课本：

1. **数据仓库的两种实现方式**：关系实现方式，多维实现方式
2. **主题域的两个特点**：独立性，完备性
3. **为什么说数据仓库的数据是不可更新的，且又是随时间变化的，这不是矛盾吗？**

不可更新是指，值不可更新，是只读的。

随时间不断变化是指，新数据的插入和旧数据的删除。

1. **数据仓库的构成**：前台工具，后台工具，数据仓库服务器，OLAP服务器
2. **OLAP服务器分为？**

ROLAP(Relation OLAP)：采用关系DBMS来存储和管理数据

MOLAP(Multi-dimension OLAP)：采用多维数组来存储数据

HOLAP(Hybrid OLAP)：将ROLAP和MOLAP结合起来

特殊SQL服务器

1. **什么是数据集市？**是部门级的数据仓库
2. **ODS系统可以做什么？**

ODS中只存放当前或接近当前的数据。

1. 进行企业级的联机事务处理（OLTP）
2. 即时OLAP数据处理
3. **ODS（操作数据存储）与DW（数据仓库）的差别**
4. 存放的数据内容不同

ODS：当前或接近当前的数据，细节数据，可联机更新

DW： 历史数据，细节数据和综合数据，不可变快照

1. 数据量是不同等级的

ODS:存放当前或接近当前的细节数据

DW:大量历史数据

1. 二者的技术支持不尽相同

ODS：要支持联机更新，又要保证数据与源数据库系统中数据一致，应而采用和面向应用的分散DB系统的技术一样复杂

DW：需要支持ETL技术和数据快速存取技术

1. 面向的需求不同

ODS:满足企业的全局应用，企业级OLTP和即时OLAP

DW:用于高层战略决策

1. 使用者不同

ODS:企业的中层管理人员

DW:DSS分析员或企业高级决策层

1. **什么样的函数是可自维护型？不可自维护型？**

分布型/代数型聚集函数 是可自维护型

整体型聚集函数 是不可自维护型

1. **数据方体的存储有哪两种形式？**

多维数组的存储形式，传统关系表的存储形式

1. **数据方体的预计算方法有哪些？**

BUC算法，流水线算法，多路数组聚集法

1. **数据方体常用的压缩技术有哪些？**

Dwarf数据方体：将具有相同前缀或后缀的数据单元压缩在一起

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祖聚集而来的特点，将多个数据单元压缩存放在一个物理单元中。

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的原理，识别出事实表中的单元组，并将由该单元组聚集而成的数据单元进行压缩存储，从而达到了对数据方体进行缩减的目的。

1. **和树索引相比，bitmap索引的两个优点？**

较紧凑，节约空间。 可使用有效的位操作来快速回答查询。

1. **数据挖掘算法的组件化思想包括哪几个组件？** P117

模型或模式结构

数据挖掘的任务

评分函数

搜索和优化方法

数据管理策略

1. **关联规则挖掘：**交易表 转换成 二维矩阵
2. **必考：算支持度，置信度。**

支持度：指项集X和Y在数据库D中同时出现的概率。

置信度：指在项集X出现的情况下，项集Y在数据库D中同时出现的条件概率。

1. **频繁项集：**如果一个项集S是频繁的（项集S的出现频度大于最小频度），那么S的任意非空子集也是频繁的。反之，如果一个项集S的某个非空子集不是频繁的，则这个项集也不可能是频繁的。

