(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 104899199 A
(43)申请公布日 2015.09.09

(21)申请号 201410076445.8

(22)申请日 2014.03.04

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼

(72)发明人 徐玉鹏

(74)专利代理机构 北京三友知识产权代理有限公司 11127
代理人 党晓林

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

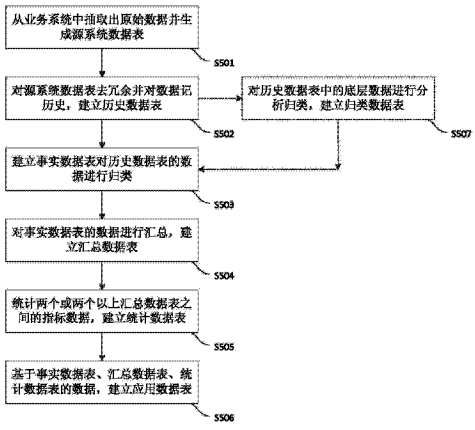
权利要求书3页 说明书13页 附图3页

(54)发明名称

一种数据仓库数据处理方法和系统

(57)摘要

本发明提供一种数据仓库数据处理方法，包括：从业务系统中抽取出原始数据并生成源系统数据表；对源系统数据表去冗余并对数据记历史，建立历史数据表；建立事实数据表对历史数据表的数据进行归类；对事实数据表的数据进行汇总，建立汇总数据表；统计两个或两个以上汇总数据表之间的指标数据，建立统计数据表；基于事实数据表、汇总数据表、统计数据表的数据，建立应用数据表。本发明提供的数据仓库数据处理方法，避免了通用维度模型层中每一层级内部的任务相互依赖，使得任务的并行数目达到最大，计算机资源能够被有效利用，从而提高数据仓库数据处理的效率。本发明还提供了相应的数据仓库数据处理系统，能够实现本发明的数据仓库数据处理方法。



CN 104899199 A

1. 一种数据仓库数据处理方法，其特征在于，包括：

从业务系统中抽取出原始数据并生成源系统数据表；

对源系统数据表去冗余并对数据记历史，建立历史数据表；

建立事实数据表对历史数据表的数据进行归类；

对事实数据表的数据进行汇总，建立汇总数据表；

统计两个或两个以上汇总数据表之间的指标数据，建立统计数据表；

基于事实数据表、汇总数据表、统计数据表的数据，建立应用数据表。

2. 如权利要求 1 所述的数据处理方法，其特征在于，所述数据处理方法还包括：

对历史数据表中的底层数据进行分析归类，建立归类数据表；

相应地，

所述建立事实数据表对历史数据表的数据进行归类，包括：建立事实数据表对历史数据表和 / 或归类数据表的数据进行归类。

3. 如权利要求 1 或 2 所述的数据处理方法，其特征在于，所述建立一个数据表称为一个当前任务；每一任务的初始任务状态为未完成状态。

4. 如权利要求 3 所述的数据处理方法，其特征在于，设置一状态标识来表示每一任务的任务状态。

5. 如权利要求 3 所述的数据处理方法，其特征在于，在所述建立任一数据表之前，还包括：

查询当前任务所依赖的父任务的任务状态；

若父任务的状态均为完成状态，执行当前任务；

若父任务的状态中至少有一个父任务是未完成状态，在预定时间间隔后，重新查询当前任务所依赖的父任务的任务状态。

6. 一种数据仓库数据处理系统，其特征在于，包括：源系统数据处理单元、企业数据仓库第一处理单元、通用维度模型第一处理单元、通用维度模型第二处理单元、通用维度模型第三处理单元、应用数据处理单元；其中，

所述源系统数据处理单元，用于从各个业务系统中抽取出原始数据，生成一个或一个以上的源系统数据表；

所述企业数据仓库第一处理单元，用于对源系统数据处理单元中的源系统数据表去冗余并对数据记历史，产生与源系统数据表相对应的一个或一个以上的历史数据表；

所述通用维度模型第一处理单元，用于建立一个或一个以上的事实表对企业数据仓库第一处理单元的数据进行归类；

所述通用维度模型第二处理单元，用于对通用维度模型第一处理单元的数据进行汇总，生成至少一个汇总数据表；

所述通用维度模型第三处理单元，用于统计通用维度模型第二处理单元中表与表的指标数据，生成至少一个统计数据表；

所述应用数据处理单元，用于根据通用维度模型第一处理单元、通用维度模型第二处理单元、通用维度模型第三处理单元的数据生成应用数据表。

7. 如权利要求 6 所述的一种数据仓库数据处理系统，其特征在于，

所述企业数据仓库第一处理单元调用源系统数据处理单元的结果；

所述通用维度模型第一处理单元调用企业数据仓库第一处理单元的结果；

所述通用维度模型第二处理单元调用通用维度模型第一处理单元的结果；

所述通用维度模型第三处理单元调用通用维度模型第二处理单元的结果；

所述应用数据处理单元，调用通用维度模型第一处理单元、通用维度模型第二处理单元、通用维度模型第三处理单元的结果。

8. 如权利要求 6 所述的一种数据仓库数据处理系统，其特征在于，所述数据仓库数据处理系统，还包括：企业数据仓库第二处理单元；

所述企业数据仓库第二处理单元，用于对企业数据仓库第一处理单元中的底层数据进行分析归类，生成归类数据表；

相应地，

所述通用维度模型第一处理单元，用于建立一个或一个以上的事实表对企业数据仓库第一处理单元、企业数据仓库第二处理单元的数据进行归类。

9. 如权利要求 8 所述的一种数据仓库数据处理系统，其特征在于，所述企业数据仓库第二处理单元调用企业数据仓库第一处理单元的结果；

相应地，

所述通用维度模型第一处理单元调用企业数据仓库第一处理单元和 / 或企业数据仓库第二处理单元的结果。

10. 如权利要求 6-9 任意一项所述的一种数据仓库数据处理系统，其特征在于，所述数据处理系统中每一处理单元建立一个数据表对应一个任务单元。

11. 如权利要求 10 所述的一种数据仓库数据处理系统，其特征在于，所述任务单元，包括：父任务单元、当前任务单元；其中，

所述父任务单元，用于记录当前任务所依赖的父任务，并查询所述父任务的任务状态；所述父任务的初始状态为未完成状态；

所述当前任务单元，用于执行当前任务，当前任务完成后，更改当前任务的任务状态为已完成状态。

12. 如权利要求 11 所述的一种数据仓库数据处理系统，其特征在于，所述父任务单元，包括：父任务状态记录单元和父任务状态查询单元；

所述父任务状态记录单元，用于记录当前任务所依赖的所有父任务；

所述父任务状态查询单元，用于查询当前任务所依赖的所有父任务的任务状态；若所有父任务的任务状态均为完成状态，则执行当前任务单元；若所有父任务中至少有一个父任务的任务状态为未完成状态，则等待预定时间后，重新执行父任务状态查询单元，直至所有父任务的任务状态均为完成状态。

13. 如权利要求 11 所述的一种数据仓库数据处理系统，其特征在于，所述当前任务单元，包括：当前任务执行单元和当前任务状态记录单元；

所述当前任务执行单元，用于执行当前任务，即建立一个数据表；

所述当前任务状态记录单元，用于记录当然任务的任务状态；所述任务状态的初始状态为未完成状态；在当前任务执行单元中当前任务执行完毕时，更改当前任务的任务状态为已完成状态。

