一、分布式系统定义。

1. 简述分布式系统定义。
2. 多 CPU 的计算机是否算分布式系统，为什么？(4)
3. 一个物理主机是否可包含多个分布式计算节点，为什么？(4)
4. 多计算节点构成的分布式系统具有什么好处。(4)

二、分布式架构模式。

1. 简述主从模式、对等模式的定义及各自的优缺点。(8)
2. Google 的 GFS 和 MapReduce 采用的哪种分布式架构模式。(2)
3. 比特币采用的哪种分布式架构模式。(2)

三、RPC。

1. 对于程序开发者来说，RPC 的主要作用是什么。(4)
2. RPC 的实现原理和 RPC 中间件的主要作用。(6)
3. IDL（接口定义语言）的作用是什么。(2)

四、MOM。

1. 消息队列模式与主题/订阅模式的区别。(4)
2. 消费者接受消息的三种方式。(3)
3. MOM 通信的优点。(4)

五、HDFS 分布式文件系统。

1. HDFS 分布式文件系统中 NameNode 节点和 DataNode 节点的主要功能是什么。(4)
2. HDFS 如何保证数据存储的可靠性。(3)
3. HDFS 文件系统读取文件时与 NameNode 和 DataNode 的交互过程。

六、数据分区。 //题目过长，大致意思为管理景点访客信息，采用分布式关系数据库，7 个节点，主键为身份证号，另一重要键值为日期，基于哈希函数，以身份证为主键进行分区。

1. 插入数据的基本思想。(4)
2. 按身份证号进行信息查找的基本思想。(2)
3. 按日期查找当日访客信息表的基本思想。(4)
4. 若不用哈希分区，采用游客来自省份(身份证号确定)进行分区存储，各省份对应 7 个物理节点(华南，华北，西北，东南，中北，XX，XX)，相对于哈希分区，这种分区方式有什么缺点。

七、MapReduce 程序设计。

微博大 V 选拔赛。

每行数据<ID1，ID2>：ID1 是 ID2 的粉丝；微博粉丝量大于 50 万的微博用户视为大 V，输出微博大 V 的 ID 和其粉丝数量。

设计 map 和 reduce 方法的伪代码。

八、Spark 程序设计。 //同样伪代码。

输入文件为学生成绩信息，包含了必修课与选修课成绩，格式如下：
班级1, 姓名1, 科目1, 必修, 成绩1
 （注：
 为换行符）
班级2, 姓名2, 科目1, 必修, 成绩2
班级1, 姓名1, 科目2, 选修，成绩3
....., ..., ..., ...
编写一个Spark程序，同时实现如下功能：
[redacted]
2. 统计学生必修课平均成绩在：90~100,80~89,70~79,60~69和60分以下这5个分数段的人数。