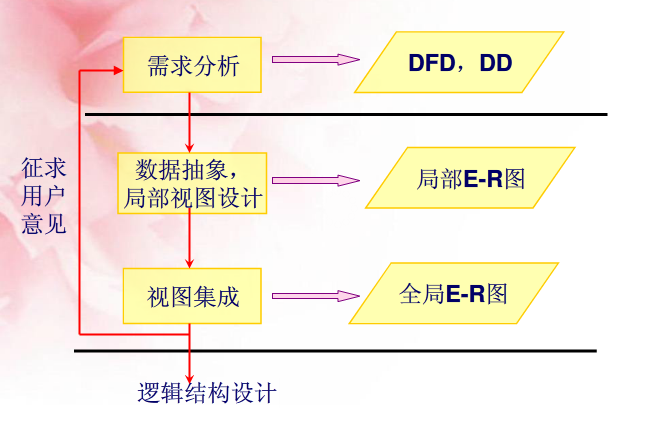
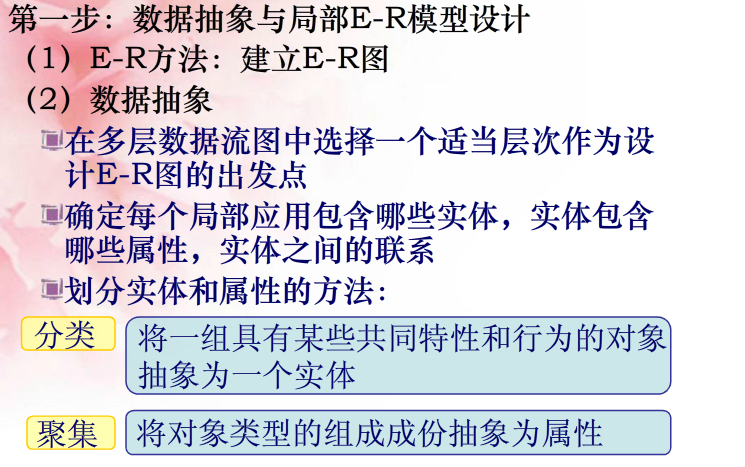
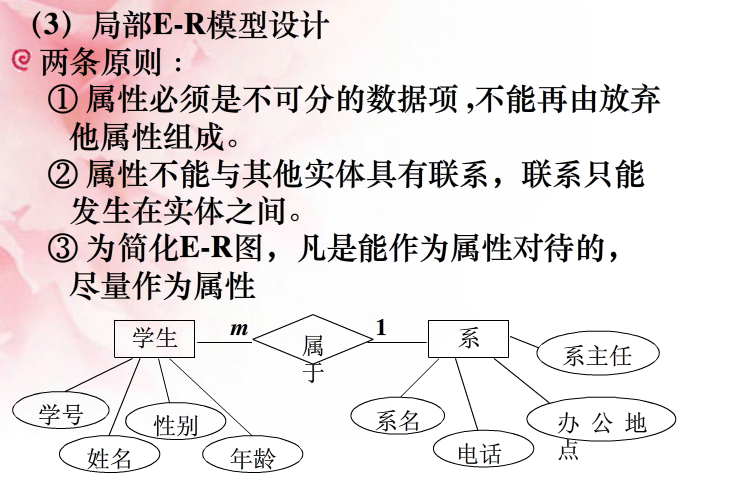
1.数据库的相关知识和数据库的设计。

