Application of Data Cleaning Method in Data Center of Electric Company

by

Zhang Xinghua

B.E. (Hexi University) 2002

A thesis submitted in partial satisfaction of the

Requirements for the degree of

Master of Engineering

in

Computer Application Technology

in the

Graduate School

of

Lanzhou University of Technology

Supervisor

Professor Chen Xu-hui

May, 2011

兰州理工大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：张兴华　　日期：2011年 6 月 10 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权兰州理工大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。同时授权中国科学技术信息研究所将本学位论文收录到《中国学位论文全文数据库》，并通过网络向社会公众提供信息服务。

作者签名：张兴华　　日期：2011年 6 月 10 日

导师签名：　　日期：2011年 6 月 10 日

目 录

摘　要

数据中心建设是国家电网“十一五”期间“SG186工程”重点实施项目之一，旨在利用数据仓库整合电力企业内部分散的业务数据，并通过便捷有效的数据访问手段，可以支持电力企业内部不同部门、不同需求、不同层次的用户随时获得自己所需的信息，并能将网络中分布的行业数据集成到一起，为决策者提供各种类型的数据分析。数据中心的主要特点是通过统一的数据定义和命名规范，保证数据的唯一性、准确性、完整性、规范性和时效性，提供一个标准的、一致的数据共享和访问平台。

本文主要研究对象是电力营销计费系统所需的底度电量数据值，针对在数据抽取过程中对电量数据中产生的“噪音数据”进行清洗，在文章中的对“噪音数据”的清洗过程是：首先利用切比雪夫原理设定一个判断区间来检测“噪音数据”，然后将这些“噪音数据”中的异常的属性值删除，然后对这些被删除的属性值当空缺值来处理。而对空缺值的处理，文中利用了基于遗传神经网络预测模型的填补空缺值的方法，该方法分利用了遗传算法的全局搜索能力和神经网络的非线性映射能力，在很大程度上提高了数据的预测精度。并验证了该方法的可行性以及在提高数据预测精度方面的有效性。

本文主要研究重点是数据中心在ETL过程中的数据清洗过程。根据国家电网“SG186工程”规划中关于数据中心的建设要求和数据中心ETL功能架构，数据中心的ETL过程主要分为抽取、清洗、转换和加载四个主要部分（较之通常的ETL过程，将清洗过程从原有的抽取过程中剥离出来单独研究），根据电力系统实际生产业务需要，本文又将数据抽取中的“清洗”过程划分为两个子流程：即先通过检测异常数据并将其值置为“NULL”，然后根据其余有效值对已置“NULL”的数据进行预测填充。

简而言之，随着计算机信息技术的不断发展和企业决策与辅助分析系统对数据质量要求的不断提高，数据抽取过程遇到的“噪音数据”或“脏数据”不能再做简单的删除处理，因为在删除这些数据时有可能丢失大量的可用的“关键数据”，因此数据清洗正在逐渐的演变为数据抽取中一个重要的环节，其理论研究与实践就显得非常具有实用价值。

关键词：ETL；数据中心；数据清洗；遗传神经网络；空缺值；电量数据；噪音数据；

Abstract

Data Center is one of "SG186 project" of the national grid during the Eleventh Five-Year. The purpose of the data center is the use of electric power enterprise data warehouse technology integration within the fragmented business data. And it supports the power of different departments within the enterprise, different needs, different levels of users to access the information they need access to data by means of convenient and effective, and it provides various types of data analysis for decision-makers by a way that the data of the industry distribution in the network to be integrated . The main features of the data center are to provide a standard consistent with the data sharing and access platforms for the enterprise through a unified data definitions and naming conventions, to ensure the uniqueness of the data, accuracy, completeness, standardization and timeliness.

The main focus of this paper is data cleaning process of the data center in the ETL (Extract Transform Load). According to the national grid "SG186 project" planning requirements on the construction of data centers and ETL functional architecture of data center, ETL process of data center is mainly divided into extraction, cleansing, transformation and loading of four main parts. In accordance with the actual production of electric power enterprise business needs, in this paper, data extraction in the "cleansing process" is divided into two sub-processes: Which is to detect abnormal data and its value is set to "NULL", and then to predict the value of these vacancies filled by other valid values.

In this paper, our main object of study is the data value of the electric energy billing system in the power marketing. In addition, the article also introduced the object model of electrical data and the reason of generating abnormal data, at the same time, proposed the way of the genetic neural network prediction models to fill the vacancy value. The effectiveness of this method was verified. And the data cleaning method mentioned in the text was applied to the construction of electric power enterprise data centers, to improve data quality dimensions of information silos in the past the problem of data for management decision-making data services to provide effective and appropriate decision support.

In short, with the continuous development of computer information technology and the business requirements of the quality of the data are constantly being

improved in the analysis systems of decision-making and assist, noise data or dirty data in the data extraction process of the encounter is gradually transformed into data extraction is an important part. To further improve the quality of data, we applied a method which based on genetic neural network to handle the missing values. This method fully used the global search ability of genetic algorithm and the nonlinear mapping ability of neural network, so that the prediction accuracy of the data was greatly improved. The experiment shows that this method is feasible and effective in improving the prediction precision of data.

Key words: ETL; Data Cleaning; Data Center; Genetic Neural Network Algorithm; Power Data; Vacancies Value;

插图索引

附表索引

第1章 绪 论

1.1 研究背景与意义

1.1.1 研究背景

目前伴随着信息技术的迅速发展，商务智能技术广泛的应用在IT的各个领域当中。尤其是在以网络技术、数据库技术为支撑的商业企业当中，规范的、系统的计算机应用建设已成为迫切的需要和发展的趋势。从总体上看信息管理的现状，绝大部分的机构及其组织都存在着多个异构系统，因此数据的组织结构和存储结构就存在着各不相同的情况，从而就形成了数据的多样性和异构型等“信息孤岛”的问题。针对这些大量存在的“信息孤岛”现象，某电力系统利用数据仓库技术，通过便捷有效的数据访问手段，来支持企业内部不同部门、不同层次、不同需求的用户随时获得自己所需的信息，来整合电力企业内部所有分散的原始的业务数据，并能将网络中分布的行业中的不同的数据源集成到一起，以便于为各应用系统提供集中的数据服务环境，并为决策者提供各种类型的数据分析。因此就要求行业内部对数据集中存储统一管理，通过统一的数据定义和命名规范，保证数据的唯一性、准确性、完整性、时效性以及规范性，提供一个标准一致的共享共用数据的平台——这就是文章中要介绍的数据中心。数据中心是解决异构环境中信息的正确性及实现信息的高效共享和交换的重要手段，而解决这些问题的有效方法就是数据仓库技术。数据清洗从字面上也看的出就是把存在问题的数据清洗成为干净的数据。因为数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来的，而且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是影响我们正确使用的数据，称为“噪音数据”。我们要按照一定的规则把“噪音数据”中的“噪音”给“洗掉”，这就是数据清洗。不符合要求的数据主要包括错误的数据、不完整的数据、重复的数据三大类。数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门。然后有专门的业务部门来确认是否过滤掉，还是由业务单位修正之后再进行抽取。要了解数据清洗我们的先知要清楚数据仓库，他们的联系是分不开的，著名的数据仓库专家 WH. Inmon 在20世纪90年代初期，在他的著作Building the Data Warehouse 一书中给出了数据仓库的概念，对数据仓库的描述如下[1] ：数据仓库是一个面向主题的、集成的、不可更新相对稳定的、来反映历史的时间变化的数据集合，用于支持管理决策。主题是与传统数据库的面向应用相对应的，是一个抽象概念。面向主题的数

据仓库不但为综合数据、历史数据的处理提供了一种行之有效的解决办法，也为有效地支持组织机构经营管理决策提供了全局一致的数据环境，也是在较高层次上将企业信息系统中的数据综合、归类并进行分析利用的抽象。每一个主题对应一个宏观的分析领域。数据仓库处理对于决策无用的数据，提供特定主题的简明视图。并最终为各级决策管理者提供科学、及时、准确、有效地辅助决策依据。

而数据的缺失、冗余、不一致、不确定等诸多情况的存在是不可避免的。我们称这些数据为“噪音数据”，而“噪音数据”又是数据挖掘或数据仓库以及数据质量管理中数据处理的重要环节。 “噪音数据”会影响着数据集中导出规则是否准确，以及抽取模式是否正确。依照“垃圾进，垃圾出”的原理[2]，利用还有“噪音”的数据进行决策分析就会得到错误的分析结果，从而误导了企业领导作出的决策。当一些企业针对一些历史的或现存的数据为对象为将来的企业发展作决策或预测时，这些“噪音数据”的清洗工作就变得非常关键了。同时，这些“噪音数据”还会造成大量时间都花费在数据的ETL阶段。因为在数据仓库建立阶段有大量的实践证明大部分时间都花在数据的ETL阶段，而数据清洗又在数据的ETL阶段占有相当一部分工作时间。随着计算机信息技术的不断发展和企业决策与辅助分析系统对数据质量要求的不断提高，数据抽取过程遇到的“噪音数据”或“脏数据”不能在做简单地删除处理，因为在删除这些数据时有可能丢失大量的可用的“关键数据”，因此数据清洗正在逐渐的演变为数据抽取中一个重要的环节，其理论研究与关键实践就显得非常具有实用价值。

数据的ETL阶段主要针对的是数据源，不同的数据源可能会有不同的抽取流程。抽取的数据源中可能存在质量问题如[3]：

1) 数据的不完整，这一类数据主要是由于一些应该有的信息缺失造成的数据不完整，如供应商的名称、客户的区域信息、分公司的名称等属性的缺失，还有业务系统中主表与明细表不匹配等的问题。对于这一类不完整的数据，要把他们过滤出来，把缺失的内容通过相应的技术手段按照要求补齐，补齐后的完整数据才能写入数据仓库。

2) 错误的数据，这一类问题产生的主要原因是企业的业务系统不够完整和健全造成的，这类错误又被认为是计算机造成的，是在接收输入后没有进行判断直接写入后台数据库造成的。这一类数据也要分类，对日期格式不正确的或者是日期越界或打了一个回车、数值数据输成全角数字字符等的这一类错误会导致ETL运行失败，这一类错误需要去业务系统数据库利用一定的技术手段挑出来，然后通过相应得方式方法修正，修正之后再进行抽取。对于一些简单的错误，如数据前后有不可见字符等的问题，只能通过写SQL语句的方式就可以找出来，然后要求客户在业务系统修正之后抽取。

3) 重复的数据，是在实际的数据处理中常见的问题，在数据仓库环境下特

别重要，因为在集成不同的系统时会产生大量的重复记录。这类问题可以是针对一个数据集的也可能是两个数据集或者一个合并后的数据集。通常情况下，指向同一个现实实体的两条记录的信息是部分冗余的，它们的数据互为补充。对于这一类数据的处理——特别是如将表中的重复数据记录的所有字段导出来，并通过判断之后确认并整理。数据清洗是一个反复的过程，一个长期的、系统的过程，不能一蹴而就，不可能在几天内完成，只有不断的发现问题，解决问题。对于是否过滤，是否修正一般要根据实际需要来确认，对于过滤掉的数据或者将过滤数据写入数据表，在ETL开发的初期将会不断地出现不同的错误数据，促使他们尽快地修正错误，同时也可以作为将来验证数据的依据。数据清洗需要注意的是不要将有用的数据过滤掉，对于每个过滤规则认真进行验证，并要用户来不断进行确认。

总之，对来自不同数据源的数据，对同一个概念有不同的表示方法。在集成多个数据源时，需要解决模式冲突的问题，主要就是为了解决数据的错误和不一致等情况。对相似或重复记录的处理问题，其主要任务是需要检测出并且合并这些记录。数据清洗过程就是我们解决上述问题的过程。数据清洗的目的是检测数据中存在的错误和不一致，剔除或者改正它们，以提高数据质量。

基本上绝大多数企业中都使用了数据仓库技术进行信息辅助决策。而数据的抽取、转换、装入(Extract Transform Load, ETL)又是创建数据仓库系统的重要环节，它从所有异构系统中采集数据，并对其进行高效的转换，它能够很好地解决组织机构内部的数据一致性与信息集成化问题，在一个数据仓库项目中，大量的工作都花费在ETL阶段。而数据清洗是保证信息源的数据质量，从而保证了辅助决策的原始数据的正确性和准确性。“噪音数据”[6] 是ETL（抽取、转换、装入）过程中不可规避的问题，同时它也是衡量企业级数据仓库优劣的基础指标之一。因此数据清洗就是提高数据质量的有效方法，选择一个高效、便捷的数据清洗算法是具有重要意义的。

1.1.2 研究意义

本课题的研究主要有以下三个方面的意义：

①.保证数据仓库的数据质量

ETL是数据仓库技术包括的内容之一，在ETL数据抽取过程中，“噪音数据”的存在是无法避免的，而数据质量问题是制约数据仓库应用的障碍之一。如果数据质量达不到要求，将直接导致数据仓库技术不能产生理想的结果，甚至会产生错误的分析结果，从而误导决策。ETL“噪音数据”清洗的任务就是通过各种措施从准确性、一致性、无冗余等方面提高数据仓库的数据质量。

②.提高ETL的工作效率

在数据仓库建设中，约70%的工作量都花费在ETL阶段，ETL程序经常会面对数据缺失、数据异常等众多“噪音数据”的情况，ETL程序算法不得不占用更多得系统软/硬件资源去处理这些问题，从而导致整个ETL处理工作量大、运行时间长等问题，因此，改进、优化ETL过程中对“噪音数据”处理算法，必定会大大的降低系统资源的占用率，从而提高整个ETL工作效率。

③.降低企业硬件投入

ETL程序算法的优劣，直接受益的另一个方面就是企业对系统硬件的投入，如果算法复杂，则需要占用较高的CPU资源，有可能还会消耗更多的内存资源，企业就必须不断的更新系统硬件满足越来越多的数据处理(在最初的数据处理过程中可能不明显，然而随着数据量的不断增加，这种现象会越来越明显)，并且会明显的感觉到投入产出比的不和谐。

1.2 国内外研究现状

数据仓库的领头设计师 W.H.Inmon 提出了数据仓库概念，数据仓库是一个面向主题的、集成的、随时间而变化的、不容易丢失的数据集合，支持管理部门的决策过程[1][2] 。数据仓库的四个重要特征就是[1]：面向主题的、集成的、与时间相关的、不可修改的数据集合等。所谓主题是指用户使用数据仓库进行决策时所关心的重点方面。面向主题性表示数据仓库中数据组织的基本原则，数据仓库中的所有数据都是围绕着某一主题组织、展开的。

其中ETL（数据抽取(Extract) 、转换(Transform) 、清洗(Cleansing) 、装载(Load)) 过程又是数据仓库建立过程中的重要环节。在数据仓库项目中，大约70%的工作量在ETL阶段，因此在此过程中难免出现大量的脏数据，为了保证高质量的数据，所以在数据进入数据仓库之前必须清洗。

数据清洗是一个较新的研究领域，对大数据集的清洗是很费时的工作，清洗过程计算量较大，很难用传统的算法操作。目前，数据清洗还没有公认的定义，不同的应用领域对其有不同的解释。为了保证应用于数据仓库前端的决策支持系统产生正确的决策分析结果，就要提高数据的质量。数据清洗技术提高数据质量方法之一，数据清洗的目的是检测数据中存在的错误、不一致数据和重复记录，并消除或改正它们，从而提高数据质量。数据清洗过程被定义为下面三个步骤[4]：(1) 定义和确定错误的类型(2) 搜寻并识别错误的实例(3) 纠正所发现的错误。数据清洗主要应用于三个领域[5]：数据仓库(DW)、数据库中的知识发现（KDD，又称为数据挖掘）和数据/信息质量管理（如全面数据质量管理TDQM）。

1.2.1 数据清洗的研究现状

从总体上看国内外的相关研究主要有以下几个方面[12]：① 在对数据集进行异常检测和清洗处理的过程中增加人工判断，来防止正确数据的错误处理。② 对大型数据集的处理时提出高效的数据异常检测算法，来避免扫描整个庞大的数据集，例如针对大型数据集进行增量处理的数据清洗算法；③ 数据集中的重复数据的清洗；④ 数据清洗时对数据集中文件的处理⑤通用的与领域无关的数据清洗框架的建立。

针对具体的应用领域和方法主要作了如下的分类概括：

(1) 对数据集进行异常检测[14]，是指对数据集中记录的属性值的清洗过程。对属性值的清洗主要有方法如下：

文献中[29]提出了采用基于距离的聚类方法来识别异常记录；采用关联规则的方法来发现数据集中不符合具有支持度的规则和高置信度的异常数据。采用基于模式的方法来发现和数据集中现有的模式不符合的异常记录；在文献[19]采用统计学方法检测数值型的属性值，其方法是计算字段值的均值和标准差，考虑每一个字段的置信区间来识别异常字段和记录。属性清洗可以针对具体问题具体分析，也可针对某类问题提供解决方案。对其值为字符型的属性利用了属性间的约束关系、模式识别技术等，难度较大。如果清洗方案能自动发掘规则，则属于自适应性属性清洗，实现难度很大，这种方案不常见[31]。

