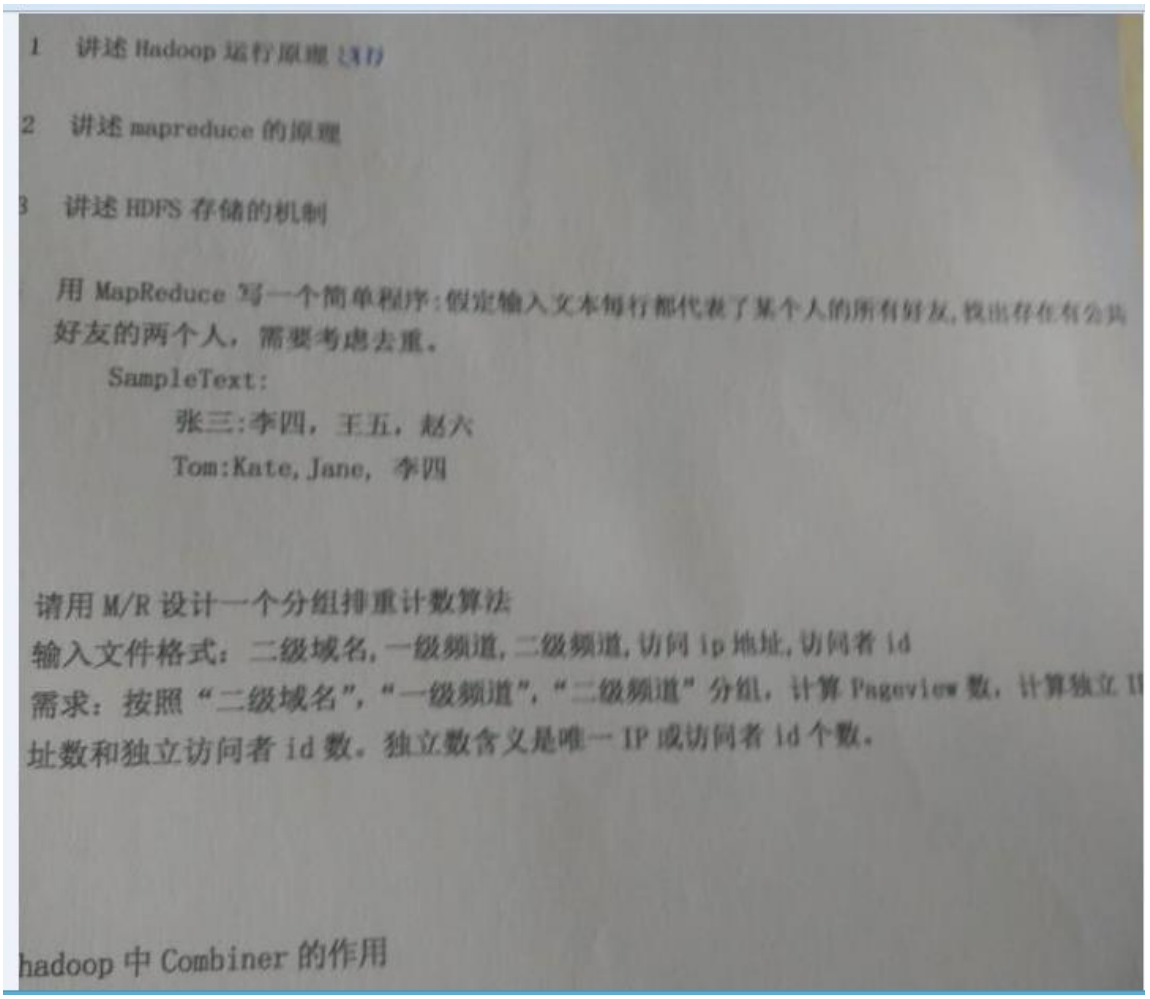
(一) Hadoop5 道题目



1 讲述 Hadoop 运行原理 (3分)

2 讲述 mapreduce 的原理

3 讲述 HDFS 存储的机制

用 MapReduce 写一个简单程序：假定输入文本每行都代表了某个人的所有好友，找出存在有公共好友的两个人，需要考虑去重。

SampleText:

张三:李四，王五，赵六

Tom:Kate,Jane, 李四

请用 M/R 设计一个分组排重计数算法

输入文件格式：二级域名,一级频道,二级频道,访问 ip 地址,访问者 id

需求：按照“二级域名”,“一级频道”,“二级频道”分组，计算 Pageview 数，计算独立 IP 地址数和独立访问者 id 数。独立数含义是唯一 IP 或访问者 id 个数。

hadoop 中 Combiner 的作用

(二) 优酷，乐视，中国电信

电信云计算：

1、单例；

2、LinkdList 和 ArrayList 的区别；

3、http 传输的时候是明文的，怎么解决安全问题；

4、spark streaming 和 storm 的区别，可以相互取代吗；

5、垃圾回收机制；

乐视：

笔试题：

1.Design a singleton class in multi-threading environment.

2.How to find middle element of linked list in one pass.

3.Reverse singly linked list in java recursively.

4.Write a java program to implement Stack in java.

面试问题：

1、scala 中的隐式函数的关键字；

2、val x=y=1 结果是什么；

3、java 内存模型；

4、spark 运行的 job 在哪里可以看到；

5、hadoop 中一个 job 提交到 resourcemanager 之后，resourcemanager 会生成一个什么样的容器来放这个 job；

优酷：

1、java 内存模型；

2、java 垃圾回收机制；

3、编译好的 scala 程序，运行的时候还需要 scala 环境吗；

4、object 中有哪些方法；

5、如何监测集群中 cpu，内存的使用情况，比如说：有一个 spark 特别占资源，特别慢，怎么排查这种情况；

6、ArrayList 中 Array 的长度超了是怎么增加的，一次增加多少；

(三) 20 道面试题

四川启明星银海科技（笔试题）

一、选择题(34 分，每题 1 分)

共 34 道，比较基础。

二、问答题（11 分，每题 1）

前 5 道是写出程序的输出结果。

6、String s = new String("xyz");创建了几个 String Object?并作说明。

7、数组有没有 length()这个方法? String 有没有 length()这个方法？

8、short s1 = 1; s1 = s1 + 1;有什么错? short s1 = 1; s1 += 1;有什么错?

short s1 = 1; s1 = s1 + 1;有错，s1 是 short 型，s1+1 是 int 型,不能显式转化为 short 型。可修改为 s1 =(short)(s1 + 1) 。

short s1 = 1; s1 += 1 正确。

9、ArrayList 和 Vector 的区别,HashMap 和 Hashtable 的区别

10、STRING 与 STRINGBUFFER 的区别。

11、ArrayList 与 Vector 的区别

三、数据库（15 分,每题 5 分）

共 3 道，比较难，写出 sql 语句。考的是索引。要求写出的 sql 语句的效率要非常的高。都是大数据量的表。

四、编程（40 分）

1、写出表达式注释部分的函数（表达式记不起了）（10 分）

2、设计 4 个线程，其中两个线程每次对 j 增加 1，另外两个线程对 j 每次减少 1。写出程序。 （15 分）

3、写一个 Singleton 出来。

Singleton 模式主要作用是保证在 Java 应用程序中，一个类 Class 只有一个实例存在。（15 分）

面试部分：

1、请自我介绍一下。

2、请简单介绍一下工作经历，以及在项目中主要负责什么？

3、你是否愿意常期出差？

4、怎样提高程序的运行效率？

5、假如你被我公司录用，我现在要派你去一个你不愿意去的地方，你该怎么办？

6、Struts 熟悉吗？

7、在你的项目中用的是什么框架？

8、你用过哪些数据库？

9、你为什么要离开你原来的公司？

10、你为什么要来我们公司？

11、 前台熟悉吗？会用 js 吗？

(四) 11 道题目

面试问题总结:

1.阿里巴巴的电话面试问到了 linux 的详细启动过程. 当时没回答出来...

1.按下电源

2.BIOS 自检

3.系统引导(lilo/grub)

4.启动内核

5.初始化系统

2.问到了还使用过其他什么开源框架.

我就把我知道的 apache 的开源框架扯了一遍

3.问到了最擅长那种技术

我就说我对 hadoop 生态系统的技术比较擅长,

然后介绍了 mapreduce 的详细过程和 hsfs 的架构.

4.问到了 hive 的优化

因为 hive 底层是 mapreduce,所以就把 maprecue 的优化介绍了一下.

比如多个 map 任务会比较消耗系统的资源,那么,在执行操作前,

应该预先把小文件合并为大文件.

reducer 的数量可以根据公式去配置(hadoop 文档中推荐的)

0.95*NUMBER_OF_NODES*mapred.tasktracker.reduce.tasks.maximum

1.75*NUMBER_OF_NODES*mapred.tasktracker.reduce.tasks.maximum

备注：NUMBER_OF_NODES 是集群中的计算节点个数；

在代码中可以通过:jobConf.setNumReduceTasks(int numOfreduceTasks);方法设置 reducer 的个数

可以写一个 UDF 函数,并且在建 hive 表的时候指定好分区，这样就提交了数据扫描的效率

可以修改配置文件,配置在 map 端的合并为 true

5.问到了在实际开发的过程中遇到了什么问题.

我回答了.hbase 的开发的过程中运行很慢,有时候会宕机.

然后通过查看日志和观察配置文件发现,hfile 的内容过小，到职 hbase 频繁的 split. 可以修改 hfile 的大小来避免这种情况.

然后把 major compact 也设置为手动,

6.然后问我 major compack 设置为手动会出现什么问题.....

没有回答出来,,感觉实际开发中应该不会这样去处理吧....

7.很多公司都会问 zookeeper 的二次开发.

8.有些公司会问的非常详细,,,比如 flume 在实际项目里面的数据采集....

当时老师讲的是通过 java 端有一个 jar 包里面有个类作为客户端,,,将数据发送到 flume,,,,

但是面试官说: 这样项目和 flume 的耦合就比较高,,,如果 flume 挂掉了,,,对项目有影响....

实际开发的过程中,,是利用 flume 的命令去定时采集数据的.

还问到了 flume 的实时采集数据和定时采集数据的方法,,,

回答的不是太好,,,

9.有公司问到了 mongoDB 和 hbase 的区别....

不是太了解,,就把 hbase 的概念给描述了一下

10.还有不少公司会问...

你做大数据的工作,,感觉自己工作里面做的最好的是哪一块..

我觉得这样的问题真心不好回答,,,这很明显是个陷阱.....一旦跳进去就等着被虐吧....

11.有不少公司会问项目组有多少人,,,人员的分工是如何的,,,数据量,,,还有集群的配置....

这些东西一定要想清楚啊,,,问的概率太高了...

我回答每天数据量 100G 人家觉得太少了..

(五) 33 道题目

单项选择题

1. 下面哪个程序负责 HDFS 数据存储。C

a)NameNode b)Jobtracker c)Datanode d)secondaryNameNode e)tasktracker

2. HDfS 中的 block 默认保存几份？ A

a)3 份 b)2 份 c)1 份 d)不确定

3. 下列哪个程序通常与 NameNode 在一个节点启动？D

a)SecondaryNameNode b)DataNode c)TaskTracker d)Jobtracker

4. Hadoop 作者 D

a)Martin Fowler b)Kent Beck c)Doug cutting

5. HDFS 默认 Block SizeB

a)32MB b)64MB c)128MB

6. 下列哪项通常是集群的最主要瓶颈 C

a)CPU b)网络 c)磁盘 d)内存

7. 关于 SecondaryNameNode 哪项是正确的？ C

a)它是 NameNode 的热备 b)它对内存没有要求

c)它的目的是帮助 NameNode 合并编辑日志，减少 NameNode 启动时间

d)SecondaryNameNode 应与 NameNode 部署到一个节点

多选题：

8. 下列哪项可以作为集群的管理工具 ABD

a)Puppet b)Pdsh c)Cloudera Manager d)d)Zookeeper

9. 配置机架感知的下面哪项正确 ABC

a)如果一个机架出问题，不会影响数据读写

b)写入数据的时候会写到不同机架的 DataNode 中

c)MapReduce 会根据机架获取离自己比较近的网络数据

10. Client 端上传文件的时候下列哪项正确 B

a)数据经过 NameNode 传递给 DataNode

b)Client 端将文件切分为 Block，依次上传

c)Client 只上传数据到一台 DataNode，然后由 NameNode 负责 Block 复制工作

11. 下列哪个是 Hadoop 运行的模式 ABC

a)单机版 b)伪分布式 c)分布式

12. Cloudera 提供哪几种安装 CDH 的方法 ABCD

a)Cloudera manager b)Tar ball c)Yum d)Rpm

判断题：0 代表错误，1 代表正确

13. Ganglia 不仅可以进行监控，也可以进行告警。（0 ）

14. Block Size 是不可以修改的。（0 ）

15. Nagios 不可以监控 Hadoop 集群，因为它不提供 Hadoop 支持。（0 ）

16. 如果 NameNode 意外终止，SecondaryNameNode 会接替它使集群继续工作。（0 ）

17. Cloudera CDH 是需要付费使用的。（ 0）

18. Hadoop 是 Java 开发的，所以 MapReduce 只支持 Java 语言编写。（ 0）

19. Hadoop 支持数据的随机读写。（0）

20. NameNode 负责管理 metadata，client 端每次读写请求，它都会从磁盘中读取或则会写入 metadata 信息并反馈 client 端。（ 1）

21. NameNode 本地磁盘保存了 Block 的位置信息。（1 ）

22. DataNode 通过长连接与 NameNode 保持通信。（ 0）

23. Hadoop 自身具有严格的权限管理和安全措施保障集群正常运行。（ 0）

24. Slave 节点要存储数据，所以它的磁盘越大越好。（ 0）

25. hadoop dfsadmin －report 命令用于检测 HDFS 损坏块。（ 0）

26. Hadoop 默认调度器策略为 FIFO（1 ）

27. 集群内每个节点都应该配 RAID，这样避免单磁盘损坏，影响整个节点运行。（ 0）

28. 因为 HDFS 有多个副本，所以 NameNode 是不存在单点问题的。（0 ）

29. 每个 map 槽就是一个线程。（0 ）

30. Mapreduce 的 input split 就是一个 block。（ 0）

31. NameNode 的 Web UI 端口是 50030，它通过 jetty 启动的 Web 服务。（ 0）

32. Hadoop 环境变量中的 HADOOP_HEAPSIZE 用于设置所有 Hadoop 守护线程的内存。它默

认是 200 GB。（ 0）

33. DataNode 首次加入 cluster 的时候，如果 log 中报告不兼容文件版本，那需要 NameNode

执行“Hadoop namenode -format”操作格式化磁盘。（0 ）

参考：http://blog.csdn.net/lulongzhou_llz/article/details/39153961

(六) 相关优化题目

hbase 优化

表设计

分区，

rowkey 设计，设置定长（64 字节）

CF 不要太多，2~3 个

Max Versio

Time To Live

Compact & Split

写表

多 HTable 并发写

HTable 参数设置 手动 flush,降低 IO

Write Buffer

批量写

多线程并发写

读表操作

多 HTable 并发读

HTable 参数设置 1）在 HBase 的 conf 配置文件中进行配置；设置二哥 cache 参数

批量读

释放资源

缓存查询结果

MpReduce 优化

1. 任务调度

I/O 方面：Hadoop 会尽量将 Map 任务分配给 InputSplit 所在的机器，以减少网络 I/O 的消耗。

2.数据预处理与 InputSplit 的大小

合理地设置 block 块大小是很重要的调节方式。除此之外，也可以通过合理地设置 Map 任务的数量来调节 Map 任务的数据输入。

3. Map 和 Reduce 任务的数量

当 Reduce 任务的数量是任务槽的 1.75 倍时，执行速度快的机器可以获得更多的 Reduce 任务，因此可以使负载更加均衡，以提高任务的处理速度。

4. Combine 函数

MapReduce 框架运行用户写的 combine 函数用于本地合并，这会大大减少网络 I/O 操作的消耗

\

Oracle 千万级别优化

1、建立视图

2、优化 sql 语句

3、分表、分库

4、存储过程

(七) openstack 常问

网络类型：

外部网络

公共网络，外部或 Internet 可以访问的网络

内部网络

私有网络，仅内部访问的网络

管理网络，用于 openstack 组件以及 mysql DB Server，RabbitMQ messaging server 之间的通信

存储网络，用于 storage/olume 节点与计算节点间的 iSCSI volume traffic

服务网络，用于租户 VLAN/subnets 中的实例的固定 IP 地址

Openstack 服务网络地址管理

固定 IP

私有 IP 地址，用于租户实例间通信

浮动 IP

公共 IP 地址，用于实例与外部或 Internet 的通信

公共 IP 地址不一定是 Internet 上可路由的地址，也可以是站点内部或局域网的地址

私有地址和公共地址的关系以及必要的路由由 nova-network 来处理，实例不必考虑此问题。

1.nova 工作原理

2.libvirt 有什么作用

3.通过程序编程对 kvm 设计过什么，

4.openstack 你认为哪个模块最复杂

5.openstack 网络流量是怎么流通的，比如 flatdhcp 比如 vlan 模式

6.你对 openstack 怎么理解，他是个什么东西

7.你对虚拟化怎么理解，为什么需要虚拟化

(八) scala 题目

1.scala 语言有什么特点，什么是函数式编、程有什么优点

2.scala 伴生对象有什么作用

3.scala 并发编程是怎么弄得，你对 actor 模型怎么理解有何优点

4.scala case class 有什么重要

5.scala akka 框架有没有接触过，有什么重要

6.scala 为什么设计 var 和 val

(九) spark 总结题目

1.spark sdd　dag ,stage 你怎么理解的

rdd:resilent destributed dataset,弹性分布式计算集，spark 的抽象数据结构类型，任何的数据在 spark 中都表示为 RDD，它的数据是分布式分区存储的，同分区的数据有可能在不同的集群节点上

dag:有向无环图,spark 里有一个 dagscheduler 专门负责在程序执行之前建立，在图论中，如果一个有向图无法从任意顶点出发经过若干条边回到该点，那么这个图就叫做有向无环图

stage:一个 job 分为很多个任务（task）,每组任务被称为一个 stage(阶段)

2.spark 宽依赖 窄依赖你怎么理解的

窄依赖指的是父 RDD 的 partition 被一个子 RDD 的 partition 所用,它属于 spark 的一类优化机制,将一组对 RDD 的宽依赖操作作为一个任务(也就是一个流水线),这样能够减少中间数据的计算过程,起到 1+1+1=3 的效果

宽依赖指的是父 RDD 的 partition 和子 RDD 的 partition 是完全打散的,这是因为它属于 shuffle 类的操作,这涉及到了大量的数据迁移操作.

3.stage 是基于什么原理分割为 task 的。。

就是宽依赖和窄依赖的原理,说的见到一点,有 shuffle 的操作就是一个宽依赖,没 shuffle 的操作就是一个窄依赖

4.血统的概念

spark 通过血统记录了 RDD 的演变过程,当一个 RDD 操作失败时会通过血统找到它依赖的 RDD 来迭代恢复数据(这句话的意思就是一层一层的往上找 RDD,直到找到保存成功的 RDD)

5.任务的概念

spark 的一个任务就是一个 pipeline

6.Application

基于 spark 的用户程序,包含 driver 和 executor

7.容错方法

检查点和日志,spark 采用的是日志容错

8.怎么理解粗粒度和细粒度

spark 操作的 RDD 就属于粗粒度的数据，而 Hadoop 操作的基本单元就是细粒度。

spark 的优越性

其中内存计算、数据本地性 （locality）和传输优化、调度优化等该居首功，也与设计伊始即秉持的轻量理念不无关系。

application 工作在两个空间中,spark RDD 空间和 Scala 原生数据空间.

