单位代码： 10293 密 级：

申请

申请



专 业 学 位 硕 士 论文



论文题目： ft东电信企业经营分析系统的设计与实现

学

号

G071204073

姓

名

贺福银

导

师

肖甫 李玉华

专业学位类别

工程硕士

类

型

在 职

专业（领域）

计算机技术

论文提交日期

2013 年 6 月

**The Design and Realization of Shandong Telecom Business Analysis System**

Thesis Submitted to Nanjing University of Posts and Telecommunications for the Degree of

Master of Engineering



By Fuyin He

Supervisor: Prof. Fu Xiao June 2013

南京邮电大学学位论文原创性声明

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，也不包含为获得南京邮电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。

本人学位论文及涉及相关资料若有不实，愿意承担一切相关的法律责任。

研究生签名： 日期：

南京邮电大学学位论文使用授权声明

本人授权南京邮电大学可以保留并向国家有关部门或机构送交论文的复印 件和电子文档；允许论文被查阅和借阅；可以将学位论文的全部或部分内容编入有关数据库进行检索；可以采用影印、缩印或扫描等复制手段保存、汇编本学位论文。本文电子文档的内容和纸质论文的内容相一致。论文的公布（包括刊登） 授权南京邮电大学研究生院办理。

涉密学位论文在解密后适用本授权书。

研究生签名：

导师签名：

日期：

摘要

随着信息管理系统的广泛应用和数据量的激增，企业经营管理人员不仅仅需要对数据进行查询和统计，还需对大量繁杂的数据进行线上数据处理和数据挖掘，进而分析出有价值的资讯，从而帮助企业决策者进行正确的决策。在此情况下，传统关系型数据库已经不能满足企业对于数据分析的需要，数据仓库技术应运而生，目前国内数据仓库在各个领域得到广泛应用。

论文针对企业发展需要，将数据仓库技术应用于ft东电信的经营分析系统的建设中，力求提供一套可行的解决方案。论文完成的主要工作包括设计实现了一个包括数据抽取、转换、清洗和转载等过程的数据仓库模型，在对企业业务需求进行探讨的基础上，通过需求分析确定用户业务分析所需的数据，并在模型设计阶段进行了数据仓库的逻辑设计，包括选择合适的主题、确定事实表、相关的维、属性等，并针对数据抽取、清洗、转载过程中凸显的数据质量问题，对数据进行检验、整理、聚集和清洗等操作，最后基于MicroStrategy实现了ft东电信的数据仓库建设和联机分析。

**关键词：** 数据仓库； 维度建模； 联机分析处理； 数据挖掘

Abstract

With the rapid increase of data amount and widely used of the information management system, enterprises managements need to query and statistics on the data, and online data processing, data mining of large amounts of complex data is also needed, and then analyzes the valuable information, so as to help business decision-makers to make correct decision. In this case, traditional relational database can not satisfy the need for data analysis, data warehouse technology emerge as the times require. The present data warehouse is widely applied in various fields.

According to the needs of enterprise development, a business analysis system using data warehouse technology in Shandong Telecom is constructed in this thesis. In order to provide a set of feasible solutions, design and implementation of a data warehouse model including a data extraction, conversion, cleaning and transfer process. Are discussed based on the enterprise business needs, determined by the demand analysis of user service of data required for the analysis, the logical design of data warehouse in the model design, including the selection of appropriate subject, determine the fact table, correlation dimensions, attributes, and according to the problems of data quality of data extraction, cleaning in the process, reproduced highlight, on the test data, sorting, aggregation and cleaning operations, and data warehouse, on-line analytical processing of Shandong telecom is constructed based on MicroStrategy analysis method

**Keywords**: data warehouse; Dimensional modeling; On-line analytical processing; Data mining

目 录

[摘要](#_Toc68699009) 2

[Abstract](#_Toc68699010) 2

[第一章 绪论](#_Toc68699011) 3

**[1.1](#_Toc68699012)** [课题背景及研究对象](#_Toc68699012) 3

**[1.2](#_Toc68699013)** [国内外研究现状](#_Toc68699013) 4

**[1.3](#_Toc68699014)** [课题主要工作及论文结构](#_Toc68699014) 4

**[1.4](#_Toc68699015)** [术语定义](#_Toc68699015) 4

[第二章 企业经营分析系统的需求和总体设计](#_Toc68699016) 5

**[2.1](#_Toc68699017)** [经营分析系统框架](#_Toc68699017) 5

**[2.2](#_Toc68699018)** [经营分析系统数据源](#_Toc68699018) 5

**[2.3](#_Toc68699019)** [系统需求](#_Toc68699019) 6

**[2.4](#_Toc68699020)** [系统功能](#_Toc68699020) 6

**[2.4.1](#_Toc68699021)** [采集管理](#_Toc68699021) 6

**[2.4.2](#_Toc68699022)** [数据处理](#_Toc68699022) 6

**[2.4.3](#_Toc68699023)** [数据查询](#_Toc68699023) 6

**[2.4.4](#_Toc68699024)** [分析展现](#_Toc68699024) 6

**[2.4.5](#_Toc68699025)** [权限管理](#_Toc68699025) 6

**[2.4.6](#_Toc68699026)** [数据仓库管理](#_Toc68699026) 6

**[2.4.7](#_Toc68699027)** [其他功能](#_Toc68699027) 7

[第三章 ft东电信数据仓库建设](#_Toc68699028) 7

**[3.1](#_Toc68699029)** [数据仓库建模分析](#_Toc68699029) 7

**[3.1.1](#_Toc68699030)** [模式选择](#_Toc68699030) 7

**[3.1.2](#_Toc68699031)** [确定主题](#_Toc68699031) 7

**[3.1.3](#_Toc68699032)** [维度表](#_Toc68699032) 8

**[3.1.4](#_Toc68699033)** [事实表](#_Toc68699033) 11

**[3.2](#_Toc68699034)****[ETL](#_Toc68699034)**[数据采集模块设计](#_Toc68699034) 12

**[3.2.1](#_Toc68699035)** [数据质量问题](#_Toc68699035) 12

**[3.2.2](#_Toc68699036)** [数据清洗问题现状](#_Toc68699036) 13

**[3.2.3](#_Toc68699037)****[ETL](#_Toc68699037)**[设计](#_Toc68699037) 13

[第四章 数据仓库的实现和](#_Toc68699038)**[OLAP](#_Toc68699038)**[分析](#_Toc68699038) 18

**[4.1](#_Toc68699039)****[MicroStrategy](#_Toc68699039)**[功能介绍](#_Toc68699039) 18

**[4.2](#_Toc68699040)** [对多维数据的访问](#_Toc68699040) 18

**[4.2.1](#_Toc68699041)****[OLAP](#_Toc68699041)**[分析的基本操作](#_Toc68699041) 18

**[4.2.2](#_Toc68699042)** [模型分析方式](#_Toc68699042) 18

**[4.2.3](#_Toc68699043)****[OLAP](#_Toc68699043)**[分析举例](#_Toc68699043) 19

[第五章 总结和展望](#_Toc68699044) 20

**[5.1](#_Toc68699045)** [总结](#_Toc68699045) 20

**[5.2](#_Toc68699046)** [展望](#_Toc68699046) 20

[参考文献](#_Toc68699047) 20

# 第一章 绪论

## **1.1** 课题背景及研究对象

本论文的研究对象是中国电信集团公司ft东省分公司（以下简称ft东电信）的企业经营分析系统。随着ft东电信在全省范围内的各大业务系统的建立和实施，各地市公司逐步建立起了支撑业务发展和分析的多套系统，这些系统的建立极大提高了公司的工作效率。但由于各业务系统互相独立，业务数据分别存放在不同的业务数据库中，不能充分联系，造成系统间的数据不能有效的共享，从而出现了业务系统“信息孤岛”的孤立情况。

在通信市场竞争日趋激烈的情况下，ft东电信迫切需要对各个业务系统的数据进行统计分析。同时全省业务日趋统一，向省公司集中模式转换，ft东电信公司需要及时地了解全省的业务发展情况，这样就会存在两个方面的问题：一方面，ft东电信分析决策的数据来源于多个分散的业务系统；另一方面，各个业务系统无法实现自身数据在整个ft东电信范围内的共享。

ft东电信为了解决这些问题，需要建立一个企业级的数据仓库，用于发布经营分析数据，在这个平台上，系统的使用者可以从平台上获得相关的统计数据和指标数据。同时，建立这样一个平台，可以有效的促进业务系统间数据处理的一致性，提高各业务系统协同工作的能力[1-3]。

未来企业的竞争，在很大程度上是其收集数据、分析数据能力的竞争。这种分析只有建立在企业内部各个环节和外部市场等方面的数据基础上，才能真正的体现企业运行的真实情况，也表现出用户的消费行为、消费方向等特征。数据仓库技术是行业竞争的必然结果。

