# 大数据笔试卷

### 填空题

1. 查看磁盘存储情况命令（）
2. 查看报告系统运行时长及平均负载命令（）
3. 8020 或者9000是哪个服务常用端口（）
4. Yarn的调度器有哪些（）
5. Zookeeper的选举策略（）
6. Kafka的机器数量计算公式（）
7. rowkey设计原则（）
8. Spark部署方式，请说出至少三种（）
9. Hadoop 2.x中HDFS块大小（）
10. ElasticSearch是基于什么来处理并发控制（）

### 简答题

1. 如何理解Spark中的血统概念？
2. LZO是Hadoop自带的压缩方式？LZO与GZIP有什么区别，且他们常使用在哪个阶段？
3. 简述Hadoop的推测执行机制
4. 简述Hive的管理表和外部表的区别
5. 描述下Zookeeper的监听机制
6. 简述RDD（cache和persist）的作用和存储级别
7. 简述Spark的宽窄依赖，以及Spark如何划分stage，每个stage又根据什么决定task个数?
8. 简述ElasticSearch相关度评分算法机制
9. 简述ElasticSearch的document写入原理
10. Flink如何实现Exactly once语义

### 优化题

1. Hive中表join时容易出现数据倾斜，如何处理？
2. 如果写入文件过量造成NameNode宕机，如何处理？
3. 如果Redis缓存穿透，如何处理?
4. 线上JAVA应用CPU过高。如何定位和解决思路？
5. Kafka消息数据积压，Kafka消费能力不足怎么处理？