14. 如权利要求 11 所述的一种数据仓库数据处理系统，其特征在于，所述源系统数据

处理单元中每一个任务对应的任务单元，不包括：父任务单元。

一种数据仓库数据处理方法和系统

技术领域

[0001] 本发明涉及数据库领域，尤其涉及一种数据仓库数据处理方法和系统。

背景技术

[0002] 数据库(Database)是按照数据结构来组织、存储和管理数据的仓库。对数据库数据的处理大致分为两类：一类是操作型处理，这类处理通常用于对数据库中的少数记录进行查询、修改；另一类是分析型处理，这类处理一般用于对历史数据进行分析，使得数据能够应用于决策，所述分析型处理后得到的面向主题的、集成的、与时间相关的、不可修改的数据集合可以称为数据仓库。数据仓库的任务主要是把信息加以整理归纳和重组，并及时提供给决策人员。目前数据仓库的数据处理通常通过分布式系统来实现，所述分布式系统可以将多台计算机联合起来，构成计算机群，并行处理大规模的数据，同时在多台计算机上运行不同任务。

[0003] 目前数据仓库中对数据进行处理一般建立在ETL数据处理理论的基础上的，ETL是指Extraction(抽取)、Transformation(转换)、和Loading(加载)。具体的ETL操作包括：将业务系统中的数据抽取出来，并将不同数据源的数据按照业务需要进行转换和整合，得出目标数据，然后将目标数据加载到数据仓库中。

[0004] 数据仓库一般是以数据表的结构存储数据，每个数据表对应一个数据对象。数据表是指一系列二维数组的集合，通常用来代表和储存数据对象之间的关系。数据库表可以由纵向的列和横向的行组成，例如一个有关作者信息的名为“作者”的表中，每个列包含的是所有作者的某个特定类型的信息，比如“姓氏”，而每行则包含了某个特定作者的所有信息：姓、名、住址等等。对于特定的数据库表，列的数目一般事先固定，各列之间可以由列名来识别。

[0005] 在数据仓库数据处理过程中，通常将建立或生成一个数据表作为一个任务，所述任务的初始状态可以是未完成状态；对每一个任务设置一个任务状态标识来表示该任务的任务状态，例如用“0”表示任务状态为未完成，用“1”表示任务状态为完成。若需要第一个任务完成后才能执行第二个任务，那么所述第一个任务称为父任务，所述第二个任务称为子任务。对于数据仓库数据处理过程，父任务和子任务分别占用调度系统的一个调度层级。调度系统可以记录各个任务之间的依赖关系。通常，数据仓库数据处理过程中，子任务每隔预定时间，主动查询其依赖的父任务的任务状态。若父任务的任务状态均为已完成状态，则可以执行子任务。

[0006] 常用的数据处理方法包括称为Inmon的企业信息化工厂式的数据处理方法和称为Kimball的维度数据仓库总线体系结式的数据处理方法构。

[0007] 所述Inmon的企业信息化工厂式的数据处理方法，该数据处理方法通过ETL将业务源系统的数据经过抽取、转换之后加载到企业数据仓库，在此企业数据仓库基础层上建立面向主题的数据集市。在主题数据集市的基础上，提供应用层服务。所述企业数据仓库基础模型遵循实体-联系模型(简称E-R模型)的原则来设计。所述企业数据仓库基础层基

于原始数据的性质，尽可能保存粒度最细的数据。所述方法在数据集市中采用维度设计的方法。

[0008] 另一种称为 Kimball 的维度数据仓库总线体系结构式的数据处理方法，该数据处理方法主要从业务源系统根据 ETL 理论建立维度数据仓库基础层。所述维度数据基础层根据维度建模的原则来设计，由一系列的星型模型和多维数据集组成。在维度数据模型的基础上建立面向主题的数据集市，数据集市同样采用维度建模的原则，对基础层重新进行维度定义和聚合。再在主题集市的基础上，建立各种应用层服务。

[0009] 上述两种数据仓库数据处理方法实现时将数据仓库划分为四层结构，分别为：源系统数据处理层、企业数据仓库数据处理层、通用维度模型数据处理层和应用数据处理层。源系统数据处理层用于从业务系统抽取原始数据，所述源系统数据处理层一般占用调度系统的一个调度层级；企业数据仓库数据处理层用于以关系模型存储各类业务数据，实现海量数据的集中、稳定、有序存贮，所述企业数据仓库数据处理层一般占用调度系统的一个或两个调度层级；通用维度模型数据处理层用于根据主题应用存贮数据集合，所述通用维度模型数据处理层对数据处理时任务比较复杂，一般需要占用调度系统的多个调度层级；应用数据处理层主要用于向用户提供业务数据，所述应用数据处理层一般占用调度系统的一个层级。

[0010] 在实现本申请过程中，发明人发现现有技术中至少存在如下问题：

[0011] 由于通用维度模型数据处理层在对数据进行处理时任务比较复杂，通用维度模型数据处理层级中的任务在该层级内部相互依赖，这样通用维度模型层在处理数据时实际会占用调度系统的多个调度层级。例如通用维度模型层中的任务可能既包含对基础信息进行描述，也包含对信息进行汇总、统计等，而对信息进行汇总依赖于对基础信息进行描述，对信息进行统计又依赖于对信息进行汇总。这样，可能多个任务依赖少数几个任务，那么在某个时间段，所述通用维度模型层中可能只有所述少数几个任务在执行，下游节点的所述多个任务都在等待所述少数几个任务结束，这样分布式系统环境下的计算机资源不能被有效利用，数据仓库的数据处理效率不高。

发明内容

[0012] 本发明的目的在于提高一种数据仓库数据处理方法和系统，以提高数据处理效率。

[0013] 一种数据仓库数据处理方法，包括：

[0014] 从业务系统中抽取出原始数据并生成源系统数据表；

[0015] 对源系统数据表去冗余并对数据记历史，建立历史数据表；

[0016] 建立事实数据表对历史数据表的数据进行归类；

[0017] 对事实数据表的数据进行汇总，建立汇总数据表；

[0018] 统计两个或两个以上汇总数据表之间的指标数据，建立统计数据表；

[0019] 基于事实数据表、汇总数据表、统计数据表的数据，建立应用数据表。

[0020] 优选方案中，所述数据处理方法还包括：

[0021] 对历史数据表中的底层数据进行分析归类，建立归类数据表；

[0022] 相应地，所述建立事实数据表对历史数据表的数据进行归类，包括：建立事实数据

表对历史数据表和 / 或归类数据表的数据进行归类。

[0023] 优选方案中，所述建立一个数据表称为一个当前任务；每一任务的初始任务状态为未完成状态。

[0024] 优选方案中，设置一状态标识来表示每一任务的任务状态。

[0025] 优选方案中，在所述建立任一数据表之前，还包括：

[0026] 查询当前任务所依赖的父任务的任务状态；

[0027] 若父任务的状态均为完成状态，执行当前任务；

[0028] 若父任务的状态中至少有一个父任务是未完成状态，在预定时间间隔后，重新查询当前任务所依赖的父任务的任务状态。

[0029] 一种数据仓库数据处理系统，包括：源系统数据处理单元、企业数据仓库第一处理单元、通用维度模型第一处理单元、通用维度模型第二处理单元、通用维度模型第三处理单元、应用数据处理单元；其中，

