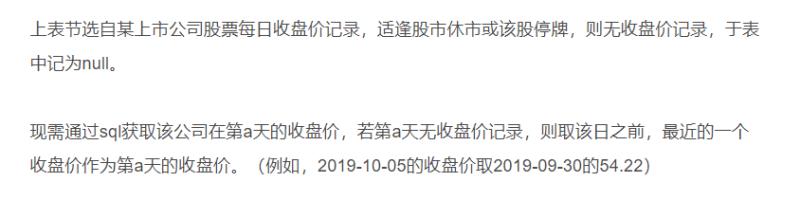
上海校区大数据企业面试真题

版本：V1.0

# 一、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？



# 二、美团外包

1. 实时方面
   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.(我说的是手动维护Kafka的offset实现一致性消费的问题)
   8. 必需要手动维护offset吗?(我转到了Flink去解决这个问题)
   9. 遇到Flink不太能解决的问题.(我没多说,其实应该说大数据量使用redis布隆过滤器实现UV去重的)
   10. 实时指标出来后的应用场景.(我提到了ES预警)
   11. 预警是怎么做到的.预警的条件.(公司内部的预警和用户行为的预警)
2. 数仓方面:
   1. 当初建模的时候应用场景是什么样的.
   2. 建模的流程(我是从数据源开始讲)
   3. 都有哪些数据同步到数仓里面(我说大概20多张表)
   4. 对这些表有过什么分类吗.(我说的同步策略)
   5. 哪些表是相应的同步策略.
   6. 跨天支付数据是怎么处理的.
   7. 用户表为什么一定是拉链表.
   8. 数仓的分层.
   9. 如何找出来用户的一天的行为轨迹.(这个问题出现在描述dws层)
   10. dws和dwt的宽表都有哪些,并且都是什么!!!(详细都说出来!!!)
   11. 出口对应的指标.(我太菜了.没说出来几个)
   12. 你们是怎么保证数据质量的.(居然没想起来监控框架.我太菜了.)
   13. 数据质量监控的角度,(我只记得数据量)
   14. 有制定一些量化的'含量指标'(应该是这个词)也就是说怎么衡量这个表好用不好用.
   15. 有没有一张表实现所有的分析.
3. SQL题(很简单)
   1. 外卖的配送ID
   2. 外卖员的ID
   3. 订单配送的City
   4. 时间的CT
   5. 一整年中,每个月每个城市订单量Top10.
4. 其他
   1. 3.1.为什么考虑换一份工作.
   2. 3.2.离线和实时更偏向哪些(我说的实时),为什么?
   3. 3.3.工作后做的最有成就感的一件事是什么.
5. flink，ck机制，内存管理，出现反压怎么处理的？
6. kylin如何直接构建cube？

# 三、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

# 四、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（你们用到了哪些框架），你们的数据量，你们的集群规模，用的是Apache还是CDH，怎么维护集群的（集群监控）
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

# 五、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？
    1. 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？

# 六、润和

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 （hbase是有事务吗求解）

# 七、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我，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？

# 八、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. 问Java和Scala谁更熟的时候，大家一定要说Scala，说了Java会被问死，我说的Scala，就拿我没辙了
12. flink窗口5种
13. flink时间语意
14. 整体数据走向，分开来说
15. 有没有hive相关的API，我说我没做过，不清楚
16. Java中GC原理
17. Java中无符号状态如何有效处理
18. 还有很多简单一点的问题，面试太久了，我没录音，主要看第11点，我给大家的建议，除非单独Java开发过的

# 九、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. 5flink监控，如何有效处理数据积压
6. 离线那些很简单，随便说下海哥整理的那一套就行了
7. 最后说是不是培训，我果断否定，flink火了别的公司用，我们也能用，我们就是半年前换了flink

# 十、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

# 十一、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，让我说一下yml语法，懵逼
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. 问我es的端口号，紧张忘了，我反问他canal端口号，我还提示说双十一相关，他也不会[666]
9. sparkstreaming怎么消费kafka（就是2种方式）
   1. 做过数据治理吗？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？（答：javaEE改表为什么不跟我们说，谁给他们勇气的）
   2. 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？（答：我们用户表建立了分区表，拉链表与分区表之间没有关系，拉链表可以建立分区表，但是我们没有建立，比如说可以按照省份进行建立分区表）

# 十二、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(不知道他当时问的是不是这个，我讲的这个)。
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，还不算太难，但是不能写错一点，不然她会觉得写的不是特别好。由这个炸裂她会引出业务中的炸裂
9. 问为什么自己定义udtf而不是直接用炸裂

# 十三、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

# 十四、金大师

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. watermark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

