并行计算

第十次课作业

姓名 刘恒星 学号 2022229044

请简要回答以下问题：

1. MapReduce编程模型为什么适于处理大规模数据？

答：因为它运行于大规模分布式集群环境 (>2000节点)，能够自动实现分布式并行计算，有很好的容错和负载平衡，并且支持任务调度和状态监控。他还有很好的扩展性，方便添加更多的节点扩展计算能力。

2、 为什么MapReduce编程模型采用key-value的数据结构？这种数据结构是否适于串行编

程？

答：这样的数据结构可以方便地进行分片和并行处理，同时也非常适合于MapReduce所采用的Map-Reduce算法。key-value对具有明显的区分性，所以在MapReduce中使用这种数据结构可以使不同的节点对不同的key-value对进行计算，从而有效地实现分布式计算。

这样的数据结构也可以用于串行编程，但是对于并行来说更加的具有优势。

3、 Hadoop封装了MapReduce并行计算过程中的哪些功能？

答：分布式存储：Hadoop提供了一个分布式文件系统（HDFS），它可以将大规模数据集存储在多台机器上，以实现数据冗余备份和高可用性。

数据划分：Hadoop将输入数据分割成小块，这些小块被称为InputSplit，并且每个InputSplit都被分配给一个Map任务进行处理。这样可以将输入数据分散到不同的计算节点上，从而有效地利用集群的计算能力。

故障恢复：Hadoop具有自动故障检测和恢复机制，当某个计算节点出现故障时，Hadoop可以自动将该节点上未完成的任务重新分配给其他可用节点进行处理，从而保证作业的完成。