数据ETL研究综述

徐俊刚 裴 莹
(中国科学院研究生院信息科学与工程学院 北京 100190)

摘 要 数据抽取、转换和装载(Extraction,Transformation and Loading,简称ETL)是数据仓库化的关键环节,对数据仓库数据质量有着至关重要的影响。随着信息化的发展,ETL已经成为当前较活跃的研究领域之一,但是ETL理论和技术的发展还不成熟。针对当前ETL研究中存在的一些问题和需要考虑的各种因素,从ETL各个阶段存在的主要问题出发,列举了各种研究方法及研究成果,并进行了分析。最后,总结并提出了ETL的未来研究方向和今后工作的建议。

关键词 ETL,数据仓库,数据质量,元数据
中图法分类号 TP391 **文献标识码** A

Overview of Data Extraction, Transformation and Loading

XU Jun-gang PEI Ying
(School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract Data extraction, transformation and loading are crucial steps of data warehousing, which influences data quality of data warehouse intensively. With the development of informationization, ETL has already become one of most popular research fields, but till now, ETL theory and technology are still not mature. As to the problems and factors appeared in ETL research, many research methods and achievements were listed according to the main problems existed in each ETL phase. Finally, several future research trends of ETL and some proposals for the future research work were summarized and presented respectively.

Keywords ETL, Data warehouse, Data quality, Metadata

1 引言

信息资源是现代企业的宝贵资源,是企业进行科学管理和决策分析的基础。随着商业智能技术(数据仓库、联机分析处理(Online Analytical Processing,OLAP)和数据挖掘等)的发展,许多企业需要通过对联机事务处理(Online Transaction Processing,OLTP)业务系统和办公自动化系统(Office Automation,OA)记录的大量业务数据进行加工和分析,找出有价值的信息,以支持企业的经营决策。数据仓库的出现使这一任务成为可能,越来越多的企业立足于多年积累的数据和自身的核心业务,构建了企业级数据仓库系统。

ETL的概念是随着数据仓库的产生而产生的。建设企业数据仓库最关键的工作就是将各业务系统中的数据按主题进行重新集成,而大多数企业的业务系统都存在平台不同、数据源异构等问题,这使得数据集成非常复杂,给数据仓库的建设带来了很大困难。为解决这类问题,人们提出了ETL的概念,它逐渐形成一个较为独立的数据集成模式。由于数据清洗也是建设数据仓库的重要步骤,并且通常在ETL过程中进行,因此研究人员将数据清洗也加入到了ETL研究中来。随着数据仓库、数据挖掘以及其他信息系统建设等对数据集成的高度要求,数据ETL逐渐成为当前信息技术中较活跃的研究领域之一。

2 ETL的概念和功能

ETL是数据抽取、转换和装载(Extract,Transformation,Loading)的英文简称,是数据仓库获取高质量数据的关键环节,是对分散在各业务系统中的现有数据进行提取、转换、清洗和加载的过程,使这些数据成为商业智能系统需要的有用数据。ETL是构建数据仓库的第一步,也是构建数据仓库最重要的步骤。

目前,对于ETL尚未有统一的定义。Vassiliadis P.认为:"ETL工具是一种专门化的工具,它的任务是处理数据仓库的同构性、数据清洗以及装载的问题"[1]。Simitsis A.将ETL工具定义为"一组负责从多个不同种类和形式的数据源中抽取数据,对数据进行清洗、客户化,进而将其装入到数据仓库中的软件"[2]。这两种定义都从功能的角度对ETL进行了解释和说明。

在ETL过程中,抽取可看作是数据的输入过程,主要解决数据源的异构问题,即从多个数据源中将数据抽取到统一的数据存储中;而数据装载可看作是数据的输出过程,即将处理后的数据从统一的数据存储装载到目标数据仓库中。二者中间的转换和清洗则主要解决数据质量问题,它通过一系列

到稿日期:2010-05-25 返修日期:2010-11-13 本文受国家863计划课题"数据中心的运行时功耗管理技术"(2009AA01Z139)资助。
徐俊刚(1972—),男,博士,副教授,主要研究方向为商业智能,E-mail:xujg@gucas.ac.cn;裴 莹(1986—),女,硕士生,主要研究方向为商业智能、数据管理。

的清洗过程将海量数据中存在的冗余、数据错误、数据缺失等检测出来并加以改正，并使用默认的或者用户定义的转换规则对数据中的某些字段进行合并、转换等操作，使得数据具有良好的正确性、一致性、完整性和可用性[3]。