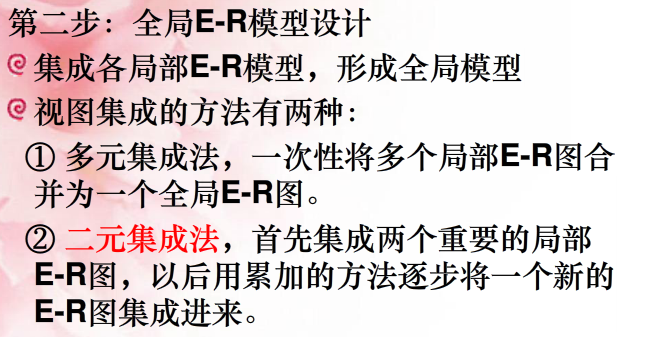
**概念结构设计**

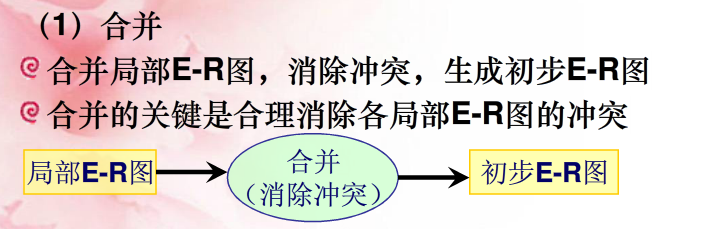


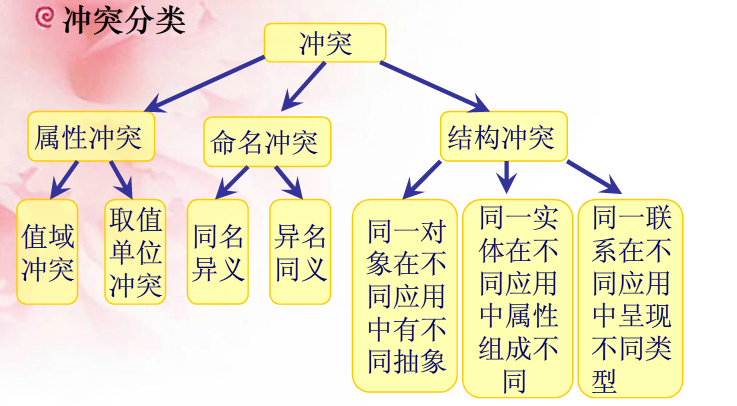


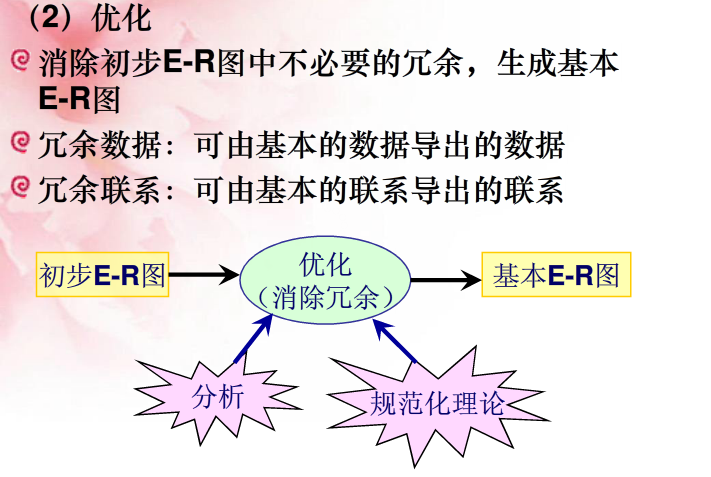




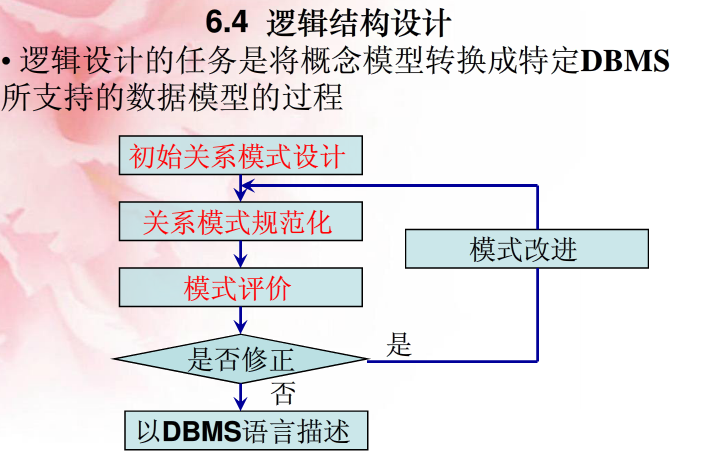


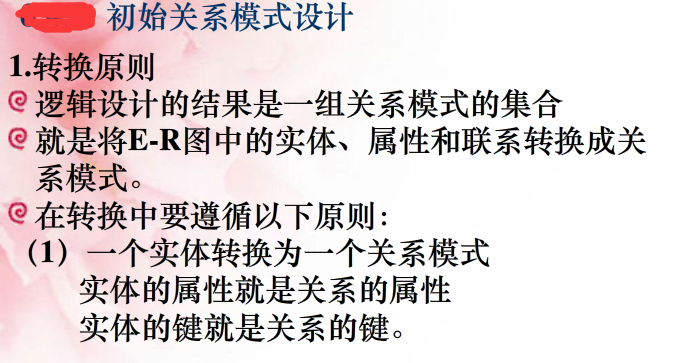


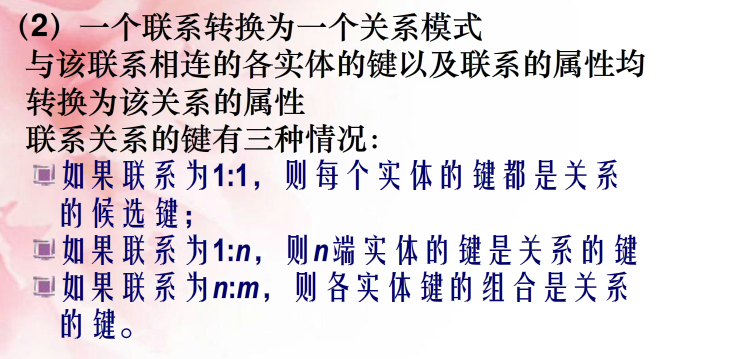


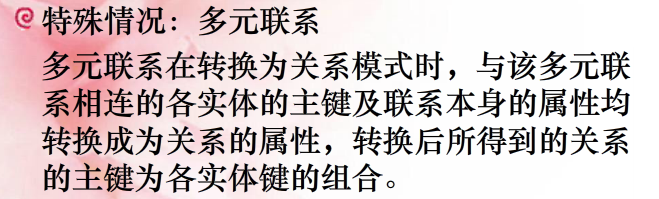


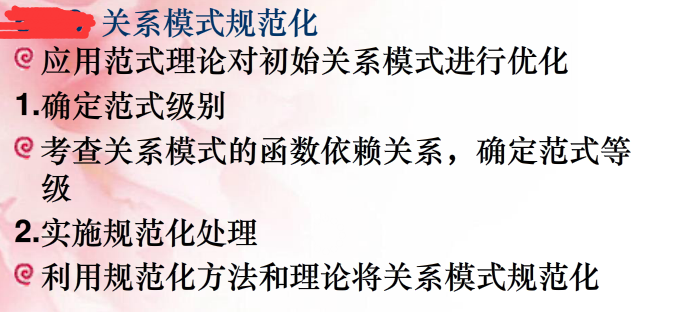
**逻辑结构设计**

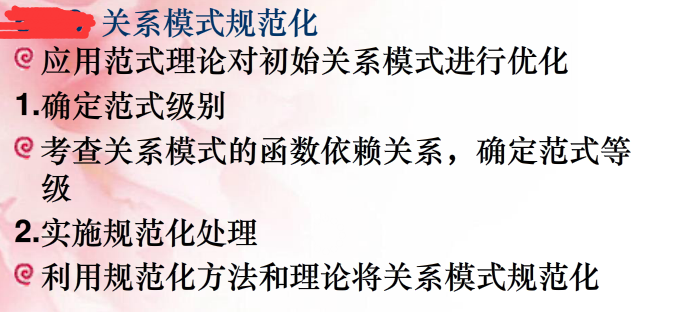


