尚硅谷大厂大数据最新面试真题

(作者：尚硅谷大数据研发部)

版本：V1.0

一、头条面试题

1.1 学长 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

1.2 学长 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）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）

1.3 学长 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）

二、腾讯面试题

2.1 学长 1

1）笔试部分

（1）有一表名 t_sh_mtt_netdisk_log，从表名可以看出该表是什么业务的，是什么周期粒度的表。

（2）怎么查看表结构，表创建语句？怎么查看表有哪些分区？怎么查看分区对应 hdfs 路径？怎么计算某个分区的数据量大小？怎么计算某个分区的文件总数？

（3）有一 hive sql，怎么计算这个 sql 会产生多少个 map 数？

（4）怎么查看 hive 有什么自带函数？怎么查看函数的详细信息？

（5）hive 支持哪些基木数据类型？

（6）请指出下面两个语句读取分区 ds 数据的不同

Select * from t where ds=20150101 and province='gd' or province='ex'

Select x fromt where ds 20150101 and（province='gd' or province='ex'）

（7）将字符串"keyl=valuel&key2=value2...keyn=valuen"进行分割放到一个字段里面，可以查出任意一个 keyx 对应的 valuex 值。请写出查询语句，并计算总共有多少个 key 值。

（8）用一个字符串比如"alblc"来测试某个函数比如 split（用|分隔符将字符串进行分割）的语句

（9）a 表和 b 表内连接，a 表为小表，只有 2000 行记录

select a.*from a Join b on a.key=b.key

可以进行怎样的优化

（10）a 表 left join b 表，b 表为小衣，可以进行怎样优化？

scleet a.*from a left join Join b ona.key=b.key

如果 a 为小表，可以怎样优化？

（11）请指出下面 sql 语句的区别

select a.* from a left Join b on a.key = b.key and a.ds=xxx and b.ds=xxx

sclect a.* from a lef Join b on a.key = b.key and b.ds=xxx

select a.* from a lef Join b on a.key = b.key and b.ds=xxx where a.ds=xxx

Select a.* from a left Join b on a.key = b.key where a.ds=xxx and b.ds=xxx

（12）多表连按的写法：a,.b,c 三个表内连接，连接字段都是 key，怎样写连接语句？

（13）两大表连接，发生了数据倾斜，有几个 reduce 无法完成，怎么查找发生数据领斜的原因？应该怎样优化？

语句：select t1.*,nvl(t2.x,1) from t1 left join t2 on t1.guid=t2.guid

（14）两大表连接，发生了数据倾斜。有一个 reduce 无法完成，检直发现 t1 中 guid=" 的记录有很多，其他 guid 都不重复，这条语句该怎样优化？

语句：select t1.*,nvl(t2.x,1) from t1 left join t2 on t1.guid = t2.guid

（15）如何用 hive q！实现 sqL 中的 exist/in 子句

mysql 语句如下：

SEL ECT a* FROM a where a.key in（select dstinct key from b where key like ‘filter%’）

（16）sort by、distribute by、cluster by 和 order by 区别

注：可参考数据工厂的“帮助中心》集群 hive》hive 使用优化》hive 使用注意事项”

2.2 学长 2

1）笔试题

用 Hivesq1 实现

（1）全量用户登员日志表 t_login_all，字段信息 ftime（登录日期）、openid（登录帐号）新增用户登录日志表 t_login_new，字段信息 ftime（登录日期）、openid（登录帐号）求每天新增用户次日、7 天、30 天留存率。

（说明：7 天留存是指当天有登录且第 7 天还登录的用户）

（2）消息流水表 t_chat all，字段信息：

Ftime（日期）、send_user id（发消息用户 id）、receive.user id（接收消息用户 id）、

chat id（消息 id）、send.time（发消息时间）

用户登录流水日志表 t_login_all，字段信息：

Ftime（日期）、user_id（用户 id）、login_id（登录 id）、login_loc（登录区服）、login_time（登录时间）

求：每天有收发消息用户最近登录时间、登录区服，输出 ftime，user_id，login_loc，login_time

用 Spark core 实现，写 scala 脚本

1、reduceByKey()、groupByKey()有什么区别？

2、DataFrame 和 RDD 有什么区别？

3、使用累加器，对数组 Array(1,2,3,4)每个元素实现累加操作，最后打印输出求和结果。

4、现有文件 file.txt，文件格式如下

Order_id，user_id，payment，productid

1、1768，50，155

2、1218，600，211

3、2239，788，242

4、3101，288，599

5、4899，25，230

6、2311，890，981

。。。。。。

求 Top10 个 payment 字段的值

2.3 学长 3

1）技术部分

（1）我看你上面写了你会 spark，那你用 spark 写一下刚刚那道题的代码（第 6 题），不能用 spark sql

（2）我看你上面有写到熟悉 hive 原理和优化，你说一下怎么优化

（3）优化有说到语句优化，然后又说回到 pv 和 uv 那个问题，假如说让你优化你怎么优化（我代码里面有 count(distinct id) uv,）

（4）问了一些 kafka 的，我只记得一点点，问我数据会不会重复消费，什么情况下会重复消费

（5）在 hive 执行语句的时候如果很慢，什么原因造成

我有提到数据倾斜的问题，他又问怎么解决数据倾斜

2）项目部分

（1）让你挑一个比较有亮点的项目详细说

（2）你负责了什么，数仓中统计了哪些指标，

（3）刚刚听你说项目里面有提到分层，如果我要你求每天商品的 pv,uv，从 ods 层到 dm 层它的具体流向，各层都统计了什么（后来让写求 pv,uv 的 sql 语句）

3）场景部分

（1）有一个分区表，表名 T，字段 qq，age，按天分区，让写出创建表的语句

（2）刚刚的分区表，求 20200221 这个分区中，年龄第 N 大的 qq 号列表

（3）有一个表，两个字段，分别是 qqa 和 qqb，数据如下：

12,34

12,56

12,78

34,56

34,12

找出所有互相关注的 qq 对

4）手写部分

（1）三个字段，timestamp，user_id，product_id，让求 pv 最大的商品，写了之后又叫用 scala 代码写一遍，然后又问，假如说只让你求 pv 数大于 100 的 top3 呢，代码又怎么写

