(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109491989 B
(45)授权公告日 2021.08.31

(21)申请号 201811338828.2
(22)申请日 2018.11.12
(65)同一申请的已公布的文献号
申请公布号 CN 109491989 A
(43)申请公布日 2019.03.19
(73)专利权人 北京懿医云科技有限公司
地址 100195 北京市海淀区玲珑路9号院西区9号楼4层1单元304
(72)发明人 陈雪松
(74)专利代理机构 北京律智知识产权代理有限公司 11438
代理人 袁礼君 阚梓瑄
(51)Int.Cl.
G06F 16/21(2019.01)
G06F 16/28(2019.01)

(56)对比文件
CN 106462583 A,2017.02.22
CN 106462583 A,2017.02.22
CN 101267349 B,2010.09.01
CN 101420419 B,2011.05.18
CN 102799682 A,2012.11.28
CN 107704436 A,2018.02.16
CN 108376564 A,2018.08.07
CN 101299198 A,2008.11.05
US 2015310035 A1,2015.10.29

审查员 黎汉杰

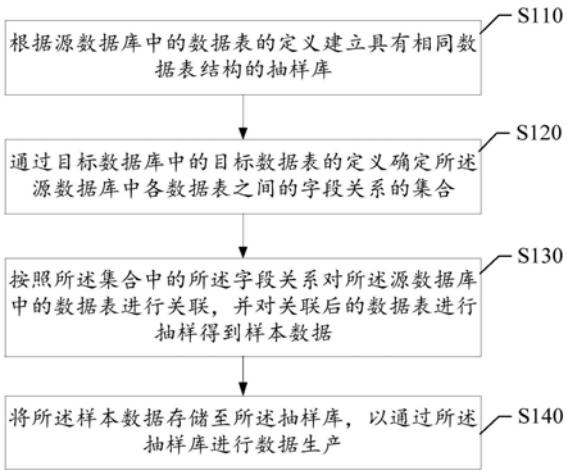
权利要求书2页 说明书9页 附图4页

(54)发明名称

数据处理方法及装置、电子设备、存储介质

(57)摘要

本公开是关于一种数据处理方法及装置、电子设备、存储介质，涉及医疗大数据技术领域，该方法包括：根据源数据库中的数据表的定义建立具有相同数据表结构的抽样库；通过目标数据库中的目标数据表的定义确定所述源数据库中各数据表之间的字段关系的集合；按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据；将所述样本数据存储至所述抽样库，以通过所述抽样库进行数据生产。本公开可以通过字段关系对源数据库中的数据表进行抽样，进而根据抽样数据提高数据生产效率。



CN 109491989 B

1.一种数据处理方法,其特征在于,包括:

根据源数据库中的数据表的定义建立具有相同数据表结构的抽样库;所述抽样库与所述源数据库的数据量不同;

通过所述源数据库中各数据表与目标数据库中各目标数据表之间的关联关系,确定所述源数据库中各数据表之间的字段关系的集合,所述字段关系用于表示不同数据表之间通过对应的字段进行关联;

按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联,并对关联后的数据表进行抽样得到样本数据;

将所述样本数据存储至所述抽样库,以通过所述抽样库进行数据生产;

其中,所述按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联,并对关联后的数据表进行抽样得到样本数据,包括:

对于所述源数据库中的第二类型表,根据所述字段关系构建关联关系树,所述关联关系树中的每个节点代表所述源数据库要抽样的一张数据表;

依次按照所述关联关系树中的各个节点对所有第二类型表进行抽样,以得到所述样本数据。

2.根据权利要求1所述的数据处理方法,其特征在于,所述源数据库中的数据表包括第一类型表和第二类型表。

3.根据权利要求2所述的数据处理方法,其特征在于,按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联,并对关联后的数据表进行抽样得到样本数据包括:

对于所述第一类型表,抽取所述第一类型表中的所有数据作为样本数据。

4.根据权利要求1所述的数据处理方法,其特征在于,根据所述字段关系构建关联关系树包括:

将所有包含预设字段的第二类型表作为起始表,并根据所述起始表与剩余的第二类型表之间的字段关系构建所述关联关系树。

5.根据权利要求4所述的数据处理方法,其特征在于,依次按照所述关联关系树中的各个节点对所有第二类型表进行抽样包括:

按照所述关联关系树中的各个节点,对所有包含所述预设字段的所述第二类型表进行抽样,得到所述样本数据。

6.一种数据处理装置,其特征在于,包括:

抽样库建立模块,用于根据源数据库中的数据表的定义建立具有相同数据表结构的抽样库;所述抽样库与所述源数据库的数据量不同;

字段关系确定模块,用于通过所述源数据库中各数据表与目标数据库中各目标数据表之间的关联关系,确定所述源数据库中各数据表之间的字段关系的集合,所述字段关系用于表示不同数据表之间通过对应的字段进行关联;

数据抽样模块,用于按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联,并对关联后的数据表进行抽样得到样本数据;

数据生产模块,用于将所述样本数据存储至所述抽样库,以通过所述抽样库进行数据生产;

其中,所述按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联,并

对关联后的数据表进行抽样得到样本数据，包括：

对于所述源数据库中的第二类型表，根据所述字段关系构建关联关系树，所述关联关系树中的每个节点代表所述源数据库要抽样的一张数据表；

依次按照所述关联关系树中的各个节点对所有第二类型表进行抽样，以得到所述样本数据。

7.一种电子设备，其特征在于，包括：

处理器；以及

存储器，用于存储所述处理器的可执行指令；

其中，所述处理器配置为经由执行所述可执行指令来执行权利要求1-5任意一项所述的数据处理方法。

8.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1-5任意一项所述的数据处理方法。

数据处理方法及装置、电子设备、存储介质

技术领域

[0001] 本公开涉及医疗大数据技术领域，具体而言，涉及一种数据处理方法、数据处理装置、电子设备以及计算机可读存储介质。

背景技术

[0002] 在使用医疗数据时，经常需要对不同来源以及异构的数据进行生产，以得到具有统一规范结构的数据，便于后续业务逻辑处理。数据生产这一过程，包含数据抽取、转换、质检等几部分工作，以保证数据质量。

