超市销售数据仓库的规划与分析

背景： 某大型连锁超市的业务涵盖于3个省范围内的1000多家门市。每个门市都有较完整的日用品 和食品销售部门，包括百货、杂货、冷冻食品、 奶制品、肉制品和面包食品等，大约5万多种，其中大约45000种商品来自外部生产厂家，并在包装上印有条形码。每个条形码代表了唯一的商品。

需求： 为该超市建立一个能够提高市场竞争能力的数据仓库，首先需要进行数据仓库的规划分析。涉及到对数据仓库的需求分析、模型构建两个过程。

1 超市销售数据仓库的需求分析

（1）超市营销售策略分析

- 超市最高层管理所关注的是如何通过商品的采购、储存与销售，最大限度地获取利润。需要通过加强对每种商品的管理，减低商品的采购成本和管理费用，吸引尽可能多的客户。其中最重要的是关于商品促销的管理决策。需要依靠合适的促销活动，应用适当的促销策略针对合适的客户，以增加超市的销售利润，是超市数据仓库建设的基本需求。
- 超市不同商品的销售利润是有差别的。希望在数据仓库中通过对商品的赢利分析，了解不同商品的销售赢利状态，以确定企业的销售重点，对那些可以为企业带来较大赢利的商品加大促销力度。

（2）超市商品库存分析

- 超市商品的库存状况对超市的利润具有巨大的影响。超市如果能够在合适的时候销售合适的商品，在不出现脱销的情况下尽可能减少商品库存的库存成本，是超市商品库存分析的主要目的。在商品库存分析中，管理人员还经常要根据商品的库存量和商品库存成本确定商品的销售价格。从超市的商品库存情况来看，库存分析实质上是对超市的价值链进行分析，分析商品库存在超市的整个价值链上所发挥的作用。

（3）超市商品采购分析

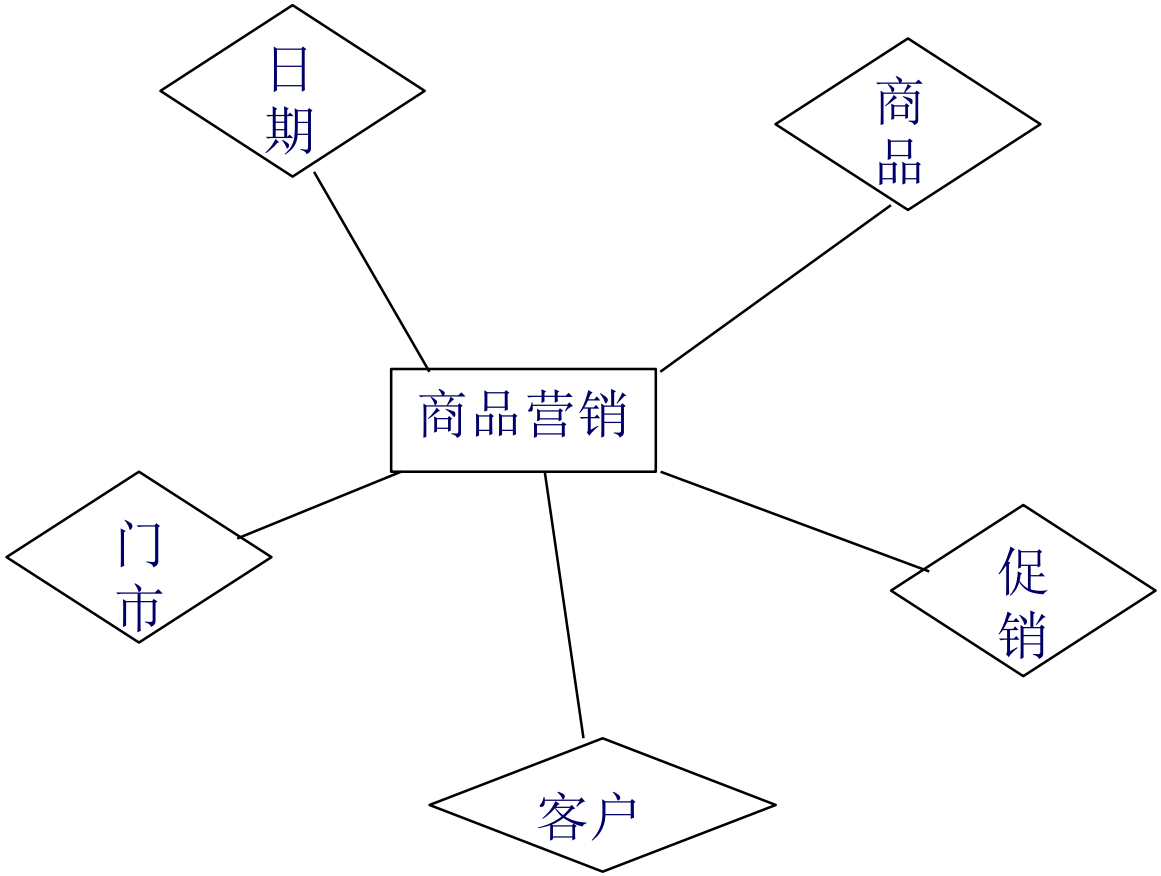
- 超市在商品采购工作中需要分析哪些商品是热销的商品，尽可能采购销售热销商品。热销商品往往是加快企业资金流动的动力，快速流动的资金可以使企业在一定的时间内取得比其他企业更多的利润。而且超市营销管理人员在了解热销商品后，可以大量采购热销商品，重新安排热销商品的货架，向更多的客户推销热销商品，便于更多客户的购买，以进一步加快企业资金的流动。