在原生数据空间中,数据表现为标量(scalar,即 scalar 基本数据类型)、集合类型和持久存储。

输入算子：将 Scala 集合或者存储数据吸入到 RDD 中，输入算子由此分为两种，操作集合的是 paralize，操作存储数据的是 textFile，它的输出就是 spark 的 RDD

变换算子（transformation）：RDD 经过变换算子生成新的 RDD，一个 RDD 被分成很多的 partition 分配到集群的各个节点中去。这里需要注意的是 partition 是一个逻辑概念，变化前后的 partition 可能是同一块内存或者存储，这是很重要的，因为它可以使得有可能对前后 RDD 的操作在同一块内存中，这样就减少了更多内存的使用。其中有些 RDD 是中间结果，它的分区不一定有相应的内存与之对应，在这时，如果需要将这个结果保存下来，就需要调用缓存算子将 RDD 物化下来。

一部分变换算子把 RDD 看做简单元素，分为如下几类：

输入输出一对一，并且 RDD 的分区结构不变，主要是 map 和 flatmap

输入输出一对一，但是 RDD 的分区结构发生改变，如 union

从输入中选择部分元素的算子，如 filter，distinct 另一部分变换算子针对 key-value 集合，又分为： 对单个 RDD 做 element-wise（元素级）的运算，如 mapValues，这是在同一个 partition 上做的操作 对单个 RDD 做重排，如 sort 和 partitionBy（实现一致性的分区划分，这个对数据本地化非常重要） 对单个 RDD 基于 key 进行重组和 reduce，如 groupByKey 和 reduceByKey 对两个 RDD 基于 key 进行 join 和重组，如 join，cogroup 需要注意的是，后三类都涉及 shuffle 操作
1.spark 为什么快，你能说出你的几点看法吗？ 2.spark 里面的 transformation action 是干什么的，有什么区别，能举例说明几个常用方法吗？ 3.sdd 你怎么理解的 4.spark 作业提交流程是怎么样的，client 和 cluster 有什么区别，各有什么作用， 5.spark on yarn 作业执行流程，yarn-client 和 yarn cluster 有什么区别 6.spark streamning 工作流程是怎么样的，和 storm 比有什么区别 7.spark sql 你使用过没有，在哪个项目里面使用的 8.spark 机器学习和 spark 图计算接触过没，，能举例说明你用它做过什么吗？
spark sdd 是怎么容错的，基本原理是什么？

(十) 3 道题目

1.spark sdd dag ,stage 你怎么理解的

2.spark 宽依赖 窄依赖你怎么理解的

3.stage 是基于什么原理分割为 task 的。。

(十一) 1 道 java 场面题目

Java HashMap 工作原理

http://www.importnew.com/16599.html

(十二) 复习计划题目

根据简历所写的内容，接下来准备做如下复习：

0.复习 hadoop 集群中各组件的功能和之间关系，并能够流畅表达出来

1.复习 MapReducer 的 shuffer 过程的具体实现，并能够流畅表达出来

2.复习 storm 体系结构，各组件功能和之间关系，并能流畅表达出来

3.了解 zookeeper 原理，以及使用场景

4.复习 hbase 性能优化，做到能够流畅的说出如何优化 hbase

5.redis 集群的搭建,能够根据企业的数据大小，给出针对性的集群搭建服务器分配的建议

6.sqoop,flume,kafka 的使用（能够说出这些工具如何使用，以及使用场景）

(十三) 7 道面试题

1．关于 string 的选择题

2．修饰符

3．两个整数相除，保存到百分位

4．写个封装，抽象，继承，多态的类集。

5．jsp 的作用域的描述。

6．如何得到 jsp 的参数。

7．写几个 SQL 语句。

面试：自我介绍。

给个产品是旧品改良 还是研发新产品的会议讨论，分为两组讨论，最后统一结果。

(十四) 6 道面试题

无双科技公司

1. spark 为什么比 hadoop 快？

首先，Spark 对分散的数据集进行抽样，创新地提出 RDD(ResilientDistributedDataset)的概念，所有的统计分析任务被翻译成对 RDD 的基本操作组成的有向无环图(DAG)。RDD 可以被驻留在 RAM 中，往后的任务可以直接读取 RAM 中的数据；同时分析 DAG 中任务之间的依赖性可以把相邻的任务合并，从而减少了大量不准确的结果输出，极大减少了 HarddiskI/O，使复杂数据分析任务更高效。从这个推算，如果任务够复杂，Spark 比 Map/Reduce 快一到两倍。

没有额外的复制，序列化，磁盘 IO 开销

DAG

2. rdd 的处理过程是什么，不要说概念

亿玛在线公司

1. 说说 hbase 的 API 都有哪些 filter?

2. 说说你用过的 storm?

3. 日志表中的数据使用 hive 怎么实现，mapreduce 怎么实现？题目见附件

美团网

1. 自己熟悉大数据的部分说一下？

2. hadoop 与 storm、spark 的比较？

3. 对一个字符串进行全排列？

4. 事务都有哪些特点？

5. hadoop 集群中的某个 block 不能 copy 数据到其他节点，怎么办？如果并发量大了，

有多个 block 不能 copy 数据，怎么办？

LeanCloud

电话面试

1. 自我介绍？

2. Scala 一些基础的问题，如：伴生对象，类的问题，有哪些 class？

筑巢新游

1. flum 是如何导入数据到 kafka?具体

无双科技公司

1. spark 为什么比 hadoop 快？

首先，Spark 对分散的数据集进行抽样，创新地提出 RDD(ResilientDistributedDataset)的概念，所有的统计分析任务被翻译成对 RDD 的基本操作组成的有向无环图(DAG)。RDD 可以被驻留在 RAM 中，往后的任务可以直接读取 RAM 中的数据；同时分析 DAG 中任务之间的依赖性可以把相邻的任务合并，从而减少了大量不准确的结果输出，极大减少了 HarddiskI/O，使复杂数据分析任务更高效。从这个推算，如果任务够复杂，Spark 比 Map/Reduce 快一到两倍。

没有额外的复制，序列化，磁盘 IO 开销

DAG

2. rdd 的处理过程是什么，不要说概念

亿玛在线公司

1. 说说 hbase 的 API 都有哪些 filter?

2. 说说你用过的 storm?

3. 日志表中的数据使用 hive 怎么实现，mapreduce 怎么实现？题目见附件

58 同城

1. 请使用 awk, sed 命令对文本文件中第二列和第三列取出来？
2. spark 原理，并写出你用过的 spark 代码？
3. 用代码写出你使用过的 mapreduce？

talkingDate

1. 请说出你在 spark 中的优化方案？
2. 你在项目中使用的技术，解决了什么问题？

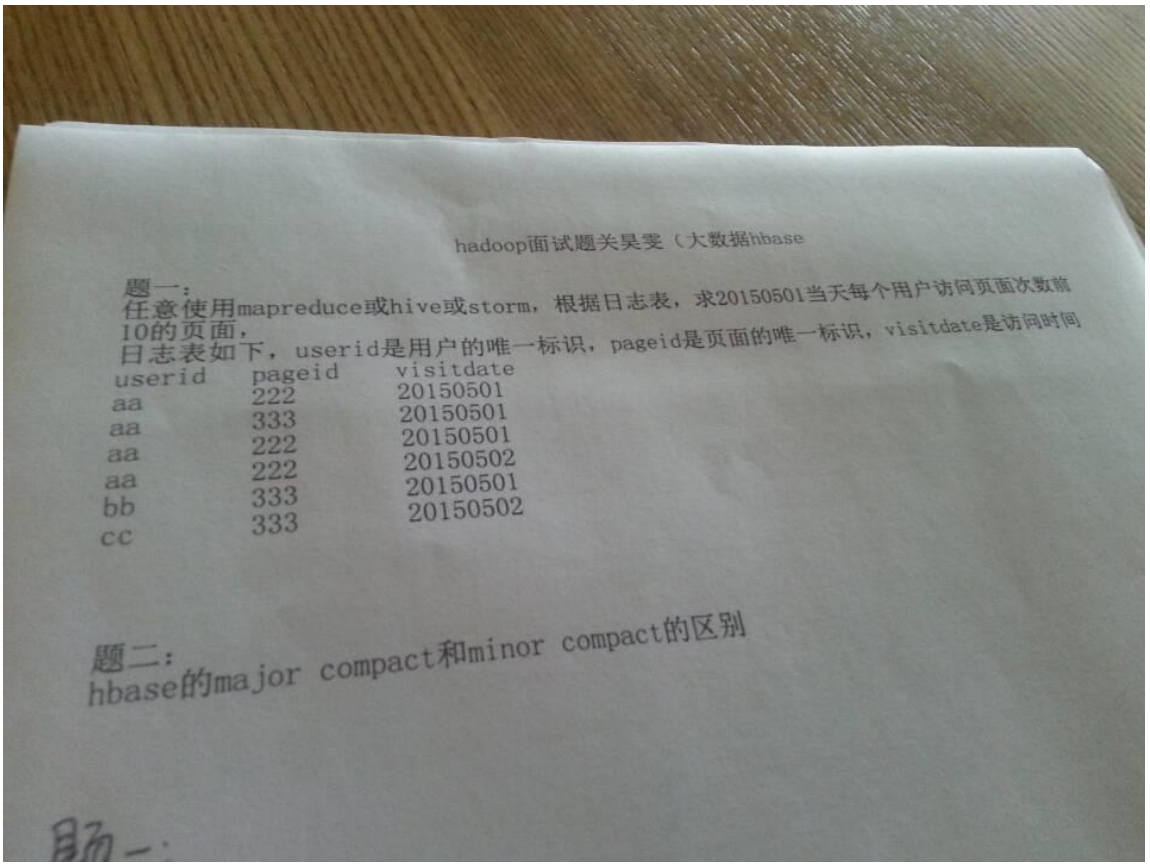
慕华信息科技有限公司

1. 用户文件 2 个属性 10 万行，课程文件 2 个属性 2 万行，日志文件 1 个属性很大，这些属性可以任意的组合查询，每秒的请求数是 10000 个，请如何最快的方式查询出数据？

2. 给你 2 个字符串，字符串最后一个字符可以循环到最前面，只要一个字符串中每一个字符在另一个字符串都有就是相等的，请用你的方法比较是否相等？

3. 在你做的项目中所使用到得技术或者工具，都是做什么的？

(十五) 20 道题目

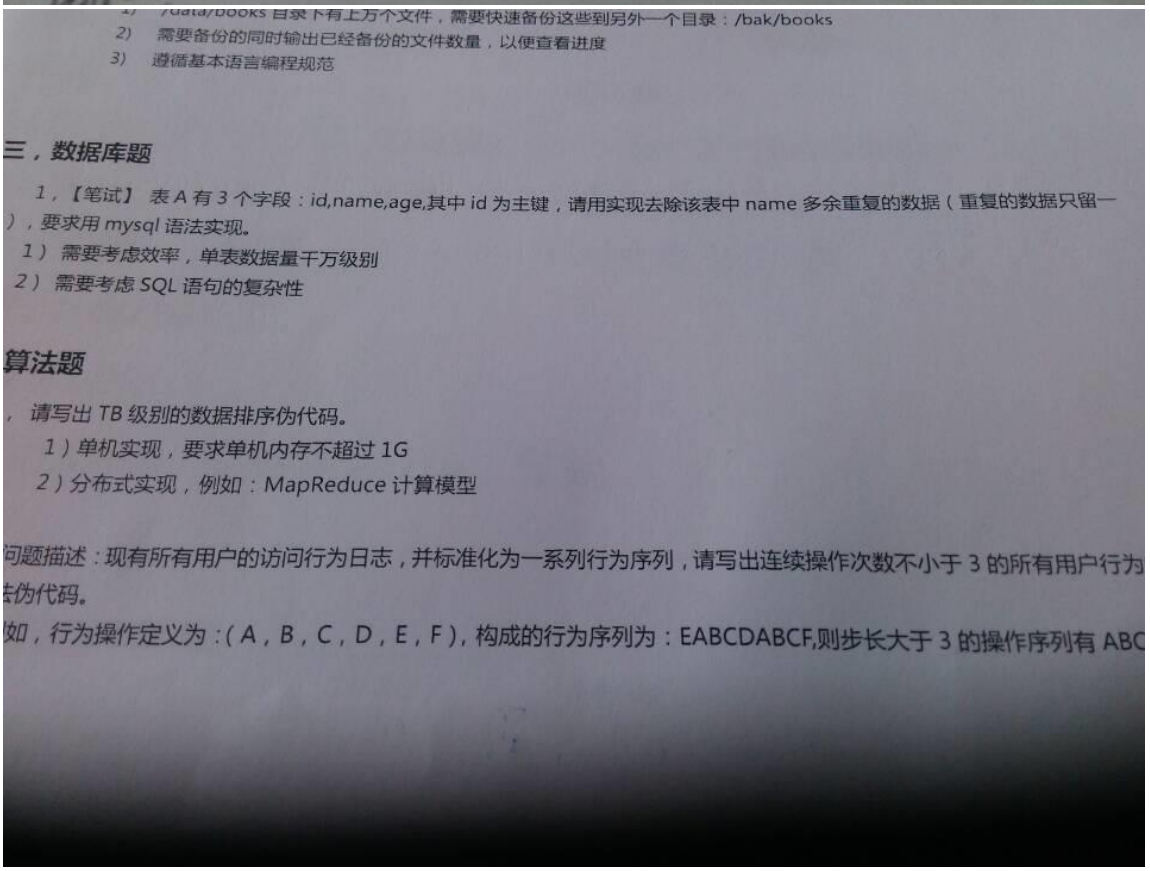


hadoop面试题关昊雯（大数据hbase

题一：
任意使用mapreduce或hive或storm，根据日志表，求20150501当天每个用户访问页面次数前10的页面，
日志表如下，userid是用户的唯一标识，pageid是页面的唯一标识，visitdate是访问时间

userid	pageid	visitdate
aa	222	20150501
aa	333	20150501
aa	222	20150501
aa	222	20150502
aa	333	20150501
bb	333	20150502
cc		

题二：
hbase的major compact和minor compact的区别



2) 需要备份的同时输出已经备份的文件数量，以便查看进度
3) 遵循基本语言编程规范

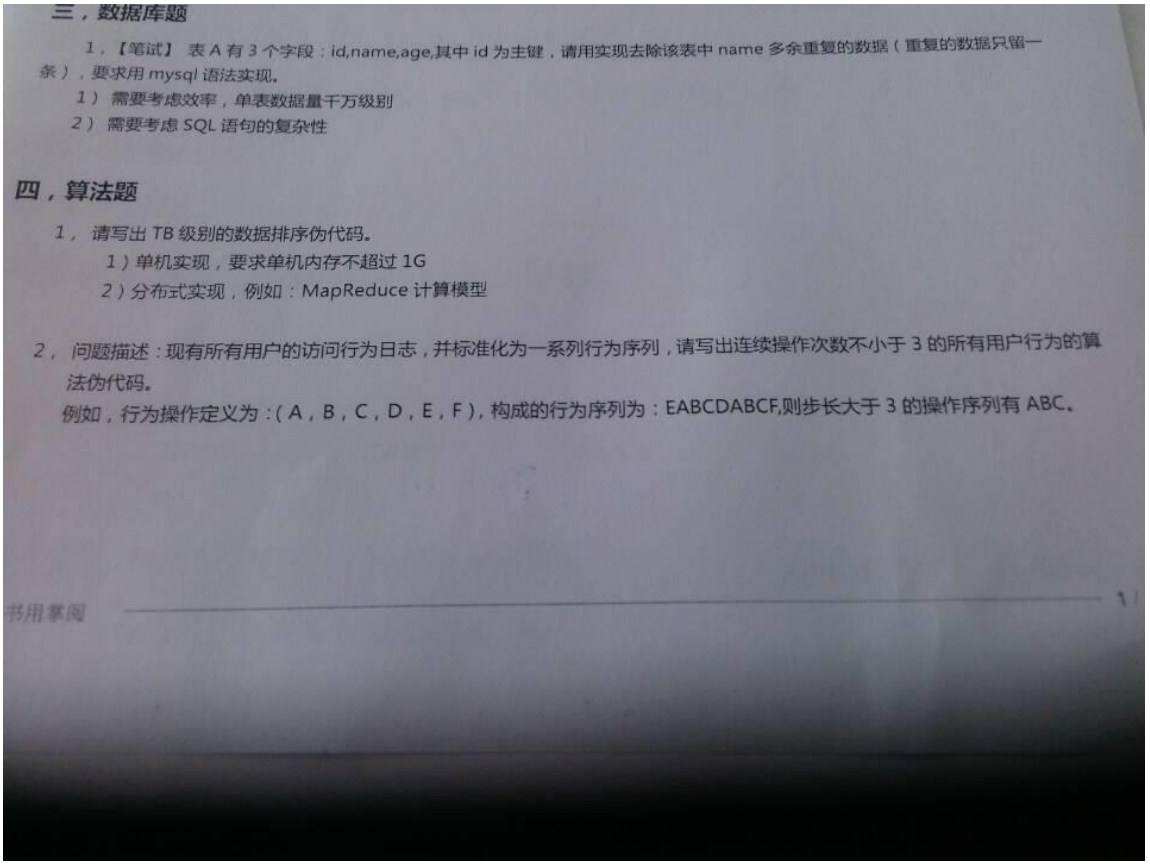
三，数据库题

1，【笔试】 表 A 有 3 个字段：id,name,age,其中 id 为主键，请用实现去除该表中 name 多余重复的数据（重复的数据只留一），要求用 mysql 语法实现。
1）需要考虑效率，单表数据量千万级别
2）需要考虑 SQL 语句的复杂性

算法题

，请写出 TB 级别的数据排序伪代码。
1）单机实现，要求单机内存不超过 1G
2）分布式实现，例如：MapReduce 计算模型

问题描述：现有所有用户的访问行为日志，并标准化为一系列行为序列，请写出连续操作次数不小于 3 的所有用户行为
法伪代码。
如，行为操作定义为：(A，B，C，D，E，F)，构成的行为序列为：EABCDABCF,则步长大于 3 的操作序列有 ABC



三，数据库题

1，【笔试】 表 A 有 3 个字段：id,name,age,其中 id 为主键，请用实现去除该表中 name 多余重复的数据（重复的数据只留一条），要求用 mysql 语法实现。