## **1.2** 国内外研究现状

数据仓库作为一个新的研究领域[4, 5]，它的研究和应用得到了广泛的关注，也很快成为数据库市场的一个新的增长点。国内外许多厂家和公司如：IBM、Oracle、微软等有实力的公司都在这个领域进行了深入的研究，并相继推出了自己的数据仓库解决方案。

在我国，数据仓库启动较晚，主要由于当时我国的应用基础尚不完善，企业意识不足。但随着我国信息化建设的逐渐完善和应用意识的提高，许多行业，如：电信、金融、税务等逐步认识到数据仓库技术对企业宏观发展带来的巨大的经济效益，纷纷建立起数据仓库系统。Oracle、IBM、Sybase、CA等厂商也

为我国提供了大型数据仓库的解决方案。目前已经成功建设并收回投资的项目已经很多[6]。和国外相比，国内的数据仓库市场还很不成熟，主要的原因有以下几个方面：

建设数据仓库的前提是要有大量的数据，特别是历史数据。近几年来，我国的数据库建设有了长足的发展，但与国外的发展程度相比仍然相差

20年的时间。目前国外的数据库已是海量，而国内，除了银行、电信等少量行业外，数据的积累都不够充分，数据库的规模也不大。并且，有的数据库建完后，企业为了节省资金，往往不投入资金进行后期的维护，就变成了“死库”；

数据仓库应用主要是一个建设的过程，实践性非常强，而从国外（包括外企）的情况看，其中能够完整实施数据仓库方案的人才严重匮乏，因此制约了国内数据仓库市场的发展；

目前数据仓库工具没有国产的，由于东西方文化的差异，数据展现不太符合国内客户要求，必须重新定制开发应用界面，从而延长了建设周期。目前的数据仓库产品已经针对中国客户的习惯进行了相关的改造。

另外，由于数据仓库技体系还不成熟，无论在中国，还是在世界上其他国家，数据仓库的应用受到了一定的限制[7, 8]。主要原因如下：

首先，数据仓库技术还没有达到数据库技术的成熟度和易用性。目前市场上的数据仓库产品都是基于一个通用的技术平台设计的，虽然能解决不同的用户的需求，但是没有将特殊领域的商业逻辑和数据仓库技术集成，因而分析效果不可能达到峰值。

其次，数据仓库的分析工具有两大类。一类是联机分析处理OLAP，另一类是作预测型分析的工具数据挖掘工具DW。它的分析过程是自动的，用户不必提出确切的问题，DW挖掘隐藏的模式预测未来的发展趋势，有利于发现未知的事实。目前基于数据库挖掘的算法层出不穷，还没有经历一个大浪淘沙的沉淀过程，而数据库中的检索技术经过人们多年的摸索，已经形成几种固定的模式。这也是数据仓库产品远没有达到数据库产品实用性的另一个原因。

因此，数据仓库技术的发展仍然处于积累阶段。

虽然数据仓库技术还需要不断的完善，但是其能带来的巨大的经济效益，使得它的发展前景不可限量。据IDC分析，从1997年至今，整个数据仓库的以平均20.5％的速度增长。数据仓库能够带来的巨大的经济效益也必将刺激中国的数据仓库市场的发展，近几年，国内已经有不少的用户在数据积累上作了大量的工作，若把这些现成的东西插上决策分析的翅膀，必将使得当今的企业如虎添翼，成为企业增长竞争实力的有力武器[9-12]。

总之，中国的数据仓库市场前景广阔，充满商机。

## **1.3** 课题主要工作及论文结构

目前ft东电信的企业信息化建设空前发展[13]，各业务系统的建设都已经比较完善，越来越多的业务系统，使得ft东电信的业务数据积累也空前的膨胀，这些数据存储在不同的业务系统中，成为一个个的信息孤岛，企业的决策需要这些数据，但是目前还没有一个大型的统一的平台来对这些数据进行分析和展现，本论文所研究的ft东电信经营分析系统的设计与实现是针对当前实际市场需求，通过统一的平台对ft东电信的海量数据进行统一的存储和展现，从而实现全省范围内经营分析数据统一架构、智能展现，通过抽取各业务系统数据，实现全省经营分析数据的数据共享，同时为集团公司数据仓库提供统一数据源。

本文主要研究了ft东电信经营分析系统的总体设计框架，数据仓库建设的过程和数据的分析和展现。通过描述经营分析系统的整个提出、立项、分析、设计、展现过程，以及我个人在主要参与数据处理、数据查询开发过程中的工作，不断学习的新知识，在完成了系统的设计的基础上，提高了个人的数据库水平及应用能力，给以后的工作打下了坚实的基础。

第一章分析了研制数据仓库建设的背景和国内外的发展现状，简要阐述了本文所需要解决的主要问题，最后指明了本课题研究的意义。

第二章对ft东电信的经营分析系统进行总体概述，着重介绍系统的目标、系统架构及系统功能模块。

第三章对ft东电信的数据仓库建设进行描述，着重描述数据仓库的建模过程。需求分析过程已经确定了用户业务分析所需要的数据，模型设计阶段将确定数据仓库系统将来的蓝图。该阶段的主要任务是进行数据仓库的逻辑设计，包括选择合适的主题，确定事实表、相关的维、属性和颗粒度划分，设计正确的表结构和主、外键关系等。对ETL数据清洗转换进行了介绍，主要阐述了数据质量问题在数据仓库解决方案中的重要地位、数据质量的分类，同时对本课题中的ETL的设计和实现进行了描述。

第四章着重介绍如何利用MicroStrategy实现数据仓库和OLAP分析。

第五章是结束语，对本文所作的工作和下一步要解决的问题进行了总结。

## **1.4** 术语定义

本论文中所用到的专门术语的定义和缩写词如表1-1所示：

表1-1 术语定义表

| 序号 | 名词 | 定义和解释 |
| --- | --- | --- |
| 1 | BI | BusinessIntelligence，商业智能 |
| 2 | ETL | Extraction&Transformation&Loading. 抽取、转换和装载的过程，  包含了数据从源数据系统到经营分析数据库中的所有步骤，其间要对数据进行过滤、修正、聚合等操作。 |
| 3 | OLAP | On-LineAnalyticalProcessing，联机分析处理。OLAP 是使分析人员、  管理人员或执行人员能够从多角度对信息进行快速、一致、交互地存取，从而获得对数据的更深入了解的一类软件技术。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。 |
| 4 | 维度 | "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。 |
| 5 | 实体 | 实体表示商务模型中的某个实例，通常由数据库中某个唯一的 ID 列  标识。 |
| 6 | 事实 | 事实是基于数据库中的物理列而定义的一种对象，它是形成度量的  基础，将度量与数据库中的数据联系起来，被广泛的应用在分析中。 |
| 7 | 度量 | 度量是对特定维度上的事实进行计算的结果。依据不同的维度，对  事实进行计算所得到的数据就是度量。 |
| 8 | 主题 | 面向一个有共同分析特征的群体，并共享基本分析方法的一个分析  类别。 |
| 9 | 专题 | 分析一类专门的业务问题，并根据发现的问题进一步分析出产生问  题的原因和规律的一个分析类别。 |

# 第二章 企业经营分析系统的需求和总体设计

## **2.1** 经营分析系统框架

经营分析系统面向公司内部各业务系统定期自动提取经营业务数据，经过对数据的清洗、转换，根据预先定义的规则加载到省公司集中的经营分析数据仓库中；全省各经营单位系统使用人员通过Web界面对业务数据进行查询和分析。同时，数据仓库的管理作为数据仓库的一个重要的组成部分，为后台数据的组织和管理提供保障。



图2-1 企业经营分析系统架构

如图2-1所示：系统主要分为数据采集子系统（ETL）、分析展现子系统（数据利用）和数据仓库管理子系统（数据仓库组织和管理）三部分。数据采集子系统程序采用C/S架构设计，确保了数据处理的高效性和安全性；分析展现子系统采用B/S方式，方便于所有用户通过网页直接进行业务数据的查询和分析。而数据仓库管理的目的是建设和维护一个高性能的数据仓库环境。

## **2.2** 经营分析系统数据源

目前ft东电信主要的数据来自于计费系统、综合客服管理系统、结算系统、天翼手机增值业务结算系统、智能网系统和客服系统，其中计费系统的话单包含用户消费行为最为详尽的信息，是进行客户管理的关键数据，同时可以从这些数据中探索用户的消费习惯，以便采用各种促销手段来更好地为用户服务；综合客服管理系统主要提供前台营业受理和业务发展量的信息；天翼手机增值

业务结算系统主要提供天翼手机业务发展和结算数据；智能网系统提供卡类业务量发展数据；结算系统提供与其他运营商的结算数据和南北结算数据，可以分析其他运营商的业务发展情况；客户服务系统提供客户服务信息，可以分析客服中心的话务量、话费查询情况和业务咨询情况等。

