(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 109669975 B
(45)授权公告日 2020.12.18

(21)申请号 201811334481.4
(22)申请日 2018.11.09
(65)同一申请的已公布的文献号
申请公布号 CN 109669975 A
(43)申请公布日 2019.04.23
(73)专利权人 成都数之联科技有限公司
地址 610000 四川省成都市一环路西一段菊乐路口1栋4层2号
(72)发明人 不公告发明人
(74)专利代理机构 成都帝鹏知识产权代理事务所(普通合伙) 51265
代理人 黎照西
(51)Int.Cl.
G06F 16/25(2019.01)

(56)对比文件
CN 104391989 A,2015.03.04
CN 101840563 A,2010.09.22
CN 104317928 A,2015.01.28
CN 102724290 A,2012.10.10
CN 106202346 A,2016.12.07

审查员 史江峰

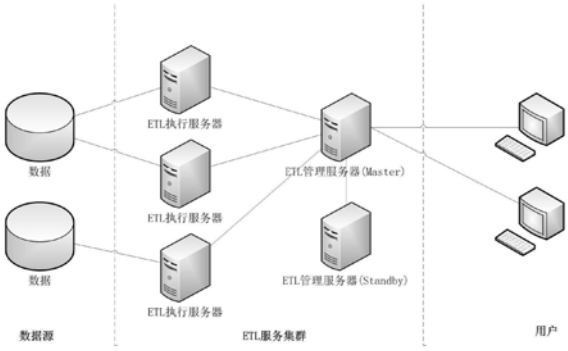
权利要求书2页 说明书5页 附图3页

(54)发明名称

一种工业大数据处理系统及方法

(57)摘要

本发明公开一种工业大数据处理系统及方法,包括ETL服务集群、数据源和用户端,所述用户端和数据源分别连接至ETL服务集群;所述ETL服务集群包括管理服务器和分布式并列的多个执行服务器,每个执行服务器连接至相同或不同的数据源,所有执行服务器均连接至管理服务器。本发明采用分布式、高可用的方式部署工业数据处理的整个过程,以确保系统的健壮性、数据的安全性和一致性;能够快速提升工业数据处理分析速度和精度,能够实现大量数据查询分析。



CN 109669975 B

1.一种工业大数据处理系统,其特征在于,包括ETL服务集群、数据源和用户端,所述用户端和数据源分别连接至ETL服务集群;所述ETL服务集群包括管理服务器和分布式并列的多个执行服务器,每个执行服务器连接至相同或不同的数据源,所有执行服务器均连接至管理服务器;

所述管理服务器包括后台的控制节点和内嵌的Web服务器,提供ETL执行程序的状态管理、分发管理和结果管理;

所述执行服务器是执行ETL程序的容器,接收管理服务器分发的任务,接收外部数据源执行ETL程序,反馈结果至管理服务器;反馈结果中包含该任务执行中的日志操作记录以及性能监控,所述性能监控包括提供执行任务过程中的任务结果和执行日志;

所述管理服务器包括元数据库、DAO模块、业务数据库和数据仓库;所述管理服务器还包括控制管理单元;所述元数据库、业务数据库和数据仓库均连接至DAO模块,所述DAO模块通过控制管理单元连接至WEB操作界面;

所述元数据库,连接数据源,记录执行服务器执行任务的情况,作为抽取工作的核心,储存数据抽取工作所需信息的全部数据;

所述DAO模块,用于指定访问的数据库和数据表;

所述业务数据库,连接数据源,获取外部业务系统数据,储存数据整合所需的全部数据;

所述数据仓库,将获取的数据源数据经过清洗和转换后,按照主题载入到根据业务逻辑和数据关联低度主题化构成的主题数据表中;

所述控制管理单元,通过WEB操作界面控制执行服务器执行ETL程序的状态管理、分发管理和结果管理。

2.根据权利要求1所述的一种工业大数据处理系统,其特征在于,所述元数据库数据包括生产履历数据、不良检测数据、特征值数据和测量数据。

3.根据权利要求2所述的一种工业大数据处理系统,其特征在于,所述数据仓库中包括实现增量的数据导入功能模块,以时间增量和字段增量导入数据;当数据增量进入到ODS层实现业务处理逻辑,整合完成后将筛选完成的数据导入数据事实表。

