上海大数据面试题汇总

# 字节跳动

1. 算法题，排列组合

示例：[1,2,3]

期望：[1,2,3],[1,3,2],[2,1,3],[2,3,1],[3,1,2],[3,2,1]

1. kafka如果创建大量的topic，对kafak会有什么影响？
2. kafka一个topic有3个partition，但是出现了数据倾斜，有2个partition中数据量很少，下游用flink消费kakfa，5秒钟一个窗口会有什么影响？
3. 谈谈你对spark Shuffle的理解？ 描述一下大表与大表join时的Shuffle机制或者过程？
4. 谈一下flink如何保证精准一次性？
5. barrier对齐会有什么危害？
6. 你在项目中都遇到过哪些问题？怎么解决的？有没有转化为经验，分享给其他同事，保证这个错误不会再犯？
7. 平时用用java代码多吗？知道线程池吗？你可以手写出一个线程池吗？

# 米哈游

1. 什么是Flink的非barrier对齐，如何实现？
2. flink的内存管理？
3. flink的序列化机制？
4. flink提交job的方式以及参数如何设置？ 页面提交和客户端提交有什么区别？
5. 你们flink集群规模？
6. flink提交作业的流程，以及与yarn是如何交互的？
7. flink的checkpoint机制以及精准一次性消费如何实现？
8. flink的状态是什么，分为几种？
9. SparkContext里面主要做了哪些工作？
10. ConcurrentHashMap的底层实现原理？
11. 什么是Watermark及主要作用？
12. flink是如何管理kakfa的offset，使用什么类型的状态保存offset？

# 华为

1. spark内存管理
2. hive分区表中，单值分区和范围分区的区别
3. 你们公司执行spark任务时，资源怎么设置的（需要直接说出来）
4. 介绍一下kafka水位线（其实就是leo和Hw）
5. 说几个指标，分别从什么数据层拿取了数据，需要直接说出来
6. 数仓采用了什么模型？为什么？
7. hive分区表，单值分区和范围分区的区别
8. spark任务切分，怎么判断有没有执行shuffle
9. 你们公司拉链表都有什么字段，拉链表出错怎么办
10. 列举几张表的同步策略
11. flink Sql 了解吗

# 喜马拉雅

1. Hive中七种join.每种区别
2. Hive的炸裂函数
3. Hive项目中遇到过哪些问题，具体业务中怎么解决的(要讲的很详细)
4. kafka的延迟队列的时间轮算法。
5. 建模的原因
6. 分层的原因
7. 命名表的时候的特殊性(比如有个表用户特别关注，如果是你怎么命名)，还有各种表的同步策略
8. 另外还有两道sql，当她的面写，一道侧写炸开，一道连续
9. 问为什么自己定义udtf而不是直接用炸裂

# 美团到店

1. 自我介绍
2. 具体介绍一下具体做了哪些工作.
3. 为什么要做sparkstreaming到Flink的转化.
4. Sparkstreaming和Flink消耗资源具体数据对比
5. 在什么场景下需要这么高的实时性.
6. 遇到SparkStreaming不太能解决的问题
7. 建模的流程
8. 数仓的分层.
9. 宽表都有哪些,？
10. 三范式知道吗，说一下？
11. 项目中遇到什么难解决的问题？
12. 有小文件和数据倾斜，这个怎么处理？
13. 空值key加随机数是一种数据倾斜解决方案，如果有单个key是热点值呢？又如果有多个key是热点值呢？用参数和代码分别怎么解决？
14. 调度工具用到哪些；
15. 数据可视化怎么做；
16. Flink怎么优化？举实际例子，数据对比
17. OLAP引擎用过哪些？
18. 用过什么工具进行数据迁移，导入导出。
19. 行存和和列存的区别？
20. OLTP和OLAP的区别
21. Flink的JobManger？
22. Flink的TaskManager
23. 为什么选择ElasticSearc，ClickHouse?
24. Spark Streaming和Flink的区别，包括计算实时指标的一个逻辑是怎样的？
25. 假设有些数据，延时了10分钟20分钟才过来，想这种数据在Spark Streaming和Flink分别做怎么处理的？
26. 算从0点累计到当前时间的DAU（日活），像这种数据，用Flink如何实现？
27. 布隆过滤器有什么缺点，哪些场景用不了？
28. 你们离线数仓是跑在什么引擎上的？
29. MapReduce从提交到最后执行大概是一个什么过程？Shuffle和Reduce有什么区别？
30. 一个任务，平常10分钟20分钟就完成了，今天1,2个小时都没完成，我们需要怎么解决？
31. 算过去30天有哪些用户是连续7天登录我们APP的，如何写SQL，思路？
32. 开窗函数有哪些？
33. 开窗函数什么情况下会有order by，什么情况下order by是必须要写的？
34. 数据报表存储这块用过哪些产品，用过哪些存储引擎？--没答上来，后来提醒的我说的HBase
35. OLAP引擎用过哪些？
36. 如何设计数据报表的存储，MySQL已经不能用了，查询效率太低，你们这时候如何存储？
37. 拉链表有什么缺点？拉链表有哪些字段必须要有的？
38. 数据和业务是怎么协作的？比如说数据对业务做一些反馈和支持？

# 七牛云

1. 实时数仓的架构？
2. kakfa中ack级别？为什么0不用
3. kafka中isr的作用？
4. kafka中rebalance？
5. hdfs中的小文件？
6. hdfs的优缺点？它支持高频率的读取数据吗？
7. kakfa的高效读写数据？零复制零拷贝（解释清楚）
8. hdfs的架构？dn、nn、2nn
9. ES的倒排索引？
10. 快排、堆排？
11. GC
12. 环形缓冲区为什么设置成环形的？

# 好未来

1. 如何删除外部表 包括原始数据
2. 建模过程
3. 如何解决用实时系统分析一下 前几个月的数据
4. sqoop 同步策略
5. 拉链表如何实现的-> 缓慢变化维的数据还有其他方式吗
6. 建模过程DWD
7. hive 跟换引擎为Spark 运行的区别
8. 实时架构，spark精准一次性消费
9. 主题如何划分的
10. hive组成

# 360数科

1. 讲一下近期做的一个项目
2. flinkCDC能不能监控mg
3. flink怎么保证精准一次
4. flink挂掉了怎么解决处理
5. 为什么用flink而不sparkstreaming
6. 如果业务上临时加一个临时需要计算的指标，是不是需要重新发布版本
7. hive怎么对JSON解析
8. flink可不可以通过数仓人员写sql去实现业务指标
9. 为什么要做dwm层，订单事实宽表做完你们放在哪里
10. flink数据倾斜怎么处理

# 韵达

1. 用过哪些调度框架，kylin刷入数据到hbase时，用的什么调度工具
2. flink的分区分配策略
3. flink各种窗口的区别
4. 时间语义
5. apply和process区别
6. flume采用什么类型组件
7. flume支持scv格式的数据么
8. 如果不想等窗口关闭才看结果，该怎么做

# 花旗

## 面1

1. Java多线程
2. 多线程的创建方式
3. Java线程池
4. 德鲁伊连接池的特点，如果我连接突然断了，会发生什么
5. Java锁，怎么加锁，用过见过哪些锁，加锁有哪些影响
6. Java数据结构，hashmap 和arraylist
7. spark client cluster
8. spark shuffle
9. spark rdd
10. mr底成和spark stage的区别，mr也是有stage的？mr stage是什么
11. spark提交参数
12. 怎么开启压缩
13. 压缩的效率有多高
14. lzo压缩以后，传输量提高了多少，把具体值说一下
15. hive表优化
16. hive各种参数
17. hive去重
18. spark数据倾斜

## 面2

1. sparkstreaming消费Kafka的相关问题：分区数、并行度怎么设置、用的哪种方式
2. sparkstreaming+Kafka，每周的某几天（如周三）数据量会剧增，你怎么处理？
3. 你说你提升消费者的batchsize，这样会导致消费者处理的数据增大，会不会影响到消费者向broker发送心跳？
4. Kafka消费者分区策略
5. 怎么修改Kafka的分区数
6. 怎么修改spark程序的并行度
7. 读取hdfs文件怎么设置分区数
8. spark job提交流程
9. spark任务切分流程
10. 说一下宽依赖和窄依赖
11. 某个spark程序启动之后跑的特别慢，你怎么定位问题？
12. spark做过哪些优化？
13. flink和sparkstreaming的区别
14. hive和MySQL的区别
15. java hashmap为什么要重写hash和equals
16. 英文自我介绍

## 面3（全英）

1. 自我介绍
2. 聊项目，实时数仓
3. 介绍一下你之前的团队
4. ODS数据存在Kafka里面？不会很大吗，Kafka这种中间件怎么放得下？
5. Maxwell是什么？说一下原理
6. kafka挂了怎么办？丢了重了怎么办？
7. 说说Hive和HBase有什么区别？
8. RDD的理解？stage的个数是多少？宽依赖是什么？
9. FlinkSQL用的多还是FlinkDstream多？
10. spark的作业流程？把提交流程那种图聊了一遍
11. 哪些算子会产生shuffle？为什么这些算子产生shuffle？

# 浦发

## 面1（银行甲方）

1. Spark的版本和对应的新特性
2. Clickhouse和ES对比
3. 介绍项目
4. 小文件问题

## 面2（浦发信用卡）

1. 说项目
2. 我看你有写clickhouse，请介绍一下
3. clickhouse这些功能是怎样实现的
4. 集群的规模，
5. kafka的优化

# 神策数据

1. 广告引流的平台上哪里，你们广告投放到了哪里？说出具体平台的名字
2. 每个投放的的平台有多少收益，为什么选择投放这些平台？
3. 你们的用户行为数据有哪些，具体说一下指标，为什么设置些？
4. 你怎么评判投放的收益标准？
5. 公司为什么设立大数据？你觉得你做了哪些指标给公司带来了什么收益？
6. 你们主要的业务是什么？
7. 你们做比价是去哪里比？那拿到这些数据你们干了什么？
8. 既然比价是点了链接到天猫，京东的平台，你们是怎么拿到用户在其他平台的下单数据的？怎么知道有没有下单？你们和淘宝有合作吗？
9. 你们每天数据量有多少？
10. 你们组内有多少人？为什么安排这么多人？
11. 你们的服务器有多少台，kafka，flume装了几台？hdfs装了几台？
12. 你们总的日活说多少？
13. 你们总的用户是数多少？
14. 你们的商品数数多少？说sku
15. 你为什么离职？

# 贝壳找房

1. 公司是否有做生命周期管理
2. 为什么要做生命周期管理
3. 为什么使用parquet列式存储？为什么不用别的？
4. orc,rc,parquet列式存储有什么区别，底层存储的内存是否是连续的？
5. 为什么orc有索引就一定快？
6. 我答了orc的构成，他随后问到的
7. hive的优化
8. 说提前使用combinehiveinputformat，那么具体是怎么实现的？这个inputformat是什么东西？有几种格式？
9. 你刚刚说开启数据倾斜时负载均衡，那么具体是怎么实现的？不能只说个大概，要说用mr是怎么实现的
10. 什么是维度建模，为什么要维度建模
11. 为什么要维度退化，维度退化有什么好处？
12. kylin的构建算法
13. 拉链表 也问的很细

# 叮咚买菜

1. 简单介绍下框架，手画框架
2. 实时和离线集群是搭建一起还是分开，占比
3. Hive如何实现去重
4. Azkaban版本，有没有二次编译过
5. HashMap
6. 有没有用springboot，编写代码流程
7. 用的框架版本（Hadoop，Saprk）
8. 快排，和归并的区别

# 旺旺集团

## 面1

1. Hadoop用的什么版本
2. hashmap底层知道嘛？
3. Maxwell 时间问题怎么处理
4. 为什么用HBASE存数据
5. 用Phoenix的原因除了SQL
6. 每天日活有多少 订单量多少
7. Redis里面有多少数据量的key
8. Maxwell初始化用过没有
9. Maxwell 遇到过哪些问题

## 面2

### 一面

两道力扣java编程题目

1. Maxwell遇到哪些问题、数仓分层问题，hive版本
2. Spark checkpoint？
3. Flink用什么写的？
4. （4）hashmap和hashtable底层？

### 二面

1. dwd维度建模怎么做的？
2. HiveOnSpark构建了四个会话，资源不够卡住了怎么办？
3. Kafka数据保留几天？
4. 实时数仓如何关联维度表？

# 微盟

## 面1

1. groupby和count（distinct）的底层机制和区别是什么
2. Spark和flink的双流join的底层原理
3. sparkstream统计每天营销额的时候，系统崩溃后，如何处理已经聚合后的数据，数据保存在哪里
4. 各种表怎么导入的，sqoop倒导表的详细步骤，累积型快照事实表，拉链表，现场写代码展示等
5. 数仓里面建的各种表，都建了哪些表，数仓每层之间同事都会有数据进行导入导出和计算，如何保证每层计算间有序状态不干涉
6. 精确一次，至多一次，至少一次对checkpoint有什么影响
7. flink里面异步IO代码具体怎么写的，每一步具体描述出来
8. 都用实时做了哪些任务

## 面2

1. 讲一下kafka中的各个组件？
2. 讲一下kafka中的分区？
3. isr的作用？
4. 数据在kafka中是怎么被处理的？
5. habse的架构？
6. 怎么获取hbase的数据？
7. 怎么设置redis中的过期时间以及hbase中怎么设置？
8. kafka中的ack级别？
9. hbase中wal的作用？以及怎么写数据？
10. 怎么设计rowkey
11. 单例模式（手敲）
12. kafka的leader挂了怎么办？

## 面3

1. 离线数仓分层的意义？
2. 维度建模的过程？
3. 退化维度？什么时候选择退化？什么时候不退化？
4. 拉链表怎么设计的？
5. 数仓里面事实表有哪些类型？以及他们的区别？
6. 累积型快照事实表怎么设计的？
7. udf，udtf有用过吗？主要的使用场景是什么？
8. 开窗之后的row\_number会产生数据倾斜嘛？底层使用的是什么udf？
9. 想要取连续三天的怎么取？
10. hive的优化
11. hive的桶表用过吗？分桶和分区的区别？桶表join什么时候使用？
12. mapreduce的过程？
13. map个数的计算公式是什么？
14. 集群用的组件有哪些？
15. yarn的资源调度策略？几种的区别？你们公司用的是哪个？
16. 大数据权限这一块你有了解过吗？
17. 数据质量这一块有搞嘛？
18. 多维数据库有用过吗？
19. hbase中rowkey的考虑主要是怎么考虑的？
20. 设计rowkey，扫描表的时候，最近的数据一直在最前面？要散列，要有顺序。（高位散列，低位顺序）
21. 一般任务调试的时候怎么调试的？
22. 业务如何保证有序性？
23. 场景三个门店，kafka里面怎么保证每个门店数据不乱？
24. 有什么写过flink sink端的自定义分区？
25. 多个分区去消费？水印一直上不去怎么办？
26. 场景：人在浏览网页，想要他在更换页面的时候触发计算？怎么设计trigger？
27. session窗口有用过吗？
28. 有用到clickhouse嘛？用的什么引擎？带副本嘛？orderby是怎么设计的？
29. 关于clickhouse写过sql嘛？clickhouse你认为优势在哪里？你在写的时候用过clickhouse里面哪些函数？
30. 拉链表的设计？
31. 订单表是怎么设计的？是怎么设计分区的？
32. ads层之后的应用是什么？
33. 有上层需求，既需要当天的也需要历史的数据怎么查数据？

