技术之面试题复习

版本：V1.6

# 第1章 技术复习

## 1.1 第1次考试（准备20分钟，考试30分钟）

**1）Linux常用高级命令**

**2）HDFS读写流程**

**3）HDFS小文件危害及解决办法**

## 1.2 第2次考试（准备15分钟，考试20分钟）

**1）Shuffle及其优化**

**2）Yarn工作机制**

**3）Yarn中各个调度器特点及生产环境中怎么选择**

**4）Zookeeper非第一次选举机制**

**5）Zookeeper符合CAP法则中哪两个**

## 1.3 第3次考试（准备20分钟，考试25分钟）

**1）解释一下零点漂移产生的原因及解决办法（Flume）**

**2）Kafka生产者发消息流程**

**3）Kafka的Broker工作流程**

**4）Kafka的消费者组消费流程**

## 1.4 第4次考试（准备20分钟，考试25分钟）

**1）Kafka挂了如何处理**

**2）Kafka怎么保证数据不丢**

**3）Kafka数据重复如何处理**

**4）Kafka数据积压如何处理**

**5）Kafka如何保证数据有序or怎么解决乱序**

## 1.5 第5次考试（准备10分钟，考试25分钟）

**1）Kafka怎么做到的高效读写**

**2）Kafka如何提高吞吐量**

**3）消费策略（Range、RoundRobin、粘性）**

## 1.6 第6次考试（准备20分钟，考试25分钟）

**1）Hive优化**

**2）Hive数据倾斜**

**3）怎么将Hive SQL转换为可以执行的MR**

## 1.7 第7次考试（准备20分钟，考试25分钟）

**1）MaxWell底层原理及为什么选择**

**2）MaxWell怎么产生的重复数据，如何解决**

**3）DataX在使用过程中遇到哪些问题，怎么解决的**

**4）Spark 转换算子10个**

**5）Spark 行动算子5个**

**6）Spark任务怎么切分**

## 1.8 第8次考试（准备20分钟，考试25分钟）

**1）Spark提交流程**

**2）SortShuffle原理**

**3）统一内存模型**

## 1.9 第9次考试（准备20分钟，考试15分钟）

**1）Flink的架构有哪些角色**

**2）Flink与Spark Streaming的区别**

**3）介绍一下时间语义，谈谈你对Watermark的理解**

**4）窗口的分类、划分、生命周期**

**5）Flink如何解决乱序问题**

## 1.10 第10次考试（准备15分钟，考试20分钟）

**1）介绍一下Flink的状态和状态后端**

**2）有没有状态比较大的作业，遇到过什么问题，怎么解决**

**3）Flink有没有遇到过反压，怎么发现、定位、分析、解决，效果如何**

**4）Flink有没有遇到过数据倾斜，怎么发现、定位、分析、解决，效果如何**

## 1.11 第11次考试（准备15分钟，考试25分钟）

**1）Flink如何保证数据精准一次**

**2）你是如何理解Flink的Checkpoint**

**3）Task的重启策略**

## 1.12 第12次考试（准备15分钟，考试20分钟）

**1）Flink有几种Join ，详细说明特点**

**2）你对keyby算子的理解**

**3）你对Interval Join的理解**

## 1.13 第13次考试（准备20分钟，考试20分钟）

**1）Flink的提交参数如何设置，怎么考虑的**

**2）Flink的提交流程**

**3）Flink的内存模型**

## 1.14 第14次考试（准备20分钟，考试20分钟）

**1）HBase存储结构**

**2）HBase的读、写流程**

**3）HBase热点问题如何解决**

## 1.15 第15次考试（准备20分钟，考试20分钟）

**1）Clickhouse的优势**

**2）Flink写入Clickhouse怎么保证一致性？**

**3）Clickhouse的常用引擎**

## 1.16 第16次考试（准备10分钟，考试20分钟）

**1）数仓项目建模准备（事实表 、维度表）**

**2）数仓项目建模（项目调研、数据域、业务矩阵、建模、指标体系建设等）**

# 第2章 项目复习

## 2.1 第17次考试（准备30分钟，考试1小时）

**1）离线数仓项目串讲**

## 2.2 第18次考试（准备30分钟，考试1小时）

**1）实时数仓项目串讲**