

## 摘要

本文提供了 pdf 版，以供打印：[商务智能-07-期末总结.pdf](#)。

# 1. 选择题

## 1.1 主题分布

1. 商务智能 overview、知识链、知识指令
2. 数据仓库
3. 数据仓库的设计
4. OLAP
5. MOLAP & ROLAP
6. 多维建模
7. 多维建模
8. 数据挖掘

## 1.2 选项判断

表述	判断
商务智能试图以信息化技术自动完成数据到信息、 <b>信息到知识</b> 的提取过程。	×
从信息技术方面来看，商务智能涵盖了数据仓库、多维建模、ETL、OLAP、仪表盘、报表查询、数据统计、数据挖掘等多种相关技术。	√
从商业应用方面来看，商务智能不仅支持最新的 IT 技术，同时也应该提供打包的商务解决方案。	√
从层次结构方面来看，可以将商务智能体系结构划分为：数据源层、数据获取层、数据存储和管理层及信息访问/展现层。	√
鉴于商务智能的重要作用，企业在进行信息化时，应以企业数据模型为蓝本，同步事务处理系统和分析处理系统。	√
由于规范化引入了查询时的额外开销，在维度表规模不大的前提下，一般不考虑采用雪花模型。然而，在应对大型快变维度时，仍然可以使用采用雪花模型/微型维度的方式节约存储开销。	√
非事实型事实表在多个维度间建立连接关系。但是由于没有独立的度量值，非事实型事实表无法独立使用，必须依附于其他事实表参与分析应用。	√
在采用累积快照进行多维建模时，事实表中记录了单个生命周期中多个关键环节所产生的信息，并使用多个日期类型的维表对这些关键环节进行标记。	√
为方便与操作型数据环境的对接，简化 ETL 的处理过程，用以连接维度表和事实表的关键字可以 <b>直接来源于</b> 操作型数据环境的关键字。	×
退化维度一般用以对事实表进行分组，并偶尔用以连接操作型数据环境。	√
数据仓库是数据库技术进一步发展的必然阶段。由于数据仓库数据容量大，数据模型先进且允许存在冗余，数据仓库正日益替代数据库，成为主流的 <b>数据存储技术</b> 。	×
原子层拥有数据仓库最低粒度的数据，因此，在数据通过 ETL 进入原子层时，应与数据源保持 <b>相同粒度</b> 。	×

表述	判断
由于数据仓库的“不可更新”特性，数据仓库中的数据实际上是“滞后”的，所以数据仓库必须定期/不定期的采用刷新方式，将数据库等数据源中最新的数据变化反应到数据仓库中来。	√
数据仓库是一种反映主题的全局性数据组织，在执行周期性分析应用或局部分析应用时，往往效率不高。在这种情况下，可以按部门或个人分别建立反映各个子主题的局部性数据组织，称作数据集市。	√
数据仓库内部以“快照”的数据结构为中心来组织。快照通常包括关键字、时间、非关键字的主要数据和二级数据四个部分，其中非关键字的主要数据是数据仓库用以存放信息的主要部分。	√
一般来说，ROLAP 查询效率 <b>优于</b> MOLAP，但装载性能 <b>劣于</b> MOLAP。（ROLAP 装载性能好，查询效率低）	×
由于采用了多维数据库（MDDDB），MOLAP 可以比 ROLAP 支持 <b>更多的</b> 维度。（MOLAP 有维度上限，ROLAP 没有维度上限）	×
在 ROLAP 中，需要采用多维综合引擎，在多维查询和结果及 SQL 查询和结果之间进行转化。	√
在 OLAP 系统中，可以融合 MOLAP 和 ROLAP 两种技术。采用关系型数据库存储细节数据，使用多维数据库来存放高层次数据或关系型数据库的查询结果。	√
在需要对多维模型进行演化时，ROLAP 相对于 MOLAP 更加灵活便利。	√
数据挖掘属于对数据的归纳。	√
对数据挖掘的结果的评价分为主观评价和客观评价。一般来说，最常见的客观评价指标是“支持度（兴趣度）” / “置信度”。	√
与分类方法不同，聚类方法不需要给定训练数据集和测试数据集，而是使用数据之间的相似程度/相异程度进行类别划分。因此，聚类方法是一种无指导的学习。	√
采用相同的数据挖掘过程、数据挖掘方法、数据挖掘算法及相同的阈值和参数，对相同数据集进行多次数据挖掘，所得到的数据挖掘结果也 <b>应该是相同的</b> 。	×
将 B 是 A 的子女记做 $A \rightarrow B$ ，如存在 $N1 \rightarrow N2, N2 \rightarrow N3, \dots, N_{k-1} \rightarrow N_k$ ，则称 $N1$ 是 $N_k$ 的祖先。如数据库/数据仓库中存放着所有市民之间的子女关系，那么通过 <b>数据挖掘</b> 能够获取所有市民之间的祖先关系。（不需要数据挖掘，常规算法可以解决）	×
商务智能的目标包括：为商务活动提供 <b>自动化解决方案</b> ；实现 <b>商务领域的人工智能</b> ；提供商务领域的专家系统；进行决策支持；改善信息访问方式。	×
依据多级体系划分属性，可以将维度表规范化，以雪花模型替代星型模型，从而节约存储空间。然而，由于规范化引入了查询时的额外开销，在维度表规模不大的前提下，一般不考虑采用雪花模型。	√
非事实型事实表在多个维度间建立连接关系。由于没有独立的度量值，非事实型事实表无法独立使用，必须依附于其他事实表参与分析应用。	√
一般来说，对事实表进行建模时，事务模型粒度最小，周期快照粒度最大，而累积快照粒度 <b>介于两者之间</b> 。（应为：事务模型 < 周期快照 < 累积快照）	×
为保持历史一致性，维度表的元组属性发生变化时，需要用额外的行或者列来记录历史信息。因此，在对大型快变维度进行处理时， <b>必须</b> 采用微型维度、预设波段等方法，将快变维度转化为渐变维度。	×
为方便与操作型数据环境的对接，并简化 ETL 的处理过程，用以连接维度表和事实表的关键字应当 <b>直接来源于</b> 操作型数据环境。	×

