**3.14号 海康研究院大数据开发面试**

面试时长：21分钟

问题1.介绍项目

1. hadoop的组件和高可用机制怎么实现
2. Java GC(答得不好)

“你能不能谈谈，java GC是在什么时候，对什么东西，做了什么事情？”

在什么时候：

1.新生代有一个Eden区和两个survivor区，首先将对象放入Eden区，如果空间不足就向其中的一个survivor区上放，如果仍然放不下就会引发一次发生在新生代的minor GC，将存活的对象放入另一个survivor区中，然后清空Eden和之前的那个survivor区的内存。在某次GC过程中，如果发现仍然又放不下的对象，就将这些对象放入老年代内存里去。

2.大对象以及长期存活的对象直接进入老年区。

3.当每次执行minor GC的时候应该对要晋升到老年代的对象进行分析，如果这些马上要到老年区的老年对象的大小超过了老年区的剩余大小，那么执行一次Full GC以尽可能地获得老年区的空间。

对什么东西：从GC Roots搜索不到，而且经过一次标记清理之后仍没有复活的对象。

做什么：

新生代：复制清理；

老年代：标记-清除和标记-压缩算法；

永久代：存放Java中的类和加载类的类加载器本身。

GC Roots都有哪些：

1. 虚拟机栈中的引用的对象

2. 方法区中静态属性引用的对象，常量引用的对象

3. 本地方法栈中JNI（即一般说的Native方法）引用的对象。

**3.21日 阿里一面**

1. spark的运行原理
2. Flink的原理
3. 数据仓库
4. 项目细节
5. 职业规划
6. 项目优化的点

**3.26远景智能：**

1. 数据结构：有哪些种链表，map
2. 多线程问题 线程池有哪些
3. jvm问题
4. Java 锁的问题 锁 有哪些

https://www.nowcoder.com/discuss/167801

1. 设计模式

https://www.cnblogs.com/zhaojinyan/p/9401010.html

1. Hbase结构 负载均衡
2. Spark rdd

**3.27 招行信用卡中心**

1.项目，项目难点，怎么解决

2.前端会写什么 spring原理

3.flink和spark的区别

4.flink的容错机制

5.hbase的底层储存结构

4.19号二面问题

1. 聚簇索引和非聚簇索引
2. 图数据库和时序数据库
3. spark的集群部署
4. Nosql 怎么实现事务

其他是常规hr面问题

**4.11 微众银行一面**

时间 1小时

1. java的值传递和引用传递 **JAVA是值传递！！**
2. 优先级队列怎么实现的 通过堆来实现的
3. 抽象类和接口
4. Static关键字
5. valitile关键字底层怎么实现的
6. 各种设计模式的实现
7. List set 的继承关系
8. 项目的源码

**4.22拼多多一面**

1. 从网页输入网址到页面展示的全过程，涉及到7层网络每层的什么协议
2. 为什么mysql索引选择B+树
3. 现场编程题：输入字符串，输出为最长的不重复子串

输入abcdcefgh 输出dcefgh

**5.6 字节跳动一面**

1.项目介绍 很详细 会仔细问每一个细节

2.写sql join left join

3.spark算子 reducebykey和groupbykey的区别

coalesce和repatition的区别

4.hive的内部表和外部表 怎么删除外部表

5.现场编程

给定一个有序数组arr {1,2,3,4,5,5,5,6,6,6,6,7}再给定int num求num在数组中的出现次数

一开始说的暴力遍历，问时间复杂度，后面问有没有优化的方法，答了二分法，时间复杂度为log2N;

在线实现出来

**5.7号360企业安全面试**

有点水 没啥技术问题 主要问能不能出差

**5.7 爱奇艺一面**

Flume介绍

Hive内部表外部表

Hive解决数据倾斜

Left join和left semi join

多线程的创建

抽象类与接口

Mysql的索引 b+树，hash树

Mysql的联合索引

Mapreduce介绍，shuffle过程介绍

5.9 爱奇艺二面

1.两列 date uid 怎么查询每日新增用户

2.mysql的联合索引问题

3.数仓介绍 数仓分层介绍

4.维度表与事实表

5.项目介绍

**5.8虎牙直播一面**

1. 项目介绍 全程被怼 被说思路有问题。。。
2. hdfs高可用的实现
3. namenode和datanode 有没有看过源码
4. 1个T的文件 求出现次数最高的10个单词 （回答了分治法，说不行，让想优化的办法）
5. 衡量大数据程序优劣的尺度（面试官提示是磁盘IO）
6. 多线程的wait notify 回答出来还是不行，要求介绍源码
7. list的接口实现类（arraylist linkedlist vector）

