2022 年上海大数据面试题汇总

（作者：尚硅谷研究院）

版本：V1.0

第1章 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]

（2）kafka 如果创建大量的 topic，对 kafak 会有什么影响？

（3）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？

（4）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？

（5）谈一下 flink 如何保证精准一次性？

（6）barrier 对齐会有什么危害？

（7）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？

（8）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？

第2章 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？

第3章 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

第4章 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，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

第5章 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？

第6章 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？

第7章 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

第8章 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

第9章 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

第10章 花旗

10.1 应聘者一

（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

10.2 应聘者二

（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

第11章 花旗一面（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？（狂喜）

（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？

第12章 浦发银行甲方

（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

第13章 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

第14章 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？

第15章 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

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。

第16章 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

第17章 旺旺集团

17.1 应聘者一

（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

17.2 应聘者二

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）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？

第18章 微盟

18.1 应聘者一

（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

18.2 应聘者二

（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？

18.3 应聘者三

（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？

第19章 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

第20章 蚂蚁

20.1 应聘者一

（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？

20.2 应聘者二

（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

第21章 蚂蚁

（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？

第22章 趣头条

22.1 应聘者一

（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？

22.2 应聘者二

（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？

第23章 得物

23.1 应聘者一

（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

23.2 应聘者二

23.2.1 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

23.2.2 二面

介绍用 flink 做的项目 大概十分钟等他问

介绍完项目就是业务场景题

（1）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

（2）开窗 1 小时 每隔十分钟输出一次计算 怎么实现

（3）你的实现 如果半夜 4 点没有数据来 会怎么样？

（4）看过哪些源码 是你自己看的那种

（5）除了做 profile 做火焰图 还有什么方法 快速定位算子问题

23.3 应聘者三

电面 问了

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。

23.4 应聘者四

（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？（他想问的更多的是如何通过 flink 本身去优化）

第24章 永辉云创

（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？

第25章 樊登读书

25.1 应聘者一

面试送两个月的会员卡

聊天形式，实时数仓是什么架构？其他的问业务，没问什么技术方面的问题，相互交流！

25.2 应聘者二

（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

第26章 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

第27章 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）

第28章 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

第29章 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？

第30章 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

第31章 晨光

（面试官一再强调他们是大公司，数据量很大，业务很复杂，让你有个心理准备）

（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。

第32章 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？

第33章 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

第34章 上海 永辉

34.1 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

34.2 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

第35章 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

第36章 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？

第37章 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

第38章 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

第39章 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

第40章 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？

第41章 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

第42章 金大师

42.1 应聘者一

（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

42.2 应聘者二

（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

42.3 应聘者三

（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

第43章 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

第44章 赢时胜

44.1 应聘者一

（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?(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?

44.2 应聘者二

（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

第45章 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

第46章 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?

第47章 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

第48章 软通

48.1 应聘者一

（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

48.2 应聘者二

（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

48.3 应聘者三

（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

48.4 应聘者四

（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。

48.5 应聘者五

（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

第49章 信也科技

49.1 应聘者一

（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？

49.2 应聘者二

（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

49.3 应聘者三

（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？

第50章 中软

50.1 应聘者一

50.1.1 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

50.1.2 华腾

（1）做了哪些业务

（2）hive 和 mysql 的区别

50.2 应聘者二

（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？

（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？

50.3 应聘者三

50.3.1 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

50.4 应聘者四

50.4.1 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

50.5 应聘者五

（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

第51章 博彦科技

51.1 应聘者一

（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

51.1.1 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

51.2 应聘者二

51.2.1 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？

51.3 应聘者三

（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

51.4 应聘者四

（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？

第52章 驰骛信息科技有限公司

52.1 应聘者一

（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

（5）这不就是尚硅谷的数仓原题吗？

52.2 应聘者二

（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？

52.3 应聘者三

（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？

第53章 精锐教育集团

53.1 应聘者一

（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？

53.2 应聘者二

（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

第54章 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

第55章 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？

第56章 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

第57章 深圳市领星网络科技有限公司

57.1 笔试

57.1.1 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+

57.1.2 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

57.1.3 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

57.1.4 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

57.1.5 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

第58章 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?