根据"Garbage in, garbage out"原则，只有在 ETL 过程中对数据进行了有效的处理，才能保证数据仓库中的数据质量，从而更好地支持 OLAP 以及数据挖掘。同时，根据调查，在企业数据仓库的建设过程中，有 60%的精力花费在数据 ETL 的设计和实施上[4]。Shilakes C. 指出，企业在 ETL 工具的采购中投入了整个数据仓库建设三分之一以上的成本预算[5]；Demarest M. 通过调查表明，ETL 实施占了整个数据仓库建设的 80%以上的时间[6]。由此可见，ETL 具有非常重要的研究价值和应用前景。

3 ETL 理论研究

3.1 ETL 建模

ETL 建模研究主要集中在概念和逻辑建模以及模型间的转换和优化等方面。

典型的 ETL 过程的概念建模是将数据源端的属性(Attributes)映像到数据仓库中各个表的属性字段中，以便在数据仓库建立的前期方便地追踪关系内的属性以及 ETL 的各个活动，并提供可定制、可扩展的建模方式，以便设计者进行 ETL 活动的概念建模。Vassiliadis P. 研究了 ETL 的初始装入情景、ETL 活动的定义以及概念表示规范[1]，Simitsis A. 在此基础上完善了 ETL 概念建模方法，并提供了一种对已有数据的结构和内容分析进行追踪的建模方法[7]。

ETL 的逻辑建模描述了从数据源端流向数据仓库的数据流，它对应于 OHM(Operator Hub Model)。OHM 是 Orchid 提供的一种抽象操作符模型及转换符机制，可以获取对于模式转换和 ETL 过程都通用的语义转换。Skoutas D. 使用语义网技术以及本体论的方法对 ETL 过程进行建模[8]。通过建立一个本体来表示一系列相关的数据源中相似的定义，从而精确而规范地定义了数据存储模式中的语义以及 ETL 转换。为了方便使用，作者提出了一系列 ETL 过程中常见的操作符以简化对设计者使用 Sellis T. 提出的模板库[9]进行 ETL 概念转换的识别工作，从而提高了 ETL 设计的自动化程度。

Simitsis A. 等人还对如何从概念建模向逻辑建模进行转化进行了研究并提出了半自动化的转换方法，但由于转换前并未对概念模型的正确性进行验证，因此无法保证逻辑模型的正确性，仅为 ETL 设计者提供了一套通用的步骤来完成 ETL 的逻辑设计。同时，Simitsis A. 将逻辑建模以及优化问题作为状态空间搜索问题，在考虑执行成本的前提下提出了如何最小化 ETL 逻辑执行时间的优化算法[2]。

3.2 ETL 过程研究

按照抽取时间的不同，数据 ETL 过程可以被划分为两种类型：全量 ETL 过程和增量 ETL 过程。全量 ETL 过程一般用于数据仓库的初始化，而增量 ETL 过程则用于数据仓库的增量维护。全量 ETL 过程包括：数据抽取、转换、清洗和装载 4 部分。增量 ETL 中除包含全量 ETL 中的 4 个部分外，还存在其特有的问题和方法。

3.2.1 数据抽取

数据抽取是从不同的网络、不同的操作系统、不同的数据库及数据格式、不同的应用中抽取数据的过程。此处的数据不仅指关系数据库系统中的数据，还涉及到半结构化的数据(如网页包含的数据等)和非结构化(如文本文件等)的数据。

对于半结构化数据，Doorenbos R. 等人采用了机器学习的技术对网页内容进行抽取[10]，但它对格式的要求比较严格；Xiaoying Gao 等人提出混合表示法对数据及数据模式进行建模[11]，它包括一个概念层次图和一套知识框架，使用基于内容以及结构框架的方法对数据进行抽取。Ling Liu 等人使用交互式的元数据知识模型建模并用引导学习的方法对 XML 文件中的数据进行说明和抽取[12]。对于非结构化数据，文献[13]使用 3 种线型模型来表示表格，并采用模糊匹配方法识别表格中直线行上的字段[13]。针对手写汉字文件的数据，Jiun-Lin Chen 等抽取提出了基于引力(gravitation)的算法，以有效识别并抽取表格中的汉字[14]。

异构数据源集成问题也是 ETL 所面临的主要挑战之一。异构数据源集成是数据集成的一种，主要处理多数据源的异构问题。多数据源异构主要表现为字段名称和类型的不一致，数据所在的定义域不同，等等。目前主要采取统一元数据来进行异构数据的统一化管理[15,16]。文献[17]从本体论的角度出发，对如何使用 CWM(Common Warehouse Metadata)[18]来对异构数据进行建模做了详细说明。在对异构数据源进行抽取时，文献[19]采用了一种基于统计方法的"空间向量积"的转换方式(称为 WHIRL)来对字段名称等进行映射，作者针对 Web 上多数据源的抽取进行了实验，并对抽取后的数据进行短查询(short queries)，证实了运用此方法映射后的准确性[19]。文献[20]开发了面向数据集成的 ETL 过程模型 DataIntegrator，以任务、通道和数据单元作为 ETL 过程的核心实现了对 ETL 数据的集成、增量复制以及任务调度[20]。