表述	判断
一般来说，MOLAP 查询效率优于 ROLAP，但装载性能 ROLAP 优于 MOLAP。	√
由于采用了多维数据库 (MDDDB)，MOLAP 可以比 ROLAP 支持更多的维度。(MOLAP 有维度上限，ROLAP 没有维度上限)	×
在需要对多维模型的推度数量等进行调整的情况下，ROLAP 相对于 MOLAP 更加灵活便利。	√
在 ROLAP 技术中，需要采用多维综合引擎，在多维查询和结果及 SQL 查询和结果之间进行转化。	√
OLAP 的基本操作包括：切片/切块、旋转、上钻 (Roll up) /下钻 (Drill down)。其它操作包括跨钻 (Drill across) /钻透 (Drill through)。人们可以通过这些操作浏览数据立方体，从而获得对分析对象的全面理解。	√
由于数据仓库的“不可更新”特性，数据仓库中的数据实际上是“滞后”的。这类问题可以通过使用操作数据存储 (ODS) 加以改善。	√
鉴于大多数的数据仓库系统仍然使用关系型数据库加以实现，在线系统需要访问数据仓库中的信息时，依然可以比照数据库的访问方式，采用 SQL 等进行 <b>直接查询</b> 。(数据仓库不能够直接查询，数据库才可以)	×
即使使用关系型数据库作为数据仓库的后台存储技术，由于目标、性能和使用对象等的不同，数据仓库仍然拥有自身所特有的数据建模思想、索引及优化方法。	√
数据挖掘属于对数据的 <b>演绎</b> 。(数据挖掘是对数据的归纳)	×
对数据挖掘的结果的评价分为主观评价和客观评价。一般来说，最常见的客观评价指标是“支持度(兴趣度)” / “置信度”。	√
与分类方法不同，聚类方法不需要给定训练数据集和测试数据集，而是使用数据之间的相似程度/相异程度进行类别划分。因此，聚类方法是一种无指导的学习。	√
特征规则挖掘可以被用来作为其他数据挖掘方法的前继步骤，通过在“基本关系表”和“概括关系表”上应用其他数据挖掘方法，可以得到不同概念层次上的数据挖掘结果。	√
采用相同的数据挖掘过程、数据挖掘方法、数据挖掘算法及相同的阈值和参数，对相同数据集进行多次数据挖掘，所得到的数据挖掘结果也 <b>应该是相同的</b> 。	×

## 2. 简答题

### 2.1 主题分布

与选择题基本相同，不包括数据挖掘。集中在 OLAP、数据仓库和建模。

需要掌握的概念：数据仓库

举例说明。

### 2.2 例题

#### 2.2.1 什么是数据仓库？其四大特色？

W.H.Inmon 在《建立数据仓库》一书中，对数据仓库的定义为：**数据仓库**就是一个面向主题的、集成的、非易失的（稳定的）、时变的（随时间不断变化的）**数据集**，用于支持经营管理过程中的决策制定。