1）需要考虑效率，单表数据量千万级别

2）需要考虑 SQL 语句的复杂性

四，算法题

1，请写出 TB 级别的数据排序伪代码。

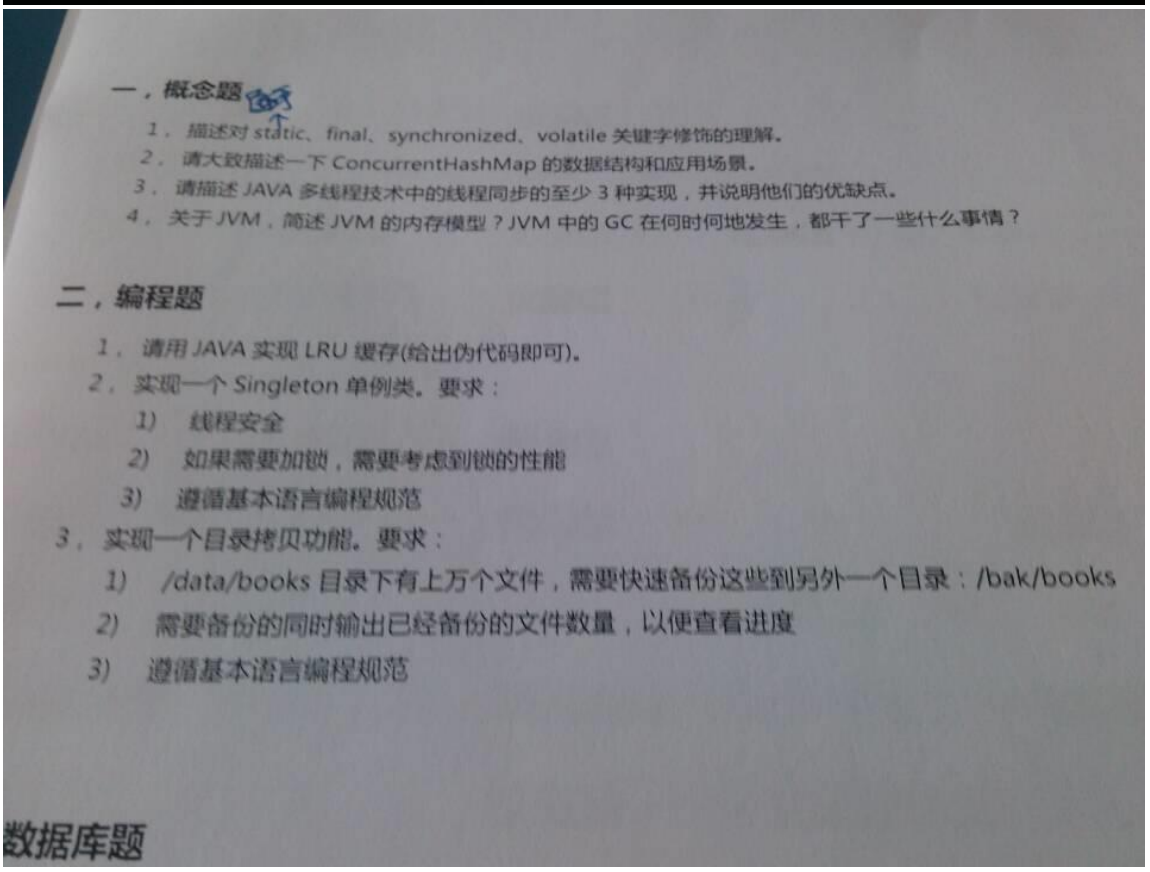
1）单机实现，要求单机内存不超过 1G

2）分布式实现，例如：MapReduce 计算模型

2，问题描述：现有所有用户的访问行为日志，并标准化为一系列行为序列，请写出连续操作次数不小于 3 的所有用户行为的算法伪代码。

例如，行为操作定义为：(A，B，C，D，E，F)，构成的行为序列为：EABCDABCF,则步长大于 3 的操作序列有 ABC。

书用掌阅



一，概念题

1，描述对 static、final、synchronized、volatile 关键字修饰的理解。

2，请大致描述一下 ConcurrentHashMap 的数据结构和应用场景。

3，请描述 JAVA 多线程技术中的线程同步的至少 3 种实现，并说明他们的优缺点。

4，关于 JVM，简述 JVM 的内存模型？JVM 中的 GC 在何时何地发生，都干了一些什么事情？

二，编程题

1，请用 JAVA 实现 LRU 缓存(给出伪代码即可)。

2，实现一个 Singleton 单例类。要求：

1) 线程安全

2) 如果需要加锁，需要考虑到锁的性能

3) 遵循基本语言编程规范

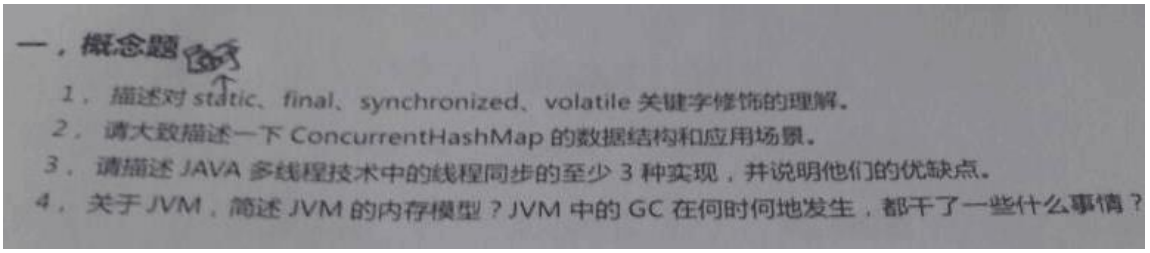
3，实现一个目录拷贝功能。要求：

1) /data/books 目录下有上万个文件，需要快速备份这些到另外一个目录：/bak/books

2) 需要备份的同时输出已经备份的文件数量，以便查看进度

3) 遵循基本语言编程规范

数据库题



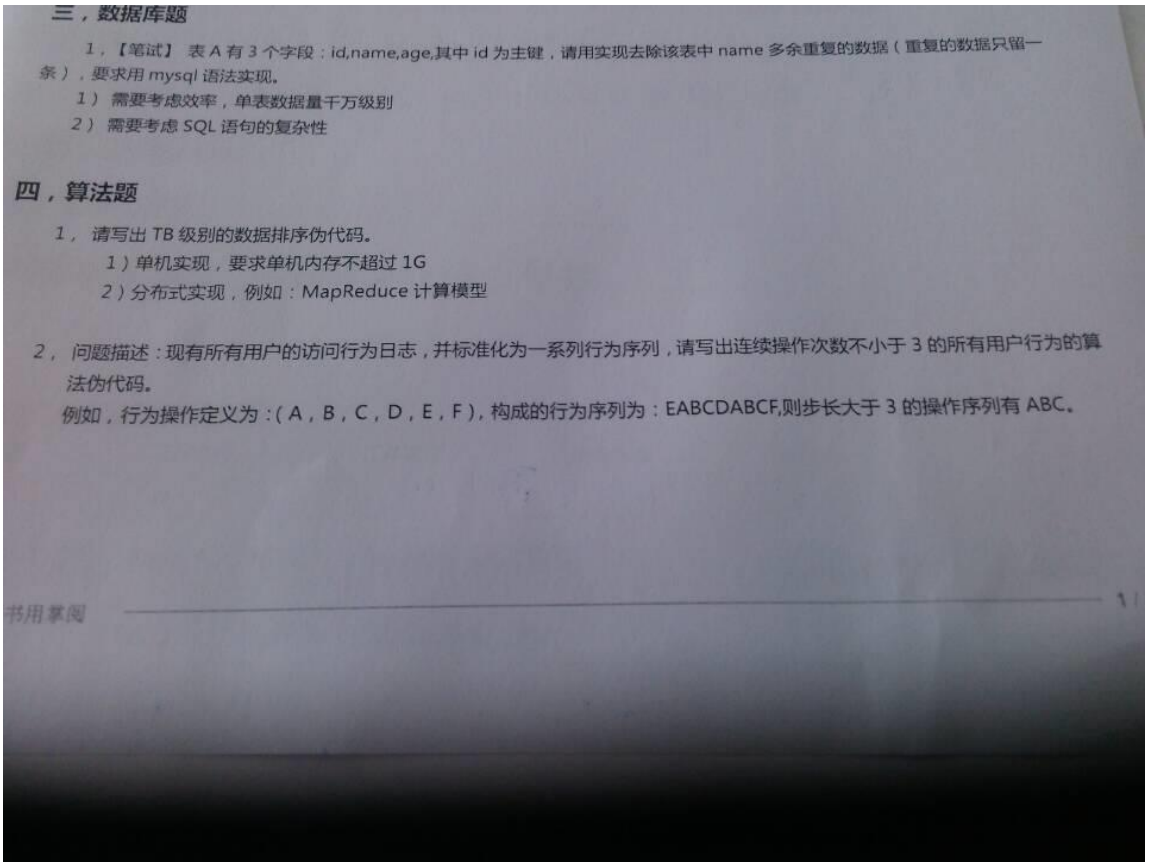
一，概念题

1，描述对 static、final、synchronized、volatile 关键字修饰的理解。

2，请大致描述一下 ConcurrentHashMap 的数据结构和应用场景。

3，请描述 JAVA 多线程技术中的线程同步的至少 3 种实现，并说明他们的优缺点。

4，关于 JVM，简述 JVM 的内存模型？JVM 中的 GC 在何时何地发生，都干了一些什么事情？



三，数据库题

1，【笔试】 表 A 有 3 个字段：id,name,age,其中 id 为主键，请用实现去除该表中 name 多余重复的数据（重复的数据只留一条），要求用 mysql 语法实现。

1）需要考虑效率，单表数据量千万级别

2）需要考虑 SQL 语句的复杂性

四，算法题

1，请写出 TB 级别的数据排序伪代码。

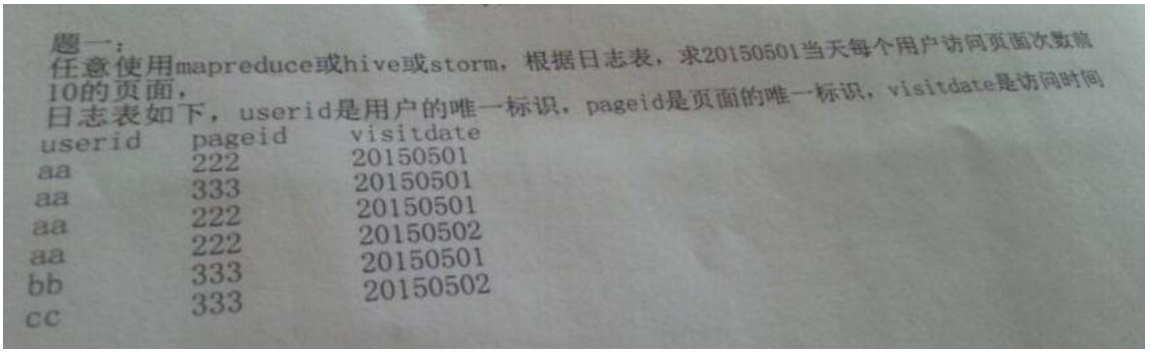
1）单机实现，要求单机内存不超过 1G

2）分布式实现，例如：MapReduce 计算模型

2，问题描述：现有所有用户的访问行为日志，并标准化为一系列行为序列，请写出连续操作次数不小于 3 的所有用户行为的算法伪代码。

例如，行为操作定义为：（A，B，C，D，E，F），构成的行为序列为：EABCDABCF,则步长大于 3 的操作序列有 ABC。

书用掌阅 1

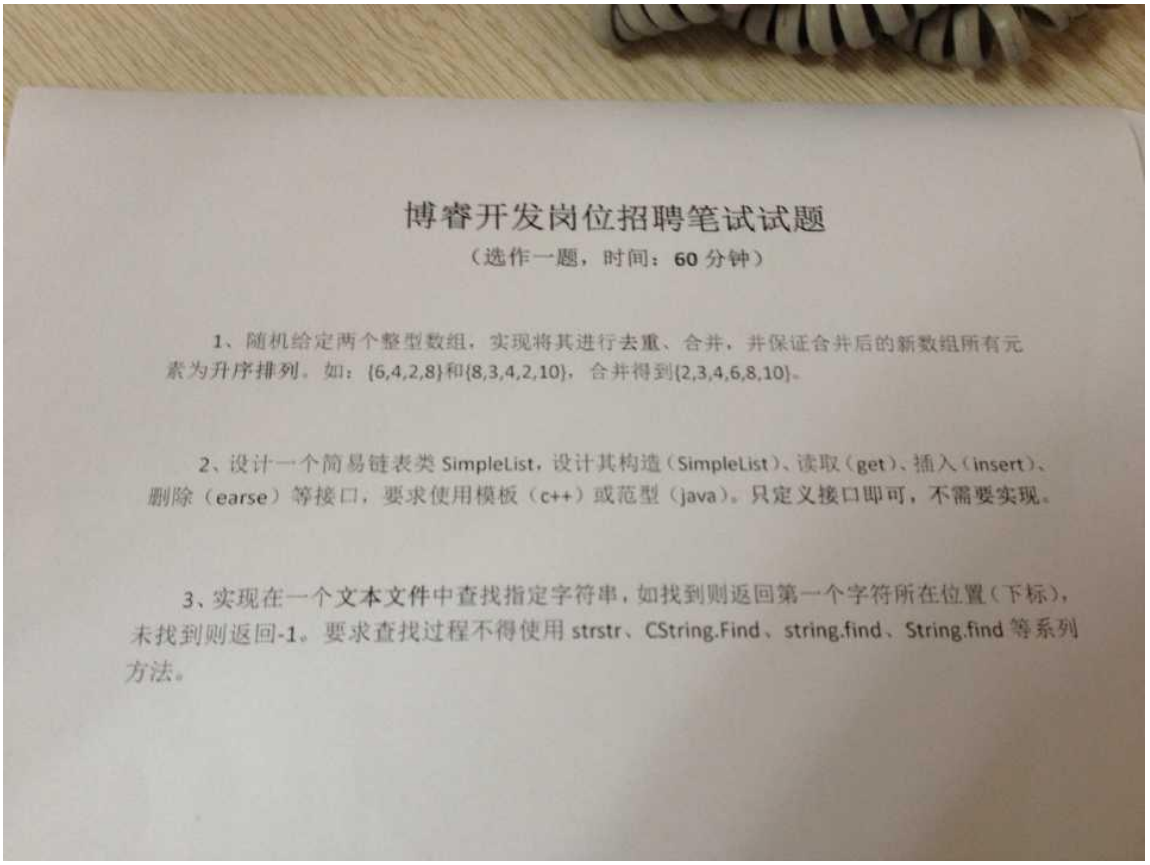


题一：

任意使用mapreduce或hive或storm，根据日志表，求20150501当天每个用户访问页面次数前10的页面，

日志表如下，userid是用户的唯一标识，pageid是页面的唯一标识，visitdate是访问时间

userid	pageid	visitdate
aa	222	20150501
aa	333	20150501
aa	222	20150501
aa	222	20150502
bb	333	20150501
cc	333	20150502



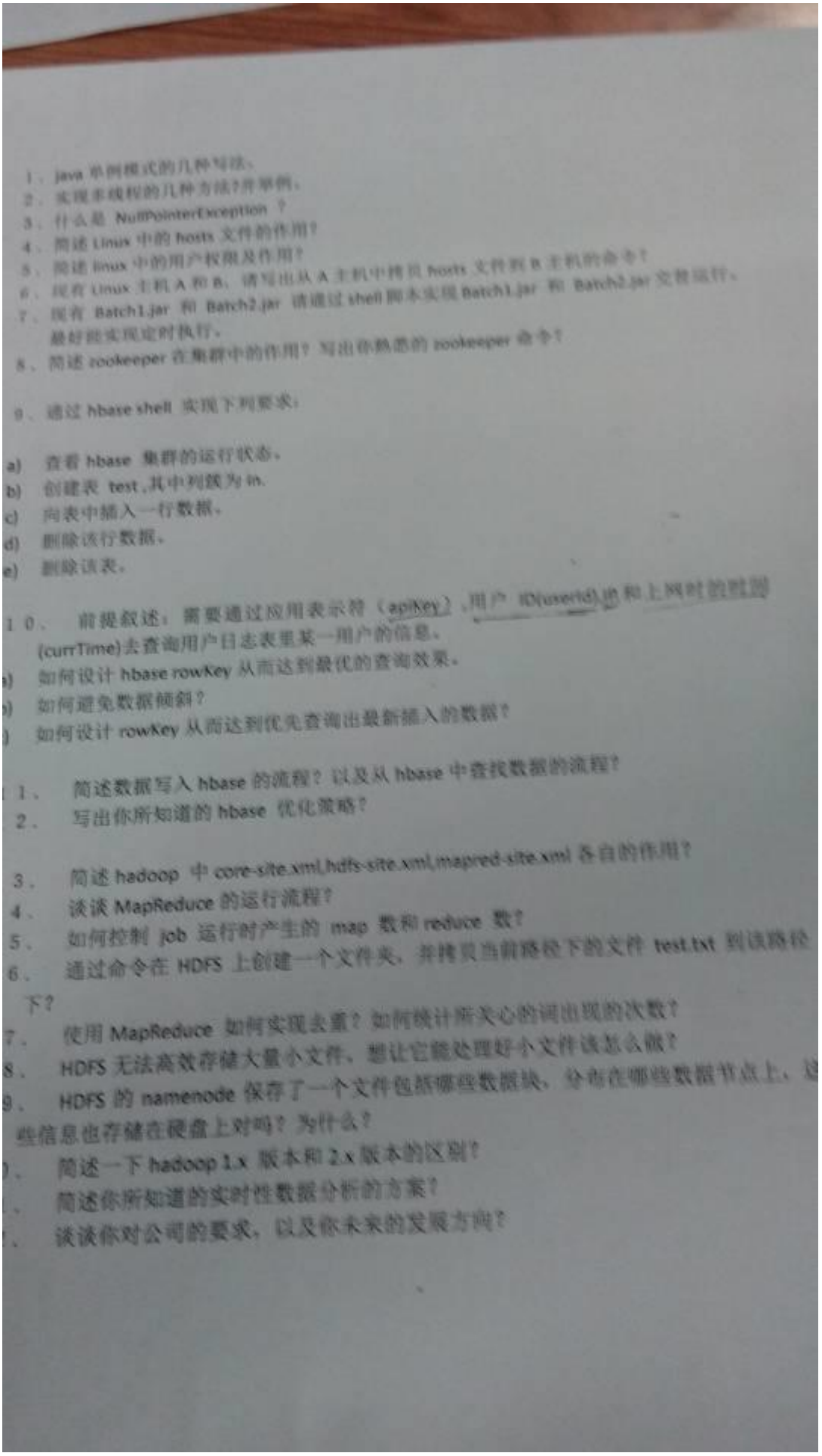
博睿开发岗位招聘笔试试题

（选作一题，时间：60 分钟）

1、随机给定两个整型数组，实现将其进行去重、合并，并保证合并后的新数组所有元素为升序排列。如：{6,4,2,8}和{8,3,4,2,10}，合并得到{2,3,4,6,8,10}。

2、设计一个简易链表类 SimpleList，设计其构造（SimpleList）、读取（get）、插入（insert）、删除（earse）等接口，要求使用模板（c++）或范型（java）。只定义接口即可，不需要实现。

3、实现在一个文本文件中查找指定字符串，如找到则返回第一个字符所在位置（下标），未找到则返回-1。要求查找过程不得使用 strstr、CString.Find、string.find、String.find 等系列方法。