3.2.2 数据转换

数据转换就是处理抽取上来的数据中存在的不一致的过程。从定义上来说，数据转换是对数据的转化(数据的合并、汇总、过滤、转换等)、数据的重新格式化和计算、关键数据的重新构建和数据汇总、数据定位的过程。

数据转换一般包括两类：一类是数据名称及格式的统一，即数据粒度转换、商务规则计算以及统一的命名、数据格式、计量单位等；另一类，数据仓库中存在源数据库中可能不存在的数据，因此需要进行字段的组合、分割或计算[23]。

针对第一类问题，Halevy A. 等人提出了"信息复写"的方法，用以处理数据转换和集成问题，通过对数据源的描述将信息"复写"成目标格式，保证了转换过程的正确性[22]。Squire C. 等人提出了数据转换自动化的问题，作者认为数据集成主要在于数据抽取和转换的自动化程度，提出了通过维护源数据文件相关的元数据来进行数据映射和转换的方法[23]。针对第二类问题，主要通过业务逻辑分析，制订并使用 ETL 转换函数来实现[24,25]。

3.2.3 数据清洗

对于创建数据仓库及其后续工作，如 OLAP、数据挖掘等，都需要数据具有良好的正确性和可用性，而现存操作型系统中的数据尚存在很多问题，容易造成"脏数据"。因此，有必要对即将进入数据仓库的数据进行全面检查及改正，使它们尽可能无差错。这一过程就称作数据清洗[3]。

Rahm E. 将数据质量问题分为 4 类：单数据源模式层问题、单数据源实例层问题、多数据源模式层问题和多数据源实

例层问题[26]。可以通过改进模式设计、模式转化和模式集成来解决模式层次上的问题(如缺少完整性约束等),而实例层的问题(如数据拼写错误、值无效、相似重复记录等)在模式层不可见,必须通过机器学习、匹配算法并与外部的查找表结合等方法来解决。

目前对于数据清洗的研究主要集中在实例层算法方面。已有的研究中已经提出了很多清洗算法,如采用数据库管理系统的集成数据清理算法[27]、增量数据清洗算法[28]等。近年来,相似重复数据检测与消除算法[29,30]、基于规则的清洗算法[31]、基于领域知识的清洗算法[32]等智能化清洗算法和方案越来越受到研究者的关注。Borkar V. 讨论了地址数据的清洗问题,将地址字段分为不同的元素,并且使用训练隐马尔可夫模型(Hidden Markov Models, HMM)来发现脏数据[33];Hernandez M. A. 提出在不同的键上多次排序,并分别计算邻近记录的相似度,最后综合多次计算的结果完成记录匹配过程[34]。Qiu Yue-feng 等人提出了一种高效的基于 N-Gram 的聚类算法,以较好地聚类相似重复记录[35]。

数据清洗与其他领域关系密切,Lee M. L. 等人将领域知识加入到规则集中,与从样本中抽取的规则相结合,使用机器学习和统计的方法进行目标数据集的匹配[32]。

由于数据清洗有很强的领域相关性,因此过去一段时间内通用的清洗框架很少有人关注。然而,随着数据 ETL 的广泛应用,通用的集成清洗方案越来越受到研究人员的重视。Raman V. 等人实现的 Potter's Wheel 数据清洗框架可向用户提供强大的交互性[36],Galhardas H. 等人提出了一种描述性语言,用以指定一些数据转化操作的参数(如记录匹配操作所使用的距离函数等),同时实现了名为 AJAX 的 ETL 框架,以较好地分离逻辑规范层和物理实现,得到结构化的通用解决方案[37,38]。鲍玉斌等人以用户为中心建立了数据清洗过程模型,详细描述了模型中各工具箱(如转换、集成、验证和统计分析等)的功能定义,以及其中的(元)数据流、处理流和工具流[39]。

3.2.4 数据装载

数据装载的主要任务是将经过清洗后的干净的数据集按照物理数据模型定义的表结构装入目标数据仓库的数据表中,并允许人工干预,以及提供强大的错误报告、系统日志、数据备份与恢复功能。整个操作过程往往要跨网络、跨操作平台。