4.根据权利要求3所述的一种工业大数据处理系统,其特征在于,所述数据仓库是主题化的随时间逐渐增长的数据集合体,包括贴源层、ODS层、数据仓库层、数据集市层和数据应用层;

所述贴源层:存储由元数据库和业务数据库中选定的表中的数据,包含抽取时间在内的时间戳;

所述ODS层:是数据到数据仓库层之前的过度存储空间,进行数据清洗和数据转换操作,完成业务逻辑筛选;

所述数据仓库层:数据进行了轻量的主题化,并按照时间戳将字段进行了分区处理;

所述数据集市层:数据进行了高度的主题化,根据维度的不同围绕事实表形成大小不同的星型结构;

所述数据应用层:提供报表数据和数据分析所需的数据,包括定制化的数据需求。

5.根据权利要求4所述的一种工业大数据处理系统,其特征在于,所述数据集市层包括不良率主题数据、停留时间主题数据、版型相关主题数据和特征值数据。

6.根据权利要求5所述的一种工业大数据处理系统，其特征在于，所述数据集市层的星型结构包括一个事实表和连接至事实表的多个支持维表；所述事实表描述数据集市中最密集的数据，根据主题的不同设置有不同的事实表。

7.根据权利要求6所述的一种工业大数据处理系统，其特征在于，在所述数据仓库储存中对数据进行压缩处理，采用snappy格式作为数据的压缩格式进行保存。

8.根据权利要求2-7中任一所述的一种工业大数据处理系统，其特征在于，所述控制管理单元包括并列设置的元数据管理模块、数据库操作模块、抽取任务配置模块和ETL管理模块，所述元数据管理模块、数据库操作模块和抽取任务配置模块均连接至所述DAO模块，所述ETL管理模块通过工作流引擎模块连接至DAO模块；

所述元数据管理模块，配合元数据库，将每一个工业执行流程中的数据记录至元数据库中；

所述数据库操作模块，控制数据仓库进行数据导入和处理；

所述抽取任务配置模块，配置任务参数；

所述ETL管理模块，通过工作流引擎模块将运行脚本的指令进行解析和运行后，进行ETL程序管理。

9.一种工业大数据处理方法，其特征在于，基于权利要求1-8所述的一种工业大数据处理系统，工业数据处理方法包括步骤：

执行服务器接收管理服务器分发的任务，接收外部数据源执行ETL程序，反馈结果至管理服务器；由管理服务器进行ETL执行程序的状态管理、分发管理和结果管理；

管理服务器通过元数据库，记录执行任务的情况，作为抽取工作的核心，储存数据抽取工作所需信息的全部数据；通过DAO模块，指定访问的数据库和数据表；通过业务数据库，获取外部数据；通过数据仓库，将元数据库和业务数据库中的数据经过清洗和转换后，按照主题载入到根据业务逻辑和数据关联低度主题化构成的主题数据表中；所有工业数据源经过各类执行服务器执行ETL程序后，汇总于数据仓库和数据集市。

一种工业大数据处理系统及方法

技术领域

[0001] 本发明属于工业数据处理技术领域，特别是涉及一种工业大数据处理系统及方法。

背景技术

[0002] 数据是现代企业的重要资源，是企业运用科学管理、决策分析的基础。目前，大多数企业花费大量的资金和时间来构建联机事务处理系统，用来记录事务处理的各种相关数据。据统计，数据量每2～3年就会成倍增长，这些数据蕴含着巨大的商业价值，而企业所利用的通常只占在总数据量的2%～4%，仍没有最大化地利用已存在的数据资源。因此，企业如何通过各种技术手段把数据转换为信息和知识，已经成为提高其核心竞争力的主要瓶颈。

[0003] 由于工厂生产工业过程比较复杂，且入库时间不能反映业务增量，无法进行业务分区转换。面对海量和复杂的工业数据，常用的数据处理方统计分析速度慢，难以进行大量数据查询分析。

发明内容

[0004] 为了解决上述问题，本发明提出了一种工业大数据处理系统及方法，工业数据处理的整个过程采用分布式、高可用的方式进行部署，以确保系统的健壮性、数据的安全性和一致性；能够快速提升工业数据处理分析速度和精度，能够实现大量数据查询分析。