三、阿里面试题

1）技术部分

（1）spark 提交 job 流程

（2）提交脚本中-jar 什么意思

（3）Excutor 怎么获取 task

（4）详解 Hadoop 的 wordcount

（5）spark 做过哪些优化，（优化说完会问你为什么？原理是什么？）

（6）spark 内存管理

2）算法部分

（1）单向链表反转

（2）实现堆栈 push Pop Min 复杂度 0（1）

（3）实现一个队列，男女均衡最长的队列 复杂度 0（N）

四、百度面试题

1）笔试部分

（1）你可以得到网站访问记录，没条记录有 user IP，设计一个程序，要随时能计算出过去 5 分钟内访问次数最多的 1000 个 IP

（2）简述 paxos 协议

（3）编程简单实现一个阻塞队列

（4）简单实现一个 LRU 算法

（5）有一个 5000 万的用户文件(user_id,name,age),一个 2 亿记录的用户看电影的记录文件(user_id,url),根据年龄段观看电影的次数进行排序？

2）技术部分

（1）Java 的 NIO 是否了解

（2）Java 多线程（什么 wait、notify）

（3）MapReduce 的 shuffle 过程（手画，边画边介绍），reduce 是怎么到 map 端获取数据的（RPC 通信是否了解）

（4）一句话介绍 MapReduce

（5） Spark 做缓存时缓存在哪里

（6）百度不用 Spark，只用 Hadoop 和 MapReduce（至少这个部门是这样，这个部门的主要业务是做用户画像），把 MapReduce 工作的各个阶段过程吃透

（7）是否看过 Hadoop 源码

（8）Linux 系统（常用命令啥的）

（9）除了这些平时还接触或者学习哪些技术

3）项目部分

（1）业务架构

（2）根据项目业务介绍技术

（3）MapReduce 数据倾斜和内存溢出怎么办

4）算法部分

（1）手写归并排序

（2）用递归实现斐波那契数列第 100 项的值（第一项和第二项分别为 0 和 1），手写

（3）算法和数据结构了解到什么程度、知道多少

（4）Java 中的数据结构，ArrayList 和 LinkedList 的区别，ArrayList 为什么查询快（为什么用数组就快）、LinkedList 为什么增删快，哪些 Map 是线程安全的

5）手写代码部分

（1）手写 MapReduce 的 WordCount（导入的包最好能记住）

6）场景部分

（1）遇到垃圾日志怎么办（假如说有人恶意制造日志，怎么处理。我说的黑名单机制，那个面试官貌似还挺满意）

五、京东面试题

5.1 学长 1

1）笔试部分

（1）列举几种数据倾斜的情况，并解释为什么会倾斜，以及如何解决？

（2）解释一下 sql 运行步骤，是否有优化空间，如果有，如何优化：

SELECT a.id, b.name FROM a LEFT OUTER JOIN b ON a.id = b.id WHERE a.dt = ‘2016-01-01’ AND b.dt = ‘2016-01-01’;

（3）已知表 a 是一张内部表，如何将它转换成外部表？请写出相应的 hive 语句

（4）用 select 做查询是，用哪个函数给值喂 null 的数据设置默认值？

（5）Hive 中，建的表为压缩表，但是输入文件为非压缩格式，会产生怎样的现象或者结果？

（6）订单详情表 ord_det(order_id 订单号，sku_id 商品编号，sale_qtty 销售数量，dt 日期分区)任务计算 2016 年 1 月 1 日商品销量的 Top100，并按销量降级排序

（7）某日志的格式如下：

pin|-|request_tm|-url|-|sku_id|-|amount

分隔符为‘|-|’，

数据样例为:

张三|-|q2013-11-23 11:59:30|-|www.jd.com|-|100023|-|110.15

假设本地数据文件为 sample.txt,先将其导入到 hive 的 test 库的表 t_sample 中，并

计算每　　个用户的总消费金额，写出详细过程包括表结构

（8）test.log 日志中内容如下左列所示，使用 awk 输出右列 4 行数据

10-3-jd-dv

2-4-jd-dv　　10-4-jd-dv

5-7-pv-click5-7-pv-click

36-24-pv-uv

37-24-pv-uv37-24-pv-uv

24-3-uv-mq 24-3-uv-mq

（9）storm 的分组策略有哪些？

5.2 学长 2

1）笔试部分

（1）hiveSQL 语句中 select from where group by having order by 的执行顺序

（2）hive 中 mapjoin 的原理和实际应用

（3）写出你常用的 hdfs 命令

（4）使用 Linux 命令查询 file1 里面空行的所在行号

（5）有文件 chengji.txt 内容如下:

张三 40

李四 50

王五 60

请使用 Linux 命令计算第二列的和并输出

（6）在 Linux 环境下有文件/home/dim_city.txt 如何加载 dim_city 外部表中，HDFS 路径/user/dim/dim_city

（7）请列出正常工作的 hadoop 集群中 hadoop 都分别需要启动哪些进程，他们的作用分别是什么，尽可能写的全面些

（8）数据仓库的整体架构是什么，其中最重要的是哪个环节

5.3 学长 3

1）笔试部分（京东金融）

1）数据按照业务含义可以分为时点数和时期数，在一般情况下，下列哪些数据属于时点数？

A.昨天的订单量　　B.昨天的库存量

C.昨天的点击量　　D.昨天的访问次数

2）About hadoop map/reduce，The right answer is？

A.reduce 的数量必须大于零

B.reduce 总是在所有 map 完成之后再执行

C.combiner 过程实际也是 reduce 过程

D.Mapper 的数量由输入的文件个数决定

3）Hive 中的元数据不包括？

A.表的名字　B.表的外键

C.表的列　　D.分区及其属性

4）Hive 中如何限制查询条数？

A、TOP　　B、limit　　C、rownum　D、only

5）关于 hivesql 以下说法正确的是：

A.cluster by 不会对字段进行排序

B order by 只保证每个 reducer 的输出有序，不保证全局有序

C sortby 是全局有序

D distribute by 制定规则字段，将相同组数据分发到同一 reducer

6 下面 sql 的运行结果是什么？

Select a.id,b.name from (select id from table_id) a left semi join (select id, name from table_nm) b on a.id=b.id