19.**必考：Apriori算法，会把频繁项目集求出，会生成关联规则。**P129

DBMS重点总结

**一 数据处理的两种基本类型：操作型，分析型。**

操作型（与业务活动关联，事务处理）

操作型处理主要完成数据的收集，整理，存储，查询和增删改等操作，由一般工作人员和基层管理人员完成。

操作型数据处理主要用于企业的日常事务处理工作，存放的是细节的数据。数据的组织方式以方便事务处理、提高事务处理的性能为主要目标。

数据存储操作频率高而每次操作处理时间短，主要针对本部门业务相关数据，对企业范围内集成应用考虑较少。

分析型（与决策活动关联）

分析型处理是对数据的再加工，往往要访问大量的历史数据，进行复杂的统计分析，从中获取信息，因此也称为信息型处理，主要由高级管理人员完成。

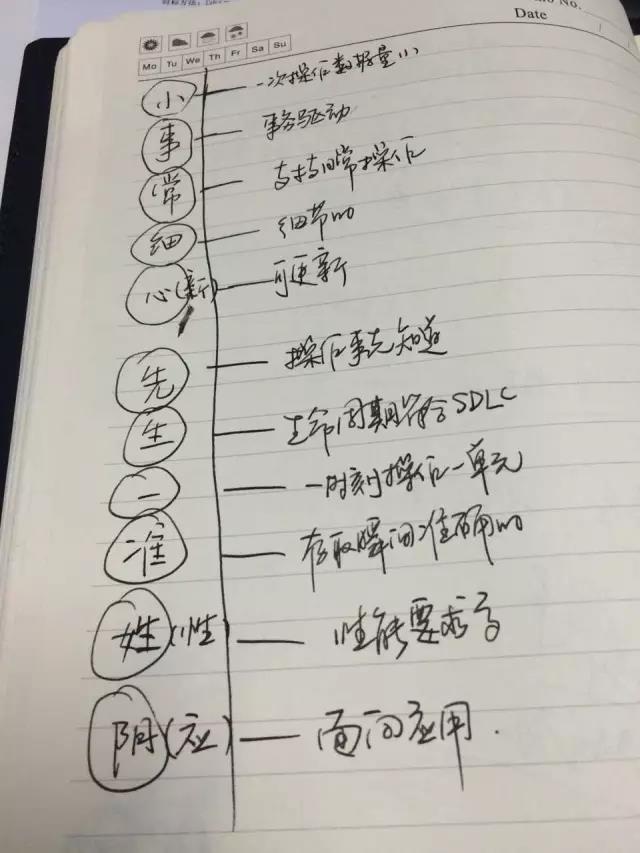
分析型处理主要用于企业的管理工作，存放的是历史和综合数据。

运行时间长，消耗系统资源多，数据全面集成，相关数据收集越完整越可靠。

**二 操作型数据和分析型数据的区别 书P4**

|  |  |
| --- | --- |
| 操作型数据 | 分析型数据 |
| 细节的 | 综合的，或提炼的 |
| 在存取瞬间是准确的（当前数据） | 代表过去的数据（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（系统生命周期） | 完全不同的生命周期 |
| 对性能要求高 | 对性能要求宽松 |
| 一个时刻操作一单元 | 一个时刻操作一集合 |
| 事务驱动 | 分析驱动 |
| 面向应用（业务处理） | 面向分析 |
| 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 |

**这个多不好记。有个小伙伴@观星自己总结的，我觉得有用，你们试试。**



**三 数据库系统的局限性 书P5**

数据的分散，“蜘蛛网”问题，数据不一致问题，数据动态集成问题，历史数据问题，数据的综合问题。

1.    数据的分散，即事务处理应用分散

由于多种原因，如经费有限，公司地点分散或企业兼并等，当前企业内部各事务处理的应用实际上几乎是是独立的，设计方案仅对于当前问题有效，没有良好的可扩展性

2.   “蜘蛛网问题”

不加限制的连续抽取导致企业的数据间形成了错综复杂的网状结构。时间基准不统一，抽取算法不相同，抽取级别不相同，参考不同的外部数据。

3.   数据不一致

同一字段不同数据类型，同一字段不同名字，同名字段不同含义

4.   数据的动态集成问题

5.   历史数据问题

联机事务处理系统中的数据往往是当前的，或近几年的。但对于决策分析，历史数据相当重要，许多分析以历史数据为依托。

6.   数据的综合问题

联机事务处理系统中的数据太细节，而决策系统中并不需要对细节数据进行分析。

**四 数据仓库的四个基本特征**

面向主题，集成，不可更新，随时间不断变化。

已有数据纪录不可更新；整体数据仓库内数据跟随时间变化不断增加新的数据，并且去除某一时间点之前的数据。只有读取，删除，插入操作，没有修改操作。

**面向主题**，根据要求将数据组织成一个完备的分析领域，即主题域。

**集成**，数据是从许多其他数据源中抽取得到的。然后再进行清洗，转换，整合，统一，最后加载到数据仓库中。

**不可更新**，数据的主要操作是查询，原始数据一旦进入数据仓库后，一般不允许再修改，且会被长期保留。只需要定期的加载和刷新。

**随时间不断变化**，   已有数据纪录不可更新，随时间变化不断增加新的数据，将不需要的数据从数据仓库中卸出，转储到其他设备，随着时间的变化不断地重新综合数据。

**五 数据仓库体系结构 书P14**

数据源，集成工具，数据仓库与数据仓库服务器，OLAP服务器，元数据与元数据管理工具，数据集市和前台分析工具等组成。

**六 ODS 书P22**

ODS主要是适应进行企业级的全局应用的需求而产生的。这种全局应用还可以大致地划分为两类：一类是进行企业级的联机事务处理，另一类可以称之为“即时OLAP”数据处理。

