填空题

第一章：

1.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。

2.元数据是描述数据仓库内数据的结构和建立方法的数据，它为访问数据仓库提供了一个信息目录，根据元数据用途的不同可将数据仓库的元数据分为技术元数据和业务元数据。

3.数据处理通常分为两大类，联机事务处理（OLTP）和联机分析处理(OLAP)。

4.多维分析是指以“维”形式组织起来的数据（多维数据集），采取钻取、切片、切块和旋转等各种分析动作，以求剖析数据，使用户能从不同角度、不同侧面观察数据仓库中的数据，从而深入理解多维数据集中的信息。

5.ROLAP是基于关系数据库的OLAP实现，而MOLAP是基于多维数据结构组织的OLAP实现。

6.数据仓库按照其开发过程，其关键环节包括数据抽取、数据存储与管理和数据表现等。

7.数据仓库系统的体系结构根据应用需求的不同，可以分为以下4种类型，两层架构、独立型数据集市、依赖型数据集市和操作型数据存储、逻辑型数据集市和实时数据仓库。

8.操作型数据存储实际上是一个集成的、面向主题的、可更新的、当前值的（但是可“挥发”的）、企业级的、详细的数据库，也叫运营数据存储。

9.“实时数据仓库”意味着源数据系统、决策支持服务和数据仓库之间以一个接近实时的速度交换数据和业务规则。

10.从应用的角度看，数据仓库的发展演变可以归纳为5个阶段：以报表为主、以分析为主、以预测模型为主、以运营导向为主和以实时数据仓库和自动决策为主。

第二章：

1.调和数据是存储在企业级数据仓库和操作型数据存储中的数据。

2.抽取、转换、加载过程的目的是为决策支持应用提供一个单一、权威数据源。因此，我们要求ETL过程产生的数据（即调和数据层）是详细的、历史的、规范化的、可理解的、即时的和质量可控制的。

3.数据抽取的两个常见类型是静态抽取和增量抽取。静态抽取用于最初填充数据仓库，增量抽取用于进行数据仓库的维护。

4.粒度是对数据仓库中数据的综合程度高低的一个衡量。粒度越小，细节程度越高，综合程度越低，回答查询的种类越多。

5.使用星型模式可以从一定程度上提高查询效率。因为星型模式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。

6.维度表一般由主键、分类层次和描述属性组成。对于主键可以选择两种方式：一种是采用自然键，另一种是采用代理键。

7.雪花型模式是对星型模式维表的进一步层次化和规范化来消除冗余的数据。

8.数据仓库中存在不同综合级别的数据。一般把数据分成4个级别：早期细节级、当前细节级、轻度综合级和高度综合级。

第三章：

1.SQL Sever SSAS 提供了所有业务数据的统一整合视图，可以作为传统报表、在线分析处理、关键性能指示器记分卡和数据挖掘的基础。

2.数据仓库的概念模型通常采用信息包图法来进行设计，要求将其5个组成部分（包括名称、维度、类别、层次和度量）全面地描述出来。

3.数据仓库的逻辑模型通常采用星型图法来进行设计，要求将星型图的各类逻辑实体完整地描述出来。

4.按照事实表中度量的可加性情况，可以把事实表对应地事实分为4种类型：事务事实、快照事实、线性项目事实和事件事实。

5.确定了数据仓库的粒度模型以后，为提高数据仓库地使用性能，还需要根据用户需求设计聚合。

6.在项目实施时，根据事实表的特点和用户的查询需求，可以选用时间、业务类型、区域和下属组织等多种数据分割类型。

7.当维表中的主键在事实表中没有与外键关联时，这样的维称为退化维，它与事实表并无关系，但有时在查询限制条件（如订单号码、出货编号等）中需要用到。

8.维度可以根据其变化快慢分为无变化维度、缓慢变化维度和剧烈变化维度三类。

9.数据仓库的数据量通常较大，且数据一般很少更新，可以通过设计和优化索引结构来提高数据存取性能。

10.数据仓库数据库常见的存储优化方法包括表的归并与簇文件、反向规范化，引入冗余、表的物理分割（分区）。

第五章：

1.分类的过程包含获取数据、预处理、分类器设计和分类器测试。

2.分类器设计阶段包含三个过程：划分数据集、分类器构造和分类器设计。

3.分类问题中常用的评价准则有精确度、查全率和查准率、F-measure、几何均值。

4.支持向量机中常用的核心函数有多项式核函数、径向基核函数、S型核函数。

第六章：

1.聚类分析包括连续型属性、二值离散型属性、多值离散型属性以及混合类属性四种类型描述属性的相似度计算方法。

2.连续型属性的数据样本之间的距离有欧式距离、曼哈顿距离和明考斯基距离。

3.略 书上P138

4.层次聚类方法包括凝聚型层次聚类和分解型层次聚类。

简答题（每小题5分，共30分）

1.什么是数据仓库的数据ETL过程？

答：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。整个过程由抽取、清洗、转换、加载和索引四个步骤组成。

