

分类号 : TP338.4
UDC : 004.4 学号 : 032117



东 南 大 学

工 程 硕 士 学 位 论 文

新一代银行股票交易系统技术研究

研究生姓名 : 杨套军

导师姓名 : 吴介一 (教授)

韩 洁 (高工)

申请学位级别 专业硕士学位 工程领域名称 软 件 工 程

论文提交日期 2005 年 6 月 2 日 论文答辩日期 2005 年 7 月 5 日

学位授予单位 东 南 大 学 学位授予日期 2005 年 7 月 5 日

答辩委员会主席 陈冠清 (教授) 评 阅 人 张家浩 (高工)

李海峰 (高工)

2005 年 7 月 5 日

THE TECHNOLOGY RESEARCH OF NEW-FASHIONED STOCK EXCHANGE SYSTEM

A Dissertation Submitted to

Southeast University

For the Academic Degree of Master of Engineering

BY

YANG Tao-jun

Supervised by

Professor WU Jie-yi

and

Senior Engineer HAN Jie

College of Software Engineering

Southeast University

June 2005

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个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，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：_____日 期：_____

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学研究生院办理。

研究生签名：_____导师签名：_____日 期：_____

摘要

本文在某国际商业银行股票交易系统构架的基础上，探讨了股票市场分析方法在该系统上的应用，以及国内外数据仓库和数据挖掘技术在各个领域的广泛应用，以提供股票投资决策为目的，研究了如何将数据仓库技术应用到股票交易系统中，从而为股票交易系统增添股票分析与预测功能提供理论基础。另外，对原有交易系统技术上进行了一些改进，从而提高系统的效率及使用上的方便。主要内容如下：

- <1> 对某国际商业银行股票交易系统的构架进行概述，并简要介绍各通信模块的功能。
- <2> 对数据仓库和数据挖掘技术及其应用进行了讨论。
- <3> 按照数据仓库的构建方法，对股票交易系统进行需求分析，将数据进行转换、粒度划分、分割等操作，实现了股票交易系统数据仓库及系统间接口的建立。
- <4> 对原股票交易系统进行技术改进，并着重叙述订单交易模块的开发及测试。

关键词： 股票交易 数据仓库 数据挖掘

ABSTRACT

In the paper, with a international commercial bank's system framework, reseach the application of stock analysis methods in the system, and extensive application of data warehouse and data mining in various fields at home and abroad. in order to provide decision-making for stock investment, the application of data storehouse in stock exchange system is studied, and the data flat roof of oriented analyzing and decision-making is established, by analyzing and forecasting stock data, consequently give client some good advices. in addition, in order to high efficiency and convenience in using, Improve old exchange system in Technology ,The following contents are studied in this dissertation:

<1> System framework of the international commercial bank is summarized, simply introduce the functions of main models.

<2> Technology and application of data warehouse and data mining are discussed.

<3> According to establishing measure of data warehouse .by analyzing of requirement in stock data exchange system, the operating such as conversing, granulari ty carving up, di vi di ng up are implemented, and data warehouse of stock exchange system is established. besides query、 statistic and so on, the flat roof of analysis and decision-making is established primarily.

<4> Improve old exchange system in Technology, and discuss the development and testing of order model stressly.

Keywords: stock exchange data warehouse data mining

目录

摘要.....	I
ABSTRACT.....	II
第一章 绪论	1
1.1 研究背景.....	1
1.2 国内外研究综述.....	1
1.2.1 网上股票交易应用研究综述	1
1.2.2 数据仓库应用研究综述	2
1.2.3 数据挖掘研究综述	3
1.3 课题来源及选题依据.....	4
1.4 本文研究的主要内容.....	4
1.5 研究意义.....	5
第二章 股票交易系统	6
2.1 股票交易相关概念.....	6
2.2 体系结构.....	7
2.3 交易流程.....	8
2.3.1 股票走势查询	9
2.3.2 填写股票订单	9
2.3.3 鉴定订单	10
2.3.4 发送订单	10
2.3.5 修改订单	11
2.3.6 执行交易	11
2.3.7 修改/鉴定交易	12
2.3.8 数据核对	12
2.3.9 结算	12
2.4 本章小结.....	13
第三章 数据仓库和数据挖掘技术分析	14
3.1 引言.....	14
3.2 数据仓库技术.....	14
3.2.1 数据仓库的定义	14
3.2.2 与数据仓库相关的几个概念 ^[12-14]	15
3.2.3 数据仓库体系结构 ^[13]	16
3.2.4 数据仓库的数据组织结构	17
3.2.5 数据仓库的实现策略 ^[15]	19
3.3 数据挖掘技术 ^[4]	20
3.3.1 数据挖掘的任务	20
3.3.2 数据挖掘对象	21
3.3.3 数据挖掘和数据仓库	21

3.3.4 数据挖掘和在线分析处理(OLAP)	21
3.3.5 数据挖掘流程 ^[16-17]	22
3.4 基于数据仓库的数据挖掘技术	22
3.5 本章小结	23
第四章 股票交易数据的数据仓库建立	24
4.1 引言	24
4.2 引言	24
4.3 主题的选取	26
4.4 逻辑结构设计	26
4.4.1 分析主题域	26
4.4.2 粒度层次划分和数据分割	27
4.4.3 数据之间关系的表示	28
4.5 物理模型设计	29
4.5.1 可适当入冗余	29
4.5.2 生成导出数据	29
4.6 数据仓库的实现 ^[20]	29
4.6.1 数据仓库的体系结构	29
4.6.2 数据获取	30
4.6.3 数据装入	31
4.6.4 技术指标计算 ^[21-22]	32
4.6.5 功能实现	34
4.7 数据仓库与系统间的连接	35
4.8 本章小结	36
第五章 股票交易系统的实现	37
5.1 开发环境	37
5.2 股票交易系统改进	37
5.3 项目开发	38
5.4 项目测试	38
5.4.1 命令行测试	39
5.4.2 在线测试	39
5.5 本章小结	42
第六章 结论	43
致 谢	44
参考文献	45

第一章 绪论

1.1 研究背景

股票至今已有将近 400 年的历史。股票是社会化大生产的产物，股票的发行与交易促进了市场经济的发展。随着人类社会进入了社会化大生产的时期，企业经营规模扩大与资本需求不足的矛盾日益突出，于是产生了以股份公司形态出现的，股东共同出资经营的企业组织；股份公司的变化和发展产生了股票形态的融资活动；股票融资的发展产生了股票交易的需求；股票的交易需求促成了股票市场的形成和发展；而股票市场的发展最终又促进了股票融资活动和股份公司的完善和发展。所以，股份公司、股票融资和股票市场的相互联系和相互作用，推动着股份公司、股票融资和股票市场的共同发展。

90 年代以来，计算机技术、尤其是数据库技术和网络技术在股票市场中得到充分应用，使得作为证券市场的重要组成部分的股票市场更加蓬勃地发展起来，逐步成为证券业乃至整个金融业的必不可少的组成部分，显示出强大的生命力。

随着 Internet 应用的普及，为客户提供网上股票交易已成为银行一种重要的业务和手段。但是网上股票交易并不是简单的银行业务网络化，它应具备以下功能、特点：充分利用现有的业务系统，减少在应用开发、测试上的投资和时间；系统应具备有增加新应用、新功能的扩展能力，并能与传统应用无缝连接；系统的安全性是在开放的网络环境下进行交易处理的基础；系统处理能力的可扩展性；系统的高可用性。如何建立一个行之有效的网上股票交易系统，已成为研究的一个重要课题。

然而，与 IT 技术的迅猛发展相比，网上股票交易的功能体系却显得不够完善、健全，例如目前多数股票交易系统的并不提供股票走势分析预测，这就不能为客户进行股票交易提供参考。即使添加该项功能，就不仅要保留交易数据，还需存储历史数据，虽然保存这些数据对于系统进行故障诊断、统计分析、数据挖掘以及决策支持都有着重大的意义，但这些数据往往是海量的，如果采用一般的数据库技术来管理和维护，并在其基础上建立分析和决策应用，其效率之低是可想而知的。因此，证券业呼唤先进的数据存储和管理技术。这些问题已成为股票交易发展进程中的一个瓶颈，成为研究中的重中之重。如何有效地解决这些问题，更好地发挥网上股票交易的作用，不仅关系到银行界业务的持续性增长，而且对股票市场的良性发展又有不可忽视的影响。种种迹象表明，这已成为当前金融界的热门话题^[1]。

基于这种原因，本课题研究将先进的数据仓库技术应用到某商业银行股票交易系统中，形成一个综合的、面向分析的环境，最终建立面向高层的决策平台，为系统增添股票走势预测功能提供理论参考，并对原股票交易系统存在的不足之处进行技术上改进，从而形成新一代的银行股票交易系统。

1.2 国内外研究综述

1.2.1 网上股票交易应用研究综述

购买股票已成为普通百姓个人投资的新热点，传统的股票交易有两种形式：一是场内现时交易，股民在交易所内通过划卡或填单委托进行交易，二是电话委托交易。但大

量的股民会有这样的感觉,采用场内现时交易受时间、空间限制较重,单就交易环境来说,几百平米的交易大厅拥挤着上千股民,股市处于牛市的时候更多的不得了。这种情况冬天还好说,遇到夏季,许多股民会感到不便,心情急躁往往会影响到股票走势的判断力,造成投资的损失。采用电话委托方式交易,虽然克服了时间、空间的限制,但部分语音信息在传输时保密性相对较弱,同样会造成不必要的损失。另外,股市行情查询与股票交易之间存在有时差,股民在进行股票交易时往往会留出一定的价位差,这样就会给股民带来收益上的影响。再有就是当股民出差或出国时,要想操作自己在居住地的股票就会有些困难。

随着 Internet 应用的普及,为客户提供网上股票交易已成为银行一种重要的业务和手段。为此,各大银行纷纷寻求一种行之有效的网上股票交易系统,这也促使相应的软件业蓬勃发展。传统的方法是采用 C/S 体系结构,银行的客户通过运行在 PC 上的浏览器访问银行的 Web Server,通过对用户的用户识别码和口令的检查,确认用户的合法性。Web Server 将客户的访问请求转变成对应的银行交易请求,并交给在 OS/390 上运行的交易处理系统处理,交易运行结果通过 Web Server 反馈给客户。目前世界各大银行大多都采用该技术。

而处于理论研究阶段的语音技术更令人遐想连篇,而且它已逐渐应用于电信的声讯信息服务领域和互联网消息收发方面。将来随着电话网与互联网的融合、网络信息项目的增多和时效性要求逐步提高,建立适合于股票交易业务的语音系统成为可能,电话用户可以通过传统的语音、传真获取互联网上无穷无尽的信息。这些业务将彻底解决传统数字录音回放技术所无法解决的海量信息库和动态变化信息的实时生成与存储的难题,因此,将语音技术与网络进行完美的结合具有强大的生命力,也为未来的股票交易系统的发展提出了一条新的分枝^[2-3]。

1.2.2 数据仓库应用研究综述

数据仓库是近几年才提出的新概念。整个 80 年代直到 90 年代初,联机事物处理(OLTP)一直是数据库应用的主流。联机事物处理系统的主要任务是将当前事务所产生的数据记录下来,在此基础上,进行日常性的处理。然而随着市场竞争的加剧和信息社会需求的发展,从大量数据中提取(检索、查询等)制定市场策略的信息就显得越来越重要了。这种需求既要求联机服务,又涉及大量用于决策的数据,而传统的数据库系统已无法满足这种需求。数据仓库是一个能很好地支持企业决策分析处理、面向主题的、集成的、随时间不断变化的数据集合。数据仓库中的数据是良构的、一致的、相对稳定的,并且它的数据量足以支持数据分析、查询、报表生成以及与历史数据对比。其建立的目的是把企业的内部数据和外部数据进行有效的集成,提供给企业的决策者和管理者使用。为了满足决策者的需求,数据仓库对数据进行结构上的重组,从便于决策分析的角度去设计,并且充分考虑今后的扩展性及与外部数据的接口。

目前数据仓库中的数据大多都存储在关系数据库中,并以 SQL 语言作为 API 接口。不同的关系数据库提供的 SQL 接口差异很大,因此一般采用三种技术途径解决:

(1) 在关系数据库上进行多维分析。为了获得多维分析功能,需要在客户机和关系数据库之间提供一个额外的语义层,用来理解多维结构及其与关系表的映射,并且把来自客户机的请求翻译成标准的 SQL 语句;

(2) 用专门的多维数据库进行多维分析。这是一种高速缓存的方法,数据从 OLTP 数据库、全局数据仓库或数据集市提取出来之后,经过预处理,加载到多维数据服务器中。多维数据库管理多维模型,并提供一种能够处理超立方体结构的 API;

(3) 在类关系数据库上进行多维分析。这也是一种高速缓存的方法。数据从业务

数据库拷贝到专用的非 OLTP 关系数据库^[4-6]。

近年来,数据仓库成为数据库行业中的一个热门话题,人们普遍认为它将是数据库技术未来发展的一个重要组成部分。在欧美发达国家,以数据仓库为基础的在线分析处理和数据挖掘应用首先在金融、保险、证券、电信等传统数据密集型行业取得成功。近年来,随着因特网和电子商务的发展,各大数据仓库产品供应商纷纷把注意力投向电子商务领域,并且通过数据仓库技术来构造商业智能(BI)平台。数据仓库技术在政府税收部门的应用带来的效益也非常可观。

目前市场上各种数据仓库产品及其解决方案品种繁多,归纳起来主要有九种主流数据仓库产品(Business Objects、Oracle、IBM、Sybase、Informix、NCR、Microsoft、SAS、CA)。根据各公司提供的数据仓库工具的功能,可将其分为三大类:单点产品、提供部分解决方案的产品、提供全面解决方案的产品。

(1) 单点产品。仅局限于数据仓库方案实施中的一部分或某一特定功能,主要是作为第三方产品或者和其它公司的产品结合起来进行使用。比较有特色的是 Business Objects。

(2) 提供部分解决方案的产品。不能提供构建数据仓库的全部工具,它们还要结合第三方供应商提供的工具才能完成整个数据仓库的构建。例如:IBM、Oracle、Sybase、NCR、Informix、Microsoft 等公司所提供的数据仓库产品都属于此类范畴。较常用的有 Microsoft 公司的 SQL Server 2000,它已经在性能和可扩展性方面确立了世界领先的地位,是一套完全的数据库和数据分析解决方案,使用户可以快速创建下一代的可扩展电子商务和数据仓库解决方案。Microsoft 将 OLAP 功能集成到 Microsoft SQL Server 中,提供可扩充的基于 COM 的 OLAP 接口。Microsoft Office 2000 套件中的 Access 和 Excel 可以作为数据展现工具,另外 SQL Server 还支持第三方数据展现工具。

(3) 提供全面解决方案的产品。其代表性产品有 CA,CA 于 1999 年收购 Platinum Technology 公司后,得到了完整的数据仓库解决方案,它具有丰富的行业知识和经验,能够向用户提供完善的数据仓库构造咨询服务、全面的数据仓库产品系列。它能够提供以“元数据”为核心的,结构化的,包容构造数据仓库各项技术的全面解决方案。它具有独特的“元数据”管理与应用特性,具有用三维可视化技术解析处理的功能。CA 解决方案提供 ODBC 接口,并将数据存储存储在第三方关系数据库(如 Oracle、Sybase、SQL Server、Informix 和 IBM DB2 等)中,其性能虽要打一些折扣,但提高了开放性。

数据仓库虽能保证对业务数据的便捷访问和以更快更好的方式回答复杂的问题,但是在数据仓库所用理论和数据仓库实现的实际过程之间,有一条坎坷的路径。数据仓库需要细致的设计规划和广泛的集成工作,因为数据仓库不是一个可以购买的产品,从大型的企业级的数据仓库到本地的面向主题的数据集市,每一个都是为特定的需求组合定制的。因此数据仓库市场前景广阔,充满了无限商机。据有关专家预测分析,从 1997 年到 2002 年,整个数据仓库市场将以平均每年 20.5%的速度增长,到 2002 年将超过 290 亿美元。相信随着现代商业模式变革的进一步深入,数据仓库应用将成为企业获得竞争优势的有力武器。

1.2.3 数据挖掘研究综述

当今数据容量规模已经达到万亿字节(TB)的水平。过量的数据被人们称为信息爆炸,带来的挑战是:一方面规模庞大、纷繁复杂的数据体系让使用者漫无头绪、无从下手;另一方面在这些大量数据的背后却隐藏着很多具有决策意义的有价值的信息。那么,如何发现这些有用的知识,使之成为管理决策和经营战略发展服务?计算机科学给出的最新回答是:数据挖掘(Data Mining)。一般说来,数据挖掘是一个利用各种分析方法和

分析工具在大规模海量数据中建立模型和发现数据间关系的过程,指的是从大量、部分、模糊、随机的实际应用数据中,提取隐含其中、人们事先不知道、但又有用的信息,同时用能被人理解的模式进行高级处理的过程,这些模型和关系可以用来做出决策和预测。它是数据库中知识发现(Knowledge Discover Database, KDD)的最核心部分。数据挖掘应建立在联机分析处理(On Line Analytical Processing, OLAP)的数据环境基础之上。数据挖掘对大量数据的探索式分析的起点是 OLAP。数据挖掘需要对大量数据进行反复查询操作,关心数据存取方式的方便性与可操作性。数据挖掘在对这些数据进行分析前,必须对这些数据进行不同程度的整合和清理,这是数据挖掘的首要环节。目前,数据挖掘技术正处在发展当中。数据挖掘涉及到数理统计、模糊理论、神经网络和人工智能等多种技术,技术含量比较高,实现难度较大。此外,数据挖掘技术还会同可视化技术、地理信息系统、统计分析系统相结合,丰富数据挖掘技术及工具的功能与性能。

