2022年北京大数据面试题汇总

# 字节跳动

## 面1

1. 自我介绍
2. 做实时用flink ？用sql还是api
3. flink反压，如何排查
4. 算子链断链有哪些方式，从代码层面说
5. 在排查flink问题的时候有没有关注过flink的metrics 比如flink自带的metrics
6. checkpoint的流程，说说理解
7. 端到端一致性，如果挂掉了怎么办
8. flink写hive是否能保证端到端一致性？怎么保证幂等写入，overwrite 不是会覆盖吗？不会丢数据吗？
9. 遇到的数据倾斜的场景和对应解决方案
10. 场景题，字段：uid用户id，vid视频id，timestamp时间戳，计算每个视频的uv，一个用户可以看多次视频，由于有热门视频，会有数据倾斜，怎么用flink应用程序实现，追问，细问加随机数怎么加，继续追问更复杂的场景状态大怎么办。Split Distinct的底层原理是什么？Id-key比较分散，
11. 补充：如果rocksDB放不开了怎么办，状态编程去重智能去掉3%左右，性能不太好，有没有更好的解决方案？
12. 常见的缩减状态的方式有哪些？
13. 旁路缓存是否能保证端到端一致性
14. Clickhouse的ReplacingMergeTree一定保证不重复吗？final和强制合并都不是可控的，还有什么好一点的方式吗？
15. Clickhouse的ReplacingMergeTree中依赖指定Order By （索引）和 版本的，如果不指定版本，在同一分片写入不同的数据，写入的时候保留那条数据？
16. Clickhouse有分片表和什么表的感念（没听清），对于ReplacingMergeTree如果我直接轮训写的话，是否能达到去重效果吗？
17. 介绍实时数仓项目？最印象深刻的点是什么？
18. 大状态如何能够复用？增量检查点启动恢复的时间是很久的，业务上是否能接受？
19. 如何理解实时离线一体架构？
20. 市面常见数据湖的优劣对比了解吗？

## 面2

1. 怎么保证flink的一致性，两阶段提交，怎么保证数据不重复，最后解决了吗。
2. clickhouse 怎么保证幂等性
3. sink端的事务输出 输出到哪些组件
4. kafka和clickhouse各自的应用方向都是哪些
5. clickhouse为什么查询快
6. clickhouse适应哪些业务场景
7. 一致性数据只是做到了数据的不丢失，怎么保证数据的准确性？
8. 你认为怎么去做才能保证数据的正确性
9. 实时数据最后是都有谁去使用了
10. 数仓怎么去做分层的，横向怎么做分数据域的
11. 你们的数据域是怎么划分的，都有哪些，详细说每个域具体是怎么去定义的，都怎么去定义这些数据范围
12. 整体分5层的优缺点都是什么
13. 数仓建模的方法都有哪些
14. 维度模型各自的特点和不同有哪些
15. 事实表的设计方法是什么
16. 事实表分为哪些
17. 对于缓慢变化的数据事实表采用什么方式存储
18. 拉链表的好处和缺点都是什么
19. sparkrdd怎么体现他的弹性，和特征
20. transform 和 action 算子有什么不同
21. spark的shuffer有哪些，讲讲各自的特点
22. 写sql 电脑上直接敲：
23. 用户连续日活表usr\_login，表中字段：userid，dt，目前该表存有1-30号内的数据，
24. 请输出30天内，连续登录三天及以上的userid，以及用户初始活跃的第一天

## 面3

1. 大数据这一段工作经验中你做的是那部分东西呢？你觉得最拿手的一个方向？
2. 你们是做什么业务的？
3. 你们有事业部概念吗？你是在哪个事业部？你负责的业务范围是什么？你在里面的角色是啥？
4. 实时的核心指标有什么？
5. M1到M7核心指标是什么？
6. 数据在ODS层，你整个实时数仓是怎么流转的，怎么计算加工的？
7. ODS层的数据都用了什么数据？数据源都用的什么数据？数据里包含什么？DWD业务明细都有什么数据？DWD中数据域都有哪些，主题域都有哪些，怎么分的，为什么这么分？
8. DWS层都有什么粒度的表？
9. DWS层提到旁路缓存和异步IO，着重讲讲异步IO？异步IO代码逻辑，用什么算子，哪些方法，里面怎么做的，和HBase的交互？
10. 实时维表是否是实时的？Redis和Hbase的一致性怎么保证？Redis的失效时间是多久？
11. 维表数据改变更新Hbase并删除redis的操作是原子的吗？线上的flink会访问数据吗？不是原子性如果产生错误数据怎么修复？
12. 你们线上容错机制都有哪些？营收这部分你们怎么做容错？
13. 类似金额，营收类统计，金银分析类东西，假如出错了，怎么做修复以及线上容错？
14. 容错方案，可以从哪几个角度来想？
15. 整个流式数据仓库的稳定性和质量怎么保证？数据怎么保证正确并且容灾？
16. 你们有实时仓库的数据监控和任务状态的运行监控吗？比如任务健康度以及数据质量相关的各个维度的监控以及异常的告警，有没有，怎么做的，没有的话，知道怎么做吗？
17. 一个算法题，用对列实现。

## 面4

1. 自我介绍
2. 问了问数据量
3. 介绍一个你参与最多的项目
4. hdfs读写流程
5. hdfs的shuffle/shuffle过程有几次排序
6. kafka乱序问题
7. kafka的acks
8. kafka的isr
9. 用户行为数据怎么划分的主题
10. 业务数据有哪些信息和指标
11. 实时用的flink吗？为什么不用spark？
12. flink的检查点怎么实现
13. java和python了解吗，主要是hive吗？
14. 连续两天登录SQL

## 面5

1. 自我介绍
2. 讲一个自己比较熟悉的系统
3. 维度数据具体是怎么跟事实数据进行关联的
4. dws层写出数据到clickhouse是多长时间一次
5. 怎么验证实时数仓指标的正确性
6. hdfs的读写流程，读写过程中有datanode节点挂掉怎么办
7. 你们那些指标是使用api写的，那些是通过sql写的 为什么要分开

## 面6

1. leetcode 162 要求二分法实现
2. kafka高效读写
3. 项目整体如何控制重复数据，哪里用的事务，哪里用的幂等
4. flink barrier的机制，如果换你来实现，如何优化对齐式barrier带来的背压问题
5. flink的内存模型
6. java的内存模型
7. 异步关联纬度，为啥加redis，不加redis可以不，在应用程序里用第三方缓存接口是否更好
8. java中array和list的区别
9. hadoop为什么默认block128m
10. hashmap的实现

## 面7

1. Clickhouse 和elasticsearch 的区别，应用场景的区别，底层的区别
2. arraylist 和linklist 有什么区别，读写上的区别
3. 创建多线程的几种方式
4. Udf,udaf,udtf,的区别，各举几个例子
5. kafka 数据是有序的吗？
6. kafka应答级别
7. 谈一下isr
8. flink 怎么保证数据不丢失
9. flink checkpoint 底层是通过什么实现的
10. 手写连续两天登陆，手写连续7天登陆

## 面8

1. 自我介绍
2. 业务表有哪些内容
3. 日志数据有哪些数据
4. 实时的ADS指标是交给哪些部门使用的，怎么使用的？
5. Clickhouse和doris的区别
6. Kafka的乱序问题
7. Kafka的isr了解吗
8. flink的状态后端说一下
9. spark的stage怎么划分的
10. distinct算子是宽依赖还是窄依赖
11. reducebykey和groupbykey的区别
12. flink的算子状态和键控状态
13. flink的端到端一致性
14. Kafka的偏移量保存在哪里
15. flink的偏移量保存在哪里
16. flink的Operato state和keyed state区别
17. 项目中使用sql多还是api多
18. sql：求连续三天登录的用户，用户一天之内可以重复登录
19. 先说思路，再写sql。
20. 算法题：三角形最长周长

题目描述

无序的数组: arr= [3, 7, 4, 5, 20]

任意选三个数作为三角形的三个边长，求能够组成三角形的最长周长

时间复杂度nlogn

数组排序

最大三个数，判断是否前两树和大于第三个

往前推移

## 面9

1. 说一下在项目中的工作
2. 说一下业务总线矩阵是怎么构建的
3. 说一下维度怎么构建的
4. 结合实际的场景说一下下钻，汇总，是怎么样的一个需求，在构建业务模型的时候是怎么满足这些需求的
5. 你们统计的周期是1天，七天，一个月，所以你们的周期是固定的吗，如果想要实现动态查看比如2天，三天怎么实现
6. clickhouse的索引，存储，怎么处理宽表，底层如何压缩的，clickouse的存在那些问题
7. flink的反压，数据倾斜怎么处理，怎么发现
8. 一道sql题，分组topN的sql题
9. 你有什么想问的吗
10. 大部分是业务方面的问题，建了哪些表，如何思考的，从哪些角度分析等等

## 面10

1. 自我介绍
2. LC 611 有效三角形的个数
3. SQL 连续四天登录
4. 介绍一下公司的业务
5. 整个公司的数据仓库都是你们那个团队搭的吧？
6. 日常使用数仓最多的人
7. 除了给管理层看的报表和面向内部研发监控数据质量的，还有没有别的场景？
8. 有没有分析师
9. 他们会有个性化的指标找你提吗？
10. 指标会落成你长期要看的东西吗？还是临时看一下
11. 除了管理指标的搭建数据还发挥了什么价值吗？
12. 用户画像是给谁用的？
13. 打广告、外部推广具体怎么用这些数据？
14. 介绍一下离线数仓如何搭建
15. 拉链的开始结束时间代表什么？
16. 用户频繁变化的维度或属性有哪些
17. 做拉链表的意义是什么？
18. 结合业务如何进行的维度建模设计
19. 事实表的具体构建过程
20. 实时数据从最原始的日志大概量级是多少？
21. 比较难以解决的问题？
22. 一致性问题是怎么解决的？
23. 你有什么要问的？
24. 咱们是属于抖音事业部的吧？属于，国内产品都合并为抖音事业部了
25. 咱们公司有多少服务器？数据量多少？日活6个亿，公司有专门做服务器采购的，服务于不同的部门和研发团队，不是专门做这个的应该不清楚
26. 做实时的有多少人？有很多团队，这个团队五六个人
27. 咱们团队主要做哪个方向？
28. 咱们公司有没有使用什么前沿的技术？比如数据湖、流批一体之类的？

## 面11

1. 介绍Flink项目
2. 为什么不用Flink Mem取代Redis
3. 维度数据量多大，热数据比例多少
4. 火焰图怎么定位反压
5. 反压的原理，反压怎么传递给上游算子的，Flink算子间数据怎么传递的
6. 解释一下数据怎么分层的
7. 解释一下minibatch
8. 集群规模，Flink几台，怎么配置的
9. spark join实现原理
10. 实时和离线资源冲突怎么应对的
11. java HashMap源码
12. 解释头插和尾插
13. 实现遍历一次链表就完成反转
14. sql连续7天登录
15. java了解多少
16. gc了解多少
17. ArrayList和Linklist怎么理解的
18. Clickhouse存数据流程

## 面12

1. 数据的链路
2. 为什么要做实时
3. 怎么保证实时数据的准确性
4. 误差一般是多少
5. kafka动态分区
6. flink中的数据倾斜
7. 离线中数据倾斜的监控，比yarn更加具体点的
8. 用户画像大概做了什么
9. 离线数仓的中间层有什么用
10. 为什么要做中间层
11. dwd表有几张
12. dwd,dim,dws怎么建立的
13. hbase在那里使用的
14. hbase里面rowkey的设计原则
15. 一个字符串(12345)把奇数放在前面把偶数放在后面,不能改变原本的顺序(13524);
16. 求商场销售总额的前10%的用户id
17. 一个uid，一个time(时间戳),一个ip。在15分钟内保证uid的ip不重复。

# 华为

## 面1

1. hbase 的rowkey设计原则
2. 讲一下kafka
3. 讲一下 zookeeper 以及选举机制
4. spark和hadoop的异同
5. 讲一下mapreduce
6. 讲一下为什么离职
7. 说一下spark 的算子
8. spark 的行动算子和其他算子有什么不同
9. map和mappartiton 的区别
10. udf udtf用过吗
11. map算子和flatmap的用法

## 面2

1. 自我介绍
2. 介绍实时、离线、画像项目
3. hdfs磁盘均衡
4. hbase冷热数据，rowkey和预分区
5. hive优化
6. flink反压倾斜
7. zk选举机制
8. hive、hbase、redis各自的优缺点、适应场景
9. 仨项目都做了哪些优化

## 面3（华为云）

1. 实时项目负责内容
2. 日志数据分流，写入kafka，发送的的流量多大，是一个flink app就完成了嘛
3. flink部署方式standalone、yarn
4. TaskManager几个并发，几核；整个任务的并发
5. 遇到过性能问题嘛，处理的慢或者写的慢，中间处理的过程，整个链路上的性能瓶颈问题
6. 反压在哪个阶段，怎么解决
7. Kafka高效读写原理，写入partition是顺序写嘛？partition映射多少文件？
8. 零拷贝：应用层是不处理嘛？什么时候落盘的？

# 新华三

1. 介绍最近做过的项目
2. 简单问了问架构设计
3. java笔试题（15道单选，几道问答，3道程序）
4. yarn提交流程
5. 项目中遇到过哪些问题，是怎么解决的
6. 数据倾斜问题

# 美团

1. 自我介绍
2. 你们维表有多大量？是否只是点查，不会范围查询或者多维度组合查询吗？
3. 有可能出现redis容量超过上限的情况吗？
4. 新增数据源或者数据格式发生变更，会停掉服务吗？
5. 大状态怎么解决？
6. 检查点怎么做对齐？
7. 两个流join时间差别比较大的化会有什么影响？
8. 实时数仓是否做过监控，稳定性的考虑？
9. 反压怎么解决，数据倾斜怎么解决？
10. 怎么计算程序所需要的资源？
11. 哪里会用到zookeeper？
12. kafka的isr和ack是什么？
13. 我有5个副本，isr只有一个会有什么影响？
14. 编程题：

数字字符串转化成IP地址

现在有一个只包含数字的字符串，将该字符串转化成IP地址的形式，返回所有可能的情况。

例如：

给出的字符串为"25525522135",

返回["255.255.22.135", "255.255.221.35"]. (顺序没有关系)

数据范围：字符串长度 0 \le n \le 120≤n≤12

要求：空间复杂度 O(n!)O(n!),时间复杂度 O(n!)O(n!)

注意：ip地址是由四段数字组成的数字序列，格式如 "x.x.x.x"，其中 x 的范围应当是 [0,255]。

**示例1**

输入： "25525522135"

输出： ["255.255.22.135","255.255.221.35"]

**示例2**

输入： "1111"

输出： ["1.1.1.1"]

**示例3**

输入： "000256"

输出： []

# 百度

## 面1

1. 介绍项目
2. gzip压缩比 10%
3. 数据倾斜
4. 累积快照事实表怎么解决累积快照事实表
5. spark怎么获得一个rdd
6. datafream可以转换成rdd吗？
7. SMB join原理
8. 广播变量原理
9. spark内存模型
10. flink内存模型
11. shuffle数据量比较大怎么调节
12. flink的shuffle分为几类
13. Kafka怎么保证数据不丢
14. kafka的leader和follower之间的数据同步
15. leader挂了怎么办，机制是什么
16. XW是啥？
17. mr的shuffle
18. mr的shuffle优化
19. 算法已知数组，1，2，3，4，5，6，7

求所有两数之和等于10，数字的下标。算法复杂度要低

## 面2

1. 数仓的概念
2. 业务域怎么分的？
3. 关于主题的说法，如何划分，
4. ADS的表如何分类，提交什么需求就叫什么表？
5. 各层有多少张表，数据量。数仓跑了多久
6. hadoop的基本原理。hadoop3跑的啥，tez还是spark
7. RDD的基本原理，弹性怎么理解
8. 宽依赖和窄依赖
9. spark数据倾斜的问题
10. SQL题1

结果表： 客服名称，客户数 ,未完成对话数， 平均响应时间

msg\_tab： from\_id, to\_id, msg\_time, msg

server\_tab server\_id,server\_name

消息表：发送消息的ID，接收消息的ID，消息的时间，消息内容

客服 和 客户都可以发送消息，接收消息

客服表：客服ID，客服名称

统计指标：统计每一个客服：客户数：服务的客户数

未完成对话数：客户发送消息，客服没有回复的

平均响应时间：客户发送消息，客服多久回复 的平均值

1. SQL题2

2-8定律，20%的人手里掌握着80%的财富

场景：一张表有两个字段，user\_Id，money ，统计掌握80%财富的 那20%的人

如何统计？

1. 对大数据岗位的理解，职业规划

## 面3

1. 自我介绍
2. 介绍一下实时
3. 介绍下你们的异步IO和旁路缓存介绍数据一致性离线实时
4. 架构是一套吗
5. 遇到的数据倾斜
6. 存储用的是什么
7. 了解多线程吗
8. 算法：整数数组个其中一个是单一的 剩下的都是两个如[9,9,1,2,2,3,3] 求出这个数是多少
9. 有什么问我的

## 面4

1. 介绍项目
2. hadoop写流程
3. shuffle讲解，几次归并？几次排序？
4. spark的广播变量，广播变量的优势？
5. 用户画像的标签分类？性别可以定义为属性标签吗？
6. 用Java写一下归并排序

## 面5

1. 自我介绍
2. 项目介绍
3. spark的使用场景
4. hadoop高可用 hdfs的启动进程
5. hive是怎么集成spark的
6. hive读取元数据
7. 多线程
8. java 高效读写用什么数据集
9. 每天跑的数据量
10. 集群规模

# 携程

1. 介绍实时数仓
2. 维度数据为什么放到Hbase
3. dws关联hbase维度表数据的意义是什么？最终输出什么？
4. 数据以什么样的形式、频率导入Hbase
5. 热数据放入redis，通过什么样的方式导入redis
6. dwd做了什么事情
7. flink挂了会怎么样？
8. flink数据乱序问题
9. 用户画像上游下游
10. 用户画像有多少标签？用了多长时间？
11. flink用了多长时间？一年（太长了？）
12. 用户画像的标签用到什么地方？说到了风险评估，追问什么样的风险评估？
13. AB测试了解的多吗？实验周期多长？你觉得实验周期跟什么有关？面试官还给解释了一下与样本量有关
14. 用flink写一个从新浪微博api去取数据，数据为posts（一条一条的数据间隙）需要做一个分词然后统计每一个词的次频，说的flinksql，又追问api方式怎么实现，
15. 在hive建模是用宽表形式吗？
16. 建模有没有用建模工具？

# 小米

1. 自我介绍
2. 你们的数据量大概有多少？
3. 实时数仓同一个操作产生的数据前端埋点的和第三方的数据时间戳不一致关联时怎么解决？
4. redis和Hbase的数据一致性怎么保证？变更高吗？
5. 你都用什么编程语言？我说java，场景题就来了

场景题：

广告抢夺更多的曝光机会，currentAds系统当前广告list的单价，candidateAds备选广告集的单价list，个数相同，当相同位置的单价高时，优先获得曝光机会，currentAds顺序不可变，重新组织candidateAds元素顺序，使得备选广告集整体获得更多曝光机会

即对于index = i, candidateAds[i] > currentAds[i] i获得曝光机会大

currentAds = [13,25,32,11]

candidateAds = [12,24,8,32]

candidateAds = [24,32,8,12]

# 58同城

1. 自我介绍
2. 离线数仓介绍，你是具体负责那一部分工作，你在工作中所认为的一些疑难点讲一下
3. 为什么分5层，对分5层的理解
4. 为什么用flume，为什么不用flink同步
5. 7天连续登录3天
6. 小文件怎么解决
7. 零点漂移怎么解决
8. shuffle里几次排序
9. udf，udaf，udtf用过吗
10. 实时：为什么用clickhouse，为什么clickhouse聚合处理快
11. 反压问题讲一下，怎么处理

# 快手

## 面1（一面）

1. 自我介绍，为什么离职，有offer吗，想做哪一块
2. 你们有几个人，分工怎么样，都和哪些人对接，数据量多大？
3. 你们平常怎么和其他产品和业务人员打交道的？如果没有产品你怎么对接业务人员？
4. 如果一个陌生领域你怎么快速熟悉业务？
5. 你们做的数据产生什么价值怎么评估？
6. 数据准确性的保障是怎么做的？是用的工具还是自研的？用什么呈现
7. 离线实时的模型设计？星形模型和雪花模型的区别？
8. 有没有看过直播，如果你来设计直播的模型，你会怎么设计数仓？
9. MR的shuffle流程
10. 快排原理？最好和最坏的时间复杂度
11. reduce个数是由什么决定的？

文本, 信件

描述已自动生成

1. 不指定分区默认是几个分区？
2. udf，udaf，udtf有什么区别？
3. 两个大表join产生数据倾斜怎么优化？
4. flink怎么保证端到端的一致性
5. flink反压是怎么样的，怎么优化
6. clickhouse解决什么问题？
7. SQL题：同时在线人数
8. SQL题：连续登录四天有哪些人，连续登录几天？
9. 数仓这块怎么挖掘更大的数据价值？怎么应用？

## 面2 （二面）

1. 介绍他们的诉求
2. 介绍我这边的项目
3. 能不能讲讲某个业务方向上是数据是解决什么问题，怎么发现的问题，又怎么解决的问题，怎么验证他的准确性，怎么去评价
4. 有没有技术上比较有挑战的项目讲一下
5. 了解客户端埋点吗？埋点这边的规范和设计咱们理解吗？怎么是好的数据埋点规范
6. 怎么是好的数据研发？
7. 咱们之前有哪些之前做的差一些，时间回溯又想要怎么提高的？
8. 如果产品的需求你觉得不靠谱，你怎么办？
9. 怎么带实习生？目标是什么？

## 面3

1. 自我介绍。
2. 讲讲实时项目。
3. Flink端到端一致性的保证。checkpoint那些。
4. CLickhouse特点，如何调优。

Sql逻辑，explain select \* from tableA;

Explain ast select \* from tableA;

建表优化：能不用string就不用string

Datetime，转化long类型，不利于性能和效率

一亿条数据单表，建议分区数 20-30，这样效率是最快的

1. 多个班级的总分数前五名sql题。
2. 互相关注的人数sql题。
3. （7）sum() over(partition by ) sum() over(partition by order by )这两个的区别。
4. （8）hive遇到过哪些问题，如何调优。
5. （9）实时有没有遇到数据倾斜，如何解决。
6. （10）加盐表怎么设计。
7. （11）数据打散的方式有哪些。
8. 散列性：hash，加盐，字符串反转
9. （12）你平时常用的sql函数有哪些？

# 度小满

1. 你们Redis用了几台服务器，我搞了6台
2. Redis中每个服务器存储的数据怎么划分的
3. Redis的RDB和AOF
4. MySQL索引优化
5. Flink检查点机制
6. 二分查找算法

# 作业帮

1. 面试官介绍了一下他们公司要做的项目，偏离线；
2. 自我介绍；
3. 项目架构；
4. 问离线项目中你做了哪些事情；
5. 问离线数仓怎么可视化的；
6. 问flink的检查点一致性；
7. sql面试，连续7天内商家评分连续下降；

# 京东物流

1. 自我介绍
2. 怎么从老师转向开发？
3. 不是本专业，又做了老师，还学了开发，可以总结一下你的学习的心得体会吗？
4. 介绍离线数仓项目，你在项目中做了什么？
5. 各层建模参与吗？
6. 数据域是怎么划分的？
7. hive sql优化
8. 遇到过数据倾斜吗？
9. 用过拉链表吗？
10. 相比离线，你觉得实时的难点在哪？你觉得离线、实时你更擅长哪个？

# 闲徕互娱

## 面1

1. 问了个foreach的累加scala代码
2. kafka系统主题了解吗
3. 消费者和消费者组的区别
4. 主题 主题域
5. 数据倾斜，mapjoin如何实现的？skewindata加随机数怎么实现的？
6. hive存储格式有哪些？为什么用orc，不用parquet？
7. kafka数据丢失，数据重复
8. 你们公司80g数据能遇到数据积压问题吗？
9. 表名：dau ，字段： user\_id, day 需求：2022-08-01 活跃的用户在之后每一天的留存数
10. 连续两天登录
11. 一道行转列sql

## 面2

1. 表名如何区分
2. kafka 的分区
3. kafka索引文件 ….index
4. kafka数据有序
5. kafka的系统有那些主题
6. kafka数据重复的原因消费者组中的消费者数量超过分区数怎么办？
7. flink的checkpoint 和savePoint 有什么区别
8. flink故障恢复的机制有那些？重启策略
9. flink 底层分区的模式 和 slot
10. flinkDAG的数据结构
11. flink有迟到数据怎么办？
12. flink乱序
13. flink状态的更改，状态就是个表，如果要添加字段怎么做？
14. 拉链表
15. 表名：dau ，字段： user\_id, day 需求：2022-06-01 活跃的用户在之后每一天的留存数
16. 连续两天登录 用户分组开窗
17. hive中的map函数

# 神州数码

## 面1

1. 自我介绍
2. 分区表分桶表按什么分区，按什么分桶？ 按时间分区
3. 怎么数仓建模的？
4. 他追问没用什么拉链表什么的？ 就用户表用了
5. 增量表与全量表？
6. 讲一下flink和kafka？
7. 经常用的是hive还是关系型数据库
8. hive，索引where不到是怎么回事？
9. 是否离职，离职原因
10. 本科专业，为什么干这个？
11. 有什么要问我的？

## 面2

1. 自我介绍
2. 问我是否会接口开发，会不会js开发
3. hive的内部表和外部表区别
4. 内部表的分区分桶有了解过吗？分桶一般使用什么字段来分效率高一点，分桶个数如何确定；分区表有哪几种分区表；使用哪种类型的分区效率比较高
5. 有没有用过hive索引，分区自带索引了解过吗？什么时候索引会失效？
6. 多条数据重复，怎么去重
7. 如果开发SQL脚本的时候，如果跑的很慢，首先怎么定位，然后怎么解决；其它层面的调优，和仅限SQL层面的调优
8. SQL的表与表关联的底层运算机制
9. 介绍一下数仓项目，怎么分层，主题划分怎么考虑的，数仓建模
10. 期望薪资多少

## 面3：

1. 问离职原因
2. 一面我提过数据湖，这次问我的理解
3. 能不能考虑出差，是否排斥和各个业务方的沟通
4. 给我介绍了他们公司发展历程和目前规划

## 面4：

1. 自我介绍
2. 实时项目介绍
3. 业务数据那几块？数据复用的地方？
4. GMV是SQL还是算子计算，状态怎么存的
5. rocksDB 状态
6. 生产中遇到反压吗
7. flink数据倾斜，我没答上来
8. 乱序，watermark
9. 实时数仓有多少主题？业务和日志分别有多少？
10. 订单金额求过吗
11. checkpoint机制
12. 两阶段提交
13. hbase的rowkey怎么设计
14. Kafka数据保存几天
15. 问离线的hive这块是不是有同样的指标，是拉姆达架构还是kafka架构
16. 指标的准确性怎么保证
17. 有什么要问的

# 中国邮政

## 面1

1. 项目用的哪些组件
2. 整体架构
3. 怎么建模
4. Hql优化
5. Flink数据一致性
6. CK替换replacingmergetree引擎 处理数据去重的延迟问题？
7. 使用框架时候怎么处理漏洞，类似一些网络安全问题
8. 框架的bug都有啥，怎么处理的

## 面2（外包）

1. 自我介绍
2. 在上家公司负责什么
3. 用到哪些技术，是怎么实现的 我说的hive
4. Hive优化
5. Hadoop集群多大
6. Hadoop遇到的问题
7. Hadoop的基础架构
8. Mr工作流程
9. 数据倾斜怎么处理
10. 怎么分区分桶
11. 你有什么想问的吗

# 人民在线

## 面1

1. 自我介绍，然后详细讲一个项目
2. 拉链表，零点漂移，
3. 行转列，列转行详细解释
4. 说说springboot常用注解
5. 讲讲flink的窗口，和join方式
6. 列举多线程安全和不安全的几个类
7. 多线程怎么实现的，用的那几个方法
8. spark reducebykey groupbykey 区别
9. 集群规模，每个节点配置，数据量

## 面2

1. 实时架构，离线架构
2. hbase 行键底层如何排序
3. flink join 大概介绍一下，几种
4. hashmap 底层如何实现
5. treeset和hashset区别
6. 列举多线程安全和不安全的几个类
7. 多线程怎么实现的，用的那几个方法
8. es大概介绍下，有部署过吗
9. flink 端到端一致性
10. flume遇到过什么问题
11. hive小文件怎么解决的
12. spark reducebykey groupbykey 区别
13. flink遇到过什么问题
14. flink 的提交方式
15. hivesql 和mysql 语法区别
16. 用到过什么压缩
17. mysql隔离机制

## 面3

1. redis kv类型的 为什么要有hash
2. redis能实现 顺序消费吗
3. redis 操作需要加锁吗 多个客户端同时使用 能保证原子操作一致性吗
4. 用过哪些锁
5. flink 事件分为几种
6. 一个任务中可以既有事件窗口和处理时间窗口吗
7. flink 一个任务需要的资源怎么计算
8. Lambda 架构 优缺点 Kappa 架构 了解吗
9. mysql 一条sql执行步骤
10. hive有索引吗

## 面4

1. 自我介绍
2. 介绍一下项目
3. NN高可用，两个NN怎么同步数据，
4. 用过哪些设计模式，多线程的问题，线程安全问题，
5. 讲下SparkRDD,转换算子和行动算子介绍及区别，一个程序中有好几个行动算子程序运行流程是啥样的，集群规模和数据量，
6. flink端到端一致性，flink数据倾斜怎么发现的、怎么解决，flink状态，flink处理几亿条数据怎么去重、处理几十万条怎么去重，flink处理数据速度过慢怎么解决，
7. scala了解吗，
8. hiveSql大概都写多少行，写的比较难的sql介绍，flinkSql了解吗，flink监控的是任务指标还是集群指标，监控哪些指标，出现什么情况会报警。

## 面5

1. 线程安全，线程锁。加锁加在哪里。
2. 开发模式有哪些，了解哪些，常用的说说。
3. 离线和实时架构。
4. flume相关。
5. Flink遇到的问题。大状态，反压，数据倾斜……
6. 实际指标有哪些，出现了以上哪些问题，能不能详细讲讲。
7. 检查点原理。顺便端到端一致性讲了。
8. 算子和Flinksql，join等等。
9. Spark了解多少，scala了解多少，能不能看spark源码，能不能二次开发
10. 说说spark算子，spark sql那些。
11. 聊聊hive在离线数仓怎么应用的。主要就写sql？
12. 数据导入流程又讲了一遍。
13. 离职没，为啥离职？
14. 之前团队多少人？人员怎么分配的？服务器多少台？数据量多大？ 公司干啥的，你的指标算出来的结果一般多少？
15. 刚刚提到状态，问我Flink状态存多大。
16. flink任务怎么提交的，资源怎么分配的？
17. 有没有遇到过oom
18. 那再讲讲Flink CEP。
19. 问我薪资预期
20. 问我现在在哪，入职时间那些。
21. 让我问问题

## 面6

1. 自我介绍
2. 介绍实时项目
3. 在关联维度时，多流join的时候遇到过什么问题
4. Flink 精准一次
5. Flink taskManager、solt、task、subtask的关联
6. HiveSQL优化
7. 线程安全的Map结构有哪些
8. Groupby和ReduceBykey的区别
9. Kafka的isr机制

## 面7

1. 自我介绍
2. 简历上写的都会问，咋们大家都大差不差。常提的那些理论问题。把Flink相关的都差不多问了。还有版本，新特性。
3. 跟实际业务相关联的去讲哪些数据处理如何使用了哪些算子。每天数据量多少，大小，内存。
4. MR有没有遇到过没有只有Map没有Reduce的业务。详细说说。
5. redis介绍一下，数据类型有哪些，用过哪些，为什么用。
6. clickhouse介绍一下，特点有哪些，有哪些优化方式。
7. 如果用Flink去处理正常的流数据时，调用了一个API，但是API会有一个200ms左右的延迟，从全链路考虑，该如何优化。
8. java用的多吗，自我感觉如何。spark怎么用的，介绍一下。Shffle介绍一下
9. 高并发如何解决。
10. hdfs能不能并发写入？是如何解决并发问题的。
11. java的锁都用过哪些？阻塞机制是什么？特点和用法。
12. java的定时器如何使用。
13. 给自己本次面试打分。
14. 薪资想要多少。
15. 还有哪些问题。

# 中国移动

1. 纬度建模和分层
2. 数据采集，问咱们为啥用这些组件
3. 了解应用程序的开发流程吗？简单说一下
4. 数据治理有了解吗

# 嘀嗒出行

1. 自我介绍
2. 如何从老师跳到大数据开发的
3. 最近的重点项目，背景，做了什么，承担了什么，遇到问题是什么，困难是什么，挑战是什么，怎么解决的
4. 怎么保证clickhouse的数据一致性，优势是什么？
5. flink的并行度了解不？
6. 如果让你来管实时的所有项目有问题吗？
7. sparkSQL如何合并小文件？
8. 什么情况下会产生小文件？
9. hivesql的执行原理？
10. 我觉得你掌握的还可以，你是这些是工作中积累的，还是准备的？
11. 开发语言都用什么语言？
12. clickhouse写过自定义函数吗？flink写过吗？

# 科大讯飞

## 面1

1. 自我介绍
2. 项目履历
3. 你负责哪些工作
4. 参与建模了吗，你们模型怎么建立的
5. 技术架构，对flink的使用 （api,sql,cdc,cep）
6. 双流jion 哪几种类型？
7. 两个队列，每个队列都是海量数据，用什么join？
8. 用过哪些分布式存储框架？调优？
9. Kafka分区下线问题？
10. 服务开发？
11. 他介绍他项目

## 面2

1. 从三方面自述一下：个人介绍，项目介绍，技术介绍
2. Flume中Channel的选择器用的哪个？
3. DWS层数据分为几个主题库，大概是怎么划分的
4. 构建指标体系所需要的数据来源
5. 数仓的建模理论是什么？表格的设计思路
6. Spark用到的算子
7. 介绍Flink的状态，怎么使用的
8. 项目中用到SpringBoot开发了什么？

# 讯飞技术股份

## 技术面

1. 了解kafka吗
2. Yarn用的那种模式
3. Yarn的调度器用的那种 ，了解他们的区别吗
4. 工作的时候调过那些yarn的参数
5. 了解spark吗
6. Flink项目主要做了那些
7. 用过sparkstreaming吗
8. 工作时用过那些关系型数据库
9. 你对mysql了解吗
10. Mysql的索引，视图有了解吗
11. 你的维度是怎么确定的，能讲讲吗

