基于地市级数据集市的结构与模块设计

张世红[1] 秦 浩[2]

(海南政法职业学院信息技术系 海口 571100)[1] (海南政法职业学院法务技术系 海口 571100)[2]

摘 要 根据地市级移动通信的实际需求,需要设计数据集市的层次结构,其结构由面向综合查询的数据层和面向明细查询的数据层构成,重点是对账户主题、业务量主题、竞争主题、用户主题、新业务主题、大客户主题等模块进行总体设计和主表设计。

关键词 数据集市,层次结构,主题模块

中图法分类号 TP311 **文献标识码** A

Designs of Structures and Modules Based on Local Data Marts

ZHANG Shi-hong[1] QIN Hao[2]

(Department of Information Technology, Hainan Vocational College of Political Science and Law, Haikou 571100, China)[1]

(Department of Law Technology, Hainan Vocational College of Political Science and Law, Haikou 571100, China)[2]

Abstract According to the actual needs of the local mobile communications, you need to design the hierarchy of data marts, and its structure consists of comprehensive query-oriented data, and data layers of detail-oriented queries, which focuses on the overall design and the primary table design for account theme, business theme, competition theme, user theme, new theme, major clients theme, such as modules.

Keywords Data marts, Hierarchy, Theme module

1 引言

建设地市级数据集市系统,旨在通过数据集市的建设,与原系统形成良好的互补,为地市分公司提供丰富、详细的数据,提高地市分公司市场经营分析水平。在地市级移动通信数据集市的建设中,数据集市的结构设计是其重要组成部分。为了充分呈现数据集市模型,需紧密结合地市级移动通信的实际需求,对层次机构进行系统设计。更重要的是,应从层次结构中提取出主要的主题模块,并对这些主题模块进行总体设计和主表设计。

2 数据集市的层次结构设计

数据集市的层次结构如图1所示,具体分层说明如下[1]:

(1)从地市级数据集市的基本分层结构可知,数据集市的数据主要由两部分数据组成:一是面向综合查询的数据层,二是面向明细数据查询的数据层。

(2)面向综合查询的数据层:本数据层为地市使用者提供一层基本完整的综合查询数据层,本层数据是为了便于最终用户使用而形成的综合查询数据表,基本是基于面向明细查询的数据层生成的数据,主要是为了提高用户综合查询的速度和效率,同时本层数据还包括根据实际经验组织的宽表数据。在本层数据模型设计上,将提供完整的综合查询数据层的模型结构。

(3)面向明细查询的数据层:本数据层为地市使用者提供一层基本完整的明细数据,本层数据基本以关系型为主,更接近原有的ODS层数据,是原有ODS层数据的扩充和完善。同时希望通过数据集市的本层数据,逐步完善数据仓库的明细数据层。本层数据在模型设计中将提供基本完整的参考模型,各省可根据本省的实际情况和现有数据仓库的情况加以调整。

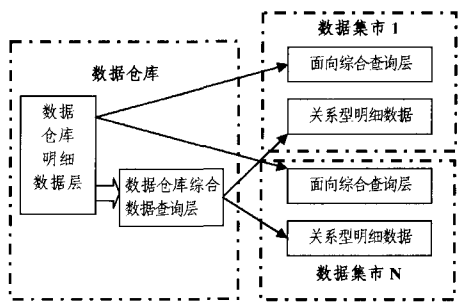

图1 数据集市的层次结构图

3 主题模块设计

由于地市级数据集市的功能模块较多,本文仅选取具有代表性的账户主题、业务量主题、竞争主题、用户主题、新业务主题、大客户主题等模块进行总体设计和主表设计。

3.1 账户主题设计

3.1.1 总体设计

账户主题主要是营业和账务出账的相关数据信息在数据

张世红(1978—),女,硕士,讲师,主要研究方向为计算机通信,E-mail:zhangshihong1978@163.com;秦 浩(1975—),男,硕士,副教授,主要研究方向为会计电算化。

集市的展示，在账户主题中，同样包含两部分数据[2]：一是底层关系型明细数据，二是便于地市用户查询的综合查询数据表。其设计原则是数据开放以底层最明细的数据为主，包括账单数据的开放，同时为了地市查询获取数据方便，将提供部分便于地市查询的综合查询数据表。