[0003] 相关技术中进行数据生产时，在具体实现层面通常采用优化SQL写法或者将MapReduce任务改为Spark任务等方式；在基础架构层面，可通过对数据的存储格式进行优化或者对调度器进行参数优化等方式加快数据生产过程。

[0004] 在上述方式中，对具体实现层面的改进，由于各生产环节的任务逻辑各异以及各生产环节的数据差异性较大，因此不具备普适性、应用范围较小；对基础架构层面改进时，由于操作难度较大，复杂度较高，不能有效提升数据生产效率。

[0005] 需要说明的是，在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解，因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

[0006] 本公开的目的在于提供一种数据处理方法及装置、电子设备、存储介质，进而至少在一定程度上克服由于相关技术的限制和缺陷而导致的数据生产效率低的问题。

[0007] 本公开的其他特性和优点将通过下面的详细描述变得显然，或部分地通过本公开的实践而习得。

[0008] 根据本公开的一个方面，提供一种数据处理方法，包括：根据源数据库中的数据表的定义建立具有相同数据表结构的抽样库；通过目标数据库中的目标数据表的定义，确定所述源数据库中各数据表之间的字段关系的集合；按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据；将所述样本数据存储至所述抽样库，以通过所述抽样库进行数据生产。

[0009] 在本公开的一种示例性实施例中，通过目标数据库中的目标数据表的定义，确定所述源数据库中各数据表之间的字段关系的集合包括：通过所述源数据库中各数据表与所述目标数据库中各目标数据表之间的关联关系，确定所述源数据库中各数据表之间的字段关系的集合。

[0010] 在本公开的一种示例性实施例中，所述源数据库中的数据表包括第一类型表和第二类型表。

[0011] 在本公开的一种示例性实施例中，按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据包括：对于所述第一类型表，抽取所述第一类型表中的所有数据作为样本数据。

[0012] 在本公开的一种示例性实施例中，按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据包括：对于所述第二类型表，根据所述字段关系构建关联关系树；依次按照所述关联关系树中的各个节点对所有第二类型表进行抽样，以得到所述样本数据。

[0013] 在本公开的一种示例性实施例中，根据所述字段关系构建关联关系树包括：将所有包含预设字段的第二类型表作为起始表，并根据所述起始表与剩余的第二类型表之间的字段关系构建所述关联关系树。

[0014] 在本公开的一种示例性实施例中，依次按照所述关联关系树中的各个节点对所有第二类型表进行抽样包括：按照所述关联关系树中的各个节点，对所有包含所述预设字段的所述第二类型表进行抽样，得到所述样本数据。

[0015] 根据本公开的一个方面，提供一种数据处理装置，包括：抽样库建立模块，用于根据源数据库中的数据表的定义建立具有相同数据表结构的抽样库；字段关系确定模块，用于通过目标数据库中的目标数据表的定义确定所述源数据库中各数据表之间的字段关系的集合；数据抽样模块，用于按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据；数据生产模块，用于将所述样本数据存储至所述抽样库，以通过所述抽样库进行数据生产。

[0016] 根据本公开的一个方面，提供一种电子设备，包括：处理器；以及存储器，用于存储所述处理器的可执行指令；其中，所述处理器配置为经由执行所述可执行指令来执行上述任意一项所述的数据处理方法。

[0017] 根据本公开的一个方面，提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述任意一项所述的数据处理方法。

[0018] 本公开示例性实施例中提供的一种数据处理方法、数据处理装置、电子设备以及计算机可读存储介质中，一方面，通过按照所述字段关系对所述源数据库中的数据表进行抽样得到样本数据，并将样本数据存储至抽样库中进行数据生产，能够快速得到精准的完备样本数据，减少数据量，相对于相关技术中改进具体实现层面的方式而言，具有普适性，应用范围更广且能够提高数据生产效率；另一方面，通过按照所述字段关系对所述源数据库中的数据表进行抽样得到样本数据，避免了对基础架构层进行改进，降低了操作复杂度，能够大幅度提升数据生产效率，以便于及时进行数据校验和质检。

[0019] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本公开。

附图说明

[0020] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开的原理。显而易见地，下面描述中的附图仅仅是本公开的一些实施例，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。

[0021] 图1示意性示出本公开示例性实施例中一种数据处理方法示意图；

[0022] 图2示意性示出本公开示例性实施例中数据生产的流程图；

[0023] 图3示意性示出本公开示例性实施例中数据表与目标数据表之间的关联关系图；

[0024] 图4示意性示出本公开示例性实施例中关联关系树的示意图；

[0025] 图5示意性示出本公开示例性实施例中一种数据处理装置的框图；

[0026] 图6示意性示出本公开示例性实施例中一种电子设备的框图；

[0027] 图7示意性示出本公开示例性实施例中一种程序产品。

具体实施方式

[0028] 现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式实施，且不应被理解为限于在此阐述的范例；相反，提供这些实施方式使得本公开将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中，提供许多具体细节从而给出对本公开的实施方式的充分理解。然而，本领域技术人员将意识到，可以实践本公开的技术方案而省略所述特定细节中的一个或更多，或者可以采用其它的方法、组元、装置、步骤等。在其它情况下，不详细示出或描述公知技术方案以避免喧宾夺主而使得本公开的各方面变得模糊。

[0029] 此外，附图仅为本公开的示意性图解，并非一定是按比例绘制。图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0030] 本示例实施方式中首先提供了一种数据处理方法，可以应用于各个医院或医疗场所的数据处理场景，可基于分布式软件框架Hadoop或其他软件框架实现。参考图1所示，该数据处理方法可以包括以下步骤：