# 合肥讯飞

1. 实时数仓的搭建过程？
2. 你们的实时元数据怎么管理的？
3. 你们有做主数据管理吗？就是里面最重要的ET实体是哪些？整个数仓最重要的实体是那些?
4. 指标体系建设是怎么做的？是产品经理直接梳理的吗？
5. 计算和加工表数据的时候主要用到的方式主要是flink sql吗还有别的吗？
6. 做过CEP规则引擎开发吗？
7. 对flink里面的状态使用过哪些？
8. 对hbase和hive这块做过调优吗，强调了下集群？
9. 遇到过热点问题吗？
10. 你在这个项目里面做了哪些部分？
11. 你们的技术选型是谁做的？
12. 你们的数据规模有多大？
13. 数据集成这部分是怎么实现的？
14. 后面的职业规划？
15. 你有哪些问题要问我？

# 我爱我家

## 初面：

1. 介绍自己
2. 介绍实时数仓
3. 开窗函数有哪些，排序那3个函数区别
4. 来公司后要学习python
5. 你有什么想问的吗

## 主管面：

1. 数据迟到怎么处理
2. 场景：有4个字段分别为 城市id，用户id，访问页面，访问时间，需求：从0点到现在，最高访问人数的一个峰值，每一分钟访问最高的前3名用户，一分钟内实现，具体到用的什么算子
3. 你懂算法吗
4. 来公司后需要很强的学习能力，需要学习算法，公司是flink结合算法来用的，会有人教，但是需要你能学明白，你觉得你的学习能力怎么样
5. 你的学习态度怎么样
6. 你还有什么想问的吗

# 中国电信

## 面1

1. 自我介绍
2. 数据采集，离线数仓 （组件的版本hadoop2.x 和3.x 的区别）
3. hive 的元数据里面有什么？
4. 维度建模每一层都干了什么？
5. 数据规模
6. hadoop的详细加载顺序
7. java arraylist 和 linklist 的区别
8. flink sql 调优（状态优化）
9. chlickhouse 的数据一致性 （实现不了 ，final ,optimized 实现）

## 面2

1. hashmap底层
2. 多线程
3. mapreduce原理及过程
4. hadoop和spaek的shuffle原理
5. flink的3种状态后端
6. kafka的分区策略
7. kafka高吞吐的原理

## 面3

1. 多线程怎么理解的
2. Flink背压定位解决，在哪发生
3. 维度关联时候发生反压
4. 数据倾斜
5. 快照说的状态太大
6. GC，流量洪峰不用处理
7. Flink消费Kafka的分区怎么理解的
8. Flink实现背压就重启的代码思路
9. Flink运行模式
10. Flink数据倾斜
11. Hive元数据存在哪
12. mysql是单节点吗
13. Hive脚本用的是jdbc吗，和客户端有啥区别
14. 有做权限管理吗
15. Hive有哪几个角色
16. 获取元数据的流程
17. 内外部表的区别
18. Hive小文件怎么产生的
19. Hive小文件的危害
20. Flume自定义拦截器的流程
21. 拦截器可以不用吗
22. Flume会丢数据吗
23. 给你个组件，比如Flink让你管理，能行吗
24. Flink怎么监控的，监控了哪些指标

## 面4

1. 自我介绍（我在中间提到了做版本选型）
2. 说一下怎么做的版本选型
3. 具体说一下你了解到的hadoop任意两个版本的区别
4. 讲一下纠删码原理
5. 集群中用到了什么脚本，手写任意一个
6. 说一下集群调优，主要结合实际业务场景说（讲了kafka）
7. 说一下hive或者spark的优化（选择的hive优化）
8. 看到了实时监控项目，具体讲一下其中的数据流转流程
9. 细说一下flink监控了哪些指标
10. java线程创建流程
11. GC垃圾回收
12. JVM常用命令
13. 手写冒泡排序

## 面5

二面：

1. 自我介绍
2. hdfs高可用切换流程
3. 集群整体监控及连接方式
4. 大数据CAP了解吗，Kafka的CAP符合哪几项？
5. kafka为啥吞吐量比较高
6. kafka的监控有没有做过，消费者积压数据怎么监控报警？
7. Spark的内存模型
8. Spark的shuffle的流程，优化，目前用到的shuffle是什么匹配流程
9. 小文件怎么避免
10. SparkSql怎么优化小文件
11. ArrayList和linkedlist的区别，工作中用那个多，ArrayList线程安全吗，怎么保证线程安全，除了加锁有没有其他的解决方法？
12. 多线程了解的多吗？
13. 集群规模，说一下版本和组件，hadoop的大版本用的那个，为什么选择3.X，参与搭建集群吗，集群的启动流程
14. hadoop相关的有哪些服务，加载哪些组件，加载流程是什么，NameNode加载哪些信息，启动过程中，元数据信息保存在哪里
15. flink做过哪些优化？(大状态的优化)，具体哪些需求用过这个优化？
16. Chickpoint状态失败了怎么恢复的，每个算子设置了id之后一定能正常恢复吗，有没有某个id下面是空的情况发生
17. 大的对象保存在哪里，这个问题双方没交流明白，尝试的沟通几次之后放弃了，直接问下一个？
18. 分布式快照原理是什么？
19. 数据是平稳过来的还是会有峰值？某个时间段峰值数据到来之后资源不够用了怎么办？怎么加资源（加内存）？
20. 具体怎么加内存，需要停止服务吗（需要停止，讲了为什么要停止），不停服务的话怎么操作，
21. 有没有考虑资源的动态调整（使用k8s）,k8s有了解过吗（没有用过，了解不深）？
22. hadoop3的纠删码了解过吗？默认是4+2模式吗(我答的3+2的模式，不确定对错)？自定义过吗？和3副本模式相比有哪些优劣呢？纠删码加速策略了解过吗？
23. 你的职业规划？
24. 你有什么问题呢？
25. 深入的了解了一下集群情况？关于自研的集群的版本、业务方向、安全资源？
26. 员工赋能培训有哪些？技术培训有吗？
27. 我的状态到公司熟悉业务大概需要多久呢？

## 面6 （二面）

1. 自我介绍
2. 你的技术栈中那个组件了解的最深？细讲一下组成？那部分是深入到源码了解的？
3. broker为什么依赖于zk，新版本有什么新特性呢？
4. 出现数据峰值问题，服务器顶不住应该怎么处理？那服役新节点可能遇到什么问题？
5. kafka的底层数据存在哪里，什么形式存储？
6. kafka有没有遇到脑裂问题？
7. 为什么这么深入的了解kafka，学习资料哪来的？下一个准备深入到底层的框架是哪个？
8. flink的算子结构分为哪几类？flink的一些细致问题忘记了，但都是大保健的拓展问题，见招拆招。
9. flink优化有哪些思路？flinkSql和FlinkAPI哪个用的多，为什么使用FlinkAPI进行开发
10. （10）使用FlinkCDC遇到过什么问题？
11. 讲一下hbase怎么优化？
12. hbase怎么设计rowkey？
13. 又给我介绍了一次公司的业务及集群环境。
14. 你还有什么问题吗？问了四面相关的内容。
15. 场景题：

有一个Bean

class Job{

Integer task;

String time;

}

列如：

1, "2022-10-31 01:12:32"

2, "2022-10-31 02:12:31"

1, "2022-10-31 01:12:32"

3, "2022-10-31 05:12:32"

2, "2022-10-31 02:12:32"

要求输出：

任务1 01时，点击数 2

任务2 02时，点击数 2

任务3 05时，点击数 1

java实现

## 面7

总结：java会涉及，平台框架问的多一些

1. 自我介绍
2. 框架用的什么？怎么分配的？hadoop几的版本？
3. 离线窜讲
4. Hive优化做啥了
5. 组件熟悉什么，我说hadoop，kafka，他问ZK
6. zk的CAP协议
7. 二分查找
8. 问了java链表和集合
9. 集群启动都干什么了？
10. 有运维吗？
11. 反问两个问题。

# 中盾安信

## 面1

这家主要做实时，应该也只会问实时

1. 自我介绍
2. 你们刚开始是使用了sparkstreaming，那后期为什么要更换为flink呢，你们对实时性的要求很高吗
3. 数据量，在你们业务当中flink在高峰期最大能够处理的数据量能够达到多少，具体是多少条
4. 每一条日志数据的大小是多少，日志数据当中具体都有哪些字段
5. 如果上游kafka接收到的数据量很多，而下游的flink只能有一个并行度，flink进行一些ETL操作能够处理数据的极限大概是每秒多少条？（反复追问，他说官网上有对应的参数）
6. 你们在使用redis做旁路缓存的时候，redis当中一共保存了多大的数据？冷数据和热数据是如何划分的，是否对数据打了标签？
7. 你们在使用flink对接kafka的时候是否遇到过什么问题，你们的dwd层是从kafka当中取数，然后再写回到kafka当中，那这样的话有没有遇到过kafka性能方面的问题，你们是怎么优化的
8. 是否针对于kafka集群进行过扩容和缩容？
9. 你是否了解过kafka都有哪些功能？不是底层原理层面的
10. 你们的flume搭了几台，每一台都是什么作用，是否针对于flume进行过一些配置
11. 你们的flume使用的是单点的吗，为什么不使用分布式的
12. flume的source，channel和sink，如果在选择channel的时候既想提高传输效率，也不想丢数据，应该如何配置
13. 你对一些nosql数据库有过了解吗，都有哪些使用场景，展开说说
14. 向他提问

## 面2

1. 介绍一下实时数仓
2. 遇关于Flume的问题了吗？
3. TaildirSource是怎么实现断点续传的？ 偏移量保持json文件
4. Kafka在什么场景出现乱序？怎么能处理的？ 维度变更 分区键
5. 你们的动态分流的具体实现？
6. 对线程池，连接池怎么理解？
7. 弄线程池是为了要做什么？
8. 怎么样批次提交事务？
9. 关于Flink的一些问题
10. Es怎么构建二级索引？ ？？ES不用建二级索引 一般es可以为别的数据库做二级索引hbase
11. 旁路缓存数据一致性问题你们怎么解决的？ 在写入数据时 通过kafka分区键 实现局部有序 同key的数据变化，串行，无并发，双写（内存+DB）
12. 对JVM有什么自己的理解？
13. 你们的数据量有多大？Kafka的数据量多大？每秒钟流处理数据量有多大？
14. 看你项目中介绍到了单例设计模式，想法是什么？遇到了什么问题？

# 计算所

1. 自我介绍
2. 讲一下项目
3. 项目详讲
4. 对那些框架比较熟悉
5. 项目中遇到过那些那些问题（数据倾斜、解决方案）
6. py会不会？
7. 写过哪些指标
8. 工作经验之类的问题

# 便利蜂

1. 你们宽表里面主要做了什么操作
2. 做了些什么指标
3. 有没有比较难的指标
4. Hive数据倾斜的现象和处理
5. 实时dws层做了什么操作
6. 维表的变化怎么传导到hbase和redis
7. flink双流join的原理是什么

# 人谷科技

## 面1

1. 自我介绍
2. 公司做什么的
3. mysql和hive区别
4. hive引擎为什么从MR换成spark，区别是什么，为什么spark快
5. hive优化
6. hive数据倾斜
7. 数仓建模怎么建的，指标体系是你分析的嘛，怎么分层的
8. ods，dwd，dim，dws，都干了什么，什么是维度数据，什么是事实数据
9. ods到dwd做了哪些处理，dwd，dws，dim都有什么区别
10. 你有什么想问的

## 面2

1. 自我介绍
2. 离线项目介绍
3. 项目中做过哪些调优？
4. 项目中都具体负责过什么？
5. order‍by和sort‍by？
6. udf，udtf，udaf怎么理解的？
7. 外部表和内部表区别？
8. 拉链表怎么理解，什么情况下会用，怎么实现？
9. rank，desne\_rank,row\_number区别？
10. 用过哪些系统函数？

## 面3

1. 自我介绍
2. 离线项目介绍
3. 你负责了哪些主题‍能讲讲dws层你建了哪些表‍为什么要建这些表
4. 数据导入到ods层你们是怎么做的
5. 拉链表的使用场景以及大概讲解一下步骤
6. 问了hive的‍UDF‍UDTF‍UDAF‍函数区别‍
7. hive用过哪些内置函数‍
8. 你有什么想问的

## 面4

1. 项目流程串讲
2. HIVE的优化，数据倾斜，小文件
3. 做过的一些指标和实现思路
4. 拉链表的实现思路
5. 数仓的建模过程
6. 常用的一些数据库

## 面5

1. Linux查看文件命令
2. HIVE的优化，数据倾斜，小文件
3. hive和mysql的比较
4. 用过哪些系统函数
5. hive优化做过哪些
6. 建模中范式建模和维度建模的区别
7. 拉链表怎样实现的
8. 用过shell什么工具
9. 写过shell什么脚本

## 面6

1. 自我介绍‍这家主要招离线
2. 讲一下宽表
3. 用哪些系统函数
4. 自定义函数三个的区别
5. 数据怎么导入‍从HDFS到0ds层‍(系统数据导入到hive)
6. HiveSQL怎么变成MR程序的
7. 两张表关联‍join的问题
8. mysql和hive的区别
9. 拉链表你们用吗？怎么实现的
10. Linux常用命令
11. 是刚刚辞职吗‍离职原因‍在北京吗‍疫苗打了吗‍现在薪资‍期望薪资
12. 面试通过什么时候到岗
13. 你还有什么要问的吗？问了‍解决什么问题‍数据量有多大

## 面7

1. 介绍项目（离线）
2. 处理数据使用的是什么？
3. 开窗函数举例
4. sql执行顺序
5. MR切片机制？
6. shuffle机制
7. linux常用命令，改变文件权限命令
8. hive sql优化
9. 数据倾斜
10. 拉链表

## 面8

1. 自我介绍
2. 项目介绍
3. ods层数据从哪里提取
4. flume使用的配置参数
5. Kafka的底层架构
6. hive优化
7. 使用Java和Scala做什么工作
8. datax通过什么方式使用Linux还是前端界面
9. Kafka的分区策略

## 面9

1. 自我介绍；
2. 介绍离线项目;
3. 数仓中有哪些主题：用户主题，流量主题，商品主题；
4. kafka积压？
5. kafka的分区数跟节点数的关系？

## 面10

1. 自我介绍；
2. 介绍实时项目架构及建模流程；
3. 维度表怎么保证一致性？答：phoenix幂等性
4. 新增指标怎么办需要历史数据？答：再写一个flink消费历史数据；
5. 数据量多少？
6. flink有过积压吗？
7. kafka积压？
8. 数据回溯？什么是回溯？丢了一条？答：kafka和flink可以保证不丢数据精确一次；
9. 集群规模多大？
10. 离线项目架构？有没有数据倾斜？
11. 数仓参与设计了吗？参与哪方面的？
12. kakfa参数优化？kafka多少节点？在职离职？期望薪资？

## 面11

1. HBASE保存的维度数据 你们公司的是怎么设计的key（具体说rowkey的设计原则）
2. 用户维度表制作的时候有遇到数据倾斜吗，怎么解决的
3. 实时和离线你都负责的什么部分，有哪些问题，怎么解决的。

## 面12

1. 自我介绍
2. 说一下最近的项目
3. 不用kafka可以吗？
4. kafka有多少个主题、为什么这么多。数据放一个主题不行吗
5. kafka的分区数是多少、怎么得出来的。
6. id、薪水。求第二高的
7. id、科目、分数 转换成id、英语、分数、语文、分数、数学、分数
8. hive中like函数你们怎么用的
9. 两个园代表两个表，不要中间交集部分，你怎么做。
10. 行转列和列转行
11. Flink的状态编程有哪些
12. Flink的提交有哪些参数
13. 算子状态编程

## 面13

1. 自我介绍
2. 介绍项目
3. 星形建模和雪花建模，星形建模的优缺点
4. 主要负责什么，sql用的多吗
5. 用过分桶表
6. hive导入导出数据怎么做
7. 怎么解析json字符串
8. kafka的分区策略
9. 薪资、疫苗相关的问题

## 面14

1. hbase的rowkey怎么设计
2. flink用的哪个版本(我回答的1.13)
3. 又追问了flink1.13有哪些新特性
4. 维度数据为什么存在hbase
5. 不存在redis是因为你们事先试验过存不下维度数据嘛
6. 每秒处理数据的条数
7. 问我什么函数能调出来状态
8. 算子和键控状态的区别
9. 富函数和普通api区别
10. 部署模式有哪些，你们用的什么
11. 为什么用yarn-per-job模式
12. 分布式快照机制原理
13. 保证数据一致性只有2pc嘛，那你们往kafka写为什么不用幂等写入
14. id、薪水，求第二高的
15. id、科目、分数 转换成id、英语、分数、语文、分数、数学、分数
16. 两个表，不要中间交集部分
17. 期望薪资

# 讯方科技

## 面1

1. spark的shuffle
2. hadoop的shuffle
3. Hadoop的shuffle和spark的shuffle有什么区别
4. 怎么优化，把shuffle过程去掉可不可以，为什么
5. ack的机制
6. flink相关的遇到什么问题，反压发生在哪种情况，怎么解决
7. hive做过哪些优化
8. 你做过项目的哪部分，负责什么
9. 你会考虑后来能不能转运维吗
10. 这家公司就是会先让你主动自我介绍，然后说做过什么项目，项目中用过哪些组件，然后展开问一些

## 面2

1. 先给我介绍他们公司的的情况，北京到处驻场，问我能不能接受
2. 让我自我介绍
3. 说一下我熟悉哪些技术栈
4. hive优化
5. 为什么spark比mr快
6. spark有相无环图解释，stage划分，宽依赖和窄依赖划分；跨stage的task是否能并行
7. hbase的rowkey设计原则及热点问题
8. 组件是自己匹配的还是CDH，组件之间的依赖关系：hive依赖啥，kafka依赖啥，hbase依赖啥，spark依赖啥
9. kafka角色和平时遇到的问题
10. SparkStreaming怎么维护offset
11. 后面就让我问他一些问题，我简单问了几个问题，驻场、几个人驻场一个公司、公司的大数据这块架构等等

## 面3

1. 先介绍自己做的
2. yarn的提交流程
3. yarn‍task失败后会怎么处理。
4. spark流程
5. spark‍shuffle种类和区别
6. spark做过那些哪些优化‍
7. kafka架构
8. broker里有什么
9. controller的作用
10. leader挂了以后怎么选举
11. isr是怎么来的
12. hbase读写流程
13. hbase一般存几个列族，能存多个列族吗，为什么
14. hbase刷写时机
15. kafka如何保证数据不丢失
16. 还问了一个‍spark‍client‍和cluster的区别

## 面4

1. 主要问沟通交流方面（项目有冲突，a客户需要你去解决问题，你正在做b项目
2. 功能没测试，然后客户需要去上线，你会这么说
3. 技术问题：Linux的权限命令，替换字符串命令，脚本定时执行

## 面5

1. yarn‍的task任务‍谁监控管理的‍挂了怎么办‍怎么调参数的
2. hive‍优化‍hive数据倾斜
3. hdfs‍小文件怎么处理的
4. spark ‍SQL‍ spark‍ Spark Streaming了解吗‍
5. ‍kafka‍架构（4张图）‍数据积压怎么解决‍分区‍结点‍副本‍怎么设计的？‍
6. Hbase‍架构‍，读写流程，‍一般存几个列族，能存多个列族吗，为什么，
7. hbase刷写时机

## 面6：（二面）

1. 先自我介绍
2. Linux用过什么命令
3. Linux怎样查看一个文件前几行
4. Hive元数据存在哪里
5. HBase的二级索引
6. 离线数仓给他串讲一部分
7. 数仓建模的意义

## 面7

1. 对哪些组件是比较了解的，包括它的运行原理
2. 任务提交到yarn中，他的执行流程是什么？
3. 只执行过程是，某个task失败，是怎么监控的？由谁进行监控？
4. hdfs避免大量小文件，为什么会产生大量小文件？怎么解决呢？
5. 在使用过程中遇到了数据倾斜了，运行慢了，执行速度没有那么高了，这些问题都是什么原因导致的，怎么处理。
6. hbase的存储架构，读写流程。
7. hbase写流程内存刷写条件？
8. hbase热点问题？
9. hbase的两种合并方式有哪些？
10. 你知道kafka的组成架构吗？
11. borker里有partition，leader，follower，controller，他们之间的关系是什么？
12. kafka 分区leader挂了怎么办？
13. kafka遇到数据积压怎么解决？
14. kafka消费数据的时候怎么保证消费的数据不丢失？
15. spark的执行流程
16. spark提交client模式和cluster模式的区别？
17. 提交spark会形成DAG有向无环图，DAG有向无环图的stage是怎么划分的？
18. 宽依赖会发生shuffle 你知道shuffle有几种吗？
19. linux常用命令

## 面8

先问我熟悉什么框架 我说flink kafka hive 几个, 然后技术总监大部分问题都是在 flink上面提问

1. 使用 kafka 时如何确定分区数?
2. 那么通过什么方法测出 消费者的吞吐量 和 生产者吞吐量, 具体说说? 我说压测 , 然后他问 具体是怎么做的?
3. flink 与 kafka 组件对接的时候有遇到什么问题吗?
4. 场景题 如何使用flink 实现一天独立访客数?
5. 如何实现flink 端到端一致性?
6. sparkStreaming 与 flink 的区别?
7. hbsae rowkey 如何设计 ?
8. 你刚才提到 rowkey 设计原则中长度原则, 这个长度原则 你怎么理解 , 具体说说为啥rowkey 不适合设计太长?
9. hbase 有哪几个版本 数据重复是直接覆盖, 而不是存为历史数据?
10. flink 提交流程
11. 有没有用过es
12. hadoop 优化 ?
13. 离职原因
14. 还有什么要问的吗?

# 柠檬微趣

1. scala语言的特质,闭包,模式匹配，变量，抽象
2. spark的rdd的特点
3. 精准一次性消费如何实现，如何手动维护
4. 为什么选用direct，以及他的并行度是怎么回事
5. 场景题，如何在大量数据中快速抓取某类数据
6. 算法加数据结构加链表，没整理，听不懂也答不上来

# 椰子传媒

1. 技术方面，直接拿我简历，翻来覆去挑着问的，没有问超过简历之外的技术，所以同学们遇到了这公司，自己简历上的得会
2. 问我薪资待遇，然后介绍他们公司业务，广告行业我能不能接受

# 嘉华颐和

1. 经历
2. 做没做过数据分析
3. 没问技术
4. 介绍了他们公司情况

# 商越网络

## 面1

### 笔试：

1. 两个时间戳相减等于多少小时
2. 时间戳转日期命令 from\_unixtime
3. 字符串中取最大和最小字符串
4. 还有一些简单的SQL题 排序，取最大最小日期，各学科成绩都80分以上的学生ID，一日留存这类

### 面试：

1. linux查询命令
2. 往HDFS上传文件的命令
3. 解压文件命令
4. 单引双引的区别
5. left join，inner join full join的用法
6. Sqoop了解吗
7. CK的引擎
8. 数仓，数据库，数据集市有什么不同
9. hive优化
10. 数仓建模，怎么分层，都有什么作用
11. 有什么想问的

# 优加健康

1. 怎么实现用户画像的标签任务，具体用了什么方法，代码怎么写的用了哪些算子
2. 怎么去划分一个用户偏好标签
3. 用户画像怎么处理数据了
4. flink 实时项目介绍
5. flink 用了什么api
6. 你做了哪些指标，说其中一个，从ods到ads 数据一步步怎么处理 每一层都用了什么方法，怎么处理的，怎么关联的。
7. 中间做了哪些优化。
8. ods层在kafka中分了几个副本
9. Kafka 丢数据
10. flink 怎么维护offset
11. Kafka积压
12. flink为什么要用动态表
13. 动态表存在哪了
14. zk都监控了哪些组件
15. shell脚本 写从1-100 的输出

# 中国软件

## 面1

1. 离线数仓分几层？
2. 离线你做了哪些？
3. 指标你怎么理解的？
4. 实时你做了哪些？
5. 实时中，数据的补偿机制？没懂，好像说的就是数据一致性，说完flink怎么保持数据一致性之后，又说Kafka参数调优保证数据一致性，他就说怎么调优都不可能保证数据一致性。
6. hive开窗函数

## 面2

1. 提到权限这块，说是央企做税务这块的东西，权限要求比较多，我说我们不怎么涉及
2. hive了解多少，数据倾斜怎么处理，两个大表join数据倾斜怎么处理，问我们数据量
3. 构建业务矩阵这块不涉及业务该怎么做，提到权限相关，不了解
4. 说一个开窗函数的具体用法
5. flink的数据一致性，重复消费的数据会迟到，说了一下数据乱序怎么处理
6. 对指标和标签的理解
7. 其他问的都是大保健上的，没啥难度

# 国研大数据

1. 处理的数据量多大
2. kafka包括实时和离线接入数据的峰值是多少
3. flink用什么版本，什么时候开始的实时项目
4. 集群规模，公司地址，部门多少人
5. Yarn页面总的计算资源，多少CPU，多少内存，离线划分多少，实时划分多少
6. 离线每天跑多少指标，每人负责多少个指标
7. sql 地区，价格两个字段，每个地区的价格TOP5，不用开窗
8. 分桶用过吗
9. Spark哪些算子触发shuffle
10. 数据质量监控用什么监控，了解吗
11. 元数据管理了解吗

# 高伟达

1. 自我介绍
2. 重点问了hadoop集群的搭建问题，spark hive怎么配置的
3. 数据量 又问了Spark内存怎么配置的
4. 数据怎么处理数据倾斜

# 汉德科技

1. 自我介绍项目
2. DATAX了解的多吗，遇到过其他问题吗
3. 增量数据怎么采集的，为什么选择maxwell
4. 写到Kafka中的话有几个TOPIC
5. maxwell的底层原理
6. Kafka下游使用什么到HDFS
7. Scoop和Datax的对比
8. Hadoop小文件的危害和解决
9. HIVE小文件的处理方式，Hive常用的存储格式
10. Spark和MR的对比
11. 了解CK吗，讲了下表引擎计算快，还了解过其他数据库吗，
12. 然后就是问我有没有其他问他的，问了数据量，项目主要做什么，公司的数据中台是怎么运行的

# 安胜

## 面1

1. 项目介绍 自己负责哪些工作
2. 公司是做什么的 面向什么客户
3. 项目上线了吗？活跃多少？
4. 集群多少台
5. 参与过的项目遇到的问题 怎么解决的
6. 实时做了什么
7. kafka默认消息时间是7天 7天以后删除的机制 是什么
8. hive的分区表和分桶表的区别
9. shell脚本$1什么意思
10. awk的用法，linux工具
11. 平常开发语言 我答得Java
12. map和flatmap有什么区别
13. flink的窗口 水位线 基本概念问题
14. 上一题的延伸 实际项目有用flink做过什么吗
15. 讲一下离线数仓的dwd dim dws层干了什么
16. 离职了吗 离职原因
17. 工作强度怎么样
18. 薪资要求
19. 有什么问题要问的吗

## 面2

1. 介绍一下项目？ 大概讲了一下采集系统以及维度建模
2. 项目团队几个人？
3. datax的工作原理？
4. datax好在哪里？
5. kafka默认消息时间是7天，大量数据批量过期怎么删除？ broker里面用segment，里面有.log,.index,timestemp, 当一个segment中的全部的ts都过期就删除
6. flink时间窗口和水印？
7. hive分区和分桶？
8. 实际生产中遇到的问题？
9. 平时加班多吗？
10. 对linux熟悉吗，awk干什么用的？
11. hdfs默认副本是3个，这么做的好处，基于什么考虑的？
12. 工作之外对什么领域有研究？讲了给大概的hudi
13. 问我跟kudu有什么区别？
14. 为什么离职？
15. 期望薪资？
16. 还有什么想问我的？

# 成都执掌天下

1. 自我介绍
2. 介绍实时数仓集群
3. flink dim怎么做的
4. clickhouse里边的多表join怎么优化
5. hbase一个表如果有6亿数据，查询优化
6. 集群服务器有多少台，集群组件各放在哪里
7. 你们数据量多少 你们业务数据量多少
8. 实时的flume传输遇到什么问题
9. dataX如何实现增量同步
10. dataX和flume有什么区别
11. hive 执行sql的优化
12. window join的本质
13. 用户量、商品量、维度最大的字段

# 集度汽车

## 面1

1. 自我介绍
2. 为什么先离职
3. 上游下游对接业务
4. 职业规划
5. 遇到的最大问题（业务和技术）
6. 两个sql留存率，路径分析（不让我用开窗最笨的方法说）
7. 思维题：有一个团购网站：领导看报表的时候发现一个烤串店12月的数据比6月多50%，那些原因

## 面2

1. 自我介绍；
2. 离线数仓分层，每一层都是怎么搭建的；
3. 离线数据域划分；
4. 自己负责的数据域的数据量；
5. 自己负责数据域中有哪些表，那张表的数据量最多，大概多少条；
6. 零点漂移问题怎么解决的；
7. hive中数据倾斜的优化
8. 什么场景下发生的数据倾斜，具体是怎么解决的
9. 对数据质量监控的理解，你做过这方面的工作吗？（我说没有），接着问，如果没有你怎么保证你计算的结果是正确的
10. sql table表，两个字段a,b，每一行表示一个用户关注了另一个用户，让求出相互关注的用户；
11. 向他提问

## 面3

1. 自我介绍
2. 随便介绍一个项目
3. 实时项目中dwd层多流join窗口等待是如何处理的？
4. 迟到数据如何处理？
5. 实时数据的结果是用什么存的？
6. 把窗口粒度比较细的情况下，是否近乎明细数据，存储是否有压力？
7. 用户画像介绍一下
8. 维度建模有什么优点？探讨了数据一致性哪种模型好
9. hive的数据倾斜？join数据倾斜怎么解决？
10. 事实表都有哪些种？
11. sql已截图

# 四川星点网络科技有限公司

1. 自我介绍
2. 讲一下实时数仓
3. 为什么你们实时维度数据没有做维度整合（涉及多个商品信息），是来一条数据都去关联吗？
4. flinkCDC读取配置表可以监控维度配置表的变化，比如业务数据库中维度表字段发生变化了会怎么样？我说我们维度表字段不会怎么变化，那岂不是使用flinkCDC矛盾了？
5. 有没有遇到多个事实表关联的情况？
6. 如果有多个事实表要关联，使用flinkSQL使用什么join
7. 窗口你们开多大，统计一天的实时指标比如GMV怎么实现的？
8. 如果要统计UV这种是不是就不能用窗口了，该怎么实现？
9. flink数据一致性，链路延迟，数据乱序
10. 实时项目中遇到什么比较困难的点吗？
11. 你们用户画像是怎么做的，里面比较有挑战的点有哪些？
12. 说一下你们离线的情况？
13. 建模用的什么？具体说一下

# 东软

1. 各项目搭建过程遇到什么问题。
2. 了解zookeeper脑裂么，怎么预防。
3. 指标GMV指标（面试官没听过）、品牌复购等。
4. Spark调优。
5. scala常用API。
6. SparkStreaming运行机制。
7. 星型模型
8. 场景题：
9. 个文件按T计量（存的都是URL），怎么实现去重
10. 如何处理单条记录消耗过大，即处理单条记录时，突然吞吐量不高怎么办。前提机器性能不高。
11. 一个40亿的数据量的数据，给你一个条数据（比如123），怎么最快查询40亿的数据里出现过没有？

# 法本

## 面1

1. 数仓为什么这么分？
2. 如果想取一个不脱敏的数据你们是怎么做的
3. 你们etl用的是什么？
4. 在这两个项目里你主要负责什么方面？都写过什么API？主要思路讲一下？
5. 你都看过Java、spark的什么源码？
6. 这数仓是你一个人搭建还是你们组一起搭的？现在要你搭一个采集平台或者数仓要多久？
7. 这些框架之间的兼容性你们是怎么解决的？有没有可以代替这些框架的？
8. 如果给你一个杂乱的数据，里面有json、txt、word。你怎么取出你想要的数据？

## 面2（大兴航空项目）

简历上写的都问了，一条一条问

1. Linux常用命令
2. flume你们是怎么用的，三个器用了哪些
3. 第二级flume hdfs sink出现过什么问题吗
4. 对于文件回滚的底层原理有了解了
5. zookeeper底层了解多少，是只是用搭建分布式框架，还是对底层算法都了解
6. 那有没有使用过zookeeper的命令，或者看过zoopkeeper上的节点信息
7. 创建一个Kafka的主题，命令行怎么写命令，口述一下
8. kafka的分区和副本有什么区别
9. Kafka的分区能被多个消费者消费吗
10. Kafka消费者和coordinator之间的心跳时间，这个知道吧，配置这个时间的参数是什么，你是怎么配置这个参数的
11. 写过Kafka的代码没有
12. Hadoop高可用怎么搭建，参数怎么配的
13. Yarn的资源是怎么配
14. 你现在还会hive吗，我现在有两个字段，一个是客机编号，一个是客机上座率，需要你先对上座率排序，再对编号排序，从建表语句开始说

drop table...;

create table ...()

partiton by ... 这里问到了，我现在没有时间字段，你觉得应该怎么进行分区

row format ...

stored as orc

location ....;

load inpath ....