table_id
id
1
2

table_id	
id	name
1	wzb

A 、(a,ok)　B、(2,null)　C、(1,ok & & 2,null)　D、以上结果均不对

7、你处理过的最大数据量是多少？处理时是否遇到问题，采取了哪些优化的方案？

8、列举 Hive 几种数据倾斜的场景以及解决方案？

5.4 学长 4

1）技术部分

（1）你们 spark 哪个版本

（2）sparkStreaming 如何和 Kafka 对接

（3）Kafka 消费过的消息如何再消费

（4）redis 版本多少

（5）cdh 版本多少

（6）集群怎么部署的

（7）数据倾斜怎么处理

（8）mr 和 spark 的 shuffle 的区别

（9）mr 环形数组怎么设置最大能设置多大

（10）redis 分布式锁怎么实现

（11）rdd 和 dateset 的区别

（12）说说 Scala 伴生对象

（13）解释 Scala 的模式匹配

（14）谈谈 Scala 隐士转换

（15）yarn 的组件说说他们的功能

（16）spark 一个 excutor 给多大内存

（17）spark 几种部署模式还有他们的区别

（18）你运行程序是 client 还是 cluster 为什么

（19）说出 git 的几个常用命名

（20）git 如何切换分支

（21）对 hive 的理解做过哪些优化

4）算法部分

（1）用 idea 写快速排序

5）手写代码部分

（1）手写 spark wc

6）项目部分

（1）整个业务数据的流向

7）场景部分

（1）现有两张大表所有字段都得保留不能再过滤了 join 操作就发生 oom 怎么解决

（2）session 日志分析如何求出用户访问的 page 路径

（3）三表的 join 求销量，order——id 的统计（记得不了）就是三表 join 求统计吧

六、58 面试题

6.1 学长 1

1）笔试部分（基础题）

（1）请用图形的方式表示出 java 运行时数据区

（2）以上哪些数据区在空间不足的情况下会抛出异常？

（3）java 工程出现内存泄露如何排查？请写出你知道的排查工具以及方法

（4）以下代码的输出结果是什么？

```
String str1=new StringBuilder("58").append("ganji").toString();
System.out.println(str1.intern == str1);
String str2=new StringBuilder("ja").append("va").toString();
System.out.println(str2.intern == str2);
```

（5）在 scala 中有一种函数其参数可以为参数，这样的函数称为高阶函数，请参照 scala 中常见 map 函数实现一个名为 mymap 的高阶函数，mymap 接收两个参数值，第一个函数（x:Int）=>3*x，第二个为 Int 型数据。在 mymap 函数体内将第一个参数作用于第二个参数。

2）笔试部分（算法）

（1）给定两个只含有 0 和 1 字符的字符串，实现二进制加法。如：

```
String str1 = "101"
String str2 = "11"
```

实现 binarySum(String s1,String s2),使得输入以上字符时返回字符串"1000"。

（2）给定一个整型数组 array（没有重复元素），再给出一个目标值 target，数组 array 中有两个元素的加和等于 target，要求返回这两个元素的下标。如 array = {1,2,4,6,3} target = 8，则返回[1,3]。

3）笔试部分（大数据相关）

（1）现有一张 hive 表，表里面有两个字段 uuid 和 ts，分隔符为\001 分别表示用户唯一标识和用户来访的时间戳。

有如下需求

（a）要求计算每个用户的最后一次来访时间，用 mapreduce 实现（写完整的代码）

（b）你所写的代码有没有优化的空间？如果有，请写出优化方法。

（2）用 spark 实现 WordCount。

（3）spark-streaming+kafka 实现实时计算这种方案中通过什么方式保证数据的准确性？

（4）分别列举 hadoop 和 spark 中的文件缓存方式。

（5）HDFS 的一致性是指什么？有几种语义？

（6）基于 yarn 的推测执行是怎么实现的？描述其大致算法。

QUESTION NO:1

```
publicclass Test1{
    public static void changeStr(String str){
        str="welcome";
    }
    public static void main(String[] args)(
        String str="1234";
        changeStr(str);
        System. out. println(str);
    }
}
```

输出结果：

QUESTION NO:2

```
public class Test2{
    static boolean foo(char c){
    System out print(c);
    return true;
}
```

```
public static void main(String[] argv）{
    int i=0;
    for(foo('A');foo('B') &&(i< 2);foo('C')){
    i++;
    foo('D');
```

输出结果：

6.2 学长 2

6.2.1 一面（学长面）

1）技术部分

（1）HashMap 如何遍历。（这个忘了，百度了一下是用 EntrySet）

（2）sparkStreaming 窗口操作的 3 个时段设定分别是什么，问到倍数这块（第二轮也问到了）

2）项目部分

（1）讲了一下推荐项目怎么回事

（2）因为是招数仓这块的，所以三轮数仓都问了（第二轮的数仓问题比较偏，我写到第二轮总结里）

3）算法部分

（1）快排或冒泡排序，我讲的是快排

4）情景部分

设定了一个情景，手写 SQL，比较简单吧

6.2.2 二面

1）技术部分

（1）flatMap 和 map 的区别

（2）A 表 joinB 表，用 MR 是如何处理的（这个真没看过，百度了之后才知道的）

（3）yarn 的资源调度

（4）Hbase 有 put 方法，那如何批量 put 数据进 hbase 里，用什么方法（这个也没答上，面试的时候想不起来了）

（5）NameNode 的 HA，一台挂掉之后，另一台会丢失数据吗，挂掉之前有一部分日志

为更新（这个我答的是 edits 会保存下来，和之前的镜像一起加载，所以不会丢）

（6）问了 kafka 的消费速度（这个我说的几十兆每秒，群里已经发了，是 1G 每秒，可自由发挥）

（7）Kafka 分区（我说的 5 个，有点多了，应该说 2 个或者 3 个吧）

（8）问了 HDFS 的块你们设置的是多大（我说的 128M，他有疑惑，我说默认是 128M，他也表示疑惑）

（9）问了 Spark 版本，说一直是这个版本吗（我说现在用的是 2.0，之前用的 1.6）

（10）Sparksql 什么时候不能读到 Hive 的数据（这个也没答上来）

2）项目部分

（1）你们数仓之间的数据是怎么调度的（我说直接写 sql，他就问说你怎么知道你的数据已经存储成功了，当时尬了一会，我说用 crontab 调度的，他问我这用 crontab 合适吗，我说那设置 Oozie 可以设置一段程序执行完再执行下一条程序，这块基本就聊炸了吧，我不知道 hive 内部数据调度还需要这样调，我以为直接读）

3）情景部分

（1）设置了一个情景，让写 sql，也相对简单吧

6.2.3 三面（项目组负责人）

（1）要招数仓和算法的，所以主要问了数仓。

（2）数仓的 DM 为何不由数据组来设计封装（因为我说的是指负责设计下面三层，然后由不同的业务组区数仓里调数据），你觉得是数据组封装有什么不妥吗

（3）HBase 里面你们都存了具体哪些业务数据（这个我说了大概的存哪类型的数据，他问的非常深，必须说出存的具体是什么数据，我后来说每个商品的点击次数统计，实时更新，），他说这些数据我觉得存到 redis 或者 mysql 中也行呀，为什么要非要用 HBase（我说 HBase 查的块，他说我觉得这样的数据量 redis，mysql 查的也不慢吧），然后问了我你认为多少的数据量适合用 HBase（我说的几百万）