# 海尔集团

1. hive都进行了哪些优化
2. 你们数仓是怎么分层的
3. 如何建模
4. 数据怎么清洗的
5. Shell中有哪些变量
6. 写个shell脚本（启/停 群发的不行）
7. 有没有哪些sql经过你的优化之后效率有大幅的提高，请详述一下内容以及前后性能对比
8. hive的小文件的处理

# 蚂蚁

## 面1

1. Hbase如何读取数据？
2. （其实想问的是怎么设计rowkey获取数据）
3. 离线数仓分层
4. flink的精准一次性消费
5. flink的一个流的数据错了怎么处理？
6. 有哪些业务线？
7. 为什么用spark？

## 面2

1. Hive数据倾斜处理，数仓建模，任务调度，ods到dwd干了什么。
2. Hive数据倾斜怎么处理

## 面3

1. 数仓分几层？
2. 你主要负责哪部分？
3. 上家公司大数据有多少人？
4. 有哪些表，字段有哪些？
5. 缓慢变化维表怎么处理的？
6. 实时和离线数仓的分层有哪些区别？
7. 怎么进行维度建模的？
8. ods到dwd层的累积型快照表怎么实现的？
9. 分析过那些指标？
10. 用没用过dataworkers？
11. Hadoop新版本有了解过吗？
12. 知道哪些大数据的新技术？
13. sqoop导入数据出现订单支付、邮递状态在一天怎么办？
14. 数仓中表加了字段怎么办？

# 趣头条

## 面1

1. 集群规模
2. 如果用多个flinkCDC监控同一个表，可能会出现什么情况？
3. Flink怎么保证一致性
4. 状态存在哪里
5. Checkpoint存在哪里，存的是什么
6. Clickhouse分布式表和本地表有什么区别
7. Rowkey设计原理
8. HBase写流程
9. 刷写时机
10. Java中HashMap和ArrayList得初始大小是多少，怎么扩容的
11. 动态分流中的分流方法是什么？

## 面2

1. 公司业务，集群规模，人数
2. 介绍一下你做得最好的一个项目
3. 什么是barrier对齐，介绍一下
4. spark你做了哪些优化，数据倾斜怎么做
5. checkpoint里面都存了什么东西你有了解吗，说一下
6. clickhouse集群规模多大
7. 列式存储数据库和行式存储数据库有什么区别
8. 介绍一下hbase的写流程
9. groupbykey和reducebykey什么区别，使用场景？
10. jvm了解吗？里面有哪几块
11. new对象的生命周期？垃圾回收器的生命周期？

# 得物

## 面1

1. 数仓数据的导入
2. 遇到过哪些问题
3. 元数据管理
4. 数据质量监控
5. 数据的权限问题
6. Flink的监控

## 面2

### 一面

1. 你用flink做了什么？讲了项目，实时数仓
2. 你用了 AsyncFuntion 讲讲这个异步IO方法
3. Flink的watermark理解
4. Ck理解
5. 你说:Barrier对齐与非对齐都能实现一致性

### 二面

1. 介绍用flink做的项目 大概十分钟等他问
2. 介绍完项目就是业务场景题
3. 我要看每小时的DAU UV 以及当天到此刻的DAU UV 结合你的数仓怎么实现,还有一些指标 比如每隔5分钟的 30分钟的
4. 开窗1小时 每隔十分钟输出一次计算 怎么实现
5. 你的实现 如果半夜 4点没有数据来 会怎么样？
6. 看过哪些源码 是你自己看的那种
7. 除了做profile 做火焰图 还有什么方法 快速定位算子问题

## 面3

1. flink异步io阻塞的问题
2. GMV写入redis访问操作太多的问题。
3. 还有提交参数 配置怎么设置。
4. check point失败的问题
5. 纬度数据慢的问题
6. water mark传递问题
7. 第一次check point什么流程。

## 面4

1. 自我介绍
2. 介绍flink项目（他们这边主要做flink）
3. 问了我做的业务产生了具体有什么收益
4. 组内多少个人，分工怎么样
5. 了不了解flink关于slot的一个sharing
6. 场景题：计算当天登陆的用户数，怎么去做？

# 永辉

## 面1（云创）

1. 介绍一下数仓分层，dwd做了哪些处理，怎么看dws层，为什么要有dws层，有什么优点
2. dws层（指标复用算不算）
3. ads给谁用，怎么用？
4. 介绍一下现在用的报表和数据产品有哪些？
5. 你主要负责哪一块业务？
6. 数据可视化用的是什么？怎么用？
7. 离线跑用什么引擎？spark的话小文件怎么处理？
8. 调度系统用什么调度？是原生的还是经过了二次开发？二次开发解决了什么问题？
9. sql问题：累加、求max，同环比
10. 大数据存储在hdfs的格式是什么？用parquet存储的话用presto有影响吗？
11. kylin为什么快？
12. flume采集日志用什么source，channel用什么？为什么？
13. 日常开发用什么语言？
14. 有没有做数据质量？怎么做？数据质量一致性，可靠性，完整性了解吗？
15. 怎么保证数据采集是准确的，业务数据采集到hdfs层，怎么保证采集的就是对的？关于采集有没有校验？
16. 数仓有哪些主题？
17. 有没有用过hive里的的grouping-sets、cube、rollup？
18. leftjoin时候where和and区别？

## 面2

### 初始（30分钟）

1. 讲了一下我主要工作内容
2. 问数仓建模
3. 问了两个简单的sql top3 7天内3天连续
4. 问了离线的业务

### 复试（30分钟）

1. 还是讲了一下数仓建模
2. 问flume到kafka到flume 端到端的一致性
3. 然后又是问业务

# 樊登读书

1. flume采集的数据是实时的吗。
2. 订单状态如何更新，之前状态是否保留。
3. sparkstreaming和flink的区别。
4. 是否做过像离线一样不论什么指标去clickhouse都能查到
5. 对樊登读书有过了解吗

# 顺丰

1. 如何保证维度数据修改后能和事实数据依然匹配
2. hbase写入数据后同步redis失败如何解决
3. clickhouse数据量多少
4. clickhouse去重失败如何解决
5. flink窗口底层如何实现
6. hbase的hflile文件合并使用哪种算法

# 中通快递

1. hbase写为什么读比读慢？
2. 读到blockcache 为什么不能直接返回？
3. 不是已经有最新数据了吗？
4. flinkcdc 2.0 比 1.0解决了大表初始化的问题，请问是怎么解决？
5. 自定义source怎么做？
6. 自己写一个框架，实现flinkcdc、canal或maxwell的功能？
7. 底层通过java的spi技术
8. flink exactly one 和 at least one区别？
9. 生产中flink资源分配，内存和cpu核数设置，最大的任务占多少资源？
10. checkpoint怎么设置？
11. jvm 垃圾回收算法？哪种容易fullGC（标记清除，因为会产生内存碎片，如果需要很大的连续内存，则就会触发fullGC）

# 极兔速递

1. 为什么采集的时候用kafka消息队列，不直接使用flume直接落盘到HDFS？
2. 现场写了sql（主要是用了开窗函数）
3. hive on spark和spark on hive有什么区别？
4. 离线数仓里面的维度建模是怎么选择的？
5. hive的优化
6. parquet，orc，textfile，针对hive里面怎么选择哪种存储会速度更加快
7. 对mapjoin的理解，mapjoin的原理（为什么能有效的避免数据倾斜问题？）
8. 维度建模的选择？有没有自己设计过模型？
9. 维度模型？
10. 实时数仓里面处理过哪些指标？有哪些比较难得指标？
11. 讲了离线数仓项目

# 德邦物流

1. 介绍一下自己
2. Sqoop导数据时候遇到啥问题了
3. hive优化知道吗？
4. 你们数据量多少？
5. 峰值数据有多少
6. 数据质量监控这一块，你们怎么做的
7. Kafka你们丢过数据嘛？
8. 按照什么方式或者业务给老板报表的
9. SQL会吗？
10. 你们宽表怎么做同步的
11. 函数用得熟练嘛？排序函数知道吗？

# 杭州 奇点云

1. spark任务提交全流程
2. 你怎么学的spark源码?debug有什么技巧?
3. 你做过开源项目吗,参与过那些提交么
4. 你项目里面有哪些你负责的比较难的实现么
5. hbase的分区是什么,需要怎么做做,里面怎么存的,数据是如何落进去的
6. clickhouse是是怎么存数据的知道么
7. acp原则知道么
8. 你知道b+树和lsm树是什么么?怎么放数据的
9. 列存的优势是什么,列底层是怎么转行的
10. 让你写一个多线程的kafka消费者,你有几种写法,分别怎么写的
11. 你看过那些框架的源码
12. 你知道一致性hash么
13. 一致性hash如何拓展知道么
14. 算法题,说一下边界条件
15. 介绍了下他们是做什么的,做到哪了,问我有什么要问的

# 晨光

1. 问公司每天数据量
2. 服务器、物理机有多少台
3. 这些数据你们是怎么用的？
4. 你在工作中遇到让你印象比较深，难解决的业务。

# Soul

1. 实时数仓每一层的数据质量怎么做？
2. flink 在join操作中，ck一直完成不了？什么原因
3. flink 一个subtask发生背压，怎么定位是哪个操作导致的？除了断开操作链？
4. kafka数据一致性？在ack设为1的时候，如何保证？
5. 谈谈JUC的多线程，semephore，countdownlatch，cyclicbarrie？
6. ThreadLocal解决什么问题？原理是什么？

# 多点生活

1. hdfs-sink小文件问题怎么处理的？
2. hdfs小文件怎么处理？
3. 遇到过flume写hdfs瓶颈吗？你是怎么处理的？
4. flune时间拦截器具体怎么实现的？
5. 除了flume还可以用啥？
6. 为什么要用kafka？讲讲kafka如何实现精准一次性，kafka分区之间的精准一次性如何保证？
7. 除了sqoop还可以用啥？
8. 讲讲hive调优，数据倾斜怎么解决？
9. hive怎么解决小文件问题？
10. dws和dwt层有什么区别？ads层主要干了啥？
11. 数仓为什么要分层？说出十个分层的原因
12. 场景题：如果有三个门店，分别卖矿泉水5瓶，可乐5瓶，可乐5瓶，相同的类型不重复计算，求解仓库还有几件货物？请使用自定义udaf函数口述伪代码
13. spark有哪些算子会走shuffle？groupbykey和reducebykey有哪些区别？除了预聚合还有啥区别？底层怎么实现的？
14. 讲讲spark内存模型
15. 讲讲flink分布式快照算法
16. 端到端的一致性如何保证？
17. 假如intervaljoin有数据join不上你怎么处理？口述伪代码

# 润和

## 面1

1. 负责几个topic
2. 每个topic有多少个分区
3. hbase列族有多少个
4. 看你Kafka写的挺多的，kafka装了几台，把你知道的关于kafka的全部说一遍
5. 我把生产者到broker到消费者给他说了一通
6. spark解决数据倾斜的手段
7. spark做了哪些优化
8. 项目中使用redis实现什么功能
9. sqoop导出一致性问题，通过什么参数增加map个数
10. hbase的组件
11. 实际中rowkey怎么设计的
12. HBASE的读写流程，如果数据已经写到了WAL还没写到MemStore挂机了，会怎么处理，有什么影响

## 面2

### HR

1. 开发语言
2. 对java熟练度怎么样？
3. 对scala熟练度？
4. 偏实时还是偏离线？
5. 在实时里面，Flink大概怎么用的？功能划分？架构设计？
6. Flink数据怎么接入的？介绍一下整个数据流？
7. 埋点数据都包括哪些数据？这些数据过来后没有标准化就直接入Flink嘛？
8. 埋点数据后续做了哪些判议？比如Flink后续对这些数据怎么进行处理的？
9. 维度表进来Flink后后续的应用流程和处理场景？
10. 项目大概多少人？
11. 什么角色？主要负责什么？
12. 实时这块做了多长时间？

### 面试官

1. 介绍一下你在实时项目里做了哪些东西？
2. 聊一下flink反压？反压相关的拥簇管理机制，有没有到源码或者对原理进行一个探究？
3. Prometheus、Grafana等这些监控告警系统有用过吗？
4. 指标这块具体实现做过的工作？
5. 对于怎么算出这些指标，算法这边用的都是比较简单的一些还是有引入比较复杂的算法去算的？（对于指标字段有没有一些模型或者一些高端算法）

# 声网

1. 怎么修改正在运行的Flink程序？如果有新的实时指标你们是怎么上线的？
2. 使用flink统计订单表的GMV，如果mysql中的数据出现错误，之后在mysql中做数据的修改操作，那么flink程序如何保证GMV的正确性，你们是如何解决？
3. 开发人员和测试人员如何保证SQL的正确性？假如这条sql就是写错误了，那么用这条SQL统计mysql中的数据，肯定也是无法发现错误，你们是如何解决？
4. 如何区分事实表和维度表？有度量值就一定是事实表吗？什么是描述性/修饰性维度？
5. 用sqoop在00：10分将mysql中的数据导入到HDFS，对于新增及变化的数据，由于存在窗口期，比如每天在00:05分的时候这条数据都修改，那么就会一直无法拉取到这条数据，怎么解决？
6. ETL清洗的规则是不超过1/10000，你们是怎么发现超过万分之一的？

# 新潮传媒集团

1. 熟悉哪些组建
2. rdd.df.ds区别
3. flink与spark区别
4. Java写99乘法表（电脑直接敲）计时
5. 字符串，首字母大写（电脑敲）
6. 手写hql（体现实力）
7. 离线架构分层
8. kafka为什么快
9. hdfs了解多少
10. hbase部分内容
11. flink窗口5种
12. flink时间语意
13. 整体数据走向，分开来说
14. 有没有hive相关的API
15. Java中GC原理
16. Java中无符号状态如何有效处理

# 卓钢链

1. flink怎么提交
2. flink提交有多少jobmanger
3. flink与spark区别
4. flink反压
5. flink监控，如何有效处理数据积压
6. 离线那些很简单

# 鸭嘴兽网络科技有限公司

1. flink窗口
2. spark手动提交offset
3. flink有什么问题
4. flink反压，如何解决
5. 为什么flink替换spark
6. spark优化
7. flink优化
8. 在flink项目中做了什么
9. flink开发哪个窗口用的最多

