2022 年深圳大数据面试题汇总

（作者：尚硅谷研究院）

版本：V1.0

第1章 字节面试题

（1）介绍项目

（2）写了一道算法题，题目是：驼峰命名转换 。

（3）二叉树打印

（4）用 List 和 Map 实现 LRU

（5）mysql 的引擎有几种，各有什么不同

（6）mysql 事务级别

（7）mysql 事务级别如何实现的

（8）有一个完全二叉树，给定其中两个节点，代码实现：两个节点哪个是父节点或者其父节点是什么？

（9）数据结构会问的比较多

（10）集合有哪些？ArrayList 和 LinkedList 有什么区别？

（11）讲一下最近的一个项目（根据架构开始问业务难点）（我的是离线数仓项目）

（12）hive 两个表 join，过滤条件在 on 后边和 where 后边有哪些不同？

（13）碰到哪些问题，怎么解决的？hive 做了哪些优化，优化后有哪些不同？

（14）问了挺多 hive 中的参数设置，大表 join 大表的解决思路，数据分块，还有一个是离线数据怎么保证及时性

（15）做了个列转行然后行转列的题，我说了两种解决方案，第二种方案，有一个函数没记清，拼接 concat_ws 函数

（16）java 多态

（17）hive map side join

（18）hive distribute by

（19）flink checkpoint exactly once

（20）mapreduce 流程

（21）mapreduce 100 亿个数取 top10

hive sql：

（22）计算连续登陆超过 5 天的人

coding：

（23）实现微信随机红包

（24）找到根节点到叶子结点值之和为 n 的所有路径

（25）clickhouse merge tree 特性，distributed 是否保存数据

（26）mysql 索引，联合索引 abc，查询条件里 abc 的索引会命中吗

（27）mysql+redis 保证数据一致性

coding：

（28）两个链表存在交点，找这个交点

第2章 字节面试题-2

（1）项目经历 3 个挨个问，flink 的机制，乱序处理，业务架构

（2）数据仓库构建基础，spark，hive 的执行过程，调优

（3）BI 系统的架构设计，负责模块

（4）sql fulljoin 的功能的实现

（5）rownumber 的底层实现

（6）输入正整数 n 和 k，n≥k，找出[1,n]范围内按照字典排序的最小第 k 个值。

（7）项目经历，细节

（8）sparkstreaming 如何实现容错，遇到问题怎么解决

（9）spark mr 执行过程，hdfs 读写过程，出错处理等

（10）判断是是否是 BST

（11）hive 的 range 分区

（12）二叉树的广度深度遍历

（13）kafka 有序

（14）shuffle 过程

（15）合并两个不同规范的城市表

（16）微信红包

（17）连续 5 天登陆

（18）一致性 hash

（19）hashmap

（20）连表倒数第 n 个节点

（21）排序 K 个大小为 N 的数组

第3章 字节面试题-3

（1）自我介绍

（2）flink exactly-once

（3）flink sql：count distinct 求 uv 场景，同一个设备刷了很多记录

（4）算法题：最长子数组和

（5）算法题：给一个矩阵，矩阵每一行单调递增，求矩阵第 K 个数

（6）结合项目问一下离线实时 sql 优化问题

（7）sql 题：新访问，留存指标如何计算

（8）介绍项目，以及项目中的一些优化情况

（9）介绍项目，及在项目中碰到的问题，承担的角色

（10）为什么离职，是否有收到其它 offer

（11）职业规划

（12）Spark 与 MapReduce 速度差异的原因

（13）Spark 任务和 Spark Streaming 任务的差别

（14）项目相关

（15）Linux 的一些基本操作指令。

Coding：

（16）两个有序数组的合并（二路归并）（编写代码、分析复杂度）

（17）拓展，K 个有序数组的合并，说说解决方案，分析复杂度。

（18）Java

（19）先大概说了下项目的情况

（20）问题：两个服务端的接口，有这样一个定义：平均耗时=（全部请求的总处理时间）/（请求次数），两个接口经过了各自的优化，使得每个服务接口的平均响应时间都变短了一点，再看整个服务端的时候，发现整个服务端的平均响应时间变长了（假设该服务端是有这两个接口），这个怎么理解。

（21）Java 平时开发用到什么数据结构，如 MAP 有哪些

（22）HashMap 的原理，JDK1.8 版本后做了哪些优化，为什么要这么优化。

（23）JAVA 开发遇到过哪些报错，比如内存溢出，怎么解决。

（24）sql 问了一个开窗函数，一个 join 函数和一个日期函数，别的都是介绍项目经历了

（25）归并排序代码，

（26）kmeans 算法 gbdt,lr,lstm,tansformer,fm,fmm,deepfm,算法题问了下原理，损失函数之类的

（27）搜索二维矩阵；

（28）循环链表找入口节点，要求无额外空间复杂；

（29）智力题：25 匹马，每次只能比较 5 匹的快慢，问至少比较多少次可以得到最快 3 匹；

第4章 字节面试题-4

（05 学长提供）

（1）讲了实时项目过程

（2）为什么有 dwm 层，这样不是增加运维成本？

（3）dws 层做了哪些聚合？有多少张表

（4）clickhouse 每张表有多少行记录，保存多久？

（5）访问 clickhouse 的表延迟是多少？

（6）数仓有什么数据，订单的数据吗？

（7）订单金额实时跟离线金额差别多少？

（8）clickhouse 重复数据怎么处理？

（9）flink 的精准一次，checkpoint 机制，两阶段提交，

（10）部门有几个人，怎么分配，实时里主要负责哪一块？

（11）flink 使用 java 多还是 sql 多？

（12）实时框架是怎么选择的？

（13）毕业时间，有没有找了别的工作？

（14）最后在算法题和 sql 题选一题来做，选了 sql，然后没做出来，很难受。

orders 表很大，5000 万订单，说出这个 sql 有什么问题，写出的的优化后的 sql。

```
select province, count (distinct  buyer_id) from orders where
date=20211001 group by province;
```

第5章 袋鼠云

（1）interval join 不上的数据，怎么处理？怎么做数据修复？

（2）maxwell bootstrap 的同时，mysql 在变化，怎么保证写到 hbase 的数据是正确的？

（3）flink 发生撤回流时，UDF 函数失效怎么办，比如级联 group 下层用上层，就会失效？

第6章 腾讯面试

（1）各种场景模拟，1 亿数据，20 亿用户，大表处理

（2）如何确定某个 java 程序将 CPU 跑满之后导致该原因的线程？

（3）Phoenix 和 HBase 是如何关联的？一个 SQL 是怎么同步写入到 HBase 的？

（4）一年的窗口怎么每 5 分钟输出一次结果。（RK 状态+CK 时间点+定时器）

（5）flink 和 spark Streaming 对比？

（6）flink 的时间语义，精准一次性这些

（7）HBase 的二级索引？大表的处理？

（8）spark SQL 和 hive on spark 有什么区别？

（9）flink 的精准一次？幂等和事务使用有什么问题？ck 连续超时有什么影响？

（10）hbase 如果问怎么快速检索 1 亿数据？布隆过滤器已经开辟了最大长度，不能在开辟了怎么处理？

（11）spark 场景题和对应的优化处理？

（12）flink 事务 sink 写文件，是怎么处理的？对应的文件存在哪里？详细说下

（13）一个 flink 任务一直在重启？怎么判断原因？

（14）CBO 和 RBO 的区别？说说常见的 RBO 的规则？

（15）flink 的反压你们是怎么处理的？怎么判定的？

（16）flink SQL 的整个底层执行过程

第7章 华为

电话面试

（1）你们的业务是什么？

（2）你最近的项目是什么是用 flink 吗？

（3）你能给我讲讲你处理过最难的指标吗？并且该指标的作用是什么？

（4）你说你使用过分流，那么你给我讲讲是怎么分流的吗？

（5）你做分流的时候，用到了什么算子。

（6）你能讲讲 JDBC 连接的整体流程吗 ，以 select 为例，并且关闭顺序是什么。

（7）你给我讲讲你们的业务吧，可以细讲一些指标吗？

（8）ArrayList 和 LinkList 的区别？

（9）你使用过 SpringBoot，你给我讲讲业务场景

（10）前端埋点，你给我讲讲，你扮演了什么角色。

第8章 百度

（1）项目中为什么考虑用两个平台来做

（2）为什么你们白天做实时，晚上做离线，白天有离线任务吗，晚上有实时任务吗

（3）你们两个平台统计的指标有什么不同

（4）你们数据量多大

（5）有多少表

（6）表都存在 hive 里吗

（7）实时数据存在哪里

（8）解释下 ods，dwd。。。。是什么意思

（9）说一下你的数仓建模

（10）为什么不在导数据到 ods 层前对数据去重

（11）数据在前面去重会有性能上的影响吗，为什么

（12）你们整个离线任务是怎么调度的

（13）任务调度出现异常重跑会不会出现数据的重复

（14）数据重复了怎么做

（15）你们 azkban 调度除了 hql 任务还有其他任务吗

（16）给个场景，azkban 调度的过程中我想把中间某张表的数据下载到本地，怎么做

（17）你还用过 azkaban 做过哪些任务的调度

（18）你们的 kafka 设置了多少 topic

（19）你们写入 kafka 的数据是什么格式的

（20）多少分区

（21）你们 kafka 集群多少，做高可用了吗，为什么可以做高可用

（22）kafka 高可用原理说一下

（23）shuffle 流程

（24）mr 的 shuffle 与 spark 的 shuffle 有什么区别

（25）为什么 group by 比 distinct 高效

（26）有 3 个 key10 个 reduce，这 3 个 key 会分到 10 个 reduce 上吗

（27）用 hql 写个问题

（28）hive 有哪些保存元数据的方式，除了 mysql

（29）flink 集群角色有哪些，flink 时间机制

（30）你刚才讲 shuffle 时提到了快速排序，可以写一个快速排序吗

（31）你 java 大概是什么水平，java SE 是什么

第9章 OPPO

（1）hdfs 读写流程，数据倾斜怎么解决，小文件问题怎么解决，

（2）yarn 资源调度原理，

（3）flink 资源调度原理，

（4）flink 容错机制，flink 数据倾斜，flink 反压和失败怎么解决，flink 内存机制，

（5）mr 的 checkpoint，

（6）kafka 为什么那么快，

（7）java 熟悉程度，

（8）Kafka 精准一次怎么做的，

（9）zk 选举机制，zk 选举时机，

（10）hive 小文件问题，hive 优化，hive 元数据管理，

（11）flume 三大组件，flume 事务流程，flume taildir source 断点续传，

（12）redis 缓存穿透，怎么解决缓存穿透，

（13）spark 为什么那么快，spark 比 mr 好在哪，spark drive 角色设定，spark 的并行度的设置，

（14）es 倒排索引，es 怎么用 api 找出年龄在给的范围内。

第10章 VIVO

（1）hdfs 的读写流程，为什么数仓要分层

（2）每层是根据什么确定的

（3）公司的业务流程

（4）为什么要用 kafka，和作用。你理解 kafka 应该是做什么的

（5）用的组件版本

（6）做过的最难的指标

第11章 VIVO-2

（1）spark 问的多，任务划分，执行流程，

（2）spark 调优，spark 数据倾斜；

（3）hive 调优，窗口函数，数据倾斜；

（4）实时维表存在 hbase，怎么查询列簇下的列名？

（5）说 hbase 集成 phoenix 二级索引好像目前不太好用；

（6）为什么用 clickhouse；

（7）sqoop 导数据失败怎么解决；

（8）flume 事务监控；

（9）Kafka 事务监控；

（10）jvm，新生代，老年代，full GC

第12章 虾皮 shopee

（1）java 的锁了解么？公平锁、非公平锁，偏向锁和非偏向锁？（纳尼？）

（2）volatile 关键字了解么？怎么用的？主存是怎么存的？那对应的非主存是怎么处理的？

（3）说说零拷贝的原理？详细的说下

（4）计算机原理的一些内容、为什么要分用户态和 core？

（5）https 的通信机制？怎么建立连接的？（好像是这么问的）

（6）信息编码的意义？

（7）SQL 的预编译的处理的底层原理了解么？

（8）为什么 C/C+ 用来写 ClickHouse、redis、Zookeeper 这些组件？和 java 有什么区别？你怎么看待

（9）说说内部排序算法的时间、空间复杂度和对应的稳定性。（没说全，有点遗忘）

（10）HBase 的读流程

（11）redis 的缓存击穿、缓存雪崩、缓存失效是什么意思和如何处理？

（12）布隆过滤器原理和 kafka 为什么快这些？

（13）内部表和外部表的区别？分桶表的原理？一些简单的优化

（14）HQL 的处理流程，B+树和 LSM 的区别？分别讲讲他们的特点？为什么 MySQL 用 B+树，HBase 用 LSM？

（15）你们数仓的建模?(对应每层的处理)

（16）留存率你从 ods 到 ads 说下各层都怎么处理获取的？最终的 SQL 要能用语言表述清除？（建议结合自己的业务）

（17）如何快速从 mysql 导数据？离线：sqoop 实时：CDC

（18）kafka 的 producer/consumer 可能会出现的问题？丢和重复，怎么避免，怎么解决？

（19）scala 的 val 和 var 各自的优缺点？为什么用 val？场景设计一堆

（20）scala 常见的集合？（可变和不可变）

（21）scala option 的底层原理是怎么设计和实现的有了解吗？

（22）flink 的精准一次性，两次事务详细说说

（23）flink 对于多个流的 join 是如何保证同时处理到的？

（24）很有其他 flink 窗口相关的问题，具体的有点忘记了，就是各种异常场景和大状态的问题

（25）一小时的数据 IP，（数据量很大，）怎么得到 top10？（只说方案和具体的实现，不敲代码）

flink 有些内容有点没回答好，java 相关的 JUC 和 LSM 忘记了。

第13章 美的

自我介绍

（1）flink cdc 怎么实现同步增量数据和全量数据，底层区别是什么？我说同步原理是 binlog 主从复制 balabala，他问同步增量数据和全量数据底层有什么不同，

（2）kafka 一般比较大的表有 10+亿数据我们希望写到不同分区，flink 消费的时候怎么实现有序，有什么方法？

（3）kafka 如果分区增加，flink 怎么在程序不停的情况下增加到同等的分区，底层怎么做？

（4）rockDB 底层内存刷写磁盘原理?