目前为止,关于数据仓库中 ETL 的研究主要集中在抽取、转换和清洗方面,对于数据装载的研究相对较少。Inmon W. H. 将 ETL 中的装载问题分为 3 类:第一类,若目标数据仓库处于同一系统中,可以将数据及相关元数据直接存入;第二类,存在 Staging Area(ETL 过程中的临时存储,一般为磁盘)中的数据,可通过异构系统的接口载入;第三类,数据仓库中已有数据的更新,可看成增量 ETL 数据的装载,在通过元数据定义的数据规则和格式检查之后更新对应数据仓库内的数据,同时将原有数据保存[40]。整个装载过程必须加时间戳。

装载中的主要问题是大数据量的装载以及与数据抽取时相似的异构数据集成的挑战。Fenk R. 等提出了使用 UB 树来装载大数据块(Bulk)的算法,它针对全量和增量过程分别给出了初始化 UB 树以及在 UB 树追加数据的方法,并考虑了系统 I/O 和 CPU 成本,具有很强的可用性[41]。

一般情况下,数据仓库的装载是以确定的周期进行的,影响了实时决策的精确性。因此,Ricardo J. S. 等提出了实时数据仓库的装载概念,将数据仓库内的信息变化定义为实时发生的,通过表结构复制、查询断言约束等方法使得数据仓库的装载最小化地影响查询响应,以提供实时的决策分析,并通过第三方标准 TPC-H 对实时数据仓库的性能进行了测试,证明了此方法的可用性[42]。

3.2.5 增量 ETL 过程

增量 ETL 用于对数据仓库的维护,主要是对数据仓库中的数据进行插入、更新和删除。相对全量 ETL 方式而言,增量 ETL 方式设计更加复杂,但从效率和性能方面来说,增量 ETL 方式比全量 ETL 方式更适合于数据仓库的日常维护。

增量 ETL 首先要获取增量数据,主要有 3 种方式:(1)基于数据库的日志;(2)基于数据库中的时间戳;(3)使用快照技术。许多现成的商业工具(例如 Oracle)提供了 CDC (Changed Data Capture)机制,首先通过比较时间戳来获取增量数据,然后通过 CDA(Capture Data Application)来将处理后的增量数据装载到数据仓库中。

张旭峰等人提出"数据仓库可以看成是数据源的物化视图"的观点,将全量 ETL 过程看成视图定义,而增量 ETL 过程看成是物化视图的增量维护,由全量 ETL 过程推导得到[28]。作者使用文献[43]中提出的方法,首先实现单个片段上的增量维护,然后采用了 MCCI(Minimal Cost of Calculating Increment)算法,从全量 ETL 推导出用于实现增量 ETL 过程的 SQL 语句的集合。MCCI 方法由于忽略了存储空间上的限制,在 ODS(Operational Data Store)等容量有限的存储媒介中运行时可能会对性能产生较大的影响。

Thomas J. 反对关于数据仓库环境下视图维护的观点[44],他认为,物化视图表示的是当前状态,而数据仓库并不是实时更新的。Thomas J. 提出了类似于关系代数学的方法,使用 OHM 操作符并在此基础上进行改进。由于许多原有系统的建设是基于 OHM 的,因此这个方法可以使原有系统无需做太大的改变即可重用。

3.3 数据质量与元数据

3.3.1 数据质量

数据仓库建立的目的是为数据分析提供准确性、一致性、完整性、有效性的数据,来辅助企业领导决策,而"脏数据"会给以后的数据挖掘、决策支持过程带来错误,导致不可靠的输出。ETL 过程是数据仓库中数据的入口,因此要保证数据仓库中数据的质量,除了尽量保证业务系统中基础数据的正确性以外,很大程度上取决于 ETL 过程对数据的处理。

要判断什么是高质量的数据,必须有一套完备的评价标准。Chapman D. 指出,数据质量的判定指标与其具体应用领域相关[45]。而在 ETL 领域内,Kimball R. 与 Caserta J. 提出的将数据的正确性、明确性、一致性、完备性、可用性和时限性这 6 个重要指标作为必须给予足够关注的数据质量指标被大多数研究者接受[46]。

3.3.2 元数据

元数据是描述数据的数据。在数据仓库中,元数据主要是对业务数据本身及其运行环境的描述与定义的数据,例如对数据库、表、列、列属性(类型、格式和约束等)等的描述。对于 ETL 过程来说,元数据的重要意义集中表现为:(1)定义数据源的位置及数据源的属性;(2)确定从源数据到目标数据的

对应规则;(3)确定相关的业务逻辑;(4)在数据实际加载前的其它必要的准备工作。元数据一般贯穿整个数据仓库项目,ETL 的所有过程必须最大化地参照元数据[47]。