# 齐数科技

1. canal传输数据这么保证不丢失
2. flink配合redis以及布隆过滤器具体怎么实现大数据量的去重
3. flume你们公司允许丢多少数据，说个范围区间
4. spark shuffle讲一下
5. azkaban任务调度怎么使用
6. canal到kafka到sparkstreaming怎么精准一次消费
7. spark的checkpoint怎么使用的，你之前公司有没有用过
8. sparkstreaming怎么消费kafka
9. 做过数据治理吗？
10. 有这种场景c表依赖于b表，b表依赖于a表，如果a表数据出错，就比如说a表是javaEE的订单表，javaEE修改了里面的字段，比如100万GMV，就可能出现只统计10万GMV，你们怎么处理？
11. 你们哪些表使用拉链表？拉链表建立分区表了吗？怎么建立的？

# 斗象

1. flink checkpoint的实现原理
2. spark checkpoint的实现原理
3. jvm如何调优
4. rdd的作用主要是干什么的
5. hive和hbase的区别
6. flink开窗五分钟过来一亿条数据你是怎么处理的
7. flink开窗5分钟被同一用户连续访问60次，需要把他的访问信息调出来 你是怎么做的
8. spark有1000个分区 他们的数据是怎么交互的
9. spark 有10万条数据 你将这些数据怎么分配到集群的

# 金大师

## 面1

1. java基本数据类型
2. 讲一下hashmap
3. hashmap为什么用到红黑树
4. 链表的时间复杂度
5. 红黑树的时间复杂度
6. ArrayList与linkedList的区别，其中链表是什么链表
7. MySQL的事务隔离级别
8. 数据库的索引，索引结构是什么？
9. ①聊一聊kafka，②zookeeper中存储的kafka中的信息的格式，③ack，④副本个数，⑤ISR，⑥kafka的存储在哪，⑦kafka的读写流程，⑧分区个数
10. HBase的读写流程，如果数据已经写到了WAL还没写到MemStore挂机了，会怎么处理，有什么影响
11. 说一下布隆过滤器怎么实现的，数据结构是什么
12. 业务中HBASE的RowKey怎么设计的
13. atermark处理迟到数据，怎么实现的
14. redis的数据类型，怎么用来去重的，存储的是什么数据
15. 说一下slot，业务中一个TaskManager设置几个slot，连接的kafka的分区数是多少

## 面2

1. hashMap和hashSet底层实现原理
2. stringbuffer和stringbuilder的区别
3. spark如何保证精准一次消费
4. hive的两个表join的工作机制
5. kafka的精准一次消费 幂等性+事务 kafka版本 事务如何实现的
6. 事务的分类
7. 如何实现多线程，线程怎么关闭
8. 你知道什么二叉树
9. 红黑树的结构
10. es如何实现更新数据 可以更新部分属性吗

## 面3

1. MySQL的隔离机制
2. MySQL的事务回滚
3. MySQL的底层索引
4. HBase的分区算法
5. HBase的预分区大小
6. HBase的如何避免分区数据倾斜、ID地区
7. HBase的请求写
8. HBase的meta 表中存储了 Hbase 集群中全部表的所有的 region 信息，在 Hbase 2.x 之后新增了表的状态信息。
9. HBase的预写日志恢复机制
10. JAVA锁
11. == 和 equal
12. Scala偏函数
13. Shell中查看上一个命令是否成功执行

# 海致星图

1. 双流join,left join,左流数据先来，右流一直没来，左流会这么样
2. 左流数据已经输出到sink了，此时右流数据来了，可以join又会这么样
3. flink故障恢复
4. Savapoint了解多少
5. 作业挂掉了，恢复上一个Checkpoint，用什么命令
6. 为什么用yarn-session
7. 说一下状态编程
8. 使用Mapstage，group by id 如何设计
9. 继续上面的Mapstage，id不放在key行不行
10. 数据积压问题
11. Kafka数据很多，内存很少，读取数据都是问题，现在想要写，怎么控制写速率（上面都是flink）
12. Spark哪一块用的多，实时，spark streaming用的是结构流还是什么，后面说到df
13. df与ds区别，课上讲的没够用
14. Task与 partition 有什么关系
15. Stage,宽依赖
16. Kafka一直说
17. 一个topic有3个分区，两个消费者，会怎么样
18. 一个topic有2个分区，三个消费者，会怎么样
19. Kafka怎么处理大量数据（为什么这么快）
20. Hdfs小文件处理，spark处理小文件
21. arraylist在 Scala有什么可以做到同样功能，比较像的
22. Hbase,redis,es选一个，我选hbase,又谈到凤凰，凤凰和hbase这么放一起的，rowkey
23. 最后闲聊，数仓分层

# 赢时胜

## 面1

1. 自我介绍
2. 详细的讲一下实时项目.
3. 为什么要用Flink替代SparkStreaming
4. 你们公司都处理过什么业务.
5. 公司的类型.公司的电商的服务还有吗?地址?大数据分析对外部提供服务是在哪个web或者app
6. 公司都卖些什么?买的最好的商品?
7. 分模块对指标进行分析.
8. 这个项目里面日常都会做哪些.
9. 指标中印象最深的就是什么?
10. 网络波动导致的支付先到了怎么办?
11. process用的种类.
12. 10个int以数组的形式保存,那么保存在什么状态好?VlaueState还是ListState?存在哪个的性能比较好?
13. 广告在没有人点击的(也就是没有数据流的时候)窗口,这个窗口存在吗?有没有对这些窗口进行校验的窗口.
14. 1小时的滚动窗口,一小时处理一次的压力比较大,想让他5分钟处理一次.怎么办?
15. 数仓项目处理的业务?
16. dws都有什么维度和和字段?
17. 维表的数据量扩大十倍会有什么问题?
18. 维度的地区增大(比如上海划分成具体的某些区),也就是改变维表的粒度会出现问题吗?

## 面2

1. yarn 资源队列
2. Hbase region 自动切分为什么不好
3. yarn session 和 pre job 区别
4. hive join 发生数据倾斜如何解决
5. 只让讲了一下采集部分，后面的是提问
6. spark sql 和 hive sql 区别
7. rdd 和 spark sql区别
8. 出了到flink实际场景的题 类似咱的Uv那个指标 ，如何解决乱序数据，后来又加了个
9. 迟到数据怎么办
10. flume 如何分渠道，就是那个mutil方式 ，记得说出拦截器加头信息
11. hashmap

# 宝尊电商

1. spring 生命周期
2. 线程生命周期
3. flink的数据量
4. azkaban几个工程
5. azkaban的任务挂了怎么办的
6. flink双流join数据延迟怎么解决？？？？？
7. hive调优

# 食哼

1. flume的故障转移.使用的是filechenal，新的flume替换掉老的flume.（使用同一个共享组，然后注释掉原flume的source.之后新的flume无缝对接）
2. sparkstreaming数据量级别.开窗1小时.和开窗1分钟系统能不能撑得住.
3. 数据质量的问题.怎么样判断各个阶段数据量是对的.比如采集判断采集的数据是对的.判断的标准是什么.
4. foreach中向kafka中发送数据.解决方案.
5. 懒加载lazy.
6. hdfs的危害.
7. datenode的迁移.
8. 需求6小时内的wc.如果6小时后没有到来则清空之前的数据.
9. ss能承载的数据量和什么有关.
10. spark处理大数据场景怎么办?
11. sparkStreaming中处理线程安全问题.
12. MR的一些优化
13. sparkStreaming处理速度太大怎么办?
14. Hive的并行度由什么决定?
15. sparkStreaming怎么处理延时数据?
16. Kylin的构建cube与什么有关?如果数据量大的话对集群有什么影响吗?

# 上海睿民

1. Hive的任务时会不会有任务的卡顿，无法完成？问数据倾斜问题？
2. 很奇怪的问题就是代码里面有没有发生什么问题造成卡顿的？
3. count distinct / group by 他是想问这个
4. 业务方面说出数据倾斜的场景
5. 有没有什么东西可以让你缩小错误的范围？如何去缩小的？
6. 程序流程图问我有没有看过，内存怎么样

# 软通

## 面1

1. 介绍一下你们公司的数仓分层。
2. 印象最深的指标，或者最难的。
3. 使用的是hive还是sparkSql
4. 常用的算子
5. 常用的行动算子。
6. foreach 和foreachPartitions 的区别
7. reduce 和 reduceBykey的区别
8. reduce 是行动算子还是转换算子
9. 广播变量（结合公司业务）
10. 你们公司的数据量（实时）
11. executor 的内存这么设置
12. 你们公司用了多大内存（他应该是指你的一个executor设置了多大内存）
13. 消费kafka的两种模式
14. 你们用了多少个executor

## 面2

1. flink实时数仓有做什么监控吗？
2. 你们是怎么提交flink任务的
3. Flink的checkpoint和spark的有什么区别
4. flink的kafka连接器和spark的有什么区别
5. flink的内存管理
6. flink的反压机制

## 面3（华为外包 南京）

1. 搭建数仓中遇到过的最难的问题，最难解决的故障
2. Spark调优，hive调优
3. Spark项目数据流程，用了哪些算子
4. Spark底层，数据倾斜
5. Shell单括号和双括号的区别，shell写过哪些脚本，有哪些常用命令
6. 分桶表和分区表的区别，分别什么时候用过
7. 有过hive事物处理吗
8. Hive文件存储形式有哪些
9. 架构可不可以用Flume+kafka或者kafka+Flume
10. 有哪些压缩方式
11. Hive数据倾斜处理

## 面4（通用汽车外包）

1. 说一下flink干了什么事？
2. 你怎么监控mysql数据库的，update数据怎么办？
3. 为什么选择hbase?
4. 为什么使用fs状态后端？
5. 为什么存hdfs上，hdfs挂了怎么办，状态不是都不存在了，为什么选择大数据。
6. link流任务运算在哪？挂了怎么办？提交流程。
7. 做过哪些优化。

## 面5

1. WaterMark源码解析
2. onEvent间歇调用
3. onPeriodicEmit周期调用 （默认间隔200ms）
4. maxTimestamp = Math.max(maxTimestamp, eventTimestamp);因为这个公式，所以WaterMark是单调递增的
5. WaterMark=maxTimestamp - outOfOrdernessMillis - 1
6. WaterMark到底是间歇生成好还是周期生成好？
7. 容灾机制/Flink on Yarn 高可用

# 信也科技

## 面1

1. Flink CEP实现了那些需求？用那些算子具体每一步怎么实现
2. HBase的架构，数据热点怎么解决，介绍下一Phoenix的协处理器机制，
3. flink各个模式，job提交流程
4. 单链表
5. flink遇到哪些问题？怎么解决的
6. 为什么选择flink？
7. flink优化？具体说几个参数优化
8. Java对象创建的几种方式？
9. ES数据库选择得什么引擎？

## 面2

1. canal搭建时候client和server装一起还是分开？
2. canal从MySQL拉过来数据什么格式
3. flink的重启策略

## 面3

1. 实时数仓分层建模描述？
2. ck机制？
3. flink监控问题？
4. 实时数仓表插入新字段？
5. 维度关联时，产生背压没关联上怎么办？
6. Kafka如何保证不重复？
7. Kafka的Ack？

# 中软

## 面1

### 国际外包太平洋

1. flink sql怎么定义的，
2. rdd与dataset的转化，
3. 双流join，数据延迟来了怎么办，
4. 你具体负责哪些业务 ，
5. 你们的业务数据加了一张表，然后再用canal倒入的时候要做什么？
6. java和scala的区别

### 华腾

1. 做了哪些业务
2. hive和mysql的区别

## 面2

1. hive优化
2. Sparkstreaming精准一致性消费
3. Map和mappartition
4. Hbase只能用phoneix建立索引吗？
5. Spark背压机制
6. Spark中使用那些算子？
7. Spark1.0和spark2.0初始化sparkcontext有什么不同？
8. 项目中遇到那些问题？
9. Canal的作用
10. 知道datax吗？

## 面3

### 中软外包平安医疗

1. Hive行列过滤是什么意思
2. 列式存储区别
3. ORC与Parquet有什么区别
4. sqoop导出的事务
5. 导入HDFS中map任务挂了怎么办
6. dws层和dwt的区别（怎么实现的，dwt是否分区）,分别是怎么做的
7. DWS层统计各个主题对象的当天行为，服务于DWT层的主题宽表
8. Dwt不分区
9. 在dwd层这些表是怎么处理的 （新增及变化表：优惠券领用表 用户表 订单表）
10. 在离线中，手动修改了数据
11. 数仓中是怎么做ETL的
12. sqoop导入到hdfs（使用几个map）又没有遇到数据不一致（比如说map失败了）
13. 新增及变化是怎么区分开的
14. 新增表、新增及变化表处理的方式有什么不一样的地方（怎么处理）
15. 数仓中分了哪些主题
16. 如何保证数仓中数据的准确性（hive在处理数仓的过程，好多逻辑，在处理过程中会不会出现误差，是怎么处理的）
17. 数仓怎么去掉错误数据（异常），在Mysql中将数据写错了（原本是1000，人为的改成了1001），错误数据应该怎么处理
18. 使用sqoop导入Hdfs,但是业务库物理删除了一点（删掉了一条不需要的订单），相邻层之间进行监控
19. cannel可以识别delete(删除语句)语句
20. 什么情况下使用拉链表

## 面4

### 太平洋项目外包

1. 自我介绍，项目情况
2. 怎么查看flink任务卡住
3. yarn任务怎么看日志
4. 为什么有时候用hive，有时候用hbase，有一张很大的hbase表想要删除中间某些数据怎么办？
5. hql优化
6. flink反压
7. 双流jion怎么做的，intervaljion什么原理
8. driver配置内存多少
9. 怎么查看本地内存情况
10. 写过的shell脚本

## 面5

1. 离线数仓每一层都干了什么？
2. 怎么区分纬度表和事实表？
3. 维度建模你们用了什么模型？
4. 纬度表中的度量值又可分为哪几种？
5. 怎么处理缓慢变化维，拉链表思路。
6. FlinkCDC工作原理，是怎么监控mysql中的数据的？
7. Flink水位线是什么，怎么设计的？
8. MR和spark的区别。
9. 纬度和纬度表什么关系

# 博彦科技

## 面1

1. orderby 和sort by
2. having的作用
3. hive中插入数据的方式
4. spark和MR的区别
5. 在spark中union会经过shuffle吗？
6. 三范式的第三个

### 平安健康险外包

1. 介绍项目，介绍到flume就被打断
2. 离线问做过哪些优化，hive做了什么优化
3. 业务问每天有多少数据，多少日活。问根据业务怎么去算指标的。
4. 使用过哪些些窗口函数。
5. 实时，问为什么使用HBase，为什么使用redis，pv、uv是怎么算的，数据算完存在哪里

## 面2

### 上海交通银行外包

1. 分区表和分桶表的区别？
2. 维度建模理论和普通建模理论的区别？为什么选了维度建模？
3. 星型模型和雪花模型的区别？
4. 拉链表的讲述
5. 为什么选用sqoop，用了sqoop做了哪些事情？

## 面3

1. 自我介绍，项目职责，为什么学大数据，怎么转行的
2. kafka、hbase介绍一下
3. spark和flink的区别
4. 写过最难的sql
5. 从0-1你能负责哪些
6. 数据没有时间戳怎么办
7. 离线发现前几天MySQL业务数据出问题了怎么办