（5）实时项目开发中有遇到什么问题？jar 包找不到、反压、数据倾斜讲一遍

（6）你们 flink 实时运行的有多少条流？我说 10+个 job，他问所有 job 加起来里面有多少条流你做开发应该很清楚，我说 30+条

（7）你们离线最大的表有多大？dwt 用户主题，最宽，但是只存今天和昨天数据不大

（8）你们每天数据量？1 亿左右，确定吗？确定！

其它大保健问题忘了

第14章 顺丰

没有问业务

（1）javaJVM 的优化

（2）你了解有哪些算法吗（然后换了个说法，说你了解有哪些排序算法）

（3）链表结构了解吗

（4）平衡二叉树了解吗

（5）kafka 的 ack

（6）kafka 怎么保证精准一次消费

（7）kafka 支持事务吗

（8）hql 熟吗

（9）内部表和外部表有什么区别，什么时候使用内部表，什么时候使用外部表

（10）你们在使用 hql 的时候做了哪些优化

（11）你们的数据量有多大

（12）你们有采用分区吗

（13）你们的分区策略

（14）分区太多和太少有什么问题

（15）你跑的最慢的一个查询是多久

（16）你在写 hql 的时候有没有做过什么优化

（17）有没有遇到什么问题

（18）join 时数据类型不一致为什么会产生数据倾斜

（19）sqoop 参数

（20）sqoop 遇到了哪些问题

（21）spark 怎么保证数据一致性

（22）spark 怎么实现高可用

（23）你们的数仓搭建原理（还是啥，有点忘了）

（24）hadoop 用的哪个版本

（25）hive 用的哪个版本

（26）你有哪些优势

（27）你们实时用的什么？（只是简单问了下，我回 flink 就没问了）

他们要找做离线的，目前没有实时需求

然后礼貌性的问了下你有什么要问的吗

第15章 丰巢

（1）你们 dws 层数据放在 clickhouse，你们不同层的数据是放在不同的存储引擎上吗？

（2）你们前面 3 层都是做短时间的 kafka 存储吗？还是说你们 kafka 也是存的全量数据

（3）明细层的话在业务库，只是做短时间的存储？最终输出到 clickhouse 做持久化的存储？

（4）离线数仓你有参与吗？离线数仓是怎么构建的

（5）你们在实时和离线的明细层分别做了两套逻辑汇总，一套是根据实时展现的指标，一套是根据后续报表进行汇总 -> 如何保证两边的口径一致?大家对一些业务口径的理解可能有些出入,如何保证两边的口径和算出来的数据是一致的

（6）不同的业务方对同名指标的描述不一致的情况,有没有考虑郭做这种指标的管理,或者一些归档?

（7）你们现在大概有多少张表, 对业务部门提供的指标有多少?

（8）数据量方便, 你们的订单表之类的大概一天有几个订单?

（9）最大的事实表是哪张? 它的量大概是什么样子? 每天的增量是多少

（10）轻度汇总是每天都会有某一个维度订单的汇总, 这个数大概是多少? 在某一个维度的事务事实的汇总的量得出来的值大概是多少呢?说的 6w 上下 -> 这个数据量的话其实用不到数仓, mysql 的一个分表就可以处理

（11）讲一下维度建模中的星型模型和雪花模型有什么区别? 模型分别的优劣势

（12）事务表用哪种方法来建? 累积快照还是事务事实? 事务事实表的建设方法

（13）事务事实和累计快照的区别, 事务事实会有什么数据, 累积快照会有什么数据?

（14）无论用事务事实还是用累积快照都可以记录订单的流转, 那累积快照是怎样去记录, 事务事实是怎样去记录

（15）SQL 与 HQL 的优化, sparkSQL 有了解过吗?

（16）reduceByKey 和 groupByKey 哪一个性能更好, 分别使用在怎样的场景下

（17）广播变量的优势是什么劣势是什么

（18）数据治理监控, 就是每天数仓跑出来的数据是正确的还是不正确的, 是否有异常, 这些怎么处理

第16章 蚂蚁金服银行科技中心

（1）自我介绍

（2）项目经历讲一下，做过哪些项目。

（3）项目规模多大，有多少张表。

（4）项目过程详细讲一下。

（5）hive 调优方法

（6）实际工作中遇到什么印象深刻的问题，怎么解决的？我说的是 join 的字段数据类型不同导致的数据倾斜

（7）数仓有多少张表，每层都做什么？

（8）dwt 有哪些主题宽表?

（9）目有什么指标，分析这些指标什么用?

小题目：

（10）有一张表，字段是设备 ID，故障开始时间，故障结束时间。求设备在工作日内的宕机时长开始时间和结束时间，可能中间隔了好几天（比如开始时间是 2021-09-28 12:12:00，结束时间是 2021-10-08 12:12:00 在工作日内的宕机时长）。

（11）有一张特别大的表怎么处理，我回答可以用 SQL 做成分区表，他又问不用分区怎么做?

（12）你的职业规划是怎样的：两年架构师，三年项目 leader

第17章 平安科技面试题

17.1 离线

（1）你们集群规模是怎样的，hadoop 集群怎么实现高可用的，各个节点都配置了那些角色

（2）yarn 的配置，怎么实现高可用

（3）离线数仓没咋问

（4）spark 没问

17.2 flink

（1）你们 flink 实时数仓是干嘛的

（2）写 flinkapi 的时候用到了那些类

（3）你们实时数仓用到了 flinksql 吗

（4）flink sql 怎么实现与表关联的，

（5）你了解哪些 connector，

（6）你们数仓都是 api 写的，那能叫数仓吗？

（7）你们 flink 没有单独的一个平台去实现写 flink sql 吗？

（8）你们数仓怎么实现 kafka 的一个 topic 对应一张表的

（9）flink 的提交流程

（10）flink job 提交后都会创建哪些类

（11）flink web 界面有那些状态

（12）flink 怎么实现精准一次

（13）kafka 是怎么实现事务的，你们 kafka 的版本有要求吗

（14）支持事务的 sink 有哪些，支持幂等的 sink 有哪些？

（15）hbase 支持幂等吗

（16）job 提交后怎么发现问题，你们遇到过很难处理的问题吗，怎么处理的分享一下

（17）你了解 flink 官网对于内存的管理吗？

（18）有了解 k8s 吗，说说 prejob 和 application 模式的区别？

（19）为啥要用 application 模式，为什么隔离性更高

17.3 java

（1）都写过哪些 java 程序

（2）你了解 hashmap 吗，说说实现原理

（3）hashmap 的 get 方法怎么实现的

（4）hashmap 和 treemap 的区别是啥

（5）你了解多线程吗，多线程会出现什么问题，你了解锁吗，悲观锁和乐观锁怎么实现的，会用到哪些类

（6）你了解开发模式吗？怎么实现单例模式

（7）你有用到 java 的一些解析工具吗比如出现了 oom 怎么解析

（8）了解 java 虚拟机吗

（9）平时有去 hadoop 社区上去接一些任务吗？

第18章 平安科技

（1）java 怎么样

（2）jvm 原理讲一下

（3）Flink 内存是怎么管理的

（4）Flink 任务提交流程

（5）Flink 组件的通讯

（6）Flink 的 checkpoint 机制

（7）Flink 的精准一次，面试官说 barrier 不对齐也可以精准一次，还给我解释了半天

（8）Flink Window 机制

（9）watermark 讲一下

（10）FlinkSQL 用过吗，FlinkSQL 是怎么把 SQL 转成任务执行的

（11）Flink 背压机制，Flink 怎么感知到背压的

（12）他应该是想问具体怎么实现的。不知道是不是 metrics-prometheus 这个东西

（13）你们平常用 prometheus 监控 Flink 的哪些指标？

（14）Kafka 读写文件为什么快？

（15）零拷贝技术，分区，顺序读写，他说还差一个页缓存，我说那不就是零拷贝吗？他说你要这样认为也行

（16）Kafka 的文件存储机制

（17）Topic：逻辑上存在 partition: 分区,物理上存在 segment:

（18）partition 的分段,逻辑上概念 index：log 文件的索引 log: 数据存储文件 timeindex: log 文件数据的时间索引 segment 的命名规则:

（19）每个分区第一个 segment 的文件名= 00000000000000000000

（20）后续第 N 个 segment 文件名 = 第 N-1 个 segment 中最后一个 offset+1

（21）segment 给 log 文件建索引的时候是每个一段范围[4k]建一个索引 如何根据 offset 找到数据位置?

（22）根据 offset 与 segment 的文件名,通过二分查找法可以确定 offset 数据处于哪个 segment

（23）通过 segment 文件的 index 索引得知 offset 处于 log 文件哪个区间

（24）扫描 log 文件对应区间得到数据

（25）kafka 的 ISR 队列

（26）Kafka 写文件是找 leader 还是 follower? leader

（27）Kafka 读取文件是找 leader 还是 follower？我说是 leader，他说你确定？然后我说 follower，他套路我，后面给我解释为什么是 leader

（28）通信方面的知识了解吗？不懂

（29）Hadoop 架构

（30）hdfs 读写流程

（31）客户端是找哪个 datanode 读取数据

（32）datanode 之间是怎么同步副本的，他说是分一块一块的,我也不懂

（33）最后讲一下 Flume 吧，他说他对 Flume 不是特别了解，让我随便扯

（34）总结：每个知识点都要问到原理的内容，他经常问：你知不知道是怎么实现的，为什么要这样

第19章 平安保险面试资料（猎头提供，仅供参考）

19.1 大数据面试题目

（1）基本是都是 spark streaming 和 Flink 的内容（数据开发）

（2）就是自己的一些项目，然后聊了一下 hive 的一些语言，还有 sql 优化的语句，然后讨论了一下运营商业绩计算规则的方案。最后要一个数据分析的报告（数据分析）

（3）SQL 题，SQL 优化，分析思路，最后要一个数据分析的报告（数据分析）

（4）基本都是问的理论方面，数仓的概念，数仓分层依据，数据质量，数据模型（数据开发）

（5）用的是什么建模技术？用的哪种建模方法论呢？抛开技术层面，从业务角度去讲解一下怎么去评估一个标签好坏的？做交叉验证的同时，有没有考虑标签的好坏，或者说

为什么好，为什么不好？（数据开发）

（6）什么叫做数仓？这几年对自己的职业规划，我说以后可能往架构师发展，然后问我做了什么准备，了解了哪些架构？提交有几层链路？公司怎么做数据治理，数据质量监控。（数据开发）

（7）spark，flink，还有数仓这一块的，一些常见故障点处理，还有技术框架的底层问的比较多，项目数据量这块有问，项目架构，数仓理论都有问，还问了对一些源码有没有研究（数据开发）

（8）nlp 的 bert transformer 之类的原理（数据挖掘）

第20章 平安产险

视频面 (约 47min)

(没问业务，基本问的都是原理概念)

（1）自我介绍

（2）有用 Java 和 Scala 吗？一上来就问我用关键字 volatile 修饰的属性有什么作用

（3）Java 设计模式有了解吗？——讲了单例模式的饿汉和懒汉

（4）Scala 的伴生对象和伴生类了解吗？样例类了解吗？

（5）Sqoop 和 Flume 有啥区别？

（6）Kafka 了解多少？为什么这么快知道吗？讲一下你知道的

（7）ISR 队列讲一下

（8）Hbase 读写流程

（9）你们分了多少个 region？有没有预分区？

（10）讲讲 Hlog 和 WAL 区别

（11）Spark 的 RDD 的特性

（12）Spark 的提交流程说一下

（13）Spark 广播变量了解吗？

（14）了解 Spark 的内存管理吗？执行内存和存储内存知道吗？

（15）你们的 Spark 版本用的多少？

（16）知道环形缓冲区吗？哪里用到归并排序

最后问了下他们那之前有好几个，现在只有 2 个人做大数据了。服务器 1000 台+ 数据量 PB 级

第21章 平安科技二面

（1）先讲项目

（2）实时项目中 Kafka 同时有多少的任务在执行

（3）项目中有哪些印象深刻的事情？

（4）工作中有哪些觉得出彩的地方？

（5）Flink 的状态机制?怎么做 checkpoint?

（6）一个算子的一次 checkpoint 会生成几个文件？

（7）从 checkpoint 恢复数据是怎么恢复的?应该是想问源码

（8）Flink 看过哪些源码?提交流程

（9）kafka 了解多少？

（10）同一个消费者组增加消费者可能消费不到数据知道是为社么吗？

（11）mysql 索引了解吗？不了解

（12）clink house 了解多少？会用而已

（13）Linux 操作系统了解多少？线程之间具体是怎么切换的？不了解

（14）一个 Flink 程序消费不到数据了怎么查找原因？看堆栈的信息，具体哪个线程在做什么

第22章 平安保险

（1）udf 函数用过吗？怎么注册的

（2）sort by 和 order by 的区别，你平时是怎么用的

（3）开窗函数有用过吗？怎么用的？

（4）旧表(全量表)join 新表，新表有新增有改变，怎么取最新的全量表？

（5）大小表的优化？怎么设置参数让小表加载到内存

（6）Hive 数据倾斜是什么？有什么表现？数据倾斜的优化？

（7）数仓数据分层是怎么分的？

（8）指标出了问题你会怎么解决？思路？

（9）sqoop 你们是怎么用的？参数有哪些？

（10）分区表你们有用吗？怎么用的？动态分区和静态分区的区别了解吗？

（11）你们数据链路多长？能从头说一说你们数据的流向吗？(从源头到报表层)

第23章 华秋电子

23.1 面试

（1）你住在哪里，你女朋友在哪块上班？想半天，直接暴露

（2）你们公司几台节点，开发多少人，部门多少人，

（3）有做元数据管理吗，

（4）集群规模，集群用的什么版本，集群角色怎么分布的？

（5）你有参与技术选型是吧，举个例子?