然后就是二级排序，在这说了下四个by的区别

1. 连续三天相关：有两个字段，一个是id，另一个是一个是否合格（取值两种，yes no），问对于同一个id，连续三次及以上合格，具体的需求有点忘了
2. 有写过spark程序吗，你们都是怎样使用spark的
3. 有写过spark程序吗，你们都是怎样使用spark的
4. 知道hive on spark 和spark on hive的区别吗
5. Flink的内存管理是怎么做的
6. Flink的内存管理，就是JobManager和TaskManager的资源分配，这个知道吗

## 面3

1. 自我介绍
2. 介绍项目
3. 业务数有没有财务数据
4. 实时数仓层级的调度
5. 实时指标不正确怎么办
6. 用过SparkStreaming嘛 没有
7. kafka数据有序
8. 业务数据增删改查的数据到hdfs 怎么出处理
9. SparkSql 写过吗
10. mr和Spark谁快
11. spark 的Shuffle一定会落盘嘛
12. 大小表优化
13. Clinkhouse用过哪些引擎 有哪些优点
14. 大量数据写到Clickhouse 优化
15. SQL题

学生表： 学生id 学生name

课程表： 课程id 课程name

成绩表： 学生id 课程id 成绩

求每个学生各科成绩排名的优化SQL

1. 还有什么想问的

## 面4

1. 自我介绍
2. 讲一讲你做的实时项目
3. 讲一下你对flink的了解
4. flink和sparkstreaming的区别
5. 问了hadoop以及yarn的一些东西
6. 讲一下hive还有他的一些优化
7. 讲一下对spark的了解
8. 大概问了一下数据胡
9. 还有什么想问我的

# 巽联科技

1. 自我介绍
2. 公司规模，团队人数，组长如何安排的任务
3. 最近工作内容介绍
4. 离职原因
5. 串讲采集系统
6. 介绍一下海豚调度器
7. 关系型数据库有哪些，用的熟练程度
8. 建模过程串讲
9. 用户画像标签的sql与代码如何实现
10. hive底层 元数据有哪些信息

# 首信云

## 面1

1. 介绍画像项目
2. flume采集日志数据，需要过滤出JSON中某个字段为空的，拦截信息给前端，需要用户看到，怎么做？
3. flume采集数据直接到数仓中，文件中包含空格换行等一些特殊字符，转换成ORC存储，会有错乱，怎么处理？

## 面2

1. 离线项目
2. Flume怎么采集
3. Flume里面的sink是什么
4. 怎么把数据从hdfs写入hive
5. 如果你要的指标，离线数仓里面没有数据怎么办

## 面3

1. 做个自我介绍
2. 介绍一下用户画像的流程
3. bitmap的作用
4. pivot的用法
5. Flume如何自定义拦截器过滤空值
6. 你在离线项目中主要做了什么？
7. Flume采集数据在HDFS中的存储格式是什么？

# 嘉一未来

1. 介绍下项目，离线实时都介绍了
2. 项目组有多少人，都是干什么的，你是干什么的
3. 每天的数据量有多大，kafka有多少个topic
4. Kafka遇到的问题（挂了、丢了、重了、积压了、乱序了都问了一遍）
5. Kafka的高效读写
6. ES了解吗
7. Flink反压
8. Spark和sparkstreaming了解的多吗

# 万古恒信

1. 画像项目的流程以及标签的定义
2. flume采集日志数据，需要过滤出JSON中某个字段为空的，拦截信息给前端，需要用户看到，怎么做？
3. flume采集到HDFS上的文件格式，DATAX采集到HDFS是什么格式
4. 如果让你独立搭建数仓系统是否可以
5. 数据治理，数据质量分析是否了解过，能否独立安装
6. flume都有哪些sink，是否了解过，能否配置使用
7. 问了下家住何方
8. 剩下的记不得了，白话了一会

# 泰康

## 面1

1. 数据量大小
2. 负责什么模块
3. 比较难的业务，怎么实现的
4. 状态怎么保存，为什么用这个方式保存
5. flink join使用场景
6. dim大小
7. 为什么引入redis
8. hbase索引种类，用了什么索引，为什么用hbase
9. 并行度设置方式，为什么设置这样的并行度
10. 检查点超时报错怎么处理

## 面2

1. 自我介绍
2. 具体介绍技术栈，实现了什么？（我说的项目采集平台和离线数仓搭建）
3. HiveSQL优化，Hive数据倾斜解决
4. 实时项目反压和数据倾斜解决？
5. 用户画像（标签）讲下（我就说了三级标签，web平台是其他同事开发的）
6. Zookeeper了解吗？什么组件用到了Zookeeper?
7. Canal了解吗？
8. Clickhouse和Doris对比？
9. 未来规划？
10. 有什么想问的？

## 面3

1. 介绍离线 实时
2. 数据最后写到哪里 离线 实时
3. yarn工作机制
4. 为啥离线不用ck 实时用ck？
5. hbase rowkey 预分区
6. habse 那些数据会写到1000-2000的分区中？
7. flink 水位线
8. hive优化
9. hive数据倾斜
10. spark什么情况会shuffle
11. kafka的数据重复
12. kafka数据丢失
13. kafka的qps是多少

# 老虎国际

## 面1

1. 介绍上一家工作项目
2. 数仓数据来源，你们的数仓指标是怎么确定的？
3. 列举下统计的指标？
4. flink状态存储类型及各自的优缺点？
5. file和rocksdb是否都支持增量？
6. flink的state类型？
7. flink的checkpoint的存储原理，若缓冲区状态满了，该如何处理？
8. 说一下flink指标的实现流程？我讲了下各省份每日订单数量，问到了维表关联用的什么函数？
9. clickhouse的数据类型有哪些？
10. hive的数据类型有哪些？
11. hive的存储类型有哪些？json，textfile，还有哪些？
12. java方面，常用的集合类型？hashmap和hashset的关系？
13. 常用的排序方法有哪些？各自的时间复杂度和空间复杂度是多少？
14. 牛客网上写一个归并排序，讲一下归并原理？

## 面2

1. 自我介绍
2. 介绍项目 介绍的实时
3. 具体负责的部分，具体是怎么实现的 ，用到了哪些flink独有的特性
4. checkpoint 的barrier 对齐 机制
5. flink数据倾斜解决办法
6. flinksql 有几种
7. 介绍ck
8. bitmap 好处和坏处
9. java 归并排序，用两个链表实现方法栈，写一个euqal方法实现比较两个二叉树是否完全相等，说出有几种排序 和 他们的时间复杂度和空间复杂度 ，说出currenthashmap原理。
10. 写出topN 的sql

## 面3

1. 自我介绍
2. 上份工作经历，项目经历
3. 是以离线为主吗
4. 离线，实时整体架构与业务划分
5. 整个数据流向与处理
6. 技术擅长领域
7. 讲一下具体擅长技术
8. hive的数据类型
9. hive架构 底层 暴露的东西
10. hive组件作用
11. spark原理
12. kf架构
13. broker作用
14. 选举过程
15. ck数据类型，表引擎
16. 算法空间，时间复杂度
17. java实现类
18. 线程安全的集合
19. 归并排序java实现
20. 子查询实现sql
21. 链表问题，具体我也不清楚了，人已经傻了
22. 有什么想问的

# 百度外包

1. 海豚调度器工作原理
2. datax页面化使用
3. flink的异步io的底层原理
4. 状态和检查点的区别
5. 对sql的优化
6. 数据治理
7. 项目难点
8. 对集群的一些优化

# 华胜天成

## 面1

1. 离线项目串讲
2. HIVE优化，小文件，数据倾斜
3. 写过的指标、疑难指标解决思路
4. 一道sql题
5. hive分区分桶表
6. Linux系统常用的命令，查看内存、磁盘、查找文件
7. 拉链表的实现思路

## 面2

1. hive的优化
2. hive数据倾斜
3. 分区分桶表
4. 拉链表
5. 一道sql题：查询左表有右表没有的数据
6. 查一个表的前五行
7. TopN

## 面3

1. 讲一下项目架构
2. 讲一下clickhouse和hive的区别
3. 平时用hbase吗？没听清
4. 讲一下脚本

## 面4

1. 介绍项目
2. Linux查看进程号
3. Linux查看文件路径
4. hive去重
5. 拉链表的制作
6. dwd层数据有问题，怎么重新导入

# 实惠多多

## 一轮技术面试（数据开发主管）

1. 你们公司的业务是什么。详细说说你们数仓怎么使用了
2. 说下怎么实时数仓的数据是怎么来的，你们离线的用户行为数据是怎么产生的，你怎么处理的，业务数据是怎么产生的
3. flink 用过状态编程么？ 你们怎么使用的，你说说你怎么用的，结合你自己做过的指标说说怎么写程序。详细说
4. 那你用的是哪一种状态， 不开窗吗？ 我说是设置ttl ，如果开窗怎么开？ 用什么窗口，开多久。
5. flink CEP 用过吗？ 怎么怎么使用，结合具体指标，数据从哪抽取，怎么处理，实现什么？详细说
6. flink 的数据写到哪里
7. 集群规模多大，数据量每天多少，用户有多少
8. 离线数仓说说你们怎么怎么采集数据的
9. kafka 怎么处理重复数据
10. Kafka怎么处理积压数据
11. kafka 的主题删除方式有哪些，有需要注意什么吗？
12. 现在有个场景，kafka出问题你怎么去排查出问题，说说你的思路
13. Doris 用过吗
14. Doris和clickhouse的区别是什么 有哪些引擎
15. 为什么你们没有用Doris
16. flink用的什么版本
17. 离线数仓建设最重要的是什么
18. hive你们做了哪些优化
19. 你们怎么去做数据治理的
20. spark 你们是怎么用了，有哪些应用场景
21. flinkcdc你是怎么用了

## 二轮面试

1. 你们公司业务板块了解吗，说说
2. 之前公司数据开发团队多少人
3. 你们有做数据中台吗，了解过吗
4. 说说你的用户画像项目
5. 你们怎么设置标签，让我说怎么分级，一直说到4级
6. 对自己有什么职业规划
7. 问薪资

# 名途信息

1. 介绍一下用过的计算框架
2. 说一下flink join
3. flink怎么避免数据乱序，端到端一致性
4. hive内部表默认存储路径
5. 简单说一下spark宽窄依赖
6. 场景：窄依赖，两个RDD，两个分区，union以后有几个分区（两个）
7. kafka和sparkstreaming精准一次如何保证
8. 介绍一下印象比较深的做过的项目（离线和实时都说了，说了hive的优化和数据倾斜，flink大状态调优，反压，数据倾斜）

# 北京芯盾时代有限公司

1. 在实时数仓中主要做了什么？提到参与了技术选型，追问具体参与了哪些技术选型？怎么进行技术选型的，做了哪些工作？感觉面试官应该是想知道框架对比，为什么选这些组件吧。
2. 行为数据是存放在日志文件还是哪里？flume的taildirsource怎么实现断点续传的
3. maxwell基于MySQL的binlog日志采集数据发往kafka中，具体是怎么实现的？
4. 对那个大数据组件有深入了解，我说的kafka，kafka怎么保证数据不丢？kafka的分区的leader怎么确定的？（broker总体流程中leader选举机制）
5. kafka往zookeeper中注册信息是写一条记录还是创建一个节点还是怎么样的？
6. zookeeper的节点有那几种类型？
7. flink消费kafka数据怎么保证幂等性？数据的精确一次性，消费数据是一条还是一批？
8. flink的checkpoint的原理，flink用过哪些算子？并行度了解吗？
9. flink支持哪些分区策略？没答上，flink的提交模式有哪些？
10. 怎么确定kafka安装部署完了？怎么确定安装好了的kafka是完好的？应该是想问kafka的压测，压测命令是什么
11. 说一下clickhouse
12. hbase的读写流程，hfile和memstore的具体关系是什么
13. 去hbase中查数据会做哪些优化
14. hdfs的读写流程，写一个1G大小文件的流程，hdfs查看有哪些文件的命令？
15. namenode高可用两个怎么切换的，hdfs高可用自动故障转移机制，ZKFC组件
16. hadoop集群安装部署过程，提到文件句柄数？你知道吗？只听其名，不知其意，质疑我这都不知道
17. 问我有没有漏掉什么大数据组件，我说hive没提，就给我出了一个统计哪些班级，学生人数大于60人SQL实现，学生id，班级id，学生id不重复

# 飞轮数据

## 面1

1. 自我介绍
2. 工作经历
3. 离线项目 技术角度
4. java怎么样
5. 写入clickhouse用了什么技术栈
6. 用户画像怎么对外提供服务，怎么查询
7. kafka底层文件格式管理
8. 其他的分析型数据库
9. es
10. ck的优化
11. flink 的watermark
12. flink端到端一致性
13. flink两阶段提交

## 面2

1. 自我介绍
2. 项目过程中的优化
3. 问kafka的单分区数据不重复，那如何保证 实时环境 下多分区数据不重复
4. 问计算引擎了解多少
5. 问olap有啥
6. 问clickhouse了解
7. 问bitmap
8. 问hashmap，为啥要用红黑树替换链表，用红黑树解决什么问题；
9. 怎么解决hashmap的线程安全问题
10. 有啥想问他的

## 面3

1. 讲讲Java多线程
2. 讲讲JVM
3. 讲讲CMS，怎么用的
4. 讲讲NIO
5. Redis主从复制
6. 介绍一下ClickHouse
7. HDFS读流程
8. Flink端到端一致性
9. 水位线介绍下
10. 列式存储和行式存储区别
11. ClickHouse更新怎么做的
12. ClickHouse的SumingMergeTree和ReplaceMergeTree区别
13. Spark如何怎么扩大并行度

## 面4

1. 介绍一下比较拿手的项目，我讲的离线
2. 介绍Flink和Spark检查点区别
3. 介绍下CheckPoint机制
4. 介绍状态
5. Java 的 TreadLocl
6. Java多线程的信号变量
7. TreeMap底层是啥
8. 介绍下红黑二叉树
9. JVM虚拟机全套

# 文创思宇

1. 自我介绍
2. 干过什么项目，具体负责了啥
3. 离职原因
4. 根据简历问的，问java用的怎么样，会哪些东西
5. flume自定义拦截器用java怎么实现的
6. 自定义函数怎么写的
7. 工作中碰到比较困难的问题以及解决方案 我说的kafka的一些
8. 问了点离线的指标实现
9. 公司的数据量，人员配置，具体负责，上家公司做什么的，数据怎么来的等杂七杂八
10. hive的优化

# 奇点云

1. 说数据怎么来源
2. 说你们离线数仓的集群规模，你们有测试集群吗，离线数仓和实时使用的是一个集群么？ 会不会有集群资源的冲突？怎么解决了？怎么做测试？
3. 离线数仓怎么采集数据， 你们在采集数据之前会对数据做什么筛选吗？ 你会去关注些什么？
4. 你们为什么做数仓？ 说说你们数仓最后输出之后哪些业务会用到？
5. 有哪些建模方法？
6. 为什么选择星座模型？都有何区别,为什么不选er
7. 你们怎么去构建业务总线矩阵？是怎么把这个矩阵做出来的？ 是你自己做的吗？
8. 都有哪些人去参与的？ 怎么个工作流程？
9. 业务总线矩阵包含了什么信息？ 你们数据域怎么划分？是谁划分的？
10. 说说什么是原子指标？什么是派生指标？ 并且当场举几个例子？然后他又说了几个指标让我判断是哪一类的指标，为什么这么判断
11. 说说你们 五个层都怎么做了， 从ods 详细说
12. ods层怎么做的压缩。
13. dim层 你们做了哪些表，说几个。
14. 如果维度变换缓慢有做哪些处理呢？
15. dwd层怎么去构建事实表，设计依据是什么？
16. dwd的都做了什么事
17. 那你们怎么选择做维度退化，为什么
18. dwd你们有没有跨数据域的表
19. hive 有没有做过数据规模的统计
20. 你会做数据量的评估吗，怎么实现？ 比如如果现在用户量10w，集群规模怎么去考虑，如果以后达到100万。
21. 数仓的数据质量监控怎么做的？
22. 怎么做元数据管理，元数据管理是干嘛的？
23. 业务之间你们怎么做优先级的？
24. 生产中你们项目是怎么开展的具体流程？详细说
25. 你在其中什么角色
26. sql语句的准确性你们怎么保证的？你会怎么判断sql语句的正确性？ 你说下你的经验
27. 你平时开发是sql多还是api多
28. sql题：当场直接说思路
29. 连续7天登陆
30. 三个字段：任务id，任务的start\_time,任务的end\_time.求输出每一分钟任务量是多少？
31. 用户画像你们做了哪些标签？用户画像得到的数据存在哪了谁去用了
32. 你们怎么去做推荐标签？ 说说你做的 我说了rfm模型 追问怎么判断重要 如何衡量 详细说

# 汉熵

1. 自我介绍
2. spring cloud 用了那些组件?
3. ck 怎么去做存储 ?
4. ck 用到了其他的表引擎嘛?
5. MergeTree 有那些要素 那些字段 用来分区 ?
6. 用过那些函数?
7. 怎么做的数据清洗和转换写到ck中的?
8. 建模过程
9. hive执行sql过程
10. hive接收执行完的结果
11. 实时数仓数据没有保存在hdfs中 保存在什么地方 数据分层都是在flink中的嘛?
12. flink中做了那些内容? ods拿数据 怎么做一些处理操作?
13. 数据倾斜怎么解决的 ?
14. 具体点,打散后怎么办?打散策略是什么? 开窗然后聚合统计 然后在聚合 说不对
15. kafka 怎么部署的 怎么做的topic 分区 副本 多少个?
16. ck几个服务器? 怎么保证服务器数据一致的 怎么同步的数据 ?

# 中软闻歌

## 画像

1. bitmap原理 和为什么用
2. 怎么与spark结合的？
3. spark集群几个节点？

## 实时

1. 数仓分层
2. cdc过程遇到的问题
3. FlinkCDC 1.x 锁表问题 ， 配置表， 配置库，对我们实时计算不产生任何影响
4. 成就比较大的地方
5. flink 提交模式

## 离线

1. datax遇到问题
2. 其他框架有调研嘛
3. Kafka， kubemq（云原生，比较窄）
4. datx内部架构
5. datax同步hive流程
6. MySQL 有逗号 怎么同步到hive
7. Hive同步mysql，分隔符
8. 字段有空格，regexp\_replace
9. flume过滤器定义
10. 拦截器
11. kafka 缺点， 依赖zk
12. 全局有序
13. 分区减少不行 为什么
14. ack
15. topic变化 生产者怎么感知
16. topic.metadata.refresh.interval.ms
17. kafka怎么与flink连接
18. hive 优化 大保健
19. hive版本

## Java

1. 设计模式
2. 内存泄露 溢出区别
3. 溢出： Out of memory 报错
4. 泄露：申请内存暂时没有被释放，暂时不会出问题，导致溢出
5. 注解类
6. spring生命周期
7. flink会落盘
8. 计算：算子，TM内存
9. 大状态：RockDB， Ck快照到hdfs
10. catlog了解过吗
11. Flinksql，表，字段

# 安胜（二面）

1. 项目中遇到的问题
2. 部门多少人
3. 技术怎么确定
4. 未来职业规划
5. 做过数据分析吗
6. 用户画像怎么做的
7. 前后端怎么配合
8. 用过阿里云，华为云这些吗
9. 之前薪资，期望薪资
10. 讲了一下他们公司的结构（面向政府的项目）
11. 如果入职可能要去厦门总部培训

# 实惠多多

## 一面

1. 自我介绍
2. 介绍实时数仓
3. mysql的去重方式？
4. hive常见函数，表函数有哪些？处理json的有哪些？
5. 实时数仓如何过滤不合法的json，或者一些不需要的数据？
6. 如何将hive中的数据读到clickhouse ？
7. 是否用过doris？
8. 数据中台怎么理解？
9. 大数仓概念

## 二面

1. 自我介绍
2. 介绍事实数仓
3. 怎么看商品的大数据搜索推荐
4. 怎么看数据库，数据仓库，数据湖？
5. 存储机制，处理介质
6. 有什么想问的？

# 汉德科技

1. 先介绍离线项目
2. 为什么加maxwell还要加datax，maxwell的底层原理
3. 数据落盘HDFS的文件格式
4. 数据量多少，组内多少人，原先公司做什么
5. hive sql和普通sql语法的区别
6. 数仓搭建过程和建模过程
7. 疑难sql的思路
8. row\_number、rank区别和使用
9. Spark中RDD的特性
10. 介绍用户画像搭建过程和日常负责工作 四个人面试的轮番问，其他的问题想不起来了

# 指南针

## 面1

1. 自我介绍
2. 你们的数据量
3. hive架构（组件，底层）
4. 技术栈
5. 数仓建模说一下
6. 集群谁搭的
7. HQL执行流程
8. 有哪些排序（4个by）
9. partition by个group by的区别
10. 开窗函数，举例子，都是干什么的，区别在哪
11. 行转列，列转行
12. 场景题：求累计百分比sql
13. RDD五大属性
14. rdd弹性体现在哪
15. spark作业提交流程
16. spark任务划分
17. sparkstreaming窗口
18. flink和sparkstreaming区别
19. flink水位线
20. flink窗口
21. flink状态
22. 主要用什么编程语言（java），不用其他的了嘛（shell写脚本没问题，了解python和scala）
23. 算法题：爬楼梯
24. 上一家待遇，期望待遇，住哪，你有什么要问的。然后让我好好看看算法

## 面2

1. 介绍项目 离线
2. 你为什么要离职
3. 团队和数据量
4. 聊聊大数据体系
5. hadoop读写流程
6. 副本默认3个
7. 纠删码原理
8. 3个节点传输怎么传 串行的还是并行
9. 100g文件 传完之后开始同步 还是怎么个流程
10. 读文件的流程 读哪个文件 就近原则
11. yarn架构
12. hive 架构
13. 操作树转换成 mr
14. 4个 by
15. 哪个by只能有升序
16. partitionby
17. patitionby 与 group by 区别
18. 对开窗函数的理解
19. join的话 需要注意些哪些
20. 谓词下推 和 列裁剪
21. kafka 和 flume
22. 数据漂移问题
23. 你对数仓规范的理解
24. 写代码在哪写
25. 集群搭建多久
26. 中间出现过什么问题
27. 都是leader负责
28. 如果说让你做数仓命名规范
29. 业务线怎么区分 大业务线 我们只有一个业务线
30. 分层分几层 哪五层

## 面3

1. 自我介绍
2. hive的优化，结合自己的指标
3. 讲一下实时的架构，用到的技术
4. 讲讲公司的数据治理
5. 讲讲离线脚本是怎么提交的，有没有集成化
6. 讲讲hive四个by，怎么用
7. 开窗的partition by 和 group by区别
8. 讲一下开窗
9. 出了一道sql题，统计店铺销量topN，累计销量等等
10. 行转列，列转行
11. RDD五大属性
12. rdd弹性体现在哪
13. spark作业提交流程
14. spark任务划分
15. 讲讲flink（面试官说自己是一个小白，给介绍一下，内存模型、cpu、水位线、窗口、状态。。。）
16. 公司用的cdh，想换成阿里云，问我知不知道我们公司搭建阿里云大概花了多少钱
17. 问了下他们公司的架构，所和我说的基本一样
18. 问了问我离职原因
19. 问了自己的发展方向，期望未来的团队什么样，期望薪资
20. 最后问了算法（二分，跳台阶，二叉树）

## 面4

1. 是否搭建过hadoop
2. 使用什么方式搭建的
3. 你们当时搭建有多少人
4. 四个人搭建了多久，都搭建了哪些组件
5. 平时用的语言格式，java，sql
6. 为什么离职
7. 集群搭建的过程中有什么问题么（指的是框架搭建，和业务中的问题关系不大）
8. 版本选型怎么选的，spark什么版本，hadoop什么版本
9. 算法能力怎么样
10. 底层原理：hadoop有哪些模块，每个模块有哪些功能
11. Hadoop读写流程具体的实现
12. Hadoop读写流程数据同步如何实现（要具体描述，写数据写三份怎么写的）
13. 不影响数据安全性的前提下，如何设置副本数量，最少的情况
14. Yarn的原理，调度的时候怎么调度
15. Ods层和dwd压缩带来的好处和坏处
16. 数据量有多大
17. 为什么不采用parquet压缩，各自的优缺点
18. Flink的水位线，时间语义，spark的时间语义
19. Spark支不支持时间语义，水位线机制，我回答之前了解的时候spark这方面不是很强没有flink清晰（告诉我现在有了）
20. Flink有哪些模块，这些模块有哪些功能
21. 实时数仓的分层，都存在哪里，如何实现
22. 一个数组abcd，求全排列（ab，ac，abc这种）问了快速排序原理，二叉树遍历的前中后序
23. 有什么想问我的

## 面5

1. 自我介绍
2. 对下一份工作有什么期待？
3. 离职原因
4. 未来想做哪方面的工作
5. 想加入什么样的团队？
6. 主要用什么语言？7.对薪资和领导的要求？
7. 算法能力怎么样？
8. 大数据了解最深的是什么？
9. 你的技术优势或者技术亮点是什么？
10. 你在工作中遇到什么问题是别人解决不了，你解决的？
11. hdfs出现小文件怎么知道改哪些参数的？
12. 底层原理了解吗？
13. Hadoop的组件有哪些，你对哪些组件了解深入？
14. Hadoop怎么管理文件和读写流程、调优
15. yarn怎么调度？
16. hive的窗口函数了解吗？
17. 学的什么专业？
18. 有什么想问我的？

# 网达软件

1. 介绍离线数仓
2. 介绍实时数仓
3. flink CDC怎么去做
4. 使用什么时间语义
5. flink精准一次性怎么保证
6. Kafka精准一次性怎么保证
7. zk里面维护了Kafka的什么
8. Kafka可以放弃zk嘛
9. hivesql的调优
10. spark参数调优
11. yarn class和yarn client 有什么区别
12. redis数据类型
13. redis存储模式rdb aof
14. hdfs读写流程？
15. hdfs写数据的话，串联通道一个节点挂了咋办
16. StringBuffer和StringBiuld区别
17. 基本数据类型和动态数据类型
18. JVM内存模型，问到这里我赶紧打住我不是干java的是干大数据的
19. 同步和异步区别 说了Kafka的异步操作
20. sleep和wait的区别 我不知道
21. 为啥用DataX，DataX底层是怎么做的。
22. 写java代码string 1.9.8<1.10<1.10.1

# 金康塞力斯

1. 实时项目串讲
2. 实时维度指标有哪些？dwd层做了哪些，dim层做了哪些，flinkCDC给他了讲一遍
3. spark的工作流程
4. 往hbase写100亿条数据怎么设计，我说的rowkey的设计原则
5. BI工具有使用哪些？
6. 项目中的痛点怎么解决的？我说的资源设置，反压
7. 公司是车企，类比一下车企的话该怎么设计数仓这些，我就把车类比成商品给他说了一下。
8. HR了解了一下个人情况巴拉巴拉

# 德特赛维技术有限公司

1. flink提交流程，数据一致性，checkpoint的理解，在项目的实际应用什么时间语义？
2. flink并行度和slot的理解
3. kafka怎么指定分区进行消费，发送数据分区策略，ISR的理解，数据重复
4. sparkSQL和hiveSQL的区别？为什么sparkSQL能使用hiveSQL的函数？sparkSQL怎么自定义函数？
5. clickhouse了解多少？使用过aggregatemergetree没有
6. clickhouse和kafka对接过没有？
7. 连续登录SQL实现思路
8. 公司主要用sparkSQL写代码开发，实时flink主要用scala代码开发，公司主要做离线的

# 枫叶汽车

1. 自我介绍
2. 离线架构(采集->分层建模) 问：知道用的这是什么架构吗？Lambda架构
3. Hadoop框架
4. HDFS文件块
5. 多人同时操作同一文件会出现什么情况？
6. HDFS块大小可以调整吗？在哪里调整？；
7. JobTracker了解吗？
8. Hive问知道orc吗？可以简单说一说吗？
9. Kafka 与传统的消息队列比较；
10. Flink与SparkStreaming区别；时间语义；WaterMark定义乱序时间；状态怎么保存？
11. 最近有学习什么吗？
12. 介绍公司情况

# 上海哲锦

1. 自我介绍
2. 介绍离线
3. 离线你负责什么
4. Linux命令熟悉吗
5. 有什么想问的

# 文创思宇

这是一家没有融资的自研公司，公司规模50多个人左右，主要给公安部做各种刑侦系统

1. hive优化
2. 指标怎么实现的
3. 离线架构
4. 实时架构
5. 画像架构
6. 面试官写算法和java的，问我你的代码中那里体现面向对象了，问了我会的设计模式

# 博英科技

1. 自我介绍，问那年毕业，专业
2. 实时flink主要的工作
3. 数据是怎么留转的ods-dwd，具体的代码api
4. flink运用什么方式运行api脚本，怎么提交
5. 有多少条数据每天
6. flink分流具体做了什么处理
7. 集群规模
8. 离线数据处理用到哪些技术
9. clickhouse，hbase都问了
10. 数据倾斜场景，处理方法的底层原理
11. 提问

# 数梦工厂

1. 自我介绍
2. 介绍项目? 从采集开始一顿讲离线
3. kafka积压了怎么监控到的？
4. 生产中遇到的问题？
5. 数据怎么来的？
6. 期望薪资？
7. **两个SQL**
8. 表A和表B有唯一主键id，重复的按表A，怎么做？
9. 如果表 有两个字段，id和名字，怎么变成[1,2,3,4,5]和[名字1，名字2名字3，名字4，名字5]

# 中软国际

## 面1

1. 自我介绍
2. 介绍一下用户画像
3. Flink 端到端一致性，ClinkHouse的读写优化
4. 你还有什么问题想问

## 面2

1. 自我介绍
2. 介绍一下用户画像
3. Flink 端到端一致性，sink端ClinkHouse是如何实现的
4. CAP法则为什么只能满足两个，zk满足了哪两个
5. 用户画像中是如何进行权限控制的
6. Springboot、mybatis掌握到什么程度了
7. 你还有什么问题想问（主要是做车企相关的数据中台）

## 面3

1. sink端ClinkHouse是如何实现的，遇到哪些问题？
2. 用户画像中是如何进行权限控制的？
3. 指标是如何定义的？
4. Flink中你主要做了哪些工作？
5. 你还有什么问题想问（主要是做车企相关的数据中台）？

## 面4

1. clickhouse、flink、spark、spring在项目中的应用
2. clickhouse的优化，还有在clickhouse中表的变化
3. clickhouse的删除

路径下面：文件

操作比较重

路径，将之前的数据迁移过来，迁移过程当中

对旧路径标记为inactive

Merttree合并时候，将旧路径删除

1. clickhouse即席查询用户留存率的一个实现

## 面5（泰康）

1. 离线集群怎么搭建？
2. 有没有因为版本不一致产生兼容问题，分别是什么版本的？
3. 即席查询presto,为什么选择这个？为什么他比较快？
4. 数仓怎么分层？实时数仓怎么分层？
5. hdfs高可用怎么搭建，有没有用2NN，高可用怎么实现（原理）？高可用各搭建了几台？
6. 场景题：现在要用flink展示当下，当周，当月的数据，怎么实现？思路--详细点--再详细点

## 面6

1. 文件存储类型，列式存储和行式存储的区别，列式存储有哪些，行式存储有哪些
2. 在工作过程中涉及到的hive优化
3. 小文件产生的原因、危害、以及具体的解决
4. 小文件已经产生了有什么解决的途径？具体设置什么参数
5. hive数据倾斜
6. 简单介绍一下数仓分层情况，各层存了什么数据
7. ads层数据是如何做可视化的
8. 调度工具用的啥
9. 数仓是如何分区的
10. 拉链表的实现思路，详细介绍一下开练闭练
11. hive的内部表和外部表有什么区别
12. 每天新增数据多少，整体数据有多少
13. 在项目中日常的工作是哪些
14. 简单介绍一下从提需求到实现需求的整个工作流程
15. 涉及到的编程语言
16. 有用过java解决数据仓库问题
17. 有了解过其他集成式的集群吗？
18. 事务表和非实物表
19. yarn的工作队列、调度器
20. 最大的并行数，起多少个进程
21. 一个不是很规范的sql 笛卡尔积 会占用大量资源 其他任务没有办法进行，如何优化
22. yarn对每个队列占用最大、最小内存参数的设置是啥
23. 查看yarn队列 的命令
24. hbase的应用场景
25. 了解元数据吗？元数据包括哪几类
26. 有了解过数据治理的体系吗

# 中创新航

1. 离线数仓项目串讲
2. flink数据一致性，flink窗口有几种
3. 提到职业规划，聊了数仓和数开分得还是挺清的
4. 有没有涉及到组件底层的优化？我说的flink的优化
5. 就没了？有什么想问的？问了他们数据量跟我说的100G差不对

# 库珀

1. 自我介绍
2. 你们部门是不是有个性毛的？
3. java掌握到什么程度？java怎么判断一个字符串是否为空
4. 数仓侧重于哪个方面？宽表是怎么建立的？能否能举一些例子
5. 怎么保证报表是正确的？比如说最后的现象这个数据非常高，你怎么排查
6. 你们怎么做数据治理的？
7. 你们各层是怎么建立的？星形模型和雪花模型怎么选择？
8. 场景sql
9. SQL的执行顺序是什么？
10. flink怎么保证分区是有序的？
11. flink的版本？spark的版本？
12. spark的union和union all有什么区别？
13. spark的动态裁剪了解吗？shuffle了解吗？
14. spark后的文件数量由什么决定的？

# 成都芯极客科技

1. 实时项目串讲
2. Maxwell和canal为什么选择Maxwell，底层原理，解析格式
3. MySQL的数据量有多大
4. 为什么选择将维度数据写入到hbase中
5. flink是使用SQL还是API？
6. 你们对历史数据是怎么处理的？
7. kafka数据保存多久
8. 你们sugar是别人做的吗？
9. 用户行为实时建模的那种是你们做的还是别人做的，涉及到算法？
10. flink调优你们做了哪些？
11. flink哪些算子容易导致反压
12. kafka数据积压怎么处理
13. flink数据丢失怎么处理的
14. flink到clickhouse的数据一致性，clickhouse查询方面的优化了解吗？
15. clickhouse有没有用到primary key，order by这些
16. flink开窗了解吗，窗口的分类
17. 异步IO是怎么做的？异步怎么保证数据有没有成功，补偿机制什么的？
18. 为什么选择clickhouse，没有用其他的
19. 有没有遇到clickhouse数据关联的情况，join性能影响有哪些方面
20. kafka的topic数据量过大遇到过没有？怎么处理，topic数据全在一个节点上
21. clickhouse有没有遇到过zk元数据不同步的问题
22. flink和sparkstreaming的区别
23. 了解kafka的shuffle机制吗？不知道
24. dataX知道吗，kettle知道吗
25. flink 的Dynamic table，动态表了解吗？
26. olap数据库的表示层工具用了哪些？BI工具sugar和superset，sugar是商业的吗？那个公司的？
27. 提问，离线实时都没有，项目比较大，帮别人做中台，大数据刚起步，

# 茶颜悦色

## 面1

1. 自我介绍
2. 讲一个你比较熟悉的项目
3. hbase有做二级索引吗？rowkey怎么设计？
4. hive的数据倾斜怎么定位？怎么解决？
5. 离线数仓有没有用过拉链表？
6. redis的缓存穿透和缓存雪崩是什么？怎么解决
7. 布隆过滤器是什么，用过吗？
8. 数据需求是和谁对接？产品还是业务方？
9. 没有产品你是否能对接？

## 面2

1. 自我介绍
2. 两个大表，都是十几个G,不使用大数据的情况下，单机模式查出两个表的相同之处
3. 业务中最大是那张表
4. spark的checkpoint跟flink的checkpoint有什么差别
5. 使用了什么特殊的指标
6. 数据量多少
7. 连续登录计算，隔一天也算计算
8. flink自己有没有开发端到端的一致性

