2022 年北京大数据面试题汇总

（作者：尚硅谷研究院）

版本：V1.0

第1章 字节跳动

1.1 面 1

（1）自我介绍

（2）做实时用 flink ？用 sql 还是 api

（3）flink 反压，如何排查

（4）算子链断链有哪些方式，从代码层面说

（5）在排查 flink 问题的时候有没有关注过 flink 的 metrics 比如 flink 自带的 metrics

（6）checkpoint 的流程，说说理解

（7）端到端一致性，如果挂掉了怎么办

（8）flink 写 hive 是否能保证端到端一致性？怎么保证幂等写入，overwrite 不是会覆盖吗？不会丢数据吗？

（9）遇到的数据倾斜的场景和对应解决方案

（10）场景题，字段：uid 用户 id，vid 视频 id，timestamp 时间戳，计算每个视频的 uv，一个用户可以看多次视频，由于有热门视频，会有数据倾斜，怎么用 flink 应用程序实现，追问，细问加随机数怎么加，继续追问更复杂的场景状态大怎么办。Split Distinct 的底层原理是什么？Id-key 比较分散，

补充：如果 rocksDB 放不开了怎么办，状态编程去重智能去掉 3%左右，性能不太好，有没有更好的解决方案？

（11）常见的缩减状态的方式有哪些？

（12）旁路缓存是否能保证端到端一致性

（13）Clickhouse 的 ReplacingMergeTree 一定保证不重复吗？final 和强制合并都不是可控的，还有什么好一点的方式吗？

（14）Clickhouse 的 ReplacingMergeTree 中依赖指定 Order By （索引）和 版本的，如果不指定版本，在同一分片写入不同的数据，写入的时候保留那条数据？

（15）Clickhouse 有分片表和什么表的感念（没听清），对于 ReplacingMergeTree 如果我直接轮训写的话，是否能达到去重效果吗？

（16）介绍实时数仓项目？最印象深刻的点是什么？

（17）大状态如何能够复用？增量检查点启动恢复的时间是很久的，业务上是否能接受？

（18）如何理解实时离线一体架构？

（19）市面常见数据湖的优劣对比了解吗？

1.2 面 2

（1）怎么保证 flink 的一致性，两阶段提交，怎么保证数据不重复，最后解决了吗。

（2）clickhouse 怎么保证幂等性

（3）sink 端的事务输出 输出到哪些组件

（4）kafka 和 clickhouse 各自的应用方向都是哪些

（5）clickhouse 为什么查询快

（6）clickhouse 适应哪些业务场景

（7）一致性数据只是做到了数据的不丢失，怎么保证数据的准确性？

（8）你认为怎么去做才能保证数据的正确性

（9）实时数据最后是都有谁去使用了

（10）数仓怎么去做分层的，横向怎么做分数据域的

（11）你们的数据域是怎么划分的，都有哪些，详细说每个域具体是怎么去定义的，都怎么去定义这些数据范围

（12）整体分 5 层的优缺点都是什么

（13）数仓建模的方法都有哪些

（14）维度模型各自的特点和不同有哪些

（15）事实表的设计方法是什么

（16）事实表分为哪些

（17）对于缓慢变化的数据事实表采用什么方式存储

（18）拉链表的好处和缺点都是什么

（19）sparkrdd 怎么体现他的弹性，和特征

（20）transform 和 action 算子有什么不同

（21）spark 的 shuffer 有哪些，讲讲各自的特点

（22）写 sql 电脑上直接敲：

用户连续日活表 usr_login，表中字段：userid，dt，目前该表存有 1-30 号内的数据，请输出 30 天内，连续登录三天及以上的 userid，以及用户初始活跃的第一天

1.3 面 3

（1）大数据这一段工作经验中你做的是那部分东西呢？你觉得最拿手的一个方向？

（2）你们是做什么业务的？

（3）你们有事业部概念吗？你是在哪个事业部？你负责的业务范围是什么？你在里面的角色是啥？

（4）实时的核心指标有什么？

（5）M1 到 M7 核心指标是什么？

（6）数据在 ODS 层，你整个实时数仓是怎么流转的，怎么计算加工的？

（7）ODS 层的数据都用了什么数据？数据源都用的什么数据？数据里包含什么？DWD 业务明细都有什么数据？DWD 中数据域都有哪些，主题域都有哪些，怎么分的，为什么这么分？

（8）DWS 层都有什么粒度的表？

（9）DWS 层提到旁路缓存和异步 IO，着重讲讲异步 IO？异步 IO 代码逻辑，用什么算子，哪些方法，里面怎么做的，和 HBase 的交互？

（10）实时维表是否是实时的？Redis 和 Hbase 的一致性怎么保证？Redis 的失效时间是多久？

（11）维表数据改变更新 Hbase 并删除 redis 的操作是原子的吗？线上的 flink 会访问数据吗？不是原子性如果产生错误数据怎么修复？

（12）你们线上容错机制都有哪些？营收这部分你们怎么做容错？

（13）类似金额，营收类统计，金银分析类东西，假如出错了，怎么做修复以及线上容错？

（14）容错方案，可以从哪几个角度来想？

（15）整个流式数据仓库的稳定性和质量怎么保证？数据怎么保证正确并且容灾？

（16）你们有实时仓库的数据监控和任务状态的运行监控吗？比如任务健康度以及数据质量相关的各个维度的监控以及异常的告警，有没有，怎么做的，没有的话，知道怎么

做吗？

（17）一个算法题，用对列实现。

1.4 面 4

（1）自我介绍

（2）问了问数据量

（3）介绍一个你参与最多的项目

（4）hdfs 读写流程

（5）hdfs 的 shuffle/shuffle 过程有几次排序

（6）kafka 乱序问题

（7）kafka 的 acks

（8）kafka 的 isr

（9）用户行为数据怎么划分的主题

（10）业务数据有哪些信息和指标

（11）实时用的 flink 吗？为什么不用 spark？

（12）flink 的检查点怎么实现

（13）java 和 python 了解吗，主要是 hive 吗？

（14）连续两天登录 SQL

1.5 面 5

（1）自我介绍

（2）讲一个自己比较熟悉的系统

（3）维度数据具体是怎么跟事实数据进行关联的

（4）dws 层写出数据到 clickhouse 是多长时间一次

（5）怎么验证实时数仓指标的正确性

（6）hdfs 的读写流程，读写过程中有 datanode 节点挂掉怎么办

（7）你们那些指标是使用 api 写的，那些是通过 sql 写的　为什么要分开

1.6 面 6

（1）leetcode 162 要求二分法实现

（2）kafka 高效读写

（3）项目整体如何控制重复数据，哪里用的事务，哪里用的幂等

（4）flink barrier 的机制，如果换你来实现，如何优化对齐式 barrier 带来的背压问题

（5）flink 的内存模型

（6）java 的内存模型

（7）异步关联纬度，为啥加 redis，不加 redis 可以不，在应用程序里用第三方缓存接口是否更好

（8）java 中 array 和 list 的区别

（9）hadoop 为什么默认 block128m

（10）hashmap 的实现

1.7 面 7

（1）Clickhouse 和 elasticsearch 的区别，应用场景的区别，底层的区别

（2）arraylist 和 linklist 有什么区别，读写上的区别

（3）创建多线程的几种方式

（4）Udf,udaf,udtf,的区别，各举几个例子

（5）kafka 数据是有序的吗？

（6）kafka 应答级别

（7）谈一下 isr

（8）flink 怎么保证数据不丢失

（9）flink checkpoint 底层是通过什么实现的

（10）手写连续两天登陆，手写连续 7 天登陆

第2章 华为

（1）hbase 的 rowkey 设计原则

（2）讲一下 kafka

（3）讲一下 zookeeper 以及选举机制

（4）spark 和 hadoop 的异同

（5）讲一下 mapreduce

（6）讲一下为什么离职

（7）说一下 spark 的算子

（8）spark 的行动算子和其他算子有什么不同

（9）map 和 mappartiton 的区别

（10）udf udtf 用过吗

（11）map 算子和 flatmap 的用法

第3章 新华三

（1）介绍最近做过的项目

（2）简单问了问架构设计

（3）java 笔试题（15 道单选，几道问答，3 道程序）

（4）yarn 提交流程

（5）项目中遇到过哪些问题，是怎么解决的

（6）数据倾斜问题

第4章 美团

（1）自我介绍

（2）你们维表有多大量？是否只是点查，不会范围查询或者多维度组合查询吗？

（3）有可能出现 redis 容量超过上限的情况吗？

（4）新增数据源或者数据格式发生变更，会停掉服务吗？

（5）大状态怎么解决？

（6）检查点怎么做对齐？

（7）两个流 join 时间差别比较大的化会有什么影响？

（8）实时数仓是否做过监控，稳定性的考虑？

（9）反压怎么解决，数据倾斜怎么解决？

（10）怎么计算程序所需要的资源？

（11）哪里会用到 zookeeper？

（12）kafka 的 isr 和 ack 是什么？

（13）我有 5 个副本，isr 只有一个会有什么影响？

（14）编程题：

数字字符串转化成 IP 地址

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

例如：

给出的字符串为"25525522135",

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

数据范围：字符串长度 0 \le n \le 120≤n≤12

要求：空间复杂度 O(n!)O(n!),时间复杂度 O(n!)O(n!)

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

示例 1

输入： "25525522135"

输出： ["255.255.22.135","255.255.221.35"]

示例 2

输入： "1111"

输出： ["1.1.1.1"]

示例 3

输入： "000256"

输出： []

第5章 百度

介绍项目

（1）gzip 压缩比 10%

（2）数据倾斜

（3）累积快照事实表怎么解决累积快照事实表

（4）spark 怎么获得一个 rdd

（5）datafream 可以转换成 rdd 吗？

（6）SMB join 原理

（7）广播变量原理

（8）spark 内存模型

（9）flink 内存模型

（10）shuffle 数据量比较大怎么调节

（11）flink 的 shuffle 分为几类

（12）Kafka 怎么保证数据不丢

（13）kafka 的 leader 和 follower 之间的数据同步

（14）leader 挂了怎么办，机制是什么

（15）XW 是啥？

（16）mr 的 shuffle

（17）mr 的 shuffle 优化

（18）算法已知数组，1，2，3，4，5，6，7

求所有两数之和等于 10，数字的下标。算法复杂度要低

第6章 携程

（1）介绍实时数仓

（2）维度数据为什么放到 Hbase

（3）dws 关联 hbase 维度表数据的意义是什么？最终输出什么？

（4）数据以什么样的形式、频率导入 Hbase

（5）热数据放入 redis，通过什么样的方式导入 redis

（6）dwd 做了什么事情

（7）flink 挂了会怎么样？

（8）flink 数据乱序问题

（9）用户画像上游下游

（10）用户画像有多少标签？用了多长时间？

（11）flink 用了多长时间？一年（太长了？）

（12）用户画像的标签用到什么地方？说到了风险评估，追问什么样的风险评估？

（13）AB 测试了解的多吗？实验周期多长？你觉得实验周期跟什么有关？面试官还给解释了一下与样本量有关

（14）用 flink 写一个从新浪微博 api 去取数据，数据为 posts（一条一条的数据间隙）需要做一个分词然后统计每一个词的次频，说的 flinksql，又追问 api 方式怎么实现，

（15）在 hive 建模是用宽表形式吗？

（16）建模有没有用建模工具？

第7章 小米

（1）自我介绍

（2）你们的数据量大概有多少？

（3）实时数仓同一个操作产生的数据前端埋点的和第三方的数据时间戳不一致关联时怎么解决？

（4）redis 和 Hbase 的数据一致性怎么保证？变更高吗？

（5）你都用什么编程语言？我说 java，场景题就来了

场景题：

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

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

currentAds = [13,25,32,11]

candidateAds = [12,24,8,32]

candidateAds = [24,32,8,12]

第8章 58 同城

（1）自我介绍

（2）离线数仓介绍，你是具体负责那一部分工作，你在工作中所认为的一些疑难点讲一下

（3）为什么分 5 层，对分 5 层的理解

（4）为什么用 flume，为什么不用 flink 同步

（5）7 天连续登录 3 天

（6）小文件怎么解决

（7）零点漂移怎么解决

（8）shuffle 里几次排序

（9）udf，udaf，udtf 用过吗

（10）实时：为什么用 clickhouse，为什么 clickhouse 聚合处理快

（11）反压问题讲一下，怎么处理

第9章 快手

9.1 一面

（1）自我介绍，为什么离职，有 offer 吗，想做哪一块

（2）你们有几个人，分工怎么样，都和哪些人对接，数据量多大？

（3）你们平常怎么和其他产品和业务人员打交道的？如果没有产品你怎么对接业务人员？

（4）如果一个陌生领域你怎么快速熟悉业务？

（5）你们做的数据产生什么价值怎么评估？

（6）数据准确性的保障是怎么做的？是用的工具还是自研的？用什么呈现

（7）离线实时的模型设计？星形模型和雪花模型的区别？

（8）有没有看过直播，如果你来设计直播的模型，你会怎么设计数仓？

（9）MR 的 shuffle 流程

（10）快排原理？最好和最坏的时间复杂度

（11）reduce 个数是由什么决定的？

$$\min\left(ceil\left(2\times\frac{totalInputBytes}{bytesPerReducer}\right),maxReducers\right)$$

