



(12)发明专利

(10)授权公告号 CN 103902653 B

(45)授权公告日 2017.08.01

(21)申请号 201410072773.0

(22)申请日 2014.02.28

(65)同一申请的已公布的文献号
申请公布号 CN 103902653 A

(43)申请公布日 2014.07.02

(73)专利权人 珠海多玩信息技术有限公司
地址 519080 广东省珠海市唐家湾镇软件园路1号会展中心1-第13层B区

(72)发明人 陈武 刘超洪

(74)专利代理机构 北京三高永信知识产权代理有限公司 11138
代理人 江崇玉

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

(56)对比文件

CN 101859303 A,2010.10.13,
CN 103186541 A,2013.07.03,
CN 102239458 A,2011.11.09,
CN 101609473 A,2009.12.23,
US 8468120 B2,2013.06.18,
杨玢玢.“数据仓库元数据的管理与运用”.
《中国优秀硕士学位论文全文数据库 信息科技辑》.2011,

衡铁刚.“面向疑点核实的数据路径追踪技术研究”.《中国优秀硕士学位论文全文数据库 信息科技辑》.2012,

审查员 王洋洋

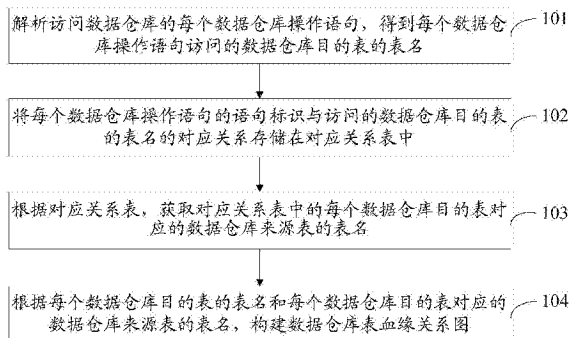
权利要求书2页 说明书9页 附图3页

(54)发明名称

一种构建数据仓库表血缘关系图的方法和装置

(57)摘要

本发明公开了一种构建数据仓库表血缘关系图的方法和装置,属于计算机领域。该方法包括:解析访问数据仓库的每个数据仓库操作语句,得到每个数据仓库操作语句访问的数据仓库目的表的表名;将每个数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中;根据对应关系表,获取对应关系表中的每个数据仓库目的表对应的数据仓库来源表的表名;根据每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图。该装置包括:解析模块、第一存储模块、第一获取模块和构建模块。本发明中服务器能够自动构建数据仓库血缘关系图。



1. 一种构建数据仓库表血缘关系图的方法,其特征在于,所述方法包括:

解析访问数据仓库的每个数据仓库操作语句,得到所述每个数据仓库操作语句访问的数据仓库目的表的表名;

获取任务类型为导入类型的数据仓库操作语句和与之对应的导入路径;

根据所述导入路径获取任务类型为分析类型且具有所述导入路径的数据仓库操作语句;

绑定所述任务类型为导入类型的数据仓库操作语句和所述任务类型为分析类型且具有所述导入路径的数据仓库操作语句;

将所述任务类型为导入类型的数据仓库操作语句的语句标识和所述任务类型为分析类型且具有所述导入路径的数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中;

对于所述对应关系表中的每条记录,获取所述记录中存储的数据仓库操作语句的语句标识和数据仓库目的表的表名;

根据所述获取的语句标识获取数据仓库操作语句;

解析所述获取的数据仓库操作语句,得到所述每个数据仓库目的表对应的数据仓库来源表的表名;

根据每个数据仓库目的表的表名和所述每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图。

2. 如权利要求1所述的方法,其特征在于,所述解析访问数据仓库的每个数据仓库操作语句,得到所述每个数据仓库操作语句访问的数据仓库目的表的表名,包括:

解析所述访问数据仓库的每个数据仓库操作语句,得到所述访问数据仓库的每个数据仓库操作语句对应的访问方式;

获取访问方式为写方式的数据仓库操作语句;

解析所述访问方式为写方式的数据仓库操作语句,得到所述访问方式为写方式的数据仓库操作语句访问的所有数据仓库目的表的表名。

3. 如权利要求1所述的方法,其特征在于,所述根据每个数据仓库目的表的表名和所述每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图,包括:

在数据仓库表血缘关系图中,构建所述数据仓库目的表的表名对应的节点,并构建所述数据仓库目的表对应的数据仓库来源表的表名对应的节点;

将所述数据仓库目的表的表名对应的节点作为所述数据仓库目的表对应的数据仓库来源表的表名对应的节点的子节点。

4. 如权利要求3所述的方法,其特征在于,所述构建所述数据仓库目的表的表名对应的节点之后,所述方法还包括:

将访问所述数据仓库目的表的数据仓库操作语句存储在所述数据仓库目的表的表名对应的节点中;

