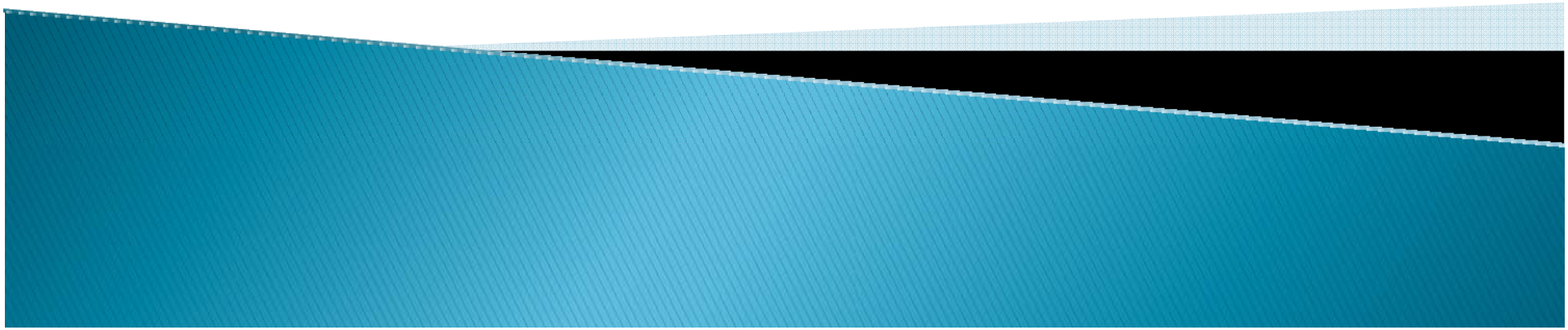
中国石油大学（北京）

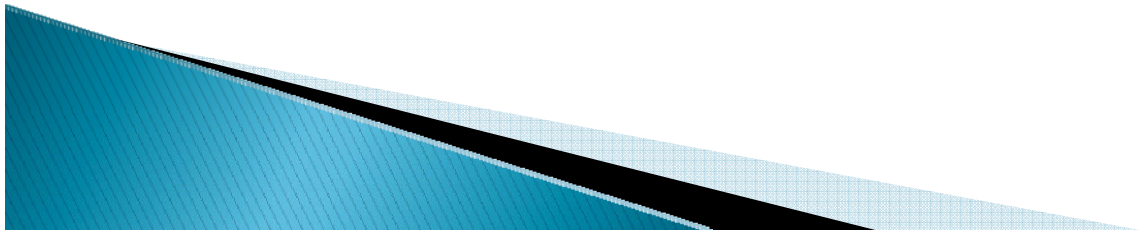
数据仓库应用实现

计算机　王莹



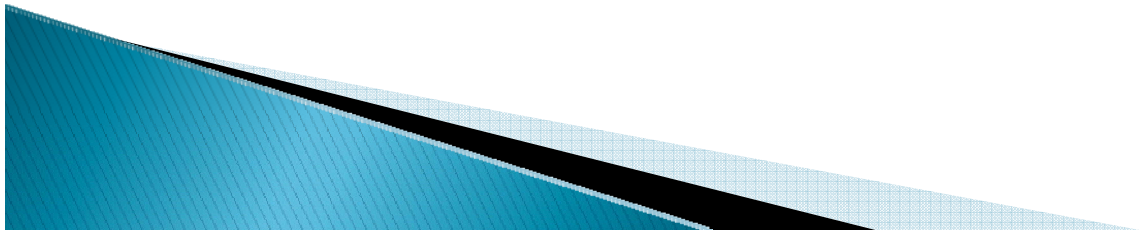
1. 实现工具

- 本例采用的是SQl Server2005所提供的商业智能服务和工具，主要包括Analysis Services(分析服务)，Integration Services(集成服务),Reporting Services(集成服务)和Bussiness Intelligence Developer Studio(BIDS)。



1. 实现工具(续)

- 分析服务(Analysis Services)
- SQL Server 分析服务（SSAS）是一个用于分析数据仓库中数据的工具，它包括了OLAP和数据挖掘工具。在SQL Server 2005数据库系统中，Analysis Services工具以服务器的方式为用户提供管理多维数据立方体的服务。Analysis Services可以把数据仓库中的数据组织起来，经过预先的聚集运算，加入到多维立方体中（即建立立方体），然后对复杂的分析型访问做出迅速的回答。

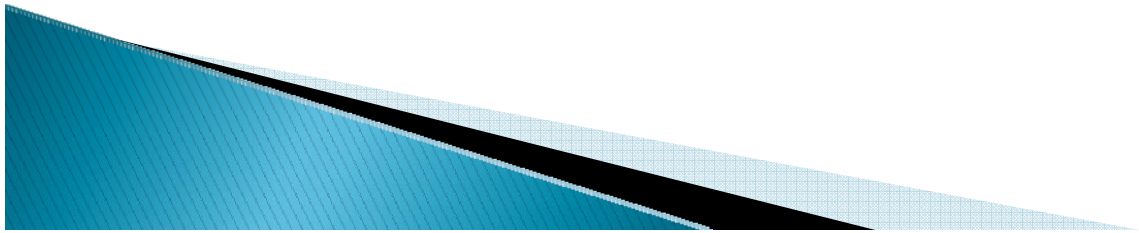


1. 实现工具(续)

- 集成服务(Integration Services)
- SQL Server 集成服务（SSIS）被定位成一个能生成高性能数据集成解决方案(包括数据仓库中数据的提取、转换和加载（ETL））的平台。其集成的含义主要就是指把ETL集成在一起。SSIS通过一个统一的环境向用户提供了数据转换服务（DTS）所能提供的所有功能，并且大大减少了用户花在编写程序和脚本上的精力和时间。
- SSIS的基本功能包括：
  - 合并来自异类数据源中的数据
  - 填充数据仓库和数据集市
  - 整理数据和将数据标准化
  - 精确和模糊的查找功能
  - 将商业智能置入数据转换过程
  - 使管理功能和数据加载自动化

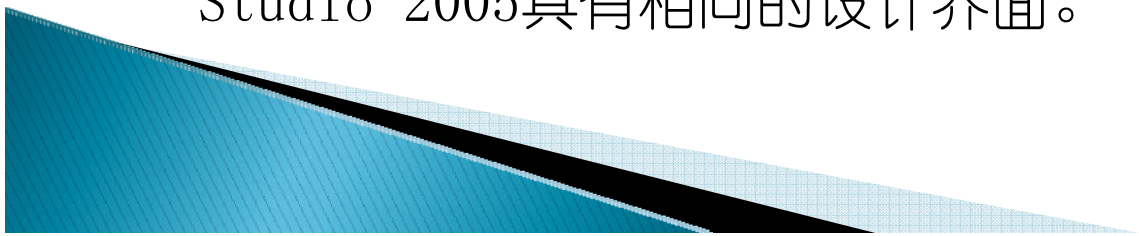
1. 实现工具(续)

- 报表服务(Reporting Services)
- SQL Server报表服务(SSRS)是一个完整的、基于服务器的平台，它可以建立、管理、发布传统的、基于纸张的报表或者交互的基于Web的报表。
- SSRS提供的主要功能有：
  - 为各类客户，跨企业提供并发访问功能
  - 为各类提供数据源支持
  - 针对个人和企业提供提供不同的数据报表分发机制
  - 生成各类形式的报表
  - 可生成多维数据报表，在此基础上可以进一步完成数据分析 工作，是真正的企业级报表生成工具。



