# 《大数据计算框架》复习

题型：选择题（30分）、填空题（10分）、分析论述题（20分）、编程题（8分）、综合题（32）

考试范围：

1. 大数据概述
2. Linux基础知识及相关命令
3. 分布式文件系统HDFS
4. 分布式数据库HBase（含MySQL）
5. NoSQL数据库
6. MapReduce
7. 大数据处理架构Hadoop
8. Spark
9. 流计算

10、数据仓库Hive

11、大数据在不同领域的应用

一、概念与分析论述（见书后习题）

1. Hadoop生态系统中的核心组件
2. 大数据决策与传统的基于数据仓库的决策
3. Hadoop特性、作用以及应用场景
4. 常用的流计算框架
5. 数据分区
6. 大数据的四大特征
7. 数据清洗概念及常见操作（步骤）以及数据清洗的必要性
8. Flink、CAP
9. HDFS HA架构构成组件及其具体功能
10. 图计算框架Pregel的应用
11. 大数据决策与传统的基于数据仓库的决策区别
12. 大数据、云计算和物联网（概念、区别与联系等）
13. HDFS和MapReduce的功能
14. HDFS中的名称节点和数据结点（包括功能）
15. HDFS的冗余数据保存策略
16. HDFS是探测错误发生以及恢复
17. HDFS的容错机制
18. HDFS在正常情况下读\写文件过程
19. HBase和BigTable
20. HBase和传统关系数据库
21. HBase中行键、列族和时间戳
22. HBase的概念视图和物理视图
23. HBase的数据分区机制
24. 批处理、实时处理概念，批处理与流处理区别
25. Nimbus进程和Supervisor进程
26. 在HBase的三层结构下，Region的大小、客户端访问数据
27. Region服务器向HDFS文件系统中读写数据的基本原理
28. HLog的工作原理
29. ETL过程概念及步骤
30. 在HBase中，每个region服务器维护一个HLog的优缺点
31. Region服务器意外终止情况的处理
32. 关系数据库和NoSQL数据库
33. 数据存储时的软状态、无状态和硬状态
34. 键值数据库、列族数据库、文档数据库和图数据库的适用场合和优缺点
35. JobTracker和TaskTracker的功能
36. MapReduce1.0体系结构的优缺点
37. HDFS 1.0的缺陷
38. Map函数和Reduce函数的输入、输出以及处理过程
39. MapReduce的工作流程（需包括提交任务、Map、Shuffle、Reduce的过程）
40. Map端和Reduce端的Shuffle过程（包括Spill、Sort、Merge、Fetch的过程）
41. Reduce的三阶段
42. BASE的具体含义
43. Combiner的灵活使用
44. MapReduce程序的中间结果存储在本地磁盘而不是HDFS上有何优缺点
45. 在YARN框架中执行一个MapReduce程序时，从提交到完成需要经历的具体步骤
46. HDFS中“块池”的概念，HDFS联邦中的一个名称节点失效会否影响为其他名称节点提供服务？
47. 单词统计实例
48. Spark的几个主要概念：RDD、DAG、阶段、分区、窄依赖、宽依赖
49. Spark对RDD的操作主要分为：行动和转换
50. Hadoop MapReduce的缺陷，并说明Spark具有哪些优点？
51. MapReduce的框架、小批量处理、以及流数据处理
52. 流计算流程与传统的数据处理流程
53. 单词统计：采用MapReduce框架与采用Storm框架区别？
54. Storm集群中的Master和Worker节点各自运行什么后台进程，这些进程又分别负责什么工作？
55. NewSQL数据库与传统的关系数据库以及NoSQL数据库的区别、常见的NoSQL数据库
56. Nimbus进程和Supervisor进程意外终止后，重启时是否能恢复到终止之前的状态，为什么？
57. 数据产生方式经历的阶段
58. 为何说MapReduce移动计算比移动数据更经济？本地计算

二、编程与综合题

Hbase创建表格、查询以及修改表格数值等；对json格式数据的处理，包括创建DataFrame、运用Scala语句的一些操作（筛选、排序、分组、查询等）；使用MapReduce编程技术对文档内容建立索引库、统计单词、合并文件、去除重复部分、以及创建表格（包括插入、求值等）；运用HBase的API创建表格、MySQL数据库创建数据库、创建表等。

三、平时做过的作业、实验等

表连接、统计单词出现的频数、HBase存储表并插入数据、