[0005] 为达到上述目的，本发明采用的技术方案是：一种工业大数据处理系统，包括ETL服务集群、数据源和用户端，所述用户端和数据源分别连接至ETL服务集群；所述ETL服务集群包括管理服务器和分布式并列的多个执行服务器，每个执行服务器连接至相同或不同的数据源，所有执行服务器均连接至管理服务器；

[0006] 所述管理服务器包括后台的控制节点和内嵌的Web服务器，提供ETL执行程序的状态管理、分发管理和结果管理；

[0007] 所述执行服务器是执行ETL程序的容器，接收管理服务器分发的任务，接收外部数据源执行ETL程序，反馈结果至管理服务器；反馈结果中包含该任务执行中的日志操作记录以及性能监控，所述性能监控包括提供执行任务过程中的任务结果和执行日志。

[0008] 进一步的是，所述管理服务器包括元数据库、DAO模块、业务数据库和数据仓库；所述管理服务器包括控制管理单元；所述元数据库、业务数据库和数据仓库均连接至DAO模块，所述DAO模块通过控制管理单元连接至WEB操作界面；

[0009] 所述元数据库，连接数据源，记录执行服务器执行任务的情况，作为抽取工作的核心，储存数据抽取工作所需信息的全部数据；

[0010] 所述DAO模块，用于指定访问的数据库和数据表；

[0011] 所述业务数据库，连接数据源，获取外部业务系统数据，储存数据整合所需的全部数据；

[0012] 所述数据仓库，将获取的数据源数据经过清洗和转换后，按照主题载入到根据业务逻辑和数据关联低度主题化构成的主题数据表中；

[0013] 所述控制管理单元，通过WEB操作界面控制执行服务器执行ETL程序的状态管理、分发管理和结果管理。

[0014] 进一步的是，所述元数据库数据包括生产履历数据、不良检测数据、特征值数据和测量数据。

[0015] 进一步的是，在所述数据仓库中包括实现增量的数据导入功能模块，以时间增量和字段增量导入数据；当数据增量进入到ODS层实现业务处理逻辑，整合完成后将筛选完成的数据导入数据事实表。本发明中采取增量导入数据，基本处理逻辑主要实现于ODS层；由于现有业务较为复杂，特别是单个产品单元的处理在一段时间内可能反复多次，所以取值时间并非由某一天来确定，而是在一个时间段内按照一定规则进行处理后再增量写入；对于不同的数据量的表，我们的增量单位也不同，有按天增量和按月增量两种方式，所对应的增量字段也不一样。

[0016] 进一步的是，所述数据仓库是主题化的随时间逐渐增长的数据集合体，包括贴源层、ODS层、数据仓库层、数据集市层和数据应用层；

[0017] 所述贴源层：存储由元数据库和业务数据库中选定的表中的数据，包含抽取时间在内的时间戳；

[0018] 所述ODS层：是数据到数据仓库层之前的过度存储空间，进行数据清洗和数据转换操作，完成业务逻辑筛选；

[0019] 所述数据仓库层：数据进行了轻量的主题化，并按照时间戳将字段进行了分区处理，提高了查询效率，减小了运行内存需求；

[0020] 所述数据集市层：数据进行了高度的主题化，根据维度的不同围绕事实表形成大小不同的星型结构，聚合程度高，极大的提高了查询速度；

[0021] 所述数据应用层：提供报表数据和数据分析所需的数据，包括定制化的数据需求。

[0022] 进一步的是，所述数据集市层包括不良率主题数据、停留时间主题数据、版型相关主题数据和特征值数据。

[0023] 进一步的是，所述数据集市层的星型结构包括一个事实表和连接至事实表的多个支持维表；所述事实表描述数据集市中最密集的数据，根据主题的不同设置有不同的事实表；通过大量的冗余来提升查询效率，支撑OLAP分析。

[0024] 进一步的是，在所述数据仓库储存中对数据进行压缩处理，采用snappy格式作为数据的压缩格式进行保存；为节省存储空间，需要对数据进行压缩处理；由于压缩效率越高，数据的查询速度越慢，为了达到压缩效率与查询速度之间的平衡，选定snappy格式作为数据的压缩格式进行保存。

[0025] 进一步的是，所述控制管理单元包括并列设置的元数据管理模块、数据库操作模块、抽取任务配置模块和ETL管理模块，所述元数据管理模块、数据库操作模块和抽取任务配置模块均连接至所述DAO模块，所述ETL管理模块通过工作流引擎模块连接至DAO模块；