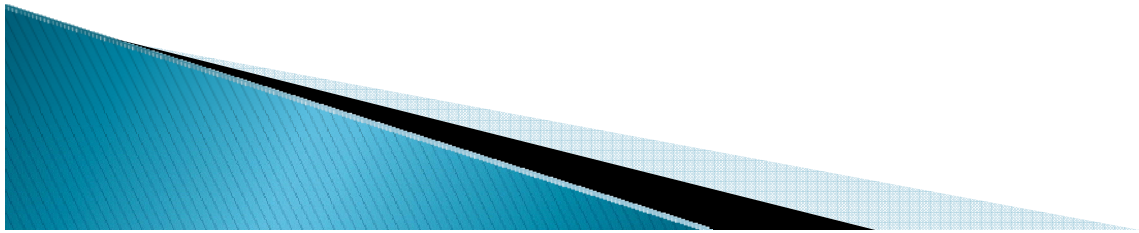
1. 实现工具(续)

- Bussiness Intelligence Developer Studio
- BIDS是SQL Server 2005新增加的一个开发环境，主要用于商业智能解决方案的开发。BIDS将开发商业智能所涉及的各个方面（例如数据转换和抽取、基于多维数据集的联机分析、数据挖掘和生成数据报表等）都集成在了一个开发平台上，也就是说商业智能开发人员可以使用BIDS开发出完整的商业智能解决方案。
- BIDS是一个基于Visual Studio 2005的开发平台，与Visual Studio 2005具有相同的设计界面。



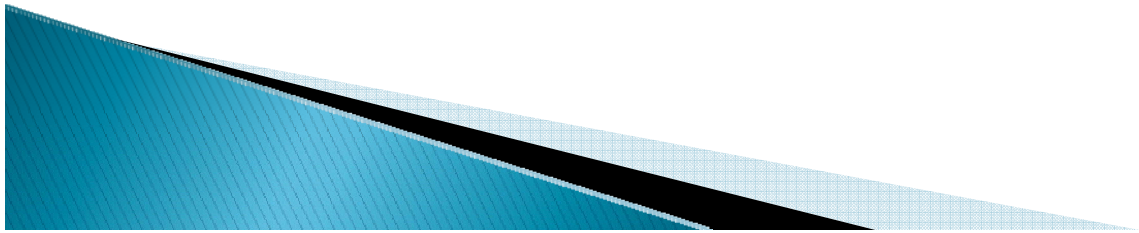
二. 数据仓库应用举例

- 1. 数据源概述
- 本例以SQL Server 2005提供的AdventureWorks数据库作为商业智能解决方案的数据源。
- AdventureWorks数据库是SQl Server 2005的范例数据库，它是一个大型的跨国自行车制造企业应用的业务数据库，其用途是帮助企业对自行车的生产和销售进行管理。
- AdventureWorks数据库主要的应用方面有人力资源、产品管理、市场销售、采购和供应商管理、生产管理。
- AdventureWorks数据库是一个比较复杂的数据库，可以使用Microsoft SQL Server Management Studio打开这个数据库，并查看其中的表格和字段。



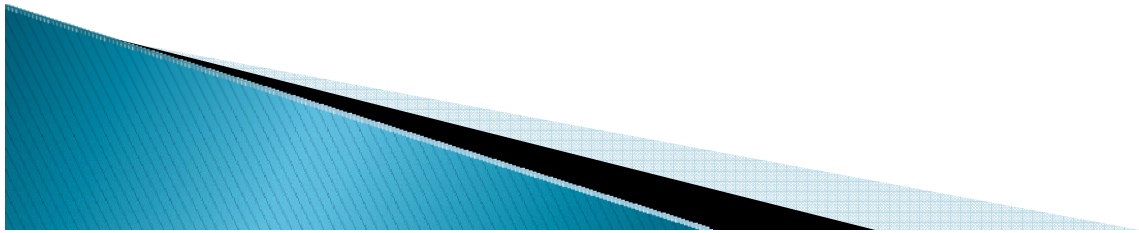
二. 数据仓库应用举例（续）

- 2. 需求分析
  - AdventureWorks数据库设计的方面很多，但是我们的目标很简单，只有以下三个：
    - 需要分析不同类别的产品通过直销在不同地区、不同时间段内销售的业绩。
    - 生成分析结果的报表。
    - 分析影响客户所有车的数量的因素。



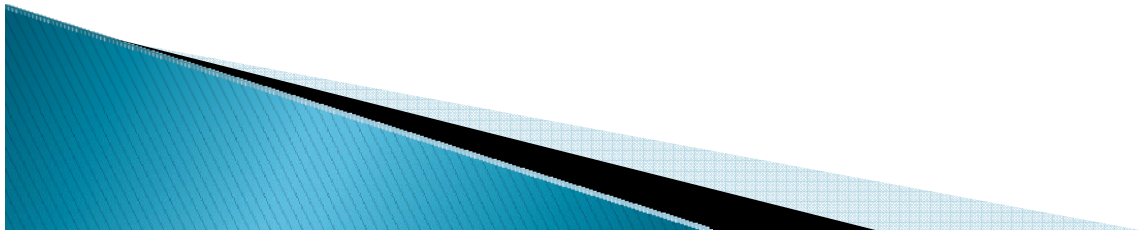
二. 数据仓库应用举例（续）

- 通过目标可以发现分析销售业绩基于的维度有三个：产品、客户和时间，事实数据则为反映销售业绩的订单。
- 对于产品我们关心的是产品的名称和分类，由于产品和产品类别之间有一对多的关系，因此可以将这个维度设计为雪花模型。
- 对于客户，我们主要需要关心客户的姓名、年龄、性别、婚姻状态、孩子的状况、是否拥有房产、拥有汽车的数量，所在的地区、国家、省和城市等信息。



