长春工业大学

硕士学位论文

OLAP技术研究及其在移动通信运营中的应用

姓名：李哲琦

申请学位级别：硕士

专业：计算机应用技术

指导教师：逄焕利

20070301

摘要

目前，数据仓库技术正处于快速发展时期，基于数据仓库的联机分析处理系统也正在成为 IT 行业新的增长点。数据仓库技术起源于对大量数据进行处理的需要，是随着业务应用的需要而产生的。与传统的数据库技术相比，数据仓库为决策分析提供了更好的支持，超出了传统联机事务处理的范畴。因此近几年来，数据仓库技术发展很快，并在各个行业都得到了很多的应用。相对于国外大中型企业，我国企业的数据仓库建设目前还处于起步和探索阶段，在电信企业这样大型的信息化产业内部建设数据仓库应用，对技术更是提出了更高的要求。

现在介绍数据仓库和 OLAP 技术的书籍和资料在概念和结构方面讨论较多，本文希望通过讨论在设计和具体实现数据仓库，以及基于数据仓库的 OLAP 的过程中遇到的一些比较实用和细节的问题，达到从实际出发、突出实用性和集成性的特点。

本文首先阐述了数据仓库的和联机分析处理的概念和发展历史，以及当前我国电信行业对传统数据库的应用情况，接下来从理论上分析了数据仓库和多维分析与传统数据库应用的不同之处，重点阐述了对数据进行多维分析的概念和方法。在应用分析部分，针对数据仓库建设的各个重点环节进行分析和讨论，结合理论知识和实际经验得出最适合当前电信行业应用的处理方案或建议，并讨论了当前可应用数据仓库和联机分析处理技术的领域。最后给出了一个数据仓库的应用实例－电信话单分析系统，通过该系统的从建模到数据抽取，到多维分析应用，展示了如何从现有业务系统上建立数据仓库应用和多维分析方法。

建设数据仓库系统能够极大地提高国内电信企业的业务支撑能力，丰富企业的业务应用内容，提高企业的市场竞争力，缩短与国际电信企业在运营管理能力方面的差距。为迎接将来更开放的、竞争更激烈的电信市场做好技术准备。

关键词：数据仓库、多维分析、联机分析处理、决策支持系统、关系数据库

Abstract

At present, the data warehouse technology is being in the fast development time, and on-line analysis processing system based on the data warehouse technology is becoming a new growing point of the IT profession. The data warehouse technology origins in carries on processing to the mass data, it is along with the service application need. Compares with the traditional database technology, the data warehouse has provided a better support for the decision analysis and jumped out the categories in traditional on-line business processes. Therefore, data warehouses technological development is very quick in the last few years, and peoples have developed many applications in each profession. Compare to the overseas middle or large scale enterprises our country enterprise's data warehouse construction is still in the exploration stage at present. In large-scale and information based enterprises like the telecommunication companies in China, building data warehouse application is a high-level request to the technology users.

Now, books and the articles which introduced the data warehouse and the OLAP technology discuss many in the concept and the structure aspect, this article hoped through the discussion in the design and the specific implementation data warehouse, and OLAP based on data warehouse in practical and detail way, achieved embarks, prominent usable from the reality and the integration characteristic. This article first elaborated the data warehouse and the on-line analysis processing concept and their development histories, and then Described the situation of the traditional database application in our current country telecommunication profession; Met down this article theoretically analyzed the deference between the data warehouse and the multi-dimensional analysis technology to the traditional database applications, and elaborated with emphasis the concept and the method of the multi-dimensional analysis to the data. In the application analysis part, carried on the analysis and the discussion in view of data warehouse construction in each key point, try to obtains the way which most suits the current telecommunication profession application with theory knowledge and the practical experience, then discussed several possible domain which the data warehouse and the online analytical processing technology can be used. Finally in this article has produced a data warehouse application example - telecommunication telephone records analysis system, through the Modeling process, data extract, to the multi-dimensional analysis application, had demonstrated how to establish a data warehouse application and the multi-dimensional analysis based on the existing

business processing systems.

Build data warehouse system can enormously enhance the service ability of domestic telecommunication enterprise and enrich enterprise's service application content, it also can enhances the competitive power in market, reduces the distance with international telecommunication enterprise. Using data warehouse for the market which more opening and the competitions which more intensions in the future.

Keywords: Data Warehouse, Multi-dimensional Analysis, On-line Analysis Processing, Decision Support System, Relational Database

原创性声明

本人郑重声明：所呈交的硕士学位论文，是本人在指导教师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：李哲峰

日期：2007 年 3 月

第一章 绪论

1.1 研究背景

1.1.1 数据仓库和 OLAP 技术的发展

随着以服务为中心的第三产业在现代社会经济中所占比重的日益增大，传统的面向工业制造业的管理模式已不再适合人们的需要。管理学家 Peter.D.Rucker 提出了知识管理的革命概念，指出企业成功的关键在于能否有效地获取和管理知识。企业从本质上说是利用知识为用户解决问题的机构，有用的知识存在于大量的原始数据中，计算机的使用使得数据得以有效的保存和组织。计算机系统的功能从数值计算扩展到数据管理距今已有三十多年，最初的数据管理形式主要是文件系统，少量的以数据片段之间增加一些关联和语义而构成层次型或网状数据库，但数据的访问必须依赖于特定的程序，数据的存取方式是固定的、死板的。

到了 1969 年，E.F.Codd 博士发表了他著名的关系数据模型的论文。此后，关系数据库的出现开创了数据管理的一个新时代。近二十多年，大量新技术、新思路涌现出来并被用于关系数据库系统的开发和实现：客户/服务器体系结构、存储过程、多线索并发内核、异步 I/O 和代价优化，等等，这使得关系数据库系统的处理能力毫不逊色于传统封闭的数据库系统。而关系数据库在访问逻辑和应用上所带来的好处则远远不止这些，数据库查询语言(SQL)的使用已成为一个不可阻挡的潮流，加上近些年来计算机硬件的处理能力呈数量级的递增，关系数据库最终成为联机事务处理系统的主宰。整个 80 年代自到 90 年代初，联机事务处理(OLTP)一直是数据库应用的主流。然而，应用在不断地进步。当联机事务处理系统应用到一定阶段的时候，企业家们便发现单靠拥有联机事务处理系统已经不足以获得市场竞争的优势，他们需要对其自身业务的运作以及整个市场相关行业的态势进行分析，而做出有利的决策。这种决策需要对大量的业务数据包括历史业务数据进行分析才能得到。21 世纪后，随着计算机上数据库技术的成熟和广泛应用，类似电信、银行和保险等公共服务企业内部积累了大量的数据，这些数据包括以往的业务历史记录以及用户资料等。长期以来，在主要进行联机事务处理(OLTP)的操作型数据库环境下，上述数据仅用于业务流程的支持和历史数据的保存，无疑是对资源的一种浪费。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使用的时候了。”

事实上，将大量的业务数据应用于分析和统计原本是一个非常简单和自然的想法。但在实际的操作中，人们却发现要获得有用的信息并非如想象的那么容易：第一，所有联机事务处理强调的是密集的数据更新处理性能和系统的可靠性，并不关心数据查询的方便与快捷。联机分析和事务处理对系统的要求不同，同一个数据库在理论上都难以做到两全；第二，业务数据往往被存放于分散的异构环境中，不易统一查询访问，而且还有大量的历史数据处于脱机状态，形同虚设；第三，业务数据的模式针对事务处理系统而设计，数据的格式和描述方式并不适合非计算机专业人员进行业务上的分析和统计。因此有人感叹：20 年前查询不到数据是因为数据太少了，而今天查询不到数据是因为数据太多了。针对这一问题，人们设想专门为业务的统计分析建立一个数据中心，它的数据从联机的事务处理系统中来、从异构的外部数据源来、从脱机的历史业务数据中来……这个数据中心是一个联机的系统，它是专门为分析统计和决策支持应用服务的，通过它满足决策支持和联机分析应用所要求的一切。这个数据中心就叫做数据仓库(Data Warehouse)。这个概念在 90 年代初被提出来，并在信息领域迅速兴起。对于数据仓库的具体定义，目前还存在较大争议。数据仓库之父 W.H. Inmon 指出："数据仓库是支持企业或组织的决策分析处理的、面向主题的、集成的、不可更新的、随时间不断变化的数据集合"[1]。在实际应用中，也存在另一种更加明晰的阐述了数据仓库和事务处理型数据库之间的关系的定义：数据仓库是为了查询(Querying)和报告(Reporting)而专门构造的事务处理型数据的副本。数据仓库定义的核心就是要支持面向主题的决策分析，数据仓库所要研究和解决的问题就是如何从数据库中获取更多、更有用的信息。

联机分析处理(OLAP)是专门设计用于对储存在数据仓库中的数据进行复杂操作的技术。它是针对特定问题的联机数据访问和分析。通过对信息的多个角度(维)进行快速、一致、稳定地交互访问，使决策分析人员可以深入地进行观察。联机分析处理(OLAP)的概念最早是由关系数据库之父 E. F. Codd 于 1993 年提出的，他同时提出了关于 OLAP 的 12 条准则。OLAP 的目标是满足决策支持或者满足在多维环境下特定的查询和报表需求，它的技术核心是"维"这个概念。维是人们观察客观世界的角度，是一种高层次的类型划分。"维"一般包含着层次关系，这种层次关系有时会相当复杂。通过把一个实体的多项重要的属性定义为多个维(dimension)，使用户能对不同维上的数据进行比较。因此 OLAP 也可以说是多维数据分析工具的集合。

1.1.2 国内电信行业的背景情况

从上世纪八十年代开始，我国电信企业开始进行大规模的信息化建设，在近二十年的时间中，电信企业已全面实现了生产及服务过程信息化。随着业务不断趋于多样

化，各电信企业都针对不同业务建立了多个生产管理系统，如中国电信建设的生产系统就包括了进行业务受理、配线配号系统；进行计费、账务及欠费处理的计费系统；114，112，180，189 等专业系统；201，IC 等卡类管理系统；基于互联网信息管理的数据业务管理系统以及交换、传输、网管系统等。中国移动也建设了综合业务支撑(ROSS)系统；用于梦网短信接入的短信网关、彩信网关、GPRS 网关等各专业管理系统[2]。目前，电信企业建设的项目仍然以生产支撑系统为主。通过这些支撑系统的建设，规范了电信企业内部管理流程，大大提高了电信企业的上作效率，增强了业务水平，提高了企业的竞争力。

但是在进入 21 世纪以后，面临迅速膨胀的业务量，电信行业在信息管理方面面临新的挑战。一方面，业务支撑系统日益复杂化，并且不断地划分为各个生产子系统，使得业务分析人员获取有效数据的难度加大；另一方面，由于业务量的迅速发展，支撑系统的各生产子系统处理负担日益加重，而统计日益复杂，仍以传统的方式，在生产系统中进行统计分析，向市场营销人员及时提供充足、准确的经营信息而又不影响生产系统的处理效率已不现实。此外，由于硬件设施的差距在缩小，竞争将最终体现在对客户的价值取向和消费心理为导向，经营模式和服务体系也从“以业务为中心”转变为“以客户为中心”[3]。基于以上几点，有必要实现操作数据与经营数据的分离，形成统一的经营信息数据源，在服务支撑系统中为统计分析等经营信息服务建设专门的处理子系统；生产子系统视本身情况，周期地备份并清理历史数据；而经营信息服务子系统所需的大量历史数据不能自接依赖于生产子系统，必须周期性地从生产子系统中抽取，独立积累、独立存储、独立管理。随着市场竞争的不断加剧，对客户资源的争夺也进入了白热化的阶段，如何发展新用户，扩大自己的用户群；如何设计出更适合用户需要的业务，将用户绑定在自己的网络上；如何合理地设定资费在用户可以接受的水平；一个个新问题摆在了电信运营商的面前。

在激烈的市场竞争面前，要想科学的决策，离不开数据的支持，从企业对于数据分析的应用已经从简单的营业报表走向了经营分析系统并进一步提出了对决策支持系统(DSS)，经理信息系统 (EIS)的需求：从简单的客户资料统计走向了客户关系管理(CRM)，这些新一代的分析决策系统都需要一个稳定可靠的，独立于生产系统数据的信息平台。基于以上需求，电信行业建立基于数据仓库的分析平台已是势在必行。

1.2 研究的目的和意义

基于上述研究背景，本文的研究目的主要是以下几点：

1. 对数据仓库和 OLAP 技术在理论上同传统数据库技术的不同之处和技术难点进行探讨和分析。

2. 对于多维查询的概念和方法进行深入分析，对技术难点提出理论解决方案。
3. 对电信企业如何实施数据仓库和运用 OLAP 技术进行分析的方法要素进行阐述和分析，并提出建议。
4. 对数据仓库和 OLAP 应用系统的实现过程和应用方式进行实践。

1.3 研究工作及论文结构

1.3.1 构建数据仓库

数据仓库构建的具体步骤如下：

1. 确定数据仓库分析主题、目标、维度和维层次。
2. 定义元数据并设计数据仓库的总体结构，确定存储方式。
3. 数据的抽取、净化和验证。

1.3.2 OLAP 前端展现工具的开发

前端展现工具的开发具体步骤如下：

1. 把已有数据仓库架构映射到多维模型。
2. 设计用户图形界面，提供向导功能，以方便决策者操作。
3. 根据决策者提出的分析目标、提供的相关数据及约束条件自动处理分析请求，并将处理的最终结果显示在用户界面，以供决策者参考。

1.3.3 论文结构

本文共分六章，其中第四、五两章是核心部分：

第一章：绪论。在绪论中，介绍了数据仓库和 OLAP 技术发展的背景，电信行业的应用情况及本文的研究目的和内容。

第二章：数据仓库理论及应用。介绍了数据仓库的相关概念，分析了数据仓库的数据组织结构和体系结构，并指出了数据仓库的技术要求。

第三章：基于数据仓库的联机分析技术。重点介绍了 OLAP 相关技术，包括基本概念，与 OLTP 的关系以及 OLAP 的数据组织和多维分析结构。

第四章：电信话单分析系统设计。以电信运营中话单业务为实现背景，构建适合通信业数据仓库模型和体系架构。

第五章：电信话单分析系统实现。介绍了事实表与维度表的生成，及基于 Open Source 开源引擎构建符合 J2EE 规范的 OLAP 分析实现系统。

第六章：论文总结。总结了研究进展，并指出有待改进和优化之处。

第二章 数据仓库理论及应用

2.1 数据仓库的定义及用户

2.1.1 数据仓库的定义

提到数据仓库(Data Warehouse)，常常有人将其与数据库混为一谈，或者将它当作一个可以从“货架”上买到的产品。其实，数据仓库既非数据库，也不是一个实实在在的产品。

数据仓库技术是近年来出现的、发展迅速的一种技术,它通过把企业大量的历史数据整理集中到一个中央仓库中,将数据加以分析并呈现给用户来支持管理者的决策。数据仓库是一个整合式的、面向主题的、历史性的以及只读性的数据集合。这一定义清楚地揭示了数据仓库和传统关系数据库的不同应用目标。传统的关系型数据库技术主要为 OLTP 提供支持，如订票系统、储蓄系统等。而数据仓库技术应决策支持需求而生。

