摘　要

自上世纪 90 年代开始，国外的商业银行就开始应用数据仓库相关技术实现对银行卡业务分析、利润分析、客户分析等，并在实施后取得了很好的利润回报。尤其在客户分析上，国外银行通过实施以客户为中心的经营理念和运用数据仓库相关技术而建立的客户关系管理（CRM, Customer Relationship Management）系统，使国内的商业银行特别是中小型商业银行，在激烈的竞争环境中将不再具有国内客户资源优势。

本文首先深入分析了目前国内商业银行面临的激烈竞争局面和数据仓库技术在国内外商业银行的实施动态，讨论了数据仓库相关技术现状发展状况和前景。在充分理解银行业相关领域知识、分析国内某银行的业务现状和发展目标的基础上，论文设计了用于该银行 CRM 系统的数据仓库。针对目前该银行处理数据量较大的问题，该系统采用的是分布式数据仓库系统和三层体系结构。

其次，本文围绕该银行 CRM 系统是用于客户分析这一目的，建立了数据仓库的数据模型。数据模型设计确认了该数据仓库系统的四个主题客户、账户、交易、产品及其数据粒度；并围绕这四个主题进行深入的分析，设计了相应的事实表、维表、索引策略、存储结构、存储策略等。

最后，本文对本数据仓库系统构建过程的关键技术 ETL（Extract－Transformation－Loading）作了重点研究。针对该银行源数据库系统中数据量大且有部分异构数据的问题，提出了该数据仓库系统的数据抽取方案、数据接口、数据映射关系和数据清洗加载的解决方案，并对该数据仓库系统的 ETL 做了部分实现。

关键字　数据仓库，客户关系管理，数据模型，ETL

ABSTRCAT

Since 1990s, oversea commercial banks have applied the related technology of data warehouse to analyse the performance evaluation, profitability analysis and customer relationship management(CRM) etc, and they have got effective profit management by do this .In the CRM, the foreign banking corporations have established the idea of "customers first" and widely applied the technology of data warehouse, these practices make the domestic commercial banks, especially the small and middle ones, do not have the superiority of customer resource in the environment of keen competition.

Firstly, this thesis analyses deeply the situation of steep competition which domestic commercial banks are facing at present and the practising state of the data warehouse technology in the domestic and foreign commercial banks. And it discusses the data warehouse related technology and their development and prospect. Based on widely mastering the knowledge of banks and analysis or the service and development targets of a domestic commercial bank, this thesis designs a data warehouse for the CRM system of this bank. There are large amount of data in the database, so the system uses distributional and the three-level structure.

Secondly, based on the objective of CRM system for the customer's analysis of bank, the thesis designs the data model of this data warehouse system. The thesis confirms four subjects and data granularity of this system, including the customer, the account, the transaction, the product. After analyzing these subjects, the thesis designs the fact table, dimension tables, index strategy, memory structure, memory strategy etc.

Finally, the thesis stresses the research of ETL, which is the key technology of this data warehouse system. Because there are large amount of data and some heterogeneous data in the bank, it proposes the system data extract strategy, data interface, data mapping relations, the data clean and load and so on, and realizes parts of ETL process.

KEY WORDS Data Warehouse, CRM, Data Model, ETL

原创性声明

本人声明，所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了论文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中南大学或其他单位的学位或证书而使用过的材料。与我共同工作的同志对本研究所作的贡献均已在论文中作了明确的说明。

作者签名：陈才　　日期：2008年5月15日

学位论文版权使用授权书

本人了解中南大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并根据国家或湖南省有关部门规定送交学位论文，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以采用复印、缩印或其它手段保存学位论文。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：陈才　导师签名谭双杉　日期：08年5月15日

第一章 绪论

1.1 课题的研究背景

在 2001 年中国加入 WTO 之后的五年过渡期内，我国银行业将逐步对外开放，这五年中，在华外资银行营业性机构从 190 家增加到 312 家，剔除机构合并等因素净增 122 家[1]。由此可见，中国银行业目前不但面临着激烈的内部竞争，还面临的国外银行业的前所未有的竞争压力。外资银行进入中国市场后，往往采用差异化服务策略，即利用标准化的产品、先进的管理和服务、领先的信息技术争取本地的黄金客户，进而扩张市场。

面对金融市场开发带来的各种挑战，我国的商业银行正处在一个非常重要的转折期。目前，国内大部分商业银行已经开发出了较为完善的会计系统、储蓄系统、国际借宿系统、资金系统、信贷系统、信用卡系统以及办公自动化系统等，使得其经营管理手段有了很大的提供。但是，国内的银行在市场结构、业务范围和经营理念上与国际先进金融企业的差距还比较大，另外，国内有一个强大的分析型应用系统的银行也很少，该系统对内能帮助金融企业加强风险管理和绩效考核，对外能够加强客户关系管理，增加赢利。因此，在信息技术上，要求国内商业银行尽快加强分析型应用系统的建设，帮助银行加强经营管理与决策分析，了解客户的需求与信用风险，开发新的产品和服务，利用现有渠道对客户进行交叉销售。

银行每时每刻都在搜集和处理大量的信息，是一个具有信息优势的部门，这些信息包括公司和个人的资料、票据信息、账户信息、贷款信息以及国内外企业、行业、产业、产品信息等。如果利用这些宝贵的信息资源和获取有益的知识，以更好的为客户服务，提高银行经营管理水平，已经是一个迫在眉睫的问题。著名的数据仓库专家 Ralph Kimball 写道："我们花了二十多年的时间将数据放入数据库，如今是该将它们拿出来的时候了"[2]。另外，银行的各种数据不能有效结合，形成了很多"信息孤岛"，使金融机构很难将各种各样的客户信息统一起来，领导决策层也很难搞清楚数据库系统的整体运作情况，不能有效的提供决策帮助。

数据仓库及相关技术就是这样一种平台，它能够为各个行业的竞争提供准确的、有效的分析型数据平台，以支持管理层的决策，CRM（Customer Relationship Management）等。数据仓库技术在许多行业有较广泛的应用，如：电信业、银行业、证券业、税务领域、保险业、图书馆等，特别是金融银行业，面对着日益激烈的竞争，国内外商业银行和金融企业都应用了数据仓库技术来优化银行的决

策支持、贷款风险控制和客户关系管理等。数据仓库是金融机构实现 CRM 的核心技术，也是金融竞争优势的来源。

1.2 国内外银行实施数据仓库的动态

数据仓库技术在银行业的应用已经越来越广泛，世界上的大银行都认识到数据仓库对于自己决策系统的重要性，比如银行卡信用卡分析系统、利润贡献度分析系统和客户关系管理系统，它们在服务提升、产品提供、成本管理、价格策略和营销费用上发挥了重要的作用。目前，国外和国内的部分银行基本上正在或已经建立了自己的数据仓库系统，以应对客户的增加和日益激烈的竞争，通过对数据的分析和加工来很快的做出正确的决策。但是，与国外银行相比，国内的各大商业银行在这方面起步较晚，而且，中国商业银行拥有更庞大的客户群，更广泛的营业网络，到目前为止，虽然他们都在逐步建立了基于自己的数据库的数据仓库，然而，这些数据仓库中的数据依然是错综复杂的互不关联，不能准确而快速的提供信息给决策层[3]。

1.2.1 国外银行业的现状

国外银行非常重视数据仓库，因此，在这方面起步早，投入大，在技术上和应用的范围和效果上都取得了非常大的进步。据美国 The Tower Group 咨询公司的有关调查资料：1998 年全球 500 家大银行中，已经有近 90 家建立了数据仓库；1999 年全球金融机构用于数据仓库建设的投资高达 54 亿美元，其中仅欧美地区就占了八成[4]。到 2002 年，估计前 500 强的银行中，正在筹备、在建或已经建成数据仓库的银行至少超过了 300 家。这其中包括了著名的美国花旗银行、美洲银行、汇丰银行，英国的 Barclays 银行，加拿大皇家银行等。在客户关系管理上，目前全球前二十大银行中，大部分都已经采取了数据仓库技术作为该管理系统的平台，平均的获利率超过 15%[5][6]。

英国 Barclays 是全球最早实施基于资产负债管理数据仓库的银行，对每个事业部资金的分配以信用、市场、企业和经营风险为考虑因素，坏账勾销显著下降[7]。

加拿大皇家银行通过实施数据仓库，大幅度提供了客户细分的准确性，在目标客户行销活动中，RSP 储蓄有 11%的提高，实现了以价值为基础的客户分析，增加了近 20%的高利润贡献客户。从而使得该银行在过去的三年中，每年市场营销带来的收入都有两位数字的增长。目前，加拿大皇家银行完全可以依靠数据的强大功能对客户进行细分，以客户的分布来细分市场，管理上已经非常扁平化。

美国美洲银行在1998年与国民银行合并后成为世界上最大的零售银行，其机构分布在全球45个国家，总资产世界排名第四位。作为如此大的银行巨鳄，每天需要处理的业务数据量巨大，业务数据库系统也积累了海量的数据。美洲银行于1986年就开始投资建立数据仓库系统，由开始数据库的容量20G逐步扩展成3.4TB的庞大系统。这套系统在1994年的洛杉矶大地震后充分显示了其价值，银行根据邮政编码进入到每个区以了解地震破坏的区域和程度，由此在短期内确定了损失。在信用卡业务的拓展方面，该数据仓库也起了很大的作用。对该数据仓库中的客户信息的分析，一方面，识别出哪些是优质客户，对这些客户提供更好的服务；另一方面，对那些信用差的客户，防止其呆帐行为，以降低风险。另外，在一次促销活动，通过分析系统中的数据找出那些使用了美洲银行的服务但没有使用信用卡服务的25万个经常有信用卡消费、收入固定、信用度较好的客户，给予一些优惠等来吸引他们，在这次促销活动吸引了其中的23%，即大约6万个高高贡献度的信用卡客户。

1.2.2 国内银行业的现状和不足

与国外银行相比，中国的商业银行拥有更庞大的客户群体、更广泛的营业网络，但是国内的商业银行是从上世纪末才开始逐渐开展数据仓库项目[3][8]。

1998年，招商银行通过与Sybase公司等合作，联合开发了招商银行数据仓库系统，包括人事、储蓄、会计等系统，成为国内银行业界第一个成功可用的数据仓库系统。

2001年，中国工商银行正是启用了数据仓库建设工程，通过建立领先的综合业务系统、数据中心和构造功能强大的数据仓库，形成了先进的金融信息技术平台，开创了经营管理信息化的新局面。

台湾地区的信托商业银行也利用数据仓库管理信用卡业务，采用了基于活动的成本核算，以进一步了解每位客户的利润贡献度，大大节省了营销成本，从而成为台湾地区最大的发卡银行和信用卡业务获利最高的银行。

其他国有银行和私营银行也逐步开始建立基于本行业务数据库的数据仓库系统，如中国银行、中国农业银行、中国建设银行等。虽然已经在信用卡管理、报表生成、客户关系管理等方面取得了一些成绩，但是仍存在着构建方法不够科学、模型不清晰、应用开发不足的等普遍问题。

1.3 数据仓库相关技术研究

1.3.1 数据仓库技术

1. 数据仓库的特点

要准确的给数据仓库下一个定义比较困难，我们可以简单的理解为数据仓库就是存放数据的地方，但这种理解比较肤浅，没有把数据仓库和数据库区别开来，更没有突出数据仓库的本质特征。目前普遍采用的定义是权威的数据仓库之父W.H.Inmon给出的：数据仓库是一个面向主题的、集成的、时变的、非易失的数据集，支持管理部门的决策过程[2]。这个定义概括了数据仓库的四个主要特征，道出了数据仓库与其他存储系统（如关系数据库系统，事务处理系统，文件系统等）的区别，同时指出了数据仓库是为支持管理部门的决策服务的，而不是普通的数据存储。

（1）面向主题（subject-oriented）

操作型数据库的数据组织面向事务处理任务，各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织。主题是一个抽象的概念，是指用户使用数据仓库进行决策时所关心的重点方面，一个主题通常与多个操作型信息系统相关。

（2）集成的（integrated）

对源数据的集成是数据仓库建设中最关键，也是最复杂的一步。数据仓库中存储的数据是从各国分散子系统中提取出来的，但并不是原有数据的简单拷贝，而是经过统一综合。由于源数据不适合用来分析，在进入数据仓库之前必须经过综合，计算，抛弃等方法分析处理不需要的数据项，并增加一些可能涉及的外部数据，以此来消除不一致和错误之处，以保证数据的质量，否则，用那些不准确或不正确的数据分析得出的结果就不能做出科学的决策。

（2）非易失（nonvolatile）

数据不可更新是从数据的事业方式上来看的。数据保存到数据仓库后，最终用户只能通过分析工具进行查询和分析而不能修改，即数据仓库的数据对最终用户而言是只读的。

（4）时变的（time-variant）

数据仓库数据的不可更新是针对应用而言，即用户分析处理时不更新数据。但不是说，数据进入数据仓库以后就永远不变，这些数据会随时间变化而定期更新。每隔一段固定的时间间隔，将抽取运行数据库系统中产生的数据，经过转换后集成到数据仓库中。随着时间的变化，数据以更高的综合层次被不断综合，以

适应趋势分析的要求。当数据超过数据仓库的存储期限，或对分析无用时，就从数据仓库中删除这些数据。

2. 数据库与数据仓库的区别

作为数据存储的工具，数据库和数据仓库是一样的，都可用于保存数据。数据库技术已经发展了几十年，并应用到了各个行业中，已经是一种较成熟的技术，而数据仓库技术则是一种新发展的技术。但是数据库关注的是事务处理的及时性、完整性与正确性，在数据的分析处理方面存在缺乏集成性、主题不明确等缺点[9][10]。数据库由于其自身条件所限，不能作为大规模数据综合分析平台，企业的决策迫彻需要一种新的技术来提供支持，这种新的技术就是数据仓库。它们之间主要有以下几点区别。

（1）系统任务不同

数据库系统和数据仓库系统的主要区别体现在 OLTP(On-Line Transaction Processing)和 OLAP(On-Line Analytical Processing)。数据库系统的主要任务是联机事务处理（OLTP），例如：日常生活中的成绩查询、购买火车票、预定飞机票等。而数据仓库系统的主要任务则是基于 OLAP 技术的数据分析，支持复杂的分析操作，其主要目的是提供决策支持。

（2）设计的面向性不同

数据库是面向事务的设计，而数据仓库是面向主题的设计。

（3）数据内容不同

数据库一般存储数据量相对小、需要在线交易的数据，但这些数据不能用于决策。而且数据仓库存储的一般是被数据库系统淘汰的“过时”的海量历史数据，OLAP 系统提供对这些海量数据进行汇总、聚集以及在不同粒度级别上进行分析的能力。

（4）设计的指导方法不同

数据库设计采用范式理论来指导，目的是尽量避免数据冗余。数据仓库在设计时则允许存在数据冗余，甚至存在较大的冗余，可以不遵守范式理论。

（5）目的不同

数据库是为常规的数据管理而设计的，以提供实时的数据检索和存储服务。数据仓库则是为分析数据而设计，企业管理层提供决策支持，其基本概念是维表和事实表。

（6）访问模式不同

对数据库的访问主要是原子事务组成，并且需要对其进行并发控制和恢复操作。对数据仓库系统的访问则绝大部分是只读访问，多体现维复杂的查询，不要考虑并发控制。

因此，数据仓库的出现，并不是要取代数据库，它们是两种目的不同、并行发展的数据处理技术。

3. OLAP 技术

联机分析处理（On-Line Analytical Processing, OLAP）是由关系型数据库技术的创始者 E.F.Codd 在 1993 年提出的。针对 OLTP 已经不能满足用户对数据库查询分析的需求，他提出了多维数据库和多维分析的概念，从而描述了信息处理技术的一个新领域[13]。OLAP 是一种软件技术，它使分析人员、经理和主管人员能够通过快速的、一致的和交互式的访问来获取并理解各种可能的信息视图的数据，这些信息由原始数据转换而来，用来反映一个企业实际的维度。其主要特征有：建模功能、存储功能、独立性、快速性、交互性和展示功能。

OLAP 与数据仓库有着密切的联系，但是二者又是两个不同的概念。OLAP 是使用客户端应用程序对数据仓库的数据进行有效分析的一种技术。通过这种技术，分析人员能够从多种角度对从原始数据进行观察和分析，以实现对数据深入的了解并从中获取潜在规律。