将所述数据仓库表血缘关系图发送给终端,由所述终端显示给用户。

5. 一种构建数据仓库表血缘关系图的装置,其特征在于,所述装置包括:

解析模块,用于解析访问数据仓库的每个数据仓库操作语句,得到所述每个数据仓库操作语句访问的数据仓库目的表的表名;

第二获取模块,用于获取任务类型为导入类型的数据仓库操作语句和与之对应的导入路径;

第三获取模块,用于根据所述导入路径获取任务类型为分析类型且具有所述导入路径的数据仓库操作语句;

绑定模块,用于绑定所述任务类型为导入类型的数据仓库操作语句和所述任务类型为分析类型且具有所述导入路径的数据仓库操作语句;

第一存储模块,用于将所述任务类型为导入类型的数据仓库操作语句的语句标识和所述任务类型为分析类型且具有所述导入路径的数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中;

第一获取模块,用于根据所述对应关系表,获取所述对应关系表中的每个数据仓库目的表对应的数据仓库来源表的表名;

构建模块,用于根据每个数据仓库目的表的表名和所述每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图;

所述第一获取模块,包括:

第一获取单元,用于对于所述对应关系表中的每条记录,获取所述记录中存储的数据仓库操作语句的语句标识和数据仓库目的表的表名;

第二获取单元,用于根据所述获取的语句标识获取数据仓库操作语句;

第三解析单元,用于解析所述获取的数据仓库操作语句,得到所述每个数据仓库目的表对应的数据仓库来源表的表名。

6.如权利要求5所述的装置,其特征在于,所述解析模块,包括:

第一解析单元,用于解析所述访问数据仓库的每个数据仓库操作语句,得到所述访问数据仓库的每个数据仓库操作语句对应的访问方式;

获取单元,用于获取访问方式为写方式的数据仓库操作语句;

第二解析单元,用于解析所述访问方式为写方式的数据仓库操作语句,得到所述访问方式为写方式的数据仓库操作语句访问的所有数据仓库目的表的表名。

7.如权利要求5所述的装置,其特征在于,所述构建模块,包括:

构建单元,用于在数据仓库表血缘关系图中,构建所述数据仓库目的表的表名对应的节点,并构建所述数据仓库目的表对应的数据仓库来源表的表名对应的节点;

作为单元,用于将所述数据仓库目的表的表名对应的节点作为所述数据仓库目的表对应的数据仓库来源表的表名对应的节点的子节点。

8.如权利要求7所述的装置,其特征在于,所述装置还包括:

第二存储模块,用于将访问所述数据仓库目的表的数据仓库操作语句存储在所述数据仓库目的表的表名对应的节点中;

发送模块,用于将所述数据仓库表血缘关系图发送给终端,由所述终端显示给用户。

一种构建数据仓库表血缘关系图的方法和装置

技术领域

[0001] 本发明涉及计算机领域,特别涉及一种构建数据仓库表血缘关系图的方法和装置。

背景技术

[0002] 数据仓库中存储有各种各样的业务数据,不同的业务数据存储在不同的业务表中。因此,数据仓库中存储有多个业务表,如何将数据仓库中存储的多个业务表构建成数据仓库表血缘关系图是迫切需要解决的问题。

[0003] 目前,都是数据仓库管理人员解析数据仓库操作语句并构建数据仓库表血缘关系图。而数据仓库管理人员构建数据仓库表血缘关系图时,容易出错;并且,数据仓库中的业务数据量很大,造成数据仓库管理人员的工作量大。

发明内容

[0004] 为了解决现有技术的问题,本发明提供了一种构建数据仓库表血缘关系图的方法和装置。所述技术方案如下:

[0005] 一方面,提供了一种构建数据仓库表血缘关系图的方法,所述方法包括:

[0006] 解析访问数据仓库的每个数据仓库操作语句,得到所述每个数据仓库操作语句访问的数据仓库目的表的表名;

[0007] 将所述每个数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中;

[0008] 根据所述对应关系表,获取所述对应关系表中的每个数据仓库目的表对应的数据仓库来源表的表名;

[0009] 根据每个数据仓库目的表的表名和所述每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图。

[0010] 进一步地,所述解析访问数据仓库的每个数据仓库操作语句,得到所述每个数据仓库操作语句访问的数据仓库目的表的表名,包括:

[0011] 解析所述访问数据仓库的每个数据仓库操作语句,得到所述访问数据仓库的每个数据仓库操作语句对应的访问方式;

[0012] 获取访问方式为写方式的数据仓库操作语句;

[0013] 解析所述访问方式为写方式的数据仓库操作语句,得到所述访问方式为写方式的访问的所有数据仓库目的表的表名。

[0014] 进一步地,所述解析访问数据仓库的每个数据仓库操作语句,得到所述每个数据仓库操作语句访问的数据仓库目的表的表名之后,所述方法还包括:

[0015] 获取任务类型为导入类型的数据仓库操作语句和与之对应的导入路径;

[0016] 根据所述导入路径获取任务类型为分析类型且具有所述导入路径的数据仓库操作语句;

[0017] 绑定所述任务类型为导入类型的数据仓库操作语句和所述任务类型为分析类型且具有所述导入路径的数据仓库操作语句。

[0018] 进一步地,所述根据所述对应关系表,获取所述对应关系表中的每个数据仓库目的表对应的数据仓库来源表的表名,包括:

[0019] 对于所述对应关系表中的每条记录,获取所述记录中存储的数据仓库操作语句的语句标识和数据仓库目的表的表名;

[0020] 根据所述获取的语句标识获取数据仓库操作语句;

[0021] 解析所述获取的数据仓库操作语句,得到所述每个数据仓库目的表对应的数据仓库来源表的表名。

[0022] 进一步地,所述根据每个数据仓库目的表的表名和所述每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图,包括:

[0023] 在数据仓库表血缘关系图中,构建所述数据仓库目的表的表名对应的节点,并构建所述数据仓库目的表对应的数据仓库来源表的表名对应的节点;

[0024] 将所述数据仓库目的表的表名对应的节点作为所述数据仓库目的表对应的数据仓库来源表的表名对应的节点的子节点。

[0025] 进一步地,所述构建所述数据仓库目的表的表名对应的节点之后,所述方法还包括:

[0026] 将访问所述数据仓库目的表的数据仓库操作语句存储在所述数据仓库目的表的表名对应的节点中;

[0027] 将所述数据仓库表血缘关系图发送给终端,由所述终端显示给用户。

[0028] 另一方面,本发明提供了一种构建数据仓库表血缘关系图的装置,所述装置包括:

[0029] 解析模块,用于解析访问数据仓库的每个数据仓库操作语句,得到所述每个数据仓库操作语句访问的数据仓库目的表的表名;

[0030] 第一存储模块,用于将所述每个数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中;

[0031] 第一获取模块,用于根据所述对应关系表,获取所述对应关系表中的每个数据仓库目的表对应的数据仓库来源表的表名;

[0032] 构建模块,用于根据每个数据仓库目的表的表名和所述每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图。

[0033] 进一步地,所述解析模块,包括:

[0034] 第一解析单元,用于解析所述访问数据仓库的每个数据仓库操作语句,得到所述访问数据仓库的每个数据仓库操作语句对应的访问方式;

[0035] 获取单元,用于获取访问方式为写方式的数据仓库操作语句;

[0036] 第二解析单元,用于解析所述访问方式为写方式的数据仓库操作语句,得到所述访问方式为写方式的访问的所有数据仓库目的表的表名。

[0037] 进一步地,所述装置还包括:

[0038] 第二获取模块,用于获取任务类型为导入类型的数据仓库操作语句和与之对应的导入路径;

[0039] 第三获取模块,用于根据所述导入路径获取任务类型为分析类型且具有所述导入

路径的数据仓库操作语句；

[0040] 绑定模块,用于绑定所述任务类型为导入类型的数据仓库操作语句和所述任务类型为分析类型且具有所述导入路径的数据仓库操作语句。

[0041] 进一步地,所述第一获取模块,包括:

[0042] 第一获取单元,用于对于所述对应关系表中的每条记录,获取所述记录中存储的数据仓库操作语句的语句标识和数据仓库目的表的表名;

[0043] 第二获取单元,用于根据所述获取的语句标识获取数据仓库操作语句;

[0044] 第三解析单元,用于解析所述获取的数据仓库操作语句,得到所述每个数据仓库目的表对应的数据仓库来源表的表名。

[0045] 进一步地,所述构建模块,包括:

[0046] 构建单元,用于在数据仓库表血缘关系图中,构建所述数据仓库目的表的表名对应的节点,并构建所述数据仓库目的表对应的数据仓库来源表的表名对应的节点;

[0047] 作为单元,用于将所述数据仓库目的表的表名对应的节点作为所述数据仓库目的表对应的数据仓库来源表的表名对应的节点的子节点。

[0048] 进一步地,所述装置还包括:

[0049] 第二存储模块,用于将访问所述数据仓库目的表的数据仓库操作语句存储在所述数据仓库目的表的表名对应的节点中;

[0050] 发送模块,用于将所述数据仓库表血缘关系图发送给终端,由所述终端显示给用户。

[0051] 在本发明实施例中,服务器解析访问数据仓库的每个数据仓库操作语句,得到每个数据仓库操作语句访问的数据仓库目的表的表名和每个数据仓库目的表对应的源数据仓库库表的表名,并根据每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名,自动构建数据仓库表血缘关系图,减少了人工工作量,并且,提高了构建数据仓库表血缘关系的速度以及准确度。