## **2.3** 系统需求

遵循ft东电信的集中、统一规划的指导思想，针对目前生产系统较多，维护、运行、计费、客户资料、用户使用情况等经营分析基础数据提供系统参差不齐的情况，在全省范围内面向上层数据需求，构筑ft东电信企业级的数据仓库，以及数据处理、展现和分析平台。

系统的建设，主要目标是将ft东电信各系统的基础数据通过收集、筛选、整理、分析、展现的步骤，整合成为给经营一线和领导层决策提供有力依据的数据。系统开发包括对系统整体结构进行设计和搭建，以及完成对本地网计费、本地网收费、营业受理（综合客服管理系统）、天翼手机增值、南北结算、网间结算等业务系统的基于指标和统计数据的分析需求。

系统整体结构上，后台通过提供灵活开放的数据采集接口和安全可控的采集管理功能实现业务经营数据的抽取、转换和装载，构筑数据仓库；前台采用强大灵活的Web访问机制，利用OLAP强大的分析引擎，满足前端用户的分析需求。

系统建设完成后，整体结构上能够实现经营分析的功能，如：可分析的经营专题、各种分析专题的联合展示、灵活的报表、灵活报表定制、各类图表分析、数据钻取等。业务上能够满足市场部门对经营数据的统计和分析需求。

除了在系统的功能上要求外，省公司对于性能需求也提出了一定的要求：按照一年的数据量计算，存储空间大约为1.5T，要求传输过程中能够进行

数据正确性校验。

对于前台数据展现、数据查询等界面的响应时间要求如下：

对于预先生成结果的查询和报表（数据量在3万条以下），响应时间应该在15秒内完成。

对于一般的数据查询分析（数据量在3万条以上，200万条以下），尽量在一分钟内完成；

大数据量查询处理生成时间视数据量大小而定，尽量在二十分钟内完成，超过二十分钟的，不建议在本系统前端处理。

## **2.4** 系统功能

系统从功能上分为六个部分：采集管理、数据处理、数据查询、分析展现、权限管理、数据仓库管理和其他功能。

各部分功能简介如下：

### **2.4.1** 采集管理

将业务系统的数据采集到基于分析的经营分析数据库，并进行相应的管理。支持任意期的数据采集，采集服务模块自动定期调度采集任务，并同时完成数据的清洗、转换处理；采集监控模块可以进行采集任务的管理，并可以监视和控制数据采集的任务处理情况。

### **2.4.2** 数据处理

将具体的业务数据抽取到经营分析数据库并进行数据清洗和转换。

### **2.4.3** 数据查询

系统中的部分需求为简单的业务数据查询，考虑查询用户较多或数据量较大的情况，为了减轻分析服务器的负担，不使用专业分析工具，采用

JSP+JavaBean的方式直接从数据库中查询返回结果。

### **2.4.4** 分析展现

总结各业务系统的专题，形成业务分析模型，用于通过界面将数据结果向最终用户展现。另外该部分实现系统的分析主界面、分析维度、自定义报表等。

### **2.4.5** 权限管理

为了更方便的进行系统用户权限的管理，系统允许设置各种角色，而每个角色都对应了系统的若干功能模块，系统管理员可以增加、修改、删除用户角色，也可以增加、修改、删除角色的权限。

系统采用分级授权的方式，权限分为功能权限和数据权限，权限管理员可以创建用户，但只能从自己拥有的权限中选取功能权限或数据权限授给新用户。

这样管理的目的是为了方便系统大量的用户管理，不至于所有的用户都由系统管理员来创建，同时又能控制系统全部的管理员不能越界授权。

### **2.4.6** 数据仓库管理

数据仓库管理包含接口管理、数据抽取管理、数据加载管理，数据备份管理，完成数据仓库描述、数据转换模型描述、分析主题的定义、分析子系统的主题结构定义，数据加载与备份等维护任务[14-17]。

数据接口管理包括数据抽取的抽取逻辑、数据校验逻辑、数据重构转换逻辑等接口逻辑构成。

数据抽取管理主要是定期从各个应用系统中抽取数据到数据仓库中来。（此项功能在数据采集模块中实现。）

数据加载以数据仓库数据为源头，进行主题数据处理加工，并装入主题数据表中。

数据备份管理对必要的细节数据、主题数据以及MetaData进行备份或恢复管理。

### **2.4.7** 其他功能

实现系统主要业务功能需求之外的其他需求，包括：外部数据源上报报表、文件上传下载、讨论区、参数维护等功能。

# 第三章 ft东电信数据仓库建设

在电信行业经营分析系统设计中，数据仓库维度建模设计至关重要。并将作为企业经营分析中最为关心的主题之一。本章将阐述如何构建数据仓库的维度模型。

## **3.1** 数据仓库建模分析

面向经营分析系统的数据仓库与面向事务处理的数据库应用需求不同，所采用的建模方法也不同。数据仓库建模的两个重要特点是面向主题和集成性。面向主题与传统数据库面向应用相对应。主题是一个在较高层次将数据归类的标准，是用户使用数据仓库进行决策分析时所关心的重点方面。每一个主题对应一个分析领域，通常与多个操作型信息系统相关。集成性是指在来源复杂的数据进入数据仓库之前，必须经过数据加工和集成，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电信行业的一致的全局信息。

### **3.1.1** 模式选择

逻辑建模是数据仓库实施中的重要一环，因为它能直接反映出业务部门的需求，同时对系统的物理实施有着重要的指导作用。目前较常用的两种建模方法是所谓的星型模式和雪花模型。

星型模式

星型模式是一种多维的数据关系，它由一个事实表（FactTable）和一组维表（DimensionTable）组成。每个维表都有一个维作为主键，所有这些维组合成事实表的主键，换言之，事实表主键的每个元素都是维表的外键。事实表的非主属性称为事实（Fact），它们一般都是数值或其他可以进行计算的数据，而维大都是时间、地域等类型的数据。

雪花模式

基本的星型模式并不能满足数据挖掘的所有需要，我们需要更复杂的维度，例如时间。分析员希望根据周、月、季度等识别模式。维度必须进行规范化。这种过程中我们得到的模式被称为雪花模式。

由于电信行业数据量非常大（达到TB级），在进行数据仓库设计时，多表连接、表的累计、数据排序、大量数据的扫描等操作是面临的主要问题。要解决这些问题，在设计模型时，就需要采取如对表进行预连接（Pre-Join）、在模型中增加有关小计数据（SummarizedData）的项、对数据事先排序、通过使用

大量的索引等措施。

基于系统的响应速度、系统的复杂度、系统的维护工作量等方面考虑，我们在实施某电信行业数据仓库系统的建设时，采用星型模式和雪花模型两种模式。星型模式的特点是速度快，它针对各个维作了大量的预处理，如按照维进行预先的统计、分类、排序等。而雪花模型允许我们设计更为复杂的维度。

### **3.1.2** 确定主题

在数据仓库中，数据是面向主题组织的，主题是在较高层次上将企业信息系统中的数据进行归并的抽象，是对分析对象的数据的一个完整的、一致的描述，体现出分析对象的各项数据相互关联，结合需求，逐步确立主题应包括的基本内容即主题的模式和所涉及的属性、数据变量，从而确定主题[18, 19]。

主题的固有信息是对该主题的属性的描述信息，即维度，以电信的产品为例，如：产品的类型；主题的业务信息是对该主题在各类业务的流动信息的描述，即度量，如产品的销售信息。描写主题的数据组织是将原来的按应用进行组织的关系模式重新以面向主题的方式加以改造，是按照确定的主题及其模式，从面向应用的各子系统中抽取相应的数据，形成关于该主题的全局一致的业务模式和信息集合，以便在此基础上分析该主题的相关信息，从面向主题的角度观察原有的数据。



电信企业的业务纷繁复杂，在信息集成中，主题通常由一些关系表表现，如图3-1所示。在本文中我们只选取其中几个进行描述。

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

图3-1 ft东电信主要业务主题分类

#### **3.1.2.1** 客户主题

客户构成分析项从客户自然属性以及业务类型、客户状态等角度对各种业务的客户数进行分析与预测，以达到充分了解不同属性的客户的总体构成及其发展趋势的目的。

##### （**A**） 客户发展专题

该专题从客户自然属性、业务类型等角度，重点对新增客户数、流失客户数、净增客户数等指标进行统计，发现ft东电信公司客户数的发展趋势及变化频率；通过一些统计分析手段，找出客户数激增的客户群体具有的属性，以及客户流失数激增的客户群体具有的属性，从而推出有针对性的营销策略，并做好应急措施。同时形成对相关业务和策略的初步评估，挖掘影响客户发展的因素，以便有针对性的采取措施吸引新客户，合理的减少客户发展的成本。通过对新增客户的分析，掌握客户发展的规律，进行既定条件下客户发展的仿真模拟，为制定科学合理的营销策略提供依据。