OLAP 的核心是“维”的概念，因此 OLAP 也可以说是一种多维数据分析的工具。进行 OLAP 分析的前提是必须建好数据仓库，也就是说，OLAP 分析是在数据仓库的基础上进行的。利用 OLAP 复杂的数据查询、数据对比、数据抽取等能力可以对数据仓库中的数据进行多维度、深层次的分析、可以在不同的粒度上对数据进行分析，得到不同形式的知识和结果，并将结果以直观的形式提供给决策层，从而实现对决策的支持[14]。为了满足决策支持或者在多维环境下特定的查询和报表需求，OLAP 的分析工具是以变量、维、维的层次、维成员、多维数组、数据单元等基本概念来展开的。

（1）OLAP 分析的基本操作

OLAP 分析主要是指采用切片、切块、旋转、钻取、等基本操作手段，对以多维形式组织的数据进行深入研究，从而使客户达到从多个角度、多个细节分析数据的目的。

（2）OLAP 的实现技术

OLAP 系统按照其存储器的数据存储格式可以分为关系 OLAP（Relational OLAP，简称 ROLAP）、多维 OLAP（Multidimensional OLAP，简称 MOLAP）和混合型 OLAP（Hybrid OLAP，简称 HOLAP）三种类型[15]

ROLAP：ROLAP 基本数据和聚合数据均存放于 RDBMS 中，ROLAP 将多维数据库的多维结构划分为两类表：一类是事实表，用来存储数据和维关键字；另一类是维表，即对每个维至少使用一个表来存放维的层次、成员类别等维的描述信息。维表和事实表通过主关键字和外关键字联系在一起，形成了“星型模式”。

对于层次复杂的维，为避免冗余数据占用过大的存储空间，可以使用多个表来描述，这种星型模式的扩展称为“雪花模式”。ROLAP 的最大好处是可以实时地从源数据中获得最新数据更新，以保持数据实时性，缺陷则在于运算效率比较低，用户等待响应时间比较长。

MOLAP：MOLAP 基本数据和聚合数据均存放于多维数据库中，多维数据在存储中将形成“数据立方体（Cube）”的结构，此结构在得到高度优化后，可以最大程度地提高查询性能。随着源数据的更改，MOLAP 存储中的对象必须定期处理以合并这些更改。两次处理之间的时间将构成滞后时间，在此期间，OLAP 对象中的数据可能无法与当前源数据相匹配。维护人员可以对 MOLAP 存储中的对象进行不中断的增量更新。MOLAP 的优势在于由于经过了数据多维预处理，分析中数据运算效率高，主要的缺陷在于数据更新有一定延滞。

HOLAP：HOLAP 基本数据存放于 RDBMS 中，聚合数据存放于多维数据库中，用户可以根据自己的业务需求，选择哪些模型采用 ROLAP，哪些采用 MOLAP。一般来说，会将非常用或需要灵活定义的分析使用 ROLAP 方式，而常用、常规模型采用 MOLAP 实现。

4. 数据仓库技术的发展前景

随着数据仓库技术的发展，其在各个企业的应用了十多年，譬如：电信业、银行业、证券业、税务领域、保险业、图书馆等。成功和失败的示例都比较多。总体来说，目前的数据仓库技术已经发展得较为成熟，如华尔街 60%的银行、保险、证券等公司采用了数据仓库技术进行风险管理，CRM 等，其中包括著名的摩根、花期银行、加拿大蒙特利尔银行、加黄银行等[17]。

（1）电信业

目前，中国加入 WTO 后，国外的电信企业将逐步进入我国市场，使得我国电信行业的竞争正在一步步迈向国际化和市场化，也使得我国电信业的竞争日趋白热化，这就需要我国的电信运营商通过更好的产品和更完善的服务来赢取更多的客户和最终的胜利。在这种形式下，我国电信行业的老大中国电信做出重大决策，决定利用最近十年发展起来的数据仓库技术及基于此技术的商业智能，深层次、多角度的挖掘，分析当前和历史的生产数据、客户信息、竞争对手的信息等相关环境的多种数据，发现其内在的联系，从而得到宝贵的决策支持信息，并对企业未来的生产计划和长远规划提供指导。

（2）银行业

通过实施先进的数据仓库技术，银行能够对全行业务数据进行集中存储和管理，合理地对信息进行详细分类，准确收集和分析信息，确保管理层随时掌握银行的经营风险、情况和目标等。银行可以通过对数据仓库中数据的关联分析，衡

量各类客户需求、满意度、潜在价值、信用度和风险度等指标，以识别客户群体，实施差别服务的策略提供技术支持。在国内，我国的商业银行也在逐步实施数据仓库技术来应对不断激烈的竞争。

1.3.2 数据挖掘技术

随着数据库的广泛应用，人们陷入了“数据丰富，知识贫乏”的尴尬处境中，因此人们急需一种新技术和自动工具可以帮助我们科学地进行决策，数据挖掘技术就是这样一类技术。数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的应用数据中，提取隐含在其中的潜在有用的信息和知识的过程[18]。从广义上来看，数据分析分为验证型分析（Verification-Driven Data Analysis）和挖掘型分析（Discovery-Driven Data Mining）。其中，多维查询和 OLAP 属于验证型范畴，它们可以很方便地观察系统的实际情况，以确定某种假设是否成立。数据挖掘则是在大量数据中发现未知的知识，属于挖掘型分析。数据挖掘是许多学科的交叉，运用了统计学，计算机，数学等学科的技术，其任务主要是关联分析、聚类分析、分类、预测、时序模式和偏差分析等[19]。

1.数据挖掘过程

在很多领域，人们把数据挖掘作为 KDD（Knowledge Discovery in Database）过程中对数据真正应用算法抽取知识的那一部分。KDD 过程是一个复杂的过程，其步骤如下：研究问题域、选择目标数据、数据预处理、数据挖掘、模式解释与评价、应用等[20]。但是在商业领域中，数据挖掘是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性信息。这个概念其实包括了整个知识挖掘过程，因此，虽然数据挖掘是整个知识挖掘过程中的一个重要步骤，但在本论文中后面提到的数据挖掘即指的知识挖掘过程。

2.数据挖掘方法

数数据挖掘技术发展的过程中，各种挖掘方法也在不断的涌现出来并不断成熟，主要有决策树、关联分析、神经网络、粗糙集、遗传算法、模糊集、概率论与数里统计等。

（1）决策树：是指利用信息论中的信息增益寻找数据库中具有最大信息量的属性字段，建立决策树的一个结点，再根据该属性字段的不同取值建立树的分支，每个分支子集中重复建立树的下层结点和分支的过程。决策树对于多峰分布类型的问题尤为有用，利用决策树采用分级的形式，能够把一个复杂的多类别分类问题转化为若干简单的分类问题来解决。

（2）关联规则

关联规则挖掘即在当前记录的各个特征间寻找内在的联系。它的典型应用就是购物篮分析，即在某个商店的销售使用数据中分析该商店的“大部分顾客会在一次购物中同时购买什么商品？”，以便对商品促销、布局等提供帮助。

（3）神经网络

神经网络是模仿人的头脑，通过一个训练数据集学习和应用所学知识来生成分类和预测的模式。它是一种最复杂的分类和回归算法，这种方法在数据是不定型的且没有任何明显模式的情况下很有效。

3.数据挖掘技术的应用现状

目前，数据挖掘技术在各个行业的应用十分广泛，如电信、银行、保险、交通、零售等领域。数据挖掘能够解决以下一些典型的商业问题：数据库营销、客户群体分析、背景分析、交叉销售、客户流失性分析、客户信用记分、欺诈发现等[21]。

数据挖掘技术在国内外金融领域有广泛应用，银行等金融需要搜集和处理大量数据，对这些数据进行分析，发现其数据模式和特征，然后挖掘出某个客户、消费群体和组织的金融或商业兴趣，并且能够观察整个金融市场的变化趋势[22]。如：世界著名的金融信息服务公司 Reuteres，它利用的数据大都是外部的数据，这些数据的质量就是公司生存的关键所在，必须从数据中检测出错误的成分。Reuteres 用 SPSS 的数据挖掘工具 SPSS/Clementine，建立数据挖掘模型，极大地提高了错误的检测，保证了信息的正确和权威性。

1.3.3 CRM 相关内容

客户是企业最重要的资源，对以服务为主要“产品”的银行来说尤其如此。在当今激烈竞争的市场环境中，能否了解客户的实际需求，并提供量身定制的个性化服务，已成为决定银行成功与否的关键因素。银行建立 CRM 系统的目的是从客户需求出发，及时准确地制定市场决策，不断维护和拓展客户群，同时优化银行内部的资源，提高银行的运作效率，挖掘更多的创收机遇，从而实现收益的持续增长。

1. CRM 概述

客户关系管理（Customer Relationship Management，CRM），最初起源于 20 世纪 80 年代初提出的“接触管理”（Contact Management），即专门收集整理客户与公司联系的所有信息[23]。CRM 的产生是在 1990 年前后，由国际上比较权威的研究组织 Gartner Group 正是提出 CRM 的概念：“客户关系管理（CRM）是为了代表增进盈利、收入和客户满意度而设计、企业范围的商业战略”。经过近 20 年的发展，CRM 日趋成熟，最终形成了一套完整的管理理论体系，并发展成为

两种 CRM 类型：运营型 CRM 和分析型 CRM。

（1）运营型 CRM（Operational CRM）

运营型 CRM 在企业成功方面起着重要的作用，它要求所有业务流程的流线化和自动化，包括由多渠道的客户“接触点”的整合、前台和后台运营之间的平滑连接和整合。

（2）分析型 CRM（Analytical CRM）

分析型 CRM 主要是分析运营型 CRM 种获取的各种数据，进而为企业的经营、决策提供可靠的量化依据。这时的分析需要用到许多先进的数据管理和数据分析工具，如数据仓库、OLAP 分析和数据挖掘等。在银行业中分析型 CRM 应用得较多。

2. 商业银行实施 CRM 的目的

随着市场竞争的日益激烈，消费者对产品选择的科技越来越大，一个公司在研究和拓展产品的物质市场时，必须研究和拓展产品在客户心中的市场。因此，CRM 是许多企业提高核心竞争的法宝，目前 CRM 在各个行业的应用非常广泛，如金融行业、电信行业、零售行业、保险行业、医药行业、旅游行业等，虽然企业实施 CRM 时也有一些失败的经历，但是在各行各业的应用 CRM 之后取得效果还是比较明显的。

特别是银行业，竞争尤其激烈，目前，全球有几百家银行，我国四大国有商业银行，其他私营银行就有几十家，要想在如此多的竞争者中取胜，更快更好的做出正确的决策而吸引更多的客户显得极为重要。银行实施 CRM 就是在这种情形下产生和发展起来的，银行 CRM 的优势主要有以下四点：

（1）挖掘客户的潜在价值

银行通过长期的营运，积累了相当丰富的客户信息资源，但这些信息都是处于相当独立的数据库或表中，形成了一个个“信息孤岛”，不通过特定技术对它们进行处理，很难发现它们之间的内在联系。这些信息被综合处理之后能比较准确的反映出客户的购买力、收入、服务历史、生活方式等，通过分析可能预知客户将来的行为等。比如，国外某银行有这样一个案例：在一次

（2）使银行更了解客户的真正需求

客户的需求是企业经营的核心，只有了解了客户的需求，才能做出令客户满意的产品，从而在竞争中吸引更多客户。银行客户需要的是方便、快捷的服务和服务人员礼貌、周到的服务。如何真正把握客户的需求，如何向客户提供一对一的优质服务，提供客户的满意程度，增加竞争力，就是客户关系管理的目标。

（3）能对盈利客户进行针对性营销

在各个行业，银行业也不例外，存在着一个著名的二八定律：占客户群 20%

的客户实现利润往往占到总利润的 80%以上，这 20%的客户就是企业的黄金客户。CRM 要做的就是根据对客户的成本利润分析，来找出这一部分客户，并对市场进行细分，针对不同的客户实施不同的策略。

（4）提高客户忠诚度

有调查表明：要争取一个新客户比保持现有客户需要多几倍的成本。银行实施 CRM，对于留着老客户，提高客户的忠诚度，使之不投向自己的竞争对手有较大的帮助。银行拥有客户的各种资料，可以通过建立 CRM 体系来了解每个老客户，从而根据客户的需求提供服务。

3. 客户满意度和忠诚度

客户满意和客户忠诚是两个相关的概念，它们之间相互促进补充。只有满意的客户才会对企业“忠诚”，同时，客户满意是以客户“忠诚”为支点。客户的非常满意并不意味着就会忠诚，大多数客户满意只是比较之后的感觉，可能是没有投诉或感觉不好的地方，但并不一定就要忠诚。只有一个高度满意的客户才会对企业忠诚更久，而一般满意的客户一旦发现更好的产品，就会很容易的选择其他供应商。科特勒的研究指出：企业可能流失 80%极不满意的客户，40%有些不满意的客户，20%一般不满意的客户和 10%一般满意的客户；公司只会流失 1%的高度满意的客户；95%的不满意客户不会投诉，仅仅是停止购买[24]。因此，要使客户有高满意度是有一定困难的。

客户满意是指一个产品的可感知效果与客户的期望值相比较后，所形成的愉悦或失望感觉状态。企业可以通过一些连续性或非连续性调查，获取客户对特定产品和服务的满意度、未满足需求、再次购买率和推荐率等指标的评价。这种调查能够对企业当前服务的质量进行量化的评估，并通过因素重要性推导模型判断服务中急需改进的因素，以此来改善产品质量、服务态度和维护扩大现有客户群基础。

保留一个老客户的成本只有开发一个新客户成本的六分之一，因此，保留老客户成为许多企业的重要目标。保留老客户的时间越长，获得的利润久越高，有研究表明，在许多企业客户的忠诚度提高 5%将带来 25%－85%的收益增长。现代社会中，企业竞争力在市场主要体现为品牌竞争力，然而，企业要提供品牌竞争力，关键在于提高客户的品牌忠诚度。企业应该通过各种途径，如和客户建立长期的关系、加强和客户之间的沟通、优化客户关系，以熟知客户的需求，并确保客户的要求贯串于企业经营过程中。

4. 国内外商业银行实施 CRM 现状

从上个世纪 90 年代开始，国外的商业银行就逐步把数据仓库技术应用到 CRM 中。最早把 CRM 引入银行业的是花旗银行，花旗银行善于运用新技术抢

占市场和潜在客户，这也是花旗银行在日益激烈的银行业竞争中取胜的一个法宝。CRM 的实施使花旗银行从“以产品为中心”的模式转移到“以客户为中心”的模式，银行关注的焦点从内部运作转移到客户关系上来。运用系统，花旗银能准确地说出谁是他们盈利来源最多的客户，并能在 10 分钟内讲清楚重要的银行客户使用了多少种银行产品，这无疑使花旗在同业中显得鹤立鸡群。另外，英国巴克利银行、加拿大皇家银行、美国美洲银行、荷兰皇家银行、日本三菱银行等一些国际银行巨头都建立了完善的数据仓库来支持自己的决策，并从中获取了比建立数据仓库的前期投资更为丰富的利润回报[25]。

相对于国外商业银行较早建立了以数据仓库为基础的 CRM 系统，我国商业银行在这方面的起步较晚。2006 年我国银行业已对外完全开放，国外银行的进入使得我国的商业银行面临的竞争也将变得更为激烈。现在国内几大商业银行都在着手调研、准备或者尝试实施各种基于数据仓库技术的 CRM 解决方案。中国工商银行进行了以个人客户关系管理（PCRM）和个人业绩价值管理（PVMS）为主题的应用试点，中国银行则全面规划了信用卡系统，其中很重要的一个子系统就是基于数据仓库技术的销售和客户服务系统[26]，中国农业银行已经在广东分行进行经营分析系统的建设。

基于数据仓库的 CRM 在商业银行的应用虽然有很多成功的案例，但是许多国内外的商业银行 CRM 系统实施的效果总体上并不理想，有些甚至以失败而告终。英国顾问公司 Butler Group 的一份报告指出，使用 CRM 的失败率高达 70%。Gartner 的调查研究表明，在所有 CRM 项目中大约 55%没有达到软件用户的预期目标[27]。因此，商业银行要利用 CRM 数据仓库实现盈利，还有许多工作要做，其实施也将是一个长期的过程。

1.4 课题来源及研究意义

本人是在导师的安排下前往北京长信通信息技术有限公司实习的，参与了该公司项目组在某银行科技部的核心系统维护和其他相关项目的开发，包括该银行核心系统维护、CRM 数据仓库项目、贵宾卡积分系统项目和第二期记账式国债项目的开发工作。在充分理解该银行业务特点和现状的基础上，本论文对该银行数据仓库项目开发进行了深入研究和设计。