附图说明

[0052] 为了更清楚地说明本发明实施例中的技术方案,下面将对实施例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的前提下,还可以根据这些附图获得其他的附图。

[0053] 图1是本发明实施例1提供的一种构建数据仓库表血缘关系图的方法流程图;

[0054] 图2是本发明实施例2提供的一种构建数据仓库表血缘关系图的方法流程图;

[0055] 图3是本发明实施例3提供的一种构建数据仓库表血缘关系图的装置结构示意图。

具体实施方式

[0056] 为使本发明的目的、技术方案和优点更加清楚,下面将结合附图对本发明实施方式作进一步地详细描述。

[0057] 实施例1

[0058] 本发明实施例提供了一种构建数据仓库表血缘关系图的方法。参见图1,其中,该

方法包括：

[0059] 步骤101:解析访问数据仓库的每个数据仓库操作语句,得到每个数据仓库操作语句访问的数据仓库目的表的表名;

[0060] 步骤102:将每个数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中;

[0061] 步骤103:根据对应关系表,获取对应关系表中的每个数据仓库目的表对应的数据仓库来源表的表名;

[0062] 步骤104:根据每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图。

[0063] 在本发明实施例中,服务器解析访问数据仓库的每个数据仓库操作语句,得到每个数据仓库操作语句访问的数据仓库目的表的表名和每个数据仓库目的表对应的源数据仓库库表的表名,并根据每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名,自动构建数据仓库表血缘关系图,减少了人工工作量,并且,提高了构建数据仓库表血缘关系的速度以及准确度。

[0064] 实施例2

[0065] 本发明实施例提供了一种构建数据仓库表血缘关系图的方法。参见图2其中,该方法包括:

[0066] 步骤201:服务器从数据仓库中获取每一个业务点的访问数据仓库的数据仓库操作语句;

[0067] 具体地,服务器根据业务类型,从数据仓库中获取属于同一业务类型的每一个业务点的访问数据仓库的数据仓库操作语句,并为每一个数据仓库操作语句分配语句标识。

[0068] 其中,服务器中存储有业务类型与访问数据仓库的数据仓库操作语句的对应关系,根据业务类型可以从业务类型与访问数据仓库的数据仓库操作语句的对应关系中,获取与业务类型对应的每一个业务点的访问数据仓库的数据仓库操作语句。

[0069] 其中,需要说明的是,在数据仓库中都是通过访问数据仓库的数据仓库操作语句对数据仓库中的数据进行操作的。访问数据仓库的数据仓库操作语句可以为:

[0070] insert overwrite table hive_table_b

[0071] Select*

[0072] From hive_table_a

[0073]

[0074] 步骤202:服务器解析访问数据仓库的每个数据仓库操作语句,得到每个数据仓库操作语句访问的数据仓库目的表的表名;

[0075] 具体地,服务器解析访问数据仓库的每个数据仓库操作语句,得到访问数据仓库的每个数据仓库操作语句对应的访问方式,该访问方式包括读方式和写方式,写方式包括写数据至数据仓库以及写数据到本地文件。服务器获取访问方式为写方式的每个数据仓库操作语句,并通过第一正则表达式匹配规则解析访问方式为写方式的每个数据仓库操作语句,搜索出每个数据仓库操作语句访问的数据仓库目的表的表名。

[0076] 其中,第一正则表达式匹配规则如下:

[0077] (1)、在字符“insert into”后面,且在“(”前面的一个不包含空白字符和“(”的字

字符串；

[0078] (2)、在字符“replace into”后面,且在字符“(”前面的一个不包含空白字符和“(”的字符串；

[0079] (3)、在字符“insert overwrite table”后面的一个不包含空白字符的字符串；

[0080] (4)、在字符“overwrite into table”后面的一个不包含空白字符的字符串。

[0081] 其中,需要说明的是,以上规则是并行关系,只要任一规则成立即可,并且,在本发明实施例中,所有字符都不区分字母大小写。

[0082] 其中,第一正则表达式代码为:

[0083] (?i) insert\\s+into\\s+([^\s\[\]]+)\s*\s*(| (?i) replace\\s+into\\s+([^\s\[\]]+)\s*\s*(| (?i) insert\\s+overwrite\\s+table\\s+(\S+)\s+| (?i) overwrite\\s+into\\s+table\\s+(\S+)\s+.