（6）在公司里你主要是做什么？

（7）Azkaban 挂过吗？ 他说他们经常挂 每天数据量多大？

（8）你们用 sqoop 从数据从 hdfs 导到 mysql 没有问题吗？

（9）你们 mysql 用的是集群吗？

（10）你在项目中做了些什么？

（11）说两个你分析过的指标 flink-cdc2.0 有什么新特性？

（12）clickhouse 数据存储方式，你们用的是集群吗？

（13）kafka 怎么保证分区有序？

（14）你们实时项目没有使用 flinksql？

第24章 储算科技

（1）自我介绍

（2）说一下实时数仓的搭建过程

（3）说一说 kafka 组件

（4）kafka 组件 consumer 消费 offset 存储

（5）kafka 的 isr 队列

（6）flume 的了解

（7）spark 和 flink 的区别

（8）是否了解数据模型

（9）go 语言写的服务器，使用 flume 和 kafka 的方式采集数据，如何优化

第25章 蔚来汽车

（1）java 并发机制 CompletionService

（2）java 异常机制——捕获异常+再次抛出异常与异常链

（3）讲讲 Kafka 的再平衡

（4）kafka 配置的正则表达式？

（5）讲讲 Spark 的 shuffe 机制 write 阶段 read 阶段

（6）讲讲 spark 数据倾斜的问题，出现原因，怎么处理。

（7）Flink 的事务是怎么实现。幂等性的实现机制，和两阶段提交的区别.

（8）讲讲对索引的理解！

（9）Hashmap 和 treemap 特点-优缺点

（10）Maxwell 和 canal 的区别！maxwell 的介绍

（11）二叉树的前序遍历—思路—在线文档写 code

第26章 碧桂园

刚刚面试完碧桂园的,总部在佛山,做 flink,给我的评价是有一定的基础,可以培养,说向他们领导申请二面,让我挑时间:

26.1 离线

（1）MySQL 在你们离线的集群中是扮演的一个什么样的角色?

（2）离线数仓 ADS 层的数据为什么不直接可视化,为什么还要再用 sqoop 导回到 MySQL 中去再实现可视化?

（3）你们离线为什么要分那么多层?

（4）你们的数仓建模的依据是什么?

（5）维度建模分为哪几种?

（6）事实是什么?维度是什么?

（7）事实跟指标有什么区别?

（8）说一说你是怎么用 Atlas 进行元数据管理和治理的? 我:....

26.2 实时

（1）你们以 HBase 存维度表,维度表有什么变化的数据?

（2）维度表变化的数据 HBase 是什么更新到的?

（3）还问了一堆 clickHouse 的... 我说不熟..

（4）问那个搜索关键词主题宽表用到的那个分词器是怎么弄的?为什么做这个?怎么做的?

（5）时间语义是什么?

（6）你说你用到 interval join, 什么是 interval join? 那个时间间隔是什么设定的?

（7）你们公司有几台服务器?kafka 有怎么分布?各个组件怎么布置?

（8）sparkSql 有没有做过? 我说自己搭集群玩过

第27章 企知道

27.1 第一个面试官

（1）感觉他不是很懂，像是通过面试学习知识的

（2）项目介绍一下

（3）hdfs 小文件怎么处理的

（4）MR 的执行过程

（5）spark 的 stage 怎么划分

（6）什么什么是宽依赖，什么是窄依赖

27.2 第二个面试官

像是个搞学术的，年级稍微大点

（1）我看你们简历写的都差不多，都写了这么多，你这些都会吗？

（2）Clickhouse 讲一下吧 我说了 Clickhouse 一些特点，他说太浅显了，每一个点都问为什么

（3）为什么 clickhouse 快，为什么 clickhouse 并发弱？

（4）为什么 mr 比 spark 快

（5）flink 的提交流程

（6）程序入口是什么？cliFrontend

（7）StreamGraph 是什么，数据怎么存的？对象封装

（8）JobGraph 是什么，数据怎么存的？对象封装

（9）flink 的 RM 向 yarn 的 RM 请求资源是请求的 slot 还是 task？

（10）为什么 Flink 要用自己的 RM，每次 jobmaster 直接向 yarn 的 RM 申请资源不是更快吗？

（11）为什么 Spark 没有自己的 RM，Flink 要有自己的 RM？从批处理和流处理的方面回答

公司办公环境很好，人都挺有礼貌

他们目前还没用 flink，大数据有十几个人

第28章 企知道

28.1 一面：技术

（1）项目内容简单讲一下

（2）你们用的技术

（3）hive 调优

（4）shuffle 过程及调优

（5）画像怎么做的

（6）上家的业务大致是什么

28.2 二面：主管

（1）你觉得最能体现你优秀的项目给我讲一下，让我觉得很优秀？

（2）ClickHouse 说下它的特点？其他的 MPP 数据库有了解么？

（3）讲了一下 Flink 提交流程？

第29章 幻创远景

我选的是电商公司 有自己的 app

（1）自我介绍

（2）上家公司是做什么的？

（3）公司项目的基本架构是什么

（4）你在其中扮演的角色是什么，

（5）遇到过什么问题

（6）数据量有多大

（7）用户增长怎么样

（8）做算法题，并和他说明解题思路

（9）hr 专门下载了 app，问 app 的东西

（10）很厉害的一家公司，对接美国硅谷的中小型公司，每天需要处理的数据量是

（离线）pb 级别，使用 hive on spark

第30章 富德

（1）自我介绍，问离职原因

（2）问 Hive 原理、分区、压缩和存储、表的管理、存储、调优、（这家主要技术为 Hive）

（3）HDFS 读写流程

（4）Flink 相较于 Spark 的优势、Flink 怎样处理数据、WaterMaker 的产生与使用，Flink 中的状态机制

（5）Java 中异常的产生与处理、悲观锁与乐观锁的理解

（6）JVM 的结构、内部运行流程图

（7）最后主要抓着业务问（比如你们公司的主营业务、你们负责那一块，用的什么框架与技术，你负责什么，主要实现了什么.....）。

第31章 滴普科技

（1）hdfs 小文件怎么处理？

（2）namenode 脑裂怎么处理？

（3）hive 有几种 join 方式？

（4）hive 怎么实现行转列

（5）hive 的优化

（6）hbase 怎么写入的？

（7）hbase 中 master 挂了怎么办？

（8）项目中有做 hbase 的优化吗？

（9）如何实现 kafka 端到端数据的一致性（producer->kafka->consumer）？

（10）还有几个问题记不清了

第32章 华胜天成

这家公司做的银行的项目，只做离线

（1）自我介绍

（2）介绍一下你简历上的项目（我简历上写的离线和实时，都介绍了）

（3）离线项目里写过的 sql

（4）问了些 sql 的问题

（5）写过 shell 脚本吗，举例几个

（6）linux 的一些命令

（7）听他吹牛逼扯淡

（8）离职原因

（9）离职前薪资，期望薪资

第33章 欢忻网络

（1）面试官先介绍了下他们公司，主做游戏业务，总部在硅谷，3 轮面试，终面要连线他们老总用英文

（2）自我介绍

（3）讲一下你们的项目

（4）开始怼 Flink

（5）讲一下 Flink 的分区策略，什么情况下要使用什么分区？

（6）Flink 的 JM 和 TM 的作用

（7）Flink 的提交流程

（8）Flink 的数据抽象

（9）FlinkSql 的执行和转化流程

（10）Flink 状态后端，精确一次

（11）Flink 的 checkpoint，详细讲讲

（12）重启策略

（13）反压机制，如何处理

（14）keyBy 等算子数据出现热点问题，如何解决

（15）开始怼 Kafka

（16）kafka 如何保证数据不丢，不重

（17）kafka 组件间一致性的协议（好像是这么个问题，记不清了）

（18）你还熟悉 hive 是吧，那跟我讲一下 hive sql 的执行流程，转化流程，结果的获取

第34章 博奥特

招全栈，需要会 java，大数据 两个技术面试官

（1）最近一次项目

（2）工作职责（我老是以用户日活举例，让我用别的举例）

（3）flume 有遇到什么问题吗

（4）kafka 有遇到什么问题吗

（5）kafka 偏移量保存

（6）kafka 副本同步机制，leader 的副本怎么同步

（7）kafka leader 选举

（8）实时数仓搭建中你做了什么

（9）集群规模有多大，多少核多少内存

（10）采集，离线和实时都在同一个集群跑吗

（11）flink 几个任务，集群能扛得住吗

（12）跑 flink 任务用的什么模式

（13）提交参数设置了哪些，为什么这样设置

（14）测试和生产是在同一个集群中吗

（15）数据量有多大

（16）数据量这么少为什么做大数据平台

（17）kafka 怎么确保一致性的

（18）为什么使用 flink 而不是 spark streaming

（19）flink 保证数据一致性

（20）说说 flink 你做的哪些

（21）数据哪来的（我说日志数据使用 flume 导入到 kafka，业务使用 maxwell 同步）

（22）maxwell 使用过程有什么问题吗

（23）为什么不适用 flinck cdc

（24）处理过表的 join 吗

（25）hive 有进行分区分桶吗

（26）说说分区和分桶

（27）有遇到过数据倾斜吗，产生的原因，怎么解决

（28）数据量这么小，考虑过 小文件的问题吗

（29）日期的话怎么获取表的新增数据

（30）那不会产生时间漂移吗

（31）窗口函数了解吗

（32）有使用自定义函数吗，需要实现哪个类，重写哪些方法

（33）如何获取 json 字符串数据

（34）有进行过优化吗，

（35）熟悉 java 吗

（36）set 去重原理

（37）线程和多线程了解吗

（38）进程和线程的生命周期有哪些

（39）了解 jvm 吗

（40）说说 jvm 中的组成及作用

（41）考虑做全栈开发吗

第35章 橙色魔方-太平保险

问的离线，实时没问

（1）自我介绍

（2）说一下项目经历

（3）数据项目架构（就是想问数据采集那一套）

（4）sqoop 是你自己配的吗（就是那些参数）

（5）任务调度用的什么

（6）java 怎么样

（7）java 基本数据类型

（8）java 三大特性

（9）有用 java 在 hive 里面写过自定义函数吗，udf udtf

（10）hive 架构

（11）hive 常用内置函数

（12）hive 优化

（13）最大的一张表

（14）你们数据量多大，内存多大，多少台服务器

（15）引擎用的什么，为什么用 hive on spark，有没有出现什么问题

（16）还有用其他的吗，对比 MapReduce 引擎怎么样

（17）文件存储格式有了解吗，你们用的什么，为什么用 parquet

（18）说一下比较难的指标，实现方法

（19）为什么离职来深圳

（20）薪资要求

第36章 火火兔

做儿童故事机，儿童机器人的主要搞嵌入式开发这一块。面试全程我讲项目。没有问啥难的问题。这家公司很抠门。而且他叫我一个人负责大数据的风控和推荐系统，公司他也嫌我要的太高。全程我在输出，他估计是不怎么懂。

第37章 豹亮科技

（1）kafka 为啥不支持读写分离

（2）kafka 的 follower 与 leader 会有延时，不好解决

（3）kafka 写多读也多，不太适合

（4）linux 哪个版本安全

（5）优化离线数仓搭建

（6）怎么理解元数据

（7）数据分析方法论

第38章 禾渤科技(上海)

公司用 spark 做的离线和实时，很奇怪的是他们所有业务日志数据放在 Redis 而没有使用 Kafka

（1）往 Redis 缓存维度数据和删除维度数据有做读写分离吗？

（2）Java/Scala 手写 WordCount

（3）手写 SQL 实现当天，最近 3 天，最近 7 天的订单总数，付款总数，退款总数

（4）HBase 高可用怎么做的

（5）Zookeeper 选举机制

（6）Flume 的结构

第39章 平安外包

（1）自我介绍和离职原因

（2）简单讲述之前的工作内容，涉及哪些框架比较熟悉，针对熟悉的框架询问。

（3）做过的指标和离线数仓每层涉及哪些表（举例说几个）

（4）hive：数据格式，数据去重，具体函数用过哪些（开窗函数，日期函数，字符串函数。基本上都要列出来，有的是提问）简单的 SQL 题现场出了一道和一些简单优化讲了下

（5）sqoop：对接的数据库或者其他组件有哪些，比如我为什么不用 sqoop 直接连接 hive，然后建表输出数据？

（6）最后问我有什么要问的，我就问项目的框架（他回答他们用的是 PG 数据库，做的离线，但是用的是 python 语言，spark 开发 最后委婉地说你可能更懂实时这块）

（7）其他都是初试，问题问的比较简单，面试大保健上有或者以前学的知识大部分都能回答上，我能想起的一个就是 mysql 建立索引答不上来，其他的不太了解的问题后续跟进。

第40章 软通动力

（1）是否熟悉 java 编程?

（2）flink 和 spark 哪个更熟悉？

（3）能说说 flink 的常用算子有哪些吗？能说说他们的作用和区别吗？

（4）说说 flink 的窗口

（5）是否设置过 flink 的重试策略

（6）很多张数据量很大的表进行 join，你能如何进行优化呢？

第41章 软通动力-2

（1）Kafka 的底层原理是什么？

（2）你们用 Kafka 主要做什么？有什么用？

（3）看你项目上写着 Flink 也做过，你给我讲一下 Flink 吧？

（4）你上一家公司现在还活着吗？

（5）你们编程语言用的什么？

（6）我们现在项目主要是做地图业务，你有什么需要了解的吗？

第42章 科脉

（1）一去就笔试，没人监考，但是外面人来人往的

（2）自我介绍，着重问了离职原因，还聊了下上家公司基本情况，我们组多少人

（3）linux 常用命令

（4）HDFS 读、写流程，着重问了写流程的块大小，块在哪里切分

（5）Shuffle 的优化，MapReduce 的全流程（包括 Shuffle 过程）

（6）yarn 的参数调优有哪些

（7）hive 的调优

（8）kafka 的架构，kafka 有哪些优点为什么用 kafka

（9）离线的数仓建模思路

（10）ods 层的数据是怎么倒过去的

（11）sqoop 怎么导增量表

（12）除了 mapjoin，其他的还有什么 join 方式

（13）dws 和 dwd 层有什么区别，ads 层是干啥的

（14）还问了数仓里面什么销售主题、用户主题怎么分的

（15）还问了其他的我题目都不懂的问题，印象不深忘记了

第43章 明源云集团

43.1 面试

（1）Java 学得怎么样：JavaSe 水平

（2）Java 内存模型了解吗

（3）人工智能，机器学习有了解过吗？不了解

（4）项目架构是怎样的，项目大概讲一遍，他会拿着简历提问

（5）同步策略讲一下

（6）拉链表怎么做的

（7）hive 怎么优化的

（8）spark 了解吗，spark 提交任务流程讲一下

（9）Flink 实时数仓介绍一下

（10）flink 背压机制讲一下

43.2 笔试

（1）HDFS 文件系统中 fsimage 和 edit 的区别

（2）HDFS 如何保证数据的安全性

（3）Hive 和 HBase 区别是什么

（4）请说明 hive 中 sort by order by cluster by distribute by 都代表什么意思

（5）Hive SQL 中 select from where group by limit order by 的执行顺序

（6）简单描述一下 Hbase 的 rowkey 设计原则

（7）flink 中水印什么概念，起到什么作用

（8）业务　有一张 order 流水表，60 张表分布在 5 个库中，order 订单状态是下定，付款，结束；更新周期最长是 15 天，如何在数仓中建立 dwd 表