(2) 针对大型数据集进行增量处理的数据清洗算法

对于大型数据集进行并行，增量处理的研究。目前对其的已有研究成果主要集中在数据ETL工具上，正对某些商业ETL工包括转换工具和清洗工具，已经具有的利用多进程、多线程、流水、多处理器等技术来进行数据的并行集成与清洗，并再次基础上提供数据的增量复制功能[36]，但它缺少姓名、地址等信息的清洗、也缺少模糊匹配和合并的功能。

(3) 识别并消除数据集中的近似重复对象，也就是重复记录的清洗。因为在集成不同的系统时会产生大量的重复记录，消除数据集中的近似重复记录问题是目前数据清洗领域研究的最多的内容[31]。

(4) 建立通用可扩展的数据清洗框架

不少数据清洗方案和算法都是针对特定应用问题的，只适用于较小的范围。而通用的与应用领域无关的、可扩展的算法和方案较少。数据清洗工具有ETL工具和专用的清洗工具。但是商业的ETL工具中虽然提供了一些数据清洗功能，但是这些清洗功能都缺乏扩展性。因此，有不少的研究人员提出了通用的数据清洗系统的框架。并且根据这些框架的要求的功能，又提出了数据清洗的模型和语言。因此在通用SQL语言基础上扩展了新的数据清洗操作。

1.2.2“噪音数据”清洗的研究现状

目前对数据的清洗主要还是“噪音数据”的清洗。而“噪音数据”清洗是数据的ETL过程中数据清洗的重要组成部分。“噪音数据”数据产生的常见原因主要有：1.在数据的采集过程中数据的采集设备有问题造成的；2. 在数据的传输过程中发生错误；3. 在对数据进行录入的过程中发生了人为或计算机的错误；4. 由于命名规则或数据代码不一致而引起的不一致。因此噪声数据[6] 就是指数据中存在着错误或异常（偏离期望值）的数据。

将目前对“噪音数据”的相关研究综合起来，可以将“噪音数据”的方法模型与过程模型组合起来，从而构建出一个清晰的多维度的过程方法的链接。在每一阶段的任务处理中都会对应相应的一种或多种方法为之服务。下面将方法模型、过程模型和构件模型都组合起来，可以得到多维度的综合映射，从而得到作“噪音数据”清洗方法模型[29] 。

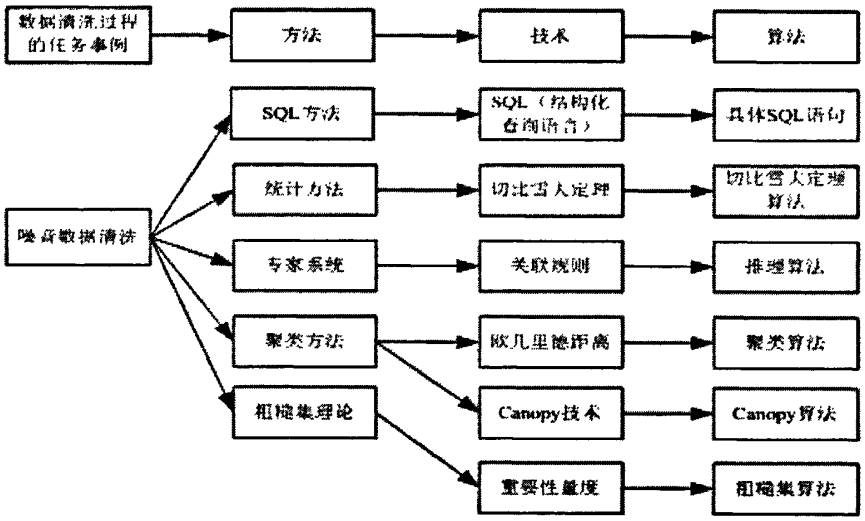

图1.1“噪音数据”清洗方法模型

文献[8] 提出了基于统计的“噪音数据”处理方法，这种方法称为基于平均值的“噪音”识别，它是根据统计学原理(切比雪夫定理)，使用数据的平均值、标准差、置信区间可以识别异常数据。有时用中值取代平均值，也称为基于中值的“噪音”识别法。基于统计的数据清洗方法比较适合数值型数据的清洗，在日常生活中被广泛使用(如奥运会上体操、跳水等比赛项目的打分制度)。该方法的缺点是在计算字段值的均值和标准差时，需要用每一个字段的置信区间来识别异常字段和记录，所以在对“噪音数据”进行处理前需要了解数据集中的“噪音”规模 。

在文献[29]中提[29] 出的基于聚类分析的在属性级别上处理噪声数据的算法。这个模型还可以为“噪音数据”的产生过程建模，这是一种在属性级别上识别噪

声数据并进行清洗的算法模型。这种方法主要用于数据质量方面的数据清洗任务。此方法是统计噪声在属性上的分布规律，并在属性级别上识别“噪音数据”，并根据其他的干净数据对其进行矫正。同时，该方法也是有缺点的，其缺点是用干净数据中的（期望）值矫正“噪音数据”的，这样的话对数值型数据来说精确度就有一定的问题。

大多数文献中提出的对于关系型数据库中“噪音数据”的处理，大量的工作集中在记录级别上发现“噪音数据”点并删除掉这些记录。这种方法有很高的执行效率，但它的缺点是一旦认定一条记录是“噪音数据”，那么这条记录中关键的干净的数据信息也将丢失。例如文献[32]中提出了一种在大型数据库中检测异常数据的线性方法，这种方法是在记录的级别上识别噪声数据，即假定整条记录要么是噪声，要么是干净的数据。因此在一些噪声数据单元很多的数据库中，要想有效地提高数据的质量来实现信息的最大化价值，这种方法显然不可行。

对于聚类分析法在噪声数据清洗中的算法还是比较多，聚类[29][44]是按某种标准将数据集分组为多个组或簇，同一簇中的数据具有高度相似性，而不同簇的数据差别较大。目前已有大量聚类算法，算法的选择取决于数据自身的特征和聚类应用的目的，典型的聚类算法如k-means(k-平均值)[44]、k-medoids（k-中心点）[45]算法。聚类可以根据大多数原则，把被分组在较小簇中的数据视为“噪音数据”来实现“噪音数据”清洗。用于“噪音”识别的聚类方法主要指基于距离的噪声识别：该方法是以数据集中两两数据间的距离为分组依据，对数据集进行聚类分析。而聚类方法的缺点是在对“噪音数据”处理之前需要了解数据集中“噪音数据”的分布情况，否则难以确定聚类分组的次数。

文献[41]该文所提出的基于遗传神经网络的数据清洗模型在预测待填补值。该方法预测精度比较高，大大提高了数据的干净程度，因为此方法充分利用了遗传算法的全局寻优特性和神经网络的非线性映射能力。遗传神经网络的数据清洗是进行数据清洗的有效方法。但此方法的缺点是缺少对噪声数据的识别。

总的来说目前对数据的清洗还没有独立出来作为一个课题，因此关于数据清洗的相关书籍还是比较少，大都是在数据仓库技术或数据挖掘技术中仅有很少的篇幅来介绍，而对数据清洗的文章只能在期刊论文和学术会议论文中看到，对“噪音数据”清洗的论文也不多。因此，要完成此项任务还是很有挑战性的。

1.3 本文的主要工作

本文根据某电力“SG186工程”数据中心建设过程中对其所需的电量数据中产生的“噪音数据”在ETL过程中的清洗技术作为主要内容。本文的主要思路：对数据中心中遇到的一些电量噪音数据如何进行处理，在文章中先是对“噪音数据”进行识别，并对识别过程中的异常数据进行删除，然后再对缺失数据进行平

滑（预测填补）。在本课题中采用统计和前馈型遗传神经网络算法相结合的数据清洗方法模型，即在噪音识别阶段采用基于统计（根据切比雪夫定理）的中值识别，在噪音平滑阶段采用基于遗传神经网络的预测模型进行缺失数据的预测填补。这样既可以有效地解决了空缺值的问题也可以用同样的方法来解决“噪音数据”的问题。

1.4 本文的内容安排工作

本文主要讲述了某电力系统在数据中心建设中如何利用遗传神经网络算法来有效的解决其在 ETL 过程中遇到的“噪音数据”和“空缺值”的问题的。本文主要分为四章。

第一章为引言部分，提出了这个项目的研发起因：

第二章详细的介绍了数据清洗和 ETL 的有关概念，从它的由来、扮演的角色、应用的方面都做了描述。

第三章是对某电力系统的数据中心整体的建设进行了描述，明确了数据中心的目标与定位，并介绍了了数据中心的系统架构和数据仓库的执行架构。

第四章主要是对数据中心的电量数据的特点进行了描述，以及电量数据的缺失和解决电量数据缺失使用到的关键算法作了阐述，并且介绍了该算法是如何解决该数据中心中的“噪音数据”问题的。数据中心的脏数据处理方案以及电量“噪音数据”清洗方法在电力系统数据中心的应用。

最后是总结和展望，对全文进行了总结，并对下一步要进行的工作进行了展望。

第 2 章　数据清洗技术

2.1 数据清洗概述

数据清洗是可以保证数据信息源的数据质量的方法之一，因此数据清洗是构建数据仓库过程中的不可缺少的重要环节。使用数据清洗技术，数据被移到数据仓库时，它们要经过转化，以确保数据的一致性。在实际的数据集成过程中，可能遇到的从异构数据源中集成数据、从网络中集成数据或者有些数据直接来自文本数据库或EXCEL表格，这些都是数据需要清洗不可避免的问题。

数据仓库是一个面向主题的、集成的、随时间而变化的、不容易丢失的数据集合。数据仓库集成包括数据抽取(Extract) 、转换(Transform) 、清洗(Cleansing) 、装载(Load) 的等集成过程。用户从数据源抽取出所需要的数据，经过数据清洗，最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去。由于各个业务应用系统在编码、数据类型、命名习惯、实际属性、属性度量等方面不一致，当数据这些业务数据进入数据仓库时，需要采用某种方法来消除这些不一致性，从而保证了辅助决策数据的正确性和准确性。

2.1.1 数据清洗的概念

数据清洗是一个减少错误和不一致性、解决对象识别问题的过程。迄今为止，不同的应用领域对其有不同的解释，因此数据清洗还没有公认的定义。数据清洗在本文中的定义是：在对数据源进行充分地分析后，利用有关技术和方法策路将从单个或者多个数据源中抽取的“噪音数据”经过一系列转化使其成为满足数据仓库质量要求的数据，我们把这样的过程称之为数据清洗。这里的相关技术指的是，如预定义的字典函数库、清洗规则、及重复记录匹配等等。数据清洗(Data Cleansing)技术是改进数据质量的有效方法之一。数据清洗问题也叫做“噪音数据”的处理。数据清洗是较新的研究领域，对大型的数据集的清洗是很费时的工作，因为清洗过程计算量较大，因此很难用传统的算法来操作。下面我们对数据清洗的解释是根据数据清洗在不同领域的中的应用来进行解释的。

(1) 在数据仓库领域中的数据清洗[28]

数据仓库的用处是对数据源的进一步分析操作，它是商务智能的基础，所谓数据仓库就是在企业管理和决策中面向主题的、稳定的、集成的、与时间相关的数据集合，用以支持经营管理中的决策制定过程。例如企业的许多数据报表就是在这一阶段生成的。数据仓库能够为多维分析和数据挖掘提供所需要的、整齐一

致的数据，因此数据仓库的重点就是能够准确、可靠、安全的从数据源中抽取数据，然后再经过加工转换成有规律的信息，以便管理人员进行分析使用，因此它为商务智能提供数据来源。数据仓库还可以实现数据的集成、传输、整合、清理和储存。为了有效地对数据仓库中的数据进行管理，我们将各部门业务应用的数据汇总到ODS，然后以此为基础再通过将ODS数据库的数据移入到数据仓库这样的形势来建立数据仓库。建立数据仓库的过程一般包括抽取操作数据并进行扩展，按要求自动更新数据仓库，以设定自动程序来定时抽取操作数据，并且能预先执行合计计算等步骤。这种数据仓库不但能够保存历史数据、阶段性数据，并从时间上进行分析，而且能够装载外部数据，接受大量的外部查询。因此在数据仓库领域中，数据清洗被定义为清除错误和不一致数据的过程，并需要解决元组重复问题。当然，数据清洗并不是简单地用优质数据更新记录，数据清洗还涉及到了数据的重组和分解的问题。

所谓ODS就是操作型数据存储[33]，在企业数据中心的主要架构中ODS是重要的功能区域。他的主要作用除了集成了来自部属于各个业务系统中的业务数据源外，ODS还起到的作用就是作为缓冲区数据在进入数据仓库区域前的，并形成一致的企业数据集成视图通过对数据ETL的过程中。因此它的主要目的是使得最终用户更好地通观全局，我们把这一区域称作ODS的统一信息视图区。通过对数据进行ETL操作形成了统一信息视图，而ODS中的统一信息视图区是有选择地来集成各类业务数据。统一信息视图是以数据主题域的方式对数据进行组织而形成的。通过这一逻辑视图，最终的用户就可以很快速获得与这个主题相关的、最近时间内的完整信息。统一信息视图区内由于统一信息视图具有准实时性的特点，因此，企业的业务系统中相关数据的更新，需要在一定时间间隔内，反映到统一信息视图区的相关主题中。这里的时间间隔确定的时候是以业务需求为驱动，因而一般情况下是准实时。通常情况下，从各业务系统向ODS缓冲区的数据移动是数据流动过程的第一步，实现了数据仓库从数据源头（即各个业务系统）将数据抽取出来并装载到ODS中的缓存区这一过程。由于考虑到各个业务系统数据库平台的异构性，数据仓库需要通过将数据放置于ODS中缓存区以得到统一的企业数据平台，方便进行后继的数据清洗/转换等操作。

数据仓库环境中，数据在准备装入数据仓库前，可从三个环节进行数据清洗，从而来提高并保证数据质量的，如图2.1所示。

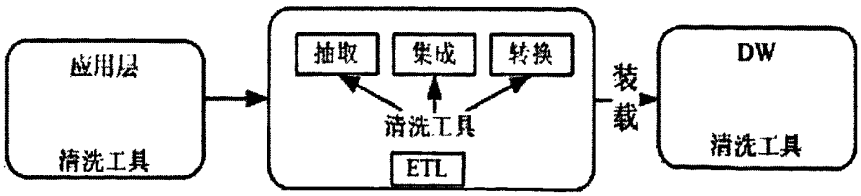

图2.1基于数据仓库的数据清洗

- 应用层：也就是数据来源，对这一层的清洗即在应用程序环境中进行数据清洗；
- ETL 层：把数据从应用系统中抽取进入整合转换层时进行数据清洗；
- 数据仓库层：在数据装入数据仓库后进行数据清洗。

(2) 数据挖掘领域中的数据清洗[23]

数据挖掘在人工智能领域，习惯上又称为数据库中知识发现，在数据挖掘的过程中，只有把要进行挖掘的数据预处理成便于挖掘的形式，才能从海量数据中得到高质量挖掘结果。数据清洗是数据挖掘的第一个步骤，即对数据进行预处理的过程。各种不同的知识发现和数据仓库系统都是针对特定的应用领域进行数据清洗的。文献[26]认为，信息的模式被用于发现“垃圾模式”，即没有意义的或错误的模式，这属于数据清洗的一种。

(3) 数据质量管理领域中的数据清洗

随着企业信息化发展，企业数据的管理是一个重要的方面，因此数据质量管理就被越来越多的企业所关注。全面数据质量管理解决整个信息业务过程中的数据质量及集成问题。通过数据清洗，将逐渐形成完整和准确的企业数据视图，为经营分析和生产支撑提供可靠的数据来源。数据质量问题是客观存在的，数据质量问题的管控工作将贯穿数据中心系统建设的整个过程。

2.1.2 数据清洗的基本原理

数据清洗的基本原理[12]就是通过对 “脏数据”或“噪音数据”的产生原因和存在形式进行分析之后，再利用现有的方法策略和技术手段对存在的“噪音数据”进行合理有效的清洗，这样“噪音数据”就被转化成立了能满足相应的应用要求或数据质量要求数据，从而提高数据集的数据的可靠性和准确性。而对“噪音数据”产生的原因和存在的形式进行分析时，就需要对数据集流经的每一个过程进行深入考察，从中提出数据清洗的策略和清洗规则。最后在数据集上应用这些策略和清洗规则来发现和清洗“噪音数据”。这些清洗规则和策略的合理与否，决定了清洗后数据的质量。数据清洗的原理可总结为如图所示。

