(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 105224527 A
(43)申请公布日 2016.01.06

(21)申请号 201410226998.7

(22)申请日 2014.05.27

(71)申请人 北京宸瑞科技有限公司
地址 100036 北京市海淀区复兴路甲65号A座16层

(72)发明人 孙二林

(74)专利代理机构 北京康思博达知识产权代理事务所(普通合伙) 11426
代理人 路永斌 余光军

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

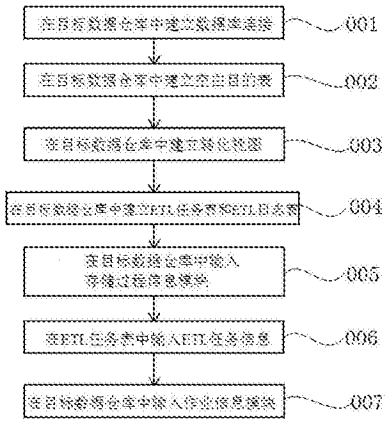
权利要求书3页 说明书17页 附图5页

(54)发明名称

适用于多种目的表更新方式的通用ETL方法

(57)摘要

本发明公开了一种适用于多种目的表更新方式的通用ETL方法，该方法既避免了ETL工具软件所存在的成本高、速度慢、故障多、实施工作量大的缺点，又克服了现有数据库脚本所存在的不通用、不能实现多种表更新方式的问题，该方法设置用于记录数据仓库中的各个ETL任务的ETL任务表、用于记录ETL存储过程运行时产生的日志的ETL日志表、用于执行某个作业所包含的ETL任务的存储过程模块以及在存储过程模块执行时，用于把目的表的索引和备注复制到影子表的索引和备注信息存储过程亚模块，通过设置上述四个数据库对象，调整其顺序，再设置数据库连接、作业等必须程序，从而完成抽取源表数据进行转化并最终将数据存储至目标数据仓库的ETL过程。



CN 105224527 A

1. 一种适用于多种目的表更新方式的通用 ETL 方法，其特征在于，该方法包括如下步骤：

步骤 1：在目标数据仓库中建立数据库连接，数据库连接用于访问源表；

步骤 2：在目标数据仓库中建立空白目的表；

步骤 3：在目标数据仓库中建立转化视图，所述转化视图的数据结构与目的表的数据结构一致；

步骤 4：在目标数据仓库中建立 ETL 任务表和 ETL 日志表，其中，ETL 任务表用于记录 ETL 任务的内容，ETL 任务表至少包括更新方式栏、作业名称栏、更新序号栏、更新状态栏，ETL 日志表用于记录 ETL 存储过程在运行时产生的日志信息；

步骤 5：在目标数据仓库中输入存储过程信息模块，所述存储过程信息模块包括存储过程模块和索引和备注信息存储过程亚模块，其中，存储过程模块用于执行作业信息模块所包含的 ETL 任务；

步骤 6：在 ETL 任务表中输入 ETL 任务信息，所述输入的 ETL 任务信息至少包括从源表到目的表的更新方式和作业信息模块名称；

步骤 7：在目标数据仓库中输入作业信息模块，每个作业信息模块中都包括一个以上 ETL 任务，所述作业信息模块用于在预定的时间通过向接收子模块发送参数的方式来调用存储过程模块；

其中，存储过程模块包括：

接收子模块，其用于接收作业信息模块发出的参数信息，并通知统计子模块查找并记录该作业信息模块中包含的所有 ETL 任务，同时，通知 ETL 日志表记录开始日志；

统计子模块，其用于查找并记录作业信息模块包含的所有 ETL 任务，将各个 ETL 任务按照更新序号排序，其中，当统计子模块开始查找并记录 ETL 任务时，通知 ETL 任务表将其中的更新状态栏改为等待执行；

更新方式选择子模块，其用于查询在 ETL 任务表中预设的从源表到目的表的 ETL 任务的更新方式，并根据预设的从源表到目的表的更新方式通知相应的更新子模块对目的表进行更新，同时通知 ETL 任务表将其中的更新状态栏改为正在执行；

全表数据替换更新子模块，其用于在预设的从源表到目的表的更新方式为全表数据替换时对目的表进行更新，其对目的表的更新过程包括如下步骤：根据源表创建目的表的影子表，所述影子表中只有数据；删除目的表；把影子表重命名为目的表；

全表数据替换并重建索引子模块，其用于在预设的从源表到目的表的更新方式为全表数据替换并重建索引时对目的表进行更新，其对目的表的更新过程包括如下步骤：根据源表创建目的表的影子表，所述影子表中只有数据；调用用于把目的表中的索引信息和备注信息复制到影子表的索引和备注信息存储过程亚模块，通过索引和备注信息存储过程亚模块把目的表的索引信息和备注信息复制到影子表；删除目的表；把影子表重命名为目的表；

差异添加子模块，其用于在预设的从源表到目的表的更新方式为差异添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：向目的表中插入源表中有而目的表中没有的数据；

时间修改添加或字符串修改添加子模块，其用于在预设的从源表到目的表的更新方式

为时间修改添加或字符串修改添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：删除目的表中数据更新时间与源表中数据更新时间不一致的所有数据；向目的表中插入源表中有而目的表中没有的数据；

时间添加子模块，其用于在预设的从源表到目的表的更新方式为时间添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：分别计算源表和目的表最大添加时间；把源表中数据添加时间介于上述两个时间点之间的数据添加到目的表中；

和

循环处理子模块，其用于在更新模块更新结束后通知ETL任务表将其中的更新状态栏改为已完成，并记录ETL任务结束时间和新增数据条数；进行循环处理过程，即通知更新方式选择子模块处理下一个ETL任务，直至作业信息模块所包含的所有ETL任务都执行完毕，当作业信息模块所包含的所有ETL任务都执行完毕后，存储过程模块运行结束，通知ETL日志表记录结束日志。

2. 根据权利要求1所述的适用于多种目的表更新方式的通用ETL方法，其特征在于，该方法的步骤2为：在目标数据仓库中建立空白目的表，在所述目的表中添加索引栏和备注栏，或调用已有目的表。

3. 根据权利要求1所述的适用于多种目的表更新方式的通用ETL方法，其特征在于，转化视图用于把源表的数据结构转化为目的表的数据结构。

4. 根据权利要求1所述的适用于多种目的表更新方式的通用ETL方法，其特征在于，ETL任务表包括：

目的表栏，其用于记录ETL任务中目标数据仓库的表名；

源表栏，其用于记录ETL任务中源表的表名；

更新方式栏，其用于记录ETL任务中从源表到目的表的更新方式；

作业名称栏，其用于记录该ETL任务所属的作业信息模块名称；

更新序号栏，其用于记录该ETL任务在所属作业信息模块中的序号；

更新状态栏，其用于记录该ETL任务当前的执行状态，ETL任务的执行状态包括：编辑、等待执行、正在执行、报错退出和已完成；

主键名栏，其用于记录ETL任务中源表和目的表的主键名；

和

增量字段名栏，其用于记录ETL任务中源表和目的表的更新时间字段名或添加时间字段名。

5. 根据权利要求4所述的适用于多种目的表更新方式的通用ETL方法，其特征在于，从源表到目的表的更新方式包括：手工更新、全表数据替换、全表数据替换并重建索引、差异添加、时间修改添加、字符串修改添加和时间添加。

6. 根据权利要求1所述的适用于多种目的表更新方式的通用ETL方法，其特征在于，ETL日志表包括：

序号栏，其用于记录自动增长的ETL任务序列号；

存储过程名栏，其用于记录在ETL任务中调用的存储过程信息模块名称及为调用该存储过程信息模块所输入的参数；

错误号栏，其用于记录在存储过程模块和/或索引和备注信息存储过程亚模块运行时

产生的错误号，用 0 表示无错误；

错误信息栏，其用于记录在存储过程模块和 / 或索引和备注信息存储过程亚模块运行时产生的错误信息；

开始时间栏，其用于记录存储过程模块和 / 或索引和备注信息存储过程亚模块开始运行的时间；

和

结束时间栏，其用于记录存储过程模块和 / 或索引和备注信息存储过程亚模块运行结束的时间。

7. 根据权利要求 1 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，索引和备注信息存储过程亚模块包括：

接收亚模块，其用于接收全表数据替换并重建索引子模块发出的调用信息，通知检查子模块检查影子表是否有索引，同时，通知 ETL 日志表记录开始日志；

检查子模块，其用于检查影子表是否有索引，如果发现影子表有索引，则删除该索引；

查找子模块，其用于查找目的表的各个索引；

循环建立子模块，其用于参照目的表的索引建立影子表的各个索引；

查询子模块，其用于分别查询目的表索引的类型，查询目的表索引的字段，查询目的表的索引是否为函数索引；

创建子模块，其用于根据查询子模块查询到的信息创建影子表索引；

和

复制子模块，其用于将目的表的表备注复制到影子表，将目的表的字段备注复制到影子表；通知 ETL 日志表记录结束日志。

8. 根据权利要求 1 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，步骤 6 为：在 ETL 任务表中输入 ETL 任务，所述输入的 ETL 任务包括输入每个 ETL 任务的目的表表名、源表表名、主键名、更新方式、作业名称、更新序号和增量字段名。

9. 根据权利要求 1 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，在步骤 7 中，预定的时间是指在每天的预定时间点和 / 或每隔预定的时间段。

10. 根据权利要求 1 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，在步骤 7 中，在目标数据仓库中输入多个作业信息模块，每个作业信息模块中都记载一个以上 ETL 任务，传输所述参数以调用存储过程模块，所述参数为在 ETL 任务表的作业名称栏内填写的各作业信息模块名称。