合理的元数据能够有效地描绘出信息的关联性,将它与数据质量结合起来,可以更加有效地指导 ETL 过程。Gomes P. 等人提出了一个基于 CWM 的数据质量元模型[48],此模型及其规则可以良好地支持数据仓库中元数据的管理以及数据转换、清洗的要求,并可以针对具体清洗算法的要求进行参数化的扩展。

4 典型 ETL 工具

4.1 商业 ETL 工具

随着 20 世纪末以来商业智能的广泛应用以及 ETL 软件市场的持续稳定发展,很多厂商致力于商业 ETL 工具的开发。当前国内外 ETL 商业工具主要有:Oracle 公司的 Oracle Warehouse Builder(OWB)、微软公司的 Data Transformation Services(DTS)、Informatica 公司的 Informatica、SAS 公司的 Enterprise ETL Server、IBM 公司的 Data Stage、iWay Software 的 DataMigrator、DataMirror 的 Transformation Server 等。

4.2 开源 ETL 工具

目前较常用的开源 ETL 工具有 Kettle, Talend, CloverETL, Octopus 等。Kettle 是当前最热门的开源 ETL 工具,它是 Pentaho 组织使用元数据驱动的方法进行设计和开发而成的;Talend 基于 Eclipse 平台,提供全功能的 Data Integration 解决方案,可以实现商业流程建模、数据流程建模等功能;CloverETL 提供了一组 API,用 XML 来定义 ETL 过程,但它同时提供了一个不开源的 CloverGUI 来进行图形化的 ETL 开发(需要购买商业许可证);Octopus 是 Enhydra 组织的 ETL 工具,它支持任何 JDBC 数据源,用 XML 定义,也支持 JDBC-ODBC, XML, EXCEL 等。

4.3 学术 ETL 工具

ETL 的学术研究成果中比较著名的是由法国 INRIA 开发的 AJAX 系统、Berkeley 开发的 Potter's wheel 以及 Vassiliadis P. 等人的 Arktos 原型系统。AJAX 主要面向数据清洗,用来处理典型的数据质量问题[37],例如:对象同一性问题、拼写错误,以及记录之间数据矛盾的问题。Potter's wheel 系统[36]可以向用户提供交互式的数据清洗过程。这两种原型都基于代数学(algebras),尤其适用于数据是均匀的网络数据的情况;Arktos[49]提供一个用于 ETL 流程的标准元模型,并支持以客户化的可扩展的方式对 ETL 过程进行建模。

5 理论和技术的研究方向

通过总结和分析,ETL 的发展方向及研究重点主要有以下几点:

(1)ETL 建模

不管是在商业 ETL 工具还是在开源或学术 ETL 工具中,现有的 ETL 模型都使用临时的脚本语言或者可视化用户接口,没有形成一种统一的 ETL 语言,仅有 Orchid 的 OHM (Operator Hub Model)提供了一种平台无关的建模方法,用于获取各种 ETL 工具的数据转换。Vassiliadis P. 等人提出了概念和逻辑建模的方法,但其尚未得到广泛的应用。因此,如何提供一种统一的 ETL 语言来规范各厂商对 ETL 流程以及相关元数据的设计,还需要进一步探索。

(2)ETL 过程的研究

在数据抽取、转换、清洗和装载 4 个步骤中,都有非常值得深入研究的问题。对于数据抽取、异构数据源集成问题一直是数据 ETL 所面临的主要困难。随着信息系统的发展,ETL 工具应当支持尽可能多的 DBMS、文件系统以及数据采集、处理系统,能够跨网络、跨平台使用,这些都给数据抽取带来了很大的挑战。数据清洗过程中,丢失值填充与相似重复记录处理,是实例层次上基于语义的数据集成,也是当前数据清洗领域研究最活跃的分支。另外,如何建立通用的可交互的数据清洗框架,提高清洗的自动化程度,使其既支持与领域专家的有效交互,同时又能减少开发人员数据清洗过程的手工作业量,也是非常值得探索的课题。数据装载与数据抽取类似,其跨平台、跨网络等特性使异构问题成为其最主要的瓶颈。

在 ETL 过程中,错误的转换、清洗以及装载过程等都可能带来错误数据。因此,要求 ETL 工具需要强大的错误恢复功能。此时,必须通过数据 ETL 过程的辅助管理工具,如日志管理、ETL 工作调度管理等对错误的操作进行恢复。目前各大厂商开发的 ETL 工具,虽然都提供了错误恢复的功能,但很不成熟。因此,这一领域非常值得研究者关注。

