2019 年大数据计算基础练习题

注：此复习题仅供同学们练习所学内容用，与考试题没有任何联系

一、简答题

1. 什么是近似算法？应该以何种方式衡量近似解代价与优化解代价的差距？
2. 大数据算法设计中，分别采用什么算法解决“主存不足”、“规模过大”的问题？
3. 什么是众包？众包算法有哪些应用？
4. 大数据和 SPARK 的关系是什么？
5. 大数据的几个 V 是什么？
6. HDFS 的核心模块都有哪些
7. MapReduce 是什么？
8. 简述 MapReduce 框架完成单词计数（WordCount）的过程。
9. 什么是 NoSQL？NoSQL 有哪些特点？

二、问答题

1. 对于无重边的简单无向图 G，顶点集为 V、边集为 E，内存大小为 M，磁盘块大小为 B，讨论求解连通分量和最小生成树问题：

(1) 简述$|V| \le M$、$|V| > M$ 时分别如何求解连通分量，并分析计算图 G 连通分量的 I/O 复杂度（可举例说明）。

(2) 简述求解图 G 连通性的算法如何扩增为求其最小生成树(MST)的算法，并分析求解最小生成树的 I/O 复杂度。

2. 给定长度为 N 的 0,1 数组（即元素只包含 0 和 1），给出判定该数组是否含有 1 的亚线性时间判定算法，并分析该算法判定的精确性。

3. Redis 作为一个高性能内存键值存储，通常被用来作为常规数据库，如 MySQL 的缓存（Cache）使用。假设小王在开发应用程序时，为了加快数据访问速度，在读取数据时只要需要的数据在 Redis 中，则直接读取而不再访问 MySQL，否则，从 MySQL 中读取数据然后存到 Redis 中；在写数据时，先将数据写入 MySQL，再将 Redis 中相应数据项删除。

(1) 请举例说明，当数据有多个副本存储在不同的用网络相连节点中时，即使发生了网络分区，系统也能够同时达成最终一致性和可用性。

(2) 在不考虑故障的情况下，请问小王构建的系统在并发读写时是否是强一致（strongly consistent）的？如果是，请说明理由，如果不是，请给出一个并发读写操作的序列，使得最终读取到的数据不满足强一致性。

(3) 在不考虑故障的情况下，请问小王构建的系统在并发读写时是否是最终一致（eventually consistent）的？如果是，请说明理由，如果不是，请给出一个并发读写操作的序列，使得最终读取到的数据不满足最终一致性。

4. 考虑外存上维护有序列表的问题。假设：数据都是整数，数据量为 n，外存的存取块（磁盘块）的大小是 B，内存大小为 m。

(1) 给出 MergeSort 算法的描述，它的 I/O 代价是多少。
(2) 针对排好序的数据，给出搜索算法（查找一个元素是否出现在列表中），它需要哪些额外的存储结构，它的 I/O 代价是多少。
(3) 考虑外存上维护有序列表的问题。假设：数据都是整数，数据量为 n，外存的存取块（磁盘块）的大小是 B，内存大小为 m。
(4) 给出 MergeSort 算法的描述，它的 I/O 代价是多少。
(5) 针对排好序的数据，给出搜索算法（查找一个元素是否出现在列表中），它需要哪些额外的存储结构，它的 I/O 代价是多少。

6. Bloom filter 是由 Howard Bloom 在 1970 年提出的二进制向量数据结构，它具有很好的空间和时间效率，被用来检测一个元素是不是集合中的一个成员。

(1) 在垃圾邮件检查问题中使用白名单方法，白名单中有 s 个邮件地址。使用 Bloom filter 进行检查时，初始化和检查流程分别如何进行？

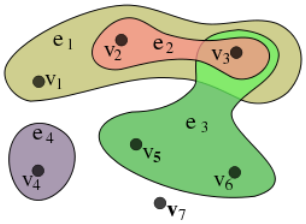
(2) 在 Bloom filter 的过程中，会出现“False positive”、“False negative”中何种错误？用何种方式可以降低这种错误发生的概率？

7. 容错是大数据系统的一个重要课题。主流的大数据计算框架都实现了容错功能。

(1) MapReduce 系统是如何实现容错的？
(2) Spark 系统是如何实现容错的？它克服了 MapReduce 系统的哪些缺点？
(3) Spark 实现容错的方式在当计算变得非常复杂时可能会出现什么问题？有哪些方法可以缓解这一问题？

8. 图论中，超图（HyperGraph）是一种广义的图，特点是一条超边可以连接多个点。超图 H 是一个集合组 H=(X,E)，其中的 X 是顶点的集合，E 是 X 的非空幂集。

下图是一个超图的例子 X = {v_1, v_2, v_3, v_4, v_5, v_6, v_7} E = {e_1,e_2,e_3,e_4} ={{v_1, v_2, v_3}, {v_2,v_3},{v_3,v_5,v_6},{v_4}}.