(1)子模块划分：账户主题按照数据信息内容分为账单、缴费、销账、账户4个子模块。

(2)粒度划分：在账户子模块中，数据的粒度为最明细的，所有数据表中最大粒度是账户级别。

具体子模块设计为：①账单子模块需根据账单的特点，按照数据粒度将本模块划分为账户账单、用户账单和用户明细账单，最小粒度是“用户＋科目”级别，最大粒度是账户级别；②缴费子模块的数据最小粒度是“用户＋账务月＋科目”，最大粒度是缴费记录；③账户信息子模块的最大粒度是账户级别，最小粒度是用户级别；④销账信息子模块的最大粒度是账户级别，最小粒度是用户账目级别。

(3)数据分割：由于账务相关信息表的数据量非常大，大省级的用户明细账单一个月的数据量基本达到8000万左右，因此账务相关表的分表原则都是“时间＋地市”，时间是在数据仓库内部的分表原则，地市是数据集市的分表原则。

(4)数据组织：账户子主题的数据以月数据为主，日数据除临时出账信息之外，其它全部为月数据。

3.1.2 主表设计

在账户主题中，主表设计包括：

(1)三级账单表。包括账户级、用户级以及用户明细级账单。在三级账单中为了数据处理以及查询访问方便，又把每一级账单分为3张主表：

当前月账单表：它是后付费用户的账单，当前月账单主要是指当月出账、抵销预存后形成的账单，本账单等待用户在下个缴费周期缴费。因此后付费用户不计算在欠费用户中，同时由于本账单和历史账单的意义不同，需单独存放，以便于数据处理和查询。包括：账户当前月账单表(DMD_ACCT_BILL_YYYYMM)、用户当前月账单表(DMD_ACCT_USERBILL_YYYYMM)和用户明细当前月账单表(DMD_ACCT_USERBILLITEM_YYYYMM)。

历史月账单表：它是后付费用户的账单，主要是指除当前出账月的账单之外的其它月账单，本部分账单是后付费用户欠费计算的主要来源表。包括：账户当前月账单表(DMD_ACCT_BILL_H_YYYYMM)、用户当前月账单表(DMD_ACCT_USERBILL_H_YYYYMM)和用户明细当前月账单表(DMD_ACCT_USERBILLITEM_H_YYYYMM)。

预付费用户账单表：是全球通预付费用户的账单，由于全球通预付费用户是实时扣费的，因此本部分账单和预付费用户的账单分开存放。包括：预付费用户账户当前月账单表(DMD_ACCT_PREBILL_YYYYMM)、用户当前月账单表(DMD_ACCT_USERPREBILL_YYYYMM)和用户明细当前月账单表(DMD_ACCT_USERPREBILLITEM_YYYYMM)。

(2)账务缴费表：主要包括缴费信息表、银行缴费信息表、个人资金变化情况表，具体解释如下：

缴费信息表：由两张主表构成，一张是缴费记录表，记录用户缴费时的记录信息，一笔缴费一条记录，另一张是缴费明细表，记录具体冲账的记录明细表。包括：用户缴费记录表(DMD_ACCT_BUSIREC_YYYYMM)记录账务缴费的欠费信息；用户缴费明细表(DMD_ACCT_BUSIFEE_YYYYMM)；银行缴费信息表(DMD_ACCT_BANKREC_YYYYMM)，本表同时在缴费记录表中有相应的记载；用户资金平衡记录表(DMD_ACCT_BUSIFUND_YYYYMM)，本表记录用户每笔账务资金的变化情况，包括每笔缴费以及每一笔的充销情况。

(3)账户相关信息表：主要包括全部账户和个人账户的相关信息情况表，包含：①账户资料信息表(DMD_ACCT_MSG_YYYYMM)，本表记录全部账户包括个人和公用账户的资料信息表；②非公用资金信息表(DMD_ACCT_PREPAY_YYYYMM)，纪录个人账户的全部资金信息，包括预付费和后付费的个人预存信息；③个人账户沉淀资金表，记录个人账户在离网或者其它异常离网情况下的账户资金情况。

(4)销账相关信息表：主要记录账务销账单相关信息表，主要包括销账户账单表和销用户账单表。①账户销账信息表(DMD_ACCT_DERREC_YYYYMM)，记录全部销账用户账单的数据信息；②用户销账信息表(DMD_ACCT_DERITEM_YYYYMM)，记录全部销账用户账单的数据信息。

3.2 业务量主题设计

3.2.1 总体设计

业务量主题主要是计费语音详单的相关数据信息在数据集市的展示，在业务量主题中，数据集市的开放数据主要包含：一是语音详单的相关数据，二是便于地市用户查询的由详单生成的综合查询数据层。在业务量设计中，数据按照粒度分为最明细数据层详单数据、业务量级粒度的明细表、用户级业务量粒度明细表3层数据。

