大数据2023春

填空

一空1.5分，共45分

1. 按程序模型并行分类：共享内存变量、（）、（）
2. MPI点对点通信分为（）和（）两类
3. 什么是MapReduce：（）、（）、（）
4. 评估可并行度：（）定律（或直接默写公式）
5. Map的中间结果通过（）传递给Reduce，Map节点共享文件可通过（），全局传递参数可通过（）
6. Combiner的作用是在（）阶段减少（）
7. Partitioner的作用是（）
8. 在资源充足时，主节点调度Map节点执行作业的原则是（）
9. HDFS中，管理元数据的是（），管理实际数据的是（）
10. 分布式Join表时，将大表和小表Join的方法是（）
11. HBase的数据模型是（），记录和检索的关键字由（）、（）、（）组成
12. Hive的Driver由（）、（）、（）组成
13. Spark的操作有（）、（）两种，主要区别是（）
14. Spark RDD的两种容错方法是（）、（）

简答

一问6分，共30分

1. Hadoop的主要设计思想和技术特点
2. HDFS的NameNode和DataNode在出错时分别怎么恢复
3. 在计算、通信或负载均衡的角度，讲三个性能优化方面的处理措施和原理
4. Hive的Table、Partition、Bucket分别是什么
5. Spark程序的构成、执行过程，RDD DAG比起MapReduce有什么优点

代码

1. （12分）加权KNN，思路和伪代码
2. （13分）倒排索引，输出格式为`单词 <文档1，词频1>，<文档2，词频2>，...`，要求根据文档号降序排序，写出思路和伪代码