(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 111159272 A
(43)申请公布日 2020.05.15

(21)申请号 201911418960.9

(22)申请日 2019.12.31

(71)申请人 青梧桐有限责任公司
地址 200241 上海市闵行区紫星路599号2幢1128室

(72)发明人 李松前 李昭 陈浩 高靖 崔岩 卢述奇 陈呈 张宵

(74)专利代理机构 北京晟睿智杰知识产权代理事务所(特殊普通合伙) 11603
代理人 于淼

(51)Int.Cl.
G06F 16/25(2019.01)
G06F 16/28(2019.01)
G06F 16/215(2019.01)
G06F 16/26(2019.01)

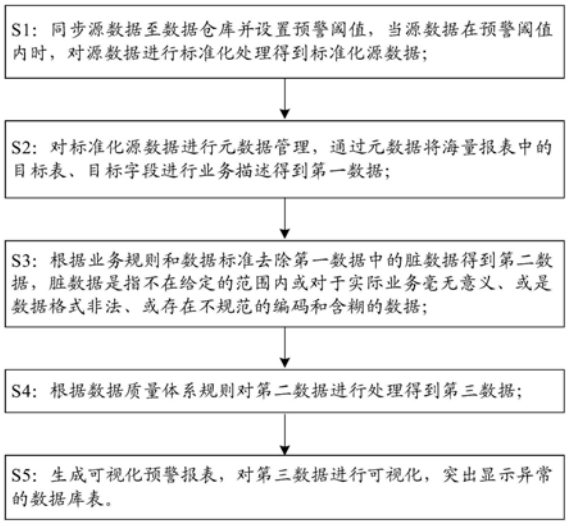
权利要求书2页 说明书7页 附图2页

(54)发明名称

基于数据仓库及ETL的数据质量监控及预警方法和系统

(57)摘要

本发明公开了一种基于数据仓库及ETL的数据质量监控及预警方法和系统，方法包括步骤：同步源数据至数据仓库并设置预警阈值，当所述源数据在所述预警阈值内时，对所述源数据进行标准化处理得到标准化源数据；对标准化源数据进行元数据管理得到第一数据；根据业务规则和数据标准去除第一数据中的脏数据得到第二数据；根据数据质量体系规则对第二数据进行处理得到第三数据；生成可视化预警报表，对第三数据进行可视化，突出显示异常的数据库表。本发明能够在建立数据仓库和ETL过程进行监控数据质量，提高数据的易读性，准确性。



CN 111159272 A

1.一种基于数据仓库及ETL的数据质量监控及预警方法,其特征在于,包括步骤:

同步源数据至数据仓库并设置预警阈值,当所述源数据在所述预警阈值内时,对所述源数据进行标准化处理得到标准化源数据;

对所述标准化源数据进行元数据管理,通过元数据将海量报表中的目标表、目标字段进行业务描述得到第一数据;

根据业务规则和数据标准去除所述第一数据中的脏数据得到第二数据,所述脏数据是指不在给定的范围内或对于实际业务毫无意义、或是数据格式非法、或存在不规范的编码和含糊的数据;

根据数据质量体系规则对所述第二数据进行处理得到第三数据;

生成可视化预警报表,对所述第三数据进行可视化,突出显示异常的数据库表。

2.根据权利要求1所述的基于数据仓库及ETL的数据质量监控及预警方法,其特征在于,所述ETL规则处理包括:在收房明细数据中当收房合同中没有房间编号,则过滤掉此数据;根据身份证号计算年龄时,当身份证号不是正确的格式将其视为垃圾数据并用0替代。

3.根据权利要求1所述的基于数据仓库及ETL的数据质量监控及预警方法,其特征在于,所述数据质量体系规则验证,根据数据质量体系规则对所述第二数据进行处理得到第三数据包括:

获取所述第二数据的数据类型和/或属性;

根据所述第二数据的数据类型和/或属性配置检测规则组合,其中所述配置检测规则组合至少包括一个检测规则;

根据所述规则组合对所述第二数据进行质量检测得到第三数据发送至目的端。

4.根据权利要求3所述的基于数据仓库及ETL的数据质量监控及预警方法,其特征在于,所述检测规则组合包括:主键检查、代码标准检查和业务规则检查。

5.根据权利要求1所述的基于数据仓库及ETL的数据质量监控及预警方法,其特征在于,还包括在配置表中设置人员管理,将所述可视化预警报表中的异常的数据库表发送至所述人员。