1、java 单例模式的几种写法。
2、实现多线程的几种方法？并举例。
3、什么是 NullPointerException ？
4、简述 Linux 中的 hosts 文件的作用？
5、简述 linux 中的用户权限及作用？
6、现有 Linux 主机 A 和 B，请写出从 A 主机中拷贝 hosts 文件到 B 主机的命令？
7、现有 Batch1.jar 和 Batch2.jar 请通过 shell 脚本实现 Batch1.jar 和 Batch2.jar 交替运行。
最好能实现定时执行。
8、简述 zookeeper 在集群中的作用？写出你熟悉的 zookeeper 命令？

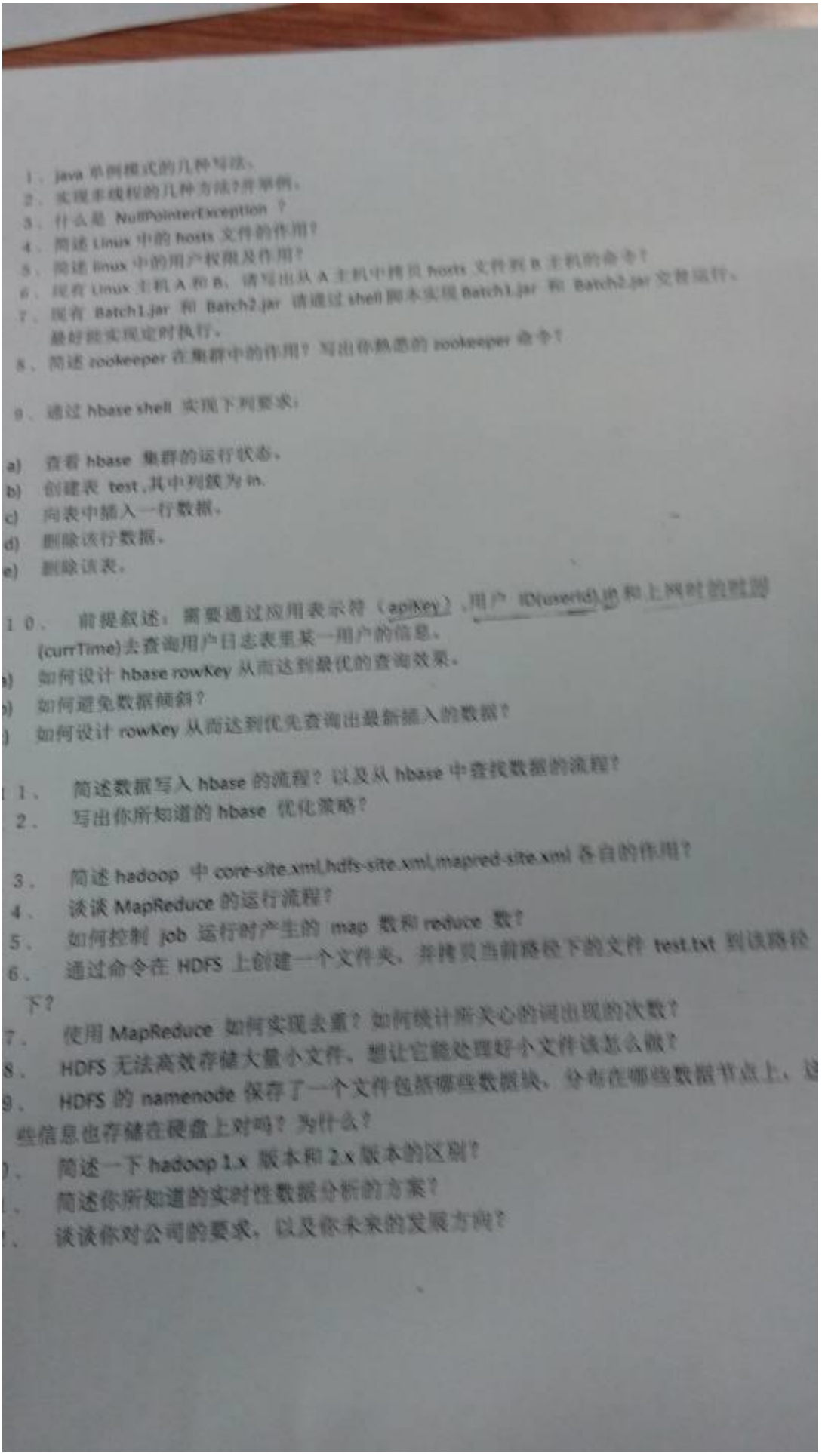
9、通过 hbase shell 实现下列要求：

a) 查看 hbase 集群的运行状态。
b) 创建表 test，其中列簇为 in.
c) 向表中插入一行数据。
d) 删除该行数据。
e) 删除该表。

10、前提叙述：需要通过应用表示符（apiKey）、用户 ID(userId)以及上网时的时间(currTime)去查询用户日志表里某一用户的信息。
a) 如何设计 hbase rowKey 从而达到最优的查询效果。
b) 如何避免数据倾斜？
c) 如何设计 rowKey 从而达到优先查询出最新插入的数据？

11、简述数据写入 hbase 的流程？以及从 hbase 中查找数据的流程？
12、写出你所知道的 hbase 优化策略？

13、简述 hadoop 中 core-site.xml,hdfs-site.xml,mapred-site.xml 各自的作用？
14、谈谈 MapReduce 的运行流程？
15、如何控制 job 运行时产生的 map 数和 reduce 数？
16、通过命令在 HDFS 上创建一个文件夹，并拷贝当前路径下的文件 test.txt 到该路径下？
17、使用 MapReduce 如何实现去重？如何统计所关心的词出现的次数？
18、HDFS 无法高效存储大量小文件，想让它能处理好小文件该怎么做？
19、HDFS 的 namenode 保存了一个文件包括哪些数据块，分布在哪些数据节点上，这些信息也存储在硬盘上对吗？为什么？
20、简述一下 hadoop 1.x 版本和 2.x 版本的区别？
21、简述你所知道的实时性数据分析的方案？
22、谈谈你对公司的要求，以及你未来的发展方向？



1、java 单例模式的几种写法。
2、实现多线程的几种方法？并举例。
3、什么是 NullPointerException ？
4、简述 Linux 中的 hosts 文件的作用？
5、简述 linux 中的用户权限及作用？
6、现有 Linux 主机 A 和 B，请写出从 A 主机中拷贝 hosts 文件到 B 主机的命令？
7、现有 Batch1.jar 和 Batch2.jar 请通过 shell 脚本实现 Batch1.jar 和 Batch2.jar 交替运行。
最好能实现定时执行。
8、简述 zookeeper 在集群中的作用？写出你熟悉的 zookeeper 命令？

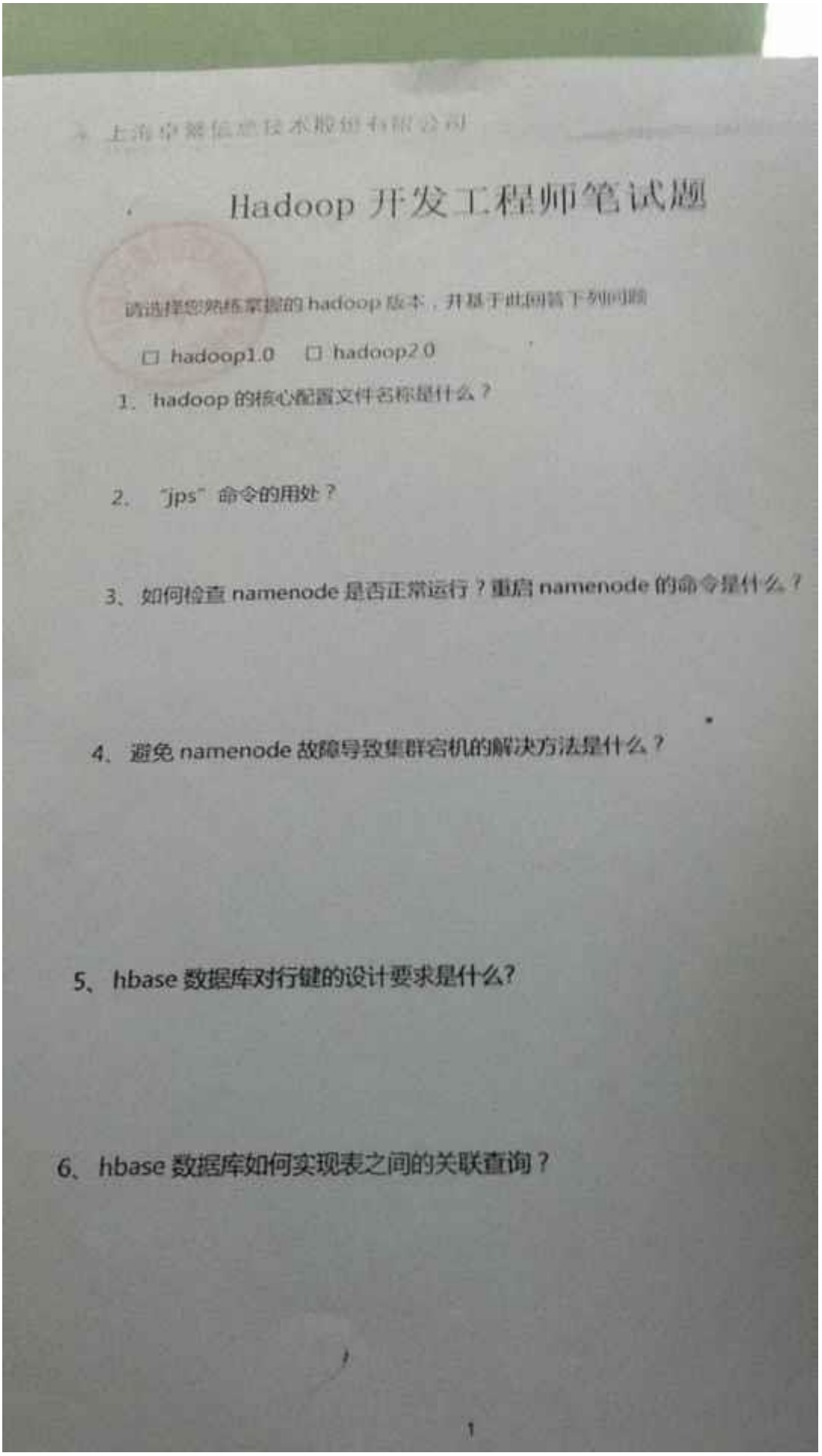
9、通过 hbase shell 实现下列要求：

a) 查看 hbase 集群的运行状态。
b) 创建表 test，其中列簇为 in.
c) 向表中插入一行数据。
d) 删除该行数据。
e) 删除该表。

10、前提叙述：需要通过应用表示符（apiKey）、用户 ID(userId)以及上网时的时间(currTime)去查询用户日志表里某一用户的信息。
a) 如何设计 hbase rowKey 从而达到最优的查询效果。
b) 如何避免数据倾斜？
c) 如何设计 rowKey 从而达到优先查询出最新插入的数据？

11、简述数据写入 hbase 的流程？以及从 hbase 中查找数据的流程？
12、写出你所知道的 hbase 优化策略？

13、简述 hadoop 中 core-site.xml,hdfs-site.xml,mapred-site.xml 各自的作用？
14、谈谈 MapReduce 的运行流程？
15、如何控制 job 运行时产生的 map 数和 reduce 数？
16、通过命令在 HDFS 上创建一个文件夹，并拷贝当前路径下的文件 test.txt 到该路径下？
17、使用 MapReduce 如何实现去重？如何统计所关心的词出现的次数？
18、HDFS 无法高效存储大量小文件，想让它能处理好小文件该怎么做？
19、HDFS 的 namenode 保存了一个文件包括哪些数据块，分布在哪些数据节点上，这些信息也存储在硬盘上对吗？为什么？
20、简述一下 hadoop 1.x 版本和 2.x 版本的区别？
21、简述你所知道的实时性数据分析的方案？
22、谈谈你对公司的要求，以及你未来的发展方向？



上海中兴信息技术股份有限公司

Hadoop 开发工程师笔试题

请选择您熟练掌握的 hadoop 版本，并基于此回答下列问题

☐ hadoop1.0 ☐ hadoop2.0

1、hadoop 的核心配置文件名称是什么？

2、"jps" 命令的用处？

3、如何检查 namenode 是否正常运行？重启 namenode 的命令是什么？

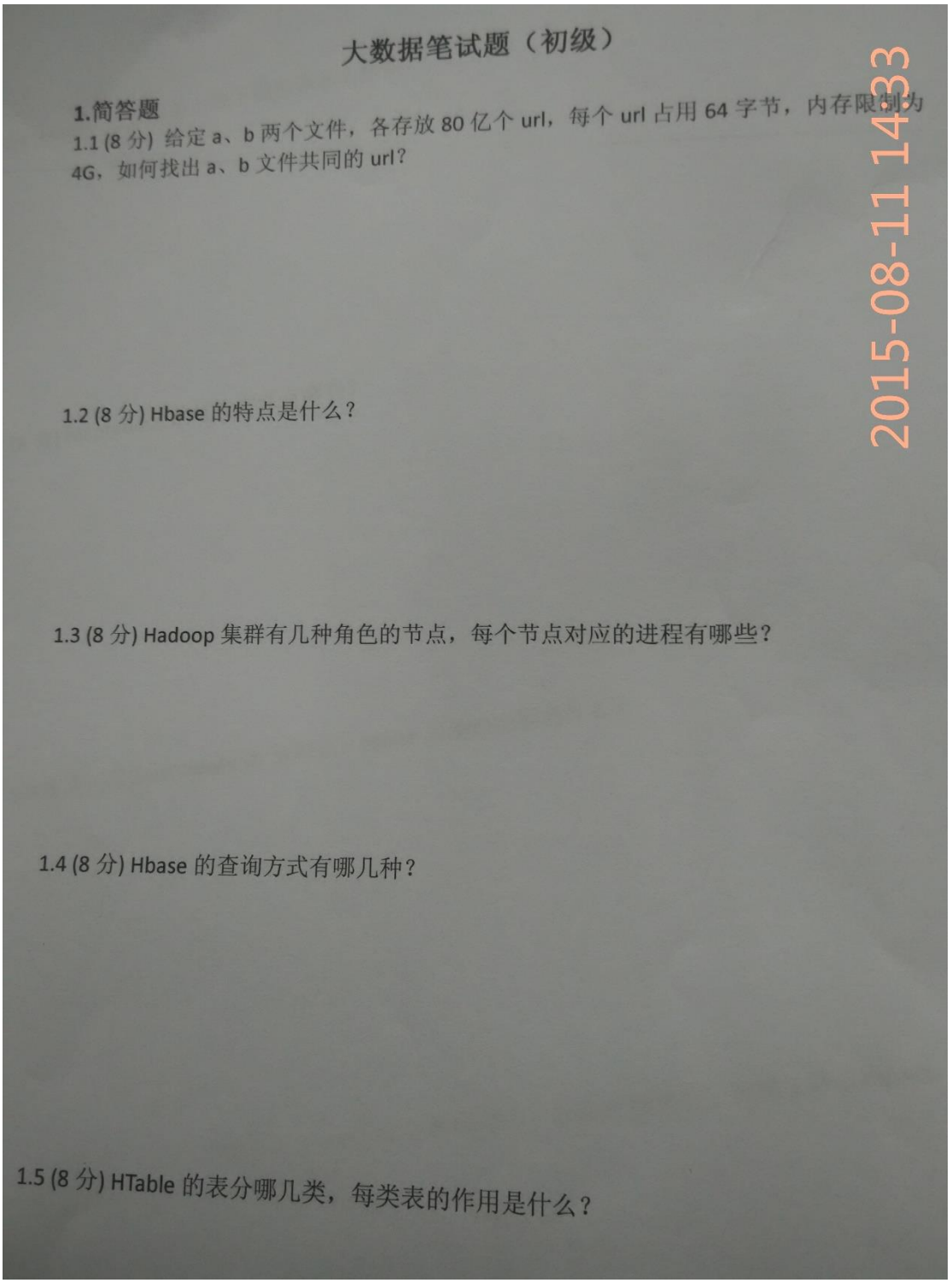
4、避免 namenode 故障导致集群宕机的解决方法是什么？

5、hbase 数据库对行键的设计要求是什么？

6、hbase 数据库如何实现表之间的关联查询？

1

(十六) 11 道笔试题



大数据笔试题（初级）

1.简答题

1.1 (8 分) 给定 a、b 两个文件，各存放 80 亿个 url，每个 url 占用 64 字节，内存限制为 4G，如何找出 a、b 文件共同的 url？

1.2 (8 分) Hbase 的特点是什么？

1.3 (8 分) Hadoop 集群有几种角色的节点，每个节点对应的进程有哪些？

1.4 (8 分) Hbase 的查询方式有哪几种？

1.5 (8 分) HTable 的表分哪几类，每类表的作用是什么？

2015-08-11 14:33

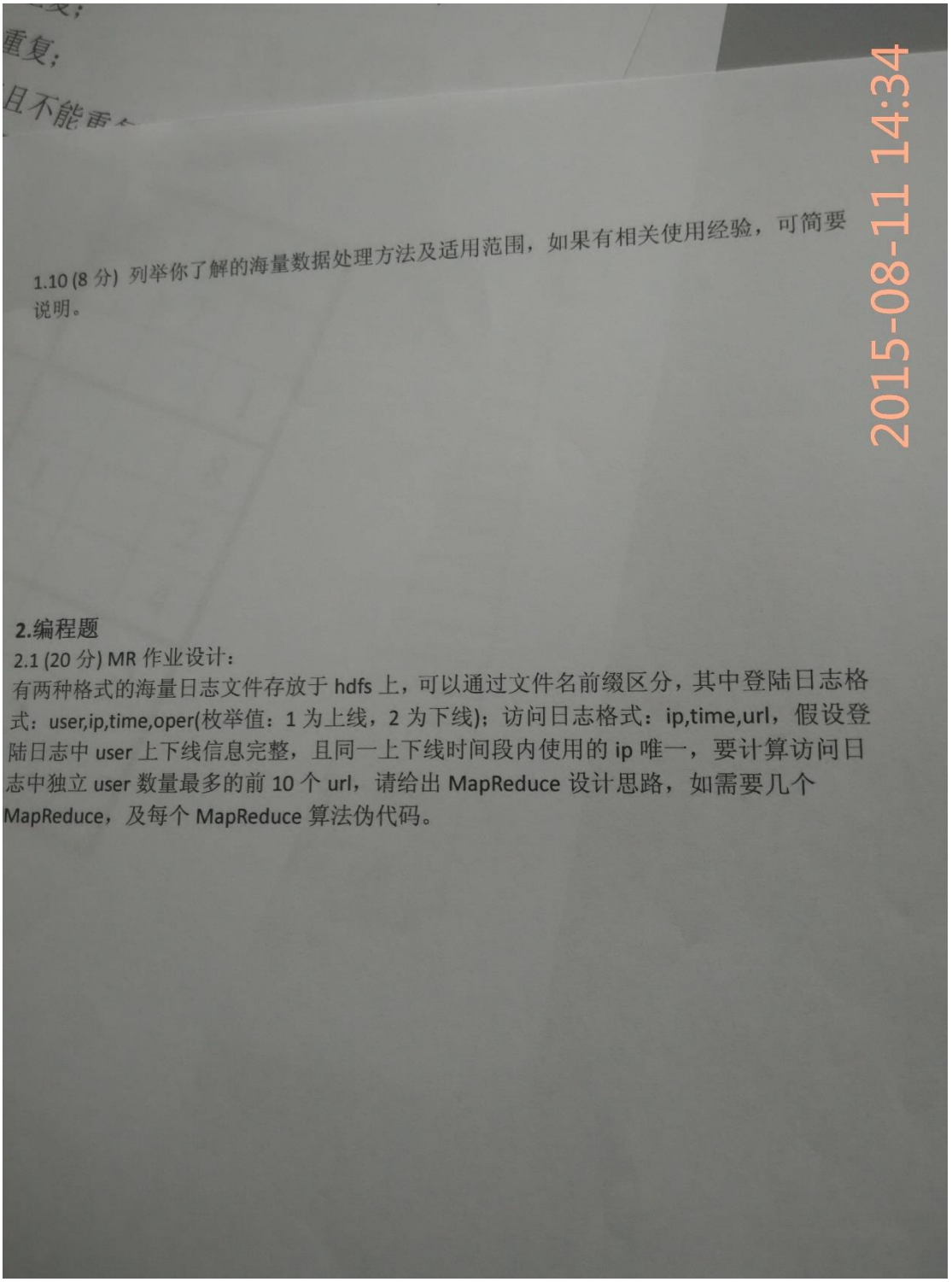
1.6 (8 分) Hadoop 节点动态上线下线应该怎么操作？

1.7 (8 分) Mapreduce 的优化方法有哪些？

1.8 (8 分) 以 start-hbase.sh 为起点，Hbase 启动的流程是什么？

1.9 (8 分) 请简述 HBASE 中 compact 用途是什么，什么时候触发，分哪 2 种 compact，有何区别，有哪些相关配置参数？

2015-08-11 14:34


重复;
且不能重
1.10 (8 分) 列举你了解的海量数据处理方法及适用范围，如果有相关使用经验，可简要说明。
2015-08-11 14:34
2.编程题
2.1 (20 分) MR 作业设计：
有两种格式的海量日志文件存放于 hdfs 上，可以通过文件名前缀区分，其中登陆日志格式：user,ip,time,oper(枚举值：1 为上线，2 为下线)；访问日志格式：ip,time,url，假设登陆日志中 user 上下线信息完整，且同一上下线时间段内使用的 ip 唯一，要计算访问日志中独立 user 数量最多的前 10 个 url，请给出 MapReduce 设计思路，如需要几个 MapReduce，及每个 MapReduce 算法伪代码。