（4）结合你这个 HBase 里存储的数据，说一下你们的 rowkey 怎么设计的（这块说的不好，因为前面说的不好，感觉这边说就强行设计了）

6.3 学长 3

1）技术部分

（1）对 hadoop 各组件之间通信 Rpc 协议了解

（2）Hadoop 的读数据流程

（3）传输过程中 DataNode 挂掉怎么办

（4）Hbase 源码

（5）Hbase 构建二级索引如何保证索引与数据的一致

（6）JVM 架构，堆内存，gc 算法，问平时会不会调 gc 算法

（7）-Xms 和-Xmx 设置一个和设置两个而且两个设置的都是一样有什么区别吗

（8）查看当前 java 内存各个代的内存比例

（9）查看线程堆栈的命令

（10）查看 java 进程里边对象的大小，

（11）堆内存泄露怎么看，除了 Dump 出来用 MAT 还有什么简单的方法

（12）Java 进程用户反馈请求服务很慢，但是代码不是你开发的你怎么排查问题，多线程，实现线程同步的方法

2）算法部分

（1）问对数据结构和算法了解吗，堆栈和链表什么的

七、新浪面试题

7.1 学长 1

7.1.1 一面

1）自我介绍

叫什么名字，来自哪里，本科哪个学校，硕士哪个学校，大数据做了多长时间，对 Hadoop 生态圈以及 Spark 生态圈中的哪些技术比较了解（很简单的一句就带过了），说完大概 2min

2）技术部分

（1）让我介绍下 sqoop，说他没见过

我说 sqoop 我用的不熟，知道怎么用，可以简单地给他介绍一下。然后就说 sqoop 是 hdfs、关系型数据、hbase 它们三者之间传输数据用的，很方便。

（2）让我介绍 MapReduce 的运行过程

二话不说要笔要纸，手绘 MapReduce 从 InputFormat 到 OutputFormat 的流程，一边画图一边说。

3）项目部分

（1）让我介绍项目我都是给他要笔要纸，手绘架构图。

离线：data->flume->kafka->hdfs->hadoop/spark->mysql->前端展示

实时：data->flume->kafka->Spark Streaming->mysql->前端展示

kafka 到 hdfs 是自定义的 consumer

（2）为什么要使用 kafka，可不可以用 flume 直接将数据放在 hdfs 上

我说可以用 flume 将数据直接放在 hdfs 上，但是很少有人这么干，为什么要用 kafka 我没有深入研究过（当时没想到咋说）

（3）让我介绍下项目中的指标，都是怎么实现的

我就挨个把我写上的项目指标说了下是什么意思，是怎么实现的（我简历上每个项目就写了几个指标，我感觉比较熟的），最后我说还有其他的很多指标，我做的就这些。

4）手写算法部分

（1）让我写链表的反转

我说链表没咋接触过，老师换个吧

（2）接下来让我手写单例

分分钟写了个双端检测单例

5）是否有问题问面试官

（1）最后说有没有问题要问他

问 1：我就问他们大数据是做什么，一天的数据量有多大

答 1：他给我说他们是做反垃圾的，一天的数据量上 T 了

问 2：我就又问他们怎么定义的垃圾数据

答 2：他就给我说一切影响用户体验的数据都是垃圾数据，比如说在用户的评论下面打广告之类的

然后我就说暂时没有了，他就让我在那等着了

7.1.2 二面

1）自我介绍

还是一面那一套

2）技术部分

（1）介绍 MapReduce 的运行过程

还是上面的一套。不过在我一边画图一边说的时候有人给他发消息，貌似挺着急，然后

我就停下来了，他说不好意思他回个消息，我就说不着急，等您忙完咱再聊。

3）项目部分

（1）项目框架介绍

老一套

（2）问我写的公司做什么的

我说做电商的

（3）他又问具体都卖什么

我说什么都卖，就跟京东和淘宝似的，只不过规模小很多

（4）问我上家公司做大数据的有多少人

我说十多个吧

（5）问我公司在哪

我说在丰台

（6）问我一天的数据量有多大

我说 APP 端有十多个 G，将近 20 多个 G，浏览器端大概有四五十 G

（7）又问我一天的访问量有多少

我当时心里就按照费老师说的 10w 个人 1G 的量大概算了下，说了个数

4）手写算法部分

（1）手写二分查找

二分是基于一个排好序的数组，我就问他数组排序用不用写，他说不用，然后就分分钟写完，最后我还说用不用手写 MapReduce 的 WordCount 以及 Spark 的 WordCount 的，一点问题没有。

（2）让我用 shell 写一个脚本，对文本中无序的一列数字排序

我说 shell 简单的我可以，比如说写个脚本，crontab 周期性调度一下，复杂的我得查下资料，也就没写

5）是否有问题问面试官

（1）最后说他没问题了，我有什么要问他的没

我说要我不，他笑了，然后他说除了这个

（2）我当时也没想到什么问题，就问了嘴他们的集群有多大

他说成千上万台

7.1.3 三面

1）自我介绍

（1）自我介绍

老一套

（2）问我哪一年毕业的，公司有多少人

我实话实说，说还没有毕业就去实习了，我说二三百人

（3）问我为什么要做这个，什么专业的

专业我实话实说的，然后说 15 年底 16 年初的时候原油价格暴跌，说以前行情好的时候最高一桶原油 120 美元，巅峰能达到 160，我找工作那年最低的时候 25，行情特别差，自己比较喜欢编程，硕士毕业的东西都是用 MATLAB 实现的，然后就自己学的 Java，找的这个工作，进去以后领导让做大数据，就跟着老师傅干了。

2）技术部分

（1）问我 kafka 用的那个版本

0.11

（2）问我 ZooKeeper 的常用命令

我说 status（基本全忘了，就这个记得很清楚），他说还有呢，我说 ls，然后我又说忘的差不多了，但是我知道它的命令和 Linux 命令差不多，都是那一套，那老师也就没再问。

（3）问我 myisam 和 innodb 的区别

我就按照阳哥讲的，先说了下 myisam 不支持主外键，不支持事物，它是表锁，不适合高并发，而 innodb 则不是，然后他又问我还有没，我说了解就这么多，他也没再问