（12）不指定分区默认是几个分区？

（13）udf，udaf，udtf 有什么区别？

（14）两个大表 join 产生数据倾斜怎么优化？

（15）flink 怎么保证端到端的一致性

（16）flink 反压是怎么样的，怎么优化

（17）clickhouse 解决什么问题？

（18）SQL 题：同时在线人数

（19）SQL 题：连续登录四天有哪些人，连续登录几天？

（20）数仓这块怎么挖掘更大的数据价值？怎么应用？

9.2 二面

（1）介绍他们的诉求

（2）介绍我这边的项目

（3）能不能讲讲某个业务方向上是数据是解决什么问题，怎么发现的问题，又怎么解决的问题，怎么验证他的准确性，怎么去评价

（4）有没有技术上比较有挑战的项目讲一下

（5）了解客户端埋点吗？埋点这边的规范和设计咱们理解吗？怎么是好的数据埋点规范

（6）怎么是好的数据研发？

（7）咱们之前有哪些之前做的差一些，时间回溯又想要怎么提高的？

（8）如果产品的需求你觉得不靠谱，你怎么办？

（9）怎么带实习生？目标是什么？

第10章 度小满

（1）你们 Redis 用了几台服务器，我搞了 6 台

（2）Redis 中每个服务器存储的数据怎么划分的

（3）Redis 的 RDB 和 AOF

（4）MySQL 索引优化

（5）Flink 检查点机制

（6）二分查找算法

第11章 作业帮

（1）面试官介绍了一下他们公司要做的项目，偏离线；

（2）自我介绍；

（3）项目架构；

（4）问离线项目中你做了哪些事情；

（5）问离线数仓怎么可视化的；

（6）问 flink 的检查点一致性；

（7）sql 面试，连续 7 天内商家评分连续下降；

第12章 京东物流

（1）自我介绍

（2）怎么从老师转向开发？

（3）不是本专业，又做了老师，还学了开发，可以总结一下你的学习的心得体会吗？

（4）介绍离线数仓项目，你在项目中做了什么？

（5）各层建模参与吗？

（6）数据域是怎么划分的？

（7）hive sql 优化

（8）遇到过数据倾斜吗？

（9）用过拉链表吗？

（10）相比离线，你觉得实时的难点在哪？你觉得离线、实时你更擅长哪个？

第13章 闲徕互娱

（1）问了个 foreach 的累加 scala 代码

（2）kafka 系统主题了解吗

（3）消费者和消费者组的区别

（4）主题 主题域

（5）数据倾斜，mapjoin 如何实现的？skewindata 加随机数怎么实现的？

（6）hive 存储格式有哪些？为什么用 orc，不用 parquet？

（7）kafka 数据丢失，数据重复

（8）你们公司 80g 数据能遇到数据积压问题吗？

（9）表名：dau ，字段： user_id, day 需求：2022-08-01 活跃的用户在之后每一天的留存数

（10）连续两天登录

（11）一道行转列 sql

第14章 神州数码（电话一面）

14.1 面 1

（1）自我介绍

（2）分区表分桶表按什么分区，按什么分桶？ 按时间分区

（3）怎么数仓建模的？

（4）他追问没用什么拉链表什么的？ 就用户表用了

（5）增量表与全量表？

（6）讲一下 flink 和 kafka？

（7）经常用的是 hive 还是关系型数据库

（8）hive，索引 where 不到是怎么回事？

（9）是否离职，离职原因

（10）本科专业，为什么干这个？

（11）有什么要问我的？

14.2 面 2

（1）自我介绍

（2）问我是否会接口开发，会不会 js 开发

（3）hive 的内部表和外部表区别

（4）内部表的分区分桶有了解过吗？分桶一般使用什么字段来分效率高一点，分桶个数如何确定；分区表有哪几种分区表；使用哪种类型的分区效率比较高

（5）有没有用过 hive 索引，分区自带索引了解过吗？什么时候索引会失效？

（6）多条数据重复，怎么去重

（7）如果开发 SQL 脚本的时候，如果跑的很慢，首先怎么定位，然后怎么解决；其它层面的调优，和仅限 SQL 层面的调优

（8）SQL 的表与表关联的底层运算机制

（9）介绍一下数仓项目，怎么分层，主题划分怎么考虑的，数仓建模

（10）期望薪资多少

14.3 面 3：（二面）

（1）问离职原因

（2）一面我提过数据湖，这次问我的理解

（3）能不能考虑出差，是否排斥和各个业务方的沟通

（4）给我介绍了他们公司发展历程和目前规划

第15章 中国邮政

（1）项目用的哪些组件

（2）整体架构

（3）怎么建模

（4）Hql 优化

（5）Flink 数据一致性

（6）CK 替换 replacingmergetree 引擎 处理数据去重的延迟问题？

（7）使用框架时候怎么处理漏洞，类似一些网络安全问题

（8）框架的 bug 都有啥，怎么处理的

第16章 人民在线

16.1 面 1

这家公司有自己的集群，想招人做维护，希望有 Java 功底和写 flinksql

（1）自我介绍，然后详细讲一个项目

（2）拉链表，零点漂移，

（3）行转列，列转行详细解释

（4）说说 springboot 常用注解

（5）讲讲 flink 的窗口，和 join 方式

（6）列举多线程安全和不安全的几个类

（7）多线程怎么实现的，用的那几个方法

（8）spark reducebykey groupbykey 区别

（9）集群规模，每个节点配置，数据量

16.2 面 2 一面

（1）实时架构，离线架构

（2）hbase 行键底层如何排序

（3）flink join 大概介绍一下，几种

（4）hashmap 底层如何实现

（5）treeset 和 hashset 区别

（6）列举多线程安全和不安全的几个类

（7）多线程怎么实现的，用的那几个方法

（8）es 大概介绍下，有部署过吗

（9）flink 端到端一致性

（10）flume 遇到过什么问题

（11）hive 小文件怎么解决的

（12）spark reducebykey groupbykey 区别

（13）flink 遇到过什么问题

（14）flink 的提交方式

（15）hivesql 和 mysql 语法区别

（16）用到过什么压缩

（17）mysql 隔离机制

16.3 二面

（1）redis kv 类型的 为什么要有 hash

（2）redis 能实现 顺序消费吗

（3）redis 操作需要加锁吗 多个客户端同时使用 能保证原子操作一致性吗

（4）用过哪些锁

（5）flink 事件分为几种

（6）一个任务中可以既有事件窗口和处理时间窗口吗

（7）flink 一个任务需要的资源怎么计算

（8）Lambda 架构 优缺点 Kappa 架构 了解吗

（9）mysql 一条 sql 执行步骤

（10）hive 有索引吗

第17章 中国移动

（1）纬度建模和分层(他们也是分五层第一层就轻度聚合了)

（2）数据采集，问咱们为啥用这些组件(他们自己研发了一个组件)

（3）了解应用程序的开发流程吗？简单说一下

（4）数据治理有了解吗

第18章 嘀嗒出行

（1）自我介绍

（2）如何从老师跳到大数据开发的

（3）最近的重点项目，背景，做了什么，承担了什么，遇到问题是什么，困难是什么，挑战是什么，怎么解决的

（4）怎么保证 clickhouse 的数据一致性，优势是什么？

（5）flink 的并行度了解不？

（6）如果让你来管实时的所有项目有问题吗？

（7）sparkSQL 如何合并小文件？

（8）什么情况下会产生小文件？

（9）hivesql 的执行原理？

（10）我觉得你掌握的还可以，你是这些是工作中积累的，还是准备的？

（11）开发语言都用什么语言？

（12）clickhouse 写过自定义函数吗？flink 写过吗？

第19章 科大讯飞

（1）自我介绍

（2）项目履历

（3）你负责哪些工作

（4）参与建模了吗，你们模型怎么建立的

（5）技术架构，对 flink 的使用 （api,sql,cdc,cep）

（6）双流 jion 哪几种类型？

（7）两个队列，每个队列都是海量数据，用什么 join？

（8）用过哪些分布式存储框架？调优？

（9）Kafka 分区下线问题？

（10）服务开发？

（11）他介绍他项目

第20章 讯飞技术股份

人事面 离职原因

20.1 技术面

（1）了解 kafka 吗

（2）Yarn 用的那种模式

（3）Yarn 的调度器用的那种 ， 了解他们的区别吗

（4）工作的时候调过那些 yarn 的参数

（5）了解 spark 吗

（6）Flink 项目主要做了那些

（7）用过 sparkstreaming 吗

（8）工作时用过那些关系型数据库

（9）你对 mysql 了解吗

（10）Mysql 的索引，视图有了解吗

（11）你的维度是怎么确定的，能讲讲吗

第21章 合肥讯飞

（1）实时数仓的搭建过程？

（2）你们的实时元数据怎么管理的？

（3）你们有做主数据管理吗？就是里面最重要的 ET 实体是哪些？整个数仓最重要的实体是那些?

（4）指标体系建设是怎么做的？是产品经理直接梳理的吗？

（5）计算和加工表数据的时候主要用到的方式主要是 flink sql 吗还有别的吗？

（6）做过 CEP 规则引擎开发吗？

（7）对 flink 里面的状态使用过哪些？

（8）对 hbase 和 hive 这块做过调优吗，强调了下集群？

（9）遇到过热点问题吗？

（10）你在这个项目里面做了哪些部分？

（11）你们的技术选型是谁做的？

（12）你们的数据规模有多大？

（13）数据集成这部分是怎么实现的？

（14）后面的职业规划？

（15）你有哪些问题要问我？

第22章 我爱我家

22.1 初面：

（1）介绍自己

（2）介绍实时数仓

（3）开窗函数有哪些，排序那 3 个函数区别

（4）来公司后要学习 python

（5）你有什么想问的吗

22.2 主管面：

（1）数据迟到怎么处理

（2）场景：有 4 个字段分别为 城市 id，用户 id，访问页面，访问时间，需求：从 0 点到现在，最高访问人数的一个峰值，每一分钟访问最高的前 3 名用户，一分钟内实现，具体到用的什么算子

（3）你懂算法吗

（4）来公司后需要很强的学习能力，需要学习算法，公司是 flink 结合算法来用的，会有人教，但是需要你能学明白，你觉得你的学习能力怎么样

（5）你的学习态度怎么样

（6）你还有什么想问的吗

第23章 中国电信

（1）自我介绍

（2）数据采集，离线数仓 （组件的版本 hadoop2.x 和 3.x 的区别）

（3）hive 的元数据里面有什么？

（4）维度建模每一层都干了什么？

（5）数据规模

（6）hadoop 的详细加载顺序

（7）java arraylist 和 linklist 的区别

（8）flink sql 调优（状态优化）

（9）chlickhouse 的数据一致性 （实现不了 ， final ,optimized 实现）

第24章 中盾安信

这家主要做实时，应该也只会问实时

（1）自我介绍

（2）你们刚开始是使用了 sparkstreaming，那后期为什么要更换为 flink 呢，你们对实时性的要求很高吗

（3）数据量，在你们业务当中 flink 在高峰期最大能够处理的数据量能够达到多少，具体是多少条

（4）每一条日志数据的大小是多少，日志数据当中具体都有哪些字段

（5）如果上游 kafka 接收到的数据量很多，而下游的 flink 只能有一个并行度，flink 进

行一些 ETL 操作能够处理数据的极限大概是每秒多少条？（反复追问，他说官网上有对应的参数）

（6）你们在使用 redis 做旁路缓存的时候，redis 当中一共保存了多大的数据？冷数据和热数据是如何划分的，是否对数据打了标签？

（7）你们在使用 flink 对接 kafka 的时候是否遇到过什么问题，你们的 dwd 层是从 kafka 当中取数，然后再写回到 kafka 当中，那这样的话有没有遇到过 kafka 性能方面的问题，你们是怎么优化的

（8）是否针对于 kafka 集群进行过扩容和缩容？

（9）你是否了解过 kafka 都有哪些功能？不是底层原理层面的

（10）你们的 flume 搭了几台，每一台都是什么作用，是否针对于 flume 进行过一些配置

（11）你们的 flume 使用的是单点的吗，为什么不使用分布式的

（12）flume 的 source，channel 和 sink，如果在选择 channel 的时候既想提高传输效率，也不想丢数据，应该如何配置

（13）你对一些 nosql 数据库有过了解吗，都有哪些使用场景，展开说说

（14）向他提问

第25章 中信百信银行

（1）自我介绍

（2）在项目中遇到了哪些问题

（3）二进制怎么转十进制

（4）离线数仓是否做过应用层（应该就是做过哪些指标），具体怎么怎么实现

（5）Superset 具体怎么使用

（6）有什么问题问面试官的

第26章 计算所

（1）自我介绍

（2）讲一下项目

（3）项目详讲

（4）对那些框架比较熟悉

（5）项目中遇到过那些那些问题（数据倾斜、解决方案）

（6）py 会不会？

（7）写过哪些指标

（8）工作经验之类的问题

第27章 便利蜂

（1）你们宽表里面主要做了什么操作

（2）做了些什么指标

（3）有没有比较难的指标

（4）Hive 数据倾斜的现象和处理

（5）实时 dws 层做了什么操作

（6）维表的变化怎么传导到 hbase 和 redis

（7）flink 双流 join 的原理是什么

第28章 人谷科技

28.1 面 1

（1）自我介绍

（2）公司做什么的

（3）mysql 和 hive 区别

（4）hive 引擎为什么从 MR 换成 spark，区别是什么，为什么 spark 快

（5）hive 优化

（6）hive 数据倾斜

（7）数仓建模怎么建的，指标体系是你分析的嘛，怎么分层的

（8）ods，dwd，dim，dws，都干了什么，什么是维度数据，什么是事实数据

（9）ods 到 dwd 做了哪些处理，dwd，dws，dim 都有什么区别

（10）你有什么想问的

28.2 面 2

（1）自我介绍

（2）离线项目介绍

（3）项目中做过哪些调优？

（4）项目中都具体负责过什么？

（5）orderby 和 sortby？

（6）udf，udtf，udaf 怎么理解的？

（7）外部表和内部表区别？

（8）拉链表怎么理解，什么情况下会用，怎么实现？

（9）rank，desne_rank,row_number 区别？

（10）用过哪些系统函数？

28.3 面 3

（1）自我介绍

（2）离线项目介绍

（3）你负责了哪些主题能讲讲 dws 层你建了哪些表为什么要建这些表

（4）数据导入到 ods 层你们是怎么做的

（5）拉链表的使用场景以及大概讲解一下步骤

（6）问了 hive 的UDFUDTFUDAF函数区别

（7）hive 用过哪些内置函数

（8）你有什么想问的

28.4 面 4

（1）项目流程串讲

（2）HIVE 的优化，数据倾斜，小文件

（3）做过的一些指标和实现思路

（4）拉链表的实现思路

（5）数仓的建模过程

（6）常用的一些数据库

28.5 面 5

（1）Linux 查看文件命令

（2）HIVE 的优化，数据倾斜，小文件

（3）hive 和 mysql 的比较

（4）用过哪些系统函数

（5）hive 优化做过哪些

（6）建模中范式建模和维度建模的区别

（7）拉链表怎样实现的

（8）用过 shell 什么工具

（9）写过 shell 什么脚本

28.6 面 6

（1）自我介绍这家主要招离线

（2）讲一下宽表

（3）用哪些系统函数

（4）自定义函数三个的区别

（5）数据怎么导入从 HDFS 到 0ds 层(系统数据导入到 hive)

