2023 年秋季学期
《数据库系统》
开箱手册

前言

本文根据邹兆年老师 2023 年春季学期对数据科学与大数据技术专业同学的授课内容总结，如果你之前对于本门课程有一定了解的话，就应该知道这门课程不同老师讲授的内容有所不同，所以请大家选择性参考。由于本门课程不同老师不同学期之间实验的内容都会有变化，所以本文不会涉及实验的相关内容，主要面向理论内容及考试。全文共约四千四百字，希望能对你的数据库学习有所帮助。另外，由于作者本人水平有限，难免会在文中出现一些错误，在这里提前致以歉意。

一、整体内容概述：

整体上这门课的内容可以分为四个大的方面，基本概念、模型与语言、数据库设计、数据库内核。在考试中这四部分占比情况大概是：前两部分：数据库设计：数据库内核=3:3:4。数据库内核是本门课程的重点、难点，也是最终决定期末考试能否上 90 分甚至 95 分的关键点。下面第二部分将会详细地介绍一下各部分知识点：

二、各部分知识点介绍：

首先声明，下文中提及的小题指的是选择、填空这样的题型，大题指的是简答题、分析题、论述题、查询题、计算题、证明题、设计题这样的题型。题型不是固定的，选择填空可能有可能没有，大题中各个题型不一定都有，这里仅是列出可能的情况供大家参考。下面的划分只是按照知识之间的逻辑联系划分的，并不一定是实际讲课的顺序。

（一）基本概念（和（二）合占约 30%）

数据模型；

数据库三层模式结构；

数据库、数据库管理系统与数据库系统；

数据库语言、用户；

数据库模式、实例；

数据独立性等。

这一部分主要讲述的就是基本概念和一些预备知识，相对来说比较重要的是数据库三层模式结构和数据独立性。我们需要准确掌握以

上基本概念。在考试中该部分主要以小题考察，出大题的可能性很低，如果出大题的话最有可能是简答题叙述某个概念或理论。

（二）模型与语言（和（一）合占约 30%）

1、基本概念：关系数据模型的若干基本概念——三要素、键、候选键、主键、外键、关系数据库的完整性约束等。

2、关系代数：关系代数表达式的书写。

3、关系演算：元组关系演算、域关系演算。

4、SQL：SQL 语句的书写，包括数据定义、数据更新、各种查询类型、以及视图的相关操作等。

这一部分的 2 和 4 必考大题，尤其是 4 中的各种 SQL 查询语句是重中之重，难点一般在于嵌套查询。3 如果考察的话也是以大题的形式，难度相对大一点，可能需要一些数理逻辑的知识，但是考察可能性很低。1 主要以小题形式考察。

（三）数据库设计（占约 30%）

1、概念数据库设计：需要掌握 ER 图的相关概念并准确设计 ER 图。

2、逻辑数据库设计：把 ER 图转化为关系模式、各种函数依赖的概念、判断函数依赖集的等价性、属性集的闭包计算、各种级别范式的定义、判读范式级别、关系模式的分解、无损连接和函数依赖保持性的判断以及相关算法、计算最小覆盖、求候选键等。

3、物理数据库设计：（**备注：**这一章在我们上课的学期主要体现在实验中，不是期末考试的考点，所以这个考点的分析以及出题的情况的总结主要结合了课件和往年经验，有很大不确定性）：不同类型索引

的比较、索引的设计、关系模式的优化（规范化关系模式的问题）、关系的划分（垂直、水平）。

对于这一部分而言，1 和 2 必考大题。一般都会出两道大题，一道大题是给定一个具体的应用场景，第一问设计 ER 图，第二问根据第一问设计的 ER 图转化成关系模式；另一道大题就是逻辑数据库设计的相关题目，会有很多个问，上面 2 中提到的考点都有可能出，有可能涉及到证明题。对于 3 而言出大题的概率不高，以选择填空为主，但是要理解其中不同索引之间的原理以及特点。

（四）数据库内核（占约 40%）

这一部分的 1、2、3、4、5 都是基本必考大题，而且考的很灵活，选择题一般也有可能涉及。具体来说如下：

1、存储管理：

（1）面向磁盘的数据库存储结构（元组存储、页布局、文件组织）；

（2）缓冲区管理（缓冲池的结构、页的请求、释放、修改、替换）。

备注：对于（2）的缓冲区管理，根据最新科研成果（Viktor Leis, Adnan Alhomssi, Tobias Ziegler, Yannick Loeck, and Christian Dietrich. 2023. Virtual-Memory Assisted Buffer Management. Proc. ACM Manag. Data 1, 1, Article 7 (May 2023), 25 pages. https://doi.org/10.1145/3588687），可以用虚拟内存来实现 DBMS 的缓冲池，这里的知识可能有较大更新，请同学们注意。