（4）超市客户关系分析

- 用80：20理论分析，占企业客户群20%左右的客户购 买金额往往占据了企业销售金额的80%。对客户群体的划分有利于企业了解企业的主要客户群体状况、主要客户群对企业销售服务的需求状况、不同客户群为企业所带来的利润状况。
- 在对客户进行类型划分的基础上，可以针对不同客户群体的特点采用不同的营销策略，对客户群体的消费进行合理的引导。
- 超市客户的流失，意味着企业赢利的降低。企业管理者希望了解哪些客户可能会流失，使企业能够提前设法加以挽留。

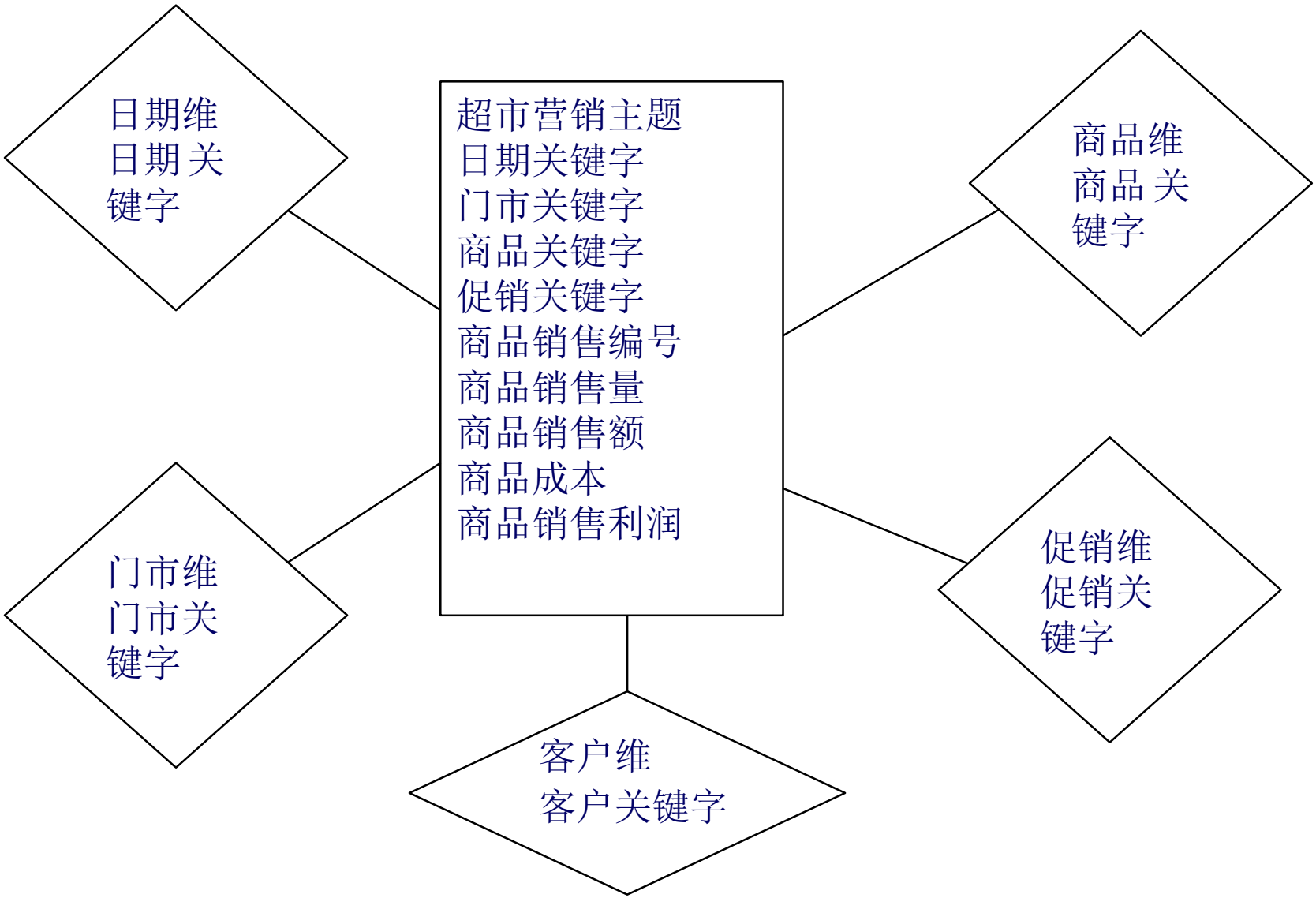
2 超市销售数据仓库E-R模型构造

- 数据仓库设计中就首先考虑营销主题的设计，确定超市营销主题模型。
- 采用了星型模型，没有采用雪花模型。因为雪花模型通过对维表的分类细化描述，对于主题的分类详细查询具有良好的响应能力。雪花模型的构造在本质上是一种数据模型的规范化处理，会给数据仓库操作带来不同表的连接困难 。但是在对维度表进行维护时，可能需要对大量重复值进行修改。
- 星型模型通过对维表的冗余应用，以牺牲维表空间来换取数据仓库的高性能与易使用的优势。


日期
商品
商品营销
门市
促销
客户

3 超市数据仓库事实表模型

- 确定在数据仓库中，怎样的粒度数据才能满足管理人员对数据仓库营销策划分析的需要。
- 超市商品销售主题中，最理想的原子数据是来自POS机上的每个销售事务数据。
- 分析超市高层管理人员通过那些角度，即需要通过那些维度来考察、选择营销方案。一般情况下，在确定超市营销策划时，超市管理人员需要通过日期、商品、门市、促销和客户五个维度对促销方案进行分析，了解促销方案的可用性和效果。


日期维
日期关
键字
超市营销主题
日期关键字
门市关键字
商品关键字
促销关键字
商品销售编号
商品销售量
商品销售额
商品成本
商品销售利润
商品维
商品关
键字
门市维
门市关
键字
促销维
促销关
键字
客户维
客户关键字