# 国网电动汽车

1. 实时数仓项目的背景，业务支撑的情况，实时数仓建设过程以及架构是什么样的？
2. Flinkcdc是一个什么样的机制？
3. 你怎么理解DIM层？
4. DIM层的维度信息是由其他团队帮忙维护好的吗？
5. 能不能说下你们在使用clickhouse这块有什么经验或者什么挑战？
6. 我自己Q了Doris，然后问Doris为什么好用，以什么机制保证查询效率高？但是由于忘了，我大概说了下它支持join，然后是目前的学习计划。
7. 什么样的数据适合做到DWS层，什么样的数据适合做到ADS层？这两个层你是怎么理解的以及它们的职责？
8. 你们的团队数仓这几层都会去维护吗？是否分层分团队去做这个项目？
9. 你们在做项目的时候是否有数据处理以及规范来确定每一个层应该有什么数据？
10. 垂直的那些主题，比如DWS层那些垂直的主题一个人维护比较方便，那如果在ADS层可能用户域关联其他域的信息，这个是你们团队怎么做共享和互动的？或者存在这种情况吗？
11. 利用FlinkApi叫上redis解决查询效率和吞吐量问题，那么在开发有没有遇到反压问题，举个例子以及怎么解决的？
12. 你们的数量级大概是什么量级？高峰时段每秒吞吐量大概多少？
13. 怎么设置flink的并行度，有哪几种方法？怎么计算一个合理的并行度？怎么验证并行度的合理性？
14. Flink中有哪几种窗口，你这边用的比较多的窗口？滑动窗口你们是什么场景下会用到？滚动窗口没办法满足你们的需求？
15. Flink中的水位线是否了解？怎么触发窗口的，以及它的机制？
16. Watermark这块用的多吗？在哪里用到？
17. 离职原因。

# 德特赛维二面

1. 数据量集群规模
2. 离线数仓的架构，离线数仓的工作和职责？
3. hive的版本，数据倾斜讲一下
4. 你们hive怎么判断小表的？是否有参数可以设置？
5. sparksql执行的时候某个任务就很慢，单不是数据倾斜导致的，该怎么进行优化。
6. 知道推测执行吗？spark的优化
7. 离线数仓的数据质量监控怎么理解的，怎么保证你的数据结果的准确性
8. 你觉得在大数据数仓开发过程中一个良好的开发流程是什么样的？
9. 你们对模型的管理做了些什么？模型开发完之后怎么管理，版本的更迭这些，
10. RDD程序写过没有，问了scala的模式匹配，尾递归，为什么scala的变量不可变的好处
11. 说一个使用flink调度场景，flink两条流怎么join
12. 你们维度数据有没有生命周期？
13. 你们用什么组件监控flink这套架构的
14. 实时项目遇到哪些问题，怎么解决的，kafka那一套讲一遍，flink调优讲一遍
15. 大状态用什么存储的，rocksDB选择第三方文件系统对效率有没有什么影响
16. 布隆过滤器了解吗？

# 中科江南

1. 自我介绍
2. 离线介绍
3. 指标怎么完成
4. 数据重复怎么办
5. oracl这种传统的关系型数据库有了解吗 ?
6. 有跟客户接触过吗？
7. 你觉得你们这种公司内部的沟通和跟客户沟通有什么区别
8. 能接受出差吗?短期的
9. 有什么想问的吗

# 东信北邮信息

1. 对数据优化怎么理解的，实时项目串讲，flink窗口
2. hiveSQL数据倾斜怎么造成的，什么现象，怎么解决的？
3. 怎么判断数据倾斜时任务卡在99.9%时不是僵死而是数据倾斜？
4. spark的shuffle原理和优化，优化后有什么效果？快？快多少？
5. hadoop搭建的时候用的什么版本？具体什么版本？
6. JVM内存加载机制？打住，聊大数据
7. Linux在某个很深的目录下找某个文件，用什么命令？没用过find
8. hbase的hmaster控制什么？
9. 用clickhouse主要做什么？
10. 换人问了，用户画像介绍
11. 标签的变更是怎么操作的？clickhouse中的表结构是什么样的？
12. hive到clickhouse怎么入库的，spark程序的具体代码逻辑是什么样的，用了写什么方法写出的
13. spark的版本是多少？spark项目使用什么构建的，idea，项目依赖管理用的什么？
14. maven中打包时把某些包剔除用的什么？怎么解决jar冲突的？依赖jar包用什么
15. spark的算子有哪些，map和mappartitions的区别，map和flatmap的区别
16. 提问，公司业务是移动旗下咪咕文化一整套互联网新媒体这种，每天数据量PB级别

# 求圣科技

## 一面

1. 介绍
2. 会不会datax，数据怎么传的，他讲了个单词没听清
3. kafka应答机制ack，你们遇到过的问题，还是自己背的
4. 问kafka分区机制，如果（3个区，4个消费者怎么消费遇到什么问题，死掉一个咋办）
5. 有没有遇到数据重复，，，，又问那5个事务怎么写的
6. flume拦截器怎么写的，能说下第一第二次的清洗，拦截
7. flink窗口，怎么触发，举个例子
8. 状态用过吗，说一下，优化呢（大状态调优）
9. shell脚本写过吗，我说了一个启停
10. 说下你们最大的表，多大
11. flink哪些api,interverjoin底层的为啥要keyby，
12. 问了个sql，简单

## 二面

1. 会不会sql
2. 会不会flink
3. 问数据量，他们300T，
4. hive知道不
5. 自学的还是公司做的，你这咋这么多技能点
6. 平时啥活动，我说学习，研究新技术
7. shell会写不，python用过吗
8. 职业规划
9. 业务转换能适应不，数据量大

# 中盾安信

1. 介绍实时数仓
2. flume采集日志的条数怎么确定不丢？
3. Kafka的数据积压，为什么会数据积压？
4. Kafka扩容是选择新增节点还是在原有节点上扩容，如果新增节点，执行负载均衡命令需要多久？
5. Kafka中设置增加吞吐量参数后，还有什么会成为瓶颈？CPU、磁盘、内存，又问100T和4各25T怎么选
6. 实时数仓中怎么进行etl的？详细说一下你们etl怎么做的，对哪些字段做了etl
7. 窗口开了多大，sugar多长时间刷新一次，这两个有什么关系吗？
8. 对数据库了解多深？熟悉存储过程、视图吗？
9. 你们集群规模多大，Kafka每秒多少数据量，flink单个并行度峰值处理能力？flink跑了多少job？多少台服务器上部署了flink？并行度多少？
10. datax是干什么的，数据源是什么，目的地又是什么？

# 德特塞维

## 面1

1. 自我介绍
2. 为啥来北京
3. 说下hadoop
4. mysql元数据，原始数据
5. spark，flink，区别，你们都用的啥
6. 有没有独立解决问题能力
7. 如何看待加班
8. 对象，家庭，薪资期望
9. 公司发展，职业道路

## 面2

boss写的总经理助理加微信，腾讯视频面

1. 自我介绍➕为什么离职
2. 期望薪资
3. 为什么女生要做大数据，考不考虑其他岗位你们公司数据量一天能有多少？
4. 实时监控你做的，那遇到问题怎么做的
5. （6）你确定不考虑其他岗位吗？

# 福卡斯特

1. 人事会大概调查你的公司，问一下你公司的情况，离职原因，为什么来北京，是否打算长期北京发展？
2. 自我介绍以及问所在大学？
3. 给自己java和大数据能力打分；
4. 自我定位与优势，最擅长的技术；
5. 公司产品、数据量以及并发量的相关问题；
6. 产品业务包含哪些？会有直播，录播这种业务？
7. 处理用户数据量是否都集中，服务器大概有多少？
8. 新产品和老产品维护你这边倾向那边？
9. 大数据的技术选型，怎么进行的选型？
10. 你这边怎么保证服务的稳定以及它的影响因数会有哪些？
11. 除了在程序和组件的优化，在硬件，数据库是否会进行进一步选择从而达到更好的稳定性？
12. ClickHouse存了多少条数据？
13. 一个sql题，查出重名的学生？
14. JVM内存模型？内存溢出报错的原因在哪？
15. 外部请求怎么通过服务器框架进行运转响应客户端？
16. HashMap里面放100条数据，初始化应该是多少？
17. Redis穿透了解吗？
18. 锁是否用过？项目中分布式锁是否用过？
19. 后台任务你们是否用过？
20. 你这考的证书对你的帮助是什么？
21. 你对自己未来的定位？

# 赛博云

## 面1

1. flink 优化
2. hive 优化
3. flink定位反压，反压场景
4. flink全数据流一致性保证
5. kafka积压处理
6. kafka怎么有序
7. hive开窗
8. hive数据倾斜场景，处理，定位
9. flink比较难的模块怎么做的
10. flink遇到的问题，怎么解决
11. flink checkpoint 超时解决
12. flink提交各个参数（容量大小，并行度，为什么这样设置），集群规模
13. flink和外部组件交互积压怎么处理的
14. 为什么要DIM，能不能和DWD合并
15. GC日志怎么看
16. flink buffer怎么调，flink taskmanager 个数怎么确定，每个taskmanager设置多少个slot
17. checkpoint和safepoint区别，怎么用的
18. 状态后端有哪些种类，选用种类，为什么选用
19. flink版本
20. hive map reduce 个数怎么设置
21. hadoop mapreduce 怎么写
22. ck监控
23. ck索引种类，原理
24. hbase二级索引
25. redis持久化方式，原理
26. redis基本类型
27. redis用过那些类型，怎么用的
28. redis集群用过吗
29. java finally final用法
30. java 多线程 sychronized是公平还是非公平，可以加在哪些上面
31. spring boot 常用注解
32. python Map原理 能否key为None
33. 集群规模，数据量，每秒数据量，每日数据增加量
34. 实时数据怎么落地？

## 面2

1. 自我介绍
2. 讲熟悉的项目
3. 数据处理（脱敏、解密）
4. kafka乱序
5. java的HashMap 底层、HashTable、多线程
6. mysql隔离级别、索引、b+tree和红黑树的区别、redis一致性（双写双删方面的）
7. zk擅长读、写？ 、zk的CAP
8. hive优化 spark会不会写
9. hql函数的优化 离线数仓解决的难题
10. flink时间语义、内存调优

# 爱普优邦（抖查查，爱盈利）

没有笔试，电话聊完直接线下面试

抖音抖查查，分析直播数据电商数据等，技术栈:spark es flink 。用的阿里云 datawork

团队4人

人事面 大学经历

## 技术面

1. kafka乱序问题怎么解决，数据量太大非要多分区下游怎么处理？ 下游具体怎么实现
2. bitmap的优势劣势
3. 为什么用flink不用sparkstrming，说说区别
4. 实时数据为什么放ck，放mysql行吗
5. 说说项目里你遇到的问题
6. spark实时es给谁用的，用户还是内部
7. 做过什么指标
8. ds不是最近几年才火的吗，为什么用，一开始就用吗

## 项目经理面

# 济南中兴协力

1. 自我介绍
2. 熟悉hsdoop吗，介绍一下
3. zk了解吗，介绍一下
4. scals/java了解吗
5. 说一下你的项目经历
6. 如果一个零基础的人想学习大数据，你会推荐他怎样学习

# 百信大数据开发 - 外包

1. 自我介绍
2. 介绍离线数仓项目
3. （讲项目过程中，会根据所讲内容随机提问，flume挂过吗？kafka的了解？整个离线有没有数据跑失败的情况？）
4. SQL题，字段userid、loc、time，每一行代表一个用户在一个时间点所在的位置，求用户在哪个位置待的时间超过30分钟（每一分钟监控一次用户位置）
5. 实时用到哪些框架？
6. ClickHouse了解多少？为什么用它？引擎？是否采用副本了？

# 国广清科

## 面试1

1. 自我介绍
2. 介绍的实时
3. 问为什么用flink cdc ，如何实现的，用了maxwell为什么还用cdc，他们俩有什么区别？
4. flume到kafka做了什么处理？
5. 数据从kafka到hdfs 文件是什么格式的，怎么转换成hive表的？

## 面试2

1. 自我介绍
2. 介绍项目 离线、实时
3. log到HDFS做了哪些工作 kafka消峰具体原理
4. dataX底层
5. maxwell原理 maxwell断点续传原理 出现故障数据重复怎么解决
6. HDFS小文件处理（我说har归档，问我具体怎么做的）
7. CDC比maxwell好在哪 DWD层做了哪些事

## 面试3

### 一面

1. 自我介绍
2. 介绍实时项目，主要负责哪几层，说一下项目架构
3. 为啥要使用旁路缓存和异步io，旁路缓存和异步io怎么实现的
4. 大状态的优化是怎么优化的，状态太大了有什么影响
5. checkpoint太慢后有什么后果
6. flink的checkpoint和spark的checkpoint有啥区别
7. cdc底层是怎么实现的，配置表怎么放到广播状态中的
8. 平时使用flinksql还是flinkapi，项目中哪里用到了flinkapi
9. flume到kafka怎么搞的，采集时你搭建的吗
10. 你做了哪些指标
11. 你在项目中解决哪些问题，比较有成就感的问题

### 二面

1. 自我介绍
2. 开发语言使用的是啥java吗？工作遇到过哪些棘手的问题，比较难以解决的问题
3. 异步io原理是啥，怎么实现，这个线程池多线程你们开了多少个，怎么设定的
4. ArrayList查找一个元素，不知下标，HashMap通过key获取元素，那个快一点
5. 假设有一个数据ArrayList里面都是整数，然后给你一个整数M，求找出两个数和等于M
6. Linux熟悉吗，我想看服务器有几核cup是啥命令，我想知道现在跑了哪些程序用啥命令
7. 两台服务器，有一台服务器监听了80端口，如何在另外一台机器上看这台服务器是否监听了80端口
8. 内网服务器可以访问外网，但是没有公网ip，那我应该怎么才能访问到这台服务器（我是内网穿透，他又问内网穿透原理）
9. 除了大数据框架，java，你还会哪些哪些语言，你会python吗，es会吗，es的分片是啥
10. 最近3-5年，你职业规划，你觉得你在那些方面有提升
11. 你离职了吗，离职原因是啥
12. git你会吗，新建一个分支会吗

# 托特私募基金（笔试题）

分5类：python、java、mysql、大数据技术（偏spark）、算法（1）

## python

没看；

## java

1. volatile、synchronized两种锁的区别；
2. 给一个文件名和一个字符串，统计文件中出现该字符串的次数，编程；
3. 列举不同线程通讯的方式
4. 用jdbc读取数据库，如何提高读取效率？如何提高更新效率？
5. 读取一几千万的数据得文件，用hashmap加载，如何优化提高加载速度？
6. 有一个10亿条数据的文件，存储用户的身份证信息和手机号，如果某一用户手机号修改了，用java什么类操作？

## mysql

1. mysql主从复制原理；
2. 数据库中常用锁及应用场景；
3. mysql查询语句select执行流程；
4. 写sql，两个有相同结构（id和ttr）的表（A和B）join，若有字段为空值， 按AB顺序取第一个非空值；

写sql，三个有相同结构（id和ttr）的表（A、B、C）union（不去重）； 并增加一列数据来源（A/B/C）

写简洁的sql，十个有相同结构（id和ttr）的表（A、B、C…I、J）join，若有字段为空值，按A、B、C…I、J顺序取第一个非空值；

1. S（DATE,STOCKID,STOCKNAME,AMOUNT,VOLUME）,

IND（STOCKID,股票所属行业ID）

SID（股票所属行业ID，股票所属行业名称）

写sql，建立合适的表索引以更方便的解决一下问题？

写sql，历史销售最高的股票名称？

写sql，股票数最高的行业？

写一个触发器，保证每日各行业股票数不少于10只？

1. 建立以x,y,z为联合索引的表；

从select…from table where x=1,y=2,z=3查询语句中用哪种表索引？

从select…from table where z=1,y=2,x=3查询语句中用哪种表索引？

从select…from table where x=1查询语句中用哪种表索引？

从select…from table where x like“%1%”查询语句中用哪种表索引？

## 大数据技术

1. 阐述spark提交流程
2. checkpoint、cache、persist区别；
3. 写出spark算子及作用；
4. 编程利用spark算子完成以下任务
5. 阐述sparksql、rdd、dataset、dataframe…的区别及计算效率；
6. 什么是宽依赖、窄依赖？
7. groupbykey和reducebykey的区别？

算法一道：没看

# 可之技

1. 介绍上一家工作项目
2. 数仓的作用？
3. 离线数仓和实时数仓是同一套，还是两套？
4. 离线数仓的技术栈？
5. 离线数仓的承担的角色
6. 同步了那些数据，数据源
7. 埋点怎么埋的？
8. 数仓搭建完成后运行过程中的问题？
9. kafka宕机加何排查？
10. kafka数据重复怎么解决
11. 增量还是全量的选择
12. 什么数据是增量同步的？
13. 增量同步了，数振更新怎么办
14. 什么是拉链表
15. 同步过程中，业务那边业务数据库的表结构发生变化，怎么办？
16. 从上家公司的离职原因？
17. 你的期望薪资？
18. 在上家公司的有什么成长？
19. 比较喜欢找一家什么样的公司
20. 面试多久了
21. 你有什么问题想了解

# 国信创新

有个笔试，三道SQL题，很简单，分组聚和、开窗、两表关联

## HR

1. 自我介绍
2. 之前工资、期望工资
3. 对自己的定位和职业发展规划
4. 对他们公司的简介

## 技术面

1. ClickHouse、ES问了一点
2. 其他也没问什么，自己讲一讲框架、组件啥的

# 有咖互动

## 面1

1. 自我介绍
2. 离线和实时整体架构
3. flink的水位线介绍一下、背压、Redis旁路缓存怎么实现的
4. CK介绍一下
5. Redis介绍一下（问了Redis的锁）
6. Flume挂了数据会丢失吗？原理是什么？后期对采集的数据有什么处理
7. 七天内连续三天下单（因为我简历写了）
8. 总之就会根据简历写的点去问；并且讲项目过程中，随时穿插会问一写细节和组件，主要就看自己怎么讲项目的
9. 还有几道他电脑上的题，直接展示，说出思路就可以
10. 注册表（id、dt）、登录表（id、dt），求每天的前三天注册的用户中登录的数量
11. 一个字符串，其中包含各种成对的<>、（）等，要求找出字符串中都是成对出现、且是嵌套关系的

## 面2

1. 说说离线数仓架构
2. 说说flume，组件之类的
3. 建模分几层
4. 维度建模，每层表都有什么
5. 怎么建的模
6. 问了谓词下推
7. hive底层
8. hive的分区，也就是hdfs分区
9. 拉链表实现
10. hql函数，开窗函数
11. 分桶表分区表外部内部表
12. kafka的leader选举，kafka优化
13. redis写入数据的api
14. hbase怎么均衡region
15. linux查找文件，自带的内部调度器
16. 手写sql题，会有侧写或者行转列的一些函数

## 面3

1. Flume读log文件，怎么知道读取完成了,读完的文件有什么变化2
2. Flume的source有哪几种了解哪几种
3. Flume时间字段在event哪里存储
4. Flume有几台
5. Datax怎么配置的，在hdfs中配置了什么
6. kafka中partition存什么格式数据
7. Kafka数据存的数据大小
8. 维度建模流程
9. 数仓哪几层用的维度建模
10. 用没用过Linux的调度
11. Linux的命令有哪些
12. 动态分区了解吗，拉链表怎么做的
13. 你们部门多少人
14. redis了解吗
15. Clickhouse了解吗

## 面4

1. 自我介绍
2. 项目串讲
3. flume的组件
4. hadoop架构
5. kafka架构
6. 事实表有哪些
7. 维度表有哪些
8. 业务总线矩阵怎么做的
9. 维度建模过程
10. 拉链表怎么做
11. 你做过哪些指标
12. 指标是谁提出的
13. linux 常用命令
14. 3个sql 列转行 行转列 topn

# 军尊

## 面1

1. 自我介绍
2. 介绍一下用户画像
3. hivesql sparksql flinksql 各自的应用场景
4. 如果说我们想定义的规则类标签有时间维度的概念 比如说我们统计10天 但是数据是11天才来的数据 那么我们怎么进行处理
5. 如果数 对用户标签的话 那么中间几天这个用户可能没有操作行为那么原有标签是消除还是保留 还是其他处理 你们是怎么做的有没有额外的业务操作或者大数据技术方面的处理
6. 项目中做过哪些优化 技术或者框架
7. flink做过哪些优化
8. flink反压
9. flink下游有没有可能消费能力不足导致出现反压
10. 介绍kafka
11. kafka新版本有什么特点，为什么要去掉zookeeper 好处是什么，同类型的消息队列有哪些 如何比较的 为什么选择kafka
12. 建议说多关注大数据技术的发展，多横向的比较框架，选出最适合的

## 面2

1. 自我介绍
2. 技术栈介绍
3. 不用flume可以吗
4. 阿里云的四层能不能简化
5. 做过哪些优化
6. 为什么选择kafka
7. 用没用到zk
8. 不用zk可以吗
9. Kafka现在不是可以不用zk吗

## 面3

1. 项目中遇到的问题
2. 介绍下用户画像
3. spaksql hivesql flinksql区别
4. Flink和spark区别
5. 实时和离线的区别
6. 小文件怎么处理的
7. 做了哪些优化

# 芯盾科技

1. 自我介绍
2. 介绍项目
3. 使用java实现wordcount
4. Kafka实时中都做了什么
5. 用的那些web（实时提交方式）
6. Redis存的什么
7. Hbase 的Rowkey怎么设计的

# 云和互动

1. 自我介绍，
2. 根据简历往下捋，集群规模，磁盘大小，服务器选取，参与了哪些集群建设
3. Hadoop有哪些组件，hdfs读写流程和yarn的运行过程，hdfs操作命令
4. 分区和分桶表区别
5. 有没有遇到过磁盘坏块
6. 有没有遇到过节点数据不均衡
7. hive调优，小文件处理
8. 8Rdd,ds,df 的转换关系，
9. 宽依赖和窄依赖
10. 有没有使用过算法提高离线数仓或实时效率
11. 用过哪些设计模式
12. 自定义函数在哪应用的

# 国汽大有时空

## 面1

1. 最近做的什么项目，我说flink，介绍下项目的实施过程。
2. flink 算子链合并的条件，
3. flink 算子间的通讯方式，
4. 状态有哪些？存储位置
5. checkpoint和savecheckpoint区别
6. checkpoint参数配置，如果并行度改变，checkpoint 可以直接恢复吗？
7. 实时有多少指标？统计过的最大的指标是哪个？
8. task和subtask、slot的关系

## 面2（偏运维）

1. 讲讲项目 我讲的离线 说kafka有几个问题 留的坑 然后他就问了我那几个坑
2. kafka一个topic几个分区 数据量大的时候几个分区
3. hdfs版本 四个进程 具体都是干啥的
4. dn nn 通信用什么 具体干啥的
5. clickhouse 单点还是集群 几个备份
6. redies单点还是集群
7. 监控器了解嘛
8. 中间器
9. 数据到hdfs 用什么组件查看
10. yarn读写流程
11. linux3剑客 都是干啥的

## 面3（实时）

1. 自我介绍
2. 问实时项目 问维度数据的每日数据量多少？为什么不用mysql
3. 实时项目用的什么架构？什么是kappa架构和lambda架构，区别是什么？有什么优缺点
4. ck的一致性，去重底层原理？（我回答主键，只保留最新的）除了主键还有什么？怎么做的？
5. flink的算子链（除了并行度相同），告诉我还有keyBy之后有重分区有shuffle
6. subtasak和task的区别

# 橙啦教育

1. 自我介绍
2. 介绍离线
3. Dws是最终指标吗
4. 离线核心框架是？
5. 有没有指标管理平台
6. 项目经理通过什么给你们指标定义（需求文档还是白皮书）
7. 有没有基于指标做过多维度查询
8. 介绍用户画像
9. 宽表存在哪
10. 标签是如何分类的
11. 用户偏好标签的业务逻辑是什么样的，怎么实现的
12. 画像的上层应用是什么
13. 数据校验怎么做的
14. 数据治理怎么做的
15. 数据安全怎么做的
16. 跑批故障保障怎么的
17. 告警的级别是如何设置的
18. 有没有做过元数据管理
19. Flink 为什么用api不用flinksql
20. Hive 优化
21. Flink state TTL
22. State清除策略
23. 常用的是哪种state
24. 反压的原因 影响 怎么解决
25. 水位线多对一 取哪个 原理是什么 咱们项目中 多对一 一对多 一对一 哪个用的多
26. Flink 端到端
27. 剩下的就不是技术问题了 还问了薪资结构

# 华熙生物

1. spark 和 flink的区别
2. flink的核心
3. flink的架构
4. flink的api分几层
5. flink的时间语义 问flink的处理时间是不是服务器的当前时间
6. 说说数仓的分层
7. 为什么用orc，orc是列存还是行存
8. flume拦截器
9. 问了一下数据量，还有数据来源是app 小程序 还是什么
10. 问对哪方面技术比较感兴趣
11. 手里几个offer 都是什么价 公司是做什么的

# 佰钧成

## 面1（vivo外包）

vivo那边负责面试，佰钧成只负责拉人，一共三面：两个技术面，一个hr面

主要去做flink指标开发和系统的升级

## 面2

1. 自我介绍
2. hive on spark中sql如何转换为spark任务的？
3. hive on spark对spark做过哪些调优？
4. kafka的offset是全局唯一的吗？
5. kafka分区数为什么只增不减？
6. spark实时系统怎么维护offset的
7. join操作如何才能避免shuffle？
8. rdd、df、ds之间的关系及转换？
9. flink实时和spark的实时的区别和共同点？
10. flink的checkpoint保存机制
11. flink使用中遇到过哪些问题？
12. spark的stage如何划分任务的？spark如何查看任务数量？我回答的依靠血缘关系，她说这可以看，但是她说问的不是这个，就想看当前stage的任务数？
13. hbase在哪用的？我说的是flink维度表，又问了数据量多少？了解多少？底层原理是啥？做过哪些优化？hbase存入数据的时候变慢了，该如何处理？
14. elasticsearch做过哪些调优？是一个指标一个索引吗? 指标统计是在es里面完成的吗？
15. 离职原因？
16. 遇到问题一般怎么解决的？
17. 上家公司遇到你没接触过的技术是给你时间学习？还是你自己私下加班学的？
18. 对未来的规划，自己有什么优缺点？
19. 最后问我之前他们出的题怎么样？让我写一个算法，她来制定，当场写。链表反转。

## 面3

1. 自我介绍
2. 为什么spark实时换成flink实时？
3. fink挂掉如何保证一致性？
4. flink程序升级，或者任务数变了，还能从状态中恢复吗？ Savepoint
5. 你一般遇到问题如何解决的？
6. 你们服务器用的什么系统？linux。你编写jar包怎么传到服务器的？为了服务器安全。
7. linux的常用命令？
8. xshell用过吗？xshell用的场景是什么？数据传输怎么传的？
9. flink或者spark是谁来维护的？你们数据量有多大？
10. flink或者sparkstreaming有没有优化的例子具体介绍下？介绍了下各省份订单量统计和外部数据库读写优化？
11. 对spark有什么优化？
12. 现场用本地idea做一个简单的树的遍历。

## 面4

机考：

二分查找

青蛙爬台阶

14道大数据相关的选择题

面试：

1. Spark join 不产生shuffle的情况 ? 除了mapjoin 还有呢？ Bucket join 提前用partitioner分区
2. Flink checkpoint以及遇到的反压问题，如何定位解决反压的？
3. Flink keyby的键值是string 型，之后改为int 会出现什么情况？
4. 每天的数据量是多少，Flink搭建开发用了多久？公司的app叫什么？
5. es 有用过吗？es mapping 与 setting 设置的参数是什么？ Shard 副本数 别名 分词器 字段类型
6. kafka 中的offset 怎么获取？kafka-consumer-groups
7. 手写几道算法题：先写一个链表反转

## 面5

1. flinkcdc原理
2. 单例怎么实现
3. 为什么用到了clickhouse

# 多益网络（游戏公司）

1. 自我介绍
2. 问了下之前公司大概是干什么的
3. 问用户体量（就是大概有多少用户，数据大概有多少）
4. 为什么离职
5. 介绍一下你最熟的一个项目，你在里面主要做了什么
6. Kafka架构了解吗，描述一下，然后中间根据我说的问了Kafka的高吞吐
7. Kafka的高效读写方面问题
8. Kafka零拷贝了解吗
9. Kafka还做了哪些优化
10. 你之前说到小文件，你能说一下小文件的危害有哪些吗
11. 问hadoop部署大概是什么样的，几台服务器
12. 我们有部署高可用吗，大概怎么配置的
13. 七天连续3天的需求怎么实现的
14. 离线数仓有遇到数据倾斜吗
15. Spark了解吗，提交流程大概是什么样的，RDD了解吗，用过吗，groupbykey和reducebykey
16. Hbase和clickhouse对比一下
17. 来了个开放问题：
18. 如果我有一个文件，里面主要是些数字，数据有几十亿设置几百亿，我想找其中数字最大的100，怎么做；

# 铂原科技

1. 三年经验的问题
2. 离职原因
3. 介绍用户画像
4. 问怎么处理3亿用户 1000个标签的问题 1000标按照我们的画像处理 宽表有1000多列 可能表就崩溃了 怎么解决这个问题
5. 有没有linux的使用 能否使用CDH
6. 介绍做的比较有难度的标签
7. 你们的标签一维度的标签，可否做二维的标签 二维的标签应该怎么设计宽表
8. 离线数仓都做了些什么 表是不是你设计的
9. 数据源的问题 日志数据是什么样的结构
10. springbroot的了解有多深 能不能进行现有spring boot代码的修改
11. 我们的标签是基于sql制作的 有没有什么办法能让不会sql的同事也能进行标签的只制作

# 时空云电话

## 面1

1. 为什么是文件块大小是128m？Hdfs中平均寻址的时间？
2. Hive做了哪些优化？
3. Hive中窗口函数RANK（）DENSE\_RANK() ROW\_NUMBER() 中的区别？
4. Hive中四个by区别？
5. Hive中行转列怎么做的？
6. Hive中字母小写转大写？
7. CONCAT\_WS、COLLECT\_SET的区别?
8. hive中如何取两位小数？

## 面2

1. 三天碳排放
2. 事实表是什么？
3. flink背压什么现象

## 面3

1. 自我介绍
2. 介绍离线
3. 主要负责的数据域
4. 三种事实表的区别
5. 拉链表的使用（今天10.31号，怎么看10.11号日期的数据，假如10.11号有变化怎么看）
6. hive的sortby orderby区别，自定义UDF函数，介绍自己写的UDF函数
7. hive的优化、数据倾斜
8. spark宽窄依赖 union是宽还是窄？
9. RDD五大属性
10. reducebykey、groupbykey、aggregatebykey区别
11. 缓存和检查点区别
12. Repartition和Coalesce区别
13. spark的Shuffle区别、bypassShuffle怎么触发的
14. zk的选举机制
15. java的GC

# 致开源

1. 介绍项目 离线
2. 数仓分层 各层的作用
3. 出现过任务失败的这种情况吗 也就是框架优化和hive优化
4. 离线的模式？T+1
5. 调度任务用的是
6. 依赖关系 用的是什么元数据管理
7. 介绍实时
8. Spark 和 flink 区别
9. Spark 模式
10. Redis 是用来做什么的
11. Kafka数据积压
12. 数据丢失
13. 画像
14. Clickhouse 好处
15. 数据迁移会做处理吗
16. Hive到clickhouse 为什么没用datax这种同步工具来同数据
17. 复杂的标签 都是写sql
18. 离线和标签不能合并使用
19. 项目中遇到问题 反压和数据倾斜
20. 做没做过数据质量管理

# VIVO外包

## 面1

1. 自我介绍
2. 介绍一下你们离线数仓项目的组件
3. 请你说一下hive的优化
4. 为什么要做实时数仓
5. 工作中有没有遇到flink反压 及 数据倾斜等问题，是如何解决的？
6. 为什么使用clickhouse
7. 请你说一下你所遇到的离线数仓的困难点
8. 使用flink api 代码过程中遇到过哪些问题？
9. spark的任务划分
10. dataframe 与 rdd 的区别
11. 如何调整spark task的数量
12. 小文件有哪些危害（两个方面）
13. 谈谈你对kafka的理解
14. 谈谈你对checkpoint的理解
15. 为什么要用hbase保存维度数据，了解hbase的rowkey吗？hbase读写流程了解吗？

## 面2

1. 离职原因
2. 自我介绍
3. SQL 题，字段 userid ，dt（登录时间）一个用户 如果登录了每天只有一行数据，求出30天内 任意7天内 有任意3天登录的用户不需要连续。
4. spark里面的join策略有哪些（就hive on spark hive的sql转为RDD算子去分析
5. 那些join 会不产生shuffle 和那些会产生shuffle
6. spark的并行度 设置方法有哪些
7. 每个stage的分区数量由什么决定的
8. 每个stage最后一个算子的分区个数是由什么决定的
9. 用过scala么
10. 实时项目的负责那些
11. 每天的数据量
12. 实时数仓用sql还是api
13. 什么是flinkCDC
14. redis 的缓存过期设置，为什么这么设置
15. flinkCDC是直接监控的数据吗（我说了一下维度层的建设过程）
16. 广播流与主流链接，主流出现积压，广播流没办法即时更新状态怎么办
17. flink反压
18. Hbase的了解
19. es 有了解吗
20. 你有什么想问我的

主要招做实时的，他们离线主要是用scala写的，用的层次很深。

# 能链集团

1. 自我介绍
2. flink实时做了多长时间？
3. flink用的sql还是其他的？
4. 实际做过的指标有没有遇到反压，怎么解决的？
5. flink的内存机制？
6. 电商遇到退单在实时中怎么处理的，我说开了一个新的退单流。他问假如想统计当天的实际交易额怎么做？
7. 你们的实时数据量有多大？实时中业务数据量有多大？
8. 数仓用的什么架构，做过哪些hive优化？
9. orc和parquet的区别？
10. 手写sql，统计连续10天登陆的用户？

# 环球网校

1. 介绍项目
2. kafka怎么做的分层
3. 是存到不同的topic
4. ck有搭建过集群吗
5. 组件是用什么安装的？apache
6. 用什么管理的 ？cdh吗
7. flink 集群模式 和yarn 有什么区别
8. 用过cdh吗
9. 了解过mery吗
10. cdh能管理哪些组件
11. pstio 和 hudu
12. hbase 作用和适用场景
13. phoenix 建表吗
14. phoenix 默认端口
15. namenode 和 datanode
16. hadoop 高可用怎么实现
17. hivesql 窗口函数 实现topn 用哪个函数
18. bitmap使用场景
19. ck宽表 有扩充的情况 会改结构吗
20. redis 默认的db 有多少个
21. 离职原因
22. 集群规模
23. impala
24. 离线有bi系统吗