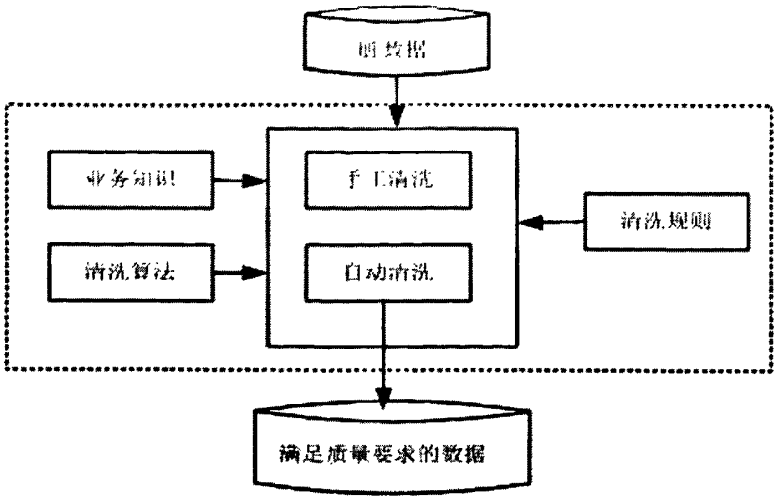

图2.2数据清洗的基本原理

2.1.3 数据清洗的步骤

在对数据进行清洗的过程中我们把数据清洗分为以下四步，其具体步骤如下[12]：

(1) 分析数据

分析数据是数据清洗的前提与基础，就是确定数据来源。数据清洗通过必要的人工检视和数据抽样调查之外还可以借助某些数据分析程序获得详细的数据质量信息（比如关于数据属性的各种元数据），然后分析检测数据中的错误或不一致情况以及发现数据集中存在的哪些具体的数据质量问题。因此首先必须对数据进行仔细的分析，数据转换流程和映射规则的定义。

数据分析主要分为如下两种：数据挖掘和数据派生。数据挖掘可以被看成是一种对数据进行搜寻的过程，这一过程不必预先提出假设或问题，但是通过这一过程就能找到那些不是预期的但却是令人关注的信息(这些信息表示了数据元素的关系和模式）。数据挖掘是从大量的数据集中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据派生可以得到属性的很多相关的信息，例如，可以得到属性值的最大值、最小值、平均值、中间值、和标准差等统计值，这些信息是通过对数据应用领域理解之后应用数理统计技术计算而来的。数据派生主要针对单独的某个属性进行实例分析。比如，数据取值区间、类型、长度、出现频率和他们的离散值、不同值的个数，出现空缺值的次数，和典型的字符串模式等。

数据分析是根据数据源的个数、数据源中不一致数据和“噪音数据”的程度等，对数据执行大量的数据转换和清洗等操作。之后再根据数据分析得到的结果来定义数据清洗、转换的规则与工作流程。要想使转换代码的自动生成变成可能，

就尽可能的为数据转换和清洗指定一种与模式相关的匹配和查询的语言。

(2) 验证

数据转换工作流的验证，及其相关定义的正确性和有效性的评估是十分必要的。例如我们可以对目标数据源的一份拷贝或者是一部分抽样数据进行实验，以此为基础，有的时候就需要对整个数据清洗流程的定义进行改进，因此就要对其进行必要的验证和评估。分析、设计、验证这三个步工作有时可能需要多次重复进行，因为只有在完成了某些数据转换之后才，有些数据质量问题会暴露出来。

(3) 执行数据清洗的过程

执行预先定义好的并且已经得到了验证的数据清洗、转换规则和工作流程等在数据源上，等经过上面三个步骤之后。利用相关的技术手段多数据执行清洗操作。为了防止需要撤销上一次或几次的清洗操作，当在数据源上直接进行清洗时，就需要备份源数据。清洗时，为了得到较为有效、准确的数据清洗流程的定义，根据噪音存在的不同形式，执行一系列的转换步骤来解决数据质量问题在模式层和实例层上的。此时即可通过相关的工具或方式方法执行整个数据清洗流程，比如数据仓库中常用的ETL工具以及其他的方法，以及各种其他的数据质量工具。

(4) 信息更新

经过数据清洗之后接下来就必须开始更新最准确的信息。即就是通过清洗已经基本去除了原数据源中存在的各种数据质量问题，因此这些干净的数据应该替换数据源中原来的“噪音数据”。使得基于这些数据的应用能够采用改善之后的数据，并避免在再次抽取数据时进行重复的清洗工作。

2.2 数据清洗的方法技术

数据清洗涉及到的应用领域很广泛，当前被提出的数据清洗的方法及其策略种类繁多，不仅涉及到了数据仓库相关领域，还涉及数据挖掘领域以及人工智能专家系统等。例如有遗传神经网络的数据清洗；基于聚类模式的数据清洗； 基于模糊匹配的数据清洗[29]；基于粗糙集理论的数据清洗；专家系统体系结构的数据清洗等等。因此我们可以把清洗子过程按照清洗的流程分为数据预处理，属性清洗，重复记录清洗三个步骤。

2.2.1 数据预处理

数据预处理指在主要的处理以前对数据进行的一些处理。在这一步中系统主要负责处理空缺值。首先我们了解处理空缺值的一般的算法(即方法，为了便于元数据对算法的管理，下面我们统称之为算法)。

(1) 空缺值的处理[8] [43]。

在获取数据的过程中，无法避免空缺值的产生。缺失值是指该值理论上是存在，但没有该值的记录。空缺值产生的原因有，1）在记录信息时，某些数据项的信息无发得到，如未成年人的身份证信息无法得到。2）有些信息在当时被认为是没有必要的。3）由于误解或检测设备出现问题导致的相关数据无法记录。4）对数据修改时忽略或与其他记录不一致而被删除等。

空缺值的处理方式基本上都是估计来进行填补的。就是将空缺值可以通过本数据源或其他数据源推倒出来。常用的方法归纳如下：

- 基于统计的数学方法。利用概率统计的方法得到的值代替空缺的值，但准确性比较低。如可用平均值，中间值、最小值、最大值等来替换空缺值，但平均值从某种意思上来说也是最准确的处理方式，是最常见和最简单的处理方式。
- 人工处理方法。一般来说，该方法是很费时，因为要靠人工尝试输入一个可接受的值。并且当数据集很大，而 数据的空缺值又很多的时，则使用该方法就会出现事倍功半的结果。
- 属性均值法。利用同一属性的平均值来填充空缺值。
- 预测填充法。当然这种办法也得结合具体的应用领域。即使用最可能的数学预测模型来预测计算得到的值来填充空缺值如：可以使遗传神经网络得预测、以及用贝叶斯形式化或用回归的基于推理的工具来预测填补。
- 忽略元组法。针对某些属性值在某些应用上根本不起作用或此属性值可有可无等情况时，就可以忽略此元组值。
- 忽略记录法。与忽略属性值法相对应的则是忽略记录法。这种清洗意义不大，当该条记录所缺少的属性占总属性的百分比超过一定的比率，则要考虑到该条数据为干扰数据还是干净数据。

(2) 错误值处理

主要是错误值的检测。其方法主要集中在以下几个方面[39]：

- 回归与聚类：基于完整的数据集，建立回归方程（模型）。回归分为线性回归和非线性回归。聚类将类似的值组织成群或“聚类”，通过聚类可以检测到孤立点。回归可以利用一部分变量的值，找到适合数据的数学方程式，可以利用其与的正常值来预测错误的值，用来消除噪声，因此回归分析可以通过让数据代入到一个适合它的函数来平滑数据。。
- 利用统计分析或人工智能的方法检测属性可能的错误值或异常值。这些方法能基本上不依赖于领域知识，自动化程度较高，能为开发高效的自动数清洗工具提供帮助。
- 分箱 (binning)：通过考察“邻居”（周围的值）来平滑存储数据的值，用“箱的深度”表示不同的箱里有相同个数的数据，用“箱的宽度”来

表示每个箱值的取值区间为常数。由于分箱方法考虑相邻的值，因此是一种局部平滑方法分箱方法通过考察周围的值来平滑存储数据的值，存储的值被分布到一些箱中，由于分箱参考临近的值，因此它是局部平滑。即箱中的每一个值被替换成平均值；也可以按箱边界平滑，即箱中的每一个数据被替换成离它最近的边界值。

- 使用简单规则库(常识性规则和业务特定规则等)检测和修正错误或使用不同属性间的约束检测和修正错误。

(3) 不一致值

不一致数据是由于系统和应用造成的不同的数据格式、类型、制式、粒度和编码方式等造成的不一致，另外由于硬件或软件故障，错误的输入，不及时更新造成的数据库状态改变等造成的数据的不一致值。多数据不一致值的清洗方法，主要是在分析不一致性数据产生原因的基础上，应用多种格式函数、变换函数、标准函数库和汇总分解函数去实现清洗。而均值填充法是指缺失数据的值用其他完整数据的算术平均值来填充，这是一种最常用的缺失值填充法。应用均值填充法将可能会影响缺失数据及其它与其他数据之间的相关性。而曾经有人提出这样的原理："在正态分布下，样本均值是估算出的最佳的可能取值"。

2.2.2 属性值级别的清洗

数据的属性清洗详细的分为异常数据检测和属性值的属性两部分。数据清洗阶段系统负责处理错误值("噪音数据")的清洗。一般来说检测过程和清洗过程就可以安排在一起，在检测时候发现错误值就可以随之进行清洗纠正，检测数据集中属性错误值的方法主要有下面几种[37]：

- 使用简单规则库(常识性规则和业务特定规则等)检测和修正错误。
- 使用外部数据源检测和修正错误。比如数据字典等。
- 分箱(binning)：分箱方法是把属性值被分布到一些等深或等宽的"箱"中，通过考察属性值周围的值，比如按照箱中属性值的平均值或中值来替换"箱"中的属性值。分箱方法考虑的是相邻的值，因此是一种局部平滑方法。
- 利用统计分析或人工智能的方法检测属性可能出现的错误值或异常值。这些方法能基本上不依赖于领域知识，自动化程度较高，能为开发高效的自动清洗工具提供帮助。统计分析的方法是指利用切比雪夫定理，根据属性值的期望、标准差，考虑每一个属性取值的置信区间来识别异常的属性和记录。这种方法是根据统计学原理(切比雪夫定理)[19]，使用数据的平均值、标准差、置信区间可以识别异常数据，称为基于平均值的

噪声识别；有时用中值取代平均值，称为基于中值的噪声识别。

2.2.3 重复数据的清洗

目前在重复记录的清洗上作了很多研究，提出了较好的算法。在关系数据库系统中，重复记录就是指两条记录在所有的属性上的值都完全相同的记录。重复记录清洗问题就是重复记录的匹配和合并也被称为对象标识问题。而相似重复记录的清除可以是针对两个数据集也可能是一个合并之后的数据集。通常情况下，两个记录的合并是指向同一个现实实体的两条记录的信息是部分冗余的，它们的数据则可以相互补充。因此，通过把这样的两条记录合并，能够更准确地反映该实体的信息。

因此，检测和消除重复记录的问题是数据清洗和数据质量领域研究的主要问题之一。

重复记录清洗的基本过程一般包括以下几个阶段[29]：

1．预处理。

(1) 选择属性。选择需要的记录相匹配的属性。一般情况下由于数据集中记录的属性很多，在进行记录匹配时需要进行属性选择，这需要数据清洗人员对数据本身的含义有深入的了解，选择出能代表记录特征的属性来。

(2)初步聚类。其主要是对数据库中的记录进行排序。在进行记录匹配操作之前通常需要采取某种方法将潜在的可能的重复记录调整到一个相邻的位置空间，从而对于特定的记录可以将记录匹配对象限制在一定的范围之内，因为数据清洗需要处理的数据集往往较大，而且记录匹配是一个复杂的计算过程，为了在有限的时间内完成重复记录的检测，同时还要保证检测结果具有较好的准确率和查全率，必须采取某种技术手段限制记录匹配的范围。因此我们称以上这个过程为初步聚类。

(3) 属性权重的分配。属性的权重表明一个属性在决定两条记录相似性中的重要程度。根据属性在决定两条记录相似性中重要程度的不同，为每个属性分配不同的权重。由于记录中不同属性对反映记录特征的贡献是不平均的，因此在衡量两条记录的相似度时，不同的属性应赋予不同的权重。

2．检测重复记录。从大型数据集中检测并消除重复记录，首要的问题就是如何判断两条记录是否是重复的。因此就需要考虑到记录匹配问题，其中计算字段相似度，也称为字段的匹配问题是其核心。记录匹配，就是在利用字段匹配算法计算比较记录的各对应字段的相似度，然后再进行记录的匹配。如果根据字段的权重，进行加权平均，计算记录的相似度，若两个记录的相似度超过了某一阈值，那么认为这两条记录是匹配的，也就是重复的，反之，就认为是指向不同实

体的记录（即不是重复记录）。经常用来解决字段匹配的算法也比较多，如基于编辑距离的字段匹配算法及缩写发现算法、Smith -Waterman 算法[26] 、递归的字段匹配算法、基本的字段匹配算法等。

3．数据库级的重复记录的聚类。检测数据库中完全重复记录的标准方法是排序-合并方法是。排序-合并方法的基本思想是，先对数据集进行排序，然后对其相邻的记录进行比较，看其是否相等。目前已有的检测重复记录的方法也大多是建立在此思想基础上的，这一方法也为在整个数据库级上重复记录的检测提供了思路基础。在数据库级应用检测重复记录的算法对整个数据集中的重复记录进行聚类。最可靠的检测重复记录的办法是比较数据仓库中每对记录，但该算法时间复杂度太大，需要N(N-1)/2次比较，其中N是数据仓库中记录的总数。

4．冲突处理。在完成数据库级的重复记录检测，对重复记录进行聚类之后。为了使数据库中每条记录都表示不同的实体信息，得到准确的数据，根据一定的规则合并重复记录，或者删除被检测出的同一重复记录聚类中的重复记录，即就是保留其中正确的那条记录。也就是需要将检测出的重复记录合并，进行冲突处理。在整个重复记录清洗的流程中，数据库级的重复记录聚类和重复记录检测是其核心步骤。

2.3 ETL 过程应用

2.3.1 ETL 过程概述

ETL是构建数据仓库中极其重要的一个环节，它是指数据抽取(Extract)、数据转换(Transform)以及数据加载(Loading)等操作。ETL首先要做的是抽取(Extract)数据仓库所需要的数据，按业务需求从源数据(业务系统 / 外部数据等)中抽取，然后对抽取出来的数据转化为数据仓库要求的统一格式之后再进行必要的处理，因此对数据进行变换(Convert)、转换(Transform)、清洗(Cleaning)去除不必要信息；最后，将数据装载(Loading)到数据仓库中按照物理数据模型定义的数据结构类型。

2.3.2 ETL 过程架构

数据批处理抽取架构由抽取、清洗、转换和加载四个主要部分组成。同时，批处理抽取架构还应提供缓存与通用数据处理服务。当然，框架内的各个模块都可以根据实现时的具体业务需求加以调整。图 2-5给出了数据ETL 批处理抽取架构图。