##### （**B**） 客户流失专题

该专题主要分析语音类业务客户流失的情况，可以从客户的自然属性角度进行分析，还可以从客户的欠费金额层次角度进行分析。从流失客户的消费行为和信用角度分析流失客户所具有的规律，还能估算出流失的客户对ft东电信的收入损失有多大；另外，从客户流失率这一指标可以看出流失客户占ft东电信总客户数的百分率。

##### （**C**） 客户行为专题

该专题从消费层次、对方通话类型、语音业务产品类型等角度来分析客户的各种通信习惯，并进行自动分群。利用客户群体特征为新产品开发、制定市场推广计划提供决策依据。根据每个客户的群体特征，可以为客户提供更贴切的个性化的服务。

##### （**D**） 客户业务办理专题

该专题通过从客户类型的角度分析各个营业区的各种业务类型的受理量÷营业收入以及受理完成情况，了解各种业务发展变化规律，分析不同类型的客户对业务的需求

#### **3.1.2.2** 业务量收主题

##### （**A**） 业务量和收入量发展专题

该专题从日期、地域两个维度以及它们之间的不同组合，对不同业务类型或产品类型的业务量情况，即通话次数、通话时长、通信流量、MOU等指标进行分析，得出各类型业务、产品的业务发展现状，变化快慢。

##### （**B**） 业务总量构成专题

该专题主要根据通信行为在物理上的总体特征，观察、分析各个业务类型、业务量发展的总体情况，以达到充分了解各项业务、产品的发展趋势、变化情况的目的。

##### （**C**） 收入总量构成专题

该专题旨在分析不同业务收入的费用构成情况，比较一次性费用、固定月租费、通信费等不同费用类型在业务收入总量中的比例情况。

##### （**D**） 缴费情况专题

该专题从不同的角度及其组合对一定统计时限下，缴费帐户数、缴费金额等指标进行多维分析，了解各种不同客户群体的缴费特征。通过对客户的缴费情况与客户的缴费行为进行深层次的挖掘分析，找出影响客户缴费行为的重要因素，为决策支持者制定相应的政策来提高缴费及时率等提供科学合理的手段。

##### （**E**） 欠费情况分析专题

该专题对欠费情况与客户欠费行为进行分析并发现影响用户欠费行为的主要因素，为降低欠费率、提高欠费回收率提供科学合理的手段。通过对客户的欠费情况与客户的欠费行为进行深层次的挖掘分析，找出影响客户欠费行为的重要因素，为决策者制定相应的政策来降低欠费率、提高欠费回收率等提供科学合理的手段。

#### **3.1.2.3** 产品主题

##### （**A**） 产品构成分析专题

该专题旨在通过不同维度和指标的组合，统计出作为最主流的业务之一的各种产品市场发展情况，从而进一步通过使用该产品的用户的相关属性发现其在市场上新的增长点。包括的产品有四类：语音产品、数据产品、智能产品和

出租产品。

主要从用户的自然属性和消费行为等维度进行观察，通过总用户数指标可看出不同时间段的用户构成情况。通过新增/流失指标可看出用户的变化情况和一些发生变化的共性。通过收入量指标有效的将用户实体和收入贡献度结合起来，从而更直接的分析收入与用户特征之间的内在联系。

##### （**B**） 消费行为分析专题

该专题以产生的固话和天翼手机的通话清单为基础，从通信时段维度发现不同客户在使用业务时的消费趋向，从通话类型的角度可看出客户在进行什么通话时更愿意使用普通电话或天翼手机，从单次通话时长层次和通话类型的组合来考察通信时长指标，可以看出客户的消费喜好。

### **3.1.3** 维度表

各种各样的维度定义是经营分析系统的核心，维度决定了系统能够进行哪些分析。所有的维度在MSTR中定义并被存储，用户不能直接察看维度的定义，只能在前台使用这些维度。

分析维度指的是从哪些角度来度量（计算）数据。举例来说，需要计算用户数的时候，首先应当指定从哪些角度来限定计算条件，确定计算什么时间、什么地方的、什么类型的用户数。维度就是这些计算角度的总称。

维由实体组成。以时间维为例，它由年、季度、月份、日期、周、旬这六个实体组成。实体之间存在一对多或者多对多关系，这种关系称为层系。层系描述了实体之间的对应关系。一个年对应了多个月，同样的一个月对应了多个日期。

在对数据进行度量（计算）的时候，首先根据维确定从那些角度计算数据，然后根据每个维指定的具体实体来确定数据从哪个层次上计算。还是以计算用户数为例，指定了从时间、地域两个角度来计算以后，还要知道时间是按月计算还是按年计算，地域是按全省计算还是按地市计算。当明确指定了实体以后，数据的计算粒度也就确定。

本部分描述公共的分析维度。

#### **3.1.3.1** 时间维度

这是每个数据中心都必须提供的一个维度，也是将来进行分类排序的首选维度。这样可以使按照时间间隔连续加载的数据能够顺序的存放到磁盘的空白存储区中。

维度层次

日期的完全描述如表3-1所示。

表 3-1 日期完全描述

| 第一层 | 第二层 | 第三层 | 第四层 | 第五层 |
| --- | --- | --- | --- | --- |
| 日期的完全描述 | 年 | 季度 | 月 | 日 |

时间维度的层次结构如图3-2所示。



图3-2 时间维度图

维度说明

日期维度是一个特殊的维度，可以有多种标识方法，粒度可以灵活掌握，可以到天，也可以到月。

维护属性是查询的约束条件。如果用户想查询周末与平时的业务量发展情况有什么不同，只有星期几这个属性是不够的，所以我们可以考虑增加“周末”和“非周末”这个属性，帮助我们解决这个问题。

还有一个与日期有关的属性，“节日与重大事件”，不能简单的标记为“是”和“否”。而是直接写入节日名称，更有利分析不同节日的业务发展情况，同时我们也可以将事件标识为某项优惠活动推出日。

所以时间维度表我们可以这样设计，如表3-2：

表3-2 时间维度表

| 日期完全描述 | 星期 | 日 | 月 | 周末 | 年 | 节日 | 重大事件 | 其他属性 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2008 年 10 月 1 日 | 星期二 | 1 | 10 | 否 | 2008 | 国庆 | 优惠套餐推出日 |  |

#### **3.1.3.2** 地域类型维度

地域维度也是每个主题中都要用到的维度，最多只是在粒度上有区别，比如有些专题只要到地市级，有些专题要到区县级。



|  |
| --- |
|  |

|  |
| --- |
|  |

地域维度的层次结构如图3-3所示：



图3-3 地域维度图

ft东电信枣庄分公司地域维度局向结构的层次如图3-4所示：

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |





|  |
| --- |
|  |

|  |
| --- |
|  |
|  |
|  |

|  |
| --- |
|  |

图3-4 ft东电信枣庄分公司地域维度局向结构图

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
|  |

#### **3.1.3.3** 业务类型维度

ft东电信目前提供的服务业务类型比较复杂，但是可以根据提供服务的业务类型分为七大类型：语音业务、基础数据业务、窄带互联网业务、宽带互联



网业务、互联网数据增值业务、电话卡业务、网元出租业务。这七大项业务包括普通固话、无线商话、公用电话等36个具体产品类型。

#### **3.1.3.4** 费用类型结构

作为电信运营商之一，ft东电信在向用户提供各种产品及附带服务的同时，将以运营及获取收益为主要目的。所以用户在使用各种产品的时候会产生各种费用，而用户费用结构类型的分析，将是市场经营方向的一个重要风向标，是领导经营决策的重要依据。目前用户产生的费用可以分为一次性费用、周期性费用、通信费用、Inernet网络费、增值新业务费、语音信息费、增值新业务费等七种类型。如下图3-5所示。