- 从销售系统中可直接获取商品销售量、销售单价、商品成本。但管理人员考察超市的营销策略时，需要考虑营销策略和相应的商品销售利润。商品销售利润可以直接通过商品销售量、销售单价和商品成本计算获得，但商品销售利润具有良好的可加性，管理人员又经常需要查看。将利润数据存放在事实表中可大大减少数据仓库工作时的工作量，还可以保证所有用户在使用商品销售利润这一重要数据时的一致性。
- 商品销售单价对于计算商品利润十分重要，但将某个商品一段时间内的所有销售单价相加是毫无意义的。管理人员可能只对某一时间段内某个商品的平均销售价感兴趣。平均销售价格可以用该时间段内的商品销售额除以商品销售量获取。在事实表中可以不用商品销售单价，代之以商品销售额，销售额也常常是管理人员衡量营销策略好坏的重要指标。

- 超市管理者还可能对商品销售的利润率感兴趣，该数据可以用商品销售利润除以销售额获得，该数据不是一个可加数据。将比率或百分数的数据进行相加，所获得的数据是没有什么意义的。管理人员在了解某一时期某些商品的利润率时，完全可以利用该时期该商品利润和销售额获得。

因此，事实表中确定度量数据为商品销售量、 商品销售额、商品成本和商品销售利润。

4 超市数据仓库维表模型设计

(1) 日期维

- 日期维模型是许多数据仓库应用中的常用维度，其设计方式与其他多数维模型有差别。具体设计时，日期维可以存放以日期表示的5到10年的数据行，也可以将3至4年的数据行作为日期维内容。如果对10年的每一天都进行存储，也只需要3650行。
- 日期维的每列由行所代表的特定日期进行定义。“星期”列含有像“星期一”这样的名称内容，该列可用于创建比较“星期一”与“星期日”销售情况对比的查询。日历日期编号从1开始取值，然后根据月份的情况取到28、29、30或者31，这一列主要用于对每个月的同一天进行比较。同样，可以给出日历周编号、和日历月编号(1，…，12)。

- 纪元表示法采用从某纪元开始连续对日期进行计数的方法给出日编号，在表中还可以给出“星期”与“月份”的绝对编号列。这些数据支持跨年度跨月份的简单数据运算。在生成报表时，经常要给出像“一月”这样的月份名称。因此，为报表确定一个“年月”(YYYY-MM) 列标题也有必要。报表中很可能需要季度编号(Q1，…， Q4)或年季度编号列。如果企业的财政年度与日历表在周期上不一致，还需要为财政年度给出类似列。
- 在“节假日”列中给出“节假日”或者“非节假日”的内容，维表属性作为数据分析的导航，简单地在“节假日”列中给出“Y”或者“N”对数据分析没有多大用处。 例如，在生成某种商品的节假日与非节假日销售情况比较查询时，列中给出“节假日”或者“非节假日”这样有意义的值要比一个简单的“Y”或者“N”之类的值有用得多。

- “星期六”与“星期日”要归入“周末”之列。当然，可以对多个日期表属性进行共同约束，从而能够实现一些像平日假期销售与周末假期销售进行比较的数据仓库应用。
- “销售时节”列应设置为销售时节的名称，例如，春节、情人节、端午节、五一节、国庆节、中秋节、重阳节、圣诞节、或者标为“不是”。
- “重大事件”列与“销售时节”列情形类似，可以标记为“周日大采购”或者“中秋合家欢”这样与日期有特殊联系的促销事件。而一般性的促销活动通常不放在日期表中处理，以促销维表的形式进行更加完整的描述。因为促销事件并不是仅仅由日期来定义，通常还需要由日期、商品与商店的组合来定义。


日期维度
日期关键字
星期
日历日期编号
日历周编号
日历月编号
纪元日编号
纪元周编号
纪元月编号
财政月日编号
年度日历周数
年度日历月数
日历年月(YYYY-MM)
日历季度
日历年季度
日历半年度
销售事实
日期关键字
商品关键字
门市关键字
促销关键字
客户关键字
POS事务编号
销售量
销售额
成本额
利润金额