[0026] 所述元数据管理模块，配合元数据库，将每一个工业执行流程中的数据记录至元数据库中；

[0027] 所述数据库操作模块，控制数据仓库进行数据导入和处理；

[0028] 所述抽取任务配置模块，配置任务参数，如任务开始时间；

[0029] 所述ETL管理模块，通过工作流引擎模块将运行脚本的指令进行解析和运行后，进行ETL程序管理。

[0030] 另一方面，本发明还提供了一种工业大数据处理方法，包括步骤：

[0031] 执行服务器接收管理服务器分发的任务，接收外部数据源执行ETL程序，反馈结果至管理服务器；由管理服务器进行ETL执行程序的状态管理、分发管理和结果管理；

[0032] 管理服务器通过元数据库，记录执行任务的情况，作为抽取工作的核心，储存数据抽取工作所需信息的全部数据；通过DAO模块，指定访问的数据库和数据表；通过业务数据库，获取外部数据；通过数据仓库，将元数据库和业务数据库中的数据经过清洗和转换后，按照主题载入到根据业务逻辑和数据关联低度主题化构成的主题数据表中；所有工业数据源经过各类执行服务器执行ETL程序后，汇总于数据仓库和数据集市。

[0033] 采用本技术方案的有益效果：

[0034] 本发明在工业数据处理的整个过程采用分布式、高可用的方式部署，以确保系统的健壮性、数据的安全性和一致性；能够快速提升工业数据处理分析速度和精度，能够实现大量工业数据查询分析；

[0035] 本发明将所有工业数据源经过各类执行服务器执行ETL程序后，汇总于数据仓库和数据集市；能够快速提升性能和处理后续ETL业务，需要在数据融合接入端进行相应的业务数据分区转换；大大提升了工业大数据处理的效率。

附图说明

[0036] 图1为本发明的一种工业大数据处理系统的结构示意图；

[0037] 图2为本发明的一种工业大数据处理系统的逻辑示意图；

[0038] 图3为本发明的数据仓库中数据集市层的结构示意图。

具体实施方式

[0039] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图对本发明作进一步阐述。

[0040] 在本实施例中，参见图1所示，本发明提出了一种工业大数据处理系统，包括ETL服务集群、数据源和用户端，所述用户端和数据源分别连接至ETL服务集群；所述ETL服务集群包括管理服务器和分布式并列的多个执行服务器，每个执行服务器连接至相同或不同的数据源，所有执行服务器均连接至管理服务器；

[0041] 所述管理服务器包括后台的控制节点和内嵌的Web服务器，提供ETL执行程序的状态管理、分发管理和结果管理；

[0042] 所述执行服务器是执行ETL程序的容器，接收管理服务器分发的任务，接收外部数据源执行ETL程序，反馈结果至管理服务器；反馈结果中包含该任务执行中的日志操作记录以及性能监控，所述性能监控包括提供执行任务过程中的任务结果和执行日志。

[0043] 作为上述实施例的优化方案，如图2所示，所述管理服务器包括元数据库、DAO模块、业务数据库和数据仓库；所述管理服务器包括控制管理单元；所述元数据库、业务数据库和数据仓库均连接至DAO模块，所述DAO模块通过控制管理单元连接至WEB操作界面；

[0044] 所述元数据库，连接数据源，记录执行服务器执行任务的情况，作为抽取工作的核心，储存数据抽取工作所需信息的全部数据；

[0045] 所述DAO模块，用于指定访问的数据库和数据表；

[0046] 所述业务数据库，连接数据源，获取外部业务系统数据，储存数据整合所需的全部数据；

[0047] 所述数据仓库，将获取的数据源数据经过清洗和转换后，按照主题载入到根据业务逻辑和数据关联低度主题化构成的主题数据表中；

[0048] 所述控制管理单元，通过WEB操作界面控制执行服务器执行ETL程序的状态管理、分发管理和结果管理。

[0049] 其中，所述元数据库数据包括生产履历数据、不良检测数据、特征值数据和测量数据。

[0050] 作为上述实施例的优化方案，在所述数据仓库ETL中包括实现增量的数据导入功能模块，以时间增量和字段增量导入数据；当数据增量进入到ODS层实现业务处理逻辑，整合完成后将筛选完成的数据导入数据事实表。本发明中采取增量导入数据，基本处理逻辑主要实现于ODS层；由于现有业务较为复杂，特别是单个产品单元的处理在一段时间内可能反复多次，所以取值时间并非由某一天来确定，而是在一个时间段内按照一定规则进行处理后再增量写入；对于不同的数据量的表，我们的增量单位也不同，有按天增量和按月增量两种方式，所对应的增量字段也不一样。