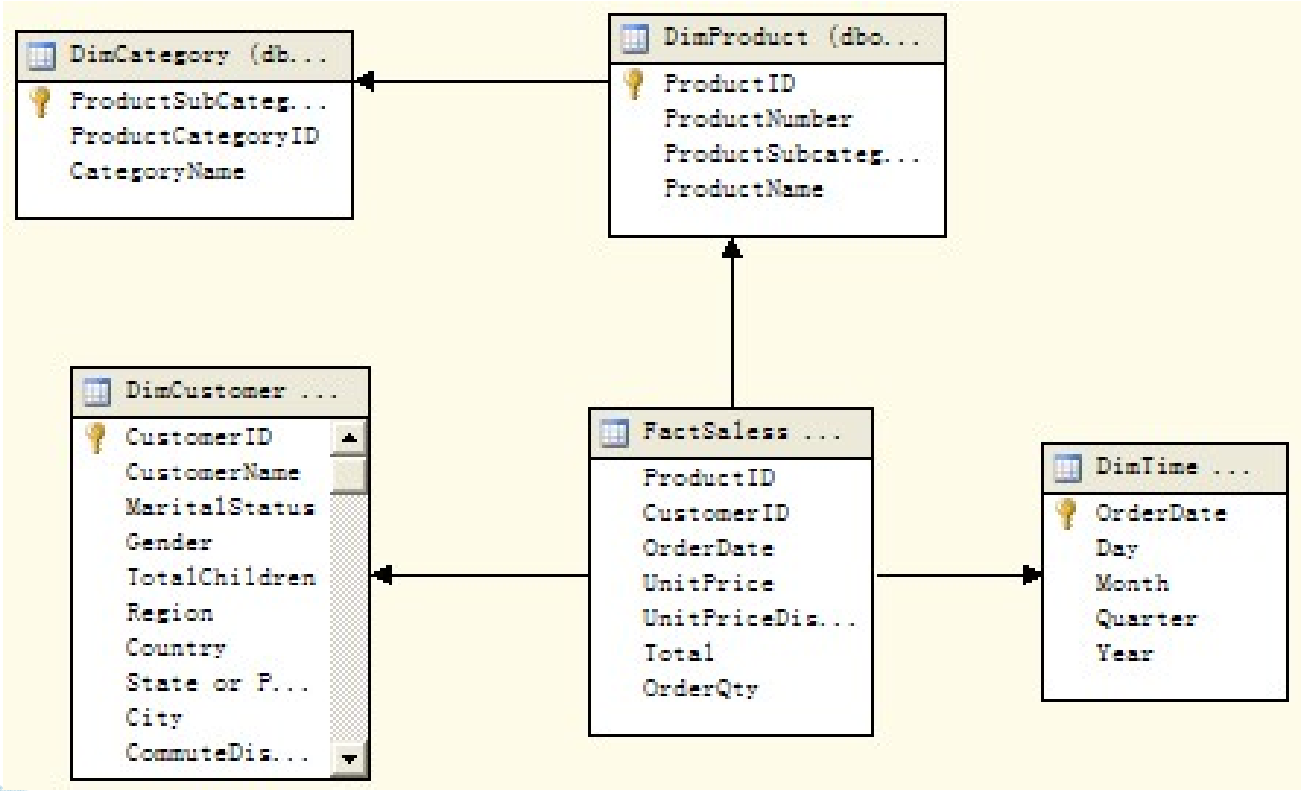
二. 数据仓库应用举例（续）

- 对于时间，我们只关心年、季度和月份，这些在数据库中不是显式存在的，但是可以从订单上的OrderDate字段中计算出来。
- 对于事实数据，我们只会关心订单中产品的价格、折扣、数量和总价的情况。
- 可以得到如下需求分析模型：



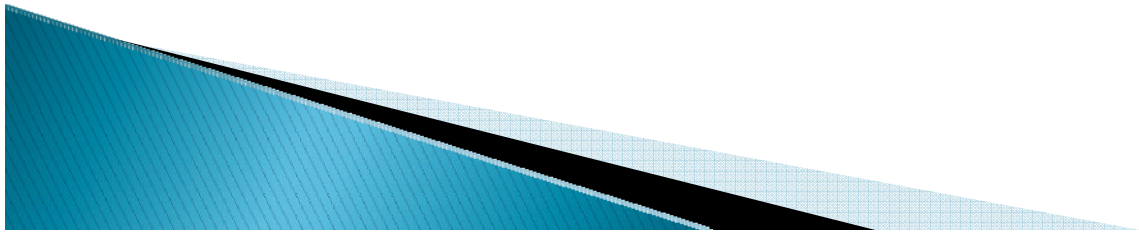
二. 数据仓库应用举例（续）

- 分析模型的事实表、维度表关系



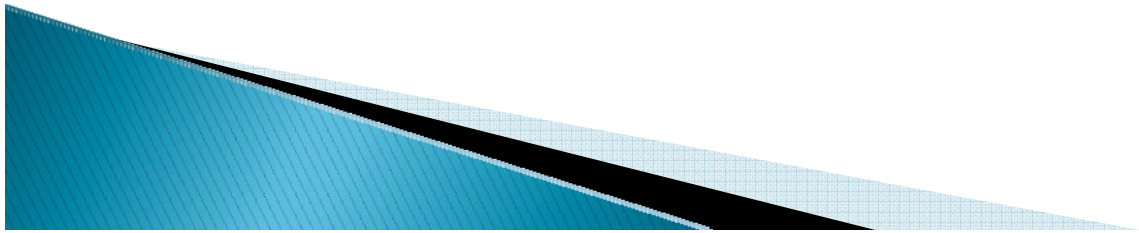
二. 数据仓库应用举例（续）

- 3. 数据转换和抽取（建立数据仓库）
  - (1)首先使用Microsoft SQL Server Management Studio新建一个数据库Sales_DW作为数据抽取的目标数据库，AdventureWorks作为源数据库。
  - (2)新建Integration Services 项目Integration Sales，并在此项目中新建一个SSIS包Integration Sales.dtsx，在此包中进行数据的抽取，整合等操作。



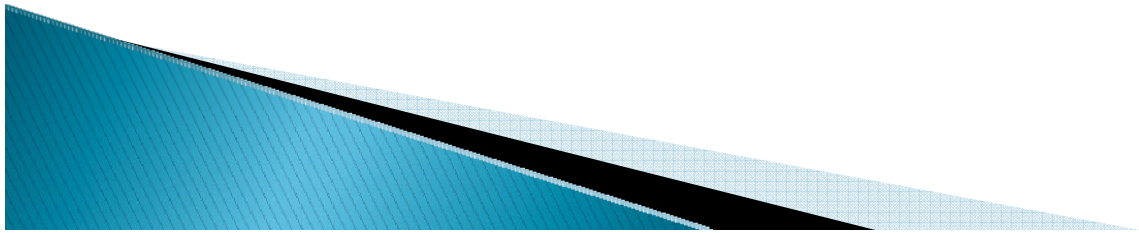
二. 数据仓库应用举例（续）

- (3)创建数据源。
  - 在Integration Sales项目下的数据源文件夹中添加两个新的数据源连接，一个连接源数据库AdventureWorks,一个连接目标数据库Sales_DW，数据源名称分别为Adventure Works和 Sales_DW 。
- (4)设计SSIS包Integration Sales.dtsx。
  - 设计包的方法是从工具箱中将需要使用的容器、任务、可执行体等工具拖拽到包的SSIS设计器窗口中，再对这些对象进行设计。
  - 由于主要执行的是数据抽取工作，因此数据流任务是我们主要设置的任务。