适用于多种目的表更新方式的通用 ETL 方法

技术领域

[0001] 本发明涉及一种数据库更新存储的 ETL 方法，具体涉及一种适用于多种目的表更新方式的通用 ETL 方法。

背景技术

[0002] ETL 是 Extract-Transform-Load 的缩写，即数据抽取、转换和装载的过程，也可以理解为数据提取、转化和加载的过程，ETL 作为数据仓库和商务智能的核心和灵魂，能够按照统一的规则集成并提高数据的价值，是负责完成数据从数据源向目标数据仓库转化的过程，是实施数据仓库的重要步骤。

[0003] 目前的 ETL 方法分为两大类 :ETL 工具软件类和数据库脚本类。

[0004] 所谓 ETL 工具软件类，是指除数据仓库本身所用到的数据库软件以外，再安装一套 ETL 工具软件，ETL 软件既可以装在数据仓库服务器上，也可以装在单独的 ETL 服务器上。ETL 软件通过统一的接口连接数据源和目标数据仓库，并通过多个配置文件和任务计划实现多个 ETL 过程。

[0005] 所谓数据库脚本类，是指利用数据仓库本身的数据库软件的远程数据库连接、表、视图、存储过程、作业等功能，通过编写和运行数据库脚本来实现 ETL 过程，不用再安装另外的 ETL 工具软件。

[0006] ETL 工具软件的缺点在于 ：

[0007] (1) 成本高，需要另外购买 ETL 软件，需要招聘能掌握 ETL 软件的人才，且需要维护 ETL 软件的正常运行。

[0008] (2) 速度慢，ETL 软件不管访问何种数据库，都采用标准的数据接口，而不同的数据库有自己的非标准的、高速的数据接口。因此 ETL 软件的 ETL 速度比经过优化的数据库脚本速度要慢。

[0009] (3) 故障多，ETL 软件方式因为涉及到两套软件，因而架构较复杂，节点较多，其中任一节点出问题，都会导致 ETL 过程失败。而数据库脚本方式的架构简单，故障少。

[0010] (4) 实施工作量大，ETL 软件采用图形化的配置方式，无法批量复制，必须逐表、逐字段的手工配置，而 ETL 所涉及的表和字段往往很多。而数据库脚本通过编程实现，它可以用传递参数和循环遍历的方法压缩工作量。

[0011] 当然，现有的数据库脚本也存在问题，如 ：

[0012] (1) 不通用，数据库开发人员往往会根据不同的项目编写不同的数据库脚本，一个项目上的数据库脚本不能完全移植到别的项目上，这就会导致重复的开发工作。

[0013] (2) 不能实现多种目的表更新方式，数据仓库中的目的表可以有多种更新方式，不同的目的表适用于不同的更新方式，数据库脚本难以囊括多种目的表更新方式。

[0014] 由于上述原因，本发明人对现有的数据库脚本的 ETL 方法进行了深入研究，以便开发出解决上述问题的 ETL 方法。

发明内容

[0015] 为了克服上述问题，本发明人进行了锐意研究，设计出一种适用于多种目的表更新方式的通用 ETL 方法，该方法既避免了 ETL 工具软件所存在的成本高、速度慢、故障多、实施工作量大的缺点，又克服了现有数据库脚本所存在的不通用、不能实现多种表更新方式的问题，该方法的核心是设置 ETL 过程的四个位于目标数据仓库的数据库对象，四个数据库对象是 :ETL 任务表、ETL 日志表、存储过程模块和索引和备注信息存储过程亚模块，其中，ETL 任务表用于记录数据仓库中的各个 ETL 任务 ;ETL 日志表用于记录 ETL 存储过程运行时产生的日志，存储过程模块用于执行某个作业信息模块所包含的 ETL 任务，索引和备注信息存储过程亚模块用于把目的表的索引和备注复制到影子表，设置上述四个数据库对象，调整其顺序，再设置数据库连接、作业等必须程序，从而完成抽取源表数据进行转化并最终将数据存储至目标数据仓库的 ETL 过程，从而完成本发明。

[0016] 具体来说，本发明的目的在于提供以下方面 :

[0017] (1) 一种适用于多种目的表更新方式的通用 ETL 方法，其特征在于，该方法包括如下步骤 :

[0018] 步骤 1 :在目标数据仓库中建立数据库连接，数据库连接用于访问源表 ;

[0019] 步骤 2 :在目标数据仓库中建立空白目的表 ;

[0020] 步骤 3 :在目标数据仓库中建立转化视图，所述转化视图的数据结构与目的表的数据结构一致 ;

[0021] 步骤 4 :在目标数据仓库中建立 ETL 任务表和 ETL 日志表，其中，ETL 任务表用于记录 ETL 任务的内容，ETL 任务表至少包括更新方式栏、作业名称栏、更新序号栏、更新状态栏，ETL 日志表用于记录 ETL 存储过程在运行时产生的日志信息 ;

[0022] 步骤 5 :在目标数据仓库中输入存储过程信息模块，所述存储过程信息模块包括存储过程模块和索引和备注信息存储过程亚模块，其中，存储过程模块用于执行作业信息模块所包含的 ETL 任务 ;

[0023] 步骤 6 :在 ETL 任务表中输入 ETL 任务信息，所述输入的 ETL 任务信息至少包括从源表到目的表的更新方式和作业信息模块名称 ;

[0024] 步骤 7 :在目标数据仓库中输入作业信息模块，每个作业信息模块中都包括一个以上 ETL 任务，所述作业信息模块用于在预定的时间通过向接收子模块发送参数的方式来调用存储过程模块 ;

[0025] 其中，存储过程模块包括 :

[0026] 接收子模块，其用于接收作业信息模块发出的参数信息，并通知统计子模块查找并记录该作业信息模块中包含的所有 ETL 任务，同时，通知 ETL 日志表记录开始日志 ;

[0027] 统计子模块，其用于查找并记录作业信息模块包含的所有 ETL 任务，将各个 ETL 任务按照更新序号排序，其中，当统计子模块开始查找并记录 ETL 任务时，通知 ETL 任务表将其中的更新状态栏改为等待执行 ;

[0028] 更新方式选择子模块，其用于查询在 ETL 任务表中预设的从源表到目的表的 ETL 任务的更新方式，并根据预设的从源表到目的表的更新方式通知相应的更新子模块对目的表进行更新，同时通知 ETL 任务表将其中的更新状态栏改为正在执行 ;

[0029] 全表数据替换更新子模块，其用于在预设的从源表到目的表的更新方式为全表数

据替换时对目的表进行更新，其对目的表的更新过程包括如下步骤：根据源表创建目的表的影子表，所述影子表中只有数据；删除目的表；把影子表重命名为目的表；

[0030] 全表数据替换并重建索引子模块，其用于在预设的从源表到目的表的更新方式为全表数据替换并重建索引时对目的表进行更新，其对目的表的更新过程包括如下步骤：根据源表创建目的表的影子表，所述影子表中只有数据；调用用于把目的表中的索引信息和备注信息复制到影子表的索引和备注信息存储过程亚模块，通过索引和备注信息存储过程亚模块把目的表的索引信息和备注信息复制到影子表；删除目的表；把影子表重命名为目的表；

[0031] 差异添加子模块，其用于在预设的从源表到目的表的更新方式为差异添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：向目的表中插入源表中有而目的表中没有的数据；

[0032] 时间修改添加或字符串修改添加子模块，其用于在预设的从源表到目的表的更新方式为时间修改添加或字符串修改添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：删除目的表中数据更新时间与源表中数据更新时间不一致的所有数据；向目的表中插入源表中有而目的表中没有的数据；

[0033] 时间添加子模块，其用于在预设的从源表到目的表的更新方式为时间添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：分别计算源表和目的表最大添加时间；把源表中数据添加时间介于上述两个时间点之间的数据添加到目的表中；

[0034] 和

[0035] 循环处理子模块，其用于在更新模块更新结束后通知ETL任务表将其中的更新状态栏改为已完成，并记录ETL任务结束时间和新增数据条数；进行循环处理过程，即通知更新方式选择子模块处理下一个ETL任务，直至作业信息模块所包含的所有ETL任务都执行完毕，当作业信息模块所包含的所有ETL任务都执行完毕后，存储过程模块运行结束，通知ETL日志表记录结束日志。

[0036] (2)根据上述(1)所述的适用于多种目的表更新方式的通用ETL方法，其特征在于，该方法的步骤2为：在目标数据仓库中建立空白目的表，在所述目的表中添加索引栏和备注栏，或调用已有目的表。

[0037] (3)根据上述(1)所述的适用于多种目的表更新方式的通用ETL方法，其特征在于，转化视图用于把源表的数据结构转化为目的表的数据结构。

[0038] (4)根据上述(1)所述的适用于多种目的表更新方式的通用ETL方法，其特征在于，ETL任务表包括：

[0039] 目的表栏，其用于记录ETL任务中目标数据仓库的表名；

[0040] 源表栏，其用于记录ETL任务中源表的表名；

[0041] 更新方式栏，其用于记录ETL任务中从源表到目的表的更新方式；

[0042] 作业名称栏，其用于记录该ETL任务所属的作业信息模块名称；

[0043] 更新序号栏，其用于记录该ETL任务在所属作业信息模块中的序号；

[0044] 更新状态栏，其用于记录该ETL任务当前的执行状态，ETL任务的执行状态包括：编辑、等待执行、正在执行、报错退出和已完成；

[0045] 主键名栏，其用于记录ETL任务中源表和目的表的主键名；

[0046] 和

[0047] 增量字段名栏，其用于记录 ETL 任务中源表和目的表的更新时间字段名或添加时间字段名。

[0048] (5) 根据上述 (4) 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，从源表到目的表的更新方式包括：手工更新、全表数据替换、全表数据替换并重建索引、差异添加、时间修改添加、字符串修改添加和时间添加。