[0084] 其中,需要说明的是,数据仓库中存储有各种各样的业务数据,服务器通过把这些海量级的业务数据加载至数据仓库hive层源头表,再根据不同的业务,对hive层源头表进行不同的抽取、清洗和转换,得到各种各样的hive层业务表。服务器再基于这些hive层业务表,根据更细层的业务需要,进行无数次的更细层地子拆分,得到更多的hive层业务表,方便不同的业务分析统计。并且,服务器还可以根据不同的业务,将hive层业务表中的业务数据进行分析统计后,导入mysql层业务表中,做网页展示。某些特殊情况,还可以对mysql层业务表的业务数据做进一步地抽取计算,以更方便做网页展示。

[0085] 其中,通过上述正则表达式匹配规则步骤(1)和(2)可以搜索出mysql层的表名,步骤(3)和(4)可以搜索出hive层的表名。

[0086] 步骤203:服务器解析访问数据仓库的每个数据仓库操作语句,得到访问数据仓库的每个数据仓库操作语句的任务类型;

[0087] 其中,任务类型包括:分析、导入和抽取。Hive层业务拆分和hive层业务分析属于分析类型;hive层业务表中的业务数据导入mysql层业务表属于导入类型;mysql层业务表的业务数据抽取计算属于抽取类型。

[0088] 步骤204:服务器获取任务类型为导入类型的数据仓库操作语句和与之对应的导入路径;

[0089] 其中,如果任务类型为导入类型时,服务器中会存储数据仓库操作语句和导入路径的对应关系,根据任务类型可以获取导入类型的数据仓库操作语句和与之对应的导入路径。

[0090] 例如,任务类型为导入类型的数据仓库操作语句为:

[0091] Insert into mysql_table_a (key1,col1,col2)

[0092] values (:key1,:col1,:col2)

[0093] ON DUPLICATE KEY UPDATE

[0094] col1=values (col1) ,

[0095] col2=values (col2) ;

[0096] 与上述数据库操作语句对应的导入路径为“/path1/path2/”。

[0097] 进一步地,服务器读取导入路径为“/path1/path2/”目录下所有的文本内容,执行以下操作语句:

[0098] Insert into mysql_table_a (key1,col1,col2)

[0099] values (:key1, :col1, :col2)

[0100] ON DUPLICATE KEY UPDATE

[0101] col1=values (col1) ,

[0102] col2=values (col2) 。

[0103] 步骤205:服务器根据该导入路径获取任务类型为分析类型且具有该导入路径的数据仓库操作语句;

[0104] 具体地,服务器获取任务类型为分析类型的数据仓库操作语句,并从获取的任务类型为分析类型的数据仓库操作语句中获取和步骤204中导入类型的数据仓库操作语句具有相同导入路径的数据仓库操作语句。

[0105] 其中,任务类型为导入类型的数据仓库操作语句的数据仓库目的表,会对应至少两个数据仓库操作语句。其中一个数据仓库操作语句为当前任务类型为导入类型的数据仓库操作语句,另外的数据仓库操作语句是和任务类型为导入类型的数据仓库操作语句具有相同导入路径的分析类型的数据仓库操作语句。

[0106] 例如,任务类型为分析类型且和导入类型的数据仓库操作语句具有相同导入路径的数据仓库操作语句为:

[0107] insert overwrite local directory '/path1/path2/path3/'

[0108] select count (1)

[0109] from hive_table_a

[0110] ...

[0111] 步骤206:服务器绑定任务类型为导入类型的数据仓库操作语句和任务类型为分析类型且具有该导入路径的数据仓库操作语句;

[0112] 其中,服务器绑定任务类型为导入类型的数据仓库操作语句和任务类型为分析类型且具有步骤204中的导入路径的数据仓库操作语句,从而可以建立起mysql层表名和hive层表名的关联关系。

[0113] 步骤207:服务器将每个数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中;

[0114] 具体地,服务器将访问的数据仓库目的表以及与数据仓库目的表对应的任务类型为导入类型的数据仓库操作语句的语句标识和任务类型为分析类型且和导入类型的数据仓库操作语句具有相同导入路径的数据仓库操作语句的语句标识存储在对应关系表中。

[0115] 其中,服务器将每个数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中,服务器根据数据仓库操作语句的语句标识可以获取访问的数据仓库目的表的表名所在的访问数据仓库操作语句。

[0116] 步骤208:服务器根据对应关系表,获取对应关系表中的每个数据仓库目的表对应的数据仓库来源表的表名;

[0117] 其中,步骤208具体可以包括如下步骤(1)至(3):

[0118] (1)、对于对应关系表中的每条记录,服务器获取记录中存储的数据仓库操作语句的语句标识和数据仓库目的表的表名;

[0119] 具体地,对于对应关系表中的每条记录,服务器依次获取对应关系表中的每条记

录中存储的数据仓库操作语句的语句标识和数据仓库目的表的表名。

[0120] (2)、服务器根据获取的语句标识获取数据仓库操作语句；

[0121] 其中,服务器中存储有语句标识和数据仓库操作语句的对应关系,根据语句标识可以从语句标识和数据仓库操作语句的对应关系中获取与语句标识对应的数据仓库操作语句。

[0122] (3)、服务器解析获取的数据仓库操作语句,得到每个数据仓库目的表对应的数据仓库来源表的表名。

[0123] 具体地,服务器通过第二正则表达式匹配规则解析获取的数据仓库操作语句,搜索出获取的数据仓库操作语句中包括的数据仓库目的表对应的源数据库仓库表的表名。

[0124] 其中,第二正则表达式代码如下:

[0125] "(?i)\\s+" + table + "(\\s+|\$|;)".