[0030] 所述源系统数据处理单元，用于从各个业务系统中抽取出原始数据，生成一个或一个以上的源系统数据表；

[0031] 所述企业数据仓库第一处理单元，用于对源系统数据处理单元中的源系统数据表去冗余并对数据记历史，产生与源系统数据表相对应的一个或一个以上的历史数据表；

[0032] 所述通用维度模型第一处理单元，用于建立一个或一个以上的事实表对企业数据仓库第一处理单元的数据进行归类；

[0033] 所述通用维度模型第二处理单元，用于对通用维度模型第一处理单元的数据进行汇总，生成至少一个汇总数据表；

[0034] 所述通用维度模型第三处理单元，用于统计通用维度模型第二处理单元中表与表的指标数据，生成至少一个统计数据表；

[0035] 所述应用数据处理单元，用于根据通用维度模型第一处理单元、通用维度模型第二处理单元、通用维度模型第三处理单元的数据生成应用数据表。

[0036] 优选方案中，

[0037] 所述企业数据仓库第一处理单元调用源系统数据处理单元的结果；

[0038] 所述通用维度模型第一处理单元调用企业数据仓库第一处理单元的结果；

[0039] 所述通用维度模型第二处理单元调用通用维度模型第一处理单元的结果；

[0040] 所述通用维度模型第三处理单元调用通用维度模型第二处理单元的结果；

[0041] 所述应用数据处理单元，调用通用维度模型第一处理单元、通用维度模型第二处理单元、通用维度模型第三处理单元的结果。

[0042] 优选方案中，所述数据仓库数据处理系统，还包括：企业数据仓库第二处理单元；

[0043] 所述企业数据仓库第二处理单元，用于对企业数据仓库第一处理单元中的底层数据进行分析归类，生成归类数据表；

[0044] 相应地，所述通用维度模型第一处理单元，用于建立一个或一个以上的事实表对企业数据仓库第一处理单元、企业数据仓库第二处理单元的数据进行归类；

[0045] 优选方案中，所述企业数据仓库第二处理单元调用企业数据仓库第一处理单元的结果；

[0046] 相应地，所述通用维度模型第一处理单元调用企业数据仓库第一处理单元和 / 或

企业数据仓库第二处理单元的结果。

[0047] 优选方案中，所述数据处理系统中每一处理单元建立一个数据表对应一个任务单元。

[0048] 优选方案中，所述任务单元，包括：父任务单元、当前任务单元；其中，

[0049] 所述父任务单元，用于记录当前任务所依赖的父任务，并查询所述父任务的任务状态；所述父任务的初始状态为未完成状态；

[0050] 所述当前任务单元，用于执行当前任务，当前任务完成后，更改当前任务的任务状态为已完成状态。

[0051] 优选方案中，所述父任务单元，包括：父任务状态记录单元和父任务状态查询单元；

[0052] 所述父任务状态记录单元，用于记录当前任务所依赖的所有父任务；

[0053] 所述父任务状态查询单元，用于查询当前任务所依赖的所有父任务的任务状态；若所有父任务的任务状态均为完成状态，则执行当前任务单元；若所有父任务中至少有一个父任务的任务状态为未完成状态，则等待预定时间后，重新执行父任务状态查询单元，直至所有父任务的任务状态均为完成状态。

[0054] 优选方案中，所述当前任务单元，包括：当前任务执行单元和当前任务状态记录单元；

[0055] 所述当前任务执行单元，用于执行当前任务，即建立一个数据表；

[0056] 所述当前任务状态记录单元，用于记录当然任务的任务状态；所述任务状态的初始状态为未完成状态；在当前任务执行单元中当前任务执行完毕时，更改当前任务的任务状态为已完成状态。

[0057] 优选方案中，所述源系统数据处理单元中每一个任务对应的任务单元，不包括：父任务单元。

[0058] 本申请提供的数据仓库数据处理方法与系统，在现有的数据仓库数据处理方法的基础上将通用维度模型数据处理层分为三层，这就避免了通用维度模型层中每一层级内部的任务相互依赖，使得任务的并行数目达到最大，这样通用维度模型数据处理层中任意一层数据处理任务完成后，数据处理结果也可以被应用层数据处理过程直接调用，这样分布式系统环境下的计算机资源就能够被有效利用，从而提高数据仓库数据处理的效率。

附图说明

[0059] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术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申请中记载的一些实施例，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。

[0060] 图 1 是本申请数据仓库数据处理系统实施例的组成结构图；

[0061] 图 2 是本申请与数据仓库数据处理系统中建立一个数据表对应的任务单元的组成结构示意图；

[0062] 图 3 是任务单元中父任务单元的组成结构图；

[0063] 图 4 是任务单元中当前任务单元的组成结构图；

[0064] 图 5 是本申请数据仓库数据处理方法实施例的流程图；

[0065] 图 6 是对用户浏览这一主题进行数据仓库数据处理的各个任务的依赖关系图。

具体实施方式

[0066] 为了使本技术领域的人员更好地理解本申请中的技术方案，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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，都应当属于本发明保护的范围。

[0067] 下面介绍本申请数据仓库数据处理系统第一实施例。图 1 是本申请数据仓库数据处理系统实施例的组成结构图。如图 1 所示，本申请数据仓库数据处理系统包括：源系统数据处理单元 101、企业数据仓库第一处理单元 102、通用维度模型第一处理单元 103、通用维度模型第二处理单元 104、通用维度模型第三处理单元 105、应用数据处理单元 106。其中，

[0068] 所述源系统数据处理单元 101，用于从各个业务系统中抽取出原始数据，生成一个或一个以上的源系统数据表；

[0069] 所述企业数据仓库第一处理单元 102，用于对源系统数据处理单元 101 中的源系统数据表去冗余并对数据记历史，产生与源系统数据表相对应的一个或一个以上的历史数据表；

[0070] 所述通用维度模型第一处理单元 103，用于建立一个或一个以上的事实数据表对企业数据仓库第一处理单元 102 的数据进行归类；

[0071] 所述通用维度模型第二处理单元 104，用于对通用维度模型第一处理单元 103 的数据进行汇总，生成至少一个汇总数据表；

[0072] 所述通用维度模型第三处理单元 105，用于统计通用维度模型第二处理单元 104 中表与表的指标数据，生成至少一个统计数据表；

[0073] 所述应用数据处理单元 106，用于根据通用维度模型第一处理单元 103、通用维度模型第二处理单元 104、通用维度模型第三处理单元 105 的数据生成应用数据表。

[0074] 所述企业数据仓库第一处理单元 102，可以调用源系统数据处理单元 101 的结果；

[0075] 所述通用维度模型第一处理单元 103，可以调用企业数据仓库第一处理单元 102 的结果；

[0076] 所述通用维度模型第二处理单元 104，可以调用通用维度模型第一处理单元 103 的结果；

[0077] 所述通用维度模型第三处理单元 105，可以调用通用维度模型第二处理单元 104 的结果；

[0078] 所述应用数据处理单元 106，可以调用通用维度模型第一处理单元 103、通用维度模型第二处理单元 104、通用维度模型第三处理单元 105 的结果。

[0079] 下面介绍本申请数据仓库数据处理系统第二实施例，本实施例与数据仓库数据处理系统第一实施例的区别在于，所述数据仓库数据处理系统，还包括：企业数据仓库第二处理单元 107；

[0080] 所述企业数据仓库第二处理单元 107，用于对企业数据仓库第一处理单元中的底

层数据进行分析归类，生成归类数据表；

[0081] 相应地，所述通用维度模型第一处理单元 103，用于建立一个或一个以上的事实表对企业数据仓库第一处理单元 102、企业数据仓库第二处理单元 107 的数据进行归类。

[0082] 所述企业数据仓库第二处理单元 107 可以调用企业数据仓库第一处理单元 102 的结果；

[0083] 相应地，所述通用维度模型第一处理单元 103 可以调用企业数据仓库第一处理单元 102、企业数据仓库第二处理单元 107 的结果。

[0084] 图 2 是本申请与数据仓库数据处理系统中建立一个数据表对应的任务单元的组成结构示意图。如图 2 所示，所述数据仓库数据处理系统中建立一个数据表对应的任务单元，包括：父任务单元 201、当前任务单元 202。其中，

