当然，在对最近一次推送时间进行获取时，可直接通过SQL、Oracle等数据库中的查询语句进行查询得到。

[0111] S212、将大于所述最近一次推送时间且小于当前系统时间的清洗数据推送至所关联的应用平台中。

[0112] 在实施例中，在服务器所关联的应用平台处于开启使用状态时，为了及时将清洗数据库中更新的清洗数据发送至服务器所关联的应用平台，可获取清洗数据的最近一次推送时间，并获取得到当前系统时间，以获取得到大于最近一次推送时间且小于当前系统时

间的所有清洗数据，然后将该所有清洗数据通过数据通信方式推送至所关联的应用平台中。其中，数据通信方式可采用无线网络、有线网络等方式，对此并不进行限定。其中，应用平台可为客户端中所安装的应用程序，其中，客户端可为台式机、笔记本电脑、智能手机等设备。

[0113] 图7是本发明实施例二提供的一种数据处理系统的结构框图。如图7所示，该数据处理系统包括：服务器310和应用平台320；其中，服务器310用于获取原始数据，并对原始数据进行处理，以得到清洗数据；应用平台320可为台式机、智能手机、笔记本电脑。在服务器310将清洗数据推送至应用平台320上，应用平台320根据清洗数据对自身数据库中的数据进行更新。

[0114] 当然，在数据处理过程中，可在数据处理的关键环节添加成功与错误判断的流程。比如，可从将原始数据写入原始数据库开始，到将清洗数据推送至所关联的应用平台结束，均可作为数据处理过程中的关键环节，当检测到数据处理过程出现错误时，直接通过Kettle中的邮件组件将错误信息发送至相关开发人员，以使开发人员对数据处理过程进行实时监控。

[0115] 实施例三

[0116] 图8是本发明实施例三提供的一种数据处理方法的流程图。本实施例是在上述实施例的基础上，以Kettle中的各个组件对数据处理过程进行说明。参考图8，该数据处理方法具体如下步骤：

[0117] S410、设置开始时间。

[0118] 其中，在Kettle的组件中包含有作业组件，并且每个作业组件可包含多个流程，同时每个流程可以实现并行操作。同时，在每个流程中可以包含多个组件，实现串行操作。图9是本发明实施例三提供的一种数据处理的组件连接示意图。在实施例中，该数据处理过程就可以认为是一个流程，并且在该流程中包含有如图9所示的多个组件，每个组件可执行不同的数据处理步骤。如图9所示，组件510为一个开始组件，用来设置任务开始的时间策略，比如，定时或时间间隔等。其中，定时可以理解为在设定的时刻开始一个任务；而时间间隔可以理解为隔一段时间开始一个任务。其中，一个任务可以理解为一个数据处理过程。

[0119] S420、对原始数据文件的个数进行统计，并判断是否为0，若为0，则执行步骤S470；若不为0，则执行步骤S430。

[0120] 在实施例中，在接收到开始组件的点击触发之后，获取原始数据文件，并对原始数据文件进行格式判断，若原始数据文件为JSON文件，对原始数据文件的个数进行统计，若原始数据文件不为0，则执行步骤S430；若原始数据文件为0，则执行步骤S470。其中，步骤S420的具体流程可通过如图9所示的组件520进行实现。

[0121] S430、将原始数据写入原始数据库中。

[0122] 在实施例中，在对原始数据文件进行解析，以得到JSON数据格式的原始数据，然后获取原始数据中的字段信息和数据内容，并将原始数据转换为字段标识，并从字段信息中提取主关键字，判断原始数据对应的字段标识与原始数据库中的字段标识是否有交集，若没有交集，则表示原始数据库中没有该原始数据，将该原始数据写入原始数据库中。其中，该步骤可通过如图9所示的组件530实现，组件530为原始数据库。

[0123] 在此需要说明的是，在将原始数据文件中的原始数据写入原始数据库中过程，可

参见现有技术中通过Kettle中的组件将JSON文件插入数据库中的过程。具体可包括:JSON解析、获取变量、字段选择、替换NULL值、增加校验列、获取系统信息等步骤。其中,JSON解析是对原始数据文件进行解析,以得到JSON数据格式的原始数据;然后获取变量组件中所设置的变量值,并通过字段选择,对所需要的字段进行重命名和筛选;并将原始数据中的空值替换成NULL字符串,以便于进行组合主关键字的设置;然后确定哈希值的字段信息,并获取系统信息,比如,系统时间信息等。