# 十五、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，让我说一下yml语法，懵逼
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. 问我es的端口号，紧张忘了，我反问他canal端口号，我还提示说双十一相关，他也不会
9. sparkstreaming怎么消费kafka（就是2种方式）

# 十六、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（1.5版本之后就又flink sql了，1.11多了hive）
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，说我概念记得挺牢的（我没理他）
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. Map与mapPartition，我说完说我让我看源码
15. Task与 partition 有什么关系
16. Stage,宽依赖
17. Kafka一直说
18. 一个topic有3个分区，两个消费者，会怎么样
19. 一个topic有2个分区，三个消费者，会怎么样
20. Kafka怎么处理大量数据（为什么这么快），说我概念背的好，但是细节不是很注意
21. Hdfs小文件处理，spark处理小文件
22. arraylist在 Scala有什么可以做到同样功能，比较像的
23. 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
24. 最后闲聊，数仓分层

# 十七、赢时胜

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(应该深入的去讲一下Flink)
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?(应该是想问网站的网址?).大数据分析对外部提供服务是在哪个web或者app(直接告诉他公司内部使用!)
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.(具体分哪些模块)
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?(自定义触发器)
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?

# 十八、花旗

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

# 十九、微盟

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（这里我解释了幂等性和事物和ack-1，面试官揪着这里问了很久，他关心的事出了问题之后怎么解决）
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
9. 都是根据你的简历，然后给你场景，让你分析，不然你介绍项目，他很抠细节，面试官年龄不大，2,30岁的样子。

# 二十、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. 光说个概念不行，会问你逐层构建每轮mr做了什么，要讲清楚，否则会一直问，我说不会，他还是追问了下。
14. 拉链表 也问的很细

面试官问的很细，都是离线的，而且每涉及到一个知识点，都会问你底层用mr是怎么实现的。不能只回答表面，会一直追问。

# 二十一、宝尊电商

1. spring 生命周期
2. 线程生命周期
3. flink的数据量
4. azkb几个工程
5. azkb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

# 二十二、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?

# 二十三、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？

# 二十三、滴滴

1. redis底层map实现的原理
2. 为什么从SparkStreaming转移到Flink？就是因为flink功能强大吗？
3. SparkStreaming的checkpoint与flink的checkpoint的区别？
4. 4.5亿用户如何海量去重，我说使用redis自定义布隆过滤？他说这么高的并发，redis抗的住吗？
5. 你懂HBase吗？那么lsm树是什么？
6. 网络传输分几层
7. TCP三次握手
8. redis为key设置过期时间的底层原理是什么？

# 二十四、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

# 二十五、软通

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

# 二十六、信也科技

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？

# 二十七、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

# 二十八、中软

（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

学长二：

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？（我那知道）
5. Spark背压机制
6. Spark中使用那些算子？（我说map，flatmap的时候他说你还用这个，我人傻了）
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？

学长三：中软外包平安医疗

1. Hive行列过滤是什么意思

答：列处理：在SELECT中，只拿需要的列，如果有，尽量使用分区过滤，少用SELECT \*。

行处理：在分区剪裁中，当使用外关联时，如果将副表的过滤条件写在Where后面，那么就会先全表关联，之后再过滤。

1. 列式存储区别

答：常用的包括：textFile，rcFile，ORC，Parquet，一般企业里使用ORC或者Parquet，因为是列式存储，且压缩比非常高，所以相比于textFile，查询速度快，占用硬盘空间少

1. ORC与Parquet有什么区别

答：1.orc不支持嵌套结构(但可通过复杂数据类型如map<k,v>间接实现)，parquet支持嵌套结构。

2.orc与hive的兼容性强，作为hive的常用存储格式。  
3.orc相比parquet的存储压缩率较高。  
4.orc导入数据和数据查询的的速度比parquet快。

1. sqoop导出的事务

答：–staging-table方式（建立临时表，通过sqoop导入到临时表，成功之后再把临时表的数据通过事务导入到mysql的业务数据表）

1. 导入HDFS中map任务挂了怎么办
2. dws层和dwt的区别（怎么实现的，dwt是否分区）,分别是怎么做的
3. 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
4. Dwt不分区
5. 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）
6. 在离线中，手动修改了数据
7. 数仓中是怎么做ETL的
8. 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） 我我回答直接删掉重新导一次
9. 新增及变化是怎么区分开的（what）
10. 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）
11. 数仓中分了哪些主题