（1）（2）都有可能出大题，可能会涉及到计算或者结合具体实例的简答等题型。

2、索引结构：

（1）各种类型索引的概念、原理；

（2）可扩展哈希表与线性哈希表（结构、查找、插入、删除）；

（3）B+树（结构、查找、插入、删除）。

（2）（3）基本必出大题，有可能挑一个出也有可能都出，基本考法就是画出插入删除数据的流程，（1）单独出大题可能性不大，有可能出小题也有可能是其他题目中涉及到的一个很小的知识点。

3、查询执行与优化

（1）查询执行时的各个算法（算法思想、IO 代价、对内存的要求）：外存排序、选择、投影去重操作、集合操作、连接操作等；

（2）查询执行方法：物化、流水线；

（3）逻辑查询优化：查询表达式树、关系表达式等价变换、选择下推和投影下推、代价模型、基数估计（选择、投影、去重、连接、集合操作的结果元组数估计方法以及使用直方图的估计方法）；

（4）物理查询优化：物理执行算法的选择以及传递中间结果的方式。

3 是基本必出一道综合性的大题，其中的（1）（2）（3）（4）基本都会涉及到，这道题通常也是我个人认为卷面中最复杂、最灵活、难度最大的题目。

4、并发控制

（1）事务：事务的概念与 ACID 性质；

（2）调度：各种调度的定义以及判断、隔离级别；

（3）基于锁的并发控制：锁的作用、锁的类型、锁管理器（整个锁的

使用过程）、锁协议（2PL 协议、SS2PL 协议）、死锁、死锁的检测与预防。

4 基本必出大题，尤其是（2）和（3）是重点，关于调度和锁的各个考点都有可能涉及到，这里还有可能会涉及到证明（证明某个调度是冲突可串行化调度等），（1）有可能出小题，也有可能涉及到简答题。

5、故障恢复

（1）故障类型、故障恢复的类型、缓冲池策略；

（2）WAL：日志记录、WAL 协议（三类）、根据日志对事务的分类；

（3）根据三类 WAL 协议进行故障恢复的原理、方法；

（4）检查点：涉及检查点的恢复原理、方法。

5 也是基本必考大题的，基本上是考察一个故障恢复的流程，补全日志记录等，各种类型的恢复都有可能涉及，一定要明白各种恢复的原理，其中一些定义和概念也有可能再出小题。

三、一些个人学习经验

根据个人学习过程中的体会总结了一点经验，未必适合所有人，请大家选择性参考。整体上来说，听课还是最重要的，考试范围跑不出老师所讲述的内容。教材是《数据库系统概念》大黑书，老师 PPT 上的内容已经足够丰富，如果没有时间也可以不看教材，吃透 PPT 即可，学有余力可以再参考教材。其次，实验要认真去完成，对我们学的理论知识是一个很好的应用过程，也会加深对理论知识的理解。下面简要介绍一下学习各个部分时候的注意事项：

第一部分基本概念是对整门课程的一个综述，可能在初学时会有所困惑，但问题不大，可以带着问题向后学，随着后面学习的深入会对其有越来越深入的理解。

第二部分关系代数和 SQL 一定要跟住进度，不要到期末再补，因为关系代数在内核部分还要用，而 SQL 在实验中要用到。最好的方法就是在电脑上多练习，SQL 练习在自己的 DBMS 中进行即可，关系代数的练习邹老师推荐了一个网站，我会放在后面。

第三部分数据库设计，相对来说算是比较简单的一部分。这里面的 ER 图的设计题很容易得到大部分的分数，但是得满分不容易，有一些细节上的选择不好确定，平时练习还是要多画一画图，不要纸上谈兵，避免考试的时候眼高手低。逻辑数据库设计的题目相对比较简单，套路性比较强，掌握了方法基本都能做对。

最后一部分数据库内核属于难度最大的一部分，这块在课后还是要下一番功夫才能掌握好的。存储管理和索引结构相比后面内容要简单一些，这一块内容需要把基本的原理、概念、方法掌握的比较熟练，尤其是 B+树和哈希索引结构这种必考题，一定要非常熟练和准确，以免浪费时间。查询执行、优化难度最大，在上课时一定要仔细听老师讲算法的 IO 代价、内存要求、以及物理优化中的不同选择是怎么分析和计算的。这里死记硬背没有用，要结合具体的情景综合各个因素分析。最后的并发控制和故障恢复的知识含量较多，我个人建议在学习时自己梳理一下各个方法、原理提出的“背景”，把它们连成线，比较有利于学习。举个例子：学习锁的相关问题时，我们在知道了锁