超图的好处可以用一个简单例子来解释，假设边是文章，点是文章作者，在简单图中，容易丢失同一篇文章的多个作者。因为简单图只能是两点一线，一篇文章只能连接两个作者；但是对于超图来说，利用其特性他能描述更多。

请设计分布式计算框架来让程序员方便地写程序处理超图上的计算(需要支持但不限于下列计算任务：寻找两点间最短路径、查找度最大的顶点、查找包含顶点 u 和 v 的边)，要求系统具有可扩展性和容错性。回答如下问题：

(1) 请设计该计算框架中表示超图的数据结构
(2) 请面向自己熟悉的语言设计 API 函数(需要标明使用何种程序设计语言)
(3) 请利用该计算框架编写程序，计算图中度最大的顶点
(4) 请设计系统架构，能够支持上述数据结构和 API 函数，并论述为何该系统架

构具有可扩展性和容错性。

(5) 以问题(3)的程序为例，假设系统中有至少三台机器，顶点和边的存储无特殊划分方法，写出图 1 在该计算框架上的运行过程。

(6) 请利用该计算框架编写程序，计算图中 u 和 v 之间最短路径

(7) 设计超图划分策略，并论述为何该划分策略能够有效支持超图的分布式计算

9. 集合上的 Jaccard 相似度是这样定义的：$Sim\ (C_1, C_2) = |C_1 \cap C_2|/|C_1 \cup C_2|$。例如：A={a,b,c}，B={b,c,d,e}那么 Sim(A,B)=2/5=0.4。

（1）解释什么是最小哈希方法（MinHash），它有什么性质。

（2）多重集合（每个元素可以出现多次）上也可以定义 Jaccard 相似度，这里多重集合的交（或者并）定义为交集（或者并集）中每个元素出现的次数等于该元素在两个多重集合中出现次数的最小值（或者最大值）。例如：{a,a,b,b}∩{a,a,a,b}={a,a,b}，{a,a,b,b}∪{a,a,a,b}={a,a,a,b,b}。相应的，可以定义多重集合上的 Jaccard 相似度。请将 MinHash 方法扩展到多重集合。

10. 设计外存有效的自相似连接算法，即给定一个字符串集合 S 和一个阈值 ε，求 S 中所有编辑距离小于 ε 的字符串对。编辑距离用于衡量字符串相似性，字符串 s1 到 s2 的编辑距离是 s1 变化到 s2 所需要的增加字符、删除字符和修改字符的最小次数。

11. 令 X,Y,Z 为三个 n*n 的矩阵。假设矩阵在磁盘中以行形式存储，现在需要计算 Z，使其等于 X*Y。

12. 括号匹配问题：给定一个“括号”序列，判断是否合法。例如：（）（）（（））是一个合法的序列；（（）（（）是一个非法的序列。

（1）基于 PRAM 模型设计一个并行算法检验给定的括号序列是否合法。（假设问题的输入很长，被按照先后顺序分块存储在有序的计算节点上，每个节点知道机器的总数以及自己的机器编号）

（2）分析所设计算法所需的时间代价、空间代价。

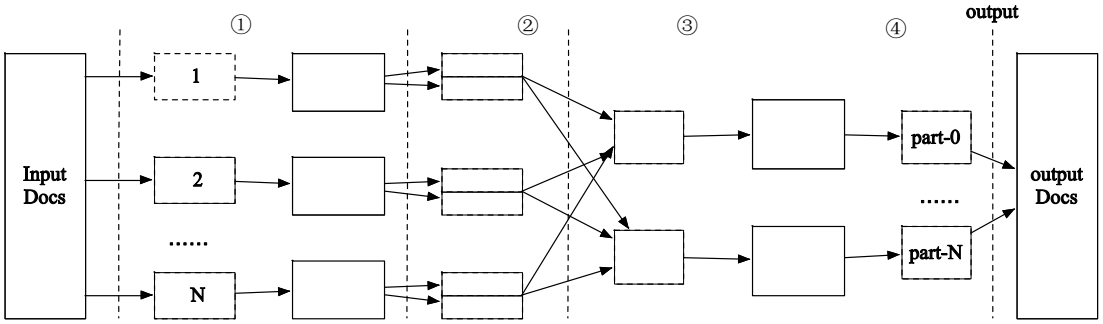
（3）PRAM 模型假设有一个共享存储，这在实际应用中通常是不现实的。尝试设计一个基于消息传递的并行算法解决括号匹配问题。

