# DBMS原理与设计 数据仓库与数据挖掘

# 从数据库到数据仓库

数据处理分为：操作型处理和分析型处理。

联机事务处理系统是操作型数据处理例子。

决策支持系统（DSS）是分析型处理例子。

|  |  |
| --- | --- |
| 操作型数据 | 分析型数据 |
| 细节的 | 综合的，或提炼的 |
| 当前数据 | 历史数据 |
| 可更新 | 不可更新 |
| 操作需求事先可知道 | 操作需求事先不知道 |
| 生命周期符合SDLC（软件开发生命周期） | 完全不同的生命周期 |
| 对性能要求高 | 对性能要求宽松 |
| 一个时刻操作一个单元 | 一个时刻操作一个集合 |
| 事务驱动 | 分析驱动 |
| 面向业务处理 | 面向分析挖掘 |
| 一次操作数据量小，计算简单 | 一次操作数据量大，计算复杂 |
| 支持日常操作 | 支持管理需求 |

为何传统数据库系统不能处理分析型应用？

1. 数据分散：事务处理应用之间独立、分散的。
2. 蜘蛛网问题：抽取导致企业数据间形成错综复杂网状结构。
3. 数据不一致问题： 抽取时间、算法等导致数据不一致
4. 数据动态集成问题：集成数据以一定周期刷新
5. 历史数据问题：分析型需要历史数据
6. 数据的综合问题：DSS不需要数据细节，需要对数据进行不同程度综合。

数据仓库出现原因：数据集成；操作型——》分析型。

数据仓库是一个面向主题的、集成的、不可更新的、随时间不断变化的数据集合，用以更好地支持企业或组织的决策分析处理。

数据仓库4个基本特征：

1. 数据是面向主题的
2. 数据是集成的
3. 数据是不可更新的
4. 数据是随时间变化的

数据仓库主要实现方式：关系实现方式和多维实现方式。

主题域特点：独立性和完备性。

为何数据仓库特征中数据是不可更新的和数据是随时间变化的，两者互斥？

答：数据不可更新指数据不可修改，而数据随时间变化指会随时间变化增加新的数据。

数据仓库系统由前后台工具、DW服务器、OLAP组成。

OLAP服务器分为：ROLAP（关系DBMS）、MOLAP（多维数组）、HOLAP以及特殊SQL服务器。

# 操作数据存储

ODS是用于支持企业日常的全局应用的数据集合。具有四个基本特点：面向主题的、集成的、可更新的、数据是当前或接近当前的。

ODS主要是适应进行企业级的全局应用的需求而产生的。全局应用分为：企业级联机事务处理（OLTP）和即时OLAP数据处理。

# 数据仓库中的数据及组织

数据粒度分为：

1. 对数据仓库中数据的综合程度高低的一个度量。它既影响数据仓库中的数据量的多少，也影响数据仓库所能回答查询的种类。粒度越小，细节程度越高，综合程度越低。
2. 特殊形式的粒度是样本数据库。粒度级别不是根据综合程度不同来划分，而是根据采样率的高低来划分。

数据分割后的数据单元称为分片。分按行水平分片和按列垂直分片。

数据仓库中数据的追加方法：时标方法、DELTA文件、前后快照文件方法、日志文件。

组织方式：简单堆积、轮转综合、简单直接、连续文件

# 概述与模型

多维数据模型分：星型、雪片和事实群模型。

聚集函数分为：分布型（sum、count、min、max等）、代数型（若干分布型函数代数运算）、整体型（median、rank等，不可自维护型）

增量维护方法针对分布型和代数型聚集函数（可自维护型）

多维分析操作：切片、切块、旋转、下钻、上卷。

# 数据方体的存储、预计算和缩减

数据方体计算方法：流水线算法、BUC算法（自底向上方式）、多路数组聚集法。

数据方体缩减方法：

1. Dwarf数据方体（小矮人法）：将具有相同前缀或后缀的数据单元压缩在一起。一般来说，数据方体中数据密集的地方前缀冗余比较多，数据稀疏的地方后缀容易比较多。
2. Condensed数据方体：利用数据方体中存在着一些数据单元是从同一个基元组（事实表中的元祖）聚集而来（从而具有相同的聚集值）的特点，将多个数据单元压缩存放在一个物理单元中，从而节省了大量的存储空间。
3. Quotient数据方体：利用多个数据单元如果都是由同一个元祖聚集而成，则一定具有相同聚集值得原理，识别出事实表中的单元组，并将由该元祖聚集而成的数据单元进行压缩存储，从而达到对数据方体进行缩减的目的。（值相等，等价划分）

# 数据方体的索引、查询和维护

和树索引相比，bitmap索引有两个优点：

1. 较紧凑，节约空间
2. 可以使用有效的位操作来快速回答查询。