[0051] 作为上述实施例的优化方案，所述数据仓库是主题化的随时间逐渐增长的数据集合体，包括贴源层、ODS层、数据仓库层、数据集市层和数据应用层；

[0052] 所述贴源层：存储由元数据库和业务数据库中选定的表中的数据，包含抽取时间在内的时间戳；

[0053] 所述ODS层：是数据到数据仓库层之前的过度存储空间，进行数据清洗和数据转换操作，完成业务逻辑筛选；

[0054] 所述数据仓库层：数据进行了轻量的主题化，并按照时间戳将字段进行了分区处理，提高了查询效率，减小了运行内存需求；

[0055] 所述数据集市层：数据进行了高度的主题化，根据维度的不同围绕事实表形成大小不同的星型结构，聚合程度高，极大的提高了查询速度。

[0056] 所述数据应用层：提供报表数据和数据分析所需的数据，包括定制化的数据需求。

[0057] 其中，所述数据集市层包括不良率主题数据、停留时间主题数据、版型相关主题数据和特征值数据。

[0058] 如图3所示，所述数据集市层的星型结构包括一个事实表和连接至事实表的多个支持维表；所述事实表描述数据集市中最密集的数据，根据主题的不同设置有不同的事实表；通过大量的冗余来提升查询效率，支撑OLAP分析。

[0059] 作为上述实施例的优化方案，在所述数据仓库储存中对数据进行压缩处理，采用snappy格式作为数据的压缩格式进行保存；为节省存储空间，需要对数据进行压缩处理；由于压缩效率越高，数据的查询速度越慢，为了达到压缩效率与查询速度之间的平衡，选定snappy格式作为数据的压缩格式进行保存。

[0060] 作为上述实施例的优化方案，所述控制管理单元包括并列设置的元数据管理模

块、数据库操作模块、抽取任务配置模块和ETL管理模块，所述元数据管理模块、数据库操作模块和抽取任务配置模块均连接至所述DAO模块，所述ETL管理模块通过工作流引擎模块连接至DAO模块；

[0061] 所述元数据管理模块，配合元数据库，将每一个工业执行流程中的数据记录至元数据库中；

[0062] 所述数据库操作模块，控制数据仓库进行数据导入和处理；

[0063] 所述抽取任务配置模块，配置任务参数，如任务开始时间；

[0064] 所述ETL管理模块，通过工作流引擎模块将运行脚本的指令进行解析和运行后，进行ETL程序管理。

[0065] 为配合本发明方法的实现，基于相同的发明构思，本发明还提供了一种工业大数据处理方法，包括步骤：

[0066] 执行服务器接收管理服务器分发的任务，接收外部数据源执行ETL程序，反馈结果至管理服务器；由管理服务器进行ETL执行程序的状态管理、分发管理和结果管理；

[0067] 管理服务器通过元数据库，记录执行任务的情况，作为抽取工作的核心，储存数据抽取工作所需信息的全部数据；通过DAO模块，指定访问的数据库和数据表；通过业务数据库，获取外部数据；通过数据仓库，将元数据库和业务数据库中的数据经过清洗和转换后，按照主题载入到根据业务逻辑和数据关联低度主题化构成的主题数据表中；所有工业数据源经过各类执行服务器执行ETL程序后，汇总于数据仓库和数据集市。

[0068] 以上显示和描述了本发明的基本原理和主要特征和本发明的优点。本行业的技术人员应该了解，本发明不受上述实施例的限制，上述实施例和说明书中描述的只是说明本发明的原理，在不脱离本发明精神和范围的前提下，本发明还会有各种变化和改进，这些变化和改进都落入要求保护的本发明范围内。本发明要求保护范围由所附的权利要求书及其等效物界定。

