# 《大数据开发技术基础》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. 大数据决策与传统的基于数据仓库的决策
2. Hadoop特性
3. 常用的流计算框架
4. Flink、CAP
5. 图计算框架Pregel的应用
6. 大数据决策与传统的基于数据仓库的决策区别
7. 大数据、云计算和物联网
8. RDD操作、Zookeeper
9. HDFS中的名称节点和数据结点
10. HDFS的冗余数据保存策略
11. HDFS是探测错误发生以及恢复
12. HDFS在正常情况下读\写文件过程
13. HBase和BigTable
14. HBase和传统关系数据库
15. HBase中行键、列族和时间戳
16. HBase的概念视图和物理视图
17. HBase的数据分区机制
18. Nimbus进程和Supervisor进程
19. 在HBase的三层结构下，客户端访问数据
20. Region服务器向HDFS文件系统中读写数据的基本原理
21. HLog的工作原理
22. 在HBase中，每个region服务器维护一个HLog的优缺点
23. Region服务器意外终止情况的处理
24. 关系数据库和NoSQL数据库
25. 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
26. JobTracker和TaskTracker的功能
27. Map函数和Reduce函数的输入、输出以及处理过程
28. MapReduce的工作流程（需包括提交任务、Map、Shuffle、Reduce的过程）
29. Map端和Reduce端的Shuffle过程（包括Spill、Sort、Merge、Fetch的过程）
30. 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
31. 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
32. 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？
33. 单词统计实例
34. Spark的几个主要概念：RDD、DAG、阶段、分区、窄依赖、宽依赖
35. Spark对RDD的操作主要分为：行动和转换
36. Hadoop MapReduce的缺陷，并说明Spark具有哪些优点？
37. MapReduce的框架、小批量处理、以及流数据处理
38. 流计算流程与传统的数据处理流程
39. 单词统计：采用MapReduce框架与采用Storm框架区别？
40. 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？
41. NewSQL数据库与传统的关系数据库以及NoSQL数据库的区别
42. 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？
43. 数据产生方式经历的阶段
44. 不一致性窗口
45. MapReduce Job与Storm Topology
46. Spark on YARN

二、编程与综合题

使用Java API对HDFS中指定文件操作（包括读写权限、大小、创建时间、路径等信息，以及创建目录，创建，删除文件）、使用Spark通过JDBC连接数据库MySQL、使用MapReduce编程技术对文档内容建立索引库、统计单词、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