[0049] (6) 根据上述 (1) 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，ETL 日志表包括：

[0050] 序号栏，其用于记录自动增长的 ETL 任务序列号；

[0051] 存储过程名栏，其用于记录在 ETL 任务中调用的存储过程信息模块名称及为调用该存储过程信息模块所输入的参数；

[0052] 错误号栏，其用于记录在存储过程模块和 / 或索引和备注信息存储过程亚模块运行时产生的错误号，用 0 表示无错误；

[0053] 错误信息栏，其用于记录在存储过程模块和 / 或索引和备注信息存储过程亚模块运行时产生的错误信息；

[0054] 开始时间栏，其用于记录存储过程模块和 / 或索引和备注信息存储过程亚模块开始运行的时间；

[0055] 和

[0056] 结束时间栏，其用于记录存储过程模块和 / 或索引和备注信息存储过程亚模块运行结束的时间。

[0057] (7) 根据上述 (1) 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，索引和备注信息存储过程亚模块包括：

[0058] 接收亚模块，其用于接收全表数据替换并重建索引子模块发出的调用信息，通知检查子模块检查影子表是否有索引，同时，通知 ETL 日志表记录开始日志；

[0059] 检查子模块，其用于检查影子表是否有索引，如果发现影子表有索引，则删除该索引；

[0060] 查找子模块，其用于查找目的表的各个索引；

[0061] 循环建立子模块，其用于参照目的表的索引建立影子表的各个索引；

[0062] 查询子模块，其用于分别查询目的表索引的类型，查询目的表索引的字段，查询目的表的索引是否为函数索引；

[0063] 创建子模块，其用于根据查询子模块查询到的信息创建影子表索引；

[0064] 和

[0065] 复制子模块，其用于将目的表的表备注复制到影子表，将目的表的字段备注复制到影子表；通知 ETL 日志表记录结束日志。

[0066] (8) 根据上述 (1) 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，步骤 6 为：在 ETL 任务表中输入 ETL 任务，所述输入的 ETL 任务包括输入每个 ETL 任务的目的表表名、源表表名、主键名、更新方式、作业名称、更新序号和增量字段名。

[0067] (9) 根据上述 (1) 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，在步骤 7 中，预定的时间是指在每天的预定时间点和 / 或每隔预定的时间段。

[0068] (10) 根据上述 (1) 所述的适用于多种目的表更新方式的通用 ETL 方法，其特征在于，在步骤 7 中，在目标数据仓库中输入多个作业信息模块，每个作业信息模块中都记载一个以上 ETL 任务，传输所述参数以调用存储过程模块，所述参数为在 ETL 任务表的作业名称栏内填写的各作业信息模块名称。

[0069] 本发明所提供的适用于多种目的表更新方式的通用 ETL 方法运行成本低、速度快、稳定可靠、实施工作量小、通用性好、能满足多种表更新方式；具体来说，本发明具有的有益效果包括：

[0070] (1) 使用本发明提供的 ETL 方法时，无需购买、安装、维护 ETL 工具软件；不会出现数据重复、错误、缺漏的情况；

[0071] (2) 本发明提供的 ETL 方法中有详细的任务状态和错误日志，以便监控和检查错误原因；

[0072] (3) 本发明提供的 ETL 方法中常见的错误易于修复，一种是某个数据源掉线了，只要它再次上线数据就能自动同步进来；另一种是某个源表的表数据结构发生改变，这时只要修改相关的转化视图和目的表的表数据结构即可；

[0073] (4) 本发明提供的 ETL 方法具有通用性，该方法适用于包括 Oracle、SQL Server、DB2 等在内的所有主流数据库，按照本发明编写出来的数据库脚本在同一种数据库内是完全可移植、可复用的，例如在一个 Oracle 项目中编写的脚本可复制到任意一个 Oracle 项目中使用；

[0074] (5) 使用本发明提供的 ETL 方法时不会影响应用程序对表的访问；

[0075] (6) 使用本发明提供的 ETL 方法时不会造成数据碎片和索引碎片，使表的性能保持最高；

[0076] (7) 本发明提供的 ETL 方法适用范围广，有无主键均的表都可采用。

附图说明

[0077] 图 1 示出根据本发明一种优选实施方式的适用于多种目的表更新方式的通用 ETL 方法的整体流程示意图；

[0078] 图 2 示出根据本发明一种优选实施方式的适用于多种目的表更新方式的通用 ETL 方法的存储过程模块结构及其各个子模块间的工作顺序；

[0079] 图 3 示出根据本发明一种优选实施方式的适用于多种目的表更新方式的通用 ETL 方法的索引和备注信息存储过程亚模块结构及其各个子模块间的工作顺序；

[0080] 图 4 示出根据本发明一种优选实施方式的适用于多种目的表更新方式的通用 ETL 方法的存储过程模块工作流程示意图；

[0081] 图 5 示出根据本发明一种优选实施方式的适用于多种目的表更新方式的通用 ETL 方法的索引和备注信息存储过程亚模块工作流程示意图。

[0082] 附图标号说明：

[0083] 001- 步骤 1

[0084] 002- 步骤 2

[0085] 003- 步骤 3

[0086] 004- 步骤 4

[0087] 005- 步骤 5

[0088] 006- 步骤 6

[0089] 007- 步骤 7

[0090] 501- 全表数据替换更新方式的更新流程

[0091] 502- 全表数据替换并重建索引更新方式的更新流程

[0092] 503- 差异添加更新方式的更新流程

[0093] 504- 时间修改添加或字符串修改添加更新方式的更新流程

[0094] 505- 时间添加更新方式的更新流程

具体实施方式

[0095] 下面通过附图和实施例对本发明进一步详细说明。通过这些说明，本发明的特点和优点将变得更为清楚明确。

[0096] 在这里专用的词“示例性”意为“用作例子、实施例或说明性”。这里作为“示例性”所说明的任何实施例不必解释为优于或好于其它实施例。尽管在附图中示出了实施例的各种方面，但是除非特别指出，不必按比例绘制附图。

[0097] 在根据本发明的一个优选的实施方式中，如图 1 中所示，提供一种适用于多种目的表更新方式的通用 ETL 方法：该方法包括如下步骤：

[0098] 如附图标号 001 所示的步骤 1：在目标数据仓库中建立数据库连接；

[0099] 如附图标号 002 所示的步骤 2：在目标数据仓库中建立空白目的表；

[0100] 如附图标号 003 所示的步骤 3：在目标数据仓库中建立转化视图；

[0101] 如附图标号 004 所示的步骤 4：在目标数据仓库中建立 ETL 任务表和 ETL 日志表；

[0102] 如附图标号 005 所示的步骤 5：在目标数据仓库中输入存储过程信息模块；

[0103] 如附图标号 006 所示的步骤 6：在 ETL 任务表中输入 ETL 任务信息；

[0104] 如附图标号 007 所示的步骤 7：在目标数据仓库中输入作业信息模块。通过作业信息模块在预定的时间调用存储过程模块，开始执行 ETL 任务，进而完成数据从源表到目标数据仓库的抽取、转换、装载过程。

[0105] 在一个优选的实施方式中，ETL 方法是指抽取源表数据进行转化并最终将数据存储至目标数据仓库的一种方法，其包括数据抽取、转化和存储等过程。

[0106] 在一个优选的实施方式中，如图 1 中的附图标号 001 所示，步骤 1：在目标数据仓库中建立数据库连接；数据库连接是数据库中的一种对象，它使得一台服务器中的数据库可以访问另一台服务器中的数据库，本发明中的数据库连接用于访问源表，本发明中，源表包括本地数据库中的表、远程数据库中的表和异构数据库中的表，由于本地数据库中的表可以直接进行访问，不必通过建立数据库连接等方式，所以步骤 1 中建立的数据连接主要用于访问远程数据库和异构数据库，具有数据库连接的目标数据仓库可以访问源表并获取源表中的数据。

[0107] 在一个优选的实施方式中，如图 1 中的附图标号 002 所示，步骤 2：在目标数据仓库中建立空白目的表，并根据情况在目的表中添加索引栏和备注栏，当然，也可以根据具体情况选择不添索引栏和备注栏，也可以只添加索引栏或只添加备注栏。

[0108] 在进一步优选的实施方式中，步骤 2 为，在目标数据仓库中调用一个已有的目的

表，用这个已有的目的表代替上述空白目的表。

[0109] 在一个优选的实施方式中，如图 1 中的附图标号 003 所示，步骤 3 ：在目标数据仓库中建立转化视图 ；每个 ETL 过程都包含一个源表和一个目的表，ETL 过程是实现 ETL 方法的过程，即抽取源表数据，经过转化以后导入到目的表中，源表和目的表的数据结构可能一致也可能不一致，如不一致就必须把源表数据结构转化成与目的表数据结构相一致的数据结构，转化视图的作用就是完成上述转化过程，即转变源表的数据结构，使源表的数据结构与目的表的数据结构一致 ；且所述转化视图的数据结构与目的表的数据结构一致。本发明中所述的数据结构为目的表和 / 或转化视图的数据结构，该数据结构是指所包含的字段的个数、名称及数据类型。

[0110] 一个优选的实施方式中，如果源表和目的表的数据结构一致就不需要经过转化视图进行转化，可以直接从源表导入到目的表中，源表和目的表的数据结构如不一致就必须把源表转化成与目的表一致的数据结构才能导入 ；源表和目的表是相对的，一个表在一个 ETL 过程中可以是目的表，而在另一个 ETL 过程中可以是源表，因此，多个简单的 ETL 过程可以组成一个复杂的 ETL 过程，一个复杂的 ETL 过程中可以包括多个转化视图。