在目前全球化竞争日益激烈的环境下，一个企业要在竞争中取胜的关键是它能否根据各种不同用户的需求做出快速而准确的决策并提供优秀的服务和产品。在金融银行业尤其如此，银行业务数据库中积累了大量的各种数据，这其中包括具有隐藏有潜在价值的客户信息，但缺乏一种机制来发现这些信息的利用价值。通过对客户及相关信息进行处理，将会发掘优质客户、控制各种风险和推出针对

性的服务等。由于业务数据库系统是针对业务处理的OLTP系统，这些数据并不能直接用来分析，这就需要建立一个 OLAP 的平台来为以上决策提供支持，这就是数据仓库。

对于该银行，为应对国内外金融业的激烈竞争，建立基于本行业务数据库的、全行统一的、数据一致的、能够对详细历史交易数据进行综合分析的数据仓库系统显得尤为重要。

1.5 论文的研究内容及结构

本文在分析了目前数据仓库技术的发展状况和国内外商业银行实施数据仓库的背景下，深入讨论的数据仓库相关技术：数据仓库的特点、OLAP；数据挖掘的过程、方法和应用现状；CRM（客户关系管理）：银行实施 CRM 的目的、客户满意度忠诚度、银行业 CRM 的现状等。并结合某银行的业务状况和发展目标，针对该银行提出了自己的解决方案，给出了该商业银行数据仓库的总体设计，概念模型，逻辑模型及物理模型的设计和数据仓库关键技术——ETL 的设计和部分实现。

本论文包括五章，其组织结构如下：

第一章绪论主要概述了本课题的研究背景，数据仓库相关技术及其在国内外商业银行的实施现状及论文的研究意义。

第二章根据目前该银行的业务发展需要，给出了基于其业务数据库系统的CRM 数据仓库总体设计方案。

第三章进行了该系统数据模型的设计，包括概念模型、逻辑模型和物理模型的设计。

第四章讨论了数据仓库的关键技术——ETL 技术，给出了该数据仓库系统中数据抽取、转换、清洗及加载的方案。并利用 Microsoft 的 SSIS 工具的 ETL 实现。

第五章分析了本论文的不足并指出了下一步将要进行的工作。

第二章 数据仓库系统的总体设计

20 世纪 90 年代，发达国家的商业银行就已经实现了业务处理的规范化、决策支持的智能化，并发展了以数据仓库技术为基础，以联机分析处理和数据挖掘工具为手段的 CRM 系统。国内的商业银行在这方面发展相对落后，除了四大国有商业银行发展 CRM 系统起步较早外，国内的其他私营股份制银行都没有建立起较为成熟的客户关系系统。因此，虽然在银行数据库中存储了大量的客户信息，但没有一套行之有效的数据挖掘工具对这些数据进行信息分析，甚至对这些数据的分别都有问题；另外，银行也不能够对客户进行分类管理，而是采取无差别的服务，这样就不能找出盈利客户，缺乏对大客户和黄金客户的有效管理。

为了在中国加入 WTO 后所面临的国际竞争环境下，保留住老客户，吸引更多优质客户，国内的商业银行急需建立一套基于数据仓库技术，OLAP 方法和数据挖掘工具的客户关系管理系统。

2.1 国内某银行业务现状分析

该银行成立有十多年，截至到 2006 年底，在全国各个大中型城市建有 23 分支行，并在一些经济发展较快地区建立了直属支行，其机构网点达到近 300 家。作为国内的私有股份制商业银行， 它在成立之处就注重信息化的建设，并逐步搭建了基于全行的科技平台。2001 年，该银行对全行实施了数据大集中，总行各业务部门、各分行的存款、贷款、同业拆借、不良资产等业务动态数据都被实时监控和跟踪，特别是对分支行数据异常的监控很好的防范和化解了潜在的金融风险。在这次数据大集中过程中，将逐步建立 CRM 系统、客户服务中心系统、个人信贷业务系统、授信风险管理系统、会计管理系统、业务流程系统等。虽然，这次数据大集中还是基于传统的 OLTP 系统平台、技术和工具，使得其功能难以随统计分析需求的不断变化而进行调整；但是，这次数据大集中过程中对总行和分行应用版本进行了统一，大大降低了数据整理工作的复杂性，为数据仓库系统的实施打下了坚实的基础。因此，该银行构建数据仓库系统的条件已经成熟；另外，该银行的管理层也充分认识到构建数据仓库系统对于应对目前激烈竞争环境的重要性。

通过构建基于银行业务的 CRM 数据仓库系统，银行业务人员可以分析客户需求，维护和发展目标客户并对客户进行分类和贡献度分析等。比如：哪些是贡献度高的优质客户，哪些是普通客户，哪些是高亏损客户，将分析成功运用在差

别服务、客户定价、渠道迁移等环节上，改善粗放型的经营模式。本数据仓库系统必须为该银行 CRM 中客户信息管理、客户综合分析、目标客户搜索、业务查询和统计功能等提供支持。

该银行业务经过十多年的发展，使得其业务系统中的数据具有海量化的特点；而且，随着业务的不断扩展和客户量的激增，这些数据都在以惊人的速度增长。另外，银行业务的复杂性也使得其统计数据多样、报表种类多，因此，该数据仓库系统将要汇集多个不同业务系统中的数据，包括综合业务系统、信贷系统、中间业务系统、网上银行系统等其他系统。针对该银行的以上业务特点，本数据仓库系统的构建过程中将逐步解决的这些问题。

2.2 系统的设计目标与原则

2.2.1 设计目标

该银行通过基于数据仓库相关技术的客户关系管理对客户进行细分，以更好地满足客户需求，并连接客户和银行来最大化客户的利润和提高客户满意度与忠诚度。该数据仓库系统建设的总体目标是建立银行企业级基础数据平台，整合客户信息资源，实现全面的客户信息管理功能：包括单一客户信息管理、客户综合分析、目标客户搜索和业务查询与统计功能，为该银行提供决策支持的管理信息，以提高其市场竞争能力。

该银行建立 CRM 数据仓库系统的目的是从客户需求出发，及时准确地制定市场决策，不断维护和拓展客户群，同时优化银行内部的资源，提高银行的运作效率，挖掘更多的创收机遇，从而实现收益的持续增长。完整、准确的客户资料是企业实施有效的客户关系管理的前提和基础，只有在深入认识、了解客户的基础上，才有可能实现对客户的有效营销和优质服务。CRM 系统中的客户信息管理所要管理的应不限于客户的名称、地址、联系电话等基本信息，而是要将对于了解客户需求、争取客户提供帮助的各种信息以客户为中心完全管理起来。对客户分类信息进行规范化管理，以便为相关客户分析提供细分客户的标准。按照一定的标准将客户进行分类，识别出每一类客户的基本消费特点，可以获得客户的真实价值和消费特征，为商业银行对客户进行有针对性地营销、销售和服务提供依据。通过建立银行的数据仓库系统，为 CRM 提供全面、准确的数据，以达到支持决策层来做出准确而快速的决策。

数据仓库系统按照两级、三层结构进行建设。两级系统是指系统数据的存储必须划分为总行数据仓库系统和省级分行数据仓库系统两级。三层结构是指系统在逻辑结构上包括数据获取层、数据存储层和数据访问层。

数据仓库系统应该通过即席查询、预定义报表、联机分析处理、数据挖掘等手段实现面向主题的业务处理；并且能根据需要进行主题内部要素的扩充、主题新增和跨主题的重构等。

2.2.2 设计原则

数据仓库系统的开发涉及的知识面较广、技术复杂、集成化程度高，而且是多学科的综合性应用。因此，系统的开发过程是一个经过不断循环、反馈而不断完善系统的过程，这对系统的体系结构的设计提出了很高的要求，要求设计的体系结构具有良好的可扩展性、灵活性，能够适应复杂的业务需求。另外，数据仓库建设不单纯为了数据集成，而是在数据集成的基础上为业务发展提供决策支持。

（1）数据仓库系统的开发需要经过一个较长时期，其中数据的抽取、清洗和加载（ETL）是该系统构建过程中最重要的一个步骤。因此，在系统设计时，要合理规划和分配各种资源和人员，以按期完成系统的建立。数据加载时需要解决数据质量问题，比较了数据仓库系统还是源数据系统中解决之后，发现在源数据系统中解决质量问题较好。

（2）银行 CRM 数据仓库系统作为银行决策支持系统的一个重要部分，建成后将为不同层次的用户提供服务，并与其他系统具有较好的兼容性，因此在系统必须在全行统一规划、部属下实现，系统的整体性必须得以保障。

（3）数据仓库系统的前期开发提出的业务分析需求远远不足，随着业务的和系统不断扩展，因此系统的开发也必须适应业务的不断变化而具有可扩展性，并为用户提供更好的分析工具和对工具的使用提供更好的培训，最终是用户能独立使用数据仓库。

2.3 系统总体构架的设计

该商业银行在大部分省市建有分行和支行，每天需要处理的数据量是很大，只在总行建立一个数据仓库来处理全行的数据太过庞大，因此，在本银行 CRM 数据仓库系统中，我们采用的是分布式数据仓库环境，总行和和省级分行都拥有自己的 CRM 数据仓库系统，并与本地的综合业务处理系统连接，另外也为总行的 CRM 数据仓库系统提供各种数据支持。图 2-1 即为数据仓库系统总行和分行的总体构架图。

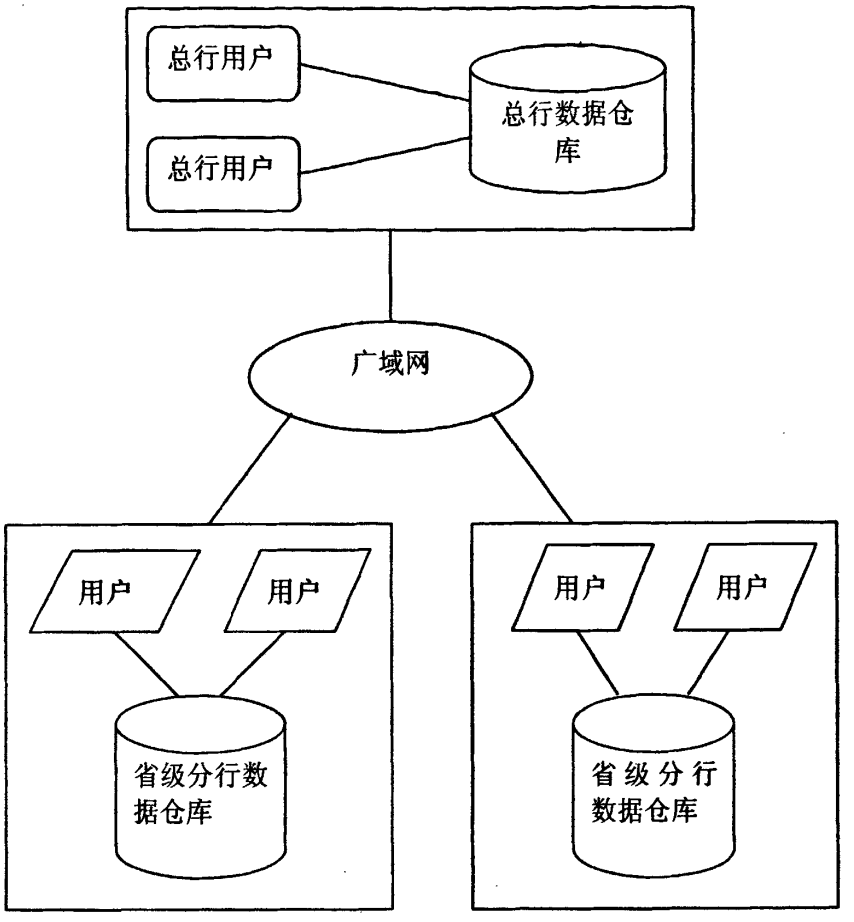

图 2-1 数据仓库系统的总体构架

该数据仓库模式下的总行和分行的数据仓库之间既具有相对的独立性，又不是相互孤立的。总行的数据仓库为全行的决策提供支持，分行的数据仓库为分行的管理层提供决策支持，这就是它们之间的相互独立性。同时，分行的数据仓库通过分布式网络为总行所共享，但是，总行数据仓库中的数据并不是分行数据仓库的简单汇总，针对总行数据仓库系统的特点和分析需求，有选择性的将分行数据仓库中对全行业务有价值的数据进行抽取。通过综合各个分行的数据，更加利用准确而快速的做出决策。

其次，为了保证总行与分行数据仓库系统数据一致性，我们采用了数据仓库总线模型构架，总行和分行数据仓库都在总线模型下构建，从而避免了数据不一致性问题。数据仓库的总线型结构是由数据仓库专家 Ralph Kimball 提出来的，它认为数据仓库的构建是在统观全局的基础上，一步步完成的，以部门级数据集市的构建为出发点，使数据集市成为完整的企业级数据仓库的一个逻辑子集。总

线模型构架必须有统一的维和事实，如统一时间维、统一客户维和具有可加性的数据型事实

2.3.1 系统的体系结构

数据仓库系统的体系结构设计是建立数据仓库系统的一个总体描述，它从宏观和整体角度对数据仓库系统的各组成部分进行总体设计，并确定在设计过程中遵循的原则。因此，体系结构的设计好坏直接影响到后期工作进程，好的体系构架保证了数据仓库的各个部分在开发过程中能够依据同样的基础和标准，为后续的数据 ETL、应用开发、系统管理等工作提供指导原则。

在有关数据仓库体系结构的多种理论中，着眼于体系部件功能的“三层结构”理论得到了最广泛的接受[11][12]。这种数据仓库系统的体系结构构建主要包括三个过程：数据的获取，数据的分析，数据的输出。系统把业务数据合外部数据通过 ETL（抽取、转换、装载）整合到数据仓库中去，再通过 OLAP（联机分析）和 DM（数据挖掘）工具对数据仓库中的数据进行处理，得到对管理层有用的决策支持信息。而根据对于局部与整体关系处理方式的不同，数据仓库的构造又可以归纳为由顶向下等六种构造模式。

一个完整的数据仓库解决方案的体系结构包括：数据源层、数据采集层、数据存储与管理层、门户管理层和最终用户层，概括起来就是三层体系结构。通过以上分析该银行的业务现状和 CRM 系统的需求和目标，该数据仓库系统采用的是三层体系结构。图 2-2 是本数据仓库系统的三层体系结构图。

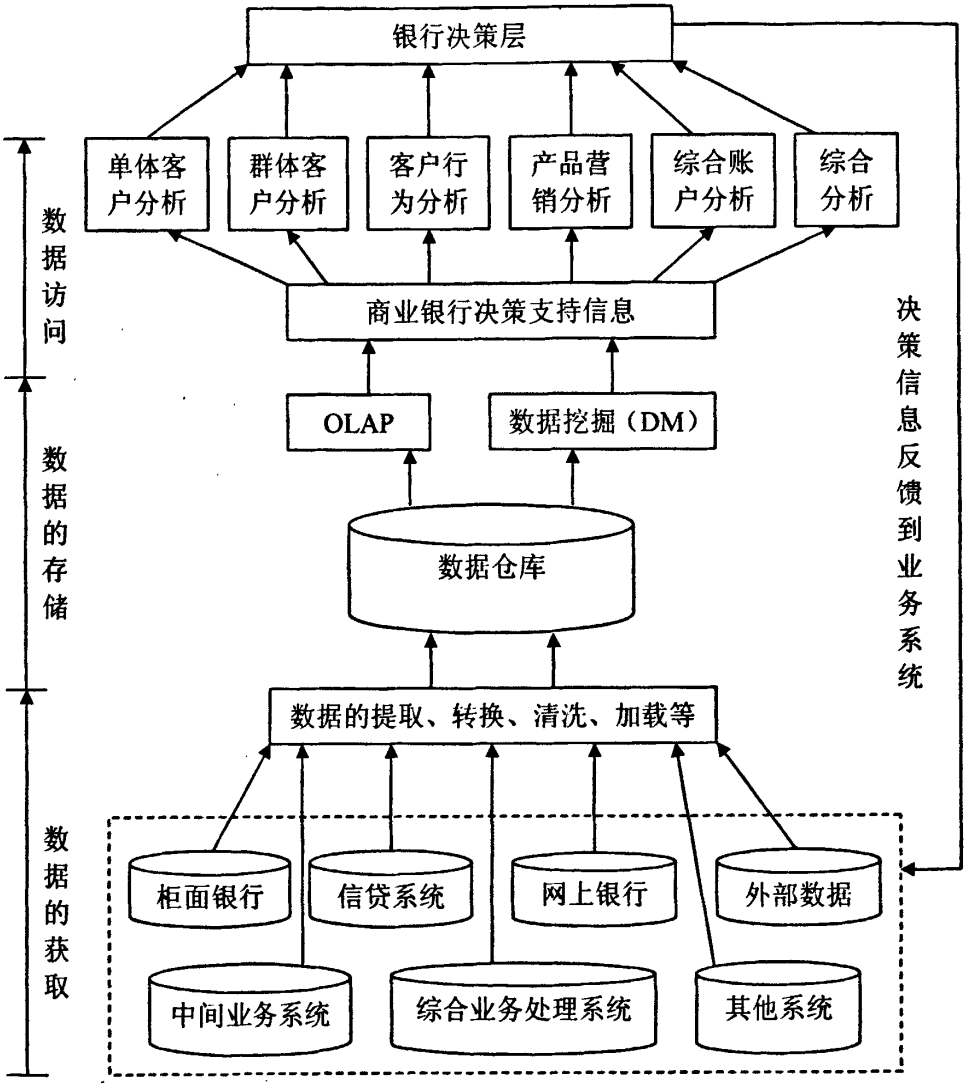

图 2-2 数据仓库系统的三层体系结构

1．数据的获取

数据的抽取过程是数据进入数据仓库的入口，抽取的数据源一般包括：OLTP系统的数据、外部数据源、脱机的数据存储介质等，数据抽取在技术上主要涉及互连、复制、增量、转换、调度和监控等几个方面。