[0085] 所述父任务单元 201，用于记录当前任务所依赖的父任务，并查询所述父任务的任务状态；所述父任务的初始状态为未完成状态；

[0086] 图 3 是任务单元中父任务单元的组成结构图。如图 3 所示，所述父任务单元 201，具体包括：父任务状态记录单元 2011 和父任务状态查询单元 2012；

[0087] 所述父任务状态记录单元 2011，可以用于记录当前任务所依赖的所有父任务；

[0088] 所述父任务状态查询单元 2012，可以用于查询当前任务所依赖的所有父任务的任务状态；若所有父任务的任务状态均为完成状态，则执行当前任务单元；若所有父任务中至少有一个父任务的任务状态为未完成状态，则等待预定时间后，重新执行父任务状态查询单元，直至所有父任务的任务状态均为完成状态。

[0089] 所述当前任务单元 202，用于执行当前任务，当前任务完成后，更改当前任务的任务状态为已完成状态。

[0090] 图 4 是任务单元中当前任务单元的组成结构图。如图 4 所示，所述当前任务单元 202，包括：当前任务执行单元 2021 和当前任务状态记录单元 2022；

[0091] 所述当前任务执行单元 2021，用于执行当前任务，即建立一个数据表；

[0092] 所述当前任务状态记录单元 2022，用于记录当然任务的任务状态；所述任务状态的初始状态为未完成状态；在当前任务执行单元中当前任务执行完毕时，更改当前任务的任务状态为已完成状态。

[0093] 需要说明的是，在数据仓库数据处理系统中，由于所述源系统数据处理单元 101 中的每一个建立数据表的任务是根节点任务，没有需要依赖的父任务，所以所述源系统数据处理单元 101 中每一个任务对应的任务单元，不包括：父任务单元 201。

[0094] 图 5 是本申请数据仓库数据处理方法实施例的流程图。如图 5 所示，所述数据仓库数据处理方法，包括：

[0095] S501：从业务系统中抽取出原始数据并生成源系统数据表。

[0096] 该步骤主要是利用源系统数据处理单元，先从各个业务系统中抽取出数据仓库数据处理所需要的原始数据，所述原始数据可以是数据仓库系统外部或内部的数据。根据抽取出的原始数据建立至少一个数据表并对所述数据表命名，所述数据表即为源系统数据表。在对所述源系统数据表命名时，为了清楚地表示所述源系统数据表为源系统数据处理单元的处理结果，可以对所述一个或多个源系统数据表的名称加一统一的标识，例如在数据表的名称前加一“odl”，所述“odl”表示源系统数据层“operational data layer”。每

个建立数据表的任务完成后，将该任务的任务状态标识由表示任务未完成的字符改为表示任务完成的字符，例如用从表示任务未完成的“0”改为表示任务已完成的“1”。

[0097] 以下述例子进行说明：

[0098] 要对用户浏览这一主题进行数据处理，要求在对数据进行处理时，可以从用户特征维度和用户浏览数据维度来分析。维度一般是指我们分析目标对象所采用的分析角度。所述的用户特征维度可以包括：用户账号信息、用户公司库信息、用户认证信息；所述的用户浏览数据维度包括：页面浏览日志、曝光点击日志。

[0099] 首先源系统数据处理单元从各业务系统中抽取出所需要的原始数据，具体包括：用户账号信息、用户公司库信息、用户认证信息、页面浏览日志、曝光点击日志。所述的用户账号信息、用户公司库信息、用户认证信息来自数据仓库外部各个不同的用户系统。所述的页面浏览日志、曝光点击日志来自专门负责采集用户点击流量数据的日志系统，所述日志系统数据来自数据仓库内部的数据库。

[0100] 根据抽取出的原始数据源系统数据处理单元建立相应的源系统数据表。根据上述抽取到的 5 个维度的数据分别建立 5 个源系统数据表，并对源系统数据处理单元建立的源系统数据表命名，为了清楚地表示所述的源系统数据表为源系统数据处理单元的处理结果，对所述源系统数据表的名称加一统一标识，例如“odl”。那么，所述的 5 个源系统数据表可以分别命名为“odl_用户账号信息”、“odl_用户公司库信息”、“odl_用户认证信息”、“odl_页面浏览日志”、“odl_曝光点击日志”。在每个源系统数据表建立后，将建立该数据表对应的任务的任务状态标识更改为表示完成状态的字符，例如从“0”改为“1”。

[0101] 所述的“odl_用户账号信息”表中，包含了用户账号 id、用户账号状态、用户账号注册日期等信息。所述的“odl_用户公司库信息”表中，包含了用户在公司的信息数据，例如职位等数据。所述的“odl_用户认证信息”表中，包含了用户在接受网站认证时产生的信息数据，例如网站注册信息数据等。所述的“odl_页面浏览日志”表中，包含了用户浏览页面产生的方法日志数据，即包含了每次点击产生一次的页面浏览量(page view，简称PV)数据、浏览页面资源的地址(Uniform Resource Locator，简称URL)数据、浏览时间数据、上个页面的 URL 数据等。所述的“odl_曝光点击日志”表中，包含了页面曝光的每个产品明细数据和点击明细数据。

[0102] S502：对源系统数据表去冗余并对数据记历史，建立历史数据表。

[0103] 由于源系统数据表中的数据是由各个业务系统中直接抽取获得的原始数据，来自不同业务系统中的信息会有重复的冗余信息，需要对信息进行去冗余。同时这些原始数据来自不同的业务系统，因此数据源地址不完全相同，需要将数据的地址变更为当前地址，即对数据记历史。

[0104] 在执行每一建立历史数据表任务之前，企业数据仓库第一处理单元主动查询该建立历史数据表任务所依赖的一个或多个父任务的任务状态，若所述一个或多个父任务的任务状态标识均为表示完成状态的“1”，则开始执行建立历史数据表的任务；若所述一个或多个父任务的任务状态标识至少有一个不是表示完成状态的“1”，则在预定时间间隔后再次查询所述一个或多个父任务的任务状态，直至所述父任务的任务状态标识均为表示完成状态的“1”再执行建立历史数据表的任务。所述任务包括：企业数据仓库第一处理单元先将不同源数据数据表中的冗余信息删除，保证信息的完整、简洁；采用对数据记历史的方式

来更改数据当前地址，保证数据的地址相同，在对数据记历史过程中建立与源系统数据表相对应的一个或多个历史数据表；并对所建立的历史数据表命名；每个建立历史数据表的任务完成后，企业数据仓库第一处理单元将该任务的任务状态标识更改为表示任务完成的字符，例如“1”。

[0105] 在对所述历史数据表命名时，为了清楚地表示所述历史数据表为企业数据仓库第一处理单元的处理结果，可以对所述一个或多个历史数据表的名称加一统一的标识，例如在数据表的名称前加一“edw1”，所述“edw1”中，edw 表示企业数据仓库“enterprise data warehouse”。

[0106] 所述记历史的方法可以是历史拉链的方式，例如：数据 x 从 2000 年 01 月 01 日至 2013 年 05 月 31 日都存放在数据库 1 中，2013 年 06 月 01 数据 x 从数据库 1 搬到数据库 2，则原来关于数据 x 的地址的记录可以是：

[0107] “x, 数据库 1”

[0108] 2013 年 06 月 01 日后，更新地址后的数据 x 的地址的记录可以是：

[0109] “x,数据库 2”

[0110] 在实际应用中，通常还在数据地址记录上增加 begin_date 和 end_date 来表示数据地址有效期的时间，这样数据 x 原来的地址记录可以是：

