by *Dseid*

# 哈尔滨工业大学（深圳）
# 2022 春季学期
# 大数据计算基础考试试卷回忆版

2022 年 6 月 08 日　　考试时间：2 小时

**前言**

1. 选择题绝大部分均来自课后习题、且未调换选项顺序。此处**只回忆个别课后习题中没出现过的题目**。标 * 代表不确定原题具体表述，只给出类似题目。
2. 大题部分涉及内容大部分在最后一节课强调过。大体部分可能无标准答案，这里只给出比较确定的题目的参考答案。
3. 单击此处访问大数据专业 Github 仓库

---

1. 选择题，共计 50 分，分为单选与多选。

    1) * javac.exe 将 java 源文件编译为的字节码文件的后缀名是?
    （A）.java
    （B）.class
    （C）.exe

    (答案：B)

2) 下列哪一项说法是错误的?

(A) UserCF 算法推荐的是那些和目标用户有共同兴趣爱好的其他用户所喜欢的物品

(B) ItemCF 算法推荐的是那些和目标用户之前喜欢的物品类似的其他物品

(C) ItemCF 算法的推荐更偏向社会化，而 UserCF 算法的推荐更偏向于个性化

(D) ItemCF 算法倾向于推荐与用户已购买商品相似的商品，往往会出现多样性不足、推荐新颖度较低的问题

(答案：C)

2. * 简答题, 共 6 道题目占 30 分

1) HDFS 具有很好的容错能力，并且能够兼容廉价的硬件设备。请阐述 HDFS 具体是如何通过廉价硬件设备而实现高性能的?

2) 请阐述 Hbase 中 region 的具体定位过程。
参考答案：见课本 P77 页。

3) 请举例说明 CAP 理论的含义。

4) 请阐述在 MapReduce 中 shuffle 具体的过程以及作用?

5) 请阐述 storm 处理流数据的工作机制。

6) 请描述两种协同过滤算法的差别以及各自的优缺点。
参考答案:
UserCF 算法和 ItemCF 算法的思想、计算过程都相似
两者最主要的区别：UserCF 算法推荐的是那些和目标用户有共同兴趣爱好的其他用户所喜欢的物品 ItemCF 算法推荐的是那些和目标用户之前喜欢的物品类似的其他物品。UserCF 算法的推荐更偏向社会化，而 ItemCF 算法的推荐更偏向于个性化
UserCF 算法的推荐更偏向社会化：适合应用于新闻推荐、微博话题推荐等应用场景，其推荐结果在新颖性方面有一定的优势。UserCF 缺点：随着用户数目的增大，用户相似度计算复杂度越来越高。而且 UserCF 推荐结果相关性较弱，难以对推荐结果作出解释，容易受大

众影响而推荐热门物品。
ItemCF 算法的推荐更偏向于个性化：适合应用于电子商务、电影、图书等应用场景，可以利用用户的历史行为给推荐结果作出解释，让用户更为信服推荐的效果 ItemCF 缺点：倾向于推荐与用户已购买商品相似的商品，往往会出现多样性不足、推荐新颖度较低的问题。

3. * 综合题, 共 2 道题目各占 10 分。

   (a) Spark 具体是怎么样做到优于 MapReduce 的?

   (b) 举出一两个大数据在实际生活中的应用。