（6）HiveSQL 怎么变成 MR 程序的

（7）两张表关联join 的问题

（8）mysql 和 hive 的区别

（9）拉链表你们用吗？怎么实现的

（10）Linux 常用命令

（11）是刚刚辞职吗离职原因在北京吗疫苗打了吗现在薪资期望薪资

（12）面试通过什么时候到岗

（13）你还有什么要问的吗？问了解决什么问题数据量有多大

第29章 讯方科技

29.1 面 1

（1）spark 的 shuffle

（2）hadoop 的 shuffle

（3）Hadoop 的 shuffle 和 spark 的 shuffle 有什么区别

（4）怎么优化，把 shuffle 过程去掉可不可以，为什么

（5）ack 的机制

（6）flink 相关的遇到什么问题，反压发生在哪种情况，怎么解决

（7）hive 做过哪些优化

（8）你做过项目的哪部分，负责什么

（9）你会考虑后来能不能转运维吗

（10）这家公司就是会先让你主动自我介绍，然后说做过什么项目，项目中用过哪些组件，然后展开问一些

29.2 面 2

（1）先给我介绍他们公司的的情况，北京到处驻场，问我能不能接受

（2）让我自我介绍

（3）说一下我熟悉哪些技术栈

（4）hive 优化

（5）为什么 spark 比 mr 快

（6）spark 有相无环图解释，stage 划分，宽依赖和窄依赖划分；跨 stage 的 task 是否能并行

（7）hbase 的 rowkey 设计原则及热点问题

（8）组件是自己匹配的还是 CDH，组件之间的依赖关系：hive 依赖啥，kafka 依赖啥，hbase 依赖啥，spark 依赖啥

（9）kafka 角色和平时遇到的问题

（10）SparkStreaming 怎么维护 offset

（11）后面就让我问他一些问题，我简单问了几个问题，驻场、几个人驻场一个公司、公司的大数据这块架构等等

29.3 面 3

（1）先介绍自己做的

（2）yarn 的提交流程

（3）yarntask 失败后会怎么处理。

（4）spark 流程

（5）sparkshuffle 种类和区别

（6）spark 做过那些哪些优化

（7）kafka 架构

（8）broker 里有什么

（9）controller 的作用

（10）leader 挂了以后怎么选举

（11）isr 是怎么来的

（12）hbase 读写流程

（13）hbase 一般存几个列族，能存多个列族吗，为什么

（14）hbase 刷写时机

（15）kafka 如何保证数据不丢失

（16）还问了一个sparkclient和 cluster 的区别

29.4 面 4

（1）主要问沟通交流方面（项目有冲突，a 客户需要你去解决问题，你正在做 b 项目

（2）功能没测试，然后客户需要去上线，你会这么说

（3）技术问题：Linux 的权限命令，替换字符串命令，脚本定时执行

29.5 面 5

（1）yarn的 task 任务谁监控管理的挂了怎么办怎么调参数的

（2）hive优化hive 数据倾斜

（3）hdfs小文件怎么处理的

（4）spark SQL spark Spark Streaming 了解吗

（5）kafka架构（4 张图）数据积压怎么解决分区结点副本怎么设计的？

（6）Hbase架构，读写流程，一般存几个列族，能存多个列族吗，为什么，

（7）hbase 刷写时机

29.6 面 6：（二面）

（1）先自我介绍

（2）Linux 用过什么命令

（3）Linux 怎样查看一个文件前几行

（4）Hive 元数据存在哪里

（5）HBase 的二级索引

（6）离线数仓给他串讲一部分

（7）数仓建模的意义

第30章 柠檬微趣

（1）scala 语言的特质,闭包,模式匹配，变量，抽象

（2）spark 的 rdd 的特点

（3）精准一次性消费如何实现，如何手动维护

（4）为什么选用 direct，以及他的并行度是怎么回事

（5）场景题，如何在大量数据中快速抓取某类数据

（6）算法加数据结构加链表，没整理，听不懂也答不上来

第31章 椰子传媒

（1）公司很小，看了看就十人左右工位

（2）技术方面，直接拿我简历，翻来覆去挑着问的，没有问超过简历之外的技术，所以同学们遇到了这公司，自己简历上的得会

（3）问我薪资待遇，然后介绍他们公司业务，广告行业我能不能接受

第32章 嘉华颐和

（1）经历

（2）做没做过数据分析

（3）没问技术

（4）介绍了他们公司情况

第33章 商越网络

33.1 面 1

33.1.1 笔试：

（1）两个时间戳相减等于多少小时

（2）时间戳转日期命令 from_unixtime

（3）字符串中取最大和最小字符串

还有一些简单的 SQL 题 排序，取最大最小日期，各学科成绩都 80 分以上的学生 ID，一日留存这类

33.1.2 面试：

（1）linux 查询命令

（2）往 HDFS 上传文件的命令

（3）解压文件命令

（4）单引双引的区别

（5）left join，inner join full join 的用法

（6）Sqoop 了解吗

（7）CK 的引擎

（8）数仓，数据库，数据集市有什么不同

（9）hive 优化

（10）数仓建模，怎么分层，都有什么作用

（11）有什么想问的

第34章 优加健康

（1）怎么实现用户画像的标签任务，具体用了什么方法，代码怎么写的用了哪些算子

（2）怎么去划分一个用户偏好标签

（3）用户画像怎么处理数据了

（4）flink 实时项目介绍

（5）flink 用了什么 api

（6）你做了哪些指标，说其中一个，从 ods 到 ads 数据一步步怎么处理 每一层都用了什么方法，怎么处理的，怎么关联的。

（7）中间做了哪些优化。

（8）ods 层在 kafka 中分了几个副本

（9）Kafka 丢数据

（10）flink 怎么维护 offset

（11）Kafka 积压

（12）flink 为什么要用动态表

（13）动态表存在哪了

（14）zk 都监控了哪些组件

（15）shell 脚本 写从 1-100 的输出

第35章 中国软件

35.1 面 1

（1）离线数仓分几层？

（2）离线你做了哪些？

（3）指标你怎么理解的？

（4）实时你做了哪些？

（5）实时中，数据的补偿机制？没懂，好像说的就是数据一致性，说完 flink 怎么保持数据一致性之后，又说 Kafka 参数调优保证数据一致性，他就说怎么调优都不可能保证数据一致性。

（6）hive 开窗函数

35.2 面 2

（1）提到权限这块，说是央企做税务这块的东西，权限要求比较多，我说我们不怎么涉及

（2）hive 了解多少，数据倾斜怎么处理，两个大表 join 数据倾斜怎么处理，问我们数据量

（3）构建业务矩阵这块不涉及业务该怎么做，提到权限相关，不了解

（4）说一个开窗函数的具体用法

（5）flink 的数据一致性，重复消费的数据会迟到，说了一下数据乱序怎么处理

（6）对指标和标签的理解

（7）其他问的都是大保健上的，没啥难度

第36章 国研大数据

（1）和上一位面试者的内容，简历。架构很相似，你认识***吗

（2）处理的数据量多大

（3）kafka 包括实时和离线接入数据的峰值是多少

（4）flink 用什么版本，什么时候开始的实时项目

（5）集群规模，公司地址，部门多少人

（6）Yarn 页面总的计算资源，多少 CPU，多少内存，离线划分多少，实时划分多少

（7）离线每天跑多少指标，每人负责多少个指标

（8）sql 地区，价格两个字段，每个地区的价格 TOP5，不用开窗

（9）分桶用过吗

（10）Spark 哪些算子触发 shuffle

（11）数据质量监控用什么监控，了解吗

（12）元数据管理了解吗

第37章 高伟达

（1）自我介绍

（2）重点问了 hadoop 集群的搭建问题，spark hive 怎么配置的

（3）数据量 又问了 Spark 内存怎么配置的

（4）数据怎么处理数据倾斜

第38章 汉德科技

（1）自我介绍项目

（2）DATAX 了解的多吗，遇到过其他问题吗

（3）增量数据怎么采集的，为什么选择 maxwell

（4）写到 Kafka 中的话有几个 TOPIC

（5）maxwell 的底层原理

（6）Kafka 下游使用什么到 HDFS，我说的 Flume 还有解决小文件问题的配置

（7）Scoop 和 Datax 的对比

（8）Hadoop 小文件的危害和解决

（9）HIVE 小文件的处理方式，Hive 常用的存储格式

（10）Spark 和 MR 的对比

（11）了解 CK 吗，讲了下表引擎计算快，还了解过其他数据库吗，说了 ES、redis、Hbase

（12）然后就是问我有没有其他问他的，问了数据量，项目主要做什么，公司的数据中台是怎么运行的

第39章 安胜

39.1 面 1

（1）项目介绍 自己负责哪些工作

（2）公司是做什么的 面向什么客户

（3）项目上线了吗？活跃多少？

（4）集群多少台

（5）参与过的项目遇到的问题 怎么解决的

（6）实时做了什么

（7）kafka 默认消息时间是 7 天 7 天以后删除的机制 是什么

（8）hive 的分区表和分桶表的区别

（9）shell 脚本$1 什么意思

（10）awk 的用法，linux 工具

（11）平常开发语言 我答得 Java

（12）map 和 flatmap 有什么区别

（13）flink 的窗口 水位线 基本概念问题

（14）上一题的延伸 实际项目有用 flink 做过什么吗

（15）讲一下离线数仓的 dwd dim dws 层干了什么

（16）离职了吗 离职原因

（17）工作强度怎么样

（18）薪资要求

（19）有什么问题要问的吗

39.2 面 2

（1）介绍一下项目？ 大概讲了一下采集系统以及维度建模

（2）项目团队几个人？

（3）datax 的工作原理？

（4）datax 好在哪里？

（5）kafka 默认消息时间是 7 天，大量数据批量过期怎么删除？ broker 里面用 segment，里面有.log,.index,timestemp, 当一个 segment 中的全部的 ts 都过期就删除

（6）flink 时间窗口和水印？

（7）hive 分区和分桶？

（8）实际生产中遇到的问题？

（9）平时加班多吗？

（10）对 linux 熟悉吗，awk 干什么用的？

（11）hdfs 默认副本是 3 个，这么做的好处，基于什么考虑的？

（12）工作之外对什么领域有研究？讲了给大概的 hudi

（13）问我跟 kudu 有什么区别？

（14）为什么离职？

（15）期望薪资？

（16）还有什么想问我的？

第40章 成都执掌天下

（1）自我介绍

（2）介绍实时数仓集群

（3）flink dim 怎么做的

（4）clickhouse 里边的多表 join 怎么优化

（5）hbase 一个表如果有 6 亿数据，查询优化

（6）集群服务器有多少台，集群组件各放在哪里

（7）你们数据量多少 你们业务数据量多少

（8）实时的 flume 传输遇到什么问题

（9）dataX 如何实现增量同步

（10）dataX 和 flume 有什么区别

（11）hive 执行 sql 的优化

（12）window join 的本质

（13）用户量、商品量、维度最大的字段

第41章 集度汽车

41.1 面 1

（1）自我介绍

（2）为什么先离职

（3）上游下游对接业务

（4）职业规划

（5）遇到的最大问题（业务和技术）

（6）两个 sql 留存率，路径分析（不让我用开窗最笨的方法说）

（7）思维题：有一个团购网站：领导看报表的时候发现一个烤串店 12 月的数据比 6 月多 50%，那些原因

41.2 面 2

（1）自我介绍；

（2）离线数仓分层，每一层都是怎么搭建的；

（3）离线数据域划分；

（4）自己负责的数据域的数据量；

（5）自己负责数据域中有哪些表，那张表的数据量最多，大概多少条；

（6）零点漂移问题怎么解决的；

（7）hive 中数据倾斜的优化

（8）什么场景下发生的数据倾斜，具体是怎么解决的

（9）对数据质量监控的理解，你做过这方面的工作吗？（我说没有），接着问，如果没有你怎么保证你计算的结果是正确的

（10）sql table 表，两个字段 a,b，每一行表示一个用户关注了另一个用户，让求出相互关注的用户；

（11）向他提问

41.3 面 3

（1）自我介绍

（2）随便介绍一个项目

（3）实时项目中 dwd 层多流 join 窗口等待是如何处理的？

（4）迟到数据如何处理？

（5）实时数据的结果是用什么存的？

（6）把窗口粒度比较细的情况下，是否近乎明细数据，存储是否有压力？

（7）用户画像介绍一下

（8）维度建模有什么优点？探讨了数据一致性哪种模型好

（9）hive 的数据倾斜？join 数据倾斜怎么解决？

（10）事实表都有哪些种？

（11）sql 已截图

第42章 四川星点网络科技有限公司

（1）自我介绍

（2）讲一下实时数仓

（3）为什么你们实时维度数据没有做维度整合（涉及多个商品信息），是来一条数据都去关联吗？

（4）flinkCDC 读取配置表可以监控维度配置表的变化，比如业务数据库中维度表字段发生变化了会怎么样？我说我们维度表字段不会怎么变化，那岂不是使用 flinkCDC 矛盾了？

（5）有没有遇到多个事实表关联的情况？

（6）如果有多个事实表要关联，使用 flinkSQL 使用什么 join

（7）窗口你们开多大，统计一天的实时指标比如 GMV 怎么实现的？

（8）如果要统计 UV 这种是不是就不能用窗口了，该怎么实现？

（9）flink 数据一致性，链路延迟，数据乱序

（10）实时项目中遇到什么比较困难的点吗？

（11）你们用户画像是怎么做的，里面比较有挑战的点有哪些？

（12）说一下你们离线的情况？

（13）建模用的什么？具体说一下

第43章 东软

（1）各项目搭建过程遇到什么问题。

（2）了解 zookeeper 脑裂么，怎么预防。

（3）指标 GMV 指标（面试官没听过）、品牌复购等。

（4）Spark 调优。

（5）scala 常用 API。

（6）SparkStreaming 运行机制。

（7）星型模型

场景题：

（1）个文件按 T 计量（存的都是 URL），怎么实现去重

（2）如何处理单条记录消耗过大，即处理单条记录时，突然吞吐量不高怎么办。前提机器性能不高。

（3）一个 40 亿的数据量的数据，给你一个条数据（比如 123），怎么最快查询 40 亿的数据里出现过没有？

第44章 法本

（1）数仓为什么这么分？

（2）如果想取一个不脱敏的数据你们是怎么做的

（3）你们 etl 用的是什么？

（4）在这两个项目里你主要负责什么方面？都写过什么 API？主要思路讲一下？

（5）你都看过 Java、spark 的什么源码？

（6）这数仓是你一个人搭建还是你们组一起搭的？现在要你搭一个采集平台或者数仓要多久？

（7）这些框架之间的兼容性你们是怎么解决的？有没有可以代替这些框架的？

（8）如果给你一个杂乱的数据，里面有 json、txt、word。你怎么取出你想要的数据？

Demo， 切割，正则

第45章 巽联科技

（1）自我介绍

（2）公司规模，团队人数，组长如何安排的任务

（3）最近工作内容介绍

（4）离职原因

（5）串讲采集系统

（6）介绍一下海豚调度器

（7）关系型数据库有哪些，用的熟练程度

（8）建模过程串讲

（9）用户画像标签的 sql 与代码如何实现

（10）hive 底层 元数据有哪些信息

第46章 首信云

46.1 面 1

（1）介绍画像项目

（2）flume 采集日志数据，需要过滤出 JSON 中某个字段为空的，拦截信息给前端，需要用户看到，怎么做？

（3）flume 采集数据直接到数仓中，文件中包含空格换行等一些特殊字符，转换成 ORC 存储，会有错乱，怎么处理？

46.2 面 2

（1）离线项目

（2）Flume 怎么采集

（3）Flume 里面的 sink 是什么

（4）怎么把数据从 hdfs 写入 hive

（5）如果你要的指标，离线数仓里面没有数据怎么办

46.3 面 3

（1）做个自我介绍

（2）介绍一下用户画像的流程

（3）bitmap 的作用

（4）pivot 的用法

（5）Flume 如何自定义拦截器过滤空值

（6）你在离线项目中主要做了什么？

（7）Flume 采集数据在 HDFS 中的存储格式是什么？

第47章 嘉一未来

（1）介绍下项目，离线实时都介绍了

（2）项目组有多少人，都是干什么的，你是干什么的

（3）每天的数据量有多大，kafka 有多少个 topic

（4）Kafka 遇到的问题（挂了、丢了、重了、积压了、乱序了都问了一遍）

（5）Kafka 的高效读写

（6）ES 了解吗

（7）Flink 反压

（8）Spark 和 sparkstreaming 了解的多吗

第48章 万古恒信

（1）画像项目的流程以及标签的定义

（2）flume 采集日志数据，需要过滤出 JSON 中某个字段为空的，拦截信息给前端，需要用户看到，怎么做？

（3）flume 采集到 HDFS 上的文件格式，DATAX 采集到 HDFS 是什么格式

（4）如果让你独立搭建数仓系统是否可以

（5）数据治理，数据质量分析是否了解过，能否独立安装

（6）flume 都有哪些 sink，是否了解过，能否配置使用

（7）问了下家住何方

（8）剩下的记不得了，白话了一会

第49章 泰康

（1）数据量大小

（2）负责什么模块

（3）比较难的业务，怎么实现的

（4）状态怎么保存，为什么用这个方式保存

（5）flink join 使用场景

（6）dim 大小

（7）为什么引入 redis

（8）hbase 索引种类，用了什么索引，为什么用 hbase

（9）并行度设置方式，为什么设置这样的并行度

（10）检查点超时报错怎么处理

第50章 老虎国际

50.1 面 1

（1）介绍上一家工作项目

（2）数仓数据来源，你们的数仓指标是怎么确定的？

（3）列举下统计的指标？

（4）flink 状态存储类型及各自的优缺点？

（5）file 和 rocksdb 是否都支持增量？

（6）flink 的 state 类型？

（7）flink 的 checkpoint 的存储原理，若缓冲区状态满了，该如何处理？

（8）说一下 flink 指标的实现流程？我讲了下各省份每日订单数量，问到了维表关联用的什么函数？

（9）clickhouse 的数据类型有哪些？

（10）hive 的数据类型有哪些？

（11） hive 的存储类型有哪些？json，textfile，还有哪些？

（12）java 方面，常用的集合类型？hashmap 和 hashset 的关系？

（13）常用的排序方法有哪些？各自的时间复杂度和空间复杂度是多少？

（14）牛客网上写一个归并排序，讲一下归并原理？

50.2 面 2

（1）自我介绍

（2）介绍项目 介绍的实时

（3）具体负责的部分，具体是怎么实现的，用到了哪些 flink 独有的特性

（4）checkpoint 的 barrier 对齐 机制

（5）flink 数据倾斜解决办法

（6）flinksql 有几种

（7）介绍 ck

（8）bitmap 好处和坏处

（9）java 归并排序，用两个链表实现方法栈，写一个 euqal 方法实现比较两个二叉树是否完全相等，说出有几种排序 和 他们的时间复杂度和空间复杂度 ，说出 currenthashmap 原理。