[0031] 在步骤S110中，根据源数据库中的数据表的定义建立具有相同数据表结构的抽样库；

[0032] 在步骤S120中，通过目标数据库中的目标数据表的定义确定所述源数据库中各数据表之间的字段关系的集合；

[0033] 在步骤S130中，按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据；

[0034] 在步骤S140中，将所述样本数据存储至所述抽样库，以通过所述抽样库进行数据生产。

[0035] 在本示例性实施例中提供的数据处理方法中，一方面，通过按照所述字段关系对所述源数据库中的数据表进行抽样得到样本数据，并基于样本数据进行数据生产，能够快速得到精准的样本数据，减少数据量，相对于相关技术中改进具体实现层面的方式而言，具有普适性，应用范围更广且能够减少数据生产时间，提高数据生产效率；另一方面，按照所述字段关系对所述源数据库中的数据表进行抽样得到样本数据，避免了对基础架构层进行改进，降低了操作复杂度，能够大幅度提升数据生产效率，以便于及时进行数据校验和质检。

[0036] 接下来，结合附图对本示例性实施例中的数据处理方法进行进一步解释说明。

[0037] 在步骤S110中，根据源数据库中的数据表的定义建立具有相同数据表结构的抽样

库。

[0038] 本示例性实施例中，参考图2所示，源数据库可以为ETLDR层的数据库，该ETLDR层指的是数据生产的初始层，可将从目标医疗系统恢复的数据在本层映射到统一结构的数据表中。目标医疗系统例如可为目标医院或者是诊所的医疗信息系统。源数据库中可以包括第一类型表和第二类型表。其中，第一类型表可以为字典表，第二类型表可以为记录信息表。第一类型表和第二类型表虽然功能和存储的数据类型不同，但是二者的结构可以相同。字典表中用户的增加、删除、修改等各种操作不会对其记录信息产生影响，字典表可用于存放一些与用户不相关的信息。记录信息表例如可以为实例表，可用于存储与用户的信息相关的一些信息。源数据库中每个第一类型表、每个第二类型表以及每个第一类型表和第二类型表之间均没有相同的字段。

[0039] 数据表的定义例如可以包括数据表的结构、表的主键、关键字以及索引等等。抽样库也可以建立在ETLDR层。抽样库中可用于存储源数据库中的少量完备样本数据。一般而言，可从ETLDR层的各数据表中进行抽样得到抽样小数据集存储在抽样库中，进一步将抽样库中的样本数据即抽样小数据集带入Schema层得到完备数据集，进而带入Schema层及后续的各层例如PP层和SOAR层等数据生产质检过程中以进行数据生产。对于整个数据生产流程而言，从ETLDR层经过Schema层、PP层到SOAR层，前一层的输出是后一层的输入。

[0040] 需要说明的是，此处的抽样库与源数据库中的数据表的结构完全相同，例如，源数据库中包括字典表和记录信息表，则生成的抽样库中也必须包括字典表和记录信息表，以保证数据抽样的准确性和完整性。除此之外，数据表的数量也完全相同，只是抽样库中并非每个数据表中都存在数据，因此抽样库与源数据库的差别仅在于数据量不同。例如，源数据库中包括100个数据表，抽样库中也包括结构完全相同的100个数据表，但是抽样库中只有50个数据表中包括数据，以减小进入数据生产的数据量，提高数据生产效率。

[0041] 在步骤S120中，通过目标数据库中的目标数据表的定义确定所述源数据库中各数据表之间的字段关系的集合。

[0042] 本示例性实施例中，目标数据库可以为源数据库的下游数据库，举例而言，ETLDR层的下游为Schema层，因此针对ETLDR层的源数据库而言，其目标数据库可为Schema层的数据库。目标数据表可以为目标数据库中要生产的数据表，以图2中所示的Schema层数据库中的目标数据表为例进行说明。图2中，ETLDR层数据库中各数据表间的字段原本是没关系的，由于要进行下游数据层Schema的数据生产，因此需要确定各数据表之间的字段关系。

[0043] Schema数据层可由初步产出的源数据库中ETLDR层通过人为定义的关系和条件关联得到。Schema数据层属于数据生产的一层，可按照逻辑定义对ETLDR层数据进行关联，得到存在嵌套的数据结构。Schema数据层定义了一个多维数据库，可包含一个逻辑模型，并定义了逻辑模型到物理模型的映射。在Schema数据层中包含多维数据的存储方式，例如事实表、维表及其结构等。

[0044] 目标数据库中的目标数据表的数量可根据实际生产需求而确定，且这些所有目标数据表中均包括预设字段，该预设字段例如可以为患者ID字段，可用PID字段来表示。在目标数据库中各目标数据表均包括预设字段的基础上，可以设定对ETLDR层的数据进行抽取的目标。可例如，对给定的患者ID字段集合，从ETLDR层各数据表抽取该患者ID字段集合关联的数据，从而在Schema层可以得到基于患者ID字段集合的完备数据集，即样本数据。

[0045] 本示例性实施例中，源数据库中各数据表和目标数据库中的目标数据表之间可存在关联关系，该关联关系可为表关联关系，例如图3所示。源数据库中，即ETLDR层的每一个数据表可对应Schema层的一个或多个目标数据表，Schema层的一个目标数据表可分别与ETLDR层的多个数据表相互对应。需要说明的是，由于字典表中的信息大多为与用户不相关的信息，因此步骤S120中的源数据库中的各数据表主要代表的是源数据库中的第二类型表，即记录信息表。

[0046] 参考图3中所示，ETLDR层中可包括多个记录信息表，例如表A、表B、表C、表D、表E，Schema层可包括多个目标数据表，例如表x、表y。其中，ETLDR层中的表A、表B和表D对应Schema层中的表x，ETLDR层中的表C、表D和表E对应Schema层中的表y。