[0124] S440、将清洗数据写入清洗数据库中。

[0125] 在实施例中,在将原始数据写入原始数据库之后,查询原始数据对应的批次标识,对最新批次标识对应的原始数据进行字段筛选、格式转换和数据校验等一系列的标准化预处理,以得到清洗数据,然后将清洗数据的主关键字与清洗数据库中的主关键字进行比对分析,若清洗数据的主关键字不存在于清洗数据库中的主关键字,则将清洗数据写入清洗数据库中。其中,如图9所示的组件540为清洗数据库。可以理解为,该步骤是在从组件530中获取最新批次标识对应的原始数据,到写入组件540清洗数据库的过程中实现的。

[0126] S450、对清洗数据库中新增或更新的清洗数据的数据量进行统计,并判断是否为0。

[0127] 在实施例中,对清洗数据库中的数据进行查询,以确定清洗数据库中新增或更新的清洗数据的数据量,新增或更新的清洗数据的数据量为0,则执行步骤S470;若新增或更新的清洗数据的数据量不为0,则执行步骤S460。其中,该步骤是通过如图9所示的组件550来实现的。

[0128] S460、将新增或更新的清洗数据推送至所关联的应用平台。

[0129] 在实施例中,在清洗数据库中的清洗数据更新或新增之后,在服务器所关联的应用平台处于开启状态时,自动将新增或更新的清洗数据推送至所关联的应用平台,以更新应用平台中的数据信息。其中,该步骤是通过如图9所示的组件560来实现的。

[0130] S470、退出作业流程。

[0131] 在实施例中,若原始数据文件的个数为0,则表明没有获取到新的原始数据,则直接退出作业流程。同时,在清洗数据库中新增或更新的清洗数据的数据量为0时,表明清洗数据库中没有新增或更新的清洗数据,则直接退出作业流程。其中,该步骤是通过如图9所示的组件570来实现的。

[0132] S480、将错误提示信息发送至相关开发人员。

[0133] 在实施例中,为了保证开发人员能及时了解到数据处理过程中的错误位置,在数据处理过程的关键环节添加成功与错误判断,比如,从步骤420-步骤S460,中添加成功与错误判断,当数据处理过程中出现错误,则将错误提示信息发送至相关开发人员。

[0134] 本实施例的技术方案,在数据处理过程中,无需将原始数据库中所有数据对应的主关键字和清洗数据库中的主关键字进行比对,提高了数据处理效率。

[0135] 实施例四

[0136] 图10是本发明实施例四提供的一种数据处理装置的结构框图。本实施例的数据处理装置可配置于服务器中,参考图10,该数据处理装置包括:第一获取模块610、转换提取模块620、第一写入模块630、预处理模块640和第二写入模块650。

[0137] 其中,第一获取模块610,用于获取两个或两个以上的原始数据,该原始数据包括

字段信息和数据内容；

[0138] 转换提取模块620，用于将原始数据转换为字段标识，并且从字段信息中提取主关键字；

[0139] 第一写入模块630，用于原始数据对应的字段标识与原始数据库中的字段标识无交集，则将原始数据写入原始数据库；

[0140] 预处理模块640，用于将写入原始数据库的原始数据进行标准化预处理，得到清洗数据；

[0141] 第二写入模块650，用于清洗数据的主关键字与清洗数据库中的主关键字无交集，则将清洗数据写入清洗数据库。

[0142] 本实施例提供的技术方案，通过获取两个或两个以上的包含有字段信息和数据内容的原始数据；将原始数据转换为字段标识，并且从字段信息中提取主关键字；若原始数据对应的字段标识与原始数据库中的字段标识无交集，将原始数据写入原始数据库；将写入原始数据库的原始数据进行标准化预处理，得到清洗数据；若清洗数据的主关键字与清洗数据库中的主关键字无交集，将清洗数据写入清洗数据库，在数据处理过程中，无需将原始数据库中所有数据对应的主关键字和清洗数据库中的主关键字进行比对，从而提高了数据处理效率