(3)增量 ETL 过程中数据准备区的应用

通常,包含数据准备区(Data Staging Area)的 ETL 系统可以将抽取后的数据统一放在数据准备区中,然后进行转换、清洗和装载。但在增量 ETL 过程中由于全量 ETL 数据已经被装载到目标数据仓库,给增量的数据转换和清洗带来了很大的困难,现有的 ETL 工具对于增量 ETL 的处理都无法提供非常简便并且实用的方法。因此研究人员可以从数据准备区入手,使用数据准备区作为临时存储,将部分变化的数据在其变化的时候被记录在此处,减少一部分 CDC(Changed Data Capture)的工作,从而提高增量 ETL 的效率和准确率。但是,使用数据准备区会增加 ETL 过程中的 I/O 成本,因此需要研究在哪种情况下可以将其引入。

(4)ETL 性能和智能化的研究

ETL 工具需要处理海量数据,因此效率极为重要。当前的 ETL 工作调度和优化尚未成熟,如何自动地分配和调度 ETL 工作流程是今后研究的又一重要课题。同时,随着数据仓库 ETL 的广泛应用,未来的 ETL 应当具备高度的可伸缩性,不仅能运行在昂贵的主机系统上,还能应用到工作站或 PC 机上,在保证 ETL 工作效率的基础上为企业减少硬件的投入,同时保障 ETL 操作的可靠性和稳定性。另外,智能化也是 ETL 发展的重要方向。ETL 需要与领域专家交互,但是随着专家系统、机器学习、人工神经网络等领域成果的出现,如何应用这些成果,使得数据源管理、ETL 规则定制、数据质量保证等工作更加智能化、自动化地完成也是今后的重点研究方向。

(5)元数据在 ETL 过程中的应用

元数据是 ETL 过程的"指挥中心",元数据的选取、规范以及管理都直接影响到 ETL 过程的正确性和效率。随着元数据标准 CWM 的出台,如何将标准化的元数据管理引入到 ETL 过程中,并指导数据仓库的 ETL 过程将成为整个数据 ETL 过程研究的先导。

6 今后研究工作的建议

对于 ETL 的研究工作有以下建议，供参考：

(1)理论与实践紧密结合

从目前的研究工作来看，在 ETL 理论和实际应用之间存在鸿沟，为了更好地在企业数据仓库建设中应用，需要计算机专家与行业专家紧密合作，研究并开发面向行业的更加实用的 ETL 工具。

(2)规范 ETL 框架是当务之急

正如软件工程中的需求分析一样，对 ETL 建立完整可行的框架是非常重要的。如果这方面的工作没有做好，会影响 ETL 的各个环节。而当前 ETL 还没有统一的规范，因此这是一项非常有意义的研究工作。

(3)加强 ETL 中数据清洗的研究

数据清洗可以保证数据的高质量，这对数据仓库以及后续的数据挖掘、决策分析的正确性有至关重要的影响，当前 ETL 中的数据清洗尚未形成完善的清洗流程，应该与其他研究领域(如数据治理、知识发现、神经网络、专家系统)相结合，研究高效的数据清洗方法。

(4)加强 ETL 自动化的研究

ETL 涉及海量数据，应尽可能使其自动化地完成数据集成工作，减少或简化人工干预，以达到降低系统风险、实现良好数据集成的效果。

参 考 文 献

[1] Vassiliadis P, Simitsis A, Skiadopoulos S. Conceptual Modeling for ETL Processes [C]// Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP. New York: ACM, 2002: 14-21

[2] Simitsis A. Mapping Conceptual to Logical Models for ETL Processes [C]// Proceedings of the 8th ACM International Workshop on Data Warehousing and OLAP. New York: ACM, 2005: 67-76

[3] 郭志懋，周傲英. 数据质量和数据清洗研究综述[J]. 软件学报，2002，13(11)：2076-2083

[4] Inmon W H. The Data Warehouse Budget [J/OL]. DM Review Magazine. http://www.datawarehouse.inf.br/Papers/inmon%20budget-1.pdf, 2010-4-12

[5] Shilakes C, Tylman J. Enterprise Information Portals [R]. New York: Merrill Lynch, 1998

[6] Demarest M. The Politics of Data Warehousing [EB/OL]. http://www.hevanet.com/demarest/marc/dwpol.html, 2009-6-12

[7] Simitsis A, Vassiliadis P. A Methodology for the Conceptual Modeling of ETL Processes [C]// Proceedings of the Decision Systems Engineering Workshop. Klagenfurt: CAiSE, 2003: 501-505

[8] Skoutas D. Designing ETL Processes Using Semantic Web Technologies [C]// Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP. New York: ACM, 2006: 67-74

[9] Sellis T. Formal Specification and Optimization of ETL Scenarios [C]// Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP. New York: ACM, 2006: 1-2

[10] Doorenbos R, Etzioni O, Weld D. A Scalable Comparison-shopping Agent for the World Wide Web [C]// Proceedings of the First International Conference on Autonomous Agents. New York: ACM, 1997: 39-48

