数据仓库就是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合。

数据处理通常分成两大类：联机事务处理OLTP和联机分析处理OLAP。

简述对数据仓库未来发展趋势的看法

数据仓库技术的发展包括数据抽取、存储管理、数据表现和方法论等方面。 在数据抽取方面，未来的技术发展将集中在系统集成化方面。进行标准化的统一管理，以适应数据仓库本身或数据源可能的变化，使系统更便于管理和维护。

在数据管理方面，未来数据库厂商明确推出数据仓库引擎，作为数据仓库服务器产品与数据库服务器并驾齐驱。在这一方面，带有决策支持扩展的并行关系数据库将最具发展潜力。

在数据表现方面，数理统计的算法和功能将普遍集成到联机分析产品中，并与Internet/Web技术紧密结合。按行业应用特征细化的数据仓库用户前端软件将成为产品作为数据仓库解决方案的一部分。

数据仓库实现过程的方法论将更加普及，将成为数据库设计的一个明确分支，成为管理信息系统设计的必备。

列出三种数据仓库产品，说明优缺点

1. IBM公司提供了一套基于可视化数据仓库的商业智能（BI）解决方案，其中，VW是一个功能很强的集成环境，既可用于数据仓库建模和元数据管理，又可用于数据抽取、转换、装载和调度。Essbase/DB2 OLAP Server支持“维”的定义和数据装载。Essbase/DB2 OLAP Server不是ROLAP（Relational OLAP）服务器，而是一个（ROLAP和MOLAP）混合的HOLAP服务器，在Essbase完成数据装载后，数据存放在系统指定的DB2 UDB数据库中。

2）Oracle数据仓库解决方案主要包括Oracle Express和Oracle Discoverer两个部分。Oracle Express利用多维模型，存储和管理多维数据库或多维高速缓存，同时也能够访问多种关系数据库；Oracle Express Web Agent通过CGI或Web插件支持基于Web的动态多维数据展现；Oracle Express Objects前端数据分析工具（目前仅支持Windows平台）提供了图形化建模和假设分析功能，支持可视化开发和事件驱动编程技术，提供了兼容Visual Basic语法的语言，支持OCX和OLE；Oracle Express Analyzer是通用的、面向最终用户的报告和分析工具（目前仅支持Windows平台）。Oracle Discoverer即席查询工具是专门为最终用户设计的，分为最终用户版和管理员版。 在Oracle数据仓库解决方案的实施过程中，通常把汇总数据存储在Express多维数据库中，而将详细数据存储在Oracle关系数据库中，当需要详细数据时，Express Server通过构造SQL语句访问关系数据库。

3）Microsoft将OLAP功能集成到SQL Server数据库中，其解决方案包括BI平台、BI终端工具、BI门户和BI应用四个部分，如图1.1。 ① BI平台是BI解决方案的基础，包括ETL平台SQL Server 2005 Integration Service(SSIS)、数据仓库引擎SQL Server 2005 RDBMS以及多维分析和数据挖掘引擎SQL Server 2005 Analysis Service、报表管理引擎SQL Server 2005 Reporting Service。 ② BI终端用户工具，用户通过终端用户工具和Analysis Service中的OLAP服务和数据挖掘服务进行交互来使用多维数据集和数据挖掘模型，终端用户通常可使用预定义报表、交互式多维分析、即席查询、数据可视化、数据挖掘等多种方法。 ③ BI门户提供了各种不同用户访问BI信息的统一入口。BI门户是一个数据的汇集地，集成了来自不同系统的相关信息。用户可以制定个性化的个人门户，选择和自己相关性最强的数据，提高信息访问和使用的效率。 ④ BI应用是建立在BI平台、BI终端用户工具和BI统一门户这些公共技术手段之上的满足某个特定业务需求的应用，例如零售业务分析、企业项目管理组合分析等。