[0047] 通过表关联关系可得到源数据库中各数据表之间的字段关系，该字段关系可以为字段关联关系。具体地，由图3中数据表与目标数据表之间的对应关系可得到ETLDR层到Schema层的表关联关系，例如ETLDR层中的表A、表B和表D与Schema层中的表x关联，ETLDR层中的表C、表D和表E与Schema层中的表y关联。进一步地可得到源数据库ETLDR层各数据表之间的字段关系，即各个记录信息表之间的字段关联关系。可用SQL(Structured Query Language，结构化查询语言)来生成Schema层的目标数据表，例如，生成Schema层表x的SQL为：

[0048] SELECT…FROM A left join B on(A.a＝B.b1)left join D on(B.b2＝D.d)；

[0049] 通过上述SQL定义，可以得到源数据库ETLDR层中表A、表B和表D间字段关联关系，例如：A.a:B.b1，可表示表A通过字段a与表B的字段b1关联；B.b2:D.d，可表示表B通过字段b2与表D的字段d关联。如此一来，可根据Schema层中目标数据表的所有定义，获取ETLDR层中各数据表之间的字段关系的集合U。

[0050] 接下来，在步骤S130中，按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据。

[0051] 在本示例性实施例中，由于源数据库包括第一类型表和第二类型表，因此针对第一类型表和针对第二类型表的抽样方式可不同。具体地，对于第一类型表即字典表而言，字典表中的数据与患者ID字段pid无关，且字典表代表的是定义型数据，同时字典表中的数据量较小，因此需要直接将字典表中的所有数据作为样本数据从源数据库拷贝至抽样库，以保证后续抽样过程的正常进行。

[0052] 对于第二类型表即记录信息表而言，首先可对记录信息表进行检查，排除其中的空表，以避免空表对数据抽样结果的影响，提高数据抽样准确性。对于非空的记录信息表而言，由于在步骤S120中确定了每个数据表之间的字段关联关系，因此可根据各数据表之间的字段关联关系构建包含所有数据表的关联关系树；进一步依次按照所述关联关系树中的各个节点对源数据库ETLDR层中的第二类型表进行抽样得到样本数据，进而将样本数据用于数据生产、数据质检等过程。

[0053] 具体地，可根据Schema层目标数据库中所有目标数据表的定义，获取ETLDR层数据表之间的字段关联关系的集合U。在集合U中，可以找到字段是预设字段，即患者ID字段且以患者ID字段为主键的至少一个左表，同时可将包含患者ID字段的这些左表作为起始表，将源数据库中剩余的其他记录信息表根据字段关系与这些起始表关联起来。例如表A为起始表，则表B通过b1字段与表A的a字段关联起来。又例如，针对与表A没有直接字段关联关系的

表D,可通过d字段与表B的b2字段关联,从而将表D间接与表A关联起来。通过这种方式,可在ETLDR层所有数据表全部建立起字段关联关系后,将患者ID字段集合作为根,生成一棵关联关系树。

[0054] 参考图4所示,对于pid集合表R而言,用R.pid＝A.pid确定了起始表A,用R.pid＝B.pid确定了起始表B。对于起始表A而言,通过A.a1＝C.c使得表C中的字段c与起始表A中的字段a1关联。除此之外,通过A.a2＝D.d1使得表D中的字段d1与起始表A中的字段a2关联。对于表D而言,通过D.d2＝F.f使得表F通过字段f与表D中的字段d2关联,另外通过D.d3＝G.g使得表G通过字段g与表D中的字段d3关联。除此之外,通过B.b＝E.e使得表E通过字段e与起始表B中的字段b关联。在所有数据表A、表B、表C、表D、表E、表F、表G均通过字段关系进行关联后,可生成如图4所示的关联关系树。

[0055] 在生成关联关系树之后,可基于该关联关系树,从根开始逐层逐节点对源数据库中的数据表进行数据抽样。关联关系树的根可为pid集合表R。关联关系树中的每个节点均代表ETLDR层要抽样的一张数据表。每一张数据表的样本数据,可通过源数据库的数据与父节点的数据join获得。

[0056] 具体包括:根据R.pid＝A.pid从源数据库中得到起始表A的样本数据存储至抽样库;接下来,可在起始表A的样本数据的基础上,从源数据库中通过A.a2＝D.d1得到表D的样本数据存储至抽样库;进一步地,可在表D的样本数据的基础上,从源数据库中通过D.d2＝F.f得到表F的样本数据存储至抽样库。如此一来,能够自动完成基于pid患者ID字段集合的样本数据抽取,在Schema层得到基于pid集合的完备数据集。通过本示例性实施例中的方法,可实现数据抽取自动化。具体的数据抽取过程可通过程序执行,此处不作特殊限定。需要说明的是,如果父节点对应的数据表中为空,则无法对子节点对应的数据表进行抽样。可例如,表D中为空,则无法对表F和表G进行数据抽样,因此在进行数据抽样之前,首先需排除空的数据表,以避免对数据抽样过程的影响。

[0057] 在步骤S140中,将所述样本数据存储至所述抽样库,以通过所述抽样库进行数据生产。

[0058] 本示例性实施例中,通过抽样库中的样本数据,可得到完备数据集。举例而言,要生产的目标数据表为表x,例如ETLDR层中与表x关联的数据表包括表A、表B和表D,进一步可根据pid在表A、表B和表D中进行数据抽样,得到完备数据集,以生产表x。

[0059] 基于这些完备数据集进行数据生产时,由于大幅度减小了数据量,因此可大幅度提高数据生产效率。除此之外,通过本示例性实施例中的方法可对所有数据进行抽样,而与各生产环节的任务逻辑无关,因此具有普适性,应用范围更广。通过抽取少量的样本数据,快速跑完各层数据生产流程,从而能够达到在短时间内验证数据生产质量,及验证目标数据结构设计是否满足业务需求的目的,减少时间和集群资源开销。

