(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 113934786 A
(43)申请公布日 2022.01.14

(21)申请号 202111147981.9

(22)申请日 2021.09.29

(71)申请人 浪潮卓数大数据产业发展有限公司
地址 214029 江苏省无锡市滨湖区高浪东路999号(软件研发大厦)707

(72)发明人 李源 刘方文

(74)专利代理机构 济南信达专利事务所有限公司 37100
代理人 孙晶伟

(51)Int.Cl.
G06F 16/25(2019.01)
G06F 16/28(2019.01)
G06F 8/20(2018.01)
G06F 11/36(2006.01)

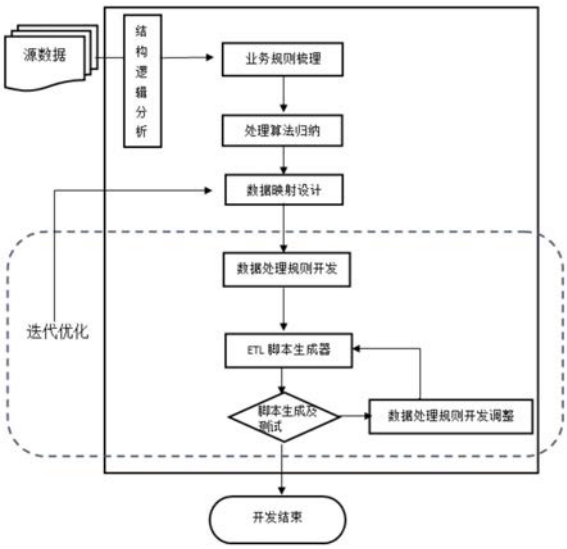
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种构建统一ETL的实施方法

(57)摘要

本发明公开一种构建统一ETL的实施方法，涉及决策分析系统的建设领域，分析数据源的数据内在逻辑结构，根据数据处理的业务目标和处理规则梳理业务规则，并根据业务规则和数据源结构分析总结数据处理算法，设计数据映射，其中参考基础映射的初始化涉及的数据处理源表和目标表基础元数据信息，选择相应的数据处理算法，开发数据规则，其中根据数据映射内容和处理规则将业务规则转化为可执行的代码表达式，根据数据映射内容以及数据处理算法利用脚本生成器生成要执行的ETL脚本，测试脚本并调整处理规则，完成构建ETL。



CN 113934786 A

1.一种构建统一ETL的实施方法,其特征是分析数据源的数据内在逻辑结构,

根据数据处理的业务目标和处理规则梳理业务规则,并根据业务规则和数据源结构分析总结数据处理算法,

设计数据映射,其中参考基础映射的初始化涉及的数据处理源表和目标表基础元数据信息,选择相应的数据处理算法,

开发数据规则,其中根据数据映射内容和处理规则将业务规则转化为可执行的代码表达式,

根据数据映射内容以及数据处理算法利用脚本生成器生成要执行的ETL脚本,

测试脚本并调整处理规则,完成构建ETL。

2.根据权利要求1所述的一种构建统一ETL的实施方法,其特征是所述分析数据源的数据内在逻辑结构之前,包括:

爬取数据源,所述爬取数据源是指只采集网页的列表页数据或采集列表页和翻页数据或者采集列表页、翻页及详情页数据。

3.根据权利要求2所述的一种构建统一ETL的实施方法,其特征是所述分析数据源的数据内在逻辑结构,包括:

判断列表页和翻页数据的逻辑结构是否一致,

判断数据合并后是否存在重复情况,

分析梳理列表页和详情页之间或者翻页与详情页之间的关联关系。

4.根据权利要求1所述的一种构建统一ETL的实施方法,其特征是所述梳理业务规则,包括:

根据源数据的数据内在逻辑结构与业务上的唯一性要求确定数据业务的唯一键,

完成数据统一视图的整合机制,

保证各业务数据之间的逻辑关联性。

5.根据权利要求1所述的一种构建统一ETL的实施方法,其特征是根据所述测试脚本的测试结果,调整处理规则,根据调整后的处理规则生成更新的ETL脚本,迭代优化ETL。