[0111] “x, 数据库 1,2000.01.01-2013.05.31”

[0112] 数据 x 新的地址记录可以是：

[0113] “x,数据库 2,2013.06.01-2999.12.31”

[0114] 记历史的方式还可以采用快照的方式。以上述的数据 x 为例，日快照的方式是将 2013 年 05 月 31 日和 2013 年 06 月 01 日的关于数据 x 的地址记录分别完整保留下来，每日存一份包含了当日地址的完整数据。

[0115] 具体的记历史的方法一般视情况选择一种合理的记历史方式，比如，如果数据地址变化的不频繁但数据本身的数据量很大，一般采用历史拉链的方式，而如果数据地址变化频繁但数据本身的数据量小，则一般采用快照的方式。

[0116] 以上述对用户浏览这一主题数据处理为例：

[0117] 需要根据“odl_用户账号信息”表中的数据建立历史数据表，所述建立历史数据表的任务为当前任务；那么 S501 中建立“odl_用户账号信息”表即为当前任务所依赖的父任务；首先查询当前任务所依赖的父任务的任务状态，若父任务的任务状态为未完成状态，例如表示任务状态的标识为表示未完成状态的“0”，则等待预定间隔时间后，再次查询父任务的任务状态；当所述父任务的任务状态为完成状态时，例如表示任务状态的标识为表示完成状态的“1”，则开始执行当前任务。所述当前任务包括：

[0118] 将上述“odl_用户账号信息”中重复的内容删除。例如，建立“odl_用户账号信息”表时从业务系统 A 中选择了用户姓名为 M 的信息，从业务系统 B 中又选择了用户姓名为 M 的信息，那么“odl_用户认证信息”表中用户姓名为 M 的信息就存在冗余的信息，需要删除。

[0119] 对上述去冗余的数据表“odl_用户账号信息”中的数据记历史，建立相应的历史数据表。对所述历史数据库表命名，可以命名为“edw1_用户账号信息历史”。所述“edw1_用户账号信息历史”建立完成后，将建立该“edw1_用户账号信息历史”的任务状态标识改为表

示完成状态的字符“1”。用同样的方法建立“edw1_用户公司库信息历史”、“edw1_用户认证信息历史”、“edw1_页面浏览日志快照”、“edw1_曝光点击日志快照”这 4 个历史数据表。

[0120] S503 :建立事实数据表对历史数据表的数据进行归类。

[0121] 在建立历史数据表后,需要根据数据仓库数据处理的主题对一个或一个以上的历史数据表中的数据进行归类。具体地,通用维度模型第一处理单元主动查询每个建立事实数据表的任务所依赖的一个或多个建立历史数据表的任务的任务状态,所述建立历史数据表的任务即为建立事实数据表任务的父任务。若所述父任务的任务状态为完成状态则开始执行该建立事实数据表的任务 ;若所述一个或多个父任务的任务状态中至少有一个不是完成状态,则在预定时间间隔后再次查询父任务的任务状态,直至父任务的任务状态均为完成状态开始执行建立事实数据表的任务。所述事实数据表通常用来描述数据集市中最密集的数据。例如,在电话公司中,用于呼叫的数据是典型的最密集数据。

[0122] 所述建立事实数据表的任务包括 :通用维度模型第一处理单元根据数据仓库数据处理的主题对一个或一个以上的历史数据表中的数据进行归类,建立一个或多个事实数据表,并对所述事实数据表命名 ;每个建立事实数据表的任务完成后,通用维度模型第一处理单元将该任务的任务状态标识更改为表示任务完成的字符,例如“1”。

[0123] 在对所述事实数据表命名时,为了清楚地表示所述事实数据表为通用维度模型第一处理单元的处理结果,可以对所述一个或多个事实数据表的名称加一统一的标识,例如在事实数据表的名称前加一“cdm1”,所述“cdm1”中,cdm 表示通用维度模型“common dimensional model”。

[0124] 以上述用户浏览的主题为例 :

[0125] 例如要通过“edw1_用户账号信息历史”、“edw1_用户公司库信息历史”这两个数据表来对曝光点击事件进行归类,那么所述建立关于曝光点击事件的事实数据表为当前任务,建立“edw1_用户账号信息历史”的任务和建立“edw1_用户公司库信息历史”的任务即为当前任务的父任务。当所述两个父任务的任务状态均为完成状态时,例如两个父任务的任务状态标识均为“1”,则开始执行当前任务。所述当前任务包括 :对“edw1_用户账号信息历史”、“edw1_用户公司库信息历史”这两个表中的数据进行归类,建立曝光点击事件事实数据表,可以将该事实数据表命名为“cdm1_曝光点击事件”。在“cdm1_曝光点击事件”事实表建立完成后,将当前任务的任务状态标识更改为表示完成状态的“1”。用同样的方法建立“cdm1_客体基本特征”、“cdm1_浏览行为事件”、“cdm1_曝光点击关键词”三个事实数据表。

[0126] S504 :对事实数据表的数据进行汇总,建立汇总数据表。

[0127] 通用维度模型第一处理单元中建立的事实数据表仅仅是对历史数据表中的数据进行了描述,还需要对所述事实数据表的数据根据数据处理主题进行简单的数据汇总。

[0128] 具体地,通用维度模型第二处理单元主动查询该步骤中建立汇总数据表的任务所依赖的通用维度模型第一数据处理单元中一个或多个事实数据表对应任务的任务状态,所述一个或多个事实数据表对应任务即为该步骤中建立汇总数据表任务的父任务 ;若所述父任务的任务状态均为完成状态,则开始执行该步骤中的任务 ;若所述一个或多个父任务的任务状态中至少有一个不是完成状态,则在预定时间间隔后再次查询父任务的任务状态,直至父任务的任务状态均为完成状态再执行该步骤中建立汇总数据表的任务。所述建立汇

总数据表的任务包括：对通用维度模型第一处理单元中建立的一个或多个事实表的数据进行汇总，建立汇总数据库表，对所述汇总数据库表命名。在所述每一汇总数据库表建立完成后，将建立该汇总数据表任务的任务状态标识从表示未完成状态的字符更改为表示完成状态的字符，例如从“0”改为“1”。

[0129] 在对所述汇总数据表命名时，为了清楚地表示所述汇总数据表为通用维度模型第二处理单元的处理结果，与通用维度模型第一处理单元的处理结果类似，可以对所述一个或多个汇总数据表的名称加一统一的标识，例如在汇总数据表的名称前加一“cdm2”。

[0130] 以上述用户浏览数据处理主题为例：

[0131] 对用户浏览行为进行数据汇总需要“cdm1_客体基本特征”和“cdm1_浏览行为事件”这两个表的数据，建立关于用户浏览行为的汇总数据表为当前任务，则建立“cdm1_客体基本特征”的任务和建立“cdm1_浏览行为事件”的任务即为当前任务所依赖的父任务。主动查询所述父任务的任务状态，当所述父任务的任务状态均为完成状态时，对“cdm1_客体基本特征”和“cdm1_浏览行为事件”表中的数据进行汇总，建立浏览行为汇总数据表。可以将所述汇总数据表命名为“cdm2_浏览行为汇总数据”。在“cdm2_浏览行为汇总数据”表建立完成后，将建立“cdm2_浏览行为汇总数据”表的任务的任务状态更改为完成状态，例如将该任务的任务状态改为“1”。用同样的方法建立“cdm2_曝光点击效果汇总数据”表和“cdm2_关键词效果汇总数据”表。