[0060] 再参考图2所示,在实际数据生产过程中,若按照现有技术的方法从ETLDR层到SOAR层的数据生产和数据质检过程中,需要使用全量数据进行数据生产。在实际数据生产时,一个包含444904个患者ID字段的ETLDR层中的数据量为49.6G,完成Schema层及PP层数据生产的时间为70个小时。通过本示例性中的方法,若在ETLDR层与Schema层之间通过抽样库得到完备数据集,使用小数据集抽样的数据校验流程,可大大减少进入Schema层至SOAR层的数据量,例如只抽取其中500个患者ID字段得到的抽样库中的数据量为1.7G,小数据集

抽样及Schema层和PP层数据生产的时间不足5小时，大大节省了校验时间和数据生产时间，从而可提高数据生产效率。

[0061] 本公开还提供了一种数据处理装置。参考图5所示，该数据处理装置500可以包括：抽样库建立模块501、字段关系确定模块502、数据抽样模块503、数据生产模块504，其中：

[0062] 抽样库建立模501，用于根据源数据库中的数据表的定义建立具有相同数据表结构的抽样库；

[0063] 字段关系确定模块502，用于通过目标数据库中的目标数据表的定义确定所述源数据库中各数据表之间的字段关系的集合；

[0064] 数据抽样模块503，用于按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据；

[0065] 数据生产模块504，用于将所述样本数据存储至所述抽样库，以通过所述抽样库进行数据生产。

[0066] 在本公开的一种示例性实施例中，字段关系确定模块包括：确定控制模块，用于通过所述源数据库中各数据表与所述目标数据库中各目标数据表之间的关联关系，确定所述源数据库中各数据表之间的字段关系的集合。

[0067] 在本公开的一种示例性实施例中，所述源数据库中的数据表包括第一类型表和第二类型表。

[0068] 在本公开的一种示例性实施例中，数据抽样模块包括：第一抽样模块，用于对于所述第一类型表，抽取所述第一类型表中的所有数据作为样本数据。

[0069] 在本公开的一种示例性实施例中，数据抽样模块包括：关系树建立模块，用于对于所述第二类型表，根据所述字段关系构建关联关系树；第二抽样模块，用于依次按照所述关联关系树中的各个节点对所有第二类型表进行抽样，以得到所述样本数据。

[0070] 在本公开的一种示例性实施例中，关系树建立模块包括：构建控制模块，用于将所有包含预设字段的第二类型表作为起始表，并根据所述起始表与剩余的第二类型表之间的字段关系构建所述关联关系树。

[0071] 在本公开的一种示例性实施例中，第二抽样模块包括：抽样控制模块，用于按照所述关联关系树中的各个节点，对所有包含所述预设字段的所述第二类型表进行抽样，得到所述样本数据。

[0072] 需要说明的是，上述数据处理装置中各模块的具体细节已经在对应的数据处理方法中进行了详细描述，因此此处不再赘述。

[0073] 应当注意，尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元，但是这种划分并非强制性的。实际上，根据本公开的实施方式，上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之，上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0074] 此外，尽管在附图中以特定顺序描述了本公开中方法的各个步骤，但是，这并非要求或者暗示必须按照该特定顺序来执行这些步骤，或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的，可以省略某些步骤，将多个步骤合并为一个步骤执行，以及/或者将一个步骤分解为多个步骤执行等。

[0075] 在本公开的示例性实施例中，还提供了一种能够实现上述方法的电子设备。

[0076] 所属技术领域的技术人员能够理解，本发明的各个方面可以实现为系统、方法或程序产品。因此，本发明的各个方面可以具体实现为以下形式，即：完全的硬件实施方式、完全的软件实施方式(包括固件、微代码等)，或硬件和软件方面结合的实施方式，这里可以统称为“电路”、“模块”或“系统”。

[0077] 下面参照图6来描述根据本发明的这种实施方式的电子设备600。图6显示的电子设备600仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0078] 如图6所示，电子设备600以通用计算设备的形式表现。电子设备600的组件可以包括但不限于：上述至少一个处理单元610、上述至少一个存储单元620、连接不同系统组件(包括存储单元620和处理单元610)的总线630。

[0079] 其中，所述存储单元存储有程序代码，所述程序代码可以被所述处理单元610执行，使得所述处理单元610执行本说明书上述“示例性方法”部分中描述的根据本发明各种示例性实施方式的步骤。例如，所述处理单元610可以执行如图1中所示的步骤。

[0080] 存储单元620可以包括易失性存储单元形式的可读介质，例如随机存取存储单元(RAM)6201和/或高速缓存存储单元6202，还可以进一步包括只读存储单元(ROM)6203。

[0081] 存储单元620还可以包括具有一组(至少一个)程序模块6205的程序/实用工具6204，这样的程序模块6205包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0082] 总线630可以为表示几类总线结构中的一种或多种，包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

[0083] 电子设备600也可以与一个或多个外部设备800(例如键盘、指向设备、蓝牙设备等)通信，还可与一个或者多个使得用户能与该电子设备600交互的设备通信，和/或与使得该电子设备600能与一个或多个其它计算设备进行通信的任何设备(例如路由器、调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口650进行。并且，电子设备600还可以通过网络适配器660与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器660通过总线630与电子设备600的其它模块通信。应当明白，尽管图中未示出，可以结合电子设备600使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0084] 在本公开的示例性实施例中，还提供了一种计算机可读存储介质，其上存储有能够实现本说明书上述方法的程序产品。在一些可能的实施方式中，本发明的各个方面还可以实现为一种程序产品的形式，其包括程序代码，当所述程序产品在终端设备上运行时，所述程序代码用于使所述终端设备执行本说明书上述“示例性方法”部分中描述的根据本发明各种示例性实施方式的步骤。

[0085] 参考图7所示，描述了根据本发明的实施方式的用于实现上述方法的程序产品700，其可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码，并可以在终端设备，例如个人电脑上运行。然而，本发明的程序产品不限于此，在本文件中，可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0086] 所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0087] 计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0088] 可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、RF等等，或者上述的任意合适的组合。