(1)子模块划分：账户主题按照数据粒度分为详单级、业务量级和业务量明细级3大子模块。

(2)粒度划分：在业务量子模块中，数据的粒度为三级划分，所有数据表中最大粒度是用户级别，最小粒度是详单级别。

(3)数据分割：由于业务量的相关信息表的数据量非常大，因此业务量相关表的分表原则都是“时间(日)＋地市”或者“时间(月)＋地市”，时间和地市基本都遵循数据仓库内部的分表原则。

(4)数据组织：业务量子主题的数据以日数据为主，最终由日数据生成为月数据。

3.2.2 主表设计

在业务量主题中，主表包括：

(1)详单相关表：详单主要包括语音详单、漫入详单以及相对应的滞后详单，具体解释为：①语音详单表(CDR_CALL_XX_YYYYMMDD)，本表按照“日＋地市”的分表原则在数据仓库中存放，同时开放到数据集市；②滞后语音详单表(CDR_LATER_CALL_XX_YYYYMM)，存放每日滞后的语音详单；③漫入详单表(CDR_CALL_ROAMIN_YYYYMMDD)，记录其它省用户漫入到本省的漫游详单；④滞后漫入详单表(CDR_LATER_CALL_ROAMIN_YYYYMM)。

(2)业务量级别相关数据表：业务量级别相关表主要包括业务量明细表和呼转明细表。具体解释为：①业务量明细表主要记录用户业务量级的数据，例如用户的长途漫游等情况

的明细数据，是一个用户一天存放多条记录的信息表，包括语音话单用户业务量日统计表和语音话单用户业务量月统计表；②呼转相关明细表主要记录详单分离出来和呼转相关的明细数据信息，包括用户呼转日明细表和用户呼转月表。

(3)用户级业务量相关表：主要记录用户的业务量，一个用户一条记录，包含用户业务量日明细表和用户业务量月明细表。

3.3 竞争主题设计

3.3.1 总体设计

竞争主题主要是计费语音及短信详单拆分出来的信息和移动用户通话的竞争对手用户信息，在竞争对手主题中，数据集市的开放数据主要包含：一是竞争对手相关资料数据，二是竞争对手网间互通数据信息[3]。

(1)子模块划分：竞争对手主题按照数据类型分为竞争对手资料数据和竞争对手网间互通信息数据两部分；

(2)粒度划分：在业务量子模块中，数据的粒度为两级划分，所有数据表中最大粒度是竞争用户级别，最小粒度是详单级别；

(3)数据分割：由于竞争对手的相关数据表的信息都相对较小，因此数据分割基本按照月来划分；

(4)数据组织：竞争对手的数据以日数据处理为主，最终由日数据生成为月数据。

3.3.2 主表设计

在竞争对手主题中，主表包括：

(1)竞争对手相关资料表，主要记录竞争对手客户及大客户的相关信息资料。①竞争对手客户日表(DM_COMP_CUST_DT)，记录截至当日为止的竞争对手的用户资料；②竞争对手客户月表(DM_COMP_CUST_YYYYMM)；③竞争对手大客户资料月表(DM_COMP_VIPCUST_YYYYMM)，存放每月竞争对手的大客户资料。

(2)竞争对手网间互通相关信息表，主要包括网间互通日累计表(DM_COMP_OPPOSITE_DT)、网间互通月表(DM_COMP_OPPOSITE_YYYYMM)和竞争对手日业务量信息表(DM_COMP_ALL_YYYYMMDD)。

3.4 用户主题设计

3.4.1 总体设计

用户主题的数据集市是以用户为中心，包括用户的基本信息、相关客户资料、用户相关选择业务信息、用户终端信息、用户业务量、用户的新增、流失和变更等。用户数据包括日数据和月数据，同时提供了底层关系型明细数据和一些综合汇总数据。

(1)粒度划分：底层关系型明细数据基本保持和 BOSS 系统一致，汇总数据到用户级。

(2)数据分割：日数据每日一张表，月数据每月一张表。

3.4.2 主表设计

(1)用户资料月全量表(DMD_PRODUCT_YYYYMM)：存放到本月底为止，在网用户资料中不包括离网用户，来源由 BOSS 系统直接抽取得到；

(2)用户离网表(DM_PRODUCT_OFFLINE_MT)：包括目前为止所有离网用户情况，包括历史离网和本月内离网的用户情况。其功能包括：统计到上月底为止所有离网用户的情况，统计月初到今天用户离网情况，统计总离网用户，来源由变更历史表统计得到。