(十七) 24 道题目

1. 请实现二分查找(C++或者 JAVA 实现)

2. 如何实现一个数组反转(C++或者 JAVA 实现)

3. 有 2 个 JAVA 线程，用什么方式实现保证一个个线程等待另外一个线程执行完成

4. HashTable HashMap CurrentHashMap 使用区别

5. 请大概描述下 JVM 的 GC 机制，常用的 JVM 调优方法，OOM 如何产生，如何处理 OOM 问题

6. Hadoop 是由哪几个组件组成并阐述 HDFS 写入数据的实现机制

7. 如何决定一个 Job 的 Map 和 Reduce 的数量

8. Hadoop 的 SEQUENCEFILE 的格式，并说明下什么是 JAVA 序列化，如何实现 JAVA 序列化

9. 简述 Hadoop1 与 Hadoop2 的架构异同

10. YARN 的新特性

11. 如何使用 MapReduce 实现 2 个表的 JOIN

13. 请描述 MapReduce 二次排序原理

14. 请描述 MapReduce 中排序发生在几个阶段

15. 请描述 MapReduce 中 shuffle 阶段的工作流程，如何优化 shuffle 阶段

16. 请描述 MapReduce 中 Combiner 的作用是什么，一般应用场景，哪些场景不适用

17. 列举在 Linux 系统下可以查看系统各项性能的工具(区分 CPU、硬盘、网络等)

18. Hive 表关联查询，如何解决数据倾斜问题

19. HBase 和 Hive 有什么区别?

20. 简单描述 HBase 的 rowkey 的设计原则?

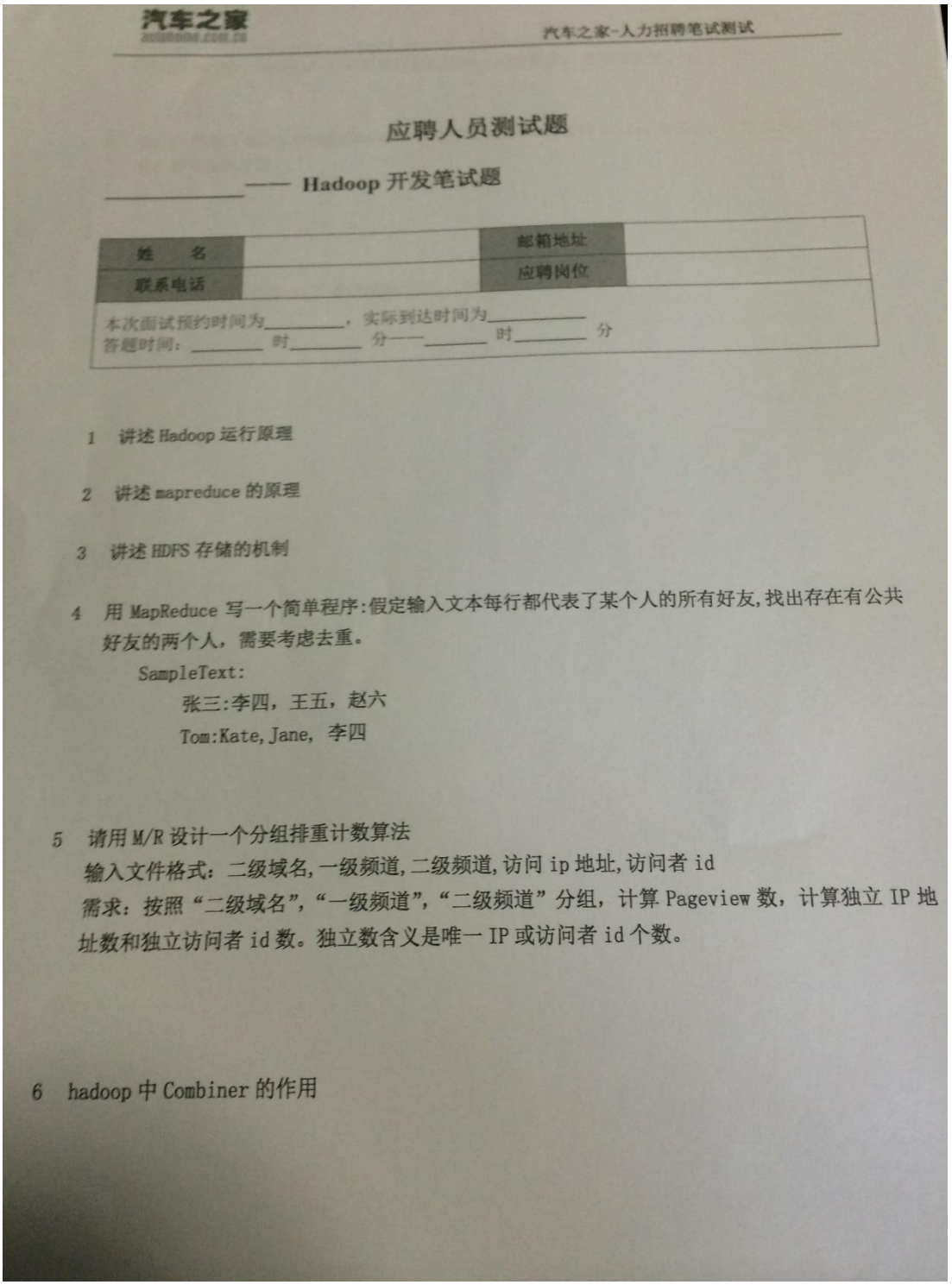
21. 请描述 HBase 中 scan 和 get 的功能以及实现的异同

22. 请描述 HBase 中 scan 对象的 setCache 和 setBatch 方法的使用

23. 请详细描述 HBase 中一个 Cell 的结构

24. 请描述如何处理 HBase 中 region 太多和 region 太大带来的冲突

(十八) 30 道题目



汽车之家

汽车之家-人力招聘笔试测试

应聘人员测试题

_____—— Hadoop 开发笔试题

姓　名		邮箱地址	
联系电话		应聘岗位	

本次面试预约时间为_____，实际到达时间为_____
答题时间：_____ 时_____ 分——_____ 时_____ 分

1　讲述 Hadoop 运行原理

2　讲述 mapreduce 的原理

3　讲述 HDFS 存储的机制

4　用 MapReduce 写一个简单程序:假定输入文本每行都代表了某个人的所有好友,找出存在有公共好友的两个人，需要考虑去重。

SampleText:

张三:李四，王五，赵六

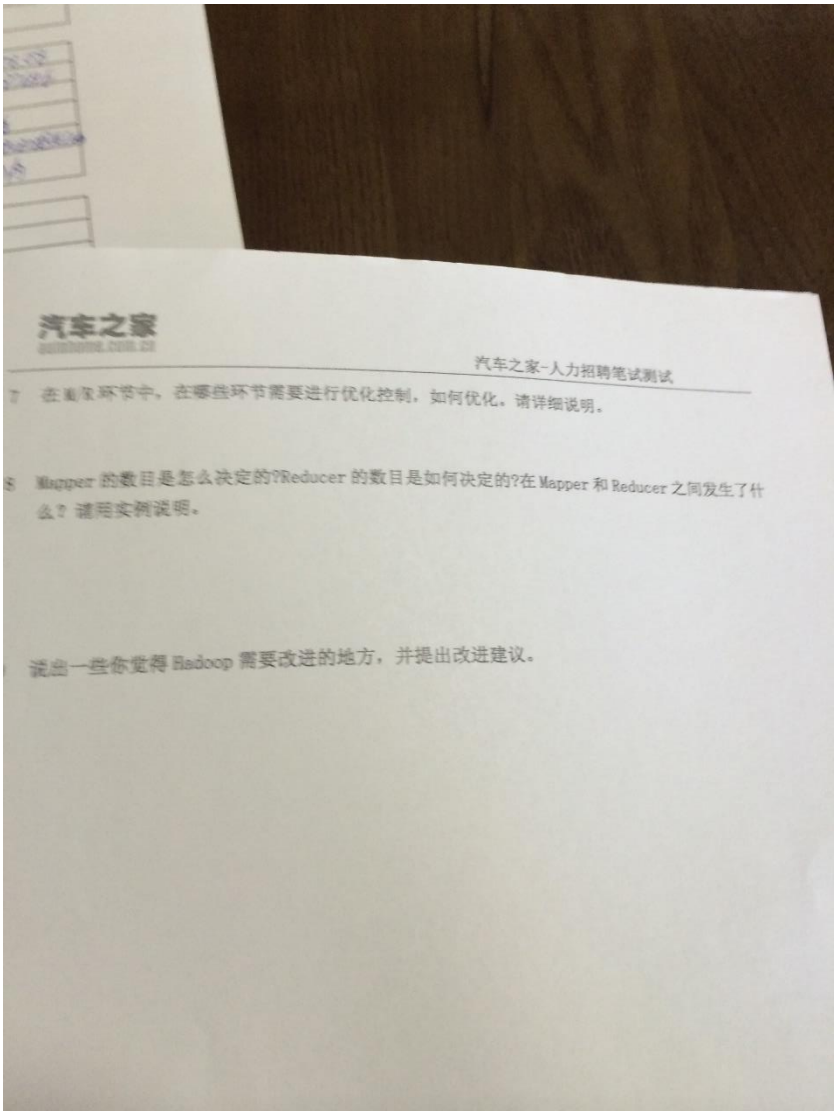
Tom:Kate,Jane, 李四

5　请用 M/R 设计一个分组排重计数算法

输入文件格式：二级域名,一级频道,二级频道,访问 ip 地址,访问者 id

需求：按照“二级域名”,“一级频道”,“二级频道”分组，计算 Pageview 数，计算独立 IP 地址数和独立访问者 id 数。独立数含义是唯一 IP 或访问者 id 个数。

6　hadoop 中 Combiner 的作用



汽车之家

汽车之家-人力招聘笔试测试

在reduce环节中，在哪些环节需要进行优化控制，如何优化。请详细说明。

Mapper的数目是怎么决定的?Reducer的数目是如何决定的?在Mapper和Reducer之间发生了什么？请用实例说明。

说出一些你觉得Hadoop需要改进的地方，并提出改进建议。

数据库题目

----------数据表说明：

- **表 1：orders 订单表**
  - 位置：orders 库，dbo 用户下
  - 表结构和记录

order_id	customer_id	order_date	City_id	Shipment_id
1001	388	2008-4-12	4	1
1002	421	2008-4-12	1	1
1007	456	2008-4-17	2	2
1032	17	2008-6-3	1	3
1177	388	2008-6-24	15	1
1222	388	2008-7-1	1	1
1234	17	2008-7-13	1	3
1344	456	2008-7-14	3	1

- **表 2：order_detail 订单明细表**
  - 位置：orders 库，dbo 用户下
  - 表结构和记录

order_id	product_id	num	price
1001	23999	1	23
1002	12000	2	7.5
1002	98001	1	45
1007	12000	15	7.5
1032	11111	3	12
1177	23999	10	23
1222	33887	1	78.2
1234	94302	2	15
1234	78665	1	3.5
1234	99008	1	17
1344	12000	3	7.5

- **表 3：city 城市字典表**
  - 位置：customer 库，wmsdata 用户下
  - 表结构和记录

City_id	City_name
1	北京
2	上海
3	天津

4	河北
15	内蒙
16	广东
17	陕西
33	广西

- **表 3：shipment 送货方式字典表**
  - 位置：customer 库，wmsdata 用户下
  - 表结构和记录

Shipment_id	Shipment_name
1	上门送货
2	邮寄
3	其他

--------**题目：**请写出完整的 SQL 查询语句

1、 按区域查找 2008 年 5 月-7 月期间订单总额占区域订单总额 20%以上的客户、顾客订单金额、区域总金额、顾客订单总金额与区域总金额占比

2、 按区域查找 2008 年 5 月-7 月期间订单总额排名前三的客户

3、 按区域查找 2008 年 5 月-7 月期间订单总额排名前 1/4 的客户

4、按照如下形式计算出 2008 年 5 月-7 月期间各城市选择上门送货、邮寄及其他送货方式的订单数

城市名称	选择上门送货订单数	选择邮寄订单数	其他送货方式订单数
北京			
上海			
天津			
河北			
内蒙			
广东			
陕西			
广西			

数据建模题目

以上述数据库题目中的表为基础，设计一下下面要求的数据模型，画出星型模型图并写出模型中涉及的表和字段；

业务需求：业务人员每天要看不同地域、不同送货方式下产生的订单数和订单金额；

UNIX 题目

1. 用 UNIX 命令找出一文档 abc.txt 的行数

2. 在 UNIX 下用什么命令更改档案/目录的权限？

3. 用一 UNIX 命令找出所有 Java 程序代码档内含有"println"的行

4. 怎样列出 UNIX 系统下某一用户正在执行的进程？

5.写出 Linux 下级联删除目录/data 的命令；

6.已知两台服务器的域名分别是 host1 和 host2，host1 的用户为 h1usr，host2 的用户为 h2usr，host1 上存在文件/home/h1usr/data.txt，现在需要将这个 data.txt 文件传输到 host2 的/home/h2usr/下，请写出传输的 rsync 命令；

Python/perl

1. 多线程与多进程的区别，请写一个多线程使用的例子

2、Python 中如何在一个函数内修改一个全局变量？

3、Python 中 pass 的作用是什么？

4、将一个元素是数值的 list，按从大到小进行排序，请用 python/perl 实现。

大数据处理

事业部转化分析:

需求：计算各事业部(bd_id)的 pv/uv。

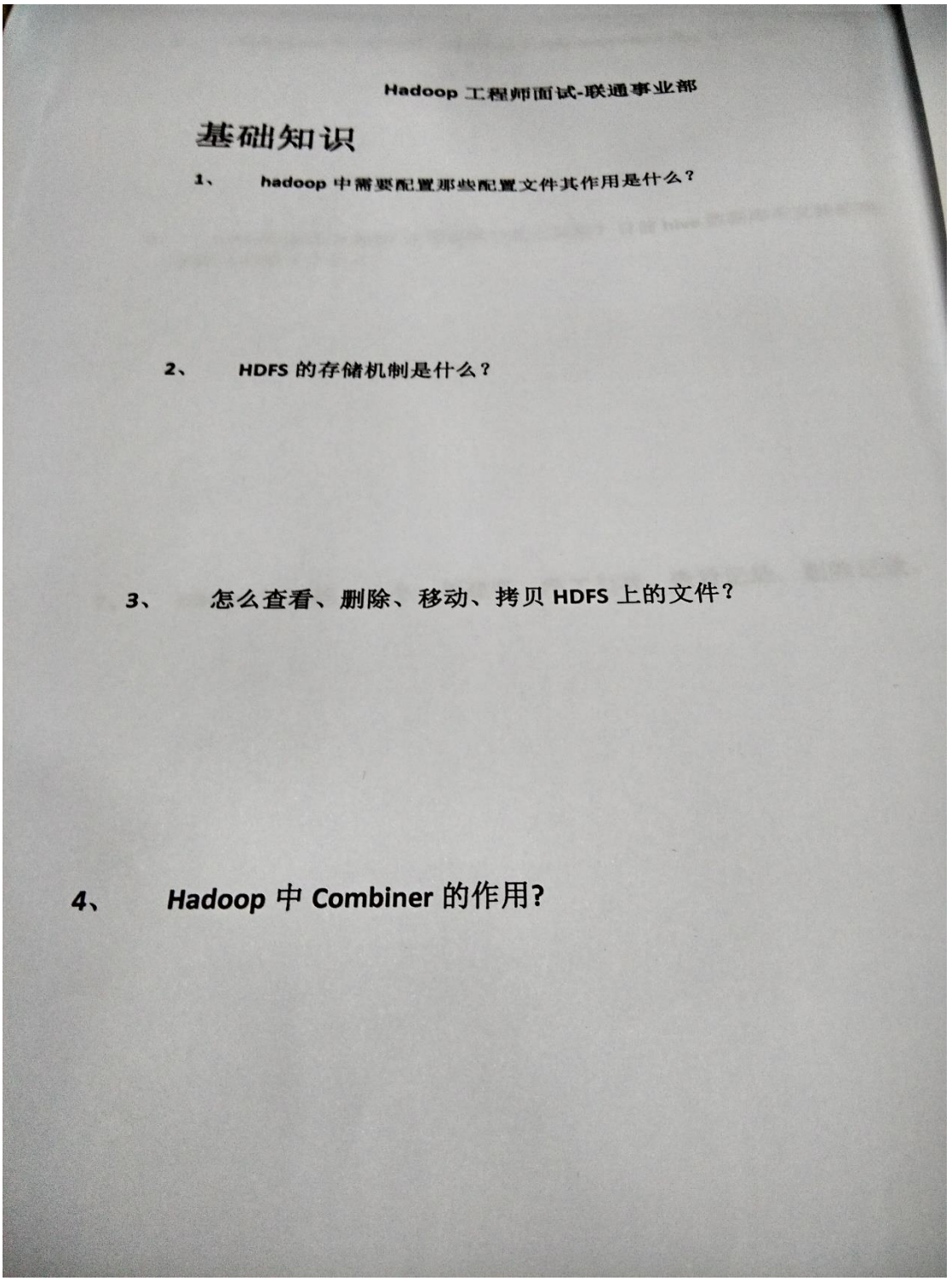
注：(pv：url 中 product_id 出现的次数，uv：permanent_id 去重个数)