6.一种构建统一ETL的实施系统,其特征是包括分析整理模块及生成模块,

分析整理模块分析数据源的数据内在逻辑结构,

根据数据处理的业务目标和处理规则梳理业务规则,并根据业务规则和数据源结构分析总结数据处理算法,

设计数据映射,其中参考基础映射的初始化涉及的数据处理源表和目标表基础元数据信息,选择相应的数据处理算法,

开发数据规则,其中根据数据映射内容和处理规则将业务规则转化为可执行的代码表达式,

生成模块根据数据映射内容以及数据处理算法利用脚本生成器生成要执行的ETL脚本,

测试脚本并调整处理规则,完成构建ETL。

7.根据权利要求6所述的一种构建统一ETL的实施系统,其特征是还包括爬取模块,在分析数据源的数据内在逻辑结构之前,爬取模块爬取数据源,所述爬取数据源是指只采集网页的列表页数据或采集列表页和翻页数据或者采集列表页、翻页及详情页数据。

8.根据权利要求7所述的一种构建统一ETL的实施系统，其特征是分析整理模块分析数据源的数据内在逻辑结构，包括：

判断列表页和翻页数据的逻辑结构是否一致，

判断数据合并后是否存在重复情况，

分析梳理列表页和详情页之间或者翻页与详情页之间的关联关系。

9.根据权利要求6所述的一种构建统一ETL的实施系统，其特征是分析整理模块梳理业务规则，包括：

根据源数据的数据内在逻辑结构与业务上的唯一性要求确定数据业务的唯一键，

完成数据统一视图的整合机制，

保证各业务数据之间的逻辑关联性。

10.根据权利要求6所述的一种构建统一ETL的实施系统，其特征是生成模块根据所述测试脚本的测试结果，调整处理规则，根据调整后的处理规则生成更新的ETL脚本，迭代优化ETL。

一种构建统一ETL的实施方法

技术领域

[0001] 本发明公开一种方法,涉及决策分析系统的建设领域,具体地说是一种构建统一ETL的实施方法。

背景技术

[0002] 随着大数据时代的蓬勃发展,决策分析系统建设质量要求不断提高,高效高质量的ETL系统直接关系到决策分析系统建设的成败。系统实施过程中数据需求指标灵活易变,业务逻辑规则纷繁复杂,数据处理规则零碎具体,但数据开发人员缺乏统一的技术规范及人员频繁流动,容易导致决策分析系统因ETL实施缺乏质量把控而效率低下,并且准确性无法保证。

发明内容

[0003] 本发明针对现有技术的问题,提供一种构建统一ETL的实施方法,可在OLAP系统的建设实施中,统一地设计和实现ETL数据处理,节约人力开发成本的同时提高开发质量,统一开发规范,让数据处理过程更加规范高效统一。

[0004] 本发明提出的具体方案是:

[0005] 一种构建统一ETL的实施方法,分析数据源的数据内在逻辑结构,

[0006] 根据数据处理的业务目标和处理规则梳理业务规则,并根据业务规则和数据源结构分析总结数据处理算法,

[0007] 设计数据映射,其中参考基础映射的初始化涉及的数据处理源表和目标表基础元数据信息,选择相应的数据处理算法,

[0008] 开发数据规则,其中根据数据映射内容和处理规则将业务规则转化为可执行的代码表达式,

[0009] 根据数据映射内容以及数据处理算法利用脚本生成器生成要执行的ETL脚本,

[0010] 测试脚本并调整处理规则,完成构建ETL。

[0011] 进一步,所述的一种构建统一ETL的实施方法中所述分析数据源的数据内在逻辑结构之前,包括:

[0012] 爬取数据源,所述爬取数据源是指只采集网页的列表页数据或采集列表页和翻页数据或者采集列表页、翻页及详情页数据。

[0013] 进一步,所述的一种构建统一ETL的实施方法中所述分析数据源的数据内在逻辑结构,包括:

[0014] 判断列表页和翻页数据的逻辑结构是否一致,

[0015] 判断数据合并后是否存在重复情况,

[0016] 分析梳理列表页和详情页之间或者翻页与详情页之间的关联关系。

[0017] 进一步,所述的一种构建统一ETL的实施方法中所述梳理业务规则,包括:

[0018] 根据源数据的数据内在逻辑结构与业务上的唯一性要求确定数据业务的唯一键,

[0019] 完成数据统一视图的整合机制，

[0020] 保证各业务数据之间的逻辑关联性。

[0021] 进一步，所述的一种构建统一ETL的实施方法中根据所述测试脚本的测试结果，调整处理规则，根据调整后的处理规则生成更新的ETL脚本，迭代优化ETL。

[0022] 一种构建统一ETL的实施系统，包括分析整理模块及生成模块，

[0023] 分析整理模块分析数据源的数据内在逻辑结构，

[0024] 根据数据处理的业务目标和处理规则梳理业务规则，并根据业务规则和数据源结构分析总结数据处理算法，

[0025] 设计数据映射，其中参考基础映射的初始化涉及的数据处理源表和目标表基础元数据信息，选择相应的数据处理算法，

[0026] 开发数据规则，其中根据数据映射内容和处理规则将业务规则转化为可执行的代码表达式，

[0027] 生成模块根据数据映射内容以及数据处理算法利用脚本生成器生成要执行的ETL脚本，

[0028] 测试脚本并调整处理规则，完成构建ETL。

[0029] 进一步，所述的一种构建统一ETL的实施系统中还包括爬取模块，在分析数据源的数据内在逻辑结构之前，爬取模块爬取数据源，所述爬取数据源是指只采集网页的列表页数据或采集列表页和翻页数据或者采集列表页、翻页及详情页数据。

[0030] 进一步，所述的一种构建统一ETL的实施系统中分析整理模块分析数据源的数据内在逻辑结构，包括：

[0031] 判断列表页和翻页数据的逻辑结构是否一致，

[0032] 判断数据合并后是否存在重复情况，

[0033] 分析梳理列表页和详情页之间或者翻页与详情页之间的关联关系。

[0034] 进一步，所述的一种构建统一ETL的实施系统中分析整理模块梳理业务规则，包括：

[0035] 根据源数据的数据内在逻辑结构与业务上的唯一性要求确定数据业务的唯一键，

[0036] 完成数据统一视图的整合机制，

[0037] 保证各业务数据之间的逻辑关联性。

[0038] 进一步，所述的一种构建统一ETL的实施系统中生成模块根据所述测试脚本的测试结果，调整处理规则，根据调整后的处理规则生成更新的ETL脚本，迭代优化ETL。

[0039] 本发明的有益之处是：

[0040] 本发明提供一种构建统一ETL的实施方法，在实施的过程中保证业务逻辑规则的统一梳理、数据映射的统一设计、处理规则的统一开发、ETL脚本地统一生成；保证从业务需求到实施落地骤统一、质量检查、开发规范从而高质量地完成决策分析系统的搭建。

附图说明

[0041] 图1是本发明方法流程示意图。

具体实施方式

[0042] 下面结合附图和具体实施例对本发明作进一步说明,以使本领域的技术人员可以更好地理解本发明并能予以实施,但所举实施例不作为对本发明的限定。

[0043] 本发明提供一种构建统一ETL的实施方法,分析数据源的数据内在逻辑结构,

[0044] 根据数据处理的业务目标和处理规则梳理业务规则,并根据业务规则和数据源结构分析总结数据处理算法,

[0045] 设计数据映射,其中参考基础映射的初始化涉及的数据处理源表和目标表基础元数据信息,选择相应的数据处理算法,

[0046] 开发数据规则,其中根据数据映射内容和处理规则将业务规则转化为可执行的代码表达式,

[0047] 根据数据映射内容以及数据处理算法利用脚本生成器生成要执行的ETL脚本,

[0048] 测试脚本并调整处理规则,完成构建ETL。

[0049] 能够在OLAP系统的建设实施中,统一地设计和实现ETL数据处理,节约人力开发成本的同时提高开发质量,统一开发规范,让数据处理过程更加规范高效统一。

[0050] 具体应用中,在本发明的一些实施例中构建统一ETL的实施时,具体过程如下:

[0051] 分析数据源的数据内在逻辑结构,其中在分析数据源的数据内在逻辑结构之前,以企业数据的天眼查网站网络爬虫数据为例,爬取源数据,主要包含三种情形,;一是数据只采集网页的列表页数据,二是数据采集了列表页和翻页数据,三是数据采集了列表页/翻页/详情页数据,

[0052] 分析数据源的数据内在逻辑结构时,包括:

[0053] 判断列表页和翻页数据的逻辑结构是否一致,若一致的,这说明两部分数据在逻辑上可以做集合UNION操作,

[0054] 判断数据合并后是否存在重复情况,可以适当删除重复数据,

[0055] 分析梳理列表页和详情页之间或者翻页与详情页之间的关联关系,梳理及分析清楚数据之间的各种关系;