[0143] 在上述实施例的基础上，该数据处理装置还包括：

[0144] 格式判断模块，用于在获取两个或两个以上的原始数据之前，获取原始数据文件，并对原始数据文件进行格式判断；

[0145] 解析模块，用于原始数据文件为JSON文件，对原始数据文件进行解析，以得到JSON数据格式的原始数据。

[0146] 在上述实施例的基础上，该数据处理装置，还包括：

[0147] 第二获取模块，用于在将写入原始数据库的原始数据进行标准化预处理之前，获取原始数据写入原始数据库的写入时间或原始数据文件的创建时间；

[0148] 确定模块，用于将写入时间或创建时间作为原始数据的批次标识。

[0149] 在上述实施例的基础上，该预处理模块640，包括：

[0150] 查询单元，用于查询原始数据库中原始数据对应的批次标识；

[0151] 预处理单元，用于对最新批次标识对应的原始数据进行标准化预处理。

[0152] 在上述实施例的基础上，该数据处理装置，还包括：

[0153] 第三获取模块，用于在将清洗数据写入清洗数据库之后，获取清洗数据的最近一次推送时间；

[0154] 推送模块，用于将大于最近一次推送时间且小于当前系统时间的清洗数据推送至所关联的应用平台中。

[0155] 在上述实施例的基础上，将所述原始数据转换为字段标识，具体用于：

[0156] 将原始数据转换为对应的哈希值。

[0157] 上述数据处理装置可执行本发明任意实施例所提供的数据处理方法，具备执行方法相应的功能模块和有益效果。

[0158] 实施例五

[0159] 图11是本发明实施例五提供的一种终端设备的结构示意图。参考图11，该终端设

备包括:处理器710、存储器720、输入装置730以及输出装置740。该终端设备中处理器710的数量可以是一个或者多个,图11中以一个处理器710为例。该终端设备中存储器720的数量可以是一个或者多个,图11中以一个存储器720为例。该终端设备的处理器710、存储器720、输入装置730以及输出装置740可以通过总线或者其他方式连接,图11中以通过总线连接为例。实施例中,该终端设备可以为服务器。

[0160] 存储器720作为一种计算机可读存储介质,可用于存储软件程序、计算机可执行程序以及模块,如本发明任意实施例所述的设备对应的程序指令/模块(例如,数据处理装置中的第一获取模块610、转换提取模块620、第一写入模块630、预处理模块640和第二写入模块650)。存储器720可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序;存储数据区可存储根据设备的使用所创建的数据等。此外,存储器720可以包括高速随机存取存储器,还可以包括非易失性存储器,例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中,存储器720可进一步包括相对于处理器710远程设置的存储器,这些远程存储器可以通过网络连接至设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0161] 输入装置730可用于接收输入的数字或者字符信息,以及产生与设备的用户设置以及功能控制有关的键信号输入。输出装置740可以包括扬声器等音频设备。需要说明的是,输入装置730和输出装置740的具体组成可以根据实际情况设定。

[0162] 处理器710通过运行存储在存储器720中的软件程序、指令以及模块,从而执行设备的各种功能应用以及数据处理,即实现上述的数据处理方法。

[0163] 上述提供的终端设备可用于执行上述任意实施例提供的数据处理方法,具备相应的功能和有益效果。

[0164] 实施例六

[0165] 本发明实施例六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一种数据处理方法,包括:

[0166] 获取两个或两个以上的原始数据,该原始数据包括字段信息和数据内容;

[0167] 将原始数据转换为字段标识,并且从字段信息中提取主关键字;

[0168] 原始数据对应的字段标识与原始数据库中的字段标识无交集,则将原始数据写入原始数据库;

[0169] 将写入原始数据库的原始数据进行标准化预处理,得到清洗数据;

[0170] 清洗数据的主关键字与清洗数据库中的主关键字无交集,则将清洗数据写入清洗数据库。

[0171] 当然,本发明实施例所提供的一种包含计算机可执行指令的存储介质,其计算机可执行指令不限于如上所述的数据处理方法操作,还可以执行本发明任意实施例所提供的数据处理方法中的相关操作,且具备相应的功能和有益效果。

[0172] 通过以上关于实施方式的描述,所属领域的技术人员可以清楚地了解到,本发明可借助软件及必需的通用硬件来实现,当然也可以通过硬件实现,但很多情况下前者是更佳的实施方式。基于这样的理解,本发明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计算机可读存储介质中,如计算机的软盘、只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random