6.一种基于数据仓库及ETL的数据质量监控及预警系统,其特征在于,包括源数据标准化处理模块、数据仓库元数据管理模块、ETL规则处理模块、数据质量体系规则验证模块、以及可视化预警报表生成模块,其中,

所述源数据标准化处理模块与所述数据仓库元数据管理模块相耦接,用于将同步到数据仓库中的源数据进行标准化处理得到标准化源数据,并将所述标准化源数据发送至所述数据仓库元数据管理模块;

所述数据仓库元数据管理模块分别与所述源数据标准化处理模块和所述ETL规则处理模块相耦接,用于对所述标准化源数据进行元数据管理,通过元数据将海量报表中的目标表、目标字段进行业务描述得到第一数据;

所述ETL规则处理模块分别与所述数据仓库元数据管理模块和所述数据质量体系规则验证模块相耦接,用于根据业务规则和数据标准去除所述第一数据中的脏数据得到第二数据,所述脏数据是指不在给定的范围内或对于实际业务毫无意义、或是数据格式非法、或存在不规范的编码和含糊的数据;

所述数据质量体系规则验证模块分别与所述ETL规则处理模块和所述可视化预警报表

生成模块相耦接，用于根据数据质量体系规则对所述第二数据进行处理得到第三数据；

所述可视化预警报表生成模块与所述数据质量体系规则验证模块相耦接，对所述第三数据进行可视化，突出显示异常的数据库表。

7.根据权利要求6所述的基于数据仓库及ETL的数据质量监控及预警系统，其特征在于，所述ETL规则处理模块用于在收房明细数据中当收房合同中没有房间编号，则过滤掉此数据；根据身份证号计算年龄时，当身份证号不是正确的格式将其视为垃圾数据并用0替代。

8.根据权利要求6所述的基于数据仓库及ETL的数据质量监控及预警系统，其特征在于，所述数据质量体系规则验证模块用于根据数据质量体系规则对所述第二数据进行处理得到第三数据包括：

获取所述第二数据的数据类型和/或属性；

根据所述第二数据的数据类型和/或属性配置检测规则组合，其中所述配置检测规则组合至少包括一个检测规则；

根据所述规则组合对所述第二数据进行质量检测得到第三数据发送至目的端。

9.根据权利要求8所述的基于数据仓库及ETL的数据质量监控及预警系统，其特征在于，所述检测规则组合包括：主键检查、代码标准检查和业务规则检查。

10.根据权利要求6所述的基于数据仓库及ETL的数据质量监控及预警系统，其特征在于，还包括可视化预警报表发送模块，用于在配置表中设置人员管理，将所述可视化预警报表中的异常的数据库表发送至所述人员。

基于数据仓库及ETL的数据质量监控及预警方法和系统

技术领域

[0001] 本发明涉及计算机领域，更具体地，涉及一种基于数据仓库及ETL的数据质量监控及预警方法和系统。

背景技术

[0002] 大数据时代的到来，带给企业很多的数据资产，企业需要在众多数据中利用有效数据来进行分析和数据挖掘，而在这个过程中，会导致很多问题。目前本领域主要存在以下问题：

[0003] 1、由于企业数据来自不同的业务系统，上游数据源数据出现异常，例如爬虫数据出现结构变化，爬虫失败等导致下游ETL数据错误；

[0004] 2、在建立数据仓库和ETL过程中元数据缺乏有效管理，导致数据的易读性很差，不能最大能力发挥元数据的功能性；

[0005] 3、数据表中数据质量问题，主要突出表现在以下几种：

[0006] 1)数据缺乏有效的主键，导致数据出现重复；

[0007] 2)数据不符合标准数据类型；

[0008] 3)数据不符合标准数据业务规则，例如数据的区间为1-100，表中出现了100以上的数据；

[0009] 4)数据表主要指标，例如每天的业务量不正常，例如在职人数每天大概在2000人左右，由于数据问题出现某天为100人左右等异常数据；

[0010] 4、企业没有有效的对数据仓库和ETL过程中进行有效的监控和预警，数据开发人员不能快速的了解目前数据的情况，数据异常得不到及时通知，引起下游相关引用该数据出现问题，给数据分析和决策人员带来错误的指导。

[0011] 有鉴于此，克服该现有技术所存在的缺陷是本领域亟待解决的问题。

发明内容

[0012] 有鉴于此，本发明提供了一种基于数据仓库及ETL的数据质量监控及预警方法和系统，使得在建立数据仓库和ETL过程进行监控数据质量，提高数据的易读性，准确性。