第44章 明源云-2

明源云视频面 (约 20+min)

(比较水，大保健上都有)

（1）自我介绍

（2）采集项目讲一下

（3）kafka 分区策略

（4）kafka 幂等原理

（5）kafka 分区设置

（6）离线数仓哪些层做了分区表

（7）离线有哪些比较难的逻辑？最近 7 天连续 3 天登录。想了 3 天 3 夜一开始想用开窗+lead 相减做不出来，后来@#￥%！&搞定了

（8）MR 用了哪些排序，流程讲一下

（9）flink 反压机制

（10）spark 了解吗？原理讲一下

（11）flink 你们用过滚动窗口吗？什么场景？解决了什么问题

我问的

（12）目前招聘的项目组日常工作是做什么

（13）目前离线跟实时做到什么阶段了

（14）所以不是因为新项目组建找人而是人员编制不足补充？

第45章 明源云客

（1）标准的自我介绍，要我简略的说一下我的项目

（2）离职还是在职，离职原因

（3）你说你喜欢学习新技术

（4）你对 flink 的一个了解

（5）Chandy-Lamport 算法有了解吗，怎么实现的

（6）watermark 机制

（7）你们工作上没有用 flink 这块是吧（因为我的项目没有写，直接说是感兴趣在自学）

（8）看着我简历，你这是一个 app 吗？是做什么的，你们公司是做那块的？

（9）离线数仓的搭建？

（10）你刚刚说采用了压缩，是哪一层，怎么压缩的

（11）你这个在 ads 层之上你的指标是怎么用的，就是你在 ads 层做了哪些指标，怎么用的这些指标

（12）可视化你们用的什么，superset 吗？

（13）你们为什么这么分层，这么分层的好处是什么？

（14）你做过哪个指标，怎么实现的，做过最难的指标是什么

（15）你这个数仓是基于 hive 来建还是 hive SQL 来建

（16）你的这个用户留存、转换率这些是怎么算的

（17）你们是怎么处理拉链表的，怎么实现的

（18）拉链表里面的 end data 和 start data 怎么来的

（19）讲下你的优化

（20）数据倾斜你怎么处理

（21）你刚刚说了在 join 的时候两张表相同的字段，不同的数据类型可能也会导致数据倾斜，那你知道为什么会导致这个现象产生吗？

（22）你怎么做二次聚合，这是为了解决什么

（23）我看你这里面用了 kylin 和 presto，说下他们的区别

（24）你事实这块是怎么做的呢，用的 spark Streaming？大概一个流程是什么？

（25）你这个实时和离线之间的一个数据关系是什么样的

（26）你这个实时又是主要分析哪些指标，拿来做什么

（27）你们 HBase 表的设计是怎么设计的

（28）你平常是些 SQL 多些还是 Java 多些

（29）你最近有没有做过数据分析和数据挖掘之类的工作

（30）个人是倾向于实时这块还是离线这块

然后就是你有什么问题想问的

第46章 明源云客-2

拿着简历问的

（1）问你技术栈—离线和实时 （用的什么框架，为什么要用这个，出现了什么问题）

（2）你在这些项目中具体干了什么

（3）工作中具体负责什么业务，

（4）hadoop 体系架构，说说你的了解

（5）spark 有用过吗，我说实时用的 flink，spark 自己学的，然后就问我在 spark 学了什么，还记得什么

（6）实时为什么用 flink 说说你对他的了解，然后他会根据你说的继续问下去

（7）工作中遇到的问题，比如数据积压，怎么处理的

（8）宕机，怎么定位和处理

（9）遇到难的指标？比如连续活跃天数这样的？说一下实现思路

总得来说就是抛出一个问题，比如对 flink 的理解，我说了水印，精准一次，反压机制等，他就会根据你说的继续问下去

第47章 明源云客-3

（1）问了个 Java 什么东西，没听懂，直接说不会

（2）项目介绍，技术选型

（3）MySQL 的引擎原理，我说不知道，但是我知道它有 MyISAM 跟 InnoDB 引擎，然后他就问我区别，就等着他问这个

（4）Hive 用的什么计算引擎，HiveSQL 底层怎么转化的

（5）Flink 反压

（6）离线中最难的一个指标，怎么实现，数据从哪一层来，还有维度是什么，连问的，

没听全

（7）Kakfa 语义，还有 Kafka 如何实现精准一次

（8）最后问我我觉简历上还有哪些没问到我的，我感觉还有很多，不知道怎么说，就随便说了个 interval join 原理，他不问这个，问我还有什么 join，为什么用 interval join，join 不上怎么办

第48章 摩比可可-1

（1）clickhouse 数据量大不大，怎么读入的

（2）flink 用的什么语言

（3）你们公司数据量大小怎么样，实时有遇到大数据量的解决办法吗

（4）Java 框架高并发度这块

（5）之前工作做了什么，具体描述

人很和蔼，一直在被动接受我的输出，和他聊了很多，业务主要做国外的推特油管等，业务数据量巨大和我说有三四亿，得考虑业务数据量大的场景，

而且说是框架用的 flink 已经投入使用了，未来打算加日志数据，说是之前做的两个人都已经走了？

想找人来熟悉及后续优化？

第49章 摩比可可-2

（1）你们 Clickhouse 中宽表有多宽？

（2）按你们这个数据量，没有必要专门在集群中加一个 Clickhouse 啊。

（3）直接问业务，你们公司实时业务都实现哪些指标？

（4）你们公司用 Flink 有没有什么调优？

（5）我去的时候，XX 正在面，可能技术问的不想问了，直接怼业务，业务问的很细，什么指标，指标的作用，为什么统计这个指标，指标是你们公司自己用还是给客户用？

（6）开始说他们自己的业务，他们主要是做广告业务，主要做流量变现，说数据量很大，其实也就 3 亿条；

（7）他们离线只做统计，离线的作用只是对实时的指标做一个校验，说白了就是离线已经数仓搭建起来了，后续的指标也不会更新，一个人就维护了。

（8）他们 Flink 实时也很离谱，他们统计的指标都是按半天算的，就是窗口开的是半

天，这一块我觉得真的有点离谱，所以就问了一句，你们这个窗口这么大，内存耗的住吗？人直接说，他们 20 太服务器全都是实时；

（9）最后问了 Kafka，架构，kafka 消息是怎么路游的？数据可靠性如何实现，数据一致性如何实现？

（10）你还有什么要问我的吗？

总结：该公司可能只招一个精通 Flink 的人，主要维护他们的 Flink 实时的正常运转就行，顺带着把离线也维护了，因为后面他给我说，他们实时现在有一个人正在负责，但是因为这个人生病了，所以想要找一个人对项目接手；

最后他问我，如果他们公司的这一套交给我来交接，我大概需要多久，我直接说起码 3 个月起步。

第50章 九章数据

（1）就 Spark 我 6 台服务器 每台 120 个核 1T 内存 80T 数据 你怎么分配资源

（2）简历上写的指标你一定要知道怎么实现，他家就是干这个的，GMV，复活，流转等等

（3）hive 我 80T 的数据量跑的特别慢，你怎么去优化

（4）给你需求，行转列，列转行之类的需求，在这过程汇总还要排序，过滤，去重，你怎么实现

（5）Altas 是拿来干嘛的，最低依赖级别能做到什么程度（字段级别）

（6）你们是怎么做同步策略的

（7）你们的宽表是怎么存储的，怎么确定他的字段

（8）拉链表，这个真的很重要，在这给他扯了十多二十分钟

（9）你业务库中的数据前一天的数据发生改变，比如说用户下完单又退单了，你怎么这些改变后的数据在同步到 hive 中

（10）你们 hive 中的宽表存在哪里，字段怎么确定

（11）你们的数据量是多大，就是 dws 层宽表的数据量

（12）hive 和 Spark 的区别

第51章 中软华为外包

（1）之前公司的日活，数据量。保存周期。

（2）半年前的 sql 当时能正常跑，现在跑不了，有哪些原因。

（3）hive 小文件的优化，hive 资源可以调哪些。

（4）Spark 的核心。

（5）Spark 会产生的 shuffle 算了。

（6）Spark 如果有需求会产生 shuffle，怎么避免，或者换种方式实现。

（7）Spark 的持久化。

（8）Sparkstreaming 和 Flink 的区别。

（9）Flink 的水印机制。

（10）Flink 的窗口。

（11）Flink 的重启策略。

（12）Flink 的分区策略。

第52章 中软国际（华为外包）

（1）自我介绍

（2）实时用的什么？Flink

（3）说一下 JVM 的结构，用的一些框架

（4）说一下 HashSet 的去重原理

（5）Flink job 的提交流程

（6）窗口说一下（时间/事件）说一下区别

（7）滑动/滚动窗口的应用

（8）Flink SQL 了解吗知道底层是怎么封装的吗

（9）说一下流批一体

（10）Flink 平时遇到的问题

（11）状态知道吗

（12）checkpoint 机制、作用、具体实现原理

（13）用过哪些数据库

（14）为什么要用 flink

（15）Flink 的常用算子有哪些

第53章 中软华为外包

（1）你能讲一下你们那边的需求是什么样子的，然后您那边开发的功能是什么样子的？承担的职责是什么样子？

（2）你简单讲一下你们那边的 flink 架构

（3）Clickhouse 的更新数据怎么考虑？数据的唯一性这块是怎么考虑的？

（4）clickhouse 里面有没有做一些简单的聚合，还是只是说打宽不做聚合吗？你们 clickhouse 使用的并发度有多少？

（5）你们 flink 有做聚合操作吗？聚合的粒度是多少？有做更长的聚合吗？几个小时？一天的指标？

（6）Flink 你负责哪些指标的计算？举例？思路？做了多久？

（7）你们公司做什么业务的？

（8）你的工作年限？薪资？

（9）采集用了什么技术？离线用了什么技术？你负责的是哪些？

（10）原则指标和延伸指标是什么？

（11）大数据的海量数据处理原理是？hive 或者 spark,flink 为什么能够处理更大的数据？原理？

（12）hadoop 相对 mysql 为什么可以做到分布式处理更大的数据量？

（13）flink sql 的 join 方式有哪几种？和流式的 join 区别？和批处理的 join 区别？它的 join 原理？怎么实现快速的 join？

（14）编程语言熟悉哪些？做过 java 开发吗？

（15）你们做开发，离线和实时都是 sql 吗？为什么？

第54章 龙通科技

（1）map join 和 reduce join 的区别？

（2）hive 里面有那几种 join，区别是什么？

（3）你最近的工作内容？

（4）你懂不懂数据分析？不懂，只做需求。

（5）你做过哪些需求？

（6）spark 的优化？

（7）Linux 怎么看进程占用的资源？

（8）kakfa 怎么处理丢数据？

（9）kafka 和其他消息队列的对比？

（10）kafka 的优点？

第55章 理想动力

（1）讲一下最近做的项目

（2）clickhouse 的引擎了解吗？

（3）MySQL 索引了解吗

（4）Linux 命令

（5）数仓搭建过程

（6）维度建模过程

（7）数仓从开始搭建到能用花了多长时间？感觉他们才开始做数仓，大数据框架都不问

（8）技术选型怎么选的?

（9）人事

（10）之前公司大数据多少人

（11）为啥离职

（12）期望薪资多少

（13）之前薪资多少

第56章 理想动力-2

56.1 技术面

（1）介绍项目 一套行云流水

（2）介绍数据源的设计思路

（3）介绍建模思路　这个海哥讲了

（4）mysql 慢查询怎么处理？我说的分表查、分治，然后面试官就走了，人事就进来了

56.2 人事面

（1）你是不是计算机专业、统招学历

（2）你在原来项目是干嘛的？

（3）人事问你是自学还是？

（4）有没有结婚？有没有女朋友？.....全无

（5）什么时候打算结婚？不着急.....

（6）期望薪资？我说的 22，原公司给我 20

（7）人事说她前两天面了一个候选人，5 年工作经验，还是计算机本科，他上家公司才给 12k，我就笑了笑:哦，这也太低了，我做爬虫的时候起步都是 12，后面项目做的好，升组长，升副经理，老板就一直涨工资，一不小心就 20 了

（8）那为什么离职？ 老板投了很多项目，我们还做了一个智慧城市的招标，我负责了其中一个模块的方案设计，但最后没中标，后来又投了 k12，结果暴雷，你懂得

（9）人事最后就说加个微信，让我等通知，说最高可能就是 21k....