Access Memory,RAM)、闪存(FLASH)、硬盘或光盘等,包括若干指令用以使得一台计算机设备(可以是机器人,个人计算机,服务器,或者网络设备等)执行本发明任意实施例所述的数据处理方法。

[0173] 值得注意的是,上述数据处理装置中,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的保护范围。

[0174] 应当理解,本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中,多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如,如果用硬件来实现,和在另一实施方式中一样,可用本领域公知的下列技术中的任一项或他们的组合来实现:具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路,具有合适的组合逻辑门电路的专用集成电路,可编程门阵列(PGA),现场可编程门阵列(FPGA)等。

[0175] 在本说明书的描述中,参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中,对上述术语的示意性表述不一定指的是相同的实施例或示例。而且,描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

[0176] 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。


获取两个或两个以上的原始数据
S110
将原始数据转换为字段标识，并且从字段信息中提取主关键字
S120
原始数据对应的字段标识与原始数据库中的字段标识无交集，则将原始数据写入原始数据库
S130
将写入原始数据库的原始数据进行标准化预处理，得到清洗数据
S140
清洗数据的主关键字与清洗数据库中的主关键字无交集，则将清洗数据写入清洗数据库
S150

图1

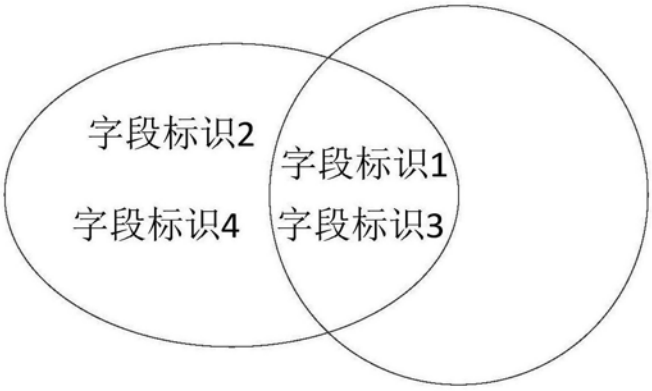
字段标识2
字段标识1
字段标识4
字段标识3

图2

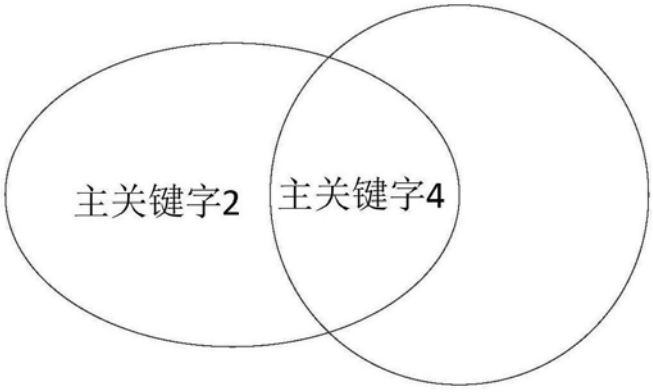

图3

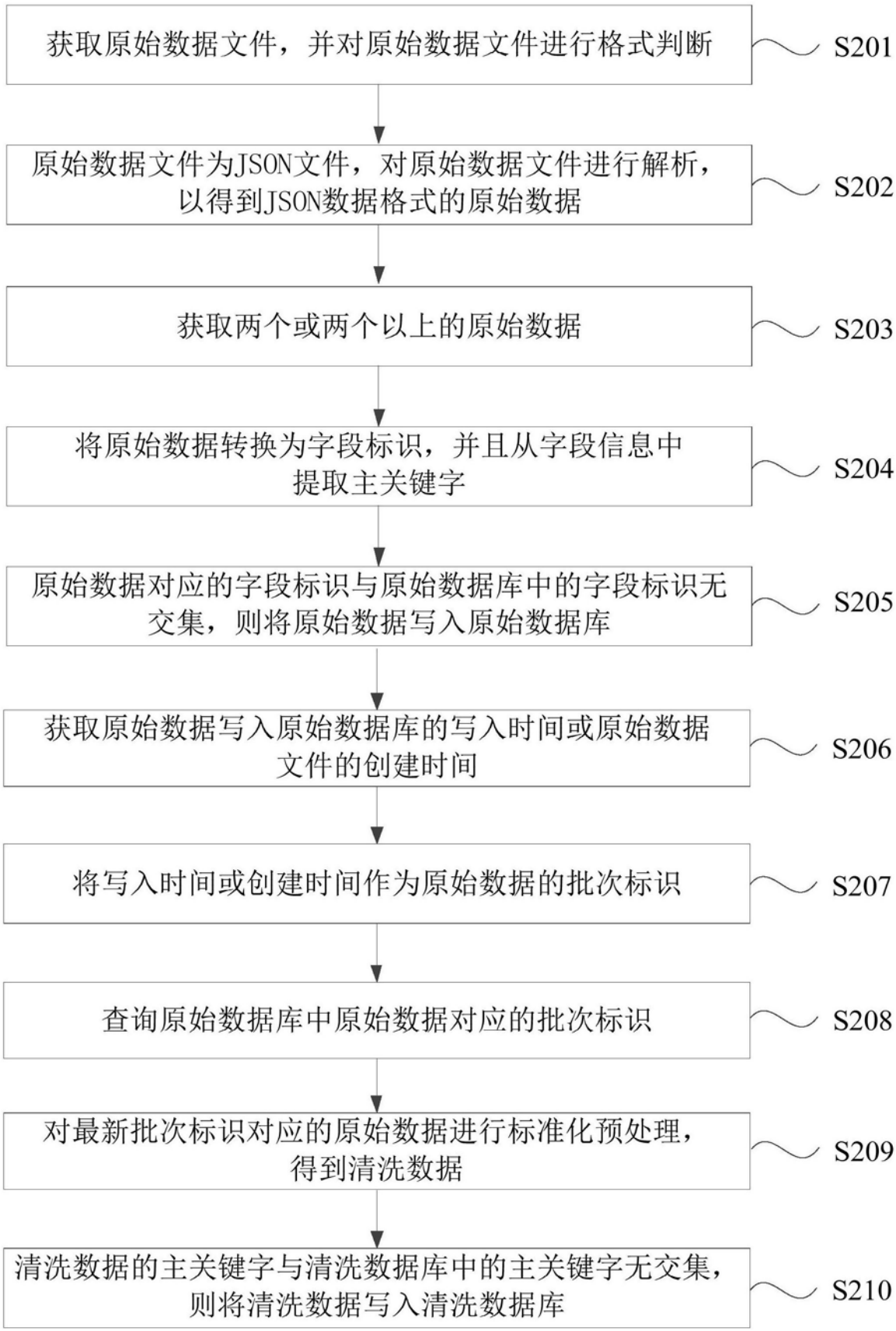
获取原始数据文件，并对原始数据文件进行格式判断
S201
原始数据文件为JSON文件，对原始数据文件进行解析，以得到JSON数据格式的原始数据
S202
获取两个或两个以上的原始数据
S203
将原始数据转换为字段标识，并且从字段信息中提取主关键字
S204
原始数据对应的字段标识与原始数据库中的字段标识无交集，则将原始数据写入原始数据库
S205
获取原始数据写入原始数据库的写入时间或原始数据文件的创建时间
S206
将写入时间或创建时间作为原始数据的批次标识
S207
查询原始数据库中原始数据对应的批次标识
S208
对最新批次标识对应的原始数据进行标准化预处理，得到清洗数据
S209
清洗数据的主关键字与清洗数据库中的主关键字无交集，则将清洗数据写入清洗数据库
S210