图3-5 ft东电信费用类型结构图

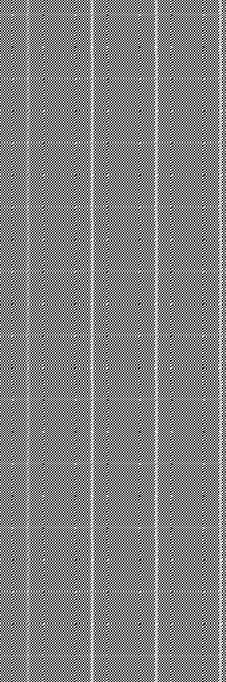
### **3.1.4** 事实表

通过对产品、服务、产生费用、设备等方面的分析，建立相应的事实表，用来存储数据，以便进行数据筛选、查询等工作。下面以用户数情况统计表为例说明事实表的设计及其结构：

表3-3 用户数情况统计表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **字段名** | **字段类型** | **缺省**  **值** | **非空** | **主键** | **外键** | **说明** |
| 0 | Id | Number(12) |  | Y | Pk\_u\_change |  | 序号 |
| 1 | quxian\_Co de | Varvchar2(5) |  |  |  | Pk\_quxian | 区县局代码 |
|  | Zhij\_code | Varvchar2(5) |  |  |  | Pk\_zhiju | 支局代码 |
| 2 | Busi\_Type | Varvchar2 (4) |  |  |  | Pk\_Busi | 业务分类 |
| 3 | Busi\_type1 | Varvchar2 (4) |  |  |  | Pk\_Busi1 | 业务细分 |
| 4 | NewAdd\_ M | NUMBER(6) | 0 |  |  |  | 本月新增 |
| 5 | NetAdd\_M | NUMBER(6) | 0 |  |  |  | 本月净增 |
| 6 | In\_M | NUMBER(6) | 0 |  |  |  | 本月移入 |
| 7 | Out\_M | NUMBER(6) | 0 |  |  |  | 本月移出 |
| 8 | Move\_Bure au | NUMBER(5) | 0 |  |  |  | 本局移 |
| 9 | Reduce\_M | NUMBER(6) | 0 |  |  |  | 本月拆机 |
| 10 | NewAdd\_ Y | NUMBER(6) | 0 |  |  |  | 本年累计新增 |
| 11 | NetAdd\_Y | NUMBER(6) | 0 |  |  |  | 本年累计净增 |
| 12 | Move\_Y | NUMBER(6) | 0 |  |  |  | 本年累计移机 |
| 13 | Reduce\_Y | NUMBER(6) | 0 |  |  |  | 本年累计拆机 |
| 14 | Stop | NUMBER(5) | 0 |  |  |  | 停机保号 |
| 15 | End\_Users | NUMBER(7) |  |  |  |  | 期末用户数 |
| 16 | GetDate | DATE |  |  |  |  | 统计时间 |

数据库中各关键事实表结构及各表之间的关系，通过以下雪花型结构图进行说明。如下图3-6所示：事实表立方体雪花形模型图。



业务类型表

区县局代码表

用户表

支局代码表

业务细分表

区县局代码表

区域代码

区域名称

业务类型编号

区县局代码

业务类型名称

局所名称所属地区

序号

区县局代码 (F K ) 支局代码 (F K )

业务类型编号 (F K 业务细分编号 (F K 本月新增

本月净增本月移入本月移出本局移 本月拆机

本年累计新增本年累计净增本年累计移机本年累计拆机停机保号

期末用户数统计时间

业务细分编号

业务细分名称

支局代码

支局名称所属区县

区县局代码 (FK)

区县局代码

局所名称所属地区区域代码

区域代码表

图3-6 事实表立方体雪花形模型图

## **3.2** **ETL**数据采集模块设计

ETL主要包括：数据的抽取（Extraction）、净化（Cleaning）、转化

（Transformation）、合成（Integration）和装载（Loading）。其功能是把数据从各种存储方式中拿出来，对数据进行检验和整理，确保数据仓库中数据的一致性，还要对部分数据进行聚集操作，提供数据仓库的查询和分析速度。这些完整的、一致的、细节的和综合的数据才可以进入数据仓库。并根据数据仓库的设计要求，对数据进行重新的组织和加工，转换成数据仓库的数据库结构和内部形式并加载到数据仓库中。

ETL的步骤：

数据清洗：同步填写空缺的值，平滑噪声数据、识别、删除孤立点，并解决不一致来“清洗”数据。

数据集成：将多个数据源中的数据结合起来存放在一致的数据存储中。他涉及到实体识别问题、冗余检测问题以及数据值冲突的检测和处理等各种问题。如地域维度，在计费数据库和在综合客服系统数据库中就存在编码和值都不一致的情况。

数据变换：将数据转换成适合挖掘的形式，可能涉及到平滑、聚集、数据概化、规范化和数据构造等内容。例如可以将季度业务发展量聚集成年发展量；可以进行概念分层，用高层次的概念如“省市”替换“区县局”。可以将数据数据按比例缩放，使之落入一个小的特定区间。还可以构造新的属性值添加到数据集中帮助分析。

数据规约：用来得到数据集的规约表示，它小的多，但仍接近于保持原数据的完整性，策略有：数据立方体聚集、维规约、数据压缩、数值压缩、离散化和概念分层产生等。

### **3.2.1** 数据质量问题

数据质量问题是创建数据仓库以及进行数据集成工作中的致命性问题。如果没有深刻的分析数据中存在的问题，就盲目的进行开发、集成，造成的后果不仅是数据仓库创建失败的经济损失，更有可能造成严重的决策失误。

数据质量的基本要素：

正确性（Correnctness）、一致性（Consistency）、完整性（Completeness）和可靠性（Reliability）。在进行大数据量集成时，由于数据源的复杂性，其中包括滥用缩写词、惯用语。数据输入错误、数据中的内嵌控制信息、重复记录、丢失值、拼写变化、不同的计量单位和过时的编码，都给数据的集成和数据仓库的创建维护工作带来非常大的困难，往往导致数据仓库建设失败。

数据质量的分类，如图3-7所示

模式层

缺少完整性约束唯一性约束引用约束

单数据源问题

数据质量分类

多数据源问题

实例层

模式层

实例层

数据记的错误拼写错误

相似重复记录互相矛盾的字

异质的数据模型和模式设计命名冲突

结构冲突

冗余、互相矛盾或者不一致的数据不一致的汇总

不一致的时间选择

图3-7 数据质量问题分类

单数据源情形中出现的问题在多数据源的情况下会变的更加严重，图中对

多数据源并没有列出单数据源情形中就已经出现的问题，模式层次上的问题也会体现在实例层次上。糟糕的模式设计、缺少完整性约束的定义以及多个数据源之间异质的数据类型、命名和结构冲突等，都属于该问题。可以通过改进模式设计、模式转化和模式集成来解决模式层次上的问题。实例层次上的问题在模式层次上不可见，一些可能的情况有数据拼写错误、无效的数据值、重复记录等。数据清洗领域的研究就是为处理上述数据质量问题而提出的。

### **3.2.2** 数据清洗问题现状

数据清洗的定义

数据清洗是指消除数据中的错误和不一致，并解决对象识别问题的过程，数据清洗需要对数据进行分解和重新组合。数据清洗主要分为检测和清洗两个步骤。国内外的相关研究主要有以下几个方面。

提出高效的数据异常检测算法，来避免扫描整个庞大的数据集。

在自动检测数据异常和进行清洗处理的步骤间增加人工判断处理，以防止对正确数据的错误处理；

数据清洗对数据集文件的处理；

如何消除合并后数据集中的重复记录；

建立一个通用的领域无关的数据清洗框架；

关于模式集成的问题；

数据清洗的工具

目前，有多种数据清洗的商用工具来支持数据检测和数据清洗转换，尤其针对数据仓库方向的，同样，它用两种形式：检测异常数据的工具和数据转换工具（针对异常数据的清洗）。ITEGRITY（Vality）使用某些模式和规则来定义和完成数据清洗转换，它进行一些语法分析、数据类型和出现频率分析。其中还有一些语言规则，包括一些列操作（比如：移动、拆分、删除）和行操作（如合并、拆分）。还有一些特殊的清洗工具和ETL工具。如SAS，它是基于DBMS来建立一个存储区，用统一的方式管理元数据。这些元数据关系着源数据、目标映射和描述语言等。

### **3.2.3** **ETL**设计

在本课题中，为完成数据的抽取转换和清洗工作，我们设计了数据采集模块。采集管理是每隔固定周期，在后台针对经营日报系统需要的原始数据进行采集、处理，并对部分数据进行统计，以生成满足经营分析活动的数据。

采集管理分为两个大模块，一是采集管理与监控，二是自动采集。采集管理与

监控主要功能是维护采集程序用到的表，分为数据源维护、静态任务维护、动态任务维护、图形化监控、日志查询五个子模块。自动采集常驻内存，实时采集数据。两大模块都与数据库联系，之间互不干扰，都采用ADO驱动，客户端安装配置十分简单。下面是对每一个小模块的描述：

#### **3.2.3.1** 数据源维护

数据源维护用来设置数据的来源的参数信息。

采集管理程序提供两种采集数据的接口，一是接口表，二是文件，所以数据源也包含两种，一是数据库服务器，二是FTP文件服务器。

数据库服务器参数包括服务名、数据库类型、数据库名称、用户名、密码。

FTP文件服务器参数包括IP地址、端口号、用户名、密码。

为了保证数据安全，数据库的密码和FTP文件服务器的密码都要经过加密之后再保存。

数据源合法性检查的流程图如图3-8所示：



图3-8 数据源合法性检查的流程图

#### **3.2.3.2** 静态任务维护

采集程序每隔固定周期就会采集数据，数据采集的周期、数据来源、导入何处等信息都记录在静态任务中。该模块功能就是维护用户定义的这些静态任务。为了方便管理，增设了任务组的概念，同一任务组的任务采用相同的采集周期，组内的任务之间可以设立先后关系。