[0089] 可以以一种或多种程序设计语言的任意组合来编写用于执行本发明操作的程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中，远程计算设备可以通过任意种类的网络，包括局域网(LAN)或广域网(WAN)，连接到用户计算设备，或者，可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0090] 此外，上述附图仅是根据本发明示例性实施例的方法所包括的处理的示意性说明，而不是限制目的。易于理解，上述附图所示的处理并不表明或限制这些处理的时间顺序。另外，也易于理解，这些处理可以是例如在多个模块中同步或异步执行的。

[0091] 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本公开的其他实施例。本申请旨在涵盖本公开的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本公开的真正范围和精神由权利要求指出。

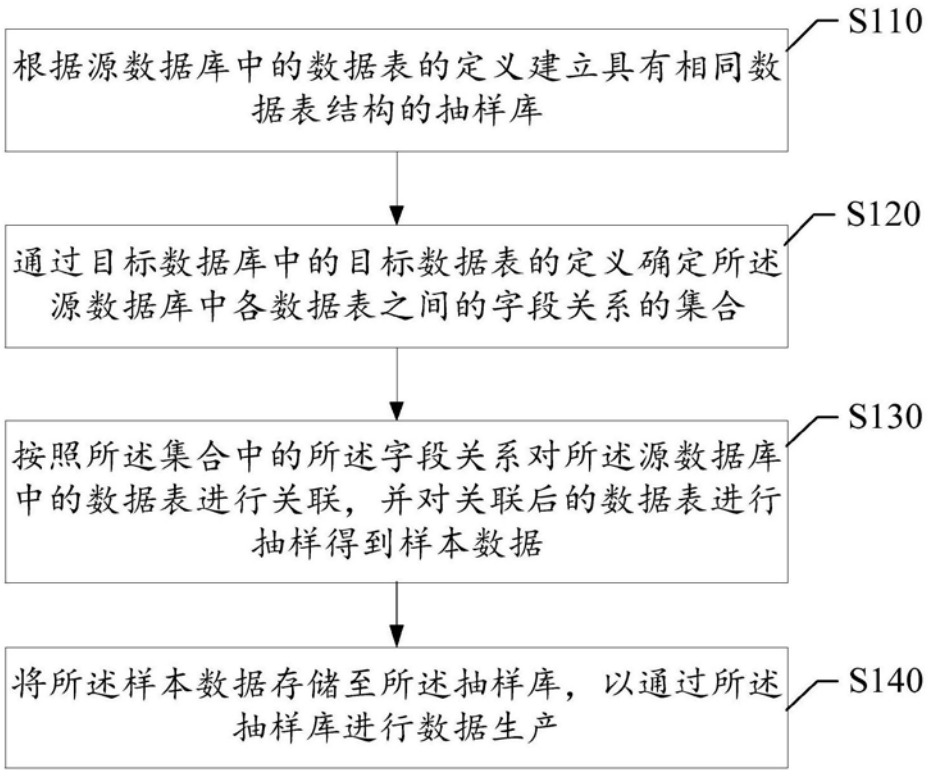

S110
根据源数据库中的数据表的定义建立具有相同数据表结构的抽样库
S120
通过目标数据库中的目标数据表的定义确定所述源数据库中各数据表之间的字段关系的集合
S130
按照所述集合中的所述字段关系对所述源数据库中的数据表进行关联，并对关联后的数据表进行抽样得到样本数据
S140
将所述样本数据存储至所述抽样库，以通过所述抽样库进行数据生产

