**商务智能复习题答案**

1、商务智能概念

**答**：工业界：商务智能可以被看作是一类技术或工具，利用它们可以对大量的数据进行收集、管理、分析和挖掘，以改善业务决策水平，增强企业的竞争力；

学术界：商务智能是一套理论、方法和应用，通过它们可以快速地发现海量数据中隐含的各种知识，有效地解决企业面临的管理和决策问题， 支持企业的战略实施。

商务智能的系统构成六个主要组成部分:数据源,数据仓库,在线分析处理,数据探查(灵活的查询、即时报表以及统计方法等, 该类方法属于被动分析方法),

数据挖掘(是从大量数据中自动发现隐含的信息和知识的过程，属于主动分析方法，不需要分析者的先验假设，可以发现未知的知识.常用的分析方法包括分类、聚类、关联分析、数值预测、序列分析、社会网络分析等),

业务性能管理

2、支持度、置信度和提升度的概念及计算方法

**答**：支持度：表示{x,y}同时出现的概率，support=同时购买{x,y}的人数/总人数

置信度：购买x的人，同时购买y的概率，confidence(x->y)=同时购买{x,y}的人数/购买x的人数

提升度：提升度表示含有X的条件下，同时含有Y的概率，与不含X的条件下却含Y的概率之比。

               Lift(X→Y) = P(Y|X) / P(Y)

3、关联分析的方法

答：逐层发现算法Apriori，无候选集发现算法FP-growth

4、分类和预测

分类（classification）是通过对具有类别的对象的数据集进行学习，概括其主要特征，构建分类模型，根据该模型预测对象的类别的一种数据挖掘和机器学习技术。

例如，电信公司的客户可以分为两类，一类是忠诚的，一类是流失的。根据这两类客户的个人特征方面的数据以及在公司的消费方面的数据，利用分类技术可以构建分类模型

答：相同点：两者都是有监督学习，解决问题的过程是相同的

不同点：分类问题预测的是类别，数值预测预测的是数值，通常是连续型的数值

5、聚类的概念

聚类（clustering）是依据物以类聚的原理，将没有类别的对象根据对象的特征自动聚集成不同簇的过程，使得属于同一个簇的对象之间非常相似，属于不同簇的对象之间不相似。 其典型应用是客户分群

答：聚类分析是指给定一组对象的描述信息，发现具有共同特性的对象构成簇

6、为什么要进行数据预处理？数据预处理都有哪些主要工作？

答：原因：1、可能存在数据缺失、错误、不一致问题；2、数据不能很好反应潜在的模式；3、数据属性存在冗余；

处理工作：数据规范化，数据离散化，数据清洗，特征提取与特征选择

7、相似性的度量方法

答：基于距离的相似度，余弦相似度，相关性相似度，Jaccard系数

8、什么是数据仓库，数据仓库都有哪些主要模型？

答：数据仓库：是在企业管理和决策中面向主题的、集成的、随时间变化的、稳定的数据集合。

模型：星型、雪片、事实星座

1、星型模型

星型模型是一种由一点向外辐射的建模范例，中间有一单一对象沿半径向外连接到多个对象。星型模型反映了最终用户对商务查询的看法:销售事实、赔偿、付款和货物的托运都用一维或多维描述(按月、产品、地理位置)。星型模型中心的对象称为“事实表”，与之相连的对象称为“维表”。对事实表的查询就是获取指向维表的指针表，当对事实表的查询与对维表的查询结合在一起时，就可以检索大量的信息。通过联合，维表可以对查找标准细剖和聚集。

2、雪花模型

雪花模型是对星型模型的扩展，每一个点都沿半径向外连接到多个点.雪花模型对星型的维表进一步标准化，它的优点是通过最大限度的减少数据存储量以及把较小的标准化表(而不是大的非标准化表)联合在一起来改善查询性能。化及维的较低的粒度，雪花模型增加了应用程序的灵活性。

3、混合模型

混合模型是星型模型和雪花模型的一种折衷模式，其中星型模型由事实表和标准化的维表组成，雪花模型的所有维表都进行了标准化。在混合模型中，只有最大的维表才进行标准化，这些表一般包含一列列完全标准化的(重复的)数据。

**8.1说明数据仓库与数据库的区别与联系**

联系表现在两个方面

- 数据仓库中的大部分数据来自于业务系统的数据库中

- 当前绝大多数数据仓库都是利用数据库系统来管理的

区别：构建目的、管理的数据、管理方法都不同

-数据库主要用于实现企业的日常业务运营，提高业务运营的效率；数据仓库的构建主要用于集成多个数据源的数据，这些数据最终用于分析

-数据库通常只包含当前数据，数据的存储尽量避免冗余，数据的组织按照业务过程涉及的数据实现，是应用驱动的。数据仓库中的数据是按照主题组织的，将某一主题的所有数据集成在一起，数据存在冗余