日历年
财政周
年度财政周数
财政月
年度财政月数
财政年月
财政季度
财政年季度
财政半年度
财政年
节假日指示符
星期指示符
销售时节
重大事件
……

(2) 商品维

- 一般超市门市可能存储60000个商品编号，但大型连锁超市保留不再销售的历史商品营销方案情况，商品维度可能至少需要150000行乃至多达百万行。
- 商品维度数据主要来源于业务系统的商品主文件。超市总部对所销售商品的主文件进行统一管理。
- 商品主文件的一个重要作用，就是维护每个商品存储标志的许多描述属性。商品维是一组重要的属性。
- 某个商品种类包含多个商品子类，商品子类包含多个商标，商标包含多个商品存储标志。
- 还应包含描述商品形状或存储位置的层次属性，例如商品的包装类型、包装尺寸、包装数量、托盘中的包装数，以及与商品存储的层次：存储类型、货架结构等维度。

(3) 门市维

- 门市维表用于描述超市的各个链锁店。门市维表是基本的地理维度，每个门市可被看成一个位置。这样，可以由门市形成诸如街道、邮政编码、县、市、省这样的任意地理属性。地理体系与门市地区体系对每个门市来说，都有良好的定义。
- 在连锁超市所使用的门市维表中有建筑面积、金融服务、最早开业时间等描述特定门市的文字描述。描述销售面积的列应该是数字型的，并且在理论上是跨门市可相加的，以表示某一地区的销售面积。它是门市的一个不变属性，通常作为报表约束或者行标题使用。而且为了能够分析不同种类商品对超市销售利润的贡献情况，还需要设立不同商品的销售面积。

(4) 促销维

- 超市的促销方案可能包含：临时降价、柜台展销、报纸广告与优惠券发放等。促销维应该可以反映商品促销方案的成效。
- 促销的成效评估因素：促销商品的销售是否在促销区间出现增长、是否在促销进行之前或者随后出现减少状况；是否发生促销商品的销售出现增长，而临近货架上的其他商品销售却呈现出相应的降低情况(同类相 食)；促销类别中所有商品的销售是否都经历了一个实际的总体增长；促销是否赢利。促销利润的计算要考虑促销类别的利润增量与时间过渡、同类调剂以及销售底线等各种情况。

- 在促销维度中为促销出现的每种组合都建立一行记录是很有意义的。在一年的销售活动中，可能出现1000个广告，5000次临时降价和1000次柜台展销，但可能只有10000个组合促销能影响任何特定的商品。例如，在某给定维度中，大多数门市都会同时运作所有促销鲁手段，而只有少数几个门市不进行柜台展销。在这种情况下，就需要两个单独的促销记录行，一个用于通常的降价并外加广告与柜台展销，而另一个用于降价并外加单纯的广告。
- 超市的促销维度可以包含促销名称、减价类型、促销媒体类型、广告类型和优惠券类型等。超市的主要促销方式是降价、广告、柜台展销与优惠券。如果将这些因素分别建立促销维度，就可以记录分析这些促销方法非常相似的信息，使用户更加容易理解促销方案的作用。但是将所有的促销因素合并在一个维表中，则能够方便用户的浏览，能够弄清各种不同的价格降低、广告、展销与优惠券是如何在一起共同发挥促销作用的。

(5) 客户维

- 超市的客户维度可以包含客户账号、姓名、地址、所在地区、 邮政编码、电子信箱、电话、日常活动范围、出生日期、收入、孩子数量、住房和汽车等内容。在客户维中的地址由于客户可能会给出其家庭地址、工作地址或其它一些常用地址，因此在维表中可以设置4个地址，对于电话的设置也是出于相同因素的考虑。在数据仓库的应用中有时需要对客户按照不同的地区进行分析，为此，在维表中就按照省、市、县（区）邮政编码进行地区的设置。性别、婚姻状况、家庭人口、住房条件和自有汽车情况均是超市销售管理人员对超市营销策略进行分析的主要依据。出于超市营销策略制定的考虑，还需要了解客户的日常活动范围，以便有针对性地进行促销广告的发送。

