中国科学院大学

考试日期：2022.06.01

课程编号：083500M02001H

试 题 专 用 纸

课程名称：大数据系统与大规模数据分析

任课教师：陈世敏　孙翼

姓名_____　学号_____　成绩_____

注意：　1、请在答题的电子文档开始写明：学号、姓名。文件命名为：学号-姓名.pdf。
2、请在课程网站>作业>期末考试，上传学号-姓名.pdf，重复上传仅保留最后一个文件。

简答题 （每题 10 分， 共 10 个题，共 100 分）

1. （10 分）请回答下述关于关系模型的问题：
   a. （2 分）什么是关系模型？
   b. （3 分）什么是主键、外键、连接键？
   c. （5 分）请说明重要的关系代数算子（选择、投影、连接、分组聚集、排序）在 select 语句中的表达形式。

2. （10 分）假设一棵基于外存的 B+-Tree 中存储了 10 亿（即 10^9）个索引项（key, recordID）。每个 key 是 8B，每个 recordID 是 8B，每个树节点是 4096B 的数据页。请回答下述问题：
   a. （3 分）每个叶子节点存储一个索引项数组、一个 8B 的整数 num_entries、一个 8B 的兄弟链表指针 next，请问叶子节点的索引项数组可以存储多少个索引项？
   b. （3 分）假设每个叶子节点存储了 200 个索引项，每个内部节点（除了根节点）有 200 个孩子，请问这棵 B+-Tree 有多少层？每层有多少节点？
   c. （4 分）对于 b 题中的 B+-Tree，已知最上面两层节点都缓存在内存缓冲池中。如果希望这棵 B+-Tree 能够支持每秒钟 1 万次（10000 op/s）随机查询操作，请问存储 B+-Tree 的外存设备应具有怎样的性能？

3. （10 分）请回答下述关于分布式文件系统的问题：
   a. （6 分）请比较 NFS 和 HDFS（设计目标、系统结构、容错恢复、并行访问等）。
   b. （2 分）如果在 NFS 的适用场景下，采用了 HDFS 存储文件，有什么问题？
   c. （2 分）如果在 HDFS 的适用场景下，采用了 NFS 存储文件，有什么问题？

4. （10 分）请回答下述关于键值存储系统的问题：

   a. 请从数据模型、系统结构、数据存储、容错恢复、及主要技术等 5 方面比较 Dynamo，BigTable，和 Memcached 这 3 个系统。

5. （10 分）图数据库通常存储顶点、边、以及属性信息。假设在图数据库中有一个顶点 A，A 有 100 条边，连接 100 个邻居。一个图查询操作 OP 需要读取 A 的所有邻居。请回答下述图数据存储的问题：

   a. （4 分）请说明 Neo4J 的存储结构。图查询操作 OP 将引起怎样的文件操作？

   b. （3 分）如果图数据库的下层存储采用关系数据库系统，请设计一组关系表，来存储图数据。图查询操作 OP 怎么用（单个）Select 查询语句来支持？

   c. （3 分）如果图数据库的下层存储采用的是键值存储系统 BigTable，请设计一种图数据映射到键值对的方式。图查询操作 OP 怎么样用键值操作来支持？

6. （10 分）请回答下述关于 MapReduce 的问题：

   a. （3 分）什么是 Map 和 Reduce？请比较 MapReduce 和关系运算 Select 的异同。

   b. （3 分）请说明 word count 的输入、中间结果、输出、及各个阶段的操作。

   c. （4 分）假设 word count 的输入是 10 亿（即 10^9）行文本，每行文本包括 20 个长度为 8B 的单词。假设不同单词的总数为 1 百万（即 10^6）个。输出每行平均 30 字节。有 100 个 Map Task 和 1 个 Reduce Task。

      - （2 分）请列式并估计中间 shuffle 结果的大小、最后输出结果的大小。
      - （2 分）如果采用了 Combiner，请列式并估计中间 shuffle 结果的大小。

7. （10 分）请回答下述关于 Spark 的问题：

   a. （3 分）Spark 的主要数据模型是什么？请解释其具体涵义。

   b. （3 分）Spark 的 Transformation 和 Action 是什么？MapReduce 中的 Map 和 Reduce 各自对应 Spark 中的什么运算？

   c. （4 分）请比较在 Spark 上实现 PageRank 与在 Pregel 同步图计算系统中实现 PageRank 的异同，包括系统结构、执行方式、内存占用等方面。

8. （10 分）完成如下问题的计算

   a. （6 分，每个距离 3 分）对于两个向量 v1 = [0; 1; 1; 0; 0; 0; 1]，v2 = [1; 0; 1; 0; 1; 0; 0] ，求它们的 Jaccard 距离和余弦距离。

   b. （4 分）给定$(0.3, 0.7, 0.7, 0.3)$-sensitive 族，对其做两次 AND 然后在做 2 次 OR 操作后所形成的（ ， ， ， ）sensitive 族。

9. （10 分）对于2×3的矩阵 $A = \begin{pmatrix} 2 & 3 & i \\ -1 & 2 & -i \end{pmatrix}$

   a. （5 分）计算矩阵A的两个奇异值。

   b. （5 分，分别 2 分、2 分、1 分）分别计算矩阵A的谱范数$||A||_2$、 F-范数$||A||_F$以及核范数$||A||_N$

   （提示，计算 AA^* 利用这个矩阵的迹和行列式计算其特征值）

10. （10 分）一个以整数构成的数据流，它由一个 1，两个 2，三个 3，依此类推，直到十个 10 等组成。请给出下列问题的答案：

    a. （2 分）计算这个数据流的零阶矩。

    b. （4 分）计算这个数据流的二阶矩（surprise number）。

    c. （4 分）定义散列函数（Hash Function）$h(i)$为 32 位二进制数（例如，$h(1) = 00 \ldots 001$, $h(2) = 00 \ldots 010$），考虑 Flajolet-Martin 算法应用于这个散列函数 h 来估计此数据流中不同元素的个数，请问数据流中不同元素的个数的估计值是多少？