第七章作业

1.

- (1)
  - 数据存储量的放大
    - 为实现数据映射所附加的额外信息（数据），如元数据，索引，日志，冗余
  - 数据访问量的放大
    - 如scan来获取数据
- (2)
  - RUM：
    - 在读放大、写放大、存储放大三个方向中，最多只能优化两个方向

2.

- (1)
  - B^ε树
    - **WOI实现**：将结点空间划分了一部分空间作为缓冲区，数据先有限缓存在当前结点的缓冲区中，采用追加写的方式，减少了原位更新，只有当缓冲区已满时才会将缓冲区的内容批量传递给下层结点。
    - **读性能下降原因**：缓冲区的追加写，导致了查找时开销的增加
  - LSM树
    - 数据分为多个层次组织，每个层次内部是一种有序结构。数据量逐层增大，顶层数据量小的层会位于内存中，下层数据量大的层位于外存。通过批量写入下一层实现了写操作效率的提高
- (2)
  1. 当收到读请求时先在内存里查询，如果查询到就返回
  2. 如果没有查询到就依次下沉，直到最坏情况下把所有的层查询一遍得到最终结果

3.

- (1)
  - 一种递归回归模型，将整个预测过程划分成多个Stage，每一个Stage的Model基于Key作为Input，然后选择下一个Stage所对应的Model，依次递归，直到最终的一个Stage能够预测出Key的数据位置（在限定的误差范围内）
- (2)
  1. 固定整个 RM-Index 的结构，包括层数、每层 Model 数量等

2. 用全部数据训练根节点，然后用根节点分类后的数据训练第二层模型，再用第二层分类后的数据训练第三层
3. 对于第三层（叶节点），如果最大误差大于预设的阈值，就换成 B 树