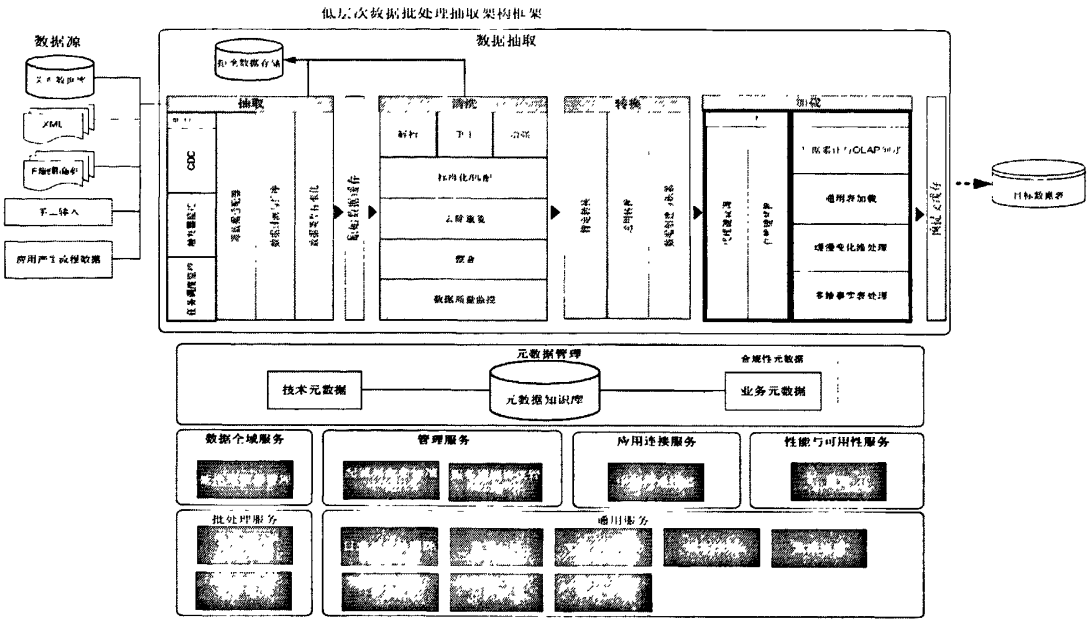

图2.3数据ETL批处理抽取架构图

2.3.3 ETL 功能定义

下面对ETL 功能架构中的抽取、清洗、转换、加载、缓存点、源数据管理服务、性能与可用性服务、批处理服务等方面对数据ETL功能框架中的关键服务进行阐述。

- 数据的抽取：抽取步骤负责把各业务系统数据源中的数据抽取到数据仓库中。抽取流程的目的是把数据从数据源抽出来，然后通过一组通用接口再把数据传送给数据抽取架构中的清洗与转换流程。由于各个不同的数据源内的数据及其质量可能各不相同，因此就有可能针对每个数据源，可能建立的抽取流程是各自独立的。
- 数据的清洗：数据清洗流程的设计可以划分为两类：一类是通用程序的设计，此类程序适用于各类业务源系统以及数据仓库内清洗需求，而另一类则是设计专用的清洗程序，它是针对特定业务源系统编写的。可以通过使用相关规则来改善和检查数据质量，来解决针对从源系统抽取数据时存在的不同程度的不一致性的问题。清洗流程负责对数据进行清洗与确认，使之与数据仓库所要求的数据标准与质量相符合。此外，还可以通过设计通用函数来解决一些多业务源系统的清洗规则和处理流程基本相同的问题，以此来提高其可重用性，因此这些关键的清洗功能就是通过这些通用函数，通过变化参数来满足不同业务源系统的清洗需求。
- 数据的转换：在这一步可以完成事实表的聚合以及关键绩效指标的计算。转换流程是数据在进入数据仓库前最后一次进行修改的环节。转换步骤

负责对源系统所抽出数据进行操作或放大。

- 数据的加载：这是 ETL 流程的最后一步。加载流程负责将数据加载到最终数据结构（这些数据结构包括是维度表、事实表或者事务表）中。代理管道是加载步骤中的主要组成部件之一，代理键管道的主要作用是将加载完成的数据表内的自然键替换成代理键。维度表的主键与外键仍然在代理键管道内得到保留。在完成加载结束以后，为了系统性能得到提升，一些约束条件将被去除而仅保留自然键。
- 缓存点：在 ETL 抽取过程中，缓存点的功能主要在于设定任务重启点以及分析数据前后沿袭关系。数据缓存既可以使用平面文件实现也可以使用数据表存放，但是通常并不直接镜像目标数据表的数据结构。
- 元数据管理服务：ETL 流程的实施关键在于设计合理的元数据使得系统能够清晰地定义 ETL 涉及的各个环节。数据抽取架构中主要包含技术和业务两类元数据。ETL 架构中的源数据管理服务必须与整个数据仓库的元数据管理服务协同一致，实现统一管理。
- 性能与可用性服务：负责监控数据抽取流程的执行统计数据和性能。性能与可用性服务能够帮助数据抽取流程满足服务级别协议的要求并且发现数据仓库与外部系统的依赖关系。
- 批处理服务：批处理服务使得 ETL 流程能够根据参数设定按批执行。但是由于批处理服务所涉及的各种系统监控机制相对复杂，因此最好使用成熟的套装工具构建。

2.3.4 ETL 过程中数据的清洗

ETL 过程中数据清洗是在集成转换过程中完成数据清洗工作，是指数据在离开应用程序后进入 ETL 平台时对数据进行的清洗操作。在数据仓库的应用中，ETL 所占的工作量占数据仓库创建过程工作量的多半以上。(一般的数据仓库系统中，ETL 中所涉及的转换往往在 75%以上)，因此在 ETL 过程中进行数据清洗是一项很有挑战性的工作.

(1) 关于 ETL 层的数据清洗模型[12]

通常情况下数据库作被作为 ETL 层中数据清洗的唯一的控制点。因为多种数据源的所有原始数据中大部分还没有被修改就被装载到数据的 ETL 工作层。不管是关系数据库还是非关系型数据集合的数据源中的数据都将被置于数据库表中，目的是为了在数据库内作进一步地转换工作，如图 2.4 所示 ETL 中的数据清洗模型。

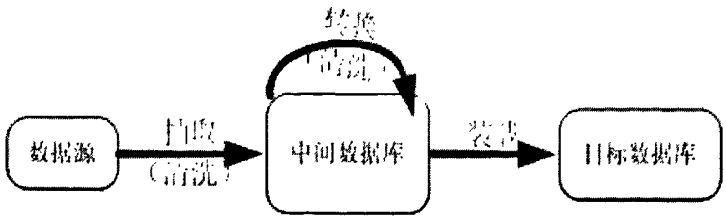

图2.4 ETL中的数据清洗模型

针对数据源是一个功能比较强的数据库管理系统，通常在数据抽取过程中可以直接使用SQL来完成数据清洗工作的一部分。但是如果是针对数据源是外部文件这样的数据源就不可能提供这种功能，所以只能直接将数据从数据源中抽取出来，然后在对数据进行转换的时候执行数据的清洗操作。因此，数据仓库中的数据清洗主要还是在数据转换的时候进行，它使用了数据库ETL处理方式中的数据库管理系统的转换清洗功能来完成大部分的清洗工作，这样数据清洗就充分利用数据库管理系统提供的功能。

(2) ETL 过程的数据清洗策略

按照数据清洗的实现方式与范围，传统的数据清洗一般包括一下四种策略；①自动清洗方式：这种方士是通过编写专门的应用程序检测、改正错误；②手工清洗方式：通过人工的方式直接修改脏数据；③正对特定应用领域的清洗：根据特定领域的数据特征采用一定的方法技术查找数值异常的记录，然后进行修正；④制定与特定应用领域无关的数据清洗：主要集中于重复记录的检测、删除。数据装载的不同环节其清洗任务不同，应为其选择不同的清洗策略，一般实际应用中的清洗策略都是以上一种或一种以上策略的德有机混合。

自动清洗方式能解决某些特定的问题，适合数据量大时使用。但同时也存在清洗过程不够灵活、反复的清洗过程导致清洗程序复杂，以及清洗过程变化时工作量很大等缺点。更为重要的是，可能存在某些潜在的错误类型在自动清洗过程中不能被发现和纠正，此时必然需要人为的参与。综合上述两种策略的优点，在 ETL 过程中把两种清洗策略[34] 相结合进行混合数据清洗将达到良好的效果。下面介绍一种混合清洗策略的流程如图 2.5 所示。

这种 ETL 过程中混合清洗策略主要以人为清洗扩展自动清洗，但以自动清洗为主。可以通过编写固定的应用程序来实现批量数据的自动清洗，在数据仓库阶段的数据初装阶段和增量数据追加阶段。但清洗模式并不能完全涵盖所有的错误类型，也不能反映语义上的正确性校正问题。因此，当某些错误类型无法按照已有模式来识别时，或者对于某些语义上不统一的数据，其修正工作就需要人工的确认和监督。此时，系统可设定异常报警功能，通过用户自身对错误的识别、理解和确认，最终实现数据清洗。

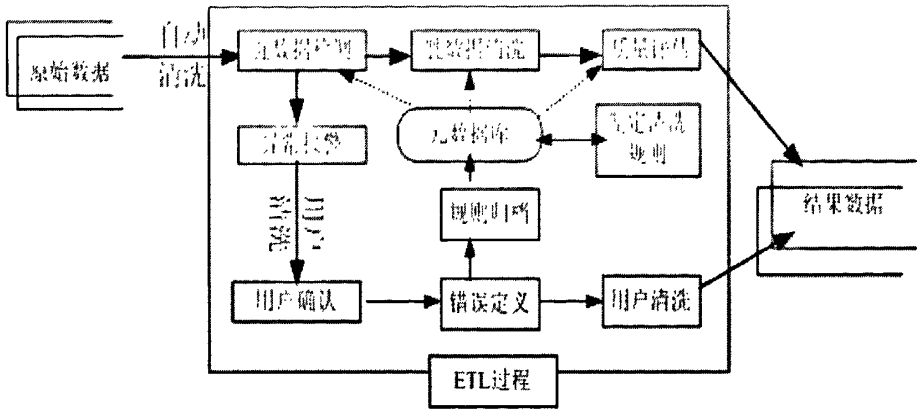

图2.5 ETL过程中混合清洗策略

2.4 本章小结

随着数据中心在企业管理系统中的深入应用，数据质量问题已经成为关系到数据仓库建设成败和决策支持系统能否提供正确决策的关键问题。数据清洗是保证数据质量的必要手段，目前仍属于较新的研究领域，本章主要对数据清洗的有关概念进行了描述，并且对 ETL 过程中的概念、架构、功能和有关清洗技术进行了介绍并介绍了一种 ETL 过程中的混合数据清洗策略，因为本文中的清洗工作主要是在数据的 ETL 过程中进行清洗的。

第 3 章　数据中心的建设

3.1 电力企业数据中心建设背景

3.1.1 “SG186 工程”简介

国家电网公司依据公司“十一五”信息发展规划，公司决定实施信息化建设工程（即“SG186 工程”）[52]，也就是在国家电网公司系统构筑由信息网络、数据中心、数据交换、应用集成、企业门户这五个部分有机组成的一体化企业级信息集成平台；其中包括建设由财务（资金）管理、安全生产管理、营销管理、人力资源管理、协同办公、项目管理、物资管理和综合管理八大业务应用系统；建立六个保障体系具体包括：健全信息化安全防护、管理调控、标准规范、技术研究、评价考核、人才队伍。“SG186 工程”的实施，重点是建设“一个系统、二级中心、三层应用”。其中一个系统指的是构筑一体化的企业级的信息系统，实现信息横向集成、纵向贯通，支撑集团化运作；二级中心指的是建设两级数据中心包括公司总部、网省公司的数据中心，实现数据资源的共享，促进集约化发展；三层应用则指的是部署和实现公司总部、网省公司、地市县公司三层业务应用的精细化全方位的管理，并且优化其业务流程。

我们要把握好“SG186 工程”实施的三个阶段中的每一步。第一步：2006～2007 年期间，开展平台及业务应用典型设计，试点先行，统一咨询，分步推广，实现初步集成；第二步：2008～2009 年期间，全面完成业务应用的推广，并基本上实现全面集成；第三步：在 2010 年，进一步完善提高，为初步建成“一强三优”现代公司提供坚强支撑。

“SG186 工程”的实现主要包括以下四个方面的目标。一是建成“横向集成、纵向贯通”的一体化企业级信息集成平台，并且能实现企业公司系统从上到下数据共享和信息畅通；二是要建成八大业务应用系统，这些业务应用必须是适应公司系统管理需求的，并且者八大业务系统是为了增强国家电网 公司系统各项业务的管理能力，提高公司的整体工作质量和效率；三是为了建设六个信息化保障体系，并且是健全规范有效的保障体系，它能够推动信息化快速、健康、可持续的发展；四是力争公司系统的信息化水平到“十一五”末期达到国内领先水平、在国际上达到先进，初步建成数字化电网、信息化企业。

一体化企业级信息集成平台的建设，其主要包括：

完善信息网络，加强基础建设。国家电网公司总部统一进行部署，制定标准

规范，分级建设能够覆盖公司总部、网省公司及公司直属单位以及地市县公司等的，以国家电网公司通信网络为基础的安全、可靠、快速、畅通的信息网络。

部署数据交换，畅通信息渠道。建立公司总部与网省公司统一的数据交换，这是由公司总部统一组织的，目的是实现公司关键业务数据的纵向快速交换，来确保企业上下信息畅通。

建设数据中心，实现数据共享。全面的把企业公司总部和网省公司的两级数据中心建设成。并且这是由公司总部统一组织，开展的典型设计和试点的，在这一部分结合数据中心的建设，来完善企业数据交换体系，并逐步实现数据中心间的数据点播和数据交换。

推进应用集成，促进流程优化。制定标准，设计架构，要按照先易后难的原则，结合业务流程的梳理，在公司总部和有关网省公司为了实现营销、财务等关键业务应用的横向集成来开展应用集成试点。之后要逐步扩大试点范围，整合资金流、物流和信息流，集成主要业务应用，实现“三流合一”，此外还要促进企业级信息系统建设，优化企业资源配置。

搭建企业门户，统一展示内容。把公司系统统一域名系统建成，来统一用户身份管理，规范业务应用界面风格。为了实现信息系统安全、统一的入口，建立公司总部和网省公司两层企业门户，这主要是为了企业范围内信息资源的综合展现，并可根据业务需要实现门户级联，定制展现内容。

八大业务应用主要包括：

财务（资金）管理业务应用。建立财务（资金）管理业务应用，统一财务标准，建设会计核算、资金管理、预算管理、分析与评价、资产管理、等核心应用模块，实现财务管理现代化提供支撑，主要是提升财务的主动控制和决策支持能力，实现公司财务信息实时共享。

营销管理业务应用。建立营销管理业务应用，服务创新和业务流程优化、推动营销管理创新。营销系统是覆盖公司总部、网省公司及基层供电公司的，加大对营销各项管理指标、服务指标的控制力、经营指标，整合服务资源。实现电能信息的自动采集、购售电环节的统一管理。实现对三级电力市场建设与运营的集约化管理。公司系统营销经营的实时分析，实现业务处理自动化、客户服务信息化、质量管理可控化、市场响应快速化、和决策支持前瞻化。客户关系管理和营销辅助决策分析和统一组织电力需求侧管理、等模块的试点和推广。

安全生产管理业务应用。建立以提高输变电设施和供电可靠性为目标，以设备管理和运行管理为基础，在检修单位定位的基础上，明确各级生产管理部门和运行，覆盖公司系统各级单位输电、调度、配电、变电全过程的安全生产管理业务应用。为了提高信息采集自动化水平及处理的实时性，必须要实现生产运营的全面信息化。完善运行、检修、技术监督、技术改造、电力设施保护管理、节能

降耗、等生产业务模块，来实现企业生产自动化系统与管理信息系统的有机结合。统一开发并推广公司系统可靠性管理模块。建立跨区电网输变电建设与运行监管模块。开发设备状态评估专家系统等高级应用，通过统一研究整合安全生产管理相关模块。

协同办公业务应用。建设综合协调、信息调研、督查督办、档案、公文、会议、通讯录、值班等业务应用模块，利用电子邮件、移动办公、即时通讯、远程视频等协同工具，提高公司综合系统管理水平。来为公司系统全面运用办公业务应用，增强协同工作能力。为了推广网络直播会议和电子签章模块，统一组织综合协调、信息调研、会议、档案、通讯录等模块的开发和实施。为了提高工作效率，逐步开展对办公业务信息的数据挖掘。

人力资源管理业务应用。实现省公司范围内的数据集中、业务整合，建立公司系统统一、集中、完整的人力资源基础信息库。建设人力资源管理业务应用，这还要基于一体化企业级信息集成平台。为了实现人力资源成本分析与战略人力资源管理，公司总部建成智能分析和人力资源查询统计模块。此外还统一组织开发和推广培训课件制作工具和远程培训模块。

物资管理业务应用。为了实现公司范围内的专家信息和招投标信息以及供应商信息的共享，完善物资管理业务应用，并构建公司系统统一的招投标管理模块。为了实现物资采购、结算和储运等全过程信息的统一管理，建设电子商务功能，。

项目管理业务应用。成熟的现代项目管理软件相结合，并统一组织开发、试点和推广覆盖项目管理业务应用的项目管理各环节。对检修、基建、技改等项目中涉及的实施、立项、监理、计划、控制等全过程以及进度、成本、质量、物资采购合同、安全等重要信息流进行有效控制和综合处理。

综合管理业务应用。为了统一组织开发计划、规划与统计应用，建设基于一体化企业级信息集成平台。实现综合计划的基础数据采集、资源共享的信息化，计划指标编制与，确保计划指标的在控、可控；按照“上级规划指导下级规划”的原则，实现规划数据滚动更新、统一、专家决策支持；为了进行造价分析、投资决策分析、项目后评价等，快速收集公司系统的投资计划信息和项目信息；为了有效整合各类统计信息资源，实现统计数据采集、传输、汇总全过程自动化拓展，使公司系统统计信息管理模块功能和应用范围。建设统一的金融信息业务应用，实现对控股金融机构业务的在线防范，监控、控制金融风险；实现公司控股金融机构的资源和信息共享，发挥协同效应，促进公司金融业务和谐、健康的发展。。整合公司系统情报资源，构建以总部为核心的公共信息资源共享体系，完善公共信息资源共享模块。完善创一流同行业对标管理等模块建。设领导综合查询与分析等辅助决策模块。统一开发和推广法律事务、审计管理、信访管理、国际合作等模块。建设整合纪检监察理、政工管、工会管理等模块。