图4

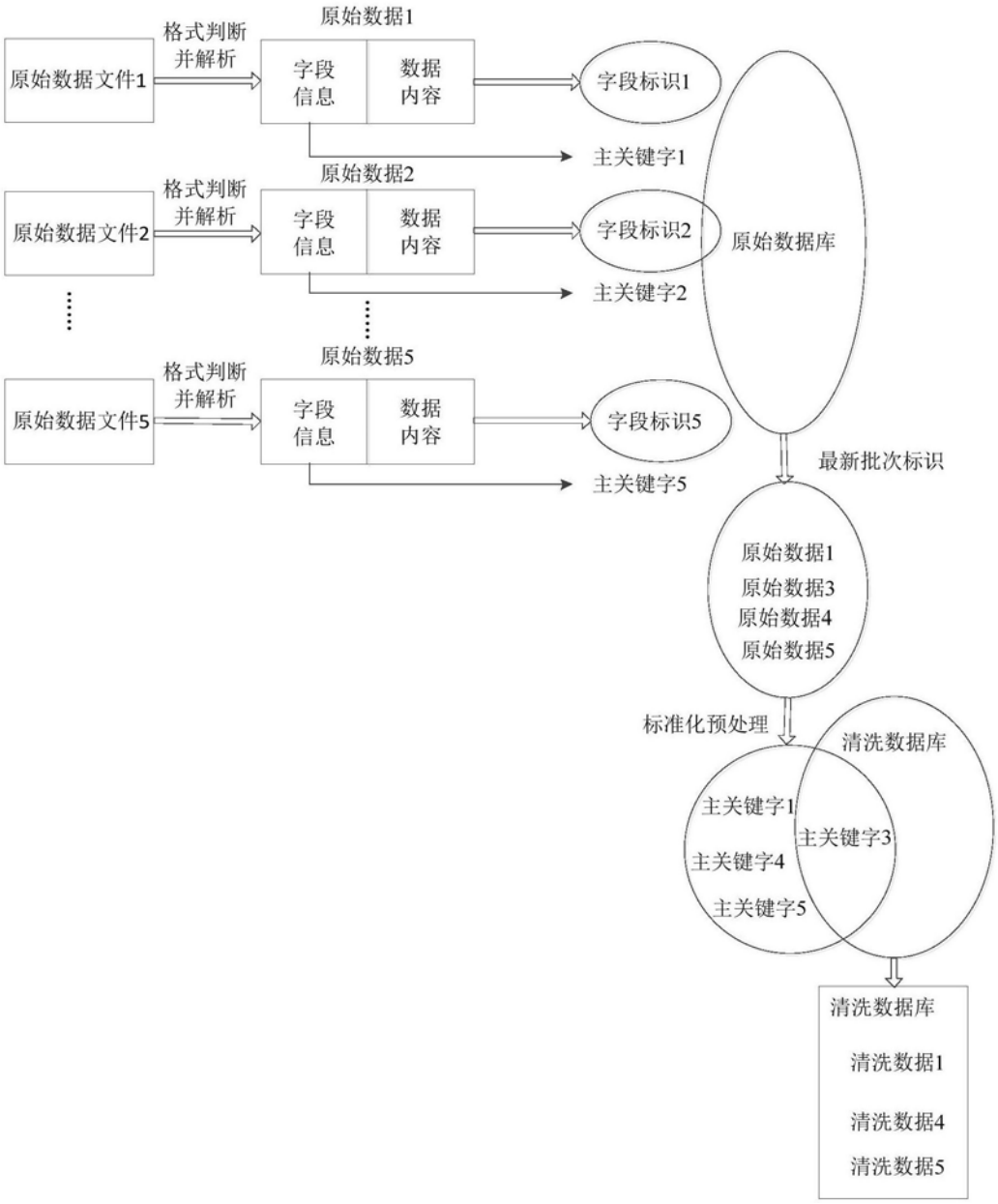
原始数据1
格式判断
并解析
原始数据文件1
字段
信息
数据
内容
字段标识1
主关键字1
原始数据2
格式判断
并解析
原始数据文件2
字段
信息
数据
内容
字段标识2
主关键字2
原始数据5
格式判断
并解析
原始数据文件5
字段
信息
数据
内容
字段标识5
主关键字5
原始数据库
最新批次标识
原始数据1
原始数据3
原始数据4
原始数据5
标准化预处理
清洗数据库
主关键字1
主关键字3
主关键字4
主关键字5
清洗数据库
清洗数据1
清洗数据4
清洗数据5

