

大数据高薪攻略——

海哥一堂课胜读十年书

江湖人称：大海哥



尚硅谷

目录



一

大厂招聘特点

二

进入大厂前准备

三

赠送《大厂面试宝典》

四

总结答疑



大厂招聘特点

1.1 梦想中的大厂

1.2 大厂大数据技术梯队

1.3 大厂真题

1.4 大厂招聘特点归纳

呐喊吧！同学们！

打出你想去的大厂XXX



有目标的人在奔跑，没有目标的人在流浪！



- 第一梯队

- 头条、阿里、腾讯

- 第二梯队

- 华为、百度、美团、京东、新浪、

- 搜狐、360、VIVO、OPPO、58、顺丰等



Document



1) 笔试部分

2) 自我介绍

3) 技术部分

4) 项目部分

5) 手写部分

6) 算法部分

7) 场景部分

8) 学习习惯部分

9) 提问面试官部分

10) HR部分



进入大厂前准备

2.1 笔试部分

2.2 自我介绍

2.3 技术部分

2.4 项目部分

2.5 手写部分

2.6 算法部分

2.7 场景部分

2.8 学习习惯部分

2.9 提问面试官部分


2.10 HR部分



- 真题在手，天下我有
- 大胆的写出你的思路



- 大大方方的聊，放松
- 体现优势，避免劣势
- 思维逻辑清晰，表达流畅
- 不说前东家或者自己的坏话



模拟面试！
专业的就业老师！



- HDFS架构理解（基础）
- HDFS源码/工作原理（高级）
- MapReduce架构理解（基础）
- MapReduce源码/原理/Shuffle原理（高级）
- MapReduce二次排序（编程，可选）
- YARN架构理解（基础）
- YARN源码/工作原理（高级）
- HBase架构理解（基础）
- HBase源码/工作原理（高级）
- HBase性能优化（高级）

大数据框架：原理、优化

- Hive原理理解（基础）
- Hive性能优化（高级）
- Flume架构理解（可选）
- Kafka架构理解（可选）
- Spark RDD理解（基础）
- Spark reduceByKey与groupByKey区别
- Spark Broadcast与Accumulator功能
- Spark工作原理（高级）
- Spark shuffle原理（高级）
- Spark源码理解/贡献（高级）
- Spark性能优化/数据倾斜（高级）



➤ JVM架构/GC算法

➤ JUC多线程

➤ HashMap源码

➤ Redis底层原理/索引

➤ MySQL底层原理/索引优化

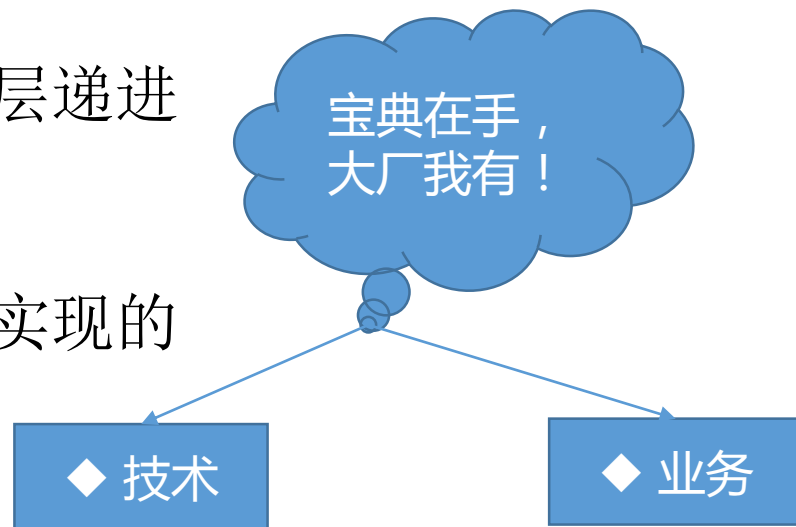
• Java框架：原理、优化





• 项目中遇到的问题，怎么解决的？

- 项目中遇到过哪些坑
- 你项目的亮点
- 常见业务难题的解决和优化，层层递进（头条追求做到极致）
- 工作中有什么指标很难的，怎么实现的





- 项目细节（集群规模、数据量、指标、具体实现等）
 - 集群规模
 - 问我一天的数据量有多大
 - 实时当天日活怎么累加
 - 问我Azkaban一天调度多少个任务
 - 哪个商品卖的好？每天卖多少？