（10）写出 topN 的 sql

50.3 面 3

（1）自我介绍

（2）上份工作经历，项目经历

（3）是以离线为主吗

（4）离线，实时整体架构与业务划分

（5）整个数据流向与处理

（6）技术擅长领域

（7）讲一下具体擅长技术

（8）hive 的数据类型

（9）hive 架构 底层 暴露的东西

（10）hive 组件作用

（11）spark 原理

（12）kf 架构

（13）broker 作用

（14）选举过程

（15）ck 数据类型，表引擎

（16）算法空间，时间复杂度

（17）java 实现类

（18）线程安全的集合

具体题目

（1）归并排序 java 实现

（2）子查询实现 sql

（3）链表问题，具体我也不清楚了，人已经傻了

（4）有什么想问的

第51章 百度外包

（1）海豚调度器工作原理

（2）datax 页面化使用

（3）flink 的异步 io 的底层原理

（4）状态和检查点的区别

（5）对 sql 的优化

（6）数据治理

（7）项目难点

（8）对集群的一些优化

第52章 华胜天成

52.1 面 1

（1）离线项目串讲

（2）HIVE 优化，小文件，数据倾斜

（3）写过的指标、疑难指标解决思路

（4）一道 sql 题

（5）hive 分区分桶表

（6）Linux 系统常用的命令，查看内存、磁盘、查找文件

（7）拉链表的实现思路

52.2 面 2

（1）hive 的优化

（2）hive 数据倾斜

（3）分区分桶表

（4）拉链表

（5）一道 sql 题：查询左表有右表没有的数据

（6）查一个表的前五行

（7）TopN

52.3 面 3

（1）讲一下项目架构

（2）讲一下 clickhouse 和 hive 的区别

（3）平时用 hbase 吗？没听清

（4）讲一下脚本

第53章 实惠多多

53.1 一轮技术面试（数据开发主管）

（1）你们公司的业务是什么。详细说说你们数仓怎么使用了

（2）说下怎么实时数仓的数据是怎么来的，你们离线的用户行为数据是怎么产生的，你怎么处理的，业务数据是怎么产生的

（3）flink 用过状态编程么？ 你们怎么使用的，你说说你怎么用的，结合你自己做过的指标说说怎么写程序。详细说

（4）那你用的是哪一种状态， 不开窗吗？ 我说是设置 ttl ，如果开窗怎么开？ 用什么窗口，开多久。

（5）flink CEP 用过吗？ 怎么怎么使用，结合具体指标，数据从哪抽取，怎么处理，实现什么？详细说

（6）flink 的数据写到哪里

（7）集群规模多大，数据量每天多少，用户有多少

（8）离线数仓说说你们怎么怎么采集数据的

（9）kafka 怎么处理重复数据

（10）Kafka 怎么处理积压数据

（11）kafka 的主题删除方式有哪些，有需要注意什么吗？

（12）现在有个场景，kafka 出问题你怎么去排查出问题，说说你的思路

（13）Doris 用过吗

（14）Doris 和 clickhouse 的区别是什么 有哪些引擎

（15）为什么你们没有用 Doris

（16）flink 用的什么版本

（17）离线数仓建设最重要的是什么

（18）hive 你们做了哪些优化

（19）你们怎么去做数据治理的

（20）spark 你们是怎么用了，有哪些应用场景

（21）flinkcdc 你是怎么用了

53.2 二轮面试（啥啥总监没听清）

（1）你们公司业务板块了解吗，说说

（2）之前公司数据开发团队多少人

（3）你们有做数据中台吗，了解过吗

（4）说说你的用户画像项目

（5）你们怎么设置标签，让我说怎么分级，一直说到 4 级

（6）对自己有什么职业规划

（7）问薪资

第54章 名途信息

（1）介绍一下用过的计算框架

（2）说一下 flink join

（3）flink 怎么避免数据乱序，端到端一致性

（4）hive 内部表默认存储路径

（5）简单说一下 spark 宽窄依赖

（6）场景：窄依赖，两个 RDD，两个分区，union 以后有几个分区（两个）

（7）kafka 和 sparkstreaming 精准一次如何保证

（8）介绍一下印象比较深的做过的项目（离线和实时都说了，说了 hive 的优化和数据倾斜，flink 大状态调优，反压，数据倾斜）

第55章 北京芯盾时代有限公司

（1）在实时数仓中主要做了什么？提到参与了技术选型，追问具体参与了哪些技术选型？怎么进行技术选型的，做了哪些工作？感觉面试官应该是想知道框架对比，为什么选这些组件吧。

（2）行为数据是存放在日志文件还是哪里？flume 的 taildirsource 怎么实现断点续传的

（3）maxwell 基于 MySQL 的 binlog 日志采集数据发往 kafka 中，具体是怎么实现的？

（4）对那个大数据组件有深入了解，我说的 kafka，kafka 怎么保证数据不丢？kafka 的分区的 leader 怎么确定的？（broker 总体流程中 leader 选举机制）

（5）kafka 往 zookeeper 中注册信息是写一条记录还是创建一个节点还是怎么样的？

（6）zookeeper 的节点有那几种类型？

（7）flink 消费 kafka 数据怎么保证幂等性？数据的精确一次性，消费数据是一条还是一批？

（8）flink 的 checkpoint 的原理，flink 用过哪些算子？并行度了解吗？

（9）flink 支持哪些分区策略？没答上，flink 的提交模式有哪些？

（10）怎么确定 kafka 安装部署完了？怎么确定安装好了的 kafka 是完好的？应该是想问 kafka 的压测，压测命令是什么

（11）说一下 clickhouse

（12）hbase 的读写流程，hfile 和 memstore 的具体关系是什么

（13）去 hbase 中查数据会做哪些优化

（14）hdfs 的读写流程，写一个 1G 大小文件的流程，hdfs 查看有哪些文件的命令？

（15）namenode 高可用两个怎么切换的，hdfs 高可用自动故障转移机制，ZKFC 组件

（16）hadoop 集群安装部署过程，提到文件句柄数？你知道吗？只听其名，不知其意，质疑我这都不知道

（17）问我有没有漏掉什么大数据组件，我说 hive 没提，就给我出了一个统计哪些班级，学生人数大于 60 人 SQL 实现，学生 id，班级 id，学生 id 不重复

第56章 飞轮数据

56.1 面 1

自我介绍

工作经历

离线项目 技术角度

java 怎么样

写入 clickhouse 用了什么技术栈

用户画像怎么对外提供服务，怎么查询

kafka 底层文件格式管理

其他的分析型数据库

es

ck 的优化

flink 的 watermark

flink 端到端一致性

flink 两阶段提交

56.2 面 2

（1）自我介绍

（2）项目过程中的优化

（3）问 kafka 的单分区数据不重复，那如何保证 实时环境 下多分区数据不重复（他说 ETL 可以，下游 flink 涉及到去重压力）

（4）问计算引擎了解多少

（5）问 olap 有啥

（6）问 clickhouse 了解

（7）问 bitmap

（8）问 hashmap，为啥要用红黑树替换链表，用红黑树解决什么问题；

（9）怎么解决 hashmap 的线程安全问题

（10）有啥想问他的

56.3 面 3

（1）讲讲 Java 多线程

（2）讲讲 JVM

（3）讲讲 CMS，怎么用的

（4）讲讲 NIO

（5）Redis 主从复制

（6）介绍一下 ClickHouse

（7）HDFS 读流程

（8）Flink 端到端一致性

（9）水位线介绍下

（10）列式存储和行式存储区别

（11）ClickHouse 更新怎么做的

（12）ClickHouse 的 SumingMergeTree 和 ReplaceMergeTree 区别

（13）Spark 如何怎么扩大并行度

56.4 面 4

（1）介绍一下比较拿手的项目，我讲的离线

（2）介绍 Flink 和 Spark 检查点区别

（3）介绍下 CheckPoint 机制

（4）介绍状态

（5）Java 的 TreadLocl

（6）Java 多线程的信号变量

（7）TreeMap 底层是啥

（8）介绍下红黑二叉树

（9）JVM 虚拟机全套

第57章 文创思宇

这公司没有大数据

（1）自我介绍

（2）干过什么项目，具体负责了啥

（3）离职原因

（4）根据简历问的，问 java 用的怎么样，会哪些东西

（5）flume 自定义拦截器用 java 怎么实现的

（6）自定义函数怎么写的，这个人是懂 java 的不懂大数据的，光想往 java 引

（7）工作中碰到比较困难的问题以及解决方案 我说的 kafka 的一些

（8）问了点离线的指标实现

（9）公司的数据量，人员配置，具体负责，上家公司做什么的，数据怎么来的等杂七杂八

（10）hive 的优化

第58章 奇点云

（1）说数据怎么来源

（2）说你们离线数仓的集群规模，你们有测试集群吗，离线数仓和实时使用的是一个集群么？ 会不会有集群资源的冲突？怎么解决了？怎么做测试？

（3）离线数仓怎么采集数据， 你们在采集数据之前会对数据做什么筛选吗？ 你会去关注些什么？

（4）你们为什么做数仓？ 说说你们数仓最后输出之后哪些业务会用到？

（5）有哪些建模方法？

（6）为什么选择星座模型？都有何区别,为什么不选 er

（7）你们怎么去构建业务总线矩阵？是怎么把这个矩阵做出来的？ 是你自己做的吗？

（8）都有哪些人去参与的？ 怎么个工作流程？

（9）业务总线矩阵包含了什么信息？ 你们数据域怎么划分？是谁划分的？

（10）说说什么是原子指标？什么是派生指标？ 并且当场举几个例子？然后他又说了几个指标让我判断是哪一类的指标，为什么这么判断

（11）说说你们 五个层都怎么做了， 从 ods 详细说

（12）ods 层怎么做的压缩。

（13）dim 层 你们做了哪些表，说几个。

（14）如果维度变换缓慢有做哪些处理呢？

（15）dwd 层怎么去构建事实表，设计依据是什么？

（16）dwd 的都做了什么事

（17）那你们怎么选择做维度退化，为什么

（18）dwd 你们有没有跨数据域的表

（19）hive 有没有做过数据规模的统计

（20）你会做数据量的评估吗，怎么实现？ 比如如果现在用户量 10w，集群规模怎么去考虑，如果以后达到 100 万。

（21）数仓的数据质量监控怎么做的？

（22）怎么做元数据管理，元数据管理是干嘛的？

（23）业务之间你们怎么做优先级的？

（24）生产中你们项目是怎么开展的具体流程？详细说

（25）你在其中什么角色

（26）sql 语句的准确性你们怎么保证的？你会怎么判断 sql 语句的正确性？ 你说下你的经验

（27）你平时开发是 sql 多还是 api 多

（28）sql 题：当场直接说思路

a) 连续 7 天登陆

b) 三个字段：任务 id，任务的 start_time,任务的 end_time.求输出每一分钟任务量是多少？

（29）用户画像你们做了哪些标签？用户画像得到的数据存在哪了谁去用了

（30）你们怎么去做推荐标签？ 说说你做的 我说了 rfm 模型 追问怎么判断重要 如何衡量 详细说

这家公司是乙方，跟着项目走。做中台。

第59章 汉熵

（1）自我介绍

（2）spring cloud 用了那些组件?

（3）ck 怎么去做存储？

（4）ck 用到了其他的表引擎嘛?

（5）MergeTree 有那些要素 那些字段 用来分区 ？

（6）用过那些函数?

（7）怎么做的数据清洗和转换写到 ck 中的?

（8）建模过程

（9）hive 执行 sql 过程

（10）hive 接收执行完的结果

（11）实时数仓数据没有保存在 hdfs 中 保存在什么地方 数据分层都是在 flink 中的嘛?

（12）flink 中做了那些内容? ods 拿数据 怎么做一些处理操作?

（13）数据倾斜怎么解决的？

（14）具体点,打散后怎么办?打散策略是什么? 开窗然后聚合统计 然后在聚合 说不对

（15）kafka 怎么部署的 怎么做的 topic 分区 副本 多少个?