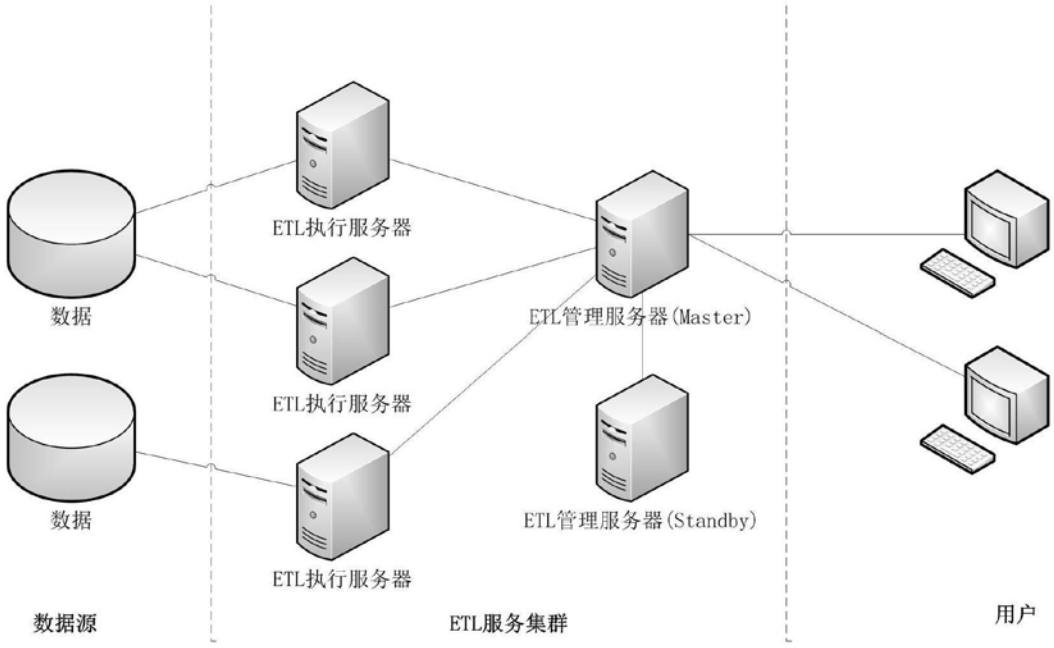

ETL执行服务器
数据
ETL执行服务器
ETL管理服务器(Master)
数据
ETL管理服务器(Standby)
ETL执行服务器
数据源
ETL服务集群
用户