数仓理论

- (1) 数据仓库分层架构（初级）
- (2) 事实表与维度表（初级）
- (3) 星型模型与雪花模型（初级）
- (4) 日增量表与日全量表（初级）
- (5) 拉链表（初级）
- (6) 缓慢变化维（初级）
- (7) 两种建模理论（初级）
- (8) data-vault理论（高级）

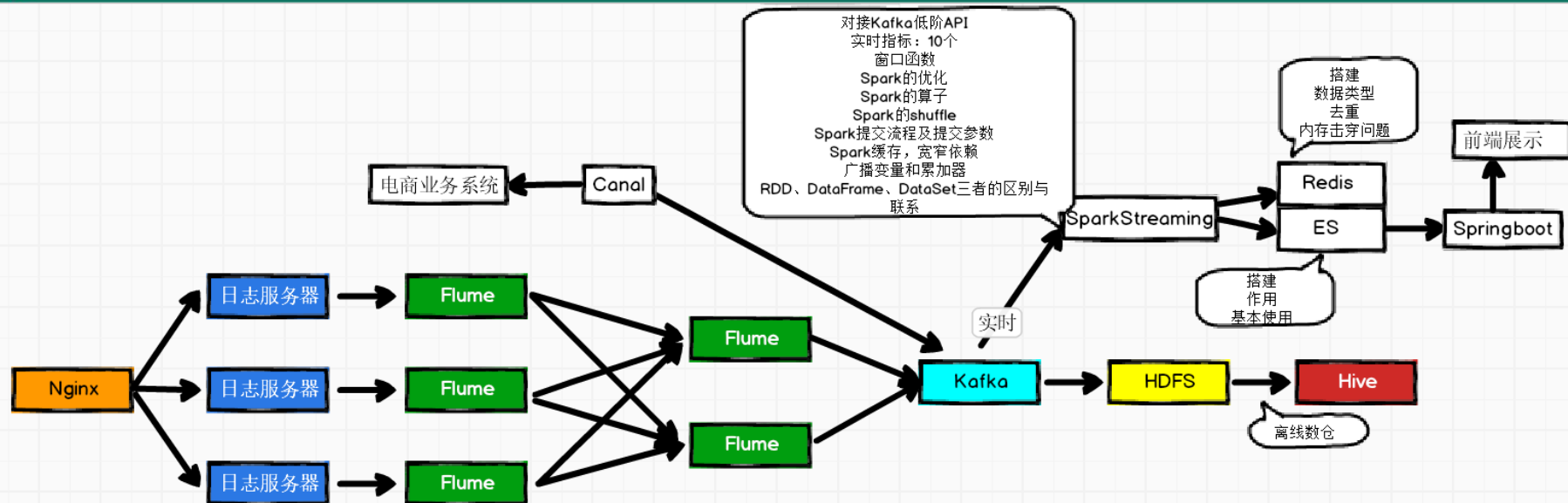
数仓周边系统

- (1) 血缘关系
- (2) 调度系统
- (3) 展示，olap（kylin）
- (4) 指标管理
- (5) 数据质量管理（高级）
- (6) 元数据管理，调度系统元数据，任务运行数据，表存储数据，数据字典等
- (7) 权限管理

数仓项目



2.4 项目部分



实现负载均衡

存放日志数据供
Flume采集
日志存放30天

1. TailDirSource实现断点续传和监控多文件的功能，
2. FileChannel保证数据的安全
3. 采用AvroSink链接下一层的Flume
4. 定义拦截器进行轻度过滤和解析日志类型
5. 事务
6. 监控器Ganglia

1. AvroSource对接第一层Flume
2. KafkaSink将数据放到Kafka
3. 负载均衡
4. Flume会丢数据吗？
5. ...
6. ...
7. ...
8. ...
9. ...
10. ...

架构，机器数量，日志存放时间
硬盘大小，监控器，每日数据量
，副本数量，有几个Topic，
几个分区，会不会丢数据？
ISR，分区分配策略，
Kafka挂了怎么办？
数据重复问题？