数据仓库整合来自企业各个业务系统的各种类型和格式的数据，进行系统加工、汇总和整理，形成一个完整而一致的企业全局信息库。数据仓库的数据按照有利于决策过程的主题进行组织，其中包含了数据的信息涵义，如销售情况、利润状况及信贷风险程度等。这样的数据集合便于信息分析和信息挖掘。除此之外，数据仓库系统中存储的数据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。

数据仓库之父 Bill Inmon 对数据仓库所下的定义是：数据仓库是面向主题的、集成的、稳定的、随时间变化的数据集合，用以支持管理决策的过程。著名的 DBS 和 MIS 专家 Rob Mattision[4]在 1996 年出版的“Data Warehouse”一书中也做如下的定义：数据仓库是一种新型的数据库，数据仓库被组织用作一个中性存储区，被 Data Mining 和其它应用程序所使用，使用这些数据将满足一组预定义的商业评判。由此可见，数据仓库是一个综合的解决方案。

一个数据仓库通常是一个分散的数据存储，在其中信息是存为这样的一种形式，它适合于业务智能化和决策支持系统。数据可能是以不同形式存储的，它并不影响 OLTP 系统的运作。数据仓库的建立是用一种循环的逐步完善的过程而不是一步完善的。数据仓库通常是与解决企业不断改变的组织问题的全过程有关。

数据仓库通常是围绕主题建立的。主题就是企业感兴趣的论题，比如部门、活动和操作结果。数据仓库的结构是由数据仓库应满足的应用决定的。快速提交信息是成功实施数据仓库的关键。由于这一点，就引入了数据集市和信息集市这些概念。数据集市是数据仓库的一个子集，它通常更为概括，以满足对关心数据的查询有比数据仓库本身更快的速度。信息集市存储可用视窗器(viewer)显示的预处理的信息。

2.1.2 数据仓库的用户

数据仓库的用户可以分为信息人员和信息使用人员。

信息人员在创建分析的时候并不知道需求。在创建数据仓库的过程中，信息人员要完成四种类型的工作：概况分析、抽取、建模和分类。信息人员要从当前成功运行的关系型数据库中查看大量的数据，要考虑数据之间的关系、关联和数据模型。

信息使用人员是数据仓库的大量用户。他们在使用数据仓库的时候，知道自己所需求，用一种可以预测的，重复性的方式来使用数据仓库平台。信息使用人员实际上是从战术上监控决策的效果。例如：医院系统中药费收入的比例问题，根据信息使用人员的报告，在一段时期内，医院的药费收入在医院的总收入中的比例过高，此时，信息人员应开始调查为什么在这段时期内，药费的收入比例会增大，得出结论后将信息提交给领导，以便领导采取相应的管理措施。

2.2 数据仓库的特征及其作用

2.2.1 数据仓库的特征

1. 面向主题的

数据库是面向应用设计的，它的数据只是为处理具体应用而组织在一起的，反映了一个单位数据的动态特征，即各个部门间的数据处理流程，这种数据组织方式具有较强的操作性，但它对于数据内容的划分不适用于分析。主题是一个在较高层次将信息系统中的数据综合、归类并进行分析利用的抽象，每一个主题基本对应某一宏观分析领域所涉及的分析对象。即主题是一个在较高层次将数据归类的标准，每一个主题基本对应一个宏观的领域，每个领域有自己的逻辑内涵互不交叉。面向主题的数据组织方式，就是在较高层次上对分析对象的数据的一个完整性、一致性的描述，能完整、统一地刻画各个分析对象所设计的各项数据，以及数据之间的联系。数据进入数据仓库之前，必然要经过加工与集成，将原始的数据结构做一个从面向应用到面向主题的转变。

2. 集成的

数据仓库中的数据来自多个外专业应用系统，但并不是对这些数据的简单归类与拷贝，它应该是对源数据的增值和统一，经必要的变换以最适合使用的方式存储起来，支持联机分析处理。

3. 非易失的(相对稳定的)

数据仓库主要是为信息分析提供综合的、集成的、面向主题的数据，这些数据原则上不允许信息分析人员直接对数据执行修改或删除操作，进入数据仓库的数据则是相对稳定的。

4. 反映历史变化的

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，信息才能发挥作用和意义。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，就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。

2.2.2 数据仓库的作用

数据仓库主要有以下几方面的作用[5]：

首先，数据仓库支持多维分析。多维分析是通过把一个实体的多项重要的属性定义为多个维度，使得用户能方便地汇总数据集，简化了数据的分析处理逻辑，并能对不同维度的值的数据进行比较，而维度则表示了对信息、的不同理解角度。应用多维分析可以在一个查询中对不同阶段的数据进行纵向或横向比较，这在决策过程中非常有用。

其次，数据仓库是数据挖掘技术的关键基础。数据挖掘技术要在已有数据中识别数据的模式，以帮助用户理解现有的信息，并在已有信息的基础上，对未来的状况做出预测。在数据仓库的基础上进行数据挖掘，就可以针对整个企业的状况和未来发展做出较完整、合理、准确的分析和预测。

2.3 数据仓库与操作型数据库分析

传统的数据库系统由于主要用于企业的商务日常事务处理工作，主要执行的是联机事务和查询处理，是为 企业的特定的应用需求而服务的，用户关心的是响应时间、数据安全性和完整性[6]。存放在数据库中的数据也就遵循了操作型数据的特点，而为适应数据分析处理需求而产生的数据仓库中所存放的数据就应该是分析型的数据。具体差异比较如下。

表 2.1 数据库与数据仓库的区别

OLTP 系统数据模型与数据仓库 OLAP 数据模型的特点比较	
OLTP 系统	数据仓库 OLAP 系统
规范化的	非规范化的
无派生数据	有派生数据
使用许多不易理解的代码	有完整的数据描述
记录中不一定有时间字段	一定要有作为关键字的时间字段，以保证历史数据的唯一性
秒级以下的响应时间	秒级到分钟级的查询响应时间
业务数据	没有"纯"业务数据

2.3.1 基本任务差异

数据仓库的基本任务与传统的数据库基本任务有很大的区别，由于数据仓库的数据源可以来自于不同的 DBMS 的数据库(内部数据源)，也可以来自于不同格式的文件中(外部数据源)。这些数据源可以看作数据仓库中输送数据的管道。在输送数据的过程中，数据仓库的设计者必须考虑如下任务：

1. 将这些数据源的模型转换成通用的描述形式。
2. 将同义的数据元素的名称、数据类型、尺寸进行统一的规范——即净化数据元素。
3. 必须从各数据源中抽取子集，为形成数据仓库的整体模型奠定基础。
4. 把相似的数据源集成为统一的资源模型。
5. 通过增加时间戳、来源戳、分割、衍生元素，提供扩展的模型用于存储聚集、概括值，从而获得数据仓库模型。

2.3.2 数据主要特征差异

数据仓库和操作型数据库在数据来源、数据内容、数据模式、服务对象、访问方式、事务管理和模型构建等方面都有不同的特点和要求，不管是在性能上，还是在功能上都存在较大的区别。事务处理通常只是针对当前和短期存储的数据，且不同数据的保存期限也不尽相同，即使有一些历史数据保存下来了，也很难得到充分利用。但对于决策分析而言，历史数据是相当重要的，许多分析方法必须以大量的历史数据为依托。没有历史数据的详细分析，难以把握未来发展趋势。数据仓库与操作型数据库的数据特征的比较如下：

1. 面向主题的结构设计。数据仓库是以最终用户的观点组织和管理数据，数据库是为了提高应用程序查询数据的效率，因而是以应用的观点设计数据库结构。
2. 管理大量的信息。由于数据仓库的设计目标是在众多的数据库中获得决策信息，因而它含有大量的历史数据。而传统的数据库为了提高系统的运行效率，通常会对历史数据进行必要的备份后，将其从运行库中清除。
   例如：在医院管理信息系统中，当一个病人住院，联机业务处理系统就要产生关于这个病人的记录，随着对病人治疗的不断进行，记录不断的被加工，当这个病人治愈出院完成手续后，病人的信息将从运行库中清除并转移到历史库中去，不能再修改。
3. 异质的数据源。由于数据仓库的数据源来自于不同种类的文件，数据存储的介质和格式会有很大的不同。因而数据仓库不仅要处理不同数据库中的信息，还必须处理不同格式的数据文件。
4. 高度概括的信息。传统的数据库存储的信息具体而且详细，但不利于用户理解，数据仓库必须从大量具体的数据中进行高度概括，并从中挖掘出准确的信息。

2.3.3 数据操作方式差异

1. 数据库支持用户对大量数据进行更新操作，由很多的短小的事务处理组成，注重于事务速率；而数据仓库中则主要是查询操作，与数据库相比，数据仓库中的数据更加稳定。
2. 数据库为用户和开发者提供的是非常庞大和复杂的结果，但是数据仓库中提供的是用于分析决策、易于理解的结果。
3. 数据库主要保存的是当前的数据，历史的数据被及时的归档后立即删除，以

提高系统的运行效率；数据仓库中则存储了大量的衍生数据，目的是为了节省工作量和提高系统的运行效率。

由于以上种种的原因，传统的数据库和数据仓库的建模方法有很大的区别。

2.3.4 数据模型与构建方法

1. 传统的数据库模型有三种：层次型、网络型和关系型。目前主要流行使用的数据库产品是关系型数据库。
2. 数据仓库的模型也有三种：星型模型、雪花模型和混合模型。

2.4 数据仓库的基本组成

2.4.1 数据仓库的基本结构

一个完整的数据仓库系统应当具备建立、管理和使用等功能。W.H. Inmon 认为，数据仓库系统可以分为三个组成部分[7]：

1. 数据源——提供源数据。
2. 数据的存储与管理——包括来自数据源数据的接收、析取、汇总、变换和储存。
3. 前端服务——面向用户的数据需求，完成数据提取和计算分析等功能。

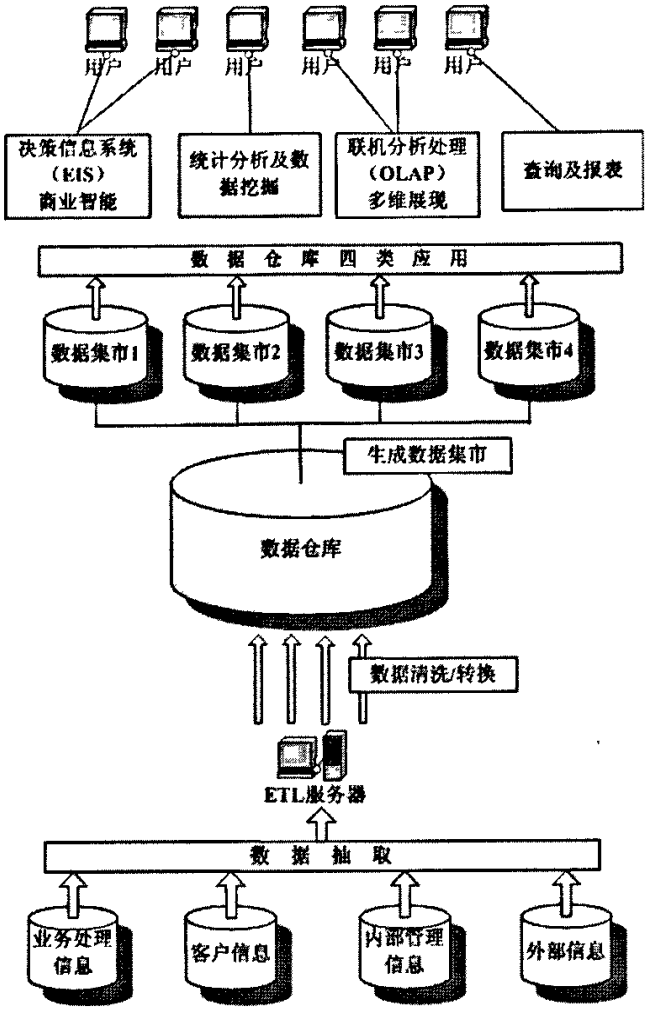


图 2.1 数据仓库构建体系结构图

如图 2.1 所示，数据仓库应具有多种工具：从多个操作型数据库和外部数据源中抽取数据的工具；清洗、转化和整合数据的工具；将数据装载到数据仓库中的工具；定期刷新数据仓库以反映数据源中的更新和从数据仓库中清除数据的工具。除了主数据仓库外，还有存在很多部门性的数据集市，数据集市实际上就是为了满足企业内各部门的分析需求而建立的微型数据仓库。数据仓库和数据集市中的数据由一个或几个数据仓库服务器存储和管理，数据仓库服务器通过前端工具将这些数据从多维角度展现出来。前端的工具包括：查询工具，报表生成器，分析工具和数据挖掘工具。最后，还有一个存储和管理元数据的元数据库以及监视和管理数据仓库系统的工具。

为了装载的平衡，较好的可测量性和较高的获取能力，数据仓库可以是分布式的。在分布式结构中，元数据库通常和数据仓库的各个片段重复并且整个数据仓库是集中

管理的。如果花费太大而不能创建一个单一的逻辑性的整合的企业级数据仓库，为了方便的实现可以构建联合的数据仓库或者数据集市，这些数据仓库和数据集市都有自己的仓库结构和各自分散化的管理。

设计和实现一个数据仓库是个复杂的过程，通常包括以下几个步骤：

1. 定义结构，选择存储的服务器，数据库和 OLAP 服务器以及工具。
2. 设计数据仓库体系结构和视图。
3. 定义数据仓库的物理组织，数据的放置、划分和获取方法。
4. 利用网关、ODBC 驱动等连接数据源。
5. 设计和实现数据抽取、清洗、转化、装载和刷新的程序脚本。
6. 利用计划和视图的定义、脚本以及其他元数据控制数据仓库。
7. 设计和实现用户端的应用程序。
8. 整理展现数据仓库和所有应用。

2.4.2 数据集市的概念

数据仓库系统中另一个重要的组件是数据集市，原始数据从数据仓库流入到不同的部门中以支持这些部门的定制化使用，这些部门级别的数据库就称为数据集市。

数据集市中包含部门决策支持处理所需要的任何数据，在数据集市中包含有多种多样的数据：即包含很多动态的概括数据也包含很多准备好的详细数据。这两类数据构成了数据集市环境中的大部分数据。

数据集市提供了一种企业视图，因为它贴近特定的财务和营销部门的重要用户。数据集市可以通过用户群来组织(物理上位于用户部门)，或者按主题域来组织即以逻辑形式组织存在数据仓库内的另外空间。数据复制和传播会在数据仓库和从属型数据集市之间实现数据同步。

数据集市是数据仓库有效的和自然的补充。数据集市延伸决策支持到部门级环境中。数据仓库提供粒状数据并且不同数据集市应用不同的方法来解释和构造这种粒状数据以满足部门决策的需要。对数据集市来说最适当的数据源是数据仓库，业务数据库不是数据集市的合适的数据源，数据集市还可以包括外部数据。