5 超市数据仓库模型的关键字设计

- 采用代理关键字技术，而不是依赖业务系统中的各种关键字（许多业务系统中的各种编码往往具有某种特定的含义）
- 代理关键字一般采用在填充维度时按需要而顺序分配的整数值。例如，为第一条商品记录分配一个值为1的商品代理关键字，第二条分配2，第n条分配n等。代理关键字仅仅用于维度表到事实表的连接。
- 代理关键字的好处还能够对数据仓库环境的操作型变化进行缓冲，不会受到商品编码生成、更新、删除、再生与重用等操作型规则的妨碍。代理关键字允许数据仓库对来自多个业务型系统的数据进行合并，即使它们之间缺乏一致的源关键字也无所谓。

- 使用代理关键字还可以获得性能上的优势。代理关键字可能只有一个整数所占据的空间大小，却能确保充裕地容纳维度行以后可能需要的序号或者最大编号。而业务型编码常常是一个混合了字母与数字的区间编码体系。
- 代理关键字还能够用于记录那些诸如“不在促销之列”这样的可能没有业务系统中编码的维度情形。通过对数据仓库的关键字施加控制，就能够做到不管是否缺少业务型编码，总可以分配一个代理关键字将这类情况标识出来。
- 将代理日期关键字处理成日期序号，可以允许事实表在日期关键字基础上进行物理分区。

- 目前在超市数据仓库中已经包含了6个实际的表：营销事实表与日期、商品、门市、促销和客户维表。每个维表有一个主关键字，而事实表除了有一个退化的销售事务编号之外，还有由五个外关键字组成的一个复合关键字。如果五个关键字都是进行了紧凑处理的连续整数，那么仅仅需要为所有五个关键字保留18个字节的小存储空间(日期、商品、促销和客户维各用4个字节，而门市用2个字节)。同时，销售事务编号可能另外需要8个字节。
- 如果事实表4类事实（销售量、销售额、成本和利润）中的任何一个都是4字节的整数，则仅仅需要再保留另外的16个字节，这样事实表只有42个字节宽。对一个10亿行的事实表也只占用大约42GB的存储空间就可以存储所有事实数据。

6 超市数据仓库元数据设计

销售主题元数据

名称	Sales
描述	整个超市中每个门市中每个POS机所记载的商品销售状况
目的	用于进行超市销售状况和促销情况的分析
联系人	各个门市销售经理
维	时间、商品、客户、商店、促销
事实	销售事实表
度量值	销售成本、销售额、销售利润、销售量

名称	Sales_Fact_年份
描述	记录每个门市每个POS机所发生的销售数据
目的	作为销售主题的分析事实
使用状况	每天平均查询次数
	每天平均查询返回行数
	每天查询平均执行时间（分钟）
	每天最大查询次数
	每天查询返回最大行数
	每天查询最大执行时间（分钟）
存档规则	每个月将前36个月的数据存档
存档状况	最近存档处理日期
	已经存档数据日期
更新规则	每个月将前60个月的数据从数据仓库中删除
更新状况	最近更新处理日期
	已更新数据日期

销售事实元数据

	数据质量要求及确认	由于从各个门市POS机上所产生的数据可能会由于极少的人工输入，而使数据质量不能得到保证，但也真实地反映了销售现状，不能随意修改，应被认可。
	数据准确性要求	必须百分百地反映各个门市销售状况
	数据粒度	要求能够反映每一项商品的销售状况，不对数据进行汇总
	表键	事实表的键是时间、商品、客户、商店和促销维中键的组合
	数据来源	超市销售业务系统中的销售表(sales_fact_年份)
	加载周期	每天一次
	加载状况	最后加载日期
		加载的行数
	加载规则	每天清晨3:00将各个超市门市中前一天的销售事实数据拷贝到本表，拷贝过程中要根据各个数据成员所定义的加载规则进行筛选和清理