数据挖掘可以应用于以下范围:

(1)自动探测以前未发现的模式。数据挖掘工具扫描整个数据库并辨认出那些隐藏着的模式。

(2)数据挖掘技术可以让现有的软件和硬件更加自动化,并且可以在升级的或者新开发的平台上执行。当数据挖掘工具运行于高性能的并行处理系统上的时候,它能在数分钟内分析一个超大型的数据库。这种更快的处理速度意味着用户有更多的机会来分析数据,让分析的结果更加准确可靠,并且易于理解。

(3)自动趋势预测。数据挖掘能自动在大型数据库里面寻找潜在的预测信息。传统上需要很多专家来进行分析的问题,现在可以快速而直接地从数据中间找到答案。一个典型的利用数据挖掘进行预测的例子就是目标营销。数据挖掘工具可以根据过去邮件推销中的大量数据找出其中最有可能对将来的邮件推销作出反应的客户。^[7]

事实上,数据挖掘目前对于各行各业来说都还是一个崭新却必要的概念。除金融行业外,电信、零售、税务、能源等行业都具有数据海量的特点和深度分析行业的需求。可以预见,这些行业的现有信息系统必然将向数据挖掘系统演化,数据挖掘的应用前景十分乐观。广阔的应用前景预示着新技术的研发价值。

1.3 课题来源及选题依据

本文的研究课题源自花旗软件的 ESP(面向欧洲国家的股票交易系统)项目。该项目是为花旗银行位于欧洲国家的分支机构提供网上股票交易服务。

1.4 本文研究的主要内容

本文的研究工作旨在建立股票交易系统增添股票走势预测功能提供理论依据,并对原交易系统不足之处进行改进,从而构建出新一代股票交易系统。

研究工作如下:

数据仓库建立:在第三章学习了数据仓库的相关技术后,在第四章将数据仓库技术应用到股票分析系统中。按照数据仓库的实现方法,对股票分析系统进行需求分析后,将各种数据转换成面向主题的格式,而后确定数据仓库中数据粒度层次划分和数据分割策略,采用星型模型表示数据之间的关系,完成数据仓库的建立,并实现与系统间的接口关系。

股票交易系统实现:首先对原系统进行分析,针对不足之处,进行技术上改进。并以订单交易模块为例,详述其开发及测试流程。

1.5 研究意义

通过课题的研究与实践，运用现代 IT 技术理论，总结并形成基于网络环境下股票交易的理论与方法，优化股票交易的体系结构，为客户提供全新便捷的电子银行服务，从而为开发出具有应用推广价值的新一代股票交易系统提供理论参考。

第二章 股票交易系统

2.1 股票交易相关概念

股票市场是上市公司筹集资金的主要途径之一。上市公司利用发行股票筹集资金，无须还本付息，只需在利润中划拨一部分出来支付红利即可，这对公司来说是相当有利的。所以发行股票来筹集资本就成为发展大企业经济的一种重要形式，而股票交易在整个证券交易中因此而占有相当重要的地位。

经营股票发行和买卖的场所称之为股票市场。一般来说，新股票的发行在初级市场，已发行的股票交易在次级市场。

初级市场(primary market) (见图 2-1)，又称发行市场。当公司初次发行股票时，必须找一家或几家综合券商承销(underwrite)其股票上市，券商会根据该公司过去三年的业绩、主要产品潜能与会计师签证报表，并替公司股票制定一个合理的价格。

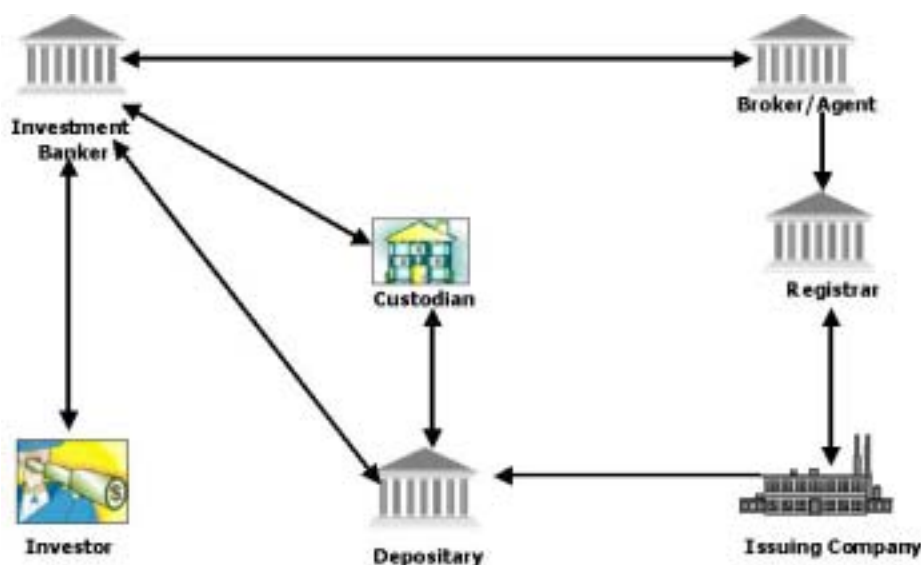


图 2-1 股票初级市场

Issuing Company: 证券发行人。证券发行人是指为筹措资金而发行债券、股票等证券的政府及其机构、金融机构、公司和企业。

Exchange: 证券交易所。证券交易所既不直接买卖证券，也不决定证券价格，而只为买卖证券的当事人提供场所和各种必要的条件及服务。

Broker: 经纪人/券商。一般客户要委托券商或经纪人进行证券交易。

Custodian: 托管人。替证券发行人保管证券并依照其指示进行处理的金融机构。

Depository: 保管机构。某国的上市公司为使其股票在外国流通，就将一定数额的股票，委托某一中间机构(通常为一银行，称为保管银行或受托银行)保管，由保管银行通知外国的存托银行在当地发行代表该股份的存托凭证，之后存托凭证便开始在外国证券交易所或柜台市场交易。

Investment Banker: 负责投资人交易资金费用处理的金融机构。

Investor: 证券投资人。证券投资者是指通过证券而进行投资各类机构法人或自然人，他们是证券市场的资金供给者。

次级市场 (secondary market) (见图 2-2), 投资者买卖已经发行的证券的市场。是指有价证券上市以后, 所有交易进行的地方, 又称流通市场。次级市场的证券价格是经过市场供需来决定, 反映出市场对各家公司的最新评价。投资者也能随时将有价证券在市场中出售, 换取现金, 去购买其他公司股票。

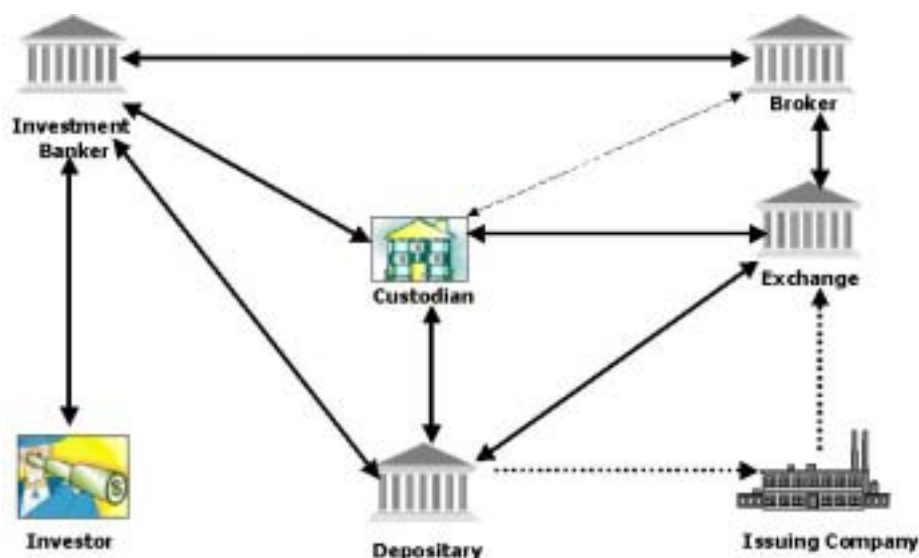


图 2-2 股票交易市场

初级市场和二级市场是相辅相成的，若发行市场不健全，发行证券的品质不好，或发行量太少又集中在少数人，则难以建立顺畅的流通市场。反之，若流通市场交易清淡，投资者需要的资金难以兑现，则认购新发行有价证券的意愿将大为降低，而使发行市场萎缩。

股票市场的变化与整个市场经济的发展是密切相关的,股票市场在市场经济中始终发挥着经济状况晴雨表的作用。

2.2 体系结构

21 世纪随着经济全球化、金融市场一体化及资产证券化的进程，世界经济运行的轴心正在转向现代金融业。这也使得金融证券业将面临者更大的挑战，市场竞争更加剧烈。现代 IT（信息技术）发展，尤其基于 Internet 技术的电子商务应用，使得整个企业运营模式正在发生着深刻变革，金融证券企业数字化运营已成为抗击市场竞争迎接挑战的必然发展趋势，营造一个高效合理的体系结构就极为重要。

花旗软件技术服务(上海)有限公司和花旗银行一样隶属于花旗集团,它专供为花旗银行提供软件开发和技术服务,在国际金融界技术领域方面处于领先地位。在长期的研究开发工作中,对各种银行应用、工具、资源库管理等方面有着深入的研究与探索,并已取得了实质性的进展和成果。

构建的股票交易系统系统结构建立在已有系统的基础上(截止到 2005 年,该系统已成功运行于花旗银行多家分支结构中),见图 2-3^[8-10]所示,系统采用 C/S 体系结构,银行的客户通过运行在 PC 上的浏览器访问银行的 Web Server,通过对用户的用户识别码和口令的检查,确认用户的合法性。Web Server 将客户的访问请求转变成对应的银行交易请求,并交给在 OS/390 上运行的交易处理系统处理,交易运行结果通过 Web Server 反馈给客户。

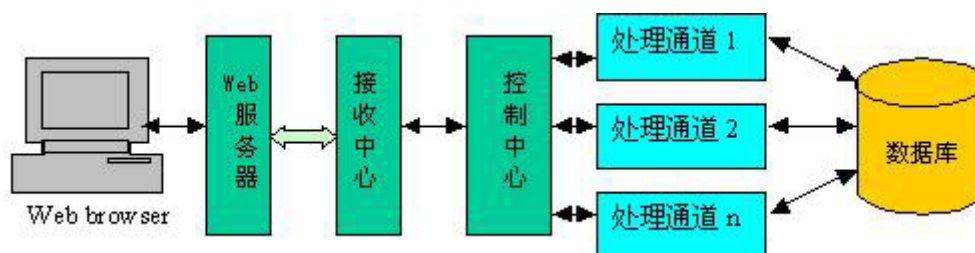


图 2-3 股票交易系统系统结构

Web browser 及 Web 服务器：为客户提供股票交易所需要的功能。在提供的界面下的功能，按照需求进行数据的输入，

接受中心：负责接受来自 Web 服务器的数据，并进行数据的格式化处理（除了对客户输入的数据进行规范化外，还在数据结尾处添加客户的请求）；当股票交易处理完成后，对产生的结果进行数据的格式化处理，反馈给前端显示。

控制中心：是整个股票交易系统的核心枢纽。不仅负责前端数据的正确性检验，还根据发送的需求为其选择一条处理通道。

处理通道：每条通道联接一项交易功能。接收交易数据并进行交易处理。

数据库：储存银行客户及各大股票交易所的股票信息。实时更新，从而保证数据的一致性。

网上股票交易系统支持股票走势查询、委托、撤单、成交回报、委托查询、成交查询、余股查询、帐户查询、网上开户、修改密码、银证转帐等功能，以及网上行情服务、个人资产管理及个性化服务等多种功能。

本系统提供全面的数字化经营运作模式，支持各业务工作系统的集成，实现综合数据共享，形成最高效率的协同工作能力。通过计算机网络数据信息快速流转，实现最高的工作效率，最终使企业运转达到高效规范的管理。

2.3 交易流程

交易流程见图 2-4^[11]。

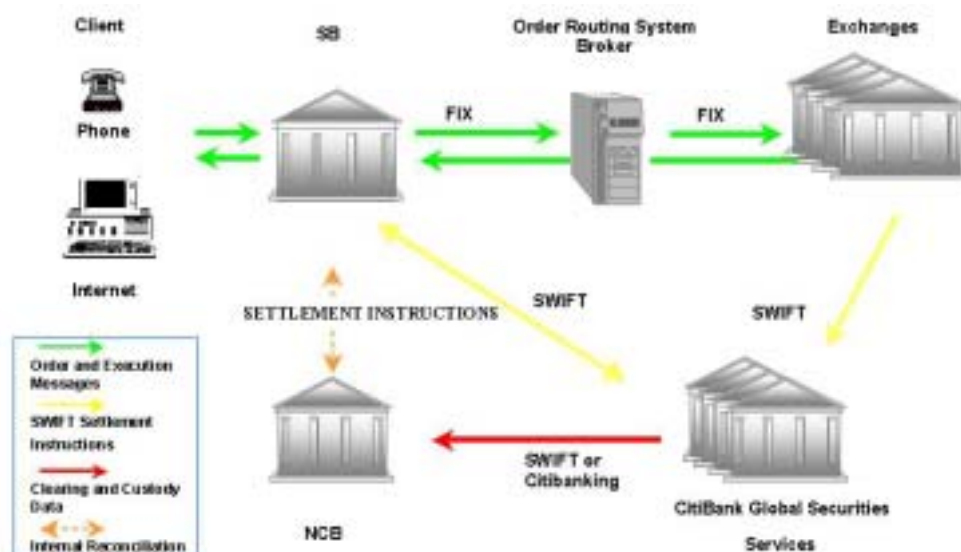


图 2-4 交易流程图

整个交易流程大致可划分交易前（开户）、交易中（下单、交割）及结算（过户）

三个阶段。现在它们大多数由七个步骤来处理：开户、下单，执行交易，确认，中台操作，清算和结算。首先，投资人应事前在券商处办理开户手续，完成开户后即可下单交易，既投资人通过电话、Internet 等通讯设备填写买卖委托单（图中通过股票交易系统），并将填好的委托单发往券商，券商收到委托后，处理后再发往证券交易所进行撮合交易，交易成交后，于第二日办理股票的转变及帐户资金费用的划分。