除了数据库之外，数据集市所使用的软件还有：访问和分析工具、自动接口生成、系统管理、净化/归档、元数据管理等。

2.5 数据仓库的建模技术

2.5.1 数据仓库建模的原则

模型是对现实事物的反映和抽象，它可以帮助我们更加清晰的了解客观世界。数据仓库建模是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，是数据仓库项目成功与否最重要的技术因素。大型企业的信息系统一般具有业务复杂、机构复杂、数据庞大的特点，数据仓库建模必须注意以下几个方面：

1. 满足不同用户的需要

大型企业的业务流程十分复杂，数据仓库系统涉及的业务用户众多，在进行数据模型设计的时候必须兼顾不同业务产品、不同业务部门、不同层次、不同级别用户的信息需求。

2. 兼顾效率与数据粒度的需要

数据粒度和查询效率从来都是矛盾的，细小的数据粒度可以保证信息访问的灵活性，但同时却降低了查询的效率并占用大量的存储空间，数据模型的设计必须在这矛盾的两者中取得平衡，优秀的数据模型设计既可以提供足够详细的数据支持又能够保证查询的效率。

3. 支持需求的变化

用户的信息需求随着市场的变化而变化，所以需求的变化只有在市场竞争停顿的时候才会停止，而且随着竞争的激化，需求变化会越来越频繁。数据模型的设计必须考虑如何适应和满足需求的变化。

4. 避免对业务运营系统造成影响

大型企业的数据仓库是一个每天都在成长的庞然大物，它的运行很容易占用很多的资源，比如网络资源、系统资源，在进行数据模型设计的时候也需要考虑如何减少对业务系统性能的影响。

5. 考虑未来的可扩展性

数据仓库系统是一个与企业同步发展的有机体，数据模型作为数据仓库的灵魂必须提供可扩展的能力，在进行数据模型设计时必须考虑未来的发展，更多的非核心业务数据必须可以方便的加入到数据仓库，而不需要对数据仓库中原有的系统进行大规模的修改。

2.5.2 数据仓库的数据模型层次

在创建数据仓库时，需要使用各种数据模型对数据仓库进行描述。数据仓库的开

发人员依据这些数据模型，才能开发一个满足用户需求的数据仓库。数据仓库的各种数据模型在数据仓库的开发中作用十分明显，主要体现在模型中只含有与设计有关的属性。这样就排除了无关的信息，突出与任务相关的重要信息，使开发人员能够将注意力集中在数据仓库开发的主要部分。模型有更好的适应性，更易于修改。当用户的需求改变时，仅对模型做出相应的变化就能反映这个改变。

数据模型是对现实世界进行抽象的工具。在信息管理中需要将现实世界的事物及其有关特征转换为信息世界的数据，才能对信息进行处理与管理，这就需要依靠数据模型作为转换的桥梁。这种转换经历了从现实到概念模型，从概念模型到逻辑模型，从逻辑模型到物理模型的转换。在数据仓库建模的过程中同样也要经历概念模型、逻辑模型与物理模型的三级模型开发。因此，数据建模可以分为三个层次：高层建模（实体关系层，概念模型），中间层建模(数据项集，逻辑模型)、底层建模(物理模型)。

概念世界是现实情况在人们头脑中的反映，人们需要利用一种模式将现实世界在自己的头脑中表达出来。

逻辑世界是人们为将存在于自己头脑中的概念模型转换到计算机中的实际物理存储过程中的一个计算机逻辑表示模式。通过这个模式，人们可以容易地将概念模型转换成计算机世界的物理模型。

物理世界是指现实世界中的事物在计算机系统中的实际存储模式，只有依靠这个物理存储模式，人们才能实现利用计算机对现实世界的信息管理。

图 2.2 数据模型层次

2.5.3 维度建模理论及方法

维度建模是一种逻辑设计技术，该技术试图采用某种直观的标准框架结构来表现数据，并且允许进行高性能存取。它必然会遵循维度方面的规范，并且坚持带有某些重要限制条件的关系模型模范。维度模型由事实和维度表构成。术语“维度”与“事实”最初是 20 世纪 60 年代在一个由 General Mills 与 Dartmouth 大学主持的联合研究计划中提出的。70 年代，AC Nielsen 和 IR 全都一致地使用这些术语描述其辛迪加数据(关于零售数据的维度数据中心)发布应用。事实表是维度模型的基本表，用于存放大量的业务性能度量值，表示维度间多对多的关系。维度值的列表则给出事实表的粒度定义，并确定度量值的取值范围是什么。实际上，维度表属性是查询约束条件与报表标签生成的基本来源。在许多方面，数据仓库不过是维度属性的体现而已。数据

仓库的能力直接与维度属性的质量和深度成正比。

理解了事实和维度表之后，就需要考虑将两个组块一起融合到维度模型中的问题，即维度建模的方法。

星型模式是建立多维模型的常用方法，它的核心思想就是在数据库中的数据之间建立简明的关系，限制必须建立的连接的数量，并降低连接的复杂程度。它是由两类基本表组成：一个事实表(fact table)和多个维表(dimension table)。事实表包含实际的业务数据。维表包含有关业务的描述信息。星型模式是面向主题的，事实表描述了主题的数据，维表则从不同角度描述了主题的分析尺度。

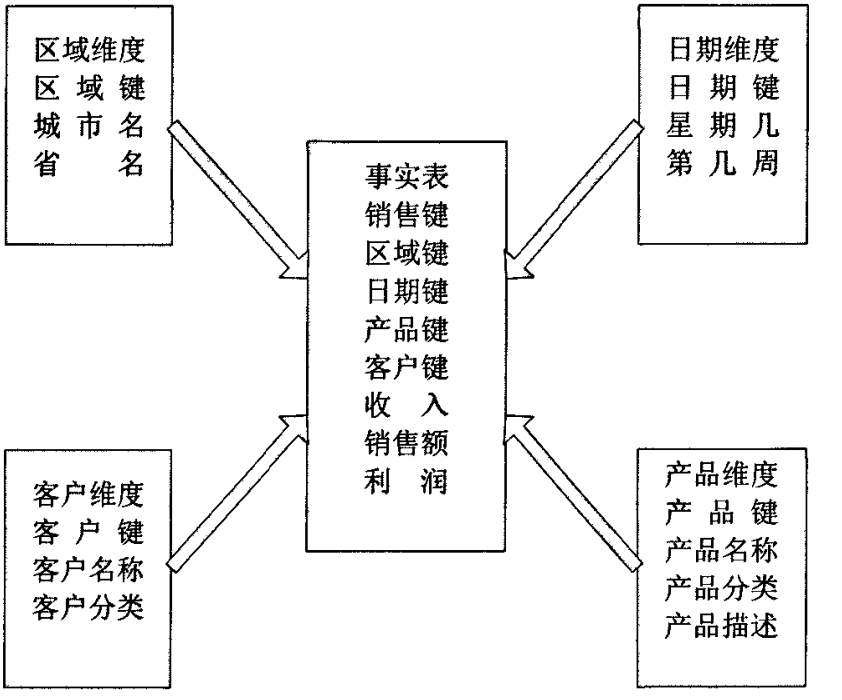


图 2.3 星型模式示例

星型模式的设计主要就是对事实表和维表进行设计。针对如何建模有两种截然不同的观点。许多人建议从维表开始，然后再到事实表。而在本文中作者将先从事实表开始，然后再到维表。当然，如果最后结果一样，以哪种方式建模并不是最重要的。从事实表设计开始，在事实表的每条记录都包含一个主键，该主键由指向维表的外键(foreign key)和通过主键独一无二的事实或度量值(measure)连接而成。事实表是满足三个范式的高度规范化的结构。在事实表的设计中，我们应注意以下几点：

1. 事实表中应包含决策所需的各种度量值(measure)；

2. 事实表中的数据粒度要根据业务需求确定，并尽可能按最高的细化级别存储数据；
3. 数据的记录方式要根据业务需求确定。

单独的一个事实表是毫无用处的，我们必须对外键进行解码。这就是维表所要做的，它们为每个事实给出相应的含义。维表采用逆规范化结构。通常维表一旦建立，内容相对固定，并且维表通常比事实表短而宽，占用存储空间较小。维表设计时要注意维的级别，并且维表的数据要尽可能地包括所有业务数据。维表的设计对于整个多维数据模型非常重要。

但有时候，维表的定义会变得复杂，例如对产品维，既要按产品种类进行划分，对某些特殊商品，又要另外进行品牌划分，商品品牌和产品种类划分方法并不一样。因此，当单张维表不是理想的解决方案时，可以采用以下方式，这种数据模型实际上是星型结构的拓展，称为雪花型模式(snow-flake schema)。

雪花模式，它是星型模式的一个变体，很像一个具有规范化维表的星型模式。雪花模式的数据结构比较清晰，但是在实际的工程应用中，推荐使用真正的星型模式。因为采用雪花模式，当需要执行一个报表按年显示销售情况，将不得不将事实表中的每一行与日期表、月表、年度表连接，而在一个真正的星型模式中，只需将事实表与日期表相连，显然这种方式的工作量小很多，这也意味着更快的查询速度。虽然难免会碰到维度定义复杂的情况，需要面对不同级别的维度表，但这时可以通过对维度表进行塌缩以将不同层级的同一系列维度表合成一个维度表。

2.6 数据仓库的数据存储和管理

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，同时也决定了其对外数据表现形式。要决定采用什么产品和技术来建立数据仓库核心，则需要从数据仓库的技术特点着手分析。

数据仓库遇到的第一个问题是对大量数据的存储和管理。这里所涉及的数据量比传统事务处理大得多，且随时间的推移而累积。从现有技术和产品来看，只有关系数据库系统能够担当此任。关系数据库经过近 30 年的发展，在数据存储和管理方面已经非常成熟，非其它数据管理系统可比。目前不少关系数据库系统已支持数据分割技术，能够将一个大的数据库表分散在多个物理存储设备中，进一步增强了系统管理大数据量的扩展能力。采用关系数据库管理数百个 GB 甚至到 TB 的数据已是一件平常的事情。

数据仓库要解决的第二个问题是并行处理。在传统联机事务处理应用中，用户访问系统的特点是短小而密集：对于一个多处理机系统来说，能够将用户的请求进行均

衡分担是关键，这便是并发操作。而在数据仓库系统中，用户访问系统的特点是庞大而稀疏，每一个查询和统计都很复杂，但访问的频率并不是很高。此时系统需要有能力将所有的处理机调动起来为这一个复杂的查询请求服务，将该请求并行处理。因此，并行处理技术在数据仓库中比以往更加重要。

数据仓库的第三个问题是针对决策支持查询的优化。这个问题主要针对关系数据库而言，因为其它数据管理环境连基本的通用查询能力都还不完善。在技术上，针对决策支持的优化涉及数据库系统的索引机制、查询优化器、连接策略、数据排序和采样等诸多部分。普通关系数据库采用 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。而扩充的关系数据库则引入了位图索引的机制，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。由于数据仓库中各数据表的数据量往往极不均匀，普通查询优化器所得出得最佳查询路径可能不是最优的。因此，面向决策支持的关系数据库在查询优化器上也作了改进，同时根据索引的使用特性增加了多重索引扫描的能力。

数据仓库的第四个问题是支持多维分析的查询模式，这也是关系数据库在数据仓库领域遇到的最严峻的挑战之一。用户在使用数据仓库时的访问方式与传统的关系数据库有很大的不同。对于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。它的特点是将数据想象成多维的立方体，用户的查询便相当于在其中的部分维(棱)上施加条件，对立方体进行切片、分割，得到的结果则是数值的矩阵或向量，并将其制成图表或输入数理统计的算法。

关系数据库本身没有提供这种多维分析的查询功能，而且在数据仓库发展的早期，人们发现采用关系数据库去实现这种多维查询模式非常低效、查询处理的过程也难以自动化。为此，人们提出了多维数据库的概念。多维数据库是一种以多维数据存储形式来组织数据的数据管理系统，它不是关系型数据库，在使用时需要将数据从关系数据库中转载到多维数据库中方可访问。采用多维数据库实现的联机分析应用，我们称之为 MOLAP。多维数据库在针对小型的多维分析应用时有较好的效果，但它缺少关系数据库所拥有的并行处理及大规模数据管理扩展性，因此难以承担大型数据仓库应用。这样的状态直到“星型模式”在关系数据库设计中得到广泛的应用才彻底改变。几年前，数据仓库专家们发现，关系数据库若采用“星型模式”来组织数据就能很好地解决多维分析的问题。“星型模式”只不过是数据库设计中数据表之间的一种关联形式，它的巧妙之处在于能够找到一个固定的算法，将用户的多维查询请求转换成针对该数据模式的标准 SQL 语句，而且该语句是最优化的。“星型模式”的应用为关系数据库在数据仓库领域打开绿灯。采用关系数据库实现的联机分析应用称为 ROLAP。目前，大多数厂商提供的数据仓库解决方案都采用 ROLAP。

在数据仓库的数据存储管理领域，从当今的技术发展来看，面向决策支持扩充的

并行关系数据库将是数据仓库的核心。

2.7 建立数据仓库所面临的问题

目前，企业在数据仓库实施过程遇到的问题主要有以下几个方面：

1. 传统业务系统不同，数据仓库是面向管理决策层应用，必须有系统自身的最终用户—企业决策层的参与。

   数据仓库应用本身并不是业务流程的再现，而是基于数据分析的管理模式的体现。在这个层次上，数据仓库对于企业决策层的意义首先不是信息技术和产品上的，而是企业经营管理模式上的。数据仓库的实施者需要加强商业智能化企业在获得市场竞争力上的作用，提供切实有效的系统实施目标和规划，使得企业决策层充分认识到数据仓库是必需的系统，在投入和配合上给予充分的支持。

2. 由于数据仓库的访问和查询往往能够通过工具来提供，因此数据仓库的功能取决于系统的规划和设计。

   在了解数据仓库应用需求的时候，主要的对象应该是企业的决策部门和管理部门，而不是信息、系统部门。因为信息系统的人员专注在计算机系统本身，缺乏对整个业务运营的管理知识和视角。了解应用的需求必须从企业如何利用信息进行管理的角度出发，需要有丰富的行业经验。

3. 数据仓库的建立缺乏完整和高质量的数据。

   许多管理决策需要的信息并不在目前的业务系统中，或者这些信息是不完整的、不准确的。由于数据仓库是独立于业务系统的，数据仓库的实施将以管理层需要的分析决策为主线，在设计中可以为不确定数据预留空间。对于数据的完整性和质量问题可通过如下方式处理：利用多种方式加载数据，可以设计专门的输入接口收集数据，如获取客户的个人资料；放宽数据的时效性，在分析中标明个别数据的有效时间；在系统中标识出低质量的数据，规范业务系统。