六个保障体系包括：

安全防护体系。建设网络信任体系和应急机制。落实安全评估制度和等级保护制度。网络与信息安全全面纳入安全生产体系。建设从设备安全到用户安全的整体安全防护体系，建立多层次的网络与信息安全防御系统，实现网络与信息安全的能控、可控、在控。

标准规范体系。为实现信息系统的标准化开发、设计、运维与管理奠定基础，完成信息系统设计、规划、建设、运行维护、管理安全、评价等方面的标准。为了实现标准的动态维护、管理与查询，建立公司信息化标准体系库。

管理调控体系。建立信息化管理调控体系，健全公司系统一体化的信息化管理机制。信息职能管理部门归口管理信息化资金和项目。完善信息职能管理部门和业务部门的分工协调机制。加强信息系统的全生命周期管理。统一公司总部和各单位“186”信息客户服务电话，并设置服务监督电话，不断提高公司系统信息服务的质量和水平。

评价考核体系。按照公司创一流同业对标工作的总体要求，推广信息化最佳业务实践，丰富对标内容和形式，积极开展信息化同业对标，完善评价指标库，将信息化工作逐步纳入公司的业绩考核体系，实现信息化水平持续提升，形成激励机制。

技术研究体系。按照公司“十一五”科技发展规划的总体要求，形成一批具有公司自主知识产权的科研成果，完善试验设施与技术手段，加强信息技术研究，推广先进适用技术，增强公司系统信息技术自主创新能力，提高公司系统信息化的整体技术水平。

人才队伍体系。结合公司“1551”人才培养计划，按照公司“十一五”教育培训规划的总体要求，培养专家和管理咨询人才，建设信息化管理、建设、运行维护和应用四支队伍；建立在岗信息技术人员培训制度，加强信息技术及应用交流，培养复合型高级人才。

3.1.2 数据中心建设目标

数据中心项目总体思路，应按国家电网公司“SG186 工程”进行数据抽取和相应建模工作，完善领导查询系统和相关业务主题域的建设。总体建设目标就是通过数据中心的建设实现对企业数据资源进行集中存储，统一管理，为省公司各个应用系统提供数据层集中服务的数据环境；通过统一的数据定义和命名规范，保证数据的唯一性、准确性、完整性、规范性和时效性，实现数据的共享共用；通过对数据中心数据的挖掘展现，向公司领导提供辅助决策支持，实现数据信息的最大价值。其的具体建设目标如下[54]：

1. 通过对数据的规范化定义，实现数据的唯一性、完整性、准确性、规范性和时效性，实现数据的共享共用，解决数据层面的信息孤岛问题；为了实现统一数据元素标准、统一信息资源层次体系和统一信息编码，建立本部及所属各单位整体数据模型；

2. 为了管理决策层提供有效的数据服务，建立电力公司本部数据仓库和数据集市；

3. 为公司各级领导提供灵活自由的数据查询和报表生成手段，建设领导查询系统；

4. 建立统一的电力公司 ODS、企业级的，满足跨业务系统的综合查询功能；

5. 利用数据交换平台，实现与国家电网公司数据中心的级联，实现与国家电网公司纵向数据交换；

6. 将领导查询系统和商业智能部分集成到企业门户中，关键指标查询可以通过企业门户进行。

3.1.3 数据中心建设理论依据

目前数据中心被定义为广义数据中心和狭义数据中心两种概念，电力企业数据中心的建设就是以此为基础理论依据的。

广义的数据中心是指：企业的数据资源与业务系统进行集成、集中、分析、共享的场地、流程、工具等的有机组合。数据中心的核心内容包括业务系统、ODS数据库、数据ETL、数据集市、数据仓库、商务智能等，除此之外也包括物理的运行环境（也就是数据中心机房）和运行维护管理服务。广义的数据中心具体的来说应至少包含以下四个方面的含义：

首先，数据中心是提供所有应用系统的运营场所。这些应用系统包括集中的业务应用系统、应用集成平台、数据交换平台。

其次，数据中心也是容纳基础设施的物理地点，这些物理点是用以支持应用系统运行的。这些物理点包括服务器、网络、存储设备。

再次，数据中心就是本身的ODS数据库、数据仓库及建立在其上的决策分析的应用。

最后，为了保证应用系统高效地不间断运行，数据被正确的访问，数据中心需要有一套成熟的运行、维护体系来支持其日常的运行。

狭义数据中心的概念：狭义的概念是相对于广义而言的，狭义的数据中心是指数据仓库和建立在数据仓库之上的决策分析应用，这些应用具体包括：数据源、ODS数据库、数据的ETL、数据仓库、元数据管理和商务智能应用等。下面分别定义狭义的数据中心中各部分：

数据源：数据源是交易处理型与操作型业务应用系统（如人力资源系统、营销系统、财务系统等）内存放和收集的数据集合，是存放为满足最终用户需求的各类信息的源头，这些信息源是被迁移到数据仓库内的，而网省数据中心的数据源包括网省业务系统以及网省内地市供电公司上传的业务数据。

数据的ETL：ETL是用于将ODS区的数据迁移到数据仓库或将数据源整合到ODS的工作，主要提供数据移动控制以及数据转储及其相关的各种流程与服务。ETL的服务主要由以下几部分组成：任务调度、批量文件控制、错误处理、错异常处理、文件与数据传输、验证与审核。

ODS数据库：（也叫做操作型数据存贮）：ODS是数据仓库架构中重要的功能区域，它集成了来自不同业务系统的数据。ODS区的主要作用是（1）为终端用户提供一致的企业数据集成视图。（2）作为业务系统间的数据共享区，通过EAI系统集成平台使各个业务系统能够从ODS访问到它们想到的数据，避免各系统间直接进行数据交换。ODS区存放数据仓库所在层次各个业务系统间需要共享的数据，（3）为了满足用户对准实时性数据的分析需，同时ODS还能够以更高的性能生成操作型报表，生成集成的、近实时的数据存储来使得最终的用户能快速查询到近期细节生产的数据。

数据仓库区[50]：是数据仓库架构中最核心的数据存储区域，数据仓库区包含一个企业级的、相对稳定的数据仓库数据模型，用以支撑大部分的数据应用。这些数据主要是历史信息与企业级数据，并且这些数据在线存储的周期一般较长。在进入数据仓库之前，这些数据必须按照业务数据主题域进行整合。数据仓库区中的数据来源于ODS或者业务源数据。数据仓库的数据模型应该满足第三范式。

元数据[51]：在数据仓库系统中，元数据可以帮助数据仓库的开发人员和数据仓库管理员非常方便地找到他们所关心的数据；元数据按照传统的定义（Metadata）是关于数据的数据。元数据分为技术元数据（Technical Metadata）和业务元数据（Business Metadata）元数据是描述数据仓库内数据的结构和建立方法的数据，元数据主要包括技术元数据与业务元数据，其中技术元数据是用于开发和管理数据仓库使用的数据，是存储关于数据仓库系统技术细节的数据。业务元数据提供了介于使用者和实际系统之间的语义层，使得不懂计算机技术的业务人员也能够“读懂”数据仓库中的数据。

商务智能应用：最终用户通过前端分析、访问工具将数据仓库内的数据展现出来，达到最终为邻导决策层提供可靠的数据分析的最终目的。

3.2 数据中心系统架构设计

3.2.1 逻辑架构

根据国家电网关于数据中心典型设计的定义，数据中心总体架构分为：数据架构、应用架构、数据仓库执行架构、基础架构、运维机构以及安全架构6个部分[51]。

■ 数据架构：数据架构是指每个应用系统模块的相互关系、数据构成和存储方式，包括数据的管控手段和数据标准等。

■ 应用架构： 指数据中心所支撑的所有的应用系统部署和它们之间的关系。在本设计中，将重点描述数据仓库应用，对于业务应用系统的设计不属于本设计范围。数据仓库的应用架构通过后面章节中的应用功能设计完成。

■ 数据仓库执行架构：主要包括ETL 架构和数据访问架构，是指数据仓库在运行时态的服务流程及关键功能。

■ 基础架构（物理架构）：主要包括服务器、网络、存储等硬件设施，它是为上层的应用系统提供硬件支撑的平台。

■ 运维架构：为整个信息系统搭建了一个统一的管理平台，并提供了相关的管理维护工具，运维架构用于管理执行架构和开发架构，它主要是面向企业的信息系统管理人员，如系统管理平台、数据备份工具和相关的管理流程。

■ 安全架构：是指提供系统软硬件方面整体安全性的所有服务和技术工具的总和。安全架构覆盖数据中心各个部分，包括数据、运维、基础设施、应用等。安全架构不做专门描述。

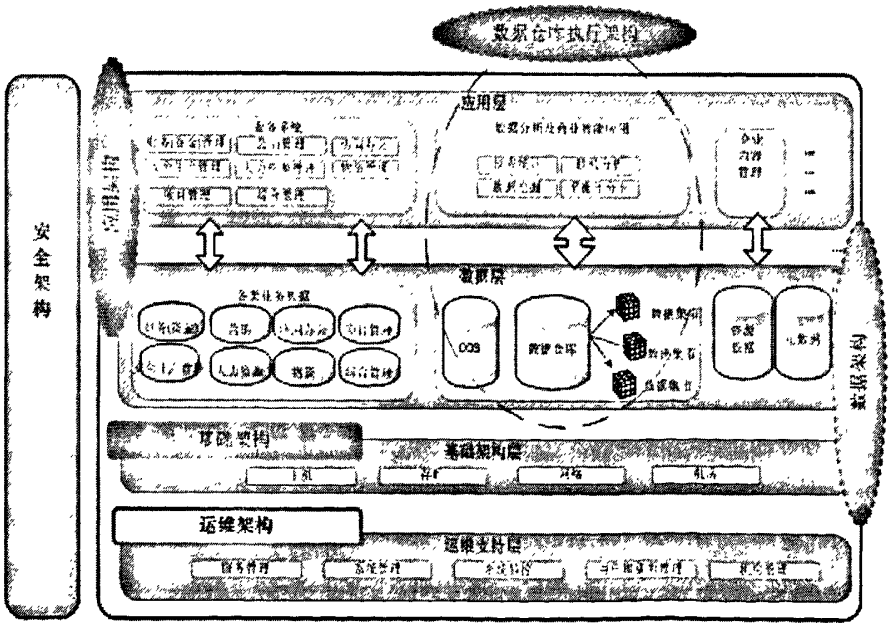

图 3.1 逻辑架构

3.2.2 数据架构

电力企业数据架构设计总体分为两种模式：即“分布式业务部署模式”和“集中式业务部署模式”。

1）分布式业务部署模式

在分布式业务部署模式中，主要针对电力企业中部分业务系统部属在地市、部分业务系统部署在网省公司的业务分部情况，因此其数据按照行政区域划分总体分布在总部、网省和地市三个层次内（如图 3.2 所示）：

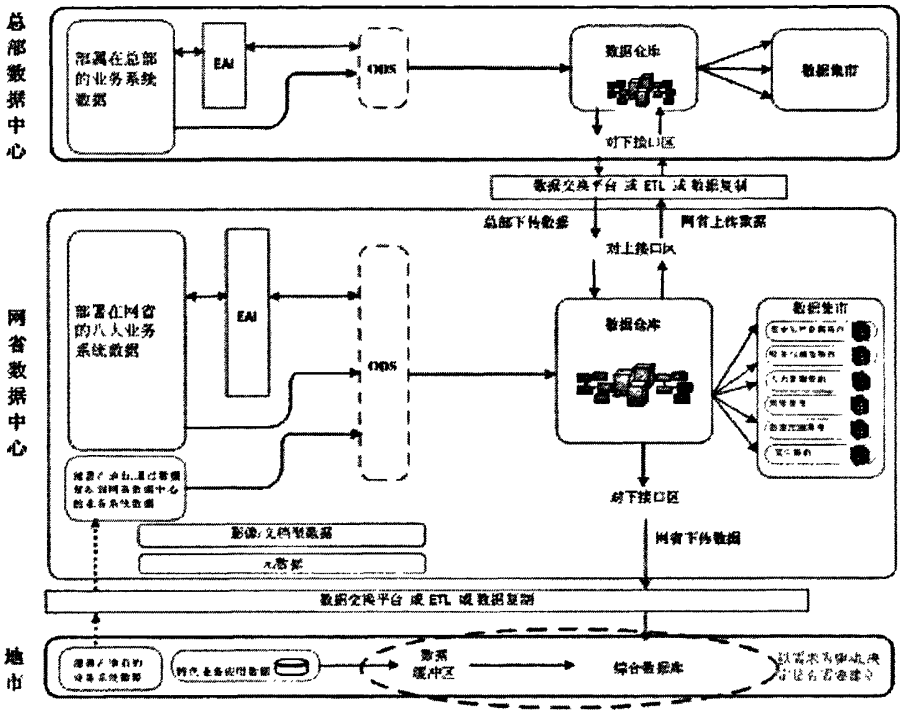

图 3.2 分布式业务部署模式数据架构图

说明：

- 总部数据仓库和网省数据仓库间通过数据交换平台进行数据交换；
- 地市业务系统数据通过数据复制上传到网省数据中心，再进入网省数据仓库；
- 地市综合数据库从网省公司数据仓库下载数据以供特色业务分析使用。

2）集中式业务部署模式

集中式业务部署模式适用于全部业务系都集中部属在国网总部和网省公司，因此其数据分布在总部、网省两个层次内；并且，总部数据仓库和网省数据仓库间可以通过数据交换平台、ETL 或数据复制的方式实现两级数据仓库间的数据移动（如图 3.3 所示）：

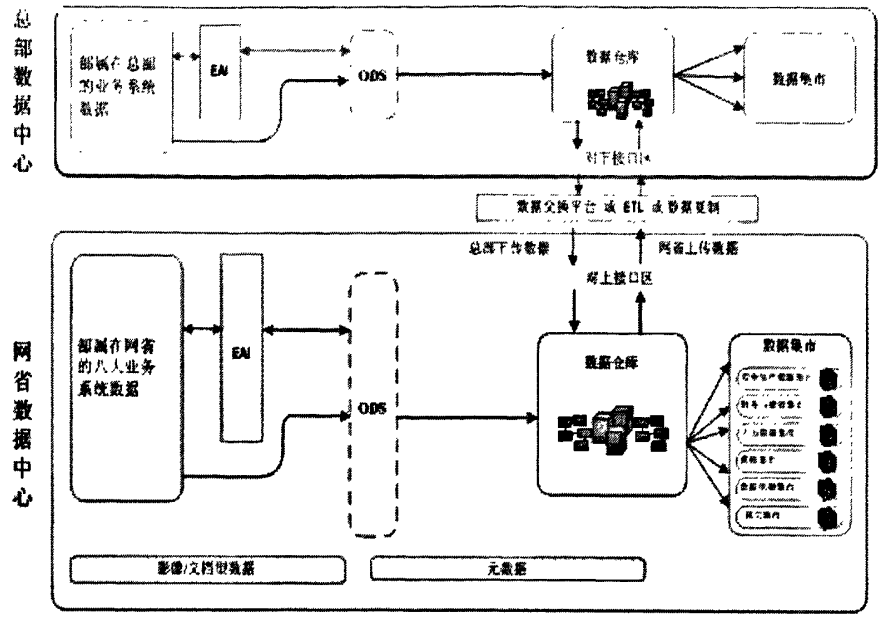

图 3.3 典型模式 I 数据架构图

3） 二者主要区别

“集中式业务部署模式”和“分布式业务部署模式”在功能上是没有区别的，主要是架构上存在以下区别如下表4.1：

表 3.1 典型模式 II 和典型模式 I 数据架构间主要区别

	集中式业务部署模式	分布式业务部署模式
行政区域	网省集中	网省和地市部署
数据仓库数据源	部署在网省的业务数据 总部数据仓库数据下载	部署在网省的业务数据 部署在地市的业务数据，通过数据复制上传至网省 总部数据仓库数据下载
网省对下接口区	无	提供数据仓库数据下载给地市综合数据库
地市综合数据库	无	根据需求驱动，确定是否建立