[0111] 在一个优选的实施方式中，如图 1 中的附图标号 004 所示，步骤 4 ：在目标数据仓库中建立 ETL 任务表和 ETL 日志表 ；其中，ETL 任务表用于记录数据仓库中的各个 ETL 任务的内容，ETL 日志表用于记录 ETL 存储过程在运行时产生的日志信息，本发明中所述的 ETL 存储过程是指 ETL 任务的执行过程，存储过程模块执行 ETL 任务的过程，包括数据从源表到目的表的更新过程。

[0112] 在一个优选的实施方式中，ETL 任务表如下表所示 ：

[0113]

字段（栏）	数据类型	简要说明
目的表	字符串型	目的表表名，唯一主键
源表	字符串型	源表表名
主键名	字符串型	源表和目的表的主键名，差异添加、时间修改添加、字符串修改添加时必填
更新方式	字符串型	手工更新 全表数据替换 全表数据替换并重建索引 差异添加 时间修改添加 字符串修改添加 时间添加
作业名称	字符串型	该 ETL 任务所属的作业信息模块名称
更新序号	数值型	该 ETL 任务在所属作业信息模块中的顺序号
更新状态	字符串型	编辑 等待执行 正在执行 报错退出 已完成
启动时间	日期型	该 ETL 任务最后一次运行的开始时间
结束时间	日期型	该 ETL 任务最后一次运行的结束时间
备注	字符串型	
增量字段名	字符串型	源表和目的表的更新时间字段名或添加时间字段名，时间修改添加、字符串修改添加、时间添加时必填
上次数据条数	数值型	目的表最后一次更新前的数据条数
当前数据条数	数值型	目的表最后一次更新后的数据条数
新增数据条数	数值型	目的表最后一次更新新增的数据条数

[0114]

[0115] 如上表中所示，ETL 任务包括：

[0116] 目的表栏，其为主键字段，其用于记录 ETL 任务中目标数据仓库中的表名，尤其是用于记录目的表的表名；该栏内容由手工输入。

[0117] 源表栏，其用于记录 ETL 任务中的源表的相关信息，包括本地数据库即目标数据仓库的表名、视图名，以及远程数据库的表名、视图名或异构数据库的表名、视图名，即源表字段（源表栏）用于记录源表的表名或视图名；该栏内容由手工输入。

[0118] 主键名栏，其用于其记录 ETL 任务中源表和目的表的主键名，且更新方式为差异添加或时间修改添加或字符串修改添加时必填；该栏内容由手工输入。

[0119] 更新方式栏，其用于记录在 ETL 任务中从源表到目的表的更新方式；更新方式包括手工更新、全表数据替换、全表数据替换并重建索引、差异添加、时间修改添加、字符串修改添加和时间添加；该栏内容由操作者预先设定并手工输入。

[0120] 作业名称栏，其用于记录该 ETL 任务所属的作业信息模块名称；该栏内容由操作者预先设定并手工输入。

[0121] 更新序号栏，其用于记录该 ETL 任务在所属作业信息模块中的序号；该栏内容预先设定好，并由操作者手工输入。

[0122] 更新状态栏，其用于记录该 ETL 任务当前的执行状态，ETL 任务的执行状态包括：编辑、等待执行、正在执行、报错退出和已完成；该栏内容在 ETL 任务运行时自动产生。

[0123] 启动时间栏，其用于记录该ETL 任务最后一次运行的开始时间；该栏内容在ETL 任务运行时自动产生。

[0124] 结束时间栏，其用于记录该ETL 任务最后一次运行的结束时间；该栏内容在ETL 任务运行时自动产生。

[0125] 备注栏，其用于记录该 ETL 任务的更详细的文字描述，可以输也可以不输，如果输入，该栏内容由手工输入。

[0126] 增量字段名栏，其用于记录 ETL 任务中源表和目的表的更新时间字段名或添加时间字段名；该栏内容由手工输入。其中，当目的表更新方式为时间修改添加、字符串修改添加或时间添加时，源表和目的表中一定有一个时间字段，这个字段记录了表中每条数据的更新时间和添加时间，即为更新时间字段和／或添加时间字段，如果没有上述字段，目的表无法实现这 3 种更新方式。

[0127] 上次数据条数栏用于记录目的表在最后一次更新以前的数据条数；该栏内容在 ETL 任务运行时自动产生。

[0128] 当前数据条数栏用于记录目的表在最后一次更新以后的数据条数；该栏内容在 ETL 任务运行时自动产生。

[0129] 新增数据条数栏用于记录目的表在最后一次更新时新增的数据条数，即当前数据条数与上次数据条数之差；该栏内容在 ETL 任务运行时自动产生。

[0130] 上述最后一次更新是指上述 ETL 任务在运行时将数据存储至目的表中的过程，即更新是目的表的存储过程。ETL 任务运行时自动产生的上次数据条数字段、当前数据条数字段和新增数据条数字段记录的“最后一次更新”都是指该 ETL 任务运行导致的目的表的更新。

[0131] 在一个优选的实施方式中，本发明提供的从源表到目的表的更新方式中，手工更新是指人工把源表中的数据导入到目的表中，作业不自动更新目的表，此种更新方式适用

于那些不需要自动更新的表，手工更新的 ETL 任务不由存储过程执行，而是由人工执行，因此存储过程模块中没有用于手工更新方式的更新子模块，，操作人员在输入 ETL 任务表时，有可能会遇到一些手工更新的表，这些表虽然不需要由存储过程定时自动更新，但需要登记在 ETL 任务表中，以使得整个过程完整。

[0132] 在一个优选的实施方式中，ETL 日志表如下表所示：

[0133]

字段（栏）	数据类型	简要说明
序号	数值型	自动增长的序列号，唯一主键
存储过程名	字符串型	存储过程信息模块的名称和参数
错误号	字符串型	报错的错误号，0 表示无错误
错误信息	字符串型	报错的错误信息
开始时间	日期型	存储过程信息模块开始运行的时间
结束时间	日期型	存储过程信息模块结束运行的时间
备注	长文本型	报错的 SQL 语句

[0134] 如上表所示，ETL 日志表包括：

[0135] 序号栏，其为主键字段，其用于记录自动增长的 ETL 任务序列号；

[0136] 存储过程名栏，其用于记录在 ETL 任务中调用的存储过程信息模块名称及调用存储过程信息模块所输入的参数，即存储过程模块名称或索引和备注信息存储过程亚模块名称以及调用该存储过程模块所需输入的参数；

[0137] 错误号栏，其用于记录在存储过程模块和 / 或索引和备注信息存储过程亚模块运行时产生的错误号，用 0 表示无错误；

[0138] 错误信息栏，其用于记录在存储过程模块和 / 或索引和备注信息存储过程亚模块运行时产生的错误信息；

[0139] 开始时间栏，其用于记录存储过程模块和 / 或索引和备注信息存储过程亚模块运行开始的时间；

[0140] 结束时间栏，其用于记录存储过程模块和 / 或索引和备注信息存储过程亚模块运行结束的时间。

[0141] 在一个优选的实施方式中，如图 1 中的附图标号 005 所示，步骤 5：在目标数据仓库中输入用于执行该 ETL 任务的存储过程信息模块，所述存储过程信息模块包括存储过程模块和索引和备注信息存储过程亚模块，存储过程模块用于执行作业信息模块所包含的 ETL 任务，索引和备注信息存储过程亚模块用于把目的表的索引和备注复制到影子表，索引和备注信息存储过程亚模块不单独运行，它仅在存储过程模块调用时运行；

[0142] 在一个优选的实施方式中，如图 2 中所示，图 2 中示出了存储过程模块所包含的子模块名称及各个子模块间的工作顺序，所述工作顺序由箭头示出；存储过程模块包括：接

收子模块、统计子模块、更新方式选择子模块、全表数据替换更新子模块、差异添加子模块、时间修改添加或字符串修改添加子模块、时间添加子模块和循环处理子模块。

[0143] 其中，接收子模块用于接收作业信息模块发出的参数信息，并通知统计子模块查找并记录该作业信息模块中包含的所有ETL任务，同时，通知ETL日志表记录开始日志，即在开始时间栏内记录存储过程模块开始运行的时间；其中，参数为作业信息模块的名称，该信息记录在ETL任务表的作业名称栏内；

[0144] 统计子模块用于查找并记录作业信息模块下属的所有ETL任务，将各个ETL任务按照预设的更新序号排序，其中更新序号是手工输入到ETL任务表中的，当统计子模块开始查找并记录ETL任务时，通知ETL任务表将其中的更新状态栏改为等待执行；

[0145] 更新方式选择子模块用于查询在ETL任务表中预设的从源表到目的表的ETL任务的更新方式，并根据预设的从源表到目的表的更新方式通知相应的更新子模块对目的表进行更新，同时通知ETL任务表，将其中的更新状态栏改为正在执行；

[0146] 全表数据替换更新子模块用于在预设的从源表到目的表的更新方式为全表数据替换时对目的表进行更新，其对目的表的更新过程包括如下步骤：根据源表创建目的表的影子表，所述影子表中只有数据；删除目的表；把影子表重命名为目的表；其中，在数据库中，创建一张新表需要一定的时间。假如先把目的表删掉，再创建一张新的目的表，那么在此期间应用程序就不能访问目的表。因此最佳方法是先创建一张目的表的影子表，影子表的表结构与目的表完全一致，但其中的数据是新的，而且表名与目的表略有差别，例如可以在目的表的表名后加“_YZ”以示区分。影子表创建完以后再删除目的表并把影子表重命名为目的表的表名，这个时间非常短暂，这样就不会影响应用程序访问目的表。