[0126] 其中,在本发明实施例中,服务器判断获取的数据仓库操作语句中是否包含数据仓库目的表的表名,如果包含,则获取的数据仓库操作语句中解析得到的表名即为数据仓库目的表对应的数据仓库来源表的表名。

[0127] 当数据仓库的操作语句的任务类型为分析类型时,服务器只需要匹配hive层表名;当数据仓库的操作语句的任务类型为导入类型或者抽取类型时,服务器只需要匹配mysql层表名。

[0128] 进一步地,服务器可以将每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名保存在第二对应关系表中,根据第二对应关系表中的内容,可以构建完整的数据库血缘关系图。

[0129] 步骤209:服务器根据每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图;

[0130] 其中,步骤209具体可以包括以下步骤(1)至(2):

[0131] (1)、服务器在数据仓库表血缘关系图中,构建数据仓库目的表的表名对应的节点,并构建数据仓库目的表对应的数据仓库来源表的表名对应的节点;

[0132] (2)、服务器将数据仓库目的表的表名对应的节点作为数据仓库目的表对应的数据仓库来源表的表名对应的节点的子节点。

[0133] 步骤210:服务器将访问数据仓库目的表的数据仓库操作语句存储在数据仓库目的表的表名对应的节点中;

[0134] 其中,在数据仓库表血缘关系图中,服务器将访问数据仓库目的表的数据仓库操作语句存储在数据仓库目的表的表名对应的节点中,服务器在数据仓库目的表的表名对应的节点中就可以获取访问数据仓库目的表的数据仓库操作语句。

[0135] 步骤211:服务器将数据仓库血缘关系图发送给终端;

[0136] 步骤212:终端接收服务器发送的数据仓库血缘关系图,并将数据仓库血缘关系图显示给用户。

[0137] 在本发明实施例中,服务器解析访问数据仓库的每个数据仓库操作语句,得到每个数据仓库操作语句访问的数据仓库目的表的表名和每个数据仓库目的表对应的源数据仓库库表的表名,并根据每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名,自动构建数据仓库表血缘关系图,减少了人工工作量,并且,提高了构建

数据仓库表血缘关系的速度以及准确度。

[0138] 实施例3

[0139] 本发明实施例提供了一种构建数据仓库表血缘关系图的装置。参见图3,其中,该装置包括:

[0140] 解析模块301,用于解析访问数据仓库的每个数据仓库操作语句,得到每个数据仓库操作语句访问的数据仓库目的表的表名;

[0141] 第一存储模块302,用于将每个数据仓库操作语句的语句标识与访问的数据仓库目的表的表名的对应关系存储在对应关系表中;

[0142] 第一获取模块303,用于根据对应关系表,获取对应关系表中的每个数据仓库目的表对应的数据仓库来源表的表名;

[0143] 构建模块304,用于根据每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名,构建数据仓库表血缘关系图。

[0144] 进一步地,解析模块301,包括:

[0145] 第一解析单元,用于解析访问数据仓库的每个数据仓库操作语句,得到访问数据仓库的每个数据仓库操作语句对应的访问方式;

[0146] 获取单元,用于获取访问方式为写方式的数据仓库操作语句;

[0147] 第二解析单元,用于解析访问方式为写方式的数据仓库操作语句,得到访问方式为写方式的访问的所有数据仓库目的表的表名。

[0148] 进一步地,该装置还包括:

[0149] 第二获取模块,用于获取任务类型为导入类型的数据仓库操作语句和与之对应的导入路径;

[0150] 第三获取模块,用于根据该导入路径获取任务类型为分析类型且具有该导入路径的数据仓库操作语句;

[0151] 绑定模块,用于绑定任务类型为导入类型的数据仓库操作语句和任务类型为分析类型且具有导入路径的数据仓库操作语句。

[0152] 进一步地,第一获取模块303,包括:

[0153] 第一获取单元,用于对于对应关系表中的每条记录,获取记录中存储的数据仓库操作语句的语句标识和数据仓库目的表的表名;

[0154] 第二获取单元,用于根据获取的语句标识获取数据仓库操作语句;