3.2.3 数据仓库执行架构

执行架构用于规范和定义数据仓库运行时态的功能流程。图3.4给出了国网公司数据仓库总体架构设计图：

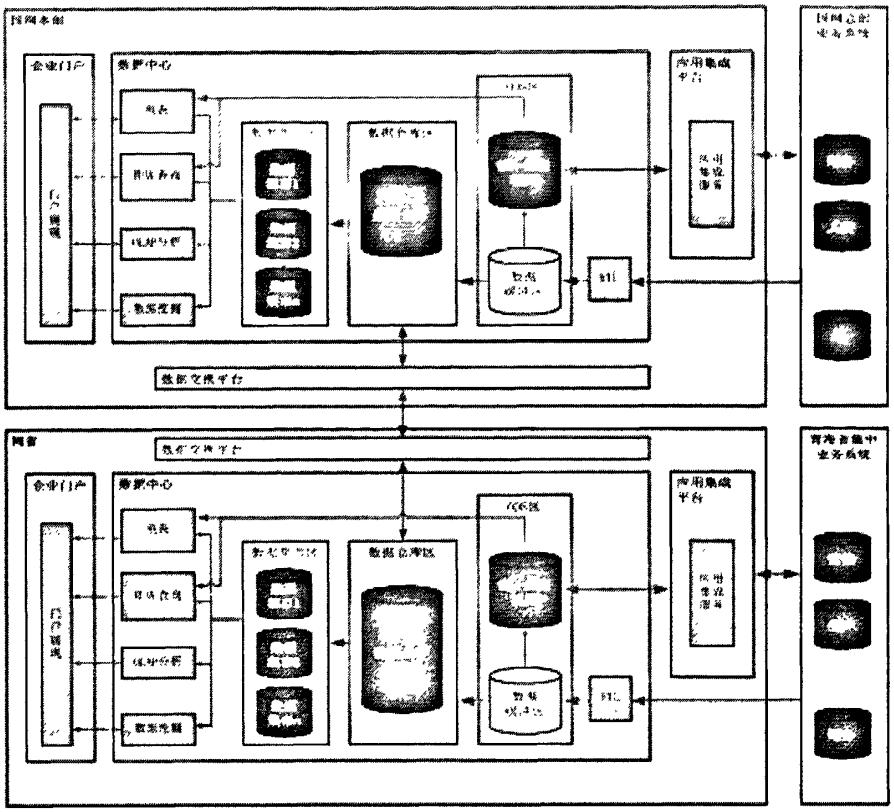

图 3.4 总体架构设计图

根据国网公司“SG186 工程”的整体规划，数据中心由以下几个部分构成：

■ ETL:ETL是用于将ODS区的数据迁移到数据仓库或将数据源整合到ODS的工作，主要提供数据移动控制以及数据转储及其相关的各种流程与服务。ETL的服务主要由以下几部分组成：任务调度、批量文件控制、错误处理、错异常处理、文件与数据传输、验证与审核

ODS区： ODS区的主要作用是（1）为终端用户提供一致的企业数据集成视图。（2）作为业务系统间的数据共享区，通过EAI系统集成平台使各个业务系统能够从ODS访问到它们想到的数据，避免各系统间直接进行数据交换。ODS区存放数据仓库所在层次各个业务系统间需要共享的数据， (3) 为了满足用户对准实时性数据的分析需，同时ODS还能够以更高的性能生成操作型报表，生成集成的、近实时的数据存储来使得最终的用户能快速查询到近期细节生产的数据。

数据仓库区：数据仓库区是电力公司数据中心架构中最核心的数据存储区域，它包含一个相对稳定的、企业级的数据仓库数据模型，支撑大部分的数据应用。数据仓库区内的数据，数据粒度与ODS一致或粗于ODS区，这些数据是按照主题存放的。在进入数据仓库之前，这些数据必须按照业务数据主题域进行整合。

■ 数据集市区：是一组针对某个主题域、特定的、部门或用户分类的数据集合。数据集集市区的数据需要针对用户的快速访问和数据输出进行优化，这种

优化的方式可以通过对数据结构进行索引和汇总。数据集市的数据按分析主题进行存放和组织，这些数据模型是星型结构的。然而数据仓库的数据是按照主题存放和组织，这里的数据模型满足第三范式。从数据仓库到数据集市的数据移动包括维表的映射、实体表向事实表，以及实体间关系到多维关系的映射，应重点考虑从规范化建模到多维建模的映射。数据仓库中所存放的数据需要经过累计汇总转换后才能加载到数据集市中，因为数据集市内需要按照地区、日期、类别等维度对数据进行汇总。

■ BI 展现：具体包括：即席查询，预定义报表，OLAP 分析等。它是通过前端分析、访问工具然后将数据集市内的数据展现给最终的用户。

■ 两级数据中心的级联：它是通过数据交换平台进行级联的。

3.3 本章小结

在本章主要介绍了某电力系统数据中心的建设背景、建设目标以及建设的理论依据，然后又分别逐个阐述了数据中心的功能架构其中包括逻辑架构、数据架构、基础架构、数据仓库执行架构、运维架构、安全架构。之后又分别介绍了数据中心的应用功能，其数据中心的应用功能主要包括：业务系统主题分析和指标查询等。数据中心系统的建设过程和方法不同于传统的操作型处理系统的过程和方法，数据中心系统建设最突出难点之一就是如何保证数据质量，使得数据准确可信。数据质量问题是客观存在的，数据质量问题的管控工作将贯穿数据中心系统建设的整个过程。数据中心系统应用来源于用户需求，来源于开发商的商业理解，应用的开发和完善也受到数据质量的制约。

因此，数据中心系统建设需要实现数据和应用的互动，不断的提升系统的数据质量，展现真实数据。从数据中心的建设、实施过程来看，它本身修复数据以提高数据质量的能力并不是很强，但是它能发现相关系统存在的一些数据质量问题从而提醒用户哪些数据有质量问题，将数据问题反馈到业务部门、业务系统，由后者做数据修正。

第 4 章　电量数据清洗技术

4.1 应用背景

电量数据从采集到最终的保存大致经历如下步骤[13]，首先是脉冲电度表（通过电流脉冲进行计量）计量并存储电量，然后通过 RS485/RS232 口将电量传输值前置采集设备，其次通过网络或 4 线/2 线 MODEM 传输之计量系统主站，最后根据特的定 ETL（抽取、清洗、转换、加载）抽取规则传输至数据中心。具体如图 4.1：

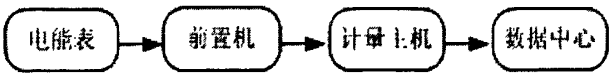

图4.1电量数据采集过程

从上图可以看到电量数据从电度表到数据中心大致有四个环节，三个传输过程，如此可将电量数据缺失、异常值产生的原因归纳为以下几个方面[9]：

■ 通信原因引起的电量数据丢失或异常

通信干扰或故障是实际生产过程中遇到的最常见的故障点，主要产生于“电能表-前置机”部分，由于电能表及前置机之间常置于强电干扰之下，加之他们之间使用的 RS485/RS232 通信规约自身的抗干扰能力及数据校验功能较弱，致使在传输数据时常常会有异常数据（这里专指数据突变）及电量采集不到的情况出现。

■ 电表更换导致的“数据异常”

虽然电表更换是实际生产过程中的常规操作，但是电表更换的结果会导致同一线路的电量底度值（电度表读数）出现归零，在系统则中表现出电量数据不一致的“异常现象”，因此在后期电量计算过程中需要“特殊处理”。

■ 旁路代替造成的“线路停电”

在实际的生产过程中，正常运行的电网线路常常会遇到停电检修、故障处理或运行方式改变时，经常会遇到用旁路线路或备用线路替代现有线路送电的情况，此时该线路的电量底度值会出现“静止”的现象，会有“线路停电”的假象出现。

4.2 数据中心脏数据的处理方案

脏数据是指源系统中的数据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，以及在源系统中存在不规范的编码和含糊的业务逻辑的数据。

在本文中脏数据特指源业务系统不规范编码的数据。

数据在抽取过程中难免产生脏数据，如果不处理脏数据，则会影响数据的完整性、有效性。如果采用手工方式处理脏数据，在大量的数据面前是一件非常困难的事情。根据国家电网“SG186 工程”中对数据质量得要求，某电力系统数据中心提出了一种好的脏数据处理方案，这样可以在提高脏数据的处理效率和统一脏数据处理模式中起到事半功倍的效果。下图 4.2 所示脏数据处理步骤：

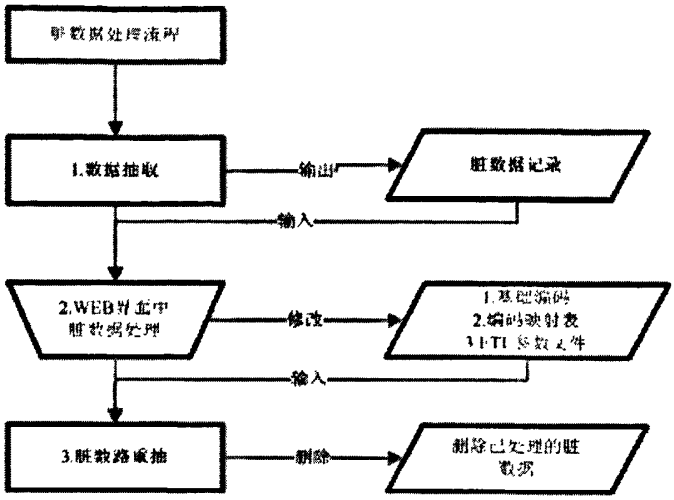

图 4.2 脏数据处理流程

①.数据抽取时产生脏数据，插入脏数据处理表中。

②.在 WEB 界面中对产生的脏数据进行处理，分别针对脏数据和非脏数据进行处理。

③.针对已经处理的脏数据进行重抽，并删除已处理的脏数据。

4.2.1 数据抽取

在数据抽取过程中所有的代码转换都通过代码转换表（CODEMAPPING）进行转换。当在代码转换表中找不到该代码时，该条数据即为脏数据。当确定数据为脏数据时需要将该信息记录在脏数据表（TB_SYS_ETL_DIRTYDATA）中。ETL 抽取脏数据处理流程图 4.3 如下：

①. 判断是否为脏数据，即在代码转换表中找不到新代码。

②. 判断该记录在脏数据表中是否存在。判断条件为代码种类 ID、业务系统 ID、源表名、源字段名，源 ID 全部相等。如果存在跳到第 4 步。

③. 将脏数据插入到脏数据表（TB_SYS_ETL_DIRTYDATA）中。数据插入后直接到第 6 步。

④. 新脏数据的时间戳的值与旧脏数据时间戳的值进行比较。如果大于则直

接跳到第 6 步。

⑤. 更新时间戳的值到脏数据表中。

⑥. 脏数据处理结束。

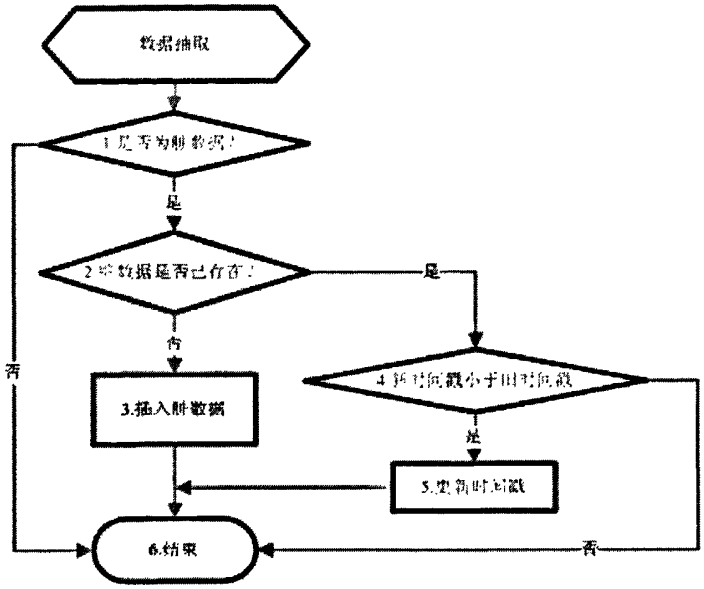

图 4.3 ETL 抽取脏数据处理流程

4.2.2 WEB 界面脏数据处理

脏数据通过 WEB 界面展现出，由脏数据管理人员进行处理决定。分为两种情况：第一种是脏数据确实为脏数据，则维护该数据抽取相关 MAPPING 的参数文件，将该脏数据排除。另一种为该代码为有效代码，此时需要在 ODS 库中相应的编码表中新建一个编码，并在代码映射表中插入相应的记录。

4.2.3 数据重抽

ETL 脏数据重抽流程图 4.4：

①. 将脏数据参数文件视图（VI_SYS_DIRTY_DATA_PARA）生成参数文件。按行号的顺序，生成到参数文件中，将列 1，列 2 的内容通过行列转换的方式输出到文本文件中。

②. 通过参数文件中的$$TABLES 变量得到是否有新的已处理脏数据。如果$$TABLES 中有内容则有脏数据，如果为空，则没有需要处理的脏数据。当没有已处理脏数据时跳回第 1 步。有脏数据则进入第 3 步。

③. 利用第一步所生成的参数文件中的$$WHERECOONDITION 变量对相应的脏数据源业务表进行重抽。在抽取过程中删除相应的脏数据记录。

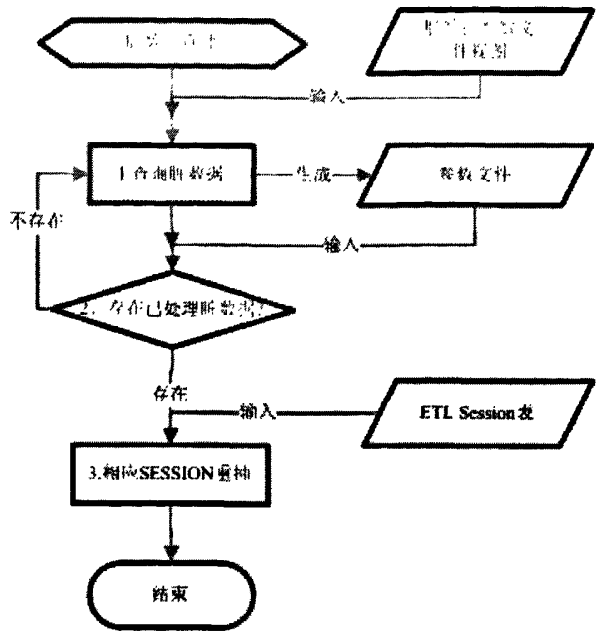

图 4.4 ETL 脏数据重抽流程

4.2.4 相关的数据字典

①. 代码映射表(CODEMAPPING)

表 4.1 代码映射表

字段名称	字段	数据类型	备注
代码种类 ID	DTYPEID	VARCHAR2(3)	来自代码种类表 CODETYPE
业务系统 ID	SYSTEMID	VARCHAR2(3)	来自代码表 CODECONTENT
源表名	SOURCETABLENAME	VARCHAR2(30)	原有业务系统源数据表
源字段名	SOURCEFIELDNAME	VARCHAR2(50)	原有业务系统源字段
源代码表名	SOURCECODETABLENAME	VARCHAR2(50)	原有业务系统对应的源代码表
源 ID 原有业务系统	SOURCEID	VARCHAR2(80)	源字段的代码 ID 值
源名称	SOURCENAME	VARCHAR2(80)	原有业务系统源字段对应的代码名
目标 ID	DESTINATIONID	VARCHAR2(36)	转换后的代码 ID 值
目标名称	DESTINATIONNAME	VARCHAR2(80)	转换后的代码名称
备注	REMARK	VARCHAR2(255)	一些代码说明

②. ETL 脏数据表(TB_SYS_ETL_DIRTYDATA)

表 4.2ETL 脏数据表

字段名称	字段	数据类型	备注
代码种类 ID	TYPEID	VARCHAR2(3)	来自代码种类表 CODETYPE
业务系统 ID	SYSTEMID	VARCHAR2(3)	来自代码表 CODECONTENT
源表名	SOURCETABLENAME	VARCHAR2(30)	原有业务系统源数据表
源字段名	SOURCEFIELDNAME	VARCHAR2(50)	原有业务系统源字段
源代码表名	SOURCECODETABLENAME	VARCHAR2(50)	原有业务系统对应的源代码表
源 ID	SOURCEID	VARCHAR2(80)	原有业务系统源字段的代码 ID 值
源名称	SOURCENAME	VARCHAR2(80)	原有业务系统源字段对应的代码名称
ETL 时间	ETLDATE	DATE	ETL 时间
处理标志	DONE_FLAG	CHAR(1)	0：未处理，1：有效数据，已处理，2：脏数据，已处理。
主键名称键	MAIN_KEY	VARCHAR2(60)	原业务系统中主要用于限制数据范围的字段名，一般为时间型主
主键值	KEY_VALUE	VARCHAR2(20)	原业务系统的主键字段的的值
主键类型	KEY_TYPE	CHAR(1)	0：日期，1：字符，2：数值
备注	REMARK	VARCHAR2(255)	一些代码说明

③. ETL SESSION 表(TB_SYS_ETL_SESS)

表 4.3 ETL SESSION 表

字段名称型	字段	数据类	备注
表名	TB_NAME	VARCHAR2(100)	原有业务系统源数据表
用途	PURPOSE	CHAR(1)	Session 的用途：0：全量，1：增量，2： 脏数据重抽，3：其它
SESSION 名称	SESS_NAME	VARCHAR2(100)	SESSION 名称
目录名	FLDR_NAME	VARCHAR2(100)	Session 所在目录
工作流名称	WFL_NAME	VARCHAR2(100)	调用 SESSION 的工作流名称
维护人员	OPRTR	VARCHAR2(20)	记录信息的维护人员
维护时间	OPRT_TIME	DATE	记录信息的维护时间
备注	REMAKR	VARCHAR2(255)	对维护进行说明