商业银行数据仓库系统中的数据来源于内部数据和外部数据。内部数据是指银行业务核心系统中的日常业务处理数据，包括信贷系统、中间业务系统、综合业务处理系统、网上银行系统中的各种数据。这些源数据经过抽取、清洗和转换、最后加载（ECTL）到数据仓库中。外部数据是指银行业相关数据、社会经济数据、

各种调查数据等有助于银行业务分析的数据。

2．数据的存储

数据仓库在存储管理上有三个问题需要解决：海量的数据、并行处理和查询的优化。数据仓库存储和管理的数据是海量的，比传统事务处理大得多，而且随着时间推移积累速度也越来越快。同时，数据仓库用户访问系统的特点与传统的联机事务处理不同，它的访问是稀疏和庞大，每一个查询或统计都非常复杂，但访问的频率并不高。这时，并行处理和查询的优化对数据仓库显得较为重要。

该银行数据仓库中存储和管理来自各种源数据系统中的数据，并为访问用户和决策层提供数据服务，是整个数据仓库的核心。这些数据按照逻辑数据模型分主题进行组织、重构和存放，包括当前数据和较长期的历史数据。

3．数据输出

用户访问数据仓库的方法与传统关系数据库有很大的不同，对数据仓库的访问往往不是简单的表和记录的查询，而是基于用户业务的分析模式，即联机分析。它将数据想象成多维的数据立方，用户查询相当于在其中的部分维上添加条件，对数据立方进行切片、分割，得到的结果则是数值的矩阵或向量，并将其绘成图表或数理统计的算法等。数据的分析主要是指多维分析、数理统计和数据挖掘等方面。

商业银行 CRM 数据仓库系统的面向的主要是客户经理和其他银行的决策层，数据仓库必须能够满足这些用户的查询需要，还要为用户提供查询、报表、统计分析、数据挖掘等服务。

2.3.2 系统工具选择

该数据仓库系统选择了 IBM RS6000 作为数据仓库的应用平台；Informix Online 7.31 For SCO 作为数据仓库引擎；使用 Microsoft SSIS 为 ETL 工具，本文将在第四章关键技术部分对此做具体的分析；BO Crystal Reports 和 Excel 为前端展现工具。

IBM RS6000： RS/6000 是 Risc System / 6000 的简写，是 IBM 公司使用其 RISC 构架的 Power 处理器设计生产的小型计算机，其通行的官方操作系统是 IBM 的 AIX。Power 处理器具有强大的处理能力，其芯片体系结构的总线由一条 32 位宽的数据总线和一条 64 位宽的数据总线组成，为实现了超标量提供了指令宽带和数据带宽。

Informix：Informix 是世界上主要的数据库厂商之一，其产品具有高效的并行处理能力、共享内存技术和易管理性等特点；目前，该银行业务系统所使用的数据库就是 Informix。Informix 公司除了在数据库市场上占据一席之地外，其数

据仓库解决方案在业界也具有领先技术，如数据库平台、开发工具和应用系统等。特别是其针对金融业的数据仓库行业解决模板在帮助中国建设银行控制信贷风险上发挥了重要作用。Informix Online 7.31 是 Informix 公司推出的新一代数据库引擎，它的目标是使具有多个物理 CPU 和大容量内存的计算机创建高性能和高稳定性的操作环境，目前 Informix Online 7.X 广泛运用于那些对于速度和安全性较高的企业和部门，如银行、电信、保险、邮政等。

BO：Business Objects 是集查询、报表和 OLAP 技术为一体的智能决策支持系统。它使用独特的“语义层”技术和“动态微立方”技术来表示数据库中的多维数据，具有较好的查询和报表功能，提供钻取等多维分析技术，支持多种数据库，同时还支持基于 Web 浏览器的查询、报表和分析决策。另外，使用了 Excel2007 来实现部分查询、报表、多维分析等。

2.4 本章小结

本章主要针对国内某中小银行的业务特点及现状进行分析，确定了该商业银行构建数据仓库系统构建的目标和原则，并给出了数据仓库构建的总体方案，设计了基于中小商业银行数据库的总体构架和体系结构等。

第三章 数据仓库系统逻辑模型及物理模型设计

模型是对现实事物的反映和抽象，是理解、分析、开发或改造事物原形的一种重要手段，它能帮助设计者更清楚的了解客观世界。数据仓库构造的过程中包括三中模型的设计：概念模型、逻辑模型和物理模型。数据模型是数据仓库建设的基础，一个完整、灵活、稳定的数据模型对于数据仓库项目的成功起着重要的作用[28]。

因此，建立数据仓库系统必须先创建数据仓库的模型，数据仓库模型的设计过程中需要完成分割、主题、粒度的设计，及其数据仓库概念模型、逻辑模型和物理模型的设计。概念模型描述的是客观世纪到主观认识的映射；逻辑模型则与数据库中的模式直接相关，逻辑模型设计需要对概念模型中的每一个主题进行设计；物理模型的设计则涉及数据仓库实现的具体细节，如物理存储方式、数据存放位置、索引结构等，这三种模型之间联系紧密，它们之间的关系如图 3-1：

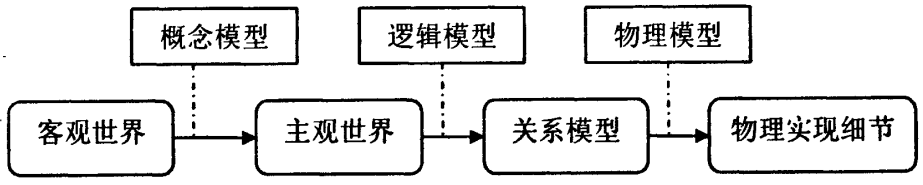

图 3-1 三种模型间的关系图

由于建立数据库与数据仓库的目的不同，它们的设计也并不相同。数据仓库系统是为企业提供一个面向主题的以提供决策目的的分析型系统，而且数据仓库系统的需求往往不太明确，因此，数据仓库建立的过程中，需求往往是不断变化的，使得数据仓库的设计和建模需要具有柔韧性和伸缩性。

3.1 概念模型设计

概念模型的设计主要是确定数据仓库的主题及相互关系，它所要完成的工作主要有二个：

（1）系统边界的界定：进行任务和环境评估、需求收集和分析，了解用户迫切需要解决的问题及解决这些问题所需的信息，即对原有数据库系统有一个完整的认识。本系统的目的是为建立 CRM 系统提供准确而及时的分析数据，以此来分析客户的需求、对银行产品的满意度等，使决策层更准更快的做出决策。目

前的该商业银行有 300 多个表（未包括交易或构件中建立的临时表），这些表主要用来存放银行的账务数据、历史数据、客户信息和登记薄等。根据现有业务的将数据库表共分为以下几类：公共业务、对公业务、结算业务、对私业务、贷款业务、批处理业务、内部帐业务、中间业务等

（2）确定主题域及其内容：确定系统所包含的主题域，然后对每个主题域的公共码键、主题间的联系等有较明确的描述。商业银行的主题包括客户、账户、交易、渠道、营销活动、资产、财务、分支机构和职员等，本系统主要使对客户的消费行为进行分析和挖掘相关信息。由此确定了 4 个基本的主题：客户、账户、交易、产品。

表 3-1 是对主题的描述：

表 3-1 系统主题的描述表

主题名	公共码键	属性组
客户	客户号	客户号，客户名，证件类型，证件号码，联系电话，工作单位，
账户	账户号	账户号，客户号，凭证号，账务机构，营业机构，支付条件
交易	交易码	交易码，交易流水号，交易时间，账户，交易量等
产品	产品代码	产品代码，产品名称，产品类型，产品收益，发行机构，发行价格，发行量等。

以上主题之间的关系如图 3-1：

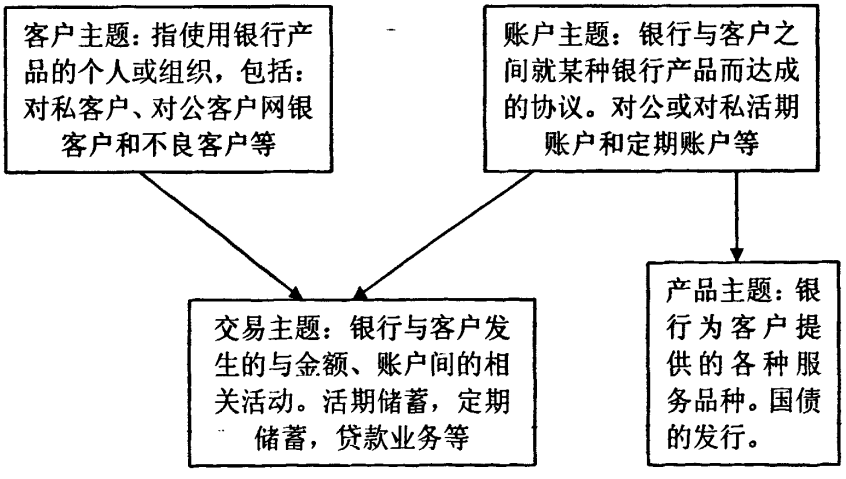

图 3-1 各个主题间关系图

3.2 逻辑数据模型设计

逻辑数据模型（Logical Data Model，LDM）是用来发现、记录和沟通业务的详细“蓝图”。银行数据仓库是一个统一、共享的基础数据平台，为各个业务部门的不同业务需求提供一致、规范的数据。作为数据仓库的基础，逻辑数据模型设计的好坏对数据仓库有着重大影响。数据仓库存储了海量的数据和分析需求不确定，理想的逻辑数据模型应该具有非冗余性、稳定性、一致性、数据使用的灵活性等特征。

3.2.1 逻辑模型的重要性

（1）逻辑数据模型使数据仓库系统建立的导航图，因此通过逻辑数据模型可以清楚的表达内部各种业务主体之间的相关性，使不同业务部门、开发和管理人员获得统一的视图。

（2）逻辑数据模型的建立可以使开发人员清楚了解数据之间的关系及作用，消除了数据仓库中数据的冗余性，通过数据模型，我们可以消除那些只是用来操作的数据，而采集用于分析的数据。

（3）数据模型使整合各种数据源的重要手段，通过数据模型，可以建立起各个业务系统与数据仓库之间的映射关系，实现数据的有效采集。

（4）逻辑数据模型的建立，能够排除数据描述的不一致性，如：同名异义、异名同义等，使得系统的各方参与人员基于相同的事实进行沟通。

（5）便于扩展：逻辑数据模型是对现有信息及信息之间的关系从逻辑层进行了全面的描述，以后业务发生变化或系统需求发生变化时，可以很容易的实现系统的扩展。

3.2.2 逻辑数据模型设计过程

下面将按照以下几步来进行逻辑数据模型设计：分析主题、设计维度表和事实表和选择数据粒度。

主题（Subject）是在较高层次上将银行信息替他中的数据进行综合、归类和分析利用的一个抽象概念，每一个主题基本对应一个宏观的分析领域。在本章第一节的概念模型设计中，已经确定了四个主题：客户，账户，交易，产品，下面将从这四个方面分析。银行的主题资源是客户，客户通过交易与产品（包括账户）发生关系，客户关系管理系统的目标是对客户进行类分，以挖掘优质客户，实施一对一的服务。因此，在本数据仓库系统中，账户、交易和产品都是以围绕客户

主题来建立。

客户主题

客户信息在原有数据库中根据客户类型的不同而存储在不同的表中，主要从以下三个方面来分析

（1）客户基本信息

客户的基本信息主要有三种类型

A．对私客户：客户号，个人中文名，性别，出生日期，证件种类，证件号码，电话号码，联系地址

B．对公客户：客户号，企业组织机构代码，客户中文名，客户简称，电话号码，传真号码，办公地址，执照号码，启用日期，法人代表身份证号码等

C．其他客户：客户号，客户名……

（2）客户账户信息

客户账户信息包括：账户号、账户类型、开户时间、销户时间、开户行、余额等

（3） 客户背景资料

客户内部资料：客户号，客户消费水平，客户服务等级，客户重要纪念日、客户地域构成，客户资本构成等

客户业务发展情况：客户号、客户业务、客户的市场地位、业务发展方向等。

围绕客户关系管理而建立的主题还有交易、账户、机构、日期、产品和渠道，因此，本系统还建立了这些主题的维表。

账户维表：账户号，货币代号，营业机构号，账户序号，科目号，开户日期，销户日期，起息日，到期日等。

机构维表：机构号，机构管理级别，机构营业级别，管理上级，账务上级，机构名称，机构类型，启用日期等。

产品维表：产品代码，产品名称，产品类型，产品收益，发行机构，发行价格，发行量等。

交易维表：交易码，交易流水号，交易时间，账户，交易量等

图 3-2 即为围绕客户主题来展开分析的逻辑模型设计图。

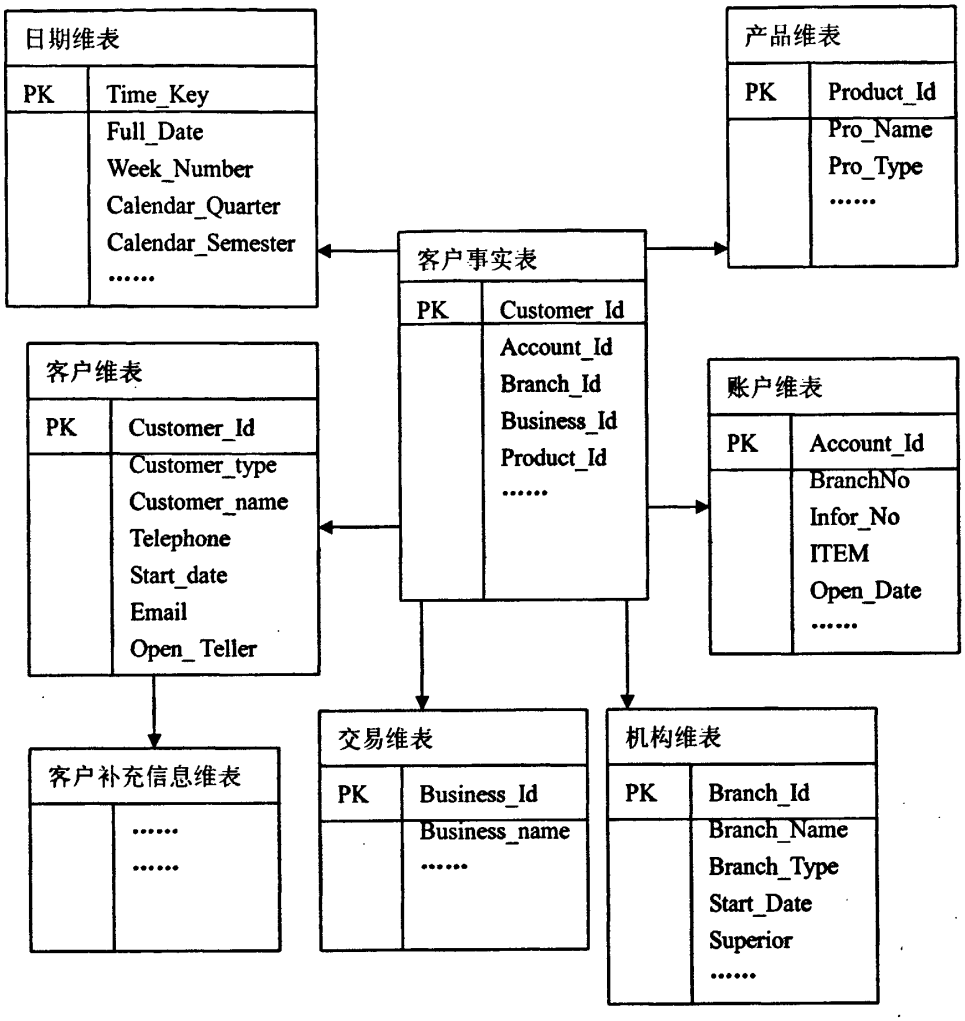

图 3-2 客户主题 LDM 设计图

原有数据库中还有很多表涉及到客户信息，如客户补充信息表等；而且表中涉及到的客户信息字段比较多，必须在后期的 ETL 来处理。

3.2.3 粒度设计

粒度是指数据仓库中数据单元的详细程度和级别，数据越详细，粒度就越小，级别也就越低；数据综合度越高，粒度就越大，级别业就越高。数据仓库的数据分为 4 个级别：早期细节级、当前细节级、轻度综合级和高等综合级。源数据经过综合后，首先进入当前细节级，并根据具体需要进一步综合，进入轻度综合级甚至高度综合级，老化的数据将进入早期细节级。数据的粒度和存储空间是一队