[0155] 第三解析单元,用于解析获取的数据仓库操作语句,得到每个数据仓库目的表。

[0156] 进一步地,构建模块304,包括:

[0157] 构建单元,用于在数据仓库表血缘关系图中,构建数据仓库目的表的表名对应的节点,并构建数据仓库目的表对应的数据仓库来源表的表名对应的节点;

[0158] 作为单元,用于将数据仓库目的表的表名对应的节点作为数据仓库目的表对应的数据仓库来源表的表名对应的节点的子节点。

[0159] 进一步地,该装置还包括:

[0160] 第二存储模块,用于将访问数据仓库目的表的数据仓库操作语句存储在数据仓库目的表的表名对应的节点中;

[0161] 发送模块,用于将数据仓库表血缘关系图发送给终端,由终端显示给用户。

[0162] 在本发明实施例中,服务器解析访问数据仓库的每个数据仓库操作语句,得到每个数据仓库操作语句访问的数据仓库目的表的表名和每个数据仓库目的表对应的源数据仓库库表的表名,并根据每个数据仓库目的表的表名和每个数据仓库目的表对应的数据仓库来源表的表名,自动构建数据仓库表血缘关系图,减少了人工工作量,并且,提高了构建数据仓库表血缘关系的速度以及准确度。

[0163] 需要说明的是:上述实施例提供的构建数据仓库表血缘关系图的装置在构建数据仓库表血缘关系图时,仅以上述各功能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模块完成,即将装置的内部结构划分成不同的功能模块,以完成以上描述的全部或者部分功能。另外,上述实施例提供的构建数据仓库表血缘关系图的装置与构建数据仓库表血缘关系图的方法实施例属于同一构思,其具体实现过程详见方法实施例,这里不再赘述。

[0164] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0165] 以上所述仅为本发明的较佳实施例,并不用以限制本发明,凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