2.贝叶斯网络的三个主要议题是什么？

答：贝叶斯的三个主要议题是预测、诊断和学习

3.什么是聚类分析？（2分）聚类分析的应用领域有哪些？（3分）

答: 聚类分析是将物理的或者抽象的数据集合划分为多个类别的过程，聚类之后的每个类别中任意两个数据样本之间具有较高的相似度，不同类别的数据样本之间具有较低的相似度。

聚类分析在科学数据分析、商业、生物学、医疗诊断、文本挖掘、Web数据挖掘等域都有广泛应用

4.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？

答：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 (CIA, 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。

5.简单遗传算法包括哪些步骤？

答：（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，否则执行第（2）步。

6.前馈网络和递归网络有什么本质区别?

答：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。

7.请比较PCA和LDA的区别。

答：LDA：有类别的数据进行降维，同一距离要小，不同距离要大

PCA：降维之后方差最大。

8.请简述数据挖掘中关联规则Apriori算法的思想。

答：其发现关联规则分两步，第一是通过迭代，检索出数据源中所有烦琐项集，即支持度不低于用户设定的阀值的项即集，第二是利用第-步中检索出的烦琐项集构造出满足用户最小信任度的规则。

9.请分析特征选择和特征提取有何区别?

答:特征选择定义为从有N个特征的集合中选出具有M个特征的子集，并满足条件M≤N。特征选择能够为特定的应用在不失去数据原有价值的基础上选择最小的属性子集，去除不相关的和冗余的属性。

特征提取广义上指的是一种变换，将处于高维空间的样本通过映射或变换的方式转换到低维空间，达到降维的目的。它可以从一组特征中去除冗余或不相关的特征来降维。

10.TF-IDF算法是什么，有什么实际意义?

答：TF:词频 IDF：逆向文件频率。 实际意义：进行文本分类和文本挖掘。

11.数据挖掘与统计的区别与联系?

答：区别：数据挖掘和统计的区别主要是数据量的区别，数据挖掘是大数据和建模，统计是小数据。

联系：都是数据分析

聚类和分类有什么区别和联系?

答：区别：聚类是所有样本两两之间距离，没有类别属性。分类是事先定义好类别，类别数不变。用每一个样本特征与类别特征的距离，谁小用谁。

联系：处理过程一样

12.什么是分类？（2分）分类的应用领域有哪些？（3分）

答：分类是指把数据样本映射到一个事先定义的类中的学习过程，即给定一组输入的属性向量及其对应的类，用基于归纳的学习算法得出分类。

分类问题在商业、银行业、医疗诊断、生物学、文本挖掘和因特网筛选等领域都有广泛应用。

13.什么是信息包图法？（2分）它为什么适用于数据仓库的概念模型的设计？（3分）

答：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，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。

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。

14.关联规则的分类有哪些？（2分）关联规则挖掘的步骤包括什么？（3分）

答：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。

15.什么是关联规则？（2分）关联规则的应用有哪些？（3分）

答：用来发现超级市场中用户购买的商品的隐含关联关系，并用规则的形式表示出来，称为关联规则。

应用：发现超级市场中用户购买的商品的隐含关联关系、文本挖掘、商品广告邮寄分析和网络故障分析。

计算题（每小题10分，共30分）

1.给定下表所示的一个事物数据库，写出Apriori算法生成频繁项目集，强关联规则的过程（假定最小支持度=0.5，最小置信度=0.5）。

|  |  |
| --- | --- |
| TID | 项目集 |
| 1 | a,b,c |
| 2 | a,c |
| 3 | a,d |
| 4 | b,e,f |

2.根据下表，利用ID3算法生成决策树，即选择根节点的属性。

答：(1)由I={a,b,c,d,e,f}的所有项目直接产生1-候选集C1,计算其支持度，去除支持度小于supmin的项集，形成1-频繁集L1。

|  |  |  |  |
| --- | --- | --- | --- |
| 项集C1 | 支持度 | 项集L1 | 支持度 |
| {a} | 3/4 | {a} | 3/4 |
| {b} | 2/4 | {b} | 2/4 |
| {c} | 2/4 | {c} | 2/4 |
| {d} | 1/4 |  |  |
| {e} | 1/4 |  |  |