（16）ck 几个服务器? 怎么保证服务器数据一致的 怎么同步的数据？

第60章 中软闻歌

60.1 画像

bitmap 原理 和为什么用

怎么与 spark 结合的？

spark 集群几个节点？

60.2 实时

数仓分层

cdc 过程遇到的问题

FlinkCDC 1.x 锁表问题， 配置表， 配置库，对我们实时计算不产生任何影响

成就比较大的地方

flink 提交模式

60.3 离线

datax 遇到问题

其他框架有调研嘛

Kafka， kubemq（云原生，比较窄）

datx 内部架构

datax 同步 hive 流程

MySQL 有逗号 怎么同步到 hive

Hive 同步 mysql，分隔符

字段有空格，regexp_replace

flume 过滤器定义

拦截器

kafka 缺点， 依赖 zk

全局有序

分区减少不行 为什么

ack

topic 变化 生产者怎么感知

topic.metadata.refresh.interval.ms

kafka 怎么与 flink 连接

hive 优化 大保健

hive 版本

60.4 Java

设计模式

内存泄露 溢出区别

溢出： Out of memory 报错

泄露：申请内存暂时没有被释放，暂时不会出问题，导致溢出

注解类

spring 生命周期

flink 会落盘

计算：算子，TM 内存

大状态：RockDB， Ck 快照到 hdfs

catlog 了解过吗

Flinksql，表，字段

第61章 安胜（二面）

（面试官不怎么懂技术）

（1）项目中遇到的问题

（2）部门多少人

（3）技术怎么确定

（4）未来职业规划

（5）做过数据分析吗

（6）用户画像怎么做的

（7）前后端怎么配合

（8）用过阿里云，华为云这些吗

（9）之前薪资，期望薪资

（10）讲了一下他们公司的结构（面向政府的项目）

（11）如果入职可能要去厦门总部培训

第62章 实惠多多

62.1 一面

（1）自我介绍

（2）介绍实时数仓

（3）mysql 的去重方式？

（4）hive 常见函数，表函数有哪些？处理 json 的有哪些？

（5）事实数仓如何过滤不合法的 json，或者一些不需要的数据？

```
Where  a <>
Not  in ('')
<  >  =
Is  not  null
```

（6）如何将 hive 中的数据读到 clickhouse ？

（7）是否用过 doris？

（8）数据中台怎么理解？

（9）大数仓概念

62.2 二面（一个做 java 的）

（1）自我介绍

（2）介绍事实数仓

（3）怎么看商品的大数据搜索推荐

（4）怎么看数据库，数据仓库，数据湖？

（5）存储机制，处理介质

（6）有什么想问的？

第63章 汉德科技（二面）

（1）先介绍离线项目

（2）为什么加 maxwell 还要加 datax，maxwell 的底层原理

（3）数据落盘 HDFS 的文件格式

（4）数据量多少，组内多少人，原先公司做什么

（5）hive sql 和普通 sql 语法的区别

（6）数仓搭建过程和建模过程

（7）疑难 sql 的思路

（8）row_number、rank 区别和使用

（9）Spark 中 RDD 的特性

（10）介绍用户画像搭建过程和日常负责工作　四个人面试的轮番问，其他的问题想不起来了

第64章 指南针

64.1 面 1

（1）自我介绍

（2）你们的数据量

（3）hive 架构（组件，底层）

（4）技术栈

（5）数仓建模说一下

（6）集群谁搭的

（7）HQL 执行流程

（8）有哪些排序（4 个 by）

（9）partition by 个 group by 的区别

（10）开窗函数，举例子，都是干什么的，区别在哪

（11）行转列，列转行

（12）场景题：求累计百分比 sql

（13）RDD 五大属性

（14）rdd 弹性体现在哪

（15）spark 作业提交流程

（16）spark 任务划分

（17）sparkstreaming 窗口

（18）flink 和 sparkstreaming 区别

（19）flink 水位线

（20）flink 窗口

（21）flink 状态

（22）主要用什么编程语言（java），不用其他的了嘛（shell 写脚本没问题，了解 python 和 scala）

（23）算法题：爬楼梯

（24）上一家待遇，期望待遇，住哪，你有什么要问的。然后让我好好看看算法

64.2 面 2

（1）介绍项目 离线

（2）你为什么要离职

（3）团队和数据量

（4）聊聊大数据体系

（5）hadoop 读写流程

（6）副本默认 3 个

（7）纠删码原理

（8）3 个节点传输怎么传 串行的还是并行

（9）100g 文件 传完之后开始同步 还是怎么个流程

（10）读文件的流程 读哪个文件 就近原则

（11）yarn 架构

（12）hive 架构

（13）操作树转换成 mr

（14） 4 个 by

（15）哪个 by 只能有升序

（16）partitionby

（17）patitionby 与 group by 区别

（18）对开窗函数的理解

（19）join 的话 需要注意些哪些

（20）谓词下推 和 列裁剪

（21）kafka 和 flume

（22）数据漂移问题

（23）你对数仓规范的理解

（24）写代码在哪写

（25）集群搭建多久

（26）中间出现过什么问题

（27）都是 leader 负责

（28）如果说让你做数仓命名规范

（29）业务线怎么区分 大业务线 我们只有一个业务线

（30）分层分几层 哪五层

第65章 网达软件

（1）介绍离线数仓

（2）介绍实时数仓

（3）flink CDC 怎么去做

（4）使用什么时间语义

（5）flink 精准一次性怎么保证

（6）Kafka 精准一次性怎么保证

（7）zk 里面维护了 Kafka 的什么

（8）Kafka 可以放弃 zk 嘛

（9）hivesql 的调优

（10）spark 参数调优

（11）yarn class 和 yarn client 有什么区别

（12）redis 数据类型

（13）redis 存储模式 rdb aof

（14）hdfs 读写流程？

（15）hdfs 写数据的话，串联通道一个节点挂了咋办

（16）StringBuffer 和 StringBiuld 区别

（17）基本数据类型和动态数据类型

（18）JVM 内存模型，问到这里我赶紧打住我不是干 java 的是干大数据的

（19）同步和异步区别 说了 Kafka 的异步操作

（20）sleep 和 wait 的区别 我不知道

（21）为啥用 DataX，DataX 底层是怎么做的。

（22）写 java 代码 string 1.9.8<1.10<1.10.1

第66章 金康塞力斯

（1）实时项目串讲

（2）实时维度指标有哪些？dwd 层做了哪些，dim 层做了哪些，flinkCDC 给他了讲一遍

（3）spark 的工作流程

（4）往 hbase 写 100 亿条数据怎么设计，我说的 rowkey 的设计原则

（5）BI 工具有使用哪些？

（6）项目中的痛点怎么解决的？我说的资源设置，反压

（7）公司是车企，类比一下车企的话该怎么设计数仓这些，我就把车类比成商品给他说了一下。

（8）面试官应该不太懂大数据，随便问的

（9）HR 了解了一下个人情况巴拉巴拉

第67章 德特赛维技术有限公司

（1）flink 提交流程，数据一致性，checkpoint 的理解，在项目的实际应用什么时间语义？

（2）flink 并行度和 slot 的理解

（3）kafka 怎么指定分区进行消费，发送数据分区策略，ISR 的理解，数据重复

（4）sparkSQL 和 hiveSQL 的区别？为什么 sparkSQL 能使用 hiveSQL 的函数？sparkSQL 怎么自定义函数？我说的 hive 自定义函数

（5）clickhouse 了解多少？使用过 aggregatemergetree 没有

（6）clickhouse 和 kafka 对接过没有？

（7）连续登录 SQL 实现思路

（8）上来就共享屏幕，连忙处理桌面，都忘记录音了，一些问题记不清了

（9）公司主要用 sparkSQL 写代码开发，实时 flink 主要用 scala 代码开发，公司主要做离线的

第68章 枫叶汽车

（1）自我介绍

（2）离线架构(采集->分层建模) 问：知道用的这是什么架构吗？Lambda 架构

（3）Hadoop 框架

（4）HDFS 文件块

（5）多人同时操作同一文件会出现什么情况？

（6）HDFS 块大小可以调整吗？在哪里调整？；

（7）JobTracker 了解吗？

（8）Hive 问知道 orc 吗？可以简单说一说吗？

（9）Kafka 与传统的消息队列比较；

（10）Flink 与 SparkStreaming 区别；时间语义；WaterMark 定义乱序时间；状态怎么保存？

（11）最近有学习什么吗？

（12）介绍公司情况

第69章 上海哲锦

（1）自我介绍

（2）介绍离线

（3）离线你负责什么

（4）Linux 命令熟悉吗

（5）有什么想问的

第70章 文创思宇

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

（1）hive 优化

（2）指标怎么实现的

（3）离线架构

（4）实时架构

（5）画像架构

（6）面试官写算法和 java 的，问我你的代码中那里体现面向对象了，问了我会的设计模式

第71章 指南针

（1）自我介绍

（2）hive 的优化，结合自己的指标

（3）讲一下实时的架构，用到的技术

（4）讲讲公司的数据治理

（5）讲讲离线脚本是怎么提交的，有没有集成化

（6）讲讲 hive 四个 by，怎么用

（7）开窗的 partition by 和 group by 区别

（8）讲一下开窗

（9）出了一道 sql 题，统计店铺销量 topN，累计销量等等

（10）行转列，列转行

（11）RDD 五大属性

（12）rdd 弹性体现在哪

（13）spark 作业提交流程

（14）spark 任务划分

（15）讲讲 flink（面试官说自己是一个小白，给介绍一下，内存模型、cpu、水位线、窗口、状态。。。）

（16）公司用的 cdh，想换成阿里云，问我知不知道我们公司搭建阿里云大概花了多少钱

（17）问了下他们公司的架构，所和我说的基本一样

（18）问了问我离职原因

（19）问了自己的发展方向，期望未来的团队什么样，期望薪资

（20）最后问了算法（二分，跳台阶，二叉树）

第72章 博英科技

（1）自我介绍，问那年毕业，专业

（2）实时 flink 主要的工作

（3）数据是怎么留转的 ods-dwd，具体的代码 api

（4）flink 运用什么方式运行 api 脚本，怎么提交

（5）有多少条数据每天

（6）flink 分流具体做了什么处理

（7）集群规模

（8）离线数据处理用到哪些技术

（9）clickhouse，hbase 都问了

（10）数据倾斜场景，处理方法的底层原理

（11）提问

第73章 数梦工厂

（1）自我介绍

（2）介绍项目？从采集开始一顿讲离线

（3）kafka 积压了怎么监控到的？

（4）生产中遇到的问题？

（5）数据怎么来的？

（6）期望薪资？

两个 SQL

（1）表 A 和表 B 有唯一主键 id，重复的按表 A，怎么做？

（2）如果表 有两个字段，id 和名字，怎么变成[1,2,3,4,5]和[名字 1，名字 2 名字 3，名字 4，名字 5]

第74章 中软国际

74.1 面 1

（1）自我介绍

（2）介绍一下用户画像

（3）Flink 端到端一致性，ClinkHouse 的读写优化

（4）你还有什么问题想问

74.2 面 2

（1）自我介绍

（2）介绍一下用户画像

（3）Flink 端到端一致性，sink 端 ClinkHouse 是如何实现的

（4）CAP 法则为什么只能满足两个，zk 满足了哪两个

（5）用户画像中是如何进行权限控制的

（6）Springboot、mybatis 掌握到什么程度了

（7）你还有什么问题想问（主要是做车企相关的数据中台）

74.3 面 3

（1）sink 端 ClinkHouse 是如何实现的，遇到哪些问题？

（2）用户画像中是如何进行权限控制的？

（3）指标是如何定义的？

（4）Flink 中你主要做了哪些工作？

（5）你还有什么问题想问（主要是做车企相关的数据中台）？

第75章 中创新航

（1）离线数仓项目串讲

（2）flink 数据一致性，flink 窗口有几种

（3）提到职业规划，聊了数仓和数开分得还是挺清的

（4）有没有涉及到组件底层的优化？我说的 flink 的优化

（5）就没了？有什么想问的？问了他们数据量跟我说的 100G 差不对，但是整体集群多到 100 多台，真有钱

第76章 库珀

（1）自我介绍

（2）你们部门是不是有个性毛的？

（3）java 掌握到什么程度？java 怎么判断一个字符串是否为空

（4）数仓侧重于哪个方面？宽表是怎么建立的？能否能举一些例子

（5）怎么保证报表是正确的？比如说最后的现象这个数据非常高，你怎么排查

（6）你们怎么做数据治理的？

（7）你们各层是怎么建立的？星形模型和雪花模型怎么选择？

（8）场景 sql

（9）SQL 的执行顺序是什么？

（10）flink 怎么保证分区是有序的？

（11）flink 的版本？spark 的版本？

（12）spark 的 union 和 union all 有什么区别？

（13）spark 的动态裁剪了解吗？shuffle 了解吗？

（14）spark 后的文件数量由什么决定的？

第77章 成都芯极客科技

（1）1.实时项目串讲

（2）2.Maxwell 和 canal 为什么选择 Maxwell，底层原理，解析格式

（3）MySQL 的数据量有多大

（4）为什么选择将维度数据写入到 hbase 中

（5）flink 是使用 SQL 还是 API？

（6）你们对历史数据是怎么处理的？我说我们没有处理过历史数据

（7）kafka 数据保存多久

（8）你们 sugar 是别人做的吗？我说的是前端那边做的，我们负责发布接口

（9）用户行为实时建模的那种是你们做的还是别人做的，涉及到算法？我说不是我们做的

（10）flink 调优你们做了哪些？

（11）flink 哪些算子容易导致反压，我说的 flatmap

（12）kafka 数据积压怎么处理

（13）flink 数据丢失怎么处理的

（14）flink 到 clickhouse 的数据一致性，clickhouse 查询方面的优化了解吗？

（15）clickhouse 有没有用到 primary key，order by 这些，忘记了没答上

（16）flink 开窗了解吗，窗口的分类

（17）异步 IO 是怎么做的？异步怎么保证数据有没有成功，补偿机制什么的？

（18）为什么选择 clickhouse，没有用其他的

（19）有没有遇到 clickhouse 数据关联的情况，join 性能影响有哪些方面

（20）kafka 的 topic 数据量过大遇到过没有？怎么处理，topic 数据全在一个节点上，我说的负载均衡

（21）clickhouse 有没有遇到过 zk 元数据不同步的问题

（22）flink 和 sparkstreaming 的区别

（23）了解 kafka 的 shuffle 机制吗？不知道

（24）dataX 知道吗，kettle 知道吗

（25）flink 的 Dynamic table，动态表了解吗？

（26）olap 数据库的表示层工具用了哪些？BI 工具 sugar 和 superset，sugar 是商业的吗？那个公司的？

（27）提问，离线实时都没有，项目比较大，帮别人做中台，大数据刚起步，就 2 人，数据量半年 1800 亿更新数，具体不太懂

第78章 茶颜悦色

78.1 面 1

（1）自我介绍

（2）讲一个你比较熟悉的项目

（3）hbase 有做二级索引吗？rowkey 怎么设计？

（4）hive 的数据倾斜怎么定位？怎么解决？

（5）离线数仓有没有用过拉链表？

（6）redis 的缓存穿透和缓存雪崩是什么？怎么解决

（7）布隆过滤器是什么，用过吗？

（8）数据需求是和谁对接？产品还是业务方？

（9）没有产品你是否能对接？

78.2 面 2

（1）自我介绍

（2）两个大表，都是十几个 G,不使用大数据的情况下，单机模式查出两个表的相同之处

（3）业务中最大是那张表

（4）spark 的 checkpoint 跟 flink 的 checkpoint 有什么差别

（5）使用了什么特殊的指标

（6）数据量多少

（7）连续登录计算，隔一天也算计算

（8）flink 自己有没有开发端到端的一致性

第79章 国网电动汽车

（1）实时数仓项目的背景，业务支撑的情况，实时数仓建设过程以及架构是什么样的？

（2）Flinkcdc 是一个什么样的机制？

（3）你怎么理解 DIM 层？

（4）DIM 层的维度信息是由其他团队帮忙维护好的吗？

（5）能不能说下你们在使用 clickhouse 这块有什么经验或者什么挑战？

（6）我自己 Q 了 Doris，然后问 Doris 为什么好用，以什么机制保证查询效率高？但是由于忘了，我大概说了下它支持 join，然后是目前的学习计划。

（7）什么样的数据适合做到 DWS 层，什么样的数据适合做到 ADS 层？这两个层你是怎么理解的以及它们的职责？

（8）你们的团队数仓这几层都会去维护吗？是否分层分团队去做这个项目？

（9）你们在做项目的时候是否有数据处理以及规范来确定每一个层应该有什么数据？

（10）垂直的那些主题，比如 DWS 层那些垂直的主题一个人维护比较方便，那如果在 ADS 层可能用户域关联其他域的信息，这个是你们团队怎么做共享和互动的？或者存在这种情况吗？