[0013] 一方面本发明提供了一种基于数据仓库及ETL的数据质量监控及预警方法，包括步骤：

[0014] 同步源数据至数据仓库并设置预警阈值，当所述源数据在所述预警阈值内时，对所述源数据进行标准化处理得到标准化源数据；

[0015] 对所述标准化源数据进行元数据管理，通过元数据将海量报表中的目标表、目标字段进行业务描述得到第一数据；

[0016] 根据业务规则和数据标准去除所述第一数据中的脏数据得到第二数据，所述脏数据是指不在给定的范围内或对于实际业务毫无意义、或是数据格式非法、或存在不规范的编码和含糊的数据；

[0017] 根据数据质量体系规则对所述第二数据进行处理得到第三数据；

[0018] 生成可视化预警报表，对所述第三数据进行可视化，突出显示异常的数据库表。

[0019] 基于同一发明思想，本发明还提供了一种基于数据仓库及ETL的数据质量监控及预警系统，包括源数据标准化处理模块、数据仓库元数据管理模块、ETL规则处理模块、数据质量体系规则验证模块、以及可视化预警报表生成模块，其中，

[0020] 所述源数据标准化处理模块与所述数据仓库元数据管理模块相耦接，用于将同步到数据仓库中的源数据进行标准化处理得到标准化源数据，并将所述标准化源数据发送至所述数据仓库元数据管理模块；

[0021] 所述数据仓库元数据管理模块分别与所述源数据标准化处理模块和所述ETL规则处理模块相耦接，用于对所述标准化源数据进行元数据管理，通过元数据将海量报表中的目标表、目标字段进行业务描述得到第一数据；

[0022] 所述ETL规则处理模块分别与所述数据仓库元数据管理模块和所述数据质量体系规则验证模块相耦接，用于根据业务规则和数据标准去除所述第一数据中的脏数据得到第二数据，所述脏数据是指不在给定的范围内或对于实际业务毫无意义、或是数据格式非法、或存在不规范的编码和含糊的数据；

[0023] 所述数据质量体系规则验证模块分别与所述ETL规则处理模块和所述可视化预警报表生成模块相耦接，用于根据数据质量体系规则对所述第二数据进行处理得到第三数据；

[0024] 所述可视化预警报表生成模块与所述数据质量体系规则验证模块相耦接，对所述第三数据进行可视化，突出显示异常的数据库表。

[0025] 与现有技术相比，本发明提供的基于数据仓库及ETL的数据质量监控及预警方法和系统，至少实现了如下的有益效果：

[0026] 本发明能够在建立数据仓库和ETL过程进行监控数据质量，提高数据的易读性，准确性；

[0027] 因为本发明采用对源数据先后经过源数据标准化处理、数据仓库元数据管理、ETL规则处理以及数据质量体系规则验证，可以提高检测质量，挺高检测效果以及检测精度，使管理者方便对目前数据资产和数据仓库对相关质量的判定，能够指导管理者和开发者对改善数据质量提供更明细的数据问题定位；能够使数据易懂易读，便于数据使用者；

[0028] 本发明可以生成可视化预警报表，突出显示异常的数据库表，方便对数据进行监控。

[0029] 当然，实施本发明的任一产品必不特定需要同时达到以上所述的所有技术效果。

[0030] 通过以下参照附图对本发明的示例性实施例的详细描述，本发明的其它特征及其优点将会变得清楚。

附图说明

[0031] 被结合在说明书中并构成说明书的一部分的附图示出了本发明的实施例，并且连同其说明一起用于解释本发明的原理。

[0032] 图1是本发明提供的一种基于数据仓库及ETL的数据质量监控及预警方法流程图；

[0033] 图2是本发明提供的一种基于数据仓库及ETL的数据质量监控及预警系统结构框

图。

具体实施方式

[0034] 现在将参照附图来详细描述本发明的各种示例性实施例。应注意到:除非另外具体说明,否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本发明的范围。

[0035] 以下对至少一个示例性实施例的描述实际上仅仅是说明性的,决不作为对本发明及其应用或使用的任何限制。

[0036] 对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论,但在适当情况下,所述技术、方法和设备应当被视为说明书的一部分。