2.3.1 股票走势查询

本系统利用数据仓库与数据挖掘技术，构建智能决策系统，进行股票业务商业分析，提供股票指标的科学预测，辅助决策者进行合理投资。

客户在前端通过可利用的通讯设备，输入所要查询股票的代码/名称，系统将以图形或数字形式显示该股票的走势，从而为客户进行交易提供参考。见图 2-5a、图 2-5b。

其中，各股票意义如下：

自选股：安装客户所设定股票，查看其走势。

热门股：查看前一交易日成交量最大的 30 支股票走势。

分类股：选择股票类别后，查看属于同类股票走势。



图 2-5a 股票走势查询图

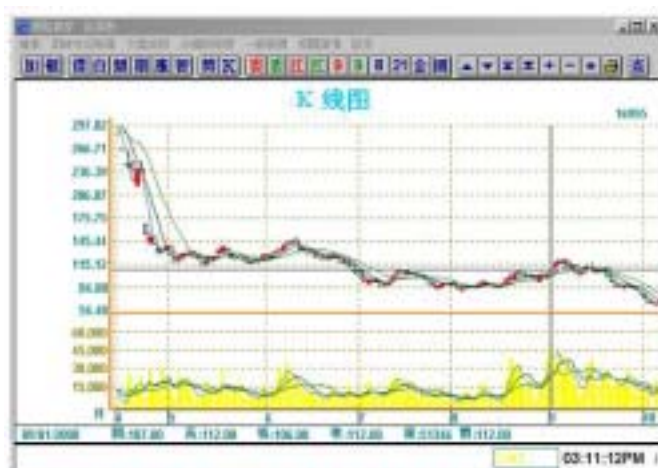


图 2-5b 股票走势查询图

2.3.2 填写股票订单

客户填写所要交易股票的代码、交易数量、限价、有效期限等信息，完成后提交给

系统处理。(见图 2-6)。



图 2-6 填写股票订单

2.3.3 鉴定订单

系统对订单信息进行授权处理，根据鉴定结果赋予订单不同状态。(见图 2-7)。

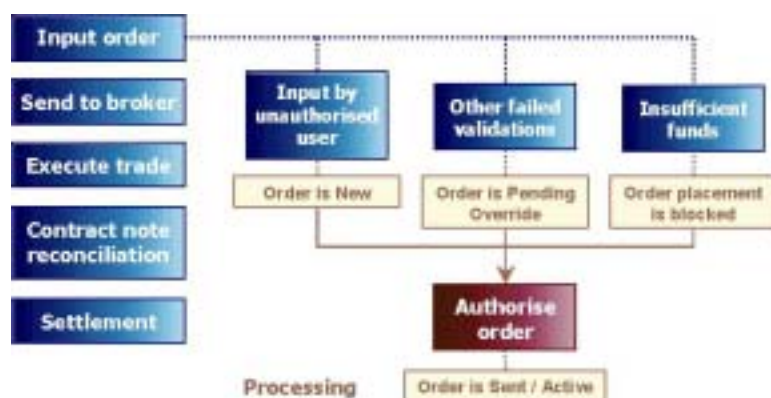


图 2-7 鉴定股票订单

2.3.4 发送订单

系统将鉴定通过的订单发往券商 (Broker)。(见图 2-8)。

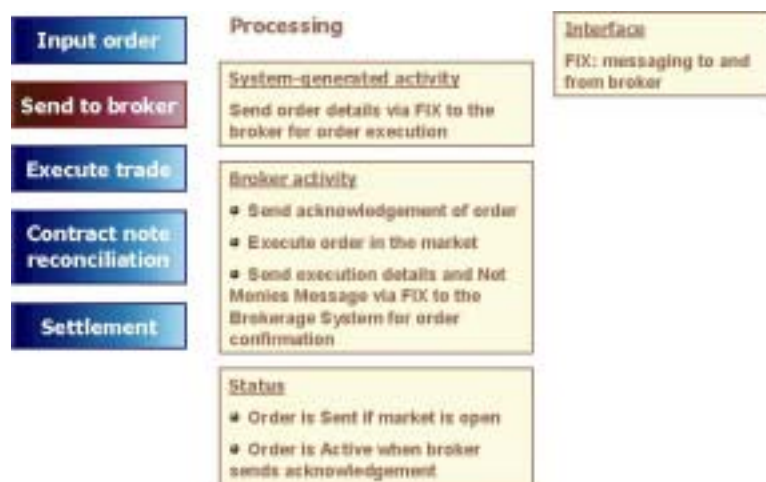


图 2-8 发送股票订单

2.3.5 修改订单

在订单未进入执行状态前，客户可以对订单进行修改。（见图 2-9）

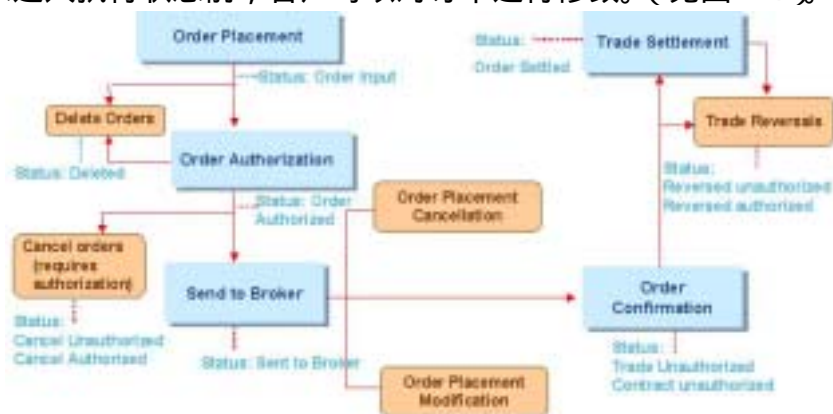


图 2-9 修改股票订单

2.3.6 执行交易

券商与证交所进行信息交互后，返回系统相关信息；系统根据信息进行交易数据的计算。（见图 2-10a、2-10b）

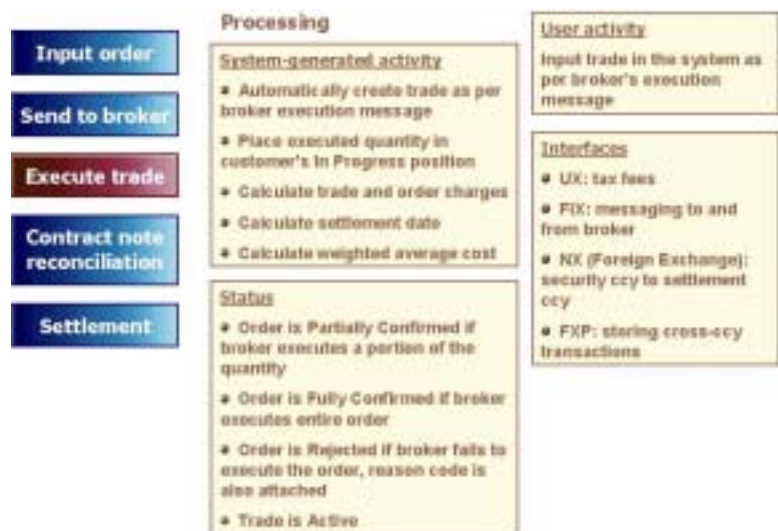


图 2-10a 执行交易

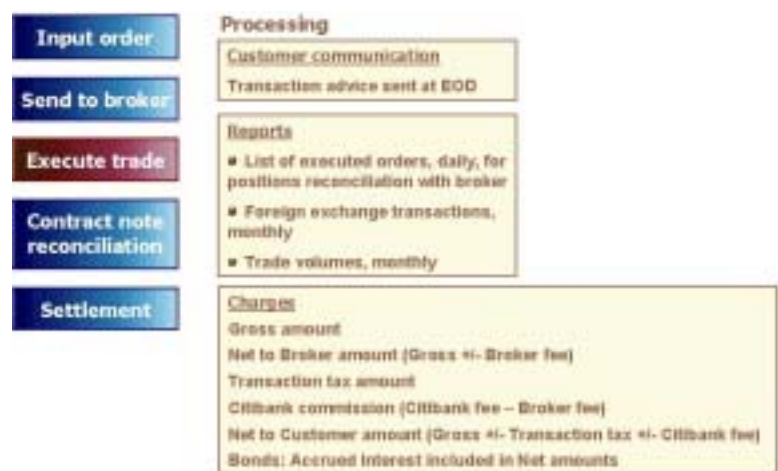


图 2-10b 执行交易

2.3.7 修改/鉴定交易

客户可更改交易所、交易日期、结算日期等信息；系统对交易数据进行鉴定。（见图 2-11）

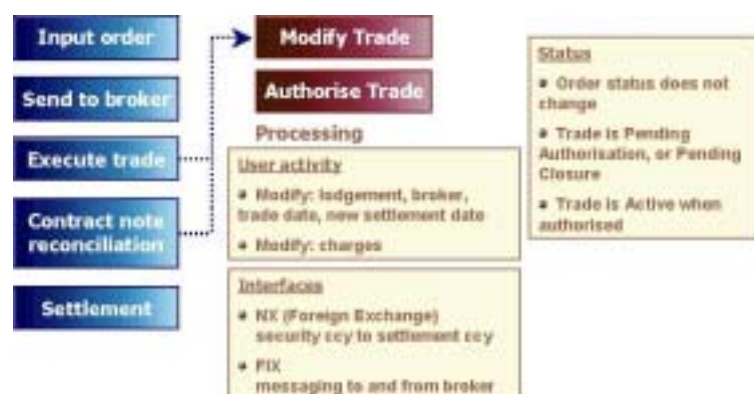


图 2-11 修改/鉴定交易

2.3.8 数据核对

系统要同托管机构/券商进行数据核对。（见图 2-12）

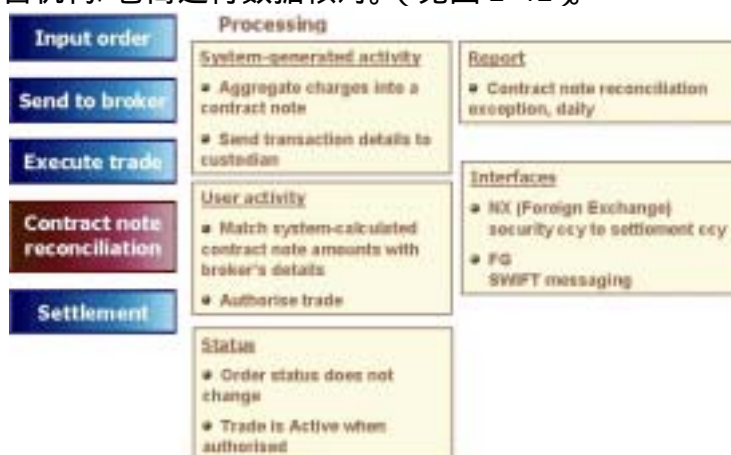


图 2-12 数据核对

2.3.9 结算

按照实际交易额等进行各对应处股票数量、客户帐户金额、费用的相关处理，完成股票交易。（见图 2-13）

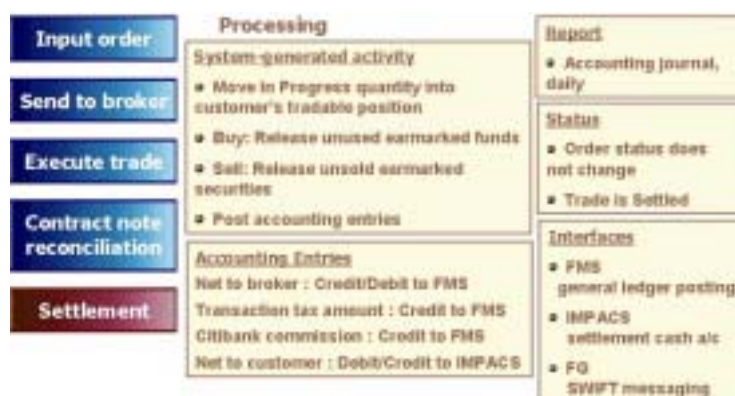


图 2-13 结算

2.4 本章小结

本章简单介绍了股票交易的相关概念,以某商业银行的股票交易系统为基础,给出了股票交易的整个流程。

第三章 数据仓库和数据挖掘技术分析

3.1 引言

目前,数据仓库和数据挖掘技术是国际上数据库、信息决策和机器学习等领域最前沿的研究方向之一。数据挖掘的研究已经和数据仓库的研究结合起来,两者协同工作,一方面,可以迎合和简化数据挖掘过程中的重要步骤,提高数据挖掘的效率和能力,确保数据挖掘中数据来源的广泛性和完整性。另一方面,数据挖掘技术已经成为数据仓库应用中极为重要和相对独立的方面和工具。

数据挖掘和数据仓库是融合与互动发展的,其学术研究价值和应用研究前景将是令人振奋的。它是数据挖掘专家、数据仓库技术人员和行业专家共同努力的成果,更是广大渴望从数据库“奴隶”到数据库“主人”转变的企业最终用户的通途。

本章主要介绍数据仓库和数据挖掘的定义及关键技术,并且讨论基于数据仓库的数据挖掘技术的新特点。

3.2 数据仓库技术

数据仓库技术所要研究和解决的问题就是从 OLTP 系统、异构分散的外部数据源、脱机的历史业务数据中获取数据,处理后为数据分析和决策提供应用服务。数据仓库技术是计算机应用发展的必然产物,是对管理信息系统开发和应用的经验总结的升华。

3.2.1 数据仓库的定义

公认的数据仓库概念是 W. H. Inmon 在《建立数据仓库》一书中提出的:数据仓库就是面向主题的、集成的、不可更新的(稳定性)随时间不断变化(不同时间)的数据集合,用以支持经营管理中的决策制定过程。

根据数据仓库概念的含义,数据仓库拥有以下四个特点:

(1) 面向主题。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。面向事务处理的操作型数据库通常与某些特定的应用相关,数据库之间相互独立,并且往往是异构的。而数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的,必须消除源数据中的不一致性,以保证数据仓库内的信息是关于整个企业的一致全局信息。

(3) 相对稳定的。操作型数据库中的数据通常实时更新,数据根据需要及时发生变化。数据仓库的数据主要供企业决策分析之用,所涉及的数据操作主要是数据查询,一旦某个数据进入数据仓库以后,一般情况下将被长期保留,也就是数据仓库中一般有大量的查询操作,但修改和删除操作很少,通常只需要定期的加载、刷新。

(4) 反映历史变化。操作型数据库主要关心当前某一个时间段内的数据,而数据仓库中的数据通常包含历史信息,系统记录了企业从过去某一时点(如开始应用数据仓

库的时点)到目前的各个阶段的信息,通过这些信息,可以对企业的发展历程和未来趋势做出定量分析和预测。

企业数据仓库的建设,是以现有企业业务系统和大量业务数据的积累为基础。数据仓库不是静态的概念,只有把信息及时交给需要这些信息的使用者,供他们做出改善其业务经营的决策,信息才能发挥作用,信息才有意义。而把信息加以整理归纳和重组,并及时提供给相应的管理决策人员,是数据仓库的根本任务。因此,从产业界的角度看,数据仓库建设是一个工程、一个过程。

需要指出的是,数据仓库中的数据并不是最新的、专有的,而是来源于其它数据库的。数据仓库的建立并不是要取代数据库,它要建立一个较全面和完善的信息服务应用的基础上,用于支持高层决策分析,而原有的事务处理数据库在总体数据环境中承担的是日常基础业务的处理任务。数据仓库是数据库技术的一种新的应用,而且到目前为止,数据仓库大部分还是用关系数据库管理系统来管理其中的数据。

与关系数据库不同的是,数据仓库至今并没有严格的数学理论基础,它更偏向于工程。由于数据仓库的这种工程特性,因而在技术上可以根据它的工作过程分为:数据的抽取、数据的存储和管理、数据的展现等关键技术。

(1) 数据的抽取

数据的抽取是数据进入仓库的入口。由于数据仓库是一个独立的数据环境,它需要通过抽取过程将数据从联机事务处理系统、外部数据源、脱机的数据存储介质中导入数据仓库。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的处理。在数据抽取方面,未来的技术发展将集中在系统功能集成化方面,以适应数据仓库本身或数据源的变化,使系统更便于管理和维护。

(2) 数据的存储和管理

数据仓库的组织管理方式决定了它有别于传统数据库的特性,也决定了其对外部数据的表现形式。数据仓库管理所涉及的数据量比传统事务处理大得多,且随时间的推移而快速累积。在数据仓库的数据存储和管理中需要解决的是如何管理大量的数据、如何并行处理大量的数据、如何优化查询等。目前,许多数据库厂家提供的技术解决方案是扩展关系型数据库的功能,将普通关系数据库改造成适合担当数据仓库的服务器。

(3) 数据的展现

在数据展现方面主要的方式有:

查询:实现预定义查询、动态查询、OLAP 查询与决策支持智能查询;报表:产生关系数据表格、复杂表格、OLAP 表格、报告以及各种综合报表;可视化:用易于理解的点线图、直方图、饼图、网状图、交互式可视化、动态模拟、计算机动画技术表现复杂数据及其相互关系;统计:进行平均值、最大值、最小值、期望、方差、汇总、排序等各种统计分析;挖掘:利用数据挖掘等方法,从数据中得到关于数据关系和模式的知识^[6]。

3.2.2 与数据仓库相关的几个概念^[12-14]

(1) 粒度

粒度是数据仓库的重要概念。粒度可以分为两种形式,第一种粒度是对数据仓库中的数据的综合程度高低的一个度量,它既影响数据仓库中的数据量的多少,也影响数据仓库所能回答询问的种类。在数据仓库中,多维粒度是必不可少的。由于数据仓库的主要作用是 DSS 分析,因而绝大多数查询都基于一定程度的综合数据之上的,只有极少数查询涉及到细节。所以应该将大粒度数据存储于快速设备如磁盘上,小粒度数据存于低速设备如磁带上。

还有一种粒度形式，即样本数据库。它根据给定的采样率从细节数据库中抽取出一个子集。这样样本数据库中的粒度就不是根据综合程度的不同来划分的，而是有采样率的高低来划分，采样粒度不同的样本数据库可以具有相同的数据综合程度。

（2）分割

分割是数据仓库中的另一个重要概念，它的目的同样在于提高效率。它是将数据分散到各自的物理单元中去，以便能分别独立处理。有许多数据分割的标准可供参考：如日期、地域、业务领域等等，也可以是其组合。一般而言，分割标准总应包括日期项，它十分自然而且分割均匀。

（3）数据集市

人们在早期开发企业级数据仓库时发现了一些问题，为了解决这些问题，人们提出了数据集市的概念。

数据集市是支持某一部门或特定商业需求的 DSS 应用的数据集合。数据集中数据仍具有数据仓库中数据的特点。只不过，数据集中的数据是为了某一个部门或某个特定商业需求所定制。数据集市和数据仓库类似，一般认为数据集中的数据来自数据仓库。

数据集市具有建设周期短(只需几个月)，投资小、见效快的优点；由于部门用户完全拥有自己的数据集市，因而可以根据需要定制其中的数据；部门用户可以自由地选择适合自己的数据集市的决策分析工具；数据集市的使用可以解决各部门对数据仓库资源的竞争，提高处理速度。

（4）元数据(Meta data)

元数据是关于数据的数据，它描述的是数据的结构、内容、码、索引等项内容。设计一个内容完善的元数据是有效管理数据仓库的重要前提。

主要有三类元数据需要管理：

管理性元数据：包括用于建立和使用数据仓库的信息，如元数据的描述、数据仓库结构的定义，派生数据的定义，预定义查询和报表的定义，数据集市定义和内容，数据提取、数据净化，转换规则，数据刷新和删除的策略等；