（11）利用 FlinkApi 叫上 redis 解决查询效率和吞吐量问题，那么在开发有没有遇到反压问题，举个例子以及怎么解决的？

（12）你们的数量级大概是什么量级？高峰时段每秒吞吐量大概多少？

（13）怎么设置 flink 的并行度，有哪几种方法？怎么计算一个合理的并行度？怎么验证并行度的合理性？

（14）Flink 中有哪几种窗口，你这边用的比较多的窗口？滑动窗口你们是什么场景下会用到？滚动窗口没办法满足你们的需求？

（15）Flink 中的水位线是否了解？怎么触发窗口的，以及它的机制？

（16）Watermark 这块用的多吗？在哪里用到？

（17）离职原因。

第80章 德特赛维二面

（1）数据量集群规模

（2）离线数仓的架构，离线数仓的工作和职责？

（3）hive 的版本，数据倾斜讲一下

（4）你们 hive 怎么判断小表的？是否有参数可以设置？这个没答上

（5）sparksql 执行的时候某个任务就很慢，单不是数据倾斜导致的，该怎么进行优化。面试官的意思是数据量一样但是就某两个任务运行很慢该怎么进行优化

（6）知道推测执行吗？spark 的优化

（7）离线数仓的数据质量监控怎么理解的，怎么保证你的数据结果的准确性

（8）你觉得在大数据数仓开发过程中一个良好的开发流程是什么样的？我就说了一下代码编写规范，其他不清楚啊

（9）你们对模型的管理做了些什么？模型开发完之后怎么管理，版本的更迭这些，我不清楚了

（10）RDD 程序写过没有，问了 scala 的模式匹配，尾递归，为什么 scala 的变量不可变的好处

（11）说一个使用 flink 调度场景，flink 两条流怎么 join

（12）你们维度数据有没有生命周期？

（13）你们用什么组件监控 flink 这套架构的

（14）实时项目遇到哪些问题，怎么解决的，kafka 那一套讲一遍，flink 调优讲一遍

（15）大状态用什么存储的，rocksDB 选择第三方文件系统对效率有没有什么影响

（16）布隆过滤器了解吗？

（17）感觉一旦涉及好多实际工作中做的一些事情，没实际干过就真不知道啊

第81章 中科江南

（1）自我介绍

（2）离线介绍

（3）指标怎么完成

（4）数据重复怎么办

（5）oracl 这种传统的关系型数据库有了解吗？

（6）有跟客户接触过吗？

（7）你觉得你们这种公司内部的沟通和跟客户沟通有什么区别

（8）能接受出差吗?短期的

（9）有什么想问的吗

第82章 东信北邮信息

（1）对数据优化怎么理解的，实时项目串讲，flink 窗口

（2）hiveSQL 数据倾斜怎么造成的，什么现象，怎么解决的？

（3）怎么判断数据倾斜时任务卡在 99.9%时不是僵死而是数据倾斜？

（4）spark 的 shuffle 原理和优化，优化后有什么效果？快？快多少？

（5）hadoop 搭建的时候用的什么版本？具体什么版本？

（6）JVM 内存加载机制？打住，聊大数据

（7）Linux 在某个很深的目录下找某个文件，用什么命令？没用过 find

（8）hbase 的 hmaster 控制什么？

（9）用 clickhouse 主要做什么？

（10）换人问了，用户画像介绍

（11）标签的变更是怎么操作的？clickhouse 中的表结构是什么样的？

（12）hive 到 clickhouse 怎么入库的，spark 程序的具体代码逻辑是什么样的，用了写什么方法写出的

（13）spark 的版本是多少？spark 项目使用什么构建的，idea，项目依赖管理用的什么？

（14）maven 中打包时把某些包剔除用的什么？怎么解决 jar 冲突的？依赖 jar 包用什么，dependency

（15）spark 的算子有哪些，map 和 mappartitions 的区别，map 和 flatmap 的区别

左表是详细地址，右表是城市，统计各城市在详细地址中出现的次数？详细地址不能拆分，说思路，我说的按照城市分组，然后使用字符串包含的方法来计数统计

（16）提问，公司业务是移动旗下咪咕文化一整套互联网新媒体这种，每天数据量 PB 级别

第83章 求圣科技

83.1 一面

（1）介绍

（2）会不会 datax，数据怎么传的，他讲了个单词没听清

（3）kafka 应答机制 ack，你们遇到过的问题，还是自己背的

（4）问 kafka 分区机制，如果（3 个区，4 个消费者怎么消费遇到什么问题，死掉一个咋办）

（5）有没有遇到数据重复，，，，又问那 5 个事务怎么写的

（6）flume 拦截器怎么写的，能说下第一第二次的清洗，拦截

（7）flink 窗口，怎么触发，举个例子

（8）状态用过吗，说一下，优化呢（大状态调优）

（9）shell 脚本写过吗，我说了一个启停

（10）说下你们最大的表，多大

（11）flink 哪些 api,interverjoin 底层的为啥要 keyby，

（12）问了个 sql，简单

83.2 二面（突然进来个人，应该是组长）

（1）会不会 sql

（2）会不会 flink

（3）问数据量，他们 300T，

（4）hive 知道不

（5）自学的还是公司做的，你这咋这么多技能点

（6）平时啥活动，我说学习，研究新技术

（7）shell 会写不，python 用过吗

（8）职业规划

（9）业务转换能适应不，数据量大

第84章 中盾安信

（1）介绍实时数仓

（2）flume 采集日志的条数怎么确定不丢？

（3）Kafka 的数据积压，为什么会数据积压？

（4）Kafka 扩容是选择新增节点还是在原有节点上扩容，如果新增节点，执行负载均衡命令需要多久？

（5）Kafka 中设置增加吞吐量参数后，还有什么会成为瓶颈？CPU、磁盘、内存，又问 100T 和 4 各 25T 怎么选

（6）实时数仓中怎么进行 etl 的？详细说一下你们 etl 怎么做的，对哪些字段做了 etl

（7）窗口开了多大，sugar 多长时间刷新一次，这两个有什么关系吗？

（8）对数据库了解多深？熟悉存储过程、视图吗？

（9）你们集群规模多大，Kafka 每秒多少数据量，flink 单个并行度峰值处理能力？flink 跑了多少 job？多少台服务器上部署了 flink？并行度多少？

（10）datax 是干什么的，数据源是什么，目的地又是什么？

第85章 德特塞维

（1）自我介绍

（2）为啥来北京

（3）说下 hadoop

（4）mysql 元数据，原始数据

（5）spark，flink，区别，你们都用的啥

（6）有没有独立解决问题能力

（7）如何看待加班

（8）对象，家庭，薪资期望

（9）公司发展，职业道路

第86章 福卡斯特

（1）人事会大概调查你的公司，问一下你公司的情况，离职原因，为什么来北京，是否打算长期北京发展？

（2）自我介绍以及问所在大学？

（3）给自己 java 和大数据能力打分；

（4）自我定位与优势，最擅长的技术；

（5）公司产品、数据量以及并发量的相关问题；

（6）产品业务包含哪些？会有直播，录播这种业务？

（7）处理用户数据量是否都集中，服务器大概有多少？

（8）新产品和老产品维护你这边倾向那边？

（9）大数据的技术选型，怎么进行的选型？

（10）你这边怎么保证服务的稳定以及它的影响因数会有哪些？

（11）除了在程序和组件的优化，在硬件，数据库是否会进行进一步选择从而达到更好的稳定性？

（12）ClickHouse 存了多少条数据？

（13）一个 sql 题，查出重名的学生？

（14）JVM 内存模型？内存溢出报错的原因在哪？

（15）外部请求怎么通过服务器框架进行运转响应客户端？

（16）HashMap 里面放 100 条数据，初始化应该是多少？

（17）Redis 穿透了解吗？

（18）锁是否用过？项目中分布式锁是否用过？

（19）后台任务你们是否用过？

（20）你这考的证书对你的帮助是什么？

（21）你对自己未来的定位？

第87章 赛博云

业务 2B，工业互联网

87.1 面 1

（1）flink 优化

（2）hive 优化

（3）flink 定位反压，反压场景

（4）flink 全数据流一致性保证

（5）kafka 积压处理

（6）kafka 怎么有序

（7）hive 开窗

（8）hive 数据倾斜场景，处理，定位

（9）flink 比较难的模块怎么做的

（10）flink 遇到的问题，怎么解决

（11）flink checkpoint 超时解决

（12）flink 提交各个参数（容量大小，并行度，为什么这样设置），集群规模

（13）flink 和外部组件交互积压怎么处理的

（14）为什么要 DIM，能不能和 DWD 合并

（15）GC 日志怎么看

（16）flink buffer 怎么调，flink taskmanager 个数怎么确定，每个 taskmanager 设置多少个 slot

（17）checkpoint 和 safepoint 区别，怎么用的

（18）状态后端有哪些种类，选用种类，为什么选用

（19）flink 版本

（20）hive map reduce 个数怎么设置

（21）hadoop mapreduce 怎么写

（22）ck 监控

（23）ck 索引种类，原理

（24）hbase 二级索引

（25）redis 持久化方式，原理

（26）redis 基本类型

（27）redis 用过那些类型，怎么用的

（28）redis 集群用过吗

（29）java finally final 用法

（30）java 多线程 sychronized 是公平还是非公平，可以加在哪些上面

（31）spring boot 常用注解

（32）python Map 原理 能否 key 为 None

（33）集群规模，数据量，每秒数据量，每日数据增加量

（34）实时数据怎么落地？

87.2 面 2

（1）自我介绍

（2）讲熟悉的项目

（3）数据处理（脱敏、解密）

（4）kafka 乱序

（5）java 的 HashMap 底层、HashTable、多线程

（6）mysql 隔离级别、索引、b+tree 和红黑树的区别、redis 一致性（双写双删方面的）

（7）zk 擅长读、写？ 、zk 的 CAP

（8）hive 优化　　spark 会不会写

（9）hql 函数的优化　离线数仓解决的难题

（10）flink 时间语义、内存调优

第88章 爱普优邦（抖查查，爱盈利）

没有笔试，电话聊完直接线下面试

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

团队 4 人

人事面 大学经历

88.1 技术面

（1）kafka 乱序问题怎么解决，数据量太大非要多分区下游怎么处理？ 下游具体怎么实现

（2）bitmap 的优势劣势

（3）为什么用 flink 不用 sparkstrming，说说区别

（4）实时数据为什么放 ck，放 mysql 行吗

（5）说说项目里你遇到的问题

（6）spark 实时 es 给谁用的，用户还是内部

（7）做过什么指标

（8）ds 不是最近几年才火的吗，为什么用，一开始就用吗

88.2 项目经理面

第89章 济南中兴协力

（1）自我介绍

（2）熟悉 hsdoop 吗，介绍一下

（3）zk 了解吗，介绍一下

（4）scals/java 了解吗

（5）说一下你的项目经历

（6）如果一个零基础的人想学习大数据，你会推荐他怎样学习

第90章 百信大数据开发 - 外包

（1）自我介绍

（2）介绍离线数仓项目

（3）（讲项目过程中，会根据所讲内容随机提问，flume 挂过吗？kafka 的了解？整个离线有没有数据跑失败的情况？）

（4）SQL 题，字段 userid、loc、time，每一行代表一个用户在一个时间点所在的位置，求用户在哪个位置待的时间超过 30 分钟（每一分钟监控一次用户位置）