[0132] S505：统计两个或两个以上汇总数据表之间的指标数据，建立统计数据表。

[0133] 实施该步骤时，通用维度模型第三处理单元主动查询该步骤中建立每一统计数据表需要的一个或多个事实数据表对应任务的任务状态，所述一个或多个事实数据表对应的任务即为建立统计数据表的任务的父任务，该步骤需要执行的任务为子任务；若所述父任务的任务状态为完成状态，则开始执行建立统计数据表的任务；若所述一个或多个父任务的任务状态中至少有一个不是完成状态，则在预定时间间隔后再次查询父任务的任务状态，直至父任务的任务状态均为完成状态，开始执行建立统计数据表的任务。所述建立统计数据表的任务包括：统计通用维度模型第二处理单元中建立的2个或2个以上汇总数据表之间的指标数据，例如用户实体与行业实体之间的指标数据等，根据指标数据建立统计数据表，并对所述统计数据表命名。当每个统计数据表建立完成后，将表示该任务的任务状态的标识从表示未完成状态的字符更改为表示完成状态的字符，例如从“0”改为“1”。

[0134] 在对所述统计数据表命名时，为了清楚地表示所述统计数据表为通用维度模型第三处理单元的处理结果，与通用维度模型第一处理单元、通用维度模型第二处理单元的处理结果类似，可以对所述一个或多个统计数据表的名称加一统一的标识，例如在统计数据表的名称前加一“cdm3”。

[0135] 以上述对用户浏览主题进行数据处理为例：

[0136] 需要根据“cdm2_用户曝光点击效果汇总数据”和“cdm2_关键词效果汇总数据”这两个汇总数据表中的数据，建立关键词交叉效果统计数据表。所述建立关键词交叉效果统计数据表为当前任务，则建立“cdm2_用户曝光点击效果汇总数据”表的任务和建立“cdm2_关键词效果汇总数据”表的任务为当前任务的父任务。主动查询所述父任务的任务状态，若所述父任务中有一个任务的任务状态为未完成状态，例如建立“cdm2_用户曝光点击效果汇总数据”表的任务的任务状态标识为表示未完成的“0”，则等待预定时间间隔后，再次查

询所述父任务的任务状态，当所述两个父任务的任务状态均为完成状态时，执行当前任务。所述执行当前任务包括：统计“cdm2_用户曝光点击效果汇总数据”和“cdm2_关键词效果汇总数据”两个汇总数据表中关键词与用户之间指标数据，例如选择关键词的人数等数据，建立关键词交叉效果统计数据表；可以将所述关键词交叉效果统计数据表命名为“cdm3_关键词用户交叉效果统计”。在“cdm3_关键词用户交叉效果统计”表建立完成后，将当前任务的任务状态更改为完成状态，例如将任务状态标识的字符更改为“1”。

[0137] S506：基于事实数据表、汇总数据表、统计数据表的数据，建立应用数据表。

[0138] 实施该步骤时，应用数据处理单元主动查询该步骤建立应用数据表需要的通用维度模型第一处理单元、通用维度模型第二处理单元、通用维度模型第三处理单元中建立一个或多个数据表对应的任务的任务状态，所述一个或多个数据表对应的任务即为该步骤中建立应用数据表任务所依赖的父任务；若所述父任务的任务状态为完成状态，则开始执行建立应用数据表的任务；若所述一个或多个父任务的任务状态中至少有一个不是完成状态，则在预定时间间隔后再次查询父任务的任务状态，直至父任务的任务状态均为完成状态，开始执行建立应用数据表的任务。所述建立应用数据表的任务包括：将建立每一应用数据表所依赖的一个或多个事实数据表和/或汇总数据表和/或统计数据表进行分析和合并，生成相应的应用数据表，对所述应用数据表命名。当所述每一建立应用数据表的任务完成后，应用数据处理单元将表示该任务的任务状态的标识从表示未完成状态的字符更改为表示完成状态的字符，例如从“0”改为“1”。

[0139] 在对所述应用数据表命名时，为了清楚地表示所述应用数据表为应用数据处理单元的处理结果，可以对所述一个或多个应用数据表的名称加一统一的标识，例如在应用数据表的名称前加一“adm”，所述“adm”表示通用维度模型“application data model”。

[0140] 以上述对用户浏览主题进行数据处理为例：

[0141] 需要根据“cdm1_浏览行为事件”和“cdm2_用户浏览行为汇总数据”两个数据表来分析出用户流量分析表；则建立用户浏览分析表为当前任务，建立“cdm1_浏览行为事件”的任务和建立“cdm2_用户浏览行为汇总数据”的任务为当前任务所依赖的父任务。首先主动查询所述父任务的任务状态，当所述父任务的任务状态均为完成状态时，开始执行当前任务，所述当前任务包括：根据表“cdm1_浏览行为事件”和表“cdm2_用户浏览行为汇总数据”分析出的用户流量，建立用户流量分析表，可以将所述用户流量分析表命名为“adm_用户流量分析”表，在“adm_用户流量分析”表建立完成后，将当前任务的任务状态更改为完成状态，例如将表示任务状态的标识更改为“1”。用同样的方法建立“adm_用户路径分析数据”和建立“adm_用户关键词研究”。上述三个应用数据表为用户浏览这一主题的相关应用提供业务数据。

[0142] 下面介绍本申请数据仓库数据处理方法第二实施例，该实施例与数据仓库数据处理方法第一实施例的区别在于，所述数据仓库数据处理方法，还包括：

[0143] S507：对历史数据表中的底层数据进行分析归类，建立归类数据表。

[0144] 在历史数据表中，有部分复杂逻辑数据需要作为后续数据处理工作的基础数据，可以对这部分数据进行归类。

[0145] 具体实施过程中，企业数据仓库第二处理单元主动查询该步骤建立归类数据表需要的企业数据仓库第一处理单元中一个或多个历史数据表对应任务的任务状态，所述历史

数据表对应任务即为该步骤中建立归类数据表任务的父任务。若所述父任务的任务状态为完成状态，则开始执行该步骤中建立归类数据表任务的任务；若所述一个或多个父任务的任务状态中至少有一个不是完成状态，则在预定时间间隔后再次查询父任务的任务状态，直至父任务的任务状态均为完成状态，开始执行该步骤中建立归类数据表的任务。所述建立归类数据表任务包括：将所述父任务建立的历史数据表中复杂逻辑数据进行分析和归类，生成归类数据表，对所述归类数据表命名；在每一个归类数据表建立完成后，将该归类数据表对应任务的任务状态标识更改为表示完成状态的字符，例如改为“1”。

[0146] 在对所述归类数据表命名时，为了清楚地表示所述归类数据表为企业数据仓库第二处理单元的处理结果，可以对所述一个或多个归类数据表的名称加一统一的标识，例如在归类数据表的名称前加一“edw2”。

[0147] 相应地，S503 中建立事实数据表时，可以调用所述归类数据表的数据，那么建立所述归类数据表的任务可以作为建立事实数据表任务的父任务。

[0148] 以数据仓库数据处理方法第一实施例中对用户浏览这一主题进行数据仓库数据处理的例子来说：

[0149] 流量来源是基础的、复杂的逻辑数据，需要对流量来源进行归类，这就需要 edw1 中的“edw1_ 页面浏览日快照”表的数据，建立流量来源归类数据表为当前任务，则建立“edw1_ 页面浏览日快照”的任务即为当前任务的父任务。首先主动查询所述父任务的任务状态，当该父任务的任务状态为完成状态时，执行当前任务。所述执行当前任务，包括：利用“edw1_ 页面浏览日快照”表中的数据对流量的来源进行区分归类；所述流量来源可以用一个标识来区分，例如用一个字段来表示流量来源，例如用字段“refer_url”来表示该流量来自哪个 url 来，或者在每个 url 后加一后缀，例如“tracelog”，来表示是否是通过其他浏览页面的链接进入当前浏览页面的；将字段相同或后缀相同的流量来源归为一类。根据对流量来源的归类，建立归类数据表，所述归类数据表可以命名为“edw2_ 流量来源归类”。在建立流量归类数据表建立完成后，可以将当前任务的任务状态更改为完成状态，例如将任务状态标识改为表示任务完成状态的字符“1”。