HDFS读写流程，
Shuffle机制，
Hadoop的优化，
YARN调度器，
YARN的任务提交流程，
集群的搭建过程

Hive的架构，内部标外部表，四个By，窗口函数，
系统时间函数，Hive的优化

数仓的输入源和输出目的地，数仓的分层，每层数据
每层之间的联系，离线指标：30个
表的分类及导入策略，
拉链表，维度表，事实表，多少张宽表
模型：雪花，星型，星座
自定义UDF，UDTF函数
Sqoop导入导出参数，Azkaban每天有多少job
秒级查询（Impala），报表可视化工具（echarts）
ETL工具（kettle），数仓命名规范
基本一个项目建一个库，表格个数为基础的原始数据表
格加上统计结果表格的总数。（一般70-80张表格）



- 手写代码

- 手写MapReduce的WordCount
- 手写Spark WordCount

- 手写设计模式

- 双端检测单例、工厂、代理、装饰模式

- 手写算法

- 详见算法部分

- 手写场景

- 详见场景部分



• 数据结构

数据结构必考，手写代码，每一面都会考。（今日头条）

老韩带你学
数据结构

- 用IDEA写快速排序
- 快排的时间空间复杂度？快排原理
- 手写归并排序
- 二叉树的前中后序遍历？
- 链表转置/二叉树转置
- 单向链表反转
- 手写二分查找
- 字符串反转
- 冒泡的时间空间复杂度？原理
- 实现堆栈Push Pop Min 复杂度O（1）



- LeetCode

多刷Leetcode，题都是有套路的（今日头条）

清华硕士带你刷题！

- 1. 两数之和
- 2. 爬楼梯
- 3. 翻转二叉树
- 4. 反转链表
- 5. LRU缓存机制
- 6. 最长回文子串
- 7. 有效的括号
- 8. 数组中的第K个最大元素
- 9. 实现 Trie (前缀树)
- 10. 编辑距离



- 写一个SQL将每个月的Top3取出来 我用了三个子查询做出来不行
- 最近七天连续三天活跃用户怎么实现的？手写一个各区域Top10商品统计程序？
- 三个字段，timestamp，user_id，product_id，让求pv最大的商品，写了之后又叫用Scala代码写一遍，然后又问，假如说只让你求pv数大于100的Top3呢，代码又怎么写
- 有一个分区表，表名T，字段qq，age，按天分区，让写出创建表的语句



- 看书

Hadoop专家、图解Spark、
Spark Streaming实时流式大数据处理实战、
基于Apache Flink的流处理、
Flink原理实战与性能优化等

- 看博客、写博客

CSDN、博客园、简书等

- 看官网、GitHub

Apache官网/GitHub



- 研究新技术

- Flink、Atlas、Griffin、Kylin、ClickHouse等

- 关注公众号

- 大海哥朋友圈，顶20个公众号



- 面试官：您还有什么想问我的吗？

- 这是体现个人眼界和层次的问题

- 参考答案

- ✓ 公司希望我入职后的3-6个月内，给公司解决什么样的问题
- ✓ 以你现在对我的了解，您觉得我需要多长时间融入公司？



- 你的优点是什么？


- 大胆的说出自己各个方面的优势和特长

- 你的缺点是什么？

- 不要谈自己真实问题；用“缺点”衬托自己的优点

- 你的离职原因是什么？

- 不说前东家坏话，哪怕被伤过
- 不要说因为加班太多
- 不要说超过1个以上的原因



专业的就业老师！



- 您对薪资的期望是多少？

- 非终面不深谈薪资，只说区间，不说具体数字
- 底线是不低于当前薪资
- 非要具体数字，区间取中间值，或者当前薪资的+20%
- 注意：和HR接触不要激动。会影响最终薪资。



赠送《大厂面试宝典》



• 新电商数仓项目

- 用户行为和业务表增加到40张表
- 分析的主题13个、指标100多个
- 增加数仓建模理论
- 增加元数据管理
- 增加数据质量
- 更新可视化框架
- 增加集群监控空间Zabbix，异常发钉钉



电商数仓指南

名企收割机！
培训小清华！



四

总结答疑

粉丝：谈恋爱吗？坐牢的那种？

宋宋：别闹宝贝，赶紧加关注！送海狗人参丸！

新来的宝贝们没点关注的请点一下关注，加关注，不迷路。

尚硅谷，www.atguigu.com

硅谷珊珊姐，qq：3408297627

硅谷蕾蕾姐，qq：1749553798



谢谢观看