第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.1 什么是数据集市

1.3.2 数据集市的类型

1.3.3 数据集市与数据仓库 的区别

1.3.4 数据集市的特点

1.3.5 数据集市的开发方法

1.3.6 数据集市的建立

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.1 什么是数据集市

数据集市是一种小型的数据仓库，主要面向部门级业务，并且只面向某个特定的主题，是为满足特定用户的需求而建立的一种分析环境。它能够快速地解决某些具体的问题，发布特定用户所需的信息。它们的投资规模比数据仓库小很多，并且更关注在数据中构建复杂的业务规则来支持功能强大的分析。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.1 什么是数据集市

一种比较常见的误解：

认为数据仓库和数据集市的差别只是数据量的大小而已。

实际上数据仓库是企业级的，数据仓库中存放的是整个企业的信息，并且数据是按照不同主题来组织的，能为整个企业各个部门的运行提供决策支持手段。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.1 什么是数据集市

- 数据集市只存放了某个主题需要的信息，一般只能为某个局部范围内的管理人员服务，因此也成为“小数据仓库”或“部门级的数据仓库”。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.1 什么是数据集市

例：假设为某个银行构建一个分行级别的数据仓库，再为该分行国际业务部构建从属型数据集市。

数据仓库的数据来源于银行的业务系统，包括储蓄、卡、个贷、外汇宝、中间业务等，分析的主题包括客户、渠道、产品等。数据仓库的数据粒度根据分析的需求而定，一般包括具体的历史记录。然后，将这些记录汇总到天、周、月、季度、年等各个层次，具体数据粒度由分析的需求而定。另外，数据仓库还存储一些为分析而计算的指标。比如，客户的价值或客户的忠诚度。这些指标的计算不能通过单一的业务系统取得，它需要从所有业务上综合考虑，这也是数据仓库系统的优点之一。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.1 什么是数据集市

假设整个分行有20万个客户，那么数据仓库将包含20万个客户所有业务的历史数据、汇总数据以及数据仓库指标数据，数据量将会达到几十甚至数百G。为了满足全行所有部门用户的查询和分析，数据仓库只能采用范式化设计。这样，不管用户有什么查询需求，只要有数据存在就能满足所需。