矛盾体，因此数据的粒度不能太大也不能太小。表 3-2 是数据粒度的大小与数据综合度、细节的所示的比较。

表 3-2 客户主题粒度级别性能比较

表名称	粒度级别	综合度	数据细节	后果
客户成交明细	低	低	最高	针对事务记录，无损失
客户成交日明细	次低	次低	较高	损失交易细节，无法揭示当日的买卖行为与行情变化的关系
客户成交月明细	较高	较高	次低	细节损失较多，无法作为客户交易行为分析的数据
客户成交年明细	最高	最高	低	高度概括，只对一些统计分析功能有意义

由于数据仓库种的数据存储是通过存取索引实现的，而索引的组织是针对表中的行来完成的，即在每个索引中，每行必有一个索引项。因此，索引的大小与划分粒度的最终依据不是表的体积规模，而是表的总行数。考虑到本系统只是用于 CRM 的数据分析，把数据仓库的粒度设定为较高级别。

3.3 物理模型设计

数据仓库的物理模型就是数据仓库逻辑模型在物理系统中的实现模式，主要解决数据从存储结构、数据的索引策略、数据的存储策略、存储分配优化等问题。与传统的事务处理数据库相比，数据仓库需要处理的数据量更大，而且随着时间的推移不断有新数据加入。因此，物理设计的目的有两个：一是提高性能，其次是更好地管理存储数据。

3.3.1 事实表与维表的设计

在物理数据模型设计过程中，首先把逻辑数据模型设计过程中确定的事实表和维表转换为物理存储结构的表。在逻辑模型确定了客户、账户、机构、日期、交易等主题，因此，在此阶段针对以上主题设计了客户事实表、账户信息维表、客户信息维表、机构维表、日期维表等，表 3-3 至表 3-7 即是这些表的详细设计情况

表 3-3 客户事实表

字段名称	中文含义	数据类型	键	是否为空
Customer_FactID	编号	Char(10)	PK	N
Customer_Id	客户号	Char(10)		N
Branch_No	分行代码	Char(8)		N
Time_Key	时间编号	Char(6)		N
Account_No	账号	Char(20)		N
Product_Id	产品编号	Char(10)		N
ID_Type	证件类型	Char(2)		N
ID_No	证件号码	Char(20)		N
Start_Date	启用日期	Date		N

表 3-4 客户信息维表

字段名称	中文含义	数据类型	键	是否为空
Customer_Id	客户号	char(10)	PK	
Customer_type	客户类型	char(2)		N
Customer_name	客户名称	char(62)		N
Telephone	电话号码	char(15)		
Office_add	办公地址	char(62)		
ID_Type	证件类型	Char(2)		N
ID_No	证件号码	Char(20)		N
Start_date	启用日期	char(8)		
Edd_date	有效日期	char(8)		
CY_NO	货币代号	char(2)		
OpenLicence_No	开户许可证号码	char(16)		
ForeignExchange_No	外汇许可证号码	char(16)		
Office_PostalCode	办公地址邮政编码	char(7)		
Email	E_MAIL 地址	char(42)		
Open_Branch_Type	开户机构代号	char(4)		
Open_Teller	开户柜员	char(8)		

表 3-5 账户信息维表

字段名称	中文含义	数据类型	键	是否为空
Account_No	帐号	char(20)	PK	N
BranchNo	营业机构号	char(4)		N
UTNO	帐务机构号	char(4)		N
CYNO	货币代号	char(2)		N
ITCD	业务代号	char(3)		N
Customer_No	客户号	char(10)		N
Name	客户名	char(22)		N
Infor_No	信息代码	char(50)		N
ITEM	科目号	char(5)		N
Open_Date	开户日期	Date		N
Modify_Date	维护日期	Date		N
Accout_Type	账户类别	Char(2)		N
Month_AvgBal	当月平均余额	Decimal(13,2)		N
Quarter_AvgBal	当季平均余额	Decimal(13,2)		N
Year_AvgBal	当年平均余额	Decimal(13,2)		N

表 3-6 日期维表

字段名称	中文含义	数据类型	键	是否为空
Time_Key	日期代码	Char(6)	PK	N
Full_Date	日期	Date		N
Week_Number	周数	Char(3)		N
Month	月	Char(3)		
Calendar_Quarter	季度	Char(1)		
Calendar_Semester	半年	Char(1)		
Calendar_Year	年	Char(5)		

表 3-7 机构维表

字段名称	中文含义	数据类型	键	是否为空
Branch_No	支行代码	Char(5)	PK	N
Branch_Name	支行名称	Char(10)		
Branch_Type	支行类别	Char(1)		
Start_Date	启用日期	Date		
Superior	上级机构	Char(5)		

3.3.2 存储结构

在物理设计时，一般要按数据的重要性、使用频率及对反应时间的要求进行分类，将类型不同的数据分别存储在不同的存储设备中。重要性高、经常存取并

对反应时间要求高的数据存放在高速存储设备上；存取频率低或对存取响应时间要求低的数据则存放在低速存储设备上。数据的存储主要有两种：分布式存储和集中式存储。

（1）分布式存储

分布式存储是指采用磁盘阵列在多个节点间分布的方式来存储数据。节点间可以通过互连 I/O 进行通信，这些磁盘阵列中存储的数据是可以相互访问的，因此，数据是可以完全共享的。

（2）集中式存储

集中式存储中所有的数据所有节点式共享的，数据的读取可以不通过节点间内部高速互连网络。因此，这种方式的好处是可以将节点从数据存储管理负担中解脱处理，实现数据存储和数据处理的分离，另外不会占有节点间的内部通信带宽。

比较两种方式，集中式存储更有利于中小型商业银行数据仓库系统数据的存储。其次，由于在对服务器进行处理时往往要进行大量的等待磁盘数据的工作，系统中加入 RAID（Redundant Array of Inexpensive Disk，廉价冗余磁盘阵列）的使用。

数据仓库的存储结构是为海量信息设计的，因此这些数据通常存储在多个硬盘上。如果遇到硬盘错误或硬盘损坏，技术用户经常作数据备份，重建硬盘和数据恢复工作也需要花费大量金钱和时间。而 RAID 是一种使用多磁盘驱动器来存储数据的数据存储系统，可以使用多种不同的存储技术来实现不同等级的冗余、错误恢复。当遇到磁盘错误时，数据仓库系统还可以继续运行，数据也不会丢失。因此，RAID 是一个符合服务器大容量硬盘和数据安全性、服务器运行速度等综合要求的廉价解决方案。商业银行对数据的准确性和容错能力是相当高，在银行的数据仓库系统中出现数据丢失所带来的损失会直接转化为经济上的损失。

除了这种 RAID 特性外，还通过购买硬件组成一个驱动器阵列，因为这种硬件实现比软件实现效率更高。

3.3.3 索引策略

索引是根据指定的一列或多列的内容对行进行排序，索引主要用于提高查询的效率。银行数据仓库存储了海量的数据，另外数据仓库中的数据一般很少更新，所以需要对数据的存取路径进行自己的设计和选择，通过建立索引来提高数据的读取和检索效率。数据仓库中的表一般要建立更多的索引，表中应用的最大索引数应该与表格的规模成正比。

1．在数据仓库的中常用的索引策略有以下几种：

（1）B 树索引

B 树索引对于 Product_Id 或者 Customer_Id 这样的高基数数列来说很有用，在许多关系数据库管理系统中经常用到 B 树索引。而且，大部分关系数据库管理系统都会为表中已经声名的主键自动创建唯一的 B 树索引，并通过这种索引形式进一步增强对主键的约束。早期的关系数据库管理系统只能建立一种索引机制，但是采用 B 树类的索引，对于性别、年龄、地区等具有大量重复值的字段几乎没有效果。目前，大多数关系数据库厂商在他们的产品中都包含了查询优化器，从而可以建立多种索引机制。

（2） 位图索引

由于 B 树索引也有它的缺陷，在本系统中不但建有 B 树索引，还加入了位图索引。关系数据库则引入的位图索引机制，以二进制位表示字段的状态，将查询过程变为筛选过程，单个计算机的基本操作便可筛选多条记录。位图索引与 B 树索引刚好相反，它更适用于低基数的列创建索引。位图索引实质上示为列中的所有可能值分别建立一个位串，位串中的每一位分别代表该列在某一行中的取值。因为位图索引通常仅针对单个列建立索引，使得数据优化器在查询中必须使用多个位图索引。

（3） 连接索引

关系数据库管理系统除了用到 B 树索引和位图索引，还有一种连接索引的使用也比较广泛。连接索引是指将事实表和维表中的索引项进行连接运算，然后将结果作为索引保留。虽然索引建立过多会使数据处理的速度下降，但是连接索引项比整个记录条目要小，也比直接对事实表和维表进行连接的结果集要小得多，所以，连接索引的引入还是对数据处理效率 Hash 索引起到较大的作用，同时这样也有利于数据的装载。图 3-3 是围绕客户主题建立的连接索引图。

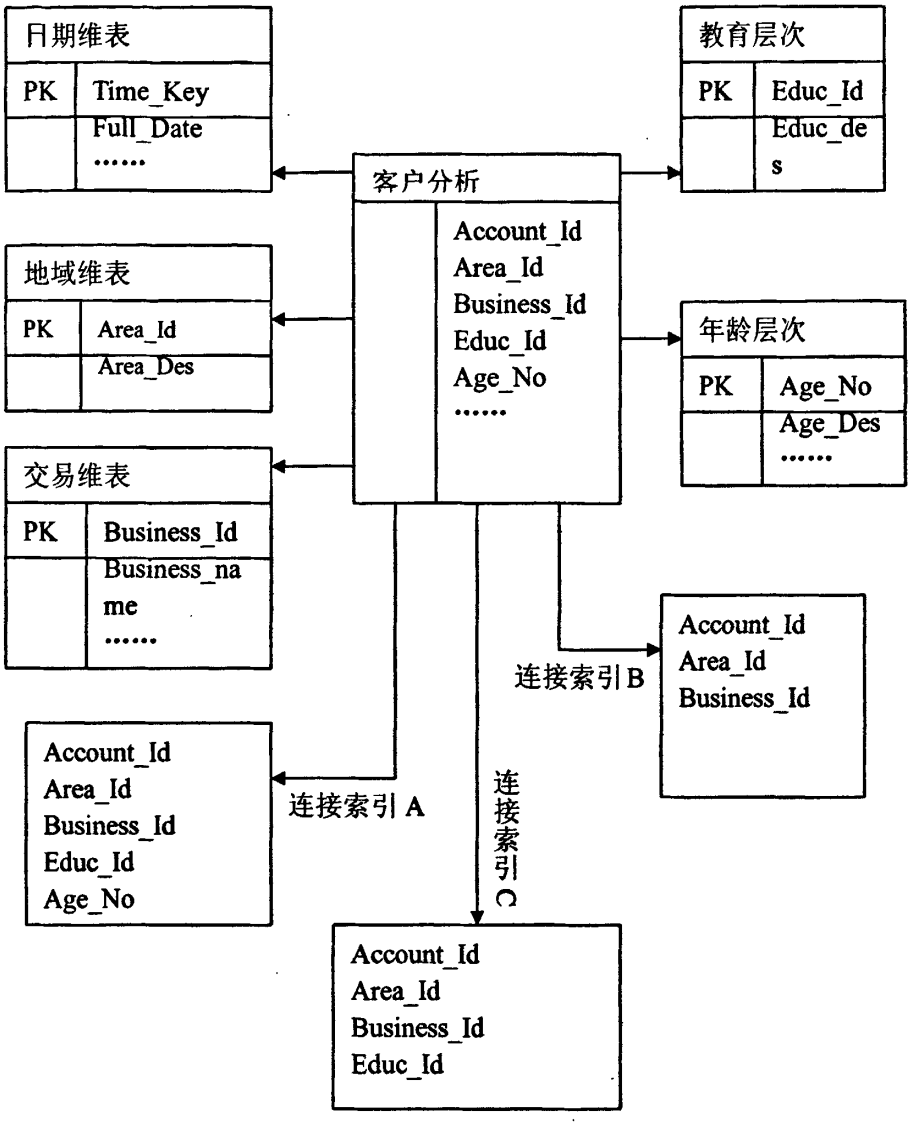

图 3-3 客户主题连接索引图

（4）Hash 索引

部分关系数据库管理系统使用到了 Hash 索引，Hash 索引是指采用 Hash 算法和简洁格式来表示复杂事务。Hash 算法将一组字符代表文本串，并对原有极大的数据进行压缩。Hash 索引描述数据的形式比较简洁，用它来查找数据行的效率较高。

2. 为事实表创建索引

在前面的逻辑数据设计中，确定了客户事实表，这节内容中将为该事实表的

主键创建 B 树索引。在为数据仓库事实表的主键声明约束时，应该按照这些列的声明次序创建一个唯一索引，因为一般来说大部分的优化器对主键的索引功能特别强大。为了在查询时能够充分利用这些主键索引，系统为事实表设置了主键并建立了主键索引。对客户的相关数据分析时会充分考虑时间的因素，因此，系统将日期键放在主键的最前面。这样做的另外一个好处是能够加快数据仓库的维护过程，因为数据仓库的增量式装载过程是以日期为依据的。表 3-8 就是客户事实维表的索引。

表 3-8 客户事实表索引设计表

索引名称	索引类型	是否唯一	列	创建理由
Customer_fact				
Customer_Id	B 树索引	Y	Customer_Id Branch_No Time_Key Account_No Product_Id	主键索引
Cf_Customer_Id	位图索引	N	Customer_Id	在大多数星型连接用户查询使用
Time_Key	位图索引	N	Time_Key	在大多数星型连接用户查询使用
Account_No	位图索引	N	Account_No	在大多数星型连接用户查询使用
ID_ No	位图索引	N	ID_Type	在大多数星型连接用户查询使用
Product_Id	位图索引	N	Product_Id	在大多数星型连接用户查询使用

3．为维度表创建索引

维度表中有一个单列的主键，首先必须为该主键建立一个唯一的 B 树索引。另外，也应该为非选择性的维度属性建立单列的位图索引，这些维度索引引用将作为过滤或者标题使用。

在确定完事实表的主键索引后，还要为建立其他索引项。目前的数据仓库系统中允许在一个查询中同时使用一个表的多个索引，我们可以为可能用到的事实表键创立多重复的索引，在本系统中是为事实表的每个键都建立一个单列的索引，这样并不会对处理效率造成很大的影响，因为优化器会按照最适合于解决查询问题的方式去组合这些索引。表 3-9 为各个维表的索引设计表