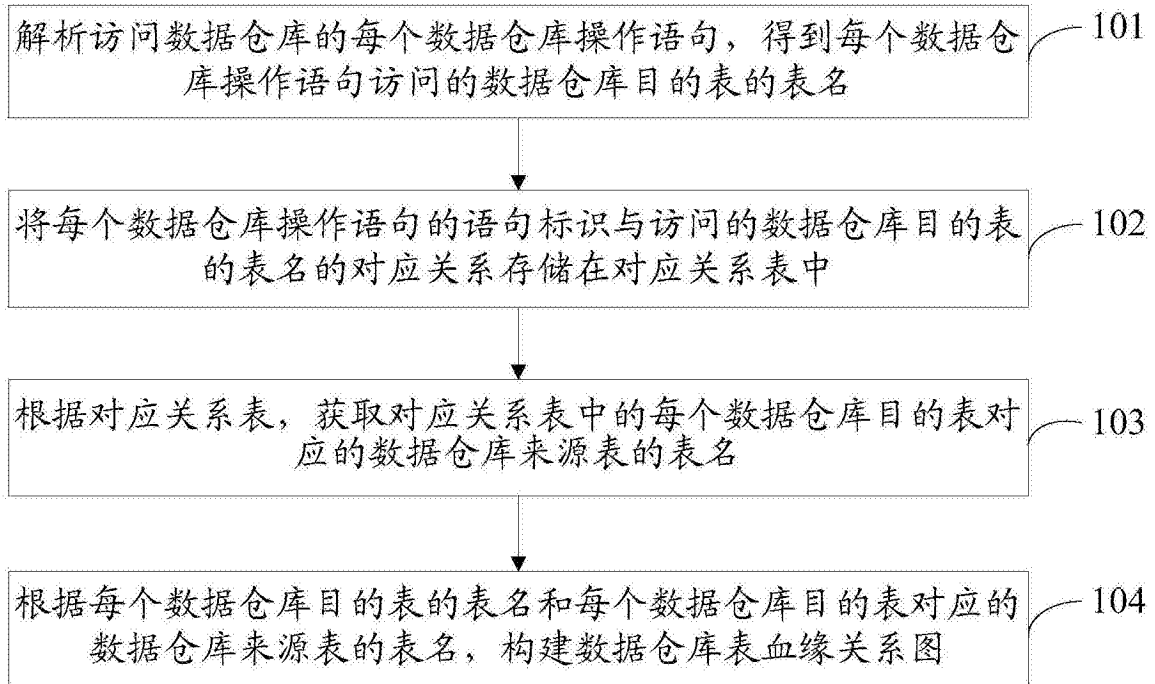


图1

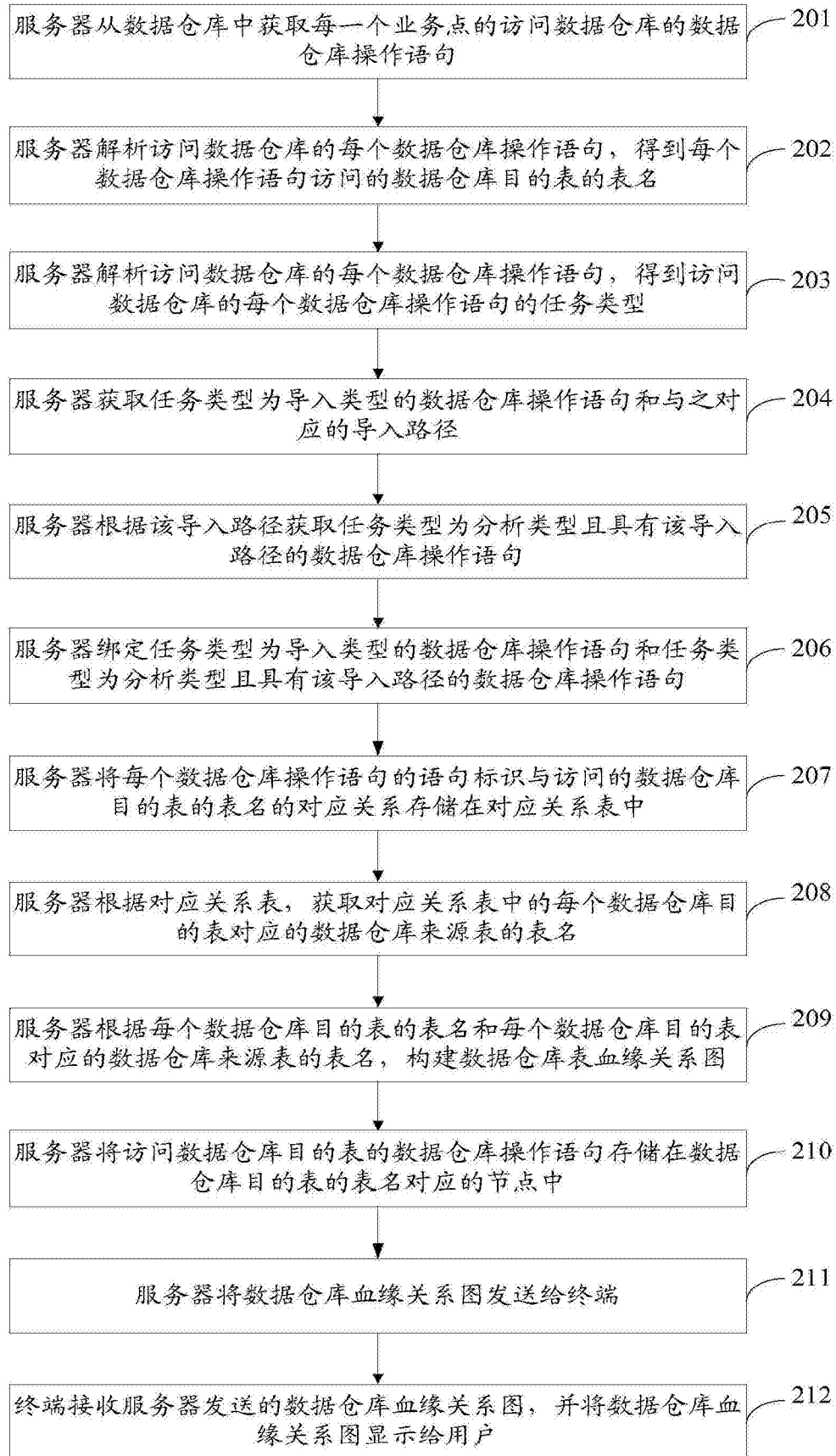


图2

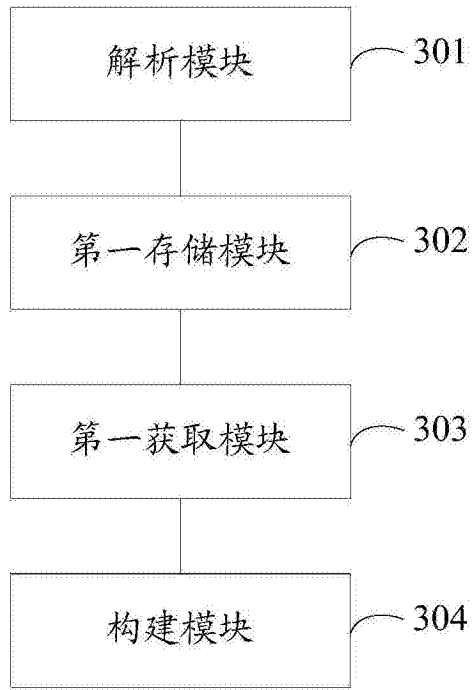


图3