（4）他又问我一句如果单线程访问，myisam 和 innodb 那个性能好一点

我说 myisam 吧，因为它不支持事物，要快一些

然后他又问我一个问题，我没答上来，我忘了那个问题是啥了

（5）问我 redis 常用的命令

我顿了一两秒甩了一句 keys*，他笑着说可以可以

（6）问我为什么要用 redis

我就说 redis 是内存型数据库，以前访问量没那么大的时候关系型数据库完全可以胜任，但是在高并发的情况下访问量一大，关系型数据就不行了，所以就有了 redis

（7）问我了解 hbase 不

我说了解不多，就知道它是一个非关系型数据库，它也就没再问

（8）然后问我机械硬盘和 SSD 硬盘的差别了解不

我说 SSD 读写快，然后他说是，他又问我底层了解不，为什么 SSD 读写快，我没答出来，他又说从硬盘读取数据首先要干什么，我说寻址（然后突然一下子好像明白了，就跟他扯了扯）

（9）然后他又问我机械硬盘每秒读取次数多少知道不

我说我不知道，没研究这么深，但是我说我知道我们大数据读取数据的时候寻址不超过 10ms（我当时心里这样想的，这些东西我没接触过，我要往大数据上靠，要引导他去问我大数据的东西），结果他说知道这个能算出来机械硬盘 1 秒能读多少次，然后我好想明白了，就 1s/10ms 呗，他说是，也就 100 来次

3）项目部分

（1）问我一天的数据量有多大

我还是按照原来的那个说的，他说不到 100G 单机就能搞定吧，我说公司就这么定的，然后尴尬不失礼貌地笑了笑（跟尽际老师学的）

然后就看着简历问了一些问题

（2）问我 azkaban 一天调度多少个任务

我说三五百个吧

（3）然后他问我这么一个问题，他说你做这些项目需要埋点日志的对吧，然后说如果后台的人不愿意给你加这个埋点，嫌麻烦，你怎么办

我愣了一会，随口说了句，吃顿饭，那老师当场笑了，然后他说这算是个办法，他又说如果这个人没时间，没工夫跟你吃饭，你怎么办，我当时傻笑了两下，没回答。

4）场景题，手写 SQL

场景 1：是这样的，一张表三个字段，分别是学生名字、科目名字、科目得分，让我用 sql 把总分最高的找出来

我说我数据掌握不太好，试着写写。

我用了个子查询，先按学生分组，然后求每个学生的总分数

然后在外部查询中求最大的

我一开始在子查询中直接对用聚合函数求的 sum_score 排序了，他看了以后说能这么做么，我说我忘了，然后我又说子查询不行的话咱们就在外部查询排序呗，然后反问他可以吧，

他说可以的

5）平时学习习惯

CSDN、博客园、Apache 官网

6）是否有问题问面试官

（1）最后他说他差不多了，我有什么要问他的没

我问了下他业务方面的一些，也没想到啥，随便问的

（2）最后我又问他，他刚才问我埋点日志后台的人不给加怎么办

他给我说这个需要各个组协调好，不然别干活了，就扯了一会

7.1.4 四面

1）自我介绍

老一套

2）技术部分

（1）介绍 MapReduce 的运行流程

本来我说给他手画的，他不让，说介绍就行。我就从 InputFormat 开始到 OutputFormat 给他说了一遍

然后他就说按着简历来问

（2）让我说下 myisam 和 innodb 的区别

老一套

（3）然后看着简历说你知道 HDFS 的存储过程，我说需不需要我说一下

然后我就把海哥 hadoop 存储过程那几个图给他说了一遍

3）项目部分

让我介绍下我写的三个项目都是干啥的，我就没画图，给他介绍了下每个项目具体是做啥的，怎么做的（什么领导层要看指标啊啥的）

4）平时学习习惯

还是 CSDN、博客园、Apache 官网，说都是晚上和周六周天自己学的

（1）问我喜不喜欢技术

我说很喜欢，我说了以后他说为什么喜欢，我就说我搭建起来一个集群很有成就感，能在我女朋友面前炫耀一番。。。

（2）问我喜欢跟人打交道还是喜欢跟机器打交道

我按照实际情况说了下，说比较宅，比较喜欢跟机器打交道，还说从小就比较喜欢玩小霸王游戏机之类的。。。

7.1.5 五面 HR

教训：和 HR 接触不要激动，不要表现的很倾向来她们公司，因为后面工资可能不好谈，要不到理想的价位。

1）自我介绍

老一套

2）人事高频问题

（1）有没有别的 offer，面试几家了

我说这周一开始投的简历，在等百度的 offer，上午面试的 58 到家，但是 58 到家做 Hive 比较多，我了解不多

（2）问我什么时候辞职的

我说十一月初

（3）问我为什么现在才找工作

说回家一趟，大半年没回家了，这两天刚回来

（4）上家薪资多少

我就按照我心里的想法和马老师说的大概说了个数

然后说什么他们很严格，如果不行会马上辞退啥的，最后说看我能力可以，问我啥时候可以上班，让我准备下薪资流水和离职证明

7.2 学长 2

1）技术部分

（1）kafka 吞吐量怎么优化

（2）hive 底层，sql、表关联、窗口函数、聚合函数、一堆场景题

（3）join 对应的 MR

（4）一张小表 join 一张大表 在 MR 上怎么 JOIN

（5）map()阶段聚合

（6）spark job 流程

2）项目部分

（1）上个月用户访问次数最大的 ip

（2）topN

（3）日活实时统计

（4）sparkStreaming 的活动窗口大小，数据重复怎么办？

（5）数据仓库 ODS-》DWD 用什么写的脚本？

3）算法部分

（1）二分查找

（2）字符串反转

7.3 学长 3

1）技术部分

（1）hdfs 的小文件处理。

（2）hive 调优，数据倾斜问题

（3）spark 任务提交流程

（4）mr 任务提交 yarn 流程

（5）hdfs 的读写流程，shuffle 流程，shuffle 过程的归并排序具体怎么做。

（6）是否用了 currenthashmap，高并发问题就高枕无忧了

（7）线程池了解吗？

（8）怎么判断一个 kafka 消费者挂了

（9）了解 redis，mysql 吗，mysql 的索引了解吗？

（10）select。。。for update 是什么锁？

（11）mysql 底层文件存储

（12）读过哪些框架的源码？

（13）jvm 底层原理

（14）spark 的 distinct()算子底层

（15）布隆过滤器

（16）zookeeper 的作用，zookeeper 单独的事务，nginx 的作用，只做收集数据吗？

（17）说一下 kafka 的架构，描述 hw leo，kafka 的优化，怎么做幂等，你们做业务幂等性，为什么消费者怎么看到 hw。消息的一致性能解决什么问题，kafka 事务怎么理解。