# 贝壳外包

1. 说一下你最近做的一个项目把
2. 为什么需要dws层，为什么不可以从dwd直接到ads层？
3. dwd为什么不可以去掉？
4. 业务总线矩阵怎么构建的？
5. 数据量是多少？
6. 什么样的数据放在dim表，什么数据放在dwd，dws？
7. 为什么用户放在dim表里？
8. 什么是退化维？
9. 指标数据都是通过dws出的吗？
10. 如果你们跨了各种数据域的数据放在哪里？
11. 手写拉链表sql
12. 手写用户首次登录的城市sql？
13. 了解三范式和维度建模吗？
14. 说一下三范式
15. 说一些维度模型

# 青岛群之脉

1. 日常写什么代码。
2. java基础范型
3. hive工作原理
4. 调优经验
5. 对spark了解吗？执行过程
6. 项目中遇到的难题，如何解决
7. 翻译英文，（一段话）
8. sql窗口函数
9. shuffer优化
10. clickhouse 优缺点
11. hive on spark怎么做的
12. 项目 dwd dws区别
13. 数据库
14. 前端可视化工具
15. Linux常用命令
16. 代码如何管理
17. Git命令
18. 编程方面的书籍
19. spark有哪些概念
20. 技术短板
21. 职业规划
22. 加班态度
23. py写过没

# 先智数元

1. 哪些自己单独完成
2. 业务数据过滤关联字典表
3. 双流join怎么匹配
4. dim层配置表，怎么写入hbase
5. interval join时间范围
6. cep
7. 端到端一致性
8. 8．检查点
9. 下游怎么实现幂等性
10. 异步io?数据库支持吗？
11. 旁路缓存双写？
12. driver端执行，ex端执行？
13. 13flink redis在哪里写代码
14. foreach在哪里执行
15. 算法

# 中信百信银行

## 面1

1. 数仓结构；
2. 各层有哪些表；
3. dwd->dws中间还有没有分层；
4. 离线数仓怎么保证数据质量；
5. 怎么验收数据；
6. 哪些数据域，各个数据域包含那些表；
7. 遇到比较难解决的建表问题

## 面2

1. 连续三天碳排放小于100
2. 用户连续登录最大天数
3. 纬度建模四个步骤
4. （我给个提示你啊…，选择业务过程）
5. 事实表分哪些
6. flink检查点流程

## 面3

1. 说一下你们的数据埋点是怎么实现的
2. hive的调优
3. sql条件where和on的区别
4. 表与表的连接经常习惯用那个，如果遇到rightJoin的话会转成leftjoin来使用吗
5. 通过sql实现在用户访问表里查找连续六天以上访问的用户数量

## 面4

1. 自我介绍
2. 在项目中遇到了哪些问题
3. 二进制怎么转十进制
4. 离线数仓是否做过应用层（应该就是做过哪些指标），具体怎么怎么实现
5. Superset具体怎么使用
6. 有什么问题问面试官的

## 面5

1. 小文件怎么解决
2. 长期使用集群还是会产生小文件，怎么彻底解决
3. har归档文件怎么用
4. hdfs写流程
5. 写流程中挂掉了datanode怎么办，挂了1个和挂了两个该怎么办（3个datanode）
6. flume中设置30min和128M，如果是30min才写的文件，他长久之后还是小文件，你怎么解决
7. 影响maptask个数的因素是什么（除了切片个数）
8. zk有几种类型数据节点
9. kafka写文件流程
10. Kafka如何保证不丢失数据
11. Kafka中ack=-1，存活的副本挂了会有影响吗
12. flink几种部署模式
13. flink在yarn中有几种提交任务模式
14. per-job和session的区别
15. flink几种时间语义
16. f两阶段提交
17. flink消费Kafka如何保证一致性
18. Kafka的幂等性是如何配置的
19. flink向hive中写数据如何保证幂等性，写到一半挂了怎么办（30分钟的任务到15分钟挂了怎么办，已经写了一亿条数据，）
20. 检查点会阻塞数据吗
21. 双流join注意事项，join不上的数据怎么办
22. 业务场景：两个系统的数据，一个数据是汽车订单，一个是这个人的征信数据，在他购买之前对这个人征信数据检查，你有什么想法
23. hbase读数据原理
24. hbase基于什么机制可以快速定位数据
25. hbase热点数据有什么解决方案
26. 线程池用过吗
27. 说一下乐观锁和悲观锁
28. 最近项目架构，以及数据流转
29. flink出现反压是怎么定位解决的
30. 查看linux资源使用情况的命令，查看剩多少内存的命令。
31. 有啥想问的。
32. 你写过SpringBoot吗

## 面6

1. 初始化线程池参数
2. post和get请求方法
3. 线程和进程区别
4. FLink-per-job和application区别
5. Flink session和per-job区别
6. session和per-job，application的client在哪里
7. 保证flink端到端一致性
8. flink checkpoint机制
9. olap数据库用了哪几种
10. spark 的shuffle及底层差别
11. 职业规划
12. 有什么想问的？

# 上海玛驷众智能科技有限公司

1. hive数据倾斜遇到过吗，怎么解决的
2. 手写topNsql（注意会屏幕共享，桌面上不要放见不得人的东西）
3. 数据采集整个流程
4. FlinkCDC版本（要用X，X会锁表）
5. 第一级flume用的什么source
6. taildir source的断点续传底层原理
7. kafka数据积压如何解决
8. 2级flume的小文件问题
9. 数仓分层
10. 用过flinkSQL吗，在哪儿用的
11. ADS层文件存储格式
12. kafka有多少分区

# 航空信息股份有限公司

1. 简单做个自我介绍
2. 说下hdfs文件系统缺点3点
3. 说下MapReduce实现jion思想
4. 用户拉链表怎么实现
5. hiveSQL和MySQL有哪三点不一样
6. 说下hbase如何删除表
7. 说下对hbase中数据使用SparkSQL分析
8. 三个词总结下自己

# 锐捷

1. 最近做过啥项目
2. 说说flink提交参数都有哪些
3. flink在集群里面怎么部署的
4. 设置了多少并行度
5. kafka多少分区
6. 为什么设置kafka分区数40
7. 还问了DS版本

# 克拉星球，有咖互动

1. mr流程及优化
2. yarn调度器
3. 维度模型
4. 统计过哪些指标
5. Hive的是个排序
6. Hive函数 3个rank,over里面有哪些东西
7. 手写七天内连续三天登录
8. 说一说kafka
9. zookeeper的作用
10. 数据一致性
11. Kafka事务加幂等性
12. Zookeeper选举机制

# 元拓公司

1. 讲项目
2. 离线自己做的吗，是你们公司自己用的吗，有自己的网站吗
3. 采集离线实时是一套吗
4. 讲一下采集
5. 小文件产生的原因
6. 离线数仓面向那些业务
7. 最终的业务数据放在哪里
8. mr shuffer 流程 调优的点

# 中软信息系统工程

## 面1

1. 讲一下实时项目，带着业务讲
2. 每层怎么做的，有哪些事实表，那些指标
3. MySQL索引优化

## 面2

1. 自我介绍
2. 离线数仓项目介绍
3. 为什么要分为5层
4. datax架构
5. sql调优
6. sql题：
7. 有哪几种方法去重
8. 排名函数有哪些
9. 取前20%的数据有哪几种方法
10. 用户留存率
11. 侧写函数实现

# 上海博彦科技

## 面1

1. 用什么组件开发
2. 创建线程池的方式
3. java三大特性
4. 根据三大特性，开发过什么
5. hashmap结构
6. 做过后端开发吗？
7. 说下隐式转换
8. 说下闭包
9. 传参方式有哪些
10. spark出现的oom场景
11. 共享变量有哪些，原理

## 面2

1. linklist和ArrayList的区别
2. list和set的区别
3. hashmap的底层
4. String s="abc"和new String对象的区别
5. 多线程你了解吗
6. 创建线程的方式
7. 对线程中锁有了解吗？底层的实现
8. 锁加级了解吗？
9. Spring boot中每层的注解
10. Flink 的checkpoint的实现原理
11. 滑动窗口和滚动窗口的区别
12. watertmark是干什么用的，它的实现原理
13. Flink里面的状态类有哪些？
14. HBase使用过程中如何避免热点的，热点问题是什么？
15. （15）mom store在Hbase中的作用？
16. hdfs的写数据流程？
17. kafka中生产数据的流程，数据丢失，ack的不同值的含义？

## 面3

1. flink一致性
2. 检查点一定能保证一致行吗
3. 窗口join源码看过吗
4. flink TTL知道吗：
5. flink 回撤流知道吗
6. flink用的api 还是sql
7. spark数据倾斜怎么办
8. mr的数据倾斜
9. 离线没问多少 溢写理论知识 建模啥的 常规问题

# 京东

## 面1（外包）

1. hive针对shuffle的优化？
2. hive的shuffle和spark的suffle的区别？
3. repartition和colesce区别？repartition是在哪儿执行的，shuffle内还是shuffle外？为什么
4. 去重的几种方式？(我回答的是distinct和group by，要三种实现方式)
5. 两种方式实现前百分之10统计
6. 连续一天登录，连续三天登录？连续间隔n天登录？
7. Flink做过的优化？
8. saprkStreaming和flink在处理业务的时候有什么场景的区别？举例说明
9. 离线数仓的统计结果和实时数仓的统计结果是如何比对的？差多少？
10. 多流join你用过吗？你们是怎么用的？
11. 用过窗口函数的区别，自己实现过窗口函数吗？
12. flink CEP怎么用?
13. 7迟到数据怎么处理？依据什么调整窗口时间？
14. interval join用法？举例说明你们interval join使用的场景?

## 面2（外包）

1. 自我介绍
2. 介绍最近项目(说的实时)
3. 用户画像怎么把宽表存储位一张bitmap表？具体怎么实现
4. bitMap最多能存储多少位？
5. bitMap底层原理？
6. 对spark怎么理解？工作原理
7. 为什么spark是基于内存计算？
8. 对flink怎么理解？
9. java多线程？
10. HashMap怎么理解？底层为什么使用数组加链表存储？
11. Mysql数据库的锁怎么理解，索引使用

## 面3（外包）

1. 自我介绍
2. 介绍一下实时 数仓项目
3. 端到端一致性
4. PV.UV具体怎么统计，我按照每层数据如何处理说的
5. 离线shell脚本怎么写的
6. 用过什么语言，不太会就别说他深挖
7. 用纯Java实现一个问题

// 一万条数据 tab键分割。 第一列是设备id， 第二列浏览页面

// 输出1： uv

// 输出2： top10 pv 的 设备id分别是什么

[(90条消息) Java实现WordCount(多线程和单线程)\_korry24的博客-CSDN博客\_java实现wordcount](https://blog.csdn.net/korry24/article/details/117532756)

## 面4（外包）

1. 数据建模理论
2. 数据装载脚本的目的，为什么进行数据装载
3. 说说维度退化
4. 用户画像做了那些事，读取MYsql的标签表怎么设置的？
5. 和业务人员怎么沟通需求，沟通那些需求？
6. 数据校验过程怎么做的？最终指标提供他们对不对怎么判断？校验机制
7. 作出的和实际业务差异比较大的时候怎么办？
8. 你在项目中是主要做沟通的吗？
9. 实时数仓dws层宽表聚合，clickhouse的引擎？
10. 1clickhouse的字段怎么设置的？
11. sql计算题：

表users：两列 name citys

原表： zs 上海

  zs 北京

  ls 杭州

结果   zs 上海 2

  zs 北京 2

  ls 杭州 1

# 东方国际

1. Hive AB表关联，A表10条 B表5条，关联后怎么有12条
2. 讲一下hive优化
3. Java
4. 维度建模方法论，每层干什么
5. 实时熟吗?怎么办证数据准确数据及时

# 柯莱特

## 面1（外包美团）

1. 离线串讲
2. 日活，年活、数仓数据总量
3. 部门人员划分、日常工作内容
4. 数仓数据质量如何保证（上线前、上线后）
5. 负责哪部分内容（数据域）
6. 最细粒度，概念

## 面2

1. MapReduce计算和流计算有什么区别
2. 种计算分别适合什么场景
3. 几种流式计算是怎么理解的，spark，flink，storm
4. Transform和action有什么区别
5. 了解类似于像dinky这种开源软件么
6. FlinkCDC用在什么场景，解决什么问题
7. 维度表增加一列需要怎么办，flinkCDC能同步过来么
8. 源端是mysql是直接读的数据库还是binlog日志
9. Zookpeer收敛性，用的什么算法做到收敛性（一种一致性算法）
10. 哪些项目是你做的
11. 数仓的分工，维度层起到了什么作用
12. .数据拿过来数据质量集合在哪里做
13. 宽表是在哪里设计的，怎么使用如何加工
14. 宽表在设计的时候未来修改了有什么影响，如何解决
15. Maxwall和DataX在使用时是什么场景，有什么优缺点
16. 这两个组件高可靠是怎么保证的，能做集群么
17. Maxwall断点续传能保证么，底层通讯要加密么

# 四川华为运维

1. hadoop默认副本
2. 副本如何写到几个datanode
3. flume到kafka到flink到hbase，9点半后，我数据查不到了，原因？
4. 某几个datanode磁盘空间不足的原因
5. 现在提交不同的应用到hdfs，有的要2副本，有的要3副本，怎么指定？

# 中华财险（外包）

1. HDFS读写流程
2. yarn提交
3. 三个调度器，你们使用的那些
4. 用户拉链表实现流程
5. 数仓建模分几层
6. flink迟到数据怎么办？
7. spark有shuffer的算子。
8. spark有排序的算子
9. spark-shuffer了解吗
10. hadoop优化，shuffer优化
11. hivesql优化
12. 你为什么统计个省份下单量和下单总额，指标给谁使用，要做什么
13. 统计个省份下单量和下单总额，用到各层哪几张表？
14. 一个sql

商品交易表

商品id 交易日期 当日总销售金额 当日总销量

求连续七天销售商品id

# 中科特瑞

1. 介绍一下自己离线
2. kafka有多少分区
3. 数据量有多大
4. 数仓建模理论
5. 能不能手写斐波那契数列算法
6. Java用的多吗
7. 手写连续三天登录，两天存活
8. 平时喜欢干什么

# 中科软科技股份有限公司

## 面1

1. 在项目什么地方用到redis
2. 怎么保持redis与hbase数据一致性
3. flink怎么保持数据一致性
4. 在项目中，你具体负责哪一部分
5. 做离线是基于什么库
6. 做实时的时候比较有挑战的地方？
7. flink断点续传
8. kafka数据保存时间短，怎么实现断点续传
9. kafka多个主题的数据需要join
10. flinkSQL在join时数据迟到怎么处理的（你说的只能解决10s内的迟到情况）
11. kafka一个topic变动，一个不变，一条流会有大量数据过来，flinkSQL怎么关联
12. kafka分区增加的话，是要重建topic吗，数据怎么流入新分区？

## 面2

1. 自我介绍
2. 说一下离线的项目
3. 数据倾斜又遇到过吗
4. 日志采集flume是否加入拦截器 两个拦截器 拦截器用什么语言实现的 对java熟悉吗 拦截器作用是什么 json不完整怎么判断的 （我说忘了）现在来说怎么判断json是否完整
5. kafka消费者是拉取还是推送 怎么知道消费者一直是存活的 是有数据的
6. sql题
7. 还有什么想问的

# 博亚君杰

1. hbase 部署在十台集群上面，挂了一台，数据会丢失吗？
2. 维度建模
3. 你负责那一块？
4. flink你写那块功能，讲讲？
5. 项目中遇到的问题
6. 你还有啥问题

# 速通科技

## 面1

1. 讲一下项目(数据采集他们还没做后期考虑做没有要问我的)
2. 纬度分层，为啥分五层，每一层都做了啥
3. sql题:

找出高速公路收费站同一收费口对同一辆汽车进行两次扣费的数据

三分钟内同一车道扣费的车辆数

## 面2

1. 讲一下你的项目
2. 为什么要用maxwell。
3. sql

# 浙江御安

1. 介绍一下flink项目
2. flink反压怎么解决
3. flink检查点说一下，
4. 分桶表与分区表的区别
5. kafka数据积压了怎么办？
6. hive的排序方式
7. hive的优化，还有你遇到的数据倾斜，groupby join
8. 离线数仓做了多久，都是几个人参与的，你们的服务器有几台。
9. 未来大数据职业规划。

# 军民融合舰船装备

1. 3-5分钟自我介绍
2. 熟悉哪些组件
3. 对数据治理这块有了解吗
4. 了解过mpp数据库吗

# 外包速科

1. 物化视图
2. clickhouse 副本机制
3. olap 数据库和oltp 数据库的区别
4. 为什么搭数仓，MySQL加磁盘也可以做
5. clickhouse 的replicingmergetree 会有重复，如何解决去重
6. lead函数
7. row between 函数

# 亿达信息

1. 介绍离线项目
2. Maxwell怎么同步的
3. 漏斗分析你们怎么实现的
4. 外部表和内部表
5. 小文件的处理
6. 数据倾斜
7. 第二级Flume里HDFS Sink配置小文件的参数
8. 怎么把数据写入ODS层
9. insert into和insert overwrite的区别
10. 怎么删除ODS分区里的污染数据
11. maxwell和DataX怎么配合工作的
12. Maxwell工作原理
13. Flink怎么处理乱序、数据
14. CEP

# 合众

1. 讲一下离线数仓吧
2. hbase在你离线数仓中怎么用的？
3. java springcloud 了解吗
4. 为什么用海豚调度器？为啥不用java中的querz
5. 你们的指标 ods层到dwd，怎么查的巴拉巴拉记不清了
6. flink怎么处理乱序数据

# 上海智租物联科技有限公司

1. 19年毕业？不是读大二吗？（出来实习了）
2. 说下你熟悉的项目，并统计过哪些指标
3. hbase的rowkey设计原则
4. 你自定义线程池为啥自定义，为啥不用java内置
5. FlinkCDC锁表问题说一下
6. 怎么处理Flink数据乱序和迟到的，极端情况下如何处理
7. 离线建模方法说一下
8. 说一下JVM结构

# 灵信互动外派中金证券ETL岗位

1. ETL的三个阶段：分别是数据抽取、数据转换、数据加载
2. 拉链表的实现
3. 如何保证ETL的数据一致性
4. 表的连接方式有几种
5. 说一下NVL函数
6. 说一下全量抽取和增量抽取
7. 日志数据分几类

# 百融云创

1. 手写熟悉的设计模式
2. 7天连续登录
3. 行转列 字段类型 班级，学生姓名（个数不定） 成绩
4. 二叉树前序遍历
5. 二分查找查找

# 新桥信通

1. 讲一下你熟悉的项目
2. 拉链表怎么做的以及装载思路
3. 你们全量同步怎么做的
4. maxwell会一直监控binlog文件嘛
5. 你们的binlog文件保存多久
6. Oracle用过没
7. datax怎么用的
8. 你写一个指标要多久
9. 你在项目里负责什么

# 苏州盈天地

1. 采集到hdfs中小文件怎么处理的
2. mr中可不可以只用map 或者只用reduce
3. mr shuffle 和spark shuffle区别
4. spark 中的shuffle算子 join 会shuffle吗
5. hbase了解吗
6. clickhouse了解吗
7. flink了解吗 说了一堆 他说我们用不到就不问了
8. java 中的集合说一下
9. hashmap 的 jdk1.7 和jdk1.8的区别
10. arraylist说一下

# 招商新智（就一面）

1. 讲一下项目，我讲了15分钟
2. 怎么知道你们的数据符不符合要求，
3. 说一下数仓建模怎么建的，五层分别作了什么。
4. 然后就聊了他们的情况

# 东方金信

1. 介绍一下实时项目？
2. 问我mysql的binlog在哪开的？
3. 说到可视化的时候，问我mapper干啥的？
4. sql写的怎么样？
5. 问我在公司学到了什么？
6. 数据峰值多少？带宽多大？
7. 搞了哪些优化？

# 九章大数据科技

## 面1

1. 自我介绍
2. 说一下实时项目
3. 实时的ADS数据存储在哪里
4. 说一下实时的指标做过哪些，然后自己挑几个进行实现
5. 离线的DWD的宽表有多少
6. 说一下DIM和DWD的建模
7. 说一下维度表和事实表的区别
8. 为什么实时使用Maxwell进行同步
9. 维度建模使用的什么模型
10. 说一下用户画像的架构
11. 用户画像使用的是什么引擎
12. 任务调度使用的是什么代码写的
13. 说一下你写过哪些指标

## 面2

1. 自我介绍
2. 说一下最近做的项目
3. 说一下维度表和事实表的区别
4. 做过哪些指标，sql怎么写
5. Flink事件语义，窗口
6. 为什么用Maxwell，sqoop了解吗能使用吗
7. 做项目遇到哪些困难
8. 写过哪些shell脚本
9. 两道sql题：

表内容

a b c

2014 B 9

2015 A 8

2014 A 10

2015 B 7

问题一：多行转多列

问题描述：将上述表内容转为如下输出结果所示：

a col\_A col\_B

2014 10 9

2015 8 7

问题二：如何将结果转成源表？（多列转多行）

问题描述：将问题一的结果转成源表，问题一结果表名为t1\_2。

# 汉克科技

## 面1

1. 自我介绍
2. 介绍公司项目（实时）
3. 用flink做什么
4. 离线用hive
5. flink算子
6. flink sql flink api
7. 项目上线遇到什么问题
8. redis的ttl设置 我说一天
9. 用java开发 不咋会
10. java掌握程度
11. 熟不熟悉kafka
12. 具体说一下kafka（组成，版本，多少台，业务数据多少，多少条数，offset消费方式）
13. 数据重复消费测试
14. kafka的leader和follower topic
15. 如果一个leader挂了，kafka
16. kafka的leader的选举流程
17. kafka的分区策略
18. 业务数据多少个分区
19. 遇到过节点挂点的情况吗
20. flink的taskma jobma 内存分配 cpu分配
21. 实时项目日志数据怎么处理维度数据
22. flink cdc 原理
23. kafka有啥作用 我说的萧峰缓存，可以重复消费，存储数据量大，
24. kafka读写快不快，为啥
25. hbase和hive对比
26. hbase数据热点
27. 关系型数据库用什么比较多 mysql
28. mysql和hbase区别 胡乱诌了一堆
29. java类加载顺序 不会
30. java版本
31. java的函数式编程 了解一点但没用过
32. A系统topic数据经过清洗后放到另一个topic，b系统没有数据，如何排查，一开始能读取，后来不能读取
33. 个人倾向实时还是离线

## 面2

1. 自我介绍
2. spark开发
3. 有没有做过spark读文件 做操作
4. flink实时批量操作有没有做过，我说做过开一天窗口的。然后问一天窗口数据量会不会很大，数据量的问题、问按天统计给多少资源
5. 数据量多少 我说日活50w，然后flink中日活50w 这个状态保存在哪，怎么统计日活，然后rocksDB 了解深吗
6. flink基于什么 k8s 还是yarn
7. rocksDB是新搭的还是内部的
8. flink挂了还有rocksDB吗？在哪存着
9. 开发语言
10. java基本数据类型
11. char几个字节
12. int字节
13. int float double 区别
14. int float区别
15. float存储数据遇到的问题？double存储数据遇到的问题？
16. java怎么判断两个对象是否相等，如果是两个student对象内容都一样 相等吗
17. 常用的集合类型
18. list set区别 set如何去重
19. list实现类
20. Arraylist 删除大于10的数字 需要说代码
21. flink写代码的流程 main函数里面怎么写 需要说出来 具体代码
22. flink读kafka 说具体代码
23. flink数据 往哪里写
24. 想把flink数据积攒缓存10s 再写入clickhouse 怎么实现
25. windowall 怎么理解的
26. flink精准一次性
27. 代码上线流程
28. 有运维团队？
29. 作业的管理权归运维还是开发？、
30. 作业跑失败 处理流程

## 面3

1. spark的RDD怎么理解？
2. 有一个文本文件，我们怎么把他转为RDD？
3. 有一个文本文件，我们把它读进来，怎么转换成DataFrame
4. java中的集合类用过哪些？
5. hashset和treeset的区别？
6. arraylist底层实现是用什么？底层数据结构是什么？
7. arraylist添加元素最多能加多少个？
8. 用java写了两个类，比如student类，怎么去比较两个对象相等？
9. 只重写.equlas不重写.hashcode行不行？ 为什么不行？
10. flink怎么保证精确一次？
11. flink读取数据后，开了一个5s的窗口，5秒种往下游写一次数据，写的时候是单条单条的写，能不能保证精准一次？
12. checkpoint记录的是什么？窗口里的数据能记录哪些信息。
13. flink开了一个窗口，一次读10条数据进行计算，读到第5条的时候挂了，重启flink，会重新发前5条，还是接着第6条读？

## 面4

1. 自我介绍
2. 介绍公司项目（实时）
3. 为什么要旁路缓存+异步io
4. redis缓存数据保留时长
5. flink开发卡住时间长的问题
6. flink作业分配资源
7. 资源不足造成的反压
8. flink用java还是scala
9. java熟悉？
10. kafka了解
11. 实时数仓几个topic
12. 业务数据一天多少量
13. kafka版本
14. 集群台数
15. 一个topic分区和副本各自作用(第一时间没反应过来)
16. 一个topic可能被两个消费者消费吗？消费者组
17. kafka集群挂掉
18. kafka精准一次消费
19. kafka offset存储在哪 还有什么高并发？
20. flink算子
21. keyby 底层原理
22. Flinkcdc
23. 离线 hive和hbase区别 问我hive是数据库吗？
24. hive删除表 数据会删除吗
25. habse行和列 都有哪些角色 或者说一个表的组成
26. mysql使用过吗
27. 场景

a系统数据库 cdc组件采集一个表到一个kafka topic

flink对kafka数据做简单清洗 写到另一个topic

b系统来读另一个topic数据 b系统读不到数据了 问题在哪？

1. broker中的leader选举机制
2. 想从事什么工作
3. 问我java脚本开发能接受吗
4. 问我java类加载顺序

# 柯莱特集团（笔试）

1. 左连接和右连接什么区别？左表 3 条记录，右表 2 条记录，2 个表做左连接，结果是几条记录？
2. 数据库中有个员工表 emp,建表语句如下: create table emp\_his (emp\_id number, emp\_name varchar(50)); 表中存在重复记录(根据 emp\_id 来判断),用 SQL 如何查出存在重复的 emp\_id？
3. 有张 student\_scores 表，存放每个学生，每门课程的分数，建表语句如下: create table student\_scores (emp\_id number ,course\_name varchar(50) ,score number); 需要把每门课程按照学生分数降序排名并添加序号，sql 如何写?
4. 有表 emp\_his,表 emp\_inc，表结构一样，关联条件是 emp\_id，建表语句如下： create table emp\_his (emp\_id number,emp\_name varchar(50),age number); create table emp\_inc (emp\_id number,emp\_name varchar(50),age number); emp\_his 表存放历史数据，emp\_inc 表存放有变化，或者新增的数据，如果要一个最新的全量数据，sql 如何写？
5. 订单表 orders(id,order\_id ,member\_id ,order\_type,store\_id,create\_time,point) Id 主键 order\_id 订单 id member\_id 会员 ID order\_type 订单类型 store\_id 门店 ID ，create\_time 创建日期 ，point 积分会员表（会员 id 唯一） members(member\_id,name,member\_level) member\_id 会员 ID ，name 会员名称 ，member\_level 会员等级门店表（store\_id 唯一） stores(store\_id,name,store\_channel) store\_id 门店 ID ，name 门店名称 ，store\_channel 门店渠道 ① 查询门店名称包含“猎洋”的所有订单信息，统计出每个会员的总积分值； ② 查询门店名称为“婴童”，且获取总积分值大于 100 小于 600 的会员姓名和积分值； ③ 查询获得总积分最多的前 3 名会员名称和积分值； ④ 查询同会员‘张三’总积分相同的会员 ID；
6. 把字符串‘foothebar’中的 th 替换为‘22’
7. 字符串 'abc’,'def’,'gh’ 通过函数拼接起来;
8. Hive 中 order by ,sort by 区别，什么场景下使用？
9. Hive 表分区的作用？动态分区的做什么用的？
10. Java 中有哪几种集合？区别是什么？适用场景是什么？
11. 写一下维度建模的理解？雪花模型和星型模型的差异？维度建模的优缺点？
12. Spark Streaming 和 Flink 的技术区别是什么？
13. Flume 和 Kafka 的技术栈是什么，适用场景是什么

# 上海睿民

## 面1

1. 自我介绍，你在公司都干啥
2. 现在找ETL的工程师，sql写的怎么样，你干了那么长时间，都写哪些指标
3. 有这样一个场景，有两张用户表，一张昨天的，一张今天的，问怎么样得出新增，变化的人数（表里没有create\_date字段），怎么修改表，怎么实现？

## 面2

1. 离线框架介绍
2. 写过什么指标
3. SQL优化怎么样？
4. DIM 建模展开说说
5. HIVE优化？小文件处理？HIVE数据倾斜
6. shell脚本写没写过

## 面3

1. 自我介绍
2. 离职原因
3. 以后想要往哪方面发展
4. 工作中主要负责的部分(大概介绍了一下实时数仓 说到数据调研）
5. 数据调研是啥
6. 比较擅长哪方面
7. sql的问题： 有张表有重复数据想去重

有的数据重复 有的数据没重复

## 面4

1. 介绍一下你的项目，我介绍的实时
2. 他给我介绍了他们的业务，是银行业务，想招离线的
3. 问我在项目里干了什么
4. sql题：

A B 表分别是用户今天的全部信息和昨天的全部信息，

怎么求今天新增的，删除，变化的数据？

1. 你有什么要问的？

## 面5

1. 离职原因，期望薪资
2. 最近一个项目架构，流程
3. Kafka数据保存多久，到期数据如何处理、Kafka优化、零拷贝解释
4. Flink CDC 是否用过、是否用过orcale、mpp数据库、Python、Shell脚本是否会写
5. 数仓建模理论，星型模型和雪花模型区别，事实表类型，事务和周期快照事实表区别

## 面6（外包民生银行）

1. 自我介绍
2. 介绍一下实时项目架构和你主要做的部分
3. 实时项目中你遇到的最棘手的问题是哪个，详细讲解下。
4. 维度数据存储在Hbase中全量同步后，只做增量更新，如果你的程序在增量同步过程中出现了故障，数据没有更新过去，你们有什么故障保证机制吗？
5. 讲解一下flink的checkpoint。
6. 接着上面的聊，那你们有遇到过程序故障,无法从检查点或者保存点恢复数据的情况，如果做这种故障保证。

## 面7

1. 自我介绍
2. 离职原因
3. 未来规划
4. 离线实时挑一个擅长的讲
5. 用到哪些数据库，为什么用
6. 项目中遇到哪些问题
7. 期望薪资
8. 有什么想问的

# 普华讯光

1. 自我介绍
2. 最近项目介绍，感受最深的地方
3. 对加班态度、离职原因
4. 技术介绍、用户画象是干啥的，我负责的部分
5. 个人优缺点

# 联洋国融

1. 讲一下hdfs的优化、
2. 讲一下namenode的RPC线程数的优化设置吧，跟什么有关
3. 讲一下namenode的高可用，zkfailercontroller为什么不放到namenode里面为什么是单独的进程
4. 做过安全配置么，系统怎么做的监控
5. 讲一下kafka分区和副本是什么，用途，不看场景的话基础配置的数量怎么定
6. kafka生产者消费者分区数一般的数量对应关系
7. kafka如何不丢数据，讲一讲ack设置成-1的内部逻辑
8. 如果一个follower挂掉会怎么样，多个follwer挂掉会怎么样
9. kafka的ISR和OSR讲一下

# 北京南天

## 面1

1. 讲一下实时数仓
2. 详细说一下数据怎么采集的
3. flume具体怎么配置的， 它怎么能取到数据文件的
4. flink怎么保证不丢数据的
5. flink程序提交的参数有哪些
6. 你们flink任务内存怎么分配的
7. 你们实时项目做出了什么东西
8. 有没有做过增量与全量的对比
9. MR怎么实现去重，hql去重底层原理是什么

## 面2

1. 实时项目串讲。
2. 你在项目中负责哪些。
3. 你们公司实时都统计过哪些指标？ 你们的日活怎么统计的？怎么统计某一时刻的访问量？
4. 你们窗口里面做了哪些处理
5. 如果窗口已经关掉了，还有数据没到怎么办
6. 日志的数据格式是怎么存的
7. flume怎么监听日志文件的变化？
8. 提交flink jar包的时候一般设置哪些参数
9. flink的并行度，jm，tm内存等都是怎么去设置的？
10. 一般flink设置多少并发？
11. 瞬时数据量最多能达到多少？
12. 日志里做没做过数据清洗？
13. 你们第二个项目中spark的作用是什么？（第二个我写的用户画像）
14. spark运行的流程是什么？
15. 读到大文件的时候怎么处理的？ 切分 跟mr的切分一样吗？
16. spark的很多算子，它怎么去划分各个阶段？
17. 哪些算子会重新分区？
18. map和map partition的区别？ map partition处理一个分区的数据，他跟map不一样吗，不是一条一条的处理吗？
19. hive的自定义函数写过吗？ 写完自定义函数怎么去应用？
20. hive的文件用哪种压缩方式？
21. gzip的文件有多大？ 10%
22. hive读gzip文件会切片吗？ gzip不支持切片
23. 切片怎么切呢？
24. 数据里面带换行符的采集的时候怎么办？
25. 你们写的shell脚本多吗？