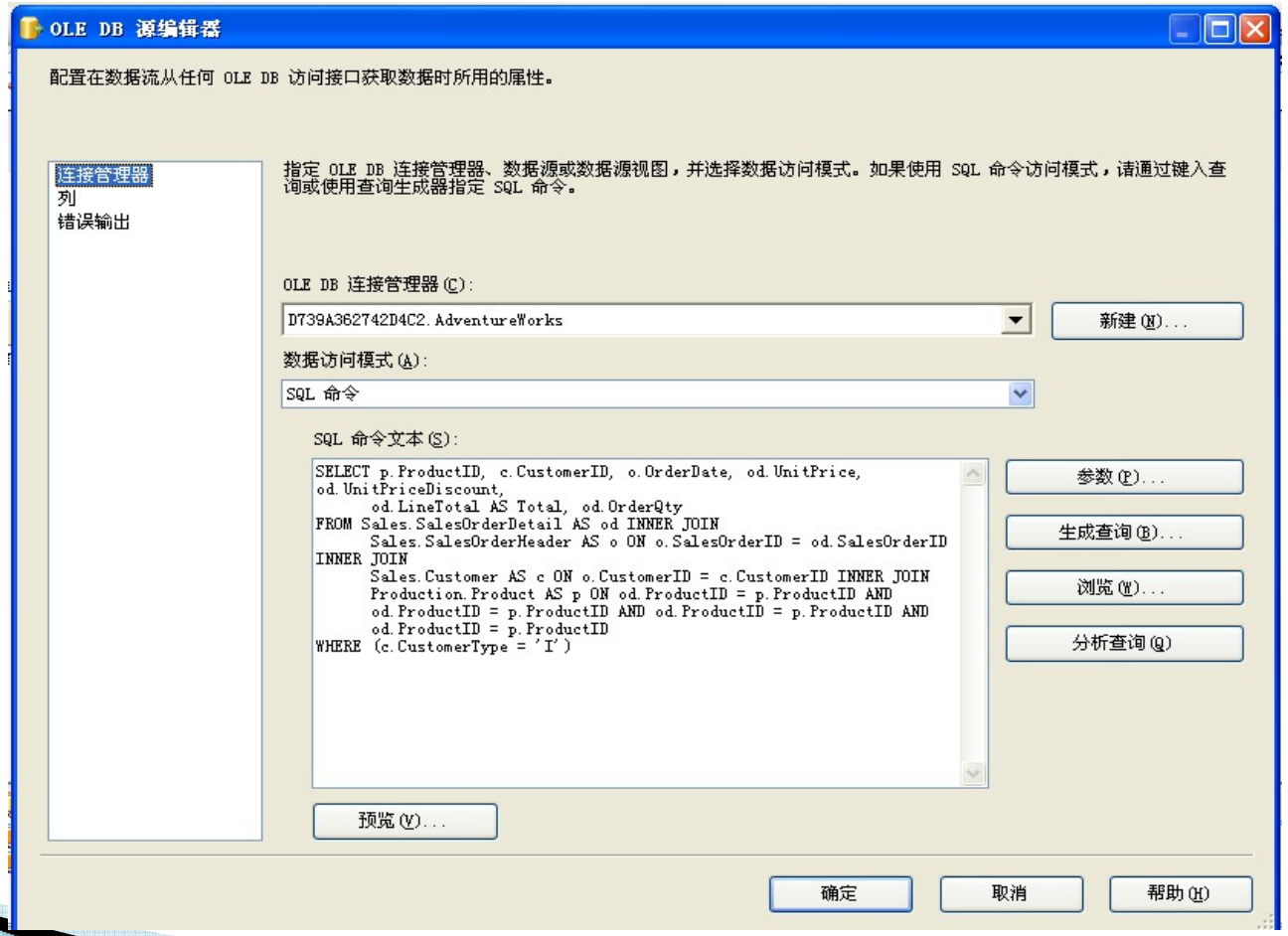
二. 数据仓库应用举例（续）

- 数据抽取中所涉及的表主要有事实表FactSales,产品信息表DimProduct，产品类别信息表DimCategory，订购时间表DimTime以及客户信息表DimCustmer。
- 抽取事实表FactSales的数据流任务的过程
  - 1)选中SSIS设计器的【控制流】标签，将工具箱中的【数据流任务】对象拖拽到SSIS设计器中，并重命名为FactSales。
  - 2)双击【数据流任务】FactSales打开【数据流】标签，将【OLE DB源】拖至SSIS设计器上。



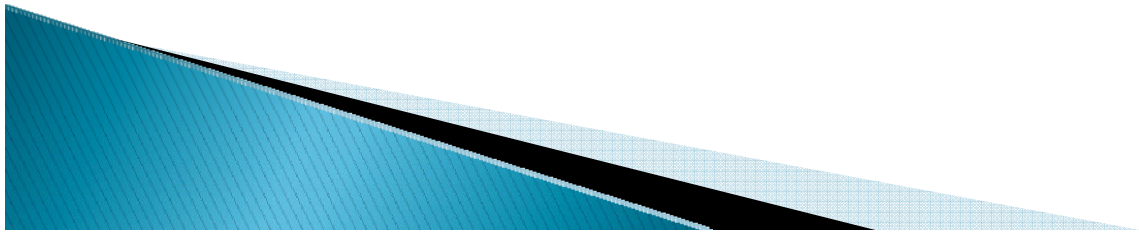
二. 数据仓库应用举例（续）

- 打开【OLE DB源编辑器】，进行【OLE DB源】对象的设置。



二. 数据仓库应用举例（续）

- 在上述设计中需选中数据源Adventure Works，并选择数据访问模式为【SQL 命令】，在【SQL 命令文本】中输入进行数据抽取的SQL语句。
- 4)完成【OLE DB源】对象设置后，从工具箱中将【SQL Server目标】对象拖至SSIS设计器上，并选中【OLE DB源】对象，将其绿色连线拖拽至新添的【SQL Server目标】对象上。打开【SQL目标编辑器】，选中数据源Sales_DW，并新建表FactSales。