(3)用户积分相关表：①DMD_PRODUCT_SCORE_YYYYMM，本表记录了用户的积分总体情况，包括截止到目前为止的总积分、已兑换积分、累计奖励积分等；②DMD_PRODUCT_CHGSCORE_YYYYMM，本表记录了用户的积分月兑换情况；③DMD_PRODUCT_CHGSCOREPROM_YYYYMM，主要记录用户每月的消费积分情况，并进行积分统计。

(4)业务受理明细表(DMD_PRODUCT_CHG_DM_YYYYMM)：业务受理明细表，记录本月每天的用户单流水，包括业务办理等。

(5)客户资料：①客户月基本资料表(DMD_CUST_MSG_YYYYMM)，存放上月底的客户公共信息，包括普通客户(企业、个人)、个人大客户、集团大客户、VPMN 客户等，其它扩张信息在各自对应的扩展表中描述。其用途：查看上月底全部的客户最新资料；统计月流失、月新增的客户；观察每个客户在网、离网时长；结合 DWD_CUST_CHGMSG_YYYYMM_DM 可以统计到目前为止所有的客户情况；②客户资料最新变更信息表(DMD_CUST_CHGMSG_DM_YYYYMM)，保留一个月内每一天的客户资料变更情况。

3.4.3 汇总表设计

(1)业务量分时段月统计表(DM_QUERY_CALL_XX_YYYYMM)：分时段统计的业务量，数据源为语音详单用户业务量统计月表 DM_CALL_XX_YYYYMM。

(2)用户宽月表(DM_PRODUCT_QUERY_YYYYMM)，数据源为用户月表、用户业务量月明细表、月用户呼转汇总表、用户新业务月明细表、IMEI 终端用户业务信息表。

(3)大客户/集团客户宽月表(DM_VIP_ENT_PRODUCT_YYYYMM)，数据源为用户宽月表 DM_PRODUCT_QUERY_YYYYMM 和 DM_VIP_CUST_YYYYMM。

3.5 新业务主题设计

3.5.1 总体设计

新业务是按照新业务产品的业务类型或承载平台来划分，并且基本上采用和 BOSS 系统设计一致的思路进行。每一类产品都包含详单、延迟详单、日汇总和月汇总表。

(1)粒度划分：所有数据表中最大粒度是用户级别，最小粒度是详单级别；

(2)数据分割：分表原则都是“时间(日)+地市”或者“时间(月)+地市”，时间和地市基本都是根据数据仓库内部的分表原则。

3.5.2 主表设计

(1)IP 记账卡业务：IP 记账卡业务详单表(DR_IP_XX)，包括了 IP 直通车业务，用来统计 IP 记账卡用户每日使用情况，包括 CDR_IP_YYYYMMDD、CDR_LATER_IP_YYYYMM、DM_NEWBUSI_IP_YYYYMMDD 和 DM_NEWBUSI_IP_YYYYMM。

(2)智能网 IP 业务：CDR_PIP_YYYYMMDD、CDR_LATER_PIP_YYYYMM、DM_NEWBUSI_PIP_YYYYMMDD 和 DM_NEWBUSI_PIP_YYYYMM。

(3)企业 PBXVOIP 业务：CDR_PBX_YYYYMMDD、

(下转第 303 页)

$=0, c_2=1$ 时，颜色迁移过程完全保持目标图像分割区域的梯度信息与式(4)一致，当 c_1 和 c_2 在[0,1]之间变化时，颜色迁移结果在标准差和细节间过渡。

结束语 本文算法综合了图像的颜色和细节信息进行局部颜色迁移。实验结果表明，本文算法能够实现对目标物体保持边界的精确分割，实现了用户可调的保持细节的局部颜色迁移。与 Reinhard 算法相比，本文算法结果更为自然，特别是在区域细节保持方面，明显优于 Reinhard 算法的结果。

参 考 文 献

[1] Reinhard E, Ashikhmin M, Gooch B, et al. Color transfer between images[J]. IEEE Computer Graphics and Applications, 2001, 21(5): 34-41

[2] Rudeman D L, Cronin T W, Chiao C C. Statistics of cone response to natural images. Implications for visual coding[J]. Journal of Optical Society of America, 1998, 15(8): 2036-2045

[3] 赵国英，向世明，李华. 高阶矩在颜色传输中的应用[J]. 计算机辅助设计与图形学报，2004，16(1)：62-66

[4] 胡国飞，傅健，彭群生. 自适应颜色迁移[J]. 计算机学报，2004，27：1245-1249

[5] 向世明，赵国英，陈睿，等. 控向金字塔颜色传递[J]. 计算机辅助设计与图形学报，2005，17(5)：948-953

[6] Pitie F, Pan Z, Dong Z. A new algorithm for adding color to video or animation clips[J]. Proceedings of WSCG-International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, 2004, 12(3): 515-519