表 3-9 为各个维表的索引设计表

索引名称	索引类型	是否唯一	列	创建理由
Calendar				
Time_Key	B 树索引	Y	Time_Key	主键索引
Full_Date	位图索引	N	Full_Date	用于维度分析
Week_Number	位图索引	N	Week_Number	用于维度分析
Month	位图索引	N	Month	经常用于浏览有限的 date_key
Quarter	位图索引	N	Calendar_Quarter	用于维度分析
Semester	位图索引	N	Calendar_Semester	用于维度分析
Year	位图索引	N	Calendar_Year	用于维度的浏览、过滤和分组操作
Customer				
Customer_id	B 树索引	Y	Customer_id	主键索引
Customer_type	位图索引	N	Customer_type	可用于区分不同类型客户
Start_date	位图索引	N	Start_date	常用于过滤器条件中
Edd_date	位图索引	N	Edd_date	常用于过滤器条件中
CY_NO	位图索引	N	CY_NO	用于维度的浏览、过滤和分组操作
Open_ Branch_Type	位图索引	N	Open_ Branch_Type	用于维度的浏览、过滤和分组操作
ID_No	B 树索引	N	ID_No	用于维度的浏览、过滤和分组操作
…… ……				

3.3.4 存储策略

确定好数据的存储结构和表的索引结构之后，为提高 I/O 系统的效率，需要进一步确定数据的存储策略。由于同一个主题的数据可能放在不同的介质上，在

这种情况下，如果这个主题的数据读取频率比较高，系统的I/O设备的负担将比较高，而且效率也会比较低。因此，系统必须按照数据的重要程度、粒度、使用频率和响应时间等要求将数据分别存放在不同的存储设备上。使用程度高、比较重要或响应时间要求高的数据存放在高速存储设备上，存取频率低或对存取响应时间要求低的数据则存放在低速的存储设备上。

为优化存储，本系统还采用了下列策略。

（1）合并表

当几个表的记录分散在几个不同的物理设备中时，多个表的存取操作的效率将会很低，这时可以将需要同时访问的表在物理上顺序存放，或者通过公共关键字将它们相互关联的记录放在一起。如客户基本信息表和账户信息表，在查询时，用户一般会查询某个客户的某个账户的消费信息。数据仓库系统中可以将这两个表存放在相邻的物理块中，则需要同时访问这两个表时，可以大大减少磁头定位的时间，从而提高I/O的效率。如表3-10至3-12所示：

表3-10 客户信息表

客户号	客户名称	证件号码	……
1000023	张三	430……	
1003022	李四	101……	
1020054	王五	201……	
……	……	……	

表3-11 账户信息表

账户号	客户号	账户类型	……
10230235	1000023	对私定期	
20230235	1003022	对公账户	
11230235	1020054	对私活期	
……	……	……	

表3-12 物理存储块

……	张三	430 ……	……	10230235	对私定期	……	……
……	李四	101 ……	……	20230235	对公账户	……	……
……							

（2）引入冗余

一些表的某些属性可能在许多地方都要用到，将这些属性复制到多个主题中，能够处理时减少存取表的个数，如在查询账户信息表时，用户不仅仅是为了

查询到相关账号所有者的客户号，还需要客户的名称等信息。由于系统查询账户信息表的频率很高、处理数据量大，因此如图 3-4，系统可以把客户名称信息加入到账户信息表中。

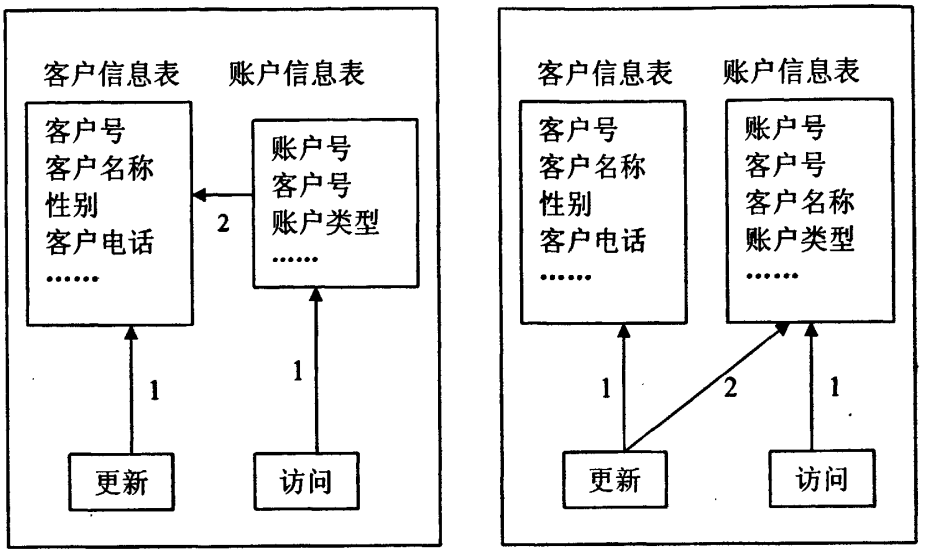

图 3-4 加入冗余信息的账户信息表对比图

（3）建立数据序列

在数据仓库中，经常按照某一个固定的顺序访问并处理一组表，但这些表最初可能分布在不同的物理块中，如果把数据按照处理顺序存放在连续的物理块中，形成数据序列。

（4）分割表的存放

数据仓库中分析的数据量巨大，为了便于访问数据，在逻辑设计中按照时间、分行、业务类型等多种标准把一个大表分割成许多较小的、可以独立管理的小表。这些表分割后，可以采用分布式的存储方式，把时间它们写进不同的磁盘阵列中。

3.4 本章小结

数据仓库建模市数据仓库构造工作正式开始的第一步，正确而完备的数据模型是用户业务需求的体现是数据仓库项目能否成功与否最重要的技术因素。本章在总体设计的基础上，完成了数据仓库构建基础部分即模型设计，包括概念模型、逻辑模型和物理模型的设计。

第四章 数据仓库系统关键技术研究

数据仓库中的数据来自于内部业务系统和外部数据源，银行的业务系统数据库中的数据结构、存储平台等具有很大的异构性，外部数据源则更加。数据仓库中的数据是面向主题方式来组织的，而业务数据库中的数据一般是围绕着一个或者几个业务处理流程组织的。因此，把业务数据库中的数据抽取、加载到数据仓库系统的过程并不是简单的过程，为了得到一致、干净的数据，往往需要对源数据进行复杂的处理才加载到数据仓库系统中。可以这样说，数据仓库的构建过程中，工作量最大（一般要占整个系统的 60%－80%）[32] [33]、日常运行中问题最多的任务是从银行业务数据库数据和外部数据向数据仓库系统中抽取、转换和加载，即 ETL（Extract－Transformation－Loading）。因此，ETL 设计是整个数据仓库系统构建过程中的关键步骤，ETL 过程的准确、高效是保证一个数据仓库数据准确、正常运行的关键。

4.1 ETL 技术讨论

4.1.1 ETL 的定义

ETL 是指数据从业务数据库系统抽取转化到数据仓库的过程，包括以下四个步骤：数据抽取（Extract）、数据清洗（Clean）、数据转换（Transformation）和数据加载（Loading），这就成了 ECTL 过程[57]，但是在部分数据仓库构建过程中数据清洗工作一般很少进行，实际就变成了 ETL，但在本数据仓库系统的源数据处理中必须对其进行清洗。

数据抽取：是指从业务数据库系统或外部数据系统中抽取源数据的过程。典型的数据抽取接口包括数据库接口和文件接口，对于不同源数据、不同性能的业务系统和不同数据量，数据仓库系统可以采取不同的接口。数据抽取的效率是该步骤中必须考虑的一个重要因素。

数据转换：是指把抽取的数据按照数据仓库系统模型的要求，进行数据的转化、拆分、汇总等处理，保证来自不同系统、不同格式的数据具有一致性、完整性。

数据清洗：数据清洗是为了获得一个适当的、统一的格式和定义。

数据加载：是指把经过前三步处理的数据装载到数据仓库系统的过程，在这一过程中，对装载工具的性能具有较高的要求。

4.1.2 ETL 的重要性

为了保证决策的及时准确，数据仓库中的数据必须清洁的，并具有良好的结构，这就是为了保证数据质量。数据质量包括数据的一致性、正确性、完整性和可靠性等。一致性：迁移后的数据和源数据保持一致；正确性：源数据要保证准确无误的迁移；完整性：在迁移过程中不能发生某一条数据被截断或遗漏。

但是，由于数据仓库中数据的来源不同（有业务数据源、外部数据源等），使得加载前的数据具有大量分散的不清洁的特点，无法直接进行加载。通过 ETL 处理能够解决业务数据库中数据模型不同、编程接口不同和数据的不兼容等问题。ETL 能够根据企业决策的需求，数据仓库将决策分析用的数据集中在一起。数据集市的数据是按照部门从数据仓库中抽取并进行处理的，使用 ETL 工具，能简化操作和提供效率。

在数据仓库的构建预算的三分之一被用于 ETL 工具与数据清理工具上，而且有 80%的时间被用于 ETL 过程的建立和执行上，在数据仓库运行代价中 ETL 过程则占 55%[40][22]。ETL 过程在整个数据仓库构建过程中的位置如图 4-1 所示，它是业务数据和其他外部数据进入数据仓库或数据集市的中间步骤，起着承上启下的作用。

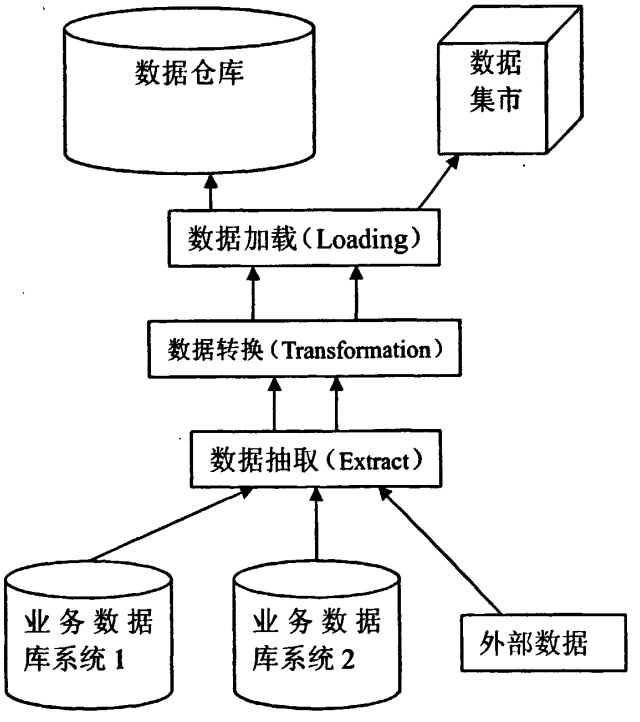

图 4-1 ETL 在数据仓库系统中的位置

因此，ETL 是数据仓库构建的核心过程，它按照统一的规则集成和整合并提高数据的价值，是负责完成源数据向目标数据仓库转化的过程。它是增进维护数据仓库的驱动力，是保障数据仓库数据质量的关键，是实施数据仓库的重要步骤。

4.2 ETL 工具的选取

由于 ETL 在整个数据仓库系统构建过程中的重要性，许多关系数据库厂商和专业 ETL 工具厂商都提供了各种 ETL 的解决方案工具，有些工具甚至是开源的。

目前，主流的 ETL 厂商提供的 ETL 产品有：Ascential DataStageXE、Informatica PowerCenter 等，这类产品提供了功能复杂和详尽的 ETL 处理，但价格一般都比较高。另外，大多数提供数据仓库存储、设计、展现工具的关系数据库厂商也提供了相应的 ETL 工具，如：IBM 的 Warehouse Manager、NCR Teradata 公司的 ETL Automation、Oracle Warehouse Builder 和 Microsoft SISS 等，这些工具对自己厂商的相关产品有较好的支持，但对其他厂商的支持相对有限。下面将分析 Informatica Powercenter、IBM Warehouse Manager、Microsoft SISS 三个产品的 ETL 处理。

（1）Informatica PowerCenter

Informatica PowerCenter 不仅是世界级的企业数据集成平台，也是业界领先的 ETL 工具。数据整合引擎 Informatica PowerCenter 拥有一个功能强大的数据整合引擎，所有的数据抽取转换、整合、装载的功能都是在内存中执行的，不需要开发者手工编写这些过程的代码。Informatica PowerCenter 的数据集成平台具有可伸缩性和扩展性，可以与 Informatica PowerConnect 一起使用。PowerCenter 提供了对特殊数据源和格式的支持，包括 SAP、Siebel、PeopleSoft、AS400 等。PowerCenter 的高性能运行能力将设计和运行环境的性能特性分离，提供了较好的灵活性，不需要重新编码，吞吐量可以通过服务器、并行引擎管理、最优化 CPU 资源等方式，尽快处理任务。PowerCenter 具有分布式体系结构，也可以单独部属。在分布式体系结构中部属时，PowerCenter 要协调和管理多个基于主题的且分别在局域网或广域网中的数据集市，这些数据集市由 Informatica PowerMart 或 Informatica PowerCenter 引擎执行的。另外，PowerCenter 通过提供 PowerCenter for Remote Data 来提高性能、安全性等。

在具体实现上，PowerCenter 通过 Mapping 表示一个 ETL 过程，运行时为 Session，绑定了具体的物理数据文件或表。在修改维护上，这两个工具都是提供图形化界面，这样的好处是直观、傻瓜式的，缺点是改动时比较麻烦（特别是批量化的修改）。

（2）IBM Warehouse Manager

IBM 的 DB2 在关系数据库市场上占据了一定的份额，其 ETL 工具和方案也占有较大的优势。IBM Warehouse Manager 在 ETL 上的优点有数据源广泛，能在大量数据的抽取中具有速度优势，提供编程接口、Cube 处理和调用外部程序的功能，还提供 agent 把数据抽取分布到工作站、小型机、大型机等各种平台。

（3）Microsoft SSIS

Microsoft 公司的 SQL Server2000 中就提供了 DTS 服务来进行 ETL 处理，虽然 DTS 能从广泛的数据源抽取数据，并提供有效的编程方式和工作流任务处理方式等，但它就数据仓库要求处理的海量数据环境而言，其处理数据的量是有限的。因此，Microsoft 在 SQL Server 2005 中推出了 SSIS（SQL Server Integration Services），SSIS 是从中的 DTS 服务升级而来的。

SISS 的功能非常强大，具有企业级数据整合工具的高性能，支持复杂数据流程设计、各类复杂数据源、Web Services 与 XML 等高级技术，而且能够进行动态调试和断点等。SSIS 具有可视化调试功能，这样开发人员能够比较清楚的看到数据管道的工作状况，另外，通过断点、变量和调用堆栈提供了非常强大的调试功能。在数据连接上，SSIS 除了对文本文件、OLE DB 和 ADO.NET 等常见的数据进行访问外，还简化了访问 SAP 中数据的方式。对 XML 和 Web Services 的支持也使得与面向服务的构架及其他非标准数据源的整合变得更轻松。SSIS 中的转换具有较高的效率，能进行数据类型、格式转换和字段解码等，在高级组件还简化了其他复杂的操作，如缓慢变化维度的装载。

在比较了 Informatica PowerCenter、IBM Warehouse Manager、Microsoft SSIS 和 Oracle 等 ETL 工具之后，并结合该银行的业务特点。另外，在这些工具的应用中，研究者普遍认为 Microsoft 公司的 SSIS（DTS）较为出色，因为它具有良好的易用性和可扩展性，使编程效率最高的数据集成工具[53]。因此，在本数据仓库系统中选择了 Microsoft SSIS 作为 ETL 工具。

4.3 数据抽取

4.3.1 数据抽取方式

源数据一旦经过抽取和发布，就可以供数据仓库所用，因此这个过程对数据处理的好坏将直接影响到源数据是否能正确进行转化[39][40]。该银行在全国各地有大量的分行，而且该银行目前有多个业务系统，包括综合业务系统、信贷系统、中间业务系统、网上银行系统等其他系统。数据主要是从业务数据库系统中抽取，由此可见抽取工作涉及到访问数十个分布的数据源，抽取量将十分巨大。如何在

这些异构的数据源中获取一致的数据是抽取的一个比较重要因素。

数据抽取的更新方式主要表现在两个方面：

（1）增量更新和批量更新

（2）实时更新和周期更新

在初次抽取业务数据库中的数据时，由于数据量的巨大，只能采取批量更新的方式。初次加载后，当数据源中的数据发生变化时，若采用刷新方式，会大大增进网络负载和处理的开销，所以在数据仓库数据的更新除了少数表需要采用全量更新外通常采用增量更新。