任务组的采集周期可以是月、周、天或者用户自定义时间间隔，还可以设置偏移量。例如：计费指标任务组采集周期为一个月，时间偏移为26天3小时，

则它会在每个月27号凌晨三点被自动执行。通过这几个固定周期和用户自定义时间间隔可以做到以任意周期执行任务，完全突破了固定周期的限制。

根据采集数据的流程不同，一个静态任务可以实现以下五个流程之一：

调用存储过程对已有数据进行处理；

直接传输导入数据；

通过接口表方式，从业务库读出数据，导入本地数据库。

文件方式导入数据；

从文件服务器下载数据文件，读出文件，导入本地数据库。

导出表中数据成为文件；

把数据库中的表导出成为文件，放到FTP文件服务器上，供其它采集程序来此下载；

BCP方式

针对地市SYBASE数据库中的超大数据量表，采用BCP方式导出数据，处理成为标准格式，再导入本地数据库。

本模块除了检查基本数据的输入，例如：数据是否唯空等，还检查任务之间的先后关系，任务之间的先后关系需要经过严格的检查，以防止死锁的产生。死锁在创建新任务时不可能产生，只会在修改任务时产生。下图3-9是修改任务时，检查任务前后关系的流程图：





图3-9 任务修改前后关系检查图



#### **3.2.3.3** 动态任务维护



静态任务加上具体执行时间即为动态任务。部分采集任务有难以自动确定运行时间的情况，比如各地市计费统计数据每个月的生成时间不确定，所以应由管理员从前台界面人工触发。其它的任务由自动采集调度，自动采集程序常驻内存，根据各个静态任务的采集周期、上次执行时间和当前时间的比较结果，生成对应的动态任务，然后根据任务之间的先后顺序依次执行。自动采集程序执行出错的任务可由动态任务维护界面手动触发再次执行。

需要的功能有：



在任务执行过程中，不能暂停正在执行的任务，但可以使未启动的动态任务不被自动执行，等待手工再次触发；

检查修正任务的错误后，可以重新生成失败的任务，自动采集程序会重新执行它；

使用本模块生成任何时间的任务，并可删除部分动态任务；生成动态任务的流程图如图3-10所示：

图3-10 动态任务流程图



本模块的难点在于根据任务的采集周期和用户选择的时间生成动态任务。计算新动态任务的时间的流程图如图3-11所示：



图3-11 任务时间计算流程图



#### **3.2.3.4** 日志查询

日志是系统的重要组成部分，对于采集的维护者来说，查看和分析，尤其是任务执行出错的日志，可以合理调整采集的时间、更快速的定位出错原因，是系统正常运行的保证。用户可以任意选择任务组、任务名称、采集方式、开始时间、结束时间等查询条件，查询采集结果。采集任务的日志详细记录了任务的编号、任务名称、源数据源名称、源数据库存储过程、源数据库接口表、目标数据源名称、目标数据库存储过程、记录数、执行时间、耗时等信息。

#### **3.2.3.5** 图形化监控

图形化监控界面用来反馈采集的信息，使用者可以通过界面提供的图形化信息了解系统采集任务的运行情况。

提供的功能包括：通过分段显示采集任务的进度；显示任务花费的时间；要能直观体现出任务的前驱后继关系；能直观看出采集任务执行情况（用不同的颜色标明）；定时刷新各个任务的状态；处理流程如图3-12所示：

图3-12 图形化显示界面流程图

#### **3.2.3.6** 自动采集

自动采集是采集管理的核心模块，其常驻内存，实时检测静态任务是否到达执行时间、生成动态任务、开始执行到达执行时间的动态任务，它与采集管理的其它模块完全分离。

自动采集既需要保证采集数据的准确性，又要保证速度。首先它具有良好的扩展性，在任务繁重的情况下，可以同时在多台机器上运行，分别执行采集任务，同一台机器也可采用多线程模式，同时执行多个采集任务。为了方便用户操作，用户可以自定义变量获取数据或者删除接口表的数据，做到多个任务利用同一个接口表并发。采集过程中进行数据校验，以保证所采集数据的正确性。

自动采集执行失败的情况下，用户可以通过采集监控界面和采集日志分析出错原因，通过静态任务维护界面修改静态任务设置，通过动态任务维护界面重置该任务的状态，使其重新执行。

针对静态任务的五种流程，自动采集采用不同的处理方法：

调用存储过程对已有数据进行处理直接调用数据库存储过程。

直接传输导入数据

采用ADO驱动，从业务库接口表读出数据，生成标准格式的文件，调用

sqlldr导入本地数据库，导入完成后，可再调用存储过程对数据进行再处理。

文件方式导入数据

从文件服务器下载数据文件，读出文件，导入本地数据库，并在导入过程中进行数据校验；导入完成后，也可再调用存储过程对数据进行再处理。

导出表中数据成为文件

把数据库中的表导出成为文件（XML格式），经过压缩形成ZIP文件（减小网络传输压力），放到FTP文件服务器上，供其它采集程序来此下载。

BCP方式

该方式只应用于从SYBASE数据库取数据，调用BCP导出数据，处理成为标准格式，调用SQLLDR导入本地数据库。

自动采集是实现动态任务自动生成并执行的功能，开机自动启动，常驻内存。它启动后生成多个线程，其中一个线程负责生成动态任务，其它线程负责执行任务。其主程序流程图如图3-13所示：

自动采集程序开机自启动

连接数据库

否

结束

等待用户操作

图3-13 自动采集程序流程图

连接成功？

是

到达最大线

程数？

是

否

创建生成动态任务的线程

显示数据库连接不成功

创建执行任务的线程

开始

创建数据库连接

到达生成动态任

务的时间？

否

是

检查各个任务组的采集周期、时间偏移量、上次运行时间

生成该组内的所有静态任务对应的动态任务

生成任务的线程循环检测各任务组，如果某个任务组到达间隔时间，则根据其静态任务的内容生成动态任务，放入动态任务表，其流程图如图3-14所示：

|  |  |
| --- | --- |
|  |  |
| 休息一段时间 | |

图3-14 任务生成流程图

否

有退出信息

是

结束

释放数据库连接

检查控制信息

其它的线程循环检测动态任务表，如果发现某一组任务已全部生成，则根据它们的前驱后继关系依次执行，流程图如图3-15所示：

开始

存在未完成的采

集任务？

否

是

否

有退出信息

是

结束

检查控制信息

找到没有前驱任务或者前驱任务执行成功的任务

调用存储过程

直接连接导数据

文件传输

导数据

导出表中数据

BCP方

式

休息一段时间

图3-15 其它线程循环检测动态任务图

#### **3.2.3.7** 数据处理模块

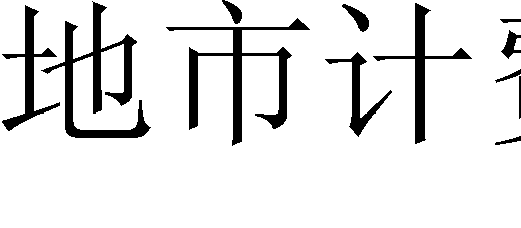
主要以采集管理模块的采集结果为基础，将数据经过抽取、筛选、聚组、求和等操作，把系统所需要、所关心的数据，整理成便于进行数据分析的格式或数据结果。

数据处理模块由统计数据采集调度管理模块进行统一调度，不需提供前台界面。模块中要调度生成数据处理监控信息的相关模块，用以提供数据处理的进度信息，这些信息再通过采集管理的监控界面进行展现。数据处理后的数据结果用于各个业务分析主题和专题进行展现和分析。

数据处理过程如下图3-16所示：



图3-16 数据处理流程



数据处理过程分为定期抽取和非定期抽取两部分，定期抽取的数据一般是一些事实表的数据，而非定期抽取的一般是一些维度表数据，从业务系统将数据抽取上来以后对于事实表数据要进行清洗转换，一般的操作有筛选、聚组、求和、重复数据检测和清除、非法数据清除等操作；而对于维度表的数据，重新抽取后要进行编码的转换，由业务系统的编码格式转换为经营分析系统的编码格式。







# 第四章 数据仓库的实现和**OLAP**分析

## **4.1** **MicroStrategy**功能介绍



图4-1 基本的决策支持架构

图4-1 的基本的决策支持架构中包含如下部分：

源系统(Source System)

数据仓库(Data Warehouse)

MicroStraregy7i商务智能平台(Business Intelligence Platform)

MicroStrategy7i是整个决策支持系统的应用平台，它包含一系列的引擎和服务器组件，用于与数据仓库进行高效的通信，并在给定的时间内精确地获得用户所需的结果。所谓“平台”有两层含义，一是指MicroStrategy7i为所有类型的前端访问（如查询、报表、OLAP分析、高级分析等）提供统一的应用逻辑模型和引擎，另一方面是指用户可以在此基础上定制任何类型的新应用，以直接访问7i平台中的各个引擎和组件[14, 20, 21]，也就是我们常说的二次开发。

