第五章作业

批注 [ww1]: 文件命名：学号-姓名-第？章作业

1.什么是 OLAP？它的特性有哪些?

答：（1）OLAP 的定义

OLAP（Online Analytical Processing，联机分析处理）是面向决策支持的联机数据访问与分析技术，核心是 “多维分析”，允许用户从多个角度交互性获取数据洞察，具体包含两种权威定义：

定义 1：针对特定问题的联机数据访问和分析，通过对维数据的多种观察形式快速、稳定一致地获取，支持管理决策人员深入观察数据；

定义 2（OLAP 委员会）：使分析人员、管理人员能从多种角度，快速、一致、交互地获取从原始数据转化而来的、反映企业维特性的信息，以深入理解数据的软件技术。

（2）OLAP 的特性

快速性：对用户大部分分析需求，系统需在 5 秒内响应，确保决策分析的高效性；

可分析性：能处理与应用相关的任意逻辑分析和统计分析，覆盖业务决策所需的各类计算场景；

共享性：支持多用户共享同一 OLAP 数据，满足团队协作分析需求；

多维性：OLAP 的关键属性，需提供数据的多维视图（如时间、地区、产品维），支持层次维与多重层次维的完整分析；

信息性：无论数据量大小、存储位置，均能及时提供指导性信息，支持大容量数据的高效管理与分析。

2.试述 ROLAP，MOLAP 和 HOLAP 体系结构的各自使用场景并解释如此选择的原因。

答：（1）ROLAP（关系型 OLAP）

使用场景：细节数据查询频繁、需要灵活定制分析逻辑的场景，如企业日常业务数据分析（需频繁钻取到具体订单、用户数据）、动态生成个性化报表的场景；

选择原因：

ROLAP 以关系数据库为核心，将原始数据、维度数据、聚合数据均存储在关系表中，通过 SQL（带 sum/group by）实现查询。其优势在于细节数据查询速度快（直接读取关系库中的明细数据）、分析逻辑灵活（支持自定义 SQL 查询），适合需要频繁访问细节数据、分析需求多变的业务场景。

（2）MOLAP（多维 OLAP）

使用场景：聚合数据查询为主、对分析性能要求高的场景，如企业固定报表生成（如月度销售汇总报表）、高管决策层的快速汇总分析（需快速获取地区、产品的总销售额）；

选择原因：

MOLAP 基于多维数组存储数据，物理上形成 “立方体” 结构，预存储聚合数据（如各维度组合的汇总值）。其优势在于聚合数据查询速度快（直接读取预计算的聚合结果，无需实时计算）、存储容量小（仅存储必要的聚合数据），适合以汇总分析为主、查询逻辑相对固定的场景。

（3）HOLAP（混合 OLAP）

使用场景：既需要频繁查询细节数据，又需高效获取聚合数据的综合分析场景，如企业级综合决策分析（既要看全国销售汇总，又要钻取到某省某产品的明细订单）；

选择原因：

HOLAP 结合 ROLAP 与 MOLAP 的优势 —— 原始数据与明细数据存储在关系数据库（保留 ROLAP 细节查询能力），聚合数据以多维结构存储（保留 MOLAP 聚合查询性能）。其优势在于平衡细节查询灵活性与聚合查询效率，避免单一结构的短板，适合对数据查询的

“灵活性” 与 “性能” 均有要求的场景。

3.OLAP 的主要分析方法有哪些？试举例说明。

答：（1）切片（Slice）

定义：在多维数组的某一维上选定一个具体值，使多维数组从 n 维降为 n-1 维，聚焦单一维度值的数据；

举例：在 “时间（年 / 季度 / 月）、地区（省 / 市）、产品（类别 / 名称）、销售额” 的四维数组中，选定 “时间 = 2025 年第 1 季度”，得到 “地区 - 产品 - 销售额” 的三维数据，即 “2025 年 Q1 各地区各产品销售额”。

（2）切块（Dice）

定义：在多维数组的某一维或多维上选定一个区间（多个值），获取该区间内的多维数据子集；

举例：在上述四维数组中，选定 “时间 = 2025 年 Q1-Q2、贷款状态 = 正常 / 次级、经济性质 = 集体 / 个人”，得到该区间内的 “地区 - 产品 - 销售额” 数据，即 “2025 年 Q1-Q2 正常 / 次级贷款、集体 / 个人客户的各地区各产品销售额”。