第一章 数据仓库原理

1.3 数据仓库与数据集市

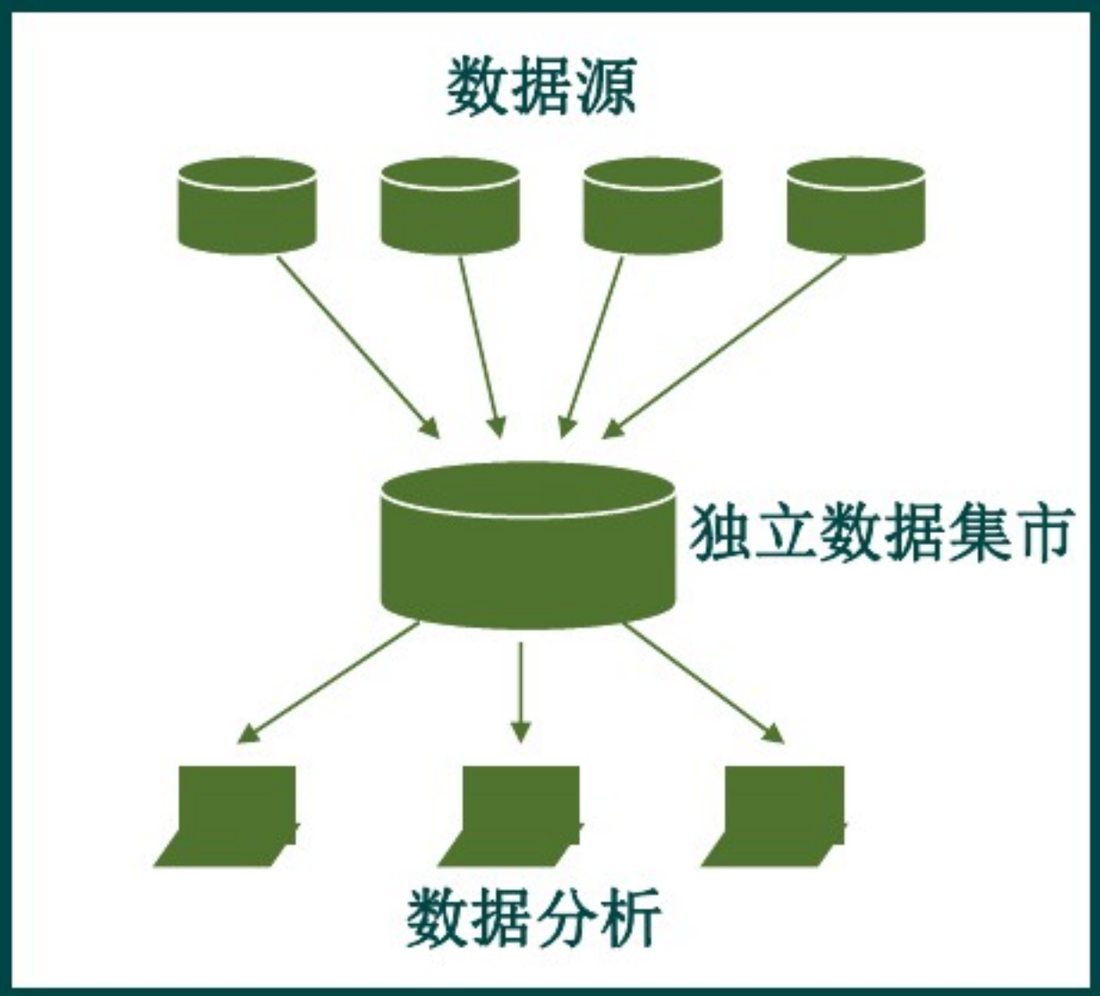
1.3.1 什么是数据集市

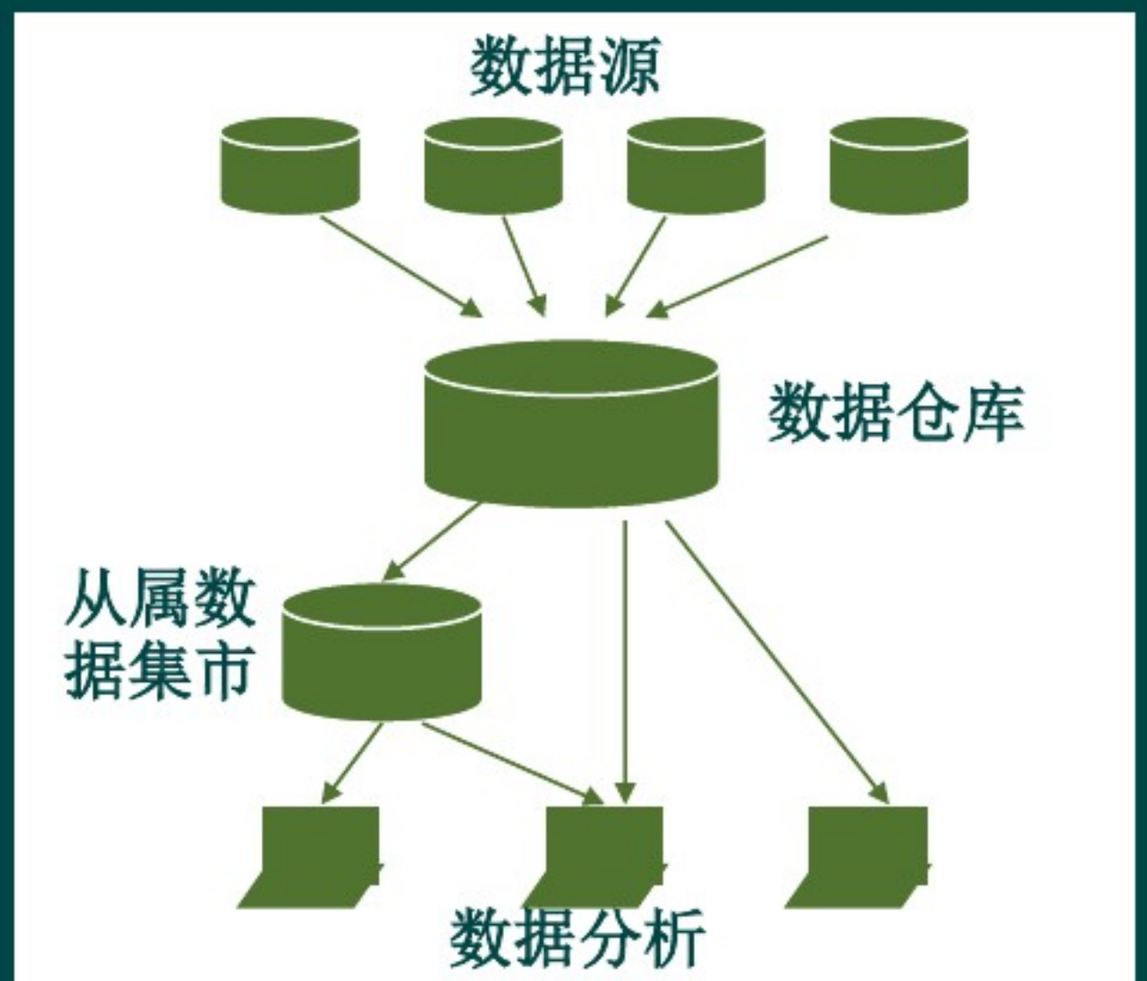
假设国际业务部门的客户有2万人。如果不构建数据集市，他们会直接在数据仓库上查询相关的信息，比如外汇宝客户去年一年外汇交易额在各种交易方式的分布。这种查询的效率和性能是非常低的，如果各个部门的所有用户都直接在数据仓库上查询相关的信息，数据仓库的性能会下降，以至于无法满足大多数用户对性能的要求。因此，构建部门级的数据集市是非常必要的。国际业务部门的数据集市，集中了数据仓库中与本部门直接相关的业务数据，例如2万个客户外汇交易的历史数据以及汇总。它采用星型模型，可以方便OLAP工具的查询和分析。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.2 数据集市的类型