应用领域元数据：如商业用术语概念，数据拥有者等；

操作性元数据：是在操作数据仓库过程中收集的信息，如迁移和改变的历史过程、数据仓库中数据的流通过程(激活、改进、删除等)。

有了元数据就等于拥有了数据仓库的一张蓝图。元数据最显著的作用就是它类似于现实生活中的地图，能指引用户在多达几百 GB 的数据仓库的数据海洋中找到自己所需要的数据，帮助用户更好地理解仓库中的数据，如果不同厂商的数据仓库及联机分析处理工具都遵循统一的元数据交换标准，则不同厂商的数据仓库及联机分析处理工具之间可以通过元数据方便地共享和交换数据。

（5）在线分析处理(OLAP)

在线分析处理是以数据库和数据仓库为基础的，其最终数据来源与 OLTP 一样，均来自底层的数据库系统。OLTP 面对的是操作人员和低层管理人员，OLAP 面对的是决策人员和高层管理人员，所使用的经综合提炼的历史数据均来自 OLTP 所依赖的底层数据库，OLAP 数据较之 OLTP 数据要多一步数据多维化或预综合处理。OLAP 是一类软件技术，它使分析人员、经理、管理人员通过对 OLAP 数据的多种可能的观察角度进行快速、一致、交互性的存取以获得对信息的深入了解。

3.2.3 数据仓库体系结构^[13]

在数据仓库系统中，数据源一般是异构的，通过网络连接。数据仓库能通过 ODBC

之类的机制，访问各数据源，从数据源选取决策所需的数据，然而在数据资源中存在大量的数据缺陷：不一致的字段长度、不一致的赋值、丢失的数据输入、违背数据的完整性等。转换过程包括：数据提取、数据净化、数据载入、数据刷新等。经过转换数据资源中的原始数据变为数据仓库中面向主题的、整合数据，但可能要做必要的转换。

为了能够从已有的数据源中提取出来数据，并经过处理，组织成可用于决策分析所需的综合数据的形式，一个数据仓库的基本体系结构（见图 3-1）中应有以下几个基本组成部分：

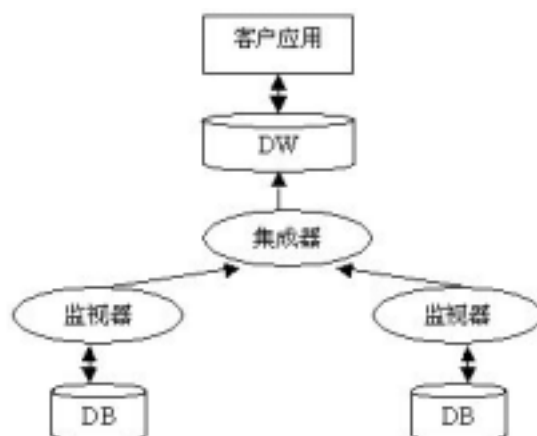


图 3-1 数据仓库基本体系结构

数据源：指为数据仓库提供最底层数据的运作数据库系统及外部数据。

监视器：负责感知数据源发生的变化，并按数据仓库的需求提取数据。

集成器：将从运作数据库中提取的数据经过转换、计算、综合等操作，集成到数据仓库中。

数据仓库：存贮已经按企业级视图转换的数据，供分析处理用。根据不同的分析要求，数据按不同的综合程度存储。数据仓库中还应存储元数据，其中记录了数据的结构和数据仓库的任何变化，以支持数据仓库的开发和使用。

客户应用：供用户对数据仓库中的数据进行访问查询，并以直观的方表示分析结果的工具。

3.2.4 数据仓库的数据组织结构

数据仓库是存储供查询和决策分析用的集成化信息仓库，是存储数据的一种组织形式。一个典型的数据仓库的数据组织结构如图 3-2 所示。

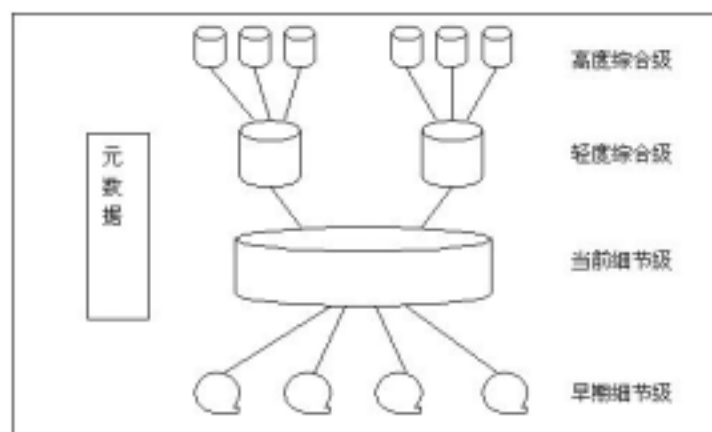


图 3-2 DW 数据组织结构

数据仓库中的数据分为四个级别:早期细节级、当前细节级、轻度综合级和高度综合级。原始数据经过综合、集成后,首先进入当前细节级(即当前基本数据层),并根据辅助决策的具体需要进行进一步地综合,从而进入轻度综合级乃至高度综合级。随着时间的推移,由时间控制机制将老化的数据转入早期细节级(即历史数据层)。从中我们可以看出,数据仓库中存在着不同的综合级别,我们称之为“粒度”。粒度越高,表示细节程度越低、综合程度越高。数据级别的划分是根据粒度进行的。

粒度是数据仓库的重要概念,它既影响到数据仓库中数据量的多少,也影响到数据仓库所能回答查询的种类。粒度越低,细节程度越高,回答查询的种类就越多。但这必然会造成数据仓库中数据的大量堆积,这样,在查询综合性问题是,需要从大量细节数据中综合并计算答案,效率将会十分低下;另一方面,粒度的提高将会提高综合查询效率,但同时也造成了回答细节问题能力的下降。

因此,为了提高数据仓库的性能,在数据仓库中,多重粒度是必不可少的。

数据仓库中还有一种重要的数据--元数据(metadata)。元数据是“关于数据的数据”,如在传统数据库中的数据字典就是一种元数据。在数据仓库环境下,主要有两种元数据:第一种是为了从操作性环境向数据仓库转化而建立的元数据,包含了所有源数据项名、属性及其在数据仓库中的转化;第二种元数据在数据仓库中是用来和终端用户的多维商业模型/前端工具之间建立映射,此种元数据称之为DSS元数据,常用来开发更先进的决策支持工具。

数据仓库需求的确定

数据仓库系统的使用者是企业各级的决策和业务人员,他们关心的问题 and 一般的操作人员不同,他们所要求的数据有:是否有异常情况;公司历来哪些客户最重要,他们如今销售如何;公司某产品今年按月计在各地销售走势如何;公司某产品历年销售中各规格比例如何,现库存状况是否满意,从这些问题中可以发现,主要是这些一些信息:

(1)事实(Facts)。如销售量、销售额、库存量、库存额和应收帐款等。这些数据是实际分析的基础数据,它们日积月累,数量庞大。

(2)维(Dimensions)。它是事实信息的属性,如销售发生的时间、客户、部门,销售的是何种产品、何种规格等,它们一般变化不大,数量也相对较小。

(3)粒度(Units)。它是维划分的单位,如时间维可按日计,也可按旬、按月、按年计;如产品维,可明细到规格或颜色,也可按款式、面料的较粗的单位来统计。这些信息一般没有变化。

数据仓库模型设计

设计数据仓库的数据模型时,常用的逻辑建模工具是以维数据建模来表示事实、维、粒度的关系,结构形式是星形模型和雪花模型。

1)星形模型。星形模型的核心是事实表(Fact table)。它是按维进行查询的中心——存储事实数据即事实的地方。事实为数字型的属性,如计算和数量,可以求和、求平均、求最大及最小,并且按照各种统计运算进行合计计算。事实属性包含了适度的、关于事实表所管理的内容的数字型值。

与传统的正规图模型相比,星形模型最适合查询。由于主要数据都在庞大的事实表中,所以只要扫描事实就可以进行查询,而不必把多个庞大的表连接起来。同时由于维表一般都很小,甚至可以放在高速缓存中,与事实表作连接时其速度较快。另一方面,对于非计算机专业的用户而言,星形模型比较直观,通过分析星形模型,很容易组合出各种查询。如通过部门和商品两个维,可分析部门在相同产品上的销售情况,便于各部门发现各自的优势,取长补短。又如通过商品、客户和时间三维,可分析不同时间的不

同要求。相应地，可以对不同的客户提供不同的服务，在不同的时间采取不同的定价。

2) 雪花模型。雪花模型是星型模型的扩展。它比星型模型增加了层次结构，体现了维的不同粒度的划分。使用雪花模型进一步增加了查询的范围，以帮助市场部门决定市场开发的策略。也能回答如某一地区某个商品去年共销售多少、人均销售如何的问题，以预计下一年不同地区的销量，为制定生产计划提供依据。但这种方式可以使系统进一步专业化和实用化，同时降低了系统的通用程度。前端工具仍然要用户在雪花的逻辑概念模式上操作，然后将用户的操作转换为具体的物理模式，从而完成对数据的查询。

3.2.5 数据仓库的实现策略^[15]

随着数据仓库技术的发展，如今数据仓库的实现策略已从原先的“自顶向下”模式发展为“自顶向下”、“自底向上”、“平行开发”三大类。

1) “自顶向下”模式

“自顶向下”的开发策略是指从原来分散存储在企业各处的 OLTP 数据库中的有用数据通过提取、净化、载入、刷新等处理步骤建立一个全局性数据仓库。这个全局数据仓库将提供给用户一个一致的数据格式和软件环境。从理论上说，决策支持所需的数据都应该包含在这个全局数据库中。数据集市存储的数据是为某个部门的 DSS 应用而专门从全局数据仓库中提取的，它是全局数据仓库中数据的一个子集。在“自顶向下”模式中，数据集市和数据仓库的关系是单方向的，即数据从数据仓库流向数据集市，如图 3-3 所示。

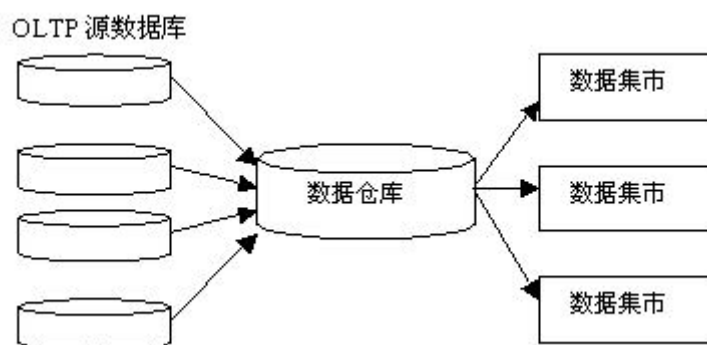


图 3-3 “自顶向下”模式

2) “自底向上”模式

“自底向上”模式是从各个部门或特定的商业问题的数据集市开始，全局性数据仓库是建立在这些数据集市的基础上。“自底向上”模式的特点是初期投资少，见效快。因为它在建立部门数据集市是只需要较少的人做出决策，解决的是较小的商业问题。“自底向上”的开发模式可以使一个单位在数据仓库发展初期尽可能少的花费资金，也可以在作出有效的投入之前评估技术的收益情况，如图 3-4 所示。

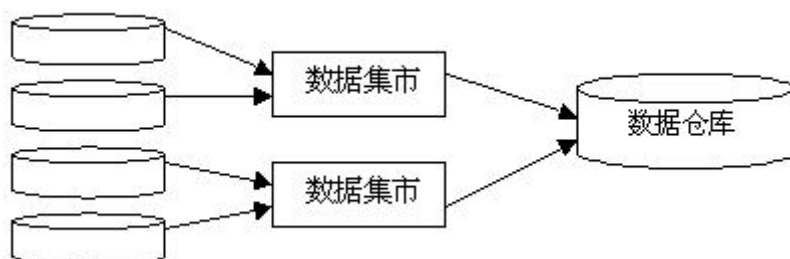


图 3-4 “自底向上”模式

3) “平行开发”模式

“平行开发”模式是指在一个全局数据仓库的数据模型的指导下,将数据集市建立和全局性数据仓库的建立同时进行。在“平行开发”模式中由于数据集市的建立是在一个统一的全局数据模型的指导下进行的,可避免各部门在开发各自的数据集市时的盲目性,减少各数据集市之间的数据冗余和不一致。事实上,一些部门在建立数据集市的过程中所遇到的问题及其解决方案、所获得的经验将导致全局性数据仓库的数据模型作出相应的改变,这些变化将使其他部门在建立集市时受益,也有助于全局性数据仓库的建设。在“平行开发”模式中数据集市的这种相对独立性有利于全局性数据仓库的建设。一旦全局性数据仓库建立好以后,各部门的数据集市将成为全局数据仓库的一个子集,全局数据仓库将负责为各部门已建成和即将要建的数据集市提供数据。如图 3-5 所示。

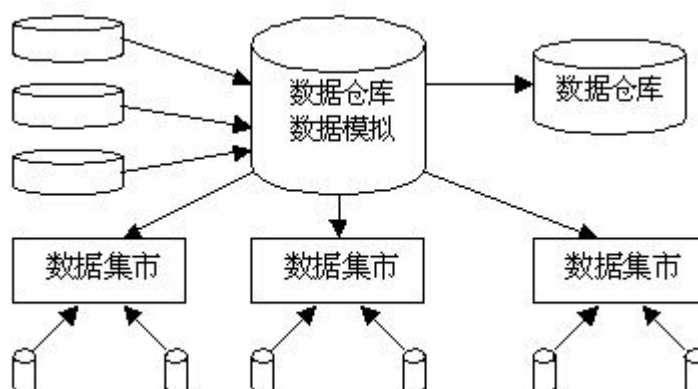


图 3-5 “平行开发”模式

通过以上的讨论,我们知道:数据仓库是一种解决问题的方案,而不是可以买到的现成产品。它是一数据库技术作为存储数据和管理资源的基本手段,以统计分析技术作为分析数据和提取信息的有效方法,以人工智能技术作为挖掘知识和发现规律的科学途径,因此它是多学科相互结合、综合应用的技术。

3.3 数据挖掘技术^[4]

数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、模糊的、随机的数据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。随着信息技术的高速发展,人们积累的数据量急剧增长,动辄以 TB 计,如何从海量的数据中提取有用的知识成为当务之急。数据挖掘就是为顺应这种需要应运而生发展起来的数据处理技术。是知识发现(Knowledge Discovery in Database)的关键步骤。对于那些决策者明确了解的信息,可以用查询、联机分析处理(OLAP)或其他工具直接获取,比如“列出各子公司在上个月的销售情况”。而另外一些隐藏在大量数据中的关系、趋势,即使是管理这些数据的专家也是没有能力发现的。这些信息对于决策可能又是至关重要的。这就需要数据挖掘来解决。

3.3.1 数据挖掘的任务

数据挖掘的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,从而从大量数据中发现知识。

(1) 关联分析(association analysis)

两个或两个以上变量的取值之间存在某种规律性,就称为关联。数据关联是数据库中存在的一类重要的、可被发现的知识。关联分为简单关联、时序关联和因果关联。关联分析的目的是找出数据库中隐藏的关联网。一般用支持度和可信度两个阈值来度量关

联规则的相关性，还不断引入兴趣度、相关性等参数，使得所挖掘的规则更符合需求。

(2) 聚类分析(clustering)

聚类是把数据按照相似性归纳成若干类别，同一类中的数据彼此相似，不同类中的数据相异。聚类分析可以建立宏观的概念，发现数据的分布模式，以及可能的数据属性之间的相互关系。

(3) 分类(classification)

分类就是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述，并用这种描述来构造模型，一般用规则或决策树模式表示。分类是利用训练数据集通过一定的算法而求得分类规则。分类可被用于规则描述和预测。

(4) 预测(predication)

预测是利用历史数据找出变化规律，建立模型，并由此模型对未来数据的种类及特征进行预测。预测关心的是精度和不确定性，通常用预测方差来度量。

(5) 时序模式(time-series pattern)

时序模式是指通过时间序列搜索出的重复发生概率较高的模式。与回归一样，它也是用已知的数据预测未来的值，但这些数据的区别是变量所处时间的不同。

(6) 偏差分析(deviation)

在偏差中包括很多有用的知识，数据库中的数据存在很多异常情况，发现数据库中数据存在的异常情况是非常重要的。偏差检验的基本方法就是寻找观察结果与参照之间的差别。

3.3.2 数据挖掘对象

根据信息存储格式，用于挖掘的对象有关系数据库、面向对象数据库、数据仓库、文本数据源、多媒体数据库、空间数据库、时态数据库、异质数据库以及 Internet 等。

3.3.3 数据挖掘和数据仓库

大多数情况下，数据挖掘都要先把数据从数据仓库中拿到数据挖掘库或数据集市。从数据仓库中直接得到进行数据挖掘的数据有许多好处。数据仓库的数据整理和数据挖掘的数据整理差不多，如果数据在导入数据仓库时已经清理过，那很可能在做数据挖掘时就没必要在清理一次了，而且所有的数据不一致的问题都已经解决了。