银行每天都会办理大量的各种业务，因此，业务数据库系统中的数据也会随时间有不断的变化。当数据仓库系统中的数据随着业务数据库的数据而实时变化，这种方式就是实时更新；按照一个固定的周期将数据源中的数据抽取到数据仓库的方式就叫周期更新。如果数据仓库中的数据每天直接从银行业务系统中实时更新，将严重占据业务数据处理的带宽，这样将影响银行业务的正常营业，而周期更新的开销比实时更新低很多。另外数据仓库中保存的主要是历史数据，分析的对象也是历史数据，除非是实时性要求很高的需求，数据仓库中一般不会包括刚刚更新的数据，对分析结果不会有影响。

该数据仓库系统只要求对远期或近期的历史数据进行分析和挖掘，而且，银行每天都会办理大量的业务，使得业务数据库系统中的数据也会随时间有不断的增长。其次就是银行业务处理时对整个业务处理系统的性能有很高的要求，不难想象如果在处理某些业务时等待的时间太长，客户对银行将会是怎样一种态度。而采用实时更新的方法无论更新数据量的大小都将会影响银行业务数据库系统的性能，特别是在大量数据更新时还有可能会产生致命的影响。因此，该系统主要采用周期更新的方式，而不采用实时更新的方法。

抽取方式确定之后，我们还要根据该银行业务现状确认周期更新方式的周期。目前该银行正处于高速发展期，通过各种优惠措施（如贵宾卡客户积分、跨行办理业务不收取手续费）来吸引新老客户来本行办理业务，所以各个业务数据库系统中每天增加的数据量非常大。这样，该数据仓库系统的数据抽取周期不能太长，否则会使得每次抽取的数据量过大，不利用保证这些数据的一致准确性等，本系统中选取的周期为每周一次。

银行业务处理是全天候的，抽取时间的选取必须选择在业务相对较少的时候，月末、季末该银行有大量的业务需要结算，另外，周末银行处理的业务量也会较多，因此该系统数据抽取的时间不能选择在这些时刻，我们选择了周一到周五业务处理少的时候。而且，在这个时间段中，相对来说晚上业务处理不会有太大的量，该数据仓库系统的周期数据抽取，可以根据上面确定的周期在当天晚上银行核心业务系统等进行日终批处理之后，将数据按照一定的接口格式转换为文

本格式，将总行需要的数据由分行的传送过去。

4.3.2 数据接口

该银行数据仓库的数据源来自于不同的业务系统，如网上银行、信贷系统、综合业务处理系统、中间业务系统等，为保证能顺利完成相关数据的抽取工作，屏蔽个别源系统物理上的差异，在数据抽取前需确定数据接口。接口数据文件终包含了相应数据接口单元本次抽取的数据内容，其名称、文件大小、记录数和数据日期应该与接口校验文件一致。

表 4-1、表 4-2 分别是“客户”接口数据文件和接口单元属性列表。

表 4-1 是“客户”接口数据文件

数据内容
接口单元名称：客户 接口单元编码：01 接口单元说明：客户包括现在已经在银行开户的对公、对私、网银和不良客户，其中包括以前在本行办理过业务的这两类客户。

表 4-2 “客户”接口单元属性列表

属性编码	属性名称	属性描述	属性类型	备注
0	记录行号	唯一标识记录在接口数据文件中的行号	Number(8)	
1	客户标识	为每个客户分配的唯一标识	Char(15)	
2	机构号标识	为客户所在分行设立的唯一标识	Char(8)	
3	客户姓名	客户开户时填写的姓名	Char(22)	
4	客户英文名	客户的英文名和中文名拼音	Char(32)	
5	证件类型	国家对证件类型的统一编码	Char(3)	
6	证件号码	唯一标识证件的号码	Char(20)	
7	电话号码	客户开户时填写的联系电话	Char(15)	
8	手机号码	客户开户时填写的手机电话	Char(15)	
9	通信地址	客户的住址或联系地址	Char(62)	
10	邮政编号	客户通信地址的邮政编号	Char(6)	
11	启用日期	客户开户的日期	Date	
12	客户状态类型编码	表示客户状态，现在是否已经销户	Char(1)	

4.4 数据转换

数据转换主要是针对数据仓库建立的模型，通过一系列的转换实现将数据从业务模型变换为分析模型，它是真正将源数据变为目标数据的关键环节[30]，它包括数据格式转换、数据类型转换、数据汇总计算、数据拼接等。

4.4.1 数据映射关系

经过多年的发展，银行业务数据库系统中积累了海量的数据，这些数据是建立数据仓库的主要来源。业务数据库系统中的数据在进行转换之前，必须进行数据映射（Source Data Mapping)；数据映射将对这些源系统的数据执行详细的分析，并将数据映射到逻辑数据模型，以确定设计和建立转换所需的工作量，并确定可能影响应用程序范围的任何缺口。在这个过程中需要明确定义数据仓库中的每个表、每个字段来自源系统或接口单元的哪张表、哪个字段等。有些映射关系比较简单，源数据表和数据仓库的目标表是一对一的关系，这样就可以直接把源表中的数据复制到目标表中。但是，这种简单的映射关系得到的分析价值不是很大，因此，大多数表间的映射关系是比较复杂的，常常要把多个源数据表经过分析合并为某一个目标表。多表间的关联一般通过主键，并进行内关联，有时还会附加某些约束条件。

数据映射包括表映射和字段映射，表映射指数据仓库中的表来源于源数据库系统中的哪个表或接口单元；字段映射指数据仓库中的每个字段来源于源表中的哪个字段以及如果通过转换、关联等操作进入数据仓库。

表4-3给出了该数据仓库系统中客户表映射和客户主题相关字段映射：

表4-3 目标客户与源表映射关系

表名	中文表明	表映射关系	抽取方式	抽取周期
Customer_Info	客户信息表	本表来源于CDSIA(对私客户信息表)，CDCIA(对公客户信息表)，CDADA客户补充信息表，CDWRA（客户服务文件）等 初始加载：取源表及各自历史表的全量数据 日常加载：取源表中增量数据	全量	每周
……	……	……	……	……

表 4-4 是客户信息表与对私客户信息表、客户补充信息表，客户服务文件字段间的映射关系

表 4-4 客户信息表字段级映射关系表

源表信息			目标表 Customer Info		
表名	字段名称	中文含义	字段名称	中文含义	字段类型
CDSIA	Customer_No	客户号	Customer_id	客户号	char(10)
		根据抽取表格来确认	Customer type	客户类型	char(2)
CDSIA	Customer Name	客户名称	Customer Name	客户名称	char(62)
CDSIA	Cantact_Phone	电话号码	Telephone	电话号码	char(15)
CDSIA		办公地址	Office_add	办公地址	char(62)
CDSIA	ID_Type	证件类型	ID_Type	证件类型	Char(2)
CDSIA	ID_No	证件号码	ID_No	证件号码	Char(20)
CDSIA	Start_date	启用日期	Start_date	启用日期	char(8)
CDWRA	CUSK	服务种类	Service_Type	服务类型	char(8)
CDADA	OpenLicence_No	开户许可证号码	OpenLicence_No	开户许可证号码	char(16)
CDADA	Foreign Exchange_No	外汇许可证号码	ForeignExchange_No	外汇许可证号码	char(16)
CDADA	Office_PostalCode	办公地址邮政编码	Office_PostalCode	办公地址邮政编码	char(7)
CDADA	Email	E_MAIL 地址	Email	E_MAIL 地址	char(42)
CDSIA	Open_ Branch Type	开户机构代号	Open_ Branch Type	开户机构代号	char(4)
CDSIA	Open_ Teller	开户柜员	Open_ Teller	开户柜员	char(8)

表 4-4 只给出了目标表客户信息表映射关系的一部分，商业银行的客户一般包括对私客户、对公客户、网银客户和不良客户等，对公客户表与目标表的映射关系不再列出。

4.4.2 数据清洗

数据清洗是指比简单变换更复杂的一种数据转换，但它并不是 ETL 中的一个单独步骤，必须与数据抽取、转换和加载统一使用。数据清洗一般包括清除无用数据和将数据标准化，SSIS 提供了 3 种机制来实现数据的清洗：一是内置转换，它能将无用数据清理和标准化、更改数据的大小写、将数据转换为不同类型或格式、根据表达式创建新列值等；二是通过精确查找和模糊查找来找到引用表中的值，并将列中的值替换为引用表中的值来清理。其三是将数据集中相似的值分组到一起来清理，如重复相似数据清理。

由于银行柜员或一些其他的人为原因，业务数据库中的数据经常会出现拼写错误、截断、缺少或插入的标记、空字段、意外的缩写语等，使得数据在加载数据仓库时出现不规范、二义性、重复和不完整的问题，这将不利用数据的加载。在该数据仓库系统构建的过程中，主要有两个比较棘手的问题：其一是由于历史原因或柜员在前台录入信息的不完全等造成了各个业务数据库系统中大量的重复记录，如客户信息的相关表、账户信息相关表等，同一客户有多个客户号等情况在核心系统中是一个普遍现象；其二是在周期抽取数据时，新增的客户信息由于各种原因使得数据不完整、格式不统一等。在初次抽取和增量抽取数据时必须把这些重复记录处理后再加载到该数据仓库系统中。

1.初次抽取前重复数据清洗

经过多年的发展，在银行核心业务系统中积累了大量的重复数据，如对私客户信息表中，由于客户的证件号码由 15 位升到 18 位等其他原因，在登记客户资料时，后期虽然对这个问题做过相关处理，但还是有大量重复记录；在对私账户信息、卡信息、信用卡信息等业务系统中也积累了许多重复记录。因此，在初次数据装载之前，必须把这些重复的脏数据进行清洗[53]。

重复记录的危害主要表现在两个方面：

一是损害信息的一致性：多条相似重复记录在数据库中以不同的主键来标识，他们的信息可能互为补充，但存在冗余，甚至矛盾。

二是资源浪费：相似重复记录不仅会造成数据库中的数据冗余，还会浪费存储空间。CRM 数据仓库中如果出现大量的重复记录，在后期 OLAP 和数据挖掘处理得到的分析结果将有很多的影响，甚至会出现错误的分析结论，使得客户经理和管理层无法做出准确的决策。

“排序-合并”算法是一种比较传统的相似重复记录方法，它先把数据库中的记录排序，然后通过比较邻近记录是否相等来检测完全重复的记录。通过各种匹配算法来进行字符串匹配，如果两条记录相似度超过了由用户预定义的某一阈

值，则认为这两条记录是匹配的，否则认为它们是指向不同实体的记录。还可以通过重复记录的聚类来快速的检索出在一个记录集中所有相互匹配的相似记录集。

2. 增量加载前数据的清洗

Microsoft SSIS 主要使用模糊查找和模糊分组转换来实现数据清洗功能。模糊查找将输入记录与引用表中无错的、标准化的记录匹配；模糊分组检测输入行之间的相似性，并通过检测输入行的字符串值确定那些行是重复项，这两种方法都对记录中的错误有复原功能。

该银行每天都会增加大量的客户，包括对公客户、对私客户和网银客户，因此，客户相关信息表中将不断的增加对客户信息的录入。由于各种原因，录入到业务数据库系统中的数据包含了信息不全、拼写错误、重复客户记录、虚假客户、同一客户的多个稍有不同的实例的数据表等。在周期抽取这些数据之后，必须对其进行清洗，对于这种类型的源数据可以通过模糊查找和模糊分组来处理，从而解决以上问题。

对于新的客户数据，为了方便处理，应该将其导入数据库中。另外，数据清洗后，要把干净数据和处理后的数据导入相关表，并通过 SQL 来创建源数据库中不存在的表。因此，在这次使用 SSIS 进行数据清洗任务中主要包括两个方面的处理：将需要的 SQL 程序集成到一个文件中，这样就不需要循环而只要执行这个组件；其次是建立一个数据流任务，这个任务主要执行模糊查找进行数据的清洗，标识出唯一的新客户、现有客户和重复客户，并将属于每种客户类型的行写入相应输出表中。

其数据流过程是：先把查询组件把源数据与原来的客户数据进行比较，完全匹配的就是老客户，如果该客户信息地址、证件号码等信息有修改则修改客户信息表中相关列或增加新列。对于不完全匹配的数据进行模糊查找操作，根据相似程度进行条件拆分等，把相似程度好的存入客户数据，匹配不好的就经过模糊分组转换，再加载到数据仓库客户信息表中。

4.4.3 异构数据合并

数据通常存储在多个不同的数据存储系统中，从各个数据源中提取数据并将其合并到单个一致的数据集中是一个较复杂的过程[54]，在该数据仓库系统的数据移植过程中也不例外。该银行在 2003 年业务大集中之后，大部分数据存储在核心业务系统中，但是在早期的业务数据库系统然后有大量有价值的数据，另外目前的网上银行系统所使用的是 Oracle 数据库。所以，在该数据仓库系统的异构数据合并过程中，是指对这两个源数据系统中数据进行移植。

在对这些数据的合并主要使用 Microsoft 公司的 SSIS 来实现转换，SSIS 能够连接到多个数据源，通过使用.NET 和 OLE DB 访问接口连接到关系数据库，也可以使用 ODBC（Open Database Connectivity，开放数据库连接）驱动程序连接到多个早期数据库。另外，通过连接到平面文件、Excel 文件和 Analysis Services 项目等，从中提取数据并转换为分析系统兼容的格式。

ODBC 是 Microsoft 公司提出的标准应用程序接口，它允许应用程序 SQL 语言为数据库的存取标准来存取不同的 DBMS 管理的数据。ODBC 屏蔽了底层数据库系统的不同，从而简化了对数据库的访问。

OLE DB 是通用数据访问技术 UDA 的系统级标准接口，它通过建立数据访问的标准接口，把多个的数据源经过抽取形式成行集的概念。另外，它还提供了一组标准的服务组件，用于提供查询、缓存、数据更新、事务处理等操作。OLE DB 方法进行数据转换，通用性较强且支持多种数据源，可以对所有的关系数据库进行转换，对于某些特定的 OLE DB 驱动程序，转换效率高。因此，在对本系统中两个异构源数据平台的转换过程中，SSIS 使用 OLE DB 作为数据源是一种比较常用的方法。

4.5 数据加载

银行业务数据库中的数据经过抽取、清洗、转换之后就是加载到数据仓库中。根据源数据的不同有三种不同的加载方式：插入、增加和刷新。

刷新：即每次加载目标表的数据时，一般会抽取源数据中的所有记录，并删除目标表原有数据，然后完全加载最新源数据。

插入：只要将抽取的数据源文件中的数据全部插入到目标表中。

增加：这种方式要求根据主键，对已有的记录进行更新，对不存在记录做插入操作。

根据模型设计部分确定的其中两个主题（即客户主题、账户主题）来选取加载方式。客户信息表的加载可以采用刷新的方式，因为客户在业务数据库中是唯一的，如果客户的某些信息做过修改，数据仓库中的原有记录必须删除，加载新的信息。账户主题中的账户信息表则要求记录抽取周期内的每个变化，所以采用插入方式。

4.6 系统 ETL 部分实现

在本章前面的内容中已经讨论了 ETL 对数据仓库构建过程的重要性，本小节将以客户信息表增量加载来说明本系统中部分 ETL 的实现过程。

本系统中采用的 ETL 工具是 Microsoft SSIS 工具。SSIS 通过包来管理复杂的数据整合任务，通过控制流、数据流和事件处理程序等组件来处理这些任务。控制流由容器、任务和优先约束等控制流元素构成。数据流由提取数据的源、修改和聚合的转换、加载数据的目标，以及将数据流组件的输出和输入连接为数据流的路径等元素构成。图 4-2 即为数据流的构成方式。

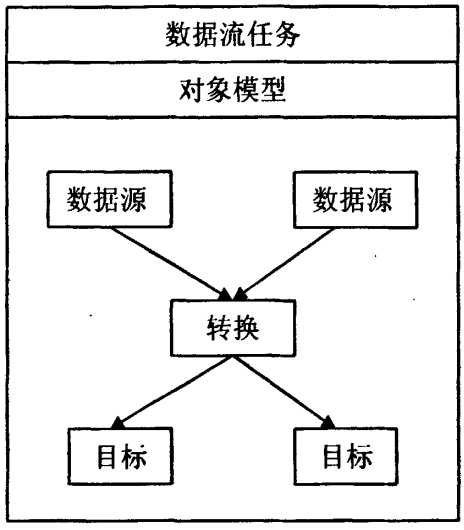

图 4-2 数据流构成方式

SSIS 的数据流任务封装数据流引擎。数据流引擎提供将数据从源移动到目标的内存中的缓冲区，并调用从文件和关系数据库中提取的数据的源。数据源系统是 Informix 数据库，数据处理全部过程在 SSIS 平台上完成，处理的结果最终导入到 Informix 平台的数据仓库中。

