尚硅谷大数据技术之面试题复习

（尚硅谷研究院）

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