北京大学数学科学学院试题

（答案写在答题纸上）

考试科目：大数据与分布式计算 姓名：_____ 学号：_____

考试时间：2022 年 12 月 25 日 任课教师：陈立军

1. 关于大数据思维，我们给出如下一些关键词：“因果性与相关性”，“可解释机器学习”，“机器学习泛化”，“科研可重复性危机”，请写一段不少于 500 字的文字叙述，把这些关键词串联起来。
2. 我们课程中介绍过几个数据结构，一致性哈希、Bloom Filter、LSM 树，请画出它们大致的结构，并列举其在大数据领域的应用场景，解决了什么问题。
3. 画出 GFS、MapReduce、Spark、Pregel 的体系结构图，标注角色名称，并简单阐述每个角色的作用。
4. 画出 GFS 的写数据流程图，按顺序标明每个步骤。
5. MR 的 Shuffle 过程会产生大量磁盘 IO，俗称泛洪。请列举一些减少 Shuffle 过程中磁盘 IO 的措施，或者说有哪些 Shuffle 类型？
6. 请列举三种不同的连接操作类型，图示出其执行过程，并回答各自适合的场合。
7. 请用 ZooKeeper 的 API 来描述一个消息订阅/发布机制的实现。
8. 什么是 ZooKeeper 中的惊群效应？什么是 Redis 中的缓存雪崩？如何避免？
9. Spark 生态系统包括哪些组件？这种全栈式大数据平台具有什么优点？
10. 什么是 RDD 的血缘，它在故障恢复中有什么作用？如何把一个 RDD DAG 划分成多个 stage？这样划分的优点是什么？
11. ABCDE 基于 Paxos 协议商量爬山时间，AE 是 Proposer，BCD 是 Acceptor。假定 A 只能与 BC 通讯，E 只能与 CD 通讯。A 提议周三，E 提议周五，版本号不同。请给出 Paxos 协议下的各种可能的商议过程。
12. 在 word.txt 文件中存储有如下数据：

To the world, you may be one person； To one person, you may be the world

在 Spark 中依次执行如下命令：

```
scala> val    lines = sc.textFile("file:///usr/local/spark/mycode/rdd/word.txt")

scala> val    words = lines.flatMap(line => line.split(" "))

scala> val pairRDD = words.map(word => (word,1))

scala> val GroupRDD = pairRDD.GroupByKey()

scala> val ReduceRDD = pairRDD.reduceByKey(_ + _)
```

请写出 GroupRDD, ReduceRDD 的结果。

13. 有如下股票交易表，Stock(stock_id, close_price, date)，分别为股票号、收盘价、交易日期（假定交易日期是单调递增的连续序列号）。所谓多头排列是五日线高于十日线，十日线高于二十日线。请用 Hive 的窗口函数找出多头排列的股票（为简单起见，只要有一天满足上述条件即可）。
14. 给定一个朋友网络，节点代表人，连边表示相互认识。现在要计算每个节点的聚类系数，也即判断他的两个朋友之间彼此是否也是朋友。请分别给出用 MapReduc 和 Pregel 解决这个问题的算法过程。
15. 在分布式计算中，有时各个节点需要共享一些全局信息，请回答在 Hadoop，Spark，Pregel 中，各自提供全局变量的方式是什么？
16. 数据倾斜是分布式系统中的一个性能瓶颈，你能想到哪些措施可以消除这方面的影响？