第57章 久谦科技

（1）自我介绍

（2）介绍一下你最近在做的项目

（3）你们用的什么语言

（4）Flink 项目是用的 api 还是用的 sql 开发的？为什么？

（5）Java 的 hashmap 底层架构，往里面的链表插数据是头还是尾

（6）单例模式怎么实现的，懒汉式里怎么整成线程安全的

（7）口述一下快排思路

（8）有若干个 0-100 的正整数，求出其中所有的两两组合和为 100 情况，口述思路

我一个干大数据的老问我 Java 干嘛？拉扯一下

（9）MySQL 的索引，什么情况下需要建索引（Java 折磨完 MySQL 折磨，继续拉扯）

（10）redis 用过吗？redis 你们保存什么类型的数据

（11）你们 Hbase 里存维度数据？维度数据不多为啥要用 hbase 存？

（12）了解 spark 吗

（13）spark 里重分区的算子

（14）spark 里的宽依赖和窄依赖了解吗

（15）kafka 的分区分配策略

第58章 跨越面试

58.1 一面

（1）自我介绍

（2）介绍一下最近的一个项目

（3）flink 算子有哪些

（4）flink 的时间语义

（5）介绍一下水印

（6）说说状态

（7）如何处理迟到数据

（8）双流 join

（9）rdd 说一下，

（10）spark 三种模式介绍一下

（11）hadoop 查看文件大小命令

（12）说说 kafka

（13）zookeeper 的选举机制了解吗

手写三道 sql 看图片

可以写不完，但是需要和他说解题思路

58.2 二面　用人部门领导

（1）自我介绍

（2）介绍一下离线数仓项目

（3）每日数据量有多少

（4）几台节点

（5）遇到过哪些难的需求

（6）数据倾斜处理过吗，怎么处理

（7）有进行过优化吗

（8）性能提升如何

（9）实时数仓用了哪些算子，为什么

（10）窗口开多大，迟到数据怎么办

58.3 三面　hr

（1）自我介绍

（2）说说上家公司

（3）离职原因是什么

（4）期望薪资和期望年薪是多少

（5）薪资流水可查吗（需要填表，包括公司证明人）

（6）什么时候入职

（7）什么时候离职

（8）什么时候涨薪

（9）只涨薪过一次吗，为什么

（10）未来职业规划能说说吗

第59章 跨越速运

（1）笔试三道 sql 题 最后一道是鹏哥那五道题中的（活动时间统计），问了具体思路

（2）介绍你做过的项目的技术选型，包括离线和实时

（3）熟悉 clickhouse 的读写流程吗

（4）问你比较熟悉的组件？ kafka

（5）rebalance 发生有哪些场景？rebalance 是针对于消费者端还是生产端？

（6）kafka 单节点最多支持多少个 topic？实际上支持多少个？为什么？

（7）kafka 支持动态修改 topic 分区数吗？怎么修改？修改时会暂停数据写入吗

（8）topic 中的数据是怎么存储的

（9）一个副本的数据是否可以存储在多块磁盘中？kafka 自身是否支持这种配置

（10）是否熟悉 mysql

（11）实时项目中的维表 join 有几种方式

（12）如果采用广播状态存储维表，修改维表中的数据后，如何让 Flink 感知到？

（13）熟悉 HQL 的转化流程吗，具体问了一个 sql 语句有多少个 mapreduce 任务

（14）项目遇到过哪些很难解决的问题？

（15）数据量多大，版本多久更新一次？

（16）你在这家公司看起来工作量不大，能不能说一个过去你觉得压力比较大的项目？

（17）说说你个人的职业规划

第60章 牧原科技

视频面试，人事和技术一起面

60.1 人事

（1）上家公司的企业文化是什么?

（2）上家公司给你带来什么样的成长?

（3）上家公司哪些地方你是觉得好的

60.2 技术

（1）问了大学专业相关的内容,虽然毫无关系,但他也要听.

（2）hdfs 的高可用模式讲一下

（3）hdfs 的高可用模式有哪些组件?JournalNoe/QJournalProtocol/nn/dn

（4）hiveOnSpark 的 join 是怎么实现的?具体详细说明

（5）hbase 的存储架构讲一下

（6）jvm 内存模型,堆内存和栈内存的区别?

（7）用过哪些数据结构,查找的空间复杂度和时间复杂度分别是?

（8）你会怎样选择数据结构?

第61章 领星

（1）你们 maxwell 数据至少一次的话，后续重复数据在 flink 里怎么进行过滤处理？

（2）质疑为什么用的 flinkCDC 技术那前面还使用 maxwell 功能，不是重复了吗，FlinkCDC 就能实现 maxwell 的功能了？

（3）Kafka 中存储三天的数据量，那你的实时数据假如统计一个月、两个月之前的跑有问题了，需要用到之前的历史数据怎么处理？重复读取这个方法不适用 有其他的吗？

（4）kafka 怎么实现 produce 端到端的一致性？以及 consumer 端到端的一致性？

（5）flink 任务在跑的过程中进行了修改，10 个算子改为了 15 个算子怎么进行动态调整，笑着和我说类似飞机飞行过程中换发动机？？

（6）flink 背压是怎么定位的除了 web？反压怎么造成的？反压会造成什么问题？

（7）flink 的回撤流写入 clickhouse 的数据怎么处理幂等性？

（8）实时数仓 kafka 中有几个分区？怎么划分的？

（9）状态后端用的哪个？最大的 checkpoint 存在 hdfs 上有多大？

第62章 中软国际

62.1 一面

技术主管电话面试

（1）是否参与了架构搭建

（2）说一下数仓搭建整个过程

（3）说一下你在项目中扮演的角色？

（4）说一下你在项目中遇到的困难，然后是怎么解决的？

（5）讲一下维度建模过程

（6）Linux 打印错误信息的命令

（7）hbase 的 rowkey 设计原则？

（8）如何调优使得 hbase 读写更快？

（9）你们每天导入数仓的数据量大概多大？

（10）你们最大的表是什么表，数据量有多大？

（11）你们使用 hbase 的时候使用过二级索引吗？

（12）你们用的是开源的还是 CDH 的？

（13）你们在写自定义函数的时候是写上函数还是下函数？

（14）你之前薪资多少？

（15）你期望多少薪资？

（16）你是哪个学校毕业的，是全日制本科吗？

62.2 二面

（1）hive 的优化有哪些？

（2）数据量比较大多个 join 执行很慢，怎么处理？

（3）spark 了解吗？spark 的核心是什么？

（4）spark action 算子有哪些？transformation 算子有哪些？

（5）你期望薪资多少，我说 19-24k

面试官：我这边给不了你这么多，最多给你 18k, 我后续再根据你的情况跟我们同事反馈一下

第63章 游禅科技

（1）什么语言写作业，python 会不会？java 和 scala 哪个熟？

（2）java 里面的反射机制？

（3）java bean 是 spring 框架里面的？

（4）scala 闭包说一下？

（5）柯里化说一下？

（6）scala 伴生类，谈谈理解

（7）谈谈隐式转换？你知道 scala 隐式转换有多少种吗？

（8）scala 模式匹配就多少种？

（9）scala 网络通信层阿卡之类的你懂吗？

（10）用 scala 写过 web 框架？

（11）能用 scala 写服务吗？

（12）说说 CDH，说说架构，里面的角色？

（13）说说 CDH 监控服务？

（14）说说 CDH 集成到 FLINK 的流程。

（15）CM sever 莫尼 ter 监控系统　之间的关系是大概什么样子？CDH 不是有基本的架构，大概是什么样子，CH 服务做高可用？

（16）还有恩 ber 锐也没听过是吧，后面会转用到恩 ber 锐

（17）hadoop 高可用以及热备份和冷备份讲讲。

（18）yarn 调度策略

（19）聊聊公平或者容量怎么实现的？

（20）然后问我几几年毕业的？

（21）说一下 redis 基本数据结构

（22）redis 哨兵模式

（23）yarn 讲讲

（24）client cluster 模式说说

（25）sparksql 只支持 client 模式那你有没有想过怎么有什么手段去优化这个支持 cluster？

（26）clinkhouse 了解多吗？ 我说查询快，面试官问所有的引擎都是查询快吗？

（27）说说物化视图是啥东西

（28）你对 dorisdb 有多少了解

（29）azkaban 会多少　我会写 flow

（30）sparkstreaming 里面怎么保证消息的精准一次

（31）说说 flink 里面怎么保证精准一次

（32）讲讲状态后端几种

（33）rocketDS 一般用在什么样里面？

（34）flink 窗口计算说说

（35）watermark　flink 里面有哪些水印生成策略？水印生产影响性能。

（36）窗口计算里面定时器，触发器，讲讲

（37）看过 flink 源码吗 入口在哪里

（38）说说实时数仓是怎么做的

（39）join 性能问题怎么去解决的。

（40）异常数据怎么处理的。

（41）遇到什么难点？性能上的难点。 然后我说了 flink 异步查询。

（42）数据没有 join 咋办，数据延迟咋处理，怎么处理更新不及时。

（43）项目中遇到哪些难点

第64章 索信达

电话面试-技术面

（1）离职原因

（2）Java 声明 volatile 变量的作用

（3）离线数仓标签有哪几种标签，这些标签有什么特性，这些标签分别采用什么技术来实现？

（4）shell 怎么知道上一行命令执行结果是成功还是失败

（5）shell 怎么得到参数个数

（6）hql 问题：三个字段 部门 薪资 姓名，想得到每个部门下面薪资最高的十个人。

（7）hive 的元数据服务和 hiveserver2 服务是什么关系，或者说 hive 的 metastore 服务可以不开吗？

（8）clickhouse 优缺点及应用场景

（9）什么是 OLAP

（10）clickhouse 每秒查多少次(QPS)——这个问题就是想问 clickhouse 的缺点：不适合

高并发

（11）Flink：时间语义

（12）Flink：对于不好直接计算得出最终结果的数据怎么处理？

（13）Flink: 有几种状态类型

（14）业务场景如何实现：Flink 消费 Kafka 的日志数据，把规则放到 MySQL 的配置表，然后需要感知配置表的变化，根据配置表规则的变化对数据流进行对应的匹配

（15）Java 的什么关键字的锁经历那几个阶段？——不懂

（16）Redis：如何实现删除 2 点到 5 点写进去的数据 —— 用 hash，key 存时间

（17）Hive 表的数据如何导到 Hbase？

第65章 索信达-2

（1）source 阶段发现读取数据慢，和并行度无关的情况下还有什么原因？ maybe kafka 只有一个分区有数据。写入的时候指定分区了

（2）es 倒排索引

（3）flink 状态 api？

（4）spark core 和 spark sql 读取 hive 表，spark core 没读取到，可能原因

（5）快排和选择排序说下、

（6）锁的状态、

（7）一些树的原理、

（8）scala 柯里化？？？

第66章 索信达-3(招商外包)

（1）flink 预提交会 commite 数据到 mysql 吗？答案是会的，然后预提交之后还要等待 mysql 的 ack，如果 ack 失败，则会全部回滚。如果 ack 成功了才会二次提交

（2）savepoint 与 checkpoint 的区别

第67章 万筑物连

67.1 一面

技术组长面 40 分钟
这一面主要是考察一下掌握的技术点，面试官还是从简历上写的项目开始问起的， 看到写的有他感兴趣的地方，就会问。

（1）自我介绍

（2）flink 状态机制

（3）flink 并行度与 slot，taskmanager 的关系

（4）flink 水印机制

（5）双流 join，join 不上怎么办

（6）对 flink 做的优化，我说的是关联维度数据，观测到反压，使用旁路缓存+flink 异步 IO，好像不太满意，最正确的回答应该是从 flink 机制上的调优，这些虽然也是优化，但是比较通用，不能体现对 flink 的了解程度。

（7）checkpoint 为什么比 spark 的好，从分布式异步快照讲了一遍

（8）精准一次性的保证

（9）flink 内存管理，这个不会，我直接就说的不知道

（10）hive 优化

（11）拉链表，从数据导入，到建模，到使用说一遍

（12）数仓建模

（13）spark 内存管理，这个也不会，就说了只知道 spark 和 flink 都有一个堆内内存一个堆外内存

（14）项目组现在是做啥的？

（15）我进去是做啥的？

（16）你们怎么数仓建模的？

67.2 二面

技术总监面 20 分钟
这一面主要考察对业务的了解情况，就是说大数据工程师还是要了解业务指标的作用和价值。答得比较磕磕巴巴。

自我介绍

（1）你们 app 是干啥的？

（2）你们做大数据的作用是什么？

（3）怎么统计这个指标的？

（4）你们怎么针对性的发放优惠卷？

我问的问题：

（5）项目组的人员配置？

（6）公司业务是啥样的？

67.3 三面

boss 10 分钟

（1）你们公司规模多大？

（2）日活？GMV？

（3）你感觉自己了解业务吗？ 我说上个面试官问的挺深的，我感觉自己还算了解把

（4）怎么来到深圳发展了？

（5）职业规划是什么样？

我问的问题：

（6）公司在研发上的投入怎么样？

（7）作为老板，怎么看待公司前景的？

第68章 先创科技

（1）kafka 原理，有哪些组件

（2）为什么用 kafka

（3）HBase 的原理

（4）HBase 优化

（5）你们用 HBase 是拿来干什么 为什么不要 ES

（6）Java 和 Scala 的区别

（7）spark 的一个优化

（8）Spark 的高阶函数用过没有 说一下你们常用的

（9）azkaban 版本是多少

（10）说一下你们的数仓建模

（11）ES 了解不 为什么用 ES

（12）我看你这上面写了 kylin 和 presto,说一下他两的区别

（13）主要还是根据你的一个简历来问，感觉自己面的稀碎

第69章 一页科技

（1）问：介绍一下项目

（2）问：介绍一下 flink

（3）问：介绍一下数据一致性

（4）问：hive 调优、数据倾斜

（5）问：介绍一下水印

（6）其他就是一些业务问题，在就没什么了，flink 的问的多一些

第70章 赢时胜

（1）自我介绍

（2）讲一下你们实时数仓的流程，根据数据流向讲

（3）讲一下你们数据怎么采集的，flume 组件，为什么不用 kafkachannel

（4）hbase 的 rowkey 怎么设计的

（5）hbase 做没做二级索引

（6）超大数据量用 hbase 怎么读写

（7）flink 提交流程，任务调度，通信原理

（8）通信原理里面你讲了 akka 和 netty，那你了解 akka 和 netty 吗

（9）kafka 如何保证不丢数据

（10）你讲了 producer 如何不丢，那 consumer 呢？

（11）两阶段提交了解吗

第71章 中地

自我介绍

（1）介绍一下你所做过的所有项目（简历上的）？

（2）数据采集为什么用 flume、kafka？

（3）你们公司都用了哪些组件？为什么要用这些组件？

（4）为什么离线和实时的可视化工具不一样？

（5）为什么离线用 superset、实时用 suger？

（6）导业务数据为什么用 maxwell，不用 sqoop？

（7）有搭建环境的经验吗？（这我理解成集群搭建了）

（8）其他的忘了，想起来了补充？

注：自研公司，目前在搭建实时平台，大数据组总共就 1 人（大佬），已经搭建到 ods 层，面试官自己都不是很懂这些

第72章 易车面试

（1）自我介绍

（2）介绍一下你们的实时数仓项目

（3）你在实时数仓开发中担任什么角色，有过哪些亮点表现，解决过什么问题

（4）有参与框架选型吗

（5）那再来介绍一下你们的离线数仓项目

（6）你在离线数仓中担任什么角色

（7）问一些简历上写的指标

（8）你的 hql 能力怎么样

（9）你写的那些 HQL 做过哪些优化（实际业务中的 HQL 语句的优化，行列过滤。。。那一套不管用）

（10）用 redis 做缓存，那考虑过穿透雪崩等问题吗

（11）你有什么问题想问的

第73章 万宝盛华

（1）Flink 底层开发做过吗，比如说 flink sql 这块

（2）说一下你用 flink sql 做了什么，有没有遇到什么问题

（3）好的，你自我介绍一下

（4）然后问了一下工作经历

（5）我们项目主要用 flink 和 hbase，你展开讲下就这两个组件你在项目中用到了哪些

（6）你们 source 是什么

（7）你们用 flink 怎么处理流数据的

（8）窗口机制

（9）你们用的 flink 是基于开源的 hadoop 吗

（10）你们部署 flink 后是用 session 模式还是 per-job 模式

（11）用 per-job 模式出现了异常你们是怎么设置自动拉起的

（12）如果重试次数达到上限后你们怎么处理

（13）Hbase 的基本原理或者是存储架构简单介绍一下

（14）表的存储结构

（15）这个 rowkey 和列簇是什么关系

（16）用 phinoix 或者 hbase api 查询表或写入数据可能出现异常，你们是怎么考虑的

（17）Java 用的怎么样，linkedlist 和 arraylist 什么区别

（18）说一下引用传递和值传递

（19）Hashset 是用什么判断两个对象是否相同

（20）删除集合元素时是用到哪个类的哪个方法

（21）Java8 的 streamAPI 会用吗

（22）Maven 常用的操作有哪些

（23）Git 的常用操作都有哪些？就是提交代码，拉取代码的命令

（24）你做过单元测试吗，做过哪些

第74章 银星智能

（1）介绍项目

（2）整个框架的理解

（3）离线数仓与实时数仓的实现思路

（4）假如有 1000 台节点，如何保证集群的正常运行

（5）讲一讲 cdh

（6）如何对 flinksql 进行封装

（7）讲一讲数据中台

（8）讲一讲数据质量监控

（9）讲一讲权限管理

（10）最后告诉我，他们是做数据中台的，类似 cdh 的那个可视化平台，供客户使用，目前他们现有节点近 1000 台，如果我去了，可能就是要做数据中台的开发.