维元数据

	名称	客户（Customer）
	定义	从超市任何一个门市购买货物的任何个人或组织都称为客户，一个客户可以与多个销售地区发生联系(即出现在地理维的不同层次体系中)
	层次结构	一个客户的数据可以在3个级别上进行统计：最低级别是出现在客户所在的县/区，其上为市、省
	更改规则	新的客户位置作为新的一行插入维中。对已有位置的修改，则在原处更新
	加载频率	每天一次
	加载统计数据	最后加载日期
		加载的行数
	使用的统计数据	每天平均查询个数
		每天查询返回的平均行数
		每天查询平均执行时间（分钟）
		每天最大的查询个数
		每天查询返回的最大行数
		每天查询执行的最长时间（分钟）

维元数据

存档规则	每个月将前36个月的数据存档
	已经存档数据日期
更新规则	每个月将前60个月的数据从数据仓库中删除
更新状况	最近更新处理日期
	已经更新数据日期
数据质量	增加一个新客户时，先检查是否已在其他地方和该客户做过交易。少数情况下，由于检查失败，会将一个客户的不同部门作为不同客户保存。直到客户注意到在不同的地方与公司交易时，以前的记录仍保持不变。地区属性并不是销售业务系统原有的，而是根据送货地址属性中的邮政编码进行区分
数据的准确程度	一个客户与其地理位置的关联出错的可能性在某一百分比以下，该百分比大小要根据对业务数据的研究情况确定
关键字	客户维的关键字是系统产生的数字

维元数据

产生关键字的方法	从销售业务系统中拷贝一个客户时，将检查转换表，检查该客户是否已经存在于数据仓库中。如果否，就产生一个新的关键字。然后将这个关键字和销售业务系统中的Custom~ID和地区ID插入转换表中。如果该客户和位置已经存在于转换表，就根据表中的关键字决定数据仓库中要更新的记录
源表名称	超市销售业务系统中的Customer表
加载规则	每天拷贝每个Customer表中的行。对于已存在的客户，进行更新。对于新客户，确定其所在地理位置之后，产生一个关键字，然后插入一行新记录。在更新／插入操作之前，需要检查是否有重复的客户名。如果有，则在客户名后增加一个顺序号，直到名字以及名字和顺序号的组合都没有重复为止。
加载规则	只选择新的和发生变化的行
源表名称	Customer_Location表
转换规则	每天拷贝一次Customer_Location表。对于已存在的客户，更新其送货地址；对于新的客户，则产生一个键，并插入一行。

数据成员元数据

名称	客户关键字（Customer_ID）
定义	用以唯一标识客户和位置的值
更新规则	一旦分配，就不改变
数据类型	数值型
值域	1—999，999，999
产生规则	由系统自动产生，将当前最大值增1
来源	系统自动生成

数据成员元数据

名称	客户名称（Customer_Name）
定义	客户的名称
更新规则	客户名称发生改变时，就在原来的记录上更新
数据类型	Char(30)
值域	保证能区分不同客户的名称。对不同而具有相同名称的客户，可在名称后依次加1来区分相同名称
来源	超市销售业务系统中Customer表中的Name
产生规则	对于零售客户，其名称由姓和名组成。对于公司，则将公司名作为客户名称

思 考 题

1 如上的设计方案有哪些优点？

2 如上的设计方案有哪些不足？

3 在某部分你的可替代的方案或思路是？

4 从以上的数据仓库设计和规划过程你得到的经验、教训或收获是？

讨 论 题

1 你的课程设计的难点或问题？

2 如上案例，你可借鉴什么？