[0147] 全表数据替换并重建索引子模块用于在预设的从源表到目的表的更新方式为全表数据替换并重建索引时对目的表进行更新，其对目的表的更新过程包括如下步骤：根据源表创建目的表的影子表，所述影子表中只有数据；调用用于把目的表中的索引信息和备注信息复制到影子表的索引和备注信息存储过程亚模块，通过索引和备注信息存储过程亚模块把目的表的索引信息和备注信息复制到影子表；最后，删除目的表；把影子表重命名为目的表；

[0148] 差异添加子模块用于在预设的从源表到目的表的更新方式为差异添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：向目的表中插入源表中有而目的表中没有的数据；

[0149] 时间修改添加或字符串修改添加子模块用于在预设的从源表到目的表的更新方式为时间修改添加或字符串修改添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：删除目的表中数据更新时间与源表中数据更新时间不一致的所有数据；向目的表中插入源表中有而目的表中没有的数据；本发明所述数据是指：存储在各个数据库及目标数据仓库中的各个表中的数据。数据库及目标数据仓库中的表具有预先定义的数据结构，即字段个数、各个字段的字段名和数据类型。表的数据结构一旦定义完毕，就可以向其中添加一条或多条数据，也可以删除、修改其中已有的数据。表中每一条数据的数据结构都与表本身的数据结构一致。

[0150] 时间添加子模块用于在预设的从源表到目的表的更新方式为时间添加时对目的表进行更新，其对目的表的更新过程包括如下步骤：分别计算源表和目的表最大添加时间；

把源表中数据添加时间介于上述两个时间点之间的数据添加到目的表中；

[0151] 循环处理子模块用于在更新模块更新结束后通知ETL任务表，将其中的更新状态栏改为已完成，并记录ETL任务结束时间和新增数据条数；进行循环处理过程，即通知更新方式选择子模块处理下一个ETL任务，直至作业信息模块所包含的所有的ETL任务都执行完毕，当作业信息模块所包含的所有的ETL任务都执行完毕后，存储过程模块运行结束，通知ETL日志表记录结束日志，结束日志包括在结束时间栏内填写存储过程模块运行结束的时间。

[0152] 在一个优选的实施方式中，如图3中所示，图3中示出了索引和备注信息存储过程亚模块所包含的子模块名称及各个子模块间的工作顺序，所述工作顺序由箭头示出；索引和备注信息存储过程亚模块包括：接收亚模块、检查子模块、查找子模块、循环建立子模块、查询子模块、创建子模块和复制子模块；

[0153] 其中，接收亚模块用于接收全表数据替换并重建索引子模块发出的调用信息，通知检查子模块检查影子表是否有索引，同时，通知ETL日志表记录开始日志，即在开始时间栏内记录全表数据替换并重建索引子模块开始运行的时间；其中，调用信息用于调用索引和备注信息存储过程亚模块，相当于作业信息模块发出的参数信息，在本发明中，该调用信息为目的表表名和影子表表名；

[0154] 检查子模块用于检查影子表是否有索引，如果发现影子表有索引，则删除该索引；

[0155] 查找子模块用于查找目的表的各个索引；

[0156] 循环建立子模块用于参照目的表的索引建立影子表的各个索引；

[0157] 查询子模块用于分别查询目的表索引的类型，查询目的表索引的字段，查询目的表的索引是否为函数索引；

[0158] 创建子模块用于根据查询子模块查询到的信息创建影子表索引；

[0159] 复制子模块用于将目的表的表备注复制到影子表，将目的表的字段备注复制到影子表；通知ETL日志表记录结束日志，即在结束时间栏内记录索引和备注信息存储过程亚模块运行结束的时间。

[0160] 在一个优选的实施方式中，如图1中的附图标号006所示，步骤6：在ETL任务表中输入ETL任务信息；所述ETL任务信息是指要完成一个ETL过程所需的基本信息，包括源表是什么、目的表是什么、如何从源表更新到目的表等，所以添加的ETL任务信息包括输入每个ETL任务的目的表名称、源表名称、主键名、更新方式、作业信息模块名称、更新序号和增量字段名，输入方式一般为手工输入，所以添加ETL任务是通过填写ETL任务表的形式完成的。

[0161] 在一个优选的实施方式中，如图1中的附图标号007所示，步骤7：在目标数据仓库中输入作业信息模块，每个作业信息模块中都包括一个以上ETL任务，所述作业信息模块用于在预定的时间通过向接收子模块发送参数的方式来调用存储过程模块，在本发明中，调用存储过程模块的输入参数就是作业信息模块名称，作业信息模块可以设置多个，虽然都是调用存储过程模块，但输入参数即作业信息模块名称不同，所以不会冲突，预定的时间是指每天的某个时间点和/或每经过一个预定的时间段以后可以理解为操作者所期望的任意一个时间点或时间段，并且可以多次重复开始，即多次调用存储过程模块，本发明中

优选的将各个作业的运行时间设定在夜里或周末，以免给数据源和数据仓库造成压力。用于调用存储过程模块的参数是指在 ETL 任务表的作业名称栏内填写的作业信息模块名称。

[0162] 在本发明的一种优选实施方式中，多个作业信息模块都调用存储过程模块，但输入参数即作业信息模块名称不同，所以不会冲突。在步骤 6 中输入每个 ETL 任务信息时，都需要输入该 ETL 任务所属的作业信息模块名称，例如输入了 ETL 任务 1 至 ETL 任务 6 六个 ETL 任务，其中 ETL 任务 1 至 ETL 任务 3 所属的作业信息模块名称为“作业信息模块 1”，ETL 任务 4 至 ETL 任务 6 所属的作业信息模块名称为“作业信息模块 2”，在步骤 7 中，在目标数据仓库中输入每个作业信息模块时，都需要输入存储过程模块名、存储过程模块的参数名、运行该作业信息模块的时间；其中存储过程模块的参数名也就是作业信息模块名称，例如输入了“作业信息模块 1”和“作业信息模块 2”两个作业信息模块，它们调用的存储过程模块是相同的，即步骤 7 中所述的存储过程模块。但它们运行的 ETL 任务是不同的，作业信息模块 1 只运行 ETL 任务 1 至 ETL 任务 3，作业信息模块 2 只运行 ETL 任务 4 至 ETL 任务 6。它们的运行时间也是不同的，这样做的好处是灵活方便，可以把相关的 ETL 任务放在同一个作业信息模块中运行，不相关的 ETL 任务放在不同的作业信息模块中运行。本发明中所述的作业信息模块包括多个 ETL 任务，其中作业信息模块包括的多个 ETL 任务，是指在 ETL 任务表的作业名称栏内填写了该作业信息模块名称的所有 ETL 任务。

[0163] 在本发明中，存储过程信息模块包括存储过程模块和索引和备注信息存储过程亚模块，在本发明的一个具体的实施例中，将存储过程模块和索引和备注信息存储过程亚模块都设定成存储过程，分别命名为存储过程 A 和存储过程 B，其定义为：存储过程 A 和存储过程 B 都为一组能完成特定功能的 SQL 语句集或流程控制语句集，其经编译后存储在数据库中，用户通过指定存储过程的名称和 / 或参数来执行该存储过程；

[0164] 在本发明的一个具体的实施例中，优选地设定作业信息模块为作业，所述作业是数据库中的一种对象，它可以按照预先设定的计划，定时自动调用一个程序或存储过程，如上述的存储过程 A 等；

[0165] 在一个优选的实施方式中，如图 4 中所示，存储过程 A 的工作流程包括：

[0166] (A-1) 作业调用存储过程 A，其输入参数为作业名称，由接收子模块接收参数信息；

[0167] (A-2) 存储过程 A 开始运行，在 ETL 日志表中记录开始日志，即在开始时间栏内记录存储过程 A 开始运行的时间；

[0168] (A-3) 通过统计子模块查找该作业下属的各个 ETL 任务，按照更新序号排序，并将 ETL 任务表的更新状态设为等待执行；

[0169] (A-4) 循环处理，即依次执行各个 ETL 任务；

[0170] (A-5) 将 ETL 任务表的更新状态设为正在执行，并在 ETL 任务表中记录启动时间；

[0171] (A-6) 根据预设的 ETL 任务的更新方式即从源表到目的表的更新方式，通过更新方式选择子模块选择选择相应的更新流程来更新目的表，每种从源表到目的表的更新方式都有与其相对应的更新流程；

[0172] (A-7) 通过循环处理子模块，将 ETL 任务表的更新状态设为已完成，并记录 ETL 任务结束时间和数据条数，回到 (A-4) 步，执行下一个 ETL 任务，直至该作业下属的所有的 ETL 任务都执行完毕；

[0173] (A-8) 存储过程 A 结束运行，在 ETL 日志表中记录结束日志，即在结束时间栏内记录存储过程 A 运行结束的时间。

[0174] 在一个优选的实施方式中，如图 4 中所示，每个从源表到目的表的更新方式（也称作目的表的更新方式）都有与其相对应的更新流程，从源表到目的表的更新方式包括：全表数据替换方式、全表数据替换并重建索引方式、差异添加方式、时间修改添加或字符串修改添加方式和时间添加方式；其中，

[0175] 目的表的更新方式为全表数据替换时，选用全表数据替换更新子模块，全表数据替换方式的更新流程 501 为：根据源表创建目的表的影子表，所述影子表中只有数据，删除目的表，把影子表重命名为目的表；

[0176] 目的表的更新方式为全表数据替换并重建索引时，选用全表数据替换并重建索引子模块，全表数据替换并重建索引方式的更新流程 502 为：根据源表创建目的表的影子表，所述影子表中只有数据，调用存储过程 B，通过执行存储过程 B 把目的表的索引和备注复制到影子表，然后，删除目的表，把影子表重命名为目的表；