第75章 中电惠安

（1）自我介绍

（2）介绍离线项目框架

（3）建立离线数仓遇到过哪些问题，怎么解决的

（4）离线数仓有做了哪些优化吗

（5）数据量有多大

（6）做过实时数仓，简单说一下整体的架构

（7）flume 熟悉吗

（8）说一说你对 kafka 的了解

（9）flink 的算子和作用

（10）说说 flink 如何解决乱序的

（11）有遇到过什么问题吗

（12）有做过哪些优化

（13）能说说 hbase 吗

（14）hadoop 写流程

（15）你所在的部门组成是如何的

第76章 视野数科

（1）IK 分词器有几种分词模式

（2）Flink 双流 join 会遇到什么问题是怎么处理的

（3）Flink 异步 I/o 是怎么做的，怎么实现的

（4）Flink 介绍开窗函数

（5）对 clickhouse 引擎有了解吗

（6）Nginx 是怎么配置的

（7）Phoenix 对 hbase 建索引有几种方式及区别

（8）spark 和 flink 的 checkpoint 的区别

（9）spark submit 提交任务会用到那些参数

（10）对于堆内和堆外内存如何了解

（11）udf udtf udaf 函数都有什么区别

（12）怎么在 hive 上使用自定义函数

（13）自定义函数上传的命令是什么

第77章 160 健康

（1）水印的作用？

（2）水印在哪些算子不算数？

（3）flink 的 slot 的 CPU 与内存是隔离的吗？

第78章 易数科技

（1）亿级表的去重

（2）hdfs 了解吗? yarn 的队列划分的方式 : 4 种

（3）hdfs 写的原理

（4）hive 的文件存储格式, parquet 与 orc 的区别

（5）mr 的 shuffle 原理

（6）shuffle 是哪个过程中产生的?在 map 端还是在 reduce 端

（7）用过 hive 的 udf 函数吗? 继承了 udf 的一个类,还能继承别的类吗?

（8）使用 hive 的时候数据的表现? 数据倾斜的原因有什么

（9）spark 了解吗

（10）flink 的反压机制,原理,怎么实现反压

（11）flink 的架构模型?source 怎么使用的?连接 clickhouse 怎么使用,说一下 API 和继承的类,用什么类

（12）map 和 richmap 有什么区别?open 方法有什么用?

（13）知道 flink 的 slot 的合并

（14）flink 的窗口? 时间语义有几种?

（15）flink 读取 kafka,按照事件时间的, 如何处理时间乱序

（16）维度建模的方法,步骤. 建模的过程中的顺序,建了哪些模型,第一步做的是什么模型

（17）ods 层做什么事,会在 ods 层输出什么东西

（18）dwd 层做什么,不是细节上的而是更加宏观的,这一次要做什么. 比如划分主题域之类

（19）clickhouse 的源码看过吗?为什么选用 clickhouse?为什么快?有几种表引擎?除了合并树还有哪些

（20）有做过数据治理吗?

第79章 十方教育

（1）hive 小文件处理，开启 combine 是查询完之后的，但是如果是那种已存在的小文件怎么处理，然后就是确实是数据量小文件小怎么处理

（2）hive 做过什么优化没？

（3）用过什么可视化软件，这个就问一下就没了，回答用了什么就行

（4）mysql 导出数据的时候数据量特别大，比如上千万级别的数据，mysql 那边是会

做分库分表的，你们 sqoop 是怎么合并的？

（5）离线的有没有遇到过失败的时候？怎么处理的？比如你 sqoop 失败了怎么处理，我答，重启就好了，然后就是有数据没有全部导出的情况，是因为没有设置--

（6）stage_table，设置一下就好了，还有就是 Azkaban 失败了，重启就是了

（7）flink 因为我说我在的时候做了第一版上了线我就跑路了，所以问的不多，问了一下纬度表的链路是怎样的，然后讲一下整个实时的情况是怎样的，离线的也问了

（8）这个东西我想了一下说这个我没有去整过，应该是我们项目经理弄得，会有问题吗？

第80章 腾云悦智

（1）namenode ha 如何切换？zkfc

（2）hadoop 集群 100 台变 50 台，纠删码原理？

（3）hadoop 怎么迁移？

（4）edit 文件存放位置，和 namenode 一起？name node ha 模式怎么同步 edit?

（5）2nn 除了秘书作用还有啥作用？恢复 nn 数据

（6）hbase region 分裂原理，不是问到哪个数据量分裂？

（7）region server 挂了两台会怎么样？

（8）spark 内存模型？spark 提交流程

（9）java -jar 运行任务，如果任务很慢怎么排查

（10）flink interval 原理，如果数据量大，里面的 map 不就是数据量会很大吗？怎么处理

（11）hive 迁移？

第81章 零碎问题汇总

（1）你们自定义过 connector 吗？怎么自定义的，用到哪些技术，是封装的 flink 的，还是自己写的。

（2）hive sql 转化成 mr 任务和 spark 任务是怎么转换的，比如 count distinct

（3）flink 的内存管理

（4）flink 的懒加载算子，spark 的算子是不是懒加载的？

（5）clickhouse 为什么性能高，如何充分利用 CPU 的

（6）yarn 有没有出现不均衡的情况，某个 nodemanager 内存多，处理核数少

（7）1.9 版本的 flink 和 1.11 版本的区别

（8）clickhouse 底层引擎

（9）sqoop 字段映射，如何处理数据库的特殊字段，比如 text 格式的字符串

（10）写到 dws 宽表 clickhouse 有数据错了，怎么定位是在哪出的问题

（11）Flink：对于不好直接计算得出最终结果的数据怎么处理？

（12）离线数仓标签有哪几种标签，这些标签有什么特性，这些标签分别采用什么技术来实现？

（13）shell 怎么知道上一行命令执行结果是成功还是失败

（14）shell 怎么得到参数个数

（15）业务表中如果遇见了批量刷数，会给 CDC 流处理程序带来什么影响？你会如何解决？

（16）业务数据库的 schema 是经常发生变化的，一旦发生变化，就会导致应用程序出错。你们是怎么管理业务元数据和数仓元数据的？

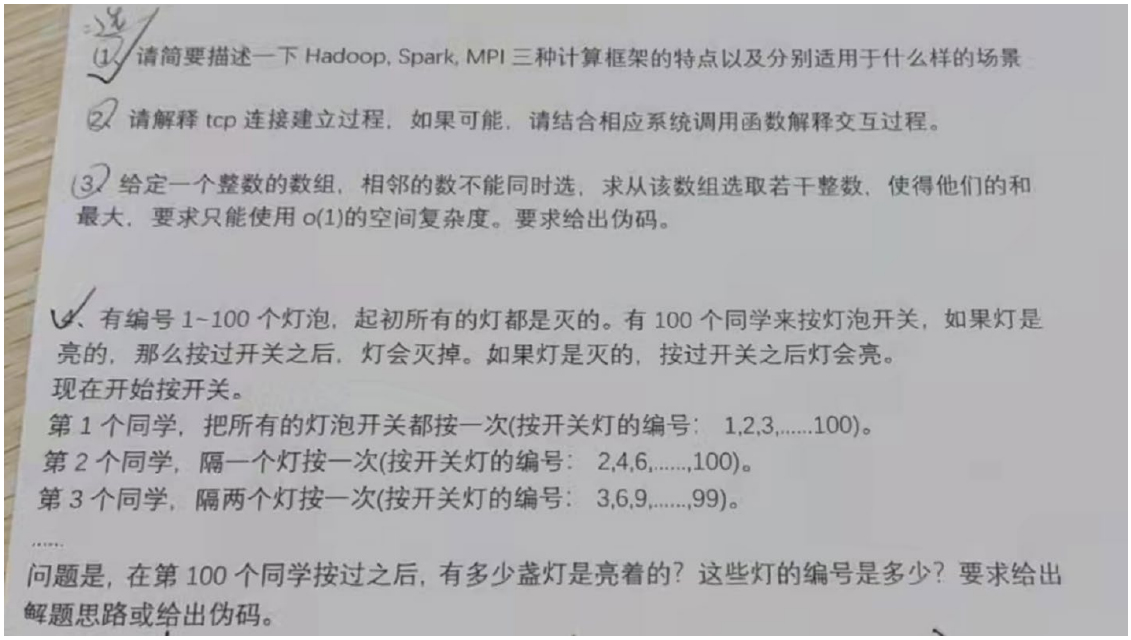
（17）用过 flink 的 cep 吗。遇到过什么问题没有

（18）有一个一年的窗口，数据量大概是十亿，如何以五分钟为单位输出一下结果

（19）公司同步数据的时候数据校验咋做的

第82章 华秋电子

（1）笔试


(1) 请简要描述一下 Hadoop, Spark, MPI 三种计算框架的特点以及分别适用于什么样的场景

(2) 请解释 tcp 连接建立过程，如果可能，请结合相应系统调用函数解释交互过程。

(3) 给定一个整数的数组，相邻的数不能同时选，求从该数组选取若干整数，使得他们的和最大，要求只能使用 o(1)的空间复杂度。要求给出伪码。

有编号 1~100 个灯泡，起初所有的灯都是灭的。有 100 个同学来按灯泡开关，如果灯是亮的，那么按过开关之后，灯会灭掉。如果灯是灭的，按过开关之后灯会亮。
现在开始按开关。
第 1 个同学，把所有的灯泡开关都按一次(按开关灯的编号：1,2,3,.....100)。
第 2 个同学，隔一个灯按一次(按开关灯的编号：2,4,6,.....,100)。
第 3 个同学，隔两个灯按一次(按开关灯的编号：3,6,9,.....,99)。
.....
问题是，在第 100 个同学按过之后，有多少盏灯是亮着的？这些灯的编号是多少？要求给出解题思路或给出伪码。

第83章 百丽零售

（1）笔试题

数仓笔试题（共三道题）

【笔试时间为 30 分钟，做完后请告知前台】

一、请写出将 /home/work/20191020.csv 数据上传到 hive 表 belle_test.dm_belle_csv_data（内部表）的脚本

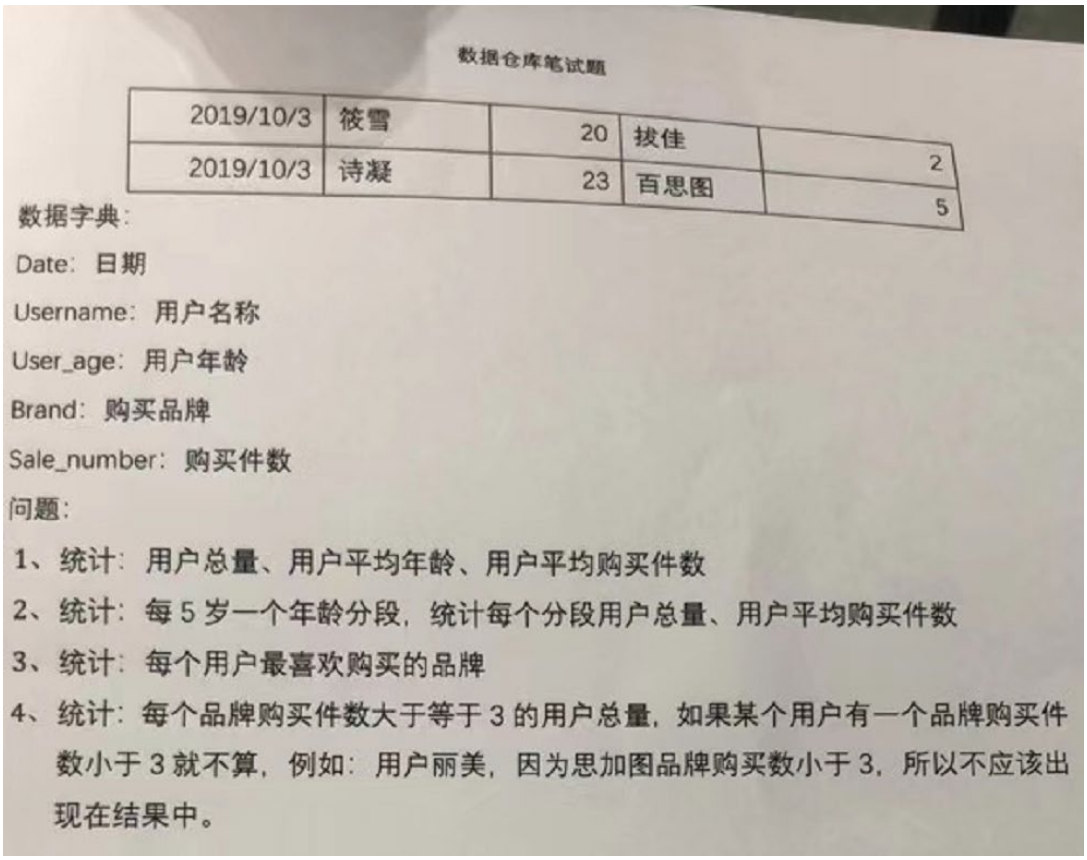
二、 请将如下 sql 修改为 Hive 的窗口分析函数来实现（不使用 union all）

```
select
invoice_type,
delivery_type,
sum(quantity) quantity,
sum(pay_total_amt) pay_total_amt
from bi_analysis.lark_ec_sale_order_parq where pay_time='2019-10-20'
group by invoice_type,delivery_type
union all
select
invoice_type,
'' delivery_type,
sum(quantity) quantity,
sum(pay_total_amt) pay_total_amt
from bi_analysis.lark_ec_sale_order_parq where pay_time='2019-10-20'
group by invoice_type
```