# 北京鼎普科技股份有限公司

1. 自我介绍
2. 离线数仓介绍，负责内容，
3. 日志数据是通过前端埋点产生的，是标准格式吗，通过逗号分割的吗？
4. Hadoop读写流程
5. NameNode元数据放在哪里？ 我说的不知道
6. SecondNameNode作用
7. 高可用脑裂问题
8. DataNode 和 NameNode之间的心跳机制 JobTrack TaskTrack
9. HDFS安全模式，数据不一致会进入安全模式，问这种情况出现过吗，什么情况下会进入安全模式？
10. Hadoop处理小文件问题
11. mr的shuffle
12. zookeeper选举机制
13. hive内外部表区别
14. 数仓建模，数据模型有哪些；问我用的什么模型？
15. 用过hive解析过json字符串吗，有没有通过自定义函数解决？
16. sort by 和order by区别
17. hive数据倾斜怎么排查的，解决方式
18. hive内置函数
19. hql转换为spark怎么转换的，hql解析
20. spark rdd持久化原理
21. 宽依赖和窄依赖
22. KafKa已经消费完的数据还能再次重复消费吗
23. Kafka怎么保证消息可靠性
24. Kafka消费能力不足，数据积压怎么办
25. 离线时是主要写hql吗，做过spark吗
26. Flume写过自定义拦截器吗？怎么实现
27. 计算指标
28. 每天多少数据量
29. 实时是写flink的吗，是直接通过hbase shell操作的吗
30. Hbase里有索引吗
31. hbase写数据
32. flink容错机制（检查点）
33. flink任务失败后，重新运行，怎么保证数据一致性？
34. 想问的问题

# 数字天堂

1. 介绍一下你们的架构，
2. flink做了多久了
3. flink都有哪些join，interv join 和 lookup join 有什么区别，
4. 具体在哪用到过 lookup join
5. spark on hive 和 hive on spark 的区别
6. 维度建模 事务性事实表，周期型快照事实表， 累积性事实表，具体用到了哪些
7. 怎么建的这些表

# 上海观安

1. Flume每日日志采集有多少数据量
2. Flume中配置了什么过滤器？
3. Flume到Kafka用的什么channel，做了哪些配置，
4. kafka你们配置了几台机器？
5. kafka假设三台机器，有一台不可抗力因素挂了，怎么保证其他两台正常运行？
6. Flink上游的数据是怎么采集来的？
7. 七日内连续三天登录统计
8. 你们Flink DWD的数据写到kafka后又写入了哪里
9. Clickhouse介绍一下
10. FlinkCDC干什么用的
11. 未来有什么打算

# 傲林科技

## 面1

1. 自我介绍
2. 项目介绍
3. 数据量+集群配置
4. 遇到过哪些问题，反正数据倾斜
5. chlikhouse的计算引擎
6. 接触过运维吗
7. flink版本
8. flink内存 ，托管内存干什么用的？
9. spark处理了哪些业务
10. flink类加载，两个类型了解吗
11. java垃圾回收机制
12. 你感觉你的优势是什么
13. 你工作的三年最有成就感的事

## 面2

1. 做一个自我介绍。
2. 你们做的大数据这块内容，简单说一下你们做的业务。
3. 离线项目和实时项目都做了多久
4. 你们实时项目都用了什么技术
5. 属于什么部门，也是分组吗？
6. 机器规模大概是多少，用的是云上的吗，是自己部署的吗
7. 组件之间在机器上怎么划分排布的，Kafka布置了几台，yarn布置了几台
8. 你们数据量有多大
9. yarn的计算资源是怎么划分的，做了哪些参数的配置
10. 你们用的什么版本的hadoop，这个版本有什么新特性
11. Hadoop的纠删码你了解吗
12. Kafka的吞吐量是怎么优化，（我说完后他问）除了参数调优还有哪些配置的优化
13. Kafka的峰值有多少
14. 数据采集到kafka的后续处理逻辑是什么
15. 你们都用了哪些flink的API（我拿Flink CEP说的），他问CEP做过哪些动态
16. Flink的并行度是怎么设置的
17. flink的slot是怎么理解的，怎么让它不共享任务插槽
18. flink的内存结构，（我答完后），问什么是托管内存
19. 你了解java的设计模式吗？简单说一下
20. 你了解Java的类加载机制吗
21. 你还有什么想问我的

## 面3

1. 介绍下你们项目的人员组成
2. 你们集群规模
3. 你们集群做了那些优化
4. 常用编程语言，java有哪些设计模式，简单介绍下
5. 简单介绍下flink内存
6. flink做了哪些优化
7. 你有什么想问的

## 面4

1. 自我介绍
2. Yarn资源配置（具体参数，具体到RS与NM的内存配置大小，）
3. Yarn调度器， 你们选择的那个，你们有几条队列
4. Yarn高可用配置了吗， 如何避免ResourceManager挂掉，挂掉之后如何处理（ResourceManager中具体参数设置）
5. 数据量 ，集群规模，
6. flink内存 模型，托管内存干什么用的？
7. java垃圾回收机制了解吗
8. Flink使用那些算子（结合具体代码讲使用流程）
9. Java设计模式了解吗， 具体举例
10. 实时跟离线共用一套吗， 如何做资源隔离
11. 大数据有多少人

# 小红书

## 面1

1. 介绍一下最近做的项目（考察真实性，会问细节）
2. flink用sql多还是api多
3. 离线和实时数据校对怎么处理
4. 实时和离线分别做过哪些优化
5. 两道sql，求平均、互相关注数量
6. sql具体是怎么执行的，执行顺序

## 面2

1. 自我介绍
2. 讲讲你们怎么做离线系统的
3. 你们怎么跟BI对接啊，给BI发的文件是什么格式的。
4. 写sql
5. 求每个用户的最长连续下单天数

要求：

1、考虑多段连续下单的场景

2、过滤掉非连续多天下单用户

3、函数名如果忘记，可以随便写一个代替，最后解释一下即可

4、请直接在下方写出实现SQL

让你同意同一时间最大在线人数，表字段有用户id，登录时间，退出时间，说出思路即可

1. 你有什么想问的吗

# 北理新源

1. 上来就说用scala版spark core开发
2. 实时数仓怎么计算的
3. 体温检测场景题
4. 实时数仓延迟
5. kafka每秒多少数据
6. 开发过flume源码吗
7. ck稀疏索引
8. hbase读流程
9. hbase二级索引的协处理器开发过吗
10. ES做二级索引
11. hive udaf写过吗
12. flume怎么使用的 写过其中的类吗，拦截器哪个线程调用的
13. spark任务划分
14. java数据结构算法，反射，多态 ，多态对应的设计模式，字典树

# 腾云天下

1. java引用数据类型和什么什么数据类型区别
2. 数据量：有多大，问我内存问我硬盘
3. spark为什么叫弹性集合
4. kafka为什么那么快
5. kafka的broke如果有一个leader挂掉了，剩下follow怎么均衡分配
6. java集合有哪些。
7. java的hashmap底层怎么可以能够不重复
8. kafka精准一次性，性能不好怎么办
9. scala什么和java的静态类很像
10. hbase：读写流程
11. 怎么调优hbase的读流程
12. blockcache有几类
13. 如果这边刷写hbase挂了怎么办怎么保证不丢失

# 润通科技

（1）DNS日志格式

clientIp | domain | time

192.168.1.2 | www.baidu.com | 2022-01-12 09:35:40

（2）求每小时每个clientIP访问量最大的100个domain及其访问量

clientIP在百万级别，domain在十万级别

（3）实现一小时的分组topn

# 北京世纪慈海科技

1. 介绍一下您在之前公司做过的项目
2. 说一下您所作的工作对你们项目有啥优化贡献？（意思就是绝的我刚刚说的都太底层，想了解一下我参与了那些业务，我和他说了用户画像，但是我负责的很少）
3. 说一下sparkStreaming和flink的区别，更都适用于那些场景
4. MySQL数据库索引
5. 职业规划

# 康明斯

1. redis缓存击穿 缓存血崩发生后如何解决
2. java接口和类的区别
3. java和scala区别，scala有什么优点
4. flink怎么处理迟到数据
5. mr shuffle流程
6. spark streaming和flink的区别
7. mysql引擎,b索引和b+tree索引
8. flink上游算子处理延迟怎么办
9. string类不用final修饰可以吗
10. 面向函数编程和面向对象编程有什么区别

# 东华软件-中邮人寿

## 面1

1. 项目介绍
2. 在数据库开发中遇到的注意事项
3. 保险行业里挑选一种业务：理赔，保全，合同，说明表结构与表链接关系
4. flink的join方法，如何保证数据一致性
5. 主要是让说保险业的数仓经验，没有经验很难说

## 面2

1. 简单介绍项目。
2. 你们hive、hbase权限怎么开通的？
3. 你们非结构化数据保存在哪里？
4. 你们用的软件是开源的吗？
5. 你们这个维度建模都是什么？
6. 你们用户拉链表最重要的字段是什么？
7. 左表3个字段，右表5个字段，left join后几个字段？

# 软通

## 面1-外派居然之家

1. 实时项目中访问时长是如何计算的？
2. 你们有几台服务器？
3. 你们实时和离线都在这几台集群上吗？
4. Hadoop集群用的什么版本？是CDH的吗？组件都是运维的撘的吗？
5. 每天的数据增量是多少？
6. 前端的埋点是你们自己做的吗？
7. 神策埋点的数据你们是怎么接回来的？
8. 神策应该有加工好的数据吧？
9. 那你们做分析的时候都用的原始日志？
10. 你们的每台机器都会部署flume吗？
11. 你们所有的日志都会到一台机器上是吗？
12. 多台日志服务器，都配flume吗？
13. Hive的计算引擎用的什么？
14. 换到Spark之后用没有遇到什么问题？
15. 有做过Hive的优化吗？
16. 数据同步用的DataX吗？
17. Kafka消息队列中的消息是有序的还是无序的？
18. kafka有遇到消息堆积吗？
19. Kafka数据丢失遇到过吗？
20. 集群挂掉过吗？
21. 你们调度是用的海豚调度器？你们平时开发任务是在什么平台上进行开发？你们写HiveSQL任务是在什么平台上去写？
22. 开发完之后测试和发布的流程是怎样的？
23. 发布流程是生成Shell脚本配到海豚调度器中吗？
24. 你们现在离线的任务有多少个？实时有多少个？
25. 实时的任务时挂在什么上面跑的？
26. 集群有做权限的管理吗？是到表的权限吗？还是具体到每个人的权限？
27. 雪花模型和星型模型各自有什么特点？
28. 用过拉链表吗？
29. 这种变化的维度你们是怎么处理的？
30. MySQL了解吗？
31. 你觉得MySQL可以做数仓吗？
32. 用户画像这块你做了那些？
33. 数仓最后的ADS层放到那里了？
34. 有没有一些明细层DWD层的数据需要做报表展示的？
35. 没有那些业务的同事需要到DWD层导一些明细数据吗？例如：订单明细数据和用户信息数据？
36. 平时业务的话，比如产品、运营等同事需要一些数据的时候是他们自己去查询吗？会写SQL的业务同事直接就可以查是吗？
37. 你在开发过程中遇到过比较难的技术问题吗？
38. 你倾向于做实时的开发还是离线的开发呢？
39. 说一个你的优点和缺点？
40. 平时开发程序主要用哪种开发语言？
41. 最后一道手写SQL题：一张login表，两个字段，username和logintime，username唯一，每个用户每天会登录多次。求出当前最新的十个登录用户。
42. 讲完主要用Java，又给了一道手写数组反转；

## 面2

1. 在职还是离职？为什么离职？哪一年毕业？为什么转行？
2. 介绍一下最近的项目；
3. kakfa怎么集成？部署？答：三台+zk；
4. kafka挂丢重乱积压？答：大保健；
5. kafka分区分配策略？答：大保健；
6. spark用过吗？答：两年前用过；现在主要用flink；
7. spark作业提交和任务划分？
8. hive数仓搭建？答：数仓建模分层离线串讲；
9. 数据倾斜？答：大保健；
10. sql优化？答：语法、建表、job；
11. java基础怎么样？答：学了SE ，框架不太懂；
12. java接口和抽象类区别？答：忘了；
13. java常用结合？ArrayList和linkedlist区别？答：大保健；
14. 住哪儿？期望薪资？外派中行

## 面3-南京雨花台

调度我们是使用yarn调度的，监控我们是使用WebUI界面监控Flink作业的

1. 数仓中什么比较重要？
2. 怎么解决redis或着其他组件可能出现挂的情况？
3. Flink你们的任务是用什么进行监控和调度的？
4. 你们的数仓为什么要分成这么多层？
5. 讲下你的FLink项目的架构？
6. 你主要负责的工作是什么？
7. 实时总共有多少指标？你做了多少指标？
8. 说一下你的日活指标是怎么实现的？

## 面4

1. 有没有用过sparksql对离线数仓中的数据进行清洗？
2. sparksql中join的策略有哪些？
3. flink中的状态有哪些？
4. 解释一下Flink的两阶段提交原理？
5. FlinkafkaProducer中怎么保证数据提交时的幂等性的？
6. Flink和Spark的区别？
7. 为什么离线的时候将报表数据写到Mysql中，但是在实时的时候写到clickhouse中？
8. 行式存储和列式存储的区别？
9. redis能介绍一下吗？
10. 怎么避免redis的频繁连接和断开？
11. 配置表发生变化的时候流里面的配置信息怎么弄？
12. 当你redis的服务断了，对你的实时有影响吗？怎么解决？
13. 怎么将不同keyd的数据提批量的提交到redis中，而不是一条一条的加？
14. 写springboot用到了Mybatis,你是怎么实现Mybatis支持的？
15. 在Sprintboot中怎么实现消息拦截的？
16. Spark中怎么解决数据倾斜的问题？
17. Spark中的广播变量应用场景？
18. 数据量太大的时候不建议使用广播变量，为什么？
19. 广播变量传输的数据量很大的话，对谁影响比较大？
20. 在离线数仓中数据量最大的是哪张表，数据量是多少？
21. spark中的调度有哪些？
22. spark中怎么实现任务满足某条件定时执行，通过注解来有实现过吗？
23. spark中怎么实现任务不重复执行？

## 面5

1. 自我介绍
2. hdfs小文件处理
3. shuffle怎么优化小文件
4. yarn的提交流程
5. kafka了解吗，有几个节点，几个副本，最大数据量，怎么设置分区数
6. kafka丢数据怎么办
7. hive sql用的多吗
8. 自己开发过UDF函数吗
9. 内部表和外部表的区别
10. 窗口函数，讲一下学生、课程、成绩三列，怎么按照课程对学生进行排名
11. 实时这边讲一下怎么拆解指标，具体什么应用场景
12. flink端到端的一致性怎么保障
13. 讲一下数仓的建设流程
14. java的hashMap和hashList的底层
15. 用户下单涉及到哪些维度

## 面6

1. 自我介绍
2. 离线和实时那个偏多
3. mr和spark都熟悉吗，更擅长那一个
4. mr 和 spark的shuffle的区别
5. sortshuffle生成的文件有什么特点，快在哪里，有什么优化，
6. mr的combiner 是什么时候操作的
7. hive数据倾斜的处理
8. flink处理什么样的事情，加工流程，链路
9. kafka数据保存多久
10. flink的精准一次
11. barrier对齐的原理，barrier到了会做什么事情
12. 双流join，flinkSQL left join 底层是怎么实现的
13. 用kafka做消息队列，数据延迟的现场，及原因（排除网络延迟故障）
14. 从flink 内部逻辑来讲的话，数据延迟的原因（他问的比较含蓄，就是反压或者与外部系统交互 等原因 会造成延迟）
15. flink的反压
16. kafka 单个partition 数据倾斜有遇到过吗
17. 数仓建模 建设流程具体怎么做的，你做了那些事情
18. 数据域划分在详细的展开说一下
19. DWS轻度汇总表，有维度的增加和减少要怎么处理
20. 一些转化率什么的基于多个业务过程怎么构建DWS层轻度汇总表。（他想问的应该就是那种衍生指标，然后我说先拆为派生指标，在抽取相同的业务过程等 构建DWS层轻度汇总表）
21. 用过HBase么，rowkey设计原则
22. rowkey散列原则怎么设计
23. rowkey怎么实现唯一的，代码上讲（怎么实现的），不唯一会怎么样。
24. 你有什么想问的
25. 离线实时都做，每天数据量上百TB。

## 面7

1. 数据量？
2. ods到ads表怎么建的？
3. hive shuffle 或者sparkshuffle ？
4. shuffle有几次排序，可能有4次么？
5. flink端到端一致性
6. 为什么选kafka？
7. flink和sparkstream区别？

## 面8

### 一面

1. 介绍一下实时数仓项目
2. 用户画像是对接到实时项目的么
3. scala会么，sparkstreaming用过么
4. hbase调优讲两点
5. 业务怎么确定的
6. 你们有哪些业务
7. 比如你们下单表用到了哪些维度
8. 事实表的粒度怎么确定的

### 二面：

1. 讲讲最拿手的项目中你主要干的事情
2. 讲讲维度建模
3. 为什么用clickhouse
4. flink leftjoin内部原理
5. flink sql 实现topN
6. count（distinct）有几次reduce

## 面9（运维相关）

1. kafka系列相关问题(kafka角色及作用，kafka底层，leader选举，controller作用)，影响分区读写速率的因素（IO）
2. region切分具体原理以及切分时是否可对外查询
3. hfile合并以及合并时是否可对外提供服务
4. hbase查数相关，get和scan的区别，查数时指定范围（scan startkey和endkey）
5. 如何避免一下子从hbase查出大量数据（hbase如何实现 ？ 我说的用phoenix的sql limit）
6. 数据一开始写往一个region，太大则会自动进行切分
7. spark角色及作用
8. spark提交流程
9. java并发安全问题

## 面10：

1. 说一下最近做的项目，你在项目里做了什么？
2. Flink CDC怎么用的？Flink CEP用过么？怎么用的？
3. Hbase和Redis数据一致性？redis中的数据多久更新一次？
4. Redis用的主备模式还是分布式？你了解Redis多么？
5. 还问了一些关于阿里云产品的了解有多少
6. Hadoop是怎么进行数据校验的？
7. kafka的事务
8. su和su - 区别
9. 数据安全你们用的什么框架？

## 面11

### 一面

1. 自我介绍
2. 谈谈对flink的理解
3. flink的一致性
4. 什么是幂等性
5. 两阶段提交的时候已经申请提交了，这时候报错了之后会发生什么事
6. flink的任务执行流程
7. 你们的数仓分层
8. 建模怎么做的
9. hive的常用优化
10. spark sql有哪几种，谈谈理解？ （我不太懂这个）
11. hive常用的函数
12. 考虑过建表时的声明周期吗
13. 云产品有哪些理解
14. 你有什么问题想问的？

### 二面

1. 自我介绍
2. 你实时和离线主要哪个做的多
3. flink描述一个业务场景
4. 说一下topn，热门商品，总体怎么实现的
5. dws层在什么平台上
6. 有涉及到多流关联的操作吗
7. 除了关联维表，有没有两个事实表的join情况，怎么关联的？
8. 什么情况下会有回撤流的现象，怎么解决的？
9. 离线的数据仓库怎么设计的？
10. olap主要用的ck吗，你们选择它的原因是什么？
11. 对比过其他的olap吗？
12. 离职原因是什么？
13. 有什么想问的？

# 北京中达恒业科技发展有限公司

1. HDFS读数据流程
2. HDFS的DataNode挂了怎么办
3. 手写冒泡法排序Java代码
4. HBASE框架，region是做什么的
5. kafka内存里的数据满了怎么办
6. 手写spark wordcount代码

# 北京博文天润

## 面1

1. 建模流程
2. sql版本
3. maxwell 底层原理和优势
4. hive on spark优化
5. 是否独自搭建阿里集群
6. 平时是怎么简单测试的
7. 遇到数据倾斜解决方案

## 面2

1. 自我介绍
2. 介绍离线项目
3. 维度建模中的星型模型和雪花模型的区别
4. hive数据倾斜
5. 介绍实时项目
6. 你有什么问题问我的

# 畅聊天下

1. 集群介绍一下
2. 给你20G内存，执行spark程序，driver和executer怎么分配
3. spark提交作业流程
4. 你们DS调度，参数怎么配置
5. kafka为什么4个分区，压测过吗
6. 你感觉sql和flink sql有什么区别
7. 建模流程

# 薪盾时代

1. 介绍项目
2. kafka和hadoop 各有几个节点
3. namenode 宕机 怎么办
4. linux常用命令查看内存，磁盘
5. zookeeper 文件导致磁盘到达80%，怎么办

# 汽车之家

1. hive中小文件怎么处理？
2. hive做过那些优化？结合具体的场景，比如遇到了什么问题？优化后的效果怎么样？数据类型不一致，join为什么会产生数据倾斜？
3. hive 中的 文件存储的格式：ORC 原理，ORC自带压缩吗？你们使用那种压缩，为什么？
4. hadoop会挂掉吗？你们遇到挂掉，怎么解决的？
5. HDFS小文件怎么解决？
6. 有没有基于sparkCore 作业的开发？
7. 基于spark原理，举1-2个在生产环境中调优的一些案例，在项目里面应用到的。
8. flink的水位，分流 和 合流
9. flink的双流join 如何保证数据一致？状态
10. flink有两条流，一条是主流，处理时间（延迟）为2000ms，另外一条是辅流处理时间（50ms），如何保证主流的数据到了再和辅流进行join，写入到表中？flink 的state.ttl设好
11. kafka 的分区如果增加了，会发生什么问题？ 开启分区感知
12. 实时数仓的指标为什么使用clickhouse，跟那些组件进行了对比？
13. 基于大数据的场景：有一个表，有200亿条数据，有三百个字段，f1--> f300 ,每一个字段取值都是0-9，做一个统计，统计每一个字段的取值分布，F1取值为0的有多少个，取值为 1的有多少个....要考虑性能的问题。 不做groupby 用sum(if())计算3000个指标，可以写代码拼接sql
14. 两个表都有100多个字段，两个表进行join，如何分别取出两个表的50个字段？
15. 在hive中有一个表，存储的格式是文本，然后把这张表的数据导入到另一个hive表中，存储格式是ORC，如何导入？怎么验证数据是否一致？这是很常见的问题，应该有三种方式，这都不会？？？你平时在做什么？insert overwrite 小表比值 ，大表比量，随机抽样本，金额求SUM
16. 平时用什么开发？java scala，问你java的基础问题：java中的list，实现类有那些，底层都是怎么实现的？扩容？

# 致美生活（红布林）

1. 数据量有多少？下单数据有多少条？支付数据有多少条？
2. clickhouse的查询性能为什么高？ 列式存储 稀疏索引 语句级多线程
3. clickhouse的数据一致性？ 至少一次+幂等
4. clickhouse都有哪些引擎？你们用的什么引擎？为什么选这个引擎？ Replacing 去重 mergeTree 正常 AggregatingMergeTree bitmap
5. ORC存储原理？条纹？是否带压缩？
6. clickhouse写入速度能达到多少？
7. 使用clickhouse 做关联查询吗？性能？ 不做关联查询
8. clickhouse能够承受多少并发查询？ 单机每秒几十 不高 qps
9. redis 分库的概念？
10. redis的 scan 和 keys 查询有什么区别？
11. java的hash表底层原理？
12. table1：时间、姓名、is\_win

is\_win 有两个值：win，lose

使用sql，统计每一个人的连续赢的次数，怎么实现？ 开窗 按时间排序 上一行和本行为胜给 1个连胜标志 最后汇总即可

如果使用java怎么实现？

# 宇信科技

## 面1

1. 介绍下实时系统
2. 用过的shell 命令有哪些？如何查询一个文件夹中某一个文件（grep -r) ，如何获取第n行到第 i行的数据？（sed -n）
3. 使用的hive的版本有哪些？（hive 3.1.2) hive 1.x 与 hive 3.x 有什么改进？
4. hive中的优化有哪些？
5. 在hive1.x中没有update,delete ,如果要更新某一条数据怎么做
6. 给定一个时间（年月日）如何确定这个时间所在季度的开始与结束时间（设置一个固定的时间表，各个时间点对应的季度）
7. 用的调度器是什么? shell吗？用过其他调度器吗？ Ds
8. 分桶表用过吗？与分区表的区别？ 外部表与内部表的区别？

## 面2

1. HBase RowKey
2. 项目介绍
3. 拦截器的具体实现方法
4. 每天数据量多少，有多少台服务器
5. hive数据倾斜优化
6. hive调整过哪些参数
7. kafka调整过哪些参数
8. kafka有多少个主题
9. 数据到HDFS做过哪些处理
10. 用什么方式查询ES数据，查询过ES的什么数据
11. 平时怎么执行脚本的

# 万达（商业）

1. 简单的说下项目
2. hsql中1-3-2 、3-1-2、2-3-1怎么修改1-3-2 顺序
3. 统计平均分大于学生的统计 统计大于平均分的学生？ 用开窗取平均分 然后比较
4. sql 6天 连续登录用户（间隔的也算） 开窗
5. 怎么处理大数据量 用datax ,mysql -->hdfs
6. 大表的话，怎么处理？
7. dwd层 dim层都做了什么。
8. 简单说下cdc
9. cdp用过吗？

# 融信数联

1. 大学期间学过计算机相关的专业吗？大数据的相关技术是怎么学的？
2. 平常开发用的是什么语言？Java？Scala？
3. 介绍一下项目
4. 你在项目中负责的业务
5. 在工作中，哪些地方用到了SparkStreaming和Spark Core的开发
6. 使用Java开发过什么？
7. spark调优
8. 工作中用到的相关技术的版本

# 上海思睿

1. 串讲一下实时项目？
2. 你们的配置表的字段都是什么？
3. 为什么选用Clickhouse，Hbase？
4. 你了解几种设计模式？项目中用到了哪几种？你当时怎么能考虑的？
5. 模板方法设计模式和抽象类，接口有什么区别？
6. Hashmap和Hashtable的底层原理？
7. Treeset和Hashset的底层原理？
8. 你了解多少种锁？怎们实现的？
9. class锁怎么实现的？
10. 双重校验锁什么时候用的？解决了什么问题？具体实现说说？
11. 谈谈Flink窗口的理解？
12. look up join存在什么问题？你怎么解决的？
13. 异步IO怎么实现的？如果想整个程序都实现异步怎么处理？ 封装
14. 主流中两次filter和侧输出流可以认为一样吗？ 侧输出实际就筛选了1次
15. Flink中数据的时间戳怎么理解？
16. Flink的状态最终存储在哪里？有几种选择？RocksDB和HDFs有什么区别？
17. 什么时候选择文件状态后端和Rocksdb后端？
18. 动态分流我认为多余，你们当时怎们考虑的？
19. Flink反压遇到过吗？结合场景说一说？ 外联查询慢

# 长安通信

1. 开发过接口吗？能简单说说吗？
2. FLink和SparkStreaming的区别？
3. Flink中项目中常用的算子？可以简单说说吗？
4. 谈谈对Clickhouse的理解？
5. 简单说说项目中的技术选型？为什么去选择他们？
6. Flink的底层API了解吗？结合场景说一说？
7. 然后问我随便说说Flink中 窗口，水位线，语义化窗口，事件驱动的理解？
8. 维表Join的几种思路？ Lookupjoin 、 旁路缓存
9. FlinkSql和Hive有甚区别吗？项目中怎么实现端到端一致性的
10. Clickhouse几种引擎介绍下？
11. 对实时计算和实时数仓有什么自己的理解吗？

# 明凡科技

1. 大数据组件选型和比较
2. Kafka的底层原理
3. Hadoop的作用
4. 集群资源监控方式
5. 集群工作流程调度
6. MySQL的调优
7. 常见的线程安全和并发容器
8. ioc容器
9. 产品上线过程

# 凯信达

## 面1

1. 两个数据量非常巨大的表怎么Join，有多少解决思路
2. SQL优化

## 面2

1. 自我介绍
2. 两个hive大表join怎么优化
3. 讲讲clickhouse
4. 你们用的什么引擎

# 菱歌科技

## 面1

1. 离线数仓串讲
2. hql的连续性问题
3. 数仓数据量
4. mr的shuffle过程
5. kafka数据积压，有序，丢不丢数据
6. 一些实际生产问题
7. 维度建模
8. 拉链表
9. hive优化

## 面2

1. 离线数仓串讲
2. hql的连续性问题
3. 数仓数据量
4. mr的shuffle过程
5. kafka数据积压，有序，丢不丢数据
6. 一些实际生产问题

# 华翔联信

1. flume能不能，做某个组件的监控， 答能，通过监控 日志文件，对关键字进行过滤。那过滤后的数据要放到哪里呢
2. 你平时负责离线这部分 工作的饱和度怎么样？
3. 写过哪些指标
4. 有的用户一直登陆的15天不下线
5. 用户标签都有啥， 项目怎么实现的
6. 为什么，ads层的数据 要再放到 mysql supset不是可以直接对接 Hive吗

# 踏歌智行

1. 项目介绍
2. 技术选型和比较
3. linux高级命令
4. shell脚本
5. Hadoop读写流程
6. zk选举
7. kafka架构
8. 提高吞吐量
9. hive优化
10. hivesql优化
11. 拉链表
12. ods层压缩
13. 维度建模
14. .java的重载和重写
15. 接口抽象类
16. 数组和链表
17. 线程转换

# 微创

1. 离线数仓的介绍
2. 平时数据量是多少
3. 业务主要是哪方面
4. 数据清洗一般是怎么做的，哪些数据会被清洗呢？
5. spark的宽窄依赖 多对一 一对一
6. transformation 和 action 算子的关系 以及运行的时候会怎么样呢？
7. 说一下顺序存储和链式存储各自的特点和优缺点
8. dwm层的宽表制作介绍一下
9. 离线数仓中的dim层的介绍

# 伯智科技

1. 离线数仓分层介绍
2. 用户拉链表怎么做的
3. hive的优化
4. hive中分区和分桶的区别
5. kafka如何保证消费到的数据全局有序
6. spark中的算子的分类 每个种类讲几个算子
7. HBase的rowkey的设计原则
8. hive中的小文件的处理

# 数字马力

1. 自我介绍
2. 项目介绍
3. flume的作用，怎么实现实时监控和断点续传的
4. MR计算流程
5. shuffle的过程
6. shuffle发生数据倾斜怎么办
7. hive数据倾斜了怎么办（我回答了mapjoin，大小表join，小文件合并，但是他还一直问还有什么）
8. 优化好了怎么从DAG图上确定没问题
9. flink的窗口有几种
10. Flink和Spark Streaming 区别
11. flink的API用过么
12. 一道sql场景题

# 中焯信息

## 面1

1. 负责哪些
2. SQL多还是API多
3. 如何保证Flink处理数据的精准一次
4. 读Kafka怎么保证有序
5. 你们flink用什么语言写的
6. 如何保证使用Kafka高并发

## 面2

只问flink

1. 端到端一致性
2. 项目做了啥
3. 怎么采集
4. 项目里的指标怎么实现的
5. 怎么出报表和展示的

# 迅策科技

## 面1

1. 几个人开发的实时数仓
2. 采集怎么做的
3. SQL多还是API多
4. SQL有遇到大数据瓶颈问题吗，多少数据量出现的，SQL怎么解决的详细说一下
5. 任务在运行中挂了，怎么保证重启后数据不丢失，怎么拉的Kafka数据，详细说一下
6. Kafka的offset怎么提交的，保存的是没消化完的offset还是消化完的offset
7. 长时间没给Kafka响应会不会从新推给你消息？
8. FlinkCDC怎么读的数据，说一下整体实现流程，支持并发吗，如何保证一致性，解释下1.x锁表原理
9. 用过Oracle吗，你们MySQL用的什么版本，你都了解过哪些版本，分别有什么大变动
10. 你接触的数据库有哪些，做过这些数据库的高可用吗
11. 介绍一下流批一体的概念，怎么将流批捆绑起来的
12. 假如一定要将流和批隔离开来，怎样实现，但是他们之间又是能够衔接上的（比如一天24小时，有8个小时是接实时数据，还有剩余的16个小时不接实时流，就想要以批的形式去获得结果，怎样才能将流批过程中自动衔接上）

## 面2

1. flink cdc可以并发吗？
2. flink cdc并发读取如何保证有序？
3. flink cdc怎么对接kafka
4. UV、PV怎么算
5. 宽表任务怎么做的
6. 宽表的关联需要保证时序吗
7. hbase和redis怎么保证的一致性

# 厦门安胜

1. 项目介绍
2. 数据质量监控
3. kafka数据重复
4. flink窗口
5. Awk
6. 有没有用过阿里的组件DW、MAXcompute
7. hive优化
8. 离线项目整个流程怎么做的监控？

# 新晨科技

1. 介绍项目，数据流转架构
2. mapreduse优化
3. hive优化
4. hive数据倾斜
5. hive架构

# 宏瓴科技

1. 自我介绍；
2. 网站日活？答：几十万，订单几万；APP有吗？有；网页多还是APP多？APP多；m站呢？什么是m站？
3. 埋点数据呢？
4. 数仓还是开发做的多？那种语言？
5. java了解怎么样？
6. flink用到哪些java序列化；java序列化用过吗？spark用过吗？
7. hashmap时间复杂度？
8. hash了解吗？
9. hashmap和treemap区别？
10. java多线程了解吗？
11. 多线程和并行度是一回事儿吗？
12. java后端写过吗？多少机器？怎么监控？
13. flink和spark的区别？
14. 怎么理解状态的？
15. datastreamapi多还是tableapi多？
16. 用窗口了吗？那种类型？都有哪些类型窗口？
17. 有基于条数的窗口的吗？
18. 水位线会导致延迟吗？原理？
19. 遇到过反压吗？怎么解决呢？
20. 数据一致性？检查点和保存点的区别？
21. flink怎么提交的；命令行怎么提交？
22. 新任务如何预估资源呢？
23. 你从哪几个方面考虑资源占用？
24. MySQL了解吗？
25. 知道服务器上的reader吗？磁盘reader？
26. CPU多级缓存呢？

# 国信司南

1. 一道算法题
2. 介绍实时项目
3. 为什么使用clickhouse？
4. 介绍watermark
5. 状态后端
6. 端到端一致性
7. 每个小时商品销售topN
8. flink集群规模？
9. flink并行度设置？
10. flink反压? source经过map算子到sink端怎么判断哪里出现了反压？
11. flink怎么定义udf函数？
12. 怎么解析数据？
13. kafka多少个分区？
14. kafka乱序问题？
15. kafka丢失数据？
16. kafka消费策略？
17. 怎么监控kafka？
18. map和maPartitions区别？mapPartitions会有什么问题
19. spark提交流程
20. Zookeeper会不会出现两个leader
21. Linux常用命令
22. 怎么在日志文件中查找一个词
23. 怎么知道一个应用运行时各个进程是否存活
24. 重写和重载的区别？
25. 重载的特点以及在项目中的应用场景
26. 可不可以使用final修饰抽象类
27. Class A继承多个类？实现多个接口？
28. ArrayList和hashMap的区别？
29. static{}和构造器的加载顺序？
30. SpringBoot怎么定义拦截器？
31. SpringBoot怎么获取日志？使用logback替换log4j需要注意什么？？
32. SpringBoot怎么读取配置中的参数信息？
33. java多线程