## 面4

（外派到外企paypal的外包公司）

一面：视频面试，2个面试官

1. 开始让你用英语介绍一下自己的情况，学历，工作经验啥的，自己介绍完了之后，一个女的用英语和你交流
2. flink是直接用还是经过二次开发的，二次开发的话配合什么使用
3. 数据在mysql里面增删改了，怎么把数据拿出来
4. 为什么选flinkcdc
5. flink中的窗口介绍一下
6. 详细讲讲会话窗口，在什么业务场景下面用到了会话窗口？

## 面5

1. 简单描述一下MapReduce
2. 为什么Spark比mr快
3. Spark提交过什么任务吗，有哪些参数
4. Spark提交任务中的那个partition参数，怎么理解的？如果1g的任务200个分区，每个文件多大
5. Spark sql的几种join，比如hash join了解吗
6. Spark sql的调优
7. Spark core和Spark sql的区别
8. Spark sql任务跑不动了怎么解决
9. Spark sql读取元数据信息什么的
10. Spark的宽窄依赖
11. map和mapPartitions区别
12. 当Spark涉及到数据库的操作时，怎么优化
13. Spark Streaming数据不丢
14. String、 StringBuilder 、StringBuffer的区别

# 驰骛信息科技有限公司

## 面1

1. 用户行为数据有那些字段
2. 如何指定hql输出的内容到一个文件
3. 如何判断计算是否正确？
4. 维度建模

## 面2

1. hive中udf函数中的方法？
2. a表往b表中写数据，如何避免小文件？
3. 如何做数据的校验在hive中？
4. 还有就是业务的问题？有哪些表？表的数据量？那些维度表事实表？怎么的同步策略？业务表和事实表是否有交叉？

## 面3

1. hive group by 数据倾斜优化
2. hive组成结构，执行时进程名是什么？
3. spark join的实现方式
4. spark executor执行流程
5. hadoop 讲一下优化
6. concurrent hashmap 高效率的原因
7. kafka 追加写为什么效率就高于随机写呢？

# 精锐教育集团

## 面1

1. 介绍一下经历，离线数仓中分层怎么分
2. kafka怎么消费不丢失数据？
3. spark优化做过哪些？
4. spark处理任务来不及怎么办？
5. 公司的数据量大小？
6. spark的双流join和flink的双流join
7. 怎么使用布隆过滤器的？
8. kylin怎么用的？
9. ES的写入过快问题怎么办？
10. 场景题：实时场景的是什么率怎么做？分子分组都开做
11. 分组topn问题，手写sql
12. HBase协处理器了解吗？

## 面2

1. 自我介绍
2. mysql索引的底层用的什么实现的 b+树
3. mysql索引的最左匹配原则是什么
4. zookeeper选举机制，选举过程中某一台挂了会怎么选举
5. 你知道哪些二叉树 讲讲红黑树结构
6. 当场出了一道sql 很简单的
7. kafka分区 副本数
8. 你有什么想问的吗

# 奇利匙

1. 数仓从采集到数仓，每层干了什么？
2. 可视化对接的哪里？是hive吗？
3. 你们dws层详细说一下他具体分析每天的什么东西？
4. kylin用过吗？说一下 （cube的聚合，剪枝）
5. flume优化做了哪些，kafka优化做了哪些，都是你做的吗？
6. 你们dwd维度建模工作是怎么分配的？按照业务线去分配

# 普洛斯

1. orc与parquet的区别？
2. 查询性能方面orc要高。
3. kakfa中的一条业务数据流？（从数据开始到结束）
4. window与processWindow中的process什么时候触发？

# 阳光午餐

1. 你们公司的是自营店还是平台
2. 你们公司的app有网页版本的吗
3. spark streaming 你会那些
4. 说一下kafka你会那些东西
5. watermark说一下
6. java自带的连接池是什么
7. 多线程说一下
8. 反射说一下。
9. 双流join说一下，对比flink的双流join
10. 你使用scala还是java多一些

# 深圳市领星网络科技有限公司

## 笔试

### Redis

1. Redis工作原理，使用场景是什么？
2. 你在项目中Redis的存储有哪些？
3. Redis支持的最大数据量是多少？Redis集群下怎么从某一台集群查key-value
4. 列举一个常用的Redis客户端的并发模型
5. Redis，传统数据库，hbase，hive每个之间的区别
6. Redis的性能瓶颈在哪里
7. Redis支持的数据格式
8. 如何使用Redis高并发可以支持10万Qps+

### Spark

1. sparksql介绍下（rdd dataframe）
2. udf和udaf都写过哪些
3. 介绍下udaf
4. spark提交流程
5. spark调优思路
6. 宽窄依赖是什么？区别是什么？
7. spark on yarn 和MapReduce中yarn有什么区别
8. spark支持的分布式部署方式

### hbase

1. hbase最主要的特点是什么？
2. 简单描述hbase的rowkey的设计原则
3. 请描述hbase中scan和get的功能以及实现的异同
4. 请描述如何处理hbase中region太多和region太大带来的冲突
5. hbase的rowkey怎么创建比较好？列族怎么创建比较好
6. hbase过滤器实现原则
7. hbase宕机如何处理
8. hbase怎么预分区
9. 请减速hbase中compact用途是什么，什么时候触发，分哪两种compact，有何区别，有哪些相关配置参数
10. 关系型数据库是怎么把数据导出到hbase里的
11. 你们用hbase存储什么数据
12. hbase如何实现模糊查询

### kafka

1. kafka中zk起到什么作用，可以不用zookeeper么
2. kafka是如何做到高吞吐量的，请分别从读写两个方面介绍一些
3. kafka的消息持久性是如何实现
4. kafka数据是如何存储
5. kafka消息数据一致性是如何保障的
6. kafka的message格式是什么样的
7. kafka中消费者组是什么概念
8. kafka中的消息是否丢失和重复消费

### flink

1. 怎么提交实时任务，有多少job manager
2. 怎么做压力测试和监控
3. 为什么用flink
4. checkpoint存在哪里
5. 如果下级存储不支持事务，flink怎么保证exctly-once
6. 说一下flink状态机制
7. flink中的Windows出现了数据倾斜，你有什么解决办法
8. flink在使用聚合函数groupby distinct keyby 等函数是出现数据热点该如何解决

# 明略科技

1. hive静态分区和动态分区区别
2. 怎么建表的？命名规则？
3. 哪些用hiveSQL、哪些用sparkSQL？
4. 小表join大表怎么实现？优化？
5. Spark DataFrame的复用？？
6. Spark持久化策略？
7. 内部表和外部表？关键字区别？
8. Java的常用集合？哪些list？
9. Java常用设计模式？
10. 数据结构？哪些map？
11. hashMap?

# 平安银行（中国平安）

## 面1

1. es读写
2. Hbase读写
3. Flink并行度设置
4. Redis悲观锁和乐观锁,又问你知道mysql的悲观和乐观锁不
5. 熟不熟悉Flink算子，像Flatmap,Map怎么用

## 面1

1. 简单的自我介绍
2. Hive sql的优化有哪些？
3. Hive的调优参数有哪些？
4. 公司多少张事实表，多少张维度表？每天多少数据量？
5. 现场写了sql(字符串中取数字求排序)
6. 维度建模是怎么选择的？写一个事实表的星型模型
7. 离线项目采用什么进行调度的
8. 简单说下实时项目
9. Flink中双流join的实现，以及还有哪些join？
10. 怎么将维度表数据关联到事实表
11. 离线项目和实时项目之间是什么关系？
12. 问了一些关于上家公司的情况，比如说CEO是谁

# 天阳科技

## 面1

1. 你们kafka数据量怎么这么大
2. 你们的kafka设多少topic？

## 面2

1. 自我介绍
2. 介绍了下实时项目
3. Flink中遇到的问题
4. Flink 的内存划分，问道参数是谁设置的，怎么设置的；
5. Flink 和sparkstreaming 的一些区别
6. Ck的机制

## 车轮互联

1. hdfs调优你都做过哪些
2. yarn调优
3. 求每小时的活跃用户，BI报表已经生成的情况下延迟数据怎么处理

# 上海小砖块网络科技有限公司

## 面1

1. HBase中rowkey怎么找数据的，连接zookeeper后怎么找到元数据表
2. 二级索引除了phoenix外还有什么方法创建
3. spark报错序列化类找不到为什么
4. spark的shuffle
5. rowkey分几段，用个实际场景说明，如何查一个用户一个月的，一个月所以用户的
6. 用的什么调度框架，里面的sql语句如何触发
7. hbase中存储的数据量多大
8. 自己员工用app有优惠吗
9. 大数据有多少人，怎么分配，整个大数据组多少人，你们的gmv多少
10. 为什么离职，你认为怎样才算更好的发展

## 面2

1. spark的shuffle流程？
2. hbase的rowkey设计？
3. 二级索引创建方法有哪些？
4. 整个数仓分层脚本如何调度执行的？
5. 脚本挂了怎么办？
6. 数仓分层之间数据处理通过什么技术实现？
7. 元数据管理是用的什么？
8. Hbase中get获取数据时是怎么个流程？

# 上海伯俊软件科技

## 面1

1. Flink实时用了Clickhouse ,说说它的优缺点
2. 对Sql熟吗，用没用过窗口函数over
3. Flink实时你们是在哪里分析的
4. SparkStreaming项目，你在里面负责那些工作
5. Flink实时项目，你负责那些工作

## 面2

1. 开窗函数
2. sparkstreaming的乱序处理

# 吉贝克

1. 你做了哪些项目
2. 你主要负责哪个项目
3. 你怎么建模的，你们的数仓架构
4. 你们用的hql，还是sparksql？
5. hive中主要用到了哪些函数
6. 你们数仓遇到的问题
7. 你们用的脚本还是jar
8. 你们azkaban的版本
9. 你们遇到过hive与mysql间的字符集乱码问题吗？怎么解决的
10. 你们hive数据倾斜遇到过吗？怎么解决的
11. 你在离线数仓中做了什么

# 安能物流

1. 什么情况下下用到的hbase，hbase存储的数据格式，有多少列族，每个列族有字段，rowkey的设计
2. sparkstreaming和flink相比实在的哪些好
3. 对flink使用java开发的看法
4. 以后的规划

# 视若飞

1. kafka结构
2. hive常用函数
3. redis穿透

# 维信金科

## 面1

1. 分层如何实现
2. Kafka的Ac
3. Kafka的数据重了、挂了、积压了、丢了
4. Spark的双流join
5. Zookeeper的选举机制

## 面2

1. 拉链表
2. 数据仓库的两种流派理论
3. hive的优化
4. parqut和orc的区别
5. parqut存储在log日志文件的反映
6. select \* from a join b on a.id＝b.id where num >2021 执行顺序

# 瑛太莱

1. 数仓数据的导入
2. 分组topN口述逻辑
3. Kafka的数据重了、挂了、积压了、丢了
4. Kafka的Ack
5. Spark数据倾斜问题

# 叠纸游戏

1. Kafka的ack，为什么为0会实时性高，什么情况用
2. ads层多少指标，在mysql里的结构什么样，写出来，是一个指标一张表吗
3. 有一张表有用户id，后面是登录日期到时分秒，求每个用户最大连续登录月是多少
4. redis什么情况用，里面的存储结构，穿透，雪崩，击穿，
5. 手写二分法查找，返回索引
6. 手写冒泡排序
7. ads层这么多指标如何规划的

# 序章科技

1. 第一面是技术
2. 问spark 和flink区别 细一点
3. kafka 23件事
4. hive下 MapReduce原理
5. 数据倾斜原因 场景 怎么解决
6. 二面Ceo
7. 问你spark和flink区别
8. 问你flink底层怎么运行 不会
9. 讲了spark底层怎么走
10. 三面 一道算法 平衡二叉树

# 数禾科技

## 面1

1. 实时数仓整体流程？
2. Flink、Storm、Spark的背压?
3. Flink有哪些窗口？
4. 流之间的关联，迟到数据怎么解决？
5. flink监控？
6. 状态后端有哪些？
7. 时间语义有哪些？
8. 状态一致性？

## 面2

1. HBase的框架及原理
2. Spark和flink的区别
3. flink的watermark
4. flink的保持exactly once语义的原理（要说到两阶段提交才行）
5. 数据倾斜怎么做的？
6. flink某个分区长时间没有更新，watermark怎么办？
7. 你们数据量不大、为什么要用Redis做旁路缓存？为什么不直接存储到Redis里面
8. HBase的rowkey设计
9. 两道现场编程题，只要说出思路，写出伪代码就行
10. 注册的时候，填写手机号了（埋点监控形成一条流）、点击注册了形成一条流。需要得到填写手机号了但是没有注册的用户（为了给这部分用户推销），时间为15min内没有注册，如何找出这部分数据（flink实时显示）
11. 一个流，三个字段：推荐人、被推荐人、时间。需要求出推荐人7天内推荐的人的数量（两种思路实现）
12. 还有就是：你们日活多少，多少人干这个，怎么分工的，你们维度表数据量多大，你负责什么模块。

# 爱回收

## 面1

1. 实时数仓整体流程？
2. 流的关联不上问题？
3. 状态一致性的保存问题？
4. 离线数仓如何建模？

## 面2

一轮面试

1. 自我介绍
2. 手写sql

--原表

brand mark ts

A 1 1616677053

A 1 1616677054

A 0 1616677055

A 0 1616677056

A 0 1616677057

A 1 1616677058

A 1 1616677059

A 1 1616677060

A 0 1616677061

B 0 1616677062

......

--结果表

brand mark ts rk

A 1 1616677053 1

A 1 1616677054 2

A 0 1616677055 1

A 0 1616677056 2

A 0 1616677057 3

A 1 1616677058 1

A 1 1616677059 2

A 1 1616677060 3

A 0 1616677061 1

B 0 1616677062 1

......

