任课教师：CLZ，WJ 考试时间：2020/1/2

复习资料：上课授课内容 平时分：期末考试=5：5

一：问答题

1.试述大数据生命周期全过程，以疫情监测应用举例

2.探索性数据分析的过程是怎样的？与传统型数据分析的区别

3.传统关系型数据库在Web2.0时代有哪些不足，试列举分析

二：设计题

1.用Map和Reduce方法展示关系型数据库的自然连接过程

2.微博后台使用redis数据库设计，主要有三张表：用户表、粉丝表以及互粉表，其中粉丝表用Redis中的set这一数据结构

(1) 试分析这一设计的好处

(2) 某微博大V可能会有成百上千的粉丝，现要查询某个用户ID是不是在大V的粉丝表里，为了避免全表扫描，

请你设计一种有效的索引机制快速返回查询结果

三：计算题

1.计算TF-IDF

TF(中国)=TF(开发)=TF(软件)=0.02

包含“的”网页数约为250亿，包含“中国”的网页数为64亿，包含“软件”网页数为0.94亿，包含“开发”的网页数为0.48亿元

(1)计算上述三个词的逆文档频率(IDF)以及TF-IDF

(2)根据TF-IDF进行判断，哪个是关键字

2：给了一张如右图一张表

table[i,j]表示User i 对 Music j 的评分

(1)将该表格转换成用户关系中二分图的形式

(2)利用图的随机游走模型，判断除了1、4、5之外，系统应该给Alice推荐什么歌？