数据库中的数据需要进行频繁的插入、删除、修改等更新操作，需要复杂的并发控制机制保证事务运行的隔离性。

- 数据仓库中的数据主要用于分析处理，除了初始的导入和成批的数据清除操作之外，数据很少需要更新操作

- 数据库中数据的更新操作的时效性很强，事务的吞吐率是个非常重要的指标。而数据仓库的数据量十分庞大，分析时通常涉及大量数据，时效性不是最关键的。数据仓库中的数据质量非常关键，不正确的数据将导致错误的分析结果。

9、什么是多维数据分析，都有哪些主要操作类型？

答：多维分析就是选择某些维度中的某些层次，交互式地探察某些度量值的过程

操作类型：切片、切块、上卷、下钻、旋转等

多维分析是按照思考业务的方式来探察数据，从而加速用户理解和作出决策的能力。

❖ 用多维分析，通过对数据的导航，简便直观的将维和度量结合起来，通过切片、切块、向上钻取、向下钻取、钻透等操作快速处理信息。

切片：固定某一维的取值为其中的一个类别，其他维不变得到的立方体称为一个切片。

切块（dice）：固定某一维的取值为其中的一个区间，其他维不变得到的立方体称为一个切块。

旋转（rotate）的三种情况：- 行列交换- 行维或列维移动到列维或行维上。- 将行维或列维用其他的维代替，就好象在旋转立方体一样。

10、OLAP、 OLTP、 ETL、 KDD 缩略语的含义。

答：OLAP：在线分析处理（online analytical processing），是数据仓库系统的主要应用，提供数据的多维分析以支持决策过程。 业务性能度量可以通过多个维度、多个层次进行多种聚集汇总，通过交互的方式发现业务运行的关键性能指标的异常之处。多维数据可以进行多种操作如切片、切块、下钻、上卷等.

OLTP：在线事务处理（online transaction processing），是数据库管理系统的主要功能，用于完成企业内部各个部门的日常业务操作。

ETL：是英文 Extract-Transform-Load 的缩写，用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程

KDD：知识发现（Knowledge Discovery in Database）是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。



15、了解商务智能系统开发的过程。

答：可以分为规划、需求分析、设计、实现四个阶段

**规划：识别业务需求、识别信息需求、时间成本规划**

每类需求，重要性和实现的难易程度

-重要性方面，可以从三个方面进行衡量Ø衡量商务智能提供的信息的可操作性；Ø衡量实施商务智能可能给企业带来的回报，Ø衡量实施商务智能可以帮助企业实现短期目标

实现的难易程度Ø商务智能的实现需要涉及的范围Ø衡量数据的可获取性

**需求分析：详细的需求分析、数据定义、技术选择**

分析阶段，针对在规划阶段最终选择要实现商务智能的业务部门或业务领域，进行详细的需求分析- 收集需要的各类数据，- 选择需要的商务智能支撑技术，如数据仓库、在线分析数据或者数据挖掘等

**设计：数仓库、数据集市、OLAP设计、数据挖掘设计**

如果要创建数据仓库，则进行数据仓库的模型设计，常用的是多维数据模型。数据集市

可以从数据仓库中抽取数据进行构建§ 在不构建数据仓库的情况下，也可以直接为

某个业务部门设计和实现数据集市。§ 如果要实现OLAP解决问题，则要设计多维

分析的聚集操作类型。§ 如果要借助数据挖掘技术，则需要选择具体的算法

**实现：构建数据仓库、构建数据集市；增强查询报表、在线分析处理、数据挖掘算法、企业绩效管理**

实现阶段，选择ETL工具实现源数据的抽取，构建数据仓库和（或）数据集市

§ 对数据仓库或数据集市的数据，选取并应用相应的查询或分析工具，包括增强型的查询、报表工具、在线分析处理工具、数据挖掘系统以及企业绩效管理工具等

§ 在具体应用该系统之前，需要完成对系统的数据加载和应用测试，设计系统的访问控制和安全管理方法

**影子价格在管理决策中的作用:**

影子价格反映了供求关系和资源的稀缺性：影子价格为零，说明供大于求，影子价格越高则资源越稀缺

影子价格告诉企业的经营管理者，增加哪种资源对提高效益有利：

若: 影子价格**＞**市场价格,则应 买进该资源

影子价格**＜**市场价格,则应 卖出该资源

影子价格告诉决策者新的产品是否可以投入生产．