[0037] 在这里示出和讨论的所有例子中,任何具体值应被解释为仅仅是示例性的,而不是作为限制。因此,示例性实施例的其它例子可以具有不同的值。

[0038] 应注意到:相似的标号和字母在下面的附图中表示类似项,因此,一旦某一项在一个附图中被定义,则在随后的附图中不需要对其进行进一步讨论。

[0039] 结合图1,图1是本发明提供的一种基于数据仓库及ETL的数据质量监控及预警方法流程图。

[0040] 图1中的基于数据仓库及ETL的数据质量监控及预警方法包括以下步骤:

[0041] S1:同步源数据至数据仓库并设置预警阈值,当源数据在预警阈值内时,对源数据进行标准化处理得到标准化源数据;

[0042] 根据各业务系统不同,在制定数据传输过程中加入预警判断;

[0043] 例如当爬虫系统异常时,根据业务人员制定的数据量标准,如低于某个最小值,系统根据条件再决定是否进行下一步;

[0044] S2:对标准化源数据进行元数据管理,通过元数据将海量报表中的目标表、目标字段进行业务描述得到第一数据;

[0045] 在海量报表中查询某些表或者相关字段,由于开发人员没有维护好元数据,导致数据使用人员难以发现自己业务对应相关数据表,及相关字段的含义;对此在本预警系统中,对各模块相关元数据的统计可视化,例如把业务系统中各库通过元数据统计出哪些表、哪些字段是没有业务描述的,通过监控预警可视化促进开发和业务相关人员完善这些数据指标,使数据更易读易懂。

[0046] S3:根据业务规则和数据标准去除第一数据中的脏数据得到第二数据,脏数据是指不在给定的范围内或对于实际业务毫无意义、或是数据格式非法、或存在不规范的编码和含糊的数据;

[0047] 对数据进行ETL(英文全称:Extract-Transform-Load,中文全称:数据仓库技术)处理,ETL处理的过程大致为从一个数据源中提取数据,将提取的数据转换为一个标准的格式,并加载到另外一个目标数据源的过程。目前,存在多种不同类型的数据源,例如:关系型Mysql、非关系型HBase、数据仓库有Hive、文件存储HDFS、具有存储功能的文件索引服务Elasticsearch;而不同的数据类型的数据源可能会具有不同的接口类型。

[0048] ETL规则处理包括:在收房明细数据中当收房合同中没有房间编号,则过滤掉此数据;根据身份证号计算年龄时,当身份证号不是正确的格式将其视为垃圾数据并用0替代。

[0049] 在数据仓库建设中ETL过程显得很重要,ETL过程中根据业务规则和数据标准首先去除某些脏数据;例如在收房明细数据中如果收房合同中没有房间编号,就可以在抽取到ODS层收房明细的过程中直接加入相关条件过滤掉此数据;也可以根据业务规则处理相关的数据,例如根据身份证号计算年龄时,当身份证号不是正确的格式时,假设身份证号前4位含有字母就视为垃圾数据,对此要加入数据处理,例如置为0等。

[0050] 可以理解的是对数据的处理行为可以划分为操作型数据处理和分析型数据处理,操作型数据处理一般放在传统的数据库(Database,DB)中进行,分析型数据处理则需要在数据仓库(Data Warehouse,DW)中进行。但是并不是所有的数据处理都可以这样划分,对数据的处理需求并不只有这两类,例如,有些操作型处理并不适合放在传统的数据库上完成,也有些分析型处理不适合在数据仓库中进行。这时候需要第三种数据存储体系,操作数据存储(Operational Data Store,ODS)系统就因此产生。它的出现,也将DB～DW两层数据架构转变成DB～ODS～DW三层数据架构。

[0051] ODS是一种数据存储系统,它将来自不同数据源的数据(各种操作型数据库、外部数据源等)通过ETL过程汇聚整合成面向主题的、集成的、企业全局的、一致的数据集合(主要是最新的或者最近的细节数据以及可能需要的汇总数据),用于满足企业准实时的OLAP操作和企业全局的OLTP操作,并为数据仓库提供集成后的数据,将数据仓库系统中的ETL过程下沉到ODS中完成以减轻数据仓库的压力。

[0052] ODS中的数据具有以下4个基本特征:面向主题的:进入ODS的数据是来源于各个操作型数据库以及其他外部数据源,数据进入ODS前必须经过ETL过程(抽取、清洗、转换、加载等);集成的:ODS的数据来源于各个操作型数据库,同时也会在数据清理加工后进行一定程度的综合;可更新的:可以联机修改。这一点区别于数据仓库;当前或接近当前的:“当前”是指数据在存取时刻是最新的,“接近当前”是指存取的数据是最近一段时间得到的。