数据挖掘库可能是数据仓库的一个逻辑上的子集，而不一定非得是物理上单独的数据库。但如果数据仓库的计算资源已经很紧张，最好还是建立一个单独的数据挖掘库。

当然为了数据挖掘不必非得建立一个数据仓库，数据仓库不是必需的。建立一个巨大的数据仓库，把各个不同源的数据统一在一起，解决所有的数据冲突问题，然后把所有的数据导到一个数据仓库内，是一项巨大的工程，可能要用几年的时间花上百万的钱才能完成。只是为了数据挖掘，你可以把一个或几个事务数据库导到一个只读的数据库中，就把它当作数据集市，然后在它上面进行数据挖掘。

3.3.4 数据挖掘和在线分析处理(OLAP)

一个经常问的问题是，数据挖掘和 OLAP 到底有何不同。下面将会解释，他们是完全不同的工具，基于的技术也大相径庭。

OLAP 是决策支持领域的一部分。传统的查询和报表工具是告诉你数据库中都有什么(what happened)，OLAP 则更进一步告诉你下一步会怎么样(What next)、和如果我采取这样的措施又会怎么样(What if)。用户首先建立一个假设，然后用 OLAP 检索数据库来验证这个假设是否正确。比如，一个分析师想找到什么原因导致了贷款拖欠，他可能先做一个初始的假定，认为低收入的人信用度也低，然后用 OLAP 来验证他这个假设。

如果这个假设没有被证实,他可能去察看那些高负债的账户,如果还不行,他也许要把收入和负债一起考虑,一直进行下去,直到找到他想要的结果或放弃。

也就是说,OLAP 分析师是建立一系列的假设,然后通过 OLAP 来证实或推翻这些假设来最终得到自己的结论。OLAP 分析过程在本质上是一个演绎推理的过程。但是如果分析的变量达到几十或上百个,那么再用 OLAP 手动分析验证这些假设将是一件非常困难和痛苦的事情。

数据挖掘与 OLAP 不同的地方是,数据挖掘不是用于验证某个假定的模式(模型)的正确性,而是在数据库中自己寻找模型。他在本质上是一个归纳的过程。比如,一个用数据挖掘工具的分析师想找到引起贷款拖欠的风险因素。数据挖掘工具可能帮他找到高负债和低收入是引起这个问题的因素,甚至还可能发现一些分析师从来没有想过或试过的其他因素,比如年龄。

数据挖掘和 OLAP 具有一定的互补性。在利用数据挖掘出来的结论采取行动之前,你也许要验证一下如果采取这样的行动会给公司带来什么样的影响,那么 OLAP 工具能回答你的这些问题。

而且在知识发现的早期阶段,OLAP 工具还有其他一些用途。可以帮你探索数据,找到哪些是对一个问题比较重要的变量,发现异常数据和互相影响的变量。这都能帮你更好的理解你的数据,加快知识发现的过程。

3.3.5 数据挖掘流程^[16-17]

数据挖掘过程一般由 3 个主要的阶段组成:数据准备、挖掘操作、结果表达和解释。知识的发现可以描述为者 3 个阶段的反复过程。

数据准备:这个阶段又可进一步分为 3 个子步骤:数据集成、数据选择、数据预处理。数据集成将多文件或多数据库运行环境中的数据进行合并处理,解决语义模糊性、处理数据中遗漏和清洗脏数据等。数据选择的目的是辨别出需要分析的数据集合,缩小处理范围,提高数据挖掘的质量,预处理是为了克服目前数据挖掘工具的局限性。

数据挖掘:这个阶段进行实际的挖掘操作,包括的要点有:

(1) 要先决定如何产生假设,是让数据挖掘系统为用户产生假设,还是用户自己对于数据库中可能包含的知识提出假设。前一种称为发现型(Discovery-driven)的数据挖掘,后一种称为验证型(Verification-driven)的数据挖掘。

(2) 选择合适的工具。

(3) 挖掘知识的操作。

(4) 证实发现的知识。

结果表达和解释:对数据挖掘的结果进行解释和评价,转换成为能够最终被用户理解的知识。并将分析所得到的知识集成到业务信息系统的组织结构中去。根据最终用户的决策目的对提取的信息进行分析,把最有价值的信息区分出来,并且通过决策支持工具提交给决策者。因此,这一步骤的任务不仅是把结果表达出来(例如采用信息可视化方法),还要对信息进行过滤处理。如果不能令决策者满意,需要重复以上数据挖掘的过程。

3.4 基于数据仓库的数据挖掘技术

大多数情况下,数据挖掘都要先把数据从数据仓库中拿到数据挖掘库或数据集市。从数据仓库中直接得到进行数据挖掘的数据有许多好处。数据仓库的数据整理和数据挖掘的数据整理差不多,如果数据在导入数据仓库时已经清理过,那很可能在做数据挖掘时就没必要在清理一次了,而且所有的数据不一致的问题都已经解决了。

数据挖掘库可能是数据仓库的一个逻辑上的子集,而不一定非得是物理上单独的数据库。但如果数据仓库的计算资源已经很紧张,最好还是建立一个单独的数据挖掘库。

当然为了数据挖掘不必非得建立一个数据仓库,数据仓库不是必需的。建立一个巨大的数据仓库,把各个不同源的数据统一在一起,解决所有的数据冲突问题,然后把所有的数据导到一个数据仓库内,是一项巨大的工程,可能要用几年的时间花上百万的钱才能完成。只是为了数据挖掘,你可以把一个或几个事务数据库导到一个只读的数据库中,就把它当作数据集市,然后在它上面进行数据挖掘。

在数据仓库中进行数据挖掘有新的特点。这些新的特点,不论是使数据挖掘更容易或更困难,都体现在数据仓库与目前数据挖掘的主要对象—数据库的本质区别上面。

规模:数据仓库中集成和存储着来自若干分布、异质的信息源的数据。这些信息源本身就可能是一个规模庞大的数据库,可以想见数据仓库会比一般数据库系统更大的数据规模。目前数据仓库的规模都超过 50GB,将来会更大。同时,数据仓库为了能进行长期趋势的分析,保留了长时间的历史数据。

数据集成和综合性:从一个企业的角度看,数据仓库集成了企业内各部门的全面的、综合的数据。数据挖掘要面对的是关系更复杂的企业全局模式的知识发现。从这一点上讲,基于数据仓库的数据挖掘能更好地满足高层战略决策的要求。而且,数据仓库机制大大降低了数据挖掘的障碍,一般进行数据挖掘要花大量的力量在数据准备阶段,而在数据仓库中数据已经被充分收集起来,进行了整理、合并,并且有些还进行了初步的分析处理。这样,注意力更集中于数据挖掘的核心处理阶段。另外,数据仓库中对数据不同粒度的集成和综合,更有效地支持了多层次、多种知识的挖掘。

查询支持:数据仓库面向决策支持,数据仓库的体系结构努力保证查询和分析的实时性。而一般的联机事物处理系统主要要求更新的实时性,对查询的性能要求相对较弱。一般的数据仓库设计成只读方式,最终用户不能直接更新数据仓库,数据更新由专门的一套机制保证,通常由系统自动更新和管理员控制来协同完成。数据仓库对查询的强大支持使数据挖掘效率更高,挖掘过程可以做到实时交互,使决策者的思维保持连续,有可能挖掘出更深入、更有价值的知识。

3.5 本章小结

从以上分析可以看出,数据仓库在纵向和横向都为数据挖掘提供了更广阔的活动空间。数据仓库完成了数据的收集、集成、存储、管理等工作,数据挖掘面对的是经初步加工的数据,使得数据挖掘能更专注于知识的发现;另一方面,由于数据仓库具有的新的特点,又对数据挖掘技术提出了更高的要求。可以说,数据挖掘技术要充分发挥潜力,就必须和数据仓库的发展结合起来。

随着数据仓库和数据挖掘的集成的进一步深化,必然给用户带来更大的利益,人们对它的研究正也日益广泛和深入。

第四章 股票交易数据的数据仓库建立

4.1 引言

成功的决策支持系统都应该有三个阶段：第一阶段是构造基础设施，即建立联机事物处理系统。其目的是让大量原始数据电子化，为以后的构造数据仓库提供物质基础。第二阶段是系统性能优化阶段。目的是在系统性能、安全性和可靠性进行必要的优化。第三阶段实现信息价值阶段。即利用数据挖掘技术挖掘数据中潜在的信息价值，以便为企业决策者提供决策上的帮助。

证券业作为计算机应用密集的行业，通过十几年的发展，每个证券公司都积累了大量的数据。怎样利用这些数据，深层次地挖掘数据资源并进行分析，使公司的决策者能及时掌握公司的运行情况，并根据这些分析结果制定长远规划，从而提高公司的管理水平和竞争优势，成为证券公司技术部门目前的努力方向^[18-19]。

为了更好地利用股票交易数据，为客户分析提供依据，我们设计并创建了一个基于数据仓库的股票交易数据分析系统。设计的目的不仅仅实现对股票交易数据的简单维护和查询，而是更希望能有效地对变化的股市环境进行分析，使信息不仅能反映市场变化的瞬间状态，而且还能反映它的历史及趋势，得到数据更深层次的信息，为客户交易提供决策支持。

本章讨论股票交易数据的数据仓库的系统设计和实现方法。

4.2 引言

一个好的项目开始于一个好的需求分析。因此，数据仓库的建设工作也就应该从分析数据入手。分析数据有两个方面：一方面是从数据中提取需求；另一方面是从需求中找数据。所谓从数据中提取需求，就是通过考察原有系统中所保存的数据，分析从这些数据中能抽取哪些对决策有用的信息，获得用户对要构建系统的需求。这种做法适合于从项目开始阶段的原型设计。而从需求中找数据，则是对用户的某一项应用需求，考察能否从业务系统中抽取以及怎样抽取所需的信息。

数据仓库设计和传统的数据库设计是不同的，数据库中处理事物的需求是较固定的，一旦构建数据库后，系统的需求基本不变。而在数据仓库建设中，对需求分析的过程贯穿在整个设计的过程中，在整个系统的建设过程中完成对系统需求的收集、分析和理解。因此数据仓库的设计是一个动态的反馈和循环的过程。一方面数据仓库的数据内容、结构以及其他物理设计根据用户所返回的信息不断地调整和完善，以提高系统的效率和性能；另一方面，通过不断地理解用户的分析需求，向用户提供更准确、更有用的决策信息。

我们先来考察我国的证券市场。近几年证券业和计算机业飞速发展，在我国已形成了较为完善的证券交易系统，在借鉴国外先进技术的基础上，已经形成了具有本国特色的较为完整的证券业解决方案。证券交易系统是以证券交易所为中心，以电脑网络技术为基础，构成一个以本地网络为核心、本地网络与远程通信相结合的集中式中央管理系统。

就目前的发展及使用状况而言，我国证券交易系统按使用对象分为交易所、券商、

股民(大户、散户)、庄家、基金管理公司、监管机构、证券交易资讯网站等几类。主要介绍如下：

交易所：上海证券交易所和深圳证券交易所利用现代科学技术已成功地建立了全国性的大型电子交易网络。这个由卫星光纤等通讯系统组成的包括分组交换网、拨号的等多种通讯方式并存的多层次立体交易网络，将全国几千个证券营业部连成一体，通过电话委托、触摸屏、自动终端等各种委托手段组成了一个高效率的交易网络。

券商：根据国内证券交易的规则，券商的证券营业部为面向股民的交易委托机构，股民对股票的交易操作必须通过证券营业部进行，因此在整个证券交易系统中券商系统处于一个非常重要的地位。

股民：在我国，股民必须通过券商营业部方能进行股票的买卖操作，股民通过与券商营业部交互完成股票的买卖。

庄家：在我国的证券市场上，对股票价格涨跌起着重要作用的便是庄家了。庄家一般是指拥有大量资金，以图通过控制某一只或某几只股票价格的涨跌来获取利润的集团，它在运作股票的时候往往有各种特殊业务需求。

基金机构：投资基金是证券市场中一种非常重要的融资工具，掌管投资基金的基金机构需要有适合自身特点的交易系统。

监管机构：1992年10月，中国证券监督管理委员会(简称证监会)宣告成立，标志着中国证券市场统一监管体制开始形成。监管机构可对发行、市场、机构、上市公司、基金实行监管功能。

证券交易资讯网站：互联网络技术发展，使证券市场信息传递有了更新的传播媒介，并正在改变证券市场的发行模式、交易模式、清算模式，这将会形成网络化的自动化交易方式。

以上我们对证券交易系统整体进行了简单的分析，现在我们对股票交易系统作进一步的分析。股票交易系统主要是针对对广大股民而设计的。对于股民而言，主要需要提供三方面的服务：一是行情信息服务，二是股票分析服务，三是交易委托服务。行情信息服务主要是从前端接受股票行情数据，经过转换后存入本系统中。本系统重点介绍股票分析服务。股票分析服务提供股民股市的动态行情，个股走势图、K线图、分析指标图等。除了这些基本的信息外，要能够根据已有的数据和知识，分析股票价格变化，预测股市未来的走势。

以我国证券交易为例，为了制定合理的买卖策略，股民所要进行的分析主要有：

分析上海、深圳两地大盘的变化趋势；

分析个股价格未来趋势走向；

分析个股的最低和最高价位；

分析个股的最佳买点和卖点。

要进行以上的分析，所需数据应包括：

股票交易历史数据；

股票实时行情数据；

各支股票相关资料。

股票交易是证券交易系统的一个分支，我们建立基于数据仓库的股票交易系统是以股票交易数据为基本数据，是以股民为核心对象，根据每日的交易行情及历史交易数据，以分析研究股票价格变化和趋势为最终实现目的。

4.3 主题的选择

主题是一个在较高层次上的数据综合、归类的标准，是数据仓库的组织核心，每个主题对应于一个客观的分析领域。面向主题的数据组织方式，即是分析对象数据的一个完整的、一致性描述。

根据以上对数据的分析，考虑到股民的分析要求，我们确定了股票交易数据仓库的两个基本主题：股票和股民。股票是股民最为关心的问题，对股票进行预测，提供各种分析需求也是以此为基础；股民是证券交易中为数最多的群体，他们不仅要了解股市信息，同时还希望了解自己在股市中的每笔交易，从中吸取教训，积累经验，以便于今后更科学地投资。因此，股票交易数据仓库选取这两个主题，以此逐步对数据进行展开描述。用 E-R 图表示如图 4-1 所示。

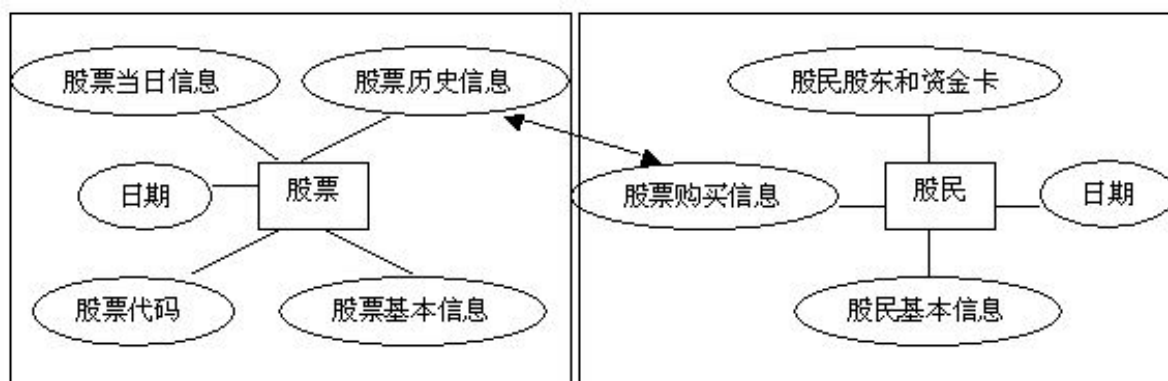


图 4-1 股票、股民之间的关系

这两个主题所应包含的内容列出如下：

股票：

股票基本信息：股票代码、股票名称、股票类型、公司资料等

股票历史信息：股票代码、日期代码、最高价位、最低价位、成交量以及各种指标信息等

股票当日信息：股票代码、日期代码、开盘价、收盘价、当日最高、当日最低、成交量及当日指标等

股民：

股民基本信息：身份证号、深圳股东卡号/上海股东卡号、深圳资金卡号/上海资金卡号、工作单位、家庭住址等

股民购买股票：身份证号、股票代码、日期代码、买入价格、购买量等

4.4 逻辑结构设计

在逻辑结构设计阶段，要进行的工作主要有：

分析主题域，确定当前要装载的主题；

确定粒度层次划分；

确定数据分割策略；

表示各类数据之间的关系。

4.4.1 分析主题域

在上节中，我们已确定了两个基本的主题域。数据仓库的设计方法是一个逐步求精

的过程,在进行设计时,一般是一次一个主题或一次若干个主题逐步完成的。应对确定的几个主题域进行分析,并选择首先要实施的主题域。

“股票”这个主题是此系统的最基本的对象,具有重要的意义和作用,我们可从“股票”主题实施。在“股票”主题中,主要是两个方面的内容:

第一,它包含了股票固有信息;

第二,应记录股票的交易数据。

主题在数据仓库中都是由一组关系表实现。主题的实现依然是基于关系数据库的。在具体实现中,一个主题可以划分为多个表,主题只是一个逻辑的概念。

4.4.2 粒度层次划分和数据分割

建立数据仓库过程中一个重要问题是如何提高系统的性能。因为数据仓库的数据量很大,分析处理时涉及的数据范围也较广,往往涉及大规模数据的查询。可以通过粒度划分和数据分割有效地解决上述问题。