[0056] 梳理业务规则,以天眼查企业数据为例,

[0057] 对数据做业务唯一性处理,即根据源数据的数据内在逻辑结构与业务上的唯一性要求确定数据业务的唯一键,对数据做去重操作,

[0058] 完成数据统一视图的整合机制,即各业务模块完成数据统一视图的整合,例如列表页数据要与详情页数据统一整合,

[0059] 保证各业务数据之间的逻辑关联性,如要天眼查企业数据各业务模块数据要有统一编码规则的主外键;

[0060] 根据业务规则和数据源结构分析总结处理算法,以天眼查企业数据为例,数据处理算法总结P0可为:只有列表页->根据业务逻辑主键去重—>公司名称标准化—>社会统一信用代码/KEY_ID信息补充—>当期数据插入目标表;P1:有列表页和翻页,翻页有公司名称->翻页和列表页合并—>根据业务逻辑主键去重—>公司名称标准化—>社会统一信用代码/KEY_ID信息补充—>当期数据插入目标表;P3:有列表页/翻页/详情页->列表页及翻页与详情页关联—>合并—>根据业务逻辑主键去重—>公司名称标准化—>社会统一信用代码/KEY_ID信息补充—>当期数据插入目标表;

[0061] 设计数据映射,参考基础映射的初始化涉及的数据处理源表和目标表基础元数据信息,选择相应的数据处理算法,

[0062] 以天眼查企业数据为例总结归纳了三种数据处理算法,数据映射设计过程中还考虑数据处理算法选择信息的存储,

[0063] 以及数据处理的映射的字段对应,大部分字段做数据的对应不需要特别的处理,

[0064] 根据业务规则附加订制信息例如天眼查企业数据要根据业务唯一键对数据做去重操作要在映射中标定唯一键;

[0065] 开发数据规则,其中根据数据映射内容和处理规则将业务规则转化为可执行的代码表达式,例如逻辑规则中标明需要多个字段编码生成新的ID,则数据规则开发要编写具体SQL语句做为脚本生成器的输入,最终生成可执行的ETL脚本;

[0066] 根据数据映射内容以及算法的选择做为输入利用脚本生成器就可以方便地生成要执行的ETL脚本,数据的映射信息、数据处理算法流程、数据去重规则等等都将作为知识库内容存储,脚本生成器会将知识库作为输入然后输出可执行的ETL脚本,该体系可以反复迭代优化,不断地将算法归纳、数据映射设计、处理规则开发优化更加方便地生成符合业务实际的程序脚本;

[0067] 脚本测试,调整处理规则,脚本测试中需不断修改处理规则开发内容,最终得到规范、准确、统一的程序,完成ETL构建。

[0068] 本发明还提供一种构建统一ETL的实施系统,包括分析整理模块及生成模块,

[0069] 分析整理模块分析数据源的数据内在逻辑结构,

[0070] 根据数据处理的业务目标和处理规则梳理业务规则,并根据业务规则和数据源结构分析总结数据处理算法,

[0071] 设计数据映射,其中参考基础映射的初始化涉及的数据处理源表和目标表基础元数据信息,选择相应的数据处理算法,

[0072] 开发数据规则,其中根据数据映射内容和处理规则将业务规则转化为可执行的代码表达式,

[0073] 生成模块根据数据映射内容以及数据处理算法利用脚本生成器生成要执行的ETL脚本,

[0074] 测试脚本并调整处理规则,完成构建ETL。

[0075] 上述系统内的各模块之间的信息交互、执行过程等内容,由于与本发明方法实施例基于同一构思,具体内容可参见本发明方法实施例中的叙述,此处不再赘述。

[0076] 同样地,本发明系统能够在OLAP系统的建设实施中,统一地设计和实现ETL数据处理,节约人力开发成本的同时提高开发质量,统一开发规范,让数据处理过程更加规范高效统一。

[0077] 需要说明的是,上述较佳实施例中各流程和各系统结构中不是所有的步骤和模块都是必须的,可以根据实际的需要忽略某些步骤或模块。各步骤的执行顺序不是固定的,可以根据需要进行调整。上述各实施例中描述的系统结构可以是物理结构,也可以是逻辑结构,即,有些模块可能由同一物理实体实现,或者,有些模块可能分由多个物理实体实现,或者,可以由多个独立设备中的某些部件共同实现。