（18）spark 的 shuffle 过程

（19）hbase 怎么预分区

（20）redis 有多少曹，redis 的持久化，你们在项目中怎么用 redis 的，如何防止雪崩。

（21）mysql update 一条语句，需要走那些流程。

（22）了解 es 吗？

（23）yarn 执行一个任务用了多少 core

（24）produce 给 kafka 的一个 topic 发送消息，topic 的一个分区挂掉，他的副本也全部挂掉，这个 topic 还能用吗

（25）有没有用 javaapi 去读 kafka 的数据

（26）spark 读取 kafka 数据的方式？

（27）hive 的内表和外表的区别，hive 为什么要做分区，hive 的元数据存在哪？

（28）列式存储和关系型数据库的区别

（29）java 中正则表达式用到哪些类？

（30）unlock 和 synchronzed 的区别？

（31）编写 sparksql 的步骤？

2）项目部分

（1）写实时项目遇到了哪些问题

（2）拉链表的创建，连续三天登陆

（3）你们公司日活，月活，留存多少？你们团队多少人？

（4）为什么加最后一层 flume，不能直接 flume 到 hdfs 吗？

（5）分析指标，对 app 有什么影响。提升了百分之多少？

（6）有没有访问 kafka 的所有 topic 的权限

（7）有没有做 hive 表的权限管理。

3）算法部分

（1）归并排序

（2）删除单链表的一个节点

4）情景题：

（1）Spark:如果消费 Kafka 中的数据时，某个分区内数据因为种种原因丢失了一条(也可理解成因为某些条件被过滤了)，现在需要重新把这条数据纳入计算结果中。怎么解决？

（2）a left join b on a.id = b.id and b.age = 20

a left join b on a.id = b.id where b.age = 20

的结果有什么区别

5）平时学习习惯

（1）你最近在研究什么？

八 搜狐面试题

1）笔试部分

（1）hdfs 原理，以及各个模块的职责

（2）mr 的工作原理

（3）shell 脚本里如何检查文件是否存在，如果不存在该如何处理？Shell 里如何检查一个变量是否是空？

（4）Shell 脚本里如何统计一个目录下（包含子目录）有多少个 java 文件？如何取得每一个文件的名称（不包含路径）

（5）哪个程序同城与 nn 在一个节点启动？哪个程序和 DN 在一个节点？如果一个节点脱离了集群应该怎么处理？

（6）列出几个配置文件优化 hadoop，怎么做数据平衡？列出步骤

（7）有序数组的二分查找

（8）编写快速排序(QuickSort)或者归并排序(MergeSort)

（9）整数数组寻找满足条件的点对

（10）给定整数数组 A[N]，求两个下标(i, j)且 0<=i<j=N-1，注意：数组中有正整数或者负整数或者 0。

（11）该(i, j),使得 A[i] + A[i] + … + A[j] 最大

（12）要求：时间复杂度尽量低，空间复杂度 O(1)

（13）请：1.写出思路的简要描述 2.用你熟悉的语言，定义函数，完成编码和测试用例

2）项目部分

搜狐面试的小伙伴注意，广告变现的业务，会问到 flume 监控的问题，必问，二面理论比较强，居然问缓慢变化维，我问一下数据仓库有拉链表么？

九 360 面试题

1）技术部分

（1）常用的 linux 命令，shell 的 awk、sed、sort、cut 是用来处理什么问题的？

（2）hive 有自带的解析 json 函数，为什么还要自定义 UDF、UDTF

（3）json 的格式，json 存的具体数据

（4）mapreduce 过程

（5）shuffle 源码？其中 reduce 的分区是怎么设置的，针对每个 key，怎么把它放到对应的分区中？

（6）你了解的 hadoop 生态圈的框架，及其大概在怎样的一个位置？

（7）mapreduce、tez 和 spark 的区别？

（8）spark 读取文件如果内存不够的话，怎么处理？

2）手写部分

（1）自己写过 mapreduce 吗？怎么写的？

（2）最近七天连续三天活跃用户怎么实现的？手写一个各区域 top10 商品统计程序？

（3）平时遇到的数据倾斜的问题，怎么解决的？

（4）大表 join 大表怎么解决？

（5）每一层大概有多少张表？表的字段也需要记

（6）UDF、UDTF、UDAF 区别？

3）算法部分

（1）二叉树的前中后序遍历？

（2）排序算法了解过吗？

（3）快排的时间空间复杂度？快排原理

（4）冒泡的时间空间复杂度？原理

4）情景部分

写一个程序获取 ip（123.123.123.123）的地理位置信息，读取配置文件，返回结果 ip \t loc_nation \t loc_pro \t loc_city

```
ip_num_start  ip_num_end   loc_nation loc_pro loc_city ISP
1910946943     1910946945     中国    辽宁    沈阳    联通
1910946947     1910946949     中国    辽宁    沈阳    联通
1910946950     1910946950     中国    辽宁    鞍山    联通
1910946951     1910946953     中国    辽宁    沈阳    联通
1910946959     1910946965     中国    辽宁    沈阳    联通
1910946966     1910946966     中国    辽宁    盘锦    联通
1910946967     1910946985     中国    辽宁    沈阳    联通
1910946986     1910946986     中国    辽宁    大连    联通
1910946995     1910947033     中国    辽宁    沈阳    联通

sc.read("input/.txt")
.mapPartition(data=>{
```

```
    val splitdata = data.split("\t")
    ip_num_start  = splitdata(0)
    ip_num_end = splitdata(1)
    loc_nation = splitdata(2)
    loc_pro = splitdata(3)
    loc_city = splitdata(4)
    ISP = splitdata(5)
    (ip_num_start, ip_num_end, loc_nation, loc_pro, loc_city)
})
.map(data=>{
    if(ip ){
        (data.loc_nation,data.loc_pro,data.loc_city)
    }
})
```