图1

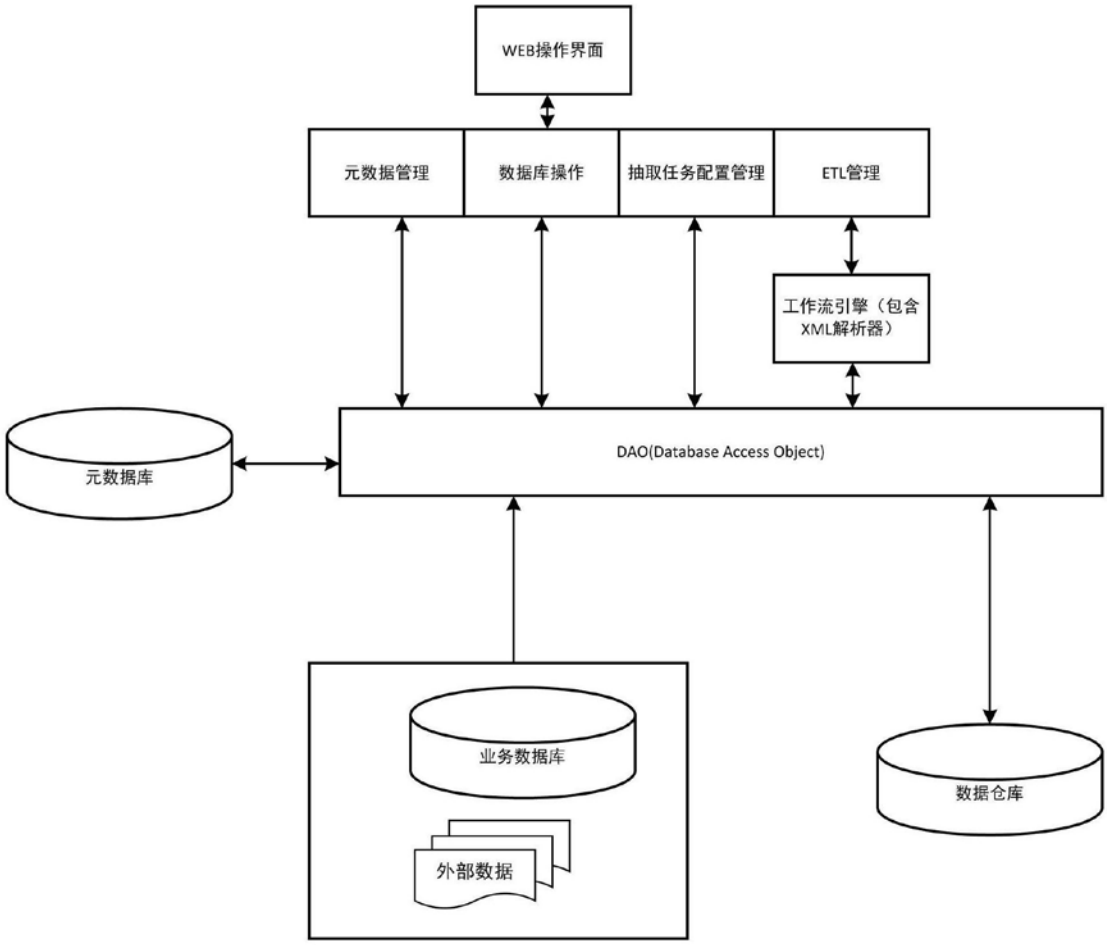

WEB操作界面
元数据管理
数据库操作
抽取任务配置管理
ETL管理
工作流引擎（包含XML解析器）
元数据库
DAO(Database Access Object)
业务数据库
外部数据
数据仓库

图2

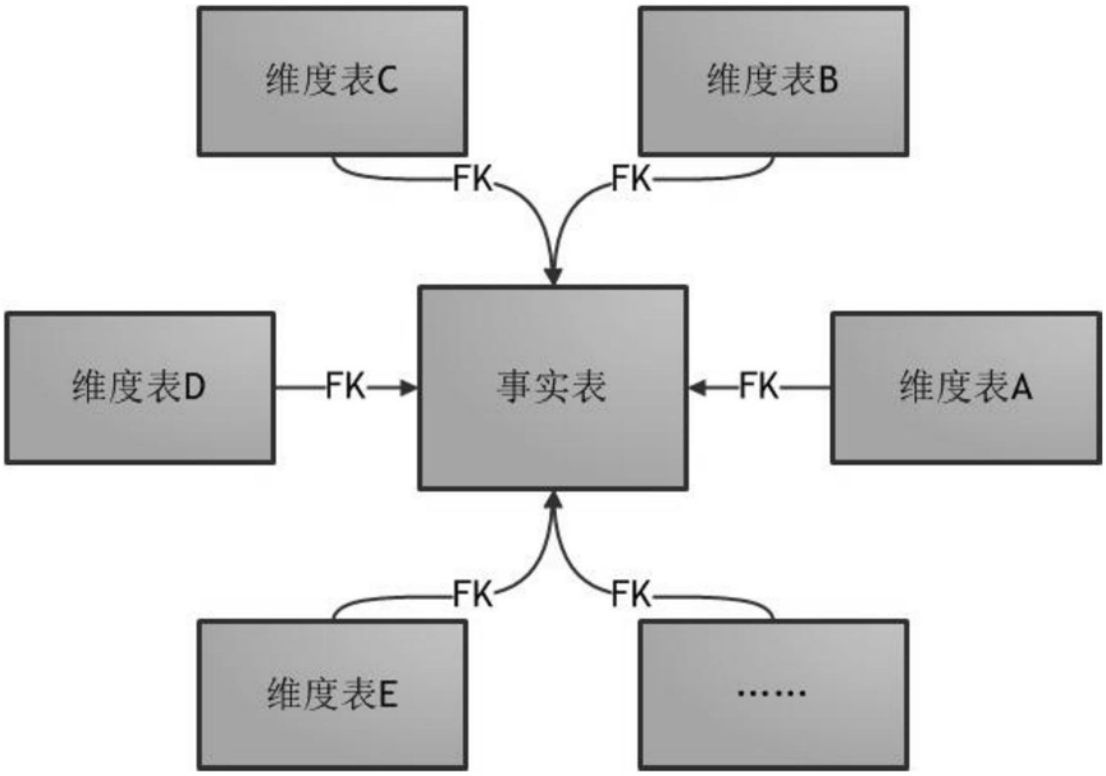


图3