[0078] 以上所述实施例仅是为充分说明本发明而所举的较佳的实施例,本发明的保护范

围不限于此。本技术领域的技术人员在本发明基础上所作的等同替代或变换，均在本发明的保护范围之内。本发明的保护范围以权利要求书为准。

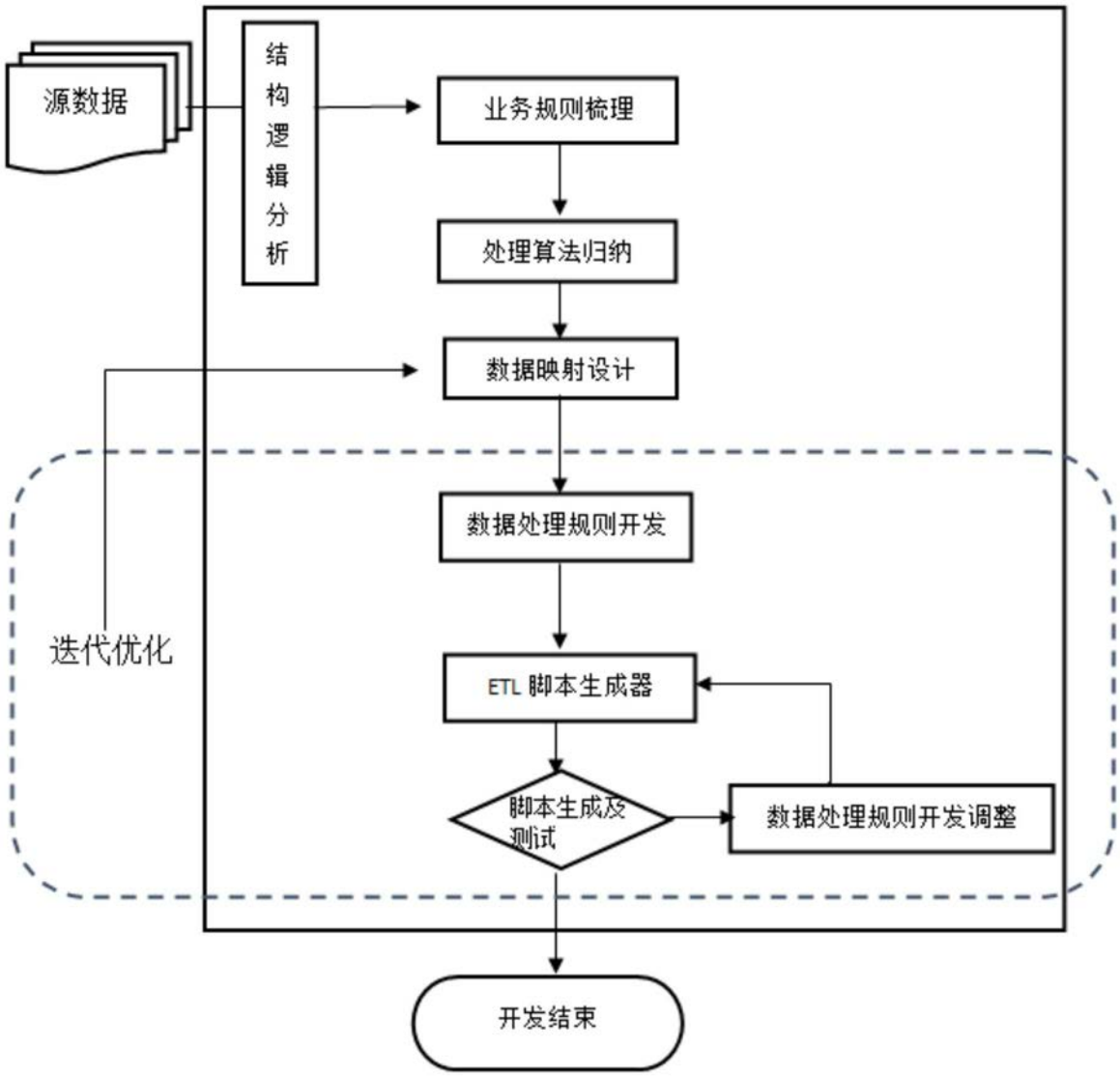

源数据
结构逻辑分析
业务规则梳理
处理算法归纳
数据映射设计
数据处理规则开发
迭代优化
ETL 脚本生成器
脚本生成及测试
数据处理规则开发调整
开发结束

图1