[7] Vieira L F M, Vilela R D, Nascimento E R, et al. Automatically choosing source color images for coloring grayscale images[C]// SIBGRAPI, 2003. 2003

[8] 孟敏，刘利刚. 勾画式局部颜色迁移[J]. 计算机辅助设计与图形学学报，2008，20(7)：838-842

(上接第 283 页)

CDR_LATER_PBX_YYYYMM、DM_NEWBUSI_PBX_YYYYMMDD 和 DM_NEWBUSI_PBX_YYYYMM。

(4) WLAN 业务：CDR_WLAN_YYYYMMDD、CDR_LATER_WLAN_YYYYMM、DM_NEWBUSI_WLAN_YYYYMMDD 和 DM_NEWBUSI_WLAN_YYYYMM。

(5)语音增值业务：CDR_ADDVALUE_YYYYMMDD、CDR_LATER_ADDVALUE_YYYYMM、DM_NEWBUSI_ADDVALUE_YYYYMMDD 和 DM_NEWBUSI_ADDVALUE_YYYYMM。

(6)梦网业务：CDR_ISMG_XX_YYYYMMDD、DM_NEWBUSI_ISMG_YYYYMMDD 和 DM_NEWBUSI_ISMG_YYYYMM

(7)短信业务：CDR_SMS_XX_YYYYMMDD、DM_NEWBUSI_SMS_YYYYMMDD 和 DM_NEWBUSI_SMS_YYYYMM。

(8)彩信业务：CDR_MMS_YYYYMMDD、DM_NEWBUSI_MMS_YYYYMMDD 和 DM_NEWBUSI_MMS_YYYYMM。

(9)WAP 业务：CDR_WAP_YYYYMMDD(本表定义了 WAP 业务的详单，包括 WAP 服务、彩铃、PDA 等)、DM_NEWBUSI_WAP_YYYYMMDD 和 DM_NEWBUSI_WAP_YYYYMM。

3.6 大客户主题设计

3.6.1 总体设计

大客户系统设计有两层结构，一层是大客户的明细级数据，一层是大客户汇总级别数据，包括大客户的异动情况、客户月发展情况等等。

(1)粒度划分：所有数据表中最大粒度是用户级别，最小粒度是日流水。

(2)数据分割：分表原则都是“时间(日)+地市”或者“时间(月)+地市”，时间和地市基本遵循在数据仓库内部的分表原则。

3.6.2 主表设计

(1)大客户流水日表(DMD_VIP_CUST_DM_YYYYMM)：保留当前月的大客户新增、流失和所有到目前为止的在网大客户。其用途主要有观察本月流失、新增大客户、观察每一天的大客户新增流失等。

(2)DMD_VIP_INFO_YYYYMM：存放所有大客户卡信息，包括历史、在用、注销。

(3)DMD_VIP_MANAGER_YYYYMM：记录大客户和客户经理的对应关系。

(4)DM_VIP_CUST_YYYYMM：大客户基本信息表，本表是大客户所有分析的基础表，包括所有在网大客户和到当前月为止的离网大客户。其用途主要有：统计当月的所有大客户基本情况；统计在网大客户情况；统计离网大客户情况、离网时长等。

(5)DM_VIP_CALLFW_YYYYMM：本表记录了大客户中月异动用户的详细明细信息、由大客户截至当日的主被叫比例及大客户业务量比例生成大客户月异动的异动组合类型维，其用途包括统计大客户异动和统计异动大客户的业务量。

结束语 建设一套完善的数据集市系统是一个长期的过程，因此本系统只是数据集市系统设计的一个探索和开端。随着数据大量增加和对深层信息的挖掘需求，数据集市的应用需要进行不断的完善，以满足用户不断变化的需要。同时个性化的数据不断增多，数据质量问题也应引起重视。另外，在数据展现部分，数据仓库中原有的 OLAP、数据挖掘和客户细分等功能也会不断地引入系统中，以便提供更深层次的数据，为决策者提供强有力的信息。

参 考 文 献

[1] 陈奕新. 无线网络中的数据集市原型设计与验证[D]. 北京：北京邮电大学，2008

[2] Chenoweth T, Corral K, Demirkan H. Seven key interventions for data warehouse success[M]. Communications of the ACM, Aug. 2006

[3] Chenoweth T, Schuff D, Louis R S. A method for developing dimensional data marts[M]. Communications of the ACM, Dec. 2009