会员

设备

商品

地区

活动

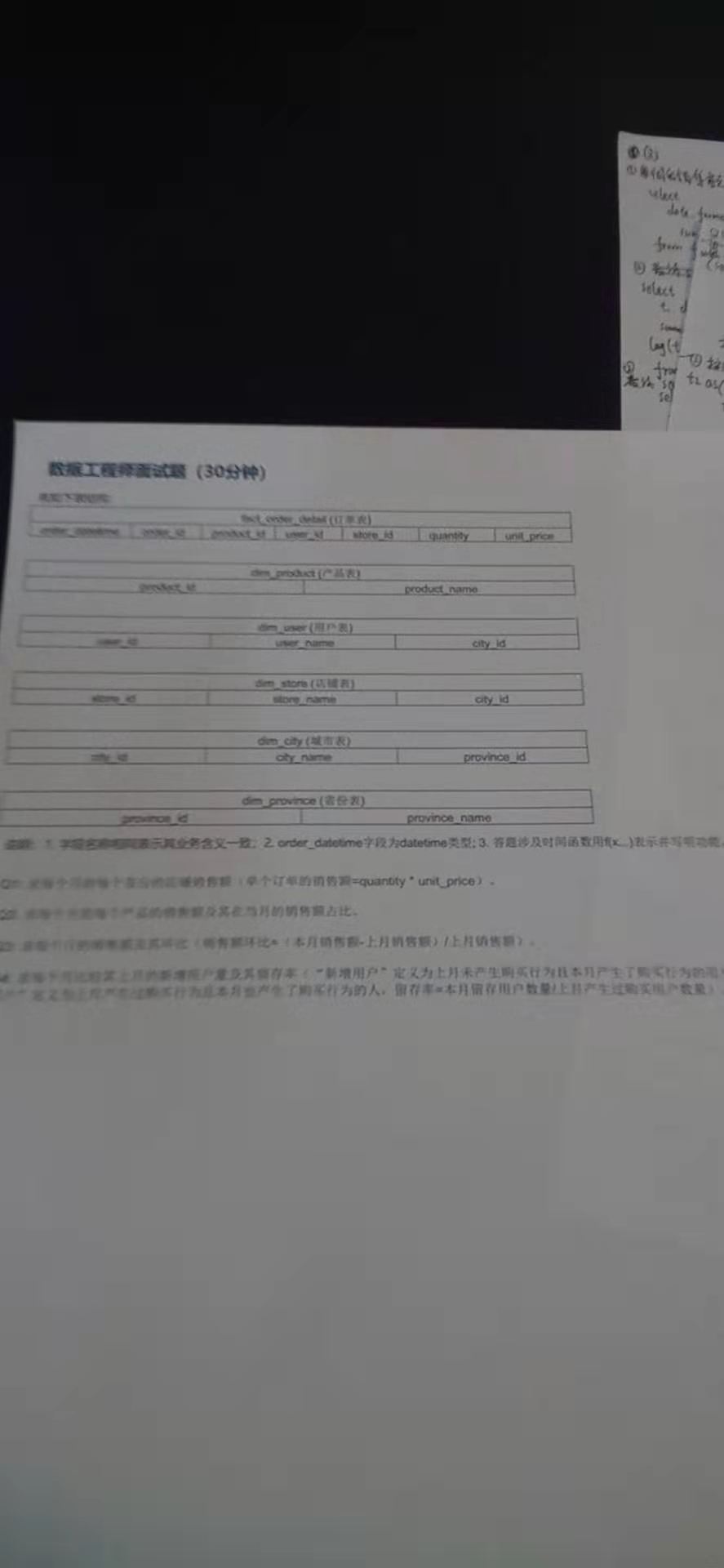
1. 如何保证数仓中数据的准确性（hive在处理数仓的过程，好多逻辑，在处理过程中会不会出现误差，是怎么处理的）
2. 数仓怎么去掉错误数据（异常），在Mysql中将数据写错了（原本是1000，人为的改成了1001），错误数据应该怎么处理
3. 使用sqoop导入Hdfs,但是业务库物理删除了一点（删掉了一条不需要的订单），相邻层之间进行监控
4. cannel可以识别delete(删除语句)语句
5. 什么情况下使用拉链表

# 二十九、博彦科技

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

# 三十、驰骛信息科技有限公司

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（我答MR写入到指定reduce，他问我hql怎么实现，spark怎么实现，我说不知道，请教他，他也没说个啥）
3. 如何判断计算是否正确？
4. 维度建模（我答维度建模的4步，他说这是比较简单的，然后说了一堆深奥的说根据相应的业务过程，没有表格啥的，就知道怎么建模，神经兮兮的）