(2)为发现频繁2-项集L2，首先利用L1中各项目组合连接，来产生2-候选集C2；然后扫描记录集，以获得C2中各项集的支持度。去除支持度小于supmin的项集，形成2-频繁集L2。

|  |  |
| --- | --- |
| 项集C2 | 支持度 |
| {a,b} | 1/4 |
| {a,c} | 2/4 |
| {b,c} | 0 |

结论：a->c:2/3 c->a:100%

|  |  |  |  |
| --- | --- | --- | --- |
| 年龄 | 收入 | 信誉度 | 买保险 |
| ≤40 | 高 | 良 | c2 |
| ≤40 | 高 | 优 | c2 |
| >50 | 中 | 良 | c1 |
| >50 | 低 | 良 | c1 |
| >50 | 低 | 优 | c2 |
| 41~50 | 低 | 优 | c1 |
| ≤40 | 中 | 良 | c2 |
| ≤40 | 低 | 良 | c1 |
| >50 | 中 | 良 | c1 |
| ≤40 | 中 | 优 | c1 |
| 41~50 | 中 | 优 | c1 |
| 41~50 | 高 | 良 | c1 |
| >50 | 中 | 优 | c2 |

I(买保险)=-5/13\*-8/13\*

I(年龄，买保险)=5/13\*（3/5\*-2/5\*）

+3/13（3/3\*）-0/3\*

+5/13（3/3\*）-0/3\*

I(收入，买保险)= 4/13\*（3/4\*-1/4\*）

+6/13（4/6\*）-2/6\*

+3/13（2/3\*）-0/3\*

I（信誉度，买保险）=7/13\*（2/7\*-5/7\*）

+6/13（3/6\*）-3/6\*

Gain（年龄）=I（买保险）-I（年龄，买保险）

Gain（收入）=I（买保险）-I（收入，买保险）

Gain（信誉度）=I（买保险）-I（信誉度，买保险）

3.某电子设备厂所用的元件是由三家元件厂提供的，根据以往的记录，这三个厂家的次品率分别为0.02、0.01、0.03，提供元件的份额分别为0.15、0.8、0.05。设这三个厂家的产品在仓库是均匀混合的，且无区别的标志。

问题1：在仓库中随机地取一个元件，求它是次品的概率。（5分）

问题2：在仓库中随机地取一个元件，若已知它是次品，为分析此次品出自何厂，需求出此元件由三个厂家生产的概率是多少？（5分）

答：设A表示“取到的是一只次品”，Bi（i=1，2，3）表示“所取到的产品是由第i家工厂提供的”

易知B1,B2,B3样本空间S的一个划分，且有

P(B1)=0.15 P(B2)= 0.8 P(B3)=0.05

P(A| B1)=0.02 P(A| B2)=0.01 P(A| B3)=0.03

(1)由全概率公式得P(A)= P(A| B1)\* P(B1)+ P(A| B2)\* P(B2)+ P(A| B3)\* P(B3)

=0.02\*0.15+0.01\*0.8+0.03\*0.05

=0.003+0.008+0.0015=0.0125

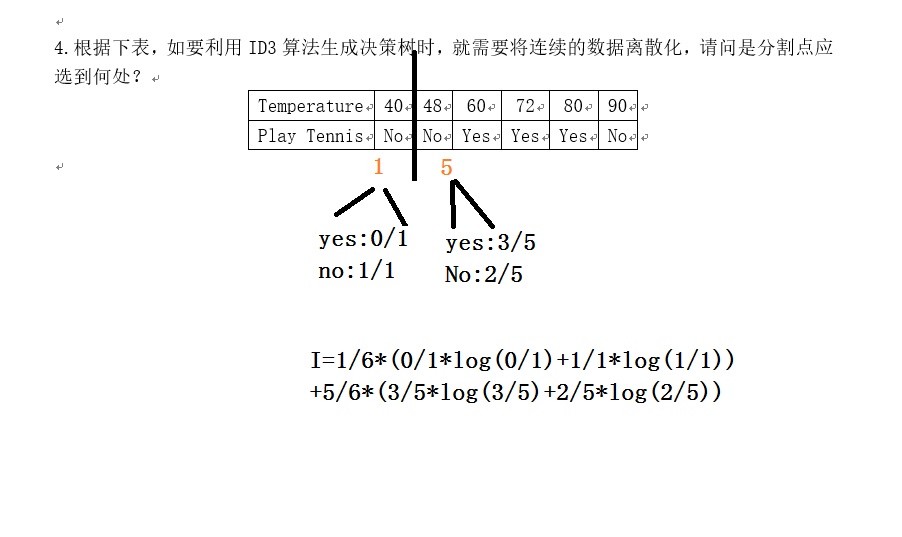
(2)P(B1|A)= P(A| B1)\* P(B1)/P(A)=0.003/0.0125=0.24

