数据仓库的作用：

数据仓库是一个面向主题的、集成的、时变的、非易失的数据集合。数据仓库通过数据清理、变换、继承、装入和定期刷新等方法，从一个或多个数据源收集信息，存放在一个一致的模式下。数据仓库能够提供大量的、按照实际要求集成的不同主题的数据，通过OLAP引擎对其进行数据挖掘，发现知识。

数据仓库的地位：

数据仓库是知识发现过程中不可或缺的一环，它是进行数据挖掘的必要基础。数据仓库能够提供非冗余的有效数据，这些数据都是面向主题的，因此能够大大提高知识发现的能力和效率。没有数据仓库，知识发现就没有数据源。

B树不适用于数据仓库的原因：

1、数据仓库中海量数据对单列而言数据重复度可能会比较高，对区分度低的属性用b-tree建立索引存储开销非常大。而bitmap正好适合。

2、b-tree要求查询语句简单，返回结果少。而数据仓库中的复杂查询b-tree往往效率很低。

3、创建b-tree存储的时间复杂度和空间复杂度过高。

数据挖掘流程：

(1)数据清理(消除噪声和删除不一致的数据)

(2)数据集成(多种数据源可以组合在一起)

(3)数据选择(从数据库中提取与分析任务相关的数据)

(4)数据变化(通过汇总或聚集操作，把数据变化和统一成挖掘的形式)

(5)数据挖掘(提取数据模式)

(6)数据评估(根据某种兴趣度度量，识别代表知识真正有趣的东西)

(7)数据表示(使用可视化和知识表示技术，向用户展示挖掘到的知识)

联机事务处理OLTP：主要任务是执行联机事务和查询处理

联系分析处理OLAP：数据仓库系统在数据分析和决策方面为用户或‘知识工人’提供服务。这种系统可以用不同的格式和组织提供数据。OLAP是一种分析技术，具有汇总、合并和聚集功能，以及从不同的角度观察信息的能力。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | 特征 | 面向 | 用户 | 功能 | DB | 数据 | 访问 |
| OLTP | 操作处理 | 事务 | DBA，办事员 | 日常操作 | 基于ER | 当前最新 | 读/写 |
| OLAP | 信息处理 | 分析 | 知识工人 | 决策支持 | 星型，雪花 | 时间跨度 | 读 |
| 特征 | 汇总 | 用户 | 操作 | 访问记录 | 优先 | DB规模 | 度量 |
| OLTP | 原始，详细 | 数千 | 主码索引 | 数十个 | 高性能可用 | 100mb-gb | 事务 |
| OLAP | 汇总，统一 | 数百 | 大量扫描 | 数百万 | 高灵活 | 100gb-tb | 查询 |

数据立方体：允许从多维对数据建模和观察。它由维和事实定义

维：关于一个组织想要保存记录的透视图和实体，每个维都有一个表与之相关联，成为**维表**

事实表：包括事实的名称和度量，以及每个相关维表的码

度量：数值函数，通过对给定点的各维-值对聚集数据，计算该点的度量值

多维数据模型:

多维数据模型将数据看作数据立方体，允许从多个维度对数据建模和观察。包含维表和事实表。

最流行的数据仓库数据模型是多维数据模型，这种模型可以是**星形**模式（事实表在中间，连接到多个维表）、**雪花**模式（星型的变种，某些维表规范化，分解到附加维表，以减少冗余）、**事实星座**模式（多个事实表共享维表）。

OLAP操作:

上卷：上卷操作通过一个维的概念分层向上攀升或者通过维规约，在数据立方体上进行聚集。

下钻：下钻是上卷的逆操作，它由不太详细的数据到更详细的数据。

切片和切块：切片对一个维进行选择。切块对两个以上维进行选择，定义子立方体。

转轴：可视化操作，转动视角，即维之间的位置互换。

钻过：跨越多个事实表。

钻透：钻到后端关系表。

数据仓库模型的不同类型：

企业仓库：收集了关于跨部门的整个组织主题的所有信息，跨越整个组织，因此是企业范围的。

数据集市：是企业仓库的一个部门子集，它针对选定的主题，对于特定的用户是有用的，因此是部门范围的，其数据通常是汇总的。

虚拟仓库：虚拟仓库是操作数据库上视图的集合，易于建立，但需要操作数据库服务器具有剩余能力。

OLAP服务器：直接实现对多维数据的操作,直接为商务用户提供来自数据仓库或数据集市的多维数据。ROLAP:多维数据操作映射到标准关系操作。MOLAP：多维数据视图映射到数组中.HOLAP:结合，历史数据ROLAP,频繁访问数据放到MOLAP。

数据仓库的设计：

(1)分析建立企业模型并映射到数据仓库概念模型

(2)逻辑模型的设计

(3)物理模型的设计

事实表的特征(与维表相比)：

(1)记录数量非常多

(2)除了度量外，其他字段都是维表或者中间维表的关键字

(3)如果事实相关的维度很多，则事实表的字段也会很多

逻辑模型设计：

(1)系统数据量估算

(2)数据粒度的选择(数据量较小的情况下使用单一的数据粒度，对于大数据量需要采用双重粒度，对于细节数据只保存近期的数据在数据仓库之中，当保留周期到达时，将距离当前较远的数据导出到磁带或存储设备上)

(3)数据的分割（数据分割是指将数据分散到各自的物理单元里以便能够独立处理，以提高数据处理的效率）

(4)表的合理划分（字段的更新频率和访问频率不一样——稳定性）

(5)删除纯操作数据（如“收款人”），增加导出字段（如“销售总量”）

(6)定义关系模式

(7)定义元数据储存(元数据：描述数据的数据，定义数据仓库对象的数据。包括数据仓库的结构、操作元数据（数据血统、流通，监控信息）、用于汇总的算法、从操作环境到数据仓库的映射；关于系统性能的数据、商务元数据)

(8)定义记录系统(记录系统指明数据仓库中关系表的各个字段来源于业务数据库何处)

物理模型的设计：

(1)确定数据的存储结构（并行RAID）

(2)索引策略（位图索引、连接索引）

(3)数据存储策略与性能优化（多路聚集优化、表的归并、分割表的存放、按列存储、存储分配优化）

(4)数据装载接口

(5)并行优化设计

多路数组聚集：

是数据立方体的高效计算方式。使用多维数组作为基本数据结构，自底向上的、共享地计算完全数据立方体。使用数组直接寻址的典型MOLAP。

方法：最大维在形成单块的平面上。最小为在形成单面的平面上，每个平面必须被排序，并按大小递增的顺序被计算。