Tim.Shelter (Informix 公司负责研究与开发的副总裁)：**数据仓库**将分布在企业网络中不同信息岛上的商业数据集成到一起，存贮在一个单一的集成关系型数据库中。利用这种集成信息，可方便用户对信息的访问，更可使决策人员对一段时间内的历史数据进行分析，研究事物发展走势。

四大特色：

1. 面向主题
2. 集成
3. 非易失 (稳定的)
4. 时变的 (随时间不断变化)

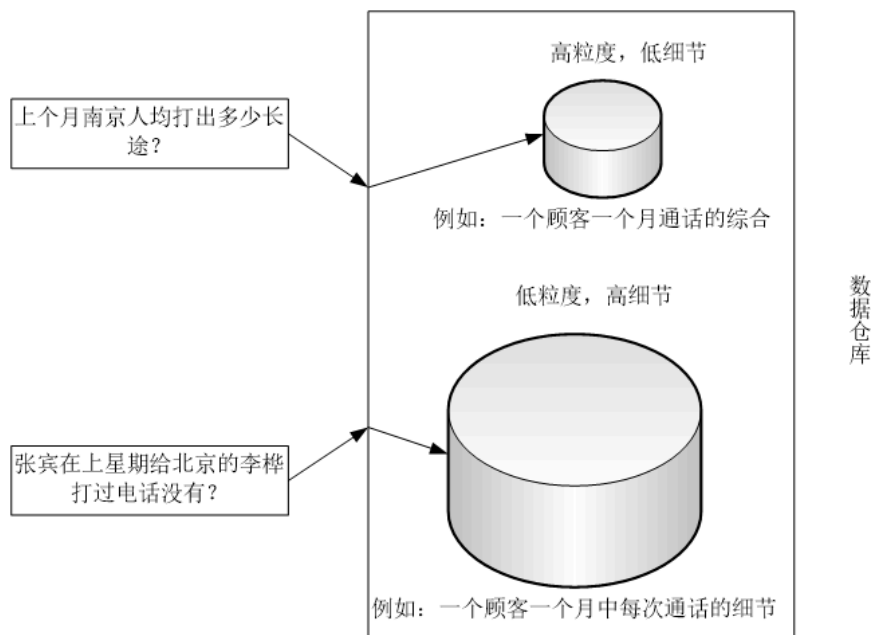
## 2.2.2 为什么在传统的以数据库为核心的事务处理环境中不适宜建立 DSS 等分析型应用？