下列的引擎构成了7i平台的核心：

SQL引擎和查询引擎

分析引擎

服务器组件（OLAP Service, Report Service）

MicroStrategy能够集成多种架构模型，包括星型、雪花型、混合型、规范、非规范、聚合架构。同时MicroStrategy中包含了一个丰富的架构抽象层，

其中的列名不受事实和属性定义的约束。使用MicroStrategy平台，层次和属性可以很方便的添加进来。比如，如果需要在项目中添加新的层次或层次中的属性，管理员不需要在数据库后台做任何附加的计算。一旦在MicroStrategy

Desktop中定义了新的属性，当这个属性在报表或钻取中用到时，MicroStrategy

SQL引擎会使用合适的连接返回正确的结果集。架构的变更对用户是屏蔽的，用户只需关注他所需要的报表。MicroStrategy对“多对多”关联和跨层次属性

（跨层次属性的之可能依赖于多个层次，比如时间层次和地理层次）提供架构支持。利用这个特性，最终用户可以管理更复杂的报表从而更有效的利用数据。

用户可以随意钻取到任何级别的细节，而无需管理员维护这些钻取路径。例如：MicroStrategy最终用户可以从某个父属性钻取到子属性然后再钻取到一个完全不同的层次。如果一个新属性添加到层次中，并不需要进行相应的数据管理使这个属性可访问。另外，使用MicroStrategy Drill Map编辑器可以自定义钻取路径[22]。

## **4.2** 对多维数据的访问

### **4.2.1** **OLAP**分析的基本操作

1、切片（slice）和切块（Dice）

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；如果有三个，则是切块。切片和切块操作实质上主要是根据维的限定做投影、选择等数据库操作从而获取数据。

2、钻取(Drill)

钻取是改变维的层次，变换分析的粒度。它包括向上钻取（roll up）和向下钻取（drill down）。roll up是在某一维上将低层次的细节数据概括到高层次的汇总数据，或者减少维数；而drill down则相反，它从汇总数据深入到细节数据进行观察或增加新维。钻取的深度与维度所划分的层次相对应。钻取是用户获得详细数据的手段。他一般能回答为什么的问题。一层一层的钻取能快速而准确的定位到问题所在。

3、旋转(Rotate)或转轴

旋转是变换维的方向，即在表格中重新安排维的放置（例如行列互换）。通过旋转可以从不同的视角观察数据。

### **4.2.2** 模型分析方式

#### **4.2.2.1** 构成分析

用户可以针对对象拥有的各种属性来分析整个数据集的构成分布情况。构成分析可以在多个维度上进行，这种分析方式经常使用数据钻取，例如：

时间构成：按照年、月、日或周、日对数据进行分析

地域构成：按照地区、营业区、局向、局号对数据进行分析业务构成：按相应业务分类对数据进行分析

用户构成：按照用户类别对数据进行分析

#### **4.2.2.2** 对比分析

要求能够提供在一个维度上的数据进行对比的功能。例如：

任意时间段对比：用户可以完成任意时间段的比较，比如：任意年份之间的比较、任意季度之间的比较。

任意地域对比

任意可比业务对比

#### **4.2.2.3** 间序列分析

用户可以在时间维度上对数据进行序列分析。如分析某项业务的发展走势，某个时间数据与历史类似时间数据的对比分析，或者某个时间在一个时间段中的比重分析等。

##### 1、 动态统计分析

年/季/月/周/日统计分析

环比分析、不同时期的对比值（比上期、比同期等）增长量/增长率分析

##### 2、 平均计算静态平均计算

动态平均计算（时期平均、时点平均、平均增长率等）3、结构比重分析

静态比重分析

月结构比重分析

季度结构比重分析。

### **4.2.3** **OLAP**分析举例

**每日关注指标分析**



图4-2 每日关注指标分析

图4-2是ft东电信枣庄分公司2013年某日的关键指标数据，显示当日移动计费收入、移动新增用户数、移动网上用户数、移动活跃用户数、宽带新增用户数、移动计费时长、宽带净增用户数、固网计费收入以及移动和联通过网用户数等指标。

从图的左上角部分可以看到，固网发展用户数、宽带发展用户数、移动计费收入、EVDO活跃用户数等项指标有较大幅度的增长，分析原因是因为套餐用户是按照天数计算每日收入。6月份比5月份少一天，因此同一套餐在6月份

就比5月份产生更多的收入，因此收入累计增幅由较大幅度的提升。同时移动发展用户数和移动计费时长有小幅度的下降。

通过点击左上角移动计费收入等项，可以在右上角显示枣庄电信公司下属各区县公司具体数据。数据显示，市公司下属6个分公司收入全部是正增长。其中薛城分公司收入增长最快，同比增幅接近4%，其次是台儿庄分公司和滕州分公司。

通过界面的中间部分，还可以看到ft东移动和ft东联通过网数据，这些数据基本上反映了ft东移动和ft东联通新发展的移动用户用户量。

**报表中心**



图 4-3 报表中心

图4-3报表中心展示的各种业务数据报表。从图4-3可以看出，有关于财务方面的电路结算明细表、手机投注金结算表，有关于创新业务方面的移动增值业务渗透率表、移动增值业务收入表、流量经营表、流量经营客户群分析表，有关于校园发展方面的大中专发展报表、校园营销中心专项日报、校园营销中心月报表、中小学客户月报表，还有关于渠道服务方面的渠道佣金政策、超标佣金、佣金类型、渠道发展数量等多种表格。

点击每一个报表都能够展示出想用的数据。例如关于财务方面的电路结算明细表，点击后查询到的数据如图4-4电路结算明细的方式展示出来。



图4-4 电路结算明细

从图4-4可以清晰地看到每个单位每月结算多少收入。

**月盘点**



图 4-5 月盘点

图4-6显示的是按照月度显示的数据。按照业务类别，有收入类分析、发展类分析、欠费类分析以及考核月报。收入类分析中有包含财务收入汇总分析、移动月收入分档、出账净收入分析。发展类分析包含越发展分析、出账净增分析、宽带分速率统计月报等内容。以收入类分析为例，点击财务收入汇总分析表，将可以查询枣庄市近四年来各月移动网收入、固网收入、预付费用户输入、后付费用户收入、年度收入、月度环比增长率等指标。如图4-7所示。



图4-6 月盘点-区县收入图

从图4-7可以清晰地查看枣庄下属六个区县公司的当月收入、当年累计收入、收入月环比等指标，还可以通过点击左上角的账期，选择不同账期来查看

历史数据。将鼠标放置到相应区县公司上面点击鼠标右键，可以继续向下钻取数据。通过向下钻取，可以继续查看更小一级营销单位乡镇支局的收入以及收入增长变化，还可以根据各类业务收入的变化来分析业务变化是因为宽带收入变化还是因为3G手机的收入变化或者是因为固话语音业务收入变化，进而分析什么原因造成的变化以及下一步需要采取的措施。

**日监控**

日监控可以提供以下功能分析，每日收入总览，每日发展总览，高额话费预警，龙手机日发展总览，苹果手机发展总览，合约智能机日发展总览等。具体内容见下图4-8。



图4-7 日监控图

举例说明：点击日监控画面上的日发展总览，将会出现图4-9。



图 4-8 日监控-日发展总览

从图4-9可以清晰地看到全业务发展用户数，还可以分业务类别看到移动业务发展用户数，无线宽带发展用户数，手机发展用户数，固话业务发展用户数，宽带业务发展用户数以及政企名单制新增用户数等。通过点击移动业务菜

单，就可以看到移动网上用户数、移动新增用户数、移动净增用户数、移动活跃用户数以及移动离网用户数等数据。还可以继续向下钻取，得到各个区县公司的数据以及每个乡镇支局的数据。

# 第五章 总结和展望

## **5.1** 总结

经过公司参与本次设计开发的所有技术人员的不懈努力，ft东电信经营分析系统已经正式运营上线，正在接受全省经营一线员工及经营决策者的检验，我个人能够成为本次开发技术人员中的一员，感觉非常荣幸，经过不断的学习、钻研，不但提高了个人的知识水平，而且实现了公司对我们的期待和要求。本次设计及系统开发的主要贡献在于：

1、实现了ft东电信数据统计需求的重定位，给出了经营分析系统的主要内容；

2、建立了系统的维度模型，围绕收入、业务量、产品等主题，构建事实表和维表，给出了部分维表的层次和事实表的表结构；

3、对本课题涉及的经营分析系统的数据采集、数据清洗和数据转换进行了设计和实现；

4、总结了经营分析系统和数据仓库建设的经验，为类似系统的建设提供可借鉴的经验；

经营分析系统应该架构在数据仓库基础之上：通过前期的调研及学习类似的系统开发经验，我们确定了经营分析系统应该架构在数据仓库基础之上而不应该直接根植于业务支撑系统，这样才能够在数据仓库这个平台上更大潜力的挖掘数据，因为经营分析系统架构在数据仓库基础上，所以对业务支撑系统的性能不会有任何影响。通过数据集成，形成企业的业务数据统一视图，统一了经营分析数据源，并且使业务数据达到良好的一致性。基于数据仓库的数据开发挖掘适应了数据源的变化，针对各种不同统计口径的经营分析查询性能得到了最基础的保障。