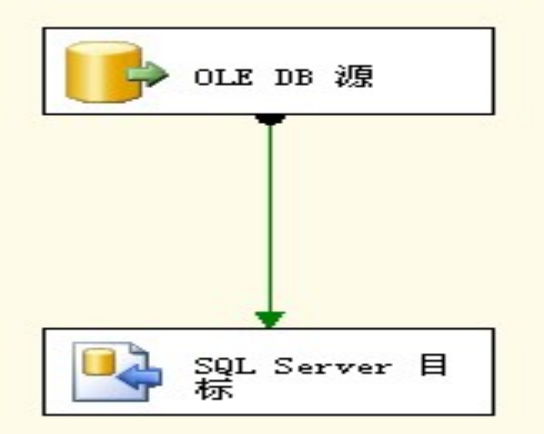


二. 数据仓库应用举例（续）

- 【SQL目标编辑器】对话框

二. 数据仓库应用举例（续）

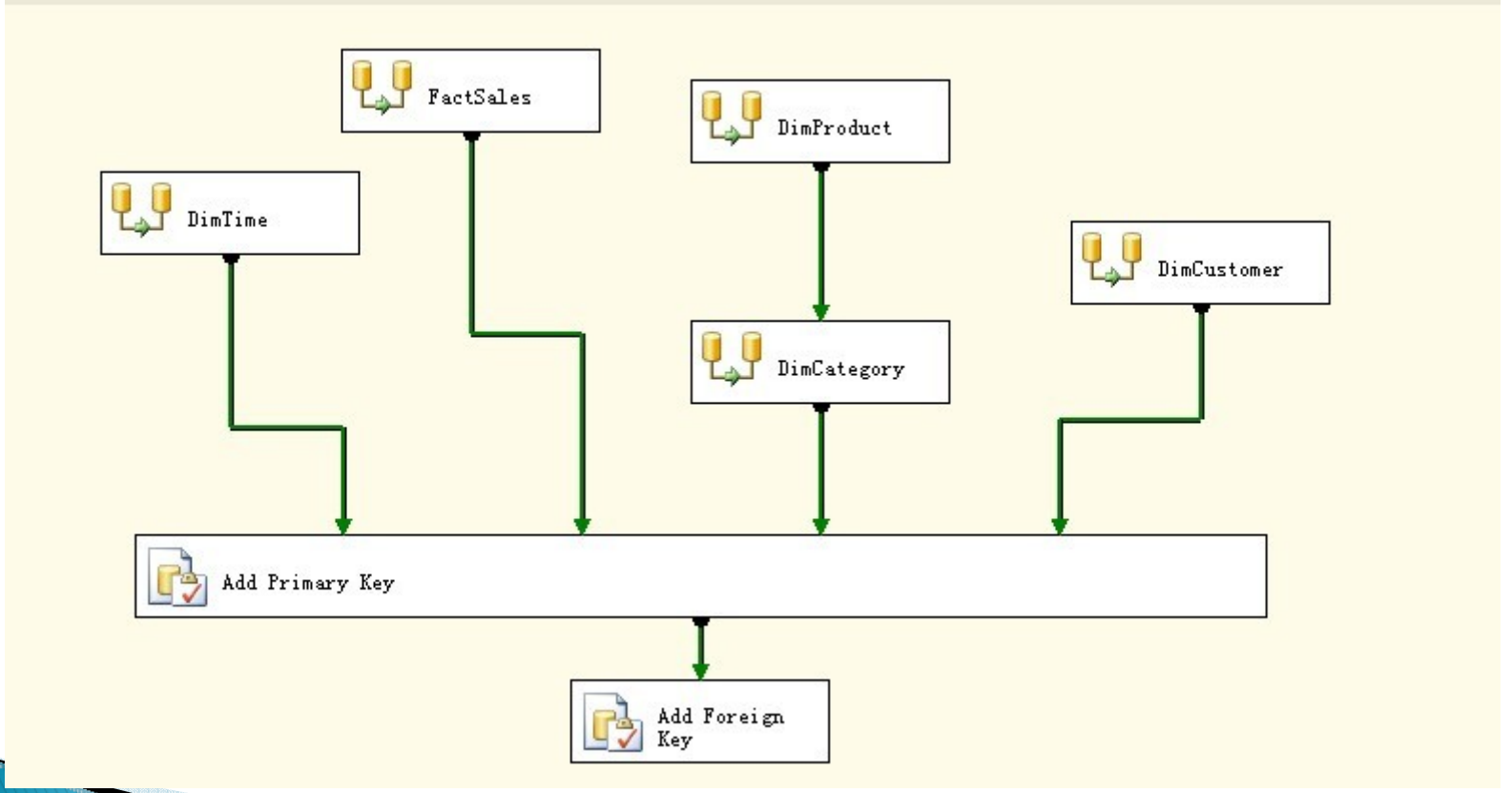
- 完成后的数据流任务视图如下所示：



- 数据仓库中其他表的数据抽取设计步骤与FactSales相同，只是输入的SQl命令不同。

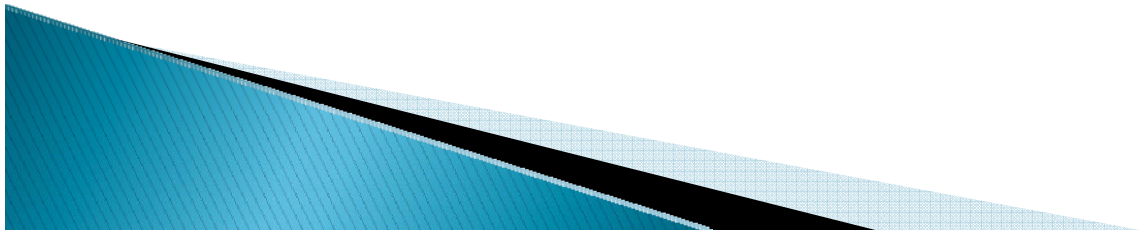
二. 数据仓库应用举例（续）

- Integration Sales 包的完整设计视图：



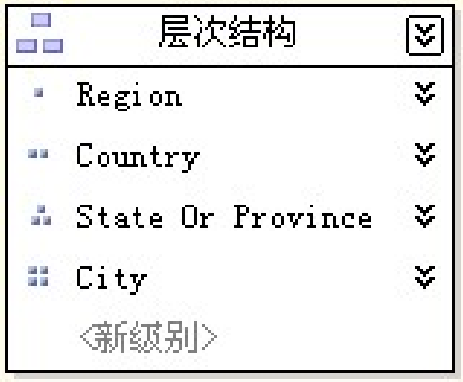
二. 数据仓库应用举例（续）

- (5)建立OLAP和挖掘模型。
  - 1)创建OLAP多维数据集数据源和数据源视图。
    - 新建项目Sales Analysis，并添加到已有解决方案中。在此项目中新建数据源选中已在Integration Sales项目中创建好的数据源Sales_DW。
    - 新建数据源视图，选中数据源Sales_DW，并将FactSales、DimCategory、DimCustomer、DimTime以及DimProduct选为视图【包含的对象】，命名此数据源视图为 Sales DW View。



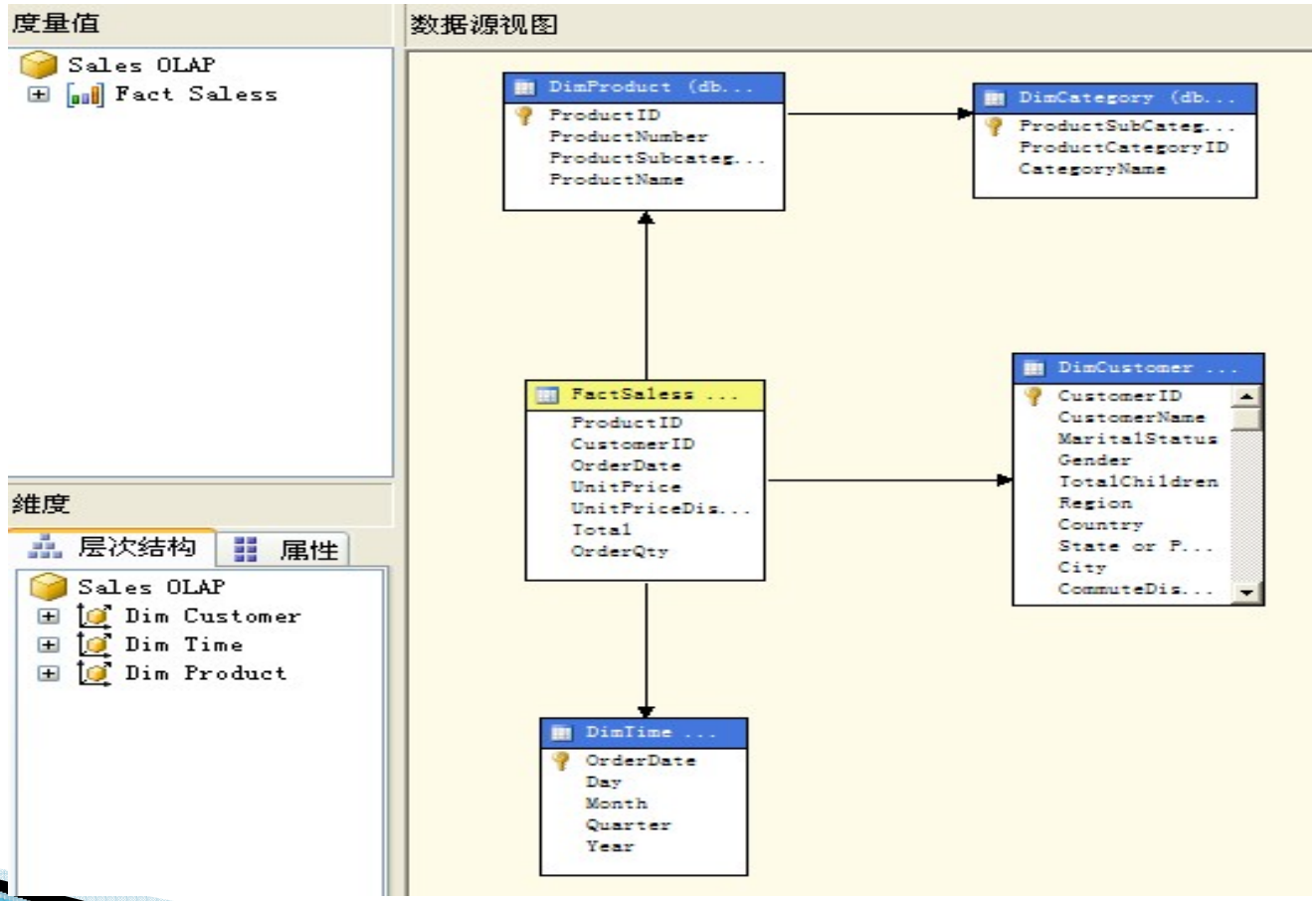
二. 数据仓库应用举例（续）

- 2)创建多维数据集
  - 新建多维数据集Sales OLAP。选中数据源视图 Sales DW View，并将FaceSales选为事实表，其他表作为维度表。
  - 由于时间维度DimTime和客户维度DimCustomer还分别具有时间层次结构和地理层次结构，因此还需要为这两个维度创建层次结构，如下所示：