**七 数据粒度 书P31**

两种形式：

第一种粒度是对数据仓库中数据的**综合程度高低的一个度量**；粒度大小影响数据仓库效率，能回答询问的种类。粒度越小，细节程度越高，综合程度越低，回答查询的种类就越多。数据仓库是多粒度的，不同粒度回答不同的查询。

另一种特殊形式的粒度是**样本数据库**。

根据采样率的高低来划分。以一定的采样率从细节数据或轻度综合数据中**随机**抽取一个子集来代替数据源进行模拟分析。

**八 数据分割（分片／数据分片） 书P32**

将数据分布到各自的物理单元中，以便能分别独立处理，提高数据分析效率。

数据分割后的数据的单元称为**分片**。

数据分割一般而言总应包括日期。

数据分割后要达到的**目的**是，使数据更易于重构、索引、重组、恢复、监控和顺序扫描。

**九 数据仓库中数据的追加 书P32**

时标方法；DELTA文件；前后快照文件的方法；日志文件（推荐的方法）。

**十 维的构成：维成员，维层，维层次，维属性。 书P45**

维是人们观察数据的特定角度，是某个事物的属性。维是商业活动中的一个基本要素。每个维都有一个唯一的名字。

**维成员：**维由一些维成员构成。维的一个取值称为维的一个成员。每个成员有一个名字。

**维层：**不同的细节程度称为不同的维层

**维层次：**维层是一种分类方法。把每种分类方法叫做一个层次。

**维属性：**说明为成员所具有的特征。

**十一 度量： 书P46**

要分析的目标或对象。输入度量和导出度量，有可累计型和不可累计型。

**十二 常见的多维数据模型：书P49**

星形，雪片，事实群模型

**星形模型**：由一个很大的中心表和一组较小的表组成。不支持维的层结构，实现时将所有的维层属性存放在这一个表中，没有进行规范化。每个层有自己的属性，有很多冗余。当不同的维层有相同属性时只能使用换名方法，影响查询。

**雪片模型**：对维表进行规范化后形成，用多张维表描述一个复杂维，支持对不同层上的相同属性查询，易于维护而且节省存储空间。执行查询时需要进行较多的链接操作，可能影响系统的性能。

**事实群模型**（星系模型）：在复杂的应用中需要多个事实表共享维表，类似于星形模型集合。

**十三 聚集函数分类：书P52**

分布型聚集函数，代数型聚集函数，整体型聚集函数

**分布型聚集函数：**将数据分成n份，对其中每一份应用该函数，可以得到n个聚集值，对这n个聚集值进行计算得到的结果和整个数据（不划分）应用该函数得出结果一致。具有可累计的特性。

**代数型聚集函数**：一个函数可以由若干个分布型函数进行代数运算得出。

**整体型聚集函数**：一个函数不能由其他函数进行代数运算得出。

**十四 常用多维分析操作：书P53**

切片，切块，旋转，下钻，上卷。

**切片**：在数据方体的某一维上选定一个维成员的动作。

**切块**：在数据方体的某一维上选定某一区间的维成员的动作。

**旋转**：改变数据方体维的次序的动作。

**下钻**：在某个分析的过程中，用户需要从更多的维或者某个维的更细层次上观察数据。操作类型有两种，第一种为在现有的维上钻取到更细一层的数据；另一种是增加更多的维。

**上卷**：在某个分析的过程中，用户需要从更少的维或者某个维的更粗层次上观察数据。操作类型有两种，第一种为上卷到现有的某个维的更高层次去进行分析；另一种是减少一个维来进行分析。

**十五 数据方体的存储 书P66**

通常将基于多维数组存储的OLAP实现方式维MOLAP；基于关系表存储的OLAP实现方式称为ROLAP。星型模型就是基于关系表存储的ROLAP。

**十六 提高数据仓库效率的方法：**

1 合并表，2建立数据序列，3引入冗余，4进一步细分数据，5    生成导出数据，6建立广义索引，7 粒度划分，8分割