13. 搜索引擎会通过日志文件把用户每次检索使用的所有检索串都记录下来，每个查询串的长度为 1-255 字节。假设目前有一千万个记录（这些查询串的重复度比较高，虽然总数是 1 千万，但如果除去重复后，不超过 3 百万个。一个查询串的重复度越高，说明查询它的用户越多，也就是越热门。），请你统计最热门的 10 个查询串，要求使用的内存不能超过 1G。

14. 现需使用 Hadoop 框架实现单词计数(WordCount)。

(1) 下图展示了 MapReduce 框架的执行流程，请指出图中(1)(2)(3)(4)这四个步骤分别执行了什么操作、结合下面这一具体的输入样例阐述每一步的作用和具体的计算流程。

```
hello world i love you
i love you my love
i love this world
```



①：
②：
③：
④：

(2) 补全 Mapper 及 Reducer 的实现代码。

```
public static class TokenizerMapper extends Mapper
   <Object, Text, Text, IntWritable>{
   private final static IntWritable one = new
IntWritable(1);
   private Text word = new Text();
   public void map(Object key, Text value, Context
      context) throws IOException, InterruptedException{
            请在此处补全代码
}}}

public static class IntSumReducer extends Reducer
   <Text,IntWritable,Text,IntWritable>{
   private IntWritable result = new IntWritable();
   public void reduce(Text key, Iterable<IntWritable>
      values, Context context) throws IOException{
            请在此处补全代码
}}}
```

(3) Hadoop 使用的数据存储机制是什么？请简要陈述它的优点和缺点。

15. 假设一个数据库中有两张表 t1, t2，t1 大小远大于 t2。现在要对两张表做连接操作，连接条件是 t1.A == t2.B。下面有两种数据划分方式：第一种，分别以 t1.A 和 t2.B 为划分属性进行散列划分；第二种，将 t1 进行循环划分，然后将 t2 不进行划分而直接复制到各个节点上。
（1）请给出一种情况，使得第一种划分方式比第二种划分方式性能更好。
（2）请给出一种情况，使得第二种划分方式比第一种划分方式性能更好。

16. 现需使用 Storm 在输入实时数据流的情况下实现单词计数(WordCount)。

(1) Hadoop 为什么不适合流计算？相比之下，Storm 有哪些特点使其更适合处理实时数据？
(2) 设计一个 Topology 可以解决实时数据流上的单词计数(WordCount)问题，画出拓扑图并说明每一个节点的作用。

17. 用$\mathcal{H} \subseteq Y^X$表示2-通用哈希函数族，对于某常量c>0,$|Y| = cM^2$。假设我们用随机函数$h \in_R \mathcal{H}$来哈希一个由X中元素组成的数据流σ，假设σ最多包含M个不同的元素，证明：发生冲突(两个不同的元素对应到哈希中的同一位置)的概率最多为1/(2c)。

18. 给定一个数据流 A[1…n],如何估计$\sum_i A[i]^2$?若有另一个数据流B[1…n],如何估计$\sum_i A[i]B[i]$?

19. 一个（有向或无向）图，包含n个顶点，m条不同的边，$m < n^2$。图的边以流的形式给出，流中的每个元素代表一条边，用一对点表示成形如（i，j）的形式。m条不同边中的每条边都可以在流中多次出现，没有边会缺失。用d_i表示顶点i的不同的邻居的数目。目标是近似$M_2 = \sum_i d_i^2$。M_2的概念类似于F_2，关键的差别是，M_2仅仅计数新的与众不同的元素（未出现过的）。

20.为了保障航运安全，旅客不允许携带危险品乘坐飞机，因此在旅客登机前需要由安检人员检测旅客是否携带危险物品。假设共有 n 名旅客需要登机，用 o(n)时间近似判断这 n 名旅客是否有人携带危险品。

21.在输入的正文串（长度为 n）查找某一字符是否出现；若出现，输出 1，否则输出 0。设计时间复杂度为 o(n)的算法求解这个问题。

22. miRNA（微小 RNA）是在生命活动中有重要功能，由约 21 个碱基构成，比如 AUGUCCUCCUUAUGCCUAUGC 可能就是一条 miRNA。如果我想知道细胞内有没有我感兴趣的 miRNA，可以通过测序解决。测序结果可能就是 n 个 21 碱基的序列组成。时间复杂度为 o(n)的算法，给定包含 n 个测序结果的集合 S（n 条序列）和感兴趣的 miRNA 序列 l，判定 S 是否包含 l。

23. 一副二值图（有 n 个像素），仅有 0 和 1 组成，设计时间复杂度为 o(n)的算法判断该图的黑色(1)和白色(0)是否为各占一半？