[11] Gao Xiao-ying, Leon S. Semi-structured Data-extraction from Heterogeneous Sources [C]// Proceedings of Internet-based Organizational Memory and Knowledge Management. Hershey, PA: IGI Publishing, 1999: 83-102

[12] Liu Ling, Calton P, Han Wei. An XML-enabled Data Extraction Toolkit for Web Sources [J]. Information Systems, 2001, 26(9): 563-583

[13] Tseng L Y, Chen R C. Recognition and Data Extraction of Form Documents Based on Three Types of Line Segments [J]. Pattern Recognition, 1998, 31(10): 1525-1540

[14] Chen Jiunlin, Lee Hsijian. Field Data Extraction for Form Document Processing using a Gravitation-based Algorithm [J]. Pattern Recognition, 2001, 34(9): 1741-1750

[15] Calvanese D, Giacomo G D, Lenzerini M, et al. A Principled Approach to Data Integration and Reconciliation in Data Warehousing [C]// Proceedings of the International Workshop on Design and Management of Data Warehouses. Picataway, NJ: DMDW, 1999: 16

[16] Sheth A P. Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics [A]// Interoperating Geographic Information Systems [C]. Norwell, MA: Kluwer Academic Publishers, 1998: 5-30

[17] Hoang A D T. An Integrated Use of CWM and Ontological Modeling Approaches towards ETL Processes [C]// Proceedings of the 2008 IEEE International Conference on e-Business Engineering. Picataway, NJ: IEEE, 2008: 715-720

[18] *Common Warehouse Metamodel (CWM) Specification (Version 1.0)*[S]. Needham, MASA: OMG, 2001

[19] William W C. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity [C]// Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1998: 201-212

[20] 钟华，冯文澜，谭红星，等. 面向数据集成的 ETL 系统设计与实现[J]. 计算机科学，2004，31(9)：87-89

[21] Orli R J, Sartos F. Data Extraction, Transformation, and Migration Tools [EB/OL]. http://www.kismeta.com/exg.html, 2010-04-28

[22] Halevy A, Rajaraman A, Ordille J. Data Integration: The Teenage Years [C]// Proceedings of the 32nd International Conference on Very Large Data Bases. New York: ACM, 2006: 9-16

[23] Squire C. Data Extraction and Transformation for the Data Warehouse Solutions [C]// Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1995: 446-447

[24] Vassiliadisl P, Simitsis A, Georgantas P, et al. A Framework for the Design of ETL Scenarios [C]// Proceedings of Conference on Advanced Information Systems Engineering (CAiSE). Klagenfurt: CAiSE, 2003: 520-535

[25] *Tziovara V, Vassiliadis P, Simitsis A. Deciding the Physical Implementation of ETL Workflows* [C]// Proceedings of the ACM 10th International Workshop on Data Warehousing and OLAP. New York: ACM, 2007: 49-56

[26] Rahm E. Data Cleaning: Problems and Current Approaches [J]. IEEE Data Engineering Bulletin, 2000, 23(4): 3-13

[27] Hernandez M A, Stolfo S J. The Merge/Purge Problem for

Large Databases [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM, 1995: 127-138

[28] Zhang Xufeng, Sun Weiwei, Wang Wei, et al. Generating Incremental ETL Processes Automatically [C]// Proceedings of the First International Multi-symposiums on Computer and Computational Sciences. Picataway, NJ: IEEE, 2006: 516-521

[29] Monge A E, Elkan C. The Field Matching Problem: Algorithms and Applications [C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1996: 267-270

[30] Monge A E. Matching Algorithm within a Duplicate Detection System [J]. IEEE Data Engineering Bulletin, 2000, 23(4): 14-20

[31] Marcus A, Maletic J I, Lin K L. Ordinal Association Rules for Error Identification in Data Sets [C]// Proceedings of the 10th International Conference on Information and Knowledge Management. New York: ACM, 2001: 589-591

[32] Lee M L, Ling T W, Low W L. IntelliClean: a Knowledge-based Intelligent Data Cleaner [C]// Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2000: 290-294

[33] Borkar V, Deshmuck K, Sarawagi S. Automatically Extracting Structure from Free Text Addresses [J]. Bulletin of the Technical Committee on Data Engineering, 2000, 23(4): 27-32

[34] Hernandez M A, Stolfo S J. Real-world Data is Dirty: Data Cleansing and the Merge/Purge Problem [J]. Data Mining and Knowledge Discovery, 1998, 2(1): 9-37

[35] Qiu Yuefeng, Tian Zengping, Ji Wenyun, et al. An Efficient Approach for Detecting Approximately Duplicate Database Records [J]. Chinese Journal of Computers, 2001, 24(1): 69-77