1. Hive优化与导致数据倾斜用到了哪些算子？
2. 开放的题目，写一条hql发现很久不出结果，会如何发现问题？
3. 基本上你会手写hql，就让你等第二面，技术大佬面
4. 数仓建模？分层？
5. 建模参与了写sql，介绍了用户拉链表的思路，问一个月之前的一个用户信息拉错了怎么办？一层一层拆！怎么拆？
6. Flink数据不丢失的三重保障
7. 如何用flink更新一个用户活跃的时间？状态编程，按照userid分组，process算子

# 格罗夫

1. 你们公司ctr是多少？
2. 你们数仓组合维度分析是怎么做的？
3. 谈谈FlinkCDC？

# 晏鼠股份

1. 数据倾斜多少数据量会发生
2. 平时做活动一天有多少数据
3. spark搭了几台

# 法本（麦当劳外包）

1. 基本技能点都会问
2. hadoop shuffle yarn
3. 卡夫卡 结构组成，ack 副本选举
4. hive优化
5. sqoop优点
6. hbase rowkey设计
7. mysql主从复制
8. spark举例算子，任务提交流程
9. flink 状态，cep 端对端一致性 水印

# 怪兽充电

1. 找地点

高德和百度地图 对一个地点有同一个标识 例如是 pid pname jid经度 wid纬度，只是pid和pname各个地图公司表达的不同，纬度和经度 2个点误差不大，你出一个方案将误差范围不大的地点全部找出来？

1. 思路题

100瓶水，某一瓶有剧毒，喝了之后，7天后必死，你如何用最少得老鼠实验出是那瓶有毒药？

# 滔博

## 面1

1. 介绍自己
2. 数据量多少
3. 客单价有点低？
4. 离线数仓如何分层搭建的，层与层分别干了哪些事？
5. 我问的问题：
6. 滔博在各大平台有店铺，里面的很多数据要怎么拿到？
7. 用到了哪些技术栈?
8. Flink数仓如何分层？
9. Clickhouse有哪些引擎，你们用的哪种引擎？

## 面2

1. flume数据重复了怎么处理？离线怎么处理？又问实时怎么处理？
2. flinkcdc监控配置表怎么保证每次数据都能全部加载？
3. flinkcep底层了解吗？用户跳出的时候如果后面永远没有数据来了你这条数据怎么获取到？
4. redis+布隆过滤器优化了uv的统计，但是造成频繁读取redis怎么优化？
5. 布隆过滤器原理，用法，用什么实现的？
6. 这个uv的缓存你设置了多久的过期时间？
7. redis缓存数据怎么保证不丢失？程序挂掉之后你怎么保证精准一次呢？
8. redis旁路缓存导致hbase里数据不一致怎么处理？
9. 实时数仓对于新增和更新的数据你怎么处理？
10. 讲一下hbase表设计？为什么要做预分区？预分区的好处？为什么不要做切片？
11. 实时数仓dim层为什么用hbase？为什么不用MySQL？这样你连同步分流都不需要做了？为什么不直接使用java那边MySQL主从的从节点呢？为什么不直接自己搭一个MySQL？
12. hbase和MySQL的区别？
13. flinkcdc监控binlog怎么保证写入Kafka的顺序性？
14. spark提交流程和任务划分
15. df，ds，rdd的区别？dataset底层做了那些优化？sparksql里一般用哪一种？
16. hive常用的优化方式
17. 你用什么格式的列式存储？为什么查询的时候只查需要的字段效率高？parquet支持什么级别的谓词下推？
18. 什么是谓词下推？谓词下推发生在什么时候（哪一步数据减少）？为什么先执行where数据就会减少？你怎么看到谓词下推生效了？
19. 讲一下维度建模，拉链表、三种事实表，讲一下维度退化
20. 离线数仓的数据怎么管理，新的需求新的数据源你怎么处理？
21. 对Java并发编程熟吗？
22. 对sprintboot很熟吗？
23. flink中一条数据更新了，进到了两个窗口，怎么保证最终聚合结果的正确性

# 百胜软件

1. 数仓建模你自己做的？数仓建模的过程？
2. 设置了一个场景，用Java去消费kafka主题时，java代码挂了，重启了java程序，数据重复了怎么办？
3. 大部分的时间相互说自己的业务

# 波克城市

## 面1

1. hashmap和hashtable哪个线程安全？
2. 手写二叉树
3. 手写快排
4. Jvm，Juc相关问题
5. Redis雪崩
6. Flink在生产上出过哪些问题？

## 面2

1. sqoop如何加快速度
2. 数仓分层每层的作用？
3. clickhouse用的多吗？
4. mysql大量的数据如何提高查询速度

# XTranfer

## 一面

1. 维度表？事实表？
2. 数仓建模过程？
3. 数仓分层？分了哪些层，每一层做了哪些事情？
4. 手写hql，五道题，比上课的题简单！

# 通联数据

1. 说说hbase。
2. 用过那些nosql数据库。
3. 安装kafka遇到的问题。
4. hashtable和hashmap的区别。
5. 什么时候用hashmap，什么时候用treemap

# 美团外包

1. 主攻哪个方向
2. 我说的实时
3. 具体介绍一下具体做了哪些工作。
4. 为什么要做sparkstreaming到Flink的转化。
5. 在什么场景下需要这么高的实时性。
6. 既然是开窗为什么一定要转FLink。
7. 遇到SparkStreaming不太能解决的问题。
8. 必需要手动维护offset吗？
9. 遇到Flink不太能解决的问题。
10. 实时指标出来后的应用场景。
11. 预警是怎么做到的.预警的条件。
12. 当初建模的时候应用场景是什么样的。
13. 建模的流程
14. 都有哪些数据同步到数仓里面
15. 对这些表有过什么分类吗。
16. 哪些表是相应的同步策略。
17. 跨天支付数据是怎么处理的。
18. 用户表为什么一定是拉链表。
19. 数仓的分层。
20. 如何找出来用户的一天的行为轨迹。
21. dws和dwt的宽表都有哪些,并且都是什么!!!
22. 出口对应的指标。
23. 你们是怎么保证数据质量的。
24. 数据质量监控的角度
25. SQL题（很简单）

外卖的配送ID

外卖员的ID

订单配送的City

时间的CT

一整年中，每个月每个城市订单量Top10。

1. 为什么考虑换一份工作。
2. 离线和实时更偏向哪些，为什么？
3. 工作后做的最有成就感的一件事是什么。
4. flink，ck机制，内存管理，出现反压怎么处理的？
5. kylin如何直接构建cube？

# 中科软外包，安盛保险

1. clickhouse数据量小为什么还用
2. 实时数仓延时多久
3. 数仓实时数仓每层多少数据
4. kylin查询需要多久
5. dwd层表多少列
6. 怎样查询更快
7. 导出到SQL多久
8. 用分布式快速计算

# 忆锦

1. flink中如何保证数据的正确性
2. 整个实时项目flink做了什么
3. clickhouse的引擎及区别
4. atlas如何关联azkaban获取元数据信息
5. 项目中如何分工

# 卫瓴

1. flume的日志数据是从日志文件采集
2. kafka宕机怎么办
3. 手敲代码
4. 组长一般做啥
5. 如何分工

# 辰龙科技

1. 搭环境，集群中高可用哪里搭了，
2. shell脚本分割文件，30个g，从clickhouse，mysql装不下30个g。
3. shell检查某个文件是否存在？
4. 单引号和双引号，区别？联合使用。
5. 变量，``返回执行命令结果 还有个写法是什么？ $()
6. 给一个文件，用shell统计数量，怎么做？回答：用awk。 不用awk呢？ tr用过吗？replace用过吗？sort用过吗？
7. python会吗？ 他们用python连hiveserver2。
8. java 垃圾回收机制了解吗？
9. Object有Equal方法和Hashcode方法。自定义一个类，重写equals方法的时候为什么要重写hashcode方法。Object中hashcode方法是干什么用的？
10. Flink依赖zk吗？ kafka依赖zk吗？ 三台分别有什么角色？ znode有四种类型，哪四种？
11. zk的客户端连接znode，有什么命令？
12. kafka 一个topic有很多分区，生产者如何确定将消息发往哪个分区下？
13. 如果kafka一个消费者挂了，启用消费者组内的其他消费者，这个过程是怎样的？
14. 如果一共有480个分区，有240个消费者，其中消费者初始化速度不一致，很可能一个消费者消费了过多的分区，（可能一个消费者分配了480个分区），这种情况怎么办？如何限制消费者的消费分区数？
15. 分区topic迁移的问题，具体命令是什么？
16. 命令行lag用过吗？看topic消费是否及时。
17. kafka 消息单条数据大小限制，可以通过配置修改。
18. hadoop dfs之类的操作
19. 链表的原理？
20. 布隆过滤器，怎么做的？
21. kafka monitor？
22. scala 接口初始化的顺序？
23. hdfs 删除文件的过程？ namenode datenode 详细是怎么操作的？

# 紫川

（平安普惠外包）

1. 离线问了hive的优化，
2. 实时问遇到过什么比较印象深刻的问题，clickhouse了解多少

# 筹远

1. 框架中使用了哪些数据处理的方式
2. hive是使用单分区还是多分区
3. clickHouse是用来干嘛的，给到多少资源
4. redis给了多少资源
5. flume、kafka如果挂了怎么办，如果需要保证不丢数据怎么办

# 博奥特

## DB

1. 写一段sql,删除表的重复记录
2. delete from tablea 和 truncate table tablea的区别
3. 选一种您熟悉的数据库，谈谈有哪几种索引类型？
4. 根据提供的表结构信息，优化以下3个SQL

基础表信息：

|  |  |  |  |
| --- | --- | --- | --- |
| 表名：AGENT | | 代理人表 | |
| 字段名 | 字段类型 | 字段含义 | 主键、外键 |
| AGENT\_NO | NUMBER(8) | 代理人编号 | PK |
| AGENT\_NAME | VARCHAR2(50) | 代理人姓名 |  |
| AGENT\_LEVEL | VARCHAR2(9) | 代理人级别 |  |
| AGENT\_SALARY | NUMBER(7,2) | 代理人工资 |  |
| BRANCH\_NO | NUMBER(4) | 代理人所属分公司号 | FK |

|  |  |  |  |
| --- | --- | --- | --- |
| 表名：BRANCH | | 分公司信息表 | |
| 字段名 | 字段类型 | 字段含义 | 主键、外键 |
| BRANCH\_NO | NUMBER(4) | 分公司号 | PK |
| BRANCH\_NAME | VARCHAR2(50) | 分公司名称 |  |
| BRANCH\_ADDRESS | VARCHAR2(500) | 分公司地址 |  |

SQL语句1：

SELECT DISTINCT B.BRANCH\_NO,B.BRANCH\_NAME

FROM BRANCH B,AGENT A

WHERE B.BRANCH\_NO = A.BRANCH\_NO;

SQL语句2：

SELECT AGENT\_LEVEL , AVG(AGENT\_SALARY)

FROM AGENT

GROUP BY AGENT\_LEVEL

HAVING AGENT\_LEVEL = 'LV3' OR AGENT\_LEVEL = 'LV4';

SQL语句3：

SELECT \*

FROM BRANCH

WHERE BRANCH\_NO IS NOT NULL;

## 问题清单

1. ETL
   1. 是否在项目中使用过Informatica？
   2. 如果(1)结果为否，请列举项目中使用的其他ETL工具
   3. 是否有处理拉链表经验？
   4. 是否有ETL工具调优经验？
2. SQL
   1. 是否在项目中使用过Oracle编写存储过程？
   2. 如果上一节结果为否，请列举项目中编写存储过程使用的其他数据库。
   3. 是否有分析数据库执行计划能力？
   4. 是否有亿级记录查询处理经验？
3. Linux
   1. 是否有独立编写shell脚本能力？
   2. 是否知道Linux文件目录权限管理机制？
4. 沟通和管理
   1. 请列举在以往项目中担任过的角色？
   2. 是否带领过团队？团队人数为多少？
   3. 请列举一个项目中遇到的困难及解决方

# 新致

1. 内部表外部表区别
2. Union和Union all区别
3. Linux查看进程方式
4. in和not in
5. 假如有两张表A和B，它两有相同的一个字段id，怎么去除B表的id而保留A表的id

# 金丘科技

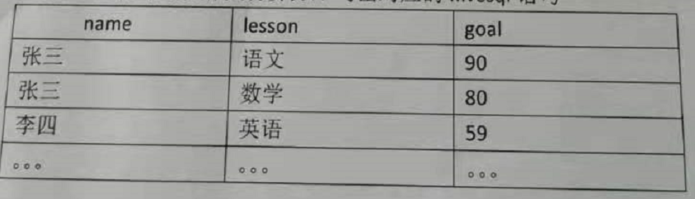
1. Kafka数据丢失问题
2. namenode工作机制
3. 服务器部署台数
4. 框架在服务器上的部署分布

# 上海子悦网络科技有限公司

1. 详细描述一下MapReduce的shuffle过程
2. 针对以下spark-sql语句画出对应的RDD的DAG图和划分的Stage

Select count(1) from table\_a a left join table\_b a on a.id=b.id;

1. Spark任务提交yarn-cluster模式跟yarn-client模式的区别
2. 以往工作中HQL的优化
3. 根据下面给到的成绩表数据写出对应的HQL语句



* 1. 找出语数英每门课前三名的学生
  2. 单科分数有低于80分的学生的总分排名

# 杭州安恒信息技术股份有限公司

1. flink spark对比
2. 数据量
3. flink容错机制
4. clickhouse常用函数去
5. 离职原因（问的挺深的，加班情况，公司在哪，住在哪
6. 懂不懂docker
7. Java堆栈使用情况
8. flink资源消耗情况

# 琚宠科技

## 面1

1. kafka数据积压，增加topic，增加batchsize还倾斜怎么办
2. kafka数据倾斜
3. flink前面的任务挂掉，怎么不计算后面的任务
4. flink容错性有哪些
5. 数据质量怎么管理的

## 面2

1. 你们公司数据质量怎么监控的？中间ETL环节出了问题，前置环节，后置环节怎么设置链路。你们表的强依赖是用什么做的？
2. 当你的数据为非正态分布的时候，你出现数据倾斜要怎么处理。
3. 你们日志数据量有多少？业务数据量有多少？

# 文思海辉

1. spark提交流程和任务切分
2. rdd、df、ds区别
3. sparksql有几种join
4. 用一条shell把当前文件夹下（有子文件夹）所有带sql的文件和文件夹拷贝到指定位置

# 珍岛集团

## 面1

1. 自我介绍
2. 之前干了什么
3. 数据量多少，技术选型，项目架构
4. flink容错，ck的流程，求了个uv
5. hbase，rowkey，regionserver挂了怎么办
6. java的储存空间
7. 几个集合类的区别

## 面2

1. 大数据部门几个人？
2. 平时集群出现过什么故障？
3. 集群搞过什么参数调优？
4. 资源队列分为多少个？资源队列任务挤栈导致任务失败出现过吗？
5. 为什么设计flume到kafka到flume？
6. 用户行为轨迹怎么存储，处理？lastpageid
7. Hbase用了phoenix对hbase的性能有什么影响
8. hbase的row key 设计
9. flink的checkpoint存在哪？
10. flink现场出题：现在有一个一个小时的数据源，我想要flink每秒获取数据怎么办？window如果每秒开个窗口？对性能有影响，有没有好的优化？
11. 谈谈对flink流批一体的理解？
12. flink延迟比较高？
13. sparkrdd如何创建？
14. hive优化？
15. utf udaf等写过吗？
16. sql数据倾斜遇到吗？
17. java常用类？集合相关，多线程，线程和进程的区别
18. scala class和object的区别？等等
19. 即席查询用什么
20. 住在哪、薪资期望，现在薪资