第59章 平安银行（中国平安）

59.1 应聘者一

（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

59.2 应聘者二

（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

第60章 天阳科技

（1）你们 kafka 数据量怎么这么大

（2）你们的 kafka 设多少 topic？

第61章 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

第62章 上海小砖块网络科技有限公司

62.1 应聘者一

（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

62.2 应聘者二

（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？

第63章 上海伯俊软件科技

63.1 应聘者一

（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

63.2 应聘者二

（1）开窗函数

（2）sparkstreaming 的乱序处理

第64章 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

第65章 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

第66章 视若飞

（1）kafka 结构

（2）hive 常用函数

（3）redis 穿透

第67章 维信金科

67.1 应聘者一

（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

67.2 应聘者二

（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

第68章 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

第69章 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

第70章 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

第71章 数禾科技

71.1 应聘者一

（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？

71.2 应聘者二

顺序记不清了。

（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

两道现场编程题，只要说出思路，写出伪代码就行

（1）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）

（2）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）

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。

第72章 爱回收

72.1 应聘者一

（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？

72.2 应聘者二

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
......
```

（3）Hive 优化与导致数据倾斜用到了哪些算子？

（4）开放的题目，写一条 hql 发现很久不出结果，会如何发现问题？

基本上你会手写 hql，就让你等第二面，技术大佬面

（1）数仓建模？分层？

（2）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？

（3）Flink 数据不丢失的三重保障

（4）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

第73章 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？

第74章 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

第75章 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

第76章 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？

（2）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？

第77章 滔博

77.1 应聘者一

（1）一轮线下面，人事与产品面

（2）介绍自己

（3）数据量多少

（4）客单价有点低？

（5）离线数仓如何分层搭建的，层与层分别干了哪些事？

（6）我问的问题：

（7）滔博在各大平台有店铺，里面的很多数据要怎么拿到？

（8）用到了哪些技术栈?

二轮电话面，技术面：

（9）Flink 数仓如何分层？

（10）Clickhouse 有哪些引擎，你们用的哪种引擎？

77.2 应聘者二

（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

第78章 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

第79章 波克城市

79.1 应聘者一

（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？

79.2 应聘者二

（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

第80章 XTranfer

80.1 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！

第81章 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

第82章 美团外包

（1）实时方面

a) 主攻哪个方向

b) 我说的实时

c) 具体介绍一下具体做了哪些工作。

d) 为什么要做 sparkstreaming 到 Flink 的转化。

e) 在什么场景下需要这么高的实时性。

f) 既然是开窗为什么一定要转 FLink。

g) 遇到 SparkStreaming 不太能解决的问题。

h) 必需要手动维护 offset 吗？

i) 遇到 Flink 不太能解决的问题。

j) 实时指标出来后的应用场景。

k) 预警是怎么做到的.预警的条件。

（2）数仓方面:

a) 当初建模的时候应用场景是什么样的。

b) 建模的流程

c) 都有哪些数据同步到数仓里面

d) 对这些表有过什么分类吗。

e) 哪些表是相应的同步策略。

f) 跨天支付数据是怎么处理的。

g) 用户表为什么一定是拉链表。

h) 数仓的分层。

i) 如何找出来用户的一天的行为轨迹。

j) dws 和 dwt 的宽表都有哪些,并且都是什么!!!

k) 出口对应的指标。

l) 你们是怎么保证数据质量的。

m) 数据质量监控的角度

（3）SQL 题（很简单）

a) 外卖的配送 ID

b) 外卖员的 ID

c) 订单配送的 City

d) 时间的 CT

e) 一整年中，每个月每个城市订单量 Top10。

（4）其他

a) 为什么考虑换一份工作。

b) 离线和实时更偏向哪些，为什么？

c) 工作后做的最有成就感的一件事是什么。

d) flink，ck 机制，内存管理，出现反压怎么处理的？

e) kylin 如何直接构建 cube？

第83章 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

第84章 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

第85章 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

第86章 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？

第87章 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

第88章 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

第89章 博奥特

89.1 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;
```