二. 数据仓库应用举例（续）

- 多维数据集Sales OLAP的数据视图和结构



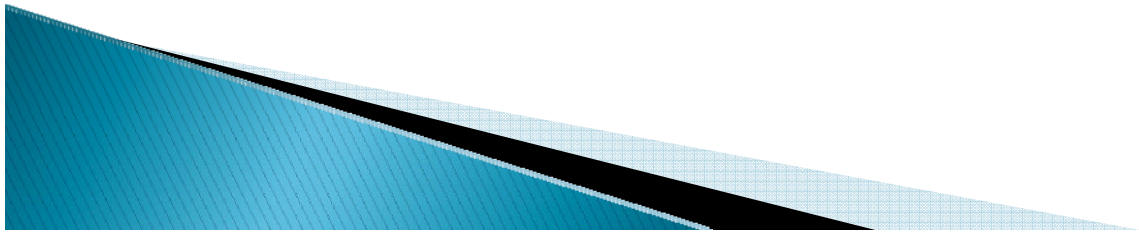
二. 数据仓库应用举例（续）

- 3）对多维数据源部署和处理。
  - 右击【多维数据集】|Sales OLAP多维数据集，选中【处理】命令。
  - 【处理】结果：

    处理完毕后，分析人员就可以使用Sales OLAP对数据进行分析了。

二. 数据仓库应用举例（续）

- 4)使用多维数据集进行销售业绩的分析
  - 双击Sales OLAP多维数据集，选中【浏览】标签，将Total等相应字段拖至浏览器选项页的正确位置，在【筛选表达式】劣种可以选择不同的产品类别，则右下侧表格中的将出现此类别产品的销售业绩。



二. 数据仓库应用举例（续）

- 所有自行车在不同地区不同时间段内销售的业绩

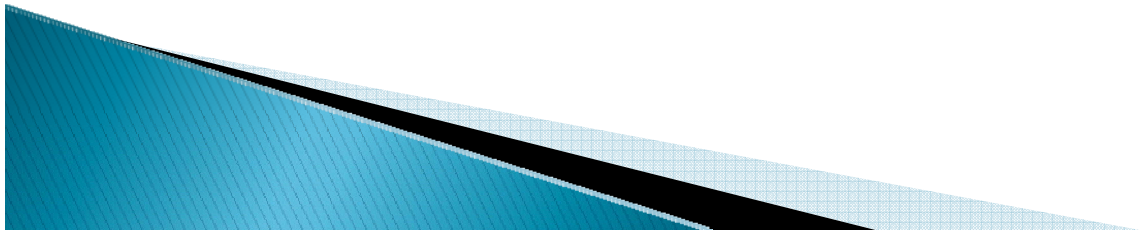
维度	层次结构	运算符	筛选表达式
Dim Product	Category Name	等于	{ Bikes }
〈选择维度〉			

将筛选字段拖至此处

		Year ▾ Quarter Month					
			⊞ 4	汇总	⊞ 2003	⊞ 2004	总计
Region ▾	Country		Total	Total	Total	Total	Total
⊟ Europe	⊞ France	6.5212	595647.0352	2059768.0524	3951360.75679998	3480887.27999998	10214302.8572
	⊞ Germany	4.652	535846.002	2084923.39	4094438.40199997	4103555.63999997	11234057.3927999
	⊞ United Kingdom	9.3792	696530.8072	2366347.416	5004318.10999998	4594343.03999998	13131370.6436
	汇总	30.5524	1828023.8444	6511038.85839999	13050117.2688005	12178785.9600004	34579730.8936009
⊟ North America	⊞ Canada	5.1904	234569.6876	2486409.5292	1885781.88959999	2325698.91999999	7285209.56839998
	⊞ United States	74.7356	1258004.9644	8506786.18399999	10709353.3800003	12381100.7600004	35999438.1232007
	汇总	29.926	1492574.652	10993195.7132001	12595135.2696005	14706799.6800006	43284647.6916012
⊞ Pacific		23.9896	1990598.7924	8617139.53399998	11791157.9324003	9763713.76000015	35408200.0176005
总计		84.468	5311197.2888	26121374.1056001	37436410.4708013	36649299.4000012	113272578.602803