[0177] 目的表的更新方式为差异添加时，选用差异添加子模块，差异添加方式的更新流程 503 为：向目的表中插入源表有而目的表没有的数据；

[0178] 目的表的更新方式为时间修改添加或字符串修改添加时，选用时间修改添加或字符串修改添加子模块，时间修改添加或字符串修改添加方式的更新流程 504 为：删除目的表中数据更新时间与源表中数据更新时间不一致的所有数据，向目的表中插入源表中有而目的表中没有的数据；

[0179] 目的表的更新方式为时间添加时，选用时间添加子模块，时间添加方式的更新流程 505 为：计算源表和目的表最大添加时间，把源表中数据添加时间介于这两个时间之间的数据添加到目的表。对于采用“时间添加”更新方式的源表和目的表来说，一定要有“添加时间字段”，该字段记录了表中每条数据的添加时间。表中各条数据的添加时间的最大值，就是该表的最大添加时间。由于目的表中的数据是从源表中更新来的，因此源表的最大添加时间一定大于或等于目的表的最大添加时间。当源表的最大添加时间等于目的表的最大添加时间时，两个表的数据相同，不需要更新。当源表的最大添加时间大于目的表的最大添加时间时，源表包含一条或多条目的表所没有的新数据。这些新数据的添加时间大于目的表的最大添加时间，小于等于源表的最大添加时间。因此只要把源表中符合此条件的数据添加到目的表中，目的表的数据就与源表相同了。

[0180] 在进一步优选的实施方式中，目的表的更新方式与存储过程 A 的多种更新流程一一对应，更新方式的选择是根据具体情况设定的，由人工输入至 ETL 任务表，更新方式及具体地更新流程包括：

[0181] 全表数据替换更新方式是指把目的表的数据全部替换成源表的数据，但不创建目的表的索引和备注，此种更新方式适用于从远程数据库复制一份数据到本地数据库，此时得到的目的表是中间表，不是应用程序访问的表；

[0182] 全表数据替换并重建索引更新方式是指把目的表的数据全部替换成源表的数据，且重建目的表的索引和备注，此种更新方式适用于对数据进行转化，转化后得到的目的表是应用程序可用的表；其中，如果目的表是百万行以下的小表，那么不管源表是在本地还是在远程，都只需运行一个全表数据替换并重建索引的任务即可，如果目的表是一个百万行

至上亿行的大表，那么需要先运行一个全表数据替换的任务，用最短的时间把远程数据库中的大表复制到本地，得到一个中间表；再运行一个全表数据替换并重建索引的任务，把中间表转化成应用程序可用的表，这样对远程数据库的访问时间短，整个 ETL 时间也短；

[0183] 差异添加更新方式是指把源表中有而目的表中没有的数据添加到目的表，此种更新方式适用于目的表中的已有数据不希望被源表删除或修改，只希望从源表中获取新增数据的场景；此种方式要求源表和目的表都有主键；其中，计算源表和目的表差异的方法采用左连接而不采用 not in，使得速度达到最快；

[0184] 时间修改添加或字符串修改添加更新方式包括时间修改添加更新方式和字符串修改添加更新方式；时间修改添加更新方式是指以源表和目的表中的更新时间字段为参照，对于同一条数据，如果源表和目的表的更新时间不一致就按照源表修改目的表，如果源表有而目的表没有就添加到目的表，此种更新方式适用于数据总量较大而添加修改量较小的情况，此种更新方式要求源表和目的表都有主键和更新时间字段；

[0185] 字符串修改添加更新方式与时间修改添加更新方式基本一致。不同之处在于：时间修改添加中的更新时间字段的数据类型是日期型，而字符串修改添加中的更新时间字段的数据类型是字符串型；

[0186] 时间添加是指以源表和目的表中的添加时间字段为参照，先计算出源表和目的表的最大添加时间，再把源表中介于这个两个时间点之间的数据添加到目的表；此种更新方式适用于数据总量较大而添加量较小，且没有主键的情况，此种方式要求源表和目的表都有添加时间字段。

[0187] 在更进一步优选的实施方式中，全表数据替换和全表数据替换并重建索引两种更新方式均不采用 delete、update、merge、truncate 等语句删除或修改目的表，它们采用的方式是：先根据源表创建一个目的表的影子表，影子表创建完成后，再把目的表删除，把影子表重命名为目的表，此种方式使得 ETL 时间短；更新过程中不会影响应用程序对表的访问；不会造成数据碎片和索引碎片，使表的性能保持最高；不管表有无主键均可采用此种方式，适用范围广。

[0188] 在一个优选的实施方式中，存储过程 B 用于把目的表的索引和备注复制到影子表，存储过程 B 的输入参数是目的表表名和影子表表名，存储过程 B 不单独运行，它仅在存储过程 A 调用时运行；如图 5 中所示，存储过程 B 的工作流程包括：

[0189] (B-1) 存储过程 A 调用存储过程 B，输入参数为目的表表名和影子表表名；

[0190] (B-2) 存储过程 B 开始运行，在 ETL 日志表中记录开始日志即在 ETL 日志表的开始时间栏内记录存储过程 B 的运行开始时间；

[0191] (B-3) 检查影子表是否有索引，并删掉索引；

[0192] (B-4) 查找目的表的各个索引；

[0193] (B-5) 进入循环，参照目的表索引依次建立影子表索引；

[0194] (B-6) 查询目的表索引的类型；

[0195] (B-7) 查询目的表索引的字段；

[0196] (B-8) 查询目的表的索引是否为函数索引；

[0197] (B-9) 根据上述查询到的信息创建影子表索引；

[0198] (B-10) 回到 (B-5) 步，建立下一个影子表索引，直至所有目的表索引都建立有对

应的影子表索引；

[0199] (B-11) 将目的表的表备注复制到影子表；

[0200] (B-12) 将目的表的字段备注复制到影子表；

[0201] (B-13) 存储过程 B 结束运行，在 ETL 日志表中记录结束日志即在 ETL 日志表的结束时间栏内记载存储过程 B 结束运行的时间。

[0202] 在一个优选的实施方式中，作业是数据库中的一种对象，它可以按照预先设定的计划，定时自动调用一个程序或存储过程，如存储过程 A 等，调用存储过程 A 的输入参数就是作业名称，作业可以设置多个，虽然都是调用存储过程 A，但输入参数即作业名称不同，所以不会冲突；在本发明中，预定的时间是指每天的某个时间点和 / 或每经过一个预定的时间段以后可以理解为操作者所期望的任意一个时间点或时间段，并且可以多次重复开始，即多次调用存储过程 A，本发明中优选的将各个作业的运行时间设定在夜里或周末，以免给数据源和数据仓库造成压力。

[0203] 实施例

[0204] 下面结合具体实例对本发明提供的 ETL 方法过程进行说明，以目的表的更新方式为“全表数据替换并重建索引”的 ETL 任务为具体实例：

[0205] 设有 S、D 两台数据库服务器，S 为远程数据库，D 为目标数据仓库，在 S 中有一张人员表，该表包含 3 个字段：身份证号、姓名、性别代码，其中身份证号是主键，性别代码中存储的是 1 和 2，1 代表男性，2 代表女性。人员表数据条数约为一千万，而且每天都会添加新数据，还会删除或修改一部分旧数据。现在需要把 S 的人员表导入到 D 中并每天自动更新，而且要把性别代码转化为有意义的男和女。为实现此任务，可按以下步骤实施：

[0206] 步骤 1：在目标数据仓库 D 中建立访问远程数据库 S 的数据库连接。该数据库连接只需建立一次，以后可以反复使用。

[0207] 步骤 2：在 D 中建立空的人员表，该表包含 4 个字段：身份证号、姓名、性别代码、性别名称。其中前 3 个字段与 S 中人员表的 3 个字段相同，而性别名称是新增字段，其中存储根据性别代码转化而来的性别名称，即 1 转化为男，2 转化为女。由于要对身份证号和姓名进行查询，因此对这两个字段建立索引。

[0208] 步骤 3：在 D 中建立把 S 中的人员表转化为 D 中的人员表的转化视图，也可称作人员视图。人员视图能够读取 S 中人员表的 3 个字段，并输出与 D 中的人员表相同的 4 个字段。人员视图本身不包含数据，它只起到数据转化的作用。

[0209] 步骤 4：在 D 中建立 ETL 任务表和 ETL 日志表。ETL 任务表和 ETL 日志表只需建立一次，以后可以反复使用，ETL 任务表和 ETL 日志表的表结构如上述实施方式中所述。

[0210] 步骤 5：在 D 中输入用于执行 ETL 过程的存储过程信息模块，存储过程信息模块包括存储过程模块和索引和备注信息存储过程亚模块，优选的设定为存储过程 A 和存储过程 B。存储过程信息模块即存储过程 A 和存储过程 B 可以反复使用。

[0211] 步骤 6：在 ETL 任务表中添加 ETL 任务，即在 ETL 任务表中添加一条数据：目的表 = D 中的人员表，源表 = 人员视图，更新方式 = 全表数据替换并重建索引，作业名称 = 人员表更新作业，更新序号 = 1。其中，由于更新方式为“全表数据替换并重建索引”，所以不需要输入主键名和增量字段名，源表中填写的不是“S 中的人员表”而是“人员视图”，这是因为 S 中的人员表需要经过人员视图转化，在这个 ETL 任务中，转化视图即人员视图作为源表