第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.3 数据集市与数据仓库的区别

	数据仓库	数据集市
范围	企业级	部门级
主题	企业主题	部门或特殊的分析主题
数据粒度	最细粒度	较粗的粒度
历史数据	大量的历史数据	适度的历史数据
优化	处理海量数据、数据探索	便于访问和分析、快速查询

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.3 数据集市与数据仓库的区别

关于数据集市，常常存在如下几个误区：

1）单纯用数据量的大小来区分数据集市和数据仓库

2）简单地理解数据集市容易建立

3）数据集市很容易升级成为数据仓库

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.4 数据集市的特点

1）规模小、灵活，可以按照多种方式来组织，如按特定的应用、部门、地域、主题等。

2）投资规模小、投资回收期短，风险小。

3）独立数据集市的构建比较快。

4）不同的数据集市可以分布在不同的物理平台上，也可以逻辑地分布在同一物理平台上。这种灵活性使得数据集市可以独立地实施，因而企业人员可以快速获取信息。

5）数据集市的思想同时提供了分布式数据仓库的思想。如果按照数据的地理分布来组织数据集市，那么就形成了一个地理上分布的数据仓库。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.4 数据集市的特点

数据集市的缺点：

1）建立各个数据集市的部门是互相隔离的，相互之间不能就标准、流程、知识及经验教训进行沟通，这将导致大量的重复劳动及重复分析。

2）这些部门一般会选择不同的工具、软件及硬件，使企业不得不为支持各种技术而维持一定数量的技术人员，造成成本增加。

3）独立数据集市直接读取操作系统的文件或表，极大限制了DSS的伸缩能力。

4）数据集市一般是为不同的部门建立的，这些数据集市没有进行集成，而且没有一个会包含了整个企业的视图。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.5 数据集市的开发方法

数据集市的开发方法有自上而下和自下而上两种。不同类型的数据集市采用不同的开发方法。

1、自上而下的开发方法

从属型的数据集市，采用自上而下的开发方法。首先建立企业级的数据仓库，然后从企业级的数据仓库中为各个部门抽取必要的数据建立部门级的数据集市。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.5 数据集市的开发方法

1、自上而下的开发方法

优点：

有利于维护全局数据的一致性。

缺点：

一步建立一个企业级的大规模数据仓库，项目实施的周期很长，难度和投资都很大，风险高。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.5 数据集市的开发方法

2、自下而上的开发方法

先从数据集市入手，就某一个特定的主题，先做独立数据集市，当数据集市达到一定规模，再从各个数据集市进行数据的再次抽取建立企业级数据仓库。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.5 数据集市的开发方法

2、自下而上的开发方法

优点：

可以先建立重要的数据集市，然后再逐步扩大，具有实时快速，失败风险小的优点。

缺点：

数据集市一般是为不同的部门建立的，每一个数据集市对数据的视角都比较窄，各数据集市中难免有矛盾和不一致的数据，因此建立数据仓库时必须进行数据的再次ETL转换。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.5 数据集市的开发方法

专家们推崇将两者结合起来的折中方法，步骤如下：

1）从整个公司的角度来计划和定义需求。

2）为完整的仓库创造一个体系结构。

3）使数据内容一致而且标准化。

4）将数据仓库作为一组超级数据集市来实施，每次一个。

在这种方法中，数据集市是整个数据仓库系统的逻辑子集，数据仓库是统一化了的数据集市。

第一章 数据仓库原理

1.3 数据仓库与数据集市

1.3.6 数据集市的建立

数据集市的建立过程如图：

商业目标驱动所需信息，而这两者将共同决定所需的基础设施。一旦数据集市构建好之后，就由基础设施来管理企业用户所需的信息并使之可以访问。

第一章 数据仓库原理

1.3 数据仓库与数据集市

本节小结：

本节介绍了数据集市的概念、类型、特点、设计方法、建立过程以及与数据仓库的区别。

数据仓库和数据集市是两个容易混淆的额概念。一种比较常见的误解是认为两者的差别只是数据量的大小而已。事实上，数据仓库是企业级的，能为整个企业各个部门的运行提供决策支持手段；而数据集市则是部门级的，一般只能为某个局部范围内的管理人员服务，因此也称之为部门级数据仓库。

数据集市有独立的数据集市和从属的数据集市两种。数据仓库和数据集市各有优缺点，到底是先建企业级的数据仓库还是先建部门级的数据集市，应根据具体情况而定。

第一章 数据仓库原理

1.3 数据仓库与数据集市

本节讨论题：

1、为什么会产生数据仓库和数据集市？数据仓库和数据集市有何区别与联系？

2、数据集市怎么建？

3、有人认为“先独立地构建数据集市，当数据集市达到一定的规模再直接转换为数据仓库”，这种方法可行吗？为什么？