的定义、类型以及加锁方式后就会知道普通加锁方式面临的问题是不能形成冲突可串行化调度，所以我们提出了 2PL 协议来解决这一问题。解决了以后，2PL 协议又面临两个新的问题，级联终止和死锁。针对前者我们的方法是使用 SS2PL 协议；针对后者我们要么预防其发生，要么检测其是否发生，发生则想办法消除。这样基本上把整个有关锁的知识串联在一起，再去理解其中的细节也会更容易。

四、关于期末考试

结合个人的体验以及学长学姐的经验，试卷的最突出的特点就是题量很大，如果你掌握不熟练的话，那大概率是答不完卷的。

对于一开始就是只求及格或者期末临时突击只求及格的同学，下面是一点建议：把重心放在基本概念，关系代数、SQL 查询、ER 图、逻辑数据库设计这些比较固定的题目，尽量多得分。然后在数据库内核中，索引部分的 B+树和哈希索引结构这种很固定的题目一定要学会，这样的分数一定要拿到。查询执行和查询优化掌握一下基本的原理（记住各个算法的 IO 和内存要求）和基数估计的公式以及启发式优化方法（两个下推），可能能拿到一些基础分。并发控制掌握一下 ACID 性质、关于调度的基本理论、加锁解锁、再有时间就去掌握一下后面的各个协议以及死锁问题。而故障恢复部分在了解了日志记录的格式以及恢复类型后直接参考几道类似的题目看一下需要填的日志大概长什么样子，一般都会有几个很简单的空，突击的话就先保证把这里的分数拿到。不过不建议大家一开始就把目标定在期末突击到 60 分。“求其上者得其中，求其中者得其下，求其下者无所得”。这门

课的知识容量还是很大的，刚开始就把目标设定为平时不去认真学习，而靠期末突击到 60 很容易在时间上来不及突击，有很大的挂科风险，所以还是建议大家好好去学而不是突击。

对于目标在 90 或者 95 分以上乃至 100 分的同学，那就没什么好说的了，所有上面分析、总结的考点都需要熟练、准确地掌握。

五、一些可能有用的资源

1、学长学姐们的攻略：

https://zhuanlan.zhihu.com/p/450306463
https://blog.csdn.net/syStardust/article/details/125309796?spm=1001.2014.3001.5502
https://blog.csdn.net/nightcrystal012/article/details/131021790

2、老师推荐的学习工具网站：

关系代数练习网站：https://dbis-uibk.github.io/relax/landing
B+树可视化网站：https://www.cs.usfca.edu/~galles/visualization/BPlusTree.html

3、网络课程资源：

邹老师（B 站）：
https://www.bilibili.com/video/BV1ii4y1S7Uk/?spm_id_from=333.999.0.0
战老师（MOOC）（分上中下）：
上：
https://www.icourse163.org/course/0809HIT026A-1001516002?outVendor=zw_mooc_pclszykctj_
中：
https://www.icourse163.org/course/0809HIT026B-1001554030?outVendor=zw_mooc_pclszykctj_
下：
https://www.icourse163.org/course/0809HIT026C-1001578001?outVendor=zw_mooc_pclszykctj_

4、两套往年期末考试真题（网上都有，我只是省去了大家找的时间）：

https://blog.csdn.net/ymd2002/article/details/132721976?spm=1001.2014.3001.5502

5、2023 年春季学期金牌讲师团的讲义/笔记（由于当时我开展的讲座的形式是线下板书，所以没有课件和录屏，还请大家理解）：

https://download.csdn.net/download/ymd2002/88062214?spm=1001.2014.3001.5503

后记

其实一开始没有想写这么多字，写到 2000 字的时候觉得是不是有点太啰嗦了，要精简一下语言，后来仔细想了一下还是要尽可能详细地给大家叙述，把这份开箱手册的作用发挥到最大。希望这个手册中的内容能够对你的数据库学习有所帮助。最后，祝愿志在 90/95/100 的同学都能得偿所愿，60 分万岁的同学都能顺利通过。

哈工大计算学部金牌讲师团

杨明达

2023 年 9 月 6 日