（5）实时用到哪些框架？

（6）ClickHouse 了解多少？为什么用它？引擎？是否采用副本了？

第91章 国广清科（谨慎）

91.1 面试 1

（1）自我介绍

（2）介绍的实时

（3）问为什么用 flink cdc ，如何实现的，用了 maxwell 为什么还用 cdc，他们俩有什么区别？

（4）flume 到 kafka 做了什么处理？

（5）数据从 kafka 到 hdfs 文件是什么格式的，怎么转换成 hive 表的？

91.2 面试 2

（1）自我介绍

（2）介绍项目 离线、实时

（3）log 到 HDFS 做了哪些工作 kafka 消峰具体原理

（4）dataX 底层

（5）maxwell 原理 maxwell 断点续传原理 出现故障数据重复怎么解决

（6）HDFS 小文件处理（我说 har 归档，问我具体怎么做的）

（7）CDC 比 maxwell 好在哪 DWD 层做了哪些事

91.3 面试 3

91.3.1 一面

（1）自我介绍

（2）介绍实时项目，主要负责哪几层，说一下项目架构

（3）为啥要使用旁路缓存和异步 io，旁路缓存和异步 io 怎么实现的

（4）大状态的优化是怎么优化的，状态太大了有什么影响

（5）checkpoint 太慢后有什么后果

（6）flink 的 checkpoint 和 spark 的 checkpoint 有啥区别

（7）cdc 底层是怎么实现的，配置表怎么放到广播状态中的

（8）平时使用 flinksql 还是 flinkapi，项目中哪里用到了 flinkapi

（9）flume 到 kafka 怎么搞的，采集时你搭建的吗

（10）你做了哪些指标

（11）你在项目中解决哪些问题，比较有成就感的问题

91.3.2 二面

面试官不是搞大数据的，主要面试基础怎么样

（1）自我介绍

（2）开发语言使用的是啥 java 吗？工作遇到过哪些棘手的问题，比较难以解决的问题

（3）异步 io 原理是啥，怎么实现，这个线程池多线程你们开了多少个，怎么设定的

（4）ArrayList 查找一个元素，不知下标，HashMap 通过 key 获取元素，那个快一点

（5）假设有一个数据 ArrayList 里面都是整数，然后给你一个整数 M，求找出两个数和等于 M

（6）Linux 熟悉吗，我想看服务器有几核 cup 是啥命令，我想知道现在跑了哪些程序用啥命令

（7）两台服务器，有一台服务器监听了 80 端口，如何在另外一台机器上看这台服务器是否监听了 80 端口

（8）内网服务器可以访问外网，但是没有公网 ip，那我应该怎么才能访问到这台服务器（我是内网穿透，他又问内网穿透原理）

（9）除了大数据框架，java，你还会哪些哪些语言，你会 python 吗，es 会吗，es 的分片是啥

（10） 最近 3-5 年，你职业规划，你觉得你在那些方面有提升

（11）你离职了吗，离职原因是啥

（12）git 你会吗，新建一个分支会吗

第92章 托特私募基金（笔试题）

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

92.1 python

没看；

92.2 java

（1）volatile、synchronized 两种锁的区别；

（2）给一个文件名和一个字符串，统计文件中出现该字符串的次数，编程；

（3）列举不同线程通讯的方式

（4）用 jdbc 读取数据库，如何提高读取效率？如何提高更新效率？

（5）读取一几千万的数据得文件，用 hashmap 加载，如何优化提高加载速度？

（6）有一个 10 亿条数据的文件，存储用户的身份证信息和手机号，如果某一用户手机号修改了，用 java 什么类操作？

92.3 mysql

（1）mysql 主从复制原理；

（2）数据库中常用锁及应用场景；

（3）mysql 查询语句 select 执行流程；

（4）写 sql，两个有相同结构（id 和 ttr）的表（A 和 B）join，若有字段为空值，按 AB 顺序取第一个非空值；

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

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

（5）S（DATE,STOCKID,STOCKNAME,AMOUNT,VOLUME），

IND（STOCKID,股票所属行业 ID）

SID（股票所属行业 ID，股票所属行业名称）

写 sql，建立合适的表索引以更方便的解决一下问题？

写 sql，历史销售最高的股票名称？

写 sql，股票数最高的行业？

写一个触发器，保证每日各行业股票数不少于 10 只？

（6）建立以 x,y,z 为联合索引的表；

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

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

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

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

92.4 大数据技术

（1）阐述 spark 提交流程

（2）checkpoint、cache、persist 区别；

（3）写出 spark 算子及作用；

（4）编程利用 spark 算子完成以下任务

（5）阐述 sparksql、rdd、dataset、dataframe…的区别及计算效率；

（6）什么是宽依赖、窄依赖？

（7）groupbykey 和 reducebykey 的区别？

算法一道：没看

第93章 可之技

（1）介绍上一家工作项目

（2）数仓的作用？

（3）离线数仓和实时数仓是同一套，还是两套？

（4）离线数仓的技术栈？

（5）离线数仓的承担的角色

（6）同步了那些数据，数据源

（7）埋点怎么埋的？

（8）数仓搭建完成后运行过程中的问题？

（9）kafka 宕机加何排查？

（10）kafka 数据重复怎么解决

（11）增量还是全量的选择

（12）什么数据是增量同步的？

（13）增量同步了，数振更新怎么办

（14）什么是拉链表

（15）同步过程中，业务那边业务数据库的表结构发生变化，怎么办？

（16）从上家公司的离职原因？

（17）你的期望薪资？

（18）在上家公司的有什么成长？

（19）比较喜欢找一家什么样的公司

（20）面试多久了

（21）你有什么问题想了解

第94章 国信创新

有个笔试，三道 SQL 题，很简单，分组聚和、开窗、两表关联

94.1 HR

（1）自我介绍

（2）之前工资、期望工资

（3）对自己的定位和职业发展规划

（4）对他们公司的简介

94.2 技术面

（1）ClickHouse、ES 问了一点

（2）其他也没问什么，自己讲一讲框架、组件啥的

第95章 有咖互动

95.1 面 1

（1）自我介绍

（2）离线和实时整体架构

（3）flink 的水位线介绍一下、背压、Redis 旁路缓存怎么实现的

（4）CK 介绍一下

（5）Redis 介绍一下（问了 Redis 的锁）

（6）Flume 挂了数据会丢失吗？原理是什么？后期对采集的数据有什么处理

（7）七天内连续三天下单（因为我简历写了）

（8）总之就会根据简历写的点去问；并且讲项目过程中，随时穿插会问一写细节和组件，主要就看自己怎么讲项目的

（9）还有几道他电脑上的题，直接展示，说出思路就可以

（10）注册表（id、dt）、登录表（id、dt），求每天的前三天注册的用户中登录的数量

（11）一个字符串，其中包含各种成对的<>、（）等，要求找出字符串中都是成对出现、且是嵌套关系的

95.2 面 2

（1）说说离线数仓架构

（2）说说 flume，组件之类的

（3）建模分几层

（4）维度建模，每层表都有什么

（5）怎么建的模

（6）问了谓词下推

（7）hive 底层

（8）hive 的分区，也就是 hdfs 分区

（9）拉链表实现

（10）hql 函数，开窗函数

（11）分桶表分区表外部内部表

（12）kafka 的 leader 选举，kafka 优化

（13）redis 写入数据的 api

（14）hbase 怎么均衡 region

（15）linux 查找文件，自带的内部调度器

（16）手写 sql 题，会有侧写或者行转列的一些函数

95.3 面 3

（1）Flume 读 log 文件，怎么知道读取完成了,读完的文件有什么变化 2

（2）Flume 的 source 有哪几种了解哪几种

（3）Flume 时间字段在 event 哪里存储

（4）Flume 有几台

（5）Datax 怎么配置的，在 hdfs 中配置了什么

（6）kafka 中 partition 存什么格式数据

（7）Kafka 数据存的数据大小

（8）维度建模流程

（9）数仓哪几层用的维度建模

（10）用没用过 Linux 的调度

（11）Linux 的命令有哪些

（12）动态分区了解吗，拉链表怎么做的

（13）你们部门多少人

（14）redis 了解吗

（15）Clickhouse 了解吗

95.4 面 4

（1）自我介绍

（2）项目串讲

（3）flume 的组件

（4）hadoop 架构

（5）kafka 架构

（6）事实表有哪些

（7）维度表有哪些

（8）业务总线矩阵怎么做的

（9）维度建模过程

（10）拉链表怎么做

（11）你做过哪些指标

（12）指标是谁提出的

（13）linux 常用命令

（14）3 个 sql 列转行 行转列 topn

第96章 军尊

96.1 面 1

（1）自我介绍

（2）介绍一下用户画像

（3）hivesql sparksql flinksql 各自的应用场景

（4）如果说我们想定义的规则类标签有时间维度的概念 比如说我们统计 10 天 但是数据是 11 天才来的数据 那么我们怎么进行处理

（5）如果数 对用户标签的话 那么中间几天这个用户可能没有操作行为那么原有标签是消除还是保留 还是其他处理 你们是怎么做的有没有额外的业务操作或者大数据技术方面的处理

（6）项目中做过哪些优化 技术或者框架

（7）flink 做过哪些优化

（8）flink 反压

（9）flink 下游有没有可能消费能力不足导致出现反压

（10）介绍 kafka

（11）kafka 新版本有什么特点，为什么要去掉 zookeeper 好处是什么，同类型的消息队列有哪些 如何比较的 为什么选择 kafka

（12）建议说多关注大数据技术的发展，多横向的比较框架，选出最适合的

96.2 面 2

（1）自我介绍

（2）技术栈介绍

（3）不用 flume 可以吗

（4）阿里云的四层能不能简化

（5）做过哪些优化

（6）为什么选择 kafka

（7）用没用到 zk

（8）不用 zk 可以吗

（9）Kafka 现在不是可以不用 zk 吗

96.3 面 3

（1）项目中遇到的问题

（2）介绍下用户画像

（3）spaksql hivesql flinksql 区别

（4）Flink 和 spark 区别

（5）实时和离线的区别

（6）小文件怎么处理的

（7）做了哪些优化

第97章 芯盾科技

（1）自我介绍

（2）介绍项目

（3）使用 java 实现 wordcount

（4）Kafka 实时中都做了什么

（5）用的那些 web（实时提交方式）

（6）Redis 存的什么

（7）Hbase 的 Rowkey 怎么设计的

第98章 云和互动

（1）自我介绍，

（2）根据简历往下捋，集群规模，磁盘大小，服务器选取，参与了哪些集群建设

（3）Hadoop 有哪些组件，hdfs 读写流程和 yarn 的运行过程，hdfs 操作命令

（4）分区和分桶表区别

（5）有没有遇到过磁盘坏块

（6）有没有遇到过节点数据不均衡

（7）hive 调优，小文件处理

（8）8Rdd,ds,df 的转换关系，

（9）宽依赖和窄依赖

（10）有没有使用过算法提高离线数仓或实时效率

（11）用过哪些设计模式

（12）自定义函数在哪应用的，我说 ip 转省份，他说用的列表？还是算法自己算的？记不太清了

第99章 国汽大有时空

（1）最近做的什么项目，我说 flink，介绍下项目的实施过程。

（2）flink 算子链合并的条件，

（3）flink 算子间的通讯方式，

（4）状态有哪些？存储位置

（5）checkpoint 和 savecheckpoint 区别

（6）checkpoint 参数配置，如果并行度改变，checkpoint 可以直接恢复吗？

（7）实时有多少指标？统计过的最大的指标是哪个？

（8）task 和 subtask、slot 的关系

第100章 橙啦教育

（1）自我介绍

（2）介绍离线

（3）Dws 是最终指标吗

（4）离线核心框架是？

（5）有没有指标管理平台

（6）项目经理通过什么给你们指标定义（需求文档还是白皮书）

（7）有没有基于指标做过多维度查询

（8）介绍用户画像

（9）宽表存在哪

（10）标签是如何分类的

（11）用户偏好标签的业务逻辑是什么样的，怎么实现的

（12）画像的上层应用是什么

（13）数据校验怎么做的

（14）数据治理怎么做的

（15）数据安全怎么做的

（16）跑批故障保障怎么的

（17）告警的级别是如何设置的

（18）有没有做过元数据管理

（19）Flink 为什么用 api 不用 flinksql

（20）Hive 优化

（21）Flink state TTL

（22）State 清除策略

（23）常用的是哪种 state

（24）反压的原因 影响 怎么解决

（25）水位线多对一 取哪个 原理是什么 咱们项目中 多对一 一对多 一对一 哪个用的多

（26）Flink 端到端

（27）剩下的就不是技术问题了 还问了薪资结构

第101章 华熙生物

（1）spark 和 flink 的区别

（2）flink 的核心

（3）flink 的架构

（4）flink 的 api 分几层

（5）flink 的时间语义 问 flink 的处理时间是不是服务器的当前时间

（6）说说数仓的分层

（7）为什么用 orc，orc 是列存还是行存

（8）flume 拦截器

（9）问了一下数据量，还有数据来源是 app 小程序 还是什么

（10）问对哪方面技术比较感兴趣

（11）手里几个 offer 都是什么价 公司是做什么的

第102章 佰钧成（vivo 外包）

vivo 那边负责面试，佰钧成只负责拉人，一共三面：两个技术面，一个 hr 面

主要去做 flink 指标开发和系统的升级

102.1 一面

（1）自我介绍

（2）hive on spark 中 sql 如何转换为 spark 任务的？我答的 sql 转 mr，然后说没有更深入的了解 spark 的原理。

（3）hive on spark 对 spark 做过哪些调优？

（4）kafka 的 offset 是全局唯一的吗？

（5）kafka 分区数为什么只增不减？

（6）spark 实时系统怎么维护 offset 的

（7）join 操作如何才能避免 shuffle？

（8）rdd、df、ds 之间的关系及转换？

（9）flink 实时和 spark 的实时的区别和共同点？

（10）flink 的 checkpoint 保存机制

（11）flink 使用中遇到过哪些问题？

（12）spark 的 stage 如何划分任务的？spark 如何查看任务数量？我回答的依靠血缘关系，她说这可以看，但是她说问的不是这个，就想看当前 stage 的任务数？没听懂她意思就过了

（13）hbase 在哪用的？我说的是 flink 维度表，又问了数据量多少？了解多少？底层原理是啥？做过哪些优化？hbase 存入数据的时候变慢了，该如何处理？

（14）elasticsearch 做过哪些调优？是一个指标一个索引吗？指标统计是在 es 里面完成的吗？

（15）离职原因？

（16）遇到问题一般怎么解决的？

（17）上家公司遇到你没接触过的技术是给你时间学习？还是你自己私下加班学的？

（18）对未来的规划，自己有什么优缺点？

（19）最后问我之前他们出的题怎么样？让我写一个算法，她来制定，当场写。链表反转。

102.2 二面

（1）自我介绍

（2）为什么 spark 实时换成 flink 实时？

（3）fink 挂掉如何保证一致性？

（4）flink 程序升级，或者任务数变了，还能从状态中恢复吗？ Savepoint

（5）你一般遇到问题如何解决的？

（6）你们服务器用的什么系统？linux。你编写 jar 包怎么传到服务器的？为了服务器安全。

（7）linux 的常用命令？

（8）xshell 用过吗？xshell 用的场景是什么？数据传输怎么传的？

（9）flink 或者 spark 是谁来维护的？你们数据量有多大？

（10）flink 或者 sparkstreaming 有没有优化的例子具体介绍下？介绍了下各省份订单量

统计和外部数据库读写优化？

（11）对 spark 有什么优化？

（12）现场用本地 idea 做一个简单的树的遍历。

第103章 多益网络（游戏公司）

（1）自我介绍

（2）问了下之前公司大概是干什么的

（3）问用户体量（就是大概有多少用户，数据大概有多少）

（4）为什么离职

（5）介绍一下你最熟的一个项目，你在里面主要做了什么

（6）Kafka 架构了解吗，描述一下，然后中间根据我说的问了 Kafka 的高吞吐

（7）Kafka 的高效读写方面问题

（8）Kafka 零拷贝了解吗

（9）Kafka 还做了哪些优化

（10）你之前说到小文件，你能说一下小文件的危害有哪些吗

（11）问 hadoop 部署大概是什么样的，几台服务器

（12）我们有部署高可用吗，大概怎么配置的

（13）七天连续 3 天的需求怎么实现的

（14）离线数仓有遇到数据倾斜吗

（15）Spark 了解吗，提交流程大概是什么样的，RDD 了解吗，用过吗，groupbykey 和 reducebykey

（16）Hbase 和 clickhouse 对比一下

（17）来了个开放问题：

（18）如果我有一个文件，里面主要是些数字，数据有几十亿设置几百亿，我想找其中数字最大的 100，怎么做；

第104章 铂原科技

（1）三年经验的问题

（2）离职原因

（3）介绍用户画像

（4）问怎么处理 3 亿用户 1000 个标签的问题 1000 标按照我们的画像处理 宽表有 1000 多列 可能表就崩溃了 怎么解决这个问题

（5）有没有 linux 的使用 能否使用 CDH

（6）介绍做的比较有难度的标签

（7）你们的标签一维度的标签，可否做二维的标签 二维的标签应该怎么设计宽表

（8）离线数仓都做了些什么 表是不是你设计的

（9）数据源的问题 日志数据是什么样的结构

（10）springbroot 的了解有多深 能不能进行现有 spring boot 代码的修改

（11）我们的标签是基于 sql 制作的 有没有什么办法能让不会 sql 的同事也能进行标签的只制作

第105章 时空云电话

（1）为什么是文件块大小是 128m？Hdfs 中平均寻址的时间？

（2）Hive 做了哪些优化？

（3）Hive 中窗口函数 RANK（）DENSE_RANK() ROW_NUMBER() 中的区别？

（4）Hive 中四个 by 区别？

（5）Hive 中行转列怎么做的？

（6）Hive 中字母小写转大写？ Lower upper

（7）CONCAT_WS、COLLECT_SET 的区别?