[0053] 传统的操作型数据库往往只存放企业某一类业务或者某一个部门的数据,因此无法面向企业全局数据的OLTP,而ODS可以实现。因为ODS的数据是面向整个企业进行集成汇总的,克服了原来面向应用的操作型数据库数据分散的缺陷。

[0054] 在数据仓库上进行OALP,往往由于数据量十分庞大而需要较长的时间。而在企业实际应用中,对于一些较低层次的决策,往往并不需要太多的历史数据,可能只需要参考当前的或者接近当前的数据就可以完成,并且要求具有较快的响应时间,因此数据仓库显然无法满足这样的要求,但是ODS可以实现。ODS中不仅有面向企业全局的细节数据和汇总数据,而且规模比数据仓库小,具有较强的实时响应能力。

[0055] S4:根据数据质量体系规则对第二数据进行处理得到第三数据;

[0056] 数据仓库为企业决策提供重要的数据支撑,数据质量的好坏,直接影响公司业务的发展;为了确保数据仓库的准确性通过主键检查、代码标准检查、业务规则检查以及其他检查来进行数据质量的监控;

[0057] 数据质量体系规则验证,根据数据质量体系规则对第二数据进行处理得到第三数据包括:

[0058] 获取第二数据的数据类型和/或属性;

[0059] 根据第二数据的数据类型和/或属性配置检测规则组合,其中配置检测规则组合至少包括一个检测规则;

[0060] 根据规则组合对第二数据进行质量检测得到第三数据发送至目的端。

[0061] 检测规则组合包括:主键检查、代码标准检查和业务规则检查。

[0062] 开发人员和相关业务人员制定相关的数据质量的管理规则,录入系统中,开发人员根据这些规则转换为脚本执行,把检查记录放入检查日志表中,根据检查日志表和数据质量体系的规则处理形成检查结果表,为可视化表系统和error通知做数据支撑;在规则体系中数据量检查主要为了确保数据的完整性,判断是不是当天数据为空表,是否影响下游数据的引用;

[0063] 在一些可选的实施例中,检查规则包括:

[0064] 主键检查:确保主键的唯一性,统计该主键字段的每日的记录数据数,写进检查日志表中;根据主键的唯一性原则当记录数大于1判定错误写进检查结果表中;

[0065] 代码标准检查:当接入源数据进入数据仓库体系中时,约定了一些字段的含义和口径,当源数据系统进行了修改或者变更导致下游数据不准确;例如源数据处增加了一个字段来标示是A类型或者B类型,在抽取过程中没有取到下游,对此进行数据源列的检查,统计相关有效列数进行对比;

[0066] 业务规则检查:业务规则检查确保数据的准确性和易用性,通过制定规则来检查每条数据是否满足该规则,例如当年龄的值为0时认为这些数据是异常的,通过这些规则最后放入检查日志和检查结果表中;

[0067] 可以理解的是还包括其他检查:根据不同的业务场景制定特定的数据探查,业务把特定的规则录入质量体系中,开发人员根据规则生成相应的脚本去执行;例如在销售评分表中查看某销售成绩为空的人数占总人数的占比,当这个占比大于60%,业务就会判断这个数据是异常的,肯定是上游数据出现了异常,脚本执行完毕后预警系统根据检查结果去通知相关人员进行核对和修改。

[0068] 该步骤主要为了确保数据的完整性,判断当天数据是否为空表,是否影响下游数据的引用;例如在统计收房在职人员时,根据在职人员表中统计出收房人员当日的在职人数,存入检查日志表中;根据人力部门相关基础数据认为2000人左右为收房人员在职人数,最低人数不应该小于1500人,根据这个规则从检查日志表中获取当日收房人员数量,当这个数量为100人时判定为异常或者错误计入检查结果表。

[0069] S5:生成可视化预警报表,对第三数据进行可视化,突出显示异常的数据库表。

[0070] 对各业务数据进行分类展示,对异常指标增加警戒线的特别提醒,例如当统计元数据字段描述此指标时,认为当没有字段描述的量达到百分之三十时认为这里的数据可读性就很差,此时特别提醒,开发人员根据日志数据去完善这些数据指标。

[0071] 在一些可选的实施例中还包括在配置表中设置人员管理,将所述可视化预警报表中的异常的数据库表发送至所述人员。

[0072] 结合图2,图2是本发明提供的一种基于数据仓库及ETL的数据质量监控及预警系统结构框图。

[0073] 图2中,基于数据仓库及ETL的数据质量监控及预警系统包括源数据标准化处理模块201、数据仓库元数据管理模块202、ETL规则处理模块203、数据质量体系规则验证模块204、以及可视化预警报表生成模块205。