[36] Raman V, Hellerstein M. Potter's Wheel: An Interactive Framework for Data Cleaning and Transformation [C]// Proceedings of the 27th Conference on Very Large Data Bases. New York: ACM, 2001: 89-92

[37] Galhardas H, Florescu D, Shasha D, et al. AJAX: an Extensible Data Cleaning Tool [C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2000: 590

[38] Galhardas H, Florescu D, Shasha D, et al. Declarative Data Cleaning: Language, Model and Algorithms [C]// Proceedings of the 27th International Conference on Very Large Data Bases. New York: ACM, 2001: 371-380

[39] 鲍玉斌, 孙焕良, 冷芳玲, 等. 数据仓库环境下以用户为中心的数据清洗过程模型[J]. 计算机科学, 2004, 31(5): 52-55

[40] Inmon W H, Conklin E. Loading Data into the Warehouse [J]. Tech Topic, 1994, 11(1): 20-25

[41] Fenk R, Kawakami A, Markl V, et al. Bulk Loading a Data Warehouse Built upon a UB-Tree [C]// Proceedings of the 2000 International Symposium on Database Engineering & Applications. Picataway, NJ: IEEE, 2000: 179-187

[42] Ricardo J S, Jorge B. Real-time Data Warehouse Loading Methodology [C]// Proceedings of the 2008 International Symposium on Database Engineering & Applications. New York: ACM, 2008: 49-58

[43] Cui Yingwei, Widom J, Wiener J L. Tracing the Lineage of View Data in a Warehousing Environment [J]. Database Systems, 2000, 25(2): 179-227

[44] Thomas J. Towards Generating ETL Processes for Incremental Loading [C]// Proceedings of the 2008 International Symposium on Database Engineering & Applications. New York: ACM, 2008: 101-110

[45] Chapman A D. Principles of Data Quality (version 1.0) [R]. Copenhagen: Global Biodiversity Information Facility, 2005

[46] Kimball R, Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data [M]. New York: John Wiley & Sons, 2004: 11-23

[47] Wang R Y, Storey V, Firth C P. A Framework for Analysis of Data Quality Research [J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(4): 623-640

[48] Gomes P, Farinha J, Trigueiros M J. A Data Quality Metamodel Extension to CWM [C]// Proceedings of the 4th Asia-Pacific Conference on Conceptual Modeling. Ballarat: Australian Computer Society, 2007: 17-26

[49] Vassiliadis P, Vagena Z, Skiadopoulos S, et al. Arktos: towards the Modeling, Design, Control and Execution of ETL Processes [J]. Information Systems, 2001, 26(1): 537-561

(上接第 14 页)

[32] Deng Yong, Wang Dong, Li Qi. An improved combination rule in fault diagnosis based on dempster shafer theory [C]// ICMLC. 2008. Kunming, China, July 2008: 212-216

[33] An A J, Stefanowski J, Ramanna S, et al. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing [C]// RSFDGrC 2007. Toronto, Canada, May 2007

[34] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases [A]// VLDB 1994 [C]. Santiago de Chile, Morgan Kaufmann. Sep. 1994: 487-499

[35] Shah K, Mahajan S. Maximizing the Efficiency of Parallel Apriori Algorithm [C]// ARTCom 2009. Kottayam, Kerala, Oct, 2009: 107-109

[36] Han J, Fu Y. Discovery of multiple-level association rules from large databases [A]// VLDB 1995 [C]. Zurich, Switzerland, Mar. 1995: 420-431

[37] Yu Wan-jun, Wang Xiao-chun, Wang Fang-yi, et al. The research of improved apriori algorithm for mining association rules [C]// ICCT 2008. Hangzhou, China. Nov. 2008: 513-516

[38] Chen H. An Intelligent Broker Architecture for Pervasive Context-aware Systems [D]. University of Maryland, 2004

[39] Roman M, Campbell R. Gaia: Enabling active spaces [C]// ACM SIGOPS European Workshop 2000. Kolding, Denmark, Sep. 2000: 229-234

[40] Dey A K, Salber D, Abowd G D. A contextual framework and a toolkit for supporting the rapid prototyping of context-aware applications [J]. Human-Computer Interaction, 2001, 16(2-4): 97-166

[41] Kim H, Cho Y, Oh S. CAMUS: A Middleware Supporting Context-aware Services for Networkbased Robots [C]// IEEE ARSO 2005. Nagoya, Aichi, Japan, June 2005: 237-242

[42] Behlouli N B, Taconet C, Bernard G. An Architecture for Supporting Development and Execution of Context-aware Component Applications [C]// IEEE ICPS 2006. Lyon, France, June 2006: 57-66