**十七 支持度： 书P123**（最好把书123页后面那道题也看看）

**支持度**：指项集X和Y在数据库D中同时出现的概率。(符合集数量／总样本数量)

**置信度**： 指在项集X出现的情况下，项集Y在数据库D中同时出现的条件概率，即Pr(Y/X)=Pr(XUY)/Pr(X)。(符合集数量／X样本数量)

**十八 预测建模：回归与分析  书P150**

**十九 聚类分析**  **书P180**

填空题：

1．数据仓库的四个基本特征是指数据仓库中的数据是 面向主题的 、集成的 、不可更新的和随时间不断变化的。

2．OLAP的实现方式有以下两种：基于关系数据库系统的实现和基于多维数据组织的实现。

3．数据从操作型环境到数据仓库过程中，通常需要进行的处理操作有数据抽取（extraction）、转换（transformation）、装载（Load）和清洗（cleaning）。

4．数据仓库中数据的分割是指将数据分散到各自的物理单元中去以便能分别独立处理。数据分割后的数据单元称为分片，数据分片的类型有 水平分片、垂直分片 、混合分片和导出分片等。

5．数据仓库系统是多种技术的综合体，它是由 数据仓库的前台后台工具、数据仓库服务器和 OLAP服务器三部分组成。

6．聚集函数分为三种类型，分别是分布型聚集函数、 代数型聚集函数 、               和  整体型聚集函数 。

7．粒度是数据仓库的重要概念，粒度越小，数据的细节程度越 高  ，可以回答查询的种类就越  多  ，但是查询效率将会很低；提高粒度将会提高查询效率，在数据仓库中通常采用多重粒度。

名次解释：

1．数据集市：数据集市包含企业范围数据的一个子集，对于特定的用户是有用的。其范围限于选定的主题。例如，一个商场的数据集市可能限定其主题为顾客、商品和销售。包括在数据集市中的数据通常是汇总的。

2．数据仓库的元数据：关于数据的数据.第一种：从操作型环境向数据仓库环境转换而建立的元数据。包含：所有源数据项名、属性及其在数据仓库中的转换.第二种：与终端用户的多维商业模型／前端工具之间建立映射的DSS元数据

3．粒度：对数据仓库中的数据的综合程度高低的一个度量,粒度越小，细节程度越高，综合程度越低,粒度大小影响数据仓库效率、能回答询问的种类,数据仓库是多粒度的，不同的粒度回答不同的查询

4．分割：指将数据分散到各自的物理单元中去以便能分别独立处理。

5．聚类分析：根据数据的特征找出数据间的相似性，将相似的数据分成一个类。

6．数据仓库的主题：主题是在较高层次上将企业信息系统中的数据综合、归类并进行分析利用的抽象。逻辑意义：对应企业中某一宏观分析领域所涉及的分析对象。

7．分类：描述或识别数据类或概念的模型(或函数)，以便能够使用模型预测类标号未知的对象。导出模型是基于对训练数据集（即，其类标号已知的数据对象）的分析。

8．序列模式分析：最长最大的子过程。GSP等方法。

多维分析中关联和序列模式的发现可以用来推动电信服务的发展。

例如：假设你想发现一系列电信服务的使用模式（按用户组，月或日历分组），按客户分组的呼叫，记录可以表现为如下形式：

（customer\_id，residense,office,time,date,service\_1,service\_2,…）

为了决定呼叫是否在两个特定的城市之间或特定的人群间发生，这样的一个序列模式，“如果一个洛杉矶地区的客户在和他居住地不同的另一个城市工作，它可能在每个工作日的下午五点先使用两个地区之间的长途服务，然后在接下来的时间里使用30 分钟的蜂窝电话”，可以通过上钻和下钻检测到。这有助于促进特定的长途销售额和蜂窝电话结合，可以用于扩展某个地区的特殊服务。

9．广义索引：预先计算出来的，用来记录具有某些特殊性质数据的索引。比如最小值，top-k值等。特点：非常小， 大大提高查询效率。

最流行的数据仓库数据模型是多维数据模型。这种模型可以以星形模式、雪花模式、或事实星座模式形式存在。

10．星型模型：最常见的模型范例星形模式；其中数据仓库包括（1）一个大的、包含大批数据、不含冗余的中心表（事实表）；（2）一组小的附属表（维表），每维一个。这种模式图很象星星爆发，维表围绕中心表显示在射线上。