[0074] 源数据标准化处理模块201与数据仓库元数据管理模块202相耦接,用于将同步到

数据仓库中的源数据进行标准化处理得到标准化源数据，并将标准化源数据发送至数据仓库元数据管理模块202；

[0075] 根据各业务系统不同，在制定数据传输过程中加入预警判断；例如当爬虫系统异常时，根据业务人员制定的数据量标准，如低于某个最小值，系统根据条件再决定是否进行下一步；

[0076] 数据仓库元数据管理模块202分别与源数据标准化处理模块201和ETL规则处理模块203相耦接，用于对标准化源数据进行元数据管理，通过元数据将海量报表中的目标表、目标字段进行业务描述得到第一数据；

[0077] 在海量报表中查询某些表或者相关字段，由于开发人员没有维护好元数据，导致数据使用人员难以发现自己业务对应相关数据表，及相关字段的含义；对此在本预警系统中，对各模块相关元数据的统计可视化，例如把业务系统中各库通过元数据统计出哪些表、哪些字段是没有业务描述的，通过监控预警可视化促进开发和业务相关人员完善这些数据指标，使数据更易读易懂。

[0078] ETL规则处理模块203分别与数据仓库元数据管理模块202和数据质量体系规则验证模块204相耦接，用于根据业务规则和数据标准去除第一数据中的脏数据得到第二数据，脏数据是指不在给定的范围内或对于实际业务毫无意义、或是数据格式非法、或存在不规范的编码和含糊的数据；

[0079] 在数据仓库建设中ETL过程显得很重要，ETL过程中根据业务规则和数据标准首先去除某些脏数据；例如在收房明细数据中如果收房合同中没有房间编号，就可以在抽取到ODS层收房明细的过程中直接加入相关条件过滤掉此数据；也可以根据业务规则处理相关的数据，例如根据身份证号计算年龄时，当身份证号不是正确的格式时，假设身份证号前4位含有字母就视为垃圾数据，对此要加入数据处理，例如置为0等。

[0080] 数据质量体系规则验证模块204分别与ETL规则处理模块203和可视化预警报表生成模块205相耦接，用于根据数据质量体系规则对第二数据进行处理得到第三数据；

[0081] 在一些可选的实施例中，数据质量体系规则验证模块204用于根据数据质量体系规则对第二数据进行处理得到第三数据包括：

[0082] 获取第二数据的数据类型和/或属性；

[0083] 根据第二数据的数据类型和/或属性配置检测规则组合，其中配置检测规则组合至少包括一个检测规则；

[0084] 根据规则组合对第二数据进行质量检测得到第三数据发送至目的端。

[0085] 检测规则组合包括：主键检查、代码标准检查和业务规则检查。数据仓库为企业决策提供重要的数据支撑，数据质量的好坏，直接影响公司业务的发展；为了确保数据仓库的准确性通过主键检查、代码标准检查、业务规则检查以及其他检查来进行数据质量的监控。

[0086] 数据仓库为企业决策提供重要的数据支撑，数据质量的好坏，直接影响公司业务的发展；为了确保数据仓库的准确性通过主键检查、代码标准检查、业务规则检查以及其他检查来进行数据质量的监控；

[0087] 在一些可选的实施例中，检查规则包括：

[0088] 主键检查：确保主键的唯一性，统计该主键字段的每日的记录数据数，写进检查日志表中；根据主键的唯一性原则当记录数大于1判定错误写进检查结果表中；

[0089] 代码标准检查:当接入源数据进入数据仓库体系中时,约定了一些字段的含义和口径,当源数据系统进行了修改或者变更导致下游数据不准确;例如源数据处增加了一个字段来标示是A类型或者B类型,在抽取过程中没有取到下游,对此进行数据源列的检查,统计相关有效列数进行对比;

[0090] 业务规则检查:业务规则检查确保数据的准确性和易用性,通过制定规则来检查每条数据是否满足该规则,例如当年龄的值为0时认为这些数据是异常的,通过这些规则最后放入检查日志和检查结果表中;

[0091] 可以理解的是还包括其他检查:根据不同的业务场景制定特定的数据探查,业务把特定的规则录入质量体系中,开发人员根据规则生成相应的脚本去执行;例如在销售评分表中查看某销售成绩为空的人数占总人数的占比,当这个占比大于60%,业务就会判断这个数据是异常的,肯定是上游数据出现了异常,脚本执行完毕后预警系统根据检查结果去通知相关人员进行核对和修改。