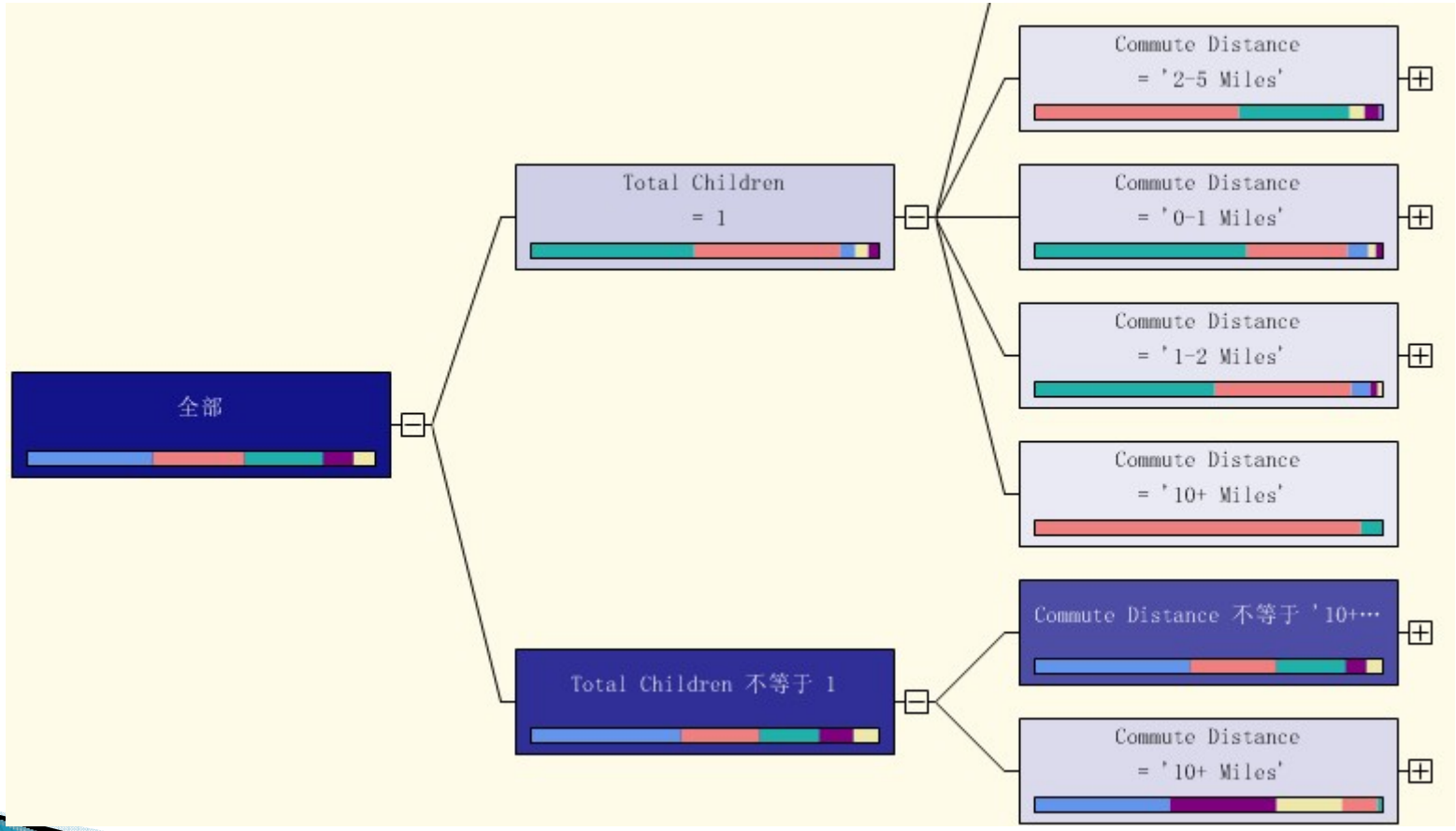
二. 数据仓库应用举例（续）

- 5）建立数据挖掘结构和数据挖掘模型
  - 从【现有多维数据集】新建挖掘结构Dim Customer，挖掘技术选择【Microsoft决策树 】。
  - 将NumberCarsOwned作为可以预测列，其他列作为输入。
  - 右击【挖掘结构】|Dim Customer，选择【处理】命令，完成对挖掘结构的部署和处理。
  - 单击设计器上的【挖掘模型查看器】标签，在设计器上会显示挖掘结果。