现在有两张表：一张流量表，属于事实表，包含 url、permanent_id 这两个字段。

：一张单品表，属于维度表，包含所有 product_id 和 bd_id 对应。

url 示例：http://product.dangdang.com/1195857607.html

(十九) 8 道题目



Hadoop 工程师面试-联通事业部

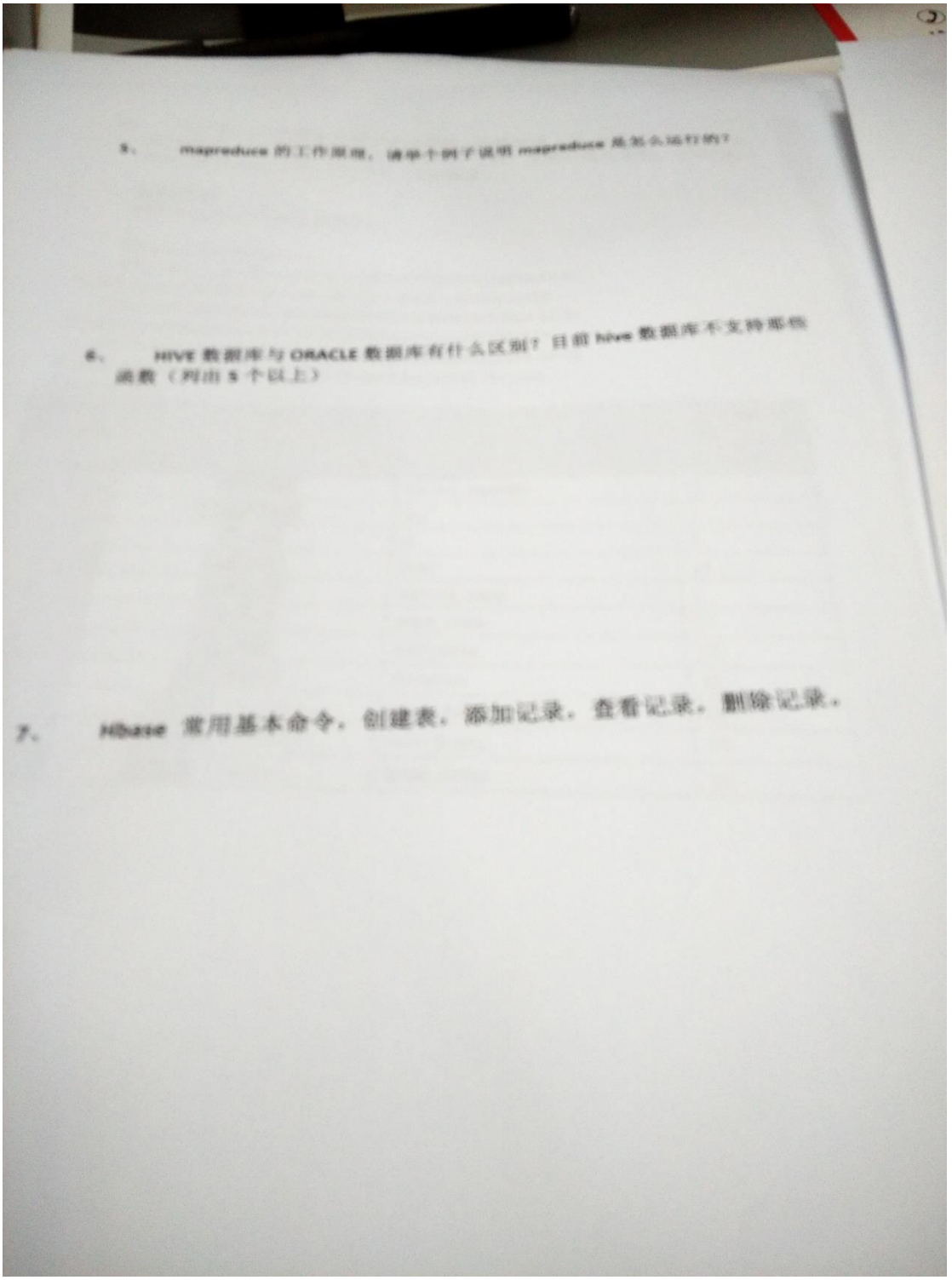
基础知识

1、 hadoop 中需要配置那些配置文件其作用是什么？

2、 HDFS 的存储机制是什么？

3、 怎么查看、删除、移动、拷贝 HDFS 上的文件？

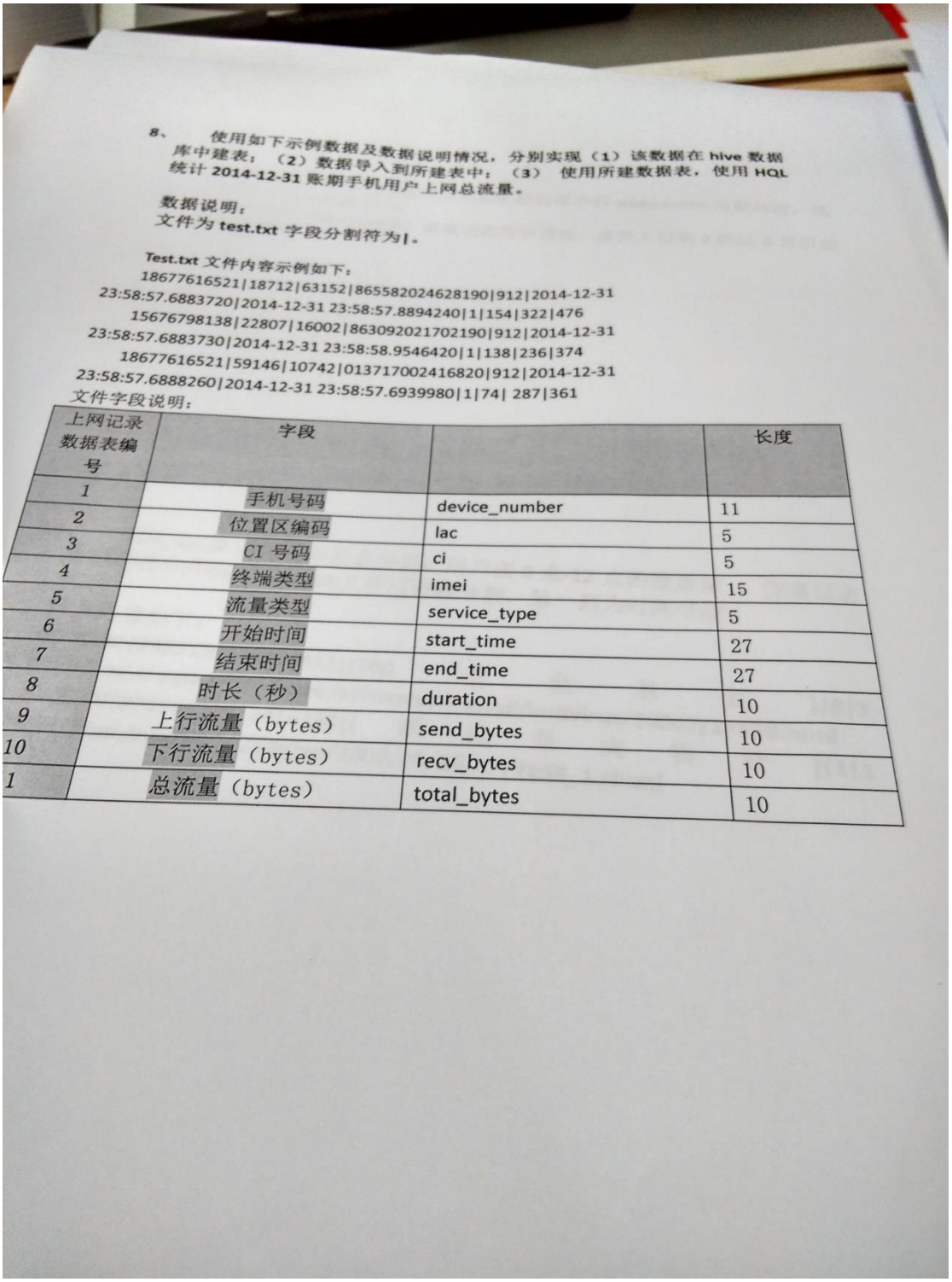
4、 Hadoop 中 Combiner 的作用？



5、 mapreduce 的工作原理，请举个例子说明 mapreduce 是怎么运行的？

6、 HIVE 数据库与 ORACLE 数据库有什么区别？目前 hive 数据库不支持那些函数（列出 5 个以上）

7、 Hbase 常用基本命令，创建表，添加记录，查看记录，删除记录。



8、 使用如下示例数据及数据说明情况，分别实现（1）该数据在 hive 数据库中建表；（2）数据导入到所建表中；（3） 使用所建数据表，使用 HQL 统计 2014-12-31 账期手机用户上网总流量。

数据说明：
文件为 test.txt 字段分割符为|。

Test.txt 文件内容示例如下：

18677616521|18712|63152|865582024628190|912|2014-12-31 23:58:57.6883720|2014-12-31 23:58:57.8894240|1|154|322|476
15676798138|22807|16002|863092021702190|912|2014-12-31 23:58:57.6883730|2014-12-31 23:58:58.9546420|1|138|236|374
18677616521|59146|10742|013717002416820|912|2014-12-31 23:58:57.6888260|2014-12-31 23:58:57.6939980|1|74| 287|361

文件字段说明：

上网记录数据表编号	字段		长度
1	手机号码	device_number	11
2	位置区编码	lac	5
3	CI 号码	ci	5
4	终端类型	imei	15
5	流量类型	service_type	5
6	开始时间	start_time	27
7	结束时间	end_time	27
8	时长（秒）	duration	10
9	上行流量（bytes）	send_bytes	10
10	下行流量（bytes）	recv_bytes	10
1	总流量（bytes）	total_bytes	10

(二十) 面试题 14 道：

1.简要描述如何安装配置一个 apache 开源版 hadoop，描述即可，列出步骤更好

第一题：1.创建 hadoop
户。

2.setup.改 IP。

3.安装 java，并修改/etc/profile 文件，配置 java 的环境变量。

4.修改 Host 文件域名。

5.安装 SSH，配置无密钥通信。

6.解压 hadoop。

7.配置 conf 文件下 hadoop-env.sh、core-site.sh、mapre-site.sh、hdfs-site.sh。

8.配置 hadoop 的环境变量。

9.Hadoop namenode -format

10.Start-all

2.请列出正常工作的 hadoop 集群中 hadoop 都需要启动哪些进程，他们的作用分别是什么？

第二题：namenode：管理集群，并记录 datanode 文件
信息。 Secondname:可以做冷备，对一定范围内数
据做快照性备份。 Datanode:存储数据

Jobtracker :管理任务，并将任务分配给 tasktracker。

Tasktracker:任务执行方。

3.启动 hadoop 报如下错误，该如何解决？

error org.apache.hadoop.hdfs.server.namenode.NameNode

org.apache.hadoop.hdfs.server.common.inconsistentFSStateExceptio

n Directory /tmp/hadoop-root/dfs/name is in an inconsistent

state storage direction does not exist or is not accessible?

第三题：可能的原因：1.hdfs 没有启动成功，通过查看 jps 确认下。
2.确认文件是否存
在。

4.请写出以下执行命令

1）杀死一个 job?

2)删除 hdfs 上的/tmp/aaa 目录

3 加入一个新的存储节点和删除一个计

算节点需要刷新集群状态命令？

第四题：hadoop job -list 拿到 job-id ,hadoop job -kill job-id

Hadoop fs -rmr /tmp/aaa

加新节点时：

Hadoop-daemon.sh start datanode

Hadoop-daemon.sh start tasktracker

删除时：

Hadoop mradmin -refreshnodes

Hadoop dfsadmin –refreshnodes

5.请列出你所知道的 hadoop 调度器，并简要说明其工作方法？

第五题：

Fifo schedular :默认，先进先出的原则

Capacity schedular :计算能力调度器，选择占用最小、优先级高的先执行，依此类推。

Fair schedular:公平调度，所有的 job 具有相同

的资源。

6.请列出在你以前工作中所使用过的开发

mapreduce 的语言？

第六题：java、python、hive

7.当前日志采样格式为

```
1. a,b,c,d
2. b,b,f,e
3. a,a,c,f
```

复制代码

请用你最熟悉的语言编写一个 mapreduce，并计
算第四列每个元素出现的个数

第七题：wordcount。第八
题：就

8.你认为用 Java，
Streaming,pipe 方式开发
mapreduce,各有哪些优缺
点？

第八题：用过 java 和
hiveQL。

Java 写 mapreduce 可以实现复杂的逻辑，如果需求简单，则显得繁琐。

HiveQL 基本都是针对 hive 中的表数据进行编写，但对复杂的逻辑很难进行实现。写 起来简单。

9.hive 有哪些方式保存元数据，各有哪些特点？

第九题：三种：内存数据库 derby，挺小，不
常用 。 本地 mysql。常用
远程端 mysql。不常
用

上网上找了下专业名称：single user mode..multi user mode...remote user mode

10.请简述 hadoop 怎么样实现二级排序？

第十题：在源码中有个例子。不过我没看。

11.简述 hadoop 实现 join 的几种方法？

第十一题：貌似好几种来着，像 mapjoin ..reducejon..还有其它的来着吧。可以去网上查 一下，我常用的就是 mapjoin，可以将小表的数据加载到内存中使用，然后匹配的大表的数据，加快效率。

12.请用 Java 实现非递归二分查找？

第十二题：用 java。我的第一思路就是排序后从中间查询呗，for 循环的事。

13.请简述 mapreduce 中，combiner，partition 作用？

第十三题：

combiner :实现的功能跟 reduce 差不多，接收 map 的值，经过计算后给 reduce，它 的 key,value 类型要跟 reduce 完全一样，当 reduce 业务复杂时可以用，不过它貌似只是 操作本机的数据。

Partition：将输出的结果分别保存在不同的文件中。。

14.某个目录下有两个文件 a.txt 和 b.txt,文件格式为（ip，username），

列如：

a.txt

127.0.0.1 zhangsan

127.0.0.1 wangxiaoer

127.0.0.2 lisi

127.0.0.3 wangwu

b.txt

127.0.0.4 lixiaolu

127.0.0.1 lisi

每个文件至少 100 万行，请使用 Linux 命令完成如下工作：

1）每个文件各自的 ip 数

2）出现在 b.txt 而没有出现在 a.txt 的 ip

3）每个 user 出现的次数以及每个 user 对应的 ip 数

第十四题：不会 shell

(二十一) 面试题 12 道：

笔试题：

（重点面试题）

1、Hive 内部表和外部表的区别？

2、Hbase 的 rowkey 怎么创建比较好？列族怎么创建比较好？

3、用 mapreduce 怎么处理数据倾斜问题？

4、hadoop 框架中怎么来优化？

5、Hbase 内部是什么机制？

6、我们在开发分布式计算 job 的,是否可以去掉 reduce()阶段？

7、hdfs 的数据压缩算法

8、mapreduce 的调度模式

9、hive 底层与数据库交互原理

10、hbase 过滤器实现原则

11、reduce 后输出的数据量有多大？

12、现场出问题测试 mapreduce 掌握情况和 HIve 的 Hql 语句掌握情况？

(二十二) 面试题 9 道：

1 、datanode 在什么情况下不会备份？

2 、combine 出现在那个过程？

3 、hdfs 得体系结构？

4 、flush 的过程？

5 、什么是队列？

6 、List 与 Set 的区别？

7、 数据库的三大范式？

8 、三个 datanode 当有一个 datanode 出现错误会怎样？

9、 sqoop 在导入数据到 mysql 中，如何让数据不重复导入？如果存在数据问题 sqoop 如何处理？

(二十三)　面试题 7 道：

1、使用 Hive 或者自定义 MR 实现如下逻辑

product_no lac_id moment start_time user_id county_id staytime city_id

13429100031 22554 8 2013-03-11 08:55:19.151754088 571 571 282 571

13429100082 22540 8 2013-03-11 08:58:20.152622488 571 571 270 571

13429100082 22691 8 2013-03-11 08:56:37.149593624 571 571 103 571

13429100087 22705 8 2013-03-11 08:56:51.139539816 571 571 220 571

13429100087 22540 8 2013-03-11 08:55:45.150276800 571 571 66 571

13429100082 22540 8 2013-03-11 08:55:38.140225200 571 571 133 571

13429100140 26642 9 2013-03-11 09:02:19.151754088 571 571 18 571

13429100082 22691 8 2013-03-11 08:57:32.151754088 571 571 287 571

13429100189 22558 8 2013-03-11 08:56:24.139539816 571 571 48 571

13429100349 22503 8 2013-03-11 08:54:30.152622440 571 571 211 571

字段解释：

product_no：用户手机号；

lac_id：用户所在基站；

start_time：用户在此基站的开始时间；

staytime：用户在此基站的逗留时间。

需求描述：

根据 lac_id 和 start_time 知道用户当时的位置，根据 staytime 知道用户各个基站的逗留时

长。根据轨迹合并连续基站的 staytime。

最终得到每一个用户按时间排序在每一个基站驻留时长

期望输出举例：

13429100082 22540 8 2013-03-11 08:58:20.152622488 571 571 270 571

13429100082 22691 8 2013-03-11 08:56:37.149593624 571 571 390 571

本能力考察

2.1 请随意使用各种类型的脚本语言实现： 批量将指定目录下的所有文件中的

$HADOOP_HOME$替换成/home/ocetl/app/hadoop

2.2 假设有 10 台主机，H1 到 H10，在开启 SSH 互信的情况下，编写一个或多个脚本实现

在所有的远程主机上执行脚本的功能

例如：runRemoteCmd.sh "ls -l"

期望结果：

H1:

XXXXXXXX

XXXXXXXX

XXXXXXXX

H2:

XXXXXXXX

XXXXXXXX

XXXXXXXX

H3:

…

3 Hadoop 基础知识与问题分析的能力

3.1 描述一下 hadoop 中，有哪些地方使用了缓存机制，作用分别是什么

3.2 请描述 https://issues.apache.org/jira/browse/HDFS-2379 说的是什么问题，最终解

决的思路是什么？

4、MapReduce 开发能力

请参照 wordcount 实现一个自己的 map reduce，需求为：

a 输入文件格式：

xxx,xxx,xxx,xxx,xxx,xxx,xxx

b 输出文件格式：

xxx,20

xxx,30

xxx.40

c 功能：根据命令行参数统计输入文件中指定关键字出现的次数，并展示出来

例如：hadoop jar xxxxx.jar keywordcount xxx,xxx,xxx,xxx(四个关键字）

5、MapReduce 优化

请根据第五题中的程序, 提出如何优化 MR 程序运行速度的思路

6、Linux 操作系统知识考察

请列举曾经修改过的/etc 下的配置文件，并说明修改要解决的问题？

7、Java 开发能力

7.1 写代码实现 1G 大小的文本文件，行分隔符为\x01\x02,统计一下该文件中的总行数，

要求注意边界情况的处理

7.2 请描述一下在开发中如何对上面的程序进行性能分析，对性能进行优化的过程。

(二十四)　hadoop 面试题 21 道：

1、设计一套系统，使之能够从不断增加的不同的数据源中，提取指定格式的数据。

要求：

1)、运行结果要能大致得知提取效果，并可据此持续改进提取方法；

2)、由于数据来源的差异性，请给出可弹性配置的程序框架；

3)、数据来源可能有 Mysql,sqlserver 等；

4)、该系统具备持续挖掘的能力，即，可重复提取更多信息

2. 经典的一道题：

现有 1 亿个整数均匀分布，如果要得到前 1K 个最大的数，求最优的算法。

（先不考虑内存的限制，也不考虑读写外存，时间复杂度最少的算法即为最优算法） 我先说下我的想法:分块，比如分 1W 块，每块 1W 个，然后分别找出每块最大值，从这最 大的 1W 个值中找最大 1K 个，
那么其他的 9K 个最大值所在的块即可扔掉，从剩下的最大的 1K 个值所在的块中找前 1K

个即可。那么原问题的规模就缩小到了
1/10。 问题：
（1）这种分块方法的最优时间复杂度。

（2）如何分块达到最优。比如也可分 10W 块，每块 1000 个数。则问题规模可降到原来 1/100。但事实上复杂度并没降低。

(3)还有没更好更优的方法解决这个问
题。

3. MapReduce 大致流程？

4. combiner, partition 作用？

5.用 mapreduce 实现 sql 语句 select count(x) from a group by b？

6. 用 mapreduce 如何实现两张表连接，有哪些方法？

7.知道 MapReduce 大致流程，map, shuffle, reduce

8.知道 combiner, partition 作用，设置 compression

9.搭建 hadoop 集群，master/slave 都运行那些服务
10.HDFS，replica 如何定位
11.版本 0.20.2->0.20.203->0.20.205, 0.21, 0.23, 1.0. 1
新旧 API 有什么不同
12.Hadoop 参数调优，cluster level: JVM, map/reduce slots, job level: reducer
#,memory, use combiner? use compression?
13.pig latin, Hive 语法有什么不同
14.描述 HBase, zookeeper 搭建过程
15.hadoop 运行的原理?
16.mapreduce 的原理?
17.HDFS 存储的机制?
18.举一个简单的例子说明 mapreduce 是怎么来运行的 ？

19.使用 mapreduce 来实现下面实例
实例:现在有 10 个文件夹,每个文件夹都有 1000000 个 url. 现在让你找出 top1000000url。

20.hadoop 中 Combiner 的作用?