## 面3

1. Flink 如何对uv实时统计1h数据/每秒数据
2. Flink checkPoint机制
3. rowKey设计
4. OLAP/OLTP
5. hive优化
6. Java静态变量 全局变量区别
7. Java8新特性

# 源犀科技

1. Sql查询关键字的执行顺序
2. hive调优，谓词下推，联合主键能不能用谓词下推，联合主键能不能用group by
3. 手写二叉树反转，不会，那写冒泡排序吧
4. 异步io怎么实现的，是nio吗
5. mysql大表join小表哪个表在前，hive呢
6. 布隆过滤器用过吗，讲一讲
7. 暂停一个线程的方法有哪几种
8. Parquet和orc的区别，为什么用parquet
9. hashmap的特点
10. gc和jvm了解吗，讲一讲
11. 实时数仓动态分流怎么做的
12. 拉链表怎么做的
13. 维度退化是怎么做的
14. kafka分区分了多少个，根据什么分的
15. sqoop增量导入，如果不给你日期字段按照主键自增，怎么做
16. 求一个月内的连续三天有交易并且三天交易额总和最大的用户，讲一下思路

# 序彦泽游戏公司

## sql题

1. 一个是计算连续七天登录的用户
2. 计算游戏关卡每一关的用户过关率以及用户在这一关放弃整个副本的几率
3. 计算每天手机授权IMEI的比率？（因为用户可以先授权然后再取消再授权。）
4. 统计连续三十天，新用户的充值的金额和拉新用户成本的比率
5. 新用户充值的数量和一次拉新活动的总人数的比值

## 面试题

1. 行为日志类型，分析一个具体业务场景的漏斗模型表
2. 对页面日志有疑问？
3. 用户不授权IMEI怎么计算日志
4. 日志中哪种类型最大
5. hive的自定义udf，如果要将用户ID转换为姓名怎么做？
6. 这个表能放到hdfs上吗？
7. 一个hivesql，优化思路是什么？
8. MySQL优化，什么索引运行最快？
9. 函数和存储过程的区别是什么？
10. clickhouse是强类型的吗？比如1=字符串1可以吗？
11. clickhouse中最常用的函数是什么？

# 晶格数字

1. 讲讲数仓分层
2. 雪花模型和星星模型的区别

# 普兰金融

1. clickhouse了解 replacingmergeTree和aggregatemergeTree有什么不同（当时就蒙了） replacingmergeTree还有怎么去的重
2. 如果表中添加一个字段flinkcdc具体怎么实现的
3. 讲项目自己负责的东西 我说的实时项目
4. 聊flinkcdc是怎么从数据库中到kafka的？比如它是怎样去拿数据库数据的，基于什么拿的等等
5. 实时怎么实现和离线数据保持一致
6. Threadlocal的了解
7. flink如何实现从source到sink端数据一致性
8. 如何使用实时我要查最近一年的新增用户
9. 离线实时两个数据会出现不一样嘛 怎么测试出来
10. 离线分层 以及每一层都干了什么
11. 离线讲用户维度怎么实现的拉链表
12. 不用拉链表还可以用什么
13. 元数据的了解

# 金仕达

1. ArrayList和LinkedList的区别，适用于什么情况；
2. 同步方法
3. 排序算法
4. 查询近30天之内带.log后缀的数据
5. Hadoop集群部署
6. HDFS的上传流程
7. Hive内部表和外部表区别
8. 分区和分桶的区别
9. Hive优化

# 途虎养车

1. 为什么使用日志打印框架，不使用消息中间键存储
2. canel，Maxwell，FlinkCDC的区别、
3. 为什么日志数据不放在消息中间键，放在日志服务器。
4. 大数据没有自己的从库吗
5. 监控指标从哪里获取

# 银科控股

1. 为什么选用kafka作为消息中间键
2. kafka是如何保证快速的
3. 为什么选用hive做分层
4. hbase的底层：HBASE可以是行存也可以是列存，看你字段有多少。HBASE底层实际上是lsm树，就类似于b+树的变种，牺牲了一部分读的性能，大大提高了写性能。clickhouse就单纯的列式存储。单表查询极快。
5. select \* from 表名 group by 执行顺序，

# 九章云极

1. Flink的运行架构
2. Flink的并行度
3. Slot个数小于并行度会怎么样？
4. 算子的优先级
5. Flink窗口的介绍，你们项目中用到了哪些窗口？
6. 用过广播流么？讲讲，并讲讲用广播流需要注意什么？
7. Flink的状态后端介绍一下，你们项目中用到的状态后端是什么？
8. Flink的提交模式你们用的什么？讲讲提交流程？
9. HDFS的读写流程讲讲
10. HDFS的常用命令讲讲
11. Yarn的提交流程讲讲
12. 讲讲Kafka的架构组成
13. Hbase的组件什么？读写流程讲讲？Rowkey的设计你们项目用到过么，讲讲
14. Zookeeper的一致性原则
15. Java面向对象的三大特性？
16. String Buffer和String Builder的区别？
17. 抽象类和接口的区别？
18. HashCode相等，值想不想等？值相等，HashCode是否相等？

# 比心

1. Redis是单线程还是多线程？
2. Flink有哪些API分层？
3. Flink并行度设置方法，优先级别？
4. Flink并行度和kafka并行度有什么区别，不相等会照成什么样的情况？
5. 如果并行度多于Kafka的分区数
6. Flink里面的窗口有哪些？
7. Flink的时间语义？
8. Watermark的迟到时间的数据怎么办？
9. 用的Datastream还是用FlinkSQL？
10. 模式匹配的规则，怎么触发的？
11. Aggregation里面包含哪些必要的方法？
12. Pocessfunction里面用了哪些方法？
13. Kafka消费数据是在哪里指定消费的？
14. Flink的反压？
15. 消费能力不足的话，增加并新度，增加消费者
16. 算子反压怎么解决？
17. Java里面的线程池有哪些参数？
18. Hashmap和Hashtable的区别？
19. Java的jvm有没有遇到查看gc的情况？
20. Flink有没有指定的垃圾回收机制？
21. 实时处理的动态配置是什么，配置信息表怎么处理的，配置信息表存在了哪里？
22. Hbase的rowkey是怎么设计的？
23. 写一个快速排序？
24. 冒泡排序的复杂度？
25. Dwm层里面关联的时候，数据来的时间不一样怎么处理？
26. Hbase的rowkey的设计，怎样会造成数据倾斜？
27. Hbase的底层存储和mysql的底层存储，hbase为什么能够增加一列而不会卡死？
28. Hive的行转列怎么使用的？
29. Stringboot在哪里用到的？
30. Suffle的优化？
31. Reducebykey和groupbykey的区别，实际带来的效果？
32. Kafka消费者消费不过来是怎么定位出来的？
33. Kafka怎么判断是数据倾斜还是消费者挂了？

# 复深蓝

1. 说一下会的技术,
2. 会不会springboot,
3. 讲一下最近的项目,
4. 行为数据怎么同步到大数据平台,
5. 除了flinkcdc还有其他监控binlog日志的工具吗?
6. 离线和实时采集过来的数据各放在哪里
7. 实时数据采集到kafka里面后怎么操作
8. 业务数据也是实时的处理吗
9. 一个数据从触发到大屏展现要多长时间?
10. flink在实时项目中起什么作用
11. 说一个主题看flink是如何实现join的
12. 跳出率是跟用户关联?怎么关联的?
13. 用过哪些join?
14. 事件有没有延迟
15. 实时项目中如何保持kafka的有序性?
16. 有这样一个场景:一张表的一条数据,先新增,再修改,那么传到库里面,是不是要先拿到新增的,再拿到修改的,这个数据才是正确的,这个场景怎么解决?
17. 如何保证kafka不丢数据?consumer那边有没有控制不丢数据的机制,比如某些数据压根就没消费到?
18. 有没有手动记offset?是把offset交给flink去管理,还是手动去维护
19. 有这样一个场景:flink有些数据跑着跑着报错了,下面你需要去回滚一部分数据,把那些报错的数据重新load进去,是怎么处理的?
20. kafka自动维护的原理是什么?怎么实现的?
21. kafkasink如何保证exactly-once的
22. 说一下hbase的热点问题?热点问题的原理是什么?
23. 预分区是如何解决热点问题的?
24. rowkey能直接用时间戳吗?
25. springboot里面常用的注解?有没有用过自定义注解?有没有用过aop?有没有用过多线程?
26. flume有没有写过自定义source?用什么语言写的?怎么写的?
27. hive这边自定义udf和udaf怎么写的？

# 国金证券

## 面1

1. 自我介绍
2. spark和flink哪个用的熟一些?
3. 对flinksql熟不熟
4. 两阶段提交
5. flink中4个slot并发去写mysql,怎么保证数据一致性?
6. 讲一下sparkstreaming实时项目的整个数据的流向以及架构
7. 业务数据指的是什么数据?
8. 基于某个页面的聚合怎么做?
9. 在clickhouse里面是一张打宽表还是多张大宽表?性能够不够?
10. 页面大宽表和事件宽表里面有多少万条?
11. 有没有调整过ck的性能?在这个过程中flink起到了什么作用?
12. flink中和维度表的join方式有几种?有没有用过broadcasthashjoin,小表广播?
13. flink有哪几类统计窗口?
14. 有没有写过flinkudf函数?
15. 说一下spark的shuffle过程?

## 面2

1. 讲项目
2. kafka数据一致性
3. 为什么DIM放到HBASE ，放到redis不行吗
4. 给你4000W条数据 放到HBASE 怎么设计
5. Hbase读写流程
6. flink 输出到没有事物的数据库 怎么保证数据一致性
7. 除了异步IO，还有什么实现方式

# 聚天时代

1. 写过的比较难的sql

**业务场景一：**一个淘宝页面,有用户id,用户页面id,点击页面的时间戳,求每一个用户在每一个页面的停留时间

**业务场景二**：有一个用户登录的全量表, 还有一个当天用户登录的增量表,字段有用户id,和最后一次登录时间,现在想把每个用户最新的登录时间写到全量里面,怎么操作

**业务场景三**：有一个销售报表,有3个字段,销售部门,月份和销售的金额,想在后面加一个字段是每个部门连续三个月的销售总额?

1. 写一个脚本,杀掉yarn上正在运行的程序
2. 说一下flink实时数仓的架构流程?整个成型之后的数据流

# 柯莱特

1. java的三大特性？
2. 什么是多态？
3. 什么重写？什么是重载？
4. 什么是JVM？JVM有哪些区？
5. java有哪些线程池？
6. java多线程怎么做？
7. java synchronized和lock有什么区别？
8. java基本数据类型？equals和==有什么区别？
9. spark有哪些算子？
10. groupbykey和reducebykey有哪些区别？除了预聚合的区别还有啥区别？
11. spark缓存和检查点的区别？
12. spark提交集群时参数怎么设置？
13. flink和sparkstreaming有哪些区别？
14. 讲讲hive有哪些调优方法？
15. mapjoin具体怎么操作？
16. SMBjoin怎么操作？除了分桶join还可以怎么处理？
17. 谓词下推是什么意思？作用是啥？
18. mapjoin具体参数怎么设置？
19. 遇到过hive数据倾斜吗？怎么处理？
20. 大表与大表join时，另一个怎么膨胀，具体怎么操作？
21. 用过哪些窗口函数？说说详细使用场景和区别？
22. 内部表和外部表的区别？
23. spark内存模型知道吗？具体讲讲
24. spark有哪几种shuffle？说说具体流程和区别？
25. 阶段怎么划分？什么是血缘关系？
26. 知道spark优化吗？
27. 项目中出现过flink反压吗？怎么处理？
28. 造成flink反压的原因是什么？讲讲flink反压的底层实现逻辑
29. watermark怎么用的？知道watermark底层是由哪个类实现的吗？具体怎么做的？
30. checkpoint常用来干嘛？除了故障恢复和精准一次性，还用来干啥？

# 和鲸科技

1. 讲讲数仓分层
2. 几道简单的sql
3. 场景题：统计会员在升到当前级之前，做了哪些行为，以及各自行为的比重？针对这个需求，怎么设计表？

# 智租

1. Flink的join方式
2. 讲一下Kafka的事务性机制
3. 拉链表的同步策略，如果更新成功插入失败怎么办

# 咪啰科技

1. 自我介绍+简单介绍下工作项目经历
2. 介绍一下离线项目的框架
3. Kafka的单条日志大小默认1M,你们是调到多少?
4. 调整HDFS的三个参数解决小文件问题,具体设置的参数是怎样的?
5. 只用到hive做数仓分层的计算的话,里面是MR计算,它的性能跟的上吗?
6. 做离线项目的时候选择Sqoop,还了解其他类似的工具吗?
7. 你整个框架的组件都是自己搭的原生的吗?还是CDH里面的?
8. 整个流程调度是用的什么工具?
9. 实时项目选用Flink,除了它是流处理之外,和SparkStreaming比还有什么优点吗?
10. Flink和SparkStreaming的窗口机制区别了解吗?
11. ClickHouse是你选的吗?为什么选?

# 生生物流

1. kafka丢数据怎么办？
2. HDFS高可用原理？
3. Flink丢数据怎么办？
4. 数据倾斜方案？
5. Hbase的rowkey设计原则
6. 知道Kudu吗
7. Spark的checkpoint原理

# 郑州UU跑腿

1. 指标有多少个，分析过的指标，
2. 技术方面，离线架构技术栈，实时架构技术栈，数据采集技术栈
3. 介绍一下hive，hive中的窗口函数，随机函数用过吗？
4. 介绍一下数据建模，
5. 需求，一个用户id，一个时间，求连续登录天数
6. olap，oltp区别
7. union和union all 区别
8. flink用的java还是sql

# 杭州米链科技

1. 离线数仓哪些表是累积型事实表，有什么特点
2. 离线架构描述
3. 多并行度下watermark的传递特点
4. 用了什么状态后端，怎么设置ck
5. 用过哪些状态算子

# 郑州华鼎供应链

1. 数仓架构，离线和实时
2. 分析过哪些指标？
3. 有没有从头到尾的分析过指标，也就是ods到ads，涉及到每层的？
4. 用的什么数据报表工具？
5. 数据存储在哪里？
6. hive用过吗，会用sql吗？
7. 听说过拉链表？
8. 了解orc吗？

# 序言网络

1. sqoop遇到了什么问题，为什么不用DATAX
2. 为什么用azkaban，有没有用过别的
3. hive你们用的外表还是内部表
4. 写过哪些脚本
5. 分桶表了解吗，怎么使用
6. 分区表和分桶表有什么区别，适用什么场景
7. sql：统计所以用户8月份最大连续天数，如果中间有断了怎么处理
8. sql：求1日到7日留存率，如何实现一条sql出结果。

# 吉祥航空

## 技术

1. 了不了解Map集合，底层？ hash冲突了解吗
2. 如何实现 Map的遍历
3. arrayList和linkList有什么区别
4. 介绍一下框架
5. 数据建模： 我说了四大步
6. 星型模型，宽表，还有个啥模型我没听过 有什么区别？
7. 为什么选用星型模型
8. 知道拉链表吗 描述一下怎么做的
9. hive知道哪些函数
10. 工作中遇到过什么较难的问题
11. 反射会用吗？
12. Flink了解多少