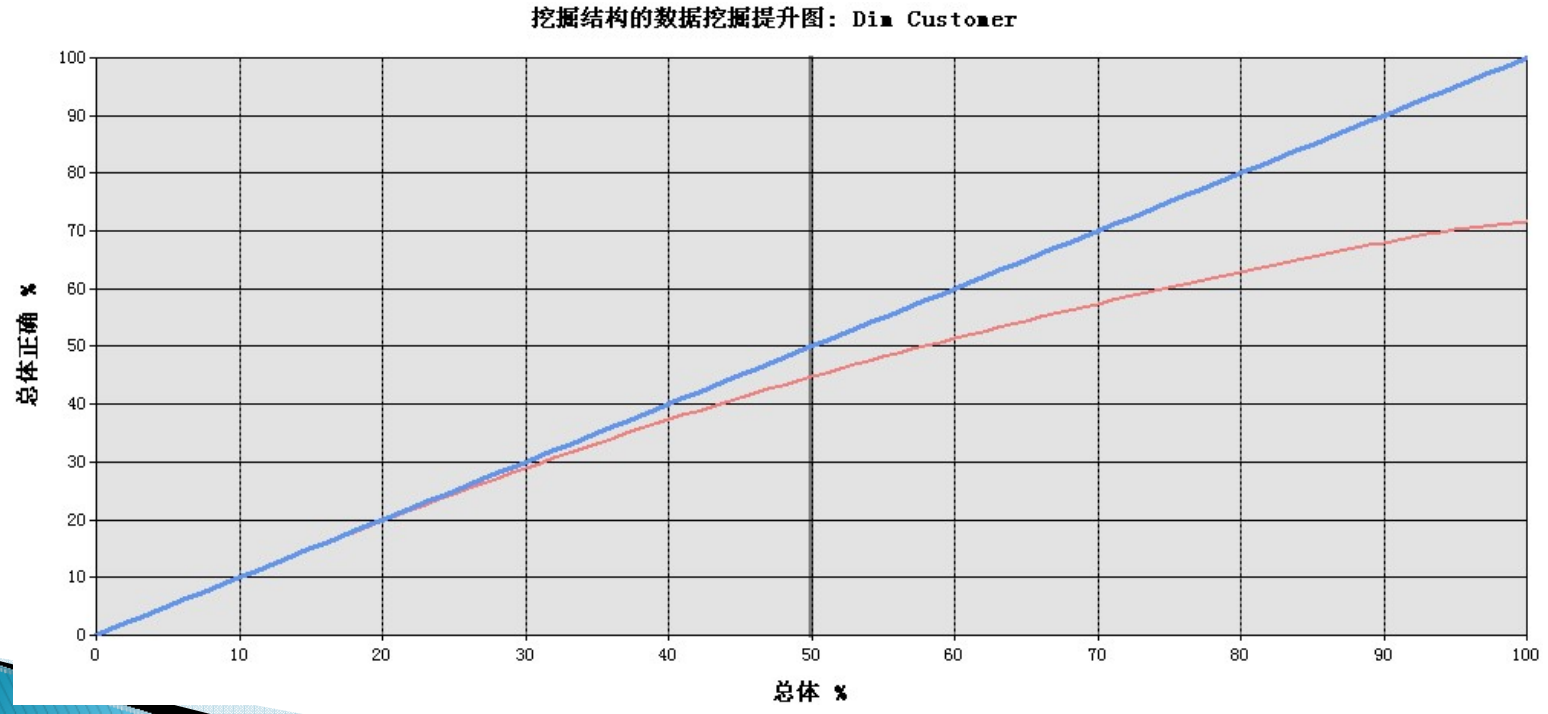
二. 数据仓库应用举例（续）

- 影响客户所有车的数量的因素挖掘模型：



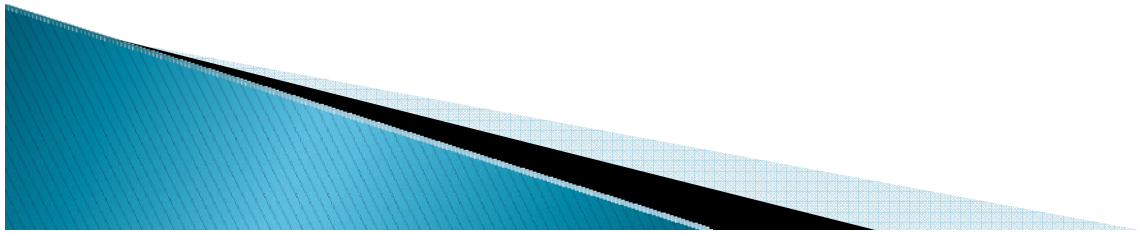
二. 数据仓库应用举例（续）

- 为了了解挖掘模型的准确度，可以单击设计器上的【挖掘准确性图标】查看挖掘结构的提升图，直线表示实际值，曲线表示预测的值，这样可以比较挖掘结构的准确度。



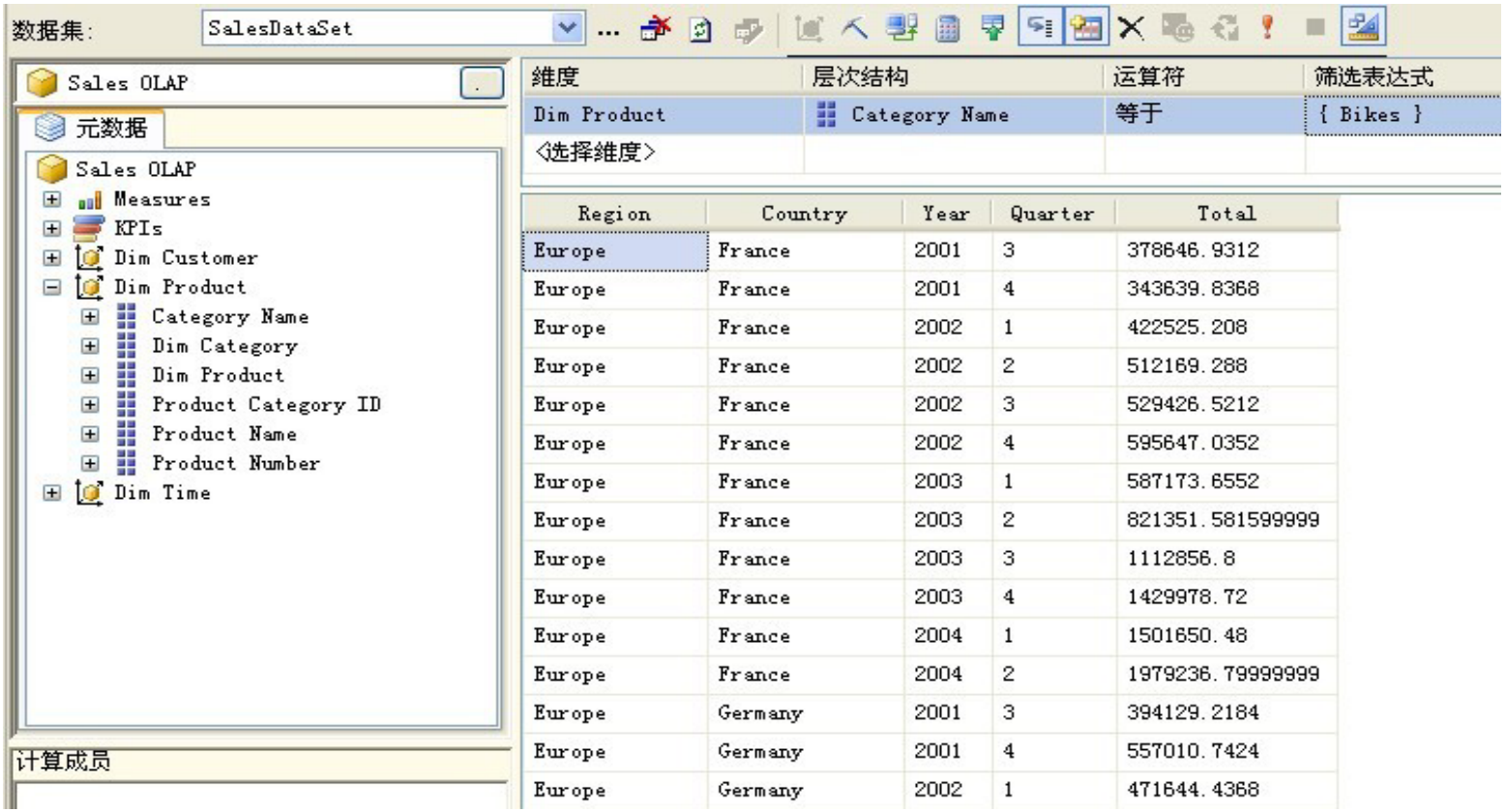
二. 数据仓库应用举例（续）

- (6)创建报表
  - 新建【报表服务项目】，项目名称为【Sales 报表】。
  - 本例采用报表设计器创建报表，选择已创建的多维数据集作为数据源，数据源名称为Sales DW。
  - 新建报表SalesReport.rdl，为报表新建一个数据集SalesDateSet，数据集的数据源选择前面已建好的数据源Sales DW。
  - 在报表设计器中可创建数据集查询视图以及所要生成的报表。



二. 数据仓库应用举例（续）

- 自行车对不同地区的客户在不同时间内的销售业绩(数据集查询视图)



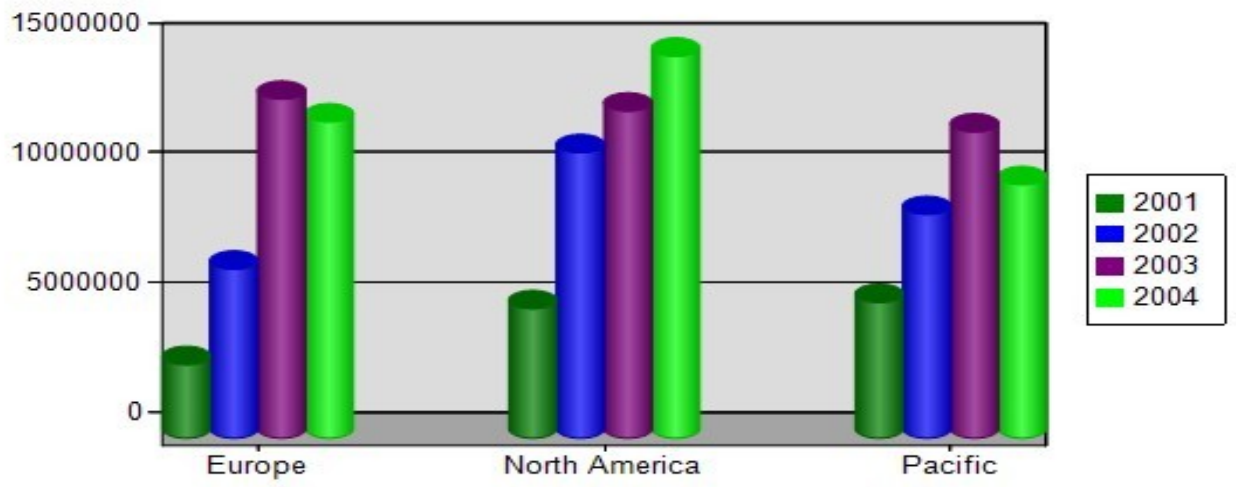
Region	Country	Year	Quarter	Total
Europe	France	2001	3	378646.9312
Europe	France	2001	4	343639.8368
Europe	France	2002	1	422525.208
Europe	France	2002	2	512169.288
Europe	France	2002	3	529426.5212
Europe	France	2002	4	595647.0352
Europe	France	2003	1	587173.6552
Europe	France	2003	2	821351.581599999
Europe	France	2003	3	1112856.8
Europe	France	2003	4	1429978.72
Europe	France	2004	1	1501650.48
Europe	France	2004	2	1979236.79999999
Europe	Germany	2001	3	394129.2184
Europe	Germany	2001	4	557010.7424
Europe	Germany	2002	1	471644.4368

二. 数据仓库应用举例（续）

- 自行车销售业绩报表【预览】结果

自行车销售业绩一览

自行车销售	2001	2002	2003	2004
Europe	￥2,839,788.81	￥6,511,038.86	￥13,050,117.27	￥12,178,785.96
North America	￥4,989,517.03	￥10,993,195.71	￥12,595,135.27	￥14,706,799.68
Pacific	￥5,236,188.79	￥8,617,139.53	￥11,791,157.93	￥9,763,713.76



中国石油大学（北京）

结束，谢谢。