21.如何确认 Hadoop 集群的健康状况。

(二十五) hadoop 面试题 9 道：

1.使用的 hadoop 版本都是什么？

2.mpareduce 原理是什么？

3.mapreduce 作业，不使用 reduce 来输出，用什么能代替 reduce 的功能

4.hive 如何调优？

5.hive 如何权限控制？

6.hbase 写数据的原理是什么？

7.hive 能像关系数据库那样，建多个库吗？

8.hbase 宕机如何处理？

9.假设公司要建一个数据中心，你会如何规划？

七、hadoop 选择判断题 33 道：

下面哪个程序负责 HDFS 数据存储。

a)NameNode b)Jobtracker c)Datanode d)secondaryNameNode e)tasktracker

2. HDfS 中的 block 默认保存几份？

a)3 份 b)2 份 c)1 份 d)不确定

3. 下列哪个程序通常与 NameNode 在一个节点启动？

a)SecondaryNameNode b)DataNode c)TaskTracker d)Jobtracker

4. Hadoop 作者

a)Martin Fowler b)Kent Beck c)Doug cutting

5. HDFS 默认 Block Size

a)32MB b)64MB c)128MB

6. 下列哪项通常是集群的最主要瓶颈哪项是正确的？

a)它是 NameNode 的热备 b)它对内存没有要求

c)它的目的是帮助 NameNode 合并编辑日志，减少 NameNode 启动时间

d)SecondaryNameNode 应与 NameNode 部署到一个节点

多选题：

8. 下列哪项可以作为集群的管理工具

a)Puppet b)Pdsh c)Cloudera Manager d)d)Zookeeper

9. 配置机架感知的下面哪项正确

a)如果一个机架出问题，不会影响数据读写

b)写入数据的时候会写到不同机架的 DataNode 中

c)MapReduce 会根据机架获取离自己比较近的网络数据

10. Client 端上传文件的时候下列哪项正确

a)数据经过 NameNode 传递给 DataNode

b)Client 端将文件切分为 Block，依次上传

c)Client 只上传数据到一台 DataNode，然后由 NameNode 负责 Block 复制工作

11. 下列哪个是 Hadoop 运行的模式

a)单机版 b)伪分布式 c)分布式

12. Cloudera 提供哪几种安装 CDH 的方法

a)Cloudera manager b)Tar ball c)Yum d)Rpm

可以监控 Hadoop 集群，因为它不提供 Hadoop 支持。（ ）

16. 如果 NameNode 意外终止，SecondaryNameNode 会接替它使集群继续工作。（ ）

17. Cloudera CDH 是需要付费使用的。（ ）

18. Hadoop 是 Java 开发的，所以 MapReduce 只支持 Java 语言编写。（ ）

19. Hadoop 支持数据的随机读写。（ ）

20. NameNode 负责管理 metadata，client 端每次读写请求，它都会从磁盘中

读取或则会写入 metadata 信息并反馈 client 端。（ ）

21. NameNode 本地磁盘保存了 Block 的位置信息。（ ）

22. DataNode 通过长连接与 NameNode 保持通信。（ ）

23. Hadoop 自身具有严格的权限管理和安全措施保障集群正常运行。（ ）

24. Slave 节点要存储数据，所以它的磁盘越大越好。（ ）

25. hadoop dfsadmin –report 命令用于检测 HDFS 损坏块。（ ）

26. Hadoop 默认调度器策略为 FIFO（ ）

27. 集群内每个节点都应该配 RAID，这样避免单磁盘损坏，影响整个节点运行。（ ）

28. 因为 HDFS 有多个副本，所以 NameNode 是不存在单点问题的。（ ）

29. 每个 map 槽就是一个线程。（ ）

30. Mapreduce 的 input split 就是一个 block。（ ）

31. NameNode 的 Web UI 端口是 50030，它通过 jetty 启动的 Web 服务。（ ）

32. Hadoop 环境变量中的 HADOOP_HEAPSIZE 用于设置所有 Hadoop 守护线程的内

存。它默认是 200 GB。（ ）

33. DataNode 首次加入 cluster 的时候，如果 log 中报告不兼容文件版本，那需要

NameNode 执行“Hadoop namenode -format”操作格式化磁盘。（ ）

(二十六) mr 和 hive 实现手机流量统计面试题 6 道：

1.hive 实现统计的查询语句是什么？
2.生产环境中为什么建议使用外部表？
3.hadoop mapreduce 创建类 DataWritable 的作用是什么？
4.为什么创建类 DataWritable？

5.如何实现统计手机流量？
6.对比 hive 与 mapreduce 统计手机流量的区别？

(二十七) 面试题 1 道：

最近去面试，出了个这样的题目，大家有兴趣也试试。

用 Hadoop 分析海量日志文件，每行日志记录了如下数据：

TableName(表名)，Time(时间)，User(用户)，TimeSpan(时间开销)。

要求：

编写 MapReduce 程序算出高峰时间段（如上午 10 点）哪张表被访问的最频繁，以及

这段时间访问这张表最多的用户，以及这个用户的总时间开销。

(二十八) 面试题 6 道：

前段时间接到阿里巴巴面试云计算,拿出来给我们共享下
1、hadoop 运转的原理?
2、mapreduce 的原理?
3、HDFS 存储的机制?
4、举一个简略的比方阐明 mapreduce 是怎么来运转的 ?
5、面试的人给你出一些疑问,让你用 mapreduce 来完成？
比方:如今有 10 个文件夹,每个文件夹都有 1000000 个 url. 如今让你找出 top1000000url。
6、hadoop 中 Combiner 的效果?
论坛某网友的回复：
1.hadoop 即是 mapreduce 的进程，服务器上的一个目录节点加上多个数据节点，将程序 传递到各个节点，再节点上进行计算。
2.mapreduce 即是将数据存储到不一样的节点上，用 map 方法对应办理，在各个节点上 进行计算，最后由 reduce 进行合并。
3.java 程序和 namenode 合作，把数据存放在不一样的数据节点上

4.怎么运转用图来表明最好了。图无法画。谷歌下

5.不思考歪斜，功能，运用 2 个 job，第一个 job 直接用 filesystem 读取 10 个文件夹作为 map 输入，url 做 key，reduce 计算个 url 的 sum，下一个 job map 顶用 url 作 key，运用-sum 作二次排序，reduce 中取 top10000000

第二种方法，建 hive 表 A，挂分区 channel，每个文件夹是一个分区.
select x.url,x.c from(select url,count(1) as c from A where channel ='' group by url)x order by x.c desc limie 1000000;
6 combiner 也是一个 reduce ,它可以削减 map 到 reudce 的数据传输 ,进步 shuff 速度。

牢记平均值不要用。需求输入=map 的输出，输出=reduce 的输入。

(二十九) 笔试和面试题 11 道:

一、笔试

1、java 基础类：

1）继承：写的一段代码，让写出结果；

2）引用对象和值对象；

Java 基础类记不太清了，有很多都是基础。

2、linux 基础：

1）find 用法

2）给出一个文本：比如 http://aaa.com

http://bbb.com

http://bbb.com

http://bbb.com

http://ccc.com

http://ccc.com

让写 shell 统计，最后输出结果：aaa 1

Ccc 2

Bbb 3

要求结果还要排序

还有别的，也是比较基础的

3、数据库类：oracle 查询语句

二、面试

讲项目经验：问的很细，给纸，笔，让画公司 hadoop 的项目架构,最后还让自己说几条业

务数据，然后经过平台后，出来成什么样子。

java 方面：io 输入输出流里有哪些常用的类，还有 webService,线程相关的知识

linux：问到 jps 命令，kill 命令，问 awk,sed 是干什么用的、还有 hadoop 的一些常用命

令

hadoop：讲 hadoop1 中 map,shuffle,reduce 的过程，其中问到了 map 端和 reduce 端

溢写的细节（幸好我之前有研究过）

项目部署：问了项目是怎么部署，代码怎么管理

Hive 也问了一些，外部表，还有就是 hive 的物理模型跟传统数据库的不同

三、某互联网公司的面试：

问到分析人行为的算法：我当时想到我们做的反洗钱项目中，有用到。我就给举例：我

们是怎么筛选出可疑的洗钱行为的。

(三十) 面试题 26 道:

****信 Hadoop 面试笔试题（共 14 题，还有一题记不住了）

1、hadoop 集群搭建过程，写出步骤。

2、hadoop 集群运行过程中启动那些线程，各自的作用是什么？

3、/tmp/hadoop-root/dfs/name the path is not exists or is not accessable.

NameNode main 中报错，该怎么解决。（大意这样 一个什么异常）

4、工作中编写 mapreduce 用到的语言，编写一个 mapreduce 程序。

5、hadoop 命令

1）杀死一个 job 任务 （杀死 50030 端口的进程即可）

2）删除/tmp/aaa 文件目录

3）hadoop 集群添加或删除节点时，刷新集群状态的命令

6、日志的固定格式：

a,b,c,d

a,a,f,e

b,b,d,f

使用一种语言编写 mapreduce 任务，统计每一列最后字母的个数。

7、hadoop 的调度器有哪些，工作原理。

8、mapreduce 的 join 方法有哪些？

9、Hive 元数据保存的方法有哪些，各有什么特点？

10、java 实现非递归二分法算法。

11、mapreduce 中 Combiner 和 Partition 的作用。

12、用 linux 实现下列要求：

ip username

a.txt

210.121.123.12 zhangsan

34.23.56.78 lisi

11.56.56.72 wanger

.....

b.txt

58.23.53.132 liuqi

34.23.56.78 liba

.....

a.txt,b.txt 中至少 100 万行。

1）a.txt,b.txt 中各自的 ip 个数，ip 的总个数。

2）a.txt 中存在的 ip 而 b.txt 中不存在的 ip。

3）每个 username 出现的总个数，每个 username 对应的 ip 个数。

13、大意是 hadoop 中 java、streaming、pipe 处理数据各有特点。

14、如何实现 mapreduce 的二次排序。

大数遇到的面试题：

15、面试官上来就问 hadoop 的调度机制；

16、机架感知；

17、MR 数据倾斜原因和解决方案；

18、集群 HA。

@找自己 提供的面试题：

19、如果让你设计，你觉得一个分布式文件系统应该如何设计，考虑哪方面内容；

每天百亿数据入 hbase，如何保证数据的存储正确和在规定的时间里全部录入完毕，

不残留数据。

20、对于 hive，你写过哪些 UDF 函数，作用是什么

21、hdfs 的数据压缩算法

22、mapreduce 的调度模式

23、hive 底层与数据库交互原理

24、hbase 过滤器实现原则

25、对于 mahout，如何进行推荐、分类、聚类的代码二次开发分别实现那些借口

26、请问下，直接将时间戳作为行健，在写入单个 region 时候会发生热点问题，为什么呢？

(三十一) 面试题 17 道：

1、hdfs 原理，以及各个模块的职责

2、mr 的工作原理

3、map 方法是如何调用 reduce 方法的

4、shell 如何判断文件是否存在，如果不存在该如何处理？

5、fsimage 和 edit 的区别？

6、hadoop1 和 hadoop2 的区别？

笔试：

1、hdfs 中的 block 默认报错几份？

2、哪个程序通常与 nn 在一个节点启动？并做分析

3、列举几个配置文件优化？

4、写出你对 zookeeper 的理解

5、datanode 首次加入 cluster 的时候，如果 log 报告不兼容文件版本，那需要 namenode

执行格式化操作，这样处理的原因是？

6、谈谈数据倾斜，如何发生的，并给出优化方案

7、介绍一下 hbase 过滤器

8、mapreduce 基本执行过程

9、谈谈 hadoop1 和 hadoop2 的区别

10、hbase 集群安装注意事项

11、记录包含值域 F 和值域 G，要分别统计相同 G 值的记录中不同的 F 值的数目，简单编写过程。

(三十二) 面试题 3 道：

1、算法题：有 2 个桶，容量分别为 3 升和 5 升，如何得到 4 升的水，假设水无限使用，写
出步骤。
2、java 笔试题：忘记拍照了，很多很基础的 se 知识。后面还有很多 sql 相关的题,常用的
查询 sql 编写，答题时间一个小时。
3、Oracle 数据库中有一个表字段 name，name varchar2(10),如何在不改变表数据的情况
下将此字段长度改为 varchar2（2）？

(三十三) 面试题 6 道：

1.说说值对象与引用对象的区
别？

2.谈谈你对反射机制的理解及其用
途？

3.ArrayList 、 Vector 、 LinkedList 的区别及其优缺点？ HashMap 、 HashTable 的区别及其

优缺点？

3.列出线程的实现方式？如何实现同步？

4.sql 题,是一个图表，具体忘了

5.列出至少五种设计模式？用代码或 UML 类图描述其中两种设计模式的原理？

6.谈谈你最近正在研究的技术，谈谈你最近项目中用到的技术难点及其解决思路。

(三十四) 面试题 1 道：

用户手机号 出现的地点 出现的时间 逗留的时间

111111111 2 2014-02-18 19:03:56.123445 133

222222222 1 2013-03-14 03:18:45.263536 241

333333333 3 2014-10-23 17:14:23.176345 68

222222222 1 2013-03-14 03:20:47.123445 145

333333333 3 2014-09-15 15:24:56.222222 345

222222222 2 2011-08-30 18:13:58.111111 145

222222222 2 2011-08-30 18:18:24.222222 130

按时间排序
期望结果是：

222222222 2 2011-08-30 18:13:58.111111 145

222222222 2 2011-08-30 18:18:24.222222 130

222222222 1 2013-03-14 03:18:45.263536 24

111111111 ~~~~~~~~~

333333333 ~~~~~~~~

(三十五) 面试题 7 道：

Hdfs:

1.文件大小默认为 64M，改为 128M 有啥影响？

2.RPC 原理？

3.NameNode 与 SecondaryNameNode 的区别与联系？

MapReduce:

4.介绍 MadpReduce 整个过程，比如把 WordCount 的例子的细节将清楚（重点讲解

Shuffle）？

5.对 Hadoop 有没有调优经验，没有什么使用心得？（调优从参数调优讲起）

6.MapReduce 出现单点负载多大，怎么负载平衡？（可以用 Partitioner）

7.MapReduce 怎么实现 Top10？

(三十六) 面试题 13 道：

xxxx 软件公司

1.你胜任该职位有什么优势

2.java 优势及原因（至少 3 个）

3.jvm 优化

4.写一个冒泡程序

5.hadoop 底层存储设计

6.职业规划

xxx 网络公司

1.数据库

1.1 第一范式，第二范式和第三范式

1.2 给出两张数据表，优化表（具体字段不记得了，是关于商品定单和供应商方面的）

1.3 以你的实际经验，说下怎样预防全表扫描

2.网络七层协议

3.多线程

4.集合 HashTable 和 HashMap 区别

5.操作系统碎片

6.zookeeper 优点，用在什么场合

7.Hbase 中的 metastore 用来做什么的？

(三十七) 面试题 18 道：

1，在线安装 ssh 的命令以及文件解压的命令？

2，把公钥都追加到授权文件的命令？该命令是否在 root 用户下执行？

3，HadoopHA 集群中哥哥服务的启动和关闭的顺序？

4，HDFS 中的 block 块默认保存几份？默认大小多少？

5、NameNode 中的 meta 数据是存放在 NameNode 自身还是 DataNode 等其他节点？

DatNOde 节点自身是否有 Meta 数据存在？

6，下列那个程序通常与 NameNode 在一个节点启动？

7，下面那个程序负责 HDFS 数据存储？

8， 在 HadoopHA 集群中国 Zookeeper 的主要作用，以及启动和查看状态的命令？

9， HBase 在进行模型设计时重点在什么地方？一张表中国定义多少个 Column Family

最合适？为什么？

10，如何提高 HBase 客户端的读写性能？请举例说明。

11，基于 HadoopHA 集群记性 MapReduce 开发时，Configuration 如何设置

hbase.zookeeper,quorum 属性的值？

12， 在 hadoop 开发过程中使用过哪些算法？其应用场景是什么？

13， MapReduce 程序如何发布？如果 MapReduce 中涉及到了第三方的 jar 包，该如何 处理？

14， 在实际工作中使用过哪些集群的运维工具，请分别阐述期作用。

15， hadoop 中 combiner 的作用?

16， IO 的原理，IO 模型有几种?

17， Windows 用什么样的模型，Linux 用什么样的模型？

18，一台机器如何应对那么多的请求访问，高并发到底怎么实现，一个请求怎么产生的， 在服务端怎么处理的，最后怎么返回给用户的，整个的环节操作系统是怎么控制的？

(三十八) 面试题 11 道：

1.hdfs 的 client 端，复制到第三个副本时宕机，hdfs 怎么恢复保证下次写第三副本?block

块信息是先写 dataNode 还是先写 nameNode?