（3）钻取（Drill）

定义：基于维度的层次结构（如时间：年→季度→月→日），在 “汇总数据” 与 “细节数据” 间切换，包括向下钻取（汇总→细节）和向上钻取（细节→汇总）；

举例：

向下钻取：从 “2025 年 Q1 总销售额” 钻取到 “2025 年 1 月、2 月、3 月的销售额”，查看季度内各月的明细数据；

向上钻取：从 “2025 年 1-6 月销售额” 汇总为 “2025 年上半年总销售额”，查看更高层次的汇总数据。

（4）探取（Drill-up/Roll-up）

定义：通过 “提升维度层次” 或 “消除维度” 实现数据概括，获取更宏观的分析结果；

举例：

提升层次：将 “2025 年 Q1-Q4 销售额” 概括为 “2025 年上半年、下半年销售额”；

消除维度：在 “时间 - 地区 - 产品 - 销售额” 数组中，消除 “产品” 维，得到 “时间 - 地区 - 销售额” 数据，即 “各时间各地区总销售额”。

（5）旋转（Pivot）

定义：改变多维数据中维度的位置关系（如交换行与列、调整维度层次），获取不同视角的数据视图；

举例：原数据视图为 “时间（行）- 产品（列）- 销售额”，旋转后变为 “产品（行）- 时间（列）- 销售额”，从 “按时间看产品销售” 切换为 “按产品看时间销售”，便于从不同维度主导分析。

（6）交叉探查（Drill Across）

定义：在星座模型（多个事实表关联同一维度表）中，基于公共维度连接多个事实表，对比分析不同事实指标；

举例：销售事实表（含 “日期 - 产品 - 销售额”）与库存事实表（含 “日期 - 产品 - 库存量”）均关联 “日期维度表”，通过 “日期” 维度连接两表，分析 “各日期各产品的销售额与库存量对比”，查看销售与库存的匹配关系。

4.OLAP 评价准则有哪些？试举例说明。

答：OLAP 的核心评价准则为 Codd 提出的 “OLAP 12 条准则”，涵盖多维分析的关键需求，具体如下（含举例）：

准则名称	核心含义	举例
1. 多维概念视图	提供数据的多维视角，支持用户从任意维度组合分析	为 “销售数据” 提供 “时间、地区、产品、客户” 的多维视图，用户可自由组合维度查询
2. 透明性	用户无需了解数据存储细节（如存在关系库还是多维库），仅关注分析需求	用户查询 “销售额” 时，无需知道数据存储在 MySQL 还是多维立方体中
3. 访问能力	确保数据易于访问，支持通过标准接口（如 SQL）获取数据	支持通过 SQL 查询 OLAP 数据，无需学习专用语法
4. 一致的报表性能	无论数据量大小、分析复杂度，查询性能保持稳定（如响应时间差异小）	查询 “1 年销售额” 与 “5 年销售额” 的响应时间均在 5 秒内
5. 客户 / 服务器架构	支持分布式部署，分为数据层（数仓）、分析层（OLAP 服务器）、展现层（前端）	数仓存储原始数据，OLAP 服务器处理分析请求，前端工具展示报表
6. 维的等同性	所有维度被平等对待，无主次之分，支持任意维度的组合分析	“时间维” 与 “地区维” 可同等参与筛选、分组，无需指定 “主维度”
7. 动态稀疏矩阵处理	高效处理稀疏数据（如某些 “地区 - 产品” 组合无销售数据），避免存储与计算浪费	对 “西北 - 奢侈品” 这类无销售数据的组合，仅标记为空，不重复存储无效数据

准则名称	核心含义	举例
8. 多用户支持	支持多个用户并发访问 OLAP 数据，且不影响性能与数据一致性	10 个业务分析师同时查询 OLAP 数据，响应时间仍保持稳定
9. 非受限的跨维操作	允许任意维度组合分析，无维度数量或组合方式的限制	可同时选择 “时间、地区、产品、客户、渠道”5 个维度进行交叉分析
10. 直观的数据操纵	操作简单直观，支持拖拽、点击等可视化操作实现分析需求	通过拖拽 “时间维” 到行轴、“产品维” 到列轴，快速生成分析视图
11. 灵活的报表生成	支持多种报表格式（如表格、图表），且可自定义报表样式	生成 “销售额明细报表”（表格）与 “销售额趋势报表”（折线图），并调整颜色、字体
12. 不受限的维度与聚合级别	支持任意维度数量（如 2 维到 10 维）与聚合级别（如日→月→年），无固定限制	可将 “时间维” 从 “日” 聚合到 “年”，也可新增 “渠道维” 扩展为 6 维分析