## HR

1. 离职原因，
2. 为什么学电气做大数据开发
3. 在上一家公司收获了什么
4. 个人优点，缺点

# 中新宽维

1. spark调度流程
2. spark提交任务参数
3. 血缘关系
4. 结合具体业务说一下spark和flink的区别
5. 结合具体业务说一下批处理
6. sparkstreaming项目分层
7. flink怎么建模的
8. hbase读写流程
9. Java直接客户端访问hbase有什么问题
10. kafka的offset怎么存储的
11. 未来发展规划

# 池鹜

## 技术

1. Hadoop生态、hive、spark
2. yarn提交流程
3. 有没有解决过生产中的问题

## HR

1. 你有什么优点
2. 问一下之前公司
3. 问一下上一次涨薪什么时候
4. 讲了一下他们的公司文化
5. 上下班时间以及福利待遇

# 中电金信

1. 讲项目，
2. 简历上的技术要点拎出来问
3. 怎么将数据写入ClickHouse
4. 离线的业务逻辑，建模思路

# weee

1. 维度建模怎么做的？你了解的说一下
2. azkban为什么选这个，你了解多说，你看中框架的什么功能特性？
3. 如果azkban中途挂了你怎么办，如果中间有脏数据或者任务跑不动你怎么解决？不断问还有吗，至少说出三种以上方法
4. 1 7 7 5 2 3 4 8 6 11 13 12 9 10这一串数据在5s窗口和2s的延迟时间下，都会进哪个窗口？
5. 第二个窗口能进的最大的数据是什么
6. 如果在flink流处理中，mysql数据（不止是增删，更多的是元数据变化，表结构改变）你怎么去保证实时流性能？
7. 像azkban这类框架最大重试时间，最大重试次数怎么设置（重点），你的原则是什么？为什么这么设置？
8. flinkcdc监控为什么选用，canal和它又什么不同？为什么会有这个不同，实现原理你懂吗？
9. 监控binlog是怎么监控的？说一下binlog的三种级别？行级别是怎么实现监控的

# 时溪信息 国泰君安证券

1. String和StringBuffer的区别
2. 序列化是什么，干什么用的，怎么实现
3. utf8带bom的utf Unicode gbk区别
4. 线程池是干什么用的，什么时候用
5. 算法题：一个字符串，在这个字符串后面最少添加几个字符能把这个字符串变成回文字符串。返回最少添加几个。
6. sql有过什么优化
7. Flink有过什么优化
8. Flink常用的算子
9. 熟悉的Hadoop组件
10. 场景题：100个文件，写个程序，怎么保证内存不挂的情况下执行完，一次放不下。怎么不断添加并执行文件，shell或者python。

# 序言泽

1. 用什么求日活，无法获取到设备id怎么办
2. 从mysql中导出数据，里面文本数据换行怎么处理
3. 添加redis缓存的时候是否会存在数据不一致的问题，具体哪里不一致？那些表变化会比较大？
4. 用户维度数据量很大，为什么redis能放得下？
5. hive创建的是什么表
6. 场景题：sql给你一个场景算最近十几天的七日留存率
7. 窗口函数用到哪些
8. HBase的架构及读写流程

# 百联

1. 介绍项目
2. hive优化
3. 数据倾斜
4. 小文件
5. 你们数据量
6. 数据质量监控怎么做的
7. 你们数据峰值多少
8. 除了MySQL，你还知道那些数据库

# 比智

1. 自我介绍
2. 介绍维度建模理论
3. hive优化
4. 数据中台和数据湖了解么
5. flink优势，比spark
6. 知不知道hudi（好像是这个词）
7. kafka优化

# GrowingIO

1. hdfs读写流程
2. hive的参数优化
3. flink的精准一次性怎么实现
4. 怎么检测hadoop集群的健康状态，怎么检测namenode是否健康
5. hive大表join大表如何优化
6. flink做过那些优化
7. 做项目遇到过那些挑战，怎么解决的
8. 什么情况下flink会挂掉，你通过什么参数调整优化flink

# 上海 瞬联科技

1. 介绍项目
2. hive数据倾斜
3. hive没有数据倾斜就是跑得慢,给你sql,你有哪些优化方案
4. spark宽依赖是什么
5. spark任务的切分情况
6. flink状态是什么,你怎么理解的
7. flink俩阶段提交是什么
8. flink与sparkstreaming的区别在哪里
9. lookupjoin是什么,怎么用的

# 苏州 盈天地

1. Mysql为什么使用B+tree
2. link checkpoint底层算法以及机制
3. ESmaster挂了重新选举产生脑裂怎么处理
4. ChlickHouse为什么快
5. hive用过哪些压缩
6. java多线程,用过哪些设计模式
7. hdfs支持多并发嘛,文件块大小128M过大和过小了分别会怎么样
8. flume文件滚动的三个参数
9. DataX与Sqoop对比
10. nn与2nn区别,为什么有2nn的存在
11. 用过FlinkCDC自己开发过读外部数据库嘛,除了你用的Mysql
12. Redis底层持久化,你们公司用那种
13. MR可不可以只要reduce不要map ？ 可不可以只要map不要 reduce？

# 上海 云智慧

1. mysql主从原理
2. awk和sed的区别
3. kafka和zookeeper的关系
4. hbase的架构
5. clickhouse的集群搭建

# 上海 捷奥

1. 自我介绍
2. 什么时候用watermark.
3. flink版本.
4. flink如何精准一次.
5. 你们用的flinksql还是datastream, 什么时候用什么.
6. 你们用的什么状态后端, 介绍一下.
7. hadoop的配置文件, 各自有什么作用, 端口
8. yarn提交流程.
9. hive写过mapreduce任务么.
10. 你们hive计算引擎为什么不用impala.
11. 内外部表区别.
12. hive数据倾斜.
13. hbase读流程.

# 上海 玛驷众

开始 3道sql 5分钟ac

## 基础

1. maxwell原理
2. hdfssink参数
3. taildirsource原理
4. kafka分区数多少
5. flinkwatermark是什么
6. 为什么用clickhouse

# 合肥 南瑞中天

1. 你项目多少人？你负责哪一块？负责什么职责？
2. 整体数据量,为什么选择大数据平台，不用系统平台？
3. 项目中承担角色，遇到困难有哪些？怎么解决？
4. 项目中的分工，项目的分层的依据？
5. 数据库范式有哪些？
6. 设计Hbase的方案和传统数据库方案的区别？
7. 两个sql题（比较简单）
8. spark和hadoop的区别？（spark为什么比hadoop快）
9. spark为什么是基于内存计算的？
10. spark的运行模式？
11. 为什么用到maxwell datax 都使用的？

# 深复蓝

1. 你们的实时数仓采集怎么保证数据的时效性和准确性
2. flink做了调优做过么?
3. 内存怎么调优的?
4. 你们的checkpoint怎么做的优化

# 南京 平安外包

1. 自我介绍
2. 根据业务讲一下你们最近的项目
3. 有宽表嘛，宽表字段丢失怎么做
4. 做一个业务要多久
5. flink和sparkstreaming的区别
6. flume遇到的问题
7. 一天100g的数据，存一年的话，数据量大了会有什么问题
8. 班级成绩排名sql，成绩相同排名相同
9. 怎么做加密，加密数据要用怎么办，我讲的md5，他问我md5怎么做恢复
10. 字段里面内容是×省×市×县，要按照省分怎么办
11. 接触过什么mpp数据库嘛，说clickhouse可能也不是那么优秀
12. 碰到过底层bug嘛

# 杭州 某医疗公司

1. 问你为什么要区分离线指标和实时指标
2. 数据分析了解吗，怎么做数据分析的
3. 问日活，问uv数据量
4. 为什么同步业务数据要使用datax和maxwell
5. 采集数据是由谁负责？业务数据具体采集方法？
6. 问我们大数据部门人数？有其他部门吗？具体做了啥
7. 你们大数据框架是咋定的？
8. 最离谱的一个问题：hadoop版本，有几个节点？

# vivo外包

## 面1

1. 你们实时做了多久?
2. 你实时做了哪些工作?
3. 你们uv怎么做的?
4. 你们怎么保证数据的准确性 还有 任务的稳定性?比如晚上机子挂了怎么办
5. 有脏数据怎么办?
6. 数据做了哪些去重,数据量特别大怎么办?
7. rocksdb做了哪些调优?

## 面2（实时）

1. 上个项目用的什么框架
2. 数仓的分层原理
3. 之前最大的qps是多少
4. 指标的口径怎么统一的（离线这边口径变了，实时这边怎么去获取的口径）
5. 做了什么优化
6. 问题怎么定位的
7. 之前项目数据量多大
8. 采集用的什么工具
9. checkpoint多次失败，怎么做恢复
10. 用的什么可视化工具，为什么用这个工具
11. 怎么保证的数据一致性
12. 之前的项目中是什么角色，主要负责什么

## 面3（离线）

1. 自我介绍
2. 数仓怎么设计
3. ck查询请求，如何查到数据和返回数据的
4. mr的shuffle和spark的shuffle有什么区别，优势在什么地方
5. 4个by区别
6. mr 的shuffle 过程
7. hive中的窗口函数有哪些

# 上海向晨

1. 怎么在海豚调度器中设置并发参数？

# 郑州 某公司

1. Flink的精准一次性
2. Flink如何保证数据有序
3. 离线数仓中Hive的数据倾斜
4. RocksDB出现之前，你们遇到大状态问题怎么解决的
5. 7天内连续三天登录
6. 为什么有Hbase了还要加入Redis，据我了解他们两个速度差不多
7. CDH的版本
8. 数据量、服务器个数、人员个数
9. HQL的底层转换
10. Hive的引擎有几种，分别有什么好处
11. FlinkCEP底层是什么

# 南京 紫金

1. 集群规模,kafka,hdfs,,,几台,怎么分布的
2. 采集通道搭建时你具体做了些什么,写了哪些东西,举几个例子
3. binlog 是什么文件格式 binlog包含什么信息 头部信息包括什么
4. 怎么理解多级分区的,你们怎么分区的,动态分区是怎么做的
5. kafka保证不丢,不重
6. kafka leader
7. kafka 有序,怎么保证全局有序
8. 有出现小文件 问题吗,怎么解决的
9. 数仓构建的理论基础,谈谈你对er与维度建模的看法

# 大连 搜配云

1. 项目如何保证精准一次？
2. 知道异步io吗
3. Flink反压
4. Flink内存管理。是使用yarn吗?提交方式
5. 知道docker吗？知道怎么用的？
6. Java的多态是什么?

# 叶子科技

## 一面

1. 阿里大数据one data是什么意思
2. 大数据之路这本书问了3个问题（不记得问题是啥）
3. 项目分层架构，数仓有几层，概述一下
4. MySQL的InnoDB引擎知道吗
5. 介绍一下MySQL的B+树
6. 介绍拉链表
7. hive调优
8. 大表驱动小表和小表驱动大表的区别
9. maxwell，canal，sqoop，flink cdc，datax的区别
10. snappy,lzo，gzip的区别
11. hive本身不自带snappy，你是如何做的
12. 老版本hadoop的重新安装
13. clickhouse的优势
14. Flink的三层API有什么区别
15. flink Watermark知道吗
16. flink 时间语义
17. flink分流是如何做的

## 二面

1. 介绍一下你的数仓分层
2. 说一下你们的数据域划分，为什么这样做
3. 不同的数据域举几个指标实现
4. 说一下你们具体的技术选型
5. 一般数据量很大才会使用hbase，为什么这样技术选型
6. 介绍一下拉链表
7. 维表有更新你们是怎么做的

# 神州信息-外包

1. 项目中遇到难题，及如何解决
2. spark架构及提交流程
3. kafka分区与副本如何设置
4. kafka如何保证数据有序
5. kafka事务如何实现，底层是如何做的
6. kafka新增消费者如何执行再平衡
7. flink checkpoint机制
8. flink状态后端
9. 二阶段提交
10. 为什么选择clickhouse
11. 说一下你了解的HBase
12. clickhouse和HBase的区别
13. 项目设计到落地具体流程及你做了哪些
14. 场景题：HBase中如何设计表，及关系型数据库中如何设计
15. 人员和角色多对多的对应关系，需求是人员和角色的新增和删除

# 亿通国际

1. 笔试手写sql，写完了面试官说最近没几个写出来，之后是闲聊
2. hive调优
3. hive分区表有什么好处
4. 说几个指标实现
5. dws层有多少表，大概有多少字段，举个例子
6. 说一下拉链表
7. flink Watermark
8. 实时和离线数据不一致，你们以哪个为准
9. MySQL索引
10. MySQL最左匹配原则了解吗，给了个简单的场景题说哪些走索引

# 南京 南瑞瑞中

1. spark架构
2. spark和hadoop区别
3. spark为什么基于内存计算,是不是完全基于内存计算的
4. DAG如何生成的
5. flink表与表join的方法
6. 谈谈对flinkcdc的理解
7. kafka精准一次
8. kafka如何保证数据有序
9. 说一下ds,为什么用ds
10. 说一下项目中的技术难点

# 北京 分贝通

1. 面试官是大佬，面了70分钟
2. 整个实时数仓中数据流是什么样的？
3. 你们数仓中有往Hbase写，里面的row\_key怎么设计的？
4. 实时里面资源具体怎么给的？
5. 你们实时数仓里的Flink任务具体怎么提交的？
6. 算法题：求二叉树最底层的那一层最左边的结点值（深度优先）？
7. Linux的查看资源的命令，top命令打开后有几个数值，分别代表的什么含义？
8. 对java的熟练度怎么样？有没有写过Spring？
9. 进程和线程怎么理解的？如果线程池有5条线程，它们共用内存嘛？会不会共用CPU？有8个核，5条线程会用其中多少资源？

# 北京 广推 数据分析师

1. 自我介绍
2. 聊一下作为数据分析师，从前到后你关心的是数据的哪些方面？
3. sql的执行顺序？
4. 题目：25匹马赛跑，1个赛道，每次5匹进行比赛，无法对每次比赛计时，但知道每次比赛结果的先后顺序，最少赛多少次可以找出前三名？（题目有没有问题？怎么解答？）
5. 题目：一家品类top10、有投放广告购买流量和自然流量的抖音电商6月份GMV环比1-5月均值下降了10%，作为业务侧的数据分析师，需要你分析原因，你分析的思路是什么？

# 上海 缔塔

1. 在实时数仓具体完成了哪些工作？
2. Flink做过哪些优化？
3. Flink出现反压怎么处理的？
4. Flink的精准一次？
5. 两阶段提交？
6. Flink具体怎么维护偏移量的？
7. watermark机制？
8. Flink跟Spark Streaming在checkpoint上的区别？
9. 存checkpoint的文件夹打开看过吗？都有什么？
10. 写数仓代码的时候并行度是按照什么定的？
11. 实时数仓里kafka的主题数和分区数是怎么设计的，怎么跟Flink做对应的？
12. HQL转换为MR流程