图5

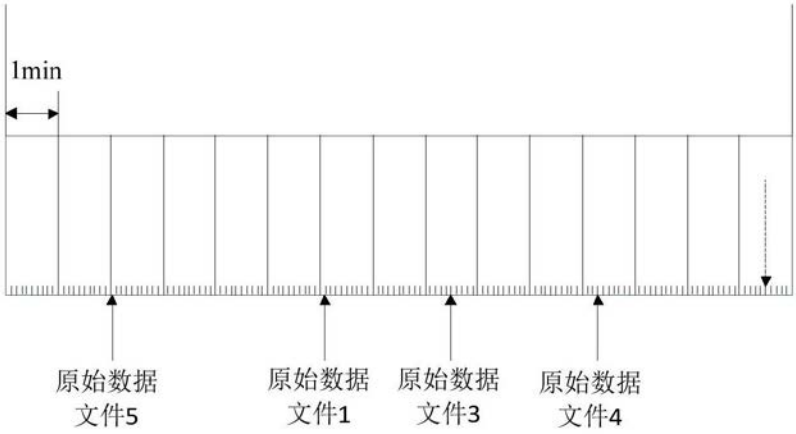

图6

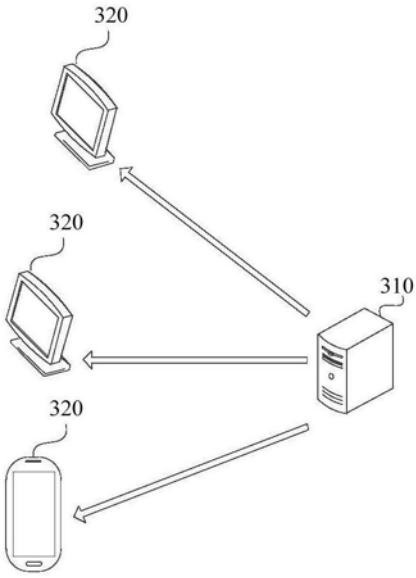

图7

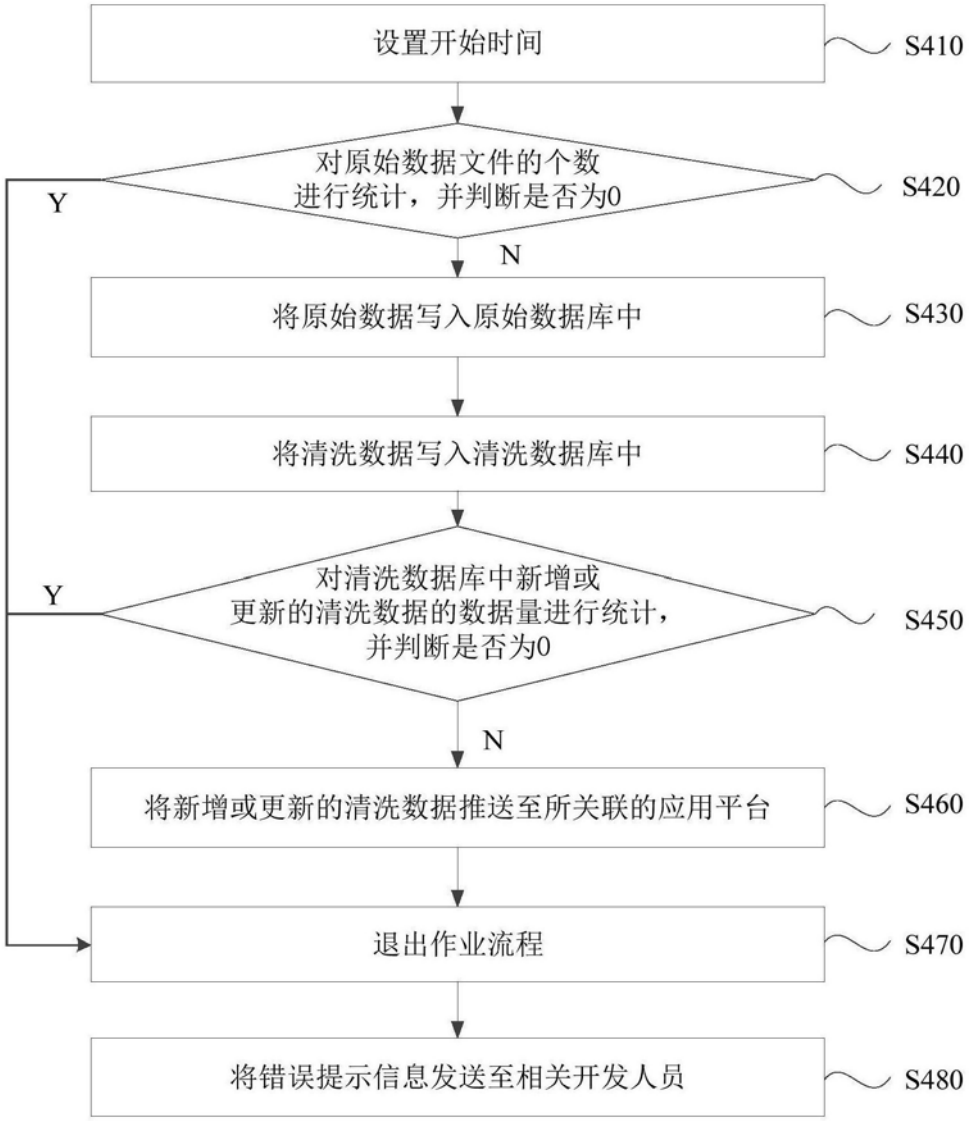
设置开始时间
S410
对原始数据文件的个数
进行统计，并判断是否为0
S420
Y
N
将原始数据写入原始数据库中
S430
将清洗数据写入清洗数据库中
S440
对清洗数据库中新增或
更新的清洗数据的数据量进行统计，
并判断是否为0
S450
Y
N
将新增或更新的清洗数据推送至所关联的应用平台
S460
退出作业流程
S470
将错误提示信息发送至相关开发人员
S480

