# 一 头条面试题

## 1

**1）技术框架部分：**

（1）会根据面试者简历上的技能点进行考察。注意技能点描述用词，不要随便用熟悉，觉得特别熟的可以说熟练使用

（2）常用的大数据框架是肯定会问的，比如Hive、Spark、Kafka等。常见的如Kafka精准消费问题的多种解决办法对比，Spark的Task调度规则

（3）有些面试官会直接让讲了解的框架最底层实现

（4）Java，MySQL，Redis必问，JVM，Hashmap，JUC相关，MySQL的索引及优化，Redis数据结构、集群、缓存淘汰

（5）常见的协议会考，租约协议，quarum原理，Zookeeper选举，acid，acp，base等

**2）项目部分：**

（1）一般会让你挑你最熟的一个项目讲。考察从项目背景到实现的把控

（2）重点是参与的部分，遇到的难点

（3）常见业务难题的解决和优化，层层递进，头条追求做到极致

（4）没有做过的项目、模块不要乱讲，容易被问懵逼或者被太简单被鄙视

**3）算法部分：**

（1）数据结构必考，手写代码，每一面都会考。常见数组、链表、二叉树、跳表的题。有些部门会先笔试（电脑实际编码运行）再面试

（2）算法部分常见动态规划、概率题、二进制一类

（3）SQL题目，行列转换。分区函数，统计连续登陆天数这一类问题

（4）多刷Leetcode，题都是有套路的

**4）HR部分：**

（1）想好跳槽理由，一般注重抗压能力和稳定性

## 2

**1）技术部分**

（1）WordCount的实现过程

（2）MR与Spark的区别

（3）Spark在Client与在集群运行的区别

（3）相同的SQL在HiveSql与SparkSQL的实现中，为什么Spark比Hadoop快

（4）自定义UDF

（5）设计HBase表需要注意的点

（6）HBase的hlog

（7）数据同样存在HDFS，为什么HBase支持在线查询

（8）数据从Hive中用SparkSql进行操作有遇到什么问题？类似兼容性的问题。

（9）SparkStream与Strom，Flink与什么区别

（10）有三个map，一个reduce来做top10，哪种方法最优。数据量特别大。

**2）项目的架构**

（1）数据仓库的模型设计

（2）数据仓库的数据清洗

（3）业务建模、数据分析方法。

（4）数据仓库是怎么设计的

（5）数仓规范设计哪些方面(字段、维度，存储压缩、数据保留机制)

（6）数仓质量怎么监控(数据质量管理系统，主键唯一、非空、数据波动)

（7）数仓主题分哪些(按照公司业务归类：申请单、客户信息、合同信息、放款、还款、余额、逾期等)

（8）数仓拉链表的原理

（9）有没有遇到数据倾斜的问题（场景、解决方式）

（10）数仓重点调度任务的保障方式（调度系统优先级）

（11）数仓任务报错和监控（调度系统捕捉错误，电话短信告之值班人员）

## 3

**1）技术部分**

（1）自我介绍

（2）写map-reduce，两份数据，一份是设备id+用户id，一份是设备id+点击的广告+点击时间，找出每个用户id每天10条最新的广告点击记录

（3）说下多线程和多进程

（4）fork命令

**2）算法部分**

（1）hashmap和map，hash实现原理（hashmap基于哈希，查找O（1）；map基于树，查找O（logn）；hash实现采用数组，偏移地址使得查找满足O（1），无论是再哈希还是链表法解决冲突，都需要存储key和value）

## 4

（1）MySQL引擎是什么，常用的是哪个innodb，知道原理吗

（2）MySQL优化，怎么建索引？选择合适的字段属性。对于某些文本字段可以设成ENUM类型（MySQL中被当做数值型数据来处理）。使用连接join代替子查询 。使用索引：innodb支持哈希索引、b+树索引、全文索引。

BTree和B+Tree

（3）手写sql...

user\_id login\_date

1 20200325

查询出用户连续三天登录的用户

表A 字段a 值1234

表B 字段b 值11235

Join 输出的结果

Left join 输出的结果

（4）Sqoop导入数据是增量怎么实现

（5）Flume事务实现

（6）Kafka消费者角度考虑是拉取数据还是推送数据

（7）Kafka中的数据是有序的吗

（8）Kafka数据推送失败怎么处理

（9）Kafka保证生产者精准一次

（10）没有接受到ack才会出现声明情况

（11）Kafka数据重复怎么处理

（12）Spark Streaming怎么里面实现精准一次消费

（13）如果offset没有发送成功数据会怎样

（14）Hive的优化，项目中怎么优化的(我说了join的一些优化)