2019.6.12 VIVO 提前批面试

手撕代码，删除链表中的重复结点

问项目，问的非常细

Spark 项目发生问题如何定位

怎么看每个结点运行时处理的文件大小

Hive运行发生错误怎么定位

数据倾斜怎么解决

Hive sql查看运行计划

Hive窗口函数，转换函数，map转string

Hdfs文件储存格式

Hbase Key 设计规范，怎么打散

数仓分层，知道哪些指标，比如留存率

Python写过什么项目

Shell写过什么程序

Sqoop导数据，有哪些参数，支持哪些数据格式

Redis我主动回答了数据结构，应用场景，回问，那怎么删除什么数据吧，这个不会

OOM了怎么定位

2019.7.25日 远景智能一面

1. 项目 自研的图数据库比jaunsgrpah改进点

①janusgraph使用es索引的时候每个图只支持一个index 这样图数据几百亿之后，es不仅性能变差，还有可能爆掉，我们改进了建索引方法，改为动态索引，初始化建立五个index 用hash取余的方式，写数据进对应的index，当单个索引大小超过阈值，自动再建5个新索引，这样保证了对于百亿级甚至千亿级别数据的支持

②janusgraph没有高可用模块，数据批量导入导出不支持断点续传，图计算性能差，我们改进后现在已经支持百亿级别的六阶关系推演，使用ssd作为计算中间数据的缓存

③janusgraph没有解决数据一致性问题，即原始数据写入hbase和索引写入es，有一个出问题最后会出现数据不一致

④janusgraph没有并发控制

1. 个人主要负责es版本升级，数据迁移，client选型，少属性写入优化

①为什么要升级

我们从es2.3.5升级到5.6.4

5.X版本新功能：

在数据结构方面，新增了多个 range 字段类型，有什么用呢，现在你可以计算连续数据的交并集，可以是时间范围，也可以是数值范围。

比如数据存放的是会议信息，航班有一个 range 字段，里面存的是会议的开始和结束时间，你通过对应的 range 查询可以很方便的查询，得到某个时间点哪些会议同时正在进行，那段时间会议室有空闲，可以预订等等。

ES 的搜索，对于一些耗时较长的查询，以前是没有办法取消的，除了干掉节点重来或者等待结束，没有办法，现在可以通过 ES 的任务管理机制来进行取消了

还有　ES5.0在Internal engine级别移除了用于避免同一文档并发更新的竞争锁，带来15%-20%的性能提升

以前的string类型被分成Text和Keyword两种类型，keyword类型的数据只能完全匹配，适合那些不需要分词的数据，

对过滤、聚合非常友好，text当然就是全文检索需要分词的字段类型了。将类型分开的好处就是使用起来更加简单清晰，以前需要设置analyzer和index，并且有很多都是自定义的分词器，从名称根本看不出来到底分词没有，用起来很麻烦。

1. 以前的hive hbase监控项目
2. Mysql的脏读幻读不可重复读，四种隔离级别Serializable (串行化)，Read committed (读已提交)Repeatable read (可重复读），Read uncommitted (读未提交)

详见 面试题文档 10页

1. redis 持久化
2. kafka的零拷贝
3. spark任务执行流程，task分配，task个数

对于数据读入阶段，例如sc.textFile，输入文件被划分为多少InputSplit就会需要多少初始Task。

在Map阶段partition数目保持不变。

在Reduce阶段，RDD的聚合会触发shuffle操作，聚合后的RDD的partition数目跟具体操作有关，例如repartition操作会聚合成指定分区数，还有一些算子是可配置的。

RDD在计算的时候，每个分区都会起一个task，所以rdd的分区数目决定了总的的task数目。

1. flink有什么了解，和spark相比的优缺点
2. 大数据实时处理存在哪些难点
3. java多线程，每种线程池的优缺点，还有参数有什么
4. hdfs读的流程
5. 2T的数全排列