P(B2|A)= P(A| B2)\* P(B2)/P(A)=0.008/0.0125=0.64

P(B3|A)= P(A| B3)\* P(B3)/P(A)=0.0015/0.0125=0.12

4.根据下表，如要利用ID3算法生成决策树时，就需要将连续的数据离散化，请问是分割点应选到何处？

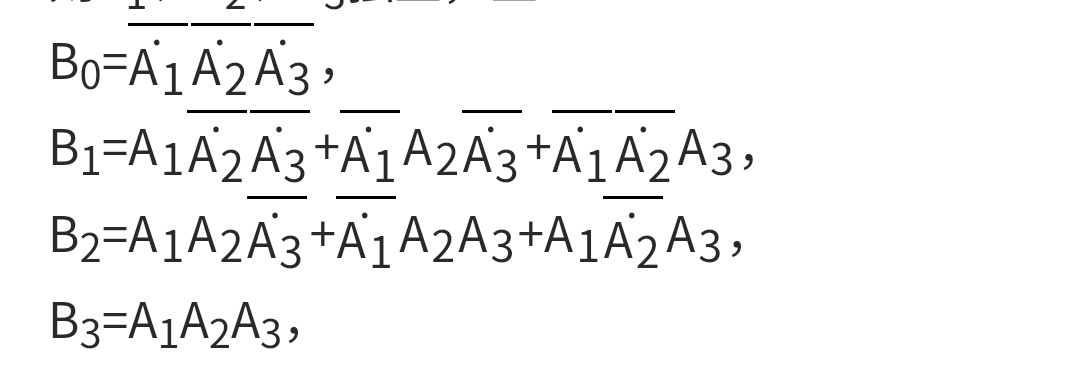
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Temperature | 40 | 48 | 60 | 72 | 80 | 90 |
| Play Tennis | No | No | Yes | Yes | Yes | No |



5.甲乙丙三人向同一飞机射击。设甲、乙、丙射中的概率分别为0.4、0.5和0.7。又设只有一人射中，飞机坠落的概率为0.2；若有两人射中，飞机坠落的概率为0.6；若有三人射中，飞机必坠落。求飞机坠落的概率。

 解：设A1、A2、A3分别是甲、乙、丙击中飞机，Bi（i=1，2，3）表示有i个人击中飞机，H表示飞机被击落。

则A1、A2、A3独立，且



于是，

P(B0)=(1-0.4)\*(1-0.5)\*(1-0.7)=0.09

P(B1)=0.4\*(1-0.5)\*(1-0.7)+(1-0.4)\*0.5\*(1-0.7)+(1-0.4)\*(1-0.5)\*0.7=0.36

P(B2)=0.4\*0.5\*(1-0.7)+(1-0.4)\*0.5\*0.7+0.4\*(1-0.5)\*0.7=0.41

P(B3)=0.4\*0.5 \*0.7=0.14

依题意有：

P(H|B0)=0, P(H|B1)=0.2 P(H|B2)=0.6 P(H|B3)=1

于是，由全概率公式有：

P(H)==0.09\*0+0.36\*0.2+0.41\*0.6+0.14\*1=0.458

论述题（每小题15分，共30分）

**1.请列出3种数据仓库产品，并说明其优缺点。**

(1)Cognos:

优点：提供可以简洁部署且具有交互性的PowerPlay Web Explorer界面。

缺点：是一个相对封闭的产品，它有自己的客户端Web Explorer。

(2)Essbase:

优点：提供丰富的API，具有几百个计算公式，支持多种计算，用户可以自己构建复杂的查询，可以充分定制开发。

缺点：开发难度较大，部署也不容易。

(3)Express:

优点：有内建的分析函数

缺点：使用比Discoverer要复杂。

**2.什么是信息包图法？（7分）它为什么适用于数据仓库的概念模型的设计？（8分）**

答：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，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。