4. 数据的抽取、转换和装载是一项非常烦琐的工作。

   由于原始数据信息的分散和非完整性，数据仓库实施过程中有 60%的费用和80%的时间与数据抽取和装载有关，数据的装载成为数据仓库实用的障碍。在系统实施过程中应该由专门小组或人员负责数据抽取的工作，将其纳入统一的管理和设计，不仅考虑原始数据源的类型，还必须考虑抽取的时间和方式。一个数据仓库系统往往同时存在多种数据抽取方式以适应原始数据的多样性，工作的原则是：简便、快捷、易维护。

5. 用户对数据仓库的认识通常从报表起步，但数据仓库并不是为业务报表而设计。

   数据仓库的分析工具在固定格式的报表上有时不如专门定制的程序。数据仓库的强项在于提供联机的业务分析手段，正因为数据仓库的使用，才使管理人员逐步摆脱对固定报表的依赖，取而代之地以丰富、动态的联机查询和分析来了解企业和市场的动态和历史信息的分析以及对其业务、客户、发展趋势的分析。

6. 系统的实施需要明确的计划和时间表。

   数据仓库的价值在于使用，如果让一些没有必要的信息去指导决策，那么数据仓库将永远停留在投资阶段。在定义实施计划时，需要明确系统的使用范围、用户的应用模式等与选择具体产品相关的重要问题。

第三章 基于数据仓库的联机分析(OLAP)技术

3.1 联机分析(OLAP)的基本概念

3.1.1 联机分析(OLAP)概念的提出

联机分析处理((OLAP)概念最早是由关系数据库之父E. F. Codd于1993年提出的，E. F. Codd认为联机事物处理(OLTP)不能满足终端用户对数据库查询分析的需要，SQL对大数据库进行的简单查询不能满足用户分析的需求。用户的决策分析需要对关系数据库进行大量计算才能得到结果，而查询的结果并不能满足决策者提出得需求。因此，E. F. Codd提出了多维数据库和多维分析的概念，即OLAP。

OLAP是独立于数据仓库的一种技术概念。其基本思想是：企业的决策者应能灵活地操纵企业的数据，以多维的形式从多方面和多角度来观察企业的状态、了解企业的变化。OLAP的目的在于共享多维信息的快速分析。OLAP系统与数据源的数据存储相分离，只要提供足够的分析数据就可以完成OLAP分析。当OLAP作为独立的使用方式时，其数据组织与数据仓库的组织方式相同。当OLAP与数据仓库结合时，OLAP的数据来源于数据仓库。数据仓库中存储的大量数据是根据多维方式组织的，是OLAP最适合的数据组织形式。多维结构是决策支持的支柱，也是OLAP的核心。OLAP展现在用户面前的是一幅幅多维视图。

3.1.2 联机分析(OLAP)的特点

Nigel Pendse提出的FASMI(Fast Analysis Of Shared Multidimensional Information)。他将OLAP所满足的特点用五个词来描述[8]：

快速性(Fast)：对用户请求的快速响应，用户对OLAP的快速反应能力有很高的要求。

分析性(Analysis)可以应用多种统计分析工具、算法对数据进行分析；

共享性(C Shared)：多个用户同时存取数据时，保证系统的安全性；

多维性(Multidimensional)：系统必须提供数据的多维视图和分析，包括对层次维和多重层次维的完全支持。

信息性(information)：指在OLAP系统中给出的不再是OLTP系统中散乱的数据，

而是能够导入具有指导意义的信息，同时要求数据能够以多种图形方式进行展示。

其中的主要特点有两方面的体现：一是在线性(On-Line)，体现为对用户请求的快速响应和交互式操作，它的实现是由 Client/Server 这种体系结构来完成的；二是多维分析(Multi-Analysis)，这也是 OLAP 技术的核心所在。

3.1.3 联机分析(OLAP)的准则

关系数据库之父 E.F.Codd 于 1993 年提出 OLAP 概念的时候，就描述了 OLAP 的 12 条准则，如下：

准则 1　OLAP 模型必须提供多维概念视图
准则 2　透明性准则
准则 3　存取能力推测
准则 4　稳定的报表能力
准则 S　客户/服务器体系结构
准则 6　维的等同性准则
准则 7　动态的稀疏矩阵处理准则
准则 8　多用户支持能力准则
准则 9　非受限的跨维操作
准则 10　直观的数据操作
准则 11　灵活的报表生成
准则 12　不受限的维与聚集层

3.2 联机分析(OLAP)技术的相关术语

3.2.1 变量(variable)

“变量”是对数据所描述具体对象的定义，它说明了数据在现实世界中的实际意义，为数据赋予了具体的含义，即数据“是什么”。同样的数字，在不同的变量定义中，具有完全不同的含义。例如，数字 81，在证券行情库中，某股票的最新价格，在学生成绩表中，表示　学生某项考试的成绩，在商品价格表中，又可以表示某种商品的零售价格。一般而言，变量是数值型试题指标，如“价格”、“数量”等，但它也可以是其他数型的指标，如字符串型（商品代码）、时间型（交易时间）等。

3.2.2 维(Dimension)

“维”是指人们观察某个数据集合的特定角度，它是以对数据的某个共性的提取为前提的。同一个问题，可以从不同的进行观察分析。维度依赖于表达业务成效的关键性能指标，能够回答类似下列问题：何时(when)，何地(where)，何人(who)等。比如运营商常常关心收入随着时间推移而产生的变化情况，这时就是从时间的角度来观察收入情况，在这里时间就是一个维(时间维)。而按所处地区进行营业额分析时，观察问题的维就变成了“地理维”。同样，还可以从“营业额”、“付款方式”等多个维开展分析工作。

3.2.3 维的层次(Dimension Hierarchies)

在同一个维度上，可以存在多个程序不同的细节，这些细节就是“维的层次”，它是对“维”的进一步细化。当人们从某个特定角度观察问题时，按所依据的细节程度（即维层次）的不同，可以得到多种描述方法。例如描述日期维时，月，季度，年构成日期维的层次(Hierarchies)，其中年、季度、月份别是日期维的第一、第二、第三层(level)。

3.2.4 维成员(Dimension Member)/类别(Categories)

“维成员”是指某个维的某个具体取值。如果该维是多层次的，那么维成员也是由在该维各层次上的取值组合而成。例如 2007 年 5 月 1 日，就是时间维上的一个维成员，它由年、月、日三个不同维层次上的取值组成。数据被组织进各个维度，并放在相应级别的层里。类别是各维度每层中数据的具体取值。下层类与上层的某个类有父子关系。(维、层和类别的关系见图 2-1)

3.2.5 多维数组(Multi-dimension Array)

如果一个数据集合可以从多个角度进行观察，即具有多个，则根据这些维度将数据组织所构成的数组，就是多维数组。多维数组是 OLAP 的核心，按其维度的数量，也可称为“数据立方体”或“数据超立方”。多维数组可以用（维 1，维 2，维 3，…，维 n，变量）来表示。

当维度的数量不超过 3 时，采用图形的方法可以很直观地表达出该数组的构造。多维数据浏览器向用户提供了透视 OLAP 立方体中的数据的功能。在浏览器中点击某个单元或维，就可以查看单元属性或维成员属性。它具有的上探和下钻功能使用户在

浏览数据时可以在维层次体系中上下转换[9]。但超过三维的结构，图形方式就无能为力了。对这种情况，可以采用表格组合的方法进行表示。

3.2.6 度量值(Measure)

给定每个维上的一个值，这些值联合起来可以唯一地确定一个数据单元，数据单元中的数据称为度量数据。度量一般可以定量评估业务成效的结果，主要说明数值性问题，如“多少?”。度量完全函数依赖于维，例如：收入为度量值，其完全依赖于时间，地域和产品类型等。

3.3 联机分析(OLAP)的功能

OLAP 系统进行分析，应该能够支持如下功能[10]：

3.3.1 切片和切块(Slice and Dice)

切片就是在某两个维度上取一定区间的维成员或全部维成员，而在其余的维上选定一个维成员的操作。切块可以看成是在切片的基础上，进一步确定各个维成员的区间得到的片段体，即是由多个切片叠合起来。切片和切块是在一部分维上选定值后，关心度量数据在剩余维上的分布。如果剩余的维只有两个，则是切片，否则是切块。

3.3.2 钻取(Drill)

钻取包含向下钻取(Drill-down)和向上钻取(Drill-up)/上卷(Roll-up)操作。钻取是改变维的层次，变换分析的粒度。即在某一维中观察上下层状况。向上钻取是通过一人维的概念分层向上攀升或通过维归约，在数据立方上进行聚集。向上钻取是获得概括的数据。而向下钻取是向上钻取的逆操作，它由不太详细的数据到更详细的数据。它可以通过沿维的概念分层向下或引入新的维来实现。钻取的深度与维度划分的层次相对应。

3.3.3 旋转(Rotate)/转轴(Pivot)

通过旋转可以得到不同视角的数据，改变观察的方向。旋转操作相当于平面数据将坐标轴旋转或变换维度的方向。例如，旋转可能包含了交换行和列，或是把某一个行维移到列维中去，或是把页面现实中的一个维和页面外的维进行交换。

3.3.4 多视图模式(Multi-view Model)

多维分析的展示不仅仅限于简单的数据表，而是采用多种格式图形化界面来显示数据(如图 2-4 所示)。

3.4 联机分析(OLAP)的分类

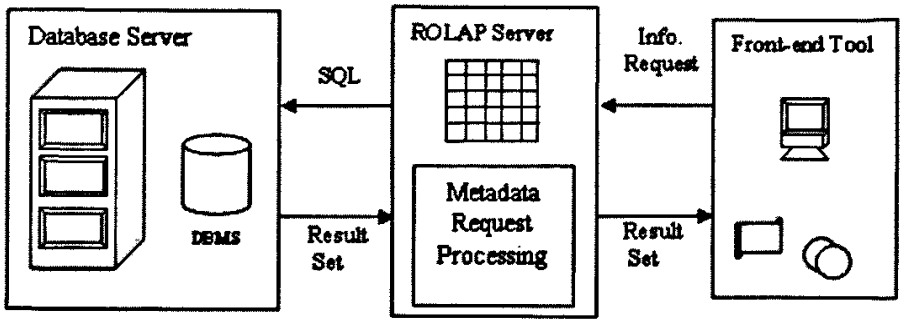
OLAP的技术类型可以分为ROLAP (Relational OLAP)、MOLAP (Multi- dimensional OLAP)两种，这两种类型的功能类似(都可以实现切片/切块、旋转、钻取等多维操作形式)，其不同之处在于底层的数据组织存储形式不同，ROLAP 以关系数据库形式组织数据，MOLAP 以多维数据库的形式组织存储数据。HOLAP 是结合以上两种类型的特点，基于混合数据组织的 OLAP 实现(Hybrid OLAP)。

3.4.1 ROLAP(Relational OLAP)

ROLAP 以关系数据库为核心，用关系型结构的二维表来组织数据，进行多维数据的表示和存储。它并不生成多维立方体，只是存储数据模型与数据仓库之间的映射关系，真正的关系物理的存放在数据仓库中。在进行多维分析时，用户通过客户端工具向 OLAP 服务器提交多维分析请求，OLAP 服务器动态地将这些请求转换成 SQL 语句执行，分析的结果经多维处理转化为多维视图返回给用户。

关系数据库将多维数据库中的多维结构划分为两类表，一个事实表，用来存储度量值及各个维的码值；另一类是维表。事实表是通过每一个维的码值和维表联系在一起的，该结构有时被称为“星型模式”(Star Schema)。数据仓库中的每个主题对应于一个星型模式结构。

ROLAP 结构的主要特点是灵活性强，用户可以动态定义统计或计算方式，不会出现稀疏问题，装载速度较之 MOLAP 要快，另外能保护在已有关系数据库上的投资。但是 ROLAP 方案的实现较为复杂，而且和一般的关系数据库的类似，对用户的分析请求处理会有较长的延迟，它这是因为 ROLAP 结构中数据库里存放了大量的细节数据和相对较少的综合信息，常以牺牲效率为代价动态地生成综合数据。ROLAP 的架构如图 3-1 所示：



ROALP Architecture

图 3.1 ROLAP 架构图

3.4.2 MOLAP(Multi-dimensional OLAP)

MOLAP 表示基于多维数据组织的 OLAP 实现(Multi-dimensional OLAP)。以多维数据组织方式为核心，也就是说，MOLAP 使用多维数组存储数据。MOLAP 利用一个专有的多维数据库来存储 OLAP 分析所需的数据，数据以多维立方体((Cube)方式存储，并以多维视图方式显示。

在 MOLAP 的结构中，关系型数据库中的数据(数据源可以是数据仓库，也可以是从多个 OLTP 系统的数据库)在被存入多维数据库时，将根据它们所属于的维进行一系列的预处理操作(计算和合并)，并把结果按一定的层次结构存入多维数据库中。用户通过客户端的应用软件的界面递交分析需求给 OLAP 服务器，再由 OLAP 服务器检索 MDDB 数据库以得到结果并返回给用户。

MOLAP 方案的主要特点是它能迅速地响应决策分析人员的分析请求并快速地将分析结果返回给用户，这是因为其多维数据库结构以及存储在其中的预处理程度很高的数据，但是，它同时具有缺乏灵活性，数据装载速度慢，且多维数据库可以被看作是一个大的稀疏多维数组，每增加一个维，数据就会剧烈增长。MOLAP 的架构如图 3-2 所示：

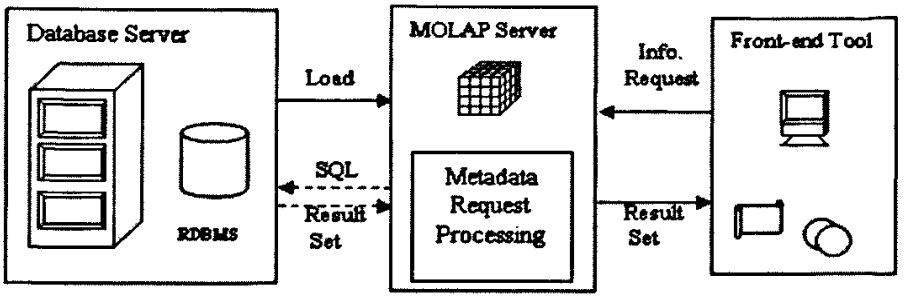


图 3.2 MOLAP 架构图

综合 MOLAP, ROLAP 我们可以得到以下结论：

1)因为 MOLAP 在数据处理方面程度很高,得到的综合数据使得相应用户的分析请求响应速度要比 ROLAP 快；

2)对维、数据动态变化的适应性方面，ROLAP 明显优于 MOLAP,且 ROLAP 技术稳定、成熟，且历史上积累了大量关系型业务数据。自然，建立基于关系的 ROLAP 更切合实际。此外， MOLAP 中产生大量的预处理结果，从而限制了它在处理大量数据的能力。由此可见，二者各有优缺点，将他们的优点结合起来，便出现了一种混合型 OLAP。

3.4.3 HOLAP(Hybrid OLAP)

由于 MOLAP 和 ROLA P 有着各自的优点和缺点，且它们的结构迥然不同，这为分析人员设计 OLAP 结构提出了难题，他们必须在两种结构之间进行选择。为此一个新的混合型 OLAP 结构——HOLAP (Hybrid OLAP)被提出，它能把 MOLAP 和 ROLAP 两种结构的优点结合起来。