# 菜根科技

## 面1

1. 自我介绍，怎么接触的大数据，选择大数据原因
2. 项目难点、遇到问题，多说几个
3. 什么是反压，原理、场景，Flink怎么做监控的，具体说一下
4. 生产服务器配置
5. 吞吐量
6. redis做缓存，提高了多少吞吐量
7. 怎么做的测试、测试周期，测试遇到问题，怎么解决
8. 场景题：做一个实时宽表，两张表随时变化，两条流驱动，能关联上，给几种实现思路，用interval join 不到最新数据怎么解决，regular join TTL怎么设置合理，回撤流原理，再说两种其他的实现方案
9. 基于8的场景，再关联几张维度表，维表在redis里面，但是现在整体吞吐量不理想，给几种方式解决，所有能优化的点都可以说
10. 线程池的作用是什么，有哪几种线程池，怎么对线程池监控，不用行不，直接开线程可以不，一条数据一个线程处理有什么弊端，Java创建线程有哪几种方式？
11. FlinkSQL多还是API多，写一个Flink应用程序完整流程（代码）

## 面2

1. 主要负责内容
2. 具体场景题,想通过离线查询出订单金额这个指标怎么查询
3. presto怎么查询？presto查询需要多少时间？有没有比这个更快的？
4. 用CK的sumingmergetree的话具体怎么配置？
5. Kafka了解多少？
6. 数据丢失和重复消费问题？
7. 为什么sink端必须支持事务？
8. 支持事务了为什么就不会重复？
9. flink了解多少

## 面3

1. 项目中遇到哪些难点 怎么解决的
2. 框架搭建的是遇到的问题
3. 有没有做过系统代码调优方面的工作
4. 你理解大数据是个什么东西吗？
5. 大数据和hadoop的关系？
6. 分区 分桶区别
7. hive 自定义函数
8. sparkstreaming 和 flink 区别
9. 各有什么优点
10. 场景题 flink实现 多长表的关联 形成宽表给个思路
11. union 和 intervalJoin区别
12. unoin 和 conncet 区别
13. yarn资源调度方面
14. hql场景题： 各科成绩前三名
15. linux常用命令
16. 数据倾斜处理

# 数秦科技

1. 大数据几年了？在职？离职？
2. SQL题，近一个月连续三天访问？
3. 开窗排序是正序还是倒序？
4. 数据脱敏？答：md5；别的方式？加\*；身份证出生年月日如何隐藏？
5. 数仓分层架构？
6. 开发流程？
7. 调度工具？
8. shell熟悉吗？
9. 如何查看日志？命令如何看日志？如何调试脚本？报错怎么办？
10. 如何修改文件在服务器？如何定位到某一行呢？
11. hive优化做过吗？
12. hive on spark 吗？是；hadoop了解吗？怎么下载文件？怎么load的？
13. 是追加还是覆盖？
14. 指定分区怎么做？
15. spark程序写过吗？
16. spark和flink的区别？哪个好？为什么？
17. Python写过吗？
18. flink 窗口类型？
19. 如何自定义一个窗口？每天00:00:00 到23:59:59？
20. 滚动、滑动窗口
21. 检查点了解吗？
22. 检查点太大，怎么处理？
23. 旁路缓存怎么做？
24. 你觉得自己性格怎么样？
25. 怎么和同事领导团队沟通？
26. 新环境如何快速融入？
27. 快速融入的方式方法？
28. 你有什么问题？

# 兰亭集势

1. 实时项目主要做了什么
2. 实时最终对外提供的应用是哪些
3. 做了实时是不是不用做单独离线数仓；流批一体，实时数仓给离线数仓提供了大部分的数据，提高实时数仓的应用
4. 实时数仓的前景，做完项目有什么收获
5. 用户画像做了哪些标签，性别、年龄、偏好数据是源自于哪
6. bitmap 和 用户分群，基于统计
7. 详细介绍数据域，流量域中做过的指标
8. Flink的压测
9. 实时的集群
10. 为什么使用Flink,Flink四大基石
11. 有界数据和无界数据的区别，常规的Flink、Kafka、Ck技术点
12. Flink中的分区策略(七八种?)
13. Spark和Flink Checkpoint 的理解
14. Hive中的分析函数，sql题：开窗口长度为7的相关需求
15. 职业规划

# 亚联

1. 介绍一下项目：
2. 说明一下 redis 和 hbase 的 一致性问题， 怎么解决的
3. spark 的 map 和 mappartition算子的区别
4. 说一下spark的优化

# 中建科研

1. 介绍实时项目
2. flink反压怎么解决
3. kafka的ack应答
4. hive内部表外部表
5. hive特点
6. hive数据更新

# 华正明天

1. 介绍数仓建模
2. 数据中台接触过吗
3. 数据治理体系了解吗
4. spark优化
5. 导入多张表性能出现问题，怎么解决

# 北京申朴

1. flume组成，会不会丢失数据，event和agent区别
2. 数仓遇到的问题，gzip特性，如何控制切分文件大小
3. 事务型事实表如何实现，单事务和多事务事实表如何设计、区别
4. hive如何实现全局排序，小文件处理方式
5. hbase存储结构、读数据原理，hmaster作用，hmaster挂了影响读吗？
6. hbase如何设计表，rowkey设计，想查最新数据怎么办？
7. click表引擎区别，删除单条数据和删除分区有什么区别？
8. hadoop shuffle优化、hive优化
9. spark提交流程，sc最核心两个组件是啥有啥用？

# 中兴金融

1. 介绍项目
2. 旁路缓存如何实现？
3. FlinkkafkaProducer一致性如何实现？
4. 如果kafka事务的超时时间小于Flink事务的超时时间，还能用两阶段提交保证一致性吗？
5. 为什么数仓分这么多层？
6. 为何要将维度数据放到HBase中？
7. （7）维度关联不上的数据如何二次处理？
8. （8）使用interval join不会关联不上吗？如何处理？
9. springboot开发过吗？Java创建线程有哪几种方式？
10. Java实现线程安全有几种方式？（饿汉式、借助内部类、普通加锁解决、双重检测）
11. JVM了解过吗？内存区域如何划分的？介绍一下堆和栈的区别。
12. hashmap底层是如何实现？hashmap和hashtable的区别。
13. 线程和进程的区别。
14. zookeeper使用过吗？介绍下
15. Linux常用命令介绍下。如何查看Java进程

# 京东乾石

1. 聊聊项目
2. hive优化 及参数
3. hdfs回收站
4. 五道sql题 分组排序题 \*# [计算工时题] [ 节假日 怎么剔除]
5. 开窗函数的考察 比如rank 三种都啥区别
6. java hashmap 结构 怎么往里面写数
7. 多线程
8. 线程池
9. 你有什么想问HR的

# 中电金信

## 面1

1. 你们sql有什么规范，怎么让他可读性更高，你们是用shell调度的，shell sql里面有没有什么注释？
2. 项目介绍
3. sql流程
4. 拿到需求怎么开发，怎么分析，介绍一下分析流程
5. （5）主要就是sql离线相关的，到银行写sql
6. （6）maxwell倒错数据怎么办？
7. （7）kafka怎么不丢数？
8. （8）我想要用户数据前三天的怎么办？
9. （9）你们kafka数据保存几天？
10. （10）你们kafka分区怎么弄的？联信宏方
11. 数据倾斜
12. 来需求，怎么分析，要不要和业务人员沟通，是不是你自己沟通？
13. 你未来3年发展规划？你想当经理要注意哪些问题？
14. flink.streaming用的多不多
15. 画像你计算过那些标签，说几个
16. 说说你熟悉的项目
17. 你们日志用flume，业务用data x这俩有啥区别？
18. sqoop用过没
19. 给你需求，你怎么实现分析它
20. 项目你自己主要负责哪部分

## 面2

### 一面:

1. MySQL中的join有哪些，
2. inner join 和leftjoin的区别，
3. 项目流程
4. 自我介绍

### 二面:

1. 问了项目流程，
2. flinksql和flinkAPI有什么区别，
3. 指标怎么写的，
4. 反压定位和解决，
5. flink的分区，
6. 你们有几台服务器，怎么划分的组件，服务器大小
7. hive中你用过的优化
8. 数据湖和数据中台的区别
9. 说说你遇到的难题

# 新致软件

## 面1

1. 做了3年大数据，拿到指标，有没有参与表结构设计？你认为什么表对你影响最大，学到最多？
2. 用户连续登陆天数大于10
3. 遇没遇到小文件问题
4. 数据倾斜场景及解决
5. 你们集群，比如说hdfs坏了，怎么解决，是停掉还是怎么？
6. DS半夜出问题，假期出问题，怎么解决？
7. bitmap好处，是不是什么场景都适用？
8. 为什么存到CK？
9. sparkSQl调优
10. 维度建模 模型几种，为什么使用星型模型？
11. superset谁做？

## 面2

1. kafka吞吐量？
2. acks isr列表的理解
3. 消费者如何保证不丢失
4. 从kafka中拉取数据写入到HDFS中程序崩溃了，在启动程序拉取，数据还会被拉倒HDFS中吗
5. kafka、hadoop版本号选取是如何考虑？
6. hive、hadoop、spark的高可用如何实现？
7. 节点出现故障，如何保证hive的SQL任务不受影响？
8. rm、nm的容灾和高可用如何配置？
9. hadoop和yarn的高可用原理？
10. rm、nn依靠什么组件或功能实现高可用？
11. 数据全量增量同步，Maxwell的配置步骤，数据源配置了多少张表？（表的数量级？千万级/亿万级）
12. nameNode分配多少内存？10台以内的服务器配置要求？
13. 行存和列存的区别？效率差异
14. Flink和Clickhouse建立实时数仓的理解？
15. 使用Maxwell对待update的表操作时是如何处理的？（细节：DWD每张表处理逻辑不同）
16. hive为什么不建议做update？在哪个地方会受到限制
17. HBase的删改操作如何实现？
18. Hive的ORC事务表了解吗？apache kudu有了解吗？
19. 什么是mpp数仓？（MPP 是将任务并行的分散到多个服务器和节点上，在每个节点上计算完成后，将各自部分的结果汇总在一起得到最终的结果。）
20. 接触过哪些商业数据库同步工具？
21. 了解Oracle DataGuard吗？

# 力石控股

1. 一面简单问了问
2. 用户画像标签分类，指的是规则、挖掘这样，不是一二三四级
3. 介绍一下技术栈
4. 谈一下你对flink、spark、hive的理解
5. 两个数仓分层
6. 介绍一下spring boot、spring cloud
7. SSM和spring boot的区别
8. 谈一下你对HBase的理解
9. redis是怎么用的

# 海尔科技

1. 实时数仓项目流程
2. flink cdc怎么用的，和maxwell对比一下
3. 动态分流：
4. maxwell采集到kafka，会不会有数据乱序问题，怎么解决
5. 宽表怎么做的
6. ck怎么用的，用到了哪些表引擎
7. 存宽表， ReplacingMergeTree
8. Final 多线程
9. ck工具类怎么写的，具体到代码，设置批次等等
10. flink cep怎么用的
11. 用户画像重点问的
12. 标签的执行流程
13. 怎么做的任务监控
14. spark任务怎么提交的，提交参数怎么设置的，为啥这么设置
15. bitmap表怎么设计的，字段哪些，类型
16. 怎么写到CK里面的，API是什么
17. ck怎么去重
18. 人群包的接口怎么写的
19. 离线数仓怎么搭的

# 华软科技

1. HA 实现的原理 (难度不止于nn 预写入日志 2nn zk 通信)
2. ZKFC原理
3. hive的事务
4. acid表
5. hive on spark的优缺点
6. dfs.ha.fencing.methods
7. hadoop.proxyuser
8. DefaultContainerExecutor 和LinuxContainerExecutor区别
9. 说说hive.txn.manager

# 全拓数据

1. 讲离线项目
2. 你们集群机器有几台
3. 业务总表有多少,hive形成的仓库表有多少,大致分了哪些主题
4. 总商城的用户有多少,日活呢?
5. hadoop启动有哪些进程?
6. ZKFC和Jounal~~是干嘛的?
7. edits是什么?
8. hadoop读写过程,有过哪些优化
9. shuffle优化有哪些东西?
10. yarn的调度,spark的调度
11. MR的程序写过么,MR底层如何写会不会写?
12. Kafka的架构?
13. kafka有过宕机么
14. Kafka的数据量,怎么宕机的,宕机的原因
15. kafka副本数,多少个topic,多少分区?
16. 分区数以后如果数据量大了怎么办
17. isr解决什么问题?具体有哪些配置
18. 延时条数和延时时间?
19. kafka数据丢失怎么配置的?
20. kafka参数的优化?
21. HIVE的基本架构?启动有哪些进程
22. hive常用参数的配置?
23. hive常用函数?
24. Rank() dense\_rank() row\_number() lag() lead
25. UDF,UDTF写过么?
26. hive调优
27. spark的计算原理,跟hadoop的区别
28. 任务的切分,咋运行?
29. spark广播变量,累加器
30. 累加器代码层怎么实现?
31. 开发具体用那种语言
32. 接口和抽象类的区别
33. java线程的状态
34. spark的语言用什么写的
35. scala的集合有哪些类型
36. 可变和不可变的集合区别
37. scala的伴生对象的用法
38. 模式匹配
39. 样例类如何用
40. 隐式转换

# 小牛冲冲冲

1. （1）实时入库和离线入库时，数据修改的策略？考虑点是什么？对订单状态有什么影响
2. join操作在哪里做的？
3. kafka存储的数据结构是什么？
4. 源数据字段变成varchar（修改表结构的先后顺序？）、字段类型int变string会有什么影响？
5. 指标需求滞后于功能的需求该如何处理。
6. DataX增量处理如何操作？
7. Where 主键范围、日期范围
8. Maxwell处理过哪些操作？insert？update？delete？
9. 日志有没有回补的逻辑？对实时有什么影响？
10. 用户行为数据在离线场景的使用？今天20日看到19号的日活和明天21号看到19号的日活一样吗？

# 医药魔方

1. 数据治理这块你做了哪些？
2. 元数据管理你们是怎么做的？
3. 你们怎么做数据质量监控的？
4. 元数据的内容有哪些？
5. 你是做数据中台的还是数据治理的？
6. 阿里的产品你会用吗？
7. Flink窗口中滑动窗口和滚动窗口的区别和使用场景？
8. 怎么实现Clickhouse中大量数据的修改？
9. 怎么实现数据一条一条的写入Clickhouse？
10. Flink中怎么保证精准一次性？
11. clickhouse用过哪记种引擎？
12. clickhouse种的表怎么关联的？
13. Concurrenthashmap你了解嘛？

# 华控智家

1. 数据治理这块你做了哪些？
2. 元数据管理你们是怎么做的？
3. 你们怎么做数据质量监控的？
4. 元数据的内容有哪些？
5. 你是做数据中台的还是数据治理的？
6. 阿里的产品你会用吗？
7. Flink窗口中滑动窗口和滚动窗口的区别和使用场景？
8. 怎么实现Clickhouse中大量数据的修改？
9. 怎么实现数据一条一条的写入Clickhouse？
10. Flink中怎么保证精准一次性？
11. clickhouse用过哪记种引擎？
12. clickhouse种的表怎么关联的？
13. Concurrenthashmap你了解嘛？

# 绿盟科技

1. hadoop在大数据领域是个什么位置
2. hdfs写流程
3. Zookeeper怎么实现分布式锁
4. redis有没有数据一致性的问题
5. redis常用数据类型
6. spark的状态和flink的状态
7. scala觉得咋样
8. 他说他们不用hive，也不会用hive
9. kafka调优，从生产者消费者方面考虑
10. hbase region的切分流程
11. java这门语言哪些部分比较重要

# 尚博信

1. dataphin、dataworks……用过吗
2. flume和kafka采集的时候有神区别
3. flume在采集的是用到了哪些相关的组件
4. clickhouse存储引擎哪些
5. clickhouse支持删除么
6. clickhouse删除一条数据怎么办
7. clickhouse列式存储还是行式存储
8. clickhouse优化
9. bitmap说下
10. 使用spark涉及的组件
11. spark shuffle 怎么处理的
12. spark任务执行慢，参数怎么调优
13. 场景：三台服务器，16内核，32内存，运行spark任务，如何规划参数

# 广州易星

1. hadoop集群在公司中是怎么搭建和规划的？
2. hdfs里面清洗数据的流程
3. hdfs的小文件是怎么处理的？
4. 小文件的危害？
5. hadoop中的shuffle中反向溢写为什么不能设置成100%？
6. hadoop中发生数据倾斜怎么解决？
7. yarn的调度器有哪些？
8. zookeeper的机器，假如有3台，isr[1,2,3],2是lead,lead挂掉后，下一个lead是谁？如果是8台zookeepper呢？为什么能安装偶数台？
9. 解释下零点漂移？
10. flume的中channel的选取？
11. kafka中lead挂了的话它怎么选举的？
12. 大概在什么业务场景下用了UDF和UDTF？
13. Hbase的读数据的流程？memstore和blockcache哪个数据更新？

# 名通科技

技术面:

1. Flink中做了哪些优化？
2. FLink与Spark Streaming的区别？
3. java的基本数据类型？
4. Flink你了解的内容？
5. 说一下你在开发过程中影响最深的地方和项目？
6. 数仓搭建时自己搭建的还是有其他人的参与的？
7. 数仓搭建的时候有什么数仓规范吗？
8. 数仓分层结构时什么样的？
9. 职业规划时什么？
10. 平时开发用的是java还是 scala？，java中重写和重载有什么区别吗？
11. StringBUffle和Stringbuilder的区别？
12. java中创建对象的方式有哪些？
13. 单例里面是怎么创建对象的呢？
14. 大数据平台的一些技术有哪些？
15. 你认为你掌握最好的3个组件？

# 浩瀚深度

1. 你在项目中遇到的问题
2. 项目中担任的角色
3. SparkStream和Flink的区别
4. 如何实现断点续传功能

# 如果科技

## 面1

1. 疑难sql的解决思路,我说了两个七天连续三天,留存
2. 你觉得做了这个离线项目对自己有什么进步
3. hive的优化

## 面2

1. 自我介绍
2. 实时项目遇到问题及 解决
3. 数仓建模的流程
4. flink的失败重启策略
5. flink和sparkstreaming区别
6. flink的分区策略
7. flink的精准一次
8. flink的并行度怎么设置的
9. 算子的中并行度具体怎么设置
10. 你有什么想问我的。

# 普元信息技术有限公司

1. 介绍项目经验，负责哪些具体工作
2. 介绍完整的ETL流程
3. Hive优化做了哪些
4. Kafka数据重复、数据丢失有没有？怎么解决的？

# 陶乐

1. 日活
2. 产生的订单
3. 数据量有多少
4. phoenix有没有做高可用,读取phoenix的时候是怎么读的,连接的zk还是?如何做的优化?
5. clickhouse TopN怎么写,引擎是什么
6. 优化前后,读取维表的时间是多久
7. flinkCDC版本,mysql版本
8. 用户维表字段

# 广微数据智能科技有限公司 阳光保险

1. 事实表有多大，事实表多少字段
2. hive总共有几台节点，节点配置
3. HQL的优化
4. hive分配多少内存
5. 维度设计要点
6. 你负责的业务是怎么划分的？
7. 维度模型有哪几种？解释下含义
8. 为什么要做拉链表？
9. 维度表多宽
10. 存量表有多大？

# 南京星邺汇捷网络有限公司

1. 说了遍离线项目流程
2. 怎么做拉链表
3. 怎么查拉链表中的当天新增数据
4. Hive内外部表 load命令怎么写
5. zookeeper的选举
6. hive的优化
7. 维度数据是纵表还是横表
8. 数据倾斜处理

# 国御科技

1. flink实时项目串讲
2. 实时处理的数据（日增）怎样的时间线（每秒多少条数据、每分钟）
3. 离线项目数据量
4. 离线一天处理多次,还是次进行处理
5. 处理多久
6. 其中有很多复杂的sql或业务逻辑在里边么?
7. 涉及任务的调度 -> azkban
8. 每个任务的耗时进行分析么
9. 优化的部分做过那些?
10. 小文件怎么处理
11. hive的内部表和外部表
12. 有没有使用的hive的一些函数
13. 自定义用过么?
14. hive解析json字符串
15. 所有的hive的任务都会使用mapreduce去执行么
16. 为什么对数据仓库进行分层
17. 数据建模了解么
18. spark的工作原理
19. 如何估算集群的机器数量
20. 写过后台接口么 使用的啥 springboot
21. spark做什么角色
22. es有了解么,大概了解多少?
23. es如何实现master的选举的,有过优化么
24. hive使用数仓搭建的时候,按照什么样的维度进行分区

# 中国联空

1. 讲项目
2. 遇见过的问题
3. kafka怎么不丢数据
4. zk客户端原理
5. zk节点的原理
6. zk存储
7. spark的广播
8. spark的reducebykey groupbykey区别
9. flink3个保存点机制
10. flin项目做过哪些 遇到什么问题
11. hive数据倾斜 小文件

# 银丰新融

1. 写过哪些宽表
2. 你做过哪些建模
3. 说一下你订单宽表的创建流程
4. 宽表创建之后你进行了哪些测试
5. 说一个指标的实现过程（从需求分析到报表展示）

# 阿里

1. 自我介绍
2. 介绍实时，离线项目
3. dataX问题
4. flink反压倾斜
5. hive倾斜优化
6. java 内存模型
7. java GC
8. 为啥用kafka，优势，遇到过啥问题、挂了咋办
9. flink提交流程

# 东方国信

1. 自我介绍
2. Hadoop集群规模
3. 数据量多少
4. Hadoop这块遇到过什么问题没有，怎么解决的
5. Hive遇到过什么问题
6. kafka遇到过什么问题
7. kafka的数据保存策略时什么
8. 讲一下离线项目，其中遇到哪些问题，怎么解决的
9. 分区表有用到吗，怎么设置的
10. 讲一下实时项目
11. 为什么用Clickhouse
12. HBase的架构
13. HBase写文件流程
14. HBase遇到过什么问题，结合项目
15. HBase 每天的数据量
16. Flink提交作业流程

# 成都国蓝中天

## 面1

1. Linux常用命令
2. （2）如何查找不同文件里面的同一个字段？
3. （3）使用Spark引擎遇到过什么问题？
4. （4）离线数仓串讲（没讲完就把我打断了）
5. （5）FlinkSql用的多还是Api用的多？
6. （6）实时数仓dim层实现方式？
7. （7）离线数仓遇到过什么问题？

## 面2（二面）

1. 自我介绍
2. 有没有负责过一个项目的搭建和落地，讲一下过程、自己的作用
3. 项目中遇到的最大的问题
4. Kafka消费能力不足怎么办？
5. 大数据工程师与业务人员应该怎么相处，是否认为自己有技术可以无视业务？
6. 大数据工程师的基本素养有哪些？是否一个后端人员会大数据组件和开发，就可以称他为大数据开发工程师了
7. 为什么选择在这个时间点离职？
8. 为什么选择我们公司？
9. 你认为你的技术能给我们公司带来什么？

# 中国华信邮电科技有限公司

1. 搭建hadoop之前机器需要做什么
2. Yarn工作机制，提交作业流程
3. Hive调优、hive底层存储使用什么
4. Hive的\t 分隔符 如果字符串中出现\t怎么办
5. Checkpoint设置、Flink做过的优化
6. Flink Watermark的生成原理
7. Flink-CDC的一些问题 采集的什么数据库
8. Linux系统字符串替换
9. SQL调度用什么、设置过什么参数
10. 有没有了解过Spark

# 平中软件

1. 自我介绍。
2. Flink的开发流程。
3. spark与Hadoop有什么不同。
4. 先问你看了什么好书，给我们推荐一下，后来又问了你现在在看什么书。
5. spark的内存模型。
6. spark的图结构
7. Linux常用命令，简单说十几个。
8. Hbase有什么优化
9. redis用过吗，有什么用
10. 说一下三个项目所用的技术，你在其中负责了那部分工作。
11. 为什么要用Kafka，有什么优势。

# 厦门宏网信息（邮政外包）

1. 自我介绍
2. 离线数仓和实时数仓主要的区别（架构，应用场景）
3. 对数据分层如何理解的，如何设计的
4. 数仓脚本的调度用的什么工具
5. 数据倾斜（结合场景）
6. 零点漂移问题（结合场景）
7. ads层后怎么做的
8. ads层做过打包成服务这种方式吗？（没有）
9. 最近在忙什么
10. 有什么想问的

# 北京易联达（邮政外包）

1. 怎么更换引擎，hive的mr怎么更换为spark
2. hive数据倾斜，深入；
3. MR流程
4. hive存文件切块，怎么切

# 纳米达克斯

1. 自我介绍
2. 组内人数
3. 离线项目串讲
4. 离线项目遇到过哪些问题
5. 如果一个表维度变化非常快，你该如何设计表
6. ES了解吗（略）
7. 架构中kafka的作用
8. 实时项目串讲
9. 架构中hbase的作用
10. flink反压
11. 反压场景中数据倾斜具体怎么解决的
12. redis为啥这么快，具体配置过哪些参数变得很快，多路复用了解过吗
13. 报表怎么做的
14. 平常大数据的项目怎么测试的

# 网易游戏

1. 项目介绍
2. spark中的shuffle有哪些
3. Hbase的优势，为什么使用了Hbase 设计rowkey？
4. Flink的四大基石都有哪些？
5. Watermark的作用是啥？如何保证数据不丢失？
6. Spark为什么比mapreduce快
7. Hdfs写数据流程介绍一下
8. Hadoop1.x和hadoop2.x的区别
9. Flink的window实现机制
10. 讲一下Flink onyarn的部署
11. 算法题

有一个数组,其中有一个数出现次数超过半数，如何找到这个数

有个长度为n的数据，对数组排序

A[n/2]

Int a ;

For(){

A+1;

}

A >a.length（）/2;

其中有两个数出现次数都超过了30%，如何找到这两个数

# 大童保险

## 面1

1. 自我介绍
2. 工作内容
3. Ods数据量多少
4. 云主机还是物理机
5. 怎么分区的
6. 增量和全量数据都同步过来吗？
7. Datax运行多久
8. Ods有多少张表
9. 具体说一下dim层
10. 具体说一下dws层
11. 开发工具
12. Ads从哪拿数据
13. Dws覆盖多少指标
14. 海豚调度器遇到什么问题
15. 调优 数据倾斜 解决
16. 我说了一下实时框架
17. Dim为什么放到hbase里
18. 期望薪资多少
19. 对未来的规划
20. 了解大厂的哪些项目
21. 自己性格怎么样
22. 和产品经理交流多吗
23. 有什么想问的吗？

## 面2

1. 自我介绍 ？
2. 最近做了那个项目，都干了什么？我说统计新增用户，回流用户。问我咋实现的，懵了，说完后说只统计这个有必要做分层实时吗？
3. 根据你简历上写的内容他详细问怎末实现的，数据从哪里取，sql如何写的，你写了他就问，一定要把简历上写的东西弄会。
4. 理论就问了Maxwell偏移量在哪里维护的，你还了解哪些数据库。

## 面3

1. 自我介绍 ？
2. 你们公司在哪？
3. 做什么的 ？在线教育
4. 用户量
5. 实时数据量多少 ： 不到一个G
6. 实时做了哪些内容
7. 独立访客，新老访客 ，
8. 维度数据怎么处理 ： 维度表 通过FlinkCDC
9. 为什么用Hbase ， 不找地方存起来不太好，为什么不用mysql？ 根据流行框架（分层处理）
10. 离线做了什么 ？
11. 关注用户转化，举个例子
12. 用户路径分析，分组，分析路径，从哪跳出，以及用户集中访问哪一块
13. 漏斗模型 ，用户流失类的
14. maxWell 用过没
15. 离线这边多少张报表， 涵盖多少个指标 ？ 指标构建体系
16. 离职？问一下原因
17. 个人规划
18. 实时主要写sql吗？
19. 题外话： 说很多简历相同，差不多都是这个项目，我说画像不是我做的
20. 对自己哪些技术比较自信？
21. Hbase key键， 键设计？ 我说 9 位 加 字母 ， 没说用Phoenix
22. 说我有什么想问的嘛？
23. 期望薪资。
24. 哪个学校毕业的

## 面4

1. maxwell底层实现
2. datax底层实现
3. 数据都存在Hbase吗
4. redis干什么的？
5. Hbase里面有多少数据，Hbase里面有多少簇
6. redis缓存最大的时候里面有多少数据

# 最右

1. 自我介绍
2. flink 端到端一致性
3. kakfa的调优
4. sparksql 使用（用户画像，）
5. 10M 数据 1M 内存 怎么定位某一个数据
6. flink 两阶段提交
7. HQL ， 作品id ，用户id ，性别， 求每个作品的性别人数
8. 算法题 矩阵的最小路径和
9. 你有什么问的

# 卡特加特

1. 自我介绍
2. 为什么离职
3. 画项目整体架构图，并且介绍项目
4. 如何导入文件这种非结构化数据，比如如何导入Excel表，5张不同来源的excel表呢？
5. 采集端介绍
6. 维度建模理论
7. 服务器有几台？
8. 参与了哪些项目？
9. 遇到了哪些问题，印象最深刻的是哪个？
10. 最快多长时间入职？
11. 你还有什么要问的吗？

# 北京恒昌利通投资管理有限公司

1. 自我介绍
2. 看简历
3. 介绍一下大数据组件及其功能（额外讲了Flume版本问题和DataX问题）
4. 知道怎么存储数据（hdfs写文件流程）
5. Zookeer是干啥的（刚才组件忘记讲了）
6. 公司常用数据源有哪些，回答就用过MySQL，会从数据源抽取数据吗（我们公司数据源多MangDb，App等等）
7. 公司数据量多大，百G左右（论坛+电商）所以大部分是日志数据
8. 学历可查吗，上家薪资多少，这家预期多少
9. 三年规划，有啥想问的
10. 能搭起来集群吗

# 瑞达沣通信技术（陕西）

1. 离线串讲
2. 写Sql提交运行的时候，会返回什么信息？
3. 数据倾斜怎么解决的？
4. Kafka的broker信息存在哪里的？
5. Hbase的刷写策略
6. Redis的三种特殊数据类型
7. 是否接受出差，期望薪资多少
8. Docker会不会？

# 东方鑫源

1. kafka分区策略
2. Kafka重复消费
3. flink插槽、subtask、task关系
4. hive的/t分隔符作用
5. hive数据倾斜
6. 集群规模，数据量
7. 旁路缓存为什么要用redis
8. checkpoint异步分界线快照
9. 平时在公司做什么
10. 采集组件，流程
11. datax怎么用
12. 为什么离职
13. flink测输出流
14. flink提交流程
15. taskmanager大小，jobmanager大小
16. flink调优

# 八维学院

1. 自我介绍
2. 介绍项目
3. flume，kafka、hive版本
4. hive就包含什么组件，各组件都是干什么的
5. hql提交之后转换流程
6. flume包含什么组件，channel就了解那些，他们之间有什么区别
7. flume中taildirsource怎么实现的
8. kafka熟悉吗？
9. kafka包含什么组件
10. spark熟悉吗
11. 然后连着问了三道SQL题（让说出实现用到的算子）
12. 项目使用的什么框架
13. 你有什么问题问我的

# 航天信息

1. 自我介绍
2. 写一个sql题，科目行转列。多了一个算总分加排序，最后要得到的字段是
3. 排名，姓名，语文成绩、数学成绩、英语成绩，总分
4. 列举三点hivesql 和标准sql三点不同的地方
5. hdfs文件系统有什么缺点，列出三点
6. 列举hbase三种不适用的情况
7. 列举redis三种不适用情况 说一下click house缺点？
8. 装过spark吗，安装spark要让spark工作起来要装那些什么，都是什么版本，大概说一下
9. 有一个网站，看排名前三的热词，假如要用spark来实现，用那些算子来实现
10. 要实现两个表的左外连接，用MapReduce如何实现，谈谈你的理解
11. 怎么给shell传参，使用shell调用系统命令或系统工具怎么使用
12. Linux系统的环境变量你知道多少？我说了my.env文件，又问了就这一个地方吗？还能再说几个？
13. 最近工作上解决了什么让你印象深刻的过程
14. 怎么把Kafka已经消费的数据找回来
15. 你有什么想问我的

# 中移金科

1. 自我介绍
2. 讲下具体项目
3. 团队多少人？
4. 和哪些部门对接需求？需求定义到什么程度？项目中谁分配需求？
5. 公司卖啥的（具体商品）？
6. 数据量多少？
7. 离线4个人，主要干什么？
8. 呆了4年，为什么想离开？
9. 以后的工作规划？
10. 有什么想了解的？

# 宝马外包

1. 自我介绍
2. sql写的怎么样
3. 为什么从开发转测试
4. 能否接受加班
5. 又没有用过jira
6. 期望薪资

# 中科院新科所

1. yarn可以调用GPU吗
2. yarn可以到达什么粒度的资源分配
3. Flume可以不通过学的那一套框架到达Hive吗，可以通过Flume直接发送到Hive吗？如果让你实现你有什么想法
4. hdfs负载均衡，其他节点90%占用，别的怎么均衡

# 杭州数理

## 面1

1. 介绍实时项目
2. FlinkSQL用的多还是Api多，我API偏多，然后他问为什么不全部用FlinkSQL
3. 介绍一个用FlinkAPI实现的复杂场景
4. 讲一下你状态调优做了什么事
5. 状态的数据量
6. 检查点恢复时候有没有遇到什么问题
7. 整个streamGraph流图改变了，在从检查点恢复还能恢复的起来吗？
8. 每天数据流量
9. flink的部署参数怎么设置
10. 具体是怎么衡量的这个参数
11. flink端到端的一致性保证
12. barrier对齐是什么样的机制
13. barrier是由谁来发起的
14. JobManager和TaskManager的心跳
15. flink的反压的处理
16. 数据倾斜的场景是求平均值类的是怎么处理。
17. 还有其他场景造成的反压么（除了数据倾斜）。
18. 访问Hbase 和redis的查询效率
19. Clickhouse的ReplacingMergeTree去重的时机。
20. ReplacingMergeTree手动final 对性能影响
21. 业务中查询Clickhouse每秒访问次数
22. 有没有遇到过双流的join
23. SQL 里面的 状态过期怎么设置
24. FLink的 几种join方式
25. 用FLinkCDC建设维表的过程
26. FLinkCDC 每张维度表 配置了 一个信息吗
27. java 里面的static的关键字
28. java里面 final 关键字的作用
29. arrayList 和Linklist的区别
30. arrayList扩容
31. 双亲委派机制
32. jvm运行区
33. jvm的 垃圾清理造成flink的反压有了解过吗