大数据量下系统性能的保障手段：详细级别的操作型集成业务数据；汇总级别的分析型经营信息，包括轻度汇总型、中度汇总型和高度汇总型。

数据质量保证手段：经营分析系统是将用户所有资料以及用户资料所衍生出来的子资料进行全方位、多角度的数据挖掘的，而ft东电信公司作为ft东省内的全业务运营商，在整个ft东省的拥有非常可观的用户群，用户数据信息将是一个非常庞大的数字。如何确保在运行大数据量下的系统性能，是一个非常重要的课题。我们做了如下工作确保系统能在最优化前提下运行：构建双重颗粒度级的数据仓库架构；系统级的性能调优，数据库索引及分区、Cube优化；应用级的性能调优，做到反规范化；及时对历史数据

进行存档，确保数据库空间在一个最优的情况。

大数据量下系统性能的保障手段：经营分析系统是将用户所有资料以及用户资料所衍生出来的子资料进行全方位、多角度的数据挖掘的，而ft东电信公司作为ft东省内的全业务运营商，在整个ft东省的拥有非常可观的用户群，用户数据信息将是一个非常庞大的数字。如何确保在运行大数据量下的系统性能，是一个非常重要的课题。我们做了如下工作确保系统能在最优化前提下运行：构建双重颗粒度级的数据仓库架构；系统级的性能调优，数据库索引及分区、Cube优化；应用级的性能调优，做到反规范化；及时对历史数据进行存档，确保数据库空间在一个最优的情况。

5、由本文描述的经营分析系统和数据仓库建设的过程和方法设计出的经营分析系统具有如下的特点：

先进性：系统设计采用当前先进稳定的数据仓库技术，充分吸收国际厂商的先进经验，并考虑到中国国情，系统不仅满足现有的需求，也抓住未来的发展方向；同时系统设计需满足大数据量处理的需求。系统的软件体系结构采用目前先进的多层B/S体系结构或B/S与C/S相结合（发挥各自的优势）。

准确性：提供多种核查或校验手段，保证系统的准确性。

安全性：保证数据不被破坏和盗用。通过操作系统、数据库、数据仓库、

OLAP服务器、WEB服务器等多层安全控制机制保护数据只被经过授权的用户进行访问。

易操作性：良好的用户操作界面、完备的帮助信息。系统参数的维护与管理通过图形化操作界面实现。

可扩展性：采用组件化设计原则，用户可以选择不同组件构成不同规模的应用系统，保持系统核心稳定性和可扩展性，有利于系统扩展功能和系统升级，应用展现环境以B/S架构为主，为未来使数据分析应用与知识查询、

OA等应用系统纳入同一窗体环境打下基础。

## **5.2** 展望

1. 关键成功要素（CSF）：

企业经营分析系统的最初出发点是有效的开发利用各系统的数据及资料，为企业经营发展提供有力的分析支撑，而这个需求的原始原因则是企业感受到来自竞争对手的压力，针对市场竞争日趋激烈，而针对不同消费群体制定不同营销政策，精细化管理和精细化营销则是目前以及以后市场营销的风向。为了将用户群颗粒更细化，更精确，就必须要有一个非常强大有效的分析系统。从

省公司企业领导层面的重视及亲自关心系统建设可以深切的感受到企业对于这样一个系统的迫切期待。而良好的技术体系的支持和开发，也是必不可少的环节，在系统成型推出后，公司在全省范围内的用户级使用测试和推广应用是将技术转化为实际生产力的切入点。 2．后期持续不断地服务（领域专家、业务分析师、数据仓库技术支持）

不断增加的新需求的维护；业务管理模式的变化对系统提出的适应性要求；深层次的挖掘型分析需要不断地维护来保持模型的可用性。

在数据仓库理论和数据仓库的实现的过程中，有一条坎坷的路途要走。在

设计本方案的过程中我们也发现，叉路越走越多，要解决的问题层出不穷，需要我们不懈的努力。同时，数据仓库的中心职责是围绕收集起来的所有数据开展一些有用的工作，它处于IT职能的中心地位。信息环境的迅速发展，使得数据仓库的构造没有静止的目标，比如：随着新技术的不断涌现，安全问题可能成为数据仓库的一个重点考虑的问题。因此数据仓库建设方案的设计也是一个摸索的过程，要需要在实践中不断的完善。

参考文献

[1]. Ralph Kimball, 数据仓库工具箱－维度建模的完全指南（第二版）[M], 谭明金译, 电子工业出版社2003 年

[2]. Ralph Kimball, 数据仓库工具箱－设计、开发和部署数据仓库的专家方法[M], 肖明, 王永红, 译, 电子工业出版社2003 年

[3]. W. H. Inmon, 数据仓库[M], 机械工业出版社, 2003 年

[4]. 段云峰吴唯宁李剑威韩洁, 数据仓库及其在电信领域中的应用[M], 电子工业出版社, 2003年10 月

[5]. 郭芝懋, 数据质量和数据清洗研究综述, 软件学报[J], 2003, 13(11): 110-122.

[6]. 陈京民, 数据仓库原理、设计与应用[M], 中国水利水电出版社, 2004年4 月

[7]. Efrem G. Mallach, 决策支持与数据仓库系统[M], 电子工业出版社, 2001年4 月

[8]. 高寅, 系统分析之路[M], 电子工业出版社, 2005年5月

[9]. 汤子瀛等, 计算机网络技术及其应用[M], 电子科技大学出版社, 1997年10月

[10]. 齐治昌, 谭庆平, 宁洪, 软件工程[M], 高等教育出版社, 2001年

[11]. 彭木根, 数据仓库技术与实现[M], 电子工业出版社, 2002年6月

[12].王珊, 萨师煊, 数据库系统概论（第三版）[M], 高等教育出版社, 2001年[13]. ft东电信公司市场经营规范及管理办法, ft东电信内部文件

[14]. 何玉洁, 数据仓库与OLAP实践教程[M], 清华大学出版社, 2008年3 月

[15].陆泉, 陈静, 决策支持系统实验教程[M], 武汉大学出版社, 2008年6 月

[16]. 邵峰晶, 数据挖掘原理与算法[M], 科学出版社, 2008年08 月

[17].张喆, 数据挖掘在客户关系管理中的应用[M], 复旦大学出版社, 2002年6 月

[18]. 池太崴, 数据仓库结构设计与实施[M], 电子工业出版社, 2009年04 月

[19]. 谢邦昌, 数据挖掘Clementine应用实务[M], 机械工业出版社, 2008年04 月

[20]. 边耐政, JavaEE平台下Web应用开发[M], 机械工业出版社, 2008年05 月

[21]. 陈文伟, 数据仓库与数据挖掘教程[M], 清华大学出版社, 2006年08 月

[22].潘华, 数据仓库与数据挖掘原理、工具及应用[M], 中国电力出版社, 2007年1 月

致**谢**

四年前，当我得知能够继续我的学生生涯，特别是我大学时期一直向往的南邮来完成我的研究生学业时，喜悦、幸福、兴奋冲淡了一切。我不断的提醒自己，要珍惜这样的求学机会，一千多个日夜如白驹过隙，转瞬即逝。当论文最终定稿时，我心中有丝丝的轻松，又有缕缕的苦涩。我希望时间能走得慢一点，再慢一点。四年了，我收获的不仅仅是论文，还有更多更宝贵的东西。我深知，仅仅靠自己的努力是无法完成论文工作的，很幸运，也很感谢，有悉心指导和关心我的南邮老师们，有甘苦与共的同学、同事们，有默默支持和鼓舞我的家人们……

本课题在选题及论文成文过程得到了尊敬的导师肖甫老师和企业导师李玉华副总的悉心指导。两位老师治学严谨，学识渊博，思想深邃，视野雄阔。在撰写论文的过程中，肖老师多次询问、指导、循循善诱，为我指点迷津，帮助我开拓研究思路，精心点拨、热忱鼓励。肖老师不仅授我以文，而且教我做人，从恩师身上我体味到了丰富的学养、严谨的作风、求实的态度，勤奋的精神，这都成为了我不断前行的动力和标杆。

在论文即将完成之际，我的心情无法平静，从开始进入课题到论文的顺利完成，有多少可敬的师长、同学、朋友给了我无言的帮助，在这里请接受我诚挚的谢意！

最后，再次感谢南邮以及尊敬的老师们，感谢南邮给我提供这么好的一次机会完成了大学时期的梦想。我要向在百忙之中抽时间对本文进行审阅、评议和参加本人论文答辩的各位师长表示忠心的感谢！