[0150] 相应地，S503 中，建立“cdm1_ 浏览行为事件”的任务可以将建立“edw2_ 流量来源归类”的任务作为父任务，则在 S503 中，建立“cdm1_ 浏览行为事件”这一事实数据表前，需要主动查询建立“cdm1_ 浏览行为事件”的任务的状态，当所有父任务的任务状态均为完成状态时，开始执行建立“cdm1_ 浏览行为事件”的任务；所述所有父任务包括建立“cdm1_ 浏览行为事件”的任务。

[0151] 图 6 是数据处理方法第二实施例中对用户浏览这一主题进行数据仓库数据处理的各个任务的依赖关系图。从图 6 中可以看出，在建立“cdm1_ 浏览行为事件”事实数据表的任务和建立“浏览行为汇总数据”汇总数据表的任务完成后，就可以执行建立“adm_ 用户流量分析”应用数据表的任务，而不必等待通用维度模型第三处理单元中所有建立统计数据表的任务结束再执行该任务。

[0152] 本申请提供的数据仓库数据处理方法与系统，在现有的数据仓库数据处理方法的基础上将通用维度模型数据处理层分为三层，这就避免了通用维度模型层中每一层级内部的任务相互依赖，使得任务的并行数目达到最大，这样通用维度模型数据处理层中任意一层数据处理任务完成后，数据处理结果也可以被应用层数据处理过程直接调用，这样分布

式系统环境下的计算机资源就能够被有效利用，从而提高数据仓库数据处理的效率。

[0153] 由于本说明书中的系统实施例基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0154] 虽然通过实施例描绘了本发明，本领域普通技术人员知道，本发明有许多变形和变化而不脱离本发明的精神，希望所附的权利要求包括这些变形和变化而不脱离本发明的精神。

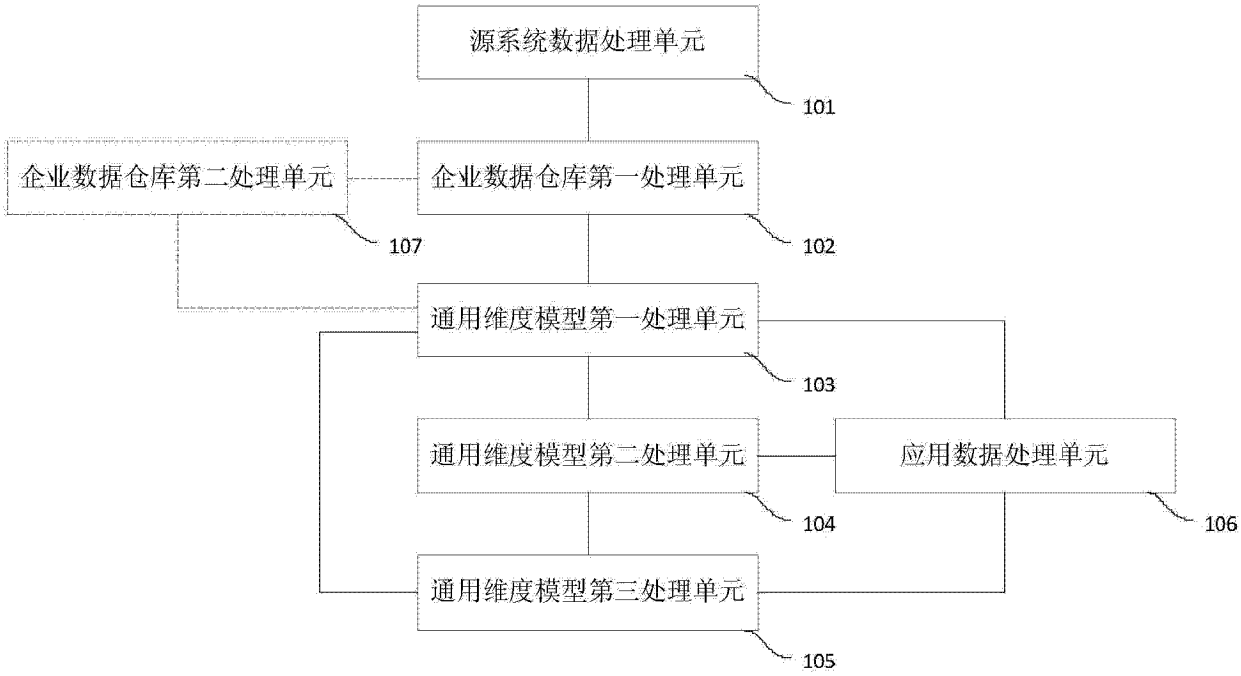

源系统数据处理单元
101
企业数据仓库第二处理单元
企业数据仓库第一处理单元
107
102
通用维度模型第一处理单元
103
通用维度模型第二处理单元
应用数据处理单元
104
106
通用维度模型第三处理单元
105