粒度层次划分

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,粒度越小级别就越低;数据综合度越高,粒度越大级别就越高。传统的操作型系统中,对数据的处理和操作都是在详细数据级别上的,但是在数据仓库环境中主要是分析型处理,粒度的划分将直接影响数据仓库中数据量以及所适合的查询类型。一般需要将数据划分为:详细数据、轻度综合、高度综合三级或更多级粒度。

对于“股票”主题,考虑主题中各项内容的特点及建立其上的数据分析要求的特点,我们将从前端得来的股票交易数据经过转换作为基础数据,按照时间粒度进行综合,如图4-2所示。

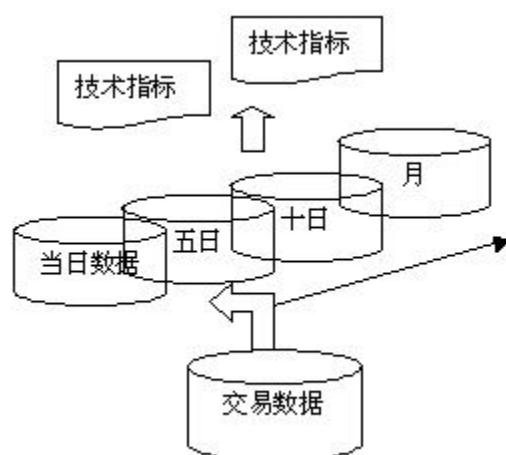


图 4-2 粒度层次划分

时间是一个有特殊意义的维度,它对决策中的趋势分析很重要。根据投资分析对交易数据不同需求,可设计以下几个交易表:当日交易数据(以日作时间段,包括最高价、最低价、开盘价、收盘价、成交量等)、五日交易数据、十日交易数据、月交易数据、年交易数据。在交易数据的基础上,还需要进一步计算预测所需要的各种技术指标,如能量潮、乖离率、威廉指标、筹码密集、分布成交等,并将结果保存在表中,用于股票交易数据分析,可给投资者有效的投资选股的依据。

以上按不同时间粒度汇总的股票交易数据可满足用户对行情的不同需求。时间粒度较细的交易数据提供信息较多,可满足短线投资需求;时间粒度较粗的交易数据则适用于长线投资,适合进行股票的趋势预测。

数据分割

数据分割是指把逻辑上是统一整体的数据分割成较小的、可以独立管理的物理单元(称为分片)进行存储,以便于重构、重组和恢复,以提高创建索引和顺序扫描的物理单元。数据的分割使数据仓库的开发人员和用户具有更大的灵活性。设计数据分割要选择适当的分割标准,一般要考虑以下几方面的因素:数据量、数据分析处理的实际情况、简单易行以及粒度划分策略等。

对于股票交易数据,可以将股票分为 A 股、B 股、Z(债券)股。再细分可分割为上证 A 股、上证 B 股、上证债券、深证 A 股、深证 B 股、深证债券。也可按板块分割:30 板块、浦东板块、汽车板块、基金板块、并购板块、庄股板块、大盘板块等。在实际的操作中,我们选择 A 股、B 股、Z 股分割,如表 4-1 所示,

表 4-1 股票交易数据的分割

2003 年	上证 A 股	上证 B 股	上证债券	深证 A 股	深证 B 股	深证债券
当日	分片 1	分片 2	分片 3	分片 4	分片 5	分片 6
五日	分片 7	分片 8	分片 9	分片 10	分片 11	分片 12
十日	分片 13	分片 14	分片 15	分片 16	分片 17	分片 18
月	分片 19	分片 20	分片 21	分片 22	分片 23	分片 24

在设计表的过程中,必须考虑到基于一个主题的所有表都含有一个称为公共码键的属性作为其主码的一部分,例如我们在“股票”主题中选取股票代码作为公共码键,此公共码键将各个表统一联系起来,从根本上体现出它们是同一个主题的。

4.4.3 数据之间关系的表示

数据仓库的目标数据库用于存放经过抽取、检验、加工和重新组织后的数据。设计数据仓库的数据模型时,常用的逻辑建模工具是以维数据建模来表示各类数据之间的关系。数据模式采用流行的星形模型和雪花型模型。星型模式是一种基本的方式,雪花型模型比星形模型增加了层次结构和类结构,体现了维的不同粒度的划分。

根据以上系统分析,我们采用图 4-3 所示的星型模型的逻辑结构表示“股票”主题。

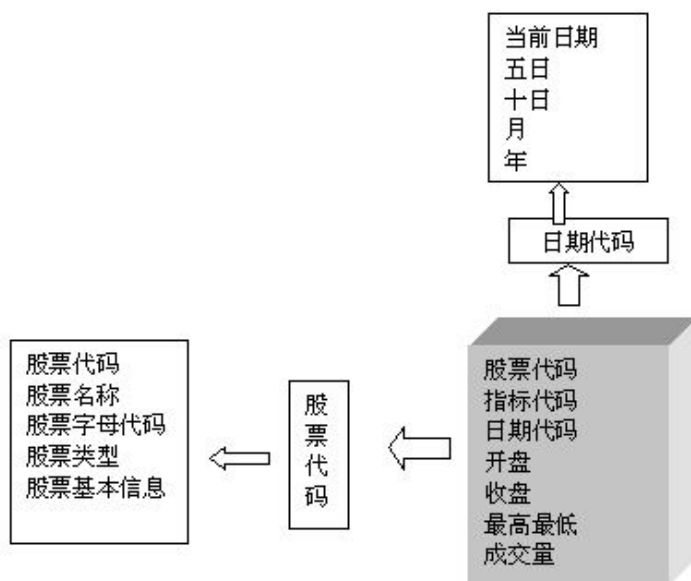


图 4-3 数据仓库逻辑结构图

按照星型模型和雪花模型等数据模式建立起来的数据表是物理上客观存在的,虽然它们之间可以通过主键或外部键建立相互关系,但是从满足分析的角度来讲,这些数据表缺乏根据多个角度动态去组织、浏览数据的能力。因此,我们必须在物理数据表的基础上创建一些用于即时查询的结构化视图。

视图看上去非常像数据表,但是视图只是一种定义,物理上并不存在,它只给用户提供了一个查看数据的窗口。通过视图,我们可以创建一些用于分析的操作,帮助用户发现不同侧面、不同层次上的有用信息,从而使用户更加直观地理解分析数据,进行决策支持。视图大大满足了用户对查询的要求。用户只要根据自己的查询要求定义好视图,就可以从视图中快速获取所需要的数据。

4.5 物理模型设计

进行数据仓库物理设计,要全面了解所选用的数据库管理系统,了解数据环境、数据的使用方式、数据规模以及响应的时间要求等,可掌握时间和空间效率进行平衡和优化的依据。

在物理模型设计中,应注意以下问题:

4.5.1 可适当入冗余

由于数据分析的处理数据范围是广泛的,通常要涉及不同表的多个属性,一些表的某些属性可能在许多地方都要用到,我们可以将这些属性复制到多个主题中,从而减少处理时被存储的表的个数。

在“股票”主题中有一个保存股票基本信息的关系表,有股票代码、股票名称、股票类型等。而其他表中只存股票代码。但在实际应用中,不是直接针对股票代码,很多情况下需要用到股票名称,必须采取表的连接或半连接操作。当涉及的数据量很大时,访问的效率很低。因此可以在其他表中,将股票的一些特定属性加入,即增加数据冗余,就可以省去这一步连接操作,减少访问的代价。引入冗余后,破坏了关系模式的规范化,同时需要维护各个表之间的一致性。但数据仓库几乎是不更新的,适当地引入冗余可提高系统的性能。

4.5.2 生成导出数据

在原始数据的基础上进行总结和计算,生成导出数据。进行预测的技术指标需在交易数据的基础上经过计算进一步得到,在应用中直接利用这些导出数据。它的另一个好处是在更高级别上建立了公用数据源。

4.6 数据仓库的实现^[20]

4.6.1 数据仓库的体系结构

关于数据仓库的体系结构,存在着不同的说法和建立。但大多包括了三部分在内。1)数据源:提供原始数据;2)后端加工:实施数据的后处理(接收、提取、汇总、交换、打包和储存等);3)前端服务:面向最终用户。

此系统是一个三层客户机/服务器结构。其中第一层为客户端,包括客户交互、格式化查询及结果等,以显示分析结果;第二层为分析服务器,存储数据仓库,集中和简化客户端和服务器的部分工作,降低系统数据的数据传输量;第三层为服务器端,通过Internet或股票机接收图文电视信号,接受股票信息,存储事务数据。其结构模型如

图 4-4 所示^[21]。

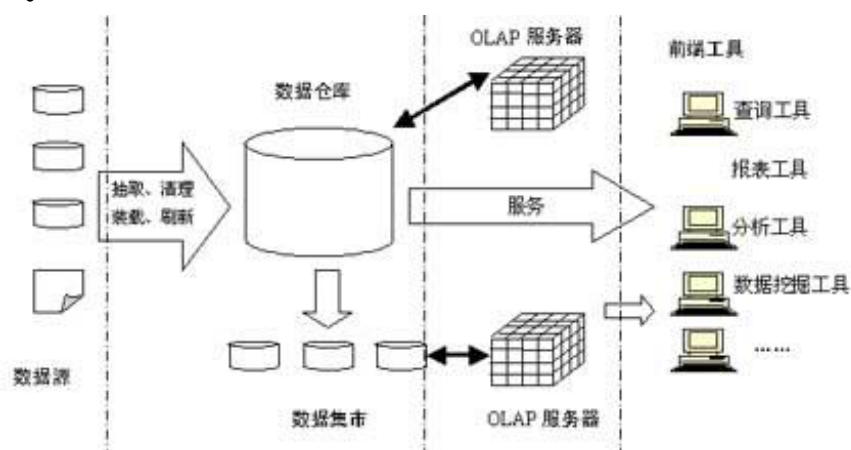


图 4-4 数据仓库体系结构

数据源：是数据仓库系统的基础，是整个系统的数据源泉。通常包括企业内部信息和外部信息。内部信息包括存放于 RDBMS 中的各种业务处理数据和各类文档数据。外部信息包括各类法律法规、市场信息和竞争对手的信息等等。

数据的存储与管理：是整个数据仓库系统的核心。数据仓库的真正关键是数据的存储和管理。数据仓库的组织管理方式决定了它有别于传统数据库，同时也决定了其对外部数据的表现形式。要决定采用什么产品和技术来建立数据仓库的核心，则需要从数据仓库的技术特点着手分析。针对现有各业务系统的数据，进行抽取、清理，并有效集成，按照主题进行组织。数据仓库按照数据的覆盖范围可以分为企业级数据仓库和部门级数据仓库（通常称为数据集市）。

OLAP 服务器：对分析需要的数据进行有效集成，按多维模型予以组织，以便进行多角度、多层次的分析，并发现趋势。其具体实现可以分为：ROLAP、MOLAP 和 HOLAP。ROLAP 基本数据和聚合数据均存放在 RDBMS 之中；MOLAP 基本数据和聚合数据均存放于多维数据库中；HOLAP 基本数据存放于 RDBMS 之中，聚合数据存放于多维数据库中。

前端工具：主要包括各种报表工具、查询工具、数据分析工具、数据挖掘工具以及各种基于数据仓库或数据集市的应用开发工具。其中数据分析工具主要针对 OLAP 服务器，报表工具、数据挖掘工具主要针对数据仓库。

4.6.2 数据获取

股票交易数据仓库中的数据来源于前端交易环境中的交易数据。可通过 Internet 接收股票信息。

股票交易数据中包含的股票基本信息有：股票数字代码 char(10)，股票字母代码 char(10)，股票名称 char(10)，股票类型 char(1)，流通总股 real (实型)等。

其中：

股票数字代码：记录该支股票的数字名称，当采用数字检索股票名称时，索引该属性列进行查询；

字母代码：记录该支股票的字母名称，当采用字母检索股票名称时，索引该属性列进行查询；

股票名称：记录该支股票的中文名称；

股票类型：记录该支股票的类型，其值可以为 A (A 股)、B (B 股)、Z(债券)；

流通总股：记录该支股票发行的上市流通的股票股数，它的意义在于日后计算换手率的需求。

动态的股票交易数据主要是指从上海、深圳传来的当日股票交易数据，它提供对股票进行数据分析的原始数据。通过数据转换后存入数据仓库中形成存储事物数据。具体的内容和数据类型有：股票数字代码 CHAR (10)，股票名称 CHAR (10)，开盘价 (REAL)，收盘价 (REAL)，最高价 (REAL)，最低价 (REAL)，成交量 (REAL)，停盘信息 CHAR (1) 以及当日技术指标 (如 OBV 指数、WMS 指数、MAV 指数、涨幅、筹码密集等)，如表 4-2 所示：

表 4-2 当日股票交易数据

属性名称	数字代码	股票名称	字母代码	开盘	收盘
数据类型	CHAR(10)	CHAR (10)	CHAR (10)	REAL	REAL
属性名称	最高	最低	成交量	停盘	
数据类型	REAL	REAL	REAL	CHAR(1)	
属性名称	OBV 指数	WMS 指数	MAV 指数	涨幅	筹码密集
数据类型	REAL	REAL	REAL	INT	INT

其中：

开盘价，收盘价，最高价，最低价，成交量：分别记录该股票在当日的开盘价位，收盘价位，最高价位，最低价位；

停盘信息：记录该股票是否被停盘的标志位，它的值是 0 或 1，当值为 0 时说明未停盘，相反则停盘，如果停盘则该股的当日数据为空。

4.6.3 数据装入

不同级别的综合数据的形成是按照不同汇总粒度计算获得，并且加上相应的时间戳。其过程如图 4-5 所示。

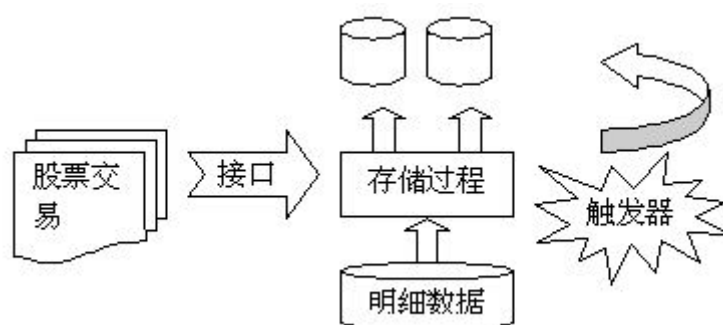


图 4-5 数据装入示意图

按时间粒度汇总对股票交易数据进行综合，生成以下几个表：当日交易数据、各支股票日记、五日交易数据、十日交易数据、月交易数据、年交易数据。

当日交易数据可由上节过程得到，在交易日则每日进行更新。除了基本股票信息外，同时计算当日的各项技术指标。这些技术指标是投资者有效投资选股的依据。

每天当交易结束后，由当日交易数据可得到各支股票的日记表，即将当日交易数据中的所有股票根据股票代码加入到各支股票的日记表中，股票日记表是记录该股票的历史信息的详细表，记录交易的最详细的信息，为细节级数据。股票的日记表中包含的信

息有：时间、股票数字代码 CHAR (10)，股票名称 CHAR (10)，开盘价 (REAL)，收盘价 (REAL)，最高价 (REAL)，最低价 (REAL)，成交量 (REAL)，停盘信息 CHAR (1) 及各项技术指标 (如 OBV，MAV，WMS，乖离度，涨幅等)。如表 4-3 所示。

表 4-3 股票的日记表

属性名称	时间	股票名称	数字代码	字母代码	股票类型
数据类型	datetime	CHAR(10)	CHAR(10)	CHAR(10)	CHAR (1)
属性名称	开盘	收盘	最高	最低	成交量
数据类型	REAL	REAL	REAL	REAL	REAL
属性名称	OBV 指数	WMS 指数	MAV 指数	成交总量	分布成交
数据类型	REAL	REAL	REAL	REAL	REAL
属性名称	乖离度	五日均线	分布界位	停盘	涨幅
数据类型	INT	REAL	INT	CHAR (1)	INT
属性名称	换手率				
数据类型	REAL				

五日、十日、月、年交易数据表为综合级数据，分别对五日、十日、月、年时间段的数据进行综合汇总，以满足对数据和预测的不同需求。每一张表中包含的信息有：时间、股票数字代码 CHAR (10)，股票名称 CHAR (10)，时间段内历史最高和历史最低，成交最高，成交总量及各项指标 (如时间段内 OBV 最高、MAV 最高、乖离度、WMS 最高等)，如表 4-4 所示。

表 4-4 五日、十日、月、年交易数据表

属性名称	数字代码	字母代码	股票名称	时间	历史最高
数据类型	CHAR(10)	CHAR (10)	CHAR(10)	datetime	REAL
属性名称	历史最低	成交最高	成交总量	OBV 最高	MAV 最高
数据类型	REAL	REAL	REAL	REAL	REAL
属性名称	乖离度	WMS 最高			
数据类型	INT	REAL			

4.6.4 技术指标计算^[21-22]

对未来股市运行趋势进行预测的基础是股市中的大量历史数据，主要是每日的开盘价、收盘价、最高价、最低价和成交量等。但是单从这些原始数据的升降很难看出股市的运行规律，因此必须对这些原始数据进行处理和加工，计算出那些可以反映股市特征和规律的指标数据。