图8

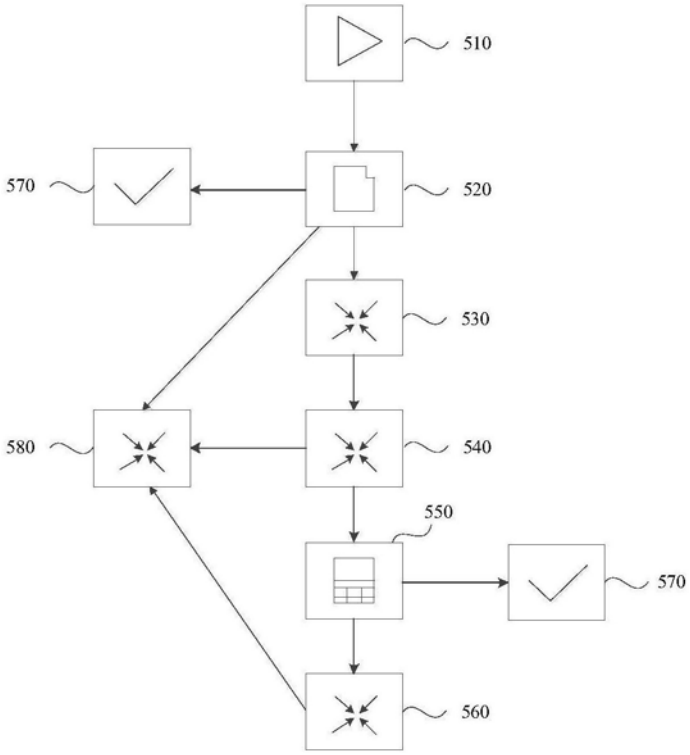
510
570
520
530
580
540
550
570
560

图9

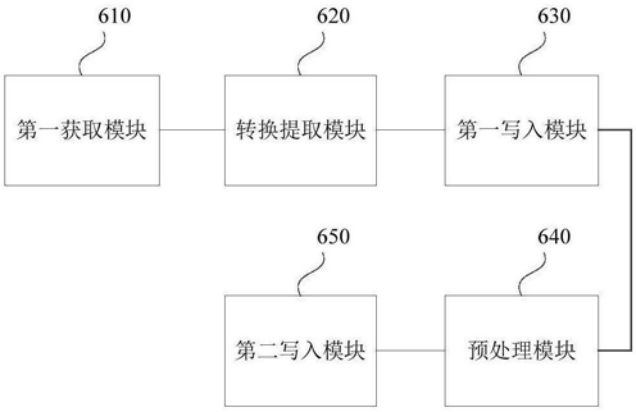
610
620
630
第一获取模块
转换提取模块
第一写入模块
650
640
第二写入模块
预处理模块

图10

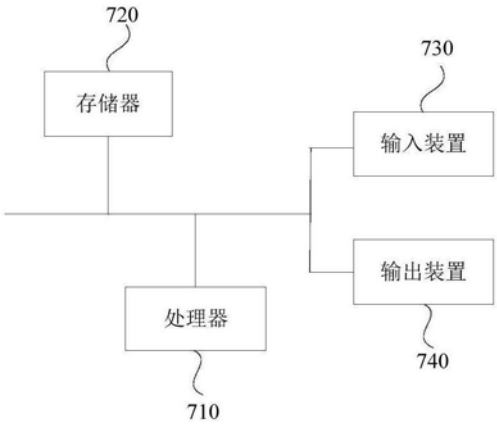

图11