④. 脏数据参数文件(VI_SYS_DIRTY_DATA_PARA)，即当有脏数据产生时，该视图将有记录，可以生成脏数据所对应的参数文件。该参数文件可用于脏数据的重抽。

表 4.4 脏数据参数文件

字段名称	字段	数据类型	备注
行号	ROW_NUM	VARCHAR2(100)	用于参数文件内容排序，按行号从小到大的输出到文件中。
参数标题	COL1	COLB	各 Workflow 或者 session 参数的标题行
参数内容	COL2	COLB	各变量的名称及内容

4.3 电量数据的对象模型

4.3.1 电量底度值对象模型

电量底度值是电度表所反映的真实读数，也是电量计量系统的基础依据（本文选取日电量底度值作为基础的数据研究对象）。由于目前电力计量系统的基础电量数据是通过对电表每 15 分钟进行一次采集得到的，因此一块电度表每天就会有 96（4×24 小时）条电量底度记录，电量底度值的数据关系可用一个多元组表示如下：

$$M_{底度值} = \begin{Bmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,k} \\ . & . & . & . \\ . & . & . & . \\ a_{j,0} & a_{j,1} & \cdots & a_{j,k} \end{Bmatrix}$$

其中：$M_{底度值}$表示电表电量底度数据多元组，$a_{j,k}$表示各属性（正向有功电量、反向有功电量、正向无功电量、反向无功电量、总、峰、平、谷等常用属性列）的电量底度值，当 j=95 时表示数据量的统计区间为天/24 小时。

4.3.2 小时电量对象模型

小时电量数据是指在单位小时内的电量值，从严格意义上说应该为“统计型”电量数据，主要原因是它是通过电量底度值计算而来的，但是作为数据中心的基础电量值它又被划分为了“基础型”，我们仍以天为计算区间，则对应的小时电量数据的记录为 24 条（1×24 小时），可用多元组表示如下：

$$P_{小时电量}=\begin{cases} b_{0,0} & b_{0,1} & \cdots & b_{0,k} \\ . & . & . & . \\ . & . & . & . \\ b_{l,0} & b_{l,1} & .... & b_{l,k} \end{cases}$$

其中：$P_{小时电量}$代表小时电量数据多元组，$b_{l,k}$代表各属性（正向有功电量、反向有功电量、正向无功电量、反向无功电量、总、峰、平、谷等常用属性列）小时电量值，当l=24时表示数据量的统计区间为天/24小时，对于$M_{底度值}$中的j与$P_{小时电量}$中的l存在着j=4l的关系，例如：$b_{0,k}=a_{4,k}-a_{0,k}$以此类推可得到$b_{1,k}=a_{4,1}-a_{0,1}$。

通过上述对$M_{底度值}$与$P_{小时电量}$数学特征的了解，可以发现$M_{底度值}$与$P_{小时电量}$的联系非常紧密，如果想要得到可靠的$P_{小时电量}$不可避免的要对电量$M_{底度值}$中可能出现的数据缺失、异常等情况进行数据“清洗”的处理。

4.4 电量数据的清洗技术

依据国网数据中心典型设计[15]规定的“清洗”流程的的功能要求，某电力数据中心在实现电量抽取时将“清洗”流程划分为了两个子流程，即“检测”流程和“空缺值的填充”流程。结合电量数据清洗流程如下图3.2所示：

图4.5电量数据清洗流程

4.4.1 电量数据检测方法

根据上述底度值数据模型$M_{底度值}$，可知每个数据元$a_{j,k}$表示第j电表对应的第k个属性数据值，为了便于说明检测方法，我们令$P_i=a_{j,k}$，且P_i为等时序列。根据电量数据特点假设以下“异常值”条件成立：

$$P_k=\begin{cases} P_k=0,且P_{k-1}\geq P_{k+1},电表更换 \\ P_k=P_{k+1},且P_{k-1}\geq P_k,旁路代替/线路停电 \\ P_k=P_{k+1}=0或P_{k-2}\leq P_{k+2},电量丢失或异常 \end{cases}$$

本文拟采用切比雪夫定理识别“噪音数据”算法与上述假设条件相结合来判断识别空缺电量及异常电量。

切比雪夫定理识别[49]“噪音数据”的思路：首先，计算出个属性值的总体均值u和方差σ。然后根据切比雪夫定理得出一个判别的置性区间，同时这个置性区间的确立还要考虑到噪音的规模。最后根据这个置信区间来识别“噪音数据”。

根据上述思路，我们通过P_i的值计算得到日电量的置信区间，然后通过这个置信区间$[u-\xi\frac{\sigma}{\sqrt{k}},u+\xi\frac{\sigma}{\sqrt{k}}]$以及我们假定的来检测电量底度值中的“异常数据”。

在利用均值和方差得到置信区间时，由于电量数据的特点具有时序性，并且它是逐渐递增的，而电量数据的异常情况通常是出现较大的值，或者是出现不能采集到的情况就是数据不增的情况，还有就是出现极小的值。因此为了既能快捷，又能高效的检测出这些“噪音数据”。本文采用局部的均值和方差来计算，即就是在判断第 i 条记录时拟采用第i-1条记录中的属性值和i+1条记录的属性值来计算均值和方差。这样减少了计算的次数，也减少了对整个数据集的扫描过程，每次只记下检测记录属性的前一个属性值和后一个属性值。因此通过这种局部计算方差和均值来得到判断区间的方法进行异常检测是高效快捷的。

4.4.2 空缺值处理技术

对于检测出复合条件的属性值我们作了“置 NULL”处理，即统一将空缺、异常值的数据置未空值，以便于下一步进行空缺数据的预测填充。根据电量数据的缺失、异常情况我们做了如下技术分类：

■ 邻近数据“平均值法”[13]

当遇到电量数据丢失或异常时采用“平均值法”来填补空缺值，如果存在大量数据丢失或异常，则需要重新对电量数据进行补采，该部分数据处理应该在数据的采集环节就可弥补，现在假设丢失或异常的数据数允许范围之内，我们采用“平均值法”进行空缺值的填充。

因为电量数据中同一属性具有相似的特点，所以可利用空缺数据所在的属性的平均值预测电量数据中相应属性的空缺数据。这是一种简单的、工程经常采用的预测空缺数据的方法。

这是一种简单的、工程经常采用的预测空缺数据的方法。

对于$j\times k$样本数据$M_{底度值}$的空间，第i个属性中的空缺数据为$a_{j,i}$，则对第i处理个属性中空缺数据的预测公式$a_{j,i}$：

$$a_{j,i}=(a_{j-1,i}+a_{j+1,i})/2 \qquad (3.1)$$

式中，$1\le i\le k$；$a_{j,k}$表示第j个样本数据的第可i个属性值。

■ “基准值”法[9]

如果遇到线路更换电表，此时电表会“清零”（假设电表更换必需进行校表清零操作），因此当日电量底度值在更换电表后需自动累加原电表“基准值”，保持电量数据的一致性。此方法也是常见方法之一。

对于 $j\times k$ 样本数据 $M_{缺度值}$ 的空间，若从 $z\times w$ 个样本开始数据清零，则令前一个样本值为基准数据 a_i 值（表示本样本所有属性值），对于清零以后数据的预测公式 $a_{z,k}$ 如下所示：

$$a_{z,w}=a_i+a_{z,w} \quad （需循环处理） \tag{3.2}$$

式 3.10 中，$1\le z\le j$；$1\le w\le k$；$1\le i\le k$；$a_{j,k}$ 表示第 j 个样本数据的第 i 个属性值。

■ 基于神经网络的预测填充方法

这也是本课题中重点研究的一种处理电量数据空缺的方法。基于遗传神经网络的数据清洗模型在预测待填补值时，充分利用了遗传算法的全局寻优特性和神经网络的非线性映射能力，这种算法的预测精度比较高，是目前进行数据清洗的有效方法。遗传算法是一种基于自然遗传和自然选择的全局优化算法。它是训练前馈型神经网络的理想方法，该方法主要是采用从自然选择机理中抽象出来的选择、交叉、变异三种操作，用基本的遗传算子对参数编码进行操作，因此通过此方法可以实现全局最优搜索。下面就详细介绍该算法。

4.5 遗传神经网络算法概述

遗传算法是一种基于自然遗传和自然选择的全局优化算法。该算法采用从自然选择机理中抽象出来的选择、交叉、变异三种基本的遗传算子对参数编码进行操作，用以实现全局最优搜索，这是一种训练前馈型神经网络 的理想方法。遗传神经网络 模型的算法首先利用遗传算法找出网络参数的最优初始值，因为遗传算法具有容易发现最优解区域的特点，之后再利用 BP 算法搜索模型参数的最优解空间，因为 BP 算法具有寻优能力。

遗传神经网络 算法的基本思想[7] [27] [40]：遗传 BP 神经网络是通过 BP 神经网络的权值进行编码，随机产生种群，利用遗传算法对该种群进行进化，搜索到最优的网络权值。基于前馈型神经网络的数据预测模型由输入层、隐层、输出层构成。各层有若干个神经元节点，神经元间通过权值相连接，如图 3.3。在进行数据预测时，输出层的输出为待填补的数据值，输入层为缺值数据相关的属性值。神经网络 的全体网络权值整体内容反映了神经网络 模型对要解决问题的知识存储情况，可用矩阵 W 表示。神经网络是通过反复地学习训练样本，而不断地修改权值，来使的网络模型的输出逐渐地逼近期望值，若网络收敛，就是当网络的输出达到用户指定的精度后，此时网络中存储的权值反映了神经网络对待解决问题的知识存储。而后，可输入待填补数据相关的属性值，则网络的输出值就是待填补数据的估计值。

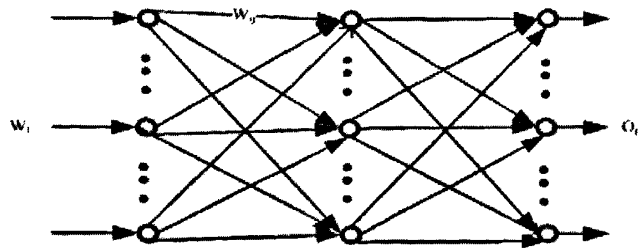

图4.6基于前馈型神经网络的数据预测模型

BP 神经网络 是由输入层， 隐藏层和输出层三部分组成。对于 N 个样本集合为离散的时间序列如：$\{(x(t),y(t))\,|\,x\in R^m,y\in R^n,t=1,2,...,N\}$，BP 网络完成可以从输入到输出的高度来进行非线性的映射， 即就是可以找到某种映射使得：$F:R^m\to R^n$. 在这里我们把全体样本分为训练样本 ϕ_1 和检测样本 ϕ_2：

$$\phi_1=\{(x(t),y(t))\,|\,x\in R^m,y\in R^n,t=1,2,...,N_1,N_{1\le}\le N\} \tag{3.3}$$

$$\phi_2=\{(x(t),y(t))\,|\,x\in R^m,y\in R^N,T=N_1+1,N_1+2,...,N\} \tag{3.4}$$

利用 ϕ_1 首先建立映射关系， 然后看网络对能否给 ϕ_2 给出正确的输入-输出，若是可以的话， 我们就认为该模型可以在实际生产中用于预测。 我们建立这样的时间序列映射，可以采用输入节点为 m 个， 输出节点为 n 个， 隐节点为 p 个的三层 BP 神经网络来实现。对于网络的输入与输出之间建立如下的关系：

$$\hat{y}_k(t)=\sum_{j-1}^{p}v_{jk}\bullet f\left[\sum_{i=1}^{m}wij\bullet x_i(t)+\theta_j\right]+r_k \tag{3.5}$$

其中 $f(x)=\dfrac{1}{1+e^{-x}},k=1,2,...,n,t=1,2,...,N_1$，$x_i$ 为网络的输入， $\hat{y}_k$ 为网络的输出，w_{jk} 为输入层 i 节点到输出层 j 节点的权值， v_{jk} 为隐层 j 节点到输出层 k 节点的权值， θ_j 为隐层 j 节点处的阀值，r_k 输出 k 节点的阀值， f 为激活函数. 设网络总的误差小于 e_1，则有：

$$E_1=\frac{1}{2}\sum_{t=1}^{N_1}\sum_{k=1}^{n}\left[y_k(t)-\hat{y_k}(t)\right]^2\le e_1 \tag{3.6}$$

设检测样本平均误差小于 e_2， 则有：

$$E_1=\frac{1}{N-N_1}\sum_{t=N_1}^{N}\sum_{k=1}^{n}\left[y_k(t)-\hat{y_k}(t)\right]^2\le e_2 \tag{3.7}$$

遗传神经网络算法的具体步骤如下[20]：

第一步，初始化。

采用用遗传算法来优化神经网络，其主要是优化神经网络中各神经元之间的

连接权值，来初始化种群 $P(t)$。本算法采用实数编码方案避免权重步进变化，因为网络的连接权是实数，编码方法就是把神经网络中的各个节点的权值形成一条染色体，按顺序排成一串。以神经网络的所有权值和阈值作为染色体的基因来进行编码的。网络隐含层转移函数为 Sigmoid 函数。

第二步，适应度函数的设计。

将训练样本逐个代入，计算出总的网络误差 E，由此计算出该染色体的适应度 f_i。因此就需要从种群中取一条染色体 i，把其中的基因（网络权值）按顺序依次带入到网络模型中。而适应度函数 f_i 的定义如下：

$$f_i = \frac{1}{1+E} \tag{3.8}$$

第三步，遗传操作。

首先输入神经网络的初始拓扑结构为：m-l-n。m 作为网络的输入层节点数，l 作为网络的隐层节点数，n 作为网络的输出层节点数。然后输入种群规模 $p(t)$，交叉概率 p_c，变异概率 p_m。为避免权值过小而使算法收敛太慢，因为遗传算法调整权值的能力较弱，在实验中我们使用[-3,3]上均匀分布的随机小数初始化种群，这些小数是十进制实数。$p=\{p_1,p_2,...,p_i\}$ 中的染色体设定选择概率 p_i.选择概率定义为：

$$p_i = \alpha*(1-\alpha)^{i-1} \tag{3.9}$$

α 是[0，1]中的随机数，一般情况下我们取 α=0.05，$i=1,2,...,t$。迭代判断是将通过把当前代中的染色体还原到网络模型中，来进行迭代判断，之后代入训练样本，来计算出该染色体的总误差 E 和适应度 f_i，然后判断网络误差是否达到最大迭代次数或指定的误差，如果是则迭代结束，否则的话继续进行遗传操作。下面将各染色体从高到低按适应度来排序。

选择操作是针对于每个染色体 i 的，计算其累积概率 q_i，q_i 的定义为：

$$q_i = \sum_{j=1}^{i} p_j \tag{3.10}$$

选择操作是使用轮盘赌算法，每轮产生一个随机数 r，若 $q_{i-1} < r <= q_i$，r 是[0，1]上均匀分布的，则此轮选中此染色体作父染色体。

交叉与变异。交叉是使用 t 次轮盘赌算法，获得 t 个父代染色体 $X_1,X_2,...,X_t$，把它们组成对，如：$(X_1,X_2),(X_2,X_3)$等。对每一对染色体，以交叉概率 p_c 确定交叉位置 k，互换两个父体在[1,k]编号间的基因，从而得到两个新的染色体 X_1',X_2'。变异是以变异概率 p_m 确定 u 个变异位置 $p_1,p_2,...,p_u$，将这些位置上的基因做变异操作，简单的做法是给原基因值加上一个[-1，1]间均匀分布的随机小数，从而得到两个新的子染色体 X_1'',X_2''。

第四步，产生新种群。

将新个体插入到种群 $P(t)$中，产生新的种群 $P(t+1)$，再把新种群个体的连接权赋予神经网络中，并计算新个体的适应度函数，若达到预定值 ε_{GA}，则进入下一步，否则继续进行遗传操作。

第五步， 再用 BP 算法训练网络权值。

达到所要求的性能指标或最大遗传代数后，将最终群体中的最优个体解码即可得到优化后的网络连接权系数。以 GA 遗传出的优化初值作为 BP 神经网络的初始权值，再用 BP 算法训练直到误差平方和达到指定精度或达到设定的最大迭代次数，算法结束。

4.6 实验分析

根据上述思路，将遗传神经网络算法应用到电量数据预测的具体方法如下：

首先对日电量底度值样本空间 j×k 样本数据 $M_{底度值}$ 的空间 P_k(令 $P_i=a_{j,k}$，且 k=96)中连续的4个数据进行分组，理想情况可分为23组（假设只有1组数据需要预测），然后从分组中尽量拟优选出时间点相近的6个分组，将其作为训练样本进行训练，如果分组达不到最终精确度，可根据试验增加分组数量。