针对上面代码的一些问题：

（1）上面那段代码中，如果文件是一个很大的文件，spark 读取的时候用一个任务处理（单机），怎样可以让它读取的效率更高一点？

（2）MapPartition 的原理是什么？

十 小米面试题

大家可以投一下小米的推荐算法。

一面：问两个有序大数组合并成一个数组的最佳方法，还有快排，问了所有项目。

二面：是问推荐算法的业务实现，还有如果你自己做一个推荐，你会怎么选择模型、算法。

三面：数据结构的查找，新增，删除的时间复杂度、还有各种数学问题，然后写了二叉树的查找、还有根据算法题写了一个递归，差不多就这样就过了

1）技术部分

（1）Java8 的新特性？

（2）Java8 的 hashMap 为什么采用数组+链表+红黑树？

（3）垃圾回收的算法，CMS 和 G1 的比较，用什么监控 JVM？

（4）Kafka 对接 sparkstreaming？

（5）线程池好处是什么，什么场合用？

（6）zookeeper 的选举机制，以及我们还可以用 zookeeper 做些什么？

（7）flume 采集数据的优缺点，还有什么可以代替它，也可以通过自己公司的业务情况自己开发。

（8）spark streaming 和 storm，Flink 的比较，以及各自的优势？

（9）怎么从 10 亿条数据中计算 TOPN？

十一 顺丰面试题

11.1 学长 1

第一面是做实时推荐同事，针对项目和知识点来问，特别喜欢问 spark，因为他用 kafka 和 spark 多，认真准备一下这关好过。

第二面是大数据这块的女总监，光跟你怼 spark，全是细节，细致到你传了那些参数

11.2 学长 2

1）技术部分

（1）kafka 原理，数据怎么平分到消费者

（2）flume hdfs sink 小文件处理

（3）flink 与 spark streaming 的差异，具体效果

（4）spark 背压机制具体原理实现

（5）spark 执行内存如何分配，执行任务时给多大内存

（6）sparksql 做了哪些功能

（7）讲一下 flink

（8）状态编程有哪些应用

（9）端到端 exactly-once 如何保证

（10）flink checkpoint 机制

（11）yarn 调度策略

2）项目部分

（1）离线指标、数据量、用户量

（2）介绍一下实时项目，哪些指标，怎么算的

（3）sparkstreaming 遇到什么问题，如何解决

（4）sparkstreaming 实现什么指标，怎么算的

（5）实时当天日活怎么累加

（6）集群规模

（7）一天的指标会用 sparksql 吗

（8）spark 手动设置偏移量，如果数据处理完后，offset 提交失败，造成重复计算怎么办

（9）项目中 es 做了哪些工作、es 实现原理、es 倒排索引怎么生成

（10）任务调度

（11）讲一下 Hbase、项目中哪些地方用了 Hbase、Hbase 写入流程、不同列族之间文件怎么划分

（12）kerberos 安全认证过程

十二 OPPO 面试题

OPPO -实时处理工程师。一面试官陈泉，他拿了多个人的简历，边翻边问。

1）技术部分

（1）sparkstreaming 消费方式及区别，spark 读取 hdfs 的数据流程

（2）kafka 高性能

（3）hive 调优，数据倾斜

（4）Zookeeper 怎么避免脑裂，什么是脑裂。

（5）Redis 的基本类型，并介绍一下应用场景

（6）最后会问一些 linux 常用命令，比如怎么查进程，查 io 运行内存等。还真有人问啊

2）项目部分

（1）hive 的分层设计

（2）还有一些 flume 和 kafka 的问题，为什么要把离线和实时搞在一起，可以做成两套系统。

十三 VIVO 面试题

13.1 学长 1

面试的数仓岗

13.1.1 一面

1）项目部分

（1）大概介绍一下你之前的工作

（2）做过哪些指标，怎么实现的(说出去的指标一定要会写 sql，会一个个问你怎么实现的）

（3）有没有做过留存用户，比如一个月，1 到 30 天的留存用户

（4）Hive 优化

（5）统计一个月内，连续登陆 7 天用户

2）情景部分

现场出了个 sql 题

uid event dt

1 点击 t1

2 收藏 t2

....

统计会话 session ：同一个用户事件时间间隔小于 1 分钟，算一次会话

13.1.2 二面

1）项目部分

（1）从数据源头开始问

（2）你们的数据源是什么，怎么埋点的，数据的格式是怎么样的

（3）是一次点击生成一个日志？，事件日志为什么会出现套 json

（4）启动日志和行为日志是怎么区分的

（5）多次点击是算一个事件还是多个，比如多次点赞，有没有做过类似点击次数的统计，你通过这个日志做了哪些指标，

（6）除了用户方面的，事件类的指标做过哪些，这些指标是怎么做的

（7）工作中有什么指标很难的，怎么实现的

vivo，好像会根据你的简历侧重，分配岗位，实时或者数仓，做好准备吧，骚年

13.2 学长 2

面试 Spark 开发岗。大家如果有去 vivo 的，一定要把简历上写到的知识点都搞懂。面试官是带着你的简历过来的，他事先有分析过你的简历。

1）项目部分

（1）先介绍你的项目，我介绍采集平台，刚说到 kafka。问双层 flume，第二层 flume 怎么对日志类型分类，分多少个类。如果一层 flume 新增了一个 topic，怎么处理。

（2）hive 表中的每层的血缘关系怎么查询

（3）每天业务量，问简历上的指标是怎么做的。

（4）hive 中用什么写 sql，hue。

（5）case class 和 class 的区别

（6）sparkStreaming 精准一次性消费，会问到你们用的什么版本的 kafka 和 sparkstreaming 。然后就是根据版本，这里 offset 的问题会展开详聊。

（7）写了 redis 缓存，会问你缓存机制。

（8）写了 hive 配置 mysql 高可用，会问你怎么配置的。

十四 华为面试题

1）华为非外包岗位，比较看重学历，要求 985，如果学历过了，技术基本不怎么卡。

2）外包总部 d1 区：

（1）namenode 内存满了，如何进行扩容，调什么参数。

（2）linux 命令怎么查看 mr 任务的 jobid

（3）kafka 消费者消费数据的流程，细节到进程和线程这一层。

（4）生产中遇到过哪些 OM 的情况，怎么解决

（5）namenode 设置多目录，在哪个配置文件里面设置，设置哪个参数？

十五 boss 直聘面试题

1）技术部分

（1）hadoop 二次排序

（2）小表关联大表怎么实现的|切片，shuffle，reduce 阶段，map 阶段，Yarn 流程

（3）combiner 使用场景 | 环形缓冲区为什么是环形的 | ETL 细节| reduce 阶段是怎么下载到本地

（4）hadoop 实现 TopN

（5）kvBuffer

（6）说一下你了解的 JVM 模型算法。为什么要使用复制算法(优势，劣势) 怎么查看 full gc 日志出现的问题。什么时候发生 full Gc。栈存储索引的大小。如何查看当前进程的 GC

（7）full GC 和 old GC 区别

（8）javaEE 说一下你擅长的框架

2）算法部分

（1）sort 快排手写

（2）io 模式 算法

（3）设计模式：单例模式、工厂、代理，代理与装饰模式区别，适用场景以及理解

（4）手写：冒泡、二分法

3）项目部分

（1）你的项目的亮点

十六 搜狗金融面试题

在茶水间面试，有很多求职者，都是一对一，并且不同面试官很可能是不同的大数据方向

1）技术部分

（1）RDD 是什么？

（2）MySQL 的索引如何理解？常用引擎是什么？有什么区别？比较 Redis 和 MySQL 的区别？说一下各自的持久化机制...