图1

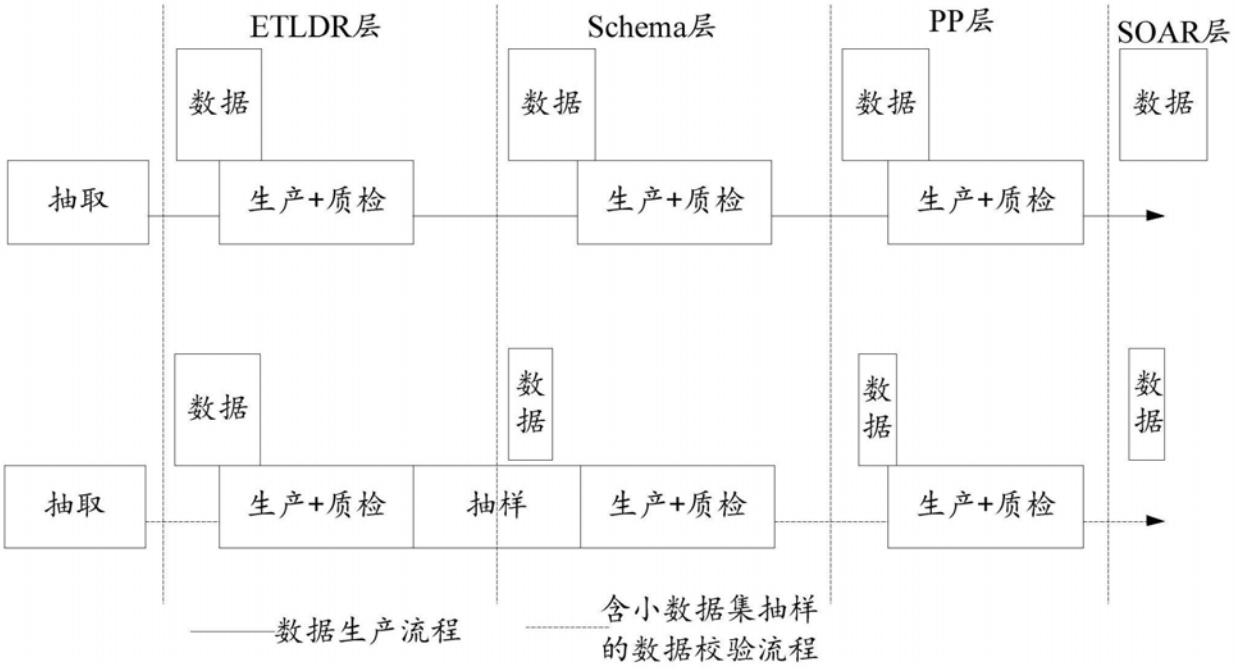

ETLDR层
Schema层
PP层
SOAR层
数据
抽取
生产+质检
抽样
数据生产流程
含小数据集抽样的数据校验流程

图2

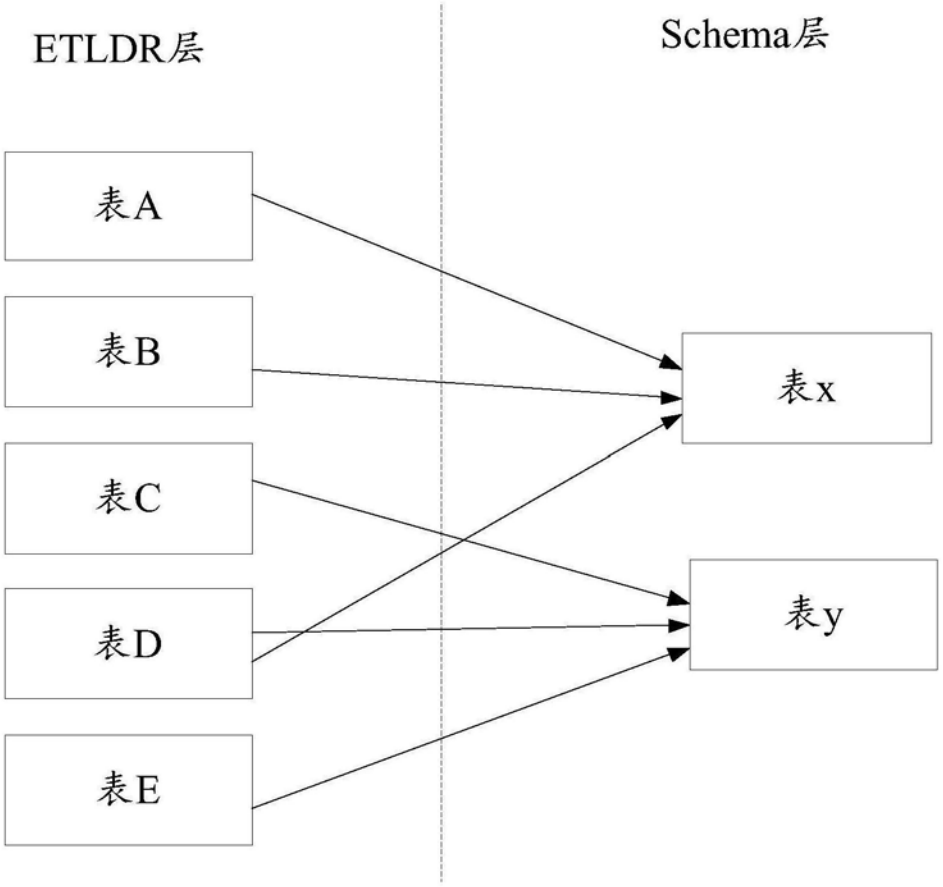


图3

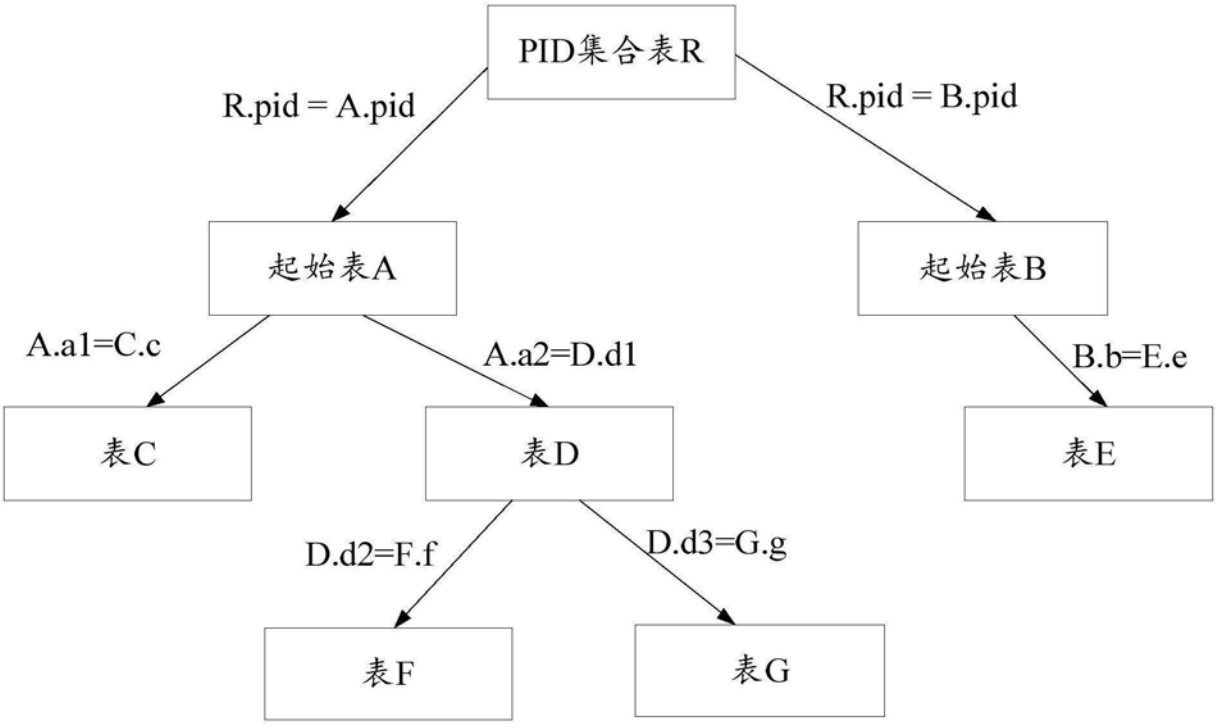

PID集合表R
R.pid = A.pid
R.pid = B.pid
起始表A
起始表B
A.a1=C.c
A.a2=D.d1
B.b=E.e
表C
表D
表E
D.d2=F.f
D.d3=G.g
表F
表G