三、 数据表如下

Date	usename	user_age	brand	sales_number
2019/10/1	依娜	28	百丽	12
2019/10/1	依娜	28	天美意	30
2019/10/2	丽美	26	百丽	6
2019/10/2	筱雪	20	他她	9
2019/10/2	丽美	26	思加图	1

1



数据仓库笔试题

2019/10/3	筱雪	20	拔佳	2
2019/10/3	诗凝	23	百思图	5

数据字典：

Date：日期

Username：用户名称

User_age：用户年龄

Brand：购买品牌

Sale_number：购买件数

问题：

1、统计：用户总量、用户平均年龄、用户平均购买件数

2、统计：每 5 岁一个年龄分段，统计每个分段用户总量、用户平均购买件数

3、统计：每个用户最喜欢购买的品牌

4、统计：每个品牌购买件数大于等于 3 的用户总量，如果某个用户有一个品牌购买件数小于 3 就不算，例如：用户丽美，因为思加图品牌购买数小于 3，所以不应该出现在结果中。

慢 SQL 如何进行优化。

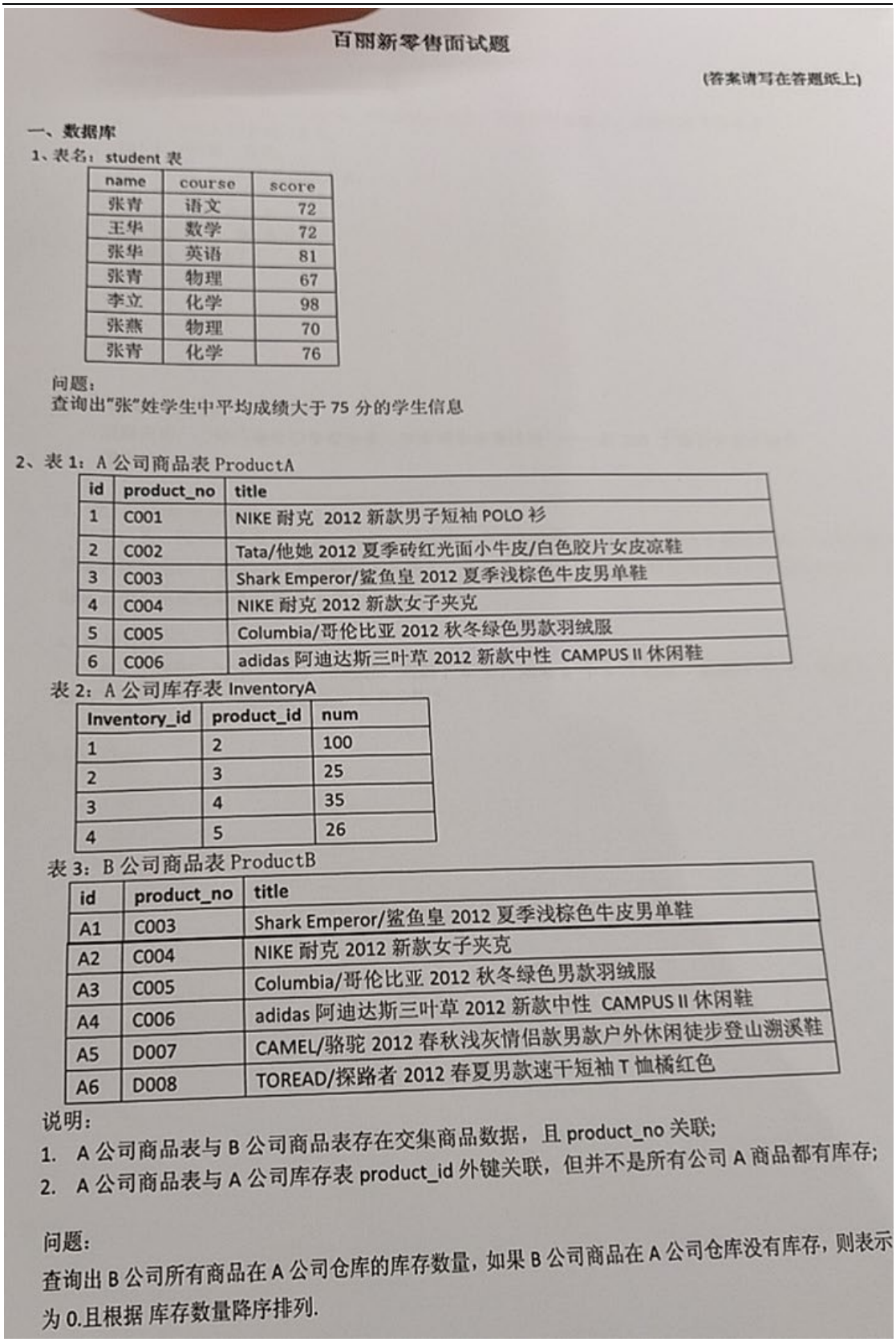
（2）explain sql 分析慢 SQL

（3）利用缓存

（4）建立分区表

第84章 百丽笔试题-2

（1）笔试



百丽新零售面试题

(答案请写在答题纸上)

一、数据库

1、表名：student 表

name	course	score
张青	语文	72
王华	数学	72
张华	英语	81
张青	物理	67
李立	化学	98
张燕	物理	70
张青	化学	76

问题：

查询出"张"姓学生中平均成绩大于 75 分的学生信息

2、表 1：A 公司商品表 ProductA

id	product_no	title
1	C001	NIKE 耐克 2012 新款男子短袖 POLO 衫
2	C002	Tata/他她 2012 夏季砖红光面小牛皮/白色胶片女皮凉鞋
3	C003	Shark Emperor/鲨鱼皇 2012 夏季浅棕色牛皮男单鞋
4	C004	NIKE 耐克 2012 新款女子夹克
5	C005	Columbia/哥伦比亚 2012 秋冬绿色男款羽绒服
6	C006	adidas 阿迪达斯三叶草 2012 新款中性 CAMPUS II 休闲鞋

表 2：A 公司库存表 InventoryA

Inventory_id	product_id	num
1	2	100
2	3	25
3	4	35
4	5	26

表 3：B 公司商品表 ProductB

id	product_no	title
A1	C003	Shark Emperor/鲨鱼皇 2012 夏季浅棕色牛皮男单鞋
A2	C004	NIKE 耐克 2012 新款女子夹克
A3	C005	Columbia/哥伦比亚 2012 秋冬绿色男款羽绒服
A4	C006	adidas 阿迪达斯三叶草 2012 新款中性 CAMPUS II 休闲鞋
A5	D007	CAMEL/骆驼 2012 春秋浅灰情侣款男款户外休闲徒步登山溯溪鞋
A6	D008	TOREAD/探路者 2012 春夏男款速干短袖 T 恤橘红色

说明：

1. A 公司商品表与 B 公司商品表存在交集商品数据，且 product_no 关联;
2. A 公司商品表与 A 公司库存表 product_id 外键关联，但并不是所有公司 A 商品都有库存;

问题：

查询出 B 公司所有商品在 A 公司仓库的库存数量，如果 B 公司商品在 A 公司仓库没有库存，则表示为 0.且根据 库存数量降序排列.

第85章 微众银行外包笔试题

（1）笔试 SQL 题

（2）基于附录 1《核额流水表》和附录 2《借据表》统计以下指标，请提供统计 SQL

指标	当日新增	昨日新增	历史累计
申请户数			
规则通过户数			
核额成功户数			
授信金额			
平均核额			
发放金额			
户均发放金额			

（3）基于附录 2《借据表》统计下述指标，请提供统计 SQL

产品类型	在贷客户数	在贷余额	不良余额	余额不良率	不良客户数	客户不良率
XX 贷						
YY 贷						
ZZ 贷						
汇总						

（4）基于附录 2《借据表》统计下述指标，请提供统计 SQL

产品类型	户数	余额	逾期率/不良率
逾期 1-30 天			
逾期 30-90 天			
逾期 90 天以上			
逾期合计			
不良合计			

（5）基于附录 3《模型输出表》统计下述指标，请提供统计 SQL（备注：value 值为 1 时即命中）

统计日期	统计指标	命中户数	命中率
2020/10/10 累计	v01		
	v02		
	v03		

	v04		
	v05		

（6）基于附录 2《借据表》统计下述指标，请提供 Vintage 统计 SQL（备注：mobX 指的是发放后第 X 月末的不良余额/发放月金额）

发放月份	放金额	MOB1	MOB2	MOB3	MOB4	MOB5	MOB6	MOB7	MOB8	MOB9	MOB10	MOB11	MOB12
2019-10													
2019-11													
2019-12													
2020-01													
2020-02													
2020-03													
2020-04													
2020-05													
2020-06													

2020-07													
2020-08													
2020-09													
2020-10													

附录：

（7）核额流水表

字段名	字段意义	字段类型
ds	日期分区，样例格式为 20200101，每个分区有全量流水	string
sno	每个 ds 内主键，流水号	string
uid	用户 id	string
is_risk_apply	是否核额申请（核额漏斗第一步）取值 0 和 1	bigint
is_pass_rule	是否通过规则（核额漏斗第二步）取值 0 和 1	bigint
is_obtain_qutoa	是否授信成功（核额漏斗第三步）取值 0 和 1	bigint
qutoa	授信金额	decimal(30,6)
update_time	更新时间，样例格式为 2020-11-14 08:12:12	string

（8）借据表

字段名	字段意义	字段类型
ds	日期分区，样例格式为 20200101，每个分区有全量借据号	string
duebill_id	借据号（每个日期分区内的主键）	string
uid	用户 id	string

prod_type	产品名称 仅 3 个枚举值：XX 贷，YY 贷，ZZ 贷	string
putout_date	发放日期，样例格式为 2020-10-10 00:10:30	bigint
putout_amt	发放金额	decimal(30,6)
balance	借据金额	decimal(30,6)
is_buliang	状态-是否不良，取值 0 和 1	bigint
overduedays	逾期天数	bigint

（9）模型输出表

字段名	字段意义	字段类型
ds	日期分区，样例格式为 20200101，增量表，部分流水记录可能有更新	string
sno	流水号，主键	string
create_time	创建日期，样例格式为 2020-10-10 00:10:13，与 sno 唯一绑定，不会变更	string
uid	用户 id	string
content	json 格式，key 值名称为 v01-v06，value 值取值为 0 和 1	string
update_time	更新日期，样例格式为 2020-10-10 00:10:13	string

第86章 法本笔试 SQL 题

（1）笔试题

一、 **编写 DDL 脚本，完成表及视图的创建（重点考察对数据类型的理解）。**

1) 创建员工信息表，包含以下字段：

员工 ID、姓名、性别、出生日期、岗位、级别（共 10 种可列举的级别）、加入公司日期、开始工作日期、当前状态（包括在职、离岗、离职）。

2) 创建在职员工信息视图，包含以下字段：

员工 ID、姓名、性别、出生日期、**年龄**、岗位、级别（共 10 种可列举的级别）、加入公司日期、开始工作日期、**工作年限**、当前状态（包括在职、离岗、离职）。

二、 请根据示例数据中已经存在的数据表，完成相关数据分析的 SQL 脚本

学生信息表：Students

字段名	字段类型	说明	备注
StudentId	UNIQUEIDENTIFIER	学生 Id	
Name	NVARCHAR(100)	学生姓名	
Birthdate	DATE	出生日期	
Gender	BIT	性别	1 代表男性、0 代表女性、NULL 代表未知
NativePlace	NVARCHAR(100)	籍贯	到省
SchoolYear	NVARCHAR(50)	年级	入学年份，如 2020 级新生
ClassNo	NVARCHAR(50)	班级	
Status	SMALLINT	状态	1 代表在读、2 代表辍学
RoomNo	NVARCHAR(50)	宿舍编号	

科目信息表：Subjects

字段名	字段类型	说明	备注
SubjectId	UNIQUEIDENTIFIER	科目 Id	
Name	NVARCHAR(255)	科目名称	

考试信息表：Examinations

字段名	字段类型	说明	备注
ExId	UNIQUEIDENTIFIER	考试 Id	
ExName	NVARCHAR(50)		
ExDate	DATE	考试日期	
StartTime	TIME	考试开始时间	当天的具体时间
TimeSpan	NUMERIC(18,4)	考试时长	单位：小时
Invigilator	NVARCHAR(255)	监考老师	姓名
Place	NVARCHAR(255)	考场	
SubjectId	UNIQUEIDENTIFIER	科目 Id	

考试成绩表：Scores

字段名	字段类型	说明	备注
StudentId	UNIQUEIDENTIFIER	学生 Id	
ExId	UNIQUEIDENTIFIER	考试 Id	
Score	NUMERIC(18,4)	分数	

科目信息表：Subjects

字段名	字段类型	说明	备注
SubjectId	UNIQUEIDENTIFIER	科目 Id	
Name	NVARCHAR(255)	科目名称	

授课表：Teach

字段名	字段类型	说明	备注
SubjectId	UNIQUEIDENTIFIER	科目 Id	
Teacher	NVARCHAR(255)	授课老师	
ClassNo	NVARCHAR(50)	班级	
TScore	NUMERIC(18,4)	均分指标	