（15）然后就问了用MR怎么实现join,手写代码

（16）数据倾斜怎么处理的，

（17）碰到过oom情况吗，什么原因导致的，怎么处理的

（18）Hbase有那些组件，

（19）什么场景会用到Hbase

（20）Hbase的读写流程，大概说一下

（21）Spark,任务提交的流程，

（22）Spark的两种核心Shuffle，未优化的和优化的

（23）常用的数据结构都有哪些，

怎么实现一个list,

怎么实现一个map

map一般什么场景使用

用过树这种结构吗，什么场景用到的

（24）算法

字符串中第一个出现三次且仅出现三次

反转一个链表

两个链表第一次交叉的节点，

（25）用Cannal监控mysql之后，Scala怎么通过SparkStreaming去消费的

（26）你们这个项目spark计算完之后存HBase怎么设计的

（27）HBase你们在项目中怎么设计rowkey的

## 5

**一面**

（1）算法：给2个有序数组，合并成一个有序数组

延伸：给n个有序数组，合并成一个有序数组

（2）sql:

班级 学号 科目 分数

求某个班级总分最高的前五名

其他：

（3）精准一次性消费

（4）数据倾斜

**二面**

（1）主要问项目中是否丢数据，怎么做的数据治理，监控，问了一个算法，一个SQL

（2）算法是有序数组查找目标值，我的做法是二分查找

（3）sql是给用户的观看视频的时长，统计观看时间在0-1分钟的用户数，1-10分钟的用户数

**三面：**

（1）选型为什么是spark而不是Flink，有哪些优缺点（我项目上写的spark）

（2）HBASE为什么快（列式存储，内存，lsm树）

（3）消费不到kafka数据怎么办

（4）kafka怎么监控，重分区

（5）g1回收器和cms区别

（6）jvm调整过没有

## 6

一面

项目类型的没啥说的，都是类似的一套东西，问的东西

手写的东西挺多的

rdd手写key的top10, 比如 user\_id, score， 求成绩的top10

spark-sql手写个dateframe的demo

sql题目

department\_id, user\_id, amount

求出来每个部门，每个人的薪资占部门薪资的比例， 一个简单的开窗函数实现的

手写个udf函数？

表的字段 report\_date, map\_info(map类型的数据类型， key是时间，value是金额)，计算大于当前report\_date的所有金额

#

我直接使用explode函数处理了

然后，继续延伸，增加个 report\_date, map\_info，count 增加计算个count数量 又怎么处理？ 其实就是max(count)就行了

然后，说那你能写个UDF函数实现吗？ 我最后就写了个轮廓；

又让写了个算法题： 最长回文子串. 给定一个字符串s，找到s中最长的回文子串

def get\_longest\_str(s):

n = len(s)

dp = [[False] \* n for \_ in range(n)]

res\_str = ""

for ele in range(n): # ele代表i和j的间隔

for i in range(n): #通过 j=i+ele 得到子串的结束位置， 对角线斜着一行一行的处理

j = i + ele

if j >= n:

break

if ele == 0:

dp[i][j] = True # 单个元素为true

elif ele == 1:

dp[i][j] = (s[i] == s[j]) # 两个元素 判断两个值是否相等

else:

dp[i][j] = (dp[i + 1][j - 1] and s[i] == s[j]) # 如果当前元素相等, 则比较之前的元素

if dp[i][j] and ele + 1 > len(res\_str): # 上次ans就是上次的j-i的长度

res\_str = s[i:j + 1]

return res\_str

## 7

|  |
| --- |
| 1.数据仓库的基础理论:建模,分层 |
| 2.flink容错机制 |
| 3.spark调优，数据倾斜的处理方式 |
| 4.sql编程考察 |
| 5.算法:聚类kmeans |
| 6.二叉树算法 |
| 7.Java 线程问题，包括线程安全，锁，和线程池，问得很细。 |
| 8.Hbasse和mysql区别和存储原理 |
| 9.kafka的底层原理，kafka如何保证全局消费数据有序 |
| 10.redis底层原理，如何预估数据量 |
| 11.yarn底层原理，执行流程和，yarn底层使用的算法 |
| 12.算法基础建模有哪些，如何进行数据挖掘 |
| 13.链表反转排序 leetcode原题 |
| 14.sql 连续日期活跃用户 |
| 15.hive udf udaf udtf spark |
| 16.任务提交过程 |
| 17.回文函数 |
| 18.hive 数据倾斜 |
| 19.redis 跳表 |
| 20.spring ioc |
| 21.spring 需要单例设置的参数 |
| 22.线程池几个配置参数含义 |
| 23.mysql 事务 a b 客户端提交事务处理 |
| 24.storm 实时 |
| 25.java 并发包 |
| 26.hive 如何转mr |
| 27.线程 可重入理解 |
| 28.算法题 一个LRU 一个深度优先搜索 |
| 29.flink 的watermark shardGroup的概念 |
| 30.kafka如何保证消息的有序行高可用 |
| 31.数据倾斜如何处理 |
| 32.flink watermark 和barries 和checkpoint |
| 33.flink和spark的区别 |
| 34.jvm常见的垃圾回收算法 |
| 35.heap怎么分带的 |
| 36.kafka调优 |