

复习资料

一、名词解释

1. 数据仓库: 数据仓库是企业所有级别的决策制定过程提供支持的所有类型数据的战略集合。它是单个数据存储, 出于分析性报告和决策支持的目的而创建。 为企业提供需要业务智能来指导业务流程改进和监视时间、成本、质量和控制。
2. 螺旋式周期性开发方法: 针对未来航空电子系统面临的挑战和航空电子系统设计的特点, 提出了基于原型仿真的航空电子系统螺旋式开发方法
3. 数据仓库元数据: 在数据仓库中, 元数据是描述数据仓库数据的结构和建立方法的数据。
- 4 数据挖掘: 是数据库知识发现中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关, 并通过统计、在线分析处理、情报检索、机器学习、专家系统和模式识别等方法来实现目标。
5. 数据集市: 一种更小、更集中的数据仓库。原始数据从数据仓库流入不同的部门以支持这些部门的定制化使用。这些部门级的数据库就称为数据集市。迎合了专业用户群体的特殊需求。

二、简答题

1. 为什么不能依靠传统的业务处理系统进行决策分析?

传统业务系统数据量少, 不能用来做出正确的预测, 是针对日常事务处理而设计的, 因此不能支持短时间内大量数据的分析和计算; 传统业务系统数据杂乱, 可能存在大量的坏数据以及异常数据, 会影响决策分析的正确性。如果在传统业务系统中进行决策分析, 那么必然会影响到系统的使用

2. 数据粒度, 数据粒度划分策略

数据仓库中数据的细化和综合程度。

划分为: 详细数据、轻度总结、高度总结三级。原则: 细化程度越高, 粒度越小; 细化程度越低, 粒度越大。

3. 数据仓库的体系结构

数据源、数据存储和管理、OLAP 服务器以及前端工具与应用四个部分

4. MOLAP 和 ROLAP 在 OLAP 的数据存储特点

MOLAP 基本数据和聚合数据存放于多维数据集中, ROLAP 基本数据和聚合数据均存放在 RDBMS 之中

5. 请列出常见的现代数据挖掘技术

知识发现技术、规则型挖掘技术、神经网络型挖掘技术、遗传算法型挖掘技术、粗糙型挖掘技术、决策树型挖掘技术、

三、简述关联规则挖掘算法 Apriori 的算法思想并说明如何利用 Apriori 性质生成频繁项集。

首先遍历目标数据库一次, 记录每个项目或属性的出现次数, 即计算每个项目的支持度, 收集所有支持度不低于用户最小支持度的项目构成频繁 1-项集 L_1 , 然后链接 L_1 中所有的元素形成候选 2 项集 C_2 , 再次遍历事务数据库, 计算 C_2 中每个候选 2-项集的支持度, 收集所有支持度不低于用户最小支持度的项目构成频繁 2-项集 L_2 , 再链接 L_2 形成 C_3 , 遍历数据库得 L_3 , 反复执行以上过程, 直到没有候选项集为止。

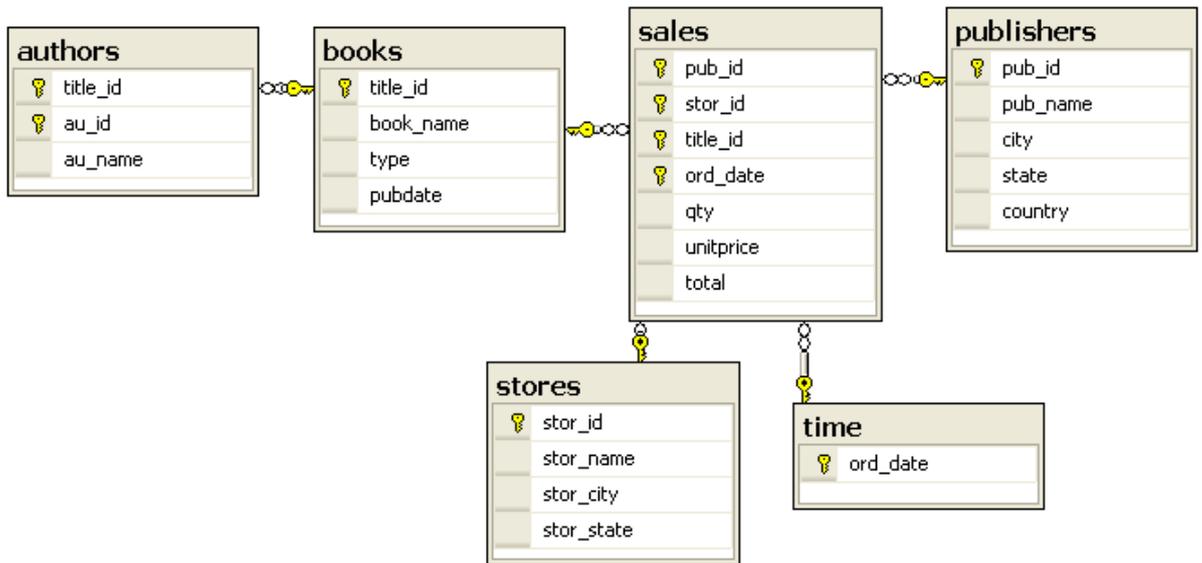
设定最小支持度 s 和最小置信度 c 。使用候选项集。首先产生出候选的项的集合, 即候选项集, 若候选项集的支持度大于或等于最小支持度, 则该候选项集为频繁项集。

四、简述数据仓库中需要的元数据，这些元数据在不同的阶段应该发挥的作用

技术元数据和业务元数据。

作用：1、描述哪些数据在数据仓库中；2、定义要进入数据仓库中的数据和从数据仓库中产生的数据；3、记录根据业务事件发生而随之进行的数据抽取工作时间安排；4、记录并检测系统数据一致性的要求和执行情况；5、衡量数据质量。

五、通过以下给出的“出版社销售图书 Pubs”数据库模型，设计一个“图书销售分析”的多维数据集模型，并详细描述如何利用 SQL Server2008 的相关组件建立该多维数据集的过程。



数据分析需求：建立出版社维度、作者维度、图书维度、书店维度、时间维度对图书销售业绩进行统计（多维数据集模型按事实表和维度表之间的星型、雪花型或星型雪花混合型的结构设计）

答：1、打开 SQL；2、附加备份的数据库文件并做出优化；3、修改数据库属性；4、建立数据仓库所需的数据库 bb（导出）；5、创建新的分析服务项目；6、新建数据源（本地服务器输入“.”）7、建立多维数据集；8、处理多维数据集，得出模型

六、在超市中的商品价格都是大于等于零的，超市的总经理只关心如何利用送一件免费商品而带来 1000 元以上的总销售量。讨论如何挖掘这种商业模式。