2.快排现场写程序实现？

3.jvm 的内存是怎么分配原理？

4.毒酒问题---1000 桶酒，其中 1 桶有毒。而一旦吃了，毒性会在 1 周后发作。问最少需要 多少只老鼠可在一周内找出毒酒？

5.用栈实现队列？

6.链表倒序实现？

7.多线程模型怎样（生产，消费者）？平时并发多线程都用哪些实现方式？

8.synchonized 是同步悲观锁吗？互斥？怎么写同步提高效率？

9.4 亿个数字，找出哪些重复的，要用最小的比较次数，写程序实现。

10.java 是传值还是传址？

11.java 处理多线程，另一线程一直等待？

(三十九) 面试题 18 道：

1.一个网络商城 1 天大概产生多少 G 的日志？

2.大概有多少条日志记录（在不清洗的情况下）？

3.日访问量大概有多少个？

4.注册数大概多少？

5.我们的日志是不是除了 apache 的访问日志是不是还有其他的日志？

6.假设我们有其他的日志是不是可以对这个日志有其他的业务分析？这些业务分析都有什 么？

7、问：你们的服务器有多少台？

8、问：你们服务器的内存多大？

9、问：你们的服务器怎么分布的？（这里说地理位置分布，最好也从机架方面也谈谈）

10、问：你平常在公司都干些什么（一些建
议） 下面是 HBASE 我非常不懂的地方：

11、hbase 怎么预分区？

12、hbase 怎么给 web 前台提供接口来访问（HTABLE 可以提供对 HTABLE 的访问，但是

怎么查询同一条记录的多个版本数据）？

13、.htable API 有没有线程安全问题，在程序中是单例还是多例？

14、我们的 hbase 大概在公司业务中（主要是网上商城）大概都几个表，几个表簇，大概 都存什么样的数据？

15、hbase 的并发问
题？ 下面的 Storm 的
问题：

16、metaq 消息队列 zookeeper 集群 storm 集群（包括 zeromq,jzmq,和 storm 本身） 就可以完成对商城推荐系统功能吗？还有没有其他的中间件？

17、storm 怎么完成对单词的计数？（个人看完 storm 一直都认为他是流处理，好像没有 积攒数据的能力，都是处理完之后直接分发给下一个组件）

18、storm 其他的一些面试经常问的问题？

(四十) 面试题 18 道：

1、你们的集群规模？

开发集群：10 台（8 台可用）8 核 cpu

2、你们的数据是用什么导入到数据库的？导入到什么数
据库？ 处理之前的导入：通过 hadoop 命令导入到
hdfs 文件系统

处理完成之后的导出：利用 hive 处理完成之后的数据，通过 sqoop 导出到 mysql 数据库 中，以供报表层使用。

3、你们业务数据量多大？有多少行数据？(面试了三家，都问这个问题)

开发时使用的是部分数据，不是全量数据，有将近一亿行（8、9 千万，具体不详，一般开

发中也没人会特别关心这个问题）

4、你们处理数据是直接读数据库的数据还是读文本数
据？ 将日志数据导入到 hdfs 之后进行处理

5、你们写 hive 的 hql 语句，大概有多
少条？ 不清楚，我自己写的时候也没有做
过统计

6、你们提交的 job 任务大概有多少个？这些 job 执行完大概用多少时间？(面试了三家 ,都 问这个问题)

没统计过，加上测试的，会与很多

7、hive 跟 hbase 的区别是？

8、你在项目中主要的工作任务
是？ 利用 hive 分析数据

9、你在项目中遇到了哪些难题，是怎么解决的？ 某些任务执行时间过长，且失败率过高，检查日志后发现没有执行完就失败，原因出在 hadoop 的 job 的 timeout 过短（相对于集群的能力来说）设置长一点即可

10、你自己写过 udf 函数么？写了
哪些？ 这个我没有写过

11、你的项目提交到 job 的时候数据量有多大？(面试了三家，都问这个问题)

不清楚是要问什么

12、reduce 后输出的数据量有多大？

13、一个网络商城 1 天大概产生多少 G 的日志？ 4tb

14、大概有多少条日志记录（在不清洗的情况下）？ 7-8 百万条

15、日访问量大概有多少个？百万

16、注册数大概多少？不清楚 几十万吧

17、我们的日志是不是除了 apache 的访问日志是不是还有其他的日志？关注信息

18、假设我们有其他的日志是不是可以对这个日志有其他的业务分析？这些业务分析都有 什么？

(四十一) 面试题 1 道：

有一千万条短信，有重复，以文本文件的形式保存，一行一条，
有重复。 请用 5 分钟时间，找出重复出现最多的前 10 条。
分析：

常规方法是先排序，在遍历一次，找出重复最多的前 10 条。但是排序的算法复杂度最低为
nlgn。可以设计一个 hash_table, hash_map<string, int>，依次读取一千万条短信，加载到 hash_table 表中，并且统计重复的次数，与此同时维护一张最多 10 条的短信表。 这样遍历一次就能找出最多的前 10 条，算法复杂度为 O(n)。

(四十二) 面试题 5 道：

1、job 的运行流程(提交一个 job 的流程)？

2、Hadoop 生态圈中各种框架的运用场景？

3、还有很多的选择题

4、面试问到的

hive 中的压缩格式 RCFile、TextFile、SequenceFile 各有什么区别？ 以上 3 种格式一样大的文件哪个占用空间大小..等等

还有 Hadoop 中的一个 HA 压缩。

5、假如：Flume 收集到的数据很多个小文件,我需要写 MR 处理时将这些文件合并（是在 MR 中进行优化,不让一个小文件一个 MapReduce）他们公司主要做的是中国电信的流量计费为主,专门写 MR。

(四十三) 面试题 2 道：

以下题目不必都做完，挑最擅长的即可。

题一：RTB 广告 DSP 算法大赛

请按照大赛的要求进行相应的建模和分析，并详细记录整个分析处理过程及各步骤成果物。

算法大赛主页：http://contest.ipinyou.com/cn/index.shtml

算法大赛数据下载地址：

http://pan.baidu.com/share/link?shareid=1069189720&uk=3090262723#dir

题二：cookieID 识别

我们有 M 个用户 N 天的的上网日志：详见 58.sample

字段结构如下：

ip string 客户端 IP

ad_id string 宽带 ADSL 账号

time_stamp string 上网开始时间

url string URL

ref string referer

ua string User Agent

dest_ip string 目标 IP

cookie string cookie

day_id string 日期

58.com 的 cookie 值如：

bangbigtip2=1; bdshare_firstime=1374654651270;

CNZZDATA30017898=cnzz_eid%3D2077433986-1374654656-http%253A%252F%252Fsh.58.com

%26ntime%3D1400928250%26cnzz_a%3D0%26ltime%3D1400928244483%26rtime%3D63;

Hm_lvt_f5127c6793d40d199f68042b8a63e725=1395547468,1395547513,1395758399,13957594

68; id58=05dvZ1HvkL0TNy7GBv7gAg==;

Hm_lvt_3bb04d7a4ca3846dcc66a99c3e861511=1385294705;

__utma=253535702.2042339925.1400424865.1400424865.1400928244.2;

__utmz=253535702.1400424865.1.1.utmcsr=(direct)|utmccn=(direct)|utmcmd=(none); city=sh;

pup_bubble=1; __ag_cm_=1400424864286; myfeet_tooltip=end; ipcity=sh%7C%u4E0A%u6D77

其中有一个属性能标识一个用户，我们称之为 cookieID。

请根据样例数据分析出 58.com 的 cookieID。

要求详细描述分析过程。

(四十四) 面试题 7 道：

1、解释“hadoop”和“hadoop 生态系统”两个概念。

2、说明 Hadoop 2.0 的基本构成。

3、相比于 HDFS1.0, HDFS 2.0 最主要的改进在哪几方面？

4、试使用“步骤 1，步骤 2，步骤 3.....”说明 YARN 中运行应用程序的基本流程。

5、“MapReduce 2.0”与“YARN”是否等同，尝试解释说明。

6、MapReduce 2.0 中，MRAppMaster 主要作用是什么，MRAppMaster 如何实现任务 容错的？

7、为什么会产生 yarn,它解决了什么问题，有什么优势？

(四十五) 面试题 6 道：

1、集群多少台,数据量多大,吞吐量是多大,每天处理多少 G 的数据？

2、自动化运维了解过吗,你们是否是自动化运维管理？

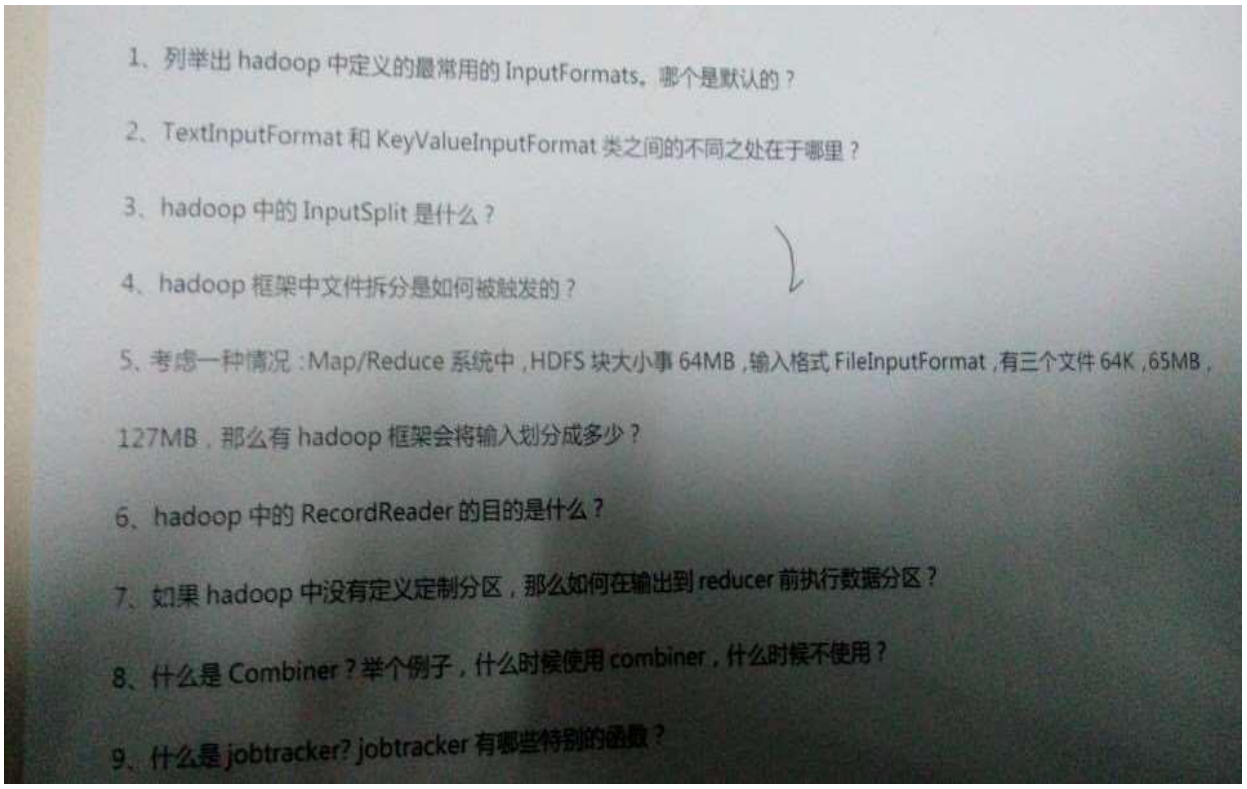
3、数据备份,你们是多少份,如果数据超过存储容量,你们怎么处理？

4、怎么提升多个 JOB 同时执行带来的压力,如何优化,说说思路？

5、你们用 HBASE 存储什么数据？

6、你们的 hive 处理数据能达到的指标是多少？

(四十六)　面试题 18 道：



1、列举出 hadoop 中定义的最常用的 InputFormats，哪个是默认的？

2、TextInputFormat 和 KeyValueInputFormat 类之间的不同之处在于哪里？

3、hadoop 中的 InputSplit 是什么？

4、hadoop 框架中文件拆分是如何被触发的？

5、考虑一种情况：Map/Reduce 系统中，HDFS 块大小事 64MB，输入格式 FileInputFormat，有三个文件 64K，65MB，127MB，那么有 hadoop 框架会将输入划分成多少？

6、hadoop 中的 RecordReader 的目的是什么？

7、如果 hadoop 中没有定义定制分区，那么如何在输出到 reducer 前执行数据分区？

8、什么是 Combiner？举个例子，什么时候使用 combiner，什么时候不使用？

9、什么是 jobtracker? jobtracker 有哪些特别的函数？

(四十七) 1 道题目

Hadoop

1、在 Hive 中有两张表：

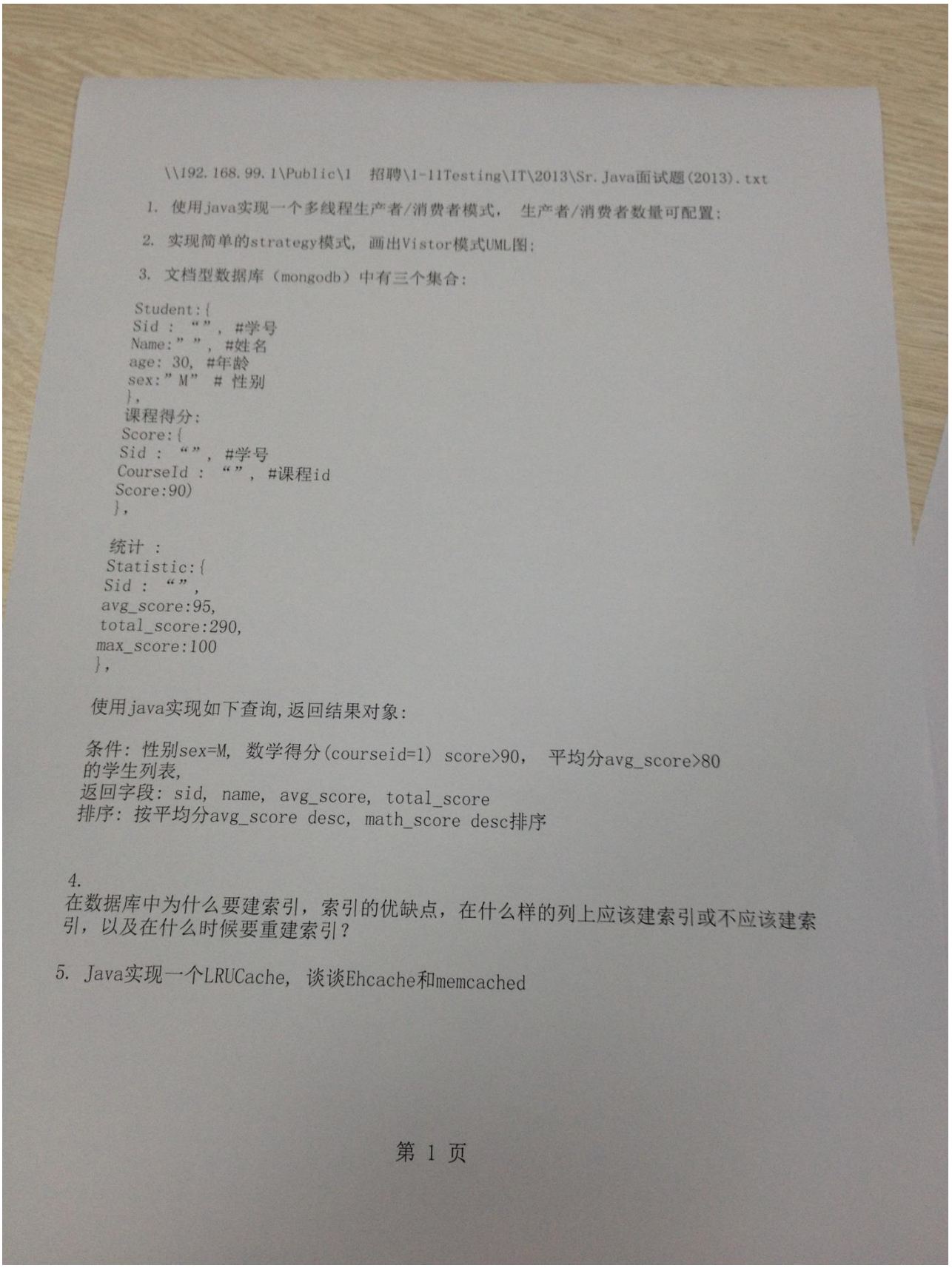
table_a(id int,name string,ip string)

table_b(id int,name string)

问：（1）select a.id from table_a a where a.id not in(select id from b)

这条语句在 hive 中能否执行？（2）如果不能如何改写？

(四十八) 5 道面试题



\\192.168.99.1\Public\1 招聘\1-11Testing\IT\2013\Sr.Java面试题(2013).txt

1. 使用java实现一个多线程生产者/消费者模式， 生产者/消费者数量可配置;

2. 实现简单的strategy模式, 画出Vistor模式UML图;

3. 文档型数据库（mongodb）中有三个集合:

```
Student:{
Sid : “”, #学号
Name:””, #姓名
age: 30, #年龄
sex:”M” # 性别
},
课程得分:
Score:{
Sid : “”, #学号
CourseId : “”, #课程id
Score:90)
},

统计 :
Statistic:{
Sid : “”,
avg_score:95,
total_score:290,
max_score:100
},
```

使用java实现如下查询,返回结果对象:

条件: 性别sex=M, 数学得分(courseid=1) score>90， 平均分avg_score>80
的学生列表,
返回字段: sid, name, avg_score, total_score
排序: 按平均分avg_score desc, math_score desc排序

4.
在数据库中为什么要建索引，索引的优缺点，在什么样的列上应该建索引或不应该建索引，以及在什么时候要重建索引？

5. Java实现一个LRUCache, 谈谈Ehcache和memcached

第 1 页

(四十九) 19 道题目

```
MR
1.有关于MR二次排序的问题
2.全局排序
3.去重
4.非等值连表查询
5.调度器
6.怎么定于多路径输出（一个reduce输出一个路径）
7.chain map/chain reduce
8.内置计数器
9.distuabute cache的用法


HIVE
1.有关于UDF、UDAF
2.SQL优化
3.大小表
4.hive全局排序
5.外部表、内部表
6.hive脚本
7.调度器
8.hive支持的格式：
9.支持哪些压缩存储：1、TEXTFILE
                     2、SEQUENCEFILE
                     3、RCFILE   (***)
                     4、ORCFILE
10.动态分区、静态分区
```

7、如果 hadoop 中没有定义定制分区，那么如何在输出到 reducer 前执行数据分区？

8、什么是 Combiner？举个例子，什么时候使用 combiner，什么时候不使用？

9、什么是 jobtracker? jobtracker 有哪些特别的函数？

12、什么是 tasktracker?

13、hadoop 中 job 和 task 之间是什么关系？

14、假设 hadoop 一个 job 产生了 100 个 task，并且其中的一个 task 失败了，hadoop 会如何处理？

15、通过划分多个节点上任务，hadoop 实现了并行处理，对少数慢节点可能会限制剩下其他程序的速率，并拖慢了整个程序。Hadoop 提供了什么机制防止这种情况的发生？

16、hadoop 推测执行是如何实现的？

17、Unux 中使用命令行，如何查看 hadoop 集群中的所有运行的任务？或是 kill 掉任务？

18、什么是 hadoop streaming?