图 1

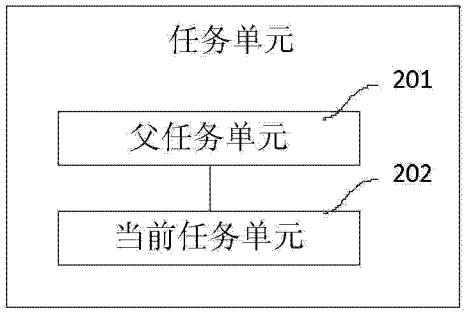

任务单元
201
父任务单元
202
当前任务单元

图 2

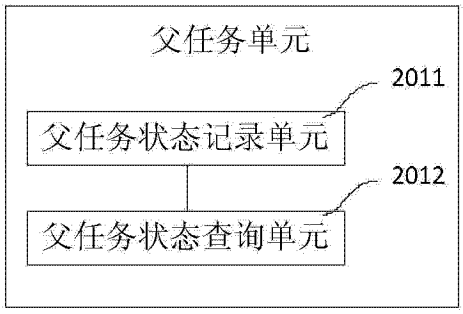

父任务单元
2011
父任务状态记录单元
2012
父任务状态查询单元

图 3

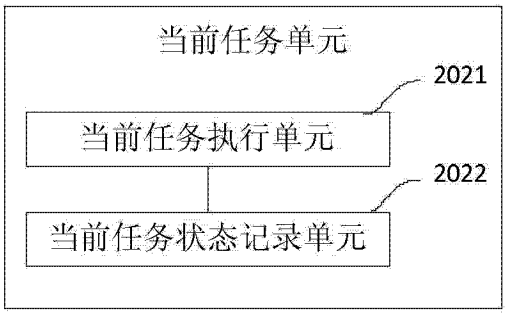

当前任务单元
2021
当前任务执行单元
2022
当前任务状态记录单元

图 4

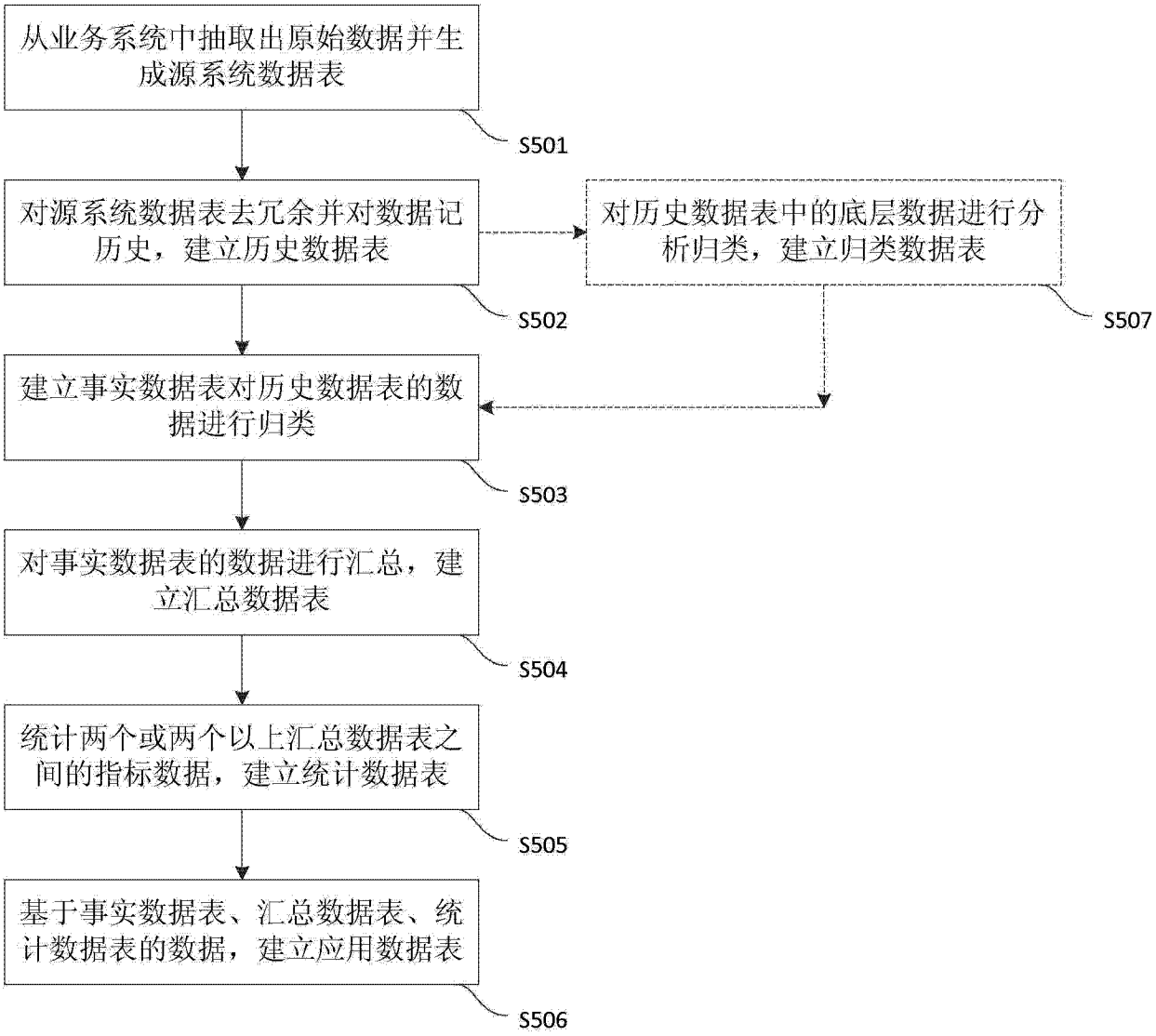

从业务系统中抽取出原始数据并生成源系统数据表
S501
对源系统数据表去冗余并对数据记历史，建立历史数据表
S502
对历史数据表中的底层数据进行分析归类，建立归类数据表
S507
建立事实数据表对历史数据表的数据进行归类
S503
对事实数据表的数据进行汇总，建立汇总数据表
S504
统计两个或两个以上汇总数据表之间的指标数据，建立统计数据表
S505
基于事实数据表、汇总数据表、统计数据表的数据，建立应用数据表
S506

图 5

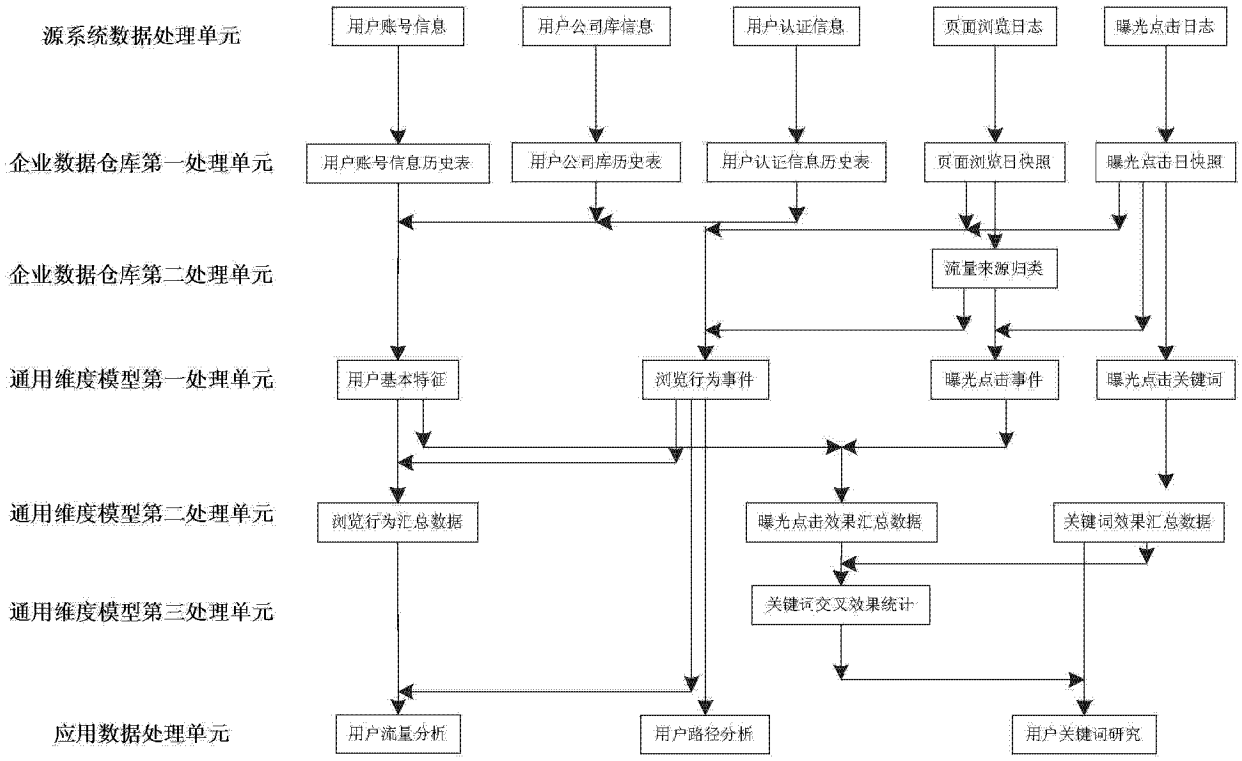

源系统数据处理单元
用户账号信息
用户公司库信息
用户认证信息
页面浏览日志
曝光点击日志
企业数据仓库第一处理单元
用户账号信息历史表
用户公司库历史表
用户认证信息历史表
页面浏览日快照
曝光点击日快照
企业数据仓库第二处理单元
流量来源归类
通用维度模型第一处理单元
用户基本特征
浏览行为事件
曝光点击事件
曝光点击关键词
通用维度模型第二处理单元
浏览行为汇总数据
曝光点击效果汇总数据
关键词效果汇总数据
通用维度模型第三处理单元
关键词交叉效果统计
应用数据处理单元
用户流量分析
用户路径分析
用户关键词研究

图 6