2）项目部分

（1）Kafka 的 offset、flume 的组成，项目中为什么用了两层？如何实时统计 pv 和 uv？

（2）给定一款产品，如何评价它的性能？

（3）对于自己统计出来的结果，如何评价结果的正确性？

（4）介绍主要的业务线，现在有没有 offer，职业规划是什么，对加班的看法，是否能接受数仓的工作....

3）手写部分

（1）手写代码 1：将两个无序数组合并，并保证合并后的数组有序，不允许用 Arrays 的方法

（2）手写代码 2：手写 wordcount，然后比较 flatmap 和 map 的区别、groupByKey 和 reduceByKey 的区别

（3）手写 HQL：给定两个表，查找 a 中有但是 b 中没有的元素

（select * from a left join b on a.id=b.id where b.id is null）

说明上述方式产生了几个 job？

4）算法部分

（1）将一个栈变成一个队列，至少用几个栈？

5）感受部分

面试大约 2 个小时，最后面试官介绍业务时才知道主要是做数仓，并且他问的这些问题，除了项目和后面的产品相关内容，基本上每一个问题都会落到数据结构上（RDD、索引、栈、队列、包括 groupByKey 和 reduceByKey...）

十七 瓜子二手车面试题

1）技术部分

（1）优化必问 hive/spark

（2）为什么 kafka 可以实现高吞吐？单节点 kafka 的吞吐量也比其他消息队列大，为什么？

（3）如果写了 JVM 调优：堆内存溢出如何查看解决，用的那些命令工具；CMS 和 G1 有什么不同？

（4）hive 自定义那些 udf 函数

（5）spark 那些算子是在 map 端聚合的

（6）kafka 的偏移量 offset 存放在哪儿，为什么？

（7）linux 命令 查看内存 磁盘 io 端口 进程

2）算法部分

（1）链表转置/二叉树转置

3）情景部分

（1）一个字符串"jasonbbtomccjackddfftomkk"，如果相邻两个字符相同视为一个切分点，实现 wordcount(注意会有多个连续相同的情况如 aabbccdd，还有要问清出现三个或者多个怎么算)

十八 猎豹移动面试题

1）技术部分

（1）对 Java 的理解

（2）Java 线程的理解

（3）对池的了解（线程池、数据库连接池），Java 线程池是怎么实现的，Java 线程池都有哪些组件，具体作用是什么

（4）对 Spring 框架的理解

（5）WebServer 端收集的日志放在哪里

2）项目部分

（1）介绍下项目，项目中用的那些技术点

3）手写部分

（1）手绘 Hadoop 架构

十九 作业盒子面试题

1）手写部分

（1）写一个 sql 将每个月的 top3 取出来 我用了三个子查询做出来不行

2）项目部分

（1）讲一下最近做的项目

（2）还有项目的 session 字段

（3）项目中的如何分辨数据（错误日志，正确日志）

（4）Hive 中的表的创建（项目中的表如何构建，创建的字段的组成）

（5）Hive 的优化，mysql 的优化

（6）Scala 算子的实际用法，场景

二十 每日优鲜

数仓岗

1）项目部分

（1）为什么用双层 flume

（2）日志标准化在哪里做最好

（3）hive 海量数据（一个分区）怎么导入 MySQL

（4）遇到过那些坑，自己认为做过有价值的地方

（5）MapReduce

（6）怎样可以让运维自己做查询

二十一 美团面试要求

1）技术部分

（1）HDFS 架构理解（基础）

（2）HDFS 源码/工作原理（高级）

（3）MapReduce 架构理解（基础）

（4）MapReduce 源码/工作原理/shuffle 原理（高级）

（5）MapReduce 二次排序（编程，必选）

（6）YARN 架构理解（基础）

（7）YARN 源码/工作原理（高级）

（8）HBase 架构理解（基础）

（9）HBase 源码/工作原理（高级）

（10）HBase 性能优化（高级）

（11）Hive 原理理解（基础）

（12）Hive 性能优化（高级）

（13）Flume 架构理解（可选）

（14）Kafka 架构理解（可选）

（15）Spark RDD 理解（基础）

（16）Spark reduceByKey 与 groupByKey 区别（基础）

（17）Spark Broadcast 与 Accumulator 功能（基础）

（18）Spark 工作原理（高级）

（19）Spark shuffle 原理（高级）

（20）Spark 源码理解/贡献（高级）

（21）Spark 性能优化/数据倾斜（高级）

2）项目部分

2.1）数仓理论

（1）数据仓库分层架构（初级）

（2）事实表与维度表（初级）

（3）星型模型与雪花模型 （初级）

（4）日增量表与日全量表 （初级）

（5）拉链表 (初级)

（6）缓慢变化维 （初级）

（7）两种建模理论（初级）

（8）data-vault 理论（高级）

2.2）数仓周边系统

（1）血缘关系

（2）调度系统

（3）展示，olap（kylin）

（4）指标管理

（5）数据质量管理 （高级）

（6）元数据管理，调度系统元数据，任务运行数据，表存储数据，数据字典等

3）手写 sql 部分

（1）点：窗口函数，行列转换，聚合函数

加一点复杂的：collect_set, split, case when，聚合函数， union 等

（2）函数考察

eg: rank, dense_rank, row_number 区别 （基础）

行列转换 （基础）

sort by, distribute by, order by, clustered by 区别

（3）3 个表关联，写下大概的 mr 过程 (中级)

（4）去重，2 个写法的区别:

select count(distinct id) from t1;

select count(1) from (select id from t1 group by id) a (高级)

map join 原理 （基础）

bloom filter 去重，原理(高级)

hive tunning

4）场景部分

（1）job，reduce 长时间百分之 99；

（2）reduce 的 3 个阶段，reduce 百分之 90+，又降下去，反复几次

（3）map 一直百分之 90 多，之后报错