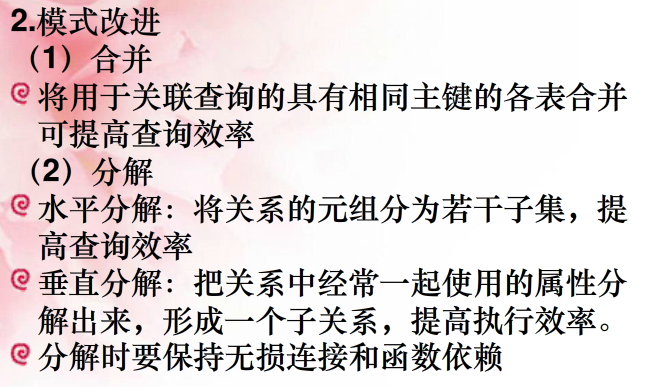




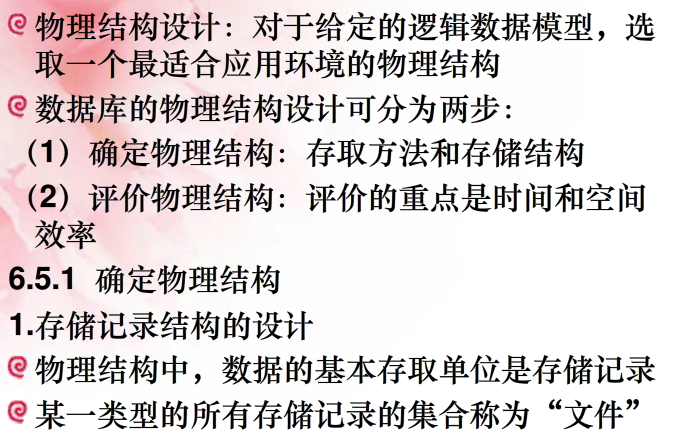


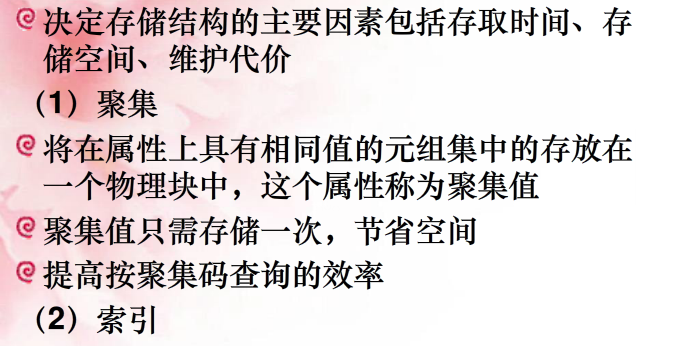


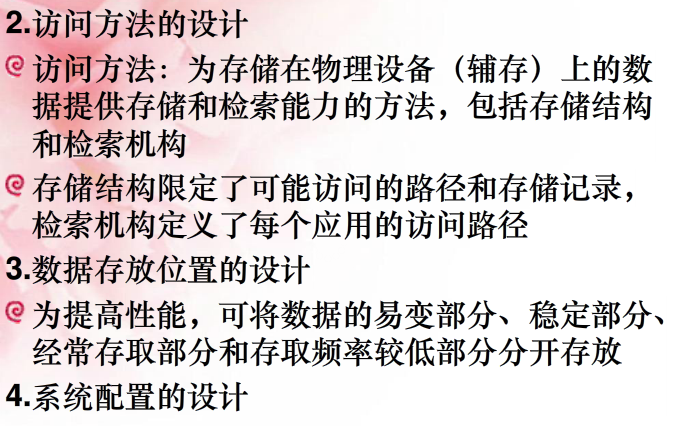




**物理结构设计**







2.需要注意问题  
　　主键与外键在多表中的重复出现, 不属于数据冗余。非键字段的重复出现, 才是数据冗余。

**防止数据库设计打补丁的方法是“三少原则”**  
　　 (1) 一个数据库中表的个数越少越好。只有表的个数少了，才能说明系统的E--R图少而精，去掉了重复的多余的实体，形成了对客观世界的高度抽象，进行了系统的数据集成，防止了打补丁式的设计；  
　　 (2) 一个表中组合主键的字段个数越少越好。因为主键的作用，一是建主键索引，二是做为子表的外键，所以组合主键的字段个数少了，不仅节省了运行时间，而且节省了索引存储空间；  
　　 (3) 一个表中的字段个数越少越好。只有字段的个数少了，才能说明在系统中不存在数据重复，很少有数据冗余，更重要的是督促读者学会“列变行”，这样就防止了将子表中的字段拉入到主表中去，在主表中留下许多空余的字段。所谓“列变行”，就是将主表中的一部分内容拉出去，另外单独建一个子表。　数据库设计的实用原则是：在数据冗余和处理速度之间找到合适的平衡点。“三少”是一个整体概念，综合观点，不能孤立某一个原则。 “文件系统集成为应用数据库，将应用数据库集成为主题数据库，将主题数据库集成为全局综合数据库。集成的程度越高，数据共享性就越强，信息孤岛现象就越少，整个企业信息系统的全局E—R图中实体的个数、主键的个数、属性的个数就会越少。　提倡“三少”原则的目的，是防止读者利用打补丁技术，不断地对数据库进行增删改，使企业数据库变成了随意设计数据库表的“垃圾堆”，或数据库表的“大杂院”，最后造成数据库中的基本表、代码表、中间表、临时表杂乱无章，不计其数，导致企事业单位的信息系统无法维护而瘫痪。