在系统中我们用到的技术指标有 OBV，MAV，WMS，乖离度，涨幅，筹码密集统计，在这些指标中除了乖离度，涨幅和筹码统计外，其他的技术指标都是以图形的方式直观的展示在用户的面前，这也正是未来软件的一个大趋势——数字图形化、直观化的

体现。

设 X_n 为时间序列变量在第 n 个交易日的交易数据, ZG 为最高价、 ZD 为最低价、 KP 为开盘价、 SP 为收盘价、 CJL 为成交量、 $LTZG$ 为流通总股。下面给出一些重要技术指标的计算方法。

OBV 指标:

能量潮指标, 为累计成交量, 如果今日股价上涨则成交量为正值, 若股价下跌则成交量为负值, 若平盘则为 0。公式如下:

$$OBV = \frac{SP - ZD - (ZG - SP)}{ZG - ZD} \times CJL$$

OBV 指标主要反映了成交量的变化, 代表了人气的指标。它比较适用于短线的进入, 与股价波动的“M 头”有密切的关系。当股价的波动区有可能形成“M 头”时, OBV 线会发挥决定性的警示作用。当股价的波动还没有形成两个高峰时, 当第二个高峰与 OBV 线同时上升时, 则表示有大量的成交量出现, 那么第二个高峰就不是高峰, 可能还有最高的价位出现。反之, 如果此时的 OBV 线无力上扬, 成交量大量减少, 那么就可以确定“M 头”的形成, 投资者也就可以抓住时机卖出。

MAV 指标:

MAV 指标又称为成交量的移动平均, 它的作用与 OBV 相似, 也是体现成交量及人气的变化的指标, 它的计算公式如下:

$$MAV = \frac{1}{K} \sum_{i=1}^k CJL_i \quad (\text{其中 } k \text{ 为移动平均的天数, 一般取十天})$$

此指标可求得最近 K 天的成交量的算术平均值, 这样可以有效的避免成交量的短期的异常对全局的影响, 并且可以更精确的显示成交量的变化。

WMS 指标:

WMS 指数的计算公式如下:

$$WMS = \frac{\max_{i=n-k-1}^n (ZG_i) - SP_n}{\max_{i=n-k-1}^n (ZG_i) - \min_{i=n-k-1}^n (ZD_i)} \quad (i, n \text{ 表示第 } i, n \text{ 交易日})$$

WMS 指标是反映当日的价位变化程度的一个指标, 它能宏观的反映该支股票在一段时间内的价位稳定程度, 如果 WMS 指数相对一段时间比较平缓, 那么说明该支股票的相对比较集中在当前价位上, 股票的波动相对比较小, 说明这支股票的投资风险相对比较小。相反, 则说明股价的浮动程度很高, 投资的风险也相对比较大。

乖离度 BIAS:

乖离度的计算公式如下:

$$BIAS = \frac{SP - \frac{KP + SP}{2}}{\frac{KP - SP}{2}} \times 100\%$$

乖离率是今日收盘价与今日均价间的差别, 用百分比表示。

涨幅指标:

涨幅指标的计算公式如下:

$$\text{涨幅} = \frac{SP_n - SP_{n-1}}{SP_{n-1}} \times 100\% \quad (n, n-1 \text{ 表示第 } n, n-1 \text{ 交易日})$$

涨幅指标为今日收盘价与昨日收盘价变化的比率，用百分比表示。

技术指标可以通过设定的存储过程和触发器来计算。存储过程是服务器上的一组预先编好并且经过编译的 SQL 语句组，使用存储过程可以提高 SQL 语句的运行性能，提高其执行效率。触发器实际上是一种特殊的存储过程。触发器的主要功能是监视用户对数据的修改。触发器可以在数据表中插入数据、修改数据或删除数据时进行检查，以保证数据完整性和一致性。触发器会根据条件自动执行。

4.6.5 功能实现

数据仓库系统主要提供以下几个一些基本功能，如图 4-6 所示。

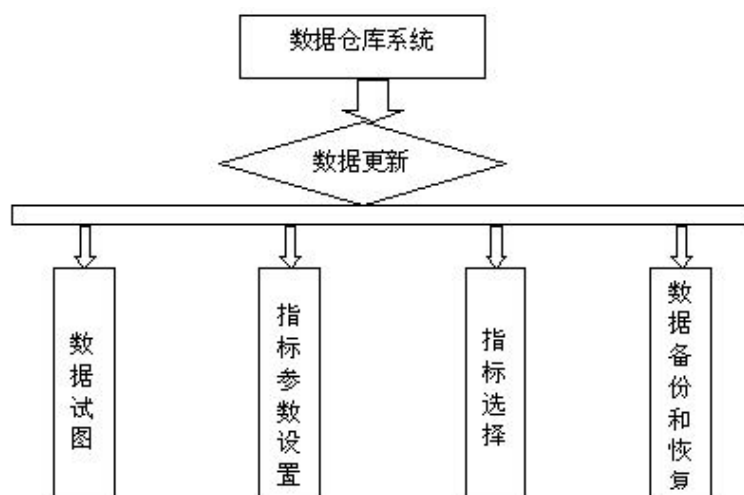


图 4-6 主要功能

(1)数据更新

在程序刚开始运行时，系统提示当前的时间以及是否要更新数据的信息，如果用户选择更新数据，那么系统将会对当日的数据进行加载，并通过程序对原始的数据进行初次的加工，存入 SQL SERVER 表中，如果用户选择不更新数据，那么系统将会跳过对当日数据的检测，而直接对历史数据进行加载。

(2)指标选择

在这里我们提供了 OBV，MAV，WMS、乖离度的选择，即在主窗体上可以同时画出日 K 线、五日均线、日均价线、OBV、成交量，或同时画出日 K 线、五日均线、日均价线、MAV、成交量或同时画出日 K 线、五日均线、日均价线、WMS、成交量或同时画出日 K 线、五日均线、日均价线、OBV，MAV，WMS、成交量；如果用户想将所有的指标都直观的显示在屏幕上，只要用户选择所有的指标，那么我们就可以在屏幕上显示所有的指标的图像，更加直观的显示了股票的动态，对股民的投资有实际意义。

(3)指标参数设置

该功能主要是对各项技术指标的参数进行动态的修改，以满足不同习惯和指标组合的要求。其中我们可以对均线的天数进行修改、对 MAV 的天数进行修改、对筹码统计的百分比进行修改。这样能够对各类指标的参数进行动态的修改的优点在于：它既避免了初学者对着满屏的五彩线陷入窘态的发生，同时也提高了程序的运行效率。

(4)数据视图

在这个模块里我们提供了三个视图：当日数据指标视图、各支股票的历史数据的视图、全体股票的视图。当日技术指标的视图是显示今天的各支股票的指标值，其中有名称、数字代码、字母代码、类型、OBV、MAV、WMS、筹码密集，日后我们选股的时候，

如果条件中有与当日数据指标有关的选项我们就可以直接对这张视图进行查询以得到有效的信息；各支股票的历史数据的视图是显示一支股票所有的历史信息视图，因为它是每支股票的详细数据的集和，所以它具有股票详细列表的所有的属性列，我们可以通过对各类线图进行验证，证实各类线图的正确性；全体股票的视图是一张对整体股票的总体的描述的视图，它所带有的属性列有数字代码、字母代码、名称、股票类型、历史最高、历史最低、成交总量、成交最高、OBV 最高、MAV 最高、乖离度、WMS 最高、流通总股。

(5)数据备份和恢复

在这个模块中，我们提供了数据的恢复及备份功能，只要选择(在股票列表中被聚焦)了需要备份(恢复)的股票名称，在选择对这支股票要进行的操作，则可将相应的数据存入外置存储介质中或从外置存储介质中恢复到系统中。

4.7 数据仓库与系统间的连接

系统间的接口见图 4-7。

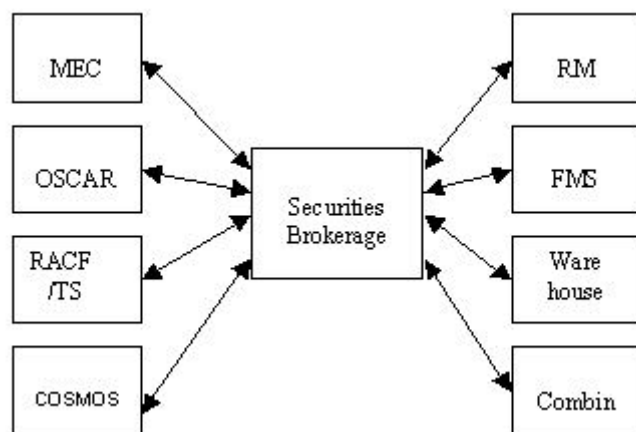


图 4-7 系统接口图

由于数据仓库是我们研究的重点，所以主要介绍数据仓库与系统间的连接。系统通过 COBOL 编写的 online 程序调用，信息的交互则通过 IBM 公司的 MLI 通信信息机制，MLI 的层次结构如下：

Request MLI

{Message Identifier group

Routing Group

Common Request group}

Function specific data group

Reply MLI

{Message Identifier group

Common Reply Group

Error Message Group }

Function specific data group

4.8 本章小结

采用数据仓库技术建立面向分析决策的交易数据平台,不仅大大提高了查询响应速度,而且还提高了随机性综合交易数据信息提取能力,增强了对交易分析决策的支持。但是,数据仓库的建立是一长期、复杂的工作,并且要进一步提高决策能力,还需引入数据挖掘技术,从股票交易的数据中得到股票交易的规则或规律,发现隐藏在数据后面的有用信息,预测股票交易数据的未来走势,提供未来情况的分析结果,这仍是将来一个需要探讨的话题。

第五章 股票交易系统的实现

5.1 开发环境

操作系统：采用 IBM 公司的 MVS 系统，该系统具有多线程、多虚拟存储技术，为大型机系统提供了高性能、高可靠性、高利用率以及兼容性的运行环境。

系统平台：IBM 公司的大型机 OS/390。目前通过专用网络连接，在线运行上海花旗软件服务有限公司位于新加坡总部的 OS/390。

开发工具：COBOL(Common Business-Oriented Language)语言、JCL (Job Control Language) 负责程序的编译连接、CICS (Customer Information Control System) 用于程序测试跟踪、DB2 (数据库管理)。

5.2 股票交易系统改进

目前，运行于该商业银行的股票交易系统，技术上还存在以下不足之处：

(1) 程序由 GOOGLen (Gen Client/Server Development Tutorial) 语言编写，虽然其使用上较为方便，操作界面见图 5-1，但在使用效率上，却存在很大的不足，因为在运行时，系统首先需要将 GOOGLen 语言转化为 COBOL 语言。另外，使用 GOOGLen 语言的另一个缺陷是，当程序出现错误时，跟踪测试上就不如 COBOL 语言方便（可方便地添加 DISPLAY 命令或 CICS 命令，详细介绍见下）。

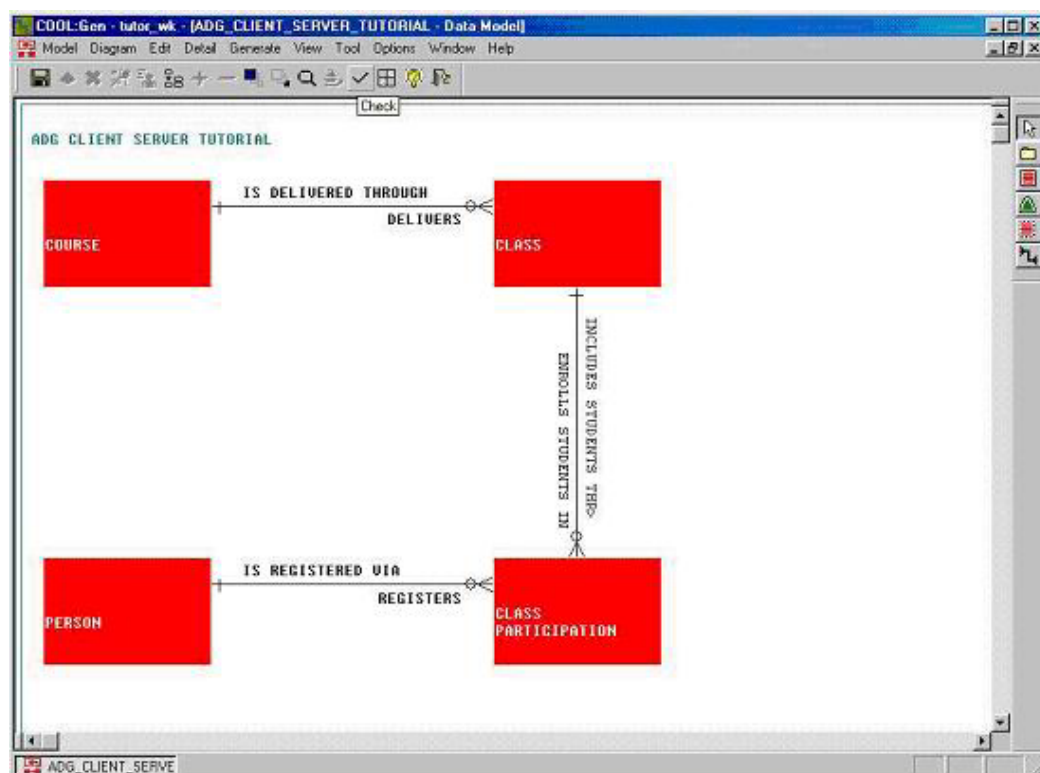


图 5-1 GOOGLen 操作界面

(2) 实际使用中, 客户在查看订单时, 系统并不支持模糊匹配。例如:

在订单号输入项, 客户输入的为: 980 (完整的订单号应为 98044), 则系统作出的反应是搜索订单号为 980 的订单, 导致的结果是出现“找不到该记录”。

在客户姓名输入项, 客户输入的为: JAM (完整的客户姓名应为 JAMES), 系统同样去搜索姓名为 JAM 的订单。

这样的情况都给客户操作上造成了很大的不便。

针对这些缺陷, 在新一代所采用的解决方法为:

(1) 程序统一使用 COBOL 语言编写, 从而有效地提高执行效率, 也便于跟踪测试。

(2) 在程序中增添模糊匹配功能。对数字型变量, 采用“不小于”该变量值的方法; 而对于字符型变量, 则采用先通过格式处理, 再使用增加通配符的方法解决。

```
INSPECT WP-VARIABLE REPLACING ALL SPACE BY '%'
```

```
.....
```

```
C_NAME LIKE : WP-VARIABLE
```

经过上述改进, 同老系统相比较, 在效率、使用灵活性上都有很大的提高。

5.3 项目开发

就股票交易系统而言, 股票走势预测功能处于理论阶段, 有待进一步深入研究。下面着重介绍 Order(订单) 项目的全过程, 整个项目大致分为四个逻辑块: View(查看订单)、Add(添加订单)、Update(修改订单)、Auth(鉴定订单)。

各逻辑块分别有多个子程序组成, 用 COBOL 语言进行编写。

程序大都遵循一定的书写标准, 从高到底依此分为 IDENTIFICATION DIVISION (标识区)、ENVIRONMENT DIVISION (环境区)、DATA DIVISION (数据区)、PROCEDURE DIVISION (过程区)。各区意义如下:

标识区: 主要为程序员以外的人员所查看使用的。一般标明程序名、程序员名和程序编写及修改日期、修改原因等。

环境区: 用于和外部文件进行连接。

数据区: 程序中用到的变量都在该区进行定义。

过程区: 程序的可执行语句书写区。按照单位大小又可分为: Section (节)、Paragraph (段)、Sentence (句)、Statement (子句)。

程序编写完后, 要进行编译连接。该环境下的编译有别于一般的程序编译连接, 它是通过一种语言 JCL (作业控制语言) 利用 JCL 组装待编译程序的模块、组织资源、调用系统中的编译连接程序对程序的各个模块进行编译和连接。一个 JCL 任务称为一个作业。该作业被提交给系统后会在系统后台运行, 运行完毕后会给出回馈信息。程序编译连接后, 会根据 JCL 的指向将编译好的程序模块放在一个 Loadmode (一块特殊的存储区) 中。

5.4 项目测试

程序编译连接通过后, 就需要对其进行跟踪测试。目前, 在股票交易系统中, 因为交易大都属于联机在线处理, 所以测试占据很大的比重, 股票交易项目整个开发过程中有一半以上的时间都用于测试。测试一般分为三个阶段: RND 测试 (自测), 一般由开发人员自己测试; SIT 测试 (系统集成测试), 一般由专业测试人员测试; UAT 测试, 这是由客户直接参与的测试。以下所叙述的都属于 RND 测试。测试方法大致分为两类:

5.4.1 命令行测试

该测试也称为 CICS 测试。首先，在 OSPF 程序界面下，执行 A DEUCICS 命令，进入 CICS 配置环境，定义并安装事务 ID 号及相连的程序名。然后，通过执行该事物 ID 号及交易的数量、价格等，就能出现如图 5-2 所示界面：

```
TRANSACTION: SBMQ PROGRAM: MFOCM000
STATUS: PROGRAM INITIATION
      COMMAREA      = 'SBLS101099999416'
      EIBTIME       = 164028
      EIBDATE       = 0105179
      EIBTRNID      = 'SBMQ'
      EIBTASKN      = 73186
      EIBTRMID      = '.....'

      EIBCPOSN      = 0
      EIBCALEN      = 5307
      EIBAID        = X'00'
      EIBFN         = X'0E02' LINK
      EIBRCODE      = X'0000000000000'
```