## 面2

1. 目前状况
2. 离职原因
3. 讲一个你认为体现你能力的，用了那些技术栈
4. maxwell+kafka 和flinkcdc的区别
5. flinkcdc和maxwell支不支持断点续传，支不支持增量采集，性能方面有什么区别
6. 要是用FLinkCDC要是监控多张表是启动一个程序还是多个程序。你们用的那个版本
7. 中间有几个问题 纠缠了 一会 ，总感觉他想diss我
8. flink的DWS层维度是怎么关联的
9. 异步IO具体是怎么做的，
10. 这个异步IO操作是你自己写的还是官网就有的
11. redis用的什么模式
12. 你还有遇到其他挑战么？我说了flink的新老访客标记的大状态问题
13. 状态后端有几种？我说1.13版本有两种，之前的版本有三种。 我说1.13版本有两种 他说不可能。
14. 窗口有用到么？用了哪几种
15. interval join 的一个场景题， a b 两条流数据过来 底层是怎么 join的
16. 窗口里面的数据是怎么老化的
17. taskManager内存了解吗，
18. 通信内存，taskmanager和jobmanager通信 的Akka通信内存是哪里分配的，栈内存是怎么分配的
19. 你有什么想问我的。

# 中国福富

1. 介绍实时项目
2. 说一下你负责的
3. flink那些组件构成，及其作用
4. flink的内存管理
5. flink的定时器，提供了几种定时器，及触发条件
6. process里面的onTimer和处理数据的方法线程是安全的么
7. 接着第6题，如果不是同步的话会有什么问题
8. 介绍一下flink的水印
9. 简单介绍一下自定义trigger（我简历里面写了）
10. Trigger里面的四个返回值
11. 用Flink去读取6个月的数据，每天的数据是2T，数据里面只有一个ip字段，对所有的数据做一个ip去重
12. 你们一天的数据量
13. flink滑动窗口的处理
14. 1亿条访问记录，内存放不下，取出当天访问 次数最多的ip
15. flink的你用的是java还是什么？
16. 你有什么想了解的。

# 成都中国软件

1. 自我介绍
2. Hive的数据倾斜遇到过吗？
3. Hive的调优有哪些？
4. 讲一下你负责过那些项目？
5. 你们离线数仓怎么分层的，你都做了什么？
6. Flink如果做到数据对账？（数据一致性）
7. 用过哪些数据库？高斯用过没有
8. 你统计过哪些指标？什么叫做指标？

# Boss

1. 自我介绍
2. 哪个项目做的比较好
3. 数据源
4. 实时整体架构
5. ETL分开做会不会导致数据不一致（这个问题聊了15分钟）
6. ck更新数据导致结果不一致，final造成所有分区合并怎么优化
7. ck数据量
8. es了解吗
9. flink写销量top10

# 北京数聚智联

1. 自我介绍
2. 挑一个项目介绍一下
3. 做过数据清洗吗？
4. 懂Java吗？
5. 做过数据治理吗？
6. 如果遇到新的组件是通过什么方式学习的？
7. 用过哪些数据库
8. clickhouse有什么好处？有什么可以替代吗？
9. 有什么要问我的？

# 成都达实旗云健康

1. 自我介绍
2. Flink遇到的问题？
3. 数据倾斜、反压怎么定位和解决？
4. 为什么RocksDB就可以解决内存问题？
5. Kafka应答等级改为-1之后，还是会丢数据，你有深究过原因吗？
6. 为什么要采用Flume、Kafka这些组件？他们的内存配置怎么分配？
7. Ck的稀疏索引、MergeTree讲一下
8. Hive的优化有哪些？
9. 数据倾斜的解决方法
10. DS的调度你们是怎么设置的Worker？有几个任务流？

# 白龙马

1. 自我介绍
2. 每个项目阶段做多久
3. 离线窜讲（主题划分 DIM拉链表实现 DWD累计快照实现 最细粒度是什么）
4. ADS需求提出到交付的开发流程（我有点蒙住了 应该是拆成衍生指标那一套）
5. 测试的同事对数据怎么验证的
6. 程序几点跑
7. ADS层产出几点出来，有没有晚的时候
8. DS报警怎么报的？半夜起来调什么？有什么场景？（说的重启和倾斜场景）
9. 实时双流join
10. 监控项目
11. 他介绍了公司（早上10点 晚上10点 ，重点告我说这不是在加班 是正常工作时间）问我能接受吗？
12. 反问3个问题，他介绍了将近15分钟

# 成都神策

1. 自我介绍
2. 讲一个你最拿手的项目
3. FlinkCDC用来做什么
4. 数据采集、分析、处理都是你们在做吗，具体做了什么，怎么保证数据质量
5. Java线程池的类型有哪些
6. 用了哪些数据库，分别在哪里使用的
7. Springboot做的是什么，还会不会其他的框架
8. Kafka的吞吐量怎么解决的
9. 你对加班怎么看？

# 成都神州数码

1. 介绍下你的项目经历
2. 介绍下Flink实时数仓的流程
3. 用过哪些数据库
4. Flink中异步操作如何实现的
5. Mysql数据库中的数据如何写到hbase的
6. Kafka的积压如何处理的
7. 对于996可以接受吗

# 成都普华永道

1. 自我介绍
2. Flume的拦截器怎么实现的
3. Kafka吞吐量如何提高
4. 给hdfs装载数据的方法有哪些？Load装载语句用在本地和用在文件系统有什么不同
5. 离线数仓的每层做了什么事情
6. 拉链表怎么实现的
7. 同比、环比做过没（没有）
8. 数据倾斜怎么发现的，如何处理
9. Flink数据倾斜了如何优化

# 华泰外包

1. 自我介绍
2. 说说离线项目和实时项目
3. hive调优
4. 实时kafka放ODS配置什么
5. 为啥用CK和Hbase
6. 为啥用redis
7. 为啥离职
8. 没有找到下家你就离职 为啥
9. 第一家公司都干啥了 涉及到大数据方面的东西没
10. 为啥转专业
11. 有啥规划
12. 有看啥新知识没
13. 有啥想问我的

# 北京宏瑞汽车

## 一面

1. 自我介绍：简单介绍下项目
2. 问了实时，可视化这块内容
3. 用的哪些组件
4. 介绍了项目中整个数据流程
5. 数据量大如何处理
6. 简单介绍了下公司业务
7. 场景题，汽车的启动，充电，等数据在数仓中的处理流程
8. 面试官简单介绍了下公司的情况，当前互联网部门的状态

## 二面

1. 自我介绍
2. 介绍实时项目
3. 问题：数据处理过程中数据量大如何处理
4. 问题：你们可视化如何做的
5. 介绍了下公司团队情况
6. 公司前期工作地点能否接受

# 讯锡科技

## 面1（京东中海油项目）

1. clickhouse主键如何设置
2. 旁路缓存具体怎么实现的
3. 遇到过哪些问题
4. kafka存几天，clickhouse存几天
5. 用户画像架构，做了那些事
6. 一道计数的简单sql
7. 给我介绍了下他们的项目

## 面2

1. Kafka的丢失，重复，积压怎么解决
2. Kafka的数据重复是由什么原因导致的
3. Kafka怎么保证有序
4. 指标是从哪里接入的
5. 拿到指标怎么分析的
6. 怎么保证指标的准确性
7. 怎么进行建模的
8. 为什么要有DWS层，直接从DWD层取数据不行吗
9. 实时数仓实现每隔1分钟实现过去10分钟的topN
10. 时间戳转为日期的函数是什么

# 峰潮信息

1. 实时简绍
2. 集群配置
3. flume 挂了怎么办
4. 为什么要分层，那层做了哪些工作
5. flume具体怎么配置的
6. 怎么控制数据速度
7. 异步io，旁路缓存怎么实现
8. flink有哪些架构
9. hbase怎么配置

# 云端微服

1. hive小文件。
2. flume底层。
3. 分层架构。
4. hbase的读流程。
5. hdfs的读流程
6. hdfs优化
7. redis穿透 数据是什么格式
8. kafka重复
9. maxwell的原理

# 德国电信

1. 对spark做一个概述
2. spark和hadoop区别
3. spark的yarn任务提交流程
4. spark重要组件
5. spark算子分类
6. 转换算子和行动算子区别
7. SparkSQL中RDD、DataFrame、DataSet
8. Repartition和Coalesce区别
9. reduceByKey与groupByKey的区别
10. 宽依赖和窄依赖概念
11. spark广播变量和累加器，实际生产应用
12. driver两种运行模式
13. 一种rdd算子实现Wordcount（Java/scala），口述，我直接口述的代码
14. spark生产环境问题数据倾斜
15. spark的OOM问题1）什么时候会产生2）在哪里查看3）怎么解决
16. Java方法重载和重写
17. 捕获异常、进程和线程区别
18. jvm的oom问题
19. 另一个面试官，英语水平怎么样，听说读写那个强，用英语介绍一下您们的项目
20. 中间问了一些Linux常用命令

1）查看内存

2）查看进程

3）查看端口占用情况

4）查看某个进程内存

5）查看隐藏文件并排序

# 华付云

1. 介绍离线流程
2. 业务数据怎么做的。更新数据和历史数据怎么处理 ，如何合并
3. ods层过来之后如何处理
4. 数仓各层都做了什么
5. Dwd层是单事务还是多事务，都有哪些业务过程
6. 公司数据量

# 重庆新致

1. 日志采集
2. Kafka中数据如何写入数仓
3. 有多少分区，分区的原因
4. 为什么出现数据积压
5. 使用什么使数据不丢失
6. ack 和副本的关系
7. DataX用来做什么的，怎么实现同步，运行模式
8. Hbase 什么时候用到的 ，怎么取得维度数据
9. 指标计算有参与嘛
10. 有设计到java的二次开发
11. 海豚监控器用来做什么的
12. hive小文件哪里遇到过怎么处理
13. 离线 中能够kafka到hadoop 用的程序还是
14. 实时 大屏接口 是做的大屏查询代码还是redis 指标数据

# 中科涟源

1. 自我介绍
2. 说一下实时架构
3. 你们的QPS是多少，每天的增量数据有多少
4. 你们的独立访客是统计一天的吗
5. 工作中是偏开发还是偏运维
6. 公司的服务器是多少台，云的还是物理的
7. 你们部门还有哪些数据部门
8. 你们部门是在公司里面担任数据中台的角色吗
9. Clickhouse使用的是分布式的吗
10. 使用Clickhouse遇到过什么问题吗
11. 你们的结果是输出给哪些部门
12. 你经历过高并发查询的问题吗
13. 你们的可视化是内部使用还是APP用户使用
14. 海豚调度器使用的时候哪里不方便，没有缺点吗
15. 除了java还会什么语言
16. 说一下你知道的窗口函数
17. hive sql的数据如果有问题，我们怎么定位？
18. 你还有什么要问的吗

# 道口金科

1. 做指标的话有没有历史数据更新的指标
2. 业务数据和日志数据是怎么采集到Kafka的
3. canal和Maxwell的底层是什么
4. 写sql的时候怎么优化的
5. 了解sparkhint吗
6. clickhouse的建表语句是怎么设计的
7. 离线数仓的sql的资源是怎么分配的

# 数数科技

1. 实时串讲
2. 为什么用ck，除了ck还有什么
3. 为什么用hbase。
4. 数据从哪儿过来
5. flink怎么保证一致性
6. 如果有同一个用户的业务数据和行为数据存在不同位置多个表，怎么判断他们是同一个人
7. 实时数仓数据量问题
8. Java怎么实现多线程
9. Java怎么保证线程安全
10. 用过啥Java框架 我说springboot
11. 细说springboot

# 银华信息

1. Flink的容错机制
2. Flink是怎么保证流式处理的
3. Flink窗口用的多吗，都用在哪里了
4. FlinkAPI用的多还是FlinkSQL用的多
5. 说一下Flink怎么处理乱序数据的
6. kafka怎么保证不丢数据
7. kafka节点挂了怎么办
8. Kafka怎么消费已经消费的数据
9. kafka数据积压怎么办
10. kafka的数据是存在内存里还是存在磁盘里，传输过程中数据每条有多大
11. 你们数仓分了几层
12. 你们项目是怎么部署的
13. Hive用的多吗
14. Java功底怎么样

# 中电福富

1. 介绍最近的一个项目
2. 需求来源是公司自己的还是用户的
3. 有没有参与过数仓模型，指标，等的分析构建
4. 怎么确保指标正确性，出厂测试
5. hive的数据倾斜，SQL的优化问了一堆，数据去重方式
6. 离职原因，加班问题

# 汝元众通

1. 自我介绍
2. 只问了一个SQL题

部门表： id， 部门名称

人员表： id， name， 部门id

订单表： id， 件数， 单价

求：每个月，各个部门的销售TOP10

1. 考虑SQL性能的优化

# 润通丰华

开始先自我介绍一下

笔试（不能用SQL相关的，可以用sparkcore，sparkstreaming，flinkapi）

题目：

1、数据类型

订单编号|商品分类|商品名称|订单金额

2、数据量：100w/s

要求：

1、统计结果

1）每个商品分类的总销售金额

2）每个商品分类的销售金额TOP10及金额

举例：

0111|手机|苹果13|13200

0112|电脑|华硕A14|5600

结果

1)

手机|300002

电脑|394843

2）

手机|苹果|234324

手机|小米|32423

手机|华为|3242

电脑|戴尔|432432423

电脑|华硕|323434232

# 深流数据

1. 离线介绍。
2. 实时dws层数据是在哪里。
3. 为什么选ck
4. ck 的分叶
5. ck是为啥是单节点。
6. ck怎么备份
7. ck的数据量 明细数据的20%`30%
8. 新老访客用户是怎么做的
9. 旁路缓存和异步IO是怎么做的
10. 为什么选用Hbase。
11. ads层数据是怎么写的
12. 用户画像你做了什么 用户画像平台开发， 标签定义， sql
13. 用户画像架构是什么
14. 你有什么要问的

# 深圳量到

1. ClickHouse中有哪几种引擎？
2. 最常用的引擎？
3. replacingmergetree的原理？
4. Hbase怎么查询？
5. rowkey设计？散列性为什么这样？怎么保证？
6. Hbase、Hadoop规模多少？集群？数据的规模？容量？
7. Hadoop多少台物理机？
8. 用过flinkCDC是吧，还有其他的相对应的技术？
9. FlinkCDC能做到exactly once，保证数据不会丢，能大概讲讲这个事吗？
10. kafka用的哪个版本？
11. binlog增删改查，最终抽到10条，kafka有多个broker，有一台挂掉，不能精确offset的位置，offset可能往前复制一下，这个时候你复制的目标数据，我这个binlog可能会....掉，flink怎么避免这个问题？
12. kafka主题有一个非常重要的概念，队列数，就是这个分区的概念，这个能大概讲讲吗？
13. 为什么需要分区？单分区什么的
14. 分区是越多越好吗？以一个什么样的规律去设计这个分区数呢？
15. flink背压
16. 不见得所有节点都出现反压，某一个步骤出现反压，怎么去定位这个问题？
17. 收到股票的行情每3s出现一次，最多的话是1分钟产生一次，这种行情数据想把它合成分钟级的数据，想合成一个1分钟的宽表，这个时候窗口函数怎么写？
18. 我们要求在分钟上对齐，flink窗口它开闭（边界上）的时候是一个什么规律？ 左闭右开
19. 股票数据要做到左开右闭怎么处理？
20. 窗口划分之后，后边要设及到窗口函数计算逻辑，选出这1分钟以内成交价格的最大、最小，这个时候你会选择用flink的哪一种window去做这件事？比如说rich....或者其他的什么窗口？
21. richFunctionWindow里面有两个方法，是哪2个？
22. 统计某一个窗口和上一个窗口改变的量，比如上一个窗口改变的量100，今天的这个量是80，之间成交量的差异是-20，用什么东西保存窗口之间的状态？
23. 这个状态能在flink物理节点上共享吗？假设需要在集群上共享这个数据，怎么做？假设股票停牌了，需要告知所有节点，怎么做？
24. 广播有什么坏处吗？
25. 平时代码写的多吗？
26. 定义一张表，这个表从csv文件来的，这个表怎么写，比如createTable....后面怎么写？
27. sink的目标存成csv文件，with后边有一大堆参数，你应该考虑哪些参数？（比如分区的参数）（源表和目标表的配置）
28. sql题

学生表，Student(StudentId,StudentName,StudentSex)，StudentSex码值：Male、Female

课程表，Course(CourseId,CourseName,CourseType)，CourseType码值：MAIN-主科、SUB-副科

成绩表，Score(ScoreId,StudentId,CourseId,Score)

1、求各科的平均得分，输出课程名称及平均得分，按平均得分降序排列；

2、求所有女生的副科平均得分，输出课程名称及平均得分，按平均得分降序排列；

3、求主科平均分低于副科平均分的学生姓名；

4、求数学得分最高的各前十名男生、女生姓名及其得分；

# 上海欣国泰信息通信

## 面一：

1. 自我介绍
2. 实时项目介绍
3. flume中的source用的什么，端点续传的信息保存在哪里，如何修改断点续传的信息，数据采到一半不用续传？
4. channel 用的什么，kafkachannel用的那些参数
5. hive建表用的内部还是外部，创建外部表用了那些字段关键字，如何解析数据
6. hive分割符指定的什么
7. 项目印象最深的技术问题，我说Flink 反压 数据倾斜
8. Flink的检查点
9. Flink的窗口函数
10. java编程用的多吗

## 面二：

1. 自我介绍
2. Flink处理的过程，使用的什么算子
3. Kafkasource如何使用
4. Kafka如何保存偏移量
5. Flink如何根据保存的偏移量继续读取
6. SQL如何进行行转列
7. sql使用过哪些系统函数
8. linux系统指令：如何提取关键字
9. 如何查看cpu使用率

# 华讯网络

1. 先自我介绍
2. hr开始问问题 为啥来成都 而现在北京 我看你之前是在杭州工作 问了好多细节 大概五六分钟
3. 面试官开始问 给我介绍你做过的项目 实现什么样的功能 具体展开讲讲
4. 你们的这一套架构是开源的吗 为什么没有用 cdh 后者说其他的集成组件
5. 拉链表用过吗 是个什么过程
6. 你们flink 是用来干什么了
7. 你们集群的搭建你参与了吗、 具体是那些呢

# 天健元达

1. 介绍实时项目
2. yarn流程。
3. hdfs小文件。
4. hdfs怎么切片的。
5. hive数据倾斜
6. spark的map和mappartition。
7. spark的cache和不知道说了个啥的区别。
8. kafka的的生产者的同步和异步API，消费者怎么订阅主题的API。
9. flink并行度。
10. sql题 姓名，科目，成绩,查每个学生成绩最高的科目
11. 反压。
12. 公司负责什么。
13. 有什么想问的

# 惠安金科

1. 二分查找：

public int binary\_search(int[] nums, int target) {

}

1. sql 题目：如下的ODS层数据表结构和数据样例:

cust\_id,event,desc,time\_stamp

123,login,192.168.3.24,1234567890

123,visit,http://xxx.com.cn,1364817288

123,visit,xxxxxxxx,126738278356

123,exit,12:AB:34:21:CD:A5,221618817288

要求将其清洗为cust\_id,event,ip,mac,time\_stamp结构的标准表（假设忽略原始数据的质量问题），标准化的数据：

cust\_id,event,ip,mac,time\_stamp

123,login,192.168.3.24,12:AB:34:21:CD:A5,1234567890

123,visit,192.168.3.24,12:AB:34:21:CD:A5,1364817288

123,visit,192.168.3.24,12:AB:34:21:CD:A5,126738278356

123,exit,192.168.3.24,12:AB:34:21:CD:A5,221618817288

1. spark提交流程
2. sprk出现数据倾斜，或出错误了，怎么解决

# 亚信科技

1. 自我介绍
2. 项目介绍
3. 公司数据量，服务器
4. 写过最难的SQL
5. hive小文件，
6. 会不会linux

# 北京友坤科技

## 面1

1. 自我介绍
2. 介绍写Hadoop
3. Hadoop写流程
4. 2NN的作用
5. 集群拉取数据到本地那个命令
6. Hive分区怎么创建的，如何动态分区
7. Hive中的四个By
8. Hive三个排序，rank， dens\_rank, rownumber
9. shell脚本后台启动任务怎么编写（具体到那个关键字，命令，函数）
10. 自己启动的脚本任务如何为关闭（log日志中有 任务id， 具体用到那个shell函数，关键字）
11. 快排跟归并排序是如何实现的
12. Java设计模式， 单例模式伪代码描述下
13. Java微服务了解吗（面试中问了多次）
14. Hive中的UDF， UDTF函数定义过那些，那些场景中年定义使用的
15. Spark宽依赖了解吗
16. spark中的shuffle具体是什么流程
17. spark部署的是集群还是客户端模式，有什么区别
18. sql题： 班级 姓名 分数

求每个班级的前三名

1. 开窗 rank

2. 没有开窗如何处理

1. 问了linux常用命令
2. Java多线程问题，

## 面2

1. 自我介绍
2. 介绍离线项目
3. hive和普通数据库的区别
4. orderby 和sortby的区别
5. hive数据倾斜
6. spark宽窄依赖
7. spark广播变量
8. hadoop常用命令
9. hadoop读写流程
10. 数组和链表的区别
11. ArrayList扩容
12. hashmap扩容
13. volatile
14. java锁
15. sql：班级 姓名 成绩 求每个班级前三，不用开窗怎么做

相关子查询

# 华腾易达

1. 先自我介绍
2. 项目串讲
3. 你在这个流程中做了那些内容呢 （我就说集群搭建起来以后主要写指标）
4. 用户流出率怎么做得呢
5. MySQL 数据库删除数据的方式有哪些
6. 你们用MySQL 用的多吗

# 惟邦新创

1. 上来先自我介绍 我串讲了五分钟实时
2. 你为什么要来成都 ？
3. 你在你们公司主要做什么 做哪一块
4. 就按照简历挨个问 一个一个问
5. 问完简历 大概半小时 然后问了kafka 选举机制
6. 你有什么想问的

# 长亮科技

（这家比较针对hadoop）

1. 自我介绍
2. 介绍你做过的项目
3. 你们hadoop集群用的开源还是商业
4. Hadoop，zookper用的那个版本
5. 磁盘坏了怎么办
6. hive优化
7. 查看集群健康情况的命令
8. string string buffer stringbuilder的区别

# 联想

## 面1（外包）

1. 项目介绍
2. hive的小文件，小文件已经在分区里面怎么解决？
3. hive去重
4. 介绍一下maxwell,maxwell挂了怎么办？
5. mysql消费kafka的数据 mysql挂了怎么从新消费？

## 面2（易宝软禁-外派联想）

1. hadoop中的小文件问题，如何产生，如何处理
2. hive中很多个分区，分区中很多小文件，如何解决，如何将小文件合并（直接新建一个表，读取原来的表，进行小文件合并，再将合并后的文件导入原来的位置）
3. oracal-->kafka--->mysql, 如果下游Mysql中的数据被误删，如何快速恢复（指定offset消费，重新生成消费者组，从头消费，mysql中去重）
4. 数据库中的数据去重问题
5. 问了rank 与 rownumber的区别

# 赢时胜

1. 介绍项目
2. hadoop的配置文件有哪些？
3. liux常用命令
4. 脚本查看最新文件
5. kafka应答机制
6. 写过Java吗，知道springboot吗？

# 易方达基金会

1. 自我介绍
2. 讲一下项目
3. 事务型事实表和累计型事实表的区别
4. 实时的dim层可以不可以用直接用fliksql从主流里面拿数据，用where条件，然后把对应的字段直接匹配到维度整合后的那个宽表里面？
5. 为什么不做维度整合？
6. 水位线是解决什么问题的？
7. Mapreduce的机制？
8. 数据倾斜在哪个阶段产生？

# 邦睿科技

1. 项目介绍
2. 开窗函数有那些
3. hadoop组件
4. ods ads层多少张表
5. hive数据倾斜
6. hive优化

# 嘉环科技

## 面1

1. 自我介绍
2. 项目中做过哪些工作？
3. 可以接受运维岗吗？
4. hive和普通数据库的比较？
5. hive调优？
6. hive内部表和外部表的区别？一般用什么表？
7. hdfs读写数据流程？
8. Kafka的ack机制？
9. Kafka的高可用？（不懂）
10. Kafka数据保证不重复？
11. Flink和sparkstreaming的区别？
12. 讲讲flink的wartermark？
13. flink是必须依赖hadoop吗？在平时工作中依赖hadoop吗
14. flink遇到数据倾斜吗？
15. 说说hbase的rowkey设计原则以及预分区？
16. 讲讲spark的rdd
17. 讲讲rdd的宽依赖和窄依赖
18. 讲讲shuffle（我说的是yarn shuffle流程）
19. shuffle的优化
20. spark里面的groupbykey和reducebykey有什么区别？哪个更好，为什么？
21. 工作中是否经常与客户交流？
22. lnuix命令经常用吗？shell编程用过吗？
23. awk怎么用的？批量替换用什么命令？假设我们有一个文件，有几千行，要替换其中一个单词，怎么做？（不会）
24. 一个目录下有上万个文件，只保留最近七天的，怎么做？
25. MySQl索引的分类？
26. 有过关系型数据库写过SQL或者优化过的经验吗？27.讲讲java的gc机制？（只说了个垃圾回收）
27. java的重载和重写的区别？

## 面2

1. 项目介绍
2. hive分区和分桶的区别
3. hive的文件格式（建表的存储格式）有哪些？，从压缩格式上看呢有什么区别？
4. 平时有遇到数据倾斜吗？
5. flume三大组件，source有哪些，channel有哪些，有啥区别，sink
6. flume拦截器
7. kafka几大要素（架构）
8. kafka数据安全的机制（消费者角度offset）
9. 怎么发现HDFS小文件，不处理怎么样，怎么处理
10. 公司hdfs上允许多少个文件块存在，达到多少会报小文件过多
11. sparkRDD说一下
12. CDH用的版本
13. linux命令
14. 写sql思路（统计每个部门的大于30的女性个数）
15. hive如何更新数据（创建临时表）
16. shell脚本有写过吗

## 面3

1. 几年经验了,那年毕业的,全日制本科吗
2. 工作中主要职责是什么
3. 最熟悉的组件是什么
4. spark提交流程
5. spark作业提交模式
6. sparkaction算子和转换算子
7. wrdcount调用算子有哪些,步骤
8. kafka的幂等性怎么理解,用在什么场景下
9. 除了幂等还有什么能够保证kafka消费数据一致性
10. FlinkCDC的原理
11. hive的优化
12. 平时用的开发语言是什么,用来多长时间
13. 对一站式集成类项目从上到下规划,比如云大数据,云大数据方案,解决架构
14. 对服务器网络层面有理解吗,服务器底层网络规划有经验吗,现在使用的服务器有哪些型号,网络出口有哪些
15. 操作系统用的那个
16. linux脚本有了解吗
17. 数据导入导出脚本涉及到的命令有哪些,列举一下
18. awk作用
19. 替换文本中指定字符串,查看进程命令
20. 查看当前节点的网络流量
21. 数据库的四种隔离级别
22. mysql数据库索引失效的情况
23. mysql主从复制的机制是什么
24. 后面问了下上家公司情况,离职原因,期望薪资等

## 面4

1. java的接口和抽象类的区别
2. hdfs的端口号
3. hbase的读写
4. hbase的合并
5. hadoop的配置文件，都配置了什么？
6. Hadoop集群都遇到了什么问题，怎么解决？
7. dfs和hdfs的区别
8. sql的join有哪几种，有什么区别
9. spark为什么mr快
10. 开发工具，用没用过Github？

# 杭州火烧云科技有限公司

1. 自我介绍
2. 介绍最近项目
3. kafka(数据乱序,重复)
4. maxwell sqoop flinkcdc...还知道那些?有什么区别?
5. datax衔接maxwell怎么做到数据不丢失
6. 数仓建模
7. 数仓没层存储方式,snappy为什么解压缩快
8. 业务总线矩阵怎么构建的
9. 合理设计rowkey，预分区,预分区怎么做的
10. 配置表那些维度放到redis?怎么知道谁是最近用过的
11. flink反压,具体做过哪些?
12. flink除了窗口,延迟时间,测输出流还有那些解决数据迟到问题
13. 为什么用ck?ck为什么快?
14. hive数据倾斜具体场景,解决?
15. select \* from A表和select\* from B 速度相差10倍 什么原因?
16. 配置表工具类怎么做的?
17. 有什么想问的

# 上海天易数聚

## 面1

1. 项目介绍
2. clickhouse，多节点多副本会造成什么问题
3. 介绍一下拉链表的设计
4. zk 的半数选举
5. 实时数仓 如何保证 双流 join 数据准确
6. hive 三种排序窗口函数的区别
7. spark 广播器、累加器
8. 说一下 hdfs 的 mapreduce
9. 一道 sql 场景题 求出 用户相互收藏

一张表，列的信息：uid,fid(收藏对方的id)

表与表自身关联

A B

A．fid=B.uid and B.fid=A.uid

1. 工作地点可以接受吗？
2. 预期薪资？
3. 有什么要问我的？

## 面2

三张表：部门表、人员表、订单表

部门表：ID，名称

人员表：ID，名称、部门ID（外键，对应部门表ID）

订单表：ID，订单日期、人员ID（外键，对应人员表ID）、产品ID、产品名称、产品单价、产品数量

1、用一个Select统计出每月部门销售额的TOP10（只用一个select，不用开窗，强调了两次）

2、结果列表显示部门ID，部门名称、销售额、排序号

# 拼多多（上海）

1. 介绍一下实时项目
2. 为什么做实时
3. 数据源的格式 我说的json文件和 mysql
4. 数据怎么到kafka
5. flink 有哪些sink 你们用什么
6. 数仓分层原理 （我说的各层分别是什么）
7. 维度层数据变动怎么办
8. Dws层会有好多个数据源（A、B、C、D）下游拉取的时候怎么确保是最新（正确的）的数据
9. 数据量有多少
10. Ads 最终输出到哪里 为什么选Clickhouse
11. 手撕算法题
12. 状态是什么时候开始的
13. 一个状态保存时间已经到了后续还有数据过来怎么处理
14. 窗口有哪些 窗口什么时候关闭
15. 如何在流式场景中找出划水的人

# 北京­纬创-海底捞

1. 自我介绍
2. 介绍离线项目
3. dim和dwd建模过程比较重要比较核心的地方
4. 多少个业务域，业务过程
5. 有多少张维度表，怎么去考虑的
6. 指标怎么去做管理，指标不断更新迭代，怎么去处理
7. 同一个指标，不同人关注点不一样怎么处理
8. 对外展示的是一个什么，针对主要人群是什么
9. 看板上线，指标数据不对是什么原因造成
10. 现在在职还是离职，离职时间及原因
11. 分析什么样的业务
12. 团队多少人，担任什么样的角色
13. 有什么要问的

# 先进数通

## 一面

1. 自我介绍
2. 实时数仓串讲
3. hbase的读流程
4. rowkey设计原则
5. 实时为什么没有用sparkStream
6. flink的端到端一致性
7. flink的乱序怎么处理

## 二面：

1. 自我介绍
2. 画像项目介绍
3. 画像平台怎么操作的，具体流程，模拟计算一个标签
4. 问了下宽表的优点
5. 画像项目hive中的宽表为什么没有做优化
6. 窄表的优缺点
7. 画像分了多少级，都是从hive中提取的吗，有没有在加工计算？
8. 介绍了一下他们目前的项目，目前主做画像，想把一些新的表迁到hudi上计算。

# 深圳高灯计算机科技有限公司

1. 自我介绍
2. 项目介绍
3. 拉链表的实现思路
4. 数仓建表用空间换时间怎么理解
5. hive和spark的应用场景 hive更实用哪些场景，spark更适用哪些场景
6. hadoop的组件有哪些
7. 谈一下你工作中的棘手的问题
8. 实时数仓用的组件，为什么用这些组件
9. 为什么用ck,ck为什么这么快 ，ck的高可用
10. 团队人员、分工、自己的工作协调能力。
11. 谈一下自己的优势、劣势
12. hive调优
13. hive数据倾斜
14. 自己的工作职责
15. 集群规模
16. yarn的工作流程
17. 如何判断数据的产出、效果，大概意思就是这个数据有什么用

# 杭州融象数科

1. 自我介绍
2. 技术能力和优势
3. 从0到1搭建哪些东西，自己承担了哪些
4. 离线数仓是怎么建的
5. 为什么参考阿里数仓分层
6. 数据量多少
7. 数仓是怎么更新的，有多少任务、多少张表
8. hdfs组成架构
9. 如果从0搭建数仓考虑到后期实时数仓会选择哪些组件
10. 这里问到ck是面向哪些程序用的 如果用于查询可以选哪些。
11. mr map 和reduce 的数量怎么设置的
12. flink 常用算子
13. springboot这块了解吗？
14. linux调优
15. 数据质量监控是怎么做的

# 上汽通用（偏离线）

1. hive的架构
2. hql点击查询的时候会发生什么？（我答的是hive解析hql过程）
3. hdfs用哪种存储格式
4. 分区表和分桶表 在hdfs上的区别
5. hive 内部表和外部表啥区别
6. sql优化 我答的是hive优化
7. mapjoin 如何设置
8. sql语句优化
9. hbase的使用，怎么用的（我答的是用于实时项目中存储维度数据）
10. hbase的rowkey是如何生成的
11. hbase热点问题
12. hbaserowkey的格式
13. 介绍一下kafka
14. flink是用来干嘛的？
15. 介绍一下实时窗口
16. 迟到数据如何处理
17. 场景题 流表join 有哪些方法进行一个join
18. hadoop安全权限相关的问题

# 天马微电子

1. 自我介绍
2. 介绍项目
3. 按照你个人想法，说下能不能对实时数仓进行改进？
4. 对数据做过什么处理吗？
5. 问了怎么调度？
6. 问了我们多少人，怎么分工的
7. clickhouse的优缺点？你们使用clickhouse有什么亮点

# 北京中创智维

1. 离线数仓串讲
2. Ods的数据和dwd数据一样吗？Ods的数据是怎么到dwd的？
3. Hive怎么引起的数据倾斜？我说可能是null值引起，怎么看出来他是null值引起的数据倾斜？怎么优化？开启Map-side，要是开启之后还不行怎么办？Hive优化
4. Flink和sparkstreaming的架构对比？
5. Kafka的数据是怎么存储的？Offset的结构？
6. Flink的分布式快照原理？如果有多个队列，Barrer怎么对齐的？
7. 你们的数据是一个数据源吗？是一个MySQL出来的吗？