**提高数据库运行效率的办法**  
　　在给定的系统硬件和系统软件条件下，提高数据库系统的运行效率的办法是：  
　　 (1) 在数据库物理设计时，降低范式，增加冗余, 少用触发器, 多用存储过程。  
　　 (2) 当计算非常复杂、而且记录条数非常巨大时(例如一千万条)，复杂计算要先在数据库外面，以文件系统方式用C++语言计算处理完成之后，最后才入库追加到表中去。这是电信计费系统设计的经验。  
　　 (3) 发现某个表的记录太多，例如超过一千万条，则要对该表进行水平分割。水平分割的做法是，以该表主键PK的某个值为界线，将该表的记录水平分割为两个表。若发现某个表的字段太多，例如超过八十个，则垂直分割该表，将原来的一个表分解为两个表。  
　　 (4) 对数据库管理系统DBMS进行系统优化，即优化各种系统参数，如缓冲区个数。  
　　 (5) 在使用面向数据的SQL语言进行程序设计时，尽量采取优化算法。  
　　总之，要提高数据库的运行效率，必须从数据库系统级优化、数据库设计级优化、程序实现级优化，这三个层次上同时下功夫。

**算法的优化**

**尽量避免使用游标，因为游标的效率较差**，如果游标操作的数据超过1万行，那么就应该考虑改写。使用基于游标的方法或临时表方法之前，应先寻找**基于集的解决方案**来解决问题，基于集的方法通常更有效。与临时表一样，游标并不是不可使用。对小型数据集使用 FAST\_FORWARD 游标通常要优于其他逐行处理方法，尤其是在必须引用几个表才能获得所需的数据时。在结果集中包括“合计”的例程通常要比使用游标执行的速度快。如果开发时间允许，基于游标的方法和基于集的方法都可以尝试一下，看哪一种方法的效果更好。  
      游标提供了对特定集合中逐行扫描的手段，一般使用游标逐行遍历数据，根据取出的数据不同条件进行不同的操作。尤其对多表和大表定义的游标（大的数据集合）循环很容易使程序进入一个漫长的等特甚至死机。   
      在有些场合，有时也非得使用游标，此时也可考虑将符合条件的数据行转入临时表中，再对临时表定义游标进行操作，可时性能得到明显提高。

3.封号、视频是否允许观看的审核机制是如制定的。

视频审核的工作一大部分由机器负责，除了规定所有用户自发上传的视频需要进行严格审查之外，网监还会对互联网上已经发布的视频进行二次审核。

2、在中国，每个公民都有一个唯一的身份证号(ID)，通过这个ID可以全方位的定位一个人的年龄、性别、出生地等，同比MD5就是视频的ID。网监有一个巨大的危险MD5库，库内藏着各种不能流于世的视频。视频网站都需要接入这个库，并在实际审核中增加自己的库存。

3、在早期，过滤视频的第一步就是横扫整个库，把MD5吻合的视频直接剔除，并将它提交给公安违禁库。

不过，匹配MD5正在逐步弃用，原因在于MD5非常容易被篡改，在视频中加入文字或者修改任意一帧就能实现。

于是，出现了更高级别的审核办法。

4、性能级:机器审核

通过MD5之后的视频，第二步会经历机器审核的过滤。市面上已经有成熟的第三方公司提供API接口，辅助视频网站进行机器审核。

5、机器审核视频是基于深度学习图像识别云，实际也是将视频截图，由机器审核每一张截图的安全性。不过具体是3秒截一张还是5秒截一张，松紧度由视频网站自己控制。

机器审核视频原理是先建模，然后导入海量的违禁视频，让多个机器同时进行深度样本学习，再标注无法通过的图像种类，进而把这些样本揉碎、旋转、添加“噪音”，提高机器识别能力。这其中拼的不止是技术，还有样本图库的大小。据了解，在图普科技的样本库中，有超过1亿的色情样本和千万级别的极端宗教主义样本特征。

6、对于不能通过的视频，机器审核视频会给出两类结果:确定不能通过的，这类的准确率几乎能达到99.5%以上;仅作参考的，准确率在95%到97%，这可能意味着该视频需要再次进行人工审核。

7、机器基本能筛过99%的视频，只有1%需要再次动用人工。如果是150万的视频，那人工只需要审核1.5万个。

8、目前视频网站的审核坐席基本都是7x24小时工作，在临近阅兵、六四等重要日子，还要加大审核人员的配置。对于机器没有通过的视频，审核的工作人员会把其每6秒截图，一页60张图，一眼扫过去就可以判别视频是否真的触了红线。