该银行由于业务发展的需要，去年在湖南省，广西省，青海省新建立了三个分行机构，因此，在数据仓库系统客户信息表中必须把这三个分行的客户信息加载进去。这三个分行建立的时间不是很长，积累的数据量不多，首先把分行的业务数据库系统中的相关数据已经汇总到一个文本文件中，然后通过 SSIS 的 ETL 过程中需要做的就是把这些数据按照分行号（Branch_No）将其分类，并按分行建立三个新表，最后把数据导入到新表中。图 4-3 即为 SSIS 中处理过程的数据流图。

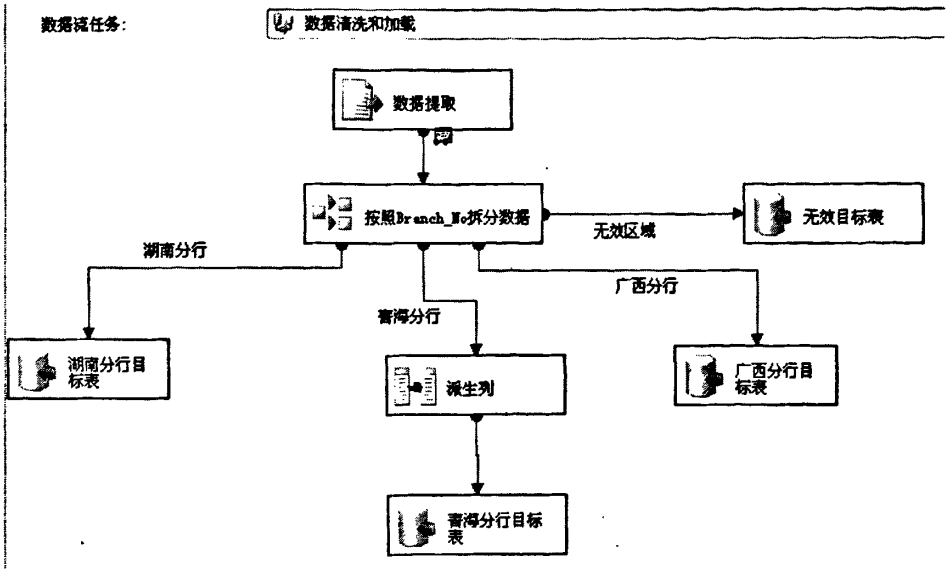

图 4-3 增量数据清洗、加载数据流图

4.7 本章小结

在数据仓库系统的构建过程中，ETL 的设计优劣将直接影响到后期的 OLAP，数据挖掘的效果，最终影响到客户分析和决策分析的正确性。因此，本论文重点讨论了数据仓库关键技术 ETL 及其工具的选取，并对该数据仓库系统数据的抽取、清洗、转换和加载进行了详细的设计，最后还以增量客户信息处理实现了 ETL 的处理过程。

第五章 总结与展望

5.1 全文总结

数据仓库技术已广泛应用于国内外银行业，用来支持银行卡信用卡业务分析、CRM、贷款风险管理和决策支持等，以应对日益激烈的市场竞争。特别是国内金融银行市场在度过了五年的过渡期后，于2006年完全对外开发，大量国外银行巨头大量涌入中国市场，国内商业银行已经感受到了国外银行给它们带来的冲击。

本文在深入分析了数据仓库技术的发展状况和国内外商业银行实施数据仓库和CRM的现状背景下，基于目前的竞争局面和某银行的业务状况，实现了针对该银行客户关系管理（CRM）数据仓库的设计与实现。该数据仓库系统的实施运行证明其结构是比较合理有效的，它能为客户信息管理、客户综合分析、目标客户搜索和业务查询与统计功能的CRM系统提供及时而准确的数据，并为后期建立全行用于银行卡信用卡业务分析、贷款风险评估、会计管理及其他决策分析等的企业级数据仓库的建立奠定了基础。

本文主要做了以下几点工作：

1．讨论了分析了数据仓库特点、数据库与数据仓库的不同、OLAP 技术的实施方法和数据仓库技术的发展前景。同时对数据挖掘技术、CRM 的特点、方法和发展等进行了相关的讨论。

2．通过参与该银行核心系统的维护和开发工作，对该银行目前的业务系统和市场发展状况作了深入调研分析，给出了一套针对该银行用于CRM数据仓库系统总体设计方案，并对该系统的工具进行了分析和选择。

3．根据总体设计的方案及体系结构，分析和设计了数据仓库的数据模型，包括概念模型、逻辑模型和物理模型。针对该数据仓库系统用于银行CRM分析的目的，深入讨论和设计了该数据仓库的主题（客户、账户、交易、产品）及主题间的关系、逻辑模型图、粒度等。在物理模型设计阶段确认了事实表、维表及其存储结构、索引策略等。

4．ETL是数据仓库构建过程中的关键技术，本文分析比较了几种目前比较流行的ETL工具。考虑到ETL的重要性和目前普遍的ETL效率不高的问题，而且该银行处理的数据量较大和存在部分异构数据，本文重点讨论和设计了围绕客户主题的 ETL 处理过程，包括数据抽取、数据接口设计、数据映射关系设计和重复数据清洗加载方案等，并在以上方案的基础上对该数据仓库系统的 ETL 过

程做了部分实现。后期的测试证明这套 ETL 处理方案的执行效率是较高，并在数据仓库实施的过程中发挥了重要的作用。

5.2 不足与展望

由于项目研究时间的仓促和人员水平有限，而且数据仓库系统的构建是一个长期和复杂的过程，本文主要对数据仓库模型的设计及数据加载、转换、清洗和加载做了深入的研究和设计，并且还有许多待改进之处。对数据仓库的下一步工作数据挖掘及相关算法没有做详细的讨论和分析。另外，商业银行构建数据仓库的目的不仅仅应用于 CRM 系统中，在资产负债管理、银行卡信用卡分析、贷款风险管理等也能够有较好的应用。最后，数据仓库的维护也是一项重要而复杂的工作，本文中也没有做相关内容的讨论。

因此，随着对商业银行数据仓库技术研究的深入，在以后的工作中将从以上几个方面做进一步的工作。

参考文献

[1]中国银行业监督管理委员会. 中国银行业对外开放报告. 北京：2007.3

[2]W.H.Inmon 著，王志海等译. Building the Data Warehouse [M]. 北京机械工业出版社.2000

[3]唐世渭，于波，孙国辉，赵征. 引入数据仓库技术加快我国银行业信息化建设[J]. 中国金融电脑，2004，2：5-10

[4]张晓东，王建民. 浅谈商业银行建立数据仓库的必要性[J]. 金融与经济，2003，1:27-28

[5]庞瑞江，李庆莉. 金融信息化的未来发展框架[J]. 中国金融电脑，2003，6：34-38

[6]李大鹏. 数据仓库和数据挖掘语言初探[J]. 计算机与通信，2004，1：36-40

[7]卜迎东. 利用数据仓库技术实现银行客户分析：[硕士学位论文]. 吉林大学. 2005

[8]林剑广. 基于数据仓库的商业营销决策支持系统开发[J].计算机应用与软件，2004，21(2)：39-42

[9]王珊等. 数据仓库技术与联机分析处理[M]. 北京：科学出版社，1998

[10]Husemann B, Lechenborger J, Vossen G. Conceptual data warehouse design[C]. Proceedings of the International Workshop on Design and Management on Data Warehouse(DMDW'2000). Sweden：2000

[11]林宇. 数据仓库原理与实践[M]. 北京：人民邮电出版社，2003

[12]张维明. 数据仓库原理与应用[M]. 北京：电子工业出版社，2002

[13]骆伟忠，陈松乔. 基于OLAP技术实现银行客户分析系统[J]. 微计算机信息，2004，20(5)：34-37

[14]江键，陈福生. OLAP 在银行数据仓库中的设计与实现[J]. 计算机工程与设计，2006，27(20)：3884-3886

[15]彭木根. 数据仓库技术与实现[M]. 电子工业出版社. 2002.6

[16]孙建玲. 基于数据仓库技术的银行卡决策支持系统：[硕士学位论文]. 广州：华南理工大学. 2003.5

[17]李庆莉. 银企对话解析金融信息化趋势—第四届中国金融信息化发展自由论坛[J]. 中国金融电脑，2003，10

[18]Jiawei Han, Micheline Kamber. Data Mining Concepts and

Techniques[M]. China Machine Press. 2005.2：3-23

[19]C.Shilakes and J.Tylman. Enterprise Information Portals. MerrillLyneh. httP://www.sagemaker.eoln/eomPany/whitePapers/eiP--indePth.pdf. 1998

[20]周龙. 基于国内银行数据仓库的ETL构造方案研究：[硕士学位论文]. 广州：华南理工大学. 2005

[21]李小东. 数据挖掘技术在银行信用卡业务中的应用研究：[硕士学位论文]. 杭州：浙江大学. 2002.11

[22] 黄华卿，张维，熊熊. 数据挖掘技术在商业银行客户关系管理中的应用分析[J]. 哈尔滨商业大学学报，2006,'3：40-43

[23]周丹晨. CRM环境下面向知识发现的数据分类技术的应用研究[J]. 计算机应用与软件，2004，21(2)：9-11

[24]叶开. 中国CRM最佳实务. 北京：电子工业出版社[M]. 2005

[25]杨文海. 中国工商银行CRM应用规划：[硕士学位论文]. 成都：西南财经大学. 2003.4

[26]刘翔. 数据仓库与数据挖掘技术[M]. 上海交通大学出版社. 2005.8

[27]李宗怡. 金融服务业中的客户关系管理战略[J]. 经济导刊， 2000，6:34-39

[28]Ralph Kimball Laura Reeves. 数据仓库生命周期工具箱：设计、开发和部署数据仓库的专家方法[M]. 北京：电子工业出版社. 2004.

[29]周四新. 银行决策支持系统中数据挖掘的研究与实现：[硕士学位论文]. 长沙：中南大学. 2004.4

[30]Wua L,Millera L,Nilakantab S.Design of data warehouses using metadata[J]. Information and Software Technology, 2001,43:109-119

[31]尤玉林，张宪民. 一种可靠的数据仓库中ETL策略与构架设计[J]. 计算机工程与应用，2005，10：172-174

[32]P.Vassiliadis, Z.Vagena, S.Skiadopoulos, N.Karayannidis, T.Sellis. ARKTOS：Towards the modeling, design, control and execution of ETL processes[J]. Information Systems. 2001.26(8)：537-561

[33]张宁,贾自艳,史忠植. 数据仓库中ETL技术的研究[J]. 计算机工程与应用，2002，38(24)：213-216

[34]张阐军. 基于数据挖掘的RCM系统关键技术研究及其应用：[硕士学位论文]. 武汉：武汉理工大学. 2005

[35]Thomas Thalhammer, Michael Schrefl, Mukesh Moohania. Active data

warehouse:complementing OLAP wich analysis rules, Data & Knowledge Engineering 2001,39: 241-269

[36]李树新. 浅谈基于数据仓库理论的银行管理信息系统建设[J]. 阴山学刊, 2006, 3: 56-58

[37]左爱群, 杜波. 基于数据仓库的银行客户关系管理系统的研究[J]. 计算机与现代化, 2006, 8: 59-63

[38]Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, EDITORS. 1996. Adavance In Knowledge Dicovery And Data Mining. AAAI/MIT Press

[39]张旭峰. ETL 若干关键技术研究: [博士学位论文]. 上海: 复旦大学. 2006.4

[40]周龙,姚耀文. 基于银行系统应用的 ETL 技术的探讨[J]. 计算机应用, 2004, 10: 147-150

[41]N. Damas, 徐德智. 数据仓库设计方法研究[J]. 企业技术开发, 2004, 23(2):7-9

[42]Calvanese D, de Giacomo G, Lenzerini M, et al. Data Integration in Data Warehousing[J]. International Journal of Cooperative Information System, 2001, 10(3):237-271

[43]陈长清, 冯玉才, 袁磊. 国产数据仓库管理系统 DM, DW 的设计[J]. 小型微型计算机系统, 2002, 23(5): 596-599

[44]李长河. 基于多层次数据库的智能 Web 挖掘系统[J] . 计算机工程, 2004, 30(5): 93-94

[45]Hongwei Zhu, Stuart E, Madnick, Michael D. Effective Data Integration in the Presence of Temporal Semantic Conflicts. Proceedings of the 11th International Symposium on Temporal Representation and Reasoning

[46]M. Demarest. The Polities of data warehousing Available from httP://www.uneg.edu/isln/ism611/Polities.Pdf. 1997

[47]Pudi V Haritsa J R. Quantifying the utility of the past in mining Large Database. Information Systems. 2000.25(5): 323—343

[48]J. Hammer, H. Garcia- Molina, J. Widom, etal. The Stanford Data Warehousing Project. IEEE Computer Application Power. 1995, 21(3): 123-141

[49]W. Juan, J. Yang, Y. Cui, etal. Performance Issues in Incremental

Warehouse Maintenance. in: Proceedings of the 26th VLDB Conferences. 2000: 461-472

[50]Milo, Tova and Sagit Zohar. Using Schema Matching to Simplify Heterogeneous Data Translation. In Proc. 24"VLDB, PP122-133, 1998

[51]徐立臻,谢鸿强,董逸生. 数据仓库系统中源数据的提取与集成[J]. 小型微型计算机系统, 2003, 5 (24): 869-873

[52]黄大荣, 黄席榔. 基于粗糙集理论的数据清洗模型[J]. 计算机工程与应用, 2004, 24(13):164-165

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

[54]Venkater Ramesh, Suda Ram. Integrity Constraint Integration in Heterogenous Database. An Enhanced Methodology for Schema Integration[J]. Information System, 1997, 22(8)

[55]吴远红. ETL 执行过程的优化研究[J]. 计算机科学, 2007, 34: 81-83

[56]花海洋,李一凡, 赵怀慈. 基于分布式数据仓库技术的 ETL 系统的研究与应用[J]. 微计算机信息, 2006, 22(10):144-147

[57]Jaideep Srivastava. Warehouse Creation-A Potential Roadblock to Data Warehousing. IEEE Transactions on Knowledge and Data Engineering, vol. 11, No. 1, Jaunary/Frbruary 1999

[58]邱云飞, 邵良杉, 那宝贵. 利用 DTS 组件实现数据仓库中 ETL 方案设计[J]. 计算机系统应用, 2007, 4: 92-96

[59]肖国荣. 银行信贷系统 ETL 技术研究[J]. 科技资讯, 2006, (10): 6-7

[60]蓝婵, 梁华金. 基于 IBM 系列平台的数据提取技术分析[J]. 现代计算机, 2005, 3: 79-82

[61]熊志正. 商业银行稽查系统的 ETL 设计及改进方法[J]. 微电子学与计算机, 2005, 10(3): 79-82

致 谢

本论文的完成，与导师谭汉松教授的悉心指导和帮助是密不可分的。从论文选题到文献检索、从初稿的形成到最终定稿，都离不开谭汉松老师的教诲和指导。三年的研究生学习中，谭汉松教授严谨的治学态度、严密的逻辑思维方式给我留下了深刻的印象。不论在学业上还是生活上，谭老师都给我以全力的支持和帮助，让我学到了很多宝贵的知识，以及很多做人的道理。在此向谭老师表示深深的谢意！

感谢北京长信通公司领导及同事在实习期间对我帮助，特别是设计事业部的章春江经理和技术总监李红女士。在整个实习过程中，章经理和李女士在工作生活上给予了我莫大的帮助。

感谢实验室的所有同学，特别是丁烨敏、黄博飞、张宗盛、邹华、王宙、刘尧、马行坡等,大家在实验室和实习期间互相帮助，愉快地度过了研究生生活。

我还要深深地向在背后一直支持我的家人道一声感谢，没有你们对我的理解和无私的关爱就不会有我实现自己梦想的机会。

最后，对百忙之中审阅我的论文和参加答辩会的老师们表示最诚挚的感谢。

攻读学位期间主要的研究成果

已发表的学术论文：

[1] 谭汉松，陈才，张宗盛. 数据仓库在商业银行 CRM 的应用研究. 科技广场, 2008, 1：93-95

[2] 谭汉松，张宗盛，陈才. 关于 MDA 技术的发展现状和研究报告. 科技广场, 2007, 11：6-8

参加的研究项目：

北京长信通信息技术有限公司

[1] 银行数据仓库系统的开发

[2] 银行贵宾卡客户积分系统的开发

[3] 银行第二期记账式国债项目的开发