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。

3.**谈一谈你对数据挖掘未来发展趋势的看法**。

答：当前，数据挖掘技术正方兴未艾，预计在21世纪还会形成更大的高潮，未来的发展趋势会是以下几个方面：

（1）形式化描述的语言，即研究专门用于知识发现的数据挖掘语言DMQL，类似SQL语言一样走向形式化和标准化；

（2）可视化的数据挖掘过程，寻求数据挖掘过程中的可视化方法，使知识发现的过程易于被用户理解和操纵，可使数据挖掘过程成为用户业务流程的一部分，也便于在知识发现的过程中进行人机交互；包括数据用户化呈现与交互操纵两部分。

（3）融合各种异构数据的挖掘技术，加强对各种非结构化数据的开采（Data Mining for Audio＆Video），如对文本数据、图形数据、视频图像数据、声音数据乃至综合多媒体数据的开采；

（4）处理的数据将会涉及到更多的数据类型，这些数据类型或者比较复杂，或者是结构比较独特。为了处理这些复杂的数据，就需要一些新的和更好的分析和建立模型的方法，同时还会涉及到为处理这些复杂或独特数据所做的费时和复杂数据准备的一些工具和软件。

（5）交互式发现；

（6）知识的维护更新。

**4.请列出3种数据挖掘中学过的分类方法，并说明其优缺点。**

（1）K-近邻分类方法：

优点：

1、是一种非常典型的分类监督学习算法，它可以解决多分类的问题

2、思想简单、效果强大。

3、可以用来解决回归问题。

缺点：

1、效率低下

2、高度数据相关

3、数据预测结果不具备可解释性

4、维数灾难

（2）决策树分类方法：

优点：

1、进行分类器设计时，决策树分类方法所需时间相对较少。

2、决策树的分析模型是树状结构，简单直观，比较符合人类的理解方法。

3、可以将决策树种到达每个叶结点的路径转换为IF-THEN形成的分类规则，这种形式更有利于理解。

缺点：

1、对连续性的字段比较难预测；

2、对有时间顺序的数据，需要很多预处理的工作；

3、当类别太多时，错误可能就会增加的比较快；

4、一般的算法分类的时候，只是根据一个字段来分类。

（3）支持向量机：

优点：

1、SVM是一种有坚实理论基础的新颖的适用小样本学习方法。

2、计算的复杂性取决于支持向量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾难”。

3、SVM学习问题可以表示为凸优化问题，因此可以利用已知的有效算法发现目标函数的全局最小值,

4、有优秀的泛化能力。

缺点：

1、对大规模训练样本难以实施

2、解决多分类问题困难

3、对参数和核函数选择敏感

**5.列举几项你知道的数据挖掘应用，并论述数据挖掘在其中的作用?**

(1)分类，根据特征判断对象属于哪个类别，有指导学习。预测肿瘤细胞是良性还是恶性;识别信用卡交易是否合法还是欺诈;电信客户流失分析;图片、音频、视频标签;蛋白质结构功能分类等。

(2)聚类，给对象归类使得同组对象尽可能相似，不同组对象尽可能不相似，无指导学习。把相关文档归并方便浏览;市场分割，细分为不同的客户群;获取价格波动相似的股票有助于决策。

(3) 关联分析，给定一组记录，分析项目之间的依赖关系。购物分析，用于促销、货价管理存货管理:医疗信息发现与某种疾病与症状的关联以便通过症状诊断病症。

(4)顾客分类，数据挖掘能够告诉我们什么样的顾客买什么产品(聚类或分类)识别顾客需求，对不同的顾客识别最好的产品，使用预测发现什么因素影响新顾客。汽车保险检测假造事故骗取保险赔偿的人。检测电话欺骗，通话距离、通话时间，每天或每周通话次数。

**6.简述你对数据仓库未来发展趋势的看法。**

答：数据仓库是数据管理技术和市场上一个方兴未艾的领域，有着良好的发展前景。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。在数据管理方面，未来的发展将使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。在数据表现方面，统计分析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。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。

计算机应用技术发展的数据仓库倾向是数据仓库发展的推动力。在市场方面，将厂商和用户两个方面看数据仓库的发展。从用户的角度看，除了数据管理的传统领域是数据仓库的主要市场之外，数据仓库的应用随着现代社会商业模式的变革而进一步普及和深入。

总之，数据仓库是一项基于数据管理和数据应用的综合性技术和解决方案，它是数据库市场的新一轮增长点，同时也是未来企业应用系统的重要组成部分。