（8）hive 中如何取两位小数？round (column_name,2)

第106章 共致开源

（1）介绍项目 离线

（2）数仓分层 各层的作用

（3）出现过任务失败的这种情况吗 也就是框架优化和 hive 优化

（4）离线的模式？T+1

（5）调度任务用的是

（6）依赖关系 用的是什么元数据管理

（7）介绍实时

（8）Spark 和 flink 区别

（9）Spark 模式

（10）Redis 是用来做什么的

（11）Kafka 数据积压

（12）数据丢失

（13）画像

（14）Clickhouse 好处

（15）数据迁移会做处理吗

（16）Hive 到 clickhouse 为什么没用 datax 这种同步工具来同数据

（17）复杂的标签 都是写 sql

（18）离线和标签不能合并使用

（19）项目中遇到问题 反压和数据倾斜

（20）做没做过数据质量管理

第107章 VIVO 外包

（1）自我介绍

（2）介绍一下你们离线数仓项目的组件

（3）请你说一下 hive 的优化

（4）为什么要做实时数仓

（5）工作中有没有遇到 flink 反压 及 数据倾斜等问题，是如何解决的？

（6）为什么使用 clickhouse

（7）请你说一下你所遇到的离线数仓的困难点

（8）使用 flink api 代码过程中遇到过哪些问题？

（9）spark 的任务划分

（10）dataframe 与 rdd 的区别

（11）如何调整 spark task 的数量

（12）小文件有哪些危害（两个方面）

（13）谈谈你对 kafka 的理解

（14）谈谈你对 checkpoint 的理解

（15）为什么要用 hbase 保存维度数据，了解 hbase 的 rowkey 吗？hbase 读写流程了解吗？

第108章 能链集团

（1）自我介绍

（2）flink 实时做了多长时间？

（3）flink 用的 sql 还是其他的？

（4）实际做过的指标有没有遇到反压，怎么解决的？

（5）flink 的内存机制？

（6）电商遇到退单在实时中怎么处理的，我说开了一个新的退单流。他问假如想统计当天的实际交易额怎么做？

（7）你们的实时数据量有多大？实时中业务数据量有多大？

（8）数仓用的什么架构，做过哪些 hive 优化？

（9）orc 和 parquet 的区别？

（10）手写 sql，统计连续 10 天登陆的用户？

第109章 环球网校

（1）介绍项目

（2）kafka 怎么做的分层

（3）是存到不同的 topic

（4）ck 有搭建过集群吗

（5）组件是用什么安装的？apache

（6）用什么管理的 ？cdh 吗

（7）flink 集群模式 和 yarn 有什么区别

（8）用过 cdh 吗

（9）了解过 mery 吗

（10）cdh 能管理哪些组件

（11） pstio 和 hudu

（12）hbase 作用和适用场景

（13）phoenix 建表吗

（14）phoenix 默认端口

（15）namenode 和 datanode

（16）hadoop 高可用怎么实现

（17）hivesql 窗口函数 实现 topn 用哪个函数

（18）bitmap 使用场景

（19）ck 宽表 有扩充的情况 会改结构吗

（20）redis 默认的 db 有多少个

（21）离职原因

（22）集群规模

（23）impala

（24）离线有 bi 系统吗

第110章 贝壳外包

（1）说一下你最近做的一个项目把

（2）为什么需要 dws 层，为什么不可以从 dwd 直接到 ads 层？

（3）dwd 为什么不可以去掉？

（4）业务总线矩阵怎么构建的？

（5）数据量是多少？

（6）什么样的数据放在 dim 表，什么数据放在 dwd，dws？

（7）为什么用户放在 dim 表里？

（8）什么是退化维？

（9）指标数据都是通过 dws 出的吗？

（10）如果你们跨了各种数据域的数据放在哪里？

（11）手写拉链表 sql

（12）手写用户首次登录的城市 sql？

（13）了解三范式和维度建模吗？

（14）说一下三范式

（15）说一些维度模型

第111章 青岛群之脉

（1）日常写什么代码。

（2）java 基础范型

（3）hive 工作原理

（4）调优经验

（5）对 spark 了解吗？执行过程

（6）项目中遇到的难题，如何解决

（7）翻译英文，（一段话）

（8）sql 窗口函数

（9）shuffer 优化

（10）clickhouse 优缺点

（11）hive on spark 怎么做的

（12）项目 dwd dws 区别

（13）数据库

（14）前端可视化工具

（15）Linux 常用命令

（16）代码如何管理

（17）Git 命令

（18）编程方面的书籍

（19）spark 有哪些概念

（20）技术短板

（21）职业规划

（22）加班态度

（23）py 写过没

第112章 先智数元

（1）哪些自己单独完成

（2）业务数据过滤关联字典表

（3）双流 join 怎么匹配

（4）dim 层配置表，怎么写入 hbase

（5）interval join 时间范围

（6）cep

（7）端到端一致性

（8）8．检查点

（9）下游怎么实现幂等性

（10）异步 io?数据库支持吗？

（11）旁路缓存双写？

（12）driver 端执行，ex 端执行？

（13）13flink redis 在哪里写代码

（14）foreach 在哪里执行

（15）算法

第113章 中信百信银行

113.1 面 1

（1）数仓结构；

（2）各层有哪些表；

（3）dwd->dws 中间还有没有分层；

（4）离线数仓怎么保证数据质量；

（5）怎么验收数据；

（6）哪些数据域，各个数据域包含那些表；

（7）遇到比较难解决的建表问题

113.2 面 2

（1）连续三天碳排放小于 100

（2）用户连续登录最大天数

（3）纬度建模四个步骤

（4）（我给个提示你啊…，选择业务过程）

（5）事实表分哪些

（6）flink 检查点流程

113.3 面 3

（1）说一下你们的数据埋点是怎么实现的

（2）hive 的调优

（3）sql 条件 where 和 on 的区别

（4）表与表的连接经常习惯用那个，如果遇到 rightJoin 的话会转成 leftjoin 来使用吗

（5）通过 sql 实现在用户访问表里查找连续六天以上访问的用户数量

第114章 上海玛驷众智能科技有限公司

（1）hive 数据倾斜遇到过吗，怎么解决的

（2）手写 topNsql（注意会屏幕共享，桌面上不要放见不得人的东西）

（3）数据采集整个流程

（4）FlinkCDC 版本（要用 X，X 会锁表）

（5）第一级 flume 用的什么 source

（6）taildir source 的断点续传底层原理

（7）kafka 数据积压如何解决

（8）2 级 flume 的小文件问题

（9）数仓分层

（10）用过 flinkSQL 吗，在哪儿用的

（11）ADS 层文件存储格式

（12）kafka 有多少分区

第115章 航空信息股份有限公司

（1）简单做个自我介绍

（2）说下 hdfs 文件系统缺点 3 点

（3）说下 MapReduce 实现 jion 思想

（4）用户拉链表怎么实现

（5）hiveSQL 和 MySQL 有哪三点不一样

（6）说下 hbase 如何删除表

（7）说下对 hbase 中数据使用 SparkSQL 分析

（8）三个词总结下自己

第116章 锐捷

（1）最近做过啥项目

（2）说说 flink 提交参数都有哪些

（3）flink 在集群里面怎么部署的

（4）设置了多少并行度

（5）kafka 多少分区

（6）我说了 40 多个，问 slot 数量太少对于 40 个分区会不会消费能力不足造成数据积压

（7） 为什么设置 kafka 分区数 40

（8）还问了 DS 版本

第117章 克拉星球，有咖互动

（1）mr 流程及优化

（2）yarn 调度器

（3）维度模型

（4）统计过哪些指标

（5）Hive 的是个排序

（6）Hive 函数 3 个 rank,over 里面有哪些东西

（7）手写七天内连续三天登录

（8）说一说 kafka

（9）zookeeper 的作用

（10）10．数据一致性

（11）11．Kafka 事务加幂等性

（12）Zookeeper 选举机制

第118章 元拓公司

讲项目

（1）离线自己做的吗，是你们公司自己用的吗，有自己的网站吗

（2）采集离线实时是一套吗

（3）讲一下采集

（4）小文件产生的原因

（5）离线数仓面向那些业务

（6）最终的业务数据放在哪里

（7）mr shuffer 流程 调优的点

第119章 中软信息系统工程

119.1 面 1

（1）讲一下实时项目，带着业务讲

（2）每层怎么做的，有哪些事实表，那些指标

（3）MySQL 索引优化

119.2 面 2

（1）自我介绍

（2）离线数仓项目介绍

（3）为什么要分为 5 层

（4）datax 架构

（5）sql 调优

（6）sql 题：

a) 有哪几种方法去重

b) 排名函数有哪些

c) 取前 20%的数据有哪几种方法

d) 用户留存率

e) 侧写函数实现

第120章 上海博彦科技

（1）用什么组件开发

（2）创建线程池的方式

（3）java 三大特性

（4）根据三大特性，开发过什么

（5）hashmap 结构

（6）做过后端开发吗？

（7）说下隐式转换

（8）说下闭包

（9）传参方式有哪些

（10）spark 出现的 oom 场景

（11）共享变量有哪些，原理

第121章 京东外包

（1）hive 针对 shuffle 的优化？

（2）hive 的 shuffle 和 spark 的 suffle 的区别？

（3）repartition 和 colesce 区别？repartition 是在哪儿执行的，shuffle 内还是 shuffle 外？为什么

（4）场景 sql：

a) 下图

b) 去重的几种方式？(我回答的是 distinct 和 group by，要三种实现方式)

c) 两种方式实现前百分之 10 统计

d) 连续一天登录，连续三天登录？连续间隔 n 天登录？

（1）Flink 做过的优化？

（2）saprkStreaming 和 flink 在处理业务的时候有什么场景的区别？举例说明

（3）离线数仓的统计结果和实时数仓的统计结果是如何比对的？差多少？

（4）多流 join 你用过吗？你们是怎么用的？

（5）用过窗口函数的区别，自己实现过窗口函数吗？

（6）flink CEP 怎么用?

（7）7 迟到数据怎么处理？依据什么调整窗口时间？

（8）interval join 用法？举例说明你们 interval join 使用的场景?

第122章 东方国际

（1）Hive AB 表关联，A 表 10 条 B 表 5 条，关联后怎么有 12 条

（2）讲一下 hive 优化

（3）Java

（4）维度建模方法论，每层干什么

（5）实时熟吗?怎么办证数据准确数据及时

第123章 柯莱特美团外包

（1）离线串讲

（2）日活，年活、数仓数据总量

（3）部门人员划分、日常工作内容

（4）数仓数据质量如何保证（上线前、上线后）

（5）负责哪部分内容（数据域）

（6）最细粒度，概念

概念: 粒度是指数据仓库中数据单元的细节程度或综合程度的级别。说白了就是表示某一个事实表中的一行数据是什么

第124章 四川华为运维

（1）hadoop 默认副本

（2）副本如何写到几个 datanode

（3）flume 到 kafka 到 flink 到 hbase，9 点半后，我数据查不到了，原因？

（4）某几个 datanode 磁盘空间不足的原因

（5）现在提交不同的应用到 hdfs，有的要 2 副本，有的要 3 副本，怎么指定？

第125章 中华财险（外包）

（1）HDFS 读写流程

（2）yarn 提交

（3）三个调度器，你们使用的那些

（4）用户拉链表实现流程

（5）数仓建模分几层

（6）flink 迟到数据怎么办？

（7）spark 有 shuffer 的算子。

（8）spark 有排序的算子

（9）spark-shuffer 了解吗

（10）hadoop 优化，shuffer 优化

（11）hivesql 优化

（12）你为什么统计个省份下单量和下单总额，指标给谁使用，要做什么

（13）统计个省份下单量和下单总额，用到各层哪几张表？

（14）一个 sql

商品交易表

商品 id 交易日期 当日总销售金额 当日总销量

求连续七天销售商品 id

第126章 中科特瑞

介绍一下自己离线

2.kafka 有多少分区

3.数据量有多大

4.数仓建模理论

5.能不能手写斐波那契数列算法

6.Java 用的多吗

7 手写连续三天登录，两天存活

8.平时喜欢干什么（

第127章 中科软科技股份有限公司

（1）在项目什么地方用到 redis

（2）怎么保持 redis 与 hbase 数据一致性

（3）flink 怎么保持数据一致性

（4）在项目中，你具体负责哪一部分

（5）做离线是基于什么库

（6）做实时的时候比较有挑战的地方？

（7）flink 断点续传

（8）kafka 数据保存时间短，怎么实现断点续传

（9）kafka 多个主题的数据需要 join

（10）flinkSQL 在 join 时数据迟到怎么处理的（你说的只能解决 10s 内的迟到情况）

（11）kafka 一个 topic 变动，一个不变，一条流会有大量数据过来，flinkSQL 怎么关联

（12）kafka 分区增加的话，是要重建 topic 吗，数据怎么流入新分区？

第128章 博亚君杰

（1）hbase 部署在十台集群上面，挂了一台，数据会丢失吗？

（2）维度建模

（3）你负责那一块？

（4）flink 你写那块功能，讲讲？

（5）项目中遇到的问题

（6）你还有啥问题

第129章 速通科技

129.1 面 1

（1）讲一下项目(数据采集他们还没做后期考虑做没有要问我的)

（2）纬度分层，为啥分五层，每一层都做了啥

（3）sql 题:

找出高速公路收费站同一收费口对同一辆汽车进行两次扣费的数据

三分钟内同一车道扣费的车辆数

129.2 面 2

（1）进一下你得项目

（2）为什么要用 maxwell。

（3）sql

第130章 浙江御安

（1）介绍一下 flink 项目

（2）flink 反压怎么解决

（3）flink 检查点说一下，

（4）分桶表与分区表的区别

（5）kafka 数据积压了怎么办？

（6）hive 的排序方式

（7）hive 的优化，还有你遇到的数据倾斜，groupby join

（8）离线数仓做了多久，都是几个人参与的，你们的服务器有几台。

（9）未来大数据职业规划。

第131章 军民融合舰船装备

（1）3-5 分钟自我介绍

（2）熟悉哪些组件

（3）对数据治理这块有了解吗

（4）了解过 mpp 数据库吗

第132章 外包速科

（1）物化视图

（2）clickhouse 副本机制

（3）olap 数据库和 oltp 数据库的区别

（4）为什么搭数仓，MySQL 加磁盘也可以做

（5）clickhouse 的 replicingmergetree 会有重复，如何解决去重

（6）lead 函数

（7）row between 函数

第133章 亿达信息

（1）介绍离线项目

（2）Maxwell 怎么同步的

（3）漏斗分析你们怎么实现的

（4）外部表和内部表

（5）小文件的处理

（6）数据倾斜

（7）第二级 Flume 里 HDFS Sink 配置小文件的参数

（8）怎么把数据写入 ODS 层

（9）insert into 和 insert overwrite 的区别

（10）怎么删除 ODS 分区里的污染数据

（11）maxwell 和 DataX 怎么配合工作的

（12）Maxwell 工作原理

（13）Flink 怎么处理乱序、数据

（14）CEP

第134章 合众

（1）讲一下离线数仓吧

（2）hbase 在你离线数仓中怎么用的？

（3）java springcloud 了解吗

（4）为什么用海豚调度器？为啥不用 java 中的 querz

（5）你们的指标 ods 层到 dwd，怎么查的巴拉巴拉记不清了

（6）flink 怎么处理乱序数据

第135章 上海智租物联科技有限公司

（1）19 年毕业？不是读大二吗？（出来实习了）

（2）说下你熟悉的项目，并统计过哪些指标

（3）hbase 的 rowkey 设计原则

（4）你自定义线程池为啥自定义，为啥不用 java 内置

（5）FlinkCDC 锁表问题说一下

（6）怎么处理 Flink 数据乱序和迟到的，极端情况下如何处理

（7）离线建模方法说一下

（8）说一下 JVM 结构

第136章 灵信互动外派中金证券 ETL 岗位

（1）ETL 的三个阶段：分别是数据抽取、数据转换、数据加载

（2）拉链表的实现

（3）如何保证 ETL 的数据一致性

（4）表的连接方式有几种

（5）说一下 NVL 函数

（6）说一下全量抽取和增量抽取

（7）日志数据分几类

第137章 百融云创

（1）手写熟悉的设计模式

（2）7 天连续登录

（3）行转列 字段类型 班级，学生姓名（个数不定） 成绩

（4）二叉树前序遍历

（5）二分查找查找

第138章 新桥信通

（1）讲一下你熟悉的项目

（2）拉链表怎么做的以及装载思路

（3）你们全量同步怎么做的

（4）maxwell 会一直监控 binlog 文件嘛

（5）你们的 binlog 文件保存多久

（6）Oracle 用过没

（7）datax 怎么用的

（8）你写一个指标要多久

（9）你在项目里负责什么

（10）还有两三个问题，记不清了

第139章 苏州盈天地

（1）采集到 hdfs 中小文件怎么处理的

（2）mr 中可不可以只用 map 或者只用 reduce

（3）mr shuffle 和 spark shuffle 区别

（4）spark 中的 shuffle 算子 join 会 shuffle 吗

（5）hbase 了解吗

（6）clickhouse 了解吗

（7）flink 了解吗 说了一堆 他说我们用不到就不问了

（8）java 中的集合说一下

（9）hashmap 的 jdk1.7 和 jdk1.8 的区别

（10）arraylist 说一下

第140章 招商新智（就一面）

（1）讲一下项目，我讲了 15 分钟

（2）怎么知道你们的数据符不符合要求，

（3）说一下数仓建模怎么建的，五层分别作了什么。

（4）然后就聊了他们的情况