图4

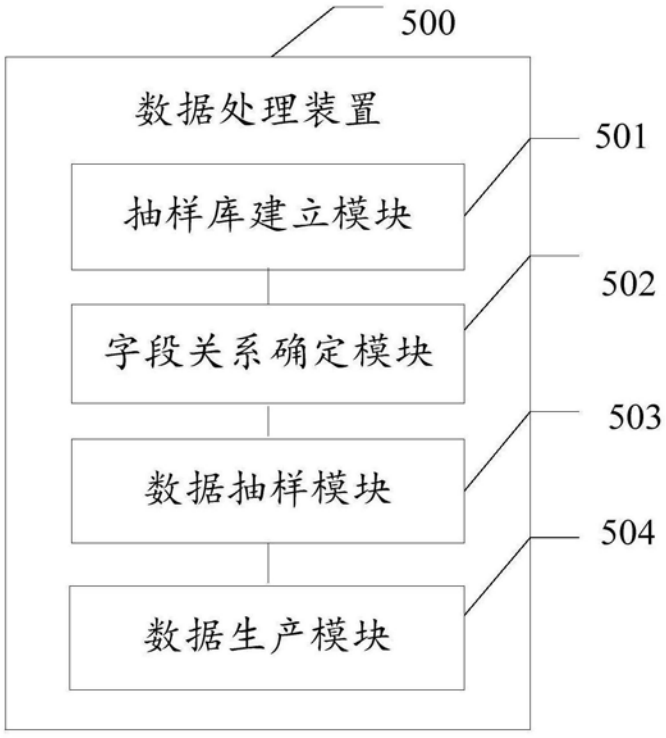

500
数据处理装置
501
抽样库建立模块
502
字段关系确定模块
503
数据抽样模块
504
数据生产模块

图5

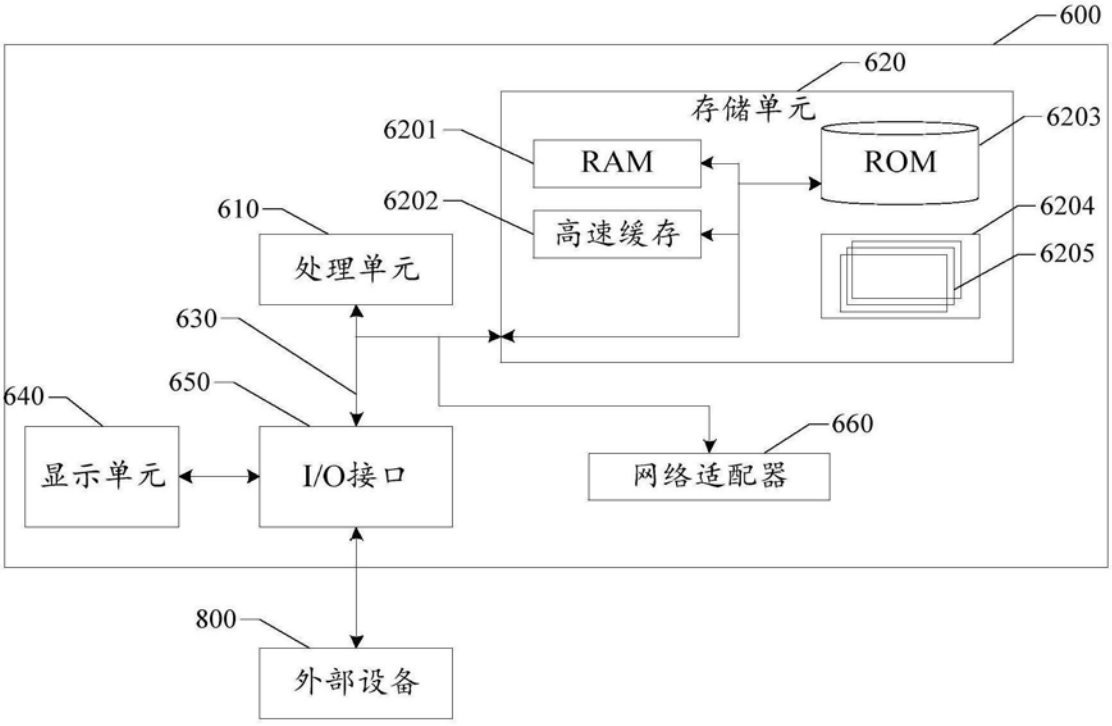

600
620
存储单元
6201
RAM
6203
ROM
6202
高速缓存
6204
6205
610
处理单元
630
650
640
显示单元
I/O接口
660
网络适配器
800
外部设备

图6

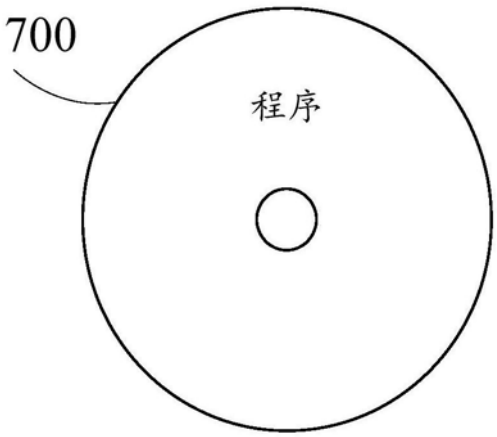

700
程序

图7