11．OLAP中的维和维层次：观察数据的角度、程度不同分层。

12．雪片模型：雪花模式是星型模式的变种，其中某些维表是规范化的，因而把数据进一步分解到附加的表中。结果，模式图形成类似于雪花的形状。雪花模式和星形模式的主要不同在于，雪花模式的维表可能是规范化形式，以便减少冗余。这种表易于维护，并节省存储空间，因为当维结构作为列包含在内时，大维表可能非常大。

13．关联分析：关联分析发现关联规则，这些规则展示属性-值频繁地在给定数据集中一起出现的条件。关联分析广泛用于购物篮或事务数据分析。

问答题：

1．  操作型数据和分析型数据的主要区别是什么？

2．  你是如何理解数据仓库的数据是不可更新的，数据仓库的数据又是随时间不断变化的。

数据仓库中的数据不可更新是针对应用而言的,用户进行分析处理时是不进行数据更新操作的.数据仓库的数据是随时间的变化不断变化的,随时间变化不断增加新的数据内容,随时间变化不断删去旧的数据内容。数据仓库中包含有大量的跟时间有关的综合数据，经常按照时间段进行综合，随着时间的变化不断地进行重新综合。

3．  举例说明数据仓库有哪三类聚集函数。

（1）分布的：一个聚集函数是分布的，如果它能以如下分布方式进行计算：设数据被划分为n 个集合，函数在每一部分上的计算得到一个聚集值。如果将函数用于n 个聚集值得到的结果，与将函数用于所有数据得到的结果一样，则该函数可以用分布方式计算。例如，count()可以这样计算：首先将数据方分割成子方的集合，对每个子方计算count()，然后对这些子方得到的计数求和。因此，count()是分布聚集函数。同理，sum(), min()和max()是分布聚集函数。一个度量是分布的，如果它可以用分布聚集函数得到。

（2）代数的：一个聚集函数是代数的，如果它能够由一个具有M（其中，M 是一个整数界）个参数的代数函数计算，而每个参数都可以用一个分布聚集函数求得。例如，avg()可以由sum()/count()计算，其中sum()和count()是分布聚集函数。类似地，可以表明min\_N(), max\_N()和standard\_deviation()是代数聚集函数。一个度量是代数的，如果它可以用代数聚集函数得到。

（3）整体的：一个聚集函数是整体的，如果描述它的子聚集所需的存储没有一个常数界。即，不存在一个具有M 个（其中，M 是常数）参数的代数函数进行这一计算。整体函数的常见例子包括median(),mode()（即，最常出现的项），和rank()。一个度量是整体的，如果它可以用整体聚集函数得到。大部分数据方应用需要有效地计算分布的和代数的度量。对于这些，存在许多有效的技术。相比之下，有效地计算整体度量是很困难的。然而，对于有些整体函数的近似计算，有效的技术是存在的。

4．为什么说naive Bayesian分类法是 naïve的？

朴素贝叶斯分类假定一个属性值对给定类的影响独立于其它属性的值。该假定称作类条件独立。做此假定是为了简化所需计算，并在此意义下称为“朴素的”。

4．  请简述数据仓库的体系结构。

a)      数据仓库的后台工具

b)      数据仓库服务器

c)      OLAP服务器

d)      前台工具

5．  举例说明多维分析操作（drill-down，roll-up）的含义是什么？

一个通常用于数据仓库多维数据方，(a) 展示AllElectronics 的汇总数据 (b) 展示数据方(a)

上的下钻与上卷结果。

7．举例说明数据仓库有哪三类聚集函数（同3）

8．试述常用的数值属性离散化方法

         等宽：每个bin的距离间隔一样。

         等深：每个bin所具有的元组的数目相等。

         等质：bin的大小决定后，每一个bin中的元组是统一分布的

9．向数据仓库追加数据时，捕捉数据变化常用的途径有哪些？

数据追加：数据仓库的数据初装完成后, 再向数据仓输入数据的过程

追加内容：上次数据追加后在OLTP数据库中变化了的数据

变化数据的捕捉途径:

1）时标方法　 （如果数据含有时标，对新插入或更新的数据记录,加更新时的时标）

2）DELTA文件 （由应用生成DELTA文件，记录应用所改变的所有内容）

3）前后映象文件（抽取数据到数据仓库之后, 本次将抽取数据之前，对数据库分别作一次快照，比较两幅快照的不同，确定追加的数据）

4）日志文件（利用DB的固有机制，数据只限于日志文件，不用扫描整个数据库）

10．试述数据仓库的设计中提高数据仓库性能的方法和技术?