HOLAP 表示基于混合数据组织的 OLAP。HOLAP 结构不应该是 MOLAP 与 ROLAP 结构的简单组合，而是这两种结构技术优点的有机结合，能满足用户各种复杂的分析请求。实现 HOLAP 的方法一般有以下两种：

在运行时把对关系型数据库的查询结果存入多维数据库。在这种方法中 HOLAP 系统按一定的先后顺序使用多维数据库和关系型数据库。HOLAP 系统利用开发人员定义一个静态结构的多维模型来暂存运行时检索出的数据。

利用一个多维数据库存储高级别的综合数据，同时，用关系型数据库的 ROLAP 存储结构存储细节数据。这种方法是如今被认为实现 HOLAP 结构较理想的方法，它结合了 MOLAP 和 ROLAP 的优点。HOLAP 的架构如图 3-3 所示：

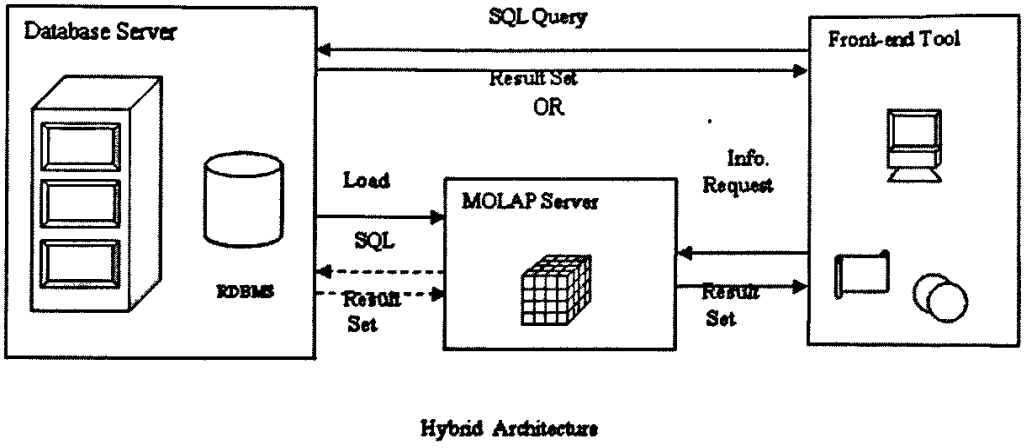


图 3.3 HOLAP 架构图

3.4.4 三种存储模式的比较

MOLAP、ROLAP 和 HOLAP 三种存储模式的比较如下表所示：

表 3.1 三种存储模式的比较

	MOLAP	ROLAP	HOLAP
查询性能	最好	最差	中等
处理性能	很好	最差	最好
额外存储空间	最大	无	最小
Cube 数据	MD	RDB	RDB
聚合数据	MD	RDB	MD

3.5 OLAP 的数据模式

3.5.1 星型模式(Star Schema)

星形模式有两类基本表组成，一个事实表(fact table)和多个维表(dimension table)。事实表包含实际事务或要分析的值，维表包含有关这些事务的描述信息，每

个维表都有一个主键，所有这些维表的主键组成事实表的主键，事实表的主键是各维表的外键，非主健属性称为事实(fact)。事实表存储用户需要查询分析的数据，是被大量载入数据的实体，一般都是数值或其他可以进行计算的数据；而在维表中存储的是不同的层次划分的相应数据，它们大都是文字、时间等类型的数据。通过数据预连接和建立有选择的数据冗余，为访问和分析过程大大简化了数据查询。

星形模式的维表是逆规范的，而事实表则不是。数据集市一般采用星形模式，中央数据仓库由于其成本，可扩展性以及易于管理采用第三范式。

星型模式的优点是：

1. 用户易于理解。
2. 优化浏览。
3. 最适合查询处理。

星型模式的缺点是[11]：

1. 占用的存储空间及数据冗余量较大。
2. 非严格意义上的规范化结构，在数据量比较大时，不易解决业务更新时须增加新维的问题。

3.5.2 雪花模式(Snow flat Schema)

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就称为雪花模式。“雪花化”是一种将星型模式维度表规范化的方法。如果你将所有维度表完全规范化，那么将得到一个以事实表为中心的雪花型结构。

雪花模式的优点是：

1. 减少(很少的)存储空间。
2. 规范化的结构更容易更新和维护。

雪花模式的缺点是[13]：

1. 增加了用户必须处理表的数量，模式比较复杂，操作专业化要求较高。
2. 浏览内容困难。
3. 额外的连接将使查询性能下降。

3.6 OLAP 与 OLTP 的区别

当前的数据库系统应用可主要分为两大类：操作型和分析型。

操作型——应用面向操作人员和低层管理人员，称为联机事务处理 OLTP(On-Line

Transaction Processing)。

分析型——应用面向高层管理人员，称为联机分析处理 OLAP(On-Line Analytical Processing)。

OLAP 概念是由 E.F.Codd 于 1993 年提出的，与 OLTP 相比较，主要存在以下差异：

1. OLTP 处理的对象是大量的事务，每个事务处理只对少量的数据进行存取与更新，查询比较简单，返回数据量少，不适应较大规模的决策支持数据分析；而 OLAP 大量的操作是查询，一般不需要修改数据，返回数据量大，重在分析，为决策支持服务，用户一般不需要很深的 SQL 知识。
2. OLTP 系统是生产系统，对时间响应要求高；而 OLAP 不必实时响应，允许合理的时间延迟。
3. OLTP 数据库只存储当前数据，数据量相对较小，历史数据的作用不大；而 OLAP 数据库存放大量的历史详细数据和当前详细数据，数据量很大。
4. OLTP 使用对象为基层操作人员，为前台柜面业务服务，用户数量大；而 OLAP 主要为业务决策和管理人员服务，用来进行营销策划、产品设计等，用户数量相对较小。
5. OLTP 表非常详细，各种操作一般都基于索引进行；OLAP 是集中的、总结性的、面向时间结构的，操作不能完全基于索引进行，而使用多维数据库技术来提高查询速度，满足查询需要。

3.7 OLAP 与数据挖掘的区别

OLAP 着力处理数据仓库中海量的数据，并将之转化为有用的信息，从而实现对数据的归纳、分析和处理，帮助企业完成决策。它可以帮助管理人员定性分析各种因素对分析目标的影响。

数据挖掘(Data Mining)旨在从海量数据中发现内在隐藏的规律，这是数据库系统应用(包括 OLTP，OLAP，DM)目前最深层次的应用。它可以根据大量的历史数据定量地给出各种因素对分析目标的影响程度。

此外，数据挖掘主要是基于人工智能、机器学习、统计学、数据库等技术。数据挖掘的分析方法主要有关联分析、序列模式分析、分类分析、聚类分析等。数据挖掘利用人工智能领域中一些已经成熟的算法和技术，如：人工神经网络、遗传算法、决策树方法、邻近搜索算法、规则推理、模糊逻辑、公式发现等来进行数据的挖掘，数据挖掘是人工智能中的成熟技术在决策支持系统中的具体应用。

OLAP 和数据挖掘是相辅相成的，但它们的侧重点不同，OLAP 侧重于与用户的交互、快速的响应速度及提供数据的多维视图。数据挖掘则能自动发现隐藏在数据中的

模式和有用信息。OLAP 的分析结果可以给数据挖掘提供分析信息作为挖掘的依据，数据挖掘可以拓展 OLAP 分析的深度，可以发现 OLAP 所不能发现的更为复杂、细致的信息。

因此，可以认为 OLAP 和数据挖掘在功能上是不交叉的。OLAP 是数据汇总、聚集工具，它帮助简化数据分析；而数据挖掘是自动发现隐藏在大量数据中的隐含模式和有趣的知识。OLAP 工具的目标是简化和支持交互数据分析，而数据挖掘的目标是尽可能自动处理，它不是用于验证某个假设的模式的正确性，而是在数据库中自己寻找模型，其本质上是一个归纳的过程，尽管期间允许用户指导这一过程。在这种意义上，数据挖掘比传统的 OLAP 前进了一步。

3.8 联机分析(OLAP)的发展现状

OLAP 的概念自 1993 年提出以来，OLAP 技术的应用得到了长足发展，无论在 OLAP 技术理论方面，还是在应用产品方面都得了重要的成果。

在 OLAP 技术领域，对 OLAP 的理解也在不断深入。如 Nigel Pendse 提出的 FASMI(Fast Analysis Of Shared Multidimensional Information)对 OLAP 做了更为简洁的定义；同时 OLAP 也逐渐结合了一些先进的技术理论，出现了面向对象的联机分析—OO OLAP（Object-Oriented OLAP），对象关系的联机分析 OROLAP（Object Relational OLAP）、分布式联机分析 DOLAP（Distributed OLAP)和时态联机分析处理 TOLAP（Temporal OLAP)等。

在 OLAP 应用产品领域取的两方面的成果：

一方面，提出了 OLAP 规范和标准：随着 OLAP 技术研究的深入和 OLAP 应用的逐渐普及，为了规范 OLAP 系统的建设同时提供对 OLAP 产品进行衡量的标准，一些公司和部门提议成立了 OLAP 的研究协会、制定了 OLAP 的标准，发布了 OLAP 产品的测试报告。如 OLAP Council 制定的 OLAP 标准 APB-1(以及对多维数据产品的研究测试报告 OLAP Report。

另一方面，出现了很多具有 OLAP 特性的应用软件：目前出现的 OLAP 产品，按实现途径可以分为三类：一类是原来的数据库产品生产厂商，在自己的产品种添加了 OLAP 的功能，这包括客户端和服务器端的支持。如 Oracle, Sybase 数据库的，Informix、SQL Server 和 DB2 等传统数据库产品均纷纷推出了新版本支持 OLAP 功能。一类是第三方厂商推出的OLAP产品，这种产品种以OLAP客户端产品居多，如Business Object、 Cognos PowerPlay 和 BrioQuery 等产品，这些产品提供了人性化的人机交互方法，有利于用户的使用。但这些产品通常和服务器端的 OLAP server 结合的不是十分紧密，效率是一个值得关注的问题。

另一类是专用的决策支持数据库厂商推出的OLAP产品，如Teradata，RedBrick等。

3.9 数据仓库及OLAP在通信领域的应用

3.9.1 客户关系管理(CRM)

长期以来，在电信业务的经营过程中，我国电信企业一直遵循“用心服务，用户至上”的宗旨。然而由于没有竞争，电信企业“以客户为中心”的经营原则并没有真正发挥作用，电信企业为客户提供均一化的业务，并不考虑单群或单个客户的特别需要，因为市场上没有其它的电信运营商提供更接近这些用户群的业务。

电信市场放开和竞争加剧的趋势，对电信企业的竞争能力提出更高要求。电信企业一般从三个方面区别于竞争对手并获得竞争优势：价格、业务和客户服务。价格战不能长期使用，业务质量的差异性也随着技术的发展将逐步消失，因此，完善客户服务成为电信企业获取最终竞争优势的重要手段。

但是电信的客户群体非常庞大，而且客户对服务的要求也越来越高，必须有专门的系统对用户进行客户关系管理分析，也就是CRM (Customer Relation Manage)。具体内容包括：利用用户资料和一切可能有助于进行客户分析管理的资料进行客户概况分析，客户忠诚度分析，客户利润分析，客户性能分析，客户未来分析，客户产品分析，客户促销分析：通过对这些数据的分析，提供既能留住老客户又能吸引新客户的决策信息。

根据调查，实施CRM可以对企业带来三个方面的好处：收入的增加：生产力的提高：客户满意度的提高。同时，权威机构的研究也发现在国外实施CRM的企业当中：有三分之一的企业其客户关系管理都没有成功，取得实施前规划中的效益。为什么会有如此多的企业陷入这样困境？分析其原因发现，这些企业没有提供详细的交易数据。没有与客户互动的数据，因此没有办法做好客户关系管理。数据是做好客户关系管理的基础，处理大量的数据有要有一个数据仓库。

数据仓库可以将各个渠道得来的数据，整理成全面、完善的客户信息库。数据仓库内存储有详细的客户轮廓的信息和客户交易行为的历史数据，通过数据挖掘和数据分析，来发现隐藏在数据后面的真实情况，才能了解客户的需求，从而提高企业的收益率和竞争力。因此数据仓库就是CRM的大脑。

3.9.2 话单分析

用户的通话话单包括了用户一次消费行为的所有信息。通过对大量历史话单的分析可以了解用户的消费习惯、消费能力等许多有价值的信息。然而，由于长话话单数量巨大，要在线保存需要占用大量的数据空间，一般电信企业只保留半年的在线话单。这给话单分析带来了麻烦，例如要比较去年五一节假期间与今年五一节假期间的长话变化情况就不太好办，而要比较最近五年的变化情况几乎就不可能。而且，保存的话单数据是为计费需要而设计的，不适应话单分析的要求，比如要统计超长话单和超短话单的地区分布情况和用户分布情况就不好办。自接在计费数据库中进行检索非常缓慢效果又不好，还可能影响正常的业务运行。

其实，一般进行分析、统计关心的是某一类用户的消费行为，而不会细到具体某个号码某次的通话记录。因此，可以事先设定一个粒度，根据粒度对同类话单数据进行汇总并加上一些我们想增添的信息，如主叫性质，整理后加载到数据仓库中。

3.9.3 优惠策略分析

电信企业经常推出各种优惠措施推广新业务和吸引扩大业务量。优惠促销固然可以开拓市场，但如果优惠策略不恰当，结果也可能适得其反。如果利用数据仓库技术建立优惠模型，实现优惠策略的仿真，根据优惠策略进行模拟计费和模拟出账，其仿真结果将提示所制定的优惠策略是否合适，并可按情况进行调整、优化，使优惠策略的出台更有科学依据。在优惠政策实行后，也可以通过数据仓库对优惠效果进行分析、评估，检查优惠效果是否达到。

3.9.4 定制报表生成

电信企业经营分析需要大量的报表。这些报表的数据都自接来源于电信企业的数据库。因此，业务人员只知道需要哪些数据，却不知道如何得到他们，虽然一般在开发软件时开发商一般都会根据企业需求开发出一些固定格式的报表程序来满足企业需要，但电信新业务层出不穷，企业经营瞬息万变，报表种类与报表格式常常要发生变化。这时候，业务人员往往求助于维护人员。但维护人员不一定理解业务人员的想法，也不一定理解数据库的每个细节，因此他给出的统计结果可能会有误差。而且，这样的查询往往是手上进行，表格也手上制作，费时费力，难以管理。