89.2 问题清单

（1）ETL

a) 是否在项目中使用过 Informatica？

b) 如果(1)结果为否，请列举项目中使用的其他 ETL 工具

c) 是否有处理拉链表经验？

d) 是否有 ETL 工具调优经验？

（2）SQL

a) 是否在项目中使用过 Oracle 编写存储过程？

b) 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。

c) 是否有分析数据库执行计划能力？

d) 是否有亿级记录查询处理经验？

（3）Linux

a) 是否有独立编写 shell 脚本能力？

b) 是否知道 Linux 文件目录权限管理机制？

（4）沟通和管理

a) 请列举在以往项目中担任过的角色？

b) 是否带领过团队？团队人数为多少？

c) 请列举一个项目中遇到的困难及解决方

第90章 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

第91章 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

第92章 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;
```

（3）Spark 任务提交 yarn-cluster 模式跟 yarn-client 模式的区别

（4）以往工作中 HQL 的优化

（5）根据下面给到的成绩表数据写出对应的 HQL 语句

name	lesson	goal
张三	语文	90
张三	数学	80
李四	英语	59
...	...	...

a) 找出语数英每门课前三名的学生

b) 单科分数有低于 80 分的学生的总分排名

第93章 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

第94章 琚宠科技

94.1 应聘者 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

94.2 应聘者 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？

第95章 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

第96章 珍岛集团

96.1 应聘者一

（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

96.2 应聘者二

（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

96.3 应聘者三

（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

第97章 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

第98章 序彦泽游戏公司

98.1 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

98.2 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？

第99章 晶格数字

（1）讲讲数仓分层

（2）雪花模型和星星模型的区别

第100章 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

第101章 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

第102章 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

第103章 银科控股

（1）为什么选用 kafka 作为消息中间键（说一下 kafka 的优点）

（2）kafka 是如何保证快速的（当时没回答上来，感觉是因为它只作为消息中间键，速度是根据生产者和消费者决定的）

（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。底层不知道。就说只是简单使用，没做深入研究

（5）select * from 表名 group by 执行顺序，from--where--group by--having--select--order by,

第104章 九章极云

（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？

第105章 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？

第106章 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？

第107章 国金证券

107.1 应聘者一

（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?

107.2 应聘者二

（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

第108章 聚天时代

（1）写过的比较难的 sql

（2）**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

（3）**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

（4）**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?

（5）写一个脚本,杀掉 yarn 上正在运行的程序

（6）说一下 flink 实时数仓的架构流程?整个成型之后的数据流

第109章 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？

第110章 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？

第111章 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

第112章 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?

第113章 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

第114章 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

第115章 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

第116章 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？

第117章 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。

第118章 吉祥航空

118.1 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

118.2 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

第119章 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

第120章 池鹭

120.1 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

120.2 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

第121章 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

第122章 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

第123章 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。

第124章 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

第125章 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

第126章 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

第127章 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

第128章 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

第129章 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？

第130章 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

第131章 上海 天阳科技

（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

第132章 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.

第133章 上海 玛驷众

开始 3 道 sql 5 分钟 ac

133.1 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

第134章 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？

第135章 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

第136章 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

第137章 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？

第138章 vivo 外包

（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?

第139章 上海向晨

（1）怎么在海豚调度器中设置并发参数？

第140章 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

第141章 南京 vivo 外包实时一面

（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

第142章 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

第143章 上海 博彦科技（做一个新加坡银行的外包项目）

（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

第144章 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?

第145章 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

第146章 神州信息-外包

（1）一百四十六、神州信息-外包

（2）项目中遇到难题，及如何解决

（3）spark 架构及提交流程

（4）kafka 分区与副本如何设置

（5）kafka 如何保证数据有序

（6）kafka 事务如何实现，底层是如何做的

（7）kafka 新增消费者如何执行再平衡

（8）flink checkpoint 机制

（9）flink 状态后端

（10）二阶段提交

（11）为什么选择 clickhouse

（12）说一下你了解的 HBase

（13）clickhouse 和 HBase 的区别

（14）项目设计到落地具体流程及你做了哪些

（15）场景题：HBase 中如何设计表，及关系型数据库中如何设计

（16）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

第147章 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

第148章 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

第149章 上海 润和

149.1 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？

149.2 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）

第150章 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？

第151章 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？

第152章 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

第153章 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,

第154章 南京 vivo 外包离线

（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

第155章 叶子科技 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

第156章 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

第157章 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? 不会！

（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？

第158章 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? 不会！

（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（ack）

（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？

第159章 谷幽（海隆软件）

159.1 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

159.2 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

第160章 孤波

160.1 公司里架构师给做的面试

（1）3 道手写 Java 题

a) 写一个 list，往里放几个 Integar，把大于 5 的数剔除

（记住要输出，要自己测试一下，被追问难道写完代码，自己不测吗）

b) 写两个 String，特别长，转换不了 Integar，如何进行相加操作？

c) 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

a) 游戏分区，求每个区用户排名 Top 5？

b) 取所有用户排名 Top5？

c) 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？

160.2 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？

第161章 春秋航空一面

（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

第162章 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

第163章 四川星点网络（仙谭酒业）

163.1 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

第164章 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

第165章 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，

第166章 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：

第167章 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

第168章 嘉环科技

（1）自我介绍

（2）Hive（Hadoop 相关）

a) 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？

b) 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

c) Hive 的小文件问题遇到过吗，怎么处理的？

d) Hive 支持的文件格式有哪些？你们是怎么做的，为什么选这种？

e) 集群的 NameNode 起不起来怎么办？一般是怎么排查的？

f) 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？

g) 用过哪些窗口函数

h) Hive 的数据倾斜是怎么处理的

i) 执行计划你们会用吗，一般怎么用（什么时候会用）？

j) 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？

k) 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？

l) Yarn 的提交流程

m) hive 的动态分区功能有没有用过？为什么要用这个功能？

（3）Flink

a) Flink 的分区策略是什么？

b) Flink 的日志文件一般存储在哪儿

（4）Kafka

a) Kafka 是怎么保证不丢不重？

（5）Hbase

a) Hbase 的写流程是什么样的？

b) Rowkey 设计是怎么设计的？

（6）Spark

a) spark 的宽窄依赖是什么？

（7）Java

a) 抽象类和接口的区别

b) 我看你说用了异步 io，那新建线程有哪几种方式呢？

（8）Linux

a) 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？

b) 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

c) var/usr 分别是存储什么文件的？

d) linux 系统的配置文件一般存放在哪个文件夹？

（9）MySQL

a) 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）

b) MySQL 你们用的版本是什么？

c) 知道 min.ux 这个函数吗？知道是什么吗？那 lag 函数呢？

第169章 精·Flink 面试总结

169.1 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？

169.2 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？

169.3 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？

169.4 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

169.5 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？

169.6 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

169.7 双流 join

（1）Spark 和 flink 的双流 join 的底层原理

（2）A 表 left join B 表

a) A 表数据来了，B 没来

b) 2）A 表数据来了，B 在规定时间内到

c) 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

169.8 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

169.9 花旗面试题

（据说有了这个可以直接入职）

169.9.1 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

169.9.2 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

169.9.3 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