# Fintrue科技

1. 公司数据量,行为数据多少?业务呢?你们dws层一共有多少张宽表?
2. 讲到kafka 打断,数据怎么消费的?实时的时候数据呢怎么消费?数据量有多少?
3. 场景：讲到dws,根据业务要求做一些宽表,假如这个表是记录用户行为,比如我们有一个用户表要进行更新这个表很大,他是一个宽表处理时间会很长,对于下游要取数据,肯定有延迟,这个时候需要给下层做一个标记,确准这个表已经更新完成,他肯定得有依赖关系,这种依赖关系业务中遇到,你是怎么做的这个标记?
4. hive 的性能在大数据领域计算能力你觉得怎么样? Impala 呢?
5. clickHouse讲一下，clickhouse 存储和hdfs 存储 的区别?clickhouse 可以按天去刷数据吗?
6. 场景:业务上汇率的问题,他重刷数据可能是一天或一个月,适不适合大数据场景，二是实时的更新某一天的某一条数据,这种场景适不适用，如果有你这种数据用什么来存,这种关于存储的框架你该怎么做技术选型?
7. superset: 是你自己搭的吗,是有专业的运维处理吗,

# 上海恒格信息(外包保险离线)

1. 说完离线跟我说以后别说这么细,简单讲就好
2. mr 的整个流程
3. hive内部表外部表,
4. hive 数据倾斜
5. hive开窗函数用了哪些
6. ow\_number、rank和dense\_rank区别
7. 行列转换
8. hive on spark 和spark on hive 区别
9. Spark中使用了哪些算子
10. spark算子,举例都用在哪，spark任务提交流程
11. spark 任务切分,Task任务调度
12. 说一下RDD,惰性求值计算机制讲一下
13. sql题:给用户,时间, 7天连续登陆思路,断一天也算思路

# 兴业银行外包

1. 说最近项目，实时·
2. 问具体flink中做了哪些工作（指标，优化）
3. 介绍Checkpoint
4. 介绍clickhouse
5. 问java会不会写微服务，让别人去ck拿数据那种，但又不能直接访问数据库?
6. flink的job提交流程？
7. 你们的容错怎么做的？
8. 数据转换数据交换怎么做的，提升它的效率
9. kafka三种消息投递语义
10. kafka哪些情况下有数据丢失的问题？
11. hive分区分桶的区别
12. 窗口函数列举一些
13. UDF、UDAF、UDTF区别
14. spark数据倾斜怎么定位，怎么解决？

# 谷幽（海隆软件）

## 初始（15分钟）

1. 说一下最近做的项目
2. 说一说为什么要数仓分层
3. 说一下数据倾斜
4. 说一下hive小文件问题

## 复试（15分钟）

1. 介绍了一下最近的项目，说的实时架构，又说了一下建模
2. 问hive优化
3. 问小文件问题
4. 问combinehiveinputformat是逻辑优化还是物理优化

# 孤波

## 公司里架构师给做的面试

1. 3道手写Java题
   * + - 1. 写一个list，往里放几个Integar，把大于5的数剔除
         2. 写两个String，特别长，转换不了Integar，如何进行相加操作？
         3. Java的流式处理会写吗？被提醒用Java 的Stream API
2. 3道Spark 题 1T的数据 100 Core 400G内存
   * + - 1. 游戏分区，求每个区用户排名Top 5？
         2. 取所有用户排名Top5？
         3. 如何把所有用户的排名按照从小到大排列，输出到 HDFS上？
3. 请讲一下函数式编程
4. 你们用户画像标签工厂是用来干什么的？
5. Flink的水印和窗口源码有看过吗？是在哪个具体的类？
6. 你们维度数据要做ETL吗？除了用户信息脱敏？没有做其他ETL吗？
7. 你怎么保证用户填写的出生日期是正确的？
8. redis会什么？
9. redis做旁路缓存怎么和HBase保持数据一致性
10. 不给你RabbitMQ，Kafka，现在让你用redis做消息队列，你要怎么做？

## 直接主管面试

1. Flink每一层都存储在哪？你觉得这样一个架构有什么优点？有什么缺点？
2. 你们用什么来监控Flink作业的？
3. 你们怎么上线这个Flink程序的？
4. 你对分布式有什么想法？
5. HDFS满足CAP原则吗？
6. Kafka怎么保证数据高可用和分区一致性的？
7. 上面提到了leader和follower，问我怎么选的，我说controller通过副本同步队列选的，他问controller挂了怎么办？
8. 你懂PAXOS算法吗？
9. 你对分区和分片是怎么理解的？
10. Buffer和cache有什么区别？

# 春秋航空

1. hive小文件
2. Java线程、集合 介绍一下
3. Kafka精准一次
4. Sql题 七连三
5. Hive基本数据类型
6. 如果源关系数据库频繁做数据变更，hive计算的比较慢，hive怎么保证跟源头数据一致
7. hbase有个别超大表，其他都是小表，怎么办
8. Hql编译过程
9. Hdfs读写流程
10. 建模，聊聊，关系型建模和维度建模优缺点
11. 冒泡排序（java的） 笑死根本不记得，就说整俩for循环
12. Hbase怎么保证数据不丢
13. Hive里timestamp怎么转date

# 江苏网进（苏州昆山）

1. HDFS读写流程
2. HDFS小文件
3. Hive SQL翻译成执行任务步骤
4. Hive数据倾斜
5. Hive分区和分桶什么时候用
6. Flink和Spark Streaming的区别
7. Flink分区

# 四川星点网络（仙谭酒业）

## 初试

1. 自我介绍
2. 讲一下flink的批流一体以及容错性
3. 讲一下hadoop读写流程
4. 讲一下数仓建模
5. 说一下离线中做了哪些指标
6. 让具体说了一下留存率是怎么做的
7. 有没有用过拉链表，讲讲
8. 说说自己的优势
9. 又问我建模完成后怎么检验整个数仓建模的质量
10. 共享屏幕写了一道留存率sql

# 中金电信 （平安信用卡项目）

1. mr和hive on sprak的区别
2. 有没有遇到过数据丢失，或者重复的问题 ，怎么解决
3. 平时数据量多少
4. ES和CK的区别
5. hive中有没有遇到过数据倾斜，怎么处理的
6. 数仓建模方面聊聊
7. 平时负责比较多的工作

# 美味不用等（南京）

1. 你们数仓架构选型
2. 你们数仓的是怎么建模的
3. 会Python嘛
4. 目前来说你们公司日活多少
5. 公司一天数据多少
6. 公司现在每天订单量有多少，
7. 公司现在的日志有过亿嘛
8. 饶了好大一圈就是问Hive的优化，

# 上海天马微电子

1. 自我介绍
2. 最近的一个项目，框架，每层做什么
3. 你日常在工作中干什么？
4. hadoop的调优
5. 实时 中选用Hbase存储维度表的数据？
6. 列举各种数据库的区别：
7. 在项目中哪些地方用Clickhouse？
8. 在项目中 为什么使用maxwell /Kafka 做一个中间件 不直接使用 Flinkcdc拉取数据？
9. 选用redis的一个原因是什么？
10. 由于某个表的数据量特别大,导致oom，节点挂了怎么处理？
11. 跑任务特别慢？
12. hudi相关？ 关于数据湖的问题 算是扩展问题 答的不好也没关系
13. 非技术类问题：

# 某网络科技

1. flume三个channel的不同
2. flume的filechannel 怎么保证数据不丢的
3. kafka的数据回溯有没有做过
4. shell 脚本a，b，c三个a和b需要同时执行，c需要等a和b执行完再执行，自己手写shell脚本实现
5. flink双流join
6. 每10秒开窗的数据量是多少
7. flink的状态后端都用哪些
8. springboot是自己写的吗

# 嘉环科技

1. 自我介绍
2. Hive（Hadoop相关）
   * + - 1. HDFS的用户权限管理是怎么做的？文件夹权限744代表的是什么？
         2. HDFS的下载和上传命令是什么？如果我要下载HDFS上Yarn的错误日志该怎么做
         3. Hive的小文件问题遇到过吗，怎么处理的？
         4. Hive支持的文件格式有哪些？你们是怎么做的，为什么选这种？
         5. 集群的NameNode起不起来怎么办？一般是怎么排查的？
         6. 你们的集群是多大，每台配置是多少？为什么要这么配置，有没有做过服务器新增或者退役？
         7. 用过哪些窗口函数
         8. Hive的数据倾斜是怎么处理的
         9. 执行计划你们会用吗，一般怎么用（什么时候会用）？
         10. 你们hive用的是什么引擎？(回答用的hive on spark)为什么不用MR？那为什么shuffle过程会导致效率的下降？
         11. 你们的副本策略是什么？有几个副本？副本的存放都是怎么存放的？
         12. Yarn的提交流程
         13. hive的动态分区功能有没有用过？为什么要用这个功能？
3. Flink
   * + - 1. Flink的分区策略是什么？
         2. Flink的日志文件一般存储在哪儿
4. Kafka
   * + - 1. Kafka是怎么保证不丢不重？
5. Hbase
   * + - 1. Hbase的写流程是什么样的？
         2. Rowkey设计是怎么设计的？
6. Spark
   * + - 1. spark的宽窄依赖是什么？
7. Java
   * + - 1. 抽象类和接口的区别
         2. 我看你说用了异步io，那新建线程有哪几种方式呢？
8. Linux
   * + - 1. 说说linux的常用命令？你们在什么时候会用top，top是看什么的？知道scp命令吗？会不会用？如果我要找关键字为err的文件日志要怎么找？具体命令是什么？
         2. vim常用吗？是干什么的？如果我要在一个很大的文件中找到port这个关键字，要怎么找？不区分大小写查找应该怎么做？怎么显示行号 ：set nu
         3. var/usr分别是存储什么文件的？
         4. linux系统的配置文件一般存放在哪个文件夹？
9. MySQL
   * + - 1. 熟悉mysql库吗?（大数据相关的hbase和clickhouse这些更熟一点，mysql只使用一些最基本的功能）
         2. MySQL你们用的版本是什么？
         3. 知道min.ux这个函数吗？知道是什么吗？那lag函数呢？

# 精·Flink面试总结

## Flink提交

1. flink怎么提交
2. flink集群规模？flink的数据量？在flink项目中做了什么？
3. flink提交作业的流程，以及与yarn是如何交互的？
4. yarn-session与Per Job优缺点
5. flink提交job的方式以及参数如何设置？ 页面提交和客户端提交有什么区别？
6. Flink的JobManger，提交有多少jobmanger
7. Flink的TaskManager
8. 说一下slot，业务中一个TaskManager设置几个slot，连接的kafka的分区数是多少
9. 怎么修改正在运行的Flink程序？如果有新的实时指标你们是怎么上线的？

## 状态编程

1. 说一下状态编程（operator state，keyed state）
2. flink的状态是什么，分为几种？
3. 10个int以数组的形式保存,保存在什么状态好?VlaueState还是ListState?存在哪个的性能比较好?
4. 使用MapStage，group by id 如何设计
5. 继续上面的MapStage，id不放在key行不行
6. flink是如何管理kafka的offset，使用什么类型的状态保存offset？
7. 一个窗口，现在只取第一帧和最后一帧，怎么做？

## 反压（背压，数据积压）

1. flink用什么监控，如何有效处理数据积压
2. 遇到Flink不太能解决的问题.(PV,UV放内存，OOM了，后面配合redis以及布隆过滤器)
3. 使用flink统计订单表的GMV，如果mysql中的数据出现错误，之后在mysql中做数据的修改操作，那么flink程序如何保证GMV的正确性，你们是如何解决？

## Spark与Flink对比

1. Spark与Flink区别
2. Flink的key By和Spark的 group by有什么区别?
3. spark有哪些优化
4. Flink怎么优化
5. 遇到SparkStreaming不太能解决的问题
6. 必需要手动维护offset吗?(转到了Flink去解决这个问题)
7. Sparkstreaming和Flink消耗资源具体数据对比
8. 为什么要用Flink替代SparkStreaming(应该深入的去讲一下Flink)
9. 在什么场景下需要这么高的实时性

## Checkpoint

1. flink checkpoint的实现原理（容错机制，故障恢复，分布式快照，checkpoint，）
2. flink的checkpoint机制以及精准一次性消费如何实现？
3. 精确一次，至多一次，至少一次对checkpoint有什么影响
4. Savapoint了解多少
5. 作业挂掉了，恢复上一个Checkpoint，用什么命令
6. 什么是Flink的非barrier对齐，如何实现？

## 窗口与Watermark

1. flink时间语义
2. 什么是Watermark及主要作用？什么时候去触发计算?
3. 消息超过watermark的时间会丢失数据吗?（允许迟到，侧输出）
4. 开窗函数有哪些？（五种）
5. flink开发哪个窗口用的最多（最好随手举一个例子表面怎么用的）
6. 既然是开窗为什么一定要转Flink（说时间语义）
7. 广告在没有人点击的(也就是没有数据流的时候)窗口,这个窗口存在吗?有没有对这些窗口进行校验的窗口.
8. 1小时的滚动窗口,一小时处理一次的压力比较大,想让他5分钟处理一次.怎么办?(自定义触发器)
9. flink开窗五分钟过来一亿条数据你是怎么处理的
10. flink开窗5分钟被同一用户连续访问60次，需要把他的访问信息调出来 你是怎么做的

## 双流join

1. Spark和flink的双流join的底层原理
2. A表left join B表
   * + - 1. A表数据来了，B没来
         2. 2）A表数据来了，B在规定时间内到
         3. 3）A表数据来了，B在规定时间后面到(此处规定时间，就可以很好的利用起来说一下两种算子优缺点)

这个问题，process中两种算子（connect，join）分别说明，Flink SQL（撤回流）可以写两种风格，种类很多，需要细细品

## 杂七杂八

1. process用的种类(8个，最好中文名都记一下，不需要都掌握，可以把最熟悉的在项目在项目中怎么用说一下)
2. flink的内存管理？
3. flink的序列化机制？
4. Kafka数据很多，内存很少，读取数据都是问题，现在想要写，怎么控制写速率
5. Rich Functions与Functions区别
6. flink里面异步IO代码具体怎么写的，每一步具体描述出来

# 花旗面试题

### Java部分

1. ==/equals/Hash code
2. Reflection反射
3. IOC
4. DI(依赖注入)
5. AOP
6. overload/override
7. PreparedStatement/Statement
8. Select\*
9. Merge(mysql)
10. In/exists
11. Union/Union all
12. 各种索引Index
13. Connection Pool
14. Link list/ArrayList/Vector/Set
15. 15.Map/HashMap/CoHashMap
16. 16.多线程实现方式
17. Junit/unit test
18. Socket网络通信
19. Transactional

### Spark部分

1. saveAsTextFile
2. saveAsTable
3. saveAs
4. Spark内存模型
5. partitionBy和Repartition
6. Repartition和coalesce
7. Cache/Persist
8. CheckPoint
9. RDD, DF, DS

### Hive部分

1. 四个By
2. 自定义函数
3. hive文件存储格式
4. Hive中数据导出表的方式
5. 分区，分桶和Index
6. hive view