利用数据仓库为自定义生成报表提供了可能。数据仓库为经营者提供多维数据源，业务人员可以根据报表样式通过报表生成系统对数据源进行调整、配置，从而自

己可以定制符合自己需求的报表。

第四章 电信话单分析系统设计

随着电信企业经营环境的变化，市场竞争越来越激烈。如何有效地利用有利的工具提高经营决策水平，成为今天每个电信企业必须面对的问题。数据仓库就是一种提高企业业务分析能力和决策水平的有效工具。业务据调查，在许多引入竞争机制的国家或地区，如美国、英国和日本，电信公司都建立了数据仓库系统作为经营决策的工具，用以提高客户满意度和经营利润。国外专业电信顾问公司研究调查发现，数据仓库和统计分析模型这两项独特能力是竞争优势的来源，是当今电信公司成功的重要原因之一[12]。电信企业一直使用计算机处理各种业务，包括设备维护(网络系统)、业务管理、财务管理等，具有丰富的历史数据，因此建立数据仓库有良好的基础。数据仓库的目的是要建立一种体系化的数据存贮环境，将分析决策所需的大量数据从传统的操作环境中分离出来，使分散的、不一致的操作数据转换成集成的、统一的信息，企业内不同单位的成员都可以在此单一的环境之下，通过运用其中的数据与信息，发现全新的视野和新的问题、新的分析与想法，进而发展出制度化的决策系统，并获取更多经营效益。

数据仓库系统和 OLAP 的建设是一项复杂的系统工程，在设计中会遇到各种各样的技术问题。一个典型的企业数据仓库系统构建过程通常包含数据仓库的设计、数据源与数据存储与管理、OLAP 前端工具四个部分。下面针对数据仓库和 OLAP 技术在电信行业中的应用的几个关键环节进行分析。

4.1 话单分析系统建模需求分析

4.1.1 通信运营分析系统概述

经营分析是企业业务管理的重要组成部分，它运用各种专门的分析方法，对输入到系统中的离散的、单一的经营数据进行进一步的加工处理，从中取得更为有用的信息，为领导决策提供可靠的依据。

电信企业因自身生产经营过程中具备的数据密集性特点，经营服务对原始数据作深层次分析的依赖性，以及电信企业历来具有的经营分析的良好传统，均对当前电信企业经营服务信息提供的实时性、准确性和深度性提出了迫切的需求。将相关的市场经营数据信息化、系统化，并利用统计分析软件建立分析模型，以展现、挖掘或分析

原始信息之间的关系以及更深层次的内容，是电信企业经营分析系统建设的主要目的。为了满足上述需求，经营分析系统的功能应包括经营发展概况、业务分析、客户分析、竞争分析、客户服务分析、专题分析等。

经营分析系统的建设应遵循“整合业务数据、面向经营分析”的原则，“整合业务数据”即工程建设必须构造面向主题的、集成的、稳定的面向经营分析的系统；“面向经营分析”是指系统必须智能地从数据中提取与企业经营相关的信息和知识，为市场经营和决策人员制定业务发展和市场竞争等策略提供科学、准确、及时的依据。从实现手段上看，系统应能通过固定/预定义报表、即席查询、联机分析处理等手段实现面向主题的业务应用；应能根据需要进行主题内部要素的扩充、主题的新增以及跨主题的重构；应能成为业务决策者专业的咨询顾问。

4.1.2 通信运营分析系统功能

从一个全省范围的电信企业市场经营管理职能来看，省公司层面的职能主要是以产品管理和营销决策管理职能为主，地市分公司则以客户营销的执行职能为主。因此，作为电信企业经营分析和营销决策支撑的数据仓库，它的功能是以产品、客户分析为主线，配合竞争、营销活动两条分析辅线，最终形成企业的市场经营概况。

4.1.3 通信运营分析系统数据来源

在数据源方面，数据仓库需采集企业内部有关市场经营活动的客户资料、各类业务计费资料、客户服务资料及其他竞争信息，在有效整合的基础上，分主题实现对经营分析、客户营销决策支撑。因此，需采集的数据源应包括客户资料、计费账务信息、客户服务信息、网间结算信息及其他信息，如网管、资源管理、统计、计财等报表，以及外部社会经营环境、竞争对手信息。

4.2 通信运营分析系统结构及实现

4.2.1 主题域的确定

电信企业建立数据仓库，目的主要是为企业的市场经营管理和营销决策提供数据分析支持，因此，系统的分析主题设计应围绕电信企业的市场经营、营销活动的构成对象和任务来进行。影响电信企业市场经营能力(或竞争能力)的几大因素是企业经营的业务或产品、企业向市场提供这些业务或产品的方式(营销活动)、企业目前所拥有的客户、现有竞争对手，因此，产品、客户、竞争对手和营销活动即是我们数据仓库

所要立足的分析对象，缺一不可。

确定了分析对象之后，还需根据企业经营管理或营销组织的实际需要将对象进一步细分，比如电信企业将客户分为大客户、商业客户、公众客户和流动客户来管理，这就需要将客户分析的主题落到每个客户群上，而且业务或产品的分析也一样需要进一步细分到各专业。

细化了分析对象后就进入分析主题内容设计阶段。这个阶段根据细化的分析对象来设计数据分析的内容，即总结和归纳企业市场经营分析人员和营销分析人员现在的数据分析上作，以更有效率地组织分析数据。

根据经验，各类分析对象的分析主题可以设计如下：

1. 业务或产品的分析主题包括：各类业务或产品发展状况分析、发展变化趋势分析、影响因素分析以及发展预测等内容；
2. 客户分析主题包括：客户价值分析、客户流失分析、客户忠诚度分析、客户信用度分析等内容；
3. 竞争分析基于网间的话单信息来设计，包括：竞争对手用户发展情况、本企业用户使用竞争对手产品情况和竞争对手用户使用本企业产品情况等内容。
4. 营销活动分析则根据营销活动的三大目的：获取客户、提高 ARPU、客户保持以及营销活动的三个环节—营销策划、营销执行和营销评估来设计相应分析内容，一般包括：营销机会判断、预期效果评估、营销效果评估、营销方案调整等内容。

4.2.2 维度划分

数据仓库中各主题的维度是为多维分析和定制报表而设计的，同时也要将报表数据分析过程中所经常要用到的分组组别考虑进来。设计维度时要强调有用性和效率的均衡，既要涵盖今后数据分析常用的角度，同时也要考虑到多加一个维度或维度值就意味着仓库中数据量的成千上万倍增长，所以必须考虑效率问题。另外，在设计每个维度的维度值时，要强调独立性和系统性。对于某个分析对象来说，每个维度的所有维度值之间是独立的，不能有交叉。数据仓库的维度可以分为以下六大类：

1. 时间维度和空间维度；
2. 业务维度：包括业务种类、流向、拨打方式、通达方式、速率等维度；
3. 客户维度：包括渠道属性、统计属性、入网时间、客户状态、城乡属性、服务等级、行业属性、计费类别等维度；
4. 用户终端(设备)维度：接入方式、终端类型等维度；
5. 营销活动维度：参加活动种类、参加活动时间等维度；

6. 运营商维度：运营商种类等维度。

4.3 数据仓库的设计步骤

数据仓库的系统设计是数据驱动的，是一个不断循环、反馈而使系统不断增长与完善的过程。

数据仓库系统设计的主要步骤如下[14]：

1. 概念模型的设计(Conceptive Model Design)：主要是要界定系统边界，确定主题域及其相关内容。常用的方法是 E-R 分析法。
2. 逻辑模型的设计(Logical Model Design)：在确定主题域后，需要对主题包含的信息进行详细定义，并对事实表和维表的关系详细定义。常用的模式是星形模式(Star Schema)。
3. 物理模型的设计(Physical Model Design)：主要考虑数据的存储方式，使得系统具有较好的性能。对于记录庞大的事实表，可以考虑分区存放。而记录很少的维表则可以集中存放于某一表空间，甚至可以让其数据在首次读取时驻留在系统内存中，以加快数据存取速度。索引的建立也在该阶段中完成。

4.4 数据仓库的概念模型设计

概念模型设计，主要采用实体关系图(ERD)，用实体以及实体之间的关系来描述。ERD 的实体主要包括指标实体(事实实体)和维度实体。

根据电信经营分析系统的需求，确定数据仓库的五个基本主题：客户主题、产品主题、话务量主题、账务主题、市场营销主题。

话务量是近期国内电信企业争夺的焦点所在。由于客户数量相对有限，再加上市场饱和的原因，使得争夺话务量成为企业规模成长的主要手段。所以对于话务量主题的建设，也是电信企业经营分析项目建设的重点所在。在建设话务量主题的过程中，我们应该牢牢抓住“有利于竞争”这个思路，力争通过话务量主题分析支撑各种针锋相对的竞争活动，用于与竞争对手的抗衡。

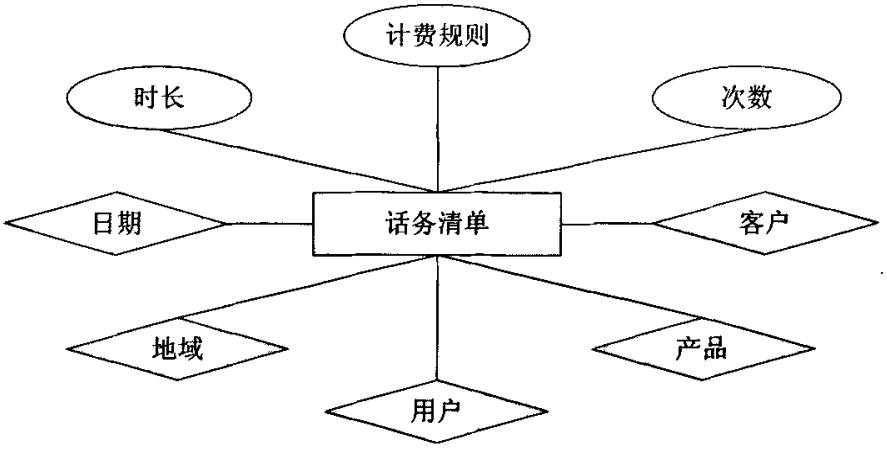


图 4.1 话单分析星型模型

如上图，话务量主题的指标实体为话务清单：话务量主要体现客户对产品的使用情况，它的相关信息包括计费规则、次数、时长等，这些信息构成星型模型的维实体；日期和地域作为公共维度，与其他模型统一。用户作为一个指标实体，将清单与产品和客户关联起来；产品指标实体作为清单产生主体，与清单紧密联系。客户指标实体描述清单归属。通过指标实体和维度实体，最终构成话务量主题概念模型。

4.5 数据仓库的逻辑模型设计

话务量主题主要包括的实体如下：

1. 话务量

主要监视时间、地域、客户维度下的话务量表现。话务量的数据非常重要，反映着市场的瞬息变化话务量的指标包括：次数(本期值/上期值/同期值)、时长(本期值/上期值/同期值…)。

2. 话务清单

话务清单主要用于对某个特定客户(有时甚至是某一个具体的产品接入号码)进行话务习惯的考察。例如来话、去话的规律，从而可以针对性地为客户“量身定做”最适合其需要的 SLA(国际通行电信服务评估标准)协议产品，在该协议产品中可以推荐一些套餐或者制定适宜的资费政策[15]。

3. 竞争对手情况

该实体主要适用于对涉及多个运营商，也就是存在竞争对手情况下的话务量的分析。在这里特别是需要关注运营商产品维度，考察的指标主要是次数和时长。对该实

体的关注，既使得电信企业可以进行重点客户的策反，也可以用来分析与竞争对手同类产品的关系，从而可以用来对自身的产品(含资费)进行调整和改良。

4.6 数据仓库的物理模型设计

为了达到分析用户通话特性的目的，以客户通话事件作为事实表的数据用星型模型建立数据集市，同时作为多维数据集的维度有数张维度表，如下：

表 4.1 FACT_PHONE_TIKEC 用户通话事实表

字段名称	字段含义	类型	字段说明	是否为空
AREA_ID(FK)	指向维表 DIM_AEA	NUMBER(8)	外主键	NOT NULL
SERV_ID(FK)	指向维表 DIM_SERV	NUMBER(8)	外主键	NOT NULL
TIME_ID(FK)	指向维表 DIM_TIME	NUMBER(8)	外主键	NOT NULL
CUST_ID(FK)	指向维表 DIM_CUST	NUMBER(8)	外主键	NOT NULL
DURATION	通话计时(秒)	NUMBER(8)	度量值	NOT NULL
COUNT	通话计数	NUMBER(8)	度量值	NOT NULL
CALLED_AREA_CODE	对方区号	VARCHAR2(4)	度量值	NOT NULL
CALLED_NUM	对方电话	VARCHAR2(8)	度量值	NOT NULL

表 4.2 DIM_SERV 设备维度表

字段名称	字段含义	类型	字段说明	是否为空
SERV_ID(FK)	设备 ID	NUMBER(8)	主键	NOT NULL
PHONE_NUM	电话号码	VARCHAR2(8)	度量值	NOT NULL
EXCHANGE_ID	局号	VARCHAR2(6)	度量值	NOT NULL
SERV_TYPE	设备类型	VARCHAR2(32)	度量值	NOT NULL
BILLING_TYPE	计费类型	VARCHAR2(32)	度量值	NOT NULL

表 4.3 DIM_CUST 客户维度表

字段名称	字段含义	类型	字段说明	是否为空
CUST_ID(PK)	客户 ID	NUMBER(8)	主键	NOT NULL
ADDRESS	客户地址	VARCHAR2(64)	度量值	NOT NULL
NAME	客户姓名	VARCHAR2(32)	度量值	NOT NULL
IDCARD	身份证号码	VARCHAR2(32)	度量值	NOT NULL
VIP_FLAG	VIP 标识	NUMBER(2)	度量值	NOT NULL
CREDIT_LEVEL	信用度	NUMBER(4)	度量值	NOT NULL
CREDIT_LEVEL	用户类型	VARCHAR2(32)	度量值	NOT NULL

表 4.4 DIM_AREA 地区维度表

字段名称	字段含义	类型	字段说明	是否为空
AREA_ID(PK)	地区 ID	NUMBER(8)	主键	NOT NULL
AREA_NAME	地址名称	VARCHAR2(64)	度量值	NOT NULL
AREA_CODE	区号	VARCHAR2(4)	度量值	NOT NULL

表 4.5 DIM_TIME 时间维度表

字段名称	字段含义	类型	字段说明	是否为空
TIME_ID(PK)	时间 ID	NUMBER(8)	主键	NOT NULL
YEAR	年	NUMBER(4)	度量值	NOT NULL
MONTH	月份	NUMBER(4))	度量值	NOT NULL
DAY	天	NUMBER(2)	度量值	NOT NULL
HOUR	小时	NUMBER(2)	度量值	NOT NULL
MINUTE	分钟	NUMBER(2)	度量值	NOT NULL
TIME_VAL	时间值	DATE	度量值	NOT NULL