使用；如果源表不需要经过视图转化，那么可以直接填写“S 中的人员表”。

[0212] 步骤 7：在 D 中设置作业信息模块，优选地在 D 中设置作业，作业信息模块的名称为“人员表更新作业信息模块”，即作业的名称为“人员表更新作业”，每天晚上 11 点作业定时调用存储过程 A，存储过程 A 的输入参数就是作业名称“人员表更新作业”。

[0213] 至此，人员表定时更新的操作过程就完成了。每天晚上 11 点，D 中的“人员表更新作业”会自动调用存储过程 A，存储过程 A 又会调用存储过程 B，它们会在 D 中创建一个经过人员视图转化后的影子人员表，这个表也就是转化后有 4 个字段的、最新的人员表。影子人员表创建完成以后会建立索引，并删除 D 中旧的人员表，最后把影子人员表重命名为人员表。这样就把 S 中的人员表全部转化并更新到了 D 中的人员表，而且又不影响应用程序访问人员表。

[0214] 以上结合了优选的实施方式对本发明进行了说明，不过这些实施方式仅是范例性的，仅起到说明性的作用。在此基础上，可以对本发明进行多种替换和改进，这些均落入本发明的保护范围内。

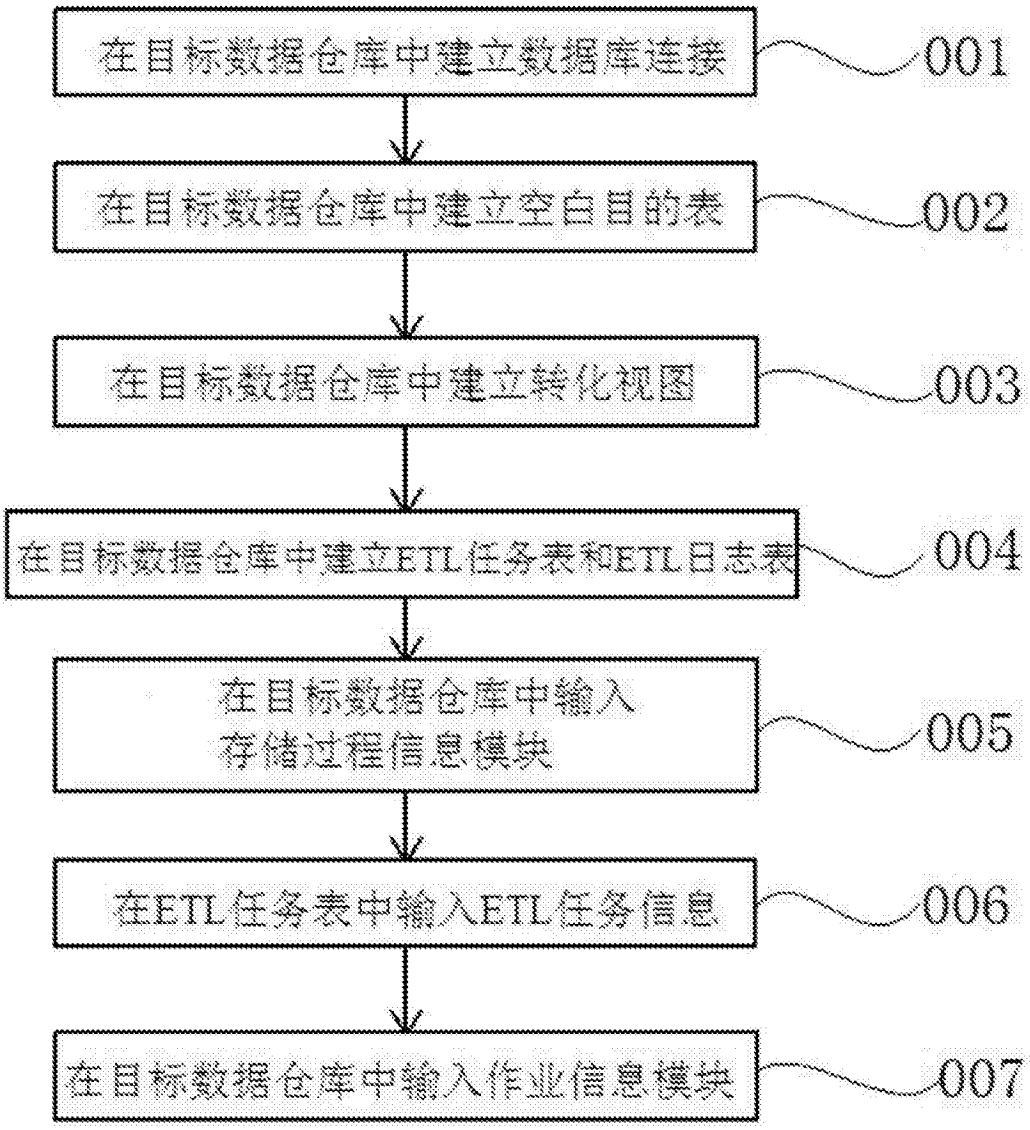

在目标数据仓库中建立数据库连接
001
在目标数据仓库中建立空白目的表
002
在目标数据仓库中建立转化视图
003
在目标数据仓库中建立ETL任务表和ETL日志表
004
在目标数据仓库中输入
存储过程信息模块
005
在ETL任务表中输入ETL任务信息
006
在目标数据仓库中输入作业信息模块
007

图 1

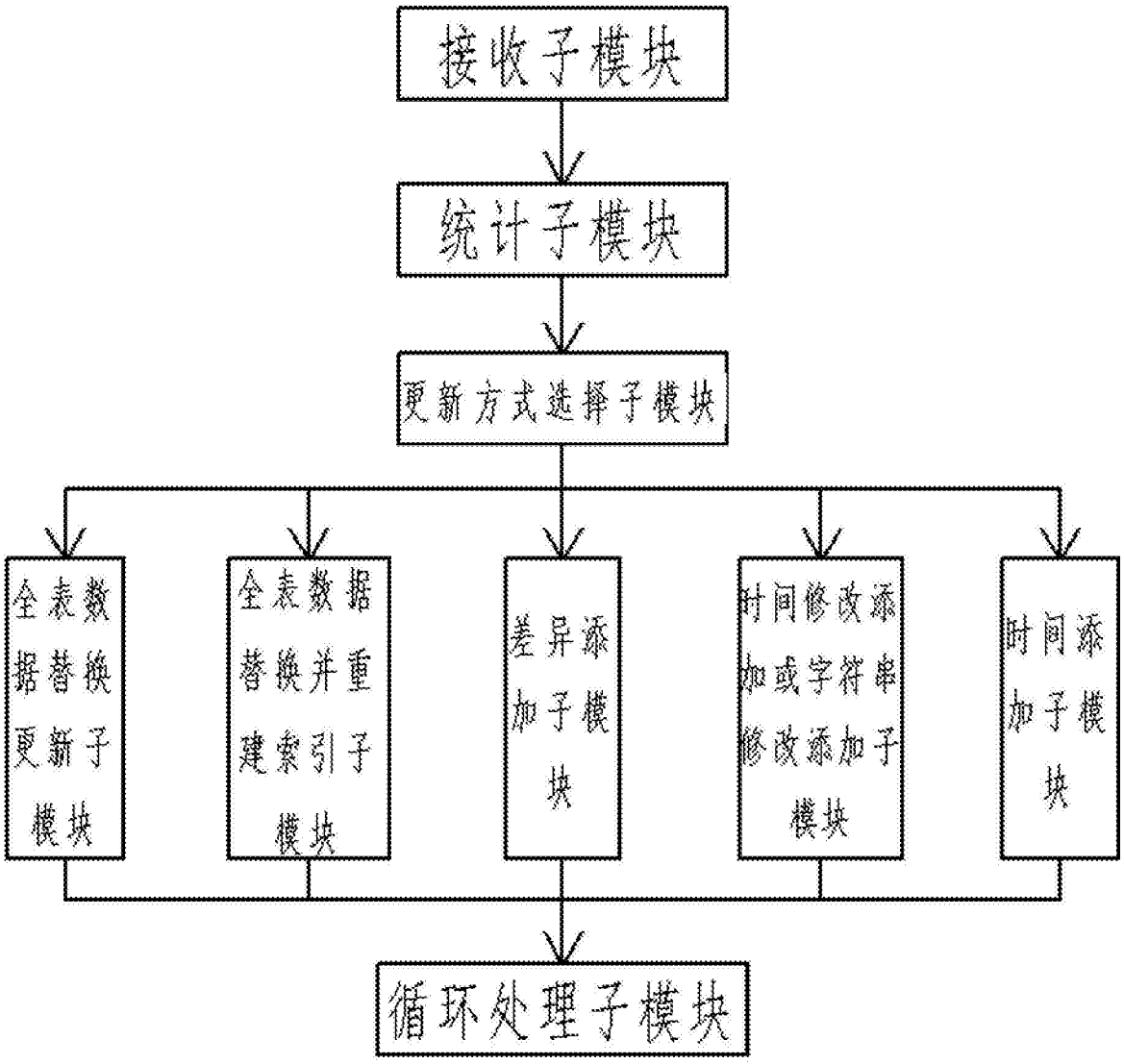

接收子模块
统计子模块
更新方式选择子模块
全表数据替换更新子模块
全表数据替换并重建索引子模块
差异添加子模块
时间修改或字符串修改添加子模块
时间添加子模块
循环处理子模块