1. 期末考试（最后一次考试）成绩已出，需要根据学生的平均成绩对教师进行评定，有两种方案进行参考：

   1) 根据教学指标完成情况评定：每位老师教授所有学生的平均分 与 设定的教学指标的百分比关系，>= 110% 成绩为 A+即优秀、>= 105%为 A 即良好、>=100%为 B 即合格、其余为 C 即不合格。请完成一段 SQL 脚本，输出每位教师的评定结果，输出信息包括教师姓名、教授课程、受教授学生的整体平均分、教学指标、完成百分比、评定结果。

   2) 请完成一段 SQL 脚本，输出每位教师所教科目最高分的学生，输出信息包括教师姓名、教授课程、所教授学生的整体平均分、该科目整体平均分、所教授学生的整体平均分年级名次、最高分分数，最高分学生姓名（如果并排分数，用逗号合并为一行显示）。

2. 需要考察学生群体的学习状况，分别从科目、班级、性别等角度进行分析：

   1) 为考察各班级在单个科目上的学习状况。完成 SQL 脚本，按照各班级、各科目的成绩进行排名，并输出每个班级、每个科目最高分（名次可重复），对应的学生姓名、年级名次（可以重复）等信息，即每个班级每个科目的最高分在年级中的排名。

   2) 为表扬学习成绩优异的同学，根据总分进行排名。完成 SQL 脚本，按照各班级、各同学总成绩进行排名，并输出班级、班级名次（可以重复）、全年级名次（可以重复）、姓名、总得分等信息，要求输出第 10 条记录到第 20 条记录。

第87章 跨越

（1）笔试

87.1 第一题

- 可使用任何数据库 SQL 实现
- 需使用单条 SQL 实现功能，如果有多种实现方式可加分
- 如不理解题目含义，可以联系 HR
- 要求在 30 分钟内完成题目

第一题

员工表结构

emp

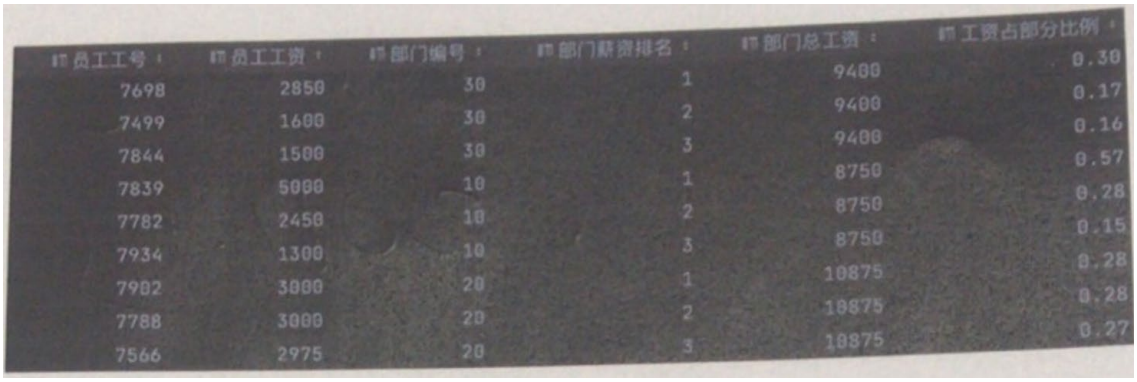
empno 员工工号	integer
ename 员工姓名	string
hiredate 入职日期	string
sal 员工薪水	integer
deptno 部门编号	integer

员工表数据

empno	ename	hiredate	sal	deptno
7521	WARD	22/2/1981	1250	30
7566	JONES	2/4/1981	2975	20
7876	ADAMS	13/7/1987	1100	20
7369	SMITH	17/12/1980	800	20
7934	MILLER	23/1/1982	1300	10
7844	TURNER	8/9/1981	1500	30
7782	CLARK	9/6/1981	2450	10
7839	KING	17/11/1981	5000	10
7902	FORD	3/12/1981	3000	20
7499	ALLEN	20/2/1981	1600	30
7654	MARTIN	28/9/1981	1250	30
7900	JAMES	3/12/1981	950	30
7788	SCOTT	13/7/1987	3000	20
7698	BLAKE	1/5/1981	2850	30

求出每个部门工资最高的前三名员工，并计算这些员工的工资占所属部门的总工资的百分比

结果：



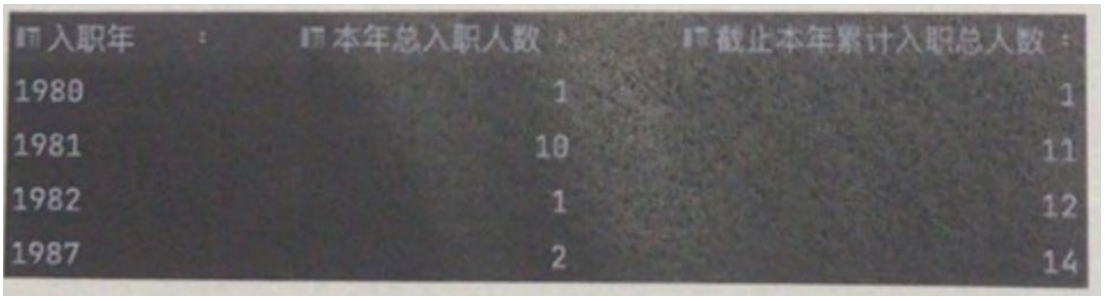
员工工号	员工工资	部门编号	部门薪资排名	部门总工资	工资占部分比例
7698	2850	30	1	9400	0.30
7499	1600	30	2	9400	0.17
7844	1500	30	3	9400	0.16
7839	5000	10	1	8750	0.57
7782	2450	10	2	8750	0.28
7934	1300	10	3	8750	0.15
7902	3000	20	1	10875	0.28
7788	3000	20	2	10875	0.28
7566	2975	20	3	10875	0.27

87.2 第二题

在第一题员工表的基础上，统计每年入职总数以及截至本年累计入职总人数

截至本年累计入职总人数=本年总入职人数+本年之前所有年的总入职人数之和

结果：



入职年	本年总入职人数	截止本年累计入职总人数
1980	1	1
1981	10	11
1982	1	12
1987	2	14

87.3 第三题

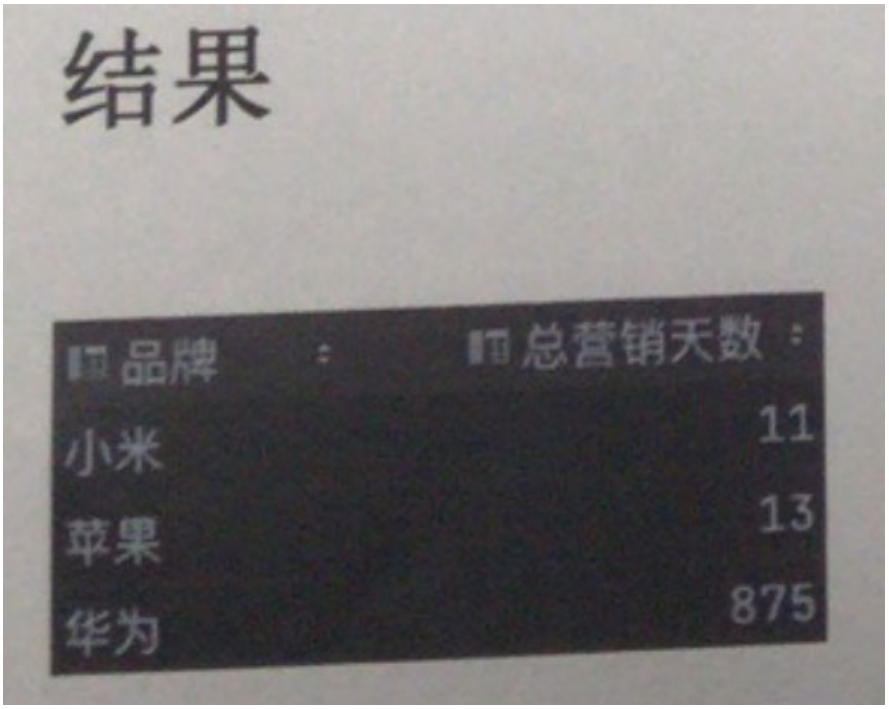
该表记录了每个品牌的营销活动开始日期以及结束日期，现需要统计出每个品牌的总营销天数。

注意

1 苹果第一行数据的营销结束日期比第二行数据的营销开始日期要晚,这部分有重叠的日期的要去重计算。

2 苹果第二行数据的营销结束日期和第三行的开始日期不连续，20190907 以及 20190908 不统计到营销天数中。

结果



第88章 招商永隆笔试题

（1）笔试

招商永隆 数据仓库笔试题

姓名：_____ 供应商：_____ 日期：_____ 考核官：_____ 分数：_____

1. 简述什么是 OLTP 和 OLAP，这二者的异同点，各自适合的应用场景。（30 分）

2. 简述对数据仓库的理解，一般有哪些体系架构，各自适合的应用场景。（40 分）

3. 简述曾经使用过的大数据平台和技术，大数据平台与传统 MPP 数据仓库平台的主要差异，各自适合什么样的应用场景。（30 分）

第89章 东信时代

89.1 技术一面(20min)

（2）自我介绍

（3）会什么开发语言？Java Scala 是自学的吗？

（4）GC 的算法 ——不太了解 只学了 JavaEE

（5）Redis 为什么这么快？

（6）Hbase 的 RowKey 设计原则

（7）RowKey 的读写流程

（8）用 Phoenix 操作 Hbase 的时候有没有遇到什么问题？——没答上

（9）Clickhouse 的优缺点

（10）MySQL 的视图和 Clickhouse 的物化视图的区别

（11）用 Clickhouse 查询最大的数据量(行数和字段数)

（12）Flink 和 SparkStreaming 的区别

（13）Flink 如何实现精确一致性

（14）有没有做过数据中台？——没有做过但讲了一下我对数据中台的理解

（15）你们部门只负责存数据，其他部门做可视化展示吗？

（16）部门多少人？做大数据多少人？

89.2 HR 二面(约 15min)

风格比较刨根问底

（1）为什么从深圳毕业要去北京工作？现在为什么从北京回来？

（2）看你简历上的每一点写得很有条理性，是否有参考别人的简历？

（3）什么时候开始做实时项目的？

（4）为什么你的简历上的项目的时间都是连贯性的 你们公司做完一个项目人员马上投入下一个项目 没有交集的吗？

（5）你觉得工作中最大的收获是什么？

（6）能否提供离职证明？

（7）你对我们公司有什么想要了解的吗？——问了下还会不会有三面，答复时间

第90章 HungryPanda 数据仓库工程师笔试题

（1）笔试题

1、T1 表 JOIN T2 表 ON T1.C = T2.C，会得到多少条数据？

表：T1

C
1
1
2

表：T2

C
1
2
2

2、如何根据三张原始表，得出目标表

表：T1

ID	C1
1	a1
2	a2
3	a3

表：T2

ID	C2
2	b2
3	b3
4	b4

表：T3

ID	C3
3	c3
4	c4
5	c5

原始表

如何根据上述三张原始表，得出下方的目标表，请写出SQL。

目标表

ID	C1	C2	C3
1	a1		
2	a2	b2	
3	a3	b3	c3
4		b4	c4
5			c5

3、如何根据原始表得出目标表

表：成绩单

原始表

姓名	科目	成绩
张三	数学	99
张三	语文	88
张三	英语	77
李四	数学	85
李四	语文	79
李四	英语	91
王五	数学	85
王五	语文	79
王五	英语	91
赵六	数学	85
赵六	语文	79
赵六	英语	91

如何根据上述的原始表，得到下方的目标表，请写出SQL

目标表

姓名	数学成绩	语文成绩	英语成绩
张三	99	88	77
李四	85	79	91
王五	85	79	91
赵六	85	79	91

4、旅店房间饱和情况

表：旅店入住离店记录表

原始表

用户ID	房间号	入住时间	离店时间
7	2004	2021-03-05	2021-03-07
23	2010	2021-03-05	2021-03-06
7	1003	2021-03-07	2021-03-08
8	2014	2021-03-07	2021-03-08
14	3001	2021-03-07	2021-03-10
18	3002	2021-03-08	2021-03-10
23	3020	2021-03-08	2021-03-09
25	2006	2021-03-09	2021-03-12

现有一分析场景，想分析出每个时间段，有客人在住的房间数量。
如何根据上述的原始表，得到下方的目标表，请写出SQL。

目标表

开始时间	结束时间	有客人在住的房间数量
2021-03-05	2021-03-06	2
2021-03-06	2021-03-07	1
2021-03-07	2021-03-08	3
2021-03-08	2021-03-09	3
2021-03-09	2021-03-10	3
2021-03-10	2021-03-12	1

第91章 地上铁笔试题

（1）笔试题

深圳地上铁数据开发岗笔试题

姓名：_____ 应聘部门：数据平台研发组 应聘岗位：_____

一、 单选题

1、下面那个程序负责 HDFS 数据存储？
A、NameNode B、JobTracker
C、DataNode D、SecondaryNameNode

2、HDFS 中的 block 默认保存几份？
A、3 份 B、2 份 C、1 份 D、5 份

3、下面关于 hive 的说法正确的是？
A、hive 是基于 hadoop 的一个数据仓库工具，可以将结构化的数据文本映射为一张数据库表，并提供 sql 查询功能。
B、hive 可以直接使用 sql 语句进行相关操作。
C、hive 能够在大规模数据集上实现低延迟快速的查询。
D、hive 可以通过索引查找数据提高查询效率。

4、在关系型数据库中，为简化用户的查询操作，而又不增加数据的存储空间，常用的方法是创建？
A、另一个表（Table） B、游标（CURSOR） C、视图（VIEW） D、索引（INDEX）

5、OLAP 技术的核心是？
A、在线性 B、对用户的快速响应 C、互操作性 D、多维分析

二、 简答题。

1、简述 ETL 中数据清洗的常见内容？

2、Hadoop 集群可以运行的三种模式分别是什么？

3、在数据平台建设过程中存在缓慢变化维情况时怎么解决？

4、Hive 中统计出每门科目成绩排名前三的学生对应的分数？
数据见右图：

name	subject	score
孙悟空	语文	87
孙悟空	数学	95
孙悟空	英语	68
大海	语文	94
大海	数学	56
宋宋	英语	84
宋宋	语文	84
宋宋	数学	64
婷婷	英语	86
婷婷	语文	84
婷婷	数学	65
婷婷	英语	85
		78