事实表的度量包括通话计时、通话计次、被叫号码等，而维度表的属性有客户类型、设备类型、时间和地区属性等。

第五章 电信话单分析系统实现

5.1 维表和事实表的数据填充

5.1.1 维的生成

以时间维为例，时间维的作用是在分析数据时提供多个时间层次的观察方式，其层次结构主要取决于分析应用的需求，而时间范围取决于分析口标的有效时间范围。以下语句利用 ORACLE PL/SQL 脚本和 Oracle 日期函数在 DIM_TIME 表生成了全年的时间维数据。

```
DECLARE
TMBEGIN DATE;
TMEND DATE;
TMP_ID NUMBER;
BEGIN
TMBEGIN := TO_DATE('2004-01-01 00:00:00','YYYY-MM-DD HH24:MI:SS');
TMEND    := TO_DATE('2004-12-31 23:59:59','YYYY-MM-DD HH24:MI:SS');

WHILE TMBEGIN<TMEND
LOOP
    SELECT TIME_SQ.NEXTVAL INTO TMP_ID FROM DUAL ;
    INSERT INTO DIM_TIME VALUES
    (TMP_ID,TO_NUMBER(TO_CHAR(TMBEGIN,'YYYY')),
            TO_NUMBER(TO_CHAR(TMBEGIN,'MM')),
            TO_NUMBER(TO_CHAR(TMBEGIN,'DD')),
            TO_NUMBER(TO_CHAR(TMBEGIN,'HH24')),
            TO_NUMBER(TO_CHAR(TMBEGIN,'MI')),
            TO_CHAR(TMBEGIN,'DAY'),
            TMBEGIN);
    TMBEGIN := TMBEGIN+ 1/(24*60);
END LOOP;
COMMIT;
END;
```

5.1.2 维度表数据的提取和插入

原生产系统的业务数据将有效数据和历史变更数据放在同一在表内，在比如在SERV 设备维度表数据发生变化时，采取原表中插入一条新纪录，并且采用需要序号累进的方式标示最新的有效数据。其情况同样存在于 ACCT 表、SERV_ACCT 表、以及 CUST 表中，这种处理方式导致运营数据库的数据量非常巨大(上述几张表记录数均在20,000,000 条以上)，在这几张表中做大量数据的关联查询并插入将会相当耗时。在将业务数据导入维表时，采取的方式是先对现有数据进行过滤清洗，将有效数据先导入临时表，再做关联查询和插入维表。在数据仓库建立完成后，每天定时对比保存有效数据的临时表和当前运营数据库，再将获得的增量变更数据用于更新维表数据，藉此来实现对运营数据库只做查询，而每次只更新临时表的数据以及维表中有变化的数据。实际应用中该方法能获得较好的执行性能。

5.1.3 事实表数据的生成

事实表中的度量数据来自通话明细单表，而主外键列的值通过电话号码的关联查询获得。由于明细单表中有记载事件的事务处理时间，故可根据时间戳实现数据的增量更新，具体实现中通过带时间参数的定时执行存储过程来实现自动更新。以下语句为装载数据的 SQL 语句。

```
INSERT INTO FACT_PHONE_TICKET
SELECT
DIM_CUST.CUST_ID,DIM_SERV.SERV_ID,
DIM_TIME.TIME_ID,DIM_AREA.AREA_ID,ET.DURATION,ET.COUNTS,ET.CA
LLED_AREA_CODE,ET.CALLED_NBR
FROM
EXTEND_TICKET
ET,DIM SERV,DIM TIME ,DIM CUST,DIM AREA,ACCT,SERV ACCT
WHERE
DIM_SERV.PHONE_NUMBER = ET.CALLING_NBR
AND DIM_SERV.SERV_ID = SERV_ACCT.SERV_ID
AND SERV_ACCT.ACCT_ID = ACCT.ACCT_ID
AND ACCT.CUST_ID = DIM_CUST.CUST_ID
AND DIM_AREA.AREA_ID = ACCT.AREA_ID
AND SERV_ACCT.STATE = 'G0A'
AND ACCT.STATE = '10A'
AND
to_char(ET.START_TIME,'YYYYMMDDHH24MI')
=to_char(DIM_TIME.THE_TIME,'YYYYMMDDHH24MI')
```

5.2 数据分析(OLAP)方案的设计与实现

5.2.1 OLAP 分析工具的介绍

生成了事实表和维表以后，就可以有了对通话明细表单进行多维分析的基础。但要使得分析者能够实际的操作多维数据集中的数据，还必须选择一个强大的 OLAP 前端工具。Microsoft 的 Office Web 组件是用于向 Web 页添加电子表格、图表和数据处理功能的 ActiveX 件的集合。其中的 PivotTable 组件实现了在使用 Microsoft Internet Explorer 浏览包含 Office Web 组件的 Web 页时，用户可以自接在 Internet Explorer 中处理显示的数据，如对数据进行排序和筛选，输入新的数值，展开和折叠明细数据，进行行列旋转以查看源数据的不同汇总信息等。而且数据透视表列表中的源数据可以来自 OLE DB 或 JDBC 数据源，该组件本身提供标准的接口，可以方便的集成到应用程序中。

5.2.2 OLAP 前端展示开发及相关开源产品

系统基于 J2EE（Java 2 Enterprise Edition）架构实现，可以部署在通用的 J2EE 应用服务器上，后台数据库支持所有能提供 JDBC（Java Database Connectivity）接口的关系型数据库。前端提供支持 OLAP 服务的 JSP 标记接口（JSP Tag Libraries）库、JOLAP（Java OLAP）标准接口和 XMLA（XML for Analysis）标准的接口。

整个系统的开发，利用了大量的开放源码软件，这些软件位于 OLAP 系统框架的不同位置，有些属于工具性的包，有的则直接属于应用层软件。表 5.1 列出了该系统中用到的主要开源软件。

表 5.1 项目中使用的开源产品清单

名称	功能	下载地址
Mondrian	MDX 分析引擎	Sourceforge.net
JPivot	呈现 OLAP 表格的 JSP 自定义标签库	Sourceforge.net
BeanUtil	操作动态操作 JavaBean 的工具	Jakarta.apache.org
Digester	将 XML 文件转换成为任意的 Java 对象，并提供灵活的扩展接口	Jakarta.apache.org
Log4j	日志管理	Jakarta.apache.org
Regexp	正则表达式处理包	Jakarta.apache.org

JFreechart	基于 Java 的图表生成器	www. jfree. org
Xalan-J	将 XML 文档转换成 HTML 等格式文档，遵循 XSLT 和 XPath 规范	Xml. apache. org
Xerces	XML 解析器	Xml. apache. org

5.2.3 OLAP 多维分析实现

在系统实现中，将累计通话时间作为重点观察的度量对象，将多维数据表中的行层次设计为“年－星期－小时”，将列层次设计为“设备类别－客户类别”，目的是考察不同的用户在不同的时间段的通话特性。

1. 功能概述：
   根据各种维度组合分析消费客户群的通话时长及所占比例情况。
2. 分析维度选择：
   时间维（年、周、小时）
   服务设备维(国内外、通话种类)
3. 分析测度选择：
   通话时长、所占比例

			SERV TYPE / CUST TYPE								
			⊞普通电话			⊞无线市话			总计		
YEAR	WEEK	HOUR	通话计时	列白分比	行白分比	通话计时	列白分比	行白分比	通话计时	列白分比	行白分比
⊟2004	⊞星期日		263995	8.48%	64.18%	147322	12.29%	35.82%	411317	9.54%	100.00%
	⊞星期一		690378	22.17%	77.98%	195000	16.27%	22.02%	885378	20.53%	100.00%
	⊞星期二		487891	15.03%	71.70%	184658	15.41%	28.30%	652549	15.13%	100.00%
	⊞星期三		505304	16.23%	75.25%	166191	13.87%	24.75%	671495	15.57%	100.00%
	⊞星期四		495768	15.92%	73.03%	183110	15.28%	26.97%	678878	15.74%	100.00%
	⊞星期五		455892	14.64%	72.46%	173285	14.46%	27.54%	629177	14.59%	100.00%
	⊞星期六		234588	7.53%	61.20%	148709	12.41%	38.80%	383295	8.89%	100.00%
	汇总		3113814	100.00%	72.21%	1198275	100.00%	27.79%	4312089	100.00%	100.00%
总计			3113814	100.00%	72.21%	1198275	100.00%	27.79%	4312089	100.00%	100.00%

图 5.1 汇总的数据

首先对高度汇总的数据进行观察，如上图，从这个层次的表中已经可以看出一些基本的通话特性，即：作为办公用途较多的设备普通电话(固话)在周末时段的通话量明显少于上作时间段，而且在上作日的第一天(周一)，通话量汇总值达到高峰(22.17%)：而作为个人用途较多的无线市话(小灵通)在一周的 7 天内，通话量汇总值的变化都不是很大。可以看出的另一个统计信息是就通话时间，当分析人员想要知道在固定电话通话量最大的周一，到底是哪些具体时间段通话最为频繁时，就可以将通话次数加到要观察的度量值当中，并对日期行进行下钻操作，获得如图 5.2 的展示结果。

			SERV TYPE / CUST TYPE					
			普通电话			无线市话		
YEAR	WEEK	HOUR	通话计时	通话次数	列白分比	通话计时	通话次数	列白分比
2004	星期日		263995	5740	8.48%	147322	1775	12.29%
	星期一	0	4625	85	.15%	4618	22	.40%
		1	1528	34	.05%	654	8	.05%
		2	2755	26	.09%	229	4	.02%
		3	5907	3	.19%	167	4	.01%
		4	27	2	.00%	180	5	.02%
		5	4986	2	.16%	194	5	.02%
		6	5322	8	.17%	622	16	.05%
		7	5222	172	.17%	1530	36	.13%
		8	51382	1175	1.65%	7124	154	.59%
		9	80232	1672	2.58%	7656	164	.64%
		10	80601	1610	2.59%	10677	181	.89%
		11	76837	1505	2.47%	9500	175	.79%
		12	28025	475	.90%	8904	170	.74%
		13	15584	234	.50%	8039	86	.67%
		14	24090	535	.77%	8885	138	.74%
		15	74763	1350	2.40%	10043	163	.84%
		16	65325	1232	2.10%	11302	182	.94%
		17	61490	1092	1.97%	13903	234	1.16%
		18	22750	399	.73%	16252	249	1.36%
		19	17388	255	.56%	13267	138	1.11%
		20	17668	252	.57%	18261	106	1.52%
		21	24528	275	.79%	13062	88	1.09%
		22	11098	169	.36%	14691	68	1.23%
		23	8287	137	.27%	15040	44	1.26%
		汇总	690378	12719	22.17%	195000	2440	16.27%
	星期二		467891	9465	15.03%	184658	2197	15.41%
	星期三		505304	9798	16.23%	166191	2037	13.87%
	星期四		495768	9353	15 92%	183110	2136	15.28%
	星期五		455892	9384	14.64%	173285	2374	14.46%
	星期六		234586	5176	7.53%	148709	1705	12.41%
	汇总		3113814	61635	100.00%	1198275	14684	100.00%
总计			3113814	61635	100.00%	1198275	14684	100.00%

图 5.2 日期钻取数据

分析人员从图 5.2 可以得知通话量在星期一急剧增加的原因确实是上作时间业务电话量的增多而造成的(通话量集中在上午 8 点至 12 点和下午 3 点至 5 点)。同时从小灵通的通话计时百分比在一天 24 小时中的分布(下午 5 点以后通话量占多数)也可以证实另一个事实：即小灵通用户的通话行为以个人用途为主。

第六章 总结

本系统基于数据仓库和联机分析处理(OLAP)技术，借鉴 J2EE 核心模式的思想，在系统地分析了通信领域业务需求的基础上，本着理论与实践相结合的方针，采用话单海量历史数据为实验环境，着重研究了数据仓库和 OLAP 方面的相关技术、设计理念及实现方法，详细阐述了数据仓库模型设计和维度建模的过程，并专门探讨了 OLAP 前端展现工具的实现机制，完成了基于数据仓库的分析系统研究过程，在数据仓库的设计、数据存储、多维分析的响应能力和结果的显示等方面均取得了较好效果。

通过本文研究，得出要成功地实施经营分析系统的建设，关键在于如何将电信企业经营分析系统业务需求，与最新数据仓库建模技术结合起来，构建一个适合国内电信运营商的企业级数据仓库模型，并使其既能够满足现有的业务需要，又具备足够的延展性。

基于对数据仓库相关理论以及电信企业经营分析活动的研究，针对电信业务的特点，本文提出了一套切实可行的方法、准则和指导思想，建立了电信企业、经营分析系统的数据仓库模型。

提出了适合省级电信企业的数据仓库分布方式，用以帮助电信企业实现其经营分析系统的数据仓库建模。

构建的数据仓库模型成功应用的实例，从实践的角度进一步体现了，完善的数据仓库模型对电信企业的经营分析系统建设的功能与效用，促使我国电信企业认识到，竞争优势的提升，不仅取决于经营分析系统的建设，完善的数据仓库模型的构建也是一个重要的因素。

在本文中，作者只是对数据仓库建模的实现进行了初步的探讨，仍然还有更深入的问题需要进一步解决:

1. 电信企业业务分析需求深入研究问题。随着电信竞争的加剧及电信业务的扩展，对业务分析的需求随着企业管理水平的提高，将要求数据仓库提供更深层次的支撑和内涵，也有待进一步的深入研究和实践。

2. 数据挖掘问题。实现经营型数据仓库分析系统转向决策支持型数据仓库，为企业的各项预测、决策提供支撑。要求不仅仅对所有业务数据进行综合及提炼，还需要更进一步研究如何对数据进行成功的挖掘。

3. 优化系统效率问题，比如数据的 ETL 过程，目前只能依赖手工编写 Oracle SQL Loader 脚本匹配维表字段的方式完成。我将在今后实践工作中继续探索，争取寻求更

好的解决方案。

数据仓库和 OLAP 新技术虽然为决策支持系统开辟了新的途径，但是，查询是数据的业务操作，发现才是数据的商业价值。只有数据仓库和 OLAP 是不够的，必须在它们的基础上建立数据挖掘才能真正驾御自己的数据。数据仓库+OLAP+数据挖掘的有机结合，必将在未来发挥巨大的作用。

致谢

本论文是在我的导师逄焕利副教授的悉心指导下完成的。在攻读硕士学位期间，导师精湛渊博的专业知识、平易近人的学者风范和严谨敬业的治学态度给我留下了深刻的印象，是使我受益终生的宝贵财富。在此论文行将完成之际，谨对导师近三年来的悉心栽培和关怀表达我最衷心的感谢！

此外，我要特别感谢胡明教授和许建潮教授。二位老师对于学科前沿课题的敏锐洞察，卓越的领导才能，宽以待人严于律己的高尚品德给我留下了深刻的印象，是我终生学习的榜样。

感谢实验室全体老师和同学，特别是刘钢老师、陈志雨老师、徐立昕老师和李新学长，曾给予我无私的关怀和指导，你们是我的人生的良师益友，你们的谆谆教导我将永远铭记于心。

感谢大连信雅达公司的张胜杰副总经理、软件事业部李征部长在我论文的研究期间提供了良好了项目实践机会和便利的生活条件。

感谢宁静峰、柳学铮、胥旭、姜朔、藏红岩和闵聚等几位师兄和同学，在上海实习期间，与你们在一起生活与学习的快乐点滴，是我人生的精彩回忆。

感谢谷钰、刘鑫、陈月、杨薇和李娜等几位同学，你们对学习和工作的热情与执着令我感动，鞭策我不断前行。

感谢参加论文评审和答辩的各位专家。

最后，我还要感谢我的父母长期以来对我的支持，他们永远是我人生前进的动力。

长春工业大学是我的母校，在计算机科学与工程学院学习期间，我的人生观与价值观有了重新的定位，我由一个懵懂的学子成长为一个有志报效祖国的青年，走上工作岗位之后，我将继续遵循母校的教诲，勤奋学习，努力工作，用我的出色的工作业绩和成果来回馈社会，让母校以我为自豪！