1. 这不就是尚硅谷的数仓原题吗？

# 三十一、精锐教育集团

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？（背压，参数记得吗？）
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？拓展（在布隆过滤器前提下，要知道具体是谁活跃过？设计bitmap的具体使用）
8. kylin怎么用的？
9. ES的写入过快问题怎么办？（他说要涉及到异步写入的场景，我没见过）
10. 场景题：实时场景的是什么什么率怎么做？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？我只回答了phoenix使用二级索引的时候？

# 三十二、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？（肯定不是，是在mysql里面）
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？（废话，都是我）
6. 你们dwd维度建模工作是怎么分配的？按照业务线去分配

# 其他

1. 你们这边数据量有多少呢，就是从ods到ads层一共多少数据量？
2. 拉链表如果有一天的数据错了，比如说到12月15号，但是发现11月10号拉链的数据错了，导致后续拉链的结果都错了，这个应该怎么修正？
3. 真实项目中，spark的core和task数量该如何设置？

task数量，至少设置成和spark application的cpu core总数量一样（但这是最理想的情况，400 cpu core，分配了400个task，同时跑，不可能差不多同时跑完）

spark官方推荐：task数量，设置为spark application的cpu core总数量的2~3倍。怎么解释呢？因为如果400 cpu core，分配了400个task，同时跑，20个先跑完了，380个还在运行，这时，就有20个cpu core 空闲出来了，就导致了浪费。那如果task数量设置为cpu core总数的2-3倍，那么一个task跑完后，另一个task立马补上来，这就避免了cpu core空闲，提高spark作业速度。

1. 如何设置并行度？

SparkConf sparkConf = new SparkConf().set(“spark.default.parallelism”, “800”);

1. java的集合了解多少?
2. scala的集合了解多少?
3. scala集合里面是怎么扩容的?
4. scala常见的算子?
5. scala中reduce by和group by有什么区别?
6. reduce by和group by 返回什么类型的值,传的是什么参数?
7. foreache和coleasce什么区别?
8. 并发读取mysql需要关注那几个参数?(sparkcontext去读去连接)
9. spark共享变量?
10. 项目中怎么用的背压机制?
11. Flink的key By和Spark的 group by有什么区别?
12. Flink的watermark什么时候去触发计算?
13. 消息超过watermark的时间会丢失数据吗?
14. Kafka数据挤压怎么办?
15. Kafka的分区分配策略?
16. 为什么不用rabitMQ要用kafka?
17. mysql中的数据是全部导入还是增量导入到hdfs?
18. java的集合了解多少?
19. scala的集合了解多少?
20. scala集合里面是怎么扩容的?
21. scala常见的算子?
22. scala中reduce by和group by有什么区别?
23. reduce by和group by 返回什么类型的值,传的是什么参数?
24. foreache和coleasce什么区别?
25. 并发读取mysql需要关注那几个参数?(sparkcontext去读去连接)
26. spark共享变量?
27. 项目中怎么用的背压机制?
28. Flink的key By和Spark的 group by有什么区别?
29. Flink的watermark什么时候去触发计算?
30. 消息超过watermark的时间会丢失数据吗?
31. Kafka数据挤压怎么办?
32. Kafka的分区分配策略?
33. 为什么不用rabitMQ要用kafka?
34. mysql中的数据是全部导入还是增量导入到hdfs?
35. 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？
36. 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？
37. 开窗函数有哪些？
38. 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？
39. 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
40. OLAP引擎用过哪些？
41. 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？
42. 拉链表有什么缺点？拉链表有哪些字段必须要有的？
43. 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？
44. 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
45. 说一下布隆过滤器怎么实现的，数据结构是什么
46. 业务中HBASE的RowKey怎么设计的
47. watermark处理迟到数据，怎么实现的
48. redis的数据类型，怎么用来去重的，存储的是什么数据
49. 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

# 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？

## 二．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？（广播状态Broadcast state）

## 反压（背压，数据积压）

<https://blog.csdn.net/weixin_49060400/article/details/112253767>

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？

## 四．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.(我说的是手动维护Kafka的offset实现一致性消费的问题)
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

## 五．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？（1.11版本）

<https://blog.csdn.net/nazeniwaresakini/article/details/107954076>

总结：异步快照

## 六．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

## 七．双流join

1. Spark和flink的双流join的底层原理
2. A表left join B表
3. A表数据来了，B没来
4. A表数据来了，B在规定时间内到
5. 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

## 八．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？

[<https://blog.csdn.net/u012151684/article/details/109439590>](https://blog.csdn.net/u012151684/article/details/109439590)

1. flink的序列化机制？

<https://blog.csdn.net/xiaopeigen/article/details/108318530>

关键字：TypeInformation、TypeInfo、TypeInfoFactory、kryo

1. 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
2. Rich Functions与Functions区别
3. flink里面异步IO代码具体怎么写的，每一步具体描述出来