[0092] 可视化预警报表生成模块205与数据质量体系规则验证模块204相耦接,对第三数据进行可视化,突出显示异常的数据库表。

[0093] 开发人员和相关业务人员制定相关的数据质量的管理规则,录入系统中,开发人员根据这些规则转换为脚本执行,把检查记录放入检查日志表中,根据检查日志表和数据质量体系的规则处理形成检查结果表,为可视化表系统和error通知做数据支撑;在规则体系中数据量检查主要为了确保数据的完整性,判断是不是当天数据为空表,是否影响下游数据的引用;

[0094] 在一些可选的实施例中还包括可视化预警报表发送模块,用于在配置表中设置人员管理,将可视化预警报表中的异常的数据库表发送至人员。

[0095] 对各业务数据进行分类展示,对异常指标增加警戒线的特别提醒,比如当统计元数据字段描述此指标时,认为当没有字段描述的量达到百分之三十时认为这里的数据可读性就很差,此时就特别提醒,开发人员根据日志数据去完善这些数据指标;然后在error通知系统中把相关业务表的开发负责人配置为该表的一个属性,当此表数据有需要报错的时候直接发送负责人,方便负责人能快速处理,查找相关问题。

[0096] 通过上述实施例可知,本发明提供的基于数据仓库及ETL的数据质量监控及预警方法和系统,至少实现了如下的有益效果:

[0097] 因为本发明采用对源数据先后经过源数据标准化处理、数据仓库元数据管理、ETL规则处理以及数据质量体系规则验证,可以提高检测质量,挺高检测效果以及检测精度,使管理者方便对目前数据资产和数据仓库对相关质量的判定,能够指导管理者和开发者对改善数据质量提供更明细的数据问题定位;能够使数据易懂易读,便于数据使用者;

[0098] 本发明可以生成可视化预警报表,突出显示异常的数据库表,方便对数据进行监控。

[0099] 虽然已经通过例子对本发明的一些特定实施例进行了详细说明,但是本领域的技术人员应该理解,以上例子仅是为了进行说明,而不是为了限制本发明的范围。本领域的技术人员应该理解,可在不脱离本发明的范围和精神的情况下,对以上实施例进行修改。本发明的范围由所附权利要求来限定。