1. 性能特性不同
2. 数据集成问题 (主要原因)
3. 数据动态集成问题
4. 历史数据问题
5. 数据综合性问题
6. 数据访问问题

## 2.2.3 什么是数据仓库中的粒度？为何要在数据仓库中采用多重粒度？试举例说明

粒度：对数据仓库中的数据的综合程度的一个度量。

举例：



## 2.2.4 数据仓库的物理模型设计优化技术有哪些？对这些技术进行简要的说明

1. 合并表
2. 建立数据序列
3. 引入冗余
4. 表的物理隔离
5. 生成导出数据
6. 建立广义索引: 与“最”有关的统计结果

## 2.2.5 什么是数据仓库中的历史完整性/一致性？为保持历史完整性/一致性，采用哪些方式？试举例说明

历史完整性/一致性：指数据可回溯，历史数据不丢失。

### 针对渐变维度

1. 改写属性值（无法保证历史一致性）
2. 添加维度行
3. 添加维度列（不适合大量变化）
4. 1+2+3

### 针对快变维度

1. 微型维度
2. 预设波段

## 2.2.6 简述数据仓库刷新的方法，并对每一种方法进行简单的说明

1. 时间戳
2. DELTA 文件：记录该应用所作的修改操作
3. 建立映像文件：快照
4. 日志文件

## 2.2.7 数据仓库中 ETL 技术是什么？在数据仓库架构中 ETL 完成什么任务？

ELT：抽取、转换、装载

任务：

1. 数据抽取
2. 数据转换
3. 数据刷新、装载

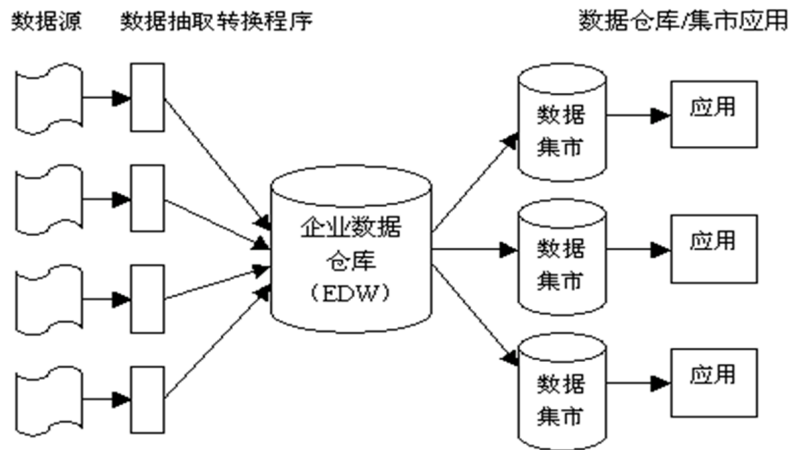
## 2.2.8 为什么在数据仓库体系中还需要建立数据集市？在企业中建立数据仓库和数据集市体系的方法主要有哪四种？请分别描述这些方法，并总结其优点和缺点。

### 2.2.8.1 原因

数据仓库太大，理解困难，效率低

### 2.2.8.2 方法

自顶向下的结构



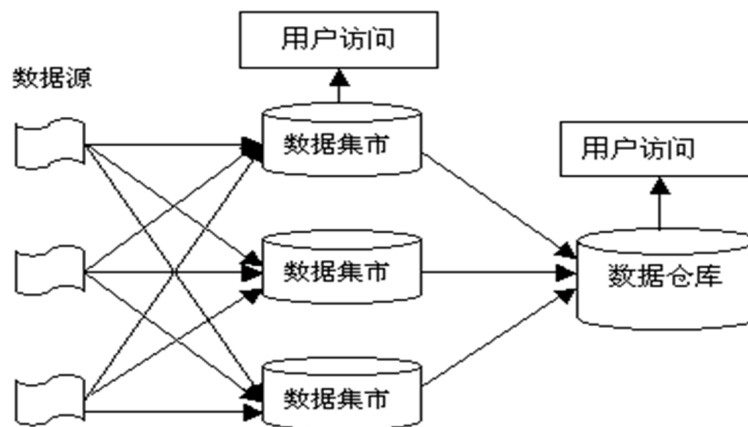
优点:

1. 建立数据集市能够减轻 DW 访问负载
2. 各部门可以任意处理数据
3. 数据转换和整合在 DW 阶段统一完成
4. 具备数据缓冲功能: 数据仓库有数据缓冲的作用

缺点:

1. 成本高
2. 见效慢: 短期之内看不到效果
3. 数据集市间不共享资源: 可能导致不一致性等困难, 数据集市具有一定的局部性。

#### 自底向上的结构

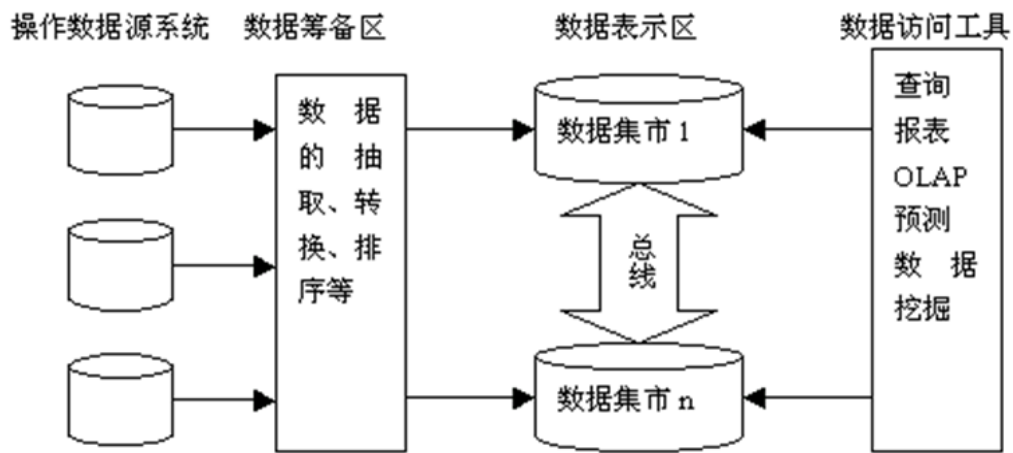


优点: 见效快、启动资金少

缺点:

1. 各个部门都要进行数据清理整合
2. 可能造成“蜘蛛网”、数据不一致等问题
3. 总体上没有节约资金

## 总线结构



优点：共享维表和事实表，解决了建立数据集市的许多问题

缺点：

1. 这种结构基于多维模型，应用限制于 OLAP
2. 多个数据源直接影响多个集市，造成数据仓库结构不十分稳定

## 企业级数据集市

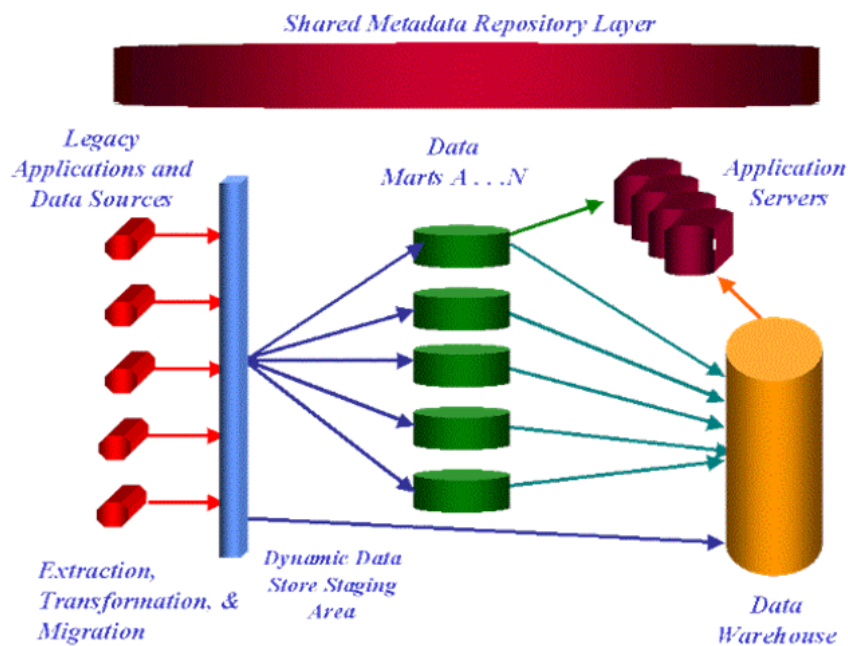


Figure Three -- EDM Architecture

优点：

1. 汇集企业信息：企业级数据仓库可以将企业内部各种信息集中在一起，方便各部门之间进行协作和共享信息，从而在决策和管理上节省时间成本。
2. 跨部门的数据访问：企业级数据仓库能够建立全面的数据存储体系，包括金融、市场、客户服务、销售分部的信息。对所有人来说，他们只需要进行一步的信息获取即可浏览原始的多部分数据内容。
3. 罗列实时数据：企业级数据仓库能够将动态上交的实时数据录入其中，这将使各部分可以根据最新信息了解市场情况或者金融市场波动情况，减少风险。
4. 协助决策：企业级数据集市把代表事实和关系采集到一张图中，提高决策效率。



### 2.2.9 数据仓库应用中数据挖掘在什么地位？和其他方式有什么差异？

### 2.2.10 各种数据挖掘方法用来解决什么问题？

### 2.2.11 星型模型和雪花模型各自有什么优缺点？

星形模型	雪花模型
非规范化	规范化
简单的表关系	复杂的表关系
记录之间存在数据冗余	<b>节省存储空间</b>
连接简单，低开销	连接的复杂，高开销
高维度浏览能力	低维度浏览能力
支持物理加速技术	不支持物理加速技术

## 3. 综合题 (16 + 16 = 32 分)

### 3.1 考试范围

1. OLAP 和数据立方体
2. 维、层
3. OLAP 操作
4. 构建数据模型
5. 模型演化

### 3.2 例题

#### 3.2.1 请给出一个具体的数据立方体模型的例子（绘图说明）

#### 3.2.2 以该数据立方体为例介绍 OLAP 中的维和层的概念

- 维是观察度量值的角度。
- 层反映了对度量值的观察深度。

#### 3.2.3 以该数据立方体中的某一维，举例说明什么是切片 (slice)、切块 (dice)、数据概括 (roll-up) 和数据细化 (drill-down) 操作

1. 切片 (Slice)：根据某一维上的某个维成员值选择统计数据进行分析
2. 切块 (Dice)：根据若干维上的维成员取值的区间选择统计数据进行分析
3. 旋转 (Pivot/Rotate)：调整维的排列次序的动作称为旋转
4. 上钻/数据概括 (roll-up)：将多维下标的取值提升到较高的概念层次上，从而形成新的统计查询结果，并进行分析。
5. 下钻/数据细化 (drill-down)：将多维下标的取值降低到较低的概念层次上，从而形成更细致的统计查询结果，并进行分析。