图 5-2 命令行测试

图中 COMMAREA 项所含的数据就包括交易的数量项等，允许在此处进行数值的修改。

该种测试类似与白盒测试，通过在程序中添加 DISPLAY 命令或 CICS 命令，能够清楚地获取变量的当前值，从而为排错提供便利。两种命令格式如下：

```
DISPLAY    变量名；
EXEC CICS WRITEQ TS
            QUEUE (程序名)
            FROM (变量名)
END-EXEC
```

该测试成功的标准是整个流程按照程序的预定逻辑执行，从头至尾执行一遍；而当有错误出现时，会出现错误提示，并调用通用错误处理程序，查找错误代码表，详细列出错误信息。

缺陷是遇到程序中存在调用多个子程序时，会逐个进入各个子程序，从而需要花费大量时间。并且只能逐个测试程序逻辑，并不能对整个交易流程进行跟踪测试，这就需要由在线测试来实现。

5.4.2 在线测试

该测试也称为 OSCAR 测试。以图形化的界面显示，允许客户直接输入数据。

(1) 添加订单。其界面如图 5-3。

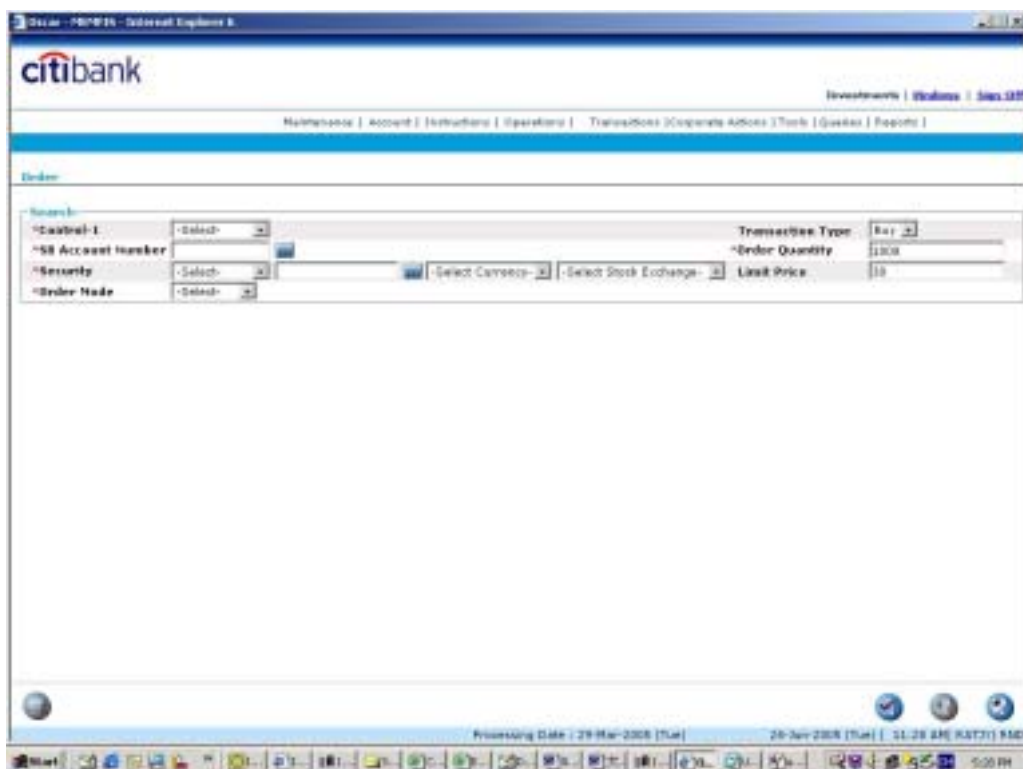


图 5-3 添加订单

其中，图中各项意义如下：

Control-1、SB Account Number: 银行分枝代码及客户资金帐号；

Transaction Type: 交易类型（买或卖）；

Security: 在此客户可以选择所交易的股票；

Order Quantity: 交易的数额；

Limit Price: 客户在此进行价格的设置；

Order Mode: 交易方式（限价交易或市价交易）。

输入各项数据后，通过点击右下方的“执行”按钮，该命令便会自动调用 Add(添加订单)模块，向系统提交请求，即请求增加一个订单。系统会先判断输入数据项是否正确，然后再按照程序流程执行。如果程序逻辑正确，将显现如下信息：

Add order 98044(系统产生的订单号) successfully.

反之，将会根据具体情况，弹出一个窗口，提示错误信息，例如如图 5-4 下例：



图 5-4 错误信息框

(2) 查看订单。其界面如图 5-5。

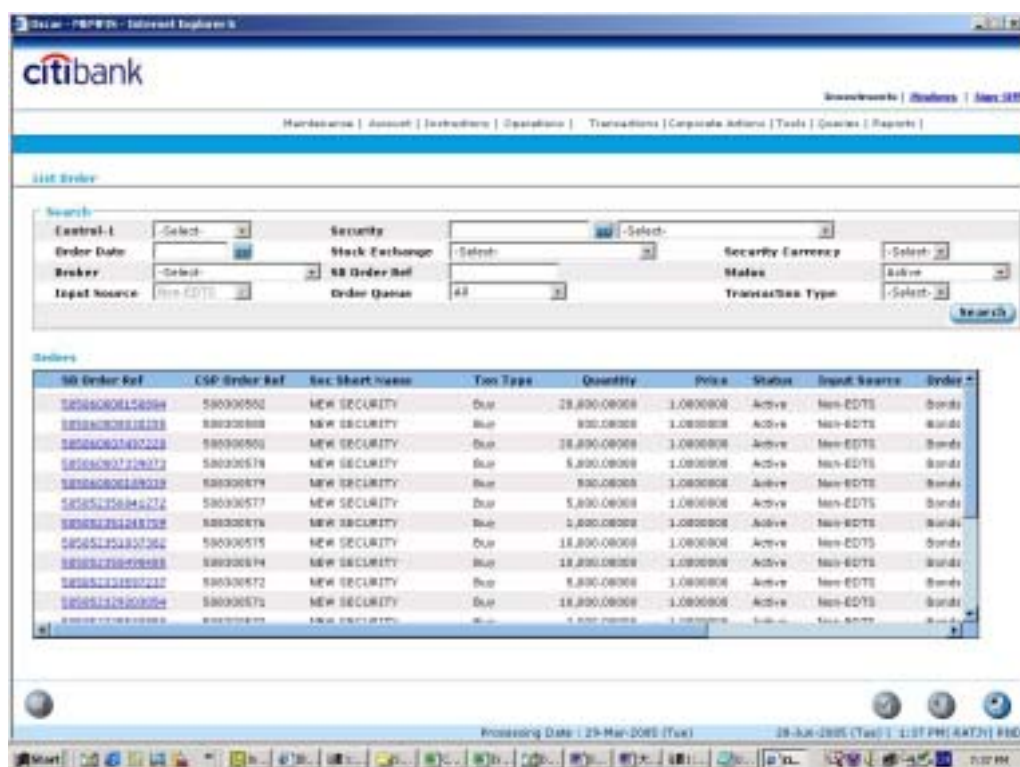


图 5-5 查看订单

通过该界面，客户可以根据 Control-1（银行分支）、SB Order ref(订单号)、Security（股票代码）、Status（订单状态）等选项，来调用 View(查看订单)模块，看是否能达到预期结果（即窗口下方将列出所搜索到的订单），并且通过点击列表框中的订单进行详细查看，见图 5-6。

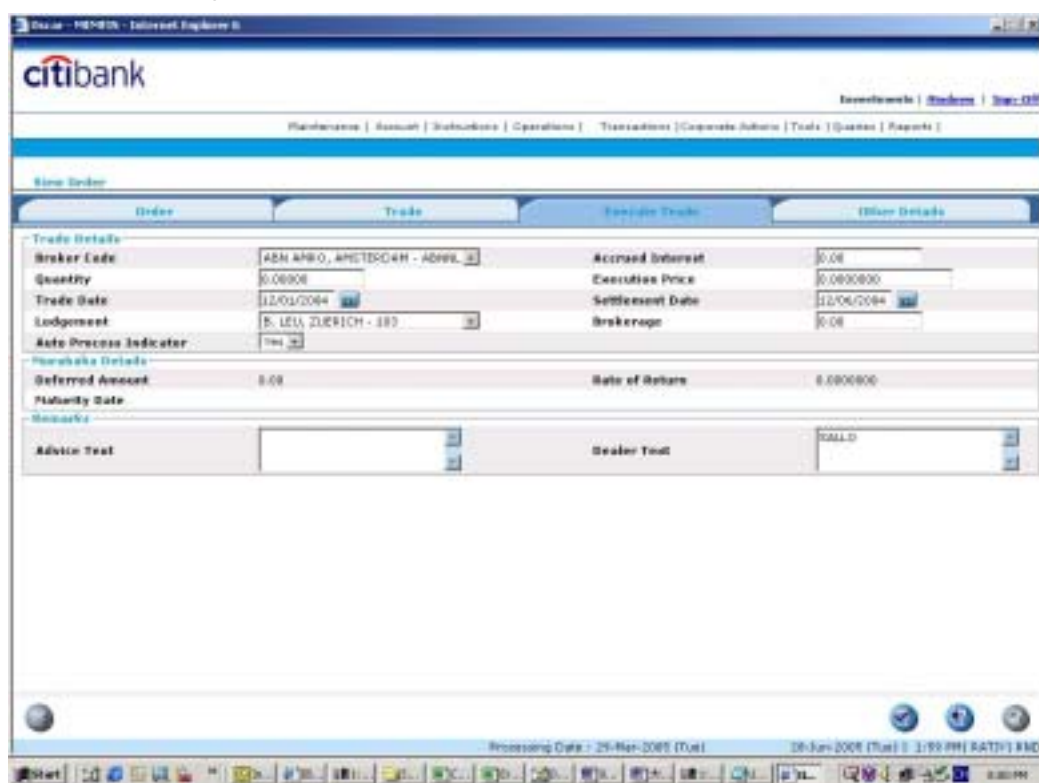


图 5-6 订单详细查看

在该窗口，客户可以修改数据项，从而调用 Update(修改订单)模块，测试其相关程序的正确性。

如果客户认为该订单正确无误，就可点击“ Authorise ”按钮(见图 5-7)，调用 Auth(签定订单)模块，测试该订单是否有误。如果无错误发生，则整个 Order (订单)过程完成，系统显示信息：

Send order 98044, please waiting!

The screenshot shows the Citibank 'New Order' window. It contains three main sections: Account Details, Order Details, and Settlement Details. The Account Details section includes fields for Control ID, Cost Center, Security Code, Security Name, Security Name, Stock Exchange, and Ledger Code. The Order Details section includes fields for Broker, Order Type, Duration Expiry Date, Cash Account, Order Date, Duration, Settlement Currency, Input Source, Order Time, Filing Date, and Cutoff Date. The Settlement Details section includes fields for Limit Price, Gross Amount, Coupon Date, Maximum Commission, Settlement Price, Order Quantity, Net Amount, Maturity Date, Commission %, Deferred Amount, Exchange Rate, Coupon % / Frequency, Accrued Interest, Commission Amount, and Rate of Return.

Account Details		Order Details		Settlement Details	
Control ID	389 CNA	Broker	Market	Limit Price	1.000000
Cost Center	461	Order Type	Market	Gross Amount	50.00
Security Code	100000000000	Duration Expiry Date	14-Apr-2005	Coupon Date	
Security Name	NEW SECURITY	Order Date	24-Mar-2005	Maximum Commission	0.00
Security Name	150	Duration	DAY	Settlement Price	1.000000
Stock Exchange	NYSE NYSE	Cash Account	000000000000	Order Quantity	50.000000
Ledger Code	000	Settlement Currency	USD	Net Amount	0.00
		Input Source	Non-BOTS	Maturity Date	
		Order Time	00:00:00	Commission %	0.000000
		Filing Date		Deferred Amount	0.00
		Cutoff Date	24-Mar-2005	Exchange Rate	1.000000
				Coupon % / Frequency	0.000000 / 00
				Accrued Interest	0.00
				Commission Amount	0.00
				Rate of Return	0.000000

图 5-7 签定订单

同样，如果错误出现，也相应弹出错误信息窗口。

OSCAR 测试优点在于直观，能够快速发现问题；但该测试往往只能给出一个大概范围，所以需要借助命令行测试，才能最终发现问题所在。

5.5 本章小结

以实际工作为例，介绍了开发背景及对原系统的改进，然后以股票交易系统 Order (订单) 项目为例，介绍了其开发及测试方法。因为该项目属于 OnLine (在线) 模式，所以其流程前后连贯性较强，这就对程序测试提出了较高的要求。

第六章 结论

股票市场是经济的“晴雨表”，其作用不仅被政府所重视，而且更受投资大众的普遍关注。近年来随着证券市场的飞速发展，不仅要利用高新科学技术，以计算机、网络为辅助工具，实现高效的全面的金融信息系统，加速信息传递，从而为客户提供便捷服务。因此，迫切需要一套完善的适用于证券业的管理信息系统，否则很难适应证券行业的发展速度。

本文首先研究将数据仓库技术应用到股票交易系统中，在进行数据交易之前，将股票交易数据集成到统一的目标数据库中，并把数据转换成面向主题的格式，从数据源中抽取、集成所需的数据，并按不同粒度、分割方式对数据进行综合、统计，从而形成数据仓库。本文探讨数据仓库技术目的为股票交易系统实现股票走势预测提供理论依据。然后通过对原交易系统的分析，分析不足，进行技术上改进，并以订单模块为例，详细叙述了其开发及测试流程。

在本文研究的基础上，未来的工作包括：

- (1) 如何将理论转变为实际，使得股票走势预测功能真正运用于该系统；
- (2) 进一步完善股票交易系统，在提高其效率的同时，也需要保证系统的稳定性。

致 谢

首先真诚地感谢培养我的学院老师们，正是你们让我学习到了更多的知识，同时学到了很多做人的道理。

感谢吴介一老师、韩洁老师、方宁生老师、张家浩老师、张颀兵老师、宋之毅老师，谢谢他们以不同的方式给予我学习上的指导及帮助。

感谢所有帮助过我的各位老师、同学。

感谢我的父母、家人对我学业的关心、理解和支持。

感谢论文评审委员会的老师们在百忙之中对我论文的悉心指正。

感谢所有关心和爱护我的人们！

参考文献

- 1 刘立强. 网上股票交易[N]. 今日电子, 1998-3-12(3)
- 2 赵波. 新加坡推出语音起动股票交易系统[N]. 东南亚南亚信息, 2000-7-18(8)
- 3 蔡莲红. 语音技术的拓展与展望[J]. 计算机世界, 2001(6): 11-12
- 4 印勇、曹长修、林景栋等. 数据仓库和数据采掘研究综述[J]. 重庆大学学报, 2000(2): 116-118
- 5 刘东波. 数据仓库技术的现状与未来[J]. 微型机与应用, 2000(7): 32-34
- 6 王珊等. 数据仓库技术与联机分析处理[M]. 北京: 科学出版社, 1999. 65-69
- 7 胡侃、夏绍玮. 基于大型数据仓库的数据采掘[J]. 软件学报, 1998(1): 53-66
- 8 Securities Brokerage System[G]. Citigroup
- 9 Building of Banking and Financial Knowledge[G]. Citigroup
- 10 Rainbow Architecture[G]. Citigroup
- 11 Securities Brokerage Orders and Trades[G]. Citigroup
- 12 吴岳强. 数据仓库技术及应用分析[J]. 江苏通信技术, 2001(1): 25-27
- 13 唐宏、聂能、熊思民等. 数据仓库实现技术[J]. 数字通信, 2000(8): 16-18
- 14 吴宏昊、陈奇. 关于数据仓库若干问题的讨论[J]. 计算机科学, 2001(6): 20-23
- 15 刘莉. 数据仓库若干问题的讨论[J]. 西南民族学院学报, 2000(2): 202-207
- 16 张寂、刘春红、敬卿. 数据仓库的建设与数据挖掘技术浅析[J]. 现代信息技术, 2000(9): 24-27
- 17 商琳、骆斌. 基于数据仓库的数据挖掘系统的结构框架[J]. 计算机应用研究, 2000(5): 13-15
- 18 张晓京、刘明吉、刘洪杰等. 证券投资交易系统技术实现中若干问题的研究[J]. 计算机工程与应用, 2001(5): 48-50
- 19 刘明吉、张晓京等. 证券交易仿真系统的设计与实现[J]. 计算机工程与应用, 1999(12): 31-33
- 20 师智斌、陈立潮、张素华. 数据仓库技术在股票交易数据分析中的应用[J]. 华北工学院学报, 2002(6): 8-10
- 21 谷岩. 基于三层结构的计算机股票交易网络系统的研究与设计[J]. 计算机应用与软件, 2004(7): 22-23
- 22 朱元编著. 证券投资学原理[M]. 北京: 立信会计图书出版社, 1993. 296-323
- 23 张龄松、陶崇恩. 股票技术分析利器[M]. 北京: 中国大百科全书出版社, 1994. 76-99
- 24 唐志宏. 股票交易中的技术分析[N]. 经济问题探索, 1994-8-3(6)