5.请列举常用的 OLAP 前端展现方式。

答：OLAP 前端展现以 “可视化 + 交互性” 为核心，常用方式包括：

多维报表：以表格形式展示多维数据，支持行 / 列维度切换（如 “时间 - 地区 - 销售额” 交叉表）；

图表类：饼图（展示各产品销售额占比）、柱状图（对比各地区销售额）、折线图（展示销售额时间趋势）、直方图（分析销售额分布）、三维图形（直观呈现多维数据关系）；

地理可视化：结合地图展示区域相关数据（如各省份销售额在地图上标注，用颜色深浅表示数值大小）；

仪表盘（Dashboard）：整合多个关键指标（KPIs）、图表，汇总展示业务状态（如电商“实时销售仪表盘”含总销售额、订单量、客单价等）；

评分卡（Scorecard）：通过定量指标评估对象表现（如金融领域的“客户信用评分卡”、电商的“商家信誉评分卡”）；

交互操作可视化：将切片、切块、旋转等分析操作可视化呈现（如通过下拉框选择“时间 = Q1”实现切片，通过拖拽维度实现旋转）。

6.数仓为什么要进行分层？典型的分层包括哪几层？各层的作用及任务又是什么？

答：**（1）数仓分层的原因**

数仓存储的原始数据杂乱、分散（如来自 CRM、ERP 等不同业务系统），分层的核心目的是：

- **数据整合与净化**：将原始数据逐步加工为干净、一致的信息，避免“脏数据”影响分析结果；
- **复用与效率**：中间层（如明细层、汇总层）数据可复用，避免重复计算，提升分析效率；
- **可维护性**：分层使数据加工链路清晰，便于问题追溯（如某指标异常时，可逐层排查加工过程）；
- **价值转化**：实现从“原始数据资产”到“可用信息资产”的转化，支撑 OLAP 分析与业务决策。

（2）典型分层及各层作用、任务

数仓典型分层为 **ODS→DWD→DWS→ADS**，各层细节如下：

分层名称	英文全称	核心作用	主要任务
1. 操作型数据存储层	Operational Data Store	存储原始业务数据，作为数仓的“数据入口”，保留数据原貌	1. 从业务系统（如 MySQL、Oracle）抽取原始数据；2. 仅做简单清洗（如去除明显错误格式），不改变数据结构与内容
2. 数据明细层	Data Warehouse Detail	整合、净化原始数据，形成结构化、一致性的明细数据，为后续分析提供基础	1. 数据清洗（处理缺失值、异常值）；2. 数据整合（关联多系统数据，如将“订单表”与“用户表”关联）；3. 数据标准化（统一字段格式，如日期格式为“YYYY-MM-DD”）
3.	Data Warehouse	对明细层数据	1. 按常用维度聚合（如按“时间

分层名称	英文全称	核心作用	主要任务
数据轻度汇总层	Summary	进行轻度聚合，减少后续分析的计算量，提升查询效率	- 地区”汇总销售额）；2. 保留一定细节度（非完全汇总），支持适度钻取；3. 存储中间计算结果（如“月度用户活跃度”）
4. 应用数据存储层	Application Data Store	为前端应用（如 OLAP 报表、BI 工具）提供直接可用的数据，是数仓的“数据出口”	1. 按前端需求加工数据（如 OLAP 分析所需的多维数据结构、报表所需的格式）；2. 生成最终指标（如“年度总销售额”“客户留存率”）；3. 适配前端工具（如 Tableau、Power BI 的数据格式）

这是最后一页，用于贴上手绘（可以用触笔在平板上手绘）的导图。图中内容包括知识点和技能。技能最好能给出具体的小例（如果限于纸张面积没法写到纸面，但至少自己要知道例子）。