参考文献

[1] [美]W.H. Inmon 著,王志海,林友芳等译.数据仓库,第三版.机械工业出版社,2003年3月,62-75

[2] 蒋力三,陈文俊.如何构建电信运营业的企业信息化.电信科学,2003年第一期

[3] QB/CU 013-2004.中国联通 UNI-CRM 系统指南

[4] Mark Sweiger,Mark R.Madsen 等著.邓昌辉,张光剑等译.点击流数据仓库.电子工业出版社,2004.1

[5] 王曙燕,耿国华,周明全.数据仓库与数据挖掘技术的研究与应用.计算机应用研究,2005年第9期,194-195

[6] 王能斌.数据库系统.北京,电子工业出版社,1995

[7] 王珊.数据仓库技术与联机分析处理,北京:科学出版社,1998

[8] 刘夫涛.从 OLAP 数据挖掘 OLAM,http://www.sqlmine.com/warehouse/htm/6.htm

[9] [美]Inmon.W.H.著,王志海译.数据仓库,北京,机械工业出版社,2000.20-22

[10] 马建红,王万森.基于数据仓库的保险管理系统的设计与实现.微机发展,2004,14(7),55-58

[11]PaulGray, HughJ.Watson. Present and Future Directions in Data Warehousing [J].The Database for Advanced in Information System 1998 Vol.29,No.3

[12] 李建中,高宏.一种数据仓库的多维数据模型.软件学报,2000,11,908-917

[13] 段云峰等著,数据仓库及其在电信领域中的应用,电子工业出版社,2003

[14] (美)Ralph Kimball 著,谭明金译,数据仓库工具箱维度建模完全指南,第二版.电子工业出版社,2003

[15] 赵越.基于数据仓库和 OLAP 的移动通信决策支持系统系统建设,2002 年第 13 期,60

攻读硕士学位期间研究成果

[1] 基于 XML 的 WEB 半结构化信息抽取，长春理工大学学报，2007 年第 1 期，第一作者

[2] 基于序列模式的 WEB 日志挖掘系统，微型计算机，2007 年上半年增刊，第二作者

OLAP技术研究及其在移动通信运营中的应用

作者： 李哲琦
学位授予单位： 长春工业大学

相似文献(10条)

1.学位论文 江涛 电信企业数据仓库Web服务的设计与实现 2006

随着数据仓库技术的发展，很多电信企业都已经成功实施了数据仓库系统。电信企业的数据仓库系统已经成为企业进行决策分析的重要工具，电信企业内部的其它系统甚至电信企业外部系统也开始有访问数据仓库系统应用的需求。问题也就随之产生，由于企业内部系统异构性、紧耦合性的特点，系统间的访问非常困难，SOA的出现恰恰解决了这个问题，它使用WebServices技术有效的封装了应用实现的细节，通过一系列的标准协议开发出与平台和编程语言无关的Web服务，从而降低了应用系统的耦合性并充分利用了现有的资源。

本论文围绕数据仓库对外提供Web服务展开。首先，概要的介绍了数据仓库技术，并结合电信领域实际的数据仓库系统进行了应用分析，总结出目前需要向外界提供服务的为报表和OLAP。然后，详细研究了Web服务技术和面向服务架构，明确了Web服务的定义、实现方式以及面向服务架构与Web服务之间的关系。接着，根据对Web服务和面向服务架构的研究和数据仓库的应用分析，设计出了基于面向服务架构的数据仓库系统Web服务解决方案并根据解决方案进行了系统的实现。该解决方案主要包括Web服务包装规范的设计和Web服务注册／发现系统的设计两个部分。Web服务包装规范设计是本文的一个创新点，它包括报表和OLAP的Web服务包装规范，作者将公共仓库元模型规范中对报表和OLAP元模型的定义引入到了规范的设计中，它与描述Web服务的WSDL规范相结合，根据元模型中定义的类以及类之间的关系，定义出包装报表和OLAPWeb服务所应该定义的数据类型、消息以及操作，这种基于已有标准的设计方式使得Web服务包装规范更具规范性和通用性，包装出的服务也更容易理解。Web服务注册／发现系统的设计依据面向服务架构，该系统集成了面向服务架构中服务注册者的角色。它的用户认证功能、Web服务查找功能、Web服务注册功能以及Web服务集成功能为数据仓库Web服务提供了基础性平台。论文最后对全文作了系统的总结，并提出下一步需要进行的一些研究工作。

2.学位论文 丁建华 数据仓库技术在电信企业中的应用 2006

数据仓库技术(DaraWarehouse)起源于对大量数据进行处理的需要，是随着业务应用的需要产生的。与传统的数据库技术相比，数据仓库为决策分析提供了更好的支持，跳出了传统联机事务处理的范畴。

数据仓库的建设目标之一，是采集企业内部生产管理系统所有市场经营相关的数据源，包括客户背景资料、产品或套餐购买行为、消费资料、客服交互行为、缴费行为等方面的信息，对其进行规范和整合，然后按业务、客户、竞争、营销活动及数据挖掘等主题，将数据按数据集市的形式存放，并提供多维报表和挖掘工具，为分析人员提供统一的数据平台和分析平台，解决此前分析人员所面对的数据分散、口径不统一、分析工作缺乏延续性等问题。数据仓库另一个建设目的是提供营销决策支撑，即在客户级数据查询或挖掘的基础上，将符合某种条件或具备某种特征的客户(用户)清单下发到各营销渠道，为客户经理执行针对性营销策略提供决策依据。

随着国内电信运营商间的竞争愈加激烈，电信企业迫切需要提高企业内部的科学决策能力，增强在市场经营等方面的判断能力，因此，电信运营商需要数据仓库技术。另一方面，电信运营商积累了大量的业务运营数据，通过数据仓库技术，可以从这些用户数据中发现很多有价值的信息，例如用户的消费行为分析特征等。

论文详细地分析了数据仓库的基本概念，以及当前研究的主要问题和一些研究成果，分析了数据仓库技术发展趋势，以及在苏州电信中的应用，详细介绍了苏州电信利用数据仓库技术建设数据集市系统的设计情况、技术思路、解决方法等，同时根据苏州电信的实际提出了一种中等电信业务规模的数据仓库解决方法和以电信企业主题分析的新思路以及在苏州电信的应用情况。

3.学位论文 严璐 电信企业数据仓库信息建模完备性的研究与实践 2004

电信行业各大运营商为了提高自身竞争力,建设决策支持系统,纷纷要求沉淀企业经营信息,积极构建企业级数据仓库,数据仓库技术在电信行业的应用越来越广泛.数据仓库的信息模型指导着数据仓库构建的整个过程,信息模型的好坏直接关系数据仓库建设的成败.由于电信行业有业务数据种类繁多、用户经营分析需求广泛多变的特点,因此,电信行业的企业级数据仓库需要构建具有信息完备性的信息模型.该文在构建电信企业信息完备的数据仓库的背景下,主要研究了完备性信息建模技术在电信行业中的应用.阐述了数据仓库体系结构和数据仓库信息建模方法,同时介绍了作者参加的中国联通统一经营信息服务系统项目的信息现状,分析了产生现状的原因.针对项目实施中的信息现状,作者提出了保证电信企业数据仓库信息模型完备性的方法,并将该方法应用到中国联通统一经营信息服务系统,保证数据仓库操作数据存储层信息模型的信息比较完备.最后,作者对所作的研究工作进行了总结,指出信息完备性研究中的一些成果,并对下一步的研究工作提出了一些看法.

4.学位论文 崔建波 电信企业数据仓库信息模型的设计与应用 2007

近年来随着数据仓库与数据挖掘技术的发展，国内各大电信企业均已经构建自己的企业数据仓库。其中，中国移动和中国联通两个具有移动牌照的运营商，均已进行了大约5年左右的建设，在商业智能方面取得了进步。数据仓库项目的开发模式是属于螺旋迭代开发模式。随着历史数据的逐步沉淀，用户需求的变化，仓库上应用的不断增加，以及数据库、存储硬件等相关技术的进步，企业数据仓库建设需要不断革新、进步。其中信息模型设计的演进是核心内容。

在本论文中，作者根据电信企业的数据特点，结合数据仓库技术、存储技术、计算技术的发展，综合考虑应用不断扩展变化、用户对系统响应效率的要求，设计并实现优化的企业数据仓库信息模型，并通过实践，给出电信企业数据仓库建设与建模的适用方法。

最后，作者对所作的研究工作进行了总结，指出在实践过程中面临的主要问题和困难，并对进一步的研究工作提出了一些建议。

5.期刊论文 胡天濡.王新娜 电信企业数据仓库的设计和应用 -硅谷2009,""(7)

为了更有效地在数据仓库的摹础上开展数据挖掘工作,首先要总结数据挖掘项目所需的客户层面的有关客户(用户)背景、购买行为等信息,从数据仓库中定期抽取,形成数据挖掘集市;然后分主题地建立包括流失预警、客户细分、交叉销售、营销预演等模型,各类模型模板化后封装至数据仓库,建立数据挖掘模型模板库.

6.学位论文 金羿 基于数据仓库的电信企业精确化营销策略探讨 2005

本论文研究了基于数据仓库和数据挖掘技术的电信市场营销管理，提出了精确化营销方法论的完整的系统体系构架。并着重在目标市场营销和客户行为预测两个目前电信市场最为关注的市场营销专题方面对精确化营销的方法论的应用进行了详细的阐述，并以实际案例加以说明。

论文着重研究了以下三个方面的内容：

- 面向电信企业的数据仓库平台建设与应用
- 精确化营销管理的方法论一基于数据分析的闭环营销管理流程体系
- 数据仓库在电信市场营销的应用，重点介绍目标市场营销和客户行为预测营销战略

论文介绍了面向电信企业的数据仓库平台的建设和应用。数据仓库在电信营销领域的应用方面，论文介绍了目标市场营销的概念思想，并对基于数据仓库和数据挖掘技术的电信目标市场营销战略作了详细介绍。另外介绍了基于数据仓库和数据挖掘技术的预测分析方法的客户行为预测市场营销战略的思想方法，及其在固话语音流失预测分析和客户保有等电信市场营销活动中的应用。两类应用分为数据准备和分析、营销活动策划、营销活动准备、营销活动执行和营销活动评估与优化几个环节。通过客户分群和预测分析，我们可以更精细地了解客户，并采取有针对性的营销活动，最终提高营销活动的投资回报率。论文对某电信企业在客户分群和客户行为预测营销战略方面实施精确化营销管理的成功案例进行了介绍。

论文着重对精确化营销管理的方法论进行了系统的介绍。精确化营销方法论体系包括市场营销环境分析、客户战略、营销战略、营销活动的设计执行和营销活动评估等方面内容。

7.学位论文 陈乐 数据仓库在电信企业中的应用研究 2005

本文在讨论了数据仓库的基本概念后，具体分析并讨论了数据仓库在电信企业中的应用，介绍了数据仓库原理和应用技术，分析了数据仓库在电信企业的应用现状，同时以山东网通为实例，具体介绍了数据仓库的应用效果及存在的问题，并研究探讨了数据仓库在电信企业中应用的可行性，为山东网通数据仓库的应用提供了一套切实可行的建设性方案。

8.学位论文 刘丹阳 基于数据仓库的电信企业EIS的研究与实现 2006

随着电信企业业务运营支撑系统的建设逐渐完善，电信企业的经营分析系统建设已经成为近年来各电信企业的信息化建设重点。基于数据仓库的经营分析系统能够为电信企业的分析与决策提供更好的支持。而一般的经营分析系统并不适合电信企业主管人员的工作习惯，在应用中还有诸多限制，所以建立专门针对电信企业高层管理者的EIS是很有必要的。

本文在对高层管理者的工作特点和信息需求考察与分析的基础上，对基于数据仓库的EIS进行了研究。本文提出了EIS应具备的功能，即信息查询、主动告警、计划实施、通讯交流。依据其功能设计出EIS整体的体系结构，并对体系结构中各模块进行了描述与详细设计。

9.学位论文 黄毅东 基于数据仓库的电信经营分析系统设计与研究 2008

通信行业是一个“数据密集型”的行业，电信企业必须处理和跟踪用户通信信息来监控网络质量，计算通信费用和制定网络建设和优化计划。而这些数据被割裂在各个业务系统中，没有被有效的开发利用。通信行业令人着迷的地方就是电信企业可以利用技术、数据和知识为客户提供更好的服务。竞争使得数据变得越来越重要。电信企业真正挑战是如何有效获取正确的数据并运用正确的数据分析工具，来为企业经营者提供决策支持。数据仓库和数据挖掘技术前景被大家看好，我们希望运用这些技术来解决电信企业现有存在的问题。数据仓库技术为电信企业提高服务质量和增强企业竞争力提供了一条捷径。数据挖掘技术是一种更智能的分析工具，不仅为我们提供各种数据报表，同时帮助我们去发现过去我们没有察觉的业务关联和商业机会。

本文给出了一个电信经营分析系统设计与实现过程，这是笔者在攻读硕士研究生期间，参与设计和开发的国内某电信运营企业经营分析系统。该系统以数据仓库为基础，我们通过系统优越性来阐述数据仓库技术在提高企业核心竞争力中的作用。数据仓库技术是通过集成各个系统的数据为系统用户更有效地提供数据。它可以加速潜在通信技术应用到经济、民生、商业和政府的进程。

本文主要研究电信经营分析系统和数据仓库，论文主要工作包含三个部分：

第一部分、电信经营分析系统设计。这部分描述了电信经营分析系统的框架，包括电信经营分析系统框架设计、系统部署、网络设计和功能模块设计等。

第二部分、电信数据仓库的建立。本阶段介绍创建电信数据仓库所依据的理论和创建方式，为创建电信数据仓库提供实用的方法。本阶段内容包括：数据仓库系统应用系统结构，数据模型设计，数据粒度设计，数据分区设计和数据集市设计。

第三部分、电信数据仓库ETL设计。ETL是数据仓库建设的一个步骤。本阶段在研究数据质量问题相关理论的基础上，面对电信企业高质量数据需求，设计了面向电信应用的数据质量控制体系，实现的以数据清理为主要功能的数据加载(ETL)系统。

10.期刊论文 彭辛庚.陈湘涛 电信企业数据仓库中元数据管理的探索与实践 -电信科学2009,25(7)

本文首先分析了电信企业数据仓库中数据管理存在的问题,引出元数据管理的目标和元数据管理系统要实现的主要功能,并设计了元数据管理系统的总体架构.在具体实现中重点探讨了元数据的自动获取技术和接口技术,然后举例说明了元数据管理的几个主要应用.最后对元数据管理系统的发展进行了展望.