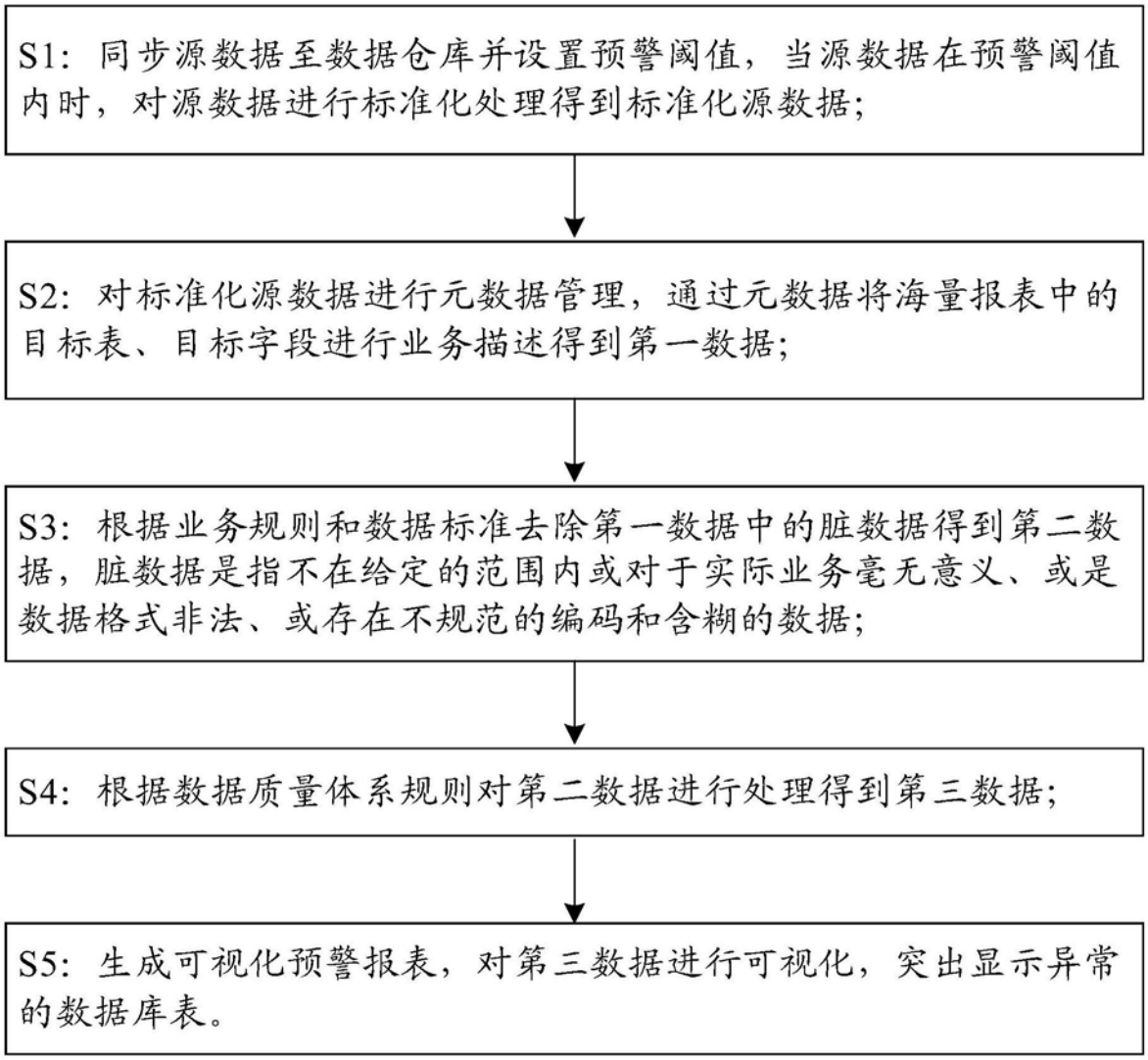

S1：同步源数据至数据仓库并设置预警阈值，当源数据在预警阈值内时，对源数据进行标准化处理得到标准化源数据；
S2：对标准化源数据进行元数据管理，通过元数据将海量报表中的目标表、目标字段进行业务描述得到第一数据；
S3：根据业务规则和数据标准去除第一数据中的脏数据得到第二数据，脏数据是指不在给定的范围内或对于实际业务毫无意义、或是数据格式非法、或存在不规范的编码和含糊的数据；
S4：根据数据质量体系规则对第二数据进行处理得到第三数据；
S5：生成可视化预警报表，对第三数据进行可视化，突出显示异常的数据库表。

图1

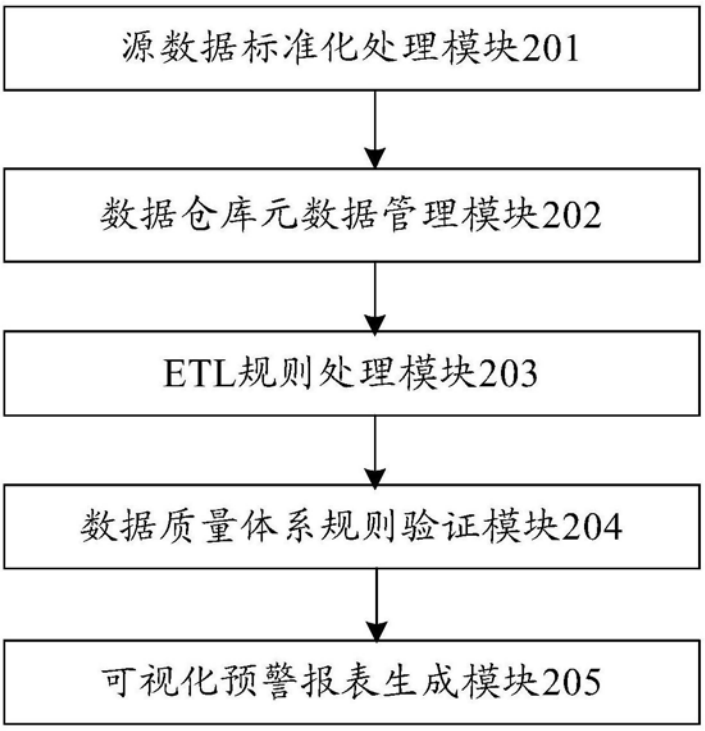

源数据标准化处理模块201
数据仓库元数据管理模块202
ETL规则处理模块203
数据质量体系规则验证模块204
可视化预警报表生成模块205

图2