4. 使用星型模式可以从一定程度上提高查询效率，因为星型模型中数据的组织已经经过预处理，主要数据都在庞大的事实表中。

什么是数据仓库ETL过程？

数据的ETL过程就是负责将操作型数据转换成调和数据的过程。整个过程由四个步骤组成：抽取、清洗、转换、加载和索引。事实上，这些步骤可以进行不同的组合，如，可以将数据抽取与清洗组合为一个过程，或者将清洗和转换组合在一起。通常，在清洗过程中发现的拒绝数据信息会送回到源操作型业务系统中，然后将数据在源系统中加以处理，以便在以后重新抽取。

什么是信息包图法?它为什么适用于数据仓库的概念模型的设计？

信息包图法，也叫用户信息需求表，就是在一张平面表格上描述元素的多维性，其中的每一个维度用平面表格的一列表示，通常的维度如时间、地点、产品和顾客等；而细化本列的对象就是类别，例如时间维度的类别可以细化到年、月、日，甚至小时；平面表格的最后一行（代表超立方体中的单元格）即为指标度量值，例如，某年在某销售点的某类产品的实际销售额。创建信息包图时需要确定最高层和最低层的信息需求，以便最终设计出包含各个层次需要的数据仓库 总之，信息包图法是一种自上而下的数据建模方法，即从用户的观点开始设计（用户的观点是通过与用户交流得到的），站在管理者的角度把焦点集中在企业的一个或几个主题上，着重分析主题所涉及数据的多维特性，这种自上而下的方法几乎考虑了所有的信息源，以及这些信息源影响业务活动的方式。

关联规则的分类有哪些？关联规挖掘的步骤包括什么？

关联规则的分类：

（1）基于规则中涉及到的数据的维数，关联规则可以分为单维的和多维的。

（2）基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。

（3）基于规则中处理的变量的类型不同，关联规则可以分为布尔型和数值型。

关联规则挖掘的步骤：

（1）找出交易数据库中所有大于或等于用户指定的最小支持度的频繁项集；

（2）利用频繁项集生成所需要的关联规则，根据用户设定的最小可信度进行取舍，产生强关联规则。

请描述K-means聚类算法的操作步骤

k-means算法基本步骤

　　(1)从数据中选择k个对象作为初始聚类中心;

　　(2)计算每个聚类对象到聚类中心的距离来划分；

　　(3)再次计算每个聚类中心

　　(4)计算标准测度函数，直到达到最大迭代次数停止，否则，继续操作。

贝叶斯网络的三个主要议题是什么？

贝叶斯网络的3个主要议题分别是：预测、诊断和历史数据训练

怎样从历史数据中训练出结点之间的条件概率或联合条件概率

要训练条件概率P(B|A)，可以在历史数据中统计A发生的次数T(A)，然后统计在A发生的数据中B发生的次数T(A,B)，条件概率P(B|A) = T(B)/T(A)。要训练联合条件概率P(C|A,B)，可以在历史数据中统计A、B共同发生的次数T(A,B)，然后在A、B共同发生的数据中统计C发生的次数T(A,B,C)，联合条件概率P(C|A,B)=T(A,B,C)/T(A,B)。以上的符号A、B、C可以表示某个事件，也可以表示该事件的相反事件

前馈网络和递归网络有什么本质的区别？

前馈网络和递归网络的本质区别是网络的某些输出是不是循环作为网络的输入。前馈网络的所有输出都不能作为输入，而递归网络的某些输出可以循环作为网络的输入。

简单遗传算法包括哪些步骤

SGA的基本流程如下：

1）初始化，产生初始种群。

2）个体评价，即计算种群中每个个体的适应度。

3）按选择概率Ps，执行选择算子，从当前种群中选择部分个体进入下一代种群。

4）按交叉概率Pc，执行交叉算子。

5）按变异概率Pm，执行变异算子。

6）若满足设定的终止条件，则输出种群中适应度最优的个体作为问题的最优解或满意解，否则执行（2）。

三道计算题：

1. 关联规则p91,p107
2. ID3算法p115
3. 贝叶斯公式p151