其次假定 p_j 为缺失数据值，选取 $\{p_{j-3}, p_{j-2}, p_{j-1}, p_j, p_{j+1}, p_{j+2}, p_{j+3}\}$ 任意中连续且缺失数据仅有 p_j 的4个数据，然后将其作为输入层数据输入，最终得到 p_j 的估计数值。

“缺失、异常数据”的识别与空缺值的填补技术在国内外理论研究都比较成熟，但是真正实用化时受到很多条件制约，诸如算法的执行效率、算法的复杂度、所需系统资源的大小等因素，本文所采用的算法的优势在于简单可行且全部通过SQL存储过程就可以实现，所付出的代价自然是精度会有所偏差，但这种偏差是在系统允许范围之内的。

本文选取110KV、220KV、330KV线路某日正向有功电量数据作为研究对象，如下表4.5（部分）：

表4.5电量数据样表

电压的等级	采集点							
	0点	0点15分	0点30分	...	16点15分	16点30分	...	23点45分
110KV（负荷大）	55.59212	55.59924	55.60576	...	56.00396	56.01936	...	56.15484
220KV（负荷小）	63.57212	63.58132	63.5902	...	64.0486	64.05508	...	64.22924
330KV(负荷平)	32.20568	32.283	32.36176	...	37.33304	37.4176	...	39.90376

说明：1. 选取了底度值相近的三组数据线路数据，目的在于反映不同线路负荷下数据变化的量。2. 不同的电压线路等级，单位时间内电量数据的变化量

增量值不同。3. 本文选取的正向有功属于电量数据的一个属性域。4. 本文在110KV、220KV、330KV的线路的采集数据中分别把16点30分这个采集点的数据作为为缺失数据。

以下分别使用“平均值”法、“基准值”法、“遗传网络神经”法分别对缺失的16点30分的采集点的数据进行预测填充，这三个点的真实值分别是：在110KV线路上的采集的值是56.01936、220KV线路上的采集值64.05508、330KV线路上的采集值37.4176。

实验中利用遗传神经网络的清洗模型时，数据清洗模型拓扑结构为4-8-1形式，即4个输入节点，8个隐层节点，1个输出节点。遗传算法的初始种群规模t=500，交叉概率 $p_c=0.5$，变异概率 $p_m=0.03$，最大迭代次数为N=300。预测结果如表4.6所示：

表4.6预测结果对比

算法比较	数据源		
	110kv线路	220kv线路	330kv线路
平均值法	56.06428	64.0548	37.41292
基准值法	56.05936	64.0549	37.4154
遗传神经网络	56.01930	64.05505	37.4173
真实值	56.01936	64.05508	37.4176

从试验结果对比就可以得出利用遗传神经网络的预测方法所预测到的结果明显它的精确度比较高，都在99.5%以上。由于不同的电压线路等级，单位时间内电量数据的变化量增量值不同但是遗传神经网络算法对电压线的等级影响不明显。由于直接从表3.2中看到的数据不是很直观，下面我们就从220kv线路上取出12个采集点作为缺失值来用前面的三种方法分别进行了预测，在接下来的试验中用的参数还是在进行前面试验中的参数值：数据清洗模型拓扑结构为4-8-1形式，即4个输入节点，8个隐层节点，1个输出节点。遗传算法的初始种群规模t=500，交叉概率 $p_c=0.5$，变异概率 $p_m=0.03$，最大迭代次数为N=300。通过实验进行预测，由于采用平均值法和基准值法在进行电量预测时受线路电压等级、负荷高低的影响很大，因此如果应用于实际生产，它们预测的值的精确度很难掌握，误差也很难控制。相反，基于遗传神经网络的预测填充方法则对它们的依赖程度很低，并且在一个较稳定的精度范围之内（精度根据增加样本训练次数可控）。因此可以看出该种算法的优越性。具体的试验结果分析如下图4.7所示。

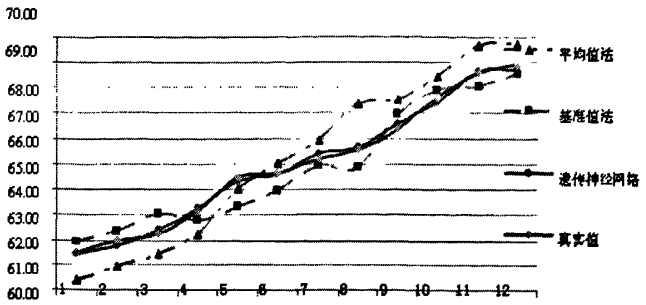

图4.7对220KV线路上数据预测结果对比

4.7 本章小结

本文中的整体清洗思路是对检测到的异常数据先删除，然后再用文中提到的清洗模型来预测填补空缺值，并且这种方法也是可行的。本文中提出的基于遗传神经网络的清洗模型充分利用了遗传算法的全局搜索能力和神经网络的非线性映射能力，在很大程度上提高了数据的预测精度，而且数据的预测精度是可控的，因此这是预测像电量数据这种精度要求高的数值型数据的理想方法，并且该算法全部通过 SQL 存储过程就可以实现。笔者的主要任务就是处理数据中心在处理电量数据时遇到的脏数据，根据本章中脏数据的处理方案，在 ETL 过程中进行数据抽取时遇到的脏数据，就用到了电量“噪音数据”的清洗方法，来提高数据的质量，以使的数据准确可信。

结　论

在了解数据清洗问题定义、特点、分类及其数据清洗的定义，环节、方法等基本概念后，对数据清洗的若干关键技术进行了深入细致地研究，在电量数据清洗问题上提出了作者的一些想法和见解，对一些关键问题提出了相关的解决方法和策略，并将电量数据的清洗方法应用于电力企业的数据中心中，达到了预期的研究效果，并对存在的问题也做出了一定的分析。

本文主要做了以下几个方面的工作：

(1) 根据电力电量数据特点，提出了相关电量数据的清洗技术方法，即将电量数据“清洗”流程划分为 “检测”和“空缺值的填充”两个子流程。并对应每个流程确定了不同的数学处理模型。其思想就是利用统计方法检测异常数据，之后把这个异常数值删除在利用预测模型进行预测填补，这样使得对空缺值的清洗和错误值得清洗合并到了一起。

(2)在空缺值的预测方面提出的基于遗传神经网络的清洗模型，充分利用了遗传算法的全局搜索能力和神经网络的非线性映射能力，在很大程度上提高了数据的预测精度。

(3) 介绍了电力系统数据中心的建设思路，根据电力系统脏数据的处理方案，并将电量数据的清洗技术应用到电力系统数据中心的建设中，提高了数据的质量和精度。

不足与展望：

数据清洗一直以来都是研究的热点问题，其中涉及许多复杂的概念和技术，还有许多问题需要进一步地研究。例如：

(1)当前数据清洗的研究重点主要放在了电量数值型数据空缺的清洗上，应用范围局限性很大，对于未来的数据清洗方法的研究应扩大数据类型的范围，由简单的数学预测发展为智能分析。

(2)在实际应用中对于大数据量的清洗工作，如果使用遗传基因算法，对单台计算机硬件系统的资源占用仍然非常的高。此外未来数据清洗算法应考虑网格化的云计算运行模式，来降低算法处理时间。

参考文献

[1] W.H.Inmon.数据仓库[M].北京机械工业出版社，2002

[2] W.H.Inmon1 Building the Data Warehouse[M]Wiley Computer Publishing 1992. 323-344

[3] 甘肃电力数据中心建设技术规范书

[4] K. Lakshminarayan, S.A. Harp, R. Goldman, and T. Samad."Imputation of missing data using machine learning techniques," [J].Proceedings: Second International Conference on Knowledge Discovery and Data Mining, pp. 140–145, 1996.

[5] 辅祥,刘云超，段智华．数据清理综述[J]．计算机应用研究，2002，29(3)：3-5.

[6] 王石，李玉忱，在属性级别上处理噪声数据的数据清洗算法[J]。计算机工程第, 2005 ,31(9): 86-88, 227

[7] 覃华,苏一丹,李陶深.基于遗传神经网络的数据清洗方法[J].计算机工程与应用2004,45-46,67.

[8] 乔珠峰,田凤占,黄厚宽.缺失数据处理方法的比较研究[J]. 计算机研究与发展, 2006, 43(Suppl)：171～175.

[9] 杨德亮,范亮星,李春平.基于表码的电量数据处理方法[J].电力系统自动化. 2008,32(4),93-96。

[9] Fayyad, Mannila, Ramakrishnan .Outlier Detection and Data Cleaning in Multivariate Non-Normal Samples[J]: The PAELLA Algorithm.

[10] Kamakshi Lakshminarayan, Steven Alex Harp, Tariq Samad.Imputation of Missing Data in Industrial Databases[J] . Applied Intelligence 11: 259–275 (1999)

[11] Taghi M. Khoshgoftaar □Jason Van Hulse. Imputation techniques for multivariate missingness in software measurement data[J]. Software Qual (2008) 16:563–600

[12] 朱宝成. ETL框架及数据清洗的研究.[硕士学位论文] 北京工业大学2007

[13] 张文准, 黄志强. 电能量采集系统异常数据分析及处理.《自动化技术与应用》2008 年第27 卷第8 期

[14] 查峰. 数据仓库化中数据清洗问题的研究[D]. 南京:东南大学,2002.

[15] 王浩，电网公司数据中心的设计与领导查询系统的实现[硕士学位论文] 东

北大学2007

[16] 周小明，赵永彬，韦明，电力企业ERP运维数据中心的设计与应用[J]，电力信息化2010, 08(7)

[17] 韩凤军，王凯夫，基于神经网络-Markov状态模型的电力电量预测[J]，信息技术，2010, (11)

[18] 曹亮，李彤，李喜旺，基于企业ERP数据中心级联的数据交换平台[J]，计算机系统应用2010, 19(11)

[19] 李隶荣. 一种基于数理统计的入侵检测算法研究［J］-微计算机信息 年2009第25卷第4. 3期 93-95

[20] Zhai LL,Zhang SC.The Feature Model of General ERP System for Discrete Manufacturing Industry.International Conference on Electronic Commerce and Business Intelligence,ECBI 2009.Beijing,2009.12-15.

[21] Kantz H. CHREIBERT.Nonlinear Time Series Analysis [J].Second edition.Cambridge University Press,2002.

[22] Zhao Y,Small M.Minimum Description Length Criterion for Modeling of Chaotic Attractors with Multilayer Perceptron Networks [J]. Transactions on Circuits and Systems 1，2006(53).

[23] 陈京民.数据仓库与数据挖掘技术[M]（第2版）电子工业出版社192-201

[24] 陈武，袁国忠翻译，企业数据仓库规划建立与实现[M]，人民邮电出版社

[25] Box M J . A New Method of Const rained Optimiza2 tion and a Comparison with Other Methods [J] .Computer Journal ,2002 ,15 (2) ,85290.

[26] Jiawei H, Micheline K. Data Mining: Concepts and Techniques[M].Copyright by Morgan Kaufmann Publishers, Inc., 2001,5:73-74

[27] 田旭光. 宋彤，刘宇新. 结合遗传算法优化神经网络的结构和参数[J]. 计算机应用与软件，2004，21(6)：69—71

[28] 田芳 ,刘震. 数据仓库清洗技术讨论-青海师范大学学报(自然科学版) ［J］2005 年第4 期 50-53

[29] 张燕. 基于聚类的数据清洗的研究与实现[硕士学位论文].保定 ：华北电力大学，2007

[30] Qiao Zhufeng，Tian Fengzhan. A Comparison Study of Missing Value Datasets Processing Methods［J］Journal of Computer Research and Development. 43(Suppl.)：171～175，2006

[31] 杨辅祥. 数据仓库下中文数据清理的研究与应用[D][硕士学位论文]. 上海：上海大学,2002

[32] 陈松，数据仓库中的数据质量问题研究及数据清洗工具Data cleaner的设计

实现[硕士学位论文] 沈阳：东北大学，2003

[33] 张悦．异构环境下ODS构建技术的研究与实现[D].[硕士学位论文] 沈阳：沈阳航空工业学院，2005

[34] 周宏广．异构数据源集成中清洗策略的研究及应用[D][硕士学位论文]．长沙：中南大学,2004

[35] Rahm,E.，Do，H. H. Data cleaning：woblems and current approaches. IEE Dam Engineering Bulletin，2000,23(4)：3-13.

[36] 克龙，王玲，等．数据仓库中ETL工具的探讨和实践．计算机应用和软件，2005；22(11)：30—78

[37] 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(11)2076-2079

[38] 周奕辛．数据清洗技术的研究山东科技大学学报(自然科学版)，2004，23(2)：55—57

[39] 周奕辛.数据清洗算法的研究与应用[D][硕士学位论文]．青岛：青岛大学,2004

[40] 鲁红英，肖思和．基于改进的遗传神经网络数据挖掘方法研究［J］计算机应用第26 卷第4 期2006 年4 月878-882

[41] 覃华 苏一丹 李陶深 基于遗传神经网络的数据清洗方法［J］计算机工程与应用 2004 45-46,67

[42] 张 宁．数据仓库中ETL 技术的研究[J]．计算机工程与应用,2002 (24) :213 - 216

[43] 赵静静．数据挖掘中空缺值预测算法的研究与实现[硕士学位论文].南京：南京航空航天大学

[44] Yong Yu, Trouve A. A Non-linear K-means Algorithm and Its Application to Unsupervised Clustering[J]. Signal Processing, 2002 6th International Conference, 2002, 2:1146-1149

[45] Chu Shu-chuan, Roddick J F, Tsong-Yi Chen,et al. Efficient Search Approaches for K-medoids-based Algorithms[J]. TENCON '02,Proceedings. 2002 IEEE Region 10 Conference on Computers,Communications, Control and Power Engineering,2002,1:712a-715a

[46] H.Galhardas, D. Florescu, D.Shasha, and E. Simon.AJAX:An Extensible Data Cleaning Tool［J］. In SIGMOD(demonstration paper),2000.

[47] LI Jue,GUO LiXin.A Research on Evaluating Project Management Maturity Based on Genetic BP Neural Network Date Mining Method.Conference of WISM-AICI2010 VOLUME Ⅲ, 2010, 09:477-480

[48] Chaudhuri S, Dayal U, Ganti V. Database Technology for Decision Support Systems[J]. Computer, 2001,34(12):48-55

[49] Jiawei H, Micheline K. Data Mining: Concepts and Techniques[M].Copyright by Morgan Kaufmann Publishers, Inc., 2001,5:73-74

[50] 张永强，数据中心架构的研究与实践[D][硕士学位论文]，广州：中山大学，2008

[51] 刘伟光，面向电信客户流失管理的数据仓库原理研究与应用[D][硕士学位论文]，大连：大连海事大学，2008

[52] 荀立坤，山东电力安全生产管理信息系统设计与实现[D][硕士学位论文]，山东：山东大学，2008

[53] 李铁成，基于系统方法论的数据库发展及建设研究[D][硕士学位论文]，沈阳：东北大学2007

[54] 王鹏 张焕水，数据仓库技术在数据服务平台中的应用［J］计算机与信息技术，2009，第3期：23-27

致　谢

三年的研究生学习生活即将结束，在这里非常感谢我的导师陈旭辉教授，陈老师在我的为我制定学习计划、建议研究方向、提供研究资料。尤其是在我撰写论文遇到困难时，陈老师更是耐心的指导，积极的鼓励和支持，并给予我莫大的鼓舞和鞭策来完成论文的写作。陈老师字逐句地为我修改攻读硕士期间发表的小论文。对于我这样的参加过工作的人来说，只有老师才能这样无私，不辞辛苦的用自己所学的所有知识来帮助每一个学生进步。在陈老师的指导教育下，我不仅学会了如何做学问，更重要的是我学到了为人处事的道理。使我在今后的工作生活中受益匪浅在此，谨向恩师表示崇高的敬意和由衷的感谢。

感谢师兄，师弟师妹们，在学术研究，论文撰写方面给予我很大的指导和帮助。还要感谢各位评审专家，花了他们很多时间来看我的论文，并给我提出了宝贵的修改意见。在这里我也衷心的感谢我的爱人和我的父母，他们多年来对我的信任、鼓励和支持是我人生中最大的一笔财富，也是我不断进取的动力。我还要感谢的人就是我的宝宝虽然妈妈没有更多的时间照顾他，但是他还是很健康、很乖、很懂事。

附录 A 攻读学位期间所发表的学术论文

[1] Xuhui Chen. Xinghua Zhang, Extract-Transform-Load of data cleaning mothed in electric Company. Conference of International Conference on Artificial Intelligence and computational Intelligence.WISM-AICI2010 VOLUME Ⅲ. p345-348 (IEEE, Springer) 978-7695-4225-6.