图 2

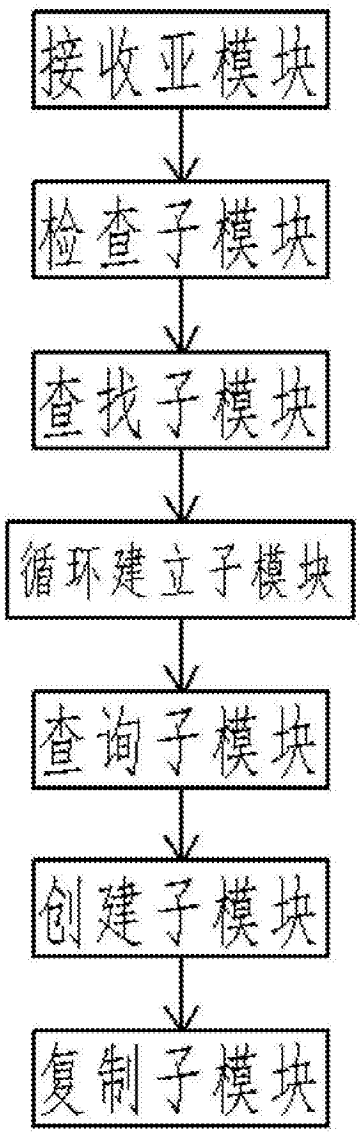


图 3

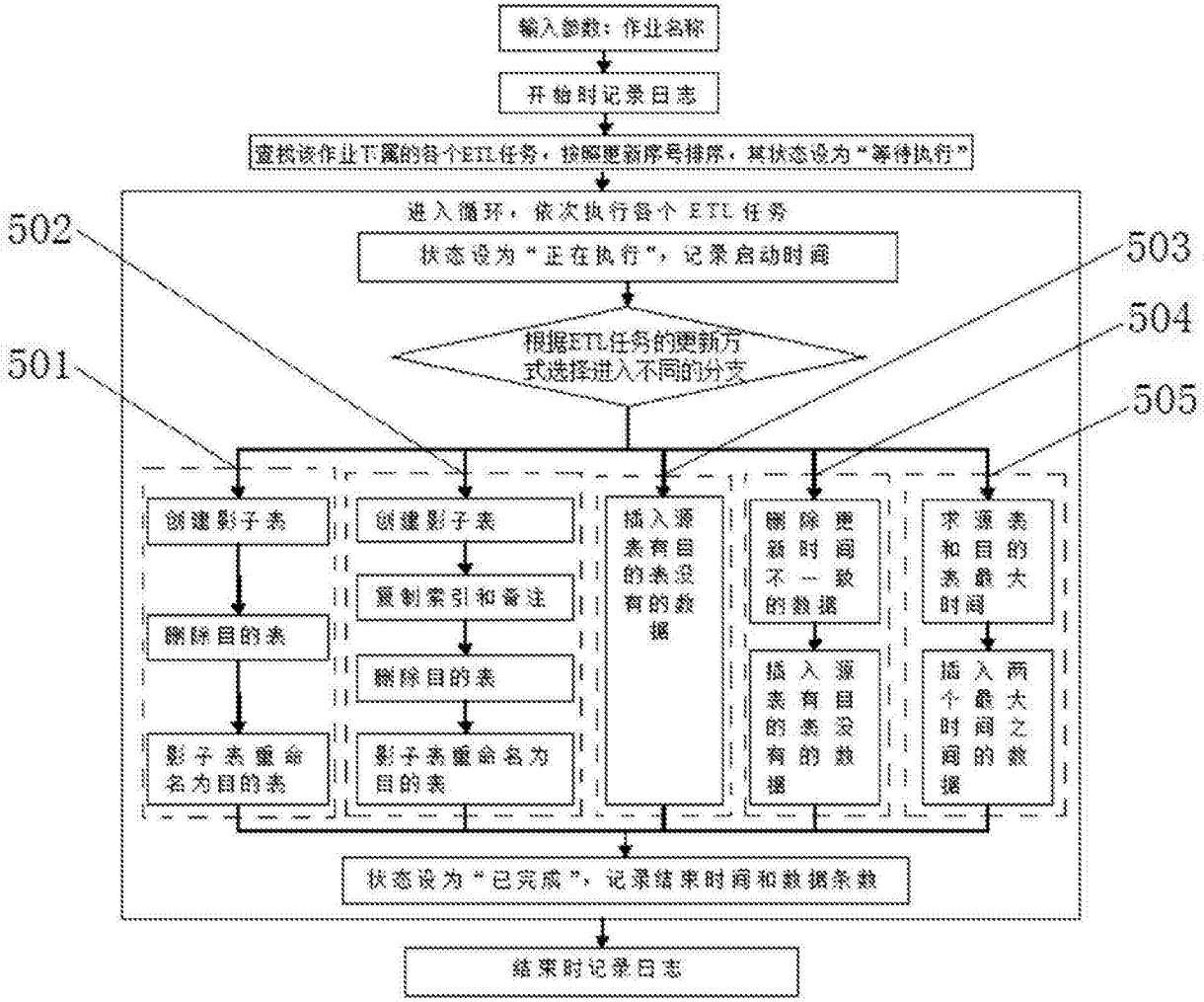

输入参数：作业名称
开始时记录日志
查找该作业下属的各个ETL任务，按照更新序号排序，其状态设为“等待执行”
进入循环，依次执行各个 ETL 任务
状态设为“正在执行”，记录启动时间
根据ETL任务的更新方式选择进入不同的分支
502
503
504
501
505
创建影子表
删除目的表
影子表重命名为目的表
创建影子表
复制索引和备注
删除目的表
影子表重命名为目的表
插入源表有目的表没有的数据
删除更新时间不一致的数据
插入源表有目的表没有的数据
求源表和目的表最大时间
插入两个最大时间之间的数据
状态设为“已完成”，记录结束时间和数据条数
结束时记录日志

图 4

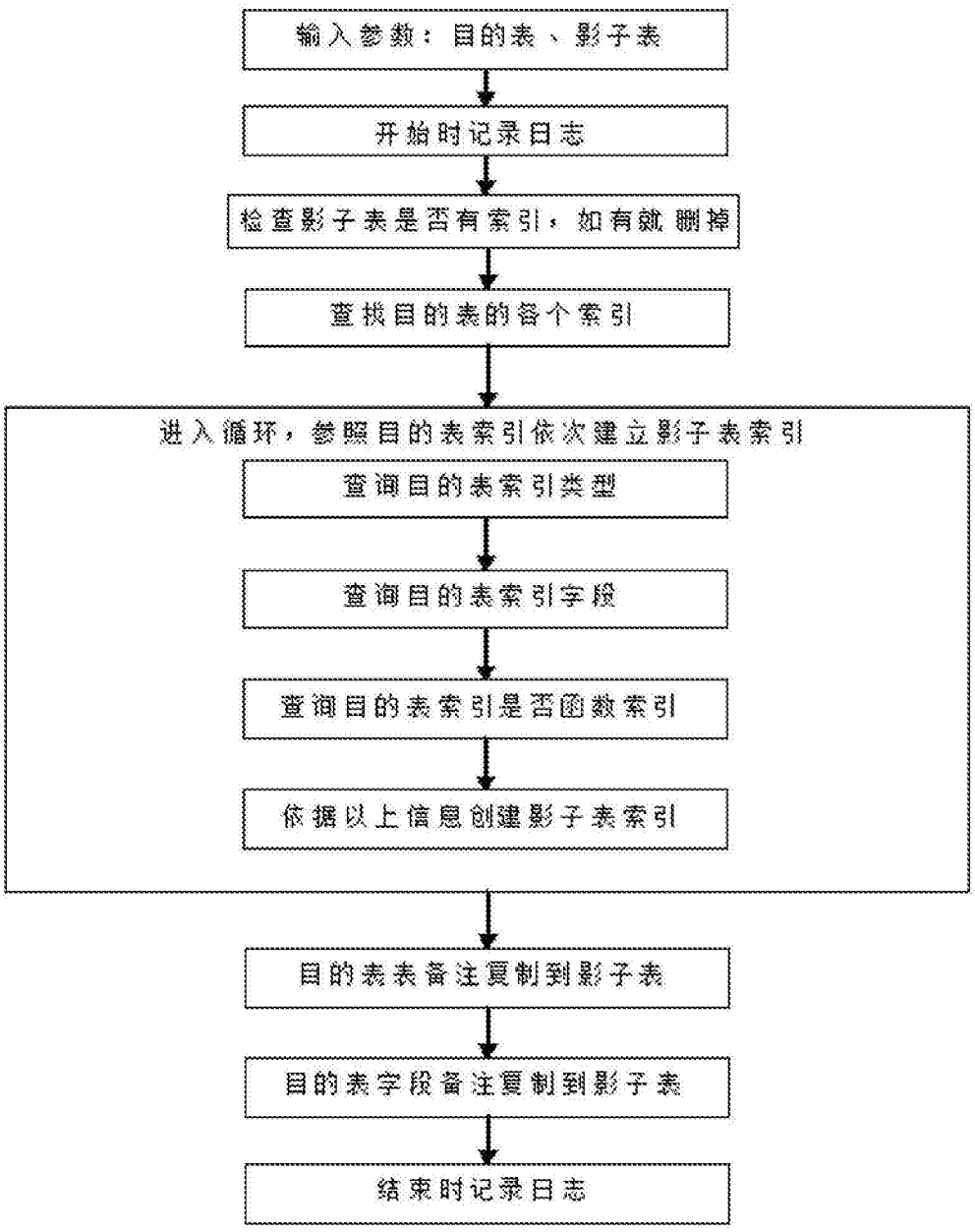

输入参数：目的表、影子表
开始时记录日志
检查影子表是否有索引，如有就删掉
查找目的表的各个索引
进入循环，参照目的表索引依次建立影子表索引
查询目的表索引类型
查询目的表索引字段
查询目的表索引是否函数索引
依据以上信息创建影子表索引
目的表表备注复制到影子表
目的表字段备注复制到影子表
结束时记录日志

图 5