由于数据仓库的数据极少甚至不再更新，可采取如下技术

来提高数据仓库的性能：

1)合并表

2)建立数据序列

3)引入冗余

4)进一步细分数据

5)生成导出数据

6)建立广义索引

7)粒度划分

8)分割

11．简要说明事务处理环境不适宜DSS应用的原因？

操作型处理也叫事务处理，是指对数据库联机的日常操作，通常是对一个或一组记录的查询和修改，主要是为企业的特定应用服务的，人们关心的是响应时间，数据的安全性和完整性。

分析型处理则用于管理人员的决策分析。例如：DSS（决策支持系统）, EIS（主管信息系统）和多维分析等，经常要访问大量的历史数据。

事务处理环境不适宜DSS应用的原因：

  (1) 事务处理和分析处理的性能特性不同（在事务处理环境中，数据的存取操作频率高而每次操作处理的时间短, 在分析处理环境中，DSS应用需要运行时间长，消耗系统资源多.）

  (2) 数据集成问题（DSS需要的数据：全面、集成、相关数据收集得越完整 结果就越可靠）

  (3) 数据动态集成问题（事务处理的数据：与本部门业务有关当前数据,对整个企业范围内的集成应用考虑少,当前企业内数据的状况[a.分散而非集成——这是事务处理环境所固有的 b.事务处理应用产生的细节数据不能成为统一的整体 c.DSS应用必须在应用程序中进行数据集成）

  (4) 历史数据问题（事务处理系统中的数据：当前数据及短期数据；决策分析的数据：必须要历史数据）

  (5) 数据的综合问题（DSS系统的分析对象:一般不对细节数据进行分析,分析前需要对细节数据进行不同程度的综合.

事务处理系统的对象:只关心细节数据，不具备综合能力,综合是一种数据冗余，需要加以限制）

12．数据仓库的设计方法与操作型环境中系统设计采用的系统生命周期法有什么不同？

SDLC-操作型环境中，业务过程和规则比较规范和固定。系统设计人员能够清晰地了解应用的需求和数据流程，系统的设计一般采取系统生命周期法 (Systems Development Life Cycle)

CLDS-分析型环境中，DSS分析对决策分析的需求不能预先作出规范说明，只能给设计人员一个抽象模糊的描述。设计人员必须在与用户不断的交流中，将系统需求逐步明确与完善。为了强调这种开发的不确定型，将此设计方法定名为CLDS方法（与SDLC相反）

13．举例说明多维分析操作（切片、切块、旋转）的含义是什么？

切片和切块：切片操作在给定的数据方的一个维上进行选择，导致一个子方。图2.10 图示了一个对维time 的切片操作，它对中心数据方使用条件time = ”Q1” 选择销售数据。切块操作通过对两个或多个维执行选择，定义子方。图2.10 图示了一个切块操作，它涉及三个维，根据如下条件对中心表切块：(location = ”Montreal”or ’Vancouver”) and (time = ”Q1” or ”Q2”) and (item = ”homeentertainment” or”computer”)。

转轴：转轴（又称旋转）是一种目视操作，它转动数据的视角，提供数据的替代表示。图2.10 给出一个转轴操作，这里item 和location 在一个2-D 切片上转动。其它例子包括转动3-D 数据方，或将一个3-D 立方转换成2-D 平面序列。

14．数据挖掘的步骤是什么？

确定挖掘对象，准备数据，建立模型，数据挖掘，结果分析，知识应用阶段

数据挖掘作为KDD（知识发现）的一个步骤。

KDD 是一个以知识使用者为中心，人机交互的探索过程，包括了在指定的数据库中用数据挖掘算法提取模型，以及围绕数据挖掘所进行的预处理和结果表达等一系列的步骤。

15．简要说明数据仓库环境中元数据的内容。

元数据是关于数据的数据。在数据仓库中，元数据是定义仓库对象的数据。对

于给定数据仓库的数据名和定义，创建元数据。其它元数据包括对提取数据添加的时间标签、提取数据的源、被数据清理或集成处理添加的字段等。

16．企业的数据库体系化环境的四个层次是什么？它们之间的关系是什么？

层次的体系化环境四个层次分别为：操作型环境、全局级数据仓库、部门级的局部仓库、个人级数据仓库。

操作型环境存放：细节的操作型数据，服务于高性能事务处理

全局级数据仓库：存放细节数据、导出数据

部门级局部仓库：一般存放导出数据

个人级数据仓库：数据一般是暂时存放，用于启发式分析。

17．简要说明数据仓库设计的步骤。

数据仓库的设计方法：CLDS方法(与SDLC相反) 参考12题

18．简要说明异常点挖掘有哪些方法？  
19．什么是元数据？简要说明数据仓库环境中元数据的内容。（参考15）

20．你是如何理解数据仓库的数据是不可更新的，数据仓库的数据又是随时间不断变化的。

21．什么是数据驱动的系统设计方法？（参考12）

22．试简述数据仓库的体系结构（参考4）

23．简述采用决策树方法进行分类的过程。

1)基本算法 (贪婪算法)

由上到下，分而治之，递归构造树

开始时，所有的训练样本都在树根

属性都是可分类的属性(如果是连续值的话，首先要对其进行离散化)

根据选择的属性，对样本递归地进行划分

在启发式或统计度量（如 information gain)的基础上选择测试属性

2)停止划分的条件

某个节点上的所有样本都属于相同的类

所有的属性都用到了– 这时采用多数有效的方法对叶子节点进行分类

没有样本了

24．请简述采用神经元网络进行分类的过程。

在开始训练之前，用户必须说明输入层的单元数、隐藏层数（如果多于一层）、每一隐藏层的单元数和输出层的单元数，以确定网络拓扑。

定义网络拓扑

向传播算法学习过程：

迭代地处理一组训练样本，将每个样本的网络预测与实际的类标号比较。

每次迭代后，修改权值，使得网络预测和实际类之间的均方差最小

25．举一个实例说明如何确定数据仓库的主题，如何确定主题所应包含的数据内容？

数据仓库中的数据是面向主题进行组织的

什么是主题

主题是在较高层次上将企业信息系统中的数据综合、归类并进行分析利用的抽象

逻辑意义：对应企业中某一宏观分析领域所涉及的分析对象

面向主题的数据组织步骤

    1、 抽取主题: 按照分析的要求来确定

    2、 确定每个主题所应包含的数据内容

例如：商场商品采购

 1）在OLTP数据库中, “订单”“订单细则” “供应商”  三个数据库模式清晰完整地描述了一笔采购业务所涉及的数据内容, 这是面向应用来进行数据组织的方式；

2）在数据仓库中,主要是进行数据分析处理, 商品采购时的分析活动主要是要了解各供应商的情况, “供应商”是采购分析时的分析对象。所以不需要组织象“订单”和“订单细则”这样的数据库模式，因为它们包含的是纯操作型的数据；但是仅仅只用OLTP数据库的“供应商”中的数据又是不够的, 因而要重新组织“供应商”这么一个主题。

26．举例说明数据仓库的多粒度。

粒度分为两种形式：

1）对数据仓库中的数据的综合程度高低的一个度量 粒度越小，细节程度越高，综合程度越低，粒度大小影响数据仓库效率、能回答询问的种类，数据仓库是多粒度的，不同的粒度回答不同的查询

2)  样本数据库

根据采样率的高低来划分粒度：以一定的采样率从细节数据或轻度综合数据中抽取的一个子集。用处：代替源数据进行模拟分析。抽样的方法：随机抽取

优点：高效率；源数据量很大的情况下，抽样数据可以大大下降，分析结果误差极小，有助于抓住主要因素和主要矛盾

实际：两种形式的粒度都存在

例：“商品”主题表的划分：

销售综合表和采购综合表是属于第一种形式的粒度（时间段上信息的综合）

库存信息的不同表：则属于第二种形式粒度划分（不同时点上的粒度）

27．举例说明什么是“星星模式”。（名词解10）

28．常用的聚